**Introduction:**

The objective of this lab was to perform a detailed analysis and clustering of a dataset containing car specifications. The process involved data cleaning, preprocessing, visualization, outlier detection, and clustering.

**Data Loading and Initial Exploration:**

A close-up of a logo

Description automatically generated

A screenshot of a graph

Description automatically generated

This code loads the car dataset from a CSV file into a Pandas Data Frame and displays the first few rows to give an initial glimpse of the data structure.

**Data Dimensions and Missing Values:**

A screenshot of a computer code

Description automatically generated

A screenshot of a computer code

Description automatically generated

Here, we obtain the dimensions of the dataset and check for any missing values in each column to understand the completeness of the data.

**Data Cleaning:**

A screenshot of a computer code

Description automatically generated

A computer screen shot of a code

Description automatically generated

A screenshot of a computer

Description automatically generated

In this step, we remove rows with any missing values and replace specific placeholder strings ("$null$") with NaN to standardize missing data representation. We then recheck for any remaining missing values.

**Data Type Conversion:**

A screenshot of a computer

Description automatically generated

A computer code with text

Description automatically generated with medium confidence



A screenshot of a computer

Description automatically generated

We convert several columns to numeric data types, coercing errors to NaN where necessary. This ensures that all numerical operations can be performed correctly on these columns.

**Handling Missing Values:**

A screenshot of a computer

Description automatically generated

Missing values in numerical columns are filled with the mean of their respective columns to maintain data integrity while not losing any data points.

**One-Hot Encoding:**

A screenshot of a computer code

Description automatically generated

We perform one-hot encoding on categorical columns (‘manufact’ and ‘model’) to convert them into a format suitable for machine learning algorithms.

**Data Normalization:**

A screenshot of a computer

Description automatically generated

Numerical features are normalized using StandardScaler to ensure they have a mean of 0 and a standard deviation of 1, which is crucial for many machine learning algorithms.

**Outlier Detection and Removal:**

A screenshot of a computer program

Description automatically generated

An Isolation Forest model is used to detect and remove outliers, ensuring the data used for clustering is of high quality.

**Data Visualization:**

A screenshot of a computer

Description automatically generated

A group of blue and white graphs

Description automatically generated

A chart with different colored squares

Description automatically generated

Histograms and box plots are used to visualize the distribution and identify any potential outliers in the numerical features of the dataset.

**Hierarchical Clustering:**

A computer code with text

Description automatically generated with medium confidence

A diagram of a clustering diagram

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer program

Description automatically generated

We perform hierarchical clustering and determine the number of clusters from the dendrogram. The resulting cluster labels are added to the dataset.

**PCA and Cluster Visualization:**

A computer code with text

Description automatically generated

A diagram of a graph

Description automatically generated with medium confidence

A computer screen shot of a program code

Description automatically generated

A screen shot of a graph

Description automatically generated

Principal Component Analysis (PCA) is used to reduce the dimensionality of the data, making it easier to visualize the clusters. The clusters are then plotted in a 2D space.

Finally, a modified input based on an existing data point is visualized within the PCA plot to see how it relates to the identified clusters.

**Streamlit App:**

A screenshot of a graph

Description automatically generated

A screenshot of a graph

Description automatically generated

Deployed on huggingface spaces: [Click](https://huggingface.co/spaces/rayyanphysicist/Vehicle_Clustering_Analysis)

--------------------------------------------------THE END-------------------------------------------------