AllLifeBank Customer Segmentation

By Jacob Siegel

Business Context

- The Marketing team for AllLifeBank wants to run personalized campaigns to target new customers as well as upsell to existing customers.
- Customers perceive the support services of the back poorly.
- Upgraded service delivery model needed to ensure that customer's queries are resolved faster.
- Goal: To identify different segments in the existing customer based on their spending patterns as well as past interaction with the bank.

Data Overview and Pre-Processing

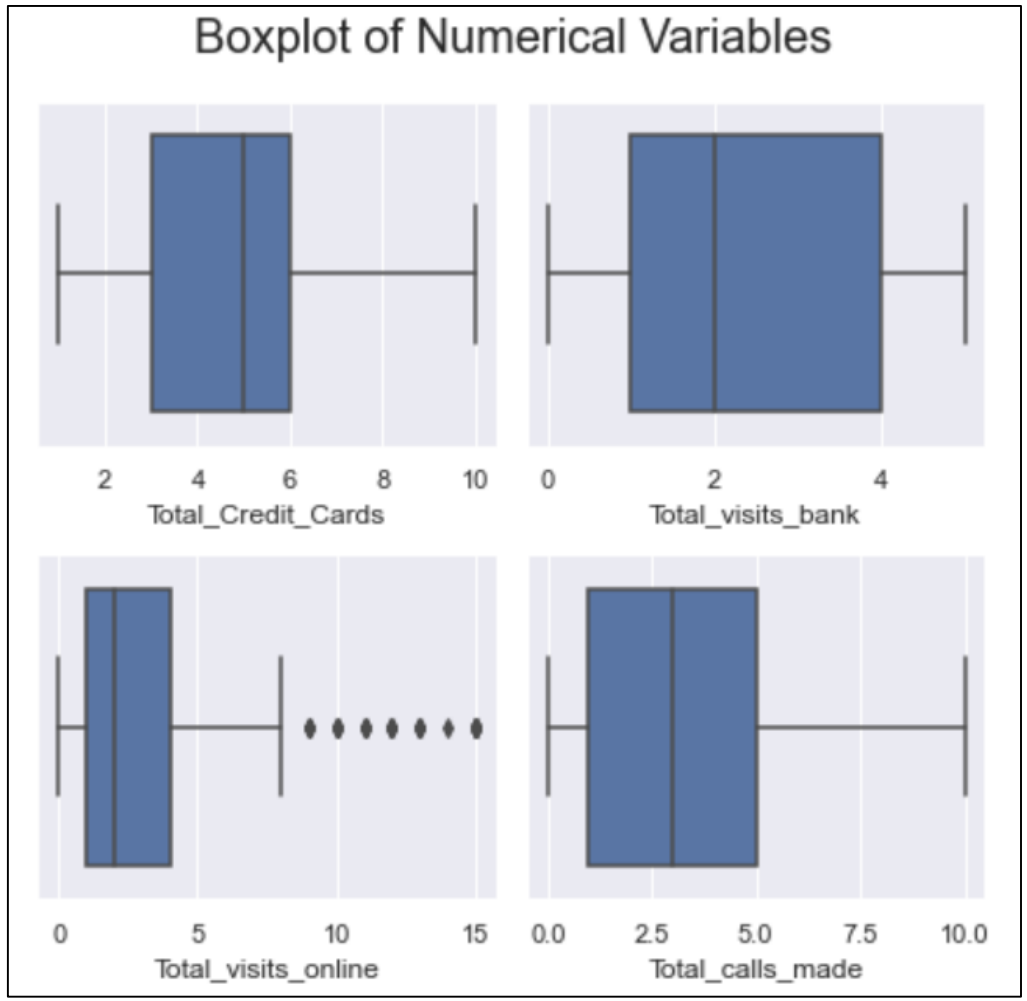
- The raw dataset contains 660 rows and 7 columns.
- 2 columns ('SI_No' and 'Customer Key') were dropped and not included in the analysis as they were all unique values.
- 11 duplicate rows were removed. There were no missing values.

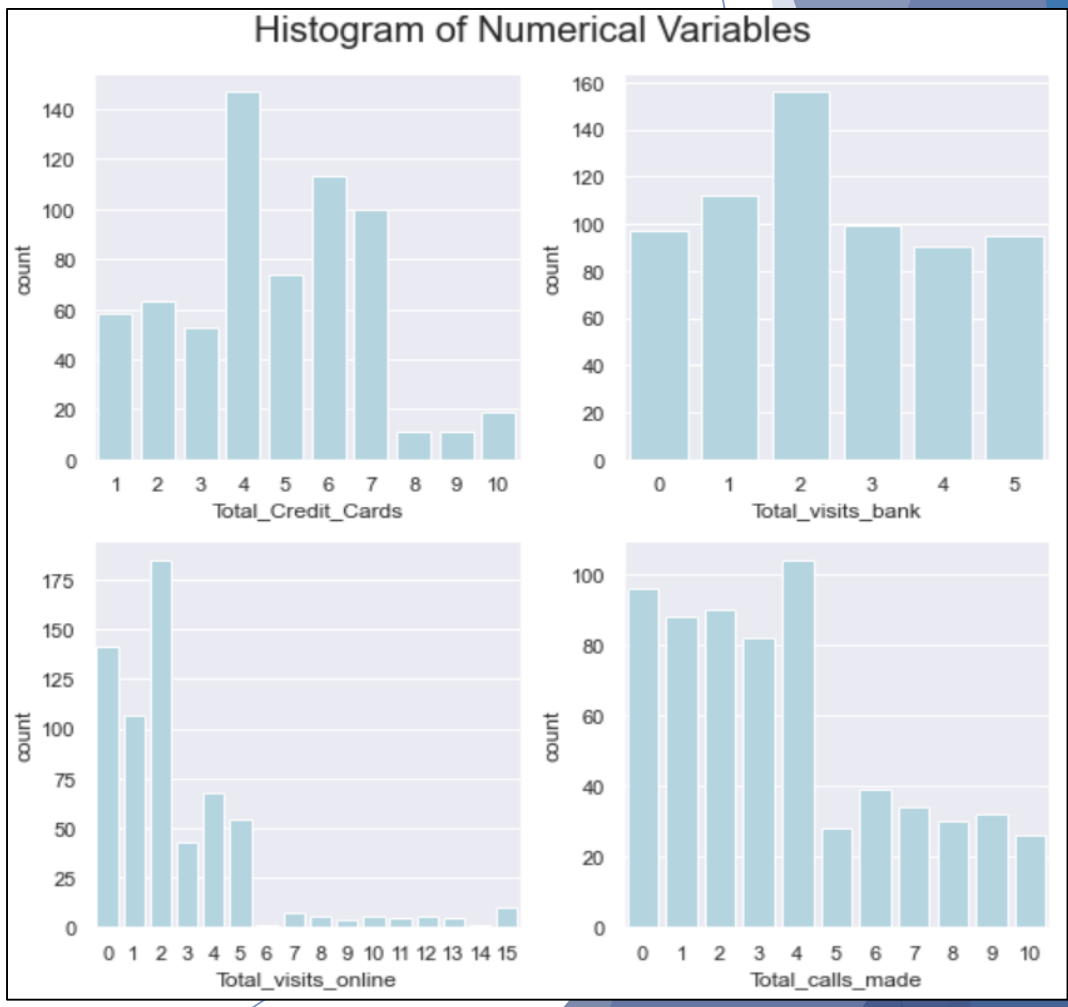
Table: Summary of Data Columns

	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made
count	649.000000	649.000000	649.000000	649.000000	649.000000
mean	34878.274268	4.708783	2.397535	2.624037	3.590139
std	37813.736638	2.173763	1.625148	2.952888	2.877911
min	3000.000000	1.000000	0.000000	0.000000	0.000000
25%	11000.000000	3.000000	1.000000	1.000000	1.000000
50%	18000.000000	5.000000	2.000000	2.000000	3.000000
75%	49000.000000	6.000000	4.000000	4.000000	5.000000
max	200000.000000	10.000000	5.000000	15.000000	10.000000

EDA: Univariate Analysis

- Total Visits to Bank is evenly distributed. Total Calls Made has more people with 4 or fewer calls.
- Total visits online is the most skewed with the majority of people having less than 5 visits, and a select few people had more than 5 visits. A few points are outside of the box plot, but they are not considered outliers.

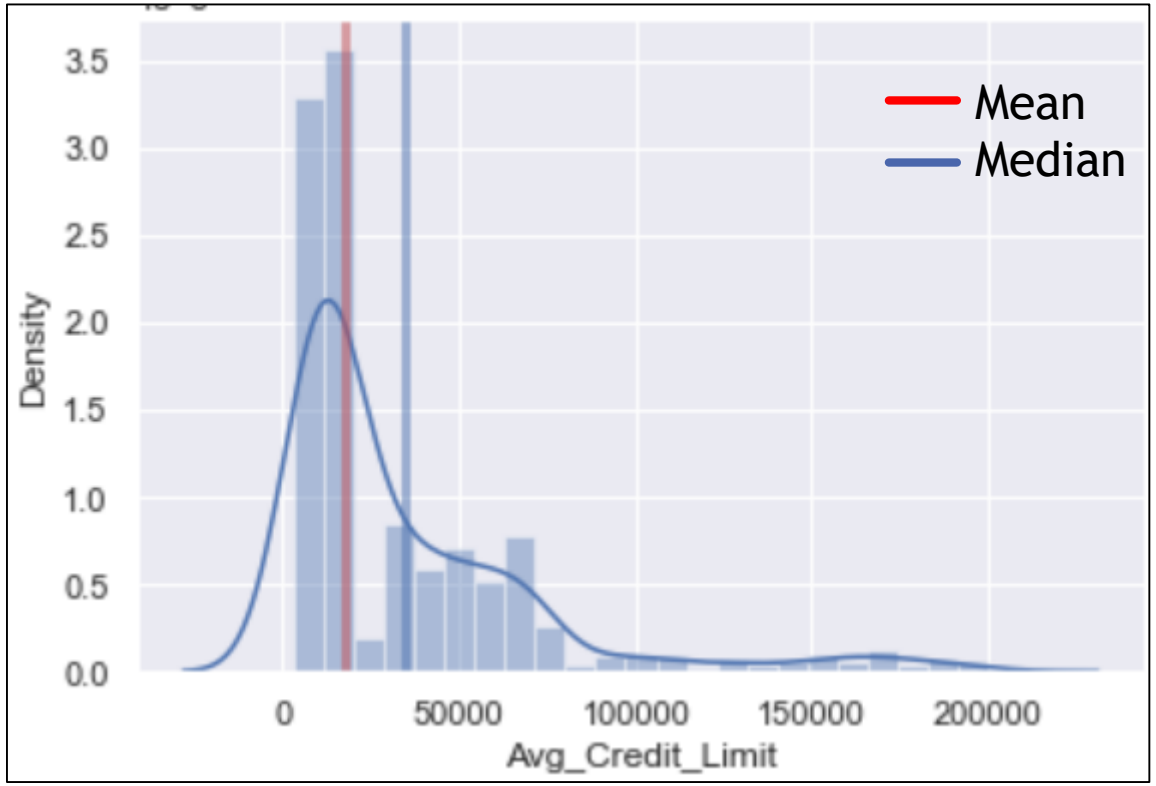




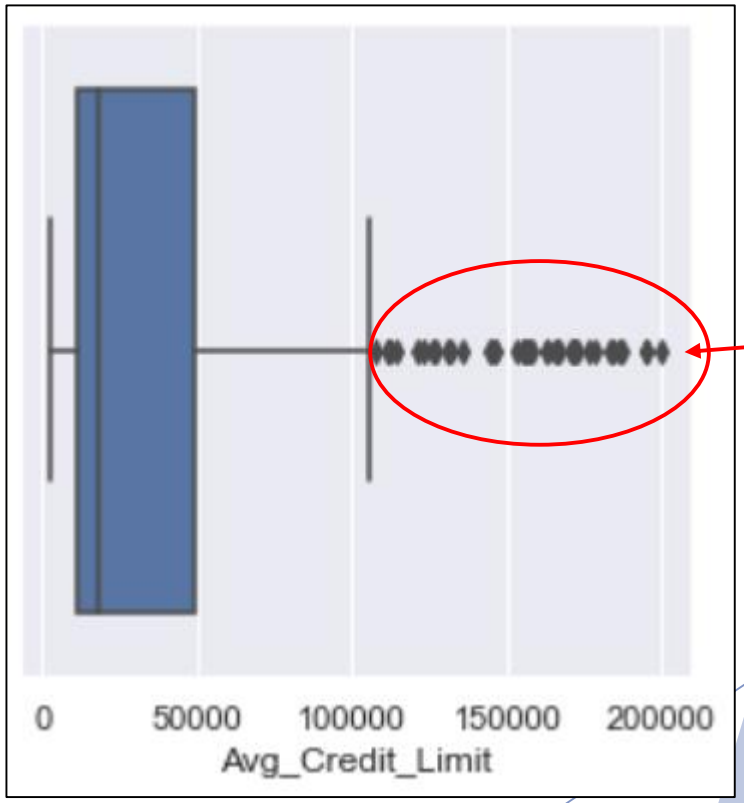
EDA: Univariate Analysis

- Average credit limit is positively skewed with a few people having a considerably higher credit limit than the average.

Distribution Plot for Average Credit Limit



BoxPlot for Average Credit Limit

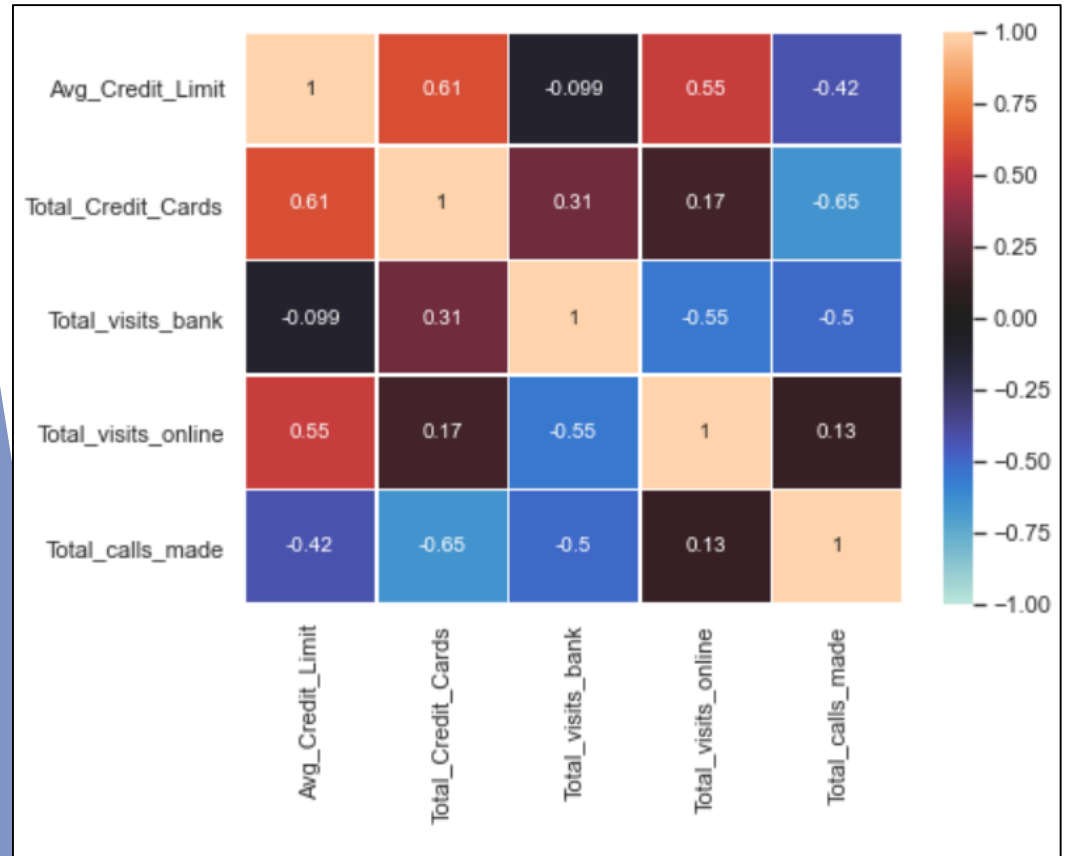


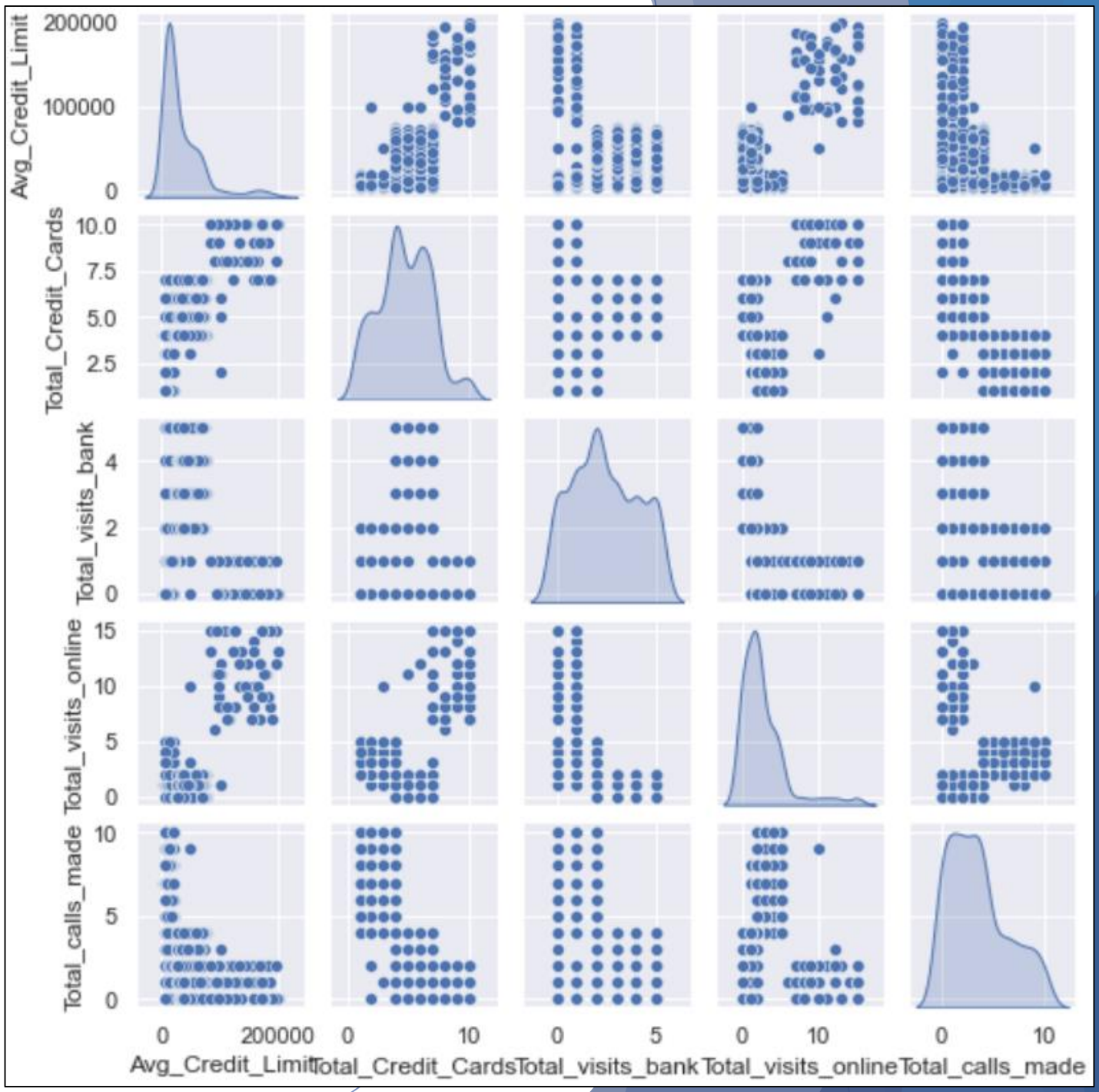
Several People have credit limit much higher than average

EDA: Bivariate Analysis

- There are no strong correlations among the variables.
- It should be noted that the data distribution is blocky and does not appear to be completely random.

Correlation Plot of All Variables

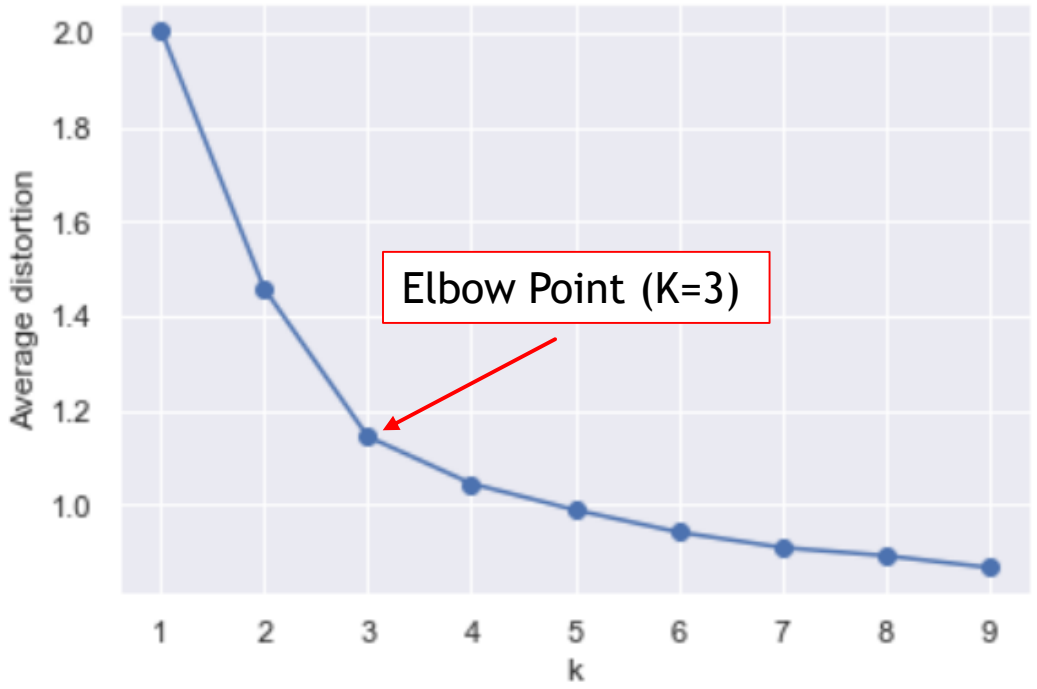




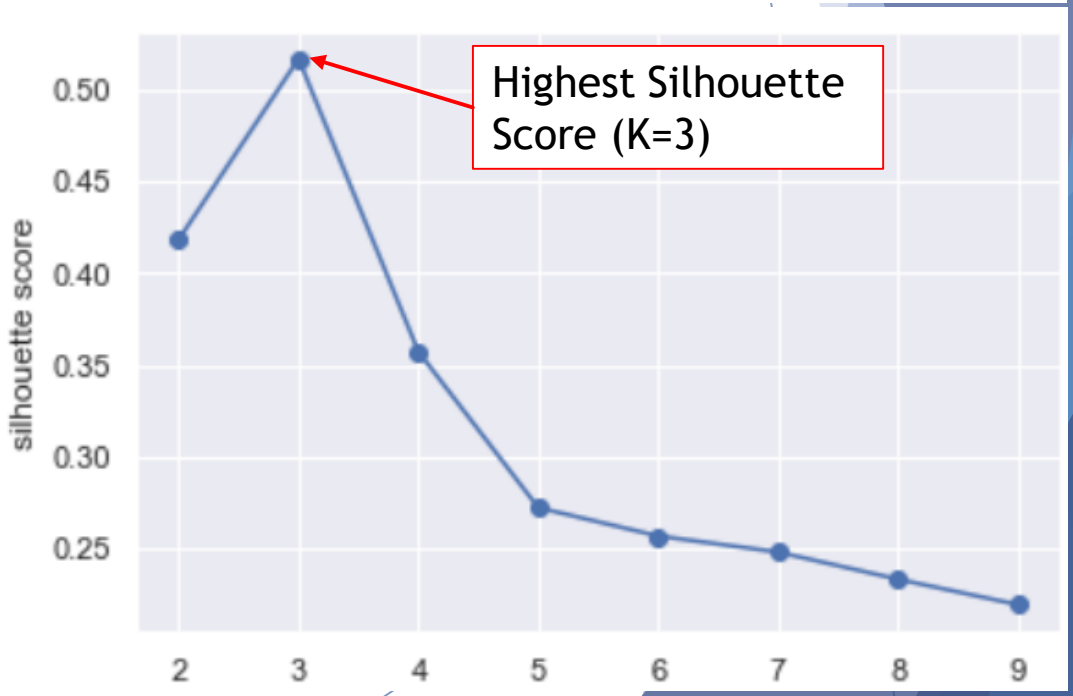
K Means Clustering Model Parameters

- A K means cluster model was run with 1 to 10 clusters. Based on the elbow plot and the silhouette score, an optimal number of clusters was determined to be 3.

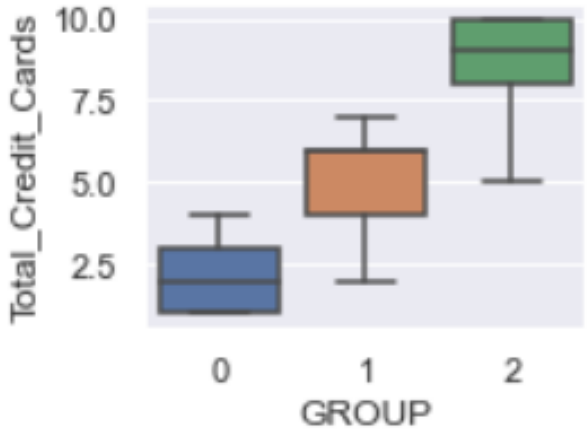
Average Distortion per K (Elbow Method)

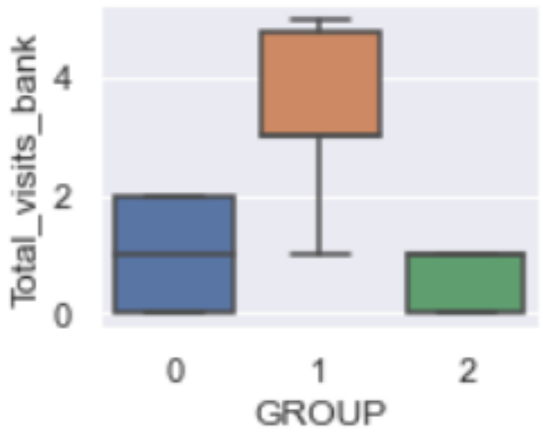


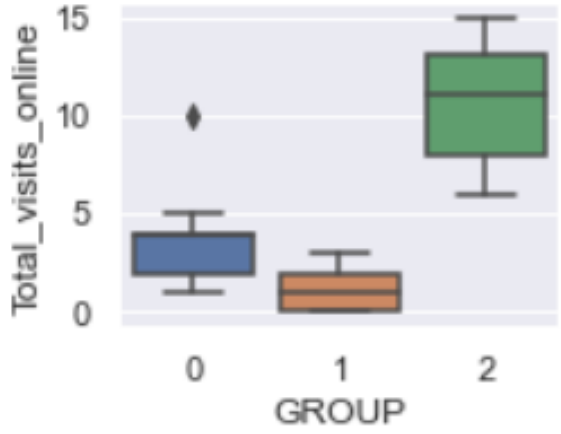
Silhouette Score per K

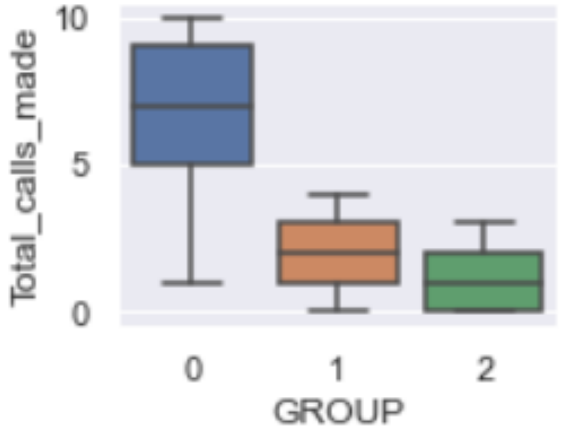


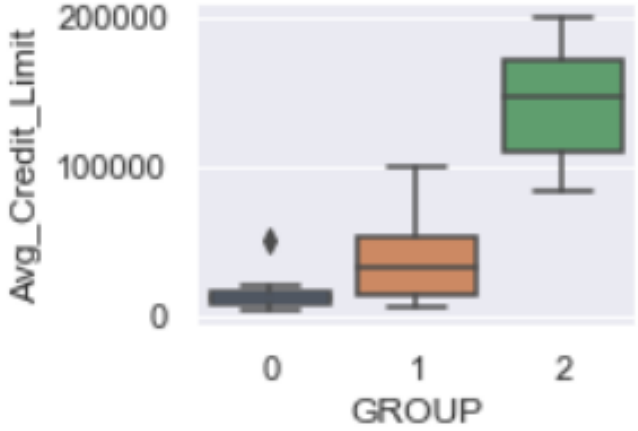
K Means Clustering Results











- GROUP 0: least credit card spend, least visits to the bank, least average credit limit.
- GROUP 1: most visits to the bank, least visits online.
- GROUP 2: Highest number of credit cards and highest credit card limit. This is also the smallest group with only 50 people.

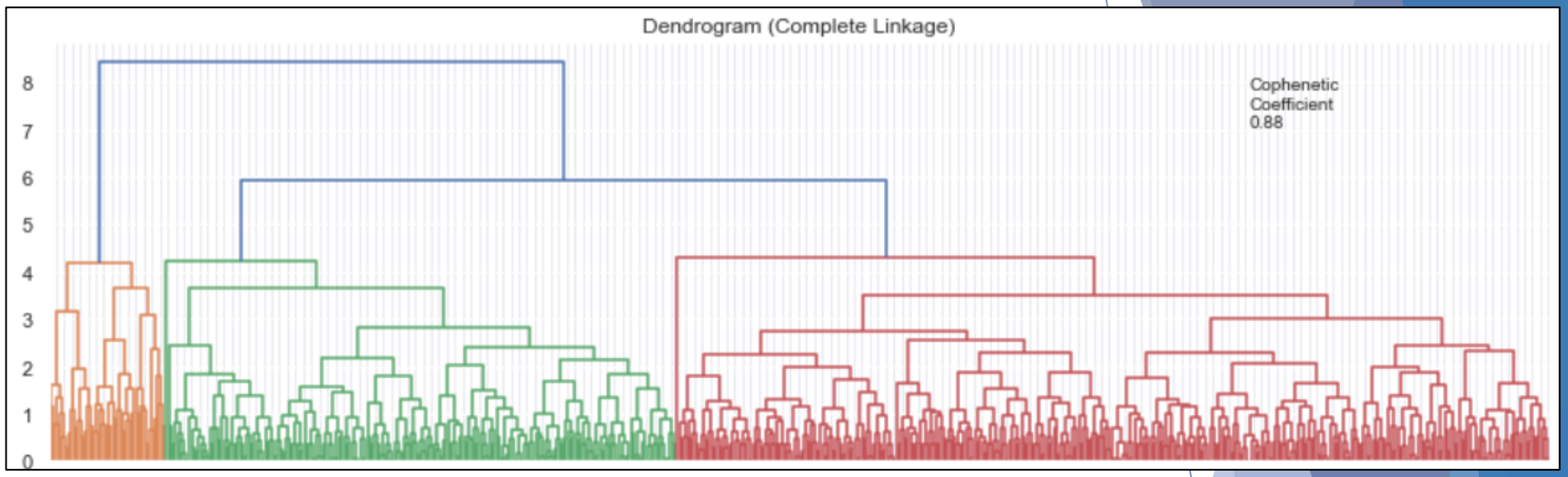
Summary Table: Average Properties of Groups

GROUP	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made	Number_In_Group
0	12239.819005	2.411765	0.945701	3.561086	6.891403	221
1	34071.428571	5.518519	3.484127	0.981481	1.992063	378
2	141040.000000	8.740000	0.600000	10.900000	1.080000	50

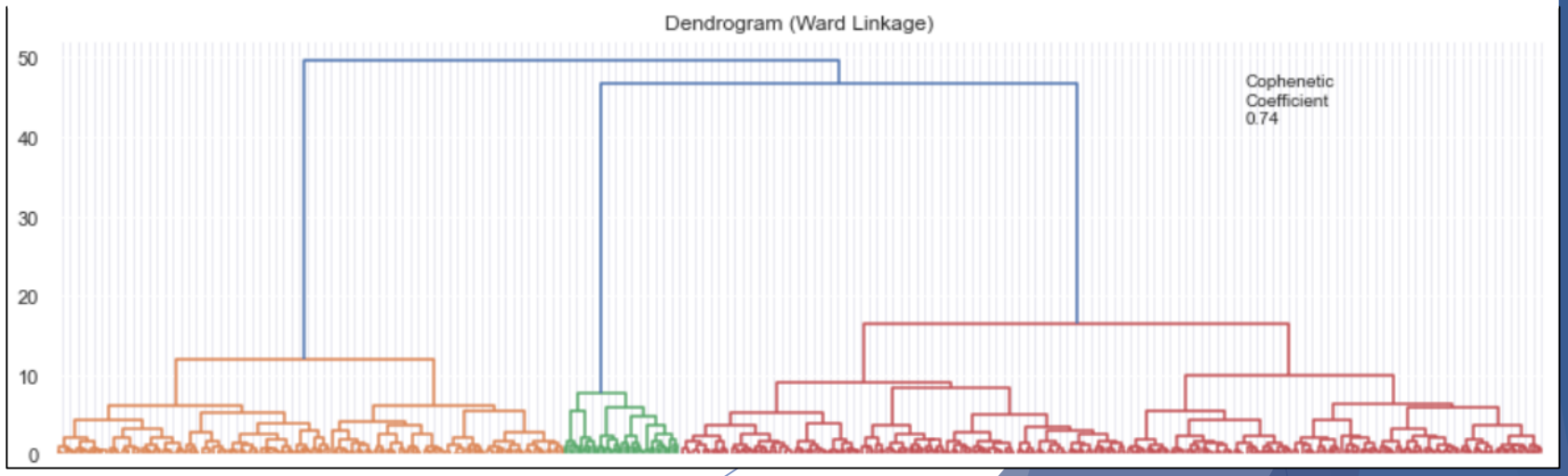
Hierarchical Clustering Model Parameters

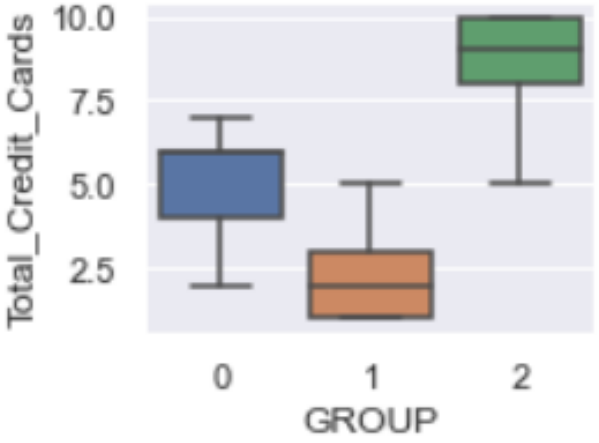
- A hierarchical clustering model was created for each of the following methods: 'single', 'average', 'complete', 'centroid',‘ ward', and'weighted‘
- Based on the distribution of the dendograms and the cophenetic score, Complete Linkage and Ward Linkage were selected as the final model with 3 clusters each.

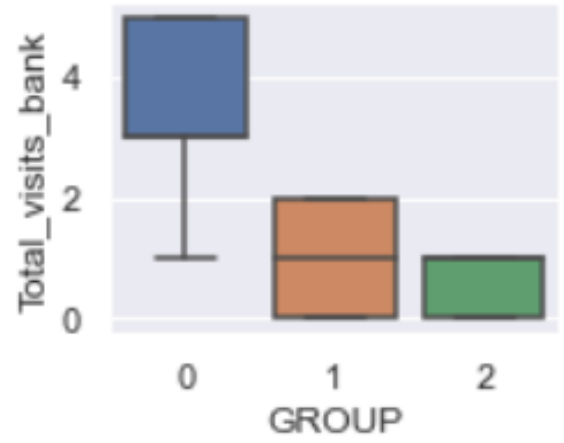
Complete Linkage Dendogram

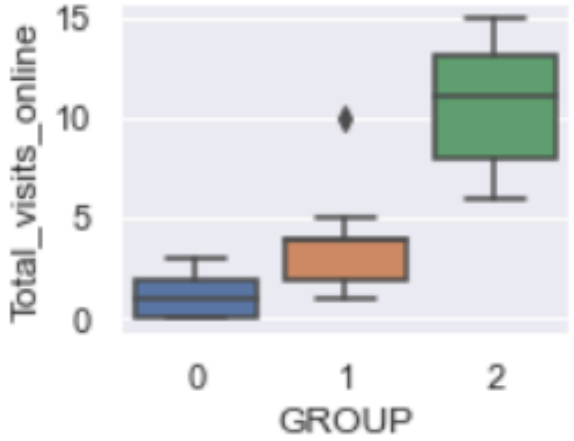


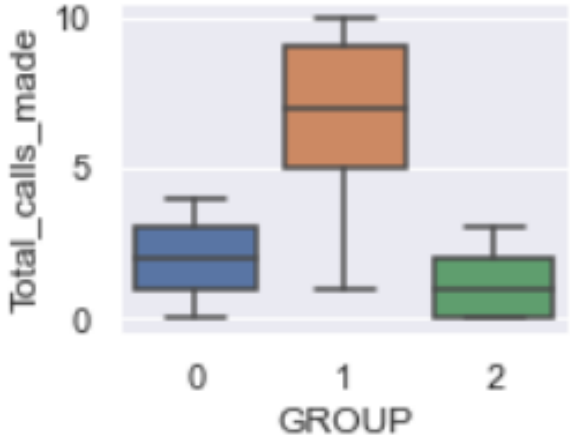
Ward Linkage Dendogram

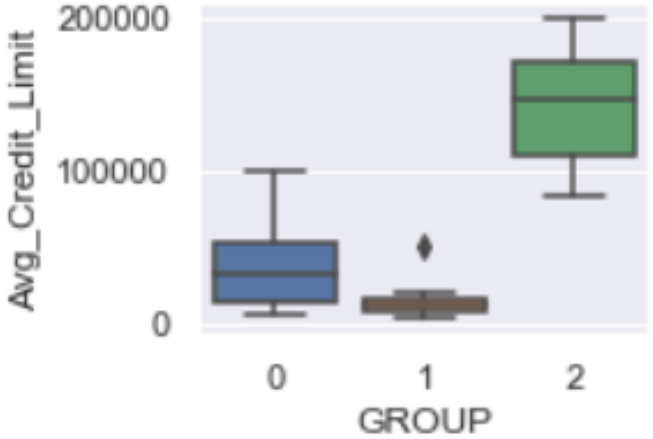












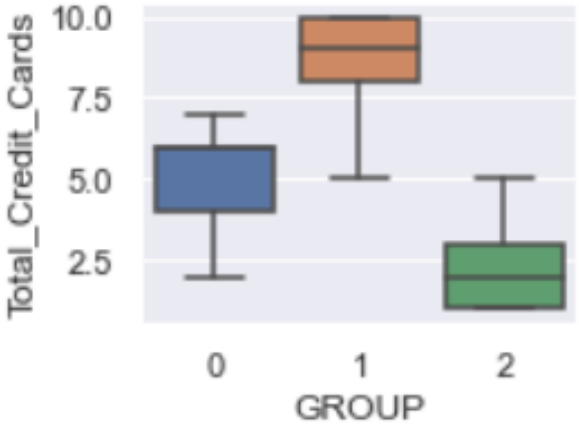
Hierarchical Clustering: Ward Linkage Results

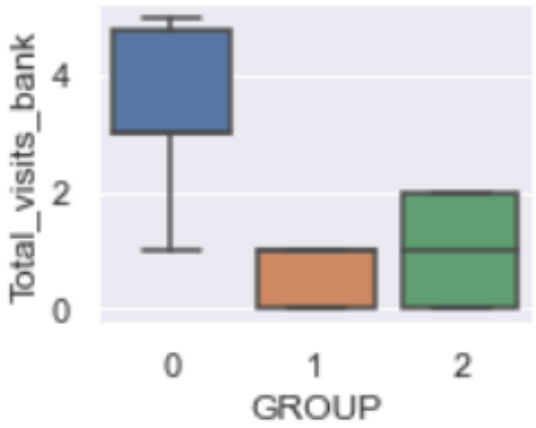
- GROUP 0: Has the most visits to the bank but the least visits online.
- GROUP 1: Has the most calls made and least calls made and least average credit limit.
- GROUP 2: Similar to the previous model, this group is the smallest with the highest number of credit cards, average credit limit and online visits.

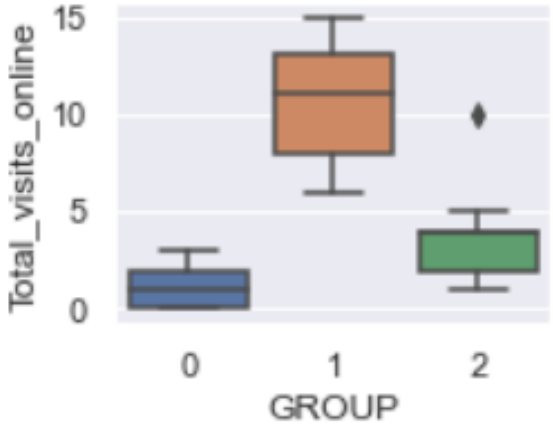
Summary Table: Average Properties of Groups

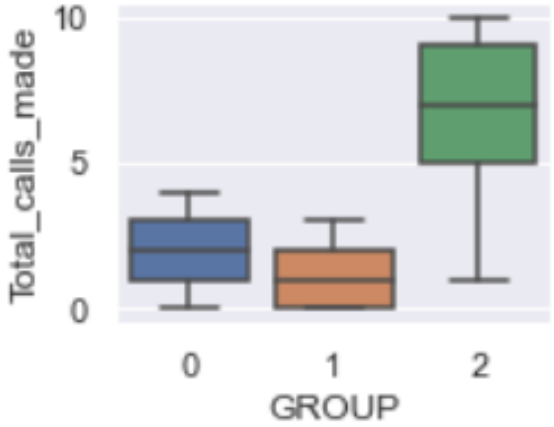
GROUP	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made	Number_In_Group
0	34143.236074	5.519894	3.488064	0.978780	1.986737	377
1	12216.216216	2.423423	0.950450	3.554054	6.878378	222
2	141040.000000	8.740000	0.600000	10.900000	1.080000	50

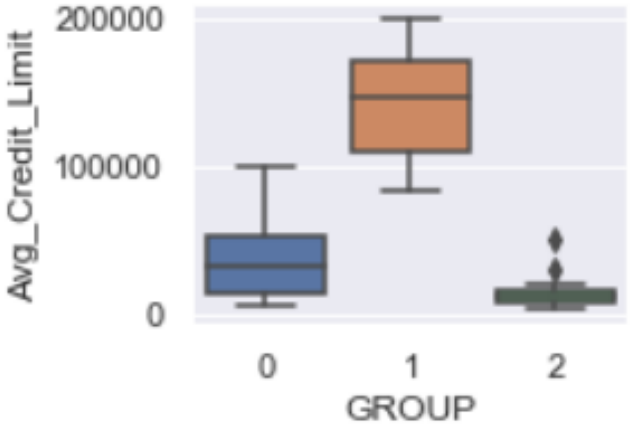
Hierarchical Clustering: Complete Linkage Results











- Group 0: has the highest total visits to the bank and the lowest total visits online.
- Group 1: Has the highest average credit limit, the highest total visits online and the highest total credit cards.
- Group 2: Has the lowest credit cards and the lowest average credit limit.

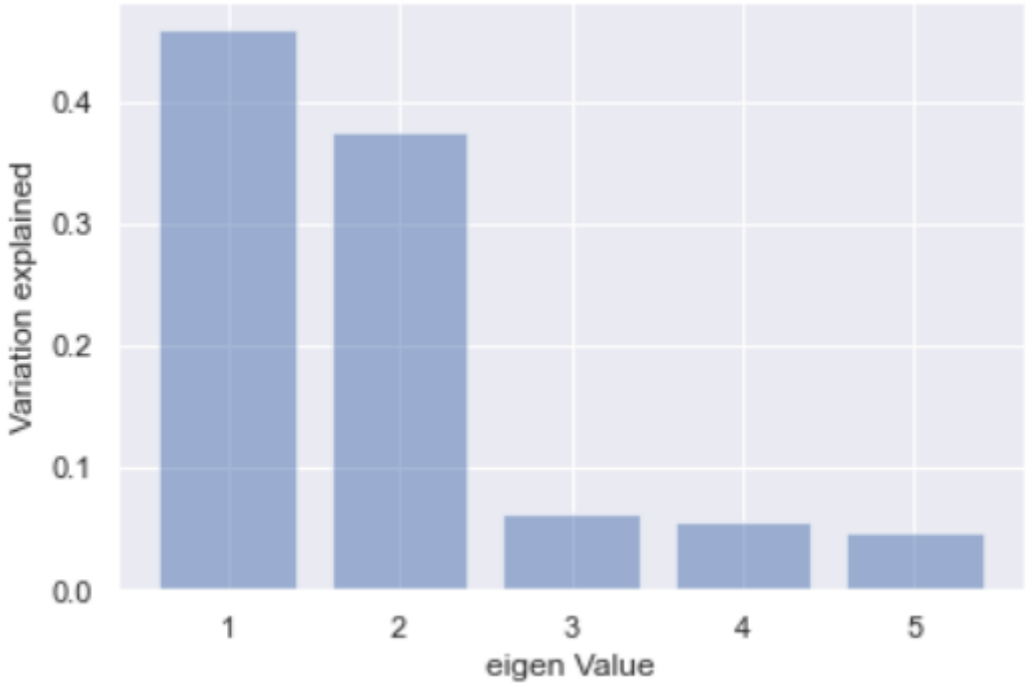
Summary Table: Average Properties of Groups

GROUP	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made	Number_In_Group
0	34010.582011	5.515873	3.486772	0.984127	1.992063	378
1	141040.000000	8.740000	0.600000	10.900000	1.080000	50
2	12343.891403	2.416290	0.941176	3.556561	6.891403	221

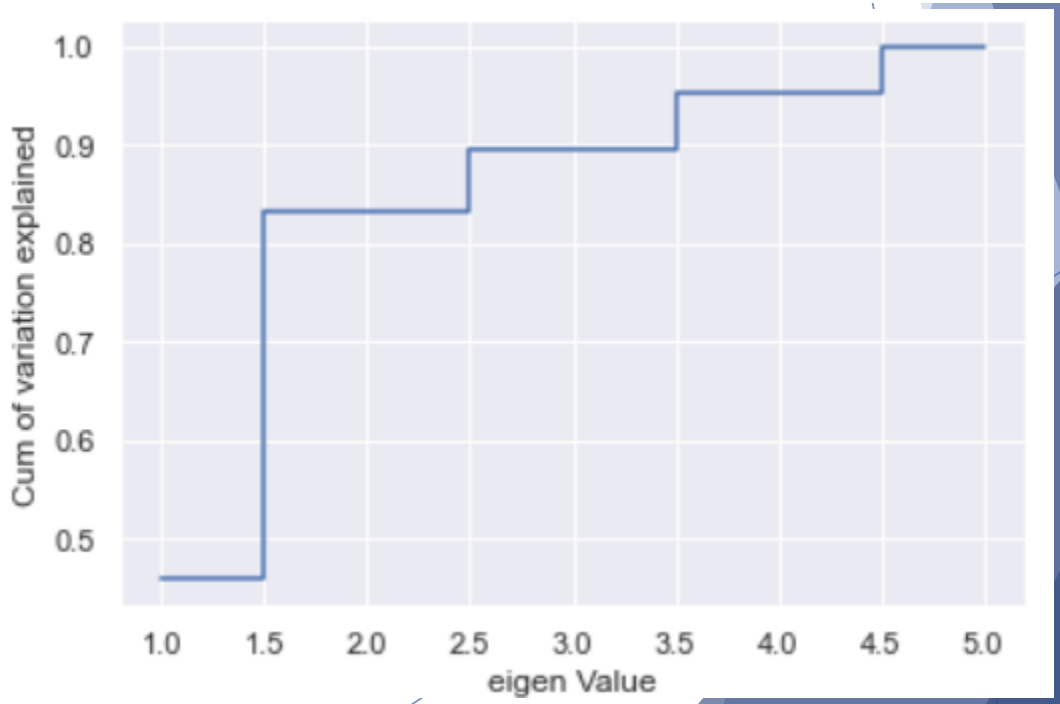
Principle Component Analysis Model Parameters

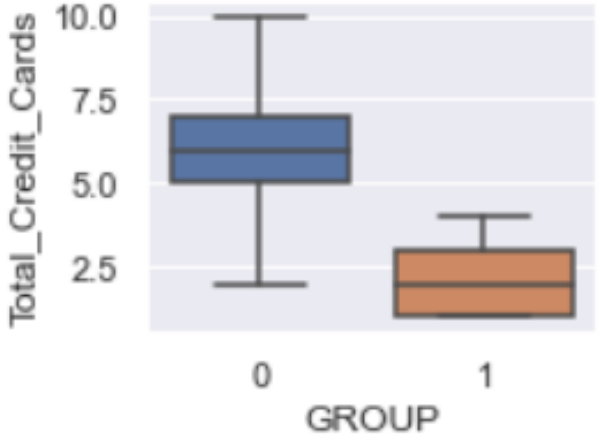
- The first two principle components are able to explain 83% of variance in the data and the final model is run with just 2 components.

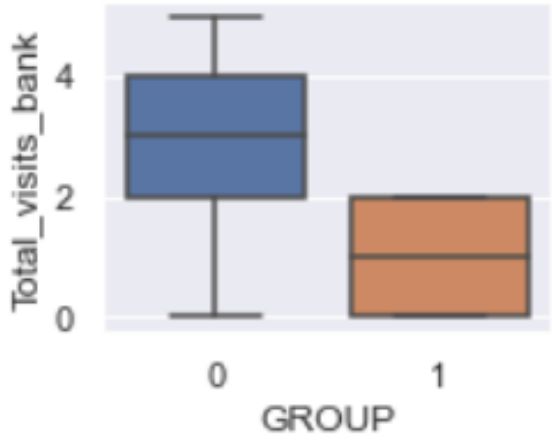
Variance Explained by each Eigen Value

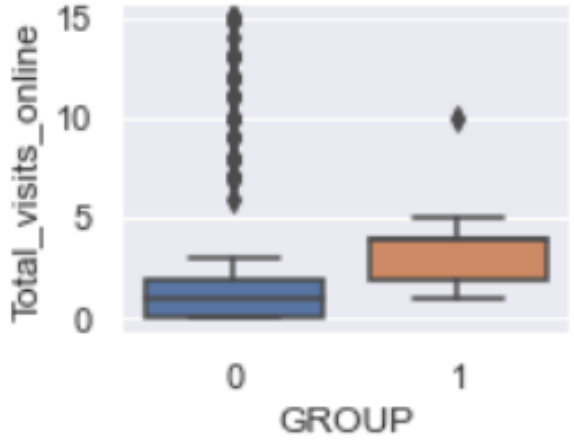


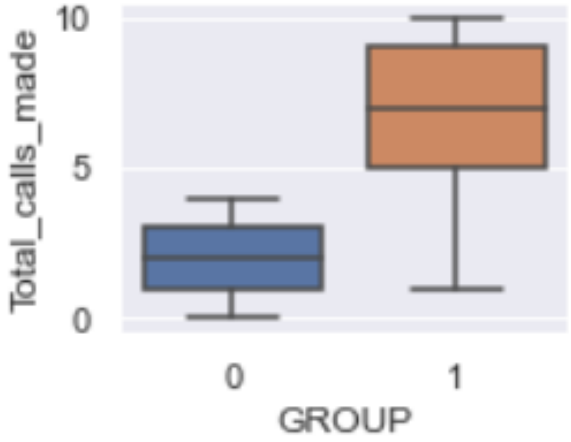
Cumulative Variance by each Eigen Value

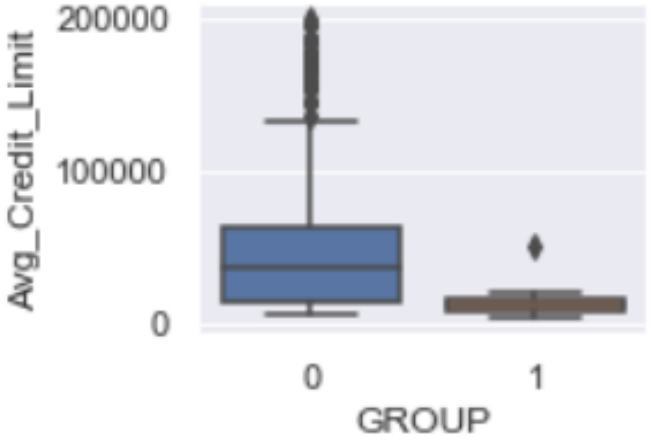












Principle Component Analysis Results

- Group 0: Higher credit limit and higher average spend.
- Group 1: has more online visits and less visits to the bank.

Summary Table: Average Properties of Groups

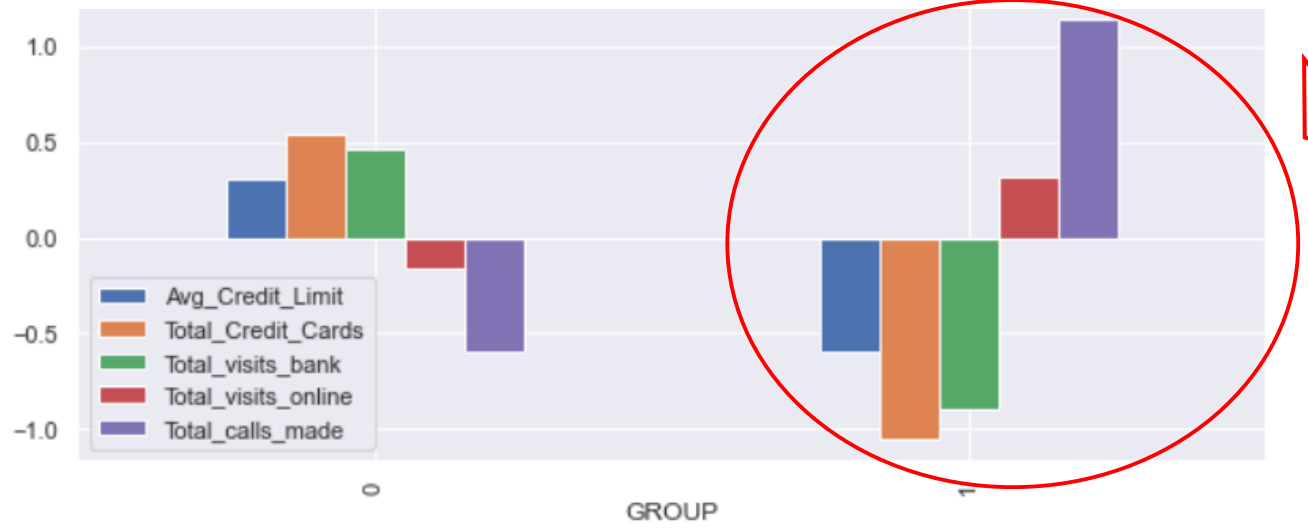
GROUP	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made	X1	X2	Number_In_Group
0	46567.757009	5.894860	3.147196	2.140187	1.885514	0.915769	-0.381988	428
1	12239.819005	2.411765	0.945701	3.561086	6.891403	-1.773526	0.739777	221

Comparing K Means Clustering to PCA

- Group 0 from the K means cluster model and Group 1 from the PCA model are nearly the same.
- Group 1 and Group 2 from the K means cluster model combine to form Group 0 in the PCA model. This removes Group 2 which was the smallest group.

K Means Summary Plot: Average Properties of Groups (scaled)

PCA Summary Plot: Average Properties of Groups (scaled)

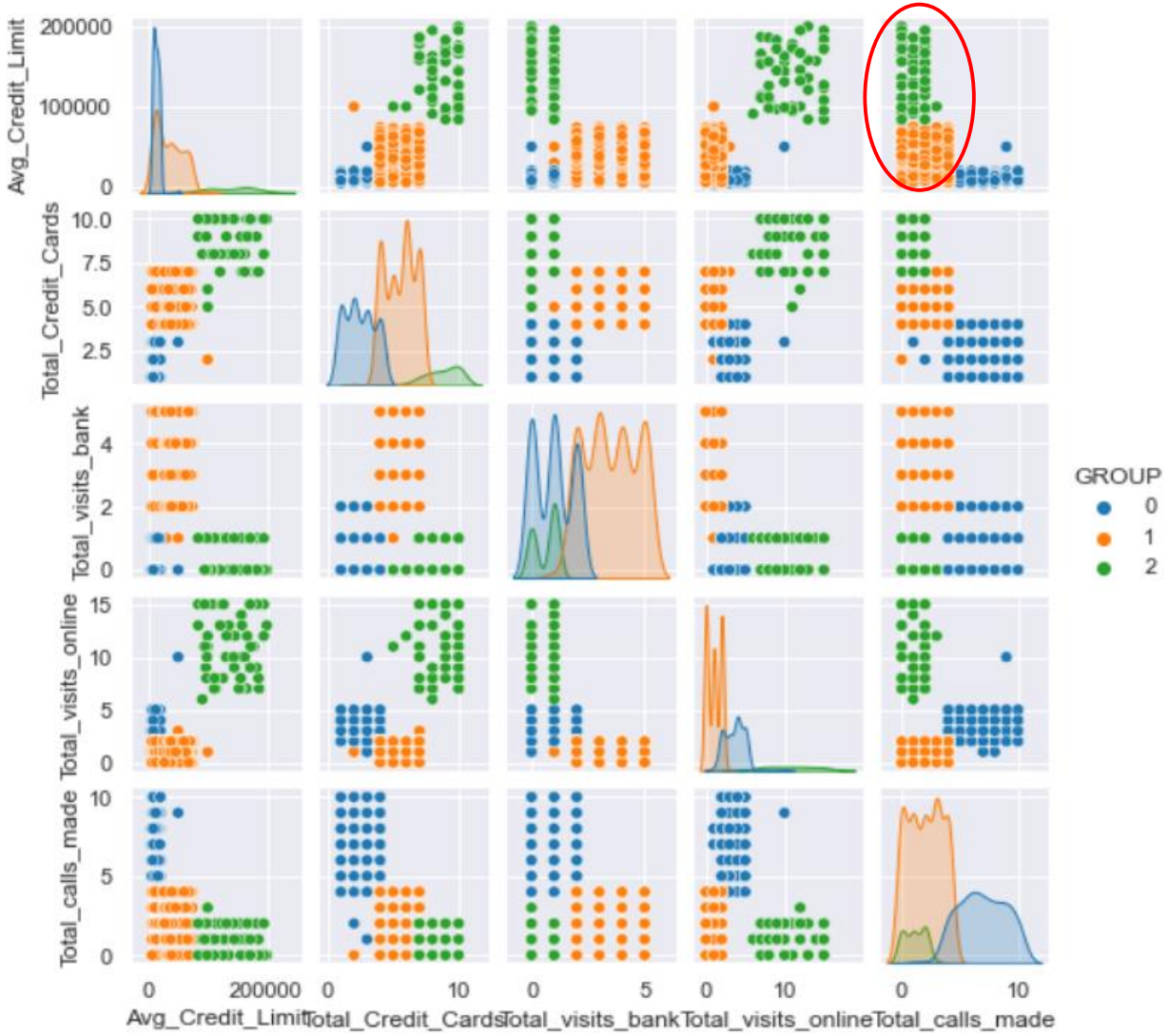


Similar Group Properties

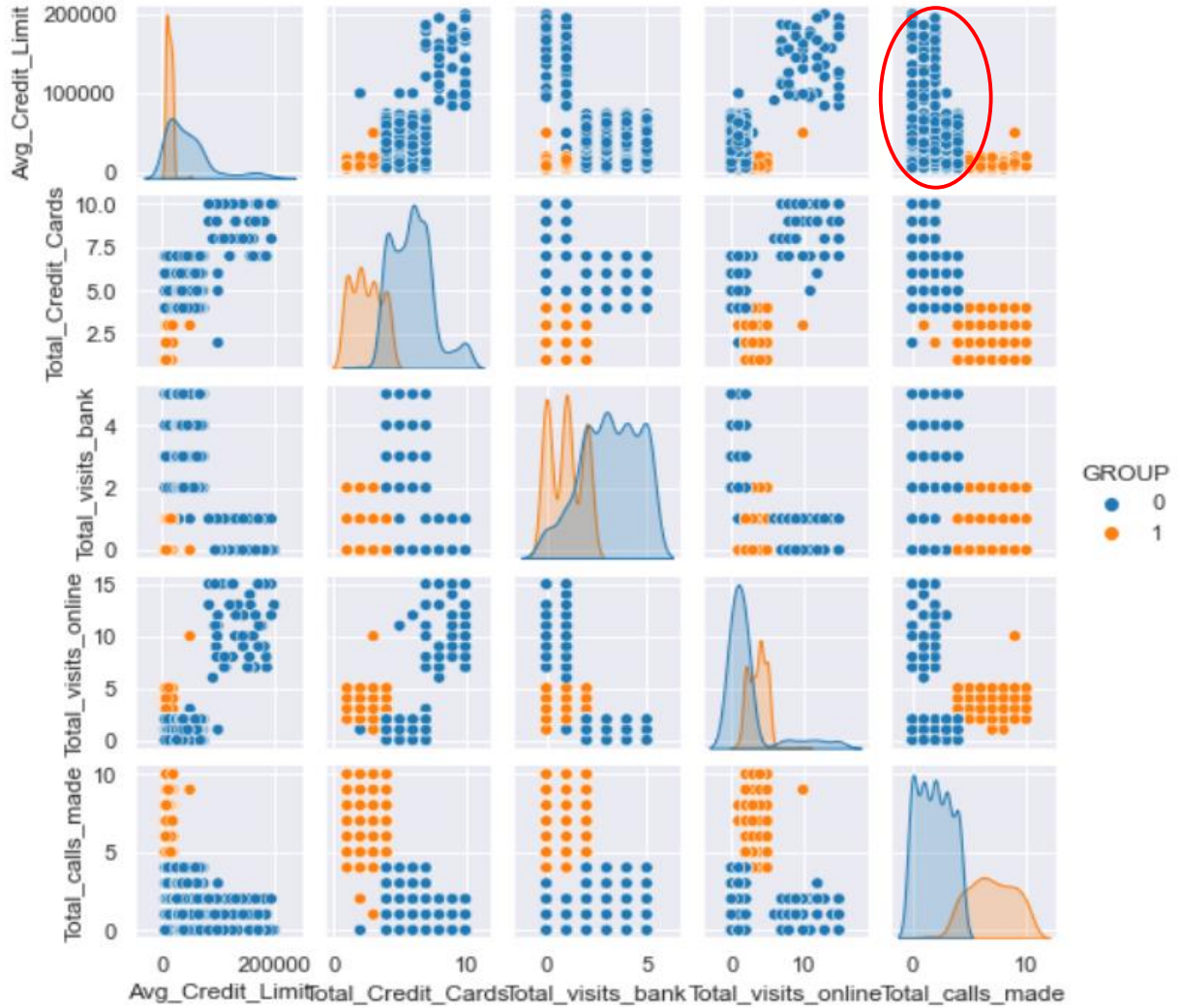
Comparing K Means Clustering to PCA (continued)

- The highlighted circle shows how groups 1 and 2 from K means combine to be the same group in PCM

K Means Distribution Plot with Groups



PCA Distribution Plot with Groups



Conclusions and Recommendations

Conclusions:

- 3 clusters was determined to be the optimal for K means and PCA while 2 clusters explains more than 80% of the variance with just 2 components.
- Both K means and Hierarchical clustering algorithms identified 3 groups that can be characterized as High, Medium, and Low average credit limit. The high average credit limit is the smallest group (50) with the highest average credit limit, highest number of credit cards and the most visits online. This group was not identified by the PCA model that was restricted to 2 components.

Recommendations:

- **High Credit Limit Group:** The smallest cluster of people with the highest credit limit and highest number of credit cards are more likely to visit the bank online, and thus they can be targeted more directly with online advertising.
- **Medium Credit Limit Group:** The largest cluster of people with the average credit limit and number of credit cards visit the bank the most. They will be better targeted with bank tellers and bank managers when the visit a bank in person.
- **Low Credit Limit Group:** The last group has the lowest average credit. This group makes the most calls to the bank and would be best served my specific marketing over the telephone.