STA360 Final Project

Martin Lim Michael Xue

Introduction

The Big Five personality trait model was developed by various individual researchers. In 1936, Gordon Allport and Henry Odbert formed a list of 4,500 terms relating to personality traits, providing the foundation for other psychologists to start investigating the basic dimensions of personality. Using factor analysis, this list was eventually narrowed down to five key personality traits, which became known as the "Big Five". Each of these five traits actually encompasses a multitude of other traits. For example, extraversion is an aggregate of gregariousness, assertiveness, activity, excitement-seeking, positive emotions, and warmth. Another key aspect of this Big Five model is that it classifies personality traits on a spectrum, rather than into specific categories. The Big Five traits are not associated with any particular personality test, as there are several different ways to measure these traits. Our project focuses on specifically on the Big-Five Factor Markers from the International Personality Item Pool developed by Lewis R. Goldberg.

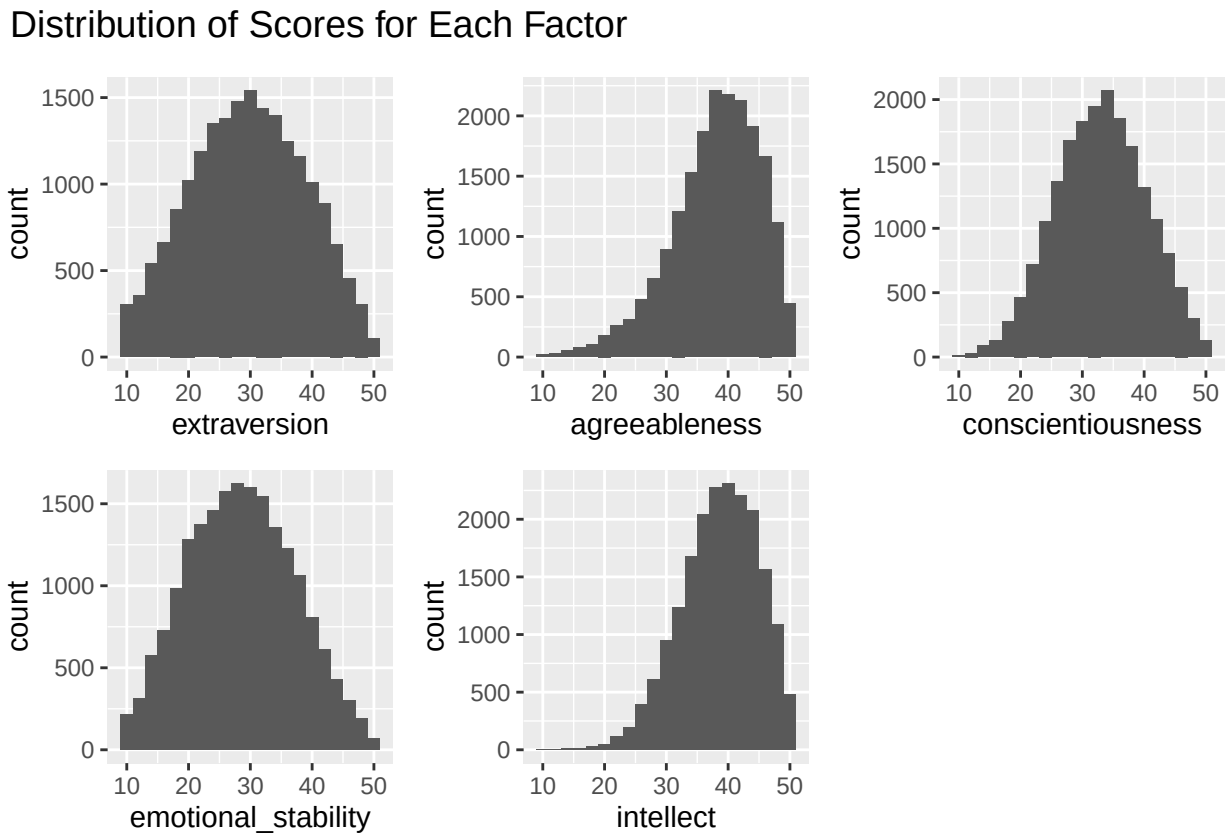
The goal of this project is to

Data

The data comes from the Open-Source Psychometrics Project, which provides a collection of interactive personality tests. The data was collected in 2012 through an online personality test of the Big-Five Factor Markers from the International Personality Item Pool on the Open-Source Psychometrics Project website. Prior to starting the test, participants were informed that their responses would be recorded and used for research purposes, and consent to record their responses was confirmed after the test. This dataset contains 57 variables: 50 of these variables are the participants' answers to the 50 question Big Five personality test and 7 of these variables are additional demographic information about the participant. The participants answer the 50 questions on a scale of 1 to 5; 1 means that the participant finds the statement "Very Inaccurate", 3 means that the statement is "Neither Accurate Nor Inaccurate", and 5 means that the statement is "Very Accurate". The 7 demographic variables that are not factored into the personality test are: race, age, gender, handedness, the participant's country location, how the participant came to the test, and whether English is the participant's native language. The Big Five factors that this 50 question test measures are "extraversion", "agreeableness", "conscientiousness", "emotional stability", and "intellect". Each question on the test corresponds to only one of these factors, and each factor is defined by 10 questions.

The original dataset included 19719 observations, but 378 observations were removed as they were missing the country variables. Given that each observation is the result of just one participant, removing these observations should not have any significant impact on our findings. From the 50 questions, additional variables were created to store the scores received in each of the Big Five factors. Scoring for the factors is computed as follows: each question corresponds to a single factor and is "+keyed" or "-keyed". For "+keyed" items, the response "Very Inaccurate" is assigned a value of 1, "Moderately Inaccurate" a value of 2, "Neither Inaccurate nor Accurate" a 3, "Moderately Accurate" a 4, and "Very Accurate" a value of 5. For "-keyed" items, the response "Very Inaccurate" is assigned a value of 5, "Moderately Inaccurate" a value of 4, "Neither Inaccurate nor Accurate" a 3, "Moderately Accurate" a 2, and "Very Accurate" a value of 1. Once values are calculated for each question, they should be summed to obtain the total score for that factor. Each factor can have a different number of ("+keyed", "-keyed") questions: extraversion (5, 5), agreeableness (6, 4), conscientiousness (6, 4), emotional stability (2, 8), intellect (3, 7). After obtaining a score for each factor on

each observation, we found that scores were normally distributed for each factor, with agreeableness and intellect being a little left skewed.



To find the continent codes for each country, we imported a data frame from an outside source mapping country code to the corresponding continent code. We merged the two data frames together and selected only the necessary variables. The continents represented in the final data are North America (NA), Europe (EU), Asia (AS), Oceania (OC), Africa (AF), and South America (SA).

```
##   Continent Count
## 1        NA  9899
## 2        EU  4039
## 3        AS  3623
## 4        OC  1140
## 5        AF   394
## 6        SA   292
```

Modeling

Results

Conclusion