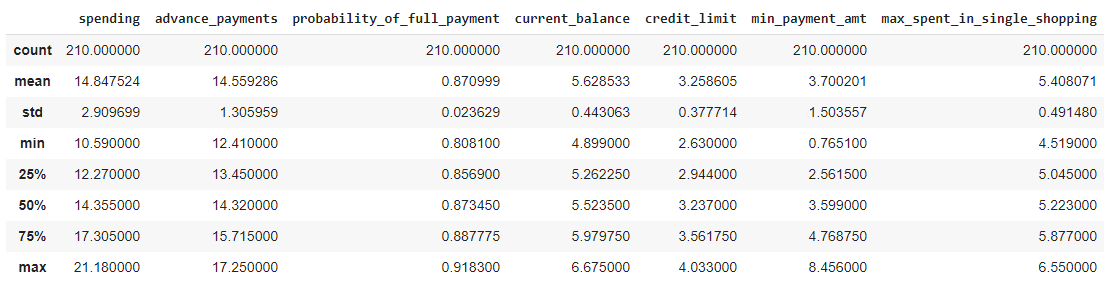
**Q1**

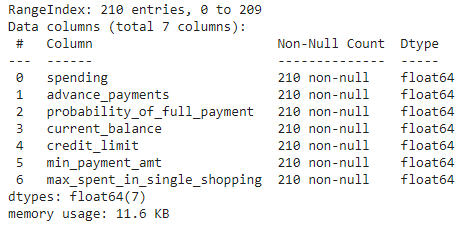
**Q 1.1**

The uni, bi and multi variate analysis are as follows:

****

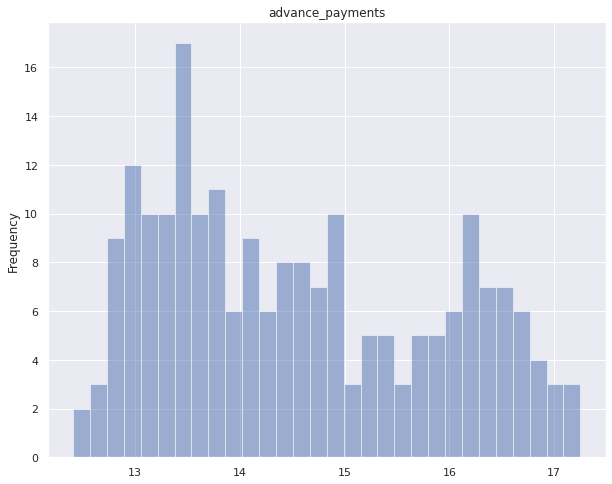
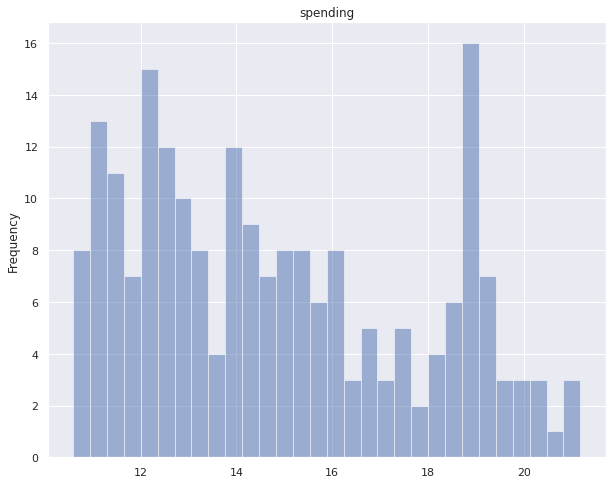
Above figures gives us the Statistics summary of the dataset.

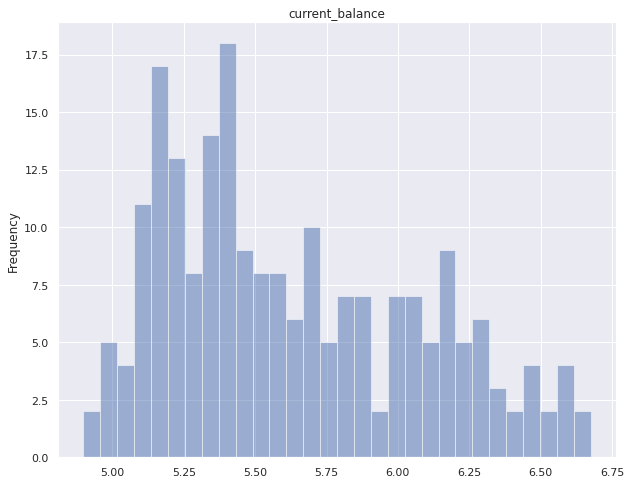
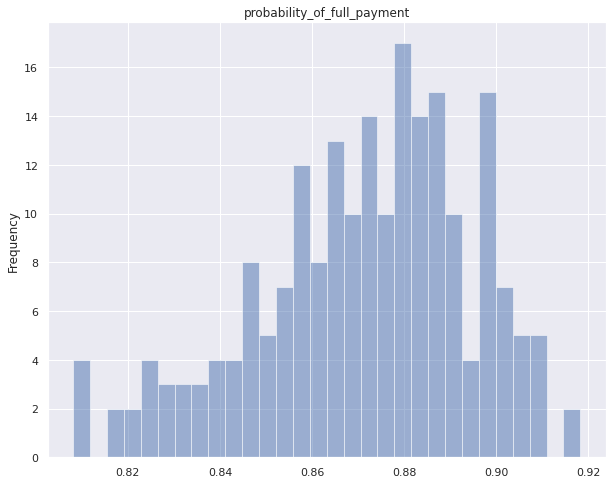
Next we will gain some information on the type of dataset and variables

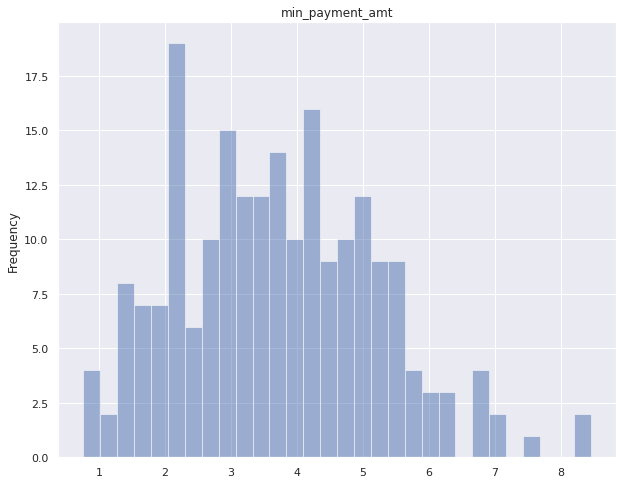
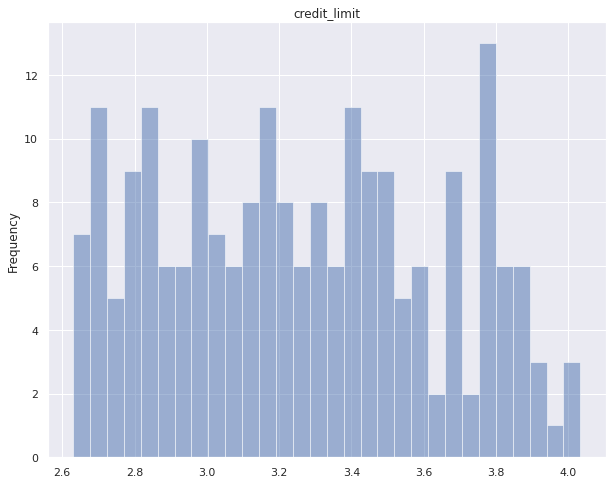


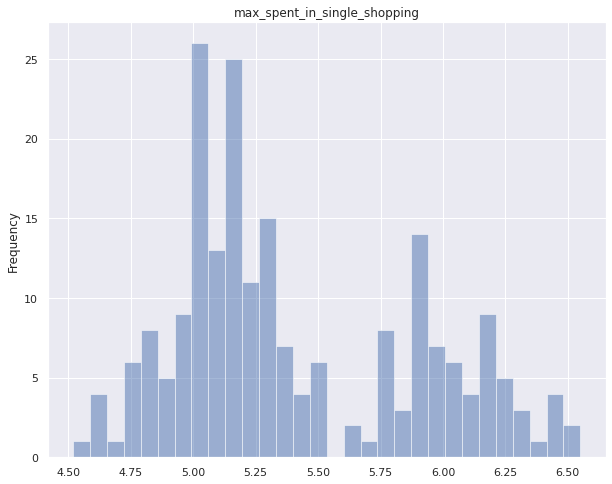
We see that all the data present are non-null and dtypes are float. No categorical values present here.

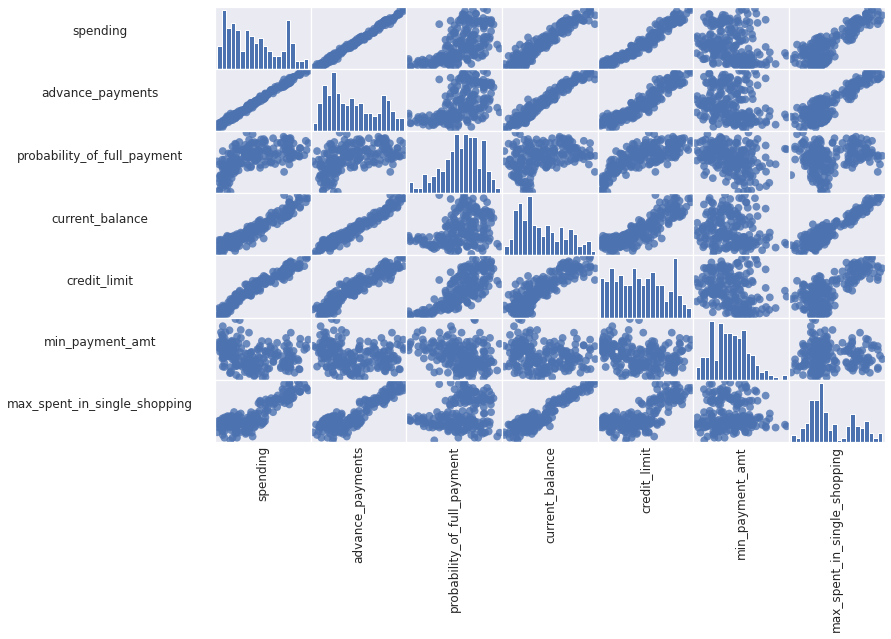
Histogram plot of all the variables (single plot)







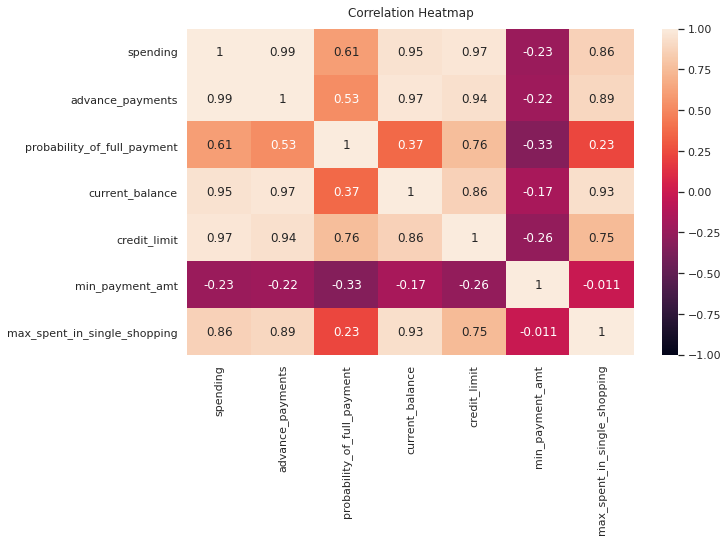




This pair plot for multivariate analysis shows us the complete relationship between different variables in terms of the distribution of the plot points. We observe an interesting pattern here. Most of the data are **linearly dependent** on each other.

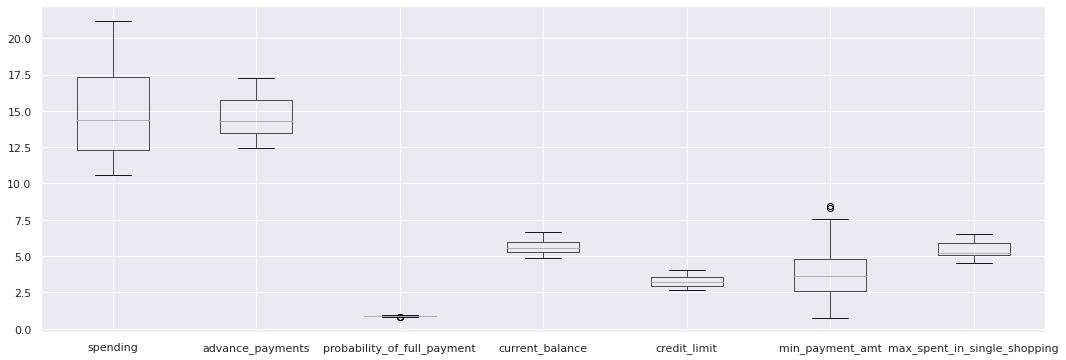
*Example*: Variables **credit\_limit, max\_spent\_in\_single\_shopping, current\_balance, and advance\_payment** are linearly dependent on **spending**

**Correlation Matrix**

****

We see high correlation values (both positive and negative but mostly positive) between all the variables

We dont find any outliers in the variables (the black scatter points represents outliers before scaling)



**Q 1.2**

Scaling: Since the **mean and std** of the variables are different from each other and their variation (difference in range) is very significant, therefore scaling is necessary for clustering in order to transform the variables into same range

**Q 1.3**

Dendrogram plot (Hierarchical clustering)



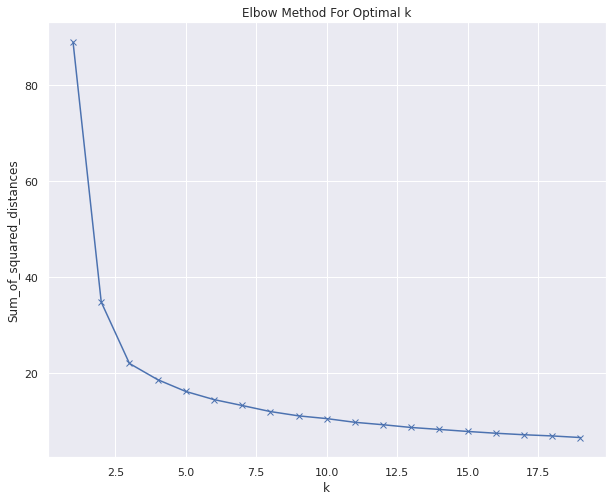
If we consider the dissimilarity to be not greater than **1.5−1.8** (y-axis of the dendrogram) sufficient enough for forming a cluster, then we find that the optimum number of clusters are **6**.

One cluster from (2) to (6), Second cluster from (3) to (18), Third one from (15) to (12), Fourth cluster from (6) to (18), then (20) to (37) and finally (5) to (28) makes a total of 6 clusters.

**Q 1.4**

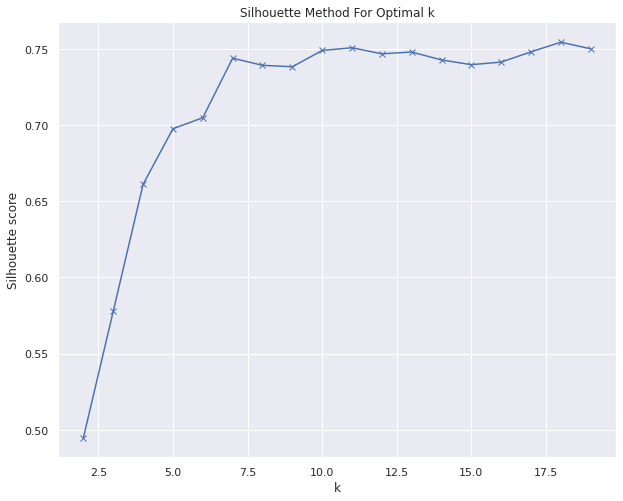
**Elbow method**

.



In the plot above the elbow is at **k=6 or k=7** indicating the optimal k for this dataset is 6 (which we can safely assume) after which distance is almost converged which is in agreement to the hierarchical clustering!

**SILHOUTTE METHOD**

****

Here also we see that **k=6** is the optimum number of clusters after which silhouette score is converged! So our finalised number of clusters are **6**

**Q 1.5**

From the business perspective, these **6** different clusters can represent different classes of customer which are segmented into different clusters based on their spending, advance\_payments, probability\_of\_full\_payment, current\_balance, credit\_limit and other features.

Maybe one cluster represents that customer is spending too much with high credit limit and high current balance, and other cluster might represents another scenario.

So based on these different clusters we can group our customers and deal with them accordingly by recommending them with different schemes like priority customer service based on their spending, priority new product services based on current balance or credit limit, multiple financing options like ease of home and vehicle loans disbursements based on another set of cluster and features, etc