An Analysis of Car Prices using SAS

Nabhan Kazi

Introduction

It feels like only yesterday that I bought my first car with my own money. Even though it has been well over 8 months now, it feels like I was visiting all the dealerships in my town and trying to find a car only yesterday. I was very confused on what kind of car to buy. Advices from people flew in from all over the place. I tried to tackle the problem step by step. I first determined how much money I could spend on the car, and whether I wanted to pay it all up front or finance it. Then I had to decide what type of car I wanted, a car, SUV, truck, etc. After that, I did some research on which makes were considered good and reliable by users. Then, I had to figure out what specifications and features I wanted in the car. For example, I knew that I wanted the transmission to be automatic, because I had no idea how to drive stick. Once I had a rough idea of what specifications I wanted the car to have, it became quite easy for me to find the ideal car. I have heard that some people spend weeks and months looking around before buying a car. Granted their cars are probably a lot more expensive than what I was in the market for. Nevertheless, I noticed that once I had my primary needs figured out and set, it was pretty easy for me to chose or reject a car as soon as I saw the specifications. As a result, I ended up buying my car 3 days after I figured out what specifications I needed.

This whole process of me trying to find a car got me thinking about what specifications people generally look for when buying cars. Almost all households have a car or two. Different people have different needs when buying cars. Good functional cars can range anywhere from $2000 to upwards of $100,000. This huge price range got me wondering what exactly determines how much a car is worth. So, I decided to try to figure out the variables that play a significant role in predicting the price of a car. Thus, in this study I will try to come up with a significant model of variables that helps determine the price range of a car. I will be looking at a website that private owners and dealers use in Canada to sell their cars. The website is autotrader.ca. It is one of the top websites in Canada for buying and selling both new and used cars. I will be gathering data from 300 different cars in Toronto and the GTA. I will collect data for both used and brand-new cars. I did some asking around and came up with the following major specifications that people usually look for when buying a car- location, make, kilometers driven, status, body type, transmission, year, number of cylinders in the engine, colour, whether it comes with car proof, and the type of fuel it intakes.

After some looking around and having bought a car myself recently, I feel like I have gained some insight on how new and old cars are valued based on their specifications. I expect that most, if not all, of my variables will play a role in determining the price of car. First and foremost, I expect that new cars in general will be more expensive than used cars. Secondly, I do not expect the exact same variables to have the same effect on prices of both new and used cars. Thirdly, on used cars, I expect make, km, body type, year and cylinders to have significant effect on the prices of used cars. Finally, I expect the prices of new cars to be affected by make, body type, cylinders and colour.

Methods

I used the website autotrader.ca to create my dataset. I used a 100km radius from where I lived so that I could get cars from Toronto and GTA. Besides that, I did not put any other filters on. When the cars where listed, I made them sort from newest posted date to oldest posted date. I did that to avoid repeating the same car twice in my data set. I also made the dataset every few days, to ensure that I did not input the same car more than once. Moreover, I did not want to sort the cars by price, kilometers, location, or year either. That is because I did not want any of my variables to be biased or have any chronological order. I did not include any cars that were not functional and were being sold just for parts. I also did not add sports cars because even sports cars from 2010 or elder were over $100,000. So, I figured that they would just be outliers in my data and mess up my tests. Many of the cars did not show the number of cylinders in their engines. I had to search those up separately, using the exact make, model, year and trim of the car. At first, I did not think of adding the locations of the cars. however, one day I was talking with my cousin and I was telling him that I think cars in the city are more expensive than the GTA or countryside. But he argued that in the city there is a lot of demand, and that there is not as much demand in the countryside, so cars in the country side probably get sold for more than the city. That really intrigued me to add a location variable by region in my dataset, and see if location plays any role in determining the price of a car. I added variable with the models of the cars as well, but that was mostly for me to keep track of the cars. I cannot really use that variable in my analyses because it has way too many variations from each brand of cars.

To do my analyses and tests, and to produce relevant graphs and outputs, I used SAS Studio. I used the default alpha value of 0.05 for all my tests.

First and foremost, I uploaded the excel file onto the SAS server. Then I imported the datafile and read it in a data-frame called “carsdata”. I displayed the first ten observations to make sure all my data were read in.

I started off by creating a side-by-side boxplot of all the cars separated by their status, to examine my initial suspicion that new cars are generally more expensive than used cars. I set the status of the cars as the independent variable and the retail price as the dependent variable. As expected, the mean price of new cars was higher in price than used cars. in both boxplots, new cars were higher in price range. There were some outliers in used cars, which were due to expensive luxurious cars.

To further confirm my suspicion about the difference in price between the two sets of cars, I ran a one-way ANOVA and Tukey (since significant difference exist). I fount that significant differences existed between the prices of used cars and new cars.

After seeing the results of the above boxplot and ANOVA test, I decided to split up new cars and used cars and conduct tests on them separately. I figured that running tests on them together would skew my results and maybe even make them insignificant. I displayed the first 10 observations of each status to make sure they were split up properly.

I wanted to get an initial idea of how each variable affects the prices, so I made plots for each variable against the prices. I did this for both new cars and used cars. I set each specification of the car as my independent variable and retail price as the dependent variable. Originally, instead of a "cylinders" variable, I had a variable that had every car's engine type. But that did not really help with running any tests and there were way too many variations. I did some research and found out that one of the most important things that matter in an engine are the number of cylinders. So, I changed my engine variable into the current cylinders variable and entered the number of cylinders each car had. All the variables seemed to have some interaction with the prices of the used cars. Some were obvious, others I knew I had to run further tests to confirm. When I made plots for the new cars, I excluded some of the variables because they would not have had any real effect on the prices. For example, the kilometers of the new cars were virtually zero, the years were either 2017 or 2018, and new cars did not need to be car-proofed. Some of the variables had obvious interaction with prices of the new cars, others did not. Either way, I knew I would have to run further tests to be completely sure.

I added a regression line to the scatterplots and then I checked to see the correlation between the independent variables with the dependent variable.

After that, I fitted the relevant variables for new cars and used cars in separate regression models. From there, I checked to see which variable had the least significant interaction with the dependent variable. I took out that variable and ran the regression again. I kept doing this until I took out all the non-significant variables. That left me with only the variables that significantly affected the price of a car. I did this for both new cars and used cars, and came up with a final model for cars of each status.

Results and Statistical Analyses

Code:

```
proc import

        datafile='/home/nabhankazi0/cars.xlsx'

        out=carsdata
        dbms=xlsx
        replace;
        getnames=yes;
proc print data=carsdata (obs=10);
```

Output:

Table 1

Analysis:

All the data and variables seem to have been read in properly. I displayed the first ten observations to verify.

Code:

```
proc sgplot data=carsdata;
        vbox price / category= status;
```

Output:

Boxplot 1

Analysis:

From the side-by-side boxplots, the difference between prices of new cars and used cars is apparent. The mean and median prices of used cars are much lower than the mean and median prices of new cars. Both boxplots are positively skewed and have few outliers. This is because luxury cars, both new and old, are usually higher in price than generic cars.

Code:

```
proc anova data=carsdata;
        class status;
        model price=status;
        means status / tukey;
```

Output:

ANOVA 1

Analysis:

I ran a one-way ANOVA and Tukey's method, to confirm that significant differences in prices do exist in used and new cars. In SAS, to get Tukey, we must run the whole ANOVA again, so I included Tukey in the same code, and would have ignored it if it were not relevant. However, we see from the ANOVA table first, that the F-test was significant with a p-value less than 0.001. This means that significant differences exist. Now, we move on to Tukey's method to determine where the differences exist. The minimum significant difference of 4729.1, says that any means differing by this much or more are significantly different. The difference of means between the new cars and used cars is 42059-19309= 22750, which is more than the minimum significant difference. Thus, Tukey's method helped me confirm that there is definite difference in prices between the two types of cars and that I should split them up and deal with them separately.

Code:

```
data usedcars;
		set carsdata;
		if status= 'Used';
proc print data=usedcars (obs=10);
```

Output:

Table 2

Analysis:

I separated the used cars in a new data frame called “usedcars”. All 117 of the used cars were read in the data frame. I displayed the first 10 observations. As can be seen in the “status” variable, all the cars say “used”.

Code:

```
data newcars;
		set carsdata;
		if status= 'New';
proc print data=newcars (obs=10);
```

Output:

Table 3

Analysis:

I created a new data frame called "newcars" that contains all the new cars. All 83 of them were input properly, I displayed the first 10 observations to verify. All the observations in the "status" variable read "new".

I will have a look at used cars first.

Code:

```
proc sgplot data=usedcars;
        vbox price / category= colour;
```

Output:

Boxplot 2

Analysis:

I created a side-by-side boxplot of the colours of used cars against their prices. The mean and median prices of black cars and white cars seem to be higher than the rest. But, the differences are not too big. All the boxplots are positively skewed. Thus, it looks like colour does not seem to have much effect on the prices of used cars. A regression model will be conducted at the end to confirm this.

Code:

```
proc sgplot data=usedcars;
        vbox price / category= bodytype;
```

Output:

Boxplot 3

Analysis:

I made a side-by-side boxplot of the types of bodies of used cars against their price to see if any significant effect is present. Trucks and SUVS seem to have the highest means and medians than the rest, with coupes following closely behind. It makes sense, since trucks are heavy duty vehicles, SUVS are fancy family friendly vehicles, and coupes fall in the stylish category. Since sedans are usually the most purchased body types, and minivans are not quite the eye catchers, it makes sense that their prices are at the bottom of the pile. Some outliers are visible throughout the plot, which are due to luxurious over-the-top cars. A regression model will be conducted at the end to determine any significance effect of body types on prices.

Code:

```
proc sgplot data=usedcars;
        vbox price / category= transmission;
```

Output:

Boxplot 4

Analysis:

A side-by-side boxplot of transmissions in used cars against price indicates that CVT are usually the more expensive ones. This is because CVT stands for Continuous Variable Transmission, which means that the driver can change between both automatic and manual transmission. This is most effective for racing or rough driving. Thus, the cars that CVTs come in must be high end cars. Automatic and manual transmissions seem to have the same means and medians. Automatic transmission has a higher positive skew and some outliers compared to manual, because automatic transmission is dominating the market right now, especially in cars in recent years and fancier cars. A regression model at the end will confirm whether transmission has any effect on price.

Code:

```
proc sgplot data=usedcars;
        vbox price / category= carproof;
```

Output:

Boxplot 5

Analysis:

I checked to see in side-by-side boxplots whether used cars being carproofed had any effect on the price. From the plot it does seem like, that cars with carproofs are generally more expensive. The boxplot for cars without carproof is positively skewed, and the plot for cars with carproof seem to be evenly distributed. I personally think it might be due to the fact that mostly only dealers provide carproofs and private owners do not, and cars from dealers tend be more expensive as well. I will run it in a regression model to see if having carproof does indeed affect price.

Code:

```
proc sgplot data=usedcars;
        vbox price / category= make;
```

Output:

Boxplot 6

Analysis:

This side-by-side boxplot of the makers of used cars against prices seem to be all over the place. But upon close observation, I can see that brands like Audi, Cadillac and Porsche have the highest means and medians than the rest of the brands. I think that makes sense because these brands are considered extremely posh and over-the-line. Brands like Acura, Hyundai, Chevrolet, Toyota, and Mazda are in the low to middle range, as they are frequently bought and considered the basic brand. BMW, Dodge, Nissan, Mercedes-Benz and Ford seem to range from low to high, as they come can come in rough-usage designs and stylish designs, with their prices varying accordingly. Finally, Honda, with no surprise at all, is at the bottom of the pile, since they are one of the most purchased and gas efficient brands. A regression model later will help confirm my suspicion whether brands affect prices.

Code:

```
roc sgplot data=usedcars;
        vbox price / category= fuel;
```

Output:

Boxplot 7

Analysis:

A side-by-side boxplot of the type of fuels of used cars against prices reveals no surprise at all. Diesel Fuel and Premium Unleaded gas have the highest means and medians by a mile. Heavy duty trucks mostly intake diesel, and high-end cars mostly intake premium gas. So, the peak in their prices is expected. Flex fuel capacity and regular unleaded gas are used by our average vehicles. Regular unleaded gas does seem to be positively skewed with a few outliers that go high. This is because nowadays, even some luxury cars and high-end cars intake regular gas. A regression model shall reveal the truth about gas types and prices later.

Code:

```
proc sgplot data=usedcars;
        vbox price / category= location;
```

Output:

Boxplot 8

Analysis:

A side-by-side boxplot of used cars by location reveals that the prices are pretty much similar in Toronto and the GTA, with a few outliers here and there. Only North York seems to sell cars that are higher in price in average. Though, the difference in mean and median does not seem to be too high compared to the rest of the location. Boxplots of Toronto, York Region and Peel Region are positively skewed since they are very populous regions and sell a broad range of cars, from average to fancy high-end cars. True effect of location on prices shall be determined by a regression model at the end.

Code:

```
proc sgplot data=usedcars;
        vbox price / category= cylinders;
```

Output:

Boxplot 9

Analysis:

I was slightly surprised by the side-by-side boxplot of number of engine-cylinders in used cars. I expected 4-cylinder and 6-cylinder cars to have around similar means and medians, with the 6-cylinder boxplot to be more positively skewed as they are usually present in most new and fancy cars. However, I was quite surprised to see the boxplot of 8-cylinder cars to have such a broad range in prices. I expected it be positively skewed and with a very high price range. But, I was not expecting it to go down at the bottom of the spectrum as well. I know that heavy duty trucks and luxurious fast cars use 8-cylinder engines in their cars, but I was quite lost as to why some 8-cylinder cars cost so cheap. So, I went back to my dataset and looked over it and found out that vehicles which uses 8-cylinder engines and are super cheap, are actually fancy vehicles from the early 2000s. A regression model will help determine whether the number of cylinders affect the price of used cars.

Code:

```
proc sgplot data=usedcars;
        scatter x=km y=price / jitter;
        reg x=km y=price / jitter;
proc corr data=usedcars;
        var km price;
```

Output:

Scatterplot 1, CORR 1

Analysis:

I ran two codes at the same time, first, to see what effect the kilometers of a used car has on price, and then to see the correlation of that effect. Firstly, from the scatterplot, it looks like a decent downward sloped graph. The regression line shows the pattern better. I also added jitter to be able to see overlapping values.

The coefficient correlation of -0.608 reveals that kilometers driven, and the price of a car does have a strong negative correlation. This means that the more kilometers it will have been driven, the lower in value and price it will be, given all other variables equal.

Code:

```
proc sgplot data=usedcars;
        scatter x=year y=price / jitter;
        reg x=year y=price / jitter;
proc corr data=usedcars;
        var year price;
```

Output:

Scatterplot 2, CORR 2

Analysis:

The scatterplot of year against price of used cars suggests a positive slope. The jitter shows overlapping values. The coefficient correlation value of 0.645 tells us that there is a strong positive relationship between the year of a used car and its price. To elaborate, the newer a car will be the more expensive it will also be, if all other variables are controlled.

Now I will have a look at new cars

Code:

```
proc sgplot data=newcars;
        vbox price / category= colour;
```

Output:

Boxplot 10

Analysis:

Colour does not seem to have much effect on the prices of new cars, as seen from the side-by-side boxplot. The means and medians are around the same average range for the new cars. Though black and grey boxplots are positively skewed, due to the fact they are usually the preferred colours. There are a few outliers, which may be due to luxurious cars. A regression model will help clear things out later.

Code:

```
proc sgplot data=newcars;
            vbox price / category= bodytype;
```

Output:

Boxplot 11

Analysis:

A side-by-side boxplot of the type of body of new cars reveals that coupes are the most expensive types. It does makes sense since the stylish and faster cars are designed as coupes. SUVs and trucks are next in line, followed by minivans and sedans. Hatchbacks are at the bottom of the pile this time. Proper effect will be determined by a regression model.

Code:

```
proc sgplot data=newcars;
	vbox price / category= transmission;
```

Output:

Boxplot 12

Analysis:

Since, automatic transmission is the most common in new cars, the boxplot reveals that automatic transmissions in new cars cost the highest and are the most widely ranged. CVT is not as popular in new cars as it was in used cars, so its price is not high either. There are almost no new cars with manual transmission. However, the ones that do exist seem to be outliers. This is because usually only sports type cars that can go very fast have manual transmission in them, causing the hike in price. Proper effect will be determined by a regression model.

Code:

```
proc sgplot data=newcars;
	vbox price / category= make;
```

Output:

Boxplot 13

Analysis:

Right off the bat we see that Cadillacs are the most expensive new cars again, due to their luxury. Buick, Chevrolet, Ram and Ford seem to be in the mid to high price range. The rest are in the low to mid price range, as they are the average everyday-use cars. true effect will be determined in a regression model.

Code:

```
proc sgplot data=newcars;
            vbox price / category= cylinders;
```

Output:

Boxplot 14

Analysis:

There seems to be an upward trend in price of number of cylinders in new cars. The more cylinders the new cars have, the higher its mean and median prices seem to be. This does make sense to me, but whether this relationship is significant or not, will be determined a regression model later.

Code:

```
proc sgplot data=newcars;
            vbox price / category= fuel;
```

Output:

Boxplot 15

Analysis:

Like with used cars, new cars that intake diesel fuel and premium unleaded gas seem to be much more expensive than cars that intake regular gas. The plot for diesel fuel is negatively skewed, while the other two are somewhat evenly spread. True effect will be determined by a regression model.

Code:

```
proc sgplot data=newcars;
        vbox price / category= location;
```

Output:

Boxplot 16

Analysis:

Similar to used car locations, locations on new cars does not seem to have much effect on prices either. The means and modes differ slightly but not by much, which will be analysed by a regression model.

The long-awaited regression models

Code:

```
proc glm data=usedcars;
        class transmission location make bodytype colour carproof fuel;
        model price= transmission location make bodytype colour carproof fuel km year
      cylinders;
```

Output:

GLM 1

Analysis:

This is what I have been leading up to. The moment of truth. This model will help determine whether the variables have any effect on the prices of used cars. From the result of the first model, I see that transmission, location, make, bodytype, colour, carproof and fuel has no effect on the price of a used car. To ensure that this is true, I started taking out the variables one by one, starting from the least significant variable. I did this because sometimes variables affect each other due to multicollinearity. So, if I take them out one by one and keep testing the model, something else that is insignificant now might become significant later.

Code:

```
proc glm data=usedcars;
		class make bodytype;
		model price= make bodytype km year cylinders;
```

Output:

GLM 2

Analysis:

I went through the model and took out the least significant variables from the new model one by one. I took out location first, then colour, then transmission, then catproof, and lastly fuel. I finally ended up with a model where all the variables had p-values that were less than 0.05, which dignifies a significant relationship. Therefore, the variables that effectively regulate the price of a used car are make, bodytype, km, year and cylinders.

Code:

```
proc glm data=newcars;
		class transmission location make bodytype colour fuel;
		model price= transmission location make bodytype colour fuel cylinders;
```

Output:

GLM 3

Analysis:

Same as I did with the used cars, I created a model to analyse which variables significantly effect the prices of new cars. I did not include km, year and carproof in my model to begin with, because they are irrelevant to new cars. So, in my model, I noticed that transmission, location and colour are non-significant variables. So, I start taking them out one by one, starting with the least significant, until I reach a model where all variables are significant.

Code:

```
proc glm data=newcars;
          class make bodytype fuel;
          model price= make bodytype fuel cylinders;
```

Output:

GLM 4

Analysis:

After taking the non-significant variables out one by one, this is what I came up with. These four variables (make, bodytype, fuel and cylinders) have p-values than are less than 0.05, meaning that they significantly affect the prices of new cars.

Conclusion

To sum it all up, I tackled the question of what exactly determines the price of a car. Though it is not a one-dimensional question that if we know a model we can affectively predict the price of a car. But, it does give us some insight into what to look for and focus on when buying a car. We can evaluate what variables are causing the price of a car to fluctuate and decide whether those variables are of importance to us. I started by comparing the prices of new cars against used cars, where I was able to confirm my first hypothesis that new cars will be significantly more expensive than used cars. Then I split up new and used cars and treated them separately. I used boxplots and scatterplots to check which variables affected the prices of the new and used cars. This helped me confirm my second suspicion that different sets of variables will have significant effects on new cars and used cars. Then I tested a model to determine the significant variables of used cars. I confirmed my hypothesis that make of a car, the body type, kilometers driven, year of the car and the number of cylinders in the engine significantly determine the price of a used car. I did the same thing for new cars, and to my surprise, came up with a model that I had not anticipated. I found out that the price of a new car is significantly determined by the make, body design, number of cylinders in the engine and the colour of the car.

These two models will help determine what to look out for when planning to buy car. There are other internal and fancy features involved as well, but those are usually secondary to these basic variables that I used in my analysis.