Big Data Analytics

Nihal Surendra Parchand

np9603@rit.edu

Homework 2

1.

Ethics:

a. I feel that studying traffic volume for road planning to maximize traffic flow is important for the safety of the people who are driving as well as the pedestrians. This maximizes the safety and keeps the reckless drivers in check. I don't have any ethical issues with doing this and I am in full support of this decision.

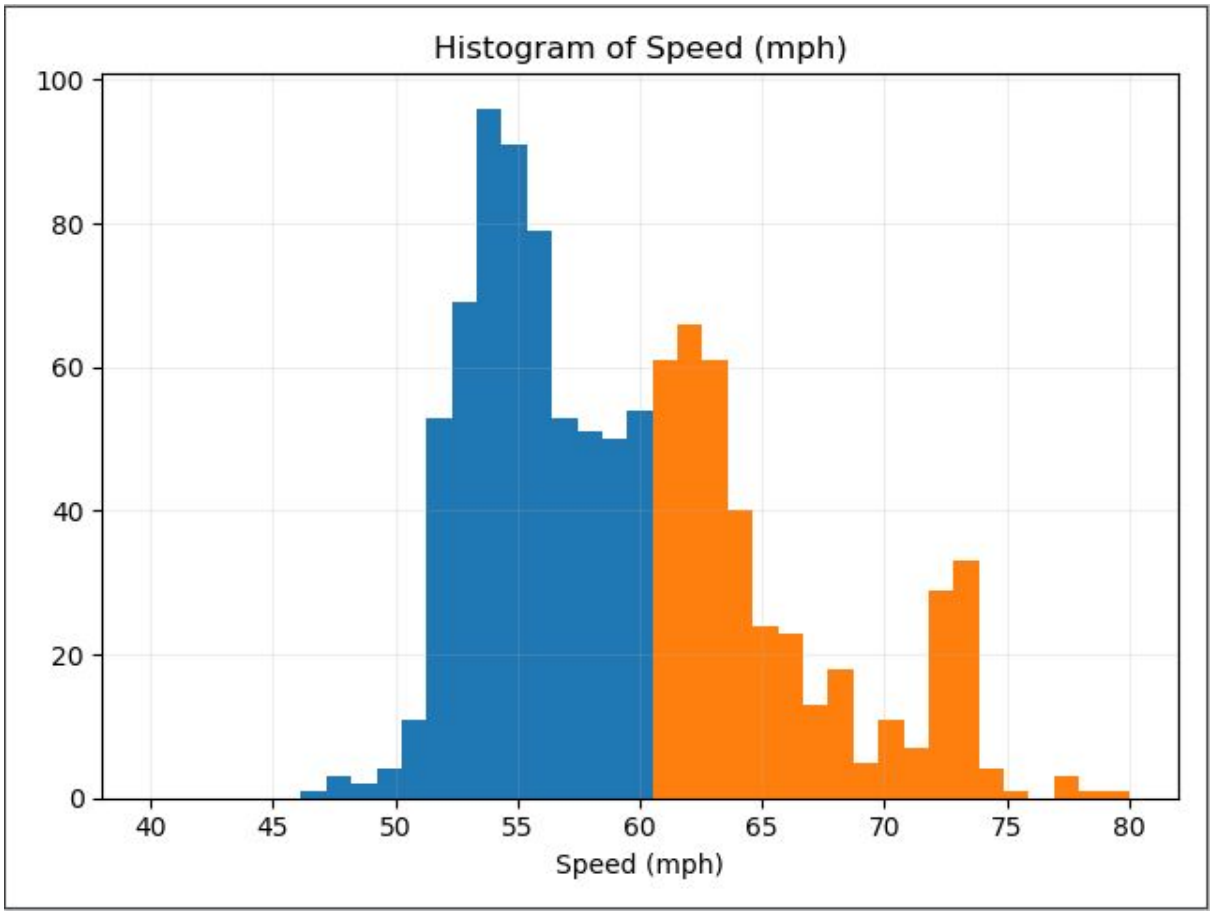
b. The people who are driving recklessly should be given a speeding ticket but it should be a reasonable enough penalty (not too high). Also the threshold limit set for the computer vision machine should be such that it does not pull down the flux rate too much (too less or too high). But there might be some cases where ambulance or police trying to catch a thief also get fined which is incorrect and should be avoided. So a better approach should be devised to tackle this approach.

c. The best speed to separate the two clusters is 60 mph

d. The minimum mixed variance is 11.68532

e. In case of a tie, the minimum mixed variance of lower threshold will be chosen according to my program. No, it does not happen in our situation.

f.
Histogram of quantized vehicle speeds



2.

Exploring Regularization

Cost_Function = Objective_Function + Regularization.

Here, the objective function is the mixed variance calculated from the previous step. We are adding a regularization term to make the two clusters of the same size.

Regularization = abs((Number of Points in Slow Group) – (Number of Points in Fast Group)) / NormFactor * alpha.

Number of Points in Slow Group is the total number of speed observations below the best threshold
Number of Points in Fast Group is the total number of speed observations above the best threshold
NormFactor is set to 50 to make regularization about 4.
Alpha ranges from [100, 1, 1/5, 1/10, 1/20, 1/25, 1/50, and 1/100, 1/1000].

For our program, we did not notice any significant changes in best mixed variance and the best threshold remained the same for every value of alpha. I feel that the regularization is not prominent enough to affect the results of our experiments.

```
Best threshold without regularization:  60
Best mixed variance without regularization:  11.685316829534742
Best threshold with regularization:  60
Best mixed variance with regularization:  11.728516829534742
```

3.
Exploratory Data Analysis

The MysterData.txt consists of pre-quantized values.

a. Original Data

```
Original Data:
Median:  14.5
Mean:  15.775
Mode:  ModeResult(mode=array([7.]), count=array([4]))
Midrange:  22.0
Average:  15.775
Standard Deviation:  8.35310570985427
----------------------------------------------------------------
```

The mode is 7 which occurs 4 times.
The median is 14.5
The mid-range is 22
The average is 15.775
The standard deviation is 8.3531

b. Removing the last value

```
Removing last value of data
Median:  14.0
Mean:  15.76923076923077
Mode:  ModeResult(mode=array([7.]), count=array([4]))
Midrange:  35.0
Average:  15.76923076923077
Standard Deviation:  8.459440299305552
----------------------------------------------------------------
```

The mode is 7 which occurs 4 times.
The median is 14.0
The mid-range is 35
The average is 15.769
The standard deviation is 8.4594

Removing the last 16 values

```
Removing last 16 values of data
Median:  11.0
Mean:  13.75
Mode:  ModeResult(mode=array([7.]), count=array([4]))
Midrange:  33.0
Average:  13.75
Standard Deviation:  7.287260573539735
```

The mode is 7 which occurs 4 times.
The median is 11.0
The mid-range is 33
The average is 13.75
The standard deviation is 7.2873

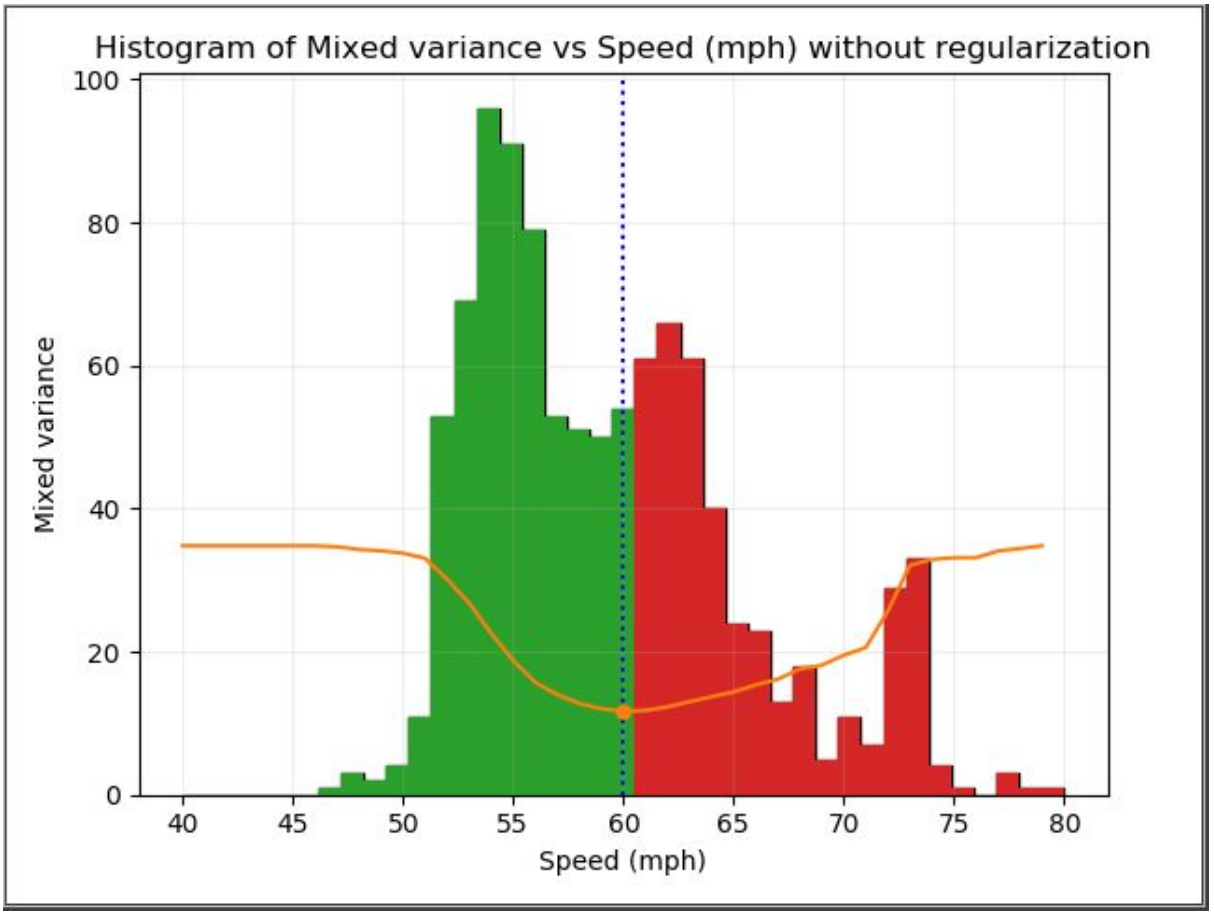
The mode remained unchanged because all the observations of 7 were before the last 16 values. So it remained the same.

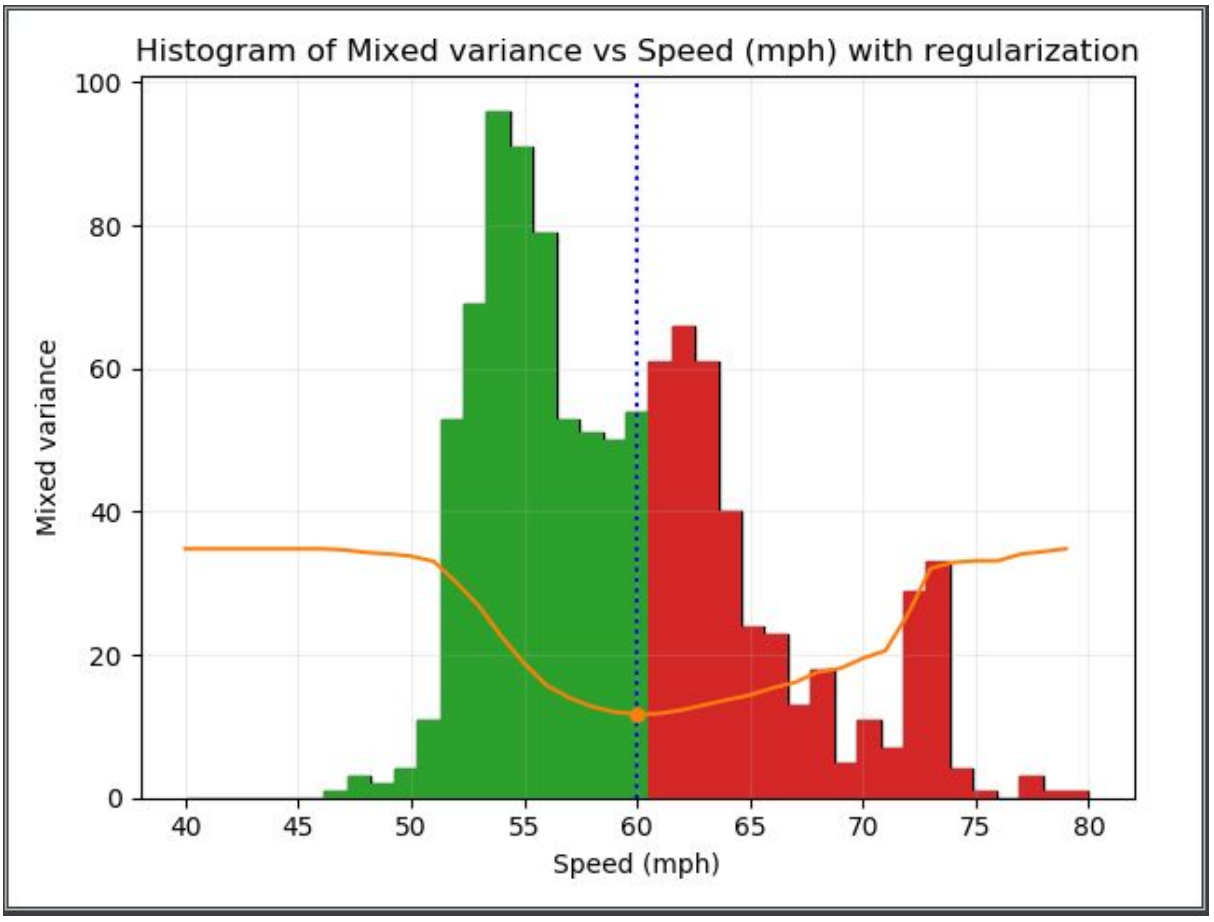
The median value did not change much if we removed only one value but it changed if we removed the last 16 value because at first the median was the 20th value (14) of the sorted data. Then after removing 1 value, it was still the 20th value (14). But after removing the last 16 values, the resulting median is the 12th value (11) of the sorted data.

The average value (mean) did change because we are removing the last 16 values which is getting reduced from the total sum of all observations/total number of observations.

4.

Graphing


Histogram of Mixed variance vs Speed (mph) without regularization
Mixed variance
Speed (mph)
0
20
40
60
80
100
40
45
50
55
60
65
70
75
80



5.

Conclusion and Discussion

In this we learn about the 1D clustering using Otsu's method. We worked on the data which had 1000 observations related to different speeds and then we tried to split the data into two clusters based on the weighted variance. Otsu's method tries to minimize the weighted variance by choosing the best threshold. The results showed that the best mixed variance was ~11 and the best threshold value for splitting the data was 60mph. In case of a tie between two minimum mixed variance, we solved the tie using the lower threshold value. Adding regularization did not impact much on the result. Working on mysterydata gave insights about the statistics related to some data. The mode, median, average(mean) was computed and surprisingly the median was affected more than the mean in our case. Usually, I feel that mean is the value which is most affected because of the outliers or the starting and ending values. Finally, we plotted the graph of speed vs variance for the data with and without regularization. The orange point is the best threshold with the minimum weighted variance. This task explored the clustering part of data analysis and showed that based on different parameters we can divide the data into two clusters. We used binning for quantizing the data into two groups.

6. Bonus

We divided the data into 2 groups by choosing a threshold speed that we calculated based on the weighted variance. The resulting threshold was 60mph. For splitting the data into more than 2 clusters we should choose the data which is below threshold (i.e. 40-60mph). Run the Otsu's method recursively with the threshold values iterating from 40 mph to 60mph and then find the minimum mixed variance and the best threshold speed. This results in 3 clusters.