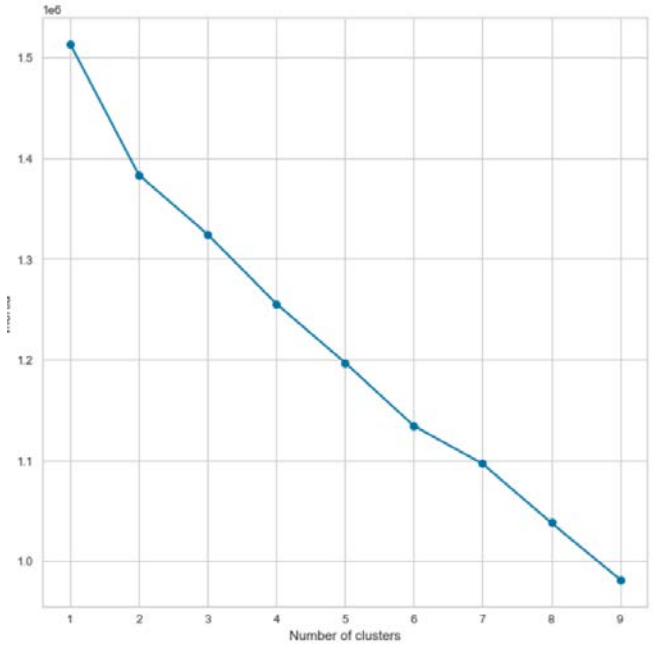
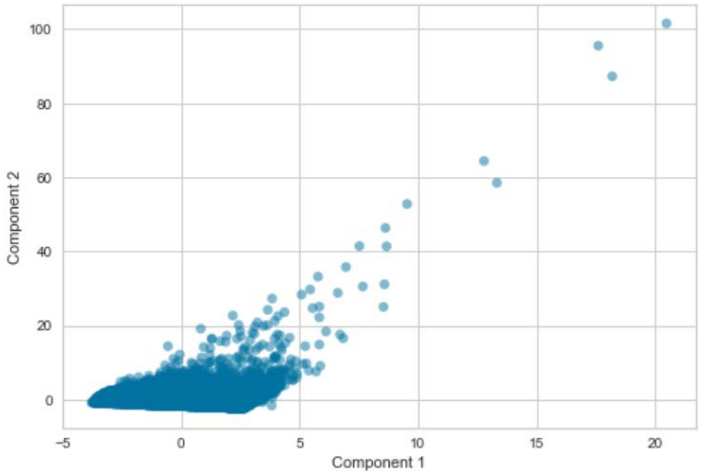
TEST CLUSTERING RIMOZIONE FEATURE

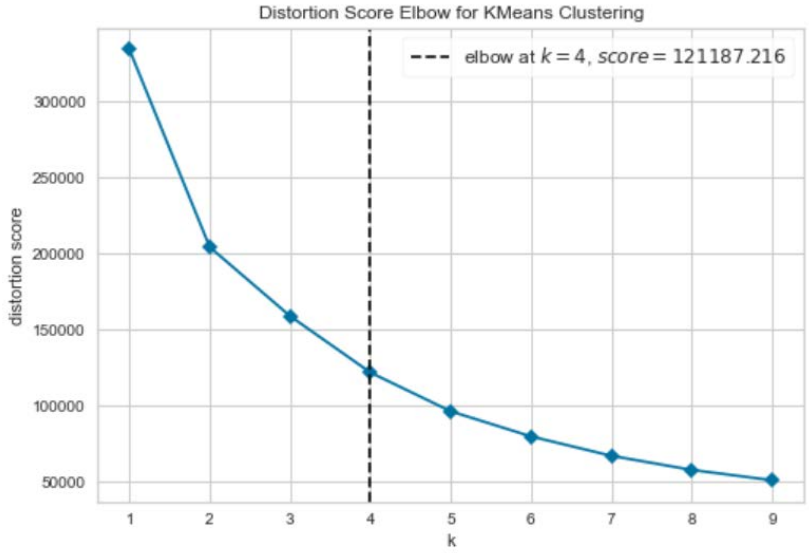
1° considero solo Interval_Age e Tot_revenue

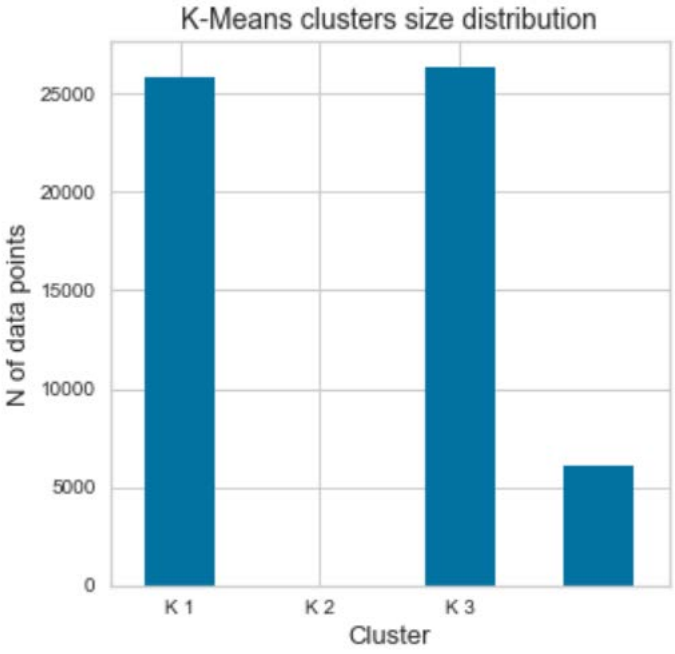
```
cluster_dataset.drop('ID', axis = 1 , inplace = True)
cluster_dataset.drop('NameHash', axis = 1 , inplace = True)
cluster_dataset.drop('DocIDHash', axis = 1 , inplace = True)
cluster_dataset.drop('Age', axis = 1 , inplace = True)
cluster_dataset.drop('LodgingRevenue', axis = 1 , inplace = True)
cluster_dataset.drop('OtherRevenue', axis = 1 , inplace = True)
```

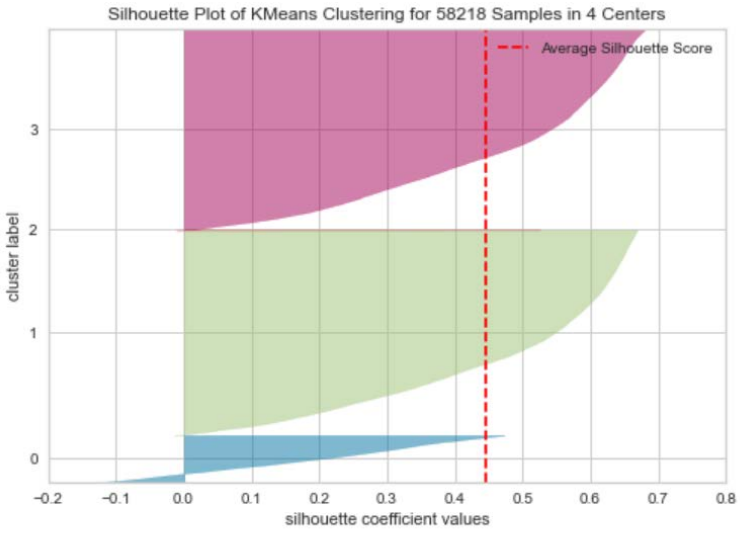


k-means no pca = 0.1497417138996520 5

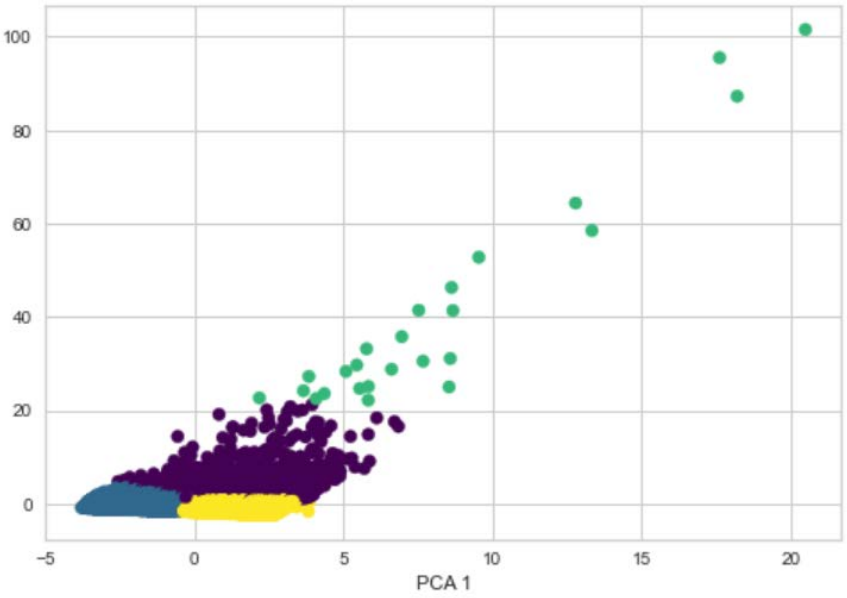




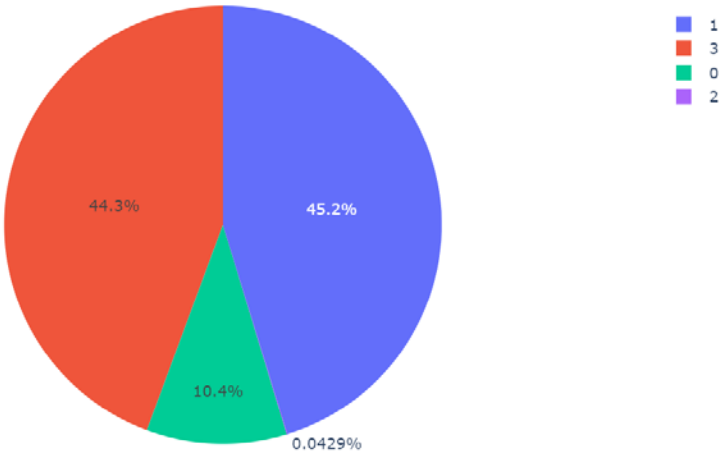




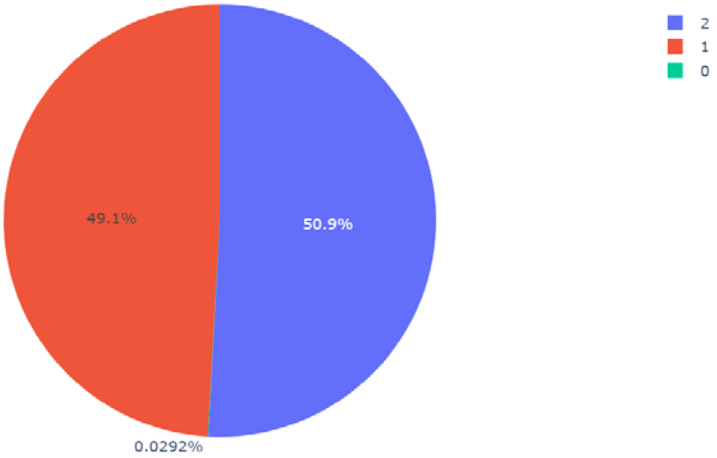

```
0.4466639922033892
```



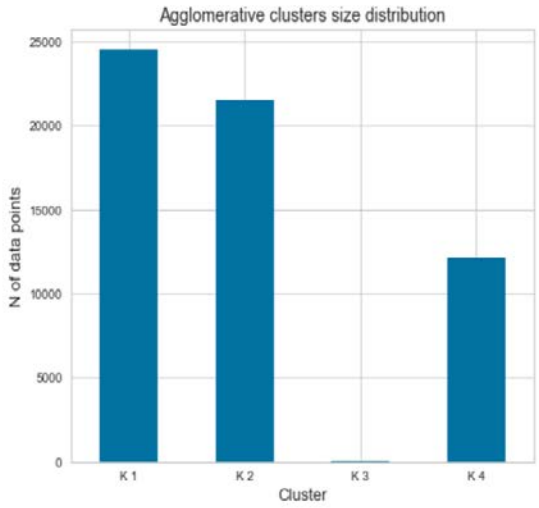
Distribution of Clusters K-means with PCA



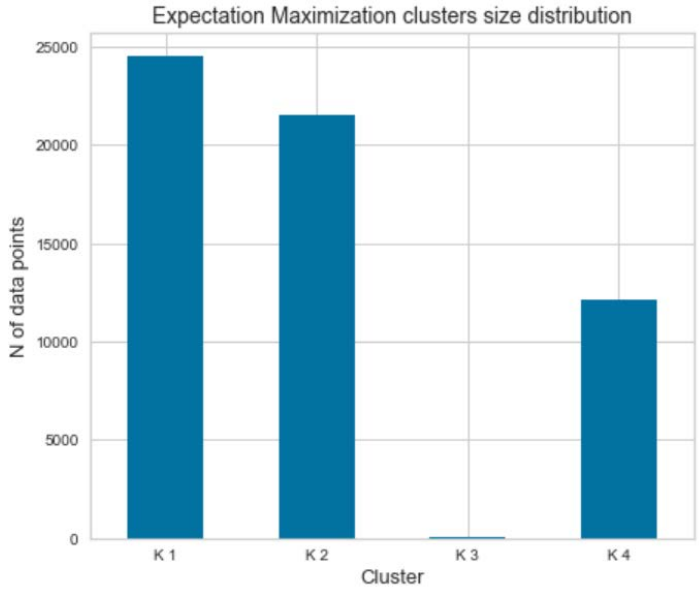
Distribution of Clusters K-means no PCA

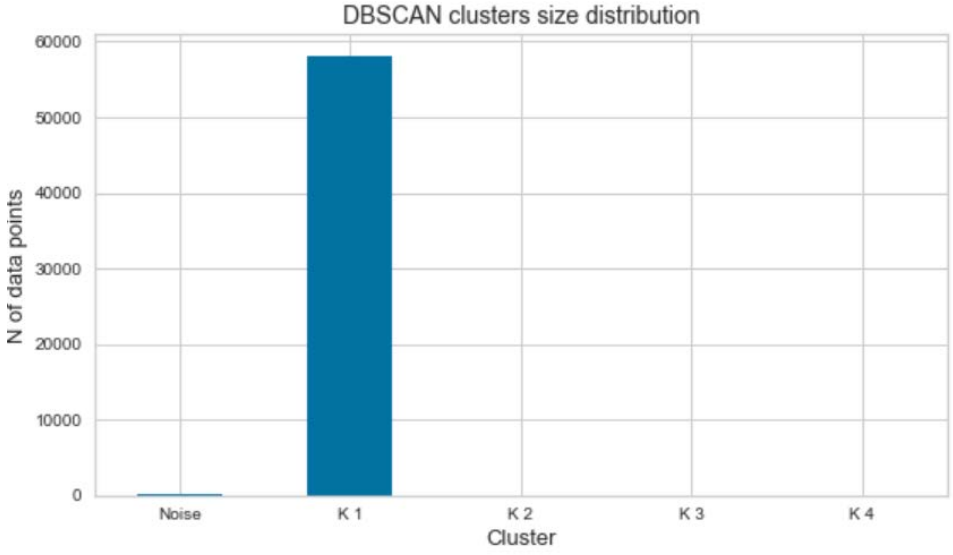


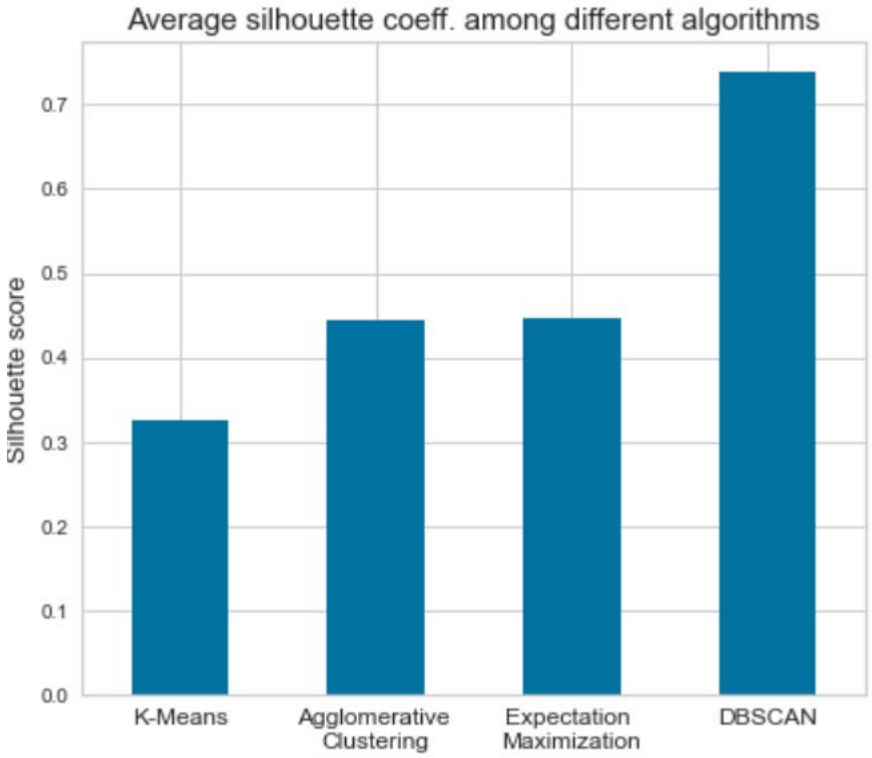
```
The Silhouette coefficient for the Agglomerative Hierarchical Clustering algorithm is 0.33
```




```
The Silhouette coefficient for the EM Clustering algorithm is 0.45
```

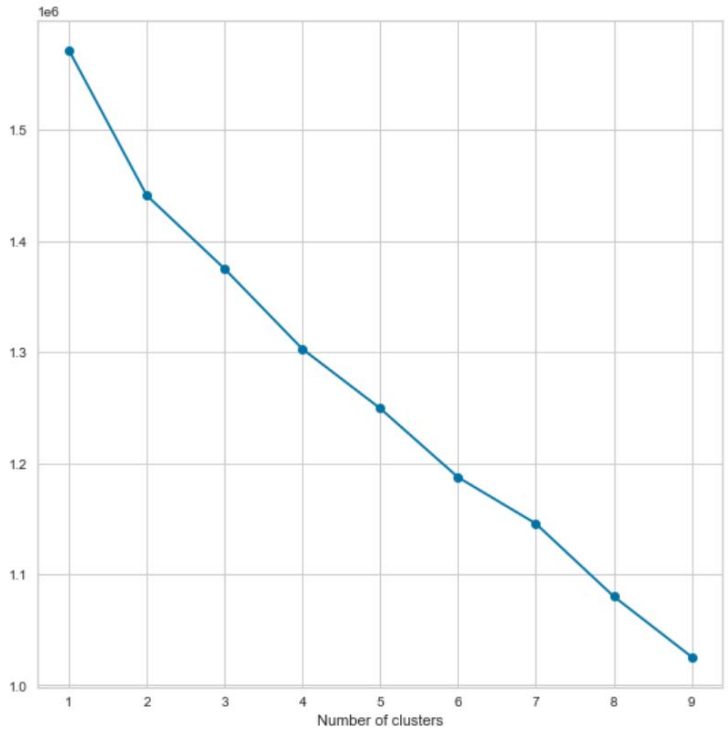




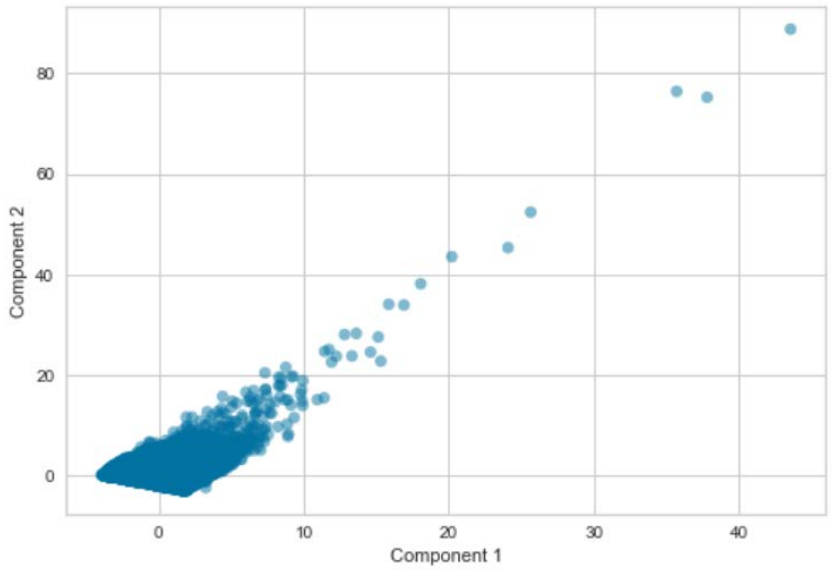

Average silhouette coeff. among different algorithms
Silhouette score
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
K-Means
Agglomerative Clustering
Expectation Maximization
DBSCAN

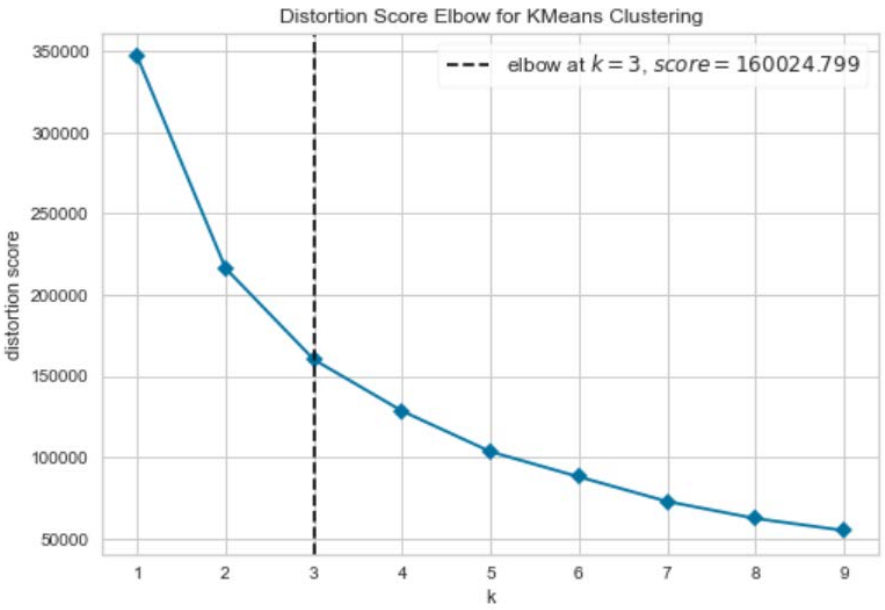
2° contrario del primo passo

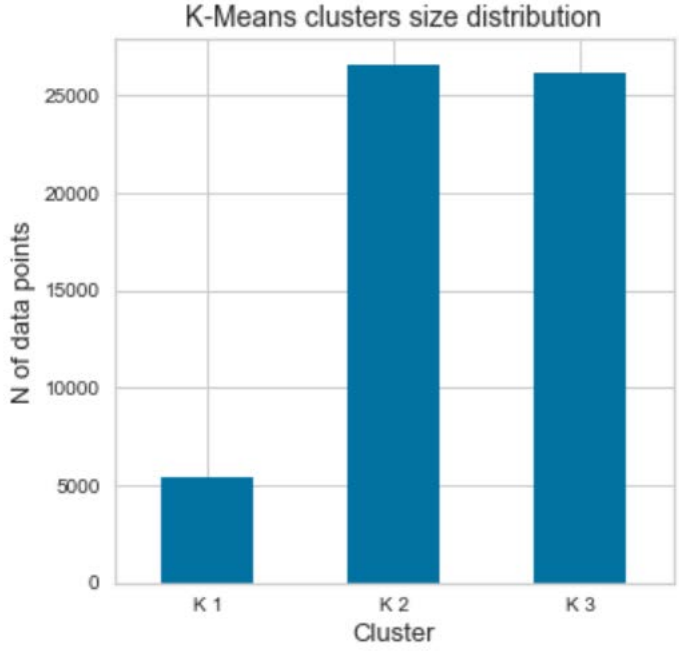
```
cluster_dataset.drop('ID', axis = 1 , inplace = True)
cluster_dataset.drop('NameHash', axis = 1 , inplace = True)
cluster_dataset.drop('DocIDHash', axis = 1 , inplace = True)
cluster_dataset.drop('Tot_Revenue', axis = 1 , inplace = True)
cluster_dataset.drop('Interval_age', axis = 1 , inplace = True)
```

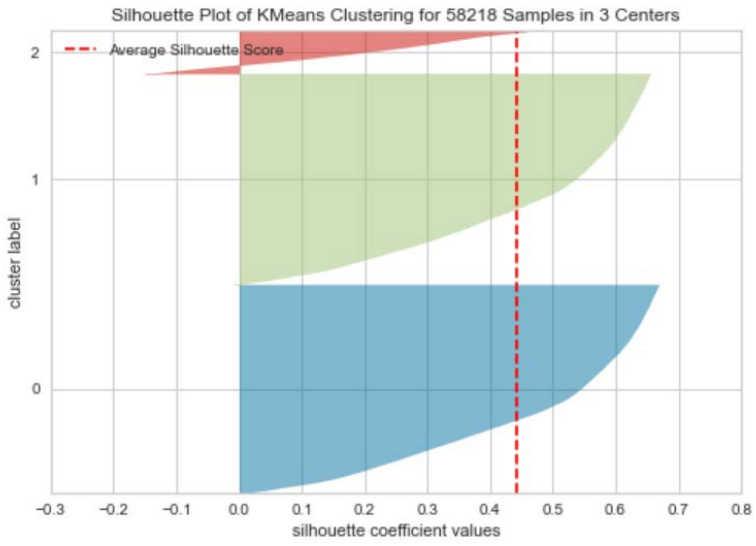


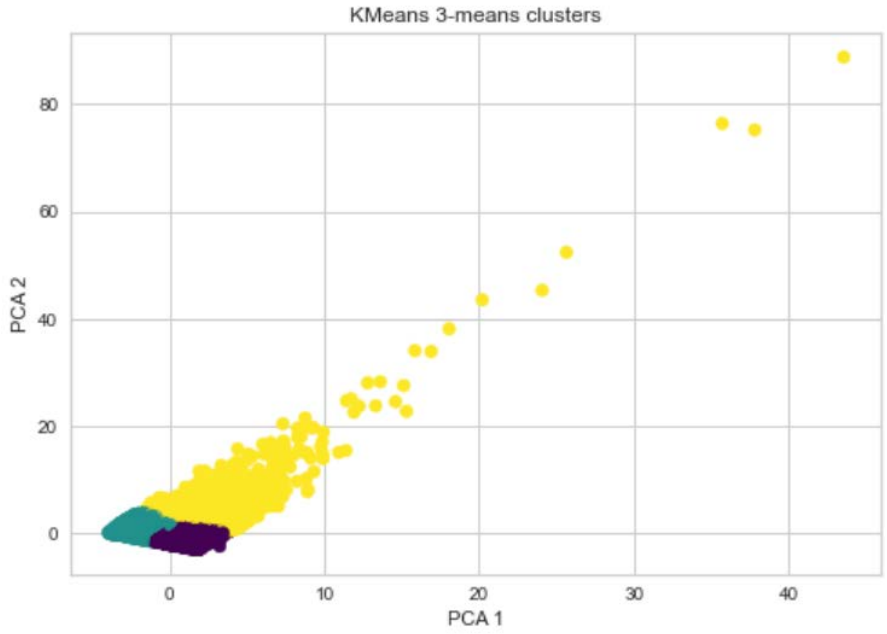
k-means no pca = 0.13490584447019968



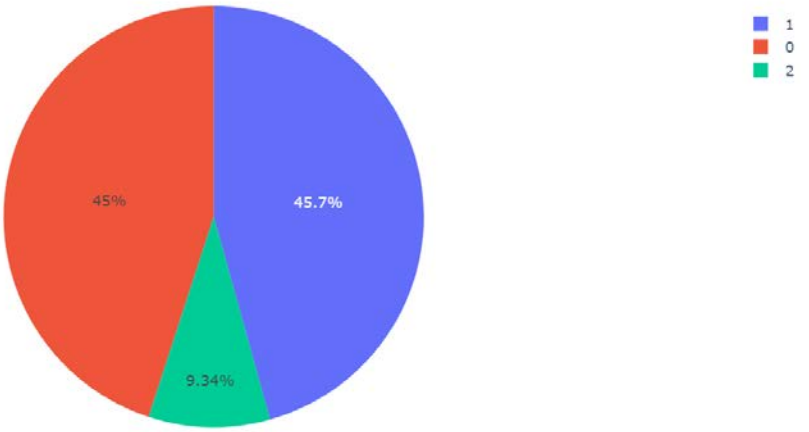

Distortion Score Elbow for KMeans Clustering
elbow at $k = 3$, $score = 160024.799$
distortion score
k
350000
300000
250000
200000
150000
100000
50000
1
2
3
4
5
6
7
8
9


K-Means clusters size distribution
N of data points
Cluster
25000
20000
15000
10000
5000
0
K 1
K 2
K 3

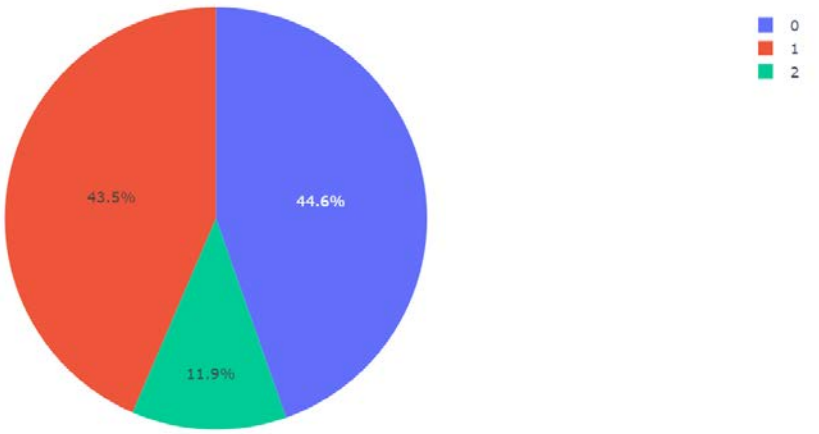

Silhouette Plot of KMeans Clustering for 58218 Samples in 3 Centers
Average Silhouette Score
cluster label
silhouette coefficient values
2
1
0
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8



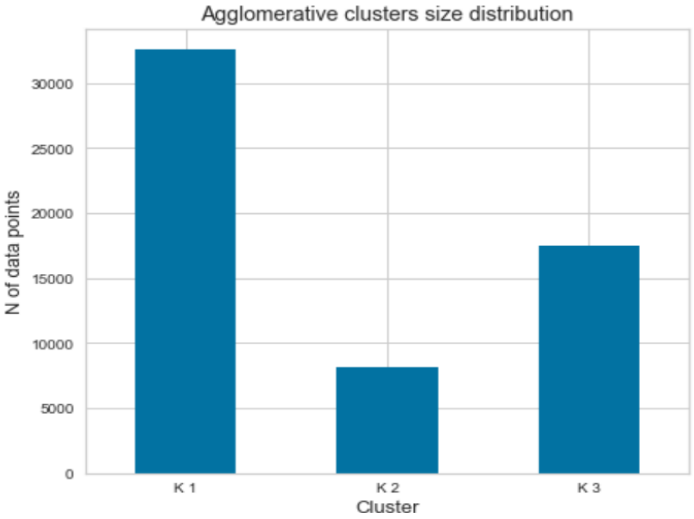
Distribution of Clusters K-means with PCA



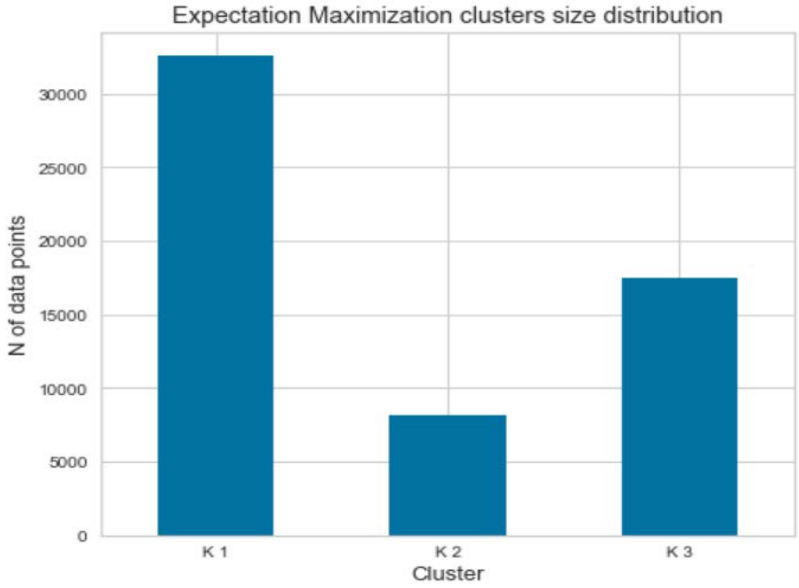
Distribution of Clusters K-means no PCA

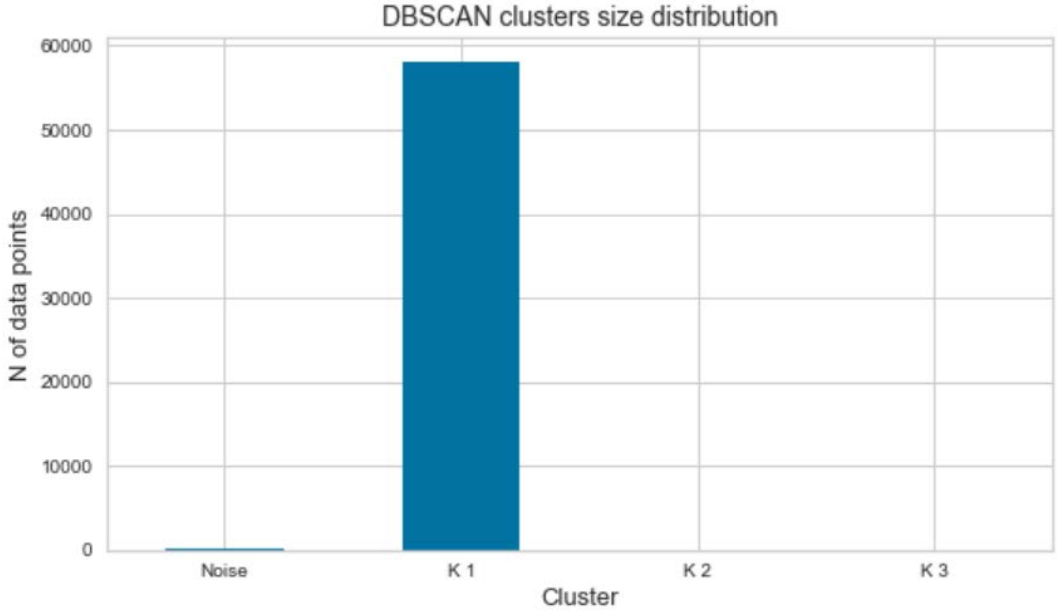


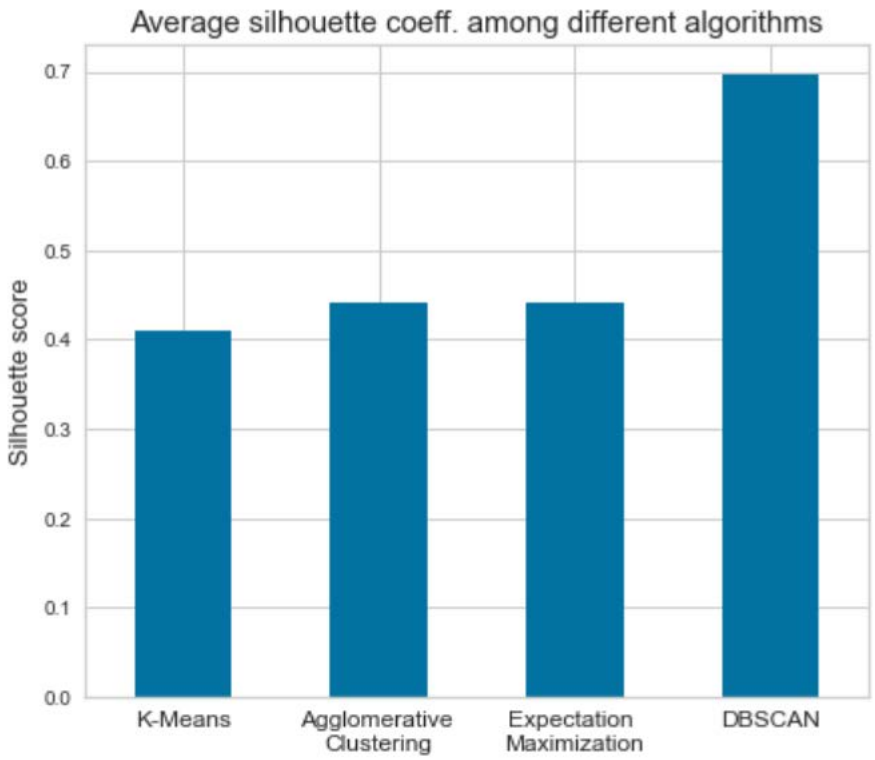
The Silhouette coefficient for the Agglomerative Hierarchical Clustering algorithm is 0.41



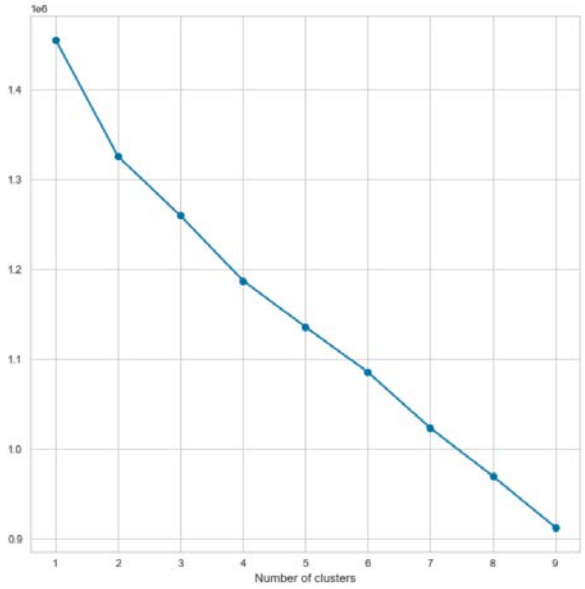
The Silhouette coefficient for the EM Clustering algorithm is 0.44





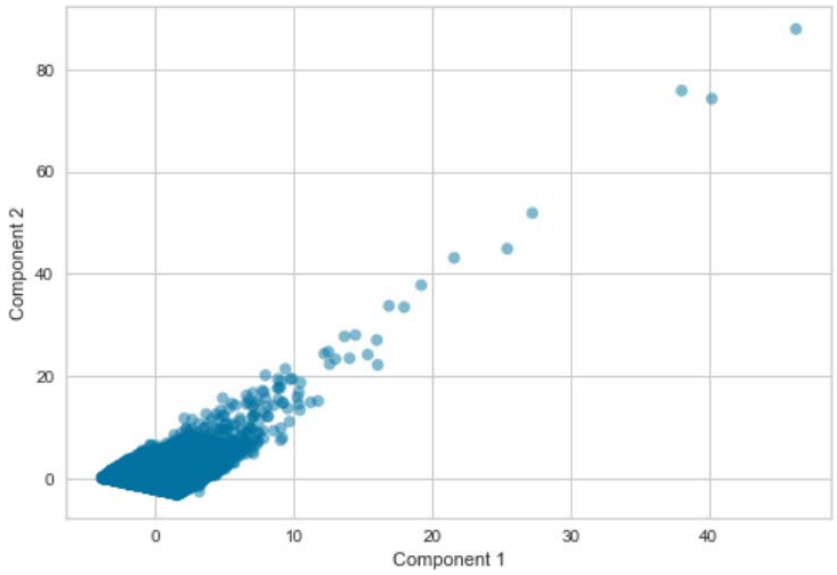

Average silhouette coeff. among different algorithms
Silhouette score
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
K-Means
Agglomerative Clustering
Expectation Maximization
DBSCAN

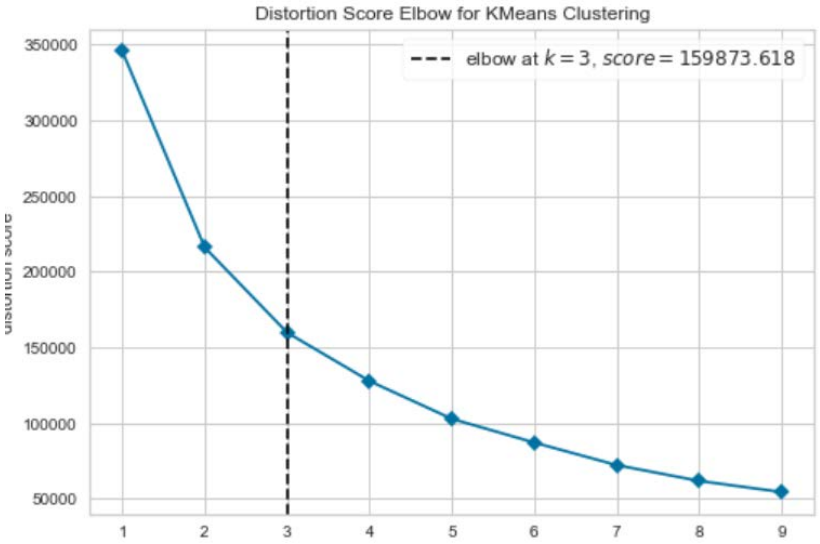
3° elimino le colonne originariamente non numeriche

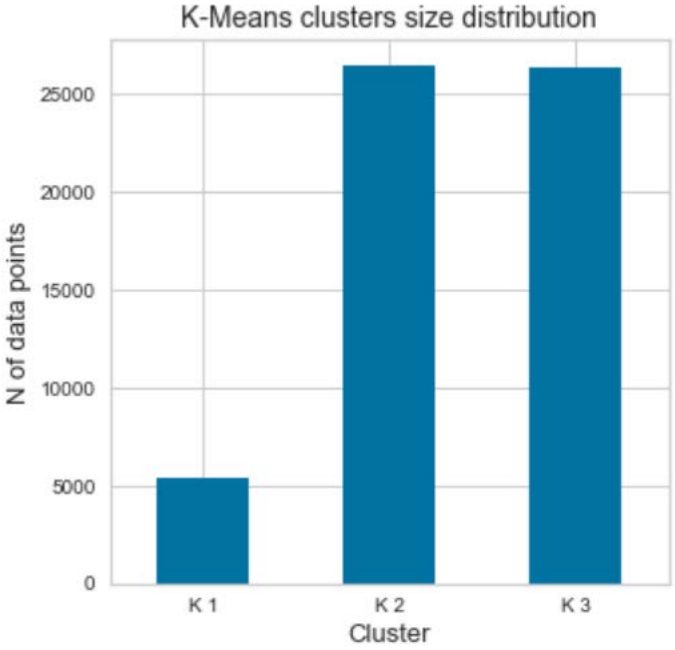


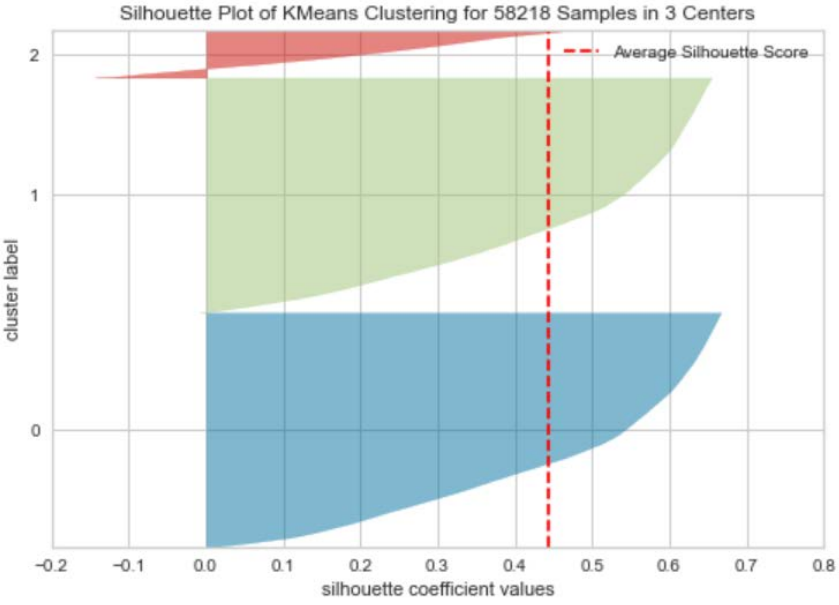

```
cluster_dataset.drop('ID', axis = 1 , inplace = True)
cluster_dataset.drop('NameHash', axis = 1 , inplace = True)
cluster_dataset.drop('DocIDHash', axis = 1 , inplace = True)
cluster_dataset.drop('Tot_Revenue', axis = 1 , inplace = True)
cluster_dataset.drop('Interval_age', axis = 1 , inplace = True)
cluster_dataset.drop('Nationality', axis = 1 , inplace = True)
cluster_dataset.drop('DistributionChannel', axis = 1 , inplace = True)
cluster_dataset.drop('MarketSegment', axis = 1 , inplace = True)
```

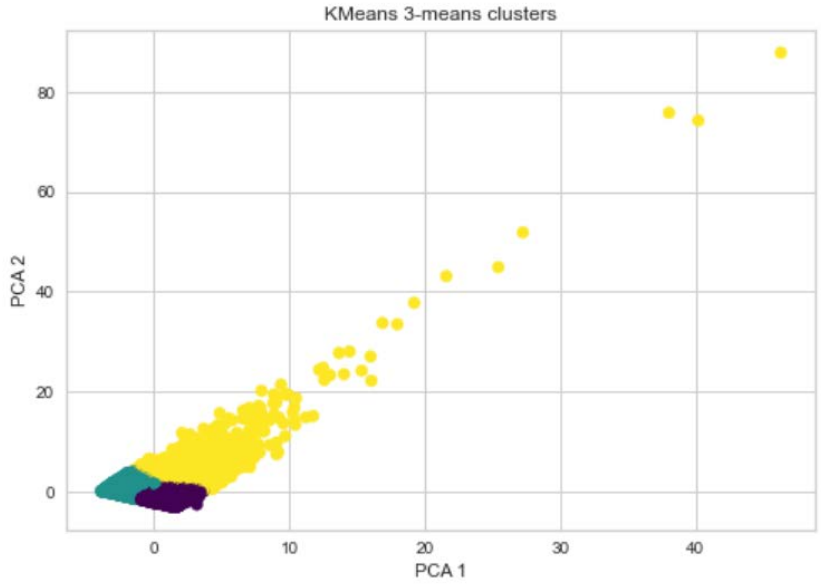
k-means no pca = 0.1620018609820730 6



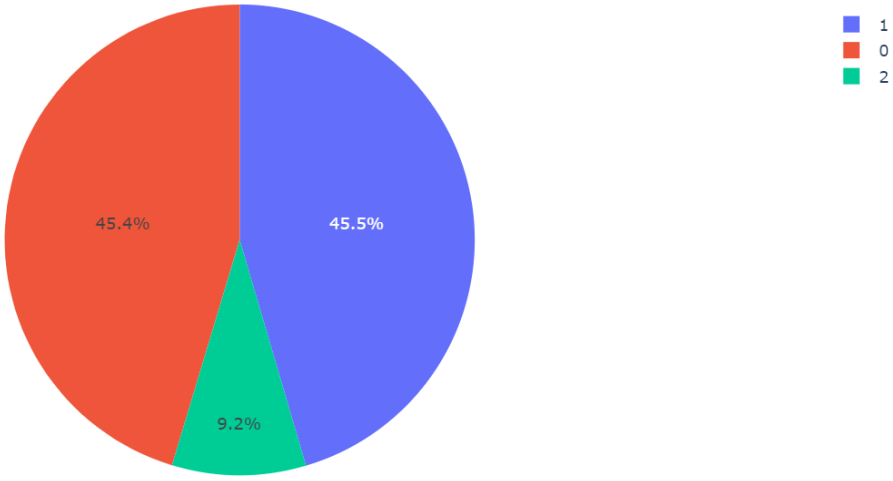

Distortion Score Elbow for KMeans Clustering
elbow at $k = 3$, $score = 159873.618$
350000
300000
250000
200000
150000
100000
50000
1
2
3
4
5
6
7
8
9


K-Means clusters size distribution
N of data points
25000
20000
15000
10000
5000
0
K 1
K 2
K 3
Cluster

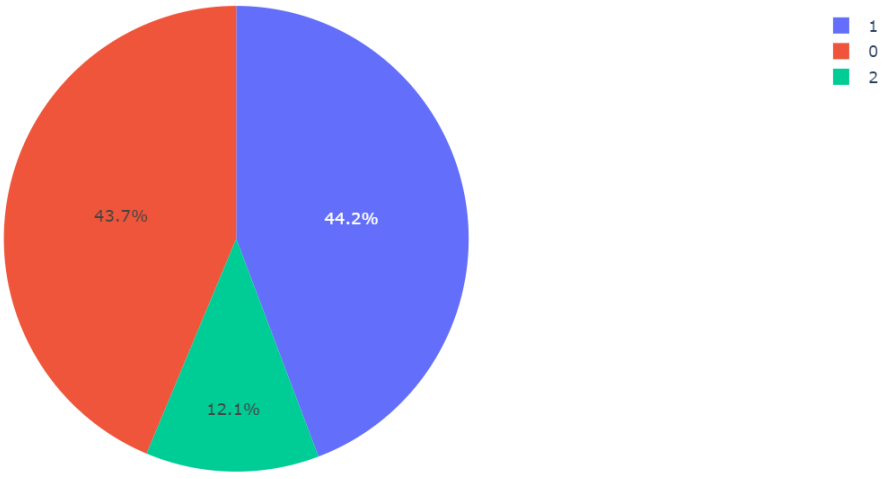

Silhouette Plot of KMeans Clustering for 58218 Samples in 3 Centers
Average Silhouette Score
cluster label
2
1
0
−0.2
−0.1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
silhouette coefficient values



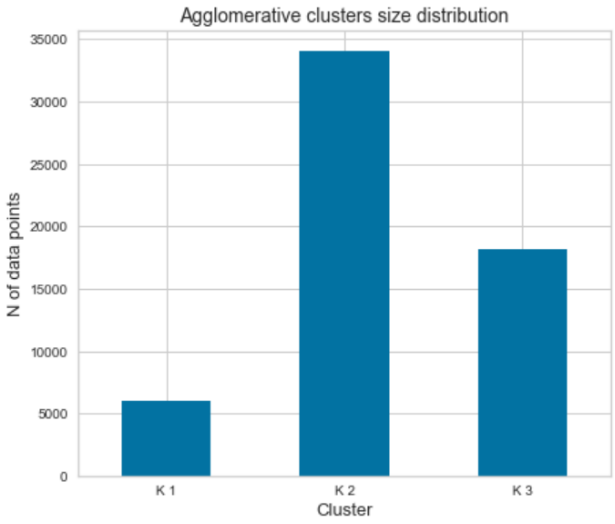
Distribution of Clusters K-means with PCA



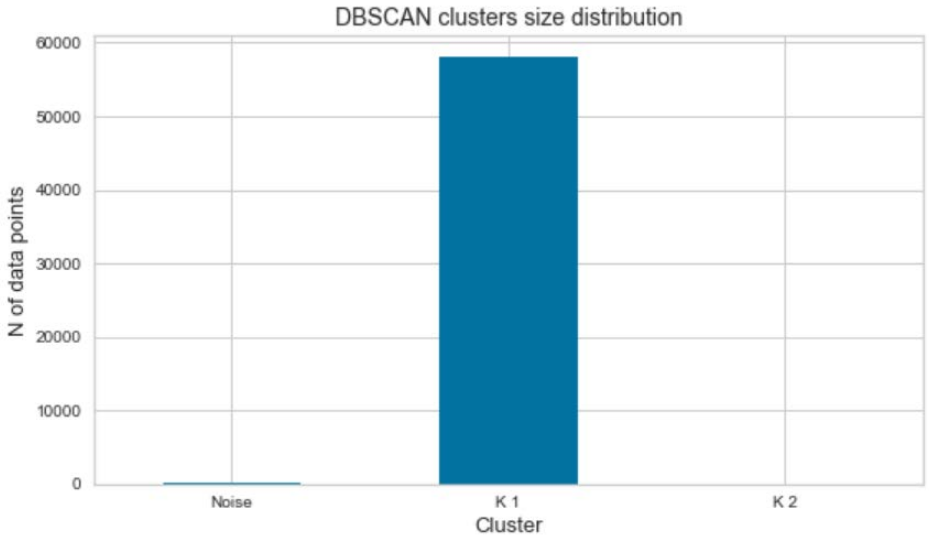
Distribution of Clusters K-means no PCA

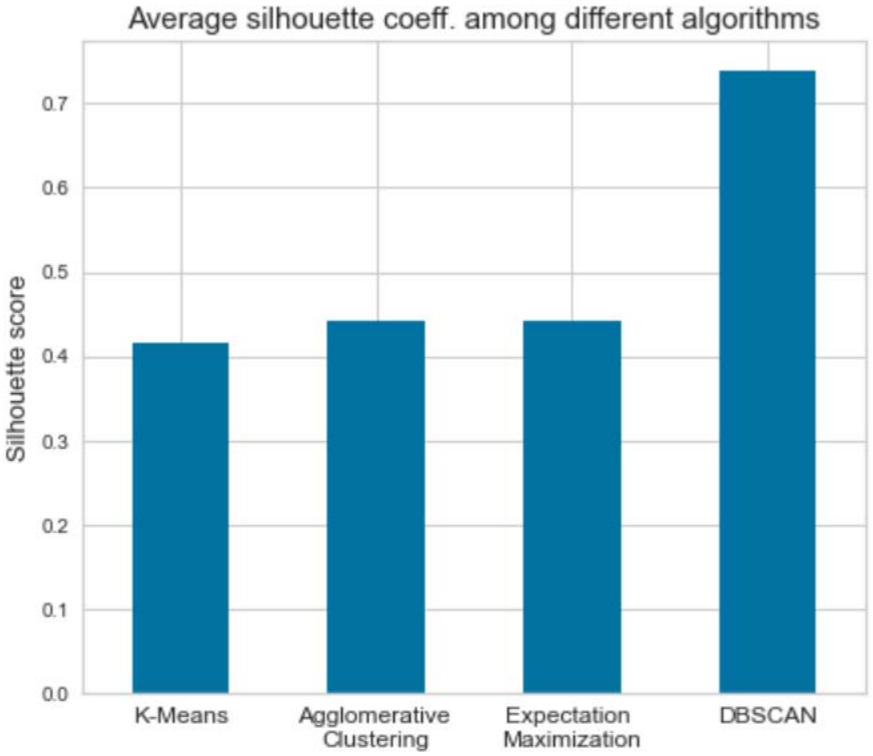


The Silhouette coefficient for the Agglomerative Hierarchical Clustering algorithm is 0.42



The Silhouette coefficient for the EM Clustering algorithm is 0.44




Average silhouette coeff. among different algorithms
Silhouette score
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
K-Means
Agglomerative Clustering
Expectation Maximization
DBSCAN

DBScan parameters

```
**************************************************
EPS 1.4 , minsample 2
The number of clusters detected by the DBSCAN algorithm is 3
The Silhouette coefficient for the DBSCAN algorithm is 0.77
**************************************************
**************************************************
EPS 1.4999999999999998 , minsample 2
The number of clusters detected by the DBSCAN algorithm is 3
The Silhouette coefficient for the DBSCAN algorithm is 0.77
**************************************************
**************************************************
EPS 1.5999999999999996 , minsample 2
The number of clusters detected by the DBSCAN algorithm is 3
The Silhouette coefficient for the DBSCAN algorithm is 0.77
**************************************************
**************************************************
EPS 1.6999999999999997 , minsample 2
The number of clusters detected by the DBSCAN algorithm is 3
The Silhouette coefficient for the DBSCAN algorithm is 0.77
**************************************************
**************************************************
EPS 1.7999999999999998 , minsample 2
The number of clusters detected by the DBSCAN algorithm is 3
The Silhouette coefficient for the DBSCAN algorithm is 0.77
**************************************************
**************************************************
EPS 1.8999999999999997 , minsample 2
The number of clusters detected by the DBSCAN algorithm is 3
The Silhouette coefficient for the DBSCAN algorithm is 0.77
**************************************************
**************************************************
EPS 1.9999999999999996 , minsample 2
The number of clusters detected by the DBSCAN algorithm is 3
The Silhouette coefficient for the DBSCAN algorithm is 0.77
**************************************************
**************************************************
EPS 2.0999999999999996 , minsample 2
The number of clusters detected by the DBSCAN algorithm is 3
The Silhouette coefficient for the DBSCAN algorithm is 0.77
**************************************************
**************************************************
EPS 2.1999999999999997, minsample 2
The number of clusters detected by the DBSCAN algorithm is 3
The Silhouette coefficient for the DBSCAN algorithm is 0.77
**************************************************
**************************************************
EPS 2.3 , minsample 2
The number of clusters detected by the DBSCAN algorithm is 3
The Silhouette coefficient for the DBSCAN algorithm is 0.77
**************************************************
**************************************************
EPS 2.3999999999999995 , minsample 2
The number of clusters detected by the DBSCAN algorithm is 3
The Silhouette coefficient for the DBSCAN algorithm is 0.77
**************************************************
**************************************************
```

```
EPS 2.4999999999999996 , minsample 2
The number of clusters detected by the DBSCAN algorithm is 3
The Silhouette coefficient for the DBSCAN algorithm is 0.77
**************************************************
**************************************************
EPS 2.5999999999999996 , minsample 2
The number of clusters detected by the DBSCAN algorithm is 3
The Silhouette coefficient for the DBSCAN algorithm is 0.77
**************************************************
**************************************************
EPS 2.6999999999999993 , minsample 2
The number of clusters detected by the DBSCAN algorithm is 3
The Silhouette coefficient for the DBSCAN algorithm is 0.77
**************************************************
**************************************************
EPS 2.7999999999999994 , minsample 2
The number of clusters detected by the DBSCAN algorithm is 3
The Silhouette coefficient for the DBSCAN algorithm is 0.77
**************************************************
**************************************************
EPS 2.8999999999999995 , minsample 2
The number of clusters detected by the DBSCAN algorithm is 3
The Silhouette coefficient for the DBSCAN algorithm is 0.77
**************************************************
**************************************************
EPS 4.1 , minsample 1
The number of clusters detected by the DBSCAN algorithm is 3
The Silhouette coefficient for the DBSCAN algorithm is 0.86
**************************************************
**************************************************
EPS 4.199999999999999 , minsample 1
The number of clusters detected by the DBSCAN algorithm is 3
The Silhouette coefficient for the DBSCAN algorithm is 0.86
**************************************************
**************************************************
EPS 4.299999999999999 , minsample 1
The number of clusters detected by the DBSCAN algorithm is 3
The Silhouette coefficient for the DBSCAN algorithm is 0.86
**************************************************
**************************************************
EPS 4.399999999999999 , minsample 1
The number of clusters detected by the DBSCAN algorithm is 3
The Silhouette coefficient for the DBSCAN algorithm is 0.86
**************************************************
```