Assignment #2: Exploratory Factor Analysis

Harini Anand

1. Introduction

In modern social sciences, survey research is one of the essential methods of collecting data for research. There are several types of surveys. The data obtained from surveys are analyzed using statistical methods to draw meaningful conclusions. However, in some cases, it is not possible to measure the concepts of primary interest directly. One such example is in the case of personality surveys. Personality surveys are used to assess the human personality constructs by collecting information on variables that are indicators of the personality constructs.

In this report, we present the results of the Exploratory Factor Analysis (EFA) conducted on the personality assessment data. EFA is a set of extraction and rotation techniques used to model the unobserved or latent constructs (in our example, the human personality constructs). The idea is to transform the original set of measured variables into a number of factors. EFA examines all the pairwise relationships between the measured individual variables and seeks to extract latent factors from the measured variables. Each measured variable is expressed as a linear combination of the underlying, latent factors. But this method does not add or remove information, but only transforms the data into a different form. Hence, we were able to use the latent factors generated by EFA to study if there are any demographic – age, gender, educational - differences in the personality traits.

2. Data

The dataset used for this analysis is the BFI (Big Five Inventory) data from the International Personality Item Pool (ipip.ori.org) as part of the Synthetic Aperture Personality Assessment (SAPA) web-based personality assessment project. The BFI data contains survey answers from 2800 subjects. Each record in the dataset corresponds to a subject. Each record contains three demographic variables (sex, education, and age) and 25 personality variables that carry the subject's self-reported answers to 25 personality questions. The personality variables in the dataset are Likert type variables measured on a scale from 1 to 6 (1 = not at all like me, and 6=totally like me) to reflect the subject's answer to the personality questions.

Among the demographic variables, sex is a nominal variable, and education is an ordinal variable. Though typically age is a continuous ratio variable, in the dataset, the age values recorded have been truncated to whole numbers with the range extending from 3 years to 86 years. All the personality variables are ordinal variables since the answers to personality questions are measured using a Likert scale.

3. Data preparation

As part of the data preparation, we removed any records in the BFI data containing missing values (NA). This resulted in the removal of 564 records bringing down the total number of records with complete information in the dataset to 2236. For EFA, as a rule of thumb, we require a minimum sample size of at least 20 records per variable. Since there are 28 variables in the dataset, overall, we would need 20*28 = 560 records in the data. With 2236 records, we have enough data to conduct the EFA analysis.

Code Snippet 01 in section 12.1 contains the R code for the data preparation.

4. Exploratory data analysis and correlation plot

In this section, we present the results of the exploratory data analysis conducted on the data. Basic EDA shows that there are 735 records from male subjects and 1501 records from female subjects. Table 01 shows the distribution of the subjects' education. The dataset has most records from subjects who reported to have some college education and the least number of records from subjects in high school. Also, though the age variable spans from 3 years to 86 years, most of the observations are from subjects whose reported age is greater than 15 years and less than 56 years old.

1 = In High School	2 = Finished High School	3 = Some College	4 = College Graduate	5 = Graduate Degree
198	250	1078	346	364

Table 01: Education distribution of the 2236 subjects

Next, we examined the correlations among the 25 personality variables by using the ***corrplot***. FIG 01 shows the corrplot.

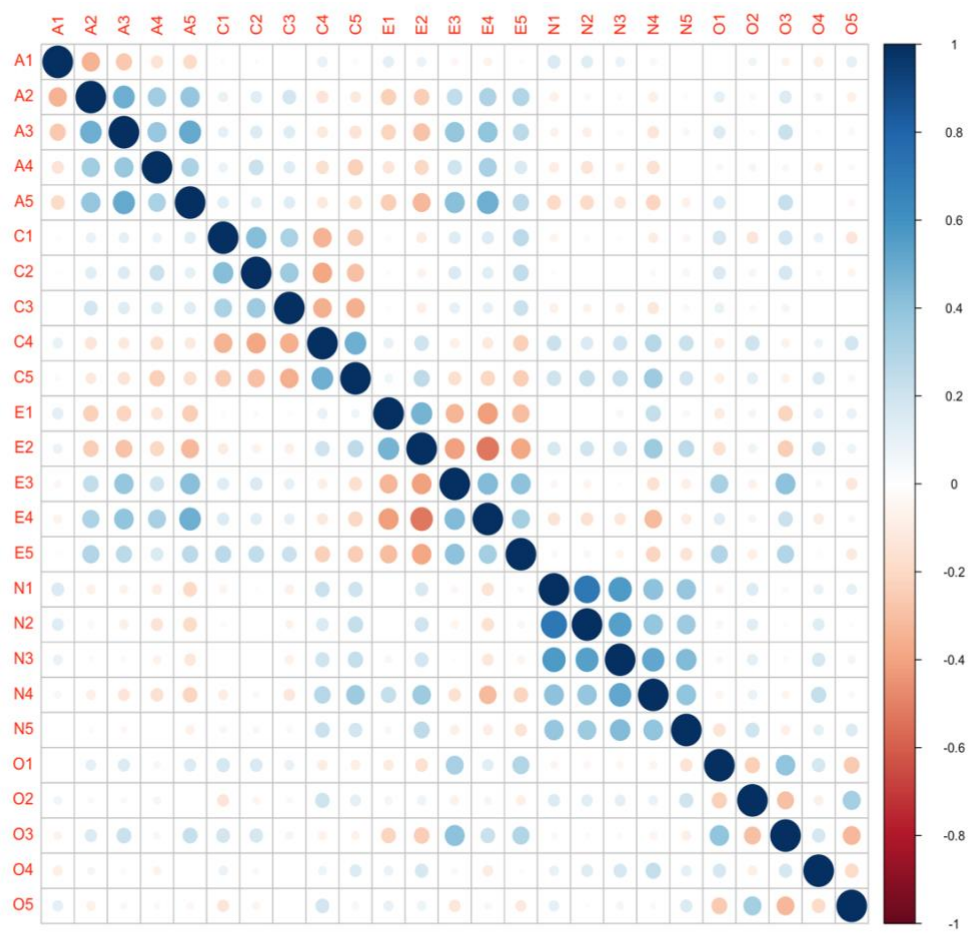

FIG 01: A plot of the correlation coefficients among the 25 personality variables

There are some patterns or constructs noted in the corrplot (Note: the diagonal values – self-correlation values which are always one - are excluded in this discussion):

- The set of personality variables A1-A5

  These variables present mild to moderate correlations with other variables in the set. The variables A2-A5 show some mild to moderate negative correlations with A1. But among the A2-A5 variables, the correlations are positive, which ranges from mild to moderate.

- The set of personality variables C1-C5

  The variables C1-C3 depict mild positive correlations with each other. On the contrary, C4-C5 variables interact with mild negative correlation with C1-C3. But C4 and C5 show a moderate positive correlation with each other.

- The set of personality variables E1-E5

  In this set, E1 has a moderately positive correlation with E2. E1 has a mild to moderately negative correlation with E3, E4, and E5. E2, too, has a mild to moderate negative correlation with E3 and E5. But notably, E2 has a strong negative correlation with E4. E3 has a moderately positive correlation with E4 and E5. E4 and E5 have a mild positive correlation.

- The set of personality variables E1-E5 and A2-A5

  E1 and E2 interact with a mild negative correlation with A2-A5. But E3-E5 share mild to a moderate positive correlation with A2-A5.

- The set of personality variables N1-N5

  N1 and N2 have a strong positive correlation. But N1 and N2 present moderate to a mild positive correlation with N3, N4, and N5. Similar mild to moderate positive correlations exists among N3, N4, and N5.

- The set of personality variables O1-O5

  O5 shows moderate (positive and negative) correlations with O1-O3. O3 also has a moderate correlation (positive and negative) with O1-O2 and O5. O1 and O3 have a moderate positive correlation.

- The set of personality variables O3 and E1-E5.

  O3 shows mild negative correlations with E1-E2. But O3 has a moderate positive correlation with E3 and some mild positive correlation with E4-E5.

Code Snippet 02 in section 12.2 contains the R code for the Exploratory Data Analysis (EDA).

5. Estimation of eigenvalues/ eigenvectors of the correlation matrix

In the section, we discuss the eigenvalues and eigenvectors computed using the correlation matrix. EFA transforms the original data in the direction of the eigenvectors. The corresponding eigenvalues for these eigenvectors indicate the total amount of variance in the measured variables explained by the common factors. An Eigenvalue of 1 implies that factor does not explain any more variance than that of a single measured variable. Table 02 shows the eigenvalues of the correlation matrix.

Factor No.	Eigen value	Factor No.	Eigen value	Factor No.	Eigen value	Factor No.	Eigen value
1	**5.0685162**	8	0.8045002	15	0.5659652	22	0.4070974
2	**2.7624793**	9	0.7140883	16	0.5448396	23	0.3888753
3	**2.1526230**	10	0.7015381	17	0.5199335	24	0.3847626
4	**1.8923330**	11	0.6808421	18	0.4938686	25	0.2681008
5	**1.5175329**	12	0.6489735	19	0.4827362		
6	**1.0788293**	13	0.6312563	20	0.4425003		
7	0.8309057	14	0.5880320	21	0.4288706		

Table 02: Eigen values of the correlation matrix

We also plotted the eigenvalues with the FA (Factor Analysis) parallel analysis, which displays both the PCA components and FA factors. In this plot, the eigenvalues are plotted from the largest to the smallest. The plot is known as the scree plot. FIG 02 shows the scree plot generated using the correlation matrix from section 4.

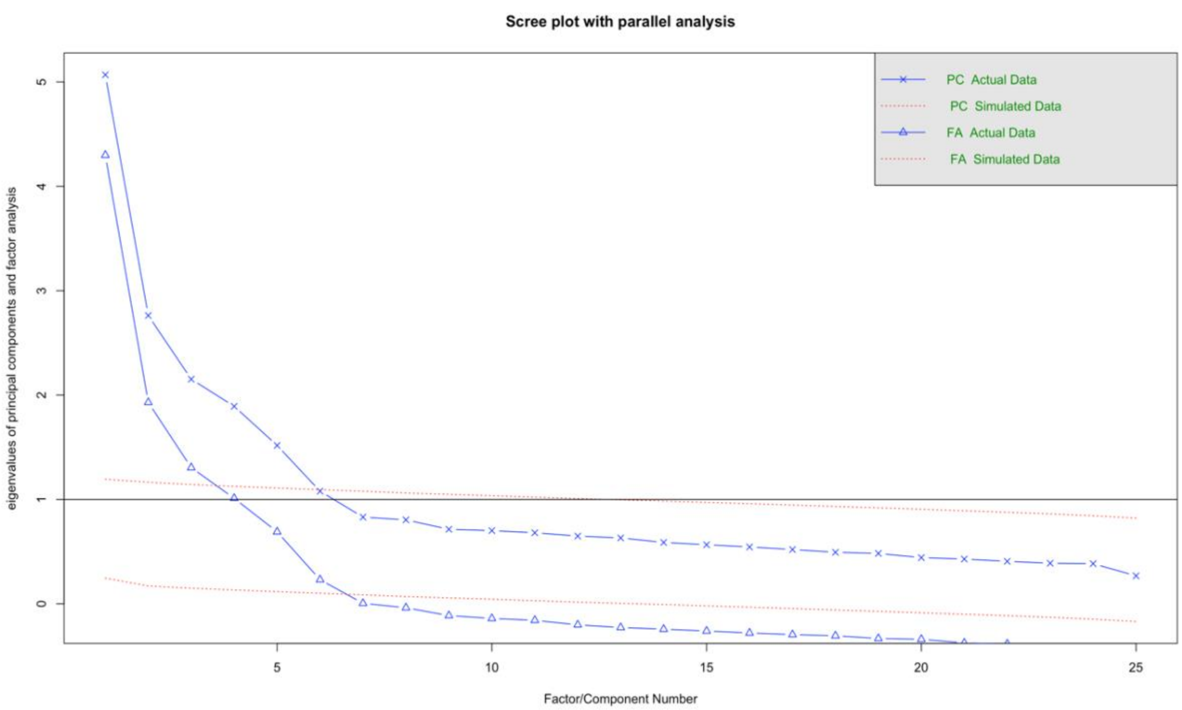

FIG 02: Scree plot with parallel analysis

Several methods exist to select the optimal number of factors for our analysis. Some of the methods we considered are:

a) Eigenvalue greater-than-one rule
b) Cattell's Scree test
c) Percent of Total Variance greater than 90%

Eigenvalue greater-than-one rule:

Since an eigenvalue of 1 implies the factor does not explain any more variance than the single measured variable, for our analysis, we are interested in eigenvalues greater than 1. By applying this rule to the data in Table 02, we determined to retain the first 6 factors.

Cattell's Scree test:

For this test, we utilized FIG 02 in which the eigenvalues are plotted from the largest to the smallest. In FIG 02, we identified the "elbow" where the eigenvalues level off. Using visual observation, we observed the "elbow" occurs at about 7th eigenvalue. That led us to retain the first 6 eigenvalues, which corresponds to the first 6 factors.

Percent of Total Variance greater than 90%:

Table 03 shows the cumulative variance and the cumulative proportion for the 25 factors generated by fa() R function. Based on the rule that percent of the total variance is greater than 90 percent, we determined we need to retain the first 9 factors. However, in this case, some of the eigenvalues are less than 1.

Factor	SS loadings /Eigen value	Proportion Var	Cumulative Var	Proportion Explained	Cumulative Proportion
Factor 1	**4.67**	**0.19**	**0.19**	**0.33**	**0.33**
Factor 2	**2.41**	**0.1**	**0.28**	**0.17**	**0.5**
Factor 3	**1.7**	**0.07**	**0.35**	**0.12**	**0.62**
Factor 4	**1.4**	**0.06**	**0.41**	**0.1**	**0.72**
Factor 5	**1.07**	**0.04**	**0.45**	**0.08**	**0.8**
Factor 6	**0.63**	**0.03**	**0.47**	**0.04**	**0.84**
Factor 7	**0.39**	**0.02**	**0.49**	**0.03**	**0.87**
Factor 8	**0.34**	**0.01**	**0.5**	**0.02**	**0.89**
Factor 9	**0.24**	**0.01**	**0.51**	**0.02**	**0.91**
Factor 10	0.24	0.01	0.52	0.02	0.93
Factor 11	0.21	0.01	0.53	0.01	0.94
Factor 12	0.16	0.01	0.54	0.01	0.95
Factor 13	0.13	0.01	0.54	0.01	0.96
Factor 14	0.12	0	0.55	0.01	0.97
Factor 15	0.1	0	0.55	0.01	0.98
Factor 16	0.1	0	0.56	0.01	0.99
Factor 17	0.07	0	0.56	0.01	0.99
Factor 18	0.05	0	0.56	0	0.99
Factor 19	0.03	0	0.56	0	1
Factor 20	0.02	0	0.56	0	1
Factor 21	0.02	0	0.56	0	1
Factor 22	0.01	0	0.56	0	1
Factor 23	0	0	0.56	0	1
Factor 24	0	0	0.56	0	1
Factor 25	0	0	0.56	0	1

Table 03: Shows cumulative variance and proportion for the first 10 factors in the fa() output

Code Snippet 03 in section 12.3 contains the R code for eigenvalue, eigenvector computation, and scree plot.

6. Maximum likelihood factor analysis with a VARIMAX rotation

In section 5, we determined by using the eigenvalue >= 1 rule that we would retain the first 6 factors. In this section, we present the results of a factor model obtained using the maximum likelihood factor analysis with a VARIMAX rotation. VARIMAX is an orthogonal factor rotation that maximizes the sum of variances of loadings of the factor matrix. The maximum likelihood (ML) factor analysis assumes the observed variables follow a multivariate normal distribution. The ML method results in estimates which most likely generate the observed correlation matrix. The correlations are weighted by each variable's uniqueness.

The resultant factor model, along with the factor loadings, is listed in Table 04. The factor loading of a factor provides the correlation between the original variable and the factor. We chose to eliminate variables that are not "strong" based on the factor loading value. The cutoff value we employed for the loadings is |0.5|. The yellow highlights in Table 04 show the loading values that meet the cutoff in each factor.

Among the six factors used in the model, after applying the cutoff values for the factor loadings, we are able to interpret the factors as follows:

- ***ML1*** - upon applying the cutoff, only the E1-E5 variables stand out. Based on these variables from the data dictionary, we interpret the ML1 factor as the **gregariousness/sociable nature** of an individual.
- ***ML2*** – After the application of the cutoff, only N1-N5 variables remain. We interpret this factor as the **depression/anxiety trait(s)** of an individual.
- ***ML3*** – After the application of the cutoff value on the loadings, only C1-C5 variables remain. Based on these variables, we interpret ML3 as a **dutifulness/organization skill** of an individual.
- ***ML5*** – The application of the cutoff leaves only the factors for the variables A1-A3. Based on this, the ML5 factor is interpreted as an **individual's compassion**.
- ***ML4*** – For this factor, the cutoff value retains only O2-O3 and O5. We interpret ML4 as an **individual's thinker/reflective nature**.
- ***ML6*** – Does not have any factor loadings above the cutoff. **Also, ML6 has a SS loadings value less than 0.**

Therefore, among the first 6 factors, we are able to provide interpretation only for the factors ML1, ML2, ML3, ML5, and ML4. **We determined that ML6 is not helpful.**

Based on Table 04, we can determine that the cumulative Var is 0.45 for the 6 factors. The Cumulative Proportion is 100% for 6 factors. The first five factors account for 94% (Cumulative Var is 0.42), which is sufficient.

Factor Analysis using method = ml
Call: fa(r = cor.matrix, nfactors = 6, rotate = "varimax", fm = "ml")
Standardized loadings (pattern matrix) based upon correlation matrix

	ML1	ML2	ML3	ML5	ML4	ML6	h2	u2	com
A1	0.03	0.1	0.05	**-0.53**	-0.11	0.12	0.33	0.67	1.3
A2	0.26	0.04	0.13	**0.65**	0.04	-0.01	0.5	0.5	1.4
A3	0.38	0.01	0.13	**0.57**	0.03	0.15	0.51	0.49	2.1
A4	0.24	-0.07	0.24	0.39	-0.15	0.1	0.3	0.7	3.1
A5	0.45	-0.14	0.11	0.43	0.02	0.23	0.47	0.53	2.8
C1	0.08	0	**0.55**	-0.01	0.19	0.08	0.35	0.65	1.3
C2	0.04	0.07	**0.67**	0.05	0.09	0.16	0.48	0.52	1.2
C3	0.04	-0.04	**0.55**	0.1	-0.01	0	0.32	0.68	1.1
C4	-0.06	0.22	**-0.63**	-0.1	-0.12	0.31	0.57	0.43	1.9
C5	-0.18	0.27	**-0.55**	-0.04	0.03	0.14	0.43	0.57	1.9
E1	**-0.57**	0.03	0.06	-0.13	-0.07	0.18	0.39	0.61	1.4
E2	**-0.67**	0.23	-0.09	-0.09	-0.05	0.12	0.54	0.46	1.4
E3	**0.59**	0	0.11	0.14	0.25	0.23	0.5	0.5	1.9
E4	**0.68**	-0.14	0.11	0.21	-0.11	0.14	0.57	0.43	1.5
E5	**0.51**	0.05	0.31	0.07	0.2	-0.06	0.41	0.59	2.1
N1	0.05	**0.82**	-0.05	-0.16	-0.07	-0.12	0.72	0.28	1.2
N2	0.01	**0.8**	-0.04	-0.12	0	-0.19	0.69	0.31	1.2
N3	-0.07	**0.71**	-0.05	-0.01	-0.01	0.08	0.52	0.48	1.1
N4	-0.34	**0.56**	-0.16	0	0.07	0.2	0.5	0.5	2.2
N5	-0.15	**0.52**	-0.04	0.09	-0.17	0.16	0.35	0.65	1.7
O1	0.23	-0.02	0.13	-0.01	**0.49**	0.16	0.33	0.67	1.9
O2	0.01	0.17	-0.09	0.05	**-0.5**	0.14	0.31	0.69	1.5
O3	0.34	0.02	0.08	0.04	**0.58**	0.18	0.5	0.5	1.9
O4	-0.16	0.21	-0.03	0.12	0.35	0.17	0.24	0.76	3
O5	-0.01	0.06	-0.05	-0.08	**-0.58**	0.15	0.37	0.63	1.2

	ML1	ML2	ML3	ML5	ML4	ML6
SS loadings	**2.73**	**2.72**	**2.05**	**1.56**	**1.54**	0.62
Proportion Var	0.11	0.11	0.08	0.06	0.06	0.02
Cumulative Var	0.11	0.22	0.30	0.36	0.42	0.45
Proportion Explained	0.24	0.24	0.18	0.14	0.14	0.06
Cumulative Proportion	0.24	0.49	0.67	0.81	0.94	1.00

Mean item complexity = 1.7
Test of the hypothesis that 6 factors are sufficient.

The degrees of freedom for the null model are 300 and the objective function was 7.41
The degrees of freedom for the model are 165 and the objective function was 0.36

The root mean square of the residuals (RMSR) is 0.02
The df corrected root mean square of the residuals is 0.03

Fit based upon off diagonal values = 0.99
Measures of factor score adequacy

	ML1	ML2	ML3	ML5	ML4	ML6
Correlation of (regression) scores with factors	0.89	0.93	0.87	0.82	0.84	0.73
Multiple R square of scores with factors	0.80	0.87	0.75	0.68	0.70	0.54
Minimum correlation of possible factor scores	0.59	0.74	0.51	0.35	0.40	0.07

Table 04: Factor model with Maximum Likelihood method and VARIMAX rotation

We also ran the ***factanal()*** on the correlation matrix to study the statistical inference of the maximum likelihood factor analysis. The assessment of whether we have the correct number of factors to describe this correlation matrix is done using the chi-square test. The null and alternative hypotheses in this case are:

Ho (Null): The factor model describes the data well, i.e., 6 factors are sufficient
Ha (Alternative): Unrestricted correlations model

The chi-square test is used to perform a statistical test that determines whether the model does not fit the data significantly worse than a model where the variables correlate freely. The null hypothesis can also be stated as - the model predicted covariance matrix is equivalent to the actual covariance matrix. Based on the test-statistic (809.11 on 165 degrees of freedom) and p-value (1.41e-85) shown for the hypothesis test in Table 05, **we reject the null hypothesis that the six-factor model is sufficient.**

```
factanal(factors = 6, covmat = cor.matrix, n.obs = 2236, rotation = "varimax")

Uniquenesses:
   A1    A2    A3    A4    A5    C1    C2    C3    C4    C5    E1    E2    E3    E4    E5    N1    N2    N3
0.672 0.496 0.490 0.700 0.530 0.650 0.516 0.683 0.429 0.570 0.611 0.456 0.498 0.430 0.591 0.283 0.306 0.477
   N4    N5    O1    O2    O3    O4    O5
0.498 0.649 0.666 0.692 0.505 0.763 0.630
```

Loadings:

	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6
A1				**-0.534**	-0.113	0.124
A2	0.26		0.128	**0.645**		
A3	0.384		0.127	**0.568**		0.153
A4	0.239		0.236	0.387	-0.152	0.102
A5	0.446	-0.137	0.107	0.435		0.227
C1			**0.549**		0.188	
C2			**0.665**			0.158
C3			**0.551**			
C4		0.222	**-0.633**	-0.101	-0.12	0.305
C5	-0.184	0.273	**-0.548**			0.137
E1	**-0.575**			-0.133		0.178
E2	**-0.675**	0.233				0.124
E3	**0.594**		0.112	0.141	0.25	0.231
E4	**0.678**	-0.14	0.114	0.215	-0.108	0.14
E5	**0.515**		0.306		0.197	
N1		**0.815**		-0.162		-0.124
N2		**0.802**		-0.122		-0.186
N3		**0.714**				
N4	-0.342	**0.562**	-0.16			0.198
N5	-0.149	**0.516**			-0.165	0.164
O1	0.232		0.134		0.487	0.158
O2		0.168			**-0.5**	0.138
O3	0.343				**0.581**	0.178
O4	-0.163	0.21		0.125	0.348	0.17
O5					**-0.58**	0.148

	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6
SS loadings	**2.728**	**2.718**	**2.049**	**1.560**	**1.537**	0.617
Proportion Var	0.109	0.109	0.082	0.062	0.061	0.025
Cumulative Var	0.109	0.218	0.300	0.362	0.424	0.448

Test of the hypothesis that 6 factors are sufficient.
The chi square statistic is 809.11 on 165 degrees of freedom.
The p-value is 1.41e-85

Table 05: Factor model with Maximum Likelihood method and VARIMAX rotation using factanal()

Code Snippet 04 in section 12.4 contains the R code for Maximum likelihood factor analysis with a VARIMAX rotation.

7. Maximum likelihood factor analysis with a PROMAX rotation

In this section, we present the results of an oblique factor rotation called PROMAX using the maximum likelihood factor analysis method. We have used ***factanal()*** R function to perform this rotation. Table 06 shows the output of ***factanal()*** for the factor model obtained with maximum likelihood factor analysis and PROMAX rotation.

Among the six factors used in the PROMAX rotation model, after applying the **cutoff value of |0.5|** for the factor loadings, we are able to interpret the factors as follows:

- ***Factor1*** - upon applying the cutoff, only the E1-E4 variables stand out. Based on these variables from the data dictionary, we interpret the factor1 as capturing the **Extroverted vs. Introverted nature** of an individual.
- ***Factor2*** – After the application of the cutoff, only N1-N3 variables remain with strong loading values. We interpret this factor as the **irritability** of an individual.
- ***Factor3*** – After the application of the cutoff value on the loadings, the C1-C5 variables remain. Based on these variables, we interpret Factor3 as an **individual's efficiency vs. lax attitude to work.**
- ***Factor4*** – The application of the cutoff leaves only the factors for the variables O2-O3 and O5. Based on this, the Factor4 factor is interpreted as an individual's **uncurious vs. intellectual** nature.
- ***Factor5*** – For this factor, the cutoff value retains only A1-A3. We interpret Factor5 as an **individual's compassionate** nature.
- ***Factor6***– For this factor, the cutoff value provides the only C4 (Do things in a half-way manner). We can interpret Factor6 **as a lack of thoroughness.**

All six factors are worth keeping because their SS loadings values are greater than 1. The PROMAX factor model **posed difficulty with providing an interpretation** of the factors particularly in the case of Factors 4, 5, and 6. Factors 1 and 3 that are obtained using the PROMAX factor rotation have similar interpretability on the factors obtained with the VARIMAX model. But Factor 6 is worth keeping in the case of the PROMAX model but not in the case of VARIMAX since its SS loadings value is less than 1. The Cumulative Var is 0.47 with the six factors.

For the statistical inference for this maximum likelihood factor analysis of whether we have the correct number of factors to describe this correlation matrix, we used the chi-square test. The null and alternative hypotheses are:

Ho (Null): The factor model describes the data well i.e., 6 factors are sufficient
Ha (Alternative): Unrestricted correlations model

Similar to the previous section, the chi-square test in this case is also used to perform a statistical test whether the model does not fit significantly worse than a model where the variables correlate freely Based on the test-statistic (809.11 on 165 degrees of freedom) and p-value (1.41e-85) shown for the hypothesis test in Table 06, **we reject the null hypothesis that the six-factor model is sufficient. The chi-square test is not affected by factor rotation.**

```
factanal(factors = 6, covmat = cor.matrix, n.obs = 2236, rotation = "promax")

Uniquenesses:
   A1    A2    A3    A4    A5    C1    C2    C3    C4    C5    E1    E2    E3    E4    E5    N1    N2    N3
0.672 0.496 0.490 0.700 0.530 0.650 0.516 0.683 0.429 0.570 0.611 0.456 0.498 0.430 0.591 0.283 0.306 0.477
   N4    N5    O1    O2    O3    O4    O5
0.498 0.649 0.666 0.692 0.505 0.763 0.630

Loadings:
```

	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6
A1	0.169			-0.109	-0.606	0.171
A2	0.124				0.658	
A3	0.289				0.513	0.165
A4	0.165		0.194	-0.19	0.332	
A5	0.394	-0.191			0.33	0.283
C1			0.592	0.134		
C2			0.735			
C3			0.61			-0.183
C4			-0.704		-0.113	0.582
C5	-0.101	0.102	-0.576			0.327
E1	-0.63	-0.21	0.183		-0.117	0.114
E2	-0.729					
E3	0.604			0.236		0.336
E4	0.726			-0.128		0.222
E5	0.497	0.225	0.233	0.169		
N1	0.136	0.926				-0.123
N2		0.938				-0.211
N3		0.654				0.101
N4	-0.359	0.347				0.24
N5	-0.148	0.38		-0.172	0.118	0.174
O1	0.184			0.48		0.199
O2				-0.504		0.181
O3	0.297			0.579		0.258
O4	-0.25			0.35	0.141	0.191
O5				-0.586	-0.124	0.183

```
               Factor1 Factor2 Factor3 Factor4 Factor5 Factor6
SS loadings      2.788   2.611   2.255   1.516   1.396   1.186
Proportion Var   0.112   0.104   0.090   0.061   0.056   0.047
Cumulative Var   0.112   0.216   0.306   0.367   0.423   0.470

Factor Correlations:
        Factor1 Factor2 Factor3   Factor4 Factor5   Factor6
Factor1   1.000  0.3461 -0.3434  0.156711   0.331 -0.147071
Factor2   0.346  1.0000 -0.0627  0.041446   0.086 -0.483835
Factor3  -0.343 -0.0627  1.0000 -0.184243  -0.272 -0.286627
Factor4   0.157  0.0414 -0.1842  1.000000   0.064 -0.000521
Factor5   0.331  0.0860 -0.2724  0.064006   1.000  0.220545
Factor6  -0.147 -0.4838 -0.2866 -0.000521   0.221  1.000000

Test of the hypothesis that 6 factors are sufficient.
The chi square statistic is 809.11 on 165 degrees of freedom.
The p-value is 1.41e-85
```

Table 06: Factor model with Maximum Likelihood method and PROMAX rotation using factanal()

Code Snippet 05 in section 12.5 contains the R code for Maximum likelihood factor analysis with a PROMAX rotation.

8. Determining the correct number of factors

In this section, we iteratively analyze the factor models from k=1 to k=9. k=9 is the number of factors to retain that was obtained based on the percent of total variation above 90% rule. The goal was to iteratively evaluate the factor models with different values of k to identify the correct number of factors to retain. As part of the evaluation, we considered ease of factor interpretability, statistical inference results, and SS loadings values. All the factor models were obtained by using maximum likelihood analysis with the orthogonal VARIMAX rotation. For this analysis, we used the factor **loading cutoff value of |0.5|.**

- **Factor model with k=1**

With k=1, we have only one factor in the model. Table 07 shows the factor model obtained using ***factanal()*** R function.

- **Factor 1:** After applying the cutoff value, only A2, A5, and E2-E5 remain. A2 & A5 pertain to compassion, and variables E2-E5 pertain to gregariousness/enthusiasm. This led us to interpret the factor as the **social interaction** of an interaction.

The SS loadings value associated with the factor is above 1.

Regarding the statistical inference for the number of factors based on the chi-square test, based on the p-value (0), we reject the null hypothesis. Therefore, the one-factor model is not sufficient.

Call:
factanal(factors = k, covmat = cor.matrix, n.obs = 2236, rotation = "varimax")

Uniquenesses:

A1	A2	A3	A4	A5	C1	C2	C3	C4	C5	E1	E2	E3	E4	E5	N1	N2	N3
0.957	0.783	0.703	0.828	0.645	0.912	0.917	0.916	0.845	0.808	0.793	0.591	0.669	0.581	0.712	0.892	0.895	0.900

N4	N5	O1	O2	O3	O4	O5
0.769	0.912	0.901	0.978	0.849	0.993	0.972

Loadings:

	Factor1
A1	-0.207
A2	0.466
A3	0.545
A4	0.415
A5	0.595
C1	0.296
C2	0.288
C3	0.291
C4	-0.394
C5	-0.438
E1	-0.455
E2	-0.639
E3	0.575
E4	0.647
E5	0.537
N1	-0.329
N2	-0.324
N3	-0.316
N4	-0.481
N5	-0.297
O1	0.314
O2	-0.15
O3	0.388
O4	
O5	-0.168

	Factor1
SS loadings	**4.278**
Proportion Var	0.171

Test of the hypothesis that 1 factor is sufficient.
The chi square statistic is 9711.62 on 275 degrees of freedom.
The p-value is 0

Table 07: Factor model (k=1) with Maximum Likelihood method and VARIMAX rotation using factanal()

- **Factor model with k=2**

With k=2, we have two factors in the model. Table 08 shows the factor model obtained using ***factanal()*** R function with VARIMAX factor rotation.

- **Factor 1:** After applying the cutoff value, only A2-A3, and A5, E2-E5 remain. Variables A2-A3 pertain to compassion, and variables E2-E5 pertain to gregariousness/enthusiasm. This led us to interpret the factor **as social interaction**.
- **Factor 2:** The cutoff value for factors leaves N1-N5. Based on that, this factor can be interpreted as capturing **irritability and anxiety.**

The values of SS loadings associated with the two factors are above 1. Regarding the statistical inference for the number of factors based on the chi-square test, based on the p-value (0), we reject the null hypothesis. The two-factor model is not sufficient.

```
Call:
factanal(factors = k, covmat = cor.matrix, n.obs = 2236, rotation = "varimax")

Uniquenesses:
   A1    A2    A3    A4    A5    C1    C2    C3    C4    C5    E1    E2    E3    E4    E5    N1    N2    N3
0.953 0.740 0.650 0.825 0.628 0.911 0.906 0.921 0.843 0.799 0.761 0.599 0.586 0.564 0.643 0.339 0.370 0.473
   N4    N5    O1    O2    O3    O4    O5
0.607 0.734 0.886 0.965 0.810 0.962 0.972

Loadings:
```

	Factor1	Factor2
A1	-0.172	0.131
A2	**0.509**	
A3	**0.589**	
A4	0.401	-0.12
A5	**0.583**	-0.182
C1	0.294	
C2	0.307	
C3	0.262	-0.103
C4	-0.286	0.275
C5	-0.317	0.318
E1	-0.488	
E2	**-0.584**	0.246
E3	**0.643**	
E4	**0.638**	-0.169
E5	**0.598**	
N1		**0.813**
N2		**0.794**
N3		**0.725**
N4	-0.263	**0.57**
N5		**0.508**
O1	0.335	
O2		0.167
O3	0.436	
O4		0.195
O5	-0.151	

	Factor1	Factor2
SS loadings	**3.734**	**2.823**
Proportion Var	0.149	0.113
Cumulative Var	0.149	0.262

```
Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 5994.58 on 251 degrees of freedom.
The p-value is 0
```

Table 08: Factor model (k=2) with Maximum Likelihood method and VARIMAX rotation using factanal()

- **Factor model with k=3**

With k=3, we have three factors in the resultant factor model obtained using maximum likelihood factor analysis with VARIMAX factor rotation. Table 09 shows the factor model obtained using ***factanal()*** R function.

- **Factor 1:** After applying the cutoff value, only A2-A3, and A5, E1-E4 remain. Variables A2-A3 pertain to compassion, and variables E1-E4 pertain to gregariousness, enthusiasm. This led us to interpret the factor **as social interaction**.
- **Factor 2:** The cutoff value for factors leaves N1-N5. Based on that, this factor can be interpreted as capturing **irritability/anxiety.**
- **Factor 3:** The cutoff value leaves behind C1-C5 variables. This factor can be interpreted as **dutifulness to work**.

The values of SS loadings associated with the three factors are above 1. The chi-square test for the correct number of the factor is rejected (p-value is 0). The three-factor model is not sufficient.

```
Call:
factanal(factors = k, covmat = cor.matrix, n.obs = 2236, rotation = "varimax")

Uniquenesses:
   A1    A2    A3    A4    A5    C1    C2    C3    C4    C5    E1    E2    E3    E4    E5    N1    N2    N3
0.943 0.731 0.623 0.819 0.582 0.663 0.610 0.720 0.530 0.628 0.714 0.577 0.583 0.494 0.627 0.344 0.366 0.474
   N4    N5    O1    O2    O3    O4    O5
0.603 0.732 0.879 0.924 0.818 0.956 0.950
```

Loadings:

	Factor1	Factor2	Factor3
A1	-0.199	0.132	
A2	**0.507**		0.109
A3	**0.608**		
A4	0.389		0.147
A5	**0.623**	-0.161	
C1	0.107		**0.57**
C2	0.106		**0.609**
C3			**0.521**
C4		0.192	**-0.653**
C5	-0.163	0.249	**-0.532**
E1	**-0.533**		
E2	**-0.599**	0.223	-0.121
E3	**0.631**		0.136
E4	**0.693**	-0.149	
E5	0.497		0.348
N1		**0.803**	
N2		**0.793**	
N3		**0.717**	-0.104
N4	-0.252	**0.546**	-0.189
N5		0.494	-0.137
O1	0.255		0.236
O2		0.134	-0.242
O3	0.374		0.202
O4		0.206	
O5			-0.208

	Factor1	Factor2	Factor3
SS loadings	**3.325**	**2.633**	**2.154**
Proportion Var	0.133	0.105	0.086
Cumulative Var	0.133	0.238	0.324

Test of the hypothesis that 3 factors are sufficient.
The chi square statistic is 4116.25 on 228 degrees of freedom.
The p-value is 0

Table 09: Factor model (k=3) with Maximum Likelihood method and VARIMAX rotation using factanal()

- **Factor model with k=4**

With k=4, we have four factors in the resultant factor model. Table 10 shows the factor model obtained using ***factanal()*** R function with VARIMAX factor rotation.

- **Factor 1:** After applying the cutoff value, only A2-A3, and A5, E1-E4 remain. Variables A2-A3 pertain to compassion, and variables E1-E4 pertain to gregariousness, enthusiasm. This led us to interpret the factor **as social interaction**.
- **Factor 2:** The cutoff value for factors leaves N1-N5. Based on that, this factor can be interpreted as capturing **irritability and anxiety.**
- **Factor 3:** The cutoff value leaves behind C1-C5 variables. This factor can be interpreted as **dutifulness to work**.
- **Factor 4:** After applying the cut off value, we were left with O1, O3, and O5 for factor 4. We interpreted this factor as an individual's **intellectual curiosity**.

The SS loadings values associated with the 4 factors are above 1. The Ho of the chi-square test for the correct no. of the factor is rejected (p-value is 0). We conclude the 4-factor model is not sufficient.

```
Call:
factanal(factors = k, covmat = cor.matrix, n.obs = 2236, rotation = "varimax")

Uniquenesses:
  A1    A2    A3    A4.   A5    C1    C2    C3    C4    C5    E1    E2    E3    E4    E5    N1    N2    N3
0.946 0.721 0.610 0.742 0.575 0.673 0.607 0.681 0.509 0.577 0.721 0.582 0.531 0.462 0.627 0.346 0.373 0.471
  N4    N5    O1    O2    O3    O4    O5
0.591 0.697 0.678 0.715 0.511 0.867 0.713
Loadings:
```

	Factor1	Factor2	Factor3	Factor4
A1	-0.196	0.124		
A2	**0.509**		0.141	
A3	**0.615**		0.109	
A4	0.422		0.218	-0.167
A5	**0.631**	-0.143		
C1			**0.528**	0.202
C2			**0.607**	0.102
C3			**0.555**	
C4		0.225	**-0.654**	
C5	-0.172	0.267	**-0.567**	
E1	**-0.517**			-0.104
E2	**-0.59**	0.219	-0.106	-0.102
E3	**0.607**			0.308
E4	**0.716**	-0.127		
E5	0.464		0.309	0.246
N1		**0.805**		
N2		**0.787**		
N3		**0.723**		
N4	-0.269	**0.549**	-0.185	
N5		**0.516**		-0.173
O1	0.198		0.108	**0.52**
O2		0.18	-0.118	**-0.483**
O3	0.319			**0.62**
O4		0.191		0.301
O5				**-0.523**

	Factor1	Factor2	Factor3	Factor4
SS loadings	**3.263**	**2.670**	**1.989**	**1.553**
Proportion Var	0.131	0.107	0.080	0.062
Cumulative Var	0.131	0.237	0.317	0.379

Test of the hypothesis that 4 factors are sufficient.
The chi square statistic is 2631.66 on 206 degrees of freedom.
The p-value is 0

Table 10: Factor model (k=4) with Maximum Likelihood method and VARIMAX rotation using factanal()

- **Factor model with k=5**

With k=5, we have five factors in the resultant factor model. Table 11 shows the factor model obtained using ***factanal()*** R function with VARIMAX factor rotation.

- **Factor 1:** The cutoff value for factors leaves N1-N5. Based on that, this factor can be interpreted as capturing **irritability/anxiety.**
- **Factor 2:** After applying the cutoff value, only E1, E2, and E4 remain. Variables E1, E2, and E4 pertain to **social nature**.
- **Factor 3:** The cutoff value leaves behind C1-C5 variables. This factor can be interpreted as **dutifulness to work**.
- **Factor 4:** After applying the cut off value, we were left with A2, A3, and A5 for factor 4. We interpreted this factor as an individual's **compassion/altruism**.
- **Factor 5:** The cutoff threshold leaves O1, O3, and O5. This factor could be interpreted as **curiosity/introspection**.

The values of SS loadings associated with the five factors are above 1.

The chi-square test for the correct number of the factor is rejected (p-value is 1.88e-177, a very small value). At a level of significance of 0.05, the null hypothesis is rejected that the five-factor model is sufficient. So, statistically, the five-factor model is not sufficient.

Call:
factanal(factors = k, covmat = cor.matrix, n.obs = 2236, rotation = "varimax")

Uniquenesses:

A1	A2	A3	A4	A5	C1	C2	C3	C4	C5	E1.	E2	E3	E4	E5	N1	N2	N3
0.843	0.602	0.485	0.694	0.525	0.669	0.579	0.675	0.516	0.561	0.640	0.454	0.543	0.461	0.585	0.277	0.341	0.474

N4	N5	O1	O2	O3	O4	O5
0.502	0.657	0.676	0.725	0.516	0.758	0.714

Loadings:

	Factor1	Factor2	Factor3	Factor4	Factor5
A1				-0.375	
A2		0.195	0.143	**0.579**	
A3		0.28	0.113	**0.649**	
A4		0.172	0.226	0.453	-0.132
A5	-0.118	0.337		**0.581**	
C1			**0.528**		0.215
C2			**0.617**	0.137	0.125
C3			**0.556**	0.12	
C4	0.222		**-0.647**		
C5	0.266	-0.193	**-0.572**		
E1		**-0.578**		-0.139	
E2	0.227	**-0.675**	-0.1	-0.157	
E3		0.498		0.326	0.311
E4	-0.123	**0.602**		0.39	
E5		0.498	0.314	0.128	0.224
N1	**0.814**			-0.208	
N2	**0.783**			-0.203	
N3	**0.717**				
N4	**0.563**	-0.374	-0.191		
N5	**0.521**	-0.183		0.109	-0.15
O1		0.176	0.112		**0.523**
O2	0.173		-0.115	0.119	-0.467
O3		0.273		0.149	**0.619**
O4	0.211	-0.221		0.13	0.36
O5					**-0.524**

	Factor1	Factor2	Factor3	Factor4	Factor5
SS loadings	**2.685**	**2.305**	**2.011**	**1.952**	**1.574**
Proportion Var	0.107	0.092	0.080	0.078	0.063
Cumulative Var	0.107	0.200	0.280	0.358	0.421

Test of the hypothesis that 5 factors are sufficient.
The chi square statistic is 1357.5 on 185 degrees of freedom.
The p-value is 1.88e-177

Table 11: Factor model (k=5) with Maximum Likelihood method and VARIMAX rotation using factanal()

- **Factor model with k=6**

With k=6, we have six factors in the resultant factor model. Table 12 shows the factor model obtained using ***factanal()*** R function with VARIMAX factor rotation.

- **Factor 1:** After applying the cutoff value, only E1-E5 remain. Variables E1-E5 pertain to **introversion**.
- **Factor 2:** The cutoff value for factors leaves the variables N1-N5. Based on that, this factor can be interpreted as capturing the **irritability and anxiety** of an individual**.**

- **Factor 3:** The cutoff value leaves behind C1-C5 variables. This factor can be interpreted as **dutifulness to work**.
- **Factor 4:** After applying the cut off value, we were left with the variables A1, A2, and A3 for factor 4. We interpreted this factor as an individual's **compassion/altruism**.
- **Factor 5:** The cutoff threshold leaves the variables O1, O3, and O5. This factor could be interpreted as **curiosity/introspection**.
- **Factor 6:** The cutoff threshold leaves no variables. **Factor 6 provides no use.**

The values of the SS loadings associated with the five factors are above 1. But the SS loading value of factor 6 is less than 1. **We conclude factor 6 is not important**.

The chi-square test for the correct number of the factor is also rejected based on the p-value is 1.41e-85, a very small value. At a level of significance of 0.05, the null hypothesis is rejected that the six-factor model is sufficient.

```
Call:
factanal(factors = k, covmat = cor.matrix, n.obs = 2236, rotation = "varimax")

Uniquenesses:
   A1    A2    A3    A4    A5    C1    C2    C3    C4    C5    E1    E2    E3    E4    E5    N1    N2    N3
0.672 0.496 0.490 0.700 0.530 0.650 0.516 0.683 0.429 0.570 0.611 0.456 0.498 0.430 0.591 0.283 0.306 0.477
   N4    N5    O1    O2    O3    O4    O5
0.498 0.649 0.666 0.692 0.505 0.763 0.630
Loadings:
```

	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6
A1				**-0.534**	-0.113	0.124
A2	0.26		0.128	**0.645**		
A3	0.384		0.127	**0.568**		0.153
A4	0.239		0.236	0.387	-0.152	0.102
A5	0.446	-0.137	0.107	0.435		0.227
C1			**0.549**		0.188	
C2			**0.665**			0.158
C3			**0.551**			
C4		0.222	**-0.633**	-0.101	-0.12	0.305
C5	-0.184	0.273	**-0.548**			0.137
E1	**-0.575**			-0.133		0.178
E2	**-0.675**	0.233				0.124
E3	**0.594**		0.112	0.141	0.25	0.231
E4	**0.678**	-0.14	0.114	0.215	-0.108	0.14
E5	**0.515**		0.306		0.197	
N1		**0.815**		-0.162		-0.124
N2		**0.802**		-0.122		-0.186
N3		**0.714**				
N4	-0.342	**0.562**	-0.16			0.198
N5	-0.149	**0.516**			-0.165	0.164
O1	0.232		0.134		0.487	0.158
O2		0.168			**-0.5**	0.138
O3	0.343				**0.581**	0.178
O4	-0.163	0.21		0.125	0.348	0.17
O5					**-0.58**	0.148

	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6
SS loadings	**2.728**	**2.718**	**2.049**	**1.560**	**1.537**	0.617
Proportion Var	0.109	0.109	0.082	0.062	0.061	0.025
Cumulative Var	0.109	0.218	0.300	0.362	0.424	0.448

Test of the hypothesis that 6 factors are sufficient.
The chi square statistic is 809.11 on 165 degrees of freedom.
The p-value is 1.41e-85

Table 12: Factor model (k=6) with Maximum Likelihood method and VARIMAX rotation using factanal()

- **Factor model with k=7**

With k=7, we have seven factors in the resultant factor model. Table 13 shows the factor model obtained using ***factanal()*** R function with VARIMAX factor rotation.

- **Factor 1:** The cutoff value for factors leaves the variables N1-N5. Based on that, this factor can be interpreted as capturing **irritability and anxiety.**
- **Factor 2:** After applying the cut off value, we were left with the variables A2, A3, and A5 for factor 2. We interpreted this factor as an individual's **compassion/sympathy**.
- **Factor 3:** The cutoff value leaves behind the variables C1-C5. This factor can be interpreted as **thoroughness with work**.
- **Factor 4:** After applying the cutoff value, only the variables E1-E2 and E4 remain. The variables E1, E2, and E4 pertain to **introversion**
- **Factor 5:** Only the variables O1, O3, and O5 are left after applying the threshold. The resultant factor can be interpreted as dealing with **curious and intellectual nature** of the individual.
- **Factor 6 and Factor 7:** The cutoff threshold leaves no variables. As a result, they are not useful.

The SS loadings values associated with the five factors are above 1. But the SS loading value of factor 6 and Factor 7 are less than 1. **We conclude factors 6 and 7 are not important**.

Examining the chi-square test for the correct number of the factors, we reject the null hypothesis based on the p-value (4.78e-51, a very small value). At a level of significance of 0.05, the null hypothesis is rejected that the seven-factor model is sufficient.

Call:
factanal(factors = k, covmat = cor.matrix, n.obs = 2236, rotation = "varimax")

Uniquenesses:

A1	A2	A3	A4	A5	C1	C2	C3	C4	C5	E1	E2	E3	E4	E5	N1	N2	N3
0.663	0.500	0.453	0.700	0.528	0.646	0.518	0.684	0.412	0.571	0.536	0.435	0.501	0.409	0.571	0.265	0.303	0.433

N4	N5	O1	O2	O3	O4	O5
0.461	0.594	0.636	0.690	0.511	0.767	0.628

Loadings:

	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6	Factor7
A1		-0.386				0.41	
A2		**0.648**	0.108	0.133		-0.205	
A3		**0.705**		0.176			
A4		0.464	0.219	0.134	-0.115		
A5	-0.143	**0.588**		0.275		0.103	
C1			**0.548**		0.201		
C2		0.12	**0.654**		0.11	0.111	0.1
C3		0.139	**0.541**				
C4	0.205		**-0.66**	-0.119		0.274	0.117
C5	0.27		**-0.553**	-0.166			0.117
E1		-0.165		**-0.639**		0.134	
E2	0.214	-0.203	-0.105	**-0.676**			
E3		0.361		0.427	0.333	0.258	
E4	-0.124	0.409	0.113	**0.597**		0.171	
E5		0.231	0.297	0.393	0.249	0.131	-0.23
N1	**0.8**	-0.113				0.101	-0.251
N2	**0.782**	-0.116					-0.264
N3	**0.739**						0.127
N4	**0.581**		-0.166	-0.324			0.245
N5	**0.546**			-0.121	-0.166		0.246
O1		0.106	0.118		**0.537**	0.19	
O2	0.159	0.138	-0.11		-0.471	0.172	
O3		0.152		0.23	**0.623**	0.13	
O4	0.203			-0.218	0.349		0.11
O5					**-0.551**	0.239	

	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6	Factor7
SS loadings	**2.724**	**2.187**	**2.037**	**1.980**	**1.650**	0.595	0.411
Proportion Var	0.109	0.087	0.081	0.079	0.066	0.024	0.016
Cumulative Var	0.109	0.196	0.278	0.357	0.423	0.447	0.463

Test of the hypothesis that 7 factors are sufficient.
The chi square statistic is 567.75 on 146 degrees of freedom.
The p-value is 4.78e-51

Table 13: Factor model (k=7) with Maximum Likelihood method and VARIMAX rotation using factanal()

- **Factor model with k=8**

With k=8, we have eight factors in the resultant factor model. Table 14 shows the factor model obtained using ***factanal()*** R function with VARIMAX factor rotation.

- **Factor 1:** The cutoff value for factors leaves only N1-N5. Based on that, this factor can be interpreted as capturing **irritability and anxiety.**
- **Factor 2:** After applying the cut off value, we were left with variables A3, A5, and E4 for factor 2. We interpreted this factor as an individual's **compassion and social nature**.
- **Factor 3:** The cutoff value leaves behind the variables C1-C5. This factor can be interpreted as **thoroughness with work**.

- **Factor 4:** After applying the cutoff value, only E1-E2 and E4 remain. The variables E1, E2, and E4 pertain to **introversion.**
- **Factor 5:** Only O1, O2, O3, and O5 variables are left after applying the threshold. The factor can be interpreted as dealing with the **curious and intellectual nature** of the individual.
- **Factor 6:** Variables A1 and A2 meet the cutoff value. This factor could be interpreted as **empathy.**
- **Factor 7 and Factor 8:** The cutoff threshold leaves no variables. **As a result, they are not useful.**

The values of SS loadings associated with the five factors are above 1. But the SS loading values of factor 6, factor 7 and factor 8 are less than 1. **We conclude factors 6, 7, and 8 are not important**.

Examining the chi-square test for the correct number of the factor, we reject the Ho. The p-value is 3.71e-31, a very small value. At a level of significance of 0.05, the null hypothesis is rejected that the eight-factor model is sufficient.

```
Call:
factanal(factors = k, covmat = cor.matrix, n.obs = 2236, rotation = "varimax")

Uniquenesses:
   A1    A2    A3    A4    A5    C1    C2    C3    C4    C5    E1    E2    E3    E4    E5    N1    N2    N3
0.694 0.385 0.428 0.688 0.518 0.632 0.516 0.669 0.387 0.544 0.512 0.444 0.496 0.405 0.496 0.236 0.325 0.431
   N4    N5    O1    O2    O3    O4    O5
0.470 0.591 0.635 0.667 0.514 0.767 0.622
 Loadings:
```

	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6	Factor7	Factor8
A1		-0.177				**0.504**		
A2		0.432	0.142	-0.162		**-0.573**		0.215
A3		**0.685**		-0.116		-0.261		
A4		0.46	0.21		-0.118	-0.162		
A5	-0.138	**0.619**		-0.22		-0.125		
C1			**0.562**		0.187			
C2		0.142	**0.662**					
C3			**0.552**					0.103
C4	0.201		**-0.643**	0.122	-0.119	0.162	0.306	
C5	0.267	-0.147	**-0.525**	0.122			0.268	
E1		-0.141		**0.663**		0.136		
E2	0.216	-0.273		**0.638**				
E3		0.486		-0.363	0.317	0.108		0.102
E4	-0.126	**0.509**	0.111	**-0.546**				
E5		0.216	0.314	-0.398	0.214			0.392
N1	**0.813**					0.15	-0.182	0.182
N2	**0.776**	-0.136					-0.129	0.185
N3	**0.739**							-0.101
N4	**0.578**	-0.106	-0.141	0.304			0.229	-0.129
N5	**0.542**				-0.18		0.218	-0.146
O1		0.158	0.124		**0.512**			0.198
O2	0.149	0.12			**-0.507**		0.141	
O3		0.226		-0.199	**0.608**		0.107	
O4	0.204			0.196	0.328	-0.107	0.181	
O5					**-0.569**	0.175		

	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6	Factor7	Factor8
SS loadings	**2.718.**	**2.104**	**2.019**	**1.758**	**1.616**	0.844	0.454	0.417
Proportion Var	0.109	0.084	0.081	0.070	0.065	0.034	0.018	0.017
Cumulative Var	0.109	0.193	0.274	0.344	0.409	0.442	0.460	0.477

```
Test of the hypothesis that 8 factors are sufficient.
The chi square statistic is 409.35 on 128 degrees of freedom.
The p-value is 3.71e-31
```

Table 14: Factor model (k=8) with Maximum Likelihood method and VARIMAX rotation using factanal()

- **Factor model with k=9**

With k=9, we have nine factors in the resultant factor model. Table 15 shows the factor model obtained using ***factanal()*** R function with VARIMAX factor rotation.

- **Factor 1:** The cutoff value for factors leaves the variables N1-N5. Based on that, this factor can be interpreted as capturing **irritability and Anxiety.**
- **Factor 2:** After applying the cut off value, we were left with the variables A3, A5, and E4 for factor 2. We interpreted this factor as an individual's **compassion and social nature**.
- **Factor 3:** The cutoff value leaves behind the variables C1-C5. This factor can be interpreted as **thoroughness with work**.
- **Factor 4:** After applying the cutoff value, only variables E1-E2 and E4 remain. E1, E2, and E4 pertain to **introversion**
- **Factor 5:** Only O1, O2, and O5 variables are left after applying the threshold. The factor can be interpreted as dealing with the **curious and intellectual nature** of the individual.
- **Factor 6:** Only the variables A1 and A2 meet the cutoff value. This factor could be interpreted as **empathy.**
- **Factor 7, Factor 8, and Factor 9:** The cutoff threshold leaves no variables. As a result, they are not useful.

The values of SS loadings associated with the five factors are above 1. But the SS loading value of factor 6, factor 7, factor 8, and factor 9 are less than 1. **We conclude factors 6, 7, 8, and 9 are not important**.

The chi-square test for the correct number of the factor is rejected (p-value is 4.99e-19, a very small value). At a level of significance of 0.05, the null hypothesis is rejected that the nine-factor model is sufficient.

In Summary,

k = 5 provides the factor model that is easiest to interpret. Based on the cutoff value of |0.5|, the interpretation of the factors we arrived at are:

- **Factor 1: irritability and anxiety.**
- **Factor 2: social nature**.
- **Factor 3: dutifulness to work**.
- **Factor 4: compassion/altruism**.
- **Factor 5: curiosity/introspection**.

However, analyzing the chi-square test results for all the models, none of the models are statistically significant, meaning the chi-square test for the correct number of factors is NOT statistically significant for any of k in the range [1,9]. So, based on that, we concluded that none of the factor models yield good fit to the data. However, it is also important to that the chi-square test is not a particularly good test. Even if the factor model is reasonably close to model the population, the chi-square test is rejected because it is not a perfect model (null hypothesis). Additionally, the chi-square test is sensitive to sample size. Small deviations in sample size can result in the test to reject the model .

Code Snippet 06 in section 12.6 contains the R code for generating k=1 to k=9 factor models using maximum likelihood factor analysis with VARIMAX orthogonal rotation.

Call:
factanal(factors = k, covmat = cor.matrix, n.obs = 2236, rotation = "varimax")

Uniquenesses:

```
  A1    A2    A3    A4    A5    C1    C2    C3    C4    C5    E1    E2    E3    E4    E5    N1    N2    N3
0.694 0.374 0.444 0.686 0.495 0.586 0.522 0.670 0.399 0.501 0.535 0.420 0.485 0.366 0.490 0.306 0.193 0.403
  N4    N5    O1    O2    O3    O4    O5
0.443 0.607 0.630 0.669 0.518 0.757 0.611
```

Loadings:

	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6	Factor7	Factor8	Factor9
A1		-0.155				**0.508**			
A2		0.397	0.143	0.168		**-0.606**	0.215		
A3		**0.657**		0.124		-0.294			
A4		0.44	0.206		-0.122	-0.192		-0.104	
A5	-0.138	**0.635**		0.208		-0.146			
C1			**0.588**		0.184			0.158	
C2		0.138	**0.66**						
C3			**0.548**				0.111		
C4	0.215		**-0.629**	-0.121	-0.136	0.15	0.131	0.273	0.104
C5	0.267	-0.143	**-0.513**	-0.119				0.361	
E1		-0.147		**-0.644**		0.13			
E2	0.215	-0.256		**-0.656**				0.125	
E3		0.493		0.355	0.304		0.183		
E4	-0.129	**0.528**	0.119	**0.553**				0.108	
E5		0.204	0.31	0.402	0.195		0.404		
N1	**0.763**					0.144	0.14		-0.213
N2	**0.762**	-0.125							-0.442
N3	**0.765**								
N4	**0.615**	-0.121	-0.14	-0.292				0.123	0.199
N5	**0.554**			-0.104	-0.18			0.134	0.117
O1		0.161	0.12		0.497		0.254		
O2	0.15	0.118			**-0.514**			0.126	
O3		0.233		0.196	**0.594**		0.124		
O4	0.207			-0.199	0.326			0.209	
O5					**-0.582**	0.158			

	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6	Factor7	Factor8	Factor9
SS loadings	**2.723**	**2.051**	**2.005**	**1.752**	**1.592**	0.902	0.426	0.398	0.348
Proportion Var	0.109	0.082	0.080	0.070	0.064	0.036	0.017	0.016	0.014
Cumulative Var	0.109	0.191	0.271	0.341	0.405	0.441	0.458	0.474	0.488

Test of the hypothesis that 9 factors are sufficient.
The chi square statistic is 297.95 on 111 degrees of freedom.
The p-value is 4.99e-19

Table 15: Factor model (k=9) with Maximum Likelihood method and VARIMAX rotation using factanal()

9. Comparison of the factors of k=5 factor model with the BFI researcher's factors

The BFI (Big Five Inventory) researchers have identified five factors to measure the personality (the latent trait). The five factors are:

a) Agreeableness (A)
b) Conscientiousness (C)
c) Extraversion (E)
d) Neuroticism (N)
e) Openness (O)

For the k=5 factor model from the previous section which has the best interpretability, our interpretation of the five factors are

- **Factor 1: Irritability and anxiety.**
- **Factor 2: Social nature**.
- **Factor 3: Dutifulness to work**.
- **Factor 4: Compassion/Altruism**.
- **Factor 5: Curiosity/introspection**.

Comparing the two, we determined that:
Factor 1 is similar to Neuroticism (N),
Factor 2 is similar to Extraversion (E),
Factor 3 is similar to Conscientiousness (C),
Factor 4 is similar to Agreeableness (A),
Factor 5 is similar to Openness (O)

10. Personality differences among gender, education, and age

In this section, we present the results of the analysis conducted using the factor scores. The factor scores were obtained using the five-factor model (the model that is determined to be the best model in section 8). We added the factor scores back to the BFI data, which also has the demographic variables (age, gender, education). Next, considering each factor score as a response variable, we analyzed its relationship with the demographic variables.

Analysis of the five personality traits versus gender:

Treating the personality traits as response variables, we plotted them against the gender (1 = Male, 2 = Female) variable. FIG 03 shows the plots with red indicating the scores of the male subjects and green showing the scores of the female subjects.

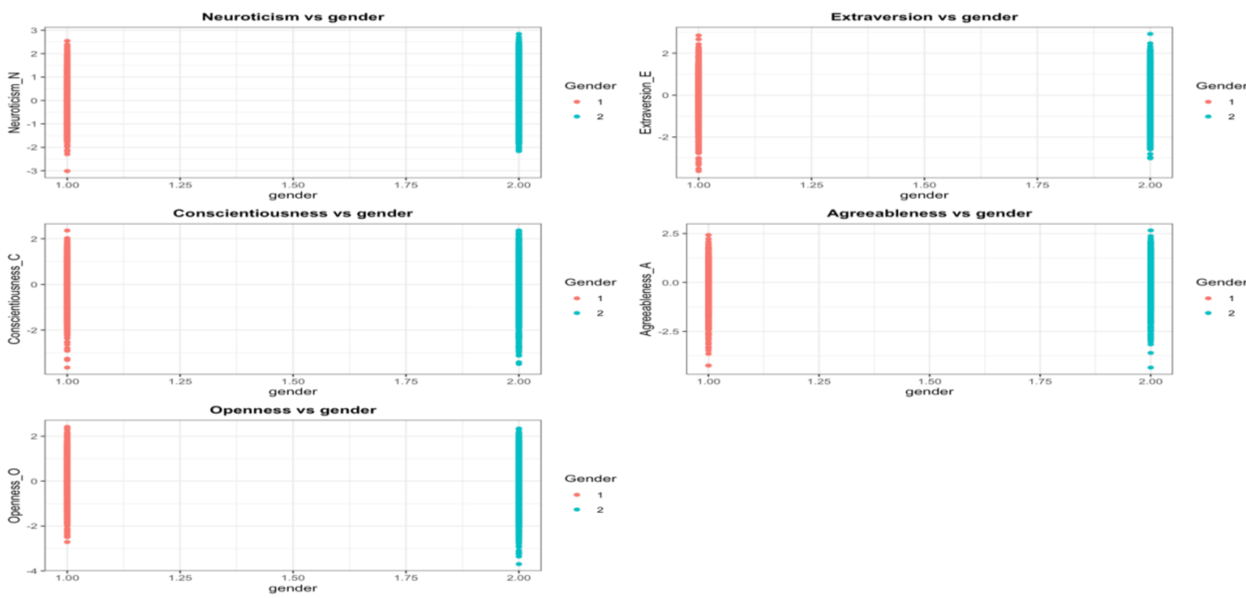

FIG 03: Personality traits versus gender

From the basic EDA, we noted we have 735 records from male subjects and 1501 records from female subjects. Though we have more than twice the number of observations for females than males, only slight variability is noticed in the scores of the personality traits. The Openness trait has some noticeable differences in the lower part of the scores obtained for males compared to females. Neuroticism and Extraversion also show the female scores start off slightly higher than the male subject scores.

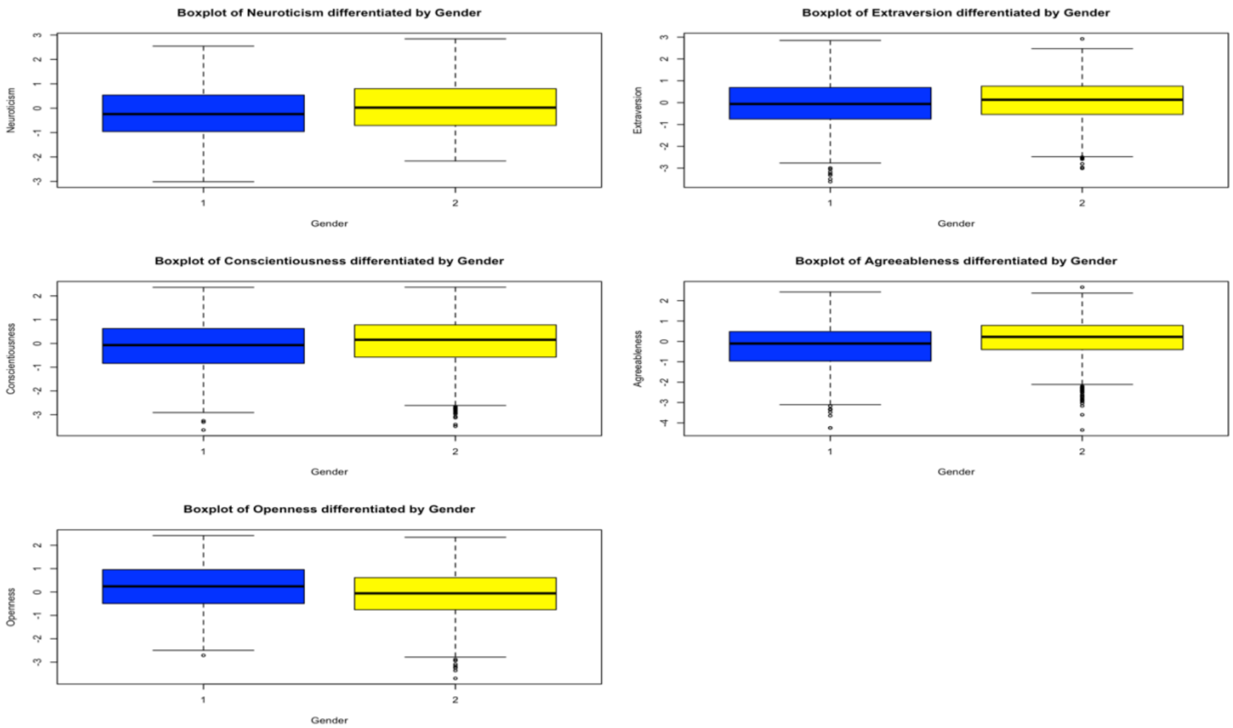

FIG 04: Boxplots of personality traits differentiated by gender

From the boxplots shown in FIG 04, we noticed that the median values of female subjects were higher than those of the male subjects for four out of the five personality traits – Neuroticism, Extraversion, Conscientiousness, Agreeableness. From the data, we noted **that male subjects have a higher median value for only Openness than the female subjects**. This may go along with the observation we made earlier about the plots that the scores for Openness for male subjects start off higher than that of the female subjects**. But, in all the cases, the male and female boxes overlap considerably, indicating there is a large overlap in the personality traits among males and females**. However, outliers are present in the case of Extraversion, Conscientiousness, Agreeableness, and Openness.

<u>Analysis of the five personality traits versus education:</u>

Next, treating the personality traits as response variables, we plotted them against the education variable (1 = HS, 2 = finished HS, 3 = some college, 4 = college graduate 5 = graduate degree). FIG 05 shows the plots.

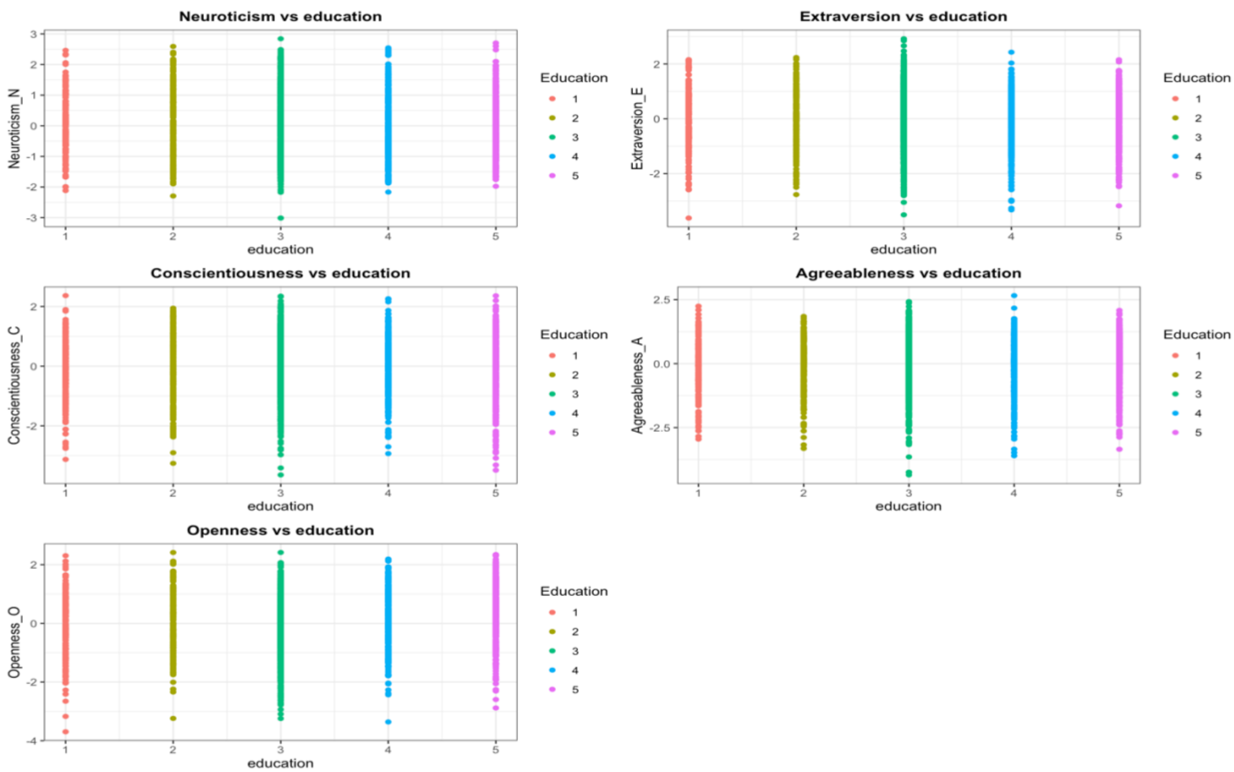

FIG 05: Personality traits versus education

From Table 16, we note that subjects with some college have the highest count followed by graduate degree and college graduate.

1 = In High School	2 = Finished High School	3 = Some College	4 = College Graduate	5 = Graduate Degree
198	250	1078	346	364

Table 16: Table of counts of subjects broken down by education

In FIG 05, we noted the largest spread of scores for the value 3 (some college). This is true for all the personality traits (Neuroticism, Extraversion, Conscientiousness, Agreeableness, and Openness). This may be due to the presence of most subjects with some college education in the data. As a result, the spread is also wider. Subjects with graduate degree also have a wider spread of scores for Conscientiousness.

From the boxplots shown in FIG 06, we can determine that the median values of all the education levels are roughly about the same with slightly higher values for “some college” (value 3) in the case of Conscientiousness and Agreeableness. Boxplots also show a median value, which is slightly higher than the rest for subjects with graduate degree (value 5) in the case of Openness. Subjects with “some college” (value 3) have a lower median value than the rest in the case of Openness. **But in all the cases, the education level boxes overlap considerably, indicating there is a large overlap in personality traits among the education levels too**. Like with gender, outliers are present in the case of Extraversion, Conscientiousness, Agreeableness, and Openness.

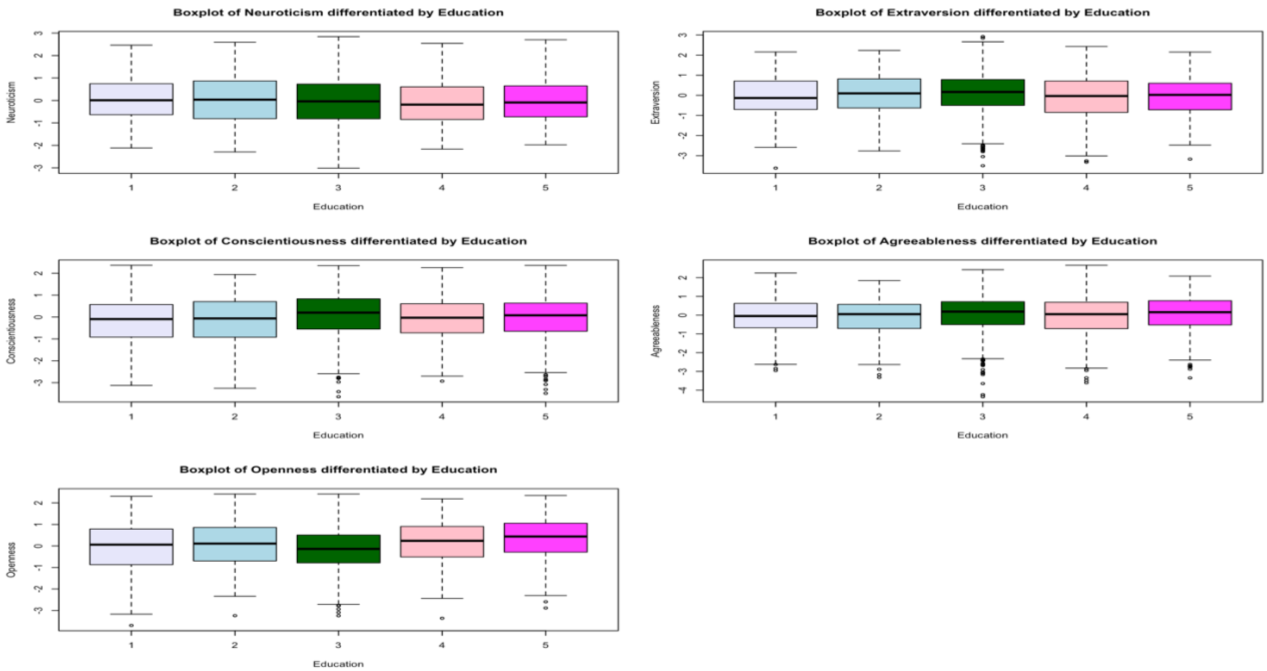

FIG 06: Boxplots of personality traits differentiated by education

Analysis of the five personality traits versus age:

Lastly, again, treating the personality traits as response variables, we plotted them against the age variable. FIG 07 shows the obtained plots.

From the basic EDA, we noted the age of the subjects ranges from age 3 years to 86 years. But ages less than 16 years and more than 56 years have records whose counts are in single digits (<10). The plots show most of the clustering from 16 years to roughly 32 years, indicating a large number of subjects belong to that age group.

Neuroticism shows the same variability or spread in the scores as the age increases until about age 45. After that, the scores for Neuroticism gradually become sparse as the age increases.

Similar clustering of values at lower ages (from 16 years to mid-30 years) is also noted with the other personality traits (Extraversion, Conscientiousness, Agreeableness, Openness). However, the decrease in variability of the scores is more pronounced and discernible in the case of Extraversion, Conscientiousness, Agreeableness, and Openness than in the case of Neuroticism.

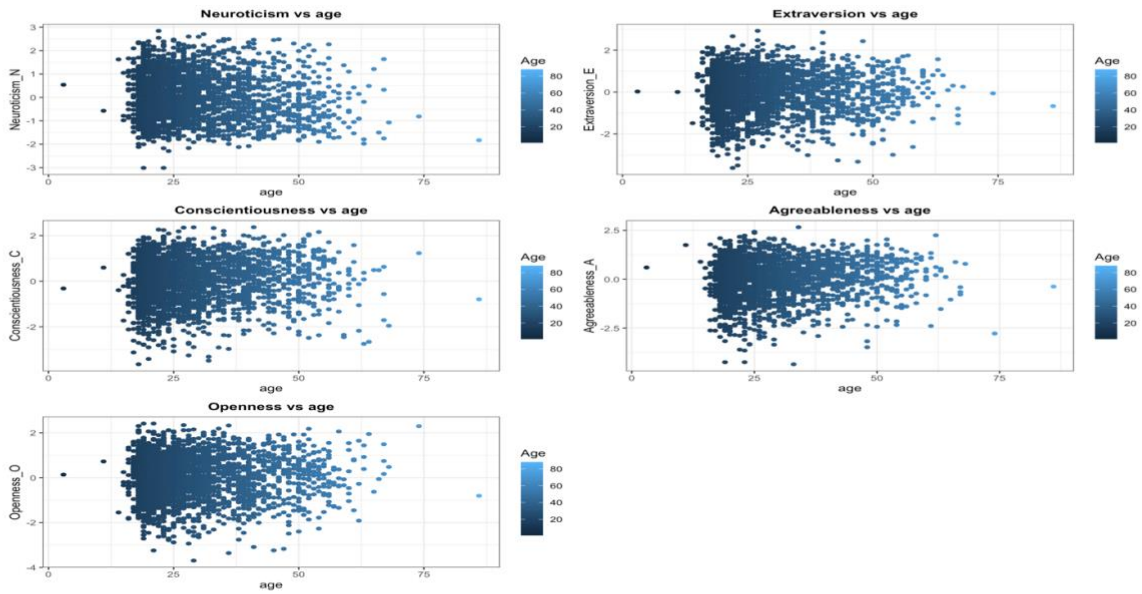

FIG 07: Personality traits versus age

From the boxplots shown in FIG 08, we can notice that **, the age boxes (with the medians) overlap substantially indicating there is large overlap in personality traits among the different ages.** Outliers are present in the case of Extraversion, Conscientiousness, Agreeableness, and Openness. Neuroticism has the fewest outliers. **For each personality trait, as the age increases, though the median values slightly fluctuate either up or down, the overall trend for each trait has been generally the same, possibly indicating that the personality of individuals only slightly varies with age.**

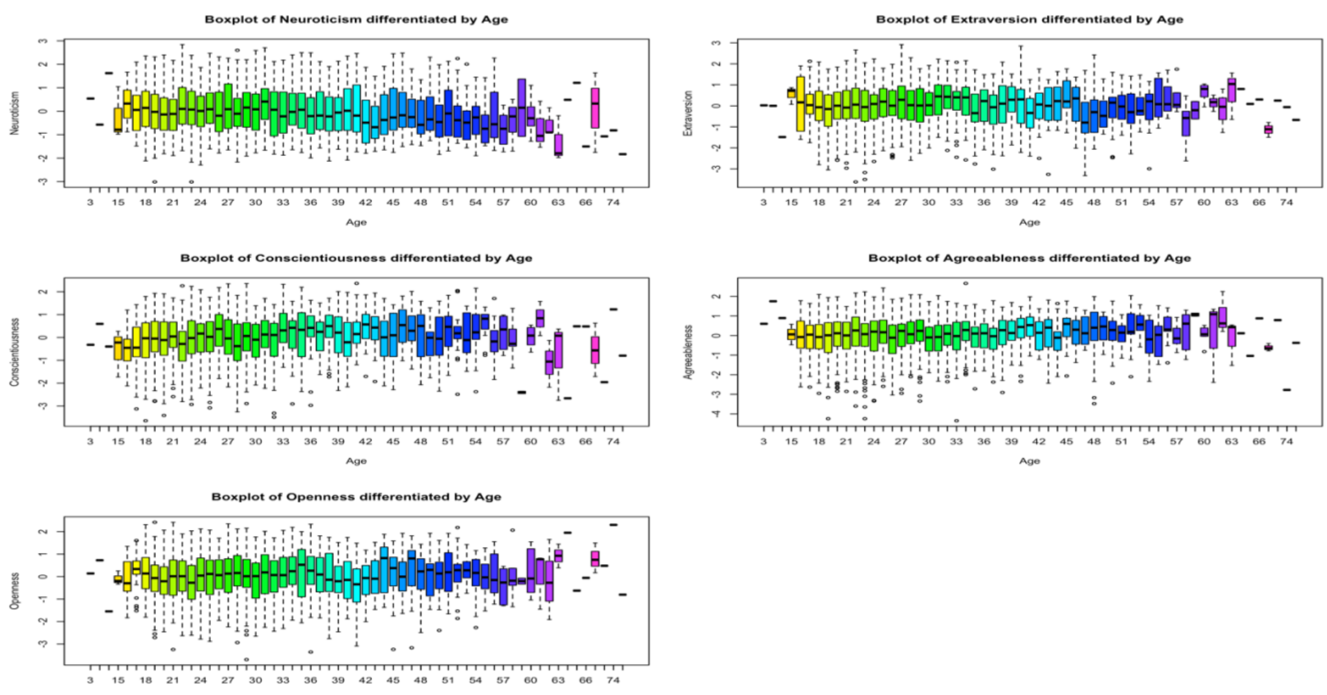

FIG 08: Boxplots of personality traits differentiated by age

Code Snippet 07 in section 12.7 contains the R code for study of the personality traits in relation to the demographic attributes.

11.Summary and Reflection

For this assignment to explore the Exploratory Factor Analysis (EFA) method, we used the BFI (Big Five Inventory) dataset from the International Personality Item Pool (ipip.ori.org) as part of the Synthetic Aperture Personality Assessment (SAPA) web based personality assessment project. The dataset has 2800 records with 25 personality variables and 3 demographic variables. Our basic exploration has shown that the dataset has some records with missing values. So, for our analysis, we removed the missing values. This has brought down the number of records in our dataset down to 2236 records.

We then proceeded to construct a correlation matrix from the reduced dataset. From the correlation plot obtained using the correlation matrix, we noted several constructs/patterns among the variables. We then obtained the eigenvalues and eigenvectors of the correlation matrix. From the eigenvalues, we determined the number of factors to retain using three rules – 1) Eigenvalue greater-than-one rule, 2) Cattell's Scree test, 3) Percent of Total Variance greater than 90%. The first rule gave us 6 factors, the second rule gave us 6 factors, but the third rule gave us 9 factors to retain. So, we learned that it is possible to obtain a different number of factors based on different rules, and we would need to select the rule that makes the most sense for the analysis.

Next, we used the Maximum Likelihood Factor Analysis method with orthogonal VARIMAX rotation to estimate a factor model with 6 factors. We used a cutoff value of |0.5| on the factor loadings. We examined the SS loadings values for the 6 factors and provided interpretation for the factors. We also studied the chi-square test to understand the null and alternative hypotheses. However, for the resultant VARIMAX model, the null hypothesis was rejected concluding that the factor model is not sufficient to describe the data.

Then, we conducted the Maximum Likelihood Factor Analysis with oblique PROMAX rotation to again estimate a factor model with 6 factors. We again used a cutoff value of |0.5| on the factor loadings and provided interpretation for the factors. However, it proved to be harder to provide the interpretability compared to the VARIMAX factor model. The chi-square test again resulted in the null hypothesis rejection. The resultant PROMAX factor model does not predict the data well.

We then iteratively evaluated factor models from k=1 to k=9 using the Maximum Likelihood Factor Analysis method with orthogonal VARIMAX rotation. By comparing the factor interpretability, SS loadings values, and chi-square test results, we determined k=5 yields the best factor model. One point to note is that the chi-square test failed for all the 9 models. We suspect this could be because of the large sample size since the test is sensitive to sample size. With a large sample, there is more chance of deviations, and finding a perfect model for the data is difficult. We also determined that the five factors from our best model align with the five factors (personality traits) put forth by the BFI researchers, namely Agreeableness, Conscientiousness, Extraversion, Neuroticism, and Openness.

Lastly, we obtained the factor scores for the best model we determined earlier. Using these five-set of factor scores as response variables, we studied them in relation to the demographic variables – age, gender, and education. Though there is some variability observed in the data between the two genders (there are more records for female subjects than for male subjects), overall, there is a large overlap in the personality traits between males and females. Similarly, among the education level, subjects with some college education also have a wider spread of scores for all the traits because their records made up the majority of the data. Subjects with graduate degrees also have a wider spread of scores for Conscientiousness. But in spite of this, overall, there is substantial overlap among all the education levels indicating the personality traits do not vary much by education. Finally, the analysis of personality traits by age also generally shows that there is only a slight variation in traits as age increases.

12.Code

12.1. Data preparation

```
library(psych)
bfi_data=bfi
bfi_data
library(dplyr)

# obtain the dimensions of the data
dim(bfi_data)
# Basic exploration of the data
str(bfi_data)
head(bfi_data)
# Check to determine if there are records with missing values
is.na(bfi_data)

# Remove rows with missing values and keep only complete cases
bfi_data=bfi_data[complete.cases(bfi_data),]
dim(bfi_data)
```

R code snippet 01: Data preparation

12.2. Exploratory Data Analysis and Correlation Plot

```
# Frequency table for gender
table(bfi_data$gender)
# Frequency table for education
table(bfi_data$education)
# Frequency table and summary for age
table(bfi_data$age)
summary(bfi_data$age)

# Remove the demographic attributes
efa_data <- bfi_data %>% select(-gender, -education, -age)
# check the dimensions
dim(efa_data)
# range of the personality variables
range(efa_data)
# check the structure
str(efa_data)

# Compute correlation matrix for returns;
cor.data <- cor(efa_data)
# Convert it to a matrix
cor.matrix <- as.matrix(cor.data)

# basic validations that the matrix is symmetric
is.matrix(cor.matrix)
isSymmetric(cor.matrix)

# Check the dimensios of the matrix
dim(cor.matrix)

# load the corrplot package
library(corrplot)
# Make correlation plot
corrplot(cor.matrix)
#corrplot(cor.data,method="number",number.cex=0.75)
```

R code snippet 02: Exploratory Data Analysis and Correlation Plot

12.3. Eigenvalue, eigenvector computation, and scree plot

```
# Compute the eigen values and eigen vectors for cor.matrix
Z<-eigen(cor.matrix)
Z$values
Z$vec

# Plot the scree plot generated using cor.matrix with both the PCA and FA methods - uses default "minRes" factor method
fa.parallel(cor.matrix, n.obs=2236, fa="both", n.iter=100, show.legend=TRUE,main="Scree plot with parallel analysis")

# Compute the cumulative variance and cumulative proportion - uses default "minRes" factor method
fa1 <- fa(r = cor.matrix, nfactors = 25, rotate="none")
fa1
```

R code snippet 03: Eigenvalues, Eigenvectors, and scree plot

12.4. Maximum likelihood factor analysis with a VARIMAX rotation

```
factors_ml_varimax1 <- fa(r = cor.matrix, nfactors = 6, fm="ml", rotate="varimax")
factors_ml_varimax1
factors_ml_varimax2 <- factanal(covmat=cor.matrix, n.obs=2236, factors=6, rotation='varimax');
factors_ml_varimax2
```

R code snippet 04: ML factor analysis for 6-factor model with VARIMAX rotation

12.5. Maximum likelihood factor analysis with a PROMAX rotation

```
library(GPArotation)
factors_ml_promax1 <- factanal(covmat=cor.matrix, n.obs=2236, factors=6, rotation='promax');
factors_ml_promax1
```

R code snippet 05: ML factor analysis for 6-factor model with PROMAX rotation

12.6. Generate k=1 to k=9 factor models using Maximum Likelihood factor analysis with VARIMAX orthogonal rotation

```
function_ml_varimax <- function (k=9) {
        factors_ml_varimax <- factanal(covmat=cor.matrix, n.obs=2236, factors=k, rotation='varimax');
        factors_ml_varimax
 }
k <- seq(1:9)
sapply(k,function_ml_varimax)
```

R code snippet 06: code for generating k=1 to k=9 factor models using Maximum Likelihood factor analysis with VARIMAX orthogonal rotation

12.7. Analysis of the personality traits in relation to the demographic attributes

```
library(ggplot2)
library(gridExtra)
# obtain the factor scores based on the five factor model
f <- factanal(covmat=cor.matrix, n.obs=2236, factors=5,rotation='varimax')
fs <- factor.scores(efa_data, f)
efa_data <- cbind(efa_data, fs$scores)

# Append the demographic attributes to the factor scores in efa_data
# create a new data frame bfi.dat
bfi.dat <- cbind(efa_data, bfi_data['age'])
bfi.dat <- cbind(bfi.dat, bfi_data['education'])
bfi.dat <- cbind(bfi.dat, bfi_data['gender'])

# Rename the factor scores columns to N, E, C, A, O
colnames(bfi.dat)[which(names(bfi.dat) == "Factor1")] <- "Neuroticism_N"
colnames(bfi.dat)[which(names(bfi.dat) == "Factor2")] <- "Extraversion_E"
colnames(bfi.dat)[which(names(bfi.dat) == "Factor3")] <- "Conscientiousness_C"
colnames(bfi.dat)[which(names(bfi.dat) == "Factor4")] <- "Agreeableness_A"
colnames(bfi.dat)[which(names(bfi.dat) == "Factor5")] <- "Openness_O"

# obtain the structure and dimension details of bfi.dat
str(bfi.dat)
dim(bfi.dat)

# scatter plots of the personality traits versus gender
# Note these scatter plots have considerable time to draw;
p1 <- ggplot(bfi.dat, aes(x=gender, y=Neuroticism_N, color=factor(gender))) + geom_point() + labs(title="Neuroticism vs gender") +
theme_bw()+theme(plot.title = element_text(hjust = 0.5, size=12,face="bold")) +labs(color= "Gender")
p2 <- ggplot(bfi.dat, aes(x=gender, y=Extraversion_E, color=factor(gender))) + geom_point() + labs(title="Extraversion vs gender") +
theme_bw()+theme(plot.title = element_text(hjust = 0.5, size=12,face="bold")) +labs(color= "Gender")
p3 <- ggplot(bfi.dat, aes(x=gender, y=Conscientiousness_C, color=factor(gender))) + geom_point() + labs(title="Conscientiousness vs
gender") + theme_bw()+theme(plot.title = element_text(hjust = 0.5, size=12,face="bold")) +labs(color= "Gender")
p4 <- ggplot(bfi.dat, aes(x=gender, y=Agreeableness_A, color=factor(gender))) + geom_point() + labs(title="Agreeableness vs gender")
+ theme_bw()+theme(plot.title = element_text(hjust = 0.5, size=12,face="bold")) +labs(color= "Gender")
p5 <- ggplot(bfi.dat, aes(x=gender, y=Openness_O, color=factor(gender))) + geom_point() + labs(title="Openness vs gender") +
theme_bw()+theme(plot.title = element_text(hjust = 0.5, size=12,face="bold"))+labs(color= "Gender")
grid.arrange(p1,p2,p3,p4,p5,nrow=3,ncol=2)

# side-by-side boxplots of personality traits differentiated by gender
par(mfrow=c(3,2))
boxplot(Neuroticism_N ~ gender, data=bfi.dat, xlab="Gender", ylab="Neuroticism",col=c("blue","yellow"))
title("Boxplot of Neuroticism differentiated by Gender")
boxplot(Extraversion_E ~ gender, data=bfi.dat, xlab="Gender", ylab="Extraversion",col=c("blue","yellow"))
title("Boxplot of Extraversion differentiated by Gender")
boxplot(Conscientiousness_C ~ gender, data=bfi.dat, xlab="Gender", ylab="Conscientiousness",col=c("blue","yellow"))
title("Boxplot of Conscientiousness differentiated by Gender")
boxplot(Agreeableness_A ~ gender, data=bfi.dat, xlab="Gender", ylab="Agreeableness",col=c("blue","yellow"))
title("Boxplot of Agreeableness differentiated by Gender")
boxplot(Openness_O ~ gender, data=bfi.dat, xlab="Gender", ylab="Openness",col=c("blue","yellow"))
title("Boxplot of Openness differentiated by Gender")

# scatter plots of the personality triats versus education
# Note these scatter plots have considerable time to draw;
p1 <- ggplot(bfi.dat, aes(x=education, y=Neuroticism_N, color=factor(education))) + geom_point() + labs(title="Neuroticism vs
education") +
theme_bw()+theme(plot.title = element_text(hjust = 0.5, size=12,face="bold")) +labs(color= "Education")
p2 <- ggplot(bfi.dat, aes(x=education, y=Extraversion_E, color=factor(education))) + geom_point() + labs(title="Extraversion vs
education") +
theme_bw()+theme(plot.title = element_text(hjust = 0.5, size=12,face="bold"))+labs(color= "Education")
```

```r
p3 <- ggplot(bfi.dat, aes(x=education, y=Conscientiousness_C, color=factor(education))) + geom_point() + labs(title="Conscientiousness vs education") +
theme_bw()+theme(plot.title = element_text(hjust = 0.5, size=12,face="bold"))+labs(color= "Education")
p4 <- ggplot(bfi.dat, aes(x=education, y=Agreeableness_A, color=factor(education))) + geom_point() + labs(title="Agreeableness vs education") +
theme_bw()+theme(plot.title = element_text(hjust = 0.5, size=12,face="bold"))+labs(color= "Education")
p5 <- ggplot(bfi.dat, aes(x=education, y=Openness_O, color=factor(education))) + geom_point() + labs(title="Openness vs education") +
theme_bw()+theme(plot.title = element_text(hjust = 0.5, size=12,face="bold"))+labs(color= "Education")
grid.arrange(p1,p2,p3,p4,p5,nrow=3,ncol=2)

# side-by-side boxplots of personality traits differentiated by education
par(mfrow=c(3,2))
boxplot(Neuroticism_N ~ education, data=bfi.dat, xlab="Education",
ylab="Neuroticism",col=c("Lavender","lightblue","darkgreen","pink","magenta"))
title("Boxplot of Neuroticism differentiated by Education")
boxplot(Extraversion_E ~ education, data=bfi.dat, xlab="Education",
ylab="Extraversion",col=c("Lavender","lightblue","darkgreen","pink","magenta"))
title("Boxplot of Extraversion differentiated by Education")
boxplot(Conscientiousness_C ~ education, data=bfi.dat, xlab="Education",
ylab="Conscientiousness",col=c("Lavender","lightblue","darkgreen","pink","magenta"))
title("Boxplot of Conscientiousness differentiated by Education")
boxplot(Agreeableness_A ~ education, data=bfi.dat, xlab="Education",
ylab="Agreeableness",col=c("Lavender","lightblue","darkgreen","pink","magenta"))
title("Boxplot of Agreeableness differentiated by Education")
boxplot(Openness_O ~ education, data=bfi.dat, xlab="Education",
ylab="Openness",col=c("Lavender","lightblue","darkgreen","pink","magenta"))
title("Boxplot of Openness differentiated by Education")

# scatter plots of the personality triats versus age
# Note these scatter plots have considerable time to draw;
p1 <- ggplot(bfi.dat, aes(x=age, y=Neuroticism_N, color=age)) + geom_point() + labs(title="Neuroticism vs age") +
theme_bw()+theme(plot.title = element_text(hjust = 0.5, size=12,face="bold")) +labs(color= "Age")
p2 <- ggplot(bfi.dat, aes(x=age, y=Extraversion_E, color=age)) + geom_point() + labs(title="Extraversion vs age") +
theme_bw()+theme(plot.title = element_text(hjust = 0.5, size=12,face="bold")) +labs(color= "Age")
p3 <- ggplot(bfi.dat, aes(x=age, y=Conscientiousness_C, color=age)) + geom_point() + labs(title="Conscientiousness vs age") +
theme_bw()+theme(plot.title = element_text(hjust = 0.5, size=12,face="bold")) +labs(color= "Age")
p4 <- ggplot(bfi.dat, aes(x=age, y=Agreeableness_A, color=age)) + geom_point() + labs(title="Agreeableness vs age") +
theme_bw()+theme(plot.title = element_text(hjust = 0.5, size=12,face="bold")) +labs(color= "Age")
p5 <- ggplot(bfi.dat, aes(x=age, y=Openness_O, color=age)) + geom_point() + labs(title="Openness vs age") +
theme_bw()+theme(plot.title = element_text(hjust = 0.5, size=12,face="bold")) +labs(color= "Age")
grid.arrange(p1,p2,p3,p4,p5,nrow=3,ncol=2)

# side-by-side boxplots of personality traits differentiated by age
colors = rainbow(length(unique(bfi.dat$age)),start=0.1,end=0.9)
names(colors) = unique(bfi.dat$age)

par(mfrow=c(3,2))
boxplot(Neuroticism_N ~ age, data=bfi.dat, xlab="Age", ylab="Neuroticism",col=colors)
title("Boxplot of Neuroticism differentiated by Age")
boxplot(Extraversion_E ~ age, data=bfi.dat, xlab="Age", ylab="Extraversion",col=colors)
title("Boxplot of Extraversion differentiated by Age")
boxplot(Conscientiousness_C ~ age, data=bfi.dat, xlab="Age", ylab="Conscientiousness",col=colors)
title("Boxplot of Conscientiousness differentiated by Age")
boxplot(Agreeableness_A ~ age, data=bfi.dat, xlab="Age", ylab="Agreeableness",col=colors)
title("Boxplot of Agreeableness differentiated by Age")
boxplot(Openness_O ~ age, data=bfi.dat, xlab="Age", ylab="Openness",col=colors)
title("Boxplot of Openness differentiated by Age")
```

R Code Snippet 08: code for study of the personality traits in relation to the demographic attributes.