|  |  |
| --- | --- |
| A picture of a winding road and trees  STATISTICS  INDIVIDUAL PROJECT REPORT | **KITHSIRI JAYAKODY**  **D19126340** |

**Section 1:**

This dataset has given 44 columns with 23 categorical variables and 21 numeric variables. According to the values of the dataset, following research question was identified.

Does the questionnaire measure a single construct Or is it possible that there are multiple aspects which is used to measure the overall average score of the professional evaluation G\_SC. ( this PCA is doing to reduce the dimensions to predict the G\_SC.)

**Section 2:**

Based on the task, first step would be to load the dataset. After observing the dataset, empty column was identified, and it was removed. Moreover, there are 21 categorical columns and 21 numeric columns. In order to apply the principle component analysis, categorical variables were first of converted to factor variables and then later it was converted to one hot encodings. Ex: if there are three types of values for one categorical variables such as category variable has normal, middle and higher as values, then it is represented as category\_normal, category\_middle and category\_higher and for each value it is represented with ([1,0,0], [0,1,0], [0,0,1]). After converting to one hot encoding, number of columns increased to 3937. It is 89-time increased compare to the 44 number of columns of the original dataset.

In addition to that, unnecessary columns were removed from the dataset. Such as Cod\_s11, Cod\_SPro. As those two columns provide unique values and it won’t help for dimensionally reduction techniques.

Then as the next step, I checked for the missing values, according to the analysis, there are no such missing values.

According to correlation values among 3937 variables, following charts shows the value distribution of variables according to the correlation values.

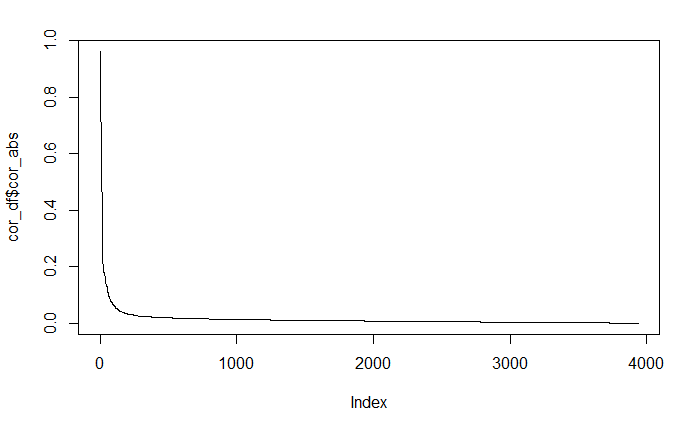


Figure 1: Value distribution of correlations

Then mainly I checked for the correlation (looking for > 0.3, < 0.8) and eliminate variables of concern and repeat . following is the output generated for the correlation matrix.

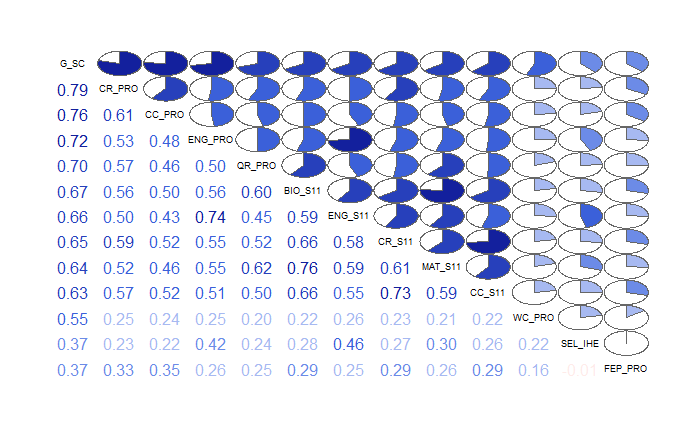


Figure 2: Correlation distribution looking for( > 0.3, < 0.8)

**Section 3:**

**3.1.**

In order to understand and check whether the difference from identity matrix is significant or not, I have already performed the bartletts test and the results showed that it is significant 0<0.05.

Moreover, when I checked for variance among factors it is higher than .6. The value is .82.

When the determinant value was tested it gave the result 9.454669e-05 > .00001. so all the relevant tests were passed and due to that we can do the principal component analysis.

**3.2**

when we have n manifest variables in our analysis, PCA transforms the data such that the total variance of the components n will be redistributed among the components and following diagrams show the eigen values of each dimensions.

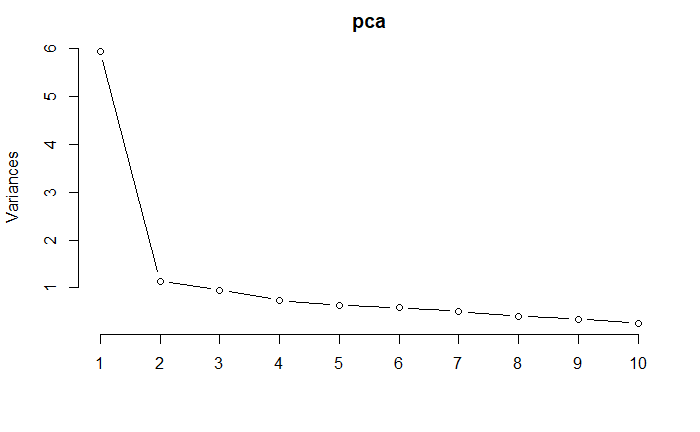
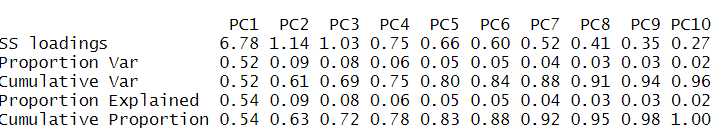
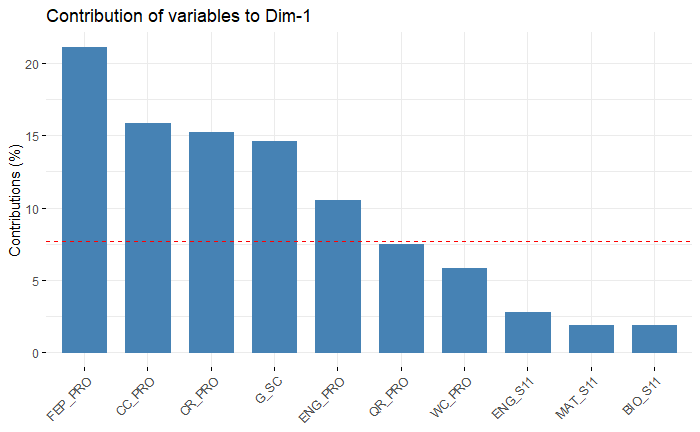
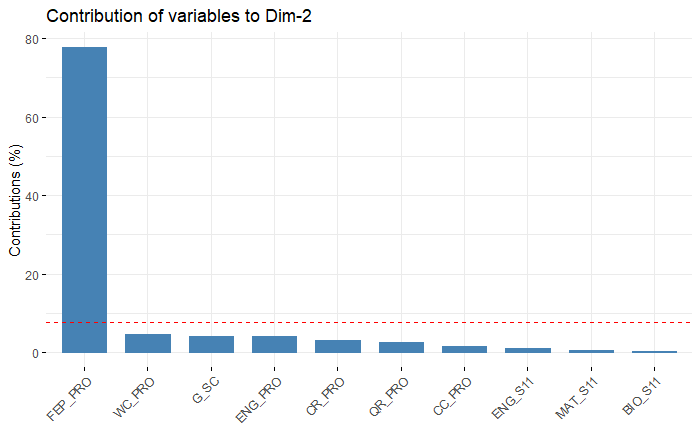


Figure 3: Eigen value distribution of PCA

According to the above Figure 3 and chart 1 it is evident that there are only three principal components which is greater than 1. (SS loading). Moreover, as table 1 depicted first principal component explain total variation which is around 52%. This was Calculated by dividing Eigenvalue by the number of factors/components. Moreover, cumulative variance Indicates how much of the variability in the data has been modelled by the extracted factors. Factor/Component 1 explains 52%, Factor/Component 1+2 explain 61% etc. Together the 4 uncorrelated factors summarise 75% of the information in the original inter-related variables. By using 10 principal components we can explain around .96. but after the fourth principal component, proportion of the variance explanation is not significant. As there are only three principal components which is higher than 1, those 3 was selected for the analysis. Moreover as Cattell (1966) suggests using the ‘point of inflexion’ of the scree plot to decide how many factors to extract, those 3 main factors were extracted. But according to the Kaiser’s Extraction with following condition, 5 factors were extracted. sample size > 250 and mean communality ≥ 0.6.

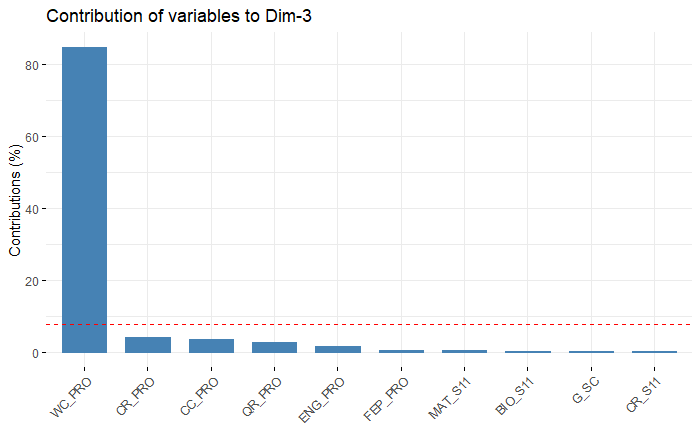
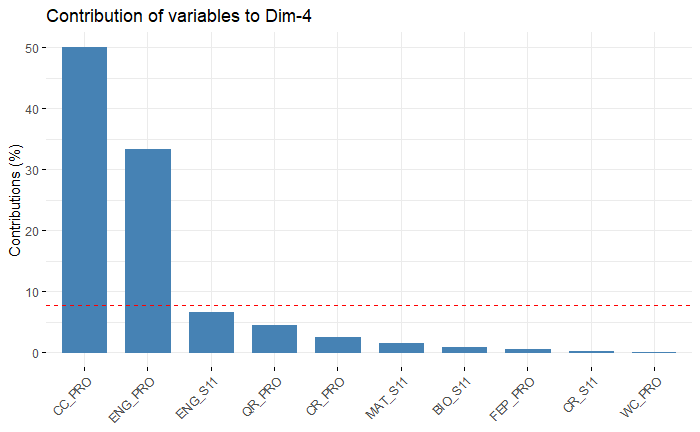
**Table 1: Statistics of the 10 main Principal Components**

Following diagram illustrates how each component varied according to the



Variable contribution of dimension 2

Variable contribution of dimension 1



Variable contribution of dimension 4

Variable contribution of dimension 3

According to George and Mallery , Cronbach’s Alpha value is greater than .87> .8. . therefore, we can consider the dataset it’s reliable.

**Final result of the analysis:**

A principal component analysis (PCA) was conducted on the 44 items with orthogonal rotation (varimax). Bartlett’s test of sphericity, Χ2(253) = 227922, p< .001, indicated that correlations between items were sufficiently large for PCA. An initial analysis was run to obtain eigenvalues for each component in the data. three components had eigenvalues over Kaiser’s criterion of 1 and in combination explained 69% of the variance. The scree plot was slightly ambiguous and showed inflexions that would justify retaining either 3 or 4 factors.

Given the large sample size, and the convergence of the scree plot and Kaiser’s criterion on three components, three components were retained in the final analysis. Component 1 represents percentitle,Feb\_pro,cc\_pro,cr\_pro, component 2 a FEP\_pro, component 3 a WC\_pro, and component 4 CC\_pro and ENG\_pro concerns.

According to the analysis all had high reliabilities, all Cronbach’s α = .87.