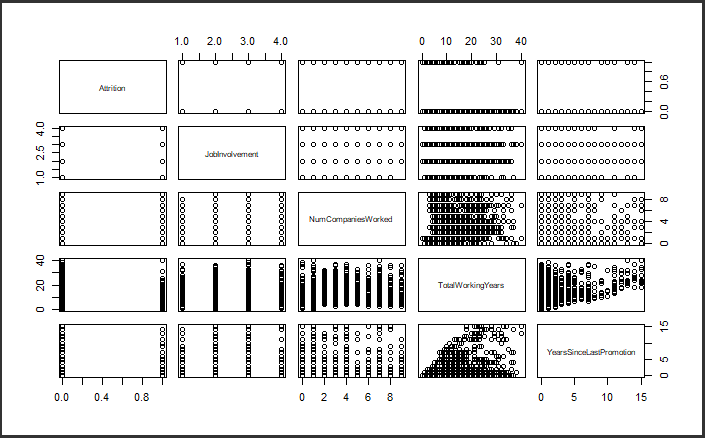
# Logistic Regression

My target variable is Attrition. I look at the scatter plot matrix from the perspective of Attrition being in the y-axis. Looking at the plot, there is hardly any highly positive correlation between target variable and predictor variables. It looks like Attrition and JobInvolvement and TotalWorkingYears are slightly negatively correlated.



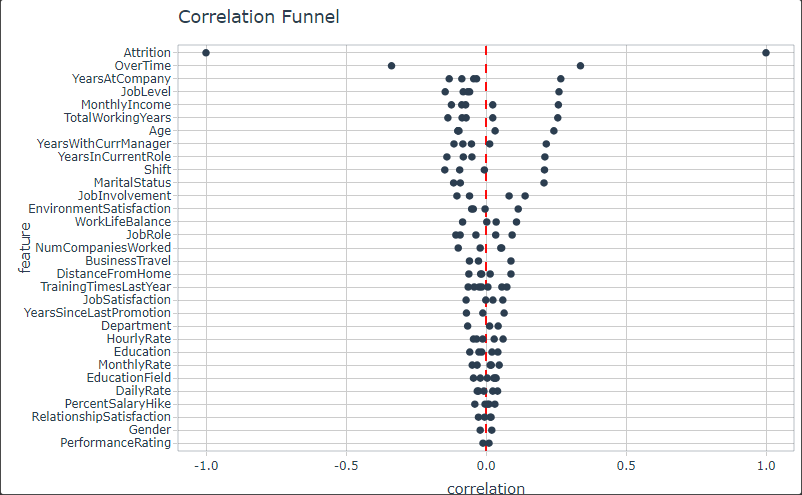
With all the variables considered while creating the regression model we achieved AIC value of 400.78 and with some improvements and removal of insignificant variables we ultimately were able to bring the AIC value down to 384.07 in 18 iterations with a Null deviance: 849.16 on 1186 degrees of freedom and Residual deviance: 328.07 on 1159 degrees of freedom

Major factors influencing the iteration rate turned out to be Business travel, Dist from home, Environmental satisfaction, Job Involvement, Job Satisfaction, Marital Status, Monthly Income, No. of companies worked, work life balance, Years in current role and Years Since Last Promotion.

The model can now predict if the person will leave the organization or not with an accuracy of 89.9%.

# Clustering

Determining the correlation between each variable in the dataset and attrition to include a variable in the study, a correlation threshold of 0.1 will be used. This is useful for a number of reasons, the most important of which is that it allows us to choose which variables to use in our cluster analysis.

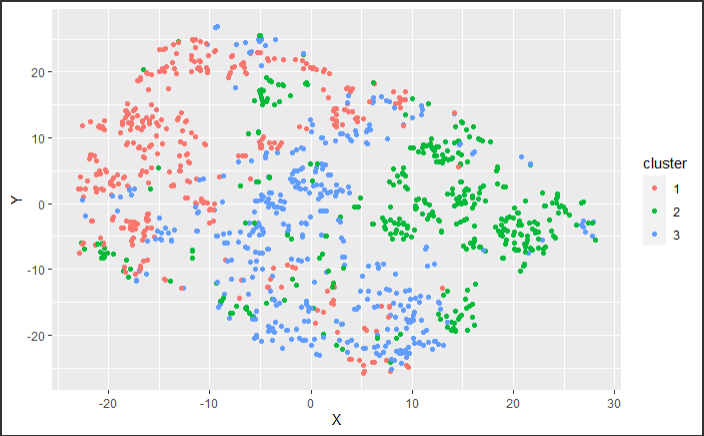


Remember that more criteria do not necessarily result in better categorization. Including more factors may make it more difficult for end users to assess and act on the results. We conclude that 14 factors, ranging from overtime to business travels, should be included in the grouping using a threshold of 0.1 defined by our correlation analysis. Before the analysis, any of these fourteen variables that are of a character data type (for example, MaritalStatus = Single) are translated into factors (more on this below).

Selected features:

"EmployeeNumber", "Attrition", "OverTime", "JobLevel", "MonthlyIncome", "YearsAtCompany", "StockOptionLevel", "YearsWithCurrManager", "TotalWorkingYears", "MaritalStatus", "Age", "YearsInCurrentRole", "JobRole", "EnvironmentSatisfaction", "JobInvolvement", "BusinessTravel".

We performed a three-cluster cluster analysis based on the average width of our silhouettes and then combined that information with our original data to determine where each person belonged. Examine the three Medoids that represent our three groups.



The clusters we discovered seem to be of diverse sorts that are causally related to company turnover. As a result, we may get a better knowledge of the reasons that drive revenue to change as a whole. Our current output is code-based, which is OK for data analysts but not for business partners or human resources stakeholders. We will present examples to make the findings more approachable to individuals without analytical backgrounds.

Using the t-SNE approach, we may observe numerous variables from the cluster analysis at once. This dimensionality reduction approach employs a sparse representation of the data to generate a more meaningful graphic. The plot allows us to graph our cluster analysis results in two dimensions, enabling end users to view what was previously code and concepts.

The average silhouette width indicates that the clusters are highly ordered. Despite the fact that each cluster has numerous outliers, the clusters as a whole seem to be coherent. We may utilize this data to advise more focused programs and activities to boost internal morale and reduce turnover. A cluster analysis, such as the one we just did, enables us to apply our intervention ideas to a much broader population—the whole cluster—increased efficiency and efficacy.

The study also reveals subgroups of the work force where turnover is lower. This information might also be useful in analysing the perks we provide to our employees.



Together with the other variables in our summary that have similarly extreme values, we can see that this group has characteristics such as having a big number of current projects, putting in long hours, and getting little to no appreciation for their work. As a result, we will refer to this statistic as the Effort-reward imbalance, and we will also introduce the Work underload meter as a result.

# Multiple Linear Regression

Initially created a linear model using the singificat variables with the reference from the previous analysis and following are the features used, Age + BusinessTravel + Department + EnvironmentSatisfaction + JobInvolvement + JobLevel + JobRole + JobSatisfaction + MaritalStatus + MonthlyIncome + OverTime+ TotalWorkingYears + YearsInCurrentRole + YearsWithCurrManager

This model had R^2 and adjusted R^2 of .737 and .732 respectively. Based on the P value of the variables we reduced the features in the next iteration with the features Age + BusinessTravel+ JobInvolvement + JobRole + TotalWorkingYears + YearsInCurrentRole + YearsWithCurrManager

And it was observed that there was no change in the value of R^2 and adjusted R^2 which are now at .734 and .731 respectively.