### Table of Contents

[**Table of Contents**](#_mvq9go2qnww) **1**

[**Hierarchy clustering Analysis**](#_c3iqb0haw4h5) **1**

[Dealing with Missing Value](#_m7aj5uqziz9) 2

[Dealing with Outliers](#_88i82mumt4fa) 2

[Grubbs Test outliers](#_ayaa9v74mpax) 4

[Dendrogram](#_gd82dzp5wk30) 6

[Characterize different cluster](#_fa7hd0sw5h3n) 8

[HeatMap](#_44k9lr6d6n5b) 9

[Reference](#_hweg6jcwho7z) 9

### 

### 

### 

### Hierarchy clustering Analysis

**Hierarchical clustering** (also called hierarchical cluster analysis or HCA) is a method of cluster analysis which seeks to build a hierarchy of clusters. Strategies for hierarchical clustering generally fall into two types:

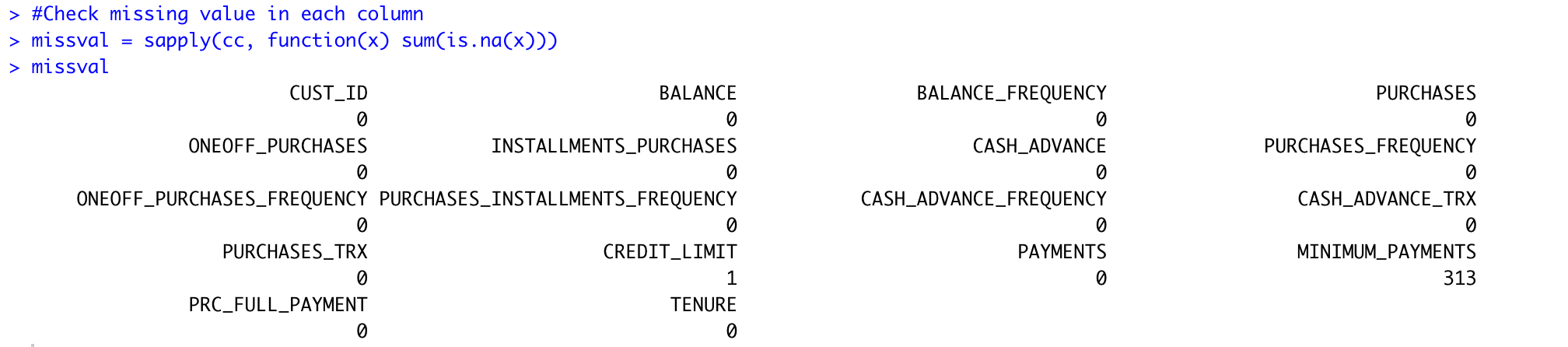
* **Agglomerative** : This is a "bottom-up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.
* **Divisive** : This is a "top-down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

In general, the merges and splits are determined in a greedy manner. The results of hierarchical clustering are usually presented in a dendrogram.

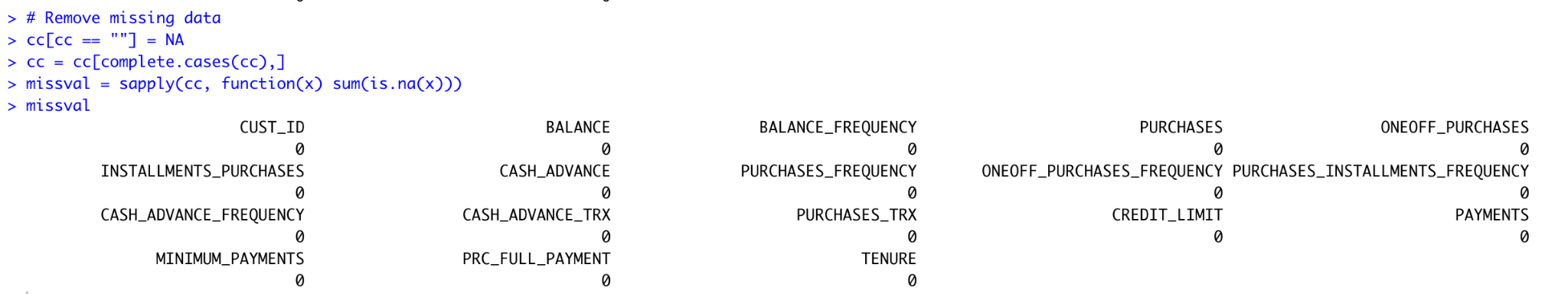
The CC GENERAL.csv data develop a customer segmentation to define marketing strategy. The Dataset summarizes the usage behavior of about 9000 active credit card holders during the last 6 months. The file is at a customer level with 18 behavioral variables.

##### Dealing with Missing Value

As usual we begin by importing the libraries and the data. We will view the dataset and We'll check for missing values.



Even if the missing value is low (the total number of samples is 8950), we will still need to remove it.



In general cases , we will check the Character Variables and Numerical Variables and summarize them to get a good visualization on the dataset.

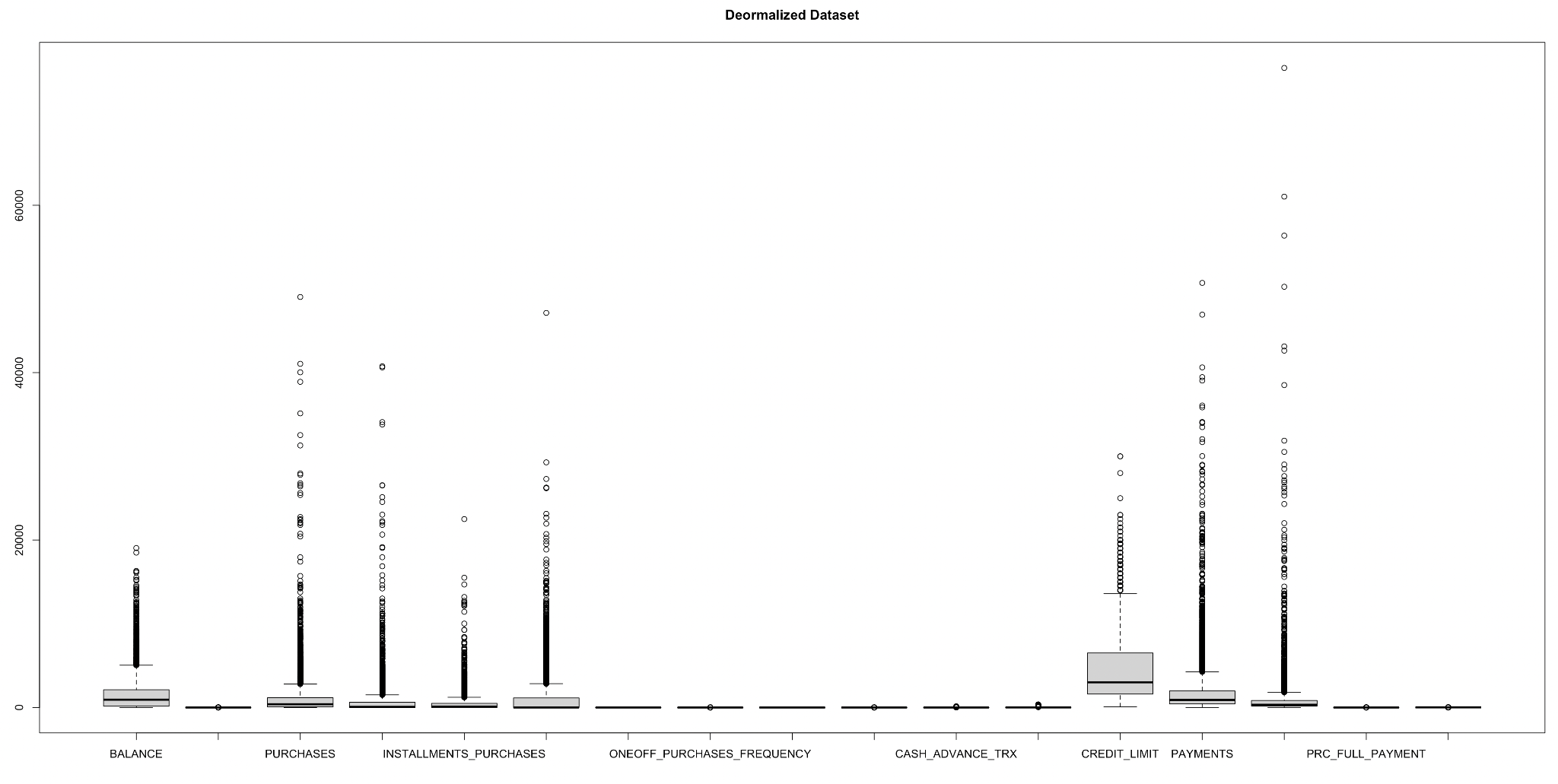
##### Dealing with Outliers

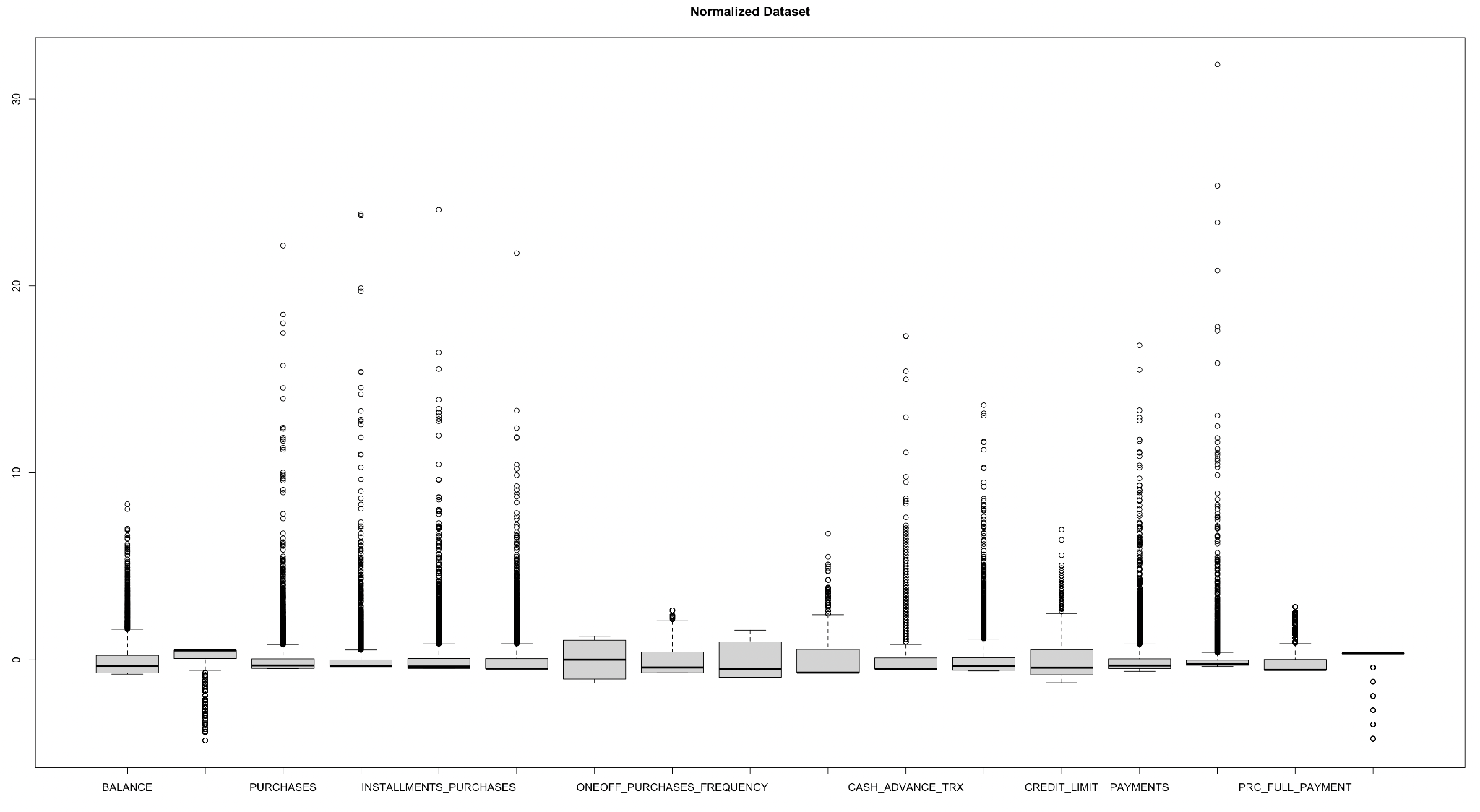
Before dealing with outliers, we are going to remove the character variable which is CUST\_ID.

Next step,We are going to check the outliers on boxplot visualization and also we are going to use the outliers package of grubbs.test function to find the outliers.

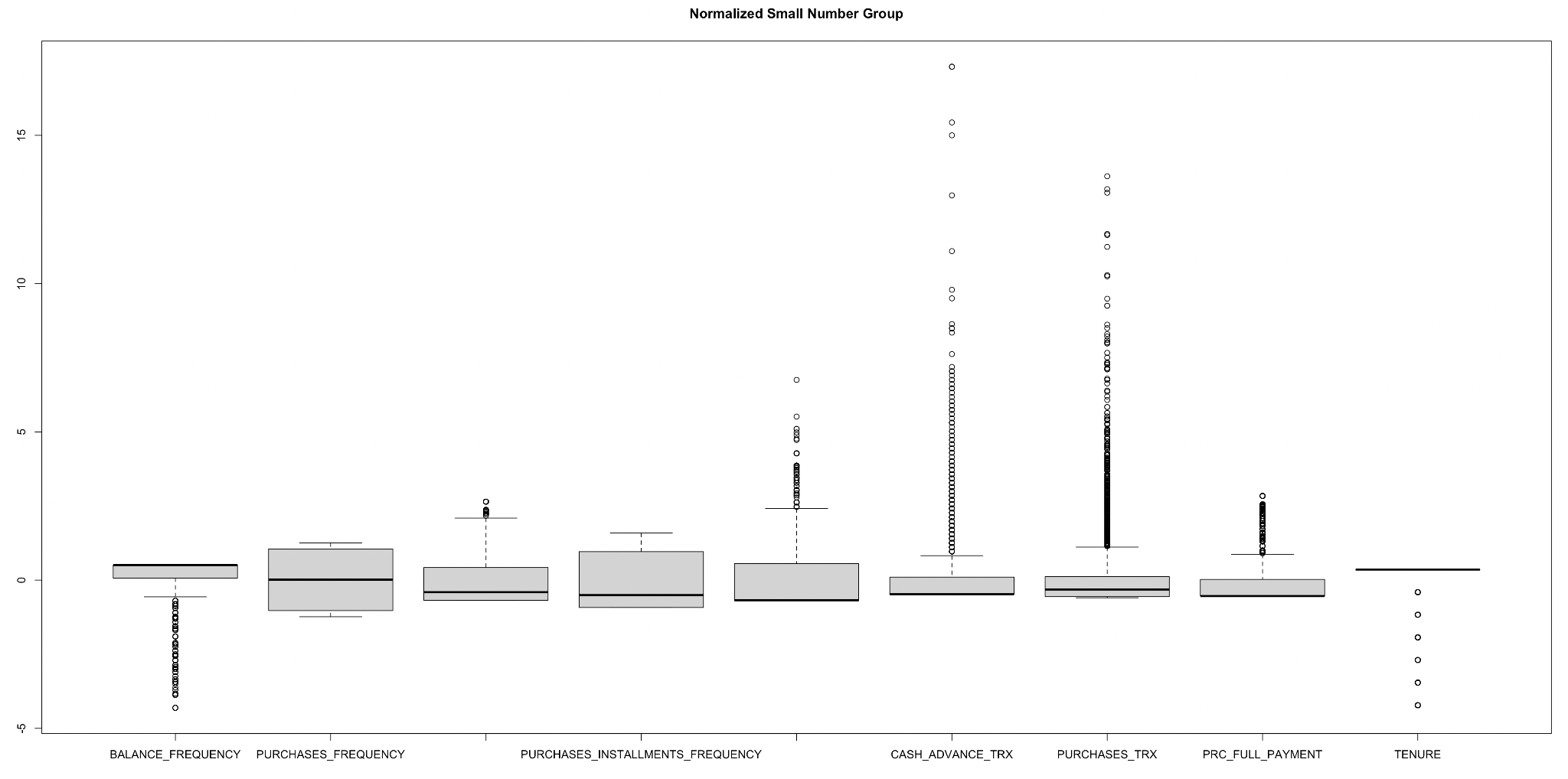
In boxplot we are going to use 3 different ways:

1. Generalize in denormalized data in boxplot
2. Generalize in normalized data in boxplot
3. Since each column has a big difference, we are going to use separate in 2 groups to visualize the dataset.





The results come out from 1 (Generalize in denormalized data in boxplot) and 2 (Generalize in normalized data in boxplot) are shown above.

The results shown below is solution 3 (Since each column has a big difference, we are going to use separate in 2 groups to visualize the dataset.) In this solution, I am using normalized data as illustrated in the boxplot to give more clear visualization, as shown in the box plot, 2 out of 17 of them doesn’t have outliers

##### Grubbs Test outliers

In grubbs.test function allows to detect whether the highest or lowest value in a dataset is an outlier.

The Grubbs test detects one outlier at a time (highest or lowest value), so the null and alternative hypotheses are as follows:

H0 : The highest value is not an outlier

H1 : The highest value is an outlier

if we want to test the highest value, or:

H0 : The lowest value is not an outlier

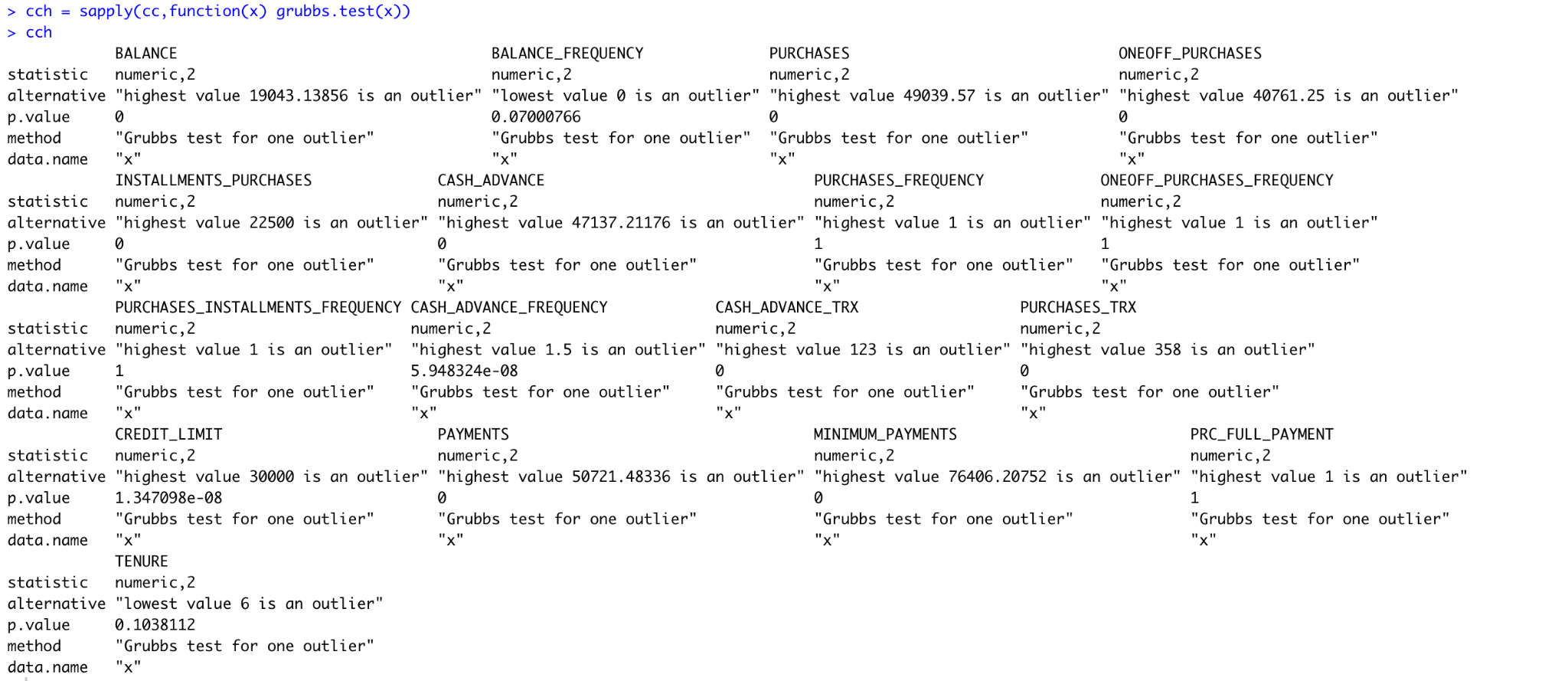
H1 : The lowest value is an outlier

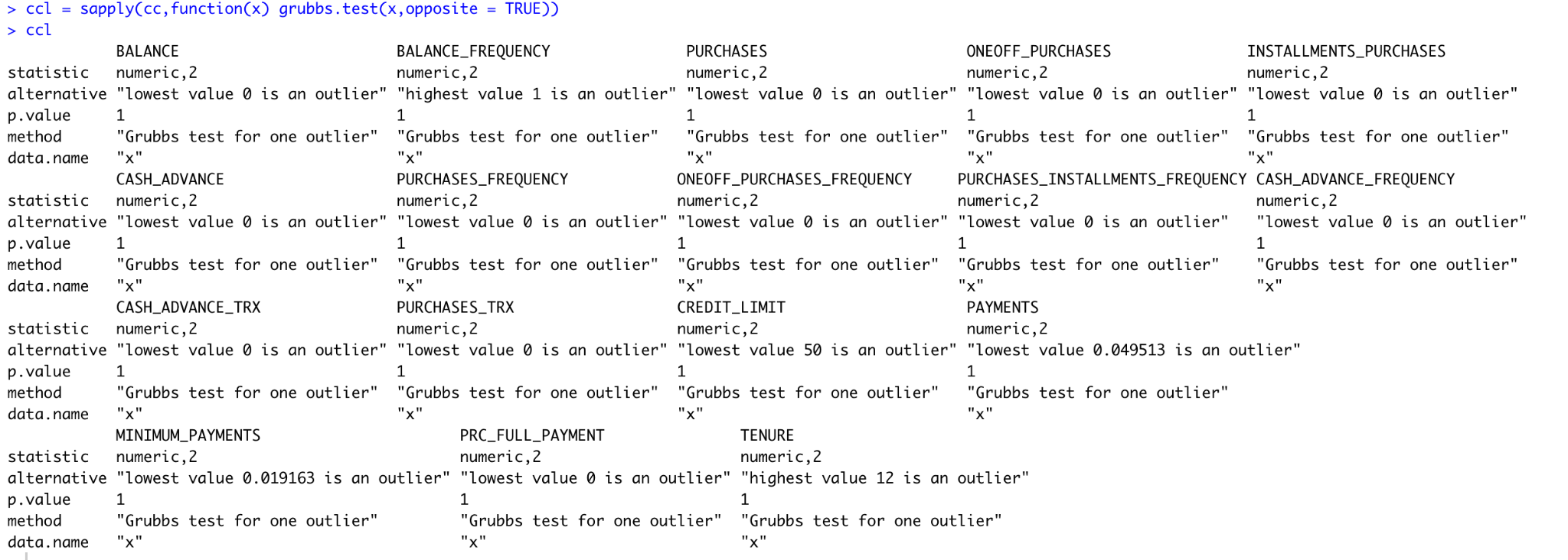
if we want to test the lowest value.

As for any statistical test, if the p-value is less than the chosen significance threshold (generally

α=0.05 ) then the null hypothesis is rejected and we will conclude that the lowest/highest value is an outlier. On the contrary, if the p-value is greater or equal than the significance level, the null hypothesis is not rejected, and we will conclude that, based on the data, we do not reject the hypothesis that the lowest/highest value is not an outlier.

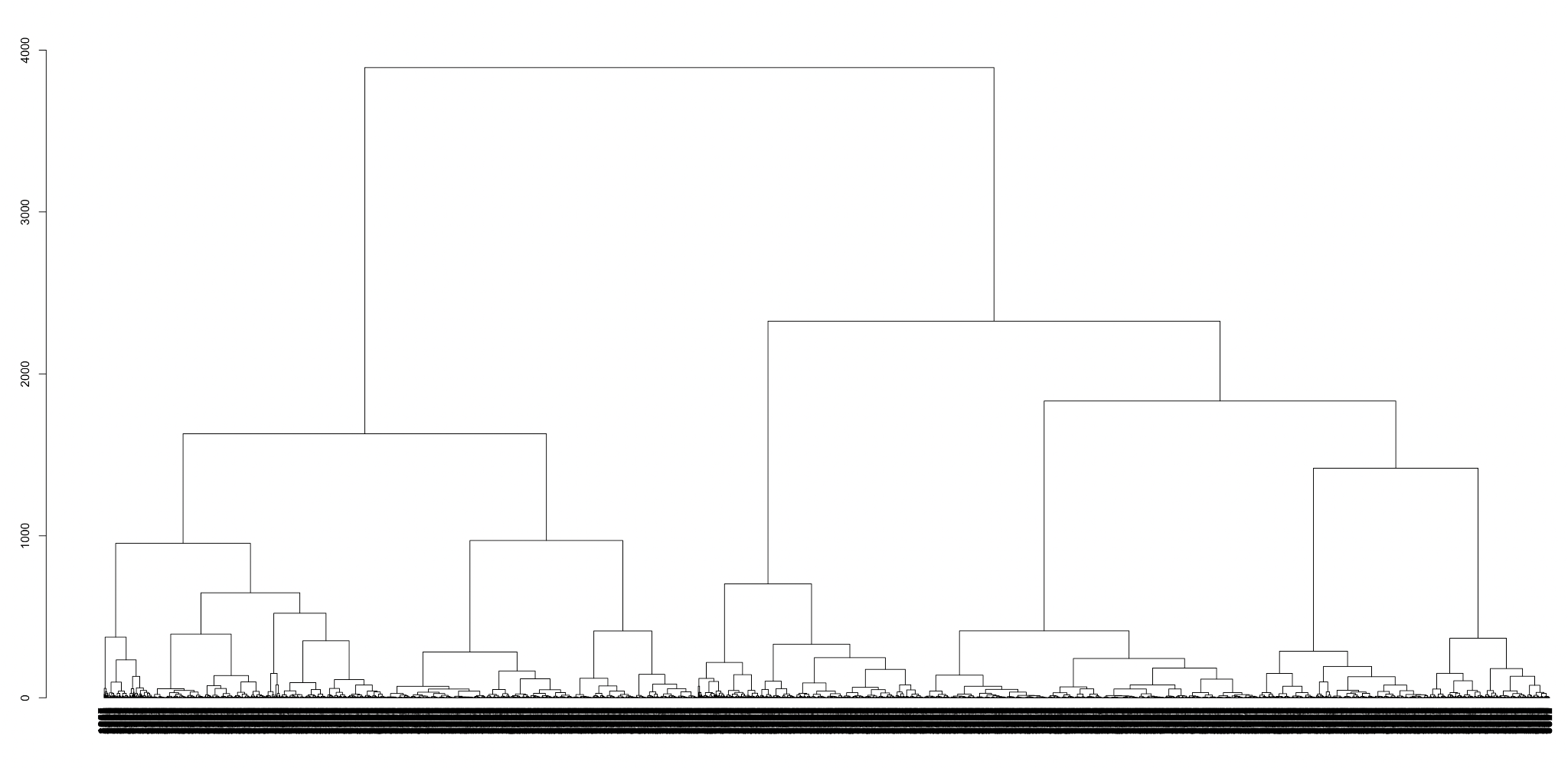
The highest outliers and lowest outliers are shown below.



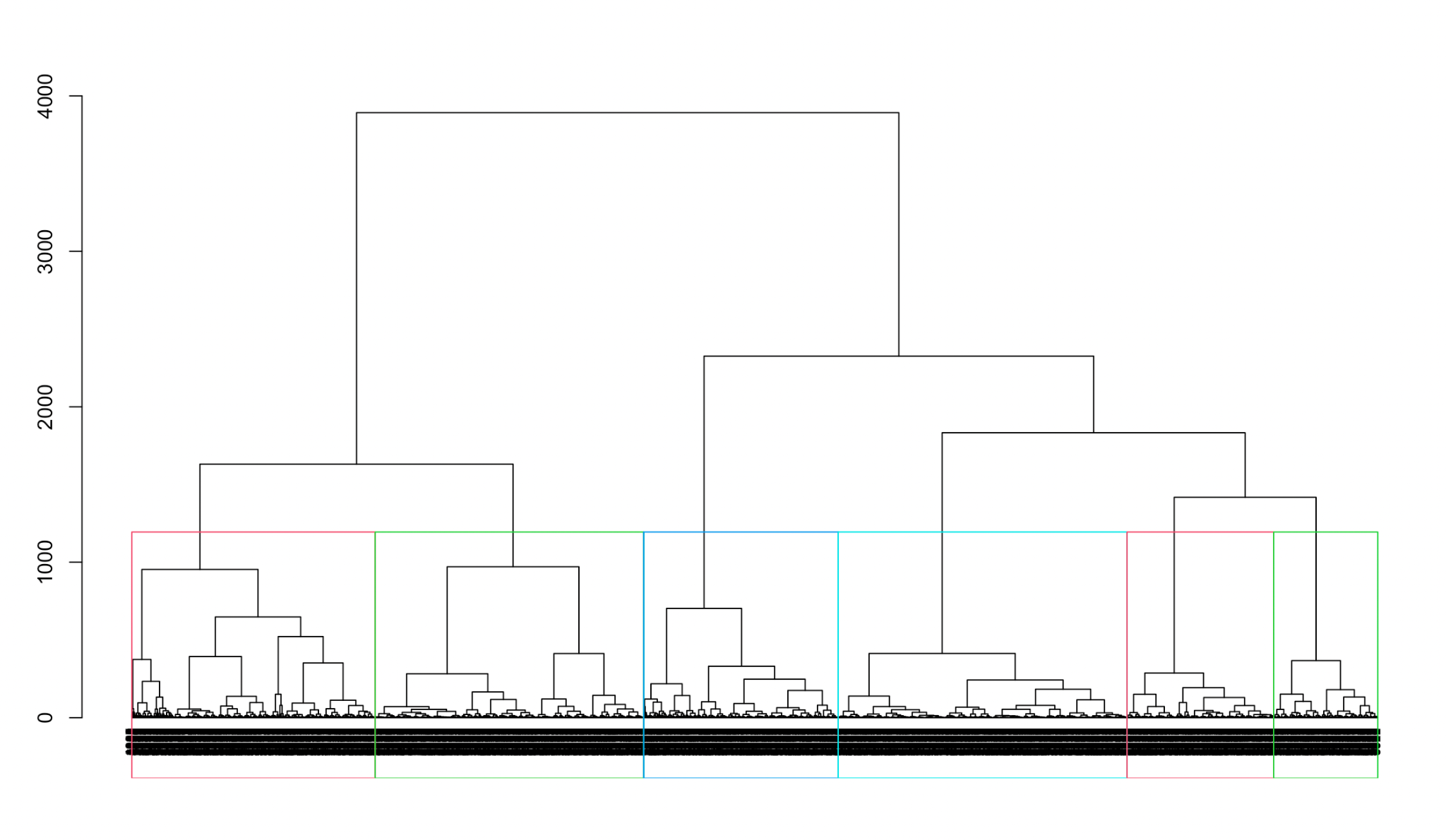


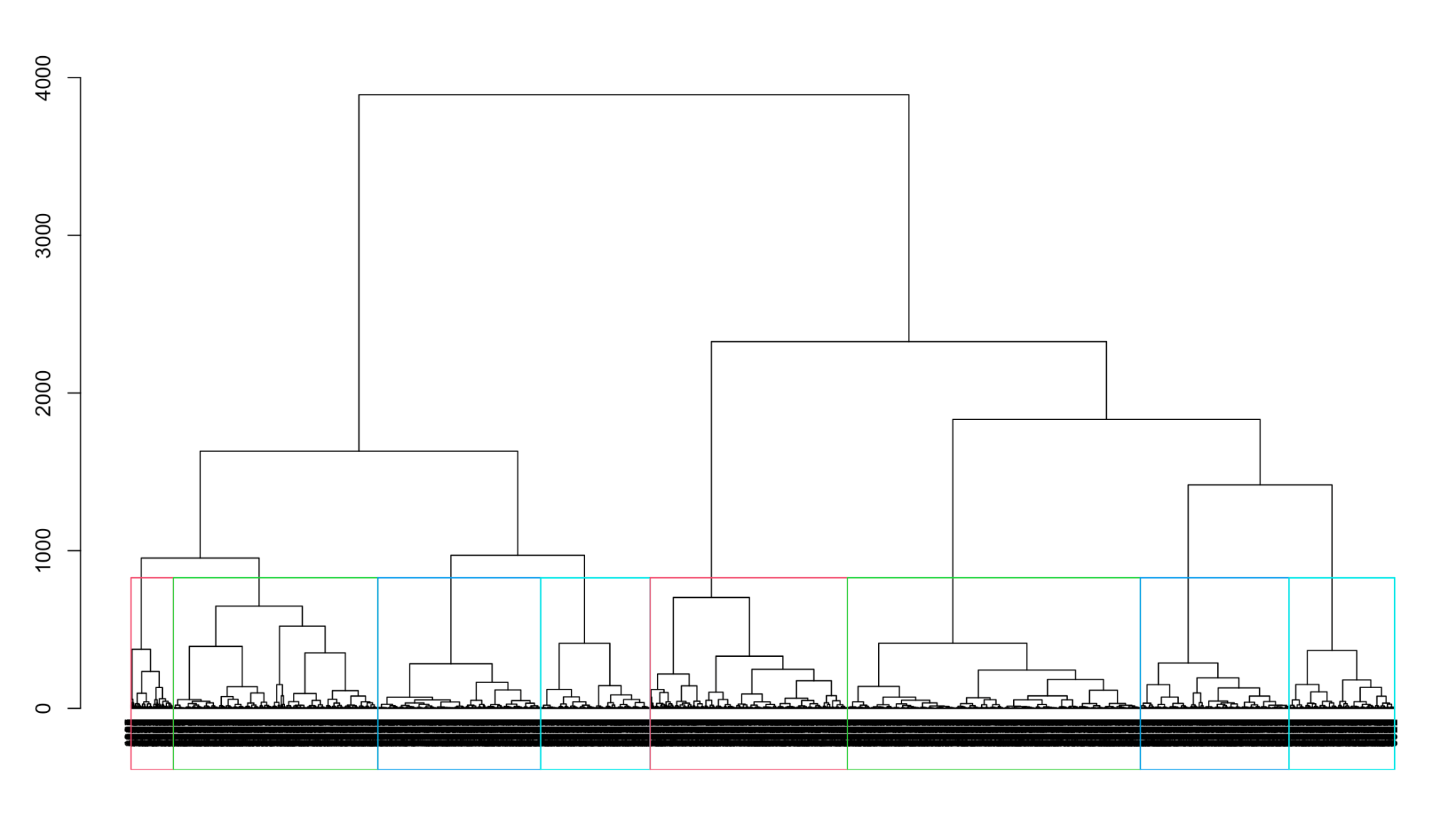
##### Dendrogram

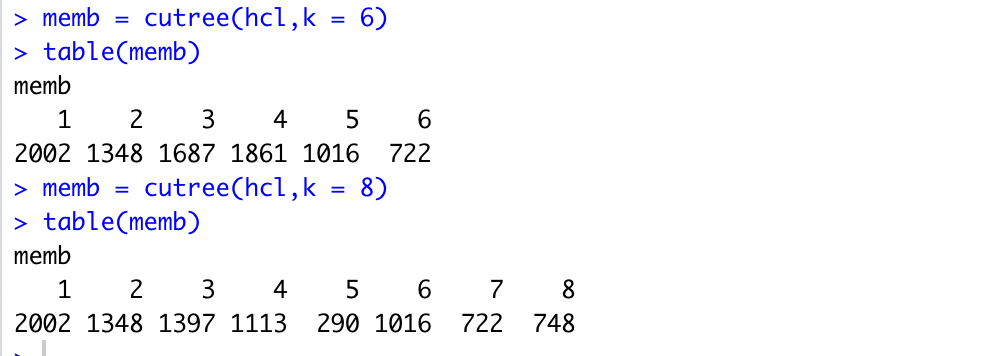
After normalizing the data, then being able to calculate the distances in different methods such as complete, single, average,centroid and ward.D. After testing all the Dendrogram method, the best illustrated method is ward.D, as you can see the result showing below.



Based on the Dendrogram, We have colored the cluster shown below visually. This is 6 clusters and 8 clusters. In 6 clusters, it divides more evenly than 8 clusters. The only difference is first 2 clusters in the first dendrogram, in the first one has 2 at the beginning, but in the second cluster has 4, as you also can see the number difference in cutree function shown below the dendrogram.







##### Characterize different cluster

Both 6 clusters and 8 clusters are works In this case study, but we have to choose one at this point, we will use 8 clusters for the characterize cluster. Both normalized data and denormalized data are going to work. Since this data has over 17 variables , we will export the cluster center to the csv file in order to do better analysis.

according to the data in the excel, we found some interesting relationships in this dataset as following:

PURCHASES\_TRX and PAYMENTS appear highly correlated

PURCHASES and PAYMENTS appear highly correlated

BALANCE and CREDIT\_LIMIT appear highly correlated

ONEOFF\_PURCHASES and PURCHASES\_TRX appear highly correlated

INSTALLMENTS\_PURCHASES and PAYMENTS appear highly correlated

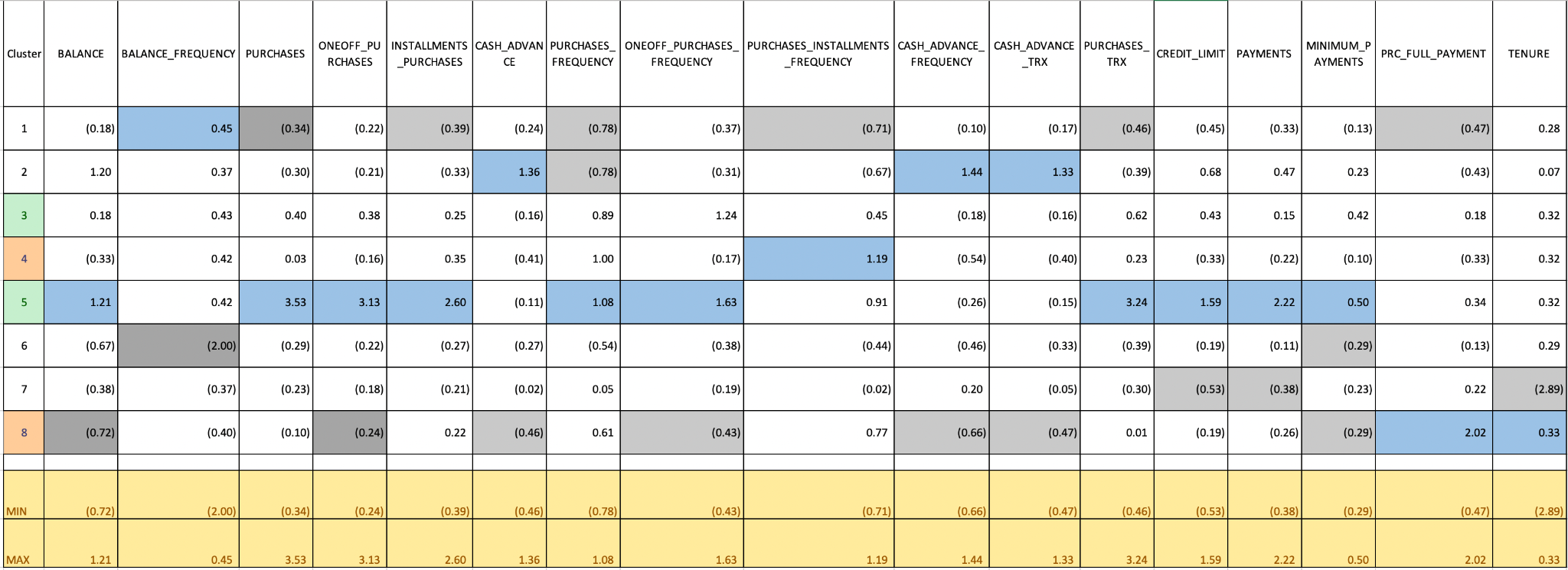
PURCHASES\_FREQUENCY and PURCHASES\_INSTALLMENTS\_FREQUENCY appear highly correlated

PURCHASES\_FREQUENCY appear highly determined by INSTALLMENTS\_PURCHASES

CASH\_ADVANCE appear highly determined by TENURE

CASH\_ADVANCE\_TRX and TENURE appear dependent on each other

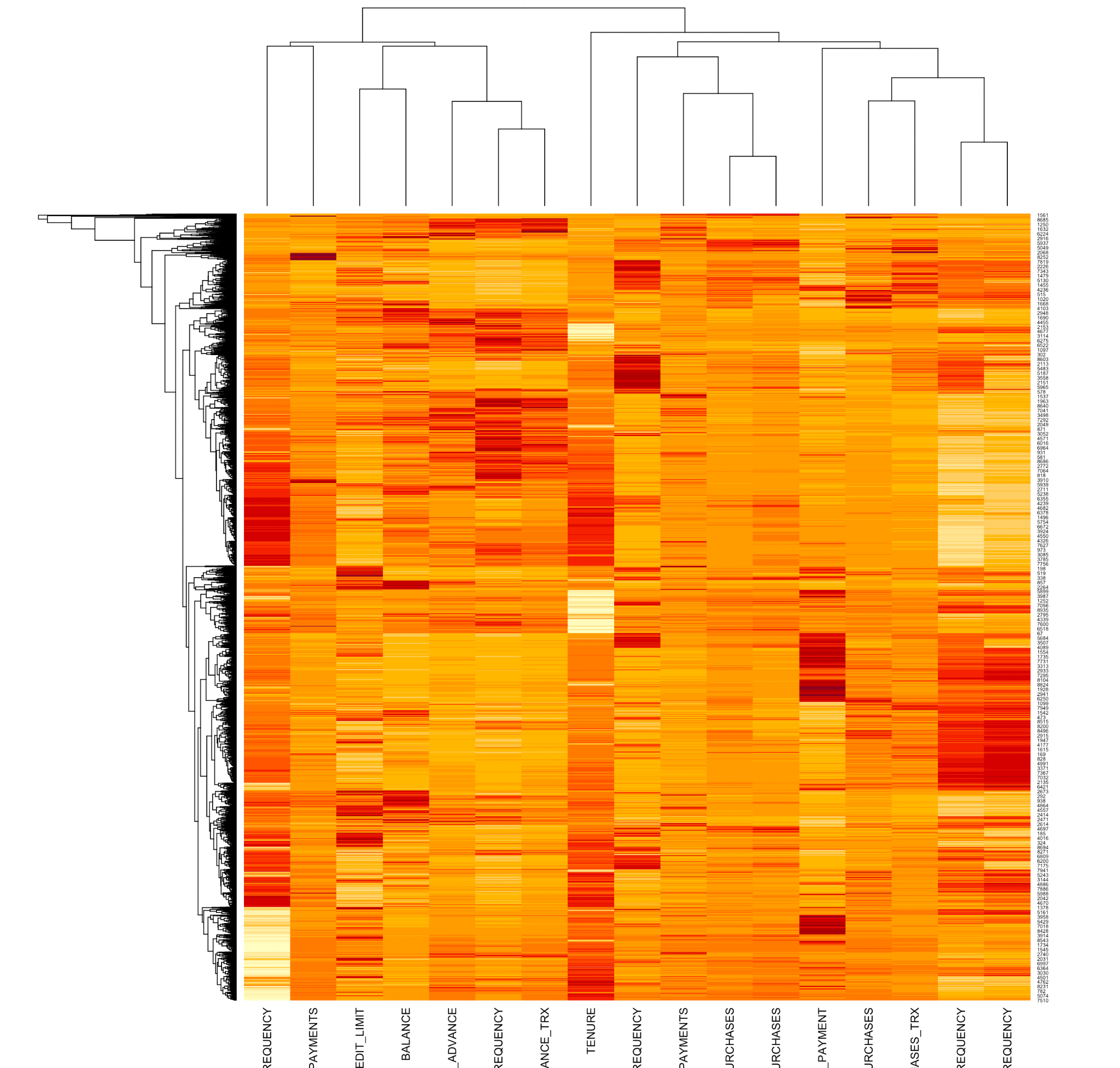
CASH\_ADVANCE\_FREQUENCY and TENURE appear highly dependent on each other



Conclusion :

| **Cluster #** | **Characterize** |
| --- | --- |
| Cluster 1 | Noticeably higher Balance\_Freq and lower Purchases\_Freq |
| Cluster 2 | Noticeably higher Cash\_Advance |
| Cluster 3 | Nothing really special |
| Cluster 4 | Noticeably higher Purchases\_Installments\_Freq |
| Cluster 5 | Noticeably higher Purchase |
| Cluster 6 | Noticeably lower Balance\_Frequency |
| Cluster 7 | Noticeably lower Tenure |
| Cluster 8 | Noticeably higher PRC\_Full\_Payment |

##### HeatMap



A **heatmap** (or **heat map**) is another way to visualize hierarchical clustering. It’s also called a false colored image, where data values are transformed to color scale.

Heat maps allow us to simultaneously visualize clusters of samples and features. First hierarchical clustering is done of both the rows and the columns of the data matrix. The columns/rows of the data matrix are re-ordered according to the hierarchical clustering result, putting similar observations close to each other. The blocks of ‘high’ and ‘low’ values are adjacent in the data matrix. Finally, a color scheme is applied for the visualization and the data matrix is displayed. Visualizing the data matrix in this way can help to find the variables that appear to be characteristic for each sample cluster.

##### Reference

Antoine Soetewey 2020-08-11 ,Outliers detection in R

[John](https://www.r-bloggers.com/author/john/),January 19, 2020,How to Remove Outliers in R

Heatmap in R: Static and Interactive Visualization,https://www.datanovia.com/en/lessons/heatmap-in-r-static-and-interactive-visualization