Title: Data Distribution

Introduction

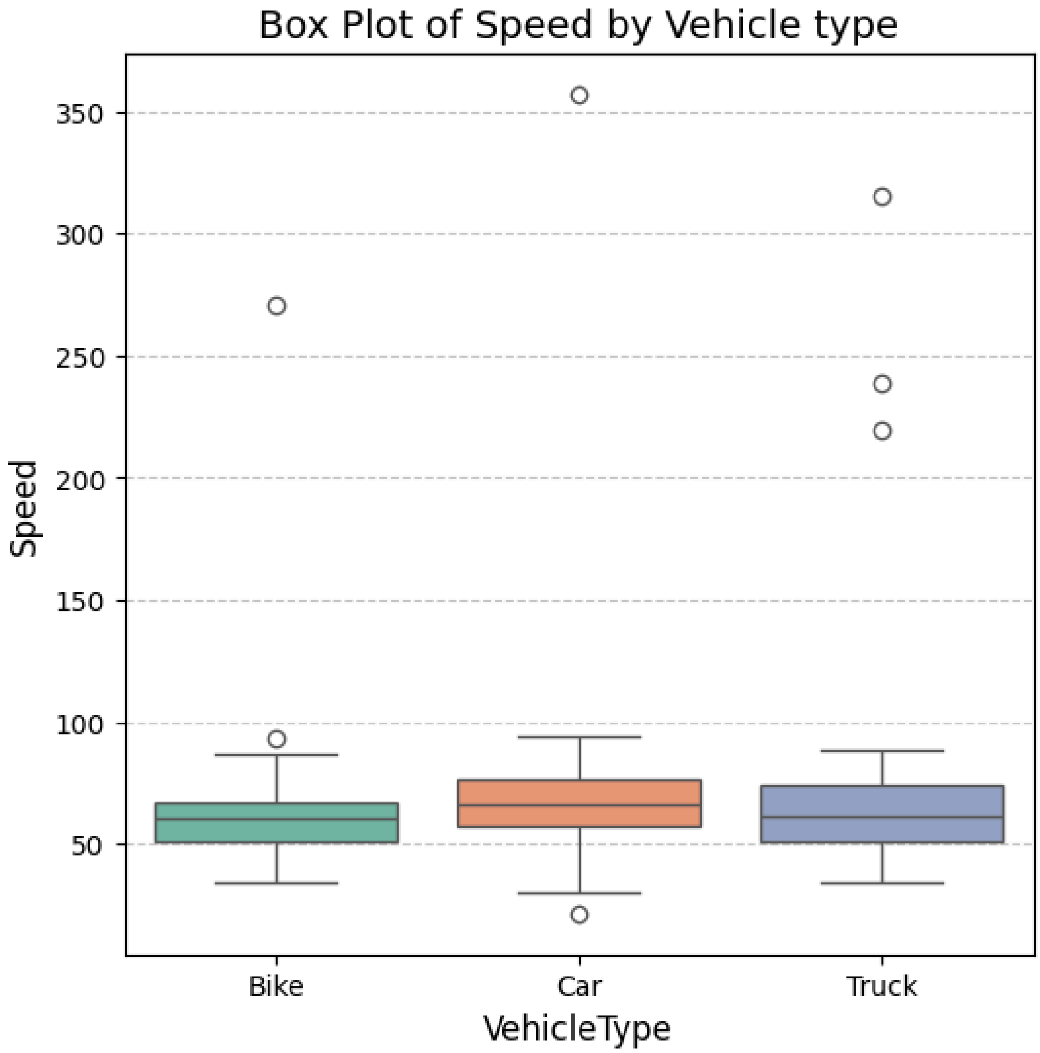
In this assignment, there will be a description of the data distribution. And there was a chosen transportation.csv file as a dataset. From this dataframe we choose features like Speed and Vehicle type to complete our calculations. We will try to visualize with a box plot and make a frequency table.

Methods

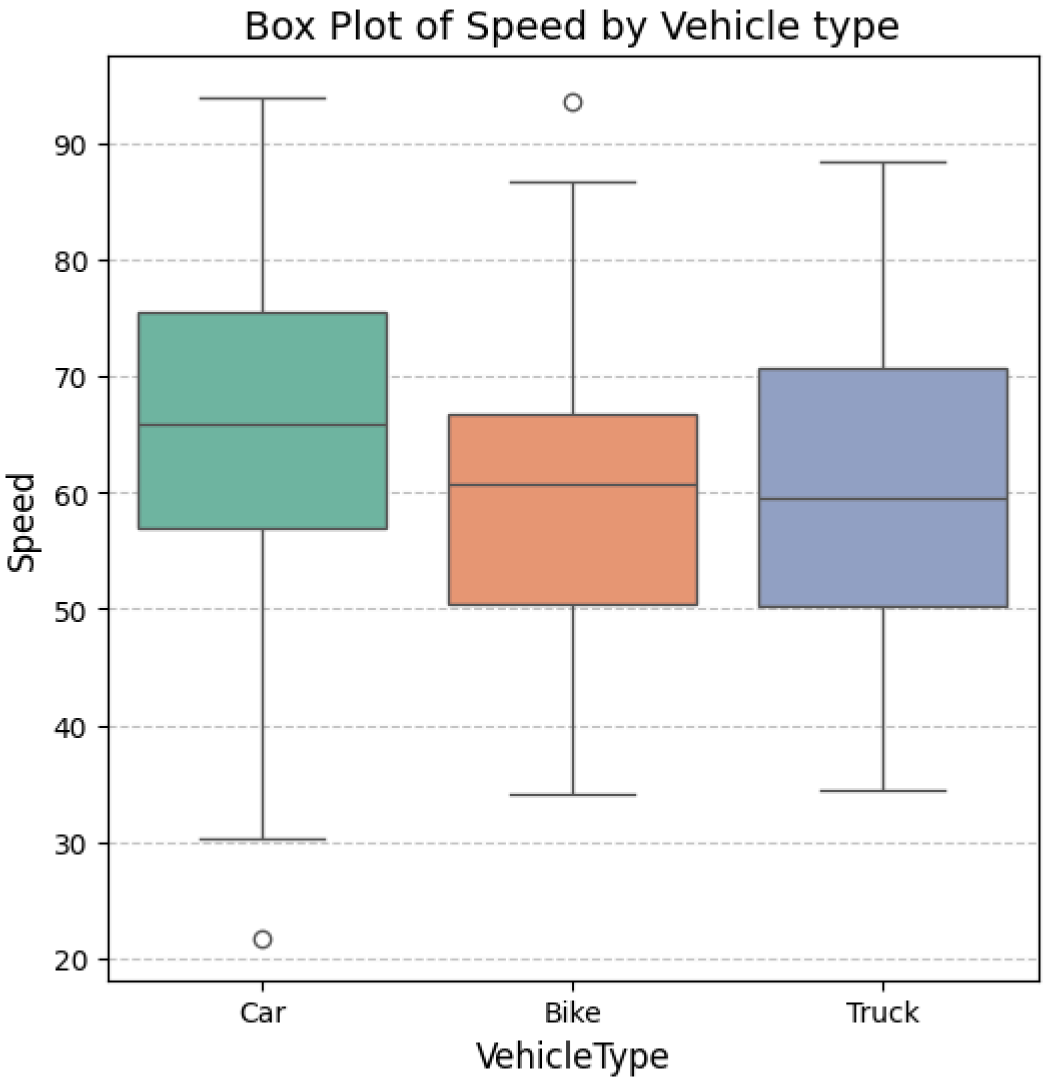
For making a Frequency table there is one method with making intervals and taking it as bins. First of all necessary to define the range of our speed feature, subtracting max from the minimum. then this range needs to divide to the bin size. After, we can add this value to min and we will define the first interval, to find other ones, we just need to just add that value to next ones.

Findings

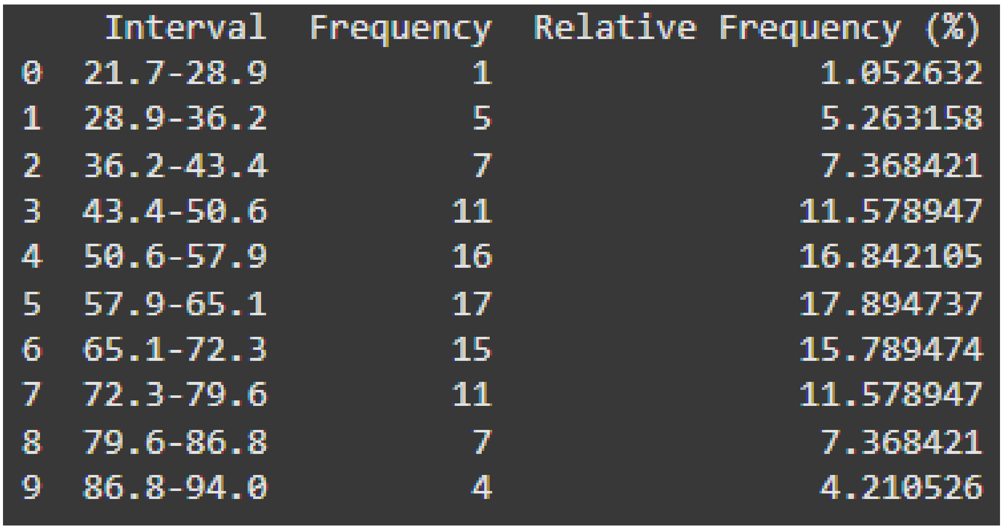
First looks of our box plot looks like this



Looking at this box plot there is clear that our outliers are much higher than normal values and they interfere very much. In this situation, normally we need to trim our data.

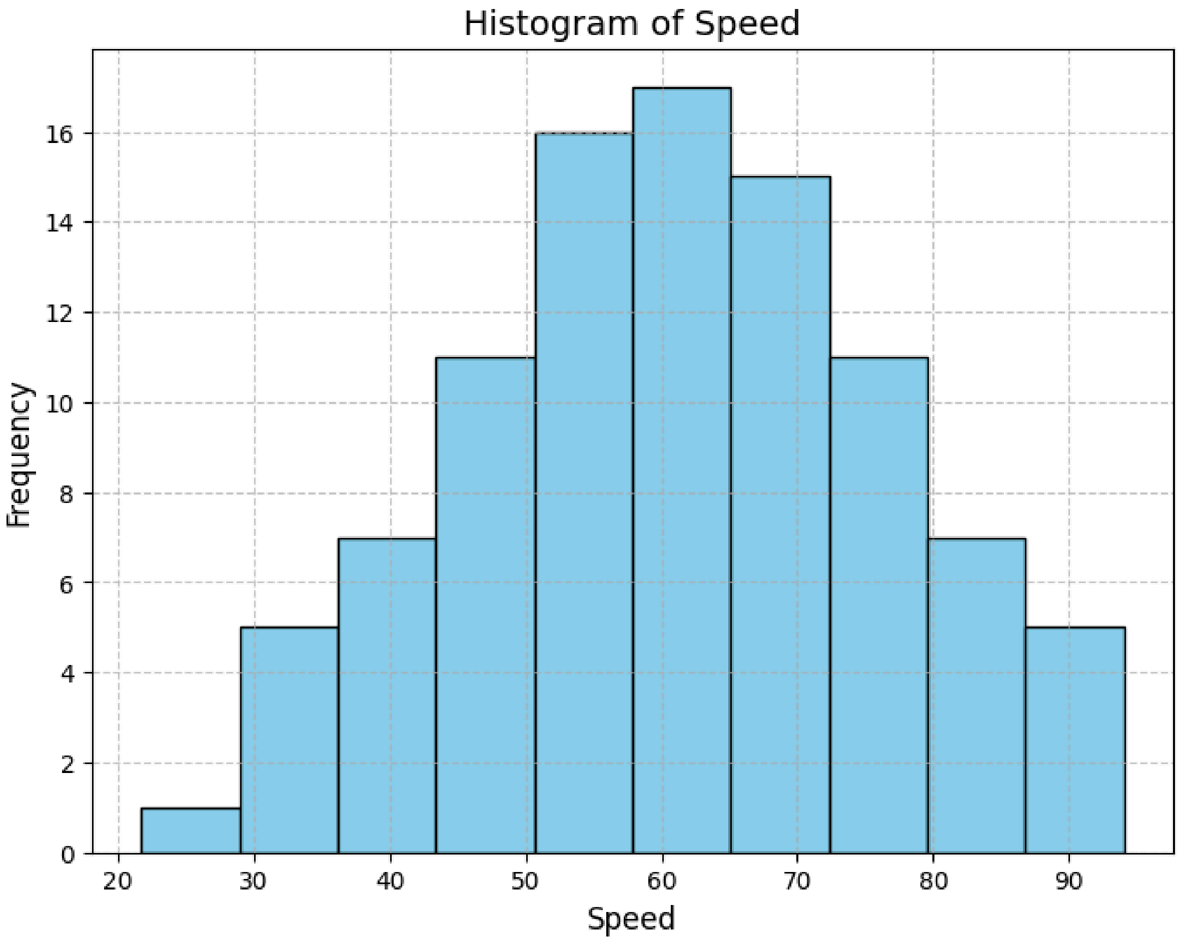


After trimming big outliers, our box plots normalized and then we can work with this dataframe. Let’s have a bin size of 10. The range is 72.34. The division is 7.234. So, now, our first interval starts from min value to min plus division like 21.71 - 28.94. With this formula all 10 intervals looks like this



```
     Interval  Frequency  Relative Frequency (%)
0  21.7-28.9          1                1.052632
1  28.9-36.2          5                5.263158
2  36.2-43.4          7                7.368421
3  43.4-50.6         11               11.578947
4  50.6-57.9         16               16.842105
5  57.9-65.1         17               17.894737
6  65.1-72.3         15               15.789474
7  72.3-79.6         11               11.578947
8  79.6-86.8          7                7.368421
9  86.8-94.0          4                4.210526
```

and the histogram of this frequency table



Looking at this histogram we can define that the data is distributed normally. and most frequencies are in the speed of 60, around 17. The last column's frequency is more than the first ones, it means that only one vehicle drives with around 20-30 speed when the 5 people drive with 90.

Conclusion

To sum up the actions made in this assignment it is clear that the shape of this data distribution is normal, because of most frequencies in the middle.