Driver Pattern Clustering and Similarity Analysis Using Driving Simulation Data

Victoria Lakomski

University of Twente

Department of Psychology

Bachelor Thesis

Specialisation in Research Methodology, Measurement and Data Analysis

28.06.2023

Apa 7th Edition

1st Supervisor: Funda Yildirim

2nd Supervisor: Willem Verwey

Abstract

This study aims to explore driving behaviour patterns, whether they represent previously identified driver types and the similarities between them which could indicate a common driver type across multiple parameters. The driving behaviour patterns emerge from using the K-Means clustering algorithm on a dataset containing the parameters speed, acceleration, brake law, steering angle and heart rate, whereby two parameters are clustered together one at a time. The dataset used in this study was previously created by researchers using a driving simulator and contains 59 participants. To assess the similarity between the patterns, the clusters of parameter pairs are compared to each other using the Rand Index. The resulting similarity measure indicates whether the driving behaviour patterns of multiple parameters combined have commonalties that could imply an underlying driver type. Reason for this study is the fact that the most common reason for road accidents is human error. It is important to understand driver behaviour to increase driving performance. While driving profiles have been studied before, their underlying behavioural components have not been analysed regarding their commonalities towards each other. This could further provide insights for driving assistance systems. The results of this study show that the previously identified driver types are represented in this study. Furthermore, there are high similarity scores for a multitude of driving behaviour patterns which can indicate one common driver type.

Contents

Driver Pattern Clustering and Similarity Analysis Using Driving Simulation Data

Road traffic accidents are a substantial threat worldwide. According to the World Health Organisation (2022), 1.3 million people die each year and between 20 to 50 million people sustain injuries due to traffic accidents. Moreover, young adults between 5-29 years are especially susceptible to fatalities as a result of road traffic crashes. These accidents cause extensive emotional and economic hardship to the affected individuals themselves and their families. Undoubtedly, it is in every nations interest to decrease the number of road traffic accidents. The General Assembly of the United Nations, the main policy-making organ of the United Nations, has set the goal to halve the global number of both deaths and injuries from road traffic accidents by 2030, relative to 2017 (WHO, 2022). To achieve this goal, it is particularly important to understand the reasons why road crashes occur. While factors such as the road infrastructure, the state of the vehicle and the current traffic laws influence the accident risk, the most common reason for accidents is human error, which stresses the importance of understanding the driver's behaviour. For instance, "not looking properly", "poor judgment of other driver's path/speed" and "being in a hurry" are the main factors leading to road accidents in Great Britain with 38%, 20% and 18% respectively (GOV.UK, 2021, in Statista). In India, "speeding" (55.9%) and "careless driving" (27.5%) were the main reasons for road accident deaths in 2021 (NCRB India, 2022, in Statista). A similar picture is painted in Nigeria, where "speeding", "traffic rule violation" and "wrongful overtaking" were the top causes for accidents in the last quarter of 2021 (NBS Nigeria, 2022, in Statista). Lastly, "speeding" accounted for about 25% of fatalities on the road in the US in 2020 (NHTSA, 2022, in Statista). These numbers emphasize the need to understand driver behaviour and followingly offer personalized solutions that increase driving performance.

Driving Profiles

Clustering drivers into different driver profiles can help to understand the essential behavioural components in driving. This is an important step in traffic research because, for

example, aggressive drivers are 2.79 times more likely to be involved in road crashes than normal drivers (Adavikottu & Velaga, 2021). Generally speaking, driving behaviour refers to how the driver manoeuvres the vehicle in the surrounding environment (Elassad et al., 2020). This includes vehicle-based measurements such as the speed, acceleration, steering angle, break law and distance to close by vehicles. In particular, clustering requires analysing groups of drivers showing similar driving behaviours and characteristics, which are distinct from other groups of drivers (Tselentis & Papadimitriou, 2023). Previous studies have investigated which driving profiles exist, what clustering and classification approaches are the most accurate and which driving behaviours are the main predictors for different driver profiles. Tselentis and Papadimitriou (2023) have conducted an extensive literature review comparing the existing methodologies to identify driver profiles. Looking at their results, it becomes apparent that while studies identified many driving profiles such as aggressive, normal, moderate, (un)safe, calm or drowsy, the aggressive and normal classification are discovered very often. Furthermore, Saleh et al. (2017) have identified the driver types aggressive, normal and drowsy.

Aggressive Driving Profile

There seems to be a general consensus that harsh breaking and harsh turning are the best indicators for the aggressive driving profile. For instance, Choi et al. (2021) conducted a study using neural networks and found that harsh turning and harsh braking were the best predictors for aggressive driving whereas harsh acceleration was not significantly different from normal drivers. Similar to this, Minglin et al. (2016) found braking to be the best predictor and acceleration worse to predict aggressive driving based on the results of their study using motion sensory data. Zhou and Zhang (2019), who identified driver types using principal component analysis, confirm these findings. Moreover, Tselentis and Yannis (2019) studied a sample of 56 drivers using a data envelopment analysis and concluded that aggressive drivers violated the speed limit on 20% to 32% of the total driving time and used

their smartphone for about 16% of the drive while normal drivers violated the speed limit 6.5% and used their smartphone 1.5% of the driving time. Finally, Mccabe et al. (2020) found a positive relationship between elevated heart rate and risky driving behaviour, specifically harsher braking.

Drowsy Driving Profile

Bergasa et al. (2019) conducted a naturalistic driving study with elderly drivers and found that drowsy driving is characterised by disproportionately slow lane changes and the inability to stay in the centre of the lane, showing that the drivers were drifting more within their own lane than normal or aggressive drivers. Additionally, Shahverdy et al. (2020) confirmed these findings and added that slow changes in acceleration and a generally slow speed are characteristic of drowsy drivers.

Normal Driving Profile

Drivers with a normal profile generally show very few of the above-mentioned risky behaviours such as harsh braking, turning or disproportionally slow turning and accelerating (Shahverdy et al., 2020). Contrary to the high level of speed limit violation and smartphone usage of aggressive drivers indicated above, normal drivers violated the speed limit 6.5% and used their smartphone 1.5% of the driving time in Tselentis' and Yannis' (2019) research. This shows that normal drivers exhibit less risky behaviour than aggressive or drowsy drivers.

Clustering and Classification Methods

It is important to distinguish between what clustering and what classification methods intend to accomplish in the context of driving behaviour research. While clustering focuses on methods which identify different driving profiles, meaning they analyse groups of drivers showing similar driving behaviours and characteristics, classification focuses on methods which can group data according to preidentified driving profiles (Avcontentteam, 2023).

Clustering

In more detail, clustering is an unsupervised machine learning approach which can

identify similar groups of observations in a dataset using algorithms (De Luca, 2022). It is therefore able to analyse unlabelled data. One method to identify driving profiles is the unsupervised K-Means algorithm. Here, the researcher has to predefine the number of clusters after which the algorithm groups all observations into the set number of clusters, so that the variance in each cluster is as small as possible and the variance between the clusters is as large as possible (Anwla, 2021). Moreover, each observation can only belong to one cluster, thus the clusters are non-overlapping.

Classification

Classification techniques are supervised machine learning methods which can assign specific labels to observations in a dataset (De Luca, 2022). This only works when the researcher has prior knowledge of what the labels represent and has therefore predefined those labels. In the context of driver classification, the chosen classification method would first learn a training dataset which entails the characteristics of the labels discovered in the previous clustering stage. Next, the chosen classification method has to label observations of a new dataset accurately according to the previously learned characteristics for each label. The current study aims to use an unlabelled, data-driven approach which is why the clustering method is used in favour of the classification method.

Promoting Driving Safety

One way of promoting driving safety is through driving monitoring and assistance systems (DMAS). DMAS are designed to monitor the driver and assist them if necessary (Khan & Lee, 2019). Examples of assistance are prompting the driver when they are driving faster than the speed limit, do not use an indicator when turning or drive on the opposite side of the road (Khan & Lee, 2019). Thus, these systems aim to enhance the driver's attention. Furthermore, DMAS tracks the surrounding environment of the vehicle and is able to warn the driver about possible collisions. Khan and Lee (2019) argue that DMAS are calibrated using average driver characteristics and are therefore not able to uniquely adapt to each

individual driver. Hence, they address that the future of DMAS should include driving style recognition and personalised assistance.

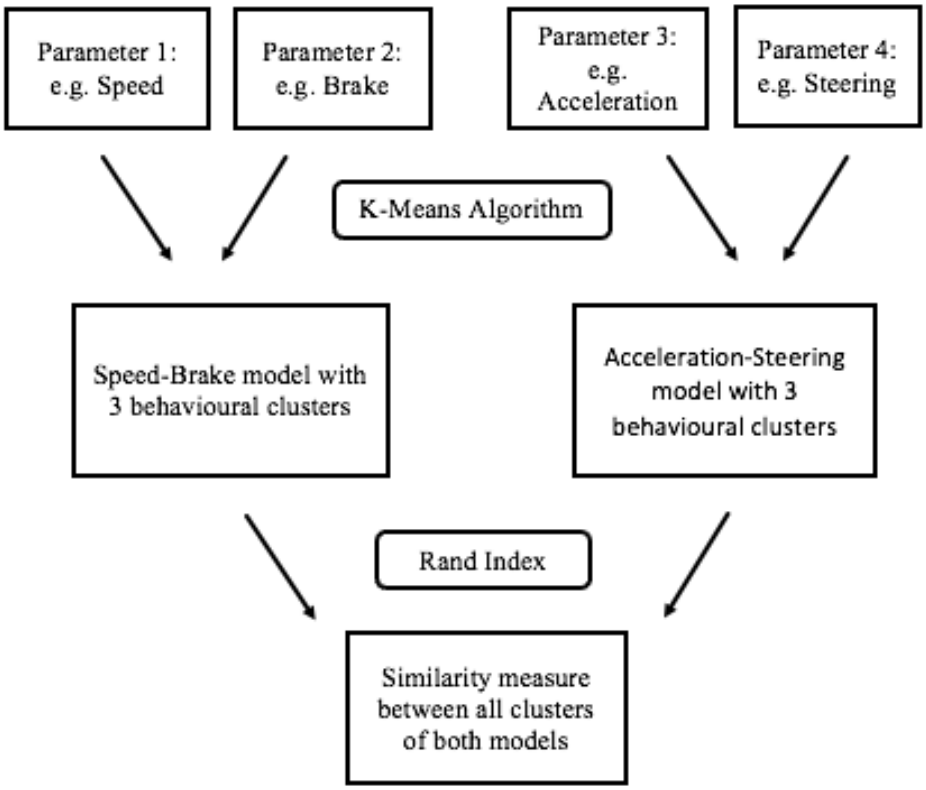
Zhang e al. (2017) developed an assistance system called SafeDrive, which is "an online and status-aware approach for detecting driving anomalies". This system is unique because it does not require labelled data and is able to identify abnormal driving behaviour in real-time via an on-board diagnostic system. It uses a state graph, which represents the normal driving style. To recognize abnormalities, it compares the real-time driving data to the state graph of a normal driving style. Zhang et al. (2017) demonstrate that SafeDrive is able to reliably identify aggressive acceleration, harsh braking and sharp turning. Moreover, SafeDrive takes into consideration the relationships and patterns between vehicle-based parameters such as speed, brake law, acceleration and steering when detecting the driving abnormalities.

Purpose of this Research

The first goal of this study is to identify drivers' behavioural patterns (clusters) that emerge from the vehicle-based measures speed, acceleration, steering angle, brake law and heart rate data. These variables are combined with each other into pairs, leading to 10 variable combinations. To create the clusters, the unsupervised K-Means clustering algorithm is used on each variable combination, which results in 10 models. This algorithm was chosen in favour of a supervised algorithm since the goal of this study is to utilise an unlabelled, data driven approach. The second aim is to review the clusters and check whether the characteristics of the aggressive, drowsy and normal driver profiles found by previous literature (Saleh et al., 2017) are represented within the data. The K-Means clustering and the reviewing step are necessary to continue with the next goal of this study. Figure 1 depicts a visualisation of this study's analysing process.

Figure 1

Visualisation of this Study's Analysing Process



Note. This process is repeated until all possible parameter combinations (10) are reached.

The third and final goal of this study encompasses a novel approach which compares the behavioural patterns (clusters) of each K-Means model to each other. To be more precise, this comparison seeks to recognise any similarities among the driving behaviour patterns across all models. Said similarity could indicate consistent characteristics across four parameters, thus relating to one common driver type. By conducting this comparison, this study provides more insights into which driving behaviour patterns are related to each other and may be associated with a certain driver type. Furthermore, analysing the similarities between driving patterns can aid assistance systems such as SafeDrive, because this system utilises relationships between driving parameters to detect anomalies in driving (Zhang et al., 2017). Instead of solely analysing the emerging patterns based on two variables and then identifying driver types, this study examines whether a pattern known to be associated with a

certain drive type is similar to another pattern, implying that the other pattern could be linked to the same driver type. The study utilises a publicly available dataset from the website osf.io which contains 59 participants who drove multiple sessions in a driving simulator.

The research questions addressed in this research are:

RQ1) What driving behaviour patterns (clusters) emerge from applying the K-Means clustering algorithm to the vehicle-based measures speed, acceleration, steering angle, brake law and physiological data heart rate? In this study, the emerging clusters serve as the dependent variables, while the input parameters are the independent variables.

RQ2) Are the driver types aggressive, normal and drowsy as proposed in previous literature (Saleh et al., 2017) represented by the emerging driving behaviour patterns (clusters) of the current study?

RQ3) Are there similarities between the emerging driving behaviour patterns (clusters) which indicate a resemblance between certain driving behaviours and subsequently imply one common driver type?

Method

Design

The dataset used in this research was originally acquired through a controlled cross-sectional study using a driving simulator.

Participants

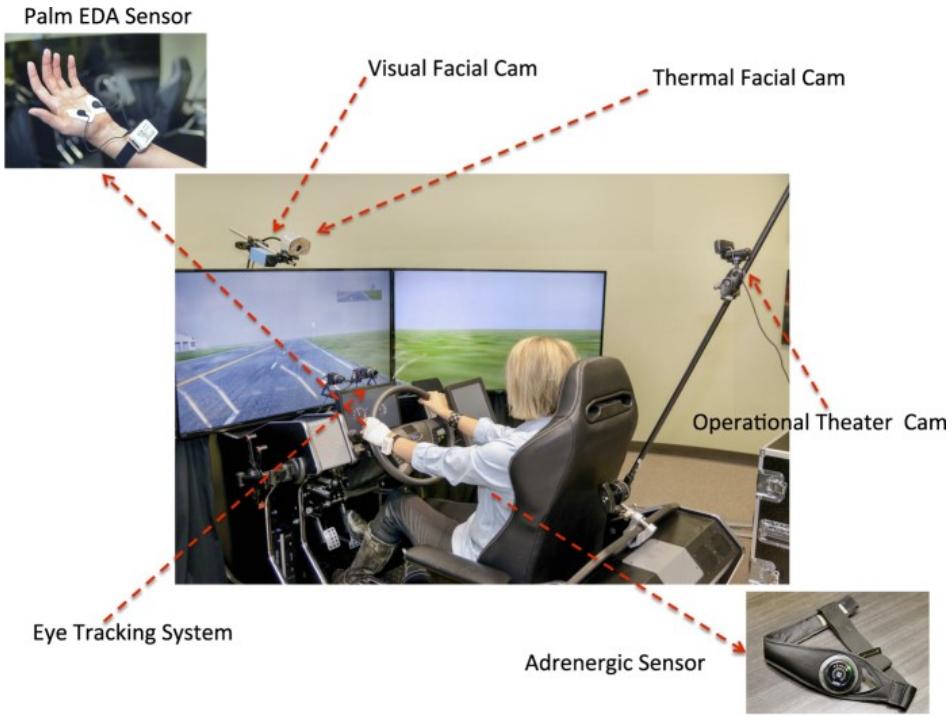
The number of total participants in the original data collection process was 68. For the purpose of this study two participants were excluded because the data of their heart rate was missing. After further exclusion of seven outliers, the total number of participants for the current study was 59, of which 28 (47.5%) were female and 31 (52.5%) were male. The mean age of participants was 44.17 (*SD* = 23.94) with an age range from 18 years to 84 years. The inclusion criteria specified that participants had to have a driving license, normal or corrected vision and at least one and a half years of driving experience (Taamneh et al., 2017).

Materials

The dataset used in this study was retrieved from the website osf.io where it was uploaded for public use. It contained measurements from participants who drove the same highway route in a driving simulator (Dcosta, 2016). Figure 2 depicts the experimental setup of the driving simulator. No further information was given by the original study about the driving simulator. According to Taamneh et al. (2017), the participants took part in the study voluntarily and were provided with an informed consent form.

Figure 2

Driving Simulator Setup.



Note. From *„A multimodal dataset for various forms of distracted driving*," by Taamneh et al., 2017, *Scientific Data*, *4*(1), doi: 10.1038/sdata.2017.110. Creative Commons 4.0 License

Procedure

Ethical approval was applied for and granted by the BMS Lab of the University of Twente (request number 230354). For the original study, participants were recruited through e-mail messages and flyer distributions from Bryan-College Station, a metropolitan area in USA Texas. Participants underwent a baseline session, during which they rested in a quiet

room. This was done to measure their baseline heart- and breathing rate. Next, the participants did a practice drive where they familiarized themselves with the driving simulator, followed by a relaxing drive with light traffic only. Finally, they did four normal drives and one drive that included a startle stimulus. All participants drove the same route on a highway and all of the measurements were recorded continuously each second while driving. Each participant completed seven drives with approximately 13 minutes, amounting to a total of 91 minutes driving time per participant and 103 hours of total driving time for all participants combined. The original study was approved by the Institutional Review Boards of the University of Houston and the Texas A&M University (Taamneh et al., 2017).

Parameters

The measurements of the dataset include speed, acceleration, brake force, steering, and lane position signals, perinasal electrodermal activity (EDA), palm EDA, heart rate, breathing rate, facial expression signals, biographical and psychometric covariates as well as eye tracking data. In this study only the variables speed (in km/h), acceleration (in °), brake force (in Newton), steering (in Radian) and heart rate (in bpm) were utilized for the analysis since analysing all of the parameters lays outside the scope of this study.

Data Analysis

Preparation of the Dataset

The dataset was analysed using the RStudio program version 1.3.1073 with the following packages: tidyverse, CataCombine, factoextra, cluster, ggpubr, ggplot2, deplyr, plotrix, fossil. Initially, each participant was represented by their own dataset which is why the first step was to merge all participants into one large dataset. Due to the requirements of the K-Means analysis, it was necessary to compute the mean of each variable for each participant. Consequently, each participant was now represented by one (mean) observation per variable instead of a large quantity of observations for each second of driving. Following that, the demographic data of the participants were integrated to the dataset containing the

driving measurements. The next step was to check the dataset for missing values which resulted in the exclusion of two participants due to the lack of heart rate values. Furthermore, the baseline session was excluded since no driving measurement were made and the eighth drive was excluded because it contained a startle stimulus. After that, the variables lane position signals, EDA, palm EDA, breathing rate, facial expression signals, biographical and psychometric covariates and the eye tracking variable were excluded since they lay outside the scope of this study.

Descriptive Statistics

The mean and standard deviation were calculated for each parameter. Furthermore, Shapiro-Wilk's normality test was performed, and distribution plots were made for each variable respectively. The R-code for the Shapiro-Wilk normality test was derived from Zach (2021a) and the R-code for the distribution plots was derived from *Plotting Distributions* (n.d.).

K-Means Analysis

In preparation for the K-Means analysis, the variables speed, brake force, steering, acceleration and heart rate were scaled, resulting in a standardised dataset represented by Z-Scores. Any participants with Z-Scores > 3 were regarded as outliers and removed. Then, a dataset was made for each possible pair of parameters. For example, the variable heart rate was paired with speed, as well as brake force, steering and acceleration separately and saved as a dataset containing two variables. With five variables this amounted to a total of 10 datasets with variable pairs. Next, the elbow method was performed on each dataset to determine the optimal number of clusters. To ensure comparability between the cluster results of each variable pair, the final number of clusters to be used for the K-Means analysis was chosen based on which optimal number was calculated most frequently by the elbow method. After the cluster number was chosen, K-Means was applied to each dataset and a graph was plotted for each K-Means model to visualise its clusters. Moreover, the explained variance for

each K-Means model was calculated to assess the model's applicability. In total, the output of each K-Means model describes the number of observations in each cluster, the cluster means, the clustering vector and the ratio of between sum squares and total sum squares. The R-code used for the elbow method and the K-Means analysis was derived from Zach (2022), James et al. (2013) and Statology (2020, in GitHub).

Cluster Comparison

Using the Rand Index, the cluster vectors of all K-Means models were compared with each other to determine their similarity. For example, the cluster vector of Speed-Acceleration was compared to the cluster vector of Speed-Break, Speed-Steering, Speed-Heart Rate and the remaining vectors of each parameter pair. The Rand Index function examines whether a participant belongs to the same cluster within two different K-Means models. The mathematical formula of the Rand Index is (Pedregosa et al., 2011):

$$RI = \frac{a + b}{C_2^{n\ samples}}$$

- a refers to the number of pairs that are the same in both clusters
- b refers to the number of pairs that are in different clusters
- $C_2^{n\ samples}$ refers to the total number of possible pairs

The resulting similarity measures were summarized in a matrix and colour-coded depending on their value. The R-code for the RI was derived from Zach (2021). The entire R-code used in this study from the data preparation to the cluster comparison is attached in Appendix A.

Results

Descriptive Statistics

The mean, standard deviation and the p-value for Shapiro-Wilk's normality test were calculated for each parameter and are summarised in Table 1. Both Speed and Heart rate are

normally distributed ($p > .05$) while Acceleration, Brake and Steering are not normally distributed ($p < 0.05$).

Table 1

Mean, Standard Deviation and Shapiro-Wilk Normality Test P-Value

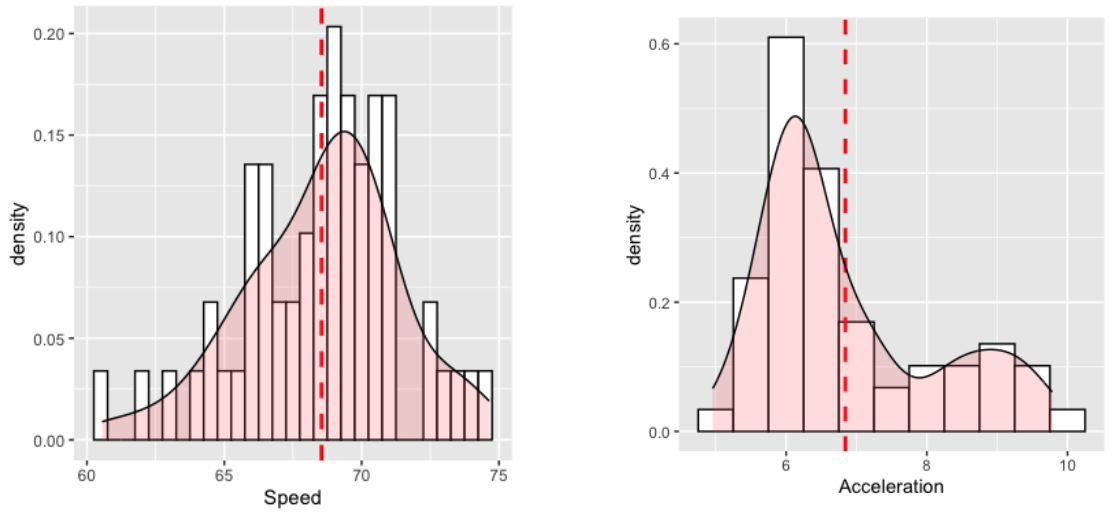
Parameter	*M*	*SD*	Shapiro-Wilk P-Value
Speed	68.5	2.6	.57
Acceleration	6.8	1.2	p < .001
Brake	12.6	9.8	.001
Steering	-0.00005	0.001	p < .001
Heart Rate	76.8	13.2	.13

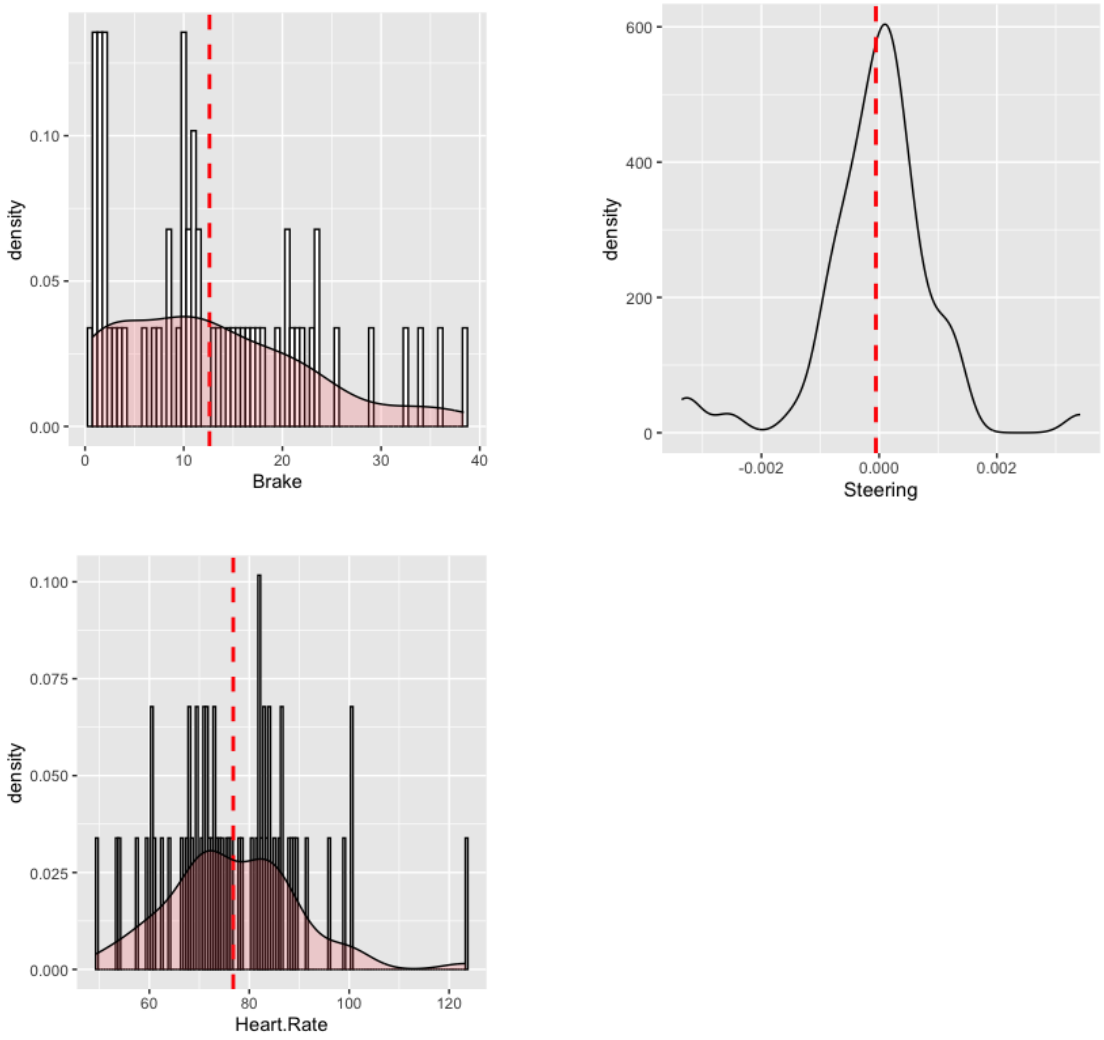
Note. A p-value > 0.05 implies that the distribution of the variable is not significantly different from normal distribution.

A histogram was plotted for each parameter to visualize their distribution. Additionally, a density curve and red line indicating the mean were added to illustrate the distribution in more detail (see Figure 3). Since the steering parameter was measured on a very small scale, it was not possible to plot a histogram for this graph.

Figure 3

Distribution plot of each Variable With Density Curve and Mean





Note. Since the steering parameter was measured on a very small scale, it was not possible to plot a histogram for this graph because each bar would be minuscule.

Elbow Method

After all the parameter pairs were made, the elbow method was applied to each pair to determine the ideal number of clusters for the K-Means analysis. The elbow graphs for each pair are attached in Appendix B. Overall, the optimal number of clusters which was found the most is three. The result for the Speed-Steering pair indicated that four clusters would be optimal and both Acceleration-Steering and Heart Rate-Steering indicate a number between three and four. This is because it can be difficult to interpret where the elbow of the graph, that is, where the total within sum of squares stop to decrease significantly, is. The result for the Brake-Steering pair was a cluster number between four and five. Since the clusters of the

K-Means models needed to be comparable, three was chosen to proceed further. The names and results for each pair are summarised in Table 2.

Cluster Variance

Following the computation of all K-Means models for each parameter pair, their explained variance, which is derived from the ratio of between sum squared and total sum squared, was calculated. The Acceleration-Brake model accounts for the highest explained variance with 81.3 % while the Heart Rate-Steering model accounts for the lowest explained variance with 52.5 %. The remaining values are summarised in Table 2 along with the elbow method results.

Table 2

Summary of Parameter Pairs, Elbow Method Results and Explained Variance of all K-Means Models

Parameter Pair	Elbow Method Result	Explained Variance
Speed-Acceleration	3	68.1%
Speed-Brake	3	58.8%
Speed-Steering	4	55%
Speed-Heart rate	3	57.2%
Acceleration-Brake	3	81.3%
Acceleration-Steering	3/4	68.1%
Acceleration-Heart rate	3	63.2%
Heart rate-Brake	3	58.7%
Heart rate- Steering	3/4	52.5%
Brake- Steering	4/5	60.9%

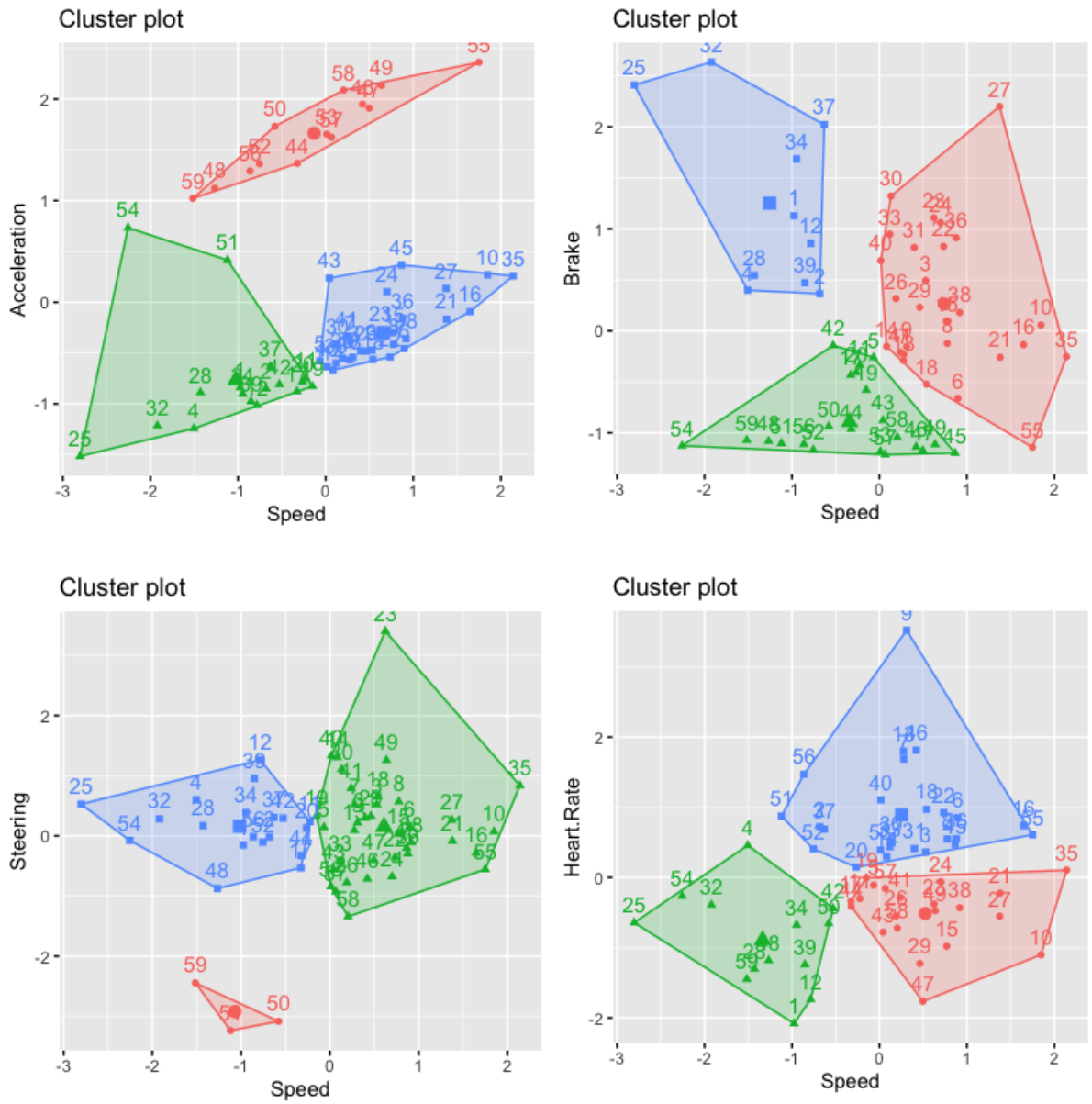
K-Means Clusters

Each K-Means clustering model was visualised and can be seen in Figure 4. The

minimum number of observations within a cluster was three, namely in cluster 1 (red) of the Speed-Steering model and cluster 2 (green) of the Acceleration-Steering model. The maximum number of observations within a cluster was 44, in cluster 1 (red) of the Acceleration-Steering model.

Figure 4

K-Means Clusters for Each Parameter Pair. Cluster 1 = Red, Cluster 2 = Green, Cluster 3 = Blue. The Naming of the Clusters is Arbitrary.

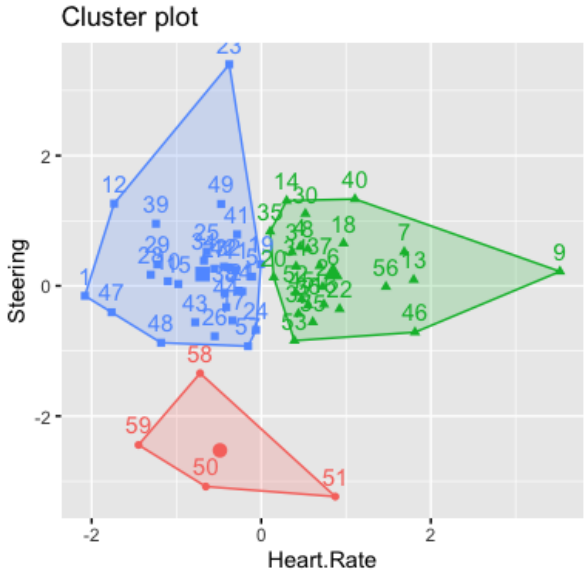


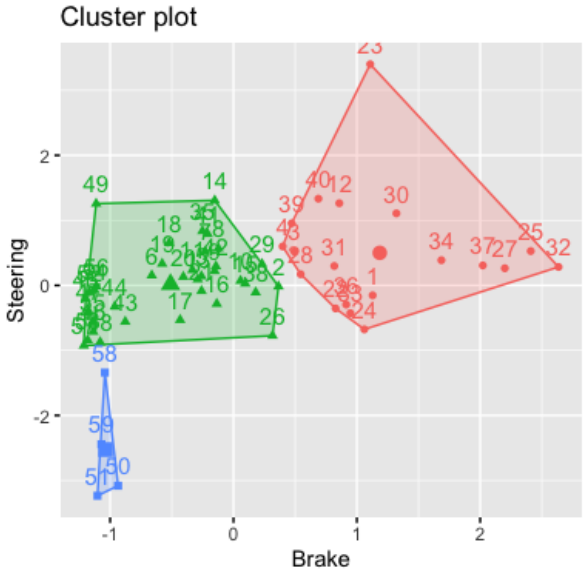

Cluster plot
Brake
Acceleration


Cluster plot
Steering
Acceleration


Cluster plot
Heart.Rate
Acceleration


Cluster plot
Brake
Heart.Rate


Cluster plot
Steering
Heart.Rate


Cluster plot
Steering
Brake

Cluster Comparison

The Rand Index (RI) was calculated for each possible combination of cluster vectors and is summarised in Table 3. The default outcome range of the RI measure is 0 to 1, with 0 indicating no similarity and 1 indicating perfect similarity. The highest similarity measure is between Speed-Brake and Speed-Acceleration (*RI* = 0.81) as well as Heart Rate-Steering and Acceleration-Heart Rate (*RI* = 0.81) and the lowest similarity measure is between Acceleration-Steering and Speed-Heart Rate (*RI* = 0.47) as well as Heart Rate-Brake and Acceleration-Steering (*RI* = 0.47).

Table 3

Cluster Similarity Matrix Using the Rand Index.

	SpAcc	SpB	SpSt	SpHR	AccB	AccSt	AccHR	HRB	HRSt	BSt
SpAcc	1.00	0.81	0.78	0.62	0.68	0.71	0.70	0.55	0.55	0.54
SpB	0.81	1.00	0.70	0.62	0.65	0.57	0.63	0.59	0.54	0.60
SpSt	0.78	0.70	1.00	0.62	0.52	0.57	0.52	0.49	0.55	0.57
SpHR	0.62	0.62	0.62	1.00	0.55	0.47	0.73	0.70	0.80	0.56
AccB	0.68	0.65	0.52	0.55	1.00	0.70	0.69	0.65	0.53	0.71
AccSt	0.71	0.57	0.57	0.47	0.70	1.00	0.69	0.47	0.55	0.55
AccHR	0.70	0.63	0.52	0.73	0.69	0.69	1.00	0.65	0.81	0.52
HRB	0.55	0.59	0.49	0.70	0.65	0.47	0.65	1.00	0.69	0.66
HRSt	0.55	0.54	0.55	0.80	0.53	0.55	0.81	0.69	1.00	0.56
BSt	0.54	0.60	0.57	0.56	0.71	0.55	0.52	0.66	0.56	1.00

Note. Colour scheme as follows: Dark Red = 1, Red ≥ 0.8, Light Red ≥ 0.7, Blue ≥ 0.6, Light Blue ≥ 0.5, White ≥ 0.4

Table 4 provides an overview of how many cluster comparison pairs are in certain Rand Index ranges. The cut-off points are 0.8, 0.7, 0.6, 0.5 and 0.4, resulting in five different ranges.

Table 4

Number of Cluster Combinations within Specific Rand Index Range

	Rand Index				
	≥ 0.8	0.79 - 0.7	0.69 - 0.6	0.59 - 0.5	0.49 - 0.4
Total number of cluster combinations	3	8	13	18	3

Discussion

This study had three goals. The first one was to examine what driving behaviour patterns (clusters) can be detected using driver's behaviour based on the performance metrics speed, acceleration, brake, steering and heart rate. The second goal was to review whether the driver types aggressive, normal and drowsy, proposed in previous literature (Saleh et al., 2017), are represented by the emerging driving behaviour clusters. The third goal aimed at analysing whether there are any similarities between the driving behaviour patterns (clusters), which indicate a resemblance between certain driving behaviours and subsequently imply one common driver type.

Through this analysis, this study enrichens the field of driving research and could be useful for providing driving safety. Driving assistence systems such as SafeDrive examine the relationships between vehicle-based parameters to detect driving behaviour that deviates from the normal driving type (Zhang et al., 2017). Since this study analyses the similarity between driving patterns across different parameters, it could bring beneficial insights for developers of assistance systems by providing more detail about the relationships of driving patterns toward each other.

The dataset used in this study was created by researchers previously conducting a driving simulation study and contained 59 participants after exclusions. To accomplish the first goal, the five parameters were combined into pairs, resulting in ten datasets each containing two variables. Then the K-Means analysis was performed on these datasets. To

achieve the second goal, the emerging clusters were visually examined to determine if the clusters exhibit specific behaviours related to the driver types aggressive, normal and drowsy. To accomplish the final goal, all K-Means models were plotted in a matrix and the RI was used to measure their similarity.

K-Means Models

Several K-Means models are highlighted in this section based on their explained variance, which informs about the applicability of the model in addition to the cluster disparity. It is clear that the Acceleration-Brake model is the most applicable of all the K-Means models due to it having the highest explained variance. In more detail, the Acceleration-Brake model has a low within-cluster variance, meaning that the observations within one cluster are highly similar to each other. Furthermore, this model has a high between-cluster variance, indicating that the three clusters have a low similarity to each other. While the K-Means algorithm makes the between-cluster variance as high as possible by default, the ratio between the within-cluster variance and the between-cluster variance leads to the highest explained variance in the Acceleration-Brake model compared to the remaining models. The visualised Acceleration-Brake model shows three clearly distinct clusters. Cluster 1 (red) shows a group of drivers who exhibit an average to slightly below average acceleration and braking force. Cluster 2 (blue) represents drivers who use a low acceleration force but high braking force. Lastly, drivers who are in cluster 3 (green), display high accelerating and low braking behaviour.

While the Speed-Acceleration model and the Acceleration-Steering model have a moderate to high variance, their clusters are not as distinctive as the Acceleration-Brake clusters. For example, cluster 1 (red) of the Speed-Acceleration model depicts drivers who show generally high accelerating behaviour but at the same time low and high speeding behaviour. Moreover, cluster 2 (green) illustrates drivers who are comparatively low in both acceleration and braking behaviour but nevertheless two drivers are visibly higher in

accelerating behaviour than the rest of cluster 2, making the cluster less compact. Cluster 3 (blue) shows drivers who have moderate accelerating behaviour and high speeding behaviour.

The Heart-Rate Steering model has the lowest explained variance. This means that the between-cluster variance is lower compared to other models, and the within-cluster variance is higher, making the drivers within one cluster less similar to each other. Looking at the cluster plot of this model it becomes apparent that many observations accumulate around a moderate heartrate and moderate use of steering angle. Cluster 2 (green) and cluster 3 (blue) merely differ from each other due to a scarce number of drivers falling further away from the cluster centroid than the rest of the drivers within that particular cluster. Cluster 1 (red) is more distinct since it embodies drivers who have a low to moderate heart rate and use a gentle steering angle. Yet, this cluster only holds four people.

Generally speaking, a high explained variance could optimize the process of data collection and analysis because one parameter is predictive of the other. In this case, accelerating and braking behaviour explain each other very well which is why it might be possible to only analyse one of both parameters and draw conclusions about the other.

Connection to Driver Types Proposed by Previous Literature

Previous research suggests that aggressive drivers are characterised by harsh braking, harsh turning as well as speeding behaviour (Choi et al, 2021; Tselentis & Yannis, 2019). Moreover, this kind of risky driving behaviour is related to an elevation of the heart rate (Mccabe et al., 2020). Looking at the cluster visualisation of the Brake-Steering model, cluster 1 (red) represents drivers who use a large brake force and slightly above average steering angle. Thus, the aggressive driver type as described in previous literature is represented in the current study. Furthermore, cluster 2 (green) of the Speed-Steering model shows drivers who are faster than average and use a large steering angle, indicating aggressive driving tendencies.

The drowsy driving profile is distinguished by slow lane changes and difficulty staying in the center of lane (Bergasa et al., 2019). Furthermore, slow changes in acceleration and slow speed are characteristic of this driving profile (Shahverdy et al., 2020). Cluster 2 (green) of the Speed-Acceleration model consists of drivers who are slower in speed than average and use a low acceleration force. While there are two drivers who exhibit a normal to slightly above average acceleration force, the 15 remaining drivers in this cluster clearly fall within the drowsy driving profile. Additionally, cluster 1 (red) of the Speed-Steering model shows drivers who are low in speed and use a very soft steering angle, also representing the drowsy driving profile.

Since the normal driving type is characterised by not expressing as many risky behaviours as aggressive or drowsy drivers, clusters representing this driver type should accumulate around the average of the parameters (Shahverdy et al., 2020). This is the case for cluster 2 (green) of the Brake-Steering model. Drivers in this cluster use an average steering angle and average to slightly below average brake force. Additionally, cluster 1 (red) of the Acceleration-Brake model represents normal drivers because they use an average to slightly below average acceleration and brake force.

Cluster Similarity

Highlighting Similarity Scores

The cluster similarity indicated by the Rand Index shows how much resemblance two driving patterns have across four parameters. If a driving pattern known to be indicative of a particular driver type is highly similar to another driving pattern, it suggests that the latter driving pattern could be associated with the same driver type. Three model combinations have a very high ($RI \geq 0.8$) similarity score. First, the clusters of the Speed-Brake model and the Speed-Acceleration model show a high degree of agreements. This indicates that the drivers' behavioural patterns for speeding together with braking in combination with speeding and accelerating are similar. Since the Speed-Brake model has a cluster representing the drowsy

driver type, the high similarity between the two models implies that certain driving patterns of the Speed-Acceleration model could also be related to drowsy drivers.

Moreover, the Heart Rate-Steering model has a very high similarity measure ($RI \geq 0.8$) with the Speed-Heart Rate model and the Acceleration-Heart Rate model. This demonstrates that there is some commonality among people's Heart Rate-Steering driving pattern and their Speed-Heart Rate and Acceleration-Heart Rate driving pattern.

Next, eight model combinations have a high (RI = 0.79 – 0.7) similarity score. The Speed-Steering driving pattern is similar to the Speed-Acceleration and Speed-Brake driving patterns. Since the Speed-Steering model has clusters indicative of aggressive and drowsy drivers, the high similarity suggests that the driving patterns of Speed-Acceleration and Speed-Brake might also be indicative of both driver types. As mentioned above, the Speed-Acceleration model entails a cluster with drowsy drivers. The Speed-Steering model also has a cluster representing drowsy drivers; thus this corresponds to the high similarity measure between the two models. Furthermore, the Brake-Steering driving patterns are similar to the Acceleration-Brake driving patterns. The Brake-Steering model entails a cluster representing aggressive drivers which subsequently implies that the behavioural patterns of the Acceleration-Brake model could also be related to the aggressive driving profile. This is an interesting finding because Choi et al. (2021) found acceleration to not be a great predictor for aggressive driving and Minglin et al. (2016) suggest that accelerating behaviour is not significantly different in aggressive drivers compared to normal drivers.

Additionally, the Acceleration-Steering driving pattern is similar to the Speed-Acceleration and the Acceleration-Brake pattern. Acceleration-Heart Rate patterns also have a high similarity to the Speed-Acceleration and Speed-Heart Rate patterns. Lastly, Heart Rate-Brake driving behaviour is similar to Speed-Heart Rate behaviour. These findings show that acceleration is an underlying factor that results in numerous high similarity scores when it is combined with different parameters. Considering that acceleration leads to many high

similarity scores this could indicate that accelerating behaviour is relatively similar across participants and combinations with other parameters. Therefore, the accelerating parameter is not able to distinguish between different driver types well because it is comparatively consistent across multiple models. This would correspond to the findings of Minglin et al. (2016), who argue that accelerating behaviour is not significantly different in aggressive drivers compared to normal drivers.

Non-Overlapping Cluster Combination

It is important to mention that all of the model combinations in the very high and high similarity range have one overlapping parameter which could account for their high similarity. However, the combination of the Acceleration-Heart Rate model with the Speed-Brake model demonstrates the highest similarity score for non-overlapping cluster combinations. While the similarity itself is only moderate (RI = 0.63), it is the only combination with a similarity this high and all different parameters. In other words, people's Acceleration-Heart Rate driving pattern has moderate commonality with their Speed-Brake driving pattern.

Limitations and Future Recommendations

The main limitation of this study is that the Rand Index is only able to measure the overall similarity between all clusters of two K-Means models and does not give an exact similarity measure for each cluster separately. For instance, one cluster representative of a particular driver type within a K-Means model cannot directly be compared to only one cluster of a different K-Means model. Calculating a similarity measure between particular clusters is useful because it would result in a more thorough cluster comparison. This way it would be possible to separately analyse the emerging driver types of driving behaviour patterns which have commonalities between each other. It is therefore recommended that future studies develop an algorithm that is able to make this type of comparison.

Nevertheless, this study contributes to the research field of driver type analysis because overall similarities between driving behaviour patterns are established and the driver

types identified by previous research are confirmed once more. Developers of assistance systems could use the established similarities to further improve their work. Future studies can benefit from this because they can further analyse patterns found to be similar according to driver types represented by individual clusters.

Conclusion

This study analysed the driving behaviour clusters emerging from the parameters speed, acceleration, brake, steering and heart-rate. This was a necessary step since the study aimed to examine whether the driving behaviour patterns of two K-Means models combined are similar to each other. A similarity between driving patterns could imply a common driver type and subsequently aid driving assistance systems such as SafeDrive because the system uses relationships between driving parameters to detect driving anomalies (Zhang et al., 2017). A secondary analysis using the K-Means algorithm and the Rand Index was conducted on a previously created dataset containing 59 participants.

The results show that the driver types aggressive, drowsy and normal, proposed by previous literature (Saleh et al., 2017), are represented in this study. Moreover, similarities have been found across various driving behaviour patterns. For instance, behavioural clusters of Speed-Acceleration could relate to drowsy drivers because they are highly similar to Speed-Brake behavioural clusters. The latter has a cluster representing drowsy drivers. The Speed-Steering model has clusters representing both aggressive and drowsy drivers and is similar to the driving patterns of the Speed-Acceleration and the Speed-Brake model. This shows that the two latter models could also be related to aggressive and drowsy drivers. The results of this study suggest that accelerating behaviour leads to multiple high similarity scores when combined with a different driving parameter, which could indicate that acceleration is a consistent behaviour across drivers. Therefore, it might not be a parameter that is able to distinguish well between driver types. Possible limitations of this study include that the Rand Index gives an overall similarity score for all clusters combined of a K-Means

model. Future research should develop an algorithm that is able to directly compare individual clusters of a K-Means model.

References

Adavikottu, A., & Velaga, N. R. (2021). Analysis of factors influencing aggressive driver behavior and crash involvement. *Traffic Injury Prevention*, *22*(sup1), S21–S26. https://doi.org/10.1080/15389588.2021.1965590

Anwla, P. K. (2021, 1. Dezember). *What is K-Means algorithm and how it works*. TowardsMachineLearning. https://towardsmachinelearning.org/k-means/

Avcontentteam. (2023). Classification vs. Clustering- Which One is Right for Your Data? *Analytics Vidhya*. https://www.analyticsvidhya.com/blog/2023/05/classification-vs-clustering/#h-what-is-classification

Bergasa, L. M., Araluce, J., Romera, E., Barea, R., Lopez-Guilen, E., Del Egido, J. & Hernanz-Mayoral, C. A. (2019). Naturalistic Driving Study for Older Drivers based on the DriveSafe App. *International Conference on Intelligent Transportation Systems*. https://doi.org/10.1109/itsc.2019.8917079

Choi, Y. D., Park, K. W., Park, E., & Kim, H. K. (2021). Unsupervised Driver Behavior Profiling Leveraging Recurrent Neural Networks. *Springer International Publishing EBooks*, 28–38. https://doi.org/10.1007/978-3-030-89432-0_3

Dcosta, M. D. (2016, November 15). *SIMULATOR STUDY I: A Multimodal Dataset for Various Forms of Distracted Driving*. osf.io. Retrieved March 15, 2023, from https://osf.io/c42cn/?view_only=

De Luca, G. (2022, 15. November). *Differences Between Classification and Clustering | Baeldung on Computer Science*. Baeldung on Computer Science. https://www.baeldung.com/cs/ml-classification-vs-clustering

Elassad, Z. E. A., Mousannif, H., Moatassime, H. A., & Karkouch, A. (2020). The application of machine learning techniques for driving behavior analysis: A conceptual framework and a systematic literature review. *Engineering Applications of Artificial Intelligence*, *87*, 103312. https://doi.org/10.1016/j.engappai.2019.103312

GOV.UK. (September 30, 2021). Distribution of contributing factors leading to road accidents in Great Britain in 2020 [Graph]. In *Statista*. Retrieved March 09, 2023, from https://www-statista-com.ezproxy2.utwente.nl/statistics/323079/contributing-factors-leading-to-road-accidents-in-great-britain-uk/

James, G. M., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning. In *Springer texts in statistics*. Springer International Publishing. https://doi.org/10.1007/978-1-4614-7138-7

Khan, M. S., & Lee, S. (2019). A Comprehensive Survey of Driving Monitoring and Assistance Systems. *Sensors*, *19*(11), 2574. https://doi.org/10.3390/s19112574

Mccabe, T., Simpkin, A., Mullins, D., Jones, E., Glavin, M. (2020, January 17-18). *The Effect of Heart Rate on Driving Style* [Paper presentation]. Bioengineering, Co Carlow, Ireland. https://www.researchgate.net/publication/338698463_The_Effect_of_Heart_Rate_on_Driving_Style

Minglin, W., Zhang, S., & Dong, Y. (2016). A Novel Model-Based Driving Behavior Recognition System Using Motion Sensors. *Sensors*, *16*(10), 1746. https://doi.org/10.3390/s16101746

NBS (Nigeria). (March 14, 2022). Leading causes of road traffic accidents in Nigeria as of 4th quarter 2021 [Graph]. In *Statista*. Retrieved March 09, 2023, from https://www-statista-com.ezproxy2.utwente.nl/statistics/1296331/main-causes-of-road-accidents-in-nigeria-by-category/

NCRB (India). (August 25, 2022). Share of road accident deaths across India in 2021, by cause [Graph]. In *Statista*. Retrieved March 09, 2023, from https://www-statista-com.ezproxy2.utwente.nl/statistics/1099025/india-share-of-road-accident-deaths-by-cause/

NHTSA. (June 24, 2022). Number of speeding-related traffic fatalities in the U.S. from 2006

to 2020 [Graph]. In *Statista*. Retrieved March 09, 2023, from https://www-statista-com.ezproxy2.utwente.nl/statistics/720326/speeding-related-fatalities-in-the-us/

Plotting distributions (ggplot2). (n.d.). http://www.cookbook-r.com/Graphs/Plotting_distributions_(ggplot2)/

Saleh, K. J., Khosravi, A., & Nahavandi, S. (2017). *Driving behavior classification based on sensor data fusion using LSTM recurrent neural networks*. https://doi.org/10.1109/itsc.2017.8317835

Shahverdy, M., Fathy, M., Berangi, R., & Sabokrou, M. (2020). Driver behavior detection and classification using deep convolutional neural networks. *Expert Systems With Applications*, *149*, 113240. https://doi.org/10.1016/j.eswa.2020.113240

Statology. (2020, December 2). *R-Guides/k_means.R at main · Statology/R-Guides*. GitHub. https://github.com/Statology/R-Guides/blob/main/k_means.R

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12, 2825–2830.*

Taamneh, S., Tsiamyrtzis, P., Dcosta, M., Buddharaju, P., Khatri, A., Manser, M., Ferris, T. K., Wunderlich, R. P., & Pavlidis, I. (2017). A multimodal dataset for various forms of distracted driving. *Scientific Data*, *4*(1). https://doi.org/10.1038/sdata.2017.110

Tselentis, D. I., & Papadimitriou, E. (2023). Driver Profile and Driving Pattern Recognition for Road Safety Assessment: Main Challenges and Future Directions. *IEEE Open Journal of Intelligent Transportation Systems*, *4*, 83–100. https://doi.org/10.1109/ojits.2023.3237177

Tselentis, D. I. & Yannis, G. (2019). Driving safety efficiency benchmarking using smartphone data. *Transportation Research Part C-emerging Technologies*, *109*, 343–357. https://doi.org/10.1016/j.trc.2019.11.006

World Health Organization: WHO. (2022, June 20). *Road traffic injuries*. https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries

Zach. (2021a). How to Perform a Shapiro-Wilk Test in R (With Examples). *Statology*. https://www.statology.org/shapiro-wilk-test-r/

Zach. (2021b). What is the Rand Index? (Definition & Examples). *Statology*. https://www.statology.org/rand-index/

Zach. (2022). K-Means Clustering in R: Step-by-Step Example. *Statology*. https://www.statology.org/k-means-clustering-in-r/

Zhang, M., Chen, C., Wo, T., Xie, T., Bhuiyan, Z. A., & Lin, X. (2017). SafeDrive: Online Driving Anomaly Detection From Large-Scale Vehicle Data. *IEEE Transactions on Industrial Informatics*, *13*(4), 2087–2096. https://doi.org/10.1109/tii.2017.2674661

Zhou, T., & Zhang, J. (2019). Analysis of commercial truck drivers' potentially dangerous driving behaviors based on 11-month digital tachograph data and multilevel modeling approach. *Accident Analysis & Prevention*, *132*, 105256. https://doi.org/10.1016/j.aap.2019.105256

Appendix A

R-Code

```
library(tidyverse)
library(DataCombine)

#Set WD#
setwd("~/Desktop/Module 11/analysis/R-Friendly Study Data")

#import all datasets#

T001  =  read.csv("T001.csv")
T002  =  read.csv("T002.csv")
T003  =  read.csv("T003.csv")
T004  =  read.csv("T004.csv")
T005  =  read.csv("T005.csv")
T006  =  read.csv("T006.csv")
T007  =  read.csv("T007.csv")
T008  =  read.csv("T008.csv")
T009  =  read.csv("T009.csv")
T010  =  read.csv("T010.csv")
T011  =  read.csv("T011.csv")
T012  =  read.csv("T012.csv")
T013  =  read.csv("T013.csv")
T014  =  read.csv("T014.csv")
T015  =  read.csv("T015.csv")
T016  =  read.csv("T016.csv")
T017  =  read.csv("T017.csv")
T018  =  read.csv("T018.csv")
T019  =  read.csv("T019.csv")
T020  =  read.csv("T020.csv")
T021  =  read.csv("T021.csv")
T022  =  read.csv("T022.csv")
T023  =  read.csv("T023.csv")
T024  =  read.csv("T024.csv")
T025  =  read.csv("T025.csv")
T026  =  read.csv("T026.csv")
T027  =  read.csv("T027.csv")
T028  =  read.csv("T028.csv")
T029  =  read.csv("T029.csv")
T031  =  read.csv("T031.csv")
T032  =  read.csv("T032.csv")
T033  =  read.csv("T033.csv")
T034  =  read.csv("T034.csv")
T035  =  read.csv("T035.csv")
T036  =  read.csv("T036.csv")
T038  =  read.csv("T038.csv")
T039  =  read.csv("T039.csv")
T040  =  read.csv("T040.csv")
T041  =  read.csv("T041.csv")
```

```
T042  =     read.csv("T042.csv")
T043  =     read.csv("T043.csv")
T044  =     read.csv("T044.csv")
T045  =     read.csv("T045.csv")
T046  =     read.csv("T046.csv")
T047  =     read.csv("T047.csv")
T050  =     read.csv("T050.csv")
T051  =     read.csv("T051.csv")
T054  =     read.csv("T054.csv")
T055  =     read.csv("T055.csv")
T060  =     read.csv("T060.csv")
T061  =     read.csv("T061.csv")
T062  =     read.csv("T062.csv")
T064  =     read.csv("T064.csv")
T066  =     read.csv("T066.csv")
T068  =     read.csv("T068.csv")
T073  =     read.csv("T073.csv")
T074  =     read.csv("T074.csv")
T075  =     read.csv("T075.csv")
T076  =     read.csv("T076.csv")
T077  =     read.csv("T077.csv")
T079  =     read.csv("T079.csv")
T080  =     read.csv("T080.csv")
T081  =     read.csv("T081.csv")
T082  =     read.csv("T082.csv")
T083  =     read.csv("T083.csv")
T084  =     read.csv("T084.csv")
T086  =     read.csv("T086.csv")
T088  =     read.csv("T088.csv")

#exclude drive 8 because of startle stimulus#

T001  =     T001  %>%  filter   (Drive != 8)
T002  =     T002  %>%  filter   (Drive != 8)
T003  =     T003  %>%  filter   (Drive != 8)
T004  =     T004  %>%  filter   (Drive != 8)
T005  =     T005  %>%  filter   (Drive != 8)
T006  =     T006  %>%  filter   (Drive != 8)
T007  =     T007  %>%  filter   (Drive != 8)
T008  =     T008  %>%  filter   (Drive != 8)
T009  =     T009  %>%  filter   (Drive != 8)
T010  =     T010  %>%  filter   (Drive != 8)
T011  =     T011  %>%  filter   (Drive != 8)
T012  =     T012  %>%  filter   (Drive != 8)
T013  =     T013  %>%  filter   (Drive != 8)
T014  =     T014  %>%  filter   (Drive != 8)
T015  =     T015  %>%  filter   (Drive != 8)
T016  =     T016  %>%  filter   (Drive != 8)
T017  =     T017  %>%  filter   (Drive != 8)
T018  =     T018  %>%  filter   (Drive != 8)
T019  =     T019  %>%  filter   (Drive != 8)
```

```
T020 =	T020 %>% filter (Drive != 8)
T021 =	T021 %>% filter (Drive != 8)
T022 =	T022 %>% filter (Drive != 8)
T023 =	T023 %>% filter (Drive != 8)
T024 =	T024 %>% filter (Drive != 8)
T025 =	T025 %>% filter (Drive != 8)
T026 =	T026 %>% filter (Drive != 8)
T027 =	T027 %>% filter (Drive != 8)
T028 =	T028 %>% filter (Drive != 8)
T029 =	T029 %>% filter (Drive != 8)
T031 =	T031 %>% filter (Drive != 8)
T032 =	T032 %>% filter (Drive != 8)
T033 =	T033 %>% filter (Drive != 8)
T034 =	T034 %>% filter (Drive != 8)
T035 =	T035 %>% filter (Drive != 8)
T036 =	T036 %>% filter (Drive != 8)
T038 =	T038 %>% filter (Drive != 8)
T039 =	T039 %>% filter (Drive != 8)
T040 =	T040 %>% filter (Drive != 8)
T041 =	T041 %>% filter (Drive != 8)
T042 =	T042 %>% filter (Drive != 8)
T043 =	T043 %>% filter (Drive != 8)
T044 =	T044 %>% filter (Drive != 8)
T045 =	T045 %>% filter (Drive != 8)
T046 =	T046 %>% filter (Drive != 8)
T047 =	T047 %>% filter (Drive != 8)
T050 =	T050 %>% filter (Drive != 8)
T051 =	T051 %>% filter (Drive != 8)
T054 =	T054 %>% filter (Drive != 8)
T055 =	T055 %>% filter (Drive != 8)
T060 =	T060 %>% filter (Drive != 8)
T061 =	T061 %>% filter (Drive != 8)
T062 =	T062 %>% filter (Drive != 8)
T064 =	T064 %>% filter (Drive != 8)
T066 =	T066 %>% filter (Drive != 8)
T068 =	T068 %>% filter (Drive != 8)
T073 =	T073 %>% filter (Drive != 8)
T074 =	T074 %>% filter (Drive != 8)
T075 =	T075 %>% filter (Drive != 8)
T076 =	T076 %>% filter (Drive != 8)
T077 =	T077 %>% filter (Drive != 8)
T079 =	T079 %>% filter (Drive != 8)
T080 =	T080 %>% filter (Drive != 8)
T081 =	T081 %>% filter (Drive != 8)
T082 =	T082 %>% filter (Drive != 8)
T083 =	T083 %>% filter (Drive != 8)
T084 =	T084 %>% filter (Drive != 8)
T086 =	T086 %>% filter (Drive != 8)
T088 =	T088 %>% filter (Drive != 8)

#select variables#
```

```
T001 =     T001 %>% select(6, 7, 9:12)
T002 =     T002 %>% select(6, 7, 9:12)
T003 =     T003 %>% select(6, 7, 9:12)
T004 =     T004 %>% select(6, 7, 9:12)
T005 =     T005 %>% select(6, 7, 9:12)
T006 =     T006 %>% select(6, 7, 9:12)
T007 =     T007 %>% select(6, 7, 9:12)
T008 =     T008 %>% select(6, 7, 9:12)
T009 =     T009 %>% select(6, 7, 9:12)
T010 =     T010 %>% select(6, 7, 9:12)
T011 =     T011 %>% select(6, 7, 9:12)
T012 =     T012 %>% select(6, 7, 9:12)
T013 =     T013 %>% select(6, 7, 9:12)
T014 =     T014 %>% select(6, 7, 9:12)
T015 =     T015 %>% select(6, 7, 9:12)
T016 =     T016 %>% select(6, 7, 9:12)
T017 =     T017 %>% select(6, 7, 9:12)
T018 =     T018 %>% select(6, 7, 9:12)
T019 =     T019 %>% select(6, 7, 9:12)
T020 =     T020 %>% select(6, 7, 9:12)
T021 =     T021 %>% select(6, 7, 9:12)
T022 =     T022 %>% select(6, 7, 9:12)
T023 =     T023 %>% select(6, 7, 9:12)
T024 =     T024 %>% select(6, 7, 9:12)
T025 =     T025 %>% select(6, 7, 9:12)
T026 =     T026 %>% select(6, 7, 9:12)
T027 =     T027 %>% select(6, 7, 9:12)
T028 =     T028 %>% select(6, 7, 9:12)
T029 =     T029 %>% select(6, 7, 9:12)
T031 =     T031 %>% select(6, 7, 9:12)
T032 =     T032 %>% select(6, 7, 9:12)
T033 =     T033 %>% select(6, 7, 9:12)
T034 =     T034 %>% select(6, 7, 9:12)
T035 =     T035 %>% select(6, 7, 9:12)
T036 =     T036 %>% select(6, 7, 9:12)
T038 =     T038 %>% select(6, 7, 9:12)
T039 =     T039 %>% select(6, 7, 9:12)
T040 =     T040 %>% select(6, 7, 9:12)
T041 =     T041 %>% select(6, 7, 9:12)
T042 =     T042 %>% select(6, 7, 9:12)
T043 =     T043 %>% select(6, 7, 9:12)
T044 =     T044 %>% select(6, 7, 9:12)
T045 =     T045 %>% select(6, 7, 9:12)
T046 =     T046 %>% select(6, 7, 9:12)
T047 =     T047 %>% select(6, 7, 9:12)
T050 =     T050 %>% select(6, 7, 9:12)
T051 =     T051 %>% select(6, 7, 9:12)
T054 =     T054 %>% select(6, 7, 9:12)
T055 =     T055 %>% select(6, 7, 9:12)
T060 =     T060 %>% select(6, 7, 9:12)
```

```
T061 = T061 %>% select(6, 7, 9:12)
T062 = T062 %>% select(6, 7, 9:12)
T064 = T064 %>% select(6, 7, 9:12)
T066 = T066 %>% select(6, 7, 9:12)
T068 = T068 %>% select(6, 7, 9:12)
T073 = T073 %>% select(6, 7, 9:12)
T074 = T074 %>% select(6, 7, 9:12)
T075 = T075 %>% select(6, 7, 9:12)
T076 = T076 %>% select(6, 7, 9:12)
T077 = T077 %>% select(6, 7, 9:12)
T079 = T079 %>% select(6, 7, 9:12)
T080 = T080 %>% select(6, 7, 9:12)
T081 = T081 %>% select(6, 7, 9:12)
T082 = T082 %>% select(6, 7, 9:12)
T083 = T083 %>% select(6, 7, 9:12)
T084 = T084 %>% select(6, 7, 9:12)
T086 = T086 %>% select(6, 7, 9:12)
T088 = T088 %>% select(6, 7, 9:12)

#exclude NAs#

T001 = T001 %>% na.omit()
T002 = T002 %>% na.omit()
T003 = T003 %>% na.omit()
T004 = T004 %>% na.omit()
T005 = T005 %>% na.omit()
T006 = T006 %>% na.omit()
T007 = T007 %>% na.omit()
T008 = T008 %>% na.omit()
T009 = T009 %>% na.omit()
T010 = T010 %>% na.omit()
T011 = T011 %>% na.omit()
T012 = T012 %>% na.omit()
T013 = T013 %>% na.omit()
T014 = T014 %>% na.omit()
T015 = T015 %>% na.omit()
T016 = T016 %>% na.omit()
T017 = T017 %>% na.omit()
T018 = T018 %>% na.omit()
T019 = T019 %>% na.omit()
T020 = T020 %>% na.omit()
T021 = T021 %>% na.omit()
T022 = T022 %>% na.omit()
T023 = T023 %>% na.omit()
T024 = T024 %>% na.omit()
T025 = T025 %>% na.omit()
T026 = T026 %>% na.omit()
T027 = T027 %>% na.omit()
T028 = T028 %>% na.omit()
T029 = T029 %>% na.omit()
T031 = T031 %>% na.omit()
```

```
T032  =     T032  %>%  na.omit()
T033  =     T033  %>%  na.omit()
T034  =     T034  %>%  na.omit()
T035  =     T035  %>%  na.omit()
T036  =     T036  %>%  na.omit()
T038  =     T038  %>%  na.omit()
T039  =     T039  %>%  na.omit()
T040  =     T040  %>%  na.omit()
T041  =     T041  %>%  na.omit()
T042  =     T042  %>%  na.omit()
T043  =     T043  %>%  na.omit()
T044  =     T044  %>%  na.omit()
T045  =     T045  %>%  na.omit()
T046  =     T046  %>%  na.omit()
T047  =     T047  %>%  na.omit()
T050  =     T050  %>%  na.omit()
T051  =     T051  %>%  na.omit()
T054  =     T054  %>%  na.omit()
T055  =     T055  %>%  na.omit()
T060  =     T060  %>%  na.omit()
T061  =     T061  %>%  na.omit()
T062  =     T062  %>%  na.omit()
T064  =     T064  %>%  na.omit()
T066  =     T066  %>%  na.omit()
T068  =     T068  %>%  na.omit()
T073  =     T073  %>%  na.omit()
T074  =     T074  %>%  na.omit()
T075  =     T075  %>%  na.omit()
T076  =     T076  %>%  na.omit()
T077  =     T077  %>%  na.omit()
T079  =     T079  %>%  na.omit()
T080  =     T080  %>%  na.omit()
T081  =     T081  %>%  na.omit()
T082  =     T082  %>%  na.omit()
T083  =     T083  %>%  na.omit()
T084  =     T084  %>%  na.omit()
T086  =     T086  %>%  na.omit()
T088  =     T088  %>%  na.omit()

# two participants have no measures for heart rate #
# add a row with mean of each variable into the beginning of each dataset#

T001  =     T001  %>%  add_row(!!! colMeans(.[]), .before = 1 )
T002  =     T002  %>%  add_row(!!! colMeans(.[]), .before = 1 )
T003  =     T003  %>%  add_row(!!! colMeans(.[]), .before = 1 )
T004  =     T004  %>%  add_row(!!! colMeans(.[]), .before = 1 )
T005  =     T005  %>%  add_row(!!! colMeans(.[]), .before = 1 )
T006  =     T006  %>%  add_row(!!! colMeans(.[]), .before = 1 )
T007  =     T007  %>%  add_row(!!! colMeans(.[]), .before = 1 )
T008  =     T008  %>%  add_row(!!! colMeans(.[]), .before = 1 )
T009  =     T009  %>%  add_row(!!! colMeans(.[]), .before = 1 )
```

```
T010 =  T010 %>% add_row(!!! colMeans(.[]), .before = 1 )
T011 =  T011 %>% add_row(!!! colMeans(.[]), .before = 1 )
T012 =  T012 %>% add_row(!!! colMeans(.[]), .before = 1 )
T013 =  T013 %>% add_row(!!! colMeans(.[]), .before = 1 )
T014 =  T014 %>% add_row(!!! colMeans(.[]), .before = 1 )
T015 =  T015 %>% add_row(!!! colMeans(.[]), .before = 1 )
T016 =  T016 %>% add_row(!!! colMeans(.[]), .before = 1 )
T017 =  T017 %>% add_row(!!! colMeans(.[]), .before = 1 )
T018 =  T018 %>% add_row(!!! colMeans(.[]), .before = 1 )
T019 =  T019 %>% add_row(!!! colMeans(.[]), .before = 1 )
T020 =  T020 %>% add_row(!!! colMeans(.[]), .before = 1 )
T021 =  T021 %>% add_row(!!! colMeans(.[]), .before = 1 )
T022 =  T022 %>% add_row(!!! colMeans(.[]), .before = 1 )
T023 =  T023 %>% add_row(!!! colMeans(.[]), .before = 1 )
T024 =  T024 %>% add_row(!!! colMeans(.[]), .before = 1 )
T025 =  T025 %>% add_row(!!! colMeans(.[]), .before = 1 )
T026 =  T026 %>% add_row(!!! colMeans(.[]), .before = 1 )
T027 =  T027 %>% add_row(!!! colMeans(.[]), .before = 1 )
T028 =  T028 %>% add_row(!!! colMeans(.[]), .before = 1 )
T029 =  T029 %>% add_row(!!! colMeans(.[]), .before = 1 )
T031 =  T031 %>% add_row(!!! colMeans(.[]), .before = 1 )
T032 =  T032 %>% add_row(!!! colMeans(.[]), .before = 1 )
T033 =  T033 %>% add_row(!!! colMeans(.[]), .before = 1 )
T034 =  T034 %>% add_row(!!! colMeans(.[]), .before = 1 )
T035 =  T035 %>% add_row(!!! colMeans(.[]), .before = 1 )
T036 =  T036 %>% add_row(!!! colMeans(.[]), .before = 1 )
T038 =  T038 %>% add_row(!!! colMeans(.[]), .before = 1 )
T039 =  T039 %>% add_row(!!! colMeans(.[]), .before = 1 )
T040 =  T040 %>% add_row(!!! colMeans(.[]), .before = 1 )
T041 =  T041 %>% add_row(!!! colMeans(.[]), .before = 1 )
T042 =  T042 %>% add_row(!!! colMeans(.[]), .before = 1 )
T043 =  T043 %>% add_row(!!! colMeans(.[]), .before = 1 )
T044 =  T044 %>% add_row(!!! colMeans(.[]), .before = 1 )
T045 =  T045 %>% add_row(!!! colMeans(.[]), .before = 1 )
T046 =  T046 %>% add_row(!!! colMeans(.[]), .before = 1 )
T047 =  T047 %>% add_row(!!! colMeans(.[]), .before = 1 )
T050 =  T050 %>% add_row(!!! colMeans(.[]), .before = 1 )
T051 =  T051 %>% add_row(!!! colMeans(.[]), .before = 1 )
T054 =  T054 %>% add_row(!!! colMeans(.[]), .before = 1 )
T055 =  T055 %>% add_row(!!! colMeans(.[]), .before = 1 )
T060 =  T060 %>% add_row(!!! colMeans(.[]), .before = 1 )
T061 =  T061 %>% add_row(!!! colMeans(.[]), .before = 1 )
T062 =  T062 %>% add_row(!!! colMeans(.[]), .before = 1 )
T064 =  T064 %>% add_row(!!! colMeans(.[]), .before = 1 )
T066 =  T066 %>% add_row(!!! colMeans(.[]), .before = 1 )
T068 =  T068 %>% add_row(!!! colMeans(.[]), .before = 1 )
T073 =  T073 %>% add_row(!!! colMeans(.[]), .before = 1 )
T074 =  T074 %>% add_row(!!! colMeans(.[]), .before = 1 )
T075 =  T075 %>% add_row(!!! colMeans(.[]), .before = 1 )
T076 =  T076 %>% add_row(!!! colMeans(.[]), .before = 1 )
T077 =  T077 %>% add_row(!!! colMeans(.[]), .before = 1 )
```

```
T079  =     T079  %>%  add_row(!!! colMeans(.[]), .before = 1 )
T080  =     T080  %>%  add_row(!!! colMeans(.[]), .before = 1 )
T081  =     T081  %>%  add_row(!!! colMeans(.[]), .before = 1 )
T082  =     T082  %>%  add_row(!!! colMeans(.[]), .before = 1 )
T083  =     T083  %>%  add_row(!!! colMeans(.[]), .before = 1 )
T084  =     T084  %>%  add_row(!!! colMeans(.[]), .before = 1 )
T086  =     T086  %>%  add_row(!!! colMeans(.[]), .before = 1 )
T088  =     T088  %>%  add_row(!!! colMeans(.[]), .before = 1 )

## keep only first observation in each dataset (row with all the means) ##

T001  =     T001  %>%  slice(1)
T002  =     T002  %>%  slice(1)
T003  =     T003  %>%  slice(1)
T004  =     T004  %>%  slice(1)
T005  =     T005  %>%  slice(1)
T006  =     T006  %>%  slice(1)
T007  =     T007  %>%  slice(1)
T008  =     T008  %>%  slice(1)
T009  =     T009  %>%  slice(1)
T010  =     T010  %>%  slice(1)
T011  =     T011  %>%  slice(1)
T012  =     T012  %>%  slice(1)
T013  =     T013  %>%  slice(1)
T014  =     T014  %>%  slice(1)
T015  =     T015  %>%  slice(1)
T016  =     T016  %>%  slice(1)
T017  =     T017  %>%  slice(1)
T018  =     T018  %>%  slice(1)
T019  =     T019  %>%  slice(1)
T020  =     T020  %>%  slice(1)
T021  =     T021  %>%  slice(1)
T022  =     T022  %>%  slice(1)
T023  =     T023  %>%  slice(1)
T024  =     T024  %>%  slice(1)
T025  =     T025  %>%  slice(1)
T026  =     T026  %>%  slice(1)
T027  =     T027  %>%  slice(1)
T028  =     T028  %>%  slice(1)
T029  =     T029  %>%  slice(1)
T031  =     T031  %>%  slice(1)
T032  =     T032  %>%  slice(1)
T033  =     T033  %>%  slice(1)
T034  =     T034  %>%  slice(1)
T035  =     T035  %>%  slice(1)
T036  =     T036  %>%  slice(1)
T038  =     T038  %>%  slice(1)
T039  =     T039  %>%  slice(1)
T040  =     T040  %>%  slice(1)
T041  =     T041  %>%  slice(1)
T042  =     T042  %>%  slice(1)
```

```
T043  =  T043  %>%  slice(1)
T044  =  T044  %>%  slice(1)
T045  =  T045  %>%  slice(1)
T046  =  T046  %>%  slice(1)
T047  =  T047  %>%  slice(1)
T050  =  T050  %>%  slice(1)
T051  =  T051  %>%  slice(1)
T054  =  T054  %>%  slice(1)
T055  =  T055  %>%  slice(1)
T060  =  T060  %>%  slice(1)
T061  =  T061  %>%  slice(1)
T062  =  T062  %>%  slice(1)
T064  =  T064  %>%  slice(1)
T066  =  T066  %>%  slice(1)
T068  =  T068  %>%  slice(1)
T073  =  T073  %>%  slice(1)
T074  =  T074  %>%  slice(1)
T075  =  T075  %>%  slice(1)
T076  =  T076  %>%  slice(1)
T077  =  T077  %>%  slice(1)
T079  =  T079  %>%  slice(1)
T080  =  T080  %>%  slice(1)
T081  =  T081  %>%  slice(1)
T082  =  T082  %>%  slice(1)
T083  =  T083  %>%  slice(1)
T084  =  T084  %>%  slice(1)
T086  =  T086  %>%  slice(1)
T088  =  T088  %>%  slice(1)

## merge all rows, now we have a dataset where each row represents a participant ##

meandata = rbind(T001,
                 T002,
                 T003,
                 T004,
                 T005,
                 T006,
                 T007,
                 T008,
                 T009,
                 T010,
                 T011,
                 T012,
                 T013,
                 T014,
                 T015,
                 T016,
                 T017,
                 T018,
                 T019,
                 T020,
```

```
T021,
T022,
T023,
T024,
T025,
T026,
T027,
T028,
T029,
T031,
T032,
T033,
T034,
T035,
T036,
T038,
T039,
T040,
T041,
T042,
T043,
T044,
T045,
T046,
T047,
T050,
T051,
T054,
T055,
T060,
T061,
T062,
T064,
T066,
T068,
T073,
T074,
T075,
T076,
T077,
T079,
T080,
T081,
T082,
T083,
T084,
T086,
T088)

#add gender, age and "group" now because we will omit participants in the next step#
```

```
meandata = meandata %>%  mutate(Gender = c("male",
                                           "female",
                                           "male",
                                           "female",
                                           "male",
                                           "female",
                                           "male",
                                           "female",
                                           "male",
                                           "female",
                                           "male",
                                           "female",
                                           "male",
                                           "female",
                                           "male",
                                           "female",
                                           "male",
                                           "female",
                                           "male",
                                           "female",
                                           "male",
                                           "female",
                                           "female",
                                           "male",
                                           "female",
                                           "male",
                                           "female",
                                           "female",
                                           "male",
                                           "female",
                                           "male",
                                           "female",
                                           "male",
                                           "male",
                                           "female",
                                           "male",
                                           "female",
                                           "male",
                                           "female",
                                           "male",
                                           "female",
                                           "male",
                                           "female",
                                           "male",
                                           "female",
                                           "male",
                                           "male",
                                           "male",
                                           "male",
                                           "female",
                                           "male",
```

```
                                    "female",
                                    "female",
                                    "female",
                                    "male",
                                    "female",
                                    "male",
                                    "female",
                                    "male",
                                    "female",
                                    "female",
                                    "female",
                                    "male",
                                    "male",
                                    "male",
                                    "male",
                                    "female",
                                    "male"))

meandata = meandata %>%  mutate(Group = c("    young  "         ,
                                    "   young  "         ,
                                    "   young  "         ,
                                    "   young  "         ,
                                    "   young  "         ,
                                    "   young  "         ,
                                    "   young  "         ,
                                    "   young  "         ,
                                    "   young  "         ,
                                    "   young  "         ,
                                    "   young  "         ,
                                    "   young  "         ,
                                    "   young  "         ,
                                    "   young  "         ,
                                    "   young  "         ,
                                    "   young  "         ,
                                    "   young  "         ,
                                    "   young  "         ,
                                    "   young  "         ,
                                    "   young  "         ,
                                    "   young  "         ,
                                    "   young  "         ,
                                    "   young  "         ,
                                    "   old    "         ,
                                    "   old    "         ,
                                    "   old    "         ,
                                    "   old    "         ,
                                    "   old    "         ,
                                    "   old    "         ,
                                    "   old    "         ,
                                    "   old    "         ,
                                    "   old    "         ,
                                    "   old    "         ,
```

```
                                  "   old     "       ,
                                  "   old     "       ,
                                  "   old     "       ,
                                  "   old     "       ,
                                  "   old     "       ,
                                  "   old     "       ,
                                  "   old     "       ,
                                  "   old     "       ,
                                  "   old     "       ,
                                  "   old     "       ,
                                  "   old     "       ,
                                  "   old     "       ,
                                  "   old     "       ,
                                  "   old     "       ,
                                  "   old     "       ,
                                  "   old     "       ,
                                  "   young   "       ,
                                  "   young   "       ,
                                  "   young   "       ,
                                  "   young   "       ,
                                  "   young   "       ,
                                  "   young   "       ,
                                  "   old     "       ,
                                  "   old     "       ,
                                  "   old     "       ,
                                  "   old     "       ,
                                  "   old     "       ,
                                  "   old     "       ,
                                  "   young   "       ,
                                  "   young   "       ,
                                  "   young   "       ,
                                  "   young   "       ,
                                  "   old     "       ,
                                  "   old     "       ,
                                  "   old     "       ,
                                  "   old     "))

meandata = meandata %>% mutate(Age = c(23      ,
                                  23    ,
                                  23    ,
                                  23    ,
                                  23    ,
                                  23    ,
                                  23    ,
                                  23    ,
                                  23    ,
                                  23    ,
                                  23    ,
                                  23    ,
                                  23    ,
                                  23    ,
                                  23    ,
```

23 ,
23 ,
23 ,
23 ,
23 ,
23 ,
23 ,
23 ,
23 ,
72 ,
65 ,
65 ,
84 ,
70 ,
70 ,
80 ,
68 ,
81 ,
65 ,
70 ,
81 ,
62 ,
72 ,
66 ,
65 ,
61 ,
70 ,
62 ,
73 ,
66 ,
73 ,
67 ,
62 ,
23 ,
22 ,
25 ,
22 ,
19 ,
25 ,
71 ,
63 ,
71 ,
75 ,
66 ,
70 ,
22 ,
22 ,
23 ,
18 ,
68 ,
78 ,

```
                    61    ,
                    86))

## remove T032 and T088 because no heart rate data is available ##

meandata = na.omit(meandata)

## add variable with participant number ##

meandata = meandata %>% mutate(Participant = c(1:66))

#reorder the variables for demographic data to be in the front#

meandata <- meandata %>% MoveFront(., c("Participant",
                                        "Gender",
                                        "Age",
                                        "Group",
                                        "Heart.Rate",
                                        "Breathing.Rate",
                                        "Speed",
                                        "Acceleration",
                                        "Brake",
                                        "Steering"))

## export new dataset as .csv file ##

write.csv(meandata, "meandata.csv", row.names = FALSE)

#make new dataframe without demographic data for outlier analysis because all variables
have to be numerical#
#and we do not want to exclude age, gender etc. #

meandata1 = meandata %>% select(5:10)

#find absolute value of z-score for each participant in each column
z_scores <- as.data.frame(sapply(meandata1, function(meandata1) (abs(meandata1-
mean(meandata1))/sd(meandata1))))

#only keep rows in dataframe with all z-scores less than absolute value of 3, then check which
participants were excluded#
no_outliers <- z_scores[!rowSums(z_scores>3), ]

#exclude 7 participants in the meandata dataframe based on the z-score analysis#
#this is because i want to keep going with the meandata dataset and not the no_outiers
dataset#

meandata <- meandata[-c(15,18,31,38,44,55,56), ]

#delete column participants and add new one because the participant number does not match
the observation number anymore#
```

```
meandata = meandata %>% select(-Participant)
meandata = meandata %>% mutate(Participant = c(1:59))

#order columns correctly#

meandata <- meandata %>% MoveFront(., c("Participant",
                                        "Gender",
                                        "Age",
                                        "Group",
                                        "Heart.Rate",
                                        "Breathing.Rate",
                                        "Speed",
                                        "Acceleration",
                                        "Brake",
                                        "Steering"))

write.csv(meandata, "meandata.csv", row.names = FALSE)


#Install packages
install.packages("factoextra")
install.packages("cluster")
install.packages("ggpubr")
#load packages
library(tidyverse)
library(factoextra)
library(cluster)
library(ggplot2)
library(ggpubr)
library(dplyr)

#Descriptive statistics of the participants#

meandata <- meandata %>% mutate(Gender = as.factor(Gender))
meandata <- meandata %>% mutate(Group = as.factor(Group))
meandata %>% select(2:4) %>% summary()
meandata %>% select(3) %>% map(sd)

#Descriptive statistics of the parameters#

meandata %>% summary()
sd(meandata$Speed)
sd(meandata$Acceleration)
sd(meandata$Brake)
sd(meandata$Steering)
sd(meandata$Heart.Rate)

#Normality Test of Parameters#

shapiro.test(meandata$Speed)
shapiro.test(meandata$Acceleration)
```

```
shapiro.test(meandata$Brake)
shapiro.test(meandata$Steering)
shapiro.test(meandata$Heart.Rate)

#Distribution plot for each variable#
#density curve#
ggplot(meandata, aes(x=Speed)) + geom_density()

#Histrogram with line for mean#
ggplot(meandata, aes(x=Speed)) +
  geom_histogram(binwidth=.5, colour="black", fill="white") +
  geom_vline(aes(xintercept=mean(Speed)), color="red", linetype="dashed", size=1)

#everything together (histrogram, line for mean and densitiy curve)
#Speed
ggplot(meandata, aes(x=Speed)) +
  geom_histogram(aes(y=..density..),binwidth=.5, colour="black", fill="white") +
  geom_vline(aes(xintercept=mean(Speed)), color="red", linetype="dashed", size=1) +
  geom_density(alpha=.2, fill="#FF6666")

#Acceleration
ggplot(meandata, aes(x=Acceleration)) +
  geom_histogram(aes(y=..density..),binwidth=.5, colour="black", fill="white") +
  geom_vline(aes(xintercept=mean(Acceleration)), color="red", linetype="dashed", size=1) +
  geom_density(alpha=.2, fill="#FF6666")

ggplot(meandata, aes(x=Acceleration)) + geom_histogram(binwidth=.5, colour="black",
fill="white")

#Brake
ggplot(meandata, aes(x=Brake)) +
  geom_histogram(aes(y=..density..),binwidth=.5, colour="black", fill="white") +
  geom_vline(aes(xintercept=mean(Brake)), color="red", linetype="dashed", size=1) +
  geom_density(alpha=.2, fill="#FF6666")

#Steering
ggplot(meandata, aes(x=Steering)) +
  geom_histogram(aes(y=..density..),binwidth=.01, colour="black", fill="white") +
  geom_vline(aes(xintercept=mean(Steering)), color="red", linetype="dashed", size=1) +
  geom_density(alpha=.2, fill="#FF6666")
  #only histogram
ggplot(meandata, aes(x=Steering)) + geom_histogram(binwidth=.5, colour="black",
fill="white")
ggplot(meandata, aes(x=Steering)) + geom_density() +
  geom_vline(aes(xintercept=mean(Steering)), color="red", linetype="dashed", size=1)

#Heartrate
ggplot(meandata, aes(x=Heart.Rate)) +
  geom_histogram(aes(y=..density..),binwidth=.5, colour="black", fill="white") +
  geom_vline(aes(xintercept=mean(Heart.Rate)), color="red", linetype="dashed", size=1) +
  geom_density(alpha=.2, fill="#FF6666")
```

```
#Beginning k-means#
#make datasets for each possible variable pair#

SpeedAcceleration = meandata %>%  select(7,8)
SpeedBrake = meandata %>% select(7,9)
SpeedSteering = meandata %>% select(7,10)
SpeedHeartrate = meandata %>% select(7,5)
AccelerationBrake = meandata %>%  select (8,9)
AccelerationSteering = meandata %>% select(8,10)
AccelerationHeartrate = meandata %>% select(8,5)
HeartrateBrake = meandata %>%  select(5,9)
HeartrateSteering = meandata %>% select (5, 10)
BrakeSteering = meandata  %>% select(9,10)

#scale all datasets to have a mean of 0 and SD of 1 -> Z-Score#

SpeedAcceleration = scale(SpeedAcceleration)
SpeedBrake = scale(SpeedBrake)
SpeedSteering = scale(SpeedSteering)
SpeedHeartrate = scale(SpeedHeartrate)
AccelerationBrake = scale(AccelerationBrake)
AccelerationSteering = scale(AccelerationSteering)
AccelerationHeartrate = scale(AccelerationHeartrate)
HeartrateBrake = scale(HeartrateBrake)
HeartrateSteering = scale(HeartrateSteering)
BrakeSteering = scale(BrakeSteering)

#find out optimal amount of clusters#
#elbow method#

fviz_nbclust(SpeedAcceleration, kmeans, method = "wss") + xlab("Number of clusters:
SpeedAcceleration")
fviz_nbclust(SpeedBrake, kmeans, method = "wss") + xlab("Number of clusters:
SpeedBrake")
fviz_nbclust(SpeedSteering, kmeans, method = "wss") + xlab("Number of clusters:
SpeedSteering")
fviz_nbclust(SpeedHeartrate, kmeans, method = "wss") + xlab("Number of clusters:
SpeedHeartrate")
fviz_nbclust(AccelerationBrake, kmeans, method = "wss") + xlab("Number of clusters:
AccelerationBrake")
fviz_nbclust(AccelerationSteering, kmeans, method = "wss") + xlab("Number of clusters:
AccelerationSteering")
fviz_nbclust(AccelerationHeartrate, kmeans, method = "wss") + xlab("Number of clusters:
AccelerationHeartrate")
fviz_nbclust(HeartrateBrake, kmeans, method = "wss") + xlab("Number of clusters:
HeartrateBrake")
fviz_nbclust(HeartrateSteering, kmeans, method = "wss") + xlab("Number of clusters:
HeartrateSteering")
fviz_nbclust(BrakeSteering, kmeans, method = "wss") + xlab("Number of clusters:
BrakeSteering")
```

```
#make this example reproducible
set.seed(250)

##SpAcc perform k-means clustering with k = 3 clusters##
kmSpAcc <- kmeans(SpeedAcceleration, centers = 3, nstart = 100)
#view results
kmSpAcc
#plot results of final k-means model
fviz_cluster(kmSpAcc, data = SpeedAcceleration)

##SpB perform k-means clustering with k = 3 clusters##
kmSpB <- kmeans(SpeedBrake, centers = 3, nstart = 100)
#view results
kmSpB
#plot results of final k-means model
fviz_cluster(kmSpB, data = SpeedBrake)

##SpSt perform k-means clustering with k = 3 clusters##
kmSpSt <- kmeans(SpeedSteering, centers = 3, nstart = 100)
#view results
kmSpSt
#plot results of final k-means model
fviz_cluster(kmSpSt, data = SpeedSteering)

##SpHR perform k-means clustering with k = 3 clusters##
kmSpHR <- kmeans(SpeedHeartrate, centers = 3, nstart = 100)
#view results
kmSpHR
#plot results of final k-means model
fviz_cluster(kmSpHR, data = SpeedHeartrate)

##AccB perform k-means clustering with k = 3 clusters##
kmAccB <- kmeans(AccelerationBrake, centers = 3, nstart = 100)
#view results
kmAccB
#plot results of final k-means model
fviz_cluster(kmAccB, data = AccelerationBrake)

##AccSt perform k-means clustering with k = 3 clusters##
kmAccSt <- kmeans(AccelerationSteering, centers = 3, nstart = 100)
#view results
kmAccSt
#plot results of final k-means model
fviz_cluster(kmAccSt, data = AccelerationSteering)

##AccHR perform k-means clustering with k = 3 clusters##
kmAccHR <- kmeans(AccelerationHeartrate, centers = 3, nstart = 100)
#view results
kmAccHR
#plot results of final k-means model
```

```
fviz_cluster(kmAccHR, data = AccelerationHeartrate)

##HRB perform k-means clustering with k = 3 clusters##
kmHRB <- kmeans(HeartrateBrake, centers = 3, nstart = 100)
#view results
kmHRB
#plot results of final k-means model
fviz_cluster(kmHRB, data = HeartrateBrake)

##HRSt perform k-means clustering with k = 3 clusters##
kmHRSt <- kmeans(HeartrateSteering, centers = 3, nstart = 100)
#view results
kmHRSt
#plot results of final k-means model
fviz_cluster(kmHRSt, data = HeartrateSteering)

##BSt perform k-means clustering with k = 3 clusters##
kmBSt <- kmeans(BrakeSteering, centers = 3, nstart = 100)
#view results
kmBSt
#plot results of final k-means model
fviz_cluster(kmBSt, data = BrakeSteering)

#only MEAN, no z-score#

install.packages("plotrix")
library("plotrix")

SpeedAcceleration = meandata %>%  select(7,8)
SpeedBrake = meandata %>% select(7,9)
SpeedSteering = meandata %>% select(7,10)
SpeedHeartrate = meandata %>% select(7,5)
AccelerationBrake = meandata %>%  select (8,9)
AccelerationSteering = meandata %>% select(8,10)
AccelerationHeartrate = meandata %>% select(8,5)
HeartrateBrake = meandata %>%  select(5,9)
HeartrateSteering = meandata %>% select (5, 10)
BrakeSteering = meandata  %>% select(9,10)

#find out optimal amount of clusters#
#elbow method#

fviz_nbclust(SpeedAcceleration, kmeans, method = "wss") + xlab("Number of clusters:
SpeedAcceleration")
fviz_nbclust(SpeedBrake, kmeans, method = "wss") + xlab("Number of clusters:
SpeedBrake")
fviz_nbclust(SpeedSteering, kmeans, method = "wss") + xlab("Number of clusters:
SpeedSteering")
fviz_nbclust(SpeedHeartrate, kmeans, method = "wss") + xlab("Number of clusters:
SpeedHeartrate")
```

```
fviz_nbclust(AccelerationBrake, kmeans, method = "wss") + xlab("Number of clusters:
AccelerationBrake")
fviz_nbclust(AccelerationSteering, kmeans, method = "wss") + xlab("Number of clusters:
AccelerationSteering")
fviz_nbclust(AccelerationHeartrate, kmeans, method = "wss") + xlab("Number of clusters:
AccelerationHeartrate")
fviz_nbclust(HeartrateBrake, kmeans, method = "wss") + xlab("Number of clusters:
HeartrateBrake")
fviz_nbclust(HeartrateSteering, kmeans, method = "wss") + xlab("Number of clusters:
HeartrateSteering")
fviz_nbclust(BrakeSteering, kmeans, method = "wss") + xlab("Number of clusters:
BrakeSteering")

#make this example reproducible
set.seed(250)

##SpAcc perform k-means clustering with k = 3 clusters##
kmSpAcc <- kmeans(SpeedAcceleration, centers = 3, nstart = 100)
#view results
kmSpAcc
#plot results of final k-means model
fviz_cluster(kmSpAcc, data = SpeedAcceleration)

##SpB perform k-means clustering with k = 3 clusters##
kmSpB <- kmeans(SpeedBrake, centers = 3, nstart = 100)
#view results
kmSpB
#plot results of final k-means model
fviz_cluster(kmSpB, data = SpeedBrake)

##SpSt perform k-means clustering with k = 3 clusters##
kmSpSt <- kmeans(SpeedSteering, centers = 3, nstart = 100)
#view results
kmSpSt
#plot results of final k-means model
fviz_cluster(kmSpSt, data = SpeedSteering)

##SpHR perform k-means clustering with k = 3 clusters##
kmSpHR <- kmeans(SpeedHeartrate, centers = 3, nstart = 100)
#view results
kmSpHR
#plot results of final k-means model
fviz_cluster(kmSpHR, data = SpeedHeartrate)

##AccB perform k-means clustering with k = 3 clusters##
kmAccB <- kmeans(AccelerationBrake, centers = 3, nstart = 100)
#view results
kmAccB
#plot results of final k-means model
fviz_cluster(kmAccB, data = AccelerationBrake)
```

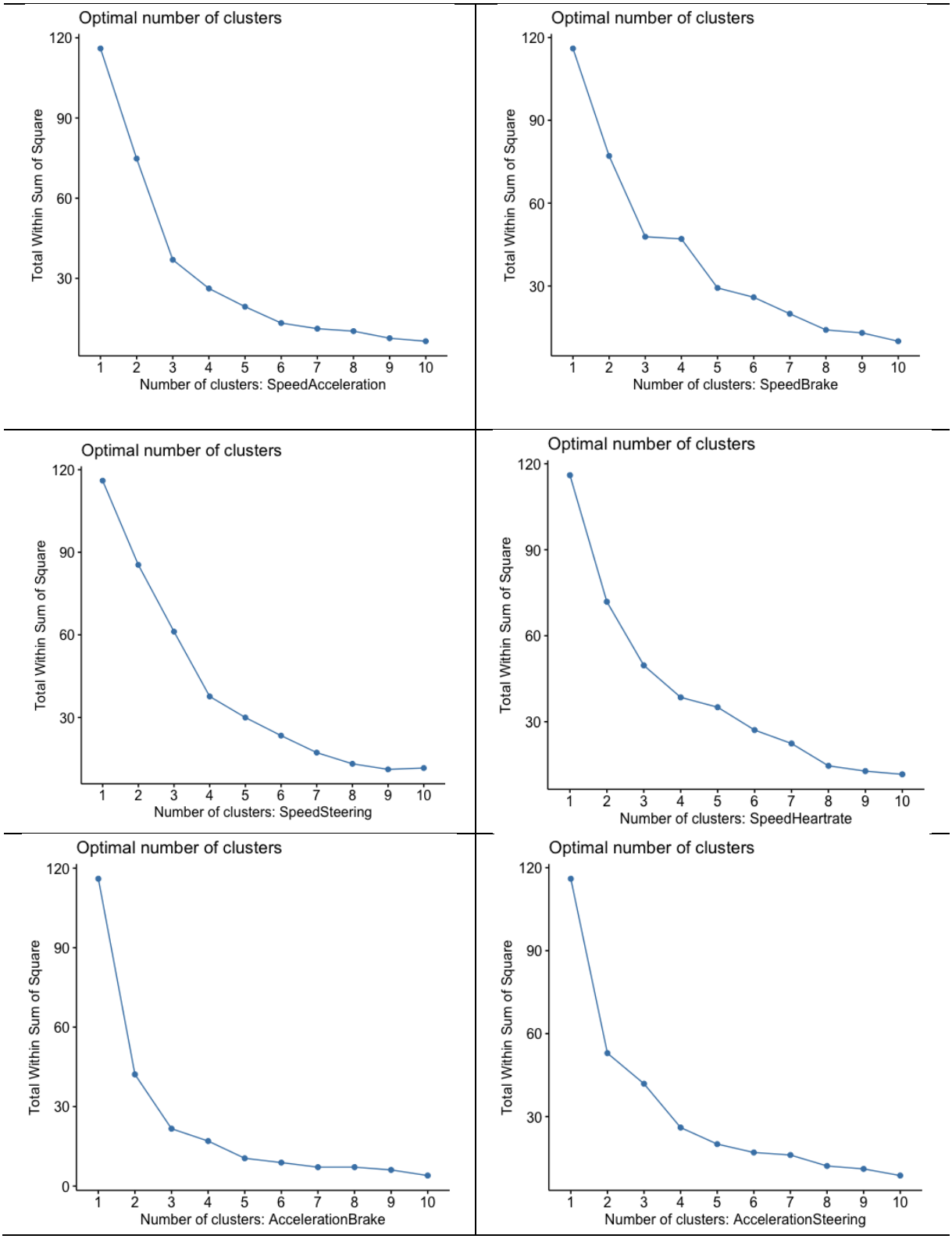
```
##AccSt perform k-means clustering with k = 3 clusters##
kmAccSt <- kmeans(AccelerationSteering, centers = 3, nstart = 100)
#view results
kmAccSt
#plot results of final k-means model
fviz_cluster(kmAccSt, data = AccelerationSteering)

##AccHR perform k-means clustering with k = 3 clusters##
kmAccHR <- kmeans(AccelerationHeartrate, centers = 3, nstart = 100)
#view results
kmAccHR
#plot results of final k-means model
fviz_cluster(kmAccHR, data = AccelerationHeartrate)

##HRB perform k-means clustering with k = 3 clusters##
kmHRB <- kmeans(HeartrateBrake, centers = 3, nstart = 100)
#view results
kmHRB
#plot results of final k-means model
fviz_cluster(kmHRB, data = HeartrateBrake)

##HRSt perform k-means clustering with k = 3 clusters##
kmHRSt <- kmeans(HeartrateSteering, centers = 3, nstart = 100)
#view results
kmHRSt
#plot results of final k-means model
fviz_cluster(kmHRSt, data = HeartrateSteering)

##BSt perform k-means clustering with k = 3 clusters##
kmBSt <- kmeans(BrakeSteering, centers = 3, nstart = 100)
#view results
kmBSt
#plot results of final k-means model
fviz_cluster(kmBSt, data = BrakeSteering)

#Rand.Index

install.packages("fossil")
library(fossil)

#copy the cluster vector from each k_means model and make it into it's own vector #
#fuction to get cluster vector#
kmSpAcc$cluster
#do this for every k-means model#

cluster_SpB = c(3, 3, 1, 3, 2, 1, 1, 1, 1, 1, 2, 3, 1, 1, 1, 1, 2, 1, 2, 2, 1, 1, 1, 1, 3, 1, 1, 3, 1, 1, 1,
3, 1, 3, 1, 1, 3, 1, 3, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 1, 2, 2, 2, 2)
cluster_SpAcc = c(2, 2, 3, 2, 3, 3, 3, 3, 3, 3, 2, 2, 3, 3, 3, 3, 2, 3, 2, 2, 3, 3, 3, 3, 2, 3, 3, 2, 3, 3,
3, 2, 3, 2, 3, 3, 2, 3, 2, 3, 3, 2, 3, 1, 3, 1, 1, 1, 1, 1, 2, 1, 1, 2, 1, 1, 1, 1, 1)
cluster_AccSt = c(1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 3, 1, 3, 3, 3, 3, 2, 2, 3, 3, 3, 3, 3, 3, 3, 2)
```

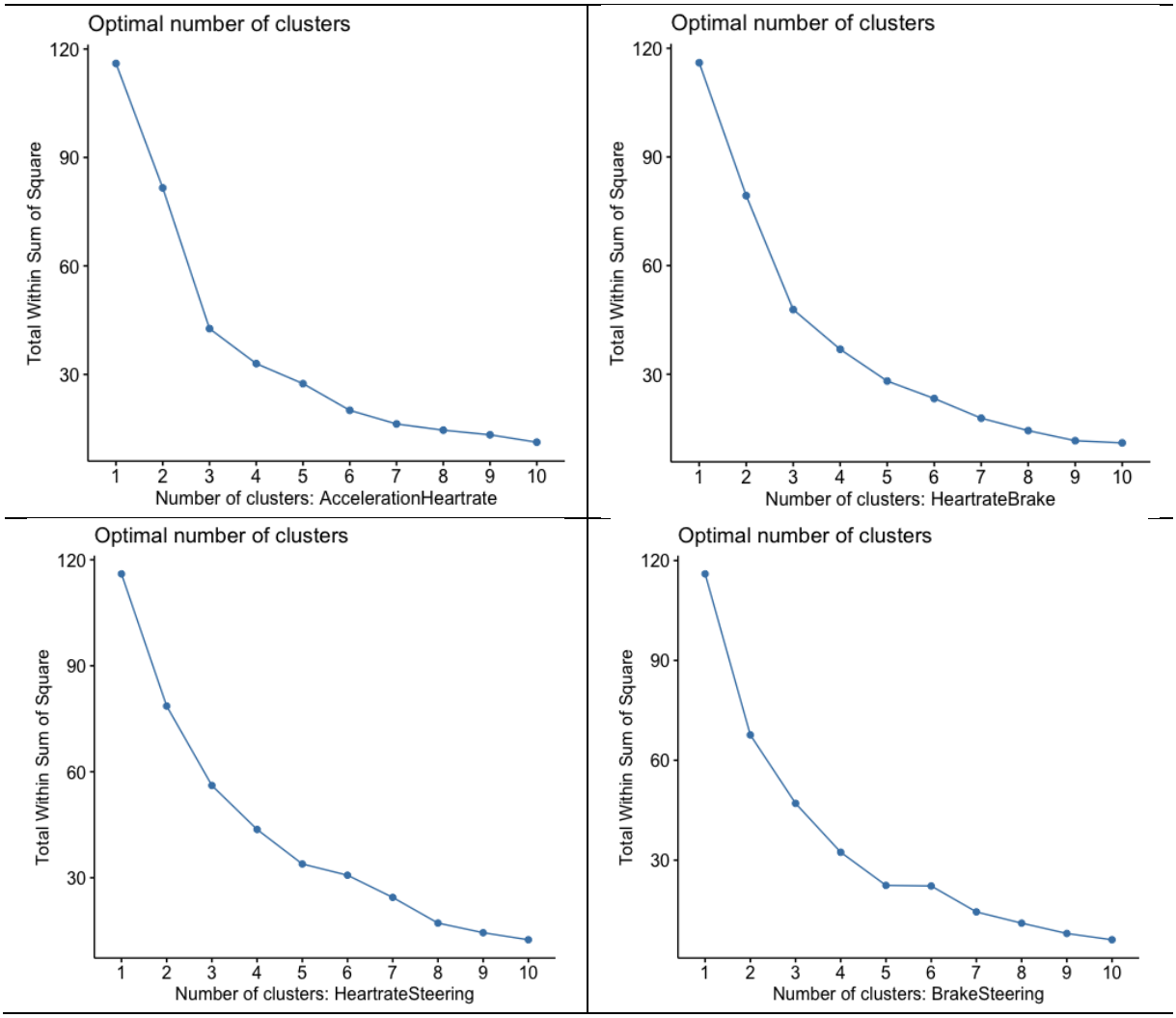
```
cluster_AccB = c(3, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 3, 1, 1, 1, 1, 1, 1, 1, 1, 1, 3, 3, 3, 3, 1, 3, 1, 1, 3,
3, 3, 3, 3, 1, 3, 3, 1, 1, 3, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2)
cluster_AccHR = c(3, 2, 2, 2, 3, 2, 2, 2, 2, 3, 3, 3, 2, 2, 3, 2, 3, 2, 3, 2, 3, 2, 3, 3, 3, 3, 3, 3, 3, 2,
2, 3, 2, 3, 2, 2, 2, 3, 3, 2, 3, 3, 3, 1, 2, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1)
cluster_BSt = c(1, 2, 1, 1, 2, 2, 2, 2, 2, 2, 2, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 1, 1, 1, 1, 2, 1, 1, 2, 1, 1,
1, 1, 1, 2, 1, 1, 2, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 2, 2, 2, 2, 2, 2, 3, 3)
cluster_HRB = c(1, 2, 2, 2, 1, 3, 3, 3, 3, 1, 1, 1, 3, 3, 1, 3, 1, 3, 1, 1, 1, 2, 2, 2, 2, 1, 2, 1, 1, 2,
2, 2, 2, 2, 1, 2, 2, 1, 1, 2, 1, 1, 1, 1, 3, 3, 1, 1, 1, 1, 3, 3, 3, 1, 3, 3, 1, 1, 1)
cluster_HRSt = c(3, 2, 2, 2, 3, 2, 2, 2, 2, 3, 3, 3, 2, 2, 3, 2, 3, 2, 3, 2, 3, 2, 3, 3, 3, 3, 3, 3, 3, 2,
2, 3, 2, 3, 2, 2, 2, 3, 3, 2, 3, 3, 3, 3, 2, 2, 3, 3, 3, 1, 1, 2, 2, 3, 2, 2, 3, 1, 1)
cluster_SpHR = c(2, 3, 3, 2, 1, 3, 3, 3, 3, 1, 1, 2, 3, 3, 1, 3, 1, 3, 1, 3, 1, 3, 1, 1, 2, 1, 1, 2, 1, 3,
3, 2, 3, 2, 1, 3, 3, 1, 2, 3, 1, 2, 1, 1, 3, 3, 1, 2, 1, 2, 3, 3, 3, 2, 3, 3, 1, 1, 2)
cluster_SpSt = c(3, 3, 2, 3, 2, 2, 2, 2, 2, 2, 3, 3, 2, 2, 2, 2, 3, 2, 2, 3, 2, 2, 2, 2, 3, 2, 2, 3, 2, 2, 2,
3, 2, 3, 2, 2, 3, 2, 3, 2, 2, 3, 2, 3, 2, 2, 2, 3, 2, 1, 1, 3, 2, 3, 2, 3, 2, 2, 1)

#apply rand index function
rand.index(cluster_SpAcc, cluster_SpAcc)
rand.index(cluster_SpB, cluster_SpAcc)
rand.index(cluster_SpSt, cluster_SpAcc)
rand.index(cluster_SpHR, cluster_SpAcc)
rand.index(cluster_AccB, cluster_SpAcc)
rand.index(cluster_AccSt, cluster_SpAcc)
rand.index(cluster_AccHR, cluster_SpAcc)
rand.index(cluster_HRB, cluster_SpAcc)
rand.index(cluster_HRSt, cluster_SpAcc)
rand.index(cluster_BSt, cluster_SpAcc)
#
rand.index(cluster_SpAcc,    cluster_SpB)
rand.index(cluster_SpB,      cluster_SpB)
rand.index(cluster_SpSt, cluster_SpB)
rand.index(cluster_SpHR, cluster_SpB)
rand.index(cluster_AccB,     cluster_SpB)
rand.index(cluster_AccSt,    cluster_SpB)
rand.index(cluster_AccHR,    cluster_SpB)
rand.index(cluster_HRB,      cluster_SpB)
rand.index(cluster_HRSt,     cluster_SpB)
rand.index(cluster_BSt,      cluster_SpB)
#
rand.index(cluster_SpAcc,    cluster_SpSt)
rand.index(cluster_SpB,      cluster_SpSt)
rand.index(cluster_SpSt, cluster_SpSt)
rand.index(cluster_SpHR, cluster_SpSt)
rand.index(cluster_AccB,     cluster_SpSt)
rand.index(cluster_AccSt,    cluster_SpSt)
rand.index(cluster_AccHR,    cluster_SpSt)
rand.index(cluster_HRB,      cluster_SpSt)
rand.index(cluster_HRSt,     cluster_SpSt)
rand.index(cluster_BSt,      cluster_SpSt)
#
rand.index(cluster_SpAcc,    cluster_SpHR)
rand.index(cluster_SpB,      cluster_SpHR)
```

```
rand.index(cluster_SpSt, cluster_SpHR)
rand.index(cluster_SpHR, cluster_SpHR)
rand.index(cluster_AccB,    cluster_SpHR)
rand.index(cluster_AccSt,   cluster_SpHR)
rand.index(cluster_AccHR,   cluster_SpHR)
rand.index(cluster_HRB,     cluster_SpHR)
rand.index(cluster_HRSt,    cluster_SpHR)
rand.index(cluster_BSt,     cluster_SpHR)
#
rand.index(cluster_SpAcc,   cluster_AccB)
rand.index(cluster_SpB,     cluster_AccB)
rand.index(cluster_SpSt, cluster_AccB)
rand.index(cluster_SpHR, cluster_AccB)
rand.index(cluster_AccB,    cluster_AccB)
rand.index(cluster_AccSt,   cluster_AccB)
rand.index(cluster_AccHR,   cluster_AccB)
rand.index(cluster_HRB,     cluster_AccB)
rand.index(cluster_HRSt,    cluster_AccB)
rand.index(cluster_BSt,     cluster_AccB)
#
rand.index(cluster_SpAcc,   cluster_BSt)
rand.index(cluster_SpB,     cluster_BSt)
rand.index(cluster_SpSt, cluster_BSt)
rand.index(cluster_SpHR, cluster_BSt)
rand.index(cluster_AccB,    cluster_BSt)
rand.index(cluster_AccSt,   cluster_BSt)
rand.index(cluster_AccHR,   cluster_BSt)
rand.index(cluster_HRB,     cluster_BSt)
rand.index(cluster_HRSt,    cluster_BSt)
rand.index(cluster_BSt,     cluster_BSt)
```

Appendix B

Elbow Method Graphs for Each Variable Pair




Optimal number of clusters
Total Within Sum of Square
120
90
60
30
1 2 3 4 5 6 7 8 9 10
Number of clusters: AccelerationHeartrate
Optimal number of clusters
Total Within Sum of Square
120
90
60
30
1 2 3 4 5 6 7 8 9 10
Number of clusters: HeartrateBrake
Optimal number of clusters
Total Within Sum of Square
120
90
60
30
1 2 3 4 5 6 7 8 9 10
Number of clusters: HeartrateSteering
Optimal number of clusters
Total Within Sum of Square
120
90
60
30
1 2 3 4 5 6 7 8 9 10
Number of clusters: BrakeSteering