UNSUPERVISED MACHINE LEARNING - SEMESTER II

Predict city-cycle fuel consumption

BY: Atharva Kulkarni, Mihir Sawant

SVKM'S NMIMS Nilkamal School of Mathematics, Applied Statistics and Analytics,

Vile Parle West, Mumbai- 400 056

Introduction:

The objective of this study is to predict the city-cycle fuel consumption in miles per gallon (mpg) for automobiles using unsupervised learning techniques. The dataset used consists of 398 instances and 9 attributes, including 3 multivalued discrete attributes (cylinders, model year, and origin) and 5 continuous attributes (displacement, horsepower, weight, acceleration, and mpg).

The dataset is first pre-processed by handling missing values and outliers. Missing values in the horsepower attribute are replaced with the median value, and outliers in the horsepower and acceleration attributes are imputed using the interquartile range (IQR) method.

Exploratory data analysis is performed to understand the distribution of the variables, identify correlations, and visualize the data using various plots and statistical techniques. The dataset is then clustered using hierarchical clustering and K-means clustering algorithms to identify patterns and group similar instances together.

The hierarchical clustering approach uses the average linkage method and creates a dendrogram to visualize the clustering process. The dataset is divided into two clusters based on the dendrogram analysis. Similarly, the K-means clustering algorithm is applied to the dataset, and the optimal number of clusters is determined using the elbow method and silhouette coefficient analysis.

Linear regression models are then trained on the original dataset and the clustered datasets (from both hierarchical and K-means clustering) to predict the 'mpg' attribute. The performance of the regression models is evaluated using appropriate metrics, and the results are compared to determine the effectiveness of the clustering techniques in improving the prediction accuracy.

The study provides insights into the application of unsupervised learning techniques, specifically clustering, in preprocessing and feature engineering for regression analysis tasks related to fuel consumption prediction.

Problem Statement:

To develop an effective machine learning model for predicting the city-cycle fuel consumption in miles per gallon (mpg) for automobiles, by leveraging unsupervised learning techniques such as hierarchical clustering and K-means clustering. The goal is to identify patterns and group similar instances together, thereby improving the accuracy of regression models trained on the clustered datasets. The study aims to explore the impact of clustering on feature engineering and preprocessing for regression analysis tasks related to fuel consumption prediction.

- ❖ **The project aims to address the following key aspects:**

1. Cluster the dataset using hierarchical clustering to create distinct datasets for regression model training.

2. Implement regression models to predict the 'mpg' attribute accurately for each cluster.

3. Handle the mix of discrete and continuous attributes through appropriate preprocessing techniques.

4. Evaluate the accuracy of the regression models and assess the impact of clustering on prediction performance.

5. Explore the correlation between variables and identify any significant relationships that can enhance predictive modelling.

Based on the key aspects, here's an elaboration on each point:

1. **Cluster the dataset using hierarchical clustering to create distinct datasets for regression model training**:
   - Hierarchical clustering is an unsupervised machine learning technique that groups similar data points together based on their proximity or dissimilarity.
   - In this project, hierarchical clustering is applied to the dataset to identify distinct clusters or groups of vehicles based on their characteristics (e.g., cylinders, displacement, horsepower, weight, acceleration).

- The rationale behind clustering is to separate the dataset into more homogeneous subgroups, which can potentially improve the performance of regression models trained on each cluster individually.
- By creating distinct datasets through clustering, the project aims to capture the inherent patterns and variations within the data, ultimately leading to more accurate predictions of fuel consumption (mpg) for each cluster.

2. **Implement regression models to predict the 'mpg' attribute accurately for each cluster**:
    - Regression analysis is a statistical technique used to model the relationship between a dependent variable (in this case, 'mpg') and one or more independent variables (e.g., cylinders, displacement, horsepower, weight, acceleration).
    - After clustering the dataset, separate regression models will be trained and evaluated for each identified cluster.
    - The goal is to build accurate predictive models that can estimate the fuel consumption (mpg) based on the vehicle's characteristics within each cluster.
    - By training cluster-specific models, the project aims to capture the unique patterns and relationships between variables that may exist within each cluster, potentially improving the overall prediction accuracy.
3. **Handle the mix of discrete and continuous attributes through appropriate preprocessing techniques**:
    - The dataset contains a mix of discrete (e.g., cylinders, model year, origin) and continuous (e.g., displacement, horsepower, weight, acceleration) attributes.
    - Appropriate preprocessing techniques are required to handle these different types of variables effectively.
    - For discrete variables, techniques like one-hot encoding or label encoding may be employed to convert them into a numerical format suitable for machine learning algorithms.

- Continuous variables may require scaling or normalization to ensure that all features are on a similar scale, preventing any one feature from dominating the others.

4. **Evaluate the accuracy of the regression models and assess the impact of clustering on prediction performance**:

    - After training the regression models on the clustered datasets, their prediction accuracy will be evaluated using appropriate metrics (e.g., mean squared error, R-squared).
    - The project will compare the performance of the cluster-specific regression models with a baseline model trained on the entire dataset without clustering.
    - This comparison will help assess the impact of clustering on prediction performance and determine whether separating the data into distinct clusters improves the overall accuracy of fuel consumption predictions.
    - The evaluation will provide insights into the effectiveness of the clustering approach and guide future decisions on whether to incorporate clustering as a preprocessing step for predictive modeling tasks.

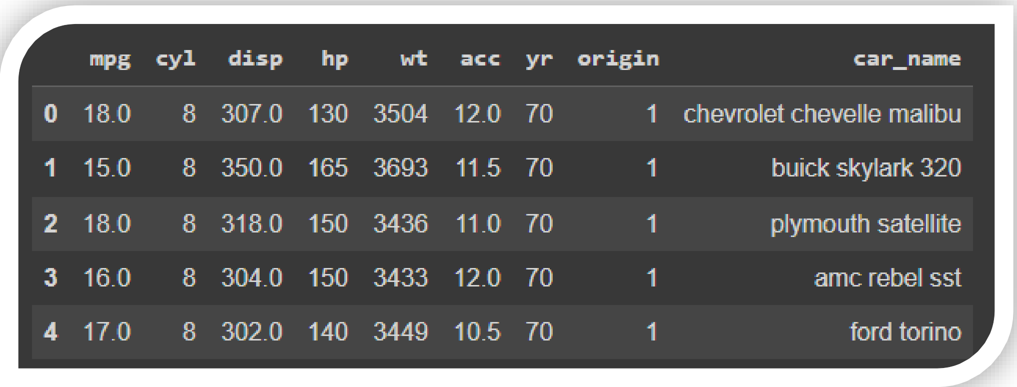
5. **Explore the correlation between variables and identify any significant relationships that can enhance predictive modeling**:

    - Correlation analysis is a statistical technique used to measure the strength and direction of the relationship between variables.
    - In this project, correlation analysis will be performed to identify any significant relationships between the independent variables (e.g., cylinders, displacement, horsepower, weight, acceleration) and the dependent variable (mpg).
    - Strong correlations between certain variables may provide insights into which features are more influential in predicting fuel consumption and could potentially be leveraged to improve the predictive models.
    - Additionally, understanding the correlations between independent variables can help detect and address multicollinearity issues, which can negatively impact the performance of regression models.

METHODOLOGY / INSIGHTS -

1. **Dataset Description**:

- The dataset used in this study comprises 398 instances of automobiles, each with 9 attributes, including 3 multivalued discrete attributes (cylinders, model year, and origin) and 5 continuous attributes (displacement, horsepower, weight, acceleration, and miles per gallon).

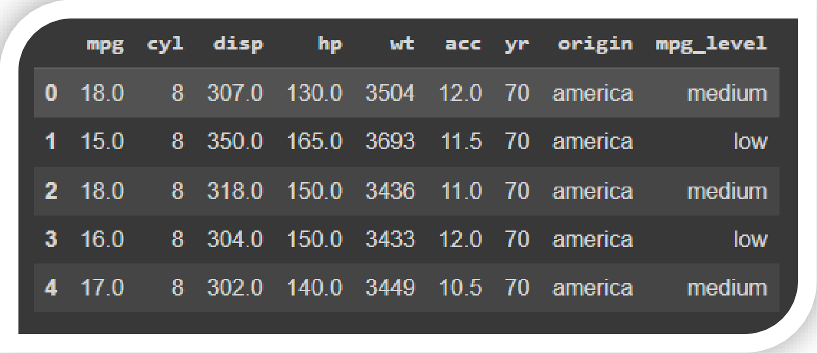


	mpg	cyl	disp	hp	wt	acc	yr	origin	car_name
0	18.0	8	307.0	130	3504	12.0	70	1	chevrolet chevelle malibu
1	15.0	8	350.0	165	3693	11.5	70	1	buick skylark 320
2	18.0	8	318.0	150	3436	11.0	70	1	plymouth satellite
3	16.0	8	304.0	150	3433	12.0	70	1	amc rebel sst
4	17.0	8	302.0	140	3449	10.5	70	1	ford torino

2. **Preprocessing:**

- Handling missing values by imputation with median values.
- Detecting and addressing outliers using the interquartile range (IQR) method.
- Encoding categorical variables for compatibility with clustering algorithms.

<u>Data after pre-processing</u>

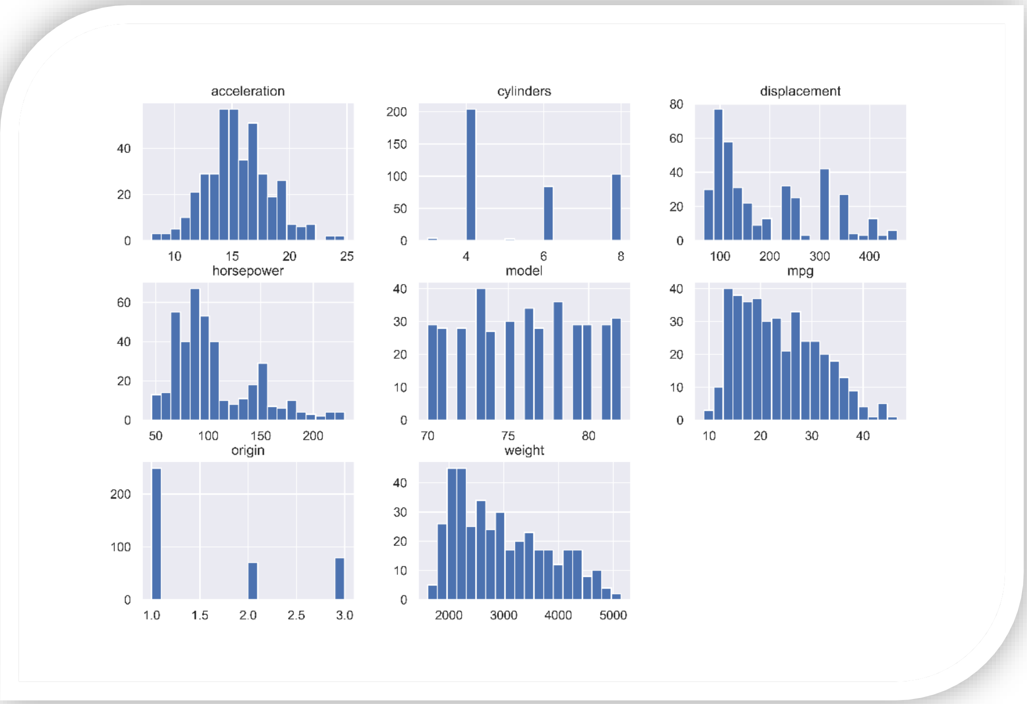


	mpg	cyl	disp	hp	wt	acc	yr	origin	mpg_level
0	18.0	8	307.0	130.0	3504	12.0	70	america	medium
1	15.0	8	350.0	165.0	3693	11.5	70	america	low
2	18.0	8	318.0	150.0	3436	11.0	70	america	medium
3	16.0	8	304.0	150.0	3433	12.0	70	america	low
4	17.0	8	302.0	140.0	3449	10.5	70	america	medium

3. Exploratory Data Analysis:

- Visualizing the distribution of variables using histograms

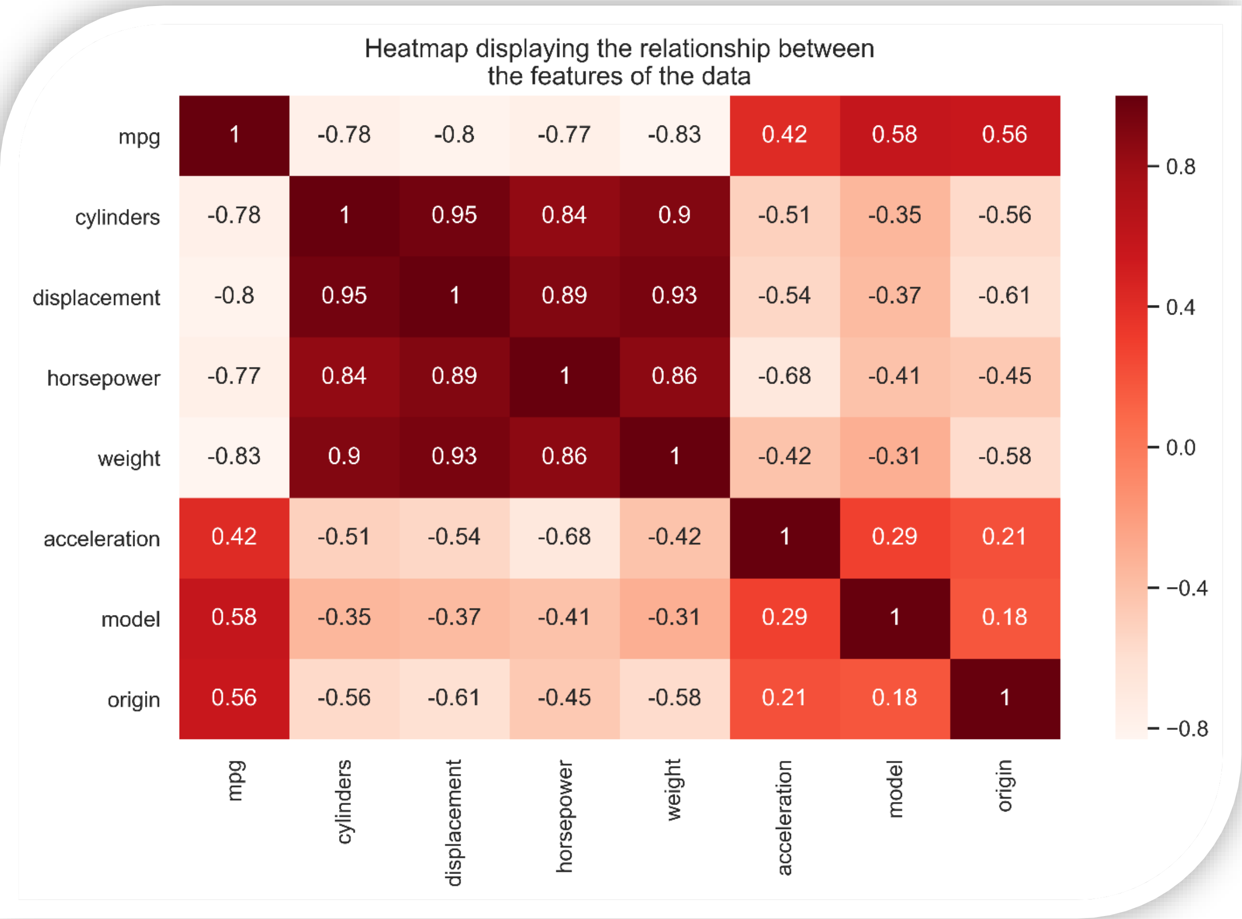
  - ➢ The histogram for acceleration shows a normal distribution, with most cars having an acceleration around 15 meters per second squared
  - ➢ The histogram for cylinders reveals that approximately 51.3% of the cars in the dataset have 4 cylinders
  - ➢ The histogram for displacement indicates a slight right skew, suggesting the presence of larger displacement values in the dataset
  - ➢ The histogram for the dependent variable mpg (miles per gallon) also exhibits a slight right skew



- Examining correlations between variables using a heatmap

  - ➢ The heatmap shows a strong negative correlation between mpg and variables like displacement (-0.8), horsepower (-0.77), weight (-0.83), and cylinders (-0.78). This implies that as these variables increase, the fuel efficiency (mpg) decreases.

- There is a strong positive correlation among displacement, horsepower, weight, and cylinders, indicating potential multicollinearity issues. Multicollinearity can hinder the performance and accuracy of linear regression models, suggesting the need for feature selection.
- The acceleration, model, and origin variables do not exhibit high correlations with each other or with the other variables.

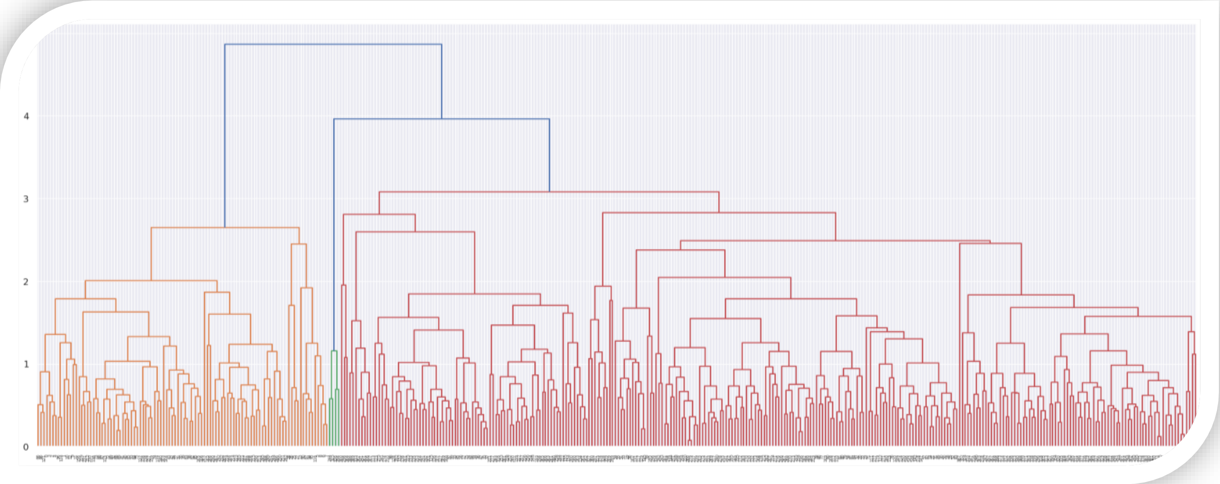


- <u>Additional insights</u>

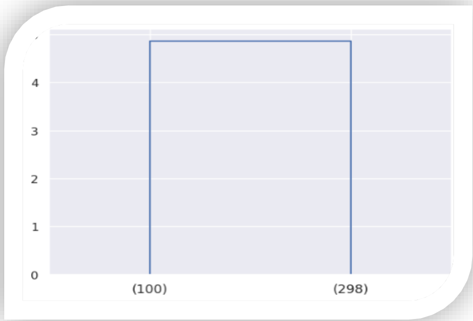
  - The relationship between mpg and other variables satisfies the linear regression assumption of a linear relationship between the dependent and independent variables.
  - The strong positive correlations among displacement, horsepower, weight, and cylinders violate the non-multicollinearity assumption of linear regression, which can negatively impact model performance and accuracy. Feature selection techniques may be necessary to address this issue.

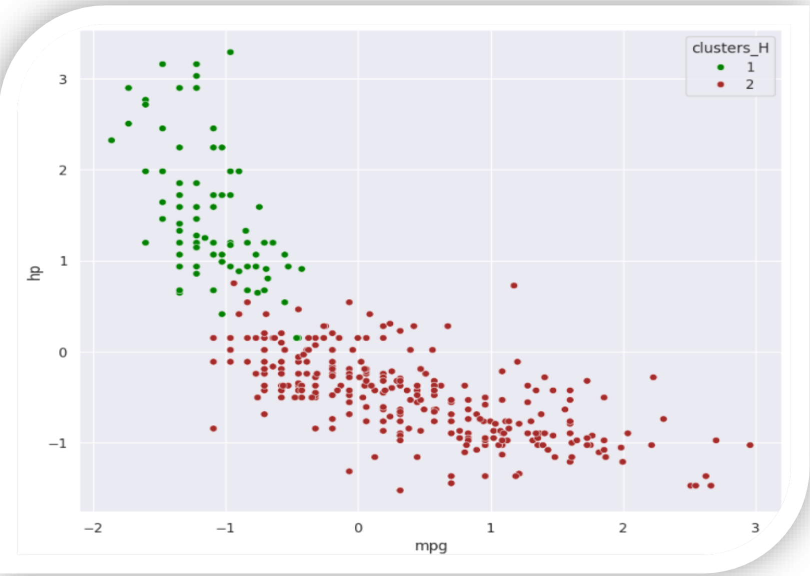
4. Hierarchical Clustering:

- Applying the average linkage method for hierarchical clustering.
- Visualizing the clustering process using dendrograms.



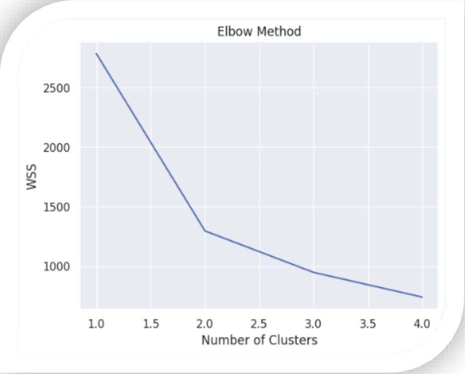
- Determining the optimal number of clusters based on domain knowledge and visual inspection.

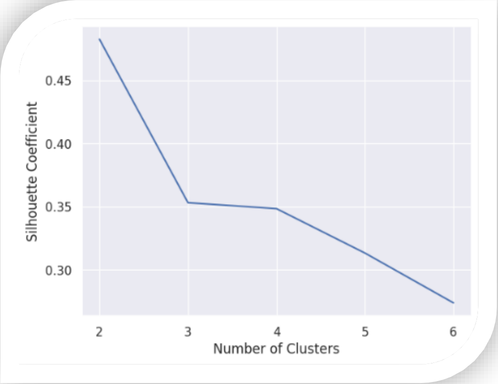




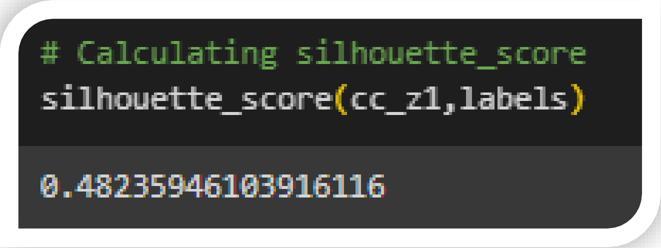
5. K-means Clustering:

- Employing the K-means algorithm for clustering.
- Determining the optimal number of clusters using the elbow method and silhouette analysis.



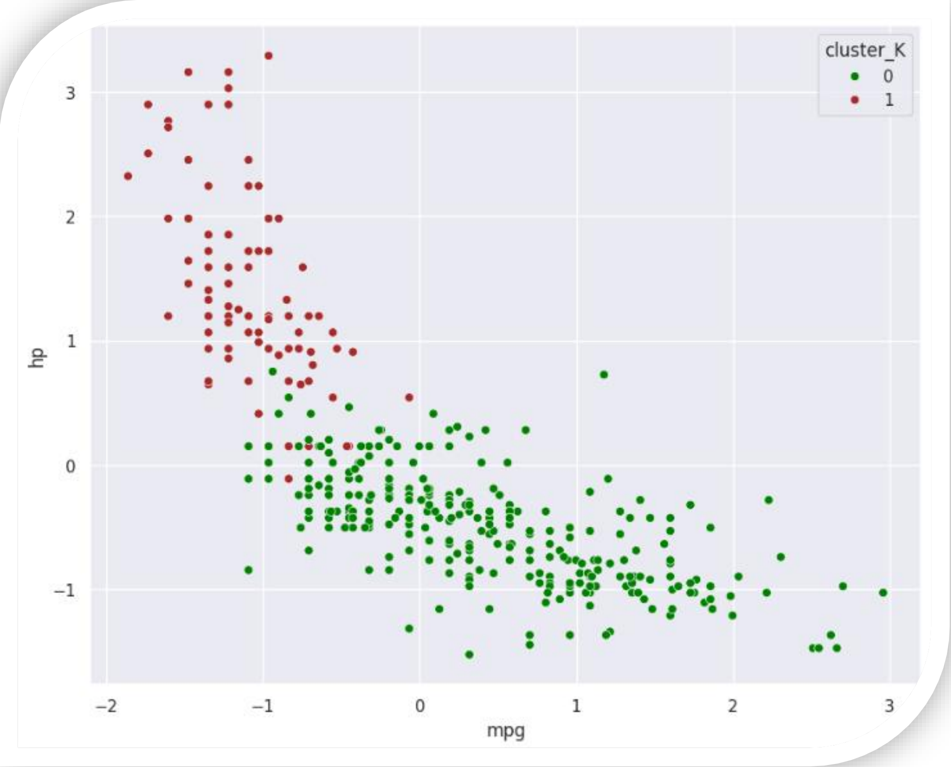


- Evaluating clustering performance using silhouette scores.



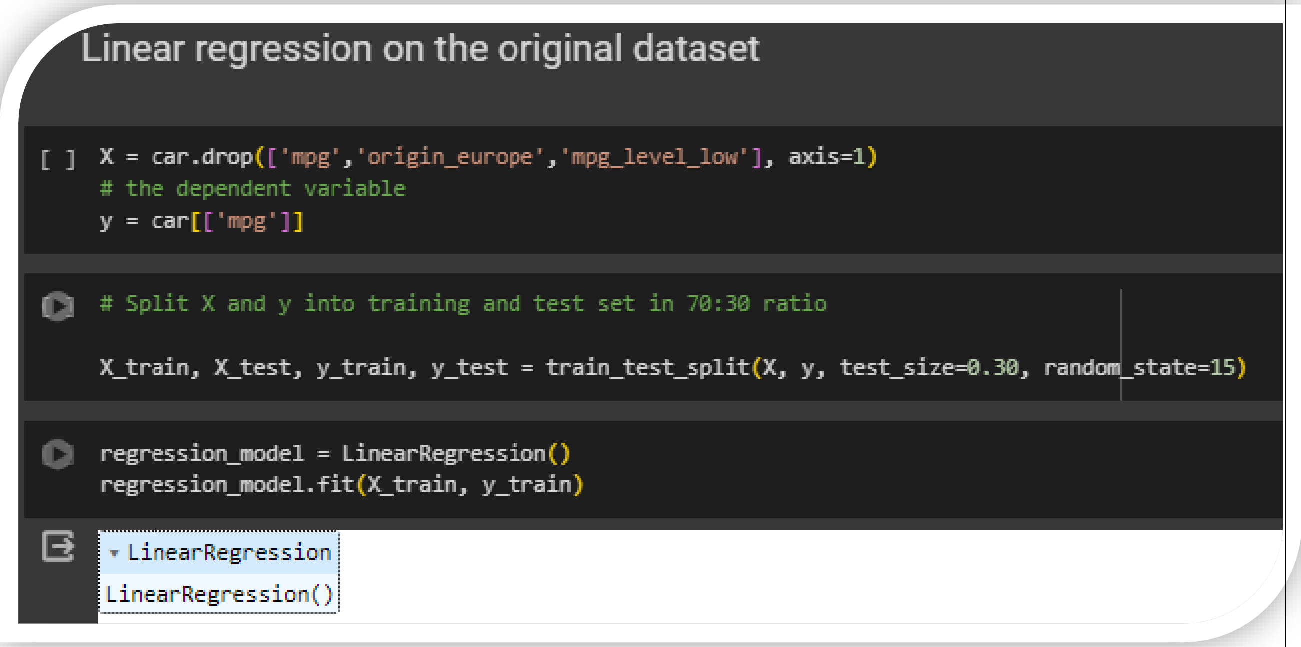
```
# Calculating silhouette_score
silhouette_score(cc_z1,labels)
```

```
0.48235946103916116
```



6. Regression Modeling:

- Training linear regression models on the original dataset.



Linear regression on the original dataset

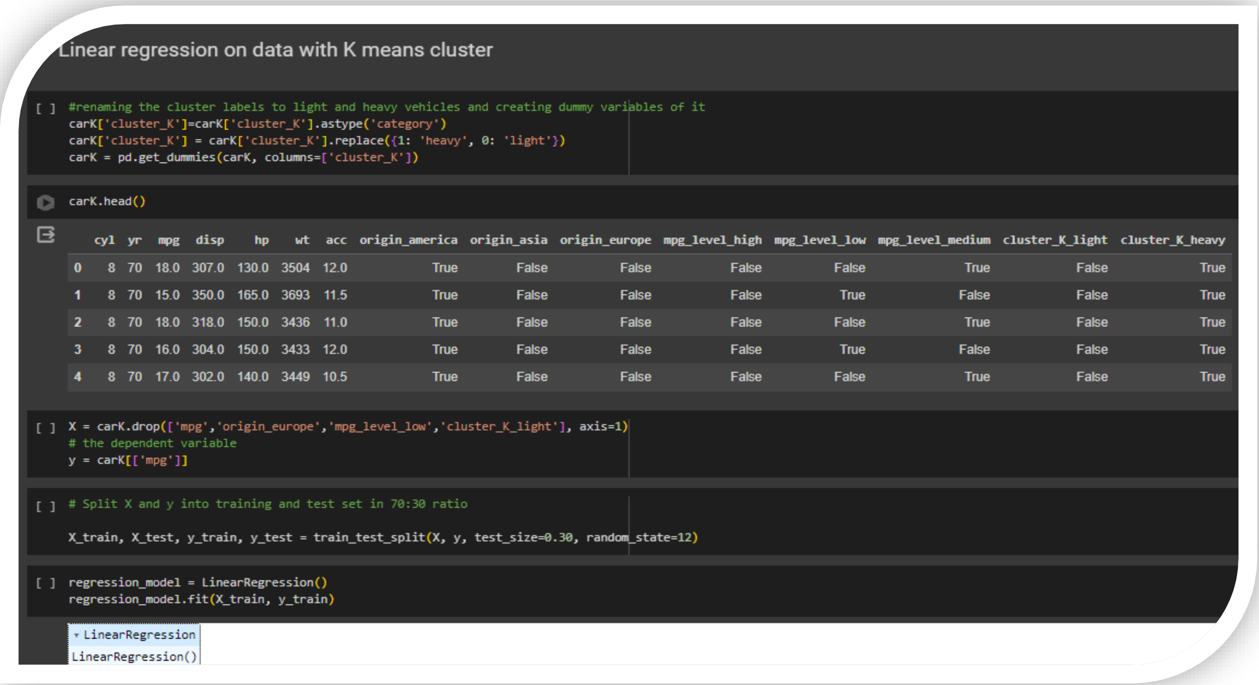
```
[ ] X = car.drop(['mpg','origin_europe','mpg_level_low'], axis=1)
    # the dependent variable
    y = car[['mpg']]
```

```
# Split X and y into training and test set in 70:30 ratio

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.30, random_state=15)
```

```
regression_model = LinearRegression()
regression_model.fit(X_train, y_train)
```

```
▾ LinearRegression
LinearRegression()
```

- Training separate regression models on the clusters obtained from hierarchical and K-means clustering.



Linear regression on data with K means cluster

```
#renaming the cluster labels to light and heavy vehicles and creating dummy variables of it
carK['cluster_K']=carK['cluster_K'].astype('category')
carK['cluster_K'] = carK['cluster_K'].replace({1: 'heavy', 0: 'light'})
carK = pd.get_dummies(carK, columns=['cluster_K'])
```

```
carK.head()
```

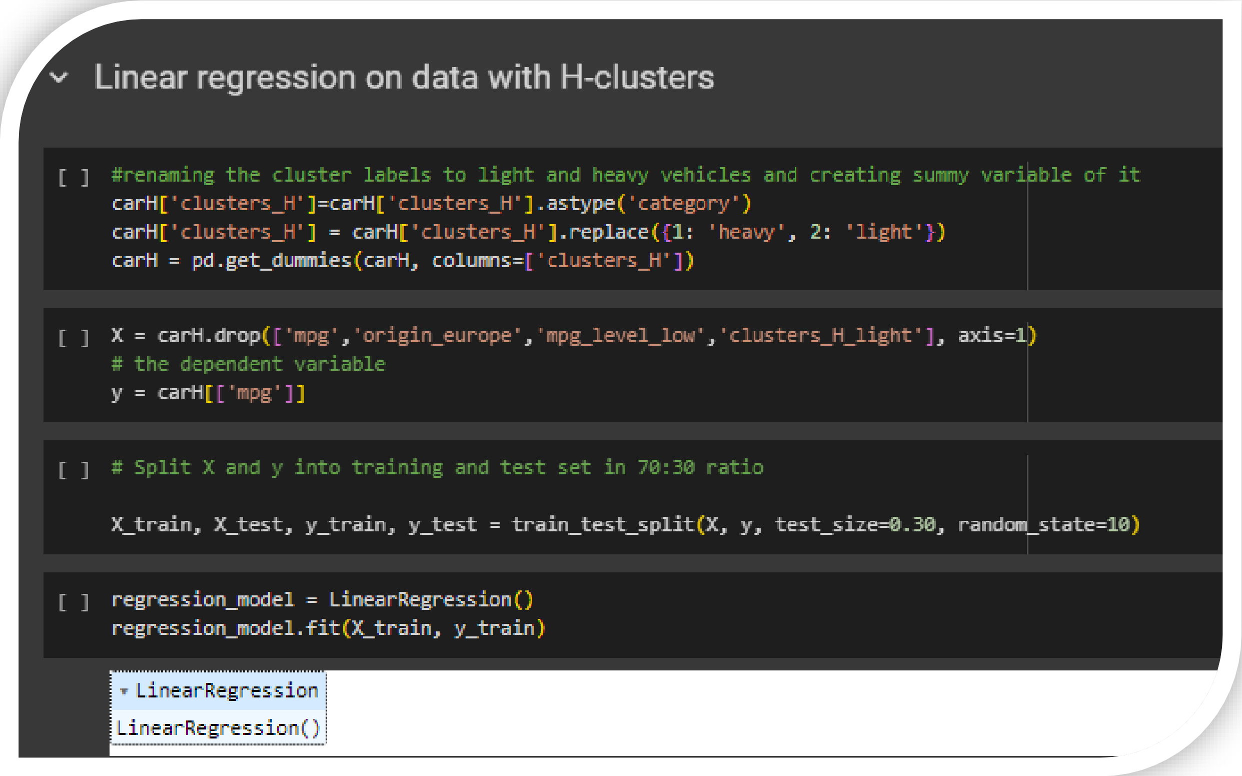
	cyl	yr	mpg	disp	hp	wt	acc	origin_america	origin_asia	origin_europe	mpg_level_high	mpg_level_low	mpg_level_medium	cluster_K_light	cluster_K_heavy
0	8	70	18.0	307.0	130.0	3504	12.0	True	False	False	False	False	True	False	True
1	8	70	15.0	350.0	165.0	3693	11.5	True	False	False	False	True	False	False	True
2	8	70	18.0	318.0	150.0	3436	11.0	True	False	False	False	False	True	False	True
3	8	70	16.0	304.0	150.0	3433	12.0	True	False	False	False	True	False	False	True
4	8	70	17.0	302.0	140.0	3449	10.5	True	False	False	False	False	True	False	True

```
X = carK.drop(['mpg','origin_europe','mpg_level_low','cluster_K_light'], axis=1)
# the dependent variable
y = carK[['mpg']]
```

```
# Split X and y into training and test set in 70:30 ratio

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.30, random_state=12)
```

```
regression_model = LinearRegression()
regression_model.fit(X_train, y_train)
```

```
▾ LinearRegression
LinearRegression()
```



Linear regression on data with H-clusters

```
#renaming the cluster labels to light and heavy vehicles and creating summy variable of it
carH['clusters_H']=carH['clusters_H'].astype('category')
carH['clusters_H'] = carH['clusters_H'].replace({1: 'heavy', 2: 'light'})
carH = pd.get_dummies(carH, columns=['clusters_H'])
```

```
X = carH.drop(['mpg','origin_europe','mpg_level_low','clusters_H_light'], axis=1)
# the dependent variable
y = carH[['mpg']]
```

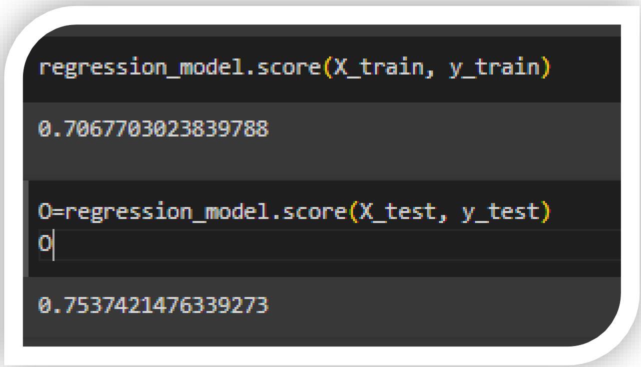
```
# Split X and y into training and test set in 70:30 ratio

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.30, random_state=10)
```

```
regression_model = LinearRegression()
regression_model.fit(X_train, y_train)
```

```
▾ LinearRegression
LinearRegression()
```

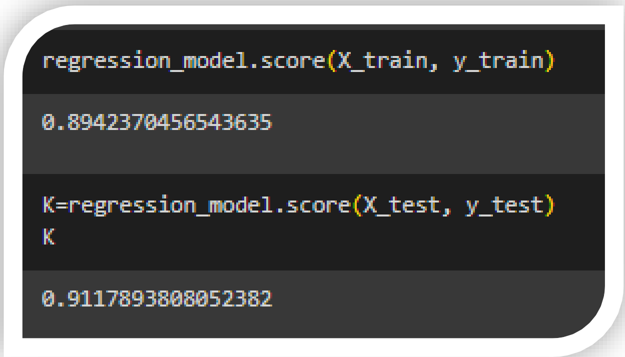
7. Comparative Analysis:

- Comparing the performance of regression models trained on the original dataset and clustered datasets.
    - ➢ Regression on original data



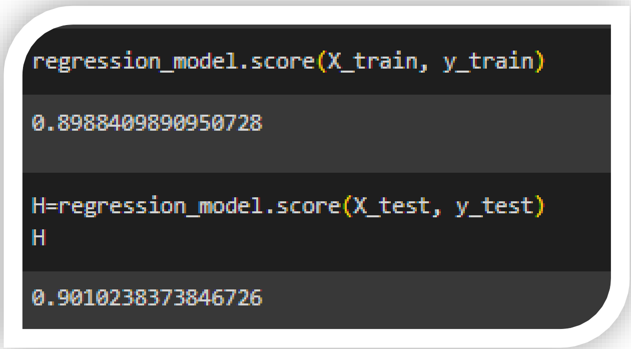
```
regression_model.score(X_train, y_train)

0.7067703023839788

O=regression_model.score(X_test, y_test)
O

0.7537421476339273
```

- ➢ Regression after k-means clustering



```
regression_model.score(X_train, y_train)

0.8942370456543635

K=regression_model.score(X_test, y_test)
K

0.9117893808052382
```

- ➢ Regression after hierarchical clustering



```
regression_model.score(X_train, y_train)

0.8988409890950728

H=regression_model.score(X_test, y_test)
H

0.9010238373846726
```

Conclusion:

The study aimed to leverage unsupervised learning techniques, specifically hierarchical clustering and K-means clustering, to enhance the accuracy of regression models for predicting city-cycle fuel consumption (mpg) in automobiles. The analysis was conducted on a dataset comprising 398 instances and 9 attributes, including multivalued discrete variables like cylinders, model year, and origin, as well as continuous variables such as displacement, horsepower, weight, and acceleration.

Through exploratory data analysis, valuable insights were gained regarding the distribution of variables, relationships between attributes, and potential issues like multicollinearity. The strong positive correlations among features like displacement, horsepower, weight, and cylinders highlighted the need for feature selection to avoid violating the non-multicollinearity assumption of linear regression.

Hierarchical clustering was performed using the average linkage method, and the dataset was divided into two clusters based on the dendrogram analysis. K-means clustering was also applied, with the optimal number of clusters determined using the elbow method and silhouette coefficient analysis. These clustering techniques allowed for the identification of patterns and groupings within the data, enabling the training of separate regression models on the clustered datasets.

Linear regression models were trained on the original dataset as well as the clustered datasets obtained from hierarchical and K-means clustering. By leveraging the clustered datasets, the study aimed to improve the accuracy of fuel consumption prediction by accounting for the inherent patterns and relationships between variables within each cluster.

Summary:

K-means appears to explain the highest variation in the datset, but with a difference of only 1% when compared with Herarchical clustering, to get more clarity a larger dataset may be used, since this is a dataset of used cars it doesn't give us how many previous owners has the cars seen which might be helful variable,the gender of the previous owners, the reason/purpose that the cars were being used is also an important factor which the dataset doen't capture. With the above mentioned features it may be possible to get a higher accuracy or explainability of the models and its variables.