Navigating the Used Car Market: Predictive Price Analytics

Keerthana J

Mahesh D

Monali Jadhav

Ajay Kumar Lakkakula

Harshith Krishna

Abstract:

This study aims to help buyers and sellers in the used car market by predicting accurate pricing for vehicles. It analyses various factors that influence the price of a used car, such as its make, model, age, mileage, and condition. By using advanced data analysis techniques, it develops a model that can estimate the fair market value of a specific used car. This model can assist buyers in negotiating better deals and help sellers price their vehicles competitively. The goal is to bring more transparency and efficiency to the used car market by providing data-driven pricing insights.

Prototype Selection:

a. Feasibility: The development of a predictive pricing model for used cars is feasible in the short-term future (2-3 years) due to several factors:

1. Data Availability: There is a vast amount of historical data available on used car sales, including information on make, model, year, mileage, condition, and sale prices. This data can be collected from various sources, such as online marketplaces, dealerships, and automotive databases.
2. Technological Advancements: Machine learning and predictive analytics techniques have made significant strides in recent years. Algorithms like random forests, gradient boosting, and neural networks can effectively handle complex data and identify patterns, making them well-suited for developing accurate pricing models.
3. Computational Power: The increasing availability of cloud computing resources and powerful hardware has made it easier to process and analyse large datasets required for training predictive models.

b. Viability: The used car market is a significant and ever-growing industry, with millions of transactions occurring annually. The demand for used cars is driven by several factors, such as:

1. Cost Savings: Purchasing a used car is often more affordable than buying a new one, making it an attractive option for many consumers.
2. Environmental Concerns: With increasing awareness of environmental issues, some consumers prefer to extend the lifespan of existing vehicles rather than contributing to the production of new ones.
3. Personal Preferences: Some buyers prefer the unique character and history of a used car over a brand-new model.

As these factors are unlikely to change significantly in the long-term future (20-30 years), a predictive pricing tool for used cars will remain relevant and viable.

c. Monetization: This product can be monetized directly through various channels:

1. Subscription-based Service: Users (buyers and sellers) can subscribe to the pricing model service, paying a monthly or annual fee for access to up-to-date pricing insights.
2. One-time Fee: Alternatively, users can pay a one-time fee for each pricing evaluation they require.

3. Integration with Existing Platforms: The predictive pricing model can be integrated into existing online marketplaces or dealership platforms. In this case, a commission or licensing fee can be charged for the pricing insights provided to users.
4. Advertising Revenue: The pricing tool can be offered as a free service to users, generating revenue through targeted advertising based on user search and pricing queries.

By addressing the criteria of feasibility, viability, and direct monetization potential, the "Navigating the Used Car Market: Predictive Price Analytics" prototype presents a promising opportunity for further development and implementation in the used car industry.

Prototype Development:

1.Data Collection and Preprocessing

1. Describe the sources of data used for building the predictive pricing model, such as online marketplaces, dealerships, automotive databases, or any publicly available datasets.
2. Explain the process of data cleaning and preprocessing, including handling missing values, removing outliers, and feature engineering.
3. Mention the specific features or variables collected for the model, such as make, model, year, mileage, condition, location, and historical sale prices.

2.Exploratory Data Analysis (EDA)

1. Provide insights gained from EDA, such as identifying patterns, trends, and correlations between different features and the target variable (sale price).
2. Include visualizations (e.g., scatter plots, histograms, box plots) to better understand the data distribution and relationships between variables.
3. Discuss any data transformations or encoding techniques applied based on the EDA findings.

3.Model Selection and Training

1. Explain the rationale behind the choice of specific machine learning algorithms or predictive modelling techniques used, such as linear regression, decision trees, random forests, gradient boosting, or neural networks.
2. Describe the model training process, including techniques like cross-validation, hyperparameter tuning, and any regularization methods employed to prevent overfitting.
3. Mention the evaluation metrics used to assess model performance, such as mean squared error (MSE), root mean squared error (RMSE), or R-squared.

4.Model Performance and Validation

1. Present the results of the trained model's performance on the test or validation dataset, including the evaluation metric scores.
2. Discuss the strengths and limitations of the model, as well as potential areas for improvement.

3. Validate the model's predictions on a subset of real-world examples or hold-out data, demonstrating its ability to provide accurate pricing estimates.

5.Prototype Implementation (Optional)

1. If a basic app or website prototype is developed, describe the user interface and functionality, including how users can input vehicle details and obtain pricing estimates.
2. Explain the integration of the trained predictive model into the prototype, ensuring seamless interaction between the front-end and back-end components.
3. Discuss any additional features or enhancements planned for the prototype, such as personalized recommendations, comparison tools, or integration with external data sources.

6.Future Enhancements and Scalability

1. Outline potential enhancements to the predictive pricing model, such as incorporating additional features, exploring ensemble methods, or leveraging more advanced deep learning techniques.
2. Discuss strategies for keeping the model up-to-date and adapting to changes in the used car market, such as periodic retraining or online learning techniques.
3. Address scalability considerations, including how the model and infrastructure can handle an increasing volume of pricing queries or data as the userbase grows.

7.**Conclusion and Next Steps**

1. Summarize the key findings and achievements of the small-scale implementation or model building phase.
2. Highlight the potential impact and benefits of the "Navigating the Used Car Market: Predictive Price Analytics" prototype for buyers, sellers, and the overall used car industry.
3. Outline the next steps for further development, testing, and deployment of the prototype into a production-ready solution.

Business Opportunity:

1.**Addressing Market Needs**

- The used car market is vast, with millions of transactions occurring annually, highlighting a significant demand for accurate pricing information.
- Consumers, including individual buyers and sellers, as well as dealerships, often struggle to determine the fair market value of a used vehicle, leading to potential overpayment or undervaluation.
- A predictive pricing solution can bridge this gap by providing data-driven insights, enabling more informed decision-making and negotiation processes.

2.**Competitive Differentiation**

- Companies that offer advanced predictive pricing systems can gain a competitive edge over traditional pricing methods or manual evaluation processes.

- By leveraging machine learning algorithms and comprehensive data analysis, these companies can deliver more accurate and reliable pricing estimates, creating a value-added service for their customers.
- This differentiation can attract a larger customer base, both from individual consumers and automotive businesses seeking to streamline their pricing strategies.

3.**Revenue Streams and Monetization**

- Predictive pricing solutions can generate revenue through various models, such as subscription-based access, pay-per-use pricing, or one-time fees for individual pricing evaluations.
- Targeted advertising based on user search and pricing queries can also be explored as a revenue stream, offering customized promotions or recommendations related to the used car market.
- Additionally, companies can monetize their data by providing insights and analytics services to third-party stakeholders, such as insurance companies, financial institutions, and automotive manufacturers.

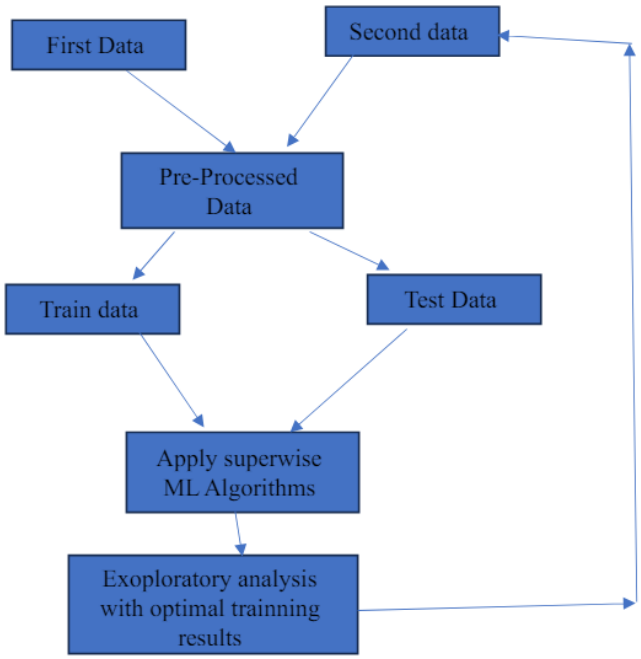
4.**Market Expansion and Diversification**

- The predictive pricing solution can be tailored and expanded to target new geographic regions or specific market segments, such as luxury or vintage vehicles.
- By adapting the system to local market dynamics, consumer preferences, and regulatory environments, companies can unlock new growth opportunities and diversify their customer base.
- Continuously refining and enhancing the pricing model can also open doors to new applications or adjacent markets, fostering long-term sustainability and growth potential.

5.**Strategic Partnerships and Collaborations**

- Collaborating with industry stakeholders, such as automotive dealerships, online marketplaces, and data providers, can significantly enhance the reach and effectiveness of the predictive pricing solution.
- Partnerships can provide access to larger datasets, facilitate seamless integration with existing platforms, and leverage complementary expertise and resources.
- By forming strategic alliances, companies can accelerate product development, gain market insights, and create innovative solutions that cater to the evolving needs of the used car market.

Flow diagram of the proposed methodology:



Data Preparation:

Dataset we have used related to used cars all about brand, model, model year, mileage, fuel type, engine, transmission, interior colour, exterior colour, accident, clean title, price.

1.Loading Data:

Import the dataset into the analysis environment (e.g., Python, R).

2.Data Cleaning:

Handle missing values, correct data types, and remove duplicates.

3.Data Transformation:

Create new variables if necessary.

Descriptive Statistics:

1.Summary Statistics:

Generate summary statistics for numerical variables (mean, median, standard deviation, etc.).

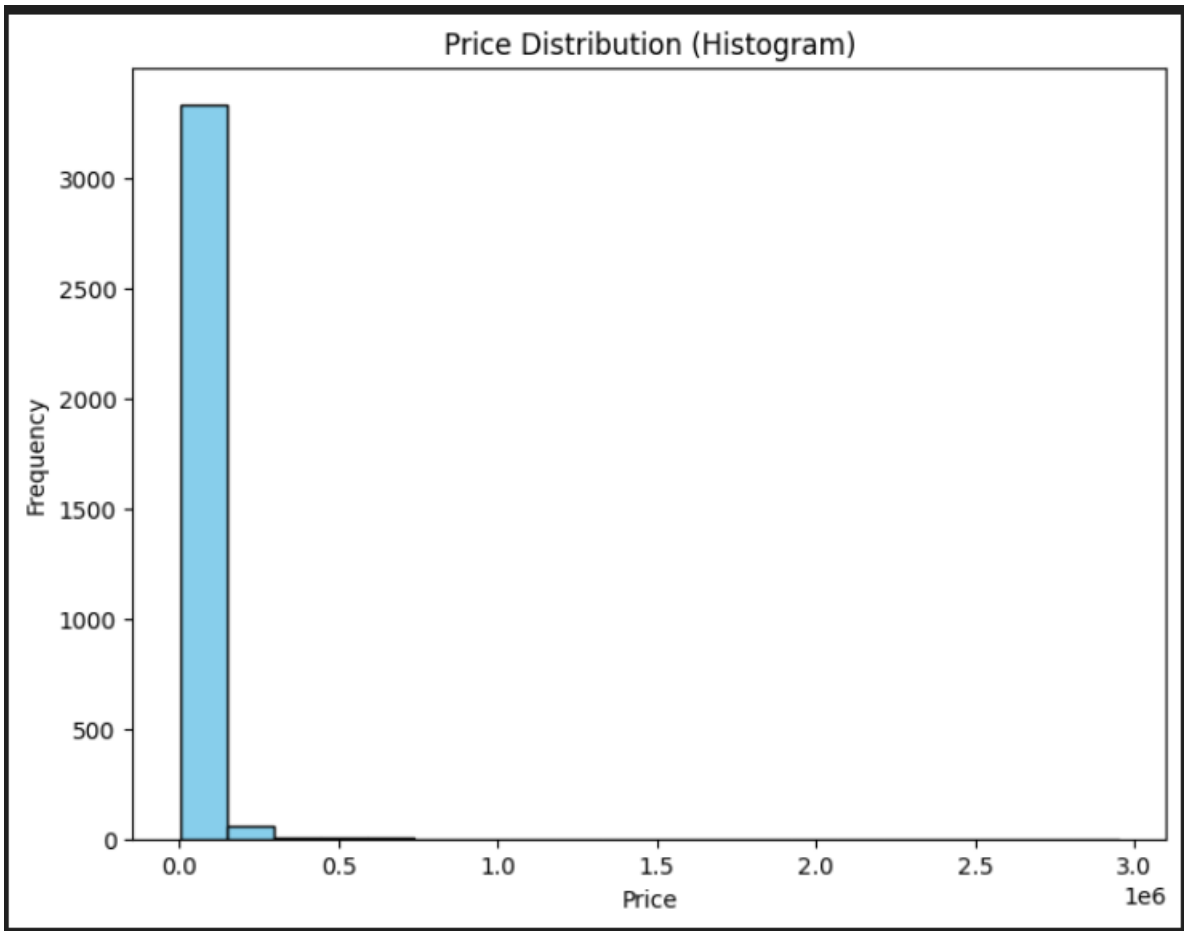
2.Distribution Analysis:

Visualize the distribution of key numerical variables using histograms or density plots.

3.Categorical Analysis:

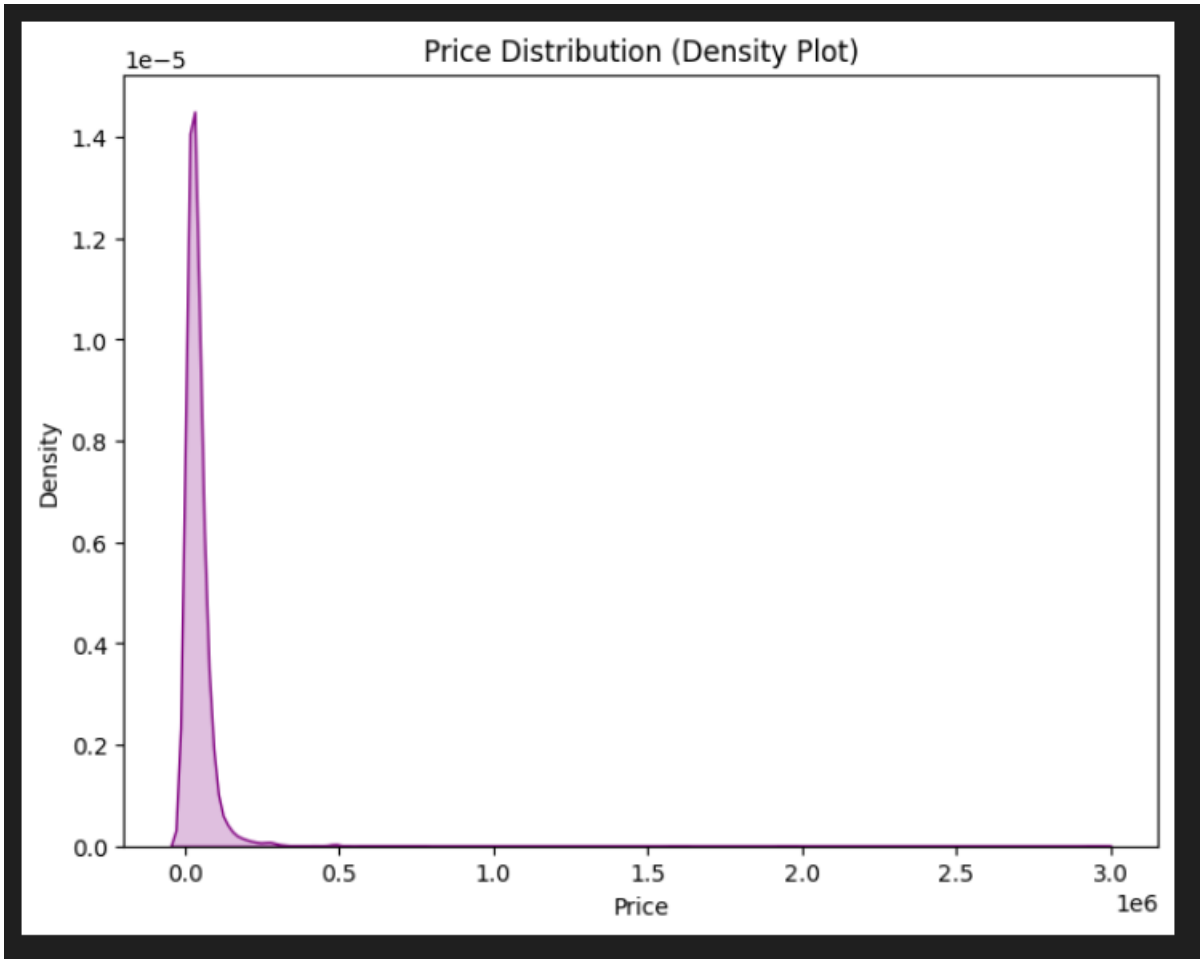
Summarize and visualize the distribution of categorical variables using bar charts or pie charts.

Explortary Data Analysis:



Above Histogram shows that most cars lie in the lower range of price.

The data has too many outliers as seen from the above graph and hence needs to be removed - to get near normal distribution.



Above plot - Density Distribution plot for price.

Again, it shows are very right skewed distribution and a very narrow line that extends to the right corner of the graph.

A strip plot for the 'price' variable

A great way to visualize the spread of numerical data.

The above plot shows how far the values are from the mean values - too many outliers.

OUTLIERS REMOVAL METHOD - IQR METHOD:

The above graph has outliers from 0.4 onwards, which need to be removed.

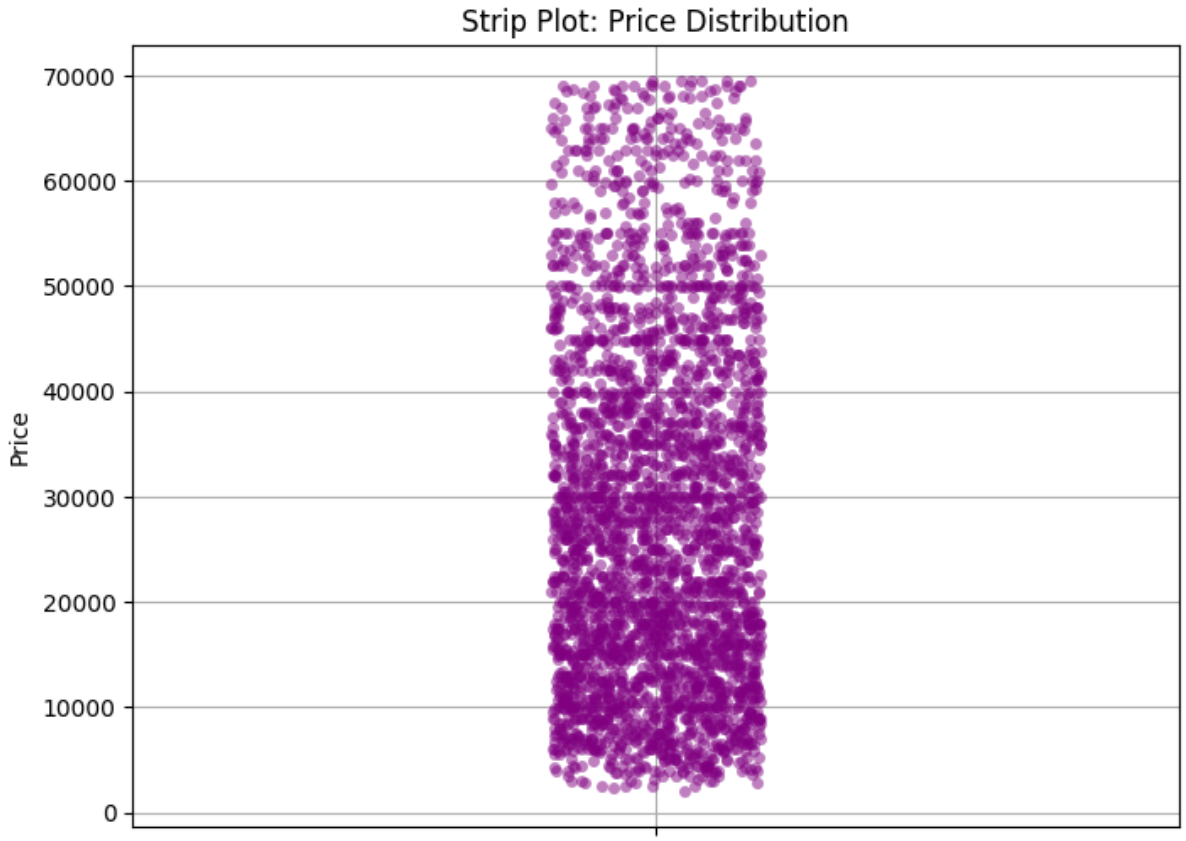
```
#identify and remove outliers from the data df in relation to price column using IQR method
Q1 = df['price'].quantile(0.25)
Q3 = df['price'].quantile(0.75)
IQR = Q3 - Q1
print(IQR)
```

```
31600.0
```

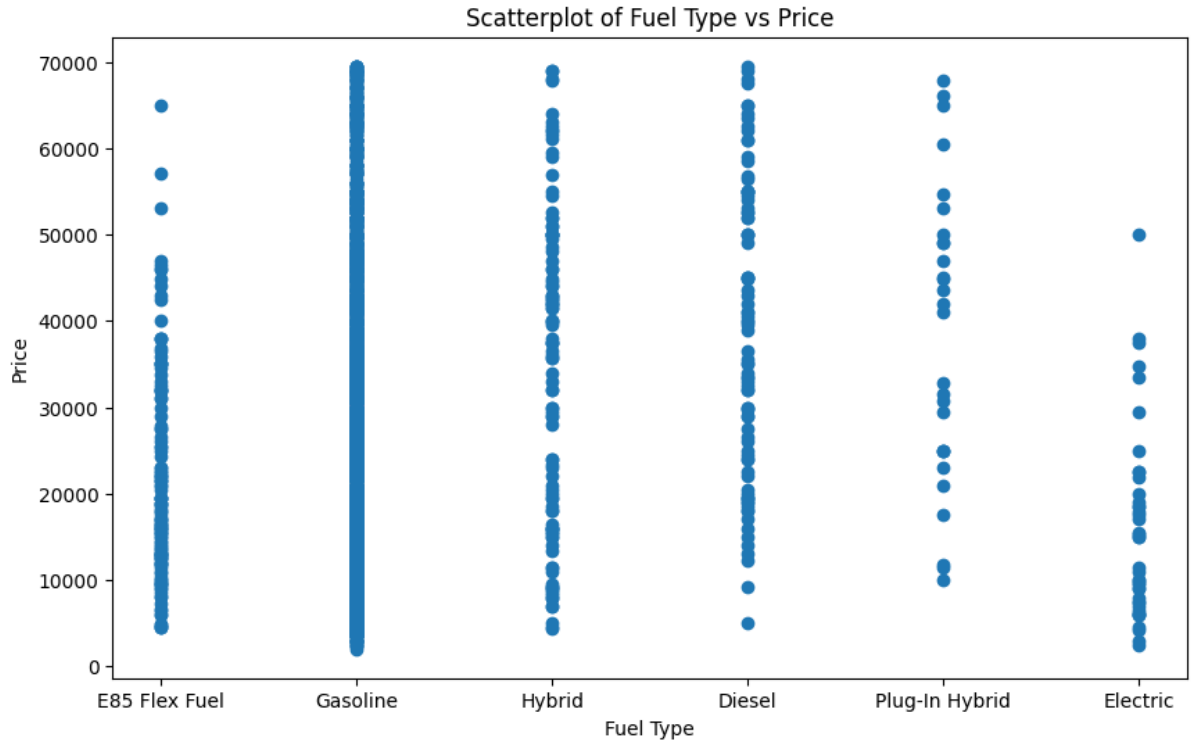
```
#use IQR value to filter out the outliers
df = df[~((df['price'] < (Q1 - 1.5 * IQR)) |(df['price'] > (Q3 + 0.7 * IQR)))]
df.shape
```

```
(3012, 13)
```

Strip plot after removing outliers.



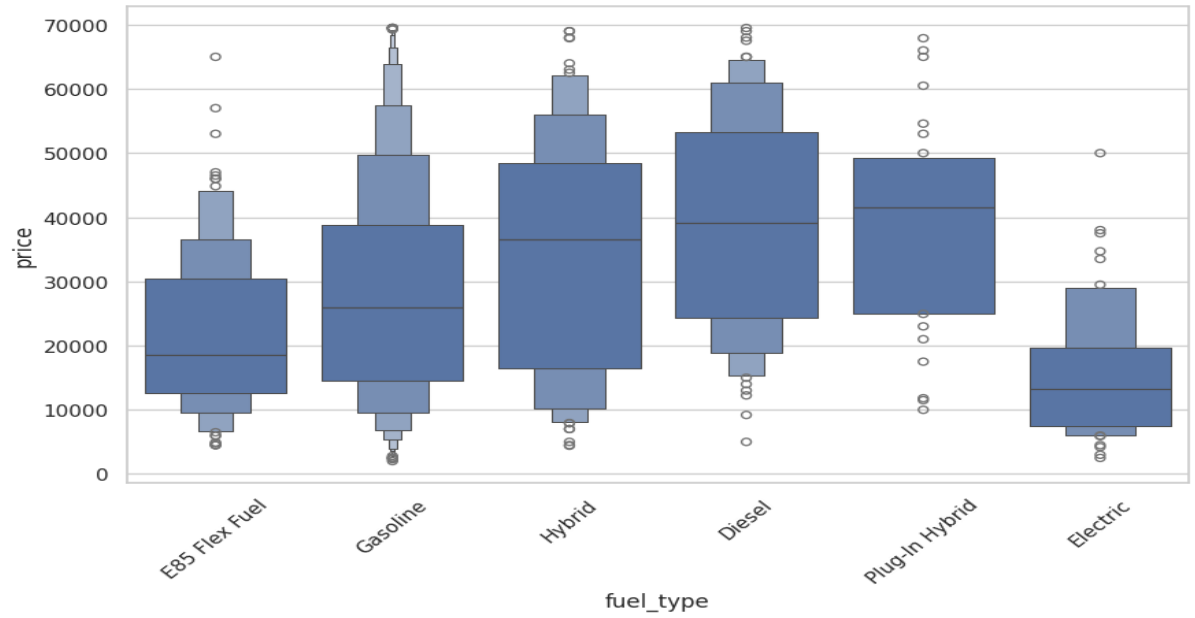
visualize fuel type column in regrads to price.



The above scatter plot shows that - Gasoline has more consistent values and have a larger range of car prices.

A near similar pattern can be seen in Hybrid type.

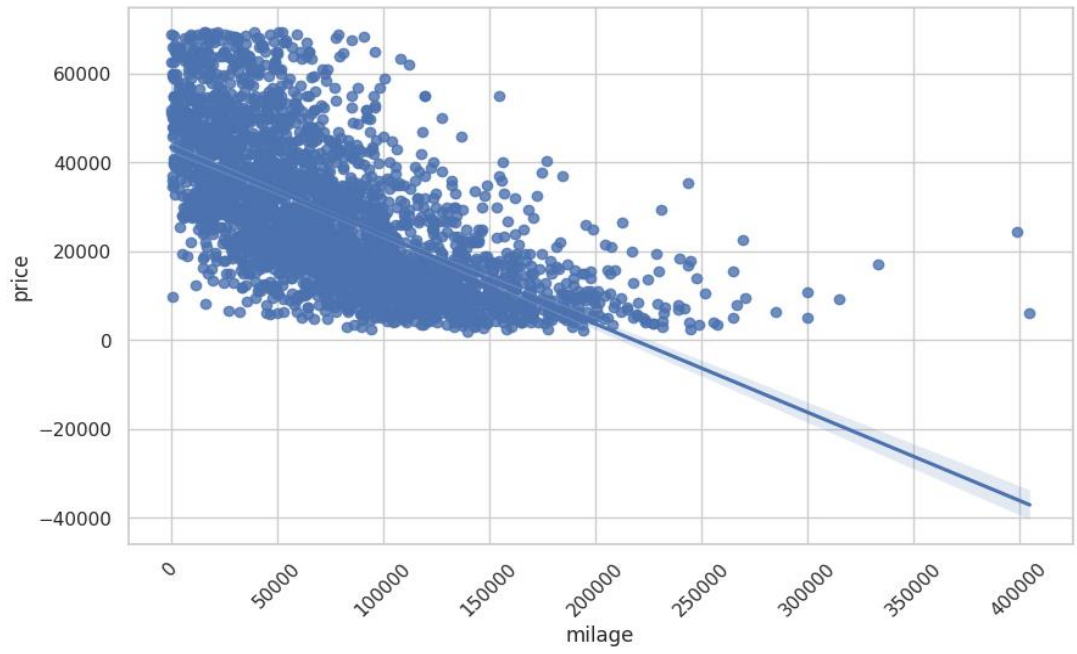
A boxplot of fuel type vs price



The above Boxplot shows that the mean price for Plug-In Hybrid is the highest among all. Meanwhile, Gasoline has a larger range.

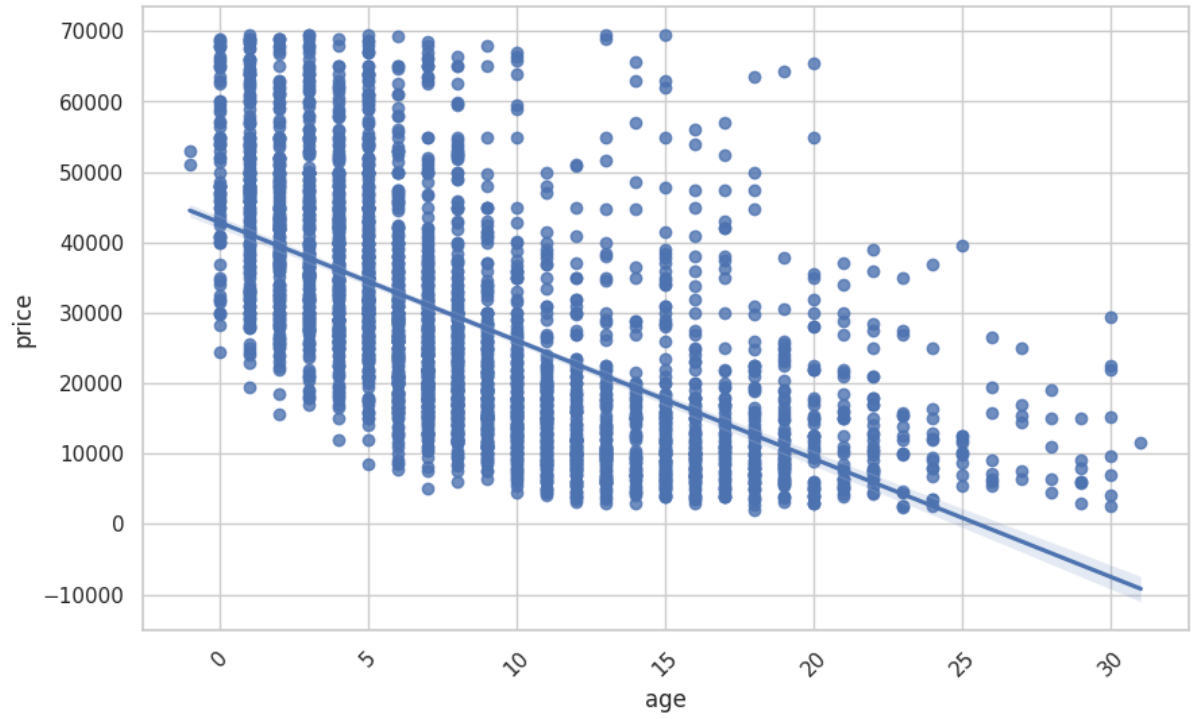
Scatterplot of clean title vs price

regression plot of milage vs price



The regression plot above shows that - the more miles the car has driven - the lesser the selling price.

regression plot of age vs price



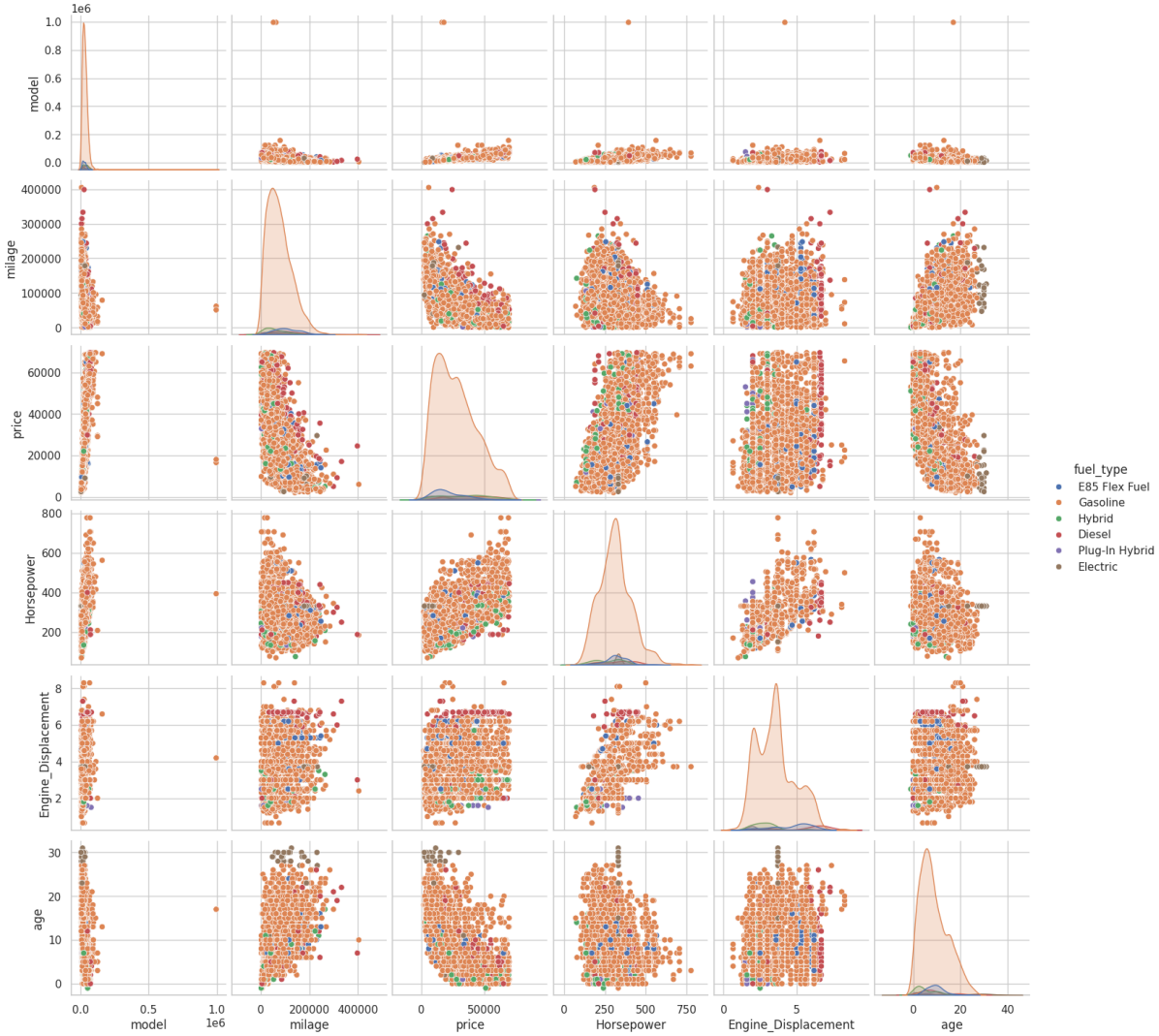
Regression Plot - 'age' vs 'price'

The older the car - the lesser the selling price.

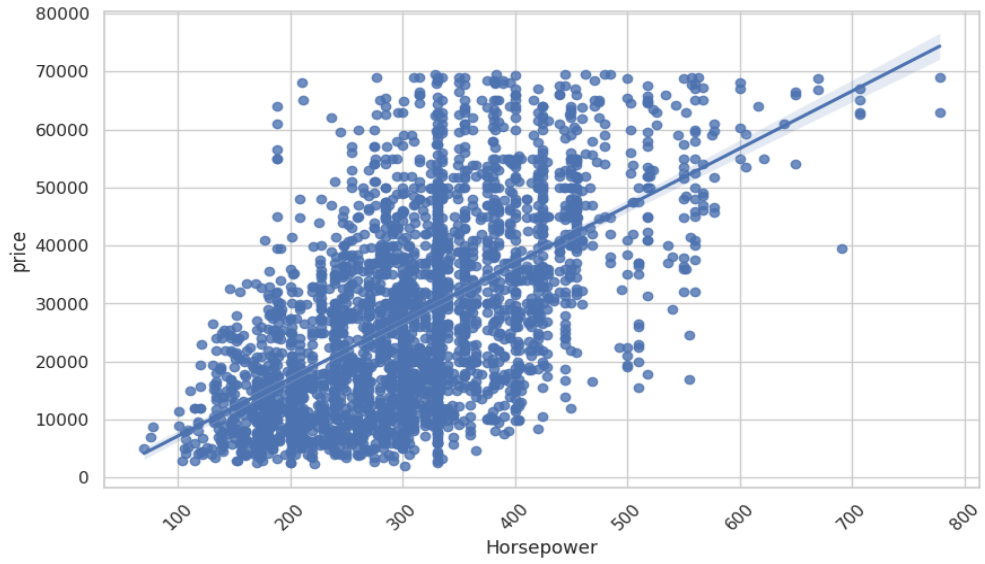
A pair plot of all the features in df except brand, transmission

The pair plots below show the dominance of Gasoline fuel type cars in relation to car price, engine displacement, age, Horsepower & milage.

It means that if one wants to sell a car - he would sell it at higher price if it were Gasoline fuel type, lower age, lower milage, higher engine displacement.

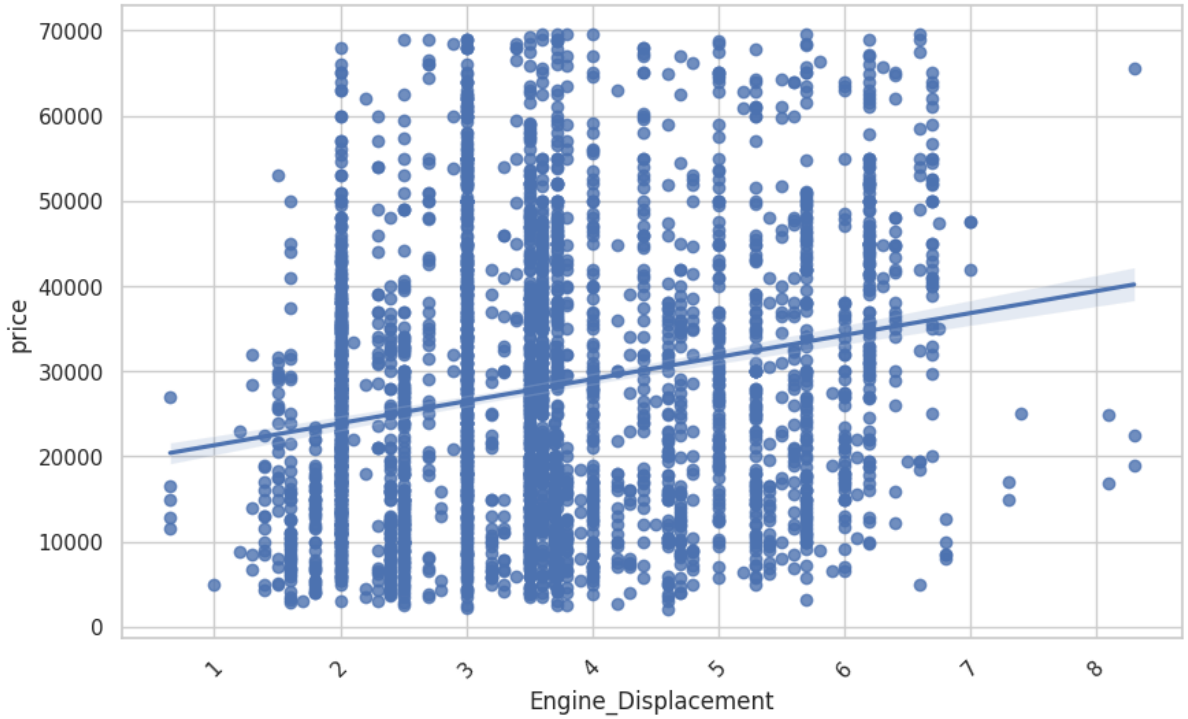


A regression plot of Horsepower vs price



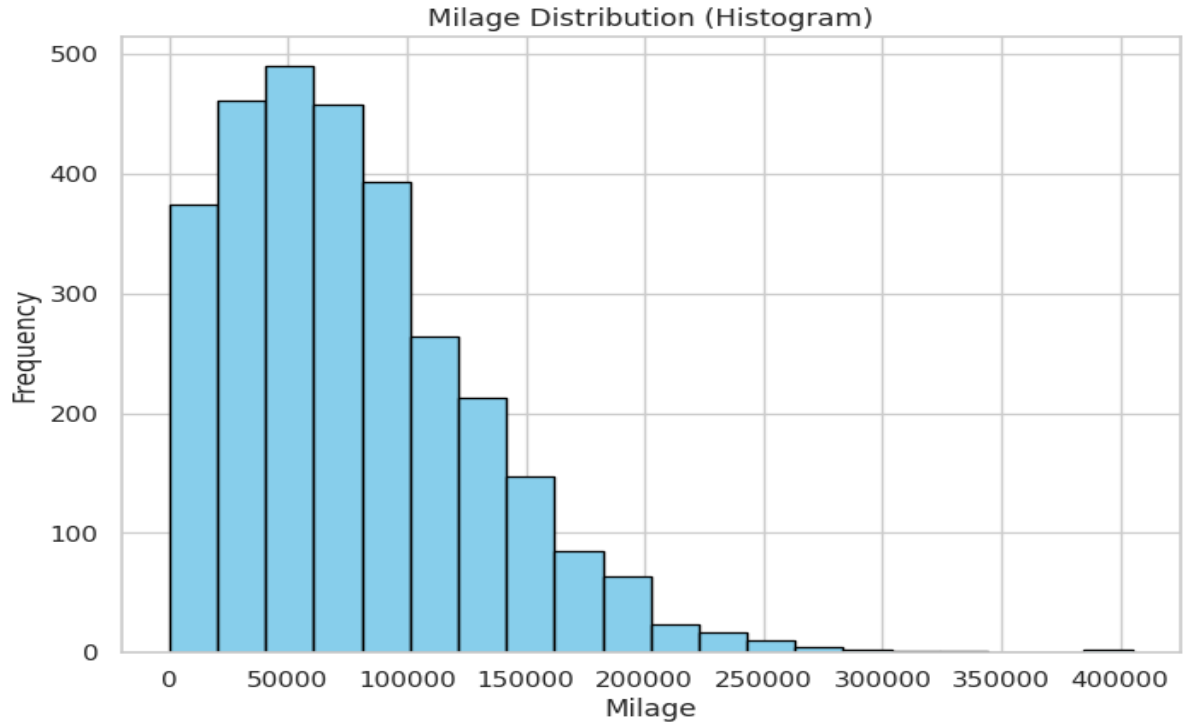
The price increases with increase in engine Horsepower.

A scatter plot of Engine Displacement vs price



The larger the engine size the higher the price.

A histogram for milage column



Majority of the cars have lesser milage.

GitHub Links for the Model and Market Segmentation

https://github.com/keerthanaj2004/Feynn-Lab-Internship/tree/main/Task3