Index

1 Description 1
2 Clustering 1
2.1 Dimensionality reduction 1
2.2 Elbow Method 4
2.3 Kmeans clustering 6
2.4 DBscan clustering 7
2.5 Difference between DBscan and Kmeans 7
3 Descriptive Analysis 8

1 Description

Analysis of users of various social media.

In the first part (using Python, sklearn and Google Colab), classes of users with similar habits and characteristics were identified using clustering algorithms.

In the second part with tableau a descriptive analysis was carried out.

2 Clustering

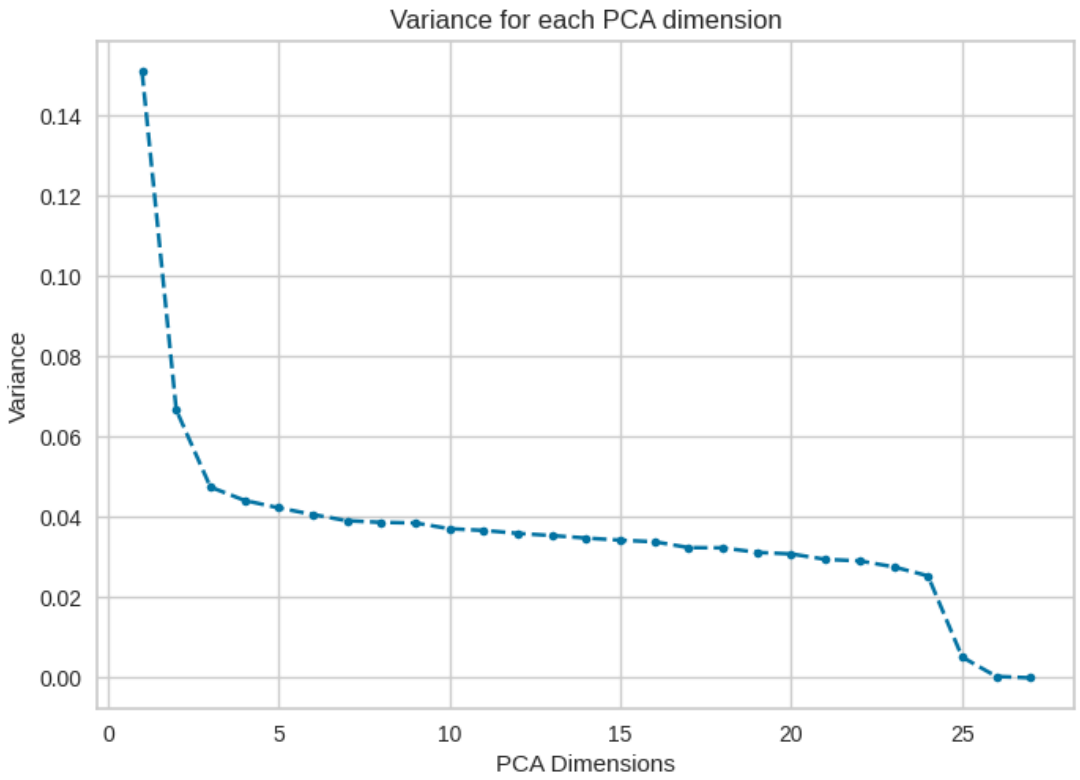
After studying the dataset with the head(), info(), and describe () commands, we moved on to preparing the data for the application of clustering algorithms.

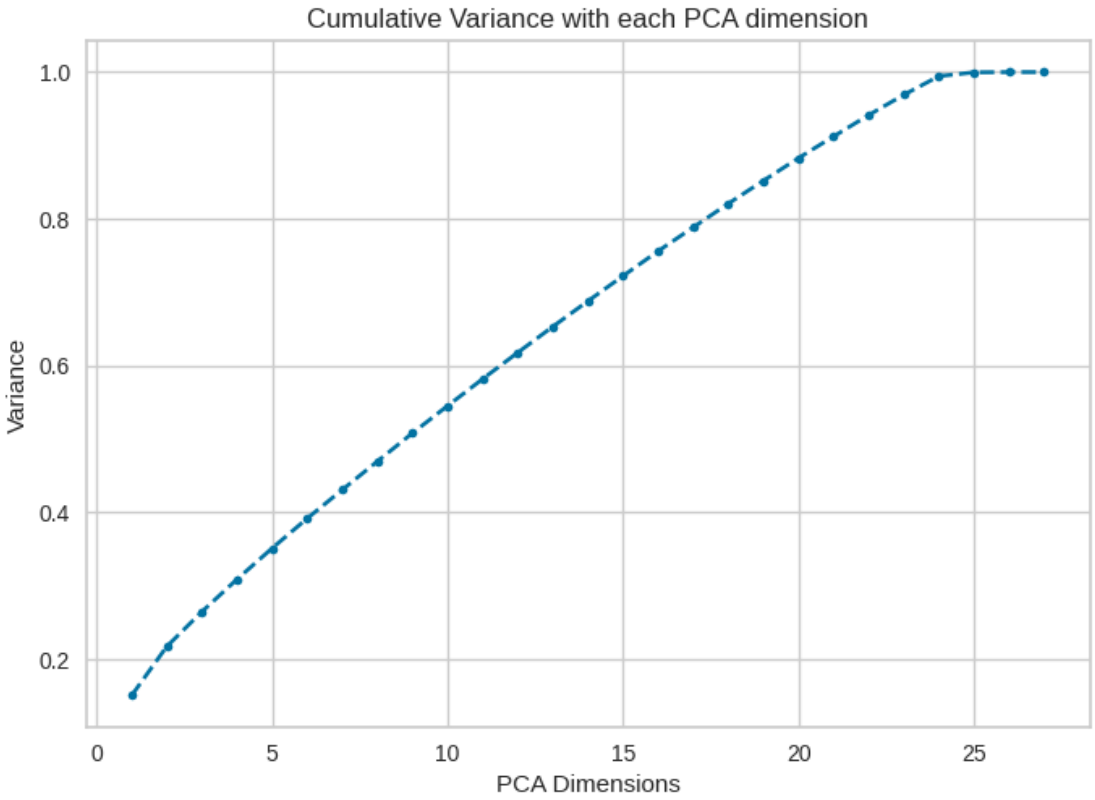
After briefly seeing the correlation between the features, the encoding of the categorical features and the scaling of the data were carried out. This was necessary to increase the performance of the algorithms applied subsequently.

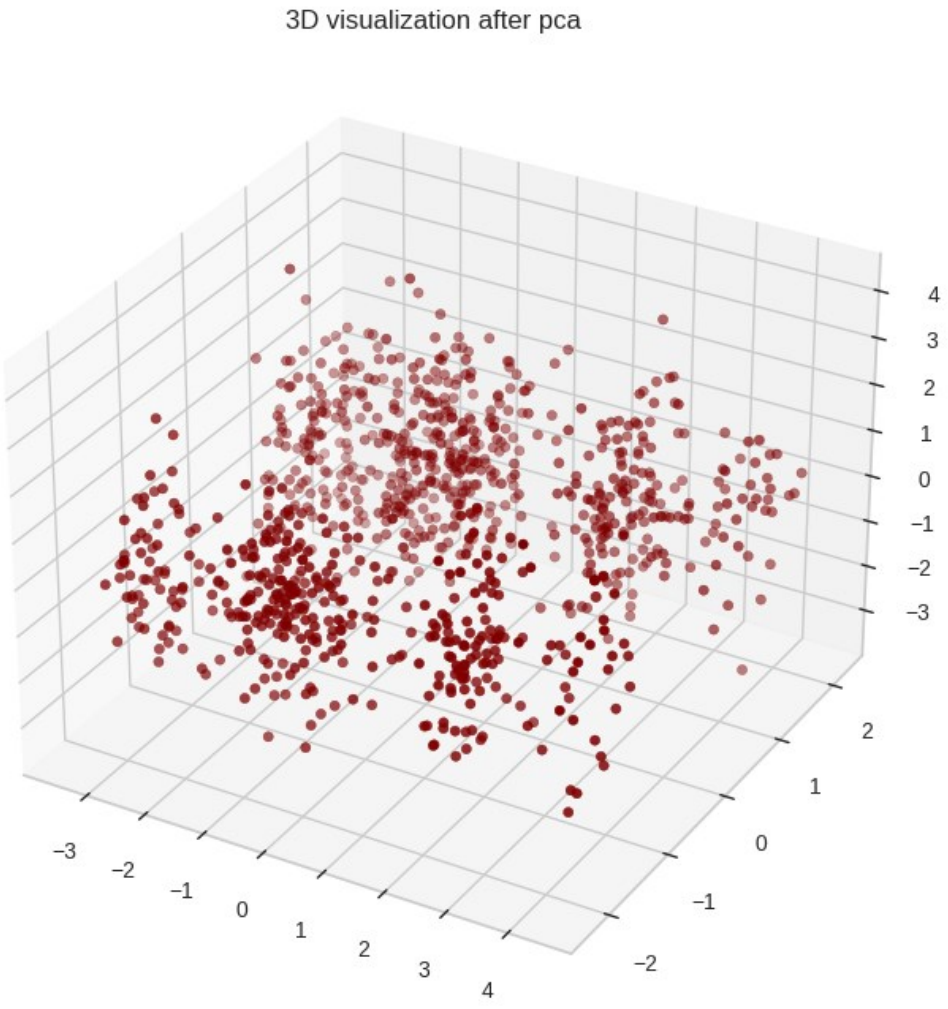
2.1 Dimensionality reduction

The dataset has many features, therefore to improve the training and performance of the algorithms it was necessary to reduce the dimensionality with the PCA method.

After seeing the results by evaluating the variance based on the number of features it was decided to proceed with the parameter n_components=3. Thus the model is simplified in the training and evaluation phase but the loss of information is minimal



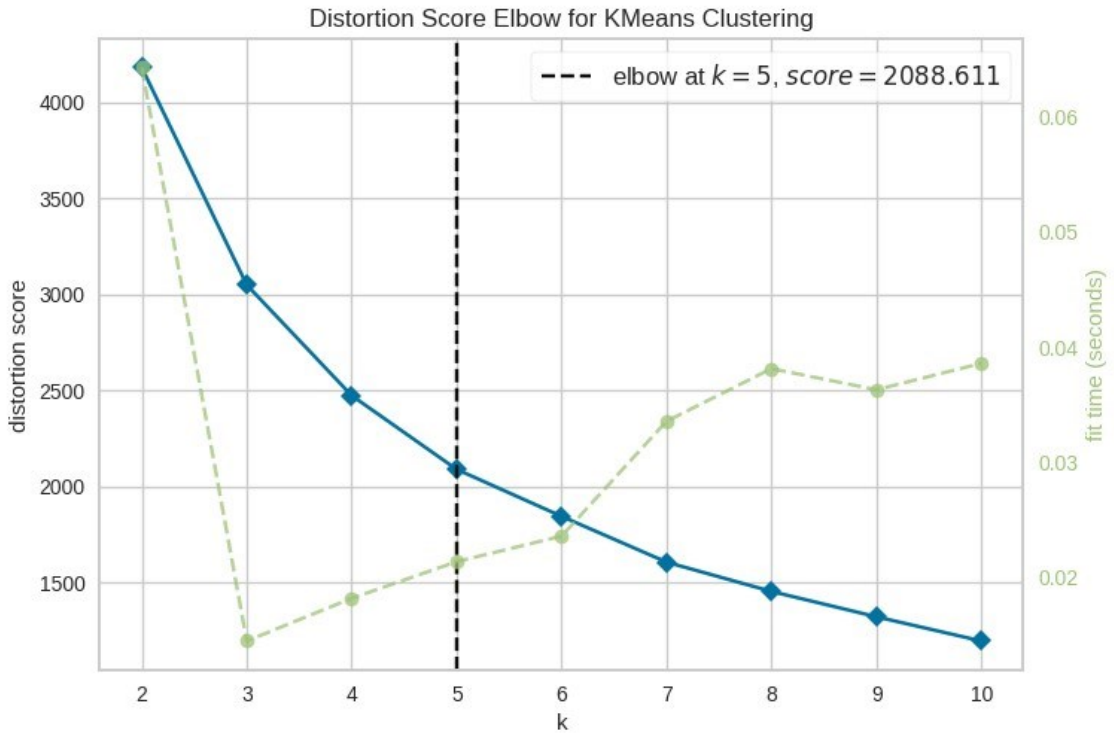

Cumulative Variance with each PCA dimension
Variance
1.0
0.8
0.6
0.4
0.2
0
5
10
15
20
25
PCA Dimensions

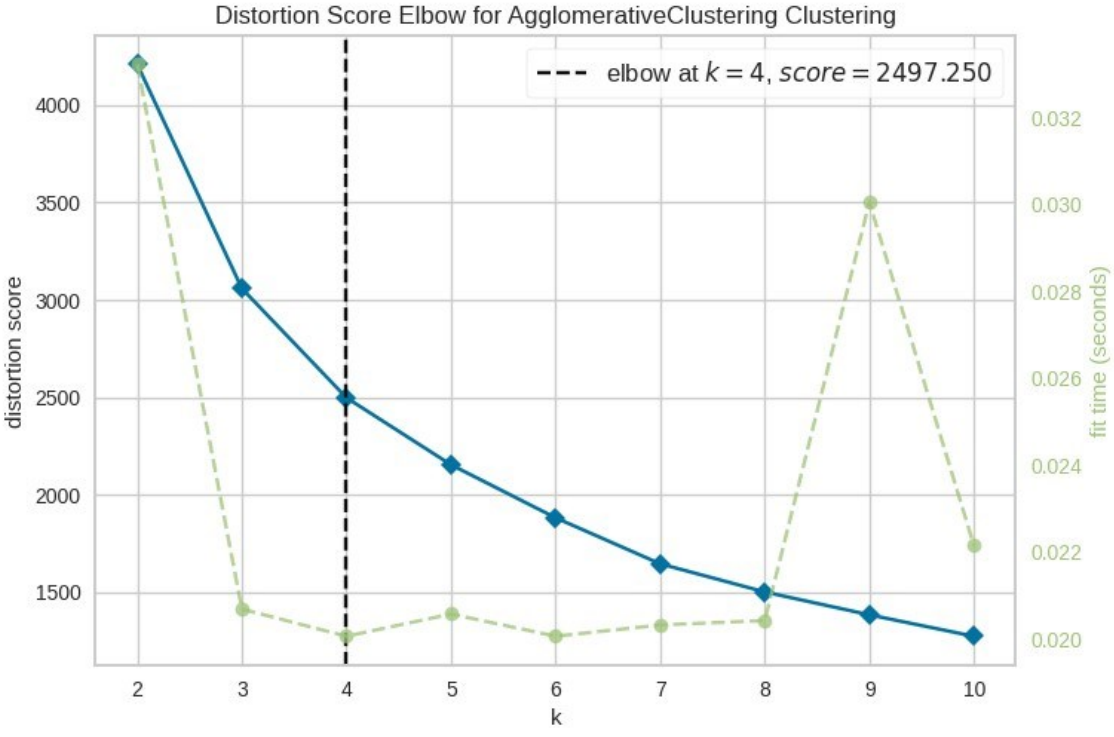


2.2 Elbow Method

For algorithms where the parameter k indicating the number of clusters must be provided, this method provides the value to obtain optimal results.

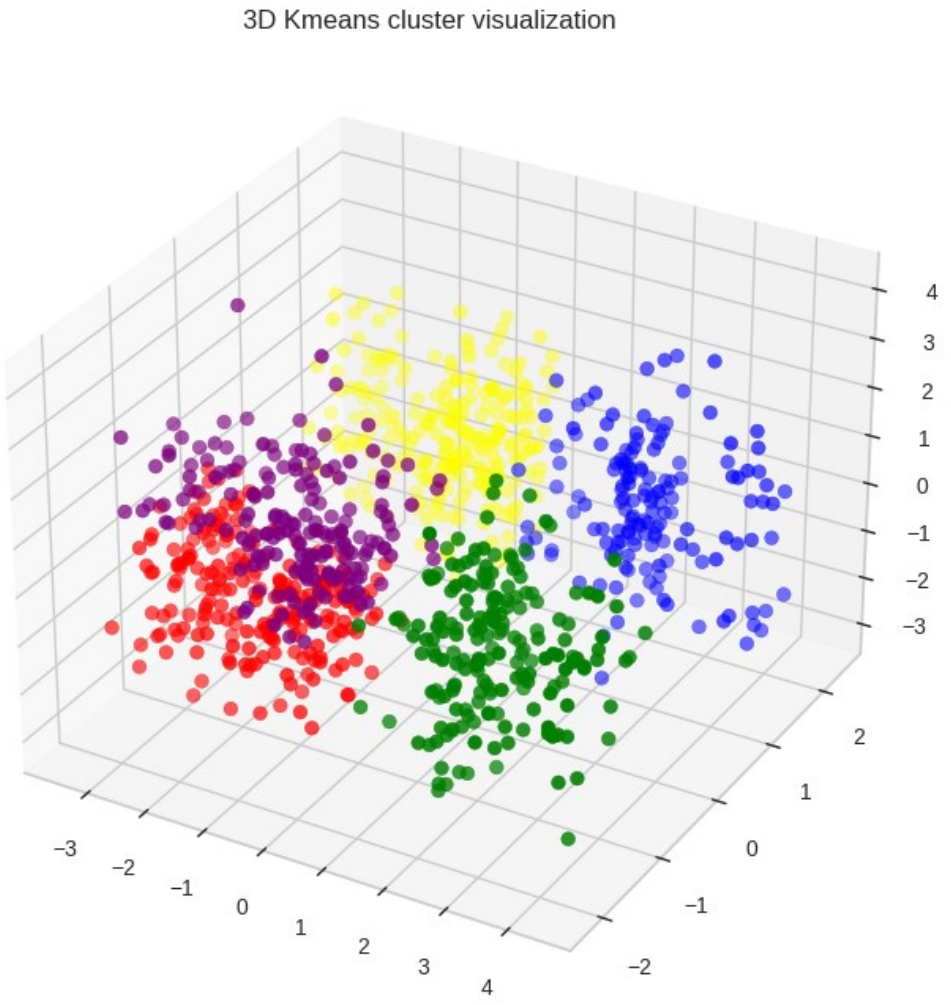
It has been evaluated for the following algorithms:


Distortion Score Elbow for KMeans Clustering
elbow at $k = 5$, $score = 2088.611$
distortion score
fit time (seconds)
k
4000
3500
3000
2500
2000
1500
0.06
0.05
0.04
0.03
0.02
2
3
4
5
6
7
8
9
10


Distortion Score Elbow for AgglomerativeClustering Clustering
elbow at $k = 4$, $score = 2497.250$
distortion score
fit time (seconds)
k
4000
3500
3000
2500
2000
1500
0.032
0.030
0.028
0.026
0.024
0.022
0.020
2
3
4
5
6
7
8
9
10

2.3 Kmeans clustering

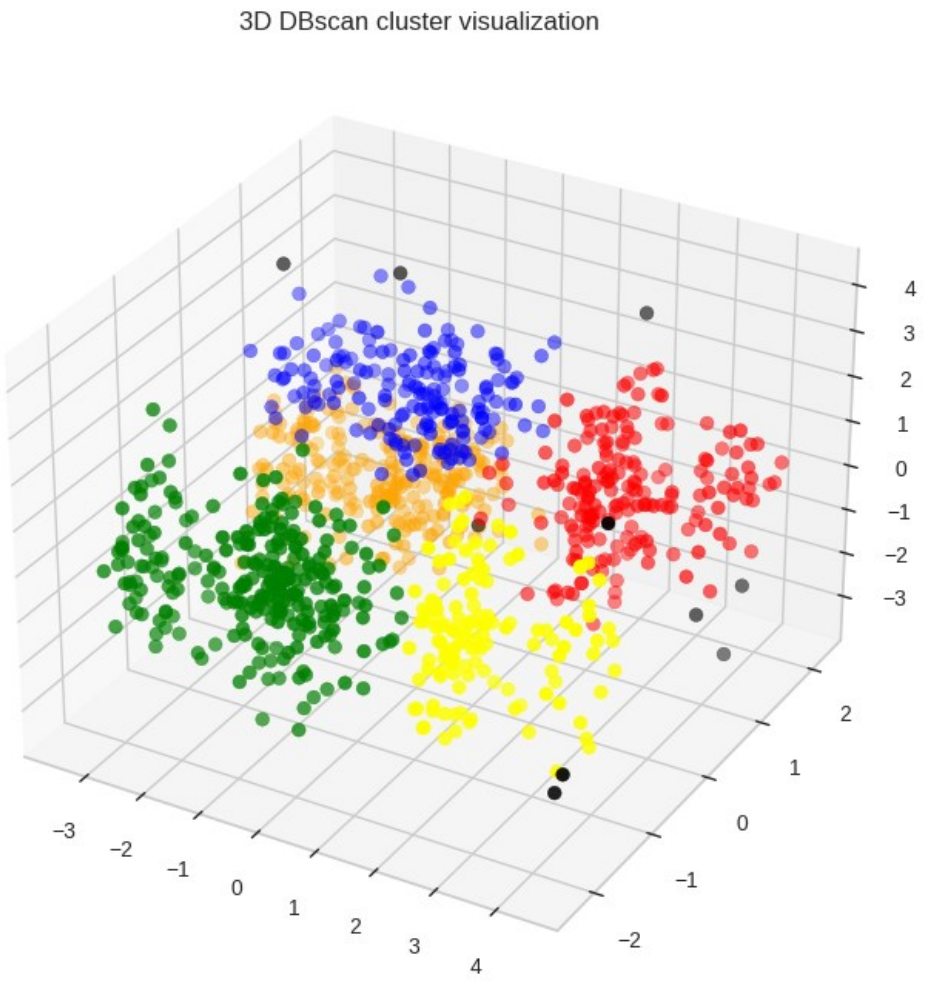
We chose to proceed with this algorithm by specifying the parameter k=5.



2.4 DBscan clustering

We also chose to see the results of an algorithm with a different logic. In black are the noise points,

which are points not assigned to any cluster as they are too distant from the clusters.



2.5 Difference between DBscan and Kmeans

Different algorithms offer different results as highlighted below

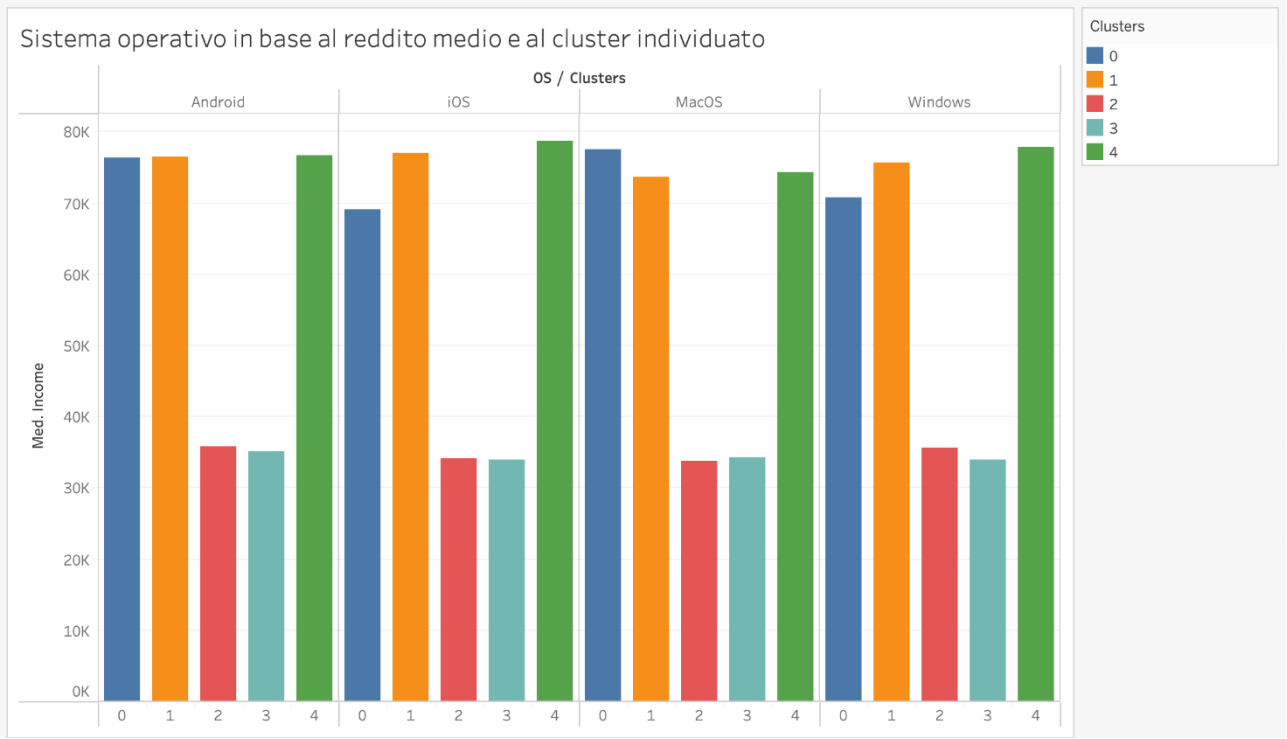
```
[72] km=dwc['Clusters']
     db=dwc['DB_Clusters']

     cont=0

     for i in km:
       if km[i]!=db[i]:
         cont=cont+1

     cont=str(cont)
     print('records with different cluster between dbcan and km are : ' +cont)
```

```
records with different cluster between dbcan and km are : 410
```

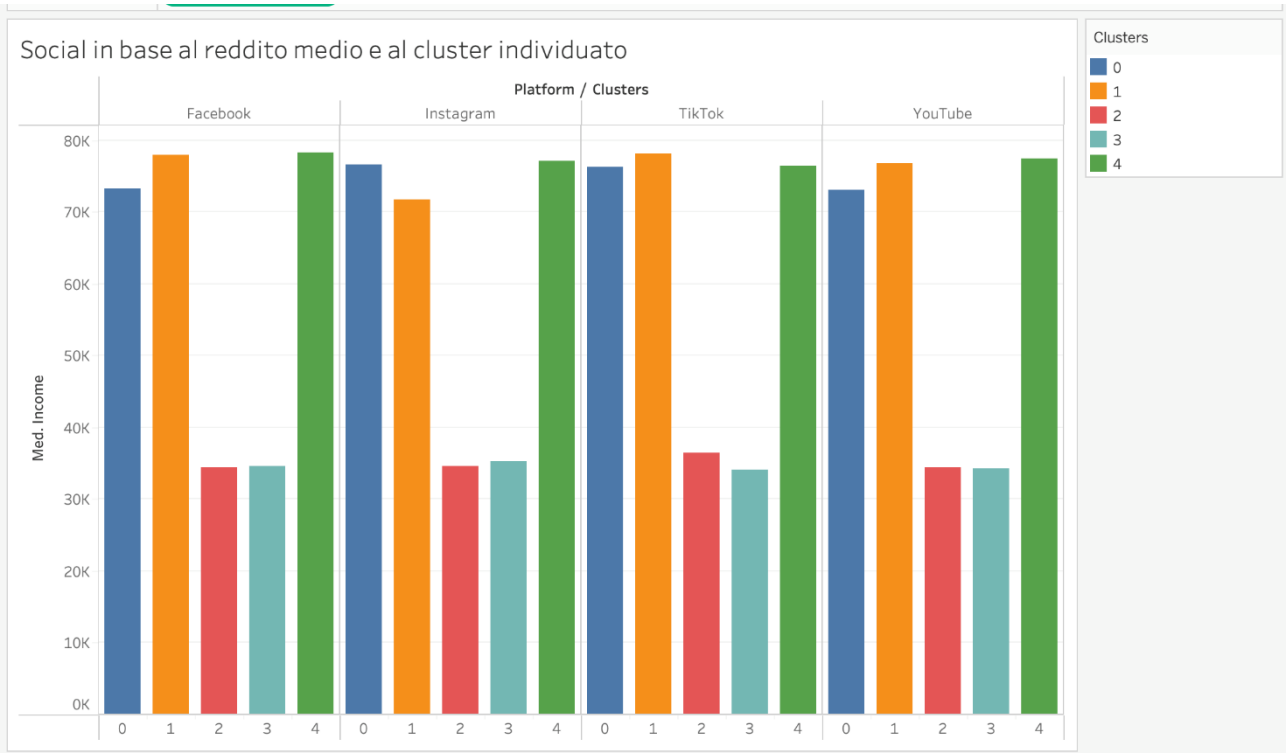
3 Descriptive Analysis

After obtaining 5 clusters, a descriptive analysis is carried out based on the features present in the database.

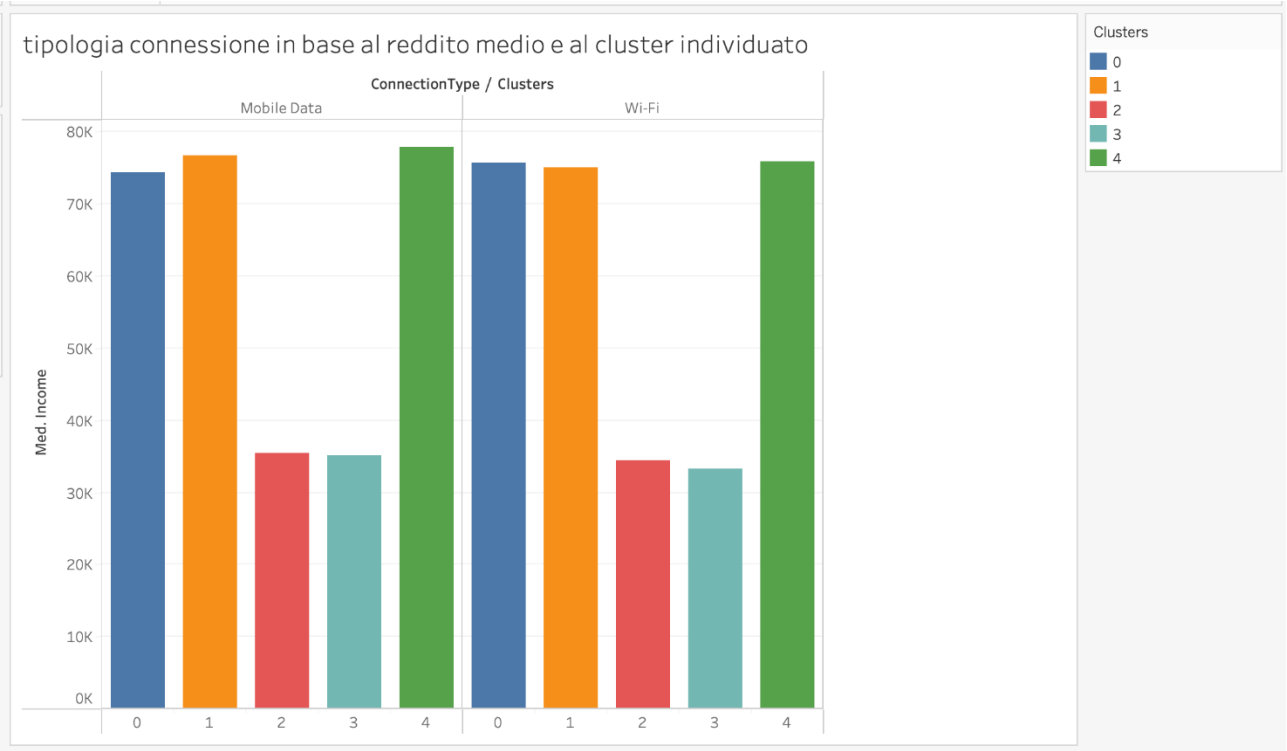
- Operating system



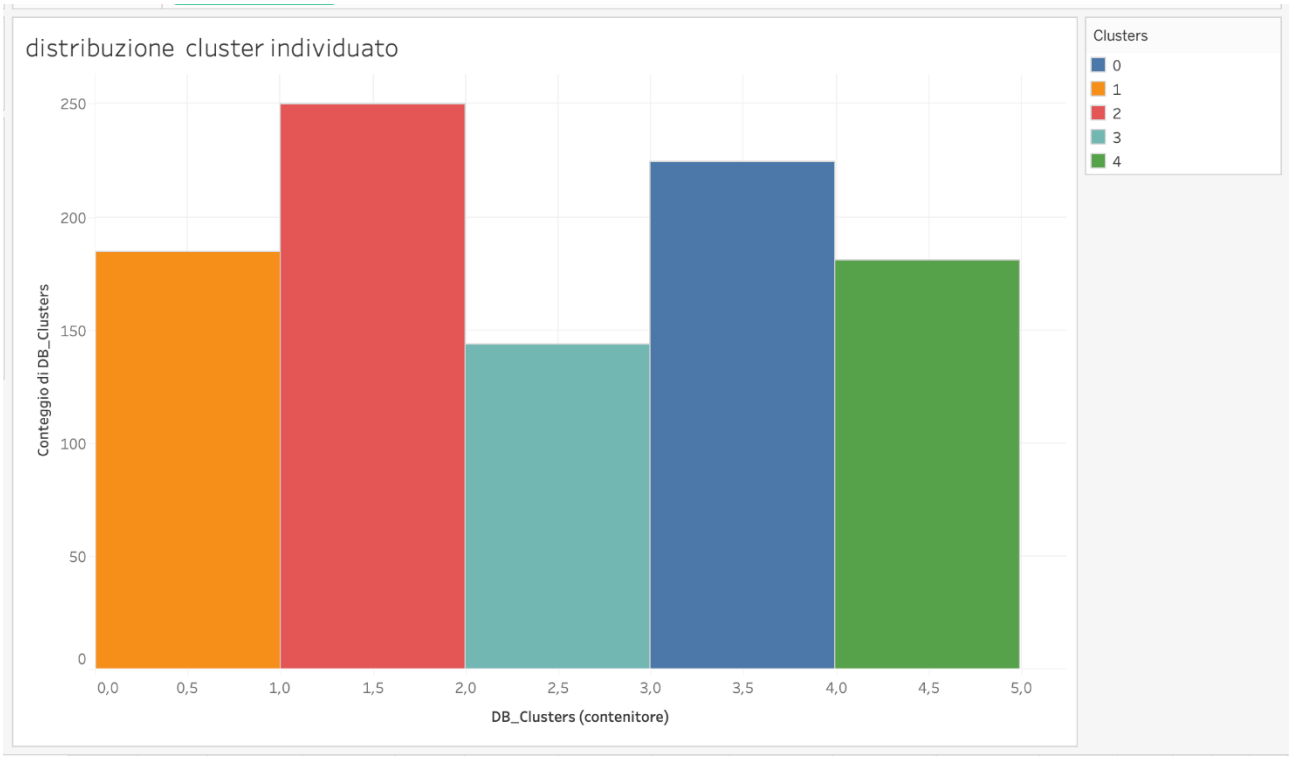
- Social Platform

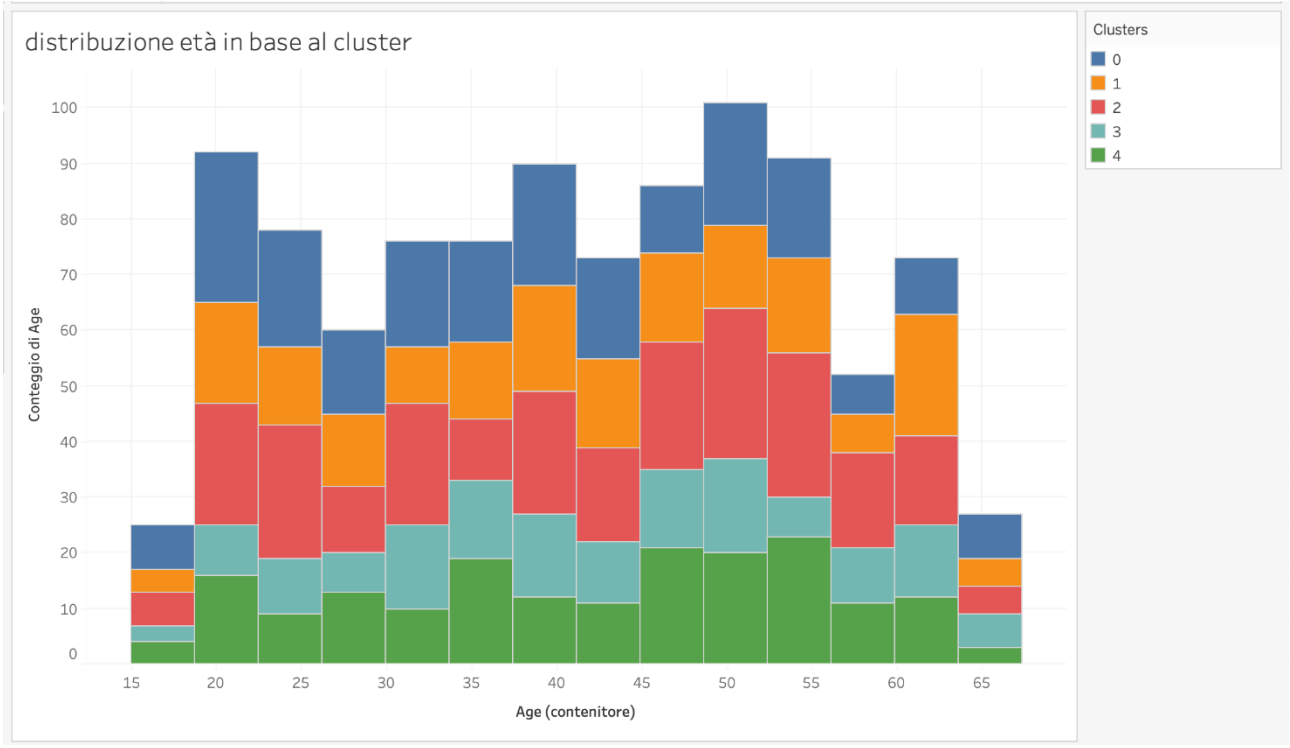


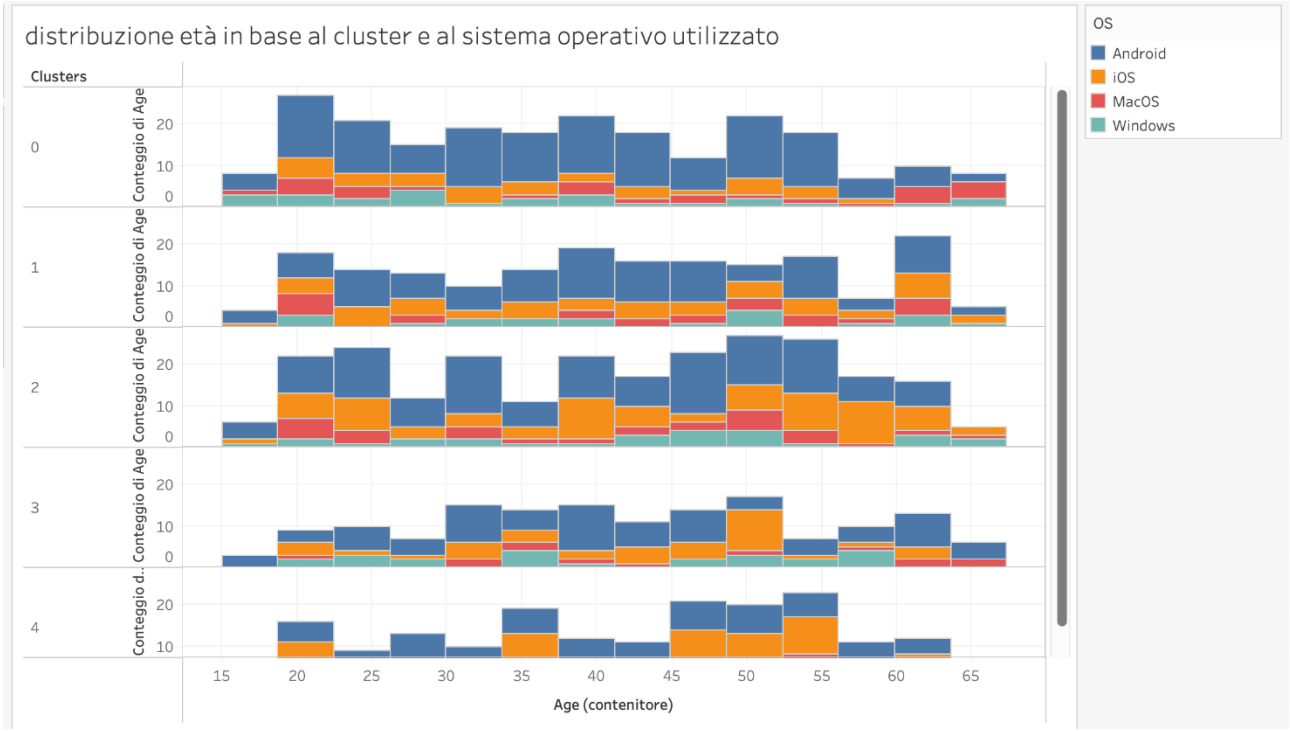
- Type of internet connection

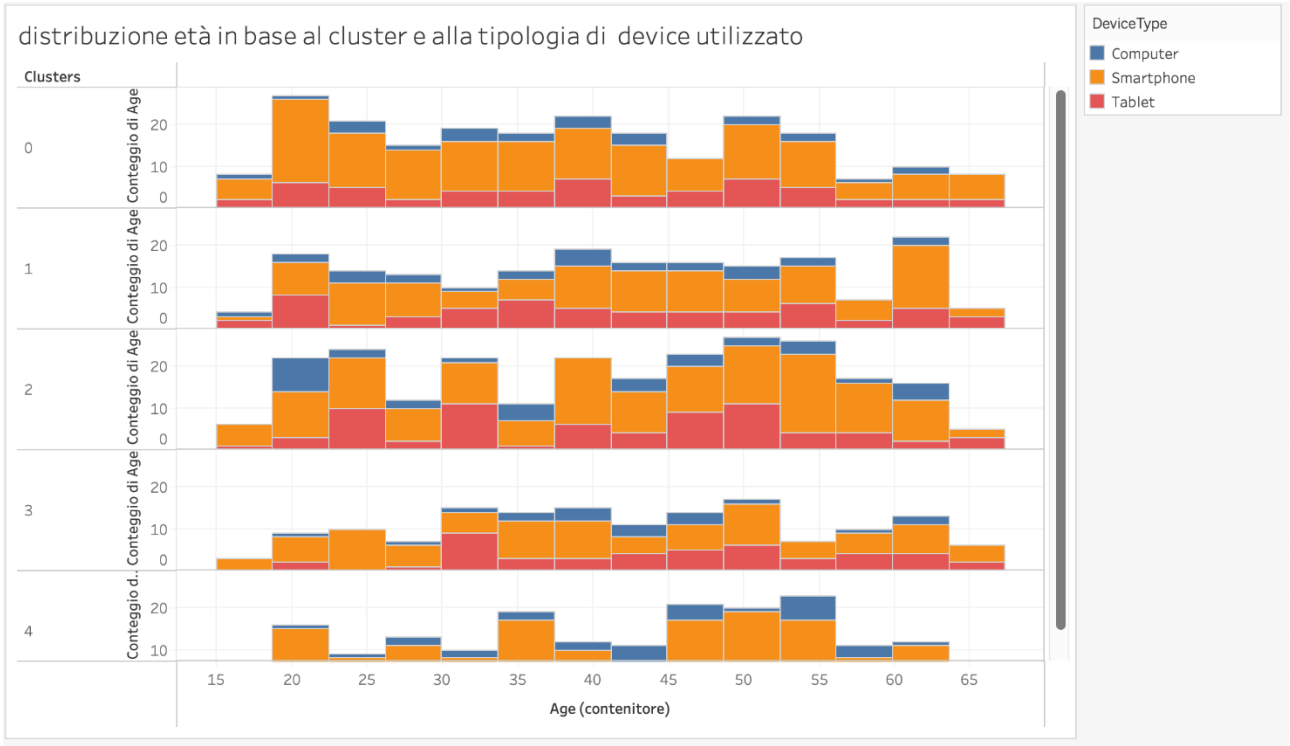


- Cluster distribution

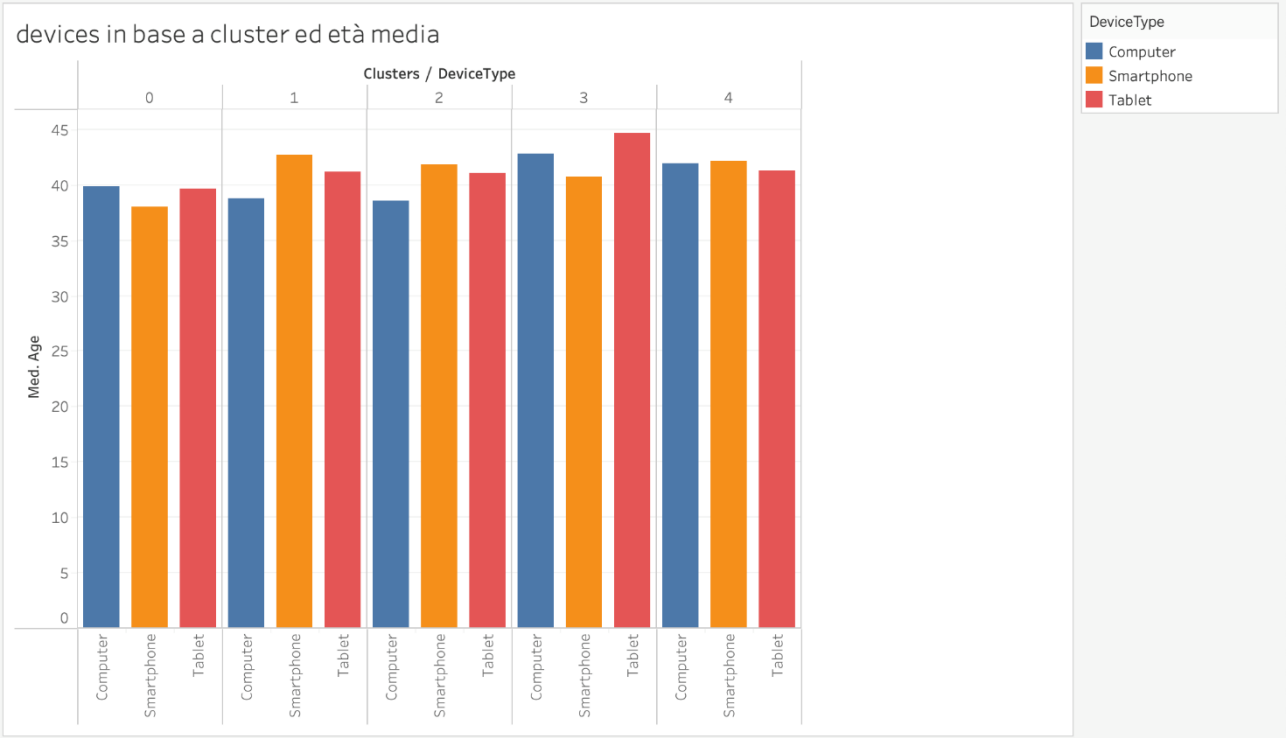

distribuzione cluster individuato
Clusters
0
1
2
3
4
Conteggio di DB_Clusters
250
200
150
100
50
0
0,0
0,5
1,0
1,5
2,0
2,5
3,0
3,5
4,0
4,5
5,0
DB_Clusters (contenitore)


distribuzione età in base al cluster
Clusters
0
1
2
3
4
Conteggio di Age
100
90
80
70
60
50
40
30
20
10
0
15
20
25
30
35
40
45
50
55
60
65
Age (contenitore)

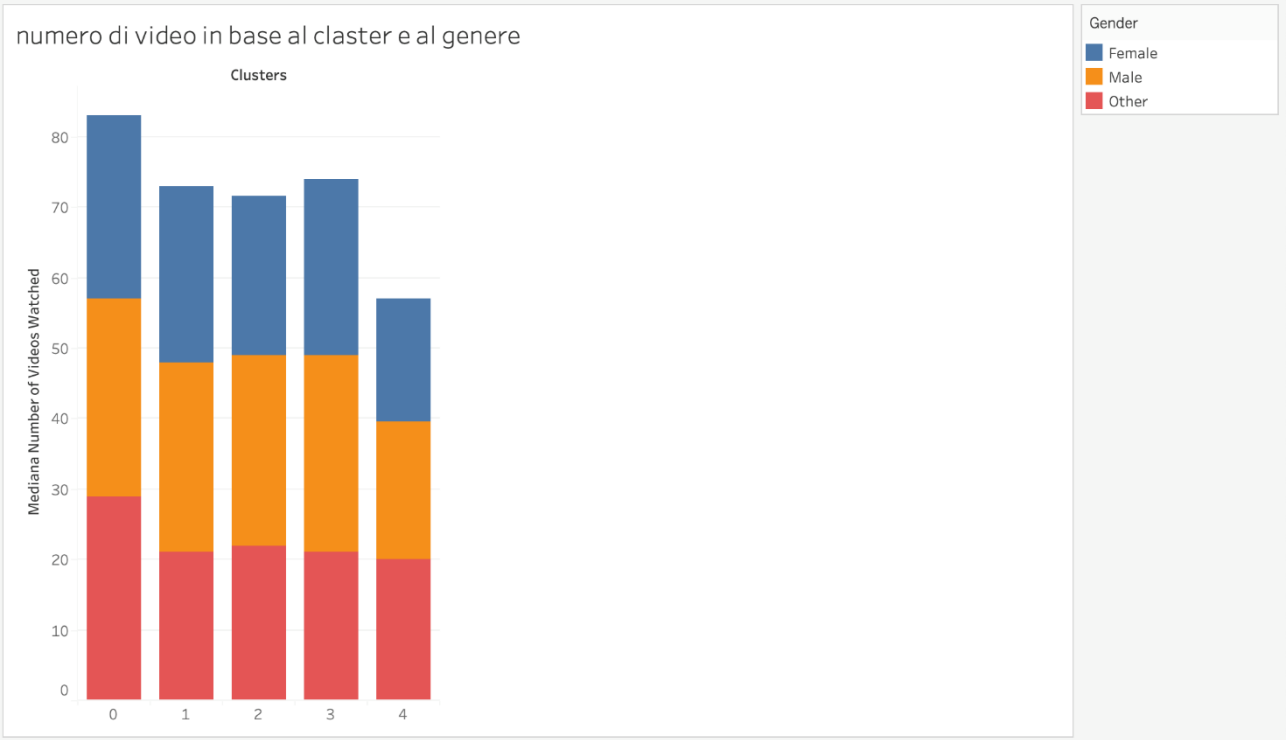




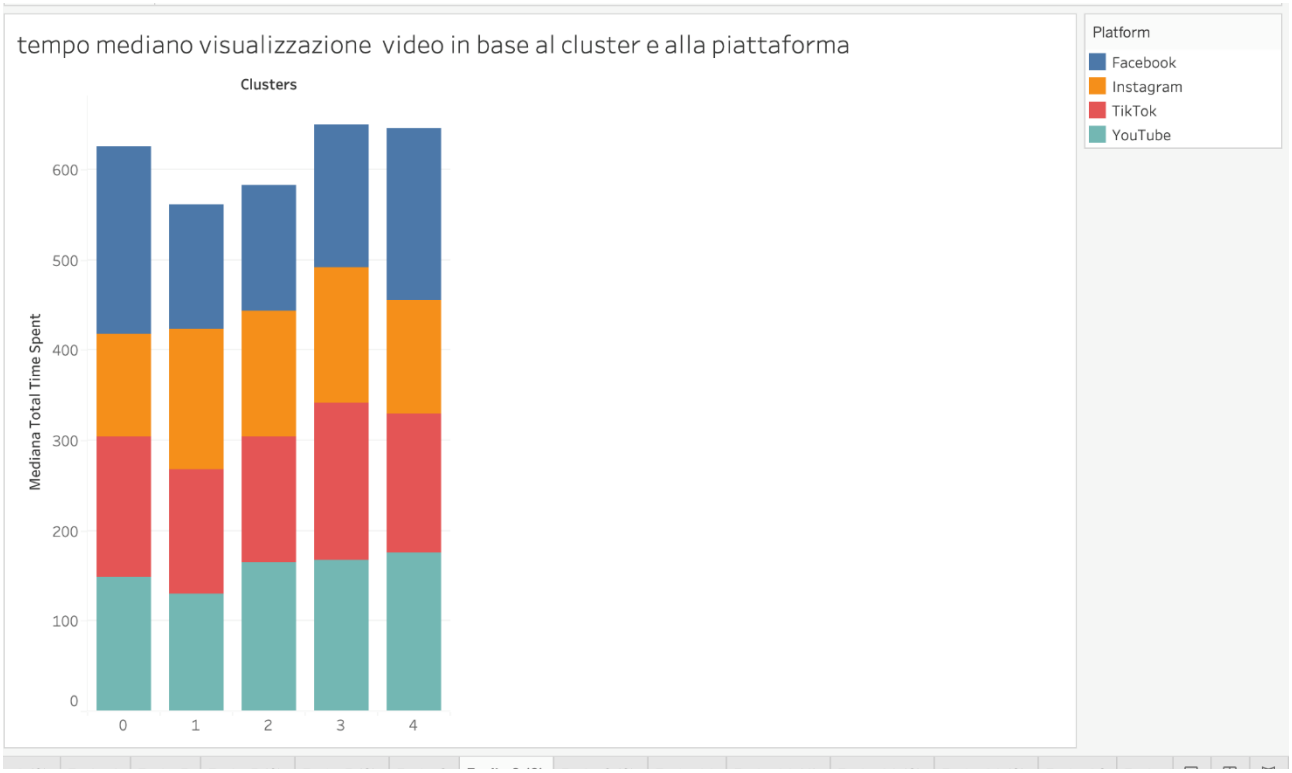
- Type of device based on user age



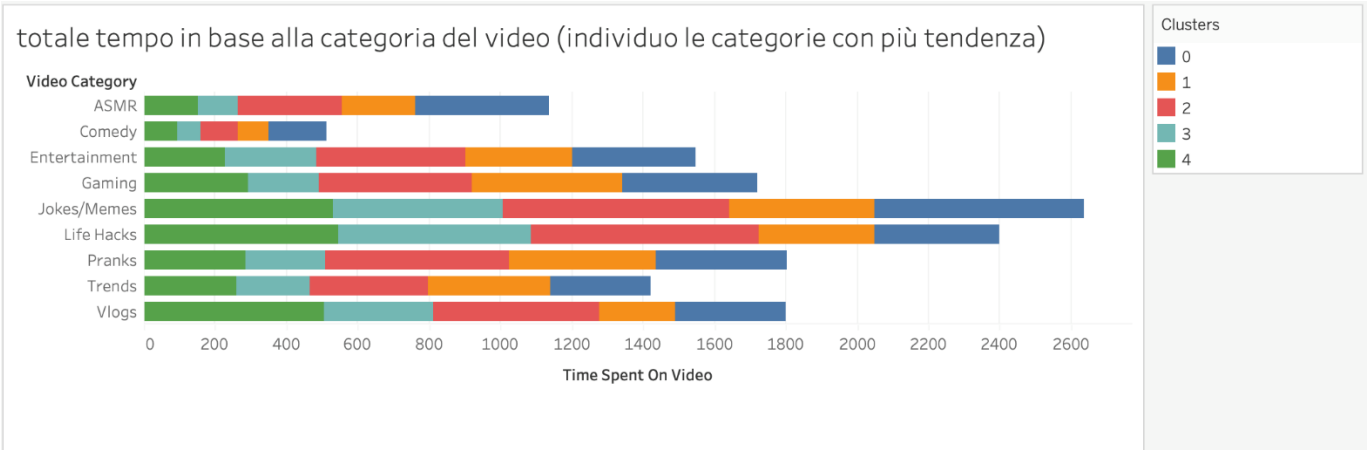
- Number of videos based on sex and cluster

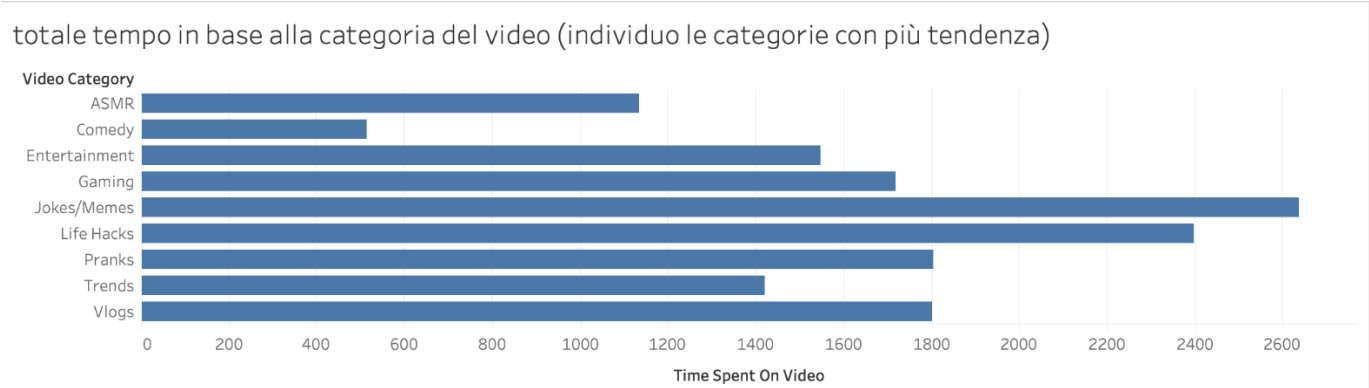


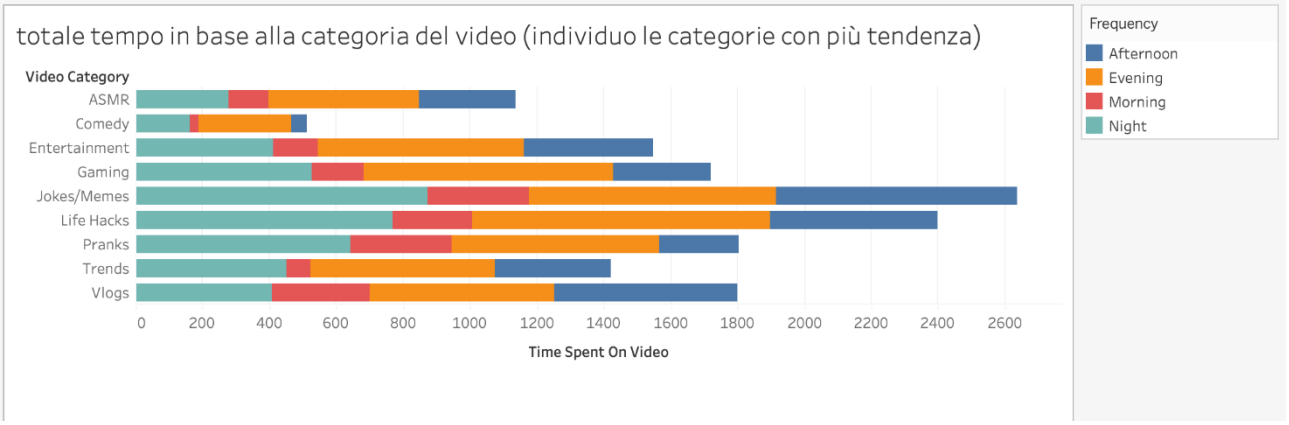
- Average time on platform

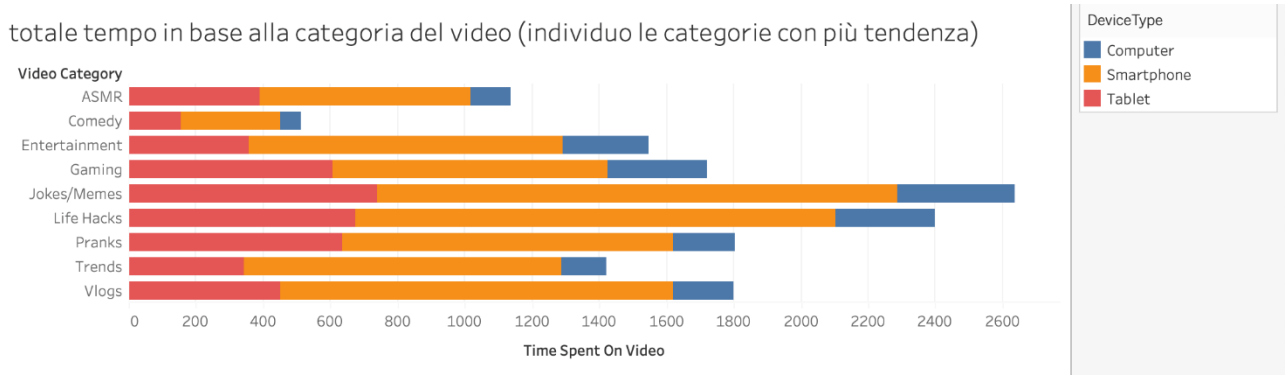


- Total time based on video category


totale tempo in base alla categoria del video (individuo le categorie con più tendenza)
Video Category
ASMR
Comedy
Entertainment
Gaming
Jokes/Memes
Life Hacks
Pranks
Trends
Vlogs
0 200 400 600 800 1000 1200 1400 1600 1800 2000 2200 2400 2600
Time Spent On Video
Clusters
0
1
2
3
4


totale tempo in base alla categoria del video (individuo le categorie con più tendenza)
Video Category
ASMR
Comedy
Entertainment
Gaming
Jokes/Memes
Life Hacks
Pranks
Trends
Vlogs
0 200 400 600 800 1000 1200 1400 1600 1800 2000 2200 2400 2600
Time Spent On Video


totale tempo in base alla categoria del video (individuo le categorie con più tendenza)
Video Category
ASMR
Comedy
Entertainment
Gaming
Jokes/Memes
Life Hacks
Pranks
Trends
Vlogs
0 200 400 600 800 1000 1200 1400 1600 1800 2000 2200 2400 2600
Time Spent On Video
Frequency
Afternoon
Evening
Morning
Night


totale tempo in base alla categoria del video (individuo le categorie con più tendenza)
Video Category
ASMR
Comedy
Entertainment
Gaming
Jokes/Memes
Life Hacks
Pranks
Trends
Vlogs
0 200 400 600 800 1000 1200 1400 1600 1800 2000 2200 2400 2600
Time Spent On Video
DeviceType
Computer
Smartphone
Tablet

- Reason for use of social media