Pratishta Prakash Rao

# Big data analytics - Homework 2

1.a. We believe that these vehicles are composed of two underlying (latent) groups: those who are intentionally speeding (reckless drivers), and those who are trying to maximize safety, conserve fuel or perhaps waste time.
You are developing a machine with studies traffic volume for road planning in order to maximize traffic flow. Do you have any ethical issues with doing this?
**Solution:**

No. Enforcing speed limits helps maintain safety on the road.

1.b. You are being paid to develop a computer vision machine that will automatically send a hefty speeding ticket and a paint ball at reckless drivers. The paint ball will permanently stain the car's paint. Do your ethical considerations change? If so, why?
**Solution:**

Yes. This method of enforcement causes damage to the individual's car. It would be alright to just send the speeding ticket.

1.c. What speed should we use to best separate the two clusters?
**Solution:**

After using Otsu's method for clustering, the best speed to separate the two cluster is 60 mph.

1.d. What is the minimum mixed variance that resulted?
**Solution:**

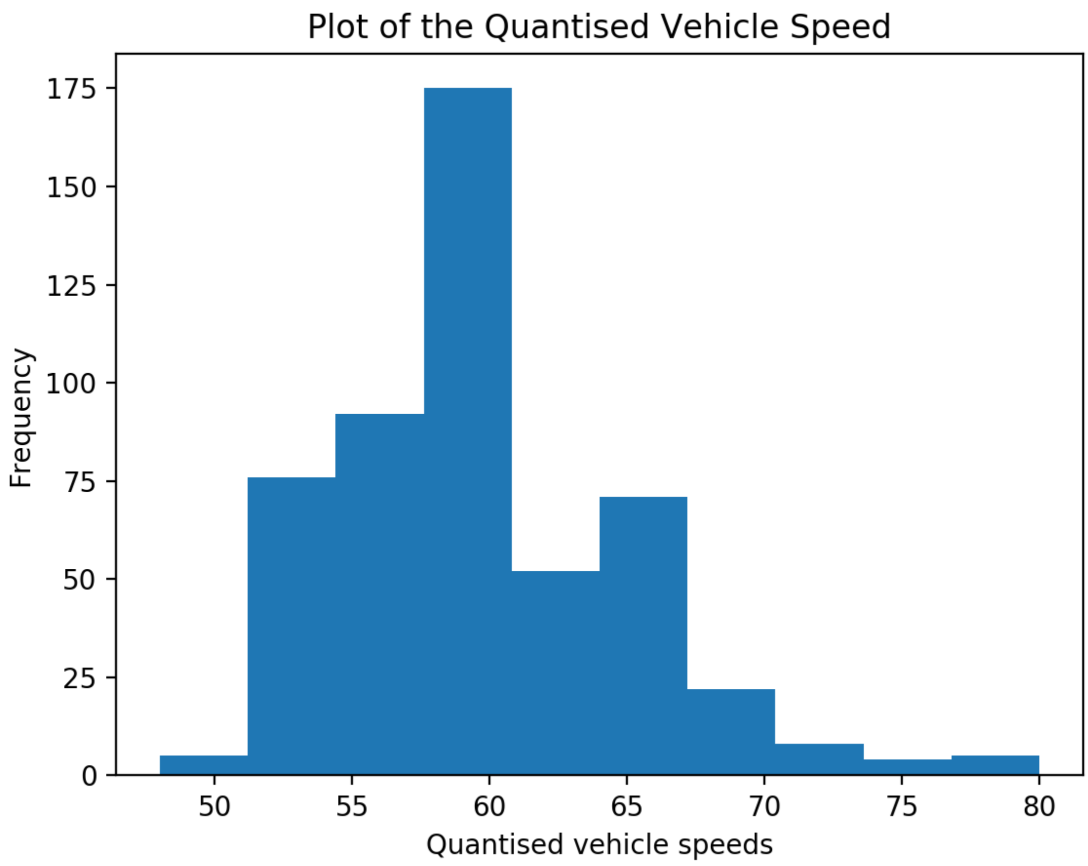
The minimum mixed variance that resulted is 9.515.

1.e. Breaking Ties: How would your program handle a situation where the minimum mixed variance occurred twice? Does this situation happen?
**Solution:**

There have been occasion where the mixed variance was same for many different speeds. The code handles such a situation. The code compares the mixed variance with the best mixed variance, it is only assigned to the best mixed variance when mixed variance is less than the best mixed variance. For cases where the there is a tie, mixed variance doesn't change until that condition is satisfied.

1.f. It might help to plot a histogram of the quantized vehicle speeds.
**Solution:**

The quantized vehicle speeds were plotted and following is the graph obtained:



2.Which settings cause the “best” point to change? What do you notice?

**Solution:**

Otsu’s method with regularization was added. The best settings to cause a change was for the value of alpha = 1. The best speed to split the data into two clusters is 58 mph. With regularization, the value for speed to split the data changed. Also, the split ratio which is calculated by fraction of the length of data points under the threshold by the length of data points over the threshold reduced from 2.148 to 1.048. the split in the data was 261 points in one cluster to 241 points in the other cluster, while for the clustering without the regularization it was 349 points in one cluster to 151 points in the other cluster.

3.a.What are the mode, median, mid-range, average and standard deviation of this data?

**Solution:**

| | |
|---|---|
| Mode | : 7, 8, 16 |
| Median | : 14.5 |
| Variance | : 69.77 |
| Mid Range. | : 22.0 |
| Standard deviation | : 8.3531 |

3.b.Remove the last value from the data, the 16, how do the mode, median, and average values change? Why do you think you observed this?
What caused this amount of change?

**Solution:**

After removing the last element from the data, there wasn't a significant change in the descriptive statistics except for the mode.

| | |
|---|---|
| Mode | : 7, 8 |
| Median | : 14.0 |
| Variance | : 71.56 |
| Mid Range | : 22.0 |
| Standard deviation | : 8.4594 |

The variance increases because of the change in the mean because an element, 16 was removed. In this case the median is now being calculated for odd set of elements .

3.c. Use your Otsu's clustering routine to split this data into two groups. What threshold best splits the data into two groups?What was the minimum mixed variance that resulted?

**Solution:**

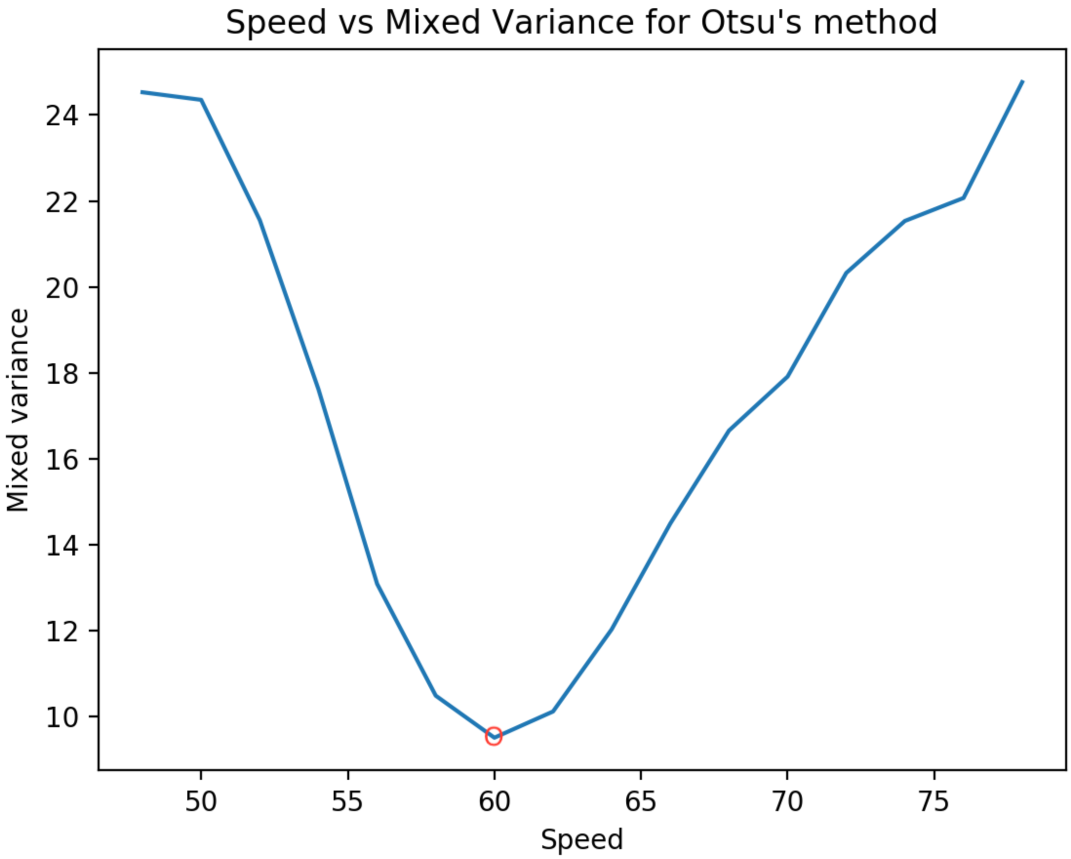
Using the Otsu's clustering to split the data into two groups, the best threshold is to split is 20, the minimum mixed variance that resulted was 21.818

4. Plot a graph of the mixed variance for the car data in question 2, versus the speed.

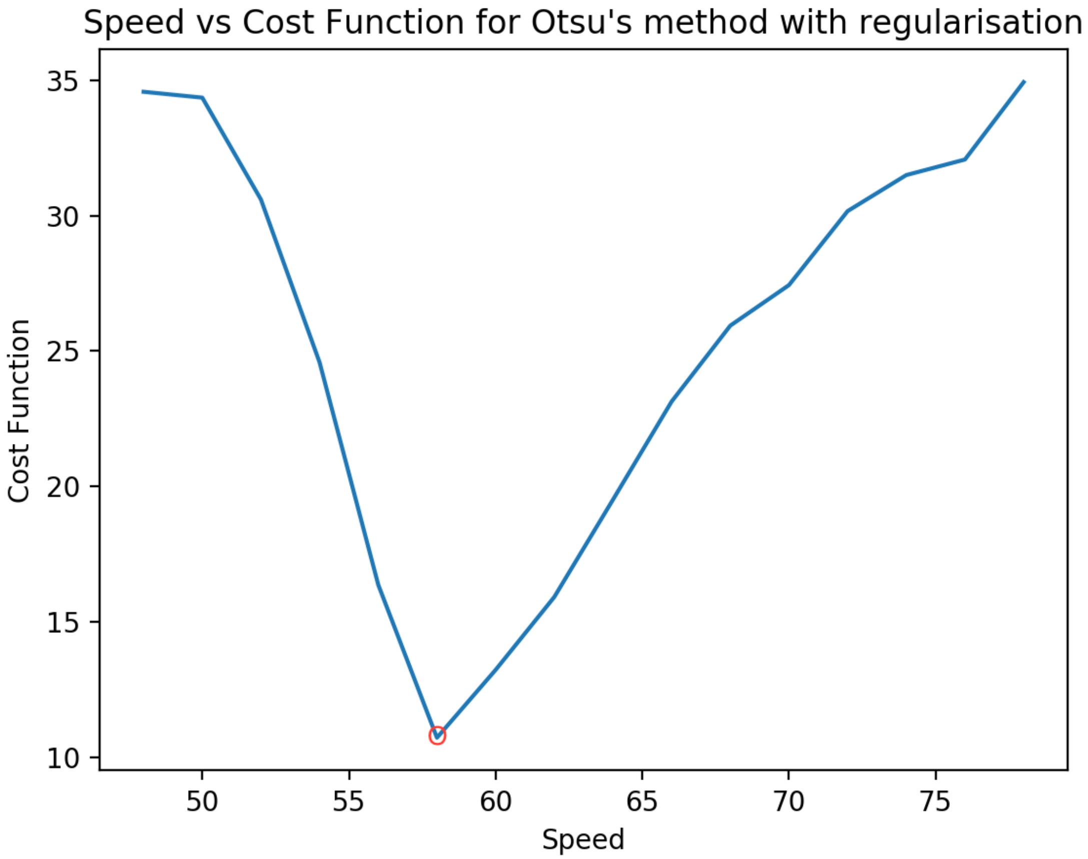
**Solution:**

The graph was plotted for speed vs mixed variance.

For Otsu's method without regularization:



For Otsu’s method with regularization:



5. a. Report how long did it actually take you to do the homework (in hours)? And report the initial guess divided by the actual result to 2 significant figures.

**Solution:**

I estimated to finish my homework in 10 hours. It took me 7 hours to finish the homework.

Ratio: 1.42

5.b. In a few sentences, what factors do you suppose that make it difficult to predict the time it takes to write software?

**Solution:**

- Software installation issues.
- Testing and debugging the code.
- Solving dependency issues between software tools
- Bureaucracy in companies
- Not having the right people for the job. It is necessary to have the people who know the technologies needed in order to maintain a deadline.