Katie Baerveldt

MK 6460

Spring B 2020

Assignment 1

**Part A: Market Basket Analysis – Online Retail data**

Part A of this report focuses on a large consumption/purchase dataset that includes several important variables for analysis

There was quite a lot of data missing, this issue wasn’t specific to one feature, so I replaced missing values with “NA” to move forward as to not discount any otherwise-complete records

Concerning the variables themselves;

* it appears that InvoiceNo acts as our ‘transaction’ feature (as we know from our class 3 practice),
* our StockCode and Description go hand-in-hand with identifying the item,
* Quantity refers to the number of that particular item purchased in one transaction,
* InvoiceDate records time series data of the transaction occurrence,
* UnitPrice shows item individual cost,
* CustomerID shows the individual consumer,
* and Country shows the country of origin for the transaction occurrence

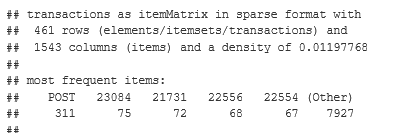
I decided to subset by France and Spain since this data set was so large

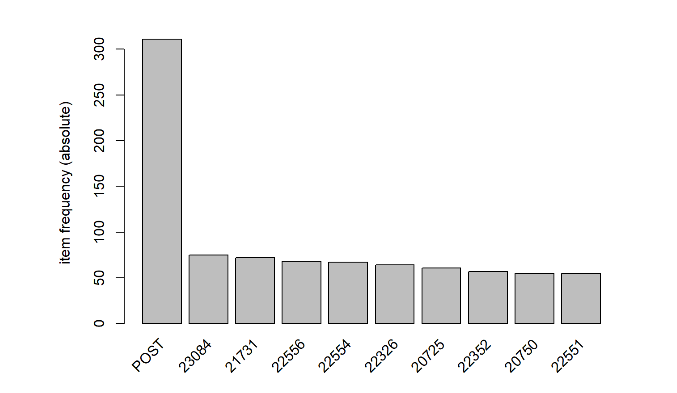
I chose to compare Spain and France since both countries are close in proximity, are a part of the EU, and seem to have a similar population size

I find it interesting that the most frequent transaction between both countries includes postage

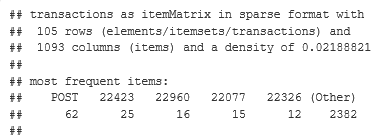
I conducted a majority of my analysis with StockCode instead of Description, so the output was cleaner

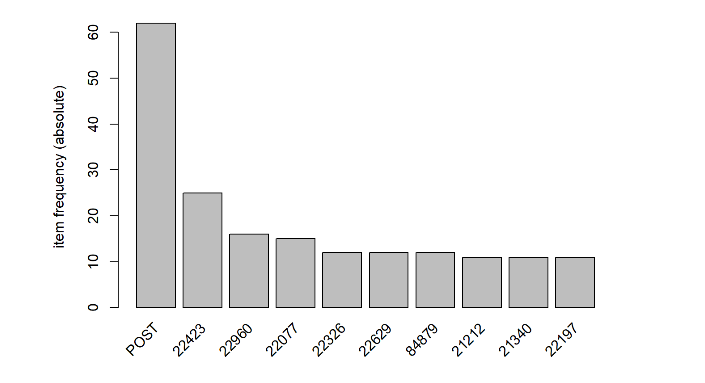
Summary of France transactions:





Summary of Spain transactions:

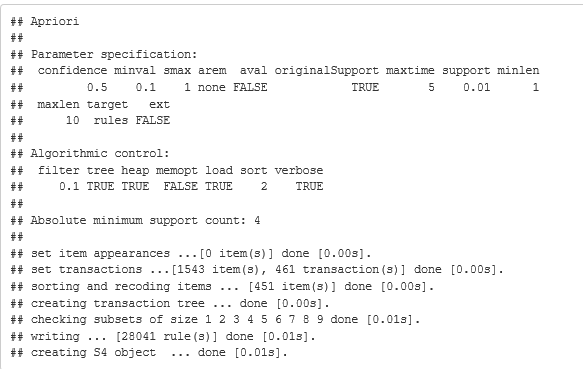




Based on item frequency and summaries of each subset, postage was the most common item in a transaction by far in both countries

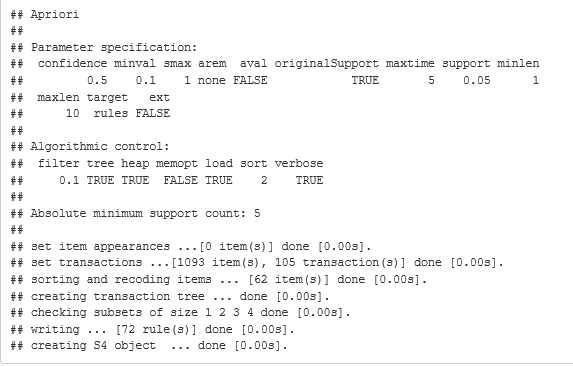
MBA France:

Confidence had to be increased to produce a usable analysis



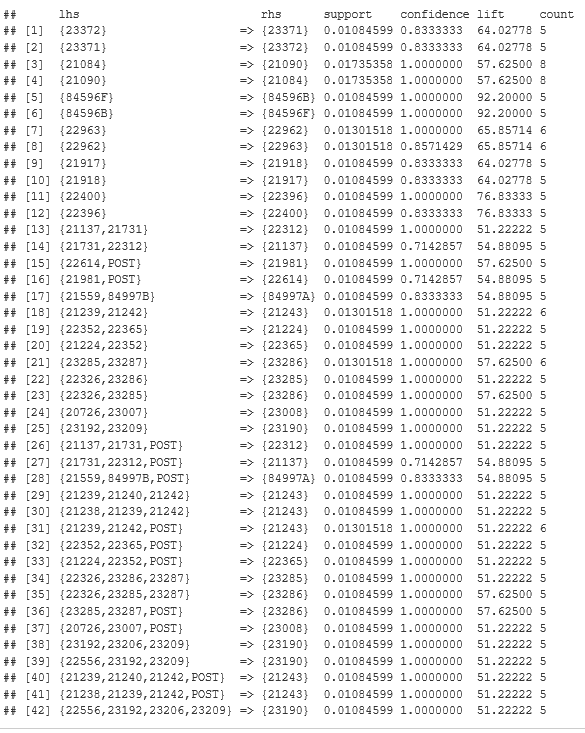
MBA Spain:

Increasing support was useful in contrast from the France subset

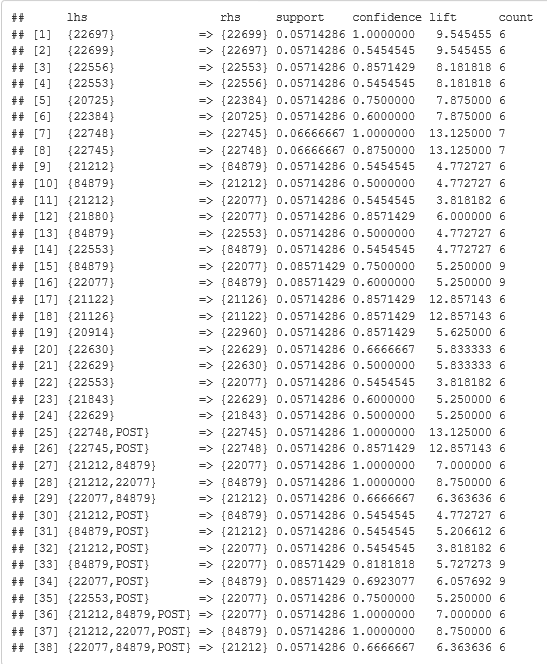


Exploring strength of relationships between itemsets using lift:

France



Spain



My data for France is much more profound, so I used a higher lift than I did with Spain

For France, the purchase of 84596F (SMALL MARSHMELLOW PINK BOWL) led to the high lift of a purchase of 84596B (SMALL DOLLY MIX DESIGN ORANGE BOWL) in the same transaction, AND vice versa for this itemset

For Spain, the purchase of 22748 (POPPY’S PLAYHOUSE KITCHEN) led to the high lift of a purchase of 22745 (POPPY’S PLAYHOUSE BEDROOM) in the same transaction, AND vice versa for this itemset

There appears to be more transaction data available for France rather than Spain, so any further analysis would need to adjust for this when conducting comparisons

The most frequently purchased item overall was postage, so further analysis would absolutely revolve around the many relationships this item has with others by country - however, postage was only a common rhs for France while it’s virtually nonexistent in the lift data I mined for Spain

The bowls mentioned previously for France seem to be complimentary goods based on their high lift values, so if not already I would explore how these can be marketed together - the same goes for the Poppy’s playhouse series for Spain, except for these I would further explore other items with “Poppy” in their description to enhance the buying power of these products

**Part B: Cluster Analysis – Insurance Policy data**

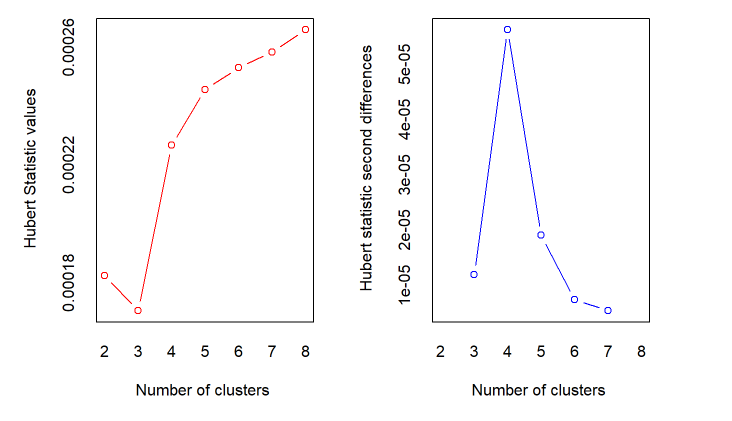
Part B of this report focuses on how an insurance firm can use company data combined with feedback data in order to better cater to its customers

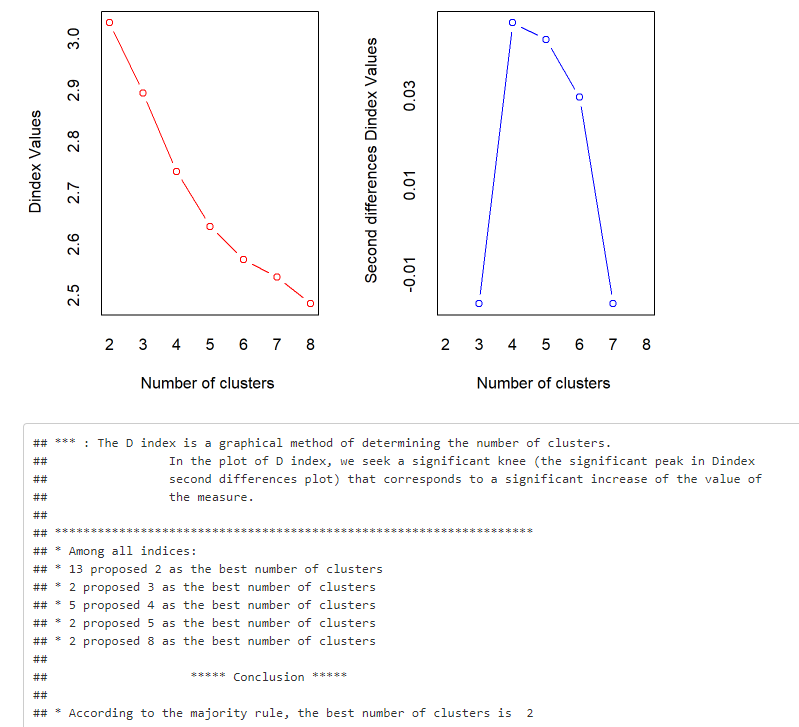
The plan is to identify appropriate segments in their customer pool

At first glance, the ‘Over 18’ feature and ‘Employee ID’ features didn’t offer much in terms of analysis since ‘Over 18’ only had one level in the data, and ‘Employee ID’ acted as a primary key

The first segmentation analysis was done by only focusing on numeric data

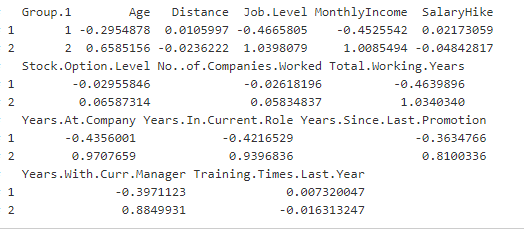
KMN Method:

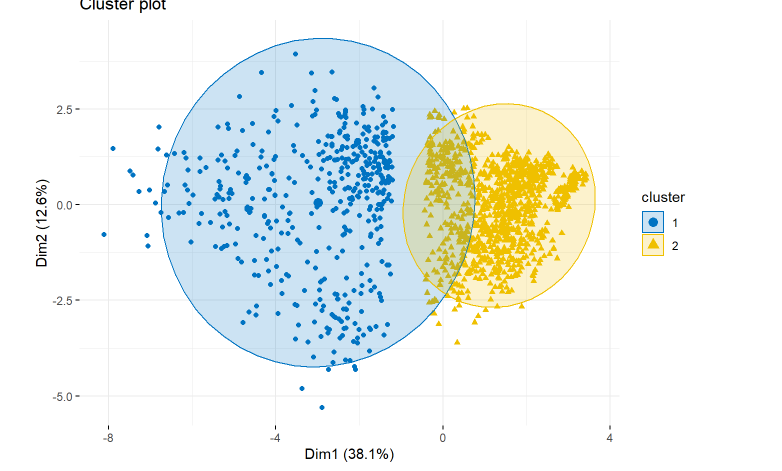




The optimal number of clusters using this method was 2, so I moved forward to analyze the breakdown of said clusters





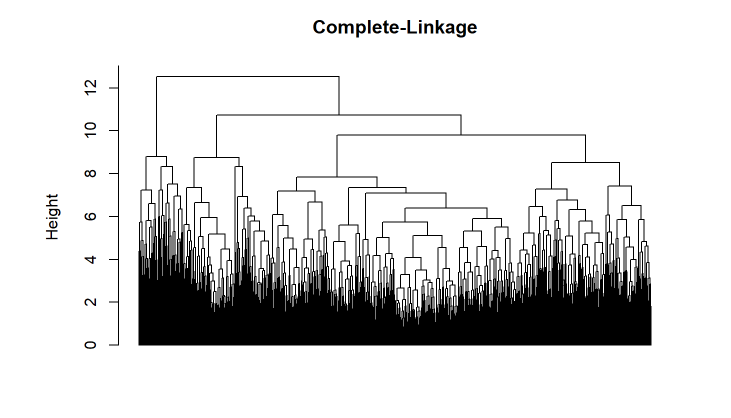


KMN Cluster 1 appears to be made up of a younger demographic with a much smaller income level, which is reflected throughout the remaining variables (i.e. less jobs held in their lives, less years at company, smaller # working years…)

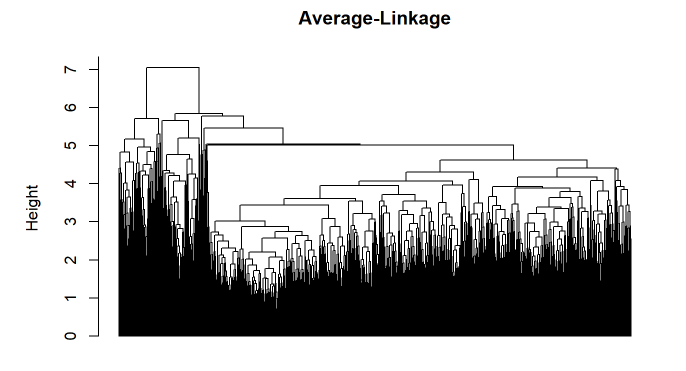
KMN Cluster 2 appears to be more top heavy and made up of older, senior-level executives who have stagnated in their likely senior-level positions, reflected in the significantly higher average salary and smaller cluster portion

HC Method:

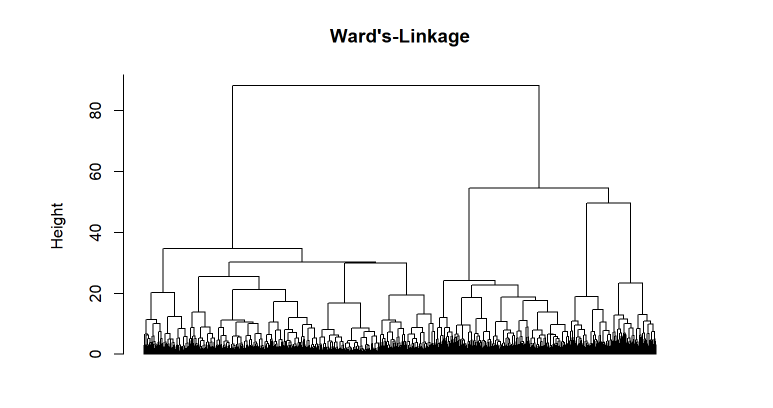
For the HC method there was more freedom to attempt different linkage options as I was testing for optimal number of clusters





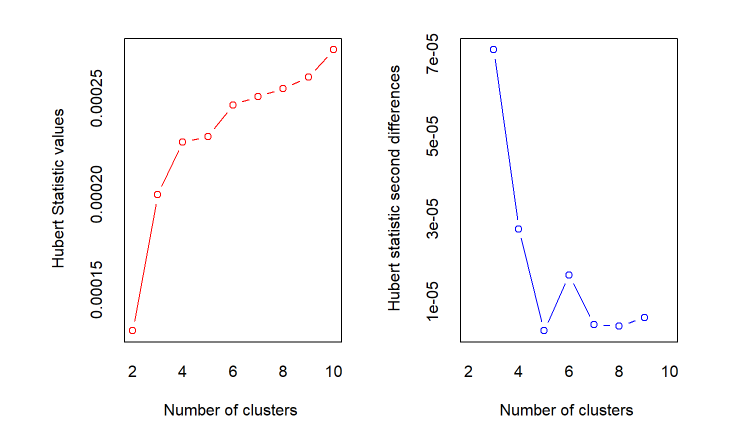


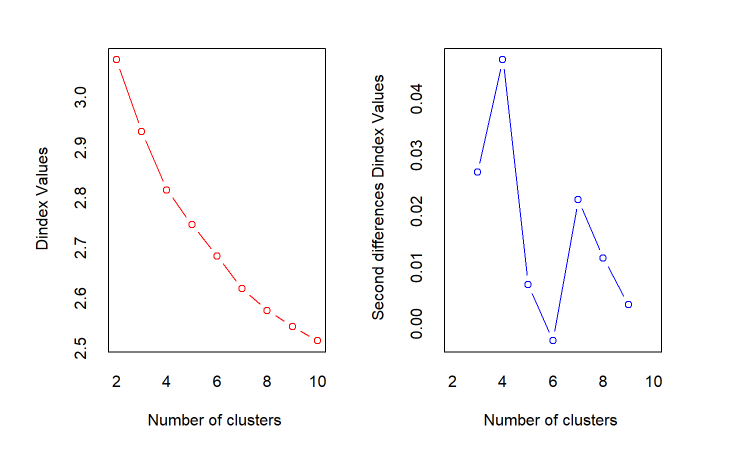


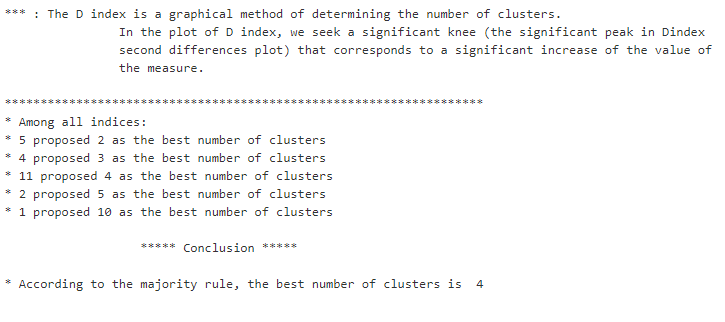




Ward’s linkage option is the cleanest approach for my HC method, it cuts the data into usable clusters (seen above in most recent example)

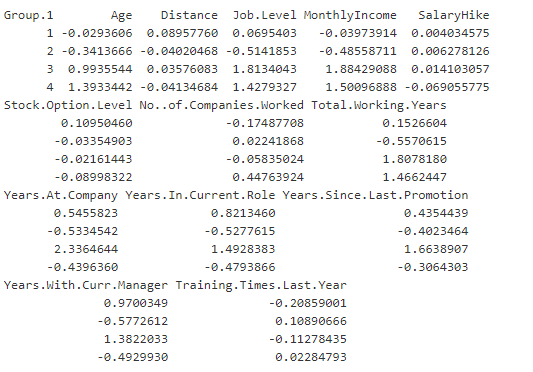


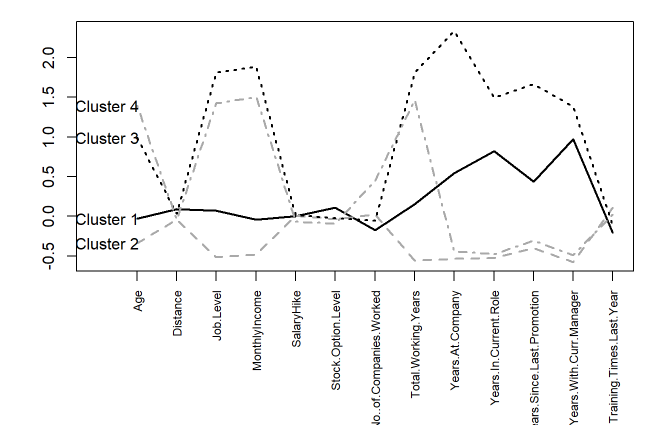




This contradicts my earlier findings of # of optimal clusters for the HC method, so I’ll reevaluate with 4 clusters and break it down (below)







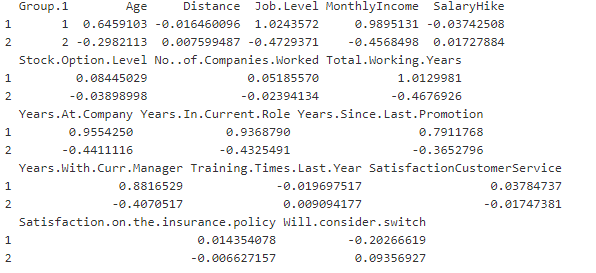
If it wasn’t apparent before, the variables where the hikes take place show where the clusters are separated

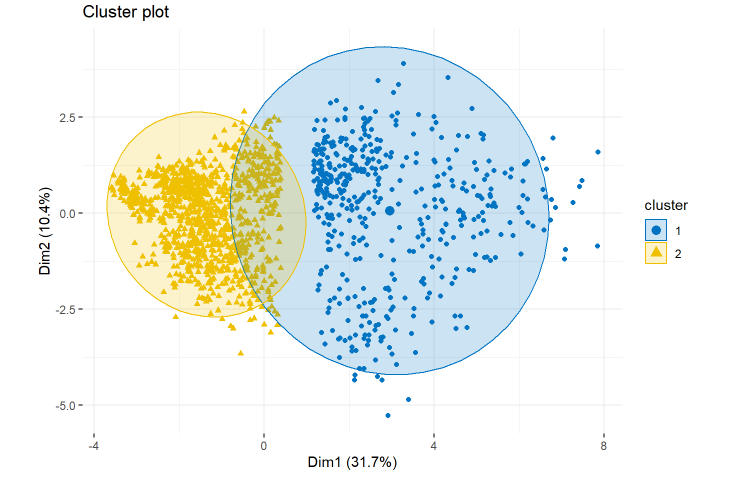
For analysis on the outcome variables, I’ll choose my best cluster option after comparing how the outcomes vary

I re-performed the KMN and HC analyses on my numeric data, but this time my outcome variables were turned into dummy variables and standardized with the rest of the data

KMN deduced that 2 clusters was still the best option for segmentation:



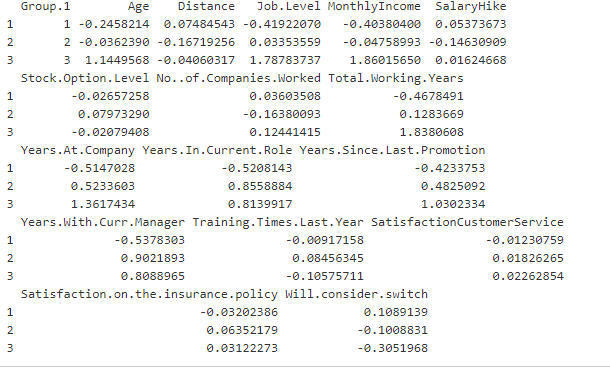


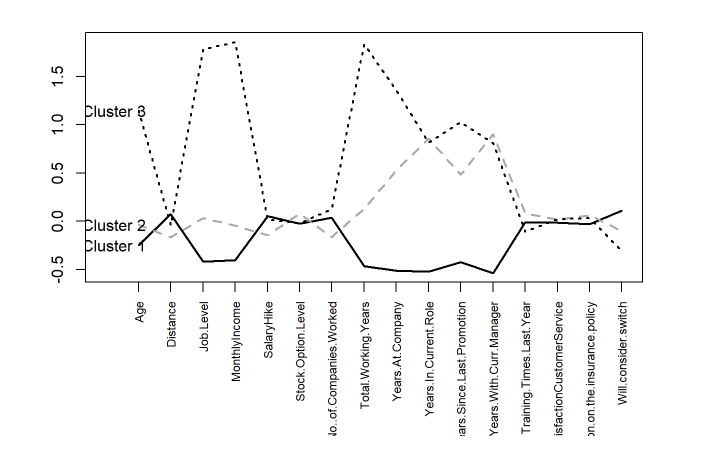


Overall with KMN, it seems so far that less-senior level professionals are less than satisfied with the customer service and satisfaction with their auto policies, are more apt to consider a switch

My HC method reduced from 4 clusters to 3 when I introduced the outcome variables:







This shows me that there is little difference in customer service or policy satisfaction between the clusters, and that these variables don’t have much predictive value in segmenting my data. However, cluster 1 is more likely to consider a switch than 2 or 3.

I like both the KMN and HC method I used with my numeric outcome data, but I would probably use HC here since it provided insights into an interesting third segment not available with KMN. Cluster 1 and 2 seem similar until we reach variables concerning tenure, so cluster 2 may refer to mid-level professionals.

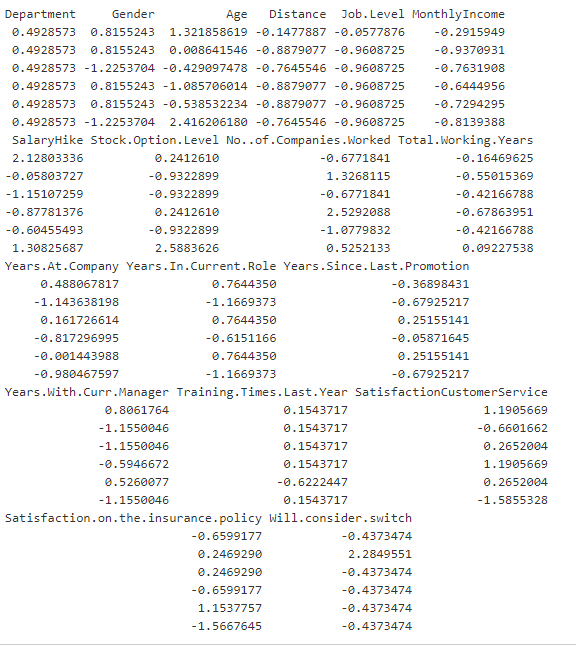
The second segmentation analysis was done by focusing on mixed data

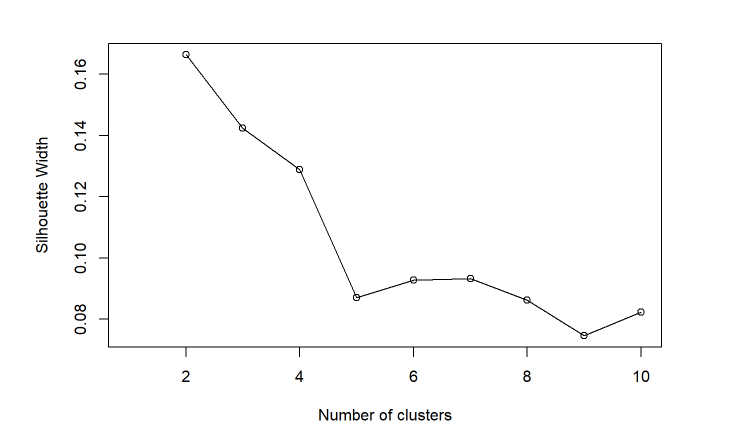
Cluster analysis using PAM():

I’ve already added in three categorial variables by including the 3 outcomes, but I’m curious how department, gender and job performance stack up in this kind of analysis

At a closer look, the data for performance review doesn’t help because there’s little variance between observations

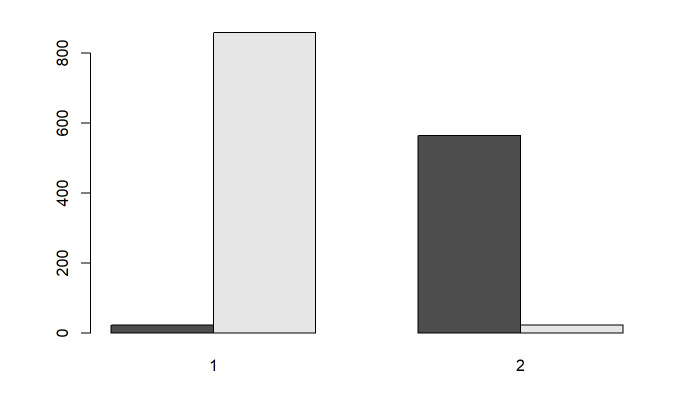
I feel like any of the other categorical variables will just supplement what I’ve already deduced from the analysis that’s already taken place



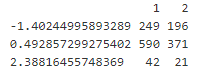


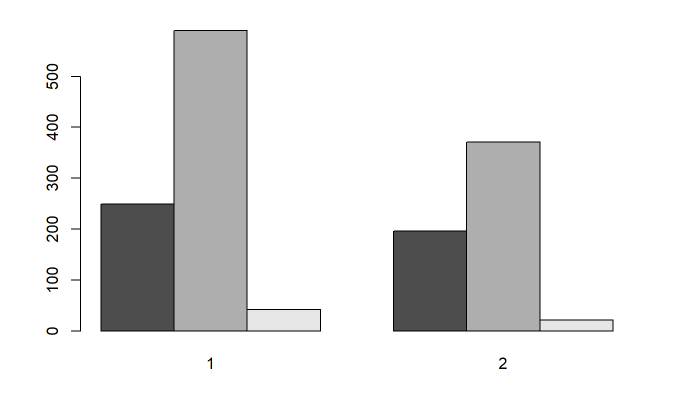
 I now have my optimal number of clusters for my categorical mixed data

There were far more men in cluster 1 and far more women in cluster 2, so this already shows me it was a huge differentiator between clusters (below)



Re-ran for the Department categorical variable:

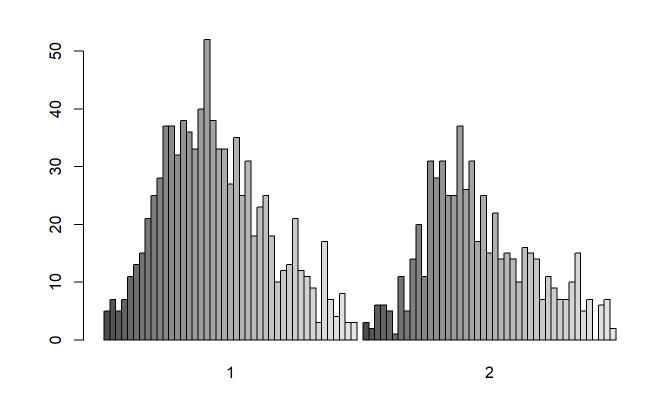




The raw data doesn’t show as much of a huge variance, but the visual allows us to see that HR doesn’t account for much of the data to begin with

The distribution of department is the same between both clusters, cluster 1 just has more people

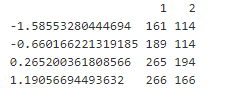
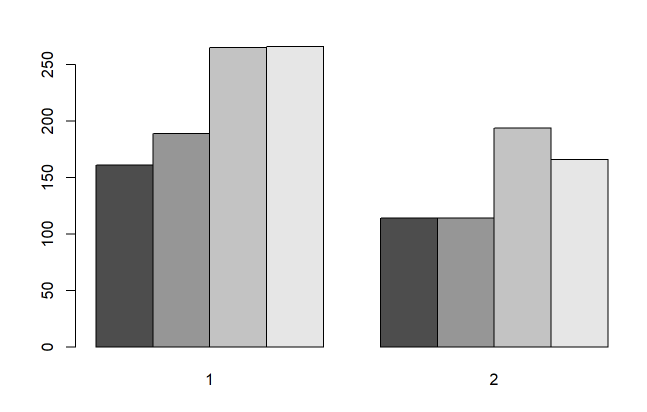
I’ll try this again with age, which we already worked with previously but I’m curious how it shows in this analysis method:



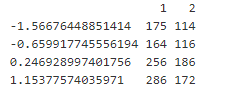
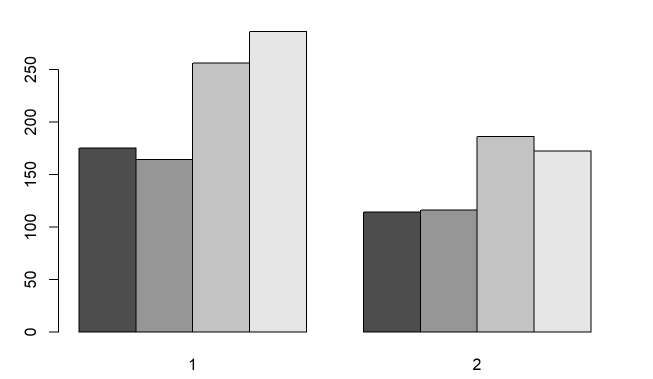
Cluster 2 might be slightly younger, but these clusters are a bit more homogenous in areas where they differed in my previous analysis before adding more categorical variables

Now I’ll look at the 3 outcome variables using this method

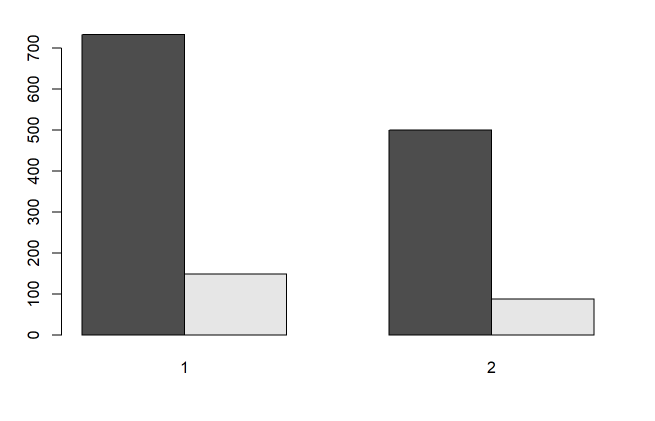
Customer service response:

Satisfaction response:

Switch response:

This is interesting, because before I added any additional categorical variables it seemed at least one cluster was more likely to consider a policy switch

It now seems that my clusters answer more positively on the surveys and are more homogenous with the outcome variables

I would absolutely test this further with more categorical data, as I feel there is more to be explored given that my numeric data told a slightly different story

Furthermore, the story here that everyone is mostly satisfied does not aid me in constructing a plan for how to address those who are at risk of churn

I want to know what are driving factors for satisfaction items - with my numeric approach, it seemed that clusters identifying mid-senior level professionals leaned more toward overall satisfaction

Variables indicating older age, more tenure, and higher salary hold more predictive value of satisfaction

In further analysis, I would be curious to perform my bar plot outside of clusters, i.e. gender vs outcome answer