Projekt 2 - EDA

Artur Żółkowski,
Mikołaj Spytek

Zbiór danych

-

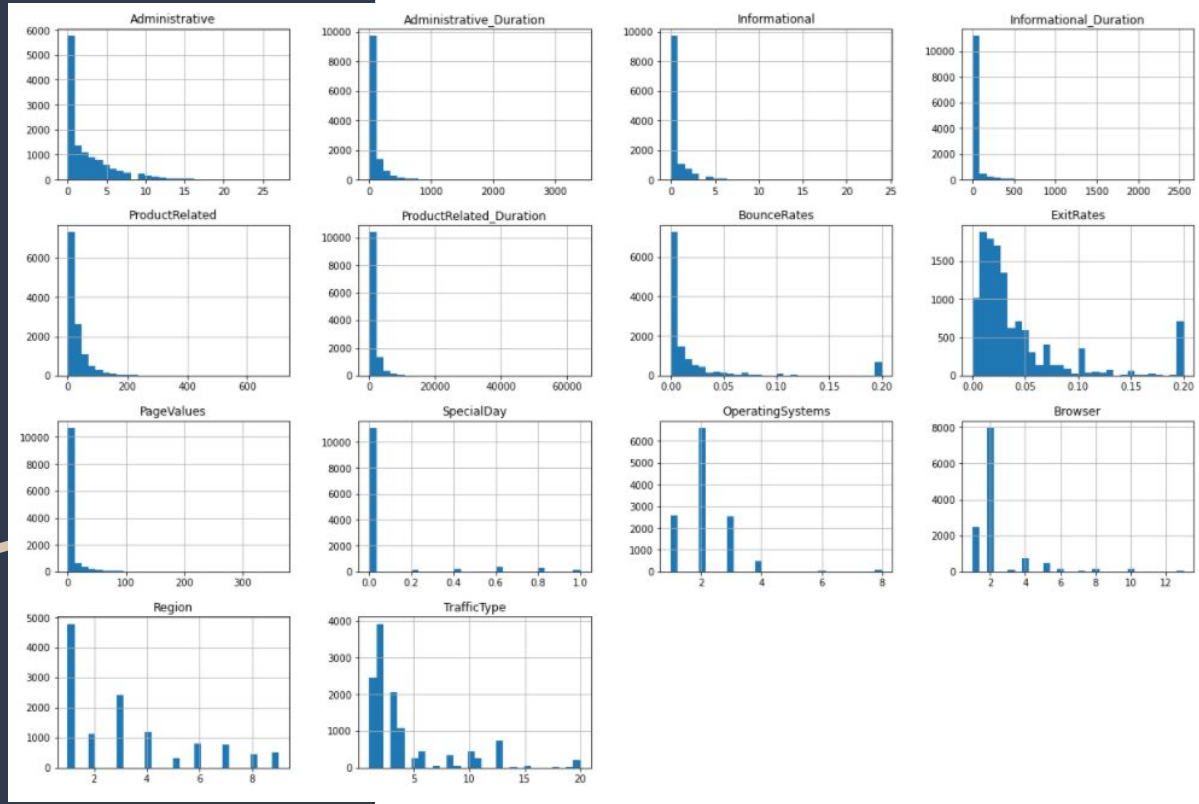
- 10 zmiennych liczbowych
- 8 zmienne kategoryczne

	Administrative	Administrative_Duration	Informational	Informational_Duration	ProductRelated	ProductRelated_Duration	BounceRates	ExitRates	PageValues
0	0	0.0	0	0.0	1	0.000000	0.20	0.20	0.0
1	0	0.0	0	0.0	2	64.000000	0.00	0.10	0.0
2	0	0.0	0	0.0	1	0.000000	0.20	0.20	0.0
3	0	0.0	0	0.0	2	2.666667	0.05	0.14	0.0
4	0	0.0	0	0.0	10	627.500000	0.02	0.05	0.0

SpecialDay	Month	OperatingSystems	Browser	Region	TrafficType	VisitorType	Weekend	Revenue
0.0	Feb	1	1	1	1	Returning_Visitor	False	False
0.0	Feb	2	2	1	2	Returning_Visitor	False	False
0.0	Feb	4	1	9	3	Returning_Visitor	False	False
0.0	Feb	3	2	2	4	Returning_Visitor	False	False
0.0	Feb	3	3	1	4	Returning_Visitor	True	False

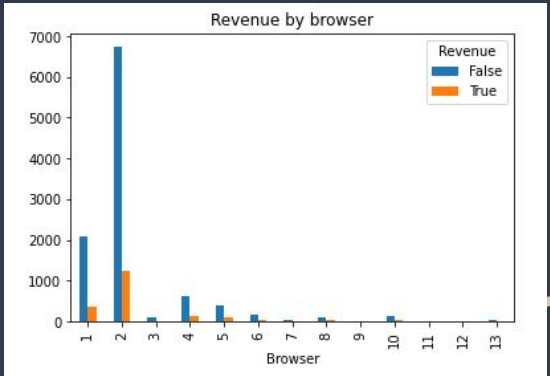
Rozkłady zmiennych

- prawie wszystkie dość mocno skośne do zera

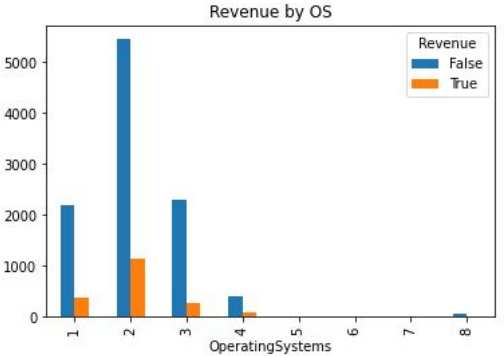


Revenue

- rozbicie revenue po różnych przeglądarkach i systemach operacyjnych



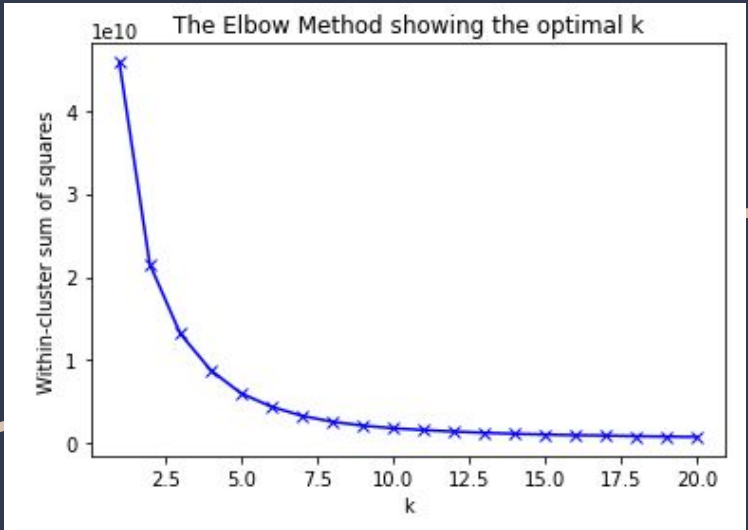
Revenue	False	True	procent
Browser			
1	2097.0	365.0	0.148253
2	6738.0	1223.0	0.153624
3	100.0	5.0	0.047619
4	606.0	130.0	0.176630
5	381.0	86.0	0.184154
6	154.0	20.0	0.114943
7	43.0	6.0	0.122449
8	114.0	21.0	0.155556
9	1.0	NaN	NaN
10	131.0	32.0	0.196319
11	5.0	1.0	0.166667
12	7.0	3.0	0.300000
13	45.0	16.0	0.262295

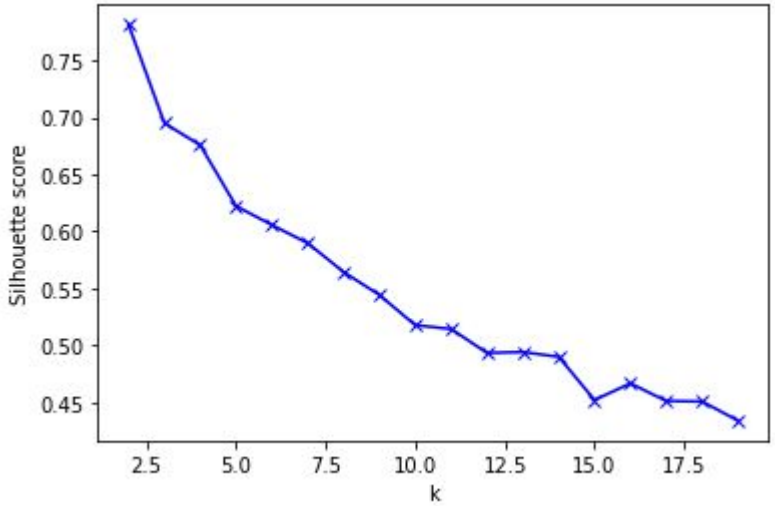


Revenue	False	True	procent
OperatingSystems			
1	2206	379	0.146615
2	5446	1155	0.174973
3	2287	268	0.104892
4	393	85	0.177824
5	5	1	0.166667
6	17	2	0.105263
7	6	1	0.142857
8	62	17	0.215190

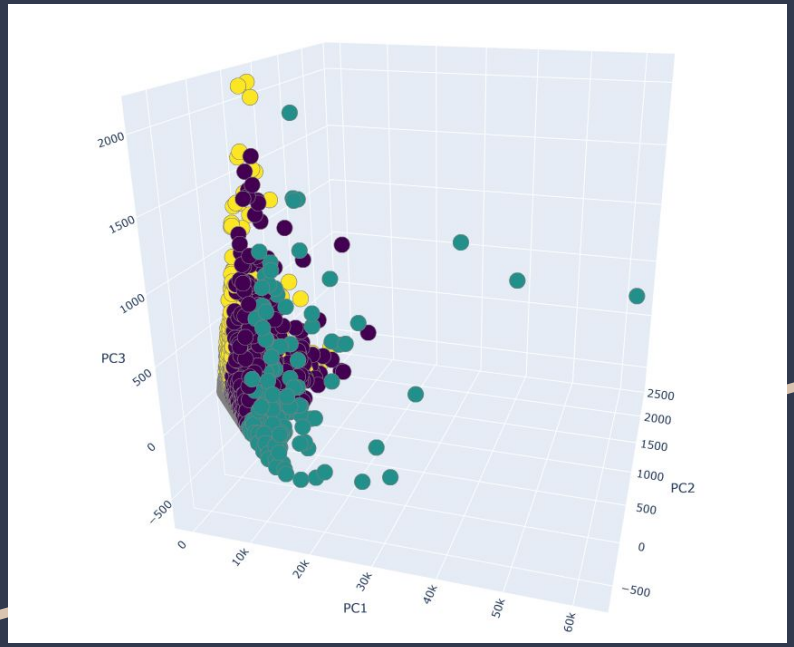
Wstępny wybór liczby klastrów

- metoda łokcia oraz silhouette score - niestety nie ma wyraźnych wskazań, będziemy próbowali korzystać też z innych metod

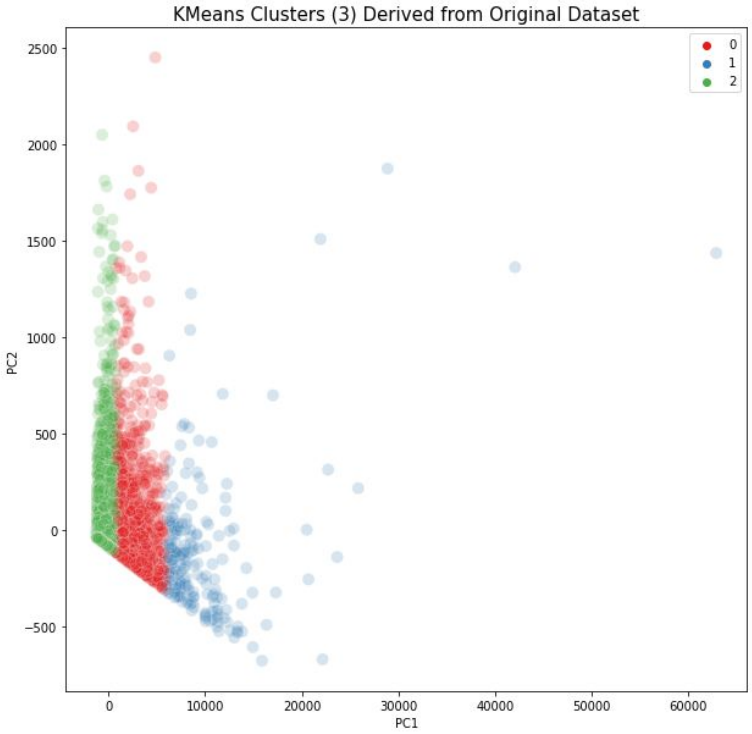




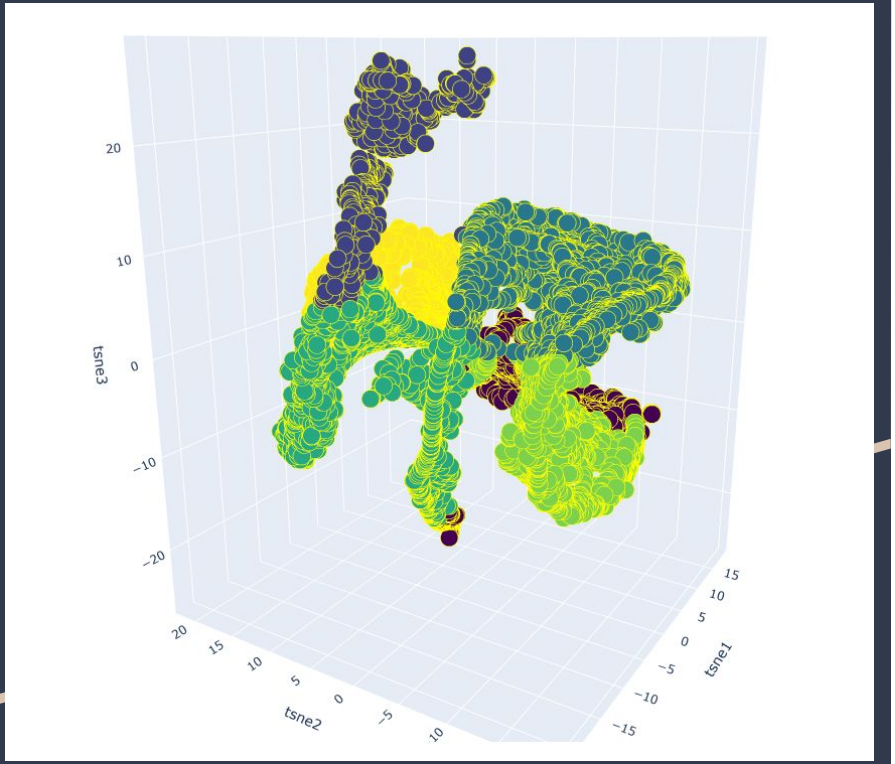
PCA

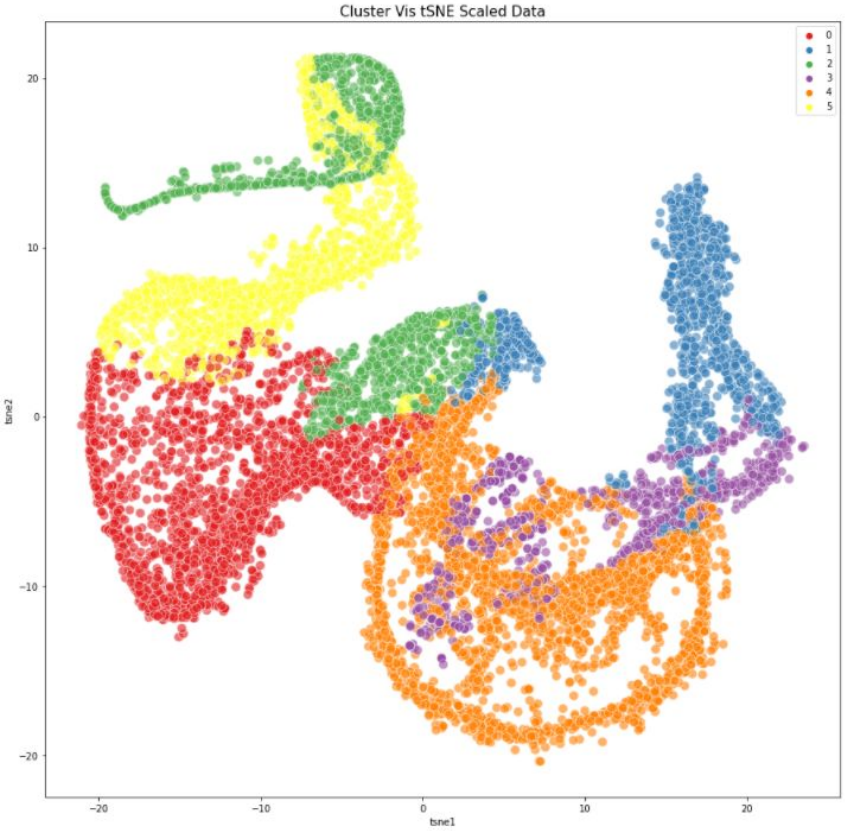


- jedna ze współrzędnych wygenerowana przez pca bardzo dobrze rozgranicza klastry



t-SNE


tsne3
tsne2
tsne1


Cluster Vis tSNE Scaled Data
0
1
2
3
4
5
tsne2
tsne1