1. 1D Clustering using Otsu's method.

You should get results that are similar to what you saw in the lectures, but not the same values, because you have different data. You are supplied with a set of measurements for a lot of car driving speeds. The data has already been quantized to 0.5 mph.

Ethics: 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.

a.) 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?

Yes, If it is a traffic volume studies then, we shouldn't capture any PII (Personal Identifying Information, including car license plate number or image of the person driving). As long as we work with only with speeds and counts of those speeds for any given time period (not anything specific to a person or group o people) then the work is fine. Basically we can use the data with timestamp to plan the traffic flow. But we shouldn't use any specific ethnicity, race, age, or driver information for this studies.

b.) You are being paid to develop a computer vision machine that will automatically send a speeding ticket to reckless drivers. Now, do your ethical considerations change from your previous answer?

What laws in the US would clash with this? Is there a legal principle you can think of that is an issue?

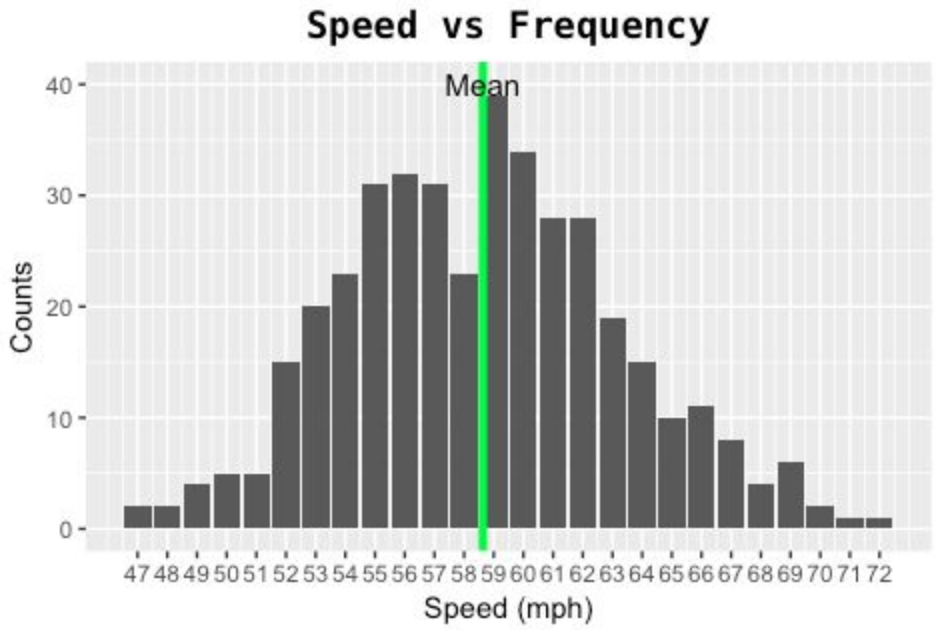
Yes, For an example in this case we may able to capture the license plate number and image of the person who is driving. Because now we are helping our legal authority for identifying reckless drivers. Again in this case, we should be very neutral about the person we are identifying as part of this look up. We cannot specify any other attributes such as sex, race, age or any other information that are not relevant for this context. (the only reason for the image is that there may be more than one person driving the car (or the car registration is for - such as husband and wife), the image could help to identify the right person).

Write a program to:
c. Quantize the vehicle speeds into bins that are [45 up to 46), [46 up to 47), ... up to 75 mph.
Please see the attached code. We have following 26 bins

Speeds	Counts
59	39
63	19
58	23
68	4
61	28
66	11
57	31
62	28
56	32
53	20
54	23
55	31
60	34
52	15
65	10
67	8
48	2
71	1
51	5
64	15
72	1
49	4
70	2
69	6
50	5
47	2

Table 1: With all the bin of speeds and counts

d.) Plot a histogram of these binned values.



Implement Otsu's method to separate the vehicles into two clusters.
e.) What speed did you determine we should use to best separate the two clusters?
Based on Otsu's method we found the cluster center where the we got the minimum total weighted variance from left and right clusters.
The speed for the cluster center is 59 mph.

f.) Breaking Ties: How would your program handle a situation where the same minimum mixed variance occurred twice?
When we have same minimum variance happened more than once then we need to check the relevant values for those minimum variances and reach out to the domain expert for guidance. But in this case, as we concerned about the safety therefore we would compare the values where the minimum weighted variance happened and find the lowest value. (Please refer to the program Line 240)

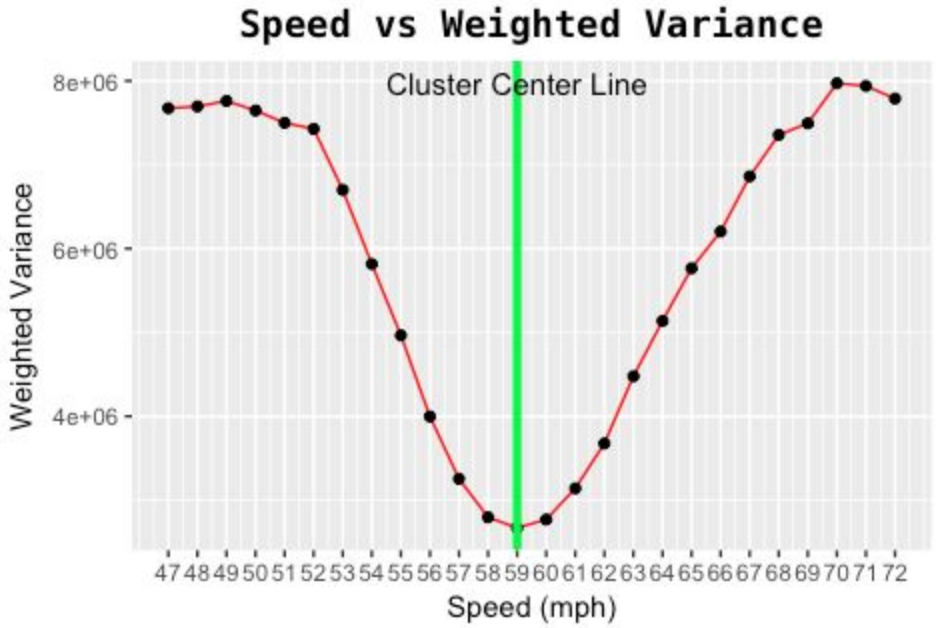
EG:

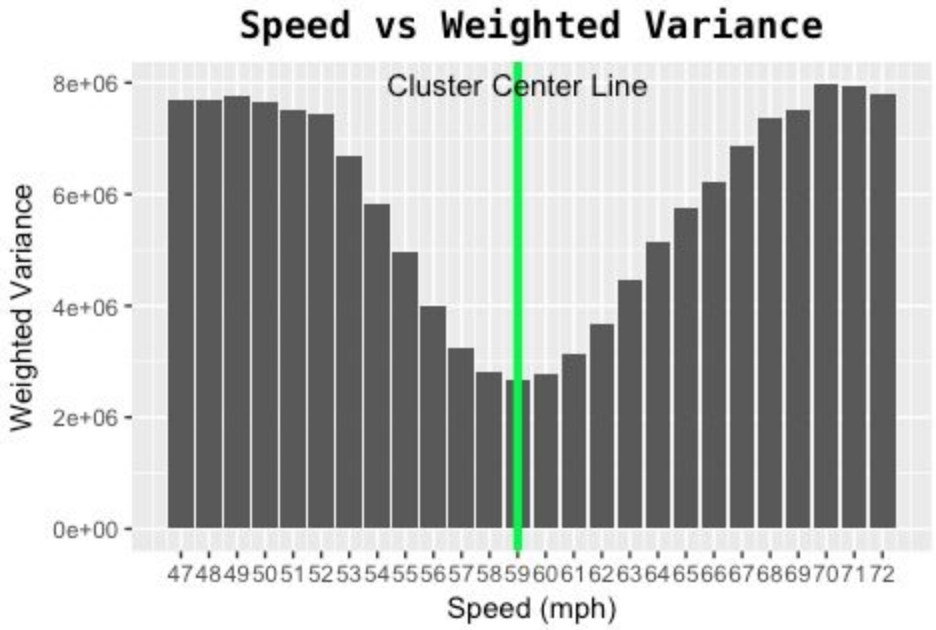
```
# Here the assumption is lower the speed higher the safety therefore if two
speeds get the
# same lowest weight variance then we pick the lower speed from them
  rowIn <- which(d$Total ==min(d$Total))
  clusterCenter <- 0
  if(length(rowIn) >1){
    #Finding the lowest value
    for (i in rowIn){
      if(clusterCenter==0){
        clusterCenter <- d[i, "Val"]
      }else if(clusterCenter > d[i, "Val"]){
        clusterCenter <- d[i, "Val"]
      }
```

```
    }
  }else{
    #The minimum weighted variance happened only once
    clusterCenter <- d[rowIn, "Val"]
  }
```

2.) The objective Function:

Plot a graph of the mixed variance for the car data in question 2, versus the value used to segment the data into two clusters. Clearly label the axes.



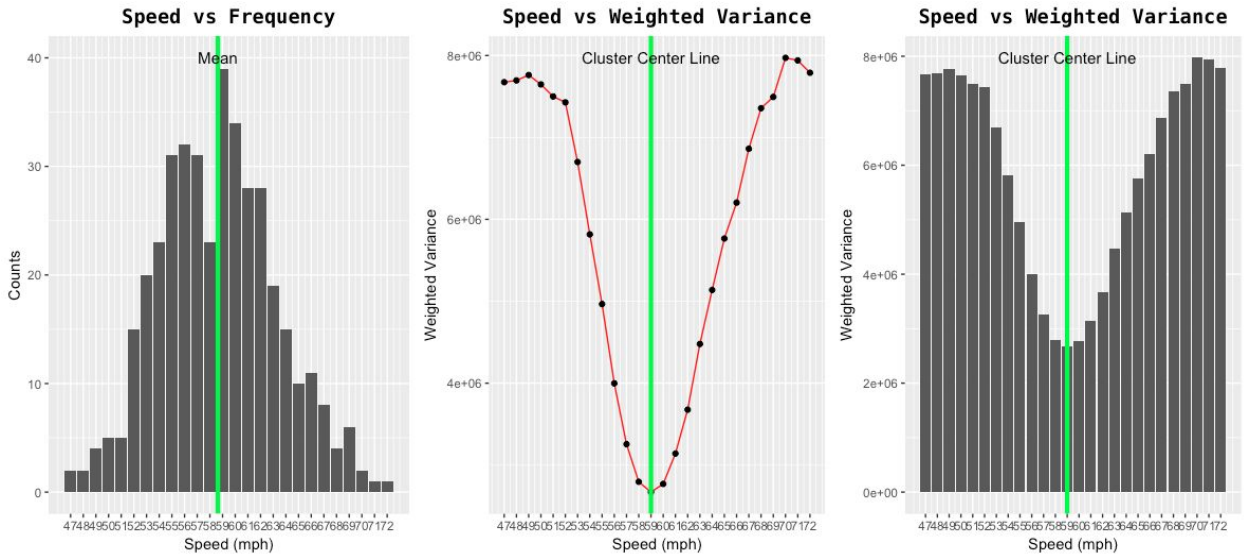


3) Write a concise overview with a conclusion.
What did you learn in this process?
From this exercise we learned how to do a clustering and how to define an objective function. Also how different statistical parameters we learned from childhood (mean, standard deviation, and variance) is useful in the real world problems. This also taught me how we can do clustering in one dimension.

Did anything go wrong?
Not that I know of. Only thing is when you want to view multiple plots with all the coordinates in x axis, it is hard to read the values. Therefore I plot them individually and I also created all of them next to each other for the instructor to view the plot at once.

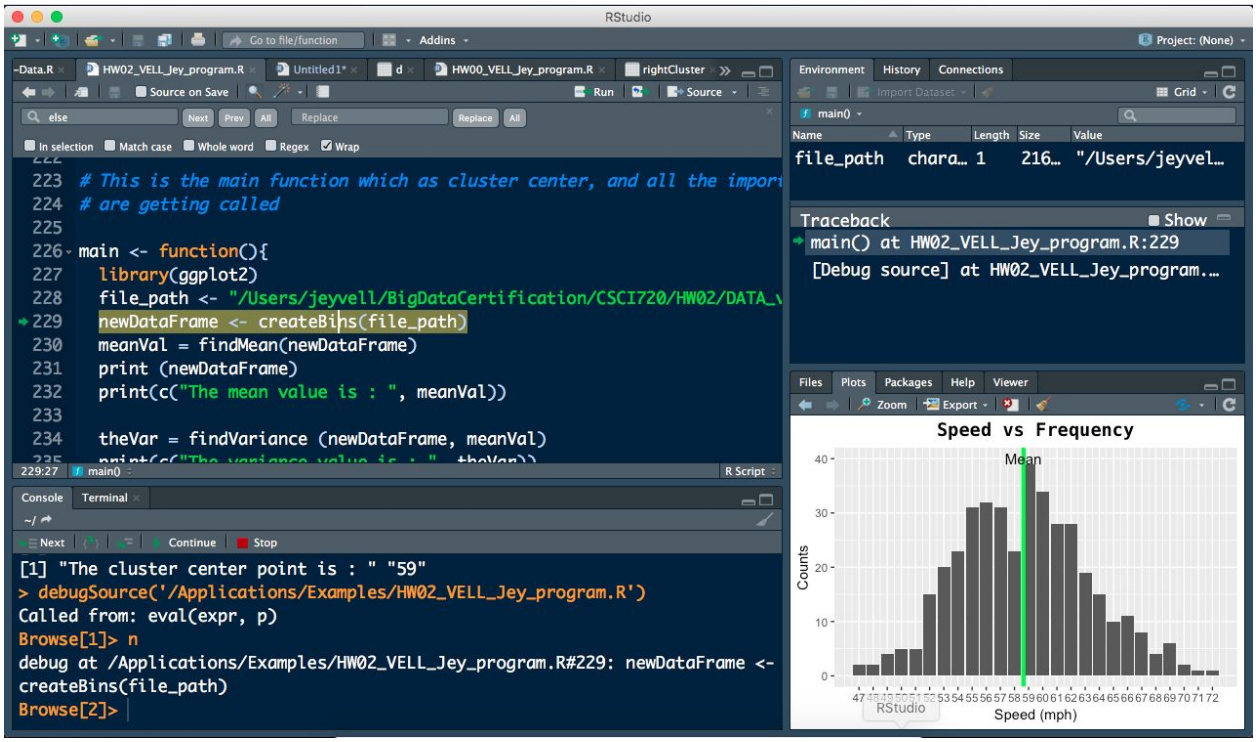
In the code I commented the “plot” function if you want to view individual plots then please comment the grid.arrange(p1, p2, p3, ncol = 3) - Line 271
And uncomment following line and substitute, p2, p3 to see other plots

#plot(p1) - Line 272



Were there any unresolved issues?
Even in this case we know that we are going to get two clusters one is for “safe drivers” and other one is for “reckless drivers”. Even here we know that we are going to get two clusters. But my question is, suppose if we have a dataset with multiple clusters (we have no idea about the structure). If so how do we even know how many clusters are exist in that data? Should we always reach out to an domain expert for advice. What if there is no SMEs around. For an example when we get a network traffic we can certainly identify a packet as harmless or harmful but how do we cluster if that packet is a totally new kind of botnet?

Appendix A


RStudio
Go to file/function
Addins
Project: (None)
HW02_VELL_Jey_program.R
Untitled1*
HW00_VELL_Jey_program.R
rightCluster
Source on Save
Run
Source
# This is the main function which as cluster center, and all the impor
# are getting called
main <- function(){
library(ggplot2)
file_path <- "/Users/jeyvell/BigDataCertification/CSCI720/HW02/DATA_
newDataFrame <- createBins(file_path)
meanVal = findMean(newDataFrame)
print (newDataFrame)
print(c("The mean value is : ", meanVal))
theVar = findVariance (newDataFrame, meanVal)
Environment
History
Connections
Import Dataset
Grid
main()
Name
Type
Length
Size
Value
file_path
chara…
1
216…
"/Users/jeyvel…
Traceback
Show
main() at HW02_VELL_Jey_program.R:229
[Debug source] at HW02_VELL_Jey_program.…
Files
Plots
Packages
Help
Viewer
Zoom
Export
Speed vs Frequency
Mean
Counts
Speed (mph)
Console
Terminal
Next
Continue
Stop
[1] "The cluster center point is : " "59"
> debugSource('/Applications/Examples/HW02_VELL_Jey_program.R')
Called from: eval(expr, p)
Browse[1]> n
debug at /Applications/Examples/HW02_VELL_Jey_program.R#229: newDataFrame <-
createBins(file_path)
Browse[2]>