**REPORT**

**Customer Segmentation**

**Introduction/Overview**

The Cow and Buffalo Milk company is a distributor of dairy products that serves the entire country. As a distributor, they provide dairy products to a network of merchants. To optimize their business operations, increase sales, and allocate advertising spend more efficiently, Cow and Buffalo Milk company aims to develop models that will help them target the right customers.

This project focuses on utilizing a merchant dataset to create a credit scoring algorithm. The credit scoring algorithm will enable Cow and Buffalo Milk company to assess the creditworthiness of their merchants. By analyzing the payment history and the cities in which the merchants operate, the algorithm will help the distributor identify which merchants are most likely to repay their credit obligations reliably, as well as identify those who may present a higher risk.

The dataset contains several relevant features that will aid in developing the credit scoring algorithm. These features include:

Merchant Id: A unique identification number assigned to each merchant, allowing for individual merchant tracking.

Annual Revenue: The annual income of the merchant, which provides insights into their financial stability and potential repayment capacity.

Spending Score: A score, ranging from 1 to 100, assigned by Ramani.io based on the merchant's spending behavior and their purchases of distributor products. This score helps assess the merchant's loyalty and engagement with Cow and Buffalo Milk company.

City: The city in which the merchant is located. Understanding the geographic distribution of merchants can provide valuable insights into regional trends, customer behavior, and market dynamics.

Most Purchased Product: This feature represents the most popular product purchased by a specific merchant from Cow and Buffalo Milk company, measured by the amount of money spent. It helps identify the merchant's preferences and highlights potential product-specific strategies.

Payment Score: A score, ranging from 1 to 5, given by a consultant based on the merchant's ability to repay inventory purchased on credit. A higher score indicates a greater likelihood of timely repayment, while a lower score indicates a higher credit risk.

**Exploratory Data Analysis (EDA) :**

**1. Introduction**  :

The dataset contains information about merchants, including their annual revenue, spending score, city, and most purchased product. The goal of the EDA is to gain insights and understand the characteristics of the merchants, which will aid in developing models to target the right customers and improve resource allocation.

**2. Dataset Overview** The dataset consists of 200 entries, each representing a merchant. It contains the following columns:

* Merchant Id: A unique identification number assigned to each merchant.
* Annual Revenue: The annual income of the merchant in thousands of dollars.
* Spending Score: A score ranging from 1 to 100, indicating the spending behavior of the merchant.
* City: The city where the merchant is located.
* Most Purchased Product: The most popular product purchased by the merchant from Cow and Buffalo Milk company.

The dataset does not contain any missing values, ensuring the reliability of the analysis.

**3. Descriptive Statistics** By examining the descriptive statistics, we gain a better understanding of the distribution and central tendencies of the numerical features:

* Merchant Id: The merchant IDs range from 1 to 200.
* Annual Revenue: The average annual revenue is approximately $60,560, with a standard deviation of $26,264. The minimum and maximum revenues are $15,000 and $137,000, respectively.
* Spending Score: The average spending score is 50.2, with a standard deviation of 25.82. The scores range from 1 to 99.

These statistics provide a snapshot of the dataset and serve as a foundation for further analysis.

**4. City Distribution** The city distribution reveals the number of merchants in each city:

* Dar es Salaam: 82 merchants
* Moshi: 76 merchants
* Mtwara: 21 merchants
* Zanzibar City: 21 merchants

This information highlights the geographical distribution of the merchants and suggests potential variations in customer behavior across different cities.

**5. Annual Revenue Distribution** Analyzing the distribution of annual revenue, we observe that some revenue values are more common than others. Notably:

* The revenue of $54,000 and $78,000 are the most common, each occurring 12 times.
* There are 64 unique revenue values, indicating a diverse range of merchant incomes.

Understanding the distribution of annual revenue helps in segmenting merchants based on their financial capacities and potential creditworthiness.

**Top of Form**

**Feature Selection and Scaling:**

Bottom of Form

Feature selection plays a crucial role in the customer segmentation project. It involves selecting the most relevant features that contribute to the clustering objective. In this project, we focused on two key features: 'Annual\_Revenue' and 'Spending\_Score'.

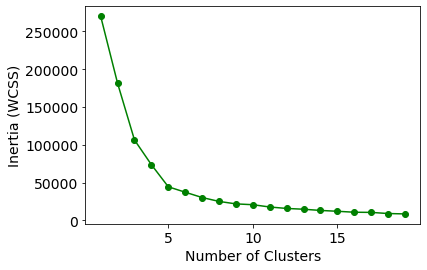
These features were chosen because they provide valuable insights into a merchant's income and spending behavior, which are critical factors for customer segmentation.

By selecting these specific features, we aimed to capture the financial aspects of a merchant's profile, which are likely to contribute significantly to their purchasing patterns and creditworthiness.

Once the relevant features were identified, further preprocessing steps, such as handling missing values or scaling, could be applied to prepare the data for clustering analysis."

**Determining Optimal Number of Clusters :**

The elbow method is used to determine the optimal number of clusters. It involves fitting K-means models with different numbers of clusters and calculating the sum of squared errors (SSE). The SSE values are then plotted against the number of clusters. The optimal number of clusters is typically identified at the "elbow" point on the plot, where increasing the number of clusters provides diminishing returns in terms of reducing SSE. This point represents a trade-off between capturing data variability and model complexity



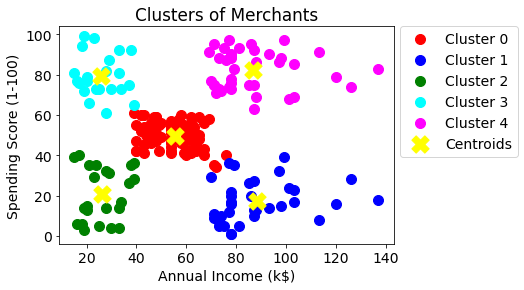
Here the optimal number of clusters are 5

**K-means Clustering:**

The K-means clustering algorithm is a popular unsupervised learning method used to group similar data points into clusters. It is applied to the scaled feature data in order to identify patterns and segment the data based on similarities.

In the implementation, the algorithm starts by randomly initializing cluster centroids. The initial positions of these centroids can affect the final clustering result, and different initialization methods can be used. However, the specific initialization method is not mentioned in the provided code snippet.

The algorithm then iteratively assigns each data point to the nearest centroid based on the Euclidean distance. After all data points have been assigned, the centroids are updated by computing the mean values of the data points within each cluster. This assignment and update process continues until convergence.

The number of iterations required for convergence can vary depending on the dataset and the specified convergence criteria. In the code snippet, the specific convergence criteria are not mentioned, so it's assumed that the algorithm continues iterating until convergence is achieved.

By analyzing the scatter plot, we can observe the grouping of data points into distinct clusters based on their 'Annual\_Revenue' and 'Spending\_Score' features. This visualization helps to understand the segmentation and identify the characteristics of each cluster.

Clusters visualization:

The scatter plot displays the clusters based on 'Annual\_Revenue' and 'Spending\_Score'.

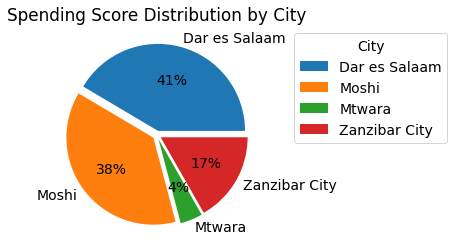
* Cluster 0 (red) represents merchants with an average income and spending score.
* Cluster 1 (blue) represents merchants with high income and low spending score.
* Cluster 2 (green) represents merchants with low income and low spending score.
* Cluster 3 (cyan) represents merchants with low income and high spending score.
* Cluster 4 (magenta) represents merchants with high income and high spending score.

**Cluster Analysis and Interpretation:**

1. General:
   * Characteristics: This segment comprises the largest group of customers, with 81 individuals. They exhibit average spending habits without any specific tendencies towards saving or splurging.
   * Business Implications: Since this segment represents a significant portion of the customer base, it is essential to maintain their satisfaction and engagement. Implementing a customer retention strategy, such as personalized offers or loyalty programs, can help maintain their loyalty and encourage repeat purchases.
2. Target:
   * Characteristics: This segment consists of 39 individuals who are likely to be the target audience for specific products or services. They show a higher level of engagement and responsiveness to targeted marketing campaigns.
   * Business Implications: For this segment, focus on developing personalized marketing campaigns based on their preferences and behavior. Utilize customer data analysis and predictive modeling to identify their needs and preferences accurately. Implement targeted advertising, personalized recommendations, and tailored promotions to increase their conversion rates.
3. Miser:
   * Characteristics: The Miser segment includes 35 individuals who tend to be frugal and cautious with their spending. They prioritize saving money and look for discounts or cost-effective options.
   * Business Implications: Tailor marketing strategies to highlight cost savings, discounts, and value-for-money propositions. Offer special promotions, bundle deals, or loyalty programs that provide incentives for this segment to make purchases. Emphasize the long-term savings and benefits of your products or services.
4. Careful:
   * Characteristics: The Careful segment consists of 23 individuals who are cautious and thoughtful about their purchasing decisions. They thoroughly research products, read reviews, and consider quality and reliability before making a purchase.
   * Business Implications: Build trust and credibility by providing detailed product information, customer reviews, and testimonials. Implement a robust review and rating system on your website or platform. Consider partnering with influencers or industry experts to provide endorsements or recommendations. Focus on delivering exceptional customer service and emphasize the quality and reliability of your products or services.
5. Spendthrift:
   * Characteristics: The Spendthrift segment includes 22 individuals who have a tendency to spend money freely and indulge in luxury or high-end products and experiences.
   * Business Implications: Develop marketing strategies that emphasize exclusivity, superior quality, and unique experiences. Collaborate with luxury brands or influencers to enhance the perceived value of your products or services. Offer personalized shopping experiences, VIP programs, and limited-edition collections. Leverage social media platforms and aspirational marketing to attract this segment.

**Insights**

Spending score Distribution:



Above pie plot represents the spending score percentage in 4 cities.

Dar es salaam: 41% spending score

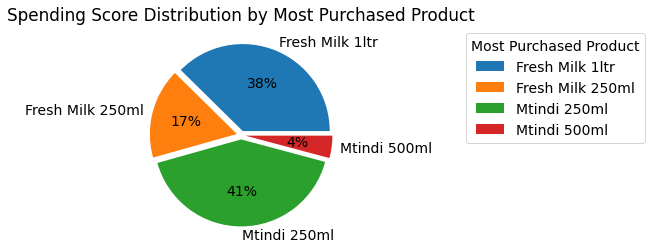
Moshi: 38% spending score

Mtwara: 4% spending score

Zanzibar city: 17% spending score

This information highlights the merchants spendings and suggests that to develop the products on the basis of this spendings.

Spending score distribution by most purchased product:



Above pie plot represents the plotting of Most purchased product in 4 cities based on merchants spendings.

Fresh milk 1ltr : 38%

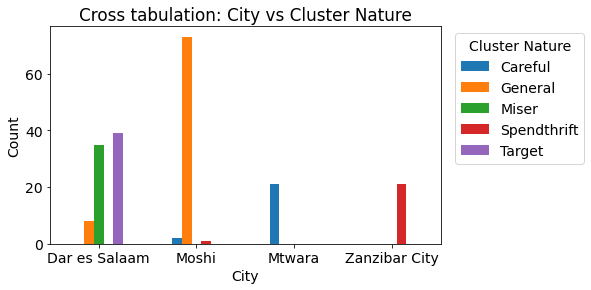
Fresh milk 250ml : 17%

Mtindi 500ml : 4%

Mtindi 250ml : 41%

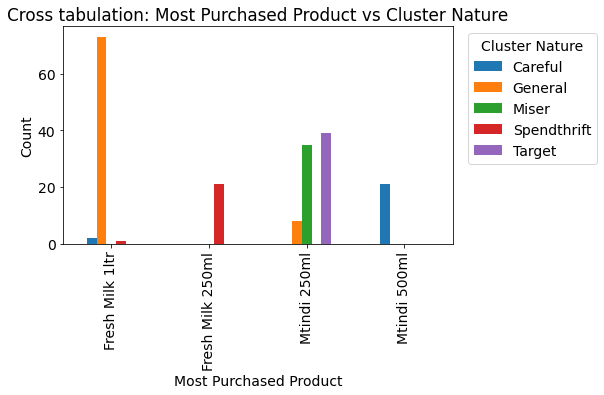
The plotting shows Mtindi 250ml is the most purchased product and Mtindi 500ml is the lowest purchased product.

Cross Tabulation: City vs Cluster nature



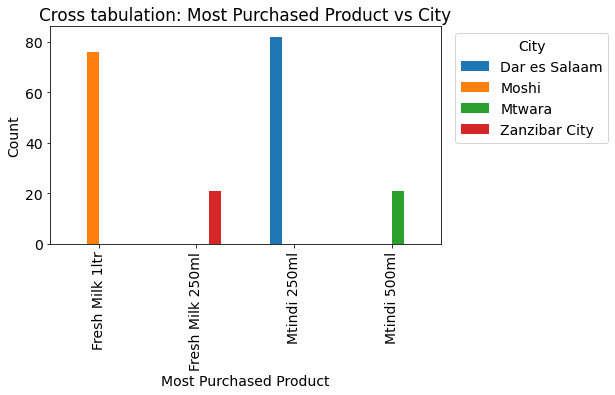
The above bar plot shows about cluster nature in the cities. In the city Dar es salaam, there are 3 natures of merchants General, Miser, Target. In Moshi General people is high. In Mtwara careful people is there. In Zanzibar city have only spendthrift people. This information helps the company to marketing the customer prefered products easily.

Cross Tabulation: Most purchased product vs Cluster Nature



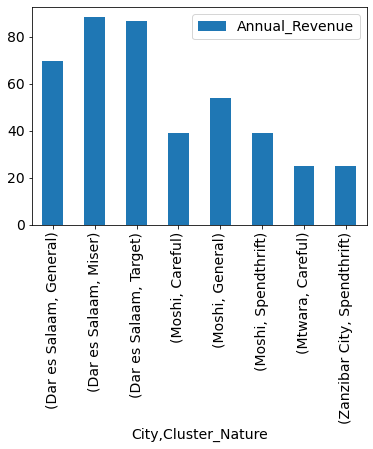
The above bar plot shows about Most purchased product based on the cluster nature. More general people use Fresh Milk 1ltr product. Only Spendthrift people are using Fresh milk 250ml product. Mtindi 250ml product is used by general, miser and Target people. Mtindi 500ml product is used by only careful people.

Cross Tabulation: Most purchased product vs city



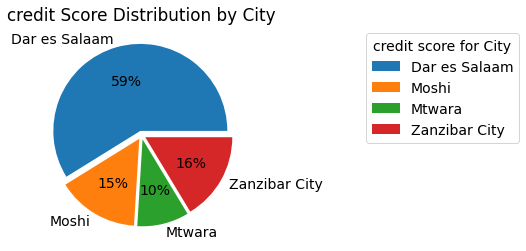
In this plot ,it shows about most purchased product in the cities. In Moshi, Freshmilk 1ltr product is most purchased. In Zanzibar city, Fresh Milk 250ml is most purchased. In Dar es salaam ,Mtindi 250ml is moat purchased. In Mtwara, Mtindi 500ml is most purchased.

Annual Revenue:



The above pivot table shows about Annual revenue. On the basis of Annual revenue, people are divided into different cluster natures. Dar es salaam city has general, Miser, Target cluster nature people. Moshi city has Careful, General, Spendthrift cluster nature people. Mtwra has only careful cluster people. Zanzibar city has Spendthrift.

Credit Score Distribution:



The above pie plot shows about credit score in cities.

Dar es salaam : 59%

Zanzibar city : 16%

Moshi : 15%

Mtwara : 10%

The highest credit score has Dar es salaam 59%. Mtwara city has least credit score 10%.