Continuous Assessment Part II

Maurício de Oliveira Reis - D21125621

Postgraduate Certificate in Data Science - TU256/1, TU Dublin

MATH9102:2022-23: Prob and Statistical Inference

R Version: 4.2.1 (2022-06-23) - Platform: aarch64-apple-darwin20 (64-bit)

R Packages: "ggplot2","dplyr","kableExtra", "caret", "Hmisc"

January 7, 2023

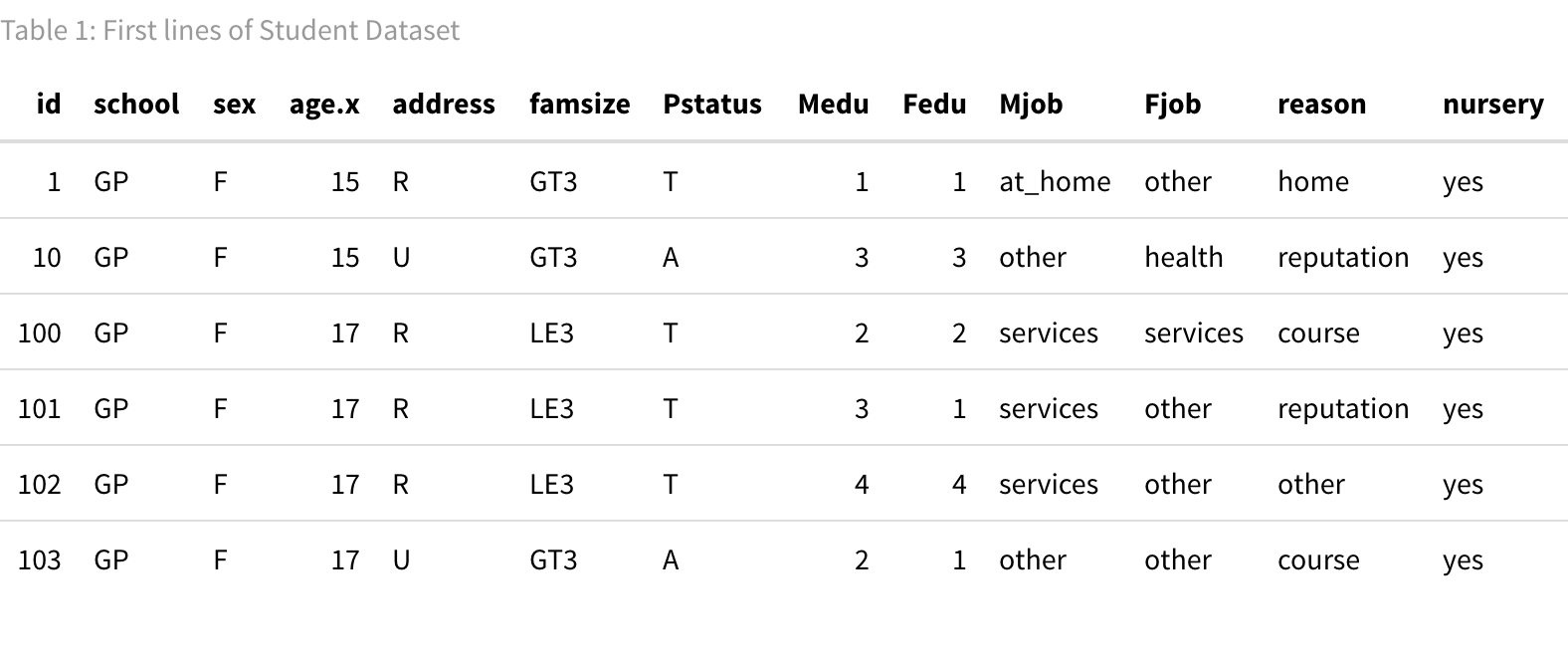
# The Big 5 Personality Traits Test on Students

The Big 5 Personality Traits Test (IPIP Big-Five 50 item Questionnaire, 2023)still now a days applied not only on academic settings but also on professional settings to access a job fit for a particular individual (The Big Five Personality Traits Model and Test, 2022). The objective of this work is, by working with a dataset of this test applied to 382 students, investigate how could we apply how Dimension Reduction to its terms and if we could predict how full of ideas a student would claim to be given other questions answered prior to this one.

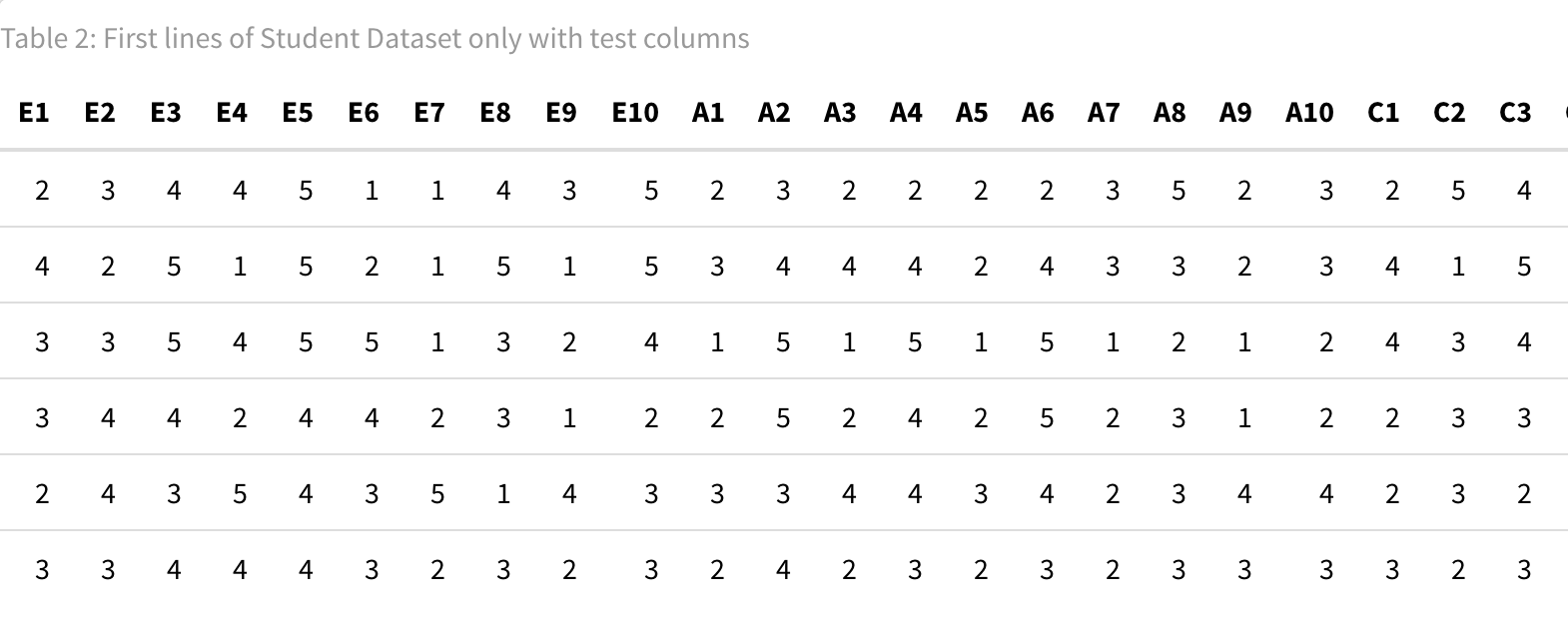
# Personality Trait Test on Students

## The test dataset cleaning

The dataset utilized on this work consists of a questionary with several questions involving not only the student’s personality test but also the student’s school, sex, age, etc. Most of those fields will not be utilized here but can be referred for future work as the dataset will be published together with this report. Table 1 shows a small portion of this dataset:



## As we are only interested on the questionnaire answers, that is what we will discuss on this report. Table 2 shows a sample of the data only for the questionnaire:

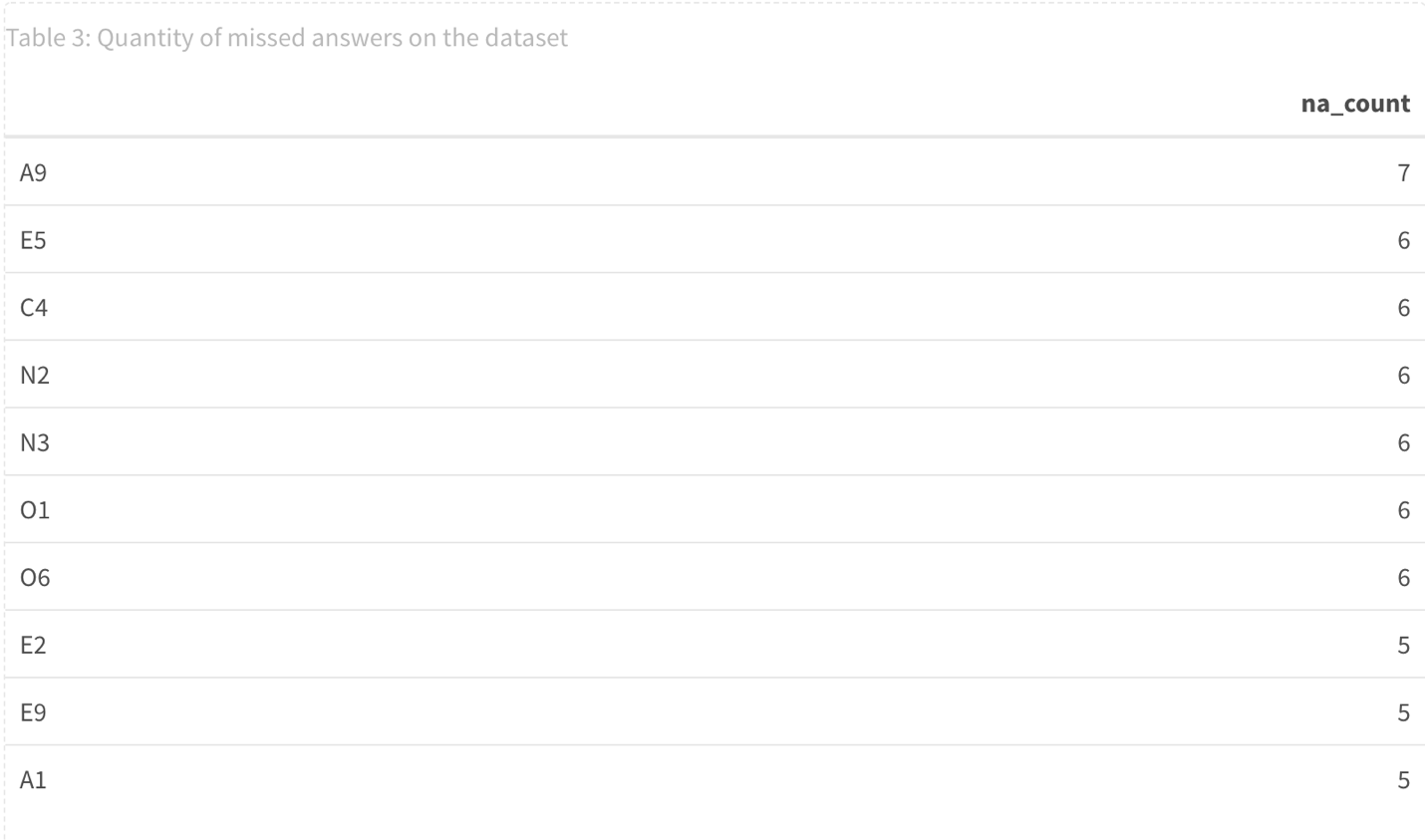


The questionnaire consists in 10 questions about each one of the following personality traits: Openness, Conscientiousness, Extraversion, Agreeableness and Neuroticism. Those questions are represented in the dataset as columns with the name being the first letter of the trait plus the number of the question. The student should answer from 1 to 5 on each of the questions, questions with the answer 0 mean that the student didn’t answer that question. Also, some of those questions were reverse coded and had to be treated like that for this work. The list below shows all the questions on the dataset and which ones were reverse-coded:

* E1: I am the life of the party.
* E2: I don't talk a lot. (Reverse Coded)
* E3: I feel comfortable around people.
* E4: I keep in the background. (Reverse Coded)
* E5: I start conversations.
* E6: I have little to say. (Reverse Coded)
* E7: I talk to a lot of different people at parties.
* E8: I don't like to draw attention to myself. (Reverse Coded)
* E9: I don't mind being the center of attention.
* E10: I am quiet around strangers. (Reverse Coded)
* A1: I feel little concern for others. (Reverse Coded)
* A2: I am interested in people.
* A3: I insult people. (Reverse Coded)
* A4: I sympathize with others' feelings.
* A5: I am not interested in other people's problems. (Reverse Coded)
* A6: I have a soft heart.
* A7: I am not really interested in others. (Reverse Coded)
* A8: I take time out for others.
* A9: I feel others' emotions.
* A10: I make people feel at ease.
* C1: I am always prepared.
* C2: I leave my belongings around. (Reverse Coded)
* C3: I pay attention to details.
* C4: I make a mess of things. (Reverse Coded)
* C5: I get chores done right away.
* C6: I often forget to put things back in their proper place. (Reverse Coded)
* C7: I like order.
* C8: I shirk my duties. (Reverse Coded)
* C9: I follow a schedule.
* C10: I am exacting in my work.
* N1: I get stressed out easily.
* N2: I am relaxed most of the time. (Reverse Coded)
* N3: I worry about things.
* N4: I seldom feel blue. (Reverse Coded)
* N5: I am easily disturbed.
* N6: I get upset easily.
* N7: I change my mood a lot.
* N8: I have frequent mood swings.
* N9: I get irritated easily.
* N10: I often feel blue.
* O1: I have a rich vocabulary.
* O2: I have difficulty understanding abstract ideas. (Reverse Coded)
* O3: I have a vivid imagination.
* O4: I am not interested in abstract ideas. (Reverse Coded)
* O5: I have excellent ideas.
* O6: I do not have a good imagination. (Reverse Coded)
* O7: I am quick to understand things.
* O8: I use difficult words.
* O9: I spend time reflecting on things.
* O10: I am full of ideas.

The fields from the list above that were marked as reverse coded had their values changed with the following formula: adjustedReverseCodingValue = maximumScore + 1 – originalScore.

Also, worth to mention that there were a few questions that had the value 0, which means that the students didn’t answer them. Since there were very few of them, 1.83% of the rows in the worst case, which was for A9, those 0 values got replaced by the mode of each question. Table 3 shows the questions with most 0 values:



## Dimension Reduction

**Chosen Data and Hypothesis**

The objective of this work was to prove that some of the values in this dataset cloud be aggregated by utilizing dimension reduction. For this piece of work, all the questions of 3 different personality traits were chosen: conscientiousness, neuroticism and openness.

The hypothesis that was tested is the following:

* Null Hypothesis: Those 30 variables within conscientiousness, neuroticism and openess, measure 30 completely distinct and unrelated characteristics or a person.
* Alternative Hypothesis: Those 30 variables within conscientiousness, neuroticism and openess, measure less than 30 completely distinct unrelated characteristics of a person.

**Data Summary**

Table 4 shows summary statistics for a piece of the data utilized in the dimension reduction:

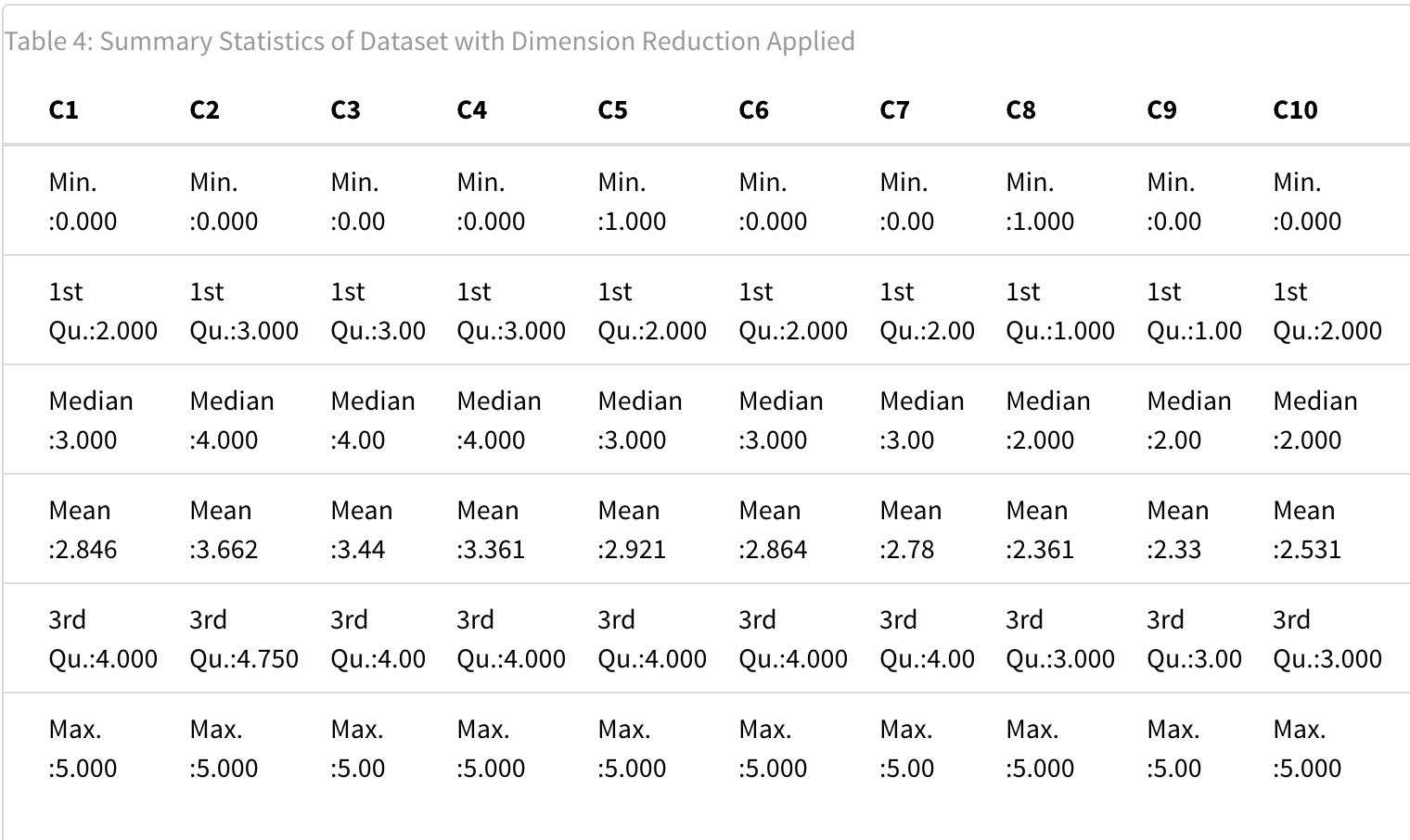


Table 4 shows that this is a very heterogeneous group of variables with first quartile ranging from 1 all the way to 4 and the 3rd quartile ranging from 3 to 5. Min and max values are 1 and 5 respectively, as expected since this data has been treated and zeros got replaced by the mode.

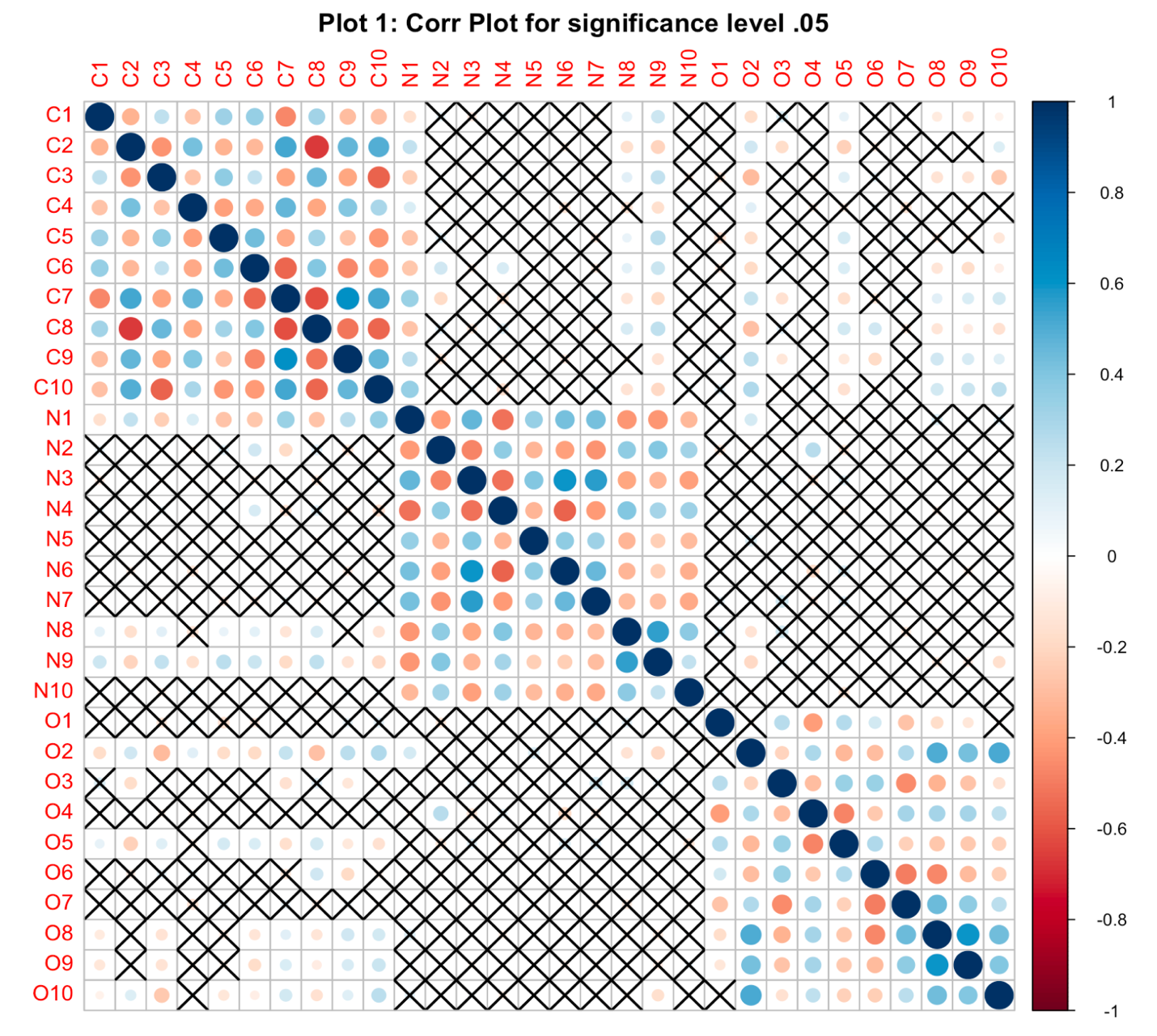
**Assessing the suitability of the dataset for dimension reduction.**

This dataset was, before anything else, accessed for suitability of dimension reduction, the criteria below were considered:

- Quality of Data: this dataset has been cleaned up and contains real world data about a IPIP Big-Five 50 item Questionnaire. This work assumes that this is real data.

- Sample size: this dataset has 382 rows which is more than enough to provide a stable factor solution (Hong, 1998). We also get here a relation rows per columns of more than 10:1 given that we have 382 rows and we are utilizing only 30 columns from this dataset.

- Correlation between items: Plot 1 shows the correlation between items, the correlation is mostly high among questions of the same letter but we can see correlations between questions of different letters too. Most importantly, a Barlett's test was conducted and yielded a p-value so low that was reported as 0, and a chisq of 4449.18, this shows that there is a significant difference in the variances of the variables on this group therefore, this dataset is adequate.

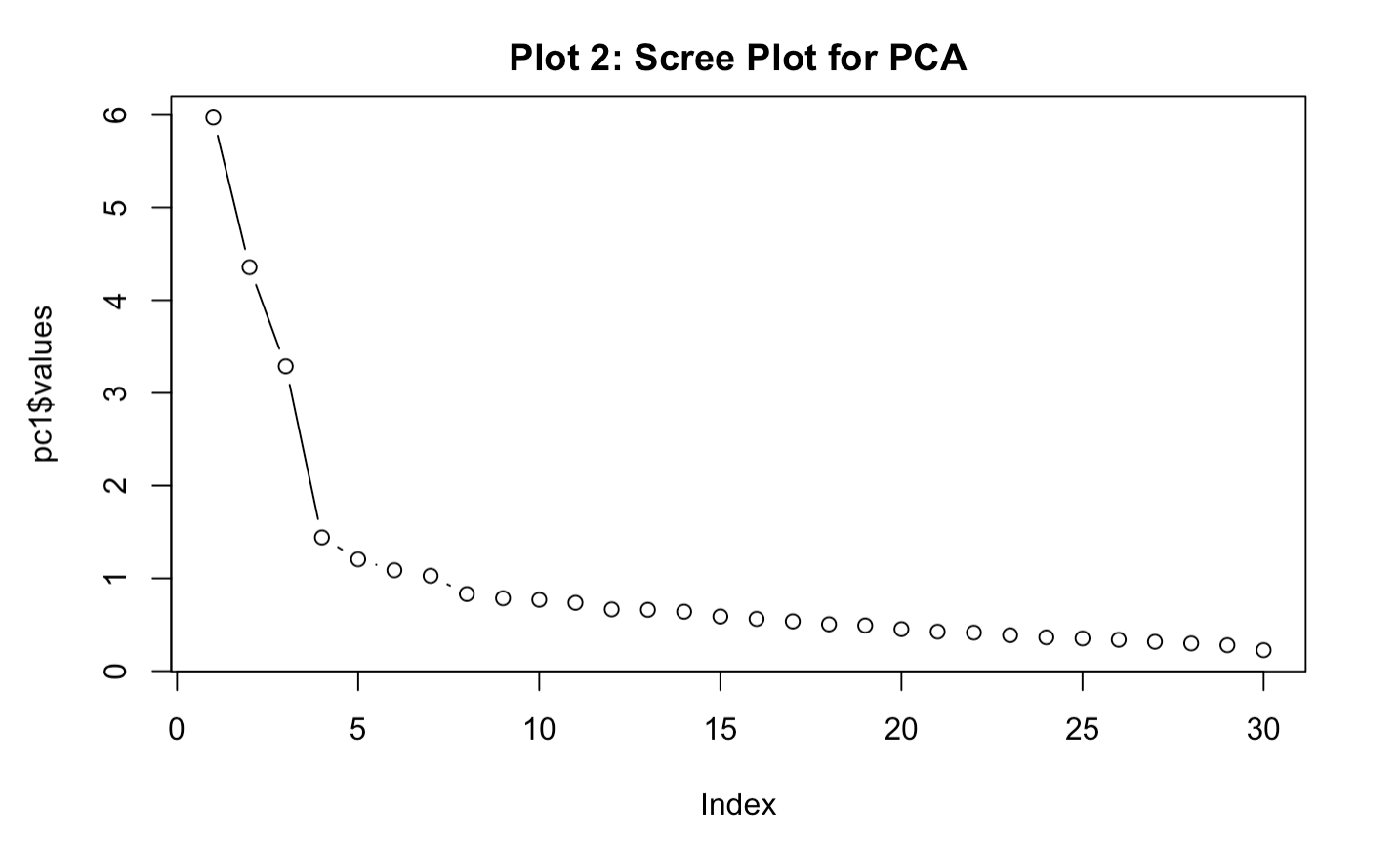


- KMO: a Kaiser-Meyer-Olkin Test was conducted and yielded a KMO criterion of 0.86, which indicates that this sampling is adequate for Factor Analysis (Kaiser-Meyer-Olkin (KMO) Test for Sampling Adequacy, 2023).

- Determinant: the determinant was also looked at, here the test yeaded a value of 0.00000603 a little bit more than just half of the threshold of 0.00001. This might indicate a completely independent item on the chosen data but, as the intention here is doing a PCA (Principal Component Analysis) (Principal component analysis , 2023), our main concern here is with the data reduction, therefore, an independent item may not be a problem. Given that, we will proceed with the analysis.

**Conducting the Dimension Reduction**

In order to conduct the dimension reduction, the adequate number of factors was looked at both with a Scree Plot (Plot 2) and looking at the Eigenvalues of each factor (Table 6).



As we can see on plot 2, the point of inflection occurs around 4.

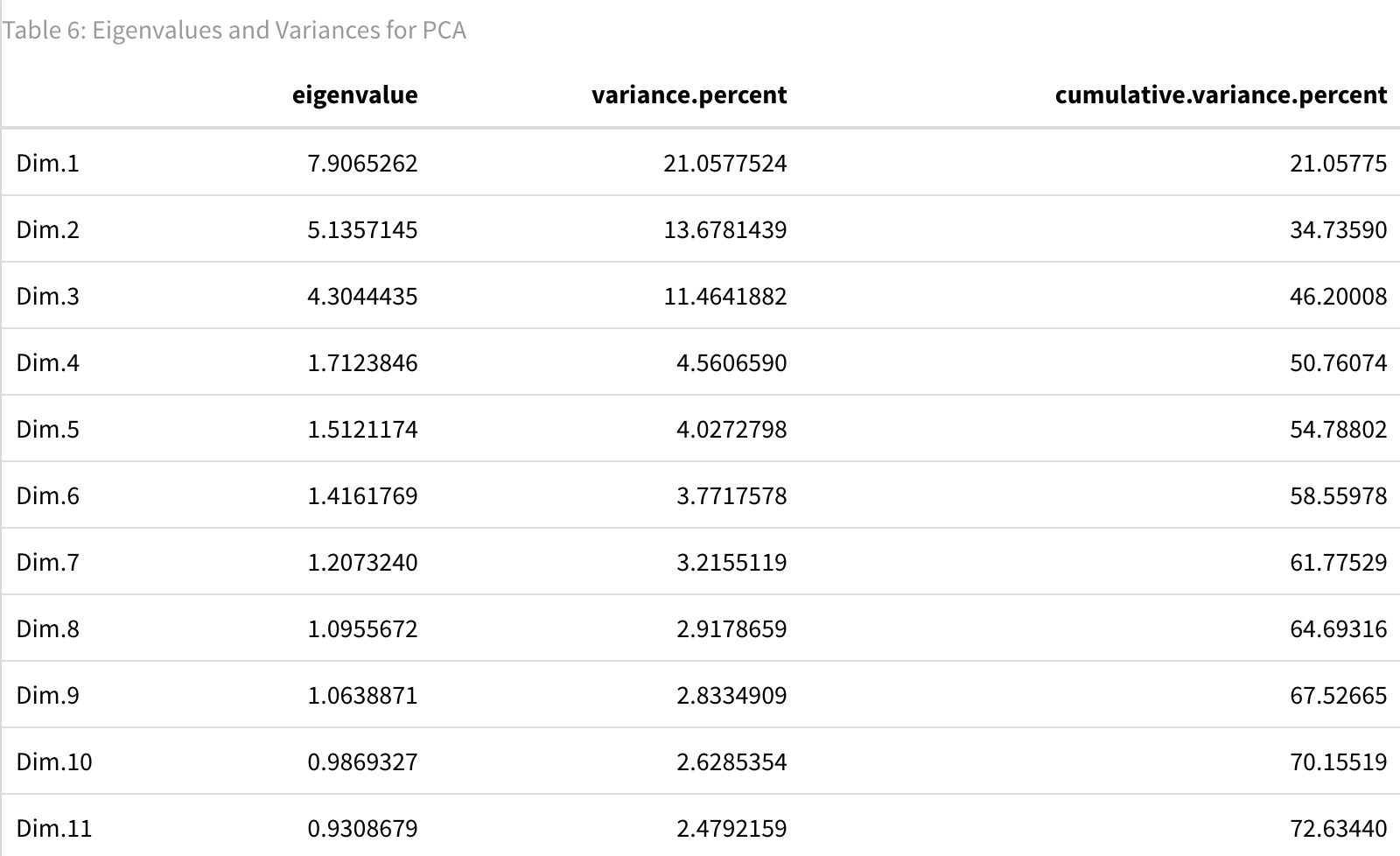
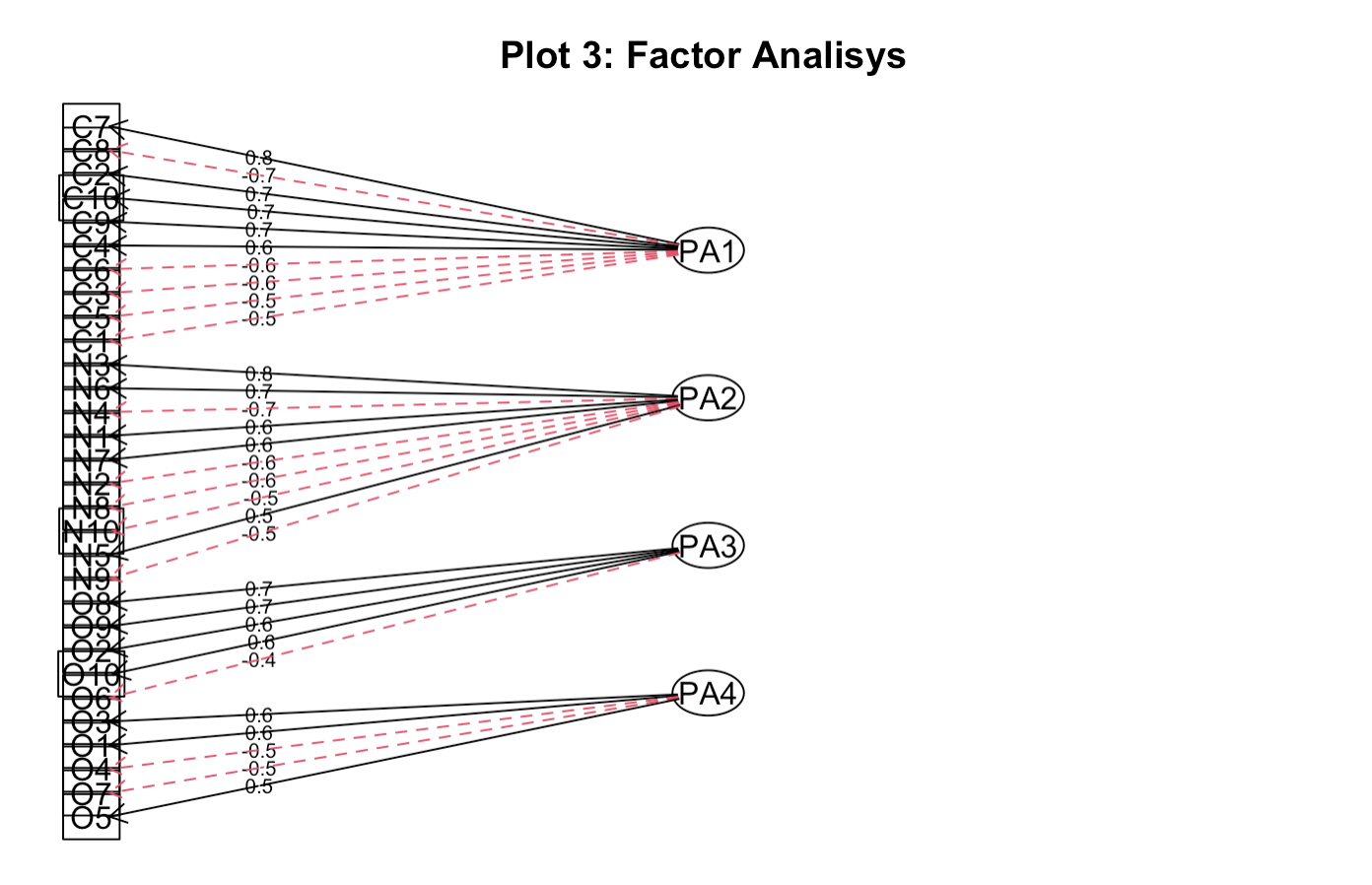


Table 6 in the other hand suggests, based on Eigenvalue > 1, that 9 components should be chosen. As this dataset has n > 300, the decision was to stay with the results from the Scree Plot and choose 4 factors, those 4 factors amount for 50.76 of the variance from the 30 variables. A Factor Analysis on Plot 3 shows how those 4 factors would be composed:



On plot 3 we can see the commonalities between the variables and the factors. As we can see, all of them are on or above 0.4

After performing the dimension reduction, a test to report the Cronbach Alpha values was done and its shown on Table 7:

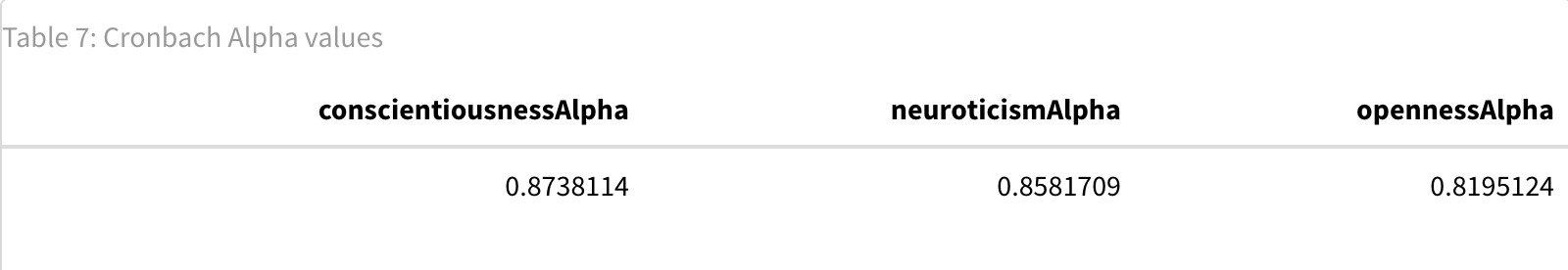


Table 7 showed that conscientiousness, neuroticism and openness had high

reliability, all Cronbach’s alpha > .80.

With this analysis, we could reject the Null Hypothesis and conclude that it’s possible group some of the characteristics captured by the questions about conscientiousness, neuroticism and openness.

**Linear Regression – Predicting Creativity**

**Hypothesis to be tested**

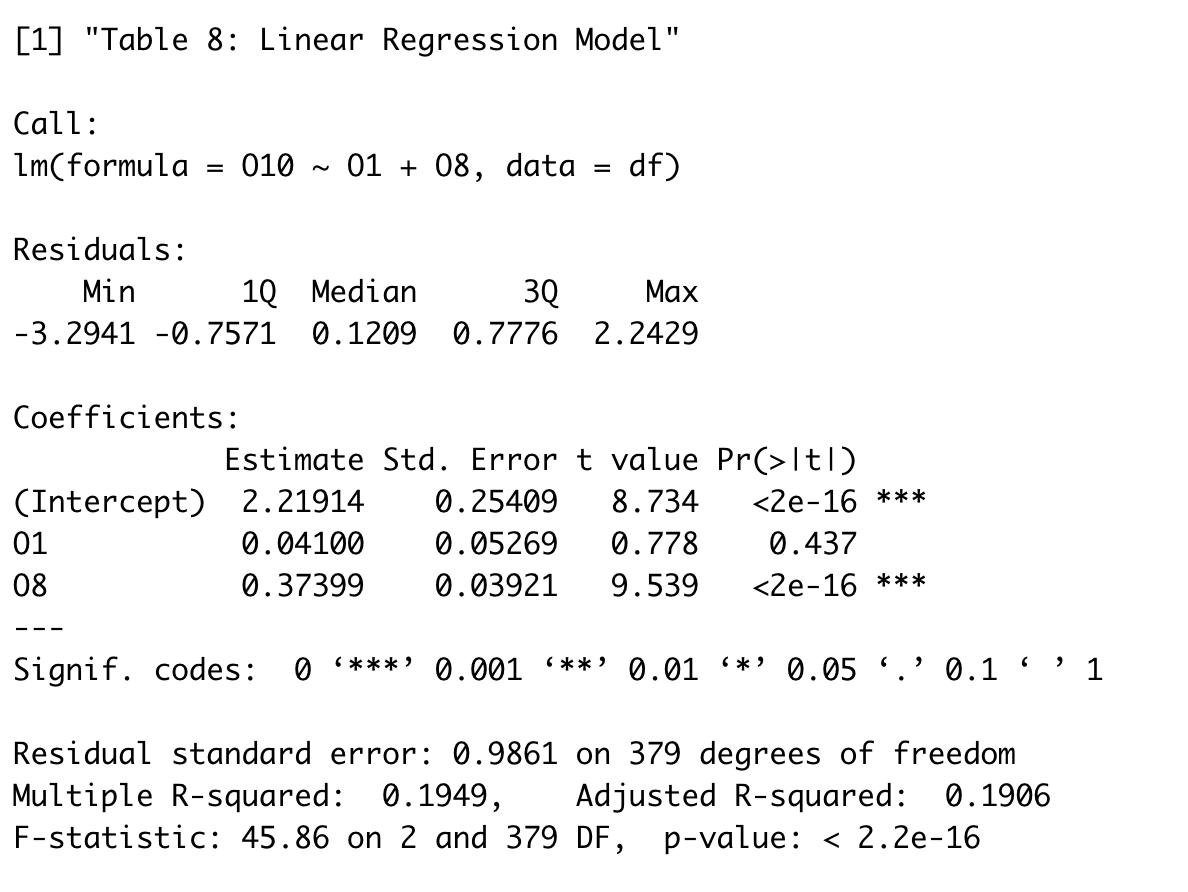
This part of the work will investigate how the possession and usage of vocabulary can impact creativity. To do that, two independent variables were used "O1 - I have a rich vocabulary" and "O8 - I use difficult words" to predict to predict the answers on "O10 - I am full of ideas".

Null Hypothesis: O1 and O8 are not related to O10, therefore you can't use it to predict what will be answered on O10.

Alternative Hypothesis: O1 and/or O8 are related to O10 and can be used to predict what will be answered on O10.

**Linear Regression Model**

This is the initial model: 𝑂10𝑖=𝛽0+𝛽1(𝑂1)+𝛽2(𝑂8)+𝜀𝑖. Table 8 shows the summary of this linear regression model.



The summary shows very low p-value for O8, 2e-16, but quite high for O1 but, before considering that using difficult words doesn’t have anything to do with being full of ideas, we should look if these isn’t an interaction between answering O1 and O8. The next model will be this: 𝑂10𝑖=𝛽0+𝛽1(𝑂1)+𝛽2(𝑂8)+𝛽3(𝑂1×𝑂8)+𝜀𝑖

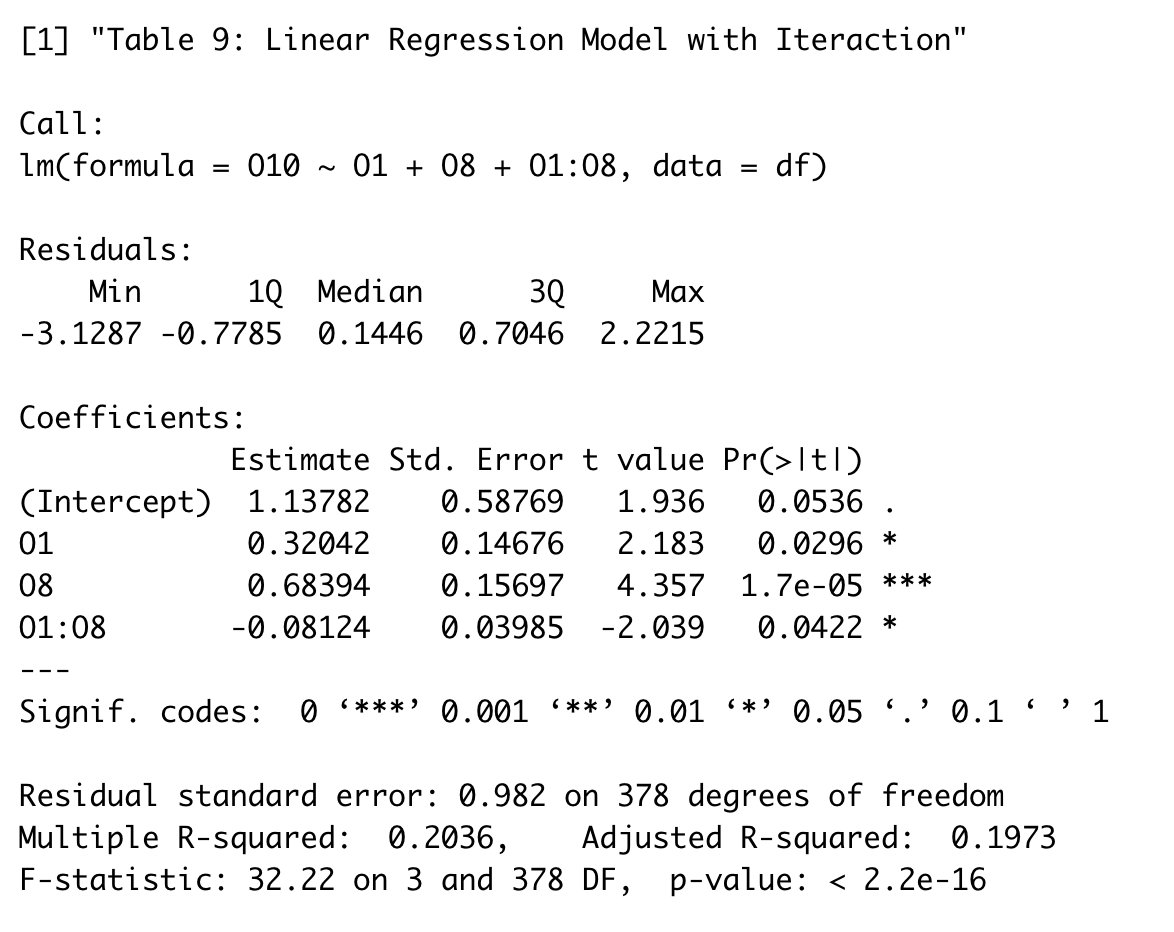
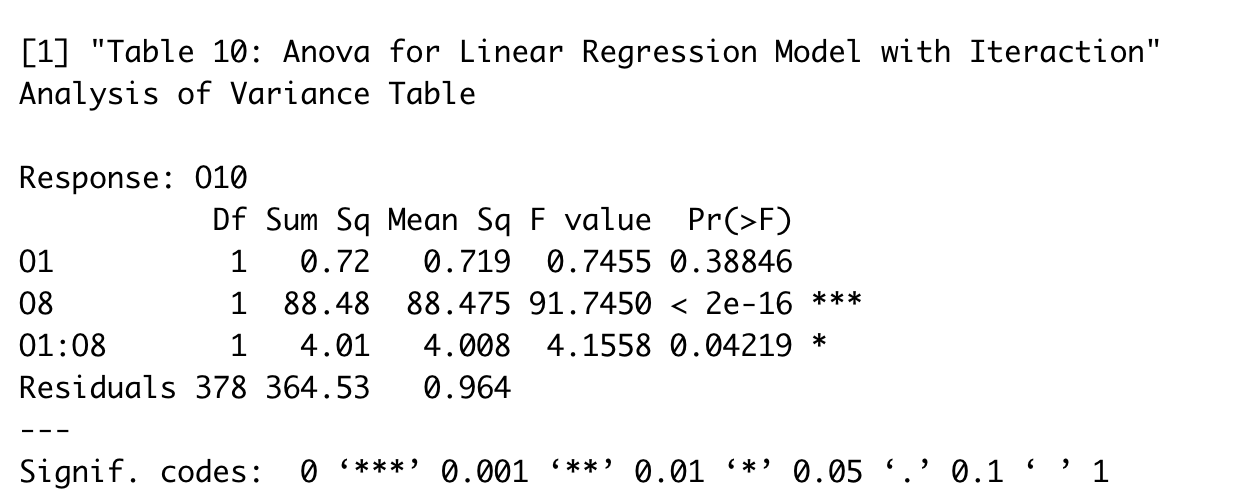


Table 9 shows that there is a relevant interaction between O1 and O8 where, the higher both get, the lower is the score on O10. Now, all terms have a p-value below 0.05. Also, this model also has a slightly higher adjusted r-squared of 0.197 compared to the 0.19 from the previous model implying that it is capable of explaining 19.7% of the variance on O10.



The reason to not choose this model is shown on Table 10. Although the regression model suggests that O1 is significant for the model, the ANOVA table revealed that this wasn’t the case, O1 here had a p-value of 0.39, way above 0.05. with that, another model without O1 had to be created.

The next model was merely: 𝑂10𝑖=𝛽0+𝛽1(𝑂1)

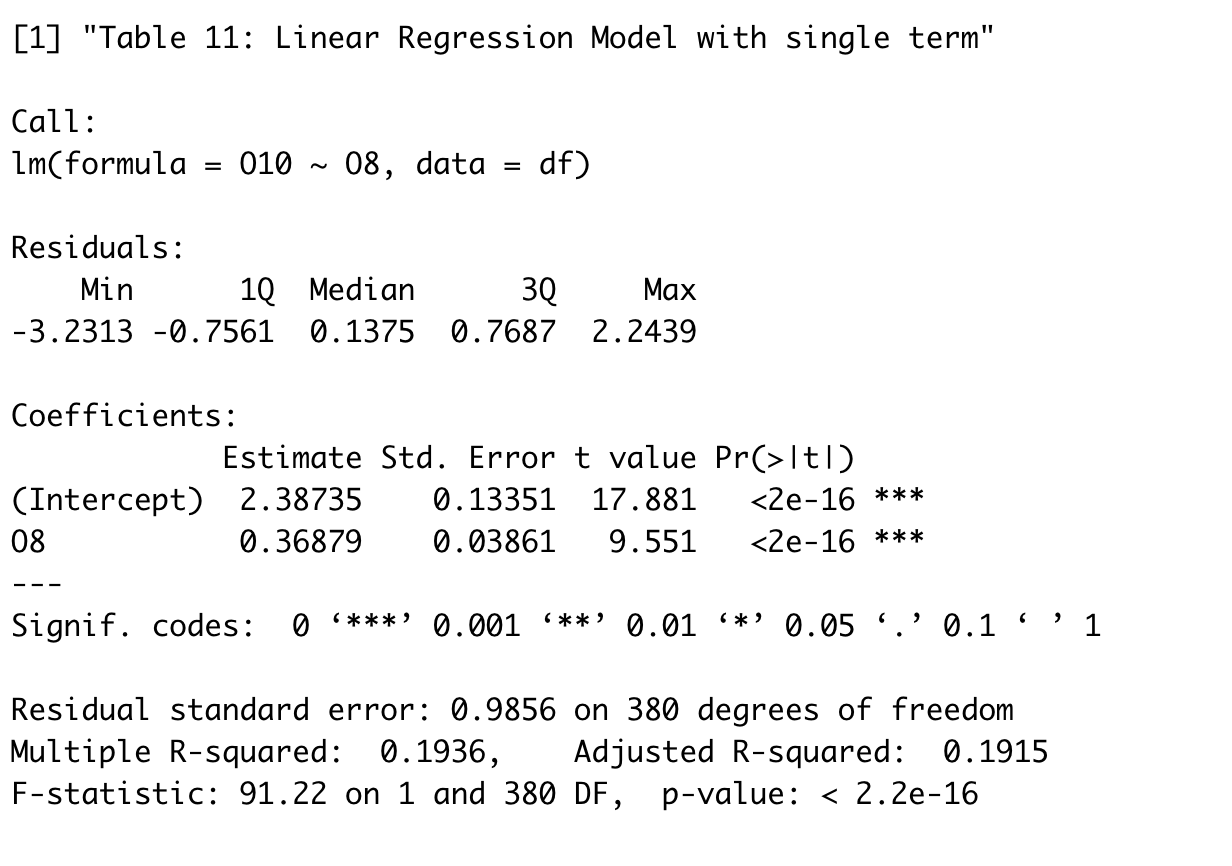


Table 11 shows that O8 stills significant (p-value of less than 2e-16 ) and this model still capable of explaining 19.15% of the variance on O10. In any case, this model is sufficient to reject the null hypothesis and conclude that O1 and/or O8 are related to O10 and can be used to predict how it will be answered.

**Linear Regression – Predicting Max Score on Creativity**

**Hypothesis to be tested**

This model is an extention of the work from the linear regression on predicting how O10 would be answered based on O1 and O8 but, the approach here will be to predict only maximum scores (5) on O10 based on the presence or absence of maximum scores on O1 and O8.

To do this work, a new binary field was created of each of those 3 fields that would take the value of max score true if the score was 5 or false otherwise.

Null Hypothesis: O1 and O8 Max Scores are not related to O10 Max Scores, therefore you can't use it to predict O10 Max Scores.

Alternative Hypothesis : O1 and/or O8 Max Scores are related to O10 Max Scores and can be used to predict it.

**Logistic Regression Model**

The Logistic Regression Model utilized was the following: 𝜂𝑖=𝛽0+𝛽1(𝑂1𝑖)+𝛽1(𝑂8𝑖)

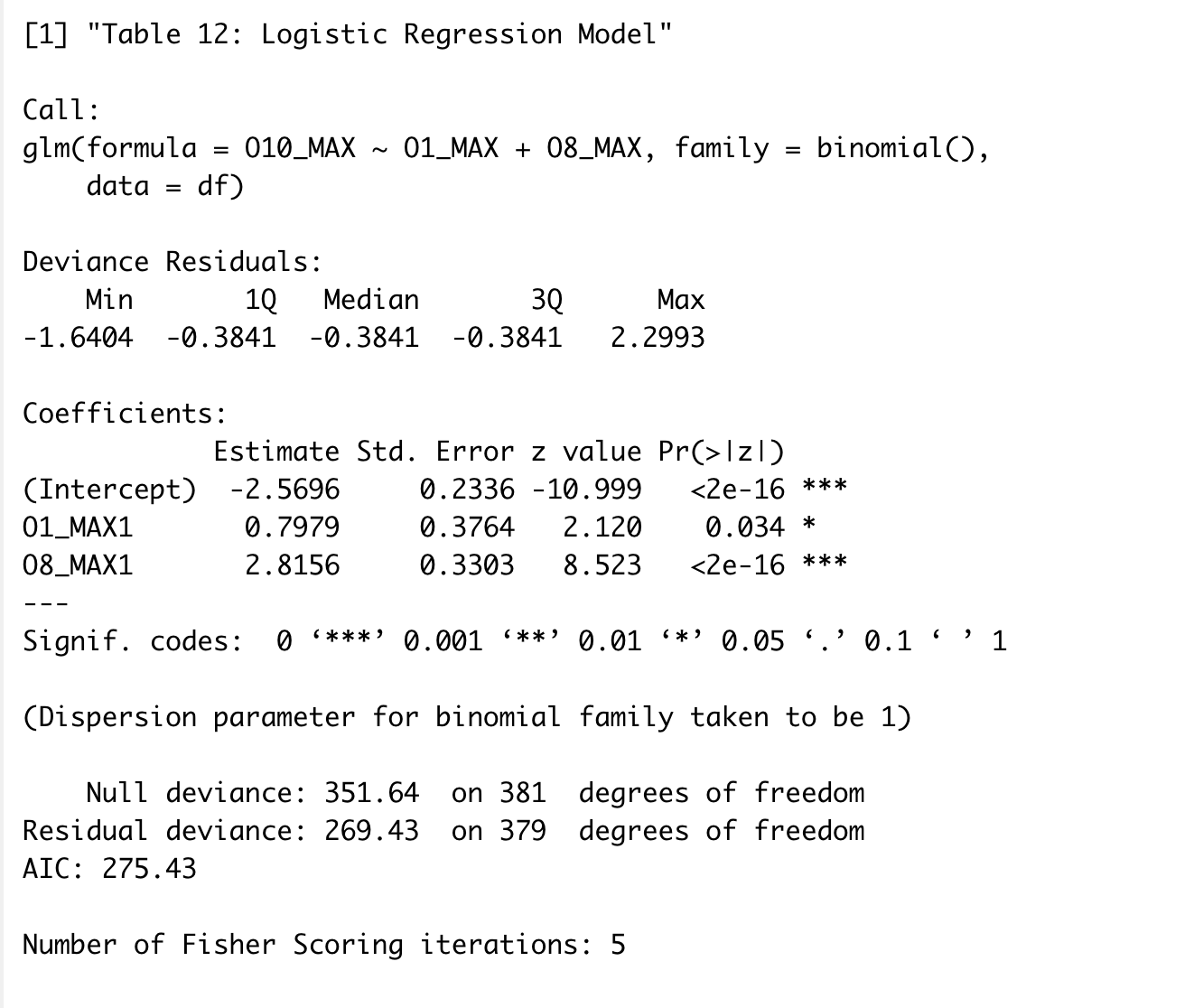


Table 12 shows the summary of model where max scores on both O1 and O8 contribute positively to have max scores on O10 but, and both are significant with p-values of 0.034 and less than 2e-16 respectively.

This model is sufficient to reject the null hypothesis and conclude that O1 and O8 max scores can be utilized to predict if O10 will get a max score.

# Conclusion

## Limitations of Existing Work and Future Research

This work just looked at simple features on a dataset related to the IPIP Big-Five 50 item Questionnaire and is not, by far a comprehensive work. Although was demonstrated that dimension reduction could be performed on this dataset, much more work would be necessary to optimize it by investigating the variables communalities and the possibility to apply factor rotation.

On the Linear and Logistic Regression work, it only showed that it may be possible to possible to predict some of the score on being full of ideas based on answers given about the person’s vocabulary but this work still a long way from proving that a person with a better vocabulary is more likely to be a more creative person since those were all answers provided by the individuals themselves and not further test was made to certify how good is their vocabulary and how full of ideas they really are.

# References

Hong, S. Z. (1998, 11 07). Retrieved from https://is.muni.cz/el/1423/podzim2014/PSY532/52195199/Reading\_4\_-\_sample\_size.pdf

*IPIP Big-Five 50 item Questionnaire*. (2023, 01 05). Retrieved from wikepedia: https://en.wikipedia.org/wiki/Big\_Five\_personality\_traits

*Kaiser-Meyer-Olkin (KMO) Test for Sampling Adequacy*. (2023). Retrieved from statisticshowto.com: https://www.statisticshowto.com/kaiser-meyer-olkin/

*Principal component analysis* . (2023, 1 6). Retrieved from wikipedia: https://en.wikipedia.org/wiki/Principal\_component\_analysis

*The Big Five Personality Traits Model and Test*. (2022). Retrieved from mindtools: https://www.mindtools.com/a2utf1j/the-big-five-personality-traits-model-and-test)