# AllLife Bank Customer Segmentation

## Elective Project

## MIT x Great Learning

November 25th, 2024

# Contents / Agenda

- Executive Summary
- Business Problem Overview and Solution Approach
- Data Overview
- EDA and Data Preprocessing
- Model Building
- Appendix

# Executive Summary

- Customers can be organized into 3 groups.
- AllLife can use Average credit limit to identify groups
- AllLife Marketing Team can use these three groups to target their ads to the correct market and to improve the support services for existing customers
- AllLife Marketing Team can also advertise less used channels of contact to existing customers to help improve the notion of support services.

# Business Problem Overview and Solution Approach

- AllLife customers perceive the support services poorly. Therefore, the Marketing Team wants to upgrade the service delivery model to ensure customers' queries are resolved faster.
- The Data Science team aims to identify different segments in the existing customer base in order to help solve this problem
- Identifying different segments as well as taking into account spending patterns and past interactions will help AllLife create faster solutions for customers within their service delivery model based on customer segment data.
- This segmentation will also help the marketing team to run personalized campaigns to target new customers.

# Data Overview

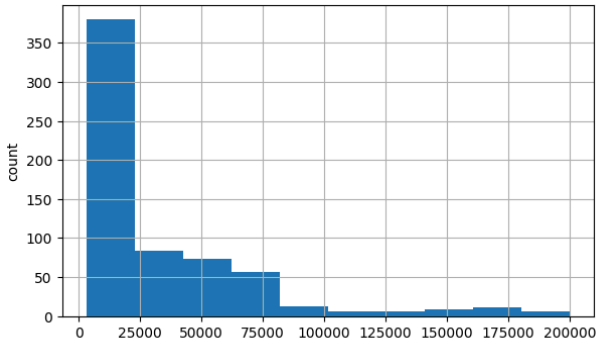
- AllLife has data including customer credit limit, the total number of credit cards the customer has, and different channels through which the customer has contacted the bank for any queries - including visiting the bank, online, and through a call center
- Key variables/features are:
    - Sl_no - Customer Serial Number
    - Customer Key - Customer Identification
    - Avg_Credit_Limit - Average credit limit
    - Total_Credit_Cards - Total number of credit cards
    - Total_visits_bank - total bank visits
    - Total_visits_online - total online visits
    - Total_calls_made - Total calls made

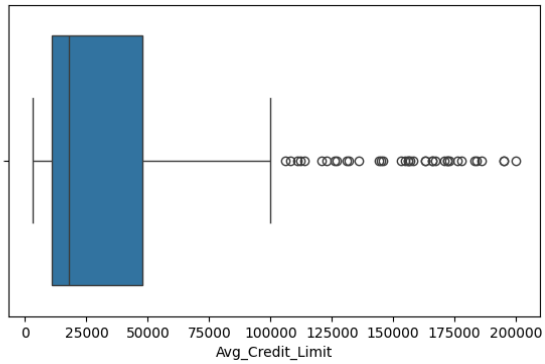
## EDA and Data Preprocessing

- There are 660 observations and 7 columns in the dataset
- There are no missing values
- All columns are the data type integer
- Customer Key had 5 duplicate observations, which were dropped before applying any algorithm. After this Customer Key and Serial Number could be dropped entirely as they aren't needed for the solution
- After dropping all the duplicated rows, the final data shape is 644 observations and 5 columns
- After summarizing the statistics,
    - Average Credit Limit has a huge standard deviation

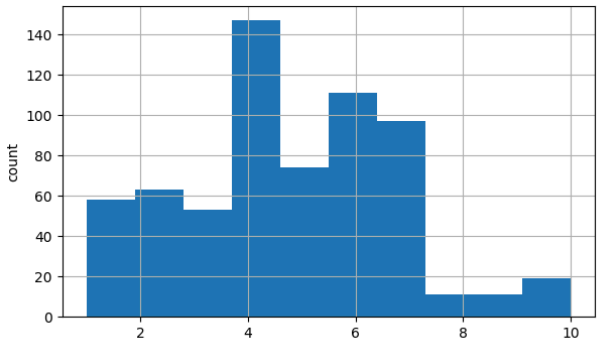
- Most customers have an average of 4.6 credit cards, visit the bank just over 2 times in person and just over 2.5 times online, while averaging around 3.5 calls.
- Some customers never visit or call at all.
- Customers tend to visit online most often over in person visits and calling.

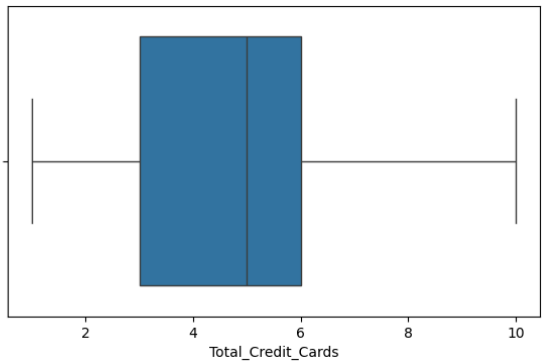
Average credit limit is heavily right skewed on the histogram. This is confirmed with a lot of outliers on the box plot. Most credit limits are somewhere between around $6,000 - $48,000



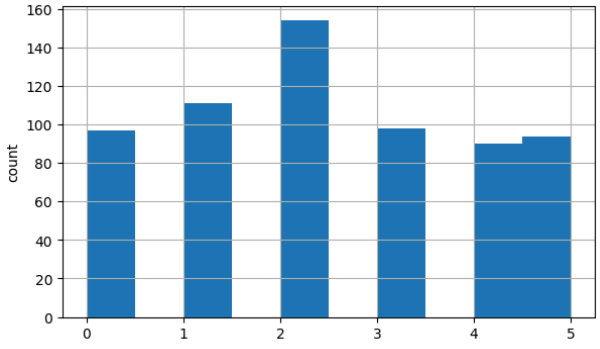



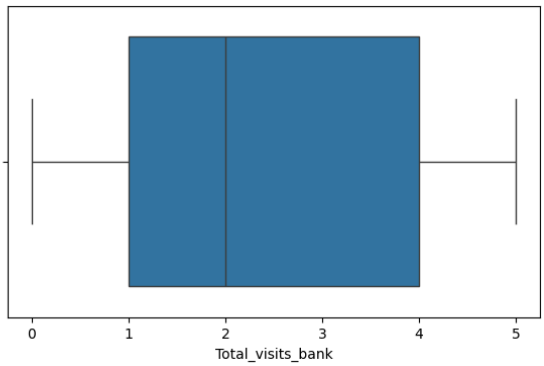

For total credit cards, most customers have 4, with the median being around 5. However, the amount ranges from 1 -10



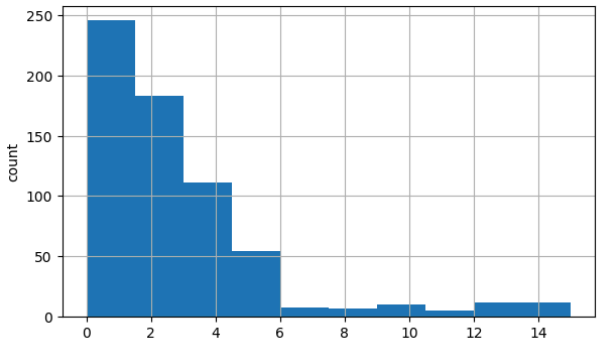



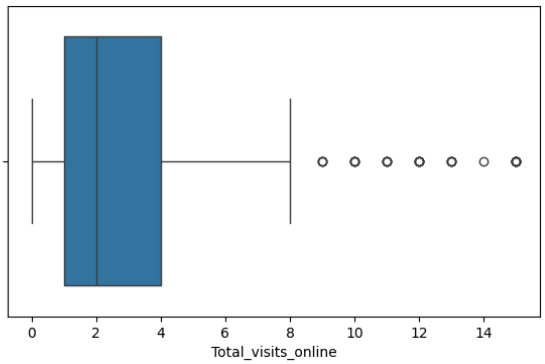

**For total bank visits, it looks like most people visited 2 times. This ranges from 1-5.**



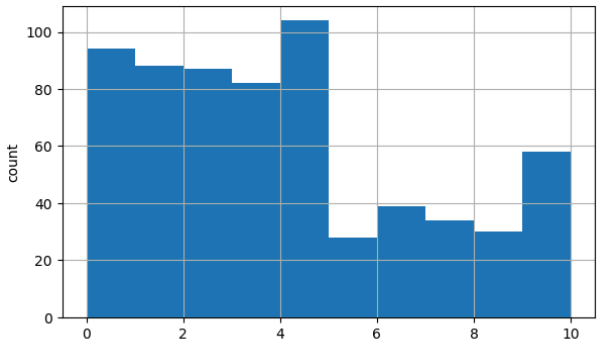



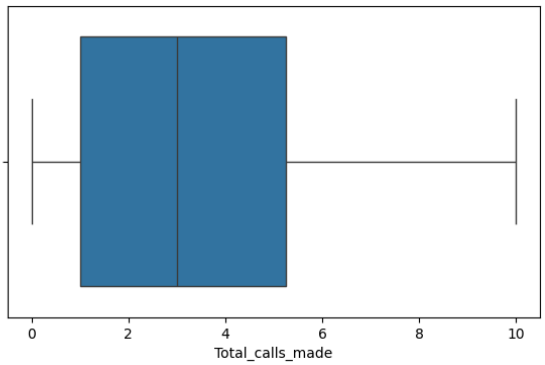

**For total bank visits online, the data is heavily right-skewed as seen in the histogram and confirmed with the outliers in the boxplot. Most people visit the online bank 0-1 times, with 2 being the median amount of visits.**




**Lastly, for total calls made, the median is around 3, but from the histogram it looks like this can widely vary, with a dip around 5 calls but a slight increase around 9/10.**

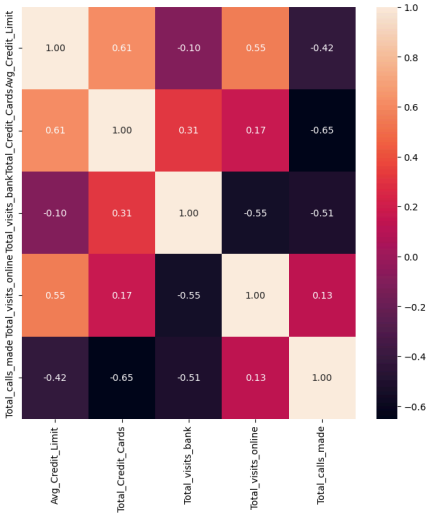



**After creating a heatmap to show correlation, it was revealed that**

- **Average credit limit is positively correlated with Total Credit cards and Total visits online which makes sense.**

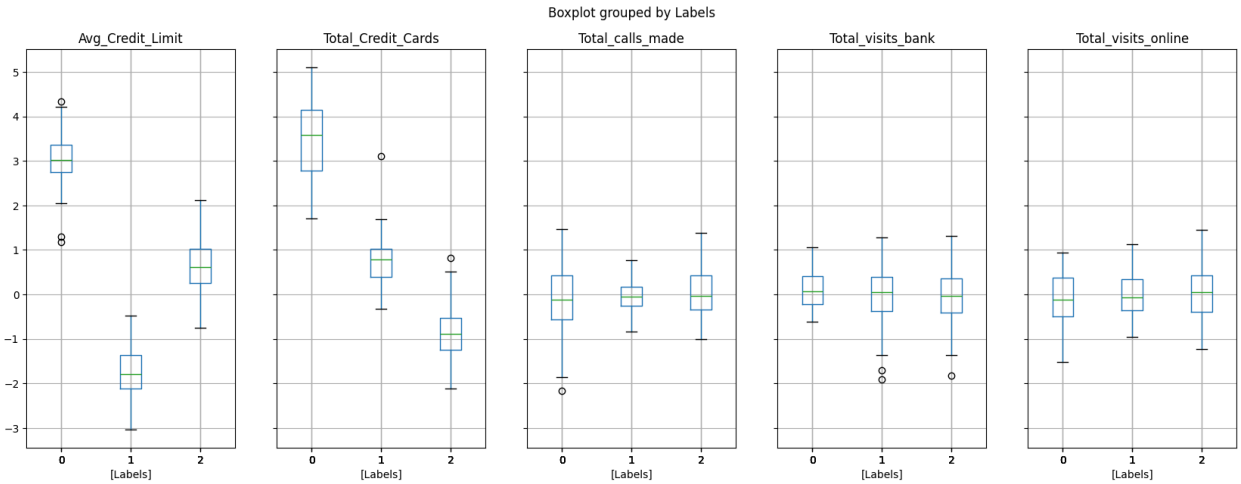
- Average credit limit is negatively correlated with total calls made and total visits to the bank
- Total bank visits, online visits, and calls made are negatively correlated which implies that the majority of customers only use one channel to contact the bank.



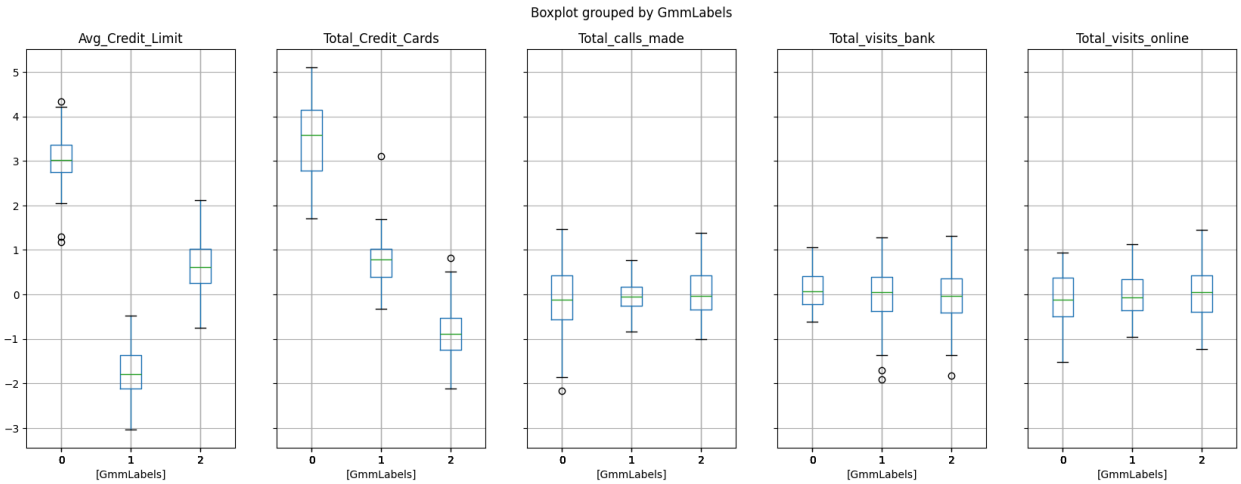

# Model Building

- After scaling the data and then applying PCA to reduce dimensionality I used the clustering algorithms K-Means, Gaussian Mixture Model, and K-Medoids to get a comprehensive understanding of the clusters in the data and what algorithm provides the most sound analysis and the optimum number of clusters to use.
- K-Means key insights
    - 3 clusters was the optimum amount with 49 observations in cluster 1, 221 observations in cluster 2, and 374 observations in cluster 3

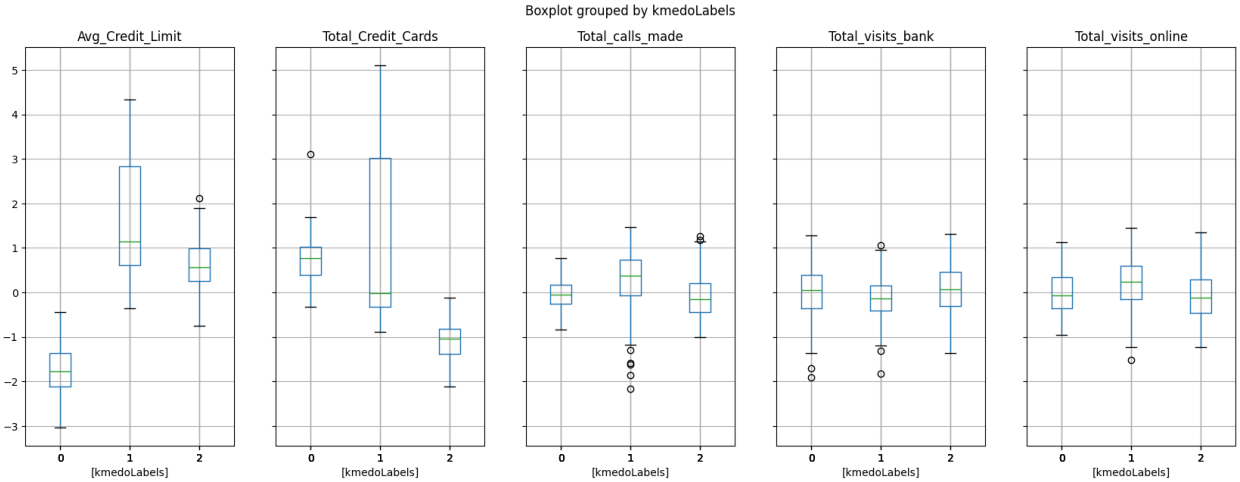
- It's clear that group 0 has the highest credit risk (highest average credit limit & total # of cards), group 1 has the lowest credit risk (lowest average credit limit), and group 2 is in the middle
- This is confirmed by the boxplot visual below
- The total calls, in person, and online visits are similar for each group




- Gaussian Mixture Model key insights
  - After applying this model with 3 components and random state 1, I received the same exact number of observations for each cluster
  - Once again it's clear that group 0 has the highest credit risk, group 1 has the lowest credit risk, and group 2 is in the middle
  - The box plot looks exactly the same as K-Means
  - The total calls, in person, and online visits are similar for each group




- K-Medoids key insights
    - After applying this model with 3 components and random state one, I received a different amount of observations for each cluster with 222 observations in Group 0, 133 observations in Group 1, and 289 observations in Group 2
    - With this model it looks like Group 0 has the lowest credit risk (lowest average credit limit), Group 1 has the highest credit risk (highest average credit limit and most # of cards), and Group 2 is somewhere in the middle.
    - The total calls, in person, and online visits are similar for each group




- When comparing these clustering algorithms it's interesting to see how K-Medoids completely flipped the results for Group 0 and Group 2.
- K-Medoids also had very different counts from K-Means and GMM.
- I believe this is due to the large amount of outliers captured in the Histogram and Box Plots for Average Credit Limit. This completely reshaped the data
- With K-Medoids being the least sensitive to outliers, it's safe to say this is the most accurate algorithm.

# Conclusion

- After analyzing this data and running the clustering algorithms, I would suggest to the marketing team to target 3 distinct segments of existing customers to improve the support services and to advertise to find new customers.
- This analysis allowed me to see that 3 clusters is the optimum amount of clusters and that you can organize customers by their credit limit as it's positively correlated to the total amount of credit cards
- The amount of in person visits, to online visits, to phone calls, while negatively correlated to each other, seemed to have little impact on the actual segments, although K-Medoids showed more outliers.
- However, this data can be used to potentially advertise other channels to customers that they might not pursue otherwise, which could then potentially increase their notion of the bank's support services.

# Appendix

Visuals and analysis included above.