PROJECT STATUS REPORT

Project Name	Data set analysis	**Reporting Period**
Project Owner	Kaan Balcı	05/12/2024 - 05/19/2024
Teacher	Burak Evrentuğ	

HIGHLIGHTS

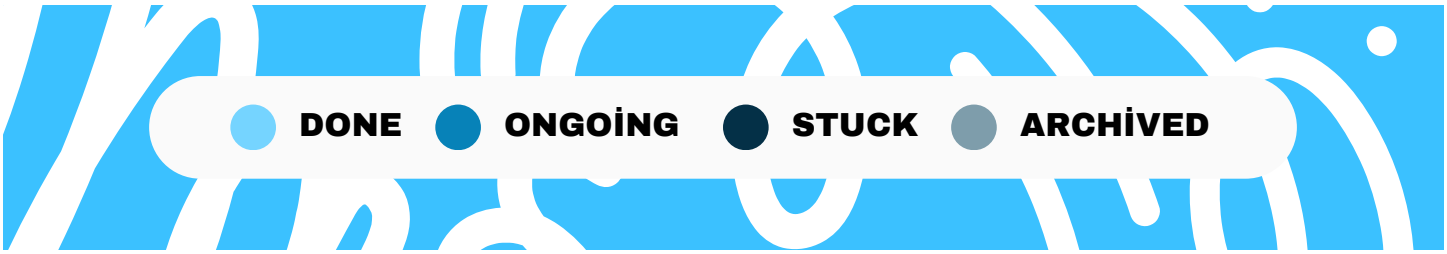
- The goal is to reinforce students' knowledge and skills in data science and machine learning by applying data preprocessing techniques and developing appropriate machine learning models on real-world data.
- Students will analyze, clean, process the data, and finally train machine learning models to derive meaningful insights from the data set they choose.

CHALLENGES

- Project Report: A detailed report explaining the research question, the chosen dataset, data preprocessing steps, the selected machine learning model(s), the model training process, the model evaluation methods, and the results obtained.
- Code Files: A Python notebook file containing data preprocessing, model training, and evaluation steps, enriched with explanations.
- YouTube Video: An approximately 5-minute video explaining the work done.
- Class Presentation: A 10-minute class presentation explaining the work done in a classroom setting.

STATUS UPDATES

Task or Deliverable	Task Owner	Status
Data Cleaning: Handling missing values by filling or removing them, and processing outliers.	KAAN BALCI	DONE
Data Transformation: Feature scaling, converting categorical data into numerical data.	KAAN BALCI	DONE
Data Visualization: Using various charts and visuals for data exploration.	KAAN BALCI	DONE
Feature Engineering: Creating and selecting new features to improve model performance.	KAAN BALCI	DONE



NEXT STEPS

Action Items

Task or Deliverable	Task Owner
Model Performance: Evaluation of model performance using metrics such as Accuracy, Precision, Recall, F1 Score for classification, and metrics like MAE, MSE, and RMSE for regression.	KAAN BALCI
Report and Presentation: Clarity and coherence of the project report, presenting analyses and results clearly.	KAAN BALCI
Code Quality: Readability of the code, presence of comments, and code organization.	KAAN BALCI

Dataset

The dataset contains information about various cars, including their specifications and selling prices.

The goal is to develop a machine learning model that can predict the selling price of a car based on its specifications. This predictive model will assist potential buyers and sellers in estimating the fair market value of a car.

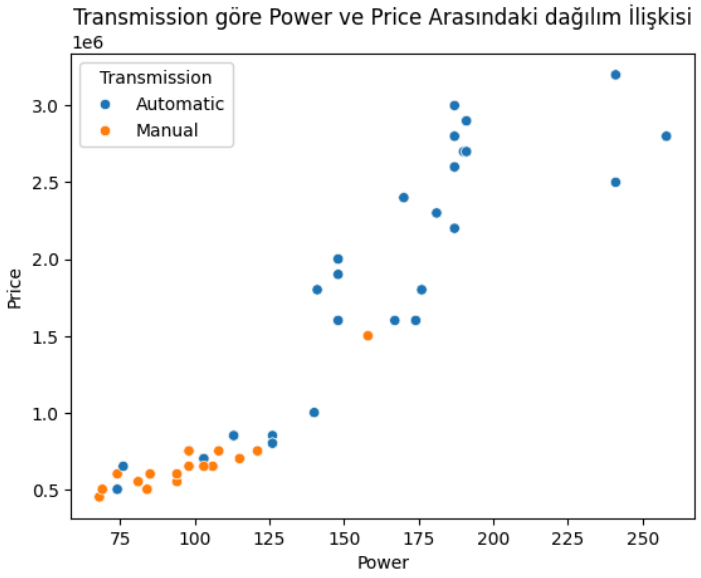
1. **Car_ID:** Unique identifier for each car listing.
2. **Brand:** The brand or manufacturer of the car (e.g., Toyota, Honda, Ford, etc.).
3. **Model:** The model of the car (e.g., Camry, Civic, Mustang, etc.).
4. **Year:** The manufacturing year of the car.
5. **Kilometers_Driven:** Total kilometers driven by the car.
6. **Fuel_Type:** The type of fuel used by the car (e.g., Petrol, Diesel, Electric, etc.).
7. **Transmission:** The type of transmission of the car (e.g., Manual, Automatic).
8. **Owner_Type:** The number of previous owners of the car (e.g., First, Second, Third).
9. **Mileage:** The fuel efficiency of the car in kilometers per liter.
10. **Engine:** The engine capacity of the car in CC (Cubic Centimeters).
11. **Power:** The maximum power output of the car in Brake Horsepower (BHP).

12. **Seats:** The number of seats in the car.
13. **Price:** The selling price of the car in INR (Indian Rupees), the target variable to be predicted.

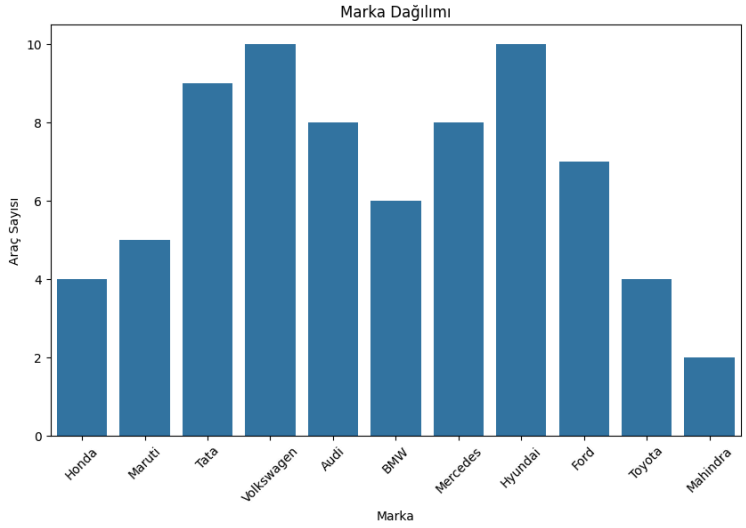
Dataset Analysis

- The dataset contains 100 rows.
- The average kilometers driven by the cars is 28,150.
- The average mileage of the cars is 17.21 kilometers per liter.
- The average engine capacity of the cars is 1855.23 CC.
- The average power output of the cars is 158.13 brake horsepower.
- The average number of seats in the cars is 5.23.
- The average selling price of the cars is 1,574,000 INR.
- Car_ID: A unique identifier ranging from 1 to 100.
- Year: The manufacturing years of the cars, ranging from 2016 to 2021.
- Kilometers_Driven: The total kilometers driven by the cars, ranging from 15,000 to 60,000, with an average of 28,150.
- Mileage: The fuel efficiency of the cars, ranging from 10 to 25, with an average of 17.21 kilometers per liter.
- Engine: The engine capacity of the cars, ranging from 1,197 to 4,951 CC, with an average of 1,855.23 CC.
- Power: The maximum power output of the cars, ranging from 74 to 396 brake horsepower, with an average of 158.13.
- Seats: The number of seats in the cars, ranging from 4 to 7, with an average of 5.23.
- Price: The selling prices of the cars, ranging from 550,000 to 4,000,000 INR, with an average of 1,574,000.
- There are no missing values.

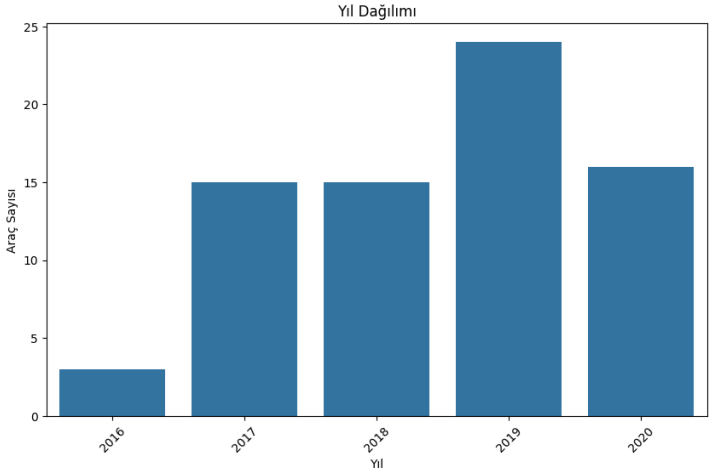
Dataset Visualization Analysis



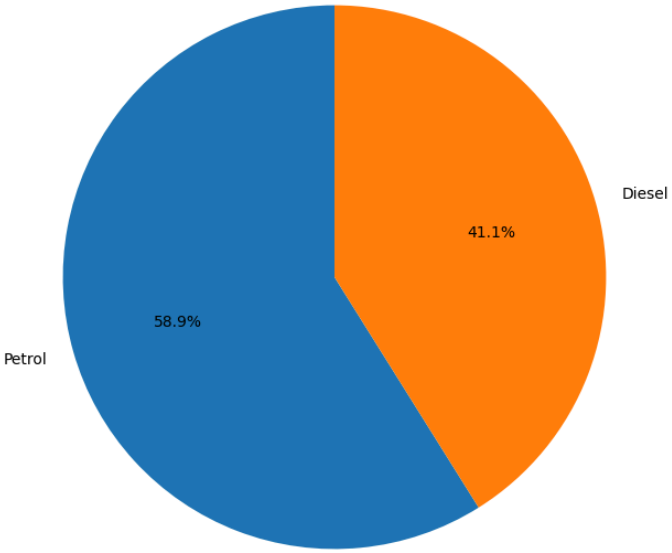
Based on the analysis, we can observe that there is a car with a power (Power) value of over 250 and a price (Price) value of over 3,000,000. Additionally, it seems that as the price of the car increases, the engine power also increases proportionally. Furthermore, there is a preference for automatic transmission (Transmission) type.



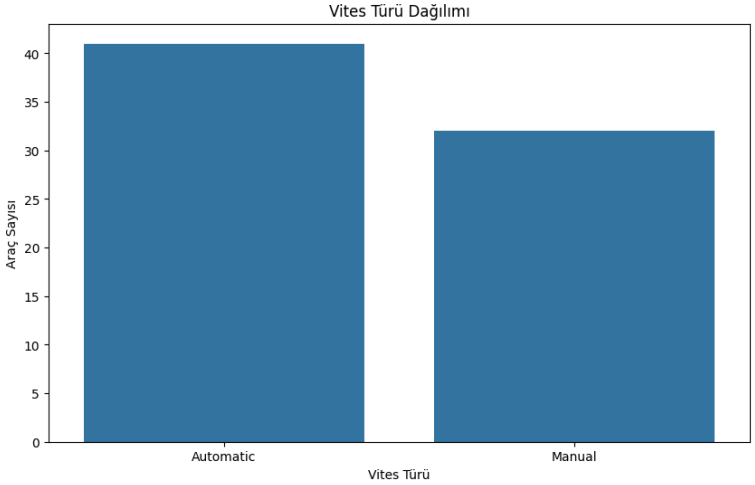
If we analyze the number of vehicles according to the brands in the dataset, we see that Hyundai and Volkswagen have the highest number of vehicles, while Mahindra has the lowest number of vehicles.



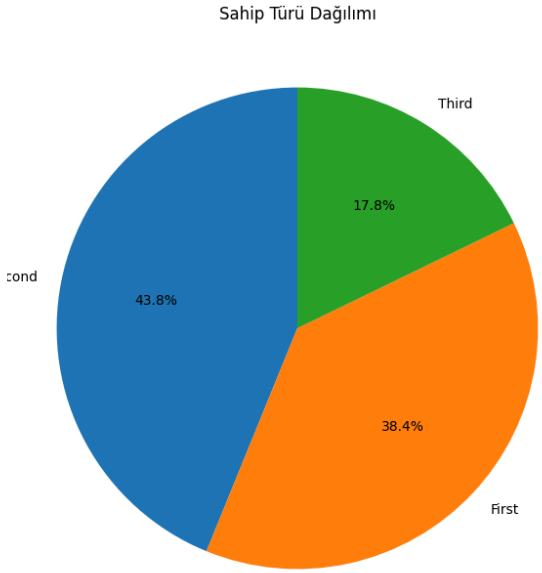
We compared the number of vehicles added to the dataset by years and analyzed that the highest number of vehicles was recorded in 2019 with a value of 20+, while the lowest number of vehicles was recorded in 2016.



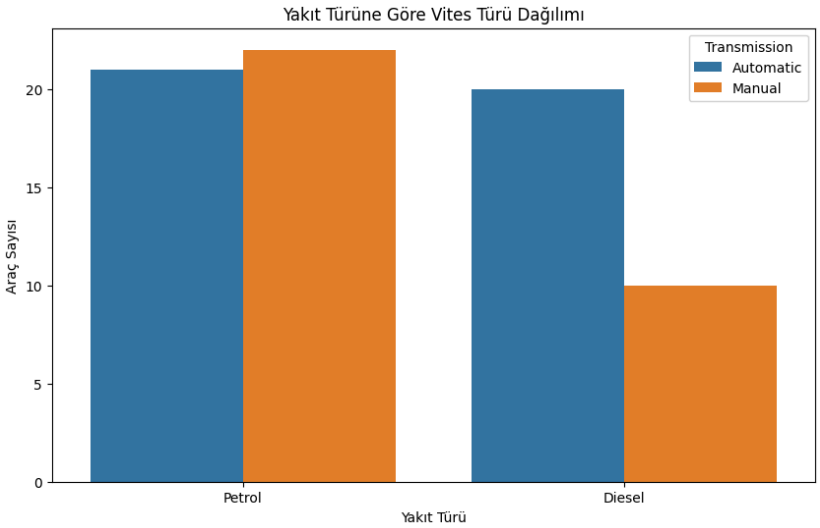
We compared the fuel types based on the number of vehicles in the dataset and found that the results are almost evenly distributed, with nearly equal numbers for each fuel type.



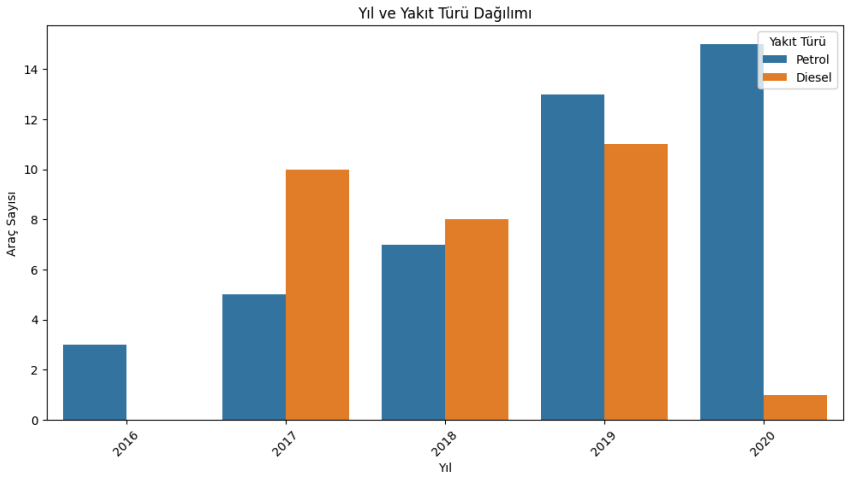
We examined the distribution of transmission types based on the number of vehicles, and as a result, we can analyze that there are more vehicles with automatic transmission.



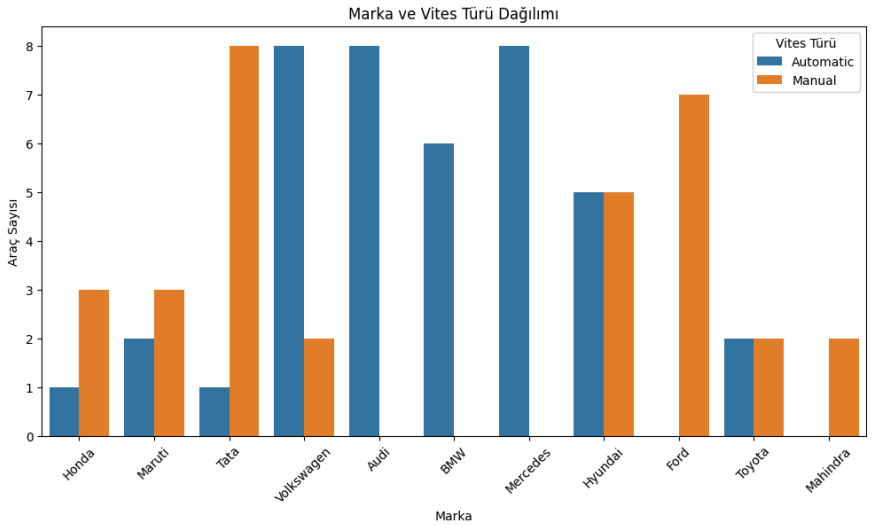
We analyzed the distribution of the number of vehicles based on their ownership type in a pie chart, and we can generally conclude that there are more used (2nd hand) vehicles in the dataset.



We compared the transmission types based on the fuel types, and we can analyze that vehicles with petrol fuel type mostly have manual transmission, while vehicles with diesel fuel type mostly have automatic transmission.

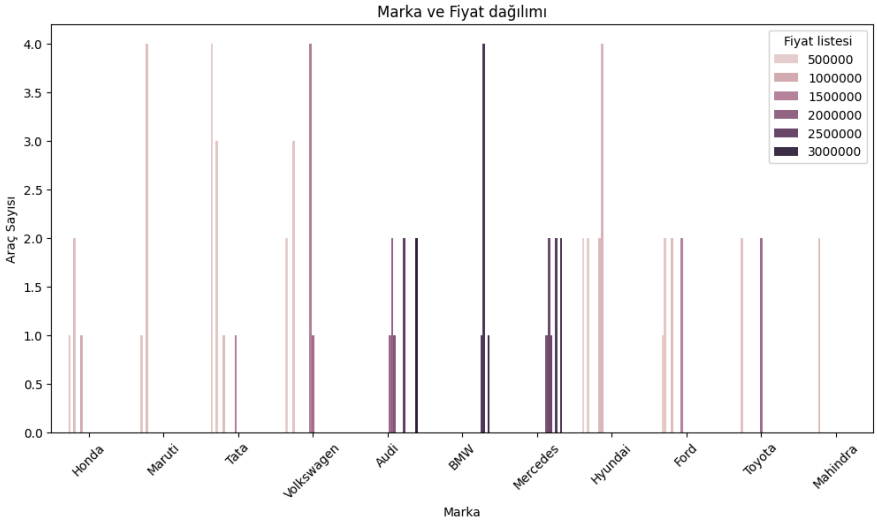


We compared the fuel types by year and analyzed that there were no diesel vehicles in 2016. In 2017, there was an increase in the number of diesel vehicles compared to petrol vehicles, but then there was a decrease. After a final increase, there was a complete return to petrol vehicles.



In this dataset, between 2016 and 2020, there are no vehicles with manual transmission in the Audi, BMW, and Mercedes brands, and no vehicles with automatic transmission in the Ford and Mahindra brands.

We observe that Tata has the highest number of vehicles with manual transmission, while Volkswagen, Audi, and Mercedes have the highest number of vehicles with automatic transmission.



In this analysis, we compared prices with brands and concluded that the most expensive vehicles are in the BMW model, while the cheapest vehicles are in the Mahindra and Maruti models.

Dataset Outlier Analysis

Car_ID

Q1: 27.75 - Q3 : 75.75 - IQR : 48.0 - low : -44.25 - high : 147.75

Year

Q1: 2018.0 - Q3 : 2019.25 - IQR : 1.25 - low : 2016.125 - high : 2021.125

Kilometers_Driven

Q1: 22000.0 - Q3 : 30500.0 - IQR : 8500.0 - low : 9250.0 - high : 43250.0

Mileage

Q1: 17.0 - Q3 : 20.0 - IQR : 3.0 - low : 12.5 - high : 24.5

Engine

Q1: 1199.0 - Q3 : 1968.0 - IQR : 769.0 - low : 45.5 - high : 3121.5

Power

Q1: 94.0 - Q3 : 175.75 - IQR : 81.75 - low : -28.625 - high : 298.375

Seats

Q1: 5.0 - Q3 : 5.0 - IQR : 0.0 - low : 5.0 - high : 5.0

Price

Q1: 600000.0 - Q3 : 2325000.0 - IQR : 1725000.0 - low : -1987500.0 - high : 4912500.0

Dataset Model Analysis

Mean Absolute Error: 276735.84383043397

Mean Squared Error: 110547144655.49069

Root Mean Squared Error: 332486.3074706847

Prediction Results

ACTUAL - PREDICTED

2200000 - 2367836.0

550000 - 386853.4

3200000 - 2981689.0

1000000 - 1503816.0

1500000 - 1828165.0

600000 - 586712.1

1800000 - 1791424.0

1800000 - 1571984.0

3000000 - 2391514.0

650000 - 889171.3

700000 - 1054986.0

700000 - 990339.5

650000 - 623728.7

1600000 - 2138442.0

2700000 - 2237813.0

Lastly, we can convey that our model is a good model because we are trying to predict the price variable, so initially, the model's error values may appear to be very high. However, it makes very good predictions with very small deviations.