PROYECTO 2

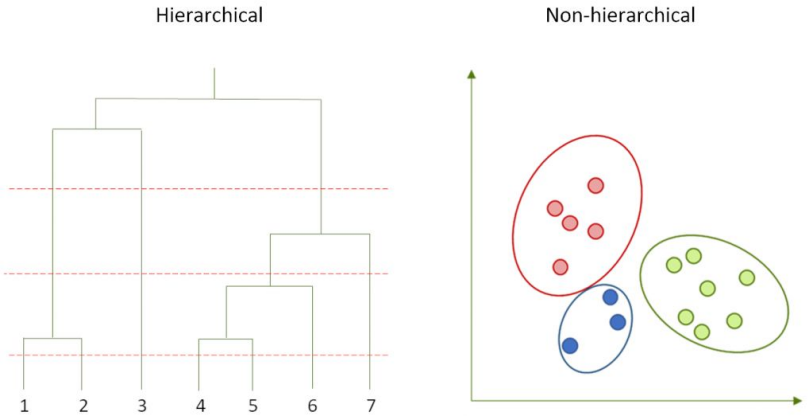
Aprendizaje no supervisado sobre perfiles de una red de contactos

Icia Carro Barallobre; Karen Salazar Gutiérrez; Laura Llorente Sanz

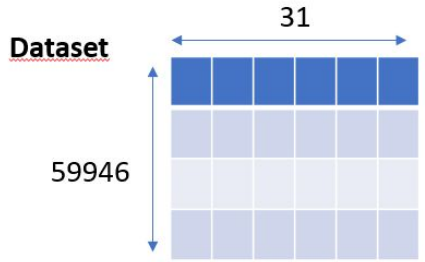
Índice

- Preprocesamiento
  - Introducción
  - NLP (LDA)
  - Decisiones

- Clustering
  - Reducción dimensional (PCA)
  - KMeans
  - DBSCAN
  - Agglomerative Clustering



Preprocesamiento: Introducción

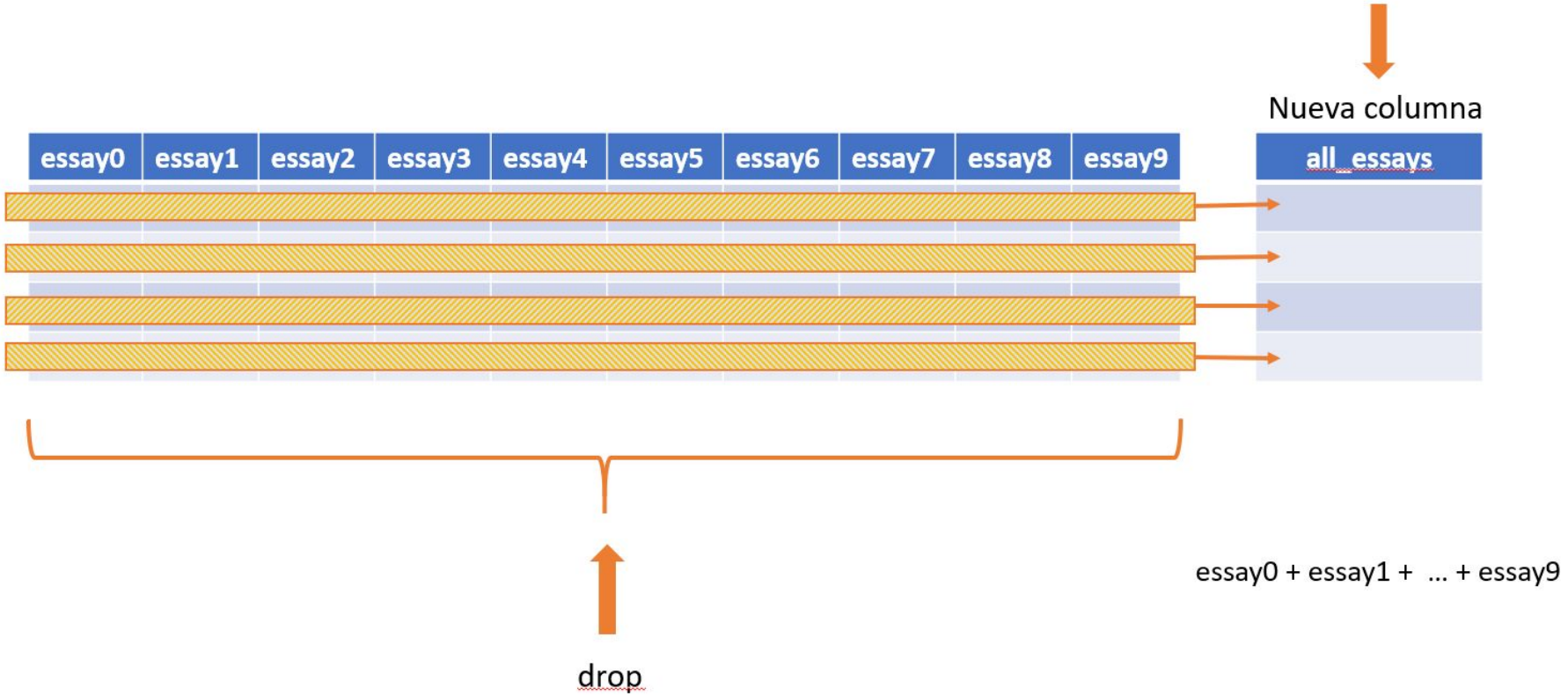


age	status	sex	orientation	Body-type	diet	...	offspring	pets	religion	sign
						...				
						...				
						...				
						...				

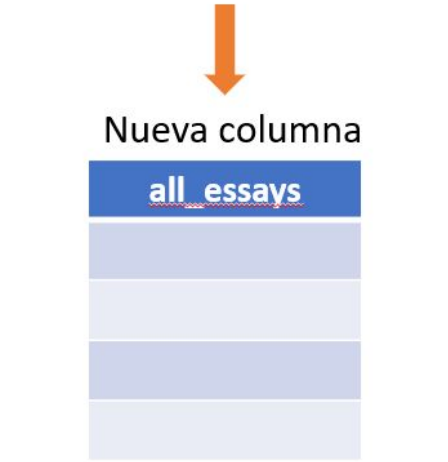
essay0	...	essay9
	...	
	...	
	...	
	...	

Metodología
1º Procesamiento del texto libre: PLN
2º Preprocesamiento
3º Clustering

PLN: Ingeniería de características



PLN: algoritmo LDA



`sklearn.feature_extraction.text`.CountVectorizer

`sklearn.decomposition`.LatentDirichletAllocation

Parameters: **n_components : *int, default=10***

Number of topics.

4 temas

PLN: algoritmo LDA

topic_matrix	n_top 1	n_top 2	n_top 3	n_top 4
n_topics 1	'friends'	'family'	'food'	'music'
n_topics 2	'read'	'reading'	'books'	'movies'
n_topics 3	'music'	'food'	'movies'	'books'
n_topics 4	'home'	'movie'	'friends'	'family'

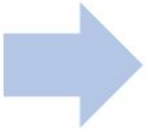
topic_names

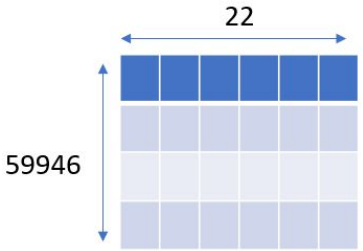
- 'Social'
- 'Book'
- 'Music'
- 'Movie'

Exploración de los datos

Nueva columna

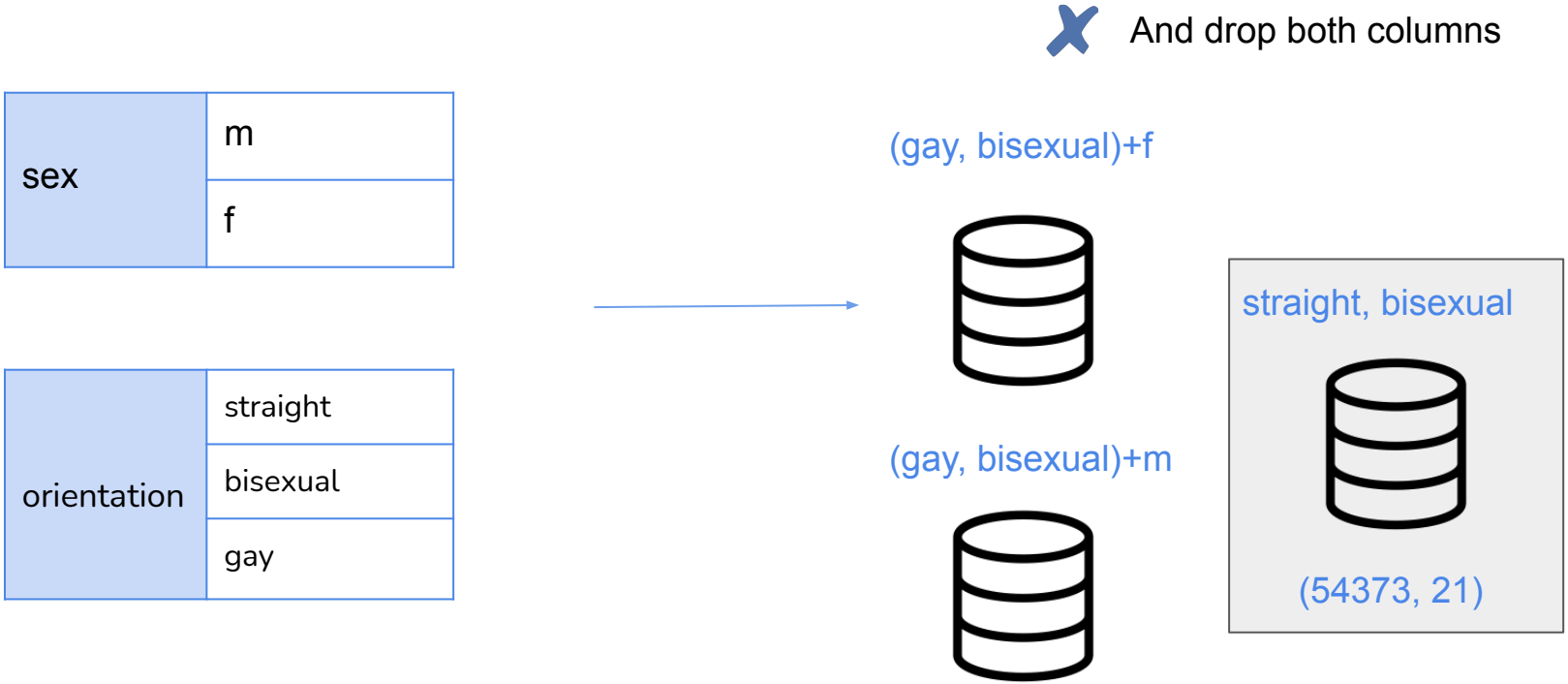
age	status	sex	orientation	...	smokes	speaks	labels
				...			'Social'
				...			'Music'
				...			'Reader'
				...			'Cuisine'





Dataset

Preprocesamiento: Decisiones (I)



Preprocesamiento: Decisiones (II)

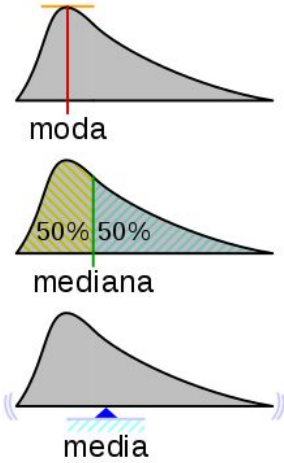
Evitar sesgos o información irrelevante:

Preprocesamiento: Decisiones (IV)

Tratamiento de nulos:

- Borrar aquellas columnas donde más del 40% de los datos sean nulos:
  - offspring

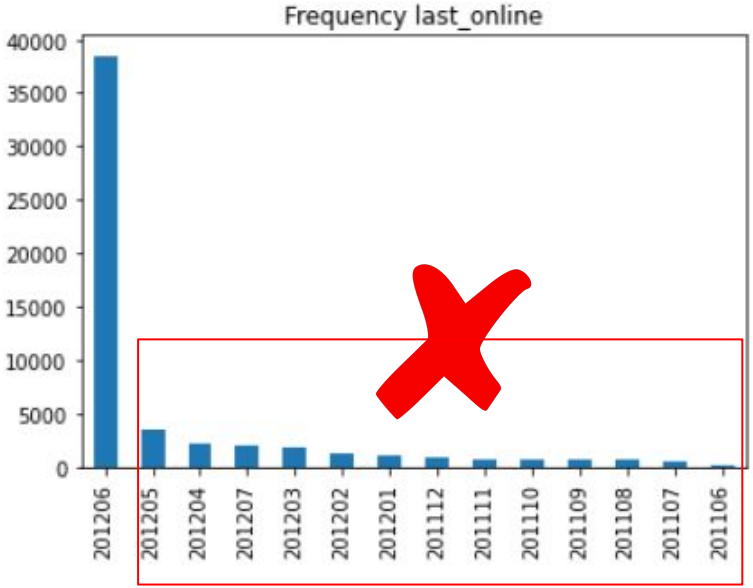
- Rellenar el resto con la moda

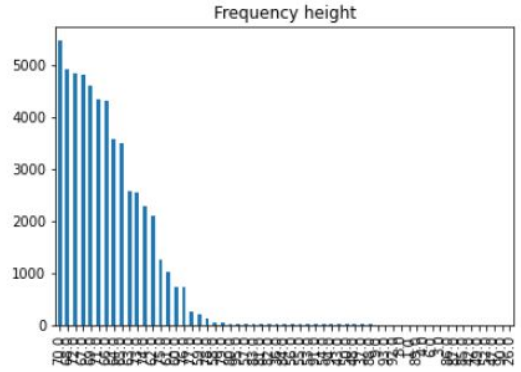


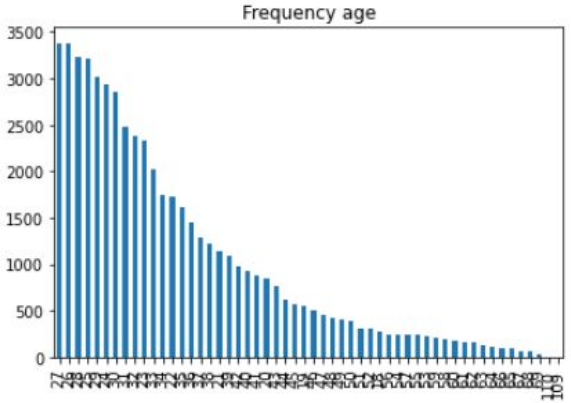
(40456, 18)

Preprocesamiento: Decisiones (V)

Exploración de las variables numéricas:



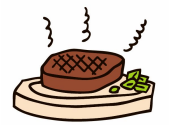




Preprocesamiento: Decisiones (VI)

Exploración de las variables categóricas (I):

Diet: No , Vegatarian, Vegan (0,1,2)



Pets: Tener o querer

- Cats: yes o no (0,1)
- Dogs: yes o no (0,1)

Jobs

- Retired
- Student
- Art
- Health
- Social
- Industry
- Other (Unemployed)

Religion: Una columna por cada religion
Puntuación en función de la seriedad [1, 0.5, 0]

Preprocesamiento: Decisiones (VI)

Exploración de las variables categóricas (II):

Location: A coordenadas, latitud y longitud

Speaks: Varias columnas:
- Number of languages
- Languages:
  - Una columna por cada uno
  - Puntuación: [1, 0.5, 0.25, 0]
  - Valores atípicos:
    - Tech: C++ o Lisp
    - Clasico: Latin, griego antiguo

Preprocesamiento: Decisiones (VI)

Exploración de las variables categóricas (III):

Drinks: no, sometimes, yes (0,1,2)

Drugs: yes o no (1,0)

Smokes: yes o no (1,0)

Offspring:

- Hijos: yes o no (1,0)
- Querer hijos: yes o no (1,0)

Preprocesamiento: Decisiones (VII)

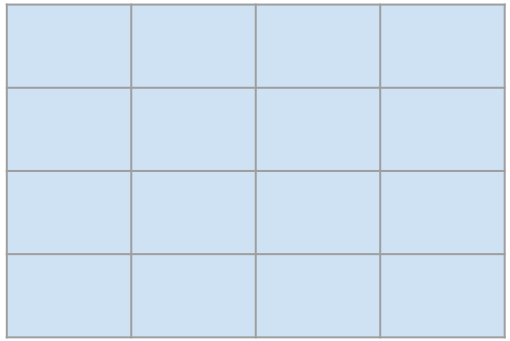
One Hot Encoder:

- Job
- Essays

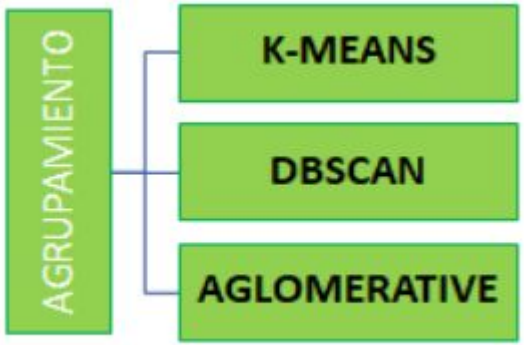
Normalización: Scale

108

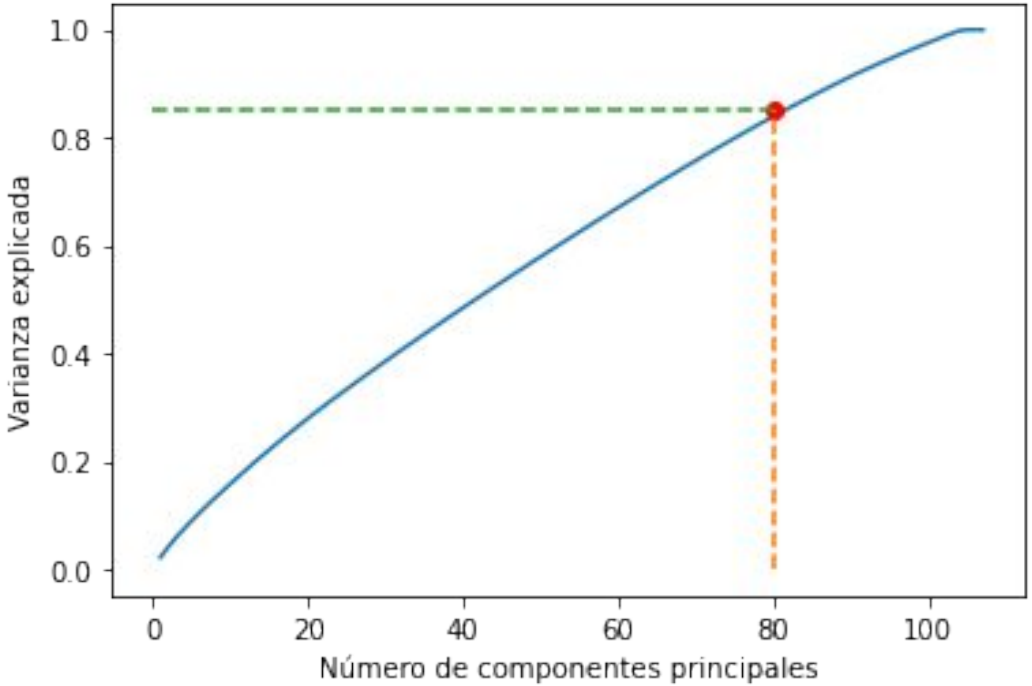
40425,



Clustering

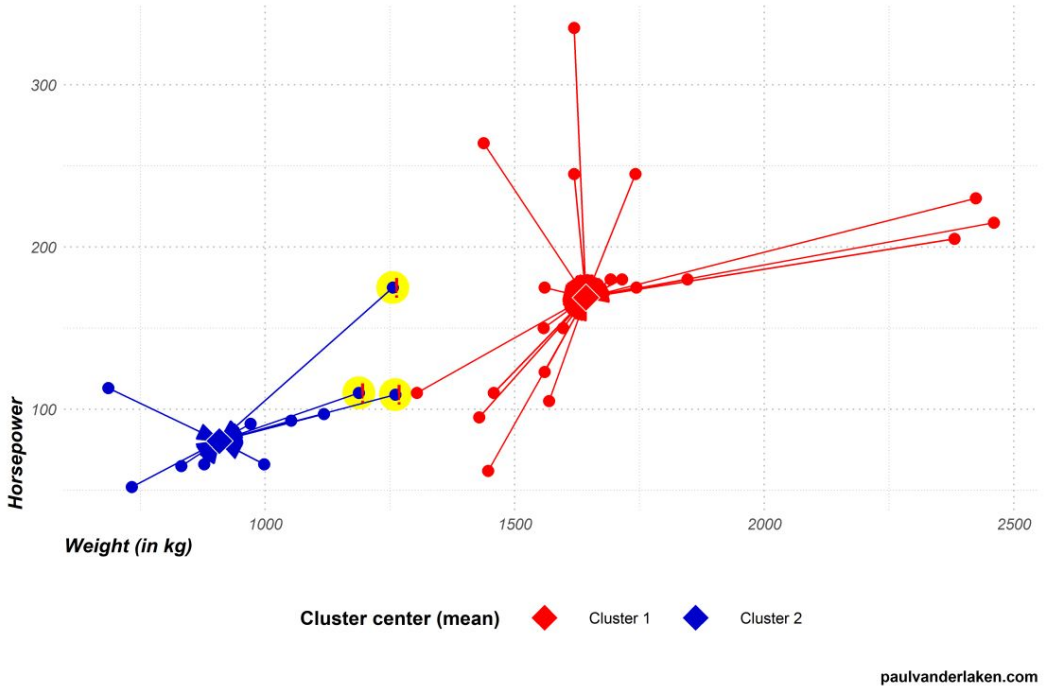


Clustering: Reducción dimensional (PCA)



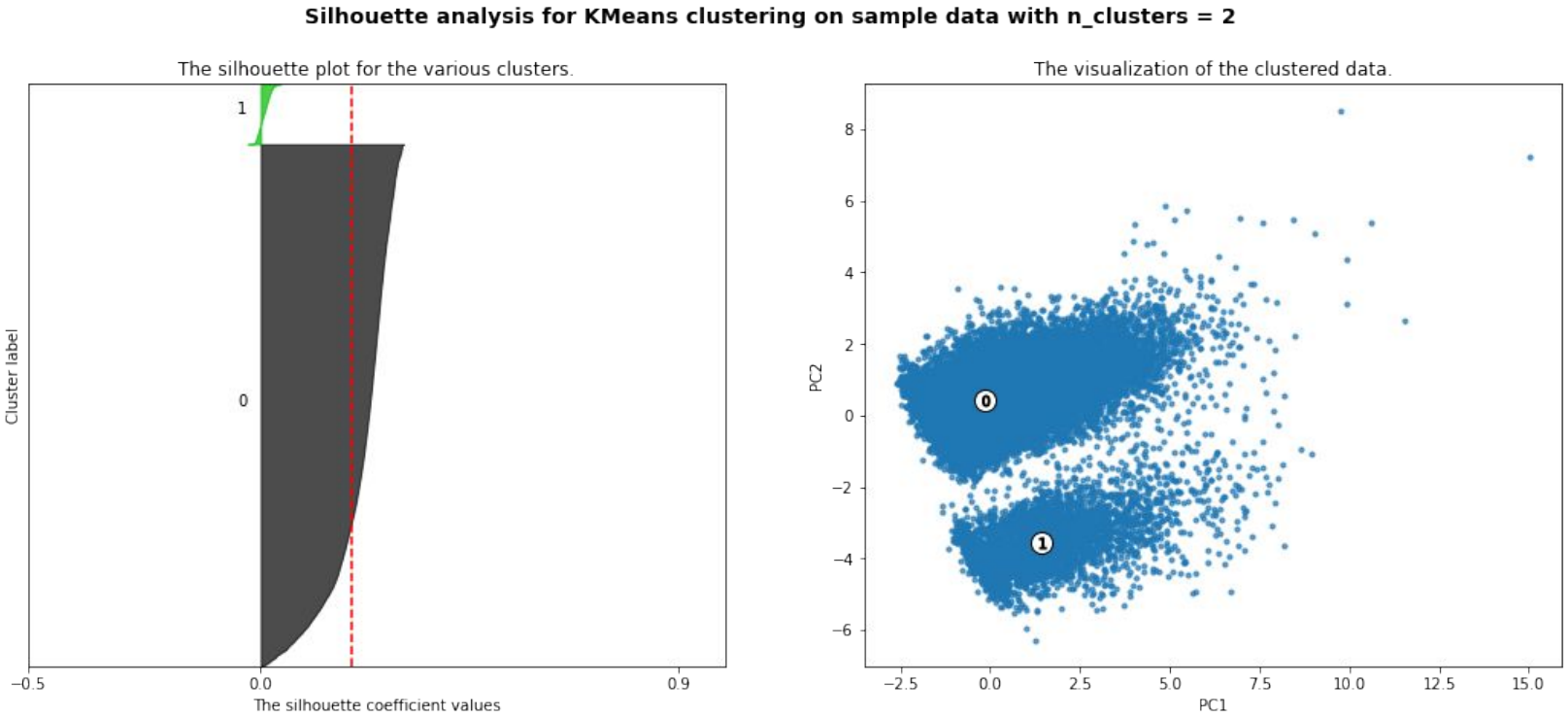
85%

Clustering: KMEANS



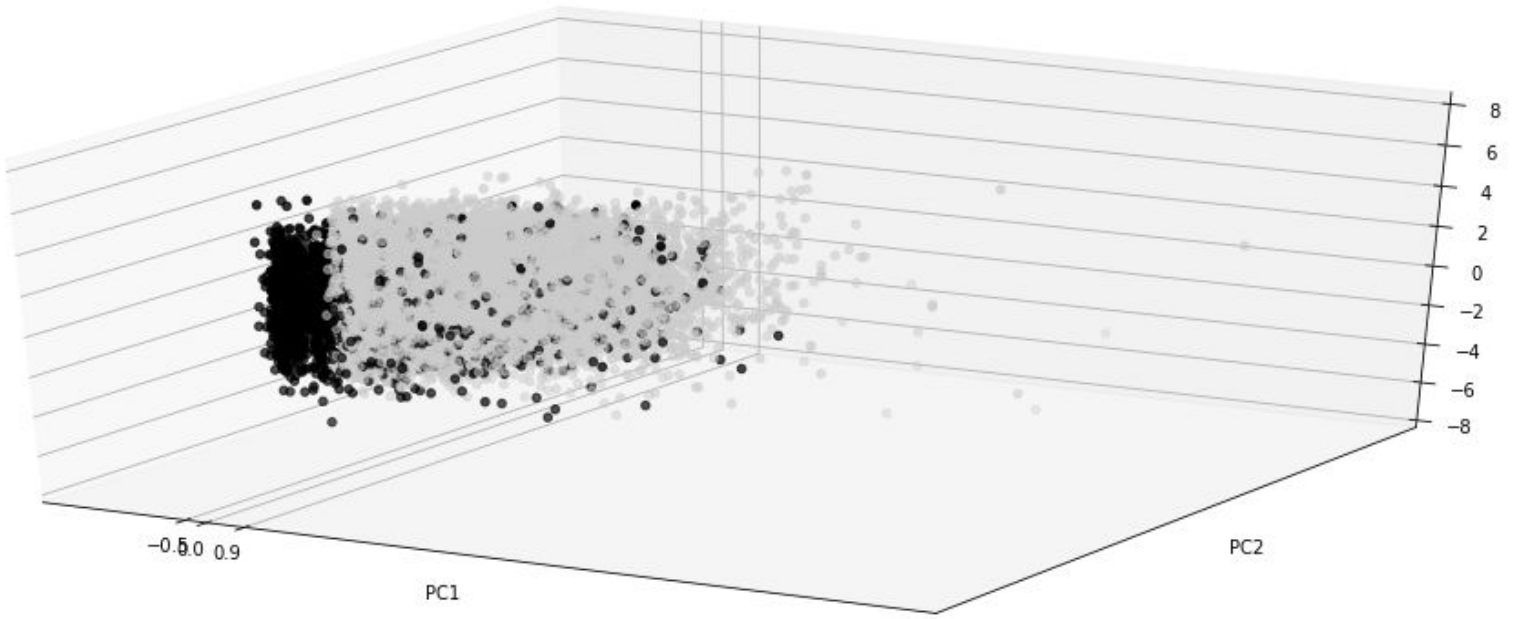
Clusters: 2

Silhouette Score: 0.20



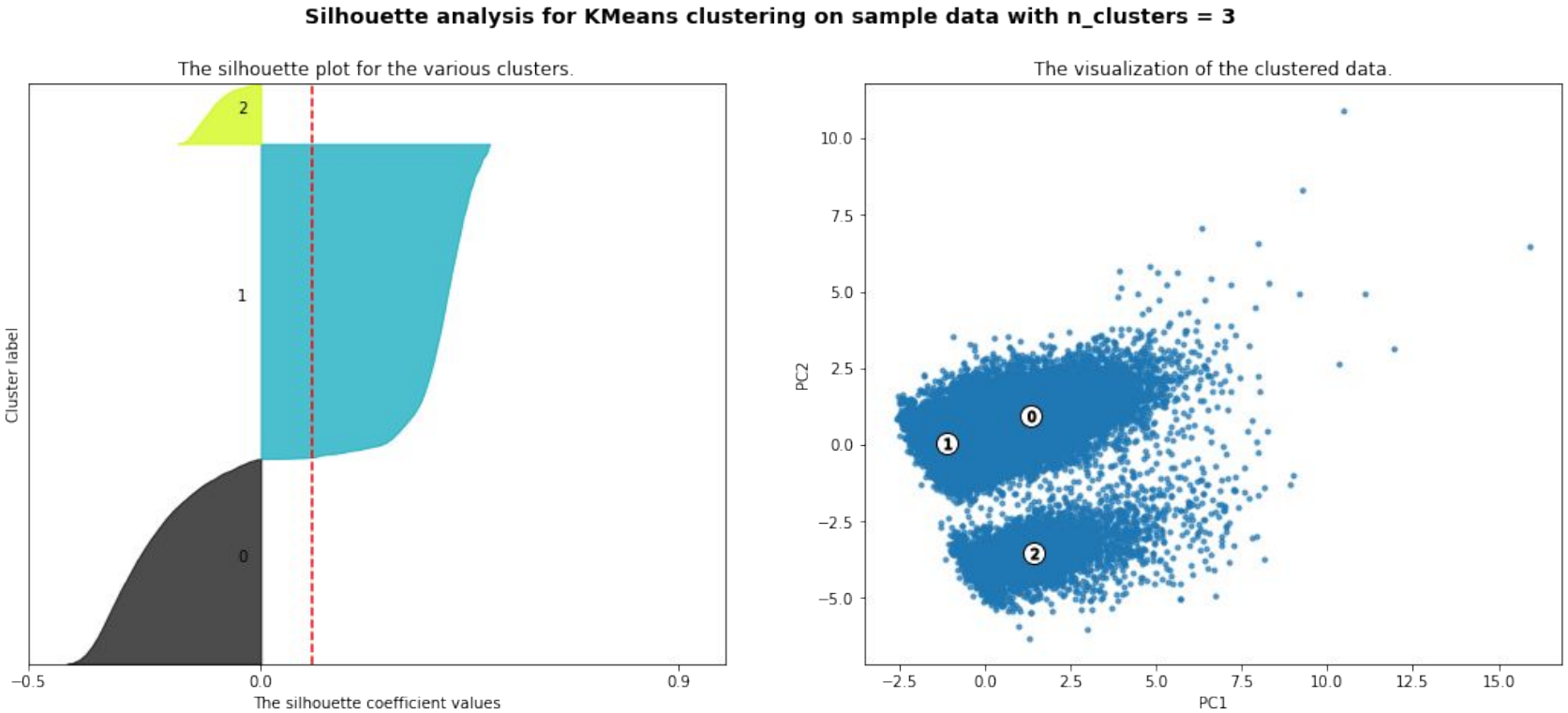
Clusters: 2

Silhouette Score: 0.20



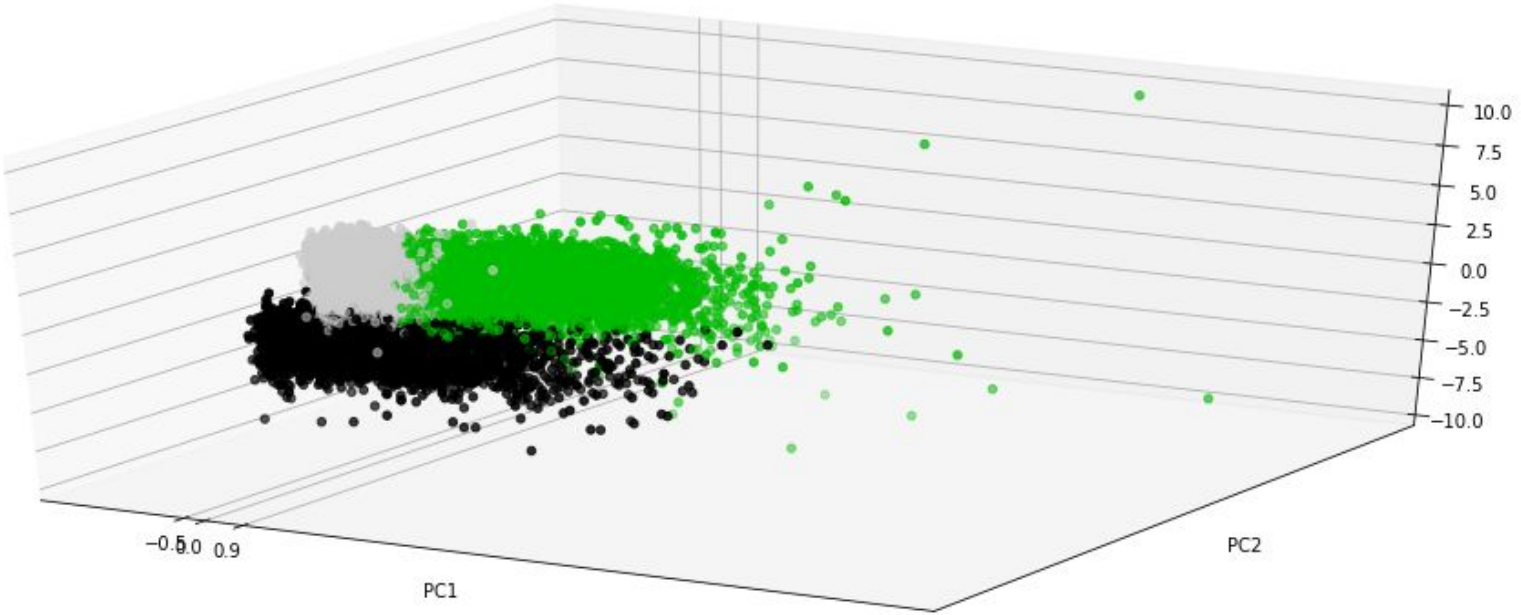
Clusters: 3

Silhouette Score: 0.12

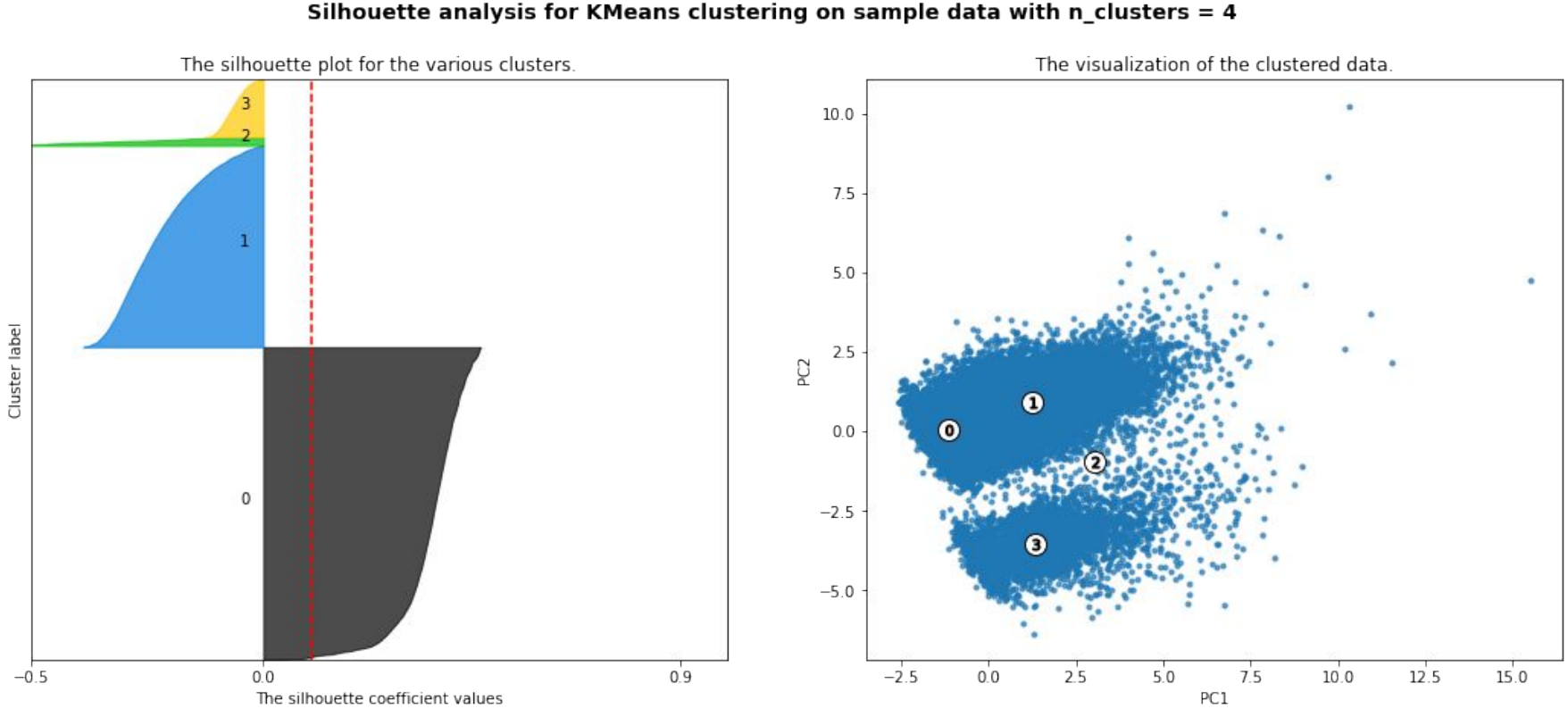


Clusters: 3

Silhouette Score: 0.12

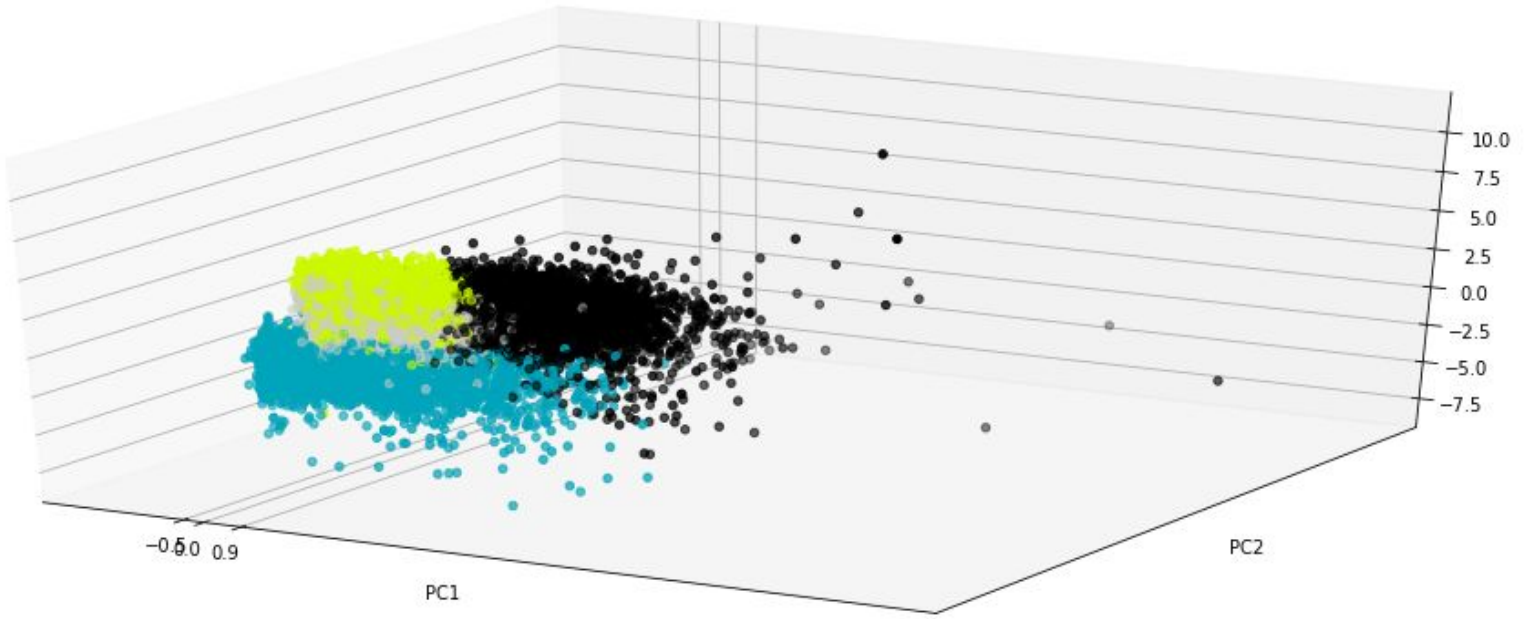


Clusters: 4 Silhouette Score: 0.10

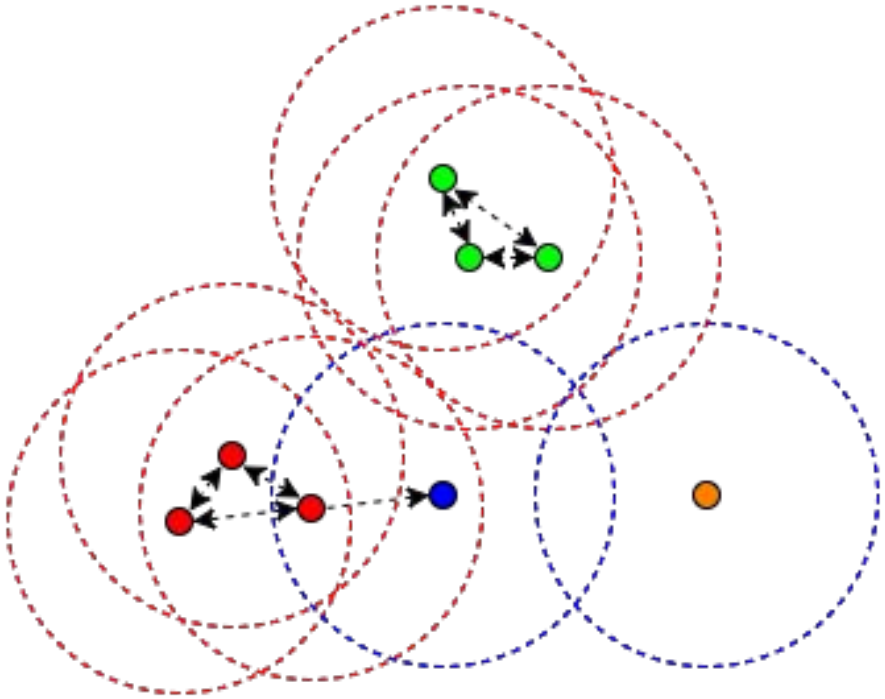


Clusters: 4

Silhouette Score: 0.10

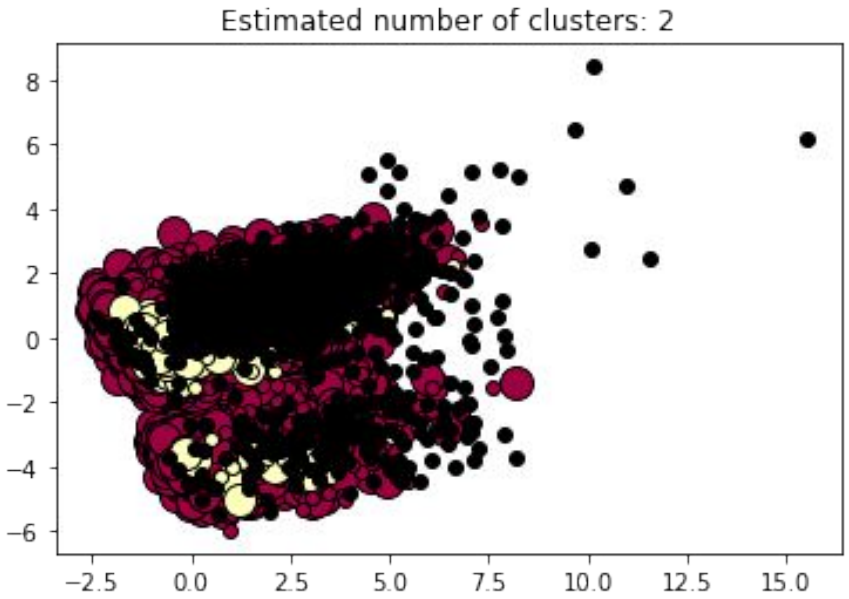


Clustering: DBSCAN

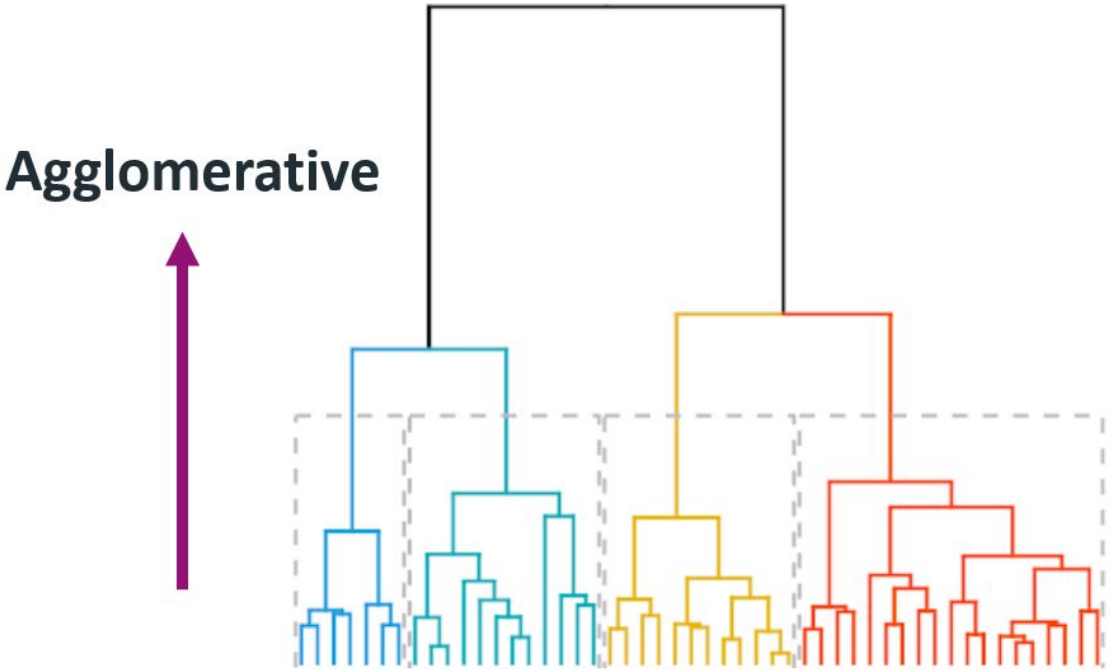


Clustering: DBSCAN

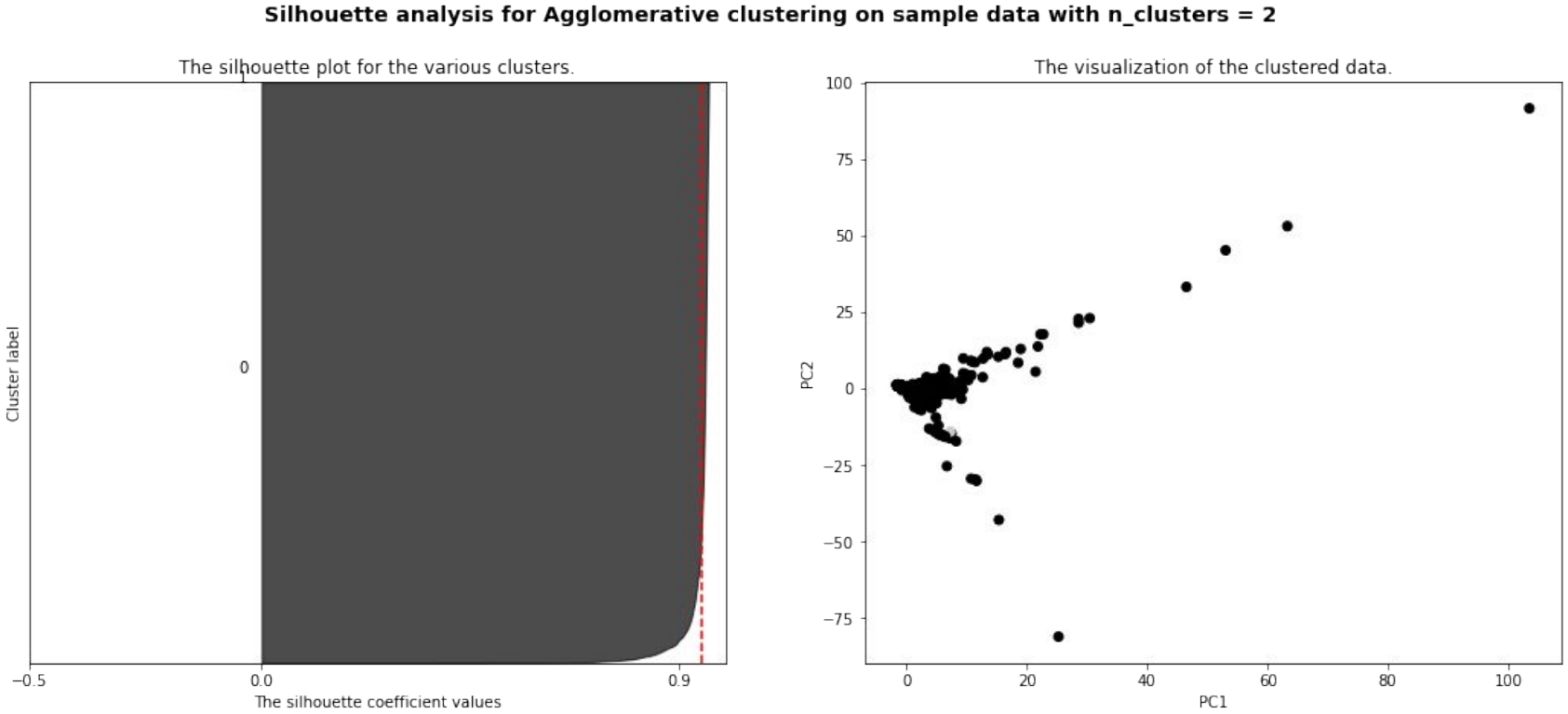
```
db = DBSCAN(eps=10, min_samples=100).fit(X)
-----------------------------
Estimated number of clusters: 2
Estimated number of noise points: 985 (BLACK)
Silhouette Coefficient: 0.389
```



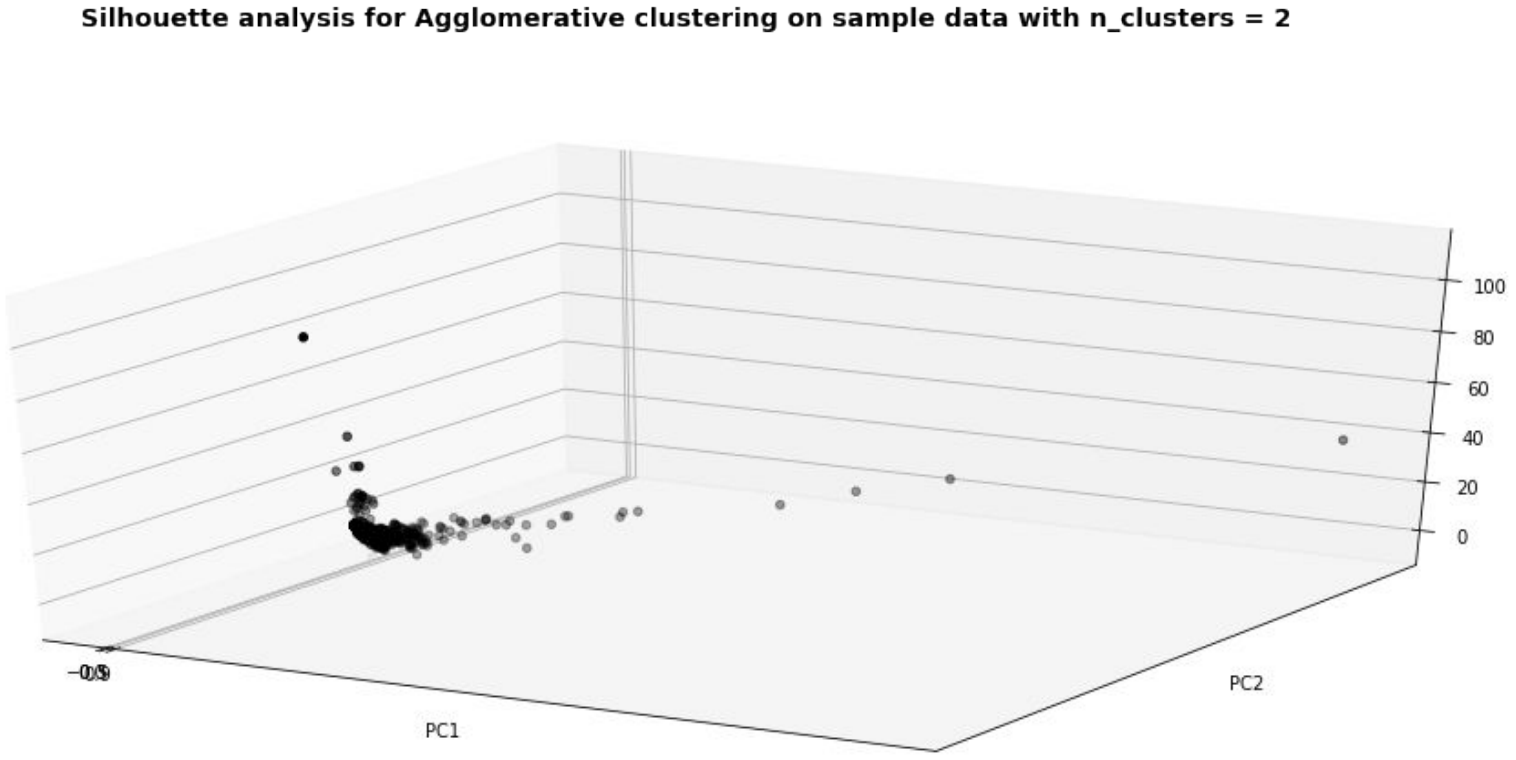
Clustering: Aglomerative (I)



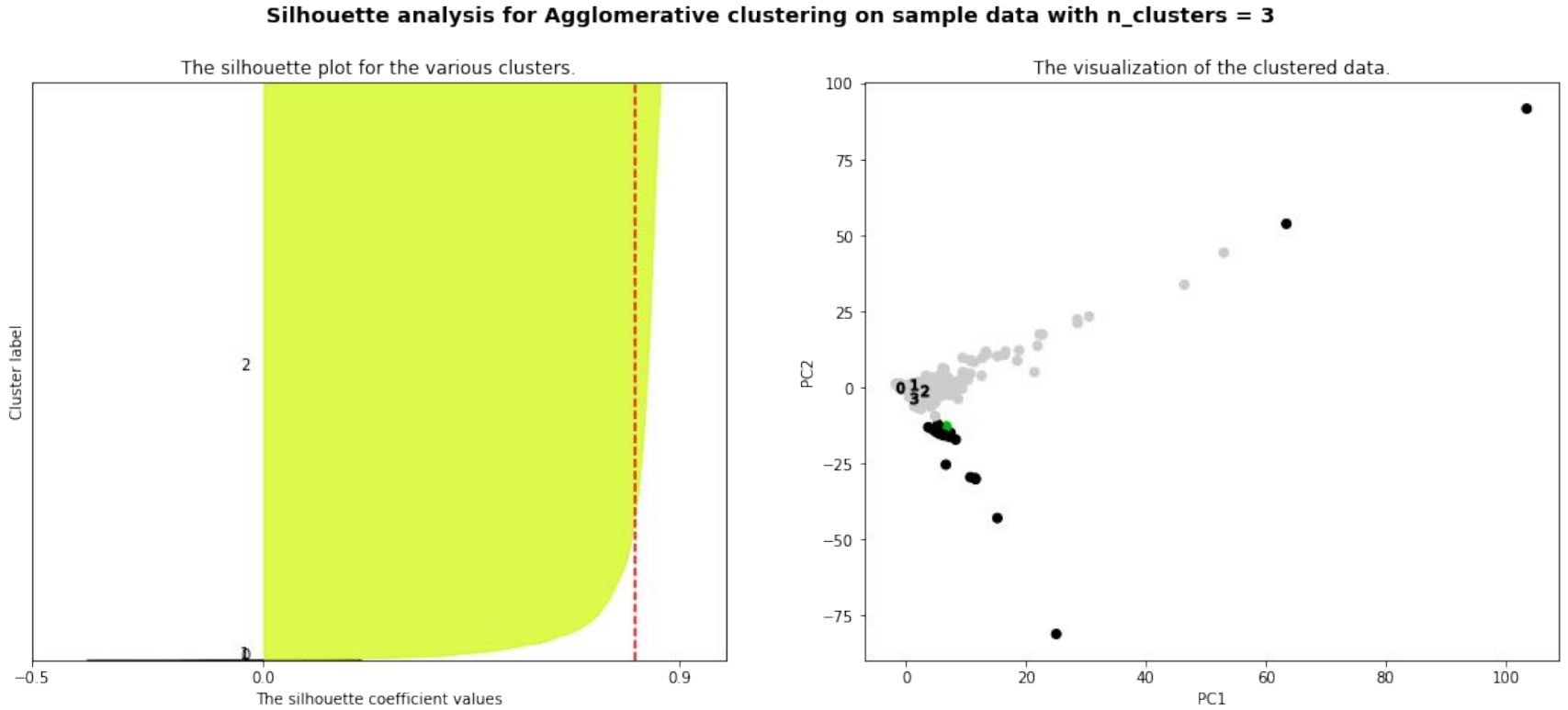
Clustering: Aglomerative (I)



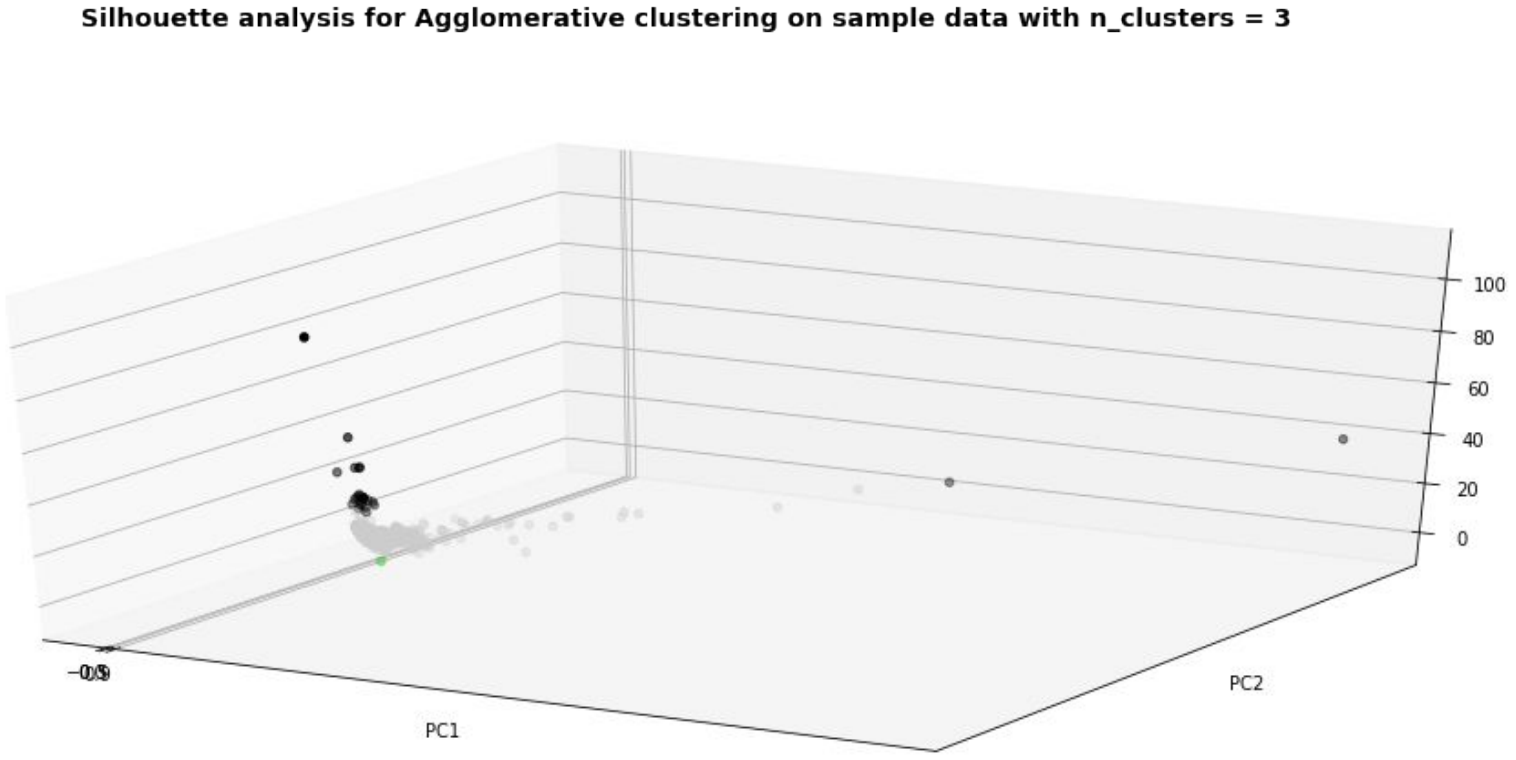
Clustering: Aglomerative (I)



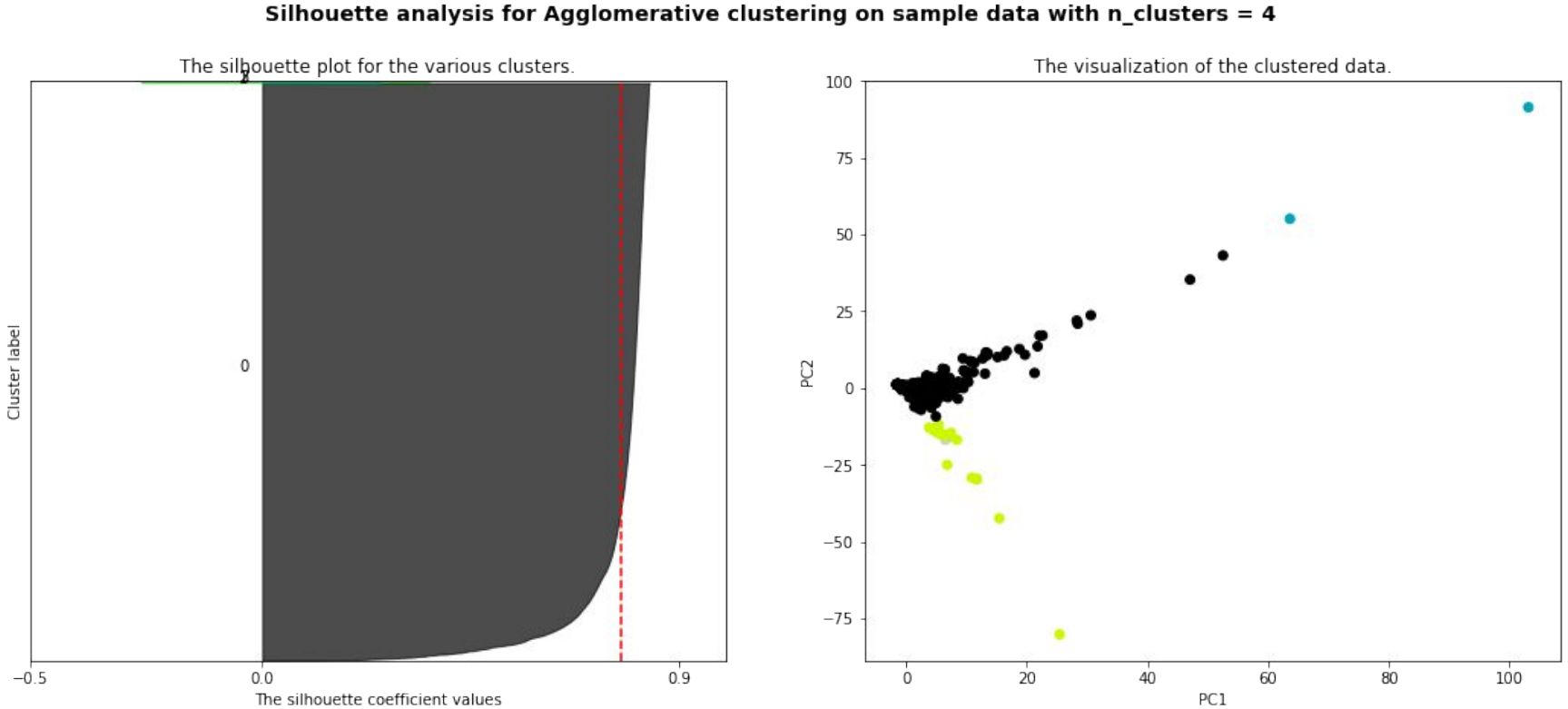
Clustering: Aglomerative (II)



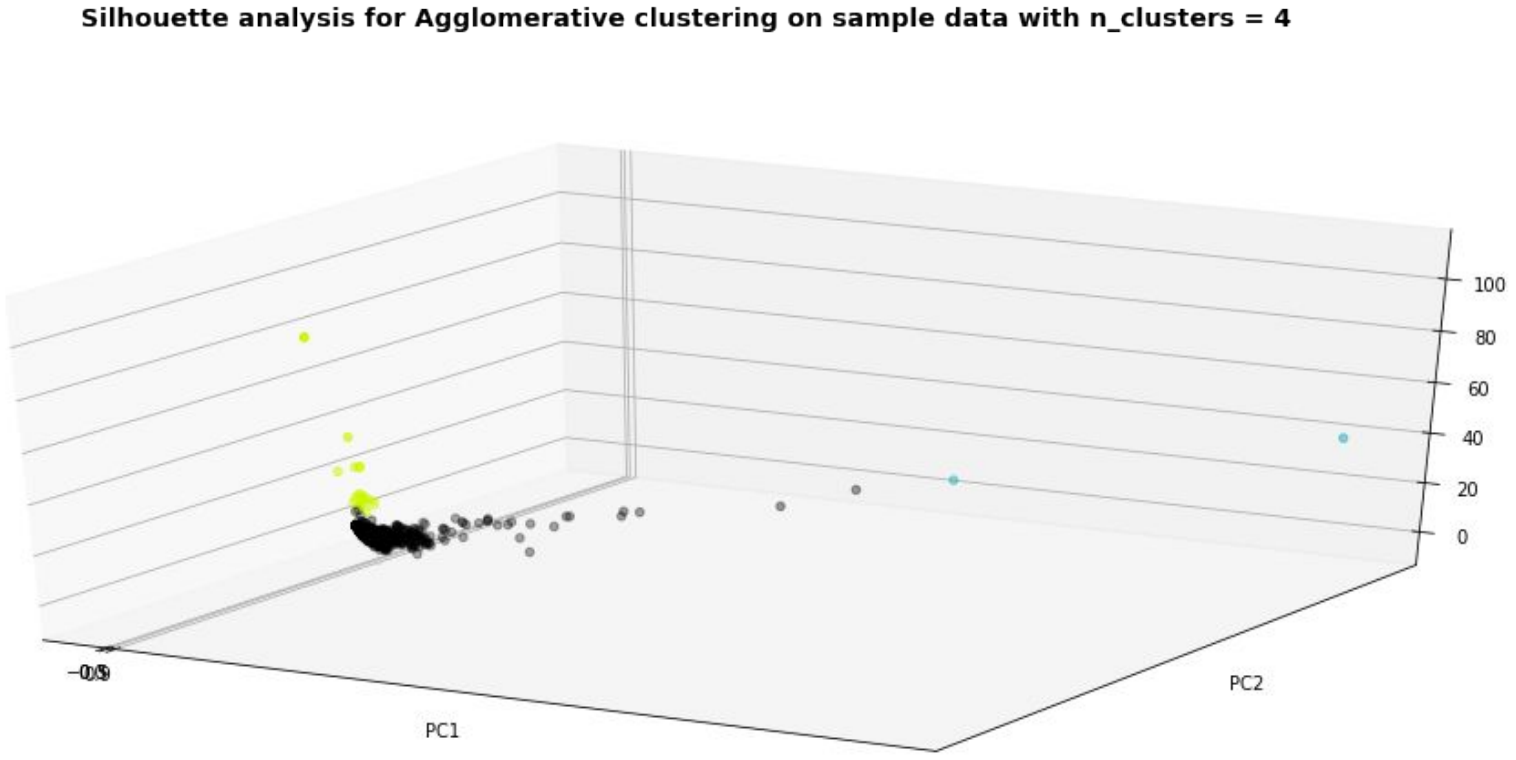
Clustering: Aglomerative (II)



Clustering: Aglomerative (II)



Clustering: Aglomerative (II)



¿Alguna pregunta?