FLIP ROBO

CAR PRICE PREDICTION

Submitted by:

Neha Kamath

ACKNOWLEDGMENT

I would like to take this opportunity to thank my mentors at FlipRobo Technologies for their guidance and support in the completion of this project.

Index

Table of Contents

INTRODUCTION

- ## Background

With the covid 19 impact in the market, we have seen lot of changes in the car market. Especially now some cars are in demand hence making them costly and some are not in demand hence cheaper.

One of our clients works with small traders, who sell used cars. With the change in market due to covid 19 impact, our client is facing problems with their previous car price valuation machine learning models.

So, they are looking for new machine learning models from new data. We have to make car price valuation model. This project contains two phase-

- Data Collection Phase –

  We have to scrape used cars data. More the data better the model.

- Model Building Phase-

  After collecting the data, we need to build a machine learning model.

Analytical Problem Review

Analytical Modeling

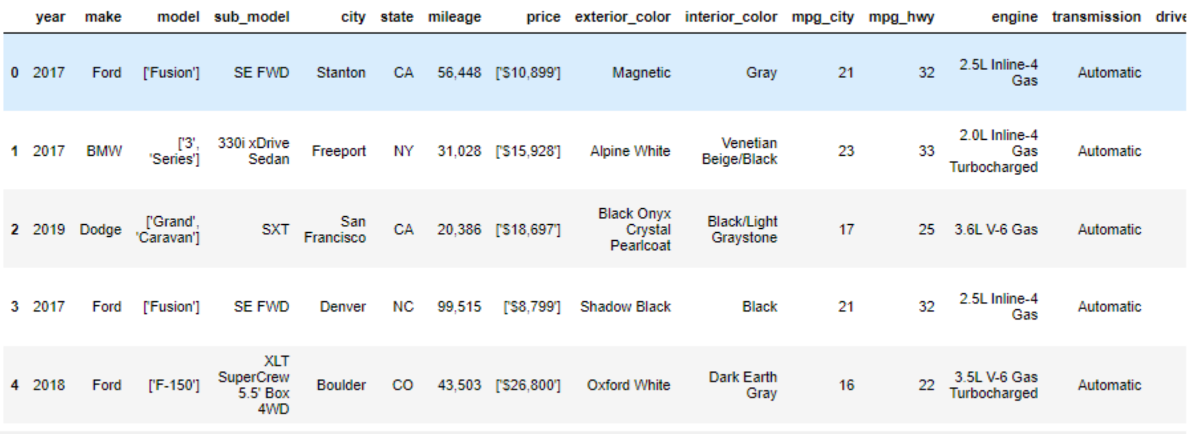
For the purpose of car price prediction analysis, the most important factors under consideration would be Brand, model, variant, manufacturing year, driven kilometers, fuel, number of owners, location and at last target variable Price of the car.

In the course of detailed analysis, we have applied various techniques such as Data Cleaning, Exploratory Data Analysis, Data Pre-Processing, Model Building, Evaluation and Selection.

Data Sources

Our primary source of data for this project has been the data collected from cars.com car reviews. It includes a total of 19 features and the total number of records is approx 9,000.

Here is a glimpse of our dataset:



	year	make	model	sub_model	city	state	mileage	price	exterior_color	interior_color	mpg_city	mpg_hwy	engine	transmission	drive
0	2017	Ford	['Fusion']	SE FWD	Stanton	CA	56,448	['$10,899']	Magnetic	Gray	21	32	2.5L Inline-4 Gas	Automatic	
1	2017	BMW	['3', 'Series']	330i xDrive Sedan	Freeport	NY	31,028	['$15,928']	Alpine White	Venetian Beige/Black	23	33	2.0L Inline-4 Gas Turbocharged	Automatic	
2	2019	Dodge	['Grand', 'Caravan']	SXT	San Francisco	CA	20,386	['$18,697']	Black Onyx Crystal Pearlcoat	Black/Light Graystone	17	25	3.6L V-6 Gas	Automatic	
3	2017	Ford	['Fusion']	SE FWD	Denver	NC	99,515	['$8,799']	Shadow Black	Black	21	32	2.5L Inline-4 Gas	Automatic	
4	2018	Ford	['F-150']	XLT SuperCrew 5.5' Box 4WD	Boulder	CO	43,503	['$26,800']	Oxford White	Dark Earth Gray	16	22	3.5L V-6 Gas Turbocharged	Automatic	

Data Preprocessing Done

The data pre-processing for this particular dataset required feature engineering, checking missing values and imputing them, encoding categorical variables to numeric, outlier checks, skewness treatment as well as scaling.

Outliers were present in most of the variables and so, the same were removed to a large extent using z-scores. Values with a z-score above 3 were eliminated.

The dataset was subject to skewness checks next and it was removed using the power transform function.

Since the values had widely different ranges, the dataset was scaled down to a standard scale using the Standard Scaler.

Hardware and Software Tools

The libraries and packages we have used on this projected are listed below:

- Data Processing- Numpy(numerical data wrangling), Pandas(data analysis)
- Data Visualization- Matplotlib, Seaborn (graphical representations)
- Text Processing- RegEx and Natural Language Toolkit libraries

Model/s Development and Evaluation

Possible problem-solving approaches (methods)

The target ‘label’ contains numeric values Hence, the logical approach to building a suitable prediction model is to use regression models such as linear regression, RFs, GBs, DTRs, KNN, XGB, etc.

Testing of Identified Approaches (Algorithms)

- Extreme Gradient Boosting Regressor (XBG Regressor)

Performance of models

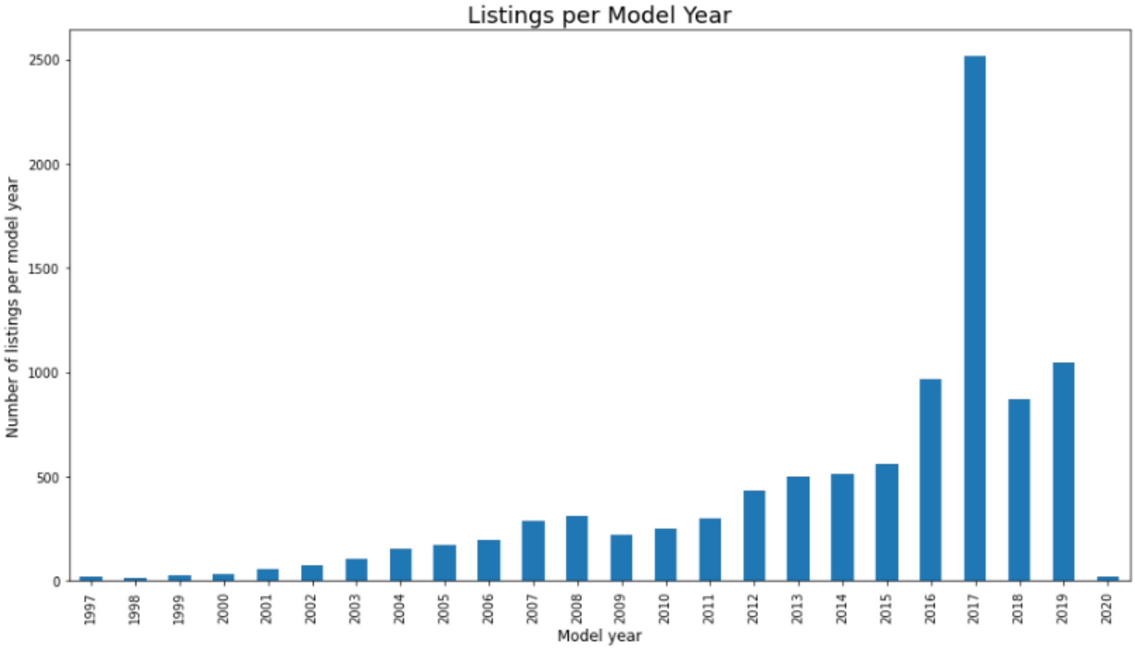
XGB Regressor:

This is a machine learning algorithm for regression purposes. Extreme Gradient Boosting (XGBoost) is an open-source library that provides an efficient and effective implementation of the gradient boosting algorithm.

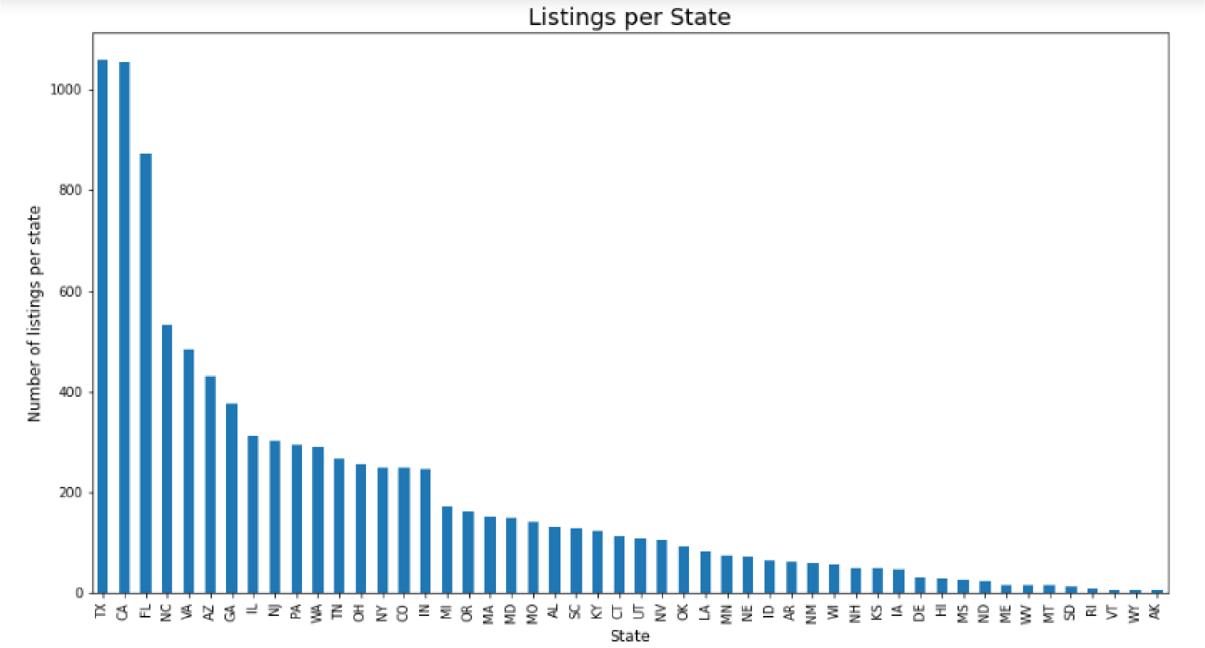
A baseline model executed gave out an RMSE score of **0.260026.**

Visualizations

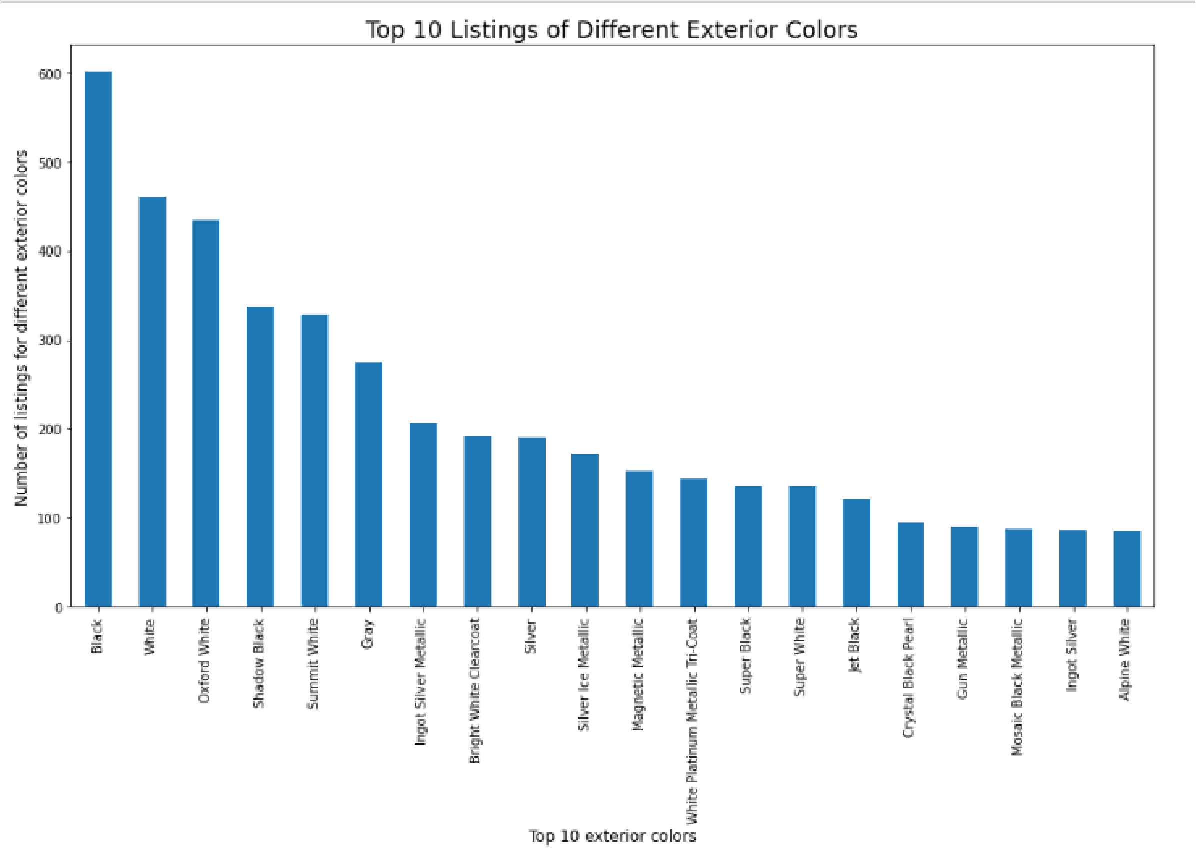
Vehicle listings for different years



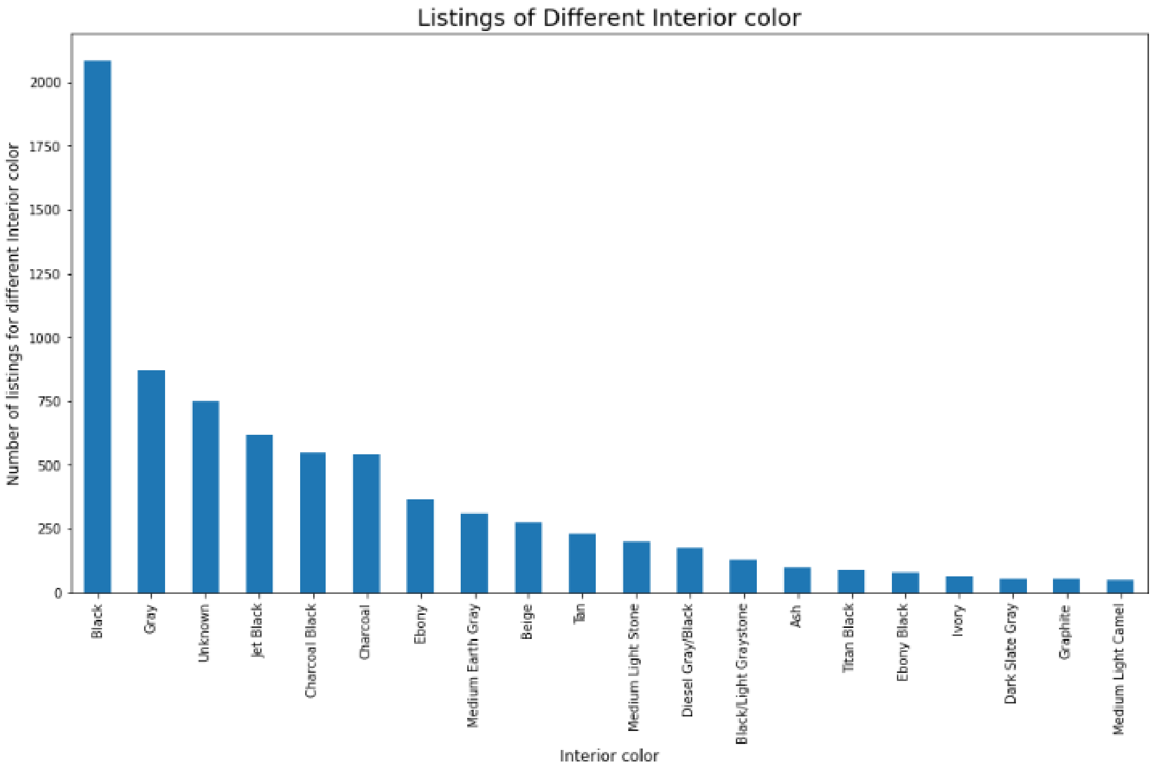
Vehicle listings from different states



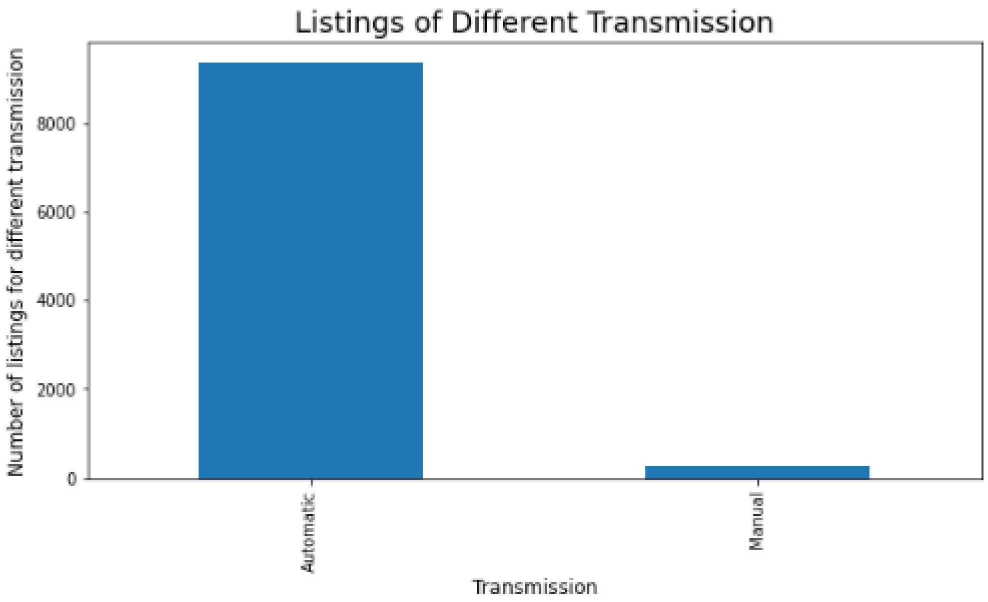
Top 10 Listings of Different Exterior Colors



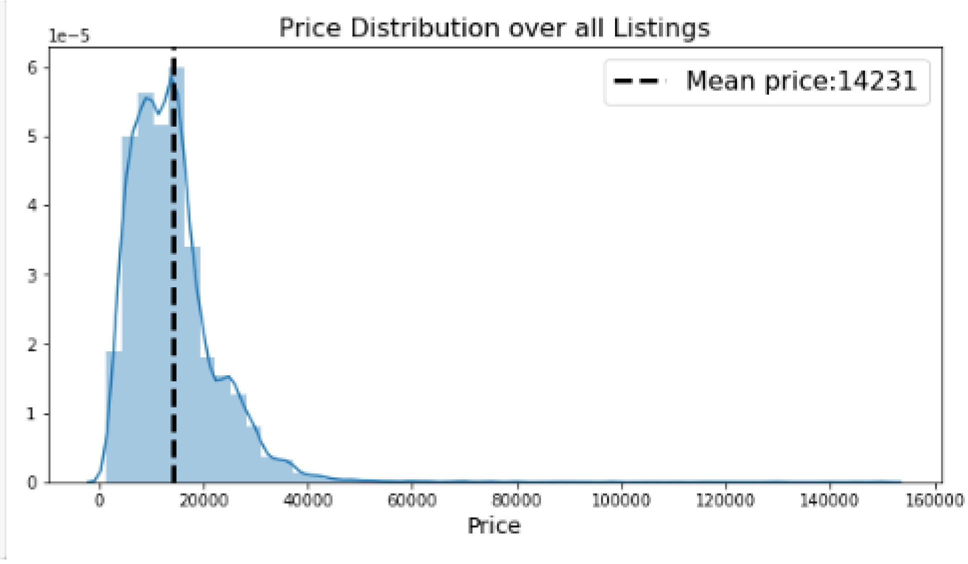
Listings of Different Interior color



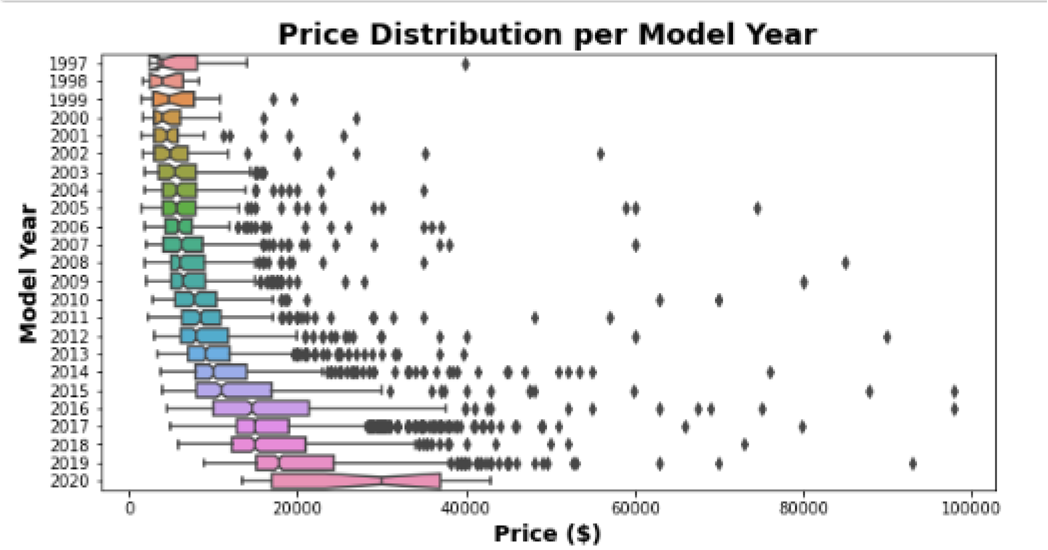
Listings of Different Transmission



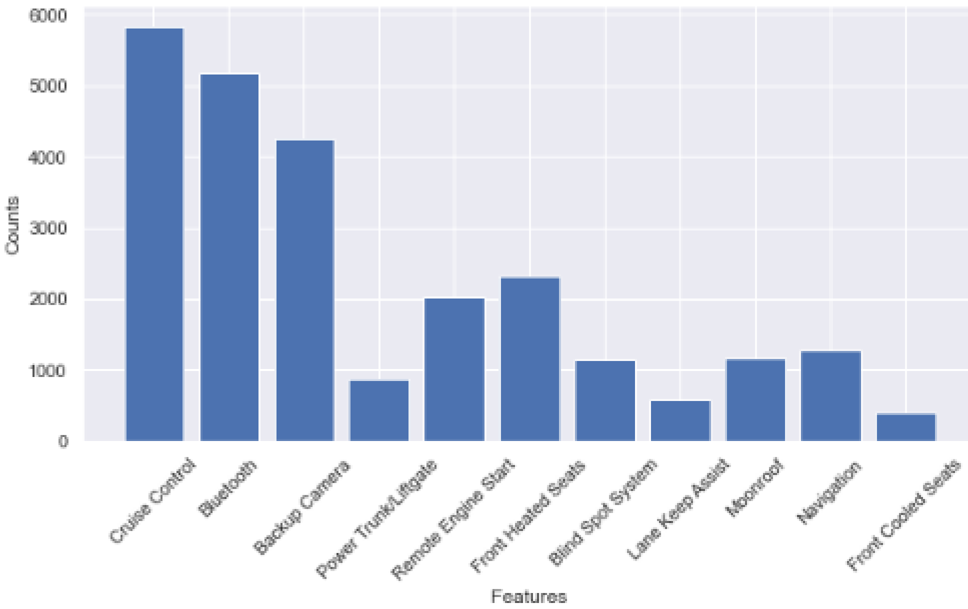
Listings of Different Transmission



Price Distribution per Model Year



Features


6000
5000
4000
3000
2000
1000
0
Counts
Cruise Control
Bluetooth
Backup Camera
Power Trunk/Liftgate
Remote Engine Start
Front Heated Seats
Blind Spot System
Lane Keep Assist
Moonroof
Navigation
Front Cooled Seats
Features

CONCLUSION

• Key Findings and Conclusions of the Study

From the visualization and machine learning modelling, I found out several conclusions below.

TX, CA and FL are the top 3 used car markets, however the car prices between CA and non-CA/TX are very close.
3-year old used cars are the largest share in the current used car markets. The possible reasons could be the popularity of 3-year leased cars.
The most popular automakers in the US are Ford, Chevrolet, Nissan and Toyota. The possible reasons could be the US-brand loyalty for Ford and Chevrolet and the popularity of pickups, and another reason is that Japanese second-hand cars are popular and reliable.
The certified pre-owned cars are more expensive than normal used cars.
When predicting the used car prices, the most important features include mileage, mpg and model year. From visualization, we see that the higher the mileage, the lower the price. And also for luxury brands vehicles, they tend to have very low mpgs. The model year is also correlated with mileage and the older the car is, the lower price it will be.

• Limitations of this work

Since there are no available api to use, web scraping and html parsing are used to extract information from Truecar.com. However, the website limit the total number of listings to be presented to users as 9900 cars, even though there are over 1 million listed used cars on the website
.
When I tried to get the 9900 cars, I used to default 'Best Match' search term to minimize the potential bias caused by the sampling and presenting procedure by the website.

One of the important assumptions I make is that I assume the 9900 vehicle listings I scraped from the Truecar.com are randomly sampled from the total population. Therefore, the following analysis can well represent the real distribution and characteristics of whole population.