

**Cars Data Analysis Report**

**Introduction**

This report presents an analysis of a cars data with the objective of understanding the factors that influence car prices and building a regression model to predict car prices based on various features. The analysis encompasses data exploration, cleaning, feature engineering, visualization, and modeling.

**Programs and Libraries**

I used Colab website to implement this project and used Python to write the codes to analyze this data. I also entered all the libraries that I would need for statistical and mathematical operations.

Below we will discuss the steps we took in this project to serve as a reference for the code steps If you need to code you found it by click on the [link](https://colab.research.google.com/drive/1-4cgcJ0R3yYyIGnvCcJTgiL9k9os__-W#scrollTo=YJUl-ZZG8iuD) .

**Data Exploration**

At this step I just explored the data to see the general shape of the data such as field types, data size, which fields have null values ​​and how many duplicate rows.

**Data Cleaning**

After exploring the data and identifying the errors that should be corrected before starting the analysis process, I handling missing values containing empty values ​​totaled no more than 70 rows, and this did not affect the total of 4960 rows, so I deleted them and replaced the illogical values ​​(other) with the most common values(mod) ​​in the transmission and fuelType columns .

**Feature Engineering**

To simplify the dataset, the brand, model, and year columns were merged into a Car Model column, and the column (carID)were dropped .

**Visualization**

Visualization is the graphical representation of data or information, making it easier to understand patterns and insights. It helps in simplifying complex data for better interpretation and communication we take examples of data car visualization .

1. A bar chart was created to show the count of the top 30 car models based on the number of listings. This helped us identify the most popular car models in the dataset.

A graph of numbers and a number of cars

Description automatically generated with medium confidence

1. Transmission and Fuel Type Distribution

Count plots were used to visualize the distribution of cars by transmission type and fuel type we notice that the Manual transmission was the most common type, followed by automatic transmission , also the Diesel and petrol were the most popular fuel types, with electric and hybrid cars having fewer entries .

A graph of cars with different colored squares

Description automatically generated

1. scatter plot visualizing the relationship between a car's mileage and its price ,This scatter plot shows the relationship between car mileage and price, with different marker sizes representing price ranges and shapes indicating transmission types. As expected, cars with lower mileage tend to have higher prices. The highest-priced car in the dataset is priced at **145,000**, and the lowest Price car in the dataset is priced at **450** , finally the most high-priced cars are found at low mileage levels.

A graph showing a number of miles per mileage

Description automatically generated

1. The line plot illustrates a positive relationship between tax and price, indicating that higher tax levels are associated with increased vehicle prices. Individual data points reveal variability, and any significant outliers may impact the model's predictions. This relationship provides valuable insights for stakeholders regarding the influence of tax policies on vehicle pricing strategies.

**A graph with red lines and numbers

Description automatically generated**

**Build Model**

A model is a simplified representation of a real-world process or system, designed to analyze relationships based on input data the first step of this data encoding find in below

1. ***Data Encoding***

In this step, we encoding the string columns to numeric that they are understandable for the model

1. ***Scalling***

After we encoding the column make all fields in close numeric range by sklearn preprocessing in StandardScaler

1. ***feature selection***

Feature selection is the process of identifying and selecting a subset of variables from a larger set to improve the performance of a predictive model

**Regression Model**

A regression model is a statistical technique used to predict a continuous outcome based on one or more predictor variables It establishes relationships through equations, with linear regression being the simplest form.

The model is trained using the fit() method, which takes the features (X\_train )and the target (y\_train) as inputs.

After make this step we calculate the error y\_test predict by mean absolute error and found the MAE = 0.4206612961930928 ,This is considered a good value that should be in the range(0-1).

calculate the y-test predict by R-squared r2= 0.6093284113555781 the value is also good .

**Conclusion**

In summary Popular car models, especially economy cars, tend to be priced lower, while luxury models with larger engines are priced higher. And Cars with **manual transmission** and **petrol/diesel fuel types** dominate the dataset. also The linear regression model, while useful, may need further tuning or the use of more advanced models to improve predictive accuracy.

This analysis provides insights into key factors affecting car prices, and further analysis can focus on optimizing the predictive models for better performance.

The linear regression model was fitted to the training data and subsequently used to predict outcomes for the test set. The Mean Absolute Error (MAE) of approximately **0.42** indicates the average magnitude of the errors in predictions, suggesting that, on average, the model's predictions deviate from the actual values by about 0.42 units.

Additionally, the R-squared value of approximately **0.61** implies that the model explains about 61% of the variance in the target variable. While this indicates a moderate level of predictive power, there is still a considerable amount of unexplained variance, suggesting potential areas for model improvement, such as incorporating additional features or exploring different modeling techniques.

Overall, while the model shows reasonable performance, further refinement and evaluation might enhance its predictive capabilities.