# MA468: Predictive Analytics & Data Mining

# Project 2 | Absenteeism Clustering

**Nathan Gonyo, Aaron Neiger, Jack Michalowski**

**Dr. Rasitha Jayasekare**

## Table of Contents

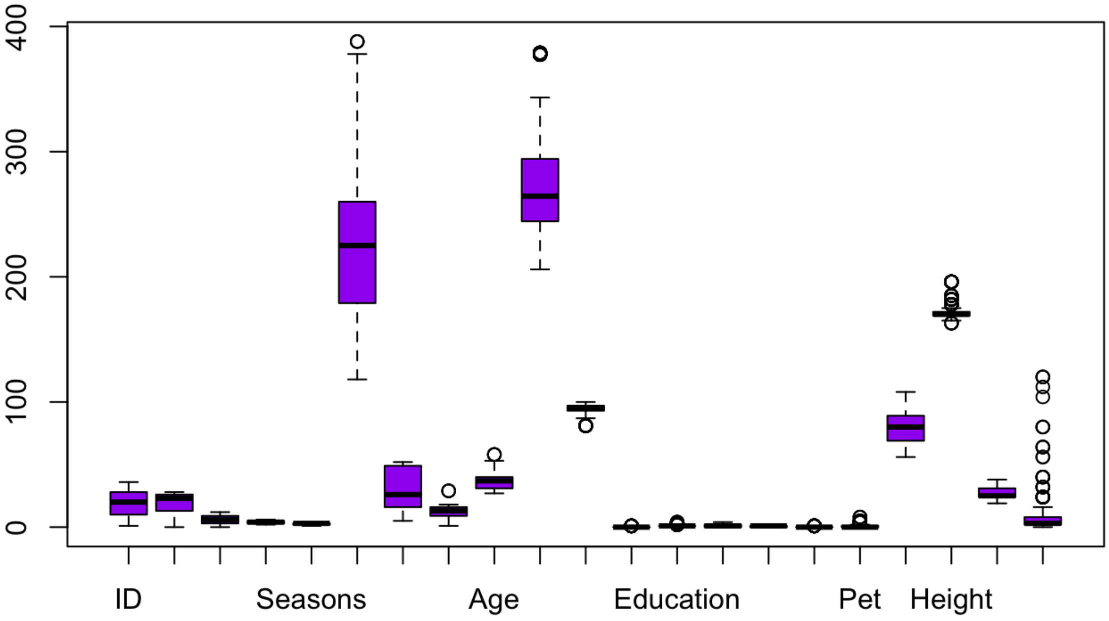
## Introduction & Background Information on Dataset

The database for this project. was created with records of absenteeism at work from July 2007 to July 2010 at a courier company in Brazil. Here is a list of variables in the dataset:

1. Individual identification (ID)
2. Reason for absence (ICD). Absences attested by the International Code of Diseases (ICD) stratified into 21 categories (I to XXI). And 7 categories without (CID)
3. Month of absence
4. Day of the week (Monday (2), Tuesday (3), Wednesday (4), Thursday (5), Friday (6)
5. Seasons (summer (1), autumn (2), winter (3), spring (4))
6. Transportation expense
7. Distance from Residence to Work (kilometers)
8. Service time
9. Age
10. Work load Average/day
11. Hit target
12. Disciplinary failure (yes=1; no=0)
13. Education (high school (1), graduate (2), postgraduate (3), master and doctor (4))
14. Son (number of children)
15. Social drinker (yes=1; no=0)
16. Social smoker (yes=1; no=0)
17. Pet (number of pet)
18. Weight
19. Height
20. Body mass index
21. Absenteeism time in hours

## Dataset Preprocessing: (Missing Values & Outliers)

To start with the preprocessing of our dataset for this project, we first created a random sample of 500 records from the dataset's original 740. Next, we checked the new dataset to see if any missing rows were present, of which there were none. Checking for the presence of outliers came next, however, we were instructed not to remove outliers for this project, so we can check for the presence of outliers for all variables at once rather than creating individual boxplots for every variable to keep it simple.



After carefully examining the outlier plot, we can see there are outliers present in variable 6, 8, 9, 10, 11, 12, 13, 16, 17, 19, 21. These variable numbers correspond to the following variables: Transportation.expense, Service.time, Age, Work.load.Average.day, Hit.target, Disciplinary.failure, Education, Social.smoker, Pet, Height, Absenteeism.time.in.hours. Again, we will not be removing outliers in this project, so we can move on to visual and numerical dataset description.

## Numeric Description of Dataset

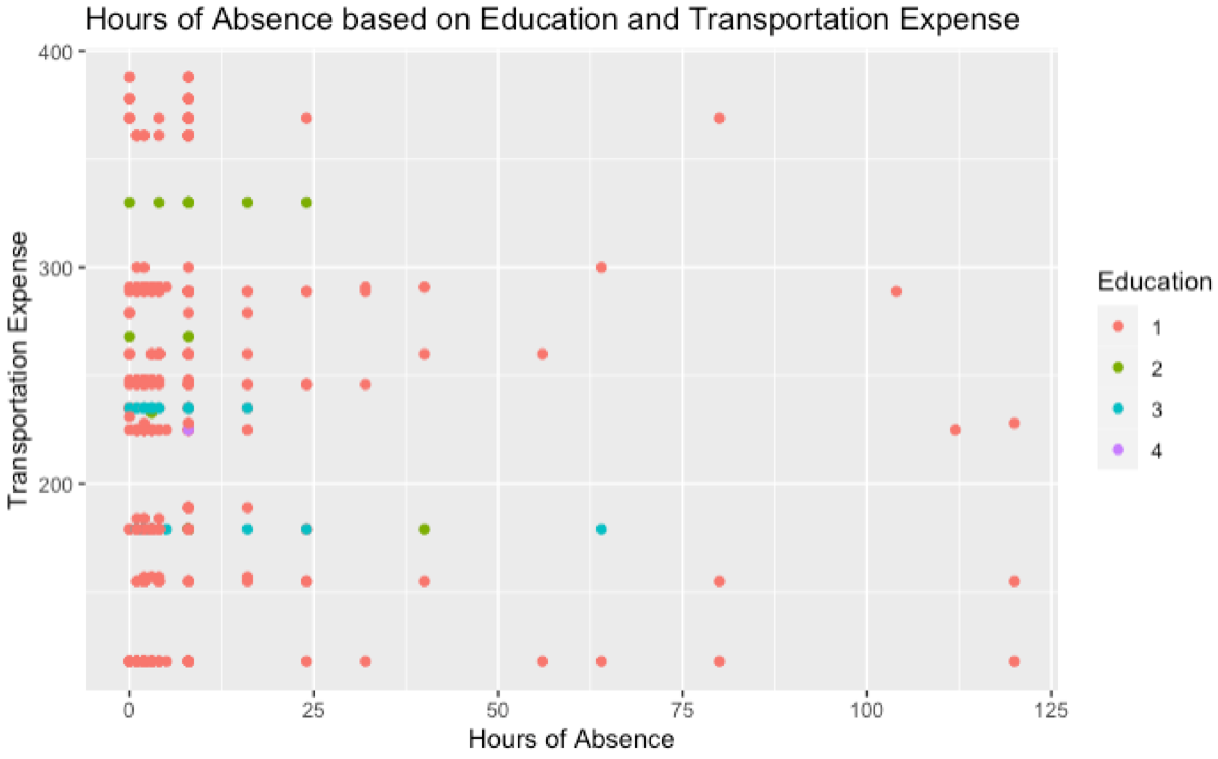
For numeric description of the dataset, we will need summary statistics for each quantitative variable contained in the dataset. We will need summary statistics for following variables: Transportation Expense, Distance from residence to Work, Service time, Age, Work load average / day, hit target, son, pet, weight, height, body mass index, absenteeism time in hours

Numeric Description of Dataset (Quantitative Variables)

| | Min | Median | Mean | Max | Std. Dev. |
|---|---|---|---|---|---|
| Transportation Expense | 118.0 | 225.0 | 221.8 | 388.0 | 67.73 |
| Distance to Work | 5.0 | 26.0 | 29.48 | 52.0 | 14.62 |
| Service Time | 1.0 | 13.0 | 12.51 | 29.0 | 4.24 |
| Age | 27.0 | 37.0 | 36.22 | 58.0 | 6.41 |
| Work Load Avg/ Day | 205.9 | 264.4 | 272.5 | 378.9 | 39.26 |
| Hit Target | 81.0 | 95.0 | 94.66 | 100.0 | 3.71 |
| Son | 0.0 | 1.0 | 1.04 | 4.0 | 1.11 |
| Pet | 0.0 | 0.0 | 0.73 | 8.0 | 1.24 |
| Weight (kg) | 56.0 | 80.0 | 78.57 | 108.0 | 12.82 |
| Height | 163.0 | 170.0 | 172.2 | 196.0 | 6.23 |
| Body Mass Index | 19.0 | 25.0 | 26.49 | 38.0 | 4.18 |
| Absenteeism Time in Hours | 0.0 | 3.0 | 7.51 | 120.0 | 14.8434 |

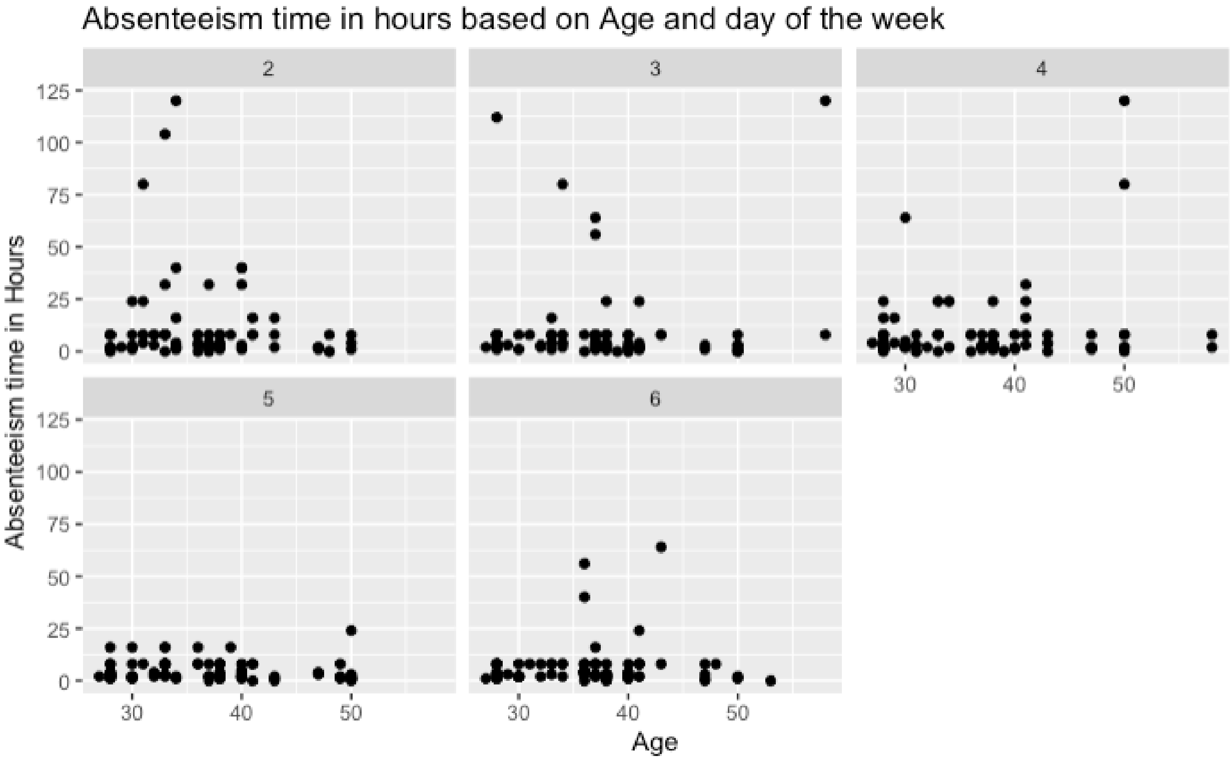
**Visual Description of Dataset**

Graph 1



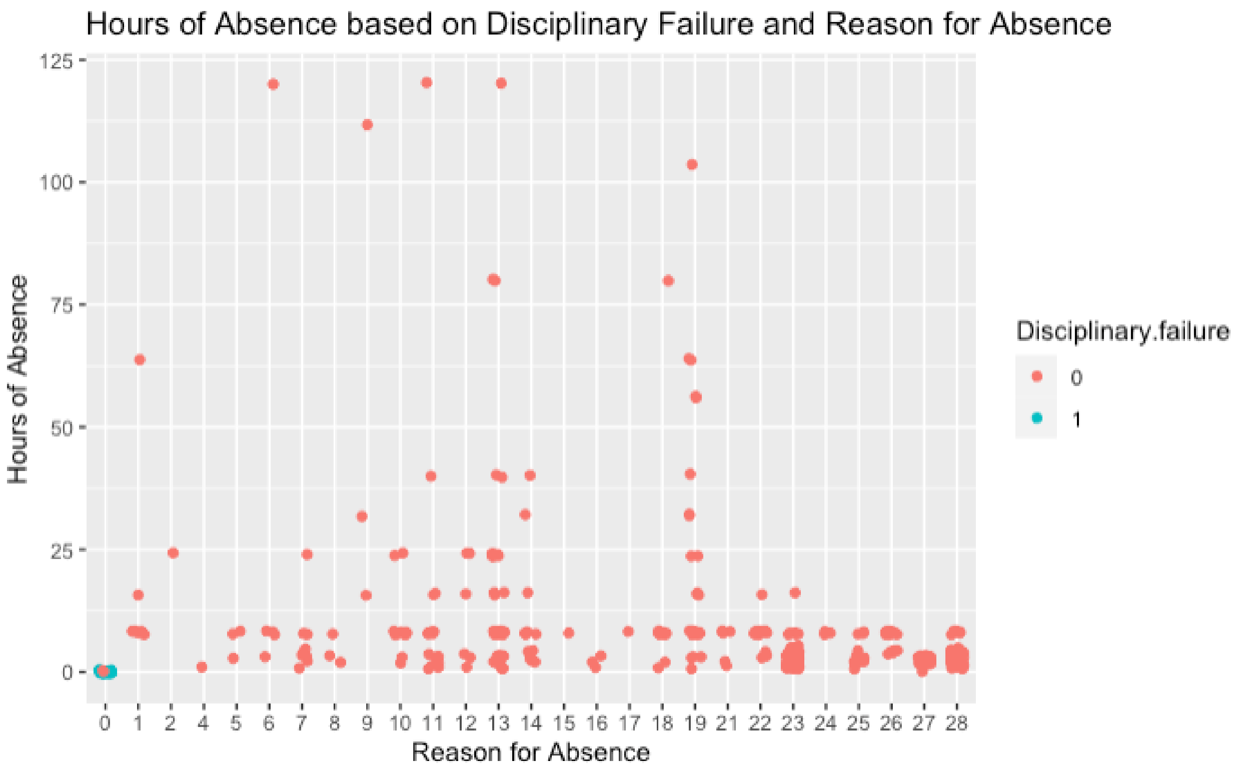
This multivariate ggplot graph shows the transportation expense and education of the clients and their hours of absence. This graph is useful to tell which education level and the employee's transportation expense have missed the most work. There are a lot of high school graduates on this and they are also the only ones who have missed more than 75 hours. One of the things I find interesting is that the employees with lower transportation expenses miss more hours of work than people with higher transportation expenses. Every employee who has a graduate education level or higher is in the middle 60ish% of transportation expenses and none have missed more than 50 hours of work except one.

Graph 2



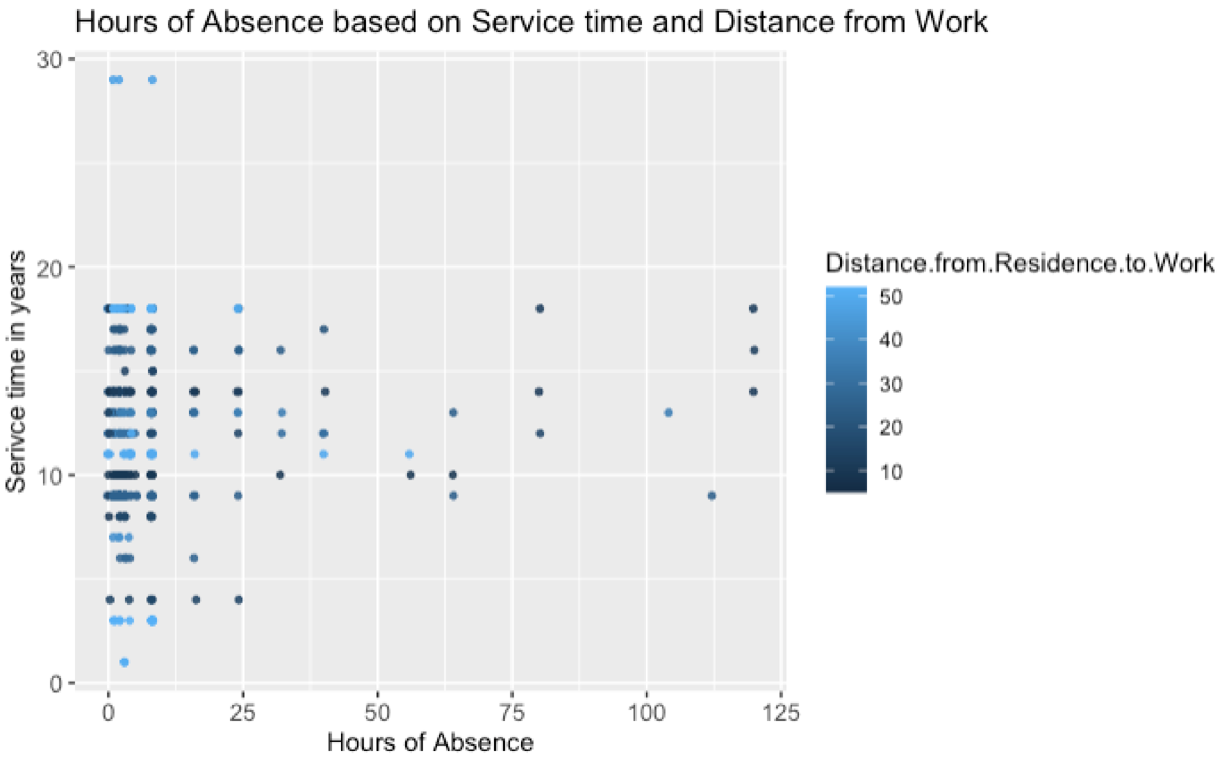
This multivariate ggplot graph shows the employees hours of absence based on their age and day of the week.. This graph is useful to tell which age and day of the week miss the most work. On graph 5, which is Thursday, almost no people miss a lot of time on that day. Then Friday is the day where people miss the second least amount of hours. I find this very interesting considering most people would prefer to miss a Friday versus another day of the week. One other notable point from this graph is that generally the people that are missing more hours of work are the younger population of employees.

Graph 3



This multivariate ggplot graph shows the employees hours of absence based on their reason and disciplinary failure. This graph is useful to tell the reason for which employees have missed work and whether they were on disciplinary failure or not. The first thing seen in this graph is that anyone who had disciplinary failure did not miss any hours of work. Of the conditions that people missed a lot of work for the four most were all for different diseases. The one that is the most spread out is for injury, poisoning and certain other consequences of external causes. This makes sense since a lot of different things can fall under this category where other categories are more specific. It appears only one person missed time for pregnancy or childbirth which I find interesting due to the fact that someone can miss a lot of time for that.

Graph 4

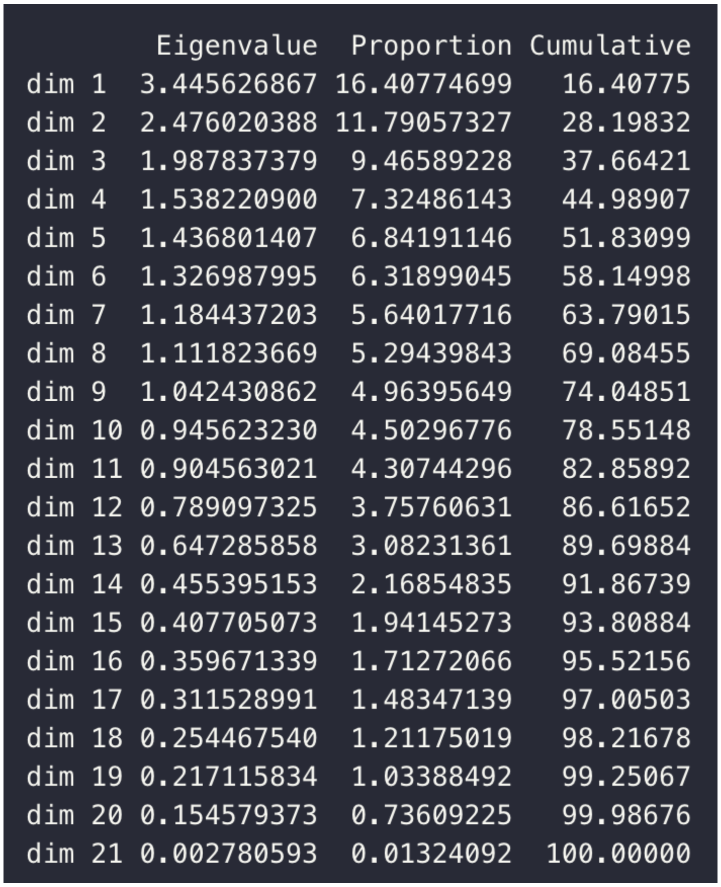


This multivariate ggplot graph shows the employees hours of absence based on their service time and how far they reside from work. This graph is useful to tell how far the employees live and how long they have worked at the company and then how much time they have missed. There are only three employees who have worked at the courier company for more than 20 years and all of them live close to 50 km away from work. They also have missed very little amount of time. Other than one employee, everyone who has missed more than 75 hours of work all appear to live within 10 km of work. This is showing that an employee's distance from their home to their work doesn't affect how much they miss work.

## PCA / Dimension Reduction

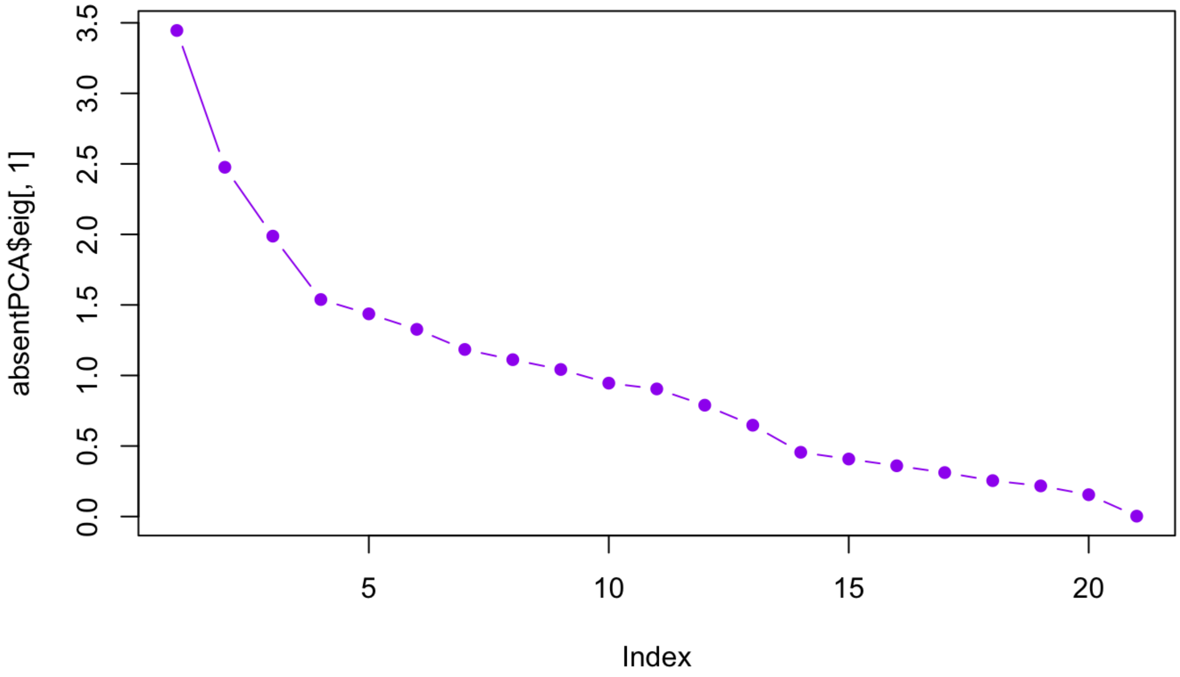
Prior to applying clustering algorithms to our dataset, we must first perform principal component analysis on our dataset. We will be applying mixed PCA as there are a mix of quantitative and qualitative variables in our dataset. If we had a response variable in our dataset, which we do not, this would have to be omitted prior to PCA.

Dimension / Components Eigenvalues Output

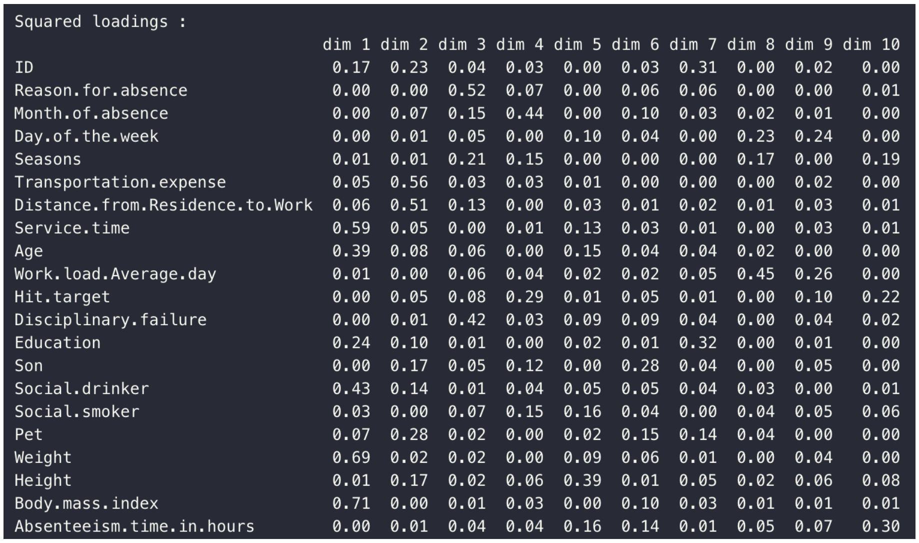


```
        Eigenvalue  Proportion Cumulative
dim 1  3.445626867 16.40774699   16.40775
dim 2  2.476020388 11.79057327   28.19832
dim 3  1.987837379  9.46589228   37.66421
dim 4  1.538220900  7.32486143   44.98907
dim 5  1.436801407  6.84191146   51.83099
dim 6  1.326987995  6.31899045   58.14998
dim 7  1.184437203  5.64017716   63.79015
dim 8  1.111823669  5.29439843   69.08455
dim 9  1.042430862  4.96395649   74.04851
dim 10 0.945623230  4.50296776   78.55148
dim 11 0.904563021  4.30744296   82.85892
dim 12 0.789097325  3.75760631   86.61652
dim 13 0.647285858  3.08231361   89.69884
dim 14 0.455395153  2.16854835   91.86739
dim 15 0.407705073  1.94145273   93.80884
dim 16 0.359671339  1.71272066   95.52156
dim 17 0.311528991  1.48347139   97.00503
dim 18 0.254467540  1.21175019   98.21678
dim 19 0.217115834  1.03388492   99.25067
dim 20 0.154579373  0.73609225   99.98676
dim 21 0.002780593  0.01324092  100.00000
```

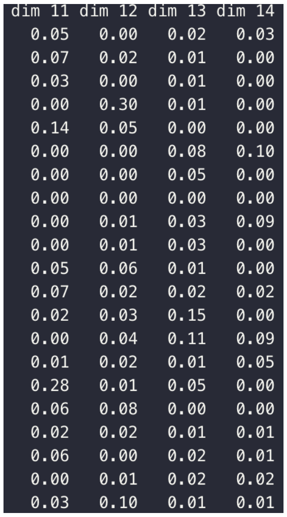
Based on the eigenvalue approach, we select PC's with eigenvalues greater than 1, so we would select the first 9 PC's. Based on the cumulative proportion approach, we would select the first 11 PC's assuming we want to preserve 80% of the dataset's variance, as the first 11 make up 82.85% of the dataset’s variance.



Based on the screeplot approach, it seems that after 14 it levels out, so we select the first 14 PC's. With these three methodologies in mind, and the fact that we wish to perform advanced clustering techniques, we will include the 14 PC’s. This was the largest number to be included based on the 3 methods. Now we will look at the first 14 PC’s in the square loadings matrix.



```
Squared loadings :
                               dim 1 dim 2 dim 3 dim 4 dim 5 dim 6 dim 7 dim 8 dim 9 dim 10
ID                              0.17  0.23  0.04  0.03  0.00  0.03  0.31  0.00  0.02   0.00
Reason.for.absence              0.00  0.00  0.52  0.07  0.00  0.06  0.06  0.00  0.00   0.01
Month.of.absence                0.00  0.07  0.15  0.44  0.00  0.10  0.03  0.02  0.01   0.00
Day.of.the.week                 0.00  0.01  0.05  0.00  0.10  0.04  0.00  0.23  0.24   0.00
Seasons                         0.01  0.01  0.21  0.15  0.00  0.00  0.00  0.17  0.00   0.19
Transportation.expense          0.05  0.56  0.03  0.03  0.01  0.00  0.00  0.00  0.02   0.00
Distance.from.Residence.to.Work 0.06  0.51  0.13  0.00  0.03  0.01  0.02  0.01  0.03   0.01
Service.time                    0.59  0.05  0.00  0.01  0.13  0.03  0.01  0.00  0.03   0.01
Age                             0.39  0.08  0.06  0.00  0.15  0.04  0.04  0.02  0.00   0.00
Work.load.Average.day           0.01  0.00  0.06  0.04  0.02  0.02  0.05  0.45  0.26   0.00
Hit.target                      0.00  0.05  0.08  0.29  0.01  0.05  0.01  0.00  0.10   0.22
Disciplinary.failure            0.00  0.01  0.42  0.03  0.09  0.09  0.04  0.00  0.04   0.02
Education                       0.24  0.10  0.01  0.00  0.02  0.01  0.32  0.00  0.01   0.00
Son                             0.00  0.17  0.05  0.12  0.00  0.28  0.04  0.00  0.05   0.00
Social.drinker                  0.43  0.14  0.01  0.04  0.05  0.05  0.04  0.03  0.00   0.01
Social.smoker                   0.03  0.00  0.07  0.15  0.16  0.04  0.00  0.04  0.05   0.06
Pet                             0.07  0.28  0.02  0.00  0.02  0.15  0.14  0.04  0.00   0.00
Weight                          0.69  0.02  0.02  0.00  0.09  0.06  0.01  0.00  0.04   0.00
Height                          0.01  0.17  0.02  0.06  0.39  0.01  0.05  0.02  0.06   0.08
Body.mass.index                 0.71  0.00  0.01  0.03  0.00  0.10  0.03  0.01  0.01   0.01
Absenteeism.time.in.hours       0.00  0.01  0.04  0.04  0.16  0.14  0.01  0.05  0.07   0.30
```



```
dim 11 dim 12 dim 13 dim 14
  0.05   0.00   0.02   0.03
  0.07   0.02   0.01   0.00
  0.03   0.00   0.01   0.00
  0.00   0.30   0.01   0.00
  0.14   0.05   0.00   0.00
  0.00   0.00   0.08   0.10
  0.00   0.00   0.05   0.00
  0.00   0.00   0.00   0.00
  0.00   0.01   0.03   0.09
  0.00   0.01   0.03   0.00
  0.05   0.06   0.01   0.00
  0.07   0.02   0.02   0.02
  0.02   0.03   0.15   0.00
  0.00   0.04   0.11   0.09
  0.01   0.02   0.01   0.05
  0.28   0.01   0.05   0.00
  0.06   0.08   0.00   0.00
  0.02   0.02   0.01   0.01
  0.06   0.00   0.02   0.01
  0.00   0.01   0.02   0.02
  0.03   0.10   0.01   0.01
```

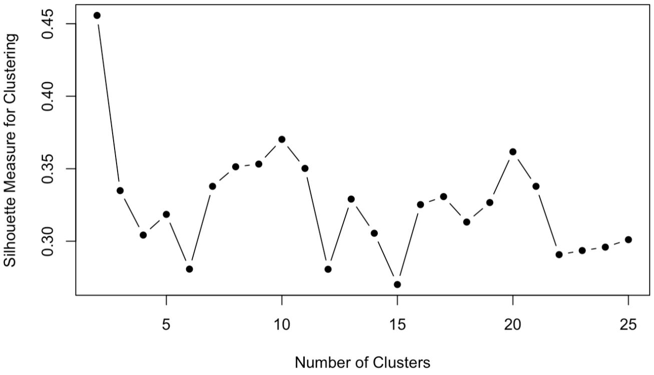
Note these are squared loadings, not loadings matrix. In the first round, we will keep variables with an |r| above .5, so we will select the variables with a squared r above .25. Based on this criteria, we only end up dropping one variable, seasons. Because this leaves 20 variables remaining, which barely changed the dataset, we will opt to keep variables with a squared r above .30. Based on our new criteria, we drop the following variables: Day.of.the.week, Seasons, Hit.target, Son, Social.smoker, Pet, Absenteeism.time.in.hours. We are left with 14 variables of the original 21 after dropping 7. Now we will move onto clustering.

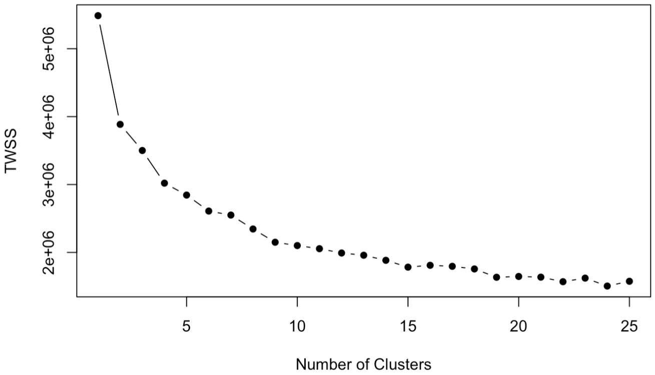
## Cluster Analysis

Since we are treating this dataset as a mixed type dataset, the clustering techniques that will be performed are the K-Prototype and Hierarchical (with gower) methods. These techniques are used to handle both quantitative and qualitative data. Since some of the variables in our dataset are quantitative, and were given numerical values, they had to be turned into factor variables for further analysis with our methods of clustering. These variables were, ID, Reason.for.absence, Month.of.absence, Disciplinary.failure, Education, and Social.Drinker. Once this part was completed, the clustering could begin. Note that the original dataset given already had converted these variables.

### *K-Prototype*

First, the K-Prototype technique was implemented. To get a visual representation of the output, TWSS and Silhouette Measure plots were created.

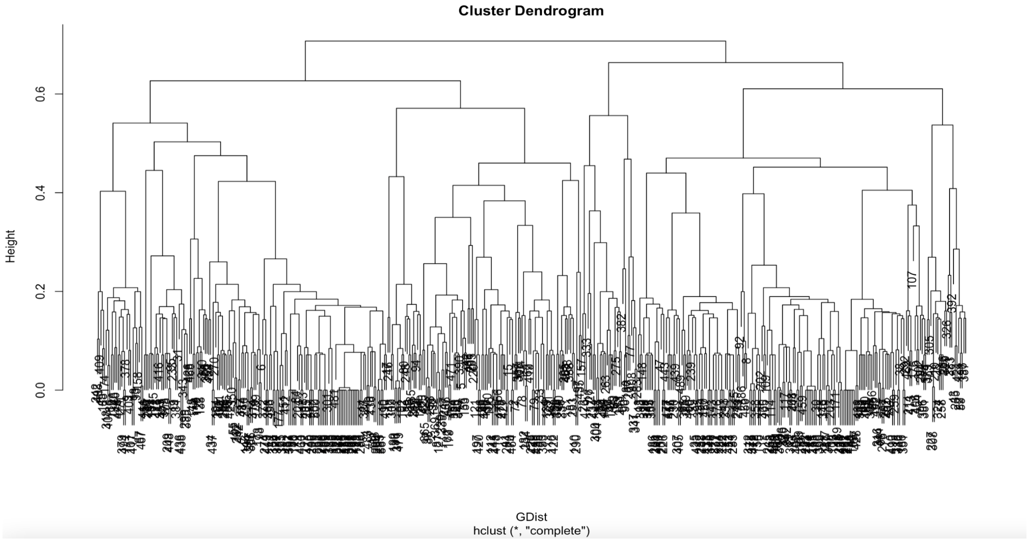




Based on the two plots, especially the Silhouette Measure, we can conclude that the optimal number of clusters should be two. This is because the max silhouette value occurs at 2, along with the elbow of the TWSS showing a similar result.

### *Hierarchical*

Next, the Hierarchical technique was implemented. A gower distance between observations was calculated and put into a distance matrix, and this distance matrix is what was used for clustering. The gower method was used because we are working with a mixed type dataset. The main visualization for this method is the Dendrogram plot. The highest clearing between the levels of this dendrogram will give us the optimal amount of clusters.



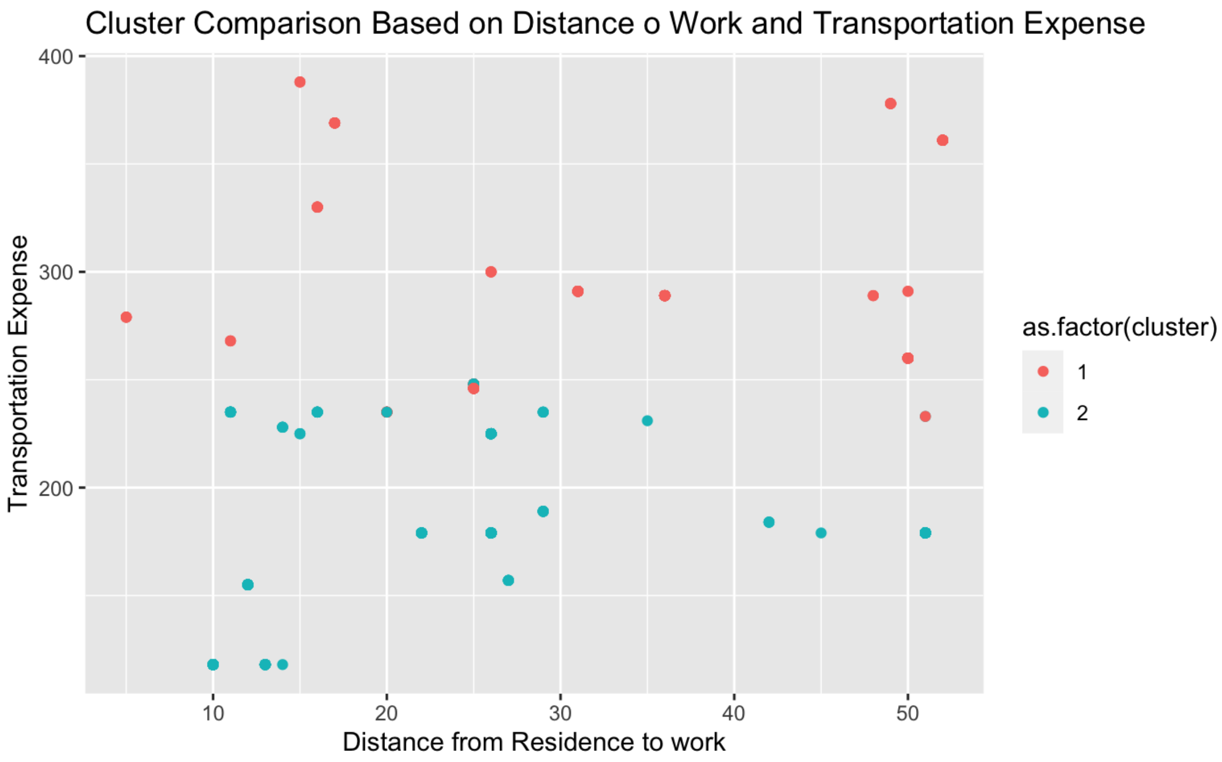
Based on the dendrogram plot, the largest clearing seemed to be between the first and second level of the cluster dendrogram, meaning the optimal number of clusters would be two, but there are very similar size clearings elsewhere on the plot that would maybe suggest three or five to be the optimal number of clusters.

### *Choosing Clustering Technique for Post Analysis*

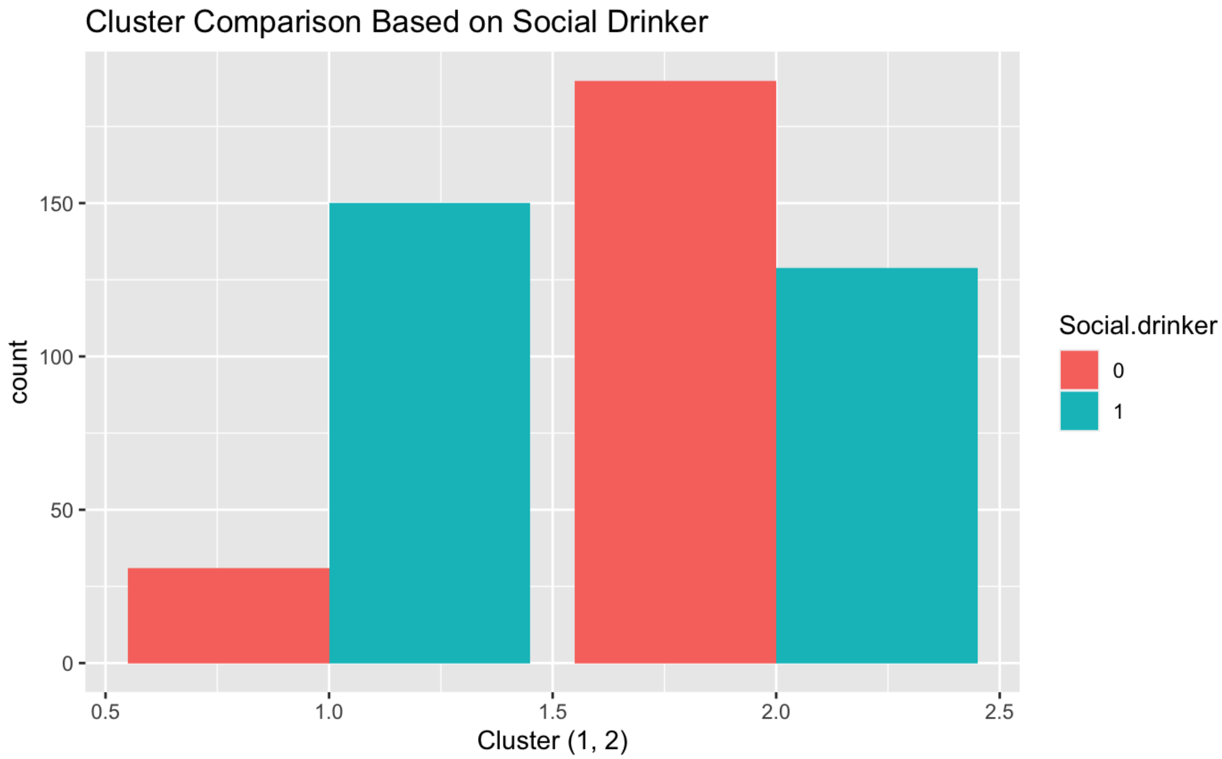
Between the K-Prototype and Hierarchical methods of clustering, for this dataset, the K-Prototype technique is being selected for post analysis. The reasoning behind this selection is that, since K-Prototype has shown a more clear cut optimal number of clusters, that number being two, it seems like the best choice. The silhouette maximum value very clearly displayed the number of clusters that we should use. Based on the dendrogram plot for the Hierarchical method, it was much more difficult to discern the optimal number of clusters. Further analysis beyond the cluster dendrogram would have to be done to truly come away with the accurate amount of clusters for this method. Therefore, K-Prototype is our choice for post analysis work.

## Postanalysis

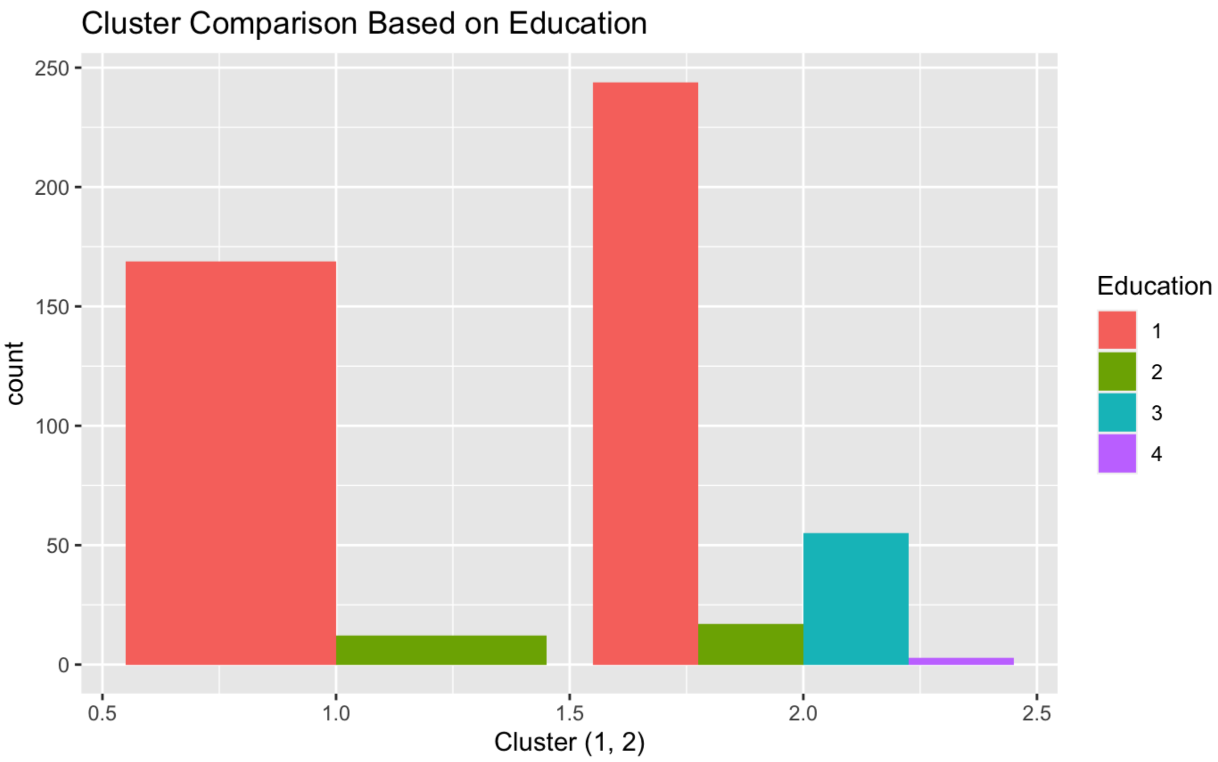
To start with the post analysis, we had to attach the cluster assignments from k prototype to the dataset as a new column to allow us to take a closer look into the patterns within the two clusters. After further investigation, a few patterns and differences within the groups became clear. The most evident differences between the 2 clusters were the following: transportation expense was much higher in the first cluster than the second. Second, education of higher levels (3,4 out of 1-4) were much more present in the second cluster. In addition, absences where social drinker was true was much more common in the first cluster. After noticing these differences between clusters, we used ggplot2 to create plots to further highlight these differences.



Cluster Comparison Based on Distance o Work and Transportation Expense



Cluster Comparison Based on Social Drinker



The above plots clearly visualize the trends between the clusters mentioned above. As mentioned, transportation expense is much higher in cluster 2 than cluster 1 (shown in chart 1). Social drinking is much more common in the first cluster (chart 2). Lastly, higher levels of education (>2) are much more prevalent in the second cluster. Summarizing these differences, an absence record placed in the second cluster is more likely to have a lower transportation expense, higher education level, and lower level of social drinking. Viewing these differences from an outsider's perspective wishing to gain insight about absenteeism, it makes sense that someone who has a higher education background also would have a lower level of social drinking as usually those who engage in these types of behavior are younger people who are less educated. It also makes sense that absences in cluster one contain a higher level of social drinking and a higher transportation expense. Logically, if someone has to spend more on transportation they probably drive to work, and nothing is worse than a long commute to work when you are hungover from a long night of drinking the night before.

Work Performed by Each Member:

Nathan: Preprocessing, PCA, post analysis

Aaron: Preprocessing, ggplot2 visuals / descriptions

Jack: Clustering, Cluster analysis / picking optimal method

Signed: *Nathan Gonyo, Aaron Neiger, Jack Michalowski*