**EWCS Dataset Unsupervised learning report:**

We are given EWCS dataset and are instructed to apply unsupervised learning methods to it in order to gain some insights into the dataset. We apply Principal Components Analysis and Cluster Analysis alongside exploratory data analysis techniques that we apply such as correlation analysis.

We begin by checking for any missing values or duplicate values in the dataset before we proceed with our analysis. After we determine that the data is clean, we conduct exploratory data analysis. Through this, we determine that there is almost an equal number of Male and Female participants in the survey, and that most of the participants of the survey are between 24 and 60 years old. Upon completion of correlation analysis, we see that all 4 subparts of Q87 are strongly correlated to one another meaning that we can determine how an individual is likely to answer one subpart of Q87 based on how they answered another. Q90 is also positively correlated to Q87, but it isn’t a strong correlation. There is no significant correlation between the age and gender of a participant and their response to the questions according to the correlation plot.

Chart, bar chart, waterfall chart, treemap chart

Description automatically generated

We calculated the frequency of all responses to the survey. In most cases, an employee is more unlikely to have indicated any ‘feelings of resentments regarding their how they feel about their job and their mood. We produce a plot showcasing the count of all responses to each question in the survey, except Q2a and Q2b.

Chart, bar chart

Description automatically generated

**Principal Components Analysis:**

Upon completion of PCA, we derive that the first two Principal Components, PC1 and PC2, explain 53 percent of the variation in the model. We produce a biplot to help us further assess. We can use a scree plot to confirm that we must use 2 factors in our PCA.

Chart, histogram

Description automatically generated

Q87 has high loadings on PC1 as seen on the biplot we have obtained, as does Q90. This indicates that PC1 heavily influences Q87 and its subparts.

Text

Description automatically generatedChart, scatter chart

Description automatically generated

**Cluster Analysis:**

We now move on to cluster analysis to help us identify clusters and associations between variables that we may otherwise not have found.

Using the elbow method, we derive that we must use 2 clusters in our analysis.

Chart, line chart

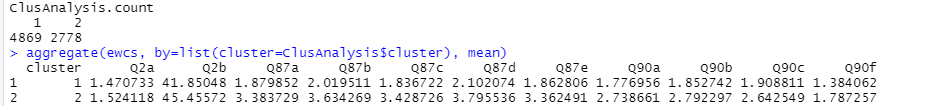
Description automatically generated

We can further confirm this by calculating the gap statistic and select a K value with the maximum gap statistic value.

We proceed with K means clustering

Chart

Description automatically generated



Through our analysis, we calculate the mean of each response and both clusters and make the following observations.

There are almost 2 times the number of members of Cluster 2 (2778 members) in Cluster 1 (4869 members). Members of Cluster 1 are slightly younger than the members of Cluster 2, and the responses are positive in general as compared to the responses in Cluster 2- which are mostly negative. Members of Cluster 1 are generally happier and more upbeat than the members of Cluster 2. People in both clusters feel that they’re good at their jobs, as seen by the mean value of the response to Q90f, which is 1.38 for Cluster 1 and 1.78 for Cluster 2.