Online shoppers purchasing intention

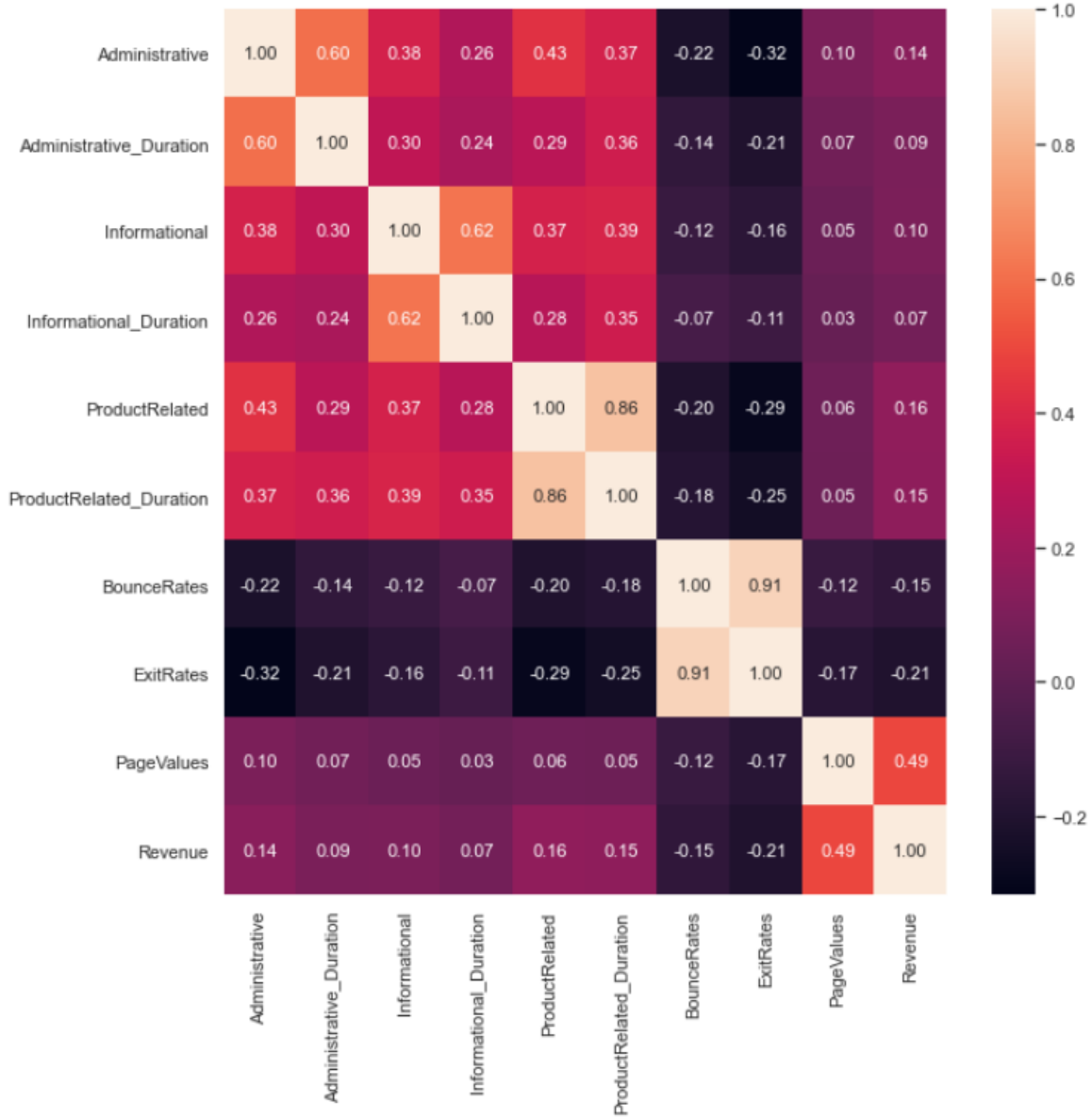
KATARZYNA SOLAWA, JAN SMOLEŃ

Podstawowe informacje

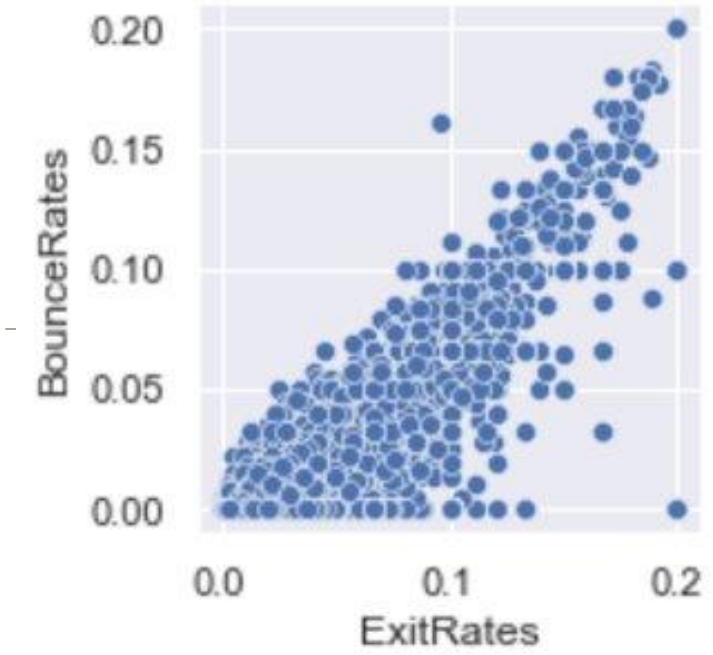
```
df=pd.read_csv("online_shoppers_intention.csv")
df=df.info()
```

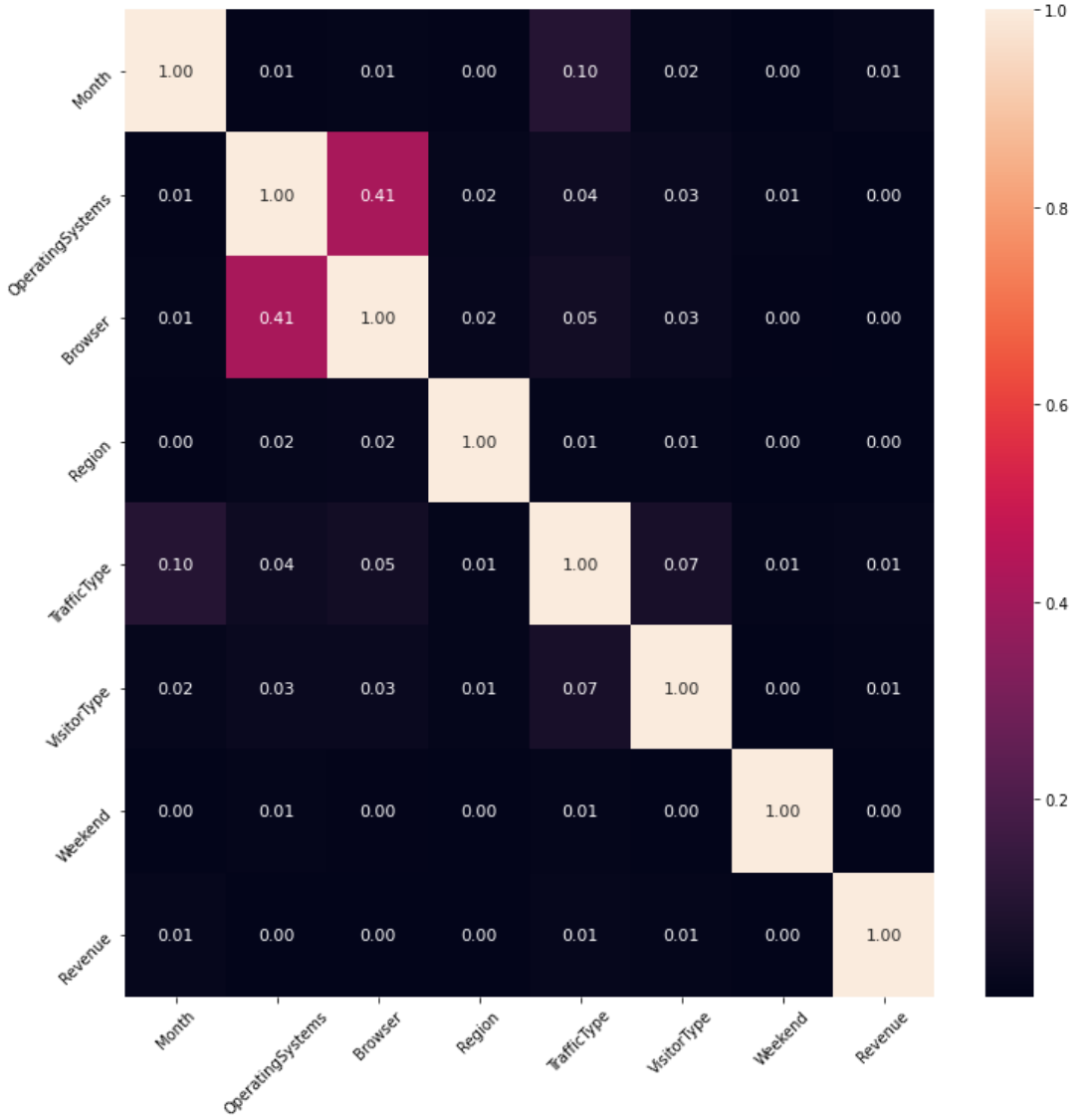
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 12330 entries, 0 to 12329
Data columns (total 18 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   Administrative           12330 non-null  int64
 1   Administrative_Duration  12330 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
```

Dataset has 125 (1.0%) duplicate rows	Duplicates
`BounceRates` is highly correlated with `ExitRates`	High correlation
`ExitRates` is highly correlated with `BounceRates`	High correlation
`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

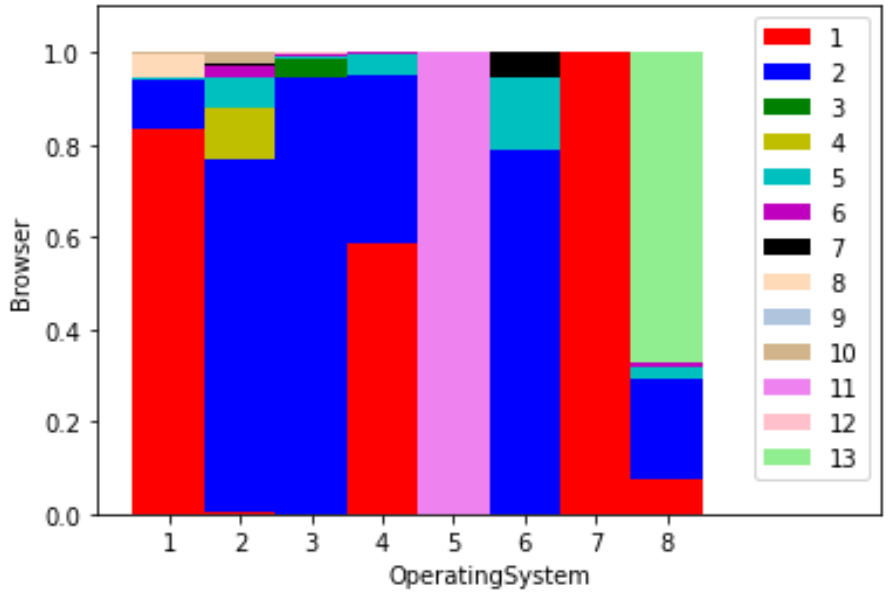


Skorelowane zmienne

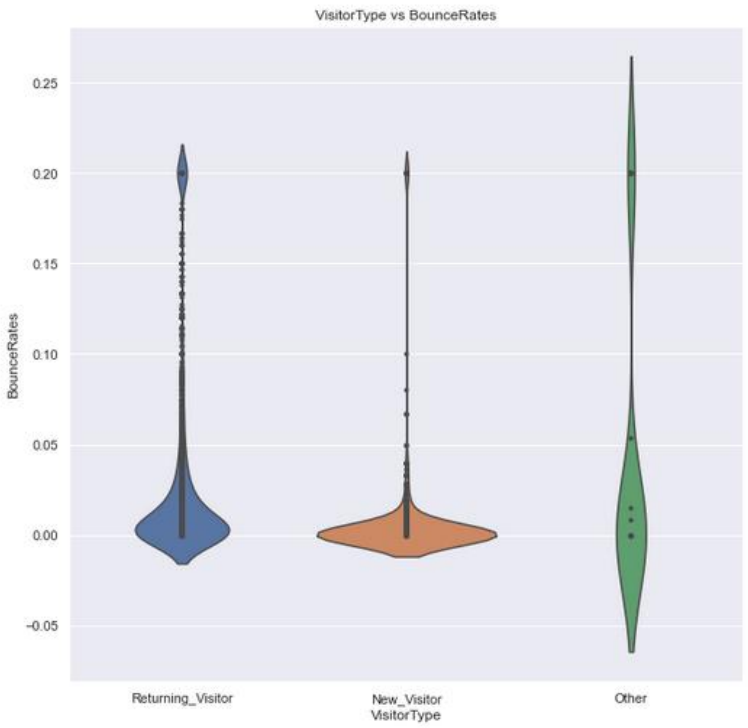


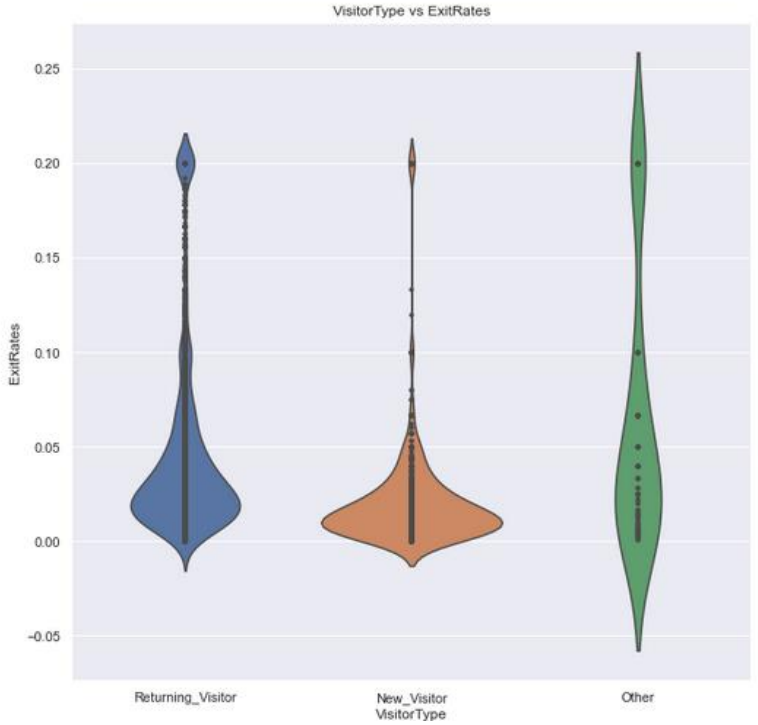


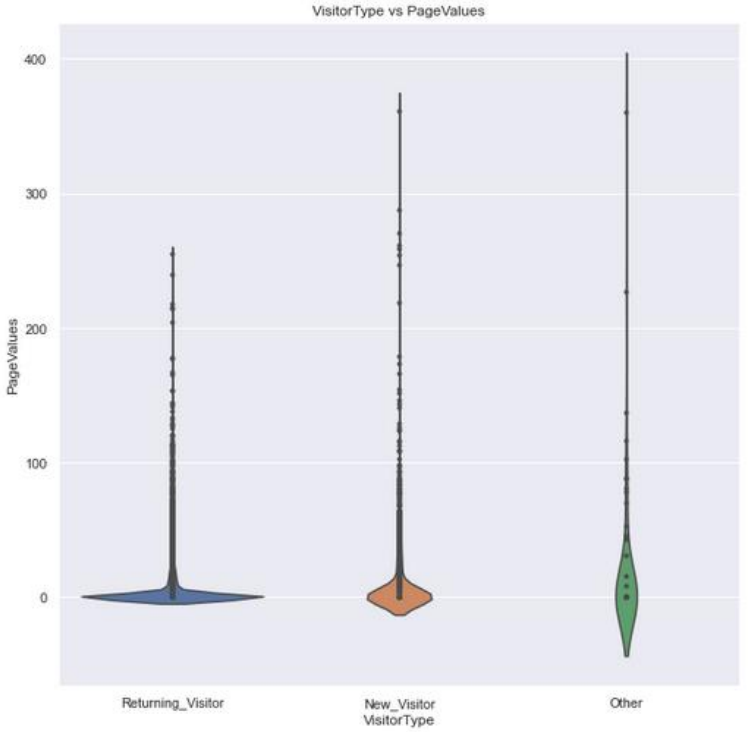
Informacja wzajemna

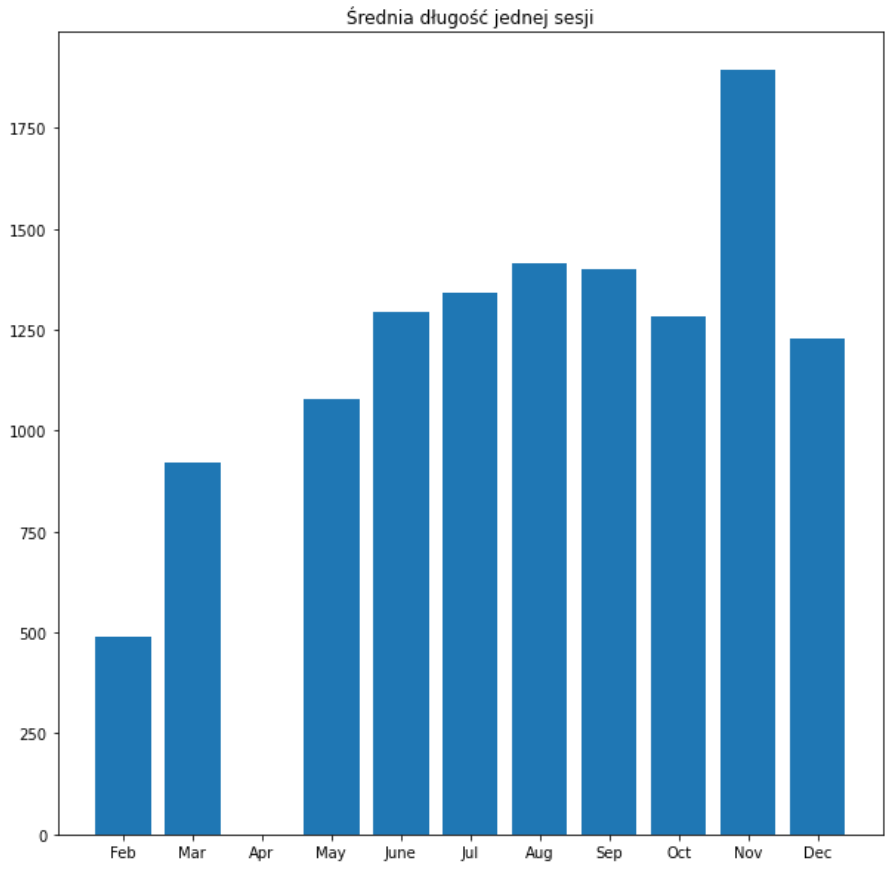


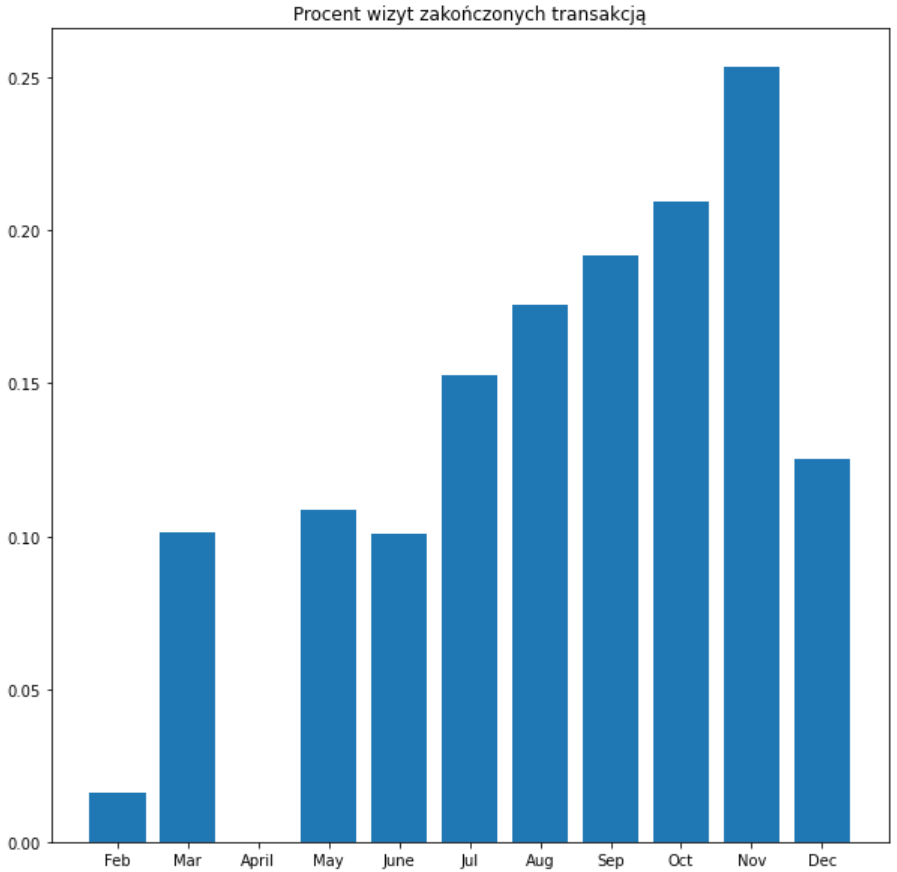
VisitorType



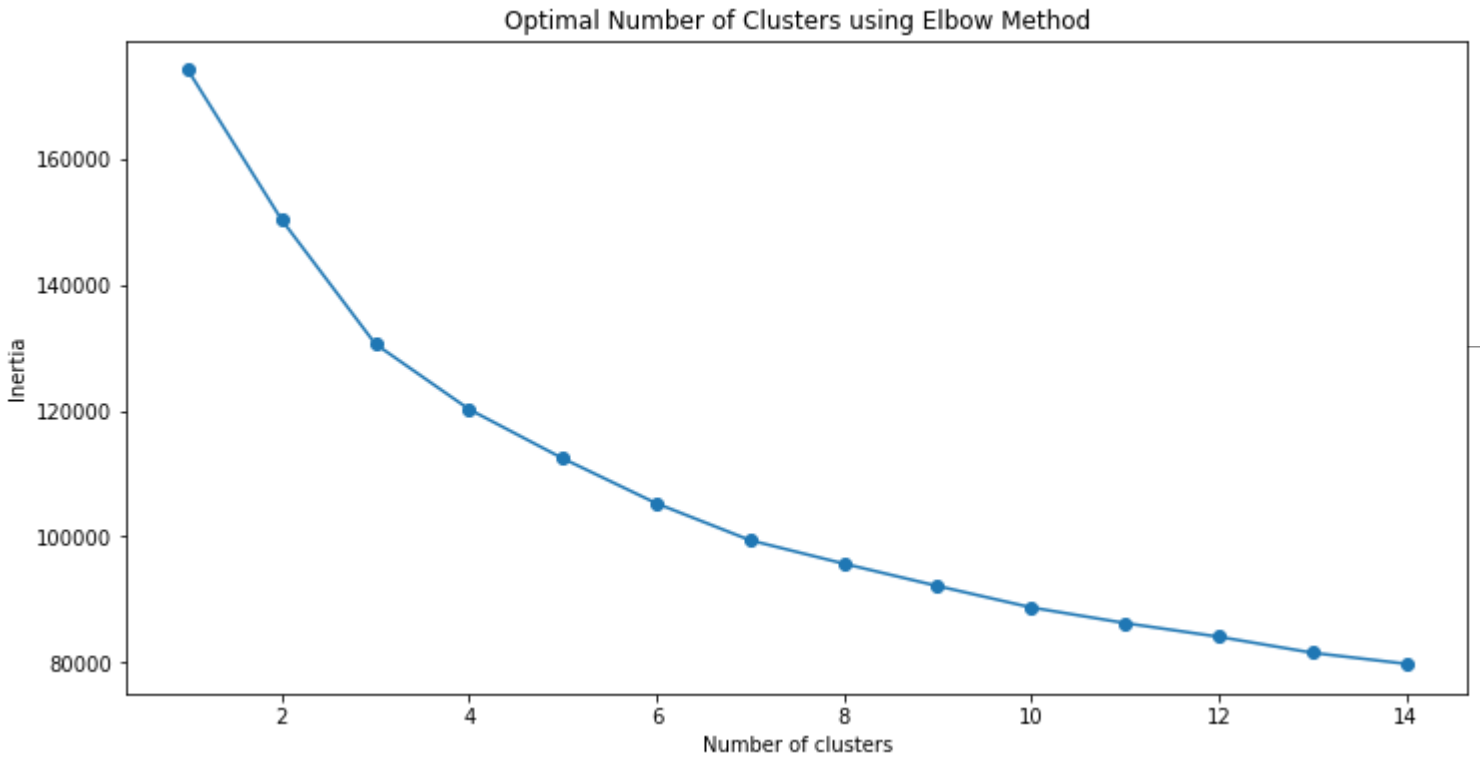




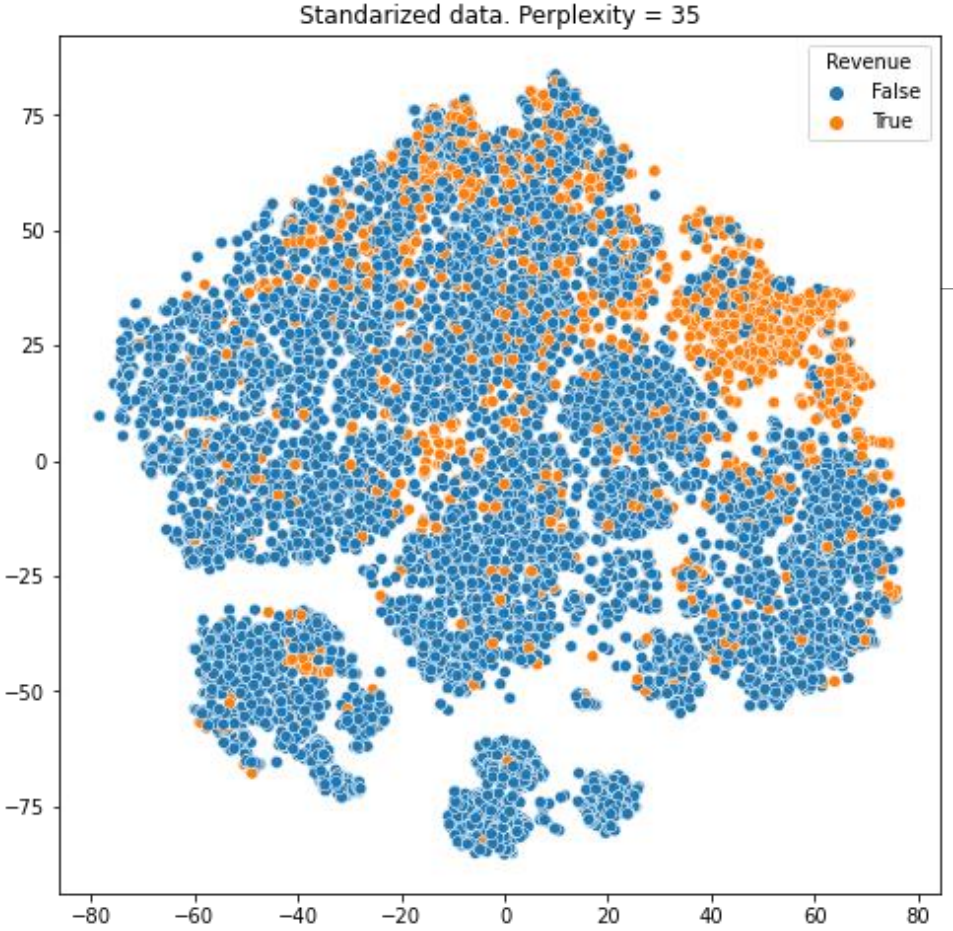


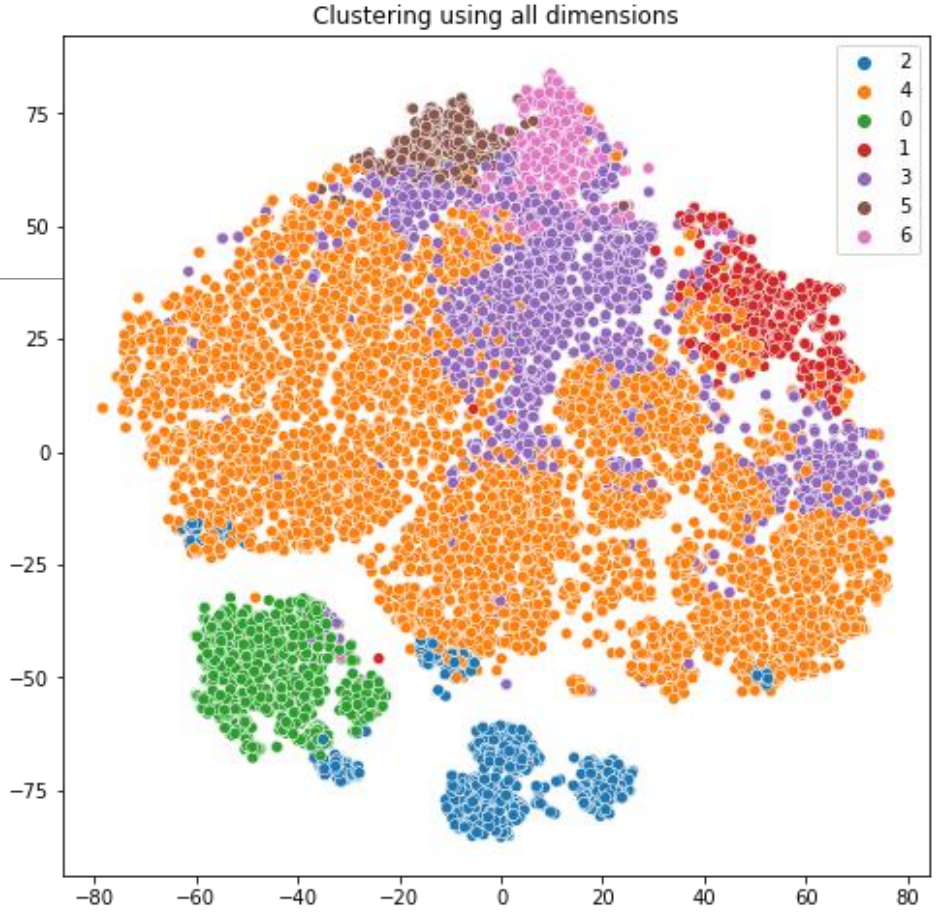


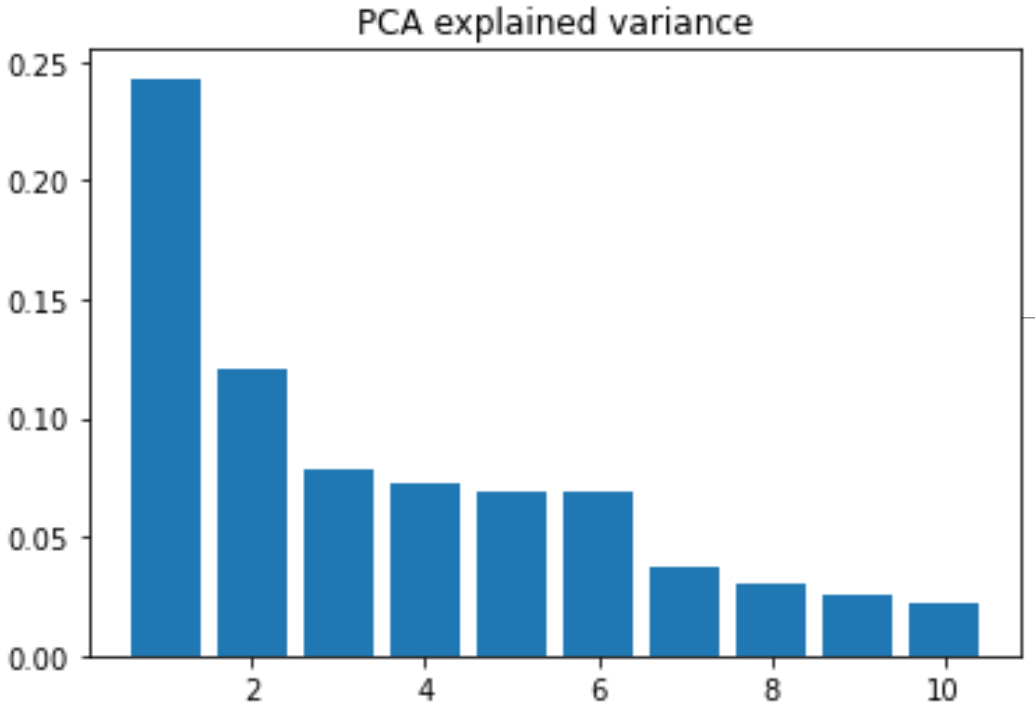
Transakcje vs miesiące

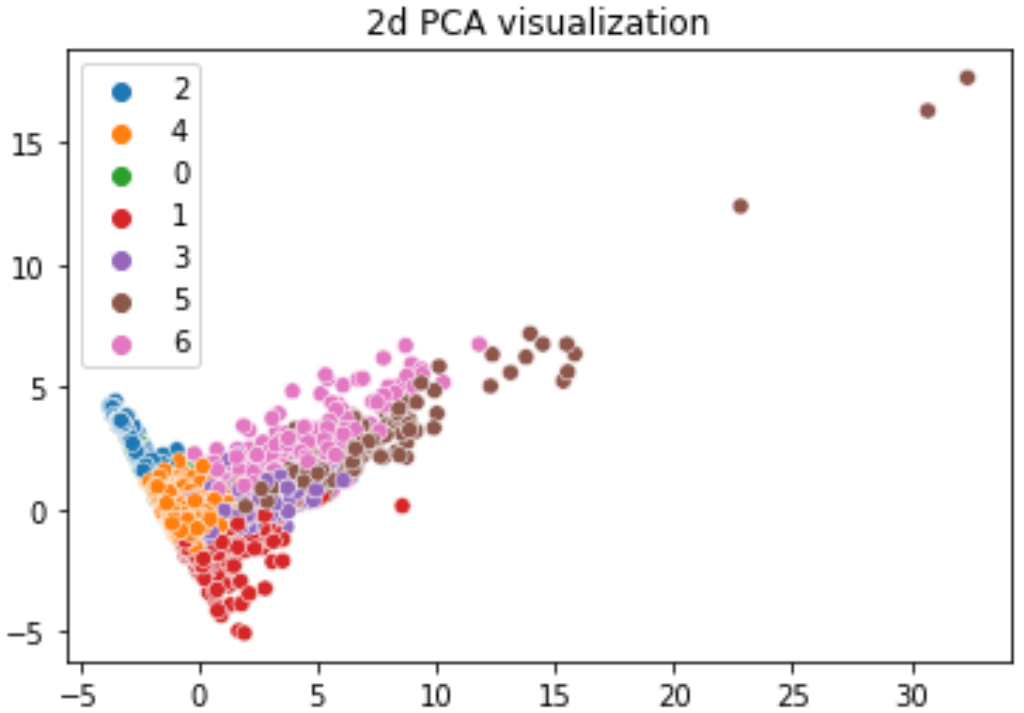


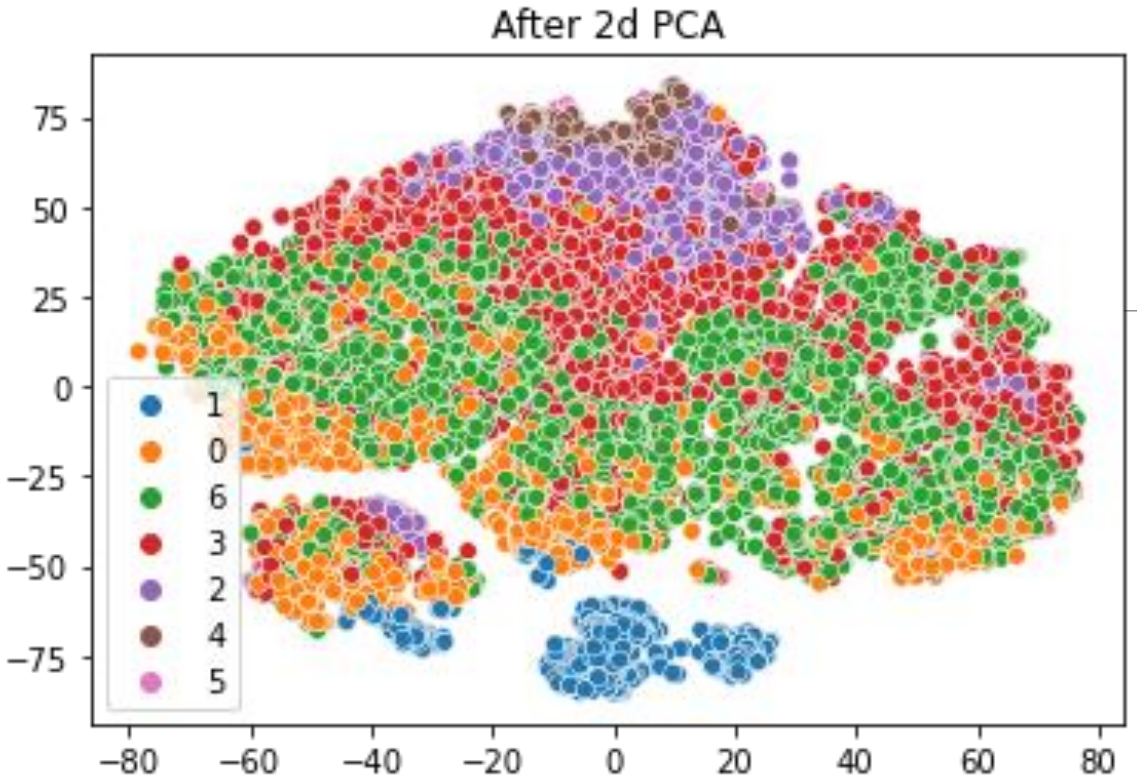
K-means clustering

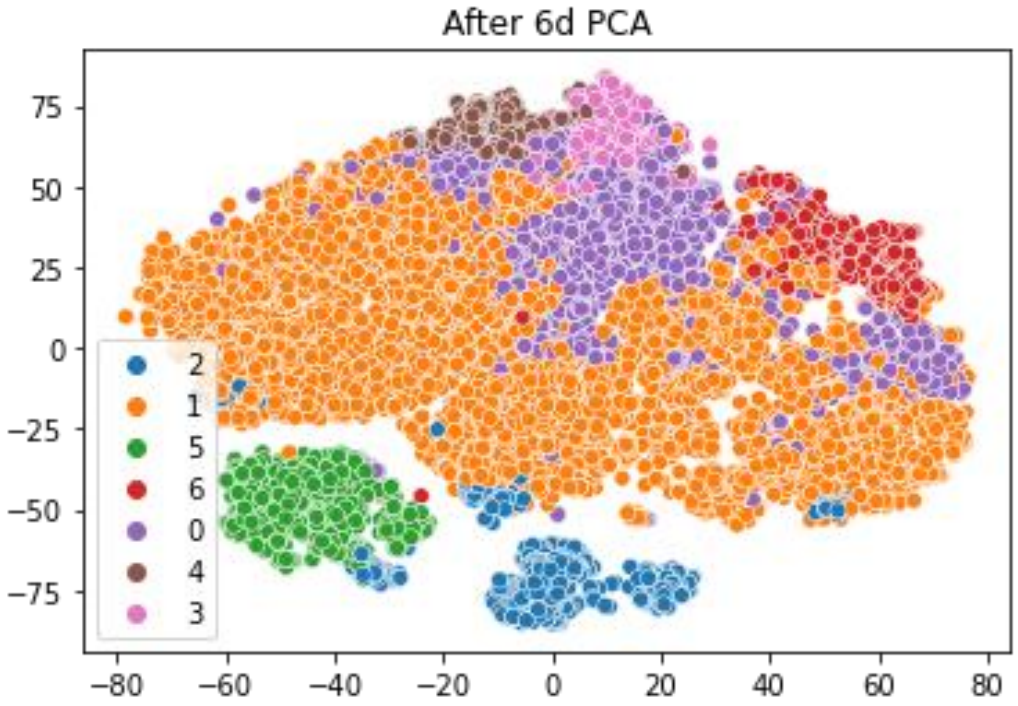

Standarized data. Perplexity = 35
Revenue
False
True
75
50
25
0
−25
−50
−75
−80
−60
−40
−20
0
20
40
60
80

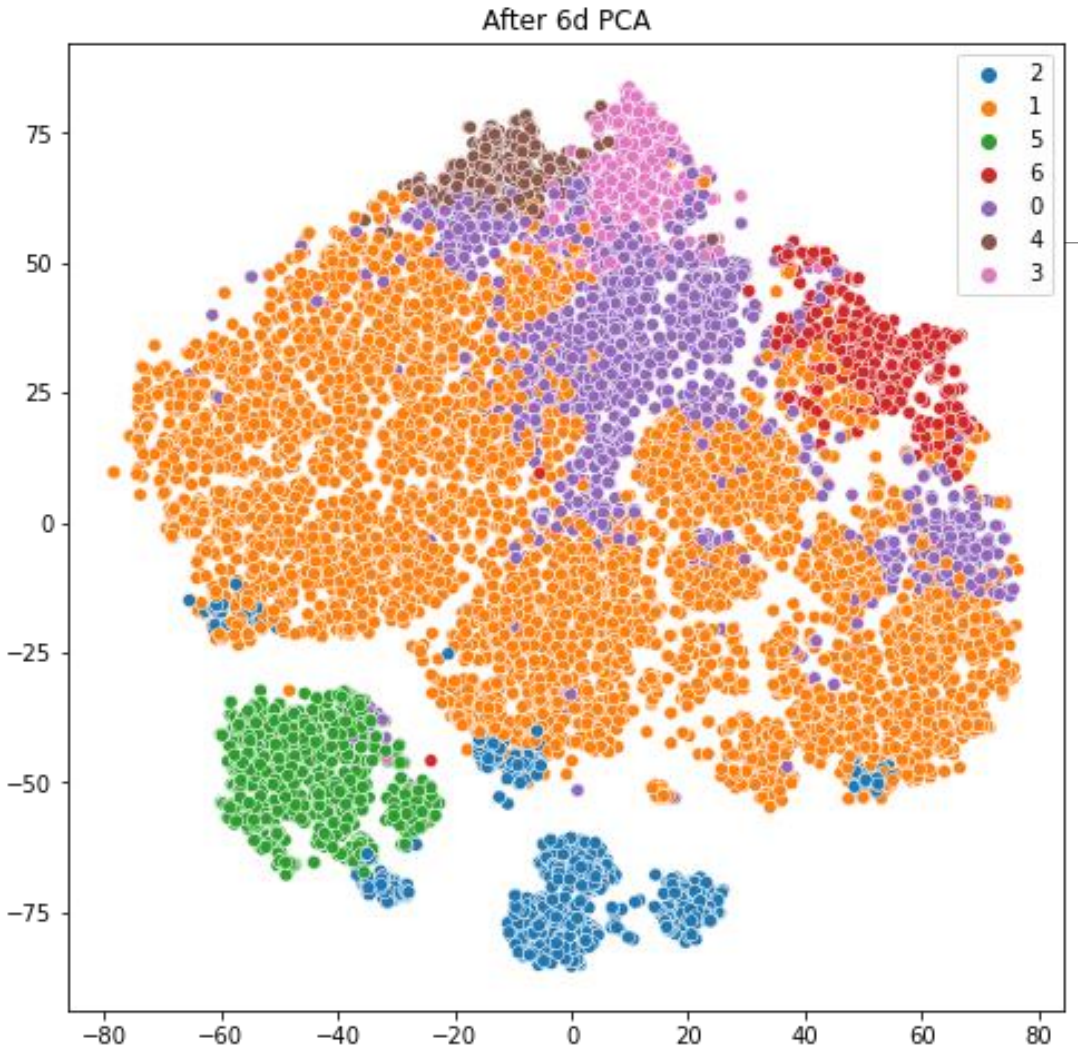

Clustering using all dimensions
2
4
0
1
3
5
6
75
50
25
0
−25
−50
−75
−80
−60
−40
−20
0
20
40
60
80

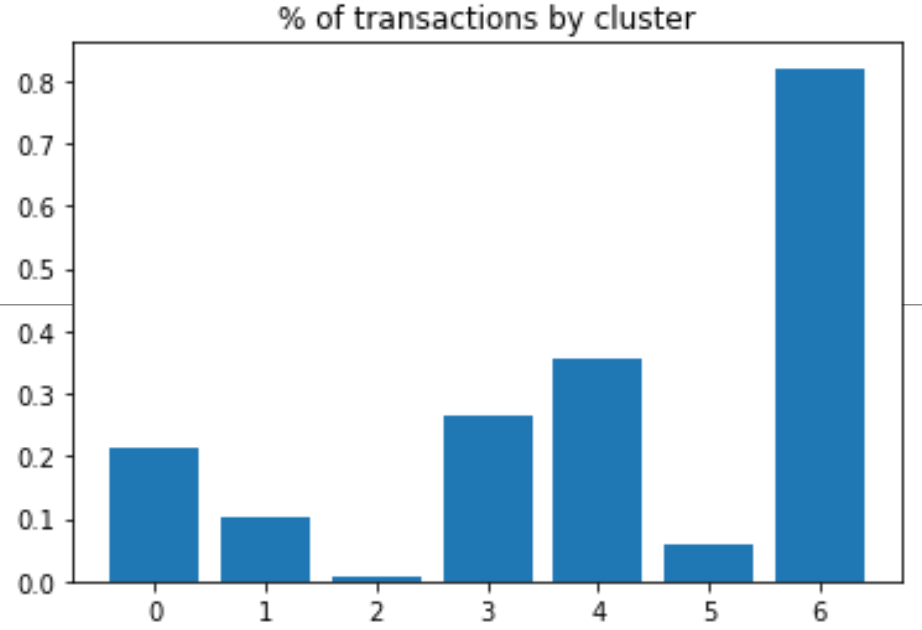

PCA explained variance
0.25
0.20
0.15
0.10
0.05
0.00
2
4
6
8
10

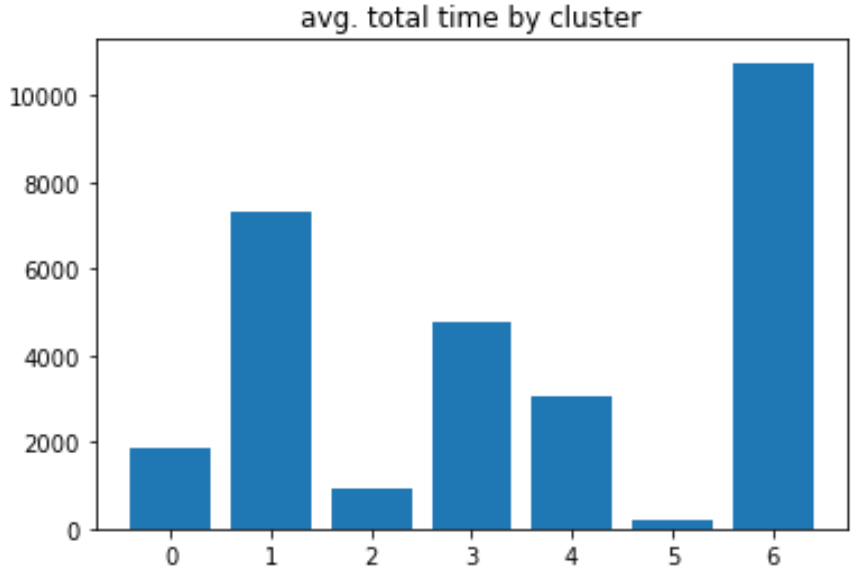

2d PCA visualization
2
4
0
1
3
5
6
15
10
5
0
−5
−5
0
5
10
15
20
25
30


After 2d PCA
75
50
25
0
−25
−50
−75
−80
−60
−40
−20
0
20
40
60
80
1
0
6
3
2
4
5


After 6d PCA
75
50
25
0
−25
−50
−75
−80
−60
−40
−20
0
20
40
60
80
2
1
5
6
0
4
3


After 6d PCA
2
1
5
6
0
4
3
75
50
25
0
−25
−50
−75
−80
−60
−40
−20
0
20
40
60
80


% of transactions by cluster
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
1
2
3
4
5
6


avg. total time by cluster
10000
8000
6000
4000
2000
0
0
1
2
3
4
5
6