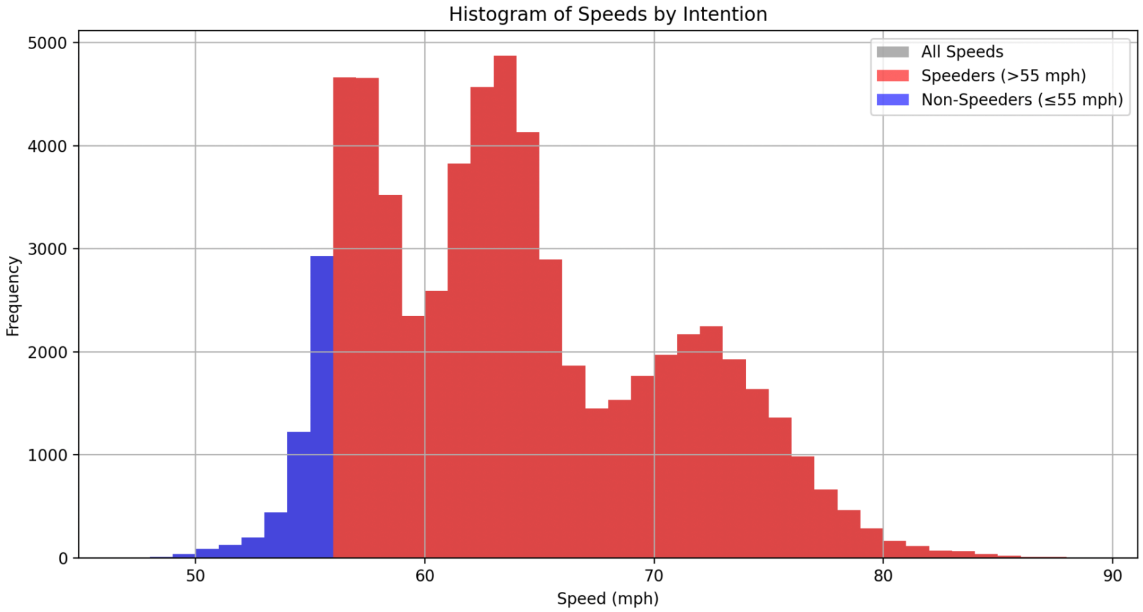
HW_02_ Anna Kurchenko CSCI 420 Data Mining

C.

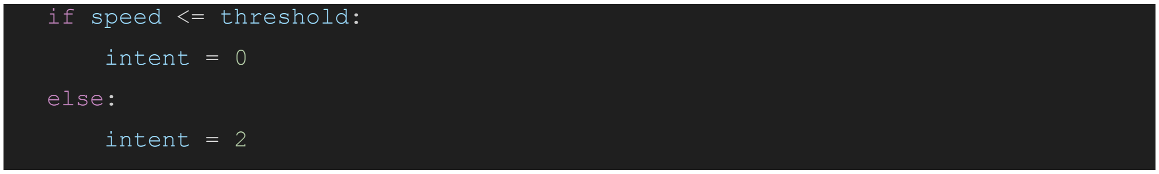
Copy your bar graph of the data in.
Looking at the associated graph of the data, how can we describe this data?
What kind of model is it?



This model almost looks like a trimodal data model, if that is a valid category when the lumps are not of equal size. We can describe this data by saying that most people are speeders, and are ‘aggressive drivers’, since the overwhelming amount of people are in the red zone. It is interesting to note that while its expected that most speeders are only 10 miles over, there is a local maximum of speeders around 75 mph, 20 miles over.

D. In your PDF, show the resulting threshold classifier code.
What was your one-rule? It is just a piece of code. (2)
(Copy and paste from your output classifier file.)

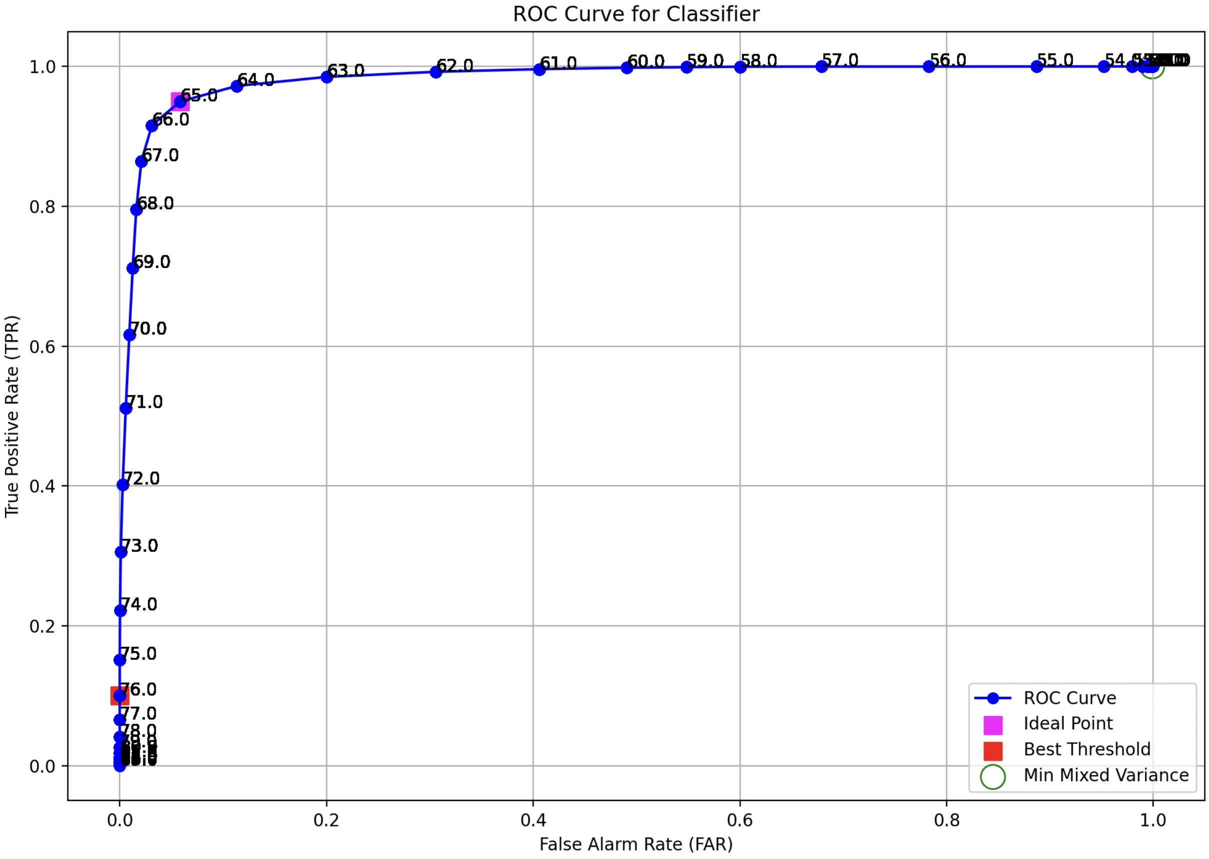
My one rule was the same as in the writeup, I didn't see a need to change it, here's my code:



```
if speed <= threshold:
    intent = 0
else:
    intent = 2
```

E. Plot the ROC curve, as specified. (4)
Show your results in your write up.
Compare the different speeds found:

a. the “ideal”,
b. the one with the lowest mistakes, and
c. the one found which finds the minimum mixed variance (Otsu’s method).
What do you observer for this data?



The ideal speed is at 65, lowest mistakes is at 76, the Otsu’s data one is about 53. From this data I can observe that the lowest mistakes’ speed being at 76 shows that this is where aggressive drivers and non-aggressive drivers are more accurately differentiated. Comparing all the speeds we can see they differ substantially, and a spread like this can mean many types of behavior are counted as being aggressive.

F. Conclusion: Write up what you learned here using at least three paragraphs. (2)
What did you discover? Were the results what you expected? What was surprising?
Was there anything particularly challenging? Did anything go wrong?
Provide strong evidence of learning.
Write a conclusion that describes what you learned in this homework.
Points are taken off for writing with bullet points or checkmarks.

In this homework I learned what TPR and FAR are, and how they differ from mixed variance/ otsu's method to describe data. Otsu's is more purely clustering data on variance and this doesn't necessarily correlate with minimizing mistakes. The ideal threshold calculation was interesting to me as it shows the best location for performance on the curve. However this result isn't practical in the real world.

I was really surprised by how different the results for best threshold, ideal point and mixed variance were, I really expected them to be somewhat grouped in the middle. It made sense that the ideal point was somewhat in the middle, it doesn't really take into account the false positives and mistakes like the best threshold does.

It was kind of challenging for me to incorporate the pieces of the last HW where we defined otsu's method and some of the preliminary exploratory data analysis and seamlessly fit it into this HW. I was a bit confused since using otsu's method for the variance result in the ROC curve wasn't mentioned till the end and I had to go back and rework some things. At first because of this I calculated the variance wrong, and my result was the same exact point as the best threshold value, and that seemed fishy to me. After re-reading the writeup I realized I should be using otsu's method and adjusted my work.