

Credit Card Customer Segmentation

By NAEEM SUFI

Background and Context

- ▶ AllLife Bank wants to focus on its credit card customer base in the next financial year. They have been advised by their marketing research team, that the penetration in the market can be improved.
- ▶ Based on this input, the Marketing team proposes to run personalized campaigns to target new customers as well as upsell to existing customers. Another insight from the market research was that the customers perceive the support services of the bank poorly.
- ▶ Based on this, the Operations team wants to upgrade the service delivery model, to ensure that customer queries are resolved faster. Head of Marketing and Head of Delivery both decided to reach out to the Data Science team for help.

Objective and Questions to be answered

- ▶ Main Objective: To identify different segments in the existing customer base. This is to be based on Customers spending patterns as well as their past interactions with the bank.
- ▶ **Key Questions**
 - ▶ How many different segments of customers are there?
 - ▶ How are these segments different from each other?
 - ▶ What are our recommendations to the bank on how to better market to and service these customers?

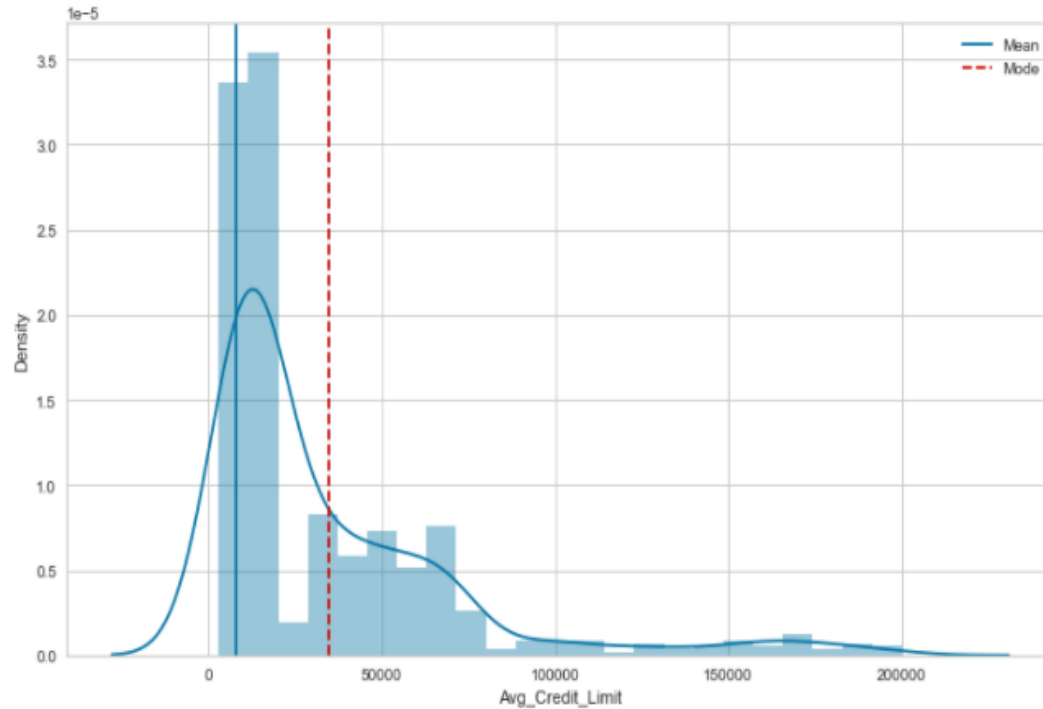
Data Description

- ▶ Data is of various customers of a bank with their credit limit, the total number of credit cards the customer has, and different channels through which customer has contacted the bank for any queries, different channels include visiting the bank, online and through a call.
- ▶ Customer key - Identifier for the customer
- ▶ Average Credit Limit - Average credit limit across all the credit cards
- ▶ Total credit cards - Total number of credit cards
- ▶ Total visits bank - Total number of bank visits
- ▶ Total visits online - total number of online visits
- ▶ Total calls made - Total number of calls made by the customer

Univariate Analysis

Average credit limit

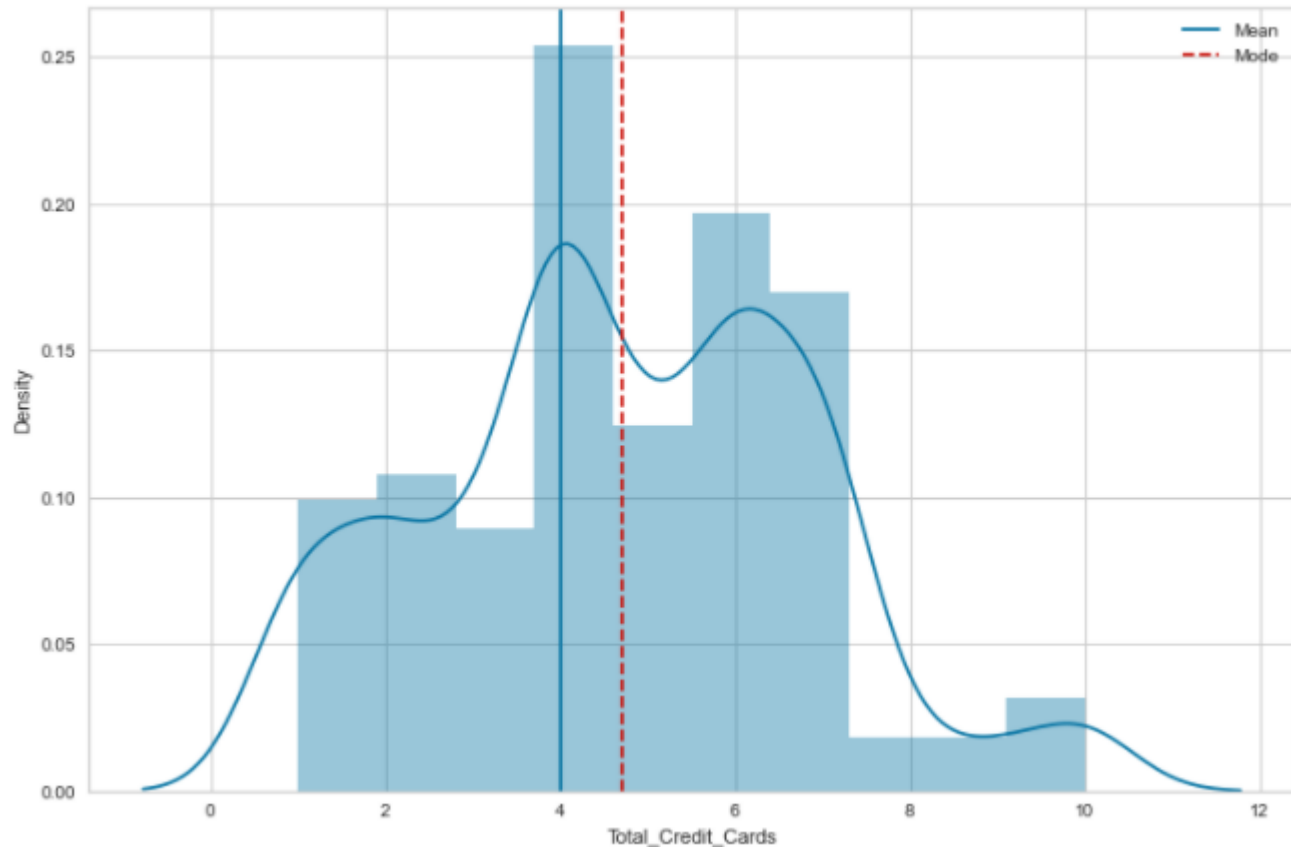
Observation: Most customers do not have credit or have a low credit limit?



- Average Credit Limit is between 10k to 20k and maximum limit of our credit card is 2K And minimum limit of card is 5k.

Total Credit Cards

Observation: . Wonder, why do so many users have more than one credit card?

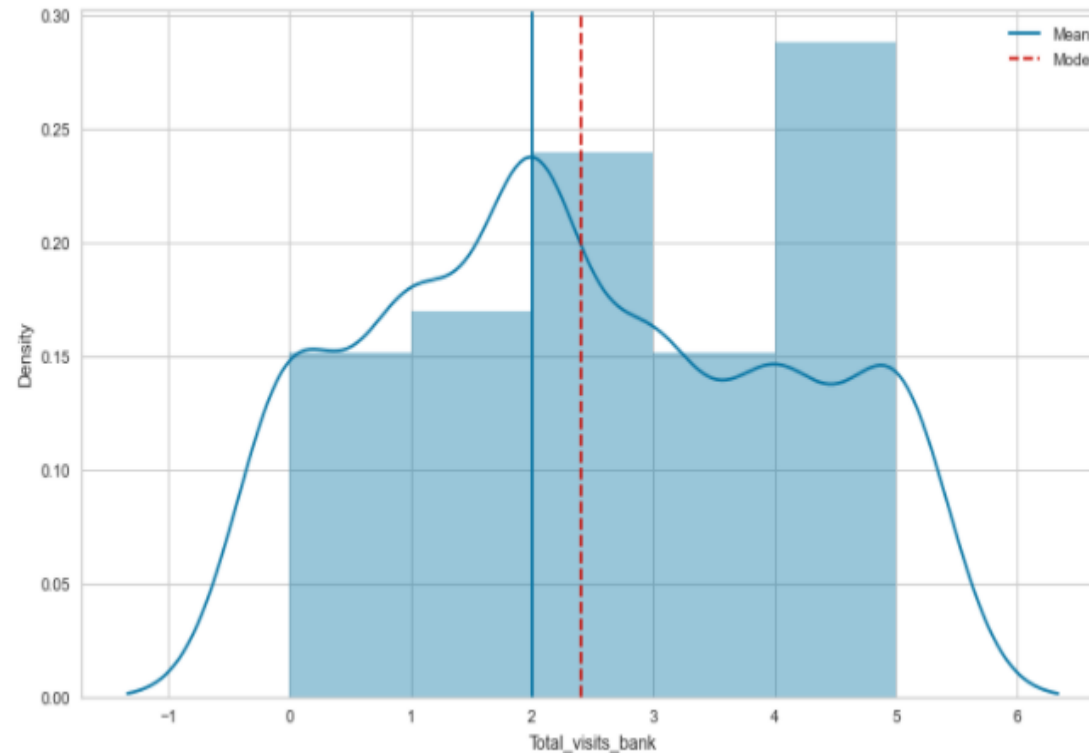


Looks to be normally distributed. The average of total_credit_cards are between 4 to 7. and maximum cards are 10. And Minimum Total_Credit_Card we have are 1 to 2.

Total Calls Made

Observation: The total number of calls made by the customer?

- Average number of calls made by the customer are between 1 to 5. And maximum number of calls made by the customer are almost 10. And minimum calls made close to 1 to 2.

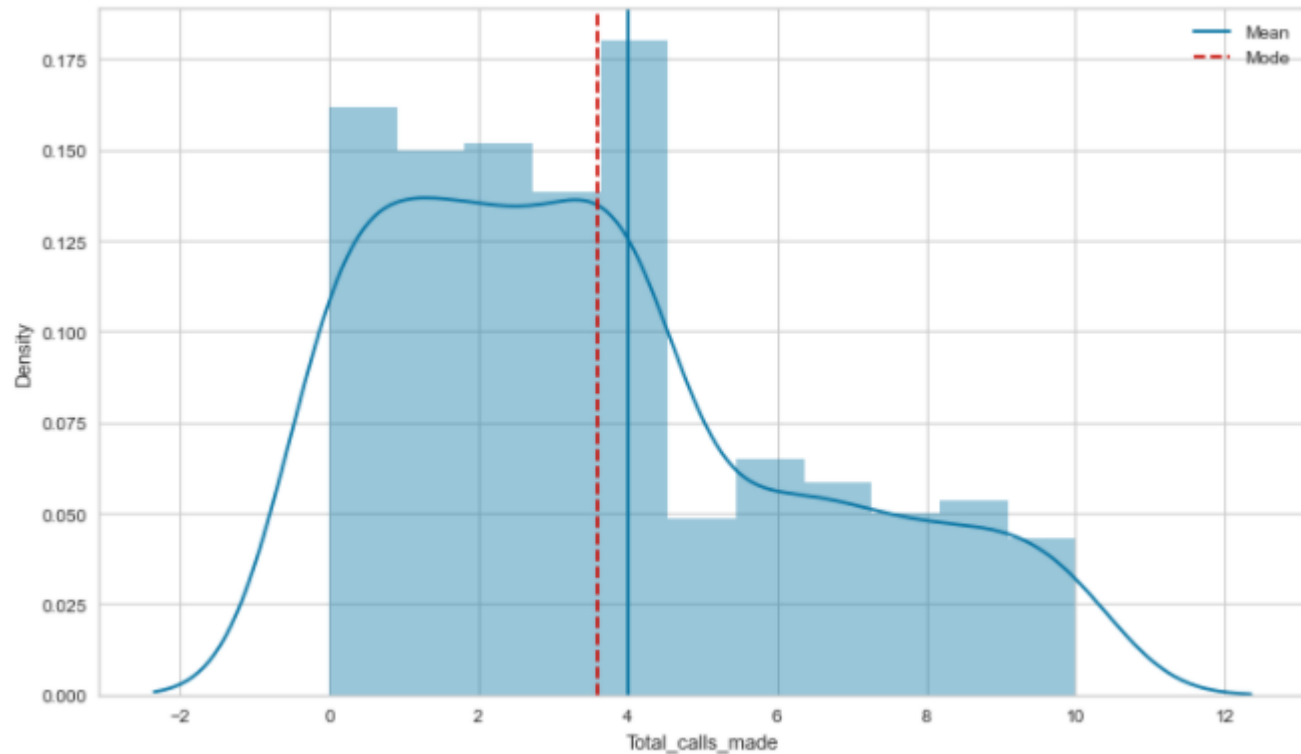


Total Bank Visits

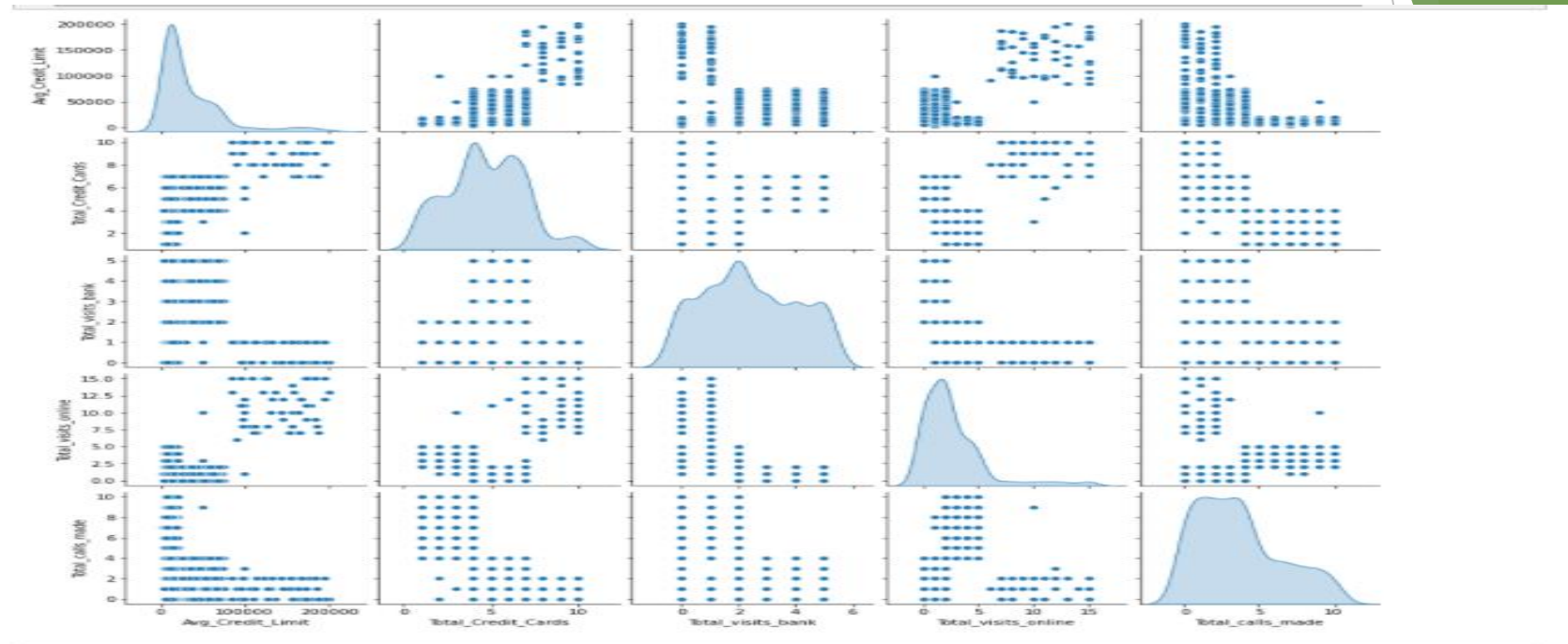
Observation:

Total number of Visits that customer made (yearly) personally to the bank are?

- . Average number of people of visiting bank is between 2 to 4 times. And the maximum number of times customer visited the bank are are 5 times.



Bivariate Analysis



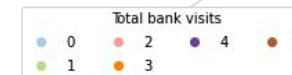
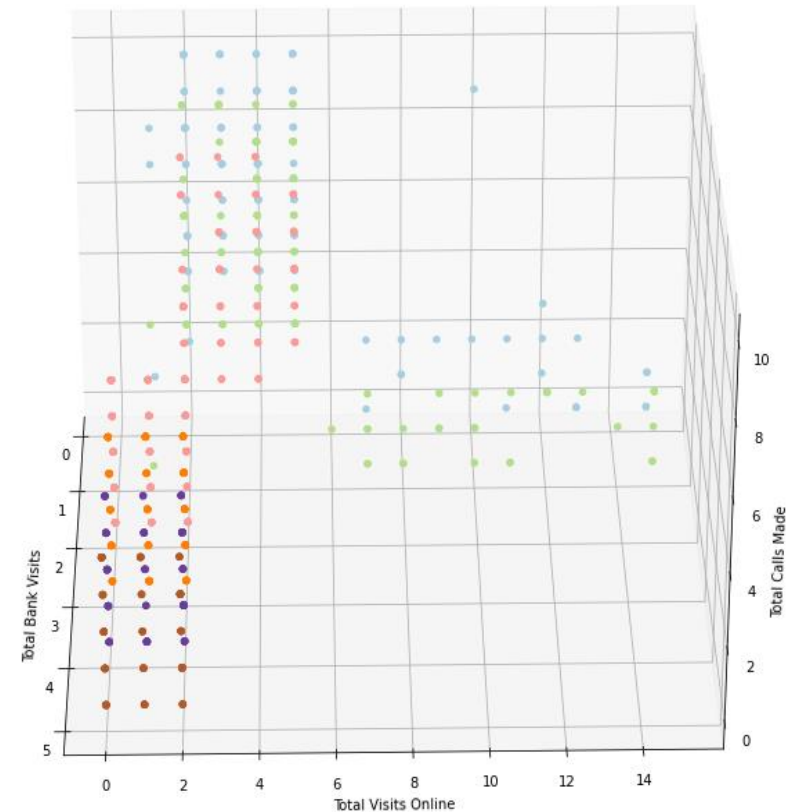
- ▶ We use the pair plot for bivariate Analysis.
- ▶ As in first plot you see the Avg_credit_limit is between 100000 to 150000.
- ▶ And the avg_credit_card are between 5 to 8.
- ▶ There are very low number of people who visit bank rather than making calls and online transactions.
- ▶ Most of the people are sort their issues by making calls.

Contact Method(3D plotting)

Observation:

Plot 3D scatter plot which shows where customers would stick to their preferred method for interacting with their bank (online, in person, and through the phone)?

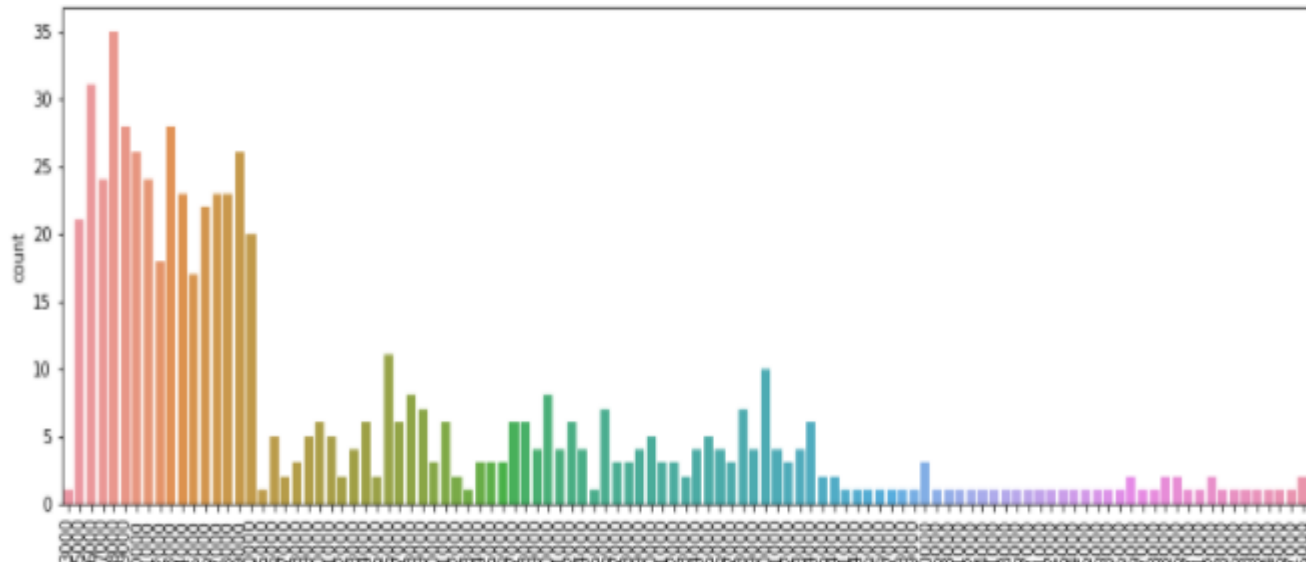
- ▶ A hypothesis that I had going into this was that there would be three clusters for contact method, where customers would stick to their preferred method for interacting with their bank (online, in person, and through the phone).
- ▶ we can see a 3D rotating scatter plot which shows my hypothesis was correct.



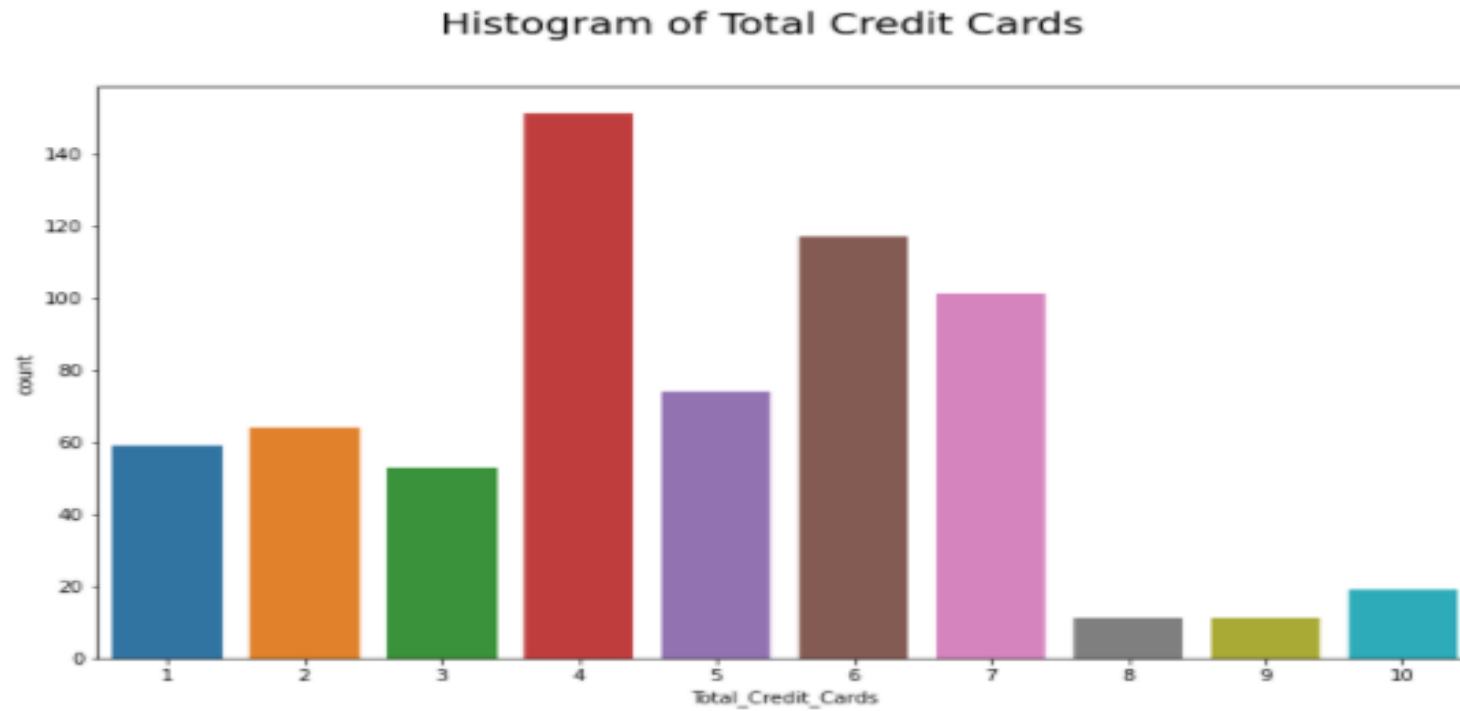
Insights Based on EDA

- ▶ There are 660 observations and 7 columns in the dataset.
- ▶ All columns have 660 non-null values i.e. there are no missing values.
- ▶ All columns are of int64 data type.
- ▶ There are no missing values.

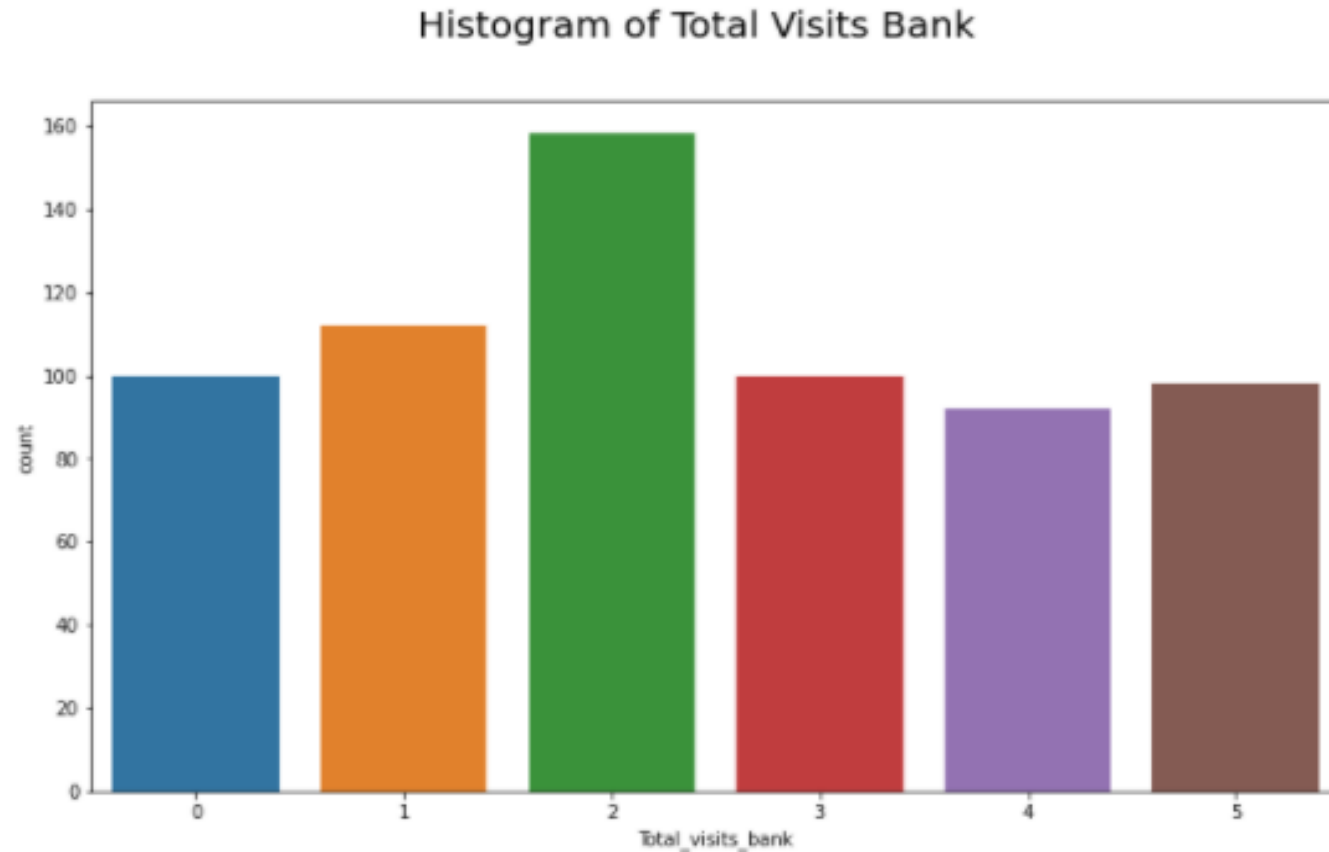
Histogram of Average Credit Limit



- ▶ - Credit limit average is around 20K with 50% of customers having a
- ▶ - credit limit less than 5K, which implies a high positive skewness.

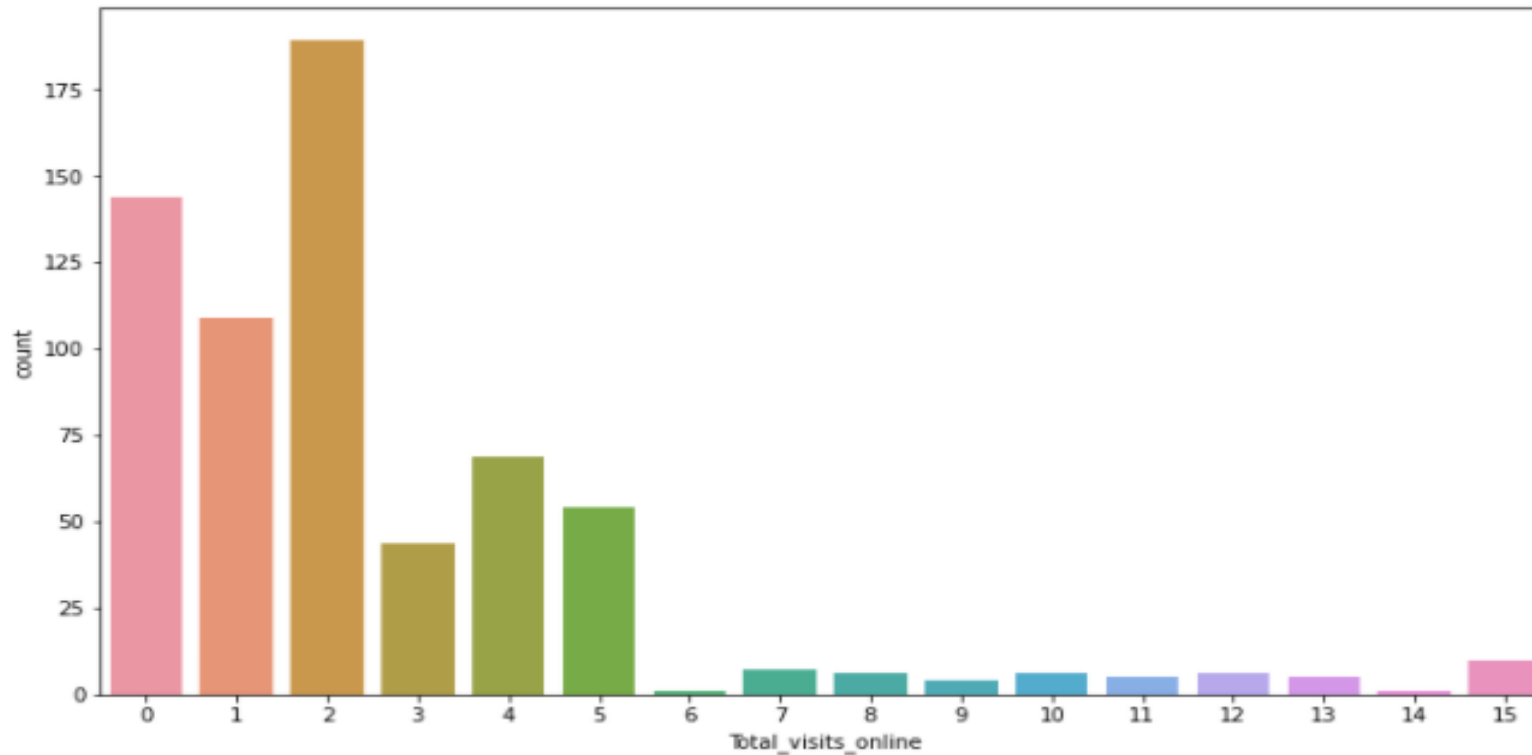


- ▶ - On average, credit cards owned by each customer are ~4. Some customers have 10.
- ▶ - On average, most customer interactions are through calls, then online. Also, some customers never contacted/visited the bank.
- ▶ - Have a higher average credit limit and each customer have up to 8 to 10 credit cards to the maximum.



- Therefore , the number of visits via online is higher but their visits to bank is lower. Almost 2 times customer visit's a bank in year. And maximum number is 5 times in a year customer visited the bank.

Histogram of Total Visits Online



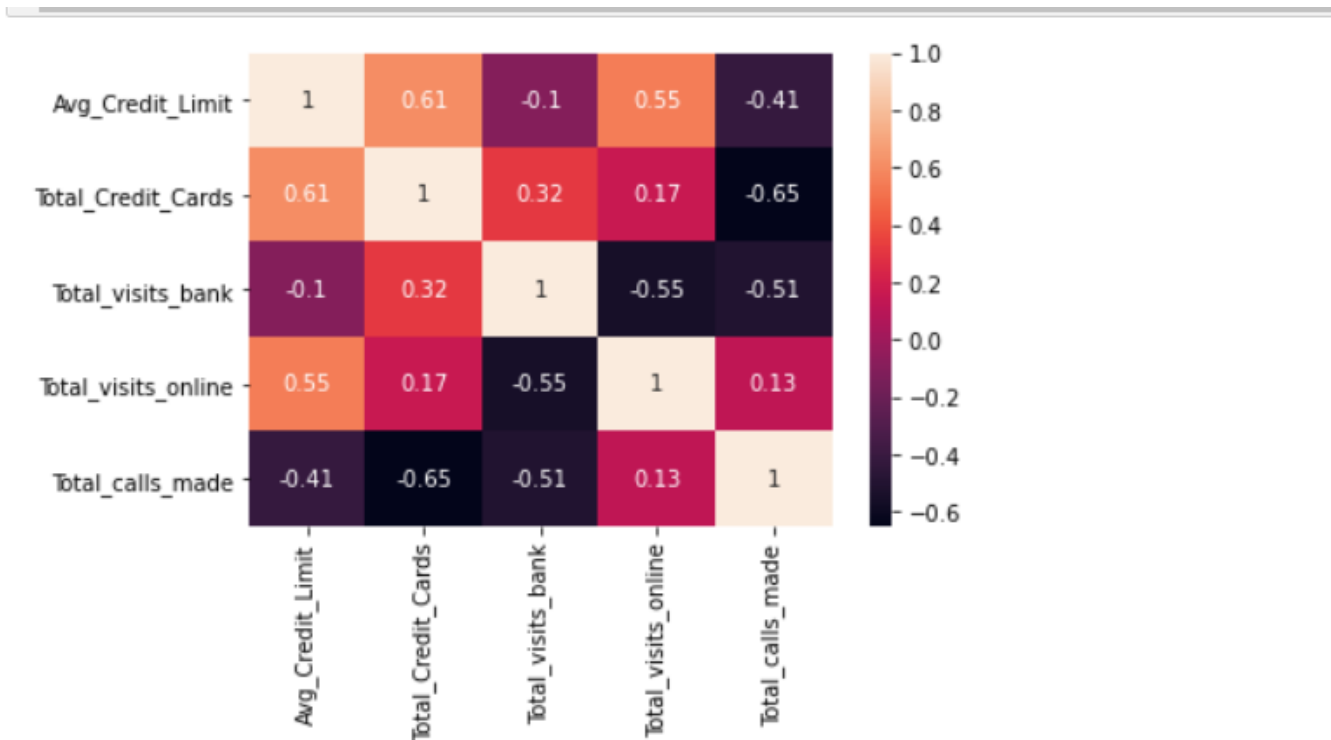
- ▶ Therefore , the number of visits via online is higher than the others like visit bank. Almost 5 times customer online visit a bank in year. And maximum number is 15 times in a year customer Online visited the bank.
- ▶ The reason most of these customers get a higher income and they the spend more and so they rely more visits online for easier payments.

	Customer Key	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made
count	660.000000	660.000000	660.000000	660.000000	660.000000	660.000000
mean	55141.443939	34574.242424	4.706061	2.403030	2.606061	3.583333
std	25627.772200	37625.487804	2.167835	1.631813	2.935724	2.865317
min	11265.000000	3000.000000	1.000000	0.000000	0.000000	0.000000
25%	33825.250000	10000.000000	3.000000	1.000000	1.000000	1.000000
50%	53874.500000	18000.000000	5.000000	2.000000	2.000000	3.000000
75%	77202.500000	48000.000000	6.000000	4.000000	4.000000	5.000000
max	99843.000000	200000.000000	10.000000	5.000000	15.000000	10.000000

- ▶ Looking at standard deviation, we can see a considerably high variation in credit limits as well.
- ▶ And in other columns like Total_Credit_Cards, Total_visits_online and Total_calls_made have very low amount of variations in data.
- ▶ Have a higher average credit limit and seen that each customer have up to 8 to 10 credit cards to the maximum.
- ▶ Therefore , the number of visits via online is higher but their visits to bank is lower .
- ▶ The reason most of these customers get a higher income and they spend more and so they rely more visits online for easier payments.
- ▶ They seldom do calls or at the most 2 per day.

Relationship Between Variables

- ▶ Total credit cards and total visit online has medium positive correlation with average credit limit : 0.61,0.55 respectively.
- ▶ Total credit cards and total visit bank has medium negative correlation with total calls made : -0.65,-0.5 respectively.
- ▶ Total visit online and medium negative correlation with total visit bank : -0.55.



Exploratory Data Analysis(EDA)

- ▶ We have a data set of **Credit Card Customer Segmentation**
- ▶ This is not a huge dataset so, we will explore the dataset to know more about the data shape, descriptive analyses, Statistical analysis, Univariate ,bivariate analysis, correlation, missing values, dtypes, column names etc.
- ▶ Let's try to look the shape of data:

```
df.shape
```

```
(660, 6)
```

- Lets check the column name and information about the data

```
['Customer Key', 'Avg_Credit_Limit', 'Total_Credit_Cards', 'Total_visits_bank', 'Total_visits_online', 'Total_calls_made']
```

Exploring column names, is an important aspect of EDA. • We can see that columns are not null. The data types of all columns are integer data type.

By closely observing the data and description given about each column attribute we can say that:

Numeric data columns are Customer Key, Avg_Credit_Limit, Total_Credit_Cards, Total_Visits_bank, Total_calls_made.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 660 entries, 1 to 660
Data columns (total 6 columns):
#   Column                Non-Null Count  Dtype
---  -
0   Customer Key          660 non-null   int64
1   Avg_Credit_Limit      660 non-null   int64
2   Total_Credit_Cards    660 non-null   int64
3   Total_visits_bank     660 non-null   int64
4   Total_visits_online   660 non-null   int64
5   Total_calls_made      660 non-null   int64
dtypes: int64(6)
memory usage: 36.1 KB
```

Descriptive Analysis

- ▶ The Exploratory Data Analysis is more important method which is describe method shows basic statistical characteristics of each numerical feature (int64 and float64 types):
- ▶ Number of non-missing values, mean, standard deviation, range, median, 0.25 and 0.
- ▶ We can see the Min, Max, mean and standard deviation for all key attributes of the dataset 75 quartiles.

	Customer Key	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made
count	660.000000	660.000000	660.000000	660.000000	660.000000	660.000000
mean	55141.443939	34574.242424	4.706061	2.403030	2.606061	3.583333
std	25627.772200	37625.487804	2.167835	1.631813	2.935724	2.865317
min	11265.000000	3000.000000	1.000000	0.000000	0.000000	0.000000
25%	33825.250000	10000.000000	3.000000	1.000000	1.000000	1.000000
50%	53874.500000	18000.000000	5.000000	2.000000	2.000000	3.000000
75%	77202.500000	48000.000000	6.000000	4.000000	4.000000	5.000000
max	99843.000000	200000.000000	10.000000	5.000000	15.000000	10.000000

Missing Values

- There is no missing value in our dataset, so we do not need to handle that, and all the columns are numeric we do not need to apply any label encoding to convert columns into numeric columns.

```
: Customer Key          0
  Avg_Credit_Limit      0
  Total_Credit_Cards     0
  Total_visits_bank      0
  Total_visits_online    0
  Total_calls_made       0
  dtype: int64
```

Duplicate Observations Check

- There are 5 number of duplicate entries for customer key. We handle this later. First you see that those are the five values SI_No which are duplicated.

There are 5 duplicate entries for Customer Key

SI_No	Customer Key	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made
-------	--------------	------------------	--------------------	-------------------	---------------------	------------------

49	37252	6000	4	0	2	8
433	37252	59000	6	2	1	2

SI_No	Customer Key	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made
-------	--------------	------------------	--------------------	-------------------	---------------------	------------------

5	47437	100000	6	0	12	3
333	47437	17000	7	3	1	0

SI_No	Customer Key	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made
-------	--------------	------------------	--------------------	-------------------	---------------------	------------------

412	50706	44000	4	5	0	2
542	50706	60000	7	5	2	2

SI_No	Customer Key	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made
-------	--------------	------------------	--------------------	-------------------	---------------------	------------------

392	96929	13000	4	5	0	0
399	96929	67000	6	2	2	2

SI_No	Customer Key	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made
-------	--------------	------------------	--------------------	-------------------	---------------------	------------------

105	97935	17000	2	1	2	10
633	97935	187000	7	1	7	0

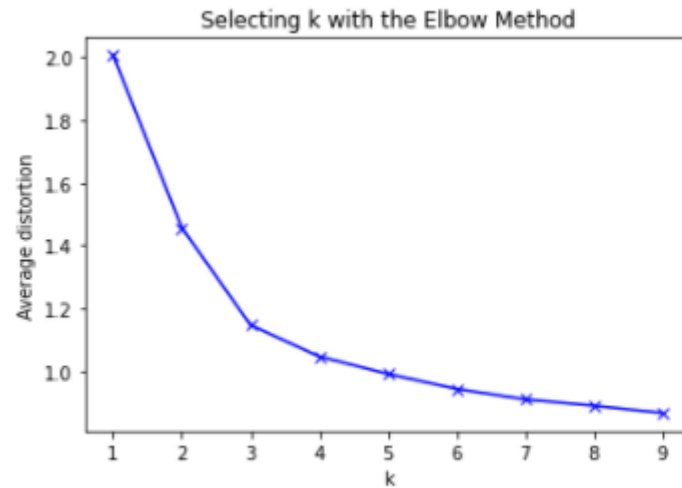
Feature Engineering

- ▶ After feature engineering we find that we have only 6 columns and all columns are important.
- ▶ And by choosing those our model gives more accuracy. And those columns are
- ▶ Customer Key, Avg_Credit_Limit, Total_Credit_Cards, Total_Visits_bank, Total_calls_made

SI_No	Customer Key	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made
1	87073	100000	2	1	1	0
2	38414	50000	3	0	10	9
3	17341	50000	7	1	3	4
4	40496	30000	5	1	1	4
5	47437	100000	6	0	12	3

K-Means Clustering

- First, we want to iterate through and view the performance of each value of K for K-means, then using a line graph to find the elbow of the plot we can select the optimal number of clusters.



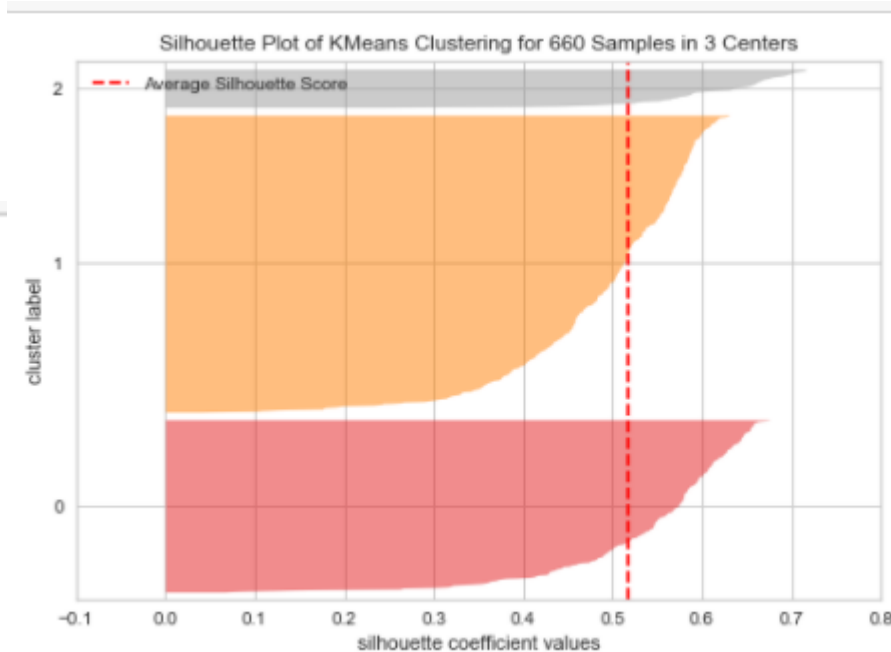
By using Elbow method, we identify that there are 3 optimal number of clusters we need to use.

Silhouette Score

- Silhouette score is a **metric used to calculate the goodness of a clustering technique**. Its value ranges from -1 to 1. 1: Means clusters are well apart from each other and clearly distinguished. And here we got 0.51 which means that clusters are not distinguished much.

clearly Clusters distinguished are

0.5157182558881063



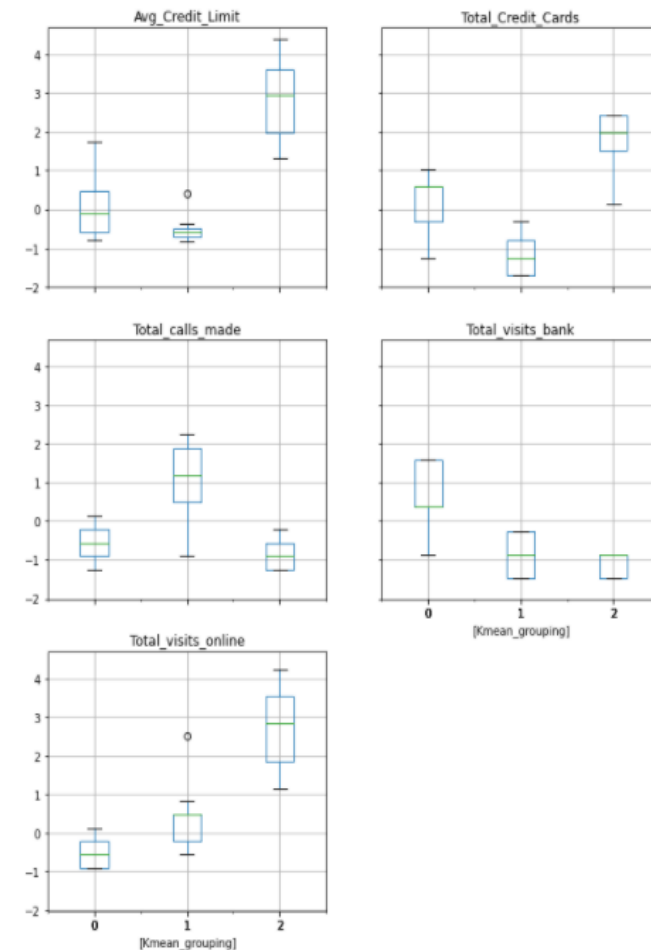
Kmean_grouping

- Here we add the predictions to the unscaled data so that we can gain some real-world interpretability.

Customer Key Avg_Credit_Limit Total_Credit_Cards Total_visits_bank Total_visits_online Total_calls_made						
Kmean_grouping						
0	224	224	224	224	224	224
1	50	50	50	50	50	50
2	386	386	386	386	386	386

- ▶ Below we see the average values for each feature in each cluster, the final row of the dataframe is the mean value for each column (since this data is unscaled, we can use this to determine how big of a step each of these means is from one another).
- ▶ It would appear my hypothesis of the clusters forming along query method has proven correct. If we look at the data, we see that there is a group which prefers online interactions with their bank, they have a much higher credit limit and have more credit cards. The customers who prefer in-person interactions tend to have the least number of credit cards and the lowest credit limit. The customers who contact via phone call are in the middle.
- ▶ One additional observation is that if we tally up the number of interactions per group (how many times they have used online, phone, or in-person services) we see the in-person customers appear to be the most active. This was the opposite of my initial expectations as visiting a bank in person has the highest friction (effort required) to complete

	CustomerKey	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made	Kmean_grouping
Phone	54881.329016	33782.383420	5.515544	3.489637	0.981865	2.000000	NaN
Online	55239.830357	12174.107143	2.410714	0.933036	3.553571	6.870536	NaN
In person	56708.760000	141040.000000	8.740000	0.600000	10.900000	1.080000	NaN
Mean	25627.772200	37625.487804	2.167835	1.631813	2.935724	2.865317	0.634068



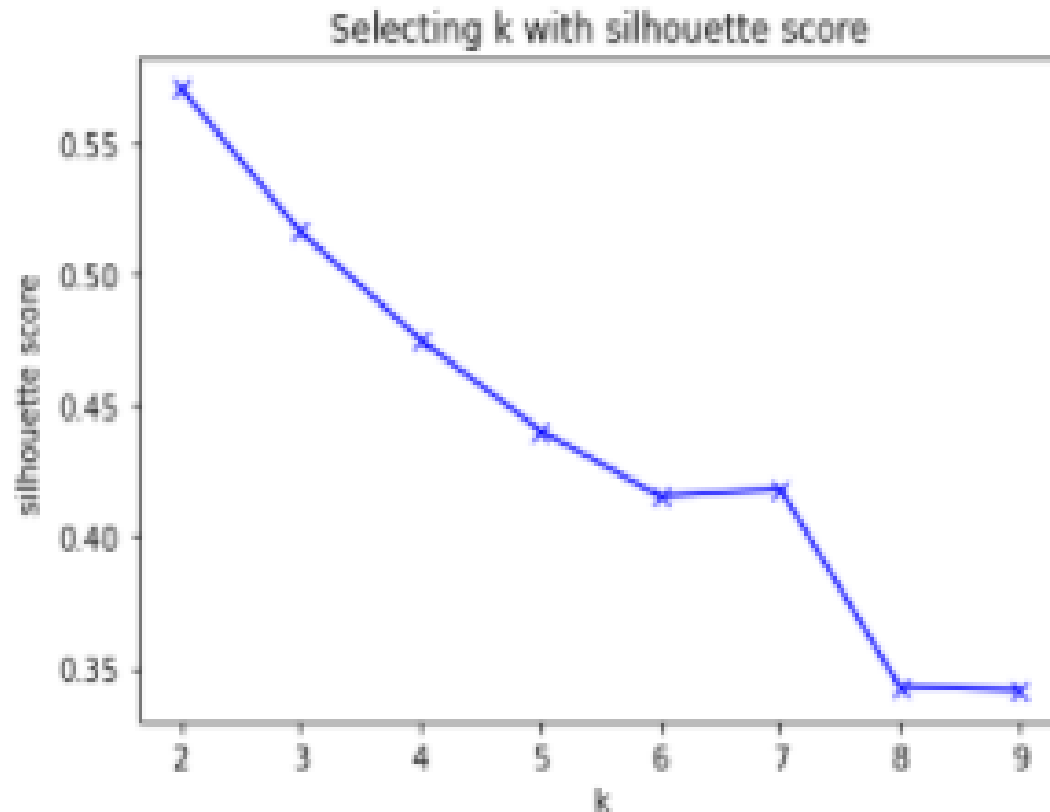
Hierarchical Clustering

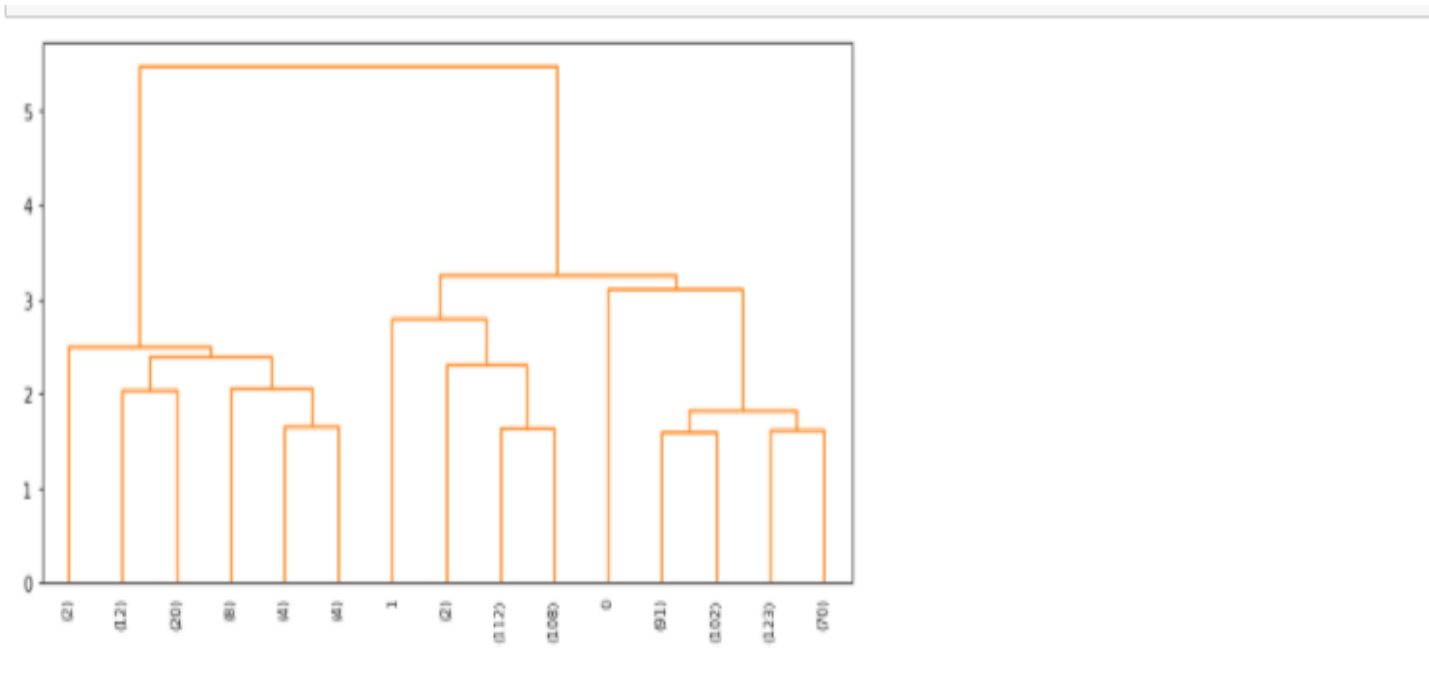
- For hierarchical clustering, we begin by evaluating the cophenetic coefficients for each linkage type and each affinity/metric. Below is a list of said outcomes, ignoring any combination that scores poorly, and ignoring any combination that will result in an error. It's worth noting that scipy has more options than sklearn does for metrics and linkages. While there are several good combinations, I will pick Euclidean for the metric and average for linkage as these are options in both scipy and sklearn.

```
affinity: braycurtis , link: single : 0.8502845518924395
affinity: canberra , link: single : 0.7566178713003985
affinity: chebyshev , link: complete : 0.8533474836336782
affinity: chebyshev , link: average : 0.8974159511838106
affinity: chebyshev , link: weighted : 0.8913624010768603
affinity: cityblock , link: complete : 0.8731477899179829
affinity: cityblock , link: average : 0.896329431104133
affinity: cityblock , link: weighted : 0.8825520731498188
affinity: euclidean , link: complete : 0.8599730607972423
affinity: euclidean , link: average : 0.8977080867389372
affinity: euclidean , link: weighted : 0.8861746814895477
affinity: euclidean , link: centroid : 0.8939385846326323
affinity: euclidean , link: median : 0.8893799537016724
affinity: mahalanobis , link: average : 0.8326994115042136
affinity: mahalanobis , link: weighted : 0.7805990615142518
affinity: minkowski , link: complete : 0.8599730607972423
affinity: minkowski , link: average : 0.8977080867389372
affinity: minkowski , link: weighted : 0.8861746814895477
affinity: seuclidean , link: complete : 0.8599730607972426
affinity: seuclidean , link: average : 0.8977080867389373
affinity: seuclidean , link: weighted : 0.8861746814895477
affinity: sqeuclidean , link: complete : 0.8820964814996479
affinity: sqeuclidean , link: average : 0.8783309583061251
affinity: sqeuclidean , link: weighted : 0.893679734763713
```

Figure out the appropriate number of clusters

- Next, I need to decide on the number of clusters. There are two methods I use here to reach my conclusion; the first plot shows the silhouette score by the number of clusters. Based on this I would not want more than 4 clusters as the score gets too low. The second plot is the dendrogram, which plots the merging of the groups based on distance. The dendrogram shows a long distance (the y axis) for the two final groups and each of the subsequent groups has a drastically shorter distance. This would suggest to me that there are two distinct groups under this method.





- This method resulted in two clusters, with one cluster containing only 8% of the total records. It is also worth noting that the online user segment matches the same online user segment from Kmeans. This division seems to be the users who prefer internet transactions and those who do not. Once again, online users have more credit cards and a larger credit limit.
- Not shown below: when using 3, 4, 5, or 6 clusters, all but two of the clusters have less than five records. Obviously, such clusters don't provide meaningful data.

Hierarchical Clustering Grouping

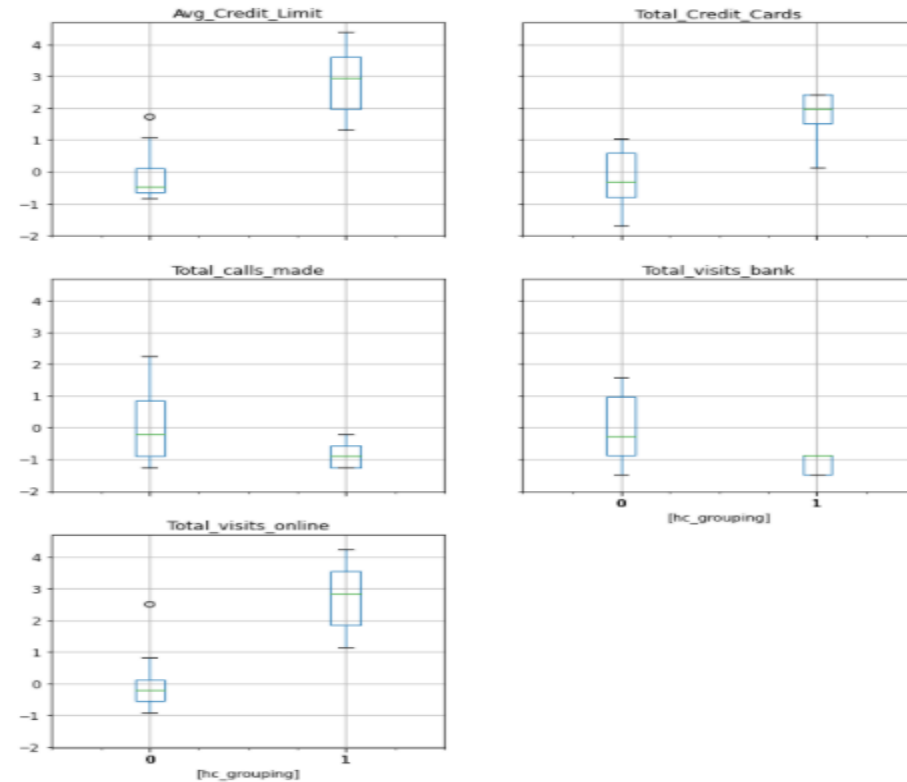
Here we add the predictions to the unscaled data so that we can gain some real-world interpretability.

	Customer Key	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made	Kmean_grouping
hc_grouping							
0	610	610	610	610	610	610	610
1	50	50	50	50	50	50	50

Below we see the average values for each feature in each cluster, the final row of the dataframe is the mean value for each column (since this data is unscaled, we can use this to determine how big of a step each of these means is from one another).

	Customer Key	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made	Kmean_grouping	hc_grouping
Offline	55012.97541	25847.540984	4.375410	2.550820	1.928230	3.788525	0.367213	NaN
Online	56708.76000	141040.000000	8.740000	0.600000	10.900000	1.080000	2.000000	NaN
Mean	25627.77220	37625.487804	2.167835	1.631813	2.935724	2.865317	0.634068	0.264811

- It would appear my hypothesis of the clusters forming along query method has proven correct. If we look at the data, we see that there is a group which prefers online interactions with their bank, they have a much higher credit limit and have more credit cards. The customers who prefer offline tend to have the least number of credit cards and the lowest credit limit.



Comparing K-means and Hierarchical Clustering

K- means Cluster Profiling

	Customer Key	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made	Kmean_grouping
Phone	55239.830357	12174.107143	2.410714	0.933036	3.553571	6.870536	NaN
Online	58708.760000	141040.000000	8.740000	0.800000	10.900000	1.080000	NaN
In person	54881.329016	33782.383420	5.515544	3.489637	0.981865	2.000000	NaN
Mean	25627.772200	37625.487804	2.167835	1.631813	2.935724	2.865317	0.930218

- ▶ **Cluster Profiles:**
- ▶ **Group 0:**
 - ▶ Customers with minimum credit limits (~ 12K in average).
 - ▶ They also have the least average number of credit cards (~ 2 cards each).
 - ▶ They tend to make phone calls rather than online and bank visits.
- ▶ **Group 1:**
 - ▶ Customers with middle credit limits (~ 34K in average).
 - ▶ They also have the middle average number of credit cards(~ 6 cards each).
 - ▶ They tend to visit the bank more often rather than making calls and online transactions.
- ▶ **Group 2:**
 - ▶ Customers with maximum credit limits (~ 140K in average).
 - ▶ They also have the maximum average number of credit cards(~ 9 cards each).
 - ▶ They tend to make online transactions rather than phone calls and bank visits.

Hierarchical Cluster Profiling

	Customer Key	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made	Kmean_grouping	hc_grouping
Offline	55012.97541	25847.540984	4.375410	2.550820	1.926230	3.788525	0.367213	NaN
Online	56708.76000	141040.000000	8.740000	0.600000	10.900000	1.080000	2.000000	NaN
Mean	25627.77220	37625.487804	2.167835	1.631813	2.935724	2.865317	0.634068	0.264811

► Cluster Profiles:

► Group 0:

- Customers with minimum credit limits (~ 25K in average).
- They also have the least average number of credit cards (~ 4 cards each).
- They tend to make phone calls rather than online and bank visits.

► Group 1:

- Customers with middle credit limits (~ 141K in average).
- They also have the middle average number of credit cards(~ 8 cards each).
- They tend to visit the bank more often rather than making calls and online transactions.

Insights About Different Clusters


- ▶ **Group 1** own less credit card than others, bank should target **group 1** to upsell credit cards services.
- ▶ Besides, bank should provide higher credit limit to target **group 0** where most of the customers are. With higher credit limit, group 0 would be able to spend more.
- ▶ Since **group 0** use the online banking the least, bank should promote more to **group 0** in order for them to use it.
- ▶ Assuming **group 1** who make most phone calls are the customers perceive the support services of the bank poorly. Bank should target **group 1** and provide better customers service by conducting feedback survey through phone.

Key Questions Answer

- ▶ There are three distinct categories of customers in AllLife Bank Credit card customer base.
- ▶ How are these segments different from each other?
- ▶ **In-person Users:** Prefer to handle bank transactions in person. They have the fewest credit cards and the lowest available credit. They are also the most active users.
- ▶ **Phone Users:** Prefer verbally handling transactions remotely.
- ▶ **Online Users:** Prefer digital transactions. They also have the most credit cards and the highest available credit.
- ▶ What are your recommendations to the bank on how to better market to and service these customers?
- ▶ **Recommendations:** We can tailor contact methods to these customer preferences. Online/phone users will probably prefer email/text notifications, while in-person users will prefer mail notifications and upselling (when at the bank location). Since online users tend to have (and presumably use) the most credit, these may be the demographic we want to target with our next ad campaign, focusing on digit recruiting.

Insights and Recommendations for the Future Business

- ▶ **Insights and Recommendations**
- ▶ **Group 0:** Business can offer better phone service to these customers as they are a sizeable percentage of population. Feedback on service delivery can be requested from customers to improve perception of delivery. New customers could be added to this category by offering cards with low credit limit. They can be offered more credit cards as their credit card utilization is low.
- ▶ **Group 1:** Business can offer better in person service to these customers as they are a sizeable percentage of population. They can be offered more information about online banking for some services, keeping their convenience in mind.
- ▶ **Group0:** Business can look to add more customers in this category, as this cluster is a small percentage of the customers population.

- 
- ▶ **Group 1:** Business can contact these customers to offer credit cards or to get feedback about service.
 - ▶ **Group 0:** We can tailor contact methods to these customer preferences. Online/phone users will probably prefer email/text notifications, while in-person users mail prefer mail notifications and upselling (when at the bank location). Since online users tend to have (and presumably use) the most credit, these may be the demographic we want to target with our next ad campaign, focusing on digit recruiting.