All Life Bank

Case Study

Sven Meydell

Areas of Focus

- **Core Business Idea:**
  - Take **existing company data** detailing historic customer spending and company interactions and **create unique customer profiles (clusters)**, allowing for **targeted marketing campaigns** and better customer service aligned with **customer patterns (preferences)** for each subgroup/profile

- **Financial Implications:**
  - The various clustering techniques should allow for better grouping of customers based on **key metrics within the data, not easily identifiable to the naked eye**
  - These newly created customer profiles should allow the bank to directly target and **market to specific subsets of customers in accordance with their clustered patterns,** offering the following benefits:
    - Efficiency and savings in marketing spend (promotions/campaigns)
    - Better conversions (sales/retention) as it relates to Credit Card usage

Solving Problems with ML

- **Problem:**
  - **Identifying different segments** in the existing customer based, based on historic spending patterns and interactions with the bank
  - These new segments will allow the Marketing team to **better target new and existing customers** run personalized campaigns and additional upsells
  - Additional insight can be gathered and presented to the Operations team regarding customer contacts to **improve customer service platform**
- **Solution:**
  - Through a variety of clustering techniques (namely k-Means and Hierarchical), unique patterns can be identified and segmented through the **combination of various data dimensions, computed simultaneously**, based on the distances within the initial clusters and nearest clusters
    - When coupled with **Exploratory Data Analysis** (graphs), unique insights can quickly be gleamed, analyzed and presented, and quickly acted upon
    - Prior clusters can be compared against with future data, **to identify new patterns (also Fraud) and ensure that the best possible groupings are achieved**

Objectives

- To identify **different segments in the existing customer base** and recommend best approaches to better service these specific customers
  - Patterns, not easily identifiable in original dataset, should be determined through **various clustering algorithms and techniques**

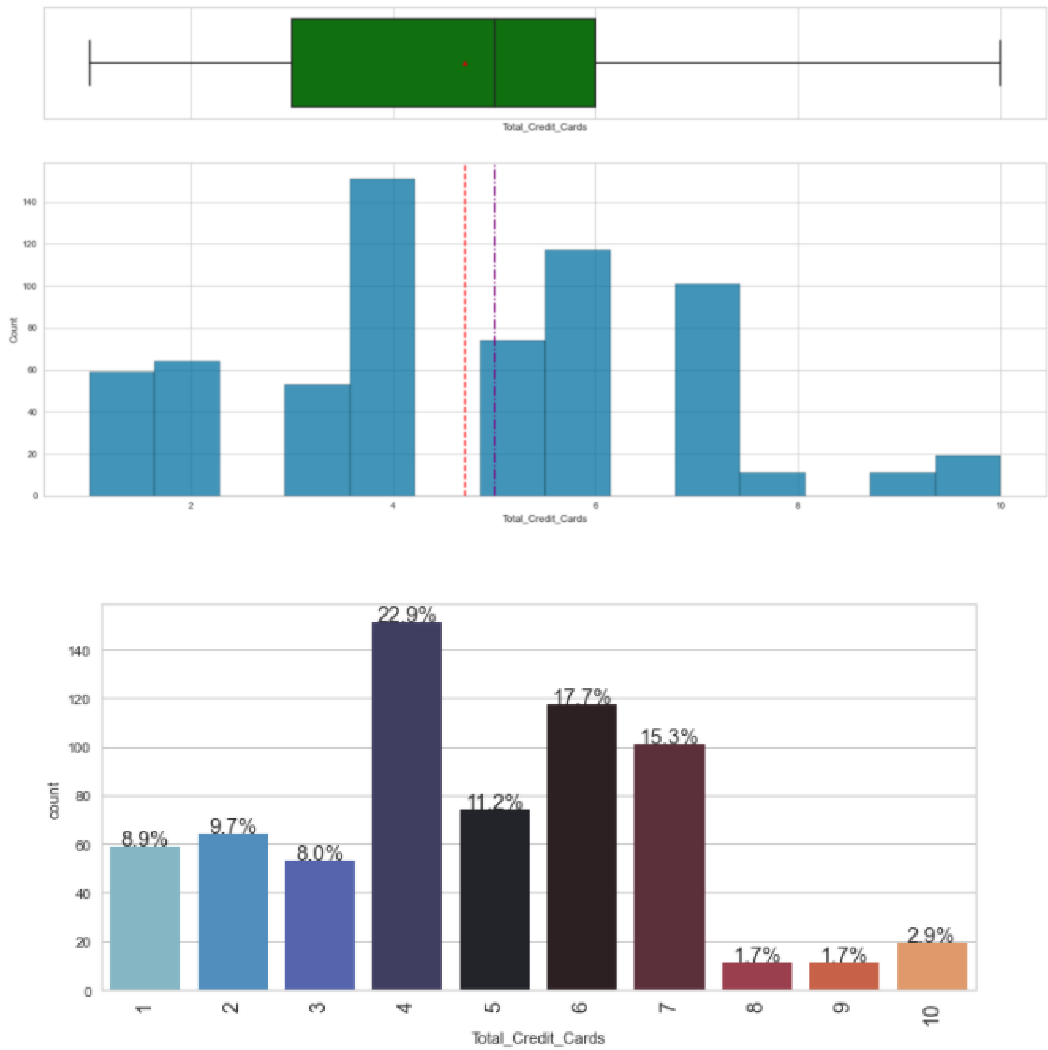
Data Provided:
Customer Details

- **SI No:** Primary index key of all records (Unique)
- **Customer Key:** Customer identification number (Unique)
- **Average Credit Limit:** Average credit limit of each customer for all credit cards
- **Total Credit Cards:** Total count of credit cards owned by each customer
- **Total Visits Bank:** Total number of annual visits made by customers to the bank
- **Total Visits Online:** Total number of annual online visits (including online logins) made by the customer
- **Total Calls Made:** Total number of annual support calls made by the customer to the bank (and/or service department)

Manipulating/Examining Raw Data

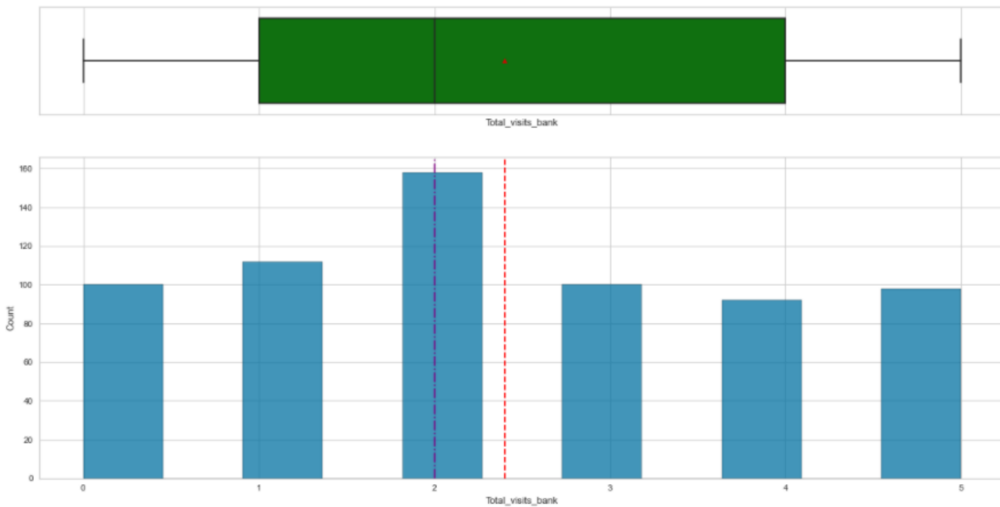
- Removal of **SI_No and Customer Key** variable as it offered little to no value and were mostly all unique
  - Customer Key had a few duplicates but functioned essentially as unique so was removed
- Inspected data for **Missing/Null** values and any **Duplicate** rows
  - Data found to be intact and non-duplicated
- Outlier values and any possible Anomalies inspected and analyzed
  - Due to the nature of Clustering within Unsupervised Learning, **all outliers left untouched**
  - **No anomalies found in data** – analysis of unique counts per column, in addition to EDA visualizations found to be normal
  - Initial **Feature Engineering** of Average Credit Limit variable attempted but abandoned so as to **not hinder (reduce) overall clustering attempts**
- Scaling performed using Standard Scaler
  - **z-Score: Mean, +- 1 Std Dev, +- 2 Std Dev**
  - **Duel analysis** provided for Regular and Scaled Data Frames

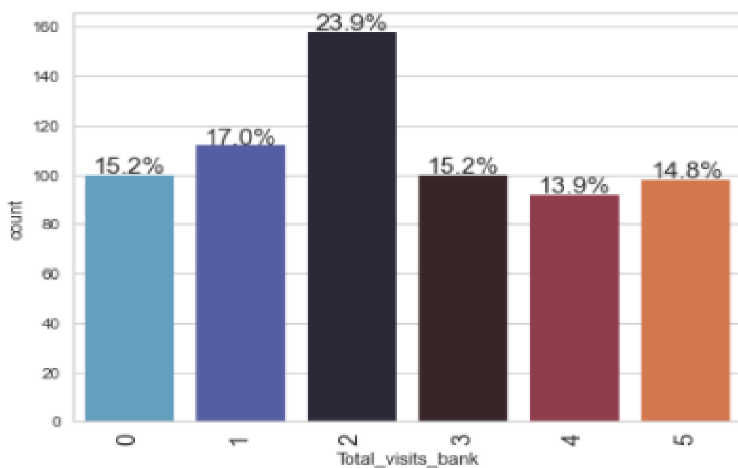
Total Credit Cards



- Relatively normally distributed (slightly left skewed - Median larger than Mean) with 50% of customers **owning at least 5 Credit Cards**
- Nearly **25%** of all customers sampled **own a total of 4 Credit Cards**, followed by customers owning 6 and 7 total Credit Cards (18% and 15% respectively)
    - Around **6%** of customers have **between 8 and 10 total Credit Cards**

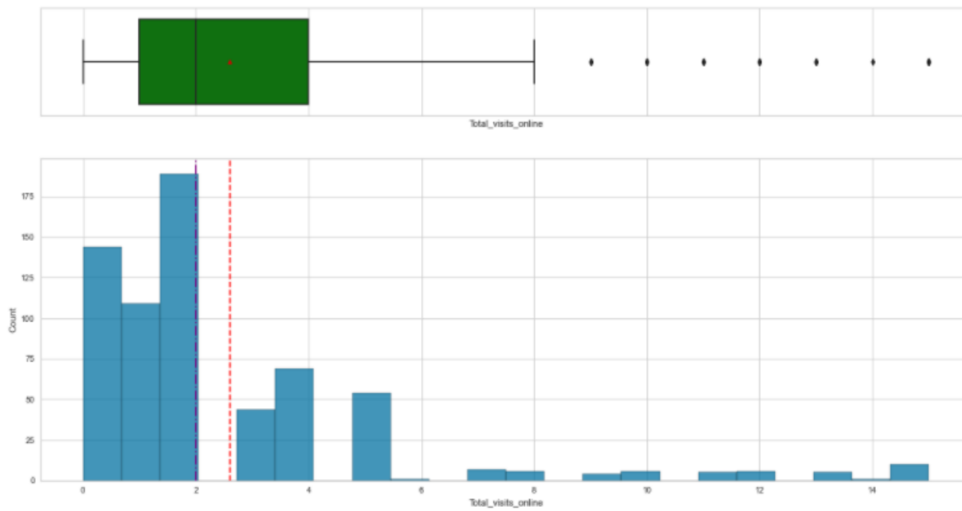
Total Visits to Bank

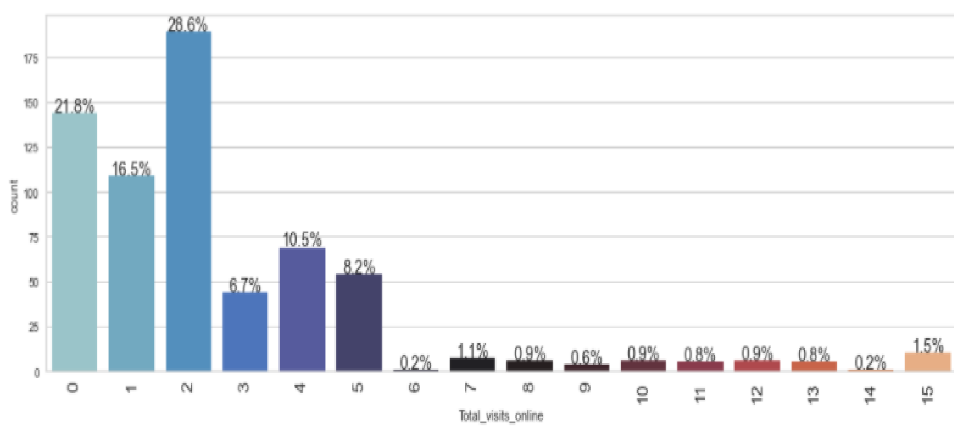




- Total Visits to Bank: Relatively normally distributed (slightly right skewed - Mean larger than Median) with 50% of customers **visiting the Bank at least 2 times a year**
- Nearly a quarter of all customers sampled visited the bank 2 times a year on average
  - **15% of customers never visited the bank in a year**
  - 44% of customers visited the bank between 3 to 5 times a year - not necessarily an indicator of any service issues, etc., but could be based on customer preference

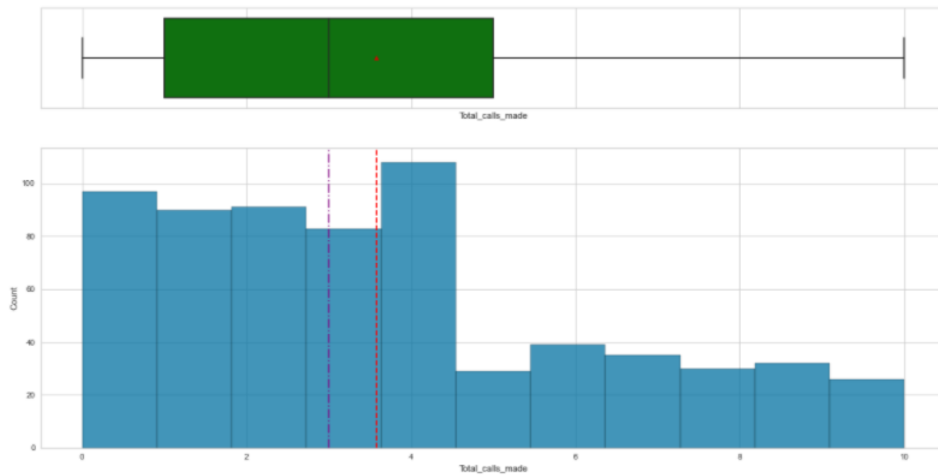
Total Visits Online

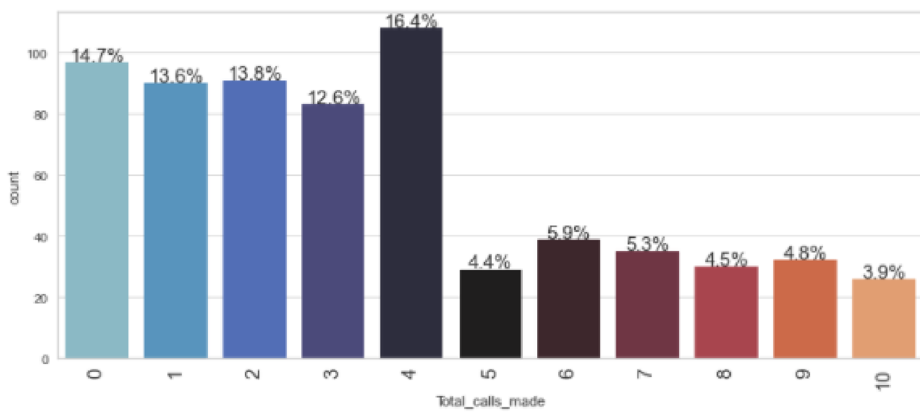




- Total Online Visits to Bank: Right skewed distribution (Mean larger than Median), with 50% of customers making at least 2 online visits to their bank accounts each year
  - There are **numerous outliers where customers visited their online accounts between 8 and 15 times a year**, however this doesn't appear to be a true 'outlier' conceptually as that is **anywhere from less than 1 to 1.25 times a month which isn't excessive**
- 67% of customers sampled visited their online bank accounts 0 to 2 times in a year which appears very low, with **22% of them never using their accounts at all in a given year**

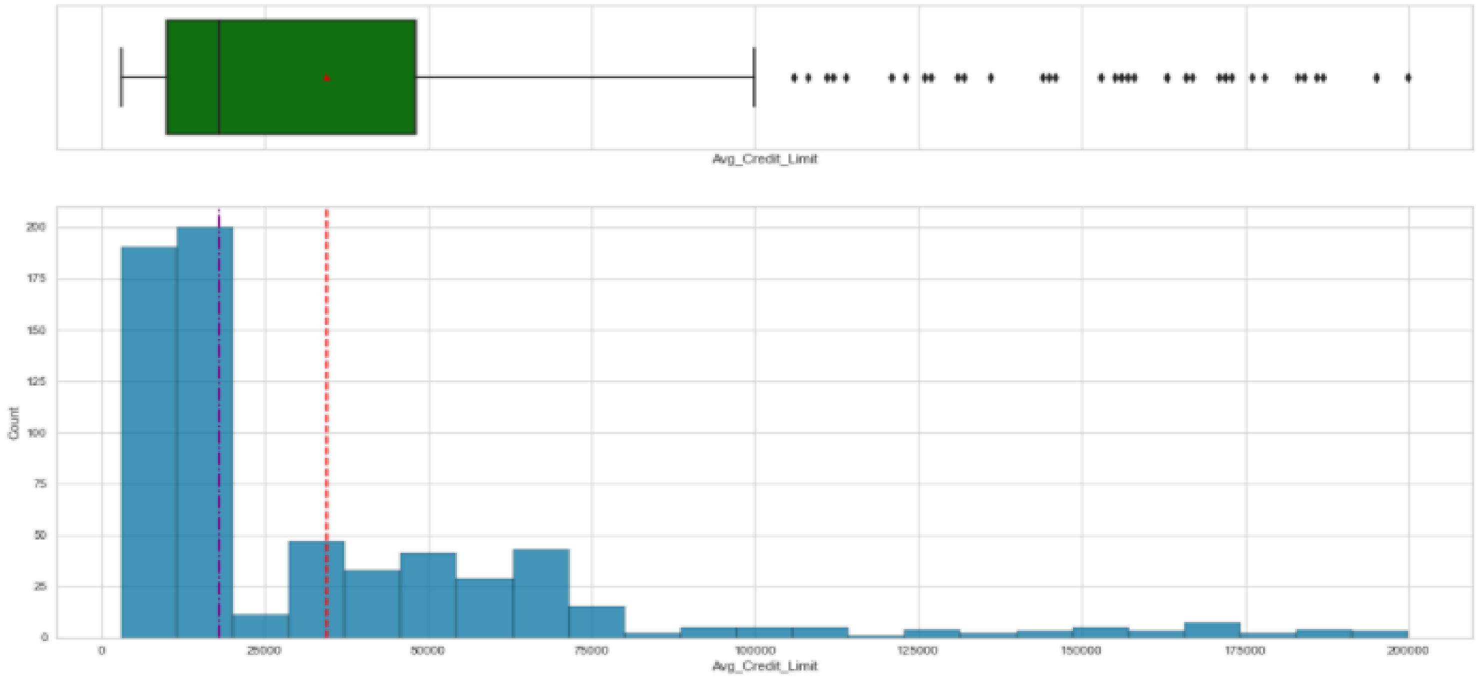
Total Calls Made





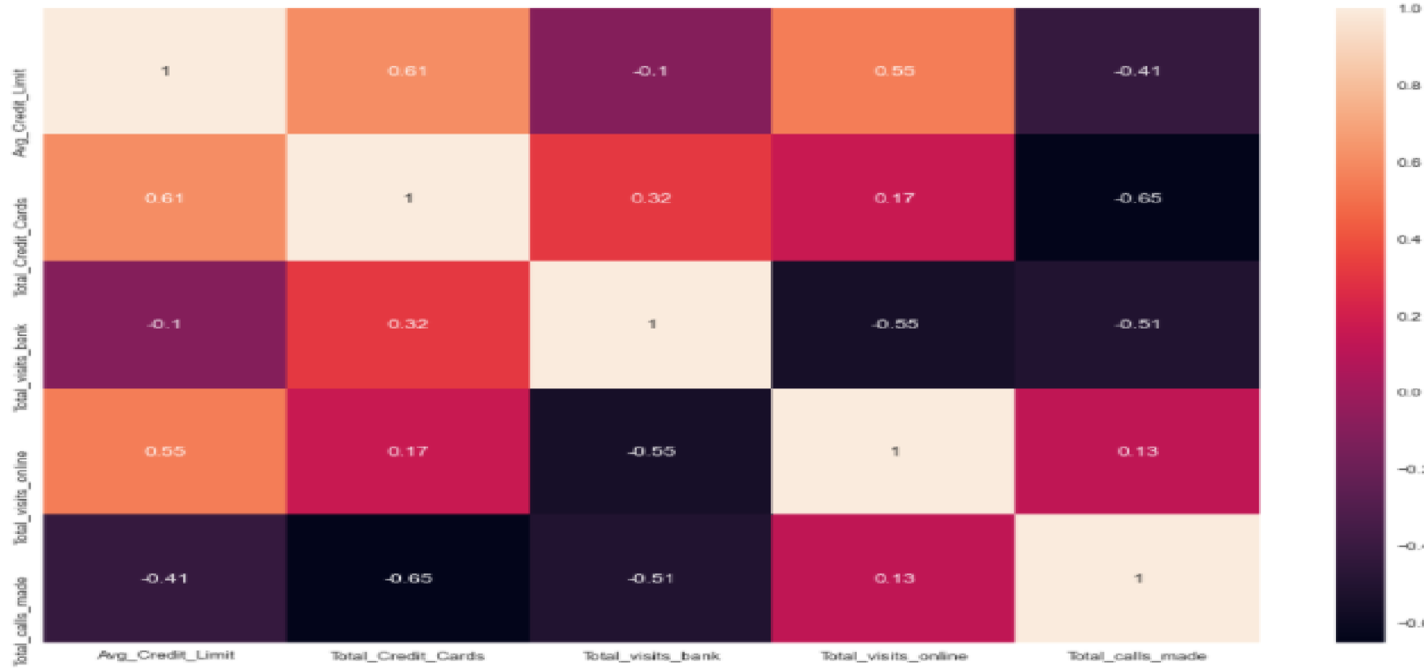
- Total Calls made to Bank: Slightly right skewed (Mean larger than Median) with 50% of customers **calling the Bank at least 3 times a year** but some customers calling between 5 and 10 times a year
  - Similar to online visits, this **doesn't appear to be excessive to call less than once a month on average**
- Over 70% of customers called the bank less than 5 times in a given year, which appears relatively low but should be further reviewed
  - **Around 15% of customers never called the bank at all**, which could **indicate satisfaction** with their service or, inversely, a **lack of interest in the bank due to dissatisfaction**, etc.

Average Credit Limit



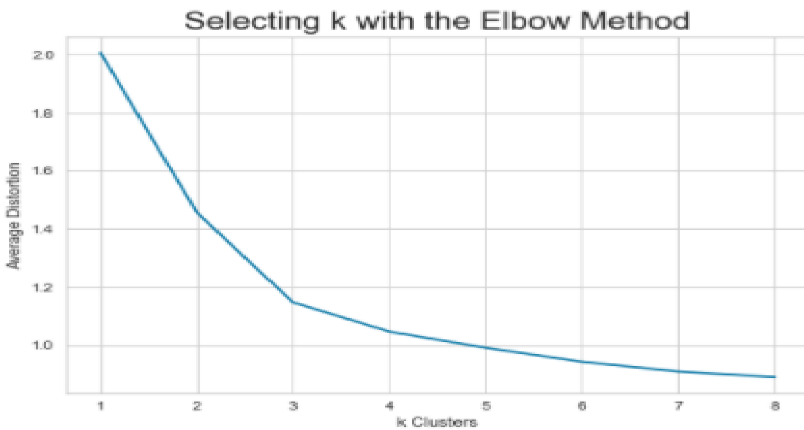
- Average Credit Limit: 50% of Customers have a Credit Limit of $18k or less, however the data is **right skewed (Mean larger than Median)** due to a larger **Mean** Limit of closer to **$35k** with numerous outlier customers with **Credit Limits well over $100k**
  - Due to a model **goal of Segmenting all Customers** into Cluster Profiles, the Average Credit Limit **outliers were left intact** and found to **contribute significantly** towards one particular Customer Cluster Profile

Correlation Summary

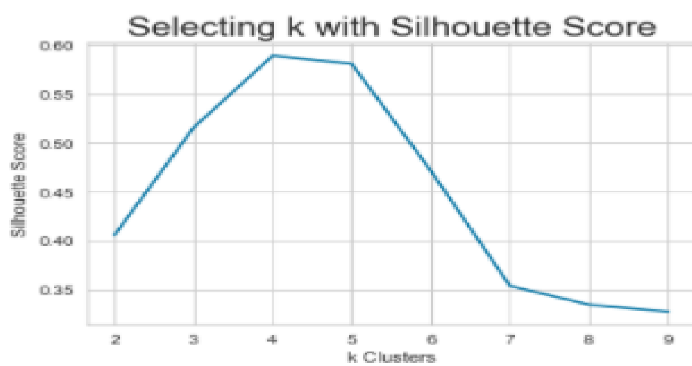


- Average Credit Limit is positively correlated with Total Credit Cards, **as additional Credit Cards increase one's overall Credit Limit available**, provided the cards aren't maxed out each time
- Total Calls Made shows a strong negative correlation to Total Credit Cards, indicating that **as customers increase their frequency of calls the bank their desire for credit with the bank stagnates or decreases**
  - This could also indicate instead that the **customers with Lower Credit Limits are those more often calling for support**, or payment forbearance services, etc.
- Conversely, Total (physical) visits to the bank is somewhat positively correlated to Total Credit Cards owned
  - This could indicate that these **in-person bank visits are improving customer loyalty and possibly converting new Credit sale**s or further limit increases for those customers engaging with the bank more frequently

K Means Clustering: Elbow Method/Silhouette Score



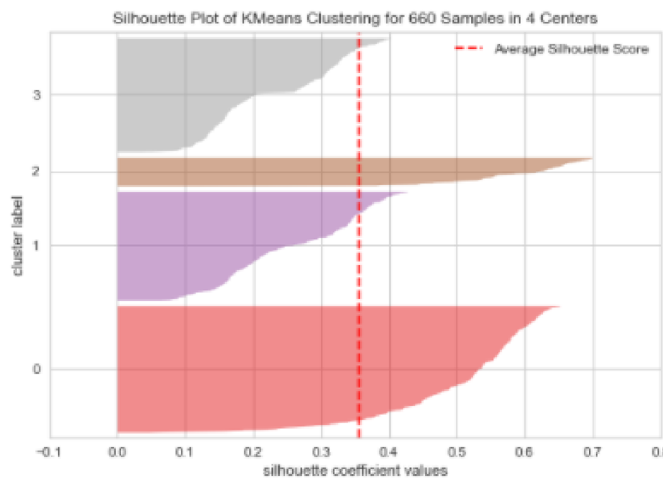

```
Number of Clusters: 1   Average Distortion: 2.00692
Number of Clusters: 2   Average Distortion: 1.45715
Number of Clusters: 3   Average Distortion: 1.14662
Number of Clusters: 4   Average Distortion: 1.04638
Number of Clusters: 5   Average Distortion: 0.99086
Number of Clusters: 6   Average Distortion: 0.94297
Number of Clusters: 7   Average Distortion: 0.90955
Number of Clusters: 8   Average Distortion: 0.89051
```

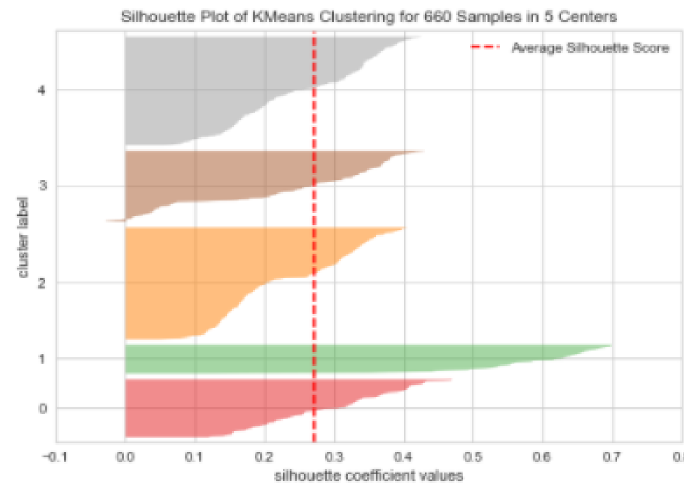


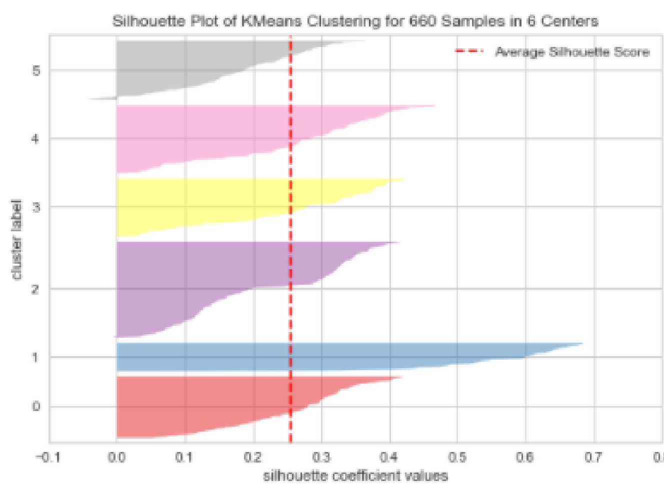

```
For n_clusters = 2, silhouette score is 0.40526
For n_clusters = 3, silhouette score is 0.51533
For n_clusters = 4, silhouette score is 0.58881
For n_clusters = 5, silhouette score is 0.58067
For n_clusters = 6, silhouette score is 0.47124
For n_clusters = 7, silhouette score is 0.35352
For n_clusters = 8, silhouette score is 0.33446
For n_clusters = 9, silhouette score is 0.32735
```

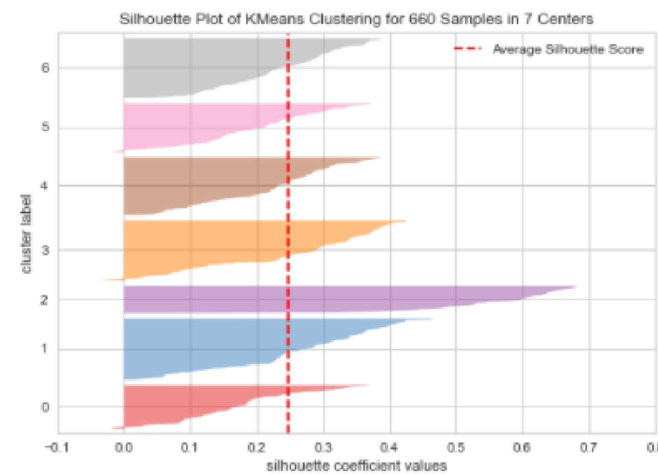
- **4 Clusters** appears to be the correct value for k from both the Elbow Curves and Silhouette Scores shown above

K Means Clustering: Visualizer (Silhouette Scores)



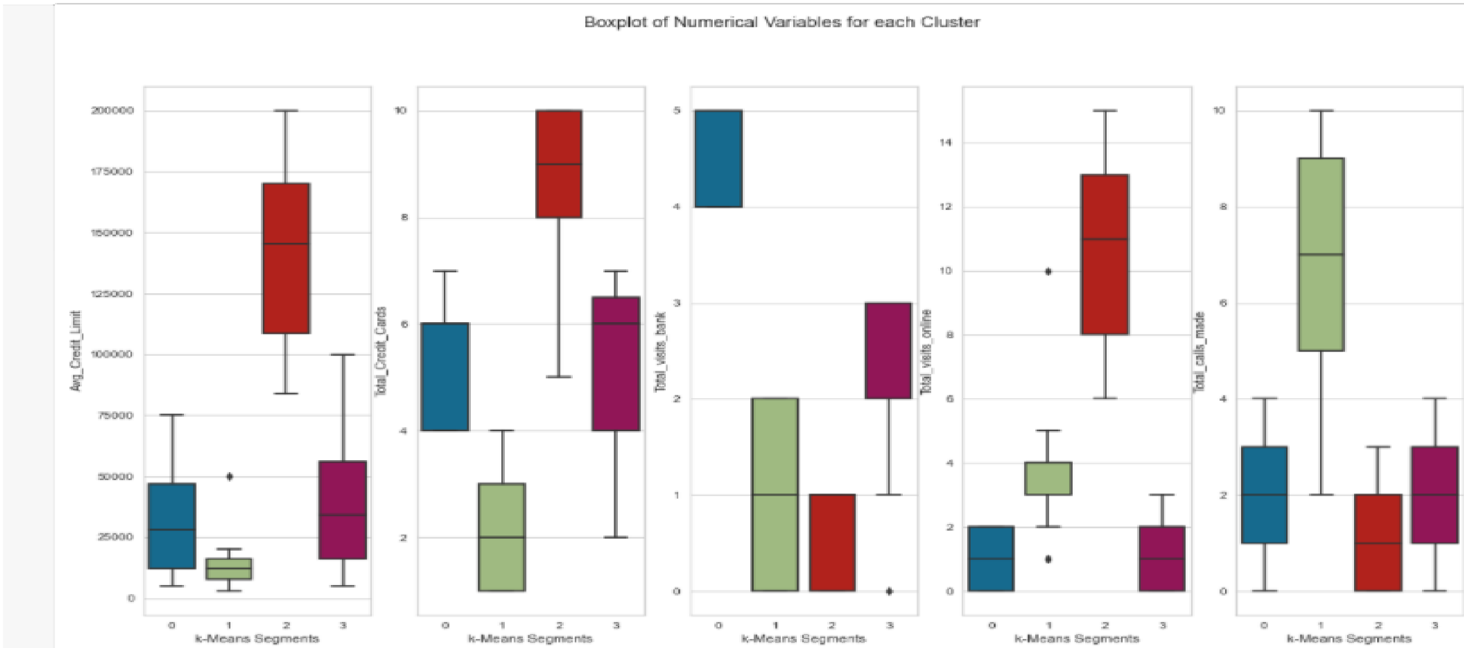






- **4 Centers/Clusters** shows the optimal/highest Silhouette Score
- As expected with a lower Cluster Split, **certain clusters are very dense** vs. larger Cluster splits with **thinner density split across a larger selection of groupings**

K Means Clustering: Cluster Profile & Insights



Recommendations

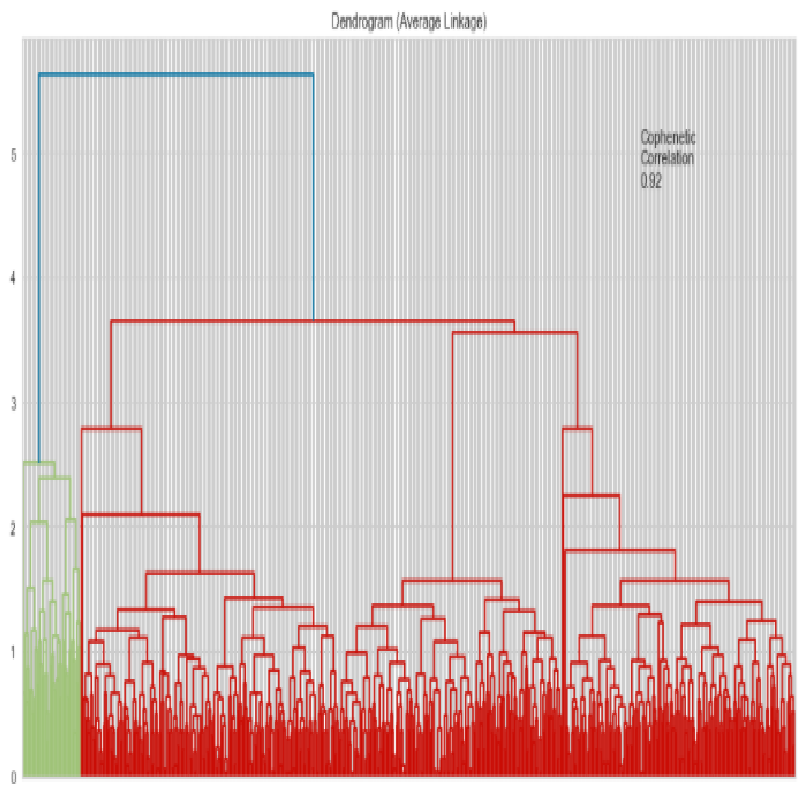
- The bank should **target customers in Cluster 2** for the **higher spending offers or specialized promotions/rewards** associated with higher spending campaigns, as this subset has the highest chance of converting and spending larger amounts between their **many Credit Cards and high Credit Limits**
- A secondary, fall back option, is to target the **next highest spending group, Cluster 3**, who are more conservative than Cluster 2 in overall spending and credit availability, but are **reliable** and somewhat consistent to forecast when it comes to credit utilization
- **Cluster groups 0 and 1 are most likely too Conservative/Traditional** in spending patterns and lifestyle to be very profitable, however should still be catered to according to their customer profile (better in person service and phone communications)

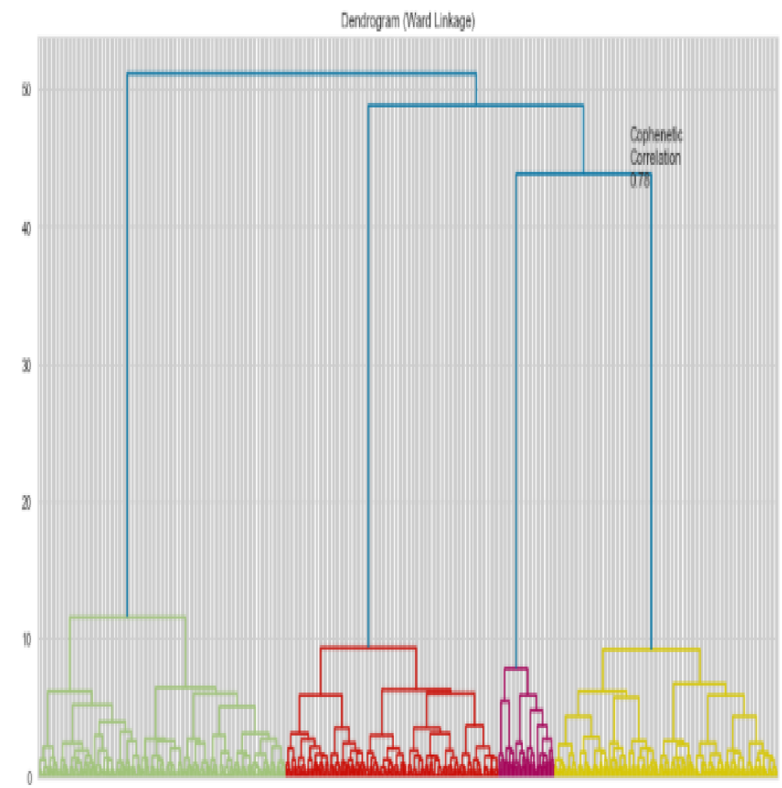
Hierarchical Clustering: Euclidean Dist./Ward Linkage

```
Cophenetic Correlation for Euclidean distance and Single linkage is 0.8003
Cophenetic Correlation for Euclidean distance and Complete linkage is 0.9129
Cophenetic Correlation for Euclidean distance and Average linkage is 0.9224
Cophenetic Correlation for Euclidean distance and Weighted linkage is 0.9003
Cophenetic Correlation for Chebyshev distance and Single linkage is 0.6978
Cophenetic Correlation for Chebyshev distance and Complete linkage is 0.8794
Cophenetic Correlation for Chebyshev distance and Average linkage is 0.9082
Cophenetic Correlation for Chebyshev distance and Weighted linkage is 0.9134
Cophenetic Correlation for Mahalanobis distance and Single linkage is 0.8277
Cophenetic Correlation for Mahalanobis distance and Complete linkage is 0.6988
Cophenetic Correlation for Mahalanobis distance and Average linkage is 0.8805
Cophenetic Correlation for Mahalanobis distance and Weighted linkage is 0.8219
Cophenetic Correlation for Cityblock distance and Single linkage is 0.9053
Cophenetic Correlation for Cityblock distance and Complete linkage is 0.9068
Cophenetic Correlation for Cityblock distance and Average linkage is 0.9134
Cophenetic Correlation for Cityblock distance and Weighted linkage is 0.897
```

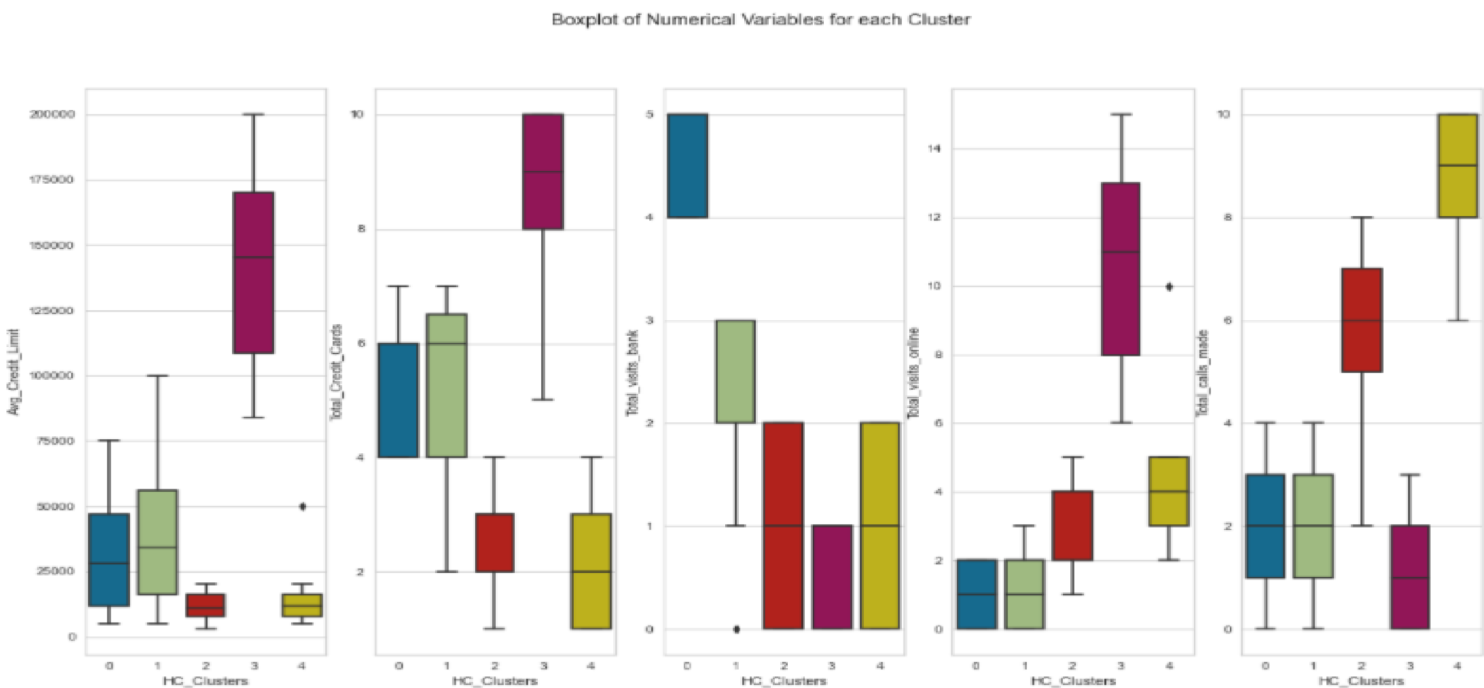
```
Cophenetic Correlation for Single linkage is 0.8003
Cophenetic Correlation for Complete linkage is 0.9129
Cophenetic Correlation for Average linkage is 0.9224
Cophenetic Correlation for Weighted linkage is 0.9003
Cophenetic Correlation for Centroid linkage is 0.9167
Cophenetic Correlation for Ward linkage is 0.7826
```

The **Ward Linkage Method** was selected over the Average Linkage Method (even though its Cophenetic Score was lower) as it offers a **much cleaner Cluster view for analysis**





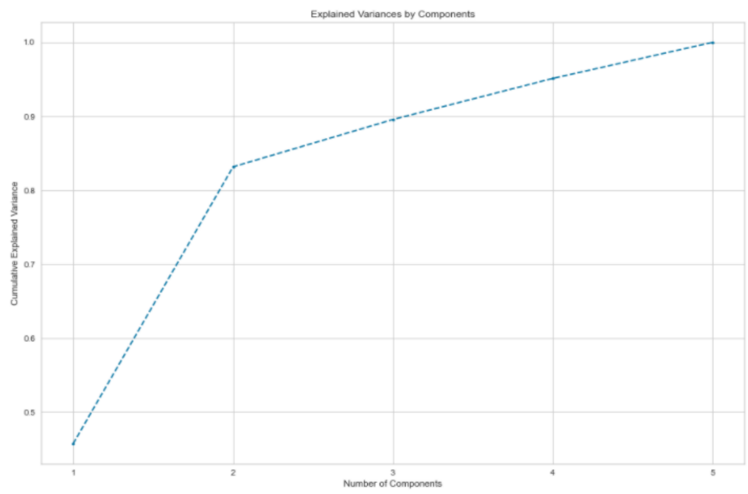
Hierarchical Clustering: Cluster Profile & Insights



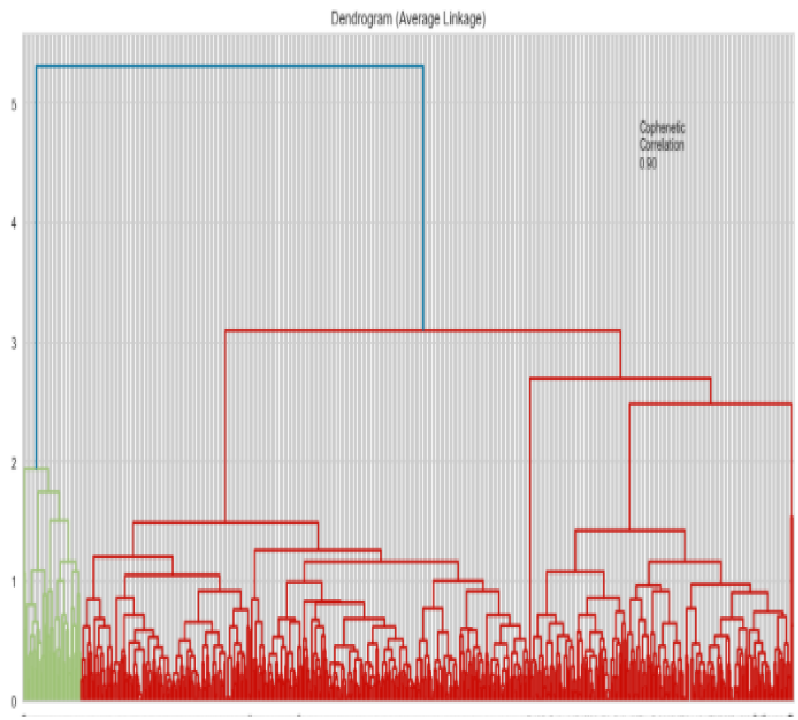
Recommendations

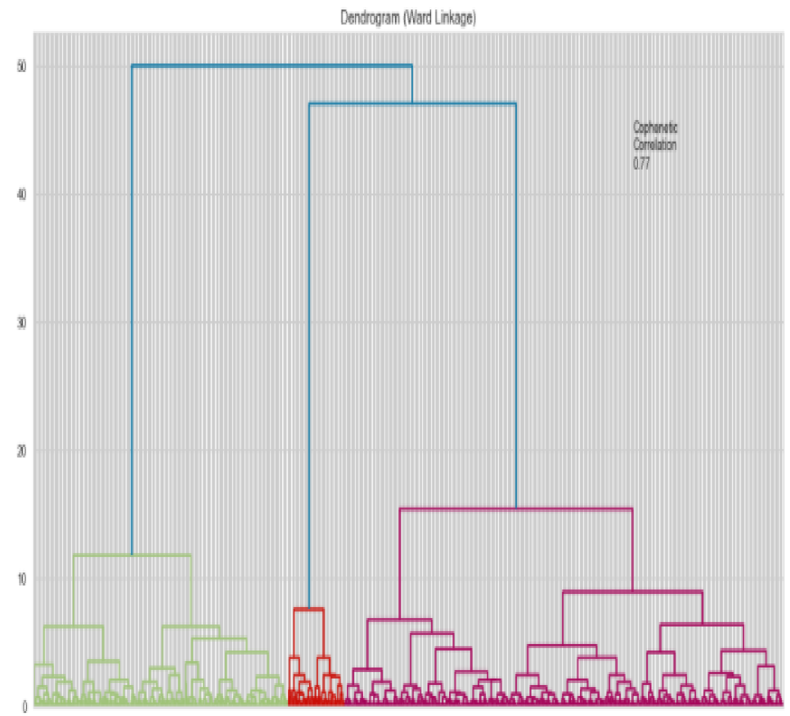
- The bank should **target customers in Cluster 3** for the **higher spending offers or specialized promotions/rewards** associated with higher spending campaigns, as this subset has the highest chance of converting and spending larger amounts between their **many Credit Cards and high Credit Limits**
- A secondary, fall back option, is to target the **next highest spending group, Cluster 1**, who are more conservative than Cluster 3 in overall spending and credit availability, but are **reliable** and somewhat consistent to forecast when it comes to credit utilization
- **Cluster groups 2 and 4, and to some extent 0,** are most likely to be low spending for either being too **Conservative/Traditional in spending patterns** or with little available Credit
  - These customers types may not be very profitable to the bank, however should still be catered to according to their customer profile (cheaper offers with lower barriers to entry, etc.)

Hierarchical Clustering: PCA (90%)

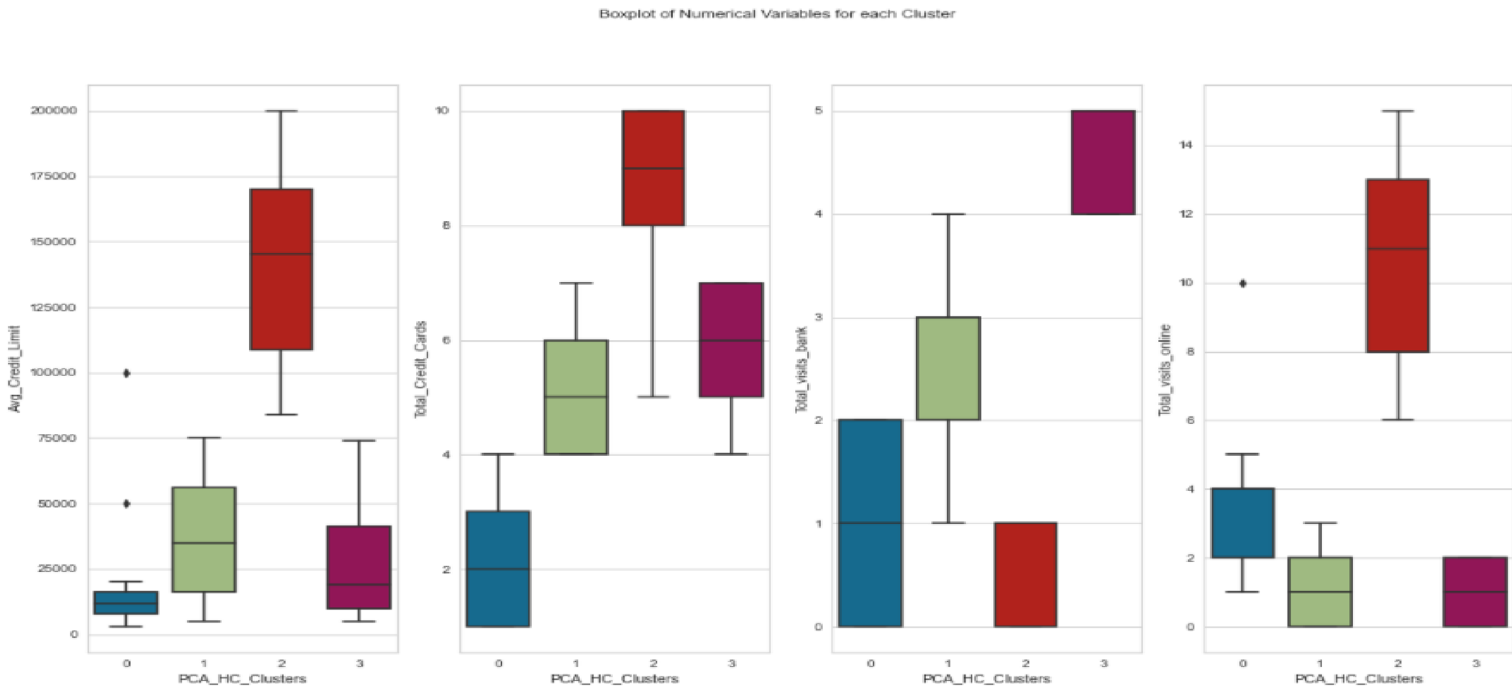


- **90%** of Variance Explained with just **3 (of 5)** components used
- Similar Dendrogram results for both the Average Linkage (highest Cophenetic Correlation) and Ward Linkage (Cleanest Cluster View) models
- Ward Linkage Model (Euclidean Distance) still selected as final model for hierarchical clustering





Hierarchical Clustering (PCA): Cluster Profile/Insights



Recommendations¶

- The bank should **target customers in Cluster 2** for the **higher spending offers or specialized promotions/rewards** associated with higher spending campaigns, as this subset has the highest chance of converting and spending larger amounts between their **many Credit Cards and high Credit Limits**
- A secondary, fall back option, is to target the **next highest spending group, Cluster 1** (though far behind), who are more conservative than Cluster 2 in overall spending and credit availability, but are **reliable** and somewhat consistent to forecast when it comes to credit utilization
- **Cluster groups 0 and 3, are most likely to be low spending for either being too Conservative/Traditional** in spending patterns or with little available Credit
  - These customers types may not be very profitable to the bank, however should still be catered to according to their customer profile (cheaper offers with lower barriers to entry, etc.)

Cluster Summary & Selection

- Of the **three Cluster Profiles created (k-Means, Hierarchical/Dendrogram, and PCA (Reduced Dimensionality) Dendrogram**, the following insights and recommendations are provided:
- Similarities existed between the 3 cluster profiles, with a **clear 'Heavy Spending' customer and various 'More Conservative/Traditional customers identified**, though the cluster ordering shifted/grew between models
  - The **Ward Linkage** model was found to score slightly **lower Cophenetic Correlations scores but showed a clearer cluster segmentation within the Dendrogram** and was therefore chosen for final Hierarchical models (Regular and PCA reduced)
  - The **Euclidean Distance Metric** was found to score the highest results and used in all Hierarchical models
- Both the **k-Means and PCA (Ward Linkage)** models were based on **4 clusters** and offered substantial segmentation in regards to Customer Profiling
  - Due to the incorporation of Primary Component Analysis (PCA) it was determined that 90% Variance Explanation could be achieve with **only 3 of the 5 components included (cumulatively)**:
    - **Average Credit Limit:** 45.7%
    - **Total Credit Cards:** 83.2%
    - **Total Visits to Bank**90%
  - Incorporating PCA with the Ward Linkage Method created 4 very clean Customer Profiles, with a clear Heavy Spending (target) customer identified
- A Hierarchical Cluster model and Dendrogram, was attempted using **5 clusters** but was found to **offer slightly redundant segmentation details and little additional value vs. the 4 cluster Profiles**
- Finally, due to the relatively similar value provided and the benefit of reduced Dimensionality (3/5 components selected), the **Ward Linkage PCA-reduced 4 Cluster Hierarchical model** was selected as the final option on which to base Customer Profiles