COLUMBIA UNIVERSITY

Price Prediction for US Used Cars

COMS 4995 Applied Machine Learning

Group 4 - Project Deliverable 2

Anqi Xia (ax2171)
Azam Khan (ak4973)
Uttam Gurram (ug2146)
Xiaokun Yu (xy2550)
Yupei Shu (ys3597)

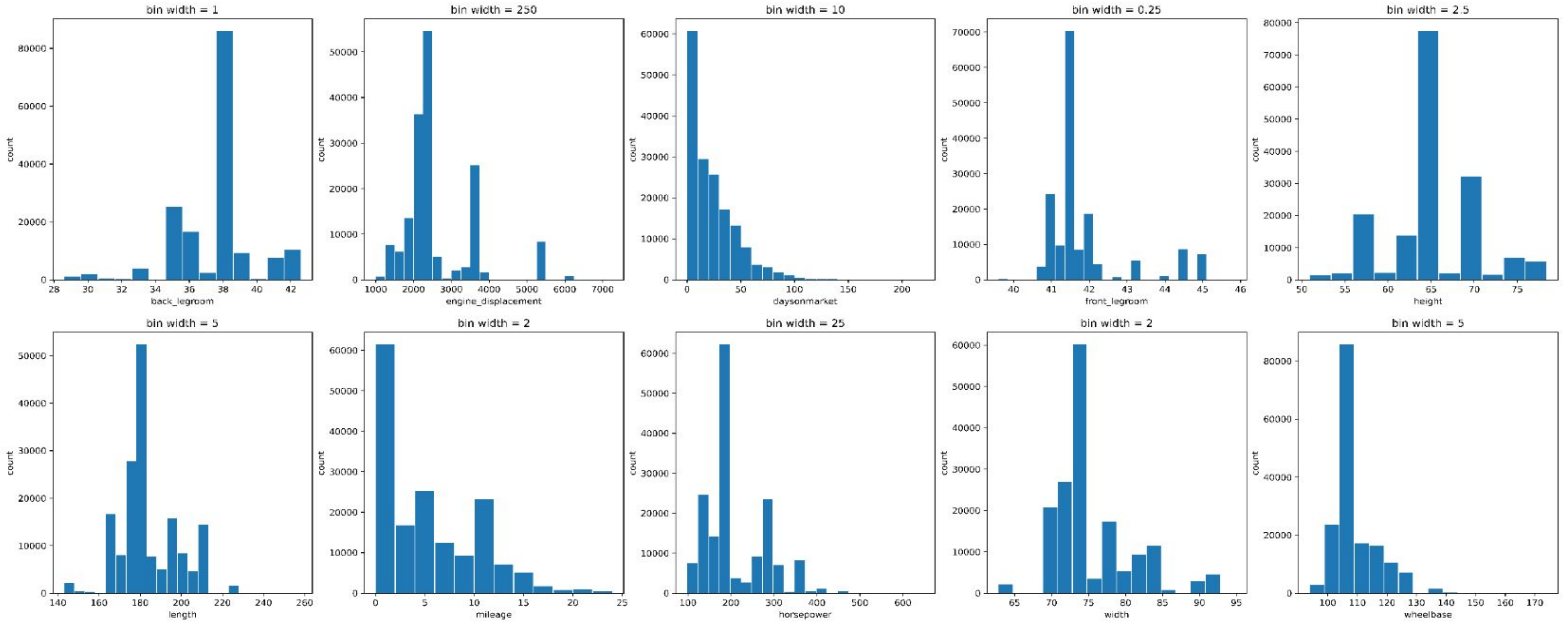
1. Data exploration with insights

COLUMBIA UNIVERSITY

Data Overview:

- The original dataset comes from Kaggle with roughly **3 million** records. However, as mentioned in our proposal, we limit our analysis to the data from year 2021.
- The dataset contains **65** features including categorical, numerical, boolean, and textual description features and roughly **176K** records.
- The target variable is **price** to be predicted

Distribution of some of the numerical features:

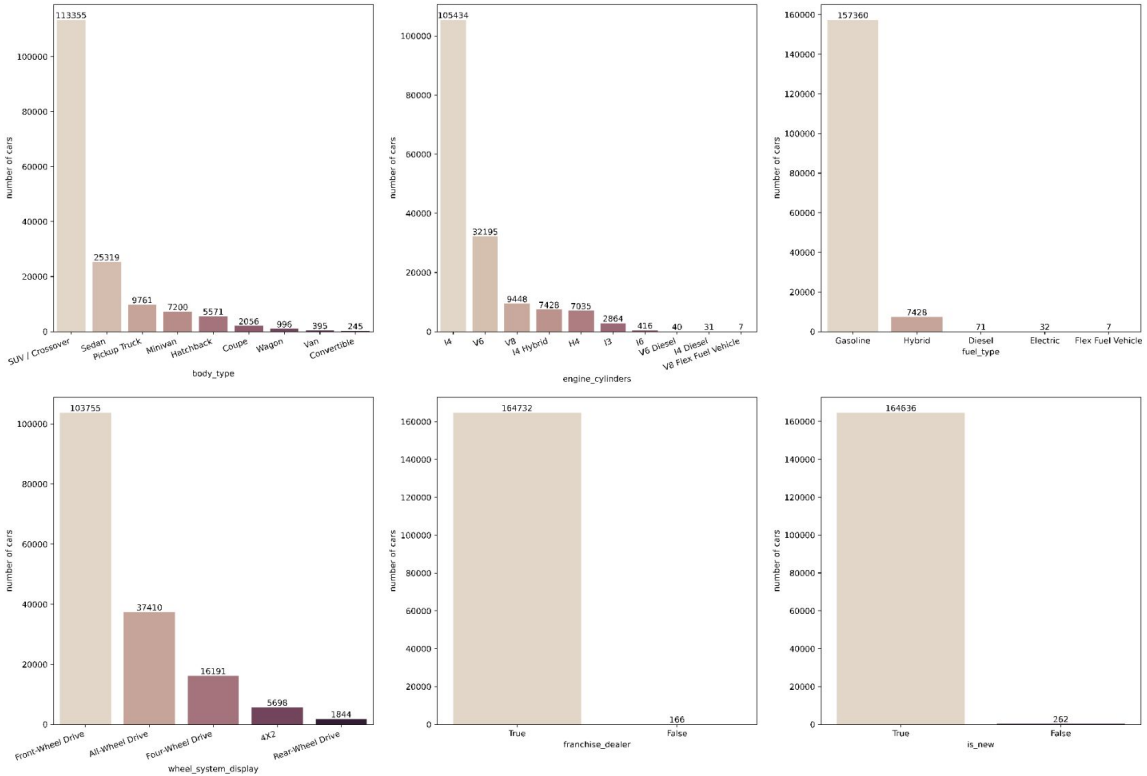


Some features such as **daysonmarket**, **mileage**, and **wheelbase** are highly skewed, so we can utilize **power or log transform** to reduce the skewness

1. Data exploration with insights

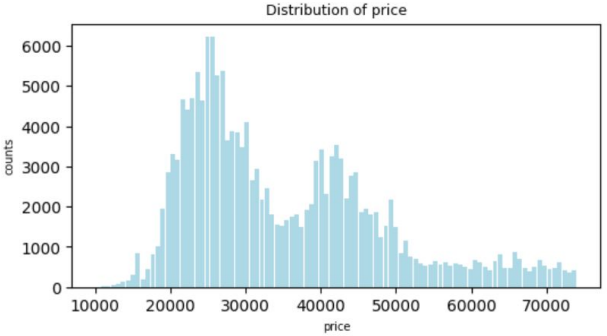
COLUMBIA UNIVERSITY

Distribution of some categorical and boolean features:



- From the distributions of some of the categorical and boolean features, we notice that there is **imbalance** in the feature values.

Distribution of target variable: price



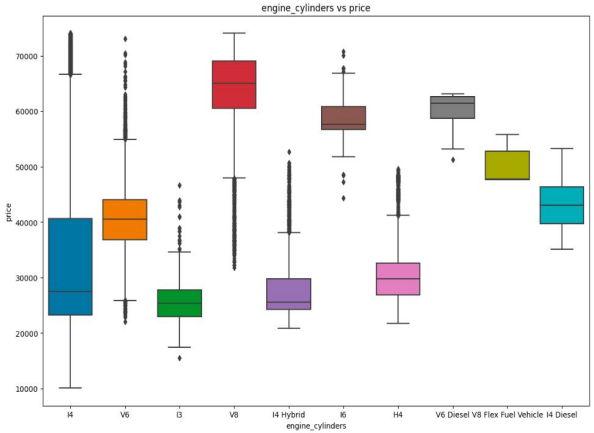
- The distribution of price is slightly skewed, with more cars being in the range of **$20000-$45000**. There are two peaks in this distribution, which suggests a bimodal distribution.

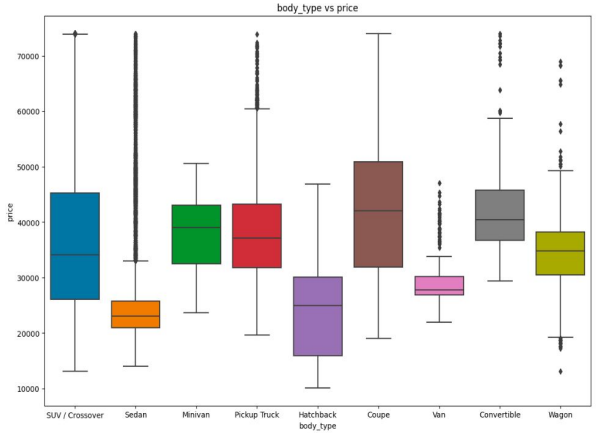
1. Data exploration with insights

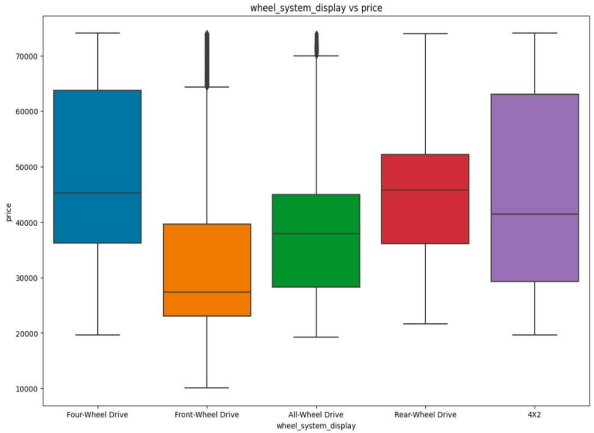
COLUMBIA UNIVERSITY

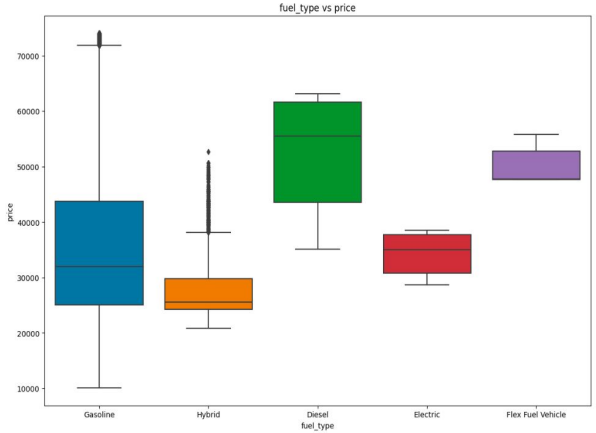
Categorical features vs price:

- Several types of engine cylinders have noticeable outliers in the car price values
- The price values of **Sedan** cars have significant number of outliers
- **Front-wheel Drive** has a lower median car price compared to other wheel system displays.
- The median car prices of **Diesel** and **Flex Fuel** type are noticeably higher than other fuel types









1. Data exploration with insights

COLUMBIA UNIVERSITY

Data processing and categorical data encoding:

- **9** categorical features with more than **10** unique values are encoded using Target encoding
- **engine_cylinders**, **fuel_type** features are encoded using Ordinal encoding and importance is assigned based on the usually observed sale prices of the cars with those different feature values
- **body_type**, **wheel_system_display** features are One Hot encoded
- **major_options** is a list of major features and is encoded as the length of list indicating the more the number of features the higher the importance
- **franchise_dealer**, **is_new** are boolean features and True is encoded as 1 and False as 0

- Some of the numerical features have metric representations like **gal, in, etc** and they are removed to make the values fully numeric

transmission_display	trim_name	wheel_system_display
Automatic	Altitude 4WD	Four-Wheel Drive
Automatic	Altitude 4WD	Four-Wheel Drive
Automatic	Altitude 4WD	Four-Wheel Drive
Automatic	Altitude 4WD	Four-Wheel Drive
Automatic	Altitude 4WD	Four-Wheel Drive

wheelbase	width
103.8 in	80 in
103.8 in	80 in
103.8 in	80 in
103.8 in	80 in
103.8 in	80 in

1. Data exploration with insights

COLUMBIA UNIVERSITY

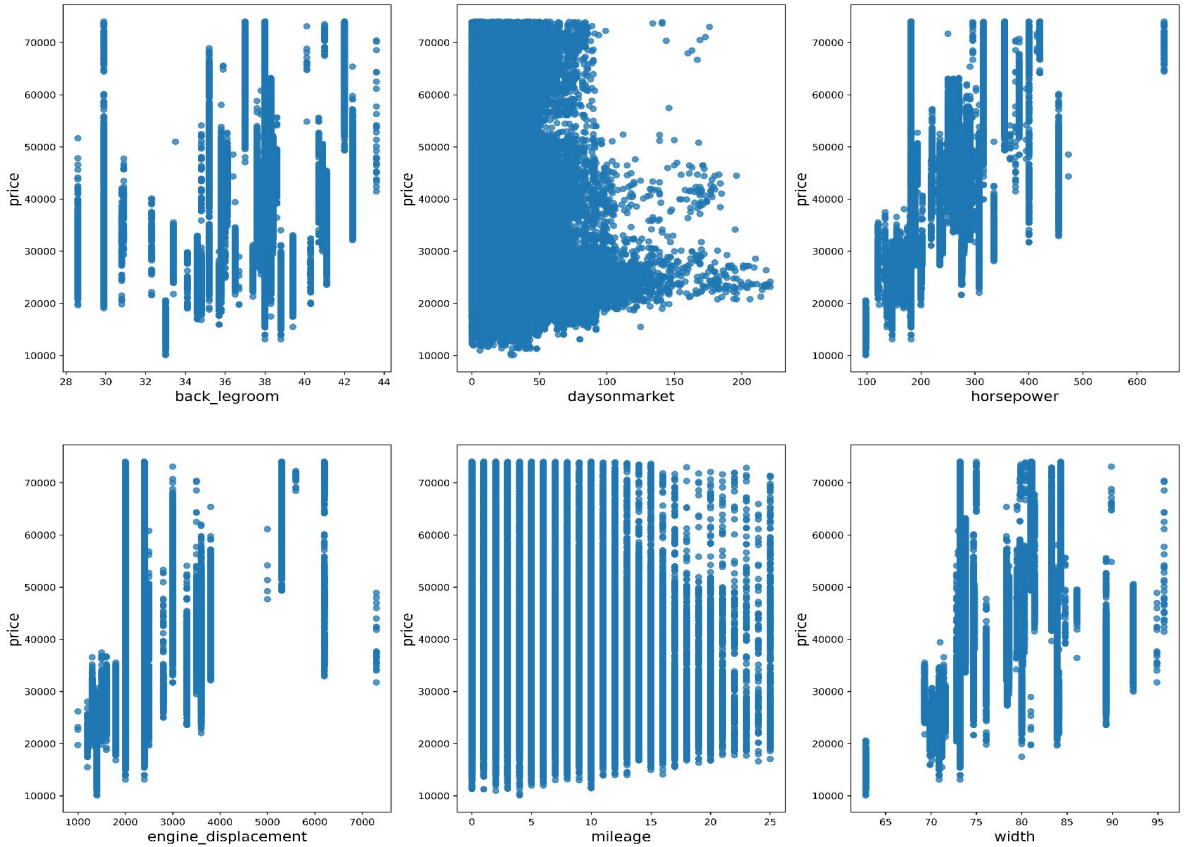
Correlation matrix:

- **franchise_make** and **make_name** are highly correlated, so **make_name** feature is dropped (0.98)
- Similarly, **length** and **wheelbase** are highly correlated, so **wheelbase** feature is dropped (0.93)

1. Data exploration with insights

COLUMBIA UNIVERSITY

Scatter plots of some of the numerical features vs price:



- Some of the numerical features are **not a linear function** of the price, so the linear regression model may not be suitable for this data

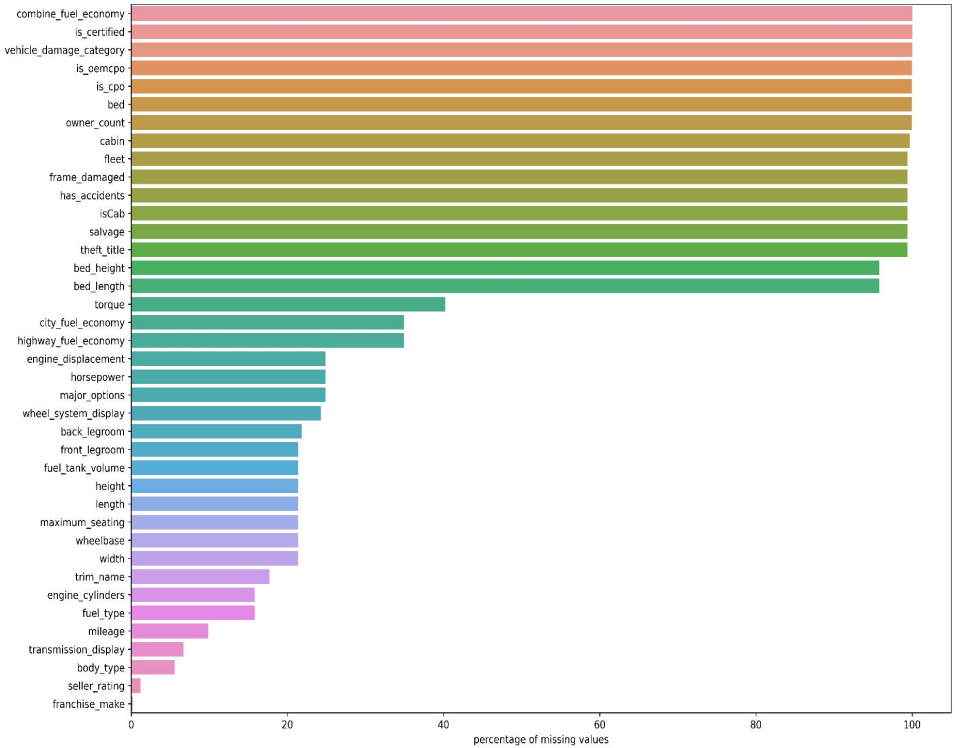
2. Cleaning and Sampling

Relevant features:

- Out of the **65** features, the **49** features which provide relevant information are considered and the rest are dropped. We also dropped textual features such as description since language processing is not our goal.

Missing values:

- All the rows in the dataset have some feature value missing. So we cannot just **drop the rows**.
- **Features** with more than **30%** values missing are dropped. In other words **19** out of the **49** features are dropped.
- **savings_amount** feature has only one unique value after outlier handling, so it is also dropped.
- All the categorical missing values are imputed using **mode** of the feature.
- Some of the numerical missing values are imputed using **median** of the features.
- **mileage** has several rows with value **0**, so the missing information is filled with **0**.
- **major_options** missing values are considered as lack of any major options, so the missing information is filled with **0**.

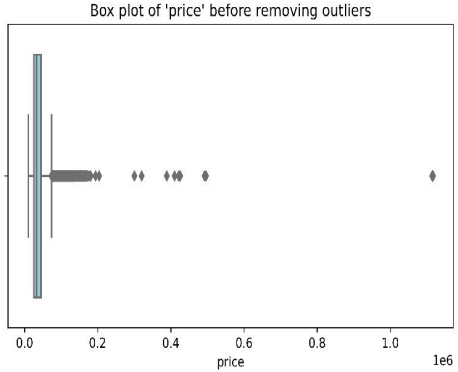


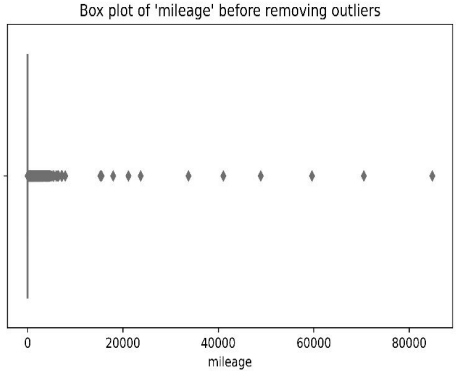
2. Cleaning and Sampling

COLUMBIA UNIVERSITY

Outliers removal:

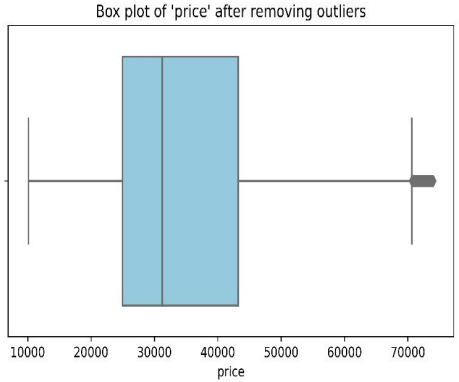
- Outliers are present in the target variable **price**, feature **mileage** as observed from the box plots, so removing this noise will improve the performance of models
- We used Interquartile Range to identify and remove the rows with **price**, **mileage** values outside the range
- Roughly **12K** records are removed (**6.6%** of the **176K** records)

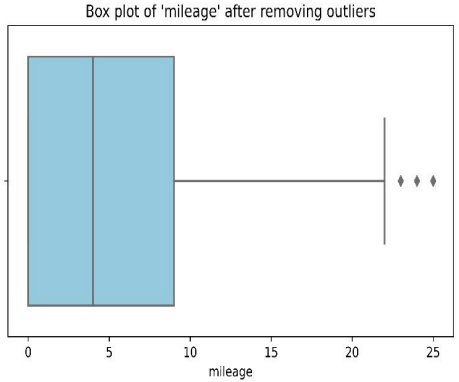




Data sampling:

- Since this is a regression task there is no relevancy of class imbalance, so we do not require any **data sampling**.





Data splitting:

- Since we do not need to worry about class imbalance we use **random splitting** to split the dataset.
- We split the dataset into training (60%), validation (20%), and test (20%) sets.
- The validation set will be used for finding the optimal hyperparameters

3. Machine learning techniques proposed

Models Choice:

- Linear Regression (baseline)
- Lasso, Ridge, and Elastic-Net
- Decision trees and Ensemble Methods (including Bagging and Gradient Boosting)
- Deep Learning Model: TabNet

Model Performance Evaluation:

- AUROC
- RMSE

Hyperparameter Tuning:

- Grid Search
- Random Search
- Hyperband