MA4829 Group Report

**Names and Matric Number:**

1. Maximilian Chan Jun Feng (U1921366K)

2. Teo Ren Hao (U1920468H)

3. Vignesh Sujith Menon (U1921506E)

4. Dylan Yeo Zhi Wei (U1922111H)

5. Josiah Lee Rong Guang (U1922674J)

6. Anthony Lee Wen Hao (U1921043F)

**Proposed Title:** Exploring correlations and trends between different variables presented by a car survey

**Abstract:**

This project proposes to draw correlations and trends between the different variables and types of data that has been collected by a survey. We first perform some data cleaning on the current dataset that is provided to us such that we can utilise it to do our analysis. Next, we move on to perform data mining on the processed dataset, specifically on selected categories of concern that we have identified: Gender, Marital Status, Ownership, Exterior Components, Interior Components and Personalisation. In this segment, we adopt a few data mining techniques such as **Principle Components Analysis, Rule Association Mining, K-means Clustering** etc. in order to draw out relations between the categories that we have identified. After which, we analyse the results to determine if there are any correlations between the categories, before linking it back to answer and establish our proposed idea.

Table of Contents

[1. Introduction 3](#_Toc127396838)

[2. Analysis and Results 6](#_Toc127396839)

[2.1 Data Exploration 6](#_Toc127396840)

[2.2 Data Mining 12](#_Toc127396841)

[2.2.1 Principle Components Analysis (PCA) / Multiple Correspondence Analysis (MCA) 12](#_Toc127396842)

[2.2.2 Associate Rule Mining 19](#_Toc127396843)

[3. Conclusion and Discussion 24](#_Toc127396844)

[4. References 24](#_Toc127396845)

Table of Figures

[Figure 1: Importing and Creation of Data Frame 4](#_Toc127396846)

[Figure 2: Dataframe descriptions 5](#_Toc127396847)

[Figure 3: Conversion of data types to string 5](#_Toc127396848)

[Figure 4: Splitting of string values 6](#_Toc127396849)

[Figure 5: Pie chart indicating the interest of survey respondents in designing their own components for car customization 7](#_Toc127396850)

[Figure 6: Pie chart showing the spending considerations of survey respondents for car customization 8](#_Toc127396851)

[Figure 7: Code for Male Marital Status vs Car Ownership 8](#_Toc127396852)

[Figure 8: Code for Female Marital Status vs Car Ownership 8](#_Toc127396853)

[Figure 9: Code for Bar plot for Gender vs Marital Status vs Car Ownership 9](#_Toc127396854)

[Figure 10: Bar Plot for Gender vs Marital Status vs Car Ownership 9](#_Toc127396855)

[Figure 11: Code for Age group vs Car Ownership 10](#_Toc127396856)

[Figure 12: Bar plot for Age group vs Car Ownership 10](#_Toc127396857)

[Figure 13: Bar Plot for Gender vs Age Group vs Car Ownership 11](#_Toc127396858)

[Figure 14: Code Exterior modifications vs Car Ownership for Males 12](#_Toc127396859)

[Figure 15: Code for Exterior modifications vs Car Ownership for Females 12](#_Toc127396860)

[Figure 16: Bar plot for Gender vs Exterior modification vs Car Ownership 12](#_Toc127396861)

[Figure 17: Estimating the Variances of all 7 components/columns 14](#_Toc127396862)

[Figure 18: Selecting 3 components for MCA from the dataset 14](#_Toc127396863)

[Figure 19: “Variances” of the Components for the first 7 entries in the dataset 15](#_Toc127396864)

[Figure 20: Graphical Representation of Component Analysis 15](#_Toc127396865)

[Figure 21: “explained\_intertia\_” values of the components 16](#_Toc127396866)

[Figure 22: The process of calculating explained\_intertia\_ 16](#_Toc127396867)

[Figure 23: Choosing the number of clusters for the MCA dataset 16](#_Toc127396868)

[Figure 24: Plotting of the WCSS against the number of components on a graph 17](#_Toc127396869)

[Figure 25: Choosing the number of clusters 17](#_Toc127396870)

[Figure 26: Concatenated data frame 18](#_Toc127396871)

[Figure 27: Mapping of segments to values 18](#_Toc127396872)

[Figure 28: Visualizing of clusters 19](#_Toc127396873)

[Figure 29: Dataframe containing binary columns 21](#_Toc127396874)

[Figure 30: Top 10 rules using Support metrics 22](#_Toc127396875)

[Figure 31: Top 10 rules using Confidence metrics 23](#_Toc127396876)

[Figure 32: Top 10 rules using Lift metrics 24](#_Toc127396877)

# 1. Introduction

Automobiles have come a long way since the early inventions of its kind in the 18th and 19th centuries. Gone are the days where these vehicles were rare and difficult to procure. Today, these vehicles are common and accessible, and we see them roaming about our streets every day. However, the interesting question lies in the gender behind the operation of such vehicles. Is there a gender disparity between the ownership and the opinions of such cars. Thus, this project aims to draw and identify correlations between gender and that of the ownership and personalisation of a car.

We first perform some **Data Cleaning** on the dataset that is required for data mining, by importing the Pandas library for data manipulation and to importing our dataset. We observe that there are 50 rows and 13 columns of data in our data frame by utilising the .shape method.



Figure : Importing and Creation of Data Frame

For the purpose of our project, we must first identify relevant columns of data and drop those that are irrelevant. We have identified said relevant columns to be Age Group, Gender, Ownership, Marital Status, Exterior, Interior and Personalisation. Consequently, we utilise the .drop method to remove the irrelevant columns and renamed our columns into the previously mentioned names. This results in a 50 row and 7 column data frame.



Figure : Dataframe descriptions

To further process our data, we require the data types of our each column to be a string. Using the .dtypes method, we can observe that our columns are all classified as an “object”. To change them into “strings” we adopted the use of the .astype method. For a more coherent dataset, we adopted the use of the .fillna method to replace any null values with the string value “NA”.

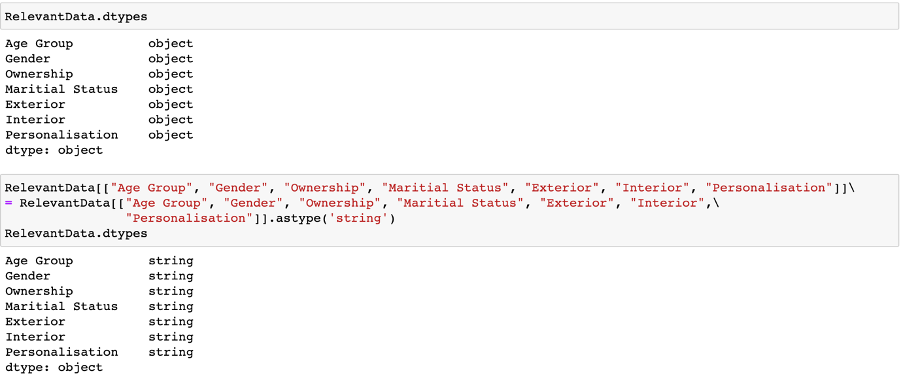


Figure : Conversion of data types to string

Finally, as the “Ownership”, “Exterior” and “Interior” columns consist of strings that are meant to represent multiple string values (as denoted by the comma), it would be convenient for our data mining that we split the strings in these columns to isolate the individual string values, i.e. “steering wheel” and “dashboard” for “Interior” instead of “steering wheel, dashboard”. This is achieved through the .str.split method and successful splits are presented in the same column as an array of string values. Our final data frame is as shown below.

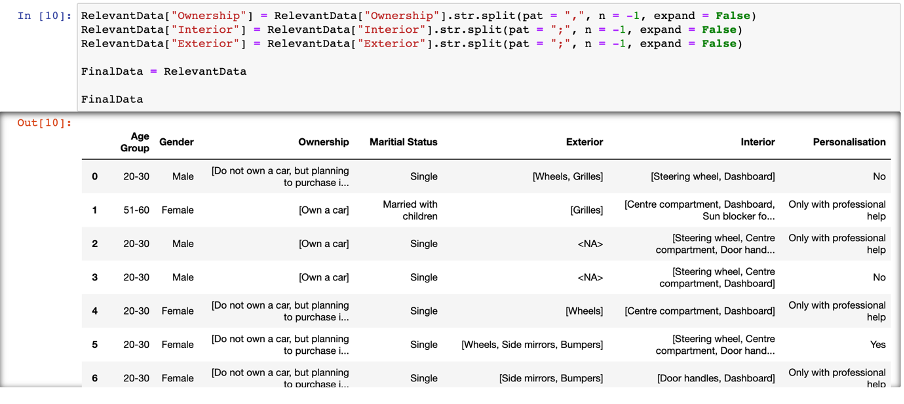


Figure : Splitting of string values

# 2. Analysis and Results

## 2.1 Data Exploration

After cleaning up the data, we move on to analyse the data and understand what insights the dataset is able to provide us. We first perform some simple **Data Exploration** to draw and identify some basic facts and areas or patterns of the data that we can dive deeper into.

We sieve through the survey results with the idea of coming up with potential business decisions that is both effective in bringing in more customers and cost-efficient for the company.

The survey results found that 96% of respondents were either “Likely” or “Very Likely” to opt for a customised vehicle if there were no extra charges involved. More than half (27) of the individuals expressed interest in designing components for the personalisation of their car, only with the enlistment of professional help [Figure 5]. Furthermore, 21 individuals indicated that they would require the services of a designer to model a sketch for the customisation of their car. Naturally, the procurement of customisation parts and the hiring of a designer would mean that the company incurs additional costs. So begs the question, is it in the company’s best interests to offer a package to allow the customer to personally customize their car?

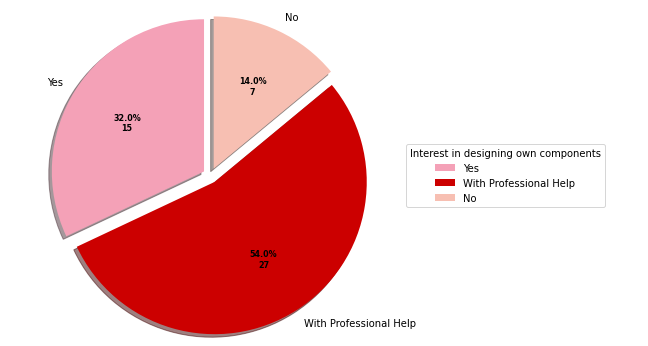


Figure : Pie chart indicating the interest of survey respondents in designing their own components for car customization

The same survey found that 90% of respondents were willing to spend on car customisation, even if surcharges were applicable. Further analysis determined that 9 of these respondents would increase their budget for car customization if personalised design options were available [Figure 6]. This number was 50% greater than the number of individuals who indicated that their spending would decrease upon the availability of the option. Therefore, we may arrive at the conclusion that the inclusion of such a package to allow for a personalised car customisation experience would be in the best interests of both the company and the consumer.

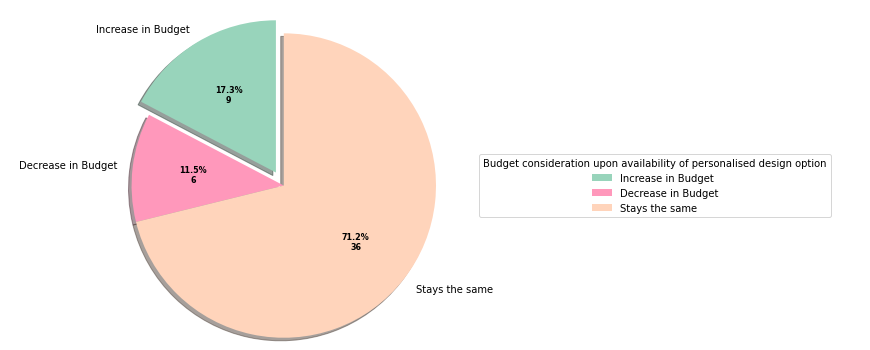


Figure : Pie chart showing the spending considerations of survey respondents for car customization

Apart from making use of pie charts, we also make use of bar plots to identify the trend between various variables. After compiling the cleaned data into a new data frame, we will need to include limitations in our search by using the .loc( ) function. We will use the “Male Marital Status and Car Ownership” variable as an example.



Figure : Code for Male Marital Status vs Car Ownership

In the first line, we include the condition whereby the male owner has to be single, owns a car or owns more than one car. The second and third line will be similar apart from the marital status of the car owner, “Married with children” and “Married with no children” respectively.

After which, the same has to be done for the females as shown below.



Figure : Code for Female Marital Status vs Car Ownership

Afterwards, we will combine all the new data frames and display them in a bar plot as shown below. The main idea behind this is to determine the length of the data frames as the length of the data frames represents the number of people within the survey who fulfils the given conditions.

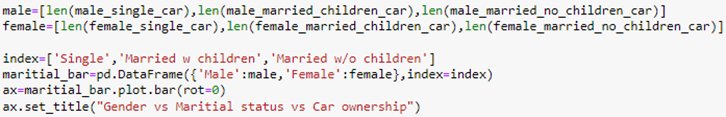


Figure : Code for Bar plot for Gender vs Marital Status vs Car Ownership

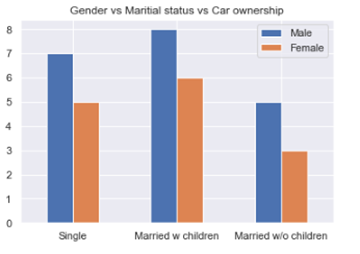


Figure : Bar Plot for Gender vs Marital Status vs Car Ownership

Based on the above we are able to identify that in general, more males have car ownership as compared to females. We can also identify that there is a greater demand for cars by males and females who are married with children, followed by singles and lastly, those that are married with no children. This trend could be due to a few reasons, for example, there is a greater need for a car when couples have children as this would significantly improve the couple’s mobility. In addition, the combined income of 2 people allows the couple to have greater purchasing power to purchase a car.[1]

Apart from marital status, we decided to classify the data based on age group and car ownership. Similarly, we make use of the .loc() function to create new data frames and display them in a bar plot.

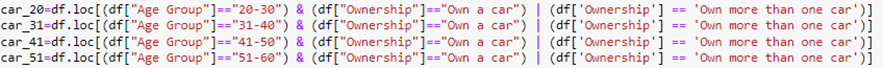


Figure : Code for Age group vs Car Ownership

The following results are produced.

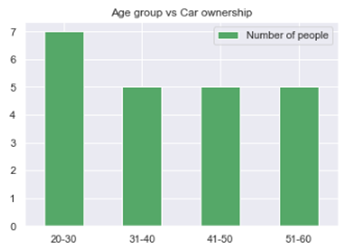


Figure : Bar plot for Age group vs Car Ownership

From the above bar plot, we can identify that those within the age group of 20-30 have higher car ownership compared to the rest of the different age groups.

We proceeded to determine the trend between gender, age group and car ownership and derived the following bar plot.

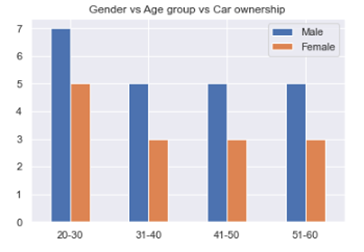


Figure : Bar Plot for Gender vs Age Group vs Car Ownership

From the above diagram, we can identify that those in the 20-30 age group have higher car ownership for both males and females as compared to those in different age groups. Another trend that we can observe is that males in general have higher car ownership than females across the different age groups.

Lastly, we proceeded to analyze which gender group are more likely to modify which exterior parts of the car. For this, we have to make use of the for loop to run through every single row in the main data frame.

From there, we will input the conditions using if statements for both gender and the exterior components being modified. For the exterior components, we will determine if the specific key words are found within the row. If all the conditions are being met, we will add a count of +1 to the relevant counts (i.e. male\_wheels=0, female\_wheels=0).

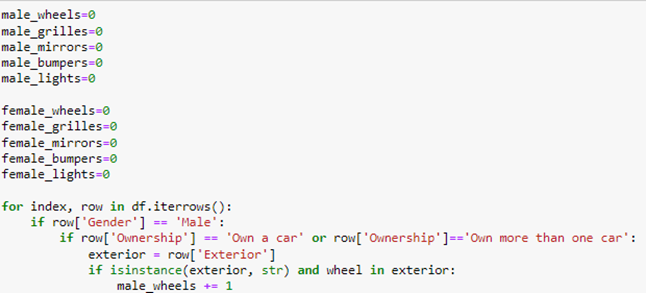


Figure : Code Exterior modifications vs Car Ownership for Males

After which, we will display all the data in a bar plot as shown below.

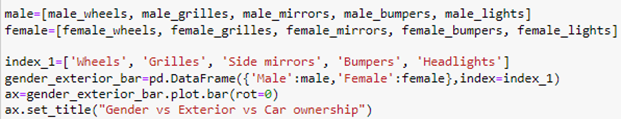


Figure : Code for Exterior modifications vs Car Ownership for Females

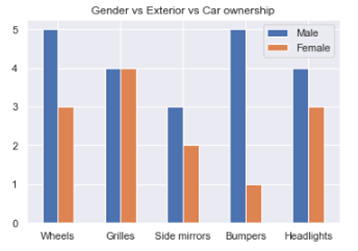


Figure : Bar plot for Gender vs Exterior modification vs Car Ownership

From the above bar plot, we can see that in general, males are more likely to make modifications to the exterior components as compared to females, with the exception of the grille component. The greatest difference would be the bumper modifications

## 2.2 Data Mining

Next, we move on to perform some **Data Mining** on our dataset to uncover more insights that are not easily seen through simple Data Exploration techniques. In this case, we performed a few types of data mining techniques: Principal Components Analysis (PCA) / Multiple Correspondence Analysis (MCA) and Associate Rule Mining.

### 2.2.1 Principle Components Analysis (PCA) / Multiple Correspondence Analysis (MCA)

Principal Components Analysis is a popular data analysis dimensionality reduction technique. Dimensional reduction helps us transform data to a state where it may be more easily available for data analysis. For example, transforming 3-dimensional data into 2-dimensional data (3 variables into 2 variables). PCA attempts to reduce the complexity of the problem by reducing the number of dependent variables, as we aim to select only a few variables while trying to preserve as much of the original information as possible. A PCA implementation is more suited for continuous or numerical variables, as it heavily relies on the calculation of the variances of the data. However, PCA is ineffective when applied on categorical variables due to the lack of a numerical variance for such variables. As most of our dataset’s columns contain categorical variables, we apply a similar Data Mining technique known as Multiple Correspondence Analysis instead of PCA, which is more suited toward categorical variables.[2]

MCA adopts the same basic principles of PCA; to maximize the extraction of important information while simultaneously reducing the dimensionality of the data. MCA is a well-recognized method when it comes to data dimension reduction for categorical data. More information and examples of MCA may be found in the references below in the papers appended in the References.[3][4]

For our analysis, we look at the data from 7 different columns, namely, “Age Group”, “Gender”, “Ownership”, “Marital Status”, “Exterior”, “Interior”, and “Personalization”. After extracting data from the relevant columns, we performed PCA/MCA analysis to reduce the dimensionality of the data from 7 to 3.

The PCA library has an ‘explained variance ratio’ attribute that may be called, to see the weightage of variance for each component (each value represents each column in the dataset). In this example, together these 7 components explain 100% of the variability in the data. A rule of thumb is to keep at least 70-80% of the explained variance. Hence in this case, picking the first three components (n = 3), preserves over 80 % of the variance[Figure 17].

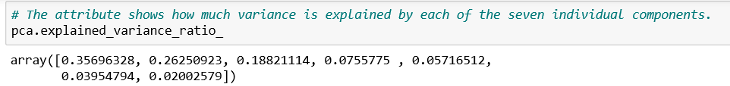


Figure : Estimating the Variances of all 7 components/columns

However, when we use MCA, there is no such attribute that may be called as variance cannot be measured. So, we assume n=3 in our analysis, meaning that the dimensionality of the data will be reduced from 7 to 3; these 3 components will then be used to explain the variability in the dataset and be tested on their ability to represent and encapsulate the variability in the dataset[Figure 18].

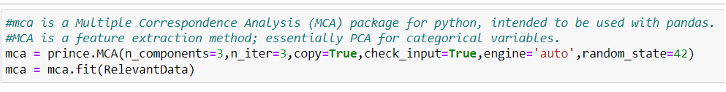


Figure : Selecting 3 components for MCA from the dataset

Once MCA has been fitted to the data, row principal coordinates may be extracted. Each column stands for a component whilst each row represents the corresponding row in the original dataset[Figure 19]. These projections then can be displayed with the ‘plot row coordinates’ method as well [Figure 20].

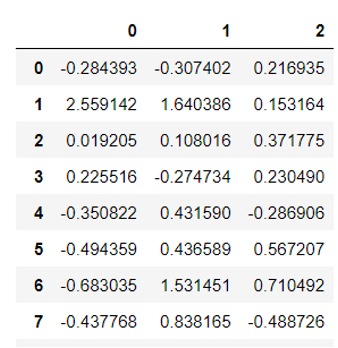


Figure : “Variances” of the Components for the first 7 entries in the dataset

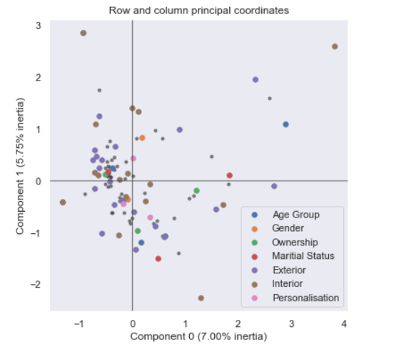


Figure : Graphical Representation of Component Analysis

With the help of a generalized SVD, MCA can decompose a vector into its orthogonal components. Each principal component explains a part of the distribution and variance in the dataset. The “explained\_intertia\_” attribute allows us to quantify this covariance. Each number in Figure 19 represents the weight/mass that the component contributes to the overall variance in the dataset[Figure 19]. The explained inertia is obtained by dividing the eigenvalues, obtained from the graph in Figure 20 and the generalized SVD, by the total inertia. Each eigenvalue represents the variance along that axis. The greater the eigenvalue, the higher the amount of information that is retained from the original dataset, along that axis. This task of calculating and obtaining the eigenvalues is similar to the steps involved in a PCA analysis.

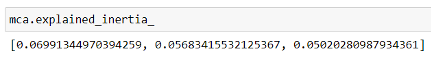


Figure : “explained\_intertia\_” values of the components

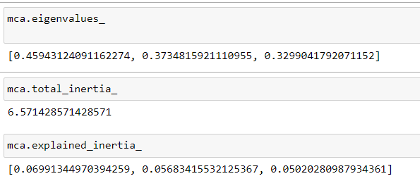


Figure : The process of calculating explained\_intertia\_

In the next section we will test k-means clustering with the MCA data that we obtained. There are no rules to choose the number of clusters in a K-means algorithm. The cluster sizes will be based on one’s dataset. In our example, we will test an algorithm for up to 10 clusters.

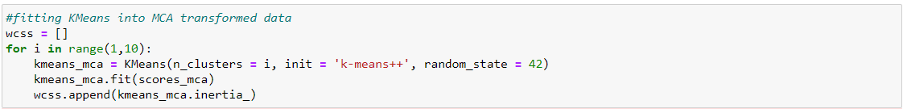


Figure : Choosing the number of clusters for the MCA dataset

Next, we will determine the Within Cluster Sum of Squares or WCSS for each solution. Based on the values of the WCSS and an approach known as the Elbow method, we will decide about how many clusters we’d like to keep.

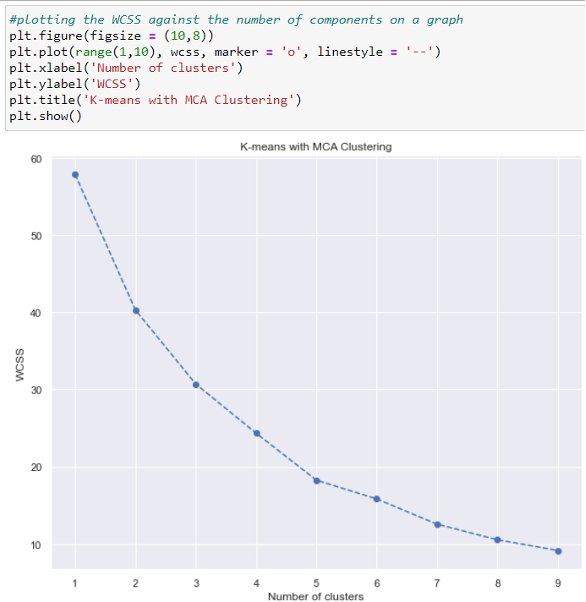


Figure : Plotting of the WCSS against the number of components on a graph

Finally, by inspecting the shape of the graph we will choose the number of clusters based on the elbow method. In this case we pick 3 clusters.



Figure : Choosing the number of clusters

The next step of the process would be to analyse the results of MCA and K-means clustering. Firstly, we will create a new data frame. This allows us to add in the variance values of the components to our original dataset ‘RelevantData’. The components’ variance which are stored in the ‘scores MCA’ variable, would be placed in the columns MC1, MC2, MC3 respectively (n=3 initially). In addition, we also append the K means MCA labels to the new data frame.

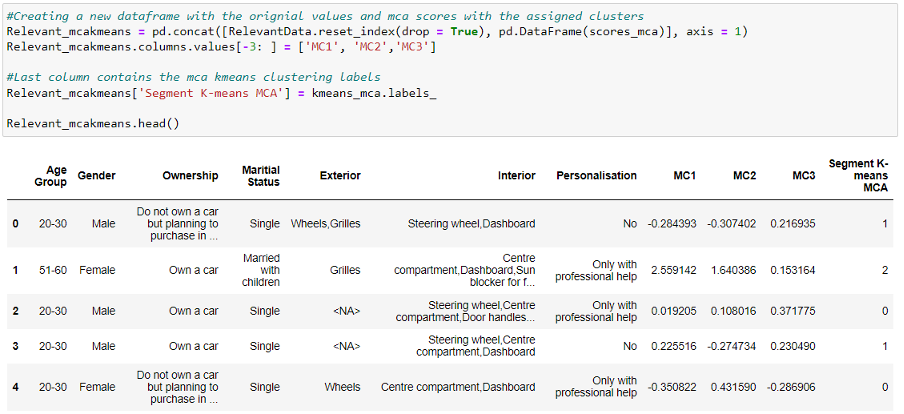


Figure : Concatenated data frame

To allow visual clarity a new column named ‘Segment’ is created to map the three clusters and make the labels of the graph.

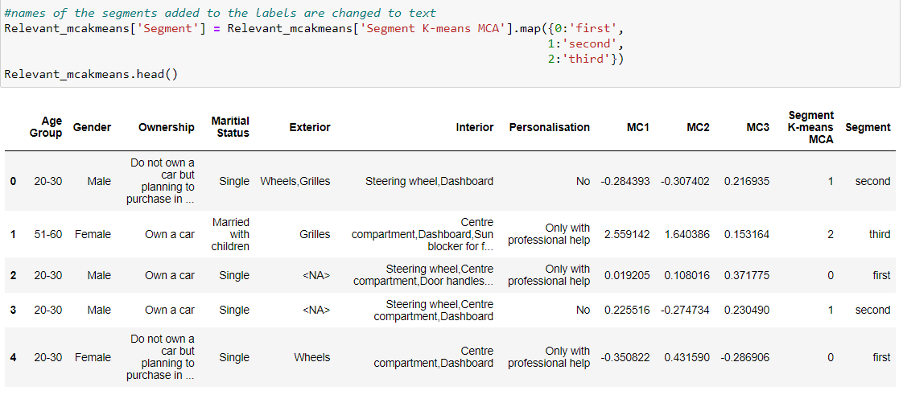


Figure : Mapping of segments to values

Now let’s visualize our clusters on a 2D plane. In a 2D visualization, so we need to choose two components and use them as axes. In doing so, we can compare if the first two components hold a higher variance as compared to the third one.

When we employ MCA prior to using K-means we can visually separate almost the entire data set. That was one of the biggest goals of MCA - to reduce the number of variables by combining them into bigger, more meaningful features. This allows the difference between components to be shown as big as possible. In the graph below, there is some overlap between the green and blue segments. But overall, all three segments are clearly separated. The spots where the two overlap are ultimately determined by the third component, which is not available on this graph.

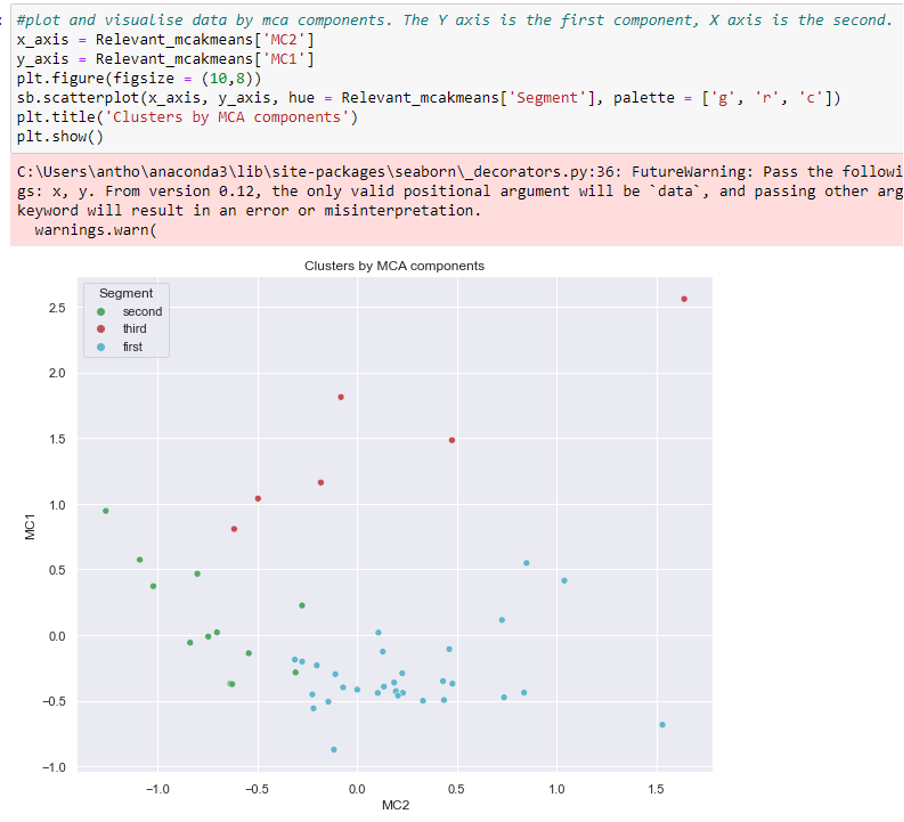


Figure : Visualizing of clusters

### 2.2.2 Associate Rule Mining

Associate rule mining is a data mining technique which observes frequent patterns, correlations or associations from datasets. Association rule mining is suitable for non-numeric and categorical data and the goal is to find the relationships between the items in a dataset.

Association rule is a learning technique that identify and describes the relationships between two or more items in a data set. For example, an association rule might identify that for customers that purchase cereal are also likely to purchase milk. With this knowledge, the association rule may be able to make recommendations to users that are planning to buy cereal.

In associate rule mining, several metrics such as support, confidence and lift are calculated to determine the strength and patterns of association between different items.

Support is the measure of how frequent an itemset appears in the whole dataset. The formula to calculate support is as follows:

Confidence is the measure of strength of association between two or more items in an itemset. The formula to calculate confidence of the rule {A} → {B} is as follows:

Lift is the degree of dependency between two items in an item set. The formula to calculate lift is as follows:

If a lift value is greater than 1, it indicates that the two items are positively correlated and if the life value is lesser than 1, it indicates that the two items are negatively correlated. A lift value of 1 would indicate that the items are independent of each other.

Several algorithms are used for associate rule mining such as the Apriori algorithm, FP-Growth algorithm and the ECLAT algorithm. In this segment, we will be utilizing the Apriori algorithm to do associate rule mining on the given dataset.

From the given dataset, we will be looking at the interior and exterior modifications that customers would like to modify if they own a car and see how the different items for both interior and exterior modications associate and affect with one another.

Most of the columns in the dataframe are dropped, leaving only the exterior and interior columns. A new dataframe would then be created with binary columns for each unique part of the car’s modifications.

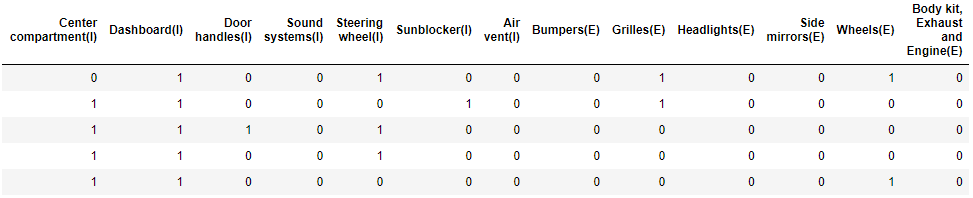


Figure : Dataframe containing binary columns

For example, in the first row, if the user owns a car, the user would modify the dashboard and steering wheel for the interior and would modify the grilles, wheels and body kit, exhaust and engine for the exterior of the user’s car.

The “mlxtend” library is used and importing “apriori” and “association\_rules” to create association rules. These association rules are then ranked accordingly to 3 metrics; support, confidence and lift.

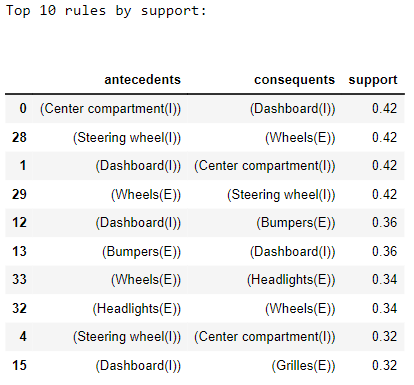


Figure : Top 10 rules using Support metrics

The figure above states the top 10 rules that is ranked using the support criterion. The first rule states that when the center compartment is present, the dashboard is likely to be present with a support of 0.42. Similarly, the second rule states that when the steering wheel is present, the wheels is also likely to be present with a support of 0.42 as well.

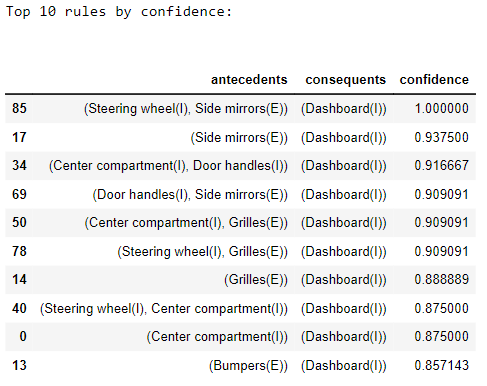


Figure : Top 10 rules using Confidence metrics

The figure above shows the top 10 rules that has the highest confidence levels. Looking at the first rule, it states that when both the side mirror and the steering wheel are present, the dashboard is present with 100% confidence. From this, we can infer that every transaction which includes both the side mirror and steering wheel, the dashboard would also be guaranteed to be in that transaction. Now looking at the second rule, without the steering wheel present now, the confidence of the dashboard being present when only the side mirror is present becomes 93.75%. This still suggests that the side mirror and dashboard are often found together but not as strong as the first rule.

Furthermore, looking at the top 10 rules, we can see that the dasboard is present as the consequent in all of the rules which may suggest that the dashboard is a very frequent choice by users and has a strong relation to many other items.

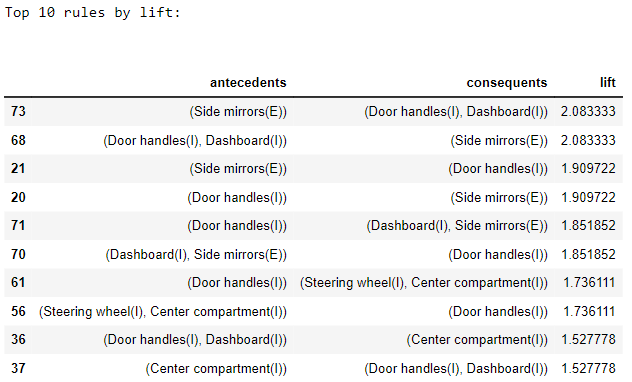


Figure : Top 10 rules using Lift metrics

The figure above shows the top 10 rules with the highest lift that shows a positive correlation between these items. From the second rule, the lift of door handles and dashboard in predicting the presence of side mirrors is 2.083. This indicates that the presence of the door handles and dashboard together makes the appearance of the side mirrors 2.083 times more likely than expected if the items were independent. This shows a strong positive correlation between the doorhandles and dashboard as antecedents and side mirrors as consequents.

In conclusion, the Apriori algorithm scans the dataset multiple times and generates larger and larger itemsets by building onto frequent singular items thus generating larger items sets through multiple iteration. This algorithm is able to handle large datasets with many items and transactions. However, as the number of items and transactions increases, this algorithm requires more computational resources and memory. Furthermore, this algorithm may generate a large number of redundant association rules as well.

# 3. Conclusion and Discussion

With the different types of Data Exploration and Data Mining performed on the dataset, we can draw a few conclusions based on the results.

From the Data Exploration, we can conclude that the personalisation of cars is one factor which individuals would be interested about, provided that there are customised packages which that are offered to them which includes factors like having designers to design their car or at a deceased and discounted cost. Furthermore, based on the results, we can also conclude that males take up a higher percentage of ownership of a car, and having customisation to their car as well.

From the Data Mining, it is typically hard to perform PCA on such a dataset as PCA is generally more suited towards numerical datasets. However, even with MCA, it is still difficult to analyse the results as we are able to draw correlations, but we are unable to tell what correlations are against each other. When we perform the Rule Association Mining technique, we look at Internal and External components of a car. We can establish that the dashboard and the steering wheel are major components of interest, especially when performing customisation on the car itself.

In conclusion, based on all the Data Exploration and Data Mining that has been performed, the results gathered indicate that there are indeed some correlations between the different types of variables and categories presented. With perhaps more data within the dataset, and perhaps with higher levels of Data Exploration or Data Mining techniques, we would be able to uncover much more results and conclusions in the future.

# 4. References

[1] N. Picard, S. Dantan, and A. de Palma, ‘Mobility decisions within couples’, *Theory Decis*, vol. 84, no. 2, pp. 149–180, Mar. 2018, doi: 10.1007/s11238-017-9648-2.

[2] finnstats, ‘PCA for Categorical Variables in R | R-bloggers’, Nov. 20, 2022. https://www.r-bloggers.com/2022/11/pca-for-categorical-variables-in-r/ (accessed Feb. 15, 2023).

[3] P. S. Costa, N. C. Santos, P. Cunha, J. Cotter, and N. Sousa, ‘The Use of Multiple Correspondence Analysis to Explore Associations between Categories of Qualitative Variables in Healthy Ageing’, *J Aging Res*, vol. 2013, p. 302163, 2013, doi: 10.1155/2013/302163.

[4] M. Dungey, F. Doko Tchatoka, and M. B. Yanotti, ‘Using multiple correspondence analysis for finance: A tool for assessing financial inclusion’, *International Review of Financial Analysis*, vol. 59, pp. 212–222, Oct. 2018, doi: 10.1016/j.irfa.2018.08.007.