Jontavius Caston

DSC 680

March 27, 2020

Buy Used not New

**The Data**

For this project I decided to use go into the domain of the used car market. While preparing for this project I plan to use sites such as Kelly Blue Book, Car Fax, websites that shows concerns that new cars have on the market, Car Guru, craigslist, CarMax, etc. as a reference to help figure out this domain completely. In particular I chose to use the craigslistVehicles.csv file that I found on Kaggle.

When I am looking for a data set it is always one of the more stressful and challenging times of the entire data science project process. This is because I have had so much bad luck when it comes to picking a data set I second guess myself many of times especially when the project is not going as I originally planned for it to go. Never the less I was excited to find this data set because I feel that it fit my hobbies so well. Cars is one of my favorite hobbies in particular I love buying used cars and working on them so that I can resale them in hopes to make a profit.

During my initial look of the data set it appeared to be even better than I originally anticipated it would be. It hosted a lot of data that I thought would make for a lot of good observations and great findings. The craigslistVehicles data set hosts over 677000 entries and 20 columns of data. There was plenty of both categorical and numerical data. While looking at the data and trying to decide what exactly did I want to get out of the data I knew for sure that there were a couple of columns that I would not need so I made it my first order of business to delete those said columns. While do some further inspections I could also see that there was a lot of data that was missing. So I had to handle that missing data in the proper way. Most of it was just filling in the missing values accordingly.

**What are we looking for**

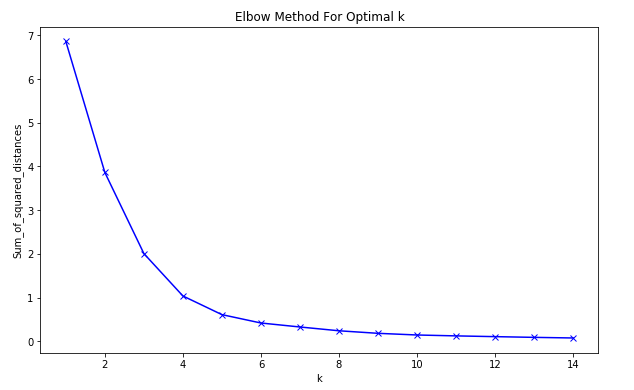
When I was originally looking for a data set I had no specific motives or questions that I was looking to have answered. It was after I had a good look at the data that I started to formulize ideas and questions that I could possibly have answered during the analysis. Some of my initial questions that I thought to look into were:

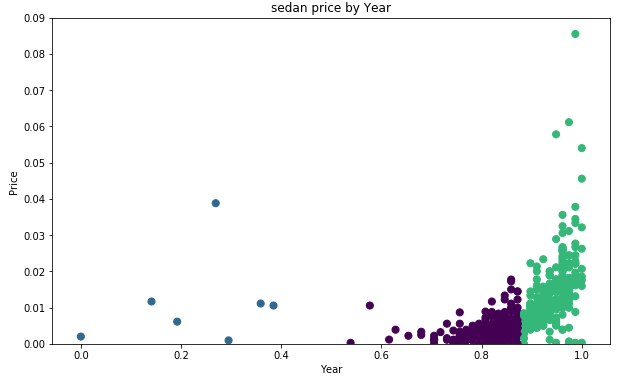
1. What model of vehicles retains the best value?
2. Most available model of vehicle in later years.
3. Least available model of vehicle in later years.
4. How does the mileage defer to say the number of possible seats available on a car such as a 2 door sports car compared to a suv.
5. How does geography affect car choices type purchases?

These initial questions were good questions that I searched for first. Most of those questions led me to answer other questions instead of the originally proposed questions that I had. As I dove deeper into the analysis I was overtaken and wanted to answer questions from the database instead of the questions that I derived without knowledge of what the data was actually telling me. One of the major pieces of data that I was looking at was the differences between sedans and coupes.

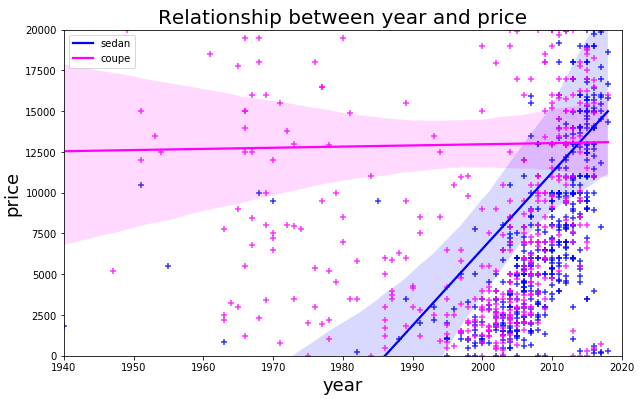
**The methods**

In order to answer some of the questions that I had around the differences in prices of sedans and cars over the years I had to figure out what methods would be needed in order to further this analysis. So far the analysis I used were:

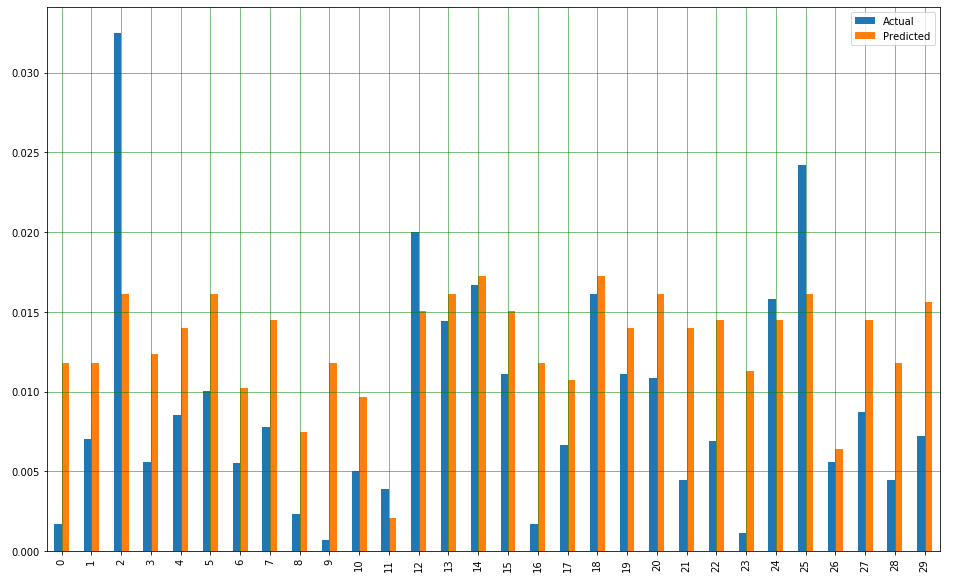
* Hypothesis T-Test in which my null hypothesis is that manual vehicles cost more than the automatic vehicles in this data set. After running the T-Test my P value was 0.022 which is less than my significance level of 0.05. So this means that we are rejecting the null hypothesis which is ultimately telling us that in this data set the manual transmission vehicles do not cost more than the automatic vehicles.
* K means clustering to see the distribution of sedans by price. For this method there was a good deal of data prepping needed in order to get the best results from using k means clustering. For the first order of business I wanted to scale the continuous features to give equal importance to all features. The next step was to find the optimum number of clusters for the clustering algorithm. So I used the elbow method and here is what that graph look like: 

This elbow plot showed me that the number of clusters should be 4. After running the algorithm with the clusters at 4 I graphed it and this is what it came out to look like: 

* Linear Regression to see the relationship in price to year in both sedans and coupes. The relationships between both the sedans and the coupes were positive, but as you can see in the graph below that the coupes relationship barely had a positive slope at all. At the same time the sedans had a very big slope that seemed to continue to rise even past the data that we have.



After analyzing this graph I thought it would be great to run a regression prediction model to predict the year to price on the sedans. I chose this because of the steady incline of the slope and I wanted to see how true that would hold up when predicting what the next set of years to value could be. To do this task I had to first reshape the data and then train the model so that we can do some predicting. Surprisingly the predicted values were pretty close to the actual values as you can see on the graph below.



**Conclusion**

To sum this all up I did enjoy this data set just as much as I hoped I would. There was a lot of good information here, and a lot of analysis that could be performed on this data. There are even areas that I can easily see where my analysis brought up other questions for me that with more time could be answered. I went into this assignment with a set of question on my mind and those are the only questions that I wanted answered. After looking into the data I saw that there were different questions that I would much rather find answers to. I am not sure if this is a common occurrence when dealing with pretty big data sets. This data set was great because it gave me a good refresher on some of those methods that I have not practiced with in a very long time. I enjoyed this assignment and look forward to do more analysis with this data set.

Works Cited

Amelia, A. (2018, September 28). K-Means Clustering: From A to Z. Retrieved from <https://towardsdatascience.com/k-means-clustering-from-a-to-z-f6242a314e9a>

Bronshtein, A. (2019, December 27). Simple and Multiple Linear Regression in Python. Retrieved from https://towardsdatascience.com/simple-and-multiple-linear-regression-in-python-c928425168f9

Brownlee, J. (2019, November 27). 17 Statistical Hypothesis Tests in Python (Cheat Sheet). Retrieved from <https://machinelearningmastery.com/statistical-hypothesis-tests-in-python-cheat-sheet/>

Editor, M. B. (2016, April 20). Understanding t-Tests: t-values and t-distributions. Retrieved from https://blog.minitab.com/blog/adventures-in-statistics-2/understanding-t-tests-t-values-and-t-distributions