

**School of InfoComm Technology**

**Applied Analytics Assignment**

Diploma in Cybersecurity & Digital Forensics

Diploma in Data Science

Diploma in Information Technology

Year 2/3 (2023/2024), Semester 3/5

**INDIVIDUAL ASSIGNMENT 1**

(30% of Applied Analytics Module)

# Deadline for Submission:

**10th Jun 2023 (Saturday), 23:59 HRS**

|  |  |
| --- | --- |
| Tutorial Group: | P03 |
| Student Name: | Huang Wen |
| Student Number: | S10241971 |

**Penalty for late submission:**

10% of the marks will be deducted every day after the deadline.

**NO** submission will be accepted after 17th Jun 2023, 23:59.

1. Summary / Overview

For this assignment, I received a dataset called “Used\_Car\_Price.csv” containing categorical and numerical data types. However, for this assignment, I focus only on the numerical data types for my cluster analysis. The goal of cluster analysis is to discover groups of similar observations and derive business insights and solutions that are purposeful to the car reseller.

To begin, I created a buyer persona and extract the cars that would be suitable for this persona. The objective was to help reseller to identify used cars with similar features to attract individuals with similar needs as the persona, such as those looking for family-friendly cars suitable for trips and neighbouring countries like Malaysia. The persona may be concerned about the interior size, safety features, child-friendly features, cargo space, fuel efficiency, connectivity, comfort, and price of the car.

Next, I performed data exploration to gain a better understanding of the dataset. I selected the variables that I would use for cluster analysis, starting off with 3 variables and gradually adding 1 variable until I hit a total of 5 variables.

Lastly, I applied K-Means and Hierarchical clustering models on datasets with 3, 4 and 5 selected variables.

I further interpret the best models to identify the cluster that prominently exhibited family-friendly cars and provided business insights and solutions to assist the car reseller. The selection of the best model was based on the cluster quality(silhouette score), as well as the number of insights generated by the model(cluster interpretability).

This data set contains 2059 rows for each variable with a total of 9 different numerical data variables. After extracting the cars that are suitable for my persona, there are only 843 rows left. There was no missing value inside the 843 rows, hence I do not need to replace the empty values with mean values. Additionally, the data were scaled using the StandardScaler() method. Scaling was necessary because variables like years(manufacturing years) were represented in four digits, while variables like kilometres (mileage) were in seven digits. Scaling ensured that all variables carried equal importance in the analysis.

The variables I selected for the persona are price, kilometer, engine capacity, fuel tank capacity and volume for the persona. Price is important as buyers have a spending budget. Kilometer(mileage) is related to the wear and tear of the car hence it is important as we can use it as a metric for safety. Volume is used as a measure of the spaciousness of the car. Lastly, engine capacity is to measure the power and fuel efficiency (lower engine capacity, higher fuel tank capacity) of the car and fuel tank capacity to see how long it can travel.

Ps(not included in word count), for lux(jupyter notebook scatter plot) look at the end of the report.

1. Building the clustering models
   1. Data Exploration

A screenshot of a computer screen

Description automatically generated with low confidence

Figure 1 Import function.

Here is the list of packages I imported for this assignment.

A picture containing text, screenshot, font

Description automatically generated

Figure 2 Data extraction, filtering and adding new columns.

I extracted the dataset from the CSV file and filter out the cars that are suitable for my persona.

Firstly, to comply with the Certificate of Entitlement (COE) regulations in Singapore, which limits car lifespan to ten years, I filtered the manufacturing year to be later or equal to 2016. This ensures that the selected cars can be used for at least three more years (2023 - 2016 = 7, and 10 - 7 = 3).

Secondly, the persona is primarily interested in spacious and family-friendly cars, with SUVs being a preferred choice. SUVs are generally taller than other car types and have a seating capacity of 5 or more. Hence, I filtered the car height to be equal to or higher than 1500mm and the seating capacity to be larger or equal to 5.

Thirdly, I applied a price filter to include cars within a specific price range so that they will be affordable and align with the persona’s budget.

Additionally, I filter kilometer, engine capacity and fuel tank capacity so that there will not have extreme values, as my persona does not need them. The persona does not prefer high kilometer(mileage) cars, as it is related to wear and tear of the car and could adversely impact the safety of the car.

Lastly, I created a new column called volume, it used as a metric for measuring the spaciousness of the car. It is calculated based on the length, width, and height of the car.



A picture containing text, screenshot, font, number

Description automatically generated

Figure 3 describe function.

I use the describe functions to show me the statistic of each numeric data, such as mean and standard deviation.

A screenshot of a computer

Description automatically generated

Figure 4 Info function

Next, I use the info function to check for any empty values in each column. All the columns have 843 rows, except for DRIVETRAIN with 840. I will not do any handling of missing values as all numeric data has no empty values.

**Further Exploration for each column**

Using the info I collected, I decide to use the following 6 variables for my first persona.

1. Price 🡪 Price of the car
2. Volume 🡪 Spaciousness of the car
3. Engine Capacity 🡪 Engine Capacity in CC (power and performance of car)
4. Fuel Tank Capacity 🡪 Fuel Tank Capacity in Litres
5. Kilometer(mileage)🡪 Mileage in KM

I will build a cluster model for 3, 4, and 5 variables. I will perform further exploration of the variables too.

Next, I will move on to further exploration of each numeric data column. The main functions I will use are boxplot, histogram, and frequency table.

The boxplot is a visualization to help me see the central tendency, spread, and any potential outliers of the variable. The histogram and frequency can show me how the data is spread across different numerical intervals.

A picture containing screenshot, text, diagram, line

Description automatically generated

A picture containing text, diagram, plot, screenshot

Description automatically generated

Figure 5 Boxplot and Histogram for Price

The boxplot tells me that the medium value is between 25000. Any values outside of around 87500 are considered outliers. Some of the outliers are very far from the medium (some more than 175000) and the minimum value is closer to the medium than the maximum value.

From the histogram, I can see that price is dominated by values before 25000, while data more than 2500 has low frequency.

A picture containing screenshot, text, diagram, line

Description automatically generated

A picture containing text, diagram, screenshot, plot

Description automatically generated

Figure 7 Boxplot and Histogram for Kilometer

The boxplot shows me the medium is around 40000. Any value after 120000 is considered an outliner.

The histogram showed me that the data is mainly concentrated around 40000.

A picture containing screenshot, text, rectangle, diagram

Description automatically generated

A picture containing text, screenshot, diagram, plot

Description automatically generated

Figure 8 Boxplot and Histogram for Engine capacity

The boxplot shows me the medium is around 1500. No outliner for engine capacity.

The histogram showed me that the data is mainly concentrated around 1200.

A picture containing screenshot, text, diagram, rectangle

Description automatically generated

A picture containing text, diagram, screenshot, plot

Description automatically generated

Figure 9 Boxplot and Histogram for Fuel Tank Capacity

The boxplot shows me the medium is around 50. There are no outliers for fuel tank capacity.

The histogram showed me that the data is mainly concentrated around 40 to 60.

A picture containing screenshot, rectangle, line, diagram

Description automatically generated

A picture containing text, diagram, screenshot, plot

Description automatically generated

Figure 10 Boxplot and Histogram for Volume

The boxplot shows me the medium is around 12. There are no outliers for volume.

The histogram showed me that the data is mainly concentrated around 10.

Next, I will do scatter plot price against other variables to find the trend of each data against prices.

A picture containing text, font, screenshot, white

Description automatically generated

A picture containing text, screenshot, diagram

Description automatically generated

Figure 12 Scatter plot for price and Kilometer

Most of the cars are concentrated at low price and low kilometer.

A picture containing text, font, screenshot, line

Description automatically generated

A picture containing text, screenshot, diagram, font

Description automatically generated

Figure 13 Scatter plot for price and Engine Capacity

The price of the car is low when engine capacity is below 2000. When the engine capacity is above 2000, the price varies.

A white card with red text

Description automatically generated with low confidence

A picture containing text, diagram, screenshot, line

Description automatically generated

Figure 14 Scatter plot for price and Fuel Tank Capacity

The price of the car is low when the fuel tank capacity is below 50. When the fuel tank capacity is above 50, the price varies.

A picture containing text, font, screenshot, line

Description automatically generated

A picture containing text, diagram, line, plot

Description automatically generated

Figure 15 Scatter plot for price and Volume

The price of the car is low when volume is below 12. When the volume is above 12, the price varies.

* 1. Clustering for 3 variables

**K-means for 3 variable**

A screenshot of a computer program

Description automatically generated with low confidence

A graph with green and blue lines

Description automatically generated with low confidence

Figure 16 Finding optimal number of clusters of 3 variable (for kmeans)

From the yellow brick, we can see that the optimal cluster is 3. Next, I will plot a 3D graph using a cluster size of 3.

A screen shot of a computer screen

Description automatically generated with low confidence

Figure 17 3D plot for 3 variables(k-means)

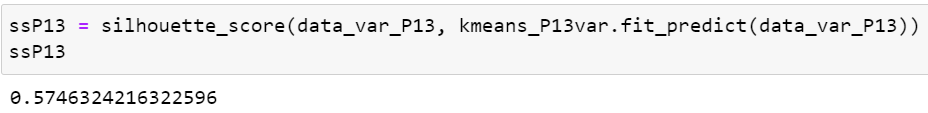
**

Figure 18 Silhouette Score for 3 variables (k-means)

*A screenshot of a computer code

Description automatically generated with medium confidenceA picture containing text, diagram, line, technical drawing

Description automatically generatedA picture containing diagram, text, line, technical drawing

Description automatically generated*

*A picture containing text, diagram, line, technical drawing

Description automatically generated*

Figure 19 Boxplot for 3 variable (k-means)

From the boxplot,

Cluster 0 has cars with low prices (around $50000), medium volume (around 15m3 – 16m3), and medium engine capacity(around 2200cc).

Cluster 1 has cars with very low prices (below $25000), low volume (around 10m3), and low engine capacity(around 1200cc).

Cluster 2 has cars with medium prices (around $110000), medium volume (around 15m3), and medium engine capacity (around 2000cc).

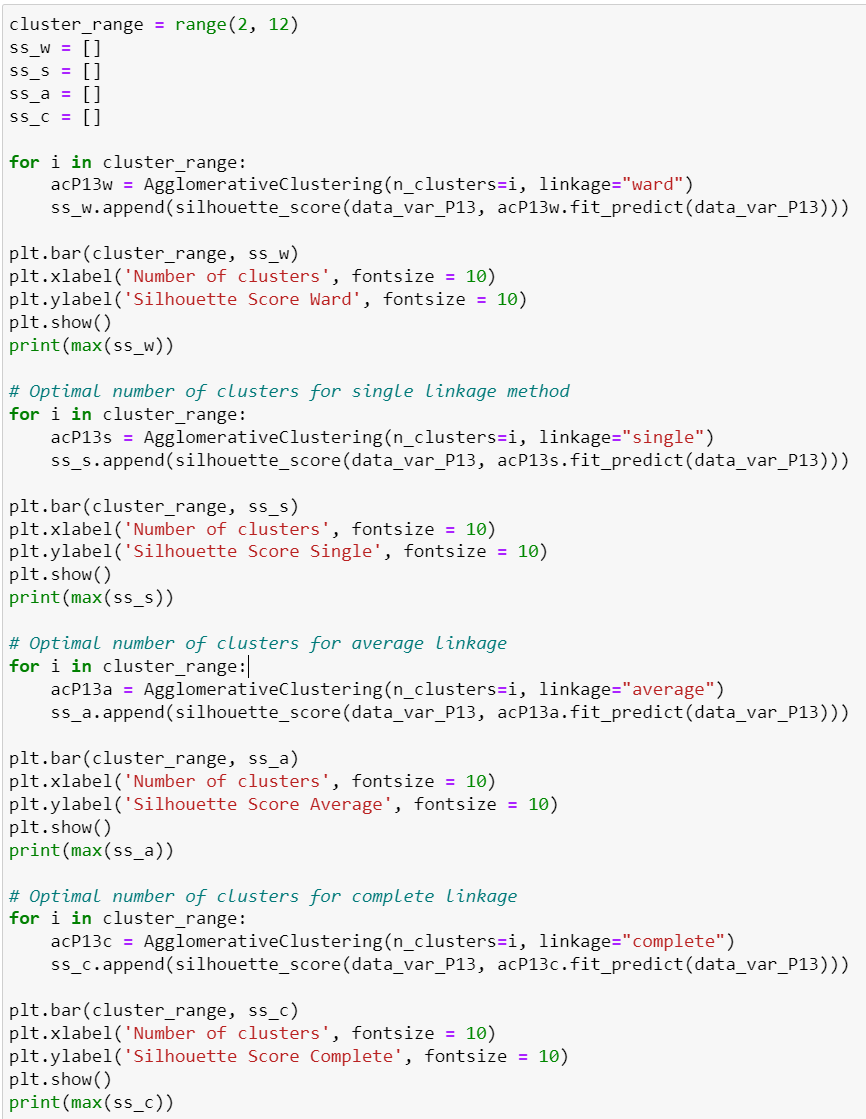
The best cluster is 0, it exhibits family-car-friendly features most prominently. The cluster offers cars that are affordable prices, spacious and has enough engine capacity. Cluster 1 is rejected as it has a small space in the car, which may not be suitable for the persona, while Cluster 2 price is high, but the volume and engine capacity is about the same as cluster 0, hence it is not recommended for the persona to pay more for little improvements.

Next, I will perform hierarchical clustering on these 3 variables.

**Hierarchical Clustering for 3 variable**

A screenshot of a computer program

Description automatically generated with medium confidence



A picture containing text, diagram, screenshot, rectangle

Description automatically generatedA picture containing text, screenshot, plot, line

Description automatically generatedA picture containing text, screenshot, diagram, rectangle

Description automatically generatedA picture containing text, screenshot, plot, diagram

Description automatically generated

A picture containing text, screenshot, diagram, line

Description automatically generatedA picture containing text, screenshot, plot, line

Description automatically generatedA picture containing diagram, text, screenshot, rectangle

Description automatically generatedA picture containing text, screenshot, plot, line

Description automatically generated**Silhouette score for ward**

A number on a white background

Description automatically generated with low confidence

**Silhouette score for single**



**Silhouette score for average**

****

**Silhouette score for complete**

****

From the above graphs, the optimal number of clusters is 2 for all linkage except for the Average of 4. The highest Silhouette score (cluster quality) is Ward linkage (0.6077637763317943).

**A screen shot of a graph

Description automatically generated with low confidence**

Figure 20 3D graph for 3 variables (K-means)

A screenshot of a graph

Description automatically generated with medium confidence

A picture containing text, diagram, line, parallel

Description automatically generatedA picture containing text, diagram, line, parallel

Description automatically generated

Figure 21 Boxplot for 3 variables (Hierarchical)

From the boxplot,

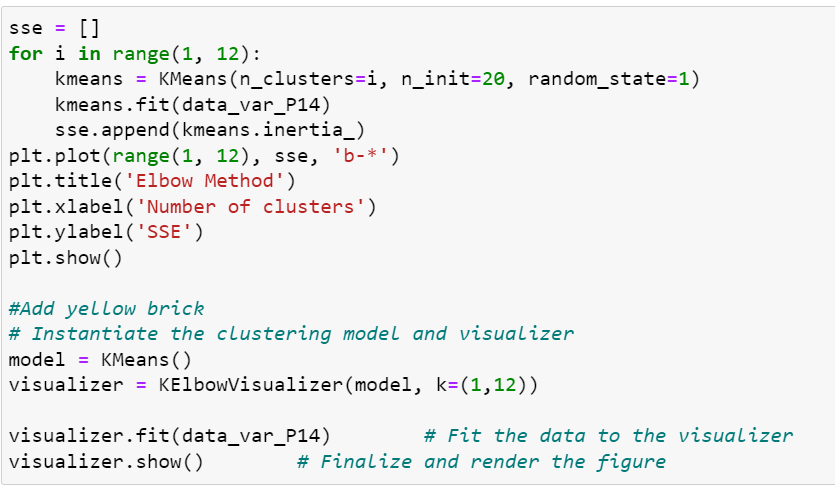
Cluster 0 has cars with low prices (around $50000-$60000), high volumes(around 15m3 – 16m3), and medium engine capacity(around 2100cc).

Cluster 1 has cars with very low prices (below $25000), low volume(around 11m3), and low engine capacity(around 1600cc).

The best cluster is 0, it exhibits family-car-friendly features most prominently. The cluster offers cars that are affordable prices, very spacious and has enough engine capacity. Cluster 1 is rejected as it has a small space in the car, which may not be suitable for the persona.

* 1. Clustering for 4 variables

**K-means for 4 variable**



A picture containing text, line, diagram, plot

Description automatically generated

From the yellow brick, we can see that the optimal cluster is 3. Since it has 4 variables, I cannot use a 3D graph or scatter plot to show the cluster, we will just move on to plotting the boxplot instead.

A picture containing text, screenshot, font, white

Description automatically generated

A screenshot of a computer code

Description automatically generated with low confidenceA picture containing text, diagram, line, parallel

Description automatically generated A picture containing diagram, line, technical drawing, plan

Description automatically generated

A picture containing text, diagram, line, parallel

Description automatically generatedA picture containing text, diagram, line, plan

Description automatically generated

From the boxplot,

Cluster 0 has cars with low prices(between $25000 to $ 50000), medium volume(around 13m3), medium engine capacity(around 2000cc) and medium fuel capacity(60 litres).

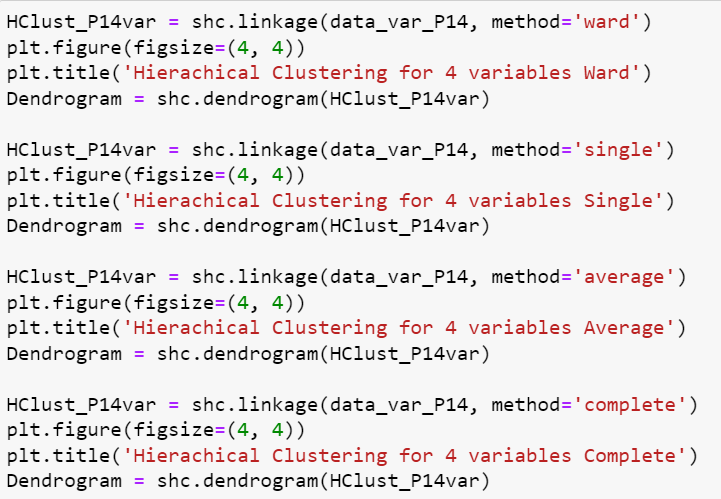
Cluster 1 has cars with very low prices (below $25000), low volume (around 11m3), low engine capacity(around 1300cc), and low fuel tank capacity(around 40 litres).

Cluster 2 has cars with medium prices(around 100000), high volume(around 16m3), medium engine capacity(around 2000cc) and high tank fuel capacity(around 70 litres).

The best cluster is 0, it exhibits family-car-friendly features most prominently. The cluster offers cars that are affordable prices, quite spacious, and have enough engine capacity and enough fuel tank capacity to last long journeys. Cluster 1 is rejected as it has a small space in the car, which may not be suitable for the persona, while cluster 2 is rejected as the paying a few times more than cluster 1 for 10 litres fuel tank capacity increase and 3m3 space increase in the car is not worth it.

Next, I will perform hierarchical clustering on these 4 variables.

**Hierarchical Clustering for 4 variable**

**

*A screenshot of a computer program

Description automatically generated with medium confidence*

*A picture containing text, diagram, screenshot, design

Description automatically generated* A picture containing text, screenshot, plot, diagram

Description automatically generated

*A picture containing text, screenshot, diagram, design

Description automatically generated* A picture containing line, diagram, plot, text

Description automatically generated

A picture containing text, screenshot, diagram, design

Description automatically generated A picture containing text, screenshot, plot, line

Description automatically generated

A picture containing text, diagram, screenshot, design

Description automatically generated A picture containing text, screenshot, plot, line

Description automatically generated

**Silhouette score for ward**



**Silhouette score for single**



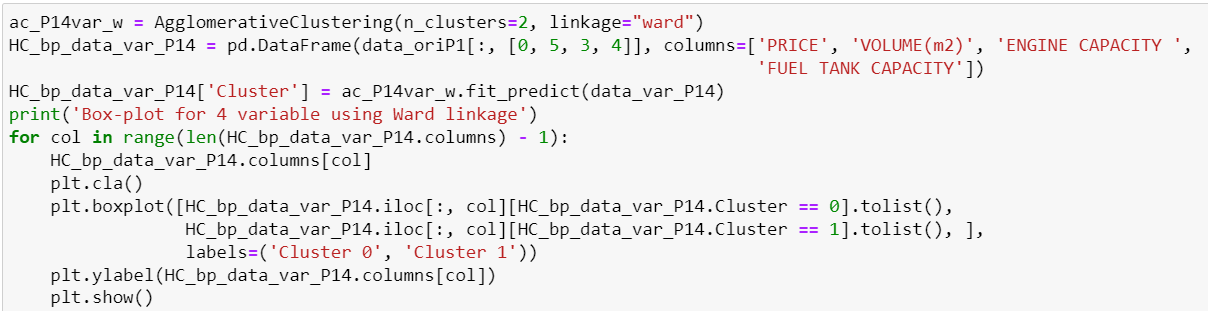
**Silhouette score for average**

****

**Silhouette score for complete**

****

From the above graphs, the optimal number of clusters is 2 for all linkage. The highest Silhouette score (cluster quality) is Ward linkage (0.571469354544218).



A picture containing text, diagram, line, parallel

Description automatically generated A picture containing text, diagram, line, parallel

Description automatically generated

A picture containing text, diagram, line, parallel

Description automatically generated A picture containing text, diagram, line, parallel

Description automatically generated

From the boxplot,

Cluster 0 has cars with low prices (between $50000 to $75000), high volumes (around 16m3), high engine capacity(around 2100cc) and high fuel tank capacity(62 litres).

Cluster 1 has cars with very low prices(below 25000), low volume(around 10m3), low engine capacity(between 1000cc and 1500cc), and low fuel tank capacity(around 45 litres).

The best cluster is 0, it exhibits family-car-friendly features most prominently. The cluster offers cars that are affordable prices, quite spacious, and have enough engine capacity and enough fuel tank capacity to last long journeys. Cluster 1 is rejected as it has a small space in the car, which may not be suitable for the persona.

* 1. Clustering for 5 variables

**K-means for 5 variable**

A screenshot of a computer code

Description automatically generated with medium confidence

A picture containing text, line, plot, diagram

Description automatically generated

From the yellow brick, we can see that the optimal cluster is 3. Since it has 5 variables, we will just plot the boxplot instead.

A screenshot of a computer code

Description automatically generated with low confidence

A screenshot of a graph

Description automatically generated with medium confidence

A picture containing diagram, line, technical drawing, plan

Description automatically generatedA picture containing text, diagram, line, parallel

Description automatically generated

A picture containing text, diagram, line, technical drawing

Description automatically generatedA picture containing text, diagram, line, parallel

Description automatically generated

From the boxplot,

Cluster 0 has cars with very low prices (below $25000), low volumes (around 10m3), low engine capacity(between 1000cc and 1500cc), low fuel tank capacity(around 65 litres) and low kilometer(mileage)(40000 km).

Cluster 1 has cars with medium price(around $100000), high volume(around 16cm3), medium engine capacity(2000cc), high fuel tank capacity(around 65 litres) and low kilometer(mileage)(around 40000km).

Cluster 2 has cars with low prices(between $25000 to $50000), medium volume(around 14m3), medium engine capacity(around 2000), medium fuel tank capacity(60 litres) and low kilometer(mileage)(between 40000km to 50000km).

The best cluster is 2, it exhibits family-car-friendly features most prominently. The cluster offers cars that are affordable prices, quite spacious, have enough engine capacity and enough fuel tank capacity to last long journeys and have low wear and tear. Cluster 0 has less space inside the car, which is not suitable for my persona. Cluster 1 is rejected as it has the price is very high, which is not suitable for my persona.

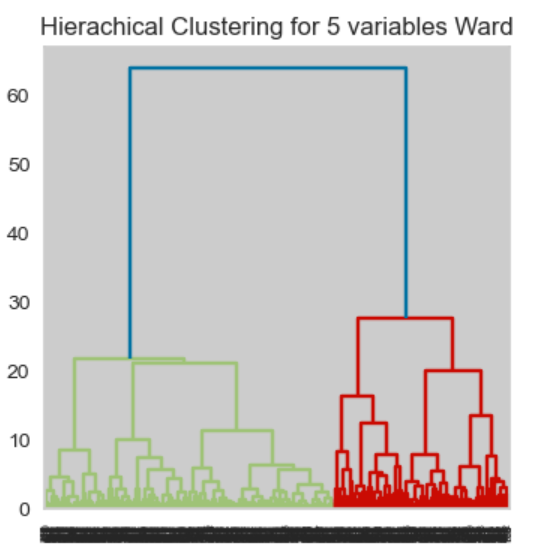
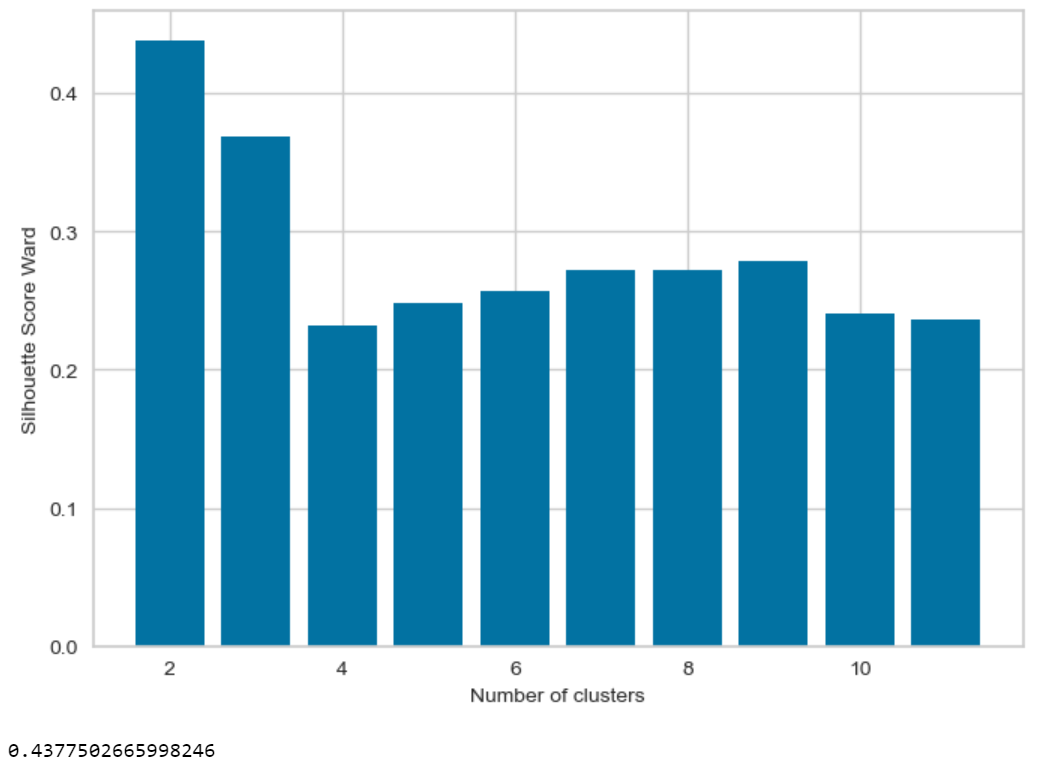
**Hierarchical Clustering for 5 variable**

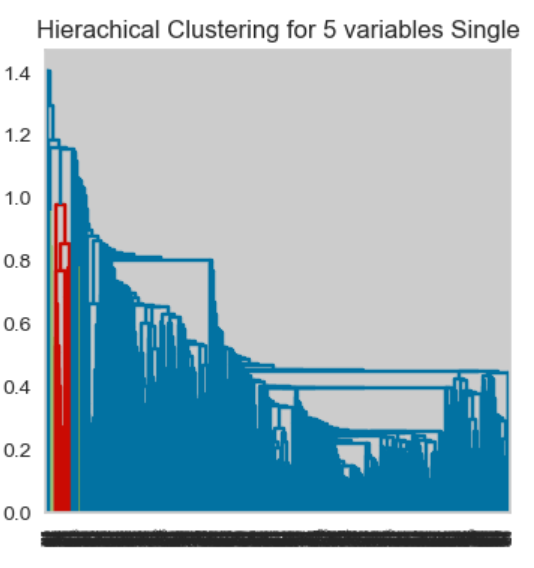
A screenshot of a computer program

Description automatically generated with low confidence

A screenshot of a computer program

Description automatically generated with low confidence

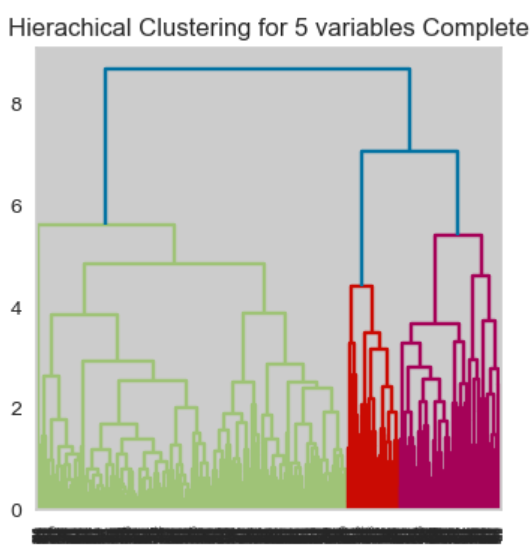
 

A picture containing text, screenshot, diagram, plot

Description automatically generated

 A picture containing text, screenshot, plot, diagram

Description automatically generated

 A picture containing text, screenshot, diagram, plot

Description automatically generated

**Silhouette score for ward**



**Silhouette score for single**



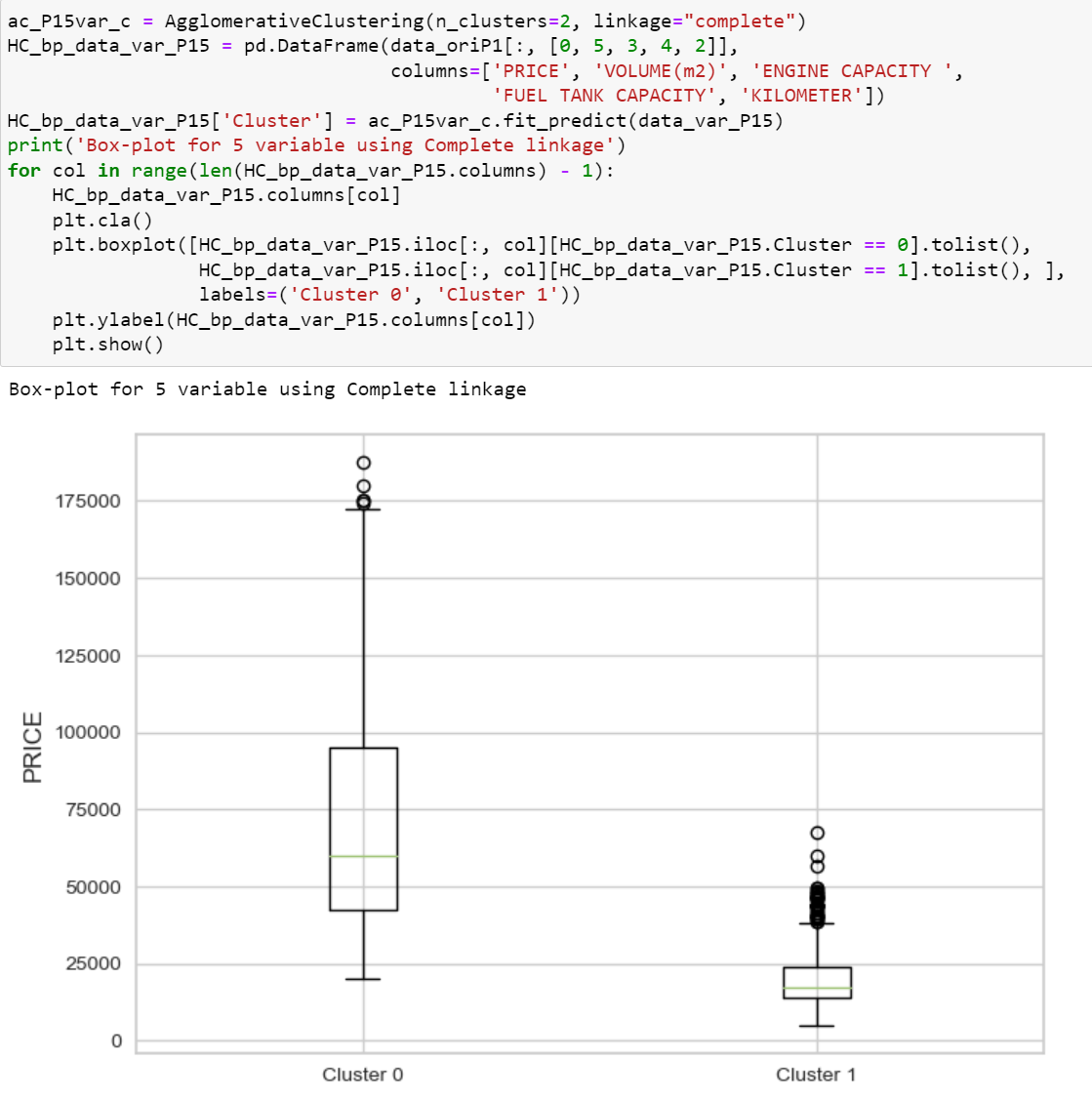
**Silhouette score for average**

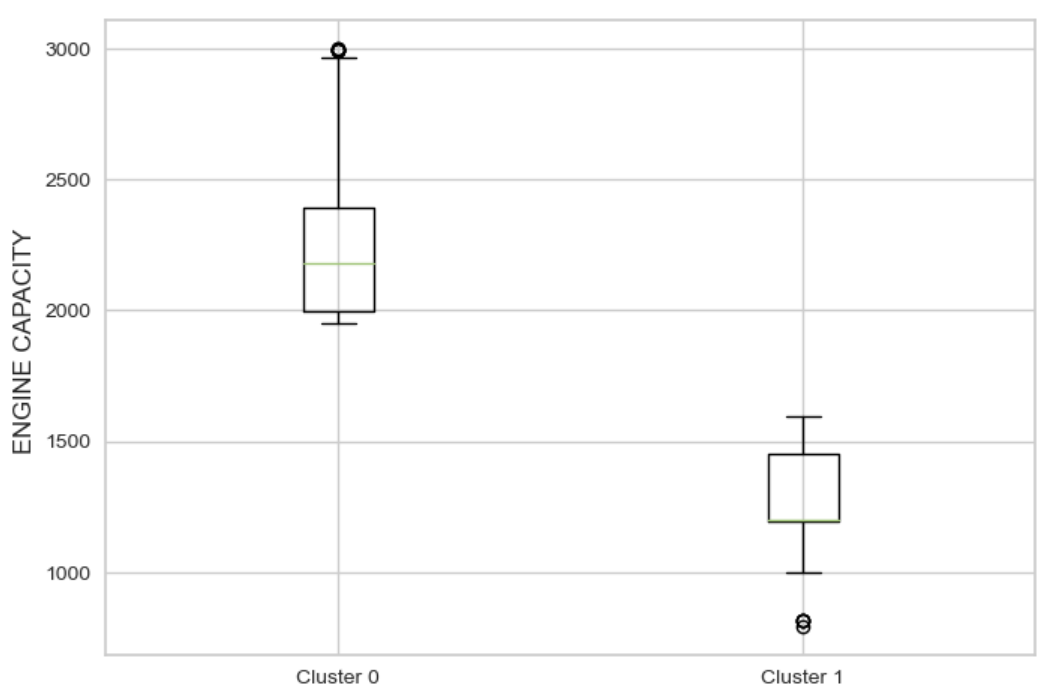
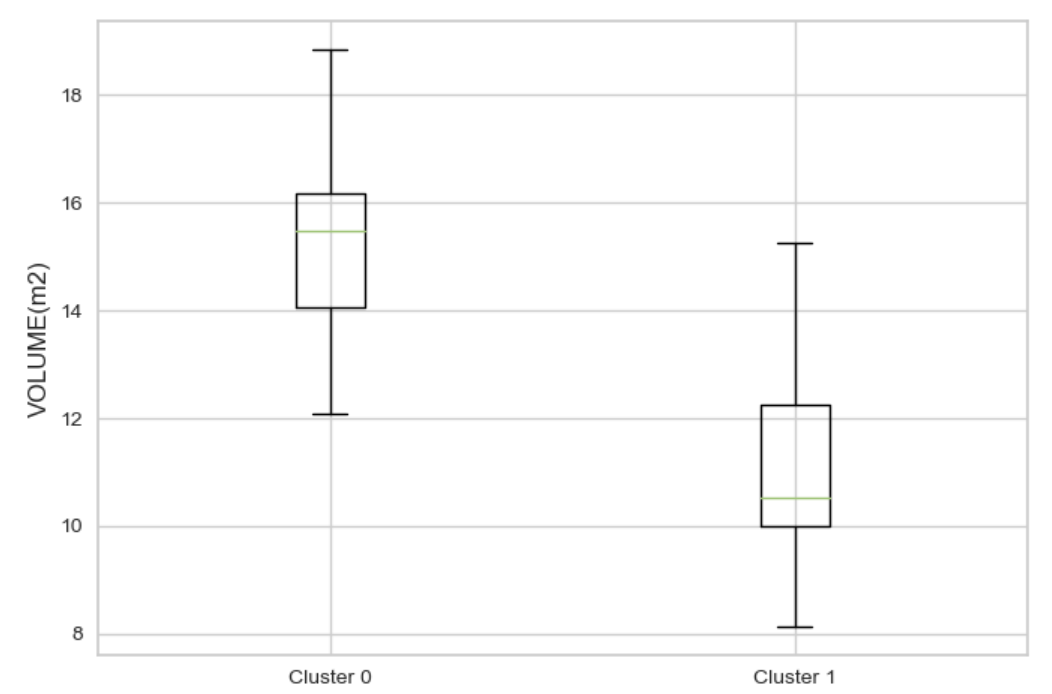


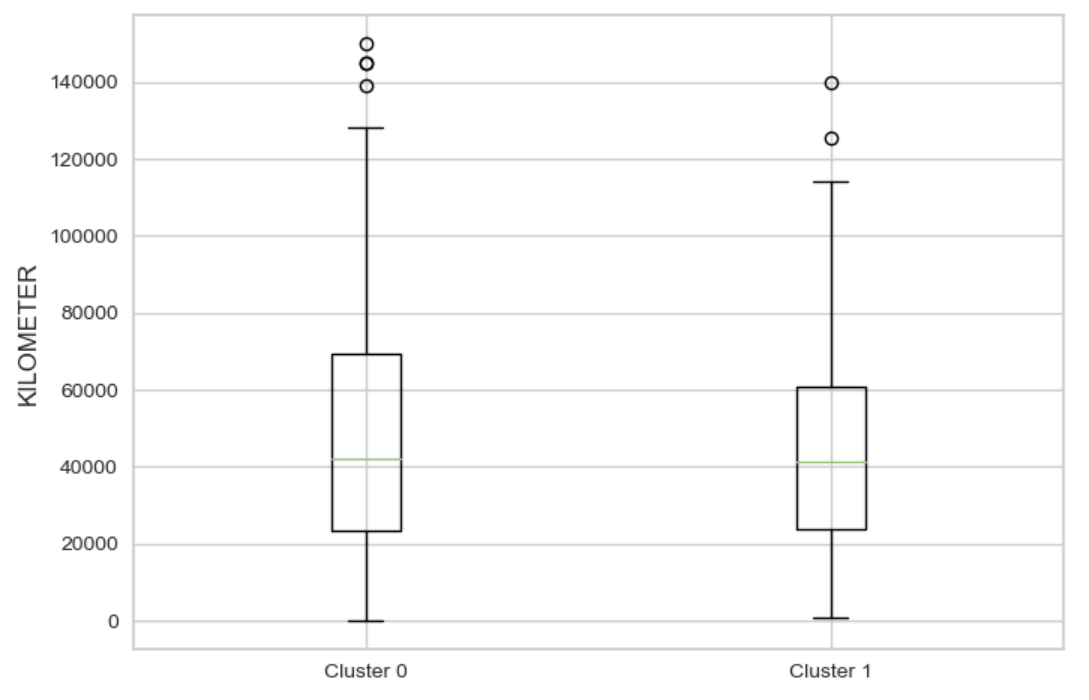
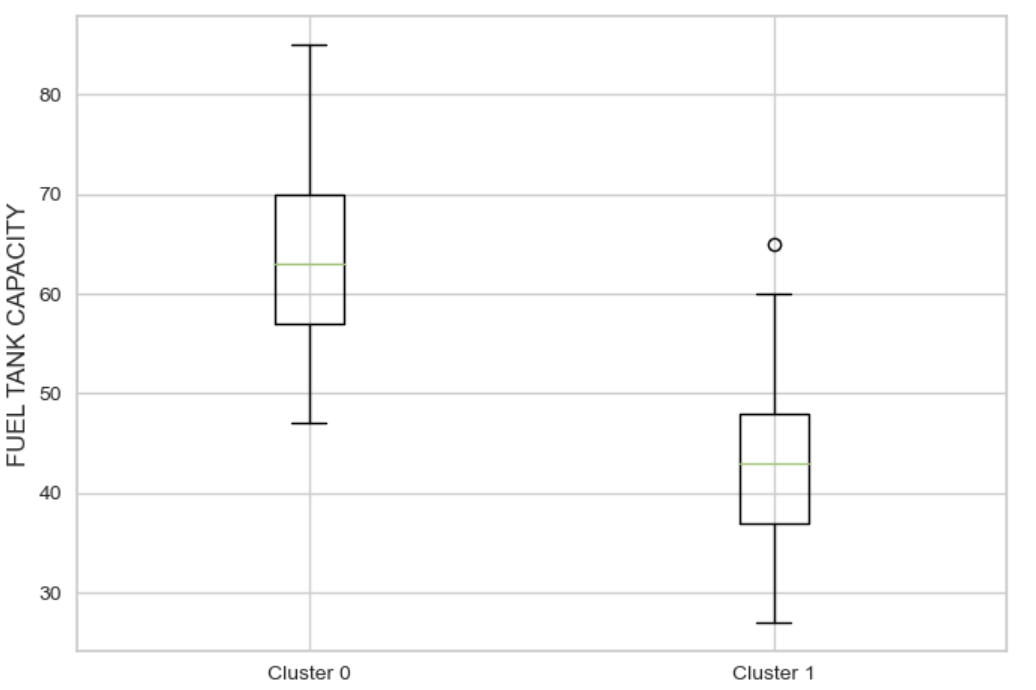
**Silhouette score for complete**



From the above graphs, the optimal number of clusters is 2 for all linkage. The highest Silhouette score (cluster quality) is Complete linkage (0.4711426145784816).

****

****

****

From the boxplot,

Cluster 0 has cars with low prices(between $50000 and $75000), high volumes (around 16m3), high engine capacity (around 2200cc), high fuel tank capacity(around 65 litres) and low kilometer(mileage)(around 40000km).

Cluster 1 has cars with very low prices(below $25000), low volume(around 10m3), low engine capacity(between 1000cc and 1500cc), low fuel tank capacity(around 45litres) and low kilometer(mileage) (around 40000km).

The best cluster is 0, it exhibits family-car friendly features most prominently, the price is affordable, very spacious, has enough engine capacity, enough fuel tank capacity to last long journeys and low wear and tear.

The best cluster is 0, it exhibits family-car-friendly features most prominently. The cluster offers cars that are affordable prices, quite spacious, have enough engine capacity and enough fuel tank capacity to last long journeys and have low wear and tear. Cluster 1 is rejected as it has the volume of the car is small, which is not suitable for my persona as the persona wanted a spacious car.

1. Summary and Interpretation

**Cluster Selected**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model type | Num of cluster | Variable Used | Linkage used | Silhouette score |
| K-means (3) | 3 | Price, Volume and Engine Capacity | NA | 0.5746324216322596 |

|  |  |
| --- | --- |
| Cluster label | Cluster 0: Cheap, strong, and big SUV  Cluster 1: Very cheap, compact, and weak SUV  Cluster 2: Expensive, strong, and big SUV |
| Optimal cluster | The persona should pick cars from cluster 0. Cars from cluster 0 are cheap and provide a good amount of space and engine capacity. Whereas cluster 1 has cars with compact space and small engine capacity and the price difference between cluster 0 and 1 is very small. Cars from Cluster 2 are expensive, but bring little improvement compared to Cluster 0. |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model type | Num of cluster | Variable Used | Linkage used | Silhouette score |
| Hierarchical Model (3) | 2 | Price, Volume and Engine Capacity | Ward | 0.6077637763317943 |

|  |  |
| --- | --- |
| Cluster label | Cluster 0: Cheap, strong, and big SUV  Cluster 1: Very cheap, compact, and weak SUV |
| Optimal cluster | The persona should pick cars from cluster 0. Cars from cluster 0 are cheap and provide a good amount of space and engine capacity. Whereas cluster 1 has cars with compact space and small engine capacity and the price difference between cluster 0 and 1 is very small. |

|  |  |
| --- | --- |
| Comparison between k-means and hierarchical cluster | The optimal cluster that the persona should pick is Hierarchical clustering. The cars from there are a little cheaper compared to K-means, even though the engine capacity is lower by 100cc, it has little impact on the persona. |
| Business Insights for Reseller | This information helps the reseller to focus on acquiring SUVs in a certain price range (between $50000 and $60000), it doesn’t need to be very cheap as the persona has certain requirements to fulfil or has expensive SUVs where there is a little number of buyers able to afford it. This way the reseller can maximize the number of buyers with similar needs as the persona attracted to buy the SUVs. |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model type | Num of cluster | Variable Used | Linkage used | Silhouette score |
| K-means (4) | 3 | Price, Volume, Engine capacity and Fuel tank capacity | NA | 0.4897857803963321 |

|  |  |
| --- | --- |
| Cluster label | Cluster 0: Cheap, strong, long-lasting, and big SUV  Cluster 1: Very cheap, compact, short-lasting, and weak SUV  Cluster 2: Expensive, strong, very long-lasting, and very big SUV |
| Optimal cluster | The persona should pick cars from cluster 0. Cars from cluster 0 are cheap and provide a good amount of space and engine capacity and fuel tank capacity. Whereas cluster 1 has cars with compact space, small engine capacity and fuel tank capacity. Additionally, the price difference between clusters 0 and 1 is very small. Cars from Cluster 2 are very expensive, and the persona does not need so many improvements of the car. |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model type | Num of cluster | Variable Used | Linkage used | Silhouette score |
| Hierarchical Model (4) | 2 | Price, Volume, Engine capacity and Fuel tank capacity | Ward | 0.571469354544218 |

|  |  |
| --- | --- |
| Cluster label | Cluster 0: Cheap, strong, long-lasting, and big SUV  Cluster 1: Very cheap, compact, short-lasting, and weak SUV |
| Optimal cluster | The persona should pick cars from cluster 0. Cars from cluster 0 are cheap and provide a good amount of space, engine capacity and fuel tank capacity. Whereas cluster 1 has cars with compact space, small engine capacity and fuel tank capacity. The price difference between clusters 0 and 1 is very small. Hence, it is better to pick cluster 0. |

|  |  |
| --- | --- |
| Comparison between k-means and hierarchical cluster | The optimal cluster that the persona should pick is k-means clustering. The cars from the hierarchical cluster require $25000 more in price compared to k-means. However, the improvement of the cars in an increase in 3m3 in space, and 100cc and 2-litre increase in engine and fuel tank capacity respectively. The price paid for improvement is not worth it for the persona. |
| Business Insights for Reseller | This information helps the reseller to focus more on the specification when acquiring SUVs. The price paid for improvement of the specification of the SUVs cannot be very high. Buyers will pay for improvements to the specifications of cars only when the price increase is reasonable. Hence, the reseller needs to take note of it when acquiring SUVs to prevent a situation where the buyers do not buy improved specifications of cars as the price is very high for improvement. This can lead to storage pressure for resellers and incur losses for them. |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model type | Num of cluster | Variable Used | Linkage used | Silhouette score |
| K-means (5) | 3 | Price, Volume, Engine capacity, Fuel tank capacity and Kilometer(mileage) | NA | 0.3767088989371512 |

|  |  |
| --- | --- |
| Cluster label | Cluster 0: Very cheap, weak, short-lasting, low wear and tear, and compact SUV  Cluster 1: Expensive, strong, very long-lasting, low wear and tear, and very big SUV  Cluster 2: Cheap, strong, long-lasting, low wear and tear, and big SUV |
| Optimal cluster | The persona should pick cars from cluster 2. Cars from cluster 0 are cheap and provide a good amount of space and engine capacity and fuel tank capacity. Whereas cluster 0 has cars with compact space, small engine capacity and fuel tank capacity. Additionally, the price difference between clusters 0 and 1 is very small. Cars from Cluster 2 are very expensive, and the persona does not need so many improvements of the car. |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model type | Num of cluster | Variable Used | Linkage used | Silhouette score |
| Hierarchical Model (5) | 2 | Price, Volume, Engine capacity, Fuel tank capacity and Kilometer(mileage) | Complete | 0.4711426145784816 |

|  |  |
| --- | --- |
| Cluster label | Cluster 0: Cheap, strong, long-lasting, low wear and tear, and big SUV  Cluster 1: Very cheap, compact, short-lasting, low wear and tear, and weak SUV |
| Optimal cluster | The persona should pick cars from cluster 0. Cars from cluster 0 are cheap and provide a good amount of space, engine capacity and fuel tank capacity. Whereas cluster 1 has cars with compact space, small engine capacity and fuel tank capacity. The price difference between clusters 0 and 1 is very small. Hence, it is better to pick cluster 0. The kilometre (mileage) is not taken into consideration as both are close results. |

|  |  |
| --- | --- |
| Comparison between k-means and hierarchical cluster | The optimal cluster that the persona should pick is k-means clustering. The cars from the hierarchical cluster require $25000 more in price compared to k-means. However, the improvement of the cars is an increase in 2m3 in space, a 200cc and 5-litre increase in engine and fuel tank capacity respectively and a reduction of 10000km for kilometer. The price paid for improvement is not worth it for the persona. |
| Business Insights for Reseller | Resellers need to focus on the attributes of SUVs acquired so that it meets the demand of the customer. Price should not be the only gauge, need to consider other features of the SUVs, does it meet the customer’s needs and so on. It is very important to balance the price of the car and the specification of the SUVs. This way, the reseller has SUVs that have affordable prices and required specifications, meeting their demand. |

1. Reflection

To improve my analysis, I could have utilized more features from the dataset to gain a deeper understanding of how different attributes impact car prices. For instance, incorporating variables like power and torque would have provided more insights into the pricing dynamics and helped identify potential outliers in the dataset. This aligns with the preprocessing stage, where handling missing values, outliers, and transforming variables is crucial.

In handling outliers, I could have employed different techniques to address them effectively. Methods such as Winsorization, which involves capping extreme values or replacing them with less extreme values, could have been applied after identifying whether the outliers were valid or erroneous data points. Additionally, transforming the data using mathematical functions like logarithmic or square root transformations, or using advanced techniques like Box-Cox transformations, could have helped normalize the data and reduce the influence of outliers. Binning, which involves grouping data into intervals or assigning outliers to separate clusters, could also have been explored to mitigate the impact of outliers on the clustering process.

By experimenting with different outlier handling techniques, I could have examined how various methods produce different results, rather than relying on a single approach. This iterative process would have allowed me to make more informed decisions about outlier treatment and explore the effects on the clustering outcomes.

Throughout this assignment, I learned the importance of data preprocessing, including the use of box plots to identify outliers and visualization tools like Lux to gain a clearer understanding of the data distribution. Undertaking these preprocessing steps is essential as even a single outlier can significantly affect the results of k-means and hierarchical clustering. Outliers can lead to the formation of low-quality clusters and low cluster interpretability, undermining the information derived from the clustering process.

In summary, incorporating more features, trying different outlier handling techniques, and understanding the impact of outliers on clustering outcomes would have enhanced the analysis process. Preprocessing the data diligently lays the foundation for reliable and meaningful cluster analysis.

Ps(not counted for word count). If there is trouble toggling with lux, re-run it(one by one), then press the button again, and it will appear. If you don’t see any scatter plot, hover over it to show.