Hyperiondev

TASK

Exploratory Data Analysis on the Automobile Data Set

Visit our website

Introduction

The dataset in question was regarding the automobile industry. Production, engine size, miles per gallon, insurance risk etc. I was curious as to what the data would return, would it confirm thoughts I've previously had or challenge my beliefs. Regardless I was highly motivated to find out the correlations between certain variables I will discuss below.

DATA CLEANING

Upon my investigation, it was found that a substantial amount of data was in the wrong format. Which would make it difficult further along the line to visualise. Every column was carefully looked into. Column by column I would convert the necessary values to numeric values. What I found odd was 2 values from horsepower were above 10 000. Which did not seem right so I removed those values from the dataset.

MISSING DATA

Firstly, my method for dealing with this particular aspect is to get a general overlook of the missing data. It all returned zero but that can be misleading. Even data filled in with "NaN", "?" and "0" can register as a filled value. So I dug a bit deeper to find exactly that. When encountering data of that source, I would measure how much of that data was missing. I would then locate the exact rows they were in. Once targeted I would calculate the mean.

I started with the normalized-losses, I felt if I could save data I would. Instead of deleting entire columns I would delete rows later if necessary. The mean was calculated for the missing data and I would continue onto the next. The price column was not so significant only 4 out of 205 were missing. I identified the missing data, filled them with the average then continued.

"num-of-doors" did not have much missing values and were captured as strings. Nothing much that you can do about that. Thankfully there was not many values missing in that column so those rows were disposed of.

DATA STORIES AND VISUALIZATIONS

When the data was finally cleaned, I proceeded to visualise. First I wanted to know how many different models did the top 10 car makers produce. It was abundantly obvious Toyota won by a

landslide. The closest competitor Nissan just about half. What was also surprising was that the top 6 were Japanese car makers. Even more surprising was that no American car makers made it to the top 10 at all! Quite a startling fact considering the first automobile was made in America the Ford Model T and Detroit was named “Motor City” because of the abundance of car manufacturing.

The symbolling column refers to a cars safety negative values such as -2 are good whereas values like 3 are not. A general overlook at the graph brings a clear indication that more cars are risky. Most car manufacturers scored 0.

The next graph indicates the distribution of wheel drives between all cars made. The predominant was front wheel drive with 58%, rear wheel drive coming in second with 37% and lastly 4-wheel drive coming in with 4%. Upon further research, each type of drive has purpose. Front wheel drive is cheaper to produce and will continue being the most due to the fact that most electric and green cars (which are all the rage right now) are front wheel drive. The more you can appeal to your consumers pocket the better. Back wheel drive cars are performance cars which are much better suited to have that set up. Why aren’t 4-wheel drive systems put in performance cars? They add weight, which is not the best thing considering speed is what you are going for when manufacturing a performance vehicle. Then lastly 4-wheel drive is specialist and serves a purpose off road and such. Most people don’t go off road and more roads are being tarred everyday especially in third world countries.

My next presentation is another pie graph differentiating between 2 variables. Turbo and normally aspirated. It was abundantly clear with a whopping 82% that normally aspirated was more popular, with turbo only having 18% of the market. It’s for good reason though. Adoption of the turbo is directly linked to price, but as with all things as time goes along the technology will get cheaper and the turbo will become the new standard. This exact same pie graph 30 years ago would look different and it would be between fuel injected and carburettor cars as time got along, the technology used in fuel injection became better and cheaper to produce. Hence the reason why you will not find a single car today that is produced with this method of fuel induction. Turbos were actually first used in the aerospace field due to the altitude of air engines would fail to get the power they needed. So, the turbine was invented to combat this problem. Air was not just blown in. It was sucked in, compressed and forced into the engine. The amount of air forced into the engine would increase the combustion and would in turn increase power. As I’ve said, it is an advanced technology that wasn’t even meant for cars. When able to be produced cheaper it will be wide spread. In fact VW has already got the ball rolling with the Polo TSI. Soon others will follow.

I produced a scatter plot to represent each value individually in regard to the next subject. I wanted to know the peak RPM in correlation to the engine size. As expected cars with bigger engines did not have a high RPM. This is due to the fact that cars with bigger engines don’t have to work as hard as smaller engines do to achieve the same results.

Similar to the chart previous I was curious to know the correlation between engine size and horse power. Not surprising cars with bigger engines performed better. Such engines would be v8's and v12's which have 8 or 12 cylinders respectively. What I did find interesting was the fact that Jaguar was the leader in this aspect and not Mercedes-Benz. Jaguar does not have the reputation for performance BMW does. I guess it's just a matter of effective marketing.

The second to last bar graph is a simple question that needed to be asked. I wanted to know how many models were produced with 2 doors and 4 doors. I didn't expect the distribution between the 2 categories to be so close.

Last but not least. I wanted to find out the economy of vehicles in highway and city conditions relative to curb weight. Curb weight is the weight of a car without occupants or baggage. City mileage was much less than highway due to the stop and start nature of the area. More petrol is used in that environment. Also weight played factor the more weight the less miles per gallon you would get.

The basic take away from the EDA are as follows:

- Toyota produces the most models of vehicle
- The top 6 model variant produces are from Japan
- Most cars have a 0-risk rating
- There are more cars with a positive than a negative risk rating
- Car producers make more front wheel drive cars
- Normally aspirated cars dominate the market
- Bigger engines need less RPM
- Jaguar is the highest performing car out of the major car manufacturers
- More 4 door cars are produced than 2 door
- City miles per gallon are lower than highway miles per gallon
- The more horsepower a car has the less mileage it will have

THIS REPORT WAS WRITTEN BY: Evan Taylor