

Online shoppers' purchasing intention

JAN SMOLEŃ, KATARZYNA SOLAWA



Introduction

The goal of this project is to use data about online sessions on the shop website to inspect whether we can extract distinct groups of customers based on their behaviour and shopping intentions, which could be useful in further development of the website and the whole brand.

Data source:

<https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset#>

Dataset has 125 (1.0%) duplicate rows	Duplica
BounceRates is highly correlated with ExitRates	High co
ExitRates is highly correlated with BounceRates	High co
Administrative has 5768 (46.8%) zeros	Zeros
Administrative_Duration has 5903 (47.9%) zeros	Zeros
Informational has 9699 (78.7%) zeros	Zeros
Informational_Duration has 9925 (80.5%) zeros	Zeros
ProductRelated_Duration has 755 (6.1%) zeros	Zeros
BounceRates has 5518 (44.8%) zeros	Zeros
PageValues has 9600 (77.9%) zeros	Zeros
SpecialDay has 11079 (89.9%) zeros	Zeros

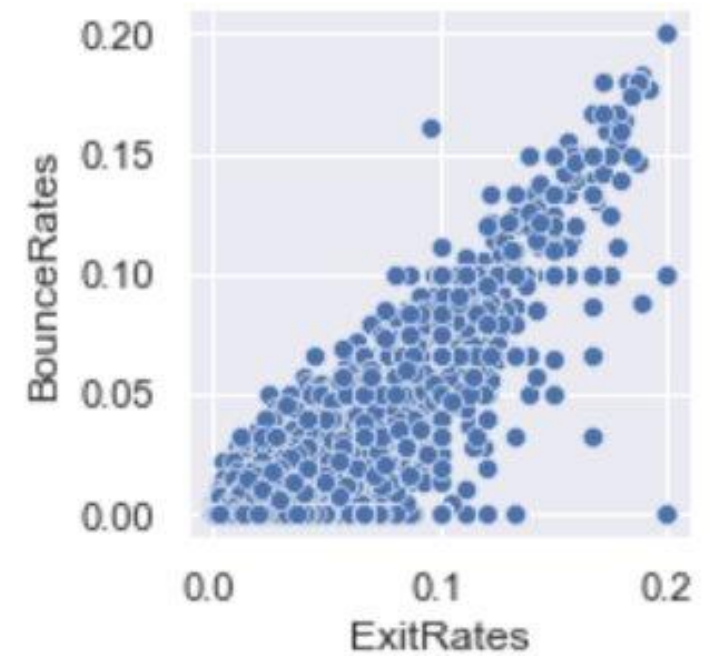
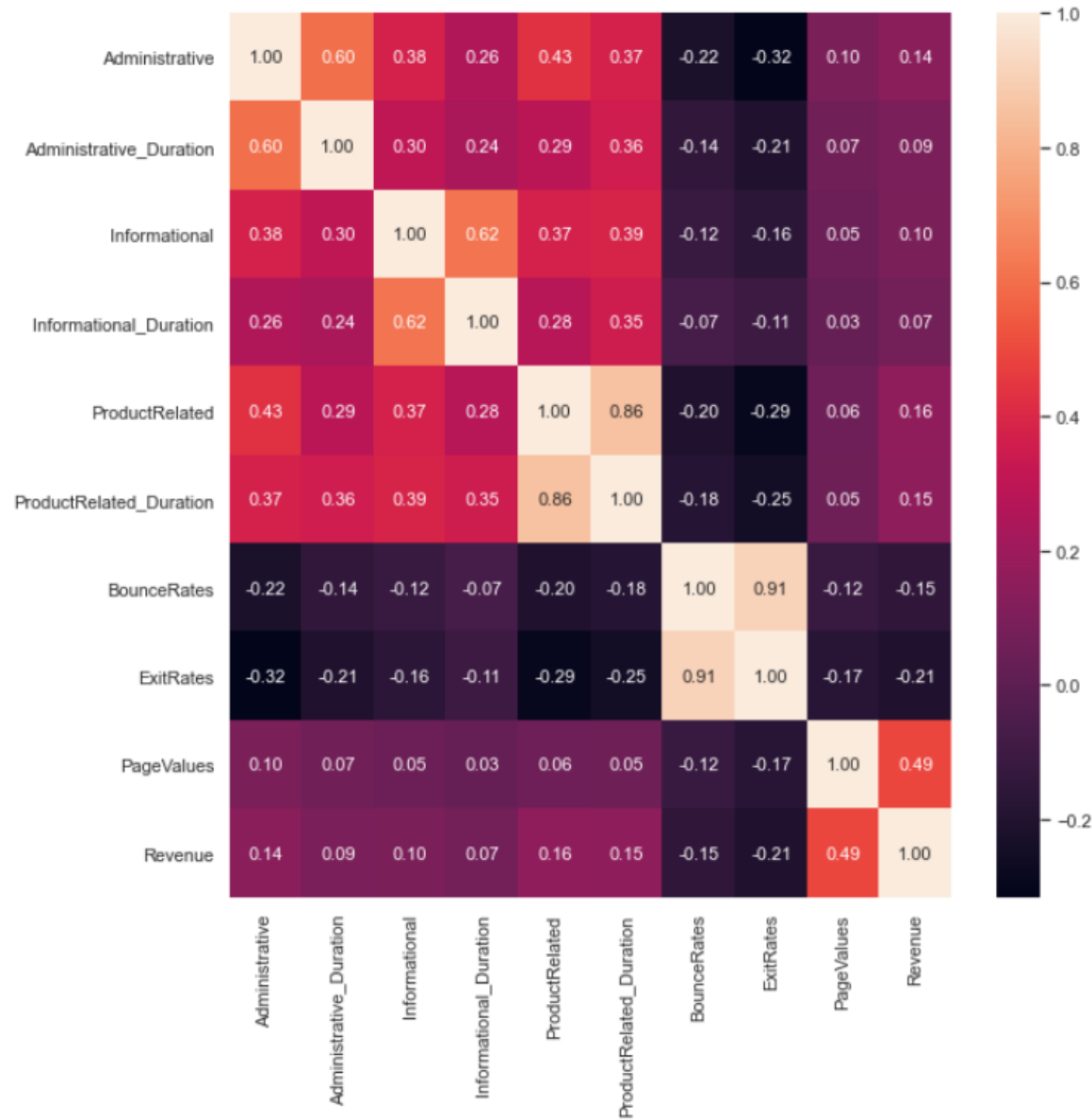
Data columns (total 18 columns):

#	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

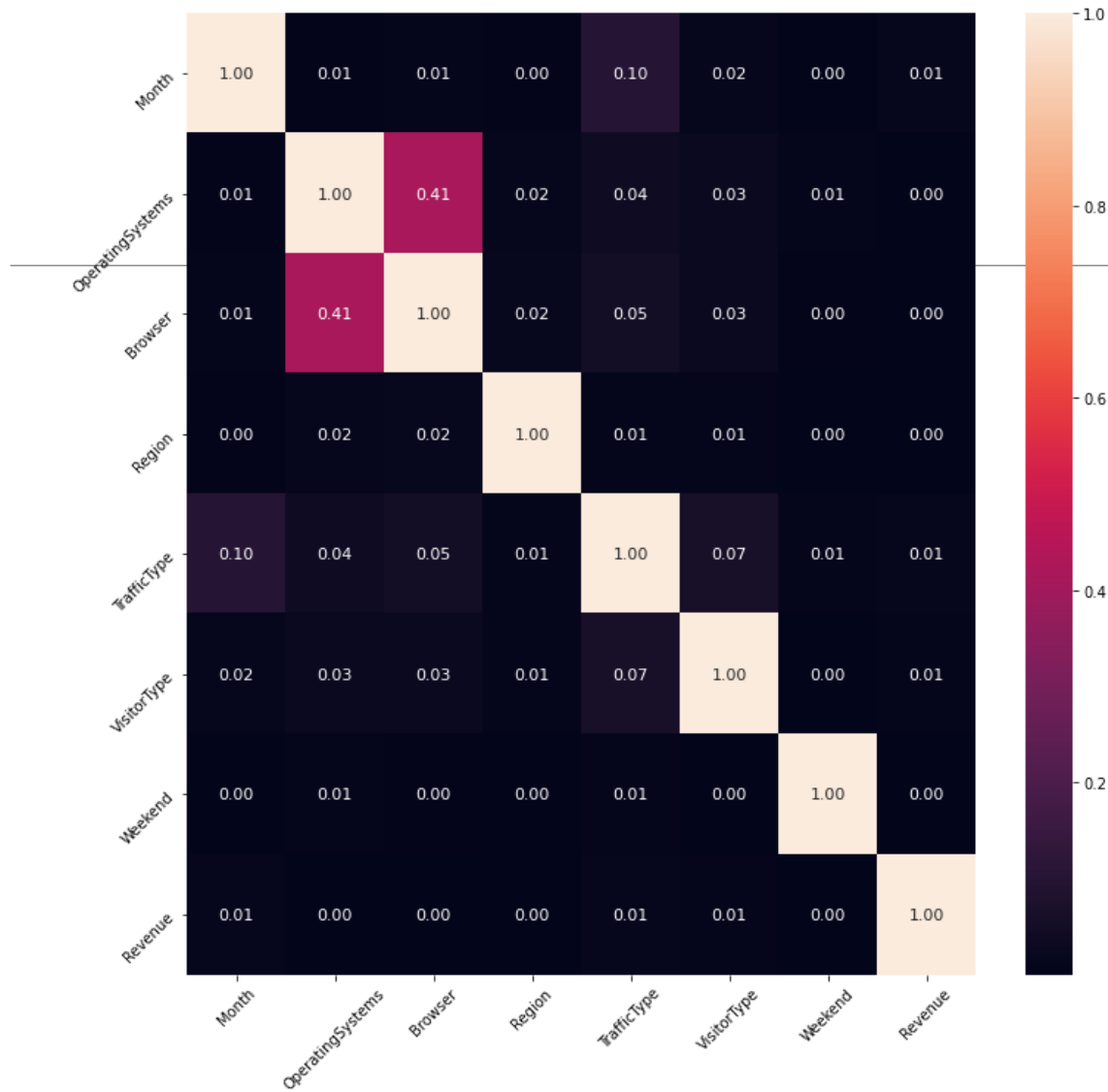
Data info

- both discrete and continuous variables
- highly correlated variables
- Many mostly zero variables

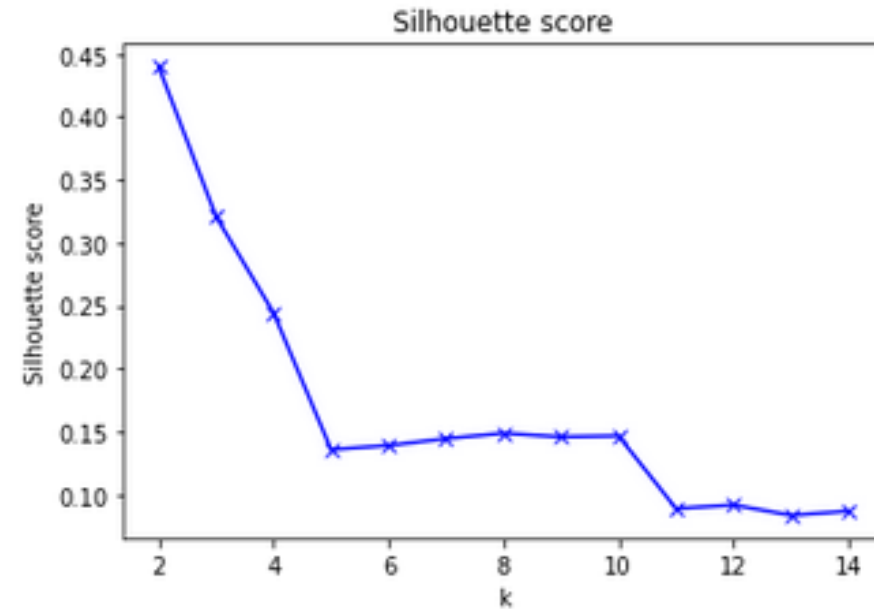
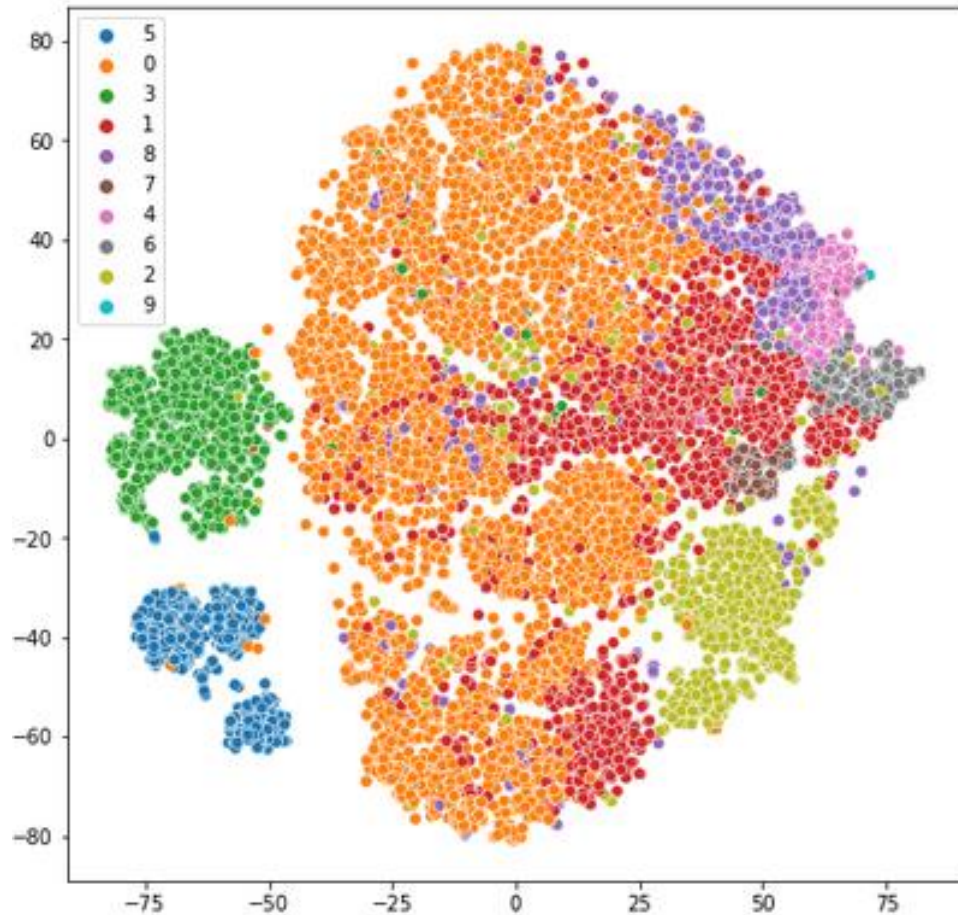
Correlation matrix



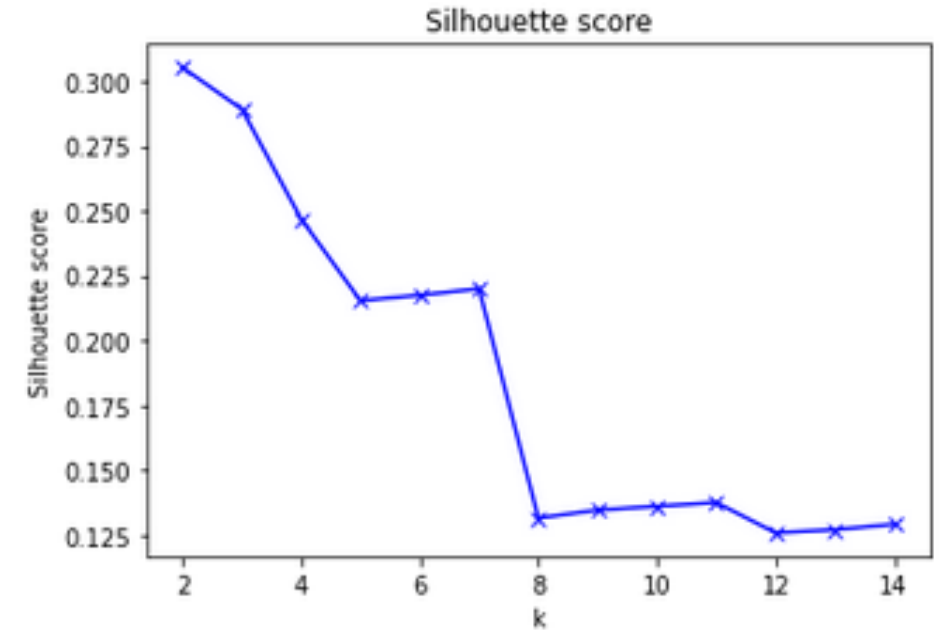
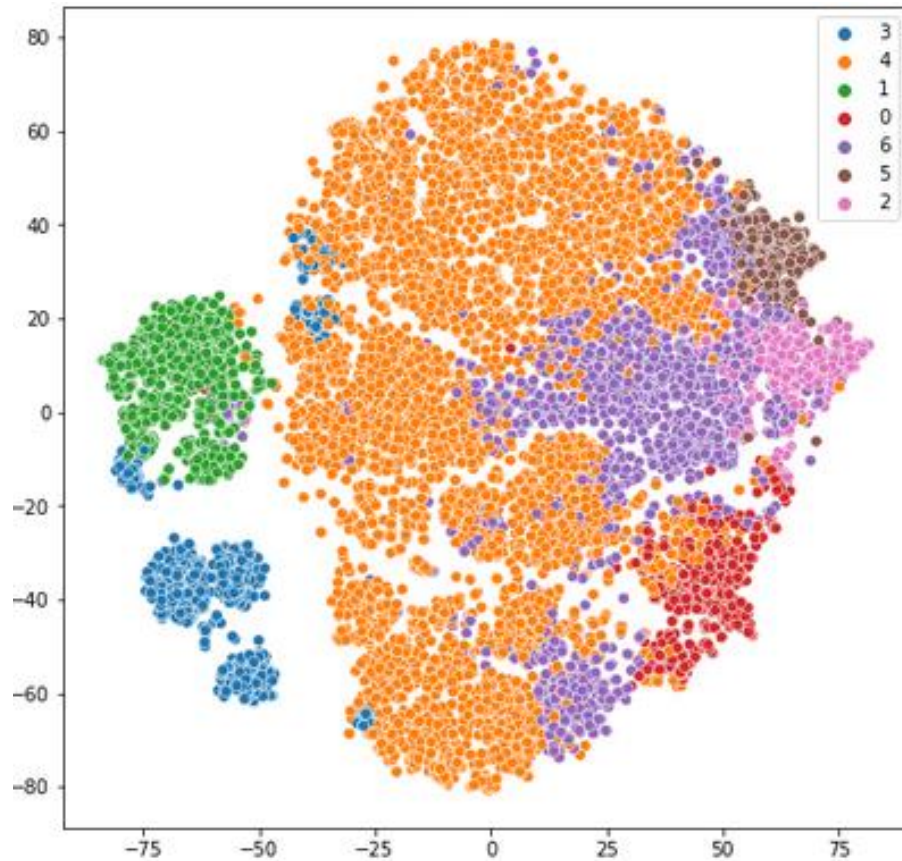
Mutual information



Hierarchical clustering

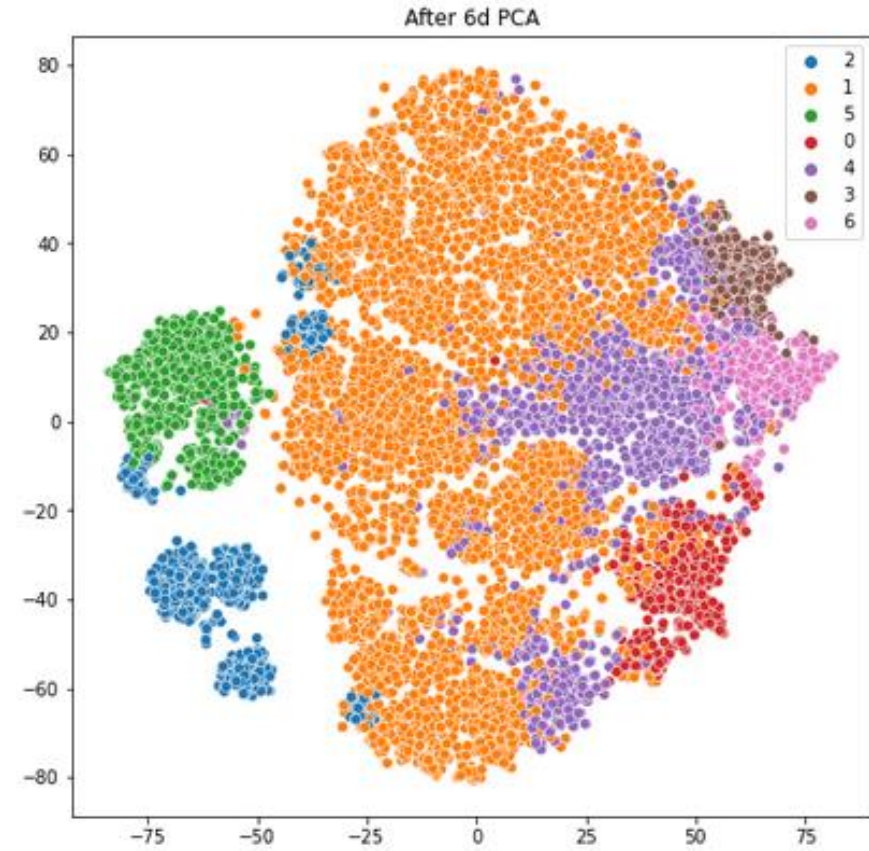
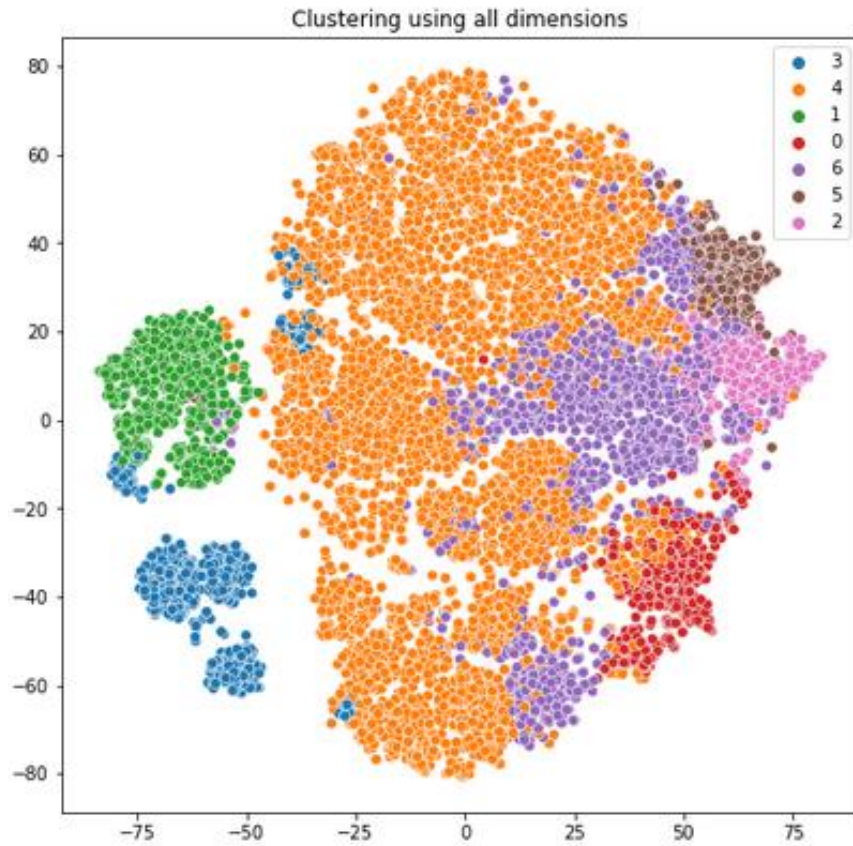


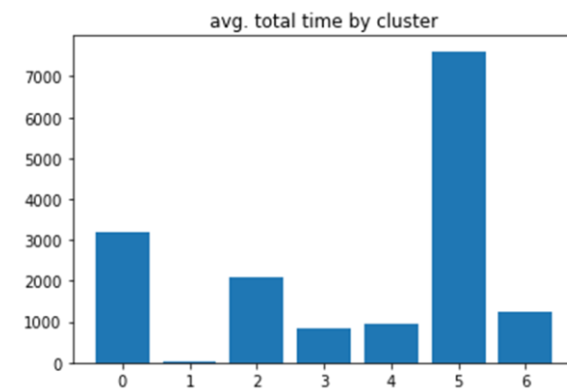
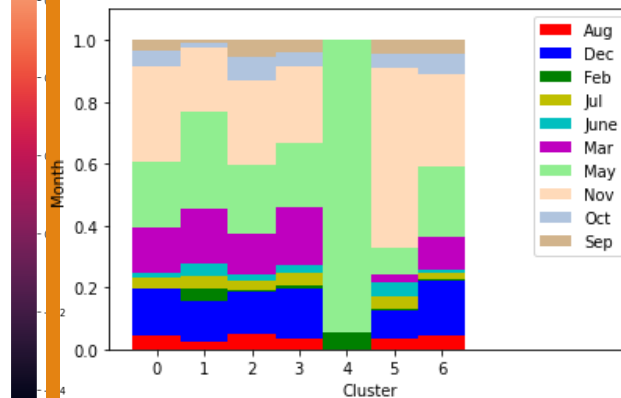
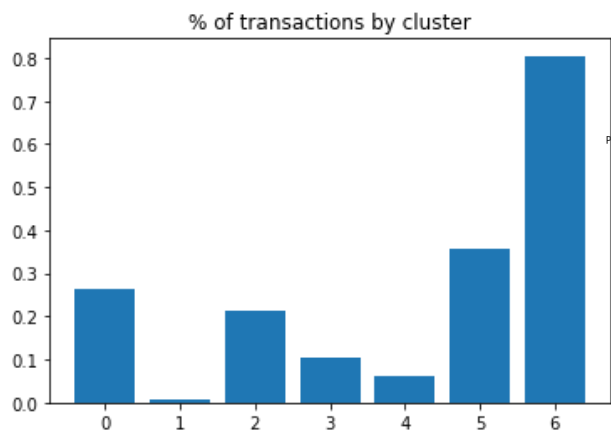
Clustering using K-means



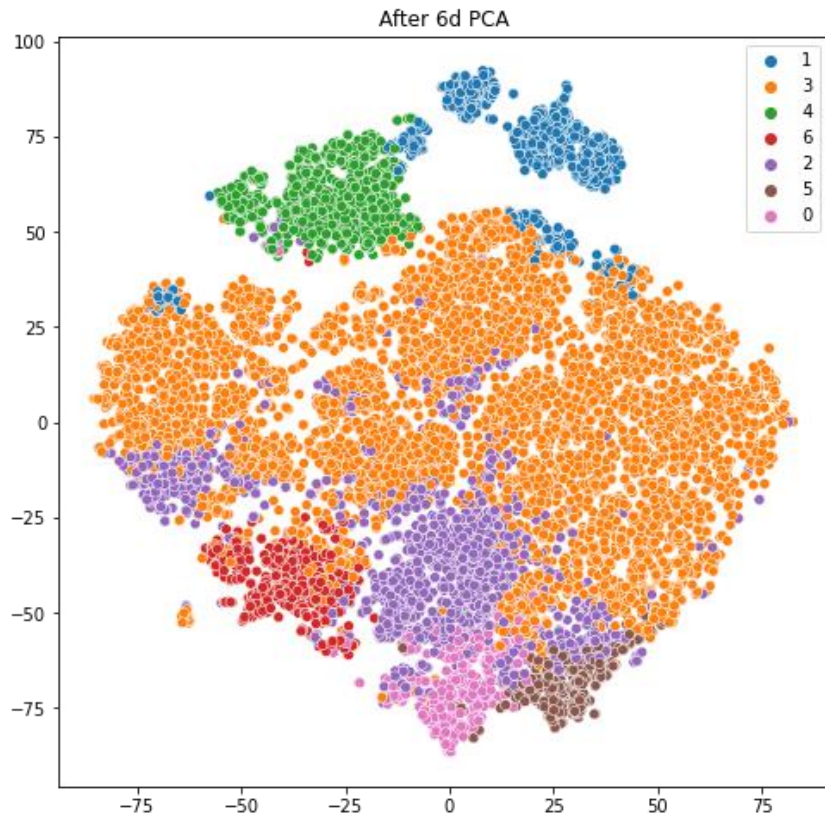
After analysis, decided to stick with K-means clustering using 7 distinct clusters.

Using PCA to reduce dimensionality





Some cluster analysis



Describing types of customers based on performed clustering

Cluster 0 – long time spent on information pages, not many transactions – maybe someone looking on FAQs?

Cluster 1 – immediate „bounce” – leaves website immediately

Cluster 2 – long time spent on administrative pages, not many transactions – someone signing up for the first time? Admins?

Cluster 3 – average customer, short time spent and not too many transactions

Cluster 4 – customers visiting before Special Days (example Valentines etc)

Cluster 5 – a long time spent on product-related pages, many transactions!

Cluster 6 – customer ending up buying the product, lots of transactions and above average time spent