Factors in Car Price Quotations

A Capstone Report
presented to the faculty of the
School of Engineering and Applied Science
University of Virginia

by

Serafettin Mert Ocal

May 10, 2023

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

Serafettin Mert Ocal

Capstone advisor: Nada Basit, Advisor Department of Computer Science

The goal of this project is to analyze the factors in car price quotations. There are many different attributes of a car, and all of these factor in when giving a quote. I used Python to clean the dataset and build plots to analyze the data.

A huge dataset was collected and cleaned in order to build plots on how important each attribute was. The data was cleaned by eliminating the unnecessary and redundant attributes. Duplicates were eliminated from the dataset by checking every entry's VIN. As type, model, manufacturer, and year were going to be used in the data analysis part, having null values for those rows was unacceptable, so all entries that had null values for any of those rows were dropped. The cars with super low or high mileage and super low or high price were dropped as they were not believable. Some entries had a price of $0 or mileage such as 123456789. At the end of the cleaning process, the number of entries dropped from 426,880 to 82,923 (%19.4 remaining) and the number of attributes dropped from 26 to 16 (%61.5 remaining).

After cleaning is finished, there's the data analysis part. To get a better understanding of the cleaned dataset, I found the number of models, miles, types, years, and manufacturers. After that, I found miles distribution over all vehicles while also calculating the average odometer. Then,I found the miles distribution over years. After that, I found miles distribution per make. Then, I did the same thing with price instead of miles. Below, I will explain my findings on each plot.

First, we have the Miles Distribution over Number of Vehicles plot. What can be seen from this plot is, the average miles is 91,939. Another point to note is, the number of vehicles with low mileage (0 to 20,000) is very low. Then, it goes up and down until it reaches its highest point. After it reaches the top at around 100,000 miles, it keeps going down as the number of

vehicles decreases as the miles increase.

The next plot is Miles Distribution per Model Year. This plot provides some interesting data. The dataset includes vehicles from 1905 to 2022. The mileage on the newer cars are obviously much less than the older ones. Then, mileage increases until year 2000. At this point, the average is around 150,000. After the year 2000, mileage actually starts going back, causing the graph to look like a parabola. The reason behind this is that after the year 2000, cars start to be classified as classic/antique. The mileage starts going down until around the 1980s down to an average of 70,000 miles. After 1980, we see more of a linear graph since pretty much all cars are antiques before this year. I realized that the odometer reached an average of around 20,000 when it reached 1928. However, I printed out the entries where the years were 1928. There were only 2 listings, and the listings were not accurate. "fort collins" and "akron" were used for 'make' which are both locations. In these big datasets, no matter how much you clean the dataset, there will be a lot of redundant data. Going back and trying to look at specific entries is definitely very helpful during analysis.

The next plot is Miles Distribution per Make. From this plot, we can see how much each make is being used. Commonly seen cars such as Toyota, Honda, Ford have an average mileage of over 100,000 miles while luxurious/sport cars such as Ferrari, Alfa-Romeo, Aston-Martin, andTesla have very low mileage (around 25,000 miles). Another thing we can learn from this plot is reliability of the cars. Cars such as Chevrolet,Ford, Honda, Ram, and Toyota have many vehicles that are closer to 300,000 mile range.

The next plot is Price Distribution over Number of Vehicles plot. What can be seen from this plot is, the average price is 19,245. There are many cars at around 10,000 dollars. And the

number of vehicles constantly decreases as the price increases. The number of vehicles with a price of around 10,000 dollars is so much that it keeps the average price at 19,245 despite the listings with 150,000 dollar prices.

The next plot is Price Distribution per Model Year. This plot is very similar to Miles Distribution per Model Year. The reason behind that is, the new cars have low mileage, so they are expensive. Similarly, the antiques have low mileage, and they are expensive. However, the cars in between (year 1995 to 2010) have high mileage (around 150,000 miles) and low price (around 8,000 dollars).

The last plot in the analysis is Price Distribution per Make. From this plot, we can easily see the expensive cars. The outliers in this dataset are Aston-Martin and Ferrari. Those two have an average price of 80,000 to 100,000 dollars. After those two are Tesla, Porsche, Ram, and Rover With an average price of around 30,000 to 40,000 dollars. The other 35 makes have an average price ranging from 3,000 to 25,000 dollars.

When a customer is trying to purchase a vehicle, they can look at the plots above and get an understanding of how the prices and the mileage ranges for a given make or year. They can read the analysis (or make their own analysis) to see if the car they are considering to buy would be a good buy by comparing their car's value with the average listings. As this dataset is very big, and the data can be redundant as I've explained in Miles Distribution per Model Year, they sometimes may need to check specific listings if they see anything weird in data.