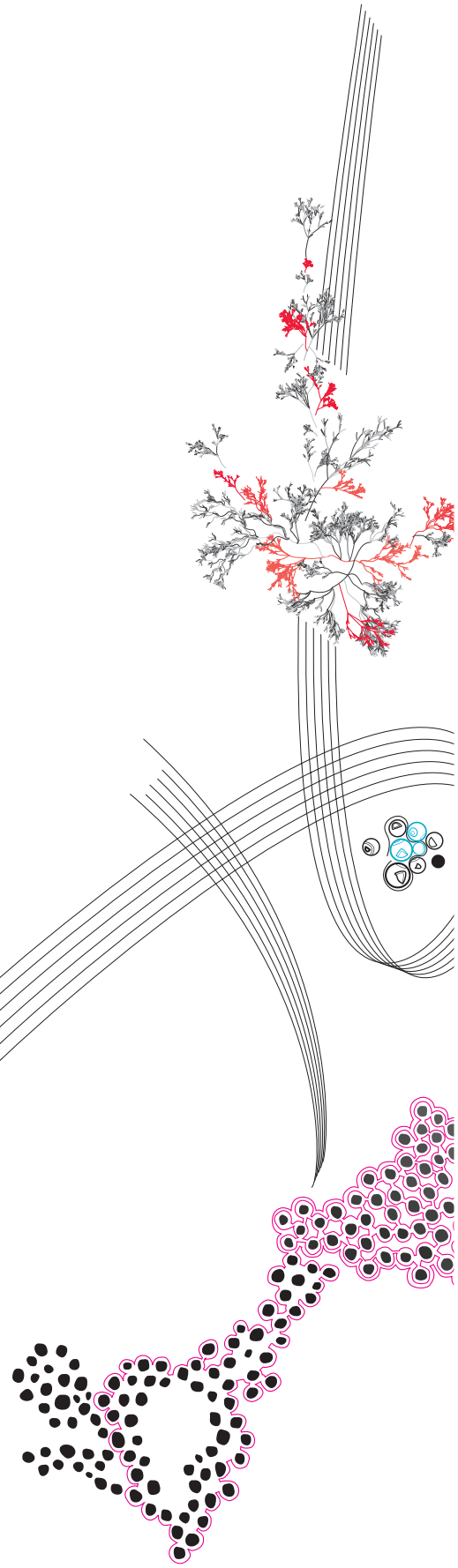


BSc Thesis Industrial Engineering and Management

Data quality analysis to support carbon footprinting

C.D.G. Maessen

Supervisor: R.L.A. Harmelink
Second supervisor: M.R.K. Mes
Company supervisor: W. Harbers

August, 2024

Department of Industrial Engineering and Management
Faculty of Behavioural, Management and Social Sciences

UNIVERSITY OF TWENTE.

Abstract

This thesis analyses the data quality of HST regarding a BigMile carbon analysis. BigMile is a tool which maps the carbon footprint of a company. With a report of the carbon footprint becoming a requirement for companies it is important to investigate the data quality to give an accurate insight.
In this thesis, the Managerial Problem Solving Method (MPSM) is used. The actions in the research are that the data quality requirements are set, data quality dimensions are selected, the data set is cleaned and the data set is analysed for possible problems. Queries are created with Excel Query to solve some of the issues and improve the data quality.
Without using the Excel Queries the data sets were not suitable for the Lean & Green input file and therefore not suitable for a BigMile analysis. When using the designed Excel Queries a maximum data quality level of Silver can be achieved. This could result in a maximum of three Lean & Green Stars when only taking data quality into account. The analysis revealed several significant issues such as the lack of data regarding loading metres, hidden information on partial shipments in the exports and missing vehicle IDs. These findings are needed for improving the data quality to give an accurate overview of the current, and past, carbon footprint of the company.
It is recommended to delve into the accuracy of loading metres, develop strategies to gather the data needed regarding Vehicle IDs and explore strategies to link HST Group datasets straight from the database to the BigMile analysis format. These possible actions could result in an improvement in the data quality to the Gold data quality level.

Management summary

This management summary gives the key findings and recommendations derived from the analysis conducted in the thesis. The purpose of this section is to provide a quick insight into the study.

Objective:
The objective of this thesis was to perform a data quality analysis of the HST group data sets regarding performing a BigMile dashboard analysis. Through research and analysis, possible problems were brought to attention. To solve a lot of the problems, and improve the data quality, queries were created. An overview of the current state, the state after using the queries, and possible future levels of the data quality are created within the boundaries of the data that is needed for the BigMile analysis.

Methodology:
Our approach involved the use of the Managerial Problem Solving Method (MPSM), which gives systematic guidance during the research. Giving structure during the research and writing of the thesis. The data analysis and data improvements are done with the help of Excel Query.

Key Findings:
Without using the Excel Queries the data sets were not suitable for the Lean & Green input file and therefore not suitable for a BigMile analysis. When using the designed Excel Queries a maximum data quality level of Silver can be achieved. This could result in a maximum of three Lean & Green Stars when only taking data quality into account. The analysis revealed several significant issues such as the lack of data regarding loading metres, hidden information on partial shipments in the exports and missing vehicle IDs. These findings are needed for improving the data quality to give an accurate overview of the current, and past, carbon footprint of the company.

Recommendations:
It is recommended to delve into the accuracy of loading metres, develop strategies to gather the data needed regarding Vehicle IDs and explore strategies to link HST Group datasets straight from the database to the BigMile analysis format. These possible actions could result in an improvement in the data quality to the Gold data quality level.

Challenges and Limitations:
It is crucial to address the limitations imposed by the time constraints of the study. The time-gated nature of the research restricted the possibility of verifying many values in the data sets such as postal codes and loading metres. These are both a study worth it on their own. There may be multiple other solutions possible for the problems presented. These aspects should be considered in the interpretation of results and in planning for future research.

In conclusion, this management summary provides the essence of the thesis. The findings and recommendations presented offer guidance for decision-making and provide a basis for future efforts to improve the data quality regarding mapping the carbon footprint of HST Group.

Preface

I am pleased to present this thesis as the result of the research and data analysis.

This thesis delves into the data quality of HST Group regarding a BigMile analysis. This analysis needs to be done to map the current and past carbon footprint of HST Group. To improve the data quality of data sets queries were created using Excel Query. The research journey has been challenging and there were a lot of hurdles, but the result is finally presented in front of you, the reader.

I would like to thank my primary supervisor, Rogier Harmelink, whose guidance has been invaluable for this study. Secondly, I would like to thank my company supervisor at HST Group, who has guided me well my during the research at HST. He understood the situations that came up during the research and made me feel welcome at the company. Lastly, I would like to thank my second supervisor Martijn Mes, who has helped me improve my thesis with his critical, constructive feedback.

With this thesis, my student time at the University of Twente will be finished. The years at the university have shaped the academic fundamentals for the rest of my career. As the pages lie before you I hope this work provides a meaningful contribution towards the challenges and current situation regarding data quality for mapping the carbon footprint of HST Group. May it spark further curiosity and inspire future research in the field of data quality in the transport sector.

Cédric Maessen
Januari 2024, Enschede

Contents

1 Introduction

1.1 Situation of HST Group

HST Group was founded due to the merger of three local transport companies – Haarman, Smit, and Thijssen – from Twente, the Netherlands in 1978. (HSTGroup, 2023) The company continued growing after its foundation and started expanding its transportation services internationally in the 2000s with the establishment of HST International in 2001 and HST Sea & Airfreight in 2005. Thus the company names itself an "all-round logistics service provider" as it provides not only (inter)national road transportation but also ocean cargo and air freight. Over the last few years, HST Group started investing in sustainability to reduce its carbon footprint. To improve their environmental performances they now use ISO 14001:2015, an international standard for environmental management systems, and have achieved a Lean & Green First Star, which is a result of reducing carbon emissions. The importance of environmentally responsible business is a well-discussed topic in today's society, therefore this study may apply to more companies in the transport sector.

1.2 Lean and Green program

Lean & Green (Connekt, 2023) is a leading CO2-reduction program in which companies in the logistical sector can participate. The program has the goal of reducing carbon emissions within the logistical sector structurally through active collaboration, sharing knowledge and deploying smart and effective measures. Lean & Green helps towards sustainable solutions and efficient logistical processes. Lean & Green helps with analysing the current and past carbon footprint of the companies with the help of a BigMile analysis. BigMile is a company which has produced a tool which helps in analysing the carbon footprint of a company. The analysis this tool produces is referred to in this thesis as a *BigMile analysis*.

1.3 Problem cluster

We start with defining the problem with the help of a problem cluster (Figure 2). HST Group is, and has been, investing in the reduction of its carbon footprint. To visualise its carbon footprint, HST Group is collaborating with Lean & Green from Connekt. On the 30th of December 2008, HST Group was awarded with the Lean & Green Award. This means conducting a baseline measurement and setting up an action plan that includes the scope and planned measures for carbon emission reduction (Connekt, 2023).

On the 15th of May 2013, the first star of the Lean & Green project was awarded to HST Group. This means a carbon emission reduction of 20% within a period of a maximum of 5 years has been achieved.

Since then, the Lean & Green project has been on hold at HST Group. The second Lean & Green Star was still in development at the time. The company BigMile, which produced the carbon footprint analysis, was founded in 2019. In a meeting in 2022, HST Group decided that they would like to continue with Lean & Green and have set the goal of a second Lean & Green star. While not tracking the progress, HST Group has invested in multiple carbon emission reduction measures over the past years. A few of them are acquiring LNG trucks, application of Long Heavy Vehicles (LHVs), tyre pressure management systems on all trailers and solar panels on their buildings.

HST Group has also made a switch in their Transport Management System (TMS). This led to a period where two systems were in use. This is visible in the Figure 1. The switch for the TMS for international shipping was on the 18th of March 2019. The switch for the TMS for national shipping was on the 16th of November 2021.

International Ritrov	Q1	2019	2020	2021	2022	2023	Ritrov (Archive)
TransPas International	Q1	2019	2020	2021	2022	2023	TransPas International
Transmission (in use)	2019		2020	2021 Q1-3	Q4 2022	2023	Transmission (Archive)
TransPas (Full usage)					Q4 2022	2023	TransPas (Full usage)

FIGURE 1: Overview of TMS

To achieve the second Lean & Green star, HST Group must meet a few requirements:

- Improving the quality of the data.
- Widening of the scope of the outgoing volume to a minimum of 65%.
- Active collaboration within the transport chain.
- Additional minimal 10% CO2 reduction over a maximum period of 3 years.
- Explanation of how the CO2 reduction is met.

The norm is that HST Group qualifies for a second Lean & Green Star. However, the reality is that they have only achieved one Lean & Green Star thus far. This action problem is at the top of Figure 2.

For choosing a core problem a problem cluster is used.

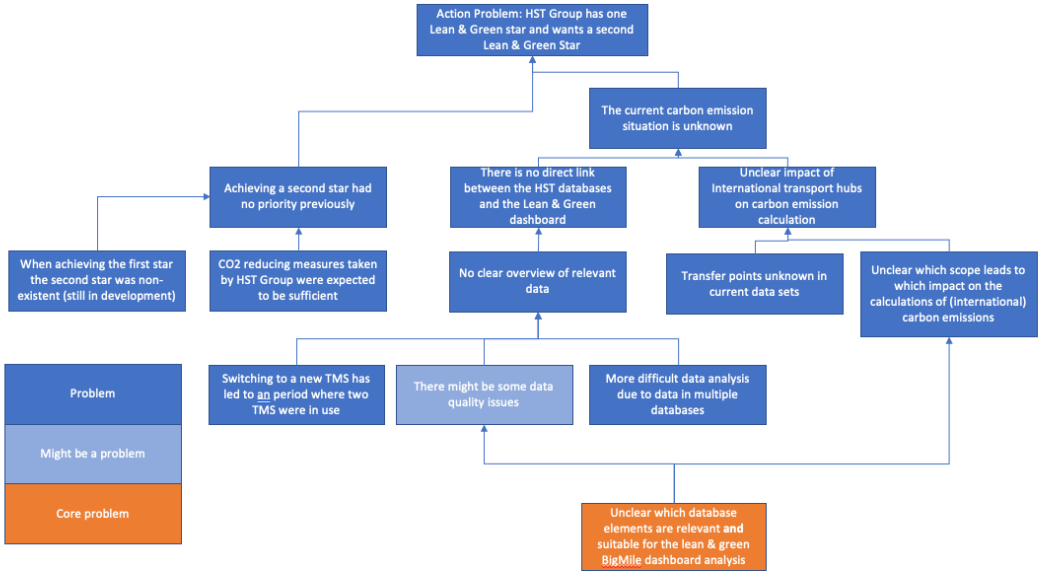


FIGURE 2: Problem Cluster

From the action problem, multiple underlying problems are derived. The first cause of why HST Group has not achieved the second Lean and Green Star yet is *"Achieving the second Lean & Green star had no priority previously"*. This is because the moment when HST Group achieved the first Lean and Green Star, a decade ago, the second Lean and Green Star was still under construction. The company also invested in greener energy

and energy-saving projects without the numerical incentive from Lean and Green reports, therefore the need to invest in the Lean & Green project was not relevant at the time.

The second cause is *"The current carbon emission situation is unknown"*. Since the current carbon emission situation is unknown, HST Group has no insight into how much they have reduced carbon emissions relatively. For this, there are multiple causes:

- *"There is no link between HST databases and the Lean and Green dashboard."* As there is no connection between the databases and the dashboard, an employee needs to be tasked with configuring and fitting the data such that it can be used in the BigMile analysis. The underlying cause is that it is unknown which data is relevant and suitable for the BigMile analysis.

- *"Unclear impact of international transport hubs on carbon emission calculation."* There are routes where HST Group does a part of the shipment, and the other part is done by another company. The total fuel consumption per month is known, however, the distance travelled to the transfer location and the number of goods is unknown. This leads to the scope needing to be determined to understand the impact on the carbon footprint since the fuel consumption and distance travelled by the contracted company are unknown.

- The underlying cause for the problems mentioned above is *"Unclear which database elements are relevant and suitable for the Lean & Green BigMile analysis."* This is the core problem that is tackled in this research. It must be known which elements in the databases can be used in the calculations (relevance) and what the data quality of the relevant data is (suitable). To know which data is suitable, a data quality analysis of selected elements in the databases needs to be performed.

Following the Managerial Problem-Solving Method (MPSM), a core problem cannot have an underlying problem itself and a core problem can be influenced (by the researcher). When more than one problem remains, the problem in which solutions have the greatest impact should be chosen (Heerkens & van Winden, 2017). This results in the core problem *"it is unclear which database elements are relevant and suitable for the Lean & Green BigMile analysis"*. To determine which data is relevant and suitable, research regarding data quality should be done with the main research question: *"What is the data quality of the HST Group database to perform a BigMile analysis?"*

1.4 Research methodology

For this research, a problem-solving method needs to be chosen. In this research, we follow the guidelines of the MPSM. This is a structured method for solving action problems. Also, this method has options for solving knowledge problems and backtracking. Following Heerkens & van Winden (2017) the seven phases of the Managerial Problem-Solving Method are:

1. Defining the problem
2. Formulating the approach
3. Analysing the problem
4. Formulating (alternative) solutions
5. Choosing a solution

6. Implementing the solution

7. Evaluating the solution

These stages are evident in the thesis, mirroring a structured progression. Chapter 1 defines the problem (1) and outlines the approach (2) with the help of research questions. Chapter 2 provides the foundation for the step analysing the problem (3) and the knowledge is provided in order to analyse the problem. Chapter 3 describes the current situation of HST Group, while in Chapter 4 the analyses of the current database, in the context of a BigMile analysis, are given. In Chapter 5 the consequences of the analysis are presented. A potential solution (4, 5) is presented in Chapter 6 to solve a few of the problems presented. In Chapter 7 other solutions to other problems are given. In Chapter 6 one solution is also implemented (6). In Chapter 6.3.5 the evaluation (7) of the data quality level with the implemented solution is presented. Chapter 8 covers the conclusion and discussion.

1.4.1 Research goal

The research objective is to analyse the data of HST Group in order to determine the data quality and describe possible data problems for the use of a BigMile dashboard analysis.

1.4.2 Research questions

In this section, the research questions are given:

- RQ: What defines data quality?
  The definition of data quality is needed to contextualize the research in the case of a BigMile analysis. In Chapter 2.1 the definition of data quality is given.
  - What are the advantages of high data quality?
    The level of data quality can have an impact on the operations of a company. By investigating what the positive results of high-quality data are, the relevance of high-quality data in regards to achieving a second Lean & Green star for HST Group can be revealed. These results are discussed in Chapter 2.2.
  - What are the challenges of data quality in big data?
    The challenges of data quality in big data need to be known to prevent complications when working with large amounts of data as much as possible in the research. These challenges are discussed in Chapter 2.3.
- RQ: According to literature, what are the data quality dimensions with the goal of a BigMile analysis?
  Selecting the relevant data quality dimensions is needed to determine the data quality level and give insight into possible problems. Selecting not relevant dimensions results in unwanted work and not selecting relevant dimensions results in an inaccurate conclusion. These dimensions are given in chapter 2.4.1 and 2.4.2.
- RQ: Which data is selected for the data quality analysis for HST Group?
  Knowing which data is included in the research and which data is out of scope is crucial for the data analysis. Chapter 3 provides guidelines for the needed scope.
  - What are the changes in the Transport Management Systems over the past years?
    Changes in the Transport Management Systems could complicate the gathering

of the needed data. An overview of the periods of the chosen data is elaborated upon in Chapter 3.1.

- RQ: What is the data quality of the relevant data?
  Knowing the data quality level of the selected data is needed to answer the main research question.
  - Which data inputs are needed for the BigMile analysis?
    The Lean & Green program requires multiple inputs for the BigMile analysis. A complete overview of the inputs is needed in order to provide a relevant data quality analysis. These data inputs are given in Chapter 3.3.
  - What data quality level does the Lean & Green program require?
    The Lean & Green program requires different data quality levels per Lean & Green star. The Lean & Green program provides the needed requirements for the current goal but also sets standards for future goals. These requirements are shown in Chapter 3.4.
  - Which data for BigMile is present in which HST Group data set?
    Knowing which data can be found in which data set is needed to provide reliability for the research, meaning other researchers could do the research again if they wanted. The relationship between the different files can be found in Chapter 3.2.

  RQ: What improvements need to be made to the data?
  - What data quality improvements need to be made to achieve the second Lean & Green star?
    By analysing the data quality it is needed to provide insight into whether there are data quality improvements needed to comply with the data quality standards set by the Lean & Green program for the BigMile analysis.
  - What data quality improvements can be made to prepare HST Group for the future?
    It is important to provide insight and give advice on the possible improvements to prepare HST Group for future requirements. This ensures that the data quality requirements can be met for possible future BigMile analyses.

First, the dataset from HST Group is accessed to know which data is available. After this, the research questions are answered with the help of literature research. When it is clear how to analyse data quality in the context of a BigMile analysis, then the data sets from HST Group are analysed. Problems found during the analysis are noted to give a complete overview of the data quality regarding a BigMile analysis for HST Group. This makes it easier to possibly implement changes in the future. When the analysis is done, the found problems are listed and possible solutions are given. These are concluded and discussed in Chapter 8.

1.4.3 Data collection

The data collection method employed in this research primarily relies on document analysis, specifically during the literature review phase. The research draws on existing literature and internal data sources to derive insights and support its objectives. It is important to note that no new data is generated during this study; instead, the analysis focuses on interpreting information from previously existing documents and data sets.

Sensitive data, particularly internal organisational data, is handled with care to ensure confidentiality and compliance with privacy standards. Access to sensitive data is restricted to the researcher via Microsoft Remote Desktop, a secure platform that ensures data remains within the confines of the HST Group organisation. This approach is crucial in safeguarding sensitive information and preventing unauthorized access.

To further enhance data security, any data stored on the researcher's laptop, even if temporarily, is deleted after the research concludes. This precautionary measure is implemented to mitigate the risk of data breaches and unauthorized access to sensitive information.

Moreover, the research takes a proactive stance in protecting the privacy of the company's clients. The data presented in the final report is anonymized and summarized to prevent any tracing back to specific clients of HST Group. This anonymization process is crucial in maintaining confidentiality and adhering to ethical considerations in research.

To correctly analyse the collected data, data quality theory is needed. In Chapter 2 the background information is given on how the data quality is determined.

2 Data quality theory

In order to assess the data quality of HST Group, in this case for the BigMile dashboard analysis, data quality theory is needed. In the first section, the definition of data quality is given. The two sections describe the advantages and challenges of data quality. At the end of the chapter, the data quality dimensions are discussed.

2.1 Defining data quality

According to Tayi & Ballou (1998) the term *data quality* can best be defined as *fitness for use*, implying that the concept of data quality is relative (Juddoo, 2015). In the case of HST Group and the Lean & Green project *data quality* is determined by what the fitness is for a BigMile analysis: "Which data is needed to perform the BigMile analysis?" In Table 2 the criteria for the data quality are given.

2.2 Advantages of data quality

Data quality, which refers to the dimensions of the data mentioned in the next chapter, offers several advantages:

> sInformed decision-making: High-quality data leads to decisions based on accurate and reliable information (Janssen et al., 2017). Trust and credibility: Organisations that maintain accurate and reliable data (high-quality data) build confidence for their customers (McGilvray, 2021). Operational efficiency: High-quality data reduces errors and inconsistencies in processes. It minimises the need for manual rework and manual data corrections (Batini et al., 2009).

For HST Group this means regarding informed decision-making that they can request a next Lean & Green star when they have the required decrease of their carbon footprint. They can also invest more in sustainability if the figures point to not enough carbon emission reduction.

Regarding operational efficiency, high-quality data ensures that HST Group does not need to invest man-hours into the Lean & Green analyses as low-quality data requires a lot of manual labour. Not only is the cost of labour an unwanted factor, but also the delay manual labour induces in the processes is unwanted.

2.3 Challenges of data quality

The data sets from HST Group are not defined as big data, however, due to the time limitations and more than one billion values in the data, big data challenges are also occurring when working with this amount of data. Cai and Zhu present the challenges of data quality in big data as listed below (Cai & Zhu, 2015, p. 2-3). For each item, the case of HST Group is described.

- *The diversity of data sources brings abundant data types and complex data structures and increases the difficulty of data integration.* For HST Group this relates to the different TMS systems in use, which complicates the analysis.
- *Data volume is tremendous, and it is difficult to judge data quality within a reasonable amount of time.* In the case of the research there are more than a billion values that need to be analysed. This makes it difficult to get a quick and accurate judgement.

- *Data change is very fast and the "timeliness" of data is very short, which necessitates higher requirements for processing technology.* Currently, this is not a problem for HST Group. In the future when near real-time analysis is expected this would induce problems regarding timeliness.

- *No unified and approved data quality standards have been formed in China and abroad, and research on the data quality of big data has just begun.* For HST Group this means, outside the data quality standards that are explained in Chapter 3.4, there are no standards set with how much data can be inaccurate in order to reach a certain level. For this research the levels described in chapter 3.4 are used.

2.4 Data quality dimensions

In Juddoo (2015) the author explains that the execution part of a data quality strategy consists of activities whose purpose is to improve the data quality based upon dimensions based on a specific context. Such activities are for example data profiling, data cleansing and constructing data quality rules. In Chapter 2.4.1 the dimensions for the research are given. The dimensions are selected from Cai & Zhu (2015) and McGilvray (2021). The dimensions that are not discussed in the research can be found in Appendix A. In Chapter 2.4.2 dimensions relevant to HST Group are briefly touched upon, however, they are not analysed.

2.4.1 Measured dimensions

The following dimensions are chosen to be included in the research. For each dimension, the definition and reason to be included in the research are given.

- **Dimension:** Integrity (Cai & Zhu, 2015)
  Definition: In a database, data with integrity are said to have a complete structure. Data values are standardised according to a data model and/or data type.
  Reason: The integrity of the data can be assessed. For example, if a loading metre value is within the range of what fits in a trailer.

- **Dimension:** Completeness (McGilvray, 2021)
  Definition: A characteristic of information quality that measures the degree to which there is a value in a field; synonymous with fill rate.
  Reason: Data needs to be complete in order to give an accurate description.

- **Dimension:** Uniqueness and deduplication (McGilvray, 2021)
  Definition: The uniqueness (positive) or unwanted duplication (negative) of data (fields, records, or datasets) existing within or across systems or data stores.
  Reason: This has an impact on how accurately the data describes the real-world situation.

2.4.2 Generally discussed dimensions

The following dimensions are briefly addressed hereby, but are not delved into in the research:

- **Dimension:** Accessibility (Cai & Zhu, 2015) (Wang & Strong, 1996)/ Access (McGilvray, 2021)
  Definition: The level of difficulty for users to obtain data.

Reason: There is a limited export function for the data in the TMS software. Therefore the time it takes to obtain the data is quite a lot. However this dimension is a problem for all fields, therefore it is only discussed generally. If HST Group wants to implement a more permanent solution they can also link a script to the database directly to create the needed files.

- **Dimension:** Timeliness (Cai & Zhu, 2015) (McGilvray, 2021)
  Definition: The time delay from data generation and acquisition to utilisation.
  Reason: As we are using a data set that is more than a year old *Timeliness* is not discussed. If in the future HST Group wants more recent updates *Timeliness* is more important. When the order is delivered the order should be updated that day in the system.

- **Dimension:** Consistency and synchronization (McGilvray, 2021)
  Definition: Equivalence of data stored or used in various data stores, applications, and systems.
  Reason: The datasets of 2022 are easily linked. The fuel usage, location energy usage, TMS national and TMS international need to be linked (copy-pasted) with an input file. Therefore there could be improvement but this has no impact on the other dimensions.

- **Dimension:** Auditability (McGilvray, 2021)
  Definition: Auditors can fairly evaluate data accuracy and integrity within rational time and manpower limits during the data use phase.
  Reason: Each auditor has another opinion on how easily data can be evaluated.

- **Dimension:** Readability (Cai & Zhu, 2015)/ Presentation Quality (McGilvray, 2021) / Concise (Wang & Strong, 1996)
  Definition: The format, appearance, and display of data and information support their collection and uses.
  Reason: This has an impact on the visual result, however it has no results on the carbon analysis in BigMile. It is important for implementation across the workplace so that the employees can use the data. If in the future HST Group wants an up-to-date file, then the *readability* becomes more important, as then employees have to find accurate information quickly.

- **Dimension:** Accuracy (McGilvray, 2021) (Cai & Zhu, 2015) (Wang & Strong, 1996)
  Definition: The correctness of the content of the data as compared to an agreed-upon and accessible authoritative source of reference.
  Reason: The accuracy of the data is always important when discussing data quality. However, the accuracy of the data is difficult to verify. Each measurement requires separate research. E.g. for loading metres, you have to measure the packages again and compare the values with the values in the data sets to provide an insight into the accuracy.

3 Current situation

In order to understand the relevance of the data and data dimensions given in the previous chapter, it is important to be aware of how to place them in the context of this research. Therefore this chapter will discuss the current situation of HST Group.

3.1 Time periods

During the research, multiple years and Transport Management Systems were analysed. For the department HST International, the switch from Ritrov to TransPas was made on the 18th of March 2019. For the department HST National, the switch from TransMission to TransPas was made on the 16th of November 2021. The Lean & Green project, and therefore also the program for a second Lean & Green star, requires a base year and an evaluation year. A base year and an evaluation year are needed to calculate the relative carbon emission reduction. The base and evaluation year are chosen such that only one Transport Management System is used for national or international shipping. For the base year July 2019 till June 2020 is chosen. For the evaluation year, 2022 is chosen. These periods were also chosen due to the implementation of an active Tire Pressure Monitoring System (TPMS) during the second half of 2020 till the first half of 2021, which is one of the measures to reduce carbon emissions. It checks and adjusts the tyre pressure of every trailer actively. In the rest of the research, the year 2022 is mainly mentioned as this is the year which gives the most accurate overview of the data quality of the systems in use. The data operations done on the data sets from the period 2019-2020 are explained in Chapter 6.1.1, Chapter 6.1.3 and Appendix C.

3.2 Data overview HST Group

This chapter explains which exports there are and how the exports were gathered. HST Group has a few different datasets that are exported from different platforms. Currently, no files are linked together. Therefore from each file data has to be gathered manually.

- TransPas National 2022 export to an Excel sheet, created by the IT department. See Figure 5.
- TransPas International 2022 and 2019-2020 export to an Excel sheet, created by the IT department. TransPas (International) is the same program as TransPas (National). However, both TransPas versions do not have the same fields. For the overview in Table 1 both versions are put together as the input values that can be found in the exports are the same. See Figure 3 and Figure 5 for an overview.
- Fuel usage exports to Excel sheets. The exports are created by the Bookkeeping department. From exports of BP and AVIA, an overview is created per fuel type per month. Which can be used as input in BigMile. See Figure 4.
- Energy usage export to Excel sheet. The exports are created by the IT department. An overview is created of the energy usage and solar energy produced per location per month. See Figure 4.

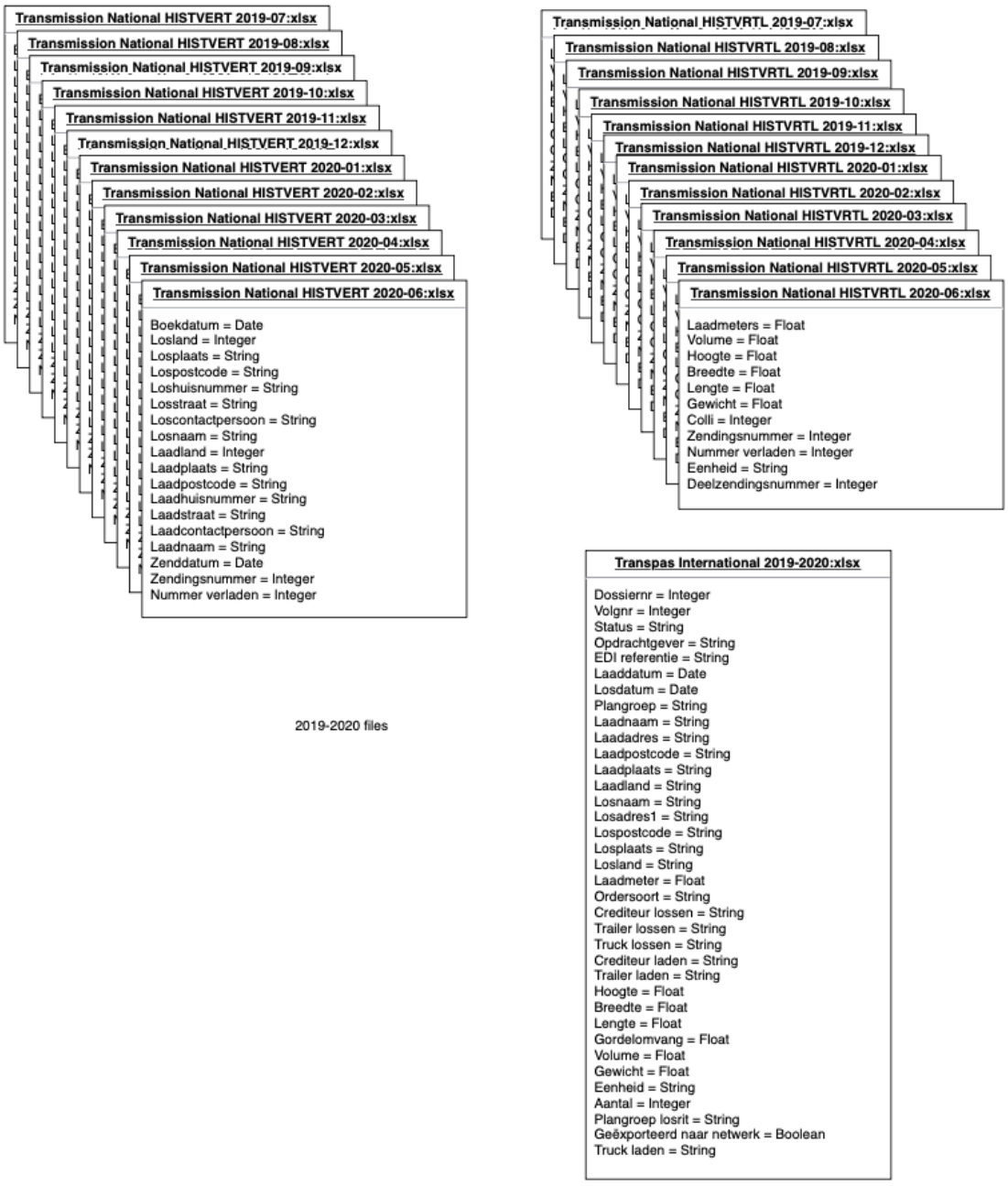


FIGURE 3: TMS files 07-2019 / 06-2020

In Figure 3 the different file exports and back-ups needed for the period 07-2019 to 06-2020 are visible. There are in total 25 different files needed to get all the data from the TMS. The HISTVERT and HISTVRTL originate from back-ups. Transpas International is exported from the TMS. These files are currently not linked.

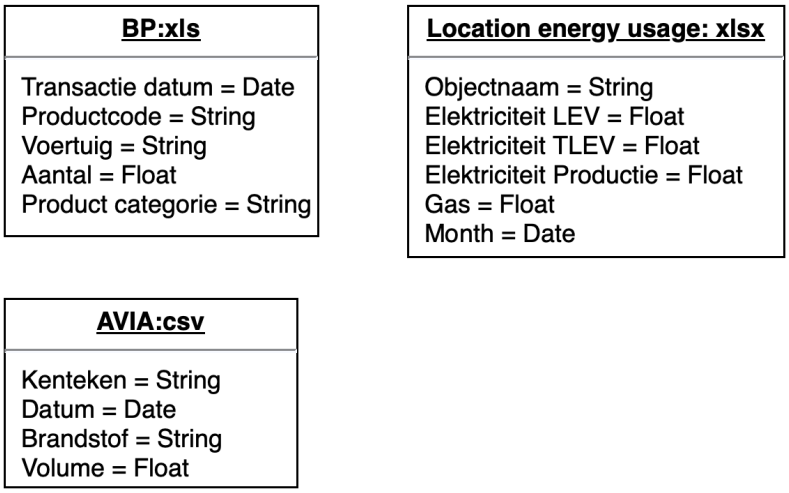


FIGURE 4: Energy and fuel file overview per year

To get all the data for fuel and energy consumption, three files are needed per year. These three files are visible in Figure 4. These files are currently not linked.

FIGURE 5: TMS file overview 2022

In Figure 5 the different files needed for the period year 2022 are visible. These files are currently not linked.

FIGURE 6: Inputfile overview

In Figure 6 an overview of which inputs the input-file needs is visible. For the BigMile analysis, the relevant data needs to be inserted into the input file.

3.3 Inputs for BigMile

In Table 1 an overview of the different data sets from HST Group for the year 2022 is visible. For each data set the possible inputs for BigMile are given. This does not mean that the field/input is complete and accurate, but it means the field is in the data set. The whole list of inputs is required as input for the BigMile tool, the carbon footprint analysis. The overview of inputs for the BigMile/Lean & Green tool is also visible in Figure 6.

Input	TransPas	Fuel	Energy
Tab "Shipments"			
• Order number, e.g. "135156 - 1"	X		
• Serial number, e.g. "1"	X		
• *Carrier*, e.g. "HST" or "Netwerk"	X		
• *Vehicle ID/ License plate*, e.g. "HST"	X	X	
• *Period fuel usage/ energy usage*, e.g. "2022-01"	X	X	X
• Date, e.g. "01-01-2022"	X	X	X
• Amount (Loading metre), e.g. "0.4"	X		
• Country code departure/ destination, e.g. "NL"	X		
• City departure/ destination, e.g. "Enschede"	X		
• Postal code departure/ destination, e.g. "7547 RW"	X		
• Customer, e.g. "HST Group"	X		
Tab "Fuel consumption			
• *Carrier*, e.g. "HST"	X		
• *Vehicle ID/ License plate*, e.g. "HST"	X	X	
• *Period fuel usage*, e.g. "2022-01"	X	X	X
• *Fuel type*, e.g. "Diesel (Liter)"		X	
• Total amount of fuel (L), e.g. "360250.12"		X	
• Emissions unknown (type "X"). If the energy usage of the shipment is unknown, e.g. network carriers.			
Tab "Locations"			
• *Period energy usage*, e.g. "2022-01"	X	X	X
• Renewable energy (kWh), eg. "75000.1234"			X
• Natural gas (m3), e.g. "12000.2504"			X

Table 1: Overview of inputs

3.4 Data quality criteria

Lean and Green presents a rulebook which mentions a few data quality levels (Simons, 2020).

- (B) Bronze: default values, estimates or GLEC figures
- (S) Silver: measured, calculated, derived (actual) values, aggregated per period (year or month)
- (G) Gold: measured, calculated, derived (actual) values per license plate/location per period
- (G+) Gold+: measured (actual) values per stop/trip

For Lean & Green the in Table 2 mentioned criteria are the requirements for the Lean & Green stars. (Simons, 2020)

	CO2 reduction	Allowed base year	Years to reach Co2 goal	Maximum years to maintain reductions	Minimum volume/CO2 scope	Minimum data quality	Minimum data completeness
1st Star	20%	Year x-3	5	3	50%	Bronze	95%
2nd Star	10%	Year x-2	3	3	65%	Silver	95%
3rd Star	5%	year x-1	2	3	75 %	Silver	95%
4th Star	Absolute reduction	Year x-1	Moving target goal*	N/A	85% (expected, in development)	Gold	95%
5th Star	Absolute reduction	Year x-1	2050 at the latest	N/A	90% (expected, in development)	Gold	95%

Table 2: Lean and Green Star criteria

At the time of writing HST Group has achieved the first Lean & Green star. The goal in the near future is, as mentioned in Chapter 1.1, to achieve the second or third Lean & Green star. Both demand a minimum data quality level of *Silver*.

4 Data quality analysis

To provide an overview of the current situation regarding data quality for HST Group data quality analysis has been performed. In this chapter, the chosen dimensions (Integrity, Completeness and Uniqueness and deduplication) are analysed per original dataset for the year 2022. Per dimension the inputs mentioned in Table 1 are analysed.

4.1 Integrity datasets year 2022

In this chapter, the integrity of the data will be analysed. Findings regarding integrity will be stated in this chapter. The consequences of the integrity of the data will be discussed in Chapter 5.1. For some fields, the integrity is explained with numbers and for other fields, the integrity is explained by examples. This is due to the time limitations of the study.

- Order number (Dossiernr) and serial number (Volgnr)
  In the current database *Order number (Dutch: Dossiernr)* is used as the order number, while *Serial number (Dutch: Volgnr)* is used to track a part of an order. For example, a company can order that every Monday of that month a shipment will take place, each with the same order number and a different serial number. Note; this does **not** mean that serial number is used for parts of a shipment (e.g. from A to B, B to C, C to D).

- Carrier
  The carrier is often unknown in the exported data. It is sometimes only known if a package is sent within the network. When a package switches from carrier within the shipment it is not visible in the export.

- Vehicle ID
  The input for *vehicle ID/ license plate* in the data is done in different ways: the field is *empty*, the field is *[name driver]/[Vehicle ID][License plate]*, the field is *[License plate]* or the field is some other name for a carrier.

- Fuel and energy usage periods and dates
  The periods of fuel or energy usage are not present in the datasets. However, the loading dates of the shipments are known.

- Amount
  For *Amount* we check if the value is not *null*, 0 or negative. Extremities are not wanted either, meaning above 13.6 loading metres as this is one full trailer. The original exported datasets from Transpas 2022 are analysed for these values. Beforehand is filtered to only include the invoiced orders.

Count	Transpas National	TransPas International
Total rows	650516	88990
Count negatives	0	1
Count *null*	3954	536
Count 0	452282	1425
Count >13.6	9	101
Count correct range	194271	86927

Table 3: Overview of Amount faults

- Country code departure or destination
  BigMile requires a standard country code of two letters (alpha-2, ISO 3166-1). In the system, there are other variants available such as a single letter (F: France, should be FR), and three letters (Est: Estonia, should be EE; MLT: Malta, should be MT, etc.).

4.2 Completeness datasets year 2022

For completeness, each file is analysed for each input value. In the column *Description* the way the completeness is counted is analysed. The percentage is the *Count* divided by the total amount of rows. When the count is not the same as all the rows the count is set to not be 100.0%. This is done such that it is visible that there is missing data. The consequences of the completeness of the data are discussed in Chapter 5.2.

- Original Transpas National 2022

Inputvalue	Count	Percentage	Description
Order number	719842	100.0%	Counted non-blank values
Serial number	719842	100.0%	Counted non-blank values
Carrier	387685	53.9%	Counted non-blank values
Vehicle ID	654259	90.9%	Counted non-blank values
Period fuel/energy usage	0	0%	Counted non-blank values
Date	719842	100.0%	Counted non-blank values
Amount	486997	67.7%	Counted values >0
Country code departure	719842	100.0%	Counted non-blank values
Country code destination	719841	99.9%	Counted non-blank values
City departure	719614	99.9%	Counted non-blank values
City destination	718390	99.8%	Counted non-blank values
Postal code departure	719842	100.0%	Counted non-blank values
Postal code destination	719841	99.9%	Counted non-blank values
Customer	719842	100.0%	Counted non-blank values

Table 4: Completeness Original Transpas National 2022

In Table 4 an overview of the completeness is given for the exported file from Transpas for the year 2022. This is the export for the shipments done by the national department. The values that stand out, when keeping the 95% data completeness in mind (see Chapter 3.4), in Table 4 are the *Carrier* values, for which 46.1% are missing, *Vehicle ID* values, for which 9.1% are missing, *Period fuel/energy usage* values, for which all are missing, and *Amount* (of loading metres) values, for which 32.3% are missing.

- Original Transpas International 2022

Inputvalue	Count	Percentage	Description
Order number	97996	99.9%	Counted non-blank values
Serial number	98000	100.0%	Counted non-blank values
Carrier	0	0%	Counted non-blank values
Vehicle ID	59509	60.7%	Counted non-blank values
Period fuel/energy usage	0	0%	Counted non-blank values
Date	97998	99.9%	Counted non-blank values
Amount	95663	97.6%	Counted values >0
Country code departure	97999	99.9%	Counted non-blank values
Country code destination	97999	99.9%	Counted non-blank values
City departure	97750	99.7%	Counted non-blank values
City destination	96931	98.9%	Counted non-blank values
Postal code departure	97993	99.9%	Counted non-blank values
Postal code destination	97998	99.9%	Counted non-blank values
Customer	98000	100.0%	Counted non-blank values

Table 5: Completeness Original Transpas International 2022

In Table 5 an overview of the completeness is given for the exported file from Transpas for the year 2022. This is the export for the shipments done by the international department. The values that stand out, when keeping the 95% data completeness in mind (see Chapter 3.4), in Table 5 are the *Carrier* values, for which all are missing, *Period fuel/energy usage* values, for which all are missing, and *Vehicle ID* values, for which 39.3% are missing.

- Fuel Usage AVIA 2022

Inputvalue AVIA	Count	Percentage	Description
Vehicle ID	363	2.2%	Counted non-blank values
Date	16432	100.0%	Counted non-blank values
Fuel type	16432	100.0%	Counted non-blank values
Amount (L)	16432	100.0%	Counted non-blank values

Table 6: Completeness Fuel usage AVIA 2022

In Table 6 the completeness of the input values that can be gathered from an export from AVIA is shown. AVIA is one of the companies which HST Group uses to get fuel for their trucks. In Table 6 it stands out that 97.8% of the *Vehicle IDs* are unknown.

- Fuel Usage BP 2022

Inputvalue BP	Count	Percentage	Description
Vehicle ID	2925	71.8%	Counted non-blank values
Date	4073	100.0%	Counted non-blank values
Fuel type	4073	100.0%	Counted non-blank values
Amount (L)	4073	100.0%	Counted non-blank values

Table 7: Completeness Fuel usage BP 2022

In Table 7 the completeness of the input values that can be gathered from an export from BP is shown. BP is the second company which HST Group uses to get fuel

for their trucks. In Table 7 it also stands out that many values of *Vehicle ID* are missing. In this case 28.2%.

- Energy Usage 2022
  In Table 8 the completeness of the energy usage values is given. These are the values for the energy usage at the location of HST Group. HST Group uses natural gas and renewable energy.

Inputvalue	Count	Percentage	Description
Month	12	100.0%	Counted non-blank values
Renewable energy (kWh)	12	100.0%	Counted non-blank values
Natural gas (m3)	12	100.0%	Counted non-blank values

TABLE 8: Completeness Energy usage 2022

4.3 Uniqueness and deduplication datasets year 2022

For *Uniqueness and deduplication* the input value *Order number* must be a unique value. For all other input values duplicates are possible and expected. In the HST database the *Order number* is "=*Dossiernr*&" - "&*Volgnr*". This is because a *Dossiernr* can have multiple shipments with different *Volgnr*.

File	Duplications
Original Transpas National 2022	0
Original Transpas International 2022	4

TABLE 9: Duplications of order number 2022

For the Transpas International export, there are 4 duplicate values. These values are all blank values.

5 Consequences of current data quality

In this chapter, the consequences of the figures presented in Chapter 4 will be discussed. In the first three sections the consequences of *integrity*, *completeness* and *uniqueness and deduplication* are given. In Section 5.4 the consequences of findings during the data analysis are given. Lastly, in Section 5.5 the result of the first four sections mentioned consequences is given, regarding the current data quality level for the Lean & Green program.

5.1 Consequences of integrity problems

In this section, the consequences of the integrity of the data will be analysed. For some fields the integrity is explained with numbers and for other fields, the integrity is explained by examples. This is due to the time limitations of the study.

- Order number (Dossiernr) and serial number (Volgnr)
  The effect that not being able to track the unique orders by the current order number (Dutch: Dossiernummer) is that there are no unique values currently and therefore the data cannot be imported to the Lean & Green input file. It will try to read multiple orders as partial shipments, as HST Group has used the serial number differently. These problems should be solved to be able to perform an analysis for two Lean & Green Stars or above. A way should be found to 1) create a unique order value, and 2) create another way to track partial shipments.

- Carrier
  When the carrier is unknown, or incorrectly labelled, in the exported data, it has as a consequence that there is no clear overview of which shipments are carried out by HST Group and which are carried out by another carrier. For the missing values in the dataset Transpas National 2022 the missing values are carried out by HST Group. For the dataset Transpas International 2022 it is more complicated. In the export, it is not evident for which shipments a carrier is used. Therefore this issue should be addressed, otherwise, the Transpas International 2022 dataset should be out of scope.

- Vehicle ID
  As the input for *Vehicle ID/ license plate* is done in different ways in the dataset, the data cannot be used easily, if at all. However currently, for the goal of two or three Lean & Green Stars for HST Group, the *Vehicle ID* is unneeded to reach that goal. Currently, this problem does not pose any challenge in reaching two or three Lean & Green Stars, however, it gives problems if HST Group wants to reach for more than three Lean & Green Stars. More information is given about the completeness of the *Vehicle IDs* in Section 5.2.

- Period fuel/energy usage and date
  Currently, there is no period assigned per order. Currently, the data cannot be assigned to a period without any data operations. The *period of fuel/energy usage* could be linked to the *loading date* for the shipment. For International one *loading date* out of almost 98,000 dates was missing and for National no dates were missing, therefore this is a viable option.

- Amount
  From Table 3 the conclusion can be made that a lot of calculations need to be made to provide an insight into the transported loading metre amount. This leads to a lot of

uncertainty for the integrity of the loading metres (/amount). Without any loading metre calculations, the data is unfit for use in the input file from Lean & Green. Therefore loading metres calculations need to be performed to meet the conditions for two or three Lean & Green Stars.

- Country code departure or destination
  In the system, there are other variants available such as a single letter (F: France, should be FR), and three letters (Est: Estonia, should be EE; MLT: Malta, should be MT, etc.). This results in that these labels have to be corrected to perform a BigMile analysis. This can be done by replacing the values with code.

5.2 Consequences of completeness problems

- Transpas National 2022
  As is visible in Table 4 there are a few problems with the data completeness. The main problem is *Amount*, which leads to a lot of guessing and calculations to perform the BigMile analysis. Most of the missing data is a result of when packages from a network are put in the system, no loading metre value is assigned.

  Note: For Lean and Green *Silver data quality level* the *Vehicle ID* is not needed. The text *HST Group* is used for the shipments transported by the company. For *Carrier* the values given are all shipments within a network. The missing values for *Carrier*, *null* values, are driven by HST Group. Therefore the percentage is not a big problem currently. In the future, to achieve *Gold data quality* (chapter 3.4) the *Vehicle ID* needs to be known. As explained in Chapter 4.1, the *Serial number* used by HST Group does not track partial shipments. In the cleaned versions this is solved with the use of code.

- Transpas International 2022
  For *Carrier*, there is no input if the shipment is done by HST Group or with a carrier. Therefore the count is 0. This is also a result of partial shipments not being split into multiple rows. This problem is elaborated upon in chapter 5.4. Another problem is that there are still values for *Amount* missing. For these values calculated guesses have to be made to give an accurate overview of the transported loading metres.

- Fuel usage AVIA 2022
  The main problem visible in Table 6 is that the *Vehicle ID* is almost always unknown. Currently, this leads to no problems, however in the future when Lean & Green star four and five are a possibility, this is required. For the last mentioned stars the data must be *measured, calculated, derived (actual) values per license plate/ location per period* (Chapter 3.4)

- Fuel usage BP 2022
  The problem mentioned in the previous section also applies to Table 7. In the case of the completeness of the *Vehicle ID* in Table 7 it is less of a problem, as in the previous Table 6 almost 98% of the data is missing regarding *Vehicle ID*. However, *Vehicle ID* is still a problem in the future as almost 30% of the data is missing.

- Energy usage 2022
  For energy usage, there were no problems regarding completeness as all of the data was present.

5.3 Consequences of uniqueness and deduplication problems

All of the duplicate values, which all four are in the dataset *Transpas International*, are blank values. Therefore there is a problem with four shipments that have no order numbers. The missing data leads to that the four orders can not be analysed.

5.4 Consequences of found problems during research

In this section, other problems found during the data cleaning period are explained. These provide an added overview of shortcomings in the data quality of HST Group. These shortcomings focus on the period of the year 2022 as that represents the most recent TMS in use.

- Non-existent postal codes
  During data cleaning, there were postal codes that were non-existent. Some were possibly a typing error and some were inexplicable. The result is that the destination or departure location is wrong and the distance between the locations cannot be calculated for this shipment. This results in fuel which is used for a shipment, but the shipment is not included by the BigMile tool. Therefore this has a negative effect on the relative emitted carbon emissions per loading metre. There were postal codes for Germany, Estonia and Finland for which the first 0 in the postal code was missing. This could be fixed to improve the data quality of the postal codes a bit.

- Charters are not categorised
  HST Group uses charters when they cannot complete the order themselves, or when it is financially responsible to use another charter. For the BigMile analysis, it is required to label the orders driven by charters, however, in the current exports of the database, no column specifies which charter drives (part of) the delivery. The main problem is that the export contains almost no details on charters when they are used. It contains details on whether a charter has loaded the shipment, thus in the first part of the trip, or has unloaded the shipment, thus in the last part of the trip. The route which the charter drives is mostly not known in the export. E.g. a trip of a shipment is from A to B by *Carrier X* and from B to C by *HST Group*. In the export, it is written as a trip from A to C, with as loading creditor *Carrier X*. It is unknown where the transfer point B is.

  A separate row per part of the shipment with a serial number (Dutch: volgnummer) such that the partial shipments are traceable to the same shipment (following the previous example one row for A to B by Carrier X and one row for B to C by HST Group) is needed such that BigMile understands that multiple order rows belong together in one order.

- DPD, Fedex and GLS shipments are not split in the TMS export
  Following the previous point DPD, Fedex and GLS have the same problem. The first part is driven by HST Group and at the depot at HST Group the packages are handed over to the providers. The problem this induces is that the original destination for the order is not the same address as the transfer point where HST Group transfers it to DPD, Fedex or GLS, namely HST Group's location in Enschede, Netherlands. For the Lean & Green input file, the destination could be changed to the depot address of HST Group.

- No export function in Transpas
  Transpas has no export function. Therefore an employee has to copy and paste everything into an Excel sheet. This leads to downtime for the other activities the employee has to do that day. It also stresses the computer servers at HST Group, on which all the Windows clients run at the company, and therefore has to be done preferably outside of office hours.

  After currently an export is produced, the orders need to be filtered on orders that are completed. The orders that are left out are orders that were cancelled. This has to be done manually. Implementing filtering on orders that are completed into an export function would reduce the steps that have to be done manually in order to perform the BigMile analysis.

5.5 Current data quality level

The current data quality level is subpar to the required level. As is mentioned in Chapter 3.4 the completeness needs to be a minimum of 95%, which cannot be achieved when so many loading metre values are missing. A unique order number value is also a necessity for the Lean & Green input file. Likewise deducing a time-period from the loading dates is needed for every shipment. Improvements can also be made regarding splitting international orders, country codes that do not have the correct format currently, postal codes that have a zero missing at the start and shipments that are handed over to DPD, Fedex and GLS. Therefore HST Group cannot improve their Lean & Green star rating without any improvements to the data.

6 Design of the Excel Queries

In order to perform needed data operations Excel Query is used, as it is a known program to the researcher and the research is time-gated. In Chapter 6.1 the problems which the queries aim to solve are presented. The designs of the different queries that aim to solve the previously mentioned problems are also explained in this chapter. In Chapter 6.3 the results of the queries are presented.

6.1 Different Queries

There are multiple different datasets on which data operations have to be performed. One time period is July 2019 - June 2020 and the second time period is the year 2022. As is visible in Figure 1, in the first time period there are two different TMS in use. One for national orders and one for international orders. In the last time period, there is only one kind of TMS in use for both departments.

In Chapter 6.1.1 the explanation of the data from Transmission (National), the merging of the backup files for the time period of July 2019 - June 2020, can be found.

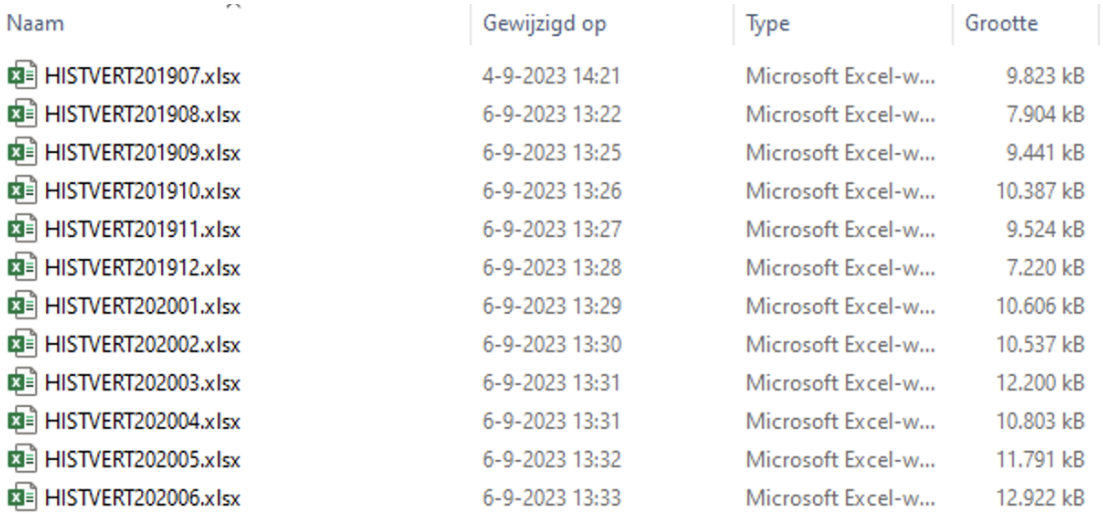
In Chapter 6.1.2 the explanation of the data of 2022 from Transpas (National) can be found.

For both time periods, the exports for TMS International use the same system. Therefore the design for both exports is explained in Chapter 6.1.3.

6.1.1 Transmission National datasets 2019-2020

- Unique order name value
  In order to merge the files each order has to have a unique number such that it can be linked to the other part of the order. The manner in which the orders are backed up meant that first, a unique number had to be created. This is done by merging the order number (Dutch: dossiernummer) and serial number (Dutch: volgnummer). For these datasets, this could be done by combining the shipment number (Dutch: zendingsnummer) and the transfer number (Dutch: nummer verladen).

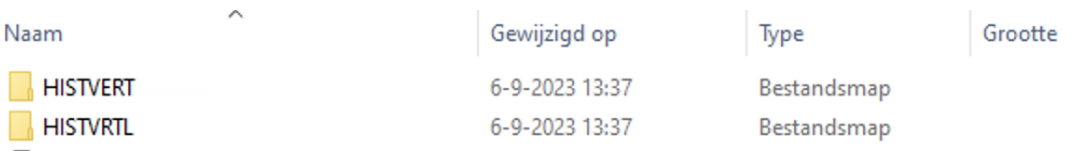
- Selecting a folder to import multiple files
  For the data from the National Department of HST Group a tool has been designed to link different time frames, each file a month of data, together and perform the same data operations on each of them. A folder is selected as the source of the files, as is visible in Appendix C.1 and Appendix C.2. This makes sure that for all of the files in the folder, the same code is used. This minimises the manual work that has to be performed.



Naam	Gewijzigd op	Type	Grootte
HISTVERT201907.xlsx	4-9-2023 14:21	Microsoft Excel-w...	9.823 kB
HISTVERT201908.xlsx	6-9-2023 13:22	Microsoft Excel-w...	7.904 kB
HISTVERT201909.xlsx	6-9-2023 13:25	Microsoft Excel-w...	9.441 kB
HISTVERT201910.xlsx	6-9-2023 13:26	Microsoft Excel-w...	10.387 kB
HISTVERT201911.xlsx	6-9-2023 13:27	Microsoft Excel-w...	9.524 kB
HISTVERT201912.xlsx	6-9-2023 13:28	Microsoft Excel-w...	7.220 kB
HISTVERT202001.xlsx	6-9-2023 13:29	Microsoft Excel-w...	10.606 kB
HISTVERT202002.xlsx	6-9-2023 13:30	Microsoft Excel-w...	10.537 kB
HISTVERT202003.xlsx	6-9-2023 13:31	Microsoft Excel-w...	12.200 kB
HISTVERT202004.xlsx	6-9-2023 13:31	Microsoft Excel-w...	10.803 kB
HISTVERT202005.xlsx	6-9-2023 13:32	Microsoft Excel-w...	11.791 kB
HISTVERT202006.xlsx	6-9-2023 13:33	Microsoft Excel-w...	12.922 kB

FIGURE 7: Example of a folder with multiple, each a month, of data

- Merging files
  The query also links two files together per month. These files are linked together by a unique value per order, which is explained at the beginning of this section. This merge of order numbers is done in each file before merging the different file types together. In Appendix C.3 the code for the merge is found. The visualisation of this is visible in Chapter 6.2.



Naam	Gewijzigd op	Type	Grootte
HISTVERT	6-9-2023 13:37	Bestandsmap	
HISTVRTL	6-9-2023 13:37	Bestandsmap	

FIGURE 8: Two folders from which the files are linked together in the Excel Query

- Decoding country codes
  The backup files from Transmission had the country codes labelled with numbers. An overview of which number represented which country was not found during the research. A list of corresponding countries was compiled with the use of the addresses and postal codes in the files. A section of codes with many if-statements is used to convert the numbers into the corresponding countries. The code is visible in Appendix C.1.

- Extracting the time period from the loading date
  In the data, there were no time periods per month given. Therefore the time period in which the order was shipped had to be extracted from the loading date. The code for this, which is visible in Appendix C.1, first duplicates the loading date column and then converts this column into the [YYYY - MM] value, such that it meets the conditions for the Lean & Green input-file for the BigMile analysis.

- Loading metre calculation
  As there are loading metre values missing, which are needed for the BigMile analysis, a script is written to calculate new loading metre values. This script is visible in

Appendix C.2. It first selects the *missing values* with the use of an if-statement as it only applies the *loading metre calculation* code to the orders where the original loading metre value is *null* or 0. For packages, it uses the assumption, which was given by the planning department, a maximum of 40 packages per cage container. This results in 50 packages per loading metre on average. For pallets, the value is directly converted into loading metres, as one pallet is 0.4 loading metres.

6.1.2 Transpas National 2022 Excel Query design

- DPD, GLS and FedEx packages
  Packages for DPD, GLS and FedEx are different kinds of shipments, meaning that these packages are collected by HST Group at the set departure location, and then transferred by HST Group to the location of HST Group in Enschede. From there the packages are picked up by the corresponding company; DPD, GLS or FedEx.

  Currently, in the exported dataset, the departure location is set correctly. However, because partial shipments are not exported currently, the destination is set in the dataset as the location to which DPD, GLS or FedEx delivers it. For the Lean & Green input file, we are interested in which part HST Group has driven. Therefore the data has to be altered to be the original location of departure, which it is already, to the location of HST Group. This means that for the destination the postal code, country code, address and city are set to the location of HST Group for these shipments.

- Packages that are imported from and exported into the network
  HST Group imports packages from a Network in The Netherlands. These packages are saved in the database correctly for the use of the Lean & Green input file. The departure location is the depot at HST Group and the destination location is the final destination of the package. This route is fully completed by HST Group. Therefore no changes are needed for these shipments.

  HST Group also uses a network to more efficiently send packages in bulk to other destinations in The Netherlands. This works in broad terms the same as how the system for DPD, GLS and Fedex works. Therefore the destination of the packages has to be altered to the depot at HST Group for the Lean & Green input file. This means that for the destination the postal code, country code, address and city are set to the location of HST Group for these shipments.

- Fixing postal codes
  There was a problem with multiple postal codes where the first 0 of the postal code seemed to be missing. This is possibly a result of converting and opening the files with Excel. This is fixed for the loading postal codes where a 0 was missing at the beginning of the postal code for the countries Germany, Estonia and Finland. This was done by adding a zero to the beginning of the string.

- Country codes that are not Alpha-2 standard
  As is mentioned in Chapter5.4, for the BigMile analysis, and therefore for the Lean & Green input-file, a two-letter country code format is needed (alpha-2, ISO 3166-1). In the current database, there is a country with only one letter (France) and countries with three letters (Estonia, Malta, San Marino). Iceland has the wrong two-letter country code. This is solved by searching for the wrong country codes and replacing

them with the right Alpha-2 standard country codes. This is visible in the code in Appendix B and D.

- Time-period from loading date
  In the data, there were no time periods per month given. Therefore the time period in which the order was shipped had to be extracted from the loading date. The code for this, which is visible in Appendix D, first duplicates the loading date column and then converts this column into the [YYYY - MM] value, such that it meets the conditions for the Lean & Green input-file for the BigMile analysis.

- Loading metre calculation
  As is visible in Table 4, around one-third of the loading metres values are missing. This needs to be improved in order to comply with the Lean & Green requirements (See Chapter 3.4). This is done by deduction and calculation.

  In the Excel Query, there are different categories for how the loading metres are calculated:

  1. Packages
     Packages are calculated by how many packages fit in a cage container on average. On average it is expected to fit 30 to 40 packages in one cage container, the source for this is a Planner for TransMission from HST Group. For the calculations for Lean & Green, it is chosen to use the 40 packages per cage container value. Converting this to loading metres means 40 packages per 0.5 loading metres. Calculation: *0.5/40 * Amount of packages* .
  2. Pallets
     One pallet is 0.4 loading metres. The calculation for the amount of loading metres is straightforward: *Amount of pallets * 0.4*.
  3. Cubic metres
     Some shipments are given with the value of cubic metres in the order. A trailer has room for 90 cubic metres and a trailer has the capacity of 13.6 loading metres. Therefore the calculation for these orders is: *Amount of cubic metres / 90 * 13.6*.
  4. Whole truck
     When a whole truck with a trailer is rented, the unit value given is *hours*. As the whole trailer is in use the loading metre value is converted to 13.6 in the script.
  5. Other items
     For other items that do not fall in the above-mentioned categories, e.g. for bikes, the calculation is: *(Length/100) * (Width/100) * 0.4*. The length and width need to be converted from centimetres into metres. Lastly, it is converted into loading metres by multiplying with 0.4.

- Unique order name value
  The Lean & Green input file requires unique order name values. As HST Group uses the serial number (Dutch: Volgnummer) as a partial order instead of a partial shipment, a unique order number has to be created. This is done by merging the order number and the serial number. As is the same for International orders.

6.1.3 Transpas International Excel Query design

As the export for both time-frames, July 2019 - June 2020 and the year 2022, is done the same, for both years the same script is used. This script can be found in Appendix B.

- Unique order name value
  The Lean & Green input file requires unique order name values. As HST Group uses the serial number (Dutch: Volgnummer) as a partial order instead of a partial shipment, a unique order number has to be created. This is done by merging the order number and the serial number. As is the same for the National orders.

- Changing country codes to Alpha-2 ISO standard
  As is mentioned in Chapter5.4, for the BigMile analysis, and therefore for the Lean & Green input-file, a two-letter country code format is needed (alpha-2, ISO 3166-1). In the current database, there is a country with only one letter (France) and countries with three letters (Estonia, Malta, San Marino). Iceland has the wrong two-letter country code. This is solved by searching for the wrong country codes and replacing them with the right Alpha-2 standard country codes. This is visible in the code in Appendix B and D.

- Splitting partial shipments into multiple rows
  As is mentioned in Chapter 5.4, partial shipments not being split is a problem for performing the analysis. It was not easily visible in the export which partial shipments were transferred to other charters and which were not. An analysis has been made for converting the information in the columns *Loading creditor*, *Unloading creditor* and *Truck unloading*. This overview is visible in Figure 9. Extracting information from these columns results only in the name of the city where the transfer point is instead of the postal code. In the query, code is written that uses the three inputs (*loading creditor*, *unloading creditor* and *Truck unloading*) to get results one, two or three outputs. This depends on which inputs are given. In the output (Action) section of Figure 9 the yellow parts are outsourced and the green parts are executed by HST Group. For the inputs red means data is not present, yellow means the data is partially present and green means that the data is present. In the *Result* column an overview is created to get an understanding of which part of the route has been executed by HST Group (green) and which part is outsourced (yellow).

Cred laad	Cred los	Truck los	Resultaat	Acties
Niet aanwezig	Niet aanwezig	Niet aanwezig	HST rijdt volledig traject	Regel 1 (HST): laadpostcode t/m lospostcode
Aanwezig	Niet aanwezig	Niet aanwezig	HST rijdt vanaf eigen depot	Regel 1 (uitbesteed): laadpostcode t/m depot HST
				Regel 2 (HST): depot HST t/m lospostcode
Niet aanwezig	Aanwezig	Land/losplaats aanwezig	HST rijdt tot extern depot	Regel 1 (HST): laadpostcode tot depot plaats extern
				Regel 2 (uitbesteed): depot plaats extern tot lospostcode
Niet aanwezig	Aanwezig	Naam aanwezig	HST rijdt tot eigen depot	Regel 1 (HST): laadpostcode tot depot HST
				Regel 2 (uitbesteed): depot HST tot lospostcode
Aanwezig	Aanwezig	Land/losplaats aanwezig	Hst rijdt van depot HST t/m extern depot	Regel 1 (uitbesteed): laadpostcode tot depot HST
				Regel 2 (HST): depot HST t/m depot plaats extern
				Regel 3 (uitbesteed): depot plaats extern t/m lospostcode
Aanwezig	Aanwezig	Naam aanwezig	Volledig uitbesteed	Regel 1 (uitbesteed): laadpostcode t/m lospostcode

FIGURE 9: Splitting shipments

- Identifying charters
  As is visible in Figure 9, the shipments where charters were used can also be identified. In the output column (Actions) the yellow rows are identified as charters and the green rows are identified as executed by HST Group. This is also done in the query.

- Creating a tracking number for partial shipments
  The Lean & Green input file requires that partial shipments be linked together by a serial number. For example, the first part of the shipment (Departure location to HST Depot) has serial number 1, the second part of the shipment (HST depot to a transfer location abroad) has serial number 2, and the last part of the shipment (transfer location to the destination location) has serial number 3. Each part of the shipment has its own row in the data, which is created by the script. For adding a serial number to each of these partial shipments code is also added to the script.

- Loading metres
  The loading metres for International are calculated in the same way as for National.

  As is visible in Table 4, around one-third of the loading metres values are missing. This needs to be improved in order to comply with the Lean & Green requirements (See Chapter 3.4). This is done by deduction and calculation.

  In the Excel Query, there are different categories for how the loading metres are calculated:

  1. Packages
     Packages are calculated by how many packages fit in a cage container on average. On average it is expected to fit 30 to 40 packages in one cage container, the source for this is a Planner for TransMission from HST Group. For the calculations for Lean & Green, it is chosen to use the 40 packages per cage container value. Converting this to loading metres means 40 packages per 0.5 loading metres. Calculation: *0.5/40 * Amount of packages* .
  2. Pallets
     One pallet is 0.4 loading metres. The calculation for the amount of loading

metres is straightforward: *Amount of pallets * 0.4.*

3. Cubic metres
   Some shipments are given with the value of cubic metres in the order. A trailer has room for 90 cubic metres and a trailer has the capacity of 13.6 loading metres. Therefore the calculation for these orders is: *Amount of cubic metres / 90 * 13.6.*
4. Whole truck
   When a whole truck with a trailer is rented, the unit value given is *hours*. As the whole trailer is in use the loading metre value is converted to 13.6 in the script.
5. Other items
   For other items that do not fall in the above-mentioned categories, e.g. for bikes, the calculation is: *(Length/100) * (Width/100) * 0.4*. The length and width need to be converted from centimetres into metres. Lastly, it is converted into loading metres by multiplying with 0.4.

- Fixing postal codes
  The same problem with some postal codes that were present in the National TMS dataset of the year 2022 was also present in the datasets for International orders. There was a problem with multiple postal codes where the first 0 of the postal code seemed to be missing. This is possibly a result of converting and opening the files with Excel. This is fixed for the loading postal codes where a 0 was missing at the beginning of the postal code for the countries Germany, Estonia and Finland. This was done by adding a zero to the beginning of the string.

- Time-period from loading date
  In the data, there were no time periods per month given. Therefore the time period in which the order was shipped had to be extracted from the loading date. The code for this, which is visible in Appendix B, first duplicates the loading date column and then converts this column into the [YYYY - MM] value, such that it meets the conditions for the Lean & Green input-file for the BigMile analysis.

6.2 Visualisation of new situation

This chapter visualises the new data situation when the Excel Queries are used. This is done to give the reader a better understanding of what has been done. The big difference is that data operations to clean and improve the data have been performed. Also data sets have been linked with the use of Excel Query for the time period 2019-2020. In each section, the figure is explained.

6.2.1 2019-2020

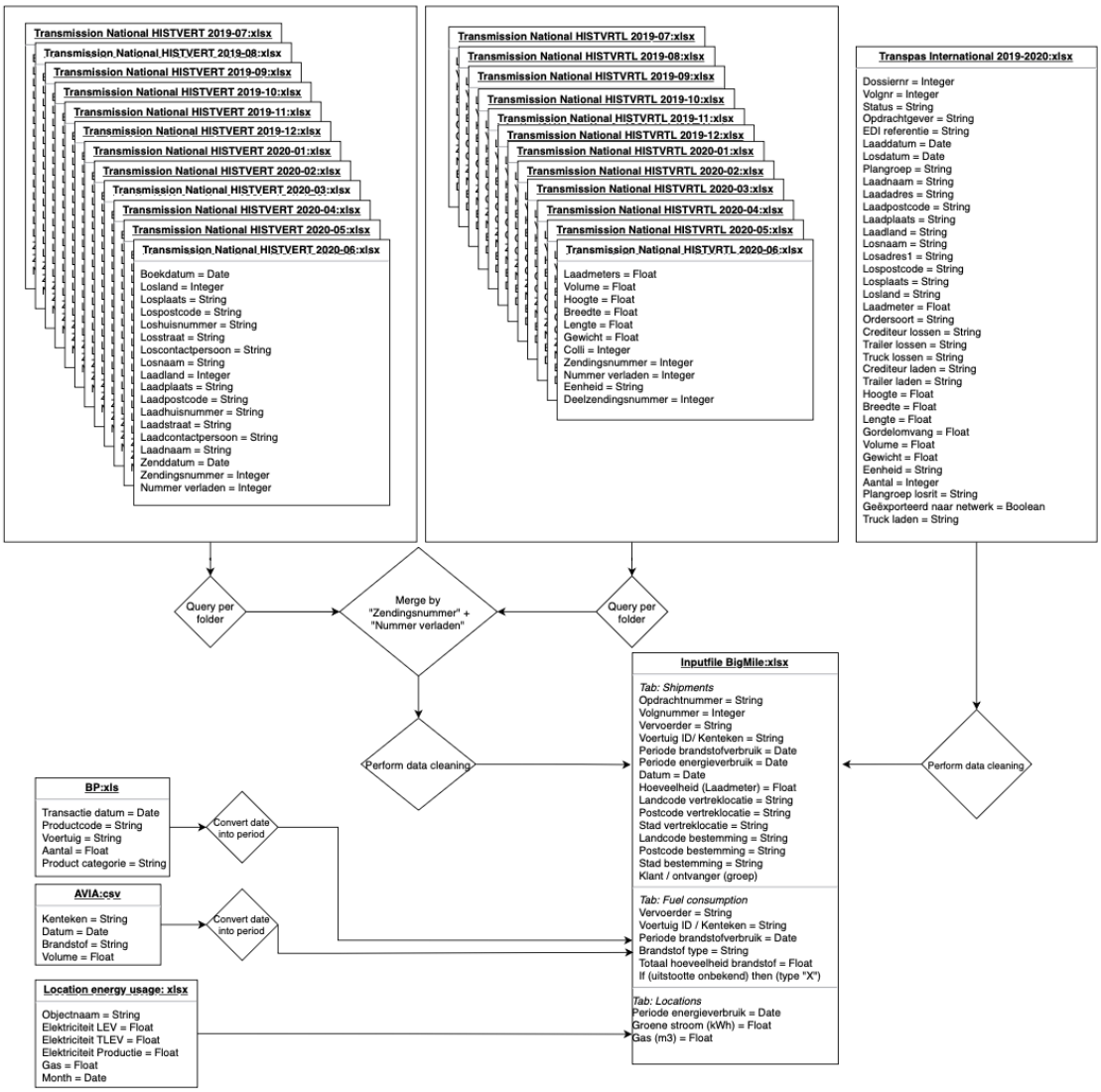


FIGURE 10: New data situation with the use of queries for data sets 2019-2020

Figure 10 shows the relations between the files for the time period 2019-2020. First, the HISTVERT and HISTVRTL backup files are converted to Excel files. After this is done the files are put into their own folder together. *HISTVERT.xlsx* files in the HISTVERT folder, and *HISTVRTL.xlsx* files in the HISTVRTL folder. For each folder, the files are linked together with the use of an Excel Query. The code for this is found in Appendix C.1 for HISTVERT and in Appendix C.2 for HISTVRTL. The merge of both queries is done with the code found in Appendix C.3. Within the three queries, the needed data operations are performed. In Figure 10 this is visible as "Perform data cleaning", which is done in order to keep the figure clear and understandable.

For *BP.xls* and *AVIA.csv*, the files in which the fuel consumption data can be found, the time periods were not known. The time periods were deduced from the dates. After this, the fuel consumption could be summed per month, per fuel type. After this, the data could be used as input for the *Inputfile BigMile.xlsx*.

The data in the *Location energy usage.xlsx* file was already usable and could directly be used as input for the *Inputfile BigMile.xlsx*.

For *Transpas International 2019-2020.xlsx* the code in Appendix B is used during the *Perform data cleaning* action, which is visible at the right part of Figure 10.

6.2.2 2022

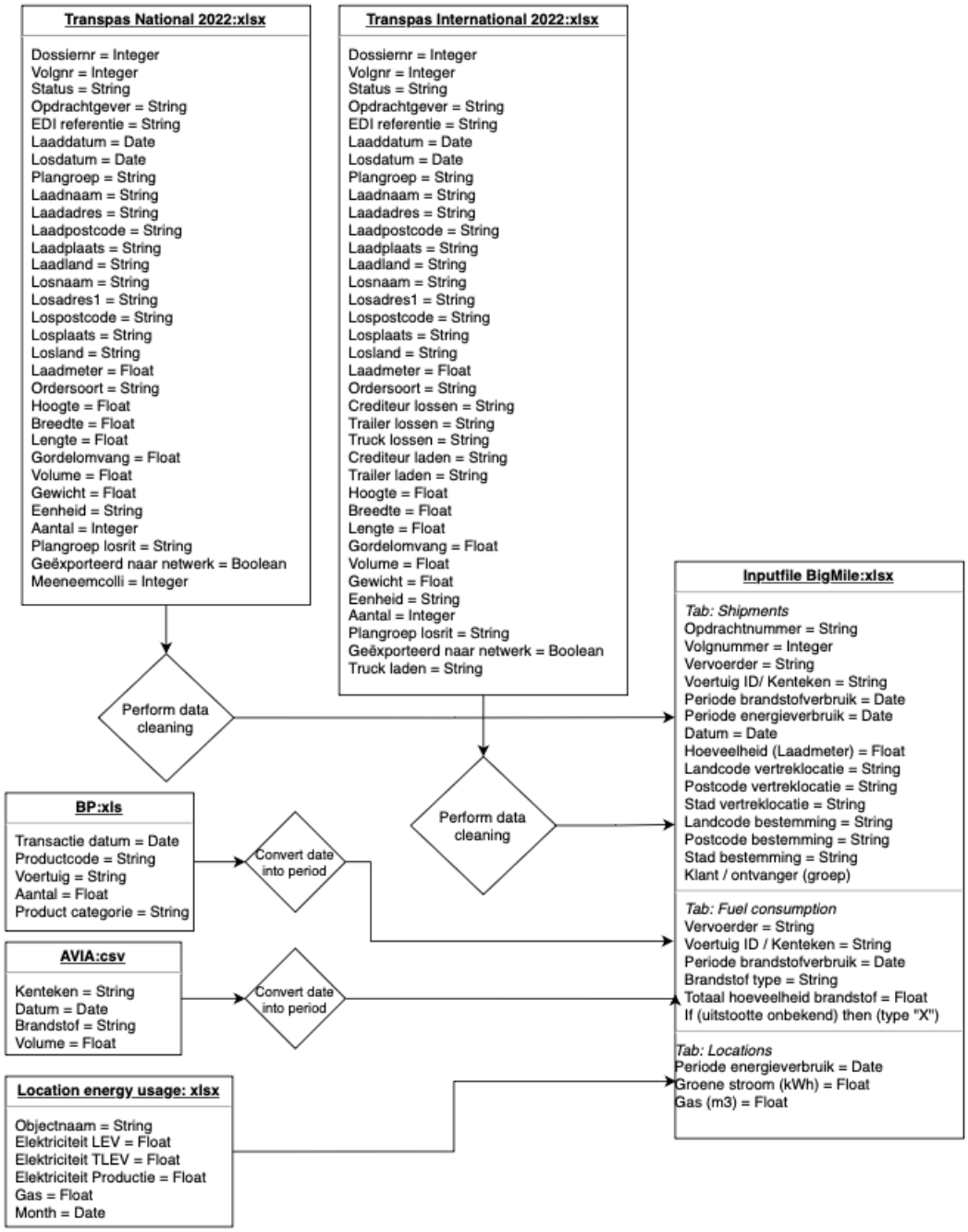


FIGURE 11: New data situation with the use of queries for data sets 2022

Figure 11 visualises what file is the source for the data in a tab (the worksheet in Excel, Dutch: tabblad) in the Excel input file for the BigMile analysis. The model explains the relationship between the different files. The arrows indicate which information per tab in the input file originates from which document. E.g. the tab (worksheet) *Fuel consumption* in the Excel input file gathers information from the BP and AVIA Excel files.

For *BP.xls* and *AVIA.csv*, the files in which the fuel consumption data can be found, the time periods were not known. The time periods were deduced from the dates. After this, the fuel consumption could be summed per month, per fuel type. After this, the data could be used as input for the *Inputfile BigMile.xlsx*.

The data in the *Location energy usage.xlsx* file was already usable and could directly be used as input for the *Inputfile BigMile.xlsx*.

For *Transpas International 2022.xlsx* the code in Appendix B is used during the *Perform data cleaning* action, which is visible at the middle of the top part of Figure 11.
For *Transpas National 2022.xlsx* the code in Appendix D is used during the *Perform data cleaning* action, which is visible at the left of the top part of Figure 11.

6.3 Result of the Excel Query

In this chapter the effect of the Excel Query on the data quality is discussed. First, the integrity when the queries are used will be explained in Section 6.3.1.In Section 6.3.2 the completeness will be discussed. In Section 6.3.3 the uniqueness and deduplication will be addressed. Section 6.3.5 will explain the data quality level when the queries are used.

6.3.1 Integrity

- Order number and serial number
  For each query, a unique order number value has been created. As is mentioned in Section 6.3.4, another way to track partial partial shipments by serial number has been created. These issues are solved with the queries.

- Carrier
  As is also mentioned in Section 6.3.4, a way to track when a charter is used has been created for the international shipments in the query. For national shipments, it is also visible when a carrier is used and when HST Group has completed the order. This issue is also solved.

- Vehicle ID
  The queries have made no improvement for the *Vehicle IDs*. This is not an issue up to and including the third Lean & Green Star. For the fourth and fifth Lean & Green Star this will be an issue.

- Period fuel/energy usage and date
  In each query code is added to deduce the time-period such that the fuel and energy usage can be linked in the Lean & Green input-file. This issue has been solved.

- Amount
  Big improvements have been made by deducing and calculating values for the amount (/loading metres). These improvements are also visible in Section 6.3.2. The queries

have improved the data enough such that the data can be used for the BigMile analysis.

- Country code departure or destination
  The queries have converted the country codes that were an issue to the correct ones. Therefore this issue has been solved.

6.3.2 Completeness

For completeness, each file is analysed for each input value. In the column *Description* the way the completeness is counted is analysed. The percentage is the *Count* divided by the total amount of rows. When the count is not the same as all the rows the count is set to not be 100.0%. This is done such that it is visible that there is missing data. For the cleaned versions the files are cleaned with the use of Excel Query. These cleaned versions were used for the input for the BigMile analysis of the year 2022. The script is in Appendix B and Appendix D.

- Cleaned Transpas National 2022

Inputvalue	Count	Percentage	Improvement	Description
Order number	650516	100.0%	0%	Counted non-blank values
Serial number	650516	100.0%	0%	Counted non-blank values
Carrier	650516	100.0%	46.1%	Counted non-blank values
Vehicle ID	623239	95.8%	-4.9%	Counted non-blank values
Period fuel/energy usage	650516	100.0%	100%	Counted non-blank values
Date	650516	100.0%	0%	Counted non-blank values
Amount	640855	98.5%	30.8%	Counted values >0
Country code departure	650516	100.0%	0%	Counted non-blank values
Country code destination	650516	100.0%	0.1%	Counted non-blank values
City departure	650491	99.9%	0%	Counted non-blank values
City destination	650509	99.9%	0.1%	Counted non-blank values
Postal code departure	650515	99.9%	-0.1%	Counted non-blank values
Postal code destination	650515	99.9%	0%	Counted non-blank values
Customer	650516	100.0%	0%	Counted non-blank values

Table 10: Completeness Cleaned Transpas National 2022

Table 10 shows that the overall number of rows has decreased. This is because the cancelled orders are filtered out. Cancelled orders are not carried out and therefore they only exist in the data. E.g. there is no fuel and storage used for a cancelled shipment. Since the shipments of these cancelled orders are not carried out, the loading metres are not included in the analysis. This also leads to a relative improvement with values such as *Vehicle ID*.

For Lean and Green *Silver data quality level* the *Vehicle ID* is not needed. The text *HST Group* is used for the shipments transported by the company. For *Carrier* the values given are all shipments within a network. In the future, to achieve *Gold data quality* the Vehicle ID needs to be known. As explained in Chapter 4.1, the *Serial number* in the original data set used by HST Group does not track partial shipments. In the cleaned versions this is solved with the use of code. In Table 10 the serial number indicates value for a partial shipment.

The big improvements are the *carrier*, *period of fuel/energy usage* and *amount* values. With these values, there are no issues regarding data quality for reaching a second or third Lean & Green Star.

- Cleaned Transpas International 2022

Inputvalue	Count	Percentage	Improvement	Description
Order number	158543	99.9%	0%	Counted non-blank values
Serial number	158549	100.0%	0%	Counted non-blank values
Carrier	158549	100%	100%	Counted non-blank values
Vehicle ID	93994	59.3%	-1.4%	Counted non-blank values
Period fuel/energy usage	158549	100.0%	100%	Counted non-blank values
Date	158547	99.9%	0%	Counted non-blank values
Amount	158351	99.9%	2.3%	Counted values >0
Country code departure	158549	100.0%	0.1%	Counted non-blank values
Country code destination	158549	100.0%	0.1%	Counted non-blank values
City departure*	34032	21.5%	See *	Counted non-blank values
City destination*	34032	21.5%	See *	Counted non-blank values
Postal code departure*	124518	78.5%	See *	Counted non-blank values
Postal code destination*	124518	78.5%	See *	Counted non-blank values
Customer	158549	100.0%	0%	Counted non-blank values

Table 11: Completeness Cleaned Transpas International 2022

There are a lot more rows after the cleaning script is run than in the original data set. This is due to splitting shipments into multiple rows when parts of a shipment are transported by another carrier. For Lean and Green *Silver data quality* the *Vehicle ID* is *Carrier* is used. In the future, to achieve *Gold data quality* the Vehicle ID needs to be known.

* For *City departure/City destination* and *Postal code departure/Postal code destination* only one of both is selected in the data cleaning script. The postal code is preferred above the city name because a postal code is more precise than a city name. Therefore if the postal code is known, no city name is given. Adding the two values together gives a total of 100% coverage.

The big improvements are the *carrier*, *period of fuel/energy usage* and *amount* values. With these values, there are no issues regarding data quality for reaching a second or third Lean & Green Star.

6.3.3 Uniqueness and deduplication

The previous missing order numbers for the export of International 2022 could not be fixed. In Table 11 six order numbers are missing instead of four. This is because shipments have been split into partial shipments. Therefore this number has increased. The problem however regarding uniqueness and deduplication did not get worse or better.

6.3.4 Other problems found during the research

- Non-existent postal codes
  For the countries Germany, Estonia and Finland the problem with a zero missing at the beginning of the postal code is solved. Other problems regarding postal codes were not solved.

- Charters are not categorised
  With the use of the queries, it is now known which shipments are transported by HST Group and which are handed over to charters. For the partial shipments a new serial number has been created, such that these orders can be tracked. Therefore this issue is solved.

- DPD, Fedex and GLS shipments are not split in the TMS export
  The queries solve this issue for the Lean & Green input-file, as the part which HST Group drives is in the dataset. This is done by changing the destination location to the depot of HST Group in Enschede. This issue is therefore solved with the use of the queries.

- No export function in Transpas
  The queries have not altered this issue. Therefore there still is no export function for Transpas.

6.3.5 Lean & Green data quality level with the use of Excel Query

The data quality level when the Excel Query script is used is Silver. The ways data values can be gathered is worded in the requirements as *measurements, calculated and derived values*. The values are currently aggregated per month. The queries have solved the problems that prevented the data quality from achieving the *Silver* data quality level. Therefore a maximum of three Lean & Green Stars can be achieved when looking at data quality.

In order to progress to the next level, Gold data quality, the data needs to be linked to license plates per period. In Tables 4, 5, 6 and 7 it is visible that a lot of data is missing regarding the license plate values (*Vehicle IDs*). Therefore the Gold data quality level cannot be achieved.

7 Other possible solutions

In Section 7.1 the proposed solutions to the in Chapter 5.4 found problems are listed. These solutions give an insight into what possibly can be done in the future. In the Section 7.2 the effect of the proposed solutions on the Lean & Green data quality level is given.

7.1 Proposed solutions

- Unique order-/ serial number
  - For *order number* and *serial number* it is possible to create an unique *order number*. This can be done for example by combining the current *order number* and *serial number*: [order number]&" - "&[serial number]. After this, the *serial number/ volgnr* field can be used for tracking the part of the shipment. E.g. A to B is serial number 1, B to C is serial number 2, and C to D is serial number 3. Advised is to refrain from using serial numbers as a tracking tool for multiple different shipments ordered in the same order. These different shipments should have a different order number value.
- Transfer point
  - Extracting the transfer point from *Crediteur laden*/*Crediteur lossen*
    As is done using the data cleaning script the transfer point can be extracted from the columns *Crediteur laden* and *Crediteur lossen*. This is however a last resort measure as this method depends on the city of the transfer location being in the name of the creditor.
  - Creating a different export function.
    Currently, Transpas has no export function. By copy-pasting the data in the overview window of the TMS the partial shipments are not exported. This leads to missing information regarding transfer points and corresponding carriers. Creating another export function that includes the data of partial shipments could be a solution to this problem.
  - Linking (one-sided) to the database.
    Another solution for this problem is that the input for BigMile is gathered directly from the database. This has an added advantage that the sheet can be updated in nearly real-time. One-sided linking is advised in order to prevent unwanted edits in the original database.
- Country codes
  - Adjust the country codes.
    It is possible to add a script, as is done with the queries, for the country codes such that France and the countries that use ISO ALPHA-3 are converted to ISO ALPHA-2 (Nations Online Project, n.d.). However, it would be preferred to adjust the original list for where the country codes are stored to the ISO ALPHA-2 standard (Nations Online Project, n.d.).
- Postal codes
  - Verifying postal codes.
    Currently, for the most frequently used countries the postal codes are verified if they exist. However, there are still several countries for which the postal codes

are not verified. Therefore applying postal code verification for all postal codes is an option to guarantee correct departure and destination postal codes.

  - DPD, Fedex and GLS.
    One solution, which is done with the query, is to alter the end destination for the shipments that use DPD, Fedex or GLS, to the location of HST Group. This is the transfer point where DPD, Fedex and GLS collect their packages and send them to their final destination. This results in the shipment in the file containing the original departure location of the package and the destination location the location of HST Group. This therefore includes the whole part driven by HST Group. However, this is not preferred. When possible this solution should be implemented in the same way as the *Transfer point* previously mentioned.

- Loading metres

  - Calculating loading metres by deduction.
    The current solution is to calculate the missing values (*null*, negative and zero values). This can be done by deduction by size, amount of pallets, amount of packages (colli) that can be placed in a pallet box on average, etc. This is not the preferred solution as this can cause a less exact figure.
  - Demanding a non-zero value as input for loading metres.
    This can be done by calculating using the dimensions of a package or by other values such as amount of pallets. However, this is done up front and therefore the planner can make a more precise assumption of the size of the order.
  - Push a warning when extremities are used.
    Currently, it is possible to give extreme values (e.g. 200 loading metres) as input. Mostly these are typographical errors. It could prevent extreme values if the user gets a *"Are you sure about this value?"* prompt when a value is used above the size of a trailer (13.6 loading metres).

- Vehicle ID

  - Register the vehicle ID at the gas station
    As is visible in table 6 and 7 there is missing data for which truck (vehicle) has fueled up. In the future, a system should be implemented to link a license plate to a trip to the gas station, e.g. usage of a fuel card which is linked to a truck/ license plate.
  - Register the vehicle ID in the database
    The vehicle ID is also not always known. A system should be implemented to link a license plate to a (partial) shipment, such that in the future the data quality can be gold level (chapter 3.4). A possible solution for this is to request the license plate name when scanning the shipment.

7.2 Lean & Green data quality level including the other proposed solutions

In Chapter 3.4 the Lean & Green data quality levels are presented (Simons, 2020). In Chapter 6.3.5 it is stated that Silver is the current Lean & Green data quality level.

With the proposed measures in Chapter 7.1 the data quality level can be improved to Gold. This is mainly due to changes in the registration of *Vehicle IDs* and the partial shipments. All other measures improve the data quality level in a way that more correct

data can be used and therefore the scope is enlarged. These changes however do not make the difference between a Silver or Gold data quality level.

With the proposed changes the data quality level up to and including a fifth Lean & Green star can be achieved.

In order to achieve the Gold+ data quality level a lot of changes have to be made. Flow measurement systems have to be installed to track the fuel usage per trip, and the way of storing data has to be adjusted to exactly know which package has been transferred with which truck and trailer. This level of data quality is currently out of scope to achieve and is not needed for a fourth or fifth Lean & Green star.

8 Conclusion and discussion

In conclusion, the data analysis in this research has brought to attention issues concerning the data quality of the HST Group data sets regarding a BigMile analysis. In the research, Excel Queries were designed to improve the data quality. Without using the Excel Queries the data sets were not suitable for the Lean & Green input-file and therefore not suitable for a BigMile analysis. When using the designed Excel Queries a maximum data quality level of Silver can be achieved. This could result in a maximum of three Lean & Green Stars when only taking data quality into account. The main challenges in this research stem from the necessity to calculate loading metres, introducing inaccuracies, and the unavailability of data for partial shipments in the Transport Management System export. The absence of partial shipments in the exported data sets results in a lack of clarity regarding whether a (partial) shipment was executed by a contracted carrier or by HST Group.

While there are other issues identified in the research, their impact on data quality is comparatively lower than the issues mentioned above or was easily fixed using the queries. Addressing these issues is needed for a more accurate and reliable foundation for future Lean & Green Star assessments. Looking ahead, the identification of Vehicle IDs in both the Transport Management System and fuel usage exports from BP and AVIA come up as a requirement for achieving the fourth and fifth Lean & Green Stars.

However, it is crucial to address the limitations imposed by the time constraints of the study. The time-gated nature of the research restricted the possibility of verifying many values in the data sets such as postal codes and loading metres. These are both a study worth it on their own.

As a roadmap for future research, it is recommended to delve into the accuracy of *loading metres*, developing strategies to gather the data needed regarding *Vehicle IDs* and explore strategies to link HST Group datasets straight from the database to the BigMile analysis format. This would also possibly make it easier to gather information from partial shipments. All aim to prevent the current manual labour needed for analysis, enhancing the data quality and improving reliability.

References

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM computing surveys (CSUR)*, *41*(3), 1–52.

Cai, L., & Zhu, Y. (2015). The challenges of data quality and data quality assessment in the big data era. *Data science journal*, *14*, 2–2.

Connekt. (2023). Programma [Computer software manual]. Retrieved from `https://lean-green.nl/programma-co2-reductieplan/`

Heerkens, H., & van Winden, A. (2017). *Solving managerial problems systematically.* Noordhoff Uitgevers bv.

HSTGroup. (2023). Historie [Computer software manual]. Retrieved from `https://www.hst.nl/over-hst-groep/historie/`

Janssen, M., Van Der Voort, H., & Wahyudi, A. (2017). Factors influencing big data decision-making quality. *Journal of business research*, *70*, 338–345.

Juddoo, S. (2015). Overview of data quality challenges in the context of big data. In *2015 international conference on computing, communication and security (icccs).* (International Conference on Computing, Communication and Security (ICCCS), Pamplemousses, MAURITIUS, DEC 04-05, 2015)

McGilvray, D. (2021). *Executing data quality projects: Ten steps to quality data and trusted information (tm).* Academic Press.

Nations Online Project. (n.d.). *Country codes list - iso alpha-2, iso alpha-3 and numerical country codes.* Retrieved from `https://www.nationsonline.org/oneworld/country_code_list.htm#E`

Simons, L. (2020). Lean and green europe rule book [Computer software manual]. Retrieved from `https://www.lean-green.eu/app/uploads/2022/11/LeanGreen_RuleBook_V3.5.pdf`

Tayi, G. K., & Ballou, D. P. (1998). Examining data quality. *Communications of the ACM*, *41*(2), 54–57.

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of management information systems*, *12*(4), 5–33.

Appendix A Ignored dimensions

- **Dimension:** Authorization (Cai & Zhu, 2015)
  Definition: Whether an individual or organization has the right to use the data.
  Reason: This is not included in the research as all data was reachable within the company.

- **Dimension:** Metadata (Cai & Zhu, 2015)
  Definition: Describing different aspects of the datasets to reduce the problems caused by misunderstanding or inconsistencies
  Reason: The meaning of the data types and structures is not analysed in this research.

- **Dimension**: Fitness (Cai & Zhu, 2015)
  Definition: Two-level requirements: 1) the amount of accessed data used by users and 2) the degree to which the data produced matches users' needs in the aspects of indicator definition, elements, classification, etc.
  Reason: As the database is used for the planning of orders, the percentage used for the export of the database is not further studied.

- **Dimension:** Structure (Cai & Zhu, 2015)
  Definition: Structure refers to the level of difficulty in transforming semi-structured or unstructured data to structured data through technology.
  Reason: The data is structured already, therefore the dimension *Structure* is excluded.

- **Dimension:** Data coverage (McGilvray, 2021)
  Definition: The comprehensiveness of data available compared to the total data universe or population of interest.
  Reason: The data in this research is not compared to another sample.

- **Dimension:** Data decay (McGilvray, 2021)
  Definition: Rate of negative change to the data.
  Reason: The data used for this research does not have high volatility. The values in the data sets are from completed orders and are less subject to change. E.g. the transported loading metres always stay the same, as it is already transported. For example, a company's name can change in the system and is subject to data decay.

- **Dimension:** Credibility (Cai & Zhu, 2015) / Perception and relevance of trust (McGilvray, 2021) / Reputation (Wang & Strong, 1996)
  Definition: Believability of a source, dependent on reliability of data sources, data normalization, and time when the data is produced.
  Reason: The credibility and perception of relevance of trust require separate research to investigate this.

- **Dimension:** Definition/Documentation (Cai & Zhu, 2015) / Representational consistency (Wang & Strong, 1996)
  Definition: Data specification, including data name, definition, ranges of valid values, standard formats, business rules, etc.
  Reason: This dimension is not included in the study as no description besides column values was given in the datasets.

Appendix B Code for Excel Query International

In this appendix, the code for Excel Query International will be explained. Per line, or section of lines the function will be mentioned.

Start function of Excel Query:

let

Selecting the source file:

Bron = Excel.Workbook(File.Contents("K:/AUTOMATISERING/Lean & Green/Input 2022/Input international 2022/Transpas-International 2022.xlsx"), null, true), Blad1 _sheet = Bron[Item = "Blad1", Kind = "Sheet"][Data], #"Headersmetverhoogdniveau" = Table.PromoteHeaders(Blad1_Sheet, [PromoteAllScalars = true]),

Assign data type per column:

#"Type gewijzigd" = Table.TransformColumnTypes(#"Headers met verhoogd niveau", {{"Dossiernr", type text}, {"Volgnr", type text}, {"Status", type text}, {"Regelfacturering", type text}, {"Klaar voor factuur", type text}, {"Opdrachtgever", type text}, {"EDI referentie", type any}, {"Laaddatum", type date}, {"Losdatum", type date}, {"Plangroep", type text}, {"Dossierreferentie", type any}, {"Laadnaam", type text}, {"Laadadres", type any}, {"Laadpostcode", type text}, {"Laadplaats", type text}, {"Laadland", type text}, {"Losnaam", type text}, {"Losadres1", type text}, {"Lospostcode", type text}, {"Losplaats", type text}, {"Losland", type text}, {"Laadmeter", type number}, {"Zendingsoort", type text}, {"CMR nr.", Int64.Type}, {"Referentie", type any}, {"Ordersoort", type any}, {"ADR-zending", type text}, {"Chauffeur lossen", type text}, {"Ritnr lossen", Int64.Type}, {"EDI provider", type text}, {"Opvolgtype", type any}, {"Afleverstatus opm.", type text}, {"# Geweigerd", type any}, {"# Beschadigd", type any}, {"# Manco", type any}, {"Afleverstatus", type text}, {"Incoterm", type text}, {"# afwijkingen", type any}, {"Consolidatienr", Int64.Type}, {"Contactpersoon", type any}, {"Voorplancode", type text}, {"Order naar Portal", type any}, {"Factuurnr", Int64.Type}, {"Losadres2", type any}, {"Geen autotarief", type text}, {"Week", Int64.Type}, {"Departement", type text}, {"Crediteur lossen", type text}, {"Trailer lossen", type text}, {"Truck lossen", type text}, {"Crediteur laden", type text}, {"Trailer laden", type text}, {"Hoogte", Int64.Type}, {"Breedte", Int64.Type}, {"Afleverstatus tijd", type any}, {"Lengte", type number}, {"Org. Gewicht", type number}, {"Gordelomvang", type any}, {"Colli", Int64.Type}, {"Palletplaatsen", type number}, {"Volume", type number}, {"Gewicht", type number}, {"Eenheid", type text}, {"Aantal", type number}, {"Losadres", type text}, {"Losreferentie", type any}, {"Lostijd", type datetime}, {"Laadreferentie", type any}, {"Laadtijd", type datetime}, {"2e chauffeur lossen", type any}, {"Foutvracht", type text}, {"Plangroep losrit", type text}, {"# Goederenregels", Int64.Type}, {"Chauffeur laadinfo", type any}, {"Chauffeur losinfo", type any}, {"Lostijd t/m", type datetime}, {"Creditnota", type any}, {"Legnr", Int64.Type}, {"Geëxporteerd naar netwerk", type any}, {"# documenten", Int64.Type}, {"Debiteur", type any}, {"Transportsoort", type text}, {"Afstand", type number}, {"Route nr.", type any}, {"Aangemaakt op", type datetime}, {"Crossdock laadadres", type text}, {"Niet Overladen", type text}, {"Chauffeur laden", type text}, {"Truck laden", type text}, {"Ritnr laden", Int64.Type}, {"Laden faxnr", type any}}),

Combining Dossiernr and Volgnr into an unique value:

#"Aangepaste kolom toegevoegd" = Table.AddColumn(#"Type gewijzigd", "Dossiernr

+ Volgnr", each[Dossiernr]&" - "&[Volgnr]),

Altering the order of the columns:

#"Volgorde van kolommen gewijzigd" = Table.ReorderColumns(#"Aangepaste kolom toegevoegd", {"Dossiernr + Volgnr", "Dossiernr", "Volgnr", "Status", "Regelfacturering", "Klaar voor factuur", "Opdrachtgever", "EDI referentie", "Laaddatum", "Losdatum", "Plangroep", "Dossierreferentie", "Laadnaam", "Laadadres", "Laadpostcode", "Laadplaats", "Laadland", "Losnaam", "Losadres1", "Lospostcode", "Losplaats", "Losland", "Laadmeter", "Zendingsoort", "CMR nr.", "Referentie", "Ordersoort", "ADR-zending", "Chauffeur lossen", "Ritnr lossen", "EDI provider", "Opvolgtype", "Afleverstatus opm.", "# Geweigerd", "# Beschadigd", "# Manco", "Afleverstatus", "Incoterm", "# afwijkingen", "Consolidatienr", "Contactpersoon", "Voorplancode", "Order naar Portal", "Factuurnr", "Losadres2", "Geen autotarief", "Week", "Departement", "Crediteur lossen", "Trailer lossen", "Truck lossen", "Crediteur laden", "Trailer laden", "Hoogte", "Breedte", "Afleverstatus tijd", "Lengte", "Org. Gewicht", "Gordelomvang", "Colli", "Palletplaatsen", "Volume", "Gewicht", "Eenheid", "Aantal", "Losadres", "Losreferentie", "Lostijd", "Laadreferentie", "Laadtijd", "2e chauffeur lossen", "Foutvracht", "Plangroep losrit", "# Goederenregels", "Chauffeur laadinfo", "Chauffeur losinfo", "Lostijd t/m", "Creditnota", "Legnr", "Geëxporteerd naar netwerk", "# documenten", "Debiteur", "Transportsoort", "Afstand", "Route nr.", "Aangemaakt op", "Crossdock laadadres", "Niet Overladen", "Chauffeur laden", "Truck laden", "Ritnr laden", "Laden faxnr"}),

Changing the type of a few columns:

#"Changed Type3" = Table.TransformColumnTypes(#"Volgorde van kolommen gewijzigd", {{"Dossiernr + Volgnr", type text}}),

Deleting the columns Dossiernr and Volgnr:

#"Kolommen verwijderd" = Table.RemoveColumns(#"Changed Type3","Dossiernr", "Volgnr"),

Filtering out the shipments that were cancelled:

#"Rijen gefilterd" = Table.SelectRows(#"Kolommen verwijderd", each ([Status] <> "Geannuleerd")),

Deleting columns which are not needed:

#"Kolommen verwijderd1" = Table.RemoveColumns(#"Rijen gefilterd",{"Regelfacturering", "Klaar voor factuur", "EDI referentie", "Dossierreferentie", "Zendingsoort", "CMR nr.", "Referentie", "Ordersoort", "ADR-zending", "Chauffeur lossen", "Ritnr lossen", "EDI provider", "Opvolgtype", "Afleverstatus opm.", "# Geweigerd", "# Beschadigd", "# Manco", "Afleverstatus", "Incoterm", "# afwijkingen", "Consolidatienr", "Contactpersoon", "Voorplancode", "Order naar Portal", "Factuurnr", "Losadres2", "Geen autotarief", "Departement", "Afleverstatus tijd", "Palletplaatsen", "Losreferentie", "Lostijd", "Laadreferentie", "Laadtijd", "2e chauffeur lossen", "Foutvracht", "# Goederenregels", "Chauffeur laadinfo", "Chauffeur losinfo", "Lostijd t/m", "Creditnota", "Legnr", "# documenten", "Debiteur", "Transportsoort", "Route nr.", "Aangemaakt op", "Crossdock laadadres", "Chauffeur laden", "Truck laden", "Ritnr laden", "Laden faxnr"}),

Filtering out the month from the Laaddatum column in order to generate the [year]-[month] format for a column. This is the period for energy usage:

#"Ingevoegde tekst tussen scheidingstekens" = Table.AddColumn(#"Kolommen verwijderd1", "Tekst tussen scheidingstekens", each Text.BetweenDelimiters(Text.From([Laaddatum], "nl-NL"), "-", "-"), type text),

Changing the name of the column to Month (Maand):

#"Namen van kolommen gewijzigd" = Table.RenameColumns(#"Ingevoegde tekst tussen scheidingstekens","Tekst tussen scheidingstekens", "Maand"),

Adding a [year]-[month] column which is needed for the period input value:

#"Aangepaste kolom toegevoegd1" = Table.AddColumn(#"Namen van kolommen gewijzigd", "Jaar en maand", each "2022 - "&[Maand]),

Adding a new calculation for loading metres:

*#"Aangepaste kolom toegevoegd2" = Table.AddColumn(#"Aangepaste kolom toegevoegd1", "Nieuwe berekening laadmeters", each if ([Eenheid] = "Colli") then (0.5 /40 * [Aantal]) else if ([Eenheid] = "Pakket") then (0.5 /40 * [Aantal]) else if ([Eenheid] = "COL") then (0.5 /40 * [Aantal]) else if ([Eenheid] = "COL zending") then (0.5 /40 * [Aantal]) else if [Plangroep] = "Ritopdracht" then 13.6 else if ([Eenheid] = "ICOL") then (0.5 /40 * [Aantal]) else if ([Eenheid] = "ICOL zending") then (0.5 /40 * [Aantal]) else if ([Plangroep] = "Nachtpendel") then 13.6 else if ([Eenheid]= "BUN") then ([Lengte]/100*[Breedte]/100*0.4)*[Aantal] else if ([Eenheid] = "ENV") then (0.5 /40 * [Aantal]) else if ([Eenheid]= "HSPL") then ([Lengte]/100*[Breedte]/100*0.4)*[Aantal] else if ([Eenheid]= "BIKE") then ([Lengte]/100*[Breedte]/100*0.4)*[Aantal] else if ([Eenheid]= "EWP") then ([Aantal]*0.4) else if ([Eenheid]= "SP") then ([Aantal]*0.4) else if ([Eenheid]= "M3") then ([Aantal]/90*13.6) else if ([Eenheid]= "Uur") then 13.6 else 0),*

If statement added to use the new calculation if the old value is null, less or equal to 0, or bigger than 13.6:

#"Voorwaardelijke kolom toegevoegd" = Table.AddColumn(#"Aangepaste kolom toegevoegd2", "Nieuwe waarde laadmeters", each if [Laadmeter] = null then [Nieuwe berekening laadmeters] else if [Laadmeter] <= 0 then [Nieuwe berekening laadmeters] else if [Laadmeter] > 0 and [Laadmeter] <= 13.6 then [Laadmeter] else if [Laadmeter] > 13.6 then [Nieuwe berekening laadmeters] else null),

Replacing incorrect country codes with the right standard:

#"Replace F FR" = Table.ReplaceValue(#"Voorwaardelijke kolom toegevoegd","F" ,"FR",Replacer.ReplaceValue,"Laadland"),
#"Replace los F FR" = Table.ReplaceValue(#"Replace F FR","F","FR", Replacer.ReplaceValue,"Losland"),
#"Replace EST EE" = Table.ReplaceValue(#"Replace los F FR","EST","EE", Replacer.ReplaceValue,"Laadland"),
#"Replace los EST EE" = Table.ReplaceValue(#"Replace EST EE","EST","EE", Replacer.ReplaceValue,"Losland"),
#"Replace MLT MT" = Table.ReplaceValue(#"Replace los EST EE","MLT","MT", Replacer.ReplaceValue,"Laadland"),
#"Replace los MLT MT" = Table.ReplaceValue(#"Replace MLT MT","MLT","MT", Replacer.ReplaceValue,"Losland"),
#"Replace RSM SM" = Table.ReplaceValue(#"Replace los MLT MT","RSM","SM", Replacer.ReplaceValue,"Laadland"),

#"Replace los RSM SM" = Table.ReplaceValue(#"Replace RSM SM","RSM","SM", Replacer.ReplaceValue,"Losland"),
#"Replace Est EE" = Table.ReplaceValue(#"Replace los RSM SM","Est","EE", Replacer.ReplaceValue,"Laadland"),
#"Replace los Est EE" = Table.ReplaceValue(#"Replace Est EE","Est","EE", Replacer.ReplaceValue,"Losland"),

Adding a new departure country:

#"Nieuw Laadland" = Table.AddColumn(#"Replace los Est EE", "Nieuw Laadland", each [Laadland]),

Adding a new departure postal code when the value length is one short, then a zero at the start is missing:

#"Nieuw Laadpostcode" = Table.AddColumn(#"Nieuw Laadland", "Nieuw Laadpostcode", each if Text.Length([Laadpostcode]) = 4 and ([Laadland] = "DE" or [Laadland] = "ES" or [Laadland] = "FI") then "0"&[Laadpostcode] else [Laadpostcode]),

If a loading creditor is used the value will be Enschede, else the previous place of departure:

#"Added Conditional Column" = Table.AddColumn(#"Nieuw Laadpostcode", "Nieuw Laadplaats", each if [Crediteur laden] <> null then "Enschede" else [Laadplaats]),

Splitting the Truck lossen column in order to possibly filter out the possible transfer location:

#"Nieuw mogelijk losplaats" = Table.SplitColumn(#"Added Conditional Column", "Truck lossen", Splitter.SplitTextByEachDelimiter("-", QuoteStyle.None, true), "Truck lossen.1", "Truck lossen.2"),

Changing the type of the split columns to text:

#"Changed Type" = Table.TransformColumnTypes(#"Nieuw mogelijk losplaats","Truck lossen.1", type text, "Truck lossen.2", type text),

Filtering out the country code from the text:

#"Nieuw mogelijk losland" = Table.SplitColumn(#"Changed Type", "Truck lossen.1", Splitter.SplitTextByPositions(0, 2, true), "Truck lossen.1.1", "Truck lossen.1.2"),

Changing the type of the split columns to text:

#"Changed Type1" = Table.TransformColumnTypes(#"Nieuw mogelijk losland","Truck lossen.1.1", type text, "Truck lossen.1.2", type text),

Adding a zero to the postal code if it is missing:

#"Nieuw Lospostcode" = Table.AddColumn(#"Changed Type1", "Nieuw Lospostcode", each if Text.Length([Lospostcode]) = 4 and ([Losland] = "DE" or [Losland] = "ES" or [Losland] = "FI") then "0"&[Lospostcode] else [Lospostcode]),

Remove not needed columns:

#"Removed Columns" = Table.RemoveColumns(#"Nieuw Lospostcode","Trailer lossen", "Week", "Trailer laden", "Afstand", "Niet Overladen", "Maand", "Nieuwe berekening laadmeters", "Hoogte", "Breedte", "Lengte", "Org. Gewicht", "Gordelomvang", "Colli",

"Volume", "Gewicht", "Laadmeter", "Status", "Geëxporteerd naar netwerk"),

Adding a temporary transfer location country code:

 # "Depotland" = Table.AddColumn(#"Removed Columns", "DepotLandTijdelijk", each if List.Contains({"NL", "DK", "DE", "CZ", "SE", "BE", "FR", "NO", "FI", "AT", "PL", "CH", "IT", "HU", "HR", "LU", "ES", "LI", "GR", "PT", "BG", "SK", "RO", "GB", "LT", "AU", "SI", "RS", "IE", "EE", "TR", "IJ", "CY", "LV", "MT"},[Truck lossen.1.2]) then [Truck lossen.1.2] else null), DepotPlaats = Table.AddColumn(Depotland, "DepotPlaats", each if [Truck lossen.2] <> null and Text.Length([Truck lossen.2]) > 2 and [DepotLandTijdelijk] <> null then [Truck lossen.2] else null),

Adding a country code for the transfer location:

#"Added DepotLand" = Table.AddColumn(DepotPlaats, "DepotLand", each if [DepotLandTijdelijk] <> null and [DepotPlaats] <> null then [DepotLandTijdelijk] else null),

Deleting the temporary transfer location country code:

#"Removed DepotLandTijdelijk" = Table.RemoveColumns(#"Added DepotLand","DepotLandTijdelijk"),

Adding a column with the if statement that when there is a loading creditor the given postal code is used:

HSTDepotLaden = Table.AddColumn(#"Removed DepotLandTijdelijk", "HSTDepotLaden", each if [Crediteur laden] <> null then "7547 RW" else null),

Adding a column with the if statement that when there is an unloading creditor the given postal code is used:

HSTDepotLossen = Table.AddColumn("HSTDepotLaden", "HSTDepotLossen", each if [Crediteur lossen] <> null and [DepotLand] = null then "7547 RW" else null),

Duplicating the departure postal code, HSTDepotLossen, HSTDepotLaden, DepotPlaats columns:

#"Duplicated Column" = Table.DuplicateColumn(HSTDepotLossen, "Laadpostcode", "Laadpostcode - Copy"),
#"Duplicated Column1" = Table.DuplicateColumn(#"Duplicated Column", "HSTDepotLossen", "HSTDepotLossen - Copy"),
#"Duplicated Column2" = Table.DuplicateColumn(#"Duplicated Column1", "HSTDepotLaden", "HSTDepotLaden - Copy"),
#"Duplicated Column3" = Table.DuplicateColumn(#"Duplicated Column2", "DepotPlaats", "DepotPlaats - Copy"),

Reordering the columns:

#"Reordered Columns" = Table.ReorderColumns(#"Duplicated Column3","Dossiernr + Volgnr", "Opdrachtgever", "Laaddatum", "Losdatum", "Plangroep", "Laadnaam", "Laadadres", "Laadpostcode", "Laadplaats", "Laadland", "Losnaam", "Losadres1", "Lospostcode", "Losplaats", "Losland", "Crediteur lossen", "Truck lossen.1.1", "Truck lossen.1.2", "Truck lossen.2", "Crediteur laden", "Eenheid", "Aantal", "Losadres", "Plangroep losrit", "Jaar en maand", "Nieuwe waarde laadmeters", "Nieuw Laadland", "Nieuw Laadpostcode", "Nieuw Laadplaats", "Nieuw Lospostcode", "DepotPlaats", "DepotLand", "HSTDepotLaden", "HSTDepotLossen", "Laadpostcode - Copy", "HSTDepotLaden - Copy", "HSTDepotLossen - Copy", "DepotPlaats - Copy"),

Unpivoting columns in order to create partial shipments:

#"Unpivoted Columns" = Table.UnpivotOtherColumns(#"Reordered Columns", "Dossiernr + Volgnr", "Opdrachtgever", "Laaddatum", "Losdatum", "Plangroep", "Laadnaam", "Laadadres", "Laadpostcode", "Laadplaats", "Laadland", "Losnaam", "Losadres1", "Lospostcode", "Losplaats", "Losland", "Crediteur lossen", "Truck lossen.1.1", "Truck lossen.1.2", "Truck lossen.2", "Crediteur laden", "Eenheid", "Aantal", "Losadres", "Plangroep losrit", "Jaar en maand", "Nieuwe waarde laadmeters", "Nieuw Laadland", "Nieuw Laadpostcode", "Nieuw Laadplaats", "Nieuw Lospostcode", "DepotPlaats", "DepotLand", "HSTDepotLaden", "HSTDepotLossen", "Attribute", "Laadplek rit"),

Adding the destination postal code of the partial shipment: *#"Lospostcode rit" = Table.AddColumn(#"Unpivoted Columns", "Lospostcode rit", each if [Attribute] = "DepotPlaats - Copy" then [Nieuw Lospostcode] else if [Crediteur lossen] = null and [Crediteur laden] = null then [Nieuw Lospostcode] else if [HSTDepotLaden] <> null and [Attribute] = "Laadpostcode - Copy" then [HSTDepotLaden] else if [HSTDepotLossen] <> null and [Attribute] = "Laadpostcode - Copy" then [HSTDepotLossen] else if [HSTDepotLossen] = null and [Attribute] = "HSTDepotLaden - Copy" and [DepotPlaats] = null then [Nieuw Lospostcode] else if [HSTDepotLossen] <> null and [Attribute] = "HSTDepotLossen - Copy" then [Nieuw Lospostcode] else if [HSTDepotLaden] <> null and [HSTDepotLossen] <> null and [Attribute] = "HSTDepotLaden - Copy" then [HSTDepotLossen] else null),*

Adding the destination city of the partial shipment. Only if there is no postal code given:

#"Losplaats rit" = Table.AddColumn(#"Lospostcode rit", "Losplaats Rit", each if [Attribute] <> "DepotPlaats - Copy" and [DepotPlaats] <> null then [DepotPlaats] else null),

Adding the departure city of the partial shipment. Only if there is no postal code given:

#"Laadplaats rit" = Table.AddColumn(#"Losplaats rit", "Laadplaats rit", each if [Attribute] = "DepotPlaats - Copy" then [Laadplek rit] else null),

Adding the departure postal code of the partial shipment:

#"Laadpostcode rit" = Table.AddColumn(#"Laadplaats rit", "Laadpostcode rit", each if [Attribute] <> "DepotPlaats - Copy" then [Laadplek rit] else null),

Adding the column to know if a charter is used for the partial shipment:

Charter = Table.AddColumn(#"Laadpostcode rit", "Charter", each if [Attribute] = "Laadpostcode - Copy" and [HSTDepotLaden] <> null then "yes" else if [Attribute] = "HSTDepotLossen - Copy" or [Attribute] = "DepotPlaats - Copy" then "yes" else "no"),

Adding the destination country for the partial shipment: *#"Losland rit" = Table.AddColumn(Charter, "Losland rit", each if [Attribute] = "DepotPlaats - Copy" or [Attribute] = "HSTDepotLossen - Copy" or ([Crediteur lossen] = null and [Crediteur laden] = null) then [Losland] else if [HSTDepotLaden] <> null and [Attribute] = "Laadpostcode - Copy" then "NL" else if [HSTDepotLossen] <> null and [Attribute] = "Laadpostcode - Copy" then "NL" else if [HSTDepotLossen] = null and [Attribute] = "HSTDepotLaden - Copy" and [DepotPlaats] = null then [Losland] else if [Attribute] = "HSTDepotLaden -*

Copy" then "NL" else if [Attribute] = "Laadpostcode - Copy" and [DepotLand] <> null then [DepotLand] else null),

Adding the departure country for the partial shipment:

#"Laadland rit" = Table.AddColumn(#"Losland rit", "Laadland rit", each if [Attribute] = "DepotPlaats - Copy" then [Nieuw Laadland] else if [Attribute] = "HSTDepotLossen - Copy" then "NL" else if [Attribute] = "Laadpostcode - Copy" then [Nieuw Laadland] else if [Attribute] = "HSTDepotLaden - Copy" then "NL" else null),

Changing the type of the columns to the correct ones:

#"Changed Type2" = Table.TransformColumnTypes(#"Laadland rit","Laadadres", type text, "Jaar en maand", type text, "Nieuwe waarde laadmeters", type number, "Nieuw Laadland", type text, "Nieuw Laadpostcode", type text, "Nieuw Laadplaats", type text, "Nieuw Lospostcode", type text, "DepotPlaats", type text, "DepotLand", type text, "HSTDepotLaden", type text, "HSTDepotLossen", type text, "Laadplek rit", type text, "Lospostcode rit", type text, "Losplaats Rit", type text, "Laadplaats rit", type text, "Laadpostcode rit", type text, "Charter", type text, "Losland rit", type text, "Laadland rit", type text),

Removing rows with errors due to empty cells: *#"Removed Errors" = Table.RemoveRowsWithErrors(#"Changed Type2", "Dossiernr + Volgnr"),*

The following code is in order to number the partial shipments: *#"Grouped Rows" = Table.Group(#"Removed Errors", "Dossiernr + Volgnr", "Count", each _, type table [#"Dossiernr + Volgnr"=nullable text, Opdrachtgever=nullable text, Laaddatum=nullable date, Losdatum=nullable date, Plangroep=nullable text, Laadnaam=nullable text, Laadadres=nullable text, Laadpostcode=nullable text, Laadplaats=nullable text, Laadland=nullable text, Losnaam=nullable text, Losadres1=nullable text, Lospostcode=nullable text, Losplaats=nullable text, Losland=nullable text, Crediteur lossen=nullable text, Truck lossen.1.1=nullable text, Truck lossen.1.2=nullable text, Truck lossen.2=nullable text, Crediteur laden=nullable text, Eenheid=nullable text, Aantal=nullable number, Losadres=nullable text, Plangroep losrit=nullable text, Jaar en maand=nullable text, Nieuwe waarde laadmeters=nullable number, Nieuw Laadland=nullable text, Nieuw Laadpostcode=nullable text, Nieuw Laadplaats=nullable text, Nieuw Lospostcode=nullable text, DepotPlaats=nullable text, DepotLand=nullable text, HSTDepotLaden=nullable text, HSTDepotLossen=nullable text, Attribute=text, Laadplek rit=nullable text, Lospostcode rit=nullable text, Losplaats Rit=nullable text, Laadplaats rit=nullable text, Laadpostcode rit=nullable text, Charter=nullable text, Losland rit=nullable text, Laadland rit=nullable text]),*
#"Added Custom" = Table.AddColumn(#"Grouped Rows", "Custom", each Table.AddIndexColumn([Count],
#"Removed Other Columns" = Table.SelectColumns(#"Added Custom","Custom"),
#"Expanded Custom" = Table.ExpandTableColumn(#"Removed Other Columns", "Custom", "Dossiernr + Volgnr", "Opdrachtgever", "Laaddatum", "Losdatum", "Plangroep", "Laadnaam", "Laadadres", "Laadpostcode", "Laadplaats", "Laadland", "Losnaam", "Losadres1", "Lospostcode", "Losplaats", "Losland", "Crediteur lossen", "Truck lossen.1.1", "Truck lossen.1.2", "Truck lossen.2", "Crediteur laden", "Eenheid", "Aantal", "Losadres", "Plangroep losrit", "Jaar en maand", "Nieuwe waarde laadmeters", "Nieuw Laadland", "Nieuw Laadpostcode", "Nieuw Laadplaats", "Nieuw Lospostcode", "DepotPlaats", "DepotLand", "HSTDepotLaden", "HSTDepotLossen", "Attribute", "Laadplek rit", "Lospostcode rit",

"Losplaats Rit", "Laadplaats rit", "Laadpostcode rit", "Charter", "Losland rit", "Laadland rit", "Index", "Dossiernr + Volgnr", "Opdrachtgever", "Laaddatum", "Losdatum", "Plangroep", "Laadnaam", "Laadadres", "Laadpostcode", "Laadplaats", "Laadland", "Losnaam", "Losadres1", "Lospostcode", "Losplaats", "Losland", "Crediteur lossen", "Truck lossen.1.1", "Truck lossen.1.2", "Truck lossen.2", "Crediteur laden", "Eenheid", "Aantal", "Losadres", "Plangroep losrit", "Jaar en maand", "Nieuwe waarde laadmeters", "Nieuw Laadland", "Nieuw Laadpostcode", "Nieuw Laadplaats", "Nieuw Lospostcode", "DepotPlaats", "DepotLand", "HSTDepotLaden", "HSTDepotLossen", "Attribute", "Laadplek rit", "Lospostcode rit", "Losplaats Rit", "Laadplaats rit", "Laadpostcode rit", "Charter", "Losland rit", "Laadland rit", "Index"),

Changing the types of two columns to the correct ones: *#"Changed Type4" = Table.TransformColumnTypes(#"Expanded Custom","Laaddatum", type date, "Losdatum", type date),*

Adding a filter to the destination city of the partial shipment:

#"Added Custom1" = Table.AddColumn(#"Changed Type4", "Losplaats rit filter", each if [Losplaats Rit] <> null and [Charter] = "no" then [Losplaats Rit] else null),

Reordering the columns:

#"Reordered Columns1" = Table.ReorderColumns(#"Added Custom1","Dossiernr + Volgnr", "Opdrachtgever", "Laaddatum", "Losdatum", "Plangroep", "Laadnaam", "Laadadres", "Laadpostcode", "Laadplaats", "Laadland", "Losnaam", "Losadres1", "Lospostcode", "Losplaats", "Losland", "Crediteur lossen", "Truck lossen.1.1", "Truck lossen.1.2", "Truck lossen.2", "Crediteur laden", "Eenheid", "Aantal", "Losadres", "Plangroep losrit", "Jaar en maand", "Nieuwe waarde laadmeters", "Nieuw Laadland", "Nieuw Laadpostcode", "Nieuw Laadplaats", "Nieuw Lospostcode", "DepotPlaats", "DepotLand", "HSTDepotLaden", "HSTDepotLossen", "Attribute", "Laadplek rit", "Lospostcode rit", "Losplaats Rit", "Losplaats rit filter", "Laadplaats rit", "Laadpostcode rit", "Charter", "Losland rit", "Laadland rit", "Index"),

Removing a column in order to create a new destination city column for the partial shipments:

#"Removed Columns1" = Table.RemoveColumns(#"Reordered Columns1","Losplaats Rit"),

Renaming a column:

#"Renamed Columns" = Table.RenameColumns(#"Removed Columns1","Losplaats rit filter", "Losplaats rit") in #"Renamed Columns"

Appendix C Code for Excel Query National 2019-2020

In this appendix, the code for Excel Query National for the time-frame 2019-2020 will be explained. Per line, or section of lines the function will be mentioned. The back-ups were stored in multiple files: HISTVERT and HISTVRTL. These files needed to be linked together using the last script, Merge.

C.1 HISTVERT

Starting the Excel Query script:
let

Selecting the source files in the folder and importing them:
Source = Folder.Files("K:\AUTOMATISERING\Lean & Green\bestanden 2019 2020 \Querymap\HISTVERT"), #"Filtered Hidden Files1" = Table.SelectRows(Source, each [Attributes]?[Hidden]? <> true), #"Invoke Custom Function1" = Table.AddColumn(#"Filtered Hidden Files1", "Transform File", each #"Transform File"([Content])), #"Renamed Columns1" = Table.RenameColumns(#"Invoke Custom Function1", "Name", "Source.Name"), #"Removed Other Columns1" = Table.SelectColumns(#"Renamed Columns1", "Source.Name", "Transform File"), #"Expanded Table Column1" = Table.ExpandTableColumn(#"Removed Other Columns1", "Transform File", Table.ColumnNames(#"Transform File"(#"Sample File"))), #"Changed Type" = Table.TransformColumnTypes(#"Expanded Table Column1",{"Source.Name", type text}, {"DOCUMENT", type logical}, {"SPOED", type logical}, {"ROUTE", type text}, {"DTBOEK", type date}, {"NRVRVEXP", Int64.Type}, {"NRPRINT", Int64.Type}, {"TEKST", type text}, {"NRWAGEN", Int64.Type}, {"NRVERV", Int64.Type}, {"VANVRL", type logical}, {"ZNDCORRECT", type logical}, {"VVAGEPRINT", type logical}, {"VERZONDEN", type logical}, {"MANCO", type logical}, {"HANGEND", type logical}, {"BETWYZ", type text}, {"BERNEG", type logical}, {"AFHSRT", Int64.Type}, {"AFHSTS", type any}, {"INCPRV", Int64.Type}, {"DEPEXTRA", Int64.Type}, {"GEAVERSCH", Int64.Type}, {"KDVALU", type text}, {"BDRRMB", Int64.Type}, {"GEAKDLAND", Int64.Type}, {"GEAPLAATS", type text}, {"GEAPOST", type text}, {"GEAHUISNR", type text}, {"GEASTRAAT", type text}, {"GEACONTPS", type text}, {"GEANAAM", type text}, {"KDADRES", type text}, {"GEAPHONE", Int64.Type}, {"AFZKDLAND", Int64.Type}, {"AFZPLAATS", type text}, {"AFZPOST", type text}, {"AFZHUISNR", type text}, {"AFZSTRAAT", type text}, {"AFZCONTPS", type text}, {"AFZNAAM", type text}, {"KDAFZ", type text}, {"NRDEPAAN", Int64.Type}, {"NRORDER", type text}, {"NRVVA", Int64.Type}, {"FRANK", Int64.Type}, {"DTZEND", type date}, {"NRZEND", Int64.Type}, {"NRVERL", Int64.Type}, {"NRDEPVTR", Int64.Type}, {"ZNDSOORT", type text}, {"TPOPRINTED", type logical}, {"HOLD", type logical}, {"LASTCOLLO", Int64.Type}, {"VRLBEREK", type logical}, {"BVRVRWVRL", type logical}, {"BVRVRWOND", type logical}, {"BVRVRWVRV", type logical}, {"AFGEHAALD", type logical}, {"WAGEN", type any}, {"BORDERO", type any}, {"RESTANTE", type logical}, {"DUBSTOP", Int64.Type}, {"PLANTIJDV", Int64.Type}, {"PLANTIJDT", type any}, {"RMBPRV", Int64.Type}, {"RMBPRVOND", Int64.Type}, {"RMBPRVBST", Int64.Type}, {"GEAHUISNR2", type any}, {"EMAILADRES", type text}, {"BTOC", type logical}, {"NIGHT", type logical}, {"ZATERDAG", type logical}),

Removing not needed columns:
#"Removed Columns" = Table.RemoveColumns(#"Changed Type",{"Source.Name", "DOCUMENT", "SPOED", "NRVRVEXP", "NRPRINT", "TEKST", "ZNDCORRECT", "MANCO",

"HANGEND", "BETWYZ", "BERNEG", "GEAVERSCH", "DEPEXTRA", "INCPRV", "KDVALU", "GEACONTPS", "KDADRES", "GEAPHONE", "AFZCONTPS", "KDAFZ", "NRORDER", "TPOPRINTED", "HOLD", "LASTCOLLO", "AFGEHAALD", "WAGEN", "BORDERO", "RESTANTE", "DUBSTOP", "PLANTIJDV", "PLANTIJDT", "RMBPRV", "RMBPRVOND", "RMBPRVBST", "GEAHUISNR2", "EMAILADRES", "BTOC", "NIGHT", "ZATERDAG", "BVRVRWVRV", "BVRVRWOND", "BVRVRWVRL", "VRLBEREK", "NRWAGEN", "NRVERV", "VANVRL", "VVAGEPRINT", "VERZONDEN", "AFHSTS", "BDRRMB", "GEASTRAAT", "AFZSTRAAT", "NRDEPAAN", "NRDEPVTR"),

Changing the column data types of two columns:

#"Changed Type1" = Table.TransformColumnTypes(#"Removed Columns","NRVERL", type text, "NRZEND", type text),

Creating an unique number which is needed to order to merge the two files:

#"Unieknr toegevoegd" = Table.AddColumn(#"Changed Type1", "Unieknr", each [NRZEND][NRVERL]),

Adding in a column in order to determine (/decode) the country of destination:

Losland = Table.AddColumn(#"Unieknr toegevoegd", "Losland", each if [GEAKDLAND] = 1 then "FR" else if [GEAKDLAND] = 2 then "BE" else if [GEAKDLAND] = 3 then "NL" else if [GEAKDLAND] = 4 then "DE" else if [GEAKDLAND] = 5 then "IT" else if [GEAKDLAND] = 6 then "GB" else if [GEAKDLAND] = 8 then "DK" else if [GEAKDLAND] = 11 then "ES" else if [GEAKDLAND] = 19 then "LU" else if [GEAKDLAND] = 28 then "NO" else if [GEAKDLAND] = 32 then "FI" else if [GEAKDLAND] = 38 then "AT" else if [GEAKDLAND] = 53 then "EE" else if [GEAKDLAND] = 56 then "LT" else if [GEAKDLAND] = 60 then "PL" else if [GEAKDLAND] = 61 then "CZ" else if [GEAKDLAND] = 63 then "SK" else if [GEAKDLAND] = 68 then "BG" else if [GEAKDLAND] = 77 then "AM" else if [GEAKDLAND] = 36 then "CH" else if [GEAKDLAND] = 288 then "PT" else if [GEAKDLAND] = 400 then "US" else if [GEAKDLAND] = 7 then "IE" else if [GEAKDLAND] = 66 then "RO" else if [GEAKDLAND] = 30 then "SE" else if [GEAKDLAND] = 37 then "LI" else if [GEAKDLAND] = 388 then "ZA" else if [GEAKDLAND] = 528 then "AR" else if [GEAKDLAND] = 92 then "HR" else if [GEAKDLAND] = 10 then "PT" else if [GEAKDLAND] = 64 then "HU" else if [GEAKDLAND] = 509 then "BR" else if [GEAKDLAND] = 720 then "CN" else if [GEAKDLAND] = 91 then "SI" else if [GEAKDLAND] = 54 then "LV" else if [GEAKDLAND] = 480 then "CO" else if [GEAKDLAND] = 21 then "ES" else if [GEAKDLAND] = 46 then "LV" else if [GEAKDLAND] = 9 then "GR" else if [GEAKDLAND] = 701 then "MY" else if [GEAKDLAND] = 404 then "CA" else if [GEAKDLAND] = 600 then "CY" else if [GEAKDLAND] = 732 then "JP" else if [GEAKDLAND] = 728 then "KR" else if [GEAKDLAND] = 75 then "RU" else if [GEAKDLAND] = 636 then "KW" else if [GEAKDLAND] = 740 then "HK" else if [GEAKDLAND] = 664 then "IN" else if [GEAKDLAND] = 800 then "AU" else if [GEAKDLAND] = 700 then "ID" else if [GEAKDLAND] = 44 then "DE" else if [GEAKDLAND] = 412 then "MX" else if [GEAKDLAND] = 736 then "TW" else if [GEAKDLAND] = 804 then "NZ" else if [GEAKDLAND] = 72 then "UA" else if [GEAKDLAND] = 0 then "NL" else if [GEAKDLAND] = 24 then "IS" else if [GEAKDLAND] = 41 then "FO" else if [GEAKDLAND] = 42 then "MC" else if [GEAKDLAND] = 55 then "TR" else if [GEAKDLAND] = 70 then "AL" else if [GEAKDLAND] = 73 then "BY" else if

[GEAKDLAND] = 76 then "GE" else if [GEAKDLAND] = 93 then "BA" else if [GEAKDLAND] = 98 then "RS" else if [GEAKDLAND] = 204 then "MA" else if [GEAKDLAND] = 220 then "EG" else if [GEAKDLAND] = 272 then "CI" else if [GEAKDLAND] = 276 then "GH" else if [GEAKDLAND] = 318 then "CD" else if [GEAKDLAND] = 324 then "RW" else if [GEAKDLAND] = 346 then "KE" else if [GEAKDLAND] = 372 then "EE" else if [GEAKDLAND] = 373 then "MU" else if [GEAKDLAND] = 382 then "CH" else if [GEAKDLAND] = 391 then "BW" else if [GEAKDLAND] = 442 then "PA" else if [GEAKDLAND] = 456 then "DO" else if [GEAKDLAND] = 462 then "MQ" else if [GEAKDLAND] = 464 then "JM" else if [GEAKDLAND] = 472 then "TT" else if [GEAKDLAND] = 474 then "AW" else if [GEAKDLAND] = 478 then "CW" else if [GEAKDLAND] = 488 then "CY" else if [GEAKDLAND] = 504 then "PE" else if [GEAKDLAND] = 512 then "CL" else if [GEAKDLAND] = 512 then "CL" else if [GEAKDLAND] = 524 then "UY" else if [GEAKDLAND] = 604 then "LB" else if [GEAKDLAND] = 624 then "IL" else if [GEAKDLAND] = 628 then "JO" else if [GEAKDLAND] = 632 then "SA" else if [GEAKDLAND] = 640 then "BH" else if [GEAKDLAND] = 647 then "AE" else if [GEAKDLAND] = 662 then "PK" else if [GEAKDLAND] = 669 then "LK" else if [GEAKDLAND] = 680 then "TH" else if [GEAKDLAND] = 706 then "SG" else if [GEAKDLAND] = 264 then "SI" else [GEAKDLAND]),

Adding in a column in order to determine (/decode) the country of departure:

Laadland = Table.AddColumn(Losland, "Laadland", each if [AFZKDLAND] = 1 then "FR" else if [AFZKDLAND] = 2 then "BE" else if [AFZKDLAND] = 3 then "NL" else if [AFZKDLAND] = 4 then "DE" else if [AFZKDLAND] = 5 then "IT" else if [AFZKDLAND] = 6 then "GB" else if [AFZKDLAND] = 8 then "DK" else if [AFZKDLAND] = 11 then "ES" else if [AFZKDLAND] = 19 then "LU" else if [AFZKDLAND] = 28 then "NO" else if [AFZKDLAND] = 32 then "FI" else if [AFZKDLAND] = 38 then "AT" else if [AFZKDLAND] = 53 then "EE" else if [AFZKDLAND] = 56 then "LT" else if [AFZKDLAND] = 60 then "PL" else if [AFZKDLAND] = 61 then "CZ" else if [AFZKDLAND] = 63 then "SK" else if [AFZKDLAND] = 68 then "BG" else if [AFZKDLAND] = 77 then "AM" else if [AFZKDLAND] = 36 then "CH" else if [AFZKDLAND] =288 then "PT" else if [AFZKDLAND] =400 then "US" else if [AFZKDLAND] =7 then "IE" else if [AFZKDLAND] =66 then "RO" else if [AFZKDLAND] =30 then "SE" else if [AFZKDLAND] =37 then "LI" else if [AFZKDLAND] =388 then "ZA" else if [AFZKDLAND] =528 then "AR" else if [AFZKDLAND] =92 then "HR" else if [AFZKDLAND] =10 then "PT" else if [AFZKDLAND] =64 then "HU" else if [AFZKDLAND] =509 then "BR" else if [AFZKDLAND] =720 then "CN" else if [AFZKDLAND] =91 then "SI" else if [AFZKDLAND] =54 then "LV" else if [AFZKDLAND] =480 then "CO" else if [AFZKDLAND] = 21 then "ES" else if [AFZKDLAND] = 46 then "LV" else if [AFZKDLAND] = 9 then "GR" else if [AFZKDLAND] =701 then "MY" else if [AFZKDLAND] =404 then "CA" else if [AFZKDLAND] =600 then "CY" else if [AFZKDLAND] = 732 then "JP" else if [AFZKDLAND] =728 then "KR" else if [AFZKDLAND] = 75 then "RU" else if [AFZKDLAND] =636 then "KW" else if [AFZKDLAND] =740 then "HK" else if [AFZKDLAND] =664 then "IN" else if [AFZKDLAND] =800 then "AU" else if [AFZKDLAND] = 700 then "ID" else if [AFZKDLAND] =44 then "DE" else if [AFZKDLAND] = 412 then "MX" else if [AFZKDLAND] =736 then "TW" else if [AFZKDLAND] =804 then "NZ" else if [AFZKDLAND] = 72 then "UA" else if [AFZKDLAND] = 0 then "NL" else if [AFZKDLAND] = 24 then "IS" else if [AFZKDLAND] = 41 then "FO" else if [AFZKDLAND] = 42 then "MC" else if [AFZKDLAND] = 55 then "TR" else if [AFZKDLAND] = 70

then "AL" else if [AFZKDLAND] = 73 then "BY" else if [AFZKDLAND] = 76 then "GE" else if [AFZKDLAND] = 93 then "BA" else if [AFZKDLAND] = 98 then "RS" else if [AFZKDLAND] = 204 then "MA" else if [AFZKDLAND] = 220 then "EG" else if [AFZKDLAND] = 272 then "CI" else if [AFZKDLAND] =276 then "GH" else if [AFZKDLAND] = 318 then "CD" else if [AFZKDLAND] = 324 then "RW" else if [AFZKDLAND] = 346 then "KE" else if [AFZKDLAND] = 372 then "EE" else if [AFZKDLAND] = 373 then "MU" else if [AFZKDLAND] = 382 then "CH" else if [AFZKDLAND] = 391 then "BW" else if [AFZKDLAND] = 442 then "PA" else if [AFZKDLAND] = 456 then "DO" else if [AFZKDLAND] = 462 then "MQ" else if [AFZKDLAND] = 464 then "JM" else if [AFZKDLAND] =472 then "TT" else if [AFZKDLAND] = 474 then "AW" else if [AFZKDLAND] = 478 then "CW" else if [AFZKDLAND] = 488 then "CY" else if [AFZKDLAND] = 504 then "PE" else if [AFZKDLAND] = 512 then "CL" else if [AFZKDLAND] = 512 then "CL" else if [AFZKDLAND] = 524 then "UY" else if [AFZKDLAND] = 604 then "LB" else if [AFZKDLAND] = 624 then "IL" else if [AFZKDLAND] = 628 then "JO" else if [AFZKDLAND] = 632 then "SA" else if [AFZKDLAND] = 640 then "BH" else if [AFZKDLAND] = 647 then "AE" else if [AFZKDLAND] = 662 then "PK" else if [AFZKDLAND] = 669 then "LK" else if [AFZKDLAND] = 680 then "TH" else if [AFZKDLAND] = 706 then "SG" else if [AFZKDLAND] = 264 then "SI" else [AFZKDLAND]),

Duplicating multiple columns in order to extract information:

#"Duplicated Column" = Table.DuplicateColumn(Laadland, "DTZEND", "DTZEND - Copy"), #"Duplicated Column1" = Table.DuplicateColumn(#"Duplicated Column", "DTZEND - Copy", "DTZEND - Copy - Copy"),

Extracting the month and year:

#"Extracted Month" = Table.TransformColumns(#"Duplicated Column1","DTZEND - Copy", Date.Month, Int64.Type), #"Extracted Year" = Table.TransformColumns(#"Extracted Month","DTZEND - Copy - Copy", Date.Year, Int64.Type),

Combining the month and year in order to create the time-period for the shipment:

#"Jaar + maand" = Table.CombineColumns(Table.TransformColumnTypes(#"Extracted Year", "DTZEND - Copy - Copy", type text, "DTZEND - Copy", type text, "nl-NL"),"DTZEND - Copy - Copy", "DTZEND - Copy",Combiner.CombineTextByDelimiter(" - ", QuoteStyle.None),"Jaar + maand")

Finishing the Query:

in
#"Jaar + maand"

C.2 HISTVRTL

Starting the Excel Query script:
let
Selecting the sources files in the folder and importing them:
Source = Folder.Files("K:\AUTOMATISERING\Lean Green\bestanden 2019 2020 \Querymap\HISTVRTL"), #"Filtered Hidden Files1" = Table.SelectRows(Source, each [Attributes][Hidden] <> true), #"Invoke Custom Function1" = Table.AddColumn(#"Filtered Hidden Files1", "Transform File (2)", each #"Transform File (2)"([Content])), #"Renamed Columns1" = Table.RenameColumns(#"Invoke Custom Function1", "Name", "Source.Name"), #"Removed Other Columns1" = Table.SelectColumns(#"Renamed Columns1", "Source.Name", "Transform File (2)"), #"Expanded Table Column1" = Table.ExpandTableColumn(#"Removed Other Columns1", "Transform File (2)", Table.ColumnNames(#"Transform File (2)"(#"Sample File (2)"))), #"Changed Type" = Table.TransformColumnTypes(#"Expanded Table Column1","Source.Name", type text, "LAADMTRS", type number, "VOLM3", type number, "HOOGTE", Int64.Type, "BREEDTE", Int64.Type, "LENGTE", Int64.Type, "GEWICHT", Int64.Type, "NTLSTUKS", Int64.Type, "NTLCOLLI", Int64.Type, "NRZEND", type text , "NRVERL", type text, "NRDEPVTR", Int64.Type, "ZNDSOORT", type text, "VRZENH", type text, "NRCOLLO", Int64.Type, "NREXTVRV", type text, "NREXTVRL", type text, "OMSVERP", type text, "GEPRINT", type logical, "AFHSRT", Int64.Type, "BERNEG", type logical, "MGEWICHT", type number, "MHOOGTE", Int64.Type, "MLENGTE", Int64.Type, "MBREEDTE", Int64.Type, "MLAADMTRS", type number, "MVOLM3", type number, "GEAPOST", type text, "OMRUILEN", type logical, "GEMETEN", type logical, "VRLEXTRA", Int64.Type, "VRLAFWIJK", type number, "DEPEXTRA", Int64.Type, "DVRLEXTRA", Int64.Type, "DVRLAFWIJK", Int64.Type, "SELECTED", type logical),

Removing not needed columns:
#"Removed Columns" = Table.RemoveColumns(#"Changed Type","Source.Name", "NRDEPVTR", "NREXTVRV", "NREXTVRL", "OMSVERP", "GEPRINT", "BERNEG", "OMRUILEN", "GEMETEN", "VRLEXTRA", "VRLAFWIJK", "DEPEXTRA", "DVRLEXTRA", "DVRLAFWIJK"),

Creating an unique number which is needed to order to merge the two files:
#"Added Custom" = Table.AddColumn(#"Removed Columns", "Unieknr", each [NRZEND] & [NRVERL]),

Calculating loading metres:
*#"Added Custom1" = Table.AddColumn(#"Added Custom", "Laadmeters", each if ([LAADMTRS] = 0 or [LAADMTRS] = null)then if ([VRZENH] = "COL" or [VRZENH] = "ICOL" or [VRZENH] ="BIKE" or [VRZENH] ="BDLS" or [VRZENH] ="CRTN" or [VRZENH] ="SPA" or [VRZENH] = "BAND" or [VRZENH] = "HSPL" or [VRZENH] = "DR" or [VRZENH] = "HANG") then [NTLCOLLI]*0.5/50 else if [VRZENH] = "EP" or [VRZENH] = "SP" then 0.4*[NTLCOLLI] else [LAADMTRS] else [LAADMTRS])*

Ending the Query:
in
#"Added Custom1"

C.3 Merge

Starting the Query Merge:
let

Selecting the other queries and merging the queries:

Source = Table.NestedJoin(HISTVRTL, "Unieknr", HISTVERT, "Unieknr", "HISTVERT", JoinKind.FullOuter), #"Expanded HISTVERT" = Table.ExpandTableColumn(Source, "HISTVERT", "ROUTE", "GEAKDLAND", "GEAPLAATS", "GEAPOST", "GEAHUISNR", "GEANAAM", "AFZKDLAND", "AFZPLAATS", "AFZPOST", "AFZHUISNR", "AFZNAAM", "NRVVA", "DTZEND", "Losland", "Laadland", "Jaar + maand", "ROUTE", "GEAKDLAND", "GEAPLAATS", "GEAPOST.1", "GEAHUISNR", "GEANAAM", "AFZKDLAND", "AFZPLAATS", "AFZPOST", "AFZHUISNR", "AFZNAAM", "NRVVA", "DTZEND", "Losland", "Laadland", "Jaar + maand"), postcode = Table.RenameColumns(#"Expanded HISTVERT", "AFZPOST", "Laadpostcode", "GEAPOST.1", "Lospostcode"), #"Reordered Columns" = Table.ReorderColumns(postcode,"LAADMTRS", "VOLM3", "HOOGTE", "BREEDTE", "LENGTE", "GEWICHT", "NTLSTUKS", "NTLCOLLI", "NRZEND", "NRVERL", "ZNDSOORT", "VRZENH", "NRCOLLO", "AFHSRT", "MGEWICHT", "MHOOGTE", "MLENGTE", "MBREEDTE", "MLAADMTRS", "MVOLM3", "GEAPOST", "SELECTED", "Unieknr", "ROUTE", "GEAKDLAND", "GEAPLAATS", "GEANAAM", "AFZKDLAND", "AFZPLAATS", "NRVVA", "AFZNAAM", "DTZEND", "Losland", "Lospostcode", "Laadland", "Laadpostcode", "Jaar + maand", "Laadmeters"), #"Changed Type" = Table.TransformColumnTypes(#"Reordered Columns","Unieknr", type text, "NRCOLLO", type text), #"Added Custom" = Table.AddColumn(#"Changed Type", "Opdrnr", each [Unieknr]" - "[NRCOLLO])

Ending the Query:
in
#"Added Custom"

Appendix D Code for Excel Query National 2022

Starting the Query:
let

Selecting the source file and importing it:
Bron = Excel.Workbook(File.Contents("K:\AUTOMATISERING\Lean & Green\Input 2022\Input Transmission 2022\Transpas-TransMission 2022.xlsx"), null, true), Blad1_Sheet = Bron[Item = "Blad1", Kind = "Sheet"][Data], #"Headersmetverhoogdniveau" = Table.PromoteHeader true]), #"Typegewijzigd" = Table.TransformColumnTypes(#"Headersmetverhoogdniveau","Dossierr

Changing the column data types of the columns "Dossiernr" and "Volgnr":
#"Type gewijzigd1" = Table.TransformColumnTypes(#"Type gewijzigd","Dossiernr", type text, "Volgnr", type text),

Adding a column which combines the "Dossiernr" and "Volgnr":
#"Dossiernr + volgnr" = Table.AddColumn(#"Type gewijzigd1", "Dossiernr + Volgnr", each [Dossiernr]&" - "&[Volgnr]),

Changing the order of the columns to get a better overview:
#"Volgorde van kolommen gewijzigd" = Table.ReorderColumns(#"Dossiernr + volgnr","Dossiernr", "Volgnr", "Dossiernr + Volgnr", "Status", "Regelfacturering", "Klaar voor factuur", "Opdrachtgever", "EDI referentie", "Laaddatum", "Losdatum", "Plangroep", "Dossierreferentie", "Laadnaam", "Laadadres", "Laadpostcode", "Laadplaats", "Laadland", "Losnaam", "Losadres1", "Lospostcode", "Losplaats", "Losland", "Laadmeter", "Zendingsoort", "CMR nr.", "Referentie", "Ordersoort", "ADR-zending", "Chauffeur lossen", "Ritnr lossen", "EDI provider", "Opvolgtype", "Afleverstatus opm.", "# Geweigerd", "# Beschadigd", "# Manco", "Afleverstatus", "Incoterm", "# afwijkingen", "Consolidatienr", "Contactpersoon", "Voorplancode", "Order naar Portal", "Factuurnr", "Losadres2", "Geen autotarief", "Week", "Departement", "Crediteur lossen", "Trailer lossen", "Truck lossen", "Crediteur laden", "Trailer laden", "Hoogte", "Breedte", "Afleverstatus tijd", "Lengte", "Origineel gewicht", "Gordelomvang", "Colli", "Palletplaatsen", "Volume", "Gewicht", "Eenheid", "Aantal", "Losadres", "Losreferentie", "Lostijd", "Laadreferentie", "Laadtijd", "2e chauffeur lossen", "Foutvracht", "Plangroep losrit", "# Goederenregels", "Chauffeur laadinfo", "Chauffeur losinfo", "Lostijd t/m", "Creditnota", "Legnr", "Geëxporteerd naar netwerk", "# documenten", "Debiteur", "Transportsoort", "Afstand", "Route nr.", "Colli Zending", "Origineel volume", "Meeneemcolli", "Aangemaakt op", "Crossdock laadadres", "Vertrek scans", "Aankomst scans"),

Removing the columns "Dossiernr" and "Volgnr" as they are combined already:
#"Kolommen verwijderd" = Table.RemoveColumns(#"Volgorde van kolommen gewijzigd","Dossiernr", "Volgnr"),

Filtering the orders to select only invoiced orders:
#"Rijen gefilterd" = Table.SelectRows(#"Kolommen verwijderd", each ([Status] = "Gefactureerd")),

Removing not needed columns:

#"Kolommen verwijderd1" = Table.RemoveColumns(#"Rijen gefilterd","Regelfacturering", "Klaar voor factuur", "Dossierreferentie", "Zendingsoort", "Referentie", "ADR-zending", "Chauffeur lossen", "Ritnr lossen", "Opvolgtype", "Afleverstatus opm.", "# Geweigerd", "# Beschadigd", "# Manco", "Incoterm", "# afwijkingen", "Consolidatienr", "Contactpersoon", "Order naar Portal", "CMR nr.", "Factuurnr", "Losadres2", "Geen autotarief", "Departement", "Crediteur lossen", "Trailer lossen", "Truck lossen", "Crediteur laden", "Trailer laden", "Afleverstatus tijd", "Losreferentie", "Lostijd", "Laadreferentie", "Laadtijd", "2e chauffeur lossen", "Foutvracht", "Chauffeur laadinfo", "Chauffeur losinfo", "Lostijd t/m", "Creditnota", "Legnr", "# documenten", "Debiteur", "Transportsoort", "Route nr.", "Vertrek scans", "Aankomst scans", "Aangemaakt op", "Losadres", "Crossdock laadadres", "# Goederenregels", "Palletplaatsen"),
#"Kolommen verwijderd2" = Table.RemoveColumns(#"Kolommen verwijderd1","Colli"),

Adding alternative calculation for the loading metres:

*#"Berekening laadmeter" = Table.AddColumn(#"Kolommen verwijderd2", "Laadmeter berekening pakket", each if ([Eenheid] = "Colli") then (0.5 /40 * [Aantal]) else if ([Eenheid] = "Pakket") then (0.5 /40 * [Aantal]) else if ([Eenheid] = "COL") then (0.5 /40 * [Aantal]) else if ([Eenheid] = "COL zending") then (0.5 /40 * [Aantal]) else if [Plangroep] = "Ritopdracht" then 13.6 else if ([Eenheid] = "ICOL") then (0.5 /40 * [Aantal]) else if ([Eenheid] = "ICOL zending") then (0.5 /40 * [Aantal]) else if ([Plangroep] = "Nachtpendel") then 13.6 else if ([Eenheid]= "BUN") then ([Lengte]/100*[Breedte]/100*0.4)*[Aantal] else if ([Eenheid] = "ENV") then (0.5 /40 * [Aantal]) else if ([Eenheid]= "HSPL") then ([Lengte]/100*[Breedte]/100*0.4)*[Aantal] else if ([Eenheid]= "BIKE") then ([Lengte]/100*[Breedte]/100*0.*
*else if ([Eenheid]= "EWP") then ([Aantal]*0.4) else if ([Eenheid]= "SP") then ([Aantal]*0.4) else if ([Eenheid]= "M3") then ([Aantal]/90*13.6) else if ([Eenheid]= "Uur") then 13.6 else 0),*

Replace empty (null) loading metre values with the value 0 in the column "Meeneemcolli":

#"Null Laadmeter" = Table.ReplaceValue(#"Berekening laadmeter",null,0,Replacer.ReplaceValue,"Meen

Adding an if statement to choose the new loading metre calculation method if the original loading metre value was null or less or equal to 0:

*#"Nieuwe waarde laadmeter" = Table.AddColumn(#"Null Laadmeter", "Nieuwe waarde laadmeter", each if [Laadmeter] = null then [Laadmeter berekening pakket]+(0.4 /40 * [Aantal]*[Meeneemcolli]) else if [Laadmeter] <= 0 then [Laadmeter berekening pakket]+(0.4 /40 * [Aantal]*[Meeneemcolli]) else if [Laadmeter] > 0 then [Laadmeter] else 0),*

Adding a column which gathers the month from the loading date:

#"Ingevoegde tekst tussen scheidingstekens" = Table.AddColumn(#"Nieuwe waarde laadmeter", "Tekst tussen scheidingstekens", each Text.BetweenDelimiters(Text.From([Laaddatum], "nl-NL"), "-", "-"), type text),

Changing the name of the column to "Maand" (Month): *#"Namen van kolommen gewijzigd" = Table.RenameColumns(#"Ingevoegde tekst tussen scheidingstekens","Tekst tussen scheidingstekens", "Maand"),*

Adding a column which represents the time-period (Year + Month):

#"Aangepaste kolom toegevoegd3" = Table.AddColumn(#"Namen van kolommen gewijzigd", "Jaar + Maand", each "2022 - "&[Maand]),

Replacing the countries that do not use the Alpha-2 country code standard:

#"Waarde vervangen" = Table.ReplaceValue(
#"Aangepaste kolom toegevoegd3","F","FR",Replacer.ReplaceValue,"Laadland"),
#"Replace EST EE" = Table.ReplaceValue(
#"Waarde vervangen","EST","EE",Replacer.ReplaceValue,"Laadland"),
#"Replace Est EE" = Table.ReplaceValue(
#"Replace EST EE","Est","EE",Replacer.ReplaceValue,"Laadland"),
#"Replace MLT MT" = Table.ReplaceValue(
#"Replace Est EE","MLT","MT",Replacer.ReplaceValue,"Laadland"),
#"Replace RSM SM" = Table.ReplaceValue(
#"Replace MLT MT","RSM","SM",Replacer.ReplaceValue,"Laadland"),
#"Replace F FR" = Table.ReplaceValue(
#"Replace RSM SM","F","FR",Replacer.ReplaceValue,"Losland"),
#"Replace IJ IS" = Table.ReplaceValue(
#"Replace F FR","IJ","IS",Replacer.ReplaceValue,"Laadland"),
#"Replace los EST EE" = Table.ReplaceValue(
#"Replace IJ IS","EST","EE",Replacer.ReplaceValue,"Losland"),
#"Replace los Est EE" = Table.ReplaceValue(
#"Replace los EST EE","Est","EE",Replacer.ReplaceValue,"Losland"),
#"Replace los MLT MT" = Table.ReplaceValue(
#"Replace los Est EE","MLT","MT",Replacer.ReplaceValue,"Losland"),
#"Replace los RSM SM" = Table.ReplaceValue(
#"Replace los MLT MT","RSM","SM",Replacer.ReplaceValue,"Losland"),
#"Replace los IJ IS" = Table.ReplaceValue(
#"Replace los RSM SM","IJ","IS",Replacer.ReplaceValue,"Losland"),

Changing the country of destination to the location of HST when the package is exported by GLS, Fedex, DPD or when the package is exported into the network. The packages are from this point transported by another company:

#"Nieuw losland toegevoegd" = Table.AddColumn(#"Replace los IJ IS", "Nieuw Losland", each if [Voorplancode] = "D - DPD" then "NL" else if [Voorplancode] = "F - Fedex" then "NL" else if [Voorplancode] = "G - GLS" then "NL" else if [Geëxporteerd naar netwerk] = "Ja" and [Plangroep losrit] = "TM Stukgoed" then "NL" else [Losland]),

Changing the place of destination to the location of HST when the package is exported by GLS, Fedex, DPD or when the package is exported into the network. The packages are from this point transported by another company:

#"Nieuw Losplaats toegevoegd" = Table.AddColumn(#"Nieuw losland toegevoegd", "Nieuw Losplaats", each if [Voorplancode] = "D - DPD" then "Enschede" else if [Voorplancode] = "G - GLS" then "Enschede" else if [Voorplancode] = "F - Fedex" then "Enschede" else if [Geëxporteerd naar netwerk] = "Ja" and [Plangroep losrit] = "TM Stukgoed" then "Enschede" else [Losplaats]),

Changing the postal code of destination to the location of HST when the package is exported by GLS, Fedex, DPD or when the package is exported into the network. The packages are from this point transported by another company:

#"Nieuw Lospostcode toegevoegd" = Table.AddColumn(#"Nieuw Losplaats toegevoegd", "Nieuw Lospostcode", each if Text.Length([Lospostcode]) = 4 and ([Losland] = "DE" or [Losland] = "ES" or [Losland] = "FI") then "0"[Lospostcode] else if [Voorplancode] = "D - DPD" then "7547 RW" else if [Voorplancode] = "F - Fedex" then "7547 RW" else if [Voorplancode] = "G - GLS" then "7547 RW" else if [Geëxporteerd naar netwerk] = "Ja" and [Plangroep losrit] = "TM Stukgoed" then "7547 RW" else [Lospostcode]),

Changing the address of destination to the location of HST when the package is exported by GLS, Fedex, DPD or when the package is exported into the network. The packages are from this point transported by another company:

#"Nieuw Losadres 1 toegevoegd" = Table.AddColumn(#"Nieuw Lospostcode toegevoegd", "Nieuw Losadres1", each if [Voorplancode] = "D - DPD" then "TRANSPORTCENTRUM 2" else if [Voorplancode] = "F - Fedex" then "TRANSPORTCENTRUM 2" else if [Voorplancode] = "G - GLS" then "TRANSPORTCENTRUM 2" else if [Geëxporteerd naar netwerk] = "Ja" and [Plangroep losrit] = "TM Stukgoed" then "TRANSPORTCENTRUM 2" else [Losadres1]),

Fixing the loading postal code when a 0 is missing at the beginning of the postal code for the countries Germany, Estonia and Finland:

#"Nieuw Laadpostcode" = Table.AddColumn(#"Nieuw Losadres 1 toegevoegd", "Nieuw Laadpostcode", each if Text.Length([Laadpostcode]) = 4 and ([Laadland] = "DE" or [Laadland] = "ES" or [Laadland] = "FI") then "0"&[Laadpostcode] else [Laadpostcode])

Ending the Query:

in
#"Nieuw Laadpostcode"