**OneBharat: Assignment for DS Interns hiring**

## **Bank Statements (P1- BankStatements.json) – 50 Marks**

### **1.  Transaction Analysis:**

   a) What is the total number of transactions made over the year?

Total number of transactions: 985

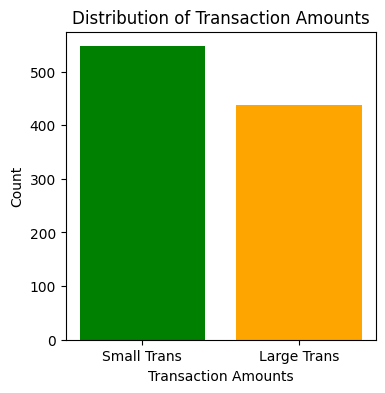
  b) What is the distribution of transaction amounts (e.g., small vs. large transactions)? (define small and large transactions by yourself)

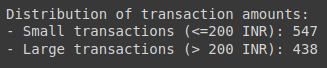
we can use multiple ways to determine small and large transactions   
1. Exploratory data analysis: **Visualize the Data:** Create visualizations like histograms or box plots to see the distribution of transaction amounts.we can determine the threshold value

2. statistical methods: **Mean and Median:** Calculate the mean and median transaction amounts. The median is particularly useful if the data is skewed.

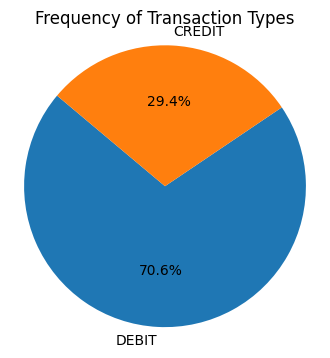
3.ML algorithms: K Means clustering is suitable for segmenting transactions into small and large categories. It iteratively groups transactions based on their amounts creating two distinct clusters.By analysing the average transactionvalues with each cluster you can effectively label them as small and large.

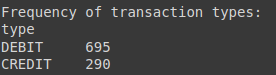
3.Naive approach: Alternatively, set fixed thresholds





   c) Analyze the frequency of different transaction types (debit vs. credit).



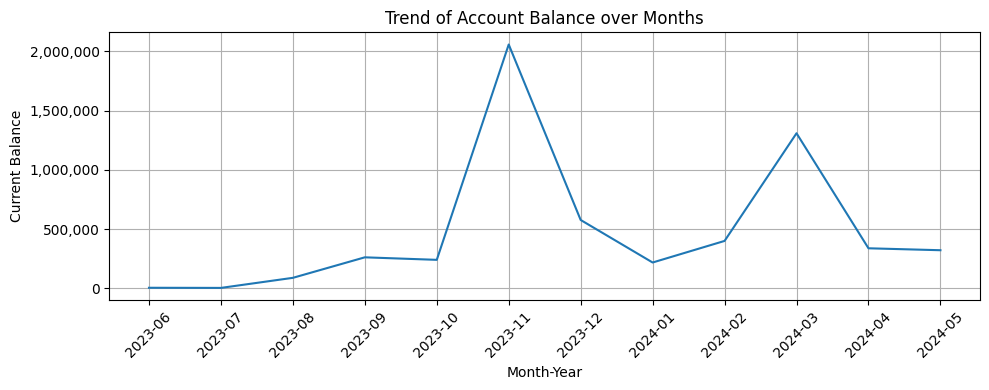


Fom the table, we can understand that Debit transactions are greater than Credit transactions.

### 

### **2.  Balance Analysis:**

1. What is the trend of the account balance over time?

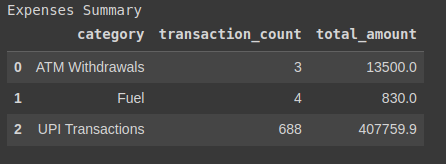


1. Identify any periods with significant changes in the account balance.

According to the Trend of Account Balance over months,The account balance saw significant increases in November 2023 and March 2024.

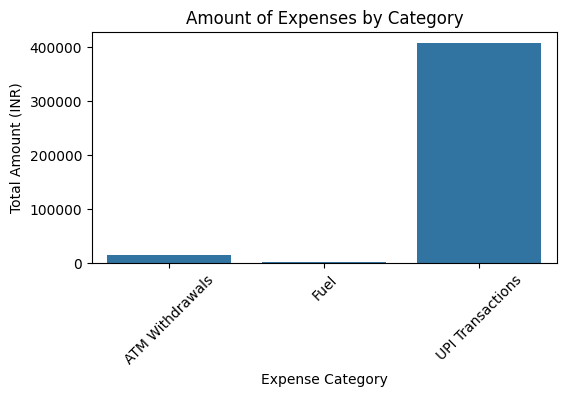
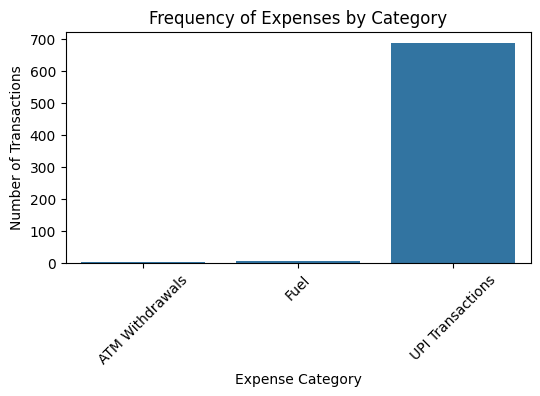
### 3**.  Spending Patterns:**

   a) What are the main categories of expenses (e.g., fuel, e-commerce, food, shopping, ATM withdrawals, UPI transactions)?



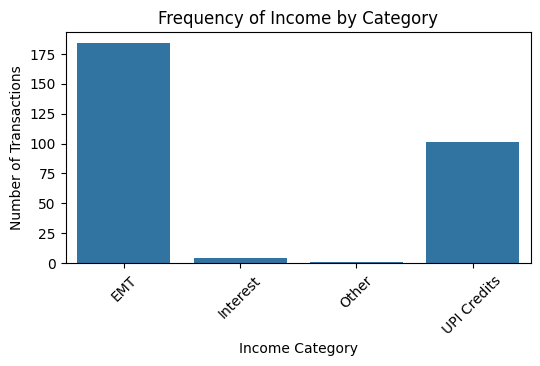
b) Analyze the frequency and amount of spending in each category.

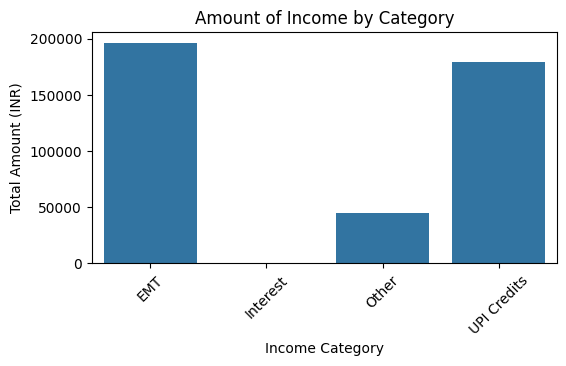
CATEGORY: EXPENSES



UPI Transactions (type = Debit) are the most frequent transactions for the expenses category.

CATEGORY: INCOME

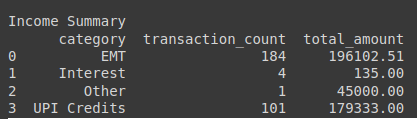




According to the Analysis, EMT and UPI credits are the main sources of Income.

### **4.  Income Analysis:**

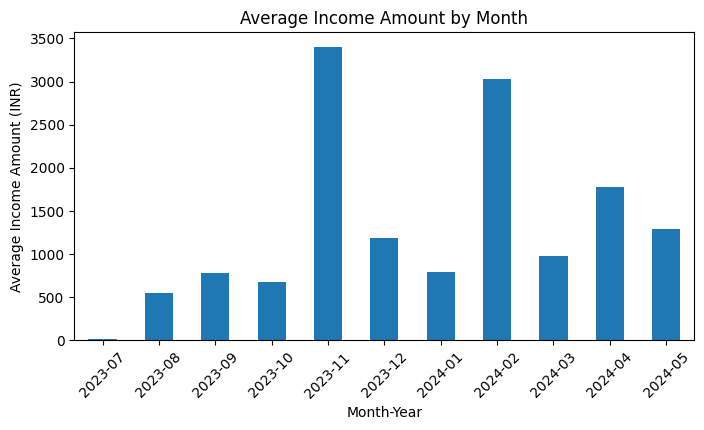
1. What are the main sources of income (e.g., salary, UPI credits)?



According to the Analysis, EMT and UPI credits are the main sources of Income.

Here I’ve considered EMT as NEFT and IMPS keywords from narration.

1. Identify any patterns in the timing and amount of income received.

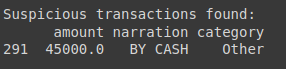


* **November 2023** and **February 2024** stand out as the months with the highest average income amounts, around 3500 INR and 3000 INR respectively.
* The income trend does not show a clear upward or downward movement, suggesting fluctuations rather than a steady increase or decrease over the observed period.
* A pattern might exist where every few months, there is a spike in income. For example, February 2024 and November 2023 are approximately three months apart.

### **5.  Alert Generation:**

1. Identify any unusual or suspicious transactions.

Method-1: Hit and Trail Method



There is one unusual or suspicious transaction. I’ve set the transaction amount threshold to be greater than 10,000 and the category to be 'Other' to identify suspicious transactions. The threshold can be adjusted based on the amount.

We can also try for different categories like setting the threshold amount as 10000 and the category as ‘Other’ or ‘ ATM withdrawals’.

Method-2: Using Anomaly Detection.

* **Feature Selection**: Select the columns ‘amount’ and ‘currentBalance’ from the ‘transaction\_df’ DataFrame as the relevant features for anomaly detection.
* **Standardize Features**: Standardization is applied to the feature matrix X to ensure that each feature has a mean of 0 and a standard deviation of 1. This is necessary because LOF is sensitive to the scale of the features.

Without standardization, features with larger scales can dominate the distance computations, leading to biased results.

* **Fit and Transform**: The fit\_transform method standardizes the data by fitting the scaler to X and transforming it. The result, X\_scaled, is the standardized feature matrix.

The Local Outlier Factor (LOF) model is a method used for detecting anomalies by measuring the local density deviation of a given data point with respect to its neighbours. It identifies points that have a significantly lower density compared to their neighbours, marking them as potential outliers.

* **Initialize LOF Model**: The LOF model is initialized with 20 neighbors (n\_neighbors=20) and a contamination level of 0.1 (contamination=0.1). The contamination parameter specifies the proportion of data points expected to be outliers.

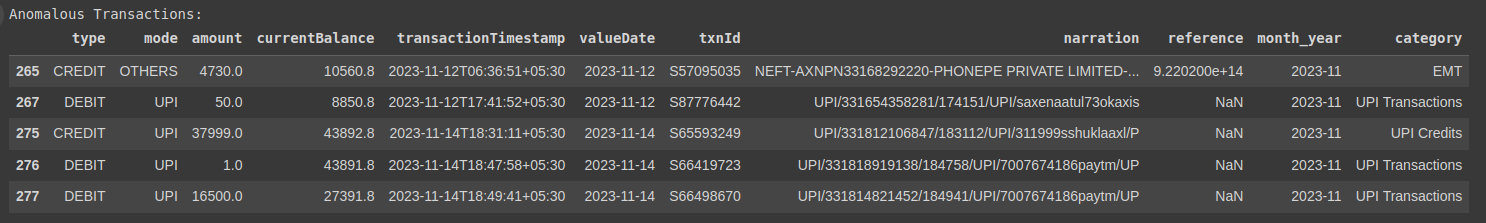
n\_neighbors: The number of neighbours to use for calculating the local density. A typical default value is 20, but it can be adjusted based on the dataset.

contamination: The proportion of outliers in the dataset. Setting this parameter helps the model understand how many points to consider as anomalies.

**Model Sensitivity**: The parameters n\_neighbors and contamination are crucial for tuning the model’s sensitivity. A lower number of neighbors or higher contamination can increase the sensitivity to outliers, but may also increase false positives.

* **Fit the Model**: The LOF model is fit to the standardized data X\_scaled.
* **Compute Anomaly Scores**: The anomaly scores for each data point are obtained using the negative\_outlier\_factor\_ attribute of the fitted LOF model. Lower scores indicate higher likelihood of being an anomaly.
* **Set Threshold**: A threshold of -1.5 is set for determining anomalies. This arbitrary value may need adjustment based on the specific dataset and desired sensitivity.
* **Flag Anomalies**: Transactions with an anomaly score below the threshold are flagged as anomalous. These transactions are selected from the original transactions\_df DataFrame and stored in anomalous\_transactions.
* **Print Results**:

There are a total of 56 anomalous transactions and here I’m showing a few transactions for your reference.



#### **1. Rule-Based Alerts**

* **Threshold-based:** Set predefined thresholds for low balance and high expenditure. When these limits are breached, send an alert.
* **Time-based:** Generate alerts for unusual spending patterns within specific timeframes (e.g., daily, weekly, monthly).

#### **2. Machine Learning-Based Alerts**

* **Anomaly Detection:** Employ algorithms to identify outliers in spending behavior. Unusual spending patterns can trigger alerts.
* **Predictive Modeling:** Build models to forecast future spending and balance levels. Alerts can be generated for predicted low balances or high expenditures.

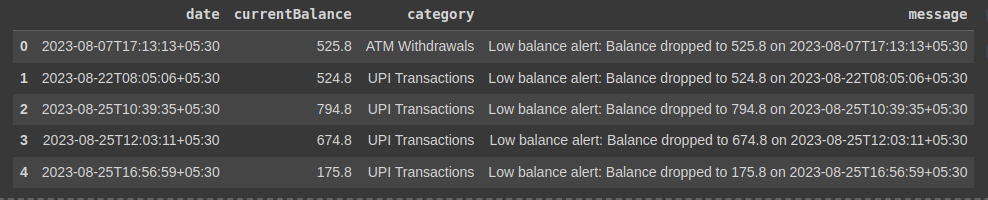
#### **3. User-Defined Alerts**

* **Custom Thresholds:** Allow users to set their own low balance and high expenditure thresholds.
* **Keyword-Based Alerts:** Enable users to receive alerts for specific transaction types or merchants.

1. Generate alerts for low balance or high expenditure periods.

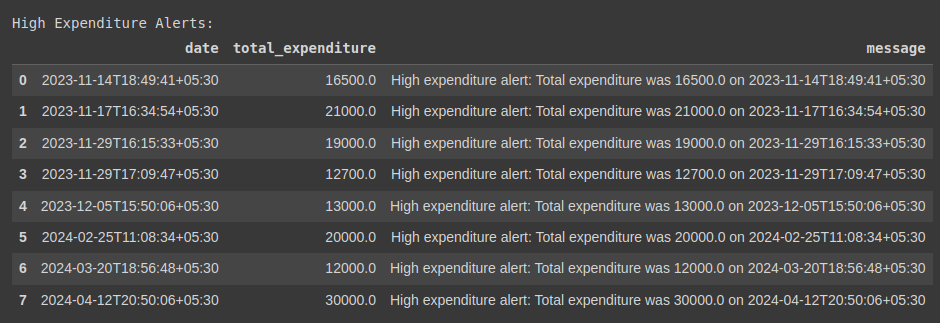
Low Balance Alerts:

There are approximately 209 rows with a low balance. This is determined by setting the low balance threshold to 1000 and checking if the 'currentBalance' is less than 1,000, which then generates an alert.



High Expenditure Periods:

There are 8 rows with High Expenditure. This is determined by setting the high expenditure threshold as 10000 and transaction type as ‘DEBIT’ .



## **Office Supplies Data (P2- OfficeSupplies Data.csv) – 20 marks**

### 

### 1. Sales Analysis:

1. What are the total sales for each product category?

| **ITEM** | **TOTAL SALES** |
| --- | --- |
| Binder | 9577.65 |
| Desk | 1700.00 |
| Pen | 2045.22 |
| Pen Set | 4169.87 |
| Pencil | 2135.14 |

1. Which product category has the highest sales?

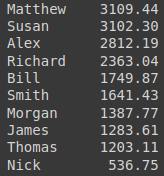
According to the above table, **Binder** product has the highest sales.

1. Identify the top 10 best-selling products.

Since there are only 5 products, Binder, Desk, Pen, Pen Set, and Pencil are the top-selling products.

### 2.  Customer Analysis:

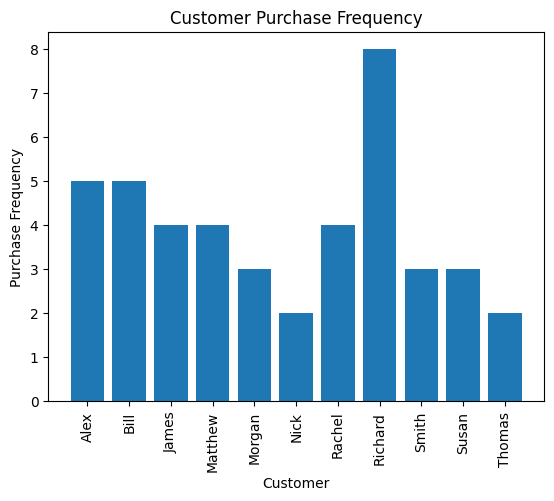
1. Who are the top 10 customers by sales?

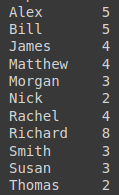


1. What is the total number of unique customers?

The total number of Unique customers is: **11**

1. Analyze customer purchase frequency.

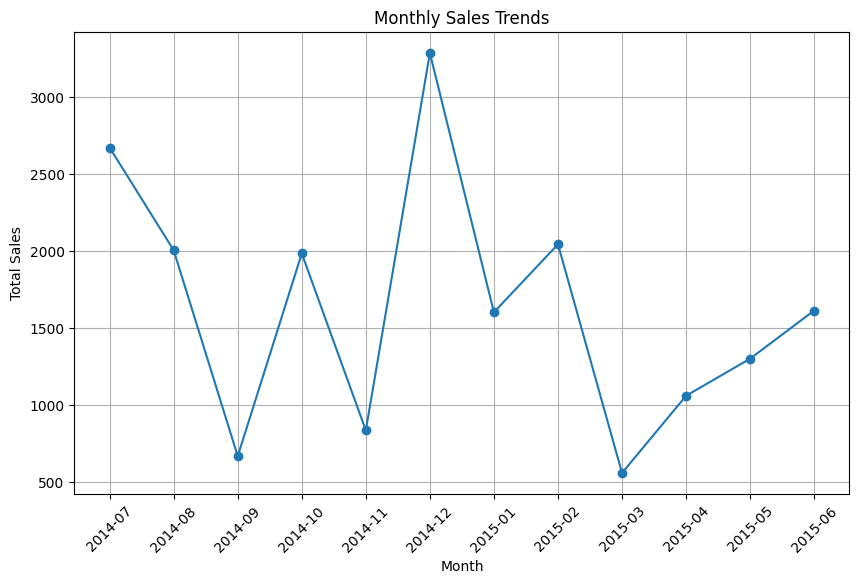




Richard has the highest purchase frequency.

### **3.  Time Series Analysis:**

1. What are the monthly sales trends over the past year?



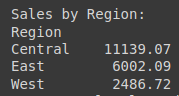
According to this graph, sales in December 2014 exceeded 3,500, while July 2014 had the second-highest sales, around 2,700.

1. Identify any seasonal patterns in the sales data.

* There is a significant peak in sales during **December 2014**. This suggests a potential increase in sales. Sales drop sharply in **August 2014** and **October 2014** but rose again in **September 2014** and **November 2014.**
* The lowest sales are observed in **January 2015,** where customers may be spending less after the end of the year.
* Peaks in sales can be seen approximately every quarter (e.g., September 2014, December 2014, March 2015).

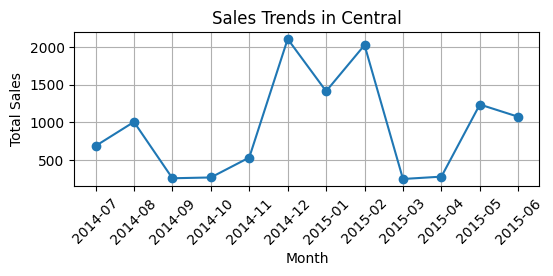
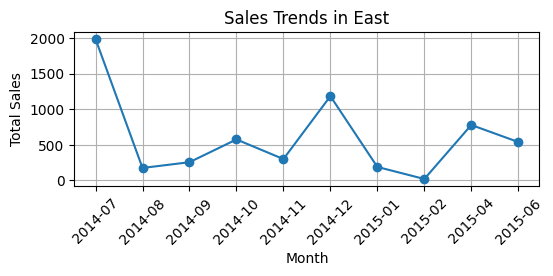
### **4.  Geographical Analysis:**

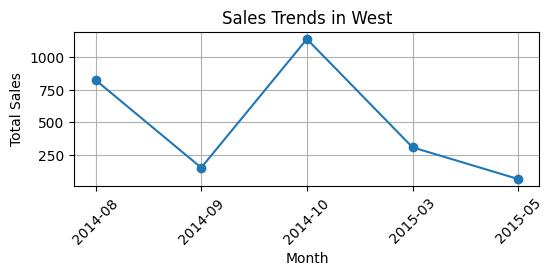
1. Which regions generate the most sales?



According to the table, **Central Region** generates the most sales.

1. What are the sales trends across different regions?

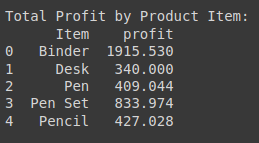




### 

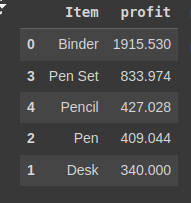
### 5.  Profit Analysis:

1. What is the total profit for each product category? (considering profit margin as 20%)



1. Identify the top 10 most profitable products.

Since there are only 5 products, The most profitable products, in order, are:



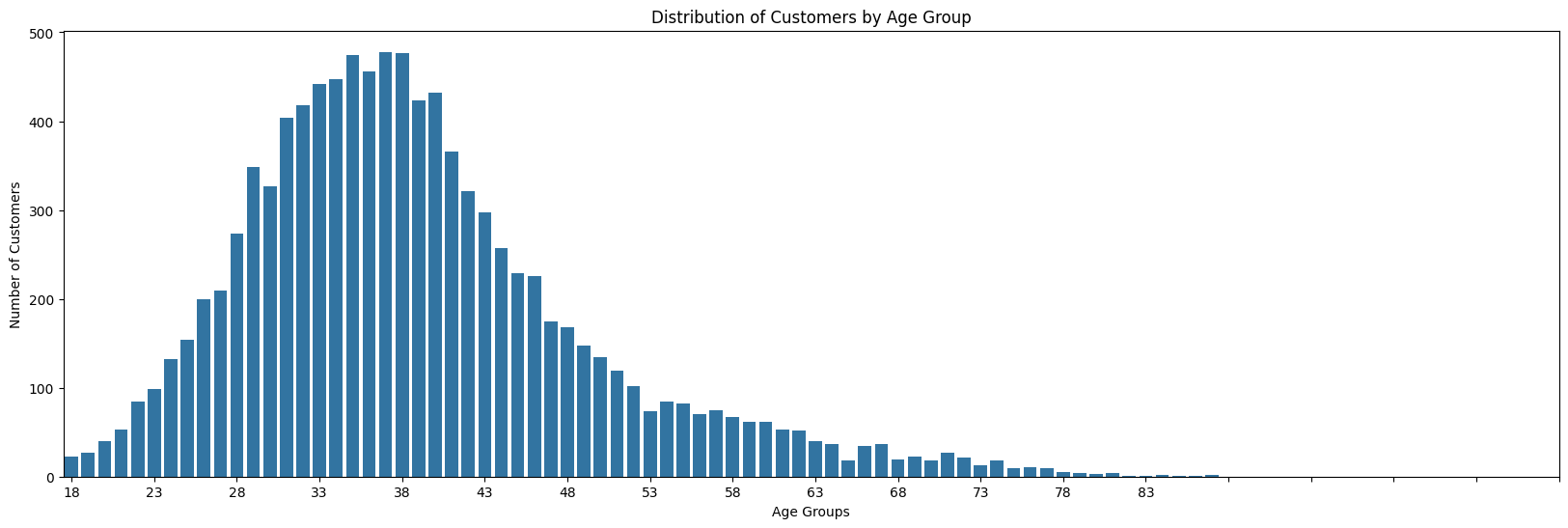
## 

## **Churn Modelling Data (P3- Churn-Modelling Data.xlsx) – 30 Marks**

### 1.  Customer Demographics:

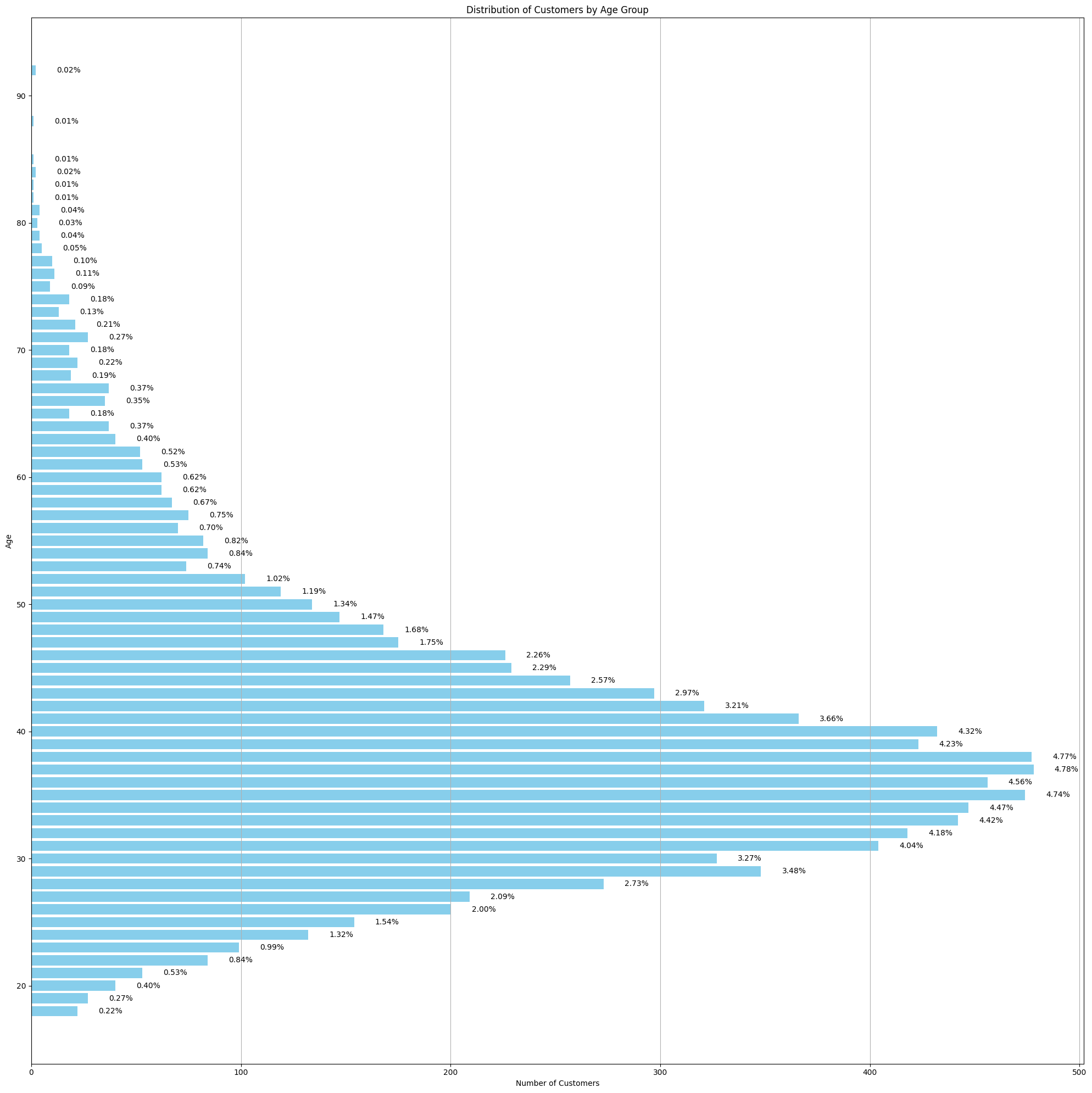
1. What is the distribution of customers across different age groups?

This graph shows the distribution of customers across different ages.

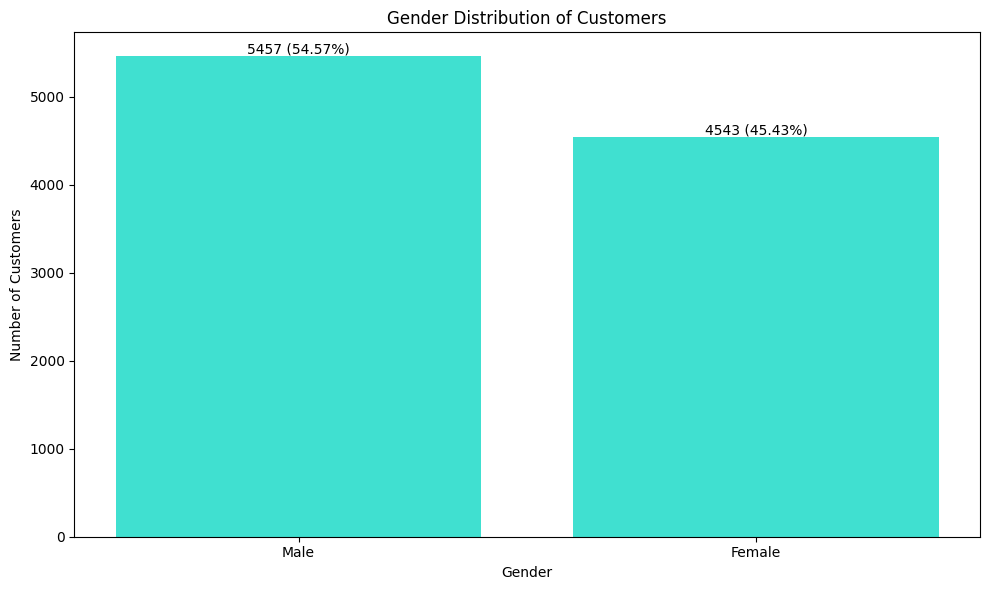


This table shows the percentage distribution of customers across different age groups

| **Age Group** | **Percentage** |
| --- | --- |
| 0 - 20 yrs | 0.89% |
| 21- 30 yrs | 18.89% |
| 31 - 40 yrs | 44.51% |
| 41 -50 yrs | 23.20% |
| 51- 60 yrs | 7.97% |
| 61 - 70 yrs | 3.31% |
| 71 - 80 yrs | 1.21% |
| 80 - 90 yrs | 0.1% |
| 90+ yrs | 0.02% |



1. Analyze the gender distribution of customers.



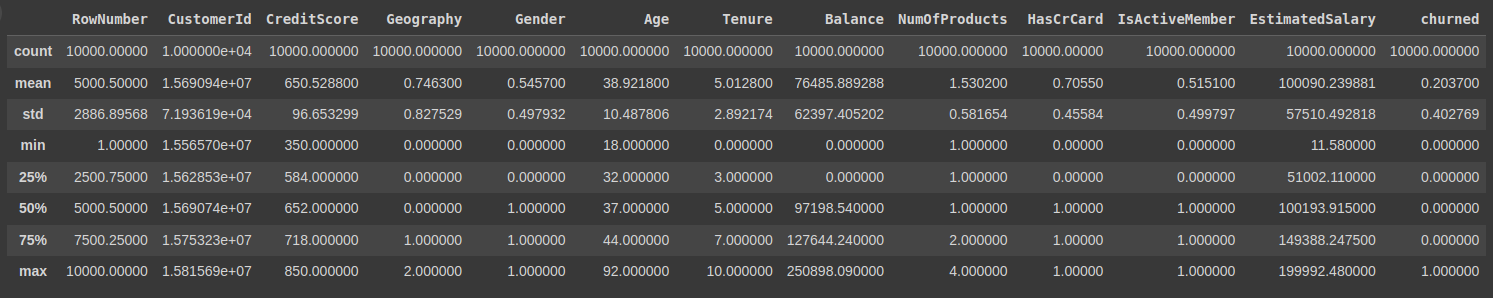
### 2.  Churn Analysis:

1. What percentage of customers have churned?

| **Churned** | **Customer** |
| --- | --- |
| 0 | 7963 |
| 1 | 2037 |

According to this table, **20.37 %** of customers have churned

1. What are the main reasons for customer churn?



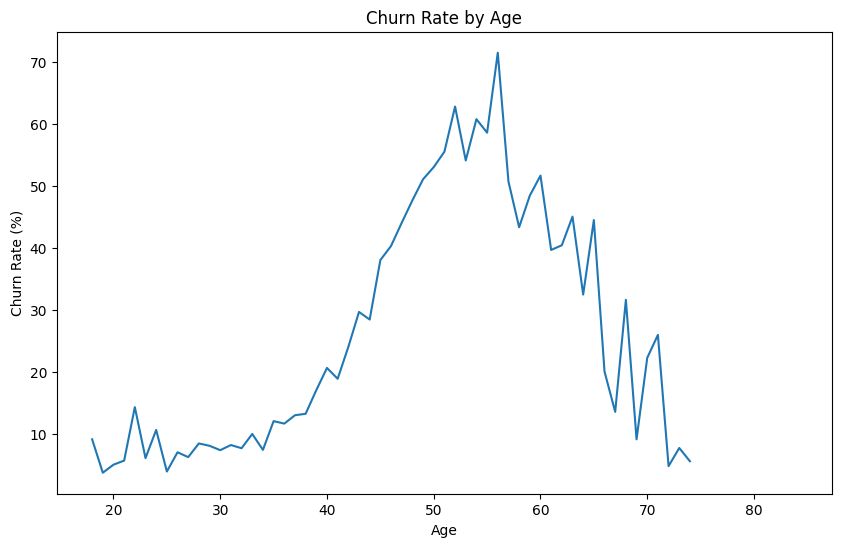
The summary statistics of the dataset provide insight into the distribution of features and can help identify potential factors contributing to customer churn. Here is an analysis of the key features:

**Analysis of Key Features:**

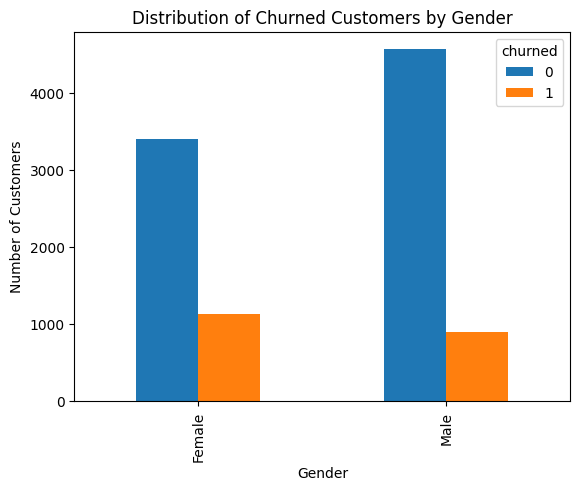
1. **CreditScore**:
   * **Mean**: 650.53
   * **Standard Deviation**: 96.65
   * **Range**: 350 to 850
   * **Interpretation**: The credit scores vary widely among customers. A lower credit score could indicate higher risk, potentially contributing to churn.
2. **Age**:
   * **Mean**: 38.92
   * **Standard Deviation**: 10.49
   * **Range**: 18 to 92
   * **Interpretation**: The customers' ages range from young adults to seniors, with a slightly higher concentration in middle age. Older customers might have different service expectations, potentially influencing churn.
3. **Tenure**:
   * **Mean**: 5.01 years
   * **Standard Deviation**: 2.89
   * **Range**: 0 to 10 years
   * **Interpretation**: Tenure indicates how long customers have been with the company. Shorter tenure might be associated with higher churn rates, as newer customers may leave before fully engaging with the services.
4. **Balance**:
   * **Mean**: 76,485.89
   * **Standard Deviation**: 62,397.41
   * **Range**: 0 to 250,898.09
   * **Interpretation**: The balance varies significantly, with some customers having no balance. Customers with higher balances might be less likely to churn due to their financial commitment.
5. **NumOfProducts**:
   * **Mean**: 1.53 products
   * **Standard Deviation**: 0.58
   * **Range**: 1 to 4 products
   * **Interpretation**: Customers typically have between 1 and 2 products. More products might indicate higher engagement and lower churn.
6. **HasCrCard**:
   * **Mean**: 0.71 (71% have a credit card)
   * **Interpretation**: Having a credit card might be a factor in customer retention.
7. **IsActiveMember**:
   * **Mean**: 0.52 (52% are active members)
   * **Interpretation**: Active members are expected to have a lower churn rate due to higher engagement.
8. **EstimatedSalary**:
   * **Mean**: 100,090.24
   * **Standard Deviation**: 57,510.49
   * **Range**: 11.58 to 199,992.48
   * **Interpretation**: Salaries vary widely. The relationship between salary and churn might depend on how well the services meet the expectations of different income groups.
9. **Churned**:
   * **Mean**: 0.20 (20.37% churn rate)
   * **Interpretation**: The dataset has a moderate churn rate, with about one-fifth of customers having churned.

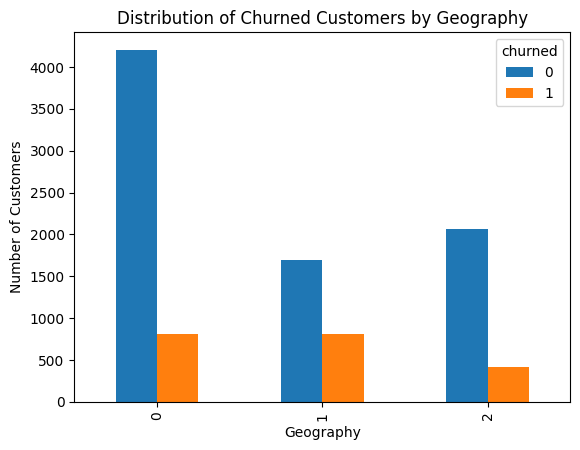
Based on the statistical summary, key factors that could contribute to customer churn include **Credit Score, Age, Tenure, Number of Products,** and **Estimated Salary.**

1. Identify any patterns or trends among customers who have churned.



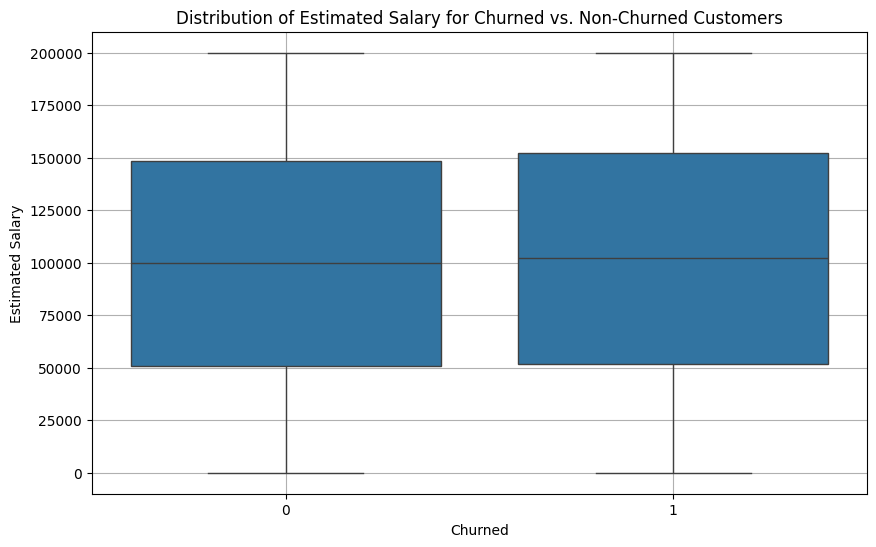
The graph indicates that customers aged 40-65 are more likely to churn.





According to this distribution of Churned Customers by Gender, females are more likely to have churned.

According to this distribution of Churned Customers by Geography, customers from Germany and France are most likely to have churned.



This box plot suggests that estimated salary does not have a strong distinguishing feature between churned and non-churned customers, as the distributions appear quite similar.

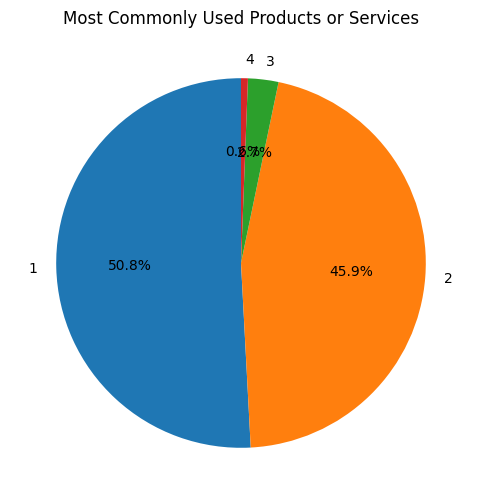
### 

### 3.  Product Usage:

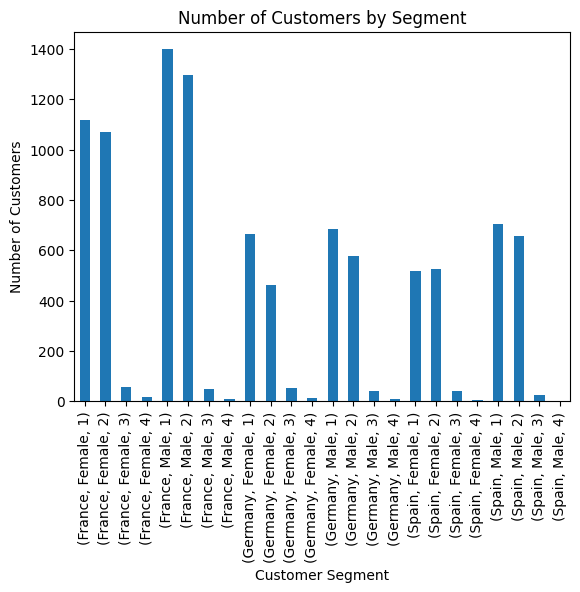
1. What are the most commonly used products or services?

In this dataset, there are no products or services. Only a number of products are mentioned.

1. Analyze the usage patterns of different customer segments.



According to this Pie chart, around 50% of customers are using only one product and 45% of customers are using 2 products. Around 95% of customers are not willing to have more than 2 products.



According to Distribution of Number of Customers by Segment, I tried to analyse the number of customers by segment, which consists of Geography, Gender and Number of Products.

### **4.  Financial Analysis:**

1. What is the average account balance of customers?

Average account balance: 76485.889288

Average balance of churned customers: 91108.53933726068

Average balance of non-churned customers: 72745.2967788522

1. Compare the financial characteristics of churned vs. non-churned customers.

Average Credit Score of churned customers: 645.3514972999509

Average Credit Score of non-churned customers: 651.8531960316463

Average of Estimated Salary of churned customers:101465.67753068237

Average of Estimated Salary of non-churned customers:99738.39177194

Average tenure of churned customers:4.932744231713304

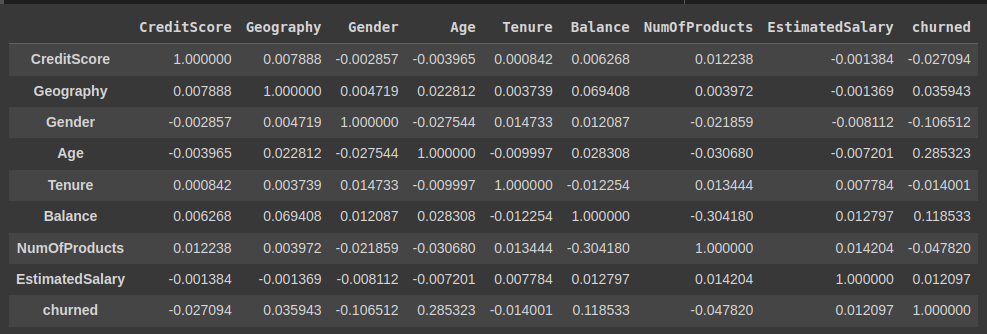
Average tenure of non-churned customers:5.033278914981791

### 

### 

### **5.  Predictive Modeling:**

1. Which factors are the most significant predictors of customer churn?



According to this correlation matrix, the top 5 features that are most significant predictors of customer churn are:  
1. Age

2. EstimatedSalary

3. CreditScore

4. Balance

5. NumOfProducts

1. Develop a predictive model to identify at-risk customers.

I developed a predictive model using a Random Forest Classifier and obtained an accuracy of **0.8565.**

In the future, I will use this predictive model to identify potential at-risk customers by examining the top significant features and testing these features with my random forest model. If the prediction is 1, indicating that the customer might churn, I will focus on retaining these customers.

Random Forest is a popular choice for this task due to several compelling reasons:

* **Handles Diverse Data Types:** It can effectively handle both numerical and categorical data, which is common in customer datasets (e.g., age, income, purchase history, demographics).
* **High Accuracy:** By creating multiple decision trees and combining their predictions, Random Forest often achieves high accuracy, reducing the risk of overfitting.
* **Feature Importance:** It provides insights into which customer attributes are most influential in predicting customer risk, helping to understand the underlying patterns.
* **Robustness to Outliers:** Random Forest is relatively insensitive to outliers, making it robust to noisy data.
* **Handles Missing Values:** It can handle missing values without requiring imputation, making it efficient for real-world datasets.
* **Interpretability:** While not as interpretable as individual decision trees, Random Forest still offers some level of explainability through feature importance.