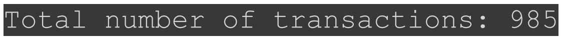
OneBharat: Assignment for DS Interns hiring

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

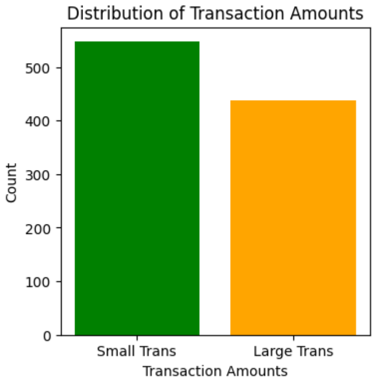
1. Transaction Analysis:

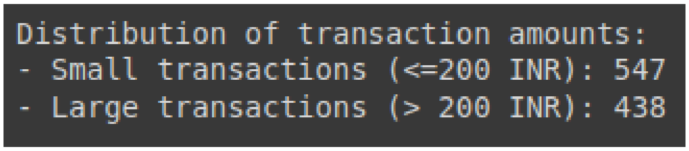
a) What is the total number of transactions made over the year?



```
Total number of transactions: 985
```

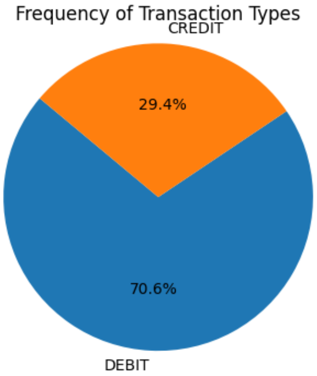
b) What is the distribution of transaction amounts (e.g., small vs. large transactions)? (define small and large transactions by yourself)

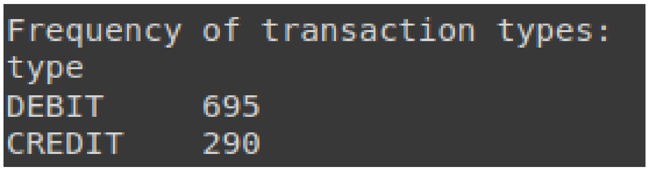




```
Distribution of transaction amounts:
- Small transactions (<=200 INR): 547
- Large transactions (> 200 INR): 438
```

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



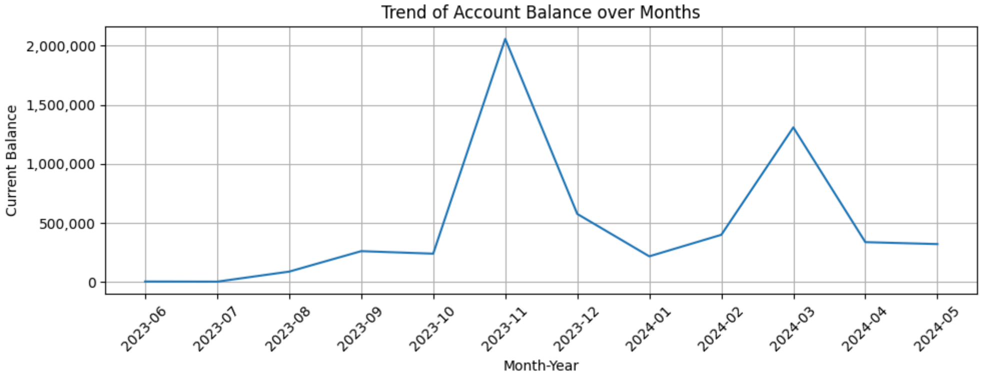


```
Frequency of transaction types:
type
DEBIT     695
CREDIT    290
```

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

2. Balance Analysis:

a) What is the trend of the account balance over time?

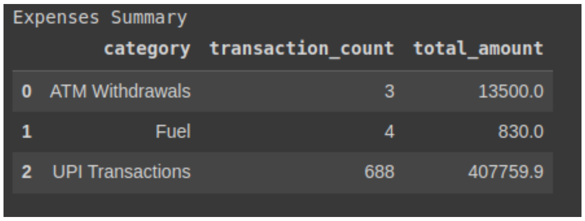


b) Identify any periods with significant changes in the account balance.

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)?

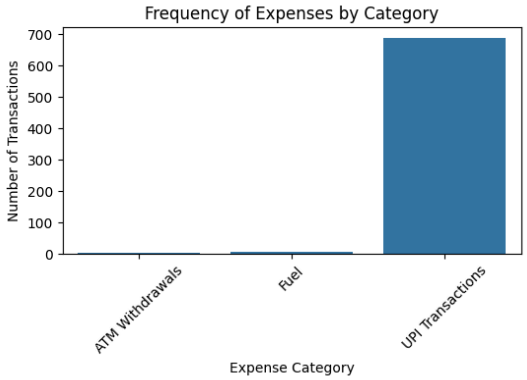


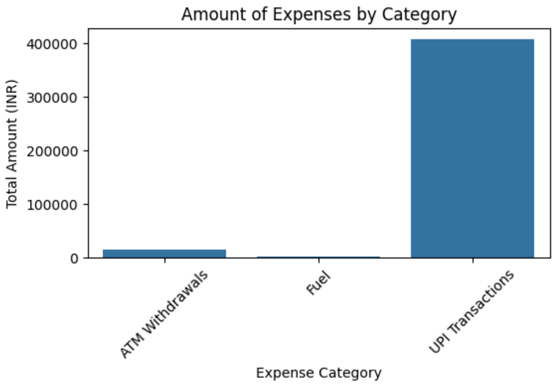
Expenses Summary

	category	transaction_count	total_amount
0	ATM Withdrawals	3	13500.0
1	Fuel	4	830.0
2	UPI Transactions	688	407759.9

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

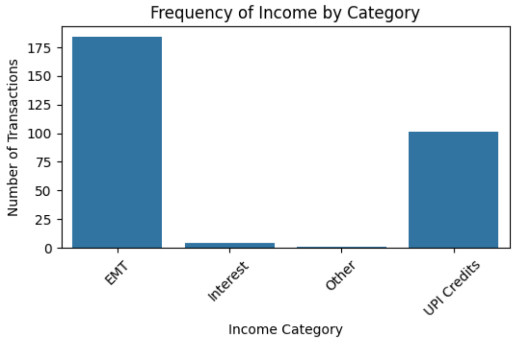
CATEGORY: EXPENSES

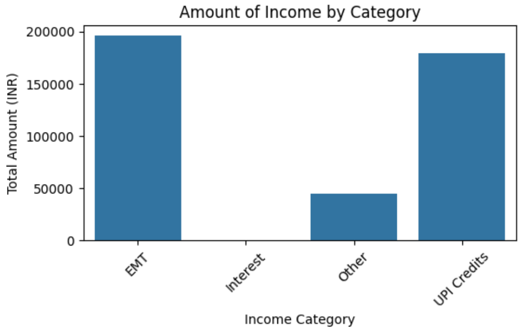




UPI Transactions (type = Debit) are the most frequent transactions for the expenses category.

CATEGORY: INCOME

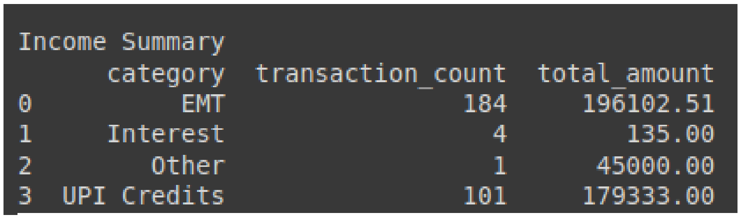




According to the Analysis, EMT and UPI credits are the main sources of Income.

4. Income Analysis:

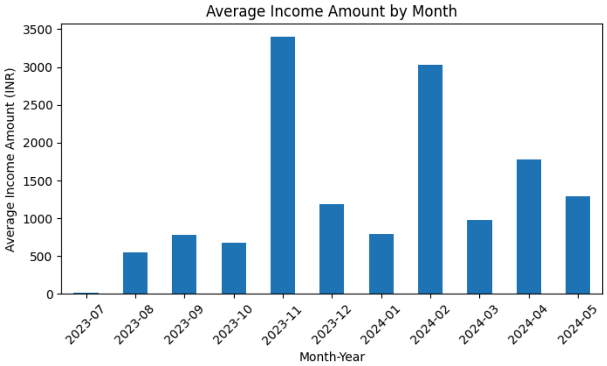
a) What are the main sources of income (e.g., salary, UPI credits)?



```
Income Summary
      category  transaction_count  total_amount
0          EMT                184     196102.51
1     Interest                  4        135.00
2        Other                  1      45000.00
3  UPI Credits                101     179333.00
```

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.

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

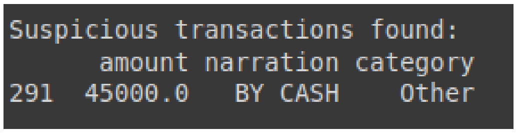


- **November 2023** and **February 2024** stand out as the months with the highest average income amounts, around 3500 INR and 3000 INR respectively.
- The income trend does not show a clear upward or downward movement, suggesting fluctuations rather than a steady increase or decrease over the observed 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:

a) Identify any unusual or suspicious transactions.

Method-1: Hit and Trail Method


```
Suspicious transactions found:
      amount narration category
291  45000.0   BY CASH    Other
```

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.

- **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.

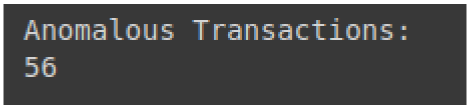
- **Initialize LOF Model**: The LOF model is initialized with 20 neighbors (`n_neighbors=20`) and a contamination level of 0.1 (`contamination=0.1`). The contamination parameter specifies the proportion of data points expected to be outliers.

- **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.

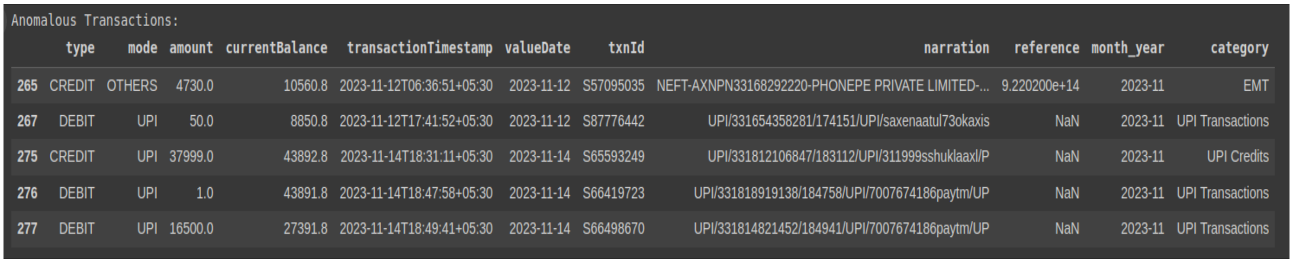
- **Set Threshold**: A threshold of -1.5 is set for determining anomalies. This arbitrary value may need adjustment based on the specific dataset and desired sensitivity.

- **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**: 

```
Anomalous Transactions:
56
```

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



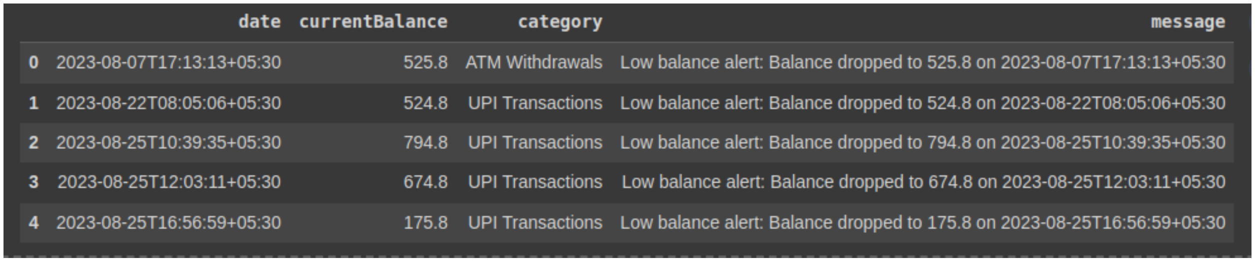
Anomalous Transactions:

	type	mode	amount	currentBalance	transactionTimestamp	valueDate	txnId	narration	reference	month_year	category
265	CREDIT	OTHERS	4730.0	10560.8	2023-11-12T06:36:51+05:30	2023-11-12	S57095035	NEFT-AXNPN33168292220-PHONEPE PRIVATE LIMITED-...	9.220200e+14	2023-11	EMT
267	DEBIT	UPI	50.0	8850.8	2023-11-12T17:41:52+05:30	2023-11-12	S87776442	UPI/331654358281/174151/UPI/saxenaatul73okaxis	NaN	2023-11	UPI Transactions
275	CREDIT	UPI	37999.0	43892.8	2023-11-14T18:31:11+05:30	2023-11-14	S65593249	UPI/331812106847/183112/UPI/311999sshuklaaxl/P	NaN	2023-11	UPI Credits
276	DEBIT	UPI	1.0	43891.8	2023-11-14T18:47:58+05:30	2023-11-14	S66419723	UPI/331818919138/184758/UPI/7007674186paytm/UP	NaN	2023-11	UPI Transactions
277	DEBIT	UPI	16500.0	27391.8	2023-11-14T18:49:41+05:30	2023-11-14	S66498670	UPI/331814821452/184941/UPI/7007674186paytm/UP	NaN	2023-11	UPI Transactions

b) Generate alerts for low balance or high expenditure periods.

Low Balance Alerts:

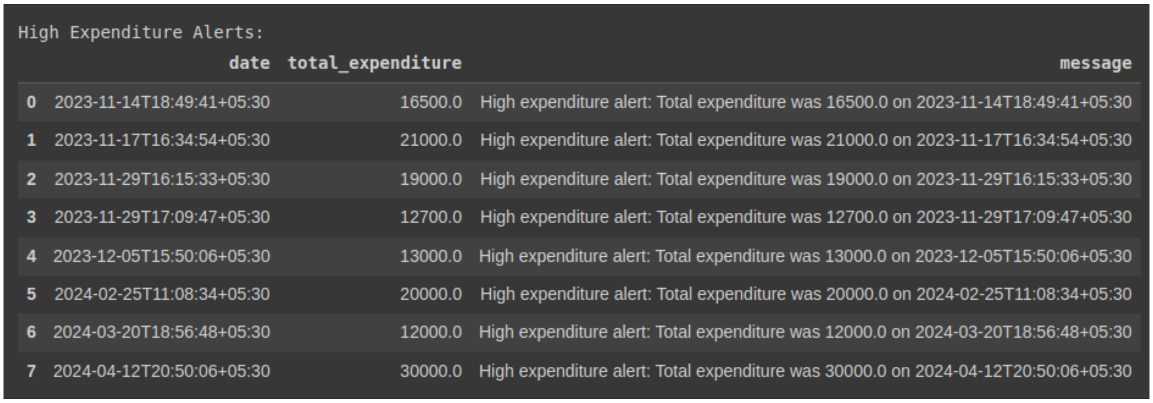
There are approximately 209 rows with a low balance. This is determined by setting the low balance threshold to 1000 and checking if the 'currentBalance' is less than 1,000, which then generates an alert.



	date	currentBalance	category	message
0	2023-08-07T17:13:13+05:30	525.8	ATM Withdrawals	Low balance alert: Balance dropped to 525.8 on 2023-08-07T17:13:13+05:30
1	2023-08-22T08:05:06+05:30	524.8	UPI Transactions	Low balance alert: Balance dropped to 524.8 on 2023-08-22T08:05:06+05:30
2	2023-08-25T10:39:35+05:30	794.8	UPI Transactions	Low balance alert: Balance dropped to 794.8 on 2023-08-25T10:39:35+05:30
3	2023-08-25T12:03:11+05:30	674.8	UPI Transactions	Low balance alert: Balance dropped to 674.8 on 2023-08-25T12:03:11+05:30
4	2023-08-25T16:56:59+05:30	175.8	UPI Transactions	Low balance alert: Balance dropped to 175.8 on 2023-08-25T16:56:59+05:30

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’ .



High Expenditure Alerts:

	date	total_expenditure	message
0	2023-11-14T18:49:41+05:30	16500.0	High expenditure alert: Total expenditure was 16500.0 on 2023-11-14T18:49:41+05:30
1	2023-11-17T16:34:54+05:30	21000.0	High expenditure alert: Total expenditure was 21000.0 on 2023-11-17T16:34:54+05:30
2	2023-11-29T16:15:33+05:30	19000.0	High expenditure alert: Total expenditure was 19000.0 on 2023-11-29T16:15:33+05:30
3	2023-11-29T17:09:47+05:30	12700.0	High expenditure alert: Total expenditure was 12700.0 on 2023-11-29T17:09:47+05:30
4	2023-12-05T15:50:06+05:30	13000.0	High expenditure alert: Total expenditure was 13000.0 on 2023-12-05T15:50:06+05:30
5	2024-02-25T11:08:34+05:30	20000.0	High expenditure alert: Total expenditure was 20000.0 on 2024-02-25T11:08:34+05:30
6	2024-03-20T18:56:48+05:30	12000.0	High expenditure alert: Total expenditure was 12000.0 on 2024-03-20T18:56:48+05:30
7	2024-04-12T20:50:06+05:30	30000.0	High expenditure alert: Total expenditure was 30000.0 on 2024-04-12T20:50:06+05:30

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

1. Sales Analysis:

a) 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

b) Which product category has the highest sales?

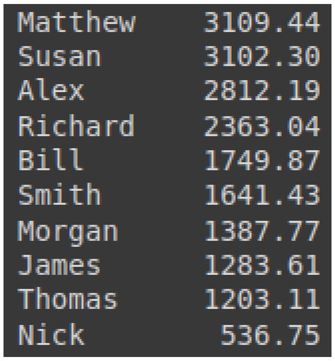
According to the above table, **Binder** product has the highest sales.

c) 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:

a) Who are the top 10 customers by sales?

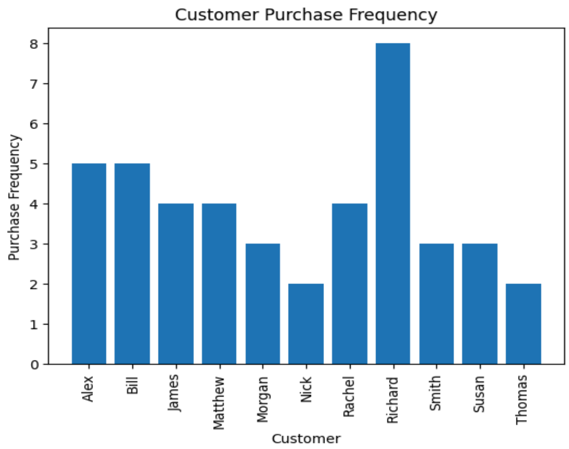


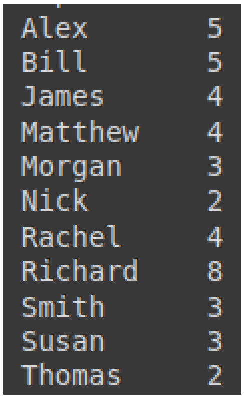
```
Matthew    3109.44
Susan      3102.30
Alex       2812.19
Richard    2363.04
Bill       1749.87
Smith      1641.43
Morgan     1387.77
James      1283.61
Thomas     1203.11
Nick        536.75
```

b) What is the total number of unique customers?

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

c) Analyze customer purchase frequency.



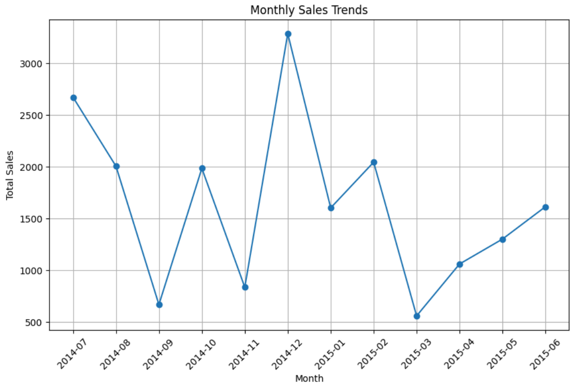


```
Alex        5
Bill        5
James       4
Matthew     4
Morgan      3
Nick        2
Rachel      4
Richard     8
Smith       3
Susan       3
Thomas      2
```

Richard has the highest purchase frequency.

3. Time Series Analysis:

a) What are the monthly sales trends over the past year?



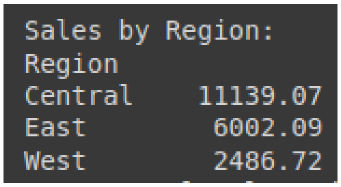
According to this graph, sales in December 2014 exceeded 3,500, while July 2014 had the second-highest sales, around 2,700.

b) 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:

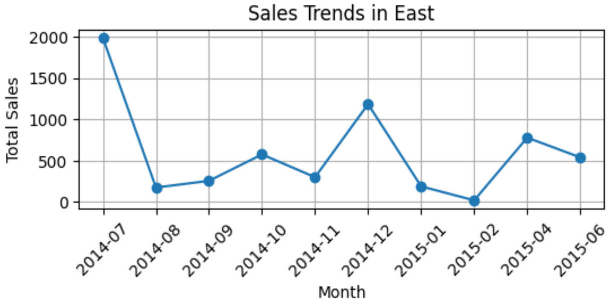
a) Which regions generate the most sales?

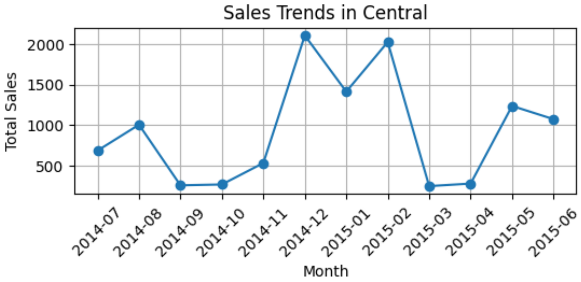


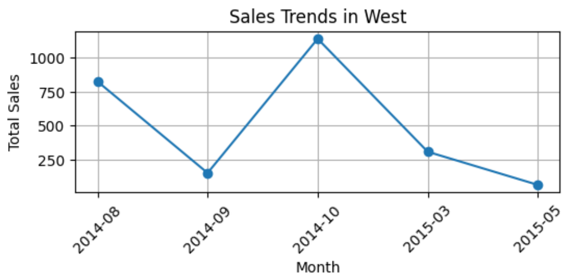
```
Sales by Region:
Region
Central     11139.07
East         6002.09
West         2486.72
```

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

b) What are the sales trends across different regions?

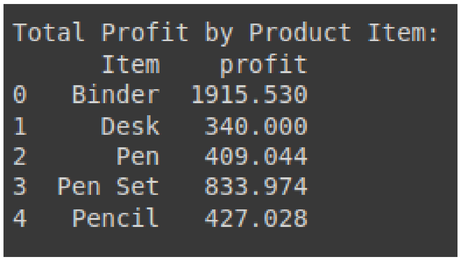






5. Profit Analysis:

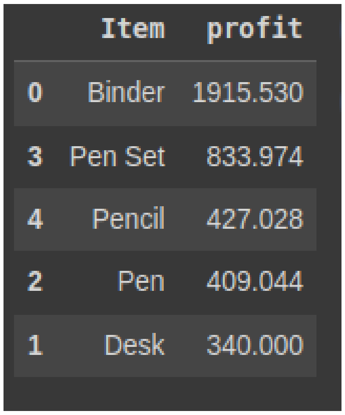
a) What is the total profit for each product category? (considering profit margin as 20%)



```
Total Profit by Product Item:
      Item    profit
0   Binder  1915.530
1     Desk   340.000
2      Pen   409.044
3  Pen Set   833.974
4   Pencil   427.028
```

b) Identify the top 10 most profitable products.

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



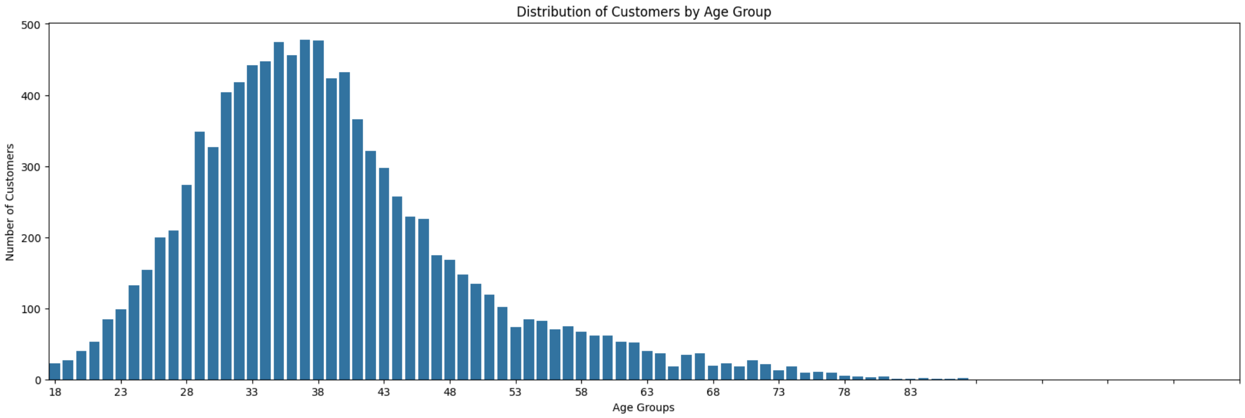
	Item	profit
0	Binder	1915.530
3	Pen Set	833.974
4	Pencil	427.028
2	Pen	409.044
1	Desk	340.000

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

1. Customer Demographics:

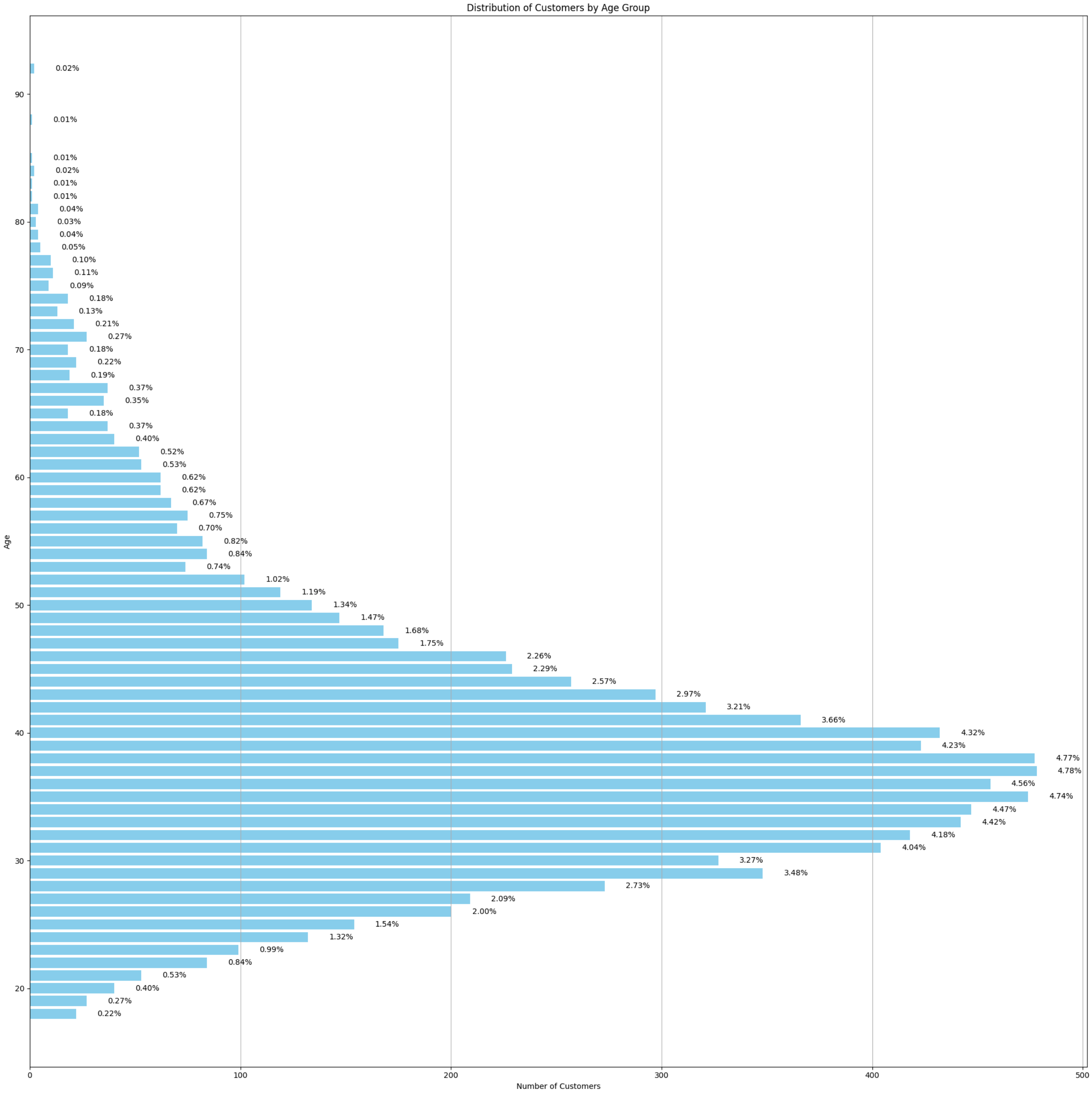
a) What is the distribution of customers across different age groups?

This graph shows the distribution of customers across different ages.

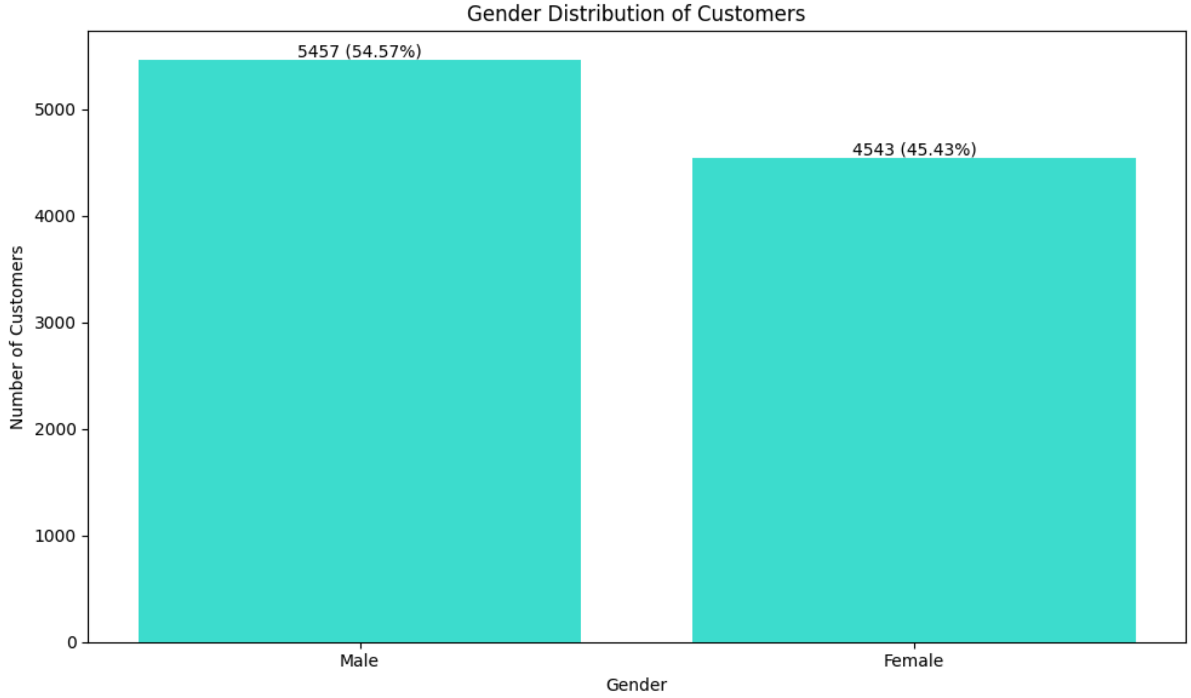


This table shows the percentage distribution of customers across different age groups

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%


Distribution of Customers by Age Group

b) Analyze the gender distribution of customers.



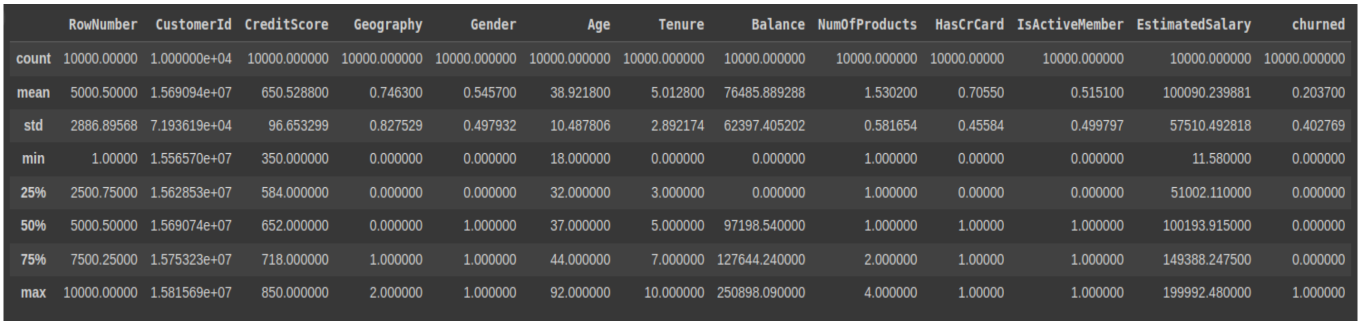
2. Churn Analysis:

a) What percentage of customers have churned?

Churned	Customer
0	7963
1	2037

According to this table, **<u>20.37 %</u>** of customers have churned

b) What are the main reasons for customer churn?



	RowNumber	CustomerId	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	churned
count	10000.00000	1.000000e+04	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.00000	10000.000000	10000.000000	10000.000000
mean	5000.50000	1.569094e+07	650.528800	0.746300	0.545700	38.921800	5.012800	76485.889288	1.530200	0.70550	0.515100	100090.239881	0.203700
std	2886.89568	7.193619e+04	96.653299	0.827529	0.497932	10.487806	2.892174	62397.405202	0.581654	0.45584	0.499797	57510.492818	0.402769
min	1.00000	1.556570e+07	350.000000	0.000000	0.000000	18.000000	0.000000	0.000000	1.000000	0.00000	0.000000	11.580000	0.000000
25%	2500.75000	1.562853e+07	584.000000	0.000000	0.000000	32.000000	3.000000	0.000000	1.000000	0.00000	0.000000	51002.110000	0.000000
50%	5000.50000	1.569074e+07	652.000000	0.000000	1.000000	37.000000	5.000000	97198.540000	1.000000	1.00000	1.000000	100193.915000	0.000000
75%	7500.25000	1.575323e+07	718.000000	1.000000	1.000000	44.000000	7.000000	127644.240000	2.000000	1.00000	1.000000	149388.247500	0.000000
max	10000.00000	1.581569e+07	850.000000	2.000000	1.000000	92.000000	10.000000	250898.090000	4.000000	1.00000	1.000000	199992.480000	1.000000

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:

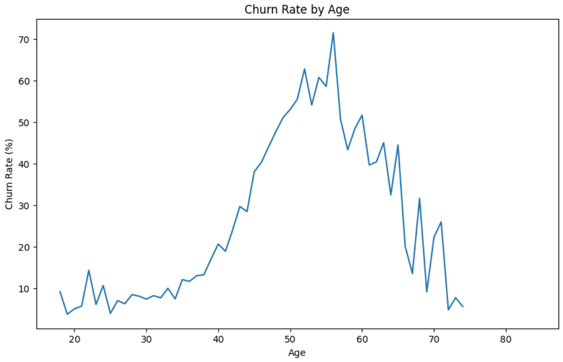
Analysis of Key Features:

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.

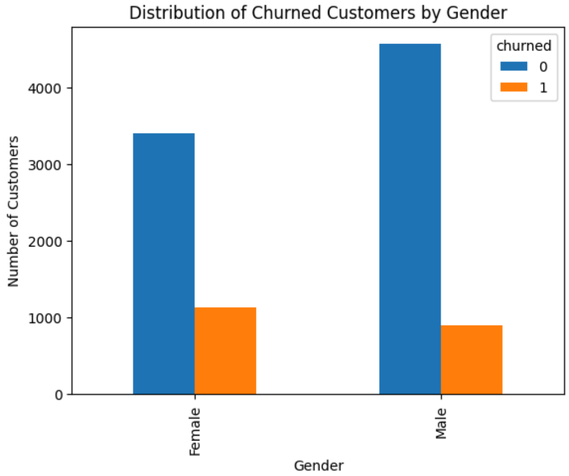
4. **Balance**:
    - **Mean**: 76,485.89
    - **Standard Deviation**: 62,397.41
    - **Range**: 0 to 250,898.09
    - **Interpretation**: The balance varies significantly, with some customers having no balance. Customers with higher balances might be less likely to churn due to their financial commitment.
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.

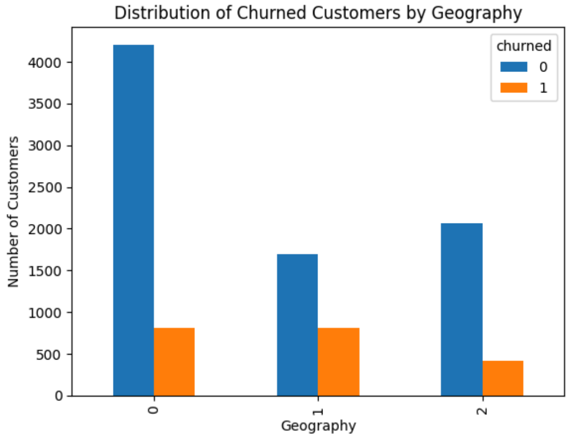
Based on the statistical summary, key factors that could contribute to customer churn include **Credit Score, Age, Tenure, Number of Products,** and **Estimated Salary.**

c) Identify any patterns or trends among customers who have churned.



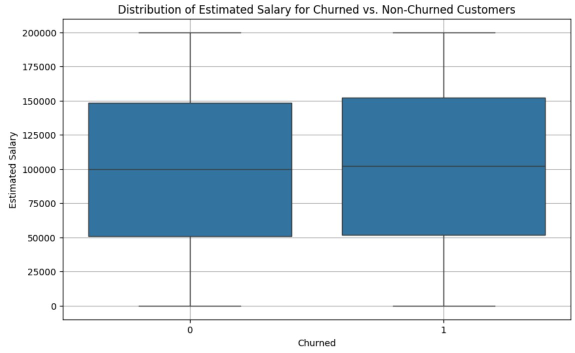
The graph indicates that customers aged 40-65 are more likely to churn.





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.



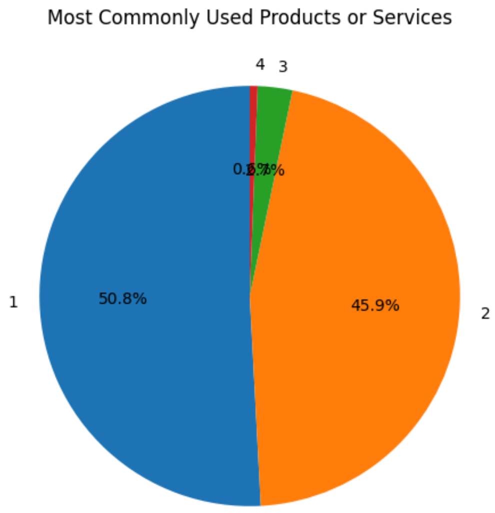
This box plot suggests that estimated salary does not have a strong distinguishing feature between churned and non-churned customers, as the distributions appear quite similar.

3. Product Usage:

a) What are the most commonly used products or services?

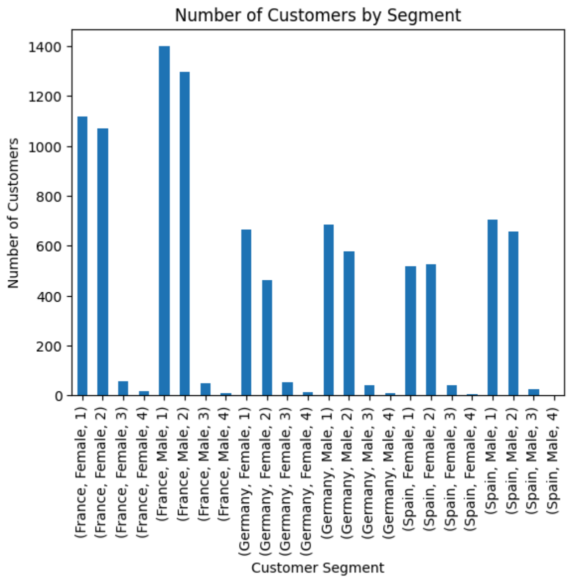
In this dataset, there are no products or services. Only a number of products are mentioned.

b) Analyze the usage patterns of different customer segments.



According to this Pie chart, around 50% of customers are using only one product and

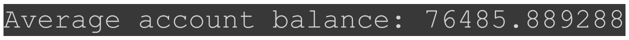
45% of customers are using 2 products. Around 95% of customers are not willing to have more than 2 products.



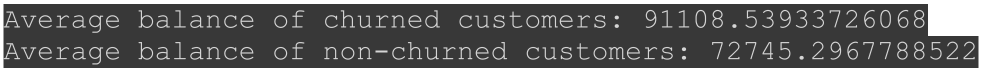
According to Distribution of Number of Customers by Segment, I tried to analyse the number of customers by segment, which consists of Geography, Gender and Number of Products.

4. Financial Analysis:

a) What is the average account balance of customers?

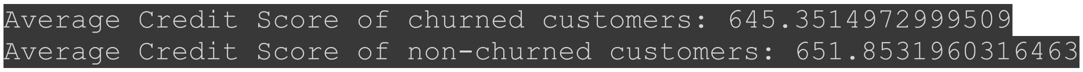


```
Average account balance: 76485.889288
```

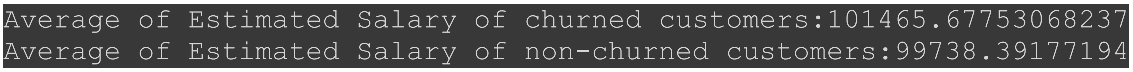


```
Average balance of churned customers: 91108.53933726068
Average balance of non-churned customers: 72745.2967788522
```

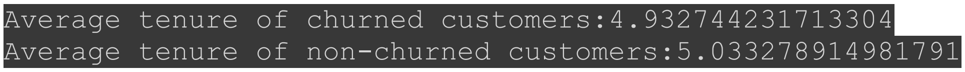
b) Compare the financial characteristics of churned vs. non-churned customers.



```
Average Credit Score of churned customers: 645.3514972999509
Average Credit Score of non-churned customers: 651.8531960316463
```



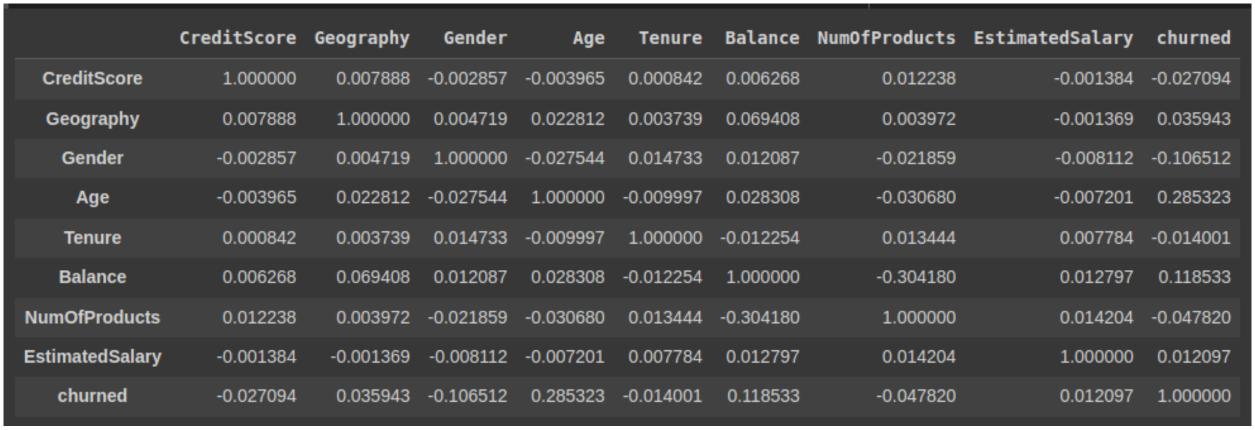
```
Average of Estimated Salary of churned customers:101465.67753068237
Average of Estimated Salary of non-churned customers:99738.39177194
```



```
Average tenure of churned customers:4.932744231713304
Average tenure of non-churned customers:5.033278914981791
```

5. Predictive Modeling:

a) Which factors are the most significant predictors of customer churn?



	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	EstimatedSalary	churned
CreditScore	1.000000	0.007888	-0.002857	-0.003965	0.000842	0.006268	0.012238	-0.001384	-0.027094
Geography	0.007888	1.000000	0.004719	0.022812	0.003739	0.069408	0.003972	-0.001369	0.035943
Gender	-0.002857	0.004719	1.000000	-0.027544	0.014733	0.012087	-0.021859	-0.008112	-0.106512
Age	-0.003965	0.022812	-0.027544	1.000000	-0.009997	0.028308	-0.030680	-0.007201	0.285323
Tenure	0.000842	0.003739	0.014733	-0.009997	1.000000	-0.012254	0.013444	0.007784	-0.014001
Balance	0.006268	0.069408	0.012087	0.028308	-0.012254	1.000000	-0.304180	0.012797	0.118533
NumOfProducts	0.012238	0.003972	-0.021859	-0.030680	0.013444	-0.304180	1.000000	0.014204	-0.047820
EstimatedSalary	-0.001384	-0.001369	-0.008112	-0.007201	0.007784	0.012797	0.014204	1.000000	0.012097
churned	-0.027094	0.035943	-0.106512	0.285323	-0.014001	0.118533	-0.047820	0.012097	1.000000

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
```

b) 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.