

**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: | P04 |
| Student Name: | Tran Huy Minh |
| Student Number: | S10223485H |

**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.

Contents

[1. Summary/Overview 3](#_Toc137308636)

[2. Build Clustering Models using Numerical Data 4](#_Toc137308637)

[3. Summary and Interpretation 7](#_Toc137308638)

[4. Reflection 11](#_Toc137308639)

# Summary/Overview

This report presents a comprehensive cluster analysis of a dataset consisting of used car transactions. The objective of this analysis is to uncover the key factors that influence sales, pricing, and demand for used cars in the market. The analysis aims to provide valuable insights and recommendations to used car resellers, facilitating the optimization of their revenue through the incorporation of the findings into their marketing, purchasing, and pricing strategies.

The dataset used in this analysis comprises 2059 records and encompasses 19 features, including details such as Make, Model, Price, Year, Kilometer, Fuel Type, Transmission, Color, Owner, Seller Type, Engine (Cc), Max Power, Max Torque, Drivetrain, Length, Width, Height, Seating Capacity, and Fuel Tank Capacity.

To approach this problem systematically, the analysis will be conducted in three distinct steps. Initially, numerical data will undergo clustering and visualization using Python libraries and the Jupyter Notebook environment. Two clustering models, namely the K-means clustering model and the Hierarchical clustering model, will be constructed to identify patterns and groupings within the dataset.

Subsequently, the models will be evaluated using appropriate metrics to assess their performance. The resulting clusters will be meticulously examined and interpreted, leveraging visualization tools to gain a deeper understanding of their characteristics and implications.

In the next step, the clusters will be summarized, interpreted, and reflected upon. Each cluster will be analyzed in detail, highlighting its defining features and presenting potential avenues for further improvement. This process will also involve reflecting on the acquired skills throughout the analysis and identifying areas that may benefit from additional development and refinement.

Through this cluster analysis, the aim is to provide valuable insights into the used car market, enabling resellers to make informed decisions and optimize their strategies. The findings of this analysis have the potential to enhance revenue generation and drive business success in the competitive used car industry.

# Build Clustering Models using Numerical Data

During the initial step of the analysis, the Used Car Price dataset was explored by loading it into Jupyter Notebook. The dataset consists of 2059 rows, and key features such as Price, Year, and Kilometer were examined. Based on the data, it can be inferred that Price represents the last sold price of the car, Year indicates the purchase year, and Kilometer signifies the remaining mileage or the car's travel capability.

A picture containing text, font, screenshot, white

Description automatically generated

Null values were observed in several columns of the dataset, including Max Power, Max Torque, Drivetrain, Length, Width, Height, Seating Capacity, and Fuel Tank Capacity. Among these, only the Length, Width, Height, and Seating Capacity features had a low percentage (3.1%) of null values, making it viable to drop these rows. For the remaining null values, logical imputation was performed to fill in the missing information.

A picture containing text, screenshot, font, number

Description automatically generated

A closer examination revealed a direct correlation between missing values in the Engine Capacity, Max Power, Max Torque, and Fuel Tank Capacity columns and the Fuel Type being Electric. These rows were considered valuable for insights, so they were retained, and the missing values were imputed as 0, allowing their inclusion in the clustering process.

For the remaining missing values without a direct correlation, numerical columns like Engine Capacity and Fuel Tank Capacity were filled with the mean values of their respective columns. On the other hand, categorical columns like Drivetrain, Max Power, and Max Torque were filled with the mode values of their respective columns.

Price

A picture containing plot, screenshot, diagram, line

Description automatically generated

Year

A picture containing diagram, plot, screenshot, line

Description automatically generated

Kilometer

A picture containing screenshot, diagram, plot, line

Description automatically generated

Diagnostic plots were generated for all the numerical columns, including Histograms, Probability Plots, and Box Plots, to visualize the data distribution. The plots indicated that the data followed a normal distribution, except for the presence of outliers. Notably, the Price, Year, and Kilometer columns exhibited significant skewness with outliers that exceeded the predefined threshold.

Upon further examination of these outliers, it was observed that a single row had a Year value of 1988, indicating an extremely old car. Despite the assumption that vintage cars may command higher prices, the actual selling price for this car was relatively low at $32,500. Consequently, capping the Year outliers was deemed appropriate due to their lack of meaningful insights. Similarly, while higher mileage might typically correlate with higher prices, the outlier rows in the Kilometer column did not exhibit exceptionally high prices.

A picture containing text, screenshot, font, number

Description automatically generated

Furthermore, the outliers in the Price column were found to be associated with specific car Makes, where ‘luxury’ cars, such as Porsche and Lamborghini, were priced significantly higher. However, due to the limited representation of luxury cars in the dataset, these high-priced instances were treated as outliers. To address this, the Price, Year, and Kilometer columns were capped using the Gaussian method, which involved setting a threshold based on the standard deviation from the mean.

For the Max Power and Max Torque columns, the data was initially formatted to include the respective units (bhp, rpm, nm), making them categorical columns instead of numerical. To properly utilize this information, the columns were separated into distinct numerical columns that contained the corresponding values. This allowed for easier analysis and handling of different formatting issues and null values, particularly for electric vehicles.

A screenshot of a computer

Description automatically generated with medium confidence

Analyzing the Make column, it became apparent that there was a wide range of car brands represented, with luxury cars having a lower frequency of occurrence. Considering the capping of prices for luxury cars, it was decided to set a threshold for the minimum number of rows required for each car make. The threshold was set at 30, which is considered a universally accepted minimum sample size for normalized data. Car makes with counts below this threshold were grouped together under a new variable called 'Others'.

A picture containing text, screenshot, font, number

Description automatically generated

The dataset included six different fuel types: Diesel, Petrol, CNG, Electric, LPG, and Hybrid, as well as a combination of Petrol and CNG. To handle the multiple fuel types, a one-hot encoding technique was applied, creating separate binary columns for each fuel type.

A white background with black text

Description automatically generated with low confidence

The Owner column represented the current number of owners based on an ordinal order. The column was encoded with ordered ordinal values based on the current order, indicating the number of previous owners. As for the remaining categorical columns, a label encoding technique was used.

After applying these data manipulations, the final dataset consisted of 1995 rows and 26 columns, ready for further analysis and clustering.

The first clustering model implemented in this analysis is the K-means clustering model. The variables selected for clustering were Price, Year, and Max Torque (Nm). These variables were chosen as they are key factors influencing the sales and pricing of used cars. Price directly affects demand, Year indicates age and condition, and Max Torque reflects performance. Analyzing these variables helps understand customer preferences, market segments, and pricing strategies.

To prepare the data for clustering, a MinMax Scaler was applied to scale the selected variables. The MinMax Scaler was chosen over the Standard Scaler to ensure that all variables were scaled to the same range (usually between 0 and 1), preserving the original distribution of the data.

A picture containing text, diagram, line, plot

Description automatically generated

The optimal number of clusters for the K-means model was determined using the elbow method in conjunction with the silhouette score. The elbow method involves plotting the number of clusters against the sum of squared distances within each cluster, and selecting the point where the curve starts to level off. In this case, the silhouette score was used as an additional criterion to assess the quality of the clustering solution.

After analyzing the elbow curve, it was observed that the curve became less steep after the point where the number of clusters reached 4. Therefore, 4 clusters were determined to be the optimal choice for this dataset.

Using the K-means clustering model with 4 clusters, a three-dimensional plot was generated to visualize the clustering results. Since there were three input features in the model (Price, Year, and Max Torque), the resulting clusters were plotted in a three-dimensional space. Each data point was color-coded based on its predicted cluster assignment.

A picture containing text, diagram, screenshot, line

Description automatically generated

The plot revealed that the clusters were well-grouped and distinct, with minimal overlap between them. This indicates that the K-means model successfully identified meaningful patterns and similarities within the dataset. The visual representation of the clusters allows for a clear understanding of how the data points are grouped based on their respective features.

To assess the quality of the clustering solution, the silhouette score was calculated. The silhouette score measures the compactness and separation of the clusters, with values ranging from -1 to 1. In this case, the K-means clustering model achieved a silhouette score of 0.386, indicating a reasonable level of clustering cohesion.

A picture containing text, diagram, plan, screenshot

Description automatically generated

Hierarchical clustering was performed as the second clustering model using the Price, Year, and Max Torque (Nm) variables. This allowed for a direct comparison with the K-means model. The data was scaled, and the hierarchical clustering was conducted using the ward linkage method. The results were visualized using a Dendrogram, which indicated the presence of three prominent clusters in the orange, green, and red colors.

A picture containing diagram, text, line, screenshot

Description automatically generated

The baseline hierarchical clustering model was implemented with three clusters based on the ward linkage. A three-dimensional scatter plot was generated to visualize the clusters. While the clusters appeared reasonably separated, there were some instances of overlapping points within each cluster. The silhouette score for this model was found to be 0.341, indicating a moderate level of cluster quality.

A picture containing text, screenshot, font, number

Description automatically generated

To determine the optimal number of clusters, a range of cluster sizes from 2 to 11 was evaluated using the silhouette score. The analysis revealed that two clusters yielded the highest silhouette score. Subsequently, another hierarchical clustering model was constructed with two clusters using the single linkage method, which considers the distance between the closest points of two clusters. This model achieved a notably improved silhouette score of 0.611.

A picture containing diagram, screenshot, design

Description automatically generated

However, upon examining the scatter plot for this two-cluster model, it became apparent that the distribution of points in each cluster was highly imbalanced. The purple cluster encompassed almost all the data points, while the yellow cluster had very few instances. Consequently, the clustering model was revised to have two clusters using the ward method once again. This adjustment resulted in a silhouette score of 0.515.

A picture containing diagram, text, screenshot, line

Description automatically generated

The subsequent visualization of the revised two-cluster model using a three-dimensional scatter plot showed a more even distribution of points with reduced overlap between clusters. This suggests that the revised model achieved a better separation of the data points into distinct clusters.

In evaluating and comparing the two clustering models, it is observed that the Hierarchical clustering model achieved a higher silhouette score compared to the K-means model. However, despite the higher silhouette score, the K-means model can be considered superior due to several factors.

One advantage of the K-means model is that it exhibits a wider range of different clusters. With the optimal number of clusters determined as four, the K-means model provides more granularity and distinct groupings of observations within the dataset. This allows for a more nuanced understanding of the underlying patterns and similarities in the used car price dataset.

On the other hand, the Hierarchical clustering model only yielded two clusters. While it achieved a higher silhouette score, the limited number of clusters restricts the level of detail and insight that can be obtained from the clustering analysis. The lack of additional clusters in the Hierarchical model may result in the loss of valuable information and hinder a comprehensive understanding of the data.

Therefore, considering both the silhouette score and the interpretability of the results, the K-means model can be deemed more suitable for the used car price dataset. Its ability to identify a greater number of distinct clusters allows for a more comprehensive exploration of the factors influencing used car prices. The K-means model strikes a balance between providing meaningful cluster divisions and maintaining a reasonable silhouette score, making it a valuable tool for uncovering insights in the dataset. As such, K-means model is the preferred model for further exploration.

# Summary and Interpretation

The K-means clustering model yielded four distinct clusters, each exhibiting different characteristics across various features. The analysis of boxplots for each feature within each cluster revealed several insights. Below is the table of findings and the suggested names for each cluster.

**Cluster 1 -** Recent Compact Cars with Mid-Range Pricing

| **Feature** | **Insights** |
| --- | --- |
| Make & Model | Wide variety of vehicles |
| Price | Mid-range pricing |
| Year | More recently purchased cars (2016-2019) |
| Kilometer | Similar spread, moderate mileage |
| Seller Type | Corporate and Individual sellers |
| Engine Capacity | Higher engine capacities |
| Length & Width | Slightly larger cars |
| Fuel Tank Capacity | Slightly larger capacities |
| Max Power (bhp) | Slightly higher horsepower |
| Max Torque (Nm) | Higher Newton-meter values |
| Fuel Types | Diesel, Petrol, and Hybrid |

**Cluster 2 -** High-End Luxury and Performance Cars

| **Feature** | **Insights** |
| --- | --- |
| Make & Model | Wide variety of vehicles |
| Price | Significantly higher pricing |
| Year | More recently purchased cars (2018-2021) |
| Kilometer | Smaller mileage range with outliers closer to the boxplot |
| Seller Type | Corporate and Individual sellers |
| Engine Capacity | Higher engine capacities |
| Length & Width | Slightly larger cars |
| Fuel Tank Capacity | Slightly larger capacities |
| Max Power (bhp) | Slightly higher horsepower |
| Max Torque (Nm) | Higher Newton-meter values |
| Fuel Types | Diesel, Petrol, and Hybrid |

**Cluster 3 -** Budget-Friendly Older Cars with High Mileage

| **Feature** | **Insights** |
| --- | --- |
| Make & Model | Wide variety of vehicles |
| Price | Very low pricing |
| Year | Recently purchased cars (2017-2019) |
| Kilometer | Similar spread, moderate mileage |
| Seller Type | Corporate, Individual, and Commercial Registration (outlier at 0) |
| Engine Capacity | Slightly lower engine capacities |
| Length & Width | Slightly smaller cars |
| Fuel Tank Capacity | Slightly smaller capacities |
| Max Power (bhp) | Slightly lower horsepower |
| Max Torque (Nm) | Lower Newton-meter values |
| Fuel Types | Diesel, Petrol, CNG, and Electric |

**Cluster 4 -** Diverse Range of Mid-Aged Cars with Varied Features

| **Feature** | **Insights** |
| --- | --- |
| Make & Model | Wide variety of vehicles |
| Price | Very low pricing |
| Year | Older cars (2012-2014) |
| Kilometer | Similar spread, moderate mileage |
| Seller Type | Corporate and Individual sellers |
| Engine Capacity | Slightly lower engine capacities |
| Length & Width | Slightly smaller cars |
| Fuel Tank Capacity | Slightly smaller capacities |
| Max Power (bhp) | Slightly lower horsepower |
| Max Torque (Nm) | Lower Newton-meter values |
| Fuel Types | Diesel, Petrol, and LPG |

A picture containing diagram, technical drawing, line, plan

Description automatically generated

Regarding the Make and Model features, there was a wide range of vehicle variety observed within each cluster, indicating that no specific Make or Model was predominantly clustered in any particular group.

A picture containing diagram, text, screenshot, technical drawing

Description automatically generated

In terms of Price, the first cluster exhibited mid-range pricing, while the second cluster showed significantly higher pricing. The third and fourth clusters, on the other hand, had vehicles with notably lower prices.

A picture containing diagram, text, plan, rectangle

Description automatically generated

For the Year feature, the first and second clusters consisted of more recently purchased cars, with an inner quartile range of 2016 to 2019 and 2018 to 2021, respectively. The third cluster also contained recent cars, but with a smaller range of 2017 to 2019. In contrast, the fourth cluster mainly comprised older cars, with the inner quartile range spanning from 2012 to 2014.

A picture containing text, diagram, screenshot, technical drawing

Description automatically generated

Examining the Kilometer column, the first, third, and fourth clusters displayed similar spreads, whereas the second cluster had a smaller range of kilometer mileage, along with outliers closer to the boxplot.

A picture containing text, screenshot, rectangle, line

Description automatically generated

Analyzing the Seller Type feature, it was observed that every cluster had both Corporate and Individual sellers. However, the presence of Commercial Registration was only noted in the third cluster, resulting in it having the sole outlier at 0.

A picture containing text, diagram, screenshot, technical drawing

Description automatically generated

Considering the Engine Capacity feature, the first and second clusters exhibited higher engine capacities, with larger outliers. In contrast, the third and fourth clusters showed slightly lower engine capacities, with a smaller spread of outliers.

A picture containing diagram, plan, technical drawing, rectangle

Description automatically generated

In terms of vehicle dimensions, the Length and Width features showed a correlation, with the cars in the first and second clusters being slightly larger compared to those in the third and fourth clusters. Similarly, the Fuel Tank Capacity feature indicated that the first and second clusters had slightly larger capacities than the third and fourth clusters.

A picture containing text, diagram, screenshot, technical drawing

Description automatically generated

Analyzing the Max Power (bhp) feature, the first and second clusters consisted of cars with slightly higher horsepower, displaying a wider range of outliers. Conversely, the third and fourth clusters exhibited slightly lower horsepower, with a more condensed range of outliers.

A picture containing diagram, text, screenshot, technical drawing

Description automatically generated A picture containing text, diagram, screenshot, plan

Description automatically generated

Regarding the Max Torque (Nm) and (rpm) features, the first and second clusters demonstrated higher Newton-meter values compared to the third and fourth clusters. Similarly, the revolutions per minute for the first and second clusters had a narrower range than the third and fourth clusters.

A picture containing text, screenshot, rectangle

Description automatically generated A picture containing text, screenshot, rectangle, diagram

Description automatically generated

A picture containing text, screenshot, rectangle, frame

Description automatically generated A picture containing text, screenshot, rectangle, display

Description automatically generated

When considering different fuel types, Diesel and Petrol were the predominant choices across all four clusters, with no noteworthy variations. However, the third and fourth clusters exclusively contained cars using CNG as a fuel type, while only the third cluster had vehicles powered by electricity. The fourth cluster uniquely consisted of cars utilizing LPG as a fuel source. Hybrid cars were found in the first and second clusters exclusively.

Finally, features such as Transmission, Color, Owner, Drivetrain, and Seating Capacity did not exhibit any distinctive patterns or noteworthy observations across the clusters.

These findings provide valuable insights into the clustering patterns and characteristics within the dataset, shedding light on the diverse attributes of the used car market.

Cluster 1: "Recent Compact Cars with Mid-Range Pricing"

Interpretation: This cluster represents a group of recently manufactured compact cars that fall within a mid-range pricing category. These cars are likely to appeal to buyers who are looking for newer models that offer a balance between affordability and modern features. The vehicles in this cluster are suitable for individuals or families seeking reliable transportation options without the high price tag associated with luxury or high-performance cars.

Cluster 2: "High-End Luxury and Performance Cars"

Interpretation: This cluster comprises high-end luxury and performance cars, indicating a segment targeted at individuals with a preference for premium vehicles. These cars are likely to have advanced features, powerful engines, and luxurious amenities. The higher pricing associated with this cluster reflects the exclusivity and premium nature of these vehicles, appealing to buyers who prioritize luxury, prestige, and high-performance driving experiences.

Cluster 3: "Budget-Friendly Older Cars with High Mileage"

Interpretation: This cluster represents budget-friendly options consisting of older cars with higher mileage. These vehicles are likely to have been on the road for a longer period and may have accumulated significant mileage. The lower pricing in this cluster reflects the age, condition, and higher usage of these cars. They are suitable for buyers on a tighter budget who prioritize affordability over newer models or lower mileage.

Cluster 4: "Diverse Range of Mid-Aged Cars with Varied Features"

Interpretation: Cluster 4 represents a diverse range of mid-aged cars with varied features. This cluster encompasses cars that fall within a mid-age range, offering a mix of specifications, features, and pricing. Buyers within this cluster have a wide range of choices and can select vehicles based on their specific preferences and requirements. This cluster provides options for buyers who seek different features, such as varying engine power, fuel types, sizes, and overall specifications.

# Reflection

Throughout this analysis, I have gained valuable insights and identified areas for improvement in the current solution. One aspect that could be enhanced is the data collection process. By incorporating a more diverse and comprehensive dataset, including variables such as vehicle condition, maintenance history, and specific features influencing customer preferences and pricing dynamics, we can achieve a deeper understanding of the used car market and make more accurate decisions as resellers.

The clustering techniques applied, namely K-means and Hierarchical clustering, have provided valuable insights into the dataset. However, there is room for exploration of alternative clustering algorithms, such as DBSCAN or Gaussian Mixture Models, which may reveal different patterns and shed new light on the data. By employing these techniques, we can potentially uncover hidden subgroups and patterns within the used car market, offering further insights to resellers.

Incorporating external data sources is another avenue for improvement. By integrating market trends, competitor pricing information, and customer reviews into the analysis, we can enhance our understanding of the used car market. This would lead to more accurate clustering results and better-informed decision-making for resellers.

Reflecting on the learning objectives of this analysis, I have successfully applied data analytics in the context of customer segmentation and clustering within the used car market. To further enhance the analysis, I could explore the use of advanced analytics techniques, such as association rule mining on organizational data, to identify valuable patterns and relationships between car features, customer preferences, and pricing strategies. Additionally, performing text mining on customer reviews and feedback would provide deeper insights to improve customer satisfaction and tailor marketing strategies accordingly.

I have developed proficiency in data preprocessing, clustering algorithms, and data visualization throughout this analysis. However, there is room for improvement in predictive modeling techniques, such as regression or classification algorithms, to enhance the predictive capabilities of the analysis and derive more actionable insights for business decision-making.

In conclusion, this analysis has demonstrated the practical application of data analytics in the used car market. By incorporating the suggested improvements and continuing to develop predictive modeling skills, I can conduct more robust analyses in the future. This will empower me, as a used car reseller, to make data-driven decisions and gain a competitive advantage in the market.