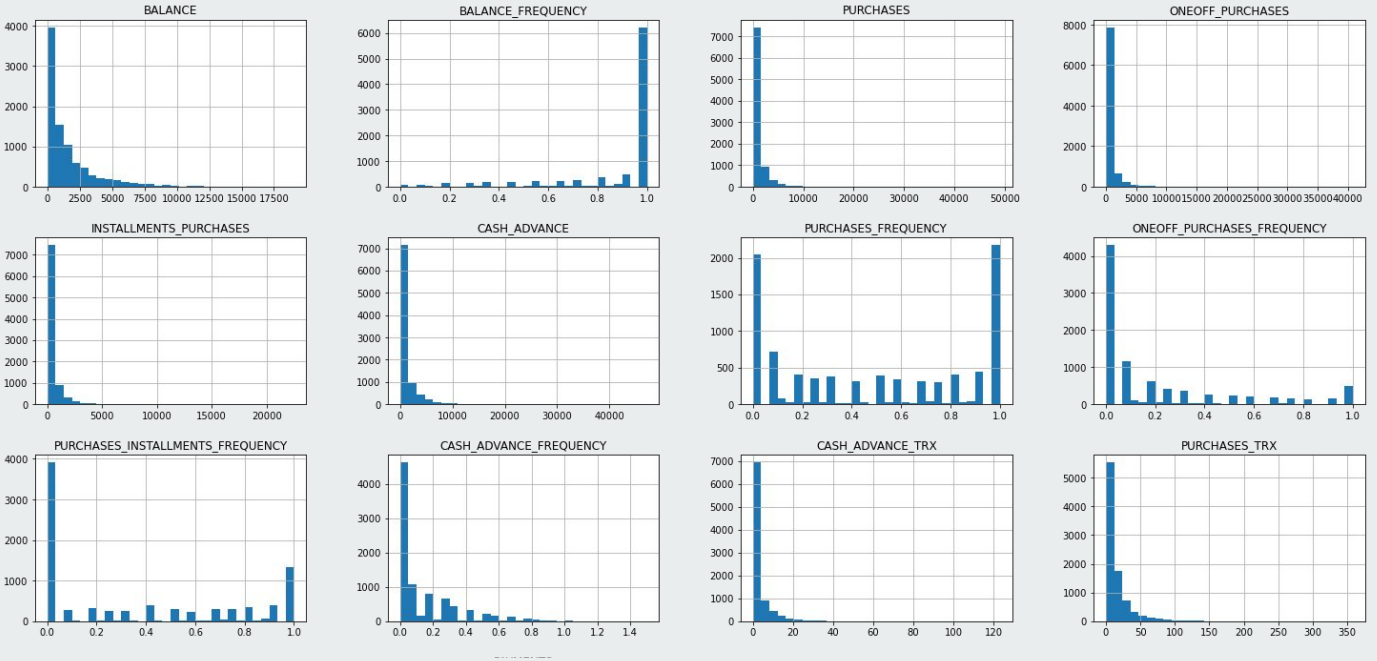
Credit card dataset

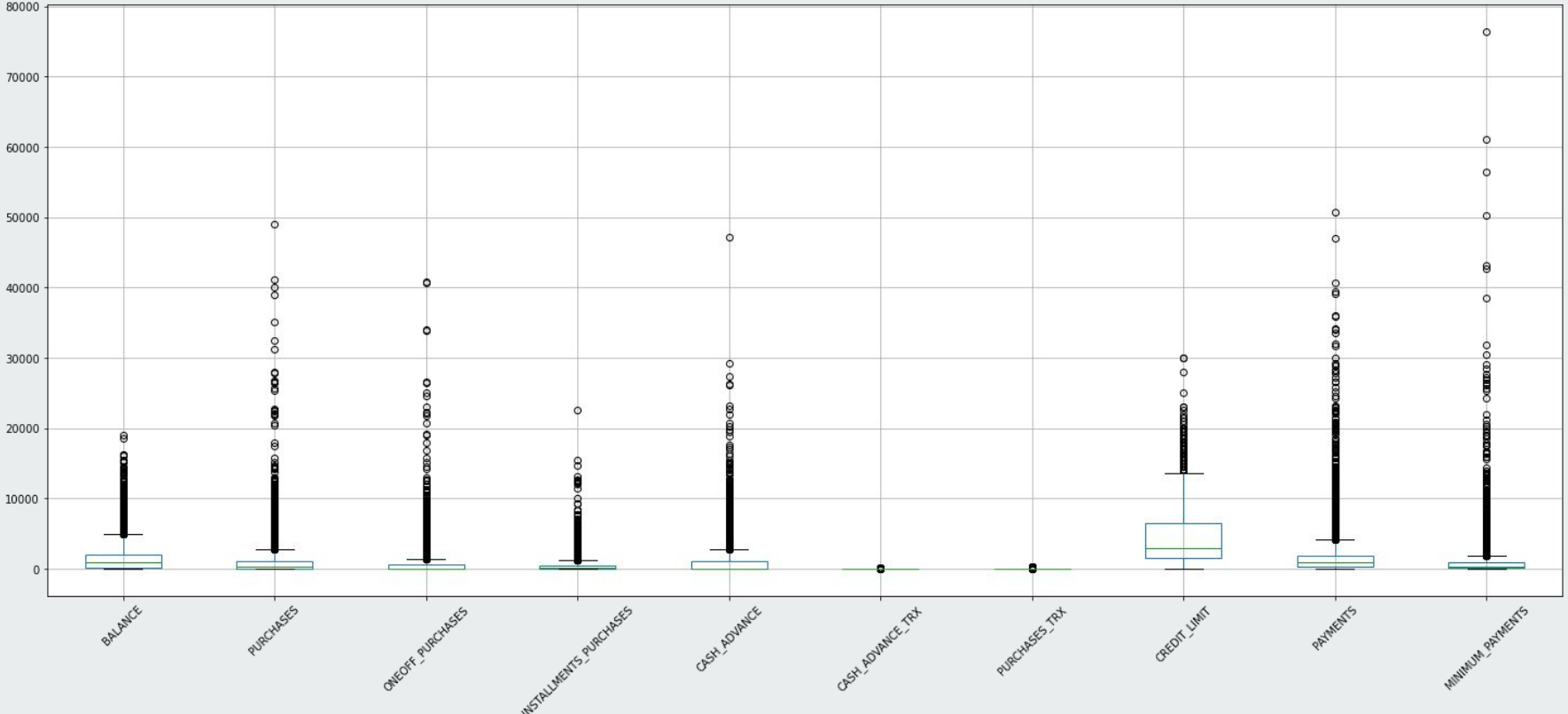
- Range: [0,1] or [0,inf]
- 17 variables
- Clustering and classification



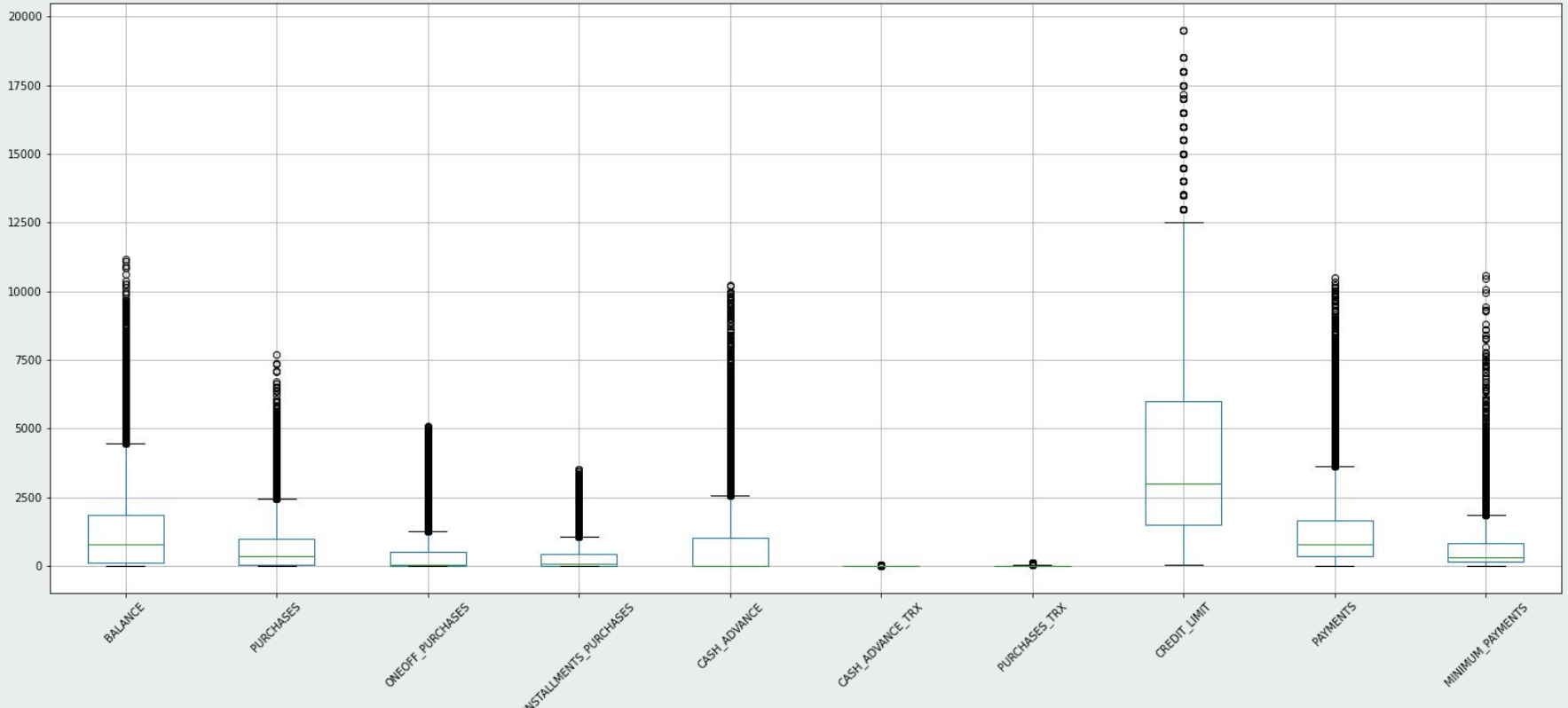
Pre-processing

- Outlier detection and removal
- Correlation matrices
- Normalization of data
- Some sampling

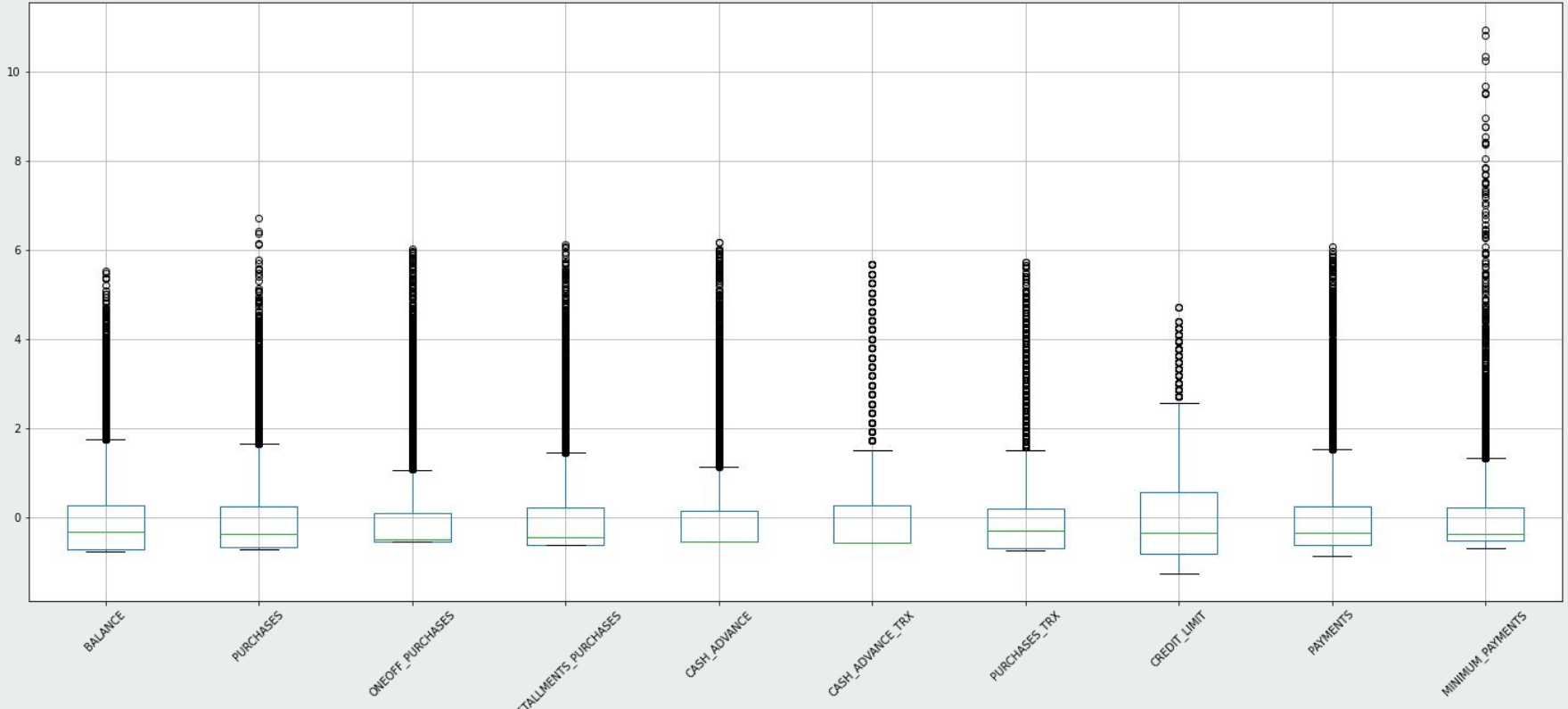
Box plots for raw data



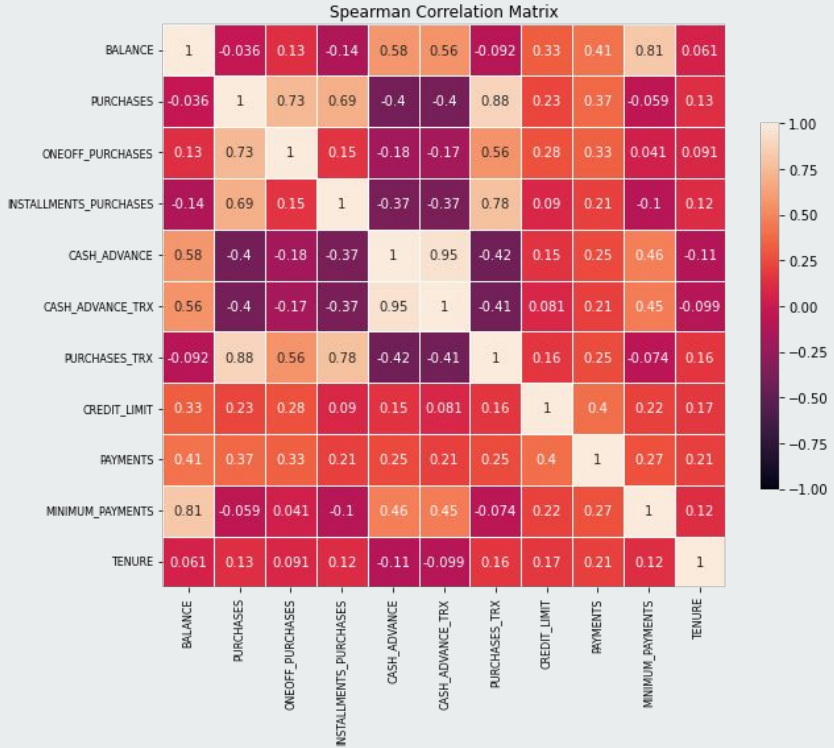
After removing outliers



With z-score normalization

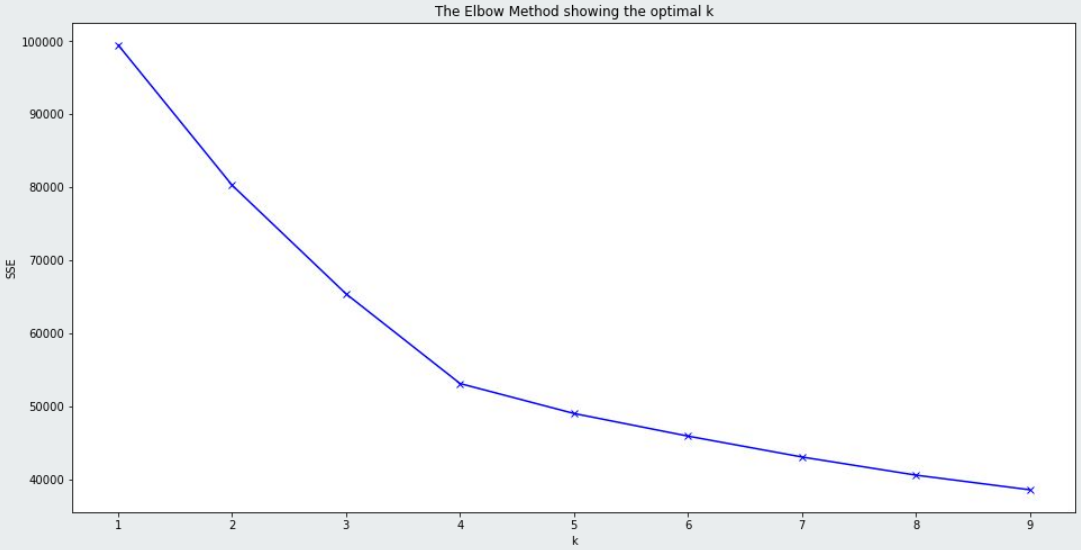


Correlation heatmaps

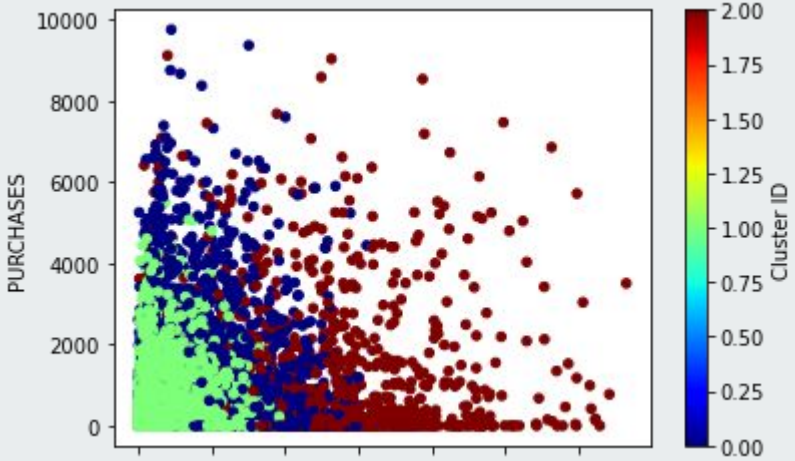


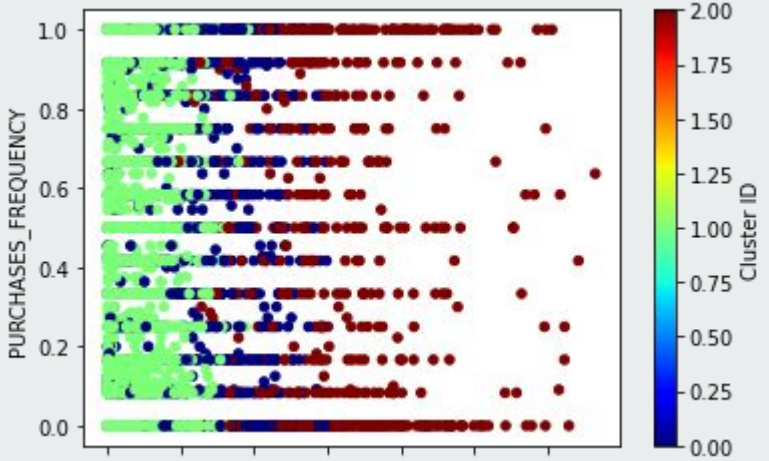
Clustering with K-means

- Elbow method to find optimal k
- Changed when normalizing: 3 → 4



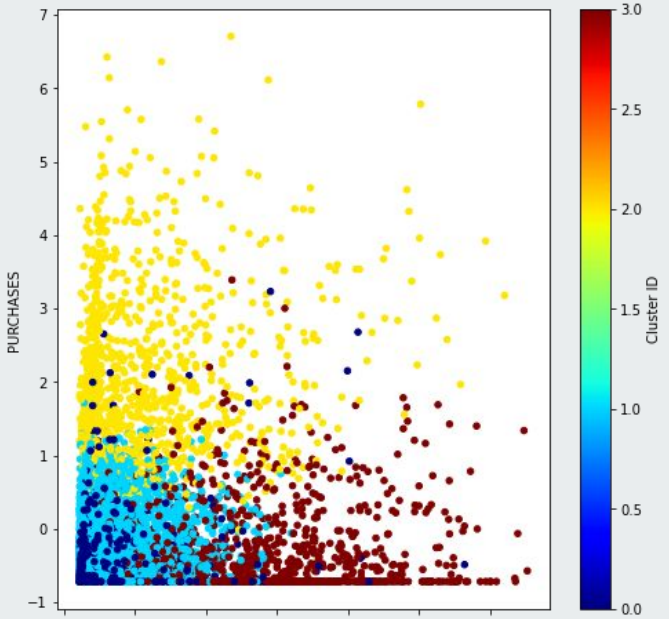
Clustering


PURCHASES
10000
8000
6000
4000
2000
0
Cluster ID
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00


PURCHASES_FREQUENCY
1.0
0.8
0.6
0.4
0.2
0.0
Cluster ID
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00

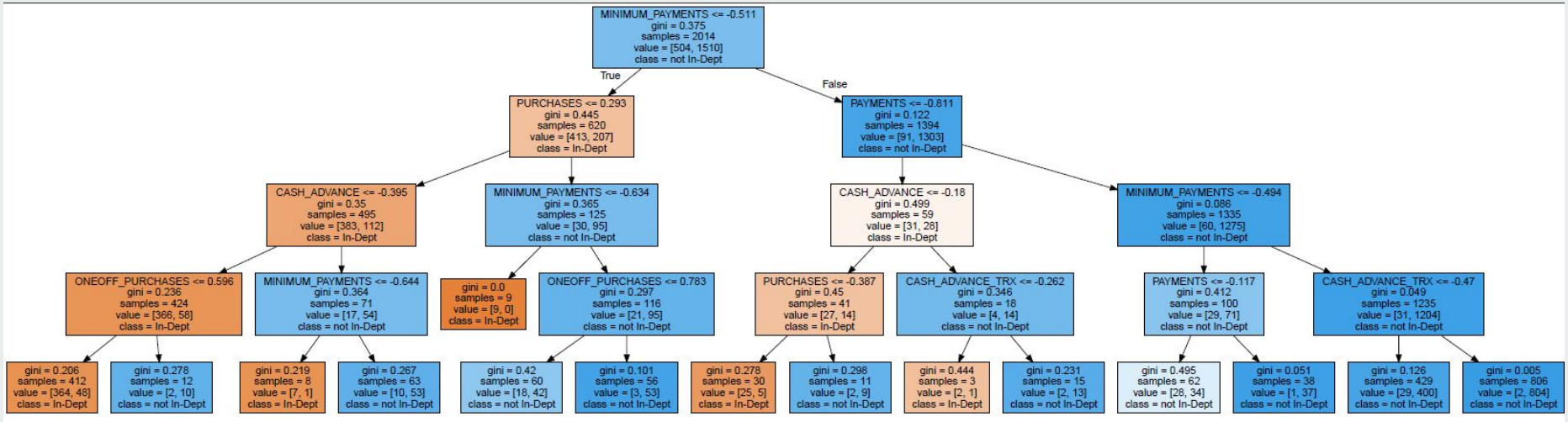
On normalized data

- Optimal number of clusters increased
- Still hard to interpret…



Classification

- decision tree classifier (Accuracy 92.84%)



Thank you

Questions?