PREDICT USED CAR PRICES

REGRESSION ANALYSIS

SECOND WIND WHEELS CAR PRICE PREDICTION

Wind Wheels is a leading online used car retailer, and wants to develop a machine learning model for predicting the accurate market value of used cars.

As Data Scientists ,our project aims to develop a machine learning model that can accurately predict used car prices.

This model will revolutionize the pricing strategy, boost profits and enhance customer satisfaction.

PROBLEM STATEMENT

Currently, Second Wind Wheels' used car pricing relies heavily on manual valuations by human appraisers. This method is prone to subjectivity, inconsistencies, and delays, leading to:

- Overpricing: Cars priced too high stagnate on the lot, incurring storage costs and lost sales.
- Underpricing: We miss out on maximizing profits by selling cars for less than their true market value.
- Customer dissatisfaction: Unfairly priced cars may discourage buyers and damage brand reputation

MAIN OBJECTIVES

- Develop a reliable and accurate price prediction model for used cars.
- Improve pricing transparency and market efficiency by providing data-driven valuations.
- Provide a user-friendly interface for individuals to input car attributes and obtain an estimated resale price.

SPECIFIC OBJECTIVES

- Collect and preprocess a comprehensive dataset of used cars, including various attributes such as make, model, year, mileage, and historical prices.

- • Explore and apply various data mining and machine learning techniques, such as regression models, to identify the most accurate price prediction model.

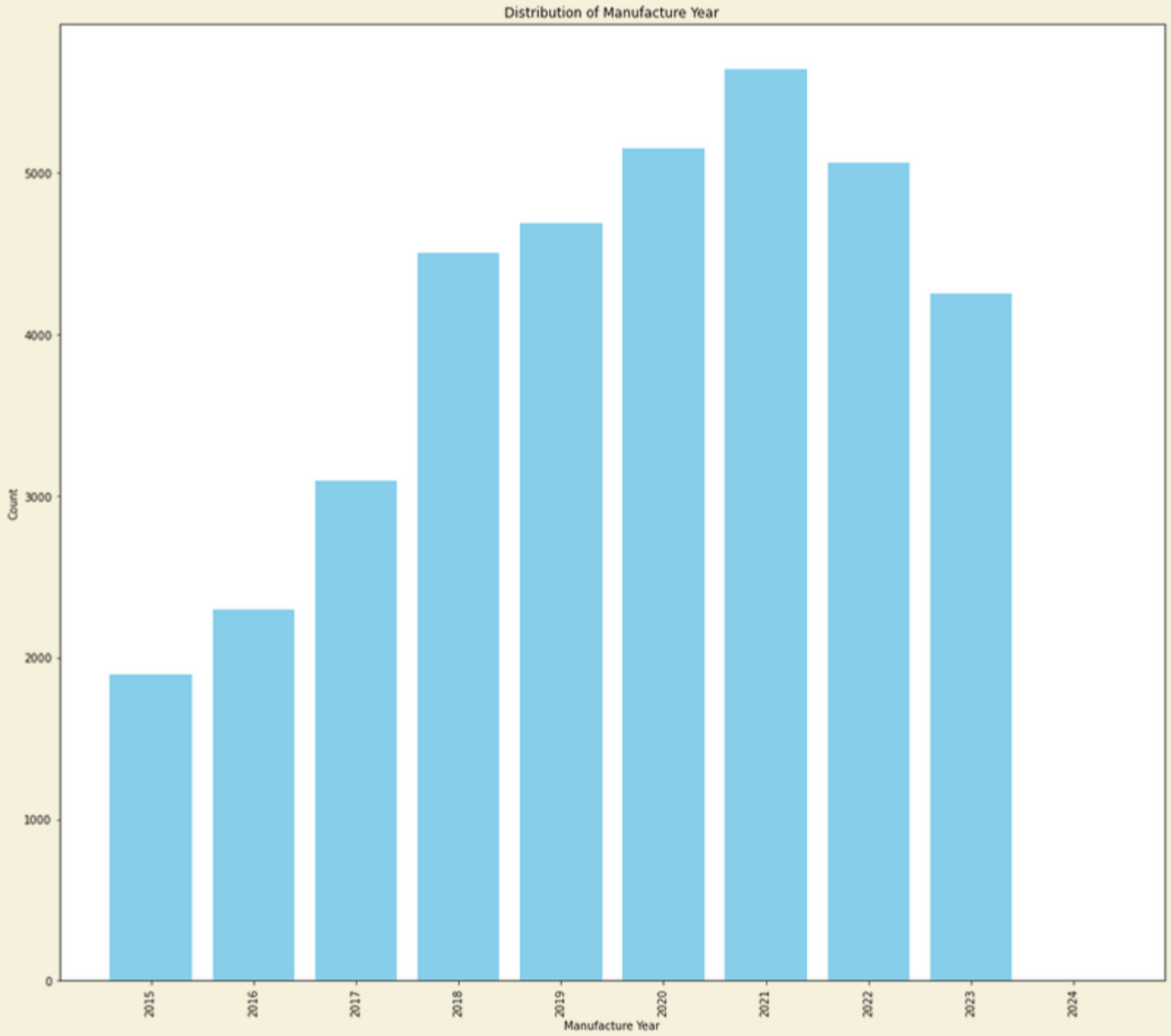
- • Evaluate the performance of the developed model and ensure it meets the predefined accuracy and reliability standards.

DATA UNDERSTANDING

We conducted web scraping on the Sbt Japan website, extracting data categorized by the vehicles' body types. Subsequently, we consolidated the gathered information into a unified data-frame, denoted as the "merged data." From this dataset, we created another dataframe, referred to as "Our_df," focusing on features that we identified as having a significant impact on the vehicle's price. The columns within Our_df were carefully chosen and renamed for enhanced clarity.

- **Make** : The make and model of a particular car and its manufacture year eg.2015/9 Toyota 86
- **mileage(km)** : The number of kilometers that the vehicle has covered
- **capacity(cc)** : The engine size ratings in cubic centimeters
- **Transmission** : The vehicles transmission, either automatic or manual transmission
- **Fuel** : The type of fuel that runs the vehicle
- **steering** : The side that the steering wheel is on eg Right Hand drive
- **drive** : The number of wheels powered by the engine
- **seats** : Number of seats in the vehicle
- **doors** : Number of doors the vehicle has
- **body_type** : The overall body type of the vehicle eg suv
- **price(Ksh)** : The total cost in Ksh of the vehicle while at port of Mombasa including the freight, insurance and inspection fees
- We introduced a new column "Manufacture_year" that shows the year a particular vehicle was manufactured. This was derived from the "Make" column

Exploratory Data Analysis

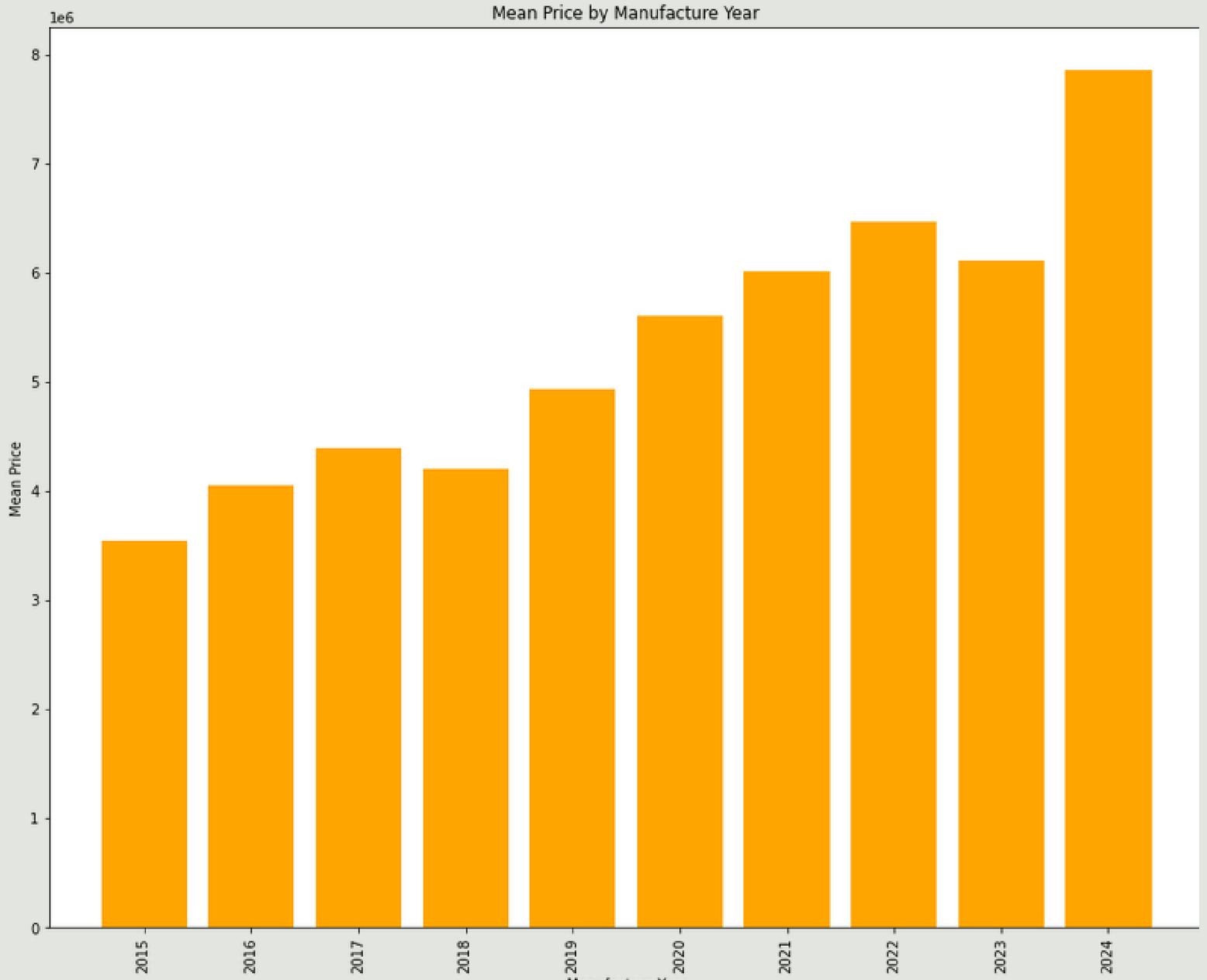


 # Extracting the Manufacture_year column values and their counts

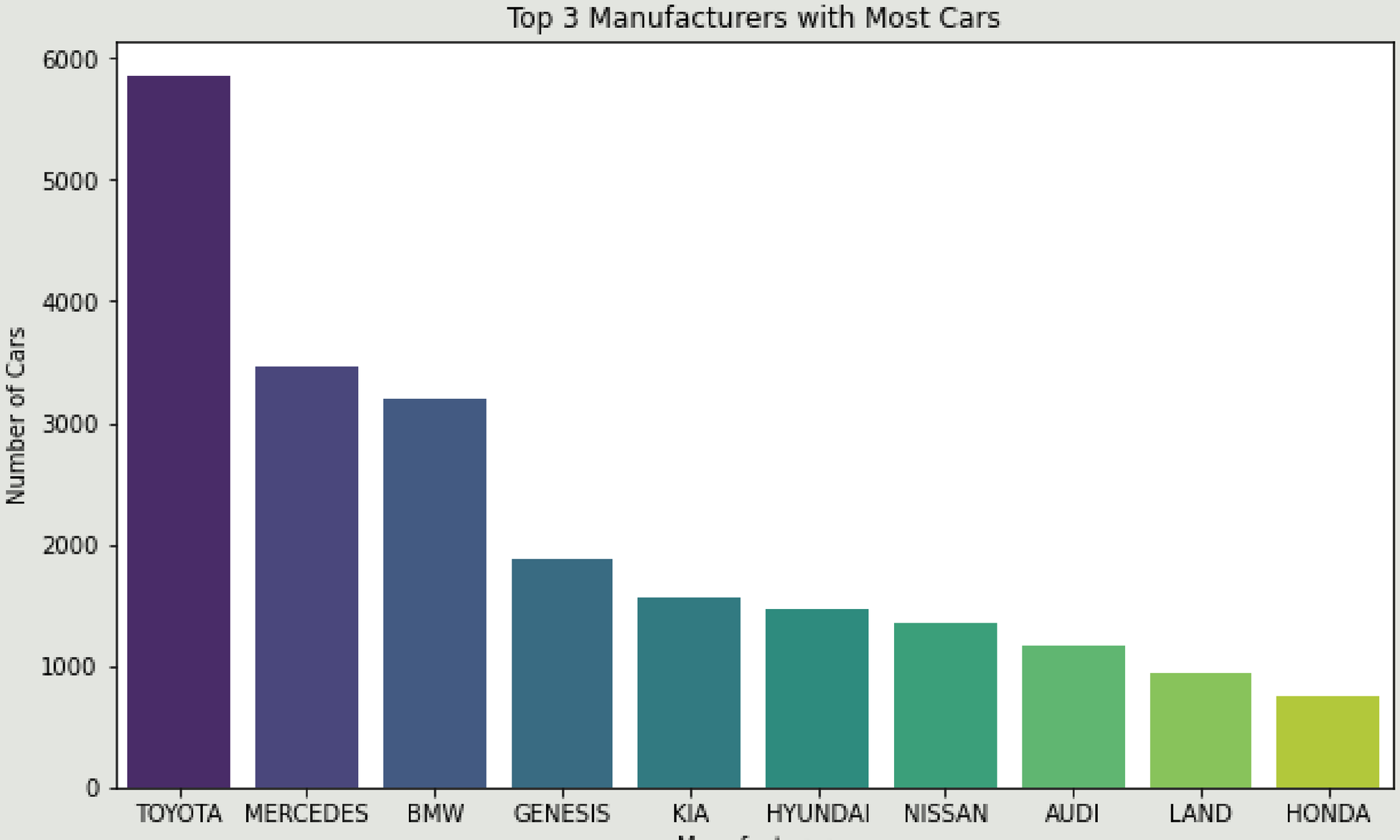
The highest values were observed in Year 2021 while the least number was in 2015

Bar graph showing the year of manufacture and the mean price

Year 2015 had the least mean price while 2024 has the highest mean price

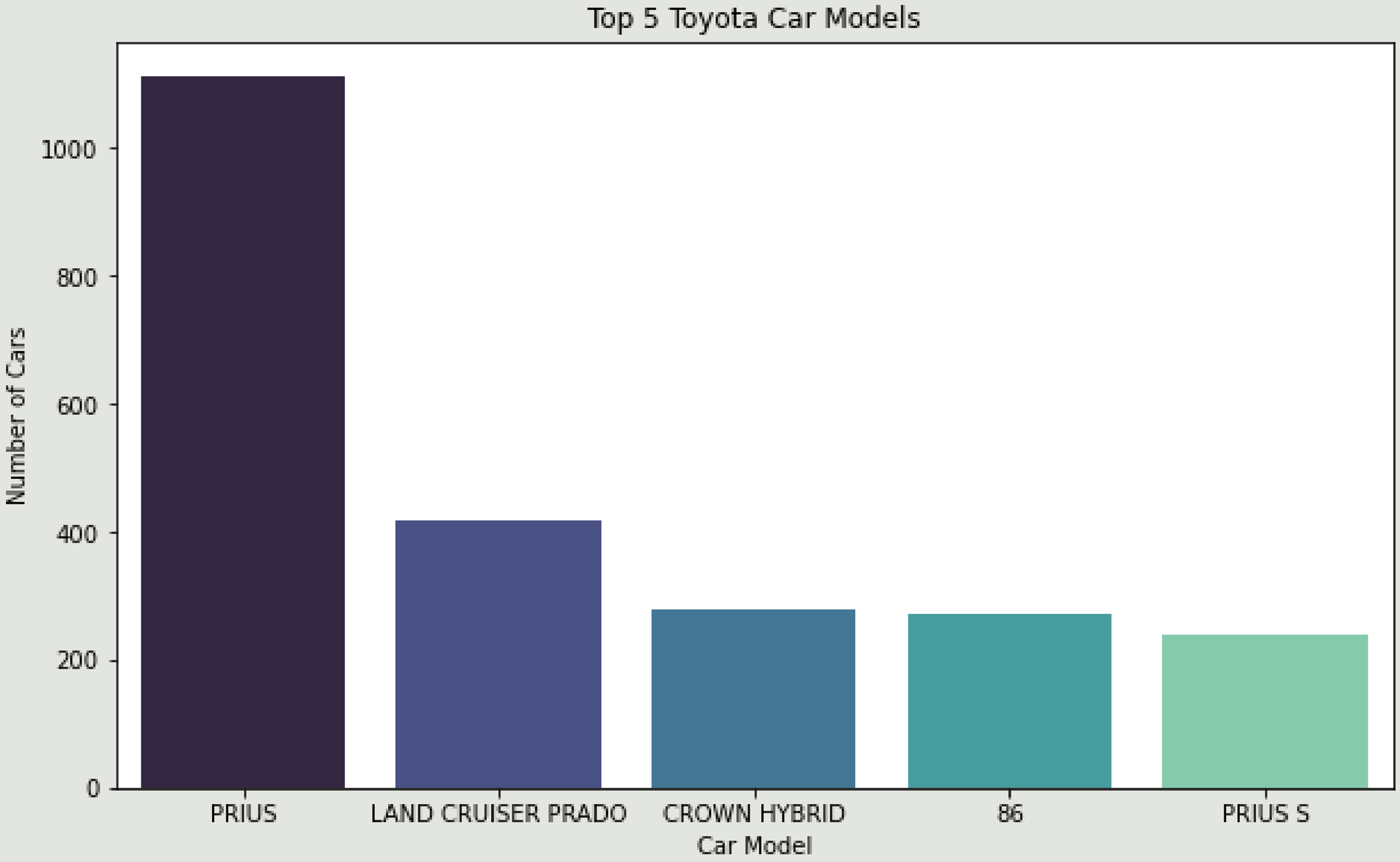


#Extracting top 10 manufacturers



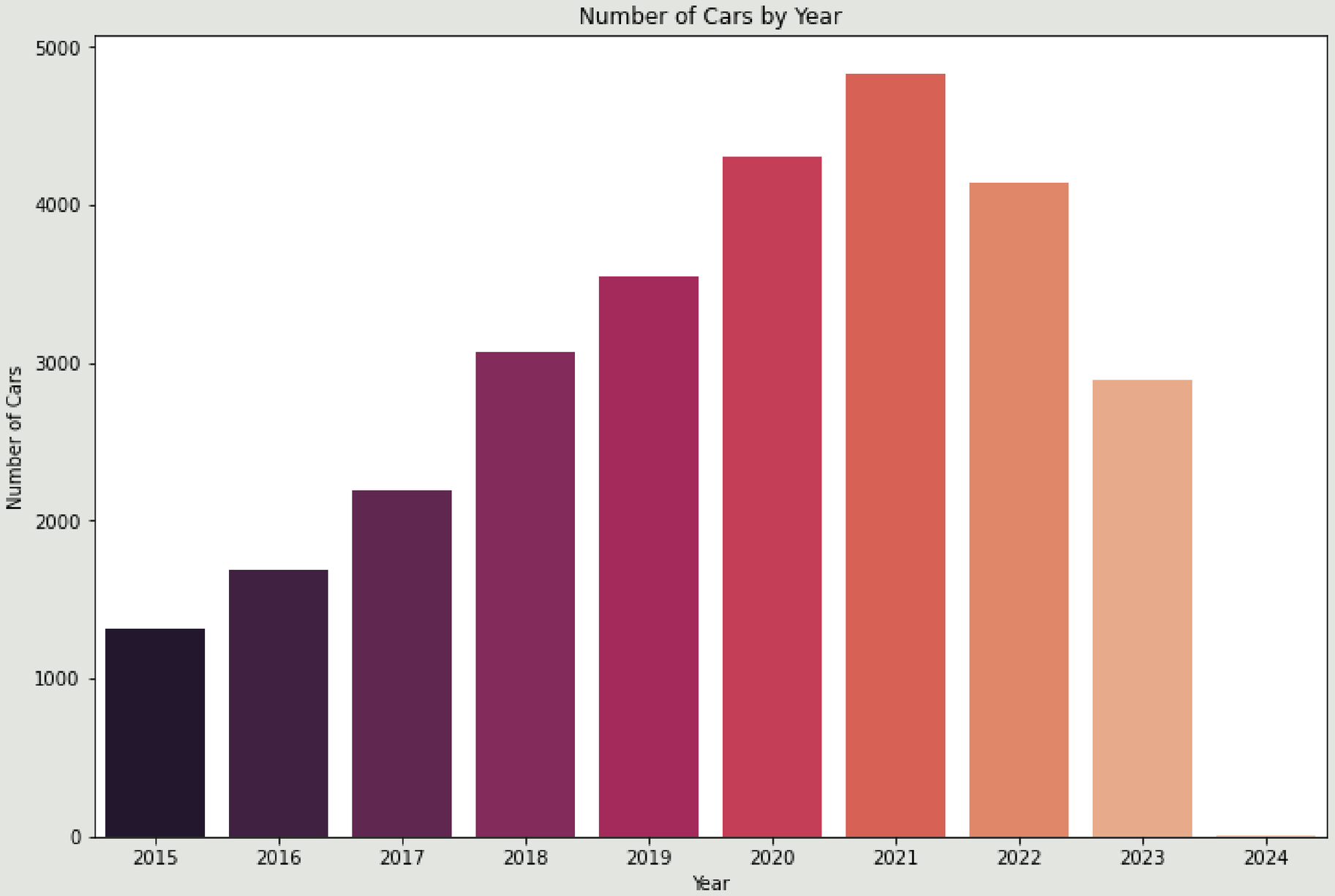
Toyota emerged as the predominant car manufacturer

Graph for Top 5 Toyota models



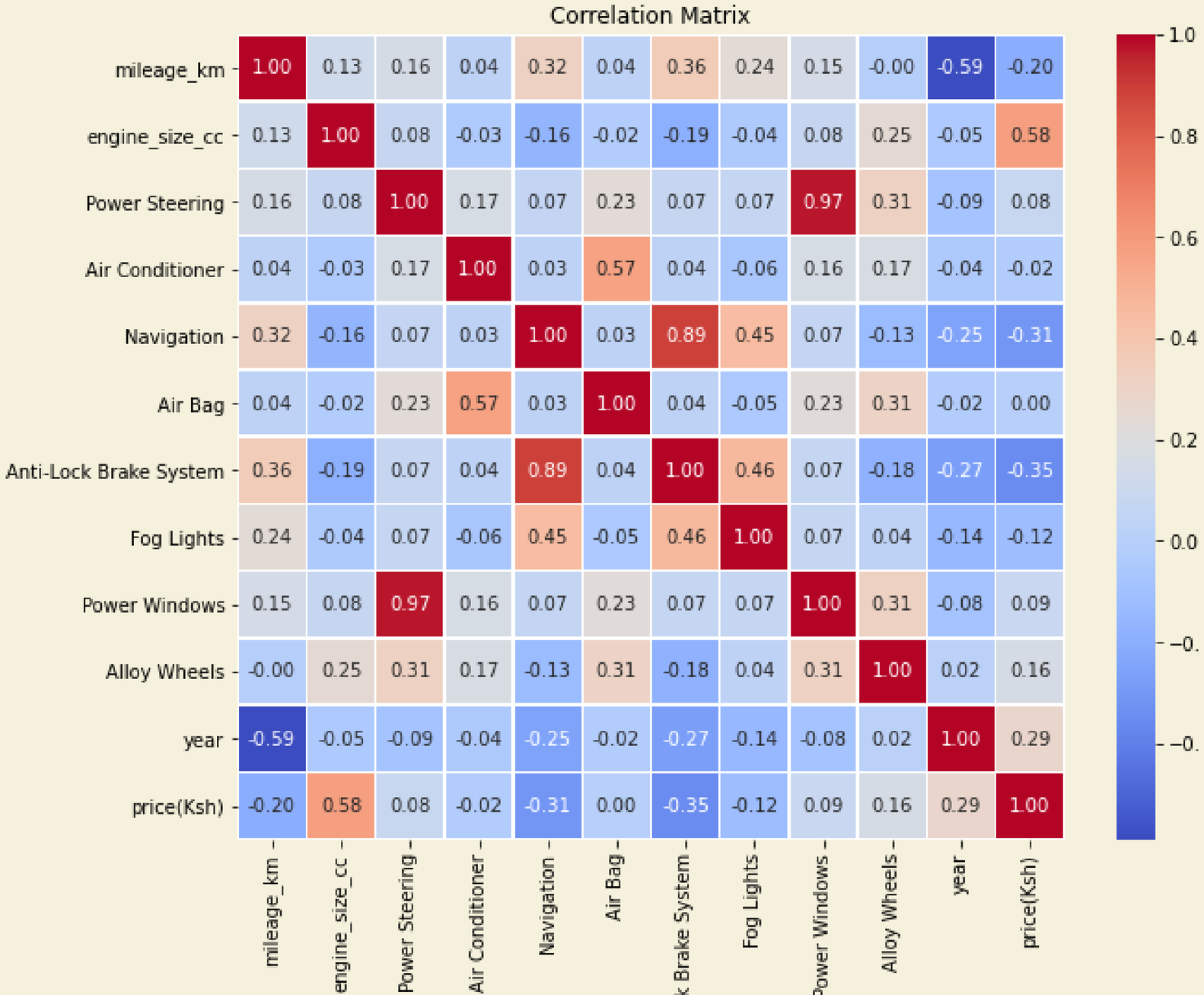
The Prius, Landcruiser Prado, and Crown hybrid are the most frequently encountered models.

Number of cars by year



The highest and least number of vehicles were manufactured in 2021 and 2015 respectively.

Correlation matrix



Correlation matrix Analysis

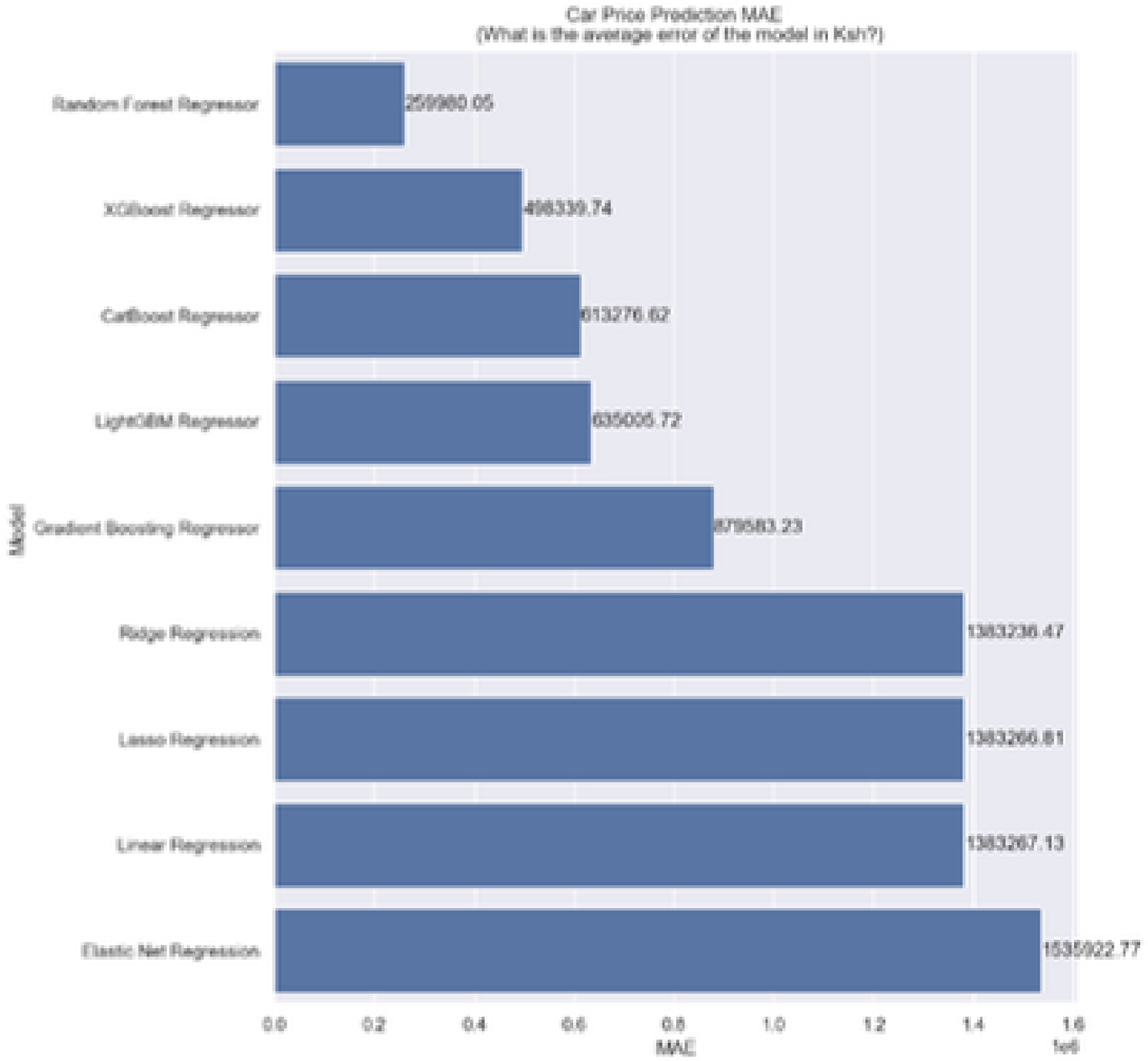
- The heatmap visually represents the interrelationships between various features in our dataset.
- Focusing on our target feature, which is the price(Ksh), it becomes evident that engine_size_cc exhibits the highest positive correlation. This indicates a significant influence of the engine capacity on the overall price of a vehicle.
- On the contrary, features such as the Anti-Lock Brake System and Navigation demonstrate the least impact on the pricing structure, as reflected by their lower correlation coefficients with the target variable.

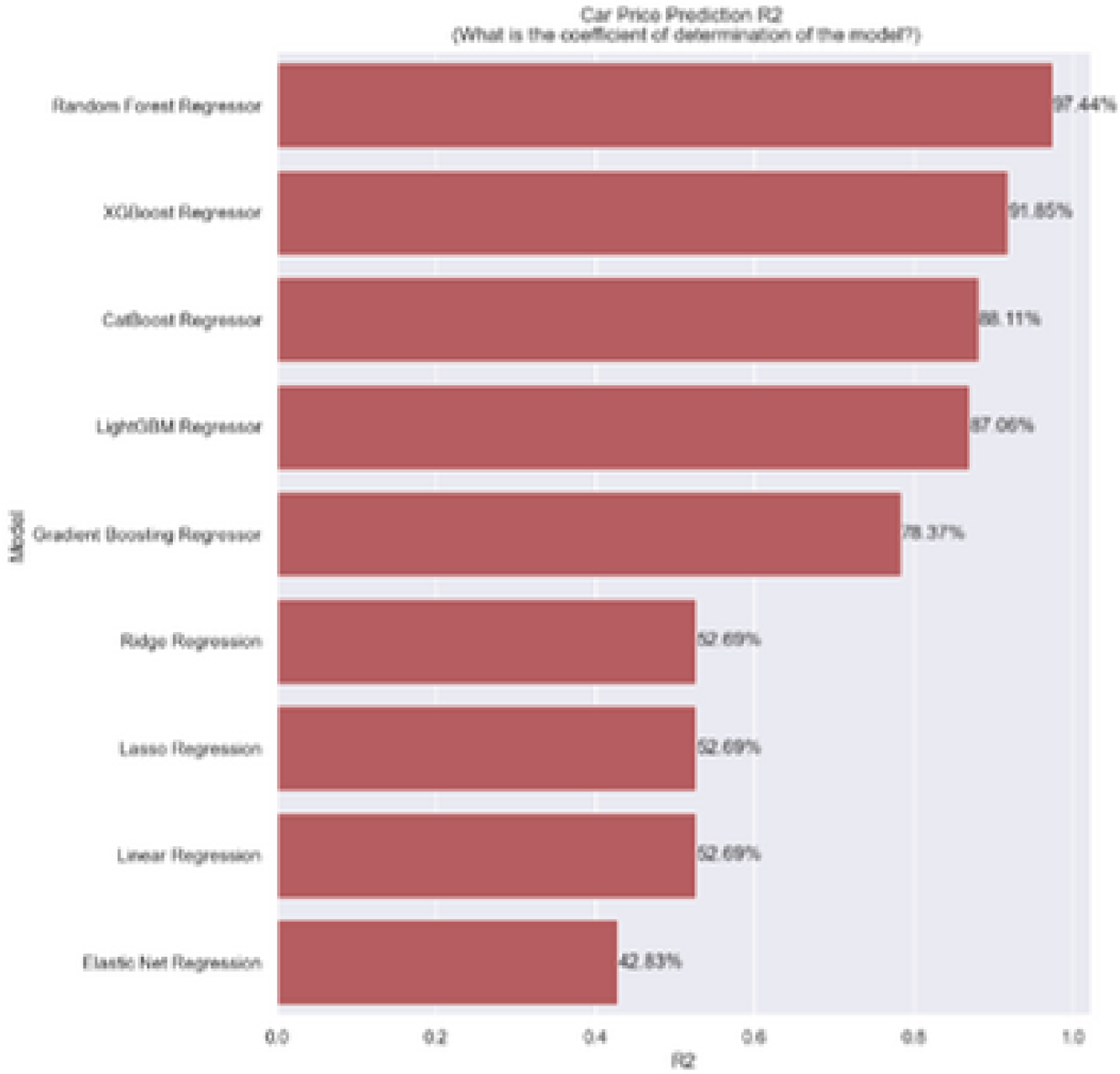
MODELLING

This involved the selection, training, and tuning of machine learning models.

	Model	MAE	R2
4	Random Forest Regressor	2.599801e+05	0.974385
6	XGBoost Regressor	4.983397e+05	0.918524
8	CatBoost Regressor	6.132766e+05	0.881084
7	LightGBM Regressor	6.350057e+05	0.870578
5	Gradient Boosting Regressor	8.795832e+05	0.783670
1	Ridge Regression	1.383236e+06	0.526867
2	Lasso Regression	1.383267e+06	0.526867
0	Linear Regression	1.383267e+06	0.526867
3	Elastic Net Regression	1.535923e+06	0.428335

Model Performance Comparison





Performance on The Training Data:

- Based on the table, we can see that the Random Forest Regressor has the best performance among all the models, as it has the lowest MAE, and the highest R2 Score.
- This means that it has the smallest average error and the best fit to the data.
- The Linear Regression model has the worst performance, as it has the highest MAE, and a zero R2 Score.
- This means that it has the largest average error and no fit to the data.
- The other models have varying degrees of performance, but none of them can match the Random Forest Regressor.

Baseline Random Forest Regressor

Evaluation: Baseline Random Forest Regressor

Training MAE : 260009.58
Training : R2 0.97
Testing MAE: 625960.75
Testing: R2 0.86
Baseline MAE: 2.046670e+06

- The model has a low MAE and a high R2 score on the training dataset, which indicates that it fits the data well and has a low prediction error.
- The model has a higher MAE and a lower R2 score on the testing dataset, which suggests that it generalizes less well to new data and has a higher prediction error.
- The model performs much better than the baseline MAE, which is the average absolute difference between the actual values and the mean value of the dependent variable. This means that the model has some predictive power and is better than a naive guess.

Baseline CatBoost Regressor

Evaluation: Baseline CatBoost Regressor

Training MAE: 563189.520349
Testing MAE :613106.379644
Baseline MAE:2.046670e+06
Training R2:0.896449
Testing R2:0.875161

Tuned CatBoost Regressor - (RandomSearchCV)

Evaluation: Tuned CatBoost Regressor - (RandomSearchCV**)**

Training MAE:540642.706753
Testing MAE:605433.891931
Baseline MAE:2.046670e+06
Training R2 :0.904717
Testing R2:0.875628

Tuned CatBoost Regressor- (GridSearchCV)

Evaluation: Tuned CatBoost Regressor- (GridSearchCV**)**

Training MAE :597401.318922
Testing MAE :640358.513114
Baseline MAE :2.046670e+06
Training R2 :0.886204
Testing R2:0.866758

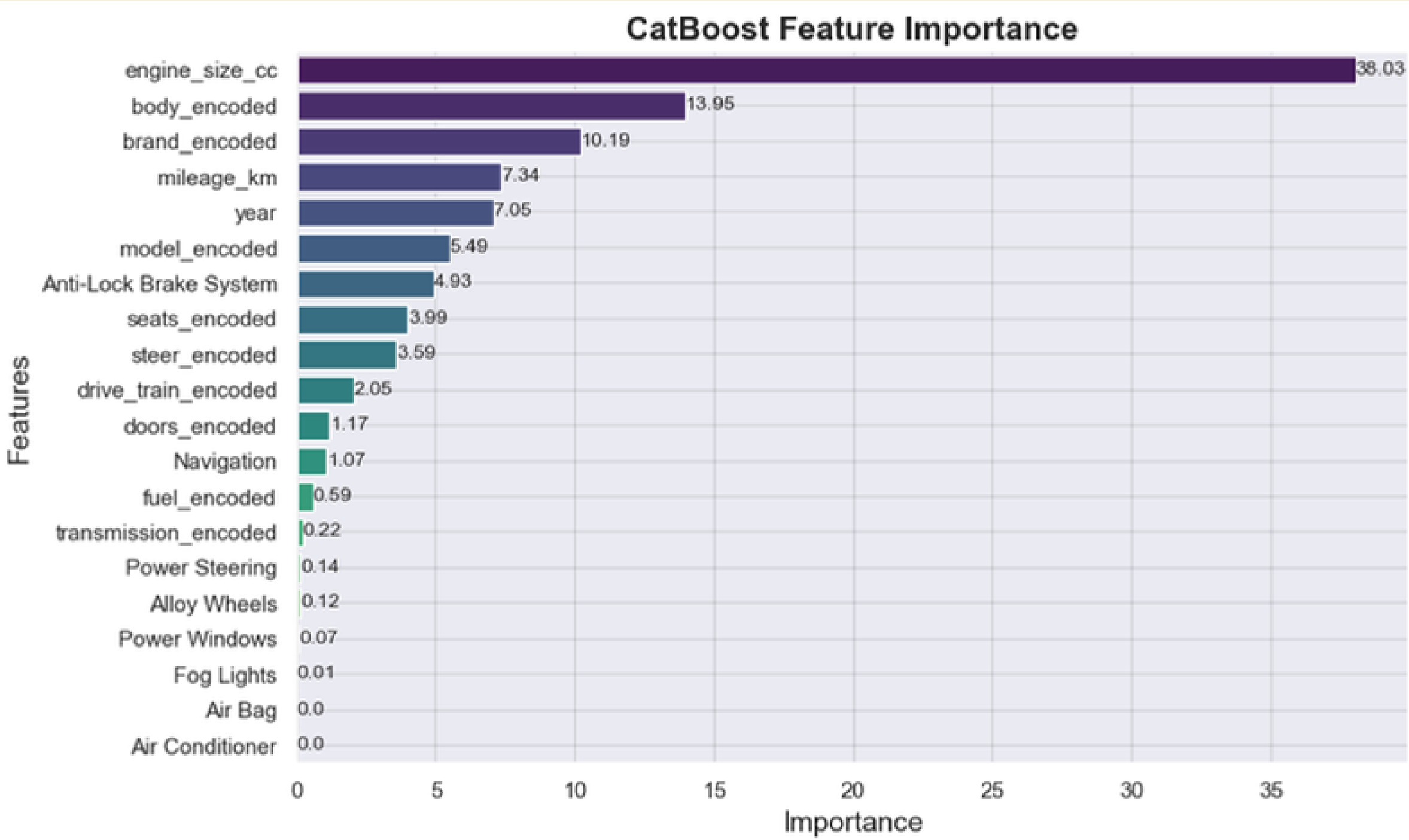

Training Data: Actual vs Predicted
1e7
Predicted Targets (y_pred_train)
Targets (y_train)
1e7
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4


Testing Data: Actual vs Predicted
1e7
Predicted Targets (y_pred_test)
Targets (y_test)
1e7
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4

	Prediction	Target	Residual	Difference%
count	5.600000e+03	5.600000e+03	5.600000e+03	5600.000000
mean	5.594490e+06	5.583627e+06	-1.086278e+04	12.760160
std	2.481781e+06	2.702921e+06	9.865684e+05	14.166943
min	2.572278e+05	5.455220e+05	-7.088808e+06	0.004615
25%	3.388812e+06	3.598264e+06	-4.515363e+05	3.820635
50%	5.573605e+06	5.219279e+06	-7.183828e+04	8.632938
75%	6.789720e+06	6.787001e+06	3.730535e+05	16.717385
max	1.365722e+07	1.215585e+07	9.278724e+06	145.146736

Tuned CatBoost Regressor- (GridSearchCV - 2)

Training MAE 800436.53
Traning Score: 0.8133501892903309
Testing MAE 813189.17
Testing Score: 0.8118500744644939


CatBoost Feature Importance
engine_size_cc
body_encoded
brand_encoded
mileage_km
year
model_encoded
Anti-Lock Brake System
seats_encoded
steer_encoded
drive_train_encoded
doors_encoded
Navigation
fuel_encoded
transmission_encoded
Power Steering
Alloy Wheels
Power Windows
Fog Lights
Air Bag
Air Conditioner
38.03
13.95
10.19
7.34
7.05
5.49
4.93
3.99
3.59
2.05
1.17
1.07
0.59
0.22
0.14
0.12
0.07
0.01
0.0
0.0
Features
0
5
10
15
20
25
30
35
Importance

- The most important feature for the model is engine_size_cc, which has a score of 38.03. This means that the size of the engine has a strong influence on the price of the car.

- The second most important feature is body_encoded, which has a score of 13.95. This means different body types may have different levels of demand, comfort, and functionality.

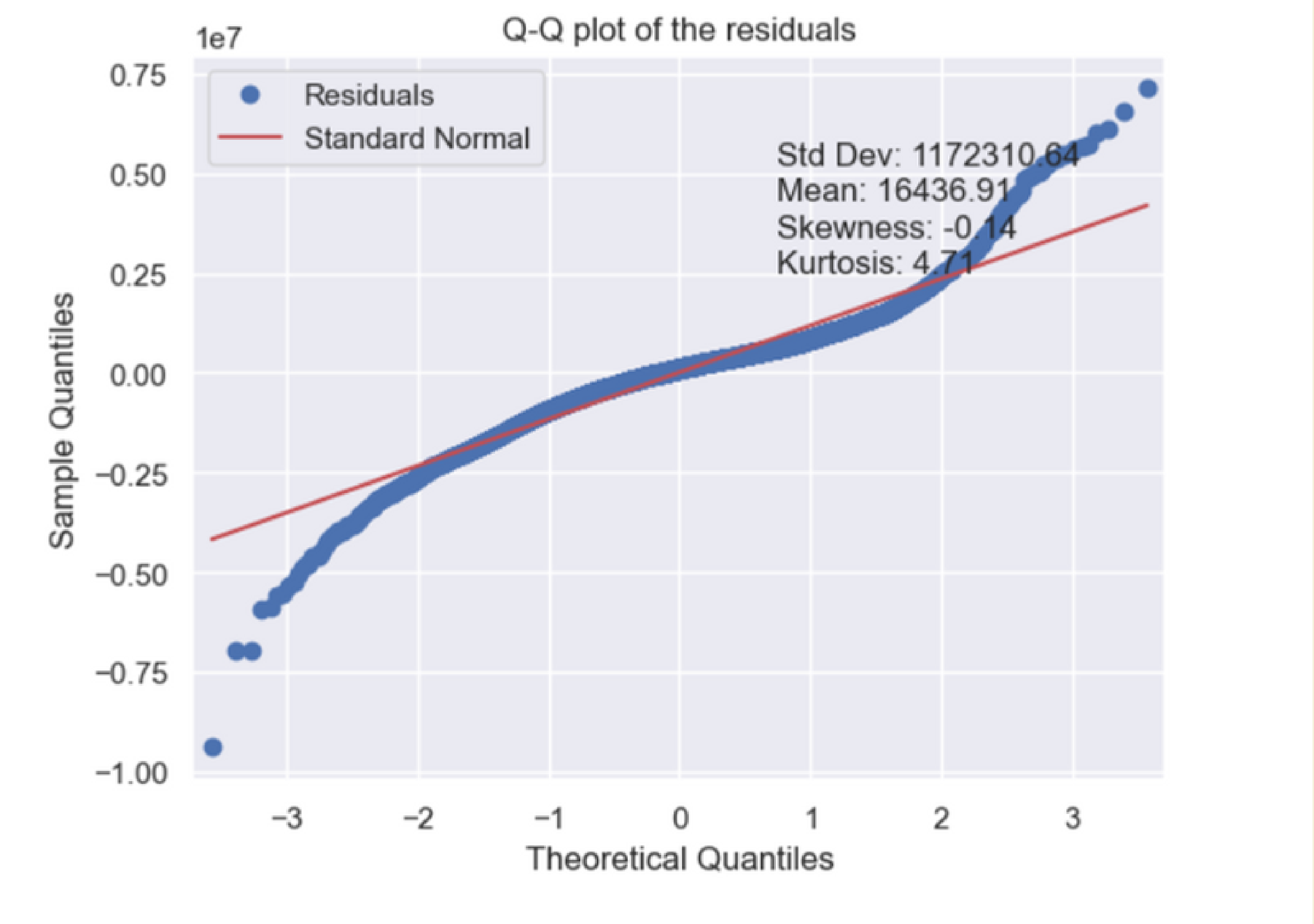
- The third most important feature is brand_encoded, which has a score of 10.19. This means that different brands may have different reputations, quality, and customer loyalty.

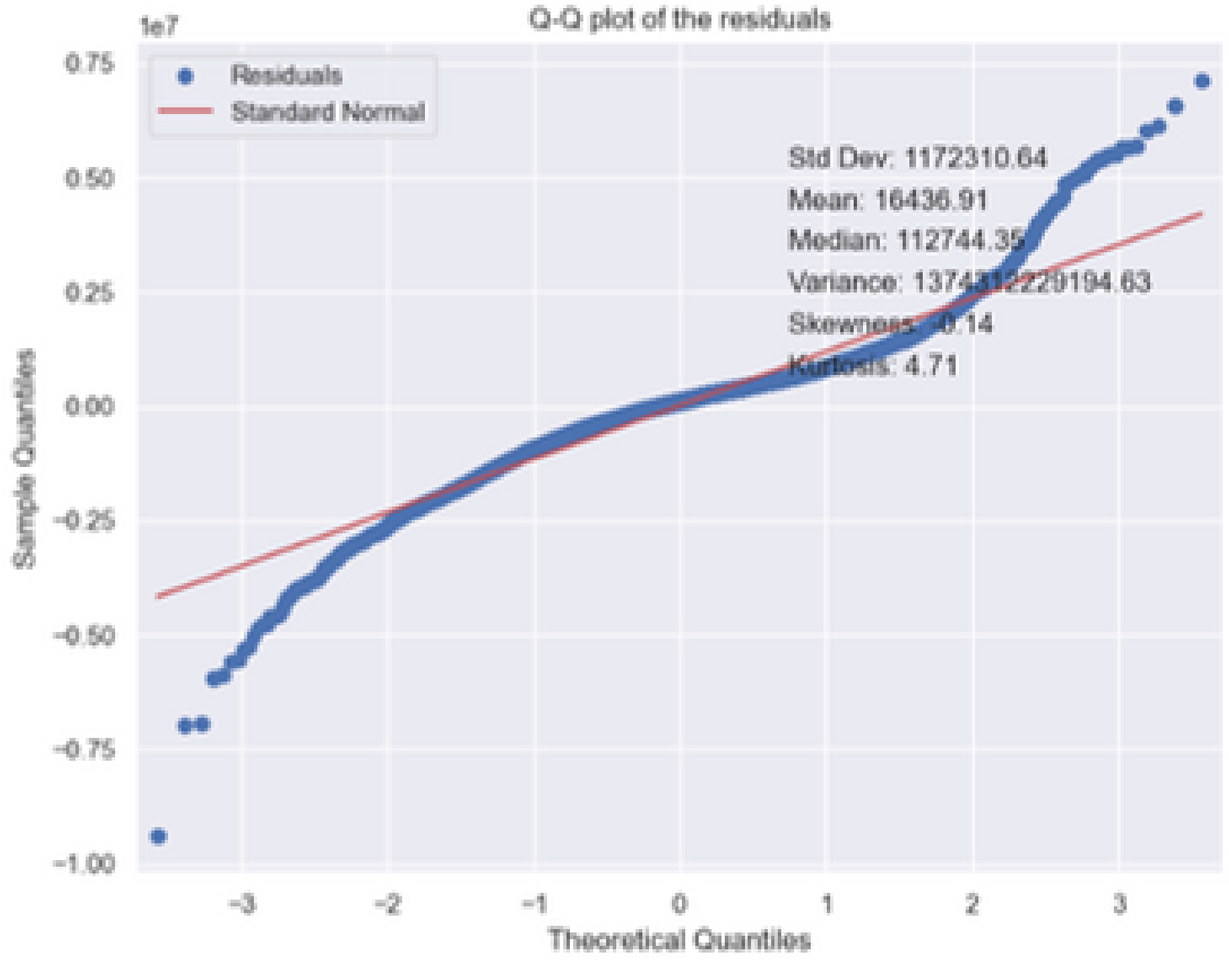
- The fourth and fifth most important features are mileage_km and year, which have scores of 7.34 and 7.05, respectively. These features represent the distance traveled by the car in kilometers and the year of manufacture of the car. These features indicate the age and condition of the car, which may affect its value and performance.

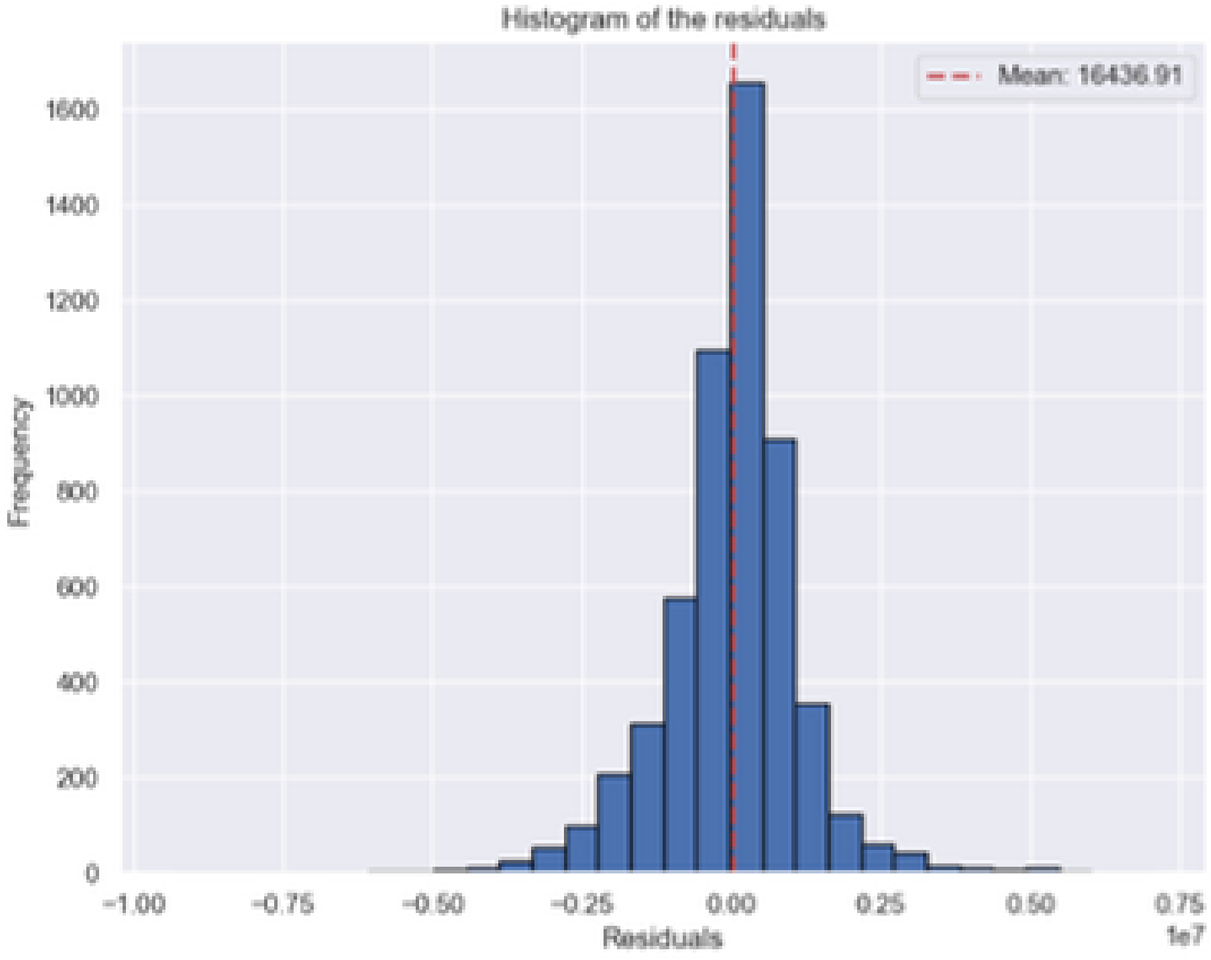
- The rest of the features have lower scores, ranging from 5.49 to 0.00. These features include the model of the car, the presence of various accessories and safety features, the number of seats and doors, the type of steering, drive train, fuel, and transmission. These features may have some impact on the price of the car, but not as much as the top five features.

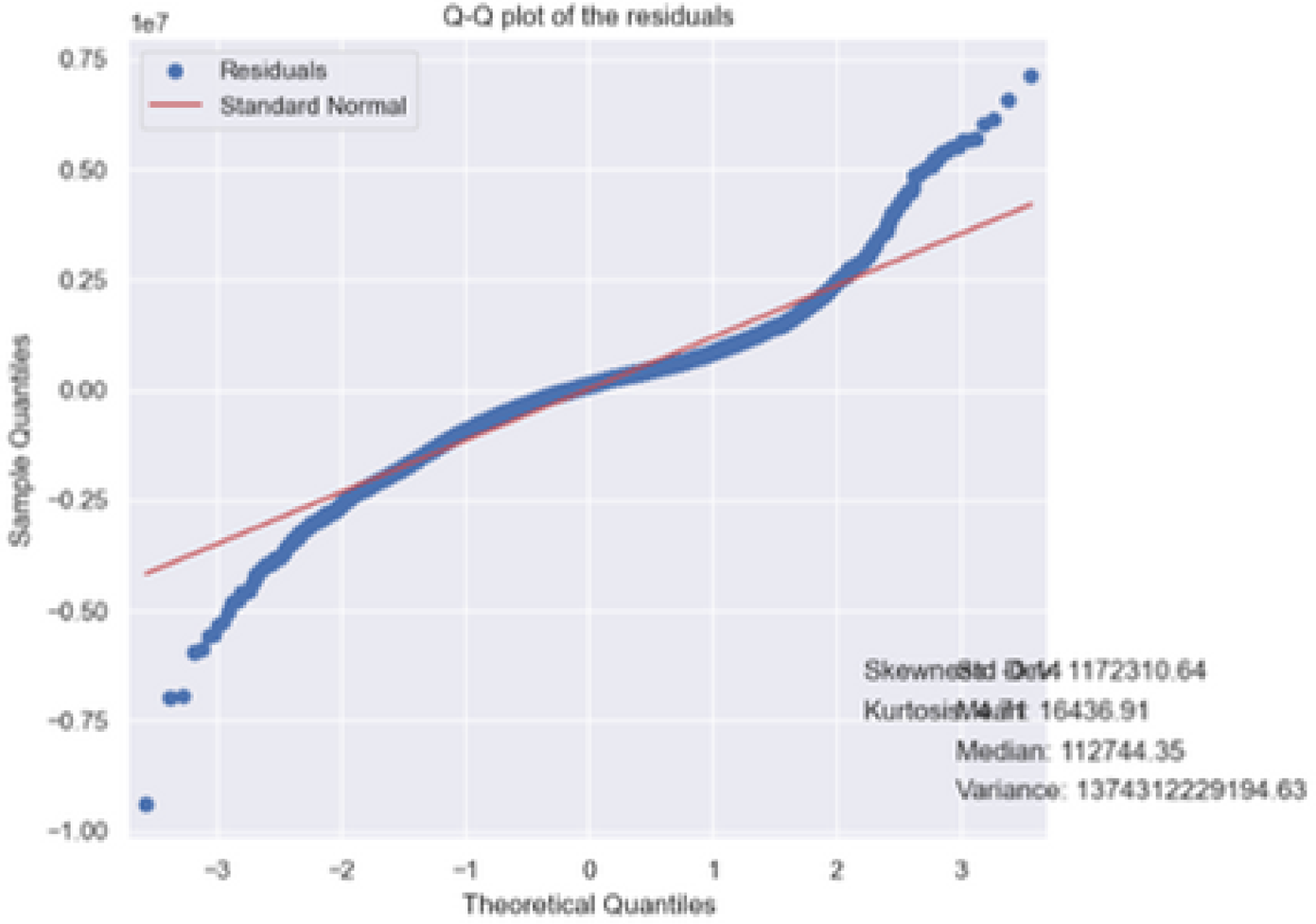
- If you are building a machine learning model to predict the price of a car, you should focus on the top five features, as they have the most influence on the outcome.
- The lower-ranked features may be included , but they may not add much value to the model.
- Pay attention to the top five features if you are buying or selling a car, as they may determine the fair price of the car.
- The lower-ranked features ay be considered, but they may not affect the price significantly.

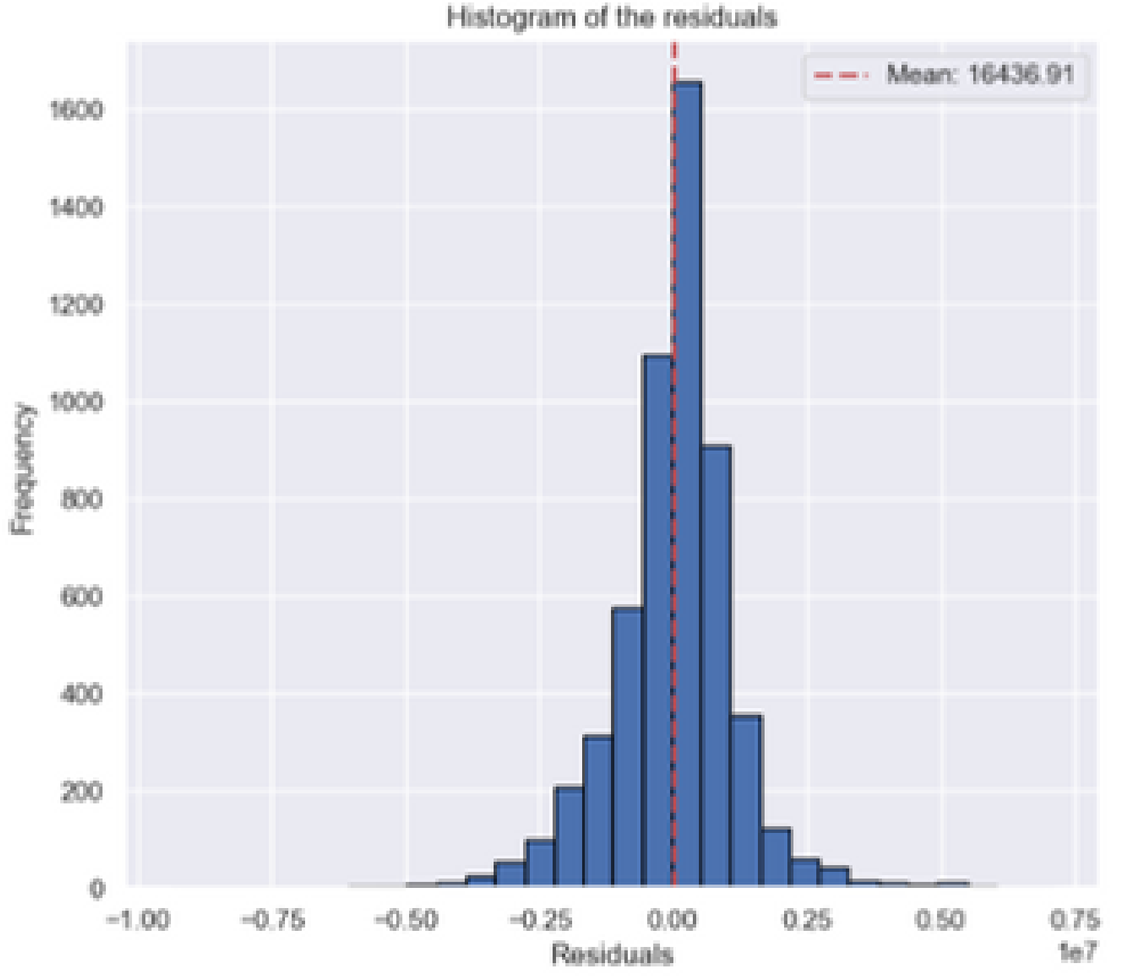
QQ-plot to check for normality of the residuals.




Q-Q plot of the residuals
1e7
Residuals
Standard Normal
Std Dev: 1172310.64
Mean: 16436.91
Median: 112744.35
Variance: 1374
229194.63
Skewness: 0.14
Kurtosis: 4.71
Sample Quantiles
Theoretical Quantiles
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
−3
−2
−1
0
1
2
3


Histogram of the residuals
Mean: 16436.91
Frequency
Residuals
1e7
1600
1400
1200
1000
800
600
400
200
0
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75


Q-Q plot of the residuals
1e7
Residuals
Standard Normal
Sample Quantiles
Theoretical Quantiles
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
−3
−2
−1
0
1
2
3
Std Dev: 1172310.64
Mean: 16436.91
Median: 112744.35
Variance: 1374312229194.63


Histogram of the residuals
Mean: 16436.91
Frequency
Residuals
1e7
1600
1400
1200
1000
800
600
400
200
0
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75

Q-Q plot: This shows how the quantiles (percentiles) of the residuals compare to the quantiles of a standard normal distribution. If the residuals are normally distributed, the blue dots should be close to the red line. In this case, the plot shows that the residuals are fairly close to normal, but deviate slightly at the tails.

Histogram: This shows the frequency of different residual values. A normal distribution has a bell-shaped curve, with most values near the mean and fewer values at the extremes. In this case, the histogram shows that most residuals are close to zero, but there are some outliers. The red dashed line indicates the mean of the residuals, which is also close to zero.