Statistical Analysis Plan

|  |  |
| --- | --- |
| TRIAL FULL TITLE | Employee Engagement |
| SAP VERSION | Version 1 |
| SAP VERSION DATE | 05/01/2024 |
| TRIAL STATISTICIAN | Gianna LaFrance, Jaiden Neff, Samantha Gouveia |
| Protocol Version (SAP associated with) | NA |
| TRIAL PRINCIPAL INVESTIGATOR | NA |
| SAP AUTHOR(s) | Gianna LaFrance, Jaiden Neff, Samantha Gouveia |

# SAP Signatures

I give my approval for the attached SAP entitled Employee Engagement dated 5/1/2024

**Statistician (Author)**

Name: Gianna LaFrance

Date: 5/1/2024

**Statistician (Author)**

Name: Jaiden Neff

Date: 5/1/2024

**Statistician (Author)**

Name: Samantha Gouveia

Date: 5/1/2024

# Table of Contents

[1 SAP Signatures 1](#_Toc2062624204)

[2 Table of Contents 1](#_Toc1561064758)

[3 Abbreviations and Definitions 2](#_Toc1834363141)

[4 Introduction 2](#_Toc404893666)

[4.1 Preface 2](#_Toc91518821)

[4.2 Scope of the analyses 2](#_Toc863224423)

[5 Study Objectives 2](#_Toc1025071773)

[6 Study Methods 2](#_Toc1041053265)

[6.1 Exploratory Data Analysis 2](#_Toc451866304)

[6.2 Study Population 2](#_Toc337673497)

[7 Data Processing 2](#_Toc441295171)

[8 General Analysis 2](#_Toc2120413977)

[8.1 Principal Component Analysis 2](#_Toc1972988244)

[8.2 Multinomial Logistic Regression 2](#_Toc1751146044)

[8.3 Missing Data 2](#_Toc2126394512)

[9 Summary of Study Data 2](#_Toc1169989477)

[9.1 Baseline Variables 2](#_Toc349357759)

[10 Reporting Conventions 2](#_Toc187915128)

[11 Summary of Changes to the Protocol and/or SAP 2](#_Toc198969872)

[12 References 2](#_Toc1650411554)

# Abbreviations and Definitions

Here is a list of abbreviations that will be used throughout the report. We will spell out abbreviated terms and indicate the abbreviation in parentheses at their first appearance in the text.

|  |  |
| --- | --- |
| PCA | Principal Component Analysis |
| SAP | Statistical Analysis Plan |

# Introduction

The data set has 1470 rows and 35 columns.

The data set contains data like age, gender, job satisfaction, environment satisfaction, education field, job role, income, overtime, percentage salary hike, tenure, training time, years in current role, relationship status, and more.

The data set is therefore great to predict turnover, or to simply find differences between the group that stayed or that left.

## Preface

Our goal for this study is to see the effect that factors have on the overall job involvement at this company. We will conduct an exploratory data analysis, a data cleaning step, and a full analysis.

## Scope of the analyses

We will then complete a PCA as well as a Multinomial logistic regression to test correlated variables and see how they affect job engagement/ involvement.

# Study Objectives

Our goal for this study is to see the effect that factors have on the overall job involvement at our company. We plan to run statistical tests and observe the data to draw conclusions that will lead us to make informed decisions about how to improve job involvement.

# Study Methods

## Exploratory Data Analysis

(ICH E3;9)

In our exploratory data analysis, we will take a deeper look at the data for job involvement to gauge the engagement of employees.

We will look at some summary statistics and frequency distributions.

We will create a frequency graph and look at how engaged our employees are overall.

We will also look at the summary statistics of other factors that we think might have a significant impact on the overall engagement of our employees such as Environment Satisfaction and Job Satisfaction

We will do a few associations graphs to see if there is any correlation between the factors to see what factors are highly correlated.

We will also describe dataset variables related to job involvement, environment satisfaction, and job satisfaction.

Address missing values and normalize/scale variables if necessary, so that we can start on data cleaning and processing.

## Study Population

The Study Population for this analysis will use these variables

Job Involvement- Likert scale 1-4 from low to very high

Environment Satisfaction- Likert scale 1-4 from low to very high

Job Satisfaction- Likert scale 1-4 from low to very high

Work Life Balance-Likert scale 1-4 from low to very high

Monthly Income-

Years Since Last Promotion-

Years At Company-

Age-

Number Companies Worked-

Department-

Job Level

Years With Current Manager-

Hourly Rate-

# Data Processing

**7.1 Data Cleaning**

Data cleaning will be done to remove any missing variables from the data set. Since the size of the data is quite large, imputing missing values is not crucial to the sample size used in the analysis. Once filtering is complete, we find there are no missing values.

Job Involvement will be turned into a categorical variable going from 1 to 4 with 1 being low engagement and 4 being high engagement.

To perform the Principal Component Analysis (PCA) we must standardize the variables onto the same scale. If variables are on different scales, some may scale may dominate others in the analysis. This can lead to a biased result. By standardizing we are ensuring each variable contributes equally to the analysis.

**7.2 Identifying Outliers**

Conducting exploratory data analysis will be done to look at the relationship between Job involvement and the explanatory variables. This will show if there are any outliers, the tendencies of the data, and show the data distribution.

A check for multicollinearity will be done on the explanatory variables using a VIF analysis. Ensuring the validity of the regression analysis results.

By looking at these aspects to process the data, this will ensure the data is prepared and validated for the exploratory and inferential analysis. An enhancement in the reliability and validity of our findings will allow for making the best decision based on the results.

# General Analysis

## Principal Component Analysis

To analyze employee engagement, the variable Job Involvement will be used as the response. To pare down the number of explanatory variables, Principal Component Analysis (PCA) will be used. PCA takes a large data set with many variables per observation and reduces them to a smaller set of summary indices. These indices retain most of the information in the original set of variables. The importance and a scree plot of the PC’s will be used to determine the optimal number of PC’s to use in our model.

## Multinomial Logistic Regression

Job Involvement is recorded from 1 to 4 with 1 being low engagement and 4 being high engagement. Because this is a categorical variable, multinomial logistic regression will be used to determine which variables influence Job Involvement. Multinomial logistic regression is used when you have a categorical [dependent variable](https://www.statisticshowto.com/dependent-variable-definition/) with two or more unordered levels (i.e. two or more discrete outcomes). It is practically identical to [logistic regression](https://www.statisticshowto.com/logistic-regression/), except that you have multiple possible outcomes instead of just one.

Some assumptions of a multinomial model are that the variables are independent and not highly correlated. Our data is independent, so this will not be an issue in our model.

A VIF factor will be used to test multicollinearity. If the VIF factor is greater than 10, one of the variables will have to be removed to get rid of redundancy in the model.

## Missing Data

First, we will see how much data is missing. If advisable, missing data can be imputed through medians if enough information is available. If there are very few missing values, the missing observations will be deleted. If a variable has too many missing values (more than 20%), then this variable will not be used in our model.

# Summary of Study Data

All continuous variables will be summarized using the following descriptive statistics: n (non-missing sample size), mean, standard deviation, median, maximum and minimum. The frequency and percentages (based on the non-missing sample size) of observed levels will be reported for all categorical measures. All summary tables will be structured with a column for Job Involvement with an explanatory variable including any missing observations.

## Baseline Variables

Job Involvement is a categorical variable with levels 1 through 4 with 1 being no/little involvement and 4 being high involvement. Our baseline for the response variable Job Involvement will be 1.

# Reporting Conventions

P-values ≥0.001 will be reported to 3 decimal places; p-values less than 0.001 will be reported as “<0.001”. The mean, standard deviation, and any other statistics other than quantiles, will be reported to one decimal place greater than the original data. Quantiles, such as median, or minimum and maximum will use the same number of decimal places as the original data. Estimated parameters, not on the same scale as raw observations (e.g. regression coefficients) will be reported to 3 significant figures.

# Summary of Changes to the Protocol and/or SAP

|  |  |
| --- | --- |
| Date change was made | Author and description of change |
| 05/1/2024 | Gianna LaFrance: Document created, Section 3,8,9,10 written |
| 05/1/2024 | Jaiden Neff: section 4,5,6 written |
| 05/1/2024 | Samantha Gouveia: section 5,7 written |
|  |  |

# References

Frost, Jim. “Principal Component Analysis Guide & Example.” *Statistics By Jim*, 29 Jan. 2023, statisticsbyjim.com/basics/principal-component-analysis/.

Tim. (2019, June 3). *Multinomial logistic regression: Definition and examples*. Statistics How To. <https://www.statisticshowto.com/multinomial-logistic-regression/>