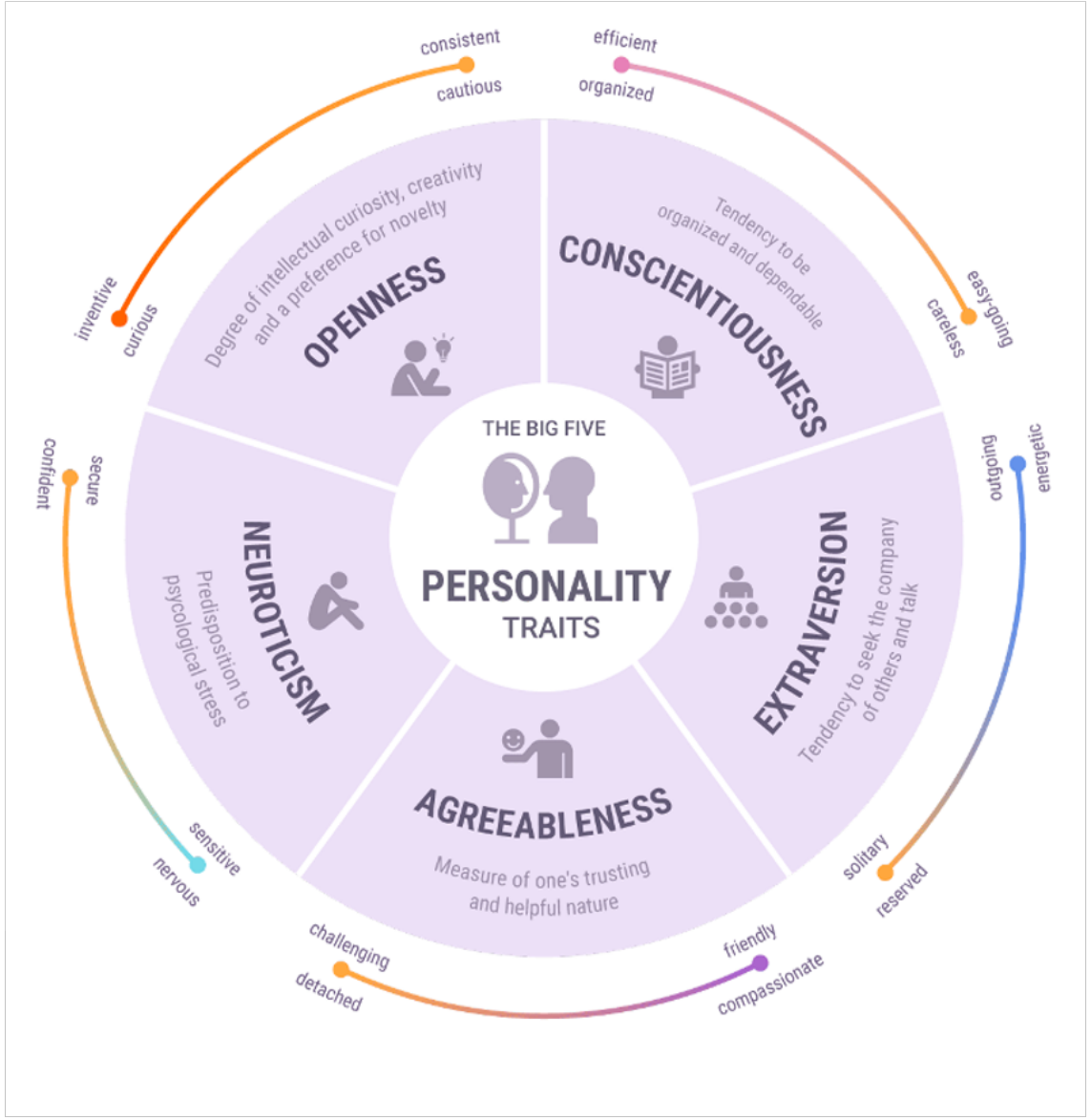
Personality Clustering

Project Idea



Many contemporary personality psychologists believe that there are five basic dimensions of personality, often referred to as the "Big 5" personality traits.

Our idea is to determine clusters of people among candidates who answered to the Big 5 Personality Test and to create an application that will allow people with similar personality to know each other.

Analysis

The dataset we used comes from Kaggle (https://www.kaggle.com/tunguz/big-five-personality-test) and was collected through an interactive on-line personality test.
In the dataset, 10 questions are available for each personality trait, with answers on a five point scale, labeled 1 = Disagree, 3 = Neutral and 5 = Agree. The corresponding time needed to answer each question is also present in ms.

Extroversion			Neurotic			Agreeableness			Openness			Conscientiousness			Time
10			10			10			10			10			50
EXT1	**…**	**EXT10**	**EST1**	**…**	**EST10**	**AGR1**	**…**	**AGR10**	**OPN1**	**…**	**OPN10**	**CSN1**	**…**	**CSN10**	**EXT1_E**	**…**
2		4	5		1	3		5	4		5	1		5	3400 ms	
4		5	1		2	2		4	5		3	2		3	1900 ms	

1015341 x 100

Data Preprocessing

Start

Data Cleaning

868933 x 100

- null values
- zeroes (not meaningful value or error in time misuration)

Numerosity Reduction

(storage and computation)

50000 x 100

Dimensionality Reduction

(Variance Threshold)

Selecting the 5 questions attributes with the highest variance for each personality, and the corresponding time

50000 x 50

Data Preprocessing

Extroversion			Neurotic			Agreeableness			Openness			Conscientiousness			Time	
5			5			5			5			5			25	
EXT1	**…**	**EXT10**	**EST1**	**…**	**EST10**	**AGR1**	**…**	**AGR9**	**OPN1**	**…**	**OPN8**	**CSN2**	**…**	**CSN9**	**EXT1_E**	**…**
2		4	5		1	3		5	4		5	1		5	3400 ms	
4		5	1		2	2		4	5		3	2		3	1900 ms	

50000 x 50

Selected questions:

['EXT1', 'EXT2', 'EXT7', 'EXT9', 'EXT10']
['EST1', 'EST6', 'EST8', 'EST9', 'EST10']
['AGR1', 'AGR3', 'AGR5', 'AGR6', 'AGR9']
['CSN2', 'CSN4', 'CSN5', 'CSN6', 'CSN9']
['OPN1', 'OPN2', 'OPN4', 'OPN6', 'OPN8']

Data Preprocessing

EXT1	…	EXT10
4		3

→

EXT
(4+..+3)/5

... for each personality …

EXT1_E	…	CSN9_E
2500		3500

→

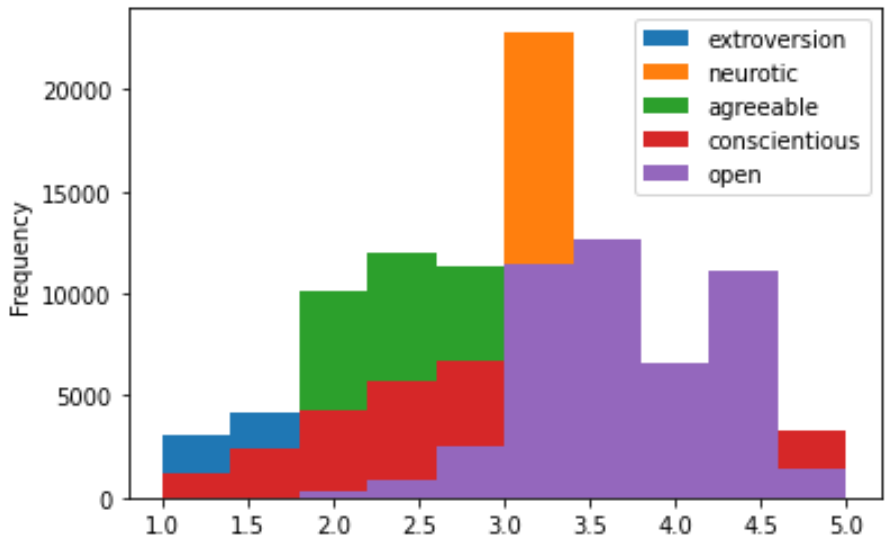
TimeSpent
(2500+..+3500)/(1000*60)

Data Visualization

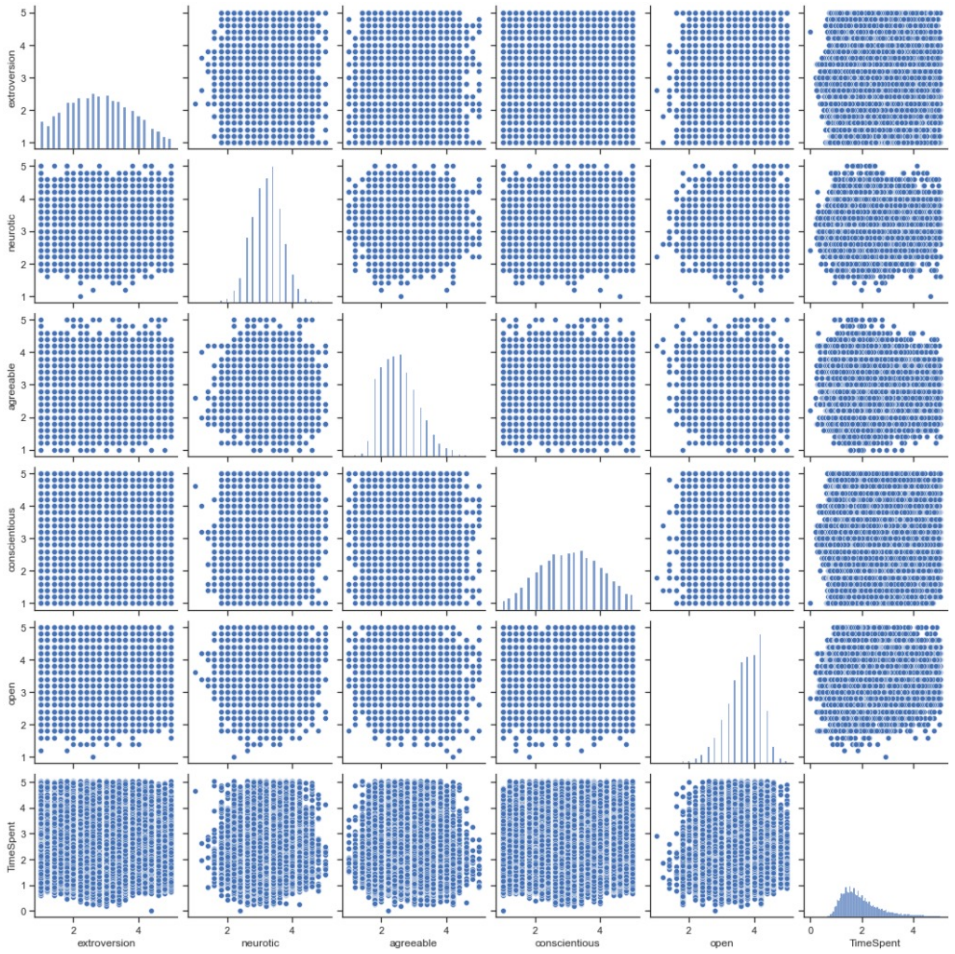
Extroversion	Neurotic	Agreeableness	Openness	Conscientiousness	TimeSpent
2.2	3.4	3.2	4.9	1.4	3.4
4.3	1.2	2.5	5	2.9	4.5
….	….	….	….	….	….

50000 x 6

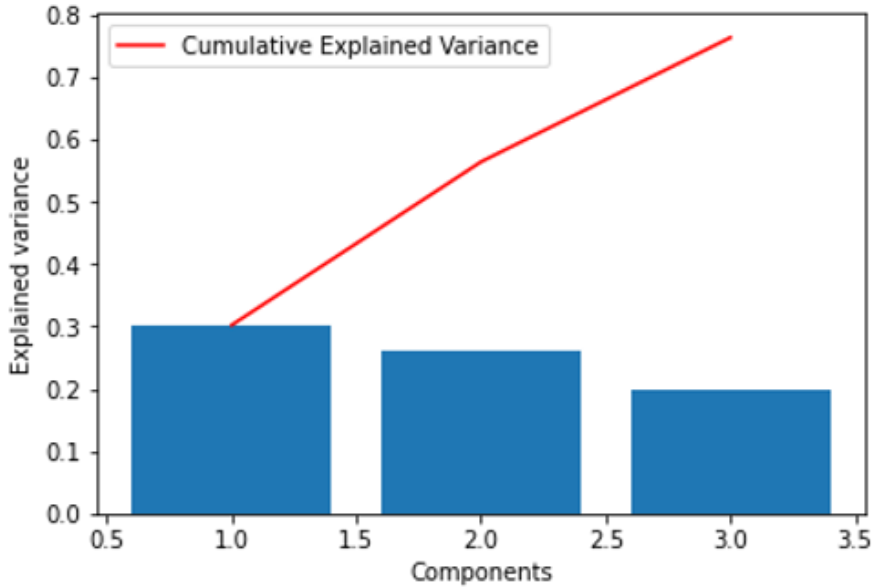
Histogram



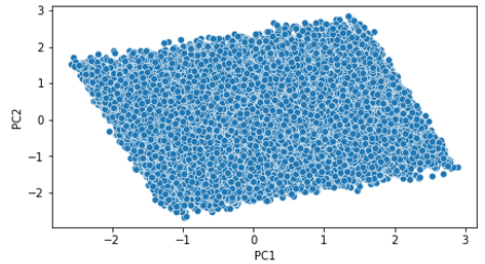
Scatterplot Matrix

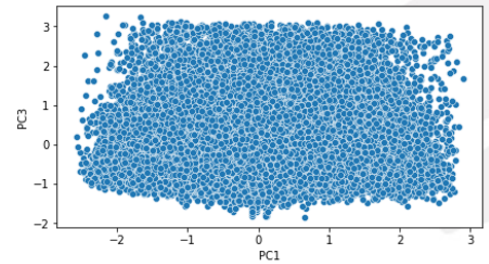


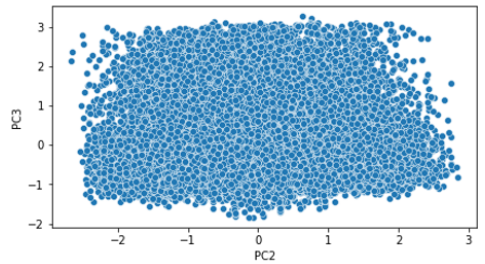
Principal Components Analysis



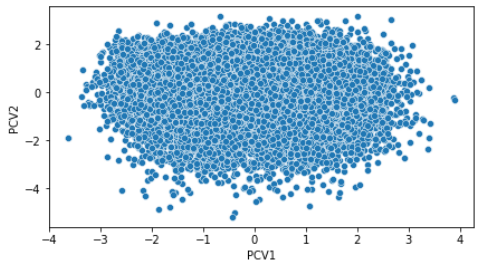
Cumulative Explained Variance using three components: **0.76**

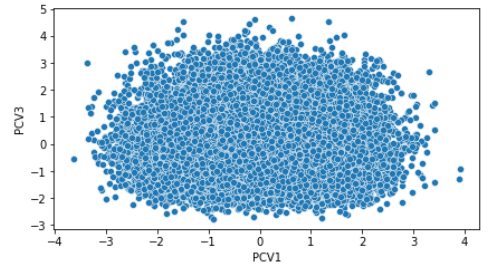


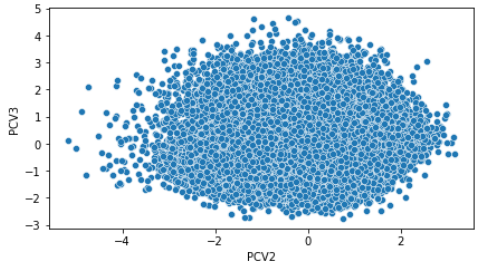




Varimax Rotation







Cumulative Explained Variance using three components and Varimax: **0.57**

Clustering Algorithms

- Partitioning Algorithms: K-Means and PAM
- Hierarchical Algorithms: AGNES and BIRCH
- Density Based Clustering Algorithms: DBSCAN

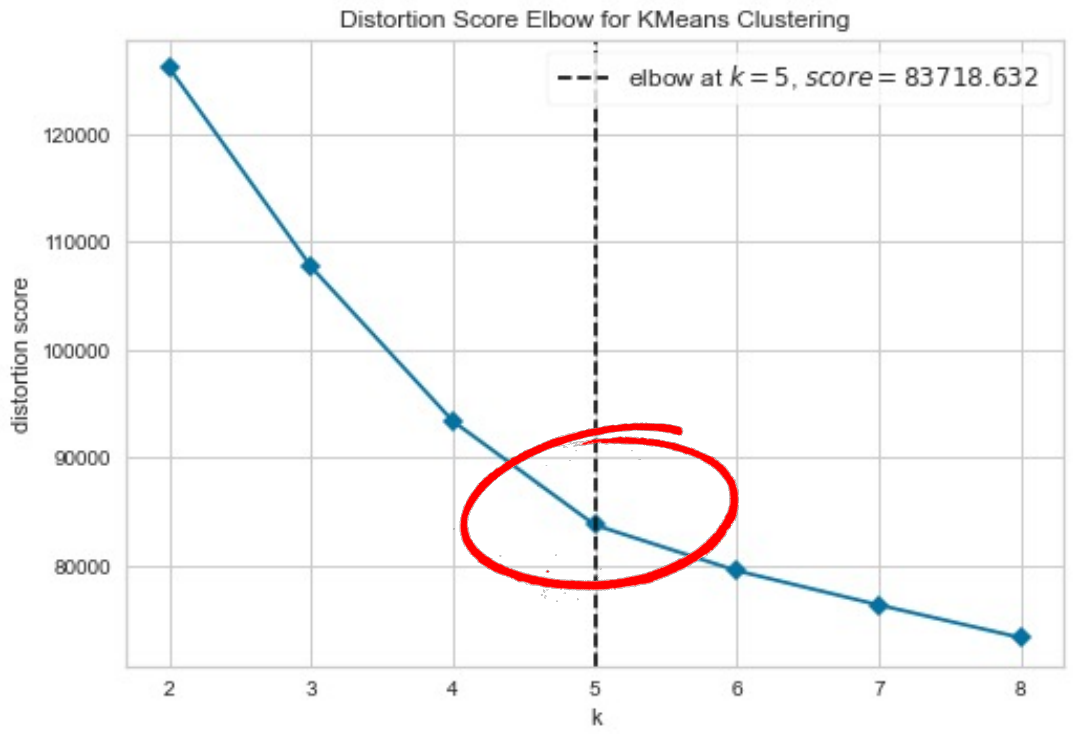
Hopkins Statistic

$$H = \frac{\sum_{i=1}^{n} yi}{\sum_{i=1}^{n} xi + \sum_{i=1}^{n} yi} = 0.73$$

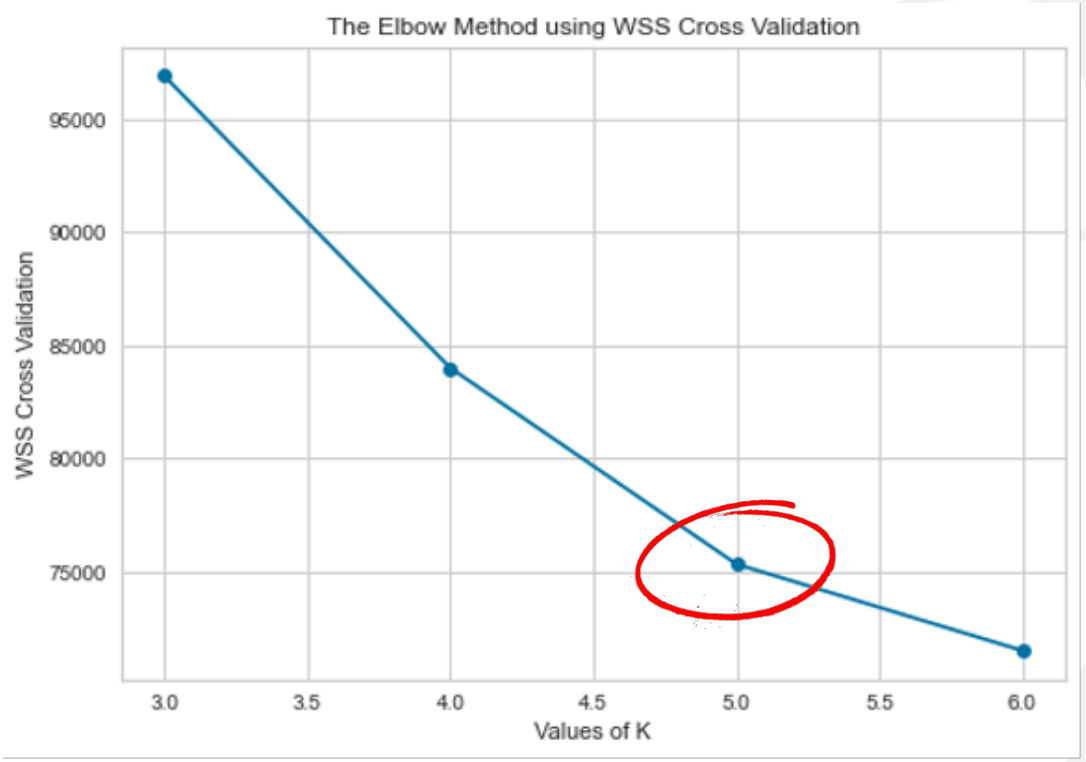
assessing clustering tendency with a confidence of approximately 90%.

K-Means : Choosing the best value of K

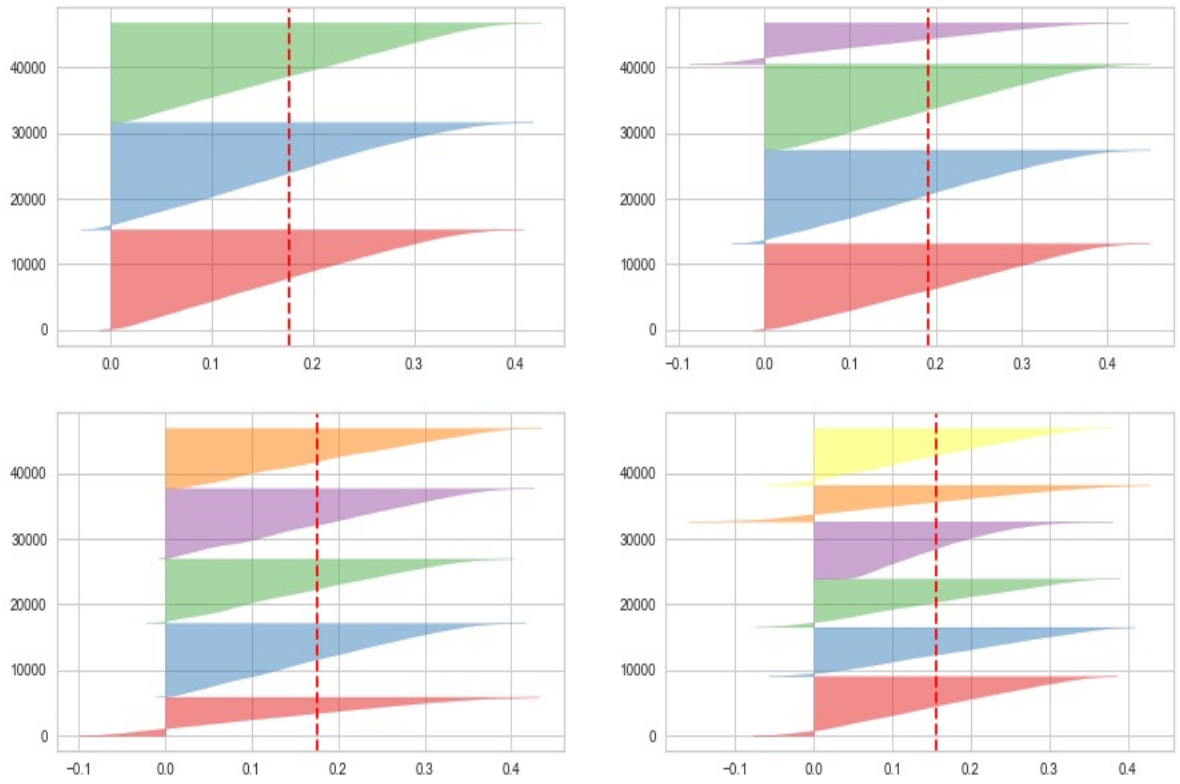
Elbow Method



K-Fold Cross Validation

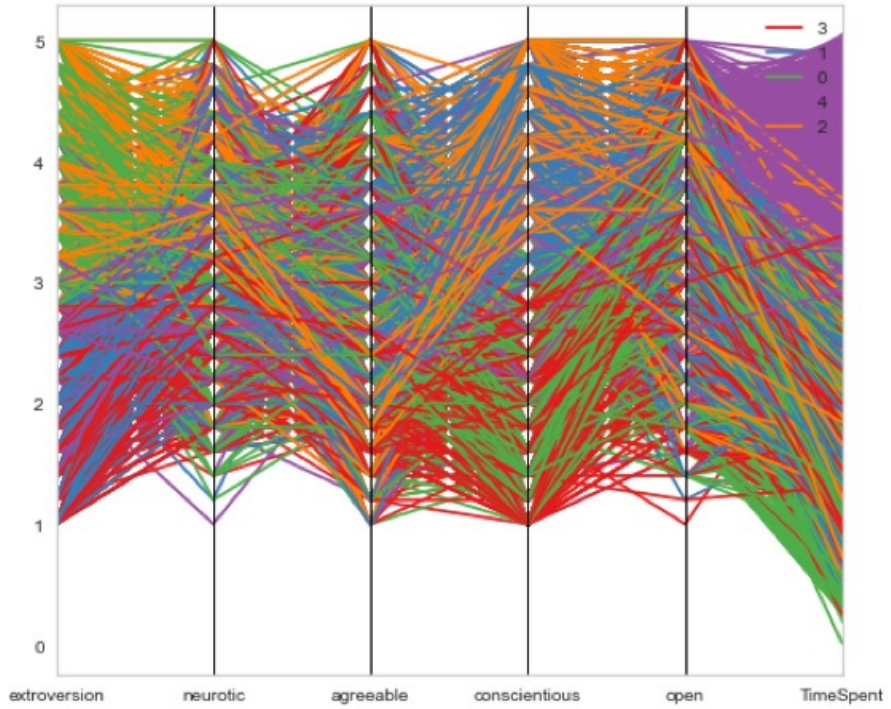


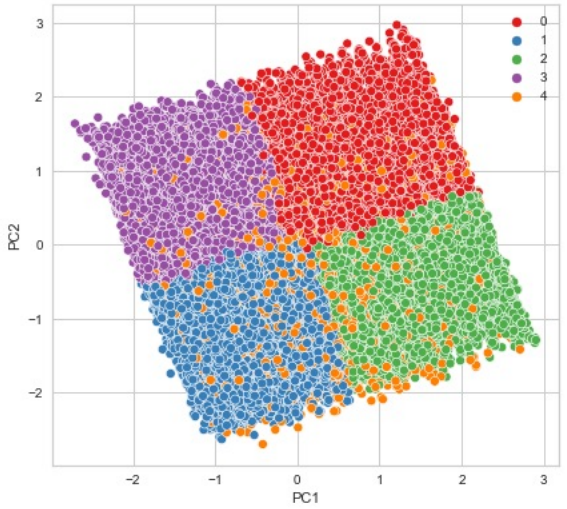
K-Means : Choosing the best value of K

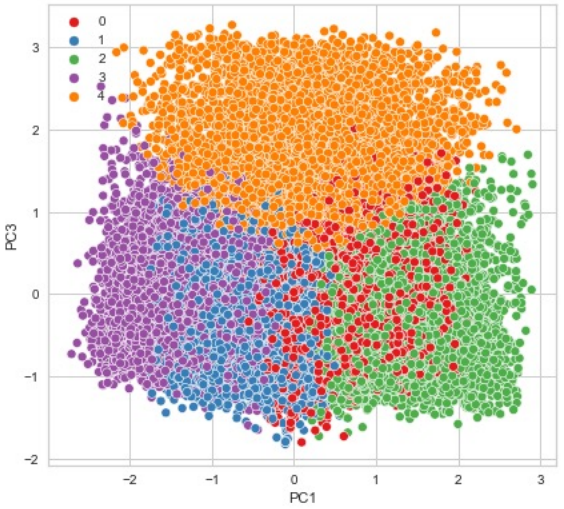


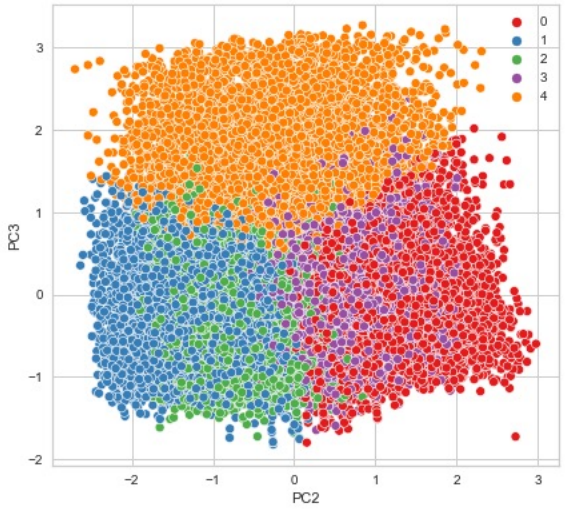
	Silhouette Score	Silhouette mean for each cluster	Std of clusters' silhouette	Number of negative single silhouette values
K = 3	0,177	[0,17, 0,17, 0,19]	0,012	961
K = 4	0,19	[0,20, 0,20, 0,18, 0,15]	0,023	1689
K = 5	0,176	[0,20, 0,18, 0,14, 0,17, 0,18]	0,019	1524
K = 6	0,156	[0,13, 0,15, 0,18, 0,16, 0,16, 0,15]	0,012	3821

K-Means with K = 5





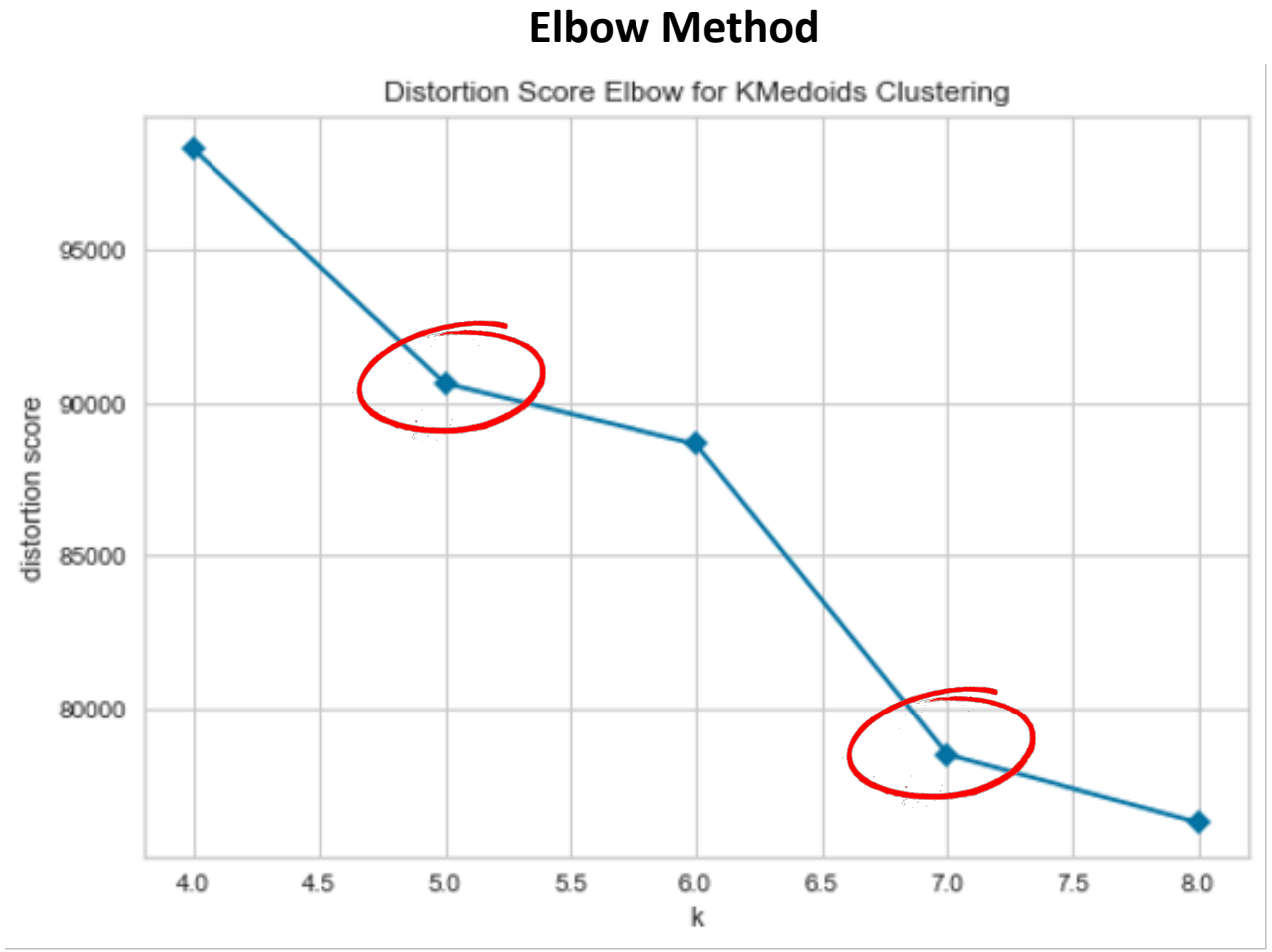




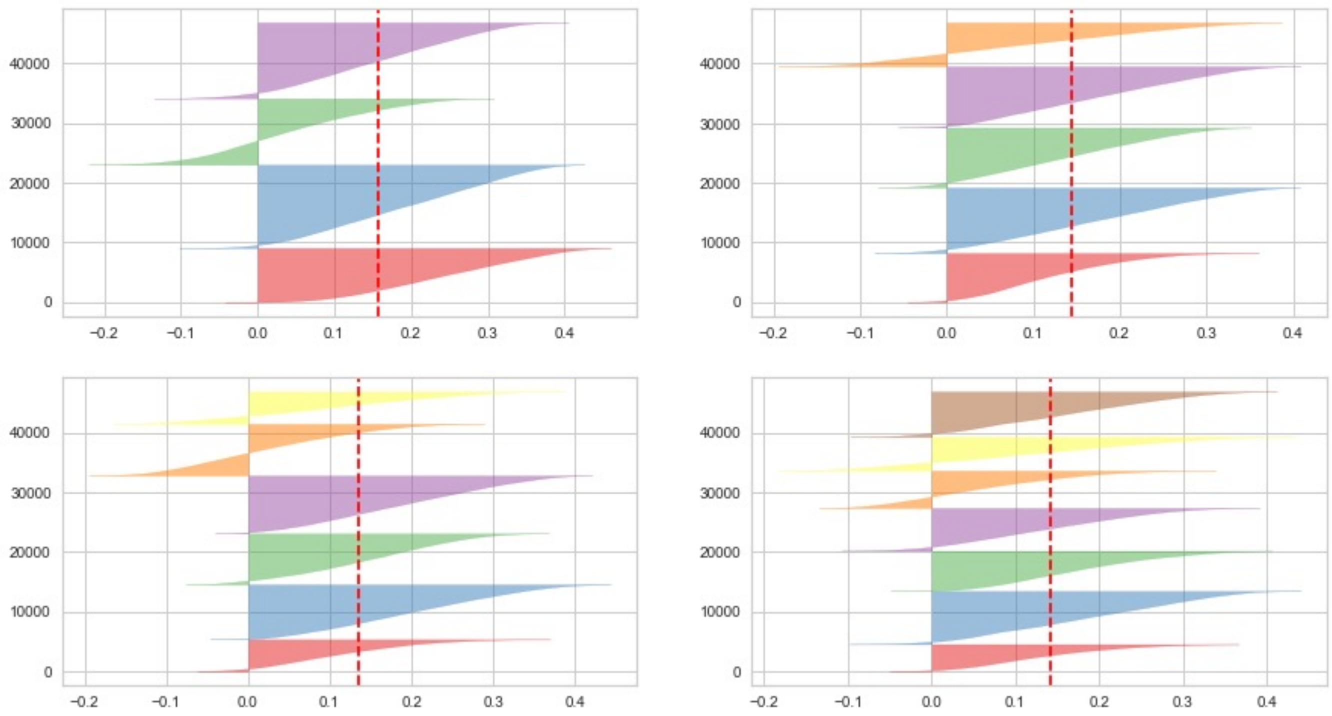
Cluster's size

Cluster 0	9126
Cluster 1	11374
Cluster 2	10142
Cluster 3	10396
Cluster 4	5837

PAM: Choosing the best value of K

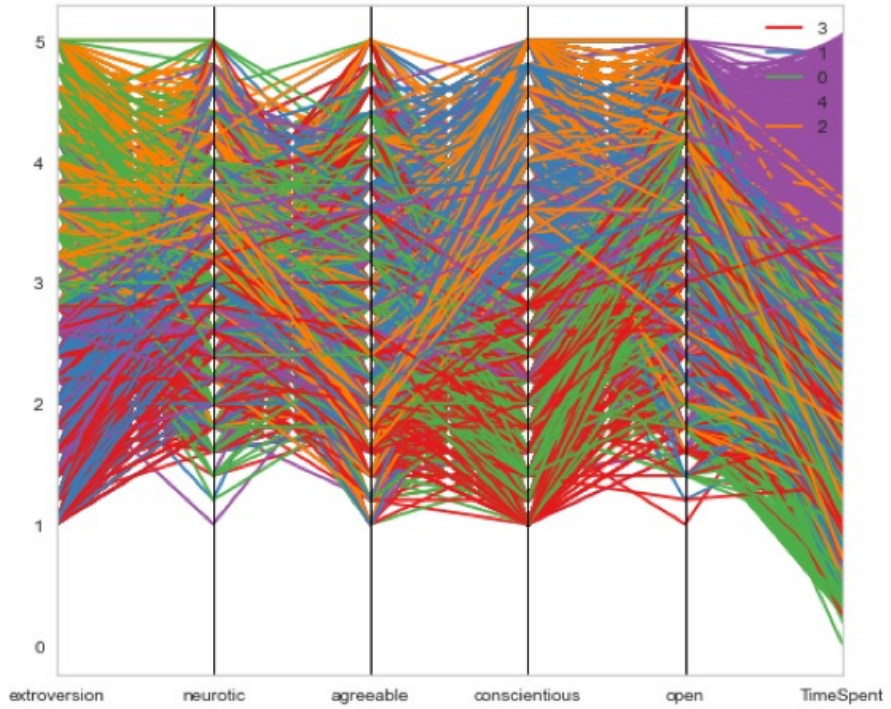


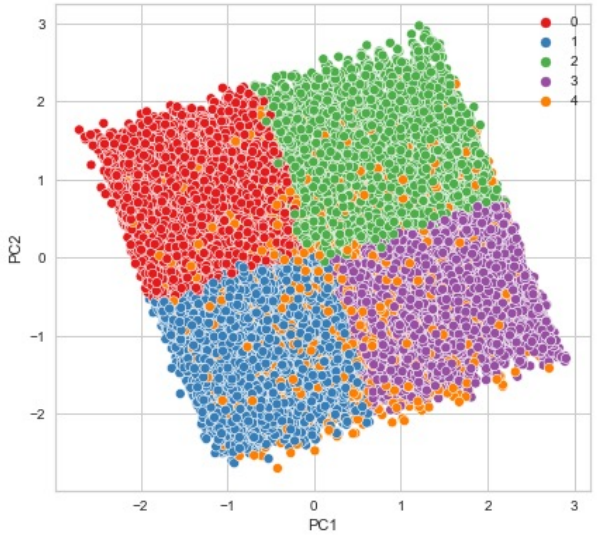
K-Means : Choosing the best value of K

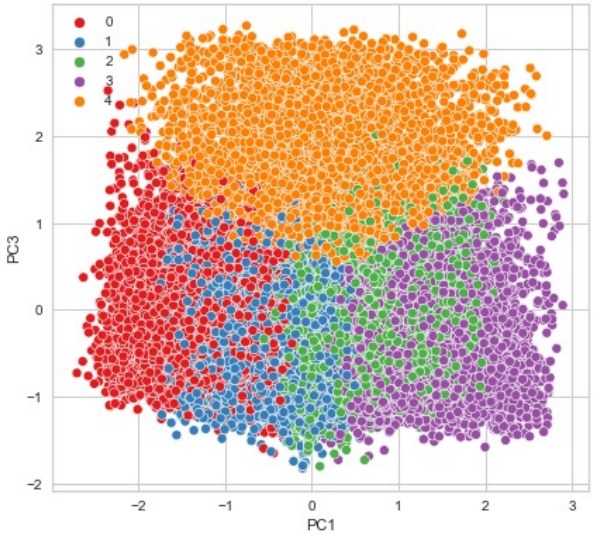


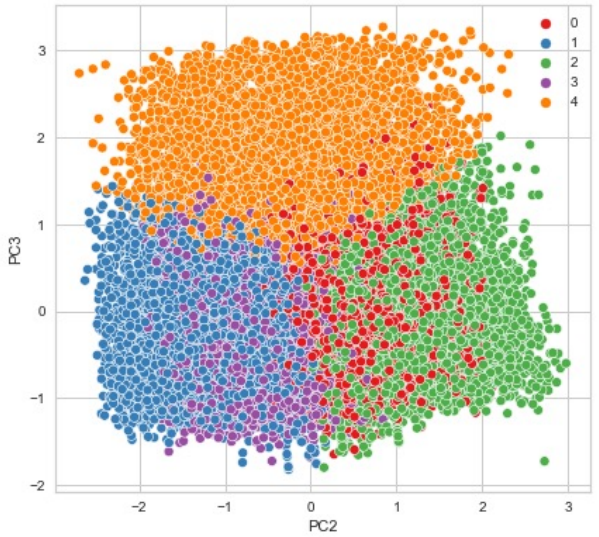
	Silhouette Score	Silhouette mean for each cluster	Std of clusters' silhouette	Number of negative single silhouette values
K = 4	0,158	[0,25, 0,19, 0,004, 0,16]	0,08	5575
K = 5	0,145	[0,12, 0,17, 0,14, 0,18, 0,09]	0,03	4194
K = 6	0,13	[0,11, 0,20, 0,15, 0,19, 0,02, 0,09]	0,06	6518
K = 7	0,14	[0,13, 0,19, 0,18, 0,14, 0,07, 0,12, 0,16]	0,04	5201

PAM with K = 5









Cluster's size

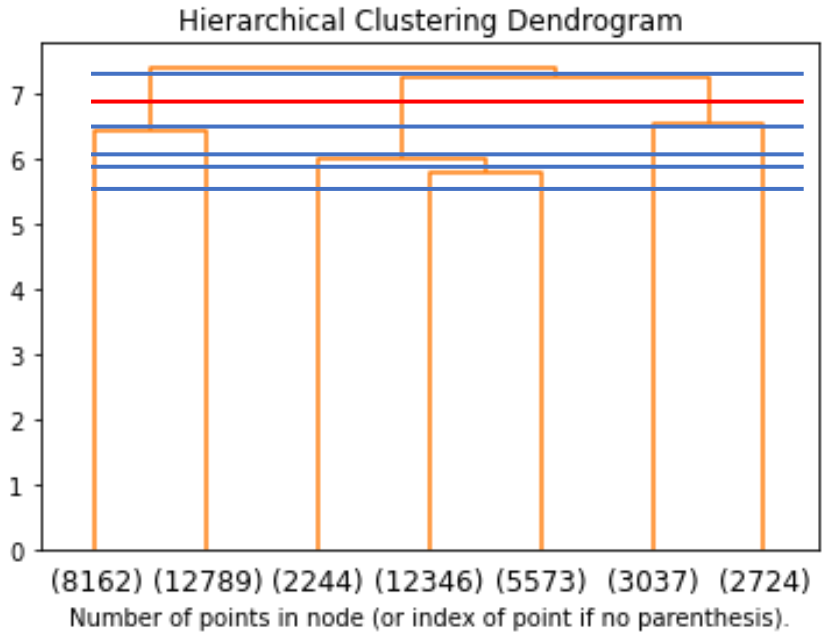
Cluster 0	10364
Cluster 1	11375
Cluster 2	10193
Cluster 3	9109
Cluster 4	5834

Comparing K-Means with PAM

99.8% of points clustered in the same way!

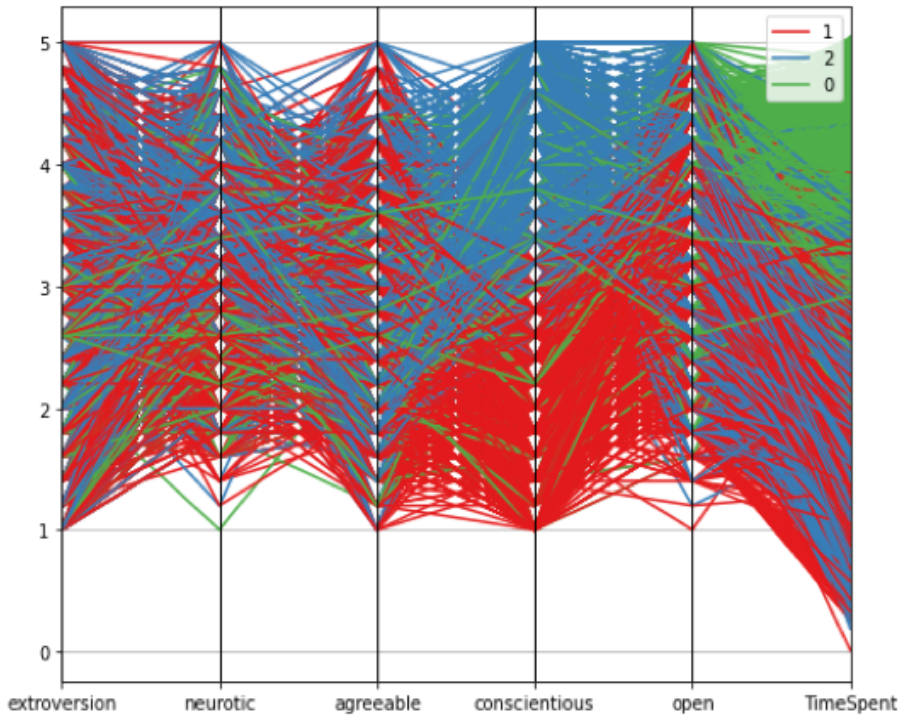
	Time Needed	Silhouette Score	DBI	CHS
K-Means	0,35 seconds	0,176	1,48	10543,06
PAM	135,15 seconds	0,145	1,48	10543,14

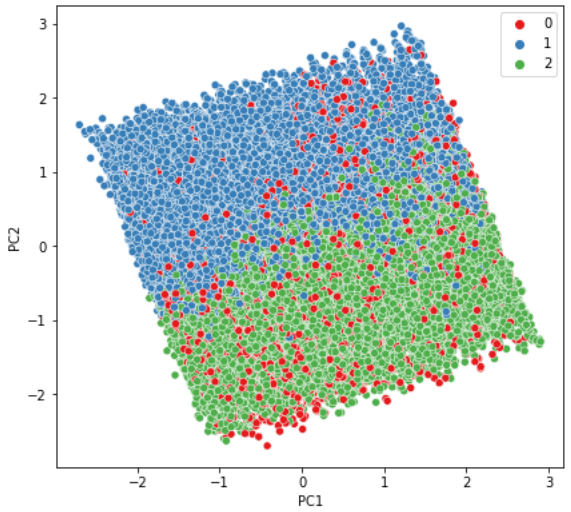
AGNES: Complete Linkage

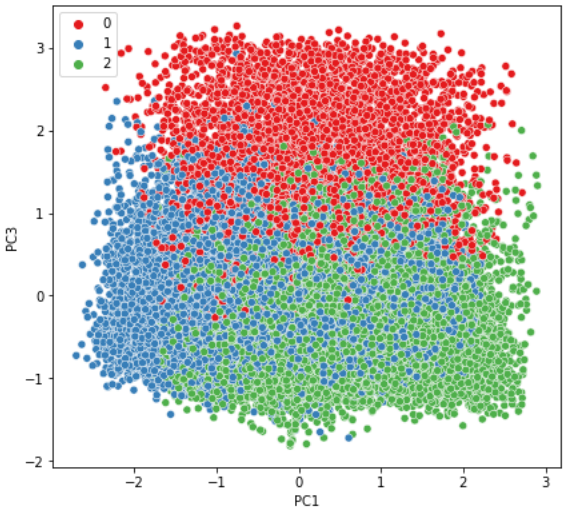


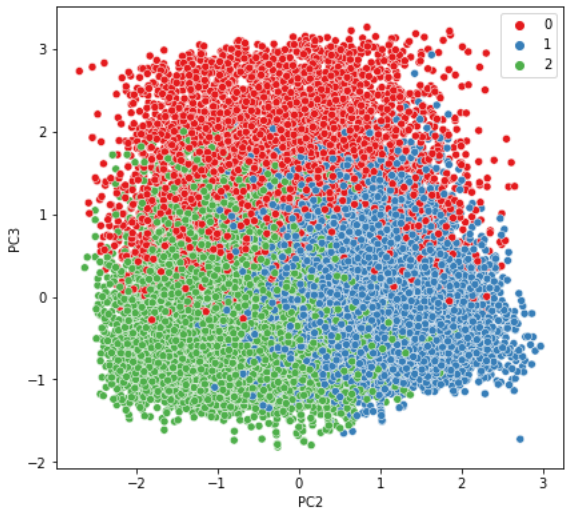
	AGNES (Complete Linkage)
Silhouette Score	0,138
Silhouette mean for each cluster	[0,09, 0,13, 0,16]
Std of cluster's silhouette	0,03
Number of negative single silhouette values	7791
CHS	7060
DBI	1,95
Time	122 seconds

AGNES: Complete Linkage





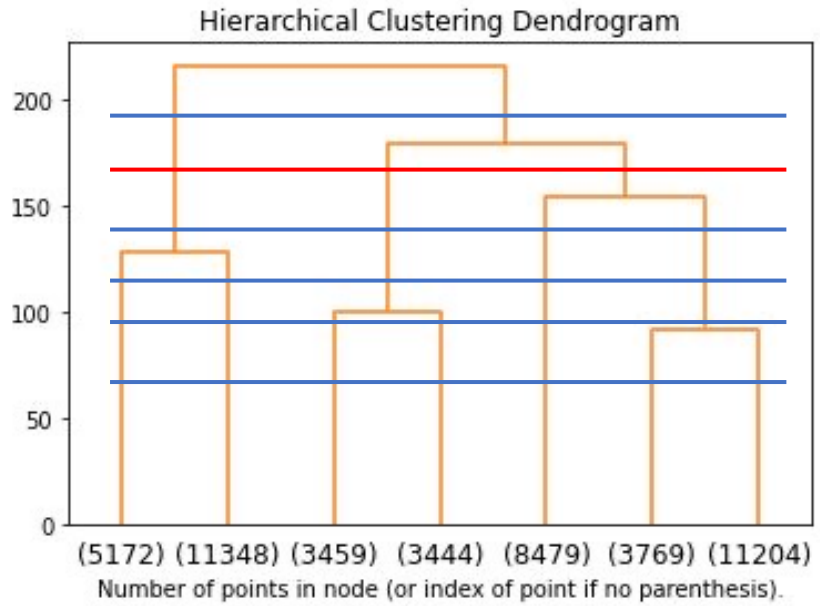




Cluster's size

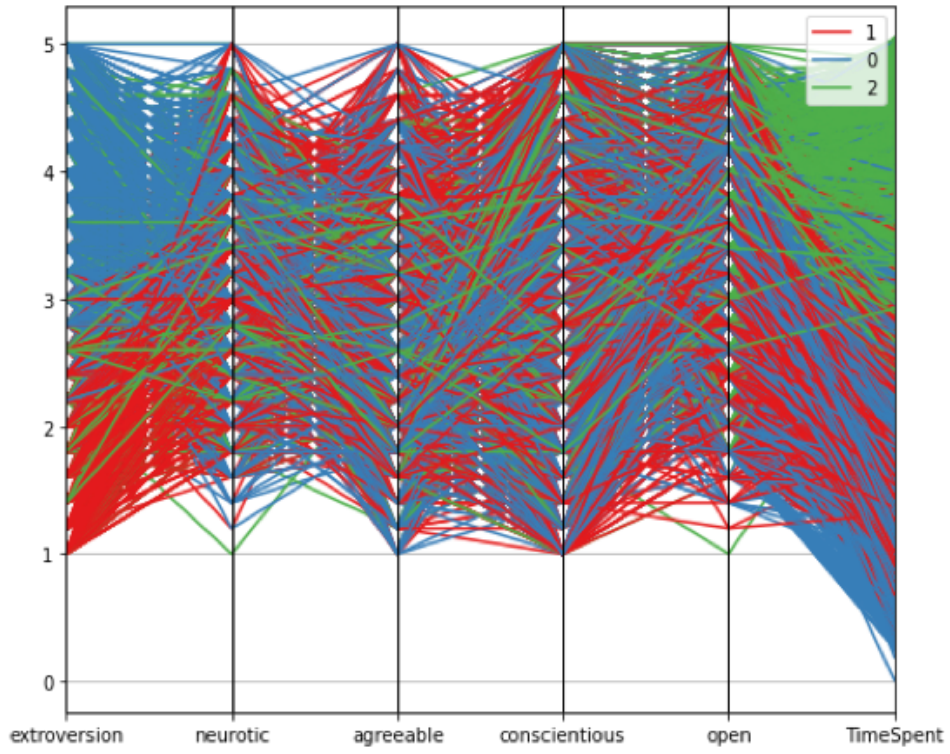
Cluster 0	5761
Cluster 1	20951
Cluster 2	20163

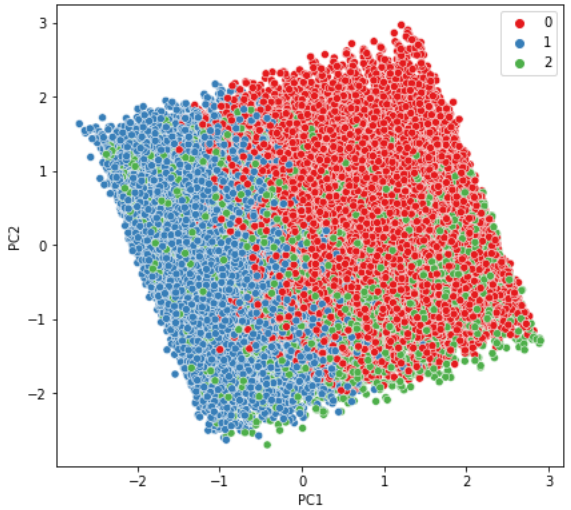
AGNES: Ward Linkage

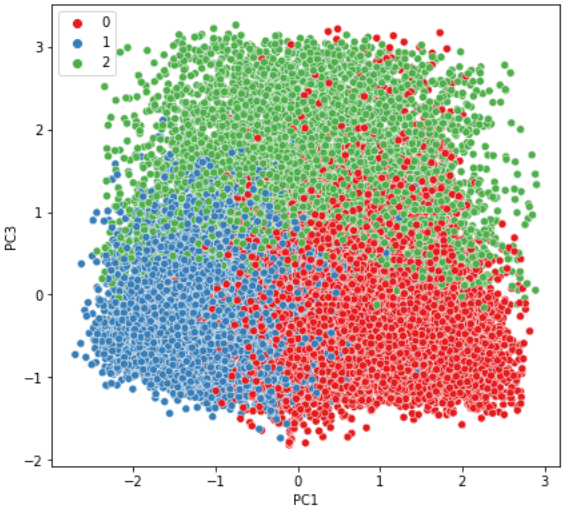


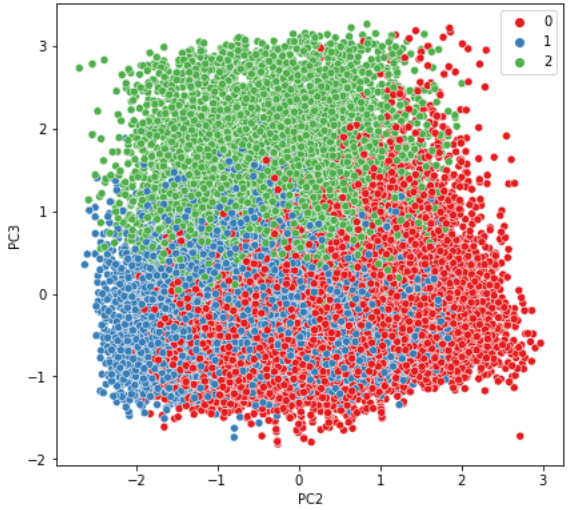
	AGNES (Ward Linkage)
Silhouette Score	0,145
Silhouette mean for each cluster	[0,14, 0,18, 0,09]
Std of cluster's silhouette	0,03
Number of negative single silhouette values	7987
CHS	7748
DBI	1,88
Time	148 seconds

AGNES: Ward Linkage





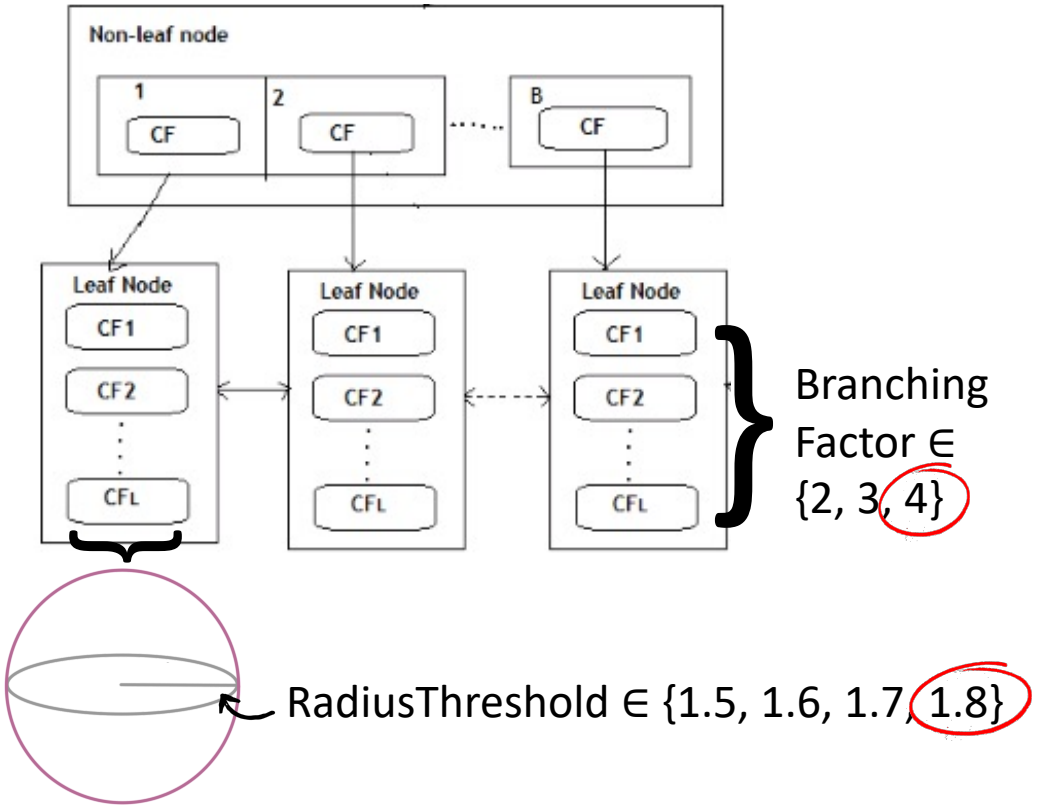




Cluster's size

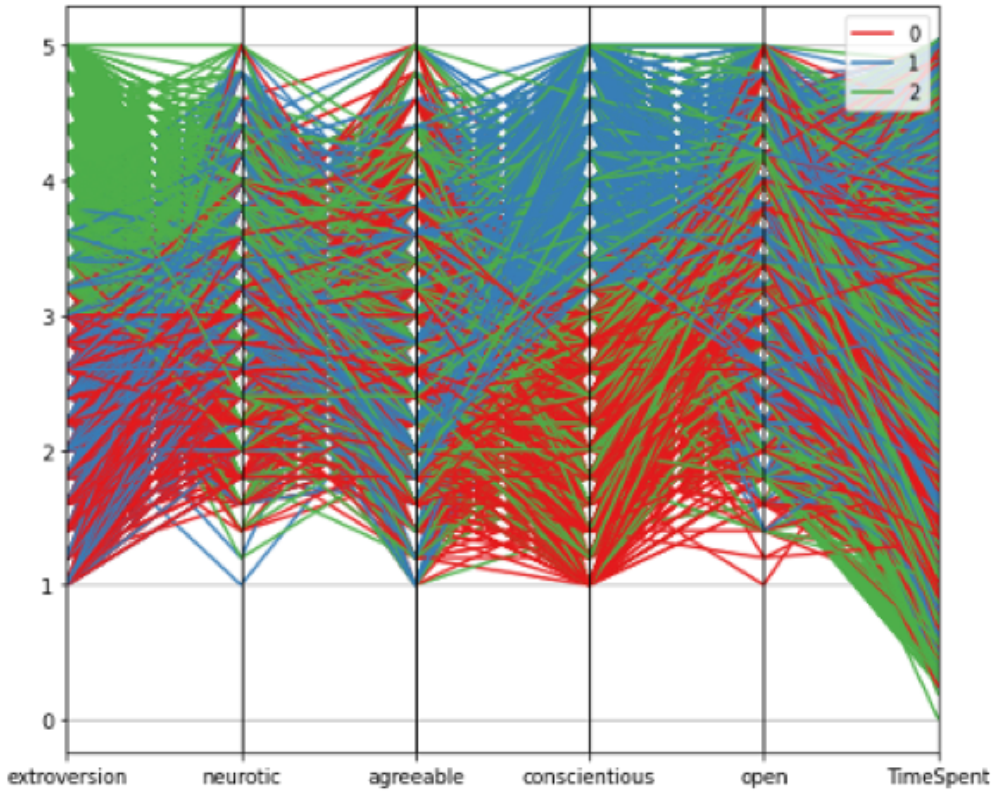
Cluster 0	23452
Cluster 1	16520
Cluster 2	6903

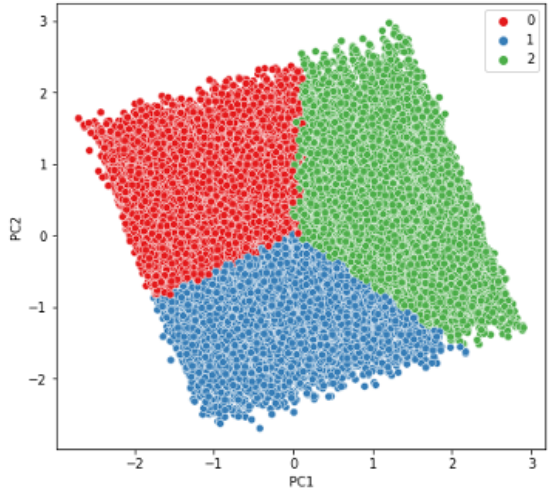
BIRCH

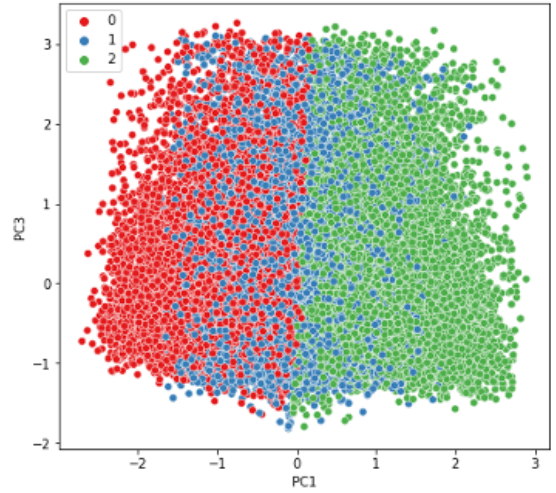


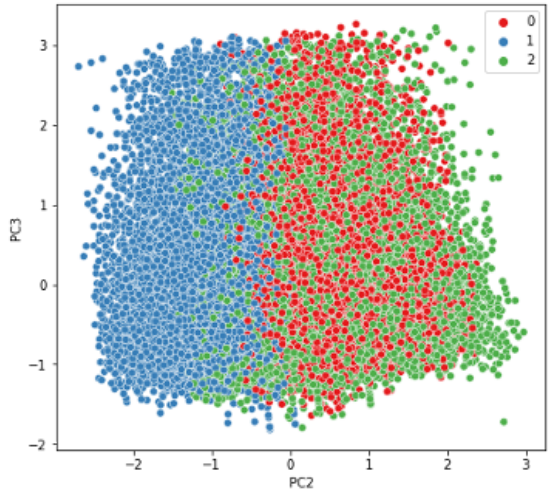
	BIRCH
Silhouette Score	0,177
Silhouette mean for each cluster	[0,20, 0,17, 0,17]
Std of cluster's silhouette	0,01
Number of negative single silhouette values	884
CHS	5087,64
DBI	1,81
Time	0,88 seconds

BIRCH









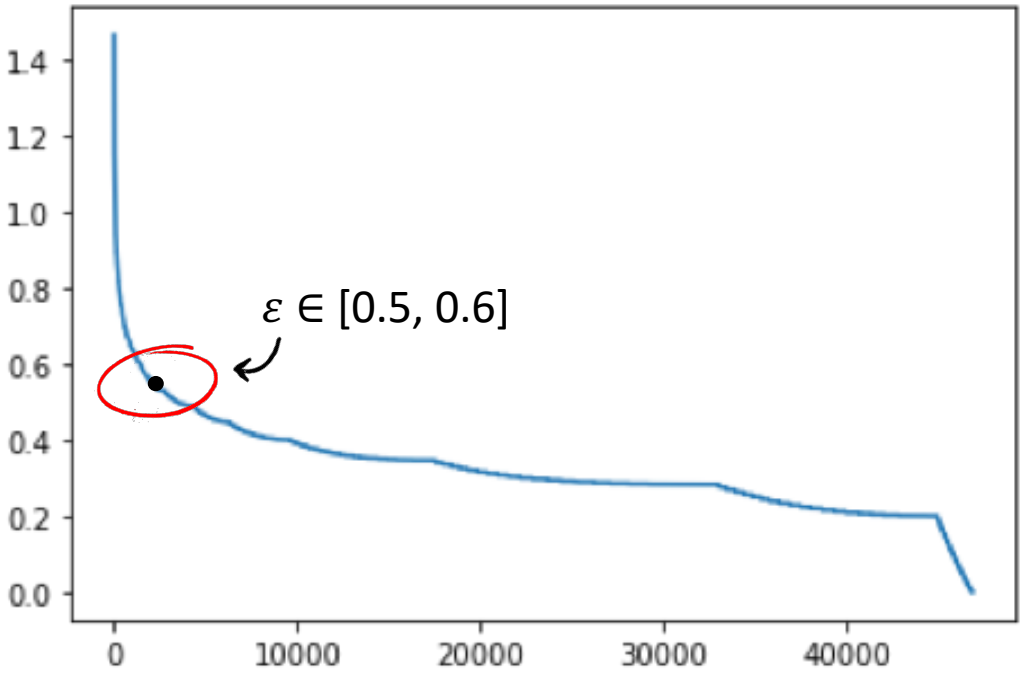
Cluster's size

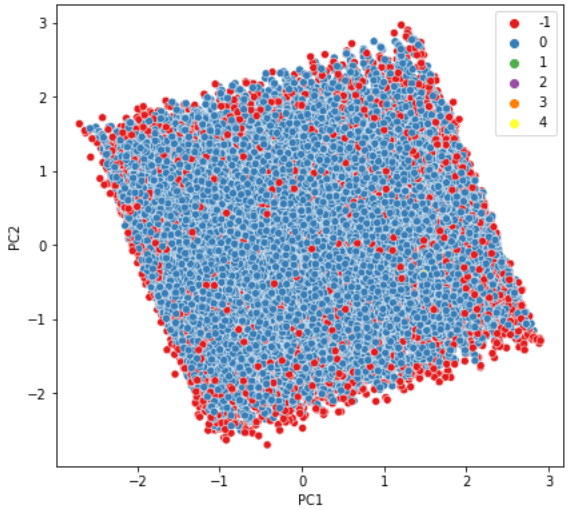
Cluster 0	15347
Cluster 1	15300
Cluster 2	16228

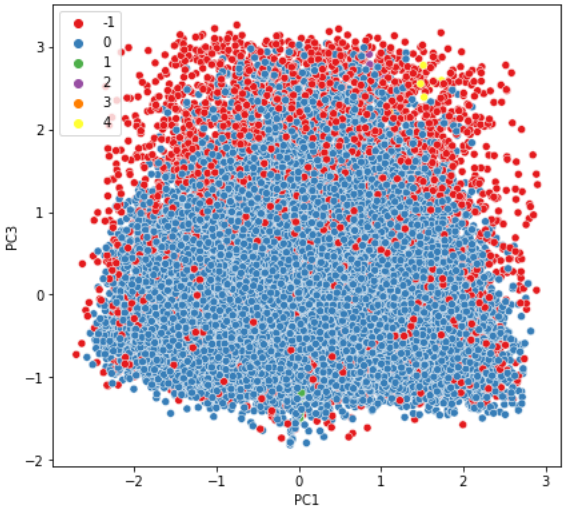
DBSCAN

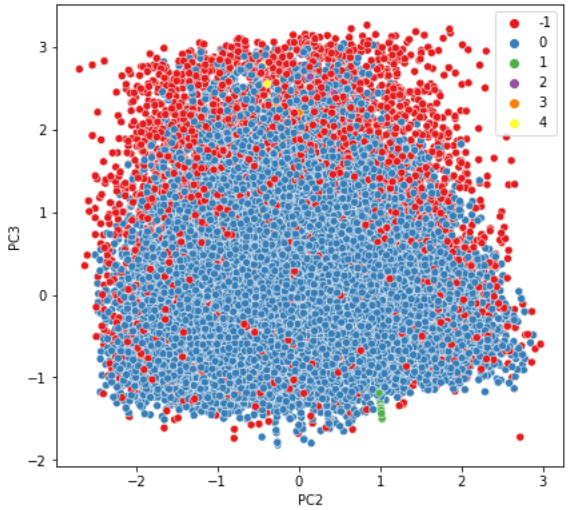
K = MinPts = 2 * numDimensions = 12

(Sander et al., 1998)









	Cluster's size
Cluster -1	4244
Cluster 0	42568
Cluster 1	27
Cluster 2	7
Cluster 3	5
Cluster 4	6

Other strategy…
Spectral Clustering & NG-Jordan Weiss Algorithm

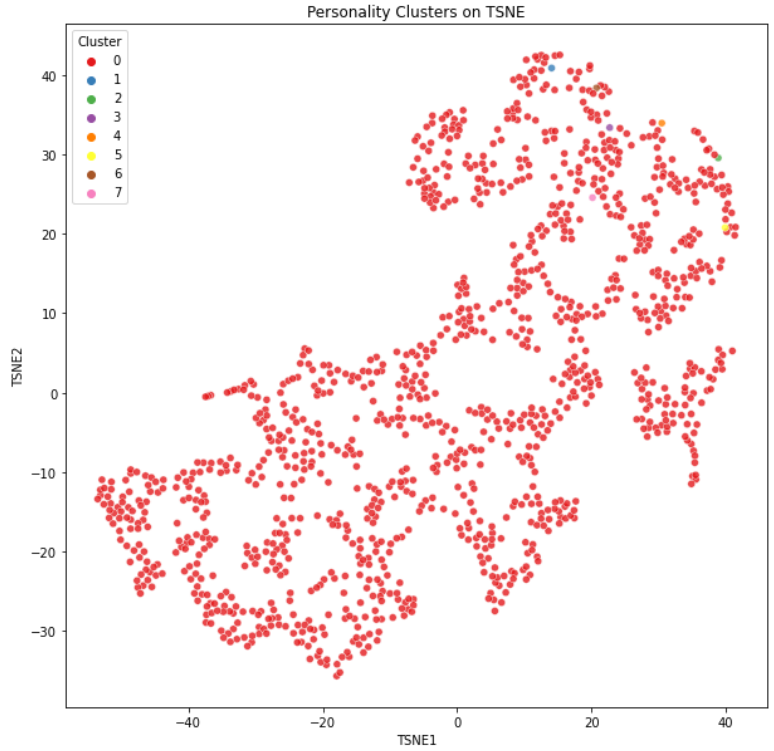
We considered all 100 attributes and applied clustering on high dimensional data.

EXT1	…	EXT10	EST1	…	EST10	AGR1	…	AGR10	OPN1	…	OPN10	CSN1	…	CSN10	EXT1_E	…
2		4	5		1	3		5	4		5	1		5	3400 ms	
4		5	1		2	2		4	5		3	2		3	1900 ms	

Hopkins Statistic

$$H = \frac{\sum_{i=1}^{n} yi}{\sum_{i=1}^{n} xi + \sum_{i=1}^{n} yi} = 0.72$$

Considering just a subset of instances for computational reasons..



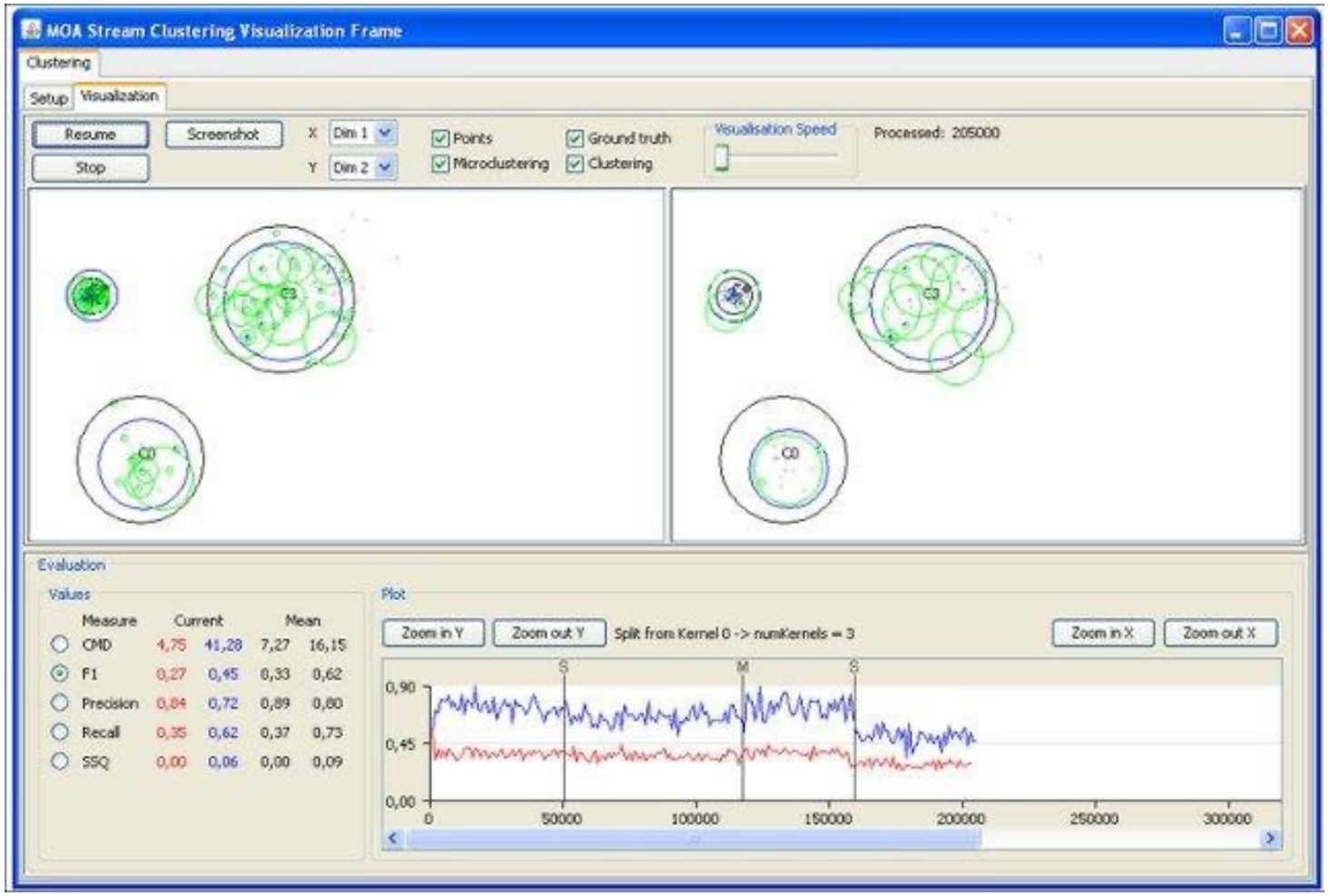
Cluster's size

Cluster 0	1155
Cluster 1	1
Cluster 2	1
Cluster 3	1
Cluster 4	1
Cluster 5	1
Cluster 6	1
Cluster 7	1

Comparison of clustering algorithms

	Silhouette Score	Silhouette mean for each cluster	Std of clusters' silhouette	Number of negative single silhouette values	CHS	DBI	Time
K-Means	0,176	[0,20, 0,18, 0,14, 0,17, 0,18]	0,02	1524	10543,06	1,48	0,35 seconds
PAM	0,145	[0,12, 0,17, 0,14, 0,18, 0,09]	0,03	4194	10543,14	1,48	135,15 seconds
AGNES (Ward Linkage)	0,145	[0,14, 0,18, 0,09]	0,03	7987	7748	1.88	148 seconds
BIRCH	0,177	[0,20, 0,17, 0,17]	0,01	884	5087,64	1,81	0,88 seconds

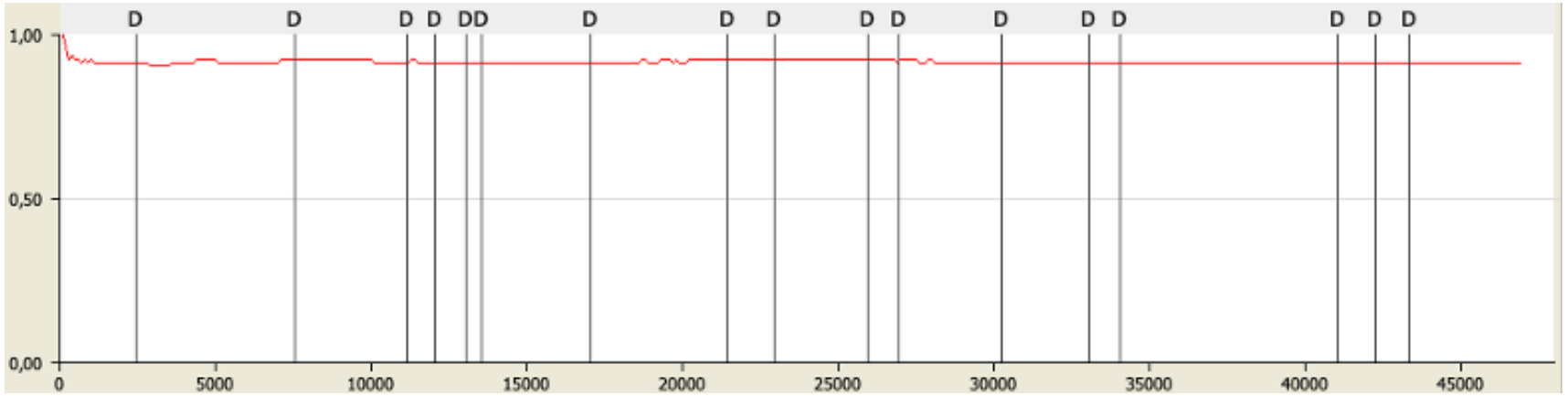
Data Streaming Clustering with MOA



THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato
NEW ZEALAND

Concept Drift Evaluator

Using the Basic Concept Drift Performance Evaluator, we can see that 17 statistical changes are detected considering the stream of our 50.000 instances.



MOA Algorithms

- StreamKM++ computes a small weighted sample of the data stream and it uses the k-means++ algorithm as a randomized seeding technique to choose the first values for the clusters. To compute the small sample, it employs coreset constructions using a coreset tree for speed up.

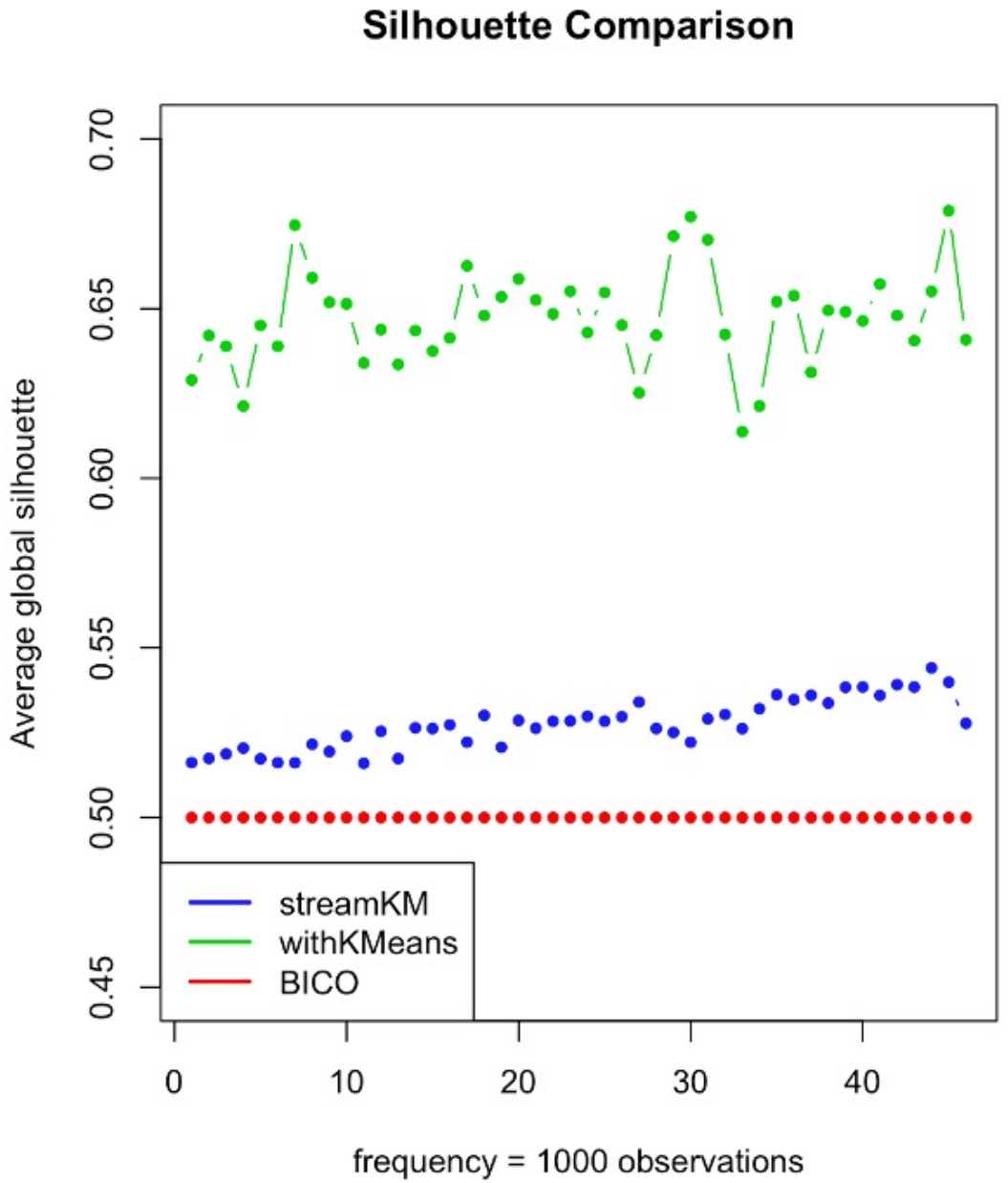
	sizeCoreset	numClusters	length
StreamKM	10000	5	100000

- BICO (an acronym for “BIRCH meets coresets for k-means clustering”) combines the data structure of BIRCH with the theoretical concept of coresets for clustering.

	numClusters	maxClusterFeatures	Projections
BICO	5	100	10

- WithKmeans determines the closest kernel and check whether instance fits into it. If the instance fits, it is put into the kernel, otherwise it applies a free-memory strategy to insert a new kernel (trying to forget oldest kernels); then, it merges the two closest kernels.

	Horizon	maxNumKernels	kernelRadiFactor	K
WithKmeans	1000	100	2	5

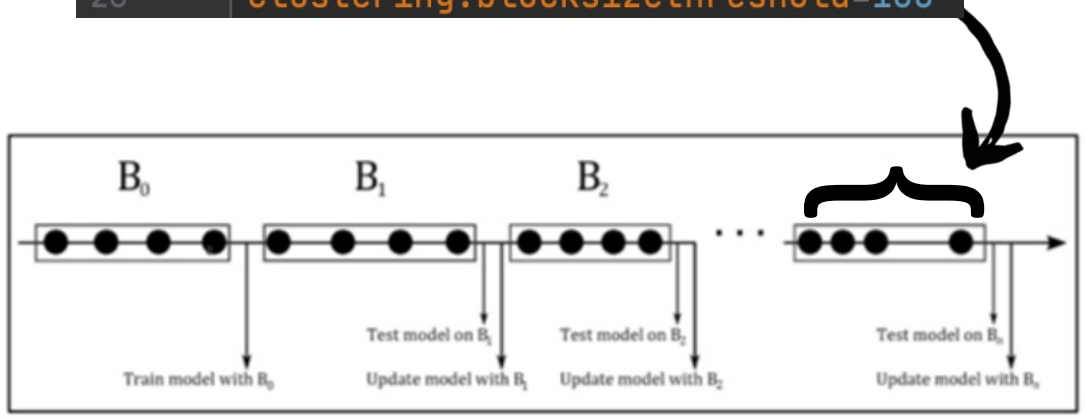


Application

Server Side: Setting Clustering mode

```
application.properties
25  clustering.mode=streaming
```

```
application.properties
25  clustering.mode=standard
26  clustering.blocksizethreshold=100
```



Block Processing

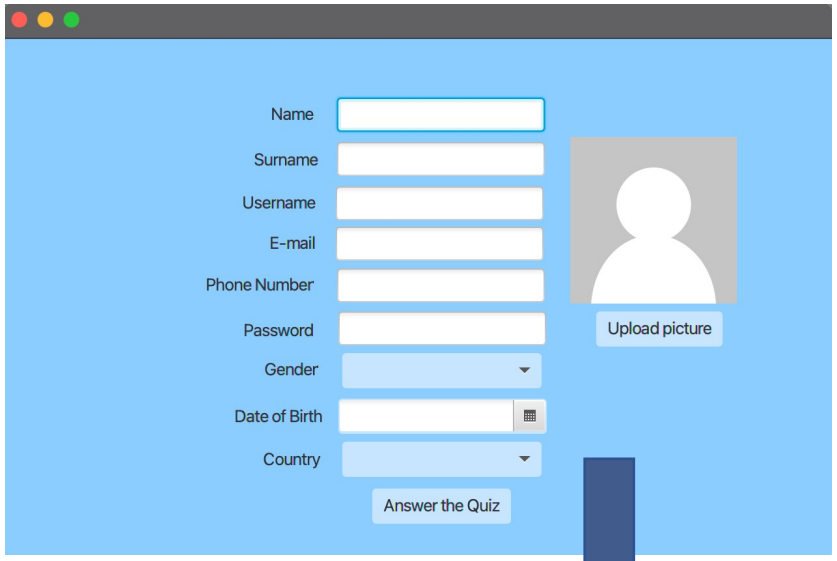
Clustering is performed at the **initialization** of the server.
If some of the users will be grouped in a different cluster than the previous, the value of the cluster ID in the database is updated.

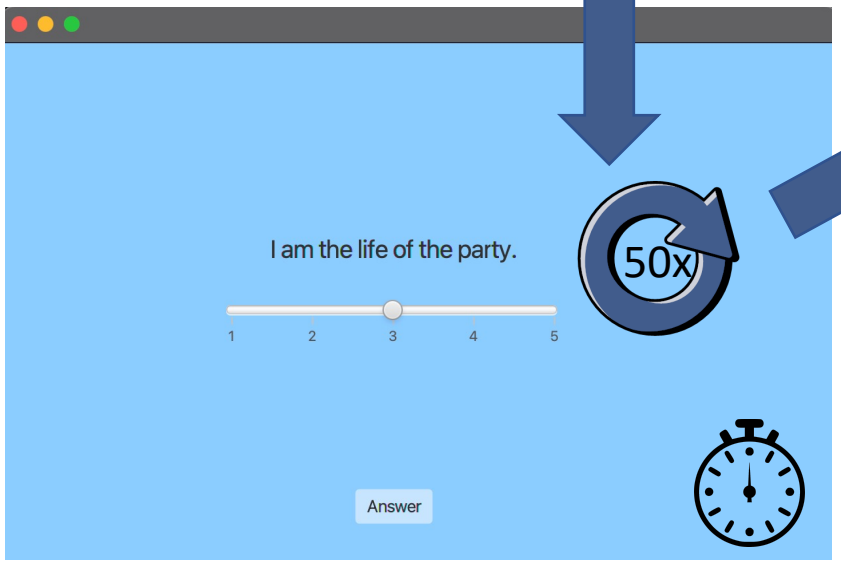
During this updating process, **continuity** of service is guaranteed, which means that a user will be able to register or login anyway, even if clusters are changing.

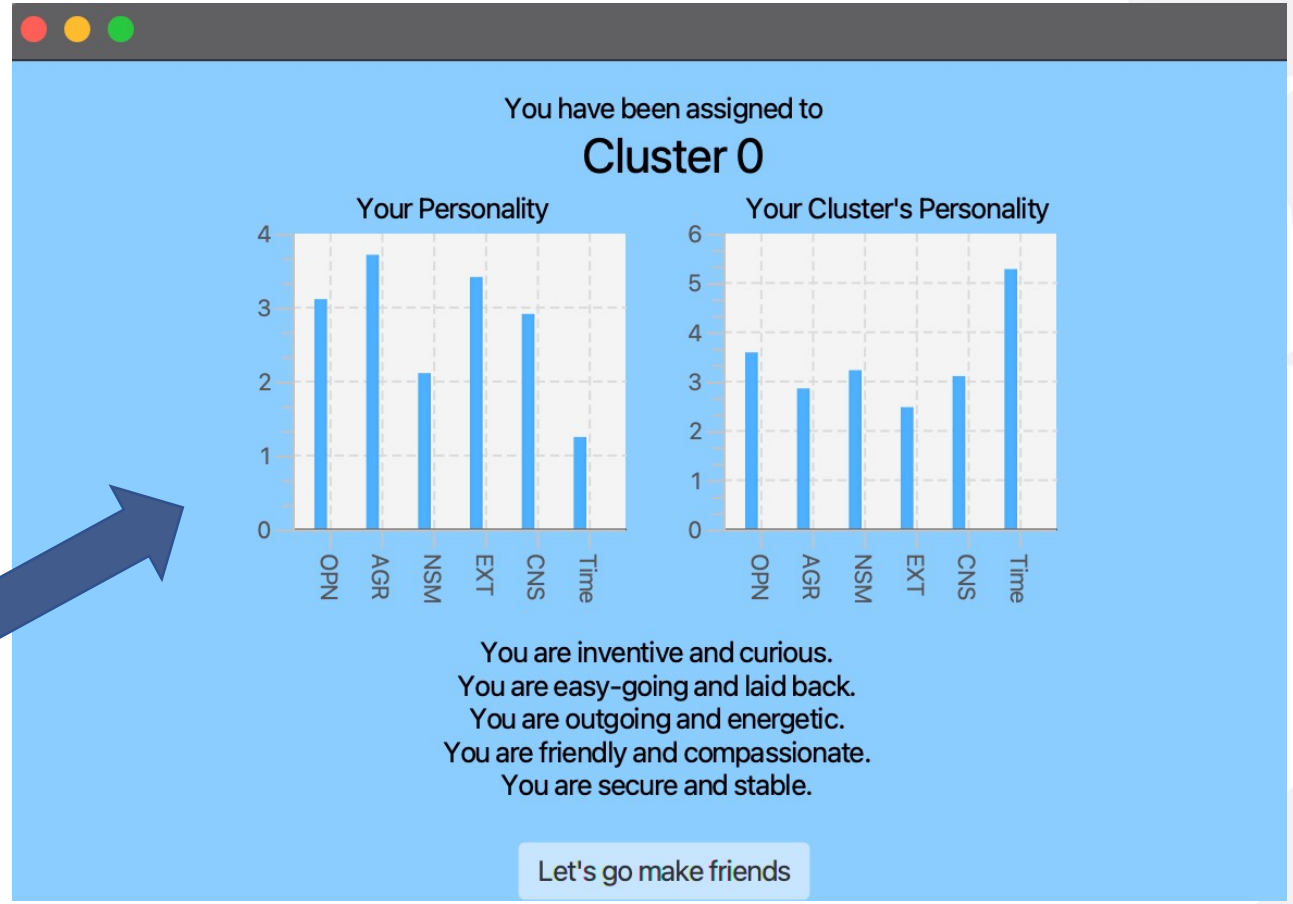
The **registration** of the user in the streaming mode will produce an evolution of the clusters, while in standard mode will produce an evolution of the clusters only if the threshold is reached.

If the streaming mode is set, when the user logins, the system checks if the cluster has changed and if so, updates the databases accordingly.

Survey

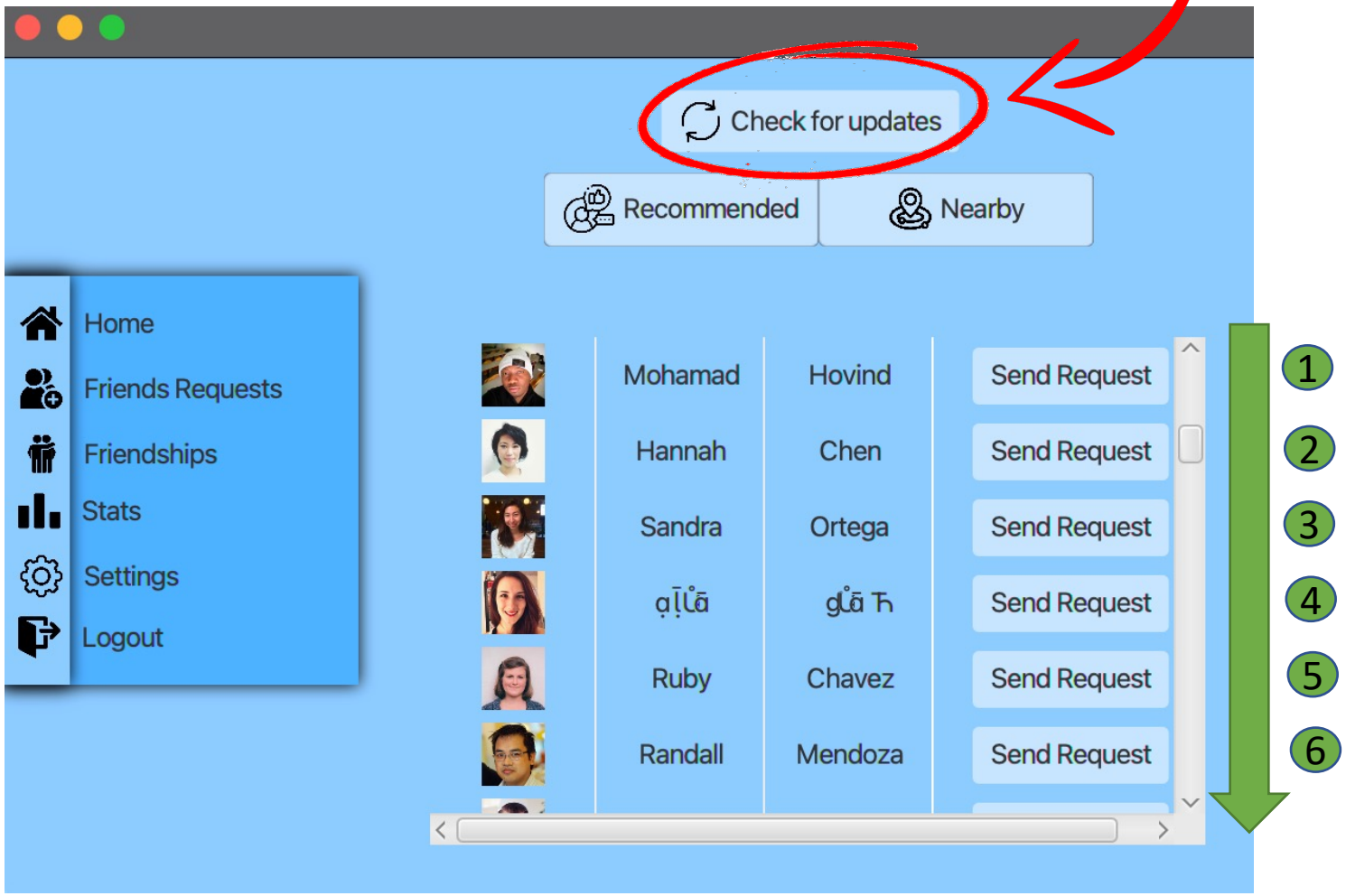

Name
Surname
Username
E-mail
Phone Number
Password
Gender
Date of Birth
Country
Upload picture
Answer the Quiz

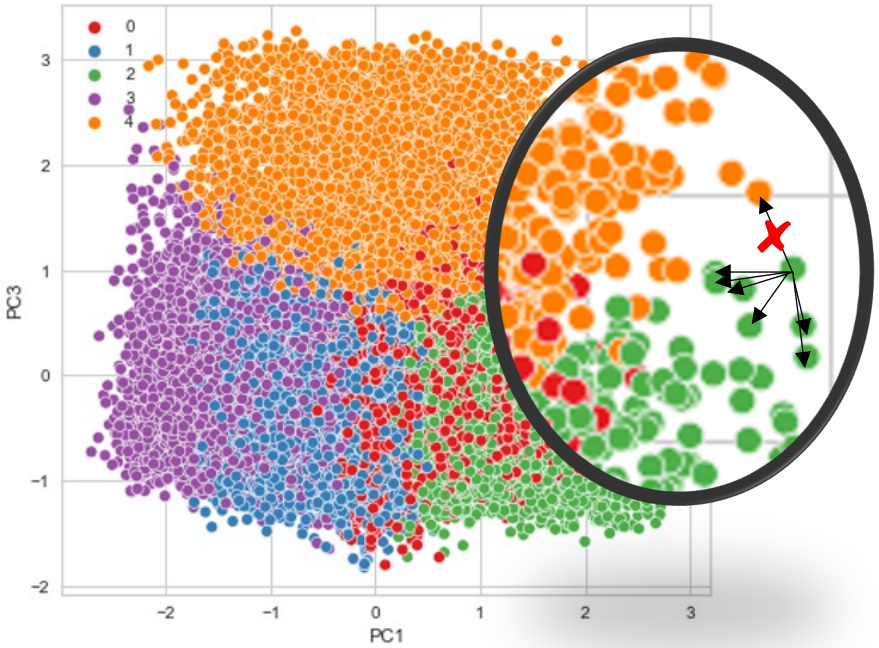

I am the life of the party.
1
2
3
4
5
50x
Answer


You have been assigned to
Cluster 0
Your Personality
Your Cluster's Personality
OPN
AGR
NSM
EXT
CNS
Time
You are inventive and curious.
You are easy-going and laid back.
You are outgoing and energetic.
You are friendly and compassionate.
You are secure and stable.
Let's go make friends

Recommended Users

Updating the list only if clusters have been updated or new users have been registered

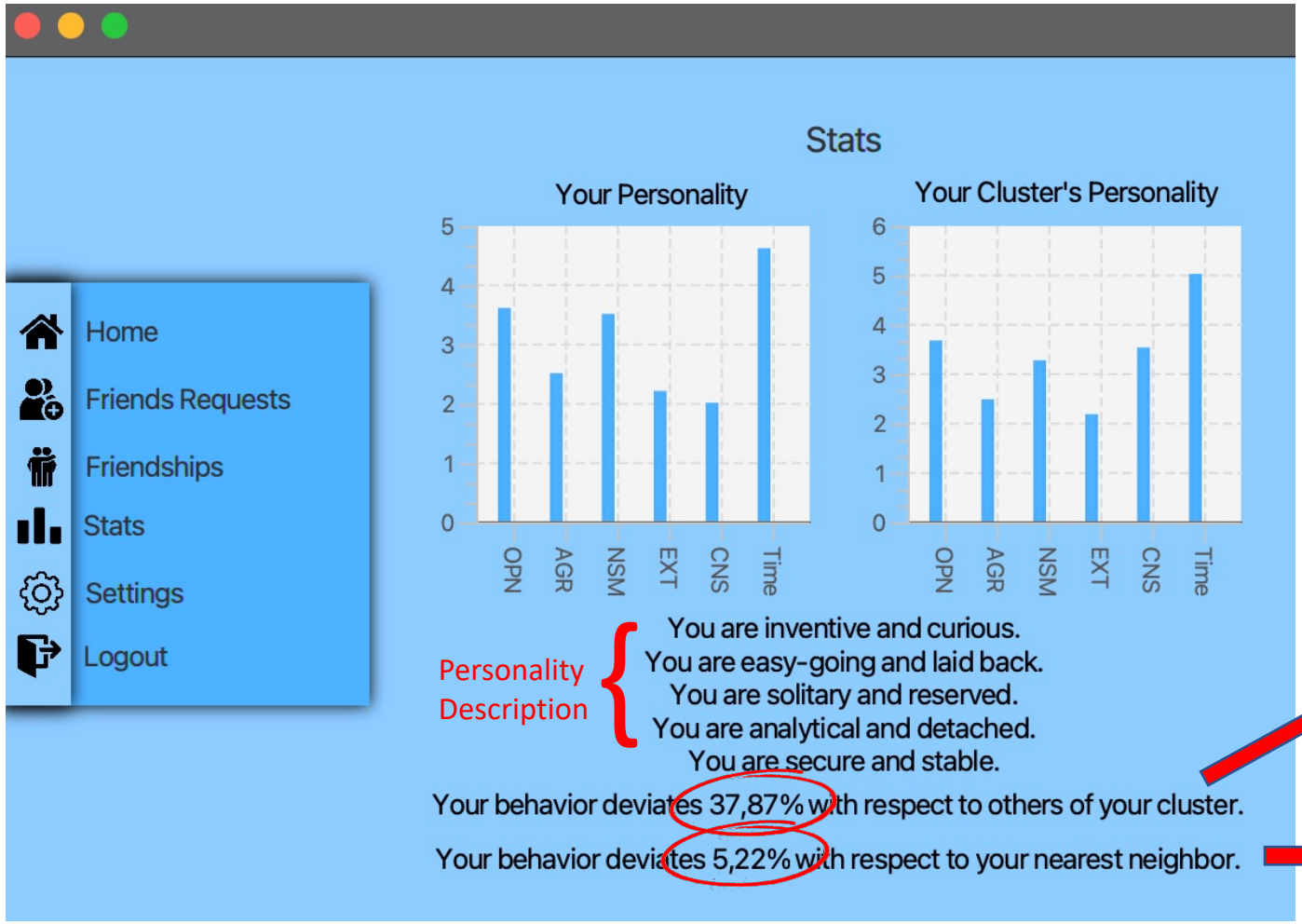




Recommended users are shown in a decreasing order of weight.

$$w_i = \frac{1}{dist^2(x,y)}$$

Statistics



Deviation from the cluster:

$$\sqrt{\sum_i (personality_i - AvgClusterPersonalityi)^2} * \frac{100}{6}$$

Deviation from the NN:

$$\sqrt{\sum_i (personality_i - personalityNNi)^2} * \frac{100}{6}$$

MongoDB Data Model

We used MongoDB to store complete information about users, including the answers to the survey.

Population of the database

EXT1	……	OPN10_E
2	…..	6448
4	…..	2331
….	….	2564
5	…..	4623

For each instance a user was generated from https://randomuser.me

The User's document structure saved in MongoDB is the following:


```
{
    "_id": {
        "$oid": "61df079e17e985790c6d7240"
    },
    "first_name": "Lærke",
    "last_name": "Thomsen",
    "date_of_birth": {
        "$date": "1965-11-05T23:00:00.000Z"
    },
    "gender": "female",
    "country": "United Kingdom",
    "username": "bigostrich858",
    "phone": "88468370",
    "email": "laerke.thomsen0@example.com",
    "password": "tiW45%E^b%",
    "registration_date": {
        "$date": "2005-10-23T22:00:00.000Z"
    },
    "picture": "https://randomuser.me/api/portraits/women/68.jpg",
    "survey": {
        "EXT": [{
            "name": "EXT1",
            "value": 1,
            "time": 2.292
        }, {…}
        }, {…}
        }, {…}
        }, {…}
        }, {…}
        }, {…}
        }, {…}
        }, {…}
        }, {…}],
        "EST": […],
        "AGR": […],
        "CSN": […],
        "OPN": […]
    }
}
```

Neo4j Data Model

We used Neo4j to store information about friendships and similarity among users. In order to maintain a connection between the same user in both databases a field «mongoId» containing the MongoDB user ObjectId was added to Neo4j.

Node Properties

User

<id>	17
country	United States
mongoId	61df079e17e985790c6d7251
name	Afşar Yeşilkaya
picture	https://randomuser.me/api/portraits/women/62.jpg

Relationship Properties

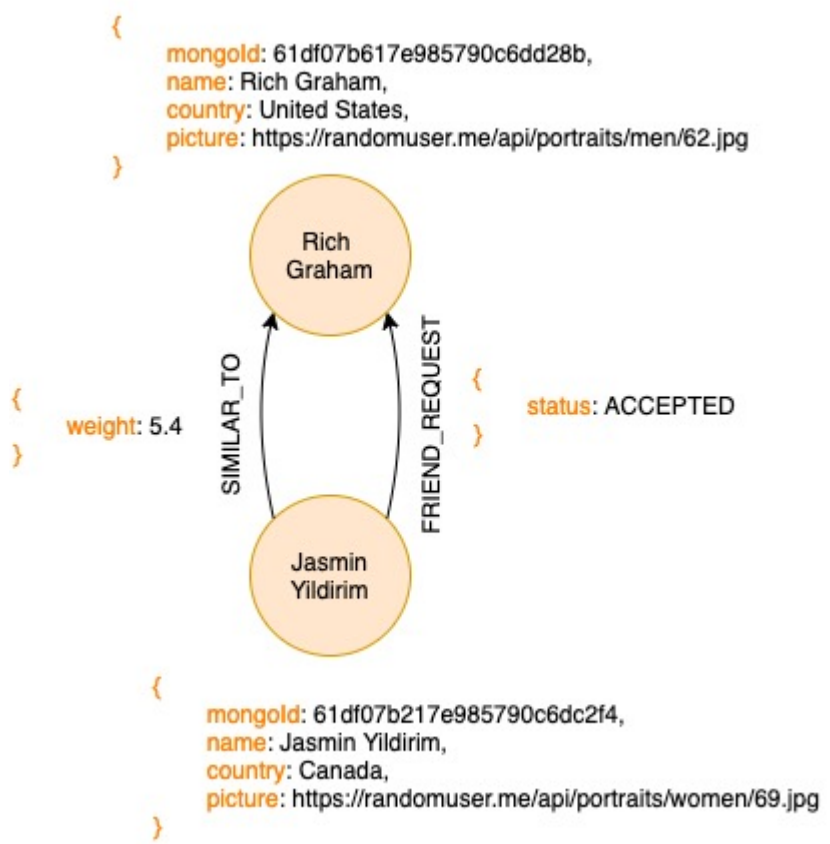
SIMILAR_TO

<id>	63
weight	4.55

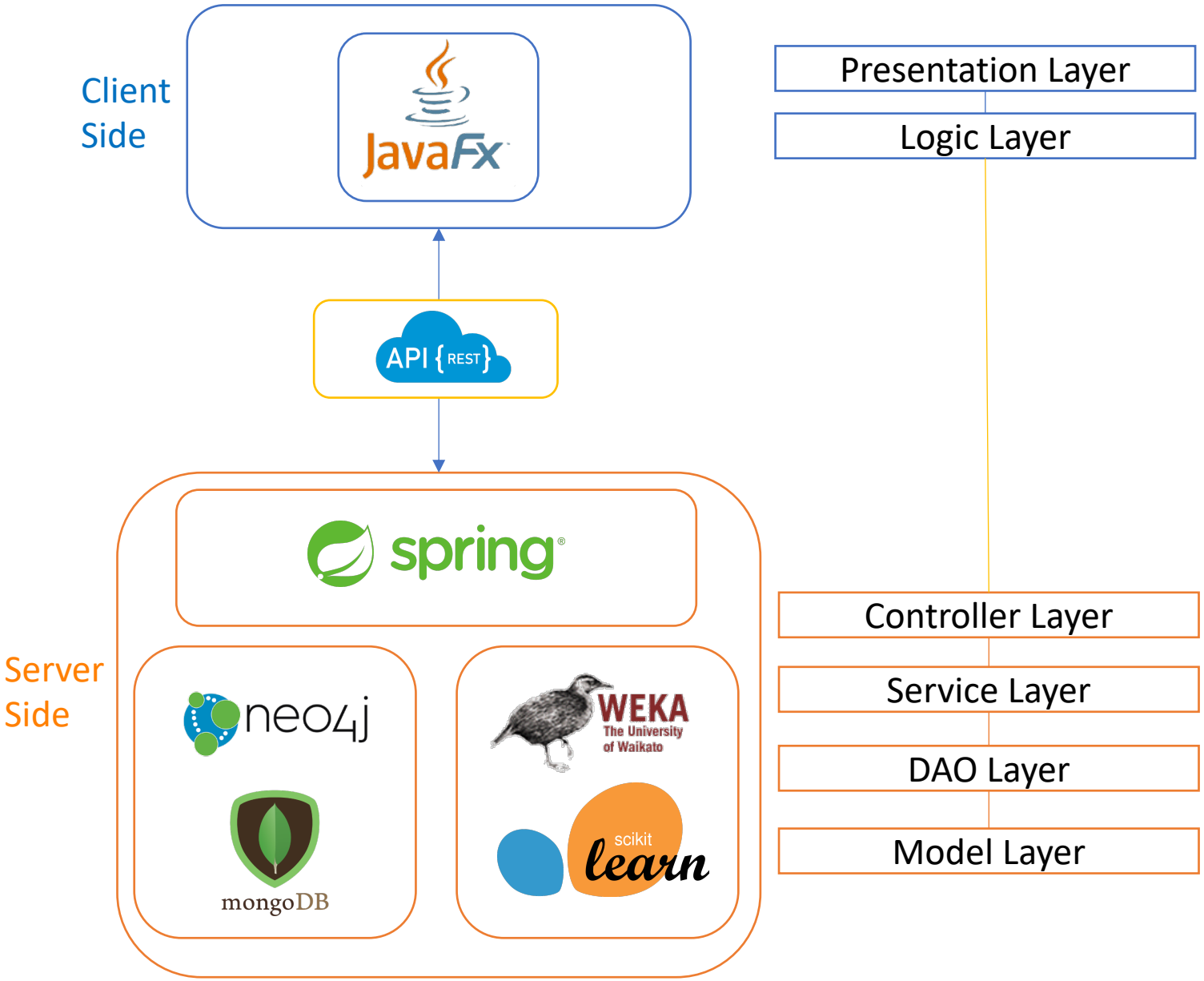
Relationship Properties

FRIEND_REQUEST

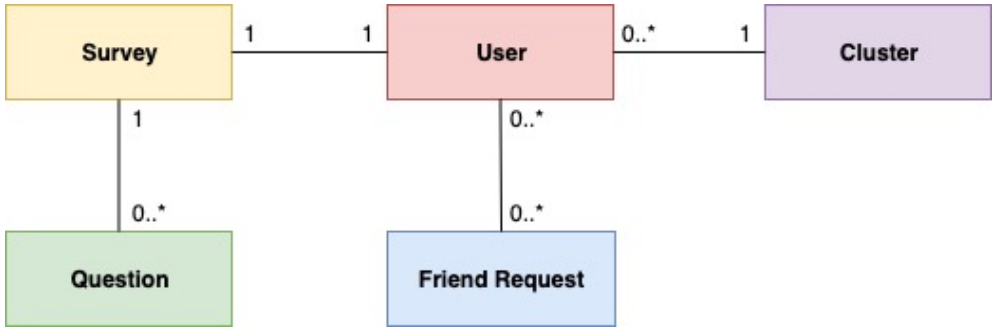
<id>	93
status	ACCEPTED



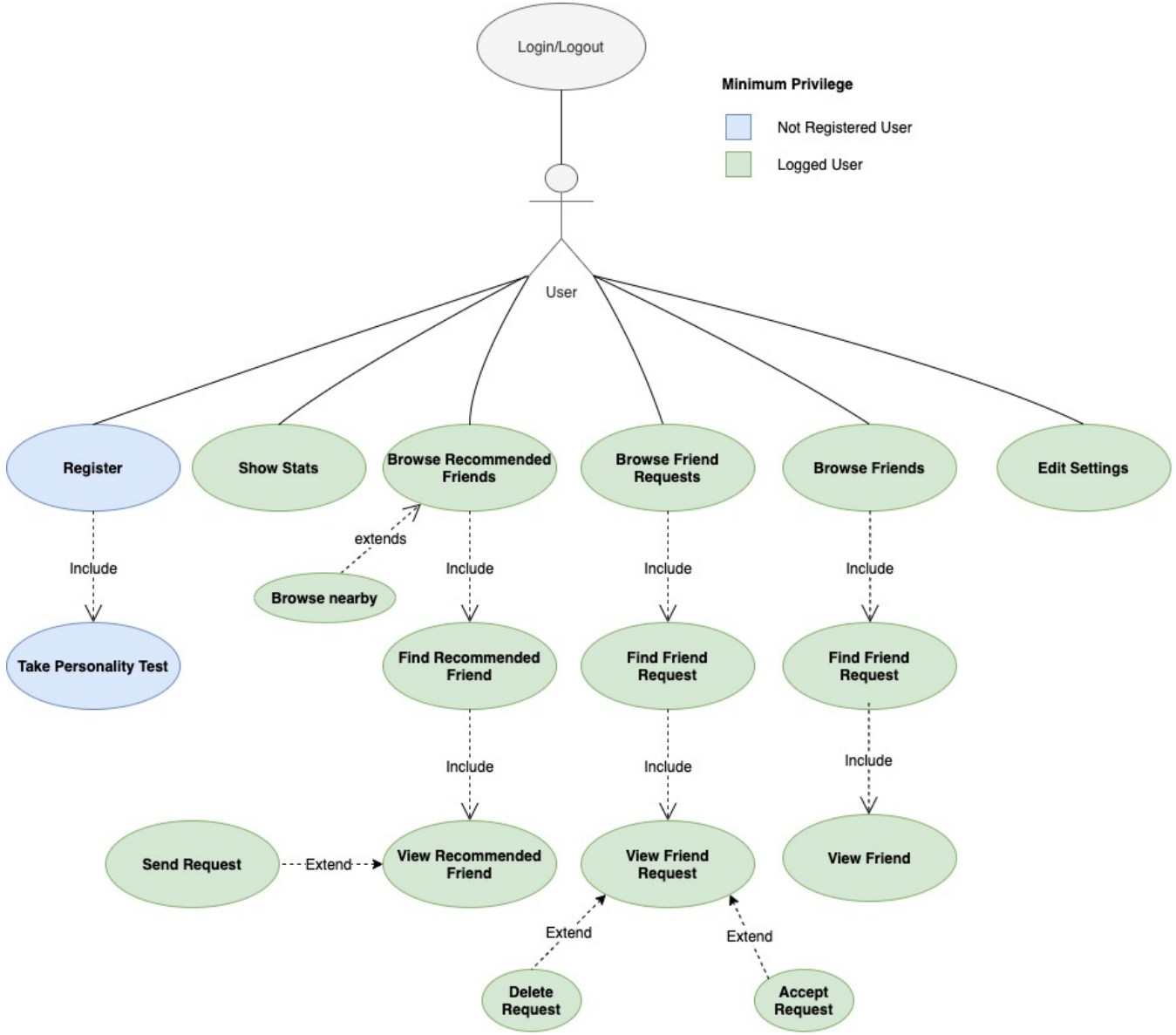
System Architecture



Analysis Classes



Use Case Diagram



Thanks for the attention!