# Intervenable Factors Predicting Employee Turnover With a Focus On Big Five Personality Traits

Authors: Joshua Zhong, Alice Xiang, Joshua Grant

## Abstract

Companies are often interested in measures of how many employees leave a company over a certain time period, also called employee turnover. Employee turnover can lead to increased costs for a company in order to refill positions, and may also lead to losses in productivity. Employee turnover is affected by a number of different factors, including wages, industry, benefit packages, and the ability to predict employee turnover may allow a company to make decisions for its future. In this study, logistic regression was used in an analysis of employee data from Russia to evaluate what intervenable factors may be used to create a model to predict employee turnover, with a particular interest in measures of an employees' personality, measured through a numeric scaling for each of their Big 5 personality traits (4). Our findings found that a statistically significant model could be generated to predict based off these demographic and personality variables, but the model performed marginally better than random chance.

## Introduction

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.

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.

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. We will conduct a logistic regression with the variable event, which measures employee turnover, as the dependent variable, to achieve outcomes for the primary and secondary research questions.

# Methods

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 16 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 |

# Variable Conversion

## Discretization

The original 'Industry' variable was transformed into an aggregated categorical variable with five levels. This derived variable categorizes the industries into 'Finance', 'Industrial', 'Retail', 'Tech', and 'Other'. 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 aggregation was done to group industries with similar corporate cultures and workflows together, allowing for more meaningful analysis of the differences between Industries. No Industry was left ungrouped

Similarly, the 'Profession' variable from the CRF was converted into a derived categorical variable with six distinct levels: 'Finance/Legal',, 'Sales/Marketing', 'Human/Public Relations', 'Technical', 'Education & Management', and 'Other'. 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 was done to group roles with similar workflows and responsibilities together, allowing for meaningful analysis of the difference between roles. No Profession was left ungrouped

The variable ‘traffic’ was also converted into a binomial categorical variable. The two categories for the derived variable are “Employee Initiated” and “Employer Initiate”. The traffic variables "KA", "advert", "recNErab", "referal", and "youjs” were coded as “Employee Initiated”. The traffic variables "empjs", "friends", and "rabrecNErab" were coded as “Employer Initiated”. These two categories were created as we determined the most practical piece of information that can be derived from the traffic variable is the difference in retention rates between employees who approached the company requesting employment and employees who were initially approached by the company. The category “KA”, which are employees who were brought to an employer through a recruiting agency, was determined to be an “Employee Initiated” hiring pipeline, as the recruiting agency is the one initiating the employment request while acting as a representative for the employee; All other categories besides “KA” fit cleanly into one of the two derived categories. No employment pipeline was left ungrouped.

For all 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. These derived variables were used to create a second data set in which they overwrote the original variables they were derived from. This second data set is the primary one we will use for the remainder of the analysis, and the derived variables will simply be referred to by the names of the variables they have replaced.

## Factor Numeric Recoding

To assist with the ease and efficiency of our analysis we recoded all categorical variables as numbers. The following gives the numeric coding for each category of all the categorical variables

**Variable “gender”**
male = 1, female = 0

**Variable “greywage”**
grey = 1, white = 0

**Variable “head_gender”**
male = 1, female = 0

**Variable “way”**
by foot = 0, by bus = 1, by car = 2

**Variable “coach”**
no coach = 0, non-boss coach = 1, boss was coach = 2

**Variable “industry”**
Finance = 1, Industrial = 2, Commercial = 3, Tech = 4, Other = 5

**Variable “profession”**
Finance/Legal = 1, Sales/Marketing = 2, Human/Public Relations = 3, Technical = 4, Education & Management = 5, Other = 0

**Variable “traffic”**
Employee Initiated = 1 Employer Initiated = 0

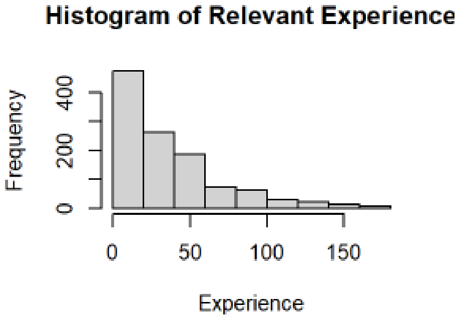
## Factor Renaming and Rescaling.

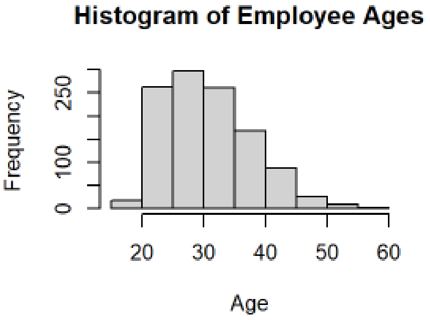
For the convenience of the analysis we renamed the big 5 trait variables to what they are most commonly referred to in the United States.
“Independ” was renamed Agreeableness
“extraversion” was renamed Extraversion
“selfcontrol” was renamed Conscientiousness
“anxiety” was renamed Neuroticism
“Novator” was renamed Openness

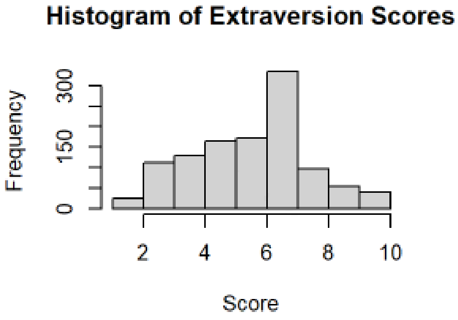
When converting “independ” to agreeableness it was also necessary to flip the scale of each employee's score, as independence and agreeableness are inverse measurements of the same trait. For example, as these traits were measured on a 10 point scale an employee who scored a 8 on independence would score a 2 on agreeableness.

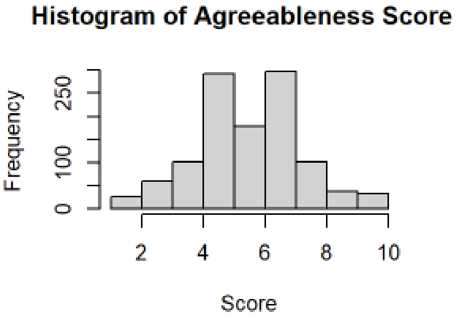
# Exploratory Data Analysis

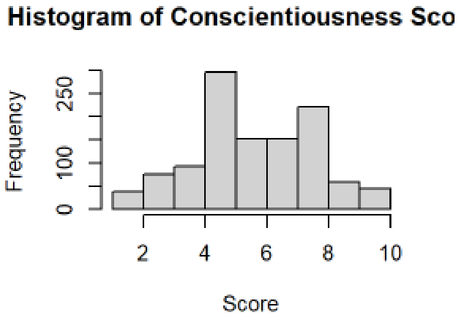
## Distributions

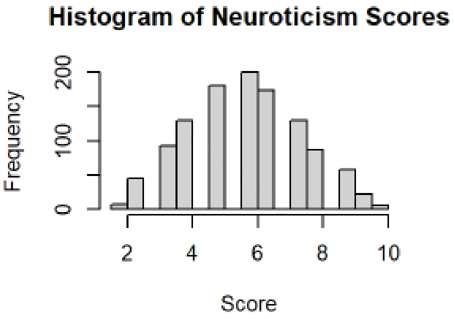


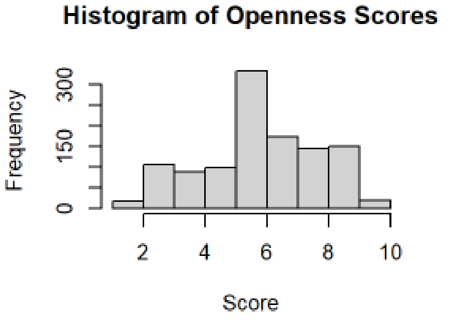






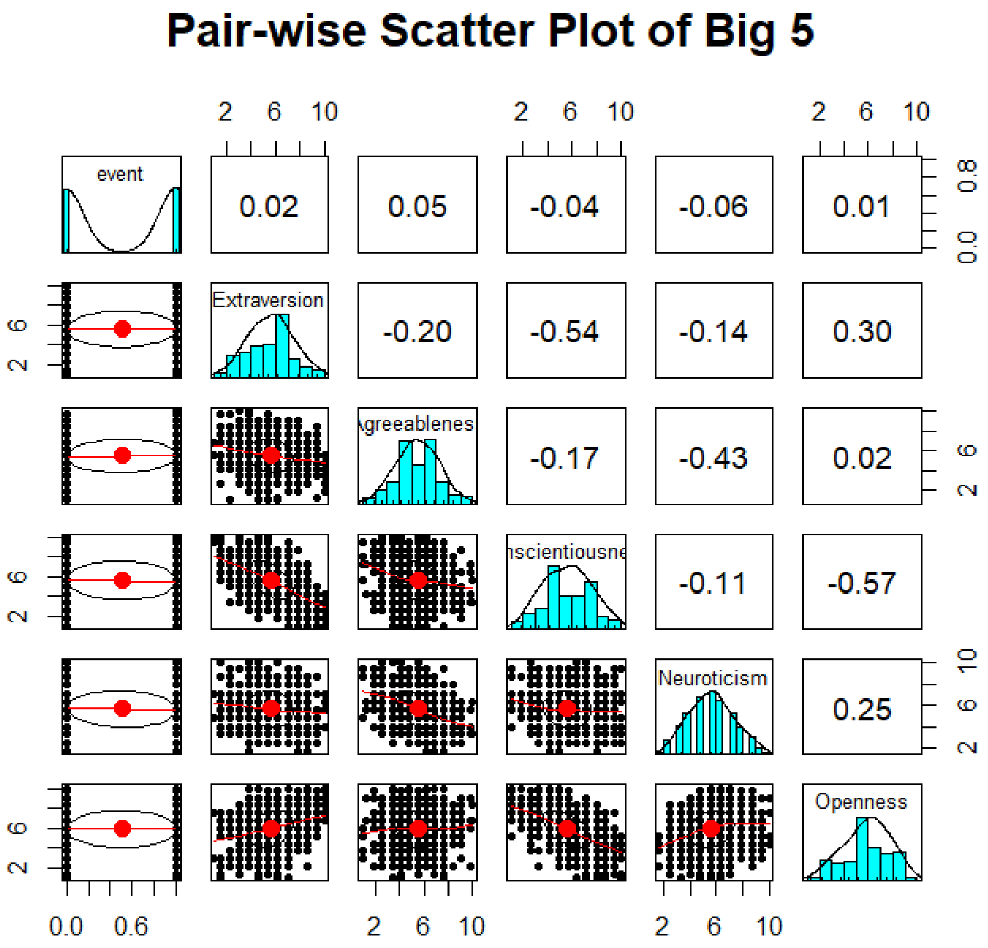




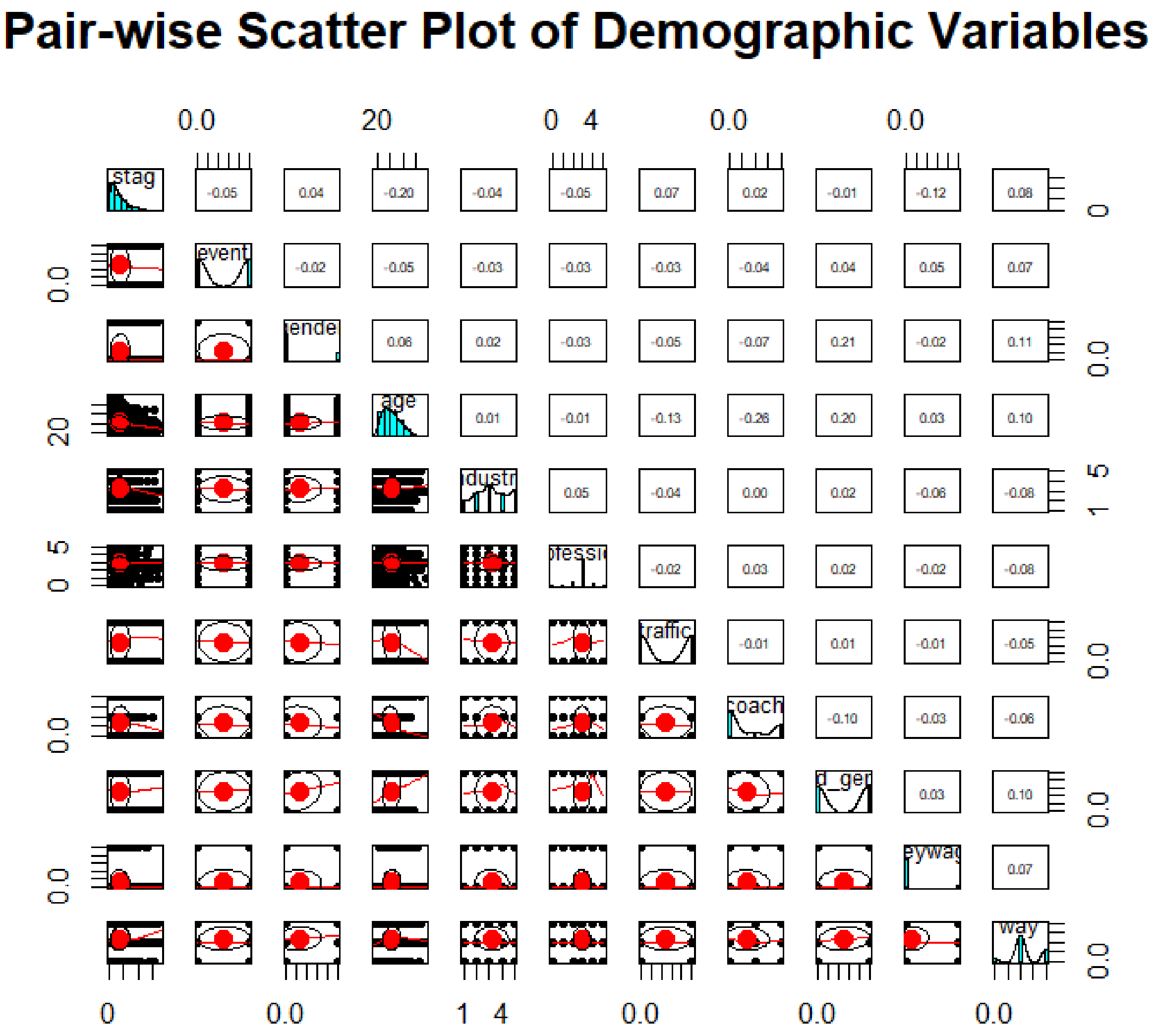


There are a number of important pieces of information revealed when we create histograms of our numeric explanatory variables. For our two demographic numeric variables, Relevant Experience and Employee Ages, there is a strong right skewness.This is entirely to be expected based on the nature of the variable. There are some surprising distributions for our “Big Five” variables. Neuroticism is normally distributed, but has some noticeable gaps in the data. This is most likely a result of how the scores were rounded, and can effectively be ignored. The other 4 “Big Five” variables all have some violations of normality.  Agreeableness and Conscientiousness are both bimodal, Extraversion has a slight right skew, and Openness has a slight left skew. This is concerning as these surveys were designed to produce a normalized response, or at the very least one without skewness. These results imply there may be a fundamental difference in big five personality scores between this sample and the one used to build the test.

## Correlation Plot



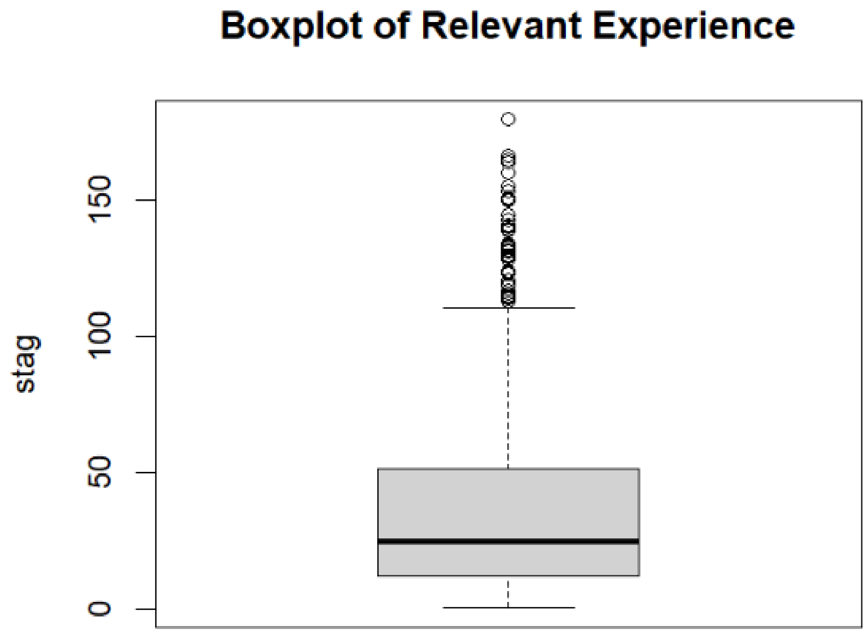
From analyzing the scatter plots and the pearson's correlations, there is a moderate correlation between a number of the Big Five traits. Extraversion and Conscientiousness, Agreeableness and Neuroticism, and Conscientiousness and Openness all have moderate negative correlations. Our response variable “event” has a near zero correlation with all the big 5 traits



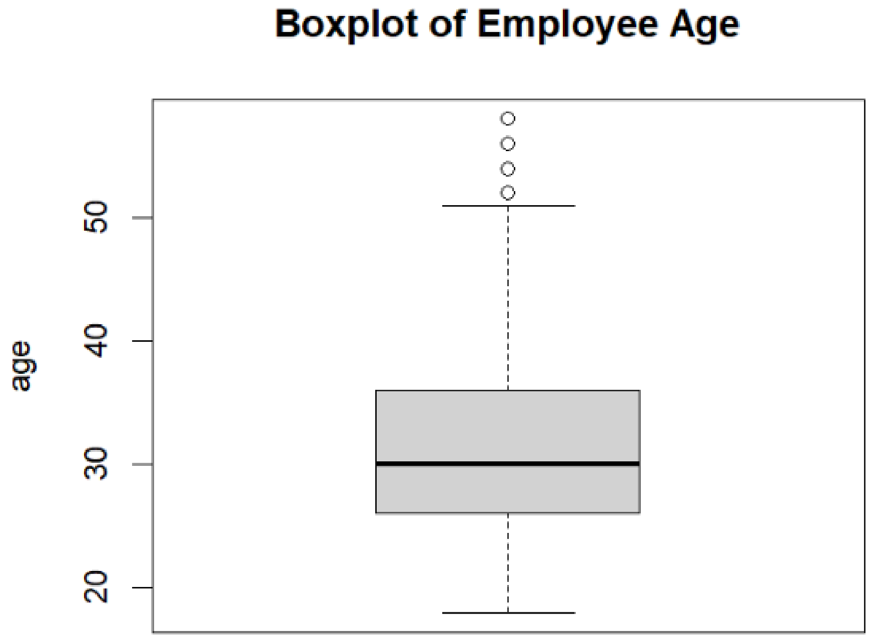
A pairwise scatter plot reveals no notable correlations between our demographic variables. Our response variable “event” is similarity uncorrelated with these variables.

## Outliers

Outliers are defined using the IQR criterion.Any observation that is either larger than the third quartile by 1.5 times the IQR, or smaller than the first quartile by 1.5 times the IQR, is considered an outlier.



The numeric demographic variable ”stag”, which is a normalized measure of the employee’s length of relevant work experience, has a large number of observations that are considered outliers. In total there are 52 outliers for the variable “stag”.

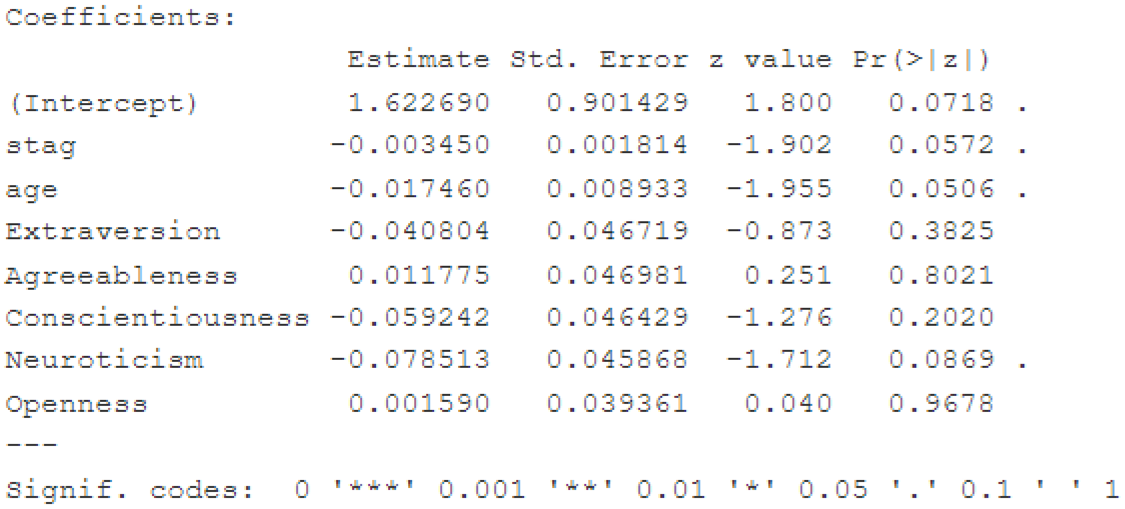


There are 7 observations of the numeric demographic variable “age” that are considered outliers.These outlier observations correspond to one observation of age 52, four of age 54, one of age 56, and one of age 58.

There were no outliers detected for any of the “Big 5” personality score variables.

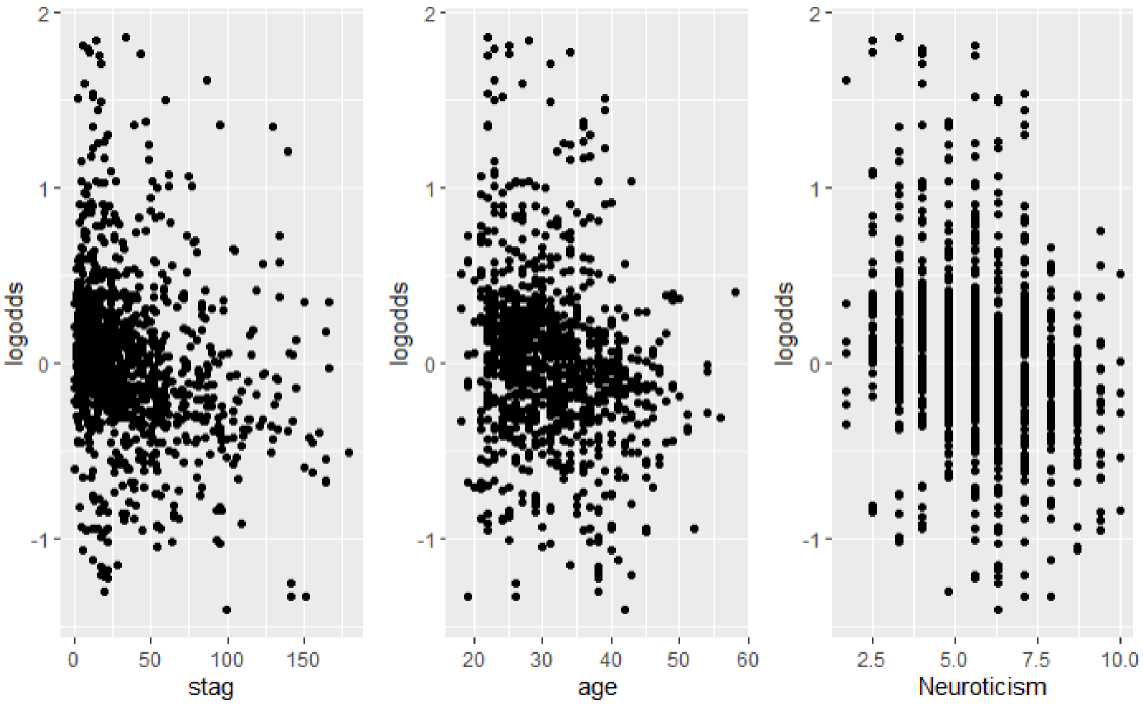
## Linearity of Log Odds:

An assumption of logistic regressions is the linearity of the log odds and the continuous independent variables. A test was run to check the linearity of the log odds against all possible continuous predictors. The results are shown in the following table.



```
Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)        1.622690   0.901429   1.800   0.0718 .
stag              -0.003450   0.001814  -1.902   0.0572 .
age               -0.017460   0.008933  -1.955   0.0506 .
Extraversion      -0.040804   0.046719  -0.873   0.3825
Agreeableness      0.011775   0.046981   0.251   0.8021
Conscientiousness -0.059242   0.046429  -1.276   0.2020
Neuroticism       -0.078513   0.045868  -1.712   0.0869 .
Openness           0.001590   0.039361   0.040   0.9678
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Three variables were marked with small, but not quite significant p-values, these being stag ($p$ = 0.0572), age ($p$ = 0.0506), and Neuroticism ($p$ = 0.0869). The log odds were then plotted against each of these predictors.



None of the above scatter plots seem to display strong linear correlations. However, none of these predictors display clear evidence of nonlinearity or curvature. It is possible that the small p-values may be in part due to the large sample size as well as the weak associations observed. Without clear evidence of curvature or transformations that must be taken to correct for blatant violations to the assumption of linearity of the log odds and considering the non significant p-values, we proceed with the logistic regression.

## Logistic Regression Models

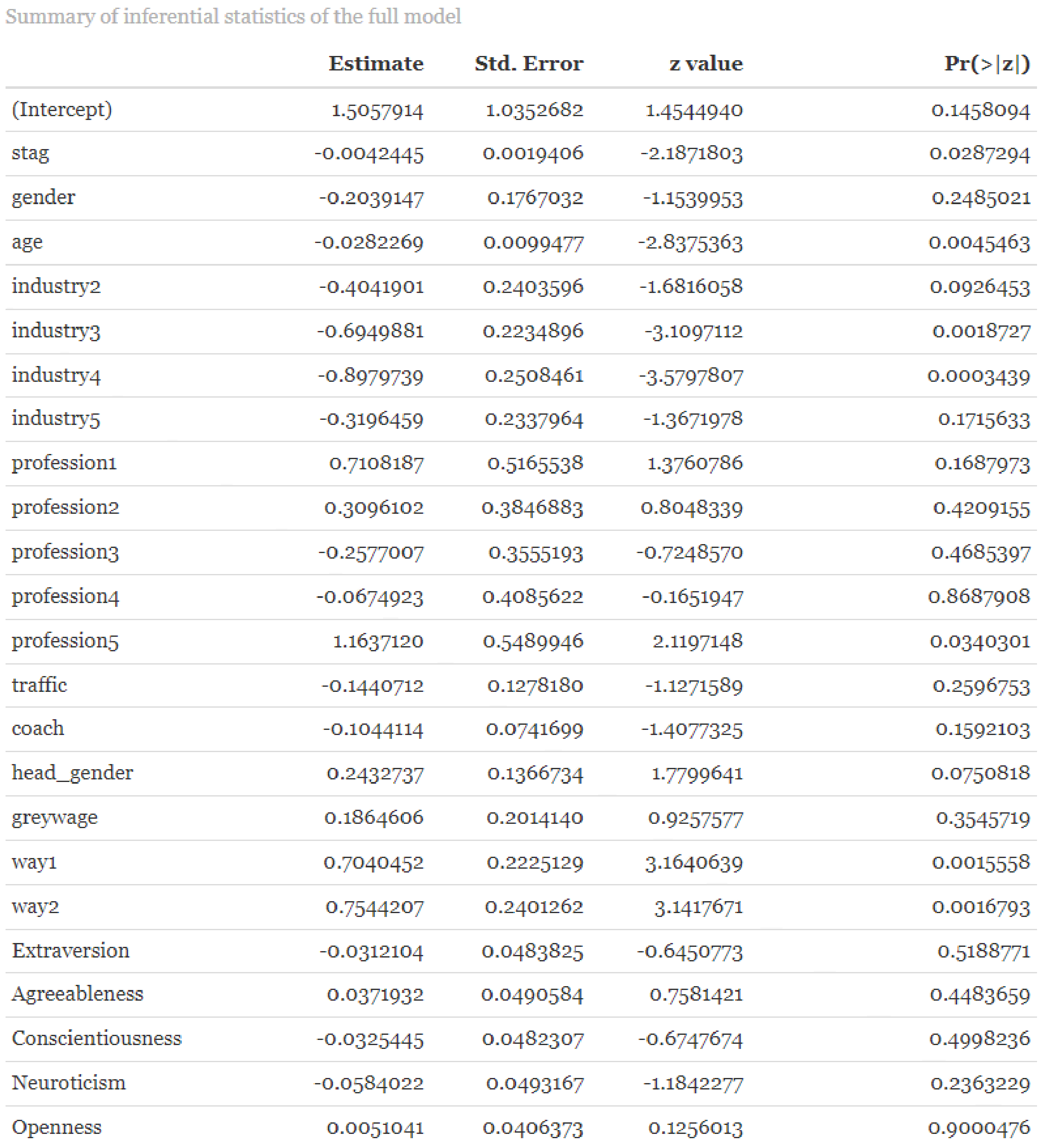
Before any logistic regression models were built, all categorical variables were converted to factor format. This was done to avoid implying an ordinality to the categories as well as to make the data compatible with the R packages we intended to use.

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.

The response variable event is binary, encoded as 1 if an employee left the company and 0 if the employee stayed. Our exploratory data analysis showed no major examples of collinearity between predictors. In the Big 5 predictors, Extraversion and Conscientiousness, Agreeableness and Neuroticism, and Conscientiousness and Openness were observed to have moderate correlations. In the demographic variables, there were no notable correlations between predictors. The linearity of log odds with the predictors was checked in the prior section, and without any clear evidence of curvature or necessary transformations for any of the predictors, this assumption seems to hold true. Since the data collection methods are unknown, it is assumed that the observations in the dataset are independent of each other. For observational studies, a sample size of at least 50*(number of predictors)+100 is recommended (5) for logistic regression; the full model for the logistic regression included 15 total predictors, so the recommended sample size was at least 50*15+100, or 850 observations. The dataset had 1130 observations, so the sample size requirement was met. Therefore, we proceeded with logistic regression.

## Logistic Regression of All Predictors:

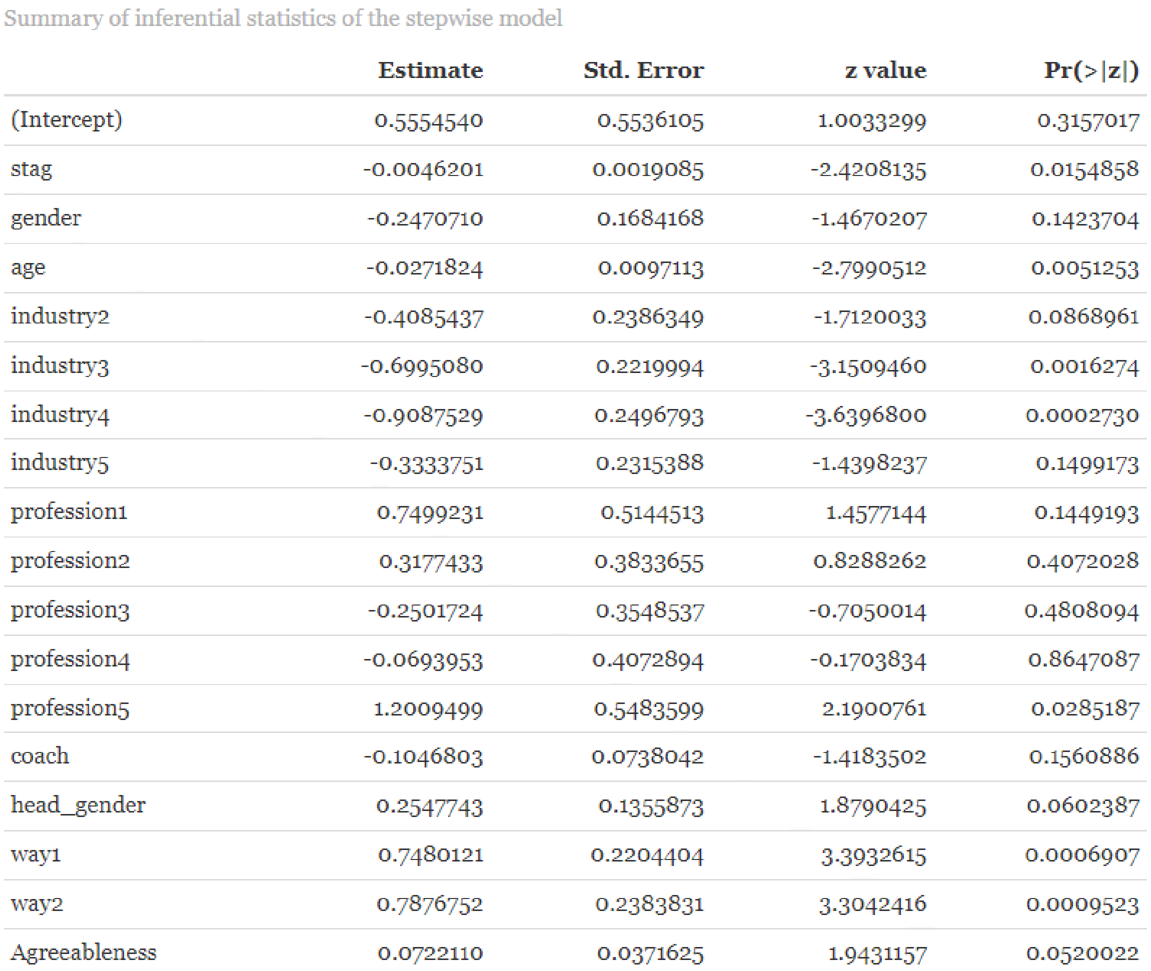
The first logistic regression model we built was designed to estimate the probability of an employee leaving their role based on all intervenable factors from the data set. In the initial model we used the event variable as the response variable and the following 15 variables as our explanatory variables:stag, gender, age, industry, profession, traffic, coach, head_gender, grey_wage, way, Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness. The summary of inferential statistics of the full model is presented below.



Summary of inferential statistics of the full model

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---:|---:|---:|---:|
| (Intercept) | 1.5057914 | 1.0352682 | 1.4544940 | 0.1458094 |
| stag | -0.0042445 | 0.0019406 | -2.1871803 | 0.0287294 |
| gender | -0.2039147 | 0.1767032 | -1.1539953 | 0.2485021 |
| age | -0.0282269 | 0.0099477 | -2.8375363 | 0.0045463 |
| industry2 | -0.4041901 | 0.2403596 | -1.6816058 | 0.0926453 |
| industry3 | -0.6949881 | 0.2234896 | -3.1097112 | 0.0018727 |
| industry4 | -0.8979739 | 0.2508461 | -3.5797807 | 0.0003439 |
| industry5 | -0.3196459 | 0.2337964 | -1.3671978 | 0.1715633 |
| profession1 | 0.7108187 | 0.5165538 | 1.3760786 | 0.1687973 |
| profession2 | 0.3096102 | 0.3846883 | 0.8048339 | 0.4209155 |
| profession3 | -0.2577007 | 0.3555193 | -0.7248570 | 0.4685397 |
| profession4 | -0.0674923 | 0.4085622 | -0.1651947 | 0.8687908 |
| profession5 | 1.1637120 | 0.5489946 | 2.1197148 | 0.0340301 |
| traffic | -0.1440712 | 0.1278180 | -1.1271589 | 0.2596753 |
| coach | -0.1044114 | 0.0741699 | -1.4077325 | 0.1592103 |
| head_gender | 0.2432737 | 0.1366734 | 1.7799641 | 0.0750818 |
| greywage | 0.1864606 | 0.2014140 | 0.9257577 | 0.3545719 |
| way1 | 0.7040452 | 0.2225129 | 3.1640639 | 0.0015558 |
| way2 | 0.7544207 | 0.2401262 | 3.1417671 | 0.0016793 |
| Extraversion | -0.0312104 | 0.0483825 | -0.6450773 | 0.5188771 |
| Agreeableness | 0.0371932 | 0.0490584 | 0.7581421 | 0.4483659 |
| Conscientiousness | -0.0325445 | 0.0482307 | -0.6747674 | 0.4998236 |
| Neuroticism | -0.0584022 | 0.0493167 | -1.1842277 | 0.2363229 |
| Openness | 0.0051041 | 0.0406373 | 0.1256013 | 0.9000476 |

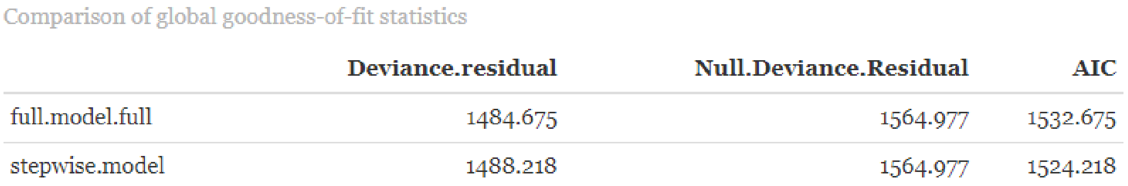
A backward stepwise AIC based variable selection algorithm was used to drop non-contributing variables from the model. The algorithm removed the variables of traffic, greywage, Extraversion, Conscientiousness, Neuroticism, and Openness. The variables stag, age, industry, profession, coach, head_gender, way, and agreeableness were retained. The inferential statistics from the final model created through this method is shown below.



Summary of inferential statistics of the stepwise model

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | 0.5554540 | 0.5536105 | 1.0033299 | 0.3157017 |
| stag | -0.0046201 | 0.0019085 | -2.4208135 | 0.0154858 |
| gender | -0.2470710 | 0.1684168 | -1.4670207 | 0.1423704 |
| age | -0.0271824 | 0.0097113 | -2.7990512 | 0.0051253 |
| industry2 | -0.4085437 | 0.2386349 | -1.7120033 | 0.0868961 |
| industry3 | -0.6995080 | 0.2219994 | -3.1509460 | 0.0016274 |
| industry4 | -0.9087529 | 0.2496793 | -3.6396800 | 0.0002730 |
| industry5 | -0.3333751 | 0.2315388 | -1.4398237 | 0.1499173 |
| profession1 | 0.7499231 | 0.5144513 | 1.4577144 | 0.1449193 |
| profession2 | 0.3177433 | 0.3833655 | 0.8288262 | 0.4072028 |
| profession3 | -0.2501724 | 0.3548537 | -0.7050014 | 0.4808094 |
| profession4 | -0.0693953 | 0.4072894 | -0.1703834 | 0.8647087 |
| profession5 | 1.2009499 | 0.5483599 | 2.1900761 | 0.0285187 |
| coach | -0.1046803 | 0.0738042 | -1.4183502 | 0.1560886 |
| head_gender | 0.2547743 | 0.1355873 | 1.8790425 | 0.0602387 |
| way1 | 0.7480121 | 0.2204404 | 3.3932615 | 0.0006907 |
| way2 | 0.7876752 | 0.2383831 | 3.3042416 | 0.0009523 |
| Agreeableness | 0.0722110 | 0.0371625 | 1.9431157 | 0.0520022 |

The model created through the stepwise AIC algorithm and the full model were compared through three different goodness of fit measures: The null deviance residual, the deviance residual, and the AIC. These are displayed in the following table.

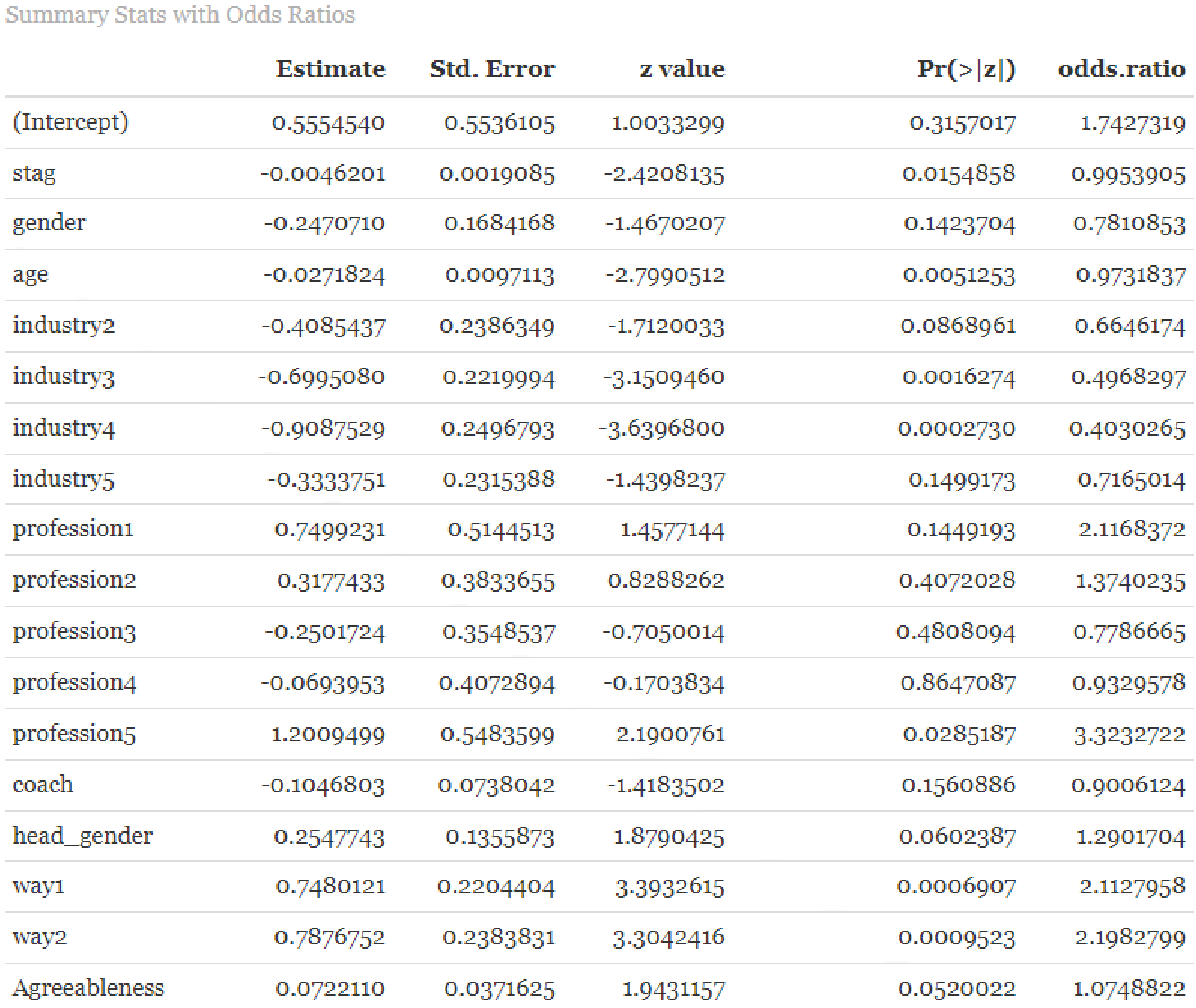


Comparison of global goodness-of-fit statistics

| | Deviance.residual | Null.Deviance.Residual | AIC |
|---|---|---|---|
| full.model.full | 1484.675 | 1564.977 | 1532.675 |
| stepwise.model | 1488.218 | 1564.977 | 1524.218 |

The null deviance residual is a measure of how well the response can be predicted with just the intercept. The deviance residual gives a measure of how well a response can be predicted with a model with *p* predictors. A lower value indicates a model that can better predict the response. A chi-squared test may be used to assess the quality of a model using the null deviance and deviance residual, where the test statistic is the difference between the null and deviance residuals and a degrees of freedom of *p,* the number of predictors in the model (6). For the full model, this gives a *p* value of χ2(80.302, 18) < 0.000001. For the stepwise model, this would be a *p* value of χ2(76.759, 17) < 0.000001. Both perform significantly better than just the intercept, with extremely low p-values from the chi-squared test. Combined with the lower AIC score, we chose the stepwise model as the better performing model..

The summary statistics of the final model are shown below with the calculated odd ratios for each predictor.

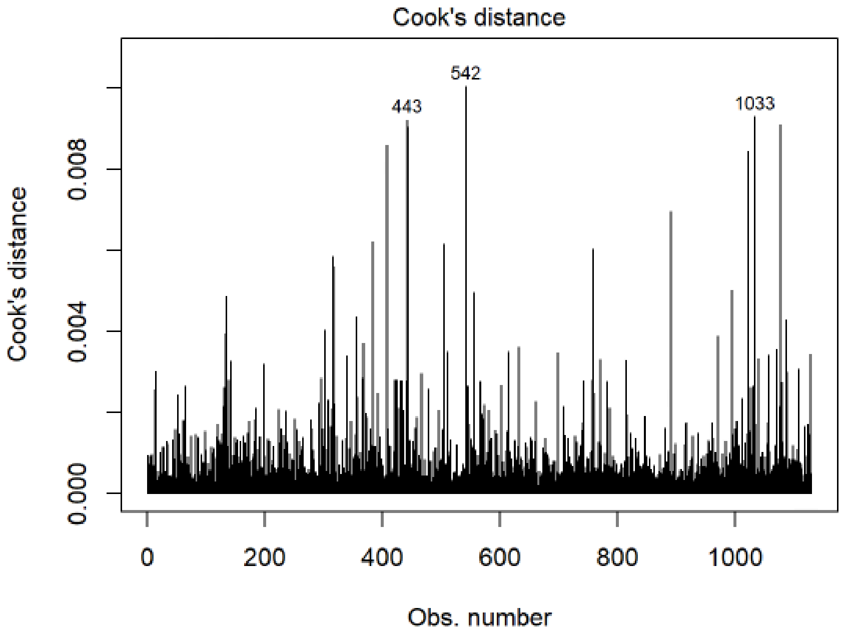


Summary Stats with Odds Ratios

| | Estimate | Std. Error | z value | Pr(>\|z\|) | odds.ratio |
|---|---|---|---|---|---|
| (Intercept) | 0.5554540 | 0.5536105 | 1.0033299 | 0.3157017 | 1.7427319 |
| stag | -0.0046201 | 0.0019085 | -2.4208135 | 0.0154858 | 0.9953905 |
| gender | -0.2470710 | 0.1684168 | -1.4670207 | 0.1423704 | 0.7810853 |
| age | -0.0271824 | 0.0097113 | -2.7990512 | 0.0051253 | 0.9731837 |
| industry2 | -0.4085437 | 0.2386349 | -1.7120033 | 0.0868961 | 0.6646174 |
| industry3 | -0.6995080 | 0.2219994 | -3.1509460 | 0.0016274 | 0.4968297 |
| industry4 | -0.9087529 | 0.2496793 | -3.6396800 | 0.0002730 | 0.4030265 |
| industry5 | -0.3333751 | 0.2315388 | -1.4398237 | 0.1499173 | 0.7165014 |
| profession1 | 0.7499231 | 0.5144513 | 1.4577144 | 0.1449193 | 2.1168372 |
| profession2 | 0.3177433 | 0.3833655 | 0.8288262 | 0.4072028 | 1.3740235 |
| profession3 | -0.2501724 | 0.3548537 | -0.7050014 | 0.4808094 | 0.7786665 |
| profession4 | -0.0693953 | 0.4072894 | -0.1703834 | 0.8647087 | 0.9329578 |
| profession5 | 1.2009499 | 0.5483599 | 2.1900761 | 0.0285187 | 3.3232722 |
| coach | -0.1046803 | 0.0738042 | -1.4183502 | 0.1560886 | 0.9006124 |
| head_gender | 0.2547743 | 0.1355873 | 1.8790425 | 0.0602387 | 1.2901704 |
| way1 | 0.7480121 | 0.2204404 | 3.3932615 | 0.0006907 | 2.1127958 |
| way2 | 0.7876752 | 0.2383831 | 3.3042416 | 0.0009523 | 2.1982799 |
| Agreeableness | 0.0722110 | 0.0371625 | 1.9431157 | 0.0520022 | 1.0748822 |

The model's diagnostics were examined to check for violations to assumptions through the presence of outliers or multicollinearity once again. The following Cook's D plot for the residuals marks the three observations with the highest measures of Cook's Distance.

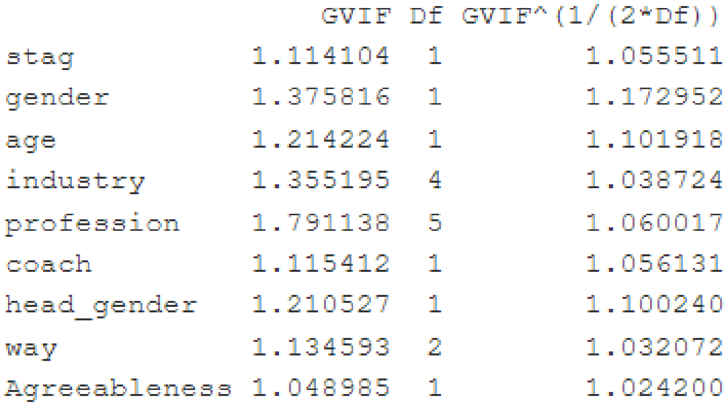
Cook’s Distance for residuals of the full model:



As there does not appear to be any one or few observations that have above and beyond large values for Cook’s distance and the overall values for Cook’s distance are quite small (<1), the data does not appear to have any extreme outliers that may need to be dealt with separately. Similarly, a look at the standardized residuals yields none with a value greater than two..

Multicollinearity was checked through values of VIF. As there are none over two, multicollinearity does not appear to be an issue for our final model.

VIF values for predictors in the full model:



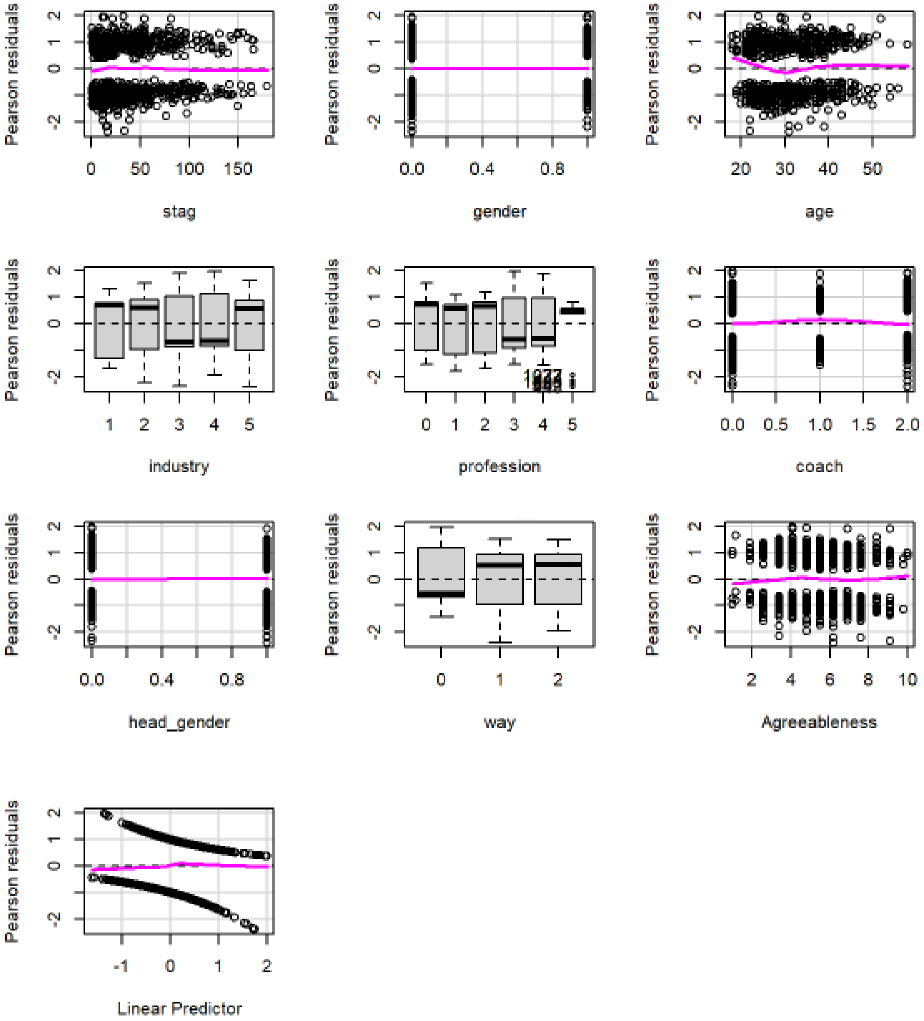
```
                  GVIF Df GVIF^(1/(2*Df))
stag          1.114104  1        1.055511
gender        1.375816  1        1.172952
age           1.214224  1        1.101918
industry      1.355195  4        1.038724
profession    1.791138  5        1.060017
coach         1.115412  1        1.056131
head_gender   1.210527  1        1.100240
way           1.134593  2        1.032072
Agreeableness 1.048985  1        1.024200
```

Finally, the Pearson's residual plots for the model were created and are shown below. Though deviances in linearity can still be observed in the predictor age, the deviance is slight enough for us to conclude that the predictors are properly specified for our demographic final model.

Lack of fit test results for final demographic regression model:

```
              Test stat Pr(>|Test stat|)
stag             0.3899          0.53236
gender           0.0000          1.00000
age              5.4136          0.01998 *
industry
profession
coach            2.6909          0.10092
head_gender      0.0000          1.00000
way
Agreeableness    0.0420          0.83757
```

Residual plots:



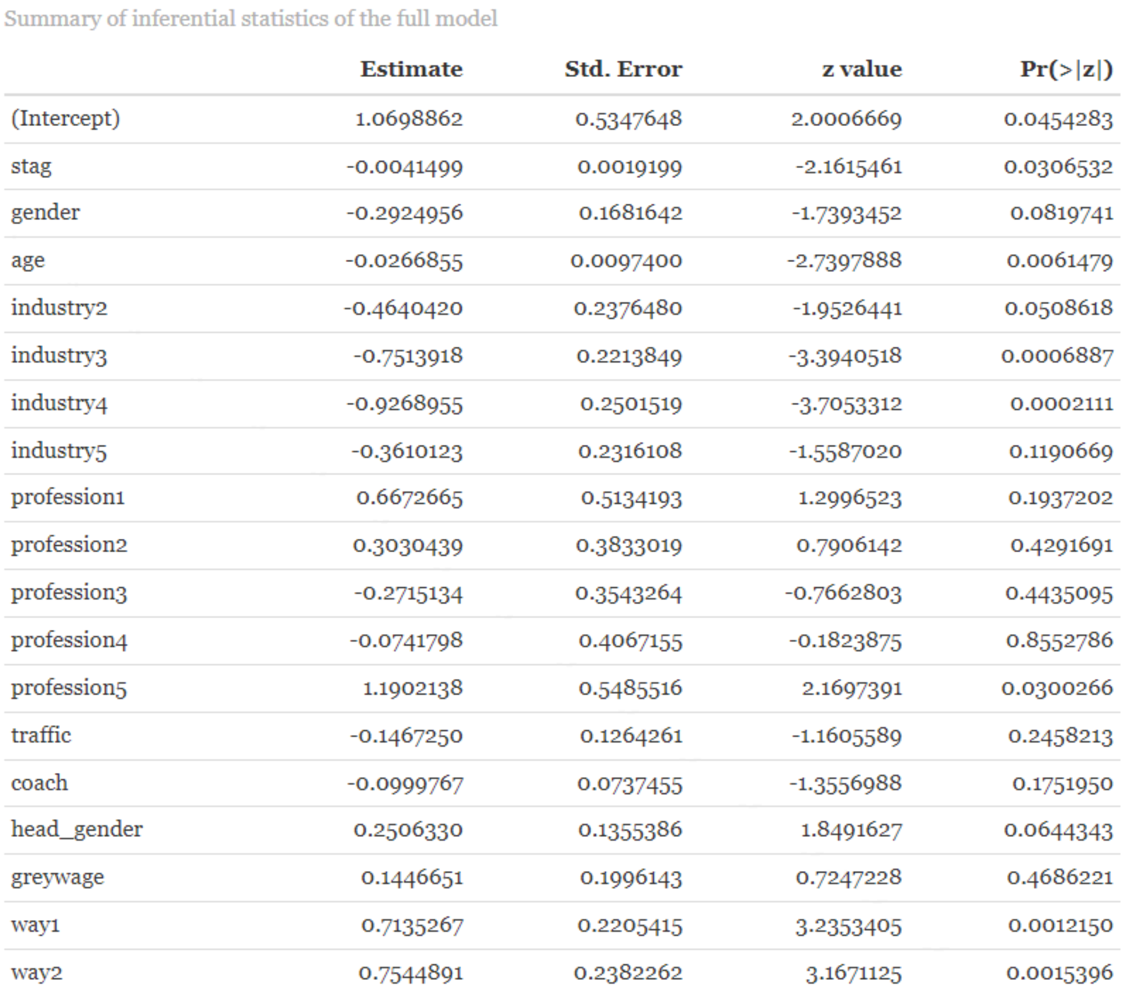
It is interesting to see that the vast majority of significant predictors selected by our stepwise model selection algorithm were from the demographic data, with only one of the big 5 traits selected. Furthermore, the *p* value of Agreeableness in our final model was found to be nonsignificant. We chose to do further analysis by breaking the data into subgroups: demographic (10 predictors: stag, gender, age, industry, profession, traffic, coach, head_gender, grey_wage, and way) and big 5 personality (5 predictors: Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness.)

## Subgroup Analysis (Demographic and Big 5 Personality):

A particular interest in this dataset was on the nature of big 5 personality and its influence on employee turnover. Therefore, to further examine this question, subgroup analysis was conducted to build two models on top of the full model: one using only the demographic predictors, and one using only the measures for the big 5 personality traits.

### Logistic Regression of the Demographic Predictors:

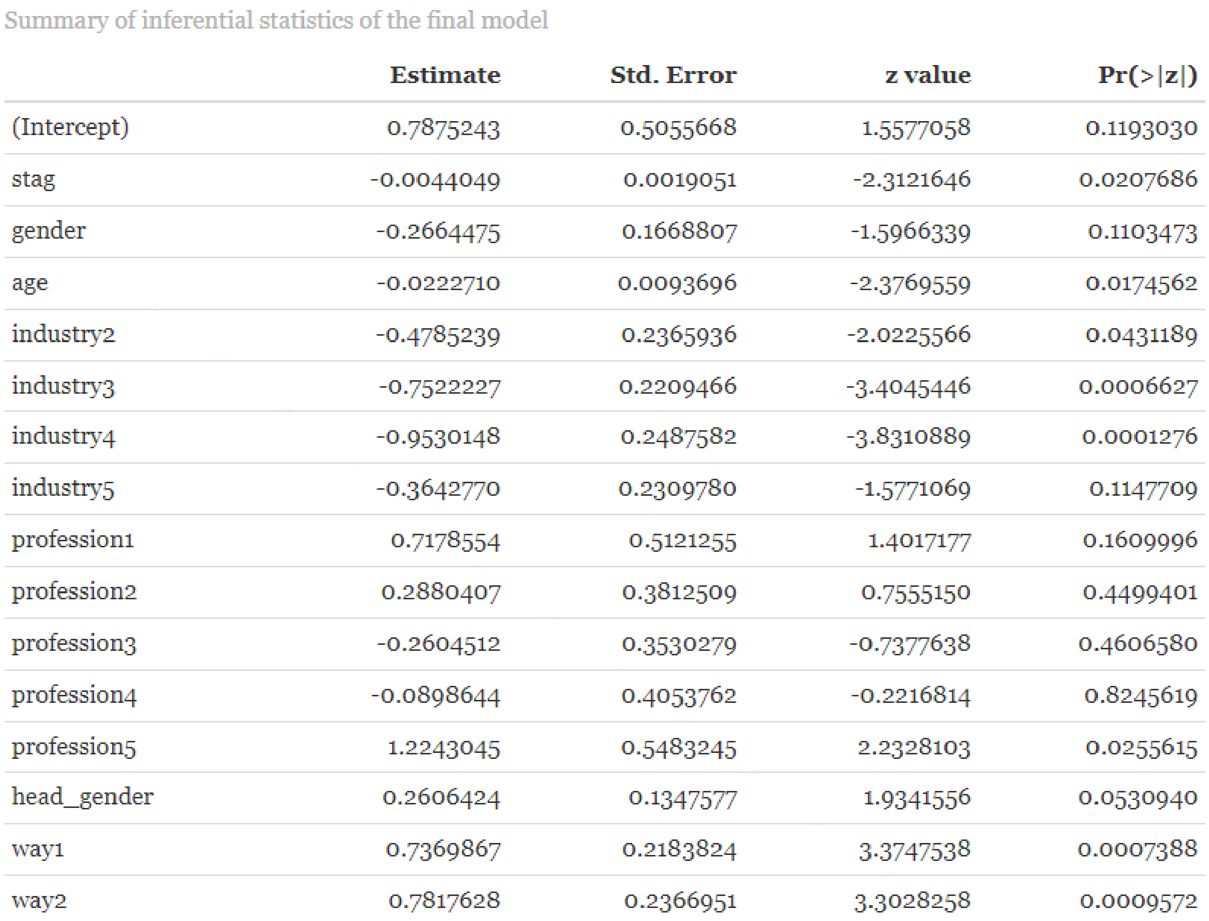
The first subgroup logistic regression model we built was designed to estimate the probability of an employee leaving their role based on the demographic variables from the data set. In the initial model we used the event variable as the response variable and the following 10 variables as our explanatory variables:stag, gender, age, industry, profession, traffic, coach, head_gender, grey_wage, and way. The summary of inferential statistics of the full model is presented below.



Summary of inferential statistics of the full model

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | 1.0698862 | 0.5347648 | 2.0006669 | 0.0454283 |
| stag | -0.0041499 | 0.0019199 | -2.1615461 | 0.0306532 |
| gender | -0.2924956 | 0.1681642 | -1.7393452 | 0.0819741 |
| age | -0.0266855 | 0.0097400 | -2.7397888 | 0.0061479 |
| industry2 | -0.4640420 | 0.2376480 | -1.9526441 | 0.0508618 |
| industry3 | -0.7513918 | 0.2213849 | -3.3940518 | 0.0006887 |
| industry4 | -0.9268955 | 0.2501519 | -3.7053312 | 0.0002111 |
| industry5 | -0.3610123 | 0.2316108 | -1.5587020 | 0.1190669 |
| profession1 | 0.6672665 | 0.5134193 | 1.2996523 | 0.1937202 |
| profession2 | 0.3030439 | 0.3833019 | 0.7906142 | 0.4291691 |
| profession3 | -0.2715134 | 0.3543264 | -0.7662803 | 0.4435095 |
| profession4 | -0.0741798 | 0.4067155 | -0.1823875 | 0.8552786 |
| profession5 | 1.1902138 | 0.5485516 | 2.1697391 | 0.0300266 |
| traffic | -0.1467250 | 0.1264261 | -1.1605589 | 0.2458213 |
| coach | -0.0999767 | 0.0737455 | -1.3556988 | 0.1751950 |
| head_gender | 0.2506330 | 0.1355386 | 1.8491627 | 0.0644343 |
| greywage | 0.1446651 | 0.1996143 | 0.7247228 | 0.4686221 |
| way1 | 0.7135267 | 0.2205415 | 3.2353405 | 0.0012150 |
| way2 | 0.7544891 | 0.2382262 | 3.1671125 | 0.0015396 |

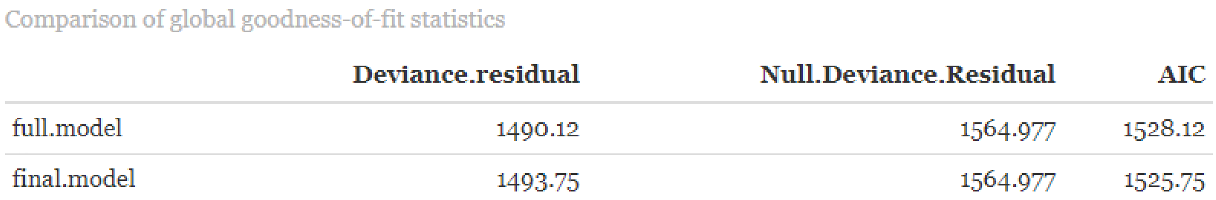
A backward stepwise AIC based variable selection algorithm was used to drop non-contributing variables from the model. The algorithm removed the variables of gender, industry, profession, traffic, and greywage. The variables stag, age, coach, head_gender, and way were retained. The inferential statistics from the final model created through this method is shown below.



Summary of inferential statistics of the final model

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---:|---:|---:|---:|
| (Intercept) | 0.7875243 | 0.5055668 | 1.5577058 | 0.1193030 |
| stag | -0.0044049 | 0.0019051 | -2.3121646 | 0.0207686 |
| gender | -0.2664475 | 0.1668807 | -1.5966339 | 0.1103473 |
| age | -0.0222710 | 0.0093696 | -2.3769559 | 0.0174562 |
| industry2 | -0.4785239 | 0.2365936 | -2.0225566 | 0.0431189 |
| industry3 | -0.7522227 | 0.2209466 | -3.4045446 | 0.0006627 |
| industry4 | -0.9530148 | 0.2487582 | -3.8310889 | 0.0001276 |
| industry5 | -0.3642770 | 0.2309780 | -1.5771069 | 0.1147709 |
| profession1 | 0.7178554 | 0.5121255 | 1.4017177 | 0.1609996 |
| profession2 | 0.2880407 | 0.3812509 | 0.7555150 | 0.4499401 |
| profession3 | -0.2604512 | 0.3530279 | -0.7377638 | 0.4606580 |
| profession4 | -0.0898644 | 0.4053762 | -0.2216814 | 0.8245619 |
| profession5 | 1.2243045 | 0.5483245 | 2.2328103 | 0.0255615 |
| head_gender | 0.2606424 | 0.1347577 | 1.9341556 | 0.0530940 |
| way1 | 0.7369867 | 0.2183824 | 3.3747538 | 0.0007388 |
| way2 | 0.7817628 | 0.2366951 | 3.3028258 | 0.0009572 |

The model created through the stepwise AIC algorithm and the full model were compared through three different goodness of fit measures: The null deviance residual, the deviance residual, and the AIC. These are displayed in the following table.

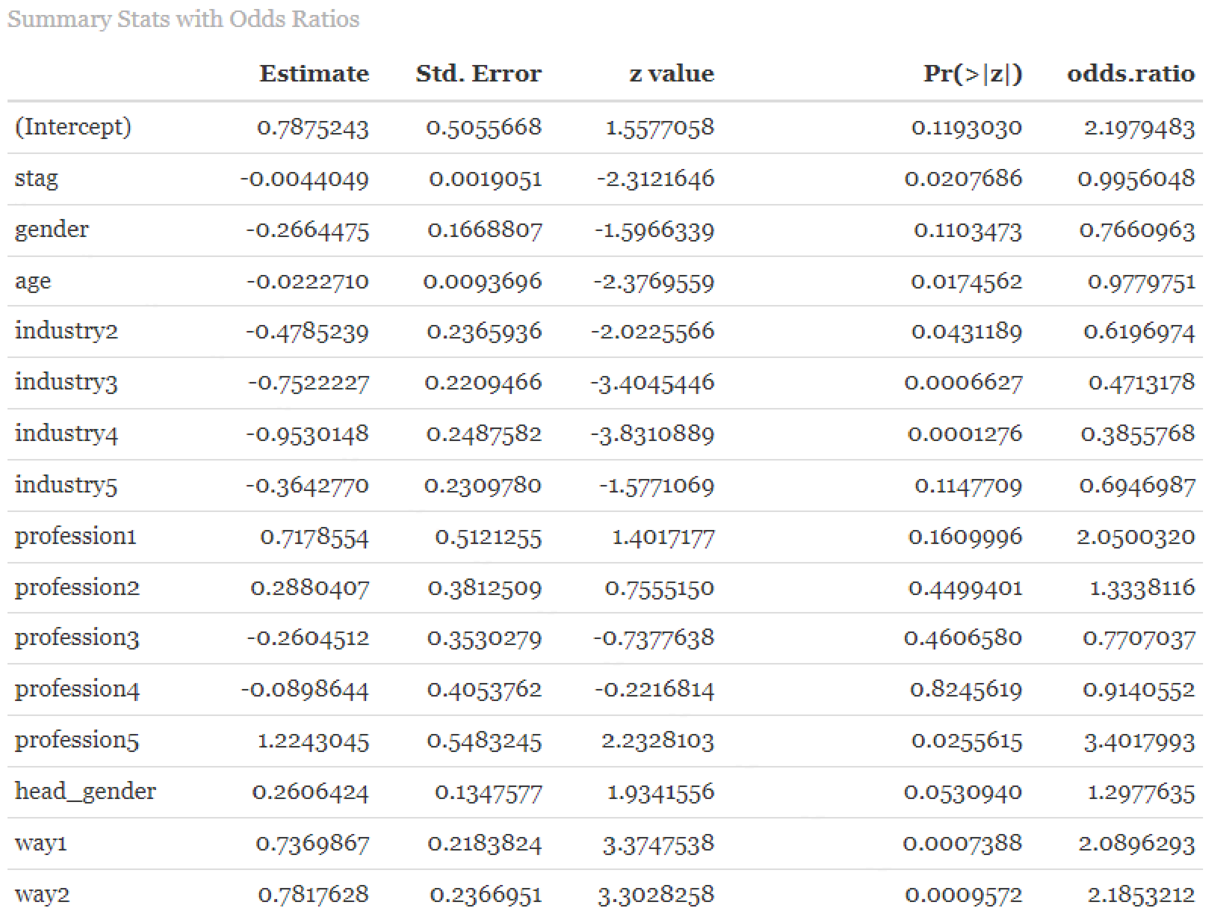


Comparison of global goodness-of-fit statistics

| | Deviance.residual | Null.Deviance.Residual | AIC |
|---|---:|---:|---:|
| full.model | 1490.12 | 1564.977 | 1528.12 |
| final.model | 1493.75 | 1564.977 | 1525.75 |

The null deviance residual is a measure of how well the response can be predicted with just the intercept. The deviance residual gives a measure of how well a response can be predicted with a model with *p* predictors. A lower value indicates a model that can better predict the response. A chi-squared test may be used to assess the quality of a model using the null deviance and deviance residual, where the test statistic is the difference between the null and deviance residuals and a degrees of freedom of *p,* the number of predictors in the model (6). For the full model, this gives a *p* value of χ2(74.857, 18) < 0.000001. For the stepwise model, this would be a *p* value of χ2(71.227, 15) < 0.000001. Both perform significantly better than just the intercept, with extremely low p-values from the chi-squared test. Combined with the lower AIC score, we chose the stepwise model as the better performing model..

The summary statistics of the final model are shown below with the calculated odd ratios for each predictor.



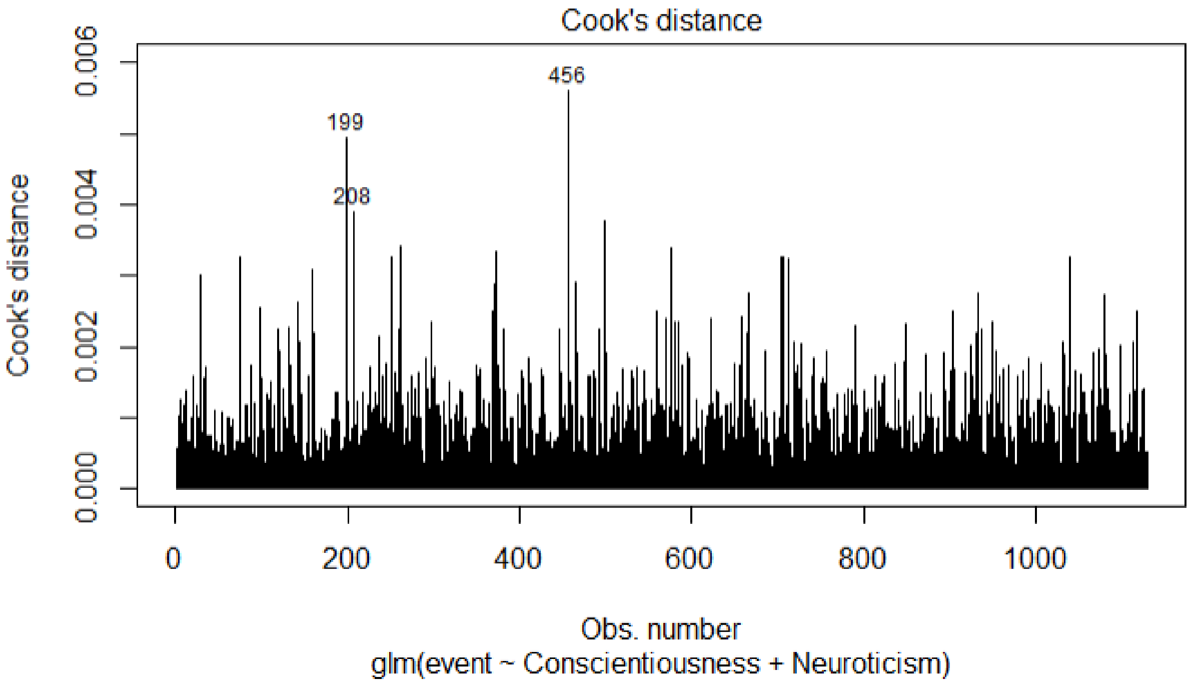
Summary Stats with Odds Ratios

| | Estimate | Std. Error | z value | Pr(>\|z\|) | odds.ratio |
|---|---|---|---|---|---|
| (Intercept) | 0.7875243 | 0.5055668 | 1.5577058 | 0.1193030 | 2.1979483 |
| stag | -0.0044049 | 0.0019051 | -2.3121646 | 0.0207686 | 0.9956048 |
| gender | -0.2664475 | 0.1668807 | -1.5966339 | 0.1103473 | 0.7660963 |
| age | -0.0222710 | 0.0093696 | -2.3769559 | 0.0174562 | 0.9779751 |
| industry2 | -0.4785239 | 0.2365936 | -2.0225566 | 0.0431189 | 0.6196974 |
| industry3 | -0.7522227 | 0.2209466 | -3.4045446 | 0.0006627 | 0.4713178 |
| industry4 | -0.9530148 | 0.2487582 | -3.8310889 | 0.0001276 | 0.3855768 |
| industry5 | -0.3642770 | 0.2309780 | -1.5771069 | 0.1147709 | 0.6946987 |
| profession1 | 0.7178554 | 0.5121255 | 1.4017177 | 0.1609996 | 2.0500320 |
| profession2 | 0.2880407 | 0.3812509 | 0.7555150 | 0.4499401 | 1.3338116 |
| profession3 | -0.2604512 | 0.3530279 | -0.7377638 | 0.4606580 | 0.7707037 |
| profession4 | -0.0898644 | 0.4053762 | -0.2216814 | 0.8245619 | 0.9140552 |
| profession5 | 1.2243045 | 0.5483245 | 2.2328103 | 0.0255615 | 3.4017993 |
| head_gender | 0.2606424 | 0.1347577 | 1.9341556 | 0.0530940 | 1.2977635 |
| way1 | 0.7369867 | 0.2183824 | 3.3747538 | 0.0007388 | 2.0896293 |
| way2 | 0.7817628 | 0.2366951 | 3.3028258 | 0.0009572 | 2.1853212 |

The odds ratio provided in the previous table, tells you weather the odds of the outcome increase (if odds ratio > 1) or decrease (if odds ratio <1 ) with a one-unit increase in the predictor variable assuming all other variables are held constant. In context, the outcome in question is whether the employee stayed at the company or left. So, for example, gender had an odds ratio fo 0.766. Since “1” corresponds to males and “0” to females, this suggests that the odds of a male employee leaving the company are about 76.6% of the odds of a female employee leaving the company, assuming all other factors are constant.

The model’s diagnostics were examined to check for violations to assumptions through the presence of outliers or multicollinearity once again. The following Cook’s D plot for the residuals marks the three observations with the highest measures of Cook’s Distance.

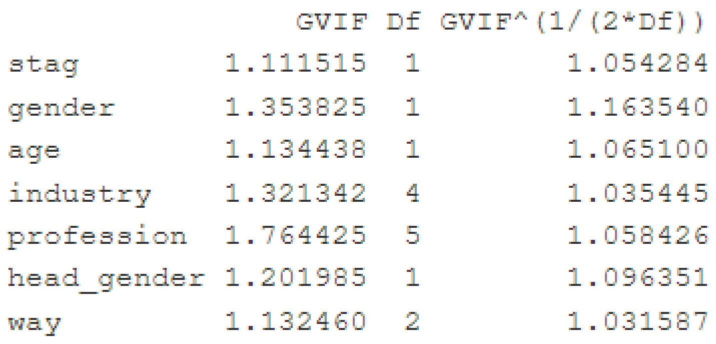
Cook’s distance for residuals of final demographic model:



As there does not appear to be any one or few observations that have above and beyond large values for Cook’s distance and the overall values for Cook’s distance are quite small (<1), the data does not appear to have any extreme outliers that may need to be dealt with separately. Similarly, a look at the standardized residuals yields that only 2 have a value of greater than 2, both of which having a standardized residual value of only about -2.01.

Multicollinearity was checked through values of VIF. As there are none over two, multicollinearity does not appear to be an issue for our final model.

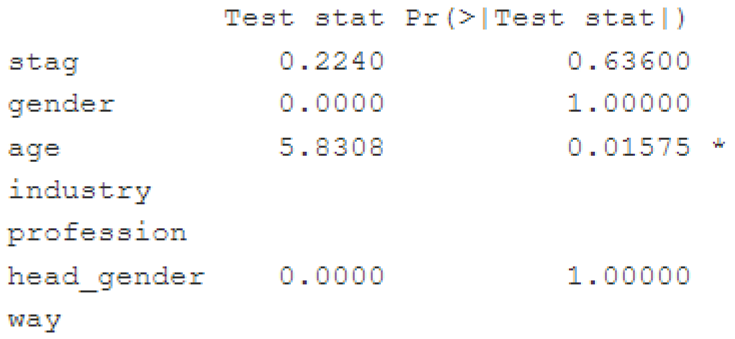
VIF values for predictors for final demographic model:



```
                GVIF Df GVIF^(1/(2*Df))
stag        1.111515  1        1.054284
gender      1.353825  1        1.163540
age         1.134438  1        1.065100
industry    1.321342  4        1.035445
profession  1.764425  5        1.058426
head_gender 1.201985  1        1.096351
way         1.132460  2        1.031587
```

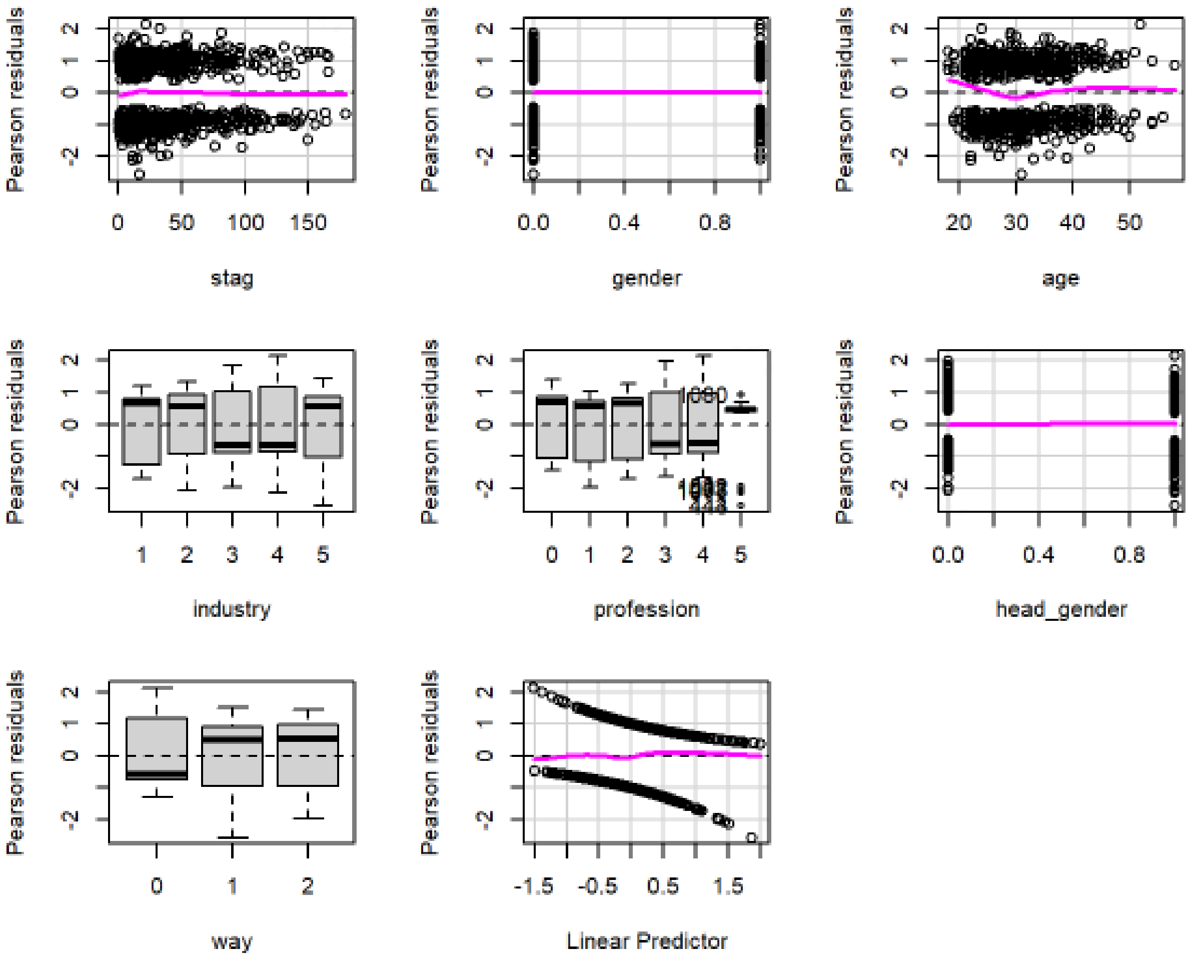
Finally, the Pearson's residual plots for the model were created and are shown below. Though deviances in linearity can still be observed in the predictor age, the deviance is slight enough for us to conclude that the predictors are properly specified for our demographic final model.

Lack of fit test results for final demographic regression model:



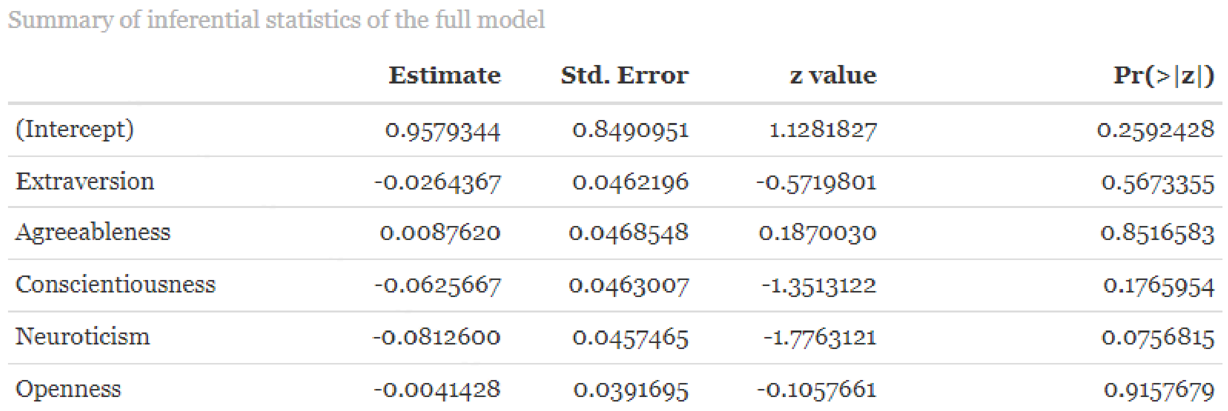
```
            Test stat Pr(>|Test stat|)
stag           0.2240          0.63600
gender         0.0000          1.00000
age            5.8308          0.01575 *
industry
profession
head_gender    0.0000          1.00000
way
```

Residual plots for final demographic model:



Logistic Regression of the Big 5 Personality Traits:

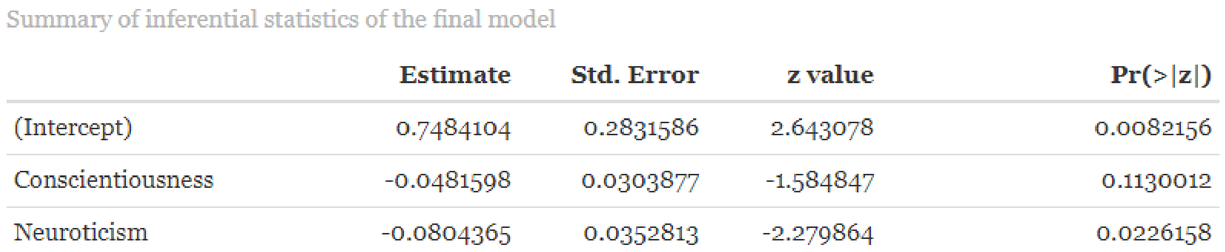
The second subgroup logistic regression model built was designed to estimate the probability of an employee leaving their role based on their big five personality trait scores.The initial logistic regression model used the variable event as the response variable and the 5 trait score variables as the explanatory variables. The summary of inferential statistics for the full big 5 model is presented below.



Summary of inferential statistics of the full model

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | 0.9579344 | 0.8490951 | 1.1281827 | 0.2592428 |
| Extraversion | -0.0264367 | 0.0462196 | -0.5719801 | 0.5673355 |
| Agreeableness | 0.0087620 | 0.0468548 | 0.1870030 | 0.8516583 |
| Conscientiousness | -0.0625667 | 0.0463007 | -1.3513122 | 0.1765954 |
| Neuroticism | -0.0812600 | 0.0457465 | -1.7763121 | 0.0756815 |
| Openness | -0.0041428 | 0.0391695 | -0.1057661 | 0.9157679 |

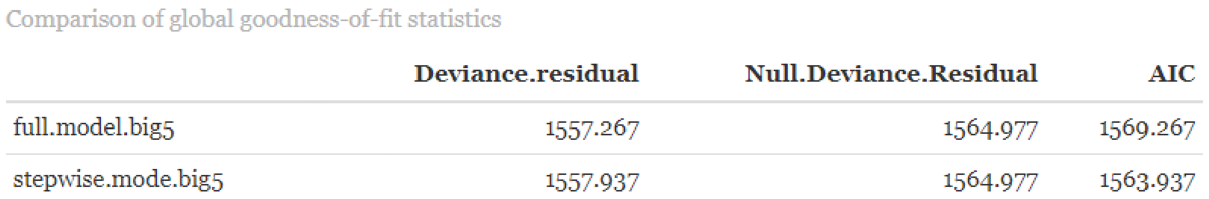
A backward AIC based variable selection algorithm, the same one used above, was used to drop non-contributing variables from the model. This algorithm removed the traits Extraversion, Agreeableness, and Openness from the model. The predictors Neuroticism and Conscientiousness were kept in the model. A summary table of the infertile statistics for the model created through the stepwise algorithm is shown below.



Summary of inferential statistics of the final model

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | 0.7484104 | 0.2831586 | 2.643078 | 0.0082156 |
| Conscientiousness | -0.0481598 | 0.0303877 | -1.584847 | 0.1130012 |
| Neuroticism | -0.0804365 | 0.0352813 | -2.279864 | 0.0226158 |

The model created through the stepwise AIC algorithm and the full model were again compared through three different goodness of fit measures: The null deviance residual, the deviance residual, and the AIC. These are displayed in the following table.

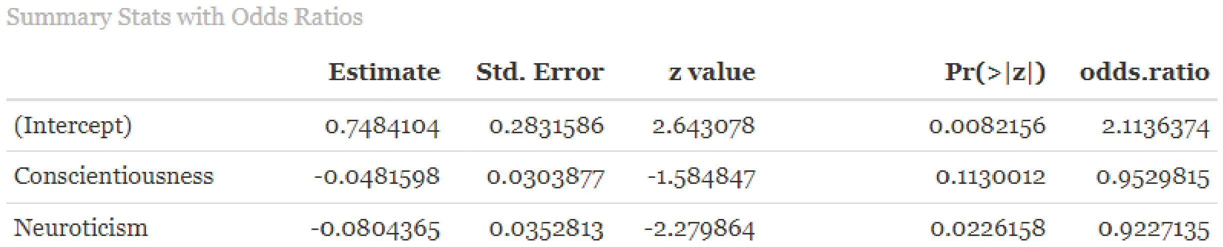


Comparison of global goodness-of-fit statistics

| | Deviance.residual | Null.Deviance.Residual | AIC |
|---|---|---|---|
| full.model.big5 | 1557.267 | 1564.977 | 1569.267 |
| stepwise.mode.big5 | 1557.937 | 1564.977 | 1563.937 |

Once again, a chi-squared test was conducted to examine the values for the null deviance and deviance residuals for the full and stepwise models. For the full model, this gives a *p* value of χ2(7.71, 5) =  0.172959. For the stepwise model, this would be a *p* value of χ2(7.04, 2) = 0.029599. The full model gives a *p* value that is greater than 0.05, indicating that the full model did not perform significantly better than a model including just the intercept. On the other hand, the created stepwise model for the personality traits gave a *p* value of 0.029599, which is less than 0.05, suggesting evidence that this model performed significantly better than that of a model without predictors. This, combined with the lower AIC score, suggest that it performs better than the full model and therefore was chosen as the final model for the demographic predictors.

The summary statistics of the final model are shown below with the calculated odd ratios for each predictor.

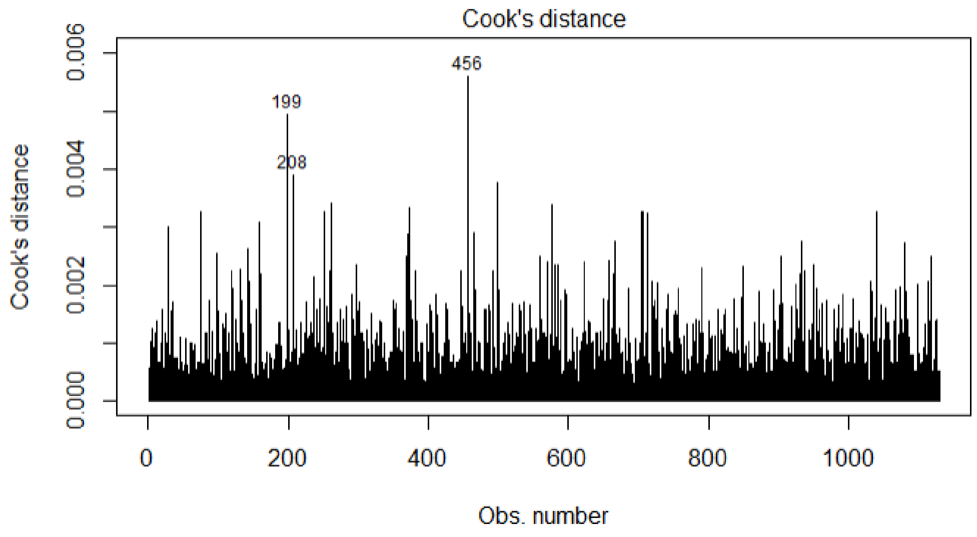


Summary Stats with Odds Ratios

| | Estimate | Std. Error | z value | Pr(>\|z\|) | odds.ratio |
|---|---|---|---|---|---|
| (Intercept) | 0.7484104 | 0.2831586 | 2.643078 | 0.0082156 | 2.1136374 |
| Conscientiousness | -0.0481598 | 0.0303877 | -1.584847 | 0.1130012 | 0.9529815 |
| Neuroticism | -0.0804365 | 0.0352813 | -2.279864 | 0.0226158 | 0.9227135 |

The model's diagnostics were examined to check for violations to assumptions through the presence of outliers or multicollinearity once again. The following Cook's D plot for the residuals marks the three observations with the highest measures of Cook's Distance.

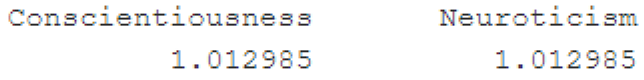
Cook's Distance for Residuals of Final Big5 Model



As there does not appear to be any one or few observations that have above and beyond large values for Cook's distance and the overall values for Cook's distance are quite small (<1), the data does not appear to have any extreme outliers that may need to be dealt with separately. Similarly, a look at the standardized residuals yields that none have a value higher than 2.

Multicollinearity was checked through values of VIF. As there are none over two, multicollinearity does not appear to be an issue for our final model.

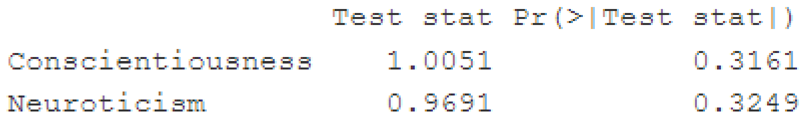
VIF values for final Big5 model predictors:



```
Conscientiousness       Neuroticism
         1.012985          1.012985
```

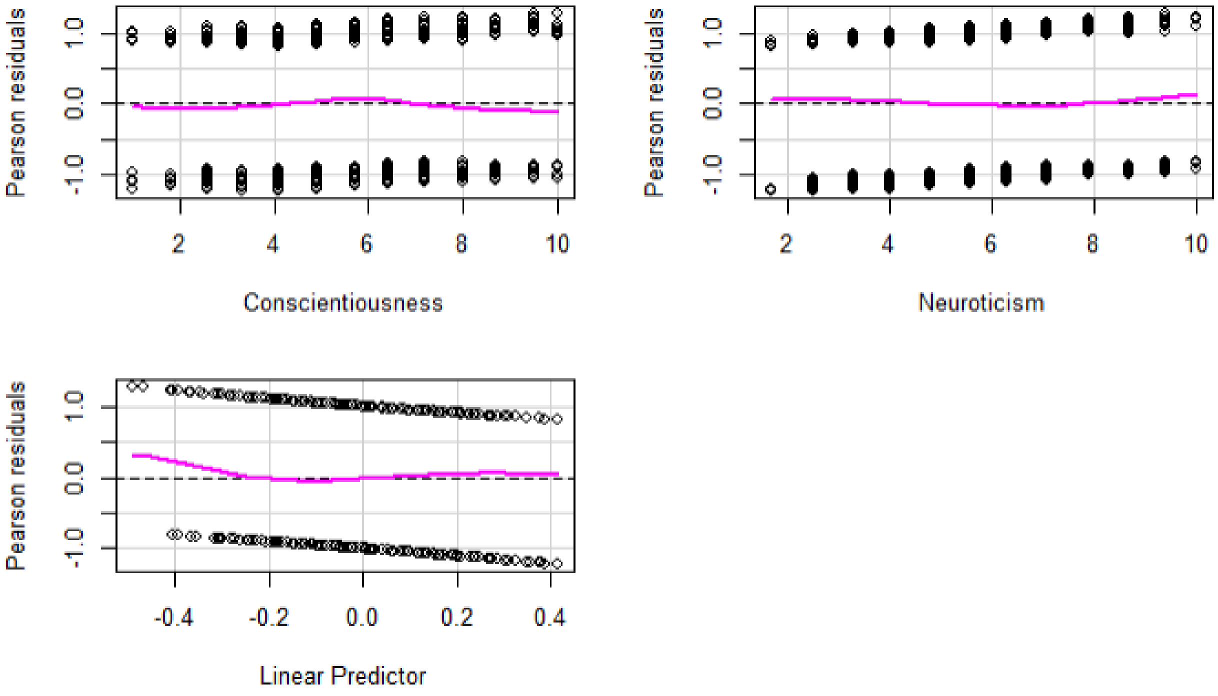
Finally, the Pearson's residual plots for the model were created and are shown below. The lack of fit test gave no significant p-values, and the residual plots show no obvious deviances in linearity for either of the predictors. Therefore, we conclude that the predictors are well-specified for the final stepwise model for the big 5 personality traits.

Lack of fit test:



```
                  Test stat Pr(>|Test stat|)
Conscientiousness    1.0051           0.3161
Neuroticism          0.9691           0.3249
```

Residual plots:



# Results

## Major Findings

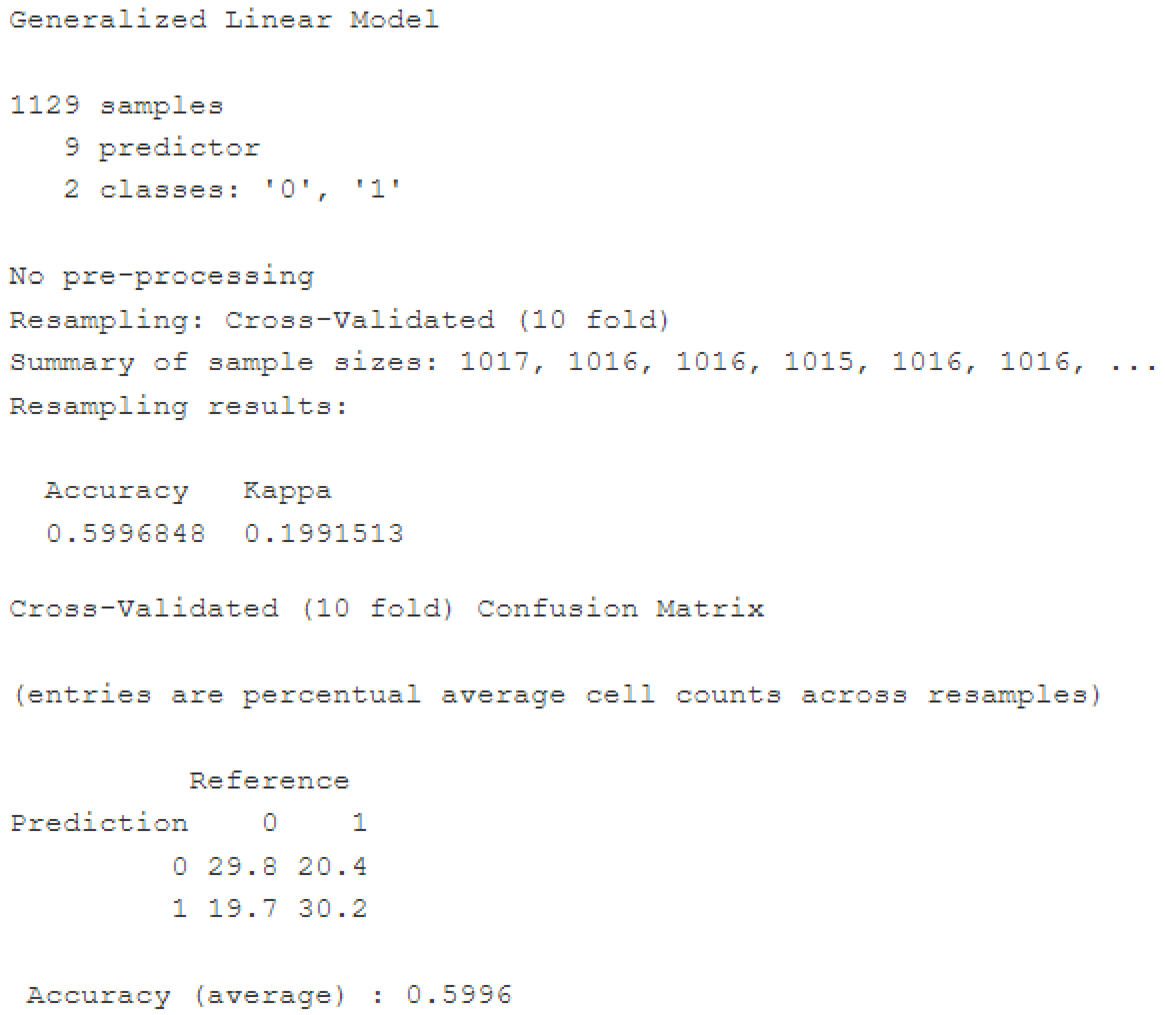
In the EDA portion of our analysis one thing that became clear was that there was near zero correlation between our response variable “event” and any of the potential predictor variables. We also noted that 4 of the 5 personality trait scores did not follow the expected normal distribution. While normality is not a necessary requirement for logistic regression, It is still concerning as normality is the expected distribution of this type of survey data.

The assumptions necessary for linear regression, such as minimal multicollinearity of the predictor variables, linearity of the relationship between predictors and the log odds ratio, and the lack of major outliers were all tested for and verified.

## Validation and Confusion Matrix

Cross validation and confusion matrices are used to evaluate the performance of our models. 1 10-fold cross-validation was performed to assess whether the model’s performance is robust and not overfitting to the data. The confusion matrix further evaluates the model performance by comparing the model’s predictions with the known values and then classifying the model’s performance into true positives, true negatives, false positives, and false negatives.

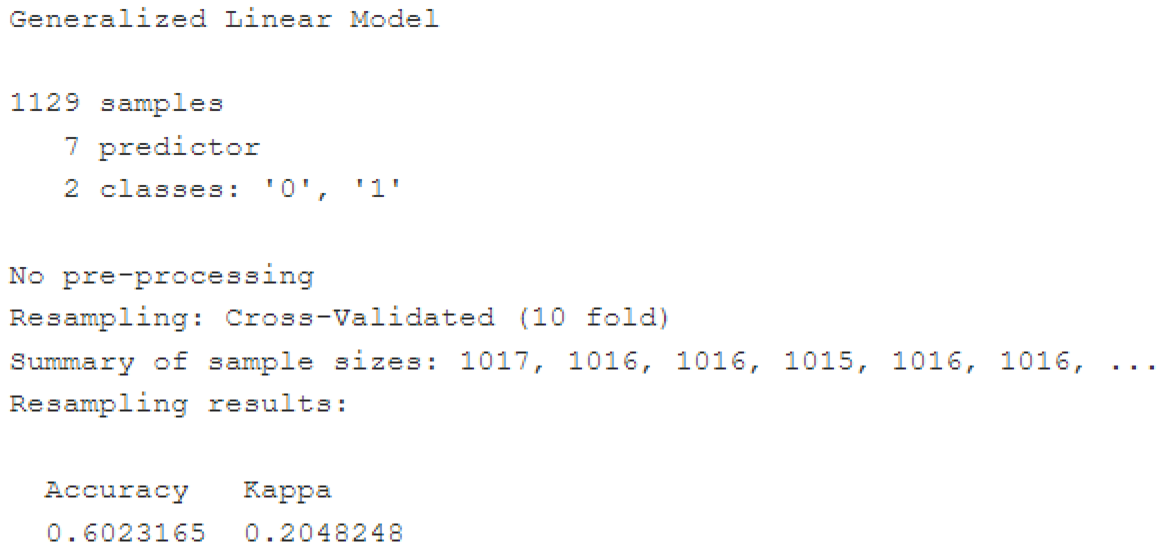
## Full Model



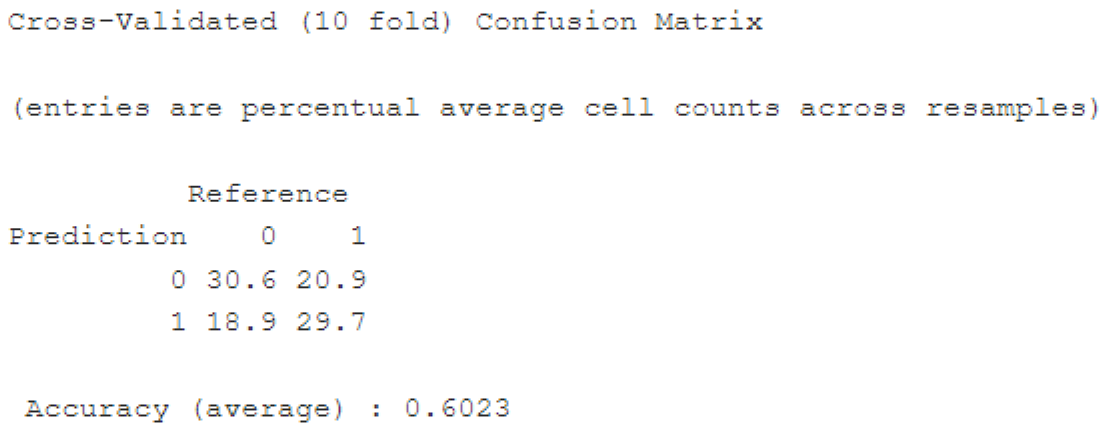
```
Generalized Linear Model

1129 samples
   9 predictor
   2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 1017, 1016, 1016, 1015, 1016, 1016, ...
Resampling results:

  Accuracy   Kappa
  0.5996848  0.1991513

Cross-Validated (10 fold) Confusion Matrix

(entries are percentual average cell counts across resamples)

          Reference
Prediction    0    1
         0 29.8 20.4
         1 19.7 30.2

 Accuracy (average) : 0.5996
```

The full model’s accuracy suggests that it can predict turnover moderately well - with the accuracy being 0.5996. However, the low Kappa score indicates limited predictive power beyond random chance. The confusion matrix reveals that while the model is fairly balanced in its error types - false positives and false negatives - there’s room for improvement, especially in reducing both types of errors to enhance predictive reliability.

## Demographic model



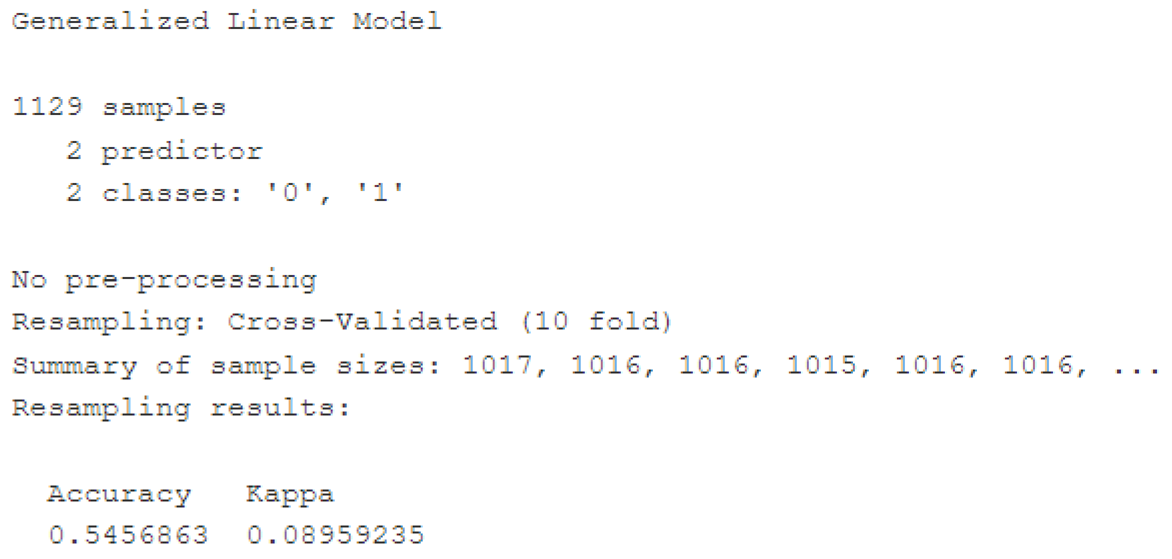
```
Generalized Linear Model

1129 samples
   7 predictor
   2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 1017, 1016, 1016, 1015, 1016, 1016, ...
Resampling results:

  Accuracy   Kappa
  0.6023165  0.2048248
```



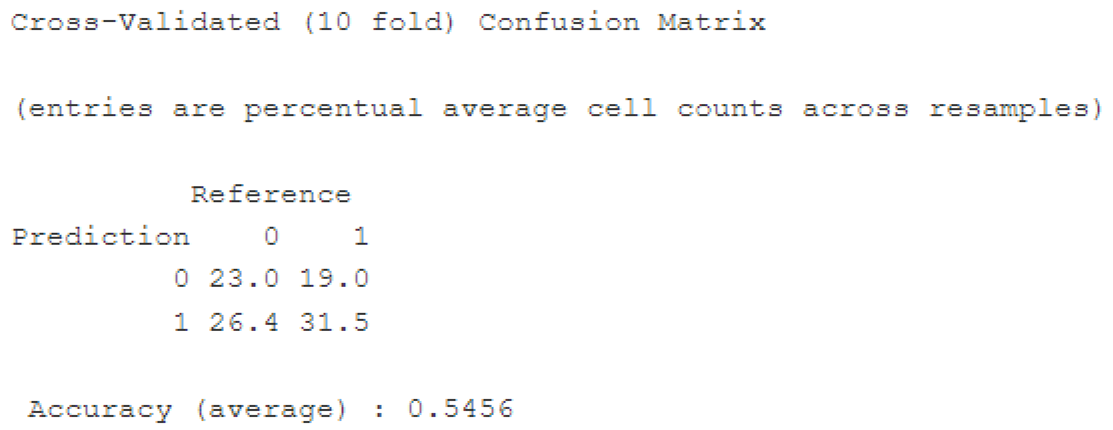
```
Cross-Validated (10 fold) Confusion Matrix

(entries are percentual average cell counts across resamples)

          Reference
Prediction    0    1
         0 30.6 20.9
         1 18.9 29.7

 Accuracy (average) : 0.6023
```

The demographic model’s accuracy after 10-fold cross-validation was conducted, showed an accuracy of 0.6023, suggesting moderate effectiveness in predicting employee turnover. The Kappa value of 0.2048, however, suggests a limited ability to predict beyond random chance. The confusion matrix shows a high rate of false positives (20.9%) and false negatives (18.9%). Again, there is area for improvement in reducing both types of errors to enhance predictive reliability.

## Big5 model



```
Generalized Linear Model

1129 samples
   2 predictor
   2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 1017, 1016, 1016, 1015, 1016, 1016, ...
Resampling results:

  Accuracy   Kappa
  0.5456863  0.08959235
```



```
Cross-Validated (10 fold) Confusion Matrix

(entries are percentual average cell counts across resamples)

          Reference
Prediction    0    1
         0 23.0 19.0
         1 26.4 31.5

 Accuracy (average) : 0.5456
```

The results from the Big5 model after a 10-fold cross-validation was performed achieved an accuracy of 54.56%. With a Kappa statistic of 0.0895, the predictive capability of the model is poor. The confusion matrix shows that the model has both a large false positive rate (19%) and false negative rate (26.4%). These figures indicate a need for improvement in predictive accuracy.

# General discussion

## Conclusion

The analysis conducted in this study utilized logistic regression to investigate the predictors of employee turnover, emphasizing the impact of demographic variables and the Big 5 personality traits. The models generated provide moderate predictive capabilities for employee turnover, with a focus on identifying the combination of factors most associated with turnover events. Notably, the study highlighted that demographic factors were generally more predictive of turnover than personality traits, suggesting the more pronounced role of concrete employee circumstances over psychological profiles. By and large, however, is that the model has very limited predictive capabilities as the models all did marginally better than random chance. There were many limitations to the study as well as some recommendations that would improve predictive power for future studies wishing to examine factors affecting employee turnover.

## Limitations

1. Data Distribution and Normality: Some of the Big 5 personality traits did not follow a normal distribution, which, while not a requirement for logistic regression, could indicate sampling or measurement issues that could affect the robustness of the findings.
2. Predictive Power: The relatively low Kappa scores across models suggest that the models' ability to predict turnover is only slightly better than random chance. This indicates a need for further refinement in model selection or the introduction of additional or different predictors.
3. Causality: The design of the study does not allow for causal inferences. The associations found are purely correlational, and causal links between the predictors and turnover cannot be established with certainty.
4. Sample Representativeness: Given that the study is based on data from Russian employees, the findings may not be generalizable to other cultural or geographic contexts.
5. Independence of Observations: Assumptions about the independence of observations might not hold if there are unaccounted-for groupings or clusters within the data, such as departments or teams that could influence turnover rates.

## Further Recommendation

1. Incorporating Additional Variables: Including more variables that capture economic conditions, job market dynamics, or individual employee performance could improve model performance.
2. Advanced Modeling Techniques: Exploring machine learning techniques such as random forests or boosted trees may provide better predictive accuracy and handle non-linearity and interaction effects more effectively.
3. Longitudinal Analysis: Implementing a longitudinal study design could help in understanding the changes in employee turnover over time and potentially offer insights into causal factors.
4. Cross-Cultural Validation: Conducting similar studies in different cultural settings could validate the findings and enhance the generalizability of the results.
5. Addressing Non-Normality: Investigating and applying transformations or alternative non-parametric methods might better capture the nature of the Big 5 personality data, potentially leading to more accurate models.

# 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) Bujang, Mohamad Adam, et al. “Sample Size Guidelines for Logistic Regression from Observational Studies with Large Population: Emphasis on the Accuracy between Statistics and Parameters Based on Real Life Clinical Data.” *The Malaysian Journal of Medical Sciences : MJMS*, U.S. National Library of Medicine, July 2018, www.ncbi.nlm.nih.gov/pmc/articles/PMC6422534/#:~:text=For%20observational%20studies%20with%20large%20population%20that%20involves%20logistic%20regression,parameters%20in%20the%20targeted%20population

(6) Bobbit, Zach. “How to Interpret Null & Residual Deviance (with Examples).” *Statology*, Statology, 1 Sept. 2021, www.statology.org/null-residual-deviance/

(7) 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