

**TASK**

**Exploratory Data Analysis on the Automobile Data Set**

[](http://www.hyperiondev.com/portal/)

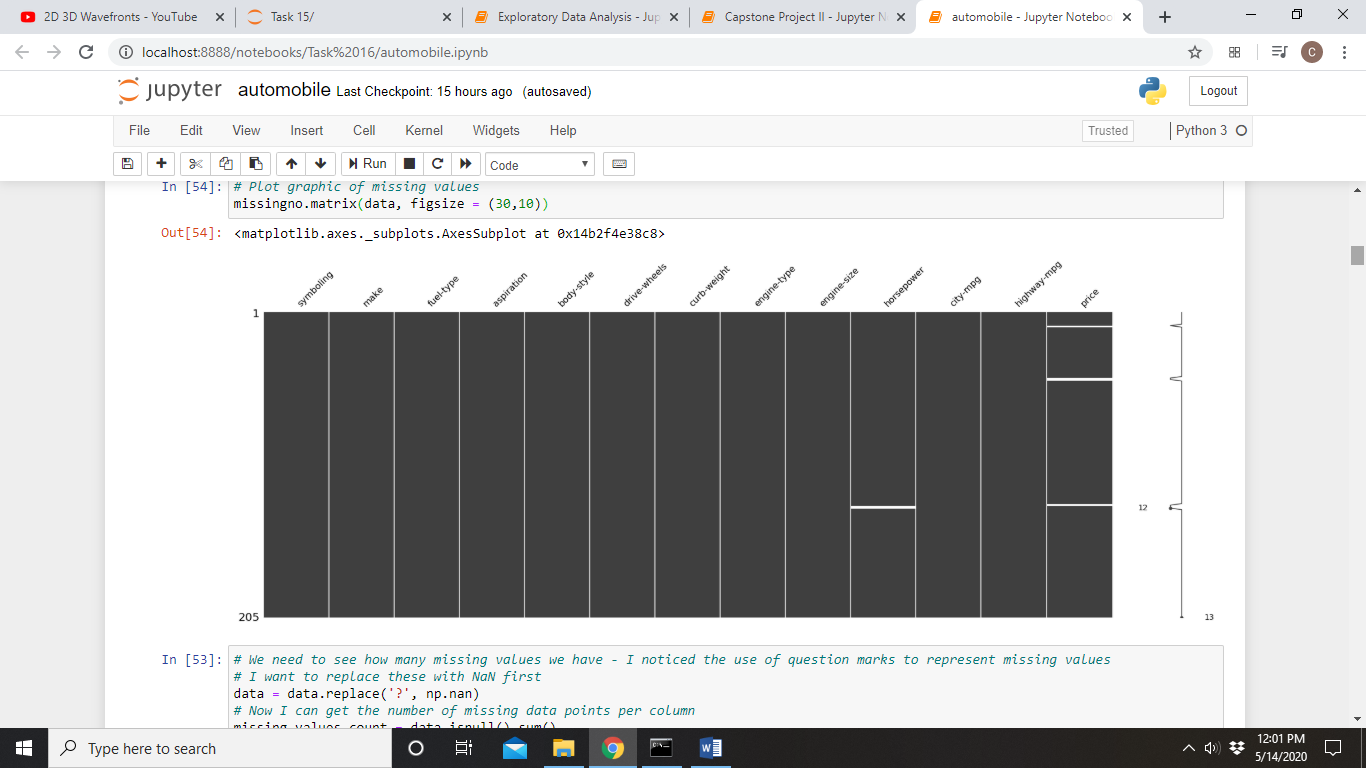
**Introduction**

The data set used for this investigation contains information on different cars including prices and numerous other specifications such as engine size, horsepower and average mileage (in miles per gallon). It contains 205 rows and 26 columns. The dataset includes quantitative, continuous data (for example price and horsepower) and qualitative, categorical data (for example make and body style).

The distribution of the price data and engine size is skewed to the right as the majority of cars are on the lower end of the spectrum for price and engine size. I was mainly interested in how different factors would influence the price of a car, so I will give a statistical summary of the price field. Prices ranged from $5 118 to $45 400 and the average price was $13 207.13. Some of the factors that were found to influence car price include the engine size, the curb weight and the make of the car.

**DATA CLEANING**

The data cleaning consisted of removing the unnecessary columns from the data frame, removing duplicate rows, identifying missing values, and making sure that the data type for each column was correct. In total, I removed 13 unnecessary columns using the pandas drop function and checked how many missing values I had. I was able to use the visualization below to locate the missing values. Lastly, I changed price and horsepower from object data type to float data type.

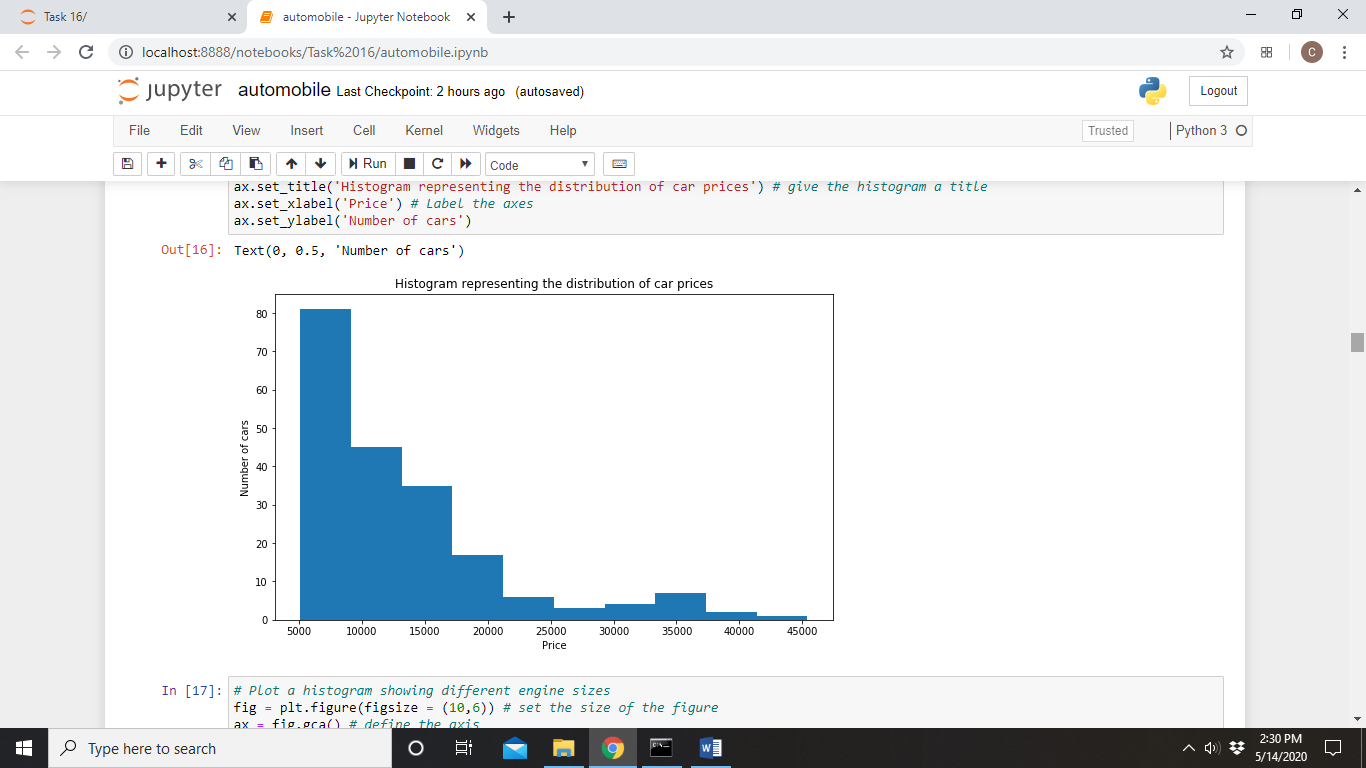


MISSING DATA

After cleaning the data frame, there was still missing data under the price and horsepower columns. For rows where there was no price, I removed the entire row. For rows where there was no value for horsepower, I filled the value with the average horsepower across the dataset.

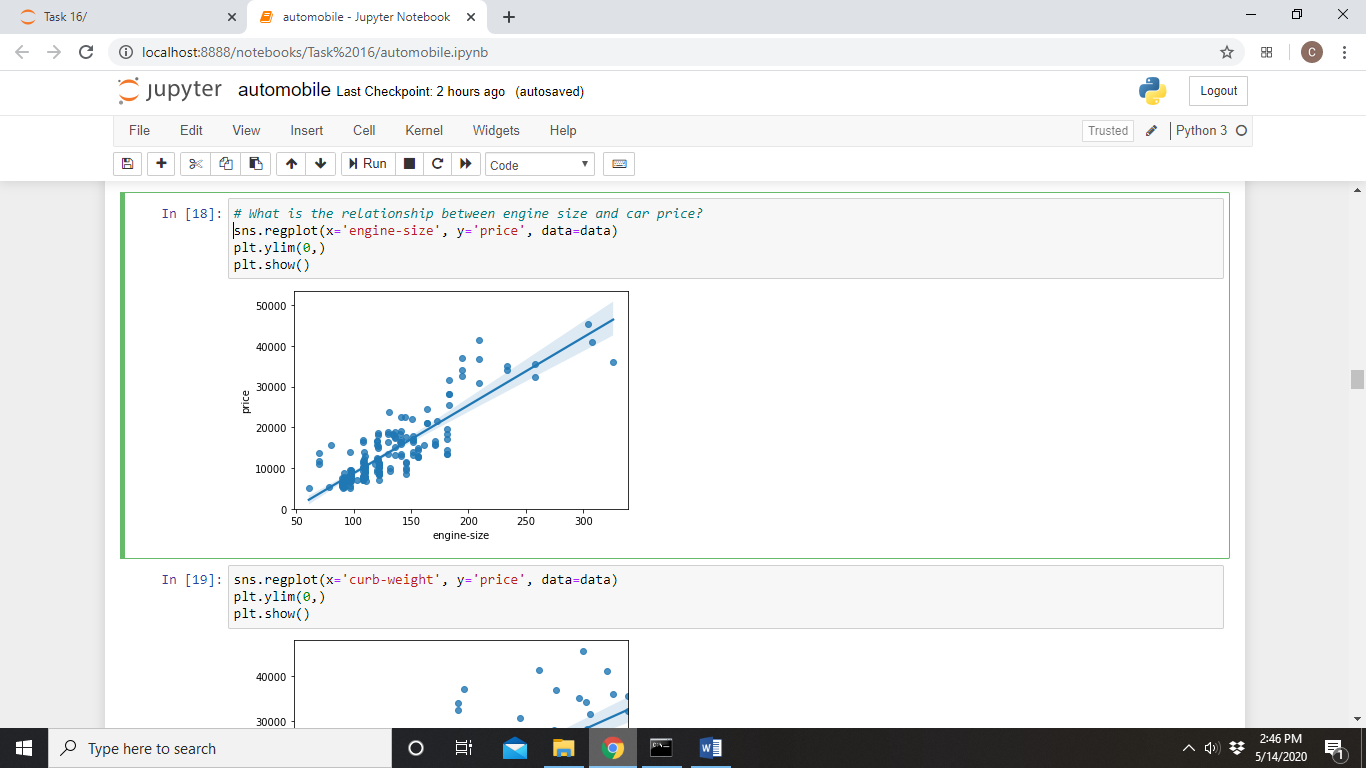
DATA STORIES AND VISUALIZATIONS

The variables price and engine size are important in this project, so I plotted histograms to visualize the spread of these values. Both variables were found to be skewed to the right, i.e. most of the values are on the lower end of the spectrum. Since we are not doing in depth statistical analysis or machine learning here, we can leave the data as is. The distribution of car prices is displayed below.

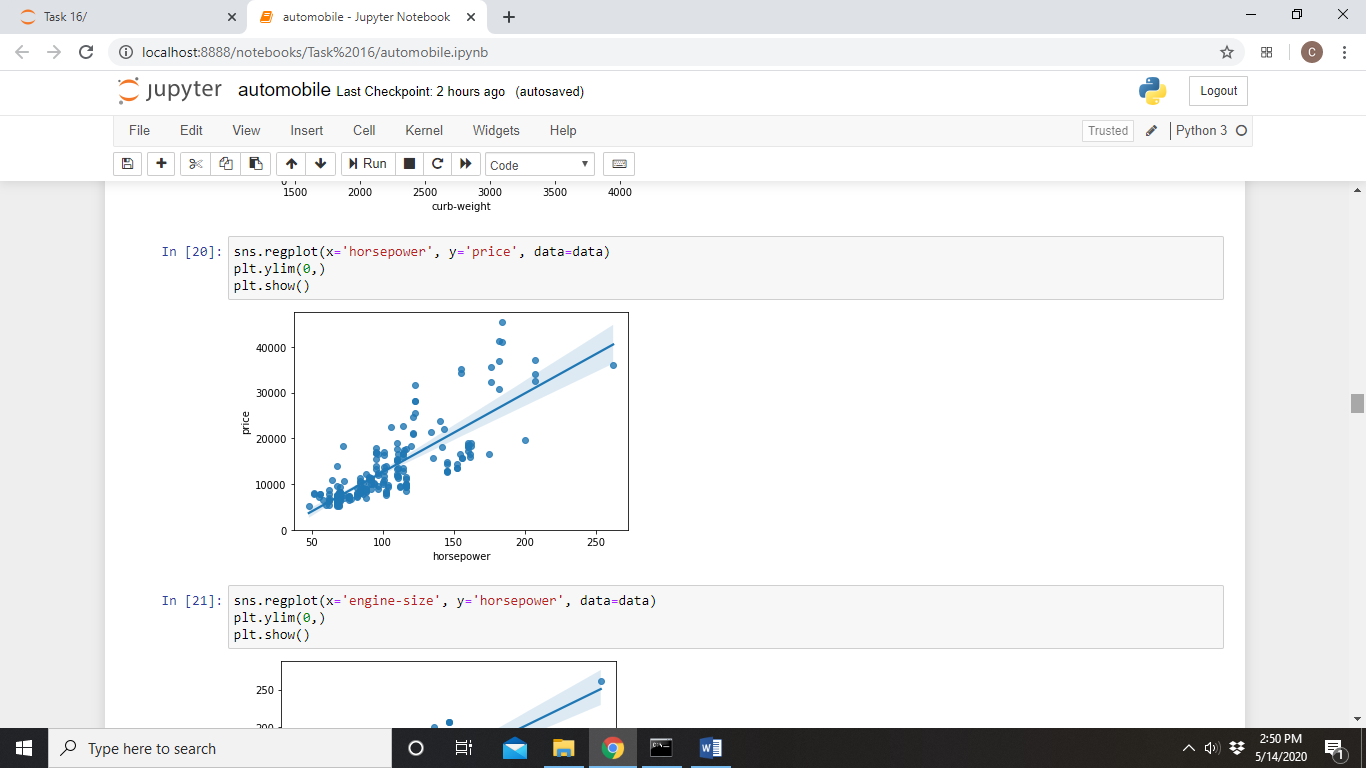
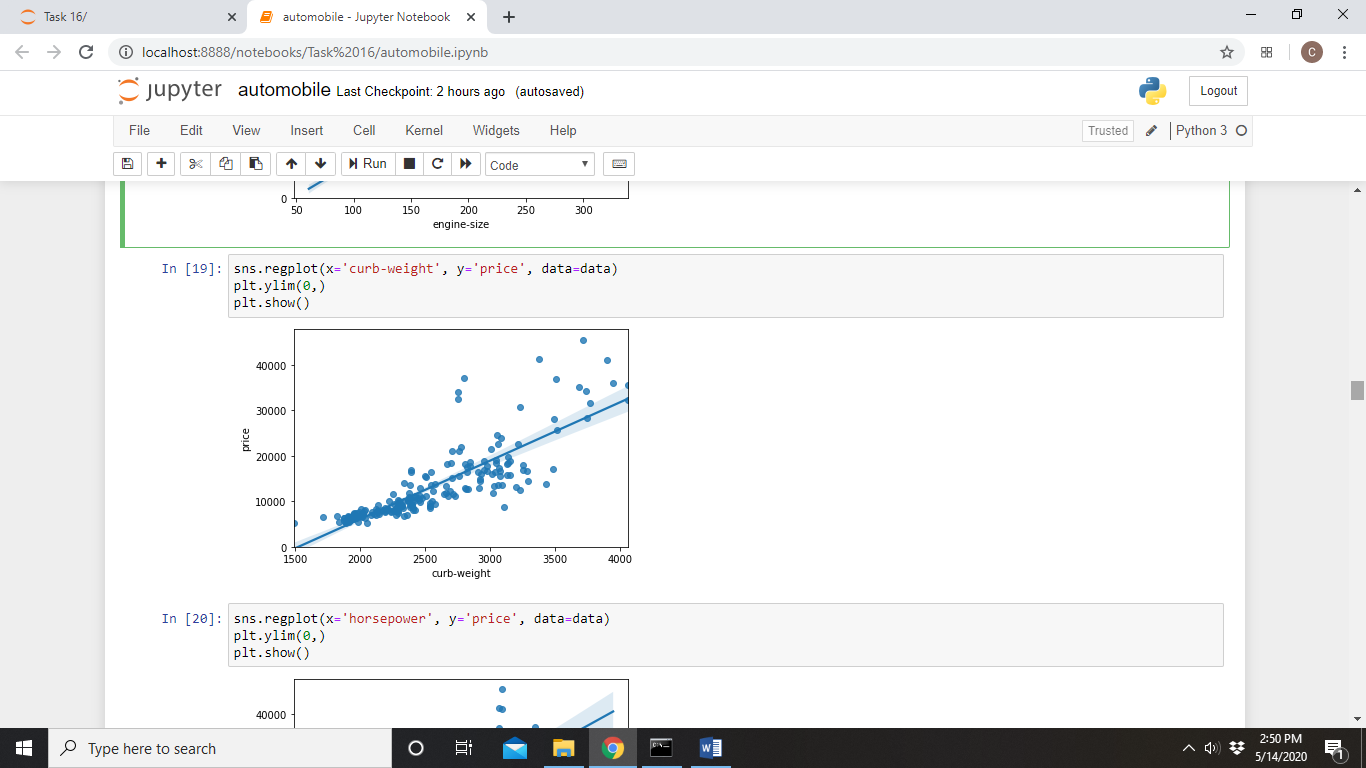


**What influences the price of a car?**

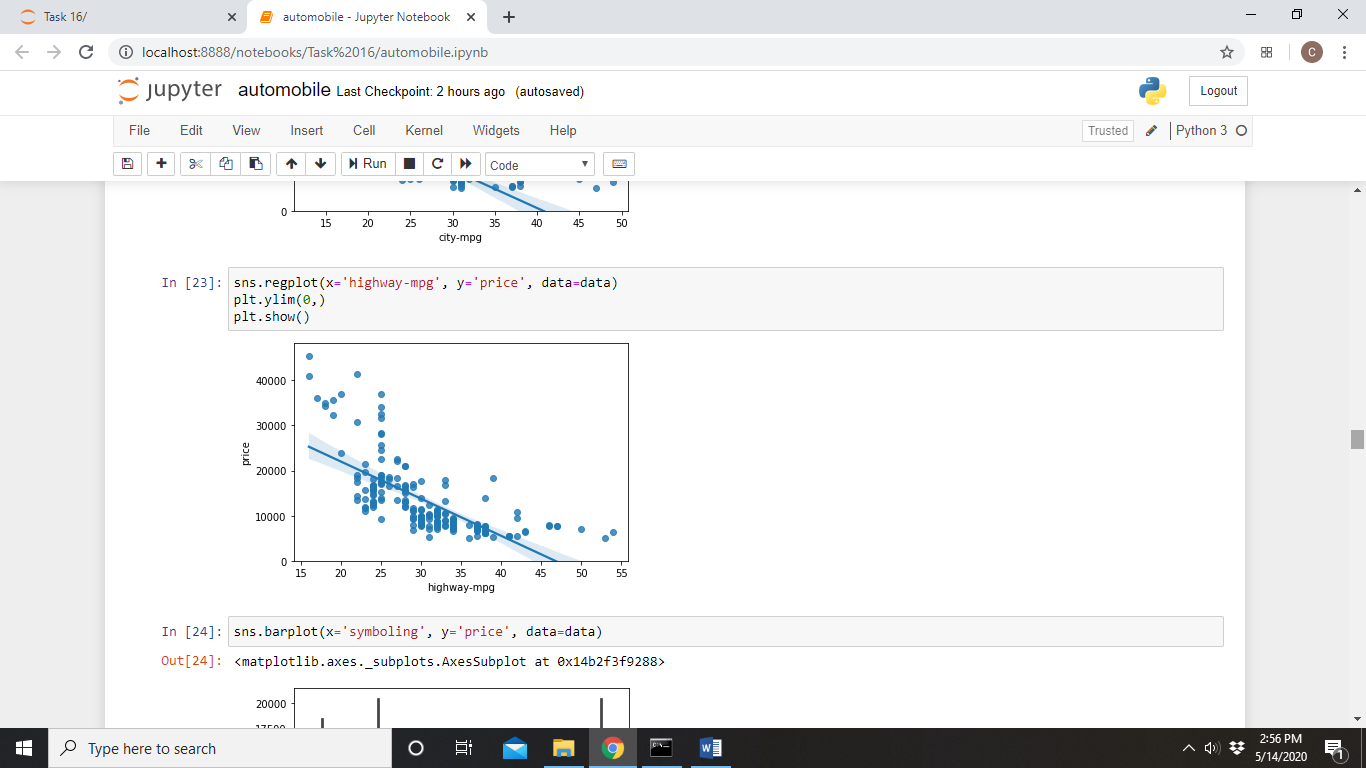
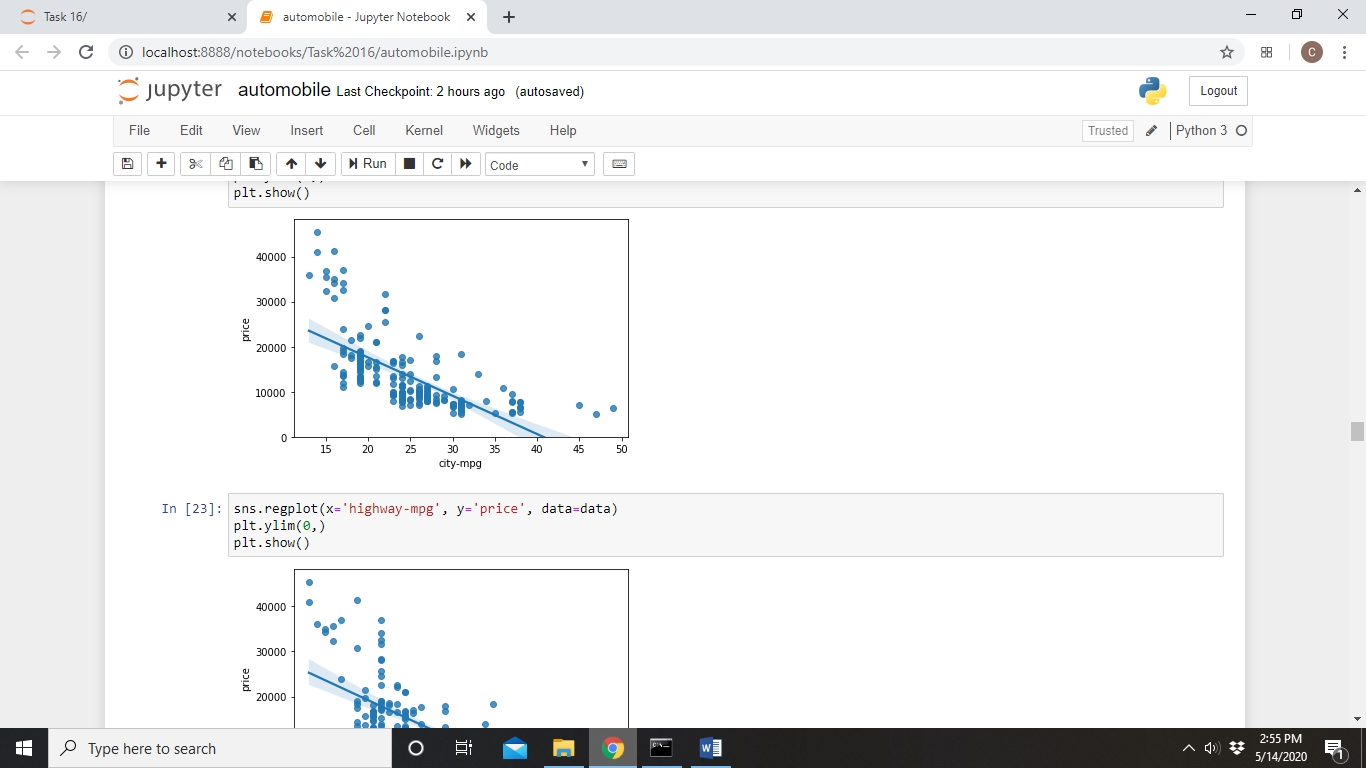
What drove this investigation was the need to determine what factors influence the price of a car. So, firstly I asked “Is car price related to engine size?” Considering the following visualization, it seems like there is a direct link between the two variables.



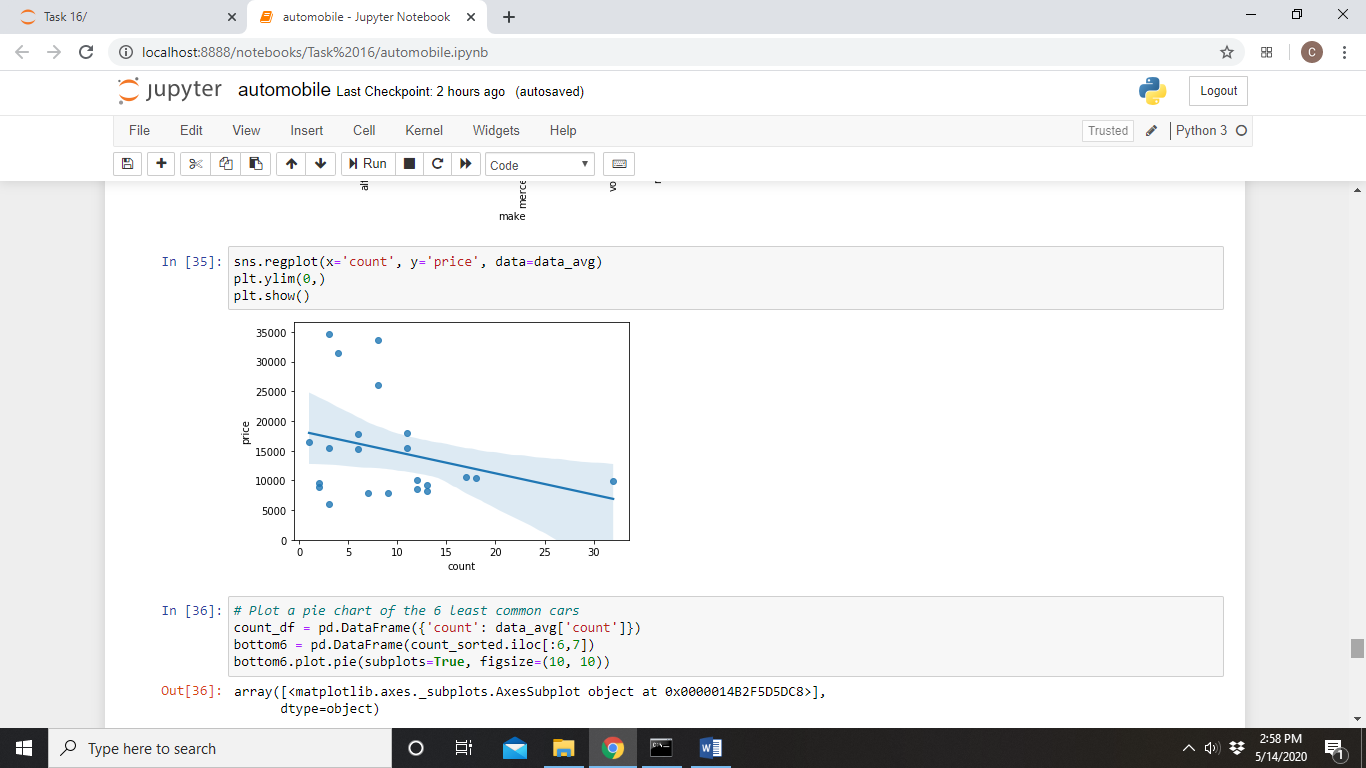
Similar plots were generated to visualize the relationship between price and curb weight as well as price and horsepower. Both plots revealed a direct relationship.



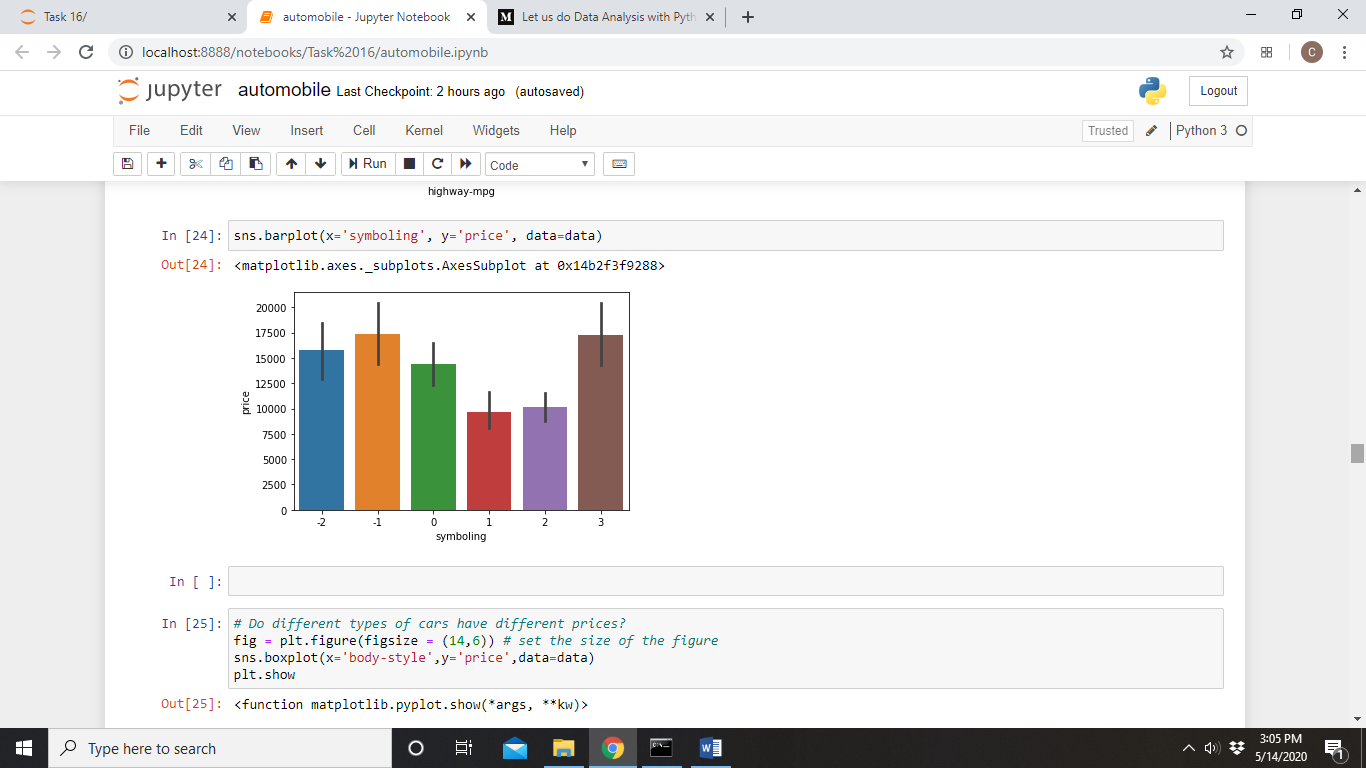
Interestingly, there is an indirect relationship car price and mileage and between car price and how common a car is. The below plots show the relationship between price and mileage in a city (left) and on a highway (right).



The visualization below reveals that car price decreases for more common cars, i.e. rarer cars tend to be more expensive.

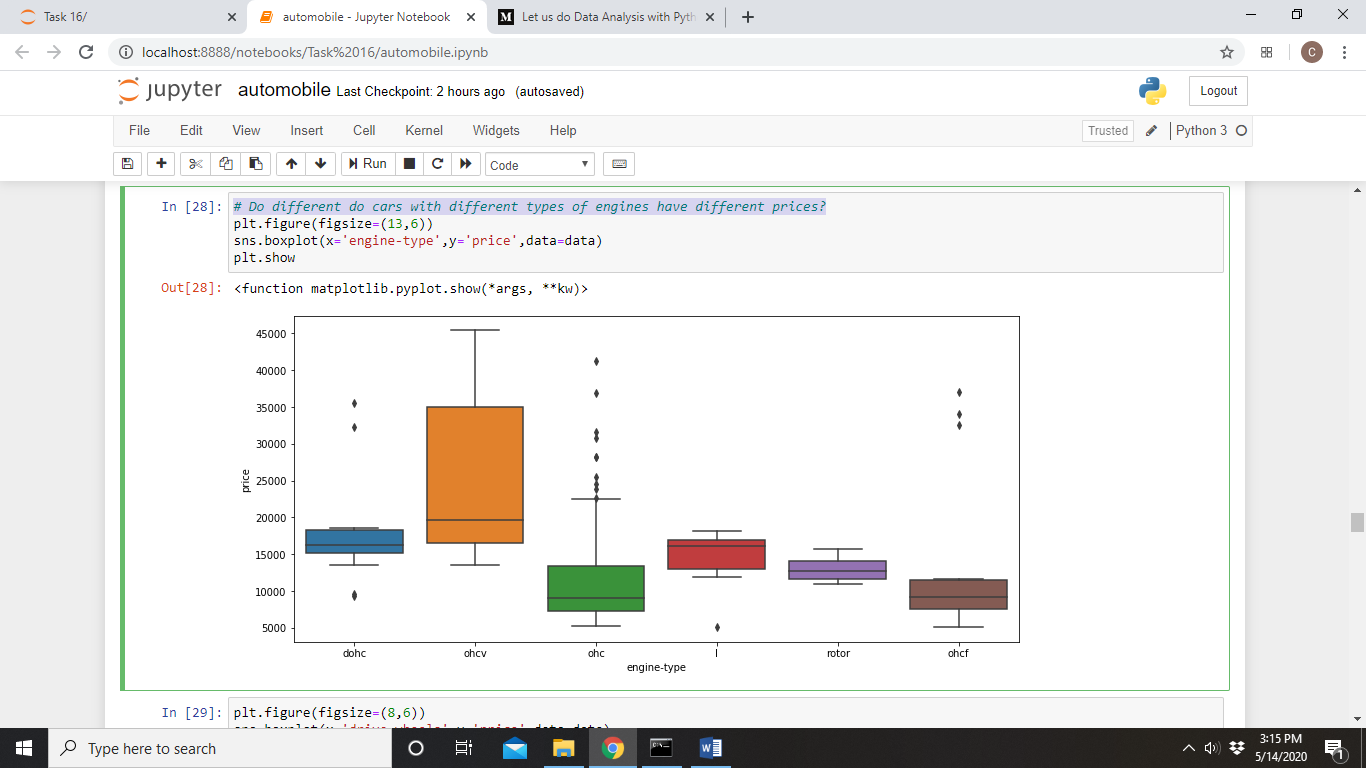


The automobile dataset had a column indicating the insurance risk level of cars. The risk levels are arranged from -3 to 3, with -3 being the safest and 3 being the riskiest. Since the risk levels are grouped categorically, I generated a visualization to view how car prices differ according to the insurance risk level.



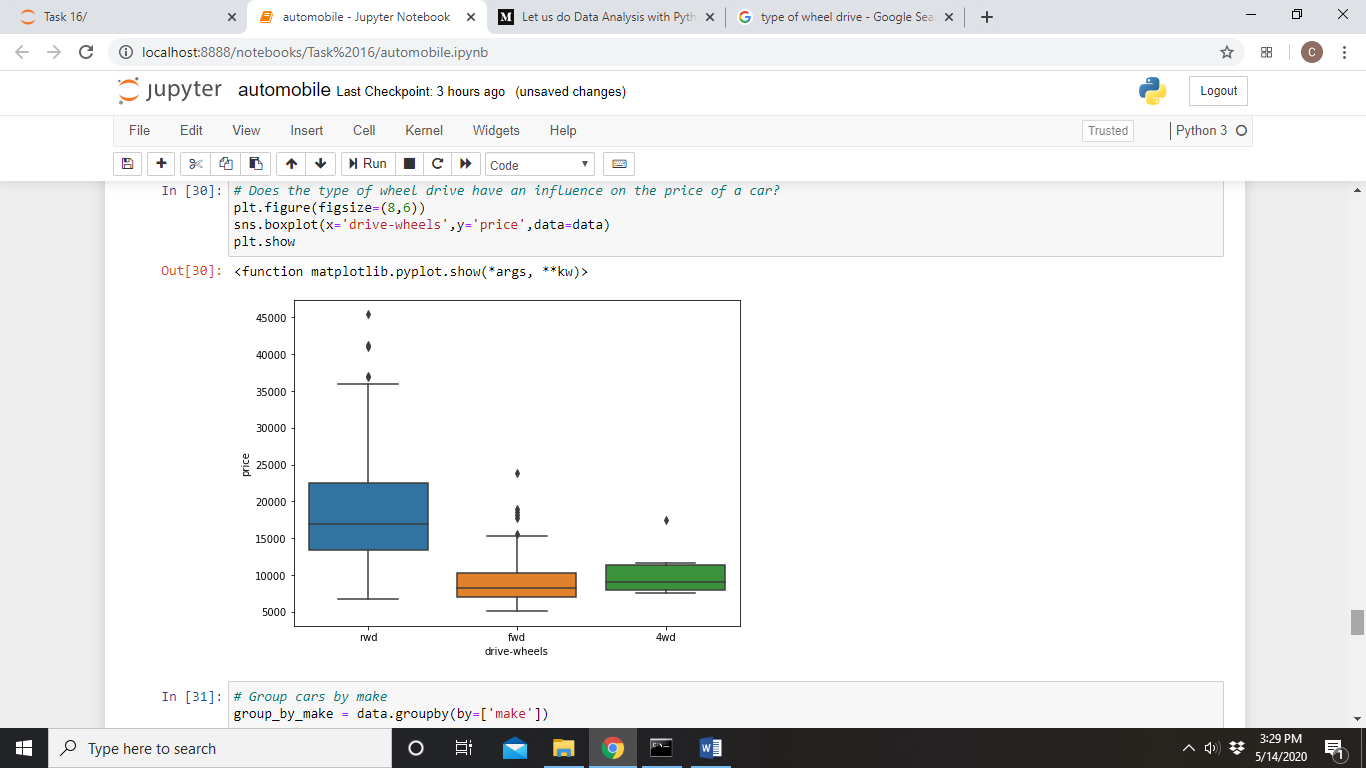
What the visualization reveals is that car price decreases from risk levels -1 to 1, and increases from risk levels -2 to -1 and levels 1 to 3. It seems there isn’t a strong link between insurance risk level and car price.

**Does the engine type have an influence on the price of a car?**



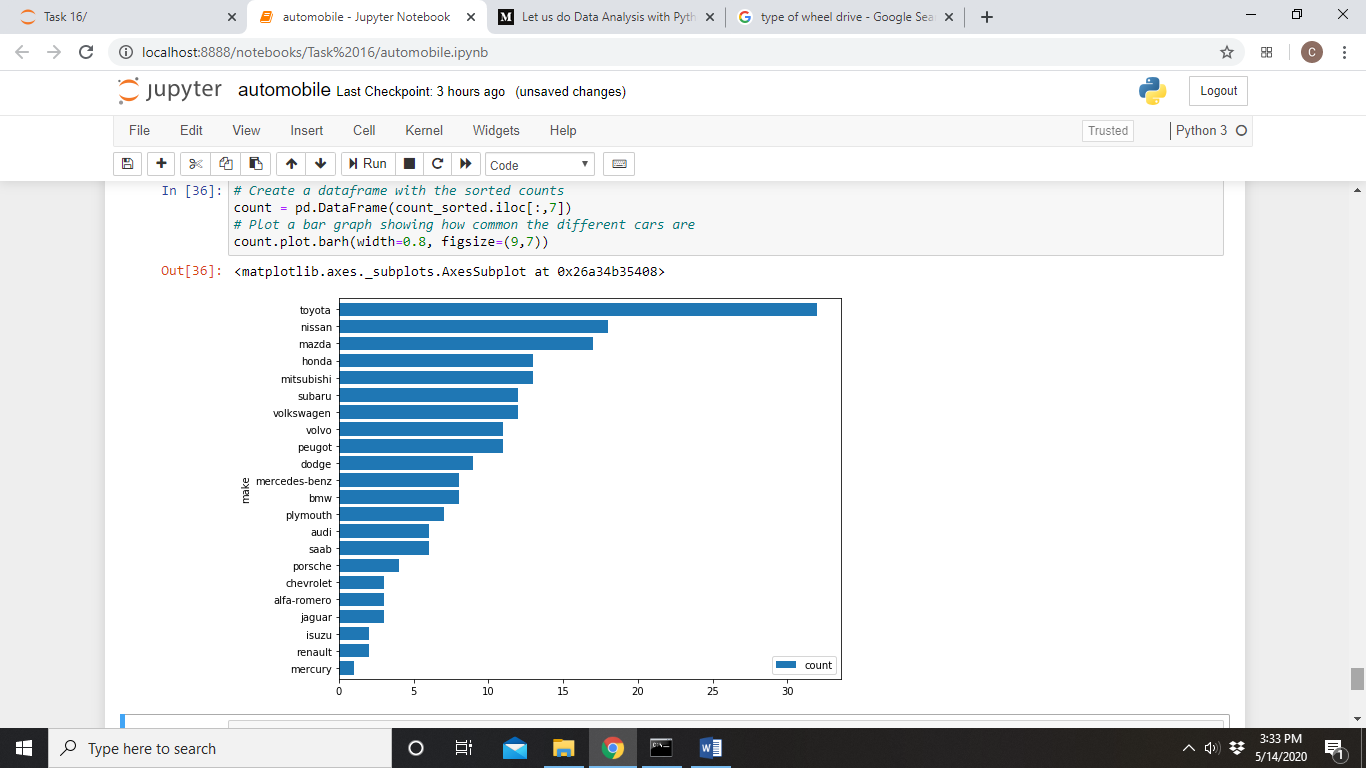
The boxplot revealed that cars with different engines do have different prices, and that the most expensive cars tend to have overhead valve engines (the orange box), although the range of prices is very wide.

**Does the type of wheel drive have an influence on the price of a car?**

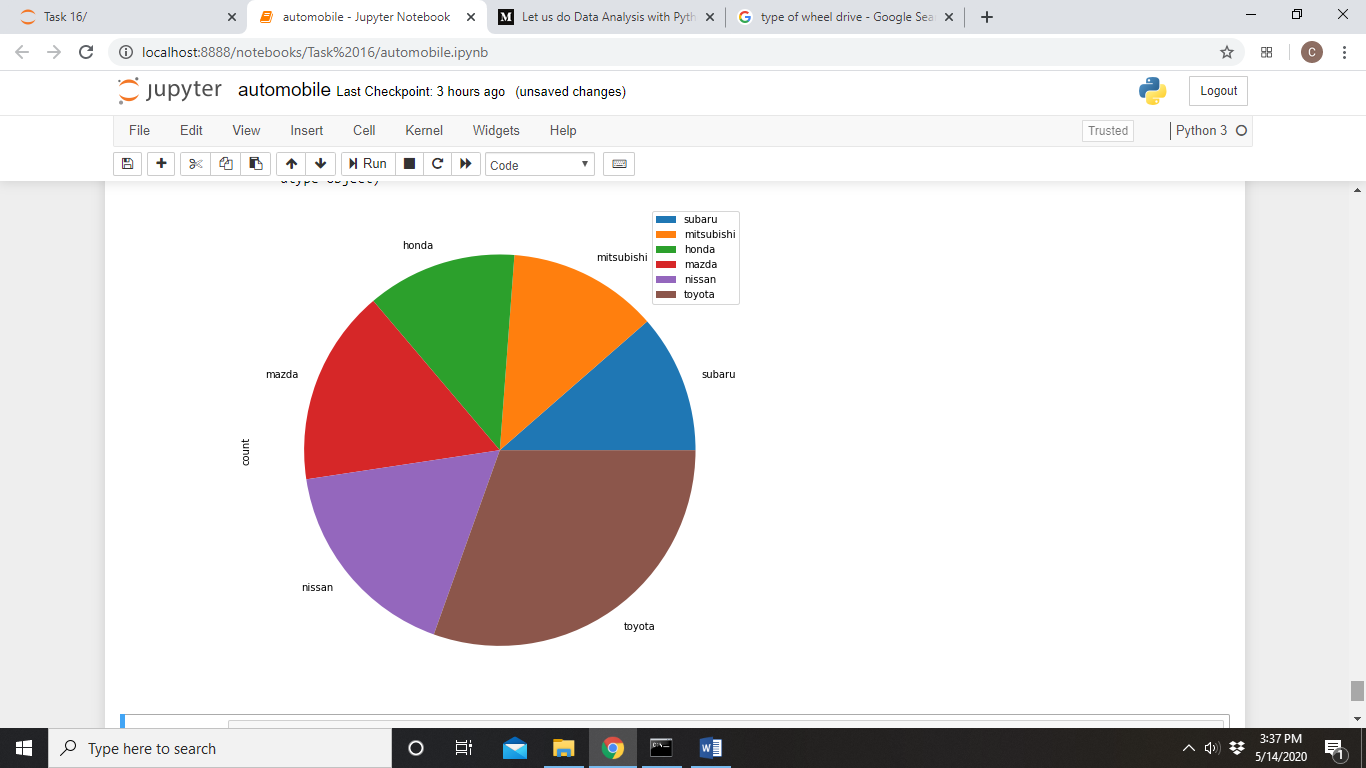


Based on this boxplot, it seems that the most expensive automobiles are those with rear-wheel drive. There is also a larger range of prices for rear-wheel drive cars.

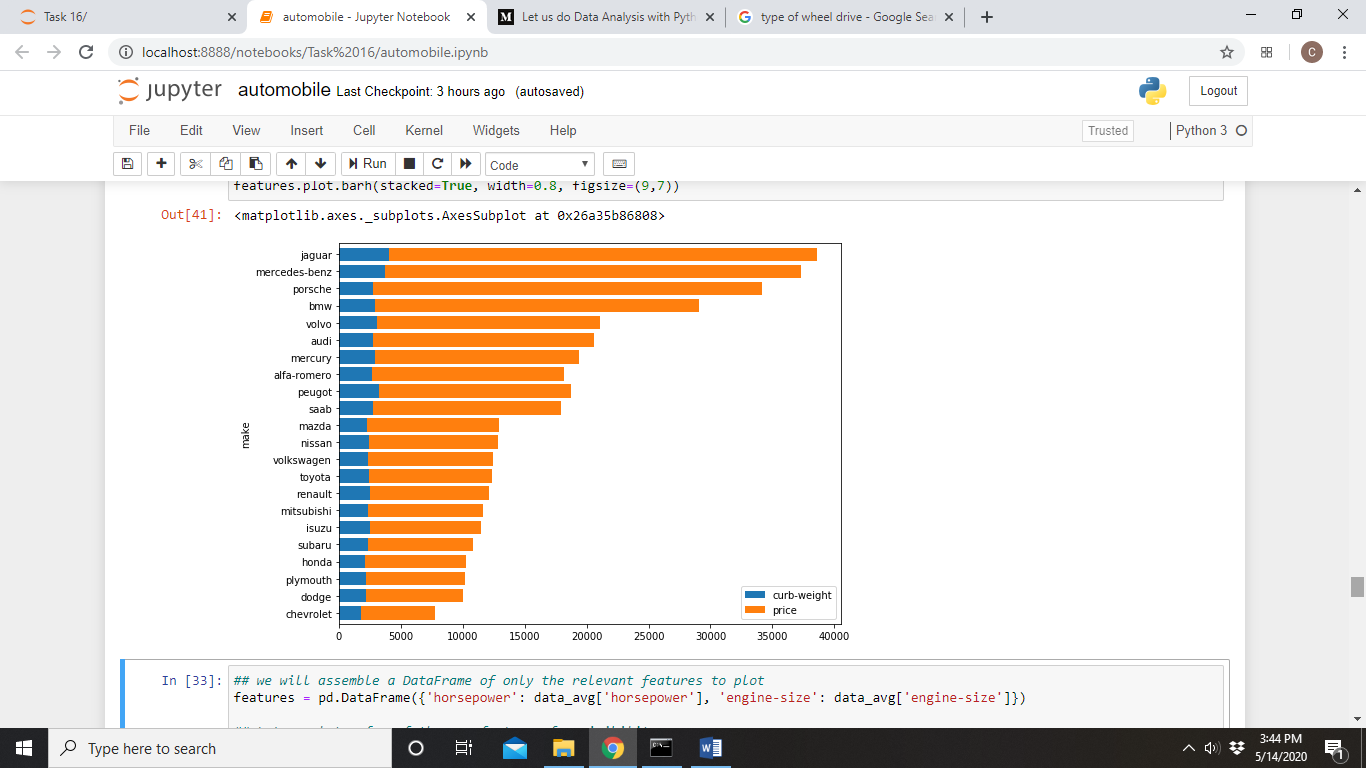
**What automobile makes are represented in the dataset and how common is each make?**



We can see that Toyota, Nissan and Mazda are among the most common automobile makes, whereas Mercury is the least common automobile. The pie chart below represents the 6 most common automobiles.

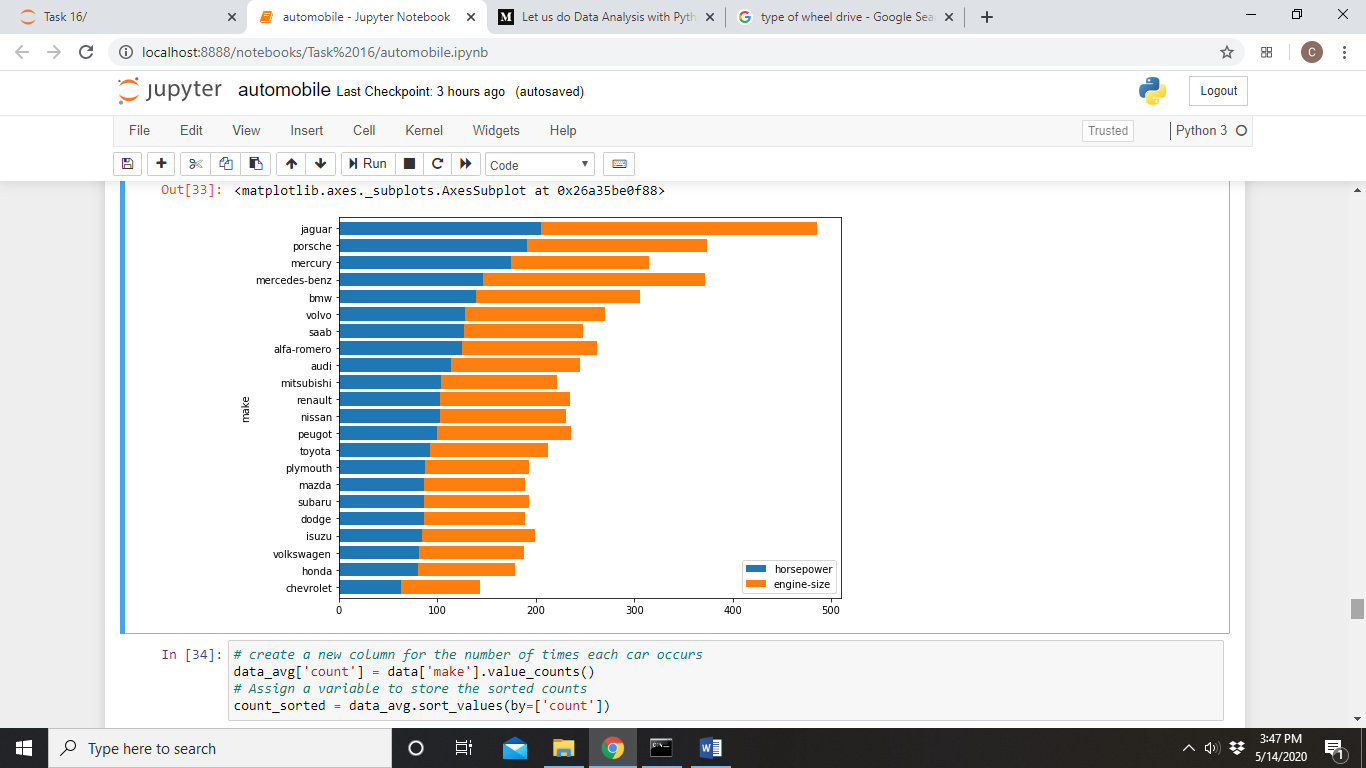


Finally, I decided to display the prices, along with the curb-weights of the different makes of car. I also decided to display the horsepower alongside the engine-size for each make of car out of interest.



If we look at the top bar, Jaguar is the most expensive car and it also has the largest curb weight.

The most expensive car makes are Jaguar, Mercedes-benz, Porsche and BMW.



Jaguar also happens to be the car with the largest engine and the most horsepower.

CONCLUSION

The exploratory data analysis conducted here reveals that there is a direct link between the price of a car and its engine size, horsepower and curb weight (each relationship considered individually). On the other hand, price seems to be indirectly related to mileage and how common an automobile is. Cars with overhead valve engines seem to be more expensive, and so do cars with rear-wheel drive.

The dataset represents automobiles of 22 makes, with Toyota being the most common. The most expensive car makes are Jaguar, Mercedes-benz, Porsche and BMW. The car with the highest curb weight, the biggest engine and the most horsepower is also the most expensive car – Jaguar.

**THIS REPORT WAS WRITTEN BY: CHRISTOPHER DELPORT**

