Title: Credit Card Holders Segmentation

1. Main objective of the analysis that also specifies whether your model will be focused on clustering or dimensionality reduction and the benefits that your analysis brings to the business or stakeholders of this data.

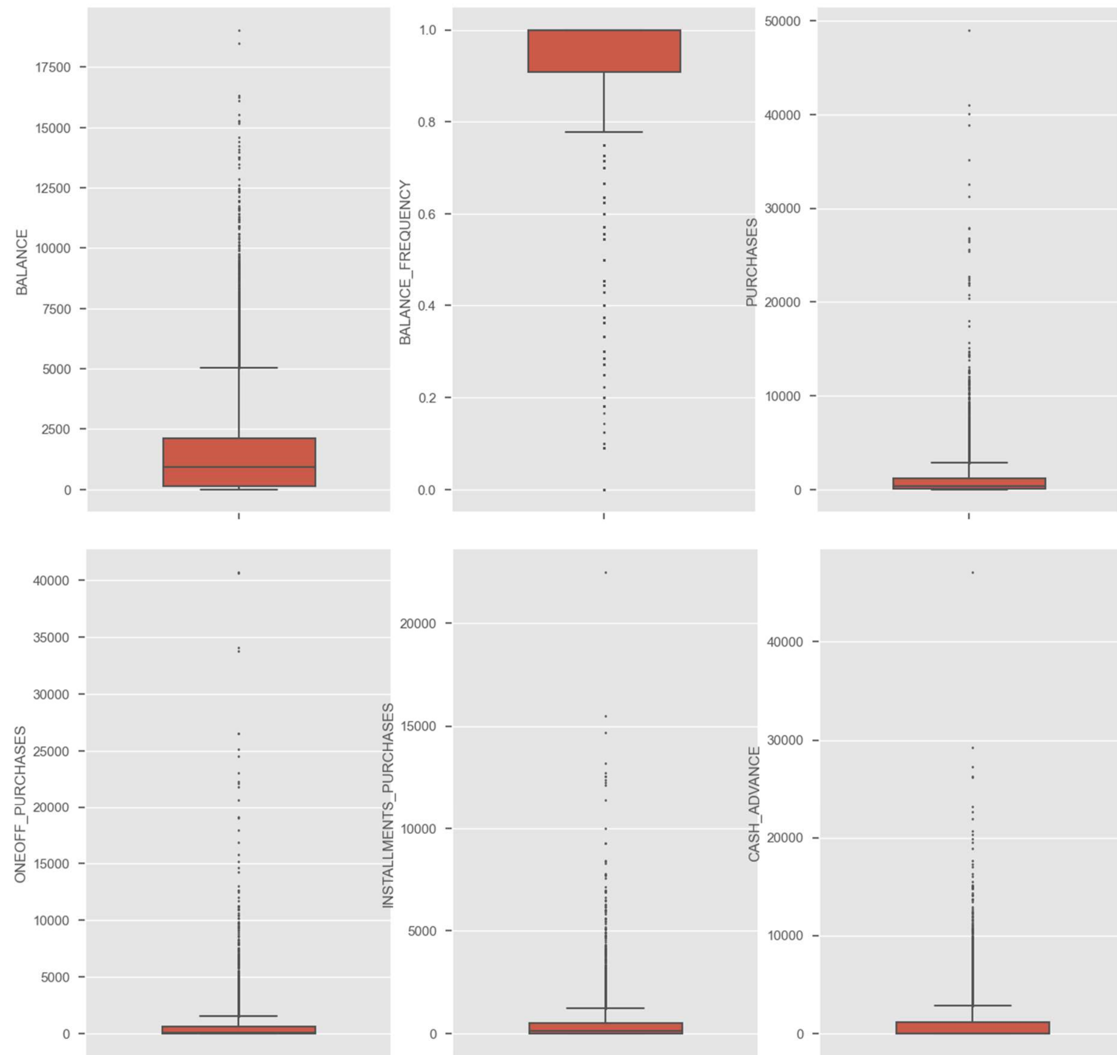
- The objective of the analysis is to segment credit card holders into similar groups based on their credit card usage behaviour over the period of six months to make defined marketing strategy targeted for this groups.

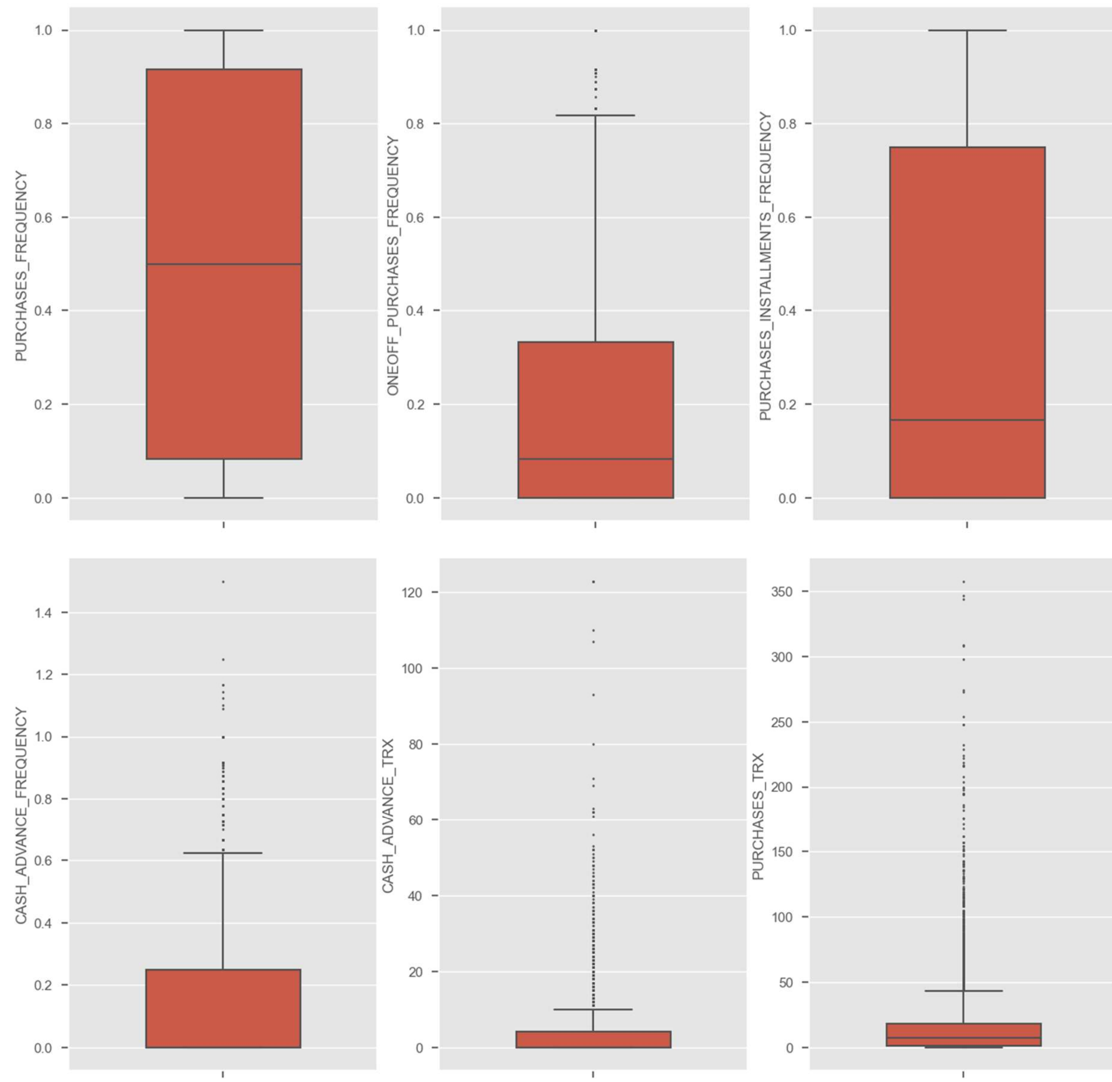
2. Brief description of the data set you chose, a summary of its attributes, and an outline of what you are trying to accomplish with this analysis.

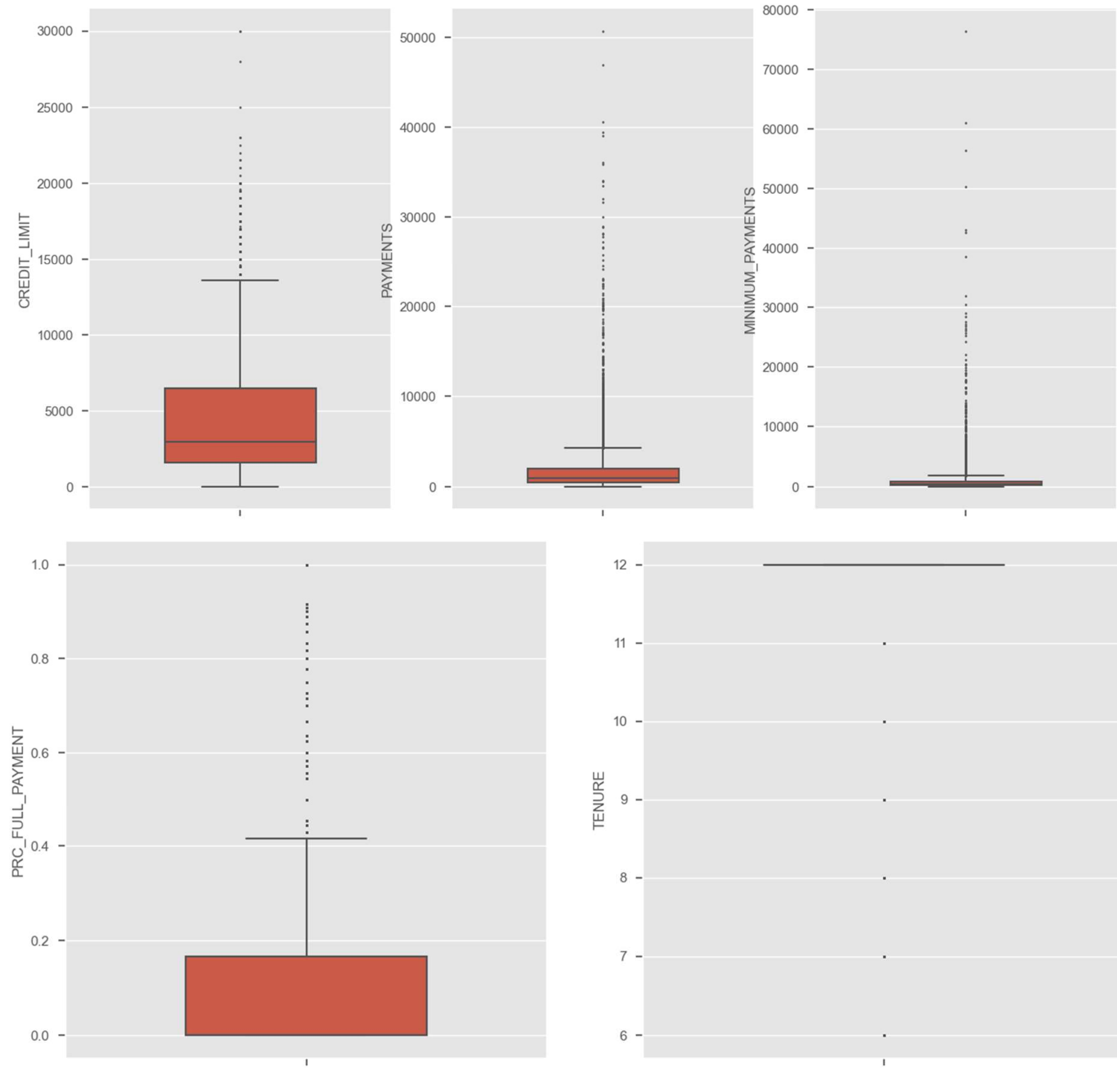
- The data used for analysis is called the "Credit Card Dataset for Clustering" version one which is available via https://www.kaggle.com/datasets/arjunbhasin2013/ccdata, uploaded by Arjun Bhasin four years ago.
- It contains 8950 entries of credit card holders with 18 columns capturing the data relating to their usage behaviour over the period of six months in CSV file format. Following are the columns:
  - CUSTID : Identification of Credit Card holder (Categorical)
  - BALANCE : Balance amount left in their account to make purchases (
  - BALANCEFREQUENCY : How frequently the Balance is updated, score between 0 and 1 (1 = frequently updated, 0 = not frequently updated)
  - PURCHASES : Amount of purchases made from account
  - ONEOFFPURCHASES : Maximum purchase amount done in one-go
  - INSTALLMENTSPURCHASES : Amount of purchase done in installment
  - CASHADVANCE : Cash in advance given by the user
  - PURCHASESFREQUENCY : How frequently the Purchases are being made, score between 0 and 1 (1 = frequently purchased, 0 = not frequently purchased)
  - ONEOFFPURCHASESFREQUENCY : How frequently Purchases are happening in one-go (1 = frequently purchased, 0 = not frequently purchased)
  - PURCHASESINSTALLMENTSFREQUENCY : How frequently purchases in installments are being done (1 = frequently done, 0 = not frequently done)
  - CASHADVANCEFREQUENCY : How frequently the cash in advance being paid
  - CASHADVANCETRX : Number of Transactions made with "Cash in Advanced"
  - PURCHASESTRX : Numbe of purchase transactions made
  - CREDITLIMIT : Limit of Credit Card for user
  - PAYMENTS : Amount of Payment done by user
  - MINIMUM_PAYMENTS : Minimum amount of payments made by user
  - PRCFULLPAYMENT : Percent of full payment paid by user
  - TENURE : Tenure of credit card service for user

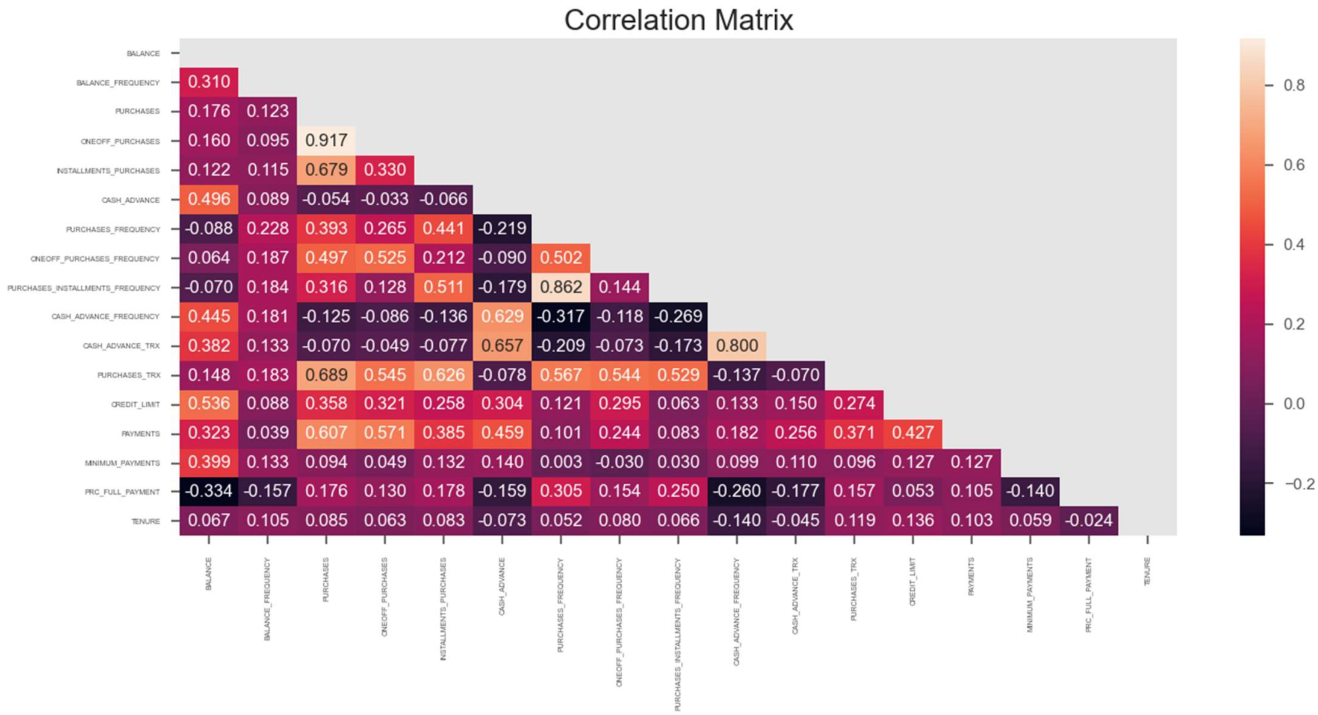
3. Brief summary of data exploration and actions taken for data cleaning or feature engineering.

- Out of 8950 entries, 314 has at least one missing value.
- No imputation done for missing data hence leaving 8636 entries with complete data to be analysed.
- No feature engineering was performed for the analysis.
- CUSTID was dropped from analysis because it has no value to the analysis.
- Boxplot of all the 17 features:




PURCHASES_FREQUENCY
ONEOFF_PURCHASES_FREQUENCY
PURCHASES_INSTALLMENTS_FREQUENCY
CASH_ADVANCE_FREQUENCY
CASH_ADVANCE_TRX
PURCHASES_TRX


CREDIT_LIMIT
PAYMENTS
MINIMUM_PAYMENTS
PRC_FULL_PAYMENT
TENURE

- 

  Above is the correlation matrix of the 17 features. There are some high corelations i.e. > 0.8. The most noticeable is between PURCHASES and ONEOFF_PURCHASES i.e. 0.917.

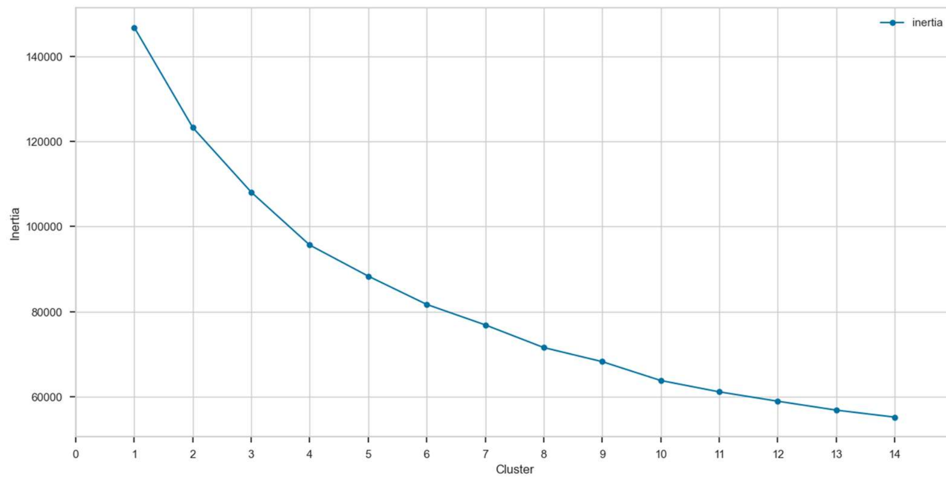
- Summary statistics of the 17 features:

	count	mean	std	min	25%	50%	75%	max
BALANCE	8636.0	1601.224893	2095.571300	0.000000	148.095189	916.855459	2105.195853	19043.13856
BALANCE_FREQUENCY	8636.0	0.895035	0.207697	0.000000	0.909091	1.000000	1.000000	1.00000
PURCHASES	8636.0	1025.433874	2167.107984	0.000000	43.367500	375.405000	1145.980000	49039.57000
ONEOFF_PURCHASES	8636.0	604.901438	1684.307803	0.000000	0.000000	44.995000	599.100000	40761.25000
INSTALLMENTS_PURCHASES	8636.0	420.843533	917.245182	0.000000	0.000000	94.785000	484.147500	22500.00000
CASH_ADVANCE	8636.0	994.175523	2121.458303	0.000000	0.000000	0.000000	1132.385490	47137.21176
PURCHASES_FREQUENCY	8636.0	0.496000	0.401273	0.000000	0.083333	0.500000	0.916667	1.00000
ONEOFF_PURCHASES_FREQUENCY	8636.0	0.205909	0.300054	0.000000	0.000000	0.083333	0.333333	1.00000
PURCHASES_INSTALLMENTS_FREQUENCY	8636.0	0.368820	0.398093	0.000000	0.000000	0.166667	0.750000	1.00000
CASH_ADVANCE_FREQUENCY	8636.0	0.137604	0.201791	0.000000	0.000000	0.000000	0.250000	1.50000
CASH_ADVANCE_TRX	8636.0	3.313918	6.912506	0.000000	0.000000	0.000000	4.000000	123.00000
PURCHASES_TRX	8636.0	15.033233	25.180468	0.000000	1.000000	7.000000	18.000000	358.00000
CREDIT_LIMIT	8636.0	4522.091030	3659.240379	50.000000	1600.000000	3000.000000	6500.000000	30000.00000
PAYMENTS	8636.0	1784.478099	2909.810090	0.049513	418.559237	896.675701	1951.142090	50721.48336
MINIMUM_PAYMENTS	8636.0	864.304943	2372.566350	0.019163	169.163545	312.452292	825.496463	76406.20752
PRC_FULL_PAYMENT	8636.0	0.159304	0.296271	0.000000	0.000000	0.000000	0.166667	1.00000
TENURE	8636.0	11.534391	1.310984	6.000000	12.000000	12.000000	12.000000	12.00000

- K-means clustering was used for segmentation. Number of clusters was decided based on Elbow method and Silhouette score for performance metric.

4. Summary of training at least three variations of the unsupervised model you selected. For example, you can use different clustering techniques or different hyperparameters.

- The 17 features were standardized by removing the mean and scaling to unit variance before fitting K-means clustering algorithm for this segmentation analysis with variations of number of clusters.
- Below is the plot of Elbow method based on inertia to get optimal number of clusters which was decided based on the Silhouette score for clustering performance evaluation.



- Following table summarizes the clustering performance evaluation when there is no information about the ground truth labels for 3, 4 and 5 clusters:

	Number of clusters		
Clustering performance evaluation	2	3	4
Silhouette score (Euclidean)	0.2089	0.2474	0.1970

- Following are the values on original scale of the centroids after applying inverse transformation.

Cluster	BALANCE	BALANCE_ FREQUENCY	PURCHASES	ONEOFF_ PURCHASES	INSTALLMENTS_ PURCHASES
0	825.492132	0.859565	519.571285	260.442347	259.463972
1	2236.823143	0.982236	4301.224269	2733.266623	1568.453105
2	4012.022317	0.96031	384.863467	247.772178	137.170262

Cluster	CASH_ ADVANCE	PURCHASES_ FREQUENCY	ONEOFF_ PURCHASES_ FREQUENCY	PURCHASES_ INSTALLMENTS_ FREQUENCY	CASH_ADVANCE_ FREQUENCY
0	331.44525	0.472205	0.135852	0.349751	0.069702
1	465.763191	0.950768	0.667714	0.749913	0.064056
2	3882.326227	0.23335	0.110901	0.145462	0.448529

Cluster	CASH_ADVANCE_TRX	PURCHASES_TRX	CREDIT_LIMIT	PAYMENTS	MINIMUM_PAYMENTS
0	1.230585	8.864653	3266.209863	947.135011	532.901337
1	1.554088	57.023947	7774.772915	4183.20924	1239.01217
2	12.469349	5.640485	6705.494601	3062.338726	1814.44752

Cluster	PRC_FULL_PAYMENT	TENURE
0	0.164012	11.501109
1	0.299224	11.921552
2	0.033485	11.359515

5. A paragraph explaining which of your Unsupervised Learning models you recommend as a final model that best fits your needs in terms.

- The final model with three clusters for the complete data of 8636 entries is recommended because it has the highest Silhouette Coefficient score (0.2474) indicating better defined clusters than both 2 and 4 clusters.

6. Summary Key Findings and Insights, which walks your reader through the main findings of your modeling exercise.

- Looking at the last table in question 4 on inverse transformed values of the centroids, the following points have been summarized:

Cluster	Label	Summary
2	Low credit card users	- Most balance in account - Least activities relating to purchases - Most activities relating to cash advance activities
0	Moderate credit cards user	- Least balance - Least cash advance - Least credit limit - Least payment made
1	Heavy credit card users	- Most activities relating to purchases - Highest credit limit - Most activities relating to minimum payments made

The above summary may be helpful to the credit card provider to design a more effective credit card marketing strategy and campaign such as cash-back offers and spend and get rewards to ensure retainment and engagement of this credit card holders to further drive business profit.

7. Suggestions for next steps in analyzing this data, which may include suggesting revisiting this model or adding specific data features to achieve a better model.

- Investigate and study the impact due to missing data and outliers
- Perform principal component analysis to reduce dimension from this 17 input features
- Reassess the K-means clustering model after incorporating the previous points
- Explore and compare other clustering algorithms

References

1. https://www.kaggle.com/datasets/arjunbhasin2013/ccdata
2. https://www.coursera.org/learn/ibm-unsupervised-machine-learning/home/welcome
3. https://www.coursera.org/learn/machine-learning-for-customer-segmentation/home/welcome
4. https://towardsdatascience.com/performance-metrics-in-machine-learning-part-3-clustering-d69550662dc6
5. https://scikit-learn.org/stable/modules/clustering.html#clustering-performance-evaluation
6. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.calinski_harabasz_score.html
7. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.davies_bouldin_score.html