**Employee Quality Predictor**

# Abstract

A Personality Predictor classifies employees to personalities based on traits that best fits the job requirements. I will be using a dataset of 1,015,342 questionnaire answers collected online by [Open Psychometrics](https://openpsychometrics.org/tests/IPIP-BFFM/). I will be using an unsupervised learning algorithm called K-means to cluster the dataset into career labels like Project Manager, Software Engineer, Human Resource Specialist and so, on. The career labels are going to be based on the Big 5 factors: Extroversion, Agreeableness, Conscientiousness, Emotional Stability and Openness. The rise of mental health and stability at work are crucial to most competitive companies. Using these factors as indicators of their positive and negative traits would enormously modify the environment and efficiency of the workplace.

***Keywords***: Python, Flask, Unsupervised Learning, K-means, OCEAN, Data Analysis

**Table of Contents**

[Abstract 1](#_Toc90461252)

[Introduction 3](#_Toc90461253)

[Algorithm 4](#_Toc90461254)

[Project Methodology 5](#_Toc90461255)

[Data Preparation 6](#_Toc90461256)

[Data Visualization 7](#_Toc90461257)

[Model Building 9](#_Toc90461258)

[Flask Deployment 10](#_Toc90461259)

[Conclusion 11](#_Toc90461260)

[Future Enhancements 11](#_Toc90461261)

[References 12](#_Toc90461262)

# Introduction

**This project consists of:**

1. Dataset of 1,015,342 questionnaire from 50 questions
2. Data preparation for training and testing
3. Visualize the results into meaningful insights
4. Deploy the model using Flask
5. Predict the cluster from the answers

*The challenge*

To predict the optimal traits like Extroversion, Agreeableness, Conscientiousness, Neuroticism and Openness for each career based on a questionnaire of 50 questions.

**Dataset consists of:**

* First 50 columns — Represents the questions asked in the questionnaire
* The time spent on each question is also recorded in milliseconds. These are the variables ending in \_E. This was calculated by taking the time when the button for the question was clicked minus the time of the most recent other button click.
* Dateload — The timestamp when the survey was started.
* Screenw — The width the of user's screen in pixels
* Screenh — The height of the user's screen in pixels
* Introelapse — The time in seconds spent on the landing / intro page
* Testelapse — The time in seconds spent on the page with the survey questions
* Endelapse — The time in seconds spent on the finalization page (where the user was asked to indicate if they have answered accurately and their answers could be stored and used for research. Again: this dataset only includes users who answered "Yes" to this question, users were free to answer no and could still view their results either way)
* IPC — The number of records from the user's IP address in the dataset. For max cleanliness, only use records where this value is 1. High values can be because of shared networks (Ex: entire universities) or multiple submissions
* Country – The country, determined by technical information (not asked as a question)
* lat\_appx\_lots\_of\_err — Approximate latitude of user. Determined by technical information noting that it is not very accurate
* long\_appx\_lots\_of\_err — Approximate longitude of user

*Note:* This data was collected (2016-2018) through an interactive on-line personality test. The personality test was constructed with the "Big-Five Factor Markers" from the IPIP. Participants were informed that their responses would be recorded and used for research at the beginning of the test, and asked to confirm their consent at the end of the test. [1]

# Algorithm

K-means is an unsupervised clustering algorithm designed to partition unlabelled data into a certain number (that’s the “K”) of distinct groupings. In other words, k-means finds observations that share important characteristics and classifies them together into clusters. Using this algorithm, it is not possible to choose the types manually as it is chosen by the model as clusters. On the other hand, choosing how many clusters and the parameters are permissible. Five clusters were chosen as the optimal number for this dataset. This algorithm has the ability to cluster the data from hidden insights that aren’t so obvious for us humans.

***Dependencies***

* Python 3.9
* Visualization libraries: Matplotlib and Seaborn
* Libraries for data preparation: Pandas and NumPy
* Machine learning libraries: Scikit-learn
* Web app library: Flask

# Project Methodology

The Big Five personality traits are a proposed typology, or grouping, for personality qualities that was created in psychological trait theory from the 1980s onwards. When component analysis (a statistical approach) is used on personality survey data, semantic correlations are revealed: some terms used to describe features of personality are frequently applied to the same person. Someone who is conscientious, for example, is more likely to be regarded as "always prepared" rather than "messy." These connections provide five major dimensions for describing human personality, demeanor, and psyche in everyday language. [2]

1. Extroversion:

Extraversion is characterized by excitability, sociability, talkativeness, assertiveness, and high amounts of emotional expressiveness.1 People who are high in extroversion are outgoing and tend to gain energy in social situations. Being around other people helps them feel energized and excited.

1. Agreeableness:

This personality dimension includes attributes such as trust, altruism, kindness, affection, and other prosocial behaviours. People who are high in agreeableness tend to be more cooperative while those low in this trait tend to be more competitive and sometimes even manipulative.

1. Conscientiousness:

Standard features of this dimension include high levels of thoughtfulness, good impulse control, and goal-directed behaviors.1﻿ Highly conscientious people tend to be organized and mindful of details. They plan ahead, think about how their behavior affects others, and are mindful of deadlines.

1. Emotional Stability/ Neuroticism:

Neuroticism is a trait characterized by sadness, moodiness, and emotional instability. Individuals who are high in this trait tend to experience mood swings, anxiety, irritability, and sadness. Those low in this trait tend to be more stable and emotionally resilient.

1. Openness:

This trait features characteristics such as imagination and insight.1﻿ People who are high in this trait also tend to have a broad range of interests. They are curious about the world and other people and eager to learn new things and enjoy new experiences. [3]

The acronyms "OCEAN" or "CANOE" might be used to recall the labels for the five components. There are a number of connected and more specific main variables beneath each proposed global component. Extraversion, for instance, is frequently related with social connectedness, assertiveness, excitement seeking, warmth, activity, and positive emotions. These characteristics aren't binary, but rather on a spectrum.

## Data Preparation

First, I imported Pandas and NumPy to use structures and data analysis tools. The “delimiter = “\t” was used to separate the data as they were originally in one cell. I dropped the unnecessary columns and kept 51 columns; the first 50 questions and 1 column for the country as features.

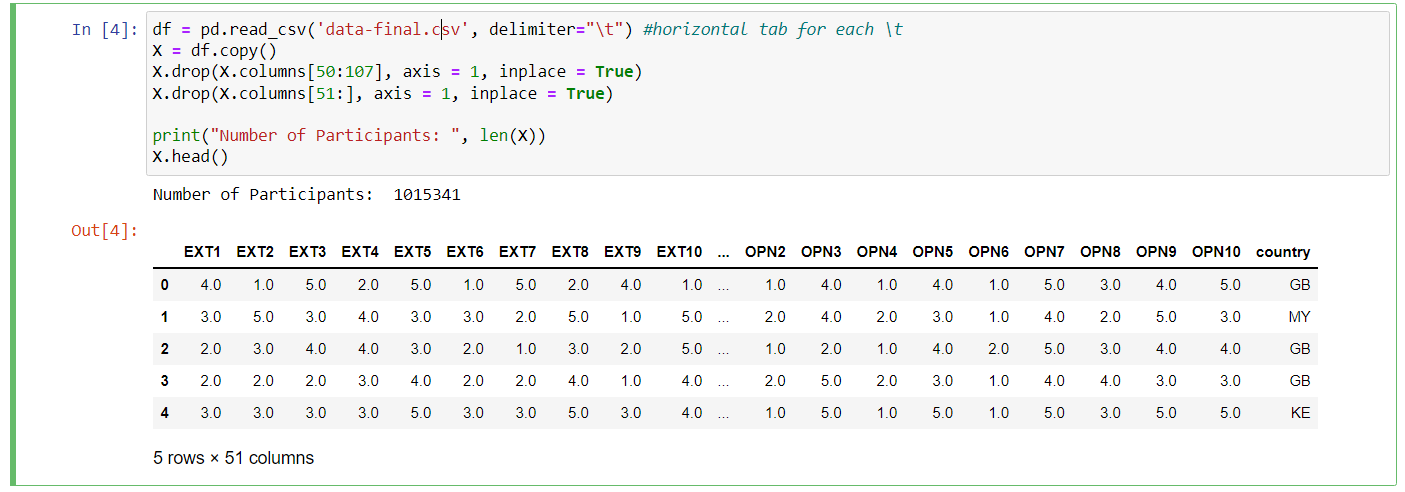


Figure 1: Data Preparation

To clean the data, I searched for null values and dropped them as shown below:

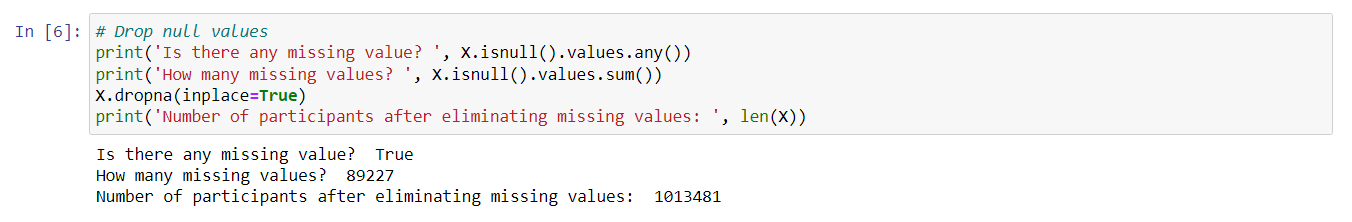


Figure 2: Data Cleaning

## 

## Data Visualization

This graph, displays the participants’ nationality distribution:

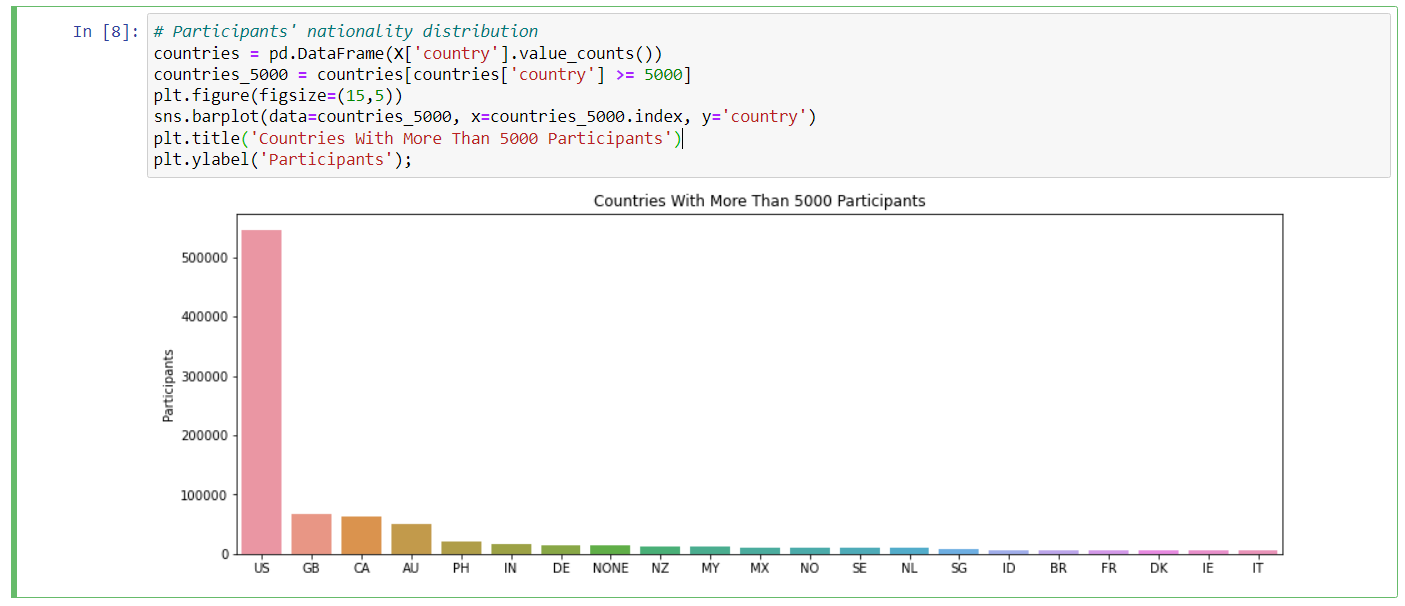


Figure 3: Participants’' Nationality Distribution

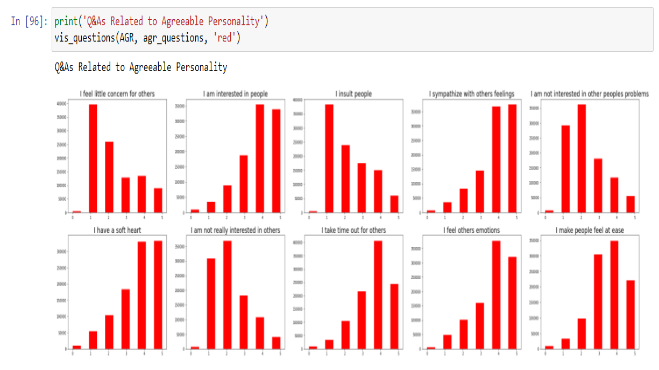
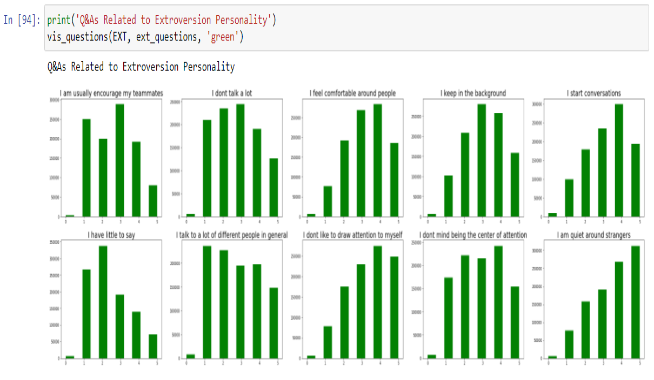
In this part, I made a dictionary that contained the 5 factors. Each factor has 10 questions and the analysis of their answers were displayed below:

Figure 4 Big 5 Factor Distribution

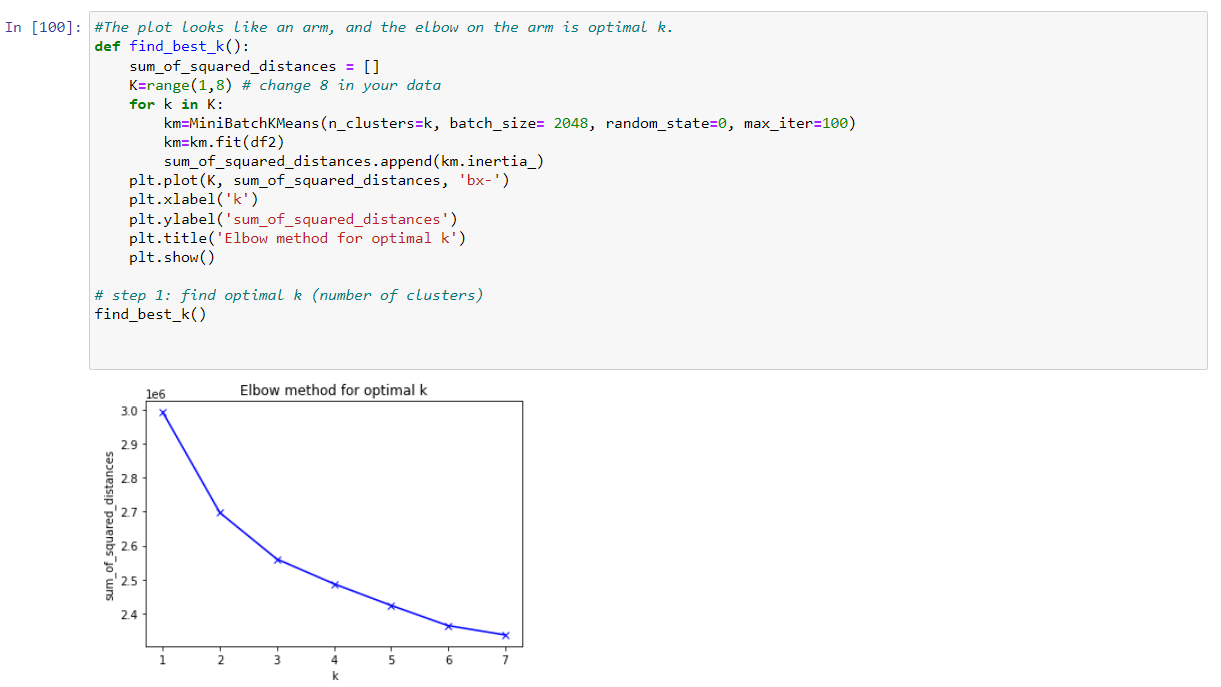
PCA was used to reduce the dimensions and noise to better display the clusters. Then it was plotted using Seaborn library:



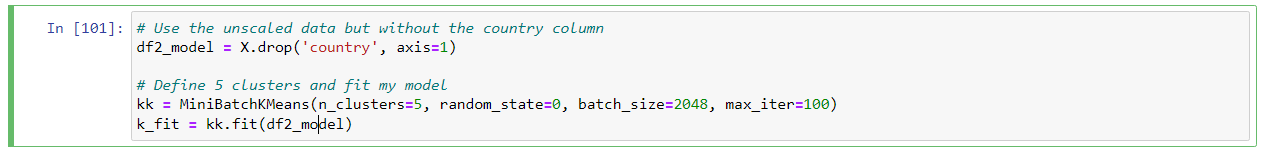
Figure 5: PCA 5 Clusters Plot

## Model Building

Elbow Plot was illustrated as shown to provide the optimal number of clusters for this dataset:



5 clusters were defined and fitted the model



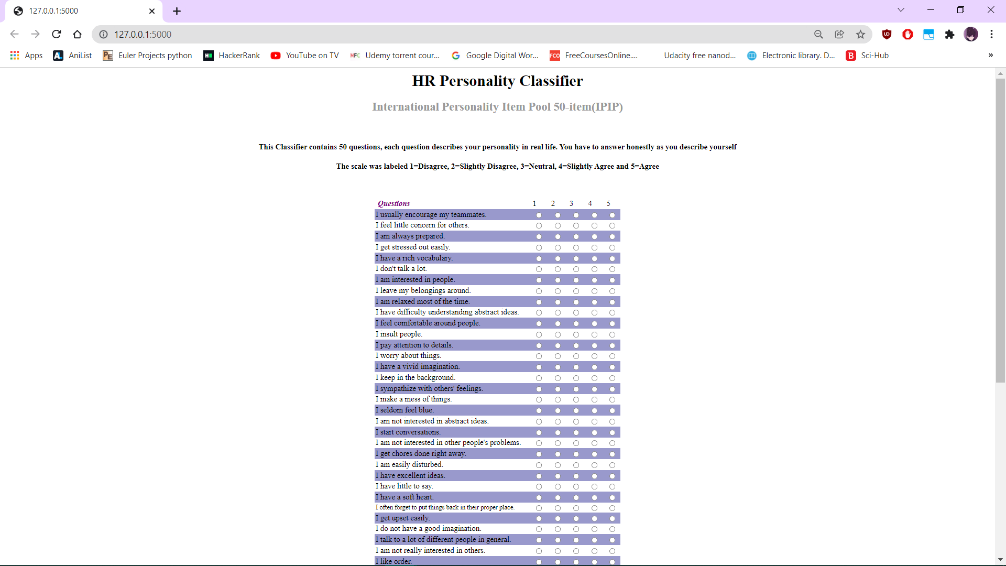
## Flask Deployment

The local hosting of the web app will be deployed in this section. Because of its ease of deployment, the flask library in Python will be used in this project to use the local server. In Part 2, the first step is to organize all of the project files into a project folder. The main project file is stored in (.py) format in the project folder, and the html files are saved in the 'templates' folder. That's all there is to running the software locally.

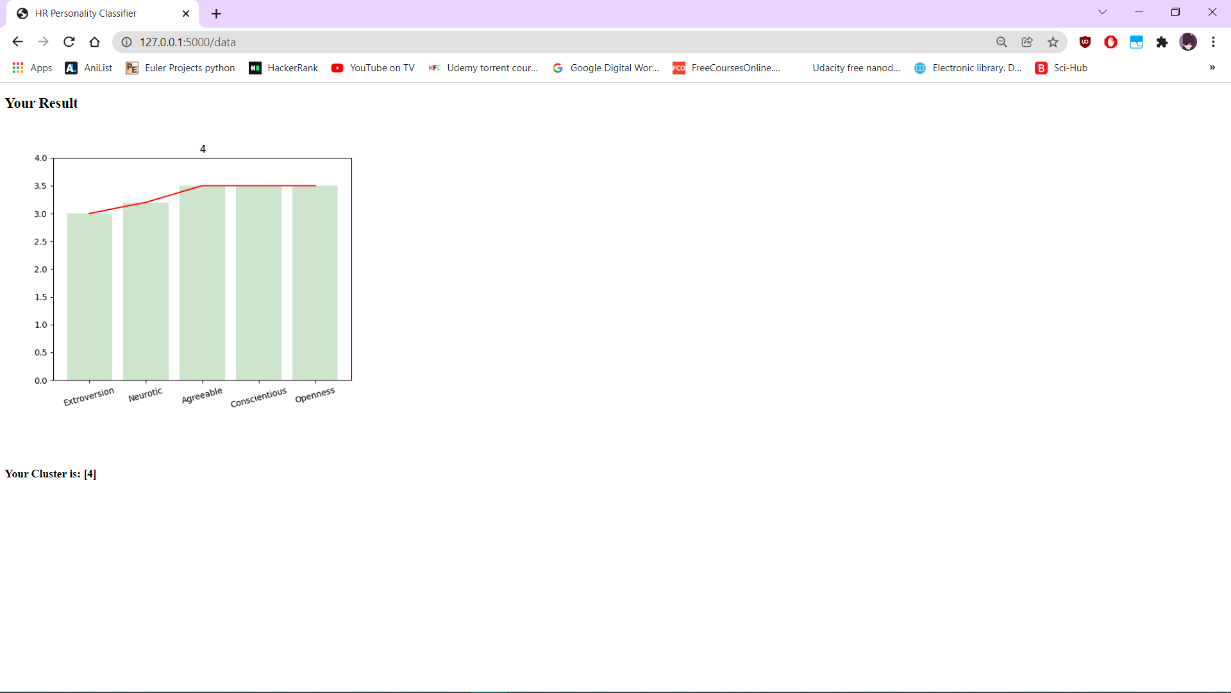
The web app is started using the Terminal in PyCharm after the project folder has been set up. For flask, a distinct python virtual environment is first built. The (.py) file was then assigned as the web app's running script.

The.html files must now be configured for correct user input. Two html files have been produced. One is called 'form.html,' and it takes user input for machine learning model training, while the other is called 'plot.html,' and it displays the anticipated plot. These two files are stored in the template’s directory.

This is the page of questionnaire that will appear when the flask app is run:



After answering the questions, the below picture shows the cluster plot and type:



The cluster type here was 4.

# Conclusion

The work demonstrates the traits and the type of clusters the user is. This can be used to know the positive and negative traits of an employee during the hiring process. The model classifies the user to clusters based on the analysis of the participants’ results in the dataset. For example, people high in Agreeableness tend to be usually warm, friendly, and tactful. As a result, choosing employees high in Agreeableness would be a good choice as they provide a supportive work environment. To sum up, the model can be used by the Human Resources department to better understand and select the optimal candidate based on the job requirements and scores of the 5 factors.

## Future Enhancements

Multiple other tests can be used together with this OCEAN test to make the selection more accurate like the [Full Scale IQ Test](https://openpsychometrics.org/tests/FSIQ/). This IQ test measures across the full spectrum of human abilities. This would give a better interpretation of the candidate cognitive abilities, reasoning and memory. [4]

# References

|  |  |
| --- | --- |
| [1] | "Open-Source Psychometrics Project," 17 9 2019. [Online]. Available: https://openpsychometrics.org/tests/IPIP-BFFM/. [Accessed 14 12 2021]. |
| [2] | Goldberg, L. R. , "The development of markers for the Big-Five factor structure.," *Psychological Assessment,* vol. 4, no. 1, p. 26–42, 1992. |
| [3] | David Susman, "Verywellmind," 20 February 2021. [Online]. Available: https://www.verywellmind.com/the-big-five-personality-dimensions-2795422. |
| [4] | "Open-Source Psychometrics Project," 30 12 2019. [Online]. Available: https://openpsychometrics.org/tests/FSIQ/. [Accessed 15 12 2021]. |
| [5] | Bojan Tunguz, "Kaggle," 2019. [Online]. Available: https://www.kaggle.com/tunguz/big-five-personality-test/activity. [Accessed 2021]. |