**Project Final Deliverable - Summary**

**The problem to solve:**

In today's world, the right marketing is becoming increasingly important. It is not only about working out the right marketing campaign, but also presenting it to the right target group. The target groups of a company must be determined through analysis and categorization. The difficulty is not to work out too many small customer segments, but on the other hand to still have larger homogeneous groups. One way to cluster customers into different groups is unsupervised machine learning. Our goal is to identify different customer groups for the selected credit card data set, which can then be used by the company for targeted marketing campaigns in the future.

**The data set:**

We chose the Credit Card [dataset](https://www.kaggle.com/arjunbhasin2013/ccdata) from Kaggle, which contains transactions about 9000 active credit card holders over 6 months. The dataset contains 8950 rows and 18 columns, and the columns contain information about the associated ID, balance and various transaction information. There are 17 variables that are numerical, and one object variable in the data set.

The reason we chose this dataset is because a big challenge for marketers is to understand whom to sell to. When you know the buyer’s personality, you can adjust your positioning and products to increase their satisfaction and earnings. When you already have a customer pool and have enough data, it is very useful to segment them. Therefore, this dataset can help us extract customer segments based on customer behavior patterns provided in the data to focus the company’s marketing strategy on specific segments. For details about EDA and Data Cleaning see deliverable 2.

**Proposed analysis methodology:**

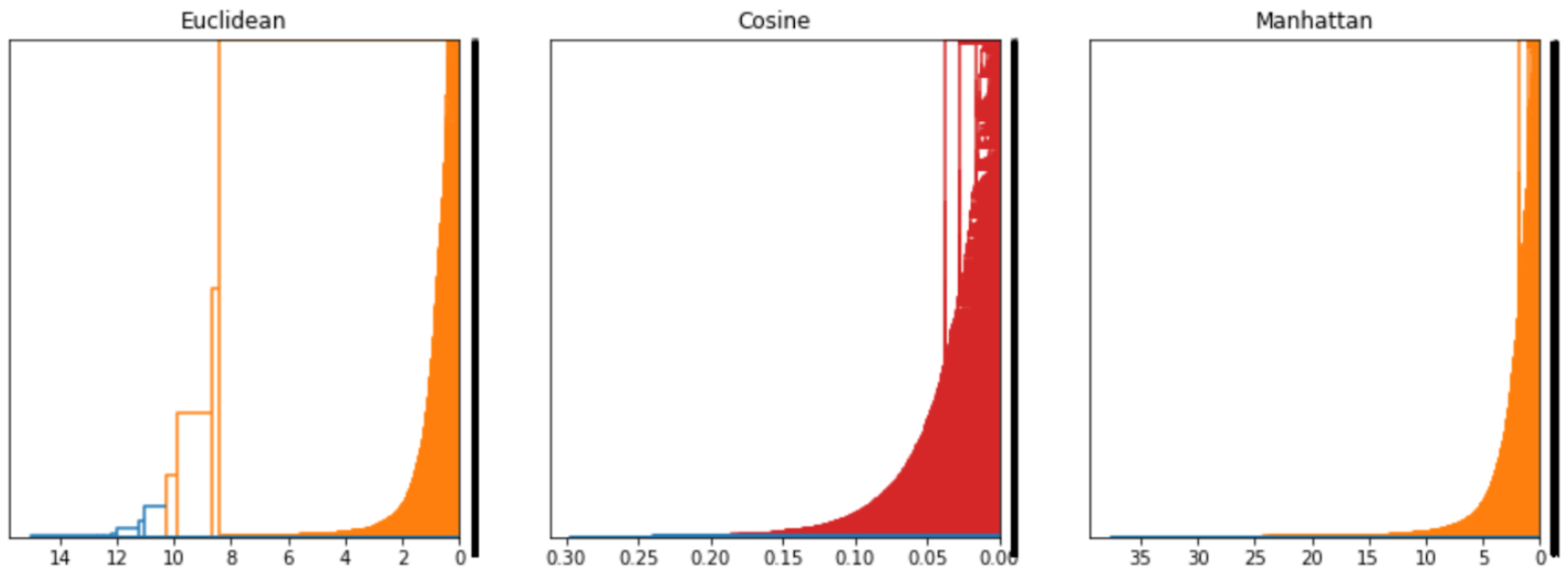
We plan to use the learning obtained from the class such as PCA, Standard Scare, Kmeans clustering, Elbow method, Hierarchical clustering and Silhouette analysis to conduct the analysis.

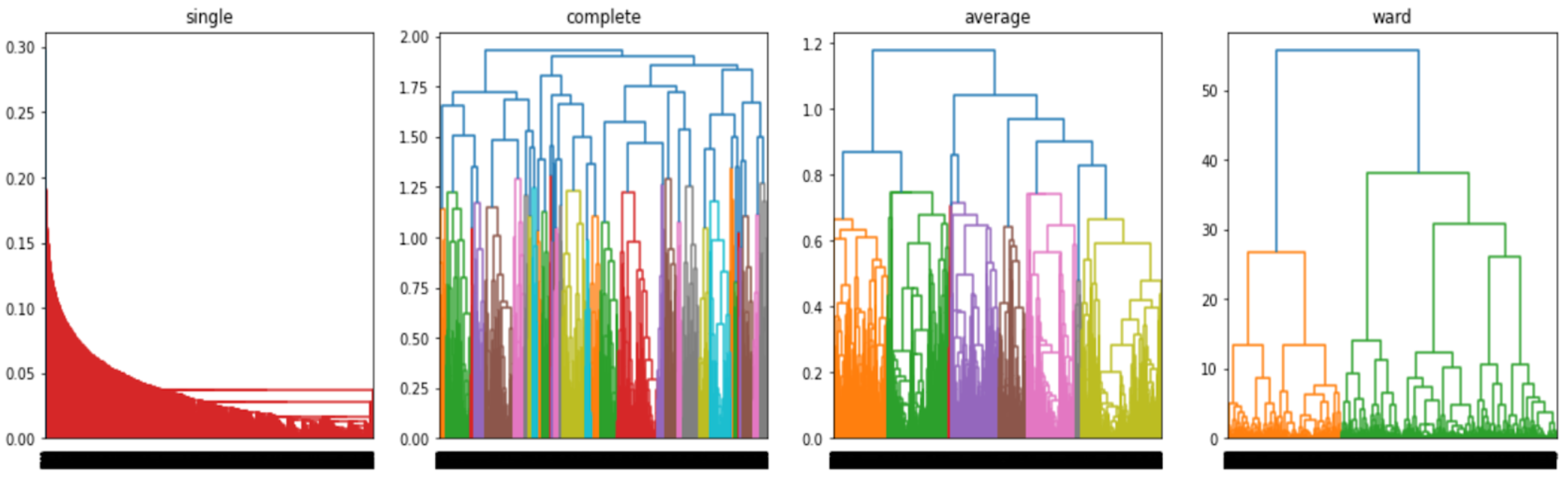
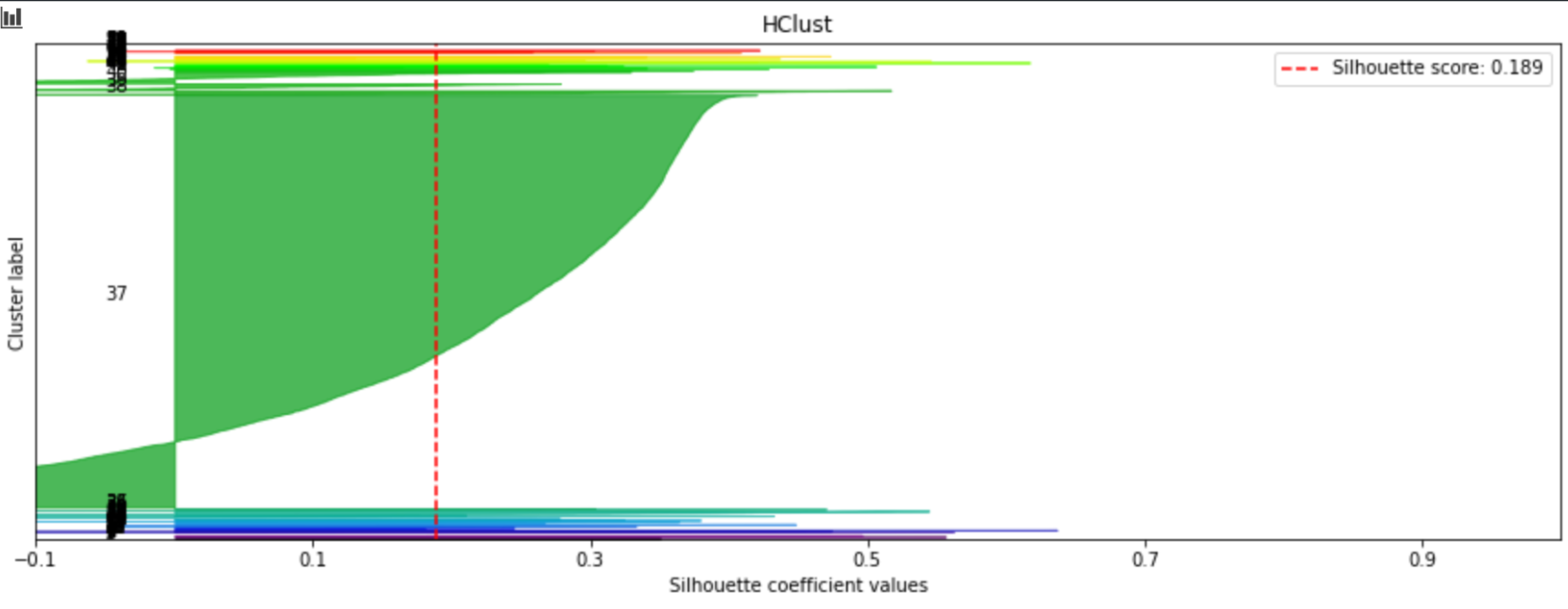
1. Data cleaning includes removing missing values and using dummy variables where possible. Also, we will look into the distributions and correlations of each variable. Data standardization and PCA might apply for the dataset depending on the distance of each variable.
2. EDA to determine structures and get to know the data itself.
3. Analyze and interpret data to understand trends in customer segments. Here we will determine the strength of our model(s) by looking at the following metrics:
4. Hierarchical clustering. Hclust is an algorithm for grouping similar objects into clusters. It treats each observation as a separate cluster. Then, it repeats the following steps: (1) identify the two closest clusters, (2) merge the two most similar clusters. Repeat this process until all clusters are merged together.
5. Kmeans clustering. It can help us divide n observations into k clusters, where each observation belongs to the cluster with the nearest mean.
6. Elbow method. In cluster analysis, the elbow method is a heuristic method to determine the number of clusters in the data set. Therefore, we will use Elbow methods to decide the number of k in order to do further clustering analysis.
7. Silhouette analysis. Silhouette analysis can be used to study the separation distance between the obtained clusters. It can be used to evaluate Kmeans clustering and Hclust. By analyzing the silhouette score and plot, we will be able to choose the best clustering model for our dataset.

**Hierarchical clustering and KMeans clustering**

Since the goal is to divide the behavioral variables of credit card users into similar groups, we decided to try two clustering techniques – Hierarchical clustering and K-Means clustering, and check which one produces better results.

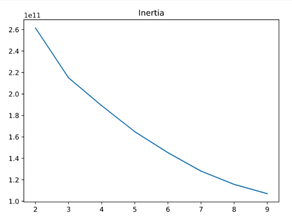
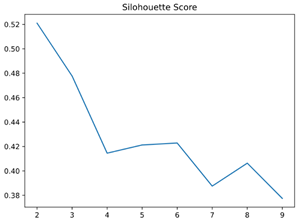
**Hierarchical clustering:**

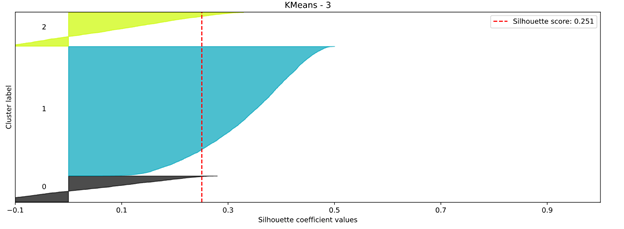
The first model is Hierarchical clustering. The dataset for our group is large, therefore, we will use a dendrogram to visualize the hierarchical relationship. We calculated the distances and then entered it into different linkage methods. We first try linkage methods: Euclidean, Cosine, and Manhattan. The result is shown left which shows that Cosine is the best among these three. And we also noticed that there are outliers in our dataset. Then, we apply the methods of single, complete, average and ward based on Cosine, with the result of graphs shown in below. 

The results of 6 clusters with average methods are the best among others. The reason we chose the average method with 6 clusters is because it forms clusters with a lower y-axis compared with other methods. And each cluster under average method is more obvious and equally distributed.

Then we apply Hierarchical clustering with 6 clusters to fit the silhouette analysis. According to the graph below, it can be known that there are negative values in most clusters. The reason for that is because some of the data goes to the wrong cluster. Also, it failed to distinguish the boundary between clusters. One of the clusters is much bigger than other clusters. The number under each cluster does not occur evenly across all clusters. Finally, the silhouette score for Hierarchical clustering with 6 clusters is 0.189.

**KMeans Clustering:**

Second, we used KMeans Model. KMeans clustering is an unsupervised machine learning model, which groups variables with the same patterns together. Because we have a large dataset, we chose to use PCA for our dimensionality reduction. We selected that in the lower dimensions, our new data could explain 95% of variance of our original data. After the PCA, for KMeans clustering, we firstly need to decide the value of k(number of clusters). To achieve our goal, we created both a plot of WSS and a plot of silhouette score. We set the range of 2 to 8 clusters to see what the two curves will look like. From the Elbow method graph, it seems that the best number of clusters is 3. To confirm our choice, we decided to double check with the silhouette score. 

The silhouette score plot shows that, when k is equal to 2, 3, 4 or 6, we may get a good result with the KMeans Model. Because we think 2 clusters are too little and we can’t effectively distinguish the different groups’ characteristics of credit card users, we just ignore the situation of k = 2. We fitted the remaining 3 k values one by one to our dataset after PCA and got that setting the k = 3 gives us the highest silhouette score, which is also an optimal k value with the Elbow method. In this case, we finally decided to select 3 as our number of clusters and get the KMeans Metric below:

From this metric, we can see that the number of credit card users in one cluster is not that equal. Most of the users are lying in cluster 1. Also, there are some negative values in the graph, which means that some data belonging to clusters may be incorrect. In addition, we can take a look at the silhouette score, which is 0.251 and pretty low. This means that the detected clusters are weak. k=3 is still our best result with KMeans clustering.

**Comparison of Hierarchical Clustering and KMeans:**

Now, having already explained individually the results of Hierarchical Clustering and KMeans, we would like to compare the results of these methods and select the best one.

For our first method, Hierarchical Clustering, we were able to discover 6 clusters. However, these clusters are partly unequal and have only a silhouette score of 0.189, which indicates that no meaningful clusters could be found.For KMeans, we found 3 clusters to be the best option and achieved a Silhouette Score of 0.252. This score indicates that the clusters found are not significant and are weak. Nevertheless, the score for KMeans is better than that for Hierarchical Clustering, which is why KMeans is the better method for our analysis.

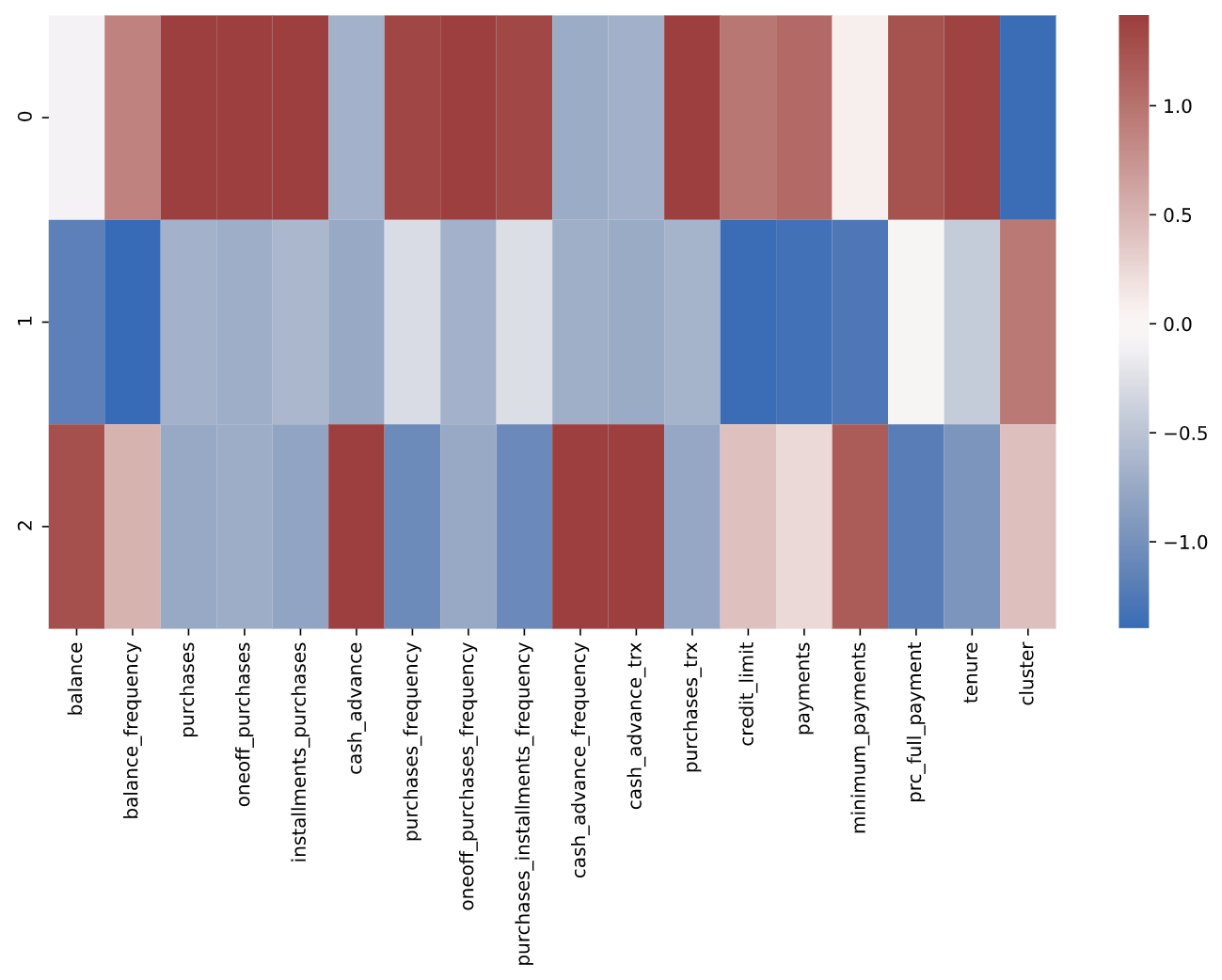
The question is why the algorithm has such poor performance. Perhaps the data would have to be collected over a longer period of time. Or it would be better if the data had been collected differently. For example, the individual transactions could have been assigned to a specific date and not all transactions summarized to the customers. This would reveal temporal patterns. Likewise, the cleaning process should be looked at again and possible other measures should be taken. The collected data could be better suited for detecting fraud attempts, as outliers can be examined more closely for this purpose. In summary, this dataset yields only limited (good) results for unsupervised machine learning. Nevertheless, in our analysis KMeans is clearly better than Hierarchical Clustering.

**Conclusion and recommendations:**

After choosing the KMeans clustering method, we assign each credit card to a cluster to view the summary statistics and try to identify the characteristics of each cluster.

* Cluster 0: Has the highest values in purchases, oneoff\_purchases, installments\_purchases, off\_purchases\_frequency, purchases\_trx, and tenure.
* Cluster 1: Has the lowest values in balance\_frequency, credit\_limit, payments, and minimum\_payment.
* Cluster 2: Has the highest values in cash\_advance, cash\_advance\_trx, and cash\_advance\_frequency.

The credit card holders who are in cluster 0 are more likely to spend money and purchases no matter if they pay by one off or installments. The bank can offer and advertise those credit card holders about their credit card purchases reward or cashback to increase their purchases and spending. In cluster 1, the credit card holders seem to have set aside all the money and have not spent much on purchases. The bank may offer them a savings account where they can earn a higher interest rate. Credit card holders under cluster 2 have the highest value in cash advance; therefore, they are more likely to borrow money to spend. The bank can offer them a longer repayment term and reduce the minimum monthly repayment.



**Challenges:**

* Working in a team with different time zones and a total differences of 13 hours (China)
* Especially, because we are working on the Capstone project simultaneously
* Visualizing the dendrogram with a bigger dataset than we saw in class