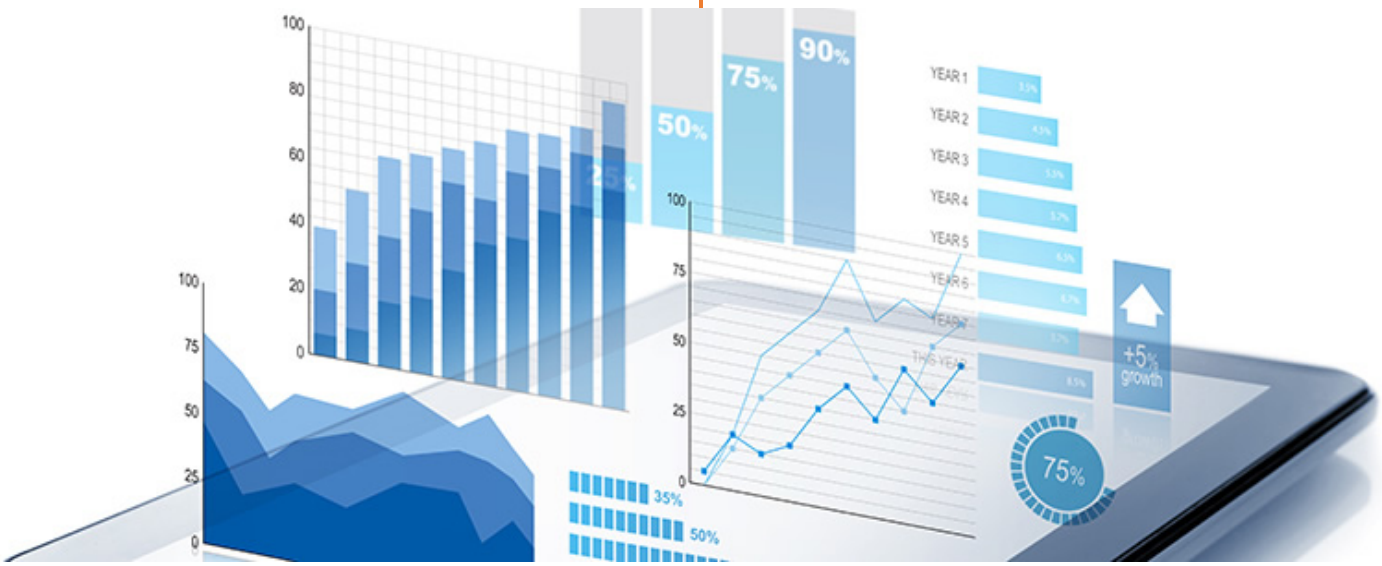
SCHOOL OF HUMAN SCIENCES & TECHNOLOGY

Jonas Hellevang

Exploratory Data Analysis

Math and Statistics for Data Analysis



I started the assignment by uploading the data in Excel where I converted the information into a table. From there I filtered out the question marks in the data to remove the rows containing missing values. What effect did it have leaving out this data? I removed in total eight rows, where four of them did not have a price, two did not have number of doors, and two did not have horsepower. Will these eight values have any effect on the table of descriptive statistics for all the data? Probably not too much on the mean, median, 1st – and 3rd quartile considering the sample size, but they most definitely can have a bigger impact on our data on our minimum and maximum value. We can also see that one brand is removed completely (Renault), and no rows are missing more than one value. I have highlighted the values that would impact the actual table the most:

Car Properties	Number of Doors	Length	Width	Height	Number of Cylinders	Engine Size	Compression Ratio	Horsepower	Price
Min	2.00	155.90	63.60	50.50	4.00	90.00	7.00	64.00	8558.00
Max	4.00	181.50	72.30	55.50	8.00	203.00	22.70	288.00	10795.00
Mean	2.67	169.89	66.35	52.29	4.63	124.75	10.49	125.67	9635.75
Quartile 1	2.00	156.95	63.75	50.58	4.00	96.00	8.43	70.00	9110.75
Median/Quartile 2	2.00	176.25	66.50	52.00	4.00	126.50	9.15	86.00	9595.00
Quartile 3	3.50	177.90	66.93	52.80	4.25	132.00	9.70	145.50	10120.00

Descriptive Statistics:
In the table below I have presented the descriptive statistics of all car properties without the rows that are missing values:

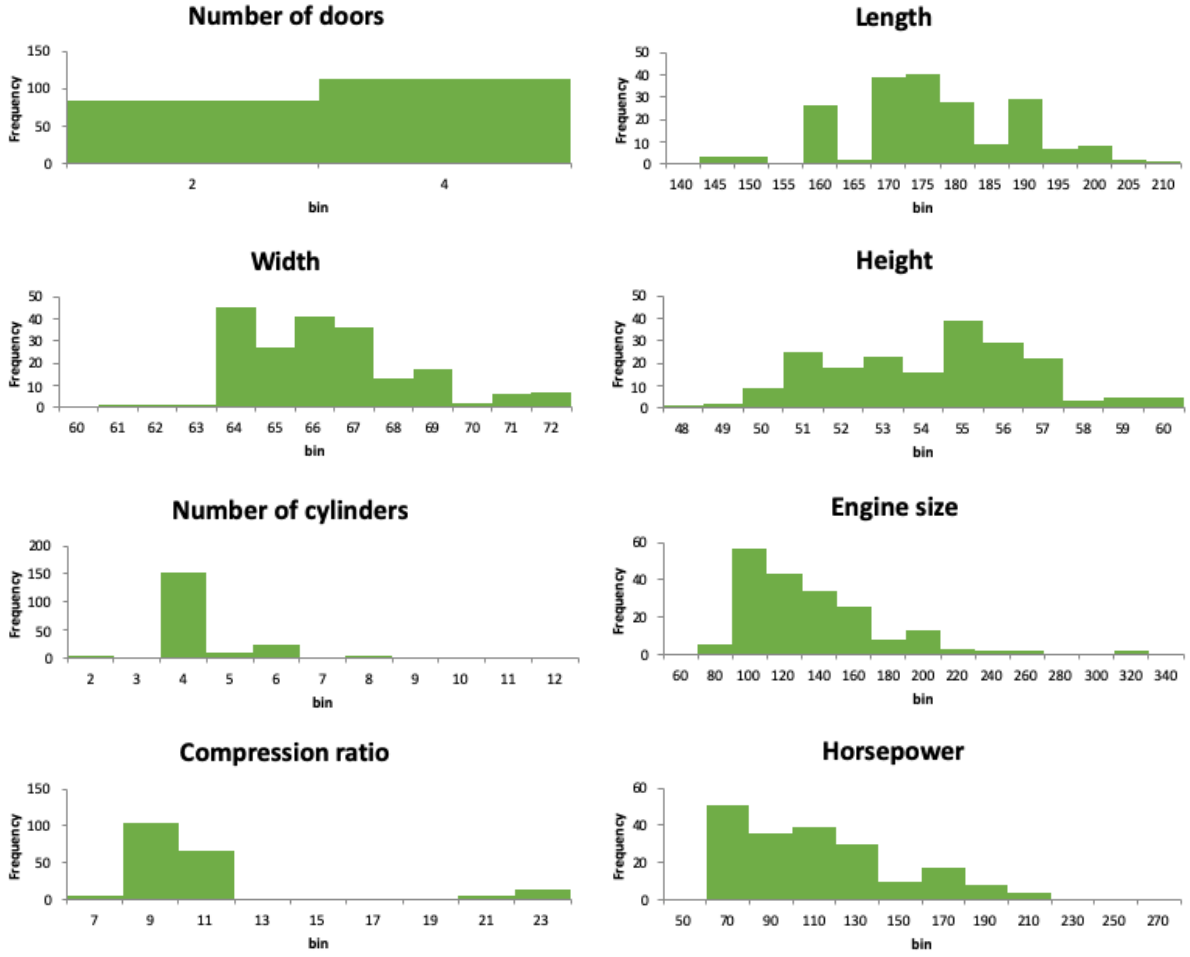
Car Properties	Number of Doors	Length	Width	Height	Number of Cylinders	Engine Size	Compression Ratio	Horsepower	Price
Min	2.00	141.10	60.30	47.80	2.00	61.00	7.00	48.00	5118.00
Max	4.00	208.10	72.00	59.80	12.00	326.00	23.00	262.00	45400.00
Mean	3.14	174.22	65.89	53.78	4.37	126.99	10.13	103.60	13279.64
Quartile 1	2.00	166.80	64.10	52.00	4.00	97.00	8.60	70.00	7775.00
Median/Quartile 2	4.00	173.20	65.50	54.10	4.00	119.00	9.00	95.00	10345.00
Quartile 3	4.00	183.50	66.90	55.60	4.00	145.00	9.40	116.00	16503.00

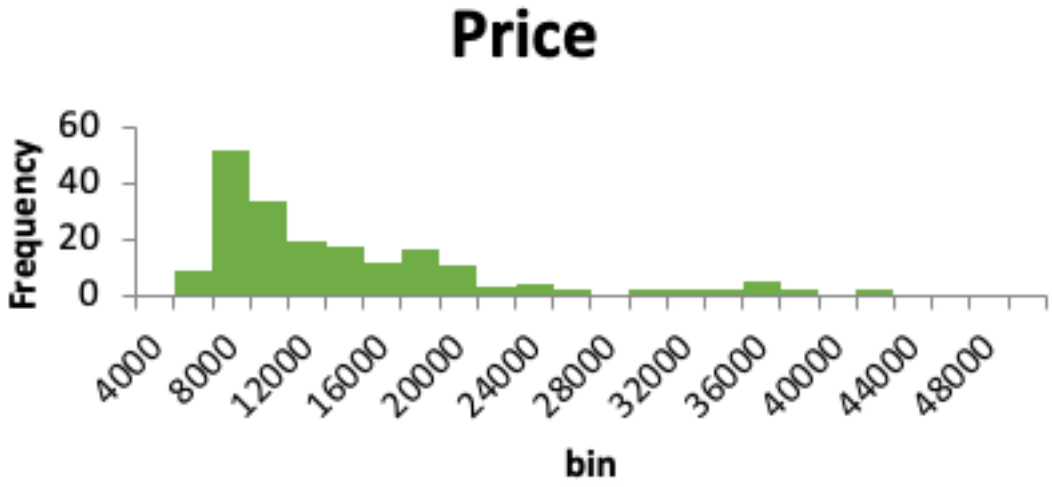
Best use of the descriptive statistics:
Since there are fewer expensive cars than cheap for the price, the median would in this case be a better fit than the mean. We can also after analyzing the data further see that the cars running on diesel are the one with the high compression ratio. In fact, the lowest compression ratio of the diesel cars is 21, which is why the maximum value is so high compared to the lowest value. Diesel cars also tend to be longer, wider and higher than cars running on gas. You can also see that the engine size is usually bigger as well.

Below you have a table of the diesel cars, marked in yellow where the number is significantly different than the average of the entire sample excluding the missing values:

Car Properties	Number of Doors	Length	Width	Height	Number of Cylinder	Engine Size	Compression Ratio	Horsepower	Price
Min	2.00	165.30	63.80	52.80	4.00	97.00	21.00	52.00	7099.00
Max	4.00	202.60	71.70	58.70	6.00	183.00	23.00	123.00	31600.00
Mean	3.68	182.23	67.48	55.85	4.32	136.42	21.97	85.53	16103.58
Quartile 1	4.00	171.70	65.50	54.90	4.00	106.50	21.25	62.00	8745.00
Median/Quartile 2	4.00	186.70	68.40	55.70	4.00	145.00	21.90	95.00	13860.00
Quartile 3	4.00	189.85	68.65	56.70	4.50	152.00	22.75	100.50	20407.00

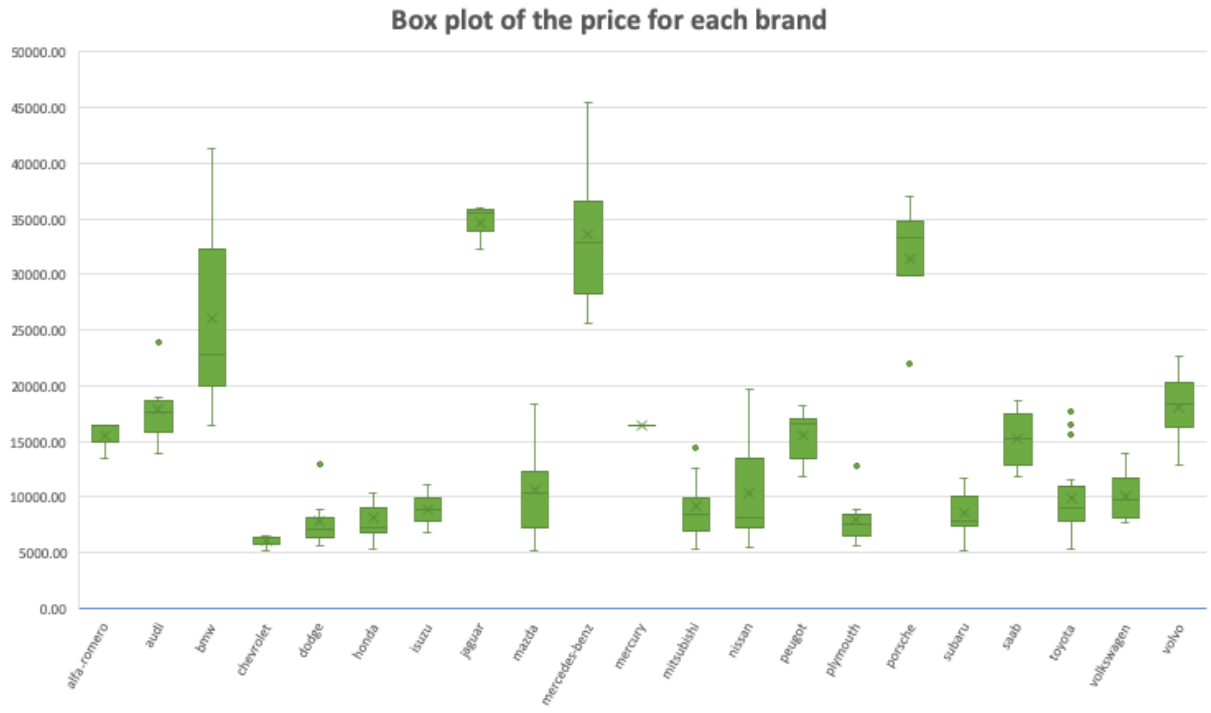
Histograms:





From the histograms above we can see that it is only length, height and width that somewhat reminds us of a bell shape and are normal distributed, while the others are skewed to the right. For these that means that the median is smaller than the mean, and which again indicates that the median is the right number to use for finding the outliers in the price on the second last page of this paper.

Box plots:



A box plot is the second tool I have used to show the numeric variables graphically. This tool is great to show range and spread of the numeric variables. In the box plot for every brand above you have:

- 1st quartile which together with the 3rd quartile give you the interquartile range which is the whole green area. This is a great tool to eliminate outliers described below.
- The middle line indicates the median, which in some instances are different than the mean.
- X indicates the mean for each brand-price.
- You have also a line going upwards called whiskers. At the top of the whisker you find the maximum value, and at the bottom you find the minimum value.

We can draw some conclusion from just having a look at the different box plots:

- BMW, Jaguar, Mercedes-Benz and Porche are in general more expensive than other brands.
- The variety of the price is bigger in BMW and Mercedes-Benz than any other brands, and Mazda and Nissan do also have some noticeable variety.
- The mean and median are very different in BMW, Mitsubishi, Nissan, Peugeot and Porche. This can be explained by the effect the few higher or lower values have on the average price of the different brands.
- Mercury does not have any variation in the numbers, but this is due to only one car.
- Some of the brands have dots outside of the box plot. These indicate outlier per brand. We could argue to take these out of the calculation, but I have decided to keep them to illustrate that even though the most expensive brands are most of the outliers described on the next page, each brand can be viewed on their own and we get a completely different picture of what is outliers and what is not.

Outliers:

make	price	Outlier
bmw	30760.00	TRUE
bmw	36880.00	TRUE
bmw	41315.00	TRUE
jaguar	32250.00	TRUE
jaguar	35550.00	TRUE
jaguar	36000.00	TRUE
mercedes-benz	31600.00	TRUE
mercedes-benz	34184.00	TRUE
mercedes-benz	35056.00	TRUE
mercedes-benz	40960.00	TRUE
mercedes-benz	45400.00	TRUE
porsche	32528.00	TRUE
porsche	34028.00	TRUE
porsche	37028.00	TRUE

IQR	8728
Upper bound	29595
Lower bound	-5317

In what way these cars are outliers or not can be discussed as all of them are luxury brands. Only one Porche, three Mercedes-Benz, and five BMW are not considered interquartile range outliers from this dataset, while all of the Jaguars are. It is not very strange considering the 3rd quartile has a value of 16,503. This tells us that since there are upper bound outliers, and no lower bound outliers, the mean for the price might not be the best tool. We can see that the median is about 3000 lower. As you can see from the box plot explained on the previous page, they would not be considered outliers based on the separation for each brand.

Why did I use the 1.5 IQ from the median? As you can see on the histogram of the price, the histogram is skewed to the right. Had it been bell-shaped, we would have calculated it differently using the mean and standard deviation times 3 instead of median and interquartile range times 1.5. The interquartile range is calculated by subtracting the 3rd quartile from the 1st quartile. From there we can calculate the upper bound and lower bound values deciding when a value is an outlier.

Conclusion regarding the dataset:

- For this dataset it is not wise to remove outliers when you take the price for the whole dataset. You can remove some outliers from the dataset from each brand where the outlier might play a role, but I decided not to because we can always use the median and not the average to eliminate outliers from the calculations. Removing luxury cars as outliers is not a good idea, but an option is to analyze the luxury brands by themselves, and the other brands by themselves.
- The luxury brands tend to have bigger variation in their boxplots, leaving you with a great deal of options for a luxury car when it comes to price.
- Diesel cars are very different from gas cars, which we can see when analyzing diesel cars by themselves. This is especially noticeable in the compression ratio where the minimum value is 21 for diesel cars.