Online shoppers clustering

Hubert Ruczyński

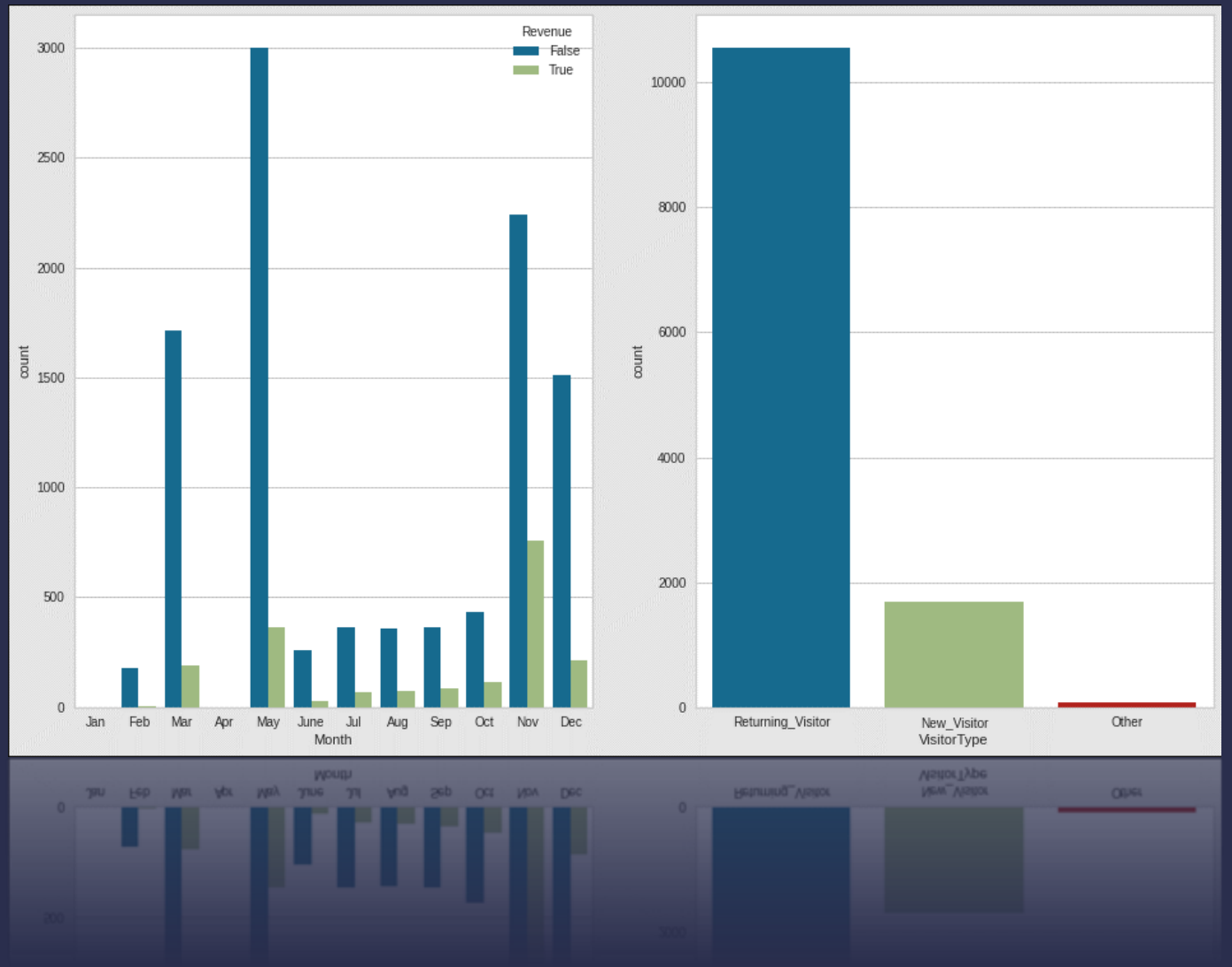
Mateusz Sperkowski

Prosty opis danych

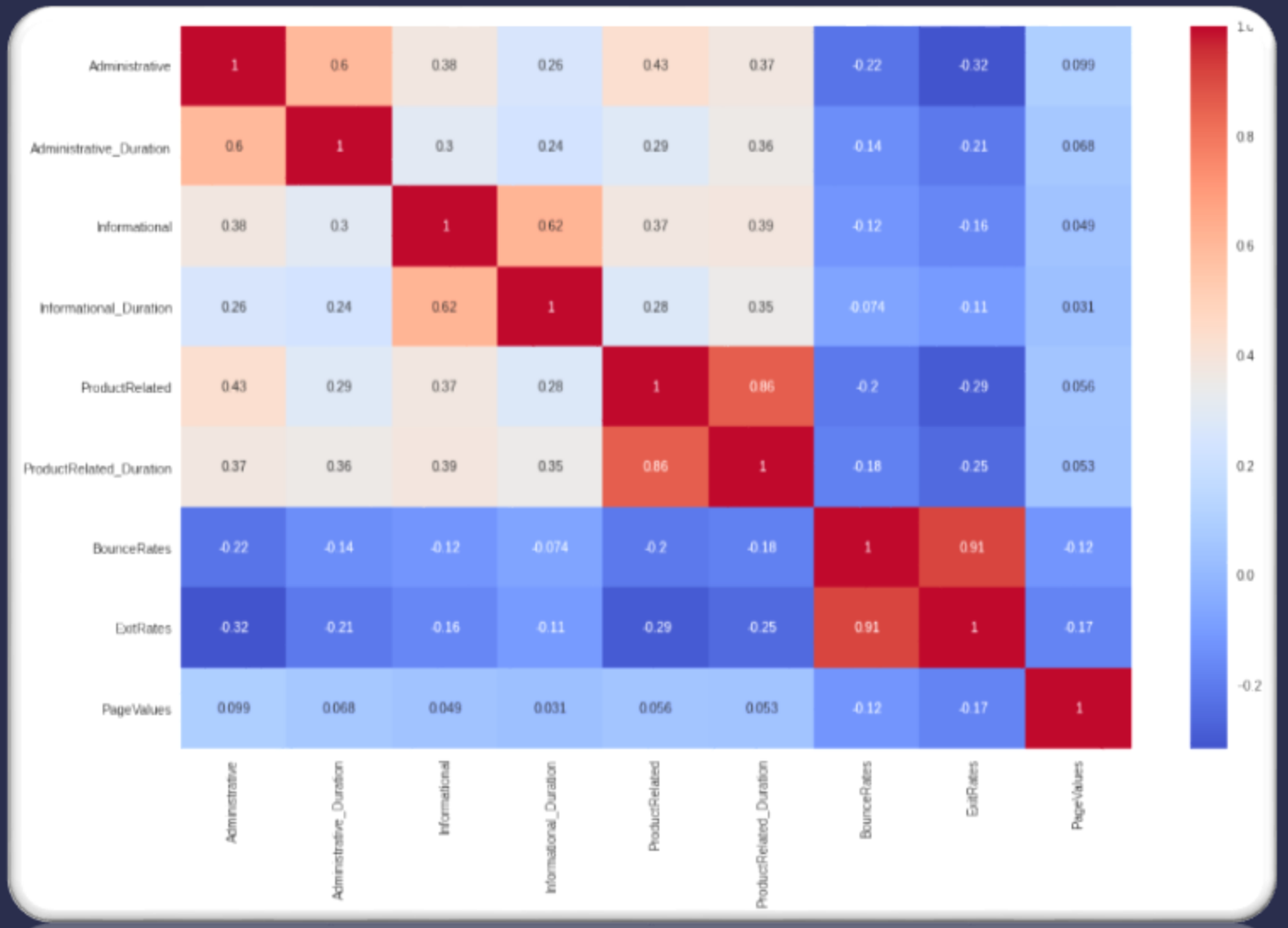
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 12330 entries, 0 to 12329
Data columns (total 18 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   Administrative           12330 non-null  int64
 1   Administrative_Duration  12330 non-null  float64
 2   Informational            12330 non-null  int64
 3   Informational_Duration   12330 non-null  float64
 4   ProductRelated           12330 non-null  int64
 5   ProductRelated_Duration  12330 non-null  float64
 6   BounceRates              12330 non-null  float64
 7   ExitRates                12330 non-null  float64
 8   PageValues               12330 non-null  float64
 9   SpecialDay               12330 non-null  float64
 10  Month                    12330 non-null  object
 11  OperatingSystems         12330 non-null  int64
 12  Browser                  12330 non-null  int64
 13  Region                   12330 non-null  int64
 14  TrafficType              12330 non-null  int64
 15  VisitorType              12330 non-null  object
 16  Weekend                  12330 non-null  bool
 17  Revenue                  12330 non-null  bool
dtypes: bool(2), float64(7), int64(7), object(2)
memory usage: 1.5+ MB
```

	Administrative	Administrative_Duration	Informational	Informational_Duration	ProductRelated	ProductRelated_Duration	BounceRates	ExitRates	PageValues	SpecialDay	OperatingSystems	Browser	Region	TrafficType
count	12330.000000	12330.000000	12330.000000	12330.000000	12330.000000	12330.000000	12330.000000	12330.000000	12330.000000	12330.000000	12330.000000	12330.000000	12330.000000	12330.000000
mean	2.315166	80.818611	0.503569	34.472398	31.731468	1194.746220	0.022191	0.043073	5.889258	0.061427	2.124006	2.357097	3.147364	4.069586
std	3.321784	176.779107	1.270156	140.749294	44.475503	1913.669288	0.048488	0.048597	18.568437	0.198917	0.911325	1.717277	2.401591	4.025169
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000	1.000000	1.000000	1.000000
25%	0.000000	0.000000	0.000000	0.000000	7.000000	184.137500	0.000000	0.014286	0.000000	0.000000	2.000000	2.000000	1.000000	2.000000
50%	1.000000	7.500000	0.000000	0.000000	18.000000	598.936905	0.003112	0.025156	0.000000	0.000000	2.000000	2.000000	3.000000	2.000000
75%	4.000000	93.256250	0.000000	0.000000	38.000000	1464.157213	0.016813	0.050000	0.000000	0.000000	3.000000	2.000000	4.000000	4.000000
max	27.000000	3398.750000	24.000000	2549.375000	705.000000	63973.522230	0.200000	0.200000	361.763742	1.000000	8.000000	13.000000	9.000000	20.000000

EDA



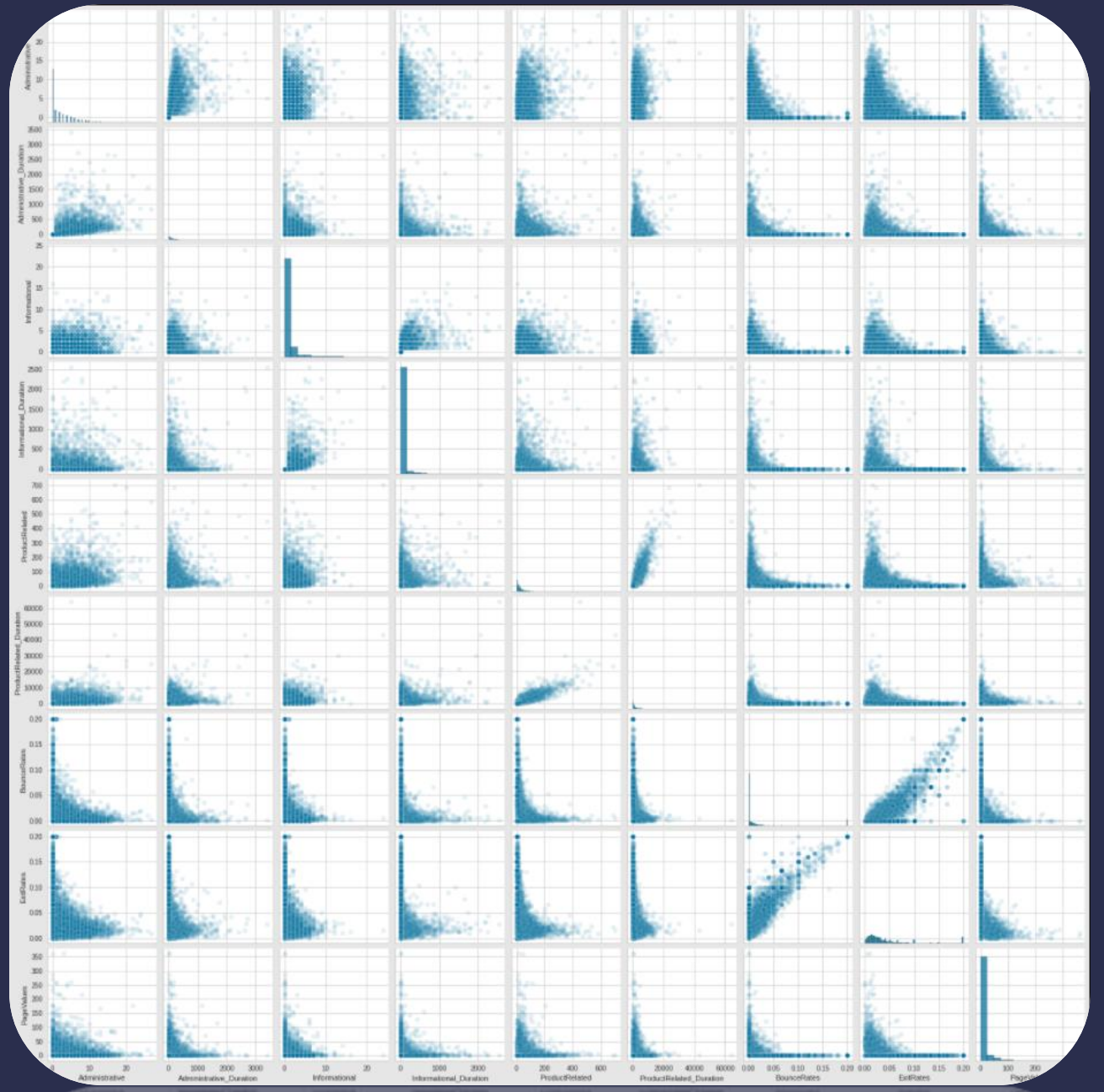
Miesięczna aktywność oraz typy odwiedzających

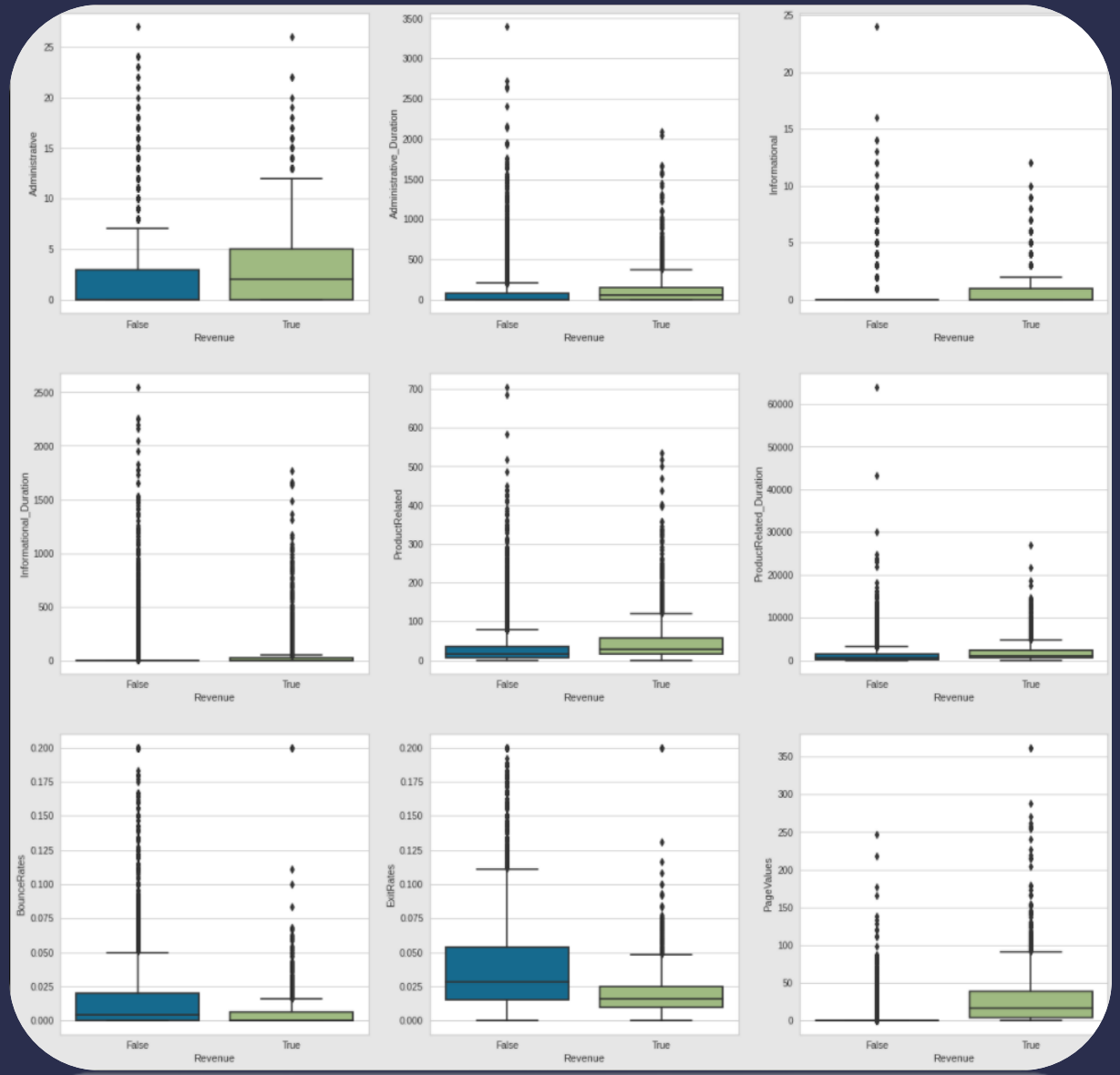


Macierz Korelacji Zmiennych Numerycznych

Macierz Korelacji Zmiennych

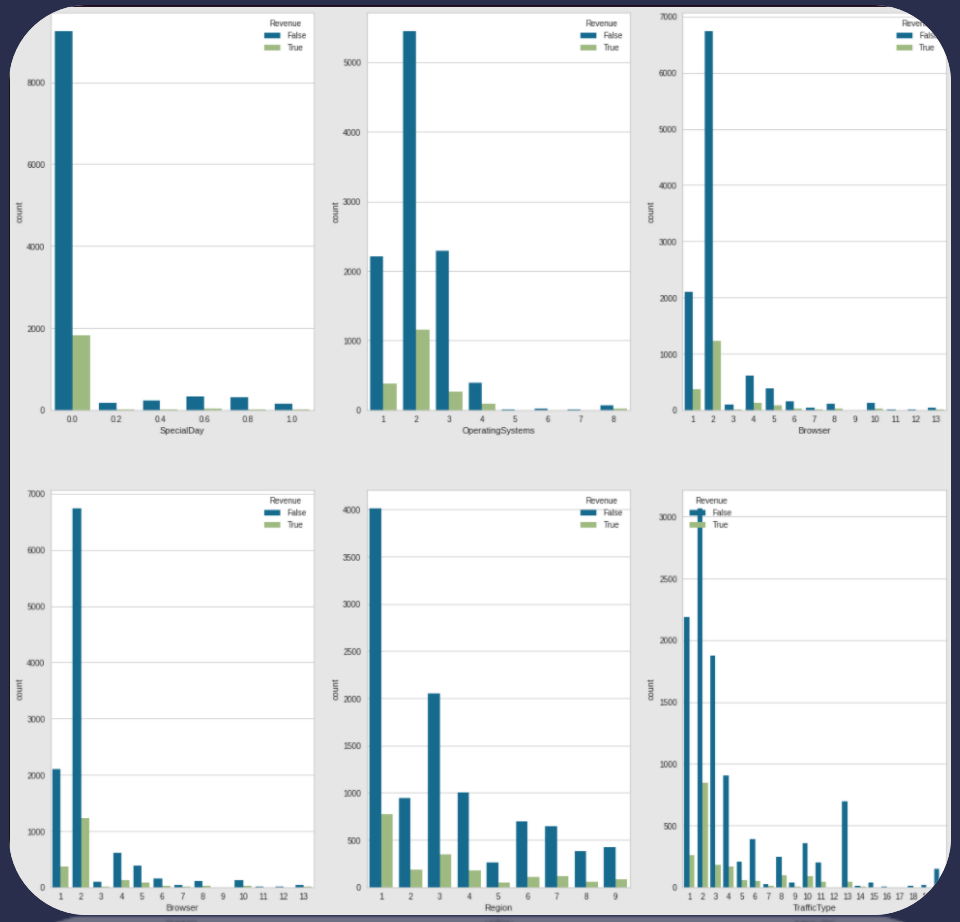
Scatterploty




Administrative
Administrative_Duration
Informational
Informational_Duration
ProductRelated
ProductRelated_Duration
BounceRates
ExitRates
PageValues
False
True
Revenue

Boxploty

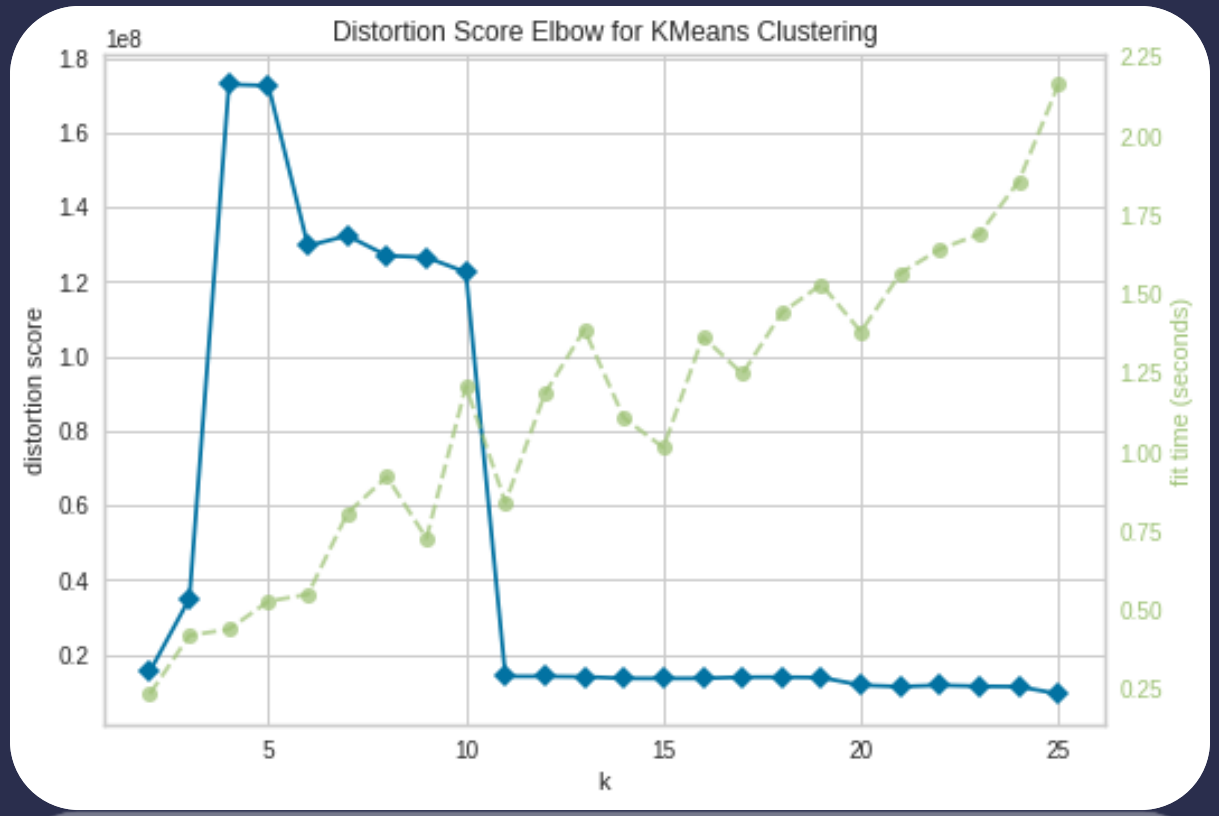
Barploty

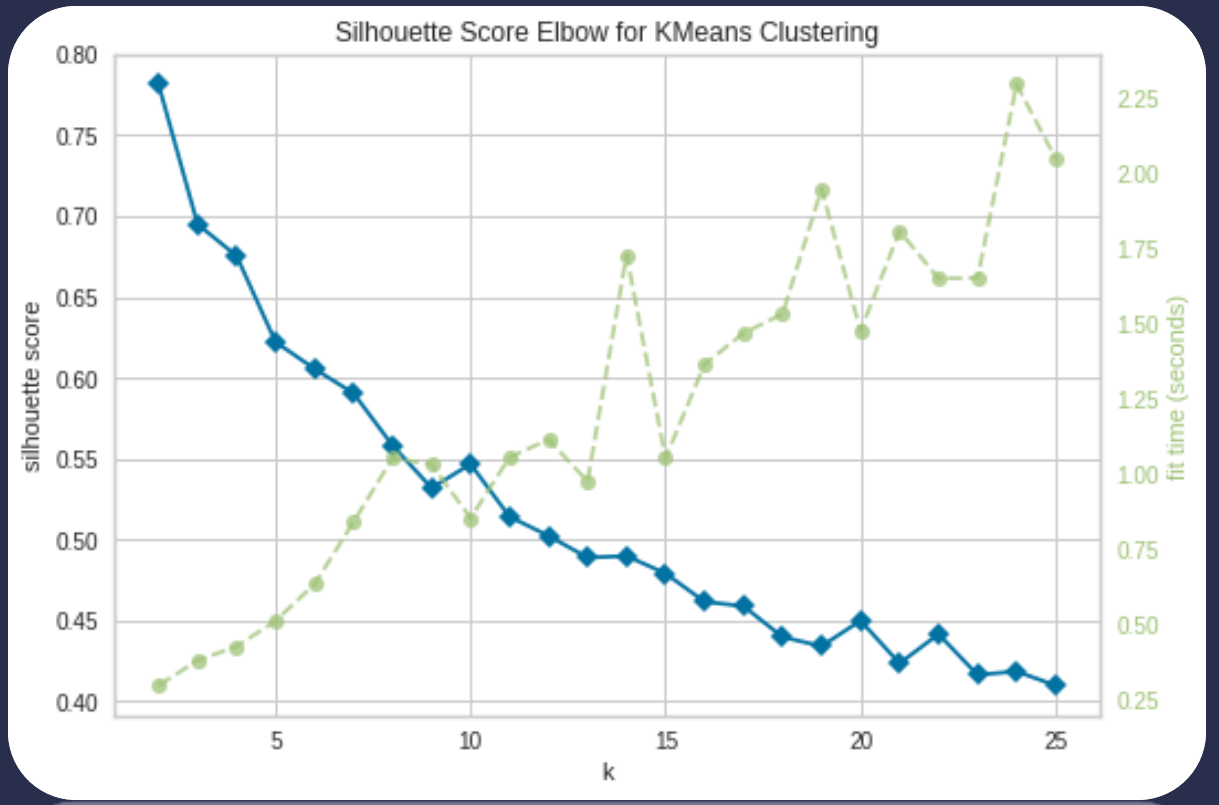

Revenue
False
True
count
SpecialDay
OperatingSystems
Browser
Region
TrafficType

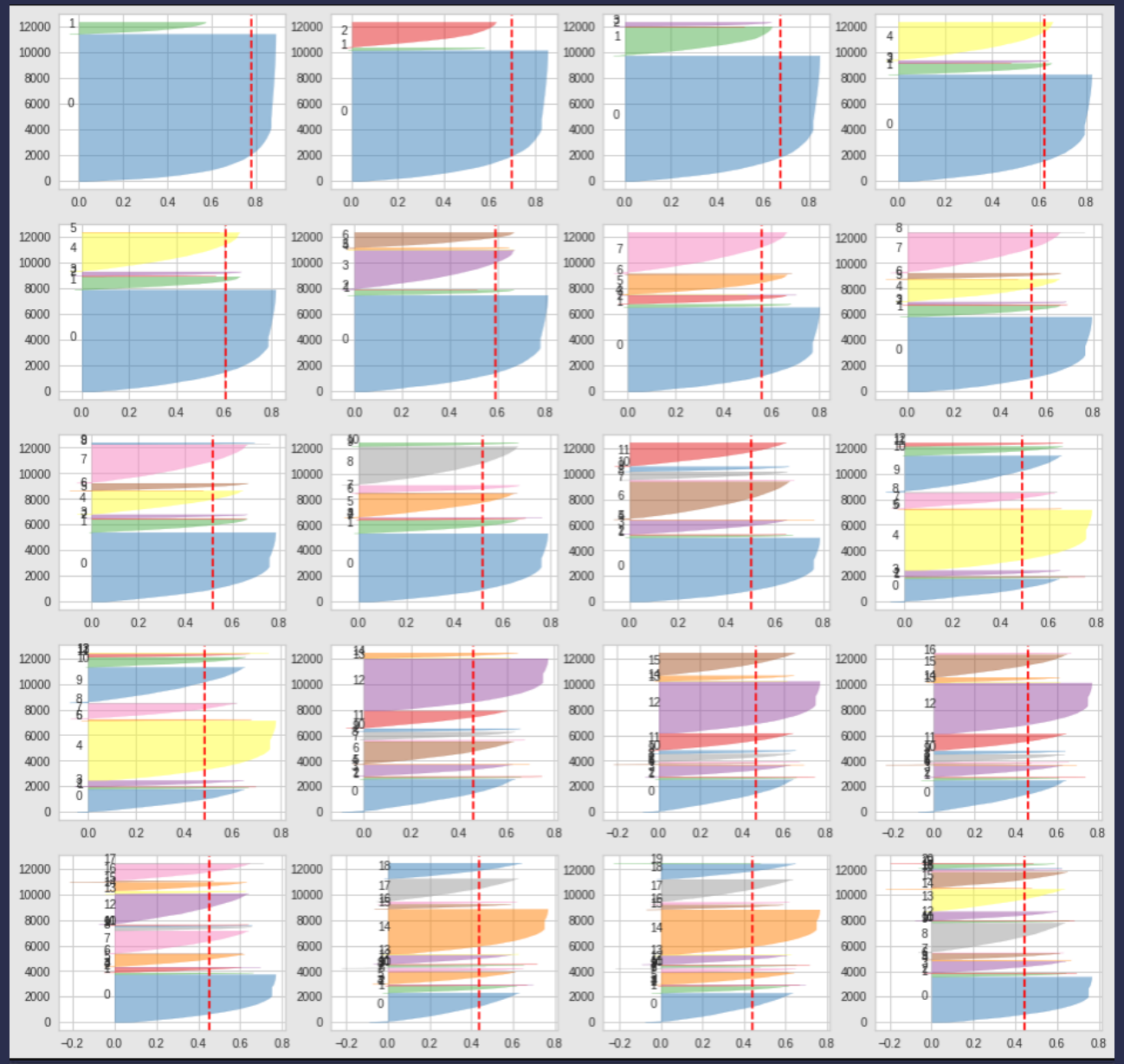
Revenue
False
True
count
SpecialDay
OperatingSystems
Browser
Region
TrafficType

Wybór liczby klastrów

Silhouette i Distortion scores



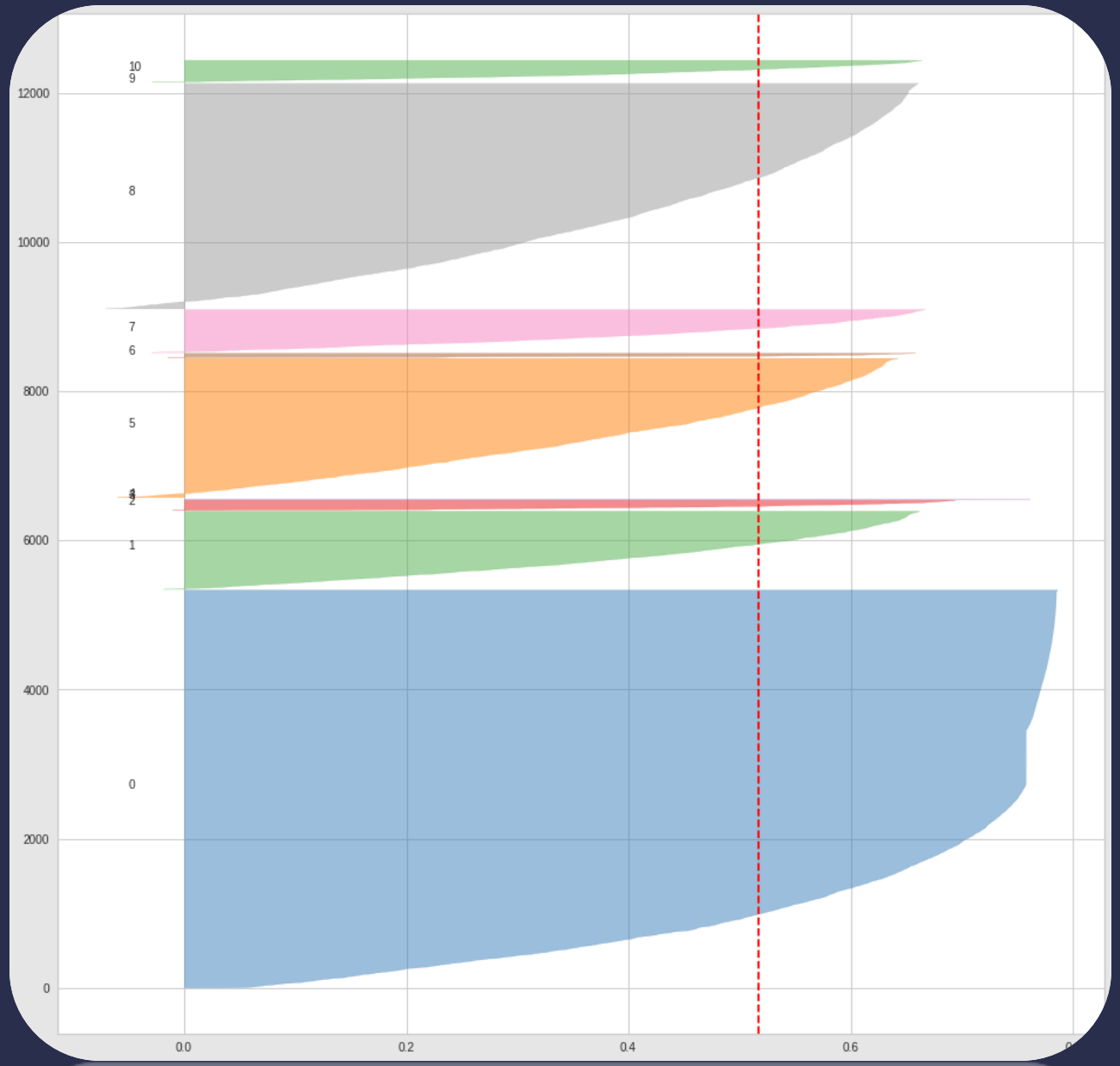


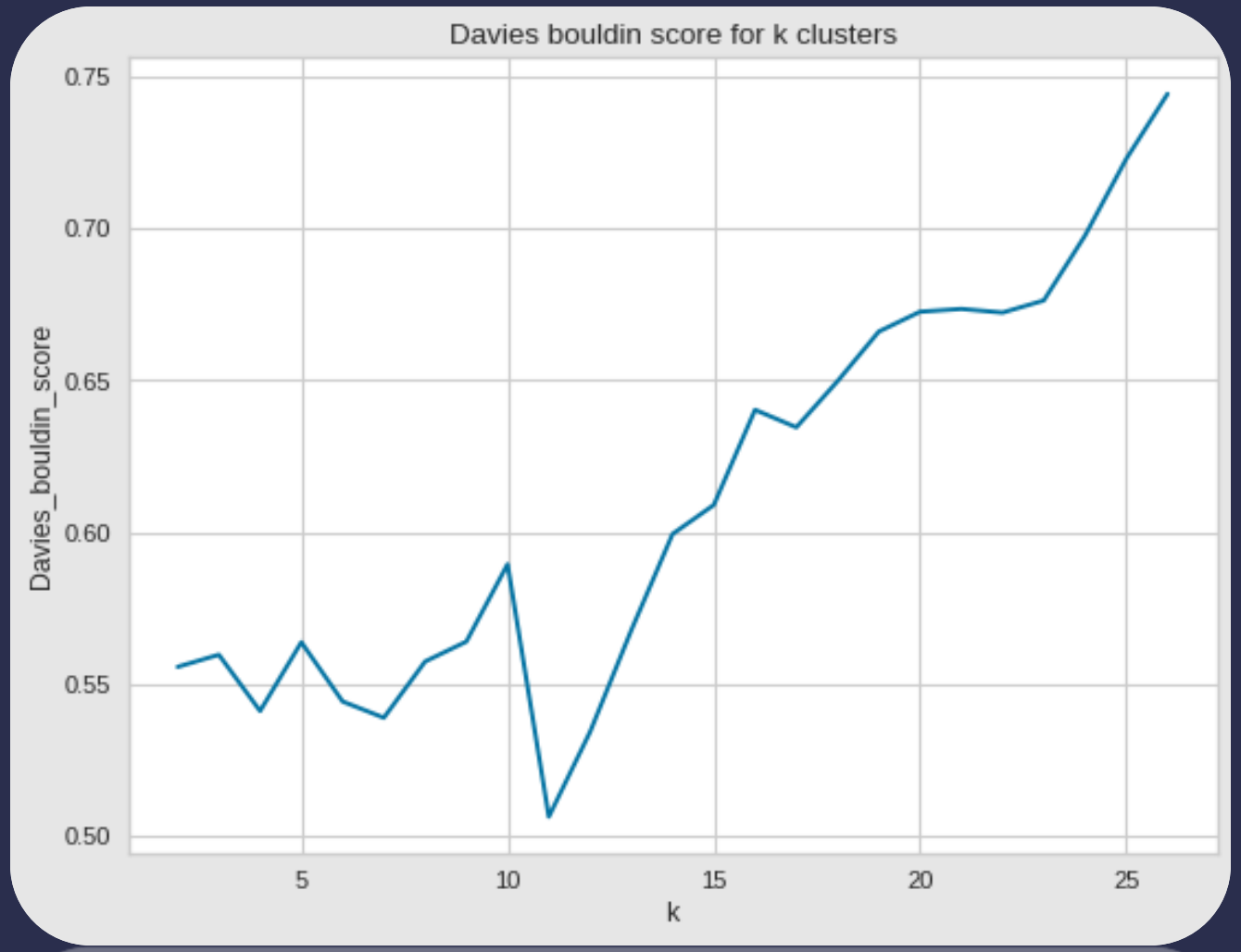


Analiza Silhouette'a

Silhouette Analysis vs Elbow Method vs Davies-Bouldin Index: Selecting the optimal number of clusters for KMeans clustering

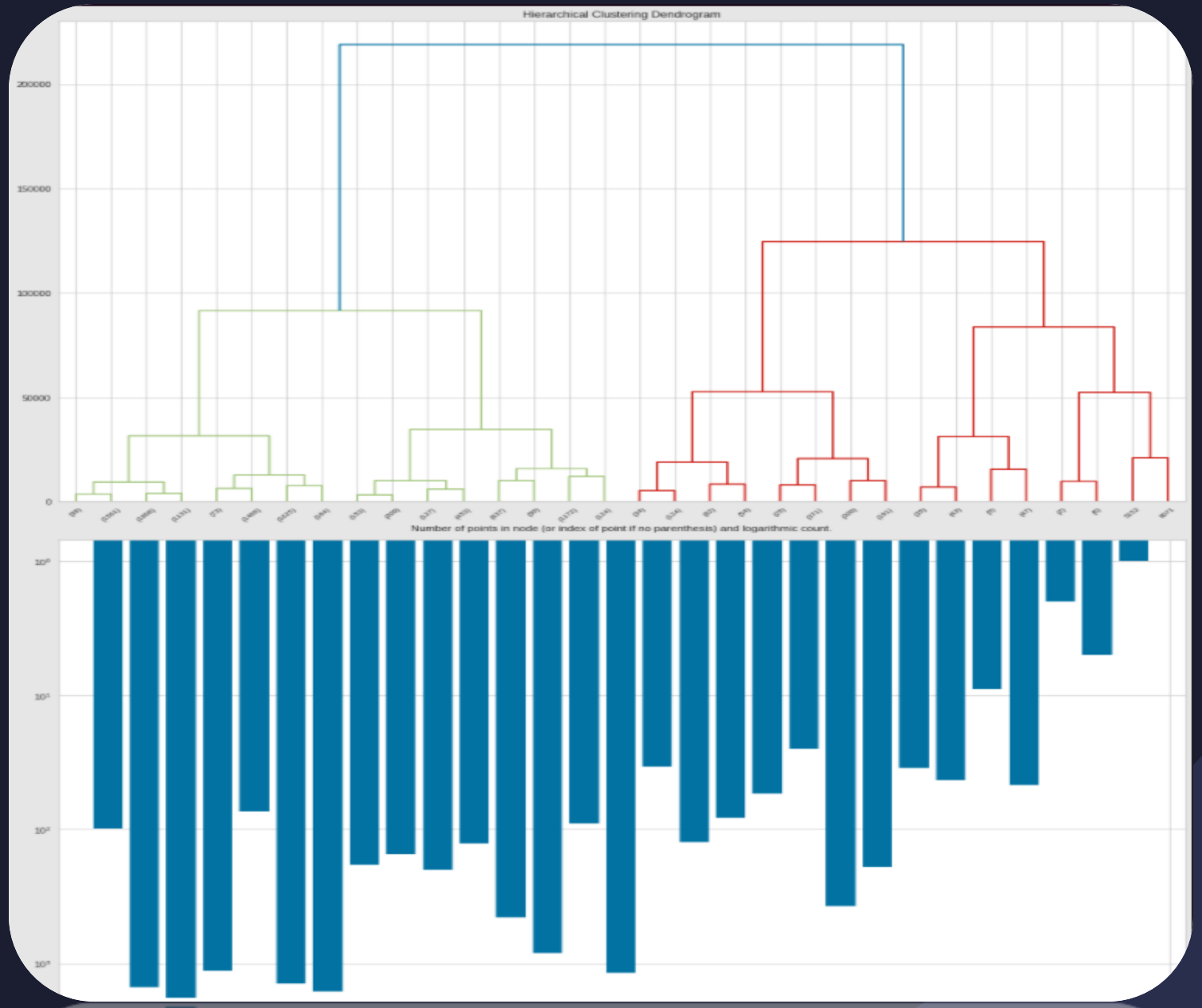
Analiza Silhouette’a dla 11 klastrów





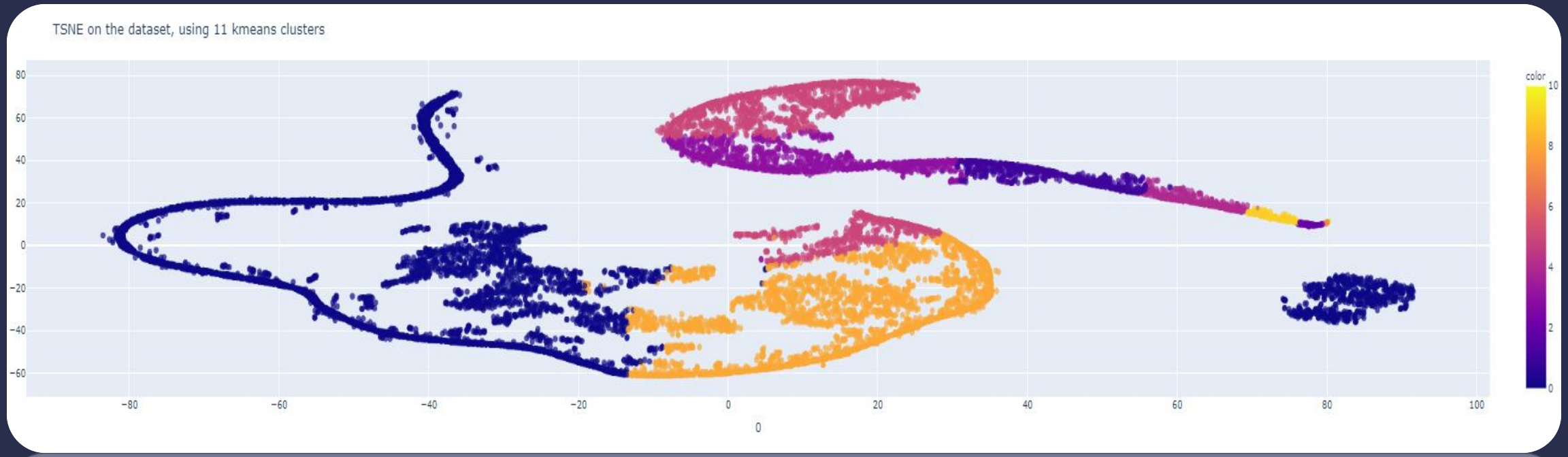
Davies-Bouldin Score

Hierarchiczny Dendrogram aglomeracyjnego klastrowania

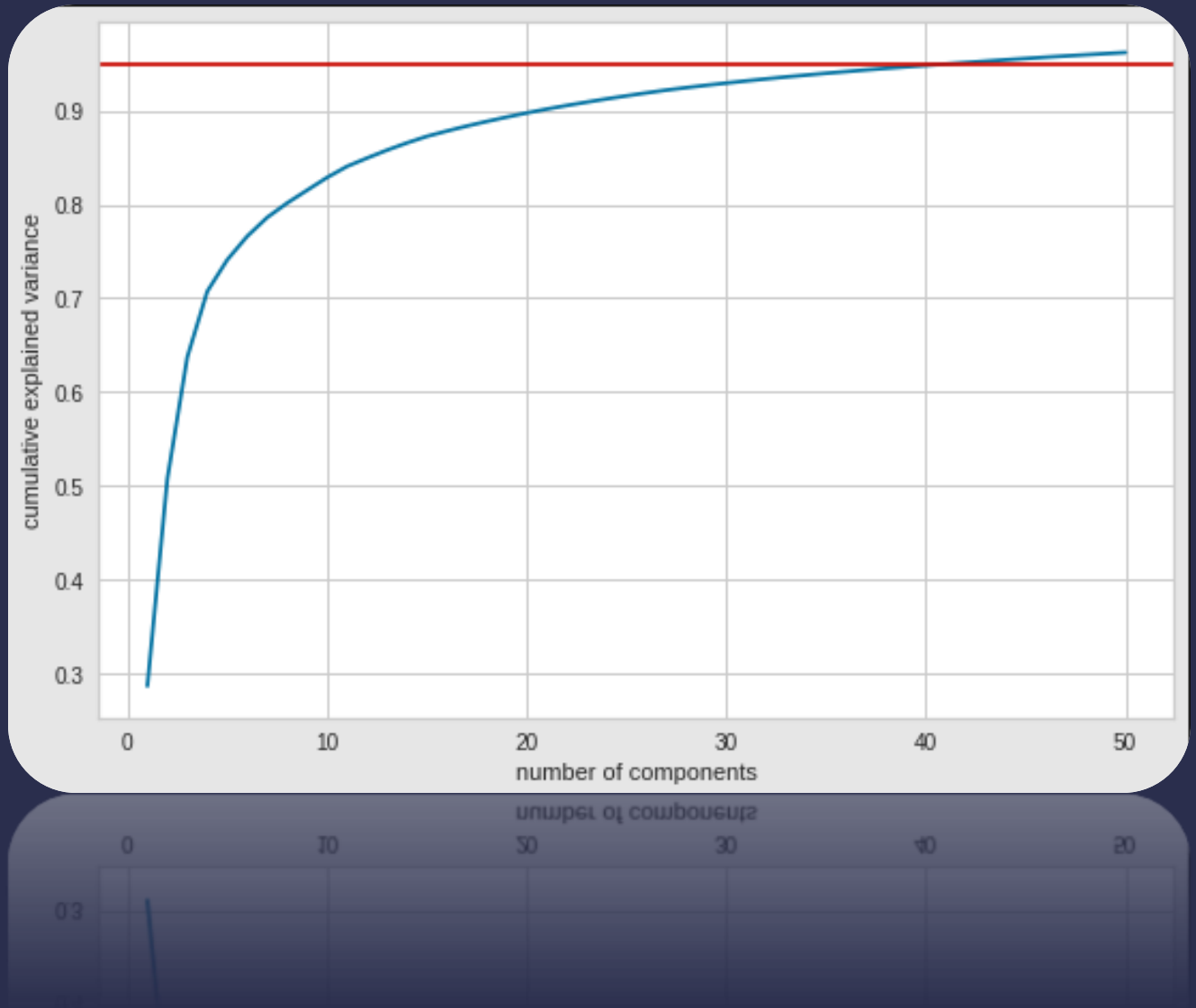


Klastry

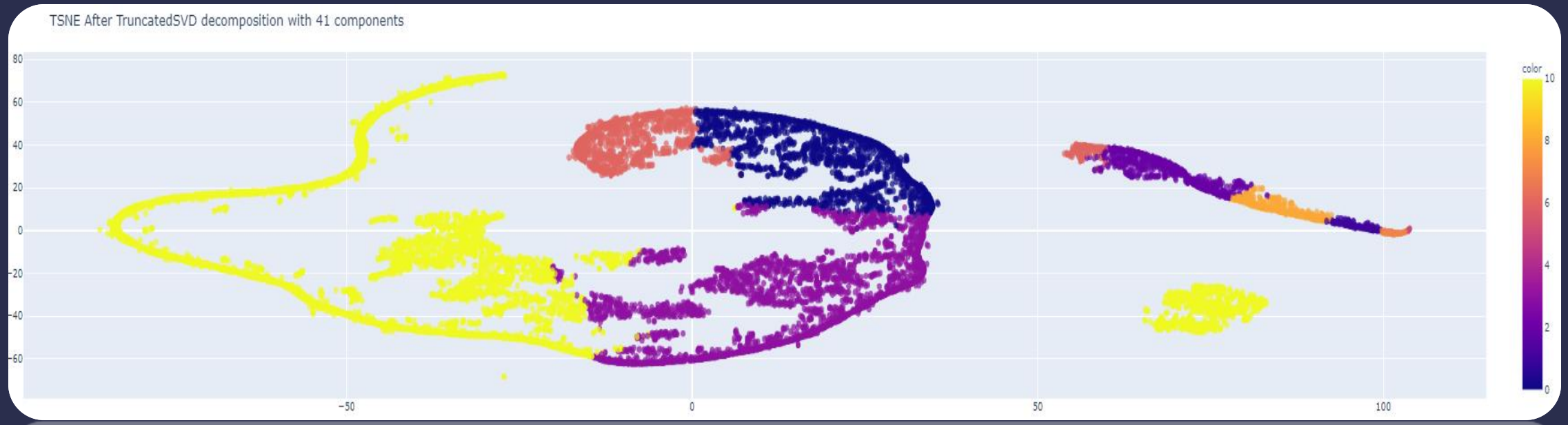
Klastrowanie z 11 klastrami



Wymagana liczba kolumn z PCA



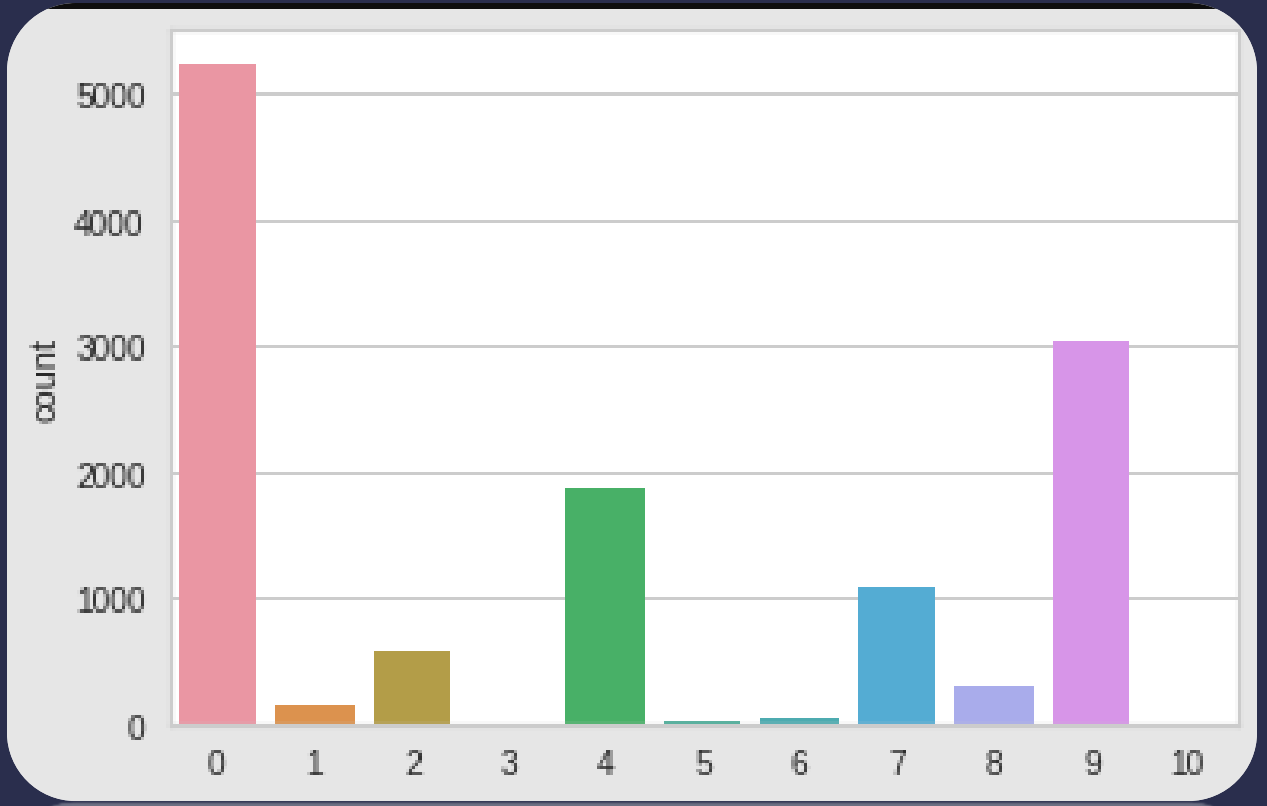
Klastrowanie z 11 klastrami i 41 komponentami

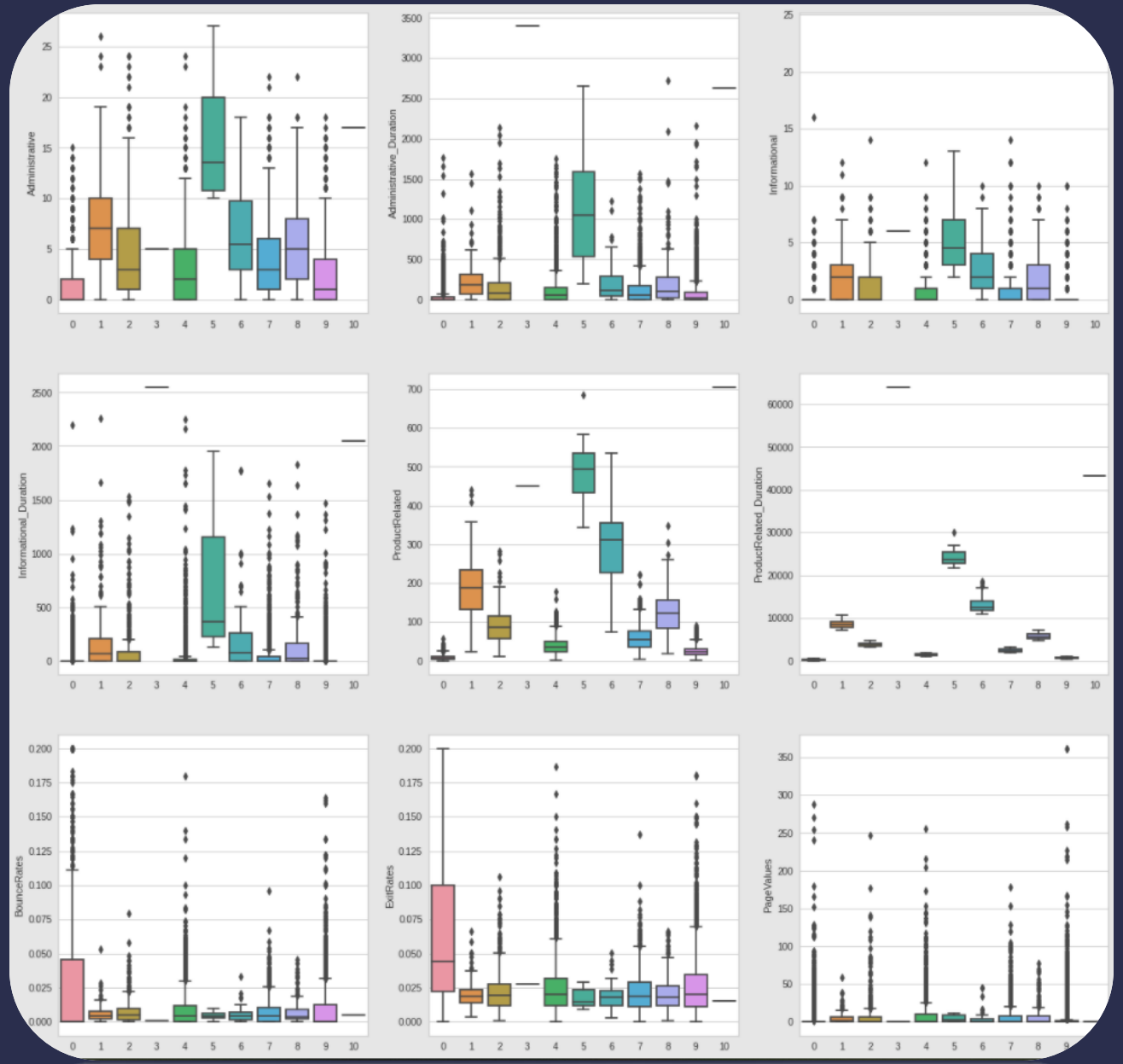


Analiza klastrów

Wyróżnianie użytkowników płacących oraz weekendowych

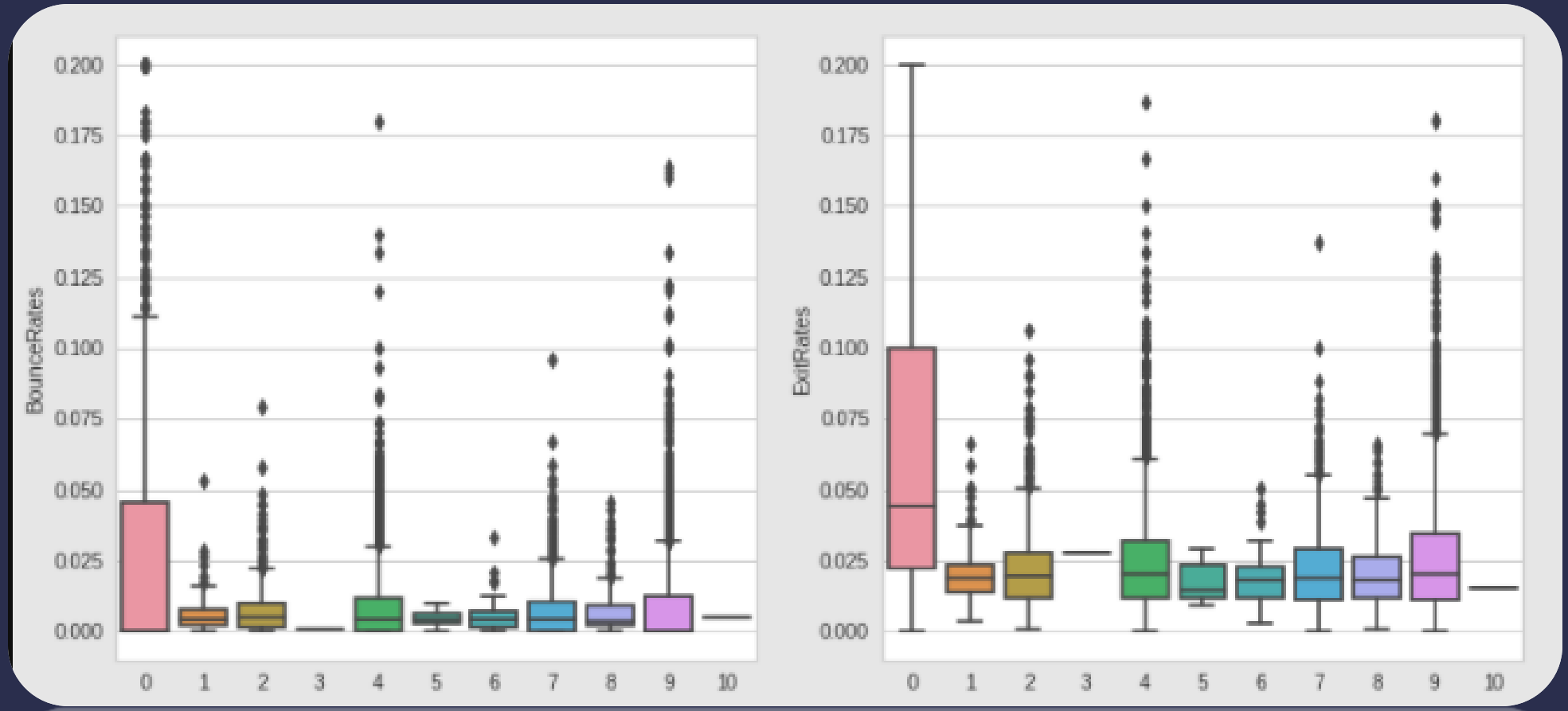
	cluster	count	revenue_fraction	weekend_fraction
0	0	5242	0.071728	0.218237
1	1	137	0.313869	0.233577
2	2	588	0.250000	0.236395
3	3	1	0.000000	0.000000
4	4	1875	0.208000	0.260267
5	5	8	0.250000	0.000000
6	6	58	0.413793	0.293103
7	7	1091	0.227314	0.205316
8	8	302	0.344371	0.238411
9	9	3027	0.189627	0.248100
10	10	1	0.000000	1.000000





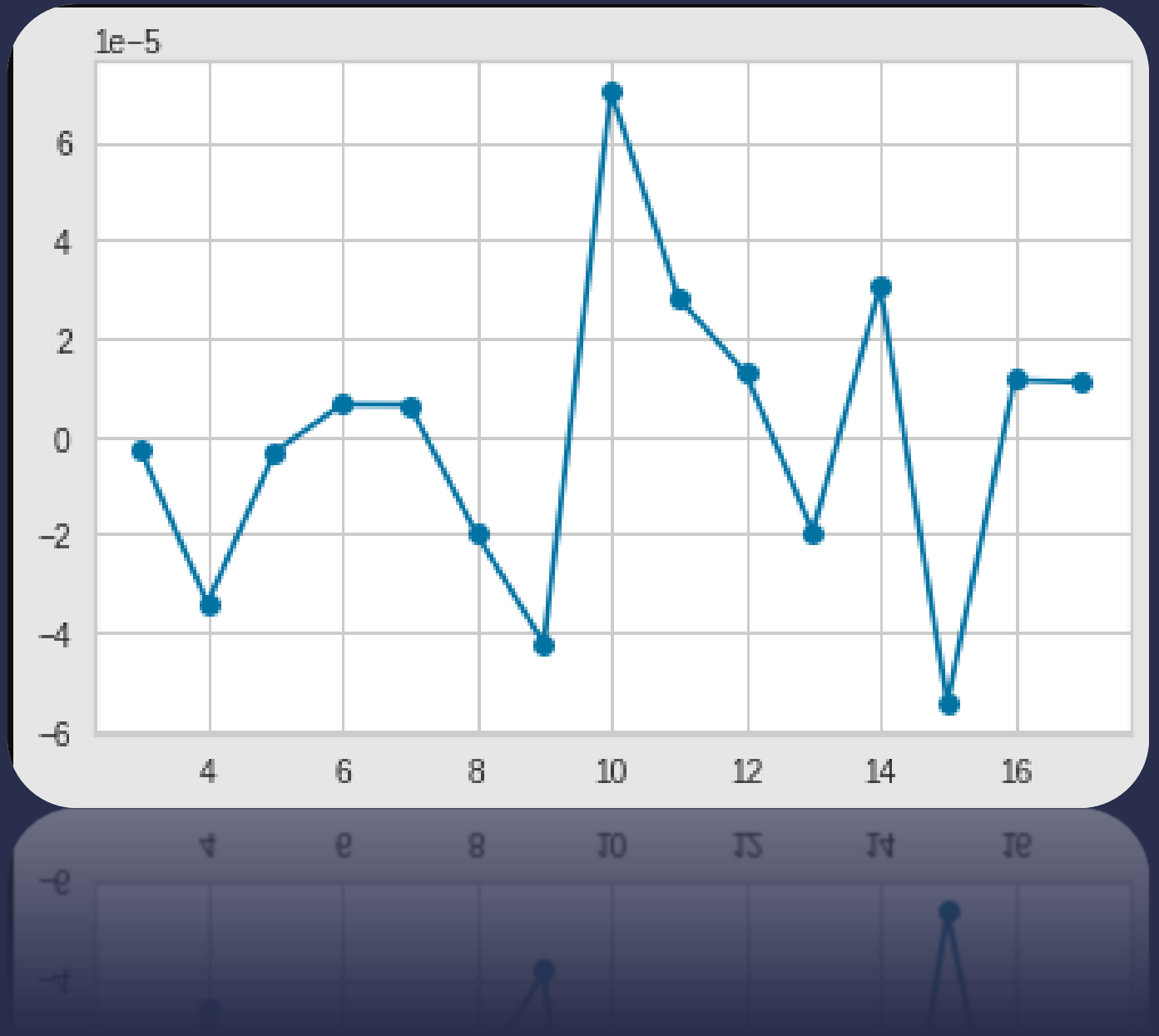
Analiza wykresami skrzynkowymi

Klaster przelotnych odbiorców

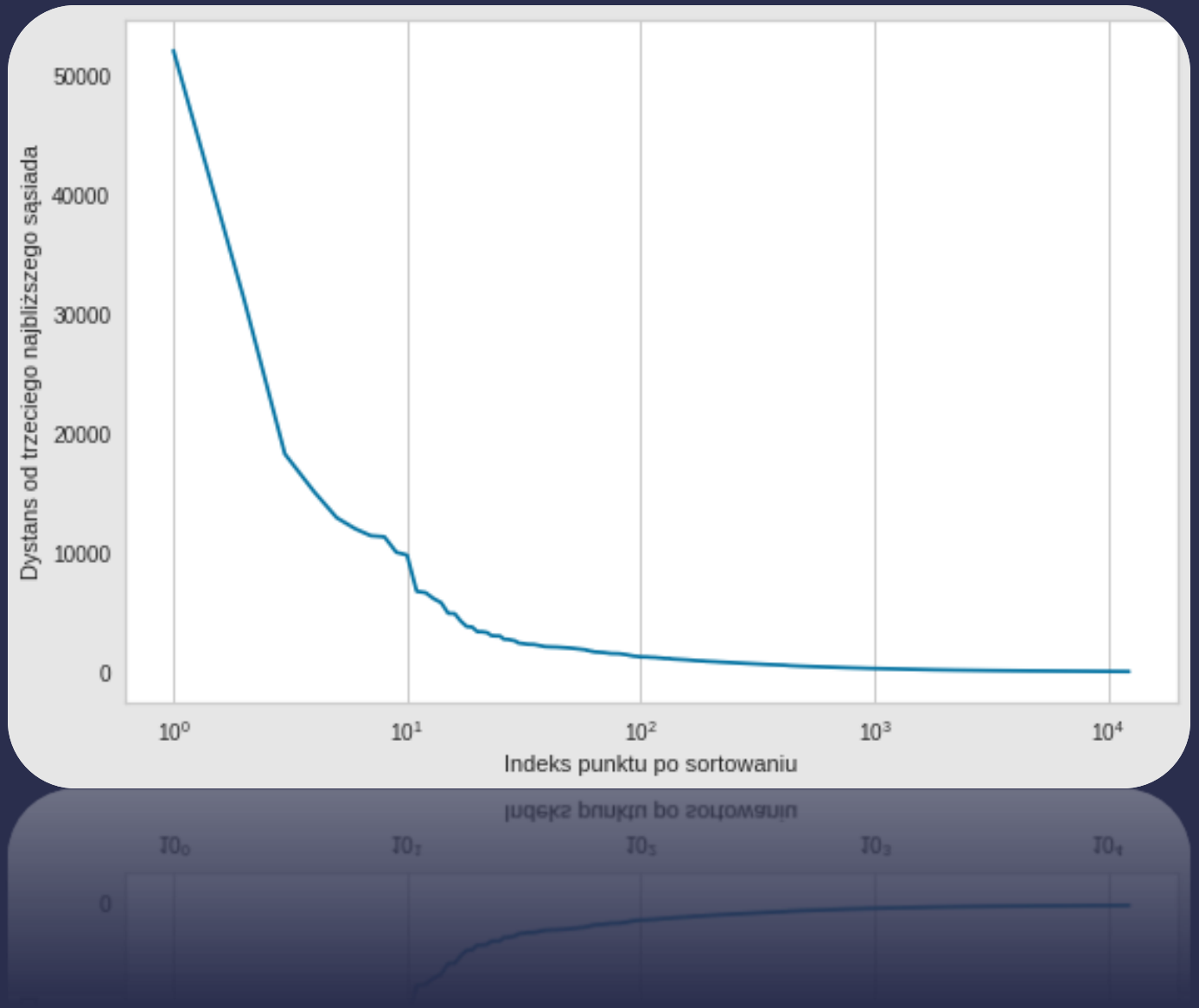


Niewypały

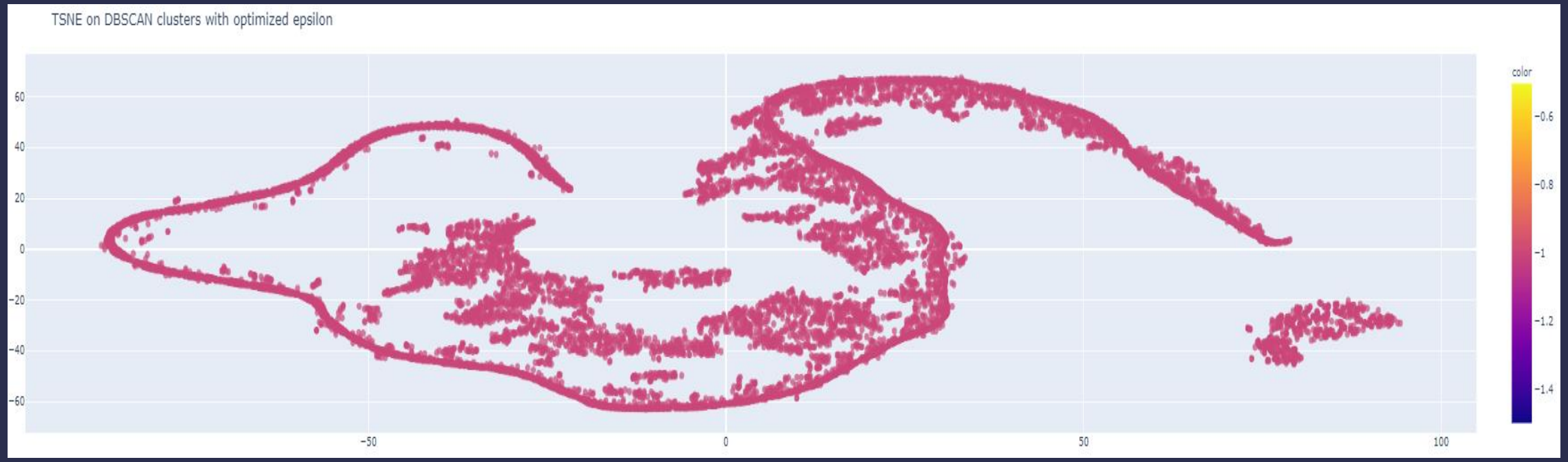
Stabilność



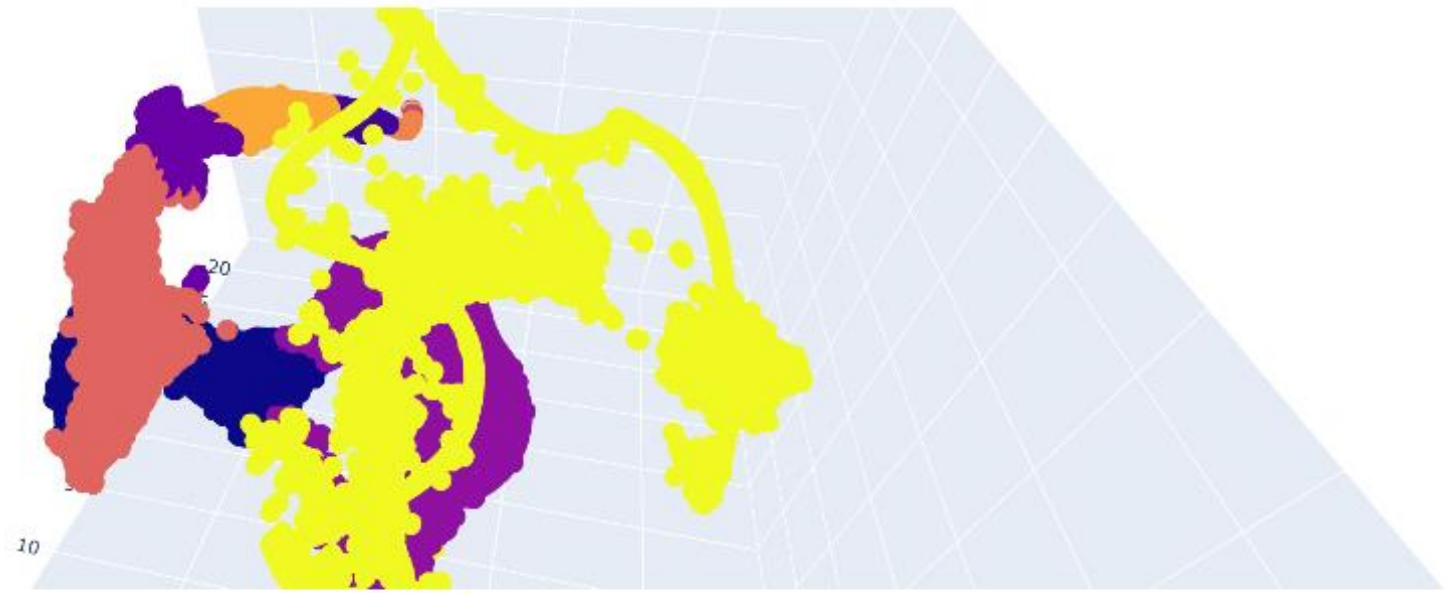
Wstęp do DBSCAN



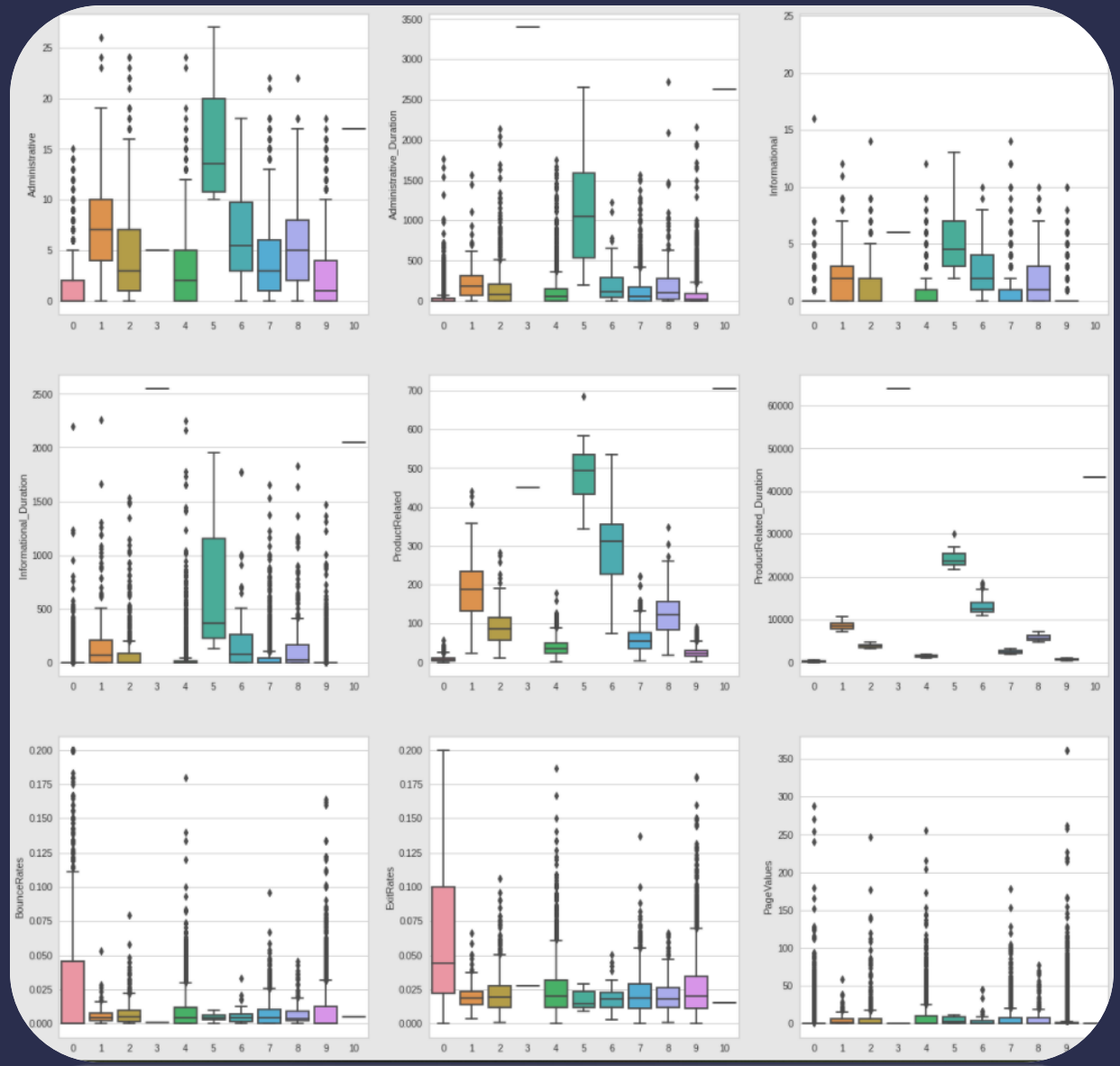
DBSCAN

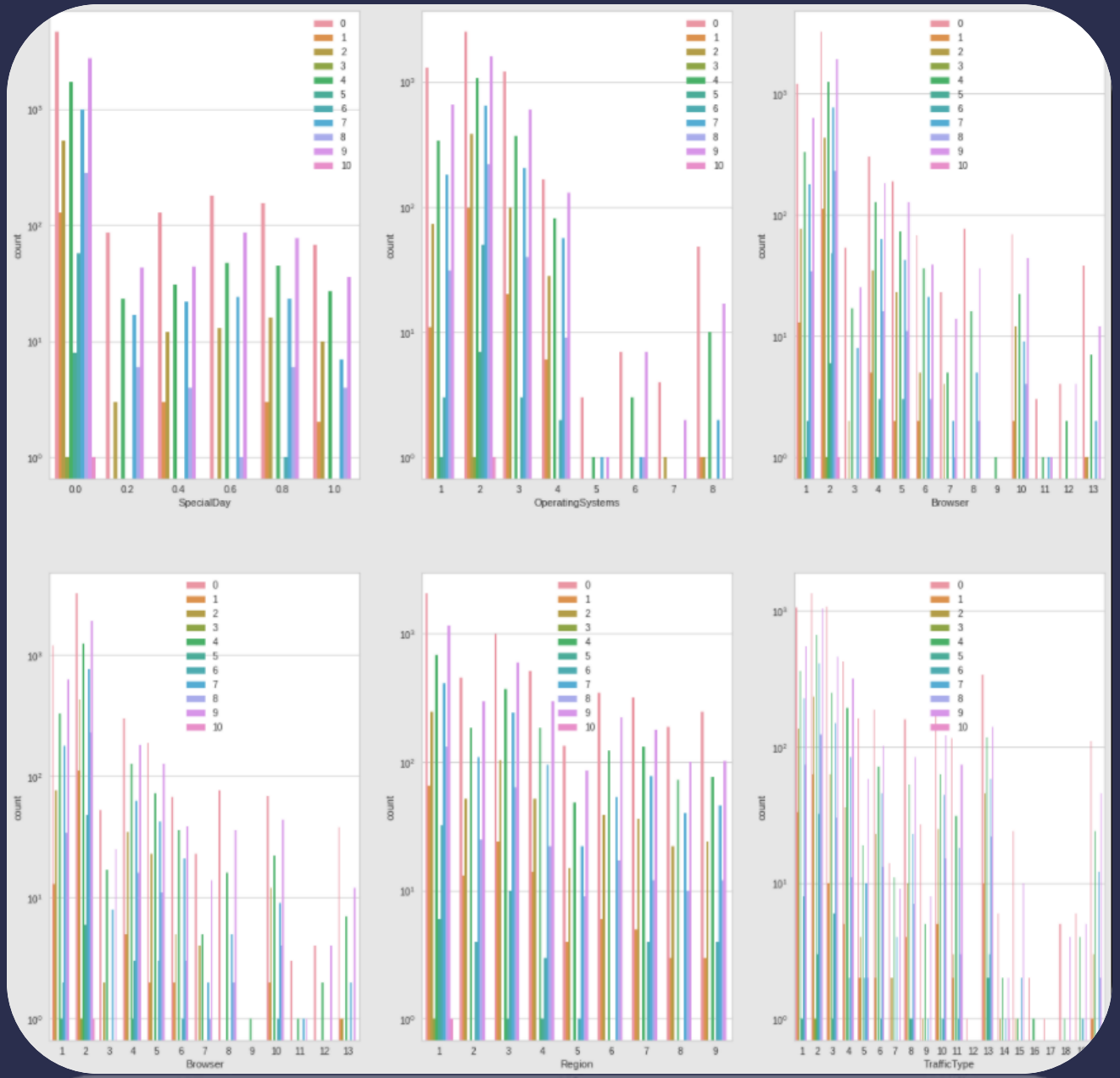

TSNE on DBSCAN clusters with optimized epsilon
color
60
40
20
0
−20
−40
−60
−50
0
50
100
−0.6
−0.8
−1
−1.2
−1.4

TSNE 3D



Analiza wykresami skrzynkowymi

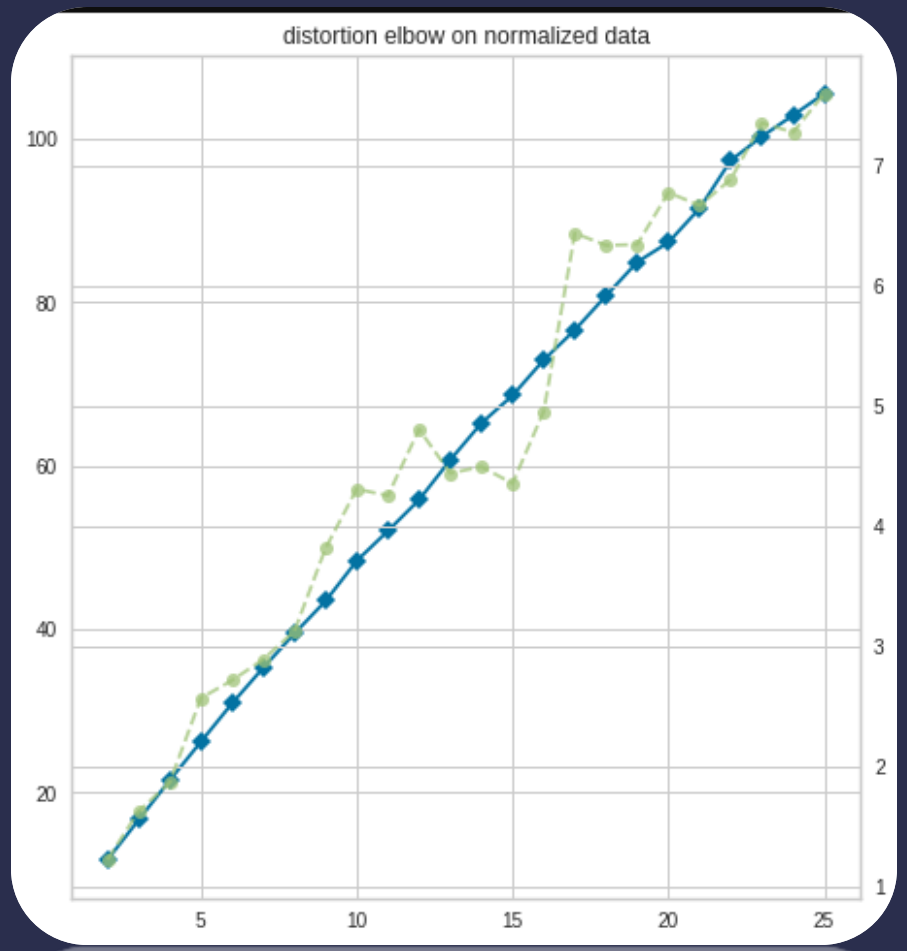


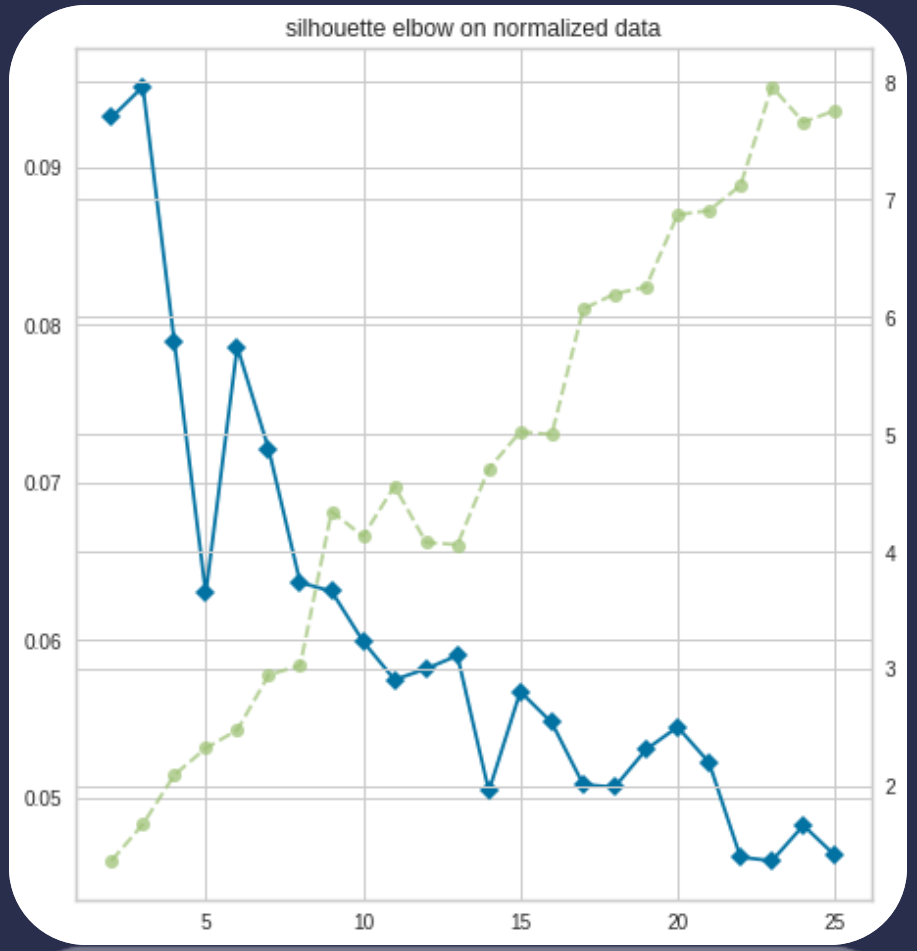


Analiza zmiennych kategorycznych po klastrach

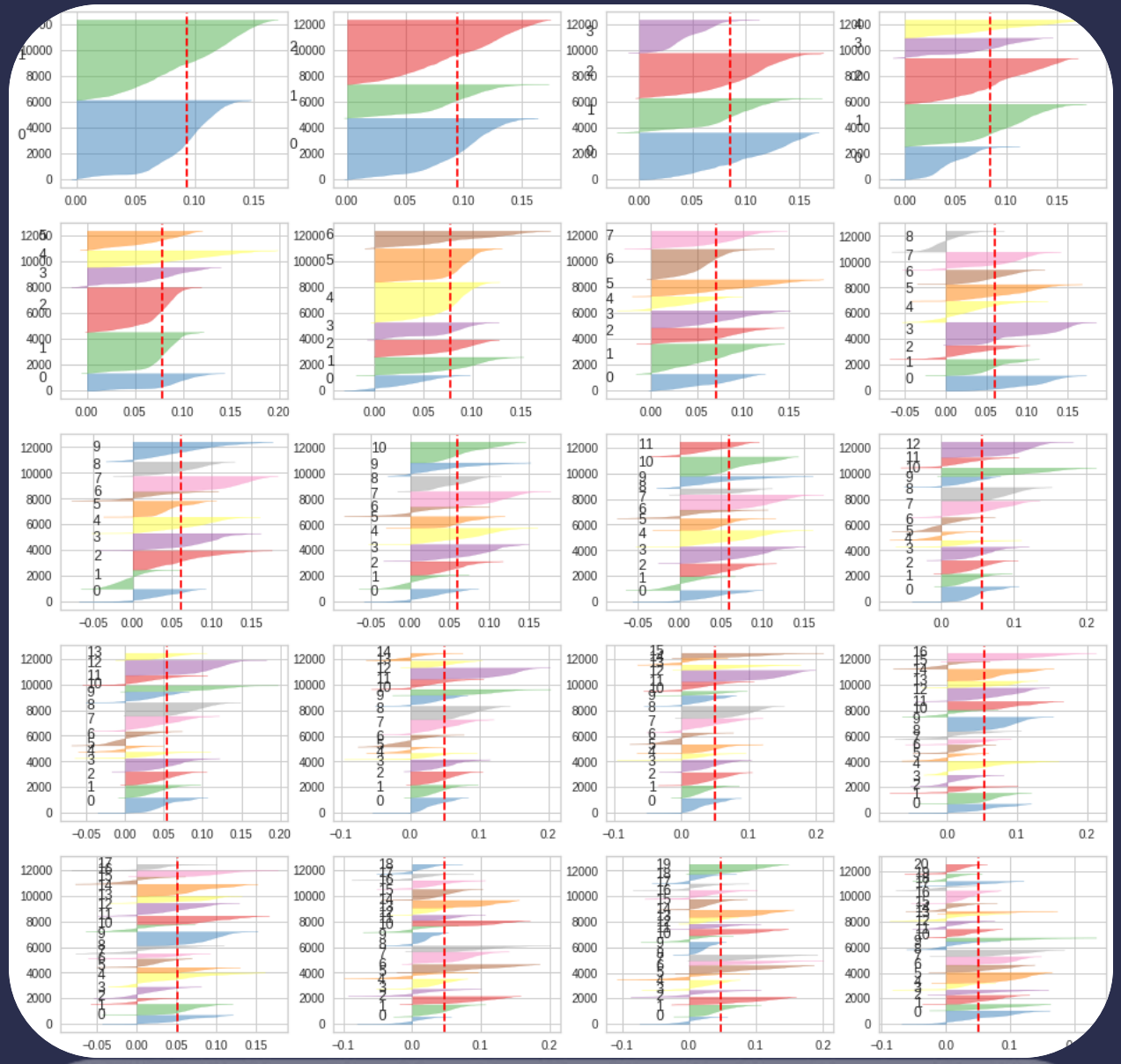
Normalizacja

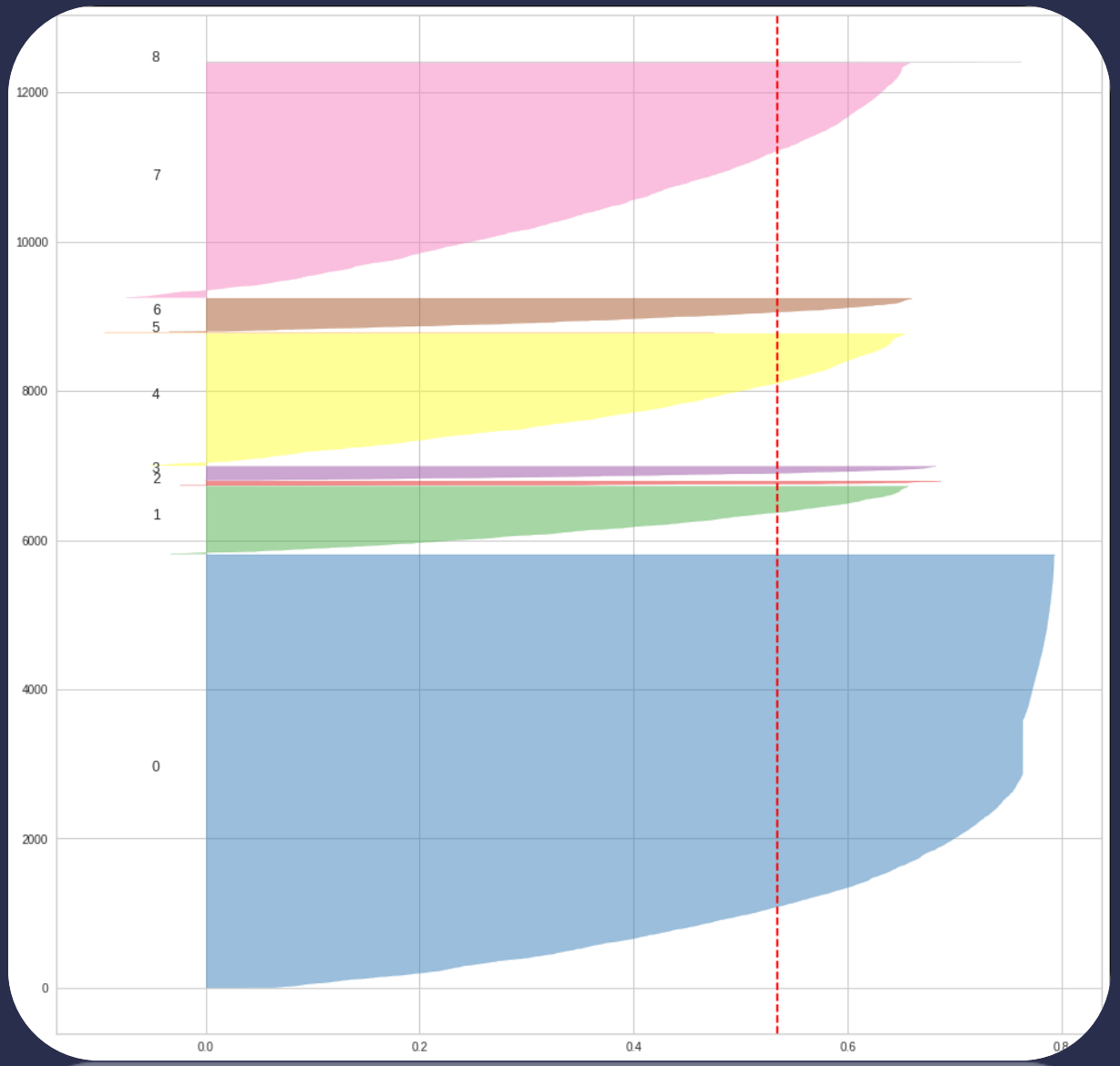
Wykresy łokciowe





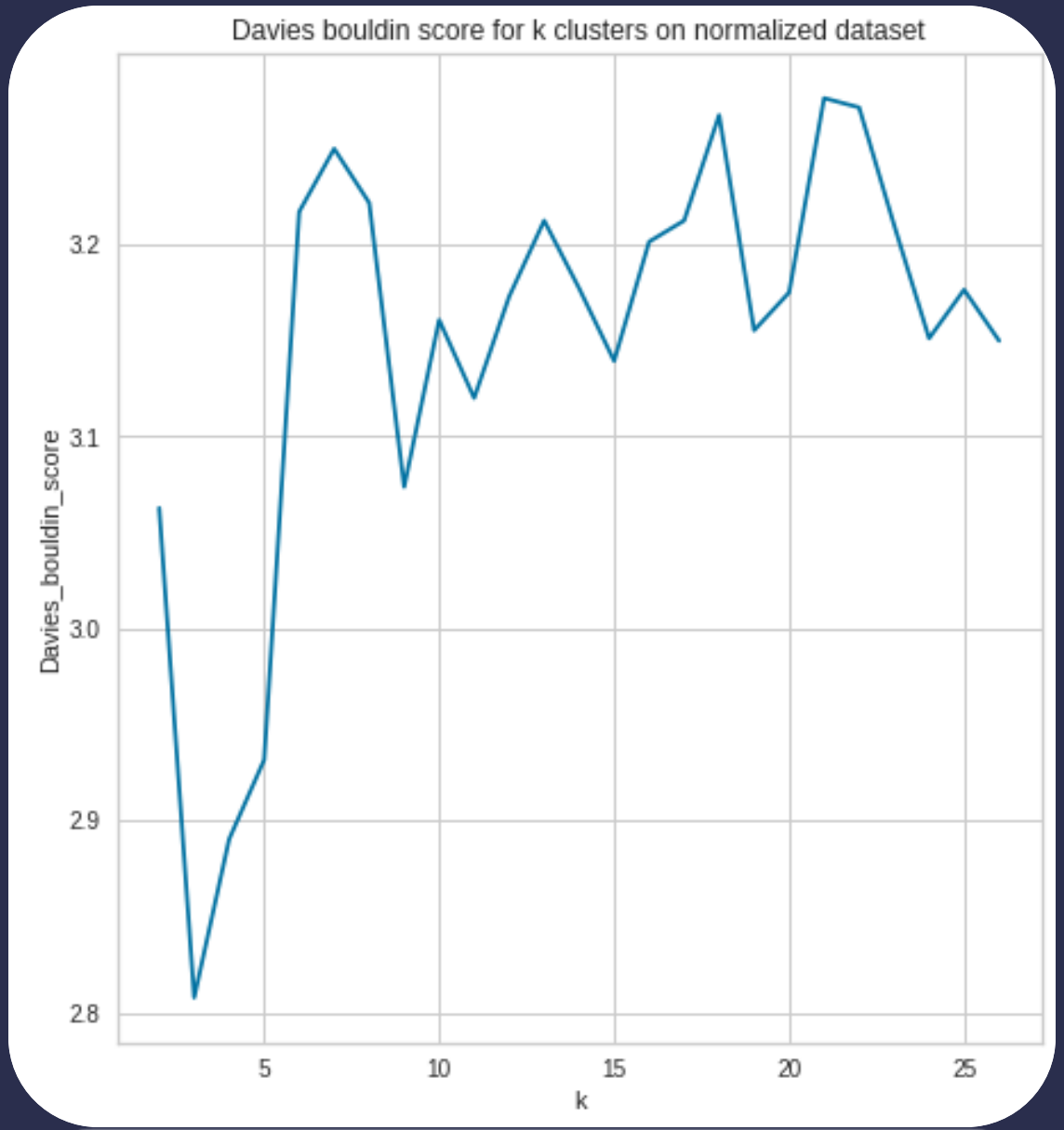
Metoda Silhouette'a

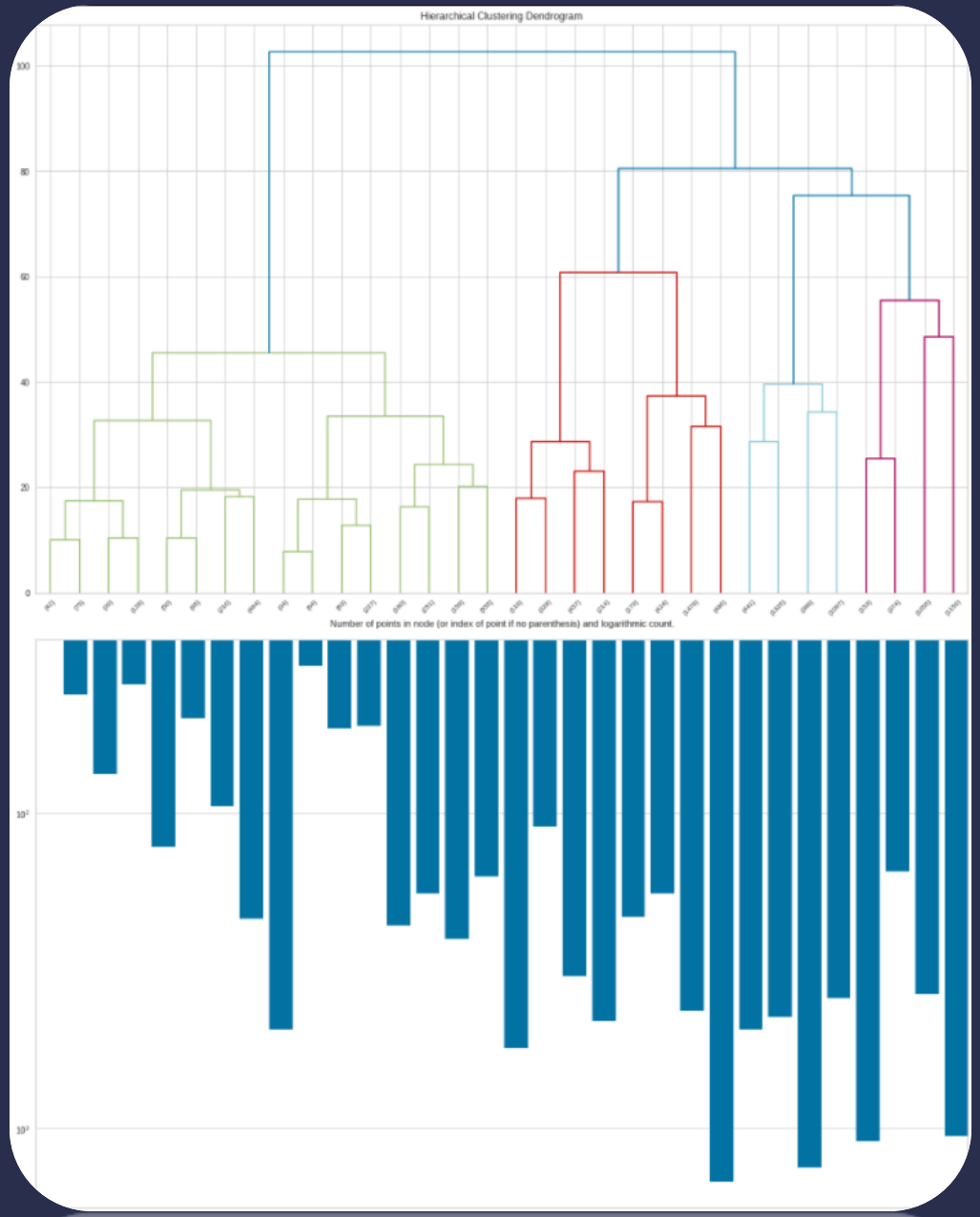




Metoda Silhouette’a dla 9 klastrów

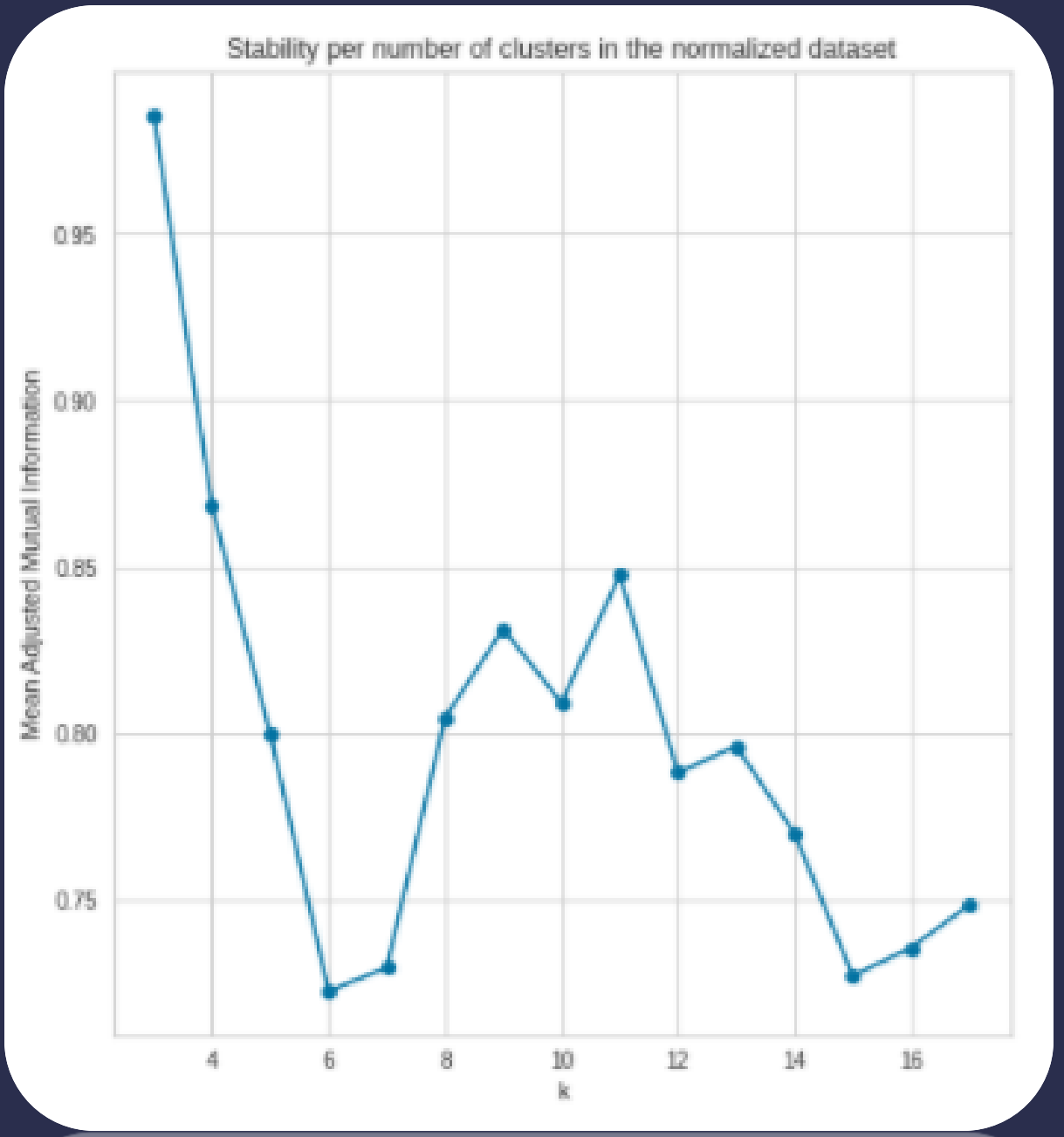
Davies-bouldin score





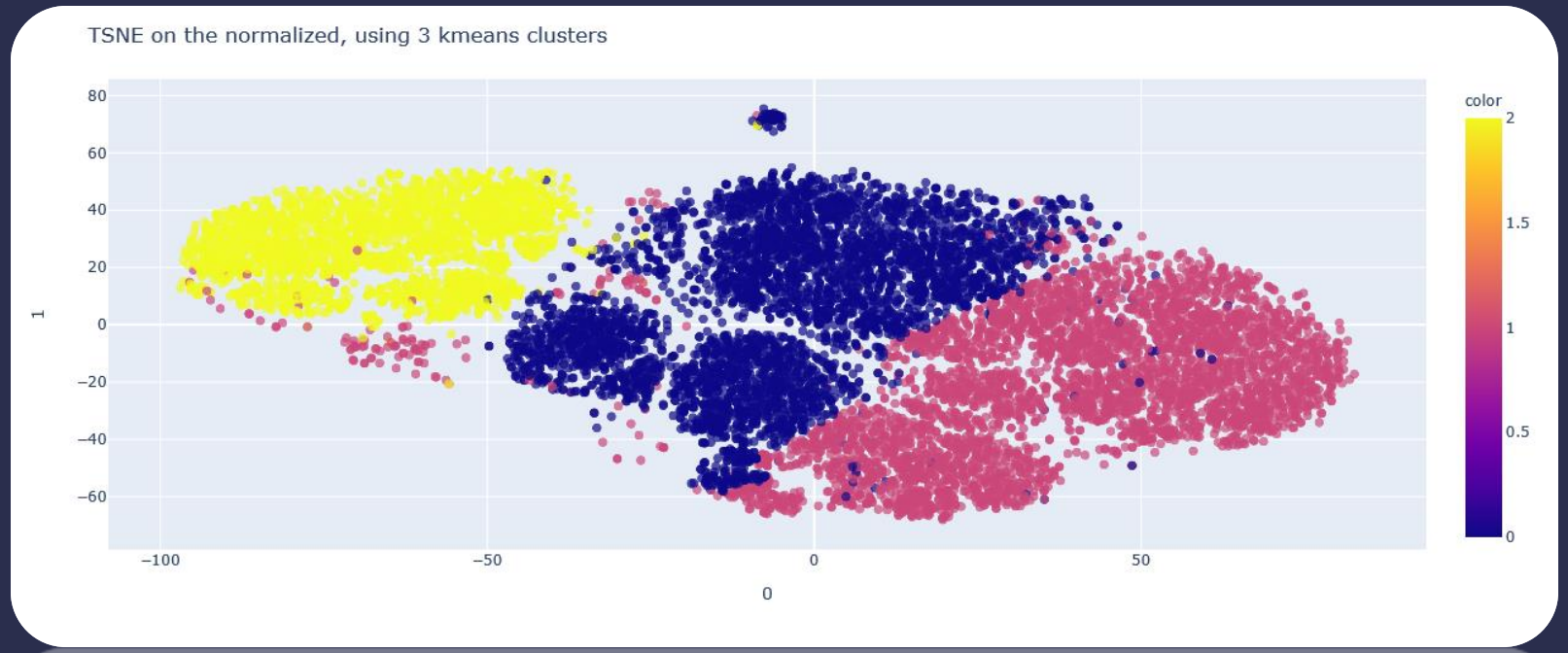
Hierarchiczny dendrogram

Stabilność

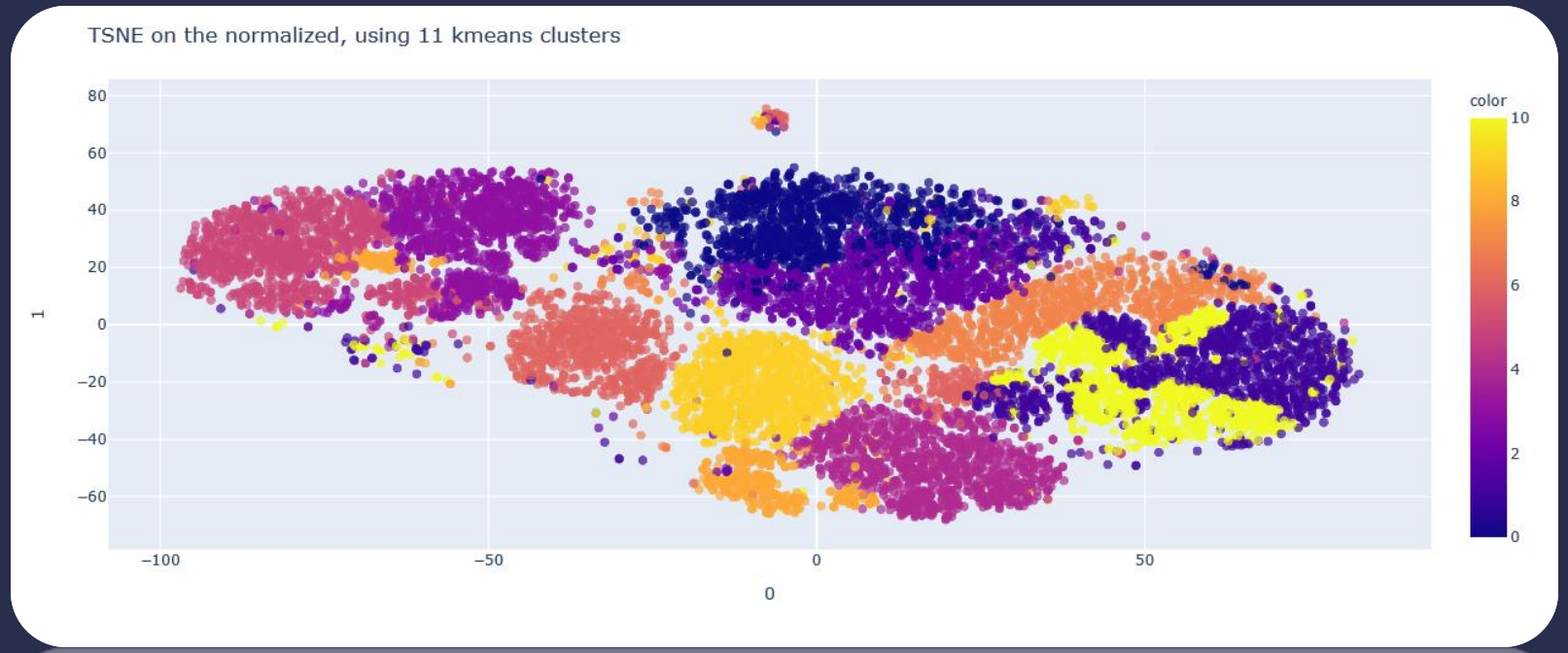

Stability per number of clusters in the normalized dataset
Mean Adjusted Mutual Information
0.95
0.90
0.85
0.80
0.75
4
6
8
10
12
14
16
k

Klastrowania Znormalizowane

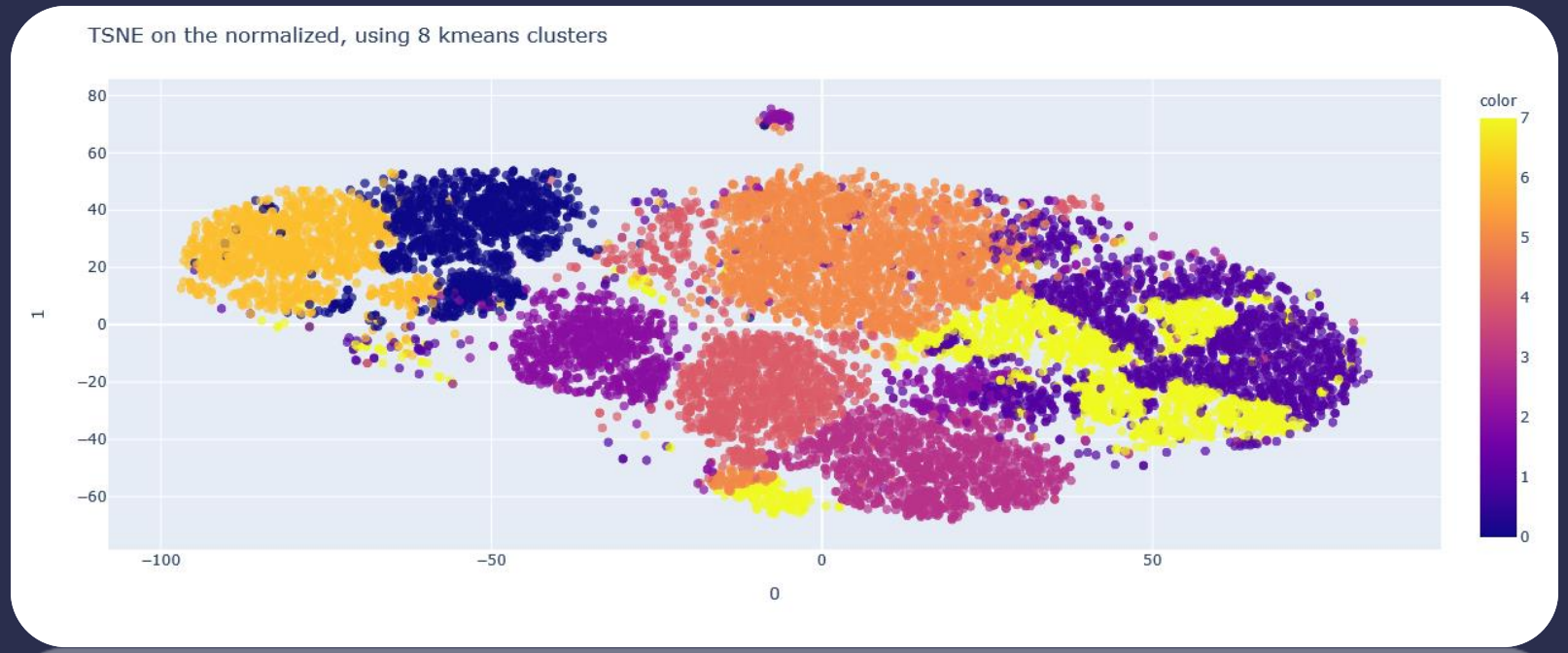
Klastrowanie z 3 klastrami



Klastrowanie z 11 klastrami



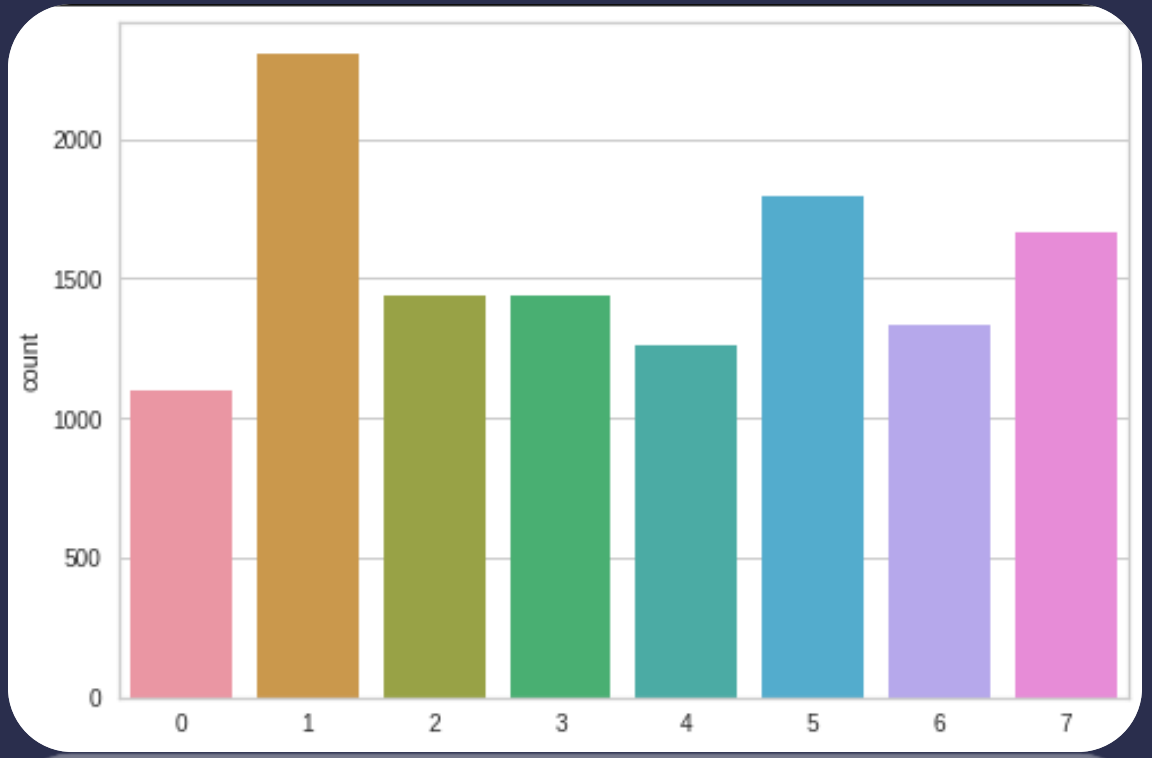
Klastrowanie z 8 klastrami



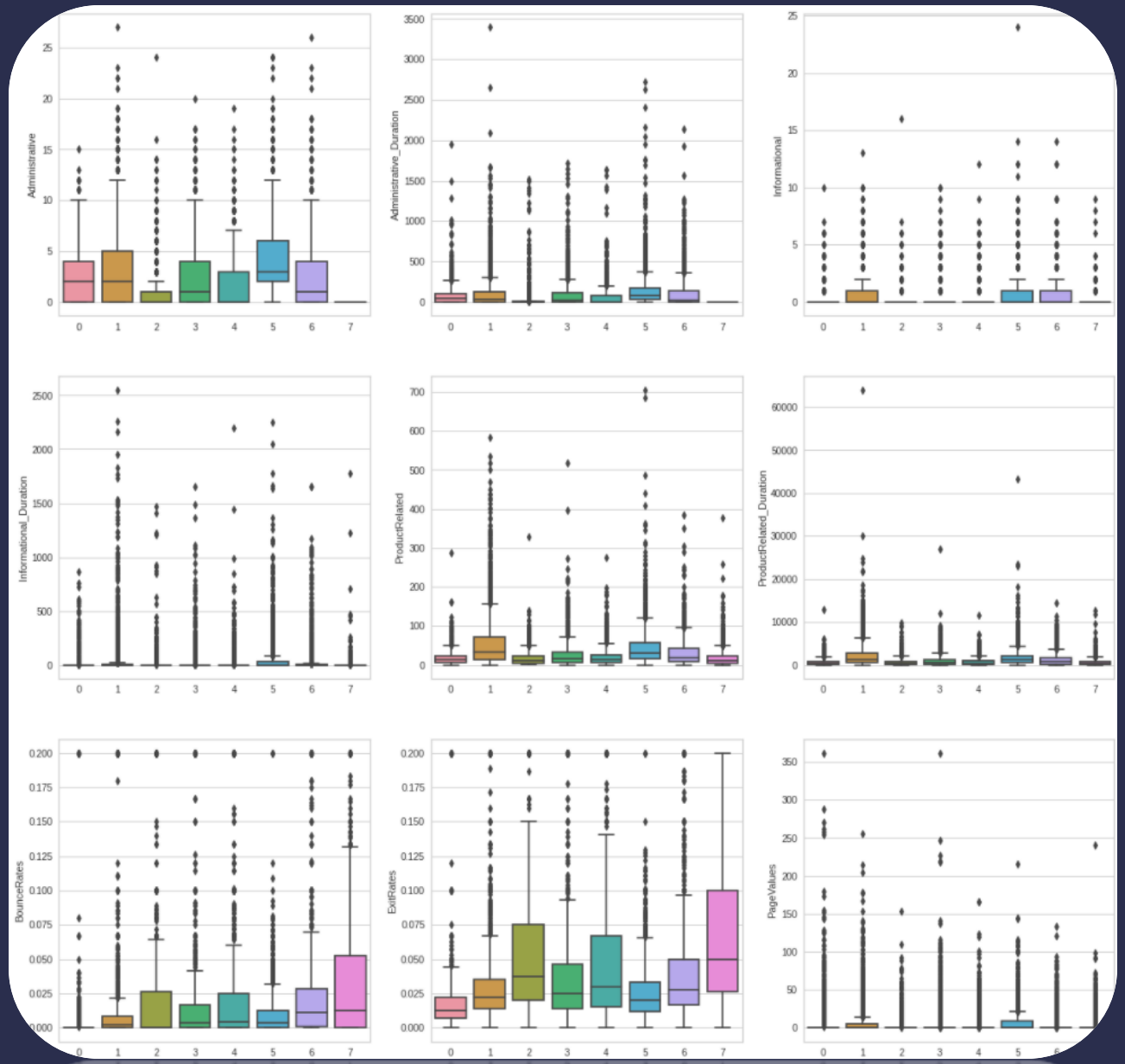
Analiza klastrów po normalizacji

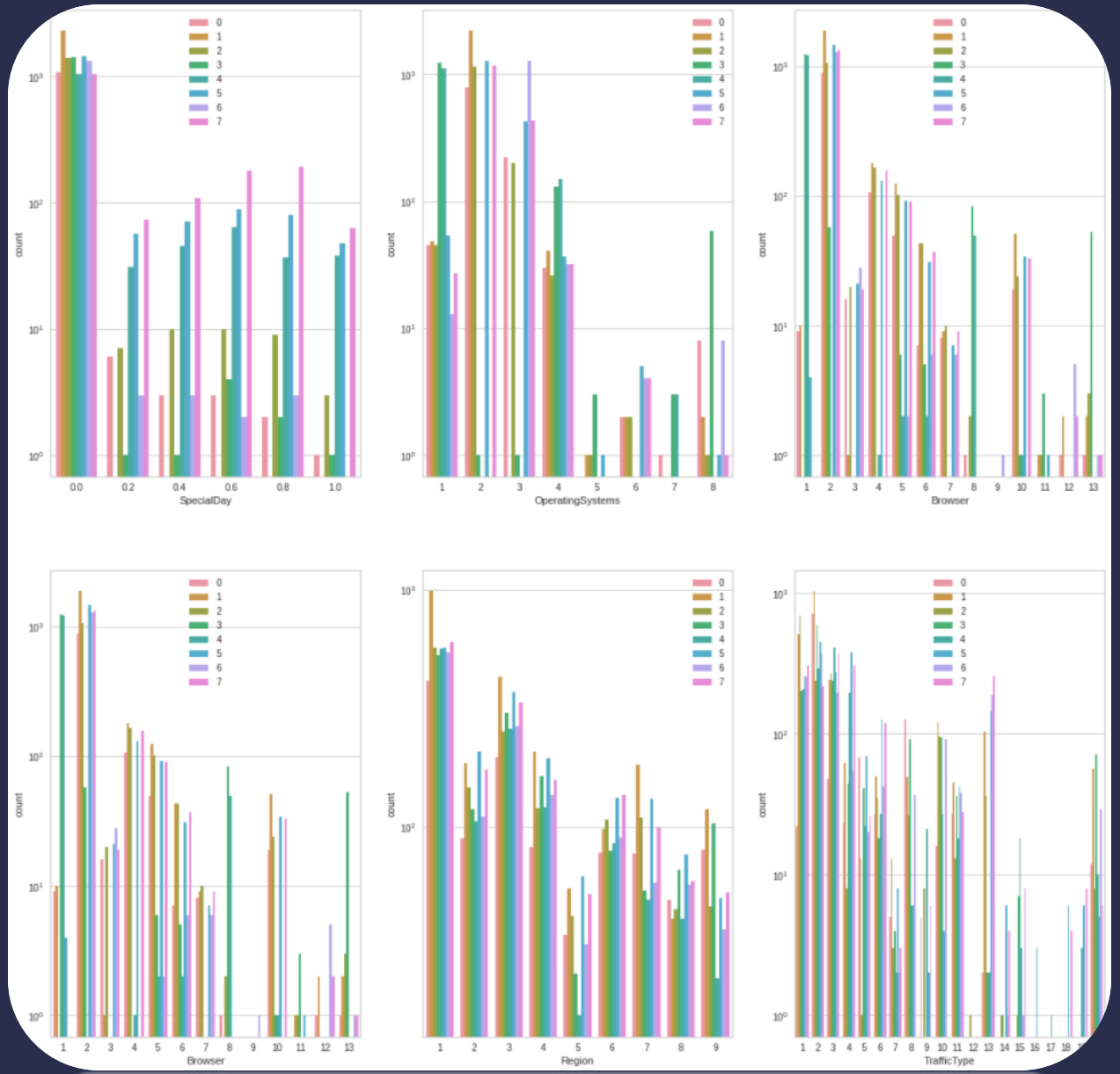
Analiza 8 Klastrów

	cluster	count	revenue_fraction	weekend_fraction
0	0	1100	0.249091	0.295455
1	1	2303	0.253582	0.193660
2	2	1436	0.054318	0.162953
3	3	1440	0.187500	0.268750
4	4	1262	0.101426	0.293978
5	5	1790	0.179330	0.222905
6	6	1331	0.129226	0.338843
7	7	1668	0.048561	0.152878



Analiza wykresów skrzynkowych





Analiza histogramów

Dziękujemy za Uwagę!