Part E: Machine Learning

Trond Zachariassen

April 2024

Code: github.com/Tronden

Introduction

The objective of this assignment is to apply advanced machine learning techniques to a dataset. In this report i do this to a dataset detailing fuel consumption characteristics across a variety of vehicles from the year 2000. Provided by the dataset "FuelConsumption.csv"

Dataset

- Title: FuelConsumption.csv
- Author: Krupa Dharmshi
- License: MIT
- Found through kaggle.com
- Link to: Dataset

1 Part 1: Pre-processing/Exploring Data

1.1 Handling Missing Values and Duplicates

Initial steps includes checking for missing values and duplicates in the dataset. Missing values are handled by imputing with the median of relevant columns, and duplicates are removed to ensure the uniqueness of data entries.

1.2 Data Transformation and Encoding

Categorical variables such as ‘MAKE‘ and ‘FUEL‘ are encoded into numerical formats using one-hot encoding to facilitate their use in machine learning models.

1.3 Info

Here is the info of the values in the dataset.

```
Data columns (total 5 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   MAKE              639 non-null    object
 1   ENGINE SIZE       639 non-null    float64
 2   CYLINDERS         639 non-null    int64
 3   FUEL CONSUMPTION  639 non-null    float64
 4   COEMISSIONS       639 non-null    int64
dtypes: float64(2), int64(2), object(1)
memory usage: 25.1+ KB
None

Dataset Description:
       ENGINE SIZE   CYLINDERS  FUEL CONSUMPTION  COEMISSIONS
count   639.000000  639.000000        639.000000   639.000000
mean      3.265728    5.805947         14.713615   296.809077
std       1.231012    1.625588          3.307044    65.504178
min       1.000000    3.000000          4.900000   104.000000
25%       2.200000    4.000000         12.500000   253.000000
50%       3.000000    6.000000         14.400000   288.000000
75%       4.300000    6.000000         16.600000   343.000000
max       8.000000   12.000000         30.200000   582.000000

Missing Values in Each Column:
MAKE                0
ENGINE SIZE         0
CYLINDERS           0
FUEL CONSUMPTION    0
COEMISSIONS         0
dtype: int64

Duplicate Rows in the Dataset:
115

Correlation Matrix:
                  ENGINE SIZE  CYLINDERS  FUEL CONSUMPTION  COEMISSIONS
ENGINE SIZE          1.000000   0.895650          0.854761     0.842569
CYLINDERS            0.895650   1.000000          0.814884     0.785582
FUEL CONSUMPTION     0.854761   0.814884          1.000000     0.985804
COEMISSIONS          0.842569   0.785582          0.985804     1.000000
```

Figure 1: Dataset info

1.4 Exploratory Data Analysis (EDA)

I conducted an extensive EDA that included:

- Histograms to visualize distributions.
- Boxplots to detect outliers.
- A correlation matrix to identify relationships between variables.

1.5 Visual Representations

Scatter plot of 'ENGINE SIZE' vs 'FUEL CONSUMPTION' are used to illustrate the relationships.

1.6 Histograms

Histograms allow us to see the frequency distribution of individual variables.

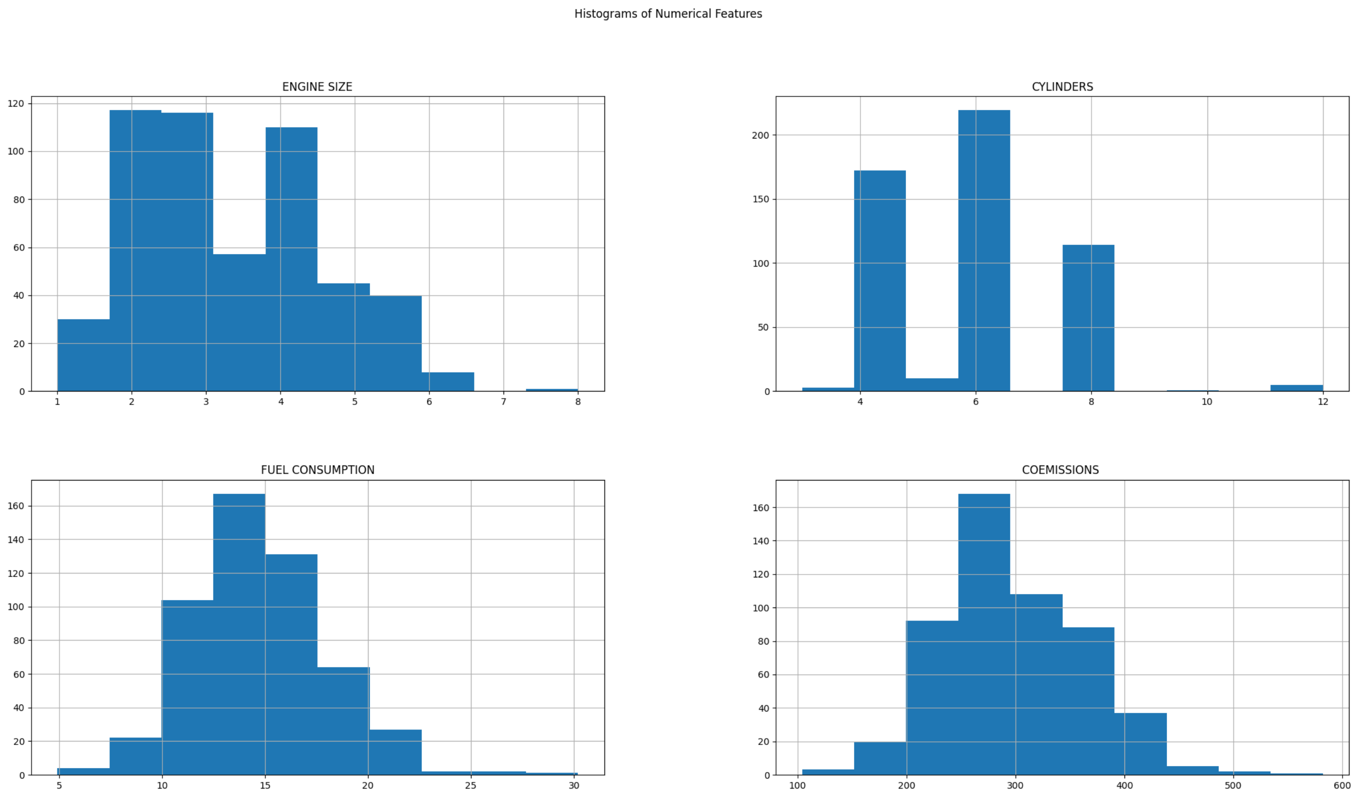


Figure 2: Histograms of numerical features for ENGINE SIZE, CYLINDERS, FUEL CONSUMPTION, and COEMISSIONS.

1.7 Boxplots

Boxplots provide a visual summary of the numerical data point distribution and help us identify outliers.

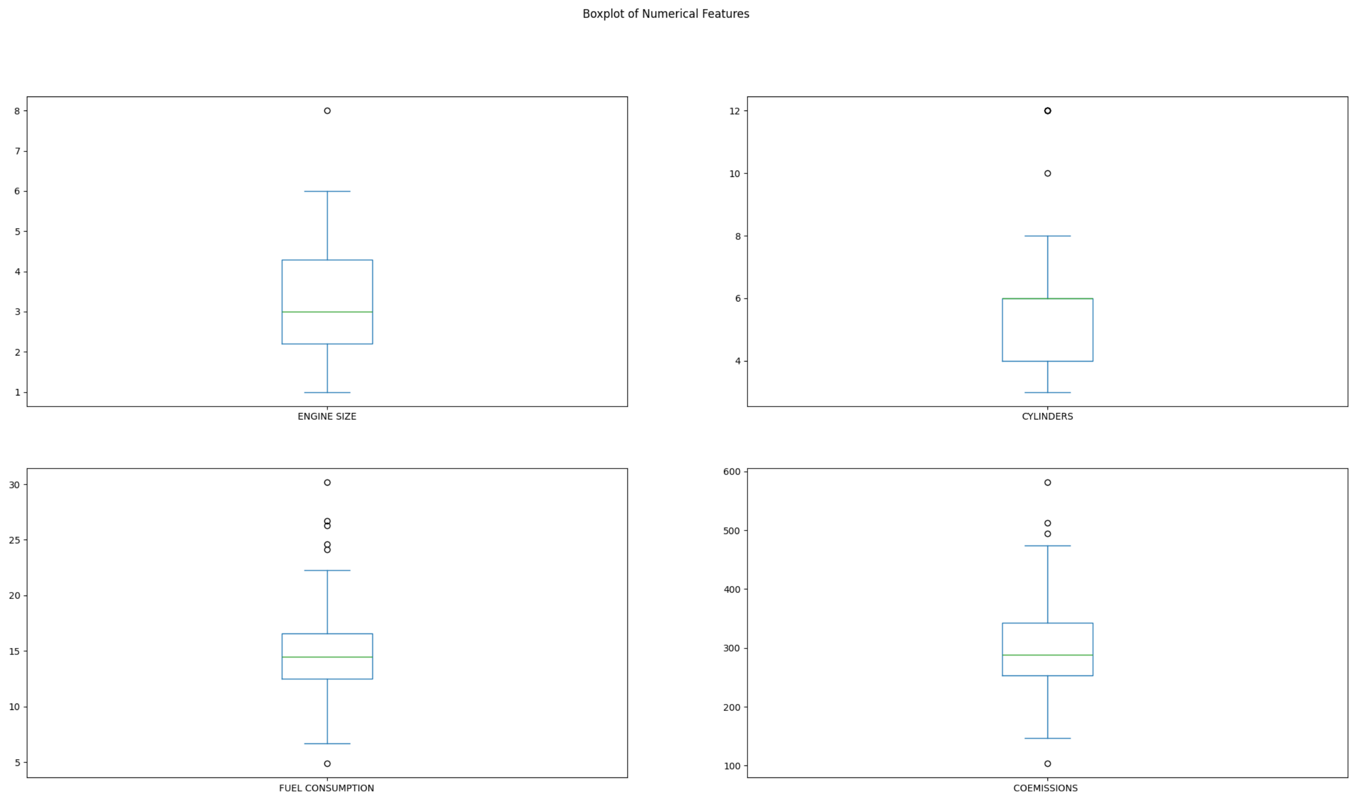


Figure 3: Boxplots of numerical features showing ENGINE SIZE, CYLINDERS, FUEL CONSUMPTION, and COEMISSIONS.

1.8 Correlation Matrix

The correlation matrix heatmap shows how features correlate with each other. High positive values indicate a strong relationship.

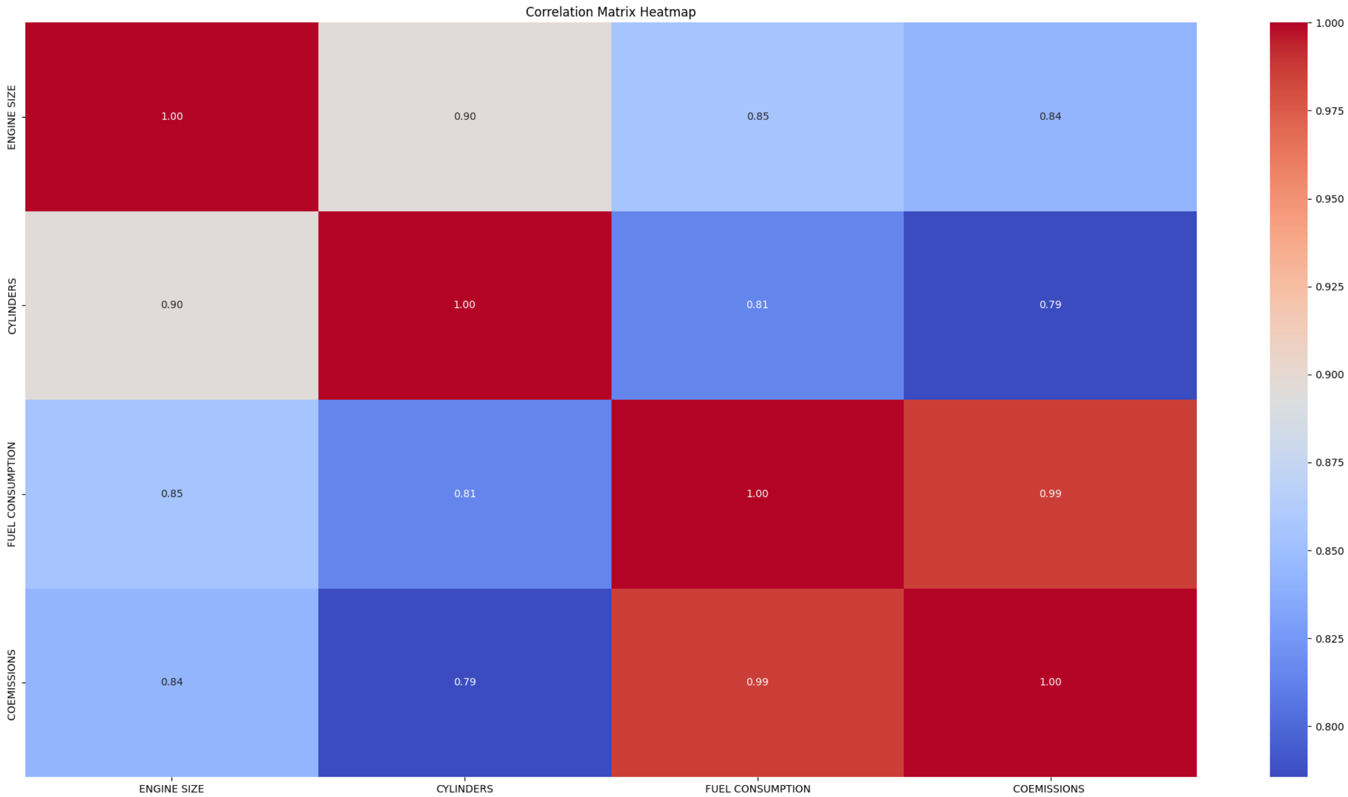


Figure 4: Correlation matrix heatmap for ENGINE SIZE, CYLINDERS, FUEL CONSUMPTION, and COEMISSIONS.

1.9 Scatter Plot

The scatter plot below illustrates the relationship between ENGINE SIZE and FUEL CONSUMPTION. This visual representation helps us to identify trends and outliers in the data. A clear trend is observable, indicating that larger engines tend to consume more fuel.

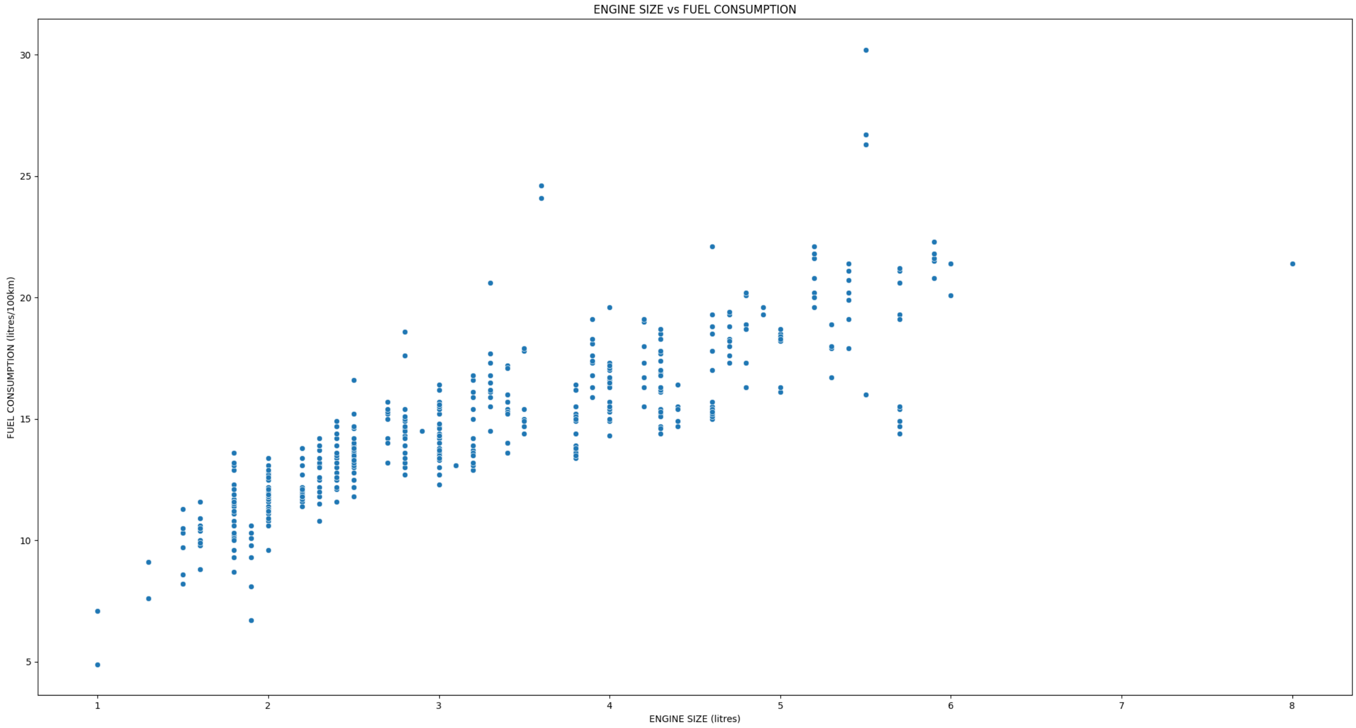


Figure 5: Scatter plot of ENGINE SIZE vs FUEL CONSUMPTION.

2 Part 2: Supervised Learning

This section discusses the implementation of supervised learning algorithms on the dataset, describing the chosen algorithms, the rationale behind their selection, the hyperparameters involved, and the composition of the training and test datasets.

2.1 Supervised Learning Algorithms

I employed Logistic Regression and Decision Tree Classifier as our supervised learning algorithms.
Logistic Regression is renowned for its simplicity and effectiveness in binary classifications. Its robustness and ease of interpretation make it an ideal choice for baseline models.
Decision Tree Classifier is favored for its ability to handle complex datasets with a mix of features. It's intuitive and the trees generated are easy to understand, which is valuable for interpretability.

2.2 Hyperparameters

For Logistic Regression, i tuned the **C** parameter, which controls the strength of regularization, and for Decision Tree Classifier, i adjusted the **max_depth**, which determines how deep the tree can grow before stopping.

2.3 Training and Test Set Distribution

The dataset was divided into an 80-20 split for the training and test sets, respectively. Here are the details of the distribution:

Training Set:

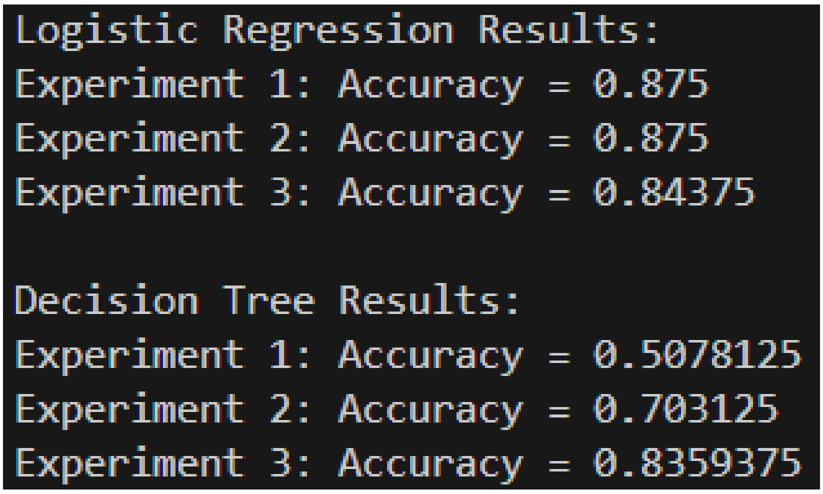
- Class 2 (Largest Engine Size): 154 instances, 30.14% of the training set.
- Class 0: 137 instances, 26.81% of the training set.
- Class 1: 133 instances, 26.03% of the training set.
- Class 3 (Smallest Engine Size): 87 instances, 17.03% of the training set.

Test Set:

- Class 2 (Largest Engine Size): 37 instances, 29.06% of the test set.
- Class 3: 32 instances, 25% of the test set.
- Class 1: 31 instances, 24.22% of the test set.
- Class 0: 28 instances, 21.88% of the test set.

2.4 Performance and Results

The models' performance was evaluated by their accuracy scores, obtained through cross-validation on the training set and applied to the test set.



```
Logistic Regression Results:
Experiment 1: Accuracy = 0.875
Experiment 2: Accuracy = 0.875
Experiment 3: Accuracy = 0.84375

Decision Tree Results:
Experiment 1: Accuracy = 0.5078125
Experiment 2: Accuracy = 0.703125
Experiment 3: Accuracy = 0.8359375
```

Figure 6: Comparison of accuracy scores for Logistic Regression and Decision Tree Classifier.

3 Part 3: Unsupervised Learning

In this part, unsupervised learning techniques were applied to explore the intrinsic structures of the fuel consumption dataset without the guidance of labeled outcomes.

3.1 Unsupervised Learning Algorithms

Two unsupervised algorithms, K-Means and Agglomerative Clustering, were chosen for this analysis.

K-Means Clustering partitions the data into k distinct clusters based on feature similarity. The primary hyperparameter, **k**, determines the number of clusters and was varied during our experiments. K-Means is particularly known for its efficiency in large datasets.

Agglomerative Clustering is a hierarchical clustering technique that builds nested clusters by merging or splitting them successively. This method uses a bottom-up approach, where each observation starts in its own cluster and pairs of clusters are merged as one moves up the hierarchy.

3.2 Hyperparameters and Experiments

The experiments with both algorithms involved varying the number of clusters. For K-Means, the values of **k** tested were 2, 5, 10, 15, and 20. For Agglomerative Clustering, in addition to testing different numbers of clusters, we explored various linkage criteria:

Ward linkage minimizes the total within-cluster variance.

Average linkage minimizes the average of the distances between all observations of pairs of clusters.

Complete linkage minimizes the maximum distance between observations of pairs of clusters.

3.3 Results and Analysis

Each clustering algorithm's performance was assessed using the silhouette score, which measures how similar an object is to its own cluster compared to other clusters.

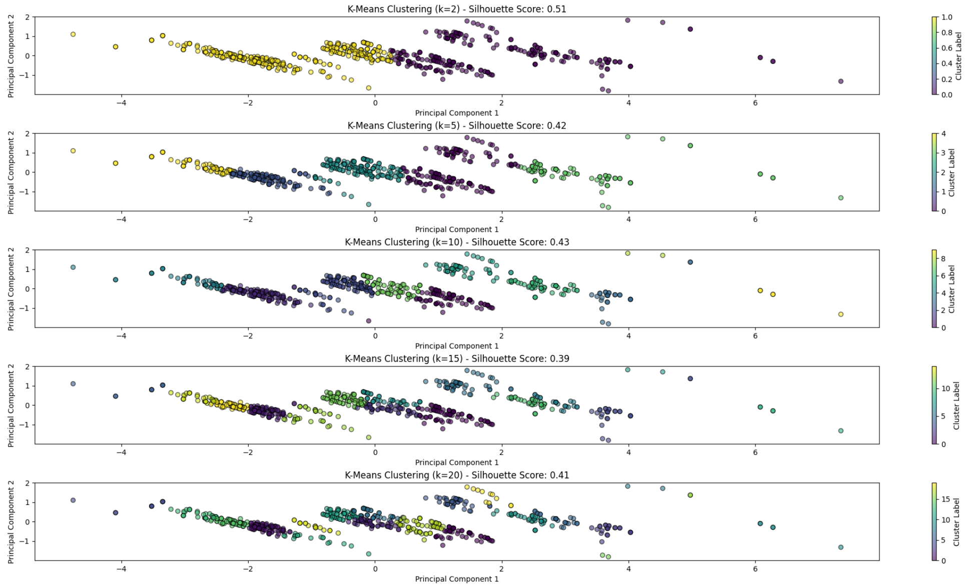


Figure 7: K-Means clustering results with different numbers of clusters, visualized using PCA-reduced features. Silhouette scores for each clustering solution are also provided.

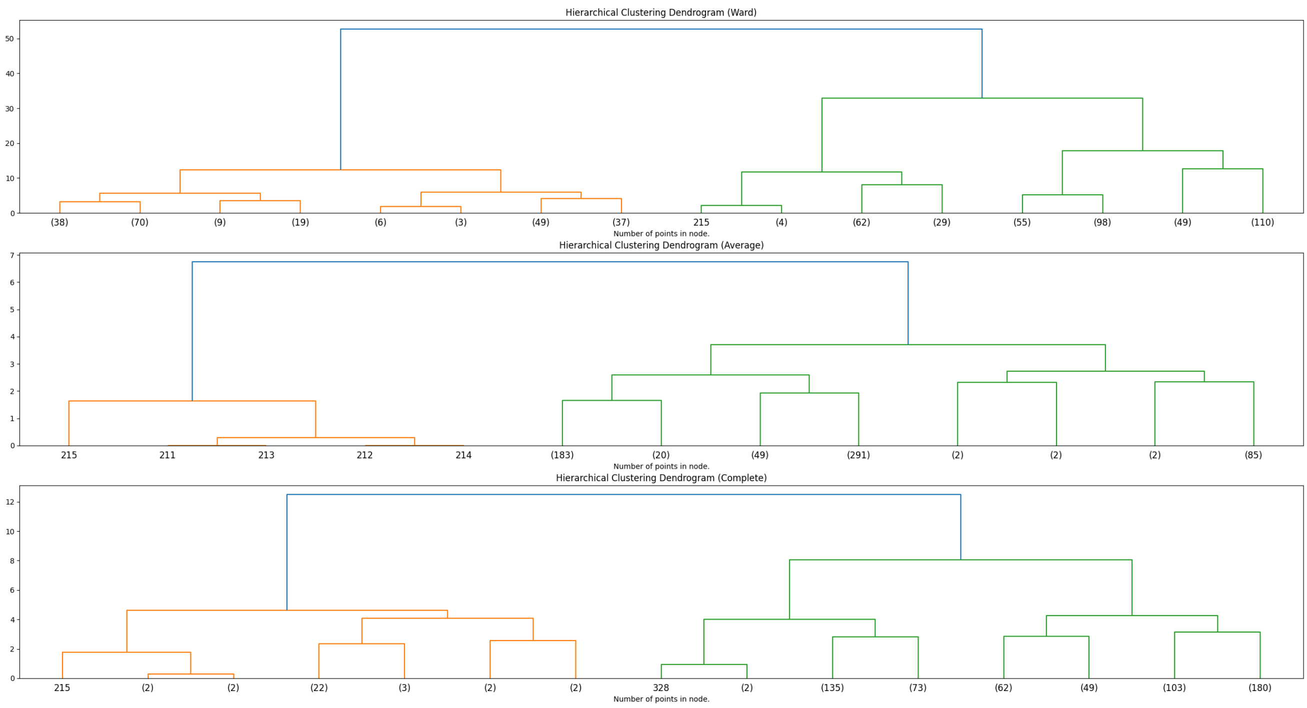


Figure 8: Agglomerative clustering dendrograms using Ward, Average, and Complete linkage methods.

```
K-Means Clustering Silhouette Scores:
Clusters: 2, Silhouette Score: 0.51
Clusters: 5, Silhouette Score: 0.42
Clusters: 10, Silhouette Score: 0.43
Clusters: 15, Silhouette Score: 0.39
Clusters: 20, Silhouette Score: 0.41
Agglomerative Clustering Silhouette Scores:
Clusters: 2, Silhouette Score: 0.47
Clusters: 5, Silhouette Score: 0.48
Clusters: 10, Silhouette Score: 0.46
Clusters: 15, Silhouette Score: 0.42
Clusters: 20, Silhouette Score: 0.41
```

Figure 9: Silhouette scores for K-Means and Agglomerative clustering methods across different numbers of clusters.

3.4 Conclusions on Data Separability

The silhouette scores and dendrogram analysis indicate the data's tendency to cluster into groups. The silhouette scores suggest that a [2] cluster solution may provide the best separation for K-Means. Meanwhile, Agglomerative clustering with [5] linkage displayed meaningful hierarchical relationships within the data.

4 References

1. Krupa Dharmshi — *FuelConsumption.csv* — Link

2. GeeksforGeeks — *Supervised and unsupervised learning* — Link