Chapter-1

Objective

The primary objective of this project is to develop a sophisticated roommate matchmaking system utilizing Big Five personality traits as a key parameter. Leveraging data-driven insights and advanced analytics, the aim is to create a robust algorithm that enhances the compatibility and overall satisfaction of individuals sharing living spaces.

By incorporating the well-established psychological framework of the Big Five personality traits — Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism — this project seeks to optimize roommate pairings, fostering harmonious cohabitation experiences.

Through the analysis of personality data and the application of machine learning techniques, the project aims to provide an innovative and reliable solution for effective roommate matchmaking, contributing to improved living environments and enhanced interpersonal relationships.

Chapter-2

Data Description

The dataset used in this project is sourced from Kaggle and focuses on the Big Five personality traits, providing a comprehensive view of individual personality characteristics. The dataset comprises responses to standardized personality assessment questions, categorizing individuals based on the following Big Five traits:

- **Openness (OPN)**: Reflects the extent to which an individual is open-minded, creative, and open to new experiences.
- **Conscientiousness (CSN)**: Measures the degree of organization, reliability, and diligence in a person's approach to tasks.
- **Extraversion (EXT)**: Indicates the level of sociability, assertiveness, and comfort in social situations.
- **Agreeableness (AGR)**: Evaluates how cooperative, empathetic, and compassionate an individual is in interpersonal relationships.
- **Neuroticism (EST)**: Measures emotional stability, stress tolerance, and response to negative stimuli.

Data Source:

The dataset was obtained from Kaggle and originally collected through surveys and psychological assessments. The responses are anonymized and aggregated to ensure privacy and confidentiality.

Dataset Size:

The dataset consists of 1 million entries, with 50 attributes capturing various aspects of each participant's personality.

Dataset Attributes:

The dataset is structured with the following attributes:

1. ID: Unique identifier for each participant in the study.
2. OPN1 to OPN10: Likert scale responses (1 to 5) for 10 survey questions related to the Openness personality trait.

3. CSN1 to CSN10: Likert scale responses (1 to 5) for 10 survey questions related to the Conscientiousness personality trait.
4. EXT1 to EXT10: Likert scale responses (1 to 5) for 10 survey questions related to the Extraversion personality trait.
5. AGR1 to AGR10: Likert scale responses (1 to 5) for 10 survey questions related to the Agreeableness personality trait.
6. EST1 to EST10: Likert scale responses (1 to 5) for 10 survey questions related to the Neuroticism personality trait.

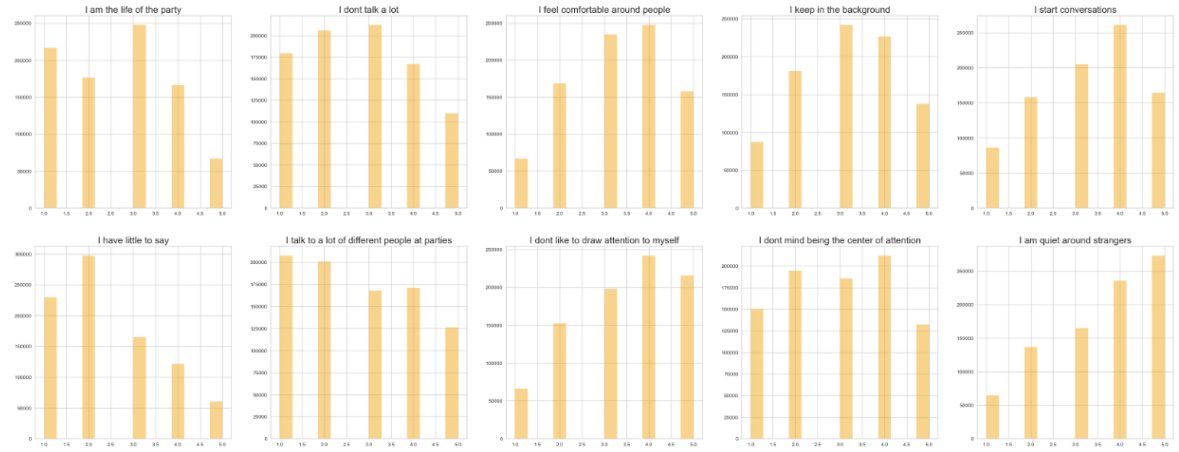
```
Data columns (total 51 columns):
 #   Column   Non-Null Count   Dtype
---  ------   --------------   -----
 0   EXT1     874366 non-null  float64
 1   EXT2     874366 non-null  float64
 2   EXT3     874366 non-null  float64
 3   EXT4     874366 non-null  float64
 4   EXT5     874366 non-null  float64
 5   EXT6     874366 non-null  float64
 6   EXT7     874366 non-null  float64
 7   EXT8     874366 non-null  float64
 8   EXT9     874366 non-null  float64
 9   EXT10    874366 non-null  float64
 10  EST1     874366 non-null  float64
 11  EST2     874366 non-null  float64
 12  EST3     874366 non-null  float64
 13  EST4     874366 non-null  float64
 14  EST5     874366 non-null  float64
 15  EST6     874366 non-null  float64
 16  EST7     874366 non-null  float64
 17  EST8     874366 non-null  float64
 18  EST9     874366 non-null  float64
 19  EST10    874366 non-null  float64
 20  AGR1     874366 non-null  float64
 21  AGR2     874366 non-null  float64
 22  AGR3     874366 non-null  float64
 23  AGR4     874366 non-null  float64
 24  AGR5     874366 non-null  float64
 25  AGR6     874366 non-null  float64
 26  AGR7     874366 non-null  float64
 27  AGR8     874366 non-null  float64
 28  AGR9     874366 non-null  float64
 29  AGR10    874366 non-null  float64
 30  CSN1     874366 non-null  float64
 31  CSN2     874366 non-null  float64
 32  CSN3     874366 non-null  float64
 33  CSN4     874366 non-null  float64
 34  CSN5     874366 non-null  float64
 35  CSN6     874366 non-null  float64
 36  CSN7     874366 non-null  float64
 37  CSN8     874366 non-null  float64
 38  CSN9     874366 non-null  float64
 39  CSN10    874366 non-null  float64
 40  OPN1     874366 non-null  float64
 41  OPN2     874366 non-null  float64
 42  OPN3     874366 non-null  float64
 43  OPN4     874366 non-null  float64
 44  OPN5     874366 non-null  float64
 45  OPN6     874366 non-null  float64
 46  OPN7     874366 non-null  float64
 47  OPN8     874366 non-null  float64
 48  OPN9     874366 non-null  float64
 49  OPN10    874366 non-null  float64
 50  country  874366 non-null  object
dtypes: float64(50), object(1)
memory usage: 340.2+ MB
```

Chapter-3

Exploratory Data Analysis

Conducting Exploratory Data Analysis on each personality trait questions through the bar plots following inferences have been made.

Extraversion Personality:



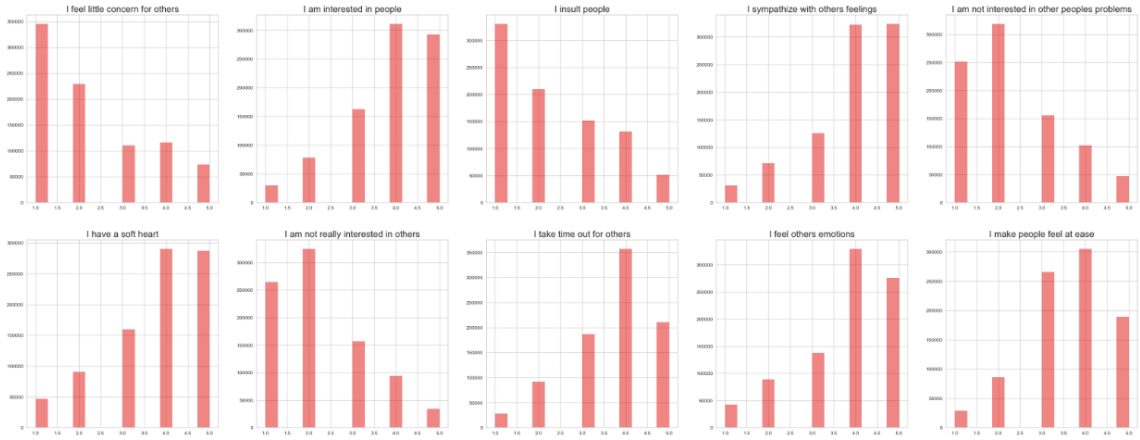
From the above plot of survey questions on Extraversion personality it can said that the data consists of a greater number of introverts rather than extroverts

Neuroticism Personality:

The plots depicting Neuroticism personality traits suggest a noticeable prevalence of candidates characterized by a tendency to overthink and exhibit neurotic tendencies.

Agreeableness Personality:

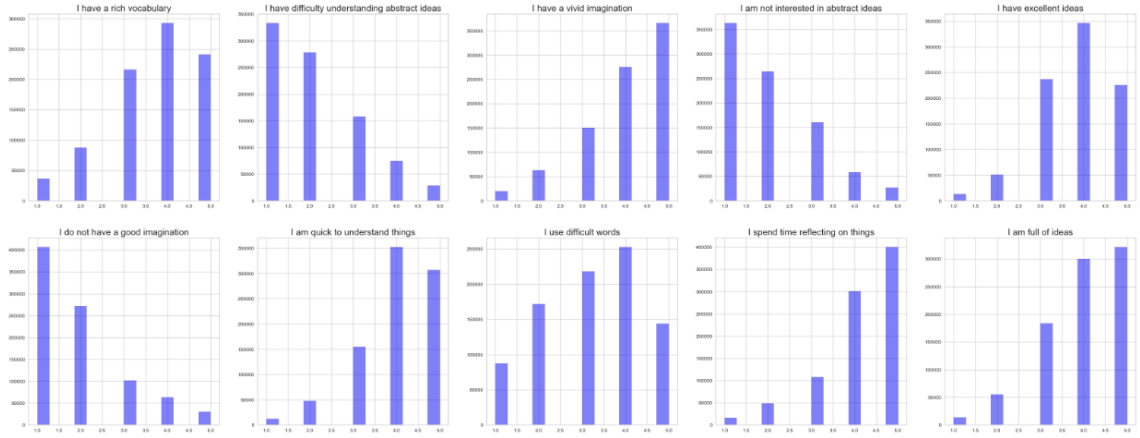
Q&As Related to Agreeable Personality



In the dataset, an equivalent number of individuals display traits associated with lower agreeableness as those demonstrating a higher degree of agreeableness.

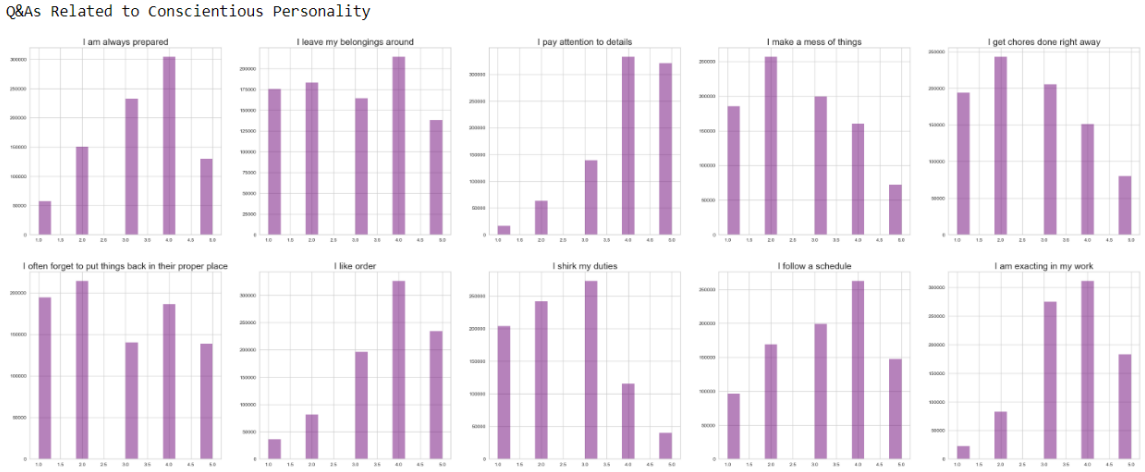
Openness Personality:

Q&As Related to Open Personality



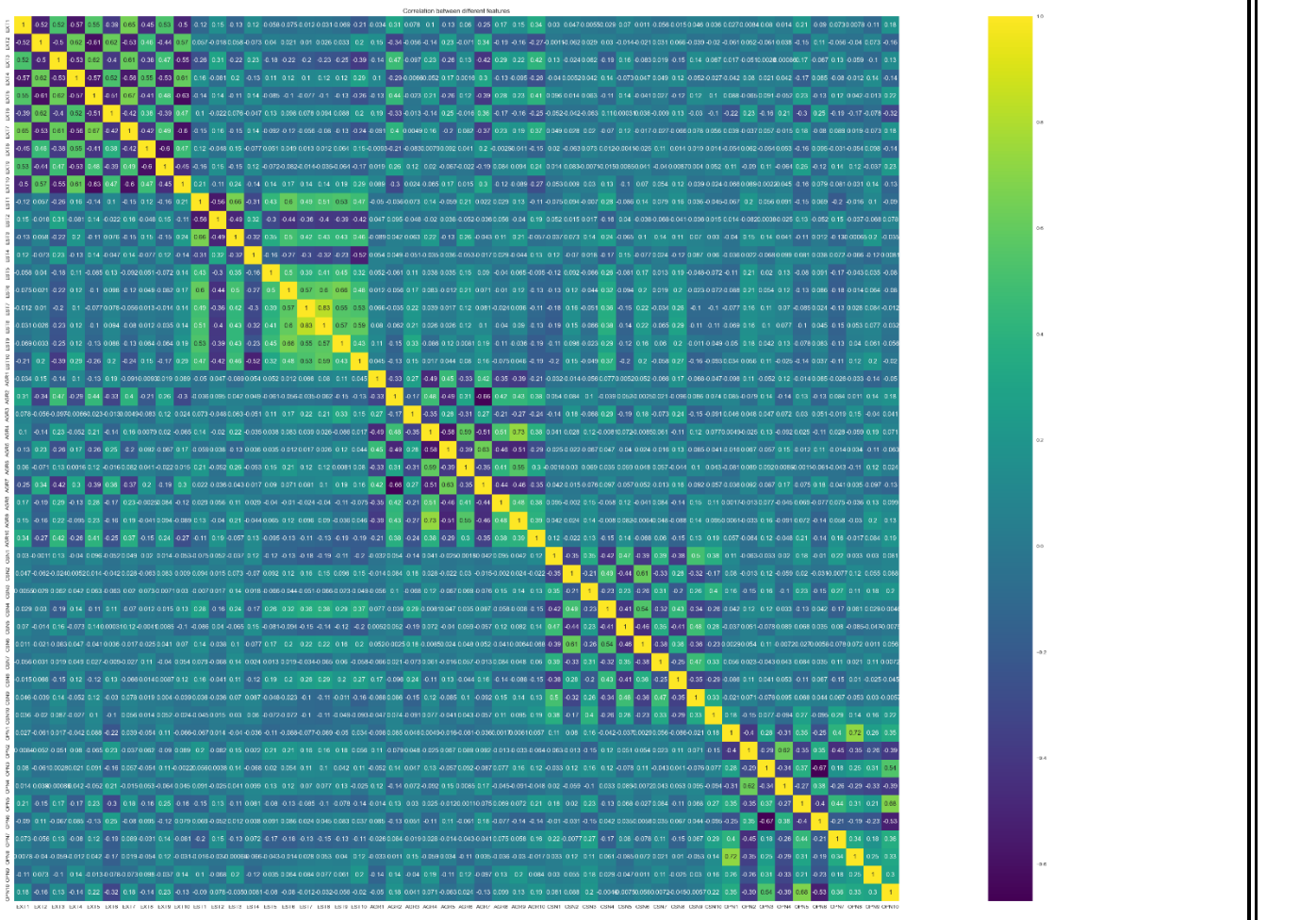
In the dataset, the number of individuals displaying traits associated with lower openness are less than those demonstrating a higher degree of openness.

Conscientiousness Personality:



Within the dataset, a higher number of individuals exhibit elevated level of conscientiousness, demonstrating a proactive and well-prepared approach towards their work and surroundings.

The correlation plot of the data looks as below



Based on the observed correlation heatmap, it can be inferred that the variables within each personality trait exhibit minimal to negligible correlation with variables from other personality traits. Conversely, a notable degree of collinearity is evident among variables belonging to the same personality trait.

Within variables representing the same personality trait, substantial negative correlations are apparent, suggesting that these variables capture the opposite characteristics of the respective personality trait.

Chapter-4

Dimensionality Reduction

In this chapter, we delve into the critical aspect of dimensionality reduction, a crucial step undertaken to address the challenge of the curse of dimensionality within our dataset. With a wealth of approximately 50 variables capturing diverse facets of personality traits, the need for dimensionality reduction becomes important for enhanced interpretability and model efficiency.

The curse of dimensionality arises when datasets encompass an abundance of features, potentially leading to increased computational complexity, greater susceptibility to noise, and challenges in model generalization.

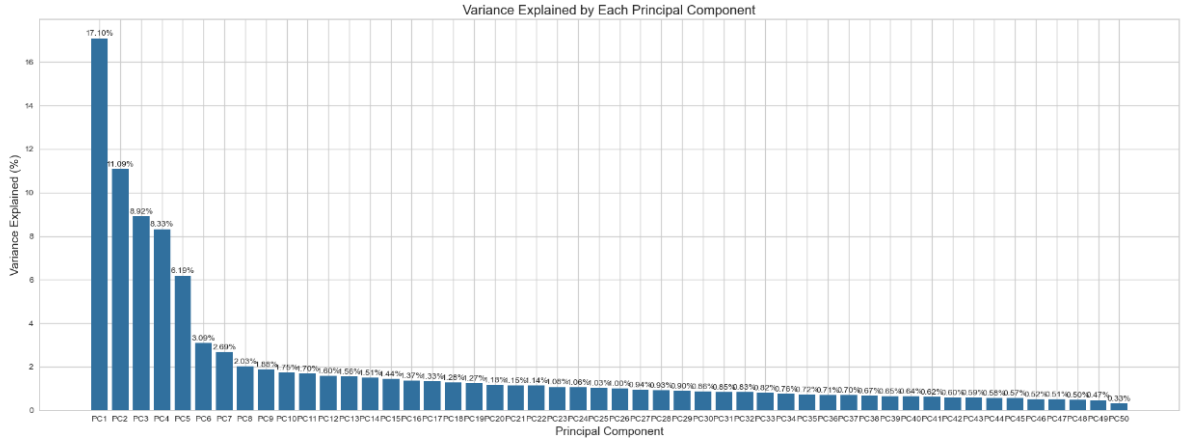
To mitigate these issues, we applied dimensionality reduction techniques, specifically Principal Component Analysis (PCA) and Factor Analysis, with the primary goal of distilling the essential information embedded in our dataset.

1. Principal Component Analysis (PCA)

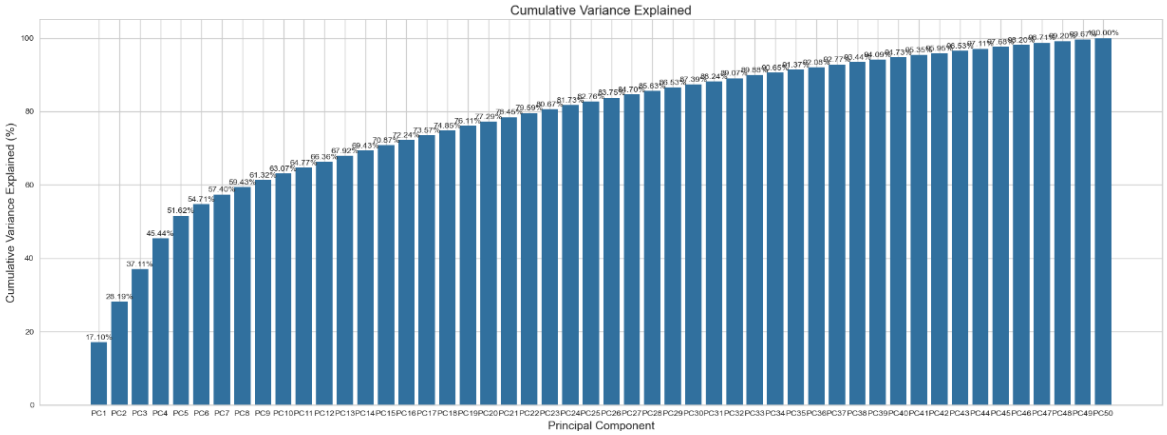
PCA is a widely-used linear transformation technique that aims to project the original variables into a new set of uncorrelated variables, known as principal components. These components are ordered by their variance, allowing us to retain the most significant information while discarding less informative dimensions.

To implement Principal Component Analysis (PCA), we initiated the process by computing the eigenvalues and eigenvectors of the correlation matrix derived from the dataset. Subsequently, we calculated the variances associated with each principal component, a crucial step in understanding the contribution of each component to the overall variance within the data.

Plotting the variance explained by each principal component as a bar plot we can see that the 1st principal component explains about 17% of the total sample variance of the data.



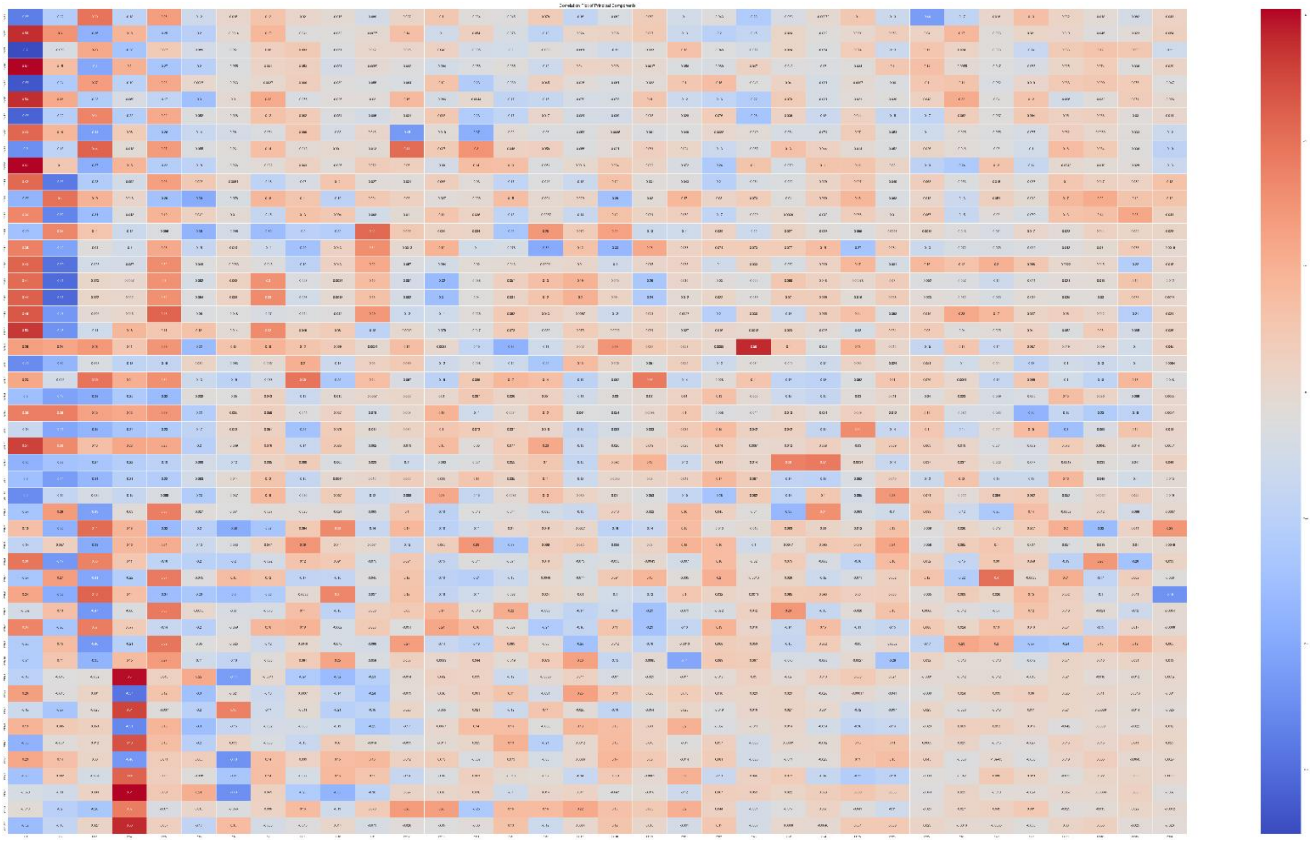
Following the computation of individual variances for each Principal Component (PC), we proceeded to calculate the cumulative variance of the PCs.



	Principal Component	Eigenvalues	Proportion	Cumulative	Eigenvalue Difference
0	PC1	8.547624	0.170952	0.170952	NaN
1	PC2	5.545133	0.110903	0.281855	3.002491
2	PC3	4.461101	0.089222	0.371077	1.084031
3	PC4	4.163873	0.083277	0.454355	0.297228
4	PC5	3.093562	0.061871	0.516226	1.070312
5	PC6	1.545445	0.030909	0.547135	1.548116
6	PC7	1.345343	0.026907	0.574042	0.200102
7	PC8	1.015306	0.020306	0.594348	0.330037
8	PC9	0.942283	0.018846	0.613193	0.073022
9	PC10	0.874643	0.017493	0.630686	0.067640
10	PC11	0.848239	0.016965	0.647651	0.026405
11	PC12	0.797534	0.015951	0.663602	0.050704
12	PC13	0.779780	0.015596	0.679197	0.017754
13	PC14	0.753968	0.015079	0.694277	0.025812
14	PC15	0.719734	0.014395	0.708671	0.034233
15	PC16	0.684044	0.013681	0.722352	0.035690
16	PC17	0.665461	0.013309	0.735661	0.018583
17	PC18	0.640961	0.012819	0.748481	0.024501
18	PC19	0.632643	0.012653	0.761134	0.008317
19	PC20	0.589405	0.011788	0.772922	0.043238
20	PC21	0.576538	0.011531	0.784452	0.012867
21	PC22	0.572316	0.011446	0.795899	0.004222
22	PC23	0.538263	0.010765	0.806664	0.034053
23	PC24	0.532338	0.010647	0.817311	0.005925
24	PC25	0.512623	0.010252	0.827563	0.019715
25	PC26	0.497892	0.009958	0.837521	0.014731
26	PC27	0.471965	0.009439	0.846960	0.025926
27	PC28	0.465689	0.009314	0.856274	0.006276
28	PC29	0.450290	0.009006	0.865280	0.015399
29	PC30	0.430744	0.008615	0.873895	0.019546
30	PC31	0.423290	0.008466	0.882361	0.007454
31	PC32	0.415475	0.008310	0.890670	0.007815
32	PC33	0.408416	0.008168	0.898838	0.007059
33	PC34	0.381645	0.007633	0.906471	0.026771
34	PC35	0.361819	0.007236	0.913708	0.019826
35	PC36	0.353143	0.007063	0.920771	0.008676
36	PC37	0.347572	0.006951	0.927722	0.005571
37	PC38	0.335938	0.006719	0.934441	0.011634
38	PC39	0.325248	0.006505	0.940946	0.010690
39	PC40	0.319112	0.006382	0.947328	0.006136
40	PC41	0.308690	0.006174	0.953502	0.010422
41	PC42	0.297527	0.005951	0.959452	0.011163
42	PC43	0.293477	0.005870	0.965322	0.004050
43	PC44	0.288697	0.005774	0.971096	0.004780
44	PC45	0.282974	0.005659	0.976755	0.005723
45	PC46	0.260261	0.005205	0.981961	0.022713
46	PC47	0.254998	0.005100	0.987061	0.005263
47	PC48	0.247606	0.004952	0.992013	0.007393
48	PC49	0.235122	0.004702	0.996715	0.012483
49	PC50	0.164246	0.003285	1.000000	0.070877

In the quest to retain a substantial portion of the dataset's variability, we strategically selected a subset of PCs. Specifically, we identified that incorporating the first 34 Principal Components is imperative to capture and explain at least 90% of the variance inherent in the dataset.

The correlation plot for the obtained Principal components with the original variables is as below.



Analyzing the correlation plot of PC's versus variables we can make inferences such as,

About First Principal Component (PC1):

- PC1 has a strong negative correlation with the OPN and AGR variables. These negative correlations indicate that as PC1 increases, these variables tend to decrease.
- It has a strong positive correlation with the EXT and EST variables. As PC1 increases, these variables also tend to increase.
- The strongest positive correlation is with the EXT variable, indicating that PC1 is primarily influenced by EXT

About Second Principal Component (PC2):

- PC2 has a strong negative correlation with the EST. As PC2 increases, these variables tend to decrease.
- It has a positive correlation with only EXT and OPN variables. These variables tend to increase as PC2 increases.

So, we have transformed the data using the principal components from 50 variables to 34 variables. The data looks as below.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	...	PC30	PC31	PC32	PC33	PC34
0	-6.740144	-9.645908	-3.570161	5.372790	5.212012	-13.661810	-8.509491	2.022773	1.644417	-1.686415	...	-2.043075	1.895830	0.396983	-0.446736	-0.882582
1	-0.930734	-7.425100	-10.058552	2.750963	1.899160	-14.953651	-6.048809	2.551719	2.313702	-2.381242	...	-2.487135	1.018172	0.344015	-0.308527	-0.304008
2	-2.016013	-7.747122	-8.249922	4.545834	2.594674	-12.156111	-8.616629	2.511561	0.017397	-2.463117	...	-2.670122	1.696684	0.933794	0.080417	-1.099263
3	-0.257502	-9.656332	-5.019804	5.355245	2.472016	-13.617841	-8.122965	1.419002	1.411677	-2.826405	...	-1.501310	0.825716	-0.107514	-1.215383	0.653088
4	-6.355684	-6.834854	-10.966968	6.056286	4.167571	-14.539712	-7.889471	2.132287	2.582928	-3.615549	...	-2.370184	2.636230	1.666525	-0.946358	-0.953288
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
874361	-1.587963	-10.258645	-4.900629	1.953408	5.548030	-16.360436	-6.690221	2.894728	-0.148174	-1.515173	...	-1.747241	2.279392	0.112009	-0.980634	-1.149979
874362	0.112310	-13.325992	-6.183535	5.587679	5.406569	-14.458970	-6.270640	4.298853	-1.108503	-2.696319	...	-2.389106	1.872226	1.001486	-1.415915	-1.025245
874363	-3.195207	-11.927986	-1.354716	7.710528	4.284800	-15.276889	-9.738225	1.219313	2.027067	-2.389295	...	-1.332829	1.261874	-0.379757	-0.712184	-1.506716
874364	0.706863	-8.857455	-9.148025	5.433366	5.277012	-13.172720	-9.207168	1.936377	0.056707	-3.538617	...	-2.313047	1.282115	0.509447	-1.332639	-0.402599
874365	-4.287251	-11.325369	-4.776425	6.057635	4.175622	-13.884846	-8.171094	1.484845	-0.018375	-2.577050	...	-1.916103	1.289488	-0.329924	-1.532864	-0.535193

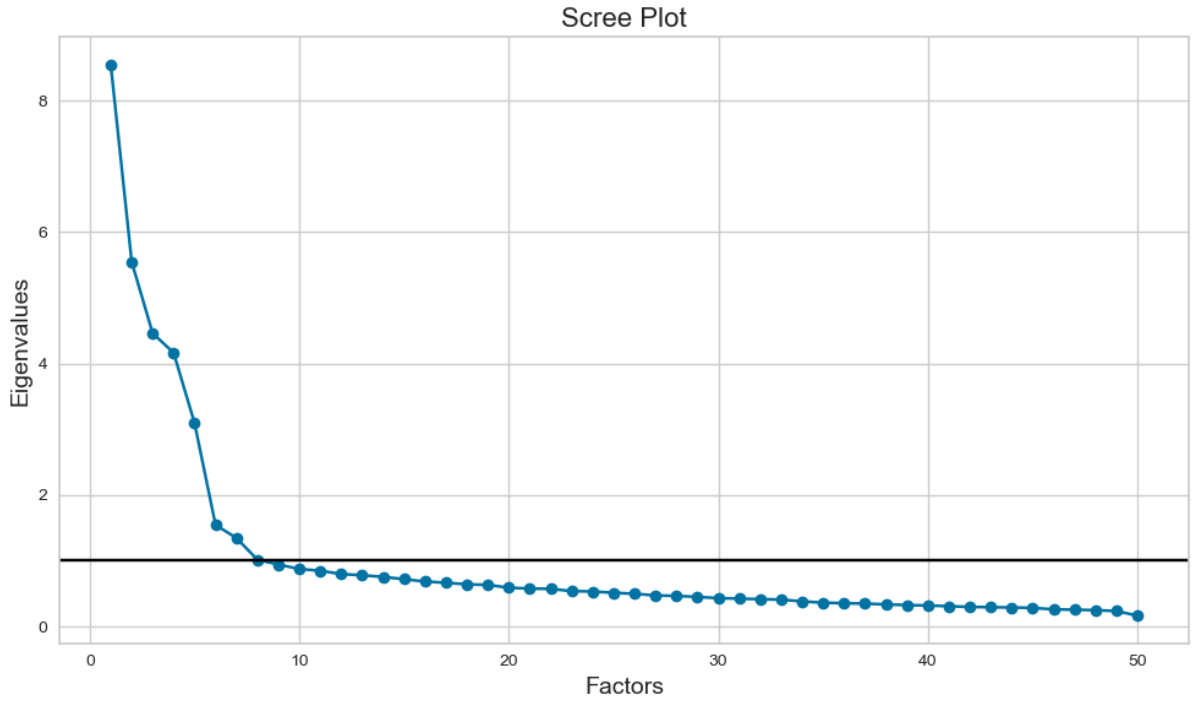
874366 rows × 34 columns

2. Factor Analysis

Factor Analysis, on the other hand, is a statistical method that explores the underlying structure of observed variables by grouping them into latent factors. These factors represent the commonalities shared among the original variables, providing a more optimal representation of the data.

In the case of Factor Analysis, our approach involved a similar foundational process. We initiated the procedure by finding the eigenvalues and eigenvectors for the correlation matrix of the original dataset.

Then we plotted the Scree plot to visualize the values of eigenvalues obtained.



Upon inspecting the scree plot, it is evident that eight eigenvalues surpass the threshold of 1. This observation leads us to the inference that employing eight latent factors adequately captures and explains the underlying patterns within the dataset.

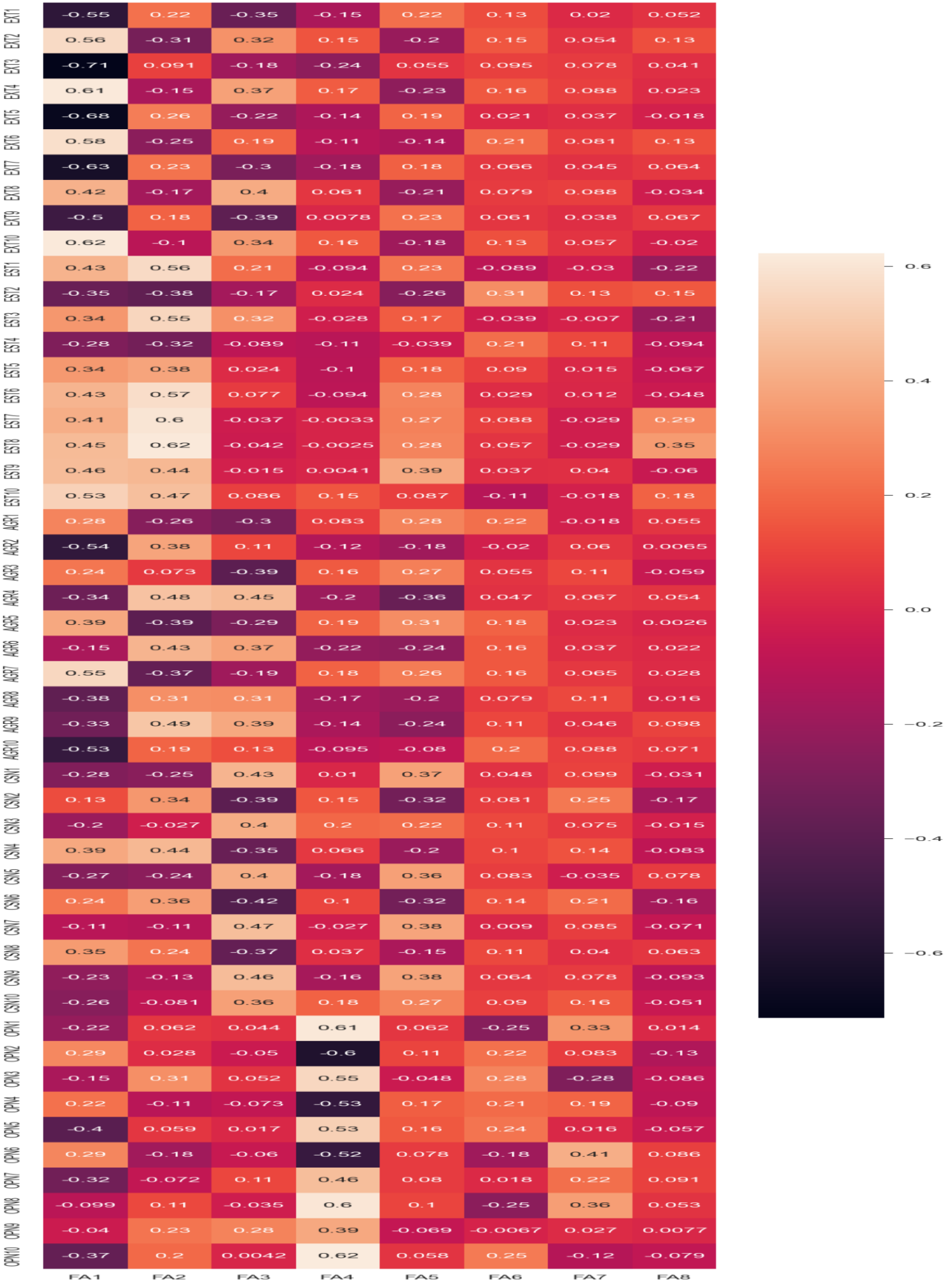
Next, we followed the procedure by estimating the factor loadings and communalities, crucial components of Factor Analysis with 8 factors. Subsequently, we determined the cumulative variance explained by each factor.

	FA1	FA2	FA3	FA4	FA5	FA6	\
Variance	8.106437	5.120328	3.955900	3.704324	2.617130	1.054934	
Proportional Var	0.162129	0.102407	0.079118	0.074086	0.052343	0.021099	
Cumulative Var	0.162129	0.264535	0.343653	0.417740	0.470082	0.491181	

	FA7	FA8
Variance	0.893547	0.595751
Proportional Var	0.017871	0.011915
Cumulative Var	0.509052	0.520967

So, the all the 8 factors combined explains about 52% of the sample variance which is satisfactory to proceed further.

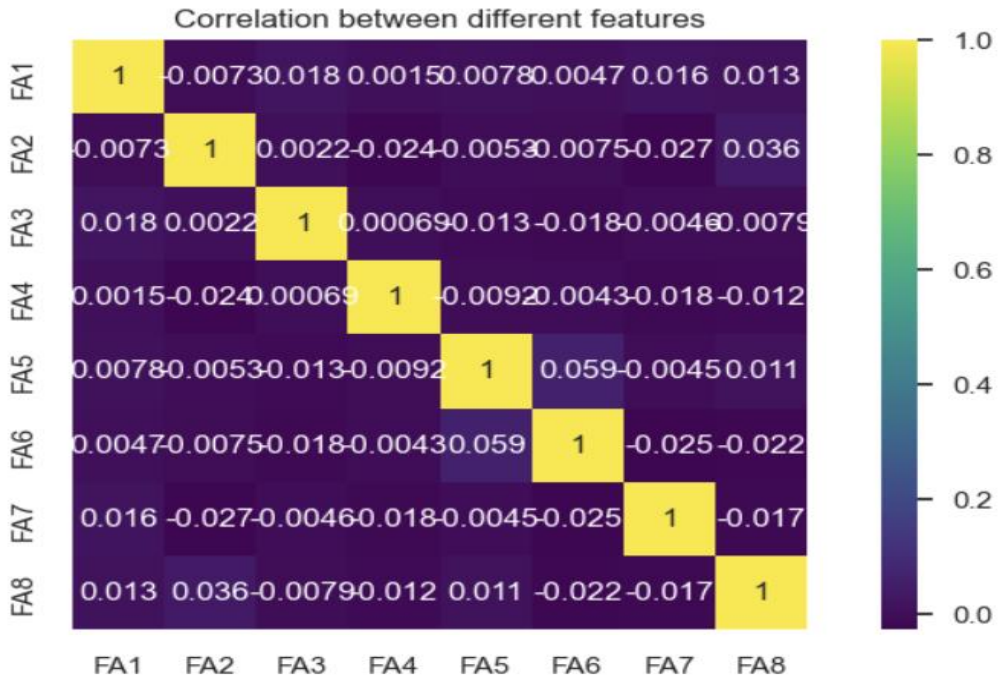
Next, we have plotted the heatmap of the loadings between the variables and the latent factors.


EXT1 -0.55 0.22 -0.35 -0.15 0.22 0.13 0.02 0.052
EXT2 0.56 -0.31 0.32 0.15 -0.2 0.15 0.054 0.13
EXT3 -0.71 0.091 -0.18 -0.24 0.055 0.095 0.078 0.041
EXT4 0.61 -0.15 0.37 0.17 -0.23 0.16 0.088 0.023
EXT5 -0.68 0.26 -0.22 -0.14 0.19 0.021 0.037 -0.018
EXT6 0.58 -0.25 0.19 -0.11 -0.14 0.21 0.081 0.13
EXT7 -0.63 0.23 -0.3 -0.18 0.18 0.066 0.045 0.064
EXT8 0.42 -0.17 0.4 0.061 -0.21 0.079 0.088 -0.034
EXT9 -0.5 0.18 -0.39 0.0078 0.23 0.061 0.038 0.067
EXT10 0.62 -0.1 0.34 0.16 -0.18 0.13 0.057 -0.02
EST1 0.43 0.56 0.21 -0.094 0.23 -0.089 -0.03 -0.22
EST2 -0.35 -0.38 -0.17 0.024 -0.26 0.31 0.13 0.15
EST3 0.34 0.55 0.32 -0.028 0.17 -0.039 -0.007 -0.21
EST4 -0.28 -0.32 -0.089 -0.11 -0.039 0.21 0.11 -0.094
EST5 0.34 0.38 0.024 -0.1 0.18 0.09 0.015 -0.067
EST6 0.43 0.57 0.077 -0.094 0.28 0.029 0.012 -0.048
EST7 0.41 0.6 -0.037 -0.0033 0.27 0.088 -0.029 0.29
EST8 0.45 0.62 -0.042 -0.0025 0.28 0.057 -0.029 0.35
EST9 0.46 0.44 -0.015 0.0041 0.39 0.037 0.04 -0.06
EST10 0.53 0.47 0.086 0.15 0.087 -0.11 -0.018 0.18
AGR1 0.28 -0.26 -0.3 0.083 0.28 0.22 -0.018 0.055
AGR2 -0.54 0.38 0.11 -0.12 -0.18 -0.02 0.06 0.0065
AGR3 0.24 0.073 -0.39 0.16 0.27 0.055 0.11 -0.059
AGR4 -0.34 0.48 0.45 -0.2 -0.36 0.047 0.067 0.054
AGR5 0.39 -0.39 -0.29 0.19 0.31 0.18 0.023 0.0026
AGR6 -0.15 0.43 0.37 -0.22 -0.24 0.16 0.037 0.022
AGR7 0.55 -0.37 -0.19 0.18 0.26 0.16 0.065 0.028
AGR8 -0.38 0.31 0.31 -0.17 -0.2 0.079 0.11 0.016
AGR9 -0.33 0.49 0.39 -0.14 -0.24 0.11 0.046 0.098
AGR10 -0.53 0.19 0.13 -0.095 -0.08 0.2 0.088 0.071
CSN1 -0.28 -0.25 0.43 0.01 0.37 0.048 0.099 -0.031
CSN2 0.13 0.34 -0.39 0.15 -0.32 0.081 0.25 -0.17
CSN3 -0.2 -0.027 0.4 0.2 0.22 0.11 0.075 -0.015
CSN4 0.39 0.44 -0.35 0.066 -0.2 0.1 0.14 -0.083
CSN5 -0.27 -0.24 0.4 -0.18 0.36 0.083 -0.035 0.078
CSN6 0.24 0.36 -0.42 0.1 -0.32 0.14 0.21 -0.16
CSN7 -0.11 -0.11 0.47 -0.027 0.38 0.009 0.085 -0.071
CSN8 0.35 0.24 -0.37 0.037 -0.15 0.11 0.04 0.063
CSN9 -0.23 -0.13 0.46 -0.16 0.38 0.064 0.078 -0.093
CSN10 -0.26 -0.081 0.36 0.18 0.27 0.09 0.16 -0.051
CPN1 -0.22 0.062 0.044 0.61 0.062 -0.25 0.33 0.014
CPN2 0.29 0.028 -0.05 -0.6 0.11 0.22 0.083 -0.13
CPN3 -0.15 0.31 0.052 0.55 -0.048 0.28 -0.28 -0.086
CPN4 0.22 -0.11 -0.073 -0.53 0.17 0.21 0.19 -0.09
CPN5 -0.4 0.059 0.017 0.53 0.16 0.24 0.016 -0.057
CPN6 0.29 -0.18 -0.06 -0.52 0.078 -0.18 0.41 0.086
CPN7 -0.32 -0.072 0.11 0.46 0.08 0.018 0.22 0.091
CPN8 -0.099 0.11 -0.035 0.6 0.1 -0.25 0.36 0.053
CPN9 -0.04 0.23 0.28 0.39 -0.069 -0.0067 0.027 0.0077
CPN10 -0.37 0.2 0.0042 0.62 0.058 0.25 -0.12 -0.079
FA1 FA2 FA3 FA4 FA5 FA6 FA7 FA8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

Upon reviewing the factor loadings depicted in the above plot, several insightful inferences can be drawn:

- Factor 1: This factor exhibits high loadings on Extroversion and Neuroticism, suggesting that it predominantly explains and captures the variance associated with these personality traits.
- Factor 2: Highlighting significant loadings on Neuroticism and Agreeableness, Factor 2 provides valuable insights into the interplay between these two personality traits within the dataset.
- Factor 3: This factor primarily explains Consciousness, with a slight influence on Agreeableness. Its high loadings on Consciousness signify its substantial contribution to understanding this particular personality trait.
- Factor 4: Factor 4 demonstrates exclusive high loadings on Openness, indicating that it singularly explains and captures the variance associated with the Openness personality trait.
- Factor 5: This factor mildly explains Consciousness and also Agreeableness.
- Factor 6 and 7: These explain more about Openness compared to any other personalities.
- Factor 8: It provides insights solely on Neuroticism with very little loadings on other personalities.

The correlation plot between the factors shows that there is very little to no correlation between the factors.



These findings provide a nuanced understanding of how specific factors contribute to the overall structure and variation within the dataset, shedding light on the relationships between different personality traits.

So, after transforming the data onto the 8 factors, the data looks as below.

	FA1	FA2	FA3	FA4	FA5	FA6	FA7	FA8
0	1.543865	-0.736605	-0.065173	-0.308339	0.357328	0.714814	-0.098322	-0.267355
1	-1.290163	-0.966562	0.975494	0.340800	0.316376	-1.517600	-0.290805	-0.107430
2	-0.636101	-0.665617	0.658827	0.060574	-0.653060	0.789287	-0.974619	-0.568594
3	-0.285760	-0.380026	0.062146	-1.206870	0.050301	0.111305	-0.549913	-0.963379
4	-0.343385	-1.074709	1.001527	1.559618	0.994322	0.721402	0.016529	-0.172788
...	...	...	...	...	...	...	...	...
872073	0.558499	0.127424	0.193634	-0.318928	0.033649	-1.341526	0.321860	0.680368
872074	0.239812	1.133581	0.042198	-0.552050	0.826398	-0.048651	-0.680654	0.882193
872075	1.083294	-0.204977	-0.495414	-1.655334	0.778304	1.234306	0.930857	-0.540020
872076	-0.896755	0.312947	0.051046	0.526283	-0.579631	0.978052	-0.264941	-0.816936
872077	0.796549	-0.181654	0.202138	-0.654581	0.768917	0.598554	0.092703	-0.980940

Chapter-5

K-Means Clustering

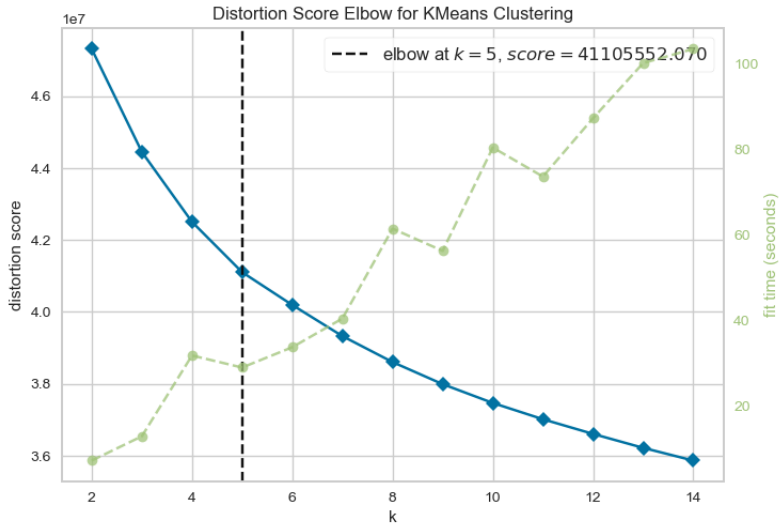
In this section, we explore the application of K-Means clustering technique to the pre-processed dataset, which underwent dimensionality reduction using PCA and Factor Analysis. The objective is to group individuals with similar behaviour or personality traits, paving the way for personalized roommate recommendations based on cluster characteristics.

K-means clustering, a versatile unsupervised learning algorithm, was chosen for its ability to group individuals with similar personality traits into cohesive clusters. By employing this method, our goal is to identify patterns and similarities in behaviour that extend beyond individual personality traits, thereby enhancing the roommate matching process.

On PCA Data

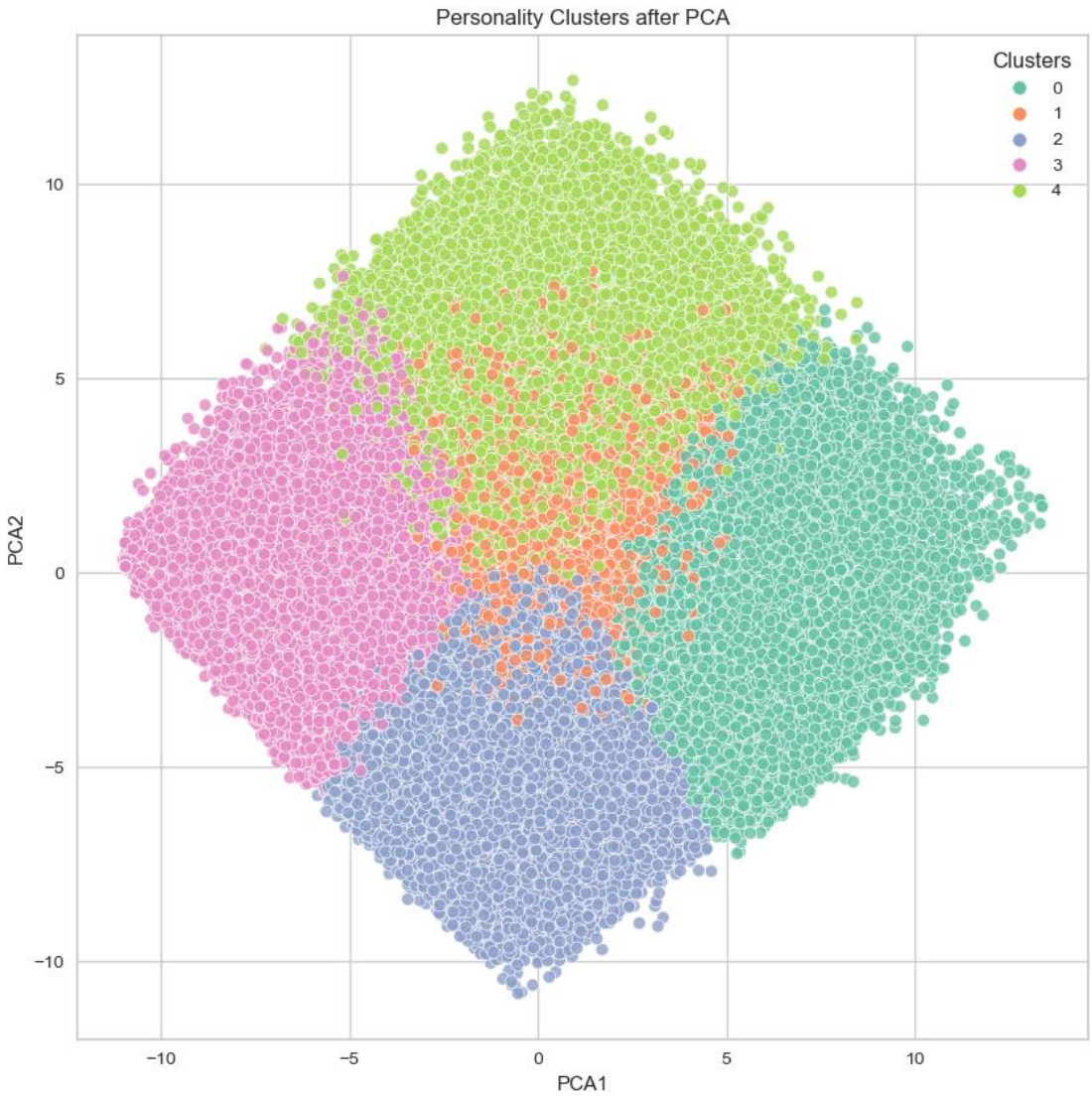
In our pursuit of meaningful clusters within the Big Five personality dataset, we employed a straightforward yet effective approach to determine the optimal number of clusters for our K-means clustering analysis, Elbow Method, which aids in finding a balance between the number of clusters and their within-cluster variance.

Upon conducting the Elbow Method on our data reduced through PCA, we observed a distinct elbow in the plot, indicating that the reduction in within-cluster variance slowed down after a certain point. In our case, this pivotal point occurred at 5 clusters.



With the optimal number of clusters identified, we proceeded to apply K-means clustering to categorize individuals into 5 distinct groups. Each cluster encapsulates a unique combination of personality traits, providing a nuanced understanding of behavioural

patterns within our dataset. The formed clusters visualized on the first two principal components seems to have formed well defined clusters.



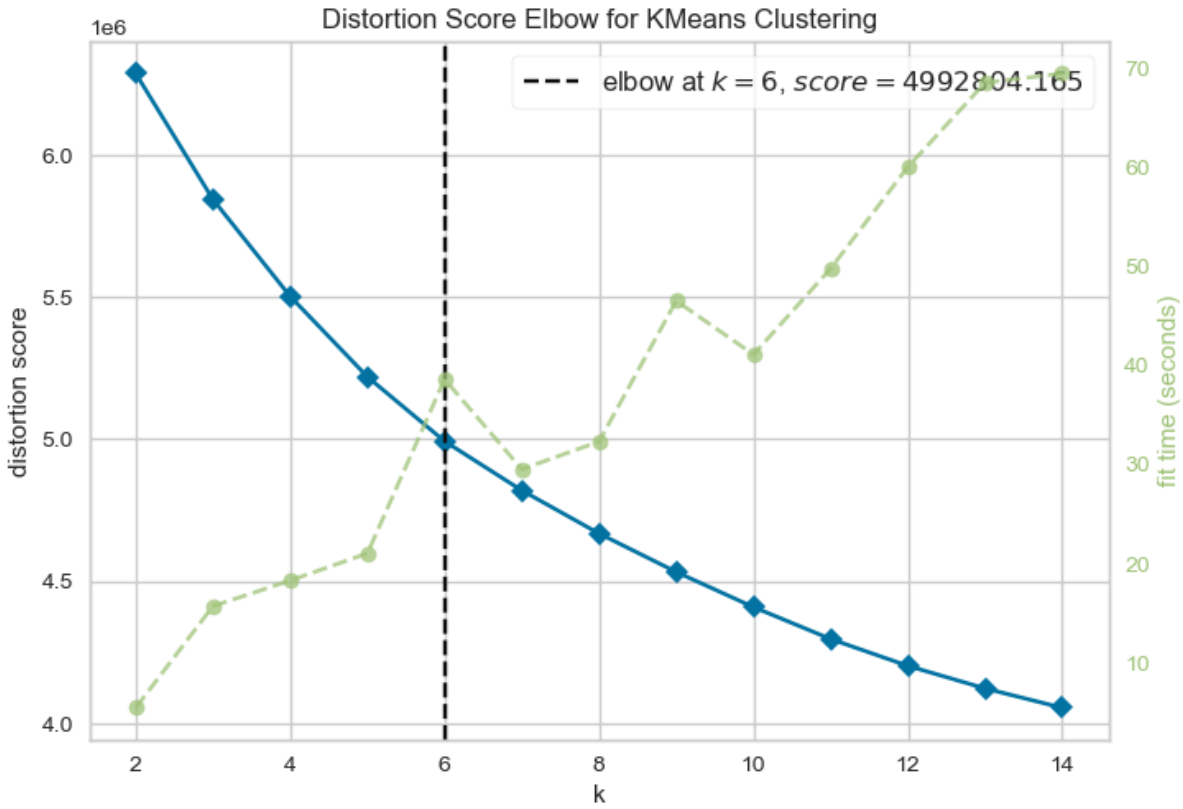
Analyzing the formed clusters properties through their mean distribution we find the individuals with low PC1 and PC2 are in cluster 2 and 3 stating those clusters contain individuals exhibiting less extroversion and neuroticism tendencies and are less in openness and agreeableness. Cluster 0 contains those that exhibit less neuroticism and are more openness and agreeableness.

```
pd.options.display.max_columns = 150
df_pca_fit.groupby('Clusters').mean()
```

Clusters	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14
0	2.777149	-10.769627	-7.015276	4.803473	4.710265	-14.815689	-8.115465	2.318966	0.896739	-2.412107	-0.659302	1.818834	1.064458	-2.330180
1	-1.291882	-8.666492	-8.996160	3.824754	4.731630	-14.556246	-7.882996	2.275844	0.895948	-2.638361	-0.734656	1.796890	1.253115	-2.320264
2	-2.229883	-12.126792	-5.120163	4.289381	5.601033	-14.747856	-8.244503	2.341595	1.032242	-2.390725	-0.757220	1.923384	1.209987	-2.393438
3	-5.754632	-8.306035	-6.158839	4.044750	4.866501	-14.545162	-7.999733	2.266904	1.001484	-2.661633	-0.861849	1.975653	1.437275	-2.365363
4	0.017563	-6.214196	-5.371759	5.464894	4.752578	-14.846188	-8.102785	2.241257	1.156080	-2.524314	-0.800084	1.915608	1.307833	-2.327936

On Factor Analysis Data

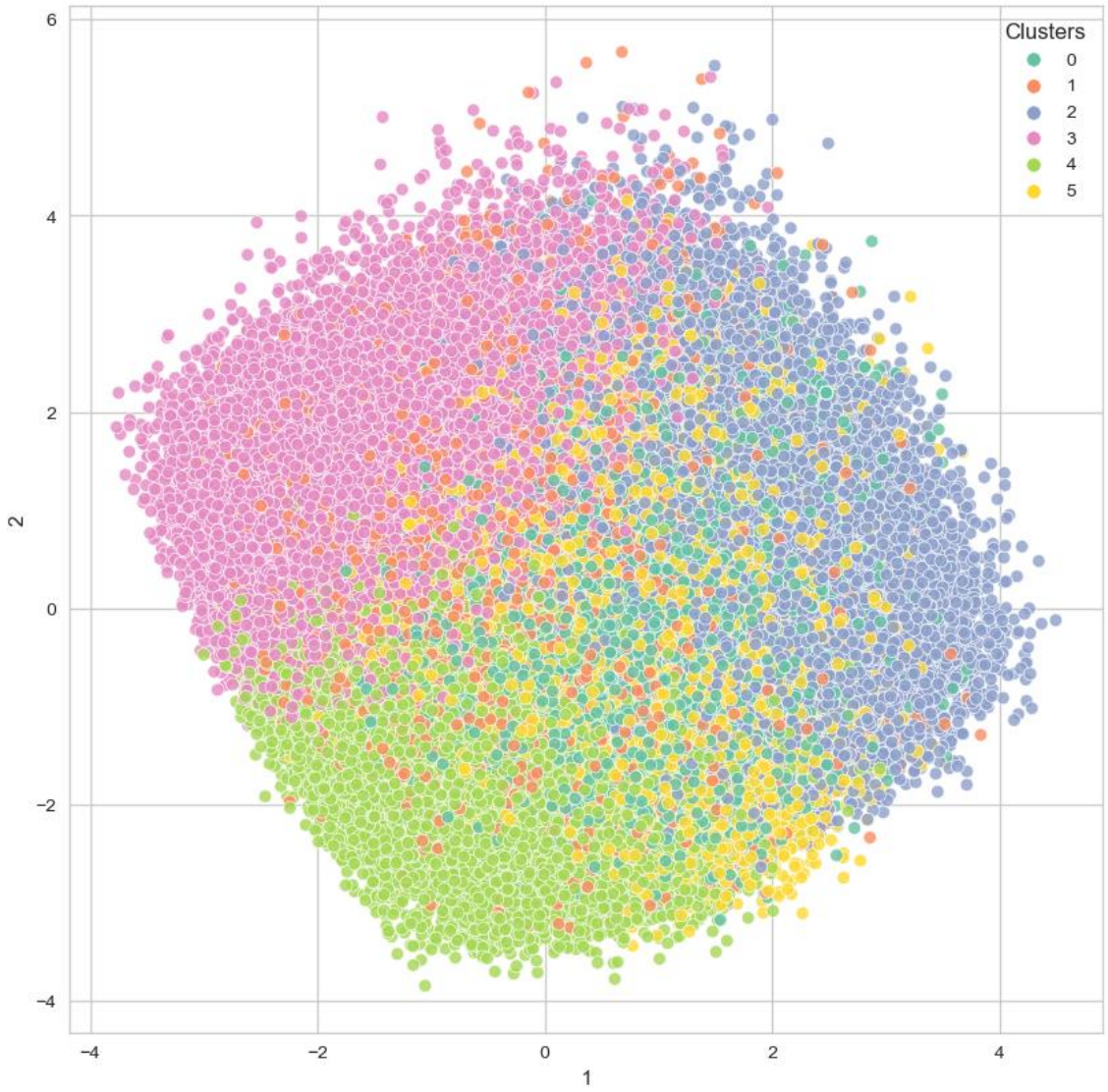
Upon conducting the Elbow Method on our data reduced through Factor Analysis, we observed a distinct elbow in the plot, indicating that the reduction in within-cluster variance slowed down after a certain point. In our case, this pivotal point occurred at 6 clusters.



With the optimal number of clusters identified, we proceeded to apply K-means clustering to categorize individuals into 6 distinct groups. Each cluster encapsulates a unique combination of personality traits, providing a nuanced understanding of behavioural patterns within our dataset.

Analysing the clusters through their centroids show that the centroids of each cluster are as below

Clusters	FA1	FA2	FA3	FA4	FA5	FA6	FA7	FA8
0	-0.221096	0.075931	-0.084693	-0.058579	-1.228506	-0.492000	0.055289	0.006779
1	-0.896867	-0.132220	0.429643	-0.562551	0.319592	0.232239	0.039655	-0.275185
2	-0.170558	0.884326	0.192831	0.550987	0.296082	-0.094454	-0.167647	0.519488
3	-0.256333	-0.219958	-1.560732	0.150706	0.327294	0.336561	0.030298	-0.135847
4	0.945258	0.315861	0.149873	-0.767335	0.192276	0.165565	0.111654	0.009300
5	0.456074	-0.803458	0.404382	0.626829	0.147202	-0.059106	-0.055636	-0.119560



The plot of Factor 2 vs Factor 3 shows that Cluster 3 has low values of Factor 1 and high values of Factor 3 and as Factor 3 primarily explains Consciousness, with a slight influence on Agreeableness, it indicates that the cluster contains individuals that exhibit consciousness traits and less extroversion and neuroticism traits.

The output of the clustering analysis revealed the formation of distinct personality clusters, each characterized by a unique combination of personality traits. This segmentation provides a nuanced understanding of individuals' behavioural tendencies, serving as a foundation for more personalized roommate recommendations.

CHAPTER-6

Roommate Recommendation

Building upon the insightful clusters formed through K-means analysis, we extended our approach to predict suitable roommates for new individuals entering our system. Leveraging the cluster assignments derived from the existing dataset, we employed a simple and effective strategy to identify compatible roommates for newcomers.

As new individuals join our dataset, we utilized the pre-established K-means clusters to predict the most suitable cluster for each new member. This predictive assignment is based on the inherent behavioural traits extracted through the clustering analysis

	EXT1	EXT2	EXT3	EXT4	EXT5	EXT6	EXT7	EXT8	EXT9	EXT10	EST1	EST2	EST3	EST4	EST5	EST6	EST7	EST8	EST9	EST10	AGR1	AGR2	AGI
0	2	4	1	4	2	3	2	4	3	5	1	4	2	3	2	3	4	3	2	3	4	2	

```
----------Using PCA----------
My Personality Cluster:  5
----------Using FA----------
My Personality Cluster:  2
```

To identify the most suitable roommates for individuals in each cluster, we employed the Euclidean distance metric to measure the similarity or dissimilarity between individuals based on their behaviour traits, allowing us to identify the nearest neighbours within each cluster.

For the new individual, we calculated the Euclidean distance to every other member within their predicted cluster. Subsequently, we ranked these distances and selected the top 10 individuals with the least distance as the most suitable roommate candidates. This approach ensures a personalized and effective roommate matching process.

```
df.iloc[idx_list1]
```

	EXT1	EXT2	EXT3	EXT4	EXT5	EXT6	EXT7	EXT8	EXT9	EXT10	EST1	EST2	EST3	EST4	EST5	EST6	EST7	EST8	EST9	EST10	AGR1	AGR2
798263	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	4.0	4.0	2.0	4.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	2.0	3.0
742658	3.0	4.0	2.0	4.0	3.0	3.0	3.0	2.0	3.0	4.0	4.0	3.0	4.0	3.0	3.0	3.0	4.0	3.0	3.0	4.0	2.0	3.0
149731	2.0	5.0	4.0	2.0	4.0	4.0	3.0	4.0	2.0	2.0	4.0	2.0	4.0	3.0	4.0	2.0	3.0	4.0	3.0	3.0	2.0	3.0
461279	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	4.0	3.0	4.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	2.0	4.0
170291	4.0	4.0	3.0	3.0	3.0	3.0	2.0	4.0	3.0	2.0	3.0	3.0	3.0	3.0	3.0	3.0	2.0	2.0	3.0	3.0	3.0	4.0
848445	3.0	3.0	4.0	2.0	2.0	3.0	4.0	4.0	4.0	4.0	3.0	2.0	4.0	4.0	2.0	3.0	2.0	1.0	3.0	4.0	1.0	4.0
438828	3.0	3.0	3.0	2.0	2.0	4.0	2.0	3.0	3.0	4.0	4.0	2.0	4.0	3.0	4.0	4.0	3.0	2.0	3.0	2.0	3.0	2.0
511373	3.0	3.0	3.0	2.0	4.0	3.0	3.0	3.0	3.0	3.0	3.0	2.0	4.0	2.0	3.0	3.0	3.0	2.0	3.0	2.0	2.0	3.0
431813	2.0	4.0	3.0	3.0	3.0	2.0	2.0	3.0	2.0	4.0	4.0	3.0	3.0	3.0	3.0	3.0	2.0	2.0	2.0	2.0	2.0	3.0
814357	3.0	2.0	2.0	3.0	3.0	2.0	2.0	3.0	3.0	2.0	4.0	2.0	4.0	2.0	3.0	4.0	3.0	3.0	4.0	4.0	1.0	3.0

```
my_data
```

	EXT1	EXT2	EXT3	EXT4	EXT5	EXT6	EXT7	EXT8	EXT9	EXT10	EST1	EST2	EST3	EST4	EST5	EST6	EST7	EST8	EST9	EST10	AGR1	AGR2	AGI
0	4	3	2	4	3	5	1	3	5	2	4	1	3	2	3	1	2	3	5	2	1	4	

CHAPTER-7

Conclusion:

After diving into the Big Five Personality dataset, we uncovered valuable insights to revolutionize how we pair roommates:

- Distinct Personalities: Identified six unique personality clusters, guiding us to understand and categorize individuals based on their behaviours.
- Slimming Down Data: PCA and Factor Analysis helped us make sense of the data overload, making our analysis more efficient and computationally easy.
- Smart Roommate Picks: Used a clever strategy to predict the best roommates for newcomers, ensuring they share similar personality traits.