Project 4: Used Car Pricing

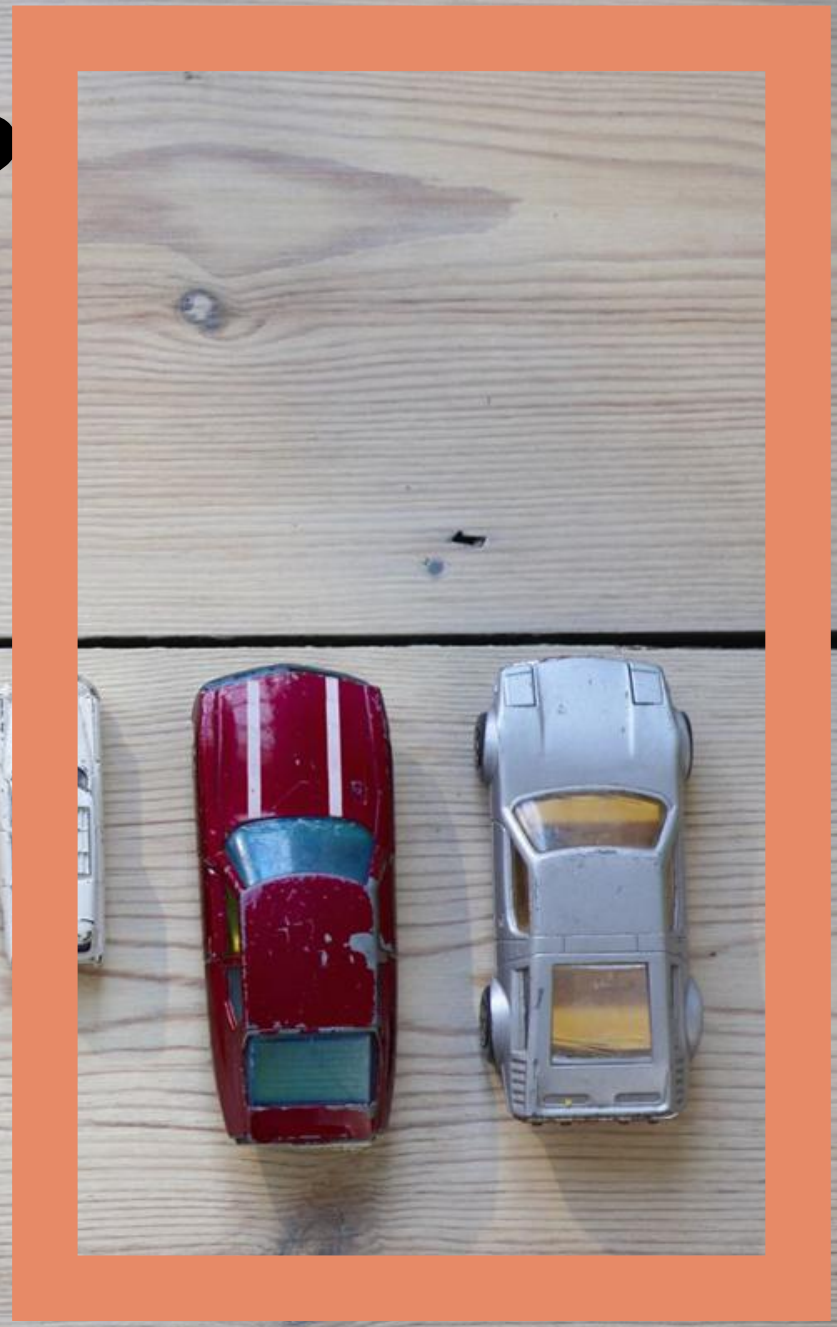
By Group 4: Gina Butler, Chase Mueller, Khatri Kriti, Thaissa Champagne, Neelam Prasad

USED CAR PRICE ANALYSIS

We used the dataset used_cars.csv from a Used Car Price Prediction analysis from Kaggle.com(https://www.kaggle.com/code/satyaprakashshukl/used-car-price-prediction/input?select=extended_data.csv, accessed: April 2025).

WHY THIS PROJECT?

Used car prices vary due to multiple factors. Our goal is to build a model that accurately predicts car prices based on relevant features.



Helps buyers and sellers estimate fair prices, reducing uncertainty in pricing negotiations.

Data Description

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4009 entries, 0 to 4008
Data columns (total 12 columns):
 #   Column        Non-Null Count  Dtype 
---  ------        --------------  ----- 
 0   brand         4009 non-null   object
 1   model         4009 non-null   object
 2   model_year    4009 non-null   int64 
 3   milage        4009 non-null   object
 4   fuel_type     3839 non-null   object
 5   engine        4009 non-null   object
 6   transmission  4009 non-null   object
 7   ext_col       4009 non-null   object
 8   int_col       4009 non-null   object
 9   accident      3896 non-null   object
 10  clean_title   3413 non-null   object
 11  price         4009 non-null   object
dtypes: int64(1), object(11)
memory usage: 376.0+ KB
```

Questions we tried to answer

What factors most significantly affect the price of a used car?

Can we predict the price of a used car based on its attributes?

How does the model year of a car affect its price?

Do accident history and clean title status impact the price of a used car?

Which fuel type (gasoline, diesel, electric, hybrid) has the highest average price?

How does engine type influence the price of used cars? ○

Methods and Approaches

- Data cleaning using python
- Tableau for analytical visualization
- Machine learning for price prediction
- Flask app for webpage design

Data Cleaning

Removal of unuseful columns for ML

- Interior Color
- Exterior Color
- Model
- Engine

Re-grouping data for Tabelau analysis

- Price
- Car age
- Model/Brand

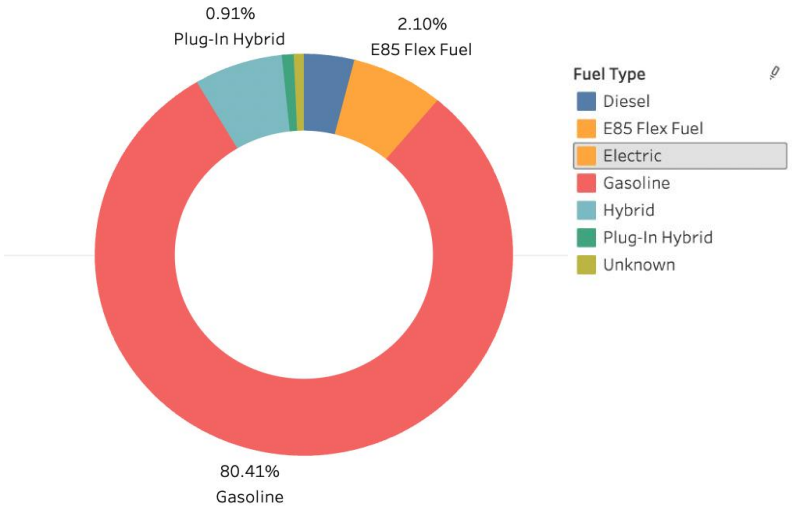
Flask app. Design

We created html pages and used Tableau to create visualizations of our data that answered our questions concerning used car price predictions by fields fuel type, model, mileage, and price.

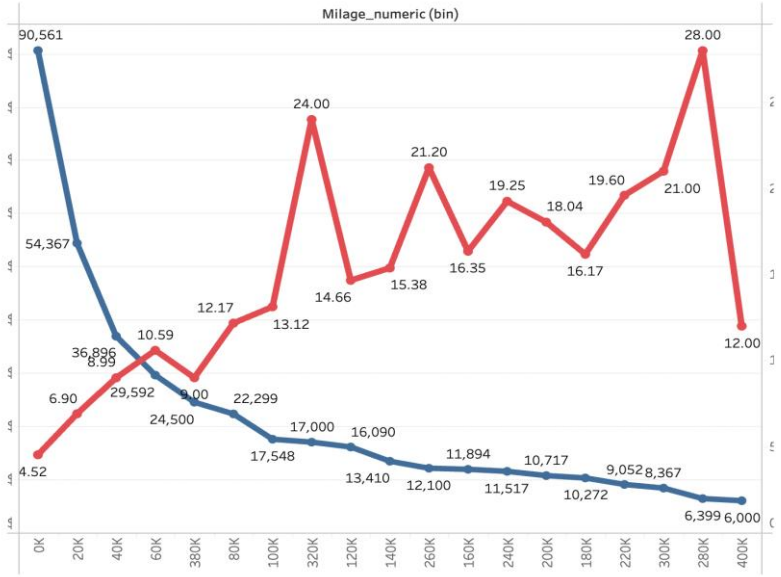
Visualizations created were bar, line, lollipop, pie and bubble charts.

Tableau for visualization

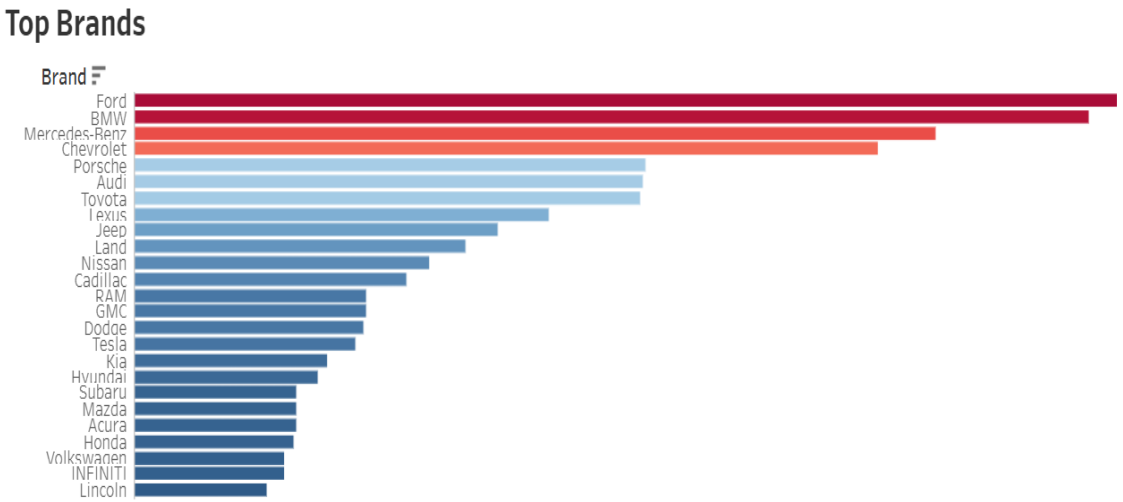
Cars by fuel type



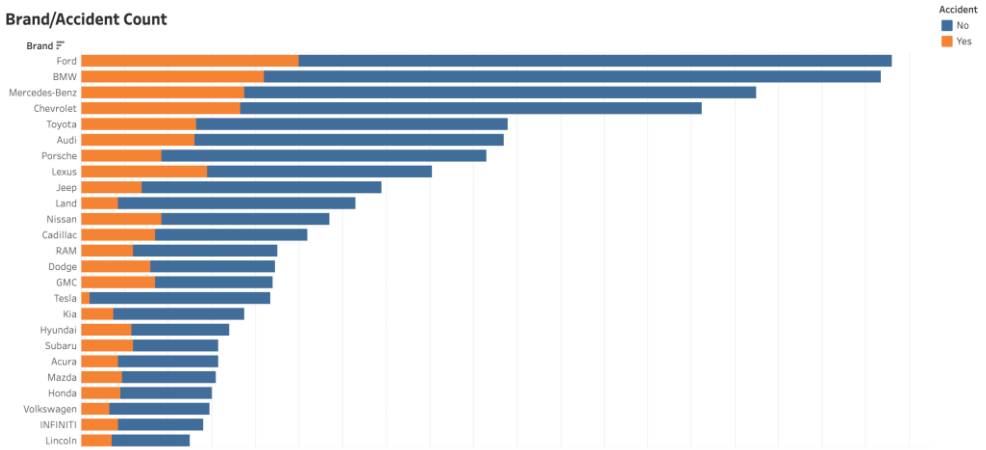
Used cars price based on mileage and car age



Used cars by brand



Used cars price based on titles



MACHINE LEARNING

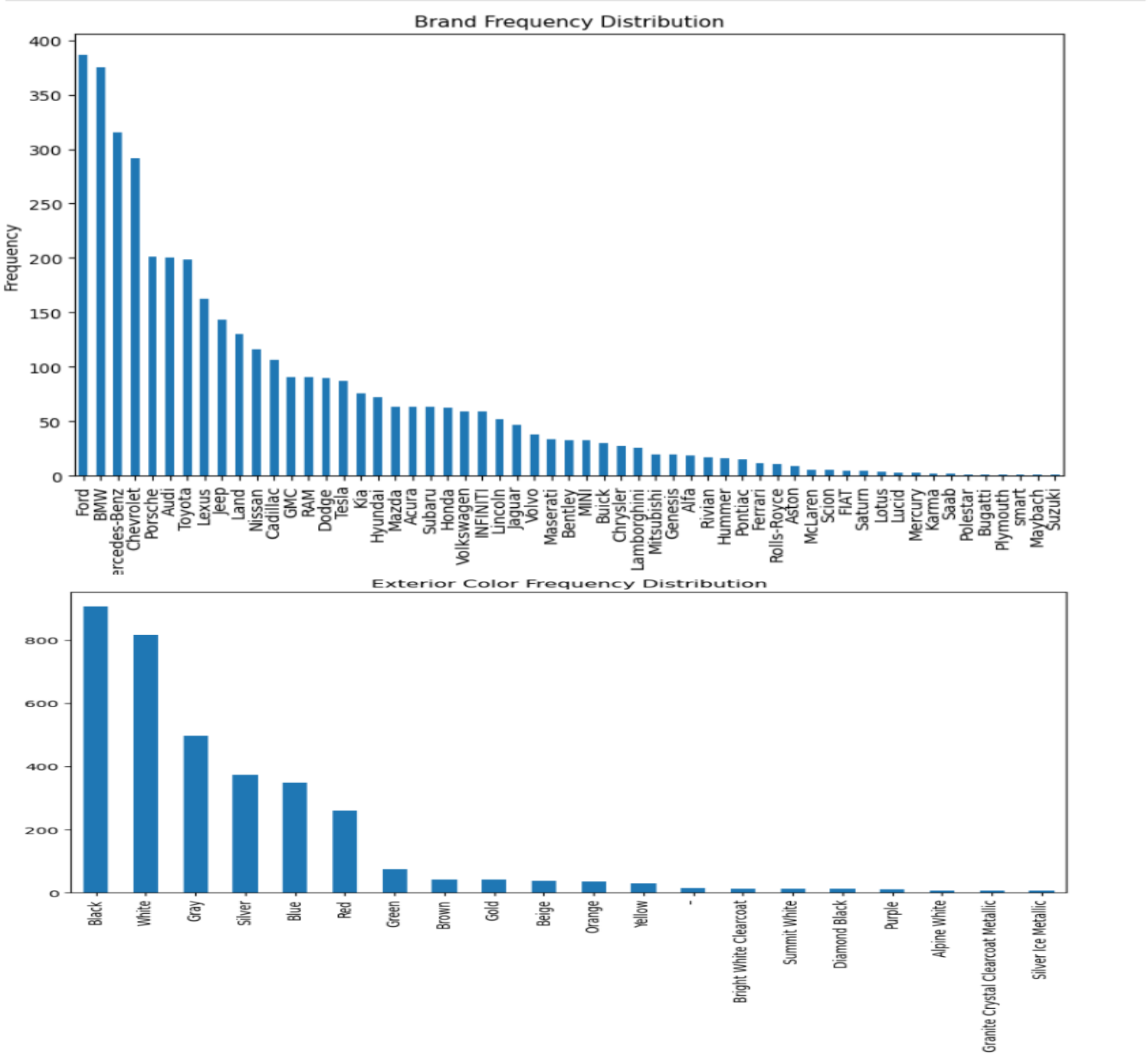
Data Preprocessing & Feature Engineering

Handling missing values & outliers

Categorical encoding (One-Hot Encoding for brand, fuel type, etc.)

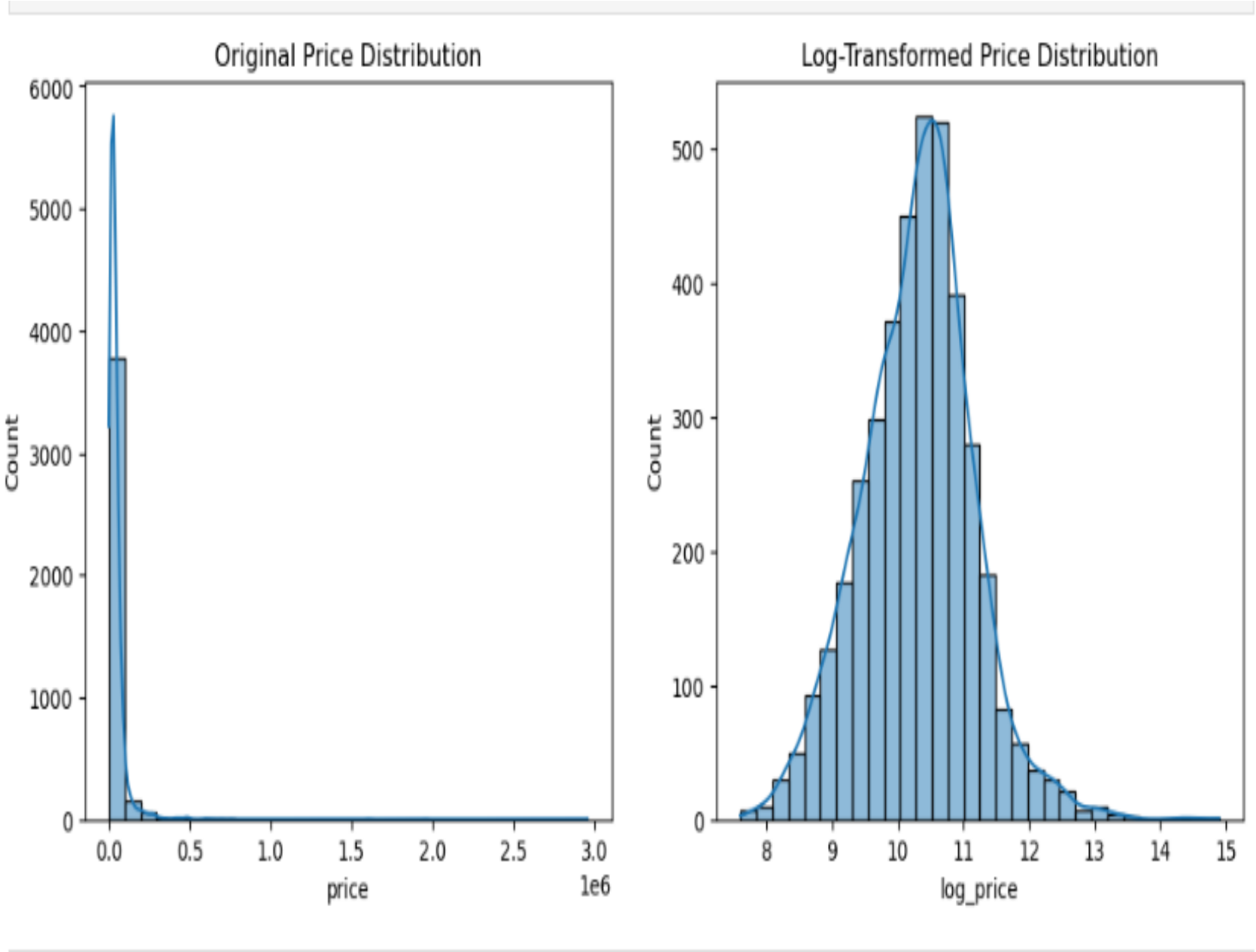
Feature scaling (Standardization for numerical features)

Derived features (e.g., **Car Age = Current Year - Model Year**)



Exploratory Data Analysis (EDA)

Price distribution (before & after log transformation)

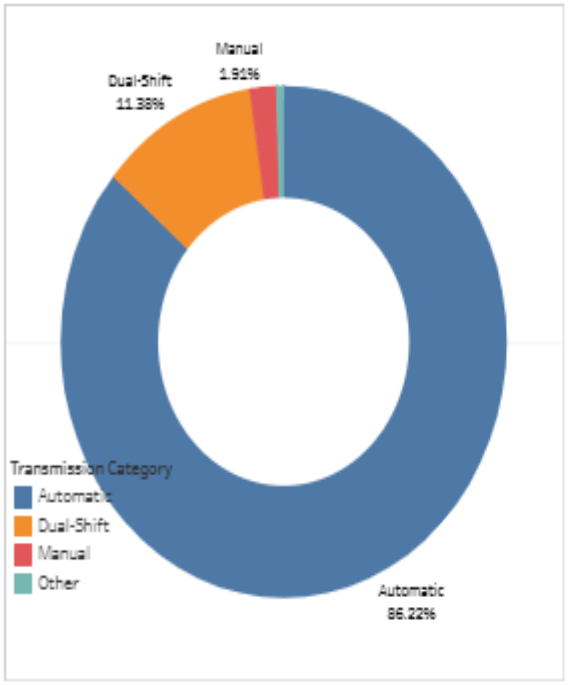


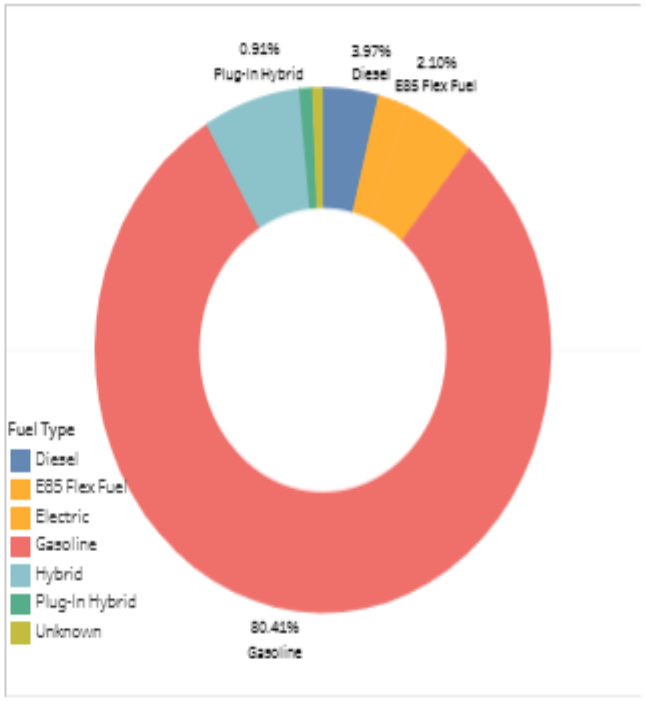
How does Tansmission category and fuel type affect the price?

Automatic cars tend to have higher prices compared to manual cars, as automatic transmission is more preferred in certain markets.

Hybrid and electric cars generally have higher resale values due to fuel efficiency and lower maintenance costs.

Transmission and fuel type are crucial factors in used car pricing. Automatic cars and fuel-efficient vehicles (hybrids/electric) command higher prices.





Model Selection & Training

We tested multiple machine learning models, including Linear Regression, Random Forest, Gradient Boosting, CatBoost, and XGBoost, to predict car prices.

Based on performance metrics such as RMSE, R^2, and MAE, Gradient Boosting, CatBoost, and XGBoost delivered the best accuracy

```
']:
# Step 1: Get the Data
X = df_final.drop(columns=["price","log_price"])
y = df_final.log_price

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.25, random_state=42) # stratify ensures same % of the

print(X.shape)
print(X_train.shape)
print(X_test.shape)

# Combine the features and target variable for training and test data
train_data = pd.concat([X_train, y_train], axis=1)
test_data = pd.concat([X_test, y_test], axis=1)

# Save the DataFrames to CSV files
train_data.to_csv('../Resources/train_data.csv', index=False)
test_data.to_csv('../Resources/test_data.csv', index=False)
```

```
(4089, 52)
(3006, 52)
(1083, 52)
```

Model	Test R^2	Test RMSE	Test MAE
Linear Regression	0.61	0.541	0.384
Ridge Regression	0.61	0.541	0.384
Lasso Regression	-0.00003	0.867	0.661
Random Forest	0.668	0.5	0.36
Gradient Boosting (GB)	0.667	0.501	0.359
XGBoost	0.682	0.489	0.356
CatBoost	0.695	0.479	0.34

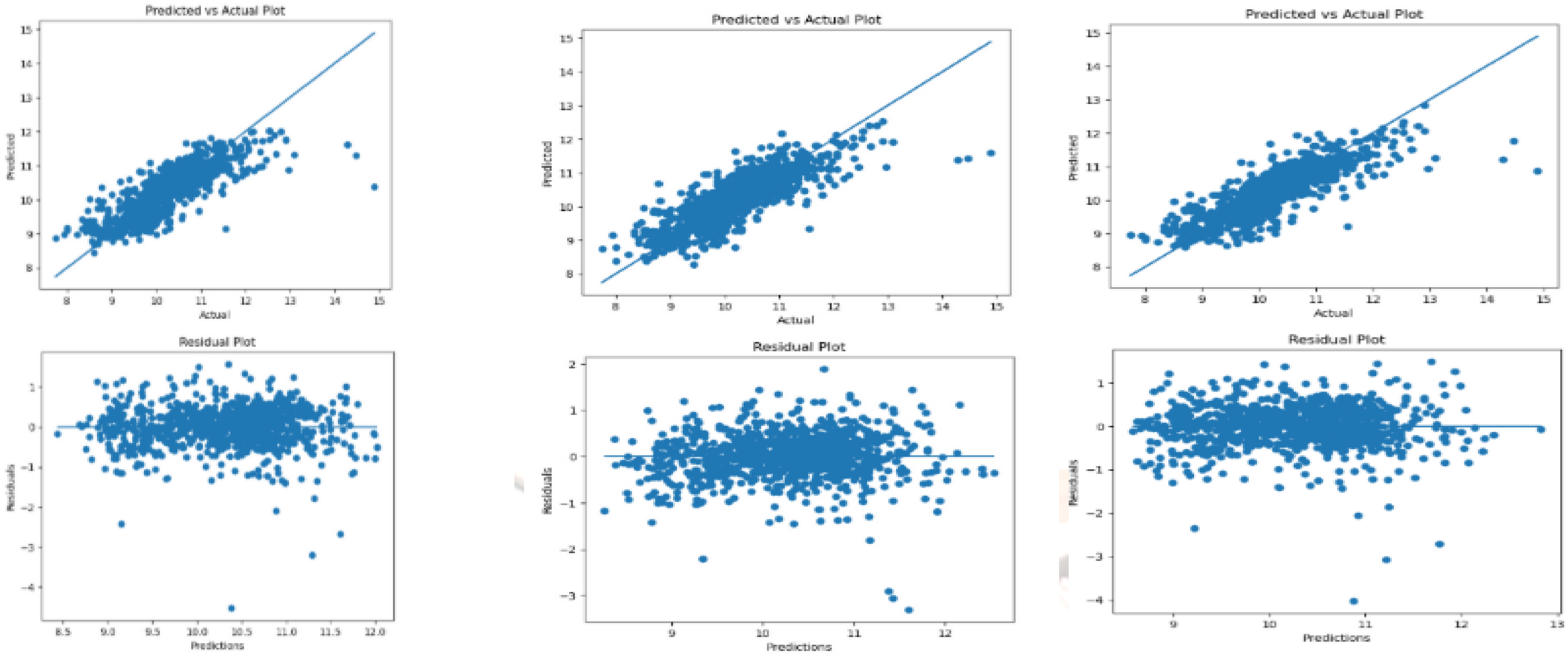
Model Evaluation

How well did model perform?

- ML models tested: **Linear Regression, Random Forest, Gradient Boosting, CatBoost, XGBoost**
- Performance comparison
- Which model performed the best?
- **CatBoost** outperformed all other models in test R^2, RMSE, and MAE, offering the best generalization performance on unseen data.

Feature	Gradient Boosting	XGBoost	CatBoost
mileage	0.6329	0.0801	39.41
model_year	0.1485	0.0343	13.11
car_age	0.0798	0	9.31
brand_Porsche	0.0383	0.0975	5.39
brand_Other	0.02	0.038	6.26
fuel_type_Diesel	0.014	0.0651	1.97
brand_Hyundai	0.0106	0.0573	1.33
brand_Mazda	0.0065	0.0392	0.81
brand_Kia	0.0057	0.0344	0.87
brand_Volkswagen	0.0051	0.0431	0.7
brand_Mercedes-Benz	0.0047	0.0231	1.31
brand_Land	0.004	0.0346	0.65
transmission_category_	0.004	0.0266	0.97

Actual vs. Predicted price plot for Gradient Boost, XGBoost and Cat Boost



Deployment & Future Improvements

Model Deployment

The trained model was deployed as:

Web Application – A user-friendly interface for price prediction

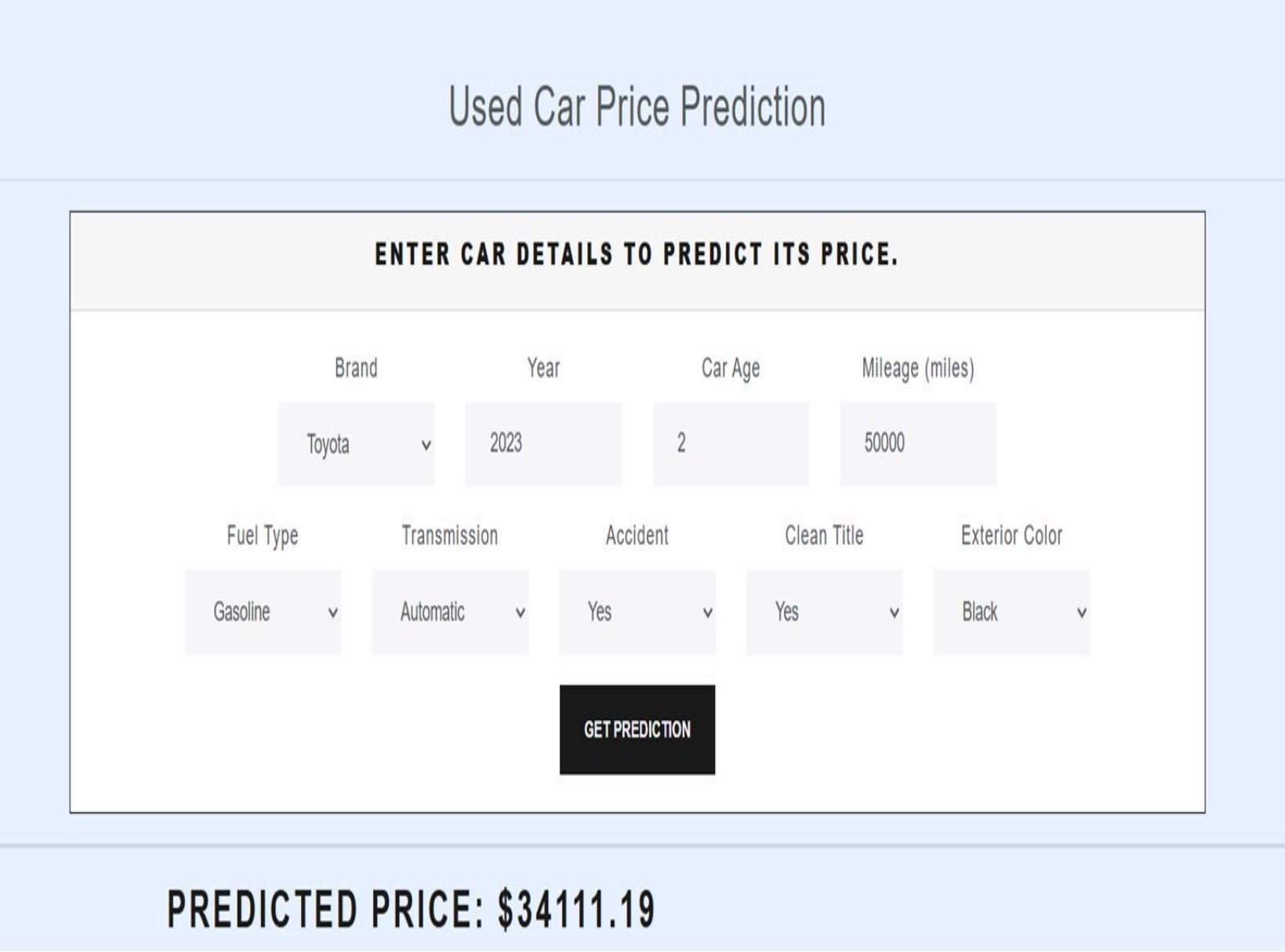
API – Allows integration with other platforms for automated predictions.

Flask – A lightweight Python web framework to create APIs for model predictions.

User sends input (e.g., car brand, model year, mileage, fuel type, etc.)

API processes the request and runs the trained ML model

API returns the predicted price as a JSON response



LIMITATIONS AND BIAS

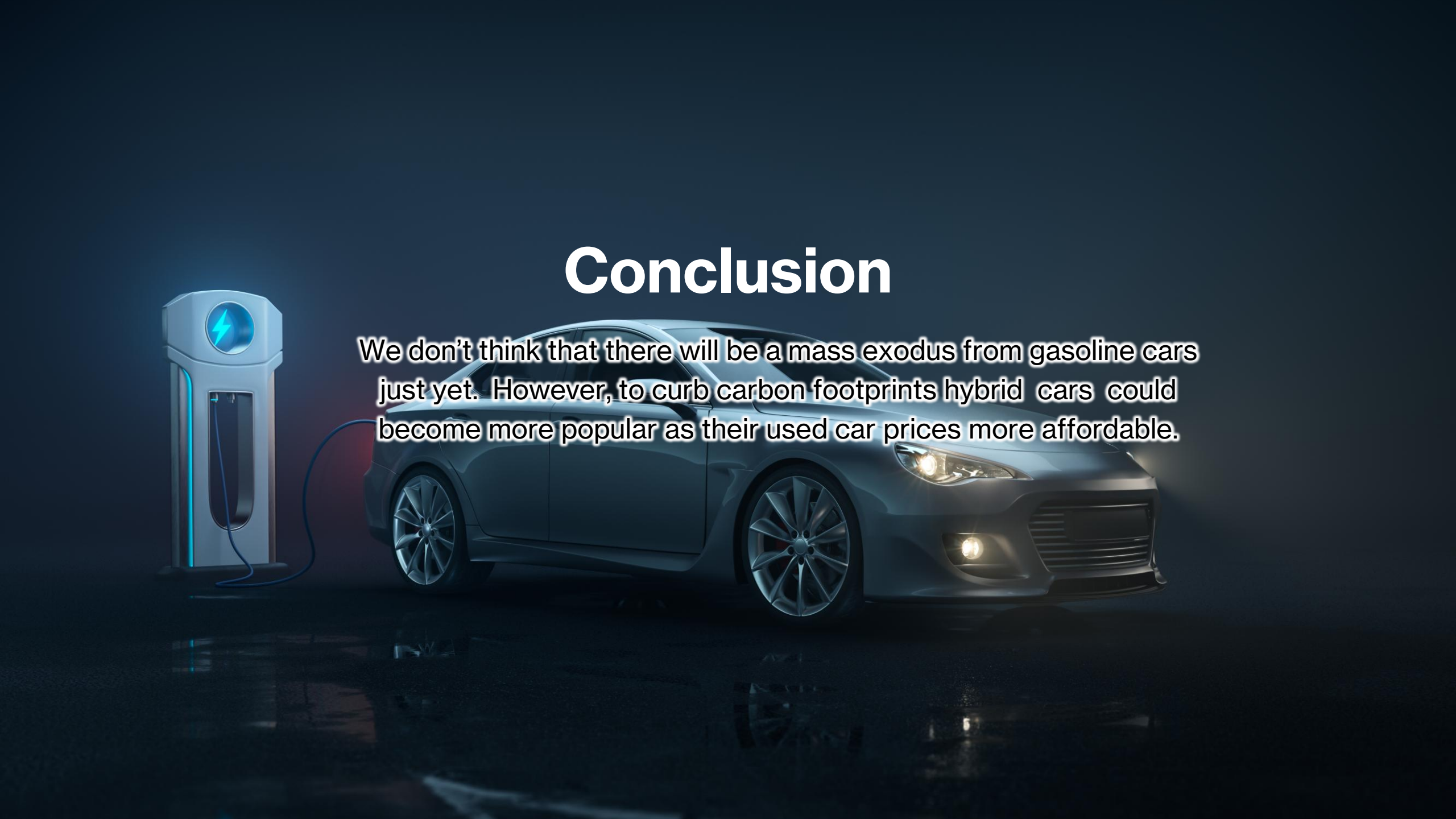
Limitations

1. **Missing or Incomplete Data**
   - Missing information (e.g., mileage, accident history, trim level), which can reduce model accuracy.
2. **Obsolete Listings**
   - Older car models/engine and model types are not representative of reflect current market.
3. **Geographical aspects**
   - Data does not includes the geographical distribution of the car prices.

Biases

1. **Make/Model/Brand/ Popularity Bias**
   - Common brands like Toyota or Honda may be overrepresented, while niche or luxury models may have insufficient data, at the same time heavily skewed the prices for accurate modeling.
2. **Temporal Bias**
   - Prices fluctuate with time due to economic factors (economic crisis/recession), seasonality, or market trends (e.g., EVs rising in popularity).
3. **Geographic Bias**
   - Cars in colder climates (which might experience more rust) or coastal areas (exposure to salt) may depreciate faster, but this might not be reflected in the dataset if location isn't properly accounted for.

Conclusion



We don’t think that there will be a mass exodus from gasoline cars just yet. However, to curb carbon footprints hybrid cars could become more popular as their used car prices more affordable.

Bibliography

Used Car Price Prediction - https://www.kaggle.com/code/satyaprakashshukl/used-car-price-prediction/input?select=extended_data.csv, accessed April 2025.

ChatGPT – www.chatgpt.com, accessed April 2025

Scikit-learn - https://scikit-learn.org/stable/,accessed April 2025

Q&A