VEHICLE DATA ANALYSIS

Libraries used:

NumPy, Pandas, Scikit learn, Seaborn, Matplotlib

Data Introduction:

Car Data with 13 features and 8128 rows. 9 categorical and 4 numerical features.

Data Cleaning:

Missing Values

mileage, engine, max_power, torque, and seats had presence of missing values. mileage, engine and seats were filled for missing values using mean, mean and median respectively as they were numerical features. After removing strings from features mileage, engine, it was converted to float type and feature seat were converted to integer type.

Duplicated Rows

Presence of 1202 duplicated rows were found. Hence it was removed from the dataset.

Data visualisations:

From the visualisations, the following insights were obtained:

1. Maruti, Hyundai and Mahindra are the three most popular brands.
2. Diesel is the fuel type which is used by 54% of cars.
3. Individual sales are the highest.
4. Most of the cars are manual.
5. Sellers are selling more first hand.
6. The cars are mostly 5 seated followed by 7 seated.
7. Most of the cars were sold highest in the year 2017.
8. The feature brand name was compared with other numerical features and results were as:
    - The brand name Ashok followed by Mitsubishi and Force were the highest when compared with km_driven.
    - Volvo followed by Jaguar had the highest selling price.
    - Renault followed by Maruti and Datsun had the topmost mileage.
    - Jaguar followed by Isuzu and Force had the topmost engine speed.

Data Pre-processing:

Label Encoding

The features max_power and torque were dropped due to irrelevance and inadequate information. The categorical features were label encoded along with numerical features for consistency.

Scaling Features

All the features were scaled using standard scaler.

Principal Component Analysis

11 components were considered and transformed & fitted for pca. From the correlation matrix, the first component was in a high correlation with fifth component and seventh with eighth component.

Segmentation:

K-means Clustering

Since labels are unknown and in such a situation unsupervised algorithm such as clustering can be used. The cluster number were found and taken to be 6.

Visualisation

From the visualisation, the first and eleventh components were compared using a scatter plot by clustering the features into 6 clusters each with different value counts and centroids.