

**MA4829 Machine Intelligence**

**Group Assignment:**

Product Survey Data Analysis

**Proposed Title:**

Decoding Car Ownership: Insights from a Categorical Analysis Dataset

**Members:**

|  |  |
| --- | --- |
| Tan Yi Ren | U2022894H |
| Mohamed Raizee Bin Mohamed Ibrahim | U2022754H |
| Nadia Houti | U2021324A |
| Arunachalam Thannirmalai | U2021049E |
| Selvam Vinothini Vibhusha | U2020585D |
| Ng Jun Wei | U2121730H |

## 

**Abstract**

The automotive industry has seen a rise in consumer demand for car customisation and personalisation. Machine learning is an effective tool that can be used by analysts to understand the underlying preferences and motivations behind this demand, in order to predict future outcomes and offer useful insights to businesses. In this report, we explore the correlations between various categorical factors influencing car ownership and customisation preferences, with the aim of optimising sales of these services through data analysis. A survey of 50 participants captured demographic information such as gender, marital status and car ownership status. Utilising techniques like Principal Component Analysis (PCA), Multiple Correspondence Analysis (MCA) and A Priori analysis, inferences were made regarding the relationship between variables affecting consumer decision on the purchase and customisation of cars. Additionally, visualisation tools such as bar charts and cluster distributions were employed to enhance the analysis. The key findings from this report were then summarised as suggestions for automotive companies to better target ideal customer groups and improve sales.

**Table of Contents**

[1. Introduction 4](#_hvx3c4z9iv26)

[1.1. Background 4](#_8y7fmvrbk6bi)

[1.2. Objective 4](#_tl19nvta8rde)

[2. Data Cleaning 5](#_gr8g3ohgif0e)

[3. Data Exploration 8](#_iobgdneg73q0)

[3.1. Demographic Distribution 8](#_rgmrcxvkkfwf)

[3.1.1. Age Group Distribution 9](#_21njg6tkih3i)

[3.1.2. Gender Distribution 9](#_2rxg2di00g1)

[3.1.3. Marital Status Distribution 10](#_899kp9e3fmbf)

[3.1.4. Car Ownership Distribution 10](#_ym0f32a58ydw)

[3.2. Customization Distribution 11](#_vwiemta4l1yl)

[3.2.1. Varied 3D Design Experience vs. Number of Parts Customization 11](#_opy0k83xmfhz)

[3.3. Correlation of Number of Car Parts to Customise vs Number of Factors to Owning a Car 12](#_7gixlnmo6nvj)

[3.3.1. Polynomial Regression 13](#_grogv7uaa042)

[3.3.2. Support Vector Regression 14](#_5d3victbrq1h)

[3.3.3. Correlation Matrix 15](#_b1978q2vu9f6)

[4. Data Mining 16](#_598im2zgym10)

[4.1. Principal Component Analysis (PCA) 16](#_hwxsa9cuxmqx)

[4.2. Multiple Correspondence Analysis (MCA) 17](#_ae08v9c9x4gj)

[4.3. Apriori 23](#_sxblwo27m6mw)

[5. Conclusion 27](#_glwwaxdtip1d)

[6. References 28](#_s231rwejkrbn)

# Introduction

## Background

In recent years, the automotive industry has witnessed a surge in the demand for car customisation and personalisation. From aesthetic modifications to performance enhancements, the consumer demand to personalise vehicles with unique characteristics reflects a growing trend among car enthusiasts. Understanding the underlying motivations and preferences driving this trend is crucial for automotive manufacturers, marketers and suppliers seeking to cater effectively to consumer needs.

By harnessing the power of machine learning algorithms, businesses can uncover intricate relationships, predict future outcomes and optimise decision-making processes, thereby gaining a competitive edge in today's dynamic marketplace.

## Objective

This report aims to explore the correlations of various categorical factors influencing car ownership and customisation preferences to optimise the sales of these services through data analysis. 50 participants were surveyed on various demographic factors such as gender, marital status and car ownership status. Machine learning techniques were used to investigate if the algorithms applied proved accurate in modelling the dataset.

# Data Cleaning

In the survey, 50 participants were asked 13 questions relating to their profile. The answers collected from the respondents are all categorical in nature. For simplification purposes, the questions were rephrased to be more concise as shown in Table 1 below.

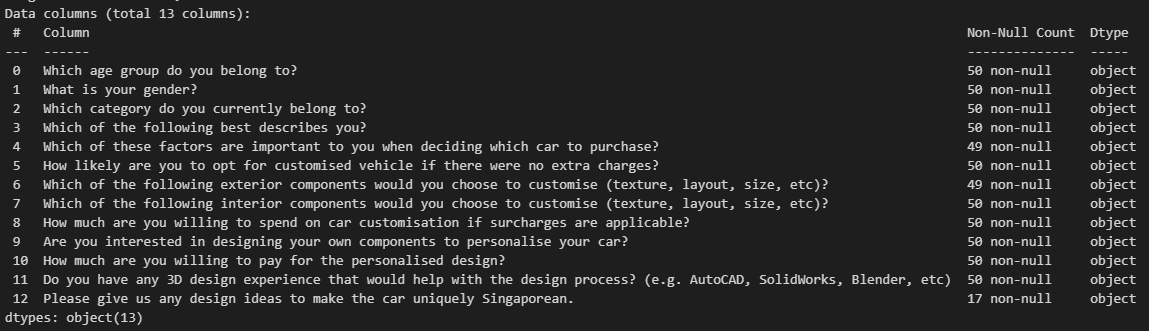


Figure 1 Non-numerical nature of the dataset

Table 1 Original vs. simplified questions after data cleaning

|  |  |
| --- | --- |
| **Original 13 questions** | **Simplified questions** |
| Which age group do you belong to? | Age group |
| What is your gender? | Gender |
| Which category do you currently belong to? | Category |
| Which of the following best describes you? | Marital status |
| Which of these factors are important to you when deciding which car to purchase? | Deciding factors to buy car |
| How likely are you to opt for a customised vehicle if there were no extra charges? | Likelihood of customised vehicle if FOC |
| Which of the following exterior components would you choose to customise (texture, layout, size, etc)? | Exterior components to customise |
| Which of the following interior components would you choose to customise (texture, layout, size, etc)? | Interior components to customise |
| How much are you willing to spend on car customisation if surcharges are applicable? | Customisation Budget |
| Are you interested in designing your own components to personalise your car? | Interest in personalising car |
| How much are you willing to pay for the personalised design? | Personalisation Budget |
| Do you have any 3D design experience that would help with the design process? (e.g. AutoCAD, SolidWorks, Blender, etc) | Any cadding experience |
| Please give us any design ideas to make the car uniquely Singaporean. | Singaporean design idea feedback |

Some columns contained “NaN” (not a number) values as the respondents left some of the questions blank. The “NaN” entries were replaced with a “-” for ease of reading as shown in Figure 2 below.

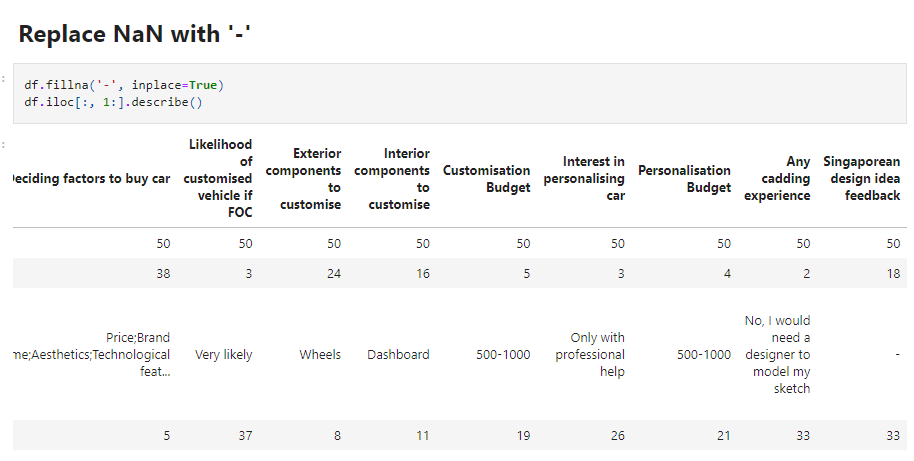


Figure 2 Code used to replace “NaN” entries

The questions on “Exterior components to customise”, “Interior components to customise” and “Deciding factors to buy a car” allowed the respondents to pick multiple answers from a given set of options. The total count of the options chosen for each question was created as new data columns with the following names respectively “Exterior Component count”, “Interior Component count” and “Factors to buy Car count”.

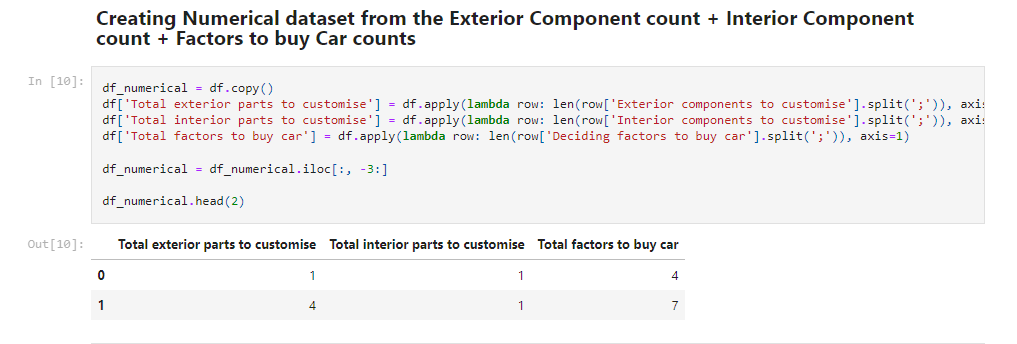


Figure 3 Code used to generate the new data columns

# Data Exploration

## Demographic Distribution

The survey dataset primarily reflects the opinions and preferences of individuals within the 20-30 age group as they make up 82% of the respondents, indicating a focus on the perspectives of younger individuals. This demographic, consisting of both males and females, is primarily composed of individuals who currently do not own a car but are contemplating a future purchase. Furthermore, the majority of survey respondents are single men, as illustrated in Figure 4.

In addition to the significant representation of the 20-30 age range, there is a small proportion of participants in the 31-40 and 41-50 age groups, most of whom already own a car. The dataset captures all major categories of marital statuses, including both single and married individuals, with and without children.

This age group distribution provides a comprehensive snapshot of preferences and considerations among individuals at different life stages, offering insights into the evolving landscape of car ownership and customization across various demographics.

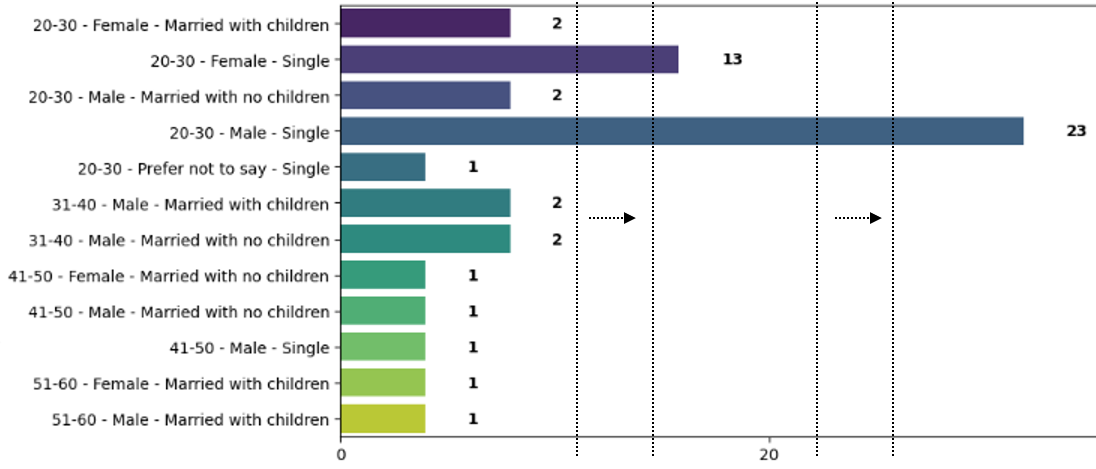


Figure 4 Detailed Grouped Chart

### Age Group Distribution

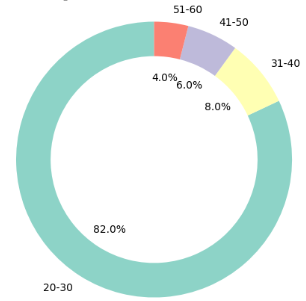


Figure 5 Age Group Distribution Chart

|  |  |
| --- | --- |
| 20-30 Age Group | 41 respondents |
| 31-40 Age Group | 4 respondents |
| 41-50 Age Group | 3 respondents |
| 51-60 Age Group | 2 respondents |

### Gender Distribution

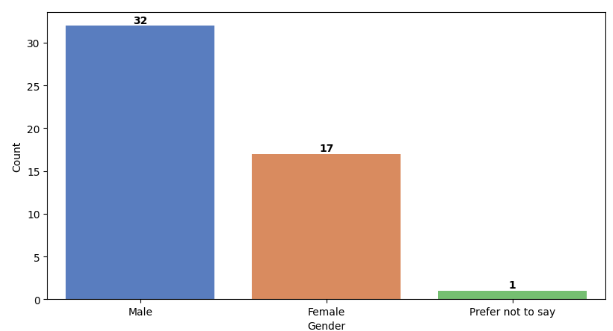


Figure 6 Gender Distribution Chart

|  |  |
| --- | --- |
| Male Gender | 32 respondents |
| Female Gender | 17 respondents |
| Undisclosed | 1 respondent |

### Marital Status Distribution

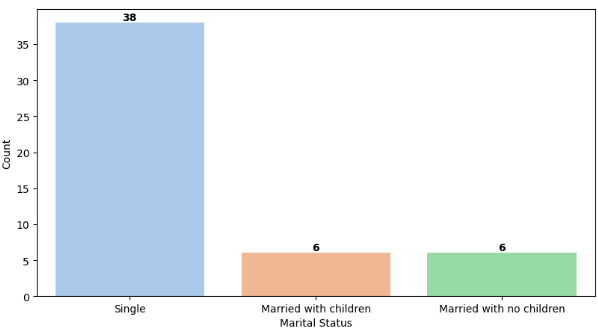


Figure 7 Marital Status Distribution Chart

|  |  |
| --- | --- |
| Single | 38 respondents |
| Married with children | 6 respondents |
| Married without children | 6 respondents |

### Car Ownership Distribution

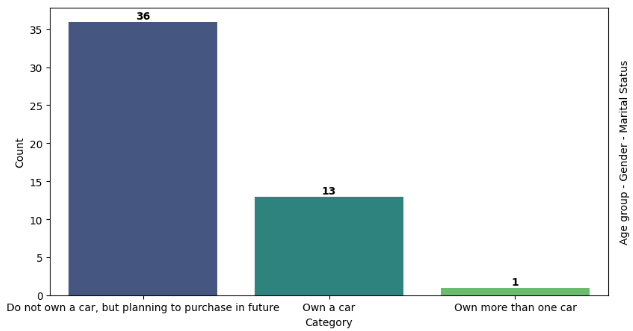


Figure 8 Car Ownership Distribution Chart

|  |  |
| --- | --- |
| Do not own a car, but planning to purchase in future | 36 respondents |
| Own a car | 13 respondents |
| Own more than 1 car | 1 respondent |

## Customization Distribution

Respondents from various demographics express a keen interest in customising their cars based on factors including aesthetics, brand name, technological features and functionality. Customization budgets range from under $500 to over $1,000 SGD. The 20-30 age group stands out as the most interested in car customization, suggesting an ideal market within this demographic.

From the dataset, respondents provide a range of design ideas for both exterior and interior components. Exterior components preferred for customization include wheels, grilles, headlights, side mirrors and bumpers while interior components chosen for customization include steering wheels, dashboard, centre compartment, door handles and brakes.

Wheels customization is preferred by many respondents, indicating a strong emphasis on the visual aspect of the automobile. Grilles, headlights, side mirrors and bumpers are the next commonly selected exteriors for customization. Both current car owners and future car owners express a strong interest in customization indicating the strong market for car customisation even after a car has been purchased by the consumer.

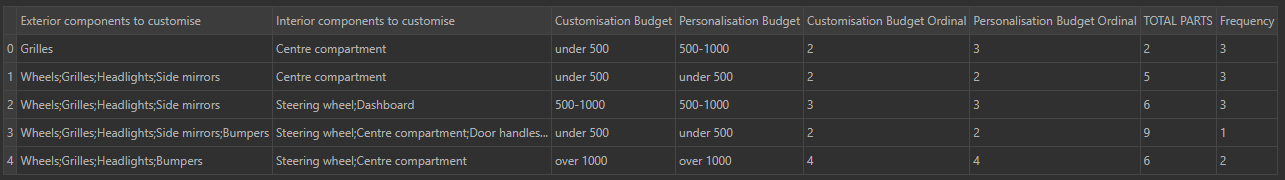


Figure 9 Components to be customised

### Varied 3D Design Experience vs. Number of Parts Customization

We explored the relationship between the total number of car parts respondents wish to customise and the total number of factors considered when purchasing a car.

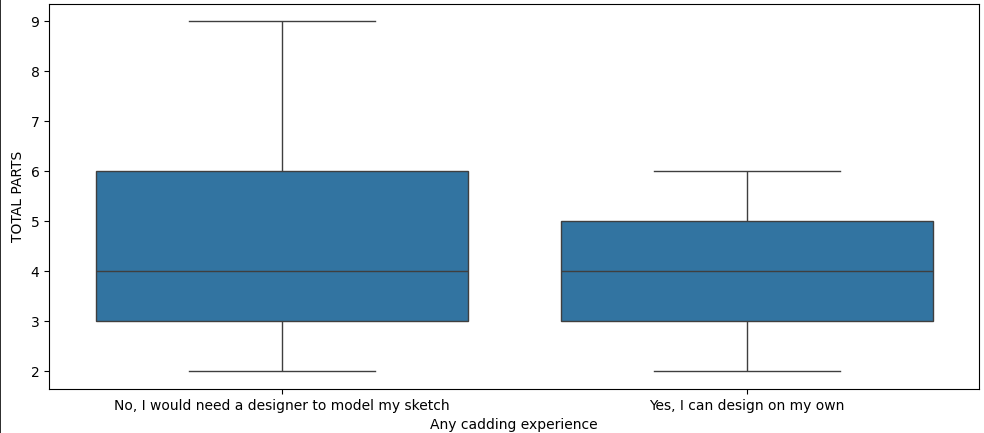


Figure 10 Relationship between customisation preferences and factors influencing car ownership

Respondents have various levels of 3D design experience. The majority of them prefer seeking professional assistance instead of designing on their own. Figure 10 shows that there is a greater spread in the number of parts respondents wish to customise when the customisation is outsourced to professionals.

## Correlation of Number of Car Parts to Customise vs Number of Factors to Owning a Car

The following section investigates if there is any relationship between the total number of car parts that respondents want to customise and the total number of factors consumers consider when purchasing a car.

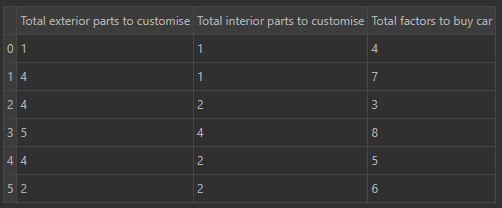


Figure 11 Correlation between the number of car parts to customise and the factors influencing car purchases

Results may reveal patterns such that respondents expressing a desire for more personalised 3D designs also tend to weigh a greater number of factors when deciding to purchase a car. Understanding this correlation could have implications for marketing strategies, product development and customer engagement within the automobile industry.

### Polynomial Regression

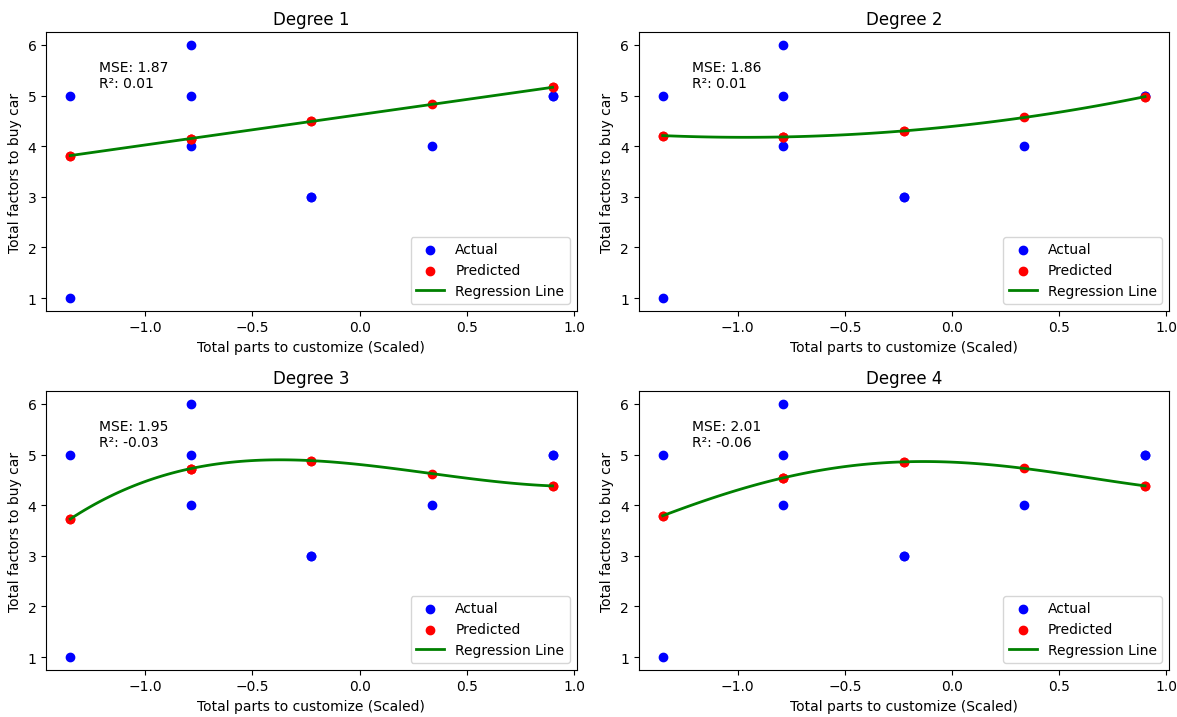


Figure 12 Polynomial Regression

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Degree | 1 | 2 | 3 | 4 |
| Mean Squared Error (MSE) | 1.87 | 1.86 | 1.95 | 2.01 |
| R-squared (R²) | 0.01 | 0.01 | -0.03 | -0.06 |

We preliminarily tested a linear regression model on the data and found that the R-squared value attained was extremely low at 0.01 suggesting that there exists no linear relationship between the two variables. Hence, we then attempted to use a polynomial regression model instead. A polynomial regression model can capture non-linear relationships between variables by fitting a non-linear regression line, allowing for a better fit of the model as compared to a linear relationship [1].

Degrees 1 and 2 exhibit similar MSE values of 1.87 and 1.86, suggesting a modest fit. However, the R squared value of 0.01 indicates that the model explains only a minimal portion of the variability in the data and is unsuitable for modelling the dataset. As the degrees were increased to 3 and 4, the R-squared value became negative and worse performing than degrees 1 and 2. The polynomial regression models were unable to comprehensively represent the data regardless of the order of the regression model.

### Support Vector Regression

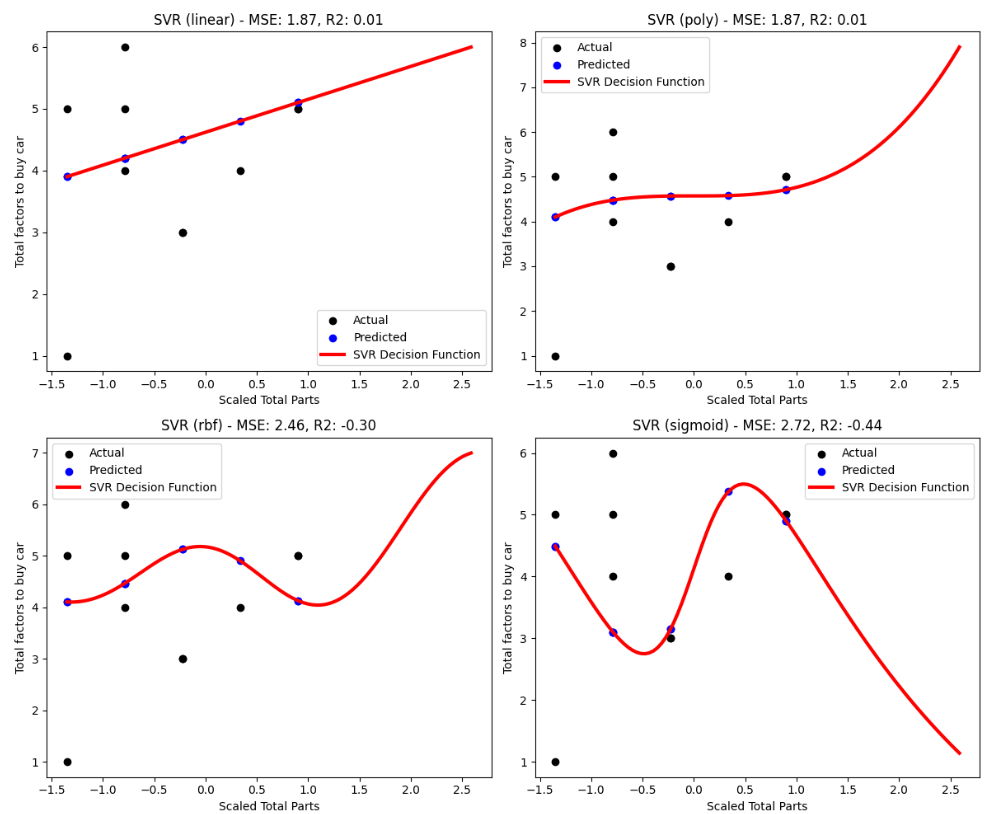
****

Figure 13 SVR Regression

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| SVR | Linear | Poly | RBF | Sigmoid |
| Mean Squared Error (MSE) | 1.87 | 1.87 | 2.46 | 2.72 |
| R-squared (R²) | 0.01 | 0.01 | -0.30 | -0.44 |

As both the linear and polynomial regression models were unable to accurately model the dataset, the support vector regression technique was explored.

The Support Vector Regression (SVR) is a machine-learning technique used for regression tasks [2]. For us to perform our analysis using SVR with the higher dimensional vectors in the transformed space, we have to use the Kernel Trick. The Kernel Trick maps the input features into a higher dimensional space where it can find a hyperplane that best separates the data points [3]. This hyperplane represents the non-linear relationship between the input features and the target variable. The choice of kernel function in SVR plays a crucial role in capturing non-linear patterns. All 4 kernel functions provided in the sklearn library were tested to find out the best-performing model. Of the 4 kernel functions, the linear and polynomial functions performed best in modelling the relationship between the variables with an R-squared value of 0.01. However, the extremely low R squared value indicates that SVR is also unable to capture the trends of the dataset.

Overall, none of the models tested were able to fully capture the relationship between the total number of car parts that respondents want to customise and the total number of factors consumers consider when purchasing a car.

### Correlation Matrix

Another statistical tool we have adopted is the correlation matrix. The correlation matrix provides an overview of the relationship between the “Total Parts” and the “Total Factors to buy a car” where 1 shows a strong relationship between the variables, 0 shows a neutral relationship and -1 shows a weak relationship.

Our correlation matrix shows us that the 2 variables have a correlation of 0.36 which tells us that the factors are only correlated to a certain extent. Hence, we can conclude that they do not have strong influences on each other when consumers are purchasing a car.

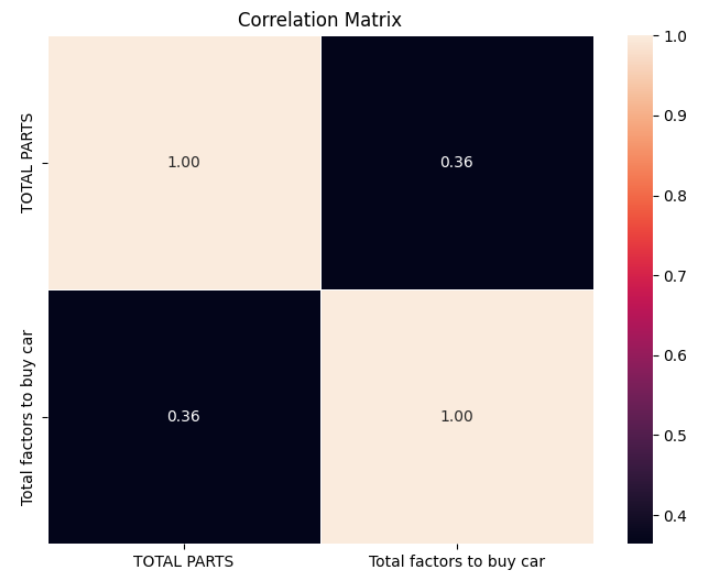


Figure 14 Correlation matrix

# Data Mining

## Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a data analysis technique which utilises supervised machine learning and data mining techniques to reduce the dimensionality of data [4]. PCA transforms original variables into a new set of uncorrelated variables called principal components based on the amount of variance they explain in the original data. Specifically, principal components can be identified with their orthogonal vectors representing variables with the most significant variance using eigenvalue-eigenvector decomposition [5]. A PCA plot allows group similarity and dissimilarity to be easily discerned .

A meaningful analysis using PCA can only be conducted on numerical data; however, we have observed that the data available is mainly categorical in nature. Hence, we applied this technique on 3 numerical variables that we have adapted from the original 13 variables to attempt a PCA analysis: “Total exterior parts to customise”, “Total interior parts to customise” and “Total factors to buy a car”.

Plotting the PC1 vs PC2 plot produces the results shown in Figure 15 – we observe that there are 5 clusters; however, all of them are scattered around and not distinct from one another. Such a result observed from the cluster plot resulting from PCA analysis could be the consequence of complex relationships between variables. The variables we have chosen to apply PCA analysis on may be related in ways that cannot be easily captured by a linear projection, making it difficult for clustering algorithms to partition the data into well-separated clusters.

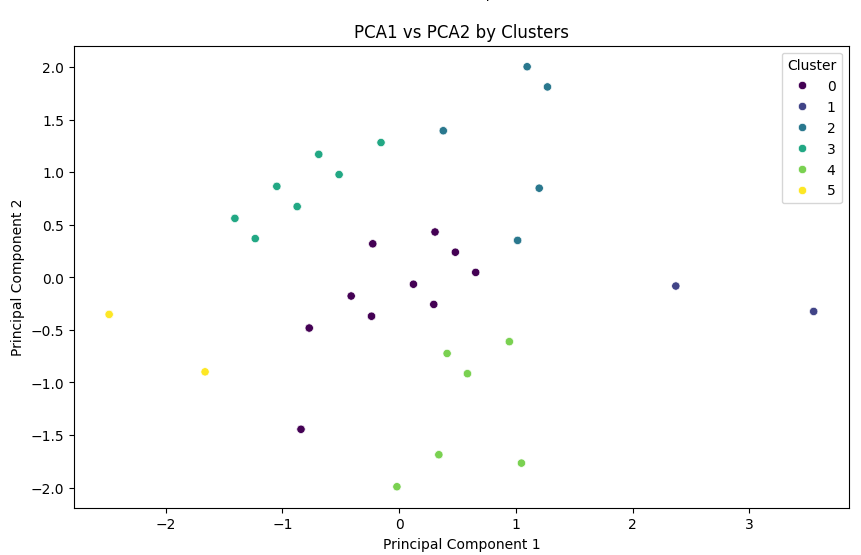


Figure 15 Cluster plot of PC2 against PC1

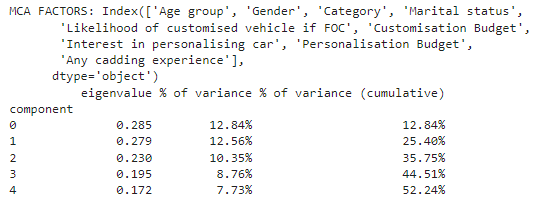
As predicted, PCA analysis was unable to provide pertinent insights on the data due to limitations of the dataset, particularly the lack of numerical data available. A more meaningful analysis could be conducted if this technique was used on a large-dimension dataset with numerical data. We explore an alternative to PCA analysis in the following section.

## Multiple Correspondence Analysis (MCA)

MCA applies techniques similar to PCA. However, it is used to conduct analysis on categorical data, highlighting relationships between qualitative variables [6]. Given the limited amount of numerical data available for this project, we have observed that MCA is a more appropriate data mining technique for our dataset.

Firstly, we used the Screeplot to determine how many components MCA should retain.

MCA was not able to retain most of its importance – as shown in Figure 16, reducing the number of MCA components to 5 brings the cumulative explained variance to 52.24%. However, for ease of plotting, we will reduce the number of dimensions to 2. We also note that as the variables are categorical data, they may have complex relationships that cannot be adequately captured in just 2 dimensions. Hence, this might pose future issues due to the inherent complexity of the data.

Figure 16 Explained variance for MCA analysis

Next, we utilised the Elbow plot shown in Figure 17 to determine the optimal number of clusters (K) for KMeans clustering. The Elbow plot is a graph of the within-cluster sum of squares (inertia) against the number of clusters. The "elbow" point, where the rate of decrease sharply changes, suggests the optimal number of clusters [7]. In this case, we have identified the optimal number of clusters within the MCA-transformed data to be 3.

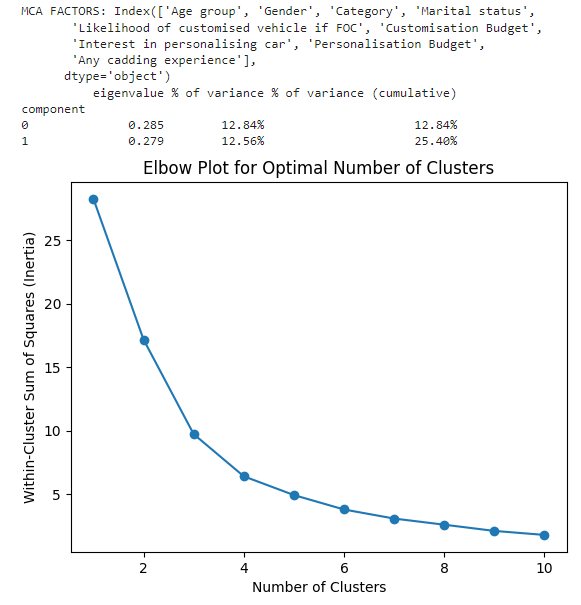


Figure 17 Elbow plot for MCA analysis

After transforming the categorical data into two-dimensional space using MCA, we obtain the scatter plot in Figure 18 which allows us to visualise the distribution of data points in the lower dimensional space. Each point represents a respondent of the survey and its colour represents the cluster it belongs to by the KMeans algorithm. This plot provides a visual overview of how well the respondents are grouped based on their categorical answers in the MCA-transformed space. Clustering the data allows us to observe patterns in the responses of survey respondents sharing similar characteristics (e.g. age group, gender, marital status, etc.). We observe the cluster distribution for these characteristics more closely by analysing the following categorical columns.

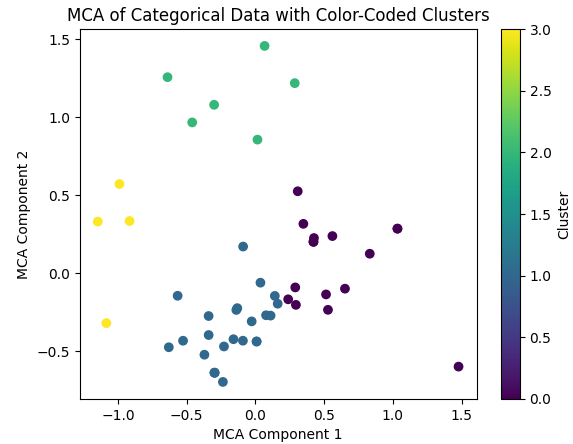


Figure 18 Post-MCA analysis cluster distribution

As shown in Figure 19, Clusters 0 and 1 comprise entirely of respondents in the age 20-30 group. In Figure 20, we observe that the majority of respondents from Cluster 0 (94%) are willing to spend the lowest amount on car customisation (between $0 to $500) whereas the majority of respondents from Cluster 1 (96%) are willing to spend the highest amount on car customisation ($500 and above). This shows that for the younger age group, Clusters 0 and 1 could represent the lower and higher end of the scale of willingness to spend on car customisation respectively.

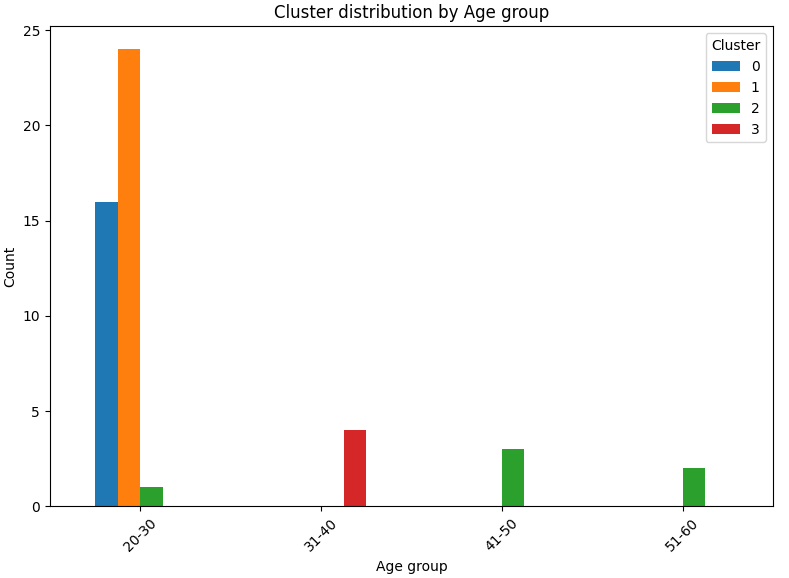


Figure 19 Cluster distribution based on age group

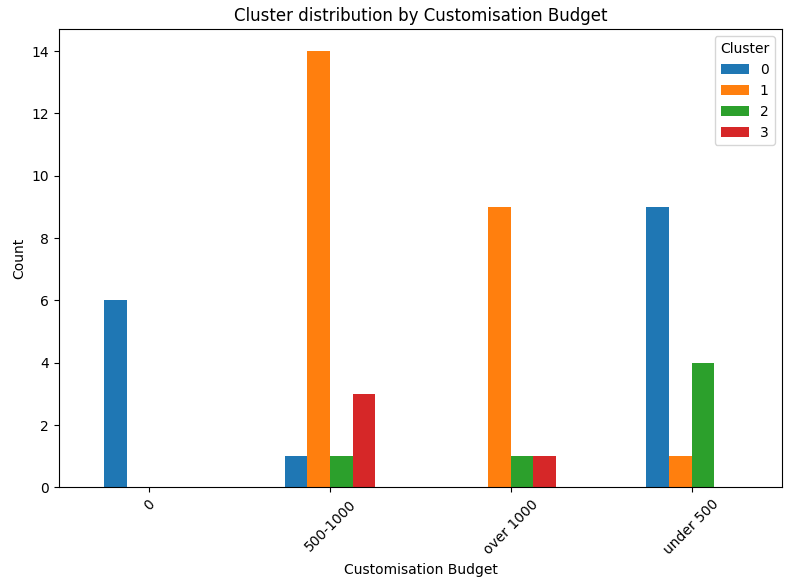


Figure 20 Cluster distribution based on customisation budget

From Figure 21, we can see that between Clusters 0 and 1, the proportion of males is more skewed towards Cluster 1 while the proportion of females is more skewed towards Cluster 0. In conjunction with observations from Figure 20, we hypothesise that a relationship between gender and willingness to spend on car customisation may exist – males appear to be more willing to spend a higher amount on car customisation compared to females.

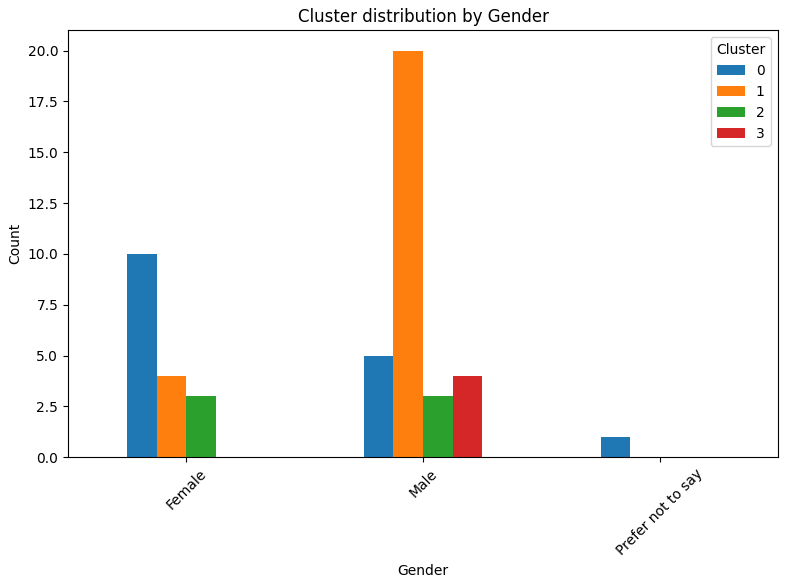


Figure 21 Cluster distribution based on gender

We further observe from Figure 22 that the group of respondents who “do not own a car, but are planning to purchase one in the future” comprises mostly of respondents from Clusters 0 and 1 (i.e. the 20-30 age group). From the observations made from Figures 19-22, a useful observation can be made: companies selling car customisation services do not necessarily have to price their products lower for the younger age group. In fact, a high proportion of non-car owners from the 20-30 age group (Cluster 1) is willing to spend the most on car customisation. Further studies and data will be necessary to ascertain the specific characteristics of this group to better target car customisation services and products towards them.

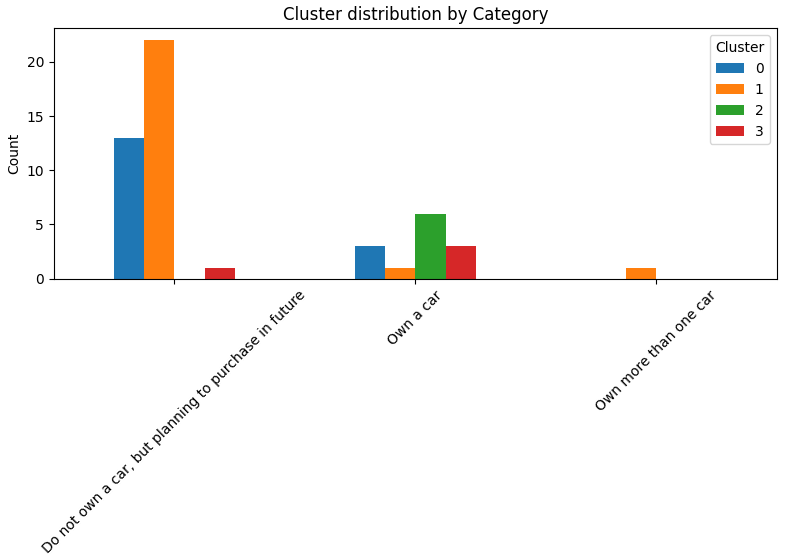


Figure 22 Cluster distribution based on car ownership status

## Apriori

## 

Another data mining technique that we explored was the Apriori Algorithm. The Apriori Algorithm uses a support-based pruning to control the growth of the search of itemsets [8]. Its mining of association rules can be broken down into 2 stages: 1) Identifying all frequent itemsets and 2) generating high-confidence rules from these frequent item sets found in the first stage. Ultimately, the high-confidence rules will then be used to generate new candidate rules.

The first variable we tested was “Deciding Factors to Buy a Car”.

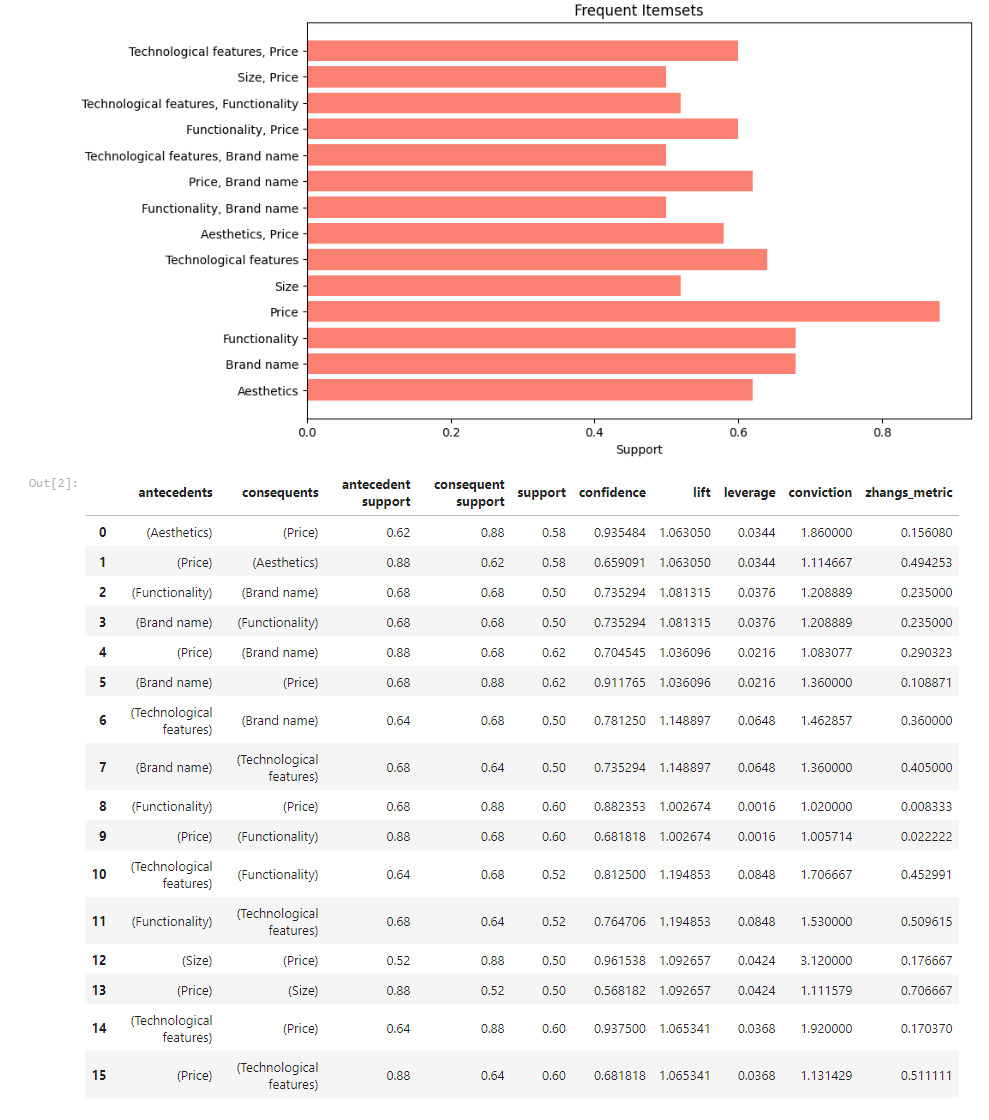


Figure 23 Frequent itemsets for deciding factors to buy a car

With the highest support and confidence, “Price” is the most significant factor in deciding to buy a car. “Branding” and “Functionality” are tied at second as the most significant factors in deciding to buy a car.

The “Lift” column measures the strength and significance of an association rule between two choices. Lift = 1 indicates independence, Lift > 1 suggests a positive association, and Lift < 1 indicates a negative association [9]. As predicted, technological features and functionality have one of the highest lifts at 1.19 indicating that if a consumer prefers functionality, it is likely that they prefer technological features as well.

Next, we analysed “Exterior Components to Customise a Car”.

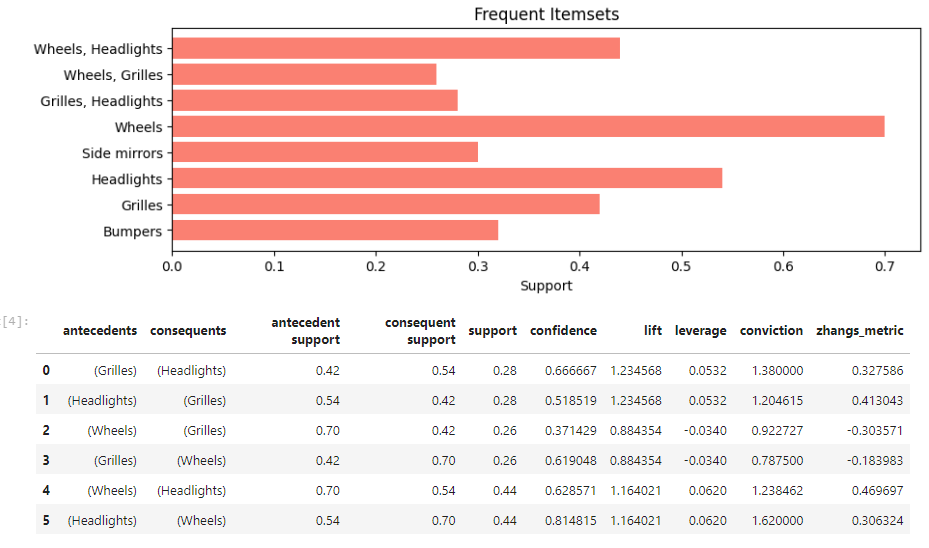


Figure 24 Frequent itemsets for exterior components to customise a car

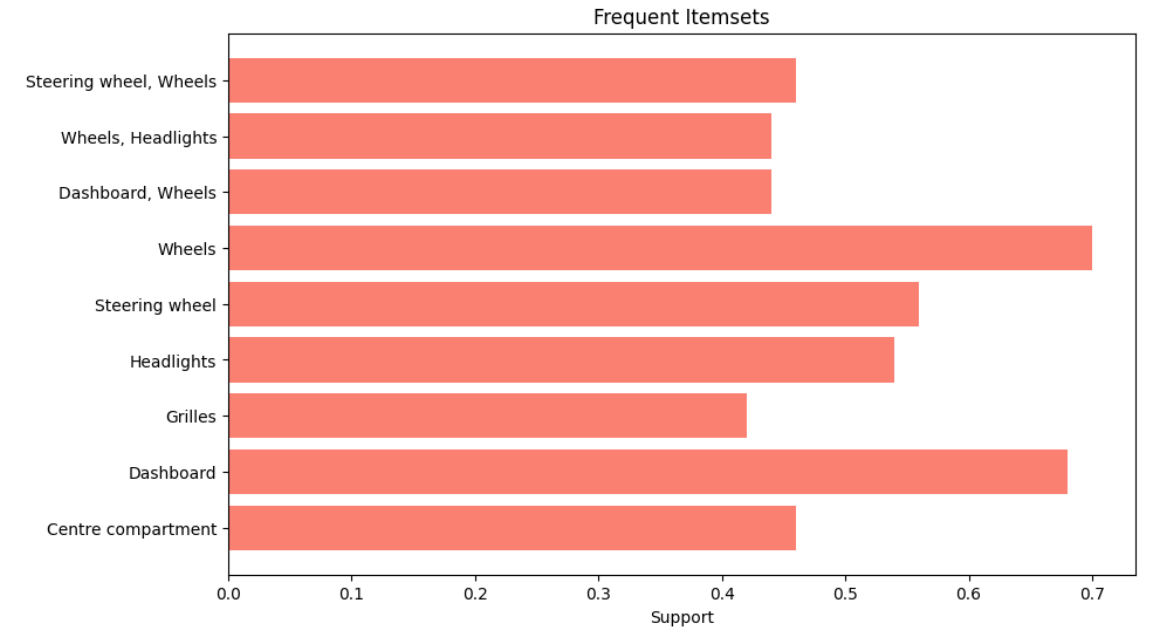
In this case, we can observe that “Wheels” are the most likely exterior component to be customised in a car, with a 54% chance of it being customised together with headlights. “Grilles” and “Headlights” are the most likely pair to be customised together with a lift of 1.234. This implies that companies looking to sell custom grilles are more likely to generate sales if custom headlights are offered to customers as well.

Next, we analysed the “Interior Components to Customise a Car”. We observe that the “Dashboard” is the most likely interior component to be customised followed by the “Steering Wheel”. It can also be seen that the lift and conviction for dashboard and centre compartments are less than 1, showing that these 2 items are less likely to be bought together.



Figure 25 Frequent itemsets for interior components to customise a car

Next, we evaluated the relationship between the interior and exterior parts customisation. If respondents chose steering wheels for interior customisation, they are likely to pick wheels for exterior customisation with a high conviction of 1.68 and a lift of 1.17. Similarly, if headlights were picked for interior customisation, wheels were likely chosen for exterior customisation as well. Hence, we can infer that customers tend to customise these features concurrently.



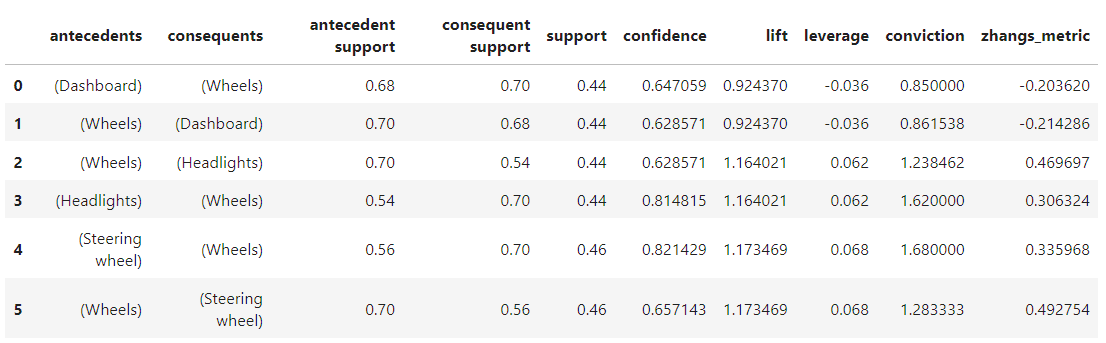


Figure 26 Frequent itemsets for interior and exterior parts customisation

The above analysis provides insights into the relationship between the variables and hence allows for better pricing models and product combinations. With these insights on consumer behaviour, companies can identify frequent item sets and apply market basket analysis to maximise sales by selling products frequently purchased together by customers.

# Conclusion

## 

The analysis conducted can be translated into the following insights for business owners to alter their marketing and sales strategies:

1. For younger customers, males tend to be more willing to spend a higher amount on car customisation compared to females. Conducting market studies on popular car customisation choices among males could potentially yield higher returns.
2. Customers are more likely to customise the wheels of their cars when the headlights and steering wheels are customised as well. Including bundle plans for the sale of customisation services for these components could boost sales.
3. Price is still the most significant factor in deciding to buy a car for most customers. These customers tend to be influenced by brand name as well, suggesting that brand reputation plays a significant role in shaping consumer preferences within the automotive market. Manufacturers of renowned automotive brands can leverage this by focusing on the development of more products under these brand names to appeal to growing consumer demand for their products.
4. Young, non-car owners are the most likely to spend a higher amount on future car customisation. Focusing on providing the necessary professional guidance in car customisation can improve the sales conversion rate for this target group.

Conclusions drawn from this dataset can be verified and extended by conducting further surveys with a more expansive and diverse population.

# References

|  |  |
| --- | --- |
| [1] | E. Ostertagová, “Modelling using polynomial regression,” *Procedia Eng.*, vol. 48, pp. 500–506, 2012. |
| [2] | M. Awad and R. Khanna, “Support Vector Regression,” in *Efficient Learning Machines*, Berkeley, CA: Apress, 2015, pp. 67–80. |
| [3] | M. N. Murty and R. Raghava, “Kernel-Based SVM,” in *Support Vector Machines and Perceptrons*, Cham: Springer International Publishing, 2016, pp. 57–67. |
| [4] | I. T. Jolliffe and J. Cadima, “Principal component analysis: a review and recent developments,” *Philos. Trans. A Math. Phys. Eng. Sci.*, vol. 374, no. 2065, p. 20150202, 2016. |
| [5] | S. Mishra *et al.*, “Principal Component Analysis,” *Int. J. Livest. Res.*, p. 1, 2017. |
| [6] | K. Kamalja and N. V. Khangar, “Multiple Correspondence Analysis and its applications,” *Electronic Journal of Applied Statistical Analysis*, vol. 10, no. 2, pp. 432–462, 2017. |
| [7] | C. Shi, B. Wei, S. Wei, W. Wang, H. Liu, and J. Liu, “A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm,” *EURASIP J. Wirel. Commun. Netw.*, vol. 2021, no. 1, 2021. |
| [8] | A. Sahu, M. Dhakar, and P. Rani, “Comparative analysis of apriori algorithm based on association rule,” *Csjournals.com*. [Online]. Available: https://www.csjournals.com/IJCSC/PDF6-2/6.%20Sahu.pdf. |
| [9] | A. Soni, A. Saxena, and P. Bajaj, “A methodological approach for mining the user requirements using Apriori algorithm,” *J. Cases Inf. Technol.*, vol. 22, no. 4, pp. 1–30, 2020. |