INSTITUTION: UNIVERSITY OF TWENTE
ACADEMIC YEAR: 2019/2020
MASTER THESIS REPORT

Using business intelligence to analyse sport associations' financial data

Author(s):
Arthur PALM - s1682660

Formal Chair:
prof. dr. Joost KOK

University supervisors:
dr. C.F. PINHO REBELO DE SA

dr. Ir. M. VAN KEULEN

Company:
STICHTING WAARBOGFONDS SPORT (SWS)

Company Supervisor:
M. BREMAN
S.R. DE WIT

February 9, 2021

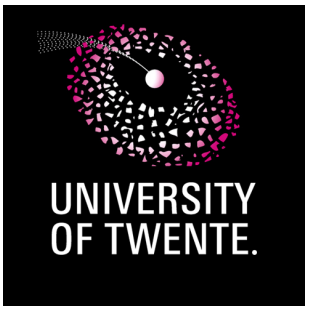


Table of Contents

1 Introduction

Informed organizations can help make better business decisions. Uninformed decision-making is more fitting when decisions do not pose serious risks, drawbacks or consequences. Here, intuition and experience are key predictors of success. Informed decision-making combines the intuitive facet of uninformed decision-making with information/data and logic (Jacknis, 2019; Ch, 2015), providing better opportunities of a successful outcome. Due to the ever-growing quantity of data in the systems and operations of the companies it is becoming more and more difficult for a (human) analyst to come up with interesting information (or patterns), just by looking at the data, and help in the decision-making process. Some companies are looking to integrate their data sources with intelligence systems, which could extract valuable information from the data-set and help them in their decision-making process. Some corporations and industry divisions are also moving their organizations towards a customer-centric perspective, gathering a huge amount data in the process. Particularly, financial data analysis is becoming more and more significant (Cai et al., 2016). In any case, processing this information storm just by looking at the raw data and using intuition and experience (the usual way), is an immense test for a human analyst, which also bound to get incorrect results and human mistakes due to the ever-growing amount of data the human analyst has to go through each year. This is a problem encountered by the Dutch organization called Stichting Waarborgfonds Sport.

1.1 Stichting Waarborgfonds Sport

Stichting Waarborgfonds Sport (SWS) is an independent guarantee institution dedicated to sport associations that is active throughout The Netherlands. SWS was founded in 1980 on the initiative of the Ministry of Health, Welfare and Sport, NOC * NSF and De Lotto, after it was found at the end of the 1970s that sports associations were increasingly struggling to obtain a bank loan for the construction, renovation or purchase of accommodation. SWS has two main activities (see official website [1]):

- ***Guarantees for sports organizations:*** SWS (usually with the cooperation of the local municipality) provide and manages guarantees for bank loans, since sport associations are often dependent on a loan from a bank for an investment in their accommodation.

[1] https://sws.nl/

- ***Financial feasibility advice:*** SWS advises the national and municipal government, but also banks, sports organizations or other stakeholders on, for example, the financial feasibility regarding the investment a sport association wants to do. This is done after the analysis of the financial situation and the financial trend of the sport association.

1.2 Domain knowledge: *Guarantee Institutions*

Domain knowledge is the area of work where the business or organization operates. This includes amongst others the knowledge and skills from the industry dynamics, the history, sectors and segments, the business model, competitive landscape, specific strategies and skills of the target enterprise[2]). The people working in that particular domain knowledge are viewed as the expert or specialists in that particular field. SWS is related to the domain knowledge of 'Guarantee institutions'. Most of the times, small and medium-sized enterprises/organizations (SMEs) want to start an investment projects, but do not have of sufficient capital to do the project. The organization tries to get a loan from the bank, but does not have enough guarantee, so the risk for the bank is too high. This means he cannot get a complete loan amount, or just small amount. Here is where a guarantee institution comes into place. The guarantee institution reviews the organization's financial position and decides to give the guarantee or not. A guarantee provided by a guarantee institution on behalf of the SME to the bank replaces this missing collateral and this makes it possible for the bank to grant the loan. Basically a financial liability by the guarantee institution to reimburse up to a specific percentage of the loan to the financial institution (banks) in the event that the SME company ought not have the option to fulfill his installments (Cui and Zhong, 2009)[3].

1.2.1 Business model and domain challenge

SWS can be considered the middle-man between sport associations and the financial institutions. They have the financial supervision over the sport association's financial data and transactions. This authority is given by a national authority, in this case the dutch government. The financial institution is usually a bank (i.e.: ING or ABNA AMRO bank), but it could also be any other company or organization giving a loan to the sport association. If a sport association wants to make a

[2]Domain knowledge: https://dnserp.wordpress.com/tag/domain-knowledge/

[3]Guarantee institutions: https://aecm.eu/guarantees/what-are-guarantee-institutions/

loan at the bank, the bank in question usually asks for a guarantee. SWS arranges all the financial details regarding the loan at the bank and provides the guarantee. Their guarantees in question are counter-guaranteed by the state or regions (the municipalities).

In addition to being a guarantee institution, SWS gathers all the financial data of the sport associations each year. This includes amongst others the club expenses, members contributions, promotion costs and fines. SWS analyse the data and a financial report is generated at the end of the financial year. Based on this financial statement, the financial stability of the sport association is determined and financial decisions and recommendations are given for the next financial year in case the financial position of the related sport association may lead to failure in the future. This guarantee scheme provides benefits not only to SWS and the financial institution giving a loan, but also to the sport associations. Some added values of this guarantee scheme are:

- For SWS: give additional support and expertise on the financial stability (including risk analysis) and access to funding to invest in economically sound projects
- For the financial institutions: reduced risk exposure when providing a loan
- For the sport associations: Cost efficient due to risk sharing

1.2.2 Stakeholders and their roles

The stakeholders involves in the guarantee business scheme of SWS are the following:

- *Stichting Waarborgfonds Sport:* SWS is the guarantee institution. SWS analyse the financial data of each sport association involved with the organization and gives recommendations based on the analysis of their financial data.
- *Sport associations in The Netherlands:* each sport association in The Netherlands associated to SWS has to provide their financial data and a financial-statement will be created at the end of the financial-year. Based on these financial data, SWS can analyse the current financial situation of the sport association and based upon this analysis SWS experts give recommendations to the sport associations related to their financial stability.

- *(sport) Community:* most of the financial decisions have an effect on the sport community. An example would be when the sport association cannot make a certain amount of loan at the bank, based on their financial history. The investment project related to the loan requested, may not be accomplished. This may (possibly) have a negative effect on the sport community, since these sport associations are non-profit organizations and are there mostly for the community.

- *The Municipalities:* each sport association is associated with the municipality of the region where it is located. A municipality may be associated with multiple sport associations at the time, depending on how many sport associations are located within its region. A municipality may also act as a guarantee source. Each municipality has one (or multiple) contact persons who are responsible for the sport associations they're associated with.

- *Government:* the work presented in this paper is part of a bigger concept which is initiated by the Dutch government. This work will possibly be linked to the main-project in the future.

- *Financial-institution:* banks are usually the financial institutions where sport associations make a loan. Normally, a bank asks for a guarantee before making a loan to a non-profit organization.

1.3 Problem statement

Over the years, more sport associations are becoming affiliated with the SWS and also more financial data is gathered each year from the sport associations. The financial data from each association is collected by SWS as flat-files (excel-files). Eventually, over the years, these increasing amount data as flat-files has become hard to analyse manually and this also makes it more difficult for SWS to accomplish one of its main tasks, which is to provide the best possible advice to the sport associations on their financial situation based on the manual analysis of the collected sport association's financial data. First, SWS would like to have a better view over the data. The amount of data SWS deals with increases each year and it will become hard to analyse this data manually and it will be hard to compare these data between the different sport associations and its different dimensions (for example per region, per city etc). We also want to move the business towards a more data-driven decision-making process which is has to do with

decision making based on the analysis of the data , instead of intuitiveness. Data-driven decision-making (DDDM) uses data to inform the decision-making process and ratifying the course of action before executing it (Foster Provost, 2013). This helps SWS experts to make decisions based on evidence-based, historical data and not only rely only on instinct. The success of this work depends on the validation and insights from the experts of SWS, whether the concept executed in this work helps SWS to provide better services to the sport associations and the stakeholders involved in the project.

1.3.1 Proposed solution

The proposed domains in data science related to this work is based on business intelligence techniques, in particular *data visualization* and *data mining* (Phongpandecha, 2018; Tripathy, 2019). We want to develop a method which could help SWS experts better understand the data and uncover unknown patterns in the financial data. Ultimately, we want this method to enable SWS experts to make business decisions based on evidence based, historical data by uncovering patterns and unknown knowledge in the data. We designed and realized an automated work process to replace the current manual work done by the SWS experts. This will include a database which will contain the data of each sport association (includes the financial data). We used data visualization and data mining techniques to analyse the data. In the future this method can be integrated in the work environment by having an interactive visualization dashboard on top, which will show the SWS experts the results, so that SWS only has to focus on making the business decisions .

1.4 Goal

The goal of this work is to create a method to show the extent of the possibility to extract useful knowledge from the financial data of the sport associations and to uncover actionable knowledge patterns in the data and transform this into understandable structure using *data visualization* and *data mining* techniques. This extracted knowledge from the data are the non-obvious indications about the financial stability of the sport associations. This knowledge will eventually help in providing data-driven decisions and help make better recommendation on the current financial position and also give suggestion on the future financial position of the sport associations if changes are necessary for the improvement of its financial stability. The application of this design project is in the domain work of guarantee

institutions. The effect of this project on the social context (stakeholders) are the following:

- *Stichting Waarborgfonds Sport:* obtain useful and previously unknown information in the financial data of the sport associations and with this information to be able to give better (evidence-based) advice / recommendations on the financial stability of the sport associations and provide better services to its customers (the sport associations) and related partners.

- *Sport associations in The Netherlands:* obtain useful information and recommendations on the current financial position of the sport association and based on recommendations given by SWS maintain a good financial stability. With a good financial stability of the sport association, it will be more likely that the bank will approve a loan and a more reasonable percentage of guarantee can be given to the sport association.

- *(sport) Community:* sport associations brings people together, foremost providing opportunities for social interaction. The sport associations involved are non-profit organizations. If loans and guarantees for projects related to the improvement of the association (for example renovations, new/more materials for the members) cannot be approved due to the unstable financial stability of the association, this may have a negative effect on the community. Sport is accepted as having crucial health, social and economic benefits on the community. Participation and interest in sport is very important for the community and the financial part to maintain these sport associations can be a hassle. A good management of the financial stability of the sport association and maintenance of the sport facilities can ensure that all people from all regions in The Netherlands are motivated and having the chance to participate in sport and enjoy its benefits. Sport and good physical activity in the community has the potential to help towards a healthier community (Donohoe, 2015).

- *The Municipalities:* will have a better view on the financial situation of the sport associations. The municipality may also act as a guarantee partner to the sport association, so it is also crucial for the municipality to be aware of the financial situation of its associated sport associations.

- *Government:* If this project obtains positive results, this can be used together with another (bigger) project initiated by the government (see section 1.2.2). With this work (and also the main-project by the government)

all the background (financial) details of the sport associations can be managed more effectively.

- *Banks:* SWS will be making decisions about the guarantee based on evidence-based data. For a bank it will be easier to approve a loan for a sport association based on more concrete and evidence-based data of the financial stability of the sport association. Banks have less worry giving loan to sport association with a positive, and also less worry of the consequences with the loan made at the bank.

1.5 Design problem & Knowledge questions

How can we get from the current situation to the goal? In a design problem we have a problem to design method/artifact. For this work, we wanted to design a method and a simple artifact to show how the method is implemented. This method helps towards the achievement of our goals. Design problems assume a context and stakeholder goals, which calls for an artifact, such that the interactions of (artifact x context) help stakeholders to achieve their goals (Wieringa, 2014). In this section the research questions and sub research questions are listed, also what observations are going to be made in this research project.

1.5.1 Research questions

From the research goal, the following main research question is formulated:

How can guarantee companies extract unusual knowledge and uncover unknown patterns in the financial historical data of the sport associations using data visualization and data-mining techniques?

In order to find an answer to the main research question, the question is decomposed into sub-questions. Therefore, the following sub-questions are formulated:

R.1 **What is an appropriate structure for the data to enable effective analysis with data visualization and data mining techniques?** Can this data set be imported in a database at its current state? Do we have to transform the data into a more structured format before analyzing the data. Given the data science principles, a recommended design of a data structure for analytical processing of data is the cube/star schema.

R.2 **Which data visualization tools/techniques are effective to use to show the data-results for SWS?** Will the visualizations of the results be generated from the prototype itself or is it be a better option to transfer the results to another system for data analysis and eventually visualize the results in a appropriate manner so that it is easy to understand for the end-users (SWS experts)?

R.3 **Which data mining techniques are effective for uncovering useful knowledge from the sport association's financial data for SWS?** Some data mining techniques works better with a certain type of data. In this research project, we are dealing with financial data (numeric data) of sport associations.

R.4 **How to design effective data visualizations for uncovering knowledge from sports associations data useful for SWS?** Depends on the tool or technique that is going to be used to analyse and visualize the data.

R.5 **To what extent is it beneficial to give SWS users freedom to navigate through the data and to perform ad hoc analysis and produce data visualization in contrast to pre-defined reporting?** Do we want to give the end-users the possibility to free to navigate the data at will or are should all the reports pre-defined? How much freedom can be given to the end users (SWS experts) to develop ad-hoc analysis?

2 Background

In this section the research that relates to this work is presented. The focus of this work is based on the following four subjects: *Data warehousing*, *Data Mining*, *Business Analytics* and *Business Intelligence*. Some use the terms in a replaceable sense, but there are strict differences amongst the four terms (Lee, 2013).

2.1 Difference between Data Warehousing, Data Mining, Business Intelligence and Business Analytics

The term *data mining* in general is used to describe a collection of different analysis techniques which includes amongst others, statistics, artificial intelligence and machine learning. In general, it has the purpose to uncover patterns and extract useful knowledge out of the data set. These sets of data may originate from one single (raw) source like a relational database or could originate from a *data warehouse*. Online Analytical Processing (OLAP) query is usually used in combination with data warehousing, where the data is stored into a multidimensional database cube where data can be analysed by viewing the data from the different dimensions. The use of these analysis techniques and decision rules to obtain critical business insights of the operational and performance characteristics of the business is the function of business analytics and the ability to generate these valuable insights based on business data is termed business intelligence (Lee, 2013, pg. 53).

2.2 Data Warehousing

A Data Warehouse is system which houses essential data into a single schema. Used particularly for reporting and data analysis, containing historical data gathered from multiple data sources. In Figure 1 you can see the data architecture. This consist of the following components[4,5]:

- The data sources: data from data warehouses originated from (multiple) sources, such a sflat files or operational databases
- Staging area / ETL process: Extraction Transform and Load. The data is copied from the various sources to a single destination system. During this

[4]https://databricks.com/glossary/unified-data-warehouse
[5]https://www.tutorialspoint.com/dwh/dwh$_{a}$*rchitecture.htm*

process calculations, concatenations and cleaning and data is transformed to appropriate format/structure to enable effective analysis on the data

- Storages (includes data marts): the data warehouse itself usesually consist of data marts. Data marts are specific data warehouse environments used for specific business purposes, for example reporting using data visualization or data mining purposes.
- The extracted models: this component consists of the query, analysis and data mining tools (also called the front-end tools)

Analysis of the data in the data warehouse is often used together with the OLAP methodology (Jarke et al., 2003; Ponniah and Reddy, 2001)[6,7]. The characteristics of the data in a data warehouse are built as following (Squire, 1995)[8]:

1. **Subject Oriented:** A data warehouse is subject-oriented. It offers information about one topic at a time i.e.: products, clients, providers, sales, revenue, etc. This is why the storage consist of data marts. In this project we focus on the subject *'Finance'*.
2. **Integrated:** A data warehouse is established by integrating data from various sources such as relational databases and flat files. Currently, all financial data from the sport associations are available only as yearly reports inside flat files (more specifically excel files).
3. **Time-Variant:** it gives information with respect to a certain period of time.
4. **Non-volatile:** Once the data is loaded into the data warehouse, it should not be updated or changed anymore. This is the historical data that is only loaded and accessed.

2.3 Data Mining techniques

Different data mining techniques can be used to extract useful information from financial data. Some common classes of tasks in data mining are *anomaly detection*, *association rule learning*, *clustering*, *classification*, *regression* and *summarization*. For this work we focused on the following two data mining techniques:

[6]https://www.guru99.com/data-warehouse-architecture.html
[7]https://www.educba.com/types-of-data-warehouse/
[8]https://www.softwaretestinghelp.com/data-warehousing-fundamentals/

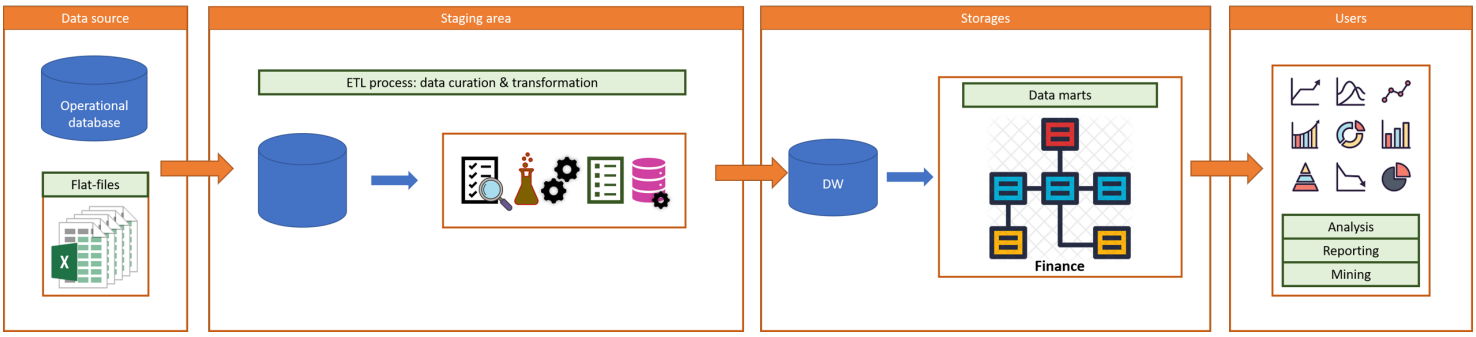


Figure 1: Data warehouse architecture

clustering and *subgroup-discovery*. Clustering is a data mining technique that is used and referred quite often (Berkhin, 2006). It is also often applied on financial data. When businesses gather more and more data from their day-to-day activities, they plan to derive valuable information from existing collected data to help make informed business decisions. Banking and financial institutions have applied various data mining techniques to boost the efficiency of their companies. Clustering is also considered to be an important tool for capturing the natural data structure (Le-Khac et al., 2012). Due to the significance of this technique and the amount of attention on using clustering data mining technique on financial data, we wanted to try this data mining technique on our financial data set This is why we considered this data mining technique in this work. Subgroup-discovery however is a less popular data mining technique and does not have a lot of research with this data mining technique on financial data. Although, subgroup discovery is applicable to different data sets out there because an important characteristic of this task, which is the combination of predictive and descriptive induction (Herrera et al., 2010). That is why we also wanted to try subgroup discovery on our data set, since we know a little about the data, and we want to uncover unknown patterns that might occur in the data over the years.

In data mining, the data can be minded by passing various process. A widely-used analytical model is the *Cross-industry standard process for data mining (CRISP-DM)*. As shown in Figure 2 (image by Kenneth Jensen (2012)) approach involves six phases: *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modeling*, *Evaluation* and *Deployment* [9]. The principles of the CRISP-DM

[9]https://towardsdatascience.com/crisp-dm-methodology-leader-in-data-mining-and-big-data-467efd3d3781

processwas used for the method and approach for this work. More information on the methodology for this work can be read at Chapter 3.

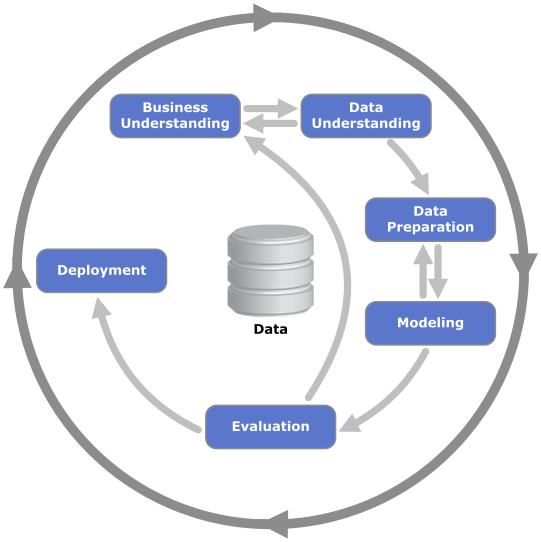


Figure 2: CRISP-DM Process Diagram (by Kenneth Jensen (2012))

2.3.1 Clustering

Clustering analysis involves grouping data points that are similar to each other within the same cluster. In a *cluster*, the degree of association between data objects is the high, and we have the lowest degree of association between the data clusters. A popular example of the usage of clustering analysis is for customer profiling. Clustering and classification are sometimes spoken interchangeably, but these have different context of data mining. Both these methods will characterize objects into groups, the primary difference between clustering and classification is the use of supervised and unsupervised learning techniques. Classification is used in supervised learning technique where *predefined* labels (target variables) are assigned to instances by properties, but clustering is used in unsupervised learning where the data is matched to the multiple clusters and it is eventually grouped into a cluster based on the difference in the similarity between them. At classification the training data is provided and the classed/clusters are defined in advance and the classification model based on this trained data set, but at clustering analysis there are no classes/clusters defined in advance and thus no training model. This exactly the case we have with the proposed research project. We have a lot of

financial data from multiple sport associations, we don't have a *target variable (classes/clusters)*, but by applying clustering analysis, we want to discover possible similarities between the financial data of the sport associations and finally uncover the classes/clusters.

There are different types of clustering techniques. Some of the most used clustering algorithms are the following (Seif, 2018; K.Kameshwaran and K.Malarvizhi, 2014; Rokach and Maimon, 2005; Xu and II, 2005):

- ***K-Means clustering:*** involves the partitioning of n data points into k amount of clusters where every data point belongs to the cluster with the closest mean. To identify the number of (k) clusters, you will have to look at the data and the center points are the mean of each cluster. By looking at the distance between the data point and the cluster's center point, the data point will be grouped into the cluster with the lowest distance between the data points. K-means algorithm is pretty fast, but because you will have to first select the number of clusters, it might not be ideally for this proposed project, since the point of the project is to gain insights from the data. Another variation of this algorithm is the K-median clustering algorithm.

- ***Mean-Shift clustering:*** involves locating the maximum of the most densed areas of data points (areas of higher point density). Data points here also represent the mean numbers. Doing this it will identify clusters in the data points. Here it is not required to know the number of clusters, since this will be uncovered by the algorithm automatically.

- ***Density Based Spatial clustering (DBSCAN):*** DBSCAN is a density-based clustered algorithm, also very equivalent to mean-shift algorithm, but not quite the same. DBSCAN checks the epsilon ε value and puts together data points that are closely fitted together. It does not expect you to define the number of clusters and it can also identify outliers as noises, unlike mean-shift, which groups these data points into a cluster, even though these points are very different. Besides, it can find inconsistently sized and inconsistently shaped clusters quite well. The disadvantage of this algorithm is that it might not perform as fast as other clusters when the clusters are of varying density. This data mining technique might not be useful for high-dimensional data, since it will become very hard to calculate the distance epsilon. Besides it is shown that density-based clustering does not suit financial data-set (Le-Khac et al., 2012).

- ***Expectation–Maximization (EM) clustering:*** EM clustering using Gaussian Mixture Models (GMMs) allows more flexibility than K-Means. K-Means uses of the mean value for the cluster center. Take a look at Figure 3. Here we observe that the mean value of the clusters are very close to each other. The mean value from the different clusters will be very difficult to distinguish from each other when using K-means. Some failure cases when using K-Means are clusters with differet sizes, clusters with different densities or non-spherical shape. GMMs can be applied to more variety of data points than K-Means. Instead of mean, the data points are Gaussian distributed, which is more sophisticated than taking a circular data set and just simply using the mean.

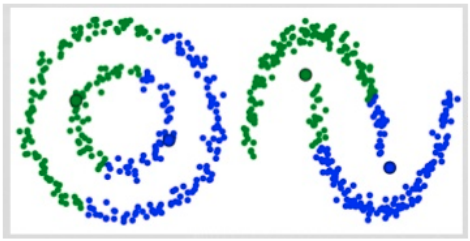


Figure 3: Two failure cases for K-Means (Seif, 2018)

- ***Hierarchical clustering:*** here you have two categories: top-down (divisive) or bottom-up (agglomerative). Top-down starts with each observation (data point) in its own cluster and then consecutively merges (or agglomerate) pairs of clusters as it is going down the hierarchical level. On the contrary, the bottom-up approach starts as one cluster and this will be split recursively as one moves down the hierarchy. A tree (or a dendogram) is used to show the hierarchy of clusters. In this algorithm the distance metric is used as the measurement of the data points and clusters are combined or split dependent on the distance between these data points. Here it is also not required to specify the number of clusters.

Ultimately, the decision of which algorithm is suitable to be applied on the data set of this project is dependent on how the data points look like. Cai et al. (2016) evaluated the different clustering algorithms with each other and discusses the advantages and disadvantages of each method. Clustering algorithms discussed in this paper are amongst other: association rules, classification, clustering, parti-

tioning methods and density-based algorithm. Some algorithms used in this research project eventually were not suitable for financial data set. This is definitely a factor to be considered for the proposed research project in this proposal.

Financial reports contain quantitative and (probably) also qualitative data. The financial reports of SWS do not contain sufficient textual data to perform text mining methods in order to see if the textual part of the reports contain useful information about the future financial performance of the sport associations. This is why the focus of this work was on the qualitative data (numbers) of the financial reports. In the paper by Kloptchenko et al. (2004), financial ratios were calculated and used for the analysis on the qualitative data. Financial ratios are the relationships which is determined from a company's financial information. The financial ratios can be used for to the following purposes:

- ***Tracking sport associations financial performance:*** By figuring out the financial ratios of a sport association (per time-period, ex. yearly) the change in the financial situation of the sport association can be tracked in the values over time and thus spot trends that may be developing at a specific sport association.

- ***Make comparative assessments between the sport associations:*** Comparing these financial rations between the different sport associations it can be identified whether a sport association is doing better or worst , financially, than other sport associations in the same category. By doing this it can also be determined if the sport associations financial assets are being used efficiently or not.

- ***Provide predictions based on previous trends in financial rations:*** If for example in a specific time-period a sport association's current financial trend is the same as a previously observed financial trend of a sport association in the same category and its financial assets were not being used efficiently and has closed down, it might be concluded that this sport association will end up the same way as the previous sport association with that same observed trend. Based on this observation, critical decisions can be made to avoid sport association close-down.

2.3.2 Subgroup Discovery

Subgroup discovery is used to uncover relations between the different properties or variables of the data set with respect to a target variable. A target variable

is the variable whose values are to be predicted or modelled by other variables available in the data set. Unknown patterns between these variables can be uncovered by this data mining technique and these patterns extracted are usually given as rules and called subgroups. Formally the extracted rule can be defined as **R:Cond** $\rightarrow Target_{value}$. For example: $R_1 : (Gender = Male\ and\ Age = Higher\ than\ 28\ AND\ Ownhouse = True) \rightarrow FamilyWithKids = True$ (Herrera et al., 2010). The main elements in a subgroup discovery algorithm are (Herrera et al., 2010; Atzmüller, 2006):

- ***Target variables:*** there are three distinct sub types of target variable which are binary (True or False), nominal (undetermined number of values) and numeric values.
- ***Description language:*** this is how the subgroups/rules are given (an example was already given above).
- ***Quality function/measures:*** these measures are used to provide an evidence of the importance and interest of the subgroups obtained for example like the coverage measurement, precision and confidence measurement.
- ***Search strategy:*** for the subgroup search strategy an efficient search is necessary. Brute-force search strategy is not used very often with subgroup discovery, although its use can be advantageous since it expands the current subgroup hypothesis further only if it outputs a better subgroup (Atzmüller, 2006).

In the paper of C.J.Carmona et al. (2012), the subgroup discovery technique has been applied to the data set obtained from the e-commerce website *OrOliveSur*[10]. SD was applied on the data of this website, with the focus on web usage mining, and with the ultimate goal to uncover unusual knowledge and behaviour of the different ways to access the website, which will allow the webmaster team to improve the design of the website, and to have more visitors and orders through the website. The paper uses the NMEEF-SD algorithm on the data-set.

In the article *'Using Subgroup Discovery to Analyse the UK Traffic Data'* by Kavšek and Lavrac (2004), a modification of the CN2 rule learner to subgroup discovery was used on data-sets acquired through the UK traffic data live feed. This shows us that subgroup discovery can also be applied on real-life- applications.

[10]Link to the website is www.OrOliveSur.com

Subgroup discovery is often used in context as association rule learning, but there is a difference between these two algorithms. The difference lies in the goals of each of the algorithms. In associative rule learning the set of classes are already defined and the goal is to generate models for each class. In contrast, subgroup discovery aims to uncover patterns in the data which will eventually lead to the discovery of rules for groups/classes in the data [11].

[11] https://docs.rapidminer.com/latest/studio/operators/modeling/predictive/rules/subgroup$_d$*iscovery.html*

3 Methodology

This section describes the method and proposed approach for this work. We chose to execute this work following the Cross Industry Standard Process (CRISP) model. The CRISP model is well known especially when using data mining techniques (Filipe et al., 2008; Chapman et al., 1999). Each phase of the CRISP approach was be applied to this research project, except *'deployment'*. The end product is a method and an artifact that will execute this method. The deployment phase is not done, since what we realized is only a proof-of-concept.

Take a look again at Figure 2. To work with the data set, one had to understand the business domain, the domain challenge and the data. One need to understand how the company works and the current problems they are facing. This part is related to the phases *business understanding* and *data understanding*. *Data preparation* involves the collection and transformation of the data into an appropriate structure for the data to enable effective analysis with data visualization and data mining techniques (*modelling phase*). The results will be assessed and justified at the evaluation phase.

Using the CRISP model, this work was executed by doing the following steps:

1. ***Business understanding and data understanding:*** the stakeholders and the research problem was defined and the goals for the stakeholders and the proposed solution. To better understand the company work domain and what we could have done with the data, some background on the subject was collected and literature papers have been reviewed. The approach for the research project was laid out and the methodology was also defined.

2. ***Data preparation:*** this step includes the collection of the data, data cleaning and data transformation. One has to design a structured format where the data will be collected and organized the analysis. If it is bad quality data, even the best algorithms could give poor quality predictions. These data from the sources are usually "dirty". For example they have errors, discrepancies between the different fields, inconsistency between the structure of the data from the different files / sources, or also inconsistency between the types of data or the data itself. When the data is "clean", we can select what part of the data is useful for data visualization and to make predictions [12].

3. ***Modelling - data visualization and data mining process:*** for the data visualization process we considered the software *Tableau*. Tableau Software

[12] https://cleverdata.io/en/clean-select-transform-data/

is an American interactive data visualization software (see Tableau official website[13]). The software is a pretty well known and used software in data science. Furthermore, SWS also has some previous experience with the software and it would be easier for the company to continue working on data visualizations using this software. Since we decided to use subgroup discovery as the primary data mining technique for this work, we choose the software Cortana. Cortana is a Data Mining tool for discovering local patterns in data (see Cortana Official website [14]). Cortana uses a generic Subgroup Discovery algorithm to implement different forms of local pattern discovery in the data.

4. ***Evaluation:*** at this step we evaluated the results. Here we justify and validate the method and artifact created based on the research questions.

[13]https://www.tableau.com/

[14]http://datamining.liacs.nl/cortana.html

4 Detailed design

In this chapter you will see how the plans and indicated steps in Chapter 3 were executed for this work. In this chapter you will also see the results of the method created and the artifact. At last the results are discussed and justified. We reflect on the strengths, difficulties, limitations of the obtained results for data visualizations and data mining, and also what the observations in the uncovered patterns mean for the company.

4.1 Data preparation: data structure

As mentioned in the beginning of Chapter 3, one has to collect and transform the data and put this into an appropriate structure for an effective analysis of the data.

4.1.1 Current state

currently, the data sources consist of flat files. Specifically, these are Excel files, each with 5 different sheets. We only had access to the Excel files for the sport type *voetbal (football / soccer)*. Each file is related to only one financial-year from one specific sport association. For the most part, every excel file has the same sheets and the data is also in the same structure. We are particularly interested in these following three sheets:

- ***Stamgegevens / Basic data:*** this sheet contains the basic data of the sport associations, like the name and project-id of the sport associations associated to SWS, its location information and amount of members (see Figure 4). This sheet also contain other information like the bank associated to the sport association, bank loan amount and the guarantee amount from the municipality, but these data are already available in another sheet. We decided to retrieve this data from the other sheet (rapport), since this data is already available over there including some other topics related to the same subject.

- ***Cijfers consol / consolidated numbers:*** an illustration of this sheet can be seen in Figure 5. This sheet contains every income and expense record (for example member bonds, sponsoring amount and the amount of money spent on advertisements. This sheet has three columns, the record id, the recordname and the amount in euros.

- ***Rapport / Report financial-year:*** this sheet contains of two sections (on two different pages). This sheet can be seen in Appendix A.1 and Appendix A.2. The first section contains some basic data about the financial report (for example the financial-year and details of the SWS expert who conducted the analysis and created the report, information about the board members of the sport association (for example name and email-address), data about the loan made from the financial institution (for example the name of the institution, date loan executed, amount of loan and rent percentage), information about the guarantee provided (for example the original amount, the release amount per year and duration in years) and information about the partners (financial institution and the municipality) and their contact information. The second part of this sheet contains the balance sheet and the conclusion about the financial position at the end of the financial-year and the recommendations from SWS to the sport association for the following financial-year. The balance sheet is an important part of a financial statement and is key to both financial modeling and accounting [15]. It shows the total assets (resources owned by the sport association) and how these assets are financed through either debt or equity.

Projectnummer SWS	#### ###
Naam vereniging	```````
Sport	voetbal
Adres	```````````````````
Postcode	```````
Stad	``````
Gemeente	`````````
Provincie	Zuid-Holland
aantal leden	######
Investering (type)	```````
Investeringsom	`######
Eigen geld	`######
Zelfwerkzaamheid	`######

Figure 4: Excel sheet - Stamgegevens

[15] https://corporatefinanceinstitute.com/resources/knowledge/accounting/balance-sheet/

5334	Donateurs	######.###
35	Sponsoring	######.###
4324	Reclameborden	######.###
323	Advertenties	######.###
564	Entreegelden	######.###
353643	Overige sponsoring	######.###
8210	Loterijen	######.###
446	Papier	######.###
8290	Overige acties	######.###
565	Subsidies	######.###
7674	Rente inkomsten	######.###
581	Boetes	######.###
3313	Schenkingen	######.###
908	Overige opbrengsten	######.###
546	Kantineverkopen	######.###
463	Overige kantine inkomsten	######.###

Figure 5: Incomes and expenses records

4.1.2 Database schemas

The structure of the data described in the previous section and the current state of the data cannot be used for analysis with data visualization and data mining techniques. For analytical purposes, we decided to create two databases instead of a data warehouse. We created two operational databases. MySQL relational database was chosen, because this way all the data can be linked to each other and this way it can be analyzed more easily. Specific data structure has been designed for each of the operational databases and we developed scripts (for data visualization, this can be done in Tableau) to transform the incoming data from the operational database to the targeted database-based data structure. We don't have specific things we are looking for in the data, creating an operational database and deriving the data from this database and creating scripts for specific data visualization gives us more flexibility to do what we want with the data. In Figure 6 and Figure 7 you can see the compact versions of the database schemas. The complete database schemas with all the details are available in Appendix A.3 and Appendix A.4. Additionally, in Appendix A.5, Appendix A.6 and Appendix A.7 you can see the *'Entity-Relationship Diagram'*, which shows the relationships be-

tween the tables displayed in the database schema related to the SWS projects in a little bit more detail. For the analysis in this paper, we only used the operational database in Figure 6.

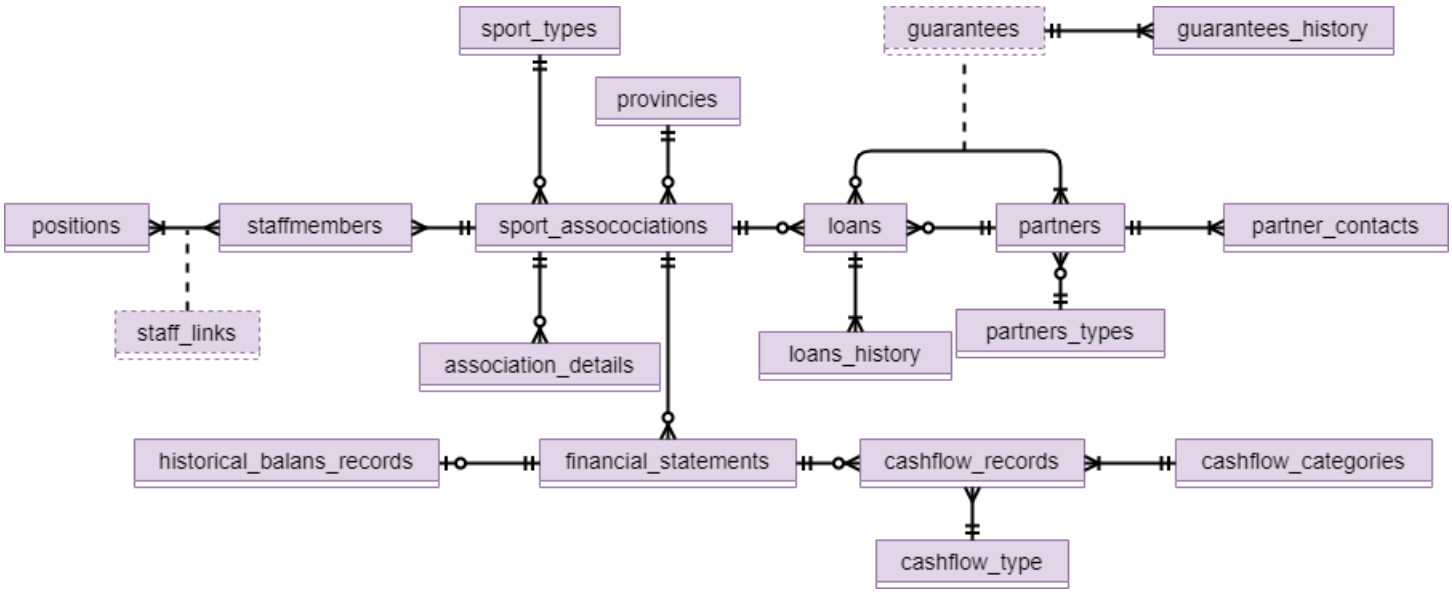


Figure 6: Operational database: SWS projects and related data

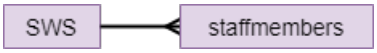


Figure 7: Operational database - SWS company data

4.2 Data preparation: data transformation

To collect and import the data from the Excel files into the databases, a system was created (Python program and using MySQL). This system collects, cleans the data and puts the data into a MySQL database. The data cleansing phase included several tasks, among other tasks like discarding unusable records, correcting spelling mistakes, format dates, remove duplicate columns, treat missing data, fix incorrect and corrupt data. During some of these tasks we found anomalies at critical fields. For example like date loan executed not indicated, so it was impossible for some loans to know when this has been conducted. Inconsistency between the

files from the same sport association, for example some data about the loan which was conducted x years ago differs from the data of the same loan which is still active and illustrated in the file for the financial-year y. We also encountered loan amount and guarantee amount values that did not add up, unexpected switch of the values between two different fields, multiple records for the income/expense with the same ID, missing income/expense record. We also encountered a record which sometimes was indicated as an expense and other times as an income. This last anomaly is intentionally made by SWS, but this gave us difficulties to import these type of records with the same ID into a relational database. Some of these anomalies were noted only after creating the data visualization, so the data transformation process had to be repeated.

Attribute	Cardinality	Type	Enabled
projectnumber_sws	13	nominal	no
name	13	nominal	no
financial_year	3	numeric	no
province	4	nominal	yes
balance_incomes_expenses	35	numeric	yes
income_avg	35	numeric	yes
expense_avg	35	numeric	yes
exp_salary_trainers	35	numeric	yes
exp_social_security_charges	9	numeric	yes
exp_travel_expenses_commuting	4	numeric	yes
exp_other_travel_accommodation_costs	9	numeric	yes
exp_other_personnel_costs	9	numeric	yes
exp_training_costs	7	numeric	yes
exp_volunteer_allowance	18	numeric	yes
exp_other_costs_volunteer	9	numeric	yes
exp_gas_water_electricity	33	numeric	yes
exp_taxes	30	numeric	yes
exp_insurances	28	numeric	yes
exp_security	16	numeric	yes
exp_cleaning_costs	26	numeric	yes
exp_maintenance	33	numeric	yes
exp_rent_loans	31	numeric	yes
exp_other_housing_costs	20	numeric	yes
exp_office_necessities	12	numeric	yes
exp_postage_costs	17	numeric	yes
exp_telecom	21	numeric	yes
exp_automation_internet	19	numeric	yes
exp_copy_costs	2	numeric	yes
exp_subscriptions	19	numeric	yes
exp_administration	8	numeric	yes

Figure 8: Tabular data structure - Cortana SD

As mentioned in the beginning of Chapter 2.3, we choose to use subgroup discovery to try uncover patterns in the data-set. The data mining too we used to apply this algorithm is called *Cortana: SubgroupDiscovery*. You can apply this algorithm on the data-set using different targets, but we studied the two targets *interest rate* and *guarantee*. To be able to use the data-set in Cortana, first the specific part of the data-set we want to apply the algorithm has to be put into a tabular structure and this way it can be imported into Cortana (for example as excel files).

The difference between the two types of analysis is their primary targets, *interest rate* and *guarantee*. Each record contains data about each sport association for a given financial-year. The columns for the subgroup discovery data-set consists of all the cashflow-records fields (which includes the incomes and the expenses records), *province name* and the following calculated fields: *average interest rate*, *end-of-the-year balance of incomes vs expenses*, and the *incomes and expenses average values*. We can see the tabular form of the data-set for subgroup discovery with the target as *interest rate* in Figure 8.

4.3 Modeling: data visualization

Several parts from the data-set were analyzed using Tableau. The most interesting results are presented and discussed here, it includes the sport associations basic data information, such as the consolidated income and expense balance (end-of-the-year results value), overview of incomes and expenses (separately), loans, interest rate and the debt evolution. For privacy reasons, the names of the sport associations, financial institutions and in some cases the provinces are indicated as numbers or blurred out [16].

4.3.1 Sport associations - Overview of the data

Figure 9 shows the count of sport associations involved with SWS for each province and the average count of members. For each province we can see the amount of sport associations working with SWS and the amount of members per province. This plot (and all future plots from Tableau shown in this dissertation) can be adjusted for more specific view of the data. We can filter the data being shown by the sport type (by default it is *voetbal*), and we can also choose to view the data for two or more financial-years together. We can also select the currently active projects related to SWS and highlight the data from a specific province. It is also possible to include more (basic) data about the sport association in the plot, for example investment-sum . In Figure 9, we can see that Utrecht has a higher average number of members per sport association, so this means bigger sport associations for example in relation to the province *Noord-Holland*, which has a lower average amount of members. This plot can be very useful to compare some data (especially over the years) between the sport associations, which would be

[16]Note: the data-set used for this work is only a partial data-set. Not all data of all the sport associations involved with the SWS was provided. Also, there were a lot of income/expense records with value zero.

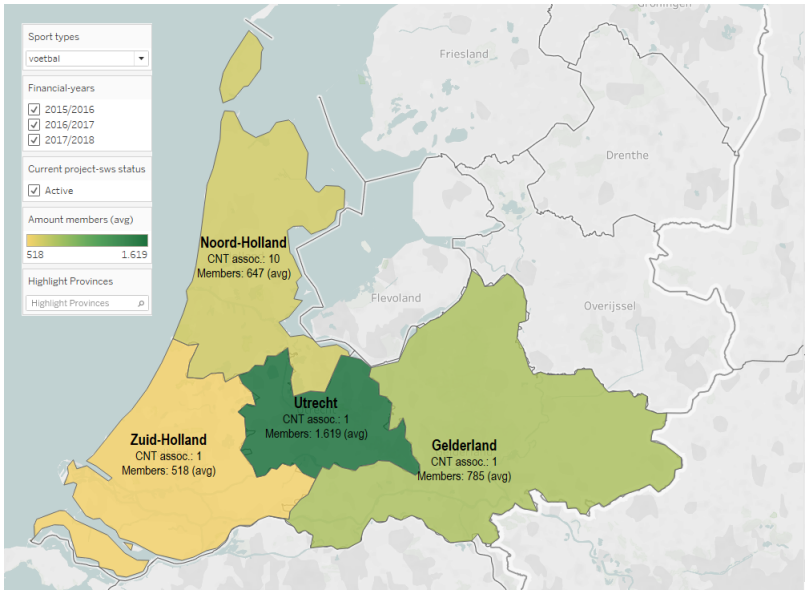


Figure 9: Sport associations - members and count associations

more and more difficult to see and distinguished from each other as the amount of data and sport associations involved with SWS keeps increasing. Also, this could not only be accomplished by province, but it is also be possible to do this by city, municipality or specific areas.

4.3.2 Consolidated income/expense balance (end-of-financial-year result)

The consolidated values refers to the sum of the values for *'foundation (stichting)'* and *'association (vereniging)'*. The visualizations related to this section we can see in the Figure 10, Figure 11, and Figure 12, each from a different perspective of the data-set: Figure 10 from the viewpoint of each financial years, Figure 11 distinguishing the data between each province over the years and in Figure 12 showing the results of each sport associations during it is financial-years. The data in this visualization shows the income/expense balance result value from a specific point of view. In Figure 13 we can see the parameters and legend. In all of these visualizations, the plots can also be adjusted, for example filtering the data only about specific sport types or only the results of each sport association for a particular year(s).

In Figure 10 we can see that the first year, the average income/expense balance from all the sport associations together was less than -3K. After the financial year

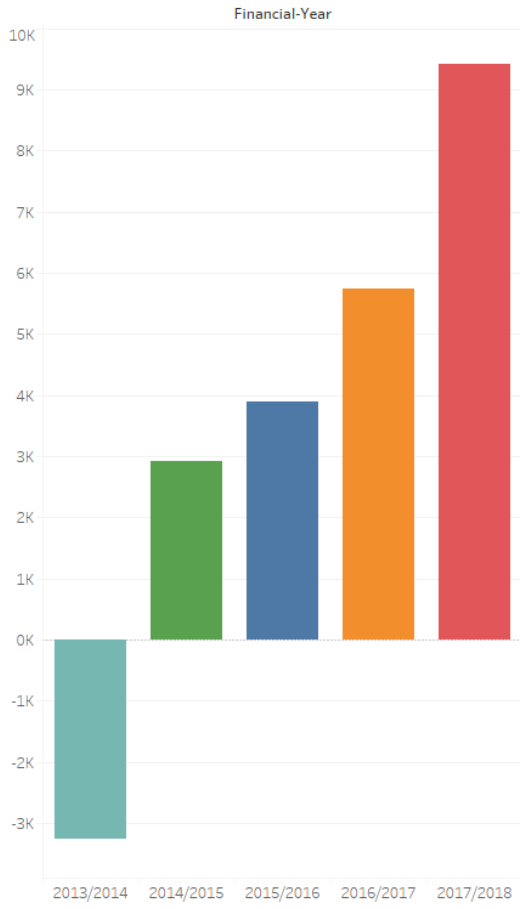


Figure 10: Consolidated income/expense balance overview: Financial-years

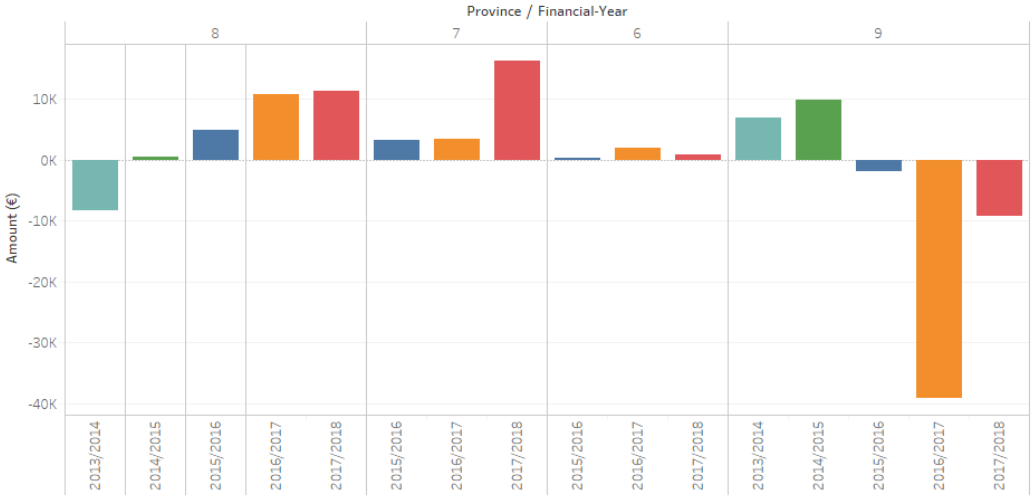


Figure 11: Consolidated income/expense balance overview: provinces (partial)

of 2014, this number was positive and kept growing a lot during the last few years.

Figure 11 compares the data between each province over the financial years. Most of the provinces have a positive average number for the income/expense balance at the end of the financial year. Although, for province number 9 this is not the case. For the first two years, this was a positive value, but the last three years it became a negative value and at the end of the financial year 2017, it almost reached the value of -40K. This value improved a little bit at the following financial year (2018), which might indicate that the financial problem is being taken care of.

Figure 12 shows the income/expense end of the year value for each sport associations during its financial-years. For most sport associations we can see that the income/expense value at the end of the financial year is between 20K and -10K and the most values being positive numbers. Although, for the sport association number 6, we can see that it had huge profits during the recent years, compared to the other sport associations. Actually, this value was more than 100K in the financial year

4.3.3 Overview incomes/expenses - per province

The results in this section are related to the incomes and expenses records. The visualizations have filters and parameters set which can customize the plots for a specific type of data (for example only for incomes data or only for expense data

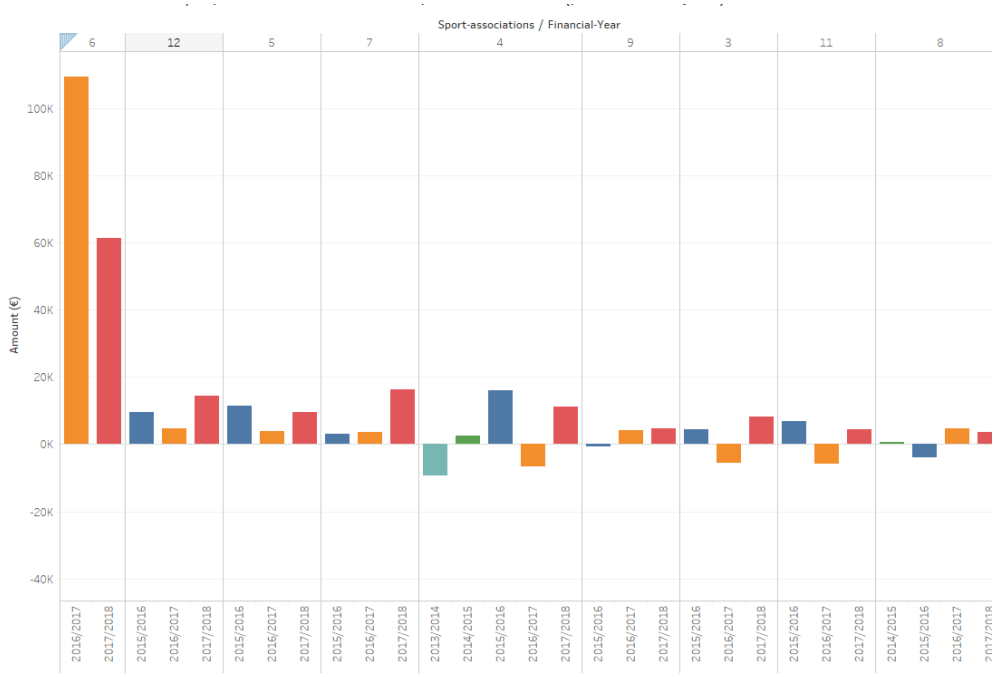


Figure 12: Consolidated income/expense balance overview: Sport associations

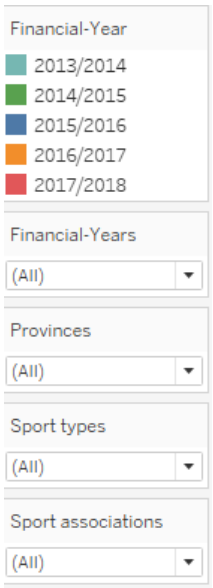


Figure 13: Filters, parameters, legend

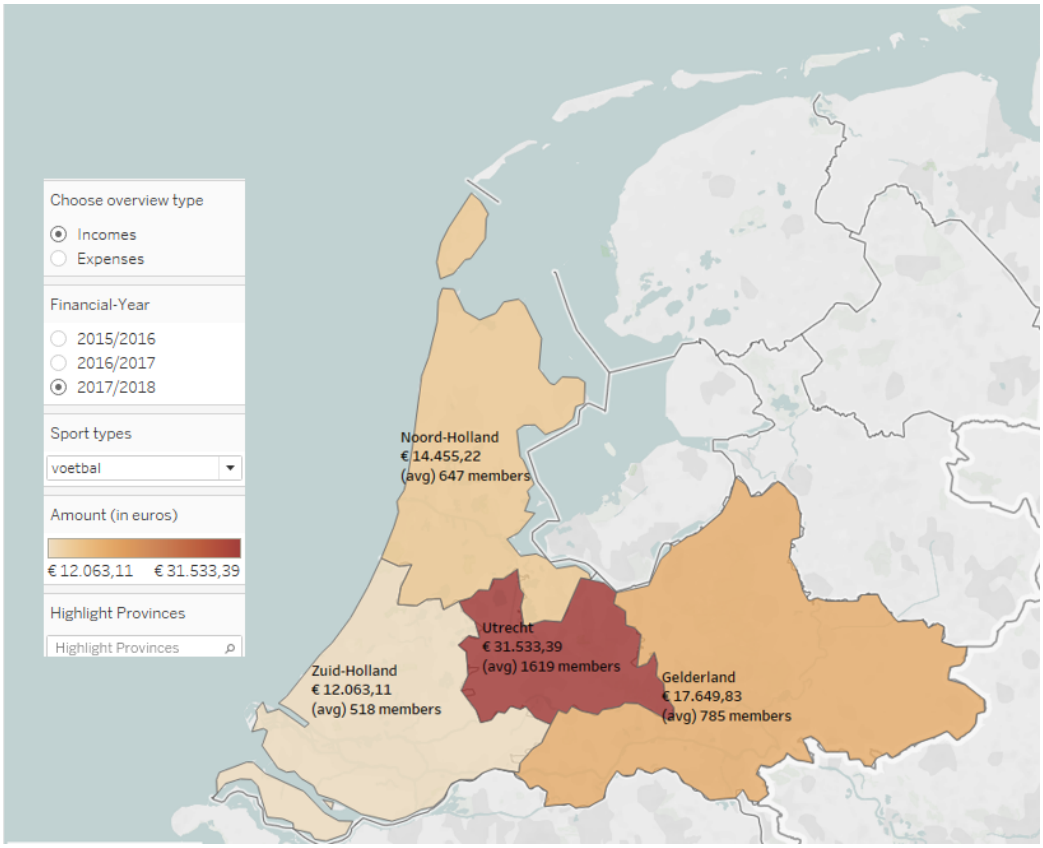


Figure 14: Incomes/Expenses (AVG) overview provinces

records), thus this means that not all the different possible visualizations can be shown in this report, but this section will cover the objective in general.

In Figure 14 we can see the average income distribution per province which have a sport association involved with SWS. We observe that the higher the average number of members, the higher the average income value per province. The color distribution shows the average amount per income record (in euros). In the figure, we can also observe that the visualization is by default set to show the average values for the financial-year 2017-2018, but this can be changed to another financial-year. In this visualization we can also set the parameter to see the average values for the expenses records, instead of the incomes. This visualization allows us to check if a province has a low or high average of incomes and expenses.

Figure 15 shows average values of income. As before, we can also change the

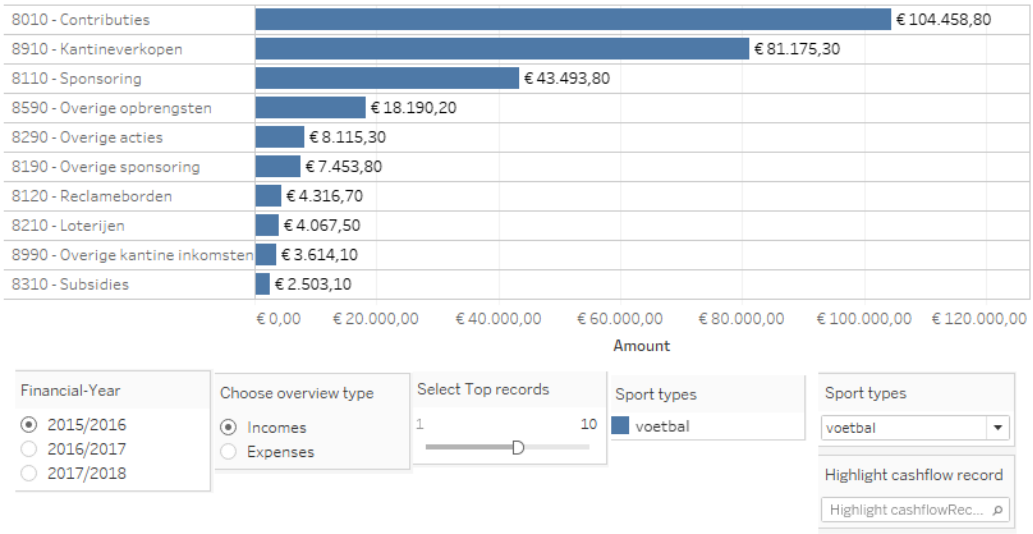


Figure 15: Incomes/Expenses (AVG) overview records

plot to view the top expenses records instead, or the top records from another year. Additionally, we can set the parameter to show only the top records related to a specific sport type (in the visualization only the records related to the sport type 'voetbal' is given), or multiple sport types at the same time. By default, only the top 10 records are shown. From this figure, we can observe that the top 3 records with the highest average income are *contributions*, *canteen-sales* and *sponsors*. The top 3 records with the highest average expense values were *salary trainers*, *purchase costs canteen* and *bond costs*. These values correspond to the financial-year 2015-2016. The top 3 records for both the income and expense do not change in the other two financial-years, but the average values can vary.

Figure 16 and Figure 17 shows the top 4 income/expense values. In Figure 16 it shows the top 5 expenses per province. We notice that for most of the provinces, the top 1 record is 'purchase costs canteen (inkoop kantine)'. For the province number 7, the top 1 expense record was 'salary trainer (salaris trainers)'. In Figure 17 we can see the top 4 expenses per sport association.

In Figure 16 we saw that 'purchase costs canteen (inkoop kantine)' was the top 1 record at almost all of the provinces, but in Figure 17 we see that this is not the same per sport association. For the sport association number 1, we cannot see the record 'purchase canteen (inkoop kantine)' in its top 5 expense records . On the contrary, for the sport association number *1*, the top expense record is

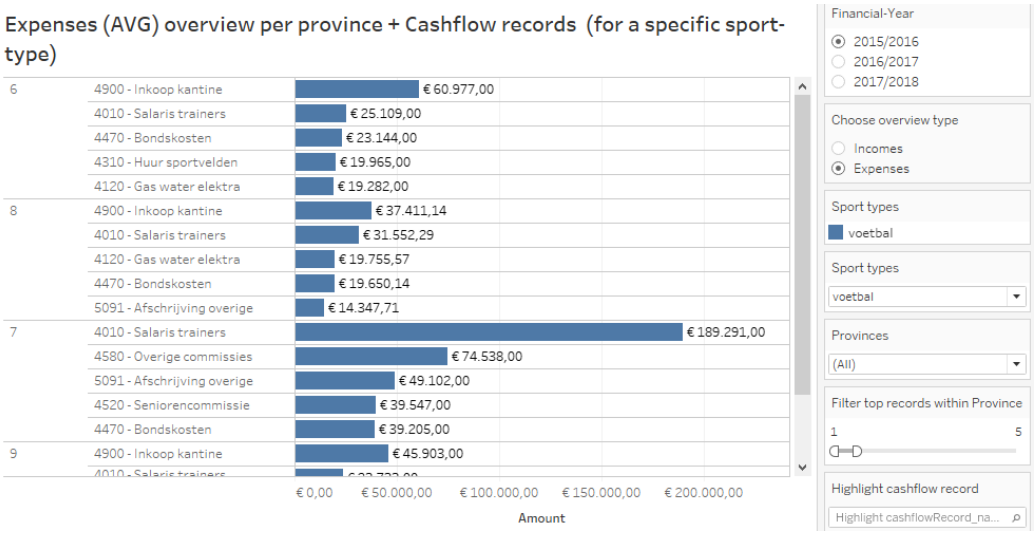


Figure 16: Incomes/Expenses (AVG) overview: Top records / province

actually *'Gas, water and electricity'*. Also most of the sport associations have *purchase costs canteen* as the top expense record and actuall y only two of the sport associations have *salary trainers* as the top expense record and the value of these records are significantly higher compared to the other sport associations.

4.3.4 Loans overview

In this section we analyze the loans of the sport associations involved with SWS. In Figure 18 we can see the loan values per bank/partner. Usually a sport association makes a loan at a bank, but there is one occurrence in the data-set where a sport association made a loan from another organization. In the figure we can see that most of the sport associations are making loans from partner number 55. The first digit in the horizontal-axis indicates a partner number and the second digit indicates the amount of loans. Partner number 55 has a total of 17 loans during the financial-years 2016 until 2018.

In this visualization it is also possible to highlight or filter out and display only with the values related to specified partners by changing the setting of the filter/parameters. There is also an option to show (currently) *active*, *non-active* or *undefined/unknown* loans. This option was made, because there were some loans in the data-set which did not have an issued date and not all information of this particular loan was available.

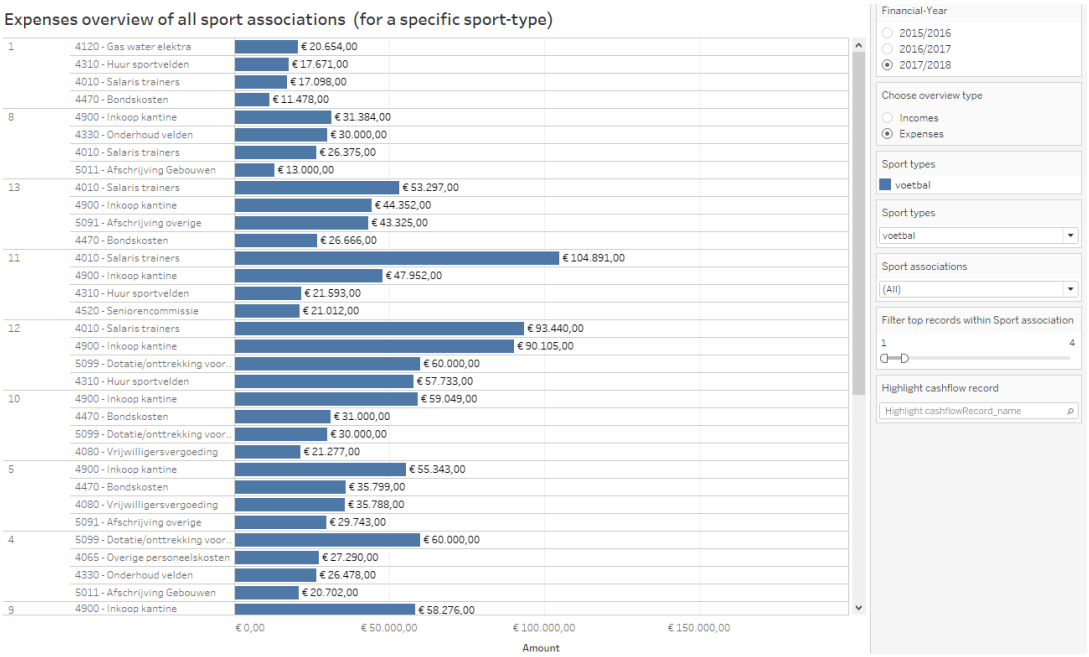


Figure 17: Incomes/Expenses (AVG) overview: Top records/Sport association

Figure 19 we can see the loan distribution per province. This visualization also shows the values for the total sum of loans for the province, the average amount of loan for each province and the count of the loans and the sport associations for that particular province. As expected, provinces with more projects involved and more loans count have a higher sum for the value of loans at that particular province. We can see that *Gelderland* has a higher value for the average loan amount compared to the other provinces. Figure 20 shows the loan overview per sport association. Besides the sum for the amount of loans per association, also the repayment amount and the partner information and loans count is shown in the visualization. The figure does not show anything unexpected, but if a sport association have for example a low count of loans and a high value for the sum of loan when comparing this to the results of the other sport associations, it can be easily seen in this visualization.

Next, in Figure 21 we can see the debt evolution for each sport association during the last financial-years. It is expected that the debt for each sport association will decrease each year (assuming no new loan has been made). We can see one example for the plot of a sport association indicated as a pink coloured line. At the financial-year 2015/2016, this sport association had a debt amount of

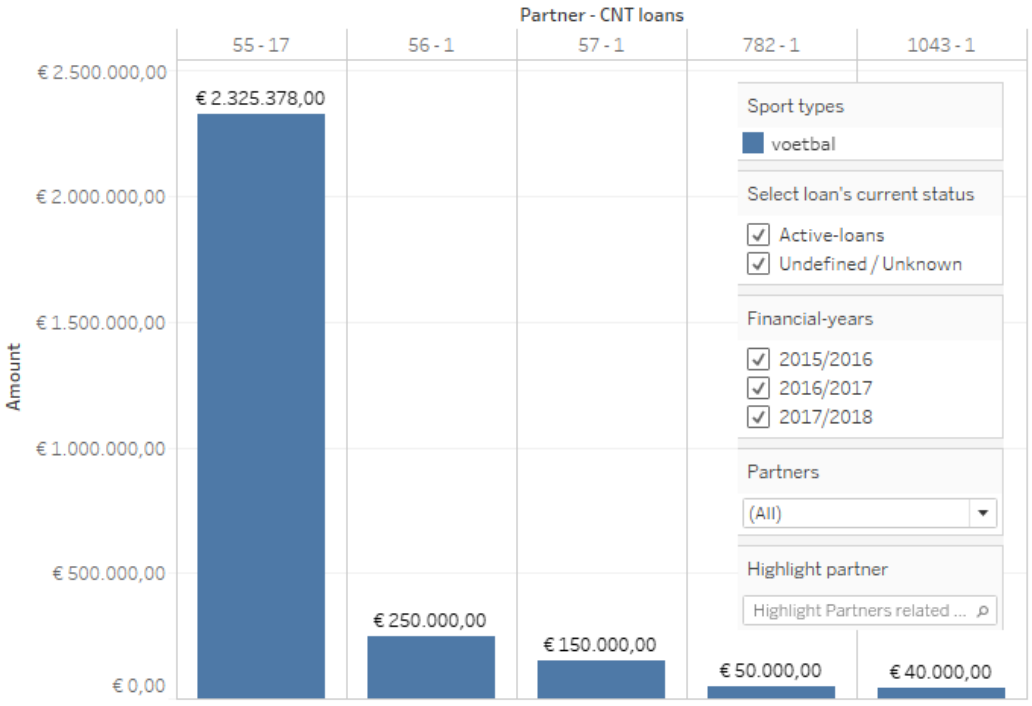


Figure 18: Loans overview (per bank/partner)

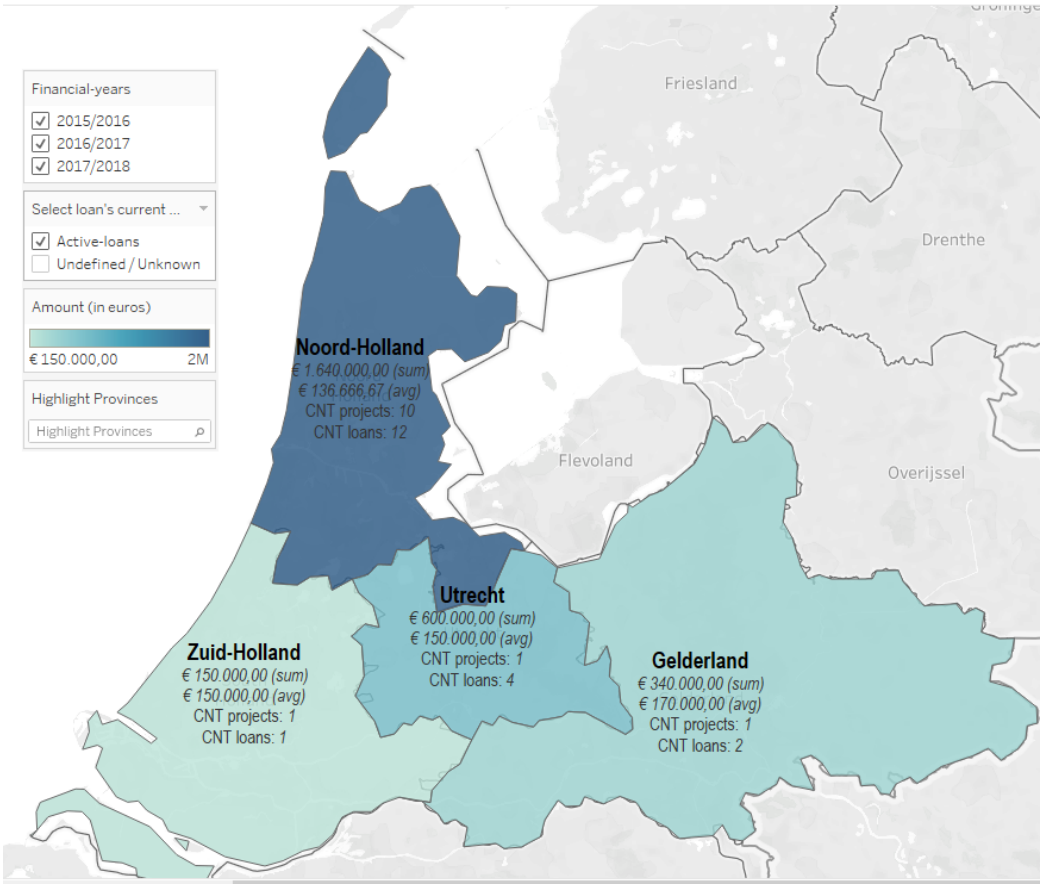


Figure 19: Loans overview per province

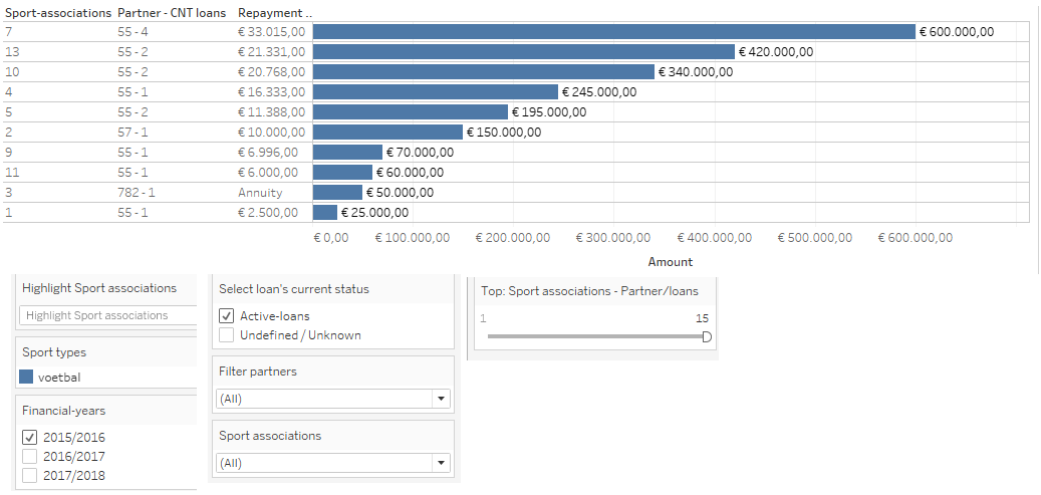


Figure 20: Loans overview: per sport association

€35.882,00. to be paid back to the bank. At the financial-year 2016/2017 this amount was less than the year before, namely €31.750,00, but at the financial-year 2017/2018, this amount was still €31.750,00 and you can also see this as a horizontal line. In other cases, a sport association was able to pay more money back at a particular financial year than the settled payment amount per year and this can produce a bigger drop in the line for that particular sport association. This can be seen at the line of the sport association indicated as the colour orange. For this sport association at the financial-year 2016-2017, the amount of debt was €184.043 and at the financial-year 2017/2018, this amount was €129.680,00. We can clearly see in the plot that this particular sport association had a bigger drop in the line, compared to the other sport associations. In this visualization is also possible to compare the debt evolution between the different sport associations. Although, when dealing with a lot of sport associations, it might become cluttered, so it will definitely be a good choice to limit the amount of sport association to display at the same time. Another visualization gives a closer look to the debt progression for each sport association. This is shown in Figure 22. In this visualization we could also see any inconsistencies related to the debt progression for each sport association and if all debts for a specific sport association are being paid accordingly.

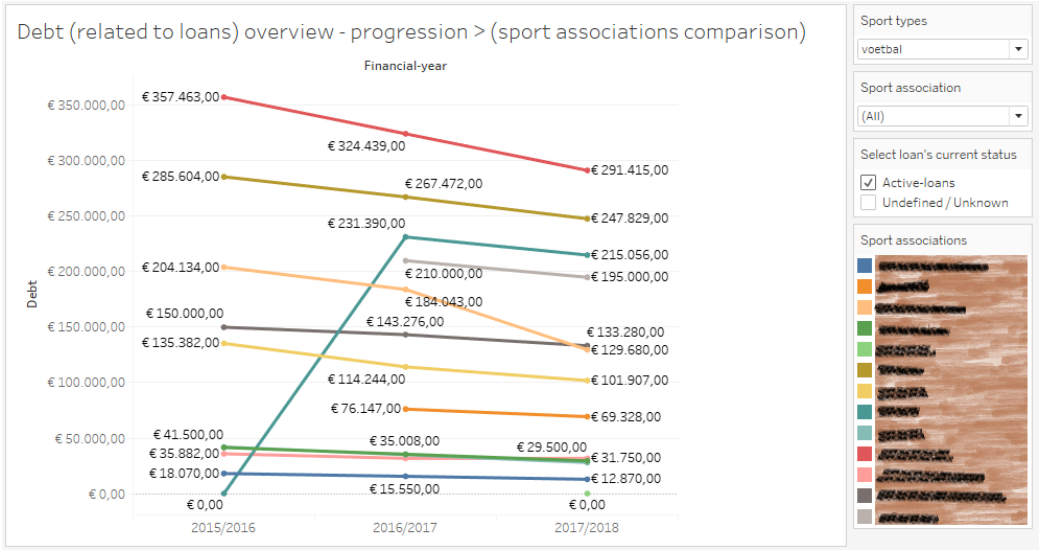


Figure 21: Debts overview

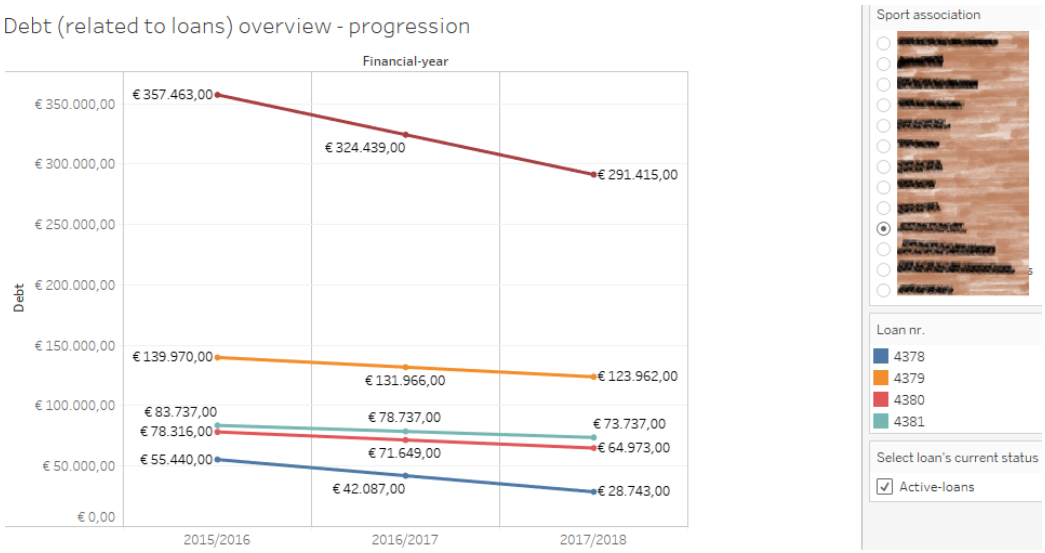


Figure 22: Debts overview (per sport association)

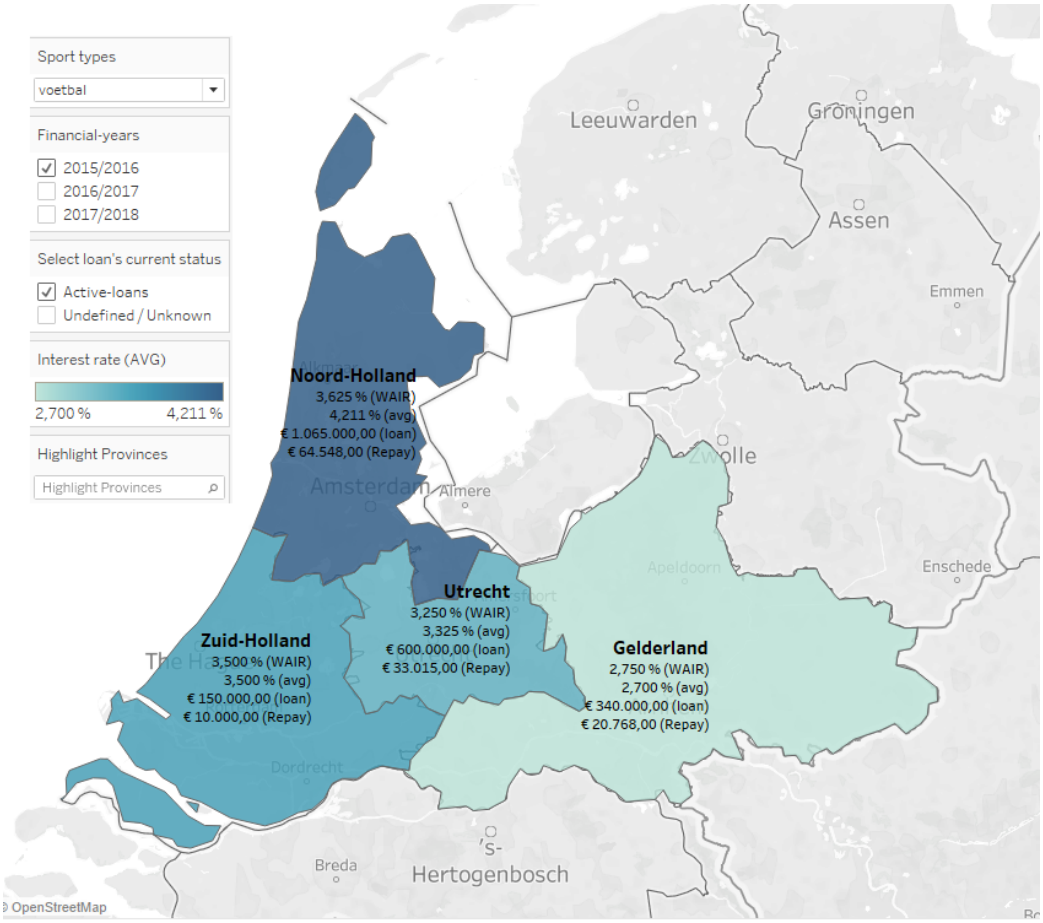


Figure 23: Interest rate overview (per province)

4.3.5 Interest rate distribution

In Figure 23 we can see the *interest rate overview* per province. Additionally the *weighted average interest rate* and the *average interest rate* per province is also shown in the visualization. In the visualization we can see that Noord-Holland has a higher average interest rate and a higher sum of loans than the other provinces.

4.4 Modeling: subgroup discovery

We used the data mining technique *Subgroup Discovery* to study more in detail the patterns in the data, which could be useful to establish the future of the financial position of a sport association. As explained in Section 4.2, the data mining tool called *Cortana: Subgroup Discovery* was used to apply subgroup discovery algorithms on the data-set. Different targets has been studied, but we focus on two: *interest rate* and *guarantee*. The following section will explain a little bit more of the settings used in Cortana for this research project.

For the subgroup discovery task in Cortana, we can select multiple parameters related to the analysis (see Figure 24). Only the parameters which were used or adjusted for this project are discussed in this section. The other constraints were left as default.

- ***Primary target and target type:*** both the interest rate and the guarantee amount are the primary target respectively. Furthermore, both these target types are single numeric
- ***Quality measure and measure minimum:*** the quality measure used for both these analyses is Z-score. Z-scores are used to measure an observation's deviation from the mean value. Z-scores reveal whether a score is typical for a specified data-set or if it is atypical (Lavrakas, 2008; Salkind, 2010). The measure minimum indicates the minimum acceptable quality measure value with respect to the analysis's results process.
- ***Refinement dept:*** the refinement dept indicates the number of columns (or attributes) to be taken into consideration for the subgroup discovery process
- ***Minimum coverage:*** the minimum coverage consist of the minimum amount of records for a subgroup for it to be considered a valid subgroup. For example, a minimum coverage of 10% on a data-set of 40 records would be a minimum of coverage value of 4. A minimum coverage of only one or two records could result into too little amount of records inside the subgroups.

- ***Strategy type:*** for the search strategy, a beam search strategy is used.

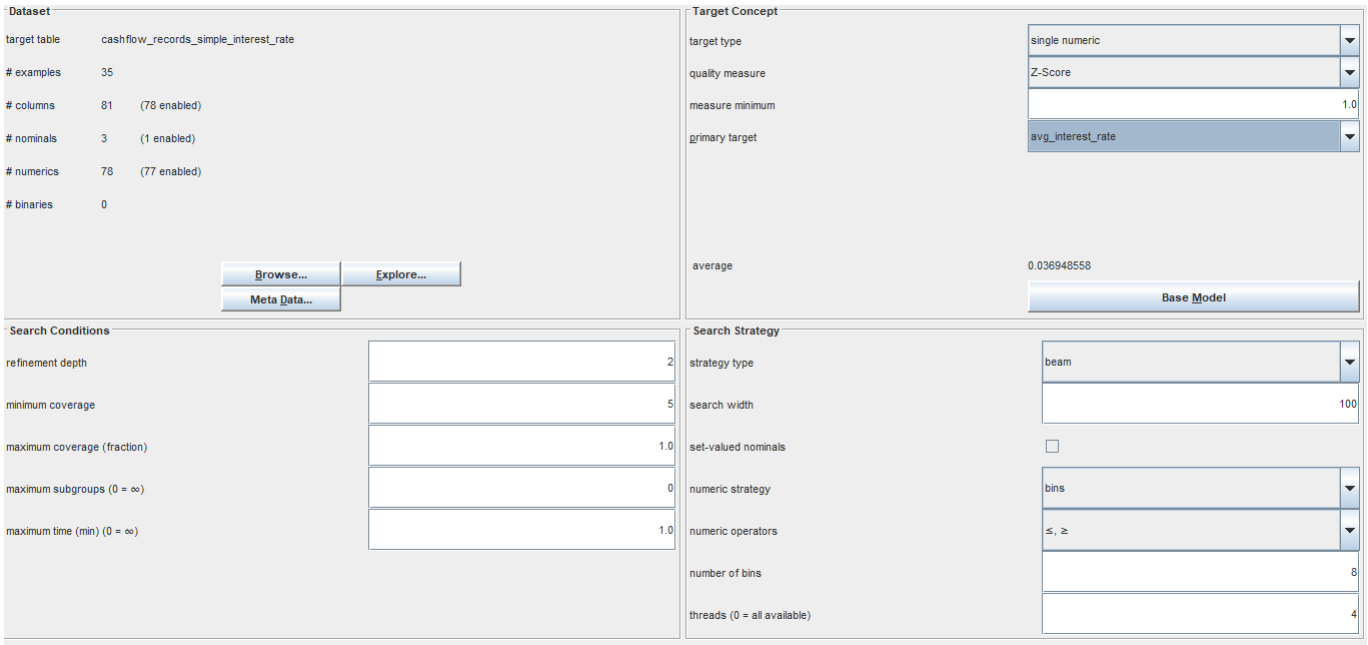


Figure 24: Constraints / parameters - Cortana SD

First the results of the data-set with the primary target as *'average interest rate'*. In Figure 25 we can see the base model which shows the density distribution of the average interest values. A higher density means a higher amount of that particular value of the target. The peaks in the plot shows illustrate the *average interest rate* values which are most common in the entire data-set. During the first search for subgroup discovery, there were 21417 results. After doing statistical validation, we were left with 21 results. These results are shown in Figure 26

In general, we can conclude that the lower the sport association's incomes and expenses, the higher interest rate they have received from the bank. Looking at the first subgroup, we can see the condition $exp\_management\_cost \leq 0$. The reason this condition is given as subgroup discovery is because a lot (but not all) records related to the expense management_costs are 0.00 during the financial-years. All rules containing these conditions (≤ 0) are ignored.

One particular subgroup which is interesting is number 2: $exp\_management\_cost \leq 2550.33$ *AND* $inc\_contributions \leq 52318.0$. The model for this particular subgroup is shown in Figure 28. The black colored line shows the density distribution over the primary

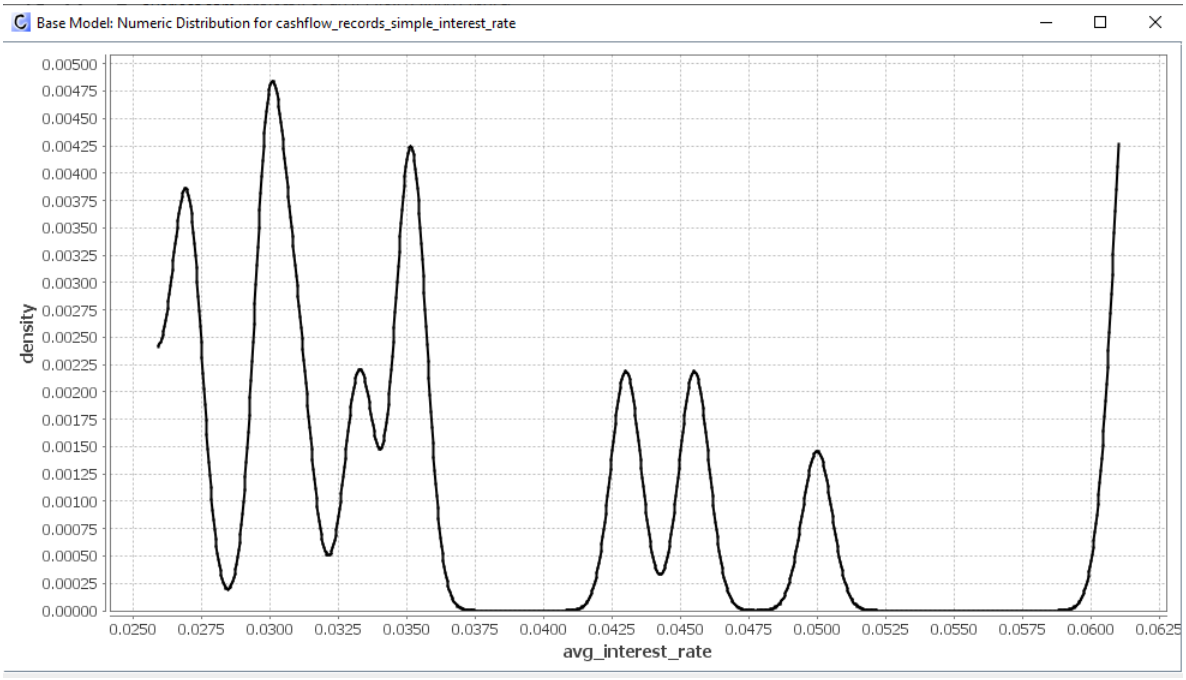


Figure 25: Base model: primary target 'Interest rate'

target (average interest rate). The red colored line shows the density distribution for the subgroup number 2. In the figure only two peaks can be observed. This means that subgroup number 2 is only observed at an average interest rate of 5% and above 6%. The model shows a higher peak at average interest above 6%. In other words. it is more likely the average interest rate for a particular sport association is 6% or higher if this subgroup/condition is met.

In Figure 27 we can see the table with the results of the analysis of the dataset using the *'average guarantee amount'* as the primary target. During the first search for subgroups, there was more than 20800 results, after statistical validation there were 135 results. In this case, when having the *avg guarantee* amount as the primary target we can see that the higher the incomes and expenses amount, the higher the average guarantee amount received for a sport association.

In Figure 29 we can see the model plot which is related to the result at number 2. This condition is interesting, because we can see that we have this subgroup when $inc\_sponsorship \geq 31376.0$ *AND* $exp\_taxes \geq 5300.0$. In the figure there are 3 peaks where this subgroup has a high density.

21 subgroups found; target table = cashflow_records_simple_interest_rate; quality measure = Z-Score

Nr.	Depth	Coverage	Quality	Average	St. Dev.	p-Value	Conditions
1	2	5	4.344888	0.0566	0.005389	-	expense_avg <= 2550.33 AND exp_management_costs <= 0.0
2	2	5	4.344888	0.0566	0.005389	-	expense_avg <= 2550.33 AND inc_contributions <= 52318.0
3	2	5	4.344888	0.0566	0.005389	-	expense_avg <= 2550.33 AND inc_other_actions <= 387.0
4	2	5	4.344888	0.0566	0.005389	-	expense_avg <= 2550.33 AND inc_other_income >= 5724.0
5	2	5	4.344888	0.0566	0.005389	-	exp_maintenance <= 3490.0 AND inc_contributions <= 52318.0
6	2	5	4.344888	0.0566	0.005389	-	exp_maintenance <= 3490.0 AND inc_canteen_sales <= 71583.0
7	2	5	4.344888	0.0566	0.005389	-	exp_telecom <= 594.0 AND inc_contributions <= 52318.0
8	2	5	4.344888	0.0566	0.005389	-	exp_subscriptions <= 0.0 AND inc_contributions <= 52318.0
9	2	5	4.344888	0.0566	0.005389	-	exp_rent_sports_fields <= 17671.0 AND exp_maintenance <= 1332.0
10	2	5	4.344888	0.0566	0.005389	-	exp_purchase_canteen <= 44352.0 AND exp_maintenance <= 1332.0
11	2	5	4.344888	0.0566	0.005389	-	exp_purchase_canteen <= 44352.0 AND inc_contributions <= 52318.0
12	2	5	4.344888	0.0566	0.005389	-	inc_contributions <= 56498.0 AND exp_maintenance <= 1332.0
13	2	5	4.344888	0.0566	0.005389	-	inc_contributions <= 56498.0 AND exp_management_costs <= 0.0
14	2	5	4.344888	0.0566	0.005389	-	inc_other_actions <= 387.0 AND income_avg <= 9985.39
15	2	5	4.344888	0.0566	0.005389	-	inc_other_actions <= 387.0 AND expense_avg <= 2550.33
16	2	5	4.344888	0.0566	0.005389	-	inc_other_actions <= 387.0 AND exp_volunteer_allowance <= 4370.0
17	2	5	4.344888	0.0566	0.005389	-	inc_other_actions <= 387.0 AND exp_maintenance <= 1332.0
18	2	5	4.344888	0.0566	0.005389	-	inc_other_actions <= 387.0 AND exp_other_housing_costs >= 1053.0
19	2	5	4.344888	0.0566	0.005389	-	inc_other_actions <= 387.0 AND exp_clothing <= 0.0
20	2	5	4.344888	0.0566	0.005389	-	inc_other_actions <= 387.0 AND exp_other_competition_costs <= 498.0
21	2	5	4.344888	0.0566	0.005389	-	inc_other_actions <= 387.0 AND exp_representation_costs <= 949.0

Figure 26: Average interest: subgroup discovery results

4.5 Discussion: data structure

The current state of the data structure was not appropriate for data analysis. For this reason, we created a new data structure and also collected and imported the data into a relational database. To process and analyze the data, the data-set had to be transformed into a relational database-schema format. This makes it more useful for business intelligence techniques. The analysis of the data using Tableau was a good choice, since Tableau can easily connect with the relational database and automatically update the visualizations when new data enters the database, so there was no need to manually transfer the data on separate files. SWS also has some basic knowledge and experience with the system. Several useful visualizations has been created on Tableau, including some geo-visualizations. These visualizations present the results into an understandable structure to the user.

Additionally, by importing the data set into a relational database, the different aspects of the data (such as the loans and guarantees) can be linked to each other. Without the relations between the different aspects of the data it would be hard to process and analyze the data digitally. Also, by having the data in a relational database, several tasks at the guarantee company can be automated. Currently, the financial data of each sport association is filled in manually into an Excel file at the end of the financial year. Eventually, this file will be analyzed, manually (the usual way) by SWS and some details (for example conclusions and recommendations about the financial position regarding the end-of-the-year balance) will be added to the file in a separate sheet and this will be the financial report. Instead, we can use the database and fill the financial data directly into the database. At the end of

134 subgroups found; target table = cashflow_records_simple_guarantee; quality measure = Z-Score

Nr.	Depth	Coverage	Quality	Average	St. Dev.	p-Value	Conditions
1	2	10	3.961604	145,250	19,667.5505...	-	exp_taxes >= 5505.0 AND inc_billboards_income <= 0.0
2	2	11	3.960484	142,272.727...	20,983.0723...	-	inc_sponsorship >= 31376.0 AND exp_taxes >= 5300.0
3	2	9	3.907665	147,777.777...	19,128.3834...	-	exp_taxes >= 5505.0 AND inc_sponsorship >= 44274.0
4	2	9	3.907665	147,777.777...	19,128.3834...	-	inc_sponsorship >= 43721.0 AND exp_taxes >= 5505.0
5	2	9	3.891252	147,500	19,472.2023...	-	exp_taxes >= 5505.0 AND exp_training_costs <= 0.0
6	2	9	3.891252	147,500	19,472.2023...	-	exp_other_canteen_costs <= 0.0 AND exp_telecom >= 590.0
7	2	9	3.891252	147,500	19,472.2023...	-	exp_other_canteen_costs <= 2191.0 AND exp_telecom >= 590.0
8	2	11	3.871407	140,909.090...	23,239.67817	-	exp_taxes >= 5505.0 AND exp_other_travel_accommodation_costs <= 5200.0
9	2	11	3.871407	140,909.090...	23,239.67817	-	exp_taxes >= 5505.0 AND exp_clothing <= 15139.0
10	2	11	3.871407	140,909.090...	23,239.67817	-	exp_taxes >= 5505.0 AND exp_youth_committee >= 2072.0
11	2	11	3.871407	140,909.090...	23,239.67817	-	exp_taxes >= 5505.0 AND inc_billboards_income <= 7225.0
12	2	11	3.871407	140,909.090...	23,239.67817	-	exp_youth_committee >= 2072.0 AND exp_taxes >= 5505.0
13	2	8	3.842792	150,625	18,403.0397...	-	exp_rent_loans >= 5122.0 AND exp_office_necessities <= 0.0
14	2	8	3.842792	150,625	18,403.0397...	-	exp_automation_internet >= 575.0 AND exp_medical_expenses >= 1330.0
15	2	8	3.842792	150,625	18,403.0397...	-	exp_medical_expenses >= 1288.0 AND exp_other_canteen_costs <= 0.0
16	2	8	3.842792	150,625	18,403.0397...	-	exp_medical_expenses >= 1288.0 AND inc_lotteries <= 5298.0
17	2	9	3.842013	146,666.666...	20,783.27259	-	inc_sponsorship >= 48409.0 AND exp_taxes >= 5300.0
18	2	9	3.842013	146,666.666...	20,783.27259	-	inc_sponsorship >= 48409.0 AND exp_automation_internet >= 641.0
19	2	9	3.8256	146,388.888...	21,085.5112...	-	exp_rent_loans >= 5122.0 AND exp_other_canteen_costs <= 0.0
20	2	9	3.8256	146,388.888...	21,085.5112...	-	exp_other_canteen_costs <= 0.0 AND exp_rent_loans >= 5400.0
21	2	9	3.8256	146,388.888...	21,085.5112...	-	exp_other_canteen_costs <= 2191.0 AND exp_rent_loans >= 5400.0
22	2	7	3.82258	155,000	15,352.9894...	-	income_avg >= 16695.11 AND exp_other_committees >= 262.0
23	2	7	3.82258	155,000	15,352.9894...	-	expense_avg >= 4249.72 AND exp_other_committees >= 262.0
24	2	7	3.82258	155,000	15,352.9894...	-	exp_taxes >= 4699.0 AND exp_other_committees >= 262.0
25	2	7	3.82258	155,000	15,352.9894...	-	exp_taxes >= 5505.0 AND exp_seniors_committee >= 4285.0
26	2	7	3.82258	155,000	15,352.9894...	-	exp_taxes >= 5505.0 AND exp_other_committees >= 262.0
27	2	7	3.82258	155,000	15,352.9894...	-	exp_taxes >= 5505.0 AND inc_sponsorship >= 51960.0

Figure 27: Average guarantee: subgroup discovery results

the financial year a financial report can automatically be generated by the guarantee company into a predefined structure. This will eliminate a lot of manual work. By having an automated system with an integrated business intelligence system and data mining techniques, it will be easier to see unknown information from the historical data. This can eventually help guarantee companies to make better evidence-based decisions about the financial situation of the sport associations.

4.6 Discussion: data visualization

The visualizations created in this work gives very well insights into the financial data from the sport associations. Using these visualization, you can also make it more easy to visualize and compare the data from the different sport associations. This would be a difficult task to do the usual way. For example in Figure 11, you would be able to see how the end-of-the-year balance progresses during the years for each of the provinces. In Figure 12 you can see the end-of-the-year balance progress during the active financial-years of each sport association. Also with the ability to adjust the visualization using the filters and parameters available you would be able to customize this specific visualization to see the different aspects of this visualized data. You can also see how the different income and

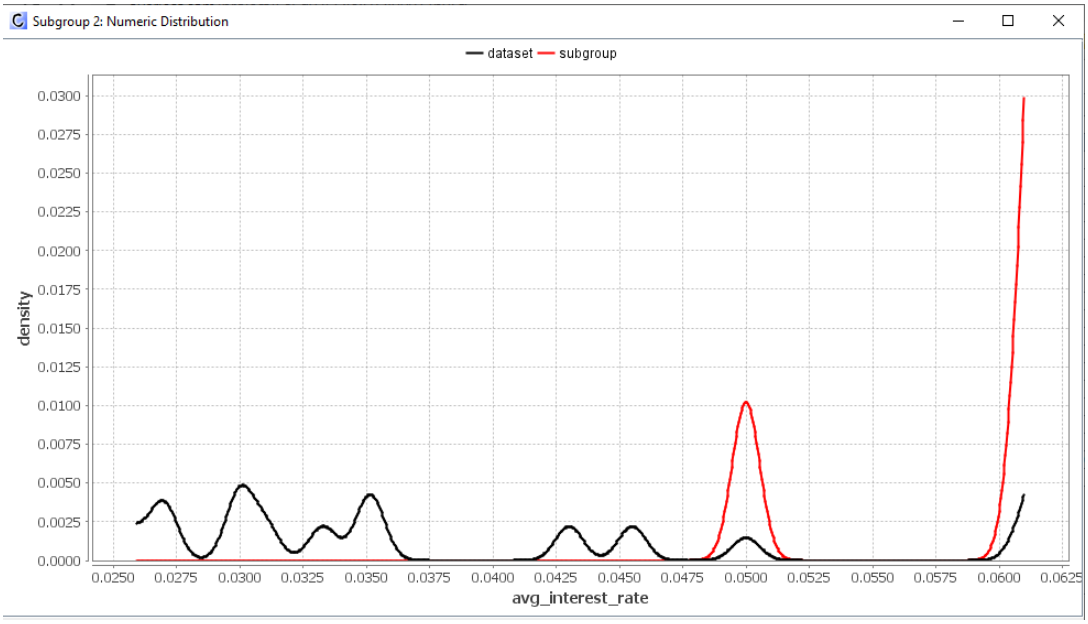


Figure 28: Model plot subgroup number 2 (primary target 'avg interest rate')

expense records are doing during the financial-years. There you can also adjust the visualization to view this specific data for a particular sport-type, viewing only the records for the incomes or expenses, or you can filter the top results (see Figure 15). This means that SWS would have the freedom to navigate through the data and peform ad-hoc analysis and produce visualization up to a certain extend of the data. It would also be possible if the company would like even more freedom to perform ad-hoc analysis and produce data visualization, although this may require some knowledge on making scripts to transform the data for specific types of visualizations

Using these visualization, it would also be easier for SWS to inspect the financial data of the sport associations, draw conclusions and make recommendations on their financial-stability and how each sport association is managing their incomes and expenses. With this visualization we can more clearly check the different top expenses and incomes records and average values and also compare with the different sport types. We can also compare the different financial-years and see if these values change to be better or maybe to the worst. The sport associations in this study are all non-profit organizations. So, it is not a good for the association for these kinds of associations to have big losses/debts at the end

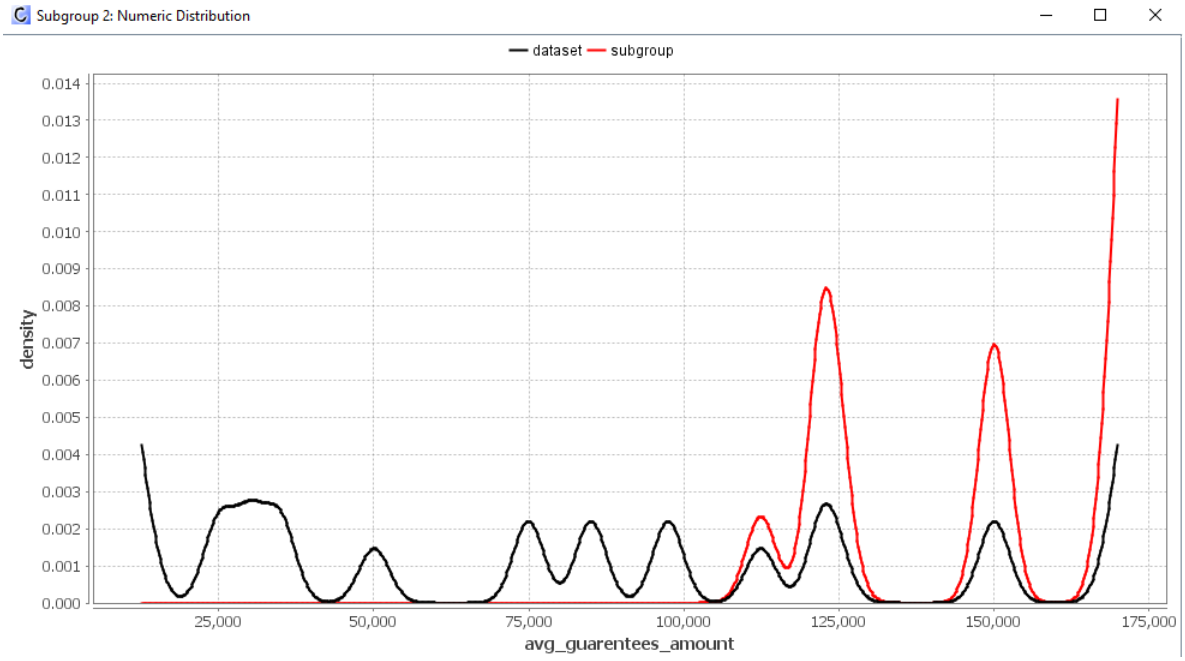


Figure 29: Model plot subgroup number 2 (primary target 'avg guarantee')

of the financial-year. It is also not usual for these kind of associations to make great profit. With these visualizations it would be easier to pick up where things are going wrong. For example, usually a company has a *key point indicator (KPI value)* to see how effectively a company is achieving key business objectives. If you take for example the end-of-the-year balance, if this value is above the KPI settled by SWS, then this is not a good view for the related sport association, since they are non-profit organizations. Furthermore, each year the number of projects are increasing, so it will be more difficult to review or look-back at these data in the Excel files and it is definitely hard to be able to compare this value between the values for each province or sport association, that's why these visualizations can be a great benefit for the company SWS to review the financial-data.

From the results of this work, we can conclude that is possible for SWS to extract useful knowledge and uncover unknown patterns in the financial historical data of the sport associations. The data should be in the right structure to be able to analyze the data using data visualization and data mining techniques such as Tableau and subgroup discovery. This knowledge can be used to more easily and accurately establish the current financial position of the sport associations and thus

make evidence based decisions about their financial situations.

4.7 Discussion: subgroup discovery

As for the data mining technique, subgroup-discovery was used, which helped to uncover some hidden patterns in the data that would not be easily detected just by looking at the data Using subgroup discovery we can see for example patterns in the income and expense data records that has big financial consequences for the sport association, like a great loss. By looking at these patterns in the data, you can predict how the financial situation of a sport association doing.

4.8 Limitations and future work: subgroup discovery

One particular limitation of this work was the amount of data available. In this project, only the data set related to the sport type 'voetbal (football)' was analyzed. Also, only the data for a limited amount of financial years was available. This definitely does have an effect on the analysis of the data using data mining techniques. Furthermore, the data is not all complete. For a lot of the loans' data, their *id* (loan_id) and the *date issued* was missing. Although there were some really interesting results, there is still room for improvements. First, it would be good to test this work on a more complete data-set. This will also give the opportunity to compare the results between the different sport associations from different sport types. It would also be good to provide visualization about the debt progression of the different sport associations and be able to compare these results between the sport associations of the same sport type. Furthermore, it would be nice to see the results when using a different data mining technique, for example *clustering*. There was not enough time to also perform clustering with this work.

5 Conclusion

Stichting Waarborgfonds Sport (SWS) is an independent guarantee institution that arranges guarantees and provide financial feasibility advice to the sport associations. SWS collects each year all the financial data from each of the sport associations. At the end of a financial year, SWS analyze the financial data of each sport association. Based on this analysis, the financial position of the sport association is determined and some recommendations are offered to the sport association in case the financial situation of the sport association is not in a good state. The financial position of the sport association is important if a sport association wants to make a loan at the bank. The bank in question asks for guarantee and this can be arranged by SWS as well. Over the years, more sport associations become affiliated with the SWS and also more financial data is gathered each year from the sport associations. Currently, SWS collects the data in Excel sheets. Analyzing increasing amount of these in a traditional manner has become more difficult and it is quite a challenge to discover trends in the data by looking at the data. It has also become impossible to avoid human errors in the data collected from the sport associations. In this project, we developed a proof-of-concept (poc) which includes a method and an artifact to have a better view of the data and be able to analyze the data and make data-driven decisions (DDDM) based on evidence-based, historical data and not only rely only on instinct.

The proposed solution is a poc based on business intelligence techniques, in particular data visualization and data mining. This method could help SWS experts to better understand the data, automate work processes and uncover unknown patterns in the financial data and aid in the decision-making process by giving non-obvious indications about the financial stability of the sport associations.

(R.1)First, we designed an appropriate structure for the data to enable effective analysis with data visualization and data mining techniques. The Excel sheets were not very effective in this situation. A system was created which imported all of the available data into a relational database where all the data from the different sheets can be linked to each other. For data visualizations, this database can be used directly. The data mining process required additional arrangement of the data set to be analyzed. Each part of the data set, analyzed to uncover trends, were exported to a separate Excel sheet to be analyzed using data mining techniques. (R.2, R.3)For the data visualization, the tool Tableau was used. An important characteristic of this work is a combination of predictive and descriptive analysis. That

is why subgroup discovery was chosen as the data mining algorithm. It is used to uncover unknown/non-obvious patterns that might occur in the data over the years. To analyze the data using subgroup discovery, the tool Cortana SD was chosen. (R.4) To achieve effective data visualizations and data mining results, the data was analyzed from different aspects of the data set. For example by sport association, by regions or through the end-of-the-year balance. (R.5)The developed artifact has an appropriate range of possibility to navigate freely through the data and could be extended even further. The generated reports are predefined, but can also be adjusted and viewed from different aspects of the data.

A Appendix

A.1 Financial statement report - Page 1

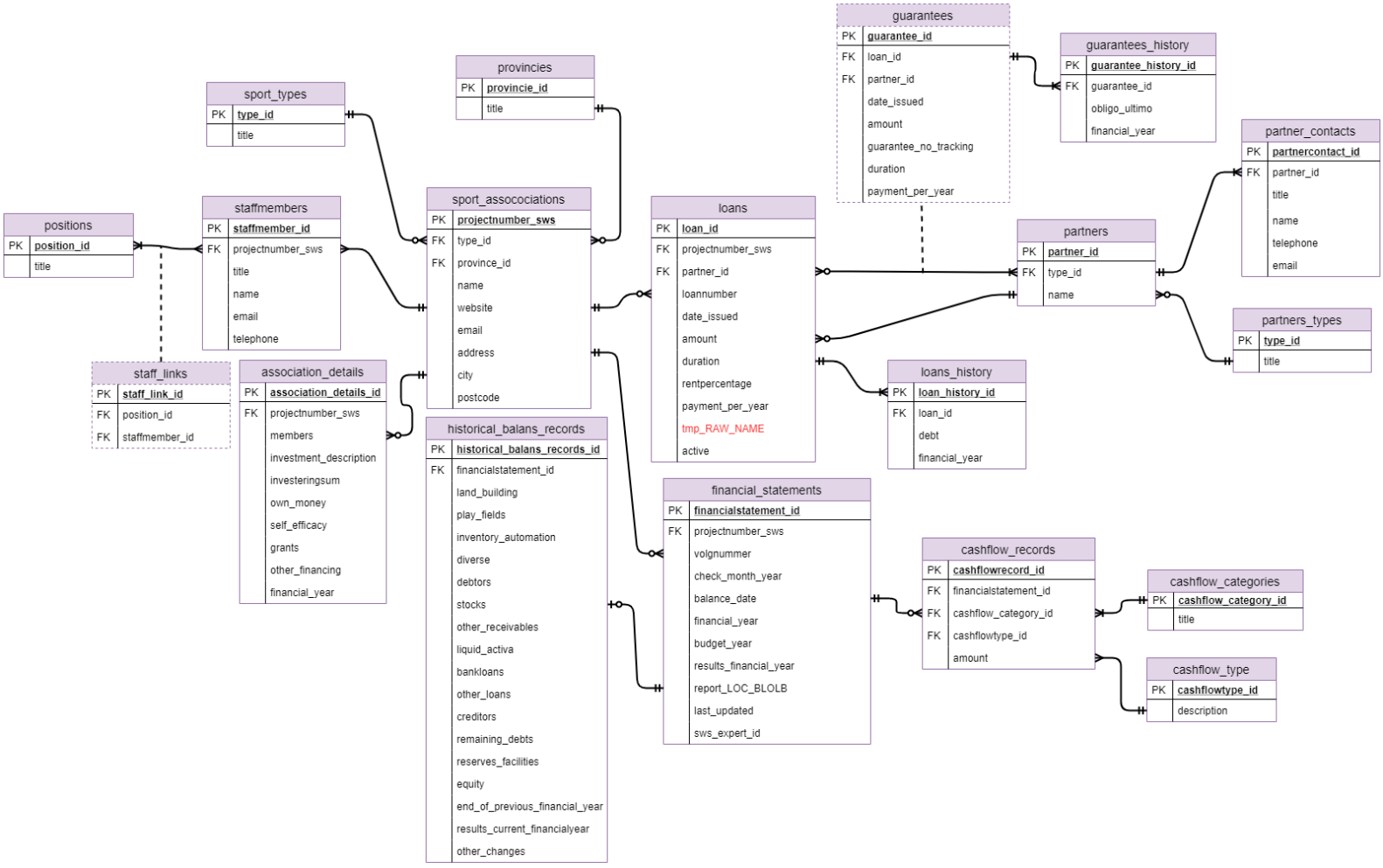
	RAPPORT CONTROLE BORGSTELLING								
	Projectnummer	**#####**		**Gecontroleerd door:**					
	Volgnummer	#		Naam	`````````````				
	Controlemaand	feb/17		Telefoon	`````````````				
	Balansdatum	30/jun/16		E-mail	`````````````				
	Boekjaar	2015/2016		website	`````````````				
	Begroting	2016/2017							
A	**Relatie SWS**								
	Naam		`````````						
	Gemeente		`````````						
	Website		`````````						
	Secretaris		`````````						
	E-mail		`````````						
	Penningmeester		`````````						
	E-mail		`````````						
	Voorzitter		`````````						
	E-mail		`````````						
B	**Leningen**		**TestBank**						
	Leningnummer								
	Datum verstrekking		23/mei/13						
	Oorspronkelijk bedrag		##.####						
	Looptijd		5 jaar						
	Rentepercentage		2.0 %						
	Aflossing per jaar		#.###						
	Schuld einde boekjaar		#.###						
C	**Zekerheden**		**SWS borgst**	**Gemeente**					
	Datum verstrekking		26/jul/13	14/sep/12					
	Oorspronkelijk bedrag		###.###	###.###					
	Looptijd		5 jaar						
	Vrijval per jaar		#.###						
	Obligo ultimo 2017		#.####						
D	**Samenvattend oordeel en advies**								
1	`````````````								
2	`````````````								
	``								
									*
E	**Partners**		**Rabobank**			**Gemeente Alkmaar**			
	Contactpersoon		`````````			````````````` / `````````````			
	Telefoon		########			06###### / 06-######			
	E-mail		`````````			`````````			
						``			

A.2 Financial statement report - Page 2

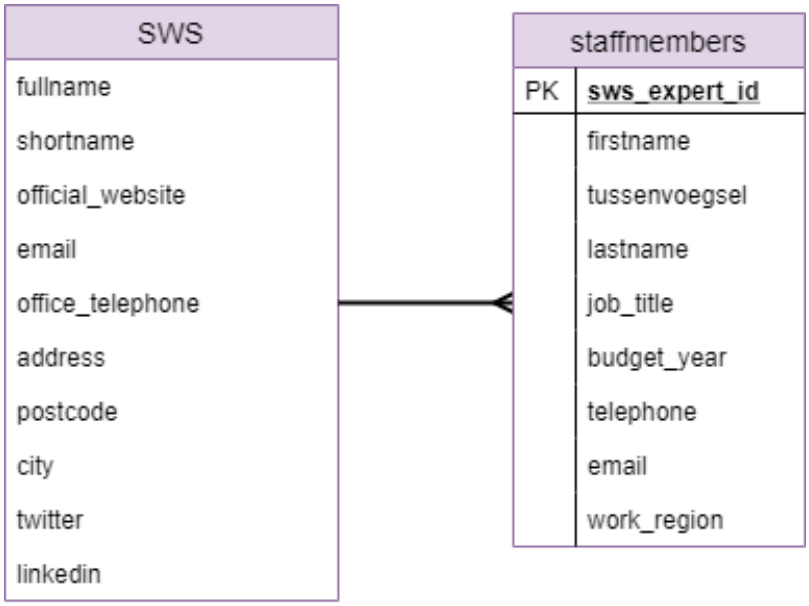
RAPPORT CONTROLE BORGSTELLING Pagina 2

	Projectnummer	#####	#	.....			Stichting Waarborgfonds Sport	
F	**Balans**			30/jun/14		30/jun/15		30/jun/16
Vaste activa								
	Grond en gebouw		#####		#####		#####	
	Velden							
	Inventaris/automatisering		#####		#####		#####	
	Diverse			#####		#####	#####	#####
Vlottende activa								
	Debiteuren		#####		#####		#####	
	Voorraden		#####		#####		#####	
	Overige vorderingen		#####	-	#####	-	#####	-
Liquide activa				#####		#####		#####
TOTAAL ACTIVA				**#VALUE!**		**#VALUE!**		**#VALUE!**
Langlopende schulden								
	Bankleningen		#####		#####		#####	
	Overige leningen			-	-	-	-	-
Kortlopende schulden								
	Crediteuren		#####		#####		#####	
	Overige schulden		#####	-		-	-	-
Eigen vermogen								
	Reserves/voorzieningen		#####		#####		#####	
	Eigen vermogen		#####	-	#####	-	#####	-
TOTAAL PASSIVA				-		-		-
Liquiditeit								
	Vlottende activa			#VALUE!		#VALUE!		#VALUE!
	Kortlopende passiva			-		-		-
	Saldo einde boekjaar			**#VALUE!**		**#VALUE!**		**#VALUE!**
Vermogen								
	Eind voorgaand boekjaar			#####		#####		#####
	Resultaat lopend boekjaar			#####		#####		#####
	Overige mutaties			-		-		-
	Reserves/voorzieningen			#####		#####		#####
	Saldo einde boekjaar			-		-		-
Toelichting resultaat en financiële positie								
1	.....							
	.....							
	.....							
2	.....							
3	.....							
4								

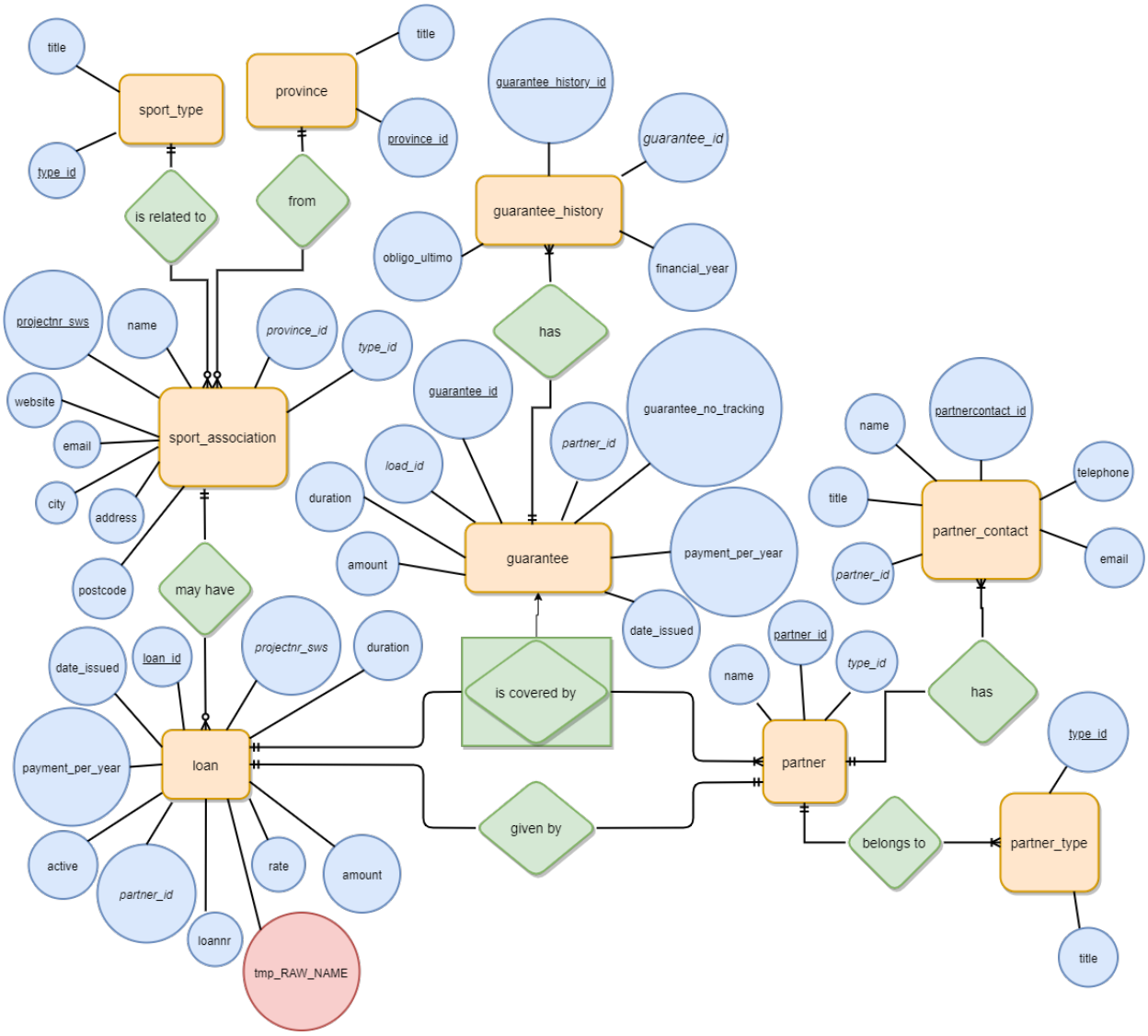
A.3 Operational database



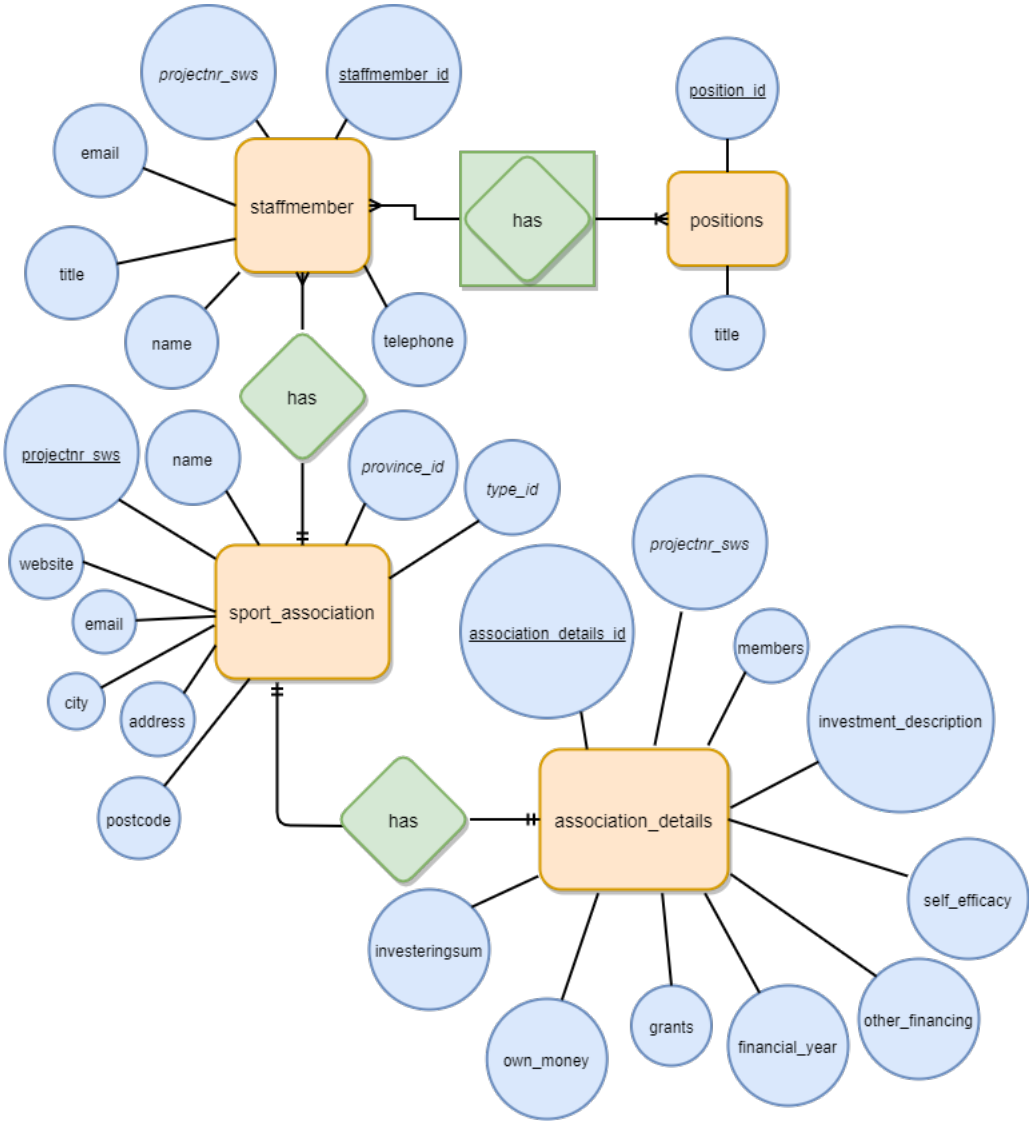
A.4 Operational database: company data



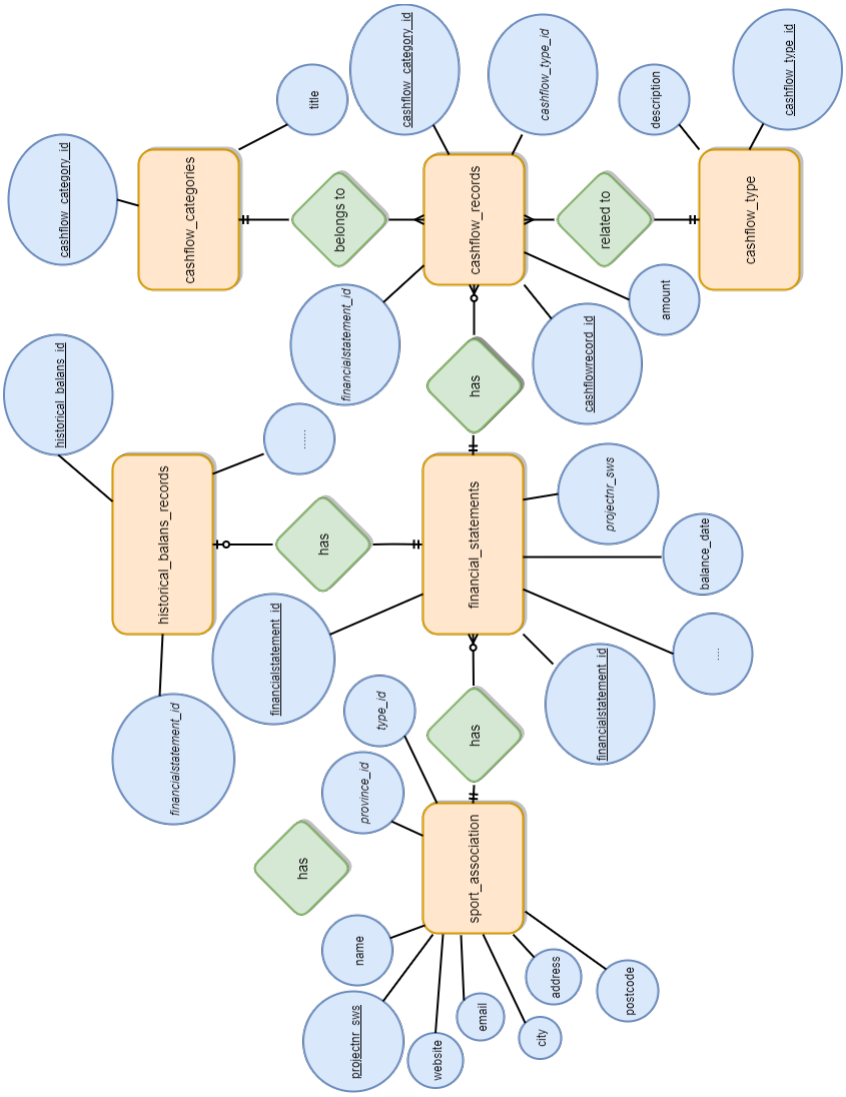
A.5 ERD-diagram (part 1)



A.6 ERD-diagram (part 2)



A.7 ERD-diagram (part 3)



References

aecm.eu. What are guarantee institutions? URL https://aecm.eu/members/what-are-guarantee-institutions/.

Antonina Kloptchenko, Tomas Eklund, Jonas Karlsson, Barbro Back, Hannu Vanharanta, and Ari Visa. Combining data and text mining techniques for analysing financial reports. *Int. Syst. in Accounting, Finance and Management*, 12(1):29–41, 2004. doi: 10.1002/isaf.239. URL https://doi.org/10.1002/isaf.239.

Ashutosh Tripathy. What are the different domains in data scientist?, 2019. URL https://www.quora.com/What-are-the-different-domains-in-data-scientist.

Branko Kavšek and Nada Lavrac. Using subgroup discovery to analyze the uk traffic data. *Metodoloski Zv.*, 1:249–264, 01 2004.

Cass Squire. Data extraction and transformation for the data warehouse. In Michael J. Carey and Donovan A. Schneider, editors, *Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data, San Jose, California, USA, May 22-25, 1995*, pages 446–447. ACM Press, 1995. doi: 10.1145/223784.223869. URL https://doi.org/10.1145/223784.223869.

C.J.Carmona, S.Ramirez-Gallego, F.Torres, E.Bernal, M.J.del Jesus, and S.Garcia. Subgroup discovery applied to the e-commercewebsite orolivesur.com. 2012. doi: 10.5220/0003982302390244.

dnserp.wordpress.com. Domain knowledge, 2013. URL https://dnserp.wordpress.com/tag/domain-knowledge/.

Fan Cai, Nhien-An Le-Khac, and Tahar Kechadi. Clustering approaches for financial data analysis: a survey. 09 2016.

Franciso Herrera, Cristóbal José Carmona, Pedro González, and María José del Jesus. An overview on subgroup discovery: foundations and applications. 29 (3):495–525, 2010. doi: 10.1007/s10115-010-0356-2.

George Seif. The 5 clustering algorithms data scientists need to know, 2018. URL https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68.

K.Kameshwaran and K.Malarvizhi. Survey on clustering techniques in data mining. 5(2), 2014.

Kenneth Jensen. Crisp-dm process diagram, 2012. URL https://commons.wikimedia.org/w/index.php?curid=24930610#/media/File:CRISP-DM_Process_Diagram.png. [Online; accessed September 11, 2020].

Lior Rokach and Oded Maimon. Clustering methods. In Oded Maimon and Lior Rokach, editors, *The Data Mining and Knowledge Discovery Handbook*, pages 321–352. Springer, 2005.

Manuel Filipe, Azevedo, and Ana Isabel Rojão Lourenço Santos. Kdd, semma and crisp-dm: a parallel overview. page 182–185, 2008. doi: 10400.22/136.

Martin Atzmüller. *Knowledge-intensive subgroup mining: techniques for automatic and interactive discovery*. PhD thesis, Julius Maximilians University Würzburg, Germany, 2006. URL http://opus.bibliothek.uni-wuerzburg.de/volltexte/2006/2100/index.html.

Marty Jacknis. 4 steps to making informed decisions, 2019. URL https://www.vistage.com/research-center/business-leadership/20190506-4-steps-to-making-informed-decisions/.

Matthias Jarke, Maurizio Lenzerini, Yannis Vassiliou, and Panos Vassiliadis. *Fundamentals of Data Warehouses*. Springer, Berlin, Heidelberg, 2003. ISBN 978-3-642-07564-3. doi: 10.1007/978-3-662-05153-5.

Neeth Ch. Decisions making: Strategic, tactical and operational decisions, 2015. URL https://www.linkedin.com/pulse/decisions-making-strategic-tactical-operational-neeth-ch/.

Neil J. Salkind. *Encyclopedia of research design*. SAGE Publications, Inc, 2010. ISBN 9781412961271. doi: 10.4135/9781412961288.

Nhien-An Le-Khac, Cai Fan, and Tahar Kechadi. Clustering approaches for financial data analysis. 07 2012.

Paschal Donohoe. The importance of sport and its benefits to communities and participants' – by paschal donohoe td, 2015. URL https://www.sportsnewsireland.com/gaa/the-importance-of-sport-and-its-benefits-to-communities-and-participants-by-paschal-donohoe-td/.

Paul J. Lavrakas. *Encyclopedia of survey research methods*. SAGE Publications, Inc, 2008. ISBN 9781412918084. doi: 10.4135/9781412963947.

Paulraj Ponniah and Pratap P. Reddy. *Data Warehousing Fundamentals*. John Wiley Sons, Inc., USA, 1st edition, 2001. ISBN 0471412546. doi: 10.5555/559231.

Pavel Berkhin. A survey of clustering data mining techniques. In Jacob Kogan, Charles K. Nicholas, and Marc Teboulle, editors, *Grouping Multidimensional Data - Recent Advances in Clustering*, pages 25–71. Springer, 2006. doi: 10.1007/3-540-28349-8_2. URL https://doi.org/10.1007/3-540-28349-8_2.

Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rüdiger Wirth. *CRISP-DM 1.0 step-by-step data mining guide*. 01 1999.

Pui Mun Lee. Use of data mining in business analytics to support business competitiveness. 17(2):53–58, 2013. doi: 10.19030/rbis.v17i2.7843.

Roel J. Wieringa. *Design Science Methodology for Information Systems and Software Engineering*. 2014. ISBN 978-3-662-43839-8.

Rui Xu and Donald C. Wunsch II. Survey of clustering algorithms. *IEEE Trans. Neural Networks*, 16(3):645–678, 2005. doi: 10.1109/TNN.2005.845141. URL https://doi.org/10.1109/TNN.2005.845141.

Tanin Phongpandecha. The three cores of data science?, 2018. URL https://towardsdatascience.com/the-three-cores-of-data-science-d58af0d7361e.

Tom Fawcett Foster Provost. Data science and its relationship to big data and data-driven decision making. 1(1), 2013. doi: 10.1089/big.2013.1508.

Xiaoling Cui and Tianli Zhong. The DEA operational efficiency evaluation of credit guarantee institutions. In Shouyang Wang, Lean Yu, Fenghua Wen, Shaoyi He, Yong Fang, and K. K. Lai, editors, *Business Intelligence: Artificial Intelligence in Business, Industry and Engineering, Proceedings of the Second International Conference on Business Intelligence and Financial Engineering, BIFE 2009, Beijing, China, 24-26 July 2009*, pages 822–825. IEEE Computer Society, 2009. doi: 10.1109/BIFE.2009.190. URL https://doi.org/10.1109/BIFE.2009.190.