Milestone 1 Report: Car Recommendation Engine

Rohith Kumar Ballem

UFID: 30969136

1. Objective of the Project :

The goal of this project is to develop a car recommendation engine that assists users in selecting a vehicle based on the customer's preferences and needs. The system will analyze various car features and attributes, the sales and purchase behaviors of customers, and their ratings to provide recommendations that are personalized to people who are planning to purchase a new car.

This recommendation engine will be beneficial for:

- **Car buyers**: people who want to get a recommendation for a car with specific requirements, budget and on the basis of review and ratings.
- **Online marketplaces:** aiming to enhance user experience by suggesting the best vehicle options to their customers.

2. Tools used:

- Language: **Python**
- Notebook: **Kaggle**
- Libraries used:

  numpy – Numerical computations

  pandas – Data manipulation and preprocessing

  matplotlib – Data visualization

  matplotlib.pyplot – Plotting functions

  seaborn – Statistical data visualization

  sklearn.preprocessing – Data normalization using MinMaxScalar

3. Data Collection

To build the car recommendation engine, three datasets have been taken from **Kaggle**

Three datasets are used to build the car recommendation system. They are:

3.1 Car Specifications:

Source: Kaggle - Car Specification Dataset 1945-2020

- **Description:** Contains the technical specifications of cars from the year 1945 to 2020.
- **Dimensions:** Multiple attributes describing car performance, dimensions, and features.

Columns of the dataset:

- **id_trim** – Unique identifier for the trim level of the car.
- **Make** – The manufacturer or brand of the car.
- **Model** – The specific model of the car.
- **Generation** – The generation of the model, representing major design changes.
- **Year_from / Year_to** – The production years for this specific trim/model.
- **Series** – The sub-line or variant of the model series.
- **Trim** – The specific variant/configuration of the car with distinct features.
- **Body_type** – The shape and style of the vehicle.
- **load_height_mm** – The height at which cargo can be loaded.
- **number_of_seats** – The seating capacity of the vehicle.
- **length_mm / width_mm / height_mm** – Dimensions of the car in millimeters.
- **wheelbase_mm** – Distance between the front and rear axles.
- **front_track_mm / rear_track_mm** – The width between the front/rear wheels.
- **curb_weight_kg** – The total weight of the car without passengers or cargo.
- **wheel_size_r14** – The wheel size, often in inches.
- **ground_clearance_mm** – The height between the car's underside and the ground.
- **trailer_load_with_brakes_kg** – Maximum trailer weight the car can tow with brakes.
- **payload_kg** – Maximum weight the vehicle can carry, including passengers and cargo.
- **back_track_width_mm / front_track_width_mm** – The distance between the wheels at the rear/front.
- **clearance_mm** – Another term for ground clearance.
- **full_weight_kg** – The total permissible weight of the vehicle.
- **front_rear_axle_load_kg** – Maximum load distribution on the front and rear axles.
- **max_trunk_capacity_l / minimum_trunk_capacity_l** – The trunk storage capacity in liters.
- **cargo_compartment_length_width_height_mm** – Cargo space dimensions.
- **cargo_volume_m3** – Total cargo capacity in cubic meters.
- **maximum_torque_n_m** – The maximum torque output of the engine.
- **injection_type** – The type of fuel injection system.

- **overhead_camshaft** – Engine design feature affecting valve timing.
- **cylinder_layout** – Arrangement of engine cylinders.
- **number_of_cylinders** – The number of cylinders in the engine.
- **compression_ratio** – Ratio of cylinder volume before and after compression.
- **engine_type** – The type of engine used.
- **valves_per_cylinder** – Number of valves per engine cylinder.
- **boost_type** – Type of forced induction, such as turbocharging or supercharging.
- **cylinder_bore_mm / stroke_cycle_mm** – Cylinder bore diameter and piston stroke length.
- **engine_placement** – Where the engine is located.
- **cylinder_bore_and_stroke_cycle_mm** – Combined metric for bore and stroke.
- **turnover_of_maximum_torque_rpm** – The RPM at which maximum torque is achieved.
- **max_power_kw** – The highest power output in kilowatts.
- **presence_of_intercooler** – Indicates if the engine has an intercooler.
- **capacity_cm3** – Engine displacement in cubic centimeters.
- **engine_hp / engine_hp_rpm** – Engine power in horsepower and the RPM at which it peaks.
- **drive_wheels** – The drivetrain type.
- **bore_stroke_ratio** – Ratio of bore diameter to stroke length.
- **number_of_gears** – Number of gears in the transmission.
- **turning_circle_m** – Minimum turning radius of the vehicle.
- **transmission** – Type of gearbox.
- **mixed_fuel_consumption_per_100_km_l** – Combined fuel consumption in liters per 100 km.
- **range_km** – The vehicle's estimated range per fuel/electric charge.
- **emission_standards** – Compliance with emission regulations.
- **fuel_tank_capacity_l** – The total fuel tank capacity in liters.
- **acceleration_0_100_km/h_s** – Time taken to accelerate from 0 to 100 km/h.
- **max_speed_km_per_h** – The top speed of the vehicle.
- **city_fuel_per_100km_l / highway_fuel_per_100km_l** – Fuel consumption in city/highway conditions.
- **CO2_emissions_g/km** – Carbon dioxide emissions per kilometer.
- **fuel_grade** – Required fuel type.
- **back_suspension / front_suspension** – Type of suspension system used.
- **rear_brakes / front_brakes** – The braking system type used on rear and front wheels.
- **steering_type** – Steering system.
- **car_class** – Classification of the car.
- **country_of_origin** – The country where the car is manufactured.
- **number_of_doors** – The number of doors in the vehicle.
- **safety_assessment / rating_name** – Safety rating and assessment details.
- **battery_capacity_KW_per_h** – Battery capacity for electric vehicles.
- **electric_range_km** – Estimated range on a full electric charge.
- **charging_time_h** – Estimated charging time for an electric vehicle

3.2 Car Sales:

Source: Kaggle - Car Sales Report

- **Description:** Records of car basic details, car sales, dealership locations and Annual income of the buyer.
- **Dimensions:** Transactional sales data from different dealers.

Columns of the dataset:

- **Car_id** – Unique identifier for each car sale.
- **Date** – The date when the car was sold.
- **Customer Name** – Name of the customer who purchased the car.
- **Gender** – Gender of the customer.
- **Annual Income** – The yearly income of the customer.
- **Dealer_Name** – Name of the car dealership.
- **Company** – The manufacturer or brand of the car.
- **Model** – The specific model of the car.
- **Engine** – Engine specifications of the car.
- **Transmission** – Type of transmission.
- **Color** – Exterior color of the car.
- **Price ($)** – The sale price of the car in dollars.
- **Dealer_No** – Unique identifier for the dealer.
- **Body Style** – Type of car body.
- **Phone** – Contact number of the dealer or customer.
- **Dealer_Region** – Geographical location of the dealership.

3.3 Car Ratings:

Source: Kaggle - Edmunds Car Review

- **Description:** User-generated car reviews and ratings from Edmunds.
- **Dimensions:** Customer review and feedback on various car models.

Columns of the dataset:

- **Company** – The car manufacturer or brand.
- **Model** – The specific model of the car.
- **Year** – The manufacturing or release year of the car.
- **Reviewer_Name** – The name of the person who reviewed the car.
- **Date** – The date when the review was written.
- **Title** – The title or headline of the review.
- **Rating** – The score or rating given to the car.
- **Review** – The detailed review or feedback given by the reviewer.

Each dataset is verified for **accessibility and licensing compliance** to ensure proper usage.

4. Data Preprocessing:

In data preprocessing stage the steps performed on the dataset are: handling missing data, detecting and handling outliers and normalizing the numerical data for future analysis and model training.

4.1 Handling Data Types and Missing Values:

- The dataset is initially examined to identify the data types and detect any missing values.
- Object-type columns were converted into appropriate numerical formats to facilitate processing and any junk is handled.
- The missing values in numerical columns were filled using **median imputation** which is resistant to any extreme values.

  (number_of_seats length_mm width_mm height_mm wheelbase_mm front_track_mm rear_track_mm curb_weight_kg ground_clearance_mm full_weight_kg max_trunk_capacity_l maximum_torque_n_m turnover_of_maximum_torque_rpm capacity_cm3 engine_hp_rpm fuel_tank_capacity_l max_speed_km_per_h fuel_grade minimum_trunk_capacity_l number_of_cylinders valves_per_cylinder cylinder_bore_mm stroke_cycle_mm number_of_gears turning_circle_m mixed_fuel_consumption_per_100_km_l acceleration_0_100_km/h_s city_fuel_per_100km_l highway_fuel_per_100km_l Year_from Year_to)
  - car_specs dataframe these columns are imputed by "**median**"

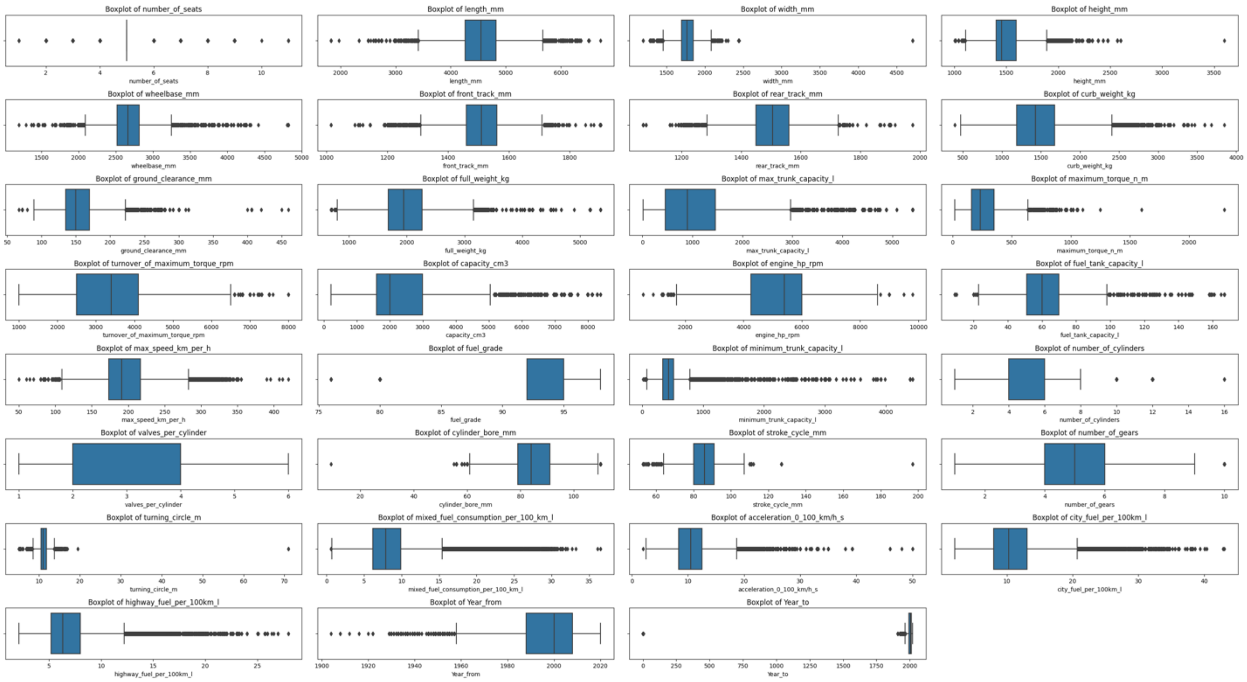
- Other types of data, missing values were imputed using **mode imputation**, preserving the most common for each column.

  (Generation Body_type injection_type cylinder_layout engine_type engine_hp drive_wheels transmission back_suspension rear_brakes front_brakes front_suspension) car_specs dataframe these columns are imputed by "**mode**"

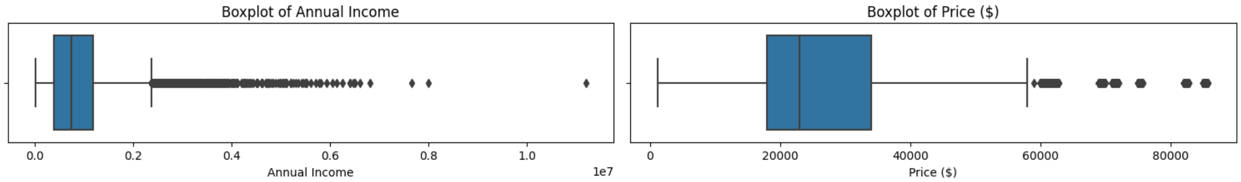
4.2 Outlier Detection and Treatment :

- **Box plots** were generated to visually analyze the presence of outliers and assess skewness in numerical variables for car specifications, car sales and car ratings datasets.

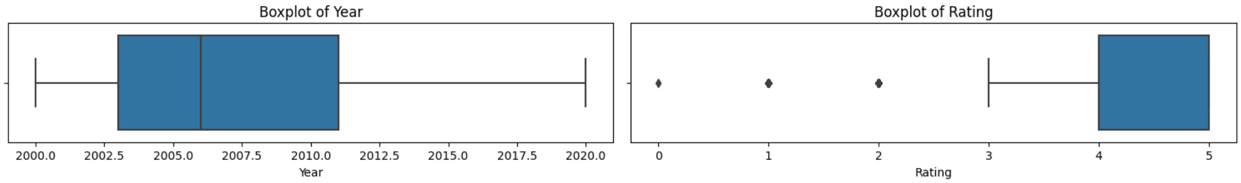
Box plots for Car Specifications data frame numerical columns:



Box plots for Car Sales data frame numerical columns:



Box plots for Car Ratings data frame numerical columns:



4.3 Handle Outliers:

Interquartile Range (IQR) method was used to identify outliers, ensuring consistency in outlier detection.

- **Log transformation** was applied to suppress extreme outliers and reduce skewness in the data as data which is segregated as outlier isn't an outlier (It's a **valid** data).Hence, chose log transformation method.

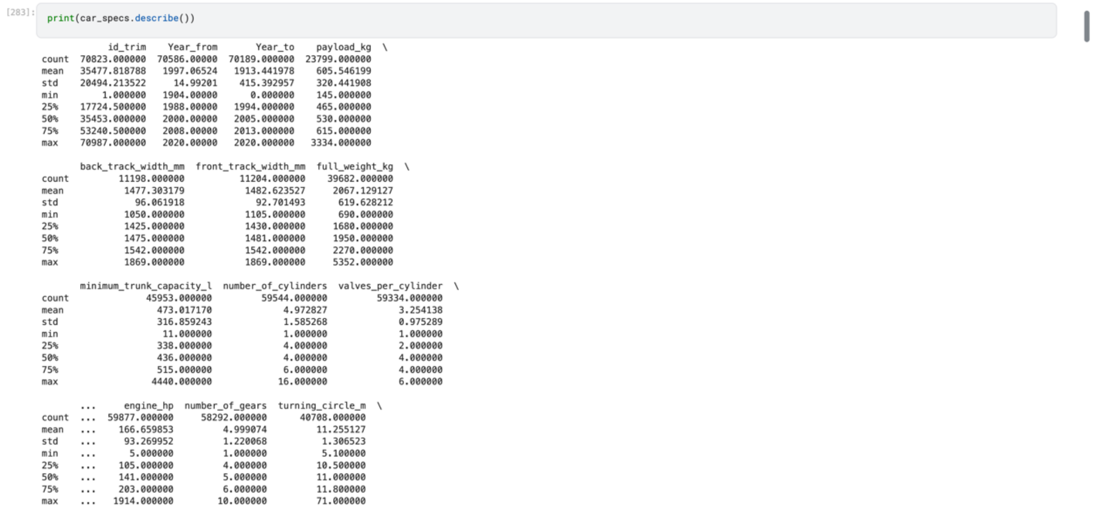
4.4 Feature Normalization

- **Min-Max Scaling** from **scikit-learn** under the **preprocessing** module was applied to normalize numerical features, ensuring values range between 0 and 1.

In conclusion, the preprocessing steps effectively cleaned and transformed the dataset for further analysis.

5. Exploratory Data Analysis (EDA):

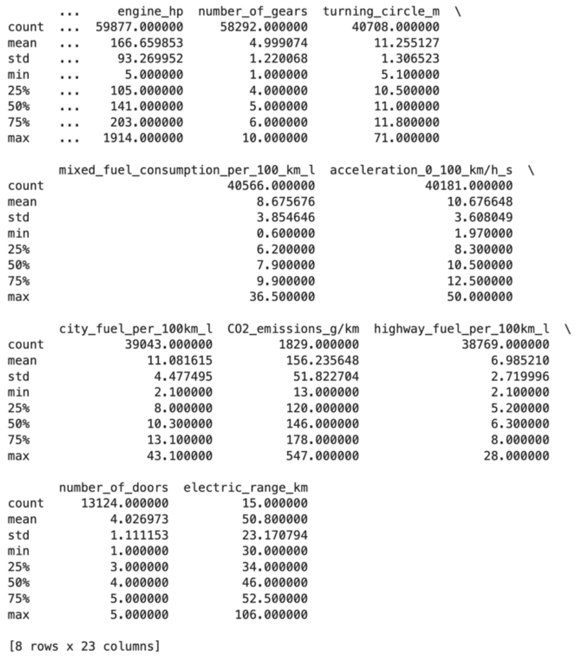
5.1 Car Specifications Dataset (car_specs dataframe)



[283]:

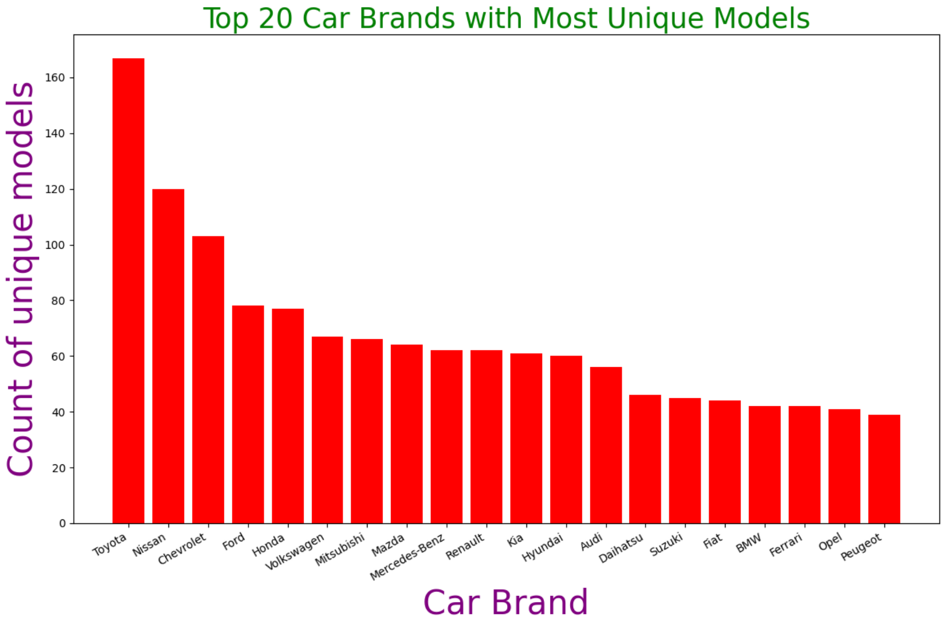
```
print(car_specs.describe())
```

```
            id_trim    Year_from       Year_to    payload_kg  \
count  70823.000000  70586.00000  70189.000000  23799.000000
mean   35477.818788   1997.06524   1913.441978    605.546199
std    20494.213522     14.99201    415.392957    320.441908
min        1.000000   1904.00000      0.000000    145.000000
25%    17724.500000   1988.00000   1994.000000    465.000000
50%    35453.000000   2000.00000   2005.000000    530.000000
75%    53240.500000   2008.00000   2013.000000    615.000000
max    70987.000000   2020.00000   2020.000000   3334.000000

       back_track_width_mm  front_track_width_mm  full_weight_kg  \
count         11198.000000          11204.000000    39682.000000
mean           1477.303179           1482.623527     2067.129127
std              96.061918             92.701493      619.628212
min            1050.000000           1105.000000      690.000000
25%            1425.000000           1430.000000     1680.000000
50%            1475.000000           1481.000000     1950.000000
75%            1542.000000           1542.000000     2270.000000
max            1869.000000           1869.000000     5352.000000

       minimum_trunk_capacity_l  number_of_cylinders  valves_per_cylinder  \
count              45953.000000         59544.000000         59334.000000
mean                 473.017170             4.972827             3.254138
std                  316.859243             1.585268             0.975289
min                   11.000000             1.000000             1.000000
25%                  338.000000             4.000000             2.000000
50%                  436.000000             4.000000             4.000000
75%                  515.000000             6.000000             4.000000
max                 4440.000000            16.000000             6.000000

       ...     engine_hp  number_of_gears  turning_circle_m  \
count  ...  59877.000000     58292.000000      40708.000000
mean   ...    166.659853         4.999074         11.255127
std    ...     93.269952         1.220068          1.306523
min    ...      5.000000         1.000000          5.100000
25%    ...    105.000000         4.000000         10.500000
50%    ...    141.000000         5.000000         11.000000
75%    ...    203.000000         6.000000         11.800000
max    ...   1914.000000        10.000000         71.000000
```

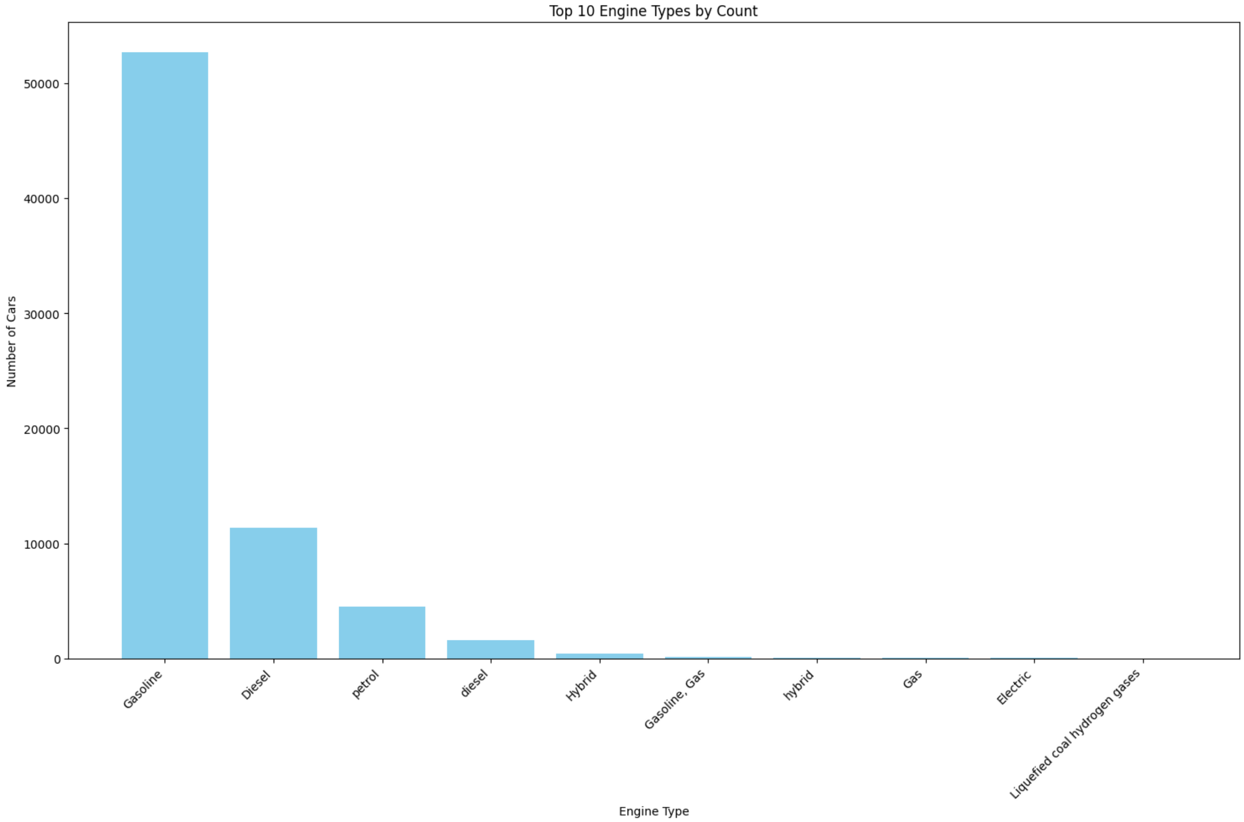


```
      ...     engine_hp  number_of_gears  turning_circle_m  \
count ...  59877.000000     58292.000000      40708.000000
mean  ...    166.659853         4.999074         11.255127
std   ...     93.269952         1.220068          1.306523
min   ...      5.000000         1.000000          5.100000
25%   ...    105.000000         4.000000         10.500000
50%   ...    141.000000         5.000000         11.000000
75%   ...    203.000000         6.000000         11.800000
max   ...   1914.000000        10.000000         71.000000

      mixed_fuel_consumption_per_100_km_l  acceleration_0_100_km/h_s  \
count                        40566.000000               40181.000000
mean                             8.675676                  10.676648
std                              3.854646                   3.608049
min                              0.600000                   1.970000
25%                              6.200000                   8.300000
50%                              7.900000                  10.500000
75%                              9.900000                  12.500000
max                             36.500000                  50.000000

      city_fuel_per_100km_l  CO2_emissions_g/km  highway_fuel_per_100km_l  \
count          39043.000000         1829.000000              38769.000000
mean              11.081615          156.235648                  6.985210
std                4.477495           51.822704                  2.719996
min                2.100000           13.000000                  2.100000
25%                8.000000          120.000000                  5.200000
50%               10.300000          146.000000                  6.300000
75%               13.100000          178.000000                  8.000000
max               43.100000          547.000000                 28.000000

      number_of_doors  electric_range_km
count    13124.000000          15.000000
mean         4.026973          50.800000
std          1.111153          23.170794
min          1.000000          30.000000
25%          3.000000          34.000000
50%          4.000000          46.000000
75%          5.000000          52.500000
max          5.000000         106.000000

[8 rows x 23 columns]
```

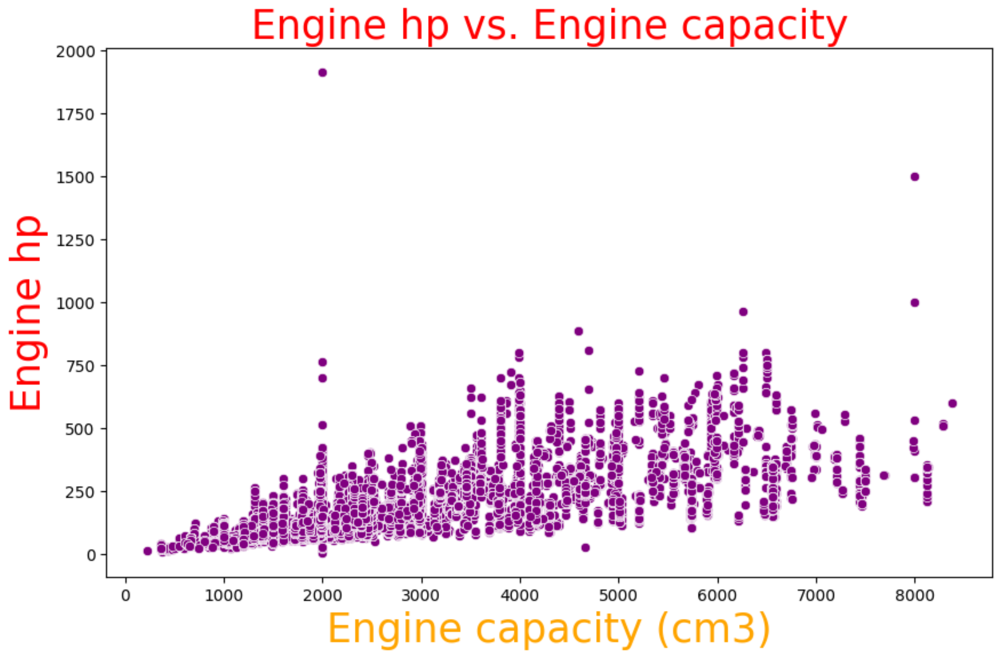
The above screenshots represents the mean, median and standard deviation for the data in the car_specs data frame.



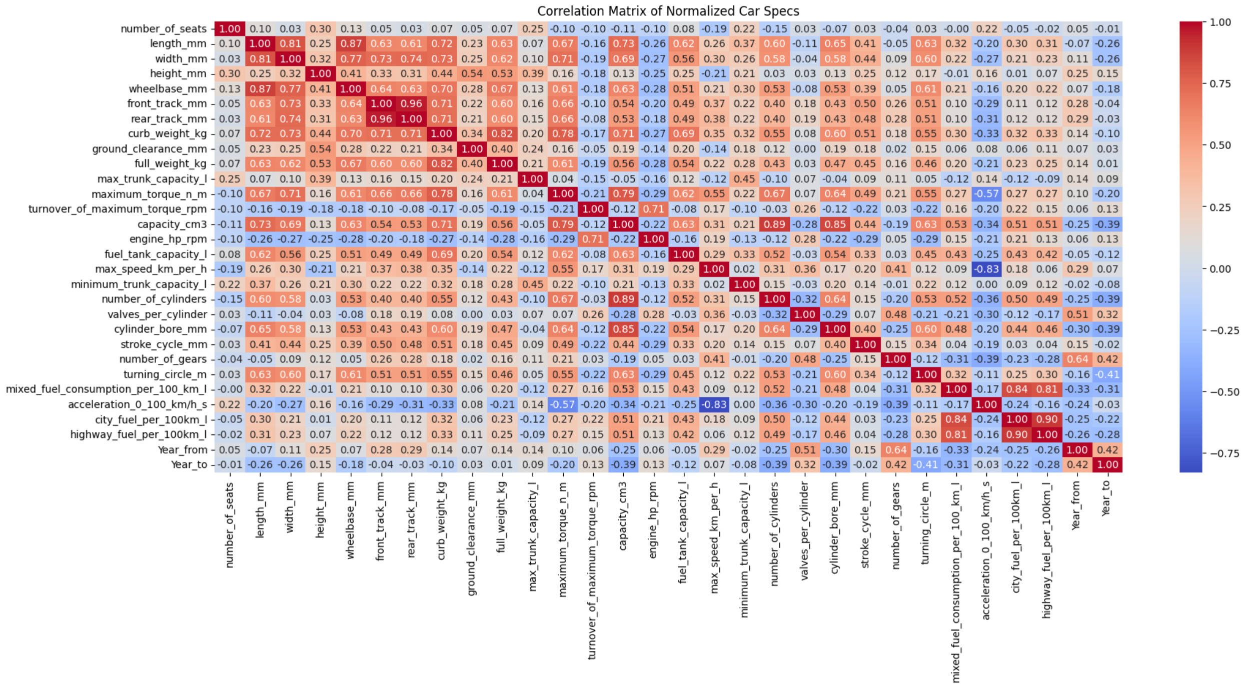
The above bar plot provides the information on “the top 20 car brands with most unique models” along with the count of the unique models for each car brand.



The above bar plot provides the information on “the top 10 engine types by count” with number of cars for each engine type.



The scatterplot above shows “Engine hp” vs “Engine capacity” of car_specs dataframe.



The above heat map represents the **correlation** between the numerical data in car_specs dataframe.

Conclusion from the heatmap:

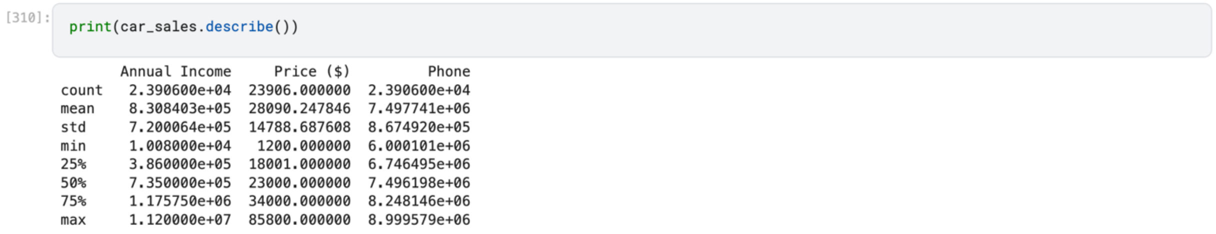
High Positive Correlations:

1. length_mm and width_mm **(corr = 0.81)**
2. length_mm and wheelbase_mm **(corr = 0.87)**
3. width_mm and length_mm **(corr = 0.81)**
4. wheelbase_mm and length_mm **(corr = 0.87)**
5. front_track_mm and rear_track_mm **(corr = 0.96)**
6. rear_track_mm and front_track_mm **(corr = 0.96)**
7. curb_weight_kg and full_weight_kg **(corr = 0.82)**
8. full_weight_kg and curb_weight_kg **(corr = 0.82)**
9. capacity_cm3 and number_of_cylinders **(corr = 0.89)**
10. capacity_cm3 and cylinder_bore_mm **(corr = 0.85)**
11. mixed_fuel_consumption_per_100_km_l and city_fuel_per_100km_l **(corr = 0.84)**
12. mixed_fuel_consumption_per_100_km_l and highway_fuel_per_100km_l **(corr = 0.81)**
13. city_fuel_per_100km_l and mixed_fuel_consumption_per_100_km_l **(corr = 0.84)**
14. city_fuel_per_100km_l and highway_fuel_per_100km_l **(corr = 0.90)**
15. highway_fuel_per_100km_l and mixed_fuel_consumption_per_100_km_l **(corr = 0.81)**
16. highway_fuel_per_100km_l and city_fuel_per_100km_l **(corr = 0.90)**

High Negative Correlations:

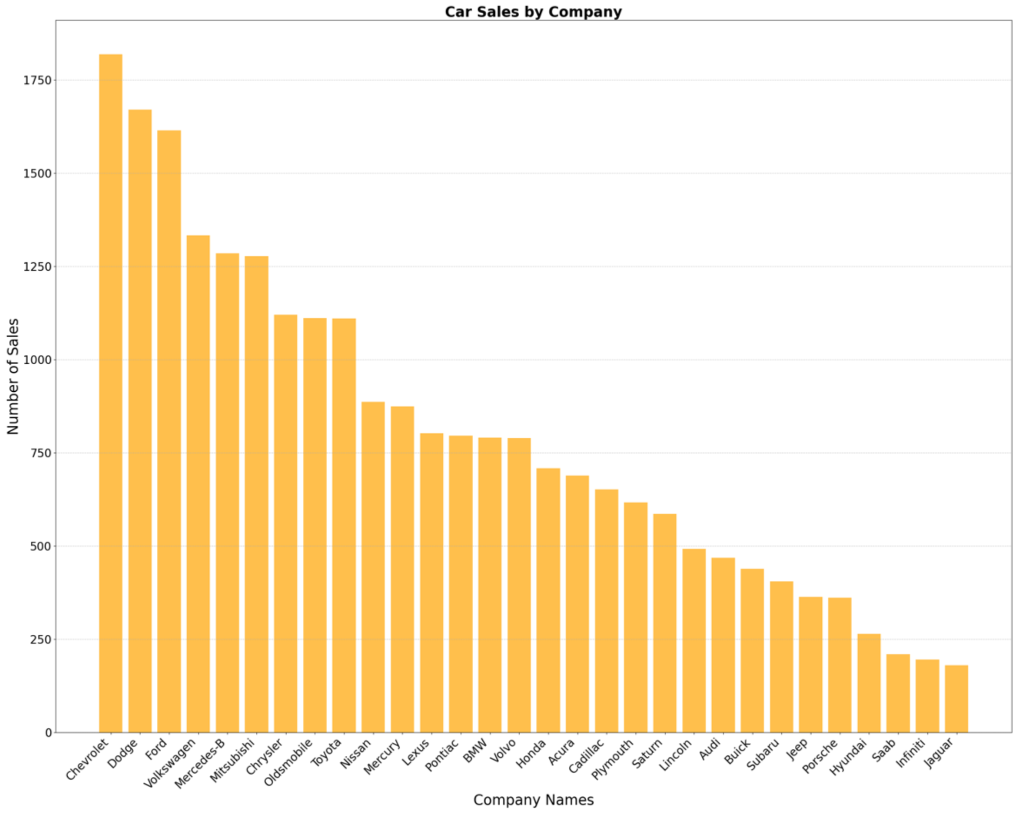
1. max_speed_km_per_h and acceleration_0_100_km/h_s **(corr = -0.83)**
2. acceleration_0_100_km/h_s and max_speed_km_per_h **(corr = -0.83)**

5.2 Car Sales (car_sales dataframe):

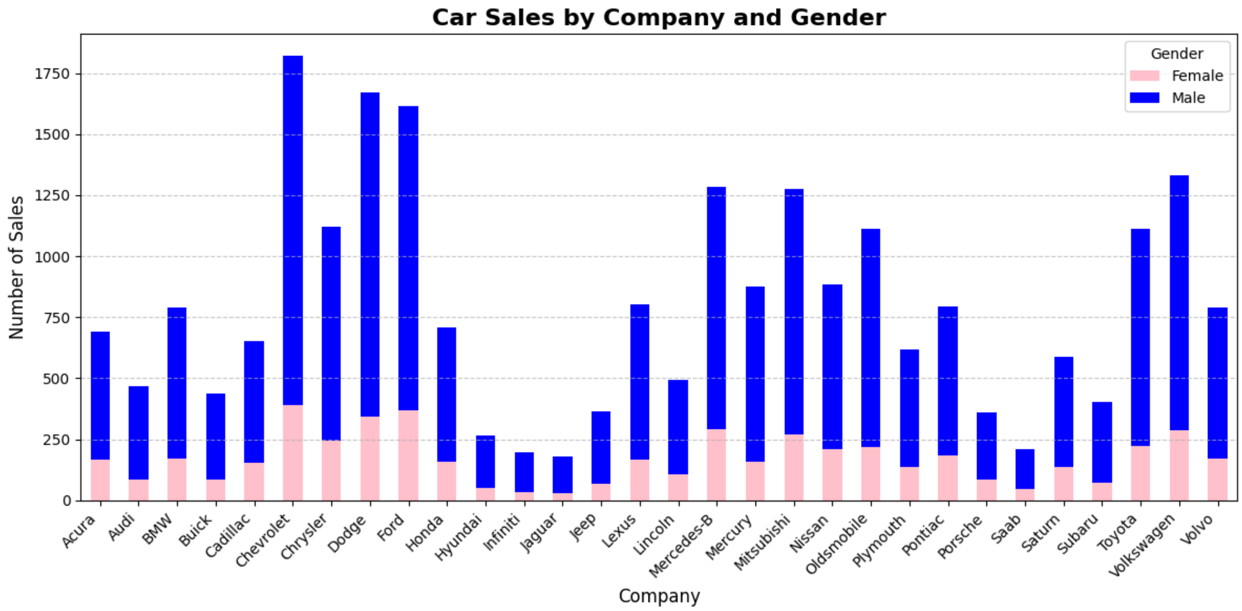


```
[310]: print(car_sales.describe())
```

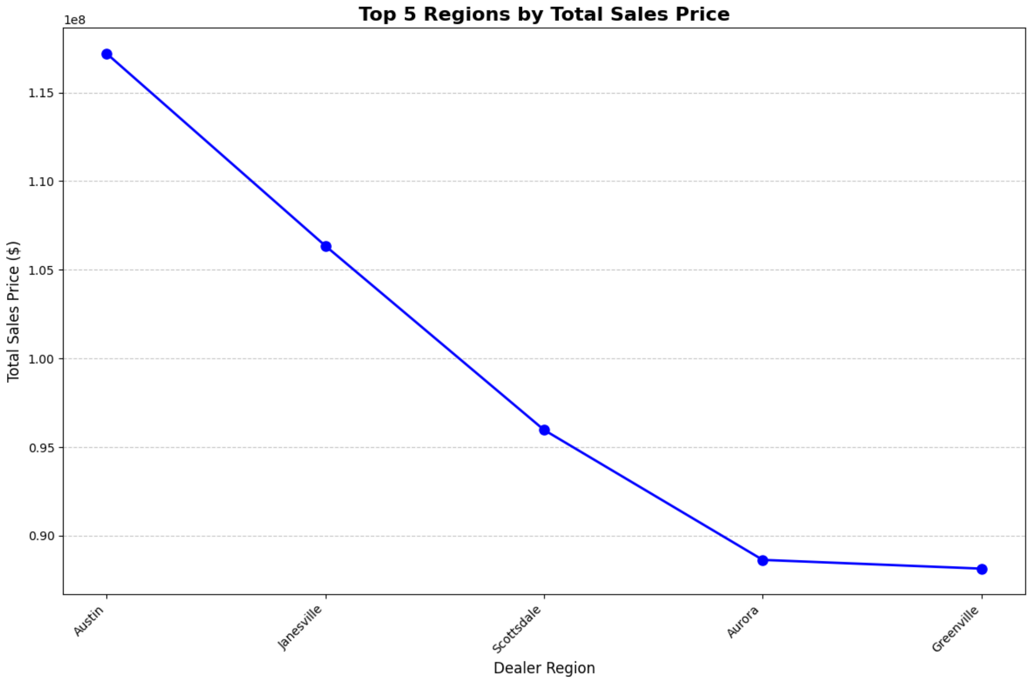
```
       Annual Income     Price ($)         Phone
count   2.390600e+04  23906.000000  2.390600e+04
mean    8.308403e+05  28090.247846  7.497741e+06
std     7.200064e+05  14788.687608  8.674920e+05
min     1.008000e+04   1200.000000  6.000101e+06
25%     3.860000e+05  18001.000000  6.746495e+06
50%     7.350000e+05  23000.000000  7.496198e+06
75%     1.175750e+06  34000.000000  8.248146e+06
max     1.120000e+07  85800.000000  8.999579e+06
```



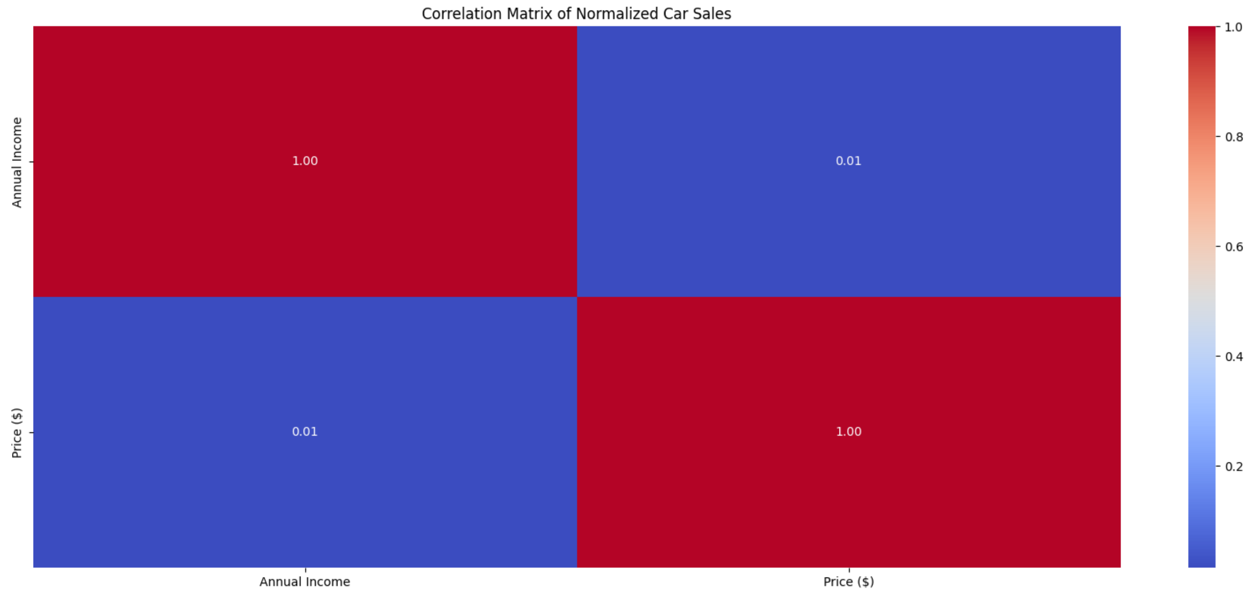
The barplot shows the “Car sales by company” wherein the y axis represents the ‘Number of Sales’ and ‘Company Names’.



The graph shows the Male and Female buyers for each car company.



The graph shows the total sales price for every dealer region in descending order.

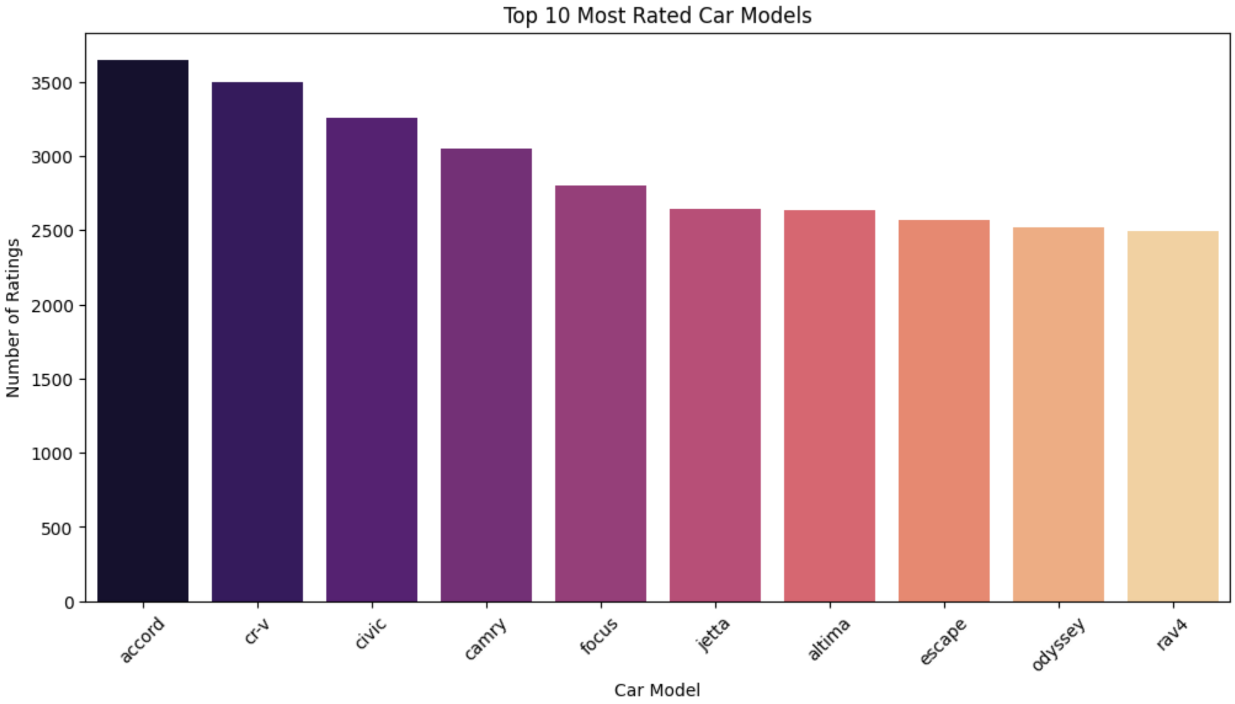


The Price and Annual income has a very weak correlation between them.

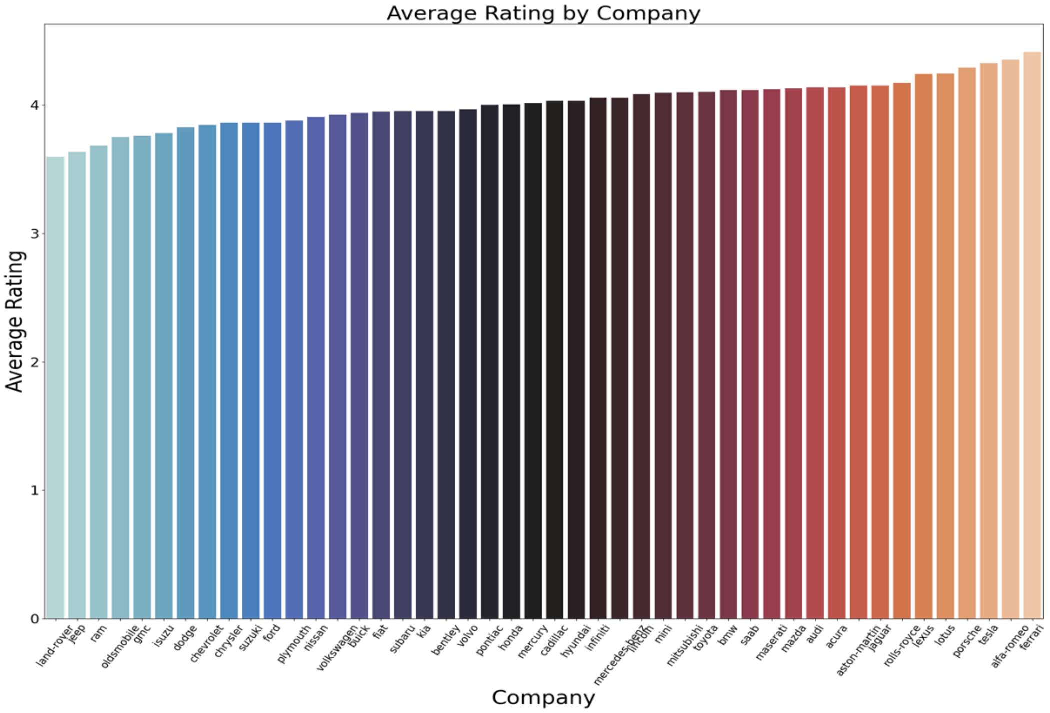
5.3 Car Ratings (car_ratings dataframe)

```
[333]:  print(car_ratings.describe())
```

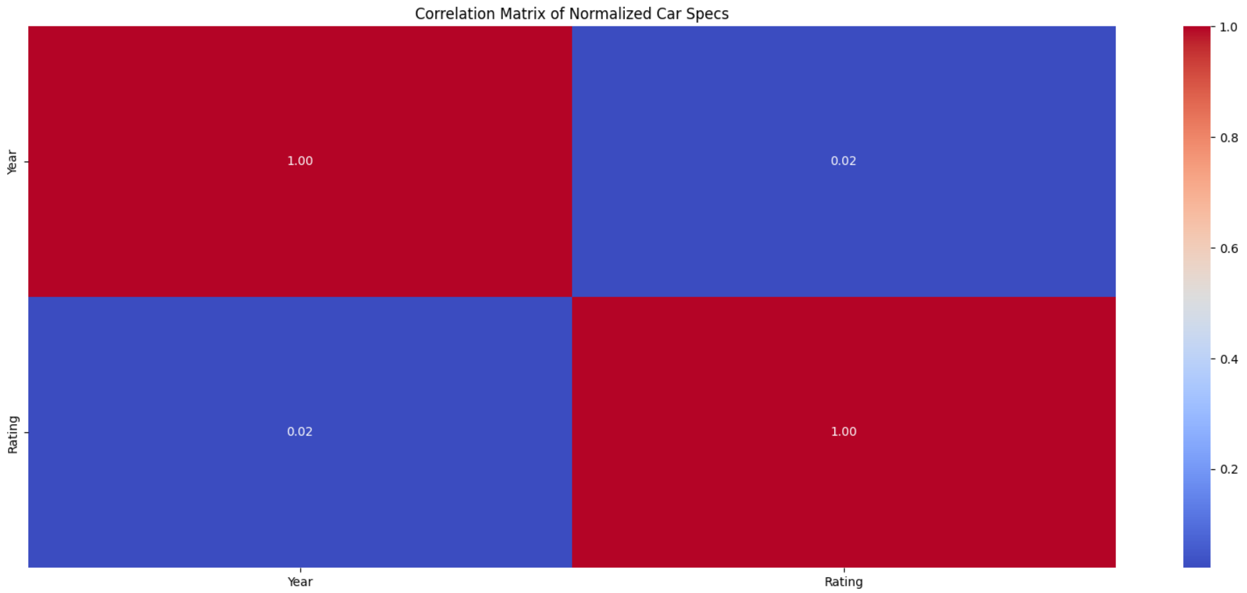
```
                Year         Rating
count  299045.000000  299045.000000
mean     2007.492247       3.980886
std         5.330847       0.993001
min      2000.000000       0.000000
25%      2003.000000       4.000000
50%      2006.000000       4.000000
75%      2011.000000       5.000000
max      2020.000000       5.000000
```



The bar plot shows top 10 most rated cars with Number of ratings for each car model.



The graph above shows the average rating of the cars sold by the respective companies.



The Year vs Rating is a very weak correlation between them based on the heatmap.

6. Project Timeline :

As part of milestone1 the datasets are taken from Kaggle which represent the Car_Specifications, Car_Sales and Car_Ratings.The distribution of the data has been identified through boxplots. The missing data and outliers have been handled for all the datasets. The data has been normalized for analysis. The visualizations are created to identify the major trends and relationships for the datasets. The correlation analysis is done on the data and the analysis is drawn.

In the upcoming milestones the focus is on encoding categorical variables using one-hot encoding. Also, evaluate the feature importance, reduce dimensionality and training of data models. The model performance and comparison of different models should be done and the prediction of the best car based on the user need should be done.

7. Conclusion:

In conclusion the datasets has been analyzed, loaded as dataframes from the three csv files(Car Specifications, Car Sales and Car Ratings) and the missing data is handled by visualizing **boxplots** and then using "**Median imputation**" and "**Mode Imputation**" to fill them.The outliers in each dataset has been analyzed using the "**IQR method"** and handled using the "**Log Transformation**" and the numeric data in all datasets are normalized using the "**MinMaxNormalization**". The **plots** and **visualizations** are created to identify the trends between the key columns in the dataset. The Correlations are analyzed on the numeric data that is normalized, and the conclusions have been drawn on the highly positive, negative correlated columns and weakly correlated columns in all the datasets.