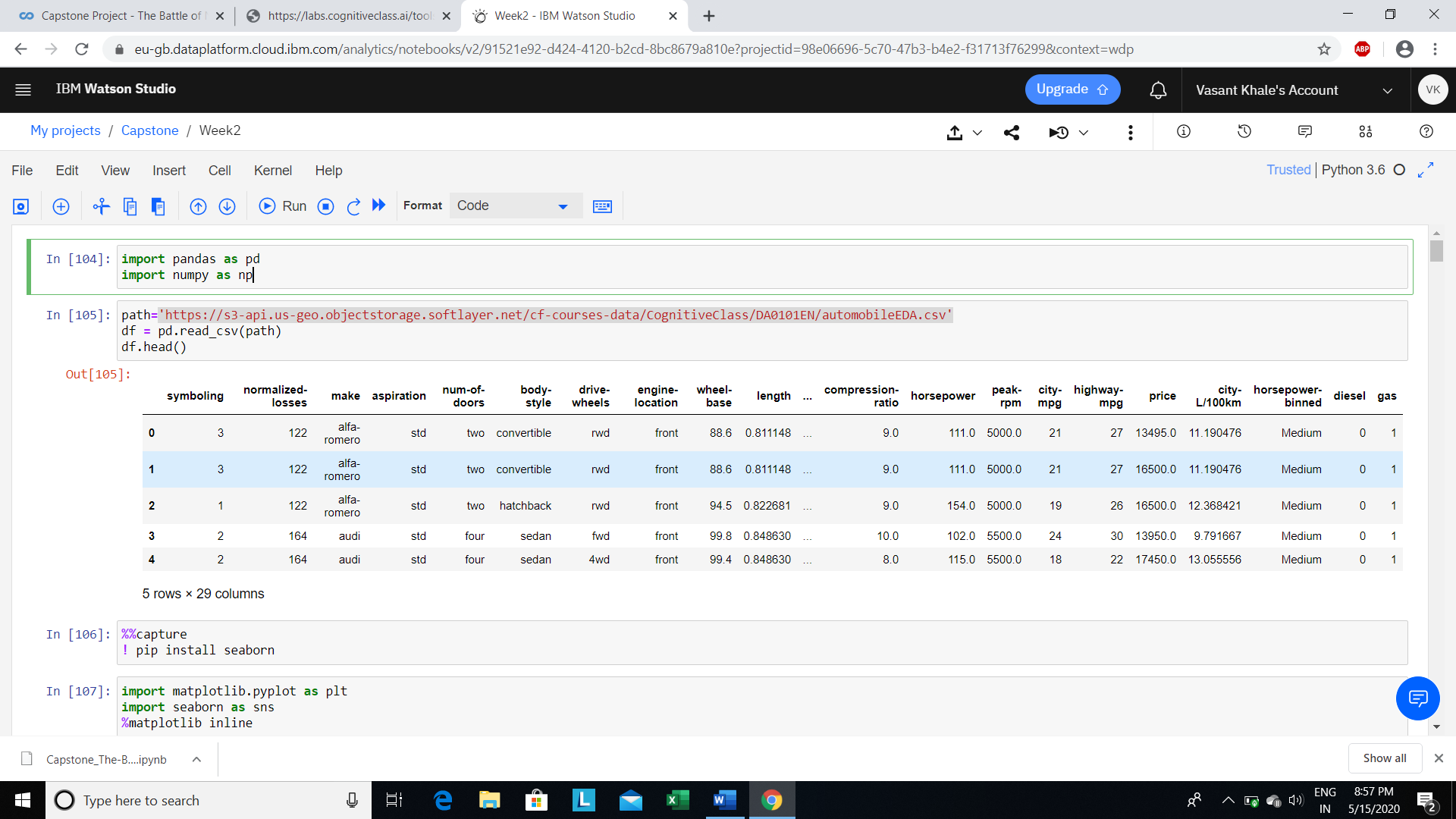
Report

1. Introduction where you discuss the business problem and who would be interested in this project.

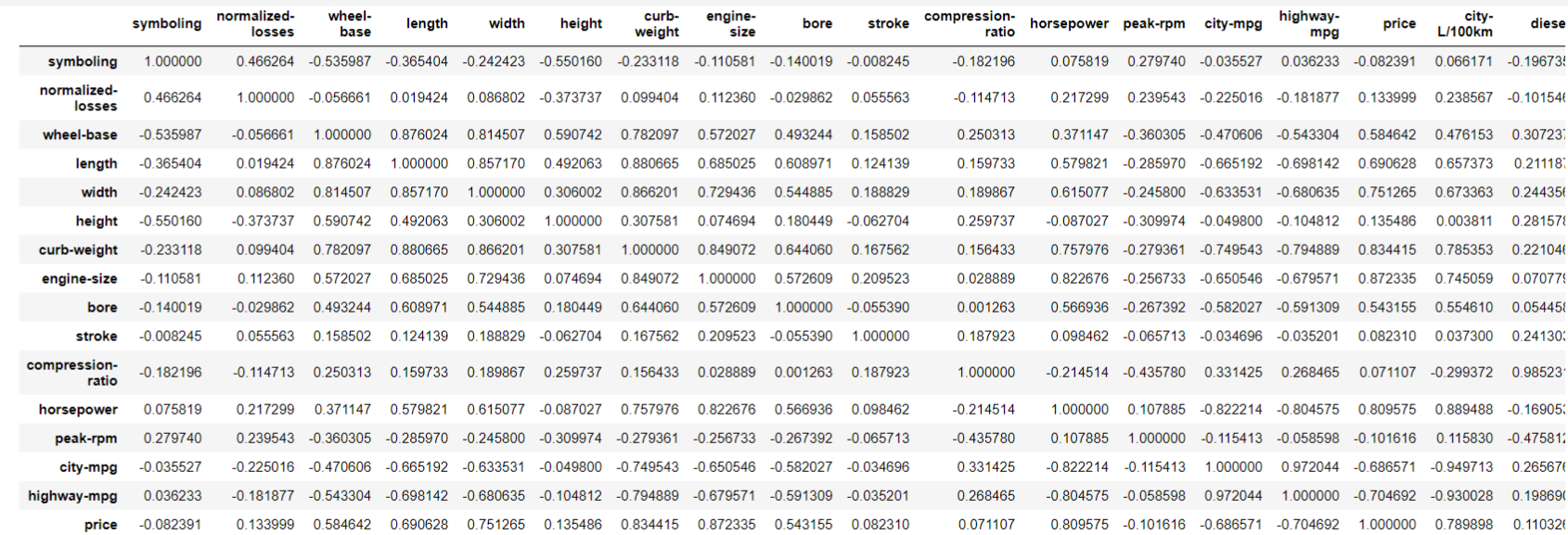
The business problem is problem regarding the purchase of a right car. Many people purchase the car and then they realise that the car is not serving the purpose for which it was meant to. Because of which they sell their car and purchase a new one which costs a heavy investment. If they are able to take the right decision after consulting the right people their money will reap benefits and that car will become useful for them. Whether the car purchased is utility vehicle or luxury vehicle its purpose should be server. Hence, I have decided to start my own consultation firm (hypothetically) to help the common man.

1. Data where you describe the data that will be used to solve the problem and the source of the data.

The data has been taken from 'https://s3-api.us-geo.objectstorage.softlayer.net/cf-courses-data/CognitiveClass/DA0101EN/automobileEDA.csv'. This csv file consist of the attributes that are to be considered while purchasing the car. Some of them have been mentioned for the sake of understanding the data. Length, width, horse power, wheel base, engine size, bore etc.



1. Methodology section which represents the main component of the report where you discuss and describe any exploratory data analysis that you did, any inferential statistical testing that you performed, if any, and what machine learnings were used and why.



**Correlation**: a measure of the extent of interdependence between variables.

**Causation**: the relationship between cause and effect between two variables.

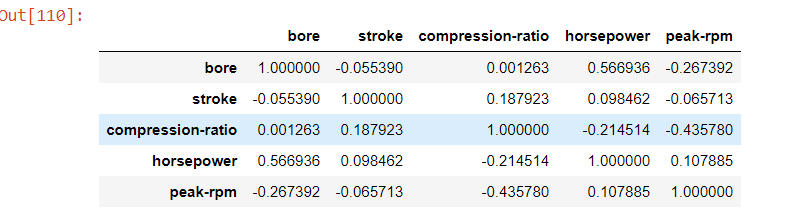
It is important to know the difference between these two and that correlation does not imply causation. Determining correlation is much simpler the determining causation as causation may require independent experimentation.

Pearson Correlation

The Pearson Correlation measures the linear dependence between two variables X and Y.

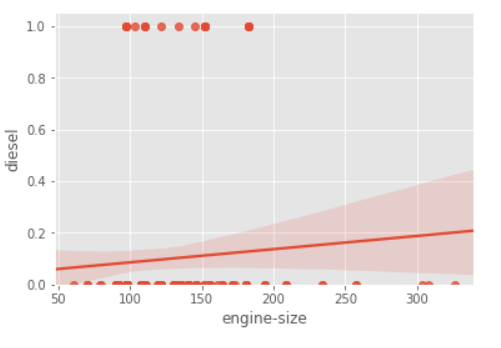
The resulting coefficient is a value between -1 and 1 inclusive, where:

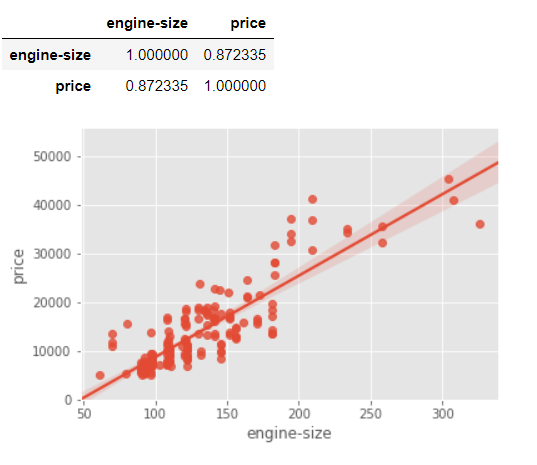
* **1**: Total positive linear correlation.
* **0**: No linear correlation, the two variables most likely do not affect each other.
* **-1**: Total negative linear correlation.

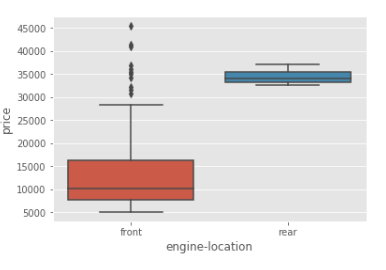


Correlation shows the relationship between the two factors which are affecting one another. Strong Correlations is always near 1 and strong negative correlation is always near -1 and weak correlation is near 0.

There is positive coefficient between diesel and engine size. Which means that as the engine size increases the diesel consumption increases but the rate of increase in the diesel consumption is very less.



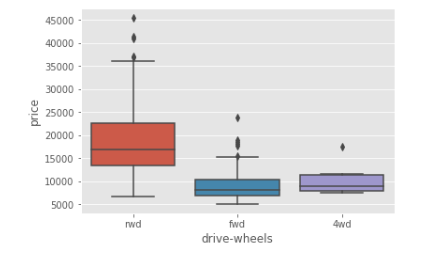


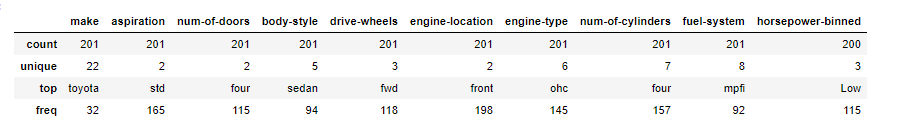


These are variables that describe a 'characteristic' of a data unit, and are selected from a small group of categories. A good way to visualize categorical variables is by using boxplots.

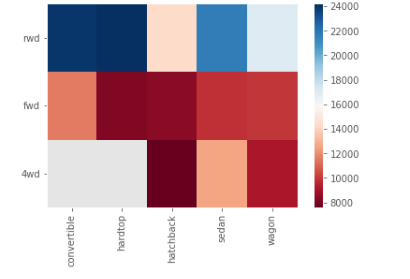
The above relationship between "price" and "engine location" in “front” and “rear”.

The above relationship between "price" and "driven wheels" in “rwd”,“fwd”, “4wd”.

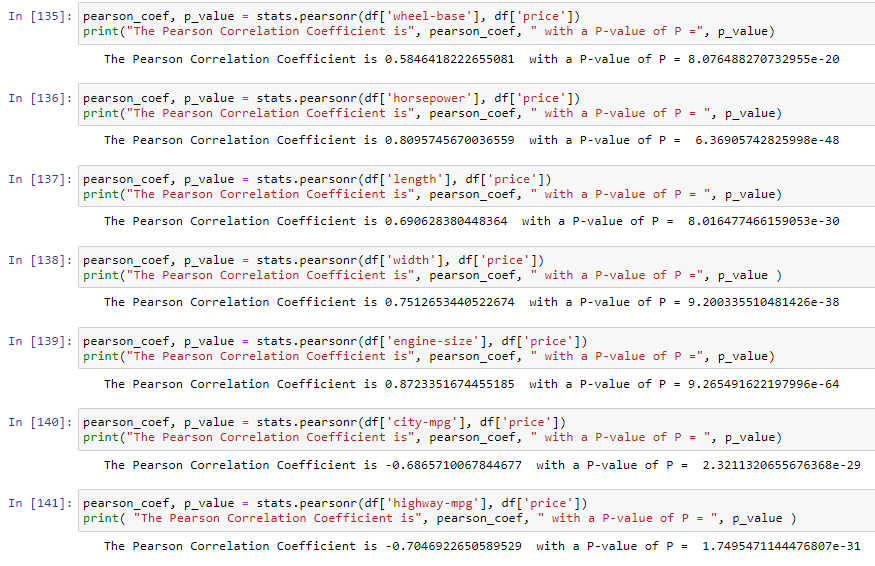




The heatmap plots the target variable (price) proportional to colour with respect to the variables 'drive-wheel' and 'body-style' in the vertical and horizontal axis respectively. This allows us to visualize how the price is related to 'drive-wheel' and 'body-style'.



1. Results section where you discuss the results.





The P-value is the probability value that the correlation between these two variables is statistically significant. Normally, we choose a significance level of 0.05, which means that we are 95% confident that the correlation between the variables is significant.

By convention, when the

* p-value is << 0.001: we say there is strong evidence that the correlation is significant.
* the p-value is << 0.05: there is moderate evidence that the correlation is significant.
* the p-value is << 0.1: there is weak evidence that the correlation is significant.
* the p-value is >> 0.1: there is no evidence that the correlation is significant.
* The Analysis of Variance (ANOVA) is a statistical method used to test whether there are significant differences between the means of two or more groups. ANOVA returns two parameters:
* **F-test score**: ANOVA assumes the means of all groups are the same, calculates how much the actual means deviate from the assumption, and reports it as the F-test score. A larger score means there is a larger difference between the means.
* **P-value**: P-value tells how statistically significant is our calculated score value.
* If our price variable is strongly correlated with the variable we are analyzing, expect ANOVA to return a sizeable F-test score and a small p-value.

1. Discussion section where you discuss any observations you noted and any recommendations you can make based on the results.

Nothing as such.

1. Conclusion section where you conclude the report.

We now have a better idea of what our data looks like and which variables are important to take into account when predicting the car price. We have narrowed it down to the following variables:

Continuous numerical variables:

* Length
* Width
* Curb-weight
* Engine-size
* Horsepower
* City-mpg
* Highway-mpg
* Wheel-base
* Bore

Categorical variables:

* Drive-wheels