Tim Miller

15.071 | HW5

**Problem 1a**

To understand the importance of variables we summed across all columns. The columns with the highest values were the ones that were most important to the most customers.

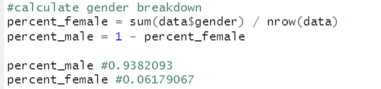




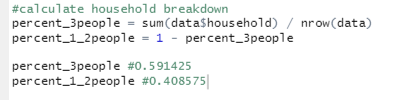
Excluding the gender and household columns, we find the ranking of the features from most important to least important (based on total number of “important” votes received)

* driving\_properties
* power
* technology
* reliability
* sporty
* safety
* comfort
* handling
* consumption
* interior

Then looking at the gender and household columns, we find that the survey consisted of ~94% males and ~6% females.

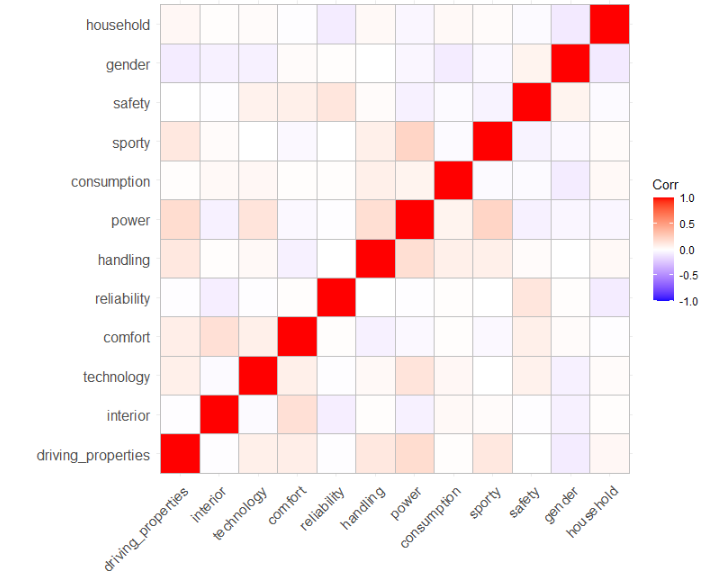


We find that ~60% of respondents have at least 3 people and the remaining have 1-2 people.



**Problem 1b**





There are a few interesting relationships in the data:

* driving\_properties is has a positive relationship with technology, comfort, handling, power, and sporty. This makes sense because the latter three are typical inputs into overall driving performance
* safety and reliability are correlated, which is intuitive. One might expect a reliable car also to be safe.
* sporty and power have a strong correlation. Again this is intuitive: a sporty car probably is powerful.
* Interestingly there is a negative correlation across gender and most of the variables. Something to continue to explore.

**Problem 1c**

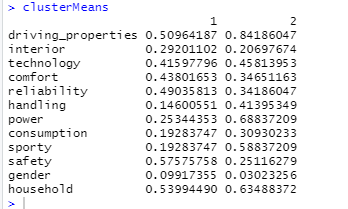
Normalizing data is most helpful when there are scaling issues across the data we are comparing. Oftne the issue is that distance between points we are analyzing is driven by variables with the larger scale, which skews the data. But since the data is all binary here, we do not have a scaling problem, therefore fine if we do not normalize.

**Problem 1d**

We use the following to calculate cluster size, finding 363 respondents in group 1 and 430 respondents in group 2.



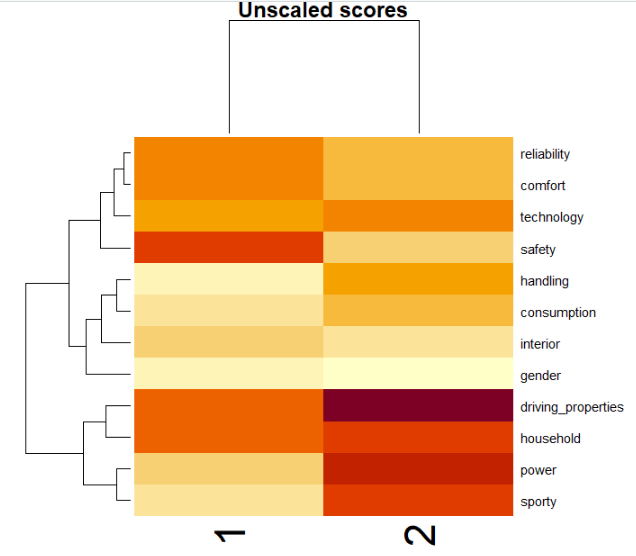
Then, printing out the summary of the clustersMeans



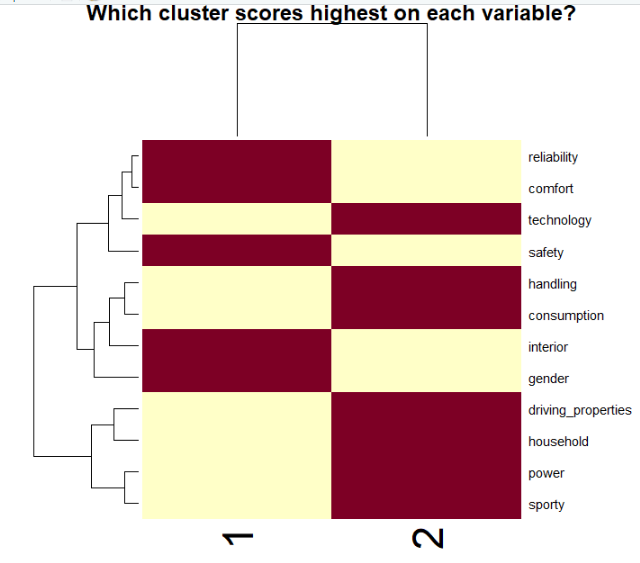
We find a few interesting takeaways:

* Cluster 1 has a slightly percentage of female respondents (higher mean value)
* Cluster 2 has slightly higher percentage of households with at least 3 people
* Cluster 2 puts more importance on driving\_properties, power, sporty, consumption
* Cluster 1 puts more importance on reliability and safety
* Relatively equal importance for interior, technology, comfort

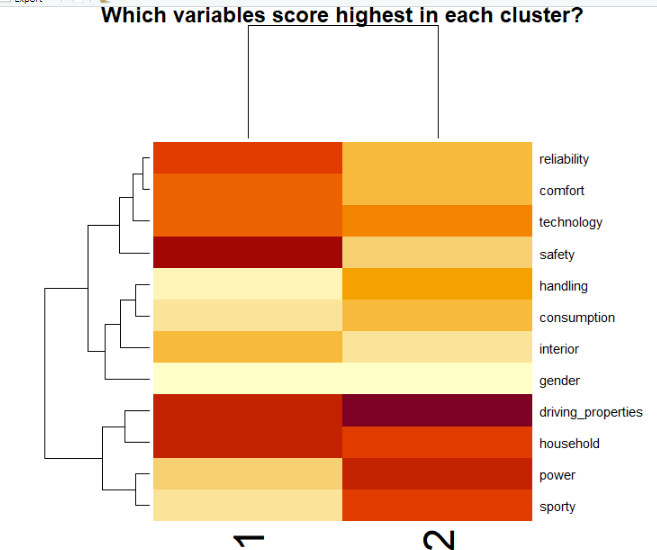












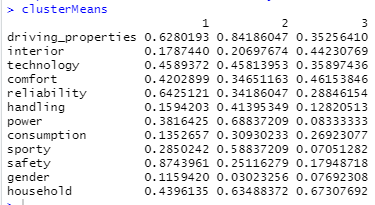
**Problem 1e**

We use the following to calculate cluster size, finding 207 respondents in group 1 and 430 respondents in group 2, and 156 respondents in group 3.



From this, group 2 in problem 1d is clearly the summation of groups 1 and 3 in problem 1e. Cluster 2 in this problem is the same as the one in problem 1d.

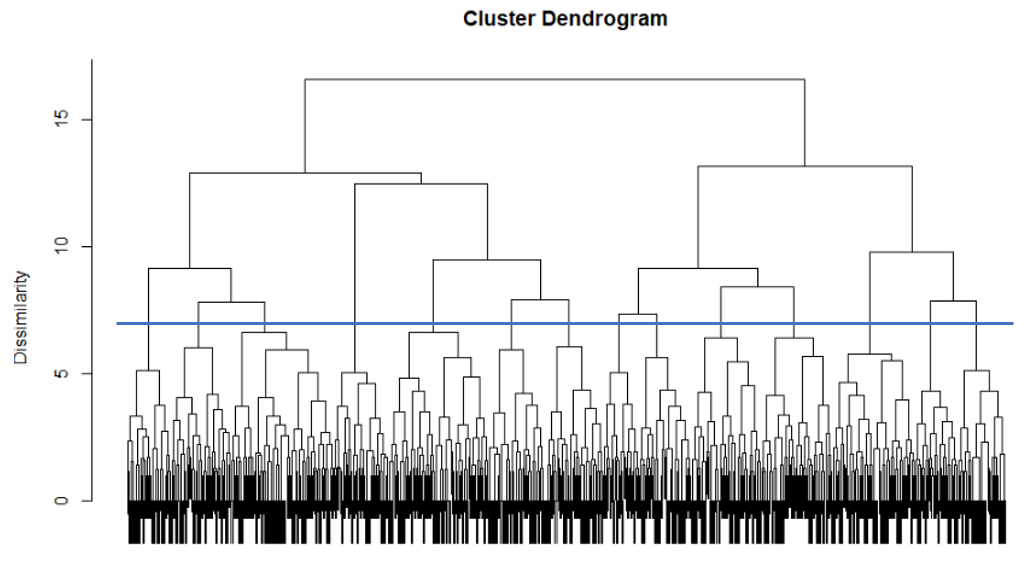
We use the following to analyze the clusters



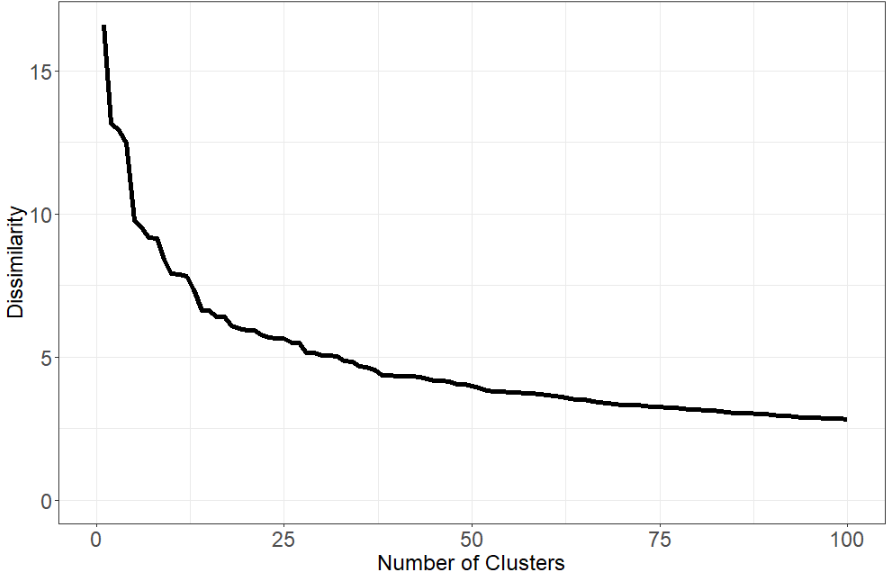
* Of the three clusters, cluster 1 has the higher portion of females, cluster 3 has second highest, and cluster 2 still has the lowest
* Cluster 1 cares more about driving\_properties, sporty, and power relative to cluster 3, but both clusters care less about those properties relative to cluster 2
* Relative to cluster 1, cluster 3 cares much more about interior and slightly more about comfort
* Cluster 1 cares much more about safety compared to the other two clusters

**Problem 1f**

Dendogram



Scree plot

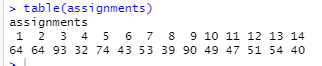


Looking at both the scree plot and the dendogram, my suggestion is we use 14 clusters. At 14 clusters, there is a reasonable amount of dissimilarity, but it is not significantly higher than the dissimilarity at 100 clusters. Furthermore, 14 clusters represents a reasonable number of clusters to work on without worrying too much about overfitting the clusters.

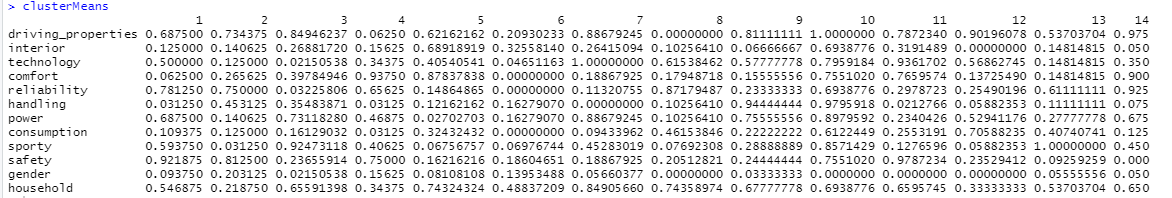
*The scree plot above shows the blue line drawn across at 14 clusters.*

**Problem 1g**

Number of respondents in each cluster below. Looks like clusters 3, 5, and 9 have larger number of respondents compared to other clusters.

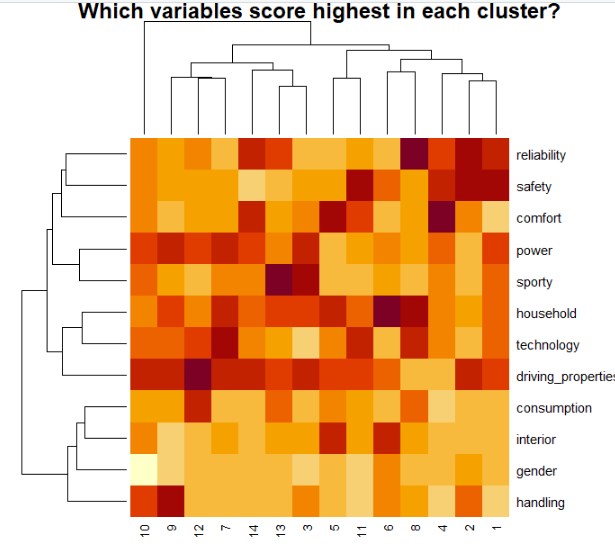


Cutting down the tree to suggested 14 clusters



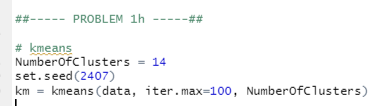
A few observations

* Clusters 8, 10 – 12 are all male
* No one in cluster 8 rated driving\_properties as important, while everyone in clusters 10 – and most people in clusters 7, 9, 12, and 14 – rated driving\_properties as important
* Interestingly, cluster 10 has some of the highest scores across all of the features. This might be a cause for checking that data to see if those respondents just answered 1 for everything and might not have provided useful data.
* From the heatmap (below) we can see clusters 1, 2, and 4 really care about reliability; across all clusters driving\_properties is often viewed as more important relative to other features; clusters 9 and 10 care more about handling – relative to other clusters – than other features; clusters 10, 9, 12, 7, and 14 care more about power – relative to other clusters – than other features



**Problem 1h**

Set up the k means clustering



Using km.size, we get the breakdown of the number of data points in each cluster. The below numbers are ordered from cluster 1 – 14 accordingly

