syllabus

[# short description about the variables (data type, unit of measure, scale of measurement) 1](#_Toc85285805)

[# analyse the depicted graphs and boxplot 2](#_Toc85285806)

[# write about outlier issues and the method of outlier filtering 4](#_Toc85285807)

[Conclusion : 5](#_Toc85285808)

## # short description about the variables (data type, unit of measure, scale of measurement)

This dataset is all about Usedcars Market , revolving different brands , Various mileage cars travelled . This essay is going to have some sights on how to analyze relation among variables , how to use codes to sketch graphs , and plots that we need , how to detect outliers that should be alerted on and be investigated more .

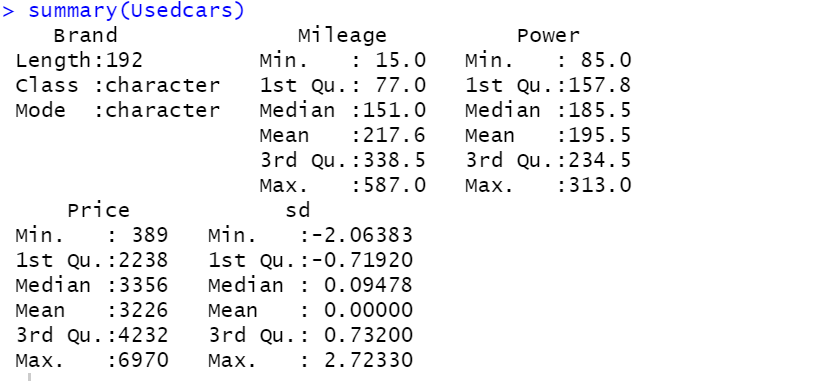
As we can see , there are 4 Variables: Brand, Mileage, Power, Price.

Brand can tell us the name of the car . It was decribed by characters . There are 3 types of the brand , they are Mercedes 、Audi、BMW. It generally could tell us : 1. Its factory / manufacture company . Different Brand could have different reputation , different avaraged price due to various time- length / financial cost took on devoloping the car technology . 2. Strength / weakness for some certain models . Though summarizing propertites of cars through brand name could be a bit subjective , it still gives us hints to have a preparation on the quality of car . For example , Benz-vichicle could possibly be better / more valued than a non-branded car .

Mileage (in 1000 km) is aggregate number of miles of cars travelled over in a given time . It is measured by numbers (double) . Among sample , it goes by a big range , from 15,000km-587,000km . It could provide a reference about how much the cars was used. All the machines（of course cars included）would go through depriciation with using it by time . As it was deducted more from original value , it would effect the price right now . But the most interesting thing could be , for different brand , the same Mileage could lead to different depriciation of cars. That’s why a multiple-variable dataset need to be searched on.

Power can mean the horse power . Data type is also number . Its range , from 85-313 , meaning the used cars have a big variety on power. The bigger it has , the more possibly it can be more valued .Price , (unit by 1000ft ) , might be the key indicator in this dataset ， as well as dependent variable for power , Mileage and Brand . The bigger power , the less Mileage , the stronger brand , can all indicate higher price .Mileage , Price , and power are all quantitative datas , while Brand is qualitative data . They are all used to describe the status of the used cars .

Table 1 sumaries of variables

.

## # analyse the depicted graphs and boxplot

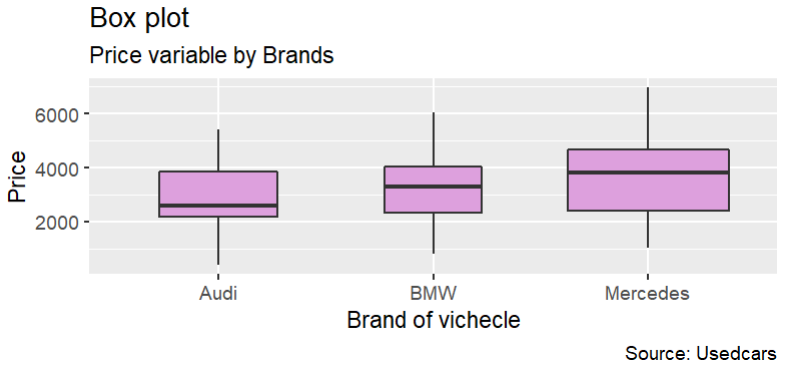


Figure 1:box plot

This box plot （Figure 1） depicits using price variable by brands . A boxplot is a standardized way of displaying the dataset based on a [five-number summary](https://en.wikipedia.org/wiki/Five-number_summary): the minimum, the maximum, the sample median, and the first and third quartiles. Seeing the location of Median , it should be noticed that Median line of Audi is more closer to bottom of the box compared to the other two Brands , this implies that middle-value price of Audi is less distanced from mimum price , the first 1/ 2 part of values are below median , 2,605k ft . For Mecedes , we see a totally opposite phenomenon . Median line is far away from the minimum and closer to maximum , menaing over 1/2 part of Mecedes cars are at price over 3,546 k ft. For Audi , the Price are spreading in a more balanced way . The median line located at approximately middle . Mecedes has the most intensive distribution for high-price cars , while Audi has the most intensive distribution for low-price cars .

From this boxplot , we could roughly evaluate that order is ： Mercedes > BMW > Audi by median , minimum and maximum. Measuring **in whole-perspective** , Mercedes can be the most valued usedcars , next is BMW , the lowest-valued brand is Audi. Besides , we can also tell Mercedes cars has widest range of price , from 1,013 to 6,970 . While BMW has a range from 800 to 6,060 . Audi has a range of 389 to 5,410 .

But in common , none of brands has narrow range of price overall. None of brands have outliers in the box plot .

# analyse and interpret the calculated desriptive statistics

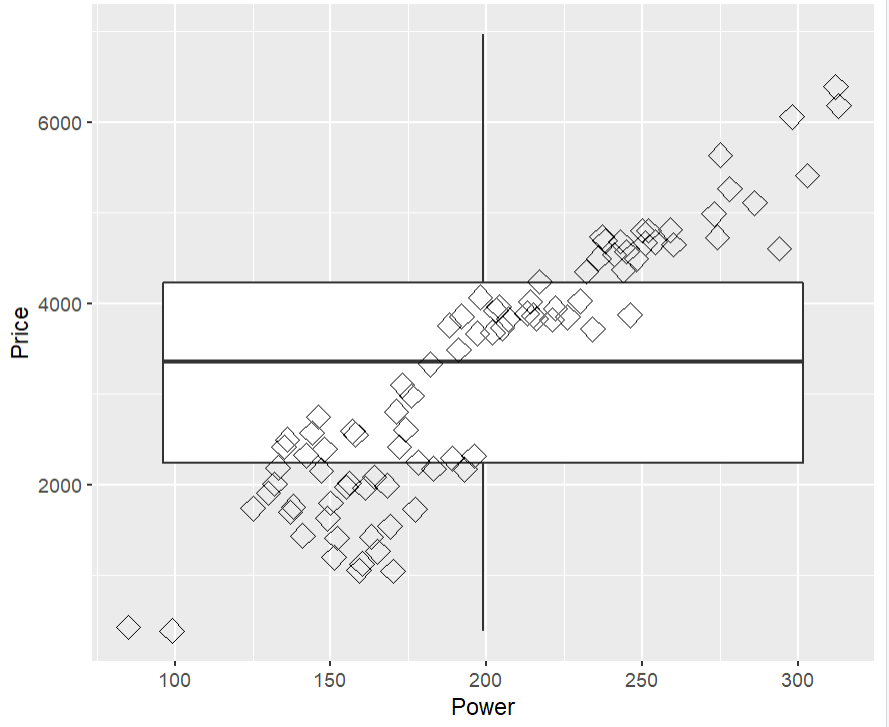


Figure 2 box plot- Power & Price

From the boxplot (Figure 2) decipting Price and Power , we could see a positive relation between these two varibale . For the value that is below the median , we see the relation is not as linear as the part that is above median . It shows that , in a certainly lower range (let us say , 3,226 k ft) , power and price is not relative enough . it is possibly be more effected by other factors , like Brands ,and Mileages . Once it exceeds a certain range (above 3,226k ft) , power would have a big wights on price of the cars . The price is more intensively distributed among 2,000-4,100 , but the power is intensive in 125 to 250 . Power has smaller range .

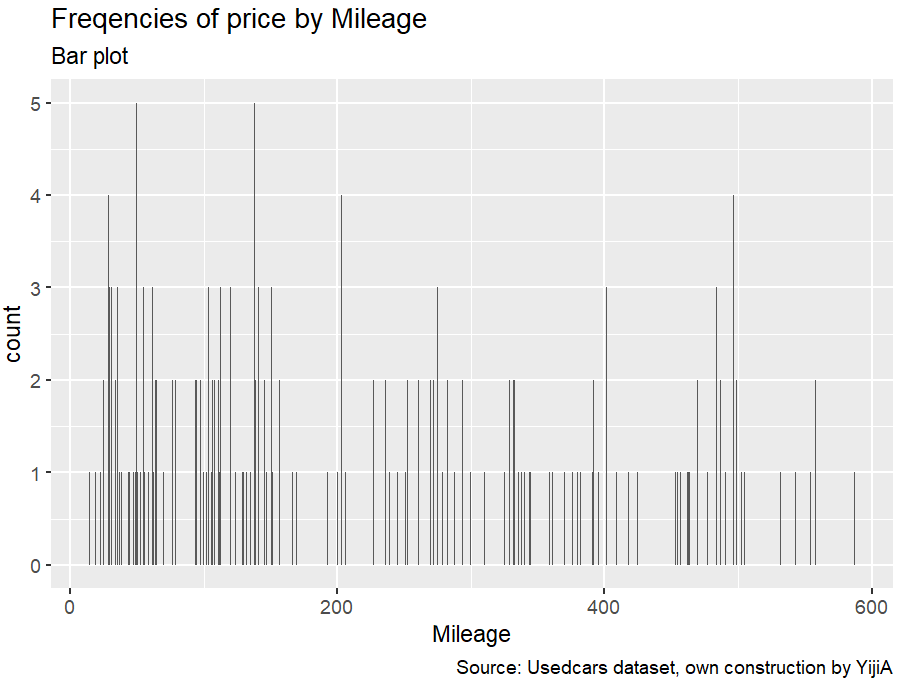


Figure 3: Mileage frequencies by Price

From figure 3 , we can analyse most of cars the less-mileage cars (below 200 ). Though These are usedcars but little of them are too extremely old / over-uesed . Besides , we can also detect , less Mileage it has before , the more likely It has higher price , from the faxt that more frequent less-mileage cars are at high price (given a count of 4 or 5 ) .

## # write about outlier issues and the method of outlier filtering

In statistics, an outlier is a data point that differs significantly from other observations. An outlier may be due to variability in the measurement or it may indicate experimental error; the latter are sometimes excluded from the data set. An outlier can cause serious problems in statistical analysis. In the Figure 1 boxplot , we didn’t find outlier .

For the outlier problem , we can take use of Z zone , to detect the part that is much more deviated by mean compared to other values . The essential varibale is Price . To calculate Z zone of the price . I use the equation : Z= （Price – Mean（Price））/s.d. . To define as an outlier , the Z >3 or Z should be < -3.

To see if there are outliers , I use “if … else …” function . it turns out the outliers exisits in the zone which is presented by Z < -3 . But if calculating it , we would find the price should be -897.184 . So the coding to detect outlier might be a bit wrong .

## #Conclusion :

We can generilze that Brands，Mileages and Power are have influences on the prices for the usedcars . Cars in different brands have different distributions on prices . Mecedes has the higher porpotion of higher-price cars , BMW comes next , Audi act as the last . Power seems would only matters on price if the prices are higher than the averaged value . Mileage could not cause very strong influence on cars , but increasing the possibilities of high price if the mileage is smaller . Most of the usedcars here in the dataset are not too old , doesn’t have too many historical mileage before . And there is no outlier in this dataset , but the way for dealing with outlier should still be waiting for investigating .