Unraveling Personalities: A Data-Driven Exploration Using 16PF Questionnaire Responses

Luca Barriviera

January 3, 2024

1 Introduction

Over the years, psychologists have tried to identify key factors, referred to as personality traits, that play a crucial role in characterizing an individual's adaptation to life. These traits are considered enduring aspects of an individual's nature, exhibiting relative stability over time.

To be able to "measure" someone's personality, the researcher Raymond Cattell developed a questionnaire, called the 16PF Questionnaire. This questionnaire comprises 160 statements grouped into 16 categories, each prompting respondents to express their level of agreement on a spectrum ranging from '*Strongly Agree*' to '*Strongly Disagree.*' An illustrative statement from this questionnaire is, for instance, *'I know how to comfort others.'*

For my project about personality analysis, I have identified two datasets sourced from openpsychometrics.org. The first dataset encompasses the questions posed in the 16PF Questionnaire, while the second dataset provides answers from over 40,000 participants, along with additional details such as gender, country of origin, and other features.

There are multiple aims for this project. The first consists of creating a new questionnaire that uses a minimal number of questions and is still able to describe someone's personality (with some approximation). The second aim is to cluster the participants into different *personality groups* using different clustering techniques. Finally, we will use the results obtained, to gather some more information about how personalities are distributed in different genders.

2 Data cleaning and exploratory analysis

The data set originates from responses to an online questionnaire, featuring a comprehensive set of data from 40,000 participants. Each participant answered 163 statements, graded on a scale from 1 to 5. Additional information includes gender, age, country of origin, and the time in seconds taken to complete the questionnaire.
It is important to acknowledge that the participants could likely answer inaccurately to the questionnaire, by inputting random values or series of consecutive numbers. It is particularly important to take these participants into account as they could strongly influence our study. Ideally, only the people who answered truthfully to the statements should be kept in order to yield a high-quality data frame to study.

2.1 Accuracy

One of the columns contains a number from 1 to 100 that represents the accuracy that each participant declared they answered the questions with. The data cleaning began by removing participants that were not precise. Throughout the whole project, the participants who input random values many times will need to be removed. To be able to identify when a value is usual we will generally study the distribution of all the data and erase those that are uncorrelated with the distribution.
After exploring the data set, we notice rows where only 1 or 2 values appear. Removing this data was

shown to have a significant effect on the analysis. I also removed the people who answered the same value consecutively for an unusual number of times.

2.2 Age

We can plot the distribution of the age of the participants. Many values are incredibly high, after removing those we get the histogram (Figure 1).

2.3 Country

We can then plot the distribution of the participants in the 10 most common countries. We notice that the bulk of the participants come from English-speaking countries. (Figure 2).

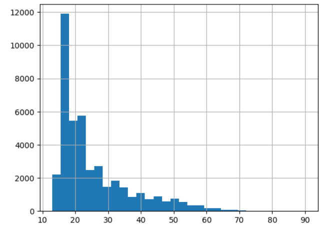

Figure 1: Distribution of ages of participants

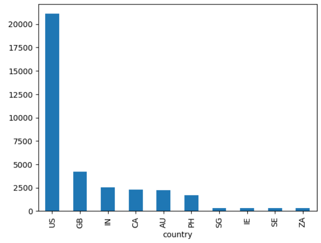

Figure 2: Distribution of countries of participants

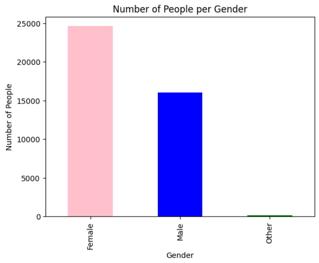

Figure 3: Distribution of participants by gender

2.4 Gender

In the column *gender* we find 4 possible values. The participants whose gender was not saved correctly (`gender = 0`) were removed and then the rest were plotted (Figure 3).

Most of the participants were women and a minority of people inserted "*Other*" as their gender.

3 Principal Component Analysis

Responding to 160 questions can be a time-intensive task, and deducing someone's personality based on those responses is even more challenging. The analysis proceeded by simplifying this process by identifying a smaller set of questions that still effectively capture an individual's personality traits. The idea is that if a person agrees with the statement *'I don't like to get involved in other people's problems*', they might also agree with *'I am not really interested in others.'*

To achieve this, I'll use Principal Component Analysis (PCA) to reduce the number of questions in my dataset, aiming to uncover a more concise set that remains indicative of a person's personality.

To understand the number of principal components that we should keep we can plot the graph of the *Cumulative explained variance* (Figure 4).

Some techniques allow you to properly pick the number of components that someone should keep, in our case, since our problem is strongly related to a real-life problem I will personally choose that number.

With the incorporation of seven principal components, the explained variance amounts to approximately 38%. Consequently, we opt to employ seven principal components as the designated number.

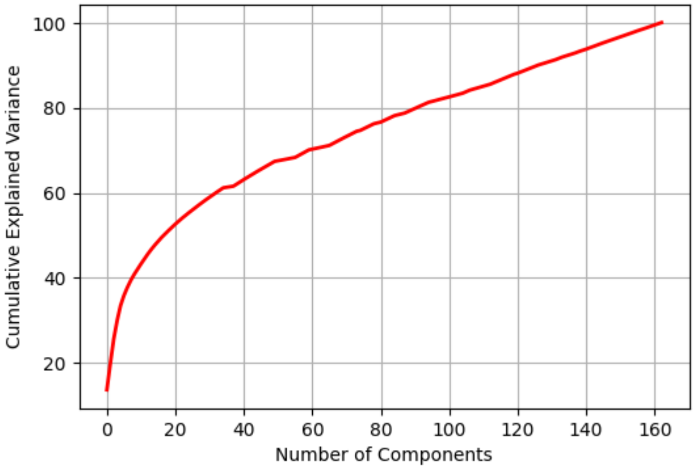

Figure 4: Cumulative explained variance

After running the algorithm we can analyze the main statements to get an idea of what each component tries to describe. Let's show the top 3 statements per principal component (Table 1).

On the left of the statement, I indicated the numerical indicator of the statement. The statements are arranged in groups of approximately 10, where each group attempts to capture a specific trait. We can notice that all the main statements in the first PC come from the same group. We can say that the first component tries to describe how confident someone is in social interactions.
The second component is of particular interest. We can notice that even though the questions come from different groups, they are closely related.

An alternative perspective on these findings is that when examining the sixth component, it can be inferred that individuals who adhere to a particular religion also tend to have greater privacy. It is crucial to check the sign of the weight of the questions that contribute to the components to properly understand if they influence positively or negatively the result.

Let's summarize what each principal component is trying to describe:

Principal Component 1: Social Interactions

Principal Component 2: Emotional Expression

Principal Component 3: Attitude Towards Authority and Rules

Principal Component 4: Reading Preferences

Principal Component 5: Tolerance for Disorder

Principal Component 6: Personal Beliefs and Privacy

Principal Component 7: Openness and Emotional Expression

4 K-Means

The project's second aim is to cluster the participants into K groups. Each group will be composed of people who have similar personalities. After comprehensive evaluations, I concluded that employing $K = 5$ clusters would be optimal for the project's objective of grouping participants based on their similar personalities.

A technique that we can use to visualize the results is to project our original data on the 2 main principal components and then display a scatter plot. (Figure 5, 6).

Table 1: Principal Components Data

Principal Component 1
68 - ”I find it difficult to approach others” 66 - ”I make friends easily” 69 - ”I often feel uncomfortable around others”
Principal Component 2
48 - ”I act wild and crazy” 28 - ”I have frequent mood swings” 46 - ”I enjoy being part of a loud crowd”
Principal Component 3
56 - ”I respect authority” 18 - ”I consider myself an average person” 59 - ”I break rules”
Principal Component 4
75 - ”I read a lot” 73 - ”I like to read” 80 - ”I do not like poetry”
Principal Component 5
149 - ”I am not bothered by disorder” 148 - ”I am not bothered by messy people” 150 - ”I leave a mess in my room”
Principal Component 6
55 - ”I believe in one true religion” 103 - ”I reveal little about myself” 107 - ”I keep my thoughts to myself”
Principal Component 7
109 - ”I am open about my feelings” 108 - ”I am open about myself to others” 111 - ”I show my feelings”

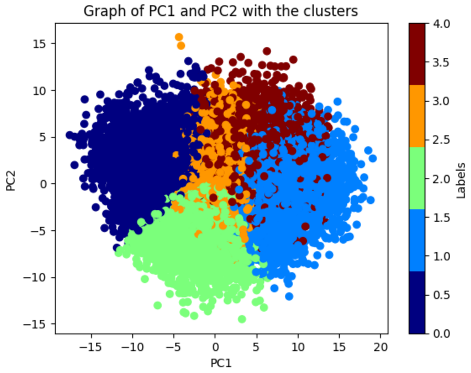

Figure 5: Scatter plot after K-Means

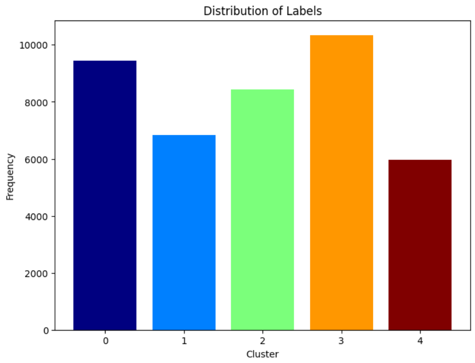

Figure 6: Label distribution

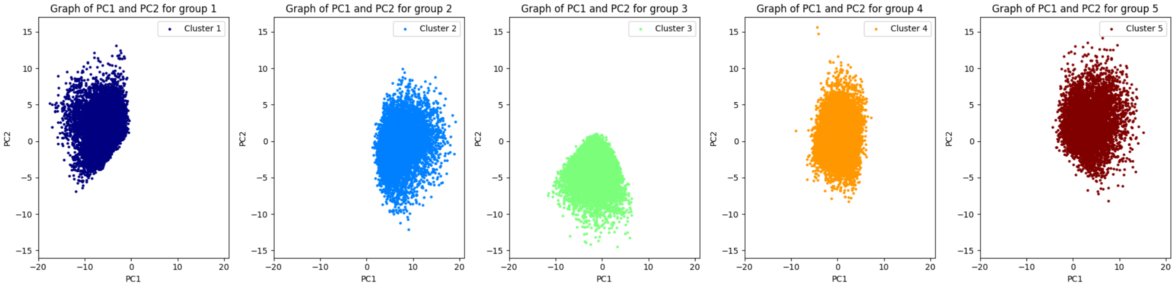

Figure 7: Scatter plots of single clusters

5 Gaussian Mixture Model

A different approach consists of clustering the participants using a *Gaussian Mixture Model*. The choice of this model can be justified by assuming that personalities are distributed as a normal distribution. This assumption is coherent but, since the support of a multivariate Gaussian distribution is $\mathbb{R}^n$ the groups will likely intersect. After applying the methods we can plot the results (Figures 8, 9).

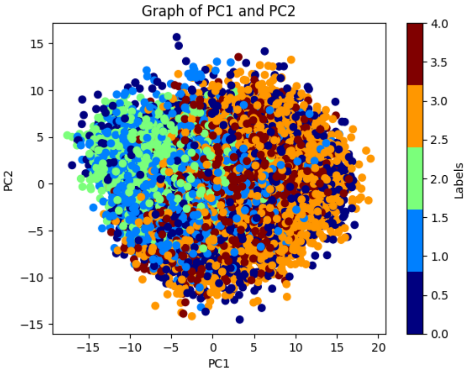

Figure 8: Scatter plot after GMM

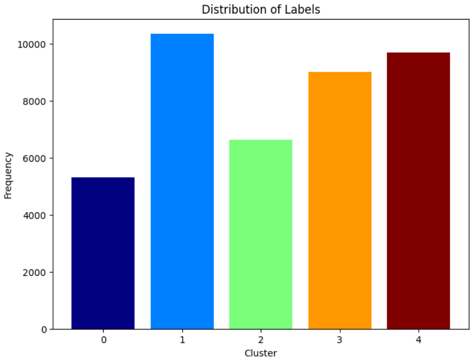

Figure 9: Label distribution

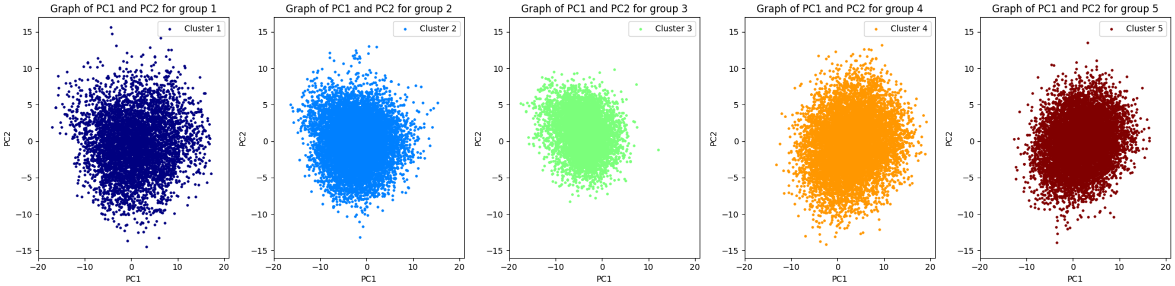

Figure 10: Distribution of ages of participants

As expected, we can notice that the division is much less noticeable in this case for the reason previously explained.
Furthermore the *K-Means* algorithm tries to divide the graph in "*balls*" that generally do not intersect. For these reasons, and for the purposes of my project, I decided to use *K-Means* for the rest of the project.

6 Personality classification

To be able to classify the 5 different groups of people into distinct personality groups we can look at the answer of each centroid concerning the main statements of the principal components. Since every component is a linear combination of questions we need to consider the most relevant questions per PC.
A graph that can be particularly useful in this case, is a radar graph (Figure 11).

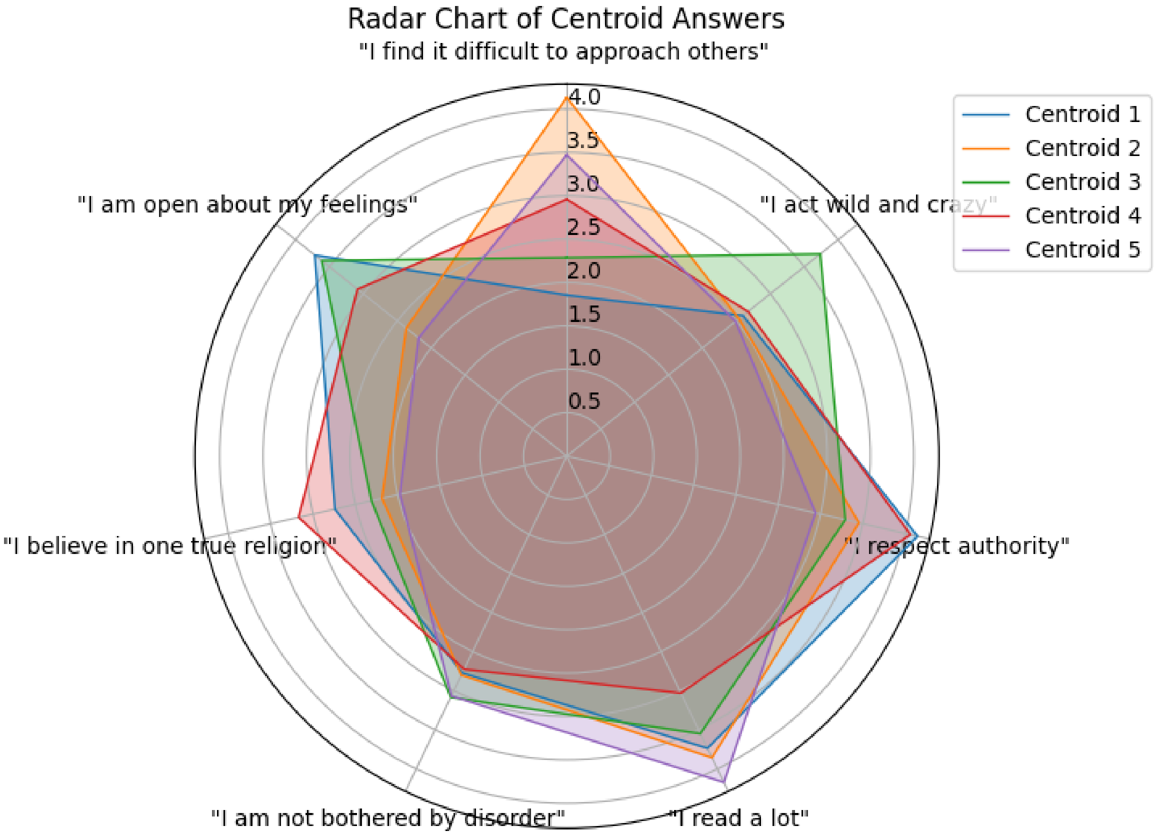

Figure 11: Radar graph with centroid's answers

After studying the results, and using some creativity, the description of the 5 personality types can be found below:

1. **The Disciplined Scholar (Cluster 1):** This individual is known for their disciplined and organized approach to life. They highly value authority and rules, often seeking intellectual pursuits through avid reading. While they may struggle with expressing emotions openly, their commitment to structure and order is admirable.

2. **The Bookish Introvert (Cluster 2):** A bookish and introverted personality, this individual is characterized by a strong affinity for reading and learning. They may be reserved in social situations, less interested in making friends easily, and prefer to keep their emotions private.

3. **The Sociable Leader (Cluster 3):** Outgoing and socially adept, this person is a natural leader who respects authority and values social connections. They are open about their feelings, making them approachable and charismatic in various social settings.

4. **The Creative Free Spirit (Cluster 4):** Unconventional and creative, this individual is a free spirit who enjoys making friends easily and engaging in wild and spontaneous activities. They are open about their feelings and express themselves freely in both thought and action.

5. **The Conscientious Traditionalist (Cluster 5):** A conscientious individual who respects authority and holds onto certain traditions, this person strikes a balance between social and introverted tendencies. They are moderate in their approach to most aspects of life, neither overly outgoing nor excessively reserved.

Finally, we can study the difference in personality between men and women, the distribution of personality types for each gender (Figure 12, 13), and the distribution of personalities for people who were very fast in finishing the test (`elapsed time < 6 mins`, Figure 14) or people that took a long time in answering (`elapsed time > 40 mins`, 15).

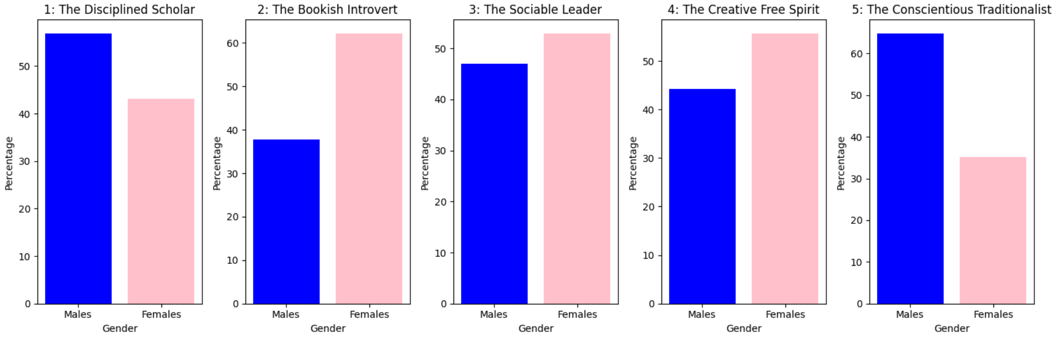

Figure 12: Distribution of gender per personality group

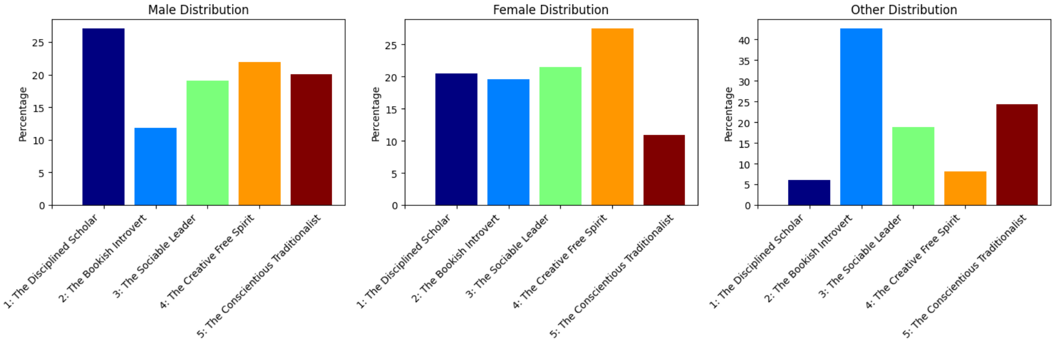

Figure 13: Distribution of personality group per gender

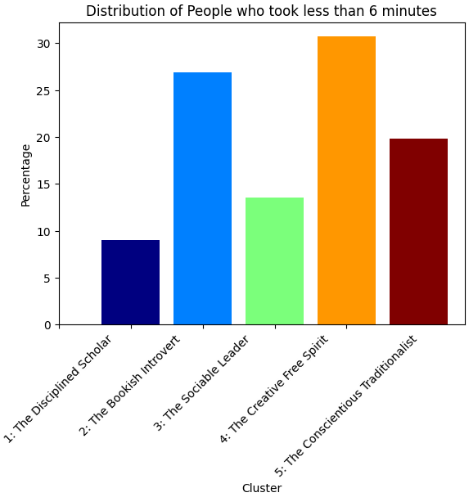

Figure 14: Personalities of people that took less than 6 minutes to finish the test

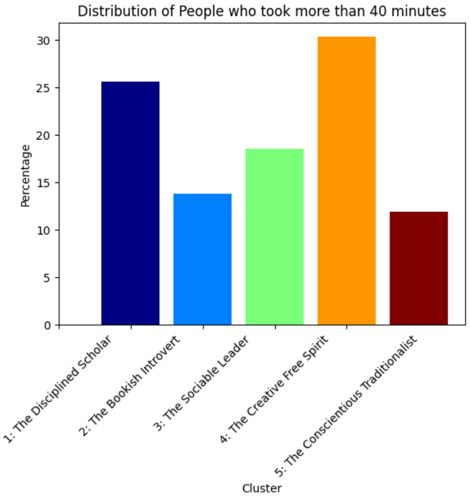

Figure 15: Personalities of people that took more than 40 minutes to finish the test

These kinds of analyses allow us to perform countless studies on personalities. From Figures 14 and 15 we could say, for example, that women tend to be more *Bookish introverts* while men more *Traditionalists*. Additionally, people who are fast at compiling questionnaires are not *Social leaders* while slower respondents are likely to be *Disciplined Scholars*.
These are just examples of how this study can be used to make inferences about individuals and reveal patterns within diverse populations.

7 Conclusion

In this data-driven journey utilizing the 16PF Questionnaire, our central aim was to distill the most pertinent questions for capturing the intricacies of an individual's personality. Employing Principal Component Analysis (PCA), we identified seven key components, such as social interaction, emotional expression, and personal beliefs, providing a concise yet comprehensive lens through which to analyze personality traits. The subsequent application of K-Means clustering illuminated five distinct personality clusters, each representing unique behavioral patterns. As we explore the world of personalities, it's clear that these insights give us an intuition into the diverse ways people behave. This helps us better understand how individuals navigate life's complexities, enriching our understanding of human behavior.