Unsupervised Machine Learning on the Big Five Personality Test

By Group #1

Nirmal Bhuvaneswari Ezhilarasan

Manoj Cn

Hitarth Patel

Shashank Sant

Shawn Young

Table of Contents

1. Introduction

The understanding of the human psyche and personality provides an advantage to any organization or individual with the readily available data on hand. On a large scale it can assist any marketing company, business, or electoral candidate in how they target a specific audience for their advertising or campaigning. On a small scale, a salesman or presenter may change their approach on how they communicate with their customers and audience.

For our analysis into human psychology, we will analyze a Big Five Personality Test Dataset that contains the responses of the Big Five Personality Test from around the globe. This includes 50 questions that are scaled from 1 (Strongly Disagree) to 5 (Strongly Agree) and the estimated country from where it was taken.

The Big Five Personality Test aims to measure and understand individual differences in personality traits, specifically these five – Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism, collectively shortened to OCEAN. This test can be used for personal and professional development to target areas for improvement, psychological research to understand human behavior, and employment to evaluate job fit and leadership potential, among many other applications.

Our primary approach will be through using unsupervised machine learning methods and ensemble techniques to model and cluster the data into distinct groups and investigate whether there are specific differences between the participants of different countries.

2. Exploratory Data Analysis

2.1. Description of Dataset

The data was collected between 2016 and 2018 through an online personality test from a website (https://ipip.ori.org/newBigFive5broadKey.htm) and involves responses to the questions listed in the next section as well as other obtainable data from the test taker, IP location to determine the country and response times. We excluded the response times from this analysis. In total without cleaning the dataset includes 110 columns, and 1.02 million observations.

2.2. List of Variables

Extroversion
EXT1 I am the life of the party.
EXT2- I don't talk a lot.
EXT3 I feel comfortable around people.
EXT4- I keep in the background.
EXT5 I start conversations.
EXT6- I have little to say.
EXT7 I talk to a lot of different people at parties.
EXT8- I don't like to draw attention to myself.
EXT9 I don't mind being the center of attention.
EXT10- I am quiet around strangers.

Emotional Stability (Neuroticism)
EST1- I get stressed out easily.
EST2 I am relaxed most of the time.
EST3- I worry about things.
EST4 I seldom feel blue.
EST5- I am easily disturbed.
EST6- I get upset easily.
EST7- I change my mood a lot.
EST8- I have frequent mood swings.
EST9- I get irritated easily.
EST10- I often feel blue.

Agreeableness
AGR1- I feel little concern for others.
AGR2 I am interested in people.
AGR3- I insult people.
AGR4 I sympathize with others' feelings.
AGR5- I am not interested in other people's problems.
AGR6 I have a soft heart.
AGR7- I am not really interested in others.
AGR8 I take time out for others.
AGR9 I feel others' emotions.
AGR10 I make people feel at ease.

Conscientiousness
CSN1 I am always prepared.
CSN2- I leave my belongings around.
CSN3 I pay attention to details.
CSN4- I make a mess of things.
CSN5 I get chores done right away.
CSN6- I often forget to put things back in their proper place.
CSN7 I like order.
CSN8- I shirk my duties.
CSN9 I follow a schedule.
CSN10 I am exacting in my work.

Openness
OPN1 I have a rich vocabulary.
OPN2- I have difficulty understanding abstract ideas.
OPN3 I have a vivid imagination.
OPN4- I am not interested in abstract ideas.
OPN5 I have excellent ideas.
OPN6- I do not have a good imagination.
OPN7 I am quick to understand things.
OPN8 I use difficult words.
OPN9 I spend time reflecting on things.
OPN10 I am full of ideas.

There are variables for each of these questions ending in _E which represent the response times for each question.
-This was excluded from this analysis.

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 had answered accurately and their answers could be stored and used for research. This dataset only includes users who answered "Yes" to this question.
IPC	The number of records from the user's IP address in the dataset. Should be limited to 1 for cleanliness, as specified from the dataset author.
lat_appx_lots_of_err	approximate latitude of user. Not accurate, as specified from the dataset author.
long_appx_lots_of_err	approximate longitude of user. Not accurate, as specified from the dataset author.

Self Generated:

Extroversion Score = 20 + EXT1 - EXT2 + EXT3 - EXT4 + EXT5 - EXT6 + EXT7 - EXT8 + EXT9 - EXT10
Agreeableness Score = 14 - AGR1 + AGR2 - AGR3 + AGR4 - AGR5 + AGR6 - AGR7 + AGR8 + AGR9 + AGR10
Conscientiousness Score = 14 + CSN1 - CSN2 + CSN3 - CSN4 + CSN5 - CSN6 + CSN7 - CSN8 + CSN9 + CSN10
Neuroticism Score = 38 - EST1 + EST2 - EST3 + EST4 - EST5 - EST6 - EST7 - EST8 - EST9 - EST10
Openness Score = 8 + OPN1 - OPN2 + OPN3 - OPN4 + OPN5 - OPN6 + OPN7 + OPN8 + OPN9 + OPN10

These equations come from https://openpsychometrics.org/printable/big-five-personality-test.pdf
This also signifies which questions are negatively associated with the score, and will need to be flipped for this analysis.

Data Cleaning Methods:

- Removed any rows with missing values in the answer columns
- Removed rows with IPC greater than 1 – as suggested by the author for max cleanliness, higher values can include shared networks or multiple test takers
- Removed any rows with a value of 0 in the answer columns
- Removed rows with NONE in the country column
- Removed rows from countries with fewer than 1000 participants
- Converted the ISO-3166 country codes in the dataset to the ISO-A3 code used by geopandas
- Renamed columns negatively associated to scores by appending a hyphen '-'.
- Removed columns:
    - 50 reaction time columns ending in _E as this was not included in this analysis
    - IPC as all remaining values are 1
    - latitude and longitude columns as these are considered unaccurate
    - Unused columns:
        - screenw, screenh, introelapse, testelapse, endelapse

Resulting cleaned dataset consists of 62 columns and 570k observations

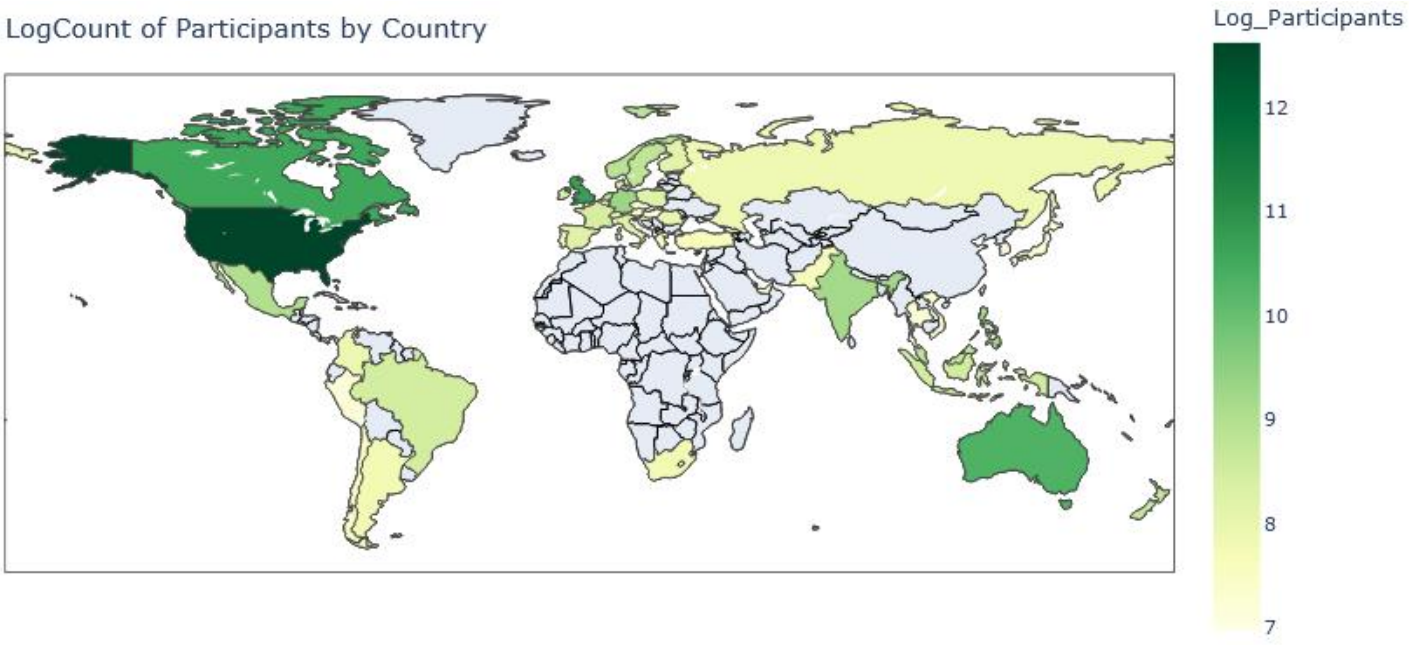
2.3. Summary Statistics

	count	mean	std	min	25%	50%	75%	max
EXT1	570277.00	2.58	1.24	1.00	1.00	3.00	4.00	5.00
EXT2-	570277.00	3.16	1.30	1.00	2.00	3.00	4.00	5.00
EXT3	570277.00	3.24	1.19	1.00	2.00	3.00	4.00	5.00
EXT4-	570277.00	2.78	1.21	1.00	2.00	3.00	4.00	5.00
EXT5	570277.00	3.25	1.25	1.00	2.00	3.00	4.00	5.00
EXT6-	570277.00	3.58	1.21	1.00	3.00	4.00	5.00	5.00
EXT7	570277.00	2.72	1.37	1.00	2.00	3.00	4.00	5.00
EXT8-	570277.00	2.53	1.24	1.00	1.00	2.00	3.00	5.00
EXT9	570277.00	2.95	1.32	1.00	2.00	3.00	4.00	5.00
EXT10-	570277.00	2.38	1.26	1.00	1.00	2.00	3.00	5.00
EST1-	570277.00	2.68	1.31	1.00	2.00	2.00	4.00	5.00
EST2	570277.00	3.17	1.19	1.00	2.00	3.00	4.00	5.00
EST3-	570277.00	2.12	1.12	1.00	1.00	2.00	3.00	5.00
EST4	570277.00	2.65	1.23	1.00	2.00	3.00	4.00	5.00
EST5-	570277.00	3.15	1.25	1.00	2.00	3.00	4.00	5.00
EST6-	570277.00	3.14	1.29	1.00	2.00	3.00	4.00	5.00
EST7-	570277.00	2.94	1.26	1.00	2.00	3.00	4.00	5.00
EST8-	570277.00	3.31	1.32	1.00	2.00	3.00	4.00	5.00
EST9-	570277.00	2.90	1.27	1.00	2.00	3.00	4.00	5.00
EST10-	570277.00	3.15	1.31	1.00	2.00	3.00	4.00	5.00
AGR1-	570277.00	3.78	1.30	1.00	3.00	4.00	5.00	5.00
AGR2	570277.00	3.86	1.08	1.00	3.00	4.00	5.00	5.00
AGR3-	570277.00	3.73	1.26	1.00	3.00	4.00	5.00	5.00
AGR4	570277.00	3.95	1.08	1.00	3.00	4.00	5.00	5.00
AGR5-	570277.00	3.70	1.15	1.00	3.00	4.00	5.00	5.00
AGR6	570277.00	3.76	1.16	1.00	3.00	4.00	5.00	5.00
AGR7-	570277.00	3.77	1.11	1.00	3.00	4.00	5.00	5.00
AGR8	570277.00	3.69	1.04	1.00	3.00	4.00	4.00	5.00
AGR9	570277.00	3.79	1.14	1.00	3.00	4.00	5.00	5.00
AGR10	570277.00	3.60	1.03	1.00	3.00	4.00	4.00	5.00
CSN1	570277.00	3.32	1.12	1.00	3.00	3.00	4.00	5.00
CSN2-	570277.00	3.00	1.37	1.00	2.00	3.00	4.00	5.00
CSN3	570277.00	4.01	0.99	1.00	4.00	4.00	5.00	5.00
CSN4-	570277.00	3.35	1.23	1.00	2.00	3.00	4.00	5.00
CSN5	570277.00	2.58	1.24	1.00	2.00	2.00	4.00	5.00
CSN6-	570277.00	3.14	1.40	1.00	2.00	3.00	4.00	5.00
CSN7	570277.00	3.74	1.07	1.00	3.00	4.00	5.00	5.00
CSN8-	570277.00	3.52	1.13	1.00	3.00	4.00	4.00	5.00
CSN9	570277.00	3.17	1.25	1.00	2.00	3.00	4.00	5.00
CSN10	570277.00	3.63	1.00	1.00	3.00	4.00	4.00	5.00
OPN1	570277.00	3.78	1.07	1.00	3.00	4.00	5.00	5.00
OPN2-	570277.00	3.98	1.08	1.00	3.00	4.00	5.00	5.00
OPN3	570277.00	4.06	1.03	1.00	3.00	4.00	5.00	5.00
OPN4-	570277.00	4.05	1.06	1.00	3.00	4.00	5.00	5.00
OPN5	570277.00	3.83	0.93	1.00	3.00	4.00	5.00	5.00
OPN6-	570277.00	4.12	1.07	1.00	4.00	4.00	5.00	5.00
OPN7	570277.00	4.06	0.92	1.00	4.00	4.00	5.00	5.00
OPN8	570277.00	3.29	1.21	1.00	2.00	3.00	4.00	5.00
OPN9	570277.00	4.23	0.94	1.00	4.00	4.00	5.00	5.00
OPN10	570277.00	4.00	0.98	1.00	3.00	4.00	5.00	5.00
screenw	570277.00	1118.49	583.41	0.00	375.00	1280.00	1440.00	13660.00
screenh	570277.00	823.42	185.69	0.00	667.00	768.00	900.00	8802.00
introelapse	570277.00	1034.09	60619.43	0.00	5.00	11.00	31.00	29443071.00
testelapse	570277.00	630.77	16675.45	1.00	169.00	218.00	300.00	5372971.00
endelapse	570277.00	3815.78	1979117.76	1.00	9.00	12.00	17.00	1493327022.00
Extroversion Score	570277.00	19.18	9.13	0.00	12.00	19.00	26.00	40.00
Agreeableness Score	570277.00	27.65	7.37	0.00	23.00	29.00	33.00	40.00
Conscientiousness Score	570277.00	23.45	7.40	0.00	18.00	24.00	29.00	40.00
Neuroticism Score	570277.00	19.21	8.62	0.00	13.00	19.00	25.00	40.00
Openness Score	570277.00	29.41	6.19	0.00	25.00	30.00	34.00	40.00

2.4. Univariate EDA

2.4.1. Participants by Country

A representation of the participating countries and their number of participants is shown in the map of the world below. The count is highly skewed with a large majority being in the USA so the color scale was set to be logarithmic and shows a clearer picture of where the participants come from. In the dataset these country names are in the ISO-A3 format. It is evident that the majority of these participants are from English-speaking countries as the test questions were likely in English.

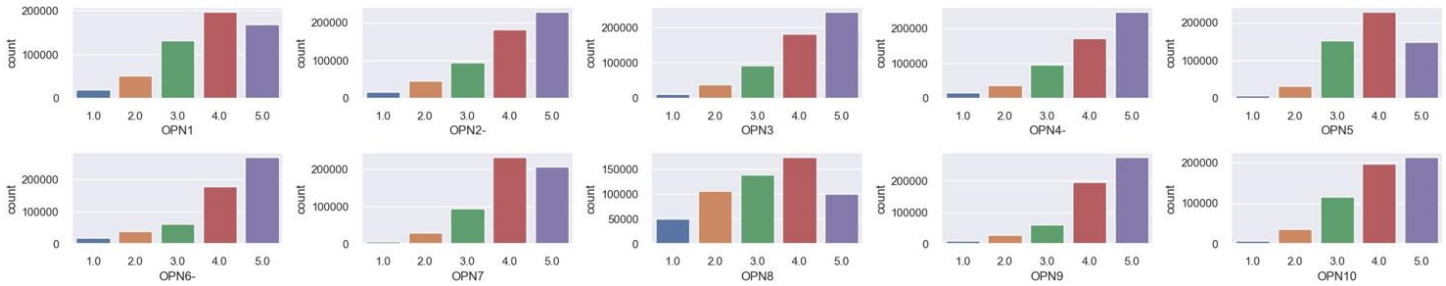


USA	300493	FRA	4497	PRT	1948
GBR	43411	ITA	3813	GRC	1931
CAN	38337	ESP	3719	CHE	1842
AUS	30302	POL	3621	ARE	1722
DEU	10895	IRL	3501	AUT	1717
IND	9847	FIN	3410	CHL	1707
PHL	9470	DNK	3345	HRV	1643
MEX	8275	ROU	2908	VNM	1471
NOR	7137	COL	2721	SRB	1431
NLD	6910	RUS	2536	JPN	1417
SWE	6526	BEL	2474	CZE	1363
MYS	6330	ARG	2430	THA	1362
NZL	5752	HKG	2421	PER	1295
IDN	4928	ZAF	2399	HUN	1169
SGP	4794	TUR	2063	KOR	1156
BRA	4767	PAK	1995	ISR	1076

2.4.2. Responses to the Questions

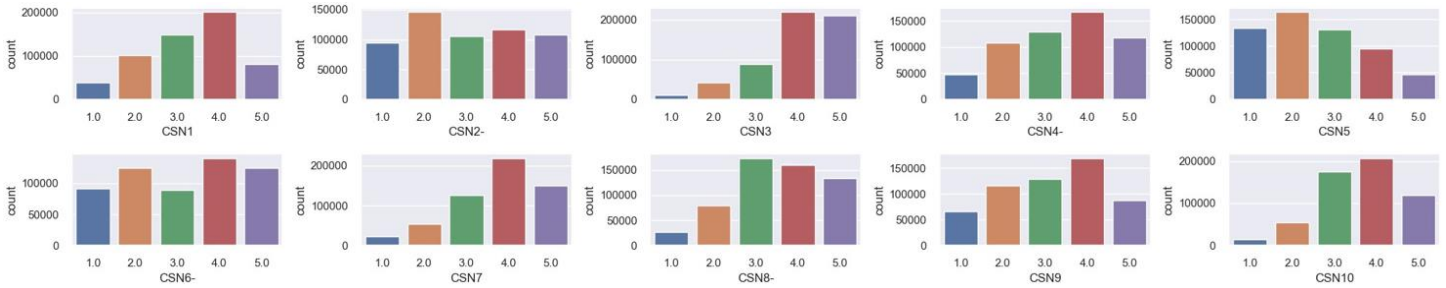
To note, the hyphen after the column name indicates the responses are flipped for positive/negative association.

2.4.2.1. Openness



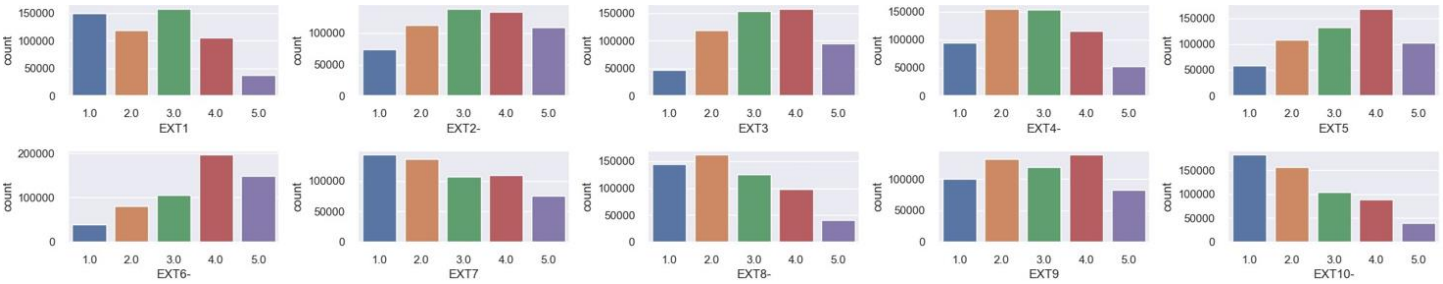
OPN1	I have a rich vocabulary.	By far, most people agree they have a rich vocabulary.
OPN2-	I have difficulty understanding abstract ideas.	By far, most people strongly disagree on difficulty with abstract ideas.
OPN3	I have a vivid imagination.	By far, most people strongly agree on having a vivid imagination.
OPN4-	I am not interested in abstract ideas.	By far, most people agree to being interested in abstract ideas.
OPN5	I have excellent ideas.	By far, most people think they have excellent ideas.
OPN6-	I do not have a good imagination.	By far, most people think they have a good imagination.
OPN7	I am quick to understand things.	By far, most people think they are quick to understand.
OPN8	I use difficult words.	Relatively even spread. More people agree they use difficult words.
OPN9	I spend time reflecting on things.	By far, most people reflect on things.
OPN10	I am full of ideas.	By far, most people think they are full of ideas.

2.4.2.2. Conscientiousness



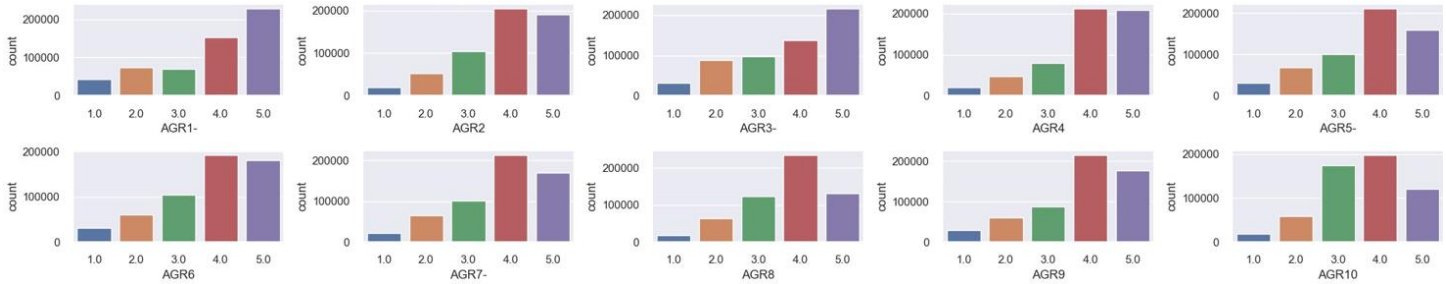
CSN1	I am always prepared.	By far, most people agree they are always prepared.
CSN2-	I leave my belongings around.	Very evenly spread, more people agree they leave their belongings around.
CSN3	I pay attention to details.	By far, most people agree/strongly agree they pay attention to details.
CSN4-	I make a mess of things.	More people disagree that they make a mess of things.
CSN5	I get chores done right away.	More people disagree they get chores done right away.
CSN6-	I forget to put things back in their proper place.	Very evenly spread, more people disagree they are forgetful.
CSN7	I like order.	By far, most people agree they like order.
CSN8-	I shirk my duties.	By far, fewer people strongly agree/agree they shirk their duties.
CSN9	I follow a schedule.	Relatively evenly spread, most people agree they follow a schedule.
CSN10	I am exacting in my work.	By far, fewer people strongly disagree/disagree they are exacting in their work.

2.4.2.3. Extroversion



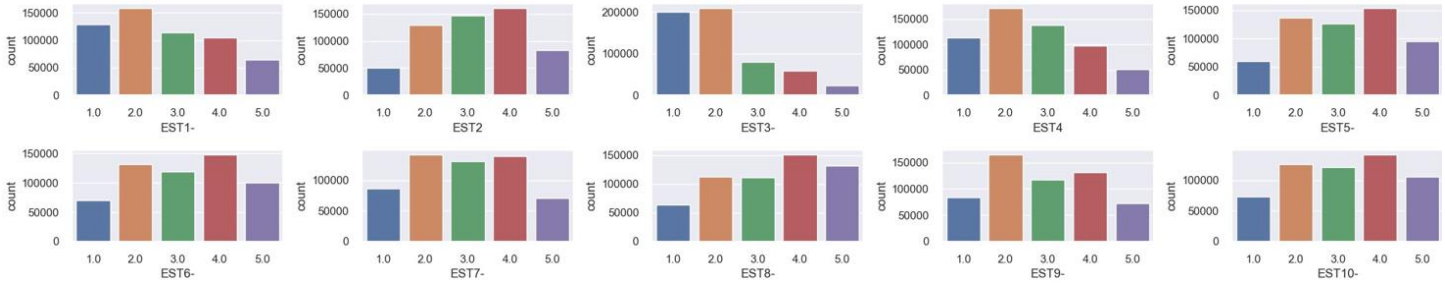
Questions for direct reference:		Interpretation:
EXT1	I am the life of the party.	By far, few people strongly agree to being the life of the party.
EXT2-	I don't talk a lot.	Relatively evenly spread, most people are neutral.
EXT3	I feel comfortable around people.	Fewest people are strongly uncomfortable around people.
EXT4-	I keep in the background.	Most people are neutral to agreeing on being in the background.
EXT5	I start conversations.	Most people agree to being conversation starters.
EXT6-	I have little to say.	By far, most people have something to say.
EXT7	I talk to a lot of different people at parties.	Relatively evenly spread, more people like to stick to themselves at parties.
EXT8-	I don't like to draw attention to myself.	By far, most people don’t like drawing attention to themselves.
EXT9	I don't mind being the center of attention.	Very evenly spread.
EXT10-	I am quiet around strangers.	By far, most people are quiet around strangers.

2.4.2.4. Agreeableness



AGR1-	I feel little concern for others.	By far, most people strongly agree to have concern for others.
AGR2	I am interested in people.	By far, most people agree/strongly agree they are interested in people.
AGR3-	I insult people.	By far, most people strongly disagree they insult people.
AGR4	I sympathize with others feelings.	By far, most people strongly agree/agree they have sympathy.
AGR5-	I am not interested in other people's problems.	By far, most people agree they’re interested in others’ problems.
AGR6	I have a soft heart.	By far, most people agree/strongly agree they have a soft heart.
AGR7-	I am not really interested in others.	By far, most people agree they are interested in others.
AGR8	I take time out for others.	By far, most people agree they take time out for others.
AGR9	I feel others emotions.	By far, most people agree they feel others’ emotions.
AGR10	I make people feel at ease.	By far, most people are neutral/agree they make people feel at ease.

2.4.2.5. Emotional Stability/Neuroticism



EST1-	I get stressed out easily.	Relatively evenly spread, more people get stressed easily.
EST2	I am relaxed most of the time.	Most people are neutral to agreeing to keeping relaxed.
EST3-	I worry about things.	By far, the majority of people agree to being worriers.
EST4	I seldom feel blue.	Most people agree to feeling blue often.
EST5-	I am easily disturbed.	Relatively evenly spread, more people disagree on being easily disturbed.
EST6-	I get upset easily.	Relatively evenly spread, more people disagree on being easily upset.
EST7-	I change my mood a lot.	Relatively evenly spread, more people agree on changing moods a lot.
EST8-	I have frequent mood swings.	More people disagree to having frequent mood swings.
EST9-	I get irritated easily.	Relatively evenly spread, more people agree on being easily irritated.
EST10-	I often feel blue.	Very evenly spread, more people disagree to feeling blue often.

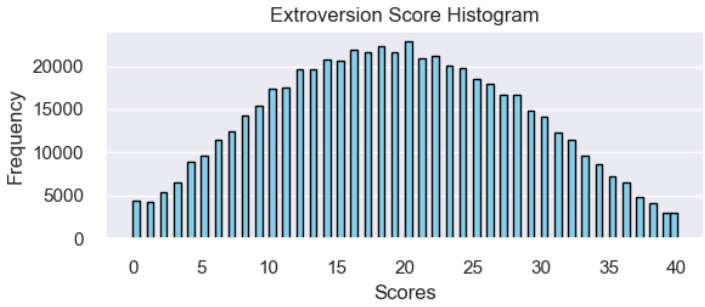
2.4.2.6. Overall Interpretation

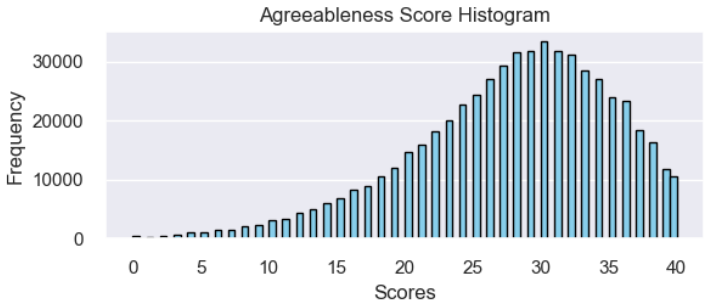
Openness:	Most people are open
Conscientiousness:	Relatively evenly spread but more people are conscientious.
Agreeableness:	Most people are agreeable
Extroversion:	Very evenly spread but most people have something to say but don't like to talk to strangers and don't like being the center of attention.
Neuroticism:	Very evenly spread but by far most people worry about things.

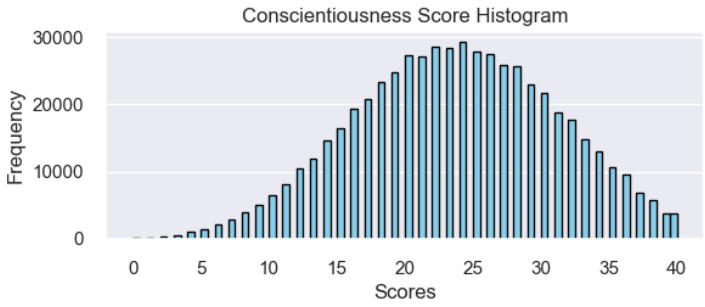
2.4.3. Personality Scores

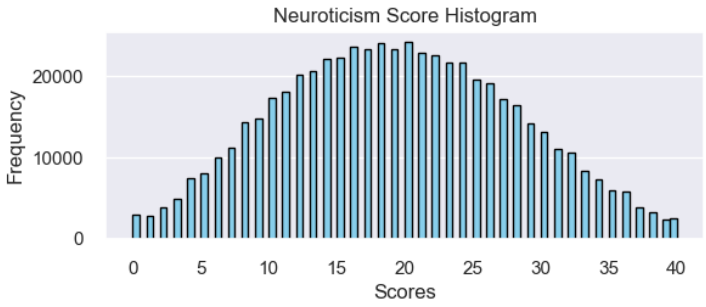
The personality scores sum up how the answers relate to the Big Five personalities. This fits very closely with the interpretation from the answers spread and paints an overall picture of where people fall on the scale.

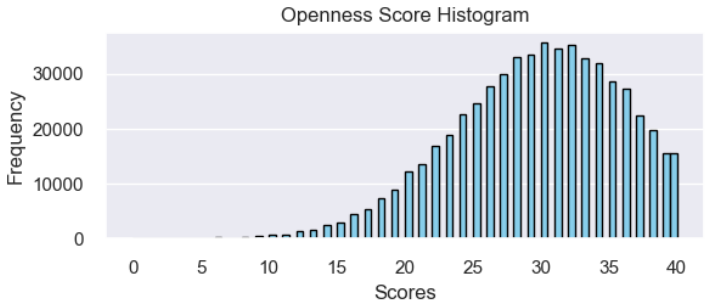
Openness:	Most people are open.
Conscientiousness:	More people are conscientious.
Extroversion:	Slightly less people are extroverted.
Agreeableness:	Most people are agreeable.
Neuroticism:	Slightly less people have a tendency towards experiencing negative emotions





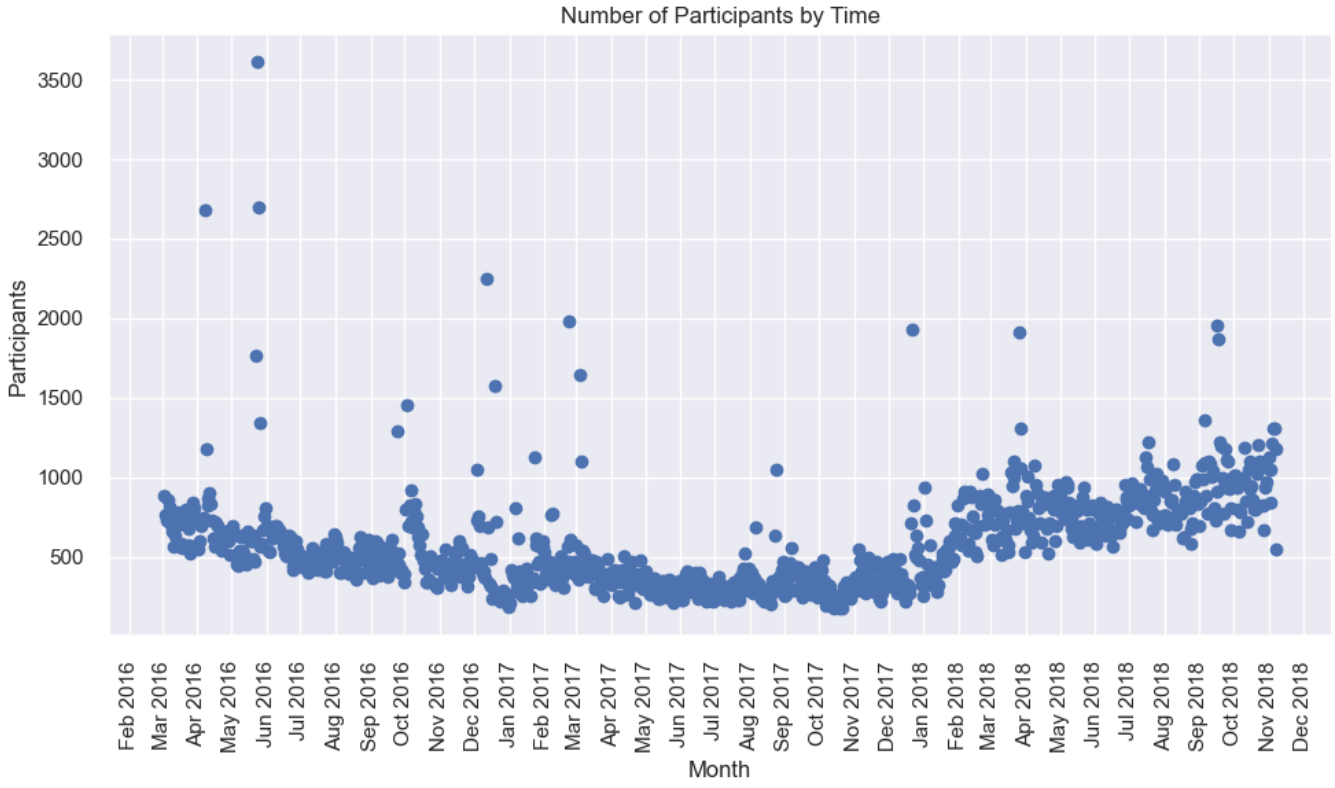






2.4.4. Timeframe of Participants

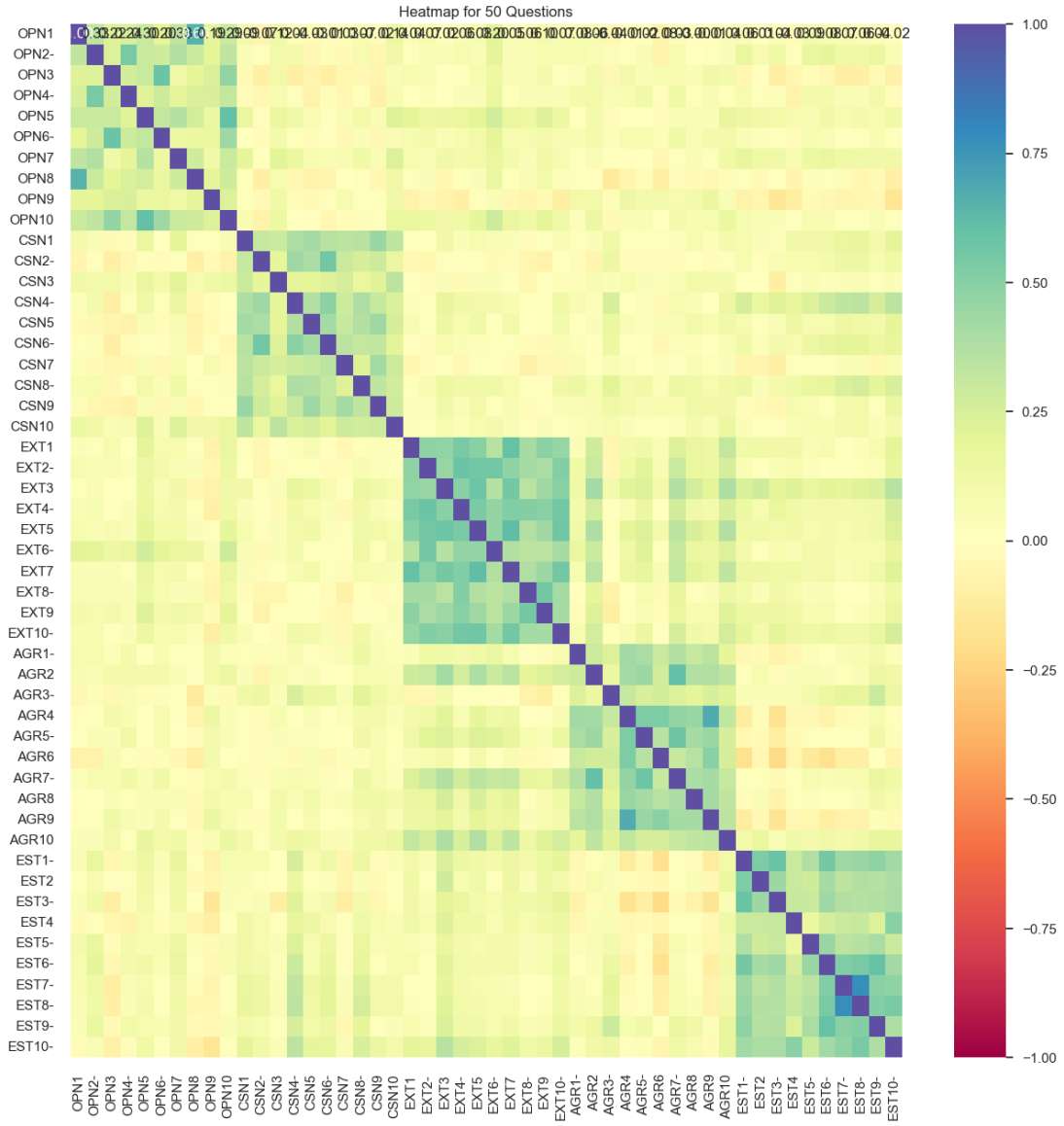
Individual test takers were decreasing from 2016 to 2017 and started increasing at the start of 2018.
It may be interesting to see how the personality scores differ based on the year and current events happening at the time but this is out of scope for this analysis.



2.5. Bivariate EDA

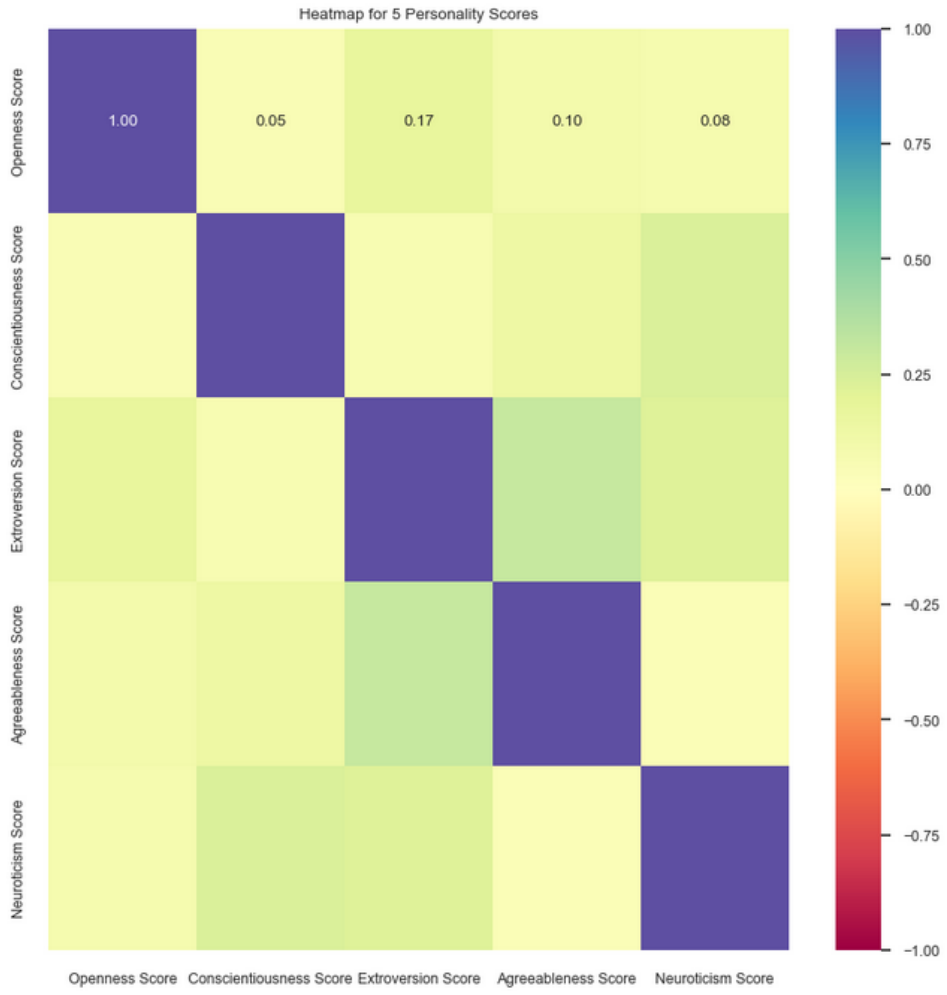
2.5.1. Correlation Table of 50 Questions

There is high correlation between the questions within the personality classification which is expected, and can cause issues with multicollinearity for the model. This suggests this might not be the best way to create distinct clusters without eliminating the specific questions that are highly correlated.



2.5.2. Correlation Table of Personality Scores

There is not anything very highly correlated within the personality scores so these should not have an issue with multicollinearity.



2.5.3. Personality Scores by Country

Refer to the code for individual charts for each country. The table below collects the mean for the personality scores by country. This can help as a guide to how to behave or communicate when you arrive to a new country for business or leisure. Further below have displayed the individual country chart for the top and bottom most countries in Openness as there is the largest differential between these scores. This would suggest when traveling to

Personality Scores Population Mean

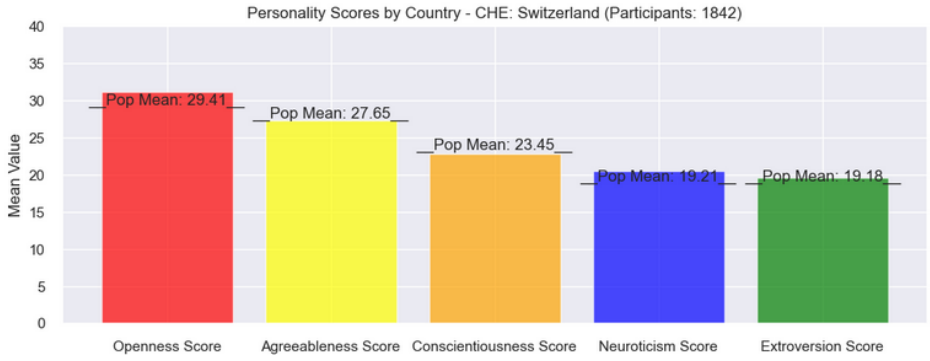
Openness	Conscientiousness	Extroversion	Agreeableness	Neuroticism
All 29.41	All 23.45	All 19.18	All 27.65	All 19.21

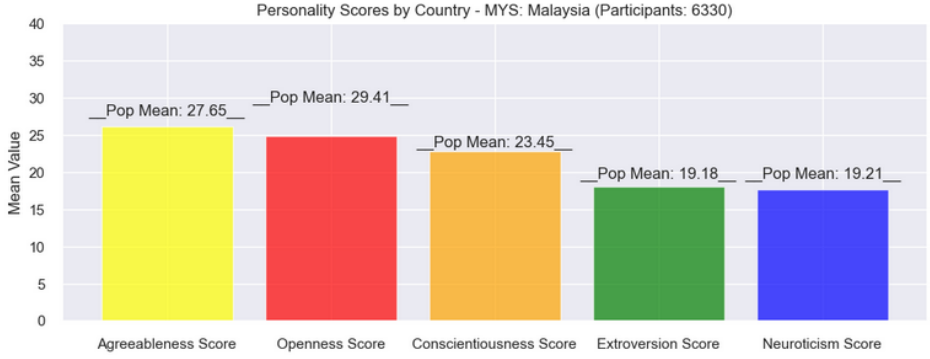
Top 5 Scores by Mean

Openness	Conscientiousness	Extroversion	Agreeableness	Neuroticism
CHE 31.17	ZAF 24.20	NLD 20.69	PAK 28.63	NLD 20.71
ISR 31.17	PAK 24.16	NOR 20.49	USA 28.34	CHE 20.50
DEU 31.11	USA 24.06	DNK 20.17	AUS 27.90	THA 20.33
FRA 31.00	ARE 23.75	IRL 19.83	ARE 27.89	SWE 20.13
AUT 30.93	NOR 23.59	PAK 19.79	CAN 27.74	AUT 20.08

Bottom 5 Scores by Mean

Openness	Conscientiousness	Extroversion	Agreeableness	Neuroticism
MYS 24.89	BRA 20.99	BRA 16.05	POL 23.89	GRC 16.71
HKG 25.92	PER 21.04	PRT 16.86	RUS 24.53	RUS 17.20
PAK 26.56	POL 21.10	POL 16.95	BRA 24.71	POL 17.31
VNM 26.67	CZE 21.53	VNM 16.95	FIN 25.11	PHL 17.41
PHL 26.72	FIN 21.54	PHL 17.64	CZE 25.36	TUR 17.42





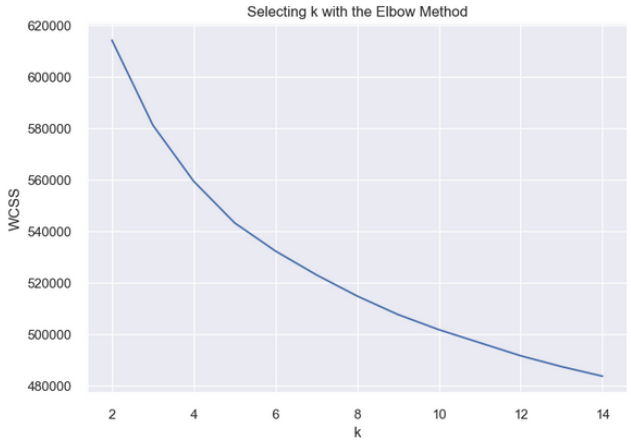
3. Machine Learning Models Used

Our dataset does not clearly have a target variable, and so unsupervised learning will be the primary approach to make sense of the data. We run these models against 10k random samples two times each over the 50 questions/answers (Answers Clusters), and the 5 personality scores (PScores Clusters). The goal is to find the best clustering method that will give clusters that are distinct from one another. The techniques we used were K-Means Clustering, Minibatch K-Means Clustering, Density Clustering with HDBSCAN after PCA, Hierarchical Clustering. We left the data unscaled as within each of the individual features modeled, the scale is the same: answers from 1 to 5, p-scores from 0 to 40.

3.1. K-Means Clustering

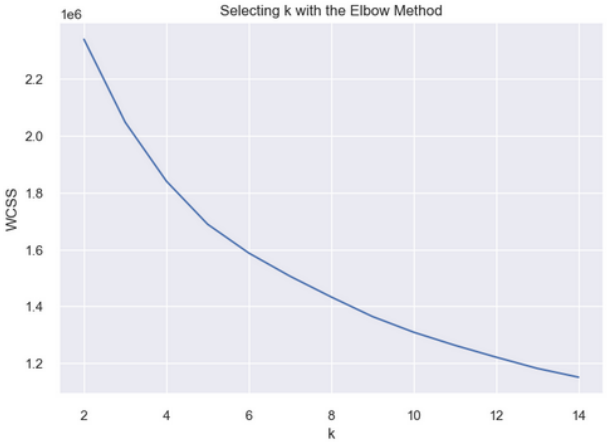
3.1.1. Traditional K-Means - Answers Clusters

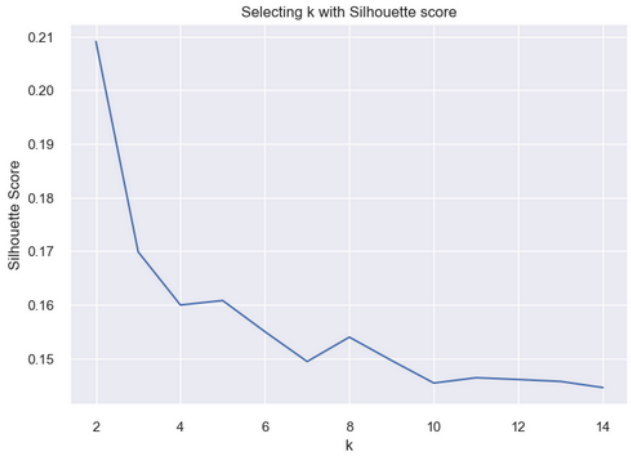
```
For n_clusters = 2, WCSS: = 614261.1594741798, Silhouette score: 0.10003438287413725)
For n_clusters = 3, WCSS: = 581334.3334318246, Silhouette score: 0.08111484700359724)
For n_clusters = 4, WCSS: = 559496.1807730757, Silhouette score: 0.06674856426467426)
For n_clusters = 5, WCSS: = 543358.3318730243, Silhouette score: 0.06213264201929137)
For n_clusters = 6, WCSS: = 532413.307059614, Silhouette score: 0.056933046839880395)
For n_clusters = 7, WCSS: = 523180.90344724234, Silhouette score: 0.05507699799133086)
For n_clusters = 8, WCSS: = 514921.11957504664, Silhouette score: 0.05126593795961181)
For n_clusters = 9, WCSS: = 507738.833716137, Silhouette score: 0.04846118420757983)
For n_clusters = 10, WCSS: = 501855.5835670504, Silhouette score: 0.04521780277097177)
For n_clusters = 11, WCSS: = 496791.404750015, Silhouette score: 0.045442858575440385)
For n_clusters = 12, WCSS: = 491730.0400089249, Silhouette score: 0.043444703614565004)
For n_clusters = 13, WCSS: = 487516.5300122177, Silhouette score: 0.04174004265625454)
For n_clusters = 14, WCSS: = 483804.77082395914, Silhouette score: 0.041347335616988015)
```

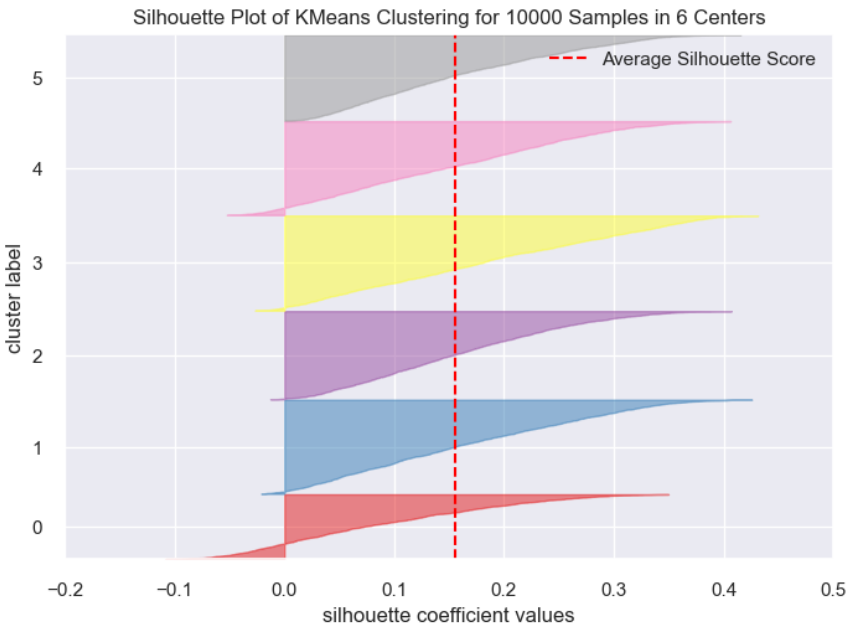


3.1.2. Traditional K-Means - PScores Clusters

```
For n_clusters = 2, WCSS: = 2338404.2991185095, Silhouette score: 0.20909656326431103)
For n_clusters = 3, WCSS: = 2047094.0543607692, Silhouette score: 0.16986271448881612)
For n_clusters = 4, WCSS: = 1840046.7728648498, Silhouette score: 0.15996314603205383)
For n_clusters = 5, WCSS: = 1688666.4845632091, Silhouette score: 0.16079836558166316)
For n_clusters = 6, WCSS: = 1587411.166379832, Silhouette score: 0.15498908779741682)
For n_clusters = 7, WCSS: = 1505987.599592246, Silhouette score: 0.14942204857502173)
For n_clusters = 8, WCSS: = 1433025.7884274612, Silhouette score: 0.15396976593027425)
For n_clusters = 9, WCSS: = 1364282.31267081, Silhouette score: 0.14960713665176564)
For n_clusters = 10, WCSS: = 1309201.143855227, Silhouette score: 0.14540946960890133)
For n_clusters = 11, WCSS: = 1263235.753099904, Silhouette score: 0.14640370279387088)
For n_clusters = 12, WCSS: = 1221546.5341310515, Silhouette score: 0.14607365096007646)
For n_clusters = 13, WCSS: = 1181956.0038577456, Silhouette score: 0.1457076722326683)
For n_clusters = 14, WCSS: = 1151110.4857856776, Silhouette score: 0.14456743211231782)
```

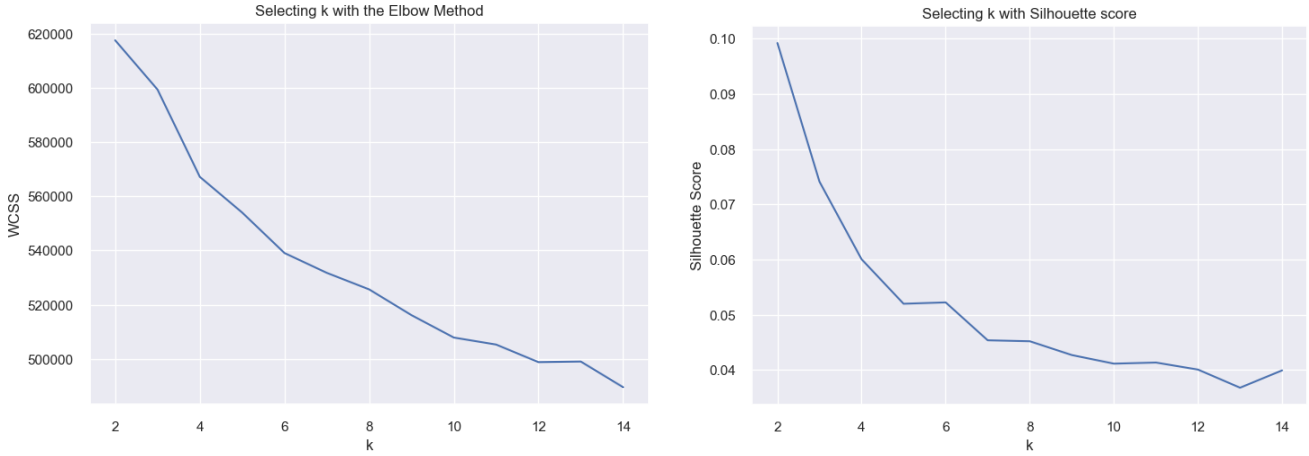






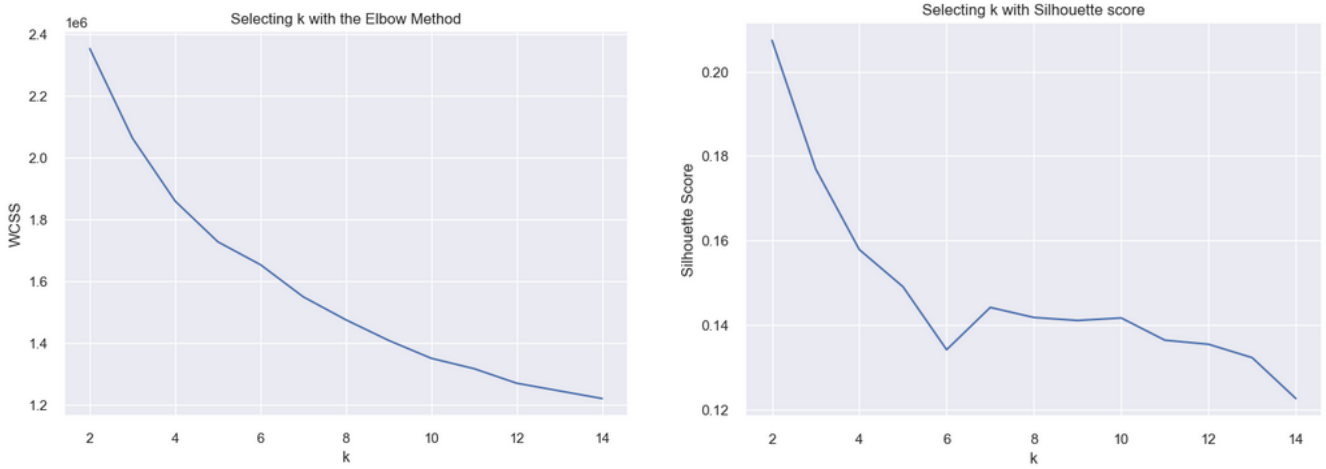
3.1.3. Minibatch K-Means Answers Clusters

```
For n_clusters = 2, WCSS: = 617590.1567382527, Silhouette score: 0.09919073364050514)
For n_clusters = 3, WCSS: = 599415.7124370943, Silhouette score: 0.07410210723989347)
For n_clusters = 4, WCSS: = 567189.3486445864, Silhouette score: 0.060064452433759484)
For n_clusters = 5, WCSS: = 554019.8757148976, Silhouette score: 0.051999655767986735)
For n_clusters = 6, WCSS: = 539078.3053117187, Silhouette score: 0.052245790815995O8)
For n_clusters = 7, WCSS: = 531761.6817203425, Silhouette score: 0.045385668995096 26)
For n_clusters = 8, WCSS: = 525659.2510211854, Silhouette score: 0.045215602656558045)
For n_clusters = 9, WCSS: = 516151.6788738613, Silhouette score: 0.04271565205306628)
For n_clusters = 10, WCSS: = 507886.9462751839, Silhouette score: 0.04115828833849128)
For n_clusters = 11, WCSS: = 505298.15974542405, Silhouette score: 0.041359201477765)
For n_clusters = 12, WCSS: = 498805.30188273045, Silhouette score: 0.04008364047373549)
For n_clusters = 13, WCSS: = 499030.6605543109, Silhouette score: 0.03680554746398148)
For n_clusters = 14, WCSS: = 489554.48698684736, Silhouette score: 0.03994207584109659)
```



3.1.4. Minibatch K-Means PScores Clusters

```
For n_clusters = 2, WCSS: = 2352487.5571859763, Silhouette score: 0.20739384614583806)
For n_clusters = 3, WCSS: = 2062720.558781524, Silhouette score: 0.17694387973602047)
For n_clusters = 4, WCSS: = 1859196.4952974226, Silhouette score: 0.15782912990034248)
For n_clusters = 5, WCSS: = 1727575.186062064, Silhouette score: 0.14903479970375694)
For n_clusters = 6, WCSS: = 1653303.2560821138, Silhouette score: 0.13411260343286394)
For n_clusters = 7, WCSS: = 1549636.5718623253, Silhouette score: 0.14413043003815723)
For n_clusters = 8, WCSS: = 1475129.8152531243, Silhouette score: 0.14174980794378414)
For n_clusters = 9, WCSS: = 1408604.011920987, Silhouette score: 0.14104300196626676)
For n_clusters = 10, WCSS: = 1350658.7737770998, Silhouette score: 0.14163727239786222)
For n_clusters = 11, WCSS: = 1317187.109929727, Silhouette score: 0.13636626923183723)
For n_clusters = 12, WCSS: = 1270095.1549621432, Silhouette score: 0.13540617655395104)
For n_clusters = 13, WCSS: = 1245204.595486747, Silhouette score: 0.13222940068588604)
For n_clusters = 14, WCSS: = 1220582.0994657965, Silhouette score: 0.1225531158931054)
```



3.1.5. Evaluation

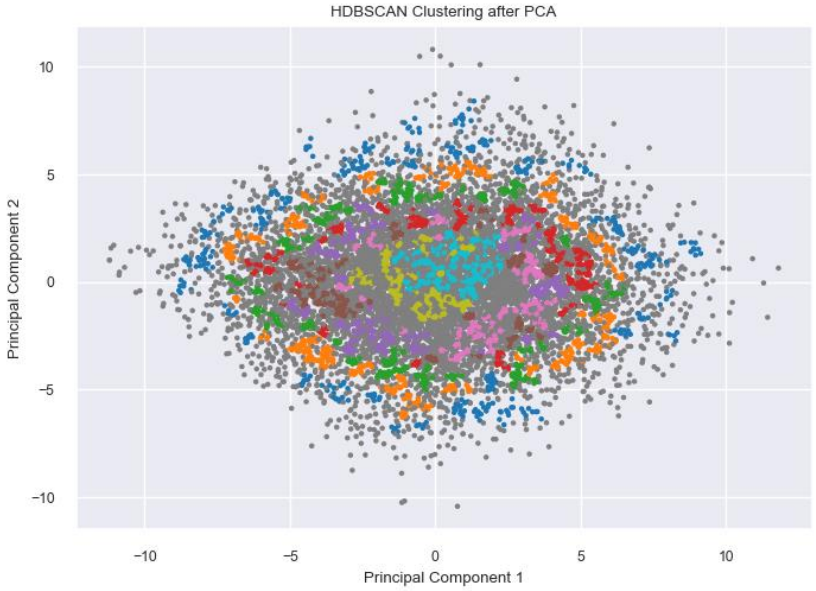
Of the K-means clustering models used, the highest quality clusters came from traditional K-means clusters on the Personality Scores on 6 clusters with a Silhouette Score of 0.155 which although not very high should be usable. The silhouette plot produced shows the clusters to be somewhat evenly distributed and distinct. For K-means on the Answers Clusters the results with the optimal number of clusters as 6 with a S-score of 0.057. Unfortunately, this low score suggests the clusters produced are not very good quality and will likely produce unclear results.

3.2. Density Clustering with HDBScan and PCA

Another approach used was Density Clustering using the HDBScan package. The number of clusters used was what the previous K-means models suggested for optimal k. For visualization, it required reducing the number of dimensions down to two components using PCA, and then clustering with HDBScan to predict the clusters. For the computer to be able to run this while still providing something interesting to visualize, the sample size is 10000, using min_cluster_size of 5 and min_samples of 5.

3.2.1. HDBScan on Answers Clusters

HDBScan on the Answers Clusters resulted in 10 clusters but with over half the observations in the sample considered as noise.



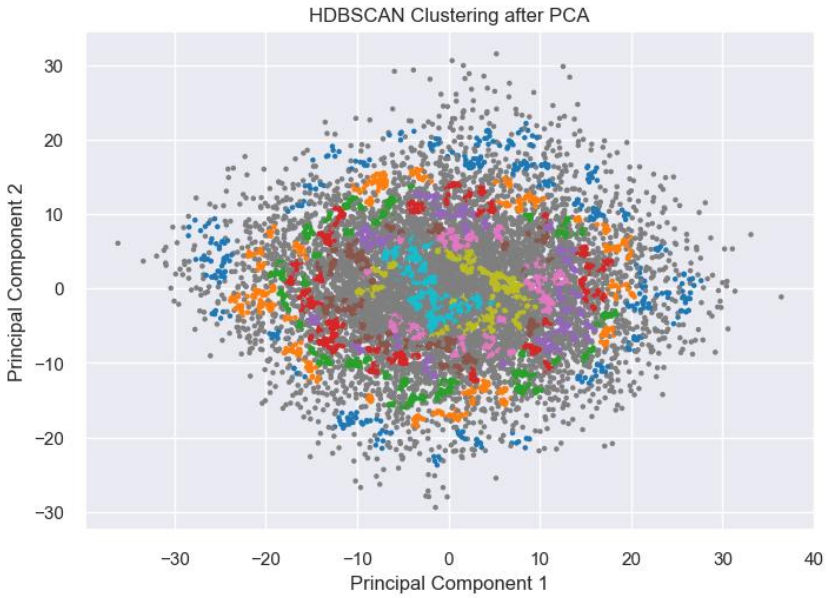
3.2.2. Evaluation

The performance was evaluated with the silhouette score which resulted with a slightly negative number, this suggests these clusters overlap and are not very distinct.

```
Silhouette Score: -0.01295
```

3.2.3. HDBScan on PScores Clusters

HDBScan on the PScores Clusters resulted in 10 clusters but with more than half of the observations in the sample considered as noise.



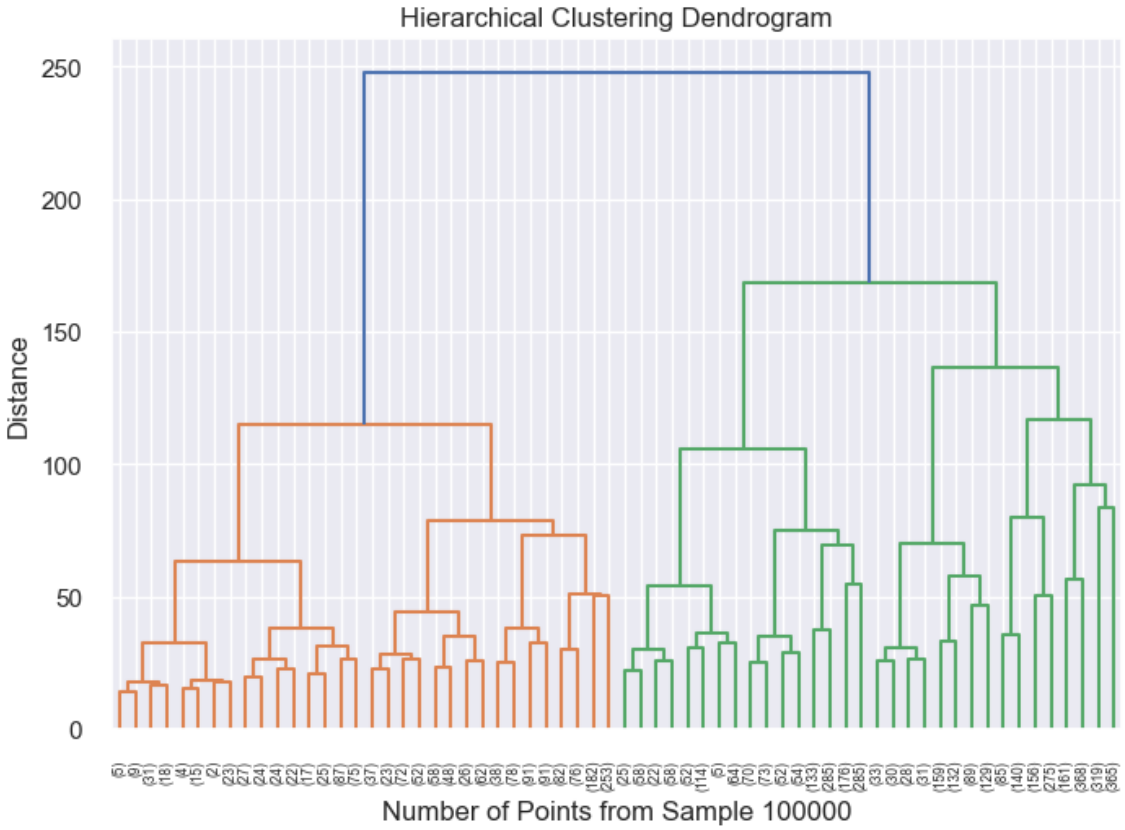
3.2.4. Evaluation

The performance was evaluated with the silhouette score which resulted with a more negative number, this suggests these clusters overlap much more so than the Answers results and are not very distinct.

```
Silhouette Score: -0.1270
```

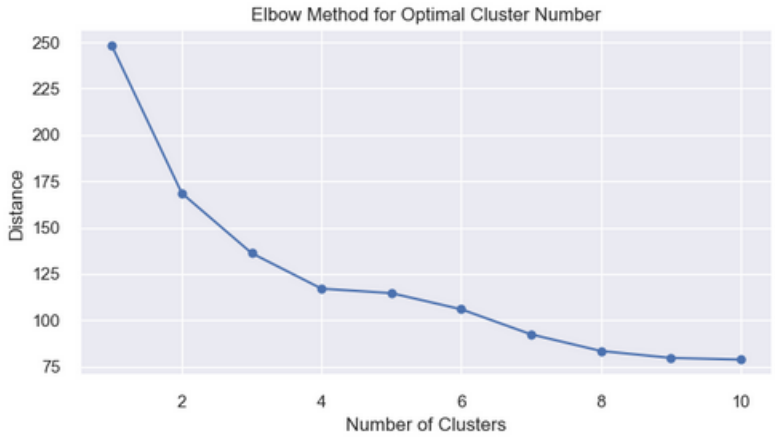
3.3. Hierarchical Clustering

3.3.1. HC on Answers Clusters

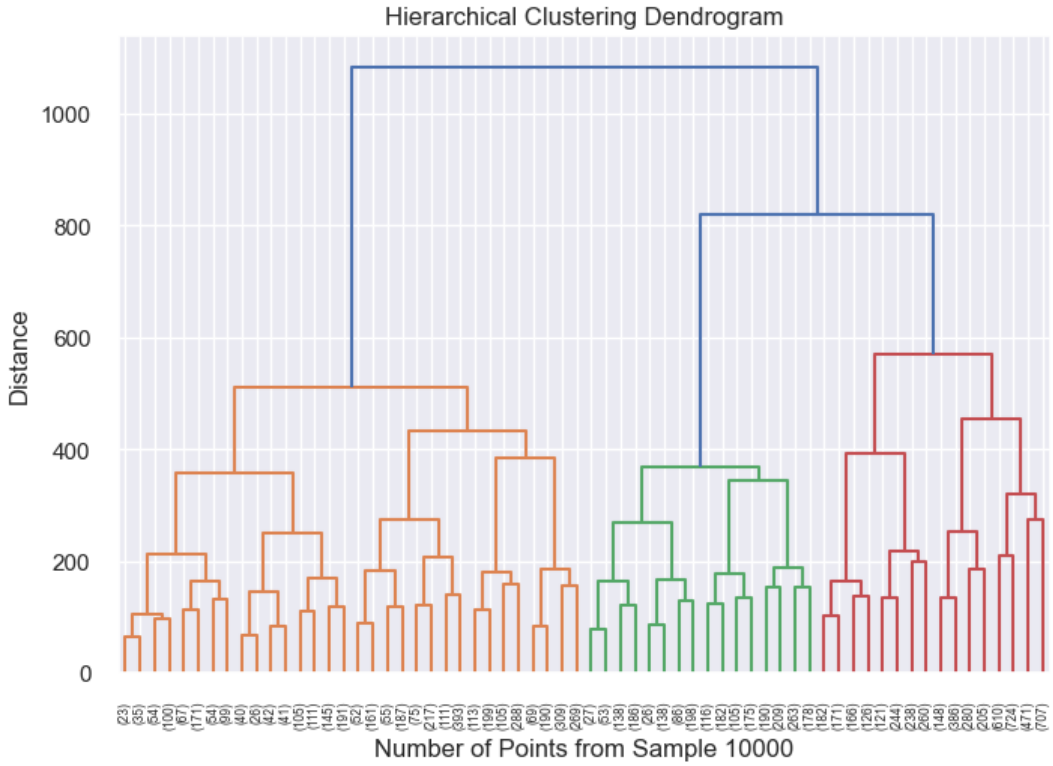


3.3.2. Evaluation

Hierarchical Clustering of the Answers results in the optimal number of clusters of 4 to 5, with a Euclidean distance of ~140.

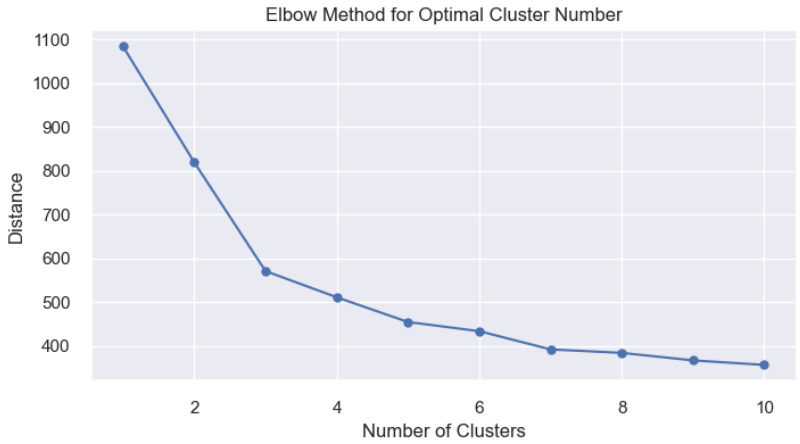


3.3.3. HC on PScores Clusters



3.3.4. Evaluation

Hierarchical Clustering of the PScores results in the optimal number of clusters of 4, with a Euclidean distance of ~500. These would be considered more distinct than the clusters from the answers based on the distance.



4. Evaluations Metrics

The evaluation metrics used are described above, but for our analysis we focused on producing more clusters that are as distinct as possible. To summarize all of the results.

	Answers Clusters		PScores Clusters	
Clustering Technique	**Optimal K**	**Silhouette Score**	**Optimal K**	**Silhouette Score**
Traditional K-Means	6	0.0569	6	0.1550
Minibatch K-Means	6	0.0522	5	0.1490
HDBScan Density	10	-0.0130	10	-0.1270
Hierarchical	4	Distance 140	4	Distance 500

Based on these results we went with Traditional K-Means using clusters generated from the Personality Scores.

5. Cluster Analysis

5.1. Cluster 0

According to the average values on the PScores:

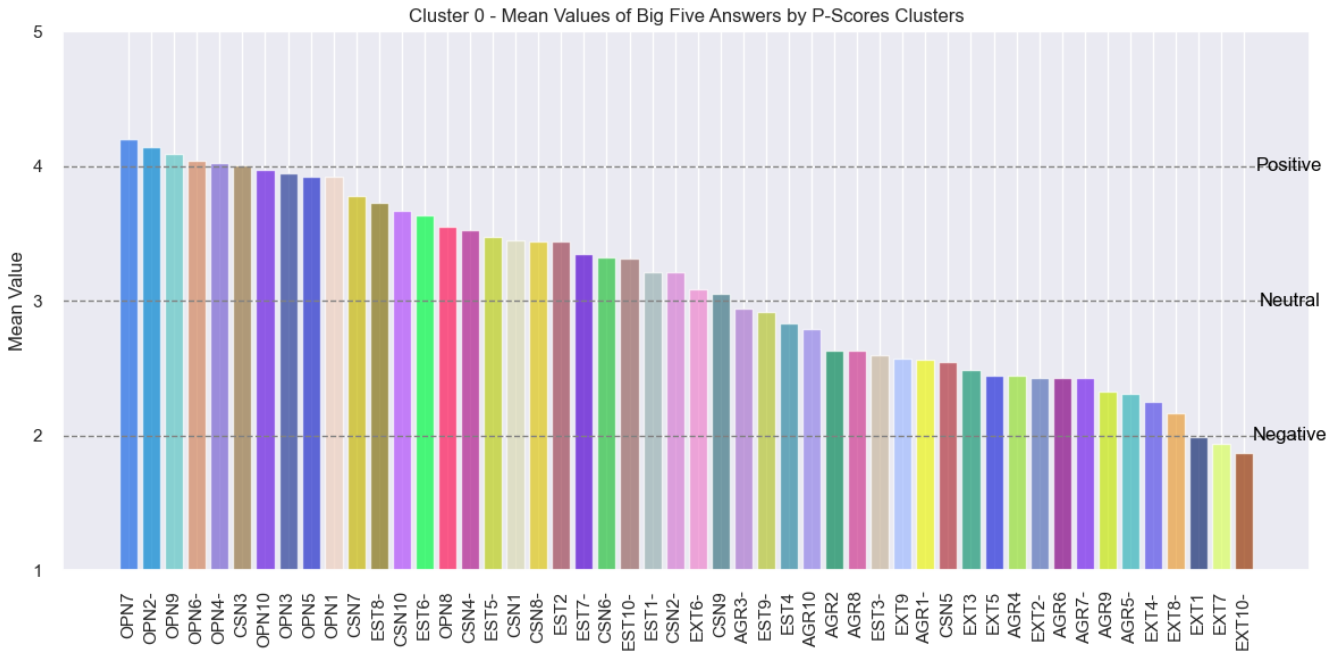
- Agreeableness is far below average.
- Conscientiousness is average.
- Extroversion is far below average.
- Neuroticism is above average.
- Openness is average.

Strongly positive to these questions:

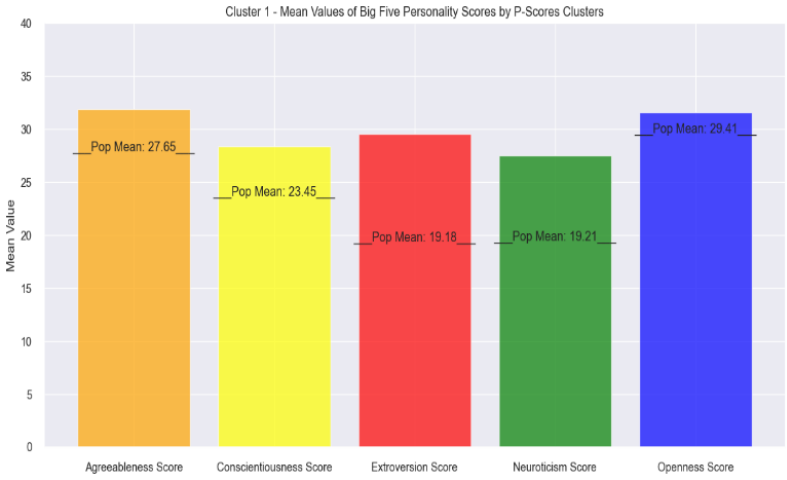
OPN7 I am quick to understand things. (Agree)
OPN2- I have difficulty understanding abstract ideas. (Disagree)
OPN9 I spend time reflecting on things. (Agree)

Strongly negative to these questions:

EXT10- I am quiet around strangers. (Agree)
EXT7 I talk to a lot of different people at parties. (Disagree)
EXT1 I am the life of the party. (Disagree)



5.2. Cluster 1



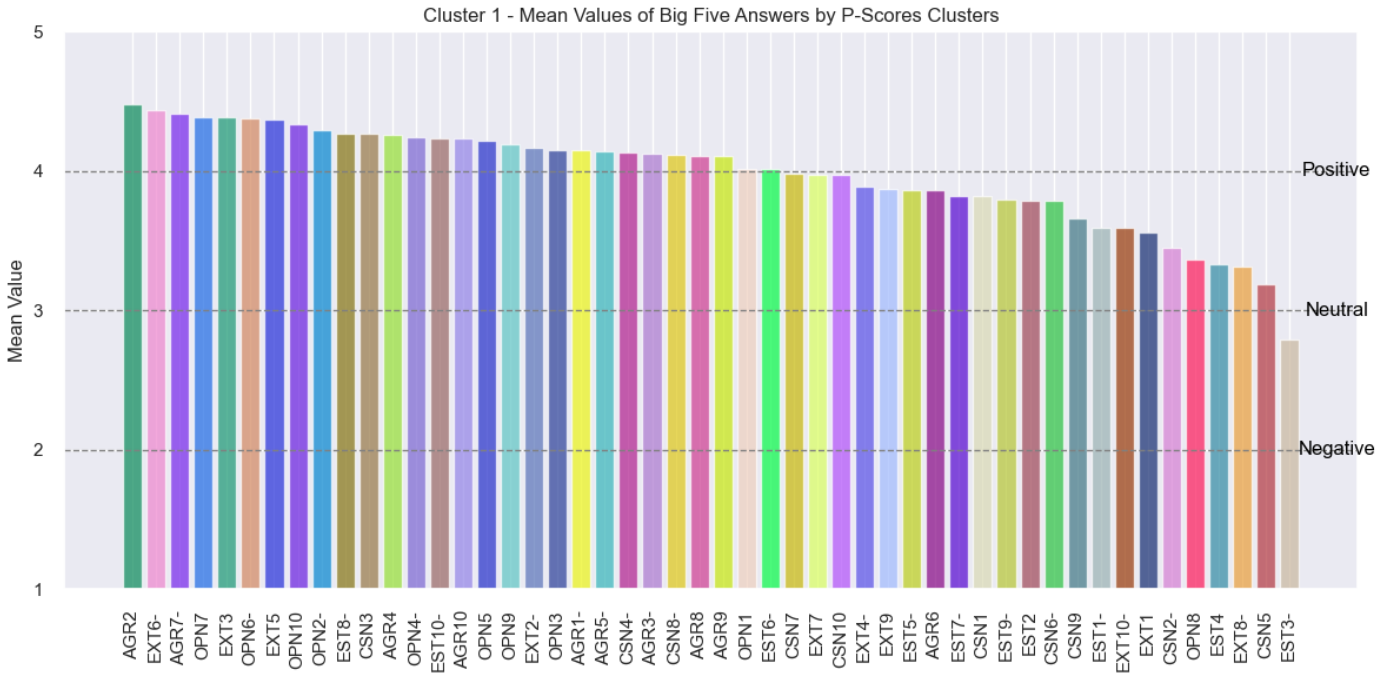
According to the average values on the PScores:

- Agreeableness is above average.
- Conscientiousness is above average.
- Extroversion is far above average.
- Neuroticism is far above average.
- Openness is above average.

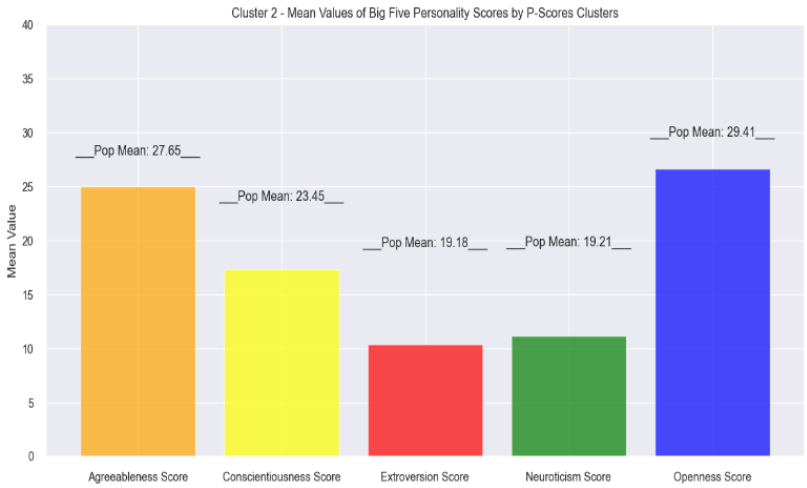
Strongly positive to these questions (Top 10):

AGR2 I am interested in people. (Agree)
EST6- I get upset easily. (Disagree)
AGR7- I am not really interested in others. (Disagree)
OPN7 I am quick to understand things. (Agree)
EXT3 I feel comfortable around people. (Agree)
OPN6- I do not have a good imagination. (Disagree)
EXT5 I start conversations. (Agree)
OPN10 I am full of ideas. (Agree)
OPN2- I have difficulty understanding abstract ideas. (Disagree)
EXT8- I don't like to draw attention to myself. (Disagree)

No negative responses to any questions.
This is the only below neutral question:
EST3- I worry about things. (Agree)



5.3. Cluster 2



According to the average values on the PScores:

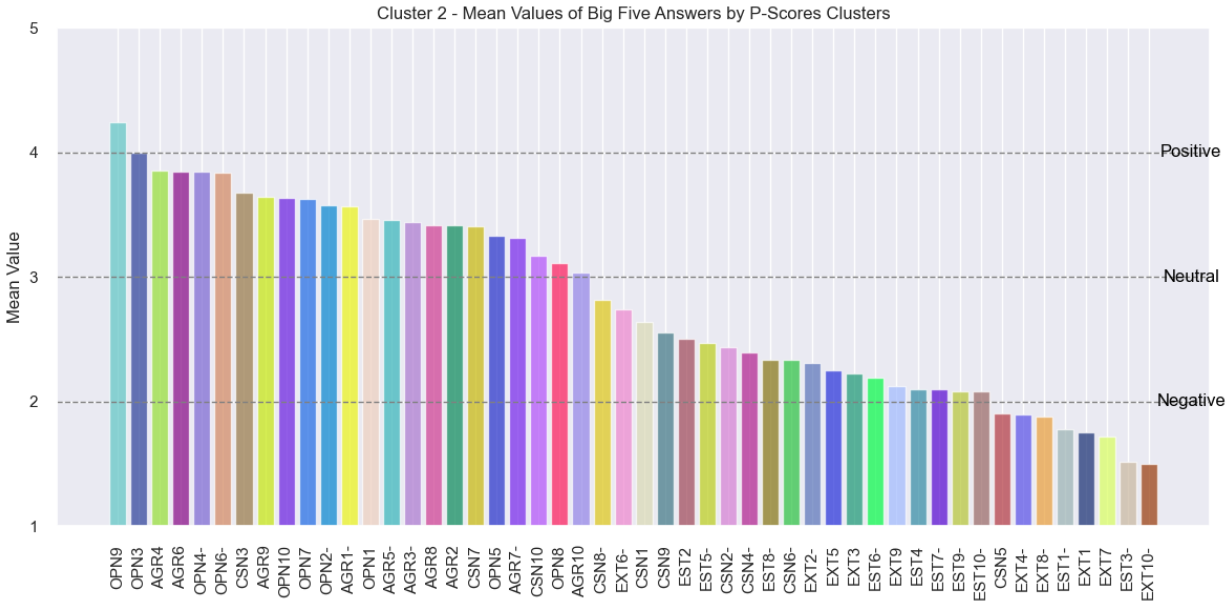
- Agreeableness is below average.
- Conscientiousness is far below average.
- Extroversion is far below average.
- Neuroticism is far below average.
- Openness is below average.

Strongly positive to these questions:

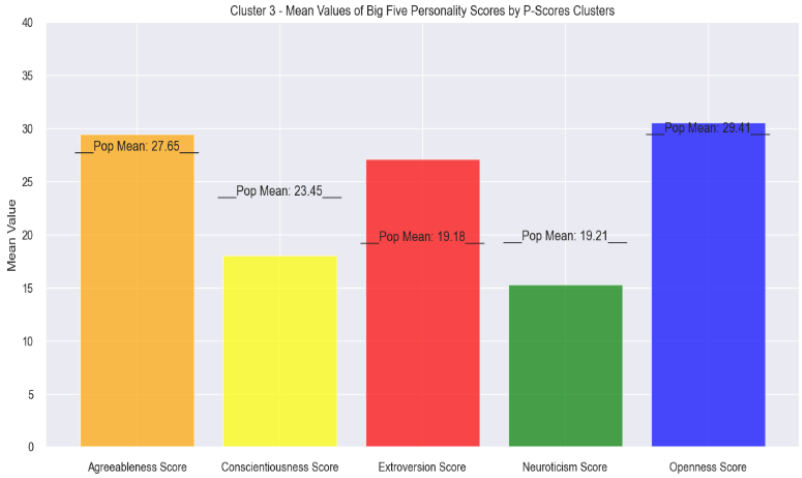
OPN9 I spend time reflecting on things. (Agree)
OPN3 I have a vivid imagination. (Agree)

Strongly negative to these questions:

EXT10- I am quiet around strangers. (Agree)
EST3- I worry about things. (Agree)
EXT7 I talk to a lot of different people at parties. (Disagree)
EXT1 I am the life of the party. (Disagree)
EXT8- I don't like to draw attention to myself. (Agree)
EXT4- I keep in the background. (Agree)
CSN5 I get chores done right away. (Disagree)



5.4. Cluster 3

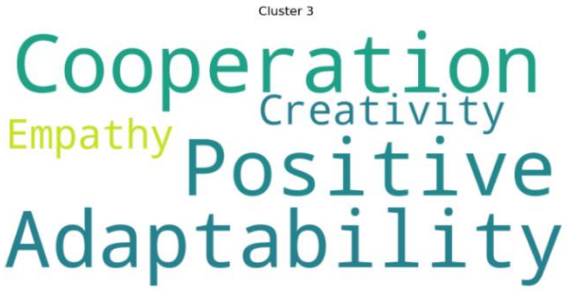


According to the average values on the PScores:

- Agreeableness is slightly above average.
- Conscientiousness is far below average.
- Extroversion is far above average.
- Neuroticism is far below average.
- Openness is average.

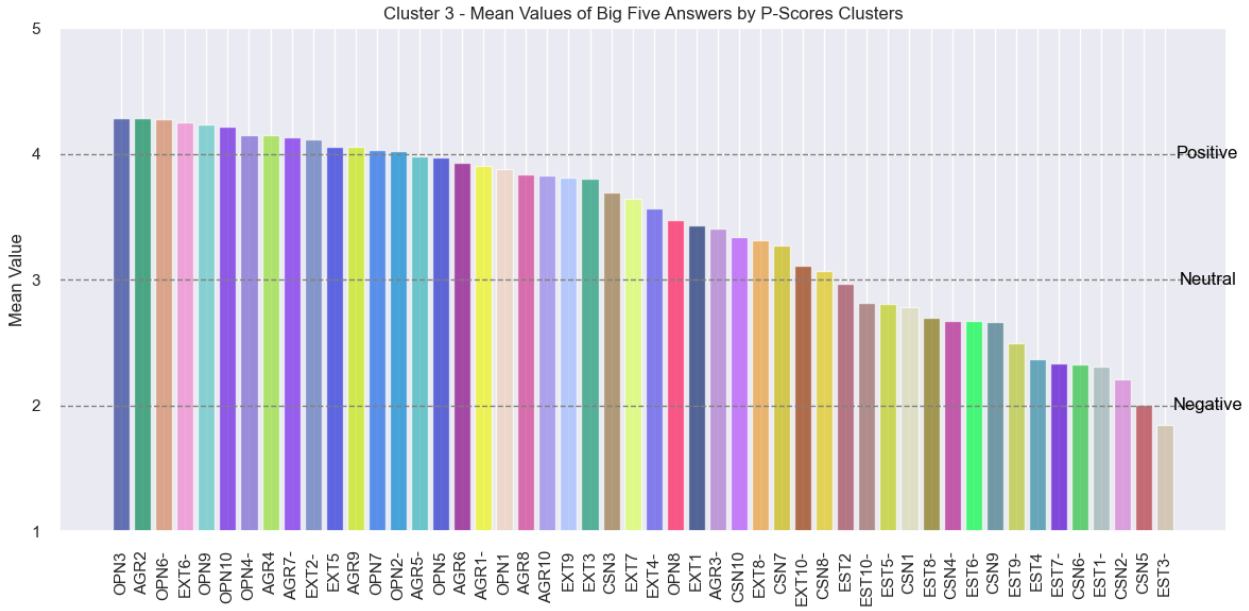
Strongly positive to these questions (Top 10):

OPN3	I have a vivid imagination. (Agree)
AGR2	I am interested in people. (Agree)
OPN6-	I do not have a good imagination. (Disagree)
EXT6-	I have little to say. (Disagree)
OPN9	I spend time reflecting on things. (Agree)
OPN10	I am full of ideas. (Agree)
OPN4-	I am not interested in abstract ideas. (Disagree)
AGR4	I sympathize with others feelings. (Agree)
AGR7-	I am not really interested in others. (Disagree)
EXT2-	I don't talk a lot. (Disagree)

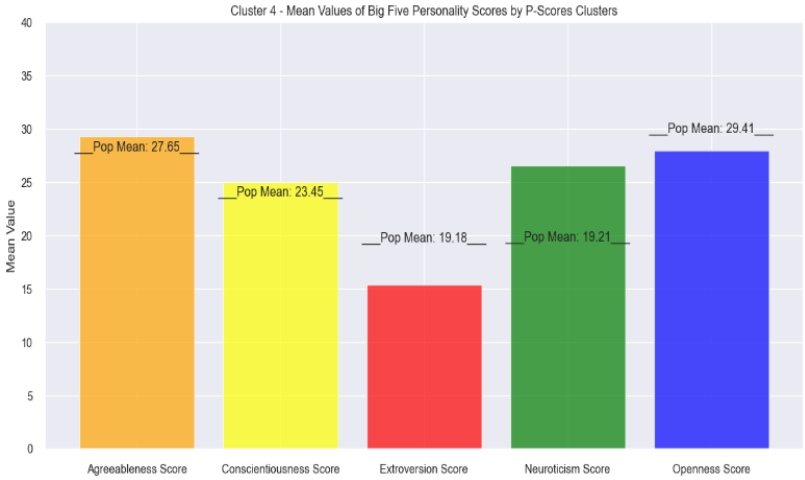


Strongly negative to these questions:

EST3-	I worry about things. (Agree)
CSN5	I get chores done right away. (Disagree)



5.5. Cluster 4



According to the average values on the PScores:

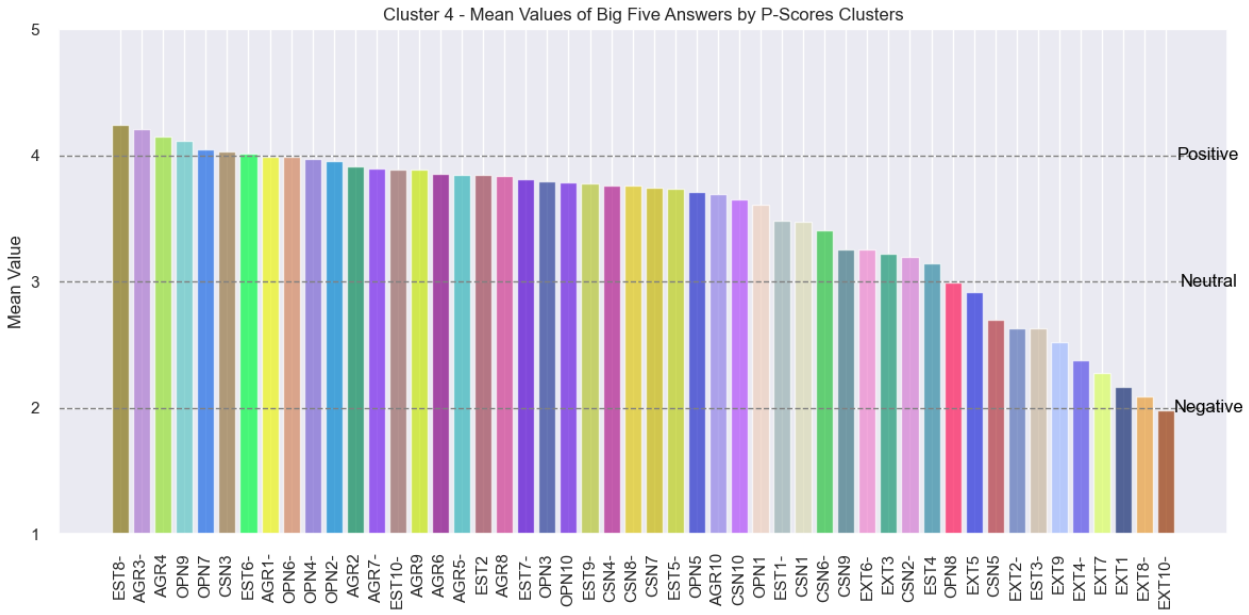
- Agreeableness is slightly above average.
- Conscientiousness is slightly above average.
- Extroversion is far above average.
- Neuroticism is far above average.
- Openness is below average.

Strongly positive to these questions:

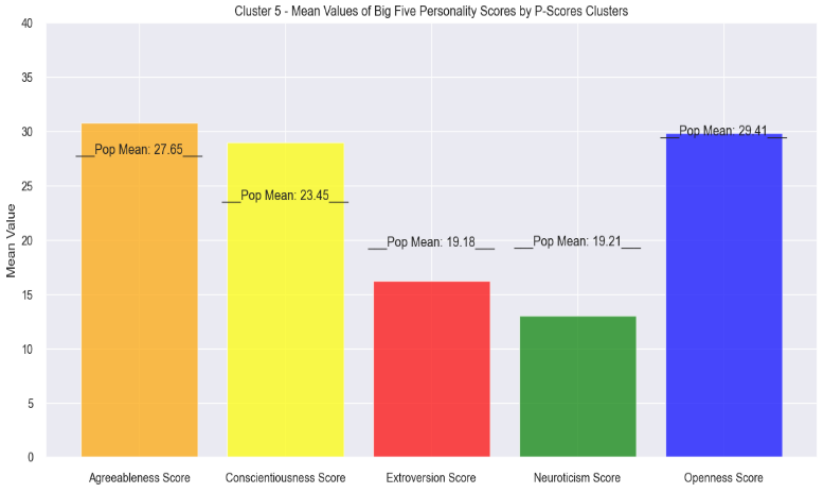
EXT6- I have little to say. (Disagree)
AGR3- I insult people. (Disagree)
AGR4 I sympathize with others feelings. (Agree)
OPN9 I spend time reflecting on things. (Agree)
OPN7 I am quick to understand things. (Agree)
CSN3 I pay attention to details. (Agree)

Strongly negative to these questions:

EXT10- I am quiet around strangers. (Agree)



5.6. Cluster 5



According to the average values on the PScores:

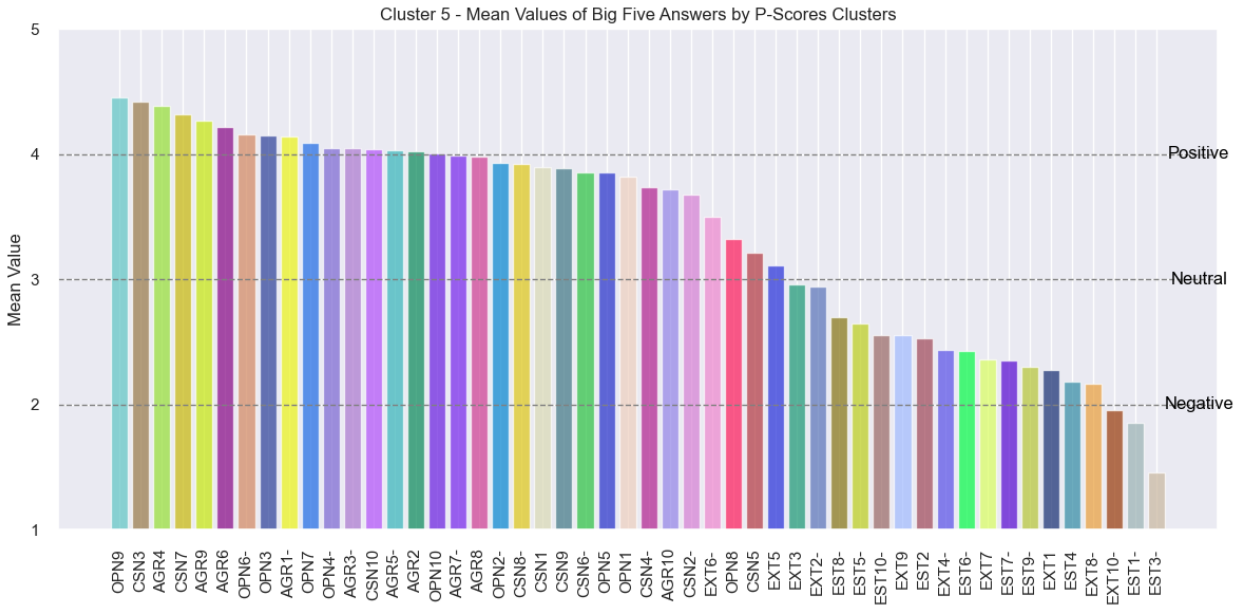
- Agreeableness is slightly above average.
- Conscientiousness is above average.
- Extroversion is below average.
- Neuroticism is far below average.
- Openness is average.

Strongly positive to these questions (Top 10):

OPN9	I spend time reflecting on things. (Agree)
CSN3	I pay attention to details. (Agree)
AGR4	I sympathize with others feelings. (Agree)
CSN7	I like order. (Agree)
AGR9	I feel others emotions. (Agree)
AGR6	I have a soft heart. (Agree)
OPN6-	I do not have a good imagination. (Disagree)
OPN3	I have a vivid imagination. (Agree)
AGR1-	I feel little concern for others. (Disagree)
OPN7	I am quick to understand things. (Agree)

Strongly negative to these questions:

EST3-	I worry about things. (Agree)
EST1-	I get stressed out easily. (Agree)
EXT10-	I am quiet around strangers. (Agree)



6. Insights and Recommendations

6.1. Cluster Descriptions and Suitability

Based on the strengths of each of the clusters, these are their suitable roles:

Cluster Number	Suitable Job Roles
Cluster 0: Innovators	Innovation Manager, Creative Director, Entrepreneur, Research Scientist, Project Manager (Innovative Projects)
Cluster 1: Collaborators	Human Resources Manager, Team Leader, Customer Service, Representative, Event Planner, Public Relations Specialist
Cluster 2: Independents	Freelancer, Independent Consultant, Graphic Designer, Researcher (Independent Projects), Archivist
Cluster 3: Visionaries	Team Leader, Creative Project Manager, Marketing Manager, Art Director, Event Coordinator
Cluster 4: Analysts	Project Manager, Financial Analyst, Data Analyst, Quality Assurance Specialist, Operations Coordinator
Cluster 5: Leaders	CEO/Executive Director, Operations Manager, Leadership, Coach, Organizational Psychologist, R&D Director

6.2. Recommendations

6.2.1. Data Quality Enhancement:

Data quality is paramount for meaningful analysis, especially in the context of personality assessments. The first aspect of enhancing data quality involves the addition of better survey questions. The effectiveness of a survey in capturing diverse personality traits relies on the clarity and relevance of its questions. By incorporating more precise and pertinent questions, the survey can better elicit accurate responses, contributing to a more comprehensive understanding of individuals' personalities.

Addressing missing values is equally crucial for robust data quality. Incomplete data can introduce biases and hinder the accuracy of analyses. Strategies such as imputation techniques or, when applicable, actively addressing missing values through additional data collection, can significantly enhance the completeness and reliability of the dataset. This dual approach ensures that the survey captures a broader range of personality dimensions and minimizes the impact of missing or incomplete information on subsequent analyses.

6.2.2. Further Analysis Opportunities:

Explore the interactions between specific personality traits:
While initial clustering provides a foundational understanding of personality patterns, there exists a valuable opportunity to deepen insights by exploring the interactions between specific personality traits. This entails identifying relationships and dependencies between pairs or groups of traits. Statistical methods such as correlation analyses or regression analyses can be employed to uncover how certain traits influence each other. This more nuanced exploration goes beyond basic clustering and provides a richer understanding of the intricate interplay between different facets of an individual's personality.

The exploration of trait interactions not only enhances the complexity of personality profiles but also sheds light on patterns that might be overlooked in a simplistic clustering analysis. Insights derived from these interactions contribute to a more accurate characterization of individuals, revealing how specific traits co-occur or potentially counterbalance each other. This deeper analysis unlocks a more sophisticated understanding of personality dynamics within the dataset.

6.2.3. Examining How the Pandemic Alters Personality Clusters:

The ongoing pandemic has introduced unprecedented challenges and stressors that may influence individual behaviors and, consequently, alter the observed patterns in personality clusters. To comprehend the impact of external events, such as the pandemic, on personality, it is essential to examine shifts in trait distributions over time. This examination may involve comparing personality profiles collected before and after the pandemic or during different phases of its progression.

Conducting a longitudinal analysis provides a valuable lens into the evolving nature of personality clusters. By tracking changes over different periods, it becomes possible to discern whether certain traits have adapted or shifted in response to the dynamic external environment. This longitudinal perspective contributes not only to understanding short-term adjustments but also to gauging the potential long-term effects of significant events on the composition of personality clusters. In essence, examining how the pandemic alters personality clusters provides insights into the adaptability and resilience of individuals in the face of external challenges.

7. Challenges and Things Learned

- **Optimizing K-means clustering due to high processing power and memory consumption:**

K-means clustering is a widely used algorithm for grouping data points, but its efficiency can be compromised by high processing power and memory consumption. The computational intensity arises from the iterative nature of K-means, where centroids are updated until convergence. To address high processing power requirements, optimization techniques such as parallel processing or distributed computing can be employed, utilizing the capabilities of multiple processors concurrently. Additionally, memory consumption can be reduced by implementing strategies like sparse data structures or sampling methods, which allow for the manipulation of large datasets with a more modest memory footprint.

- **Clusters undergo changes every time the code is run:**

A notable challenge with K-means clustering is its sensitivity to the initial placement of centroids, leading to different cluster assignments with each run of the code. To mitigate this variability, optimization techniques like K-means++ initialization can be applied. K-means++ refines the initial centroid placement, improving the chances of converging to a more optimal solution. Setting a random seed is another practice that enhances reproducibility, ensuring that the random initialization process remains consistent across different runs, facilitating comparison and interpretation of results.

- **The questions lacked distinctiveness, hindering the formation of highly unique clusters:**

The effectiveness of clustering algorithms is contingent on the distinctiveness of patterns within the data, and this is particularly relevant in the context of survey questions designed to assess personality traits. If the questions lack diversity or fail to capture nuanced aspects of personality, the resulting clusters may be ambiguous or overlapping. To address this issue, survey questions can be refined to ensure they cover different facets of personality, avoiding redundancy and enhancing distinctiveness. Feature engineering, involving the creation of additional features or modifications to existing ones, can also contribute to better differentiation. Psychologistsain experts, such as psychologists or sociologists, in the design of questions can provide valuable insights and ensure alignment with established personality models, further refining the clustering process.

8. Conclusion

- **K-means clustering with the traditional method provides the best classification results for the personality test data set:**

The selection of an appropriate clustering algorithm plays a crucial role in accurately categorizing personality traits within a dataset. In this context, the traditional K-means clustering using sample data yielded the best classification results. There is still optimization that can be further performed however we were limited in our available resources. Also comparable was the mini-batch K-means clustering method which is an efficient variant of K-means that processes a subset or "mini-batch" of the data in each iteration supposedly making it computationally less demanding and suitable for extensive datasets. By leveraging mini-batches, this method enhances the scalability of K-means, enabling it to handle larger volumes of personality test data without sacrificing classification accuracy. The mini-batch approach strikes a balance between computational efficiency and clustering effectiveness, making it well-suited for practical applications. In our analysis the traditional k-means performed slightly better however after further optimization will likely allow mini-batch k-means to surpass the traditional method.

- **Optimizing questions in every attribute can result in more distinctiveness of the clusters:**

The distinctiveness of clusters in a personality test dataset is heavily reliant on the quality and relevance of the survey questions associated with each attribute. The optimization of questions involves refining and tailoring inquiries to ensure they capture unique aspects of each attribute, preventing redundancy and ambiguity. When questions are carefully optimized, they contribute to a more granular and nuanced understanding of individuals' personalities. This results in more distinct clusters, as each attribute is precisely and comprehensively represented. By emphasizing clarity, specificity, and relevance in the survey questions for each attribute, the clustering model gains the capacity to form clusters that truly encapsulate the diversity of personality traits present in the dataset.

- **Current clustering model can be deployed to form initial interpretations to any real-world personality test application:**

The utility of a clustering model extends beyond the confines of the dataset it was trained on, especially when dealing with personality assessments. The statement suggests that the current clustering model is versatile enough to be applied to real-world personality test applications. Deploying the model in new scenarios allows for the formation of initial interpretations, providing insights into the personality profiles of individuals in diverse contexts. The adaptability of the clustering model implies its potential utility in various practical applications, such as workforce management, psychological assessments, or personalized recommendation systems. These initial interpretations serve as a valuable starting point for further analysis and exploration in real-world scenarios, offering a foundation for understanding and leveraging personality traits in different contexts.