1)

normalized : 0.528667

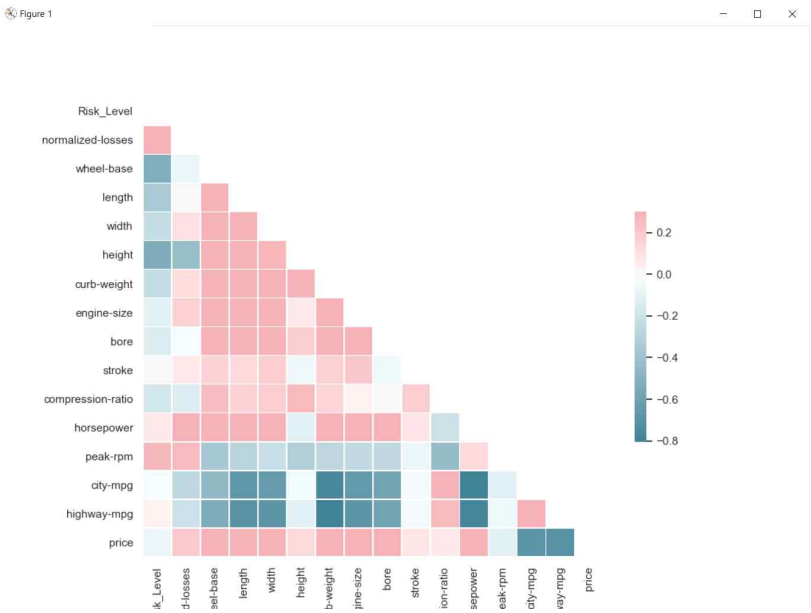
wheel-base : -0.53195

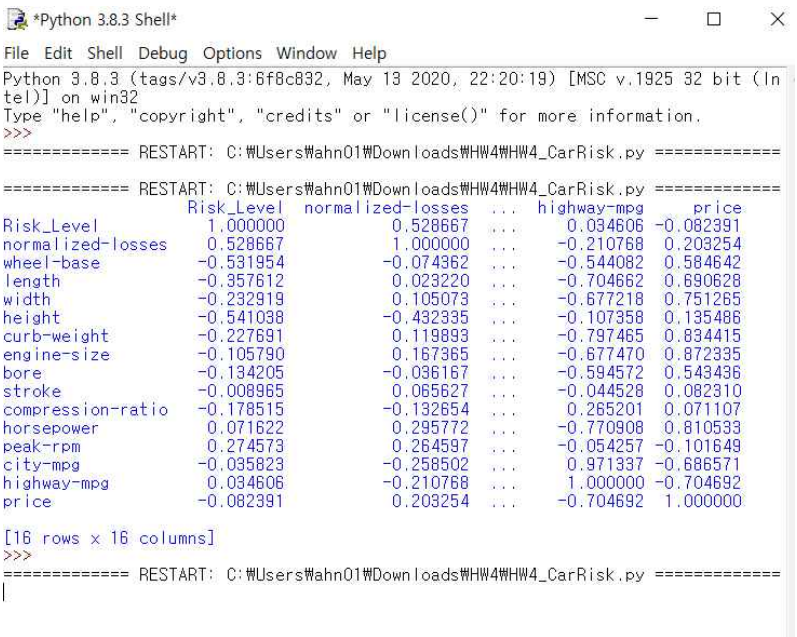
length : -0.35761

height : -0.54103

peak-rpm : 0.27457

2)





Yes. Those correlations match what I got from running them in Excel.

3)

행 레이블	평균 : Risk_Level
alfa-romero	2.333333333
audi	1.285714286
bmw	0.375
chevrolet	1
dodge	1
honda	0.615384615
isuzu	0.75
jaguar	0
mazda	1.117647059
mercedes-benz	0
mercury	1
mitsubishi	1.846153846
nissan	1
peugot	0
plymouth	1
porsche	2.6
renault	1
saab	2.5
subaru	0.5
toyota	0.5625
volkswagen	1.666666667
volvo	-1.272727273
총합계	**0.834146341**

I could observe that the risk levels of the feature 'make' were not always all the same and some such as porsche(2.6) or alfa-romero(2.3333...) were riskier than others (like jaguar(0) or volvo(-1.272727273)).

4)

```
[16 rows x 16 columns]
              Risk_Level
make
alfa-romero     2.333333
audi            1.285714
bmw             0.375000
chevrolet       1.000000
dodge           1.000000
honda           0.615385
isuzu           0.750000
jaguar          0.000000
mazda           1.117647
mercedes-benz   0.000000
mercury         1.000000
mitsubishi      1.846154
nissan          1.000000
peugot          0.000000
plymouth        1.000000
porsche         2.600000
renault         1.000000
saab            2.500000
subaru          0.500000
toyota          0.562500
volkswagen      1.666667
volvo          -1.272727
```

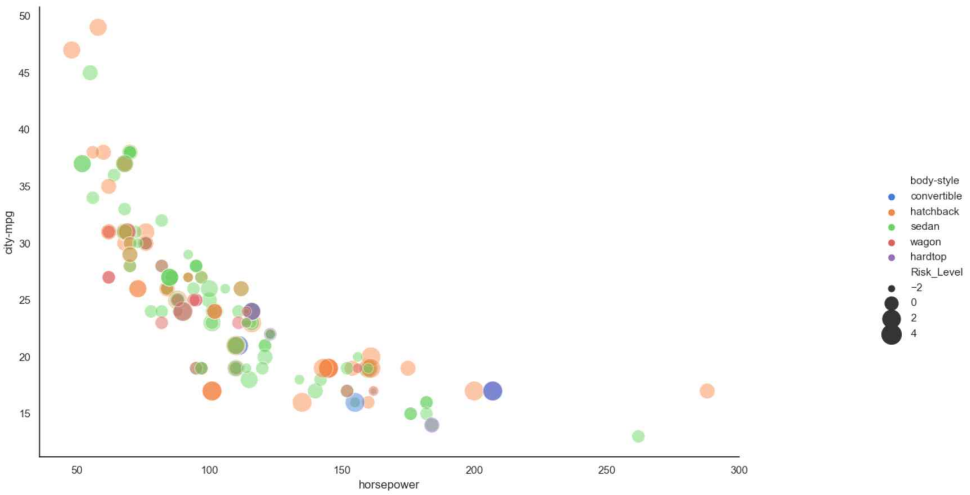
There is no difference between the averages for 'Risk Levels' for each 'make' I got above and the Excel pivot table in the previous section (completely same).

5)

The risk-level (2, -1) section marks the smallest value and the risk-level (-1,0) section marks the highest value. And this histogram is handling the integer range scale, which means that the value at the left side in the float format could be

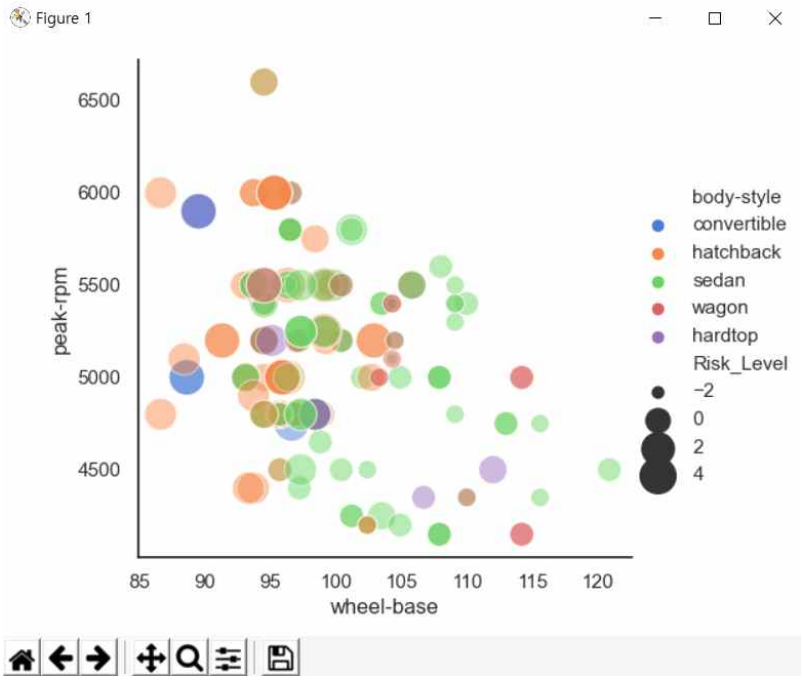
abandoned (neglected). Also, this histogram is not symmetric and not bell-shaped so not normally distributed (Gaussian) not, but I think we could do some slight manipulation such as standardization after analyzing more data to get closer-to-real-world, more accurate result than now so that we figure out whether it is actually almost normally distributed or not by increasing population.

6)



In general, 'horsepower' is inversely proportional to 'city-mpg'. And the remarkable fact is that sedan, one of the 'body-style', tends to have pretty low risk levels seeing the relative small dots than others. Thanks to the colored dots, we can know which body style prevails - sedan prevails, hardtop is scarce. Also it seems high-risk dots tends to be seen below, which means the lower city-mpg, the riskier.

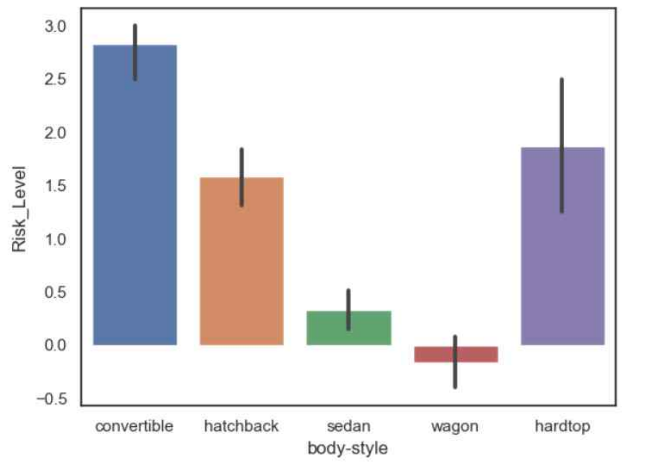
7)



Sedan tends to be evenly distributed in high or low wheel-base & peak-rpm (except almost left edge) while convertibles tend to have very low wheel-base (showed in left side). And hatchback tends to be seen at left side (low wheel-base) while wagons are usually seen at right side.

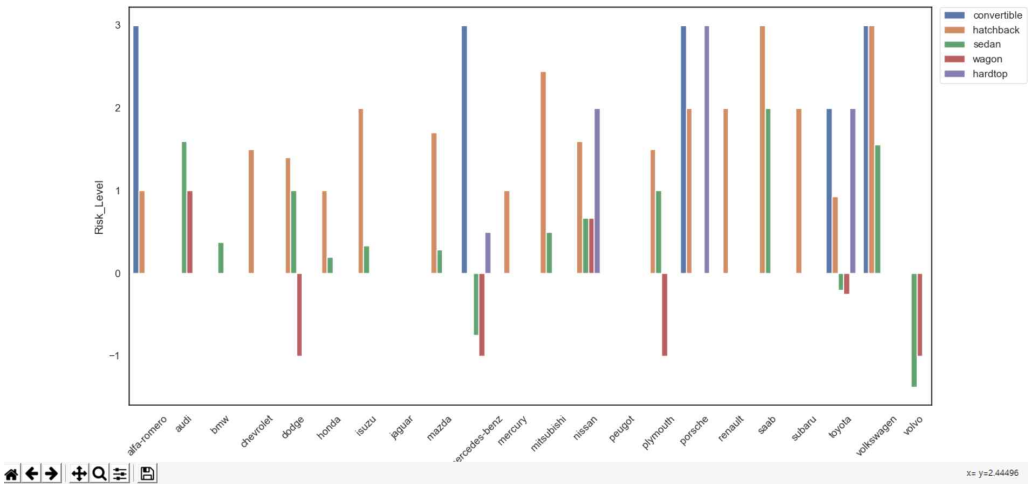
This graph also represents the relation between risk level and the two axises (wheel-base and peak-rpm). We can find this tendency ; the higher peak-rpm and the lower wheel-base, the riskier the cars become.

8)



The risk levels vary depending on 'body-style' (convertible > hardtop > hatchback > sedan > wagon). And this graph corresponds to the observed in Q7 ; In Q7, the blue dots (convertible) have large size while the green dots (sedan) are in relative small size. Likewise, the orange and violet ones are in similar, quite big size (not bigger than 'convertible' though) as the bar chart in Q8 shows. The red dots, wagon, have the smallest size among the fives.

9)



This bar chart shows quite similar information with Q3 & Q4's one (those are about risk levels according to 'make'), but is distinguished from it in that Q3 & Q4's code represented the average risk levels for each 'make' while this bar chart shows risk level according to not only 'make' but also 'body-style'. So this graph could make it clear why each make has that value of the average risk level - able to specify the causes who raise the risk level.

10)

I would give my boss these 4 fundamentally different advices ;

1) Refrain from designing car in 'convertible (just representative example)'

There was the tendency that certain car body-styles such as convertible or hardtop have higher risk than others. By replacing them with statistically safer body styles such as sedan or wagon, the car manufacturer could lower the average risk level to get better brand safety reputation.

2) Manufacture 'volvo (just representative example)'

There was the tendency that certain car designs such as porche or saab have higher risk than others. By replacing them with statistically safer designs such as volvo or peugot, the car manufacturer could lower the average risk level to get better brand safety reputation.

3) Limit the peak rpm

Seeing the graph in Q7, it is clear that the cars who have high peak-rpm tend to mark bigger dots, which means that they are much riskier than others who have low peak-rpm. So we can improve safety reputation by doing 3).

4) Make wheel-base longer

Seeing the graph in Q7, it is clear that the cars who have low wheel-base tend to mark bigger dots, which means that they are much riskier than others who have long wheel-base. So we can improve safety reputation by doing 4).

The above data exploration was just a preliminary analysis. So we could use the statistical analysis such as statistics on the accident history of each vehicle and do machine learning to search many other factors related to our target class. To do so, collecting location data (about where accidents occurred) or car-body ratio would be helpful.

11)

Technocentrism is a value system that is centered on technology and its ability to control and protect the environment. Technocentrics argue that technology can address ecological problems through its problem-solving ability, efficiency, and its managerial means and that these capabilities allow humans control over nature, allowing them to correct or negotiate environmental risks or problems.

The factors determining our mental health's state are various - they could be medical history, interpersonal relationship and etc - , and 'how they feel each day' would be one of them for sure. However, just using tech (for this time, creating app) without setting concrete plan about utilizing the data would lead to failure because it is completely depending on technology whose power is not mighty 'yet' without humans, which is like hoping ten trees grow after planting one seed. So It deserves to be called 'technocentrism'

To prevent the problems mentioned above from occurring, we have to search for other factors related to mental health more such as medical history or income level then can use our technology (including app) to calculate and analyze the correlations between them to make our way "low(er) tech" one.

12)

It is not ethical because if the tracking-gang-members-system is introduced, the downside residents could feel they are in panopticon where every move of them is surveiled and monitored and the more strongly supervised ones would feel unfairness and inequality which lead to discrimination and damage human rights. Also, the gang members have rights of privacy not to be tracked, intruded and violated. So making such system would be problematic because of the fact the system tracks someone in itself.

And the benefits would not outweigh the costs, for ;

> Developing, applying and maintaining such system cost tremendous amount. The benefits would exist for sure, however, if we target different areas of the city with more police than other areas, other areas' security would be relatively worsened, which leads to the change of criminal-target! - *benefits would be not that big*.

13)

The two are the same each other because they were made based on exactly same figures, but we cannot say that both of them are very honest (valid). Actually, none of them is completely trustworthy. Each parameter (Group A, Group B, Group C)'s value of the two graphs is in the format of percentage whose range is from 0(%) to 100(%). Then let me elaborate why each graph is *not that* valid.

<Graph #1>

In this graph, y axis's range is from 50 to 62.5, not from 0 to 100. In my opinion (subjective one), The gap between 50 and 62.5 is not that large but this graph illustrates as if the gap between them was extremely large unlike the actuality - we've already visually deceived.

<Graph #2>

Similar case with <Graph #2> but strictly we can say '#2 is better than #1' a bit. Y-axis's range is from 0 to 80 for this time, which means that the graph got pretty closer to original representation of the data although still the end of the range is not 100 - problematic.

14)

This graph is intentionally manipulated to make certain story (Internet Explorer Market share is proportional to Murders in US). Thus it is not valid.

Actually, it is fact that Murders in US and Internet Explorer Market Share are in decreasing trend as the year goes on. But each feature has different form of value - natural number and percentage which are hard to compare each other because of their different value range.

By using different value form for each and deftly sizing the minimum and maximum of y-axis, the graph-maker tried to make pseudo-correlation between them to show people that the two move similarly and related closely although drop width of each is fairly different.

15)

Every our product will be designed and devised for humans. And if we design for privacy not explicitly but implicitly or not at all, users could not recognize whether their data is now being collected or not, whether the app designer, where app developers are using that information and what information the app is trying to use (this use includes analyze and commercially sell). Thus specifying privacy issue explicitly when producing any products is necessary so that customers are not deceived in unawareness. There are no customers who want their privacy to be leaked out ,so to keep one's privacy secret is to keep one's rights. And we data scientists (also ML engineers) should make sure that we do not abuse one's privacy or even unintentionally misuse when utilizing and analyzing data.

Surely, privacy problem is included into the ethic field. However, this problem could also have to do directly with company's revenue. The product with little sense of protecting customer's privacy might yield their fear and reluctance, which leads to real damage of selling product such as boycott.

16)

First of all, extract a variety of characteristics (interaction, appearance, sound, etc.) from existing animals to create a variety of sequence pairs of those characteristics. However, each set should

> *1) not deviate significantly from the existing animal or 2) significantly be different from the existing animal.*
>
> 1) means this ; If my pet robot had puppy sounds and panda eyes in the cat's ears (indeed each is a favorite feature of many people), we would feel a sense of alienation. *Pororo* is also a penguin-based character!
>
> 2) means this ; makes it totally different from the past, giving users the feeling of meeting a new animal.

Through this algorithm, create our pet robots, then distribute them to test groups and collect user responses (either automatically through sensors or passively through observations). Or implement online software version to collect and analyze beta-test participants' responses to each model. Afterwards, analyze the features or combinations of them that elicit positive responses through ML, reapply them, and repeat the results until they meet the company's end. To do so, we should

gather many people's preference data for each existing animals, feature preference data in detail (Barking vs Bow-wow? High-tone vs Low-tone? - even they are commonly the sounds of dog), pet sales record in market and all the features that human beings are psychologically attracted to (in other words, data about psychology analyze).