

**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: | P01 |
| Student Name: | Isaac Khoo Wu Qin |
| Student Number: | S10244252C |

**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.

Table of Contents

[Deadline for Submission: 1](#_Toc137334019)

[Summary/Overview 3](#_Toc137334020)

[Building Cluster Models 3](#_Toc137334021)

[Data exploration and manipulation 3](#_Toc137334022)

[**Importing Data** 3](#_Toc137334023)

[**Understanding the Data** 4](#_Toc137334024)

[**Converting to NumPy Arrays and showing them** 5](#_Toc137334025)

[**Identifying non-numerical values and converting them to numerical values.** 5](#_Toc137334026)

[**Preparing and scaling the data** 6](#_Toc137334027)

[K-Means Clustering 6](#_Toc137334028)

[K-Means Analysis 10](#_Toc137334029)

[Hierarchical Clustering 10](#_Toc137334030)

[Hierarchical Clustering Analysis 17](#_Toc137334031)

[Evaluation and comparing of models 17](#_Toc137334032)

[Summary and Interpretation 17](#_Toc137334033)

[Reflection 18](#_Toc137334034)

# Summary/Overview

Applied Analytics teaches us to analyze customers, products and services using business analytics techniques to improve the competitiveness of businesses and organizations. Data analytics techniques would be introduced to resolve organization challenges in the area of customer segmentation and text analytics. In this assignment, sometimes we have a group of observations, and we need to split it into a number of subsets of similar observations. I will be using Cluster Analysis which is a group of techniques to help us discover these similarities between the observations. The Objective of this assignment is to use the provided dataset and split it into smaller subsets (clusters) based on similar characteristics. I be utilizing related Python Libraries (through Jupyter Notebook platform) to do this. I am also required to create cluster visualizations to help users explore the clustered data. My task is to analyze the used car data and provide insights on key factors affecting sales, pricing, and demand for used cars in the market, and used the result to advice the used car reseller how they can leverage on the result to improve revenue. For example, to improve on marketing strategy to become more customer-centric, streamline purchasing of used car, pricing strategy.

# Building Cluster Models

## Data exploration and manipulation

### **Importing Data**

Firstly we start by importing the dataset from the files and importing the required packages for later use.

A close-up of a computer screen

Description automatically generated with low confidence

### **Understanding the Data**

Next, we need to understand the data. This can be accomplished by describing and getting info on the data. This will help us to understand the data. What are its values? What are their data columns? What are their Data Type?

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

### **Converting to NumPy Arrays and showing them**

A screenshot of a computer

Description automatically generated with medium confidence

### **Identifying non-numerical values and converting them to numerical values.**

This is done by creating a for loop to loop through each and every one of the values. If the value is not a numerical value, their index will be appended to a list. A function that has a code to assign numerical values to the non-numerical values using the list of indexes to identify the values that need to be changed.

A screenshot of a computer code

Description automatically generated with medium confidence

### **Preparing and scaling the data**

We need to prepare the data again as now that we converted the data to all numerical values, we need to convert the data back into a data frame. Next we need to scale the data frame. A screenshot of a computer

Description automatically generated with medium confidence

## K-Means Clustering

For Kmeans I decided to use the columns “FUEL TANK CAPACITY”, “ENGINE CAPACITY ” and “PRICE”. I removed outlier values before starting to ensure that the outliers did not affect the data distribution. Start by trying to determine the optimal number of clusters, k to be used in my K-Means clustering, I decided to plot a graph using the elbow method to visualize the number clusters k from a range of 1 to 11 against the Sum of squared errors (SSE) value for each value of k. The graph will help me in selecting a suitable value of k which helps strikes a balance between a low sum of squared errors score (SSE) and an a good number of clusters. A picture containing text, font, white

Description automatically generatedA screenshot of a computer

Description automatically generated with medium confidence A picture containing text, diagram, line, plot

Description automatically generated

Next, Based in the graph plotted using the elbow method I have determined the optimal value of clusters where k = 2 and built the model for my K-Means clustering using that value. Where n\_clusters is the number of clusters, n\_init is the Number of time the k-means algorithm will be run with different centroid seeds and max\_iter is the Maximum number of iterations of the k-means algorithm for a single run. After which we will Fit the model and assign each data point to the cluster and predict.

A picture containing text, screenshot, font

Description automatically generated

Next, After I built my model, I then visualized the clusters using in a 3d graph. The graph shows how the data points in the 3 clusters are plotted across the 3 axes that represent the “PRICE”, “ENGINE CAPACITY “ and “FUEL TYPE” columns. A picture containing text, font, screenshot

Description automatically generated A picture containing sketch, screenshot, diagram

Description automatically generated

Lastly, We will Evalute the model using Sum of Squared Errors (SSE) and find the silhouette score to identify how good of a model it is. A screenshot of a computer

Description automatically generated with low confidence

**Clusters formed for price, engine capacity and fuel type:**

**A picture containing text, screenshot, diagram, rectangle

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

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

Description automatically generated

## K-Means Analysis

Based on the 3d graph visualization of my K-Means clusters, the clusters of cars sold are divided by 3 features. Cars with low engine capacity and low fuel tank capacity, cars with high engine capacity and low fuel tank capacity, cars that have high engine capacity and high fuel tank capacity. The clusters produced have a Sum of Squared Errors (SSE) score of 2746.77 (to 2 decimal places) and a Silhouette score of 0.542 (to 3 significant figures). From this visualization, I can conclude that customers are mostly looking for cars that have median engine capacities and fuel tank capacities around the median fuel tank capacity of all cars sold (About 60).

## Hierarchical Clustering

For Hierarchical Clustering I decided to use the columns “ENGINE CAPACITY” and “PRICE”. I removed outlier values before starting to ensure that the outliers did not affect the data distribution.

A picture containing font, text, white, typography

Description automatically generated

Firstly, To find the optimal number of clusters, k, and the optimal linkage method that can be used, I used this code to visualize the dendrogram of the clusters of the methods “ward”, “single”, “average” and “complete”.

A screenshot of a computer program

Description automatically generated with medium confidence A picture containing diagram, text, rectangle, line

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

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

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

Description automatically generated

From the bar graphs plotted, I concluded that the ”ward” linkage method is the most suitable for building the model.

Next, to find the optimal number of clusters, k, I used a code to evaluate the silhouette score of the method “ward” where k ranges from 2-10. The silhouette scores are then placed into a list and are then plotted onto a bar graph to better visualize the silhouette scores over a range of k. A screenshot of a computer

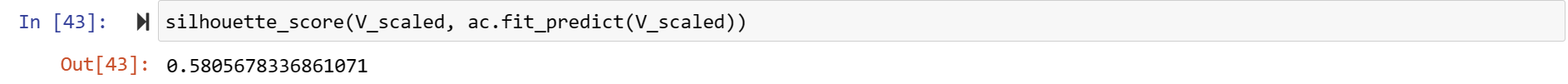
Description automatically generated with medium confidence A picture containing text, screenshot, font, line

Description automatically generated

From the list and the bar chart we are able to identify 6 as the optimal number of k. Now I will be able to build the model since I have concluded that “ward” linkage and k=6 is the most suited.

A screenshot of a computer

Description automatically generated with medium confidence A picture containing text, screenshot, diagram, line

Description automatically generated 

**Clusters formed for Engine Capacity and Price:**

**A picture containing text, screenshot, diagram, rectangle

Description automatically generated** **A picture containing text, diagram, screenshot, technical drawing

Description automatically generated**

## Hierarchical Clustering Analysis

The boxplots plotted from the hierarchical clustering using “ward” method produced an optimal 0.581 silhouette score. For Engine capacity, It shows in cluster 0 cars with higher engine capacity(about 2000) are sold, while in cluster 1 cars with lower engine capacity(about 1500) are sold. For Price, In cluster 0 cars with a lower price(about 40000) are sold, while in cluster 1 cars with a lower price(about 20000) are sold which shows that there is a demand for lower priced cars.

## Evaluation and comparing of models

The K-Means model has Sum of Squared Errors (SSE) score of 2746.77 (to 2 decimal places) and a Silhouette score of 0.542 (to 3 significant figures). Parts of the clusters overlap one another while distance between data points within thew clusters are quite close to one another with the exception of a few parts of the cluster represented by yellow data points which has some parts spread out. The cars in clusters 0 and 1 are similar in having low engine capacities with cars in cluster 1 tending to have slightly higher engine capacity. The cars in clusters 0 have lower fuel tank capacities while cars in cluster have median Fuel Tank Capacities. The selling prices of cars in clusters 0 and 1 both very low with cluster 1 having a slightly higher price. A screenshot of a computer

Description automatically generated with medium confidence  
  
 In the Hierarchical clustering model, the data points are slightly more evenly distributed throughout the clusters with a silhouette score of 0.581 (to 3 significant figures). The clusters formed using the “ward” method show that cars with higher engine capacities usually are sold at higher prices. A close-up of a computer screen

Description automatically generated with low confidence

Overall, The Hierarchical model had a higher silhouette score compared to the K-means model. I have came to the conclusion that the Hierarchical model is more appropriate model as it has a better silhouette score which further bolsters my conclusion, further more it was much easier for me to identify and come to a conclusion using the Hierarchical model compared to the K-means model. Furthermore, the clusters in the hierarchical model were less cluttered compared to the K-means model. Although the K-means model is formed using 3 columns, the ability to understand the data more easily is more important.

# Summary and Interpretation

|  |  |
| --- | --- |
| Cluster 0 | Prices range from around 30000 to 80000. Engine capacity Ranging from 2000 to 2250. |
| Cluster 1 | Prices range from around 10000 to 20000. Engine capacity Ranging from 1200 to 1500. |
| Interpretation of Cluster 0 | The cars in cluster 0 are sold at a higher price and have a higher engine capacity. |
| Interpretation of Cluster 1 | The cars in cluster 1 are sold at a lower price and have a lower engine capacity. |
| Summary of findings | To summarize my findings from the Hierarchical Clustering, I have found that the clusters are separated by Price and Engine Capacity. From the findings in the clusters formed I can come to a conclusion that the prices of the car are directly correlated to the Engine Capacity. The higher the Engine Capacity, The higher the Price. I can also come to a conclusion that there is a demand for cars with better engine capacity but lower prices. |

# Reflection

In Conclusion, I have learned a lot over the course of these few weeks on learning about python, jupyter notebook in the module Applied Analytics. I am now able to successfully do these methods of kmeans or hierarchical clustering in which I was taught. I feel like I could have done a lot better in managing my time and doing this finish faster and less last minute. However I am very thankful for the guidance Ive had over the course of this assignment by my lecturer mr joie and my classmates and seniors who have help me with guidance and making me be able to achieve and finish this assignment. For this assignment I feel like I could have done a lot better on the data cleaning part. I still feel like I have stuff in which I could have done better there which could really have helped better impact my findings in the different clustering methods. Some possible further improvements are a more condensed method for converting the objects to integers. I hope I do well for this module and this course.