SI618 Final Project Zhongqi Liu

Part I: Motivation & 4 Questions:

Nowadays, cars became one of the most essential "tools" that people need to use daily. Everyone drives to school, to work or has road trip with their family and friends. In United States, there were 17.21 million new cars and 40.42 million used cars sold in 2018. It's not hard to tell that most people will consider buying a used car since new cars depreciate so fast. However, car buying experience could be so pain to some people who doesn't know cars very well. Especially when these people decided to buy a used car. They often tricked by car dealerships by paying extra for their purchase. Therefore, I came up with one idea: I'll try to build a car valuation program. The four question I came up was:

1. What's the number of cars for a specific model listed in each production year? I think maybe drawing a histogram could help me understanding that. This question is also interesting since it could show whether the market is still active for some old car. For example, we should expect the number of listing should be small for a 1965 Ford Mustang. Moreover, what's the current price distribution for this car? This information is also extremely useful since people need to consider their budget when buying a car.

2.What's the relationship between a used car's price and mileage, production year and trim level? This question is critical for the prediction since we want to see the relationship between car's price with those factors. Drawing a scatterplot should be helpful. Also, are there any interesting facts?

3.What's the best model to predict car's price? A simple linear regression is enough, or we need more complex machine learning method such as random forest? How could we determine which model is better?

4.Are there any other interesting and worth noting facts for this car that will also influence the car's price? For example, an all-new 4th generation BMW X5 were launched in 2019 and it's already on sale, could that influence the price of previous generations X5s? That should be worth exploration.

Part II: Data Source:

If we want the result to be meaningful, instead to search datasets on website, I decided to scrape data from car-buying website such as cargurus.com or kbb.com. Those websites have cars listed nationwide and it contains thousands of listings for each car. And since one of my friends was interested in buying a used BMW X6, I decided

to work on this model. One thing needs to notice that due to legal issues of scraping, it is almost impossible to scrape all cars and all models from those websites, otherwise I'll be blocked and might be sued. Therefore, I'll only work on this specific model. And I'll also provide the data I scraped named "cars.json". I highly not recommend scrape the data again and if someone run my code. The data I scraped was from cargurus.com and kbb.com. The data is json formatted and there are 1,837 entries in the dataset. The cargurus URL is : https://www.cargurus.com/Cars/inventorylisting/viewDetailsFilterViewInventoryListing.action?sourceContext=carGurusHomePageModel&newSearchFromOverviewPage=true&inventorySearchWidgetType=AUTO&entitySelectingHelper.selectedEntity=d1137&entitySelectingHelper.selectedEntity2=&zip=48103&distance=50000&searchChanged=true&modelChanged=false&filtersModified=true#resultsPage=1, and there are total of 72 pages of listing, I scraped them all and save it locally. The URL for KBB is https://www.kbb.com/cars-for-sale/cars/usedcars/?p=1&distance=none&atcmakecode=bmw&atcmodelcode=x6&nr=100&s=kbbrank Again this URL is only for the first page. I scraped all pages and saved it locally named "cars.json". The data I plan to use contains four parts: car's listing price, model year, mileage and trim level. All those 4 variables are saved as integers. Year and trim level will be used as categorical variable for question 3.

Part III: Methods:

Q1: I made up a data frame with four columns: Year, Price, Mileage and Trim. There are total three trim levels for BMW X6: sDrive 35i, xDrive 35i and xDrive 50i. sDrive 35i is the lowest trim and xDrive 50i is the highest trim. Missing data and noisy data will be removed since majority of listings are appropriate. Remove couple listings will not change my result. To solve Q1, I will use a histogram to plot number of listings for each year. And try to discover whether the number of listings is increasing, decreasing or moving up and down. I will also plot out other variables such as distribution of price. One challenge would be filtering data, since data scraped was messy and it took me a long time to gather useful information I need. I overcame this by carefully looking at the html file and find correct information. Also, I added many if statements to determine whether the data is appropriate.

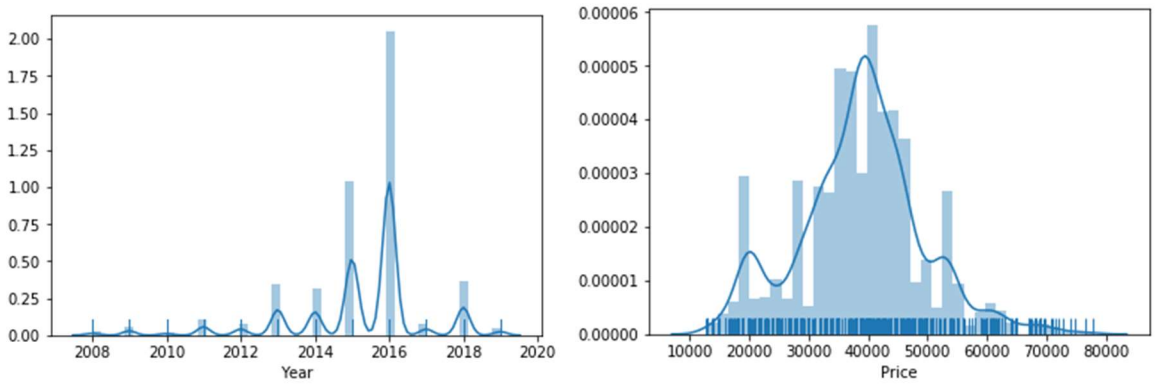
Q2: For this question, since I already have a data frame and missing data were already removed at the very first step, I'll just need to draw a pairwise scatterplot to discover the pattern. One thing I feel challenging is how should I interpret the plot between trim level against other variables. However, by looking closely to the plot, I found some interesting facts about trim level that worth mentioning.

Q3: For this question I'll split the data into training set and testing set. With 70% training and 30% testing. Again, there is no missing data since it has been cleared out at first step. I'll try to train linear regression and random forest by training set and test the fitness by testing set. After that, I'll calculate test MAE for both methods and compare the test MAE. Method with lower test MAE is better. The challenge I faced is the way to calculate test MAE. I googled and figured it out.

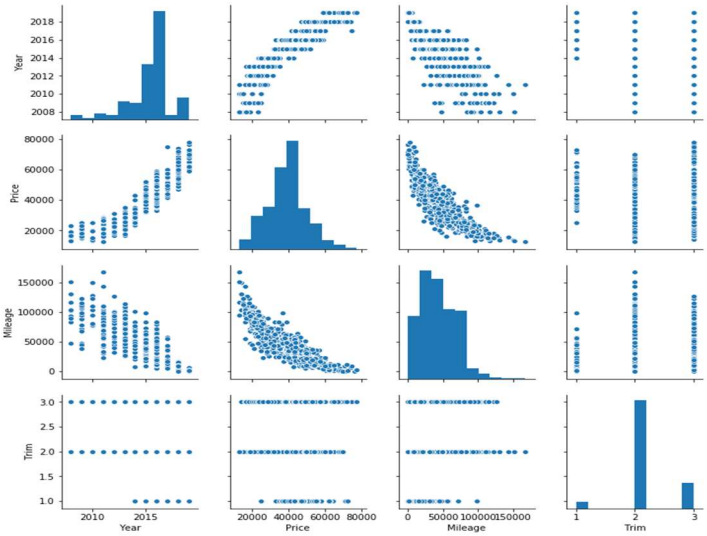
Q4: This time I'll work on my original dataset. Again, there is no missing data since it has been cleared out at first step. Since there are only 2 generations of X6 and it was facelifted in 2015. Therefore, I want to test whether the price different significantly due to facelift. First, I plotted the mean price for each year, from 2008 to 2019. But we cannot make conclusion from just looking at plots. So, I added a new column called second_Gen. If the car's model year is before, it will be the first generation and the value would be false. Otherwise, it belongs to second generation and the value would be true. After that, I ran an ANOVA test to test whether this factor (facelifted or not) will influence price significantly. The difficulty I faced was to determine the right approach. Since we need numbers to reach out to conclusion, but not looking at graph and guess. I got the idea by reviewing past lectures and did some research by myself.

Part IV: Analysis and Results:

Q1: Below are some basic information about the dataset, along with some plots. The dataset contains 1,837 entries. As we could see from the plot for number of cars in each year, most cars listed were produced in 2016. Also, as we could expect, there are not many cars listed for 2008 through 2012. Because buying a 10-year BMW is not a wise choice. So, the market for those cars is very minimal. Moreover, as we could expect, the trend for total number of listings by year increasing from 2008 to 2016, then decreasing from 2016 to 2019. This means most owners will sell their car after owning it 3 years. It's also possible the previous owner leased the car for 36 months. That could explain why the 2016 model has the greatest number of listings. Furthermore, we could also discover from the distribution plot for price that price range varies from around 14,000 to 78,000. Majority listing price is ranging from 30,000 dollars to 50,000 dollars, which means those cars depreciates a lot, since they original MSRP starts over 60,000 dollars. That's also prove the facts why people love to buy used cars.



Q2: As we could find out in the pairwise plot, the relationship between year and price is increasing, means as year increase, price will also increase. Also, the relationship between price and mileage is decreasing, meaning as mileage increase, price will decrease. Moreover, the relationship between mileage and year is also decreasing. Those all make sense. However, when I come up with trim level, I cannot provide useful interpretation. I do notice one fact that if we look at year vs trim, we could find out trim 1, which represent sDrive 35i, was not available until sometime around 2015. That means before that timestamp, there was only two trim levels. That is one interesting finding. Moreover, if we look at lower right corner, we could find out that trim level 2 have significant values comparing to trim level 1 and trim level 3, that suggest that most people purchased xDrive35i version instead of sDrive 35i and xDrive 50i. Other interesting finding is nether the relationship between price vs year or price vs mileage is highly linear. It somehow linear but not exactly. That brings to the Question 3, choosing right model to predict.

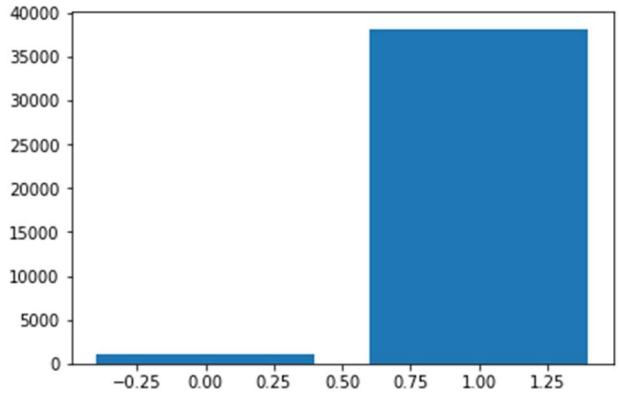


Q3: After splitting my data into training and testing set, I fitted a linear model and random forest model. I tried to use Year, Mileage and Trim to predict Price. I changed Year and Trim to categorical in the prediction. The result for linear regression is very satisfactory:

Dep. Variable:	Price	**R-squared:**	0.914
Model:	OLS	**Adj. R-squared:**	0.913
Method:	Least Squares	**F-statistic:**	959.3
Date:	Sun, 21 Apr 2019	**Prob (F-statistic):**	0.00
Time:	19:32:21	**Log-Likelihood:**	-12152.
No. Observations:	1285	**AIC:**	2.433e+04
Df Residuals:	1270	**BIC:**	2.441e+04
Df Model:	14		
Covariance Type:	nonrobust		

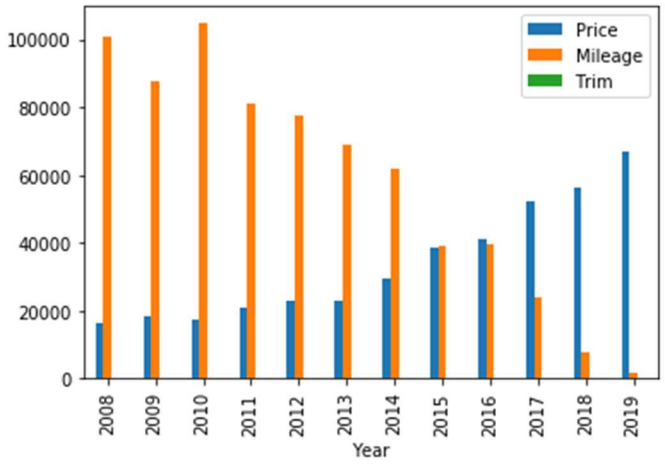
As we could see, the R-squared value is high, suggesting linear model fit the data well.

But how could I determine which method is better? Here I did some research and decided to use test Mean-Absolute-Error to measure performance for both methods. Mean absolute error was calculated by predicting test set and find the absolute difference between predicted value and real value, then sum up the difference together. This method is more appropriate than MSE because MAE could provide more meaningful interpretations. I provided interpretation below. After I calculate test MAE for both method, linear regression outperformed RF with test MAE 996.1251194982622. From the plot, the left one is MAE for linear regression and right one is MAE for random forest. Here we could see comparing to RF, linear regression did a good job, with test mean absolute error of 996.



That means this model will provide a fair price with the error of +- 996. As we could see the histogram for price above, considering for this model most car's price is ranging from 30,000 dollars to 50,000 dollars, an +- 996 dollars error is acceptable.

Q4: As we could see below, there are two major worth-noting information. First if we check the blue bar, there is the upward trend of price against model year, which make sense because newer cars worth more. Moreover, as we check the orange bar, we could see there is a downward trend for car's mileage from 2010 to 2019, which also make sense since older cars tend to have higher mileage. If we look close enough, we could see the price gap between 2014 to 2015 is noticeably larger than the price gap between 2013 to 2014. We need to keep in mind that facelift for this car happened in 2015. However, by just looking at this graph, we really cannot make any conclusion. We need numbers to prove our hypothesis, which is there is a significant difference of price between and after facelift.



Then I created another column, as described in methods section. Then I complied ANOVA test and result are below:

	sum_sq	df	F	PR(>F)
C(second_Gen)	1.016720e+11	1.0	1882.043693	1.403947e-283
Residual	9.913064e+10	1835.0	NaN	NaN

As we could observe from ANOVA table, the p-value is very small (smaller than 0.01), therefore, we will reject the null hypothesis, which is there is no significant difference for used car price based on first generation and second generation. And we come up with the conclusion that facelift DO have a crucial effect on a used car's price.