1 Administrative Information

TRIAL FULL TITLE	An observational study on the impact of personality and other explanatory variables on Employee Turnover.
SAP VERSION	1.0
SAP VERSION DATE	4/29/2024
TRIAL STATISTICIAN	Joshua Zhong, Alice Xiang, Joshua Grant
Protocol Version (SAP associated with)	N/A
TRIAL PRINCIPAL INVESTIGATOR	N/A
SAP AUTHOR(s)	Joshua Zhong, Alice Xiang, Joshua Grant

2 SAP Signatures

We give our approval for the attached SAP and approve it as adequate in scope of the main-analyses of this observational study on Employee Turnover.

Statisticians

Name: Joshua Zhong

Signature: Joshua zhong

Date: 4/29/2024

Name: Alice Xiang

Signature:

Date: 4/29/2024

Name: Joshua Grant

Signature:

Date: 4/29/2024

3 Table of Contents

4 Introduction

4.1 Background

Employee turnover is defined as the total number of employees that leave a company over a certain time period, including both employees who exit voluntarily as well as those who are fired or laid off. Attrition, on the other hand, only measures employees who exit a company voluntarily (1). Employee turnover can often refer to positions that then need to be refilled in a company, which costs valuable time and resources (2). It has been estimated that the cost of replacing an employee can be from 0.5 to 2x the worker's salary (1), and can lead to other losses of time, productivity, morale, and quality of work (1, 2).

Employee turnover can be influenced by a variety of factors, such as their wages, benefit packages, amount of time off permitted, promotions, communication, relationships with other employees, and environment (1, 2). Different industries also have different turnover rates; retail, for example, has an employee turnover rate of 37%, as opposed to the U.S. national average of 20% (1).

Employee turnover can have a large impact on a company with respect to its funds, productivity, and future. Therefore, monitoring turnover rate within a company is important for its continued functioning. The ability to predict turnover rate in a company allows for the company to make better business decisions in the future and avoid unneeded costs.

4.2 Study Objectives

The purpose of this study is to determine potential intervenable factors that are associated with employee turnover. The dataset (3) includes 1129 records of 16 variables on the gender, age, wage type, industry, profession, way of travel, source of hire, management, and big five personality of employees in Russia. Each employee has a certain variable, event, that measures if the employee stayed with the company or left, either voluntarily or involuntarily. The source of the dataset was not given.

4.3 Research Questions

The research questions associated with this study consist of a primary and secondary question. The primary question is: which combination of intervenable factors consisting on gender, age, wage type, industry, profession, way of travel, source of hire, management, and big five personality of employees in Russia is associated with employee turnover? The secondary question is whether the five personality variables from the big 5 personalities test have an association with employee turnover, and if so, which variables do.

4.4 Scope of the analyses

The SAP will be the guiding document for the analysis that will be conducted over the course of this study. Analysis will be done to find what combination of intervenable factors influence employee turnover in the dataset. The results of analysis for both the primary and secondary questions will be reported through a presentation.

5 Data Pre-processing

5.1 Industry

The data has 12 distinct industries. Some of the industries contained fewer than 25 observations. This brings up concerns of sparsity as variance drastically increases if sample size is extremely small. As a result, we grouped the industries into 5 aggregate categories: Finance, Industrial, Commercial, Tech, and Other.

5.2 Profession

The data has 15 distinct professions. Some of the professions contained fewer than 20 observations. This brings up concerns of sparsity as variance drastically increases if sample size is extremely small. As a result, we grouped the professions into 6 aggregate categories: Finance/Legal, Sales/Marketing, Human/Public Relations, Technical, Education & Management, and Other.

5.3 Recoding Categorical Variables

The following categorical variables were string values: Gender, Grey wage, head gender, way, coach, industry, profession. For these variables, we recoded them into numeric values.

For Gender, “m” = 1, “f” = 0.

For greywage, “grey” = 1, “white” = 0.

For head_gender, “m” = 1, “f” = 0.

For way, “foot” = 0, “bus” = 1, “car” = 2.

For coach, “no” = 0, “yes” = 1, “my head” = 2.

For Industry, grouped into 5 aggregate industries: Finance (1), Industrial (2), Commercial (3), Tech (4), Other (5).

For Profession, grouped into 6 aggregate professions: Finance/Legal (1), Sales/Marketing (2), Human/Public Relations (3), Technical (4), Education & Management (5), Other (0).

The Big 5 personality variables were renamed accordingly: independ changed to Agreeableness, extraversion changed to Extraversion, selfcontrol = Conscientiousness, anxiety changed to Neuroticism, novator changed to Openness.

5.4 Removing Traffic Variable

The traffic variable tracks the method that the employee found the position. However, there were many concerns with the reliability of the variable in our study. The main concern was that the variables were coded in a nonsensical manner with “KA”, “empjs”, “rabrecNErab”, “recNErab”, “youjs” being incomprehensible. In addition, the observations that had these characteristics for traffic couldn’t be removed as they make up more than 70% of all observations. As a result, the variable was removed in entirety.

6 Study Methods

6.1 General Study Design and Plan

This dataset contains the results of the big five personality test for employees in russia, as well as other employment data pertaining to turnover and employee experience. Data set was designed to test association between turnovers and personality type. Potential discoveries in predictive relationships between variables may also follow.

While this study on the data may be conducted, there are certain concerns with where this study's findings may be applied to as the data's origins are relatively vague in nature.

6.2 Sample Size and Power

This data set contains 1130 observations of individual workers. A post hoc statistical power calculation was conducted on a logistic regression from the z-test family. With a two-tailed, small effect size (odds ratio of 1.3), a conservative probability of success value - Pr(Y=1|X=1) H0 - of 0.45, alpha = 0.05, attempted power of 0.80 on a standard normal X distribution yields that we have a power of 0.8000.

7 Statistical Principles

7.1 Statistical Significance and Confidence Interval

An alpha of 0.05 and a 95% confidence level will be used for all statistical test and confidence interval building respectively. The p-value, if under the significance level of 0.05, will emphasize that our association has significance and is evidenced by the data.

7.2 Principal Component Analysis

Principal Component Analysis (PCA) is a statistical technique used to simplify the complexity in high-dimensionality data while retaining trends and patterns. However, because of the nature of the variables being studied, especially with how relationships exist between the different personality variables, we cannot conduct this procedure in order to aggregate the personality variables into an index.

7.3 Logistic Regression

Logistic regression is a statistical method used for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (where there are only two possible outcomes). It is used in various fields, including medicine, economics, and social sciences for predicting the probability of occurrence of an event by fitting data to a logistic curve. It estimates the parameters of a logistic model; it is a form of binomial regression, crucial for situations where linear regression is inappropriate due to the outcome variable's categorical nature.

7.4 Stepwise Regression

Stepwise regression is a method of fitting regression models in which the choice of predictive variables is carried out by an automatic procedure. In each step, a variable is considered for addition to or subtraction from the set of explanatory variables based on some pre-specified criterion, usually a threshold of the F-statistic from a partial F-test, or AIC/BIC criteria. Specifically, bidirectional stepwise regression, also known as stepwise selection, involves both forward selection (adding the most statistically significant variable at each step) and backward elimination (removing the least significant variable) of predictors. It is a hybrid method that aims to find a balance between fitting the model and keeping it as simple as possible.

7.5 Cronbach’s Alpha

Cronbach's Alpha is a measure of internal consistency, that is, how closely related a set of items are as a group. It is considered to be a measure of scale reliability or the reliability of a psychometric instrument. A high Cronbach's alpha (usually above 0.7) indicates that the items measure the same underlying concept and suggests that the scale is reliable.

8 General Analysis Considerations

The entire sample will be used as there is no missing data and no reason to exclude any observations from the analysis.

8.1 Missing Data

There is no missing data in this data set; however, there are values in age that were originally missing but were imputed with the mean age. Since we are not certain which observations have imputed ages, we proceed with the mindset that there were no missing values nor any imputed values.

8.2 Assumptions (PCA)

We’re not using PCA to index the 5 personality variables as it requires independence between variables. In addition, this consideration is based off the theory surrounding the Big 5 Personality Tests. The Big 5 Personality Test is based on the theory that human personality can be defined by independent, separately measurable traits of which the big 5 are the most important. The key to the big 5 theory is that these traits are independent and measured through completely separate questionnaires; a PCA would be ill-advised as loading variance from multiple traits into a single principle component would destroy the ability to derive any practical interpretations of the trait scores.

In addition, the Big 5 personality traits were derived using statistical processes and already underwent factor analysis and this aggregation was already validated. Clinical psychologists started from thousands of trait descriptors. Cattell narrowed it down to 16 factors through factor analysis, but further analysis distilled these factors down to five broader components - the 5 components widely used today (4). Further aggregation is therefore strongly discouraged in this position without further theory to back up attempts to perform additional aggregation.

8.2 Assumptions (Logistic Regression)

In logistic regression analysis, several key assumptions must be met to ensure the validity of the model's conclusions. Firstly, the dependent variable should be binary or ordinal in nature, representing the occurrence or non-occurrence of an event. Multicollinearity should be checked and mitigated, as highly correlated independent variables can distort the importance and reliability of predictors. The model assumes a linear relationship between the log odds (the logit of the probability) and the independent variables, necessitating proper transformations of predictors if this linearity is not present. Logistic regression also requires large sample sizes, especially for events with lower probabilities, to accurately estimate the model parameters. Independence of observations is essential, meaning the data points should not be related or influence each other. Lastly, the absence of influential outliers that could unduly affect the model's performance is important. These assumptions are typically examined using a variety of diagnostic measures, such as Variance Inflation Factor (VIF) and residual analysis to ensure that the model provides a reliable representation of the data.

9 Summary of Study Data

This data set contains 16 columns and 1130 rows. Each row is an observation of an individual employee's information while each column is their result for one of the variables of interest. The following table gives the 14 variables in the order they are listed in the data set.

Variable Name	**Variable Type**	**Variable Subject**	**Variable Definition**
stag	Quantitative Continuous	Demographic/ Secondary Response	Normalized measure of employment length
event	Categorical Binary	Response	If the employee is still at the company or not
gender	Categorical Binary	Demographic	Employee’s gender
age	Quantitative Discrete	Demographic	Employee’s age
industry	Categorical	Demographic	What industry the employee works in
profession	Categorical	Demographic	What type of work does the employee perform
traffic	Categorical	Demographic	Through what pipeline was the employee hired

coach	Categorical	Demographic	Does the employee have a mentor at the company (debatable definition)
head_gender	Categorical Binary	Demographic	Gender of employee’s mentor (debatable definition)
greywage	Categorical Binary	Demographic	Is the employees wages taxed (white) or are their true wages unreported (gray)
way	Categorical	Demographic	Method of transportation to work
extraversion	Quantitative Continuous	Big 5	The results of the big 5 personality trait test for extraversion
independence	Quantitative Continuous	Big 5	The inverted results of the big 5 personality trait test for agreeableness
selfcontrol	Quantitative Continuous	Big 5	The results of the big 5 personality trait test for conscientiousness
anxiety	Quantitative Continuous	Big 5	The results of the big 5 personality trait test for neuroticism
novator	Quantitative Continuous	Big 5	The results of the big 5 personality trait test for openness to experience

9.1 Study Population

The study population is workers in Russia. The population covers a varied number of industries and professions. It can not be determined if any groups or types of workers are excluded from the population as it is unknown how the data sample was acquired.

9.2 Derived variables

Industry Variable:

The original 'Industry' variable has been transformed into an aggregated categorical variable with five levels. This derived variable categorizes the industries into 'Finance' (1), 'Industrial' (2), 'Retail' (3), 'Tech' (4), and 'Other' (5). The aggregation is defined as follows: Banks, Real Estate, and Insurance are classified as Finance; Manufacture, Building, and Power Generation are categorized as Industrial; Sales, Marketica, and Telecom as Retail; IT, Consult, and Telecom as Tech; and all remaining industries are categorized as Other. This categorization was performed using a case_when function in R, ensuring no industry is left ungrouped.

Profession Variable:

Similarly, the 'Profession' variable from the CRF has been converted into a derived categorical variable with six distinct levels: 'Finance/Legal' (1), 'Sales/Marketing' (2), 'Human/Public Relations' (3), 'Technical' (4), 'Education & Management' (5), and 'Other' (0). This was achieved by grouping related professions under a common code: Accounting, Finance, and Law are coded under Finance/Legal; Sales, Marketing, Business Development, and Commercial under Sales/Marketing; HR and PR under Human/Public Relations; Engineer, IT, and Consult under Technical; and Teaching and Management under Education & Management. Professions not covered by these categories are marked as Other. This process utilized a case_when function in R to assign each profession to one of the new categories, ensuring that every recorded profession is represented.

For both derived variables, the process does not employ techniques such as carrying forward values or transforming values. The derived variables are straightforward categorical groupings of existing variables without interpolation or imputation of missing observations.

9.3 Demographic and Baseline Variables

This data set contains a number of demographic variables. Variable included that will be consider demographic include gender, age, industry, profession, traffic (source of hire), greywage (taxed wage or “under the table”), and way (method of commute)

10 Analysis

The research questions and design of the study have been specified above. The outcomes of the primary research question are:

- Primary: the association of a single or combination of factors found about employees in Russia with employee turnover
- Secondary: the test accuracy of a single or combination of factors found about employees in Russia to predict if the employee left the company or stayed

The outcomes of the secondary research question, which involves the specific impact of different personality measures on employee turnover, are:

- Primary: the association of a single or combination of big five personality trait measures for employees in Russia with employee turnover
- Secondary: the test accuracy of a single or combination of big five personality trait measures for employees in Russia with employee turnover

N, Mean Standard Deviation, Minimum and Maximum will summarize continuous variables. Number and percent will summarize categorical variables.

The model will be built using bidirectional automatic variable selection using the stepAIC() function of the MASS function. Predictors of a $p<0.05$ will be included in the final multivariable model.

The assumptions for logistic regression are no multicollinearity, linear relationship to log odds, and independence between explanatory variables. All assumptions for regression models will be assessed by viewing correlation plots, VIF, and residual plots. Internal validation will be checked through Cronbach's alpha.

We will use the Pearson *r* correlation coefficient to evaluate the degree of relationship between variables. Goodness of fit will be assessed through deviance residuals and AIC.

10.1 Primary Question

We will conduct a logistic regression with the variable event, which measures employee turnover, as the dependent variable. The variables that will be considered are listed in section 9 and include the following: stag, gender, age, industry, profession, coach, head_gender, grey_wage, way, extraversion, independ, selfcontrol, anxiety, and novator. A forward automatic variable selection process will be used and AIC and p values will be used to assess inclusion in the multivariable model. The goodness of fit of the model will be evaluated through deviance residuals and AIC.

10.2 Secondary Questions

We will conduct a logistic regression with the variable event, which measures employee turnover, as the dependent variable. The variables that will be considered are specifically the big five personality traits, listed as extraversion, independ, selfcontrol, anxiety, and novator in the dataset. A forward automatic variable selection process will be used and AIC and p values will be used to assess inclusion in the multivariable model. The goodness of fit of the model will be evaluated through deviance residuals and AIC.

10.3 Recommendations

This data set comes from a foreign population and used unknown sampling produced for collecting. We caution that because of these shortcomings a second survey of American workers would be necessary to verify the results of this survey before using the results of this analysis for any practical decision making, considering cultural differences between America and Russia.

11 Reporting Conventions

All reported p-values, correlation scores, and other statistical results will be rounded to the nearest 3rd decimal point.

95% confidence intervals will be reported as a row of three numbers. The first being the lower bound 0.05 percentile, the second being the estimate, and the third being the upper bound 0.95 percentile.

12 Summary of Changes to the SAP

4/25/2024 - Alice, Josh G., Josh Z. reformatted SAP outline using template and example SAP

4/27/2024 - Josh G further edited sections, removed template text, added table for summary of study data.

4/27/2024 – Alice X. edited Introduction sections: Background, Study Objectives, Research Questions, and Scope. edited analysis sections: analysis, primary question, secondary question.

4/28/2024 - Josh Z - completed data pre-processing section, edited study methods and statistical principles, derived populations.

4/29/2024 - Alice X. edited Analysis portion. Josh Z edited General Analysis Considerations, Statistical Principles, Data Preprocessing, Administrative Information, and References.

13 References

(1) Holiday M. (2021, January 13). What is Employee Turnover & Why It Matters for Your Business. Oracle Netsuite. https://www.netsuite.com/portal/resource/articles/human-resources/employee-turnover.shtml#:~:text=Employee%20turnover%20reference%20to%20the,%E2%80%94that%20is%2C%20involuntary%20turnover

(2) Indeed Editorial Team. (2023, February 3). Turnover vs. Attrition: Definitions, Differences and Tips. Indeed.com. https://www.indeed.com/career-advice/career-development/turnover-vs-attrition

(3) UCI-Machine Learning Repository. (n.d.). Turnover data set [Data set]. https://www.aihr.com/wp-content/uploads/2019/10/turnover-data-set.csv

(4) Boyle, G.J., Stankov, L., Cattell, R.B. (1995). Measurement and Statistical Models in the Study of Personality and Intelligence. In: Saklofske, D.H., Zeidner, M. (eds) International Handbook of Personality and Intelligence. Perspectives on Individual Differences. Springer, Boston, MA. https://doi.org/10.1007/978-1-4757-5571-8_20

(5) White, I. R., & Thompson, S. G. (2005). Adjusting for partially missing baseline measurements in randomized trials. Statistics in medicine, 24(7), 993–1007. https://doi.org/10.1002/sim.1981