Cleaning and Exploring Automobile Data

Hoang Huynh

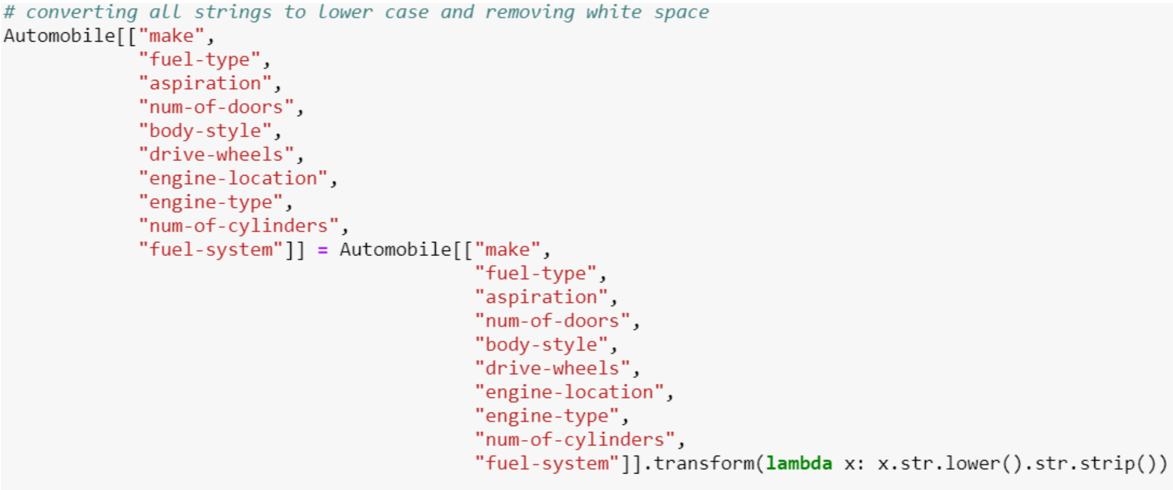
Data Preparation

Loadng the data

A copy-edit version of the automobile data from the UCI Machine Learning Repository (Archive.ics.uci.edu, 2019) was loaded. A description of the 26 attributes can be found at http://archive.ics.uci.edu/ml/datasets/Automobile?ref=datanews.io.

Cleaning the data

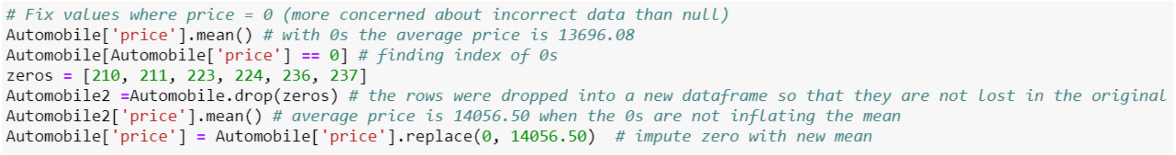
The values and variables were inspected with the original dataset. Once it was confirmed that the data was the same, sanity checks, outliers, extra white space and missing values were checked for using functions such as describe and info. To ensure that the data was valid, the data was inspected with the initial copy from the UCI Machine Learning Repository (Archive.ics.uci.edu, 2019). Judgement on incorrect data was determined by the values provided in that reference. The data was then cleaned to fit only the recommended labels. As shown below, a mass cleaning of extra whitespace and case sensitivity was undertaken on string values. Every other cleaning process was on an individual basis.



```
# converting all strings to lower case and removing white space
Automobile[["make",
            "fuel-type",
            "aspiration",
            "num-of-doors",
            "body-style",
            "drive-wheels",
            "engine-location",
            "engine-type",
            "num-of-cylinders",
            "fuel-system"]] = Automobile[["make",
                                          "fuel-type",
                                          "aspiration",
                                          "num-of-doors",
                                          "body-style",
                                          "drive-wheels",
                                          "engine-location",
                                          "engine-type",
                                          "num-of-cylinders",
                                          "fuel-system"]].transform(lambda x: x.str.lower().str.strip())
```

Despite there being many null values in normalized-losses, they were not problematic due to the lack of machine learnng and statistical inference techniques used as that was out of the scope of this project. Furthermore it was believed that quick solutions such as imputing the mean on so many missing values would be erroneous and do more harm to the authenticity of the data. Thus, missing values were taken into consideration but left untouched. The same thought porcess was made for outliers. As long as the values were in accordance with the UCI reference, it is believed that the data is real and would be more insightful to the understanding of the variables if they existed in the data especially when it comes to exploratory analysis.

What did prove to be a nuisance though were the incorrect data as they could skew summary statistics and provide inaccurate visualisations. One instance in which this was bothersome was in the price attribute, with six automobile prices listed at zero dollars. This severe error that had the potential to skew the average price of automobiles, was corrected for after being imputed with the mean price with the exception of the existing zero dollar values. By doing so the true average of all the cars could be calculated at $14056 dollars instead of the $13696 when the incorrect values were not accounted for.



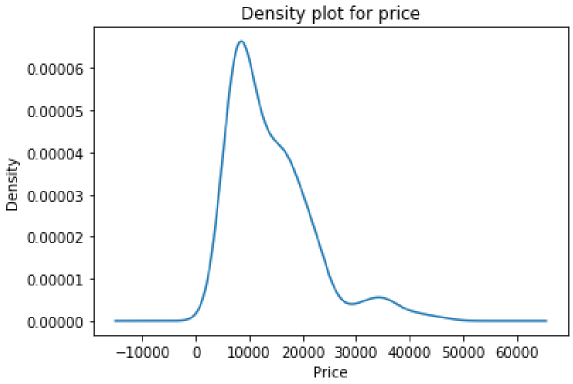
```
# Fix values where price = 0 (more concerned about incorrect data than null)
Automobile['price'].mean() # with 0s the average price is 13696.08
Automobile[Automobile['price'] == 0] # finding index of 0s
zeros = [210, 211, 223, 224, 236, 237]
Automobile2 =Automobile.drop(zeros) # the rows were dropped into a new dataframe so that they are not lost in the original
Automobile2['price'].mean() # average price is 14056.50 when the 0s are not inflating the mean
Automobile['price'] = Automobile['price'].replace(0, 14056.50)  # impute zero with new mean
```

Data Exploration

Subsection 1: Explore variables of importance

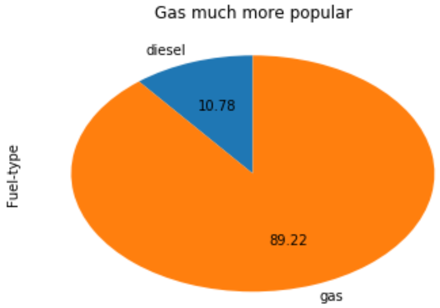
There were many variables of potential interest in the dataset. Three variable types were explored in price (continuous), fuel-type (nominal) and symboling (ordinal). A graph of each attribute was created to display the characteristics of the data.

Price



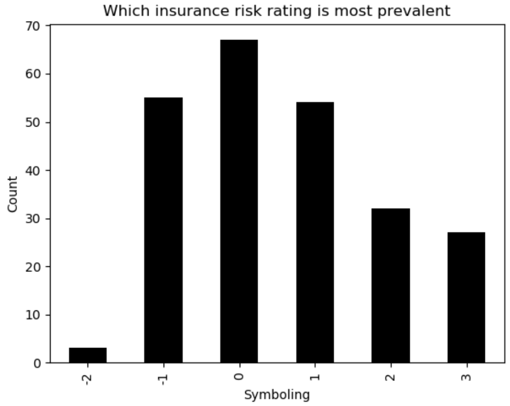
Density plots and histograms are good for showcasing the distribution of continuous data. The density plot is able to better illustrate the distribution shape through its feature to smooth out the noise by not having to produce bins like a histogram (Datavizcatalogue.com, 2019). Price can be seen being right skewed with there being a steep decline from the $15000 mark. This pattern is typical of dollar values such as income wage and can be corrected using log-transoformation. The price of automobiles seems fairly low itself but the cars from 1985 are drastically different in price, power and design to what they are today.

Fuel-type



Pie charts are notorious in the data visualisation community as a terrible graphic type. Despite its limitations, when used correctly such as starting the pie chart at the top it can become a handy tool to communicate categorical proportions on three or fewer slices (Olson, 2016) . Clearly, gas was the preferred fuel-type in this period in time by an outstanding margin. Even without knowing the year of the collected dataset, a very telling clue that the data is a little dated is the fact that there are no electric cars present in the pie chart. This says alot about the data and the state of society during this time. Because there were no electric cars being bought and sold, at least commercially, it is apparent that the environmental impact and technology for electric cars were not as evident and in the spotlight as they are in todays world.

Symboling



Alternatively a bar chart, one of the most prominent plots in the business sector to compare groups can be used to display categorical information. Symboling, which reflects a cars insurance risk rating is displayed by the number of occurence in that category. Not many cars have a very safe risk rating. The reason as of yet is not known but it would be to the benefit of an insurance company to increase the risk rating of cars. By delving into the data deeper, it would be possible to disocver certain automobile features that would be characterised as high risk or safe by an insurance company.

Subsection 2: Exploring the variables between columns

Hypothesis 1 – It is hypothesised that horsepower is positively associated with price as a high maxiumum horsepower requires greater specs such as material and engine, therefore justifying higher prices.

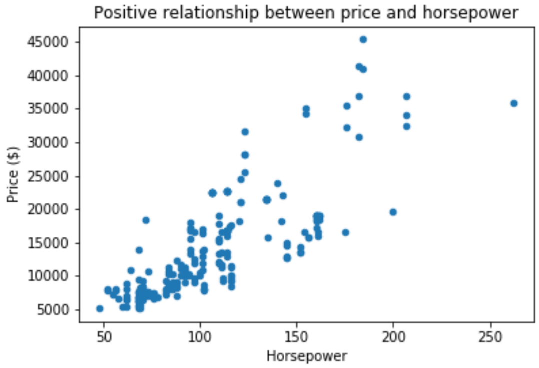


Figure 1: The scatterplot does reflect a correlation between price and horsepower as hypothesised

The pattern between horsepower and price appears to be moderately strong but as horsepower increased, the variance in the pricing range grew with it.

Hypothesis 2 – It is hypothesised that higher normalised losses (the degree to which the auto is more risky than its price indicates) create higher insurance risk (symboling). Based on the understanding of these two metrics, it would make sense for a car that is expected to incrue a big drop off in price due to economic coniditions and depreciation to be at a high risk.

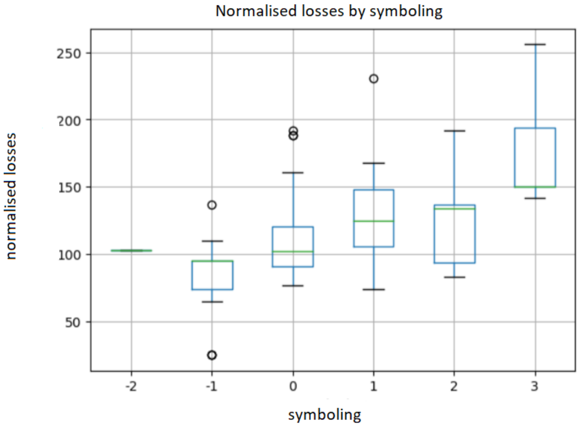
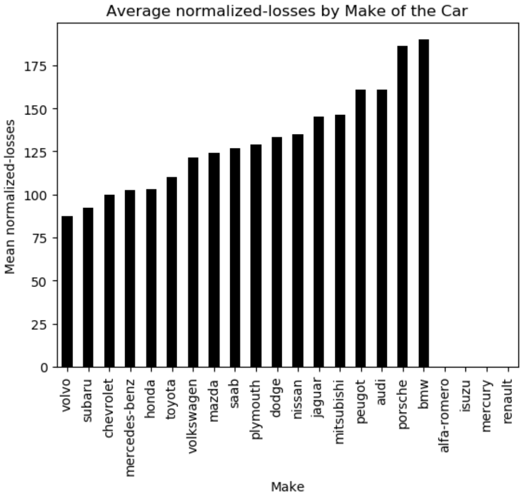


Figure 2: Higher normalised losses is shown in automobiles that have a high insurance risk rating and vise versa

The hypothesis was supported because the automobiles with the greatest insurance risk were ones that had the highest normalised losses. The difference between normalised losses and symboling had a drastic jump between the 2 and 3 insurance risk rating.

Hypothesis 3 − Based on the previous hypothesis and finding, a closer look at normalised losses may be warranted. It would fit the narrative and peoples perception of the industry if luxury European cars incur the highest normalised losses. They have an unreliable track record (Consumerreports.org, 2011), despite costing more than Japanese cars, which may well have persisted from their inception. Therefore, it is hypothesised that automobile brands that sell high-priced cars such as European luxury cars are more prone to higher normalised losses. This is because the expensive costs along with the decreased production of these cars means that the infrastructure for them are not as stable, with there being less skilled labour and parts to help maintain them.



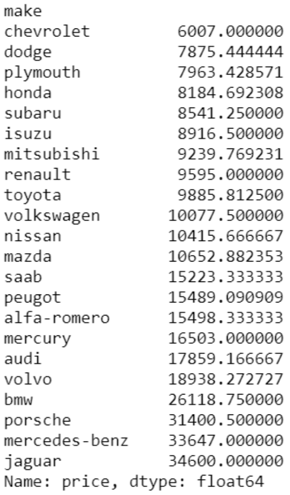
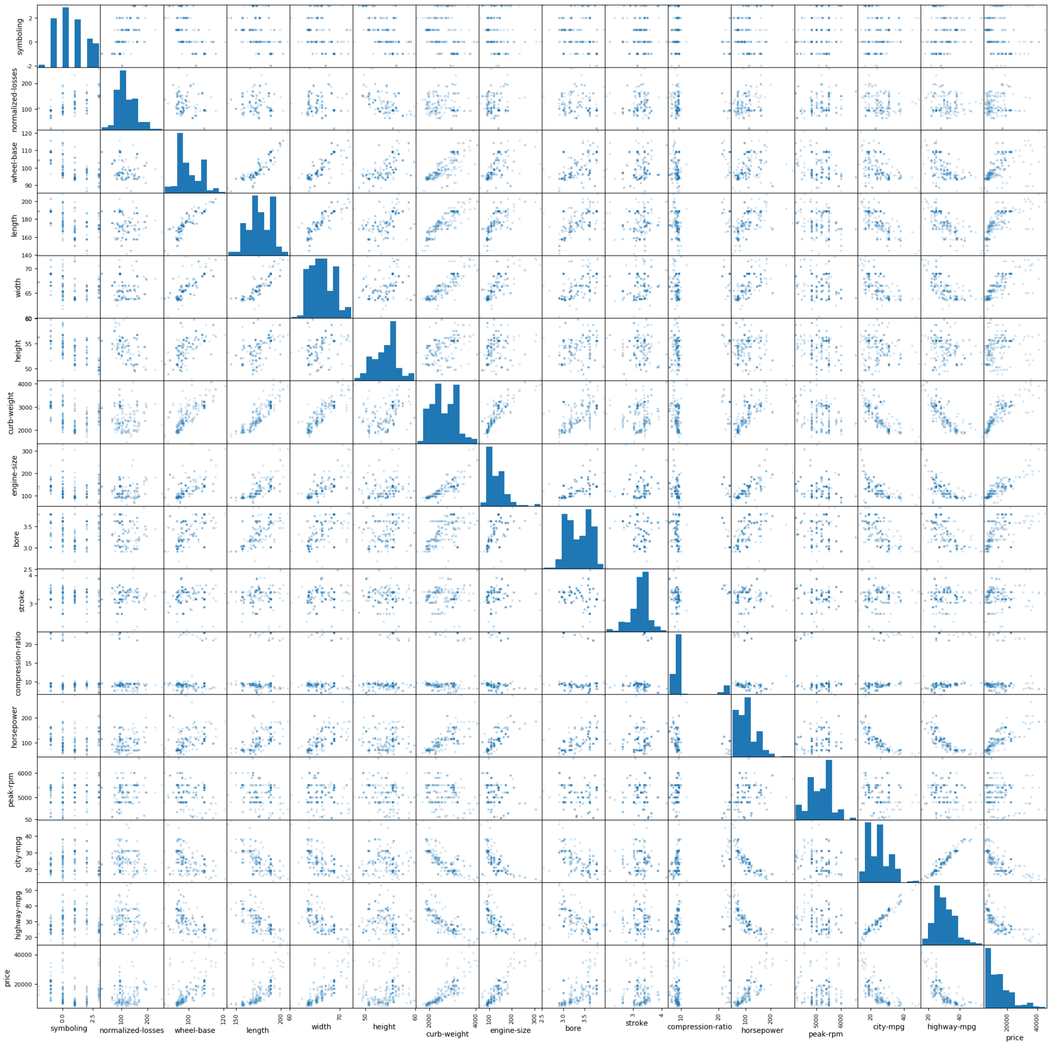

```
make
chevrolet         6007.000000
dodge             7875.444444
plymouth          7963.428571
honda             8184.692308
subaru            8541.250000
isuzu             8916.500000
mitsubishi        9239.769231
renault           9595.000000
toyota            9885.812500
volkswagen       10077.500000
nissan           10415.666667
mazda            10652.882353
saab             15223.333333
peugot           15489.090909
alfa-romero      15498.333333
mercury          16503.000000
audi             17859.166667
volvo            18938.272727
bmw              26118.750000
porsche          31400.500000
mercedes-benz    33647.000000
jaguar           34600.000000
Name: price, dtype: float64
```

Figure 3: The table on the right shows a disctintion in price between Asian and European automobile companies

It was found that European brands show higher normalised losses compared to their counterpart. Despite Volvo being a Swedish automobile company, their cars are not marketed and regarded as 'luxury'. Instead "For most of Volvo's 90-year history, the Swedish automaker offered its loyal legions of customers well-built, safe, and practical transportation" (Zhang, 2017) very much describing the automobile characteristics of the third hypothesis for low normalised losses (great availibility and low cost). What was not in line with the hypothesis however was that Volvo displayed the lowest normalised losses while still being one of the more expensive car brands as shown on the price chart with the corresponding make. The explanation for this could be the lack of globalisation at this time, which meant that cars were not being exported therefore allowng Volvo to dominate the European market as the 'cheap and practical car' for their region. Nonetheless they remain an anomaly because the other brands with low normalised losses were all relatively cheap.

<u>Subsection 3: Build a scatter matrix for all numerical columns.</u>

A scatter matrix is a useful plot that can provide a good summarisation of the relationship between numerical variables. You can see variables which have the potential to relate to one another and use this as a guide for further analyses. An interesting pattern is illustrated through the association between horsepower and city-mpg which depict a very strong relationship. This can be explored and is already useful knowledge for when it comes to linear regression as strong relationships (multicollinearity) can be problematic when it comes to fitting a model due to them both explaining similar things.



References

Consumerreports.org. (2011). Consumer Reports' 2011 Annual Car Reliability Survey: Ford drops, Chrysler rises, Scion leads. [online] Available at: https://www.consumerreports.org/cro/news/2011/10/consumer-reports-2011-annual-car-reliability-survey-ford-drops-chrysler-rises-scion-leads/index.htm

Datavizcatalogue.com. Density Plot. [online] Available at: https://datavizcatalogue.com/methods/density_plot.html

Olson, R. (2016). The correct way to use pie charts. [online] Dr. Randal S. Olson. Available at: http://www.randalolson.com/2016/03/24/the-correct-way-to-use-pie-charts/

Archive.ics.uci.edu. UCI Machine Learning Repository: Automobile Data Set. [online] Available at: http://archive.ics.uci.edu/ml/datasets/Automobile?ref=datanews.io

Zhang, B. (2017). The story of Volvo's incredible transformation into a true luxury brand. [online] Business Insider Australia. Available at: https://www.businessinsider.com.au/volvos-s90-luxury-brand-2017-5?r=US&IR=T