**Exploring Leadership Styles for Innovation**

**From Australian Public Service Census Data 2019**

**Abstract**

The Australian Public Service (APS) conducts annual census for the employees in their agencies. This analysis identifies the leadership styles which can be used to predict employees that strive for creativity and innovation within the APS agencies. Throughout exploratory factor analysis (EFA) and correspondence analysis (CA), this study identified the leadership styles as coaching, transformation, and consideration. From the result, EFA successfully determined that the employees were influenced by only the motivational characters and the considerate characters of the leaders in agencies which is related to the consideration leadership style. In addition, CA found how the employees corresponded to each of questions regarding leadership styles, innovation, organizational changes, and wellbeing. This analysis recommends that APS to continue implementing newer training and learning programs and use the coaching leadership style as it found to significantly impact the employees that strive for creativity and innovation.

1. **Descriptive Analysis**
2. **Overview**

* Data Structure

The APS employee census of 2019 data consists of 104,471 participants total. The questionnaire consists of 111 questions with sub questions under some specific questions. In total, there are 104,471 participants with 380 questions on the questionnaire.

* **Research Questions**

The major four research questions set for this analysis are

* “Can leadership styles be used to predict employees that strive for creativity and innovation?”
* “Which of the leadership styles are predominant in the APS agencies?”
* “Which of the leadership styles are regarded as significant for innovation?”
* “Understanding the employee response from leadership, innovation, organization changes and wellbeing”

1. **Demographic Profiles**

* Size of agency, gender, age, and classification level

The demographic profiles of the census participants can be classified as four variables, which are agency size, gender, age, and classification level. From the analysis, major key takeaways are

* + Around 86% of participants are from large agency
  + Around 83% of participants are under 54 years old
  + Around 59% of participants are female
  + Around 67% of participants are classified as subordinate (Trainee/Graduate/APS)
* Refer to figure 1 below for the visual representation of the demographic profiles.

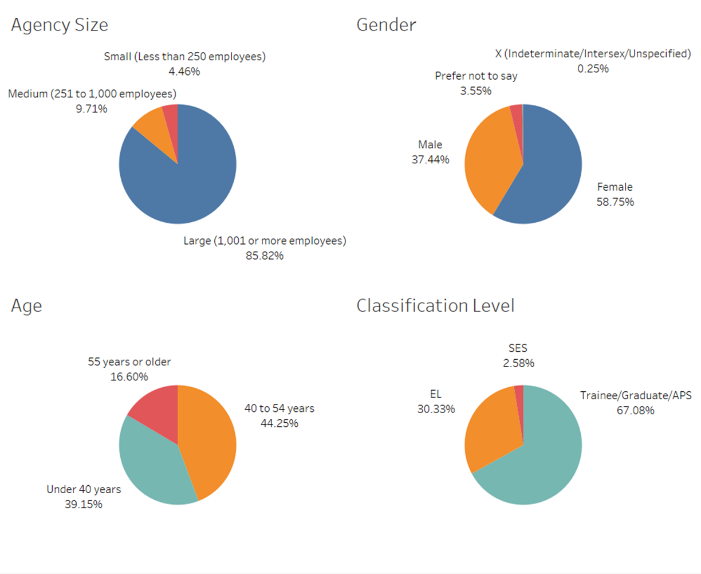


Figure 1 - Demographic Profiles of the Participants

1. **Missing Data**

* MCAR and MNAR

Because of typical characteristics of census data, high proportion of data is found missing. The types of missing data are identified as missing completely at random (MCAR) and missing not at random (MNAR). The missing values falls in the MNAR are discarded with column-wise cut-off of 15% or more. For MCAR, there is no need to compute the missing data because the proportion of missing values are small as 5.6%

Then the data was transformed with 40 variables that are selected for the further EFA and CA analysis. In addition, only Likert scale questions are selected and analyzed. Missing value imputations are not applied because of accuracy issues. The final dataset selected for the further analysis consists of 56,820 observations and 40 questions (variables).

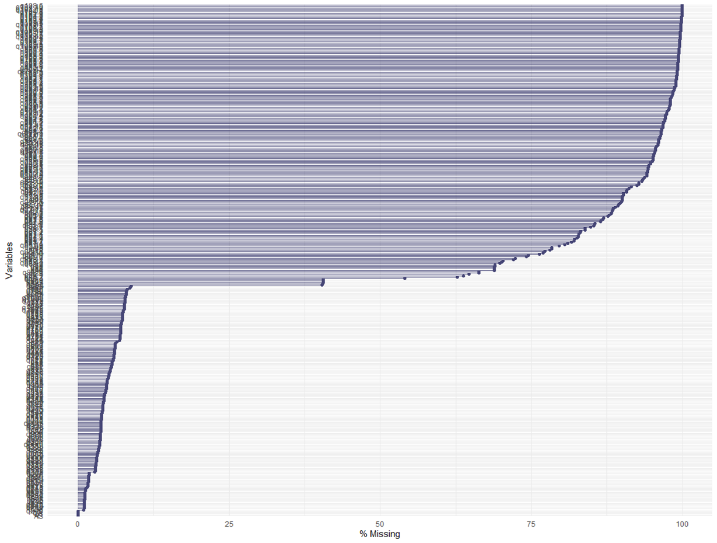
****

Figure 2 - Percentage Missing of the Data

1. **Encoding Categorical Variables to Numeric**

* Likert Scale

The data sampled from the selected questions consist of ordinal variables. Each possible response is ranged from either “Strongly Agree” to “Strongly Disagree” or “Never” to “Always”. The magnitude of agreement and frequency of event are converted from character type to numeric type. Therefore, the outcome of data consists of only the numeric values.

1. **Normality**

* Likert Scale

The dataset is only consisted with ordinal variables (Likert scale). Because the distance between responses is not measurable in Likert scale or ordinal data type, the differences between each level of scale are not necessarily equal. i.e., It is not possible to make average of “strongly agree” and “agree”. Therefore, normality of data in Likert scale does not represent the measures of central tendency for data analysis. In addition, if responses are clustered at either of extreme, the mean may be shown in the neutral or middle response. This will not make accurate representation of the data.

Because of these issues, some experts recommend using the median as the measure of central tendency for Likert scale data. However, the other experts asserted that if sample size is at least 5 to 10 observations per group in a normally distributed (or nearly normal), parametric tests can be also used in Likert scale ordinal data.

Refer to figure 3 below for the skewness and kurtosis. The value of skewness indicates not many questions are substantially skewed distribution as not many are over +1, and kurtosis shows some heavy and light tailed indications.

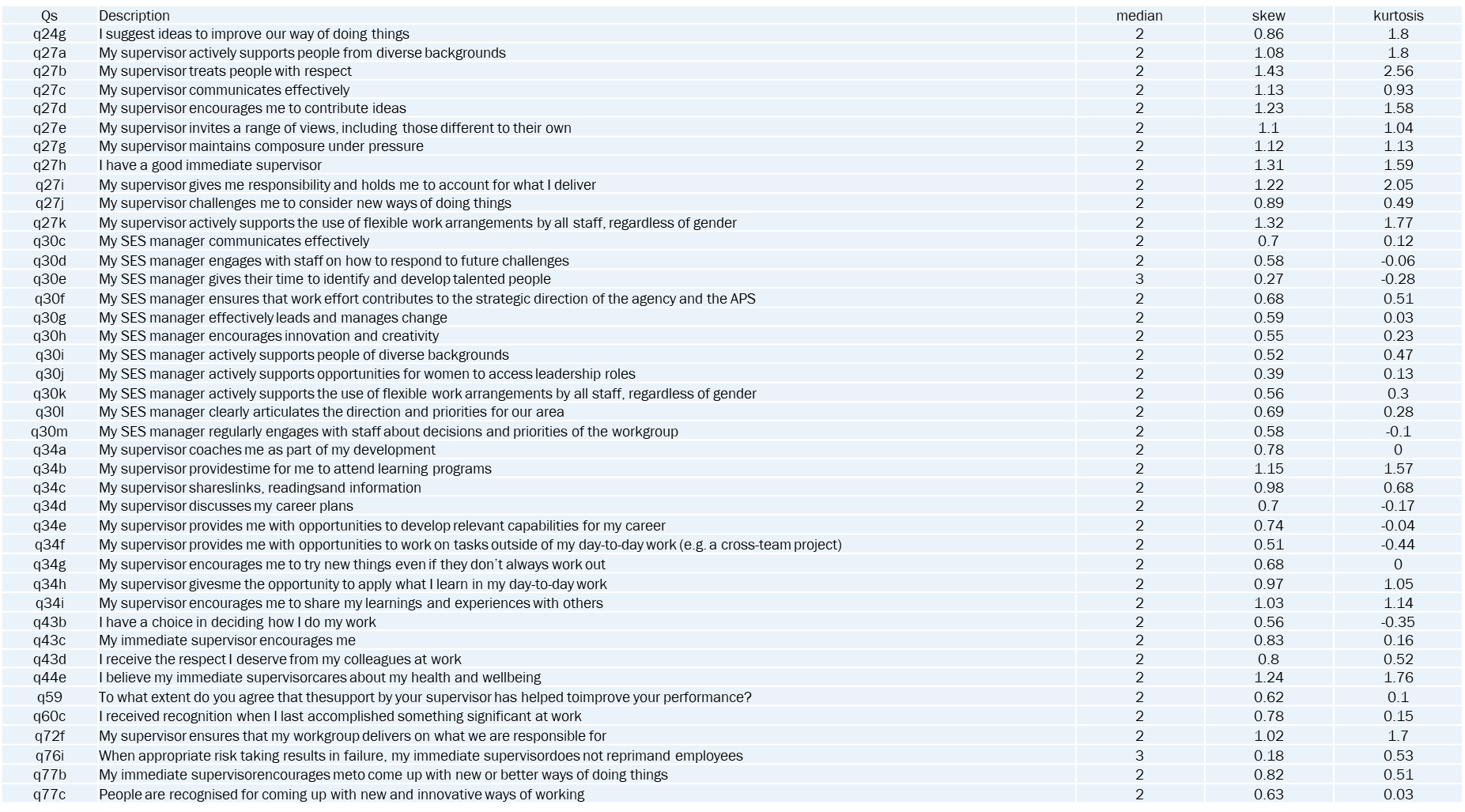


Figure 3 - Median, Skewness and Kurtosis of the Selected Variables

1. **Outliers**

* Mahalanobis Distance and Limitations

The Mahalanobis distance is widely used to detect the outliers. However, in this analysis, the outliers found by the Mahalanobis distance are not treated as outliers because the data only consists of Likert scale values. For example, values such as “Strongly disagree” and “Strongly agree” are not representation of outliers. Outliers are defined as observations that are at an abnormal distance from other values. However, the scale of responses in the dataset is pre-designed to certain number of scale points such as from 1 to 5 in “Strongly disagree” to “Strongly agree” scale. Therefore, there is no need to remove outliers, and the outliers are kept for the further analysis.

**Exploratory Factor Analysis on APSC 2019 Survey Data**

**Data Description:**

The rationale to map explanatory variables from the data was focused to satisfy two criteria:

1. Choosing all such variables which indicates an employee’s opinion about his supervisor and SES.
2. To reach the end goal of the analysis, make sure to choose the variables that explained transformation and consideration leadership styles.

40 questions were mapped from the APSC 2019 survey data to determine the leadership styles that participants regard as significant for innovation.

*\*Please refer the “Mapping of variables.xlsx” for details on the 40 selected variables.*

The chosen dataset has 104471 observations. “Q7: What is your current, actual classification level?” was used to discriminate between responses of “Trainee/Graduate/APS” and “EL and SES”. The responses of “Trainee/Graduate/APS” totaled to 70080 and the remaining were 34391.

All these responses of the candidates from the “EL and SES” level was excluded from the dataset because the agenda of the study was to determine the leadership styles and including responses of candidates who presently are at leadership level would create a bias and affect the authenticity of the study on leadership.

Additionally, all selected features had about 1% to 7% missing values. All observations with at least 1 missing value were dropped. The resulting dataset has a shape of (56820, 40).

**Exploratory Factor Analysis**

**Factorability of the data:** To perform Factor analysis, we have to satisfy 3 key assumptions: Sampling adequacy or an appropriate number of observations relative to the number of variables being examined, Sphericity or the existence of the identity matrix, Positive determinant of the correlation or variance-covariance matrices.

1. **Sampling adequacy:** The Kaiser-Meyer-Olkin (KMO) is used to measure sampling adequacy. According to KMO, data is factorable when the KMO is above the minimum acceptable level of 0.60 (thumb rule). From this test, the Overall MSA, measure of sampling adequacy is found to be 0.98 >> than thumb rule of 0.6. Therefore, the sample is statistically adequate and factor analysis can be performed on this dataset which would give fruitful results.

Text

Description automatically generated with medium confidence

1. **Bartlett’s Test of Sphericity:** Bartlett’s test of sphericity is used to determine the multivariate normality of the variables. This test also validates the hypothesis that the correlation matrix is an identity matrix (i.e., a spherical set of multivariate data), (Lattin et al., 2003)

The hypotheses for this test are as following:

* H0 = The variables correlation matrix is an identity matrix (or) the variables are orthogonal, i.e. not correlated or 0 correlation among variables.
* Ha = the variables are not orthogonal, i.e. they are correlated enough to where the correlation matrix diverges significantly from the identity matrix

A chi square value = 2521581 was obtained, p value = 0 <0.05. Therefore, the data were approximately multivariate normal. Due to rejecting the null, the result also confirms that the correlation matrix could not be construed as an identity matrix and therefore, was sufficient to test the factor analysis

Logo

Description automatically generated with low confidence

1. **Positive determinant** of the correlation or variance-covariance matrices:



Clearly, the determinant of the correlation matrix is positive. Thus, we have thus satisfied all the assumptions for factor analysis and can proceed with the analysis.

**Factor Analysis:**

**Number of Factors using Scree plot:**

We are using the eigen value criteria (choosing factors having eigen value greater than 1) to determine the number of factors. Scree plot along with parallel analysis indicates that 4 factors have an eigen value >1. However, after testing the FA with 2,3,4 factors, we determined that 4th factor does not add much (about 2%) to the cumulative proportion. Therefore, we are limiting the number of factors to 3.

Graphical user interface, application

Description automatically generated with medium confidence

Text

Description automatically generated

The Eigen value of a particular factor explains the number of variables that are decomposed into that factor and subsequently, the number of variables explained by the factor. For example, the eigen value of 1st factor is 20 which means that this factor explains the information of 20 variables of the dataset.

Factor analysis was performed using the fa() method of the psych package as well as the factanal() method using “oblimin” rotation as we believe that there is a correlation in the factors.

**Outcomes/ Inferences from the analysis:**

1. Individual **proportion of variance** explained by 3 factors: 0.33, 0.35, 0.32 respectively.
2. **Cumulative proportion** explained: 0.66. 66% of the total variance is explained by the 3 factors.
3. **Adequacy Test**:

* RMSR (root mean square of the residuals) value = 0.02. Acceptable value should be close to 0.
* Tucker Lewis Index of factoring reliability = 0.917. Acceptable value must be >0.9
* RMSEA - root mean square error of approximation = 0.065. RMSEA is an absolute fit index. It assesses how far a hypothesized model is from a perfect model. RMSEA values less than 0.05 are good, values between 0.05 and 0.08 are acceptable, values between 0.08 and 0.1 are marginal, and values greater than 0.1 are poor. Our model fits under the category of ‘Acceptable’ model.

Text

Description automatically generated

A picture containing text

Description automatically generated

1. **Loadings:** The questionnaires that are loaded onto the respective factors as determined by keeping a threshold of 0.5 are as below:

Table

Description automatically generated

The keywords / characteristics loaded onto:

**1st factor** = My supervisor coaches me, development, provides time for me to attend learning programs, shares links, readings and information, discusses my career plans, develop relevant capabilities for my career, opportunity to apply what I learn in my day-to-day work, etc.

**2nd factor** = My SES manager communicates effectively, respond to future challenges, develop talented people, contributes to the strategic direction of the agency, effectively leads and manages change, encourages innovation and creativity, supports people of diverse backgrounds, supports opportunities for women to access leadership roles, flexible work arrangements, articulates the direction and priorities, decisions and priorities of the workgroup.

**3rd factor** = My supervisor supports people, respect, communicates effectively, encourages me to contribute ideas, invites a range of view, use of flexible work arrangements, composure under pressure.

1. **Naming the Factors / Leadership styles identified:**

Based on the characteristics loaded onto the factors, we came up with the following leadership styles for the respective factors:

1. **Coaching Leadership:** A type which is characterized by collaboration, guidance, and support. Leaders with coaching style are focused on bringing out the best of their teams by guiding then through goals and obstacles.

The advantages of coaching style are, it encourages two-way communication between leaders and employees, involves lot of constructive feedback, facilitates personal and professional growth, focusses on being supportive, creates opportunities for growth and career development through learning.

1. **Transformation Leadership:** In this style, a leader inspires his or her followers with a vision and then encourages and empowers them to achieve it. The leader also serves as a role model for the vision.

The advantages of transformation style are, the leadership places high value on the corporate vision, high morale of the employees due to continuous motivation, support from employees due to inspiration from leadership and places high value on relationships.

1. **Consideration Leadership:** Consideration is the extent to which a leader exhibits for the welfare of the members of the group. This style is oriented towards interpersonal relationships and mutual trust. Some characteristics of the leaders of consideration style are being people oriented, friendly, treating all members as his/her equal, concerned about personal welfare of the team members, making him/her accessible for group members.

**FA diagram:**

A picture containing diagram

Description automatically generated

**Can leadership styles be used to predict employees that strive for creativity and innovation?**

To answer this, the scores of the 3 factors obtained by the factor analysis were placed in a data frame. The variable 24g: “I suggest ideas to improve our way of doing thing” was chosen as a response because it directly implies if an employee will strive for creativity and innovation in his workplace. The resulting data frame is as shown below:

Graphical user interface, application

Description automatically generated

The response variable has five categories which represents employee’s answer from strongly agree to strongly disagree encoded as 1 to 5. The goal is to build a predictive model to predict the outcome of this variable using the 3-leadership style latent variables. Therefore, the categories are grouped in the following manner:

* Category 1 and category 2 will be grouped into a single category '1' which means 'Agree' or 'Yes'.
* Category 4 and category 5 will be grouped into a single category '0' which means 'Disagree' or 'No'.
* Category 3 which is Neutral, has about 8864 responses will be removed from the dataset.

After grouping, the resulting response variable had 45652 observations for the category 1 and 2223 for the category 0. To balance this dataset, under sampling was performed and a final balanced dataset was obtained with 2200 observations for each category totaling to 4400 observations.

Logistic Regression was implemented in Python on a train size of 75% of the observations and the remaining 25% were used for testing. The model was found to predict the 2 categories “yes’’ and ‘no’ with an accuracy of 67.18%, sensitivity / precision of 0.71 and specificity / recall of 0.63. Therefore, our model could be used to predict employees that strive for creativity and innovation using the 3 latent variables with an accuracy of 67.18%. The confusion matrix of the predicted responses can be seen below:

Graphical user interface, text, application, email

Description automatically generated

**Correspondence Analysis**

Correspondence Analysis is geometric approach to multivariate descriptive data analysis. It is used for visualization of row points and column points in a low dimensional space. It is a dimensional reduction method applied to a contingency table.

The correspondence analysis is done to the following areas of the questionnaire to check the employee satisfaction in:

* Leadership styles
* Innovation
* Changes within the organization
* Wellbeing

Let us take a look at the above 4 analysis in detail.

**Leadership styles**

The main leadership styles considered here are

1. Consideration leadership
2. Transformational leadership.

Consideration is the degree to which the leader shows concern and support for subordinates, looks out for their welfare, treats members as equal and displays warmth and approachability.

Transformational Leaders are those who inspire subordinates to perform and recognize goals and objectives within the organization.

The questions considered for leadership styles are: q24g,q27a,q27b,q27c,q27d,q27e,q27f,q27g,q27h,q27i,q27j,q27k,q30c,q30d,q30e,q30f,q30g,q30h,q30i,q30j,q30k,q30l,q30m,q34a,q34b,q34c,q34d,q34e,q34f,q34g,q34h,q34i,q43b,q43c,q43d,q44e,q59,q60c,q72f,q76i,q77b,q77c

The results of Correspondence analysis can be summarized as follows

Text, letter

Description automatically generated

Here, the row has 5 categories namely 1 which stands for Strongly agree, 2 which means agree, 3 for neither agree nor disagree, 4 for Disagree and 5 for Strongly disagree

The column has 42 categories because we chose 42 questions from 2019 questionnaire

We then perform a chi square test to check the association between rows and columns.

Null Hypothesis (Ho): categorical variables are independent

Alternate Hypothesis (Ha): categorical variables are not independent

Text

Description automatically generated with low confidence

Since p value < 0.05, we have enough evidence to reject Ho

The factor map is displayed as follows:

Chart, scatter chart

Description automatically generated

The above plot shows a global or overall pattern in the data. The columns (indicated in red color) are the questions related to the leadership styles and rows (indicated in blue) are the likert scale numbering in which each question lies.

The co-ordinates of origin are 0,0. As observed, most of the responses are closer to 1 and 2 where 1 means strongly agree and 2 means agree

Next, we see the eigenvalues to check how much information is retained in each of the dimensions.

Text

Description automatically generated

Visualizing it using a screeplot, we get,

Chart

Description automatically generated

Here, we can see that the first axes contribute the most, with 70.7%. To see in detail the variables contributing to axes 1, we get the following:

Chart, bar chart, histogram

Description automatically generated Chart, bar chart

Description automatically generated

Here, we see q76i, q30e, q27b, q27i, q27a, q27k, q27d, q27h, q443, q59, q77c, and q27f contributes strongly to the axes 1.

The rows that contribute the most to axes 1 are: 3 and 1. This means that most of the employees either, “Strongly agree” or “Neither agree or disagree” to a certain leadership style.

Next, we need to evaluate the fit of all rows and columns.Values closer to 1 represent a good representation. So, by checking the rows and columns separately, we can see that except the columns q77b, q27f, q24g, q43c, all other columns give a good representation.

Similarly for rows, we can see that 3 and 1 give us good representation.

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Finally, the below biplot gives the overall representation of the rows and columns that gives both good and bad representation.

Chart, scatter chart

Description automatically generated

**Innovation**

The questions that were chosen to analyze innovation are: q77a, q77b, q77c, q77d, q77e. First, we perform a chi square test to determine if there is significant association between rows and columns.

Ho : categorical variables are independent

Ha : categorical variables are not independent

Text

Description automatically generated

Since p value is less than 0.05, we reject Ho

Text, letter

Description automatically generated

The factor map is displayed as follows:

Chart, scatter chart

Description automatically generated

The above plot shows a global or overall pattern in the data. The columns (indicated in red color) are the questions related to innovation and rows (indicated in blue) are the likert scale numbering in which each question lies.

Next, we see the eigenvalues to check how much information is retained in each of the dimensions.

Text

Description automatically generated

Visualizing it using a screeplot, we get,

Chart, line chart

Description automatically generated

Here, we can see that the first axes contribute the most, with 99%. To see in detail the variables contributing to axes 1, we get the following:

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated

Columns or questions q77a and q77e contribute the most to axes 1. The rows that contribute the most to axes 1 are: 3 and 1. This means that most of the employees either, “Neither agree or disagree” or “Strongly agree” to innovation.

Next, we need to evaluate the fit of all rows and columns. Values closer to 1 represent a good representation. So, by checking the rows and columns separately, we can see that actually all the columns and row values are closer to 1.

Chart, scatter chart

Description automatically generatedChart, scatter chart, bubble chart

Description automatically generated

Taking a look at the biplot below:

Chart, scatter chart

Description automatically generated

We see all the column and row values are closer to 1. So, we can say that APS in the year 2019 has a mixed reviews of agreement and disagreement from employees corresponding to innovation.

So, to go with the highest, we can say employees agree or neither agree nor disagree with correspondence to continual improvement and recognition from the agency to notice innovation.

**Organization changes**

The questions considered for Organization changes are: q86a, q86b,q86c,q86d,q86e,q87a,q87b,q87c

First, we perform a chi square test to determine if there is significant association between rows and columns.

Ho : categorical variables are independent

Ha : categorical variables are not independent

Graphical user interface, text

Description automatically generated with medium confidence

Since p value is less than 0.05, we reject Ho

Text, letter

Description automatically generated

The factor map is displayed as follows:

Chart, scatter chart

Description automatically generated

The above plot shows a global or overall pattern in the data. The columns (indicated in red color) are the questions related to the organization changes and rows (indicated in blue) are the likert scale numbering in which each question lies.

Next, we see the eigenvalues to check how much information is retained in each of the dimensions.

Text

Description automatically generated

Visualizing it using a screeplot, we get,

Chart, line chart

Description automatically generated

Here, we can see that the first axes contribute the most, with 80.8%. To see in detail the variables contributing to axes 1, we get the following:

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated

Columns or questions q87b and q86d contribute the most to axes 1. The rows that contribute the most to axes 1 are: 2 and 5. This means that most of the employees either, “Agree” or “Strongly disagree” to organization changes.

Next, we need to evaluate the fit of all rows and columns. Values closer to 1 represent a good representation. So, by checking the rows and columns separately, we can see that actually all the columns and row values are closer to 1.

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

From the below biplot we can conclude that:

Chart, scatter chart

Description automatically generated

We see all the column and row values are closer to 1. So, we can say that APS in the year 2019 has a mixed reviews of agree and strongly disagree from employees corresponding to changes.

So, to go with the highest, we can say employees agree and strongly disagree with correspondence to positive organization changing process, happiness associated with the change, and these changes are improving efficiency.

**Wellbeing**

The questions that were chosen to analyze wellbeing are: q43a,q43b,q43c,q43d,q43e,q43f,q43g,q44a,q44b,q44c,q44d,q44e,q44f,q44g,q44h. First, we perform a chi square test to determine if there is significant association between rows and columns.

Ho : categorical variables are independent

Ha : categorical variables are not independent

Text

Description automatically generated with medium confidence

Since p value is less than 0.05, we have enough evidence to reject null hypothesis.

Text, letter

Description automatically generated

The factor map is displayed as follows:

Chart, scatter chart

Description automatically generated

The above plot shows a global or overall pattern in the data. The columns (indicated in red color) are the questions related to the wellbeing and rows (indicated in blue) are the likert scale numbering in which each question lies.

Next, we see the eigenvalues to check how much information is retained in each of the dimensions.

Text

Description automatically generated

Visualizing it using a screeplot, we get,

Chart, line chart

Description automatically generated

Here, we can see that the first axes contribute the most, with 84.3%. To see in detail the variables contributing to axes 1, we get the following:

Chart, histogram

Description automatically generatedChart, bar chart

Description automatically generated

Columns or questions q43e, q43a, q44f and q44a contribute the most to axes 1. The rows that contribute the most to axes 1 are: 4 and 1. This means that most of the employees either, “Disagree” or “Strongly agree” to Wellbeing.

Next, we need to evaluate the fit of all rows and columns. Values closer to 1 represent a good representation. So, by checking the rows and columns separately, we can see that except q44h all other columns are closer to 1.

Similarly, rows 1, 4 and 2 are closer to 1.

So, the above-mentioned rows and columns give good representation.

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

From the below biplot we can conclude that:

Chart, scatter chart

Description automatically generated

Here, except q44h and q44d, all the columns are closer to 1. However, considering the major row contribution to axes 1, Questions q43d, q44g, q43f, q44e, and q43c about wellbeing are strongly agreed by the employees and questions q43a and q43e are disagreed by the employees.

The questions related to managers and supervisors considering wellbeing of the employees is agreed the most. Even the questions disagreed here corresponds to questions related to unrealistic time pressures.

This means that the managers and supervisors in 2019 do take the wellbeing of the employees into account.

**Conclusions and Recommendations**

* Until recent years, only 2 types of leadership styles were predominant in the APS agencies which were transformational and considerate. This can also be determined from the survey questionnaires of the 2014 APSC survey where the questions related to coaching leadership like teaching, learning programs, career development, skills development, readings, and information etc. are absent. Thus, the strive for innovation and creativity at work from the employees was influenced by only the motivational characters and the considerate characters of the leaders of the agencies.
* However, from the factor analysis of the 2019 survey, we can observe a new type of leadership quality in the APSC agencies. It is evident that, the APSC has implemented employee trainings and learning programs, and is trying to motivate the employees for innovation not only by motivation and considerate qualities, but also by coaching them with newer skills and technologies thereby broadening the knowledge of the employees with the technical skills required at work.
* This newer type of coaching leadership is found to be a significant latent factor in studying an employee’s willingness to be creative and innovative at work. Thus, we recommend the APSC to continue implementing newer training and learning programs, and also use the coaching leadership style as it if found to significantly impact the employees strive for creativity and innovation.

**APPENDIX – R codes**

rm(list=ls())

getwd()

setwd("C:/Users/yolee/Desktop/LANGARA/2021-3/DANA4830/Assignment 4")

# Read data

df\_aps <- read.csv("2019-aps-employee-census-dataset.csv",

header = T,

fileEncoding="UTF-8-BOM",

na.strings=c(" "))

View(df\_aps)

sum(is.na(df\_aps))

# Check data

library(vtable)

st(df\_aps)

vtable(df\_aps)

str(df\_aps)

# Demographic profile from q1 ~ 23 but in our dataset only have from AS to q7.

# Convert into factors

df\_aps$AS <- as.factor(df\_aps$AS)

df\_aps$q1 <- as.factor(df\_aps$q1)

df\_aps$q2. <- as.factor(df\_aps$q2.)

df\_aps$q7. <- as.factor(df\_aps$q7.)

str(df\_aps[c(1:4)])

st(df\_aps[c(1:4)])

# Freq barplots

library(ggplot2)

library(tidyverse)

library(scales)

## AS

AS\_bar <- ggplot(df\_aps, aes(x = AS)) + geom\_bar() + geom\_text(stat = 'count', aes(label=..count..), vjust=-1) + ggtitle("Size of Agency")

AS\_bar

## q1

Gen\_bar <- ggplot(df\_aps, aes(x = q1), ) + geom\_bar() + geom\_text(stat = 'count', aes(label=..count..), vjust=-1) + ggtitle("Gender of Respondent")

Gen\_bar

## q2.

Age\_bar <- ggplot(df\_aps, aes(x=reorder(q2., q2., function(x)-length(x)))) + geom\_bar() + geom\_text(stat = 'count', aes(label=..count..), vjust=-1) + ggtitle("Age of Respondent")

Age\_bar

## q7.

Class\_bar <- ggplot(df\_aps, aes(x=reorder(q7., q7., function(x)-length(x)))) + geom\_bar() + geom\_text(stat = 'count', aes(label=..count..), vjust=-1) + ggtitle("Classification of Respondent")

Class\_bar

# Missing Values?

missing <- sum(is.na(df\_aps))

missing # 20631763 but including missing values in the multiple choice and tick q's

# Convert all to factors (Not sure if needed this step)

#df\_aps[sapply(df\_aps, is.character)] <- lapply(df\_aps[sapply(df\_aps, is.character)], as.factor)

#str(df\_aps)

#st(df\_aps)

# Check missing values by column

testcol <- map(df\_aps, ~mean(is.na(.)))

testcol

testcol[which (testcol > 0.15)]

# Check missing values continue

st(df\_aps)

# Subset of data that are structurally missing

which(colnames(df\_aps)=="q37.1")

which(colnames(df\_aps)=="q37.15")

which(colnames(df\_aps)=="q40.1")

which(colnames(df\_aps)=="q40.11")

which(colnames(df\_aps)=="q45.1")

which(colnames(df\_aps)=="q45.4")

which(colnames(df\_aps)=="q47")

which(colnames(df\_aps)=="q52.1")

which(colnames(df\_aps)=="q52.9")

which(colnames(df\_aps)=="q61.1")

which(colnames(df\_aps)=="q65.6")

which(colnames(df\_aps)=="q67.1")

which(colnames(df\_aps)=="q67.11")

which(colnames(df\_aps)=="q69.1")

which(colnames(df\_aps)=="q69.8")

which(colnames(df\_aps)=="q78")

which(colnames(df\_aps)=="q83.2")

which(colnames(df\_aps)=="q85.1a")

which(colnames(df\_aps)=="q85")

which(colnames(df\_aps)=="q88")

which(colnames(df\_aps)=="q93.7")

which(colnames(df\_aps)=="q97")

which(colnames(df\_aps)=="q98.9")

which(colnames(df\_aps)=="q100.1")

which(colnames(df\_aps)=="q102")

which(colnames(df\_aps)=="q104.1")

which(colnames(df\_aps)=="q108.11")

which(colnames(df\_aps)=="q25")

which(colnames(df\_aps)=="q36")

which(colnames(df\_aps)=="q39")

which(colnames(df\_aps)=="q46")

which(colnames(df\_aps)=="q48")

which(colnames(df\_aps)=="q50")

which(colnames(df\_aps)=="q51")

which(colnames(df\_aps)=="q55c")

which(colnames(df\_aps)=="q57")

which(colnames(df\_aps)=="q68")

which(colnames(df\_aps)=="q73")

which(colnames(df\_aps)=="q74")

which(colnames(df\_aps)=="q84")

which(colnames(df\_aps)=="q96")

which(colnames(df\_aps)=="q99")

which(colnames(df\_aps)=="q103")

sub\_aps <- df\_aps[-c(2, 16, 83, 84:98, 100, 101:111, 127:130, 131, 132, 133, 134, 135, 136:144, 148:150, 151, 152, 159:179, 181:191, 192, 193:200, 235:253, 254, 255:268, 277:295,300, 301:310,311, 312:332,333, 334:374)]

vtable(sub\_aps)

# Drop missing values

library(DataCombine)

sub\_aps2 <- sub\_aps %>% drop\_na()

sum(is.na(sub\_aps2))

st(sub\_aps2)

vtable(sub\_aps2)

#df\_aps2 <- df\_new %>% drop\_na(q38, q39, q43a:q44h, q46, q48:q51, q53a:q60d, q66, q68, q70:q77e)

#df\_newcheck <- df\_aps %>% drop\_na(AS:q36, q38, q39, q43a:q44h, q46, q48:q51, q53a:q60d, q66, q68, q70:q77e)

# Relevel factors

# sub\_aps2$q2. <- relevel(sub\_aps2$q2., "Under 40 years")

# Convert to character and change the values to factor level orders

#sub\_aps2[sapply(sub\_aps2, is.factor)] <- lapply(sub\_aps2[sapply(sub\_aps2, is.factor)], as.character)

sub\_aps2[sub\_aps2 == "Large (1,001 or more employees)"] = 1

sub\_aps2[sub\_aps2 == "Medium (251 to 1,000 employees)"] = 2

sub\_aps2[sub\_aps2 == "Small (Less than 250 employees)"] = 3

sub\_aps2[sub\_aps2 == "Trainee/Graduate/APS"] = 1

sub\_aps2[sub\_aps2 == "EL"] = 2

sub\_aps2[sub\_aps2 == "SES"] = 3

sub\_aps2[sub\_aps2 == "Under 40 years"] = 1

sub\_aps2[sub\_aps2 == "40 to 54 years"] = 2

sub\_aps2[sub\_aps2 == "55 years or older"] = 3

sub\_aps2[sub\_aps2 == "Very satisfied"] = 1

sub\_aps2[sub\_aps2 == "Satisfied"] = 2

sub\_aps2[sub\_aps2 == "Neither satisfied nor dissatisfied"] = 3

sub\_aps2[sub\_aps2 == "Dissatisfied"] = 4

sub\_aps2[sub\_aps2 == "Very dissatisfied"] = 5

sub\_aps2[sub\_aps2 == "Always"] = 1

sub\_aps2[sub\_aps2 == "Often"] = 2

sub\_aps2[sub\_aps2 == "Sometimes"] = 3

sub\_aps2[sub\_aps2 == "Rarely"] = 4

sub\_aps2[sub\_aps2 == "Never"] = 5

sub\_aps2[sub\_aps2 == "Strongly agree"] = 1

sub\_aps2[sub\_aps2 == "Agree"] = 2

sub\_aps2[sub\_aps2 == "Neither agree nor disagree"] = 3

sub\_aps2[sub\_aps2 == "Disagree"] = 4

sub\_aps2[sub\_aps2 == "Strongly disagree"] = 5

sub\_aps2[sub\_aps2 == "Do not know"] = 6

sub\_aps2[sub\_aps2 == "Above my classification level"] = 1

sub\_aps2[sub\_aps2 == "Appropriate for my classification level"] = 2

sub\_aps2[sub\_aps2 == "Below my classification level"] = 3

sub\_aps2[sub\_aps2 == "Not at all"] = 1

sub\_aps2[sub\_aps2 == "Very little"] = 2

sub\_aps2[sub\_aps2 == "Somewhat"] = 3

sub\_aps2[sub\_aps2 == "To a great extent"] = 4

sub\_aps2[sub\_aps2 == "To a very great extent"] = 5

# Change factor to numeric

sub\_aps2[sapply(sub\_aps2, is.character)] <- lapply(sub\_aps2[sapply(sub\_aps2, is.character)], as.numeric)

vtable(sub\_aps2)

st(sub\_aps2)

## Mahalanobis

mahalanobis(sub\_aps2, colMeans(sub\_aps2), cov(sub\_aps2))

library(psych)

library(ggplot2)

library(corrplot) #plotting correlation matrices

library(GPArotation) #methods for factor rotation

library(nFactors)

describe(sub\_aps2)

## ANOVA test for satisfaction

sub\_aps2$q24c #I am satisfied with the recognition I receive for doing a good job

sub\_aps2$q24e #I am satisfied with my nonmonetary employment conditions e.g. leave, flexible work rrangements, other benefits)

sub\_aps2$q24f #I am satisfied with the stability and security of my current job

sub\_aps2$q24i #Considering everything, I am satisfied with my job

sub\_aps2$q33f #I am satisfied with the opportunities for career progression in my agency

sub\_aps2$q38 # Considering your work and life priorities, how satisfied are you with the work-life balance in your current job?

sub\_aps2$q44a #I am satisfied with the policies/practices in place to help me manage my health and wellbeing

# sample size comparison for group 1 ("Trainee/Graduate/APS") and group 2 ("EL")

nrow(sub\_aps2[sub\_aps2$q7. == "1",]) # number of subordinates

nrow(sub\_aps2[sub\_aps2$q7. == "2",]) # number of managers

sub\_aps2$q7. <- ordered(sub\_aps2$q7., levels = c("1", "2", "3"))

sub\_aps2$q7. <- ordered(sub\_aps2$q7., levels = c("1", "2"))

aps.aov <- aov(q24c ~ q7., data = sub\_aps2)

summary(aps.aov)

# Compute t-test for each

q24c.tt <- t.test(q24c ~ q7., data = sub\_aps2, var.equal = TRUE)

q24c.tt

t.test(q24c ~ q7., data = sub\_aps2, var.equal = TRUE)

t.test(q24e ~ q7., data = sub\_aps2, var.equal = TRUE)

t.test(q24f ~ q7., data = sub\_aps2, var.equal = TRUE)

t.test(q24i ~ q7., data = sub\_aps2, var.equal = TRUE)

t.test(q33f ~ q7., data = sub\_aps2, var.equal = TRUE)

t.test(q38 ~ q7., data = sub\_aps2, var.equal = TRUE)

t.test(q44a ~ q7., data = sub\_aps2, var.equal = TRUE)

library(dplyr)

group\_by(sub\_aps2, q7.) %>%

summarise(

count = n(),

mean = mean(q24c, na.rm = TRUE),

sd = sd(q24c, na.rm = TRUE)

)

library("ggpubr")

ggboxplot(sub\_aps2, x = "q7.", y = "q24c",

color = "q7.", palette = c("#00AFBB", "#E7B800", "#FC4E07"),

order = c("1", "2", "3"),

ylab = "q24c", xlab = "Classification Level")

ggboxplot(sub\_aps2, x = "q7.", y = "q24e",

color = "q7.", palette = c("#00AFBB", "#E7B800", "#FC4E07"),

order = c("1", "2", "3"),

ylab = "q24e", xlab = "Classification Level")

ggboxplot(sub\_aps2, x = "q7.", y = "q24f",

color = "q7.", palette = c("#00AFBB", "#E7B800", "#FC4E07"),

order = c("1", "2", "3"),

ylab = "q24f", xlab = "Classification Level")

ggboxplot(sub\_aps2, x = "q7.", y = "q24i",

color = "q7.", palette = c("#00AFBB", "#E7B800", "#FC4E07"),

order = c("1", "2", "3"),

ylab = "q24i", xlab = "Classification Level")

ggboxplot(sub\_aps2, x = "q7.", y = "q33f",

color = "q7.", palette = c("#00AFBB", "#E7B800", "#FC4E07"),

order = c("1", "2", "3"),

ylab = "q33f", xlab = "Classification Level")

ggboxplot(sub\_aps2, x = "q7.", y = "q38",

color = "q7.", palette = c("#00AFBB", "#E7B800", "#FC4E07"),

order = c("1", "2", "3"),

ylab = "q38", xlab = "Classification Level")

ggboxplot(sub\_aps2, x = "q7.", y = "q44a",

color = "q7.", palette = c("#00AFBB", "#E7B800", "#FC4E07"),

order = c("1", "2", "3"),

ylab = "q44a", xlab = "Classification Level")

**EFA analysis – R Code:**

# packages and libraries:

library(dplyr)

library(tidyverse)

library(caret)

library(psych)

library(corrplot)

library(ggplot2)

library(car)

## importing the selected and cleaned 2019 survey csv file from python:

# A random sample of 4521 observation was prepared and imported from python.

data\_survey <- read.csv("/Users/MADHU/Downloads/2019\_for\_EFA.csv")

view(data\_survey)

# Looking at the data before performing the analysis:

# FACTORABILITY OF THE DATA

describe(data\_survey)

# The acceptable range for skewness and kurtosis is -1 to +1. The variables having skewness and kurtosis within this range

# are considered to be univariate normal. From the describe procedure, we can see that all variables have the values under

# the normal range and mostly varies under 1.8 to 1.9. only one variable q27b.My.supervisor.treats.people.with.respect has

# a kurtosis value of 2.57

X <- data\_survey[,-c(1)]

Y <- data\_survey[,1]

# Not considering the 1st variable as it is kept as response variable

# The Kaiser-Meyer-Olkin (KMO) is used to measure sampling adequacy:

KMO(r = cor(X))

# Overall MSA: measure of smapling adequecy is found to be 0.98 >> than thumb rule of 0.6

# (data is factorable when the KMO is above the minimum acceptable level of 0.60).

# based on this test, we can probably conduct a factor analysis as the sample is adequate

# Bartlett's Test of Sphericity:

#To determine if the data multivariate normal:

# An identity matrix is a matrix in which all of the values along the diagonal are 1 and all of the other values are 0

# H0 = The variables correlation matrix is an identity matrix (or) the variables are orthogonal,

# i.e. not correlated or 0 correlation among variables

# Ha = the variables are not orthogonal, i.e. they are correlated enough to where the

# correlation matrix diverges significantly from the identity matrix

cortest.bartlett(X)

# chi square value = 2521581, p value = 0 <0.05. Therefore the data were approximately multivariate normal.

# The result also confirmed that the correlation matrix could not be construed as an identity matrix

# and therefore, was sufficient to test the factor analysis

fafitfree <- fa(data\_survey,nfactors = ncol(X), rotate = "none")

n\_factors <- length(fafitfree$e.values)

scree <- data.frame(

Factor\_n = as.factor(1:n\_factors),

Eigenvalue = fafitfree$e.values)

windows()

ggplot(scree, aes(x = Factor\_n, y = Eigenvalue, group = 1)) +

geom\_point() + geom\_line() +

xlab("Number of factors") +

ylab("Initial eigenvalue") +

labs( title = "Scree Plot",

subtitle = "(Based on the unreduced correlation matrix)")

# This scree plot shows that the first 3 factors account for most of the total variability in data (given by the eigenvalues).

# The eigenvalues for the first 3 factors are all greater than 1.The remaining factors account for a very small proportion

# of the variability and are likely unimportant

# Factor analysis using fa method:

fa.none <- fa(r=X,

nfactors = 3,

# covar = FALSE, SMC = TRUE,

fm='pa', # type of factor analysis we want to use ("pa" is principal axis factoring)

max.iter=100, # (50 is the default, but we have changed it to 100

rotate='varimax') # none rotation

print(fa.none)

# Factor analysis using the factanal method

factanal.none <- factanal(X, factors=3, scores = c("regression"), rotation = "varimax")

print(factanal.none, digits = 2, cutoff = 0.5, sort=TRUE)

# Graph Factor Loading Matrices

fa.diagram(fa.none)

factanal.none$scores

# creating a data of scores from factor analysis with response variable:

fadata <- cbind(data\_survey["q24g.I.suggest.ideas.to.improve.our.way.of.doing.things"], factanal.none$scores)

#Labeling the data

names(fadata) <- c("q24g.I.suggest.ideas.to.improve.our.way.of.doing.things","F1","F2","F3")

head(fadata)

# Splitting the data to train and test set

#Splitting the data 70:30

set.seed(100)

indices= sample(1:nrow(fadata), 0.7\*nrow(fadata))

train=fadata[indices,]

test = fadata[-indices,]

# Regression Model using train data

model.fa.score = lm(q24g.I.suggest.ideas.to.improve.our.way.of.doing.things~., train)

summary(model.fa.score)

# Our model equation can be written as: Y = 2.057074 + 0.082465\* F1 + 0.073912\* F2 + 0.135969\* F3

# Check vif

vif(model.fa.score)

EFA Analysis – Python code:

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

pd.set\_option('display.max\_columns', None)

import warnings

warnings.filterwarnings('ignore')

from collections import Counter

dataset = pd.read\_csv('C:\\E DRIVE\\LANGARA COLLEGE\\DANA 4830-001\\Team project\\2019-aps-employee-census-dataset.csv', na\_values=" ")

dataset.head()

# selecting only the questions that are related to leadership styles and which will be used in factor analysis:

selected = dataset.loc[:,['q7@','q24g','q27a','q27b','q27c','q27d','q27e','q27g','q27h','q27i','q27j','q27k'

,'q30c','q30d','q30e','q30f','q30g','q30h','q30i','q30j','q30k','q30l','q30m',

'q34a','q34b','q34c','q34d','q34e','q34f','q34g','q34h','q34i','q43b','q43c','q43d'

,'q44e','q59','q60c','q72f','q76i','q77b','q77c']]

selected.head()

selected.replace({'Strongly agree':1,

'Agree':2,

'Neither agree nor disagree':3,

'Do not know':3,

'Disagree':4,

'Strongly disagree':5}, inplace=True)

selected.replace({'Always':1,'Often':2,'Sometimes':3,'Rarely':4,'Never':5}, inplace=True)

# Question 32 c and e has an option 'Do not know'. for simplification, lets also replace this to 'Neither agree nor disagree'

selected.head()

# From question 7@, What is your current, actual classification level?, lets have only those responses who are from trainees

# to APS level. removing all the responses of executive and senior executive levels

selected = selected[selected['q7@'] == 'Trainee/Graduate/APS']

# Now we do not need to 'q7@' column in our dataset

selected.drop('q7@', axis=1, inplace=True)

selected.shape

# Now we are left with 70080 responses and 41 variables

# Determining percentage of missing values in the dataset:

# Percentage of missing values in each column

for feature in selected.columns:

print(feature, "|| Missing % =", np.round((selected[feature].isnull().sum()/len(selected[feature])\*100),2))

# Dropping all missing values observations from our dataset:

selected.dropna(inplace=True)

selected.shape

# changing all float values to integer

selected = selected.convert\_dtypes()

selected.head()

selected\_cols = selected.columns.to\_list()

selected\_cols

#to\_replace = {'q24g':'q24g I suggest ideas to improve our way of doing things',

'q27a':'q27a My supervisor actively supports people from diverse backgrounds',

'q27b':'q27b My supervisor treats people with respect',

'q27c':'q27c My supervisor communicates effectively',

'q27d':'q27d My supervisor encourages me to contribute ideas',

'q27e':'q27e My supervisor invites a range of views, including those different to their own',

'q27g':'q27g My supervisor maintains composure under pressure',

'q27h':'q27h I have a good immediate supervisor',

'q27i':'q27i My supervisor gives me responsibility and holds me to account for what I deliver',

'q27j':'q27j My supervisor challenges me to consider new ways of doing things',

'q27k':'q27k My supervisor actively supports the use of flexible work arrangements by all staff, regardless of gender',

'q30c':'q30c My SES manager communicates effectively',

'q30d':'q30d My SES manager engages with staff on how to respond to future challenges',

'q30e':'q30e My SES manager gives their time to identify and develop talented people',

'q30f':'q30f My SES manager ensures that work effort contributes to the strategic direction of the agency and the APS',

'q30g':'q30g My SES manager effectively leads and manages change',

'q30h':'q30h My SES manager encourages innovation and creativity',

'q30i':'q30i My SES manager actively supports people of diverse backgrounds',

'q30j':'q30j My SES manager actively supports opportunities for women to access leadership roles',

'q30k':'q30k My SES manager actively supports the use of flexible work arrangements by all staff, regardless of gender',

'q30l':'q30l My SES manager clearly articulates the direction and priorities for our area',

'q30m':'q30m My SES manager regularly engages with staff about decisions and priorities of the workgroup',

'q34a':'q34a My supervisor coaches me as part of my development',

'q34b':'q34b My supervisor provides time for me to attend learning programs',

'q34c':'q34c My supervisor shares links, readings and information',

'q34d':'q34d My supervisor discusses my career plans',

'q34e':'q34e My supervisor provides me with opportunities to develop relevant capabilities for my career',

'q34f':'q34f My supervisor provides me with opportunities to work on tasks outside of my day-to-day work',

'q34g':'q34g My supervisor encourages me to try new things even if they don’t always work out',

'q34h':'q34h My supervisor gives me the opportunity to apply what I learn in my day-to-day work',

'q34i':'q34i My supervisor encourages me to share my learnings and experiences with others',

'q43b':'q43b I have a choice in deciding how I do my work',

'q43c':'q43c My immediate supervisor encourages me',

'q43d':'q43d I receive the respect I deserve from my colleagues at work',

'q44e':'q44e I believe my immediate supervisor cares about my health and wellbeing',

'q59':'q59 To what extent do you agree that the support by your supervisor has helped to improve your performance?',

'q60c':'q60c I received recognition when I last accomplished something significant at work',

'q72f':'q72f My supervisor ensures that my workgroup delivers on what we are responsible for',

'q76i':'q76i When appropriate risk taking results in failure, my immediate supervisor does not reprimand employees',

'q77b':'q77b My immediate supervisor encourages me to come up with new or better ways of doing things',

'q77c':'q77c People are recognised for coming up with new and innovative ways of working'}

selected.rename(columns=to\_replace, inplace=True)

selected.head()

# importing the 2019 file with factors:

dataset\_fa = pd.read\_csv('C:\\E DRIVE\\LANGARA COLLEGE\\DANA 4830-001\\Team project\\2019\_with\_factors.csv')

dataset\_fa.head()

# Value counts for the response variable

Counter(dataset\_fa['q24g.I.suggest.ideas.to.improve.our.way.of.doing.things'])

# The variable "I suggest ideas to improve our way of doing things" indicate an employee's willingness to

# strive for creativity and innovation at workplace.

# We now want to build a model to predict the outcome of this variable using the 3 leadership style latent variables

# Lets re structure the response variable as following:

# category 1 and category 2 will be grouped into a single category '1' which means 'Agree' or 'Yes'

# category 4 and category 5 will be grouped into a single category '0' which means 'Disagree' or 'No'

# Categry 3 which has about 8864 responses will be removed from the dataset because we only want to predict if the

# employee will strive for creativity or not. Therefore considering only 2 categories

dataset\_fa = dataset\_fa[dataset\_fa['q24g.I.suggest.ideas.to.improve.our.way.of.doing.things']!=3]

dataset\_fa['q24g.I.suggest.ideas.to.improve.our.way.of.doing.things'].replace({2:1, 4:0, 5:0}, inplace=True)

Counter(dataset\_fa['q24g.I.suggest.ideas.to.improve.our.way.of.doing.things'])

# Since the '1' observations is too large compared to '0'

# lets create a random sample of size 4000 with 1 = 2000, 0 = 2000 ( Under Sampling)

dataset\_fa\_sample = dataset\_fa.groupby('q24g.I.suggest.ideas.to.improve.our.way.of.doing.things').apply(lambda x: x.sample(n=2200)).reset\_index(drop = True)

Counter(dataset\_fa\_sample['q24g.I.suggest.ideas.to.improve.our.way.of.doing.things'])

# Implementing a logistic regression to classify the responses

from sklearn.model\_selection import train\_test\_split

from sklearn.linear\_model import LogisticRegression

from sklearn.metrics import accuracy\_score, roc\_auc\_score, classification\_report, confusion\_matrix

from sklearn.neighbors import KNeighborsClassifier

from sklearn.ensemble import RandomForestClassifier

X = dataset\_fa\_sample.drop(['q24g.I.suggest.ideas.to.improve.our.way.of.doing.things'], axis=1)

y = dataset\_fa\_sample['q24g.I.suggest.ideas.to.improve.our.way.of.doing.things']

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.25, random\_state=10)

print(X\_train.shape, X\_test.shape, y\_train.shape, y\_test.shape)

model = LogisticRegression(random\_state=0)

model.fit(X\_train, y\_train)

y\_predict = model.predict(X\_test)

accuracy\_score(y\_test, y\_predict)

print(Counter(y\_test))

pd.crosstab(y\_test, y\_predict)

**Correspondence analysis:**

rm(list=ls())

ca <- read.csv("/Users/MADHU/Downloads/aps\_2019\_clean.csv")

library(tidyverse)

library(dplyr)

library(naniar)

View(ca)

str(ca)

###############

#Leadership styles compared on EFA#

###############

ca\_on\_efa <- ca %>% select(q24g,q27a,q27b,q27c,q27d,q27e,q27f,q27g,q27h,q27i,q27j,q27k,q30c,q30d,q30e,q30f,q30g,q30h,q30i,q30j,q30k,q30l,q30m,q34a,q34b,q34c,q34d,q34e,q34f,q34g,q34h,q34i,q43b,q43c,q43d,q44e,q59,q60c,q72f,q76i,q77b,q77c)

ca\_on\_efa19 <- table(stack(ca\_on\_efa))

chisq <- chisq.test(ca\_on\_efa19)

chisq

library("FactoMineR")

library("factoextra")

#symmetric plot shows overall pattern within the data of the selected questions

ca\_efa19 <- CA(ca\_on\_efa19, graph = TRUE)

print(ca\_efa19)

#eigenvalue

eig.val <- get\_eigenvalue(ca\_efa19)

eig.val

#we can use scree plot to determine the number of dimensions

fviz\_screeplot(ca\_efa19,addlabels=T) +

geom\_hline(yintercept=50,linetype=2,color="red")

#contribution of axis 1 - column

fviz\_contrib(ca\_efa19, choice="col",axes=1)

#contribution of axis 1 - row

fviz\_contrib(ca\_efa19, choice="row",axes=1)

row <- get\_ca\_row(ca\_efa19)

row

head(row$coord)

head(row$cos2)

#rows in CA

fviz\_ca\_col(ca\_efa19, col.col = "cos2")

#columns in CA

fviz\_ca\_row(ca\_efa19, col.row = "cos2",repel = TRUE)

#biplot in CA

fviz\_ca\_biplot(ca\_efa19,repel=T,map="rowprincipal", col.col = "cos2")

##Visualize the results for rows

fviz\_ca\_row(ca\_efa19)

##Visualize the results for columns

fviz\_ca\_col(ca\_efa19)

###############

#Innovation#

###############

innovation <- ca %>% select(q77a, q77b, q77c, q77d, q77e)

innovation19 <- table(stack(innovation))

chisq\_innovation <- chisq.test(innovation19)

chisq\_innovation

#symmetric plot shows overall pattern within the data of the selected questions

ca\_innovation19 <- CA(innovation19, graph = TRUE)

print(ca\_efa19)

#eigenvalue

eig.val <- get\_eigenvalue(ca\_innovation19)

eig.val

#we can use scree plot to determine the number of dimensions

fviz\_screeplot(ca\_innovation19,addlabels=T) +

geom\_hline(yintercept=50,linetype=2,color="red")

#contribution of axis 1 - column

fviz\_contrib(ca\_innovation19, choice="col",axes=1)

#contribution of axis 1 - row

fviz\_contrib(ca\_innovation19, choice="row",axes=1)

row <- get\_ca\_row(ca\_innovation19)

row

head(row$coord)

head(row$cos2)

#rows in CA

fviz\_ca\_col(ca\_innovation19, col.col = "cos2")

#columns in CA

fviz\_ca\_row(ca\_innovation19, col.row = "cos2",repel = TRUE)

#biplot in CA

fviz\_ca\_biplot(ca\_innovation19,repel=T,map="rowprincipal", col.col = "cos2")

##Visualize the results for rows

fviz\_ca\_row(ca\_innovation19)

##Visualize the results for columns

fviz\_ca\_col(ca\_innovation19)

###############

#Organisation changes#

###############

change <- ca %>% select(q86a,q86b,q86c,q86d,q86e,q87a,q87b,q87c)

change19 <- table(stack(change))

chisq\_change <- chisq.test(change19)

chisq\_change

#symmetric plot shows overall pattern within the data of the selected questions

ca\_change19 <- CA(change19, graph = TRUE)

print(ca\_change19)

#eigenvalue

eig.val <- get\_eigenvalue(ca\_change19)

eig.val

#we can use scree plot to determine the number of dimensions

fviz\_screeplot(ca\_change19,addlabels=T) +

geom\_hline(yintercept=50,linetype=2,color="red")

#contribution of axis 1 - column

fviz\_contrib(ca\_change19, choice="col",axes=1)

#contribution of axis 1 - row

fviz\_contrib(ca\_change19, choice="row",axes=1)

row <- get\_ca\_row(ca\_change19)

row

head(row$coord)

head(row$cos2)

#rows in CA

fviz\_ca\_col(ca\_change19, col.col = "cos2")

#columns in CA

fviz\_ca\_row(ca\_innovation19, col.row = "cos2",repel = TRUE)

#biplot in CA

fviz\_ca\_biplot(ca\_change19,repel=T,map="rowprincipal", col.col = "cos2")

##Visualize the results for rows

fviz\_ca\_row(ca\_change19)

##Visualize the results for columns

fviz\_ca\_col(ca\_change19)

###############

#Wellbeing#

###############

wellbeing <- ca %>% select(q43a,q43b,q43c,q43d,q43e,q43f,q43g,q44a,q44b,q44c,q44d,q44e,q44f,q44g,q44h)

wellbeing19 <- table(stack(wellbeing))

chisq\_wellbeing <- chisq.test(wellbeing19)

chisq\_wellbeing

#symmetric plot shows overall pattern within the data of the selected questions

ca\_wellbeing19 <- CA(wellbeing19, graph = TRUE)

print(ca\_wellbeing19)

#eigenvalue

eig.val <- get\_eigenvalue(ca\_wellbeing19)

eig.val

#we can use scree plot to determine the number of dimensions

fviz\_screeplot(ca\_wellbeing19,addlabels=T) +

geom\_hline(yintercept=50,linetype=2,color="red")

#contribution of axis 1 - column

fviz\_contrib(ca\_wellbeing19, choice="col",axes=1)

#contribution of axis 1 - row

fviz\_contrib(ca\_wellbeing19, choice="row",axes=1)

row <- get\_ca\_row(ca\_wellbeing19)

row

head(row$coord)

head(row$cos2)

#rows in CA

fviz\_ca\_col(ca\_wellbeing19, col.col = "cos2")

#columns in CA

fviz\_ca\_row(ca\_wellbeing19, col.row = "cos2",repel = TRUE)

#biplot in CA

fviz\_ca\_biplot(ca\_wellbeing19,repel=T,map="rowprincipal", col.col = "cos2")

##Visualize the results for rows

fviz\_ca\_row(ca\_wellbeing19)

##Visualize the results for columns

fviz\_ca\_col(ca\_wellbeing19)