**OneBharat: Assignment for DS Interns hiring**

We have 3 datasets for analyzing:

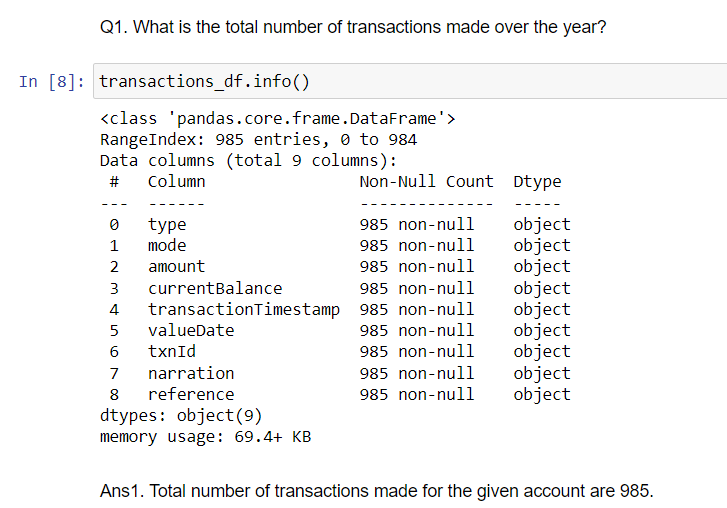
1. P1-BankStatemnets
2. P2-OfficeSupplies Data
3. P3-Churn-Modelling Data

**Bank Statements**

**Transaction Analysis:**

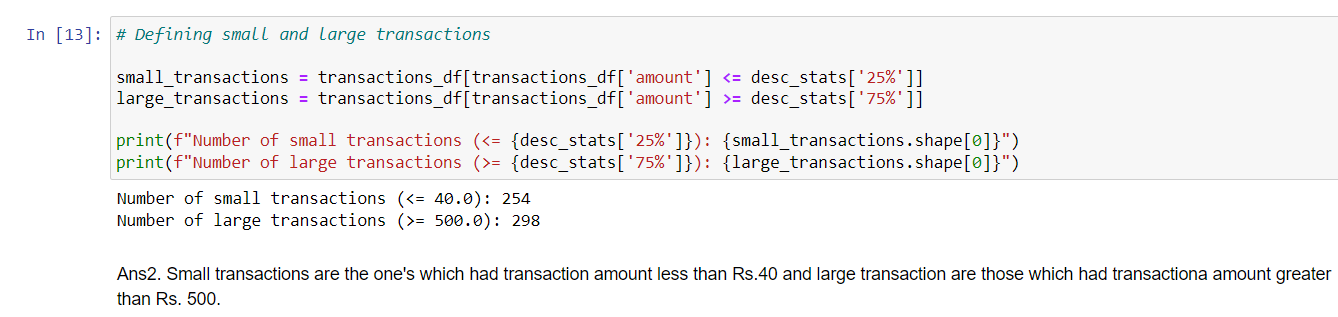
**Total Number of Transactions:**

To determine the total number of transactions made over the year, we counted the number of entries in the dataset. The total number of transactions is:

****

**Distribution of Transaction Amounts:**

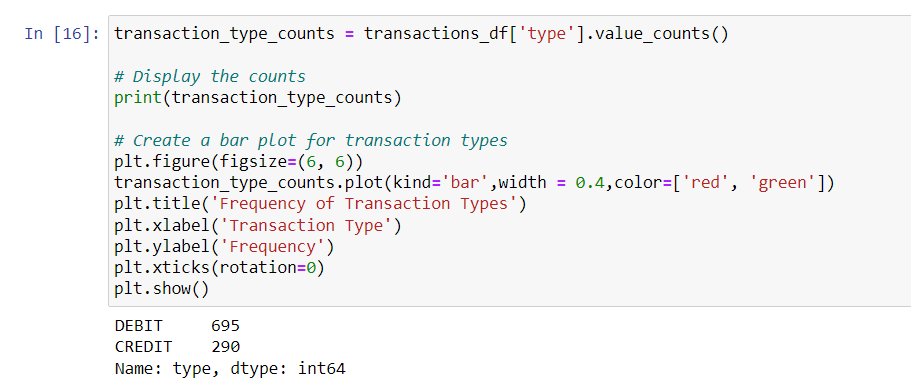
To analyse the distribution of transaction amounts, we categorized transactions into 'small' and 'large' based on their values. Small transactions were defined as those below the 25th percentile, and large transactions as those above the 75th percentile. Here are the findings:

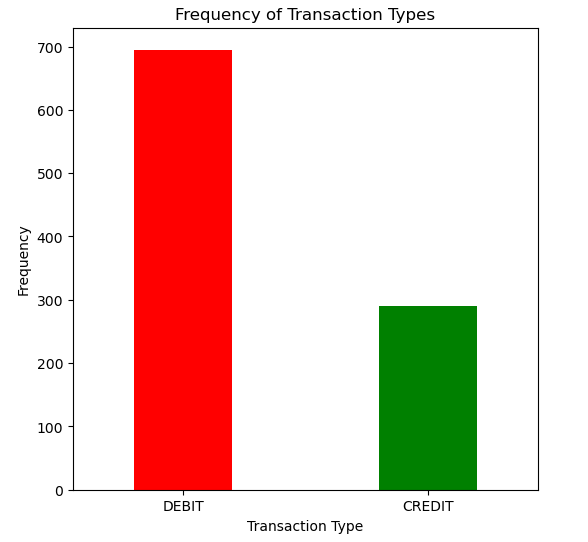
****

Small transactions are the one's which had transaction amount less than Rs.40 and large transaction are those which had transaction amount greater than Rs.500.

##### Frequency of Transaction Types

The dataset contained both debit and credit transactions. The frequency of each transaction type was calculated using:





The results indicated:

* **Debit Transactions**: 695
* **Credit Transactions**: 290

**Balance Analysis:**

**Trend of account balance over time:**

This shows the account balance after the last transaction of each month. As per the figure 1, account balance reached all-time high in the month of December while the usual trend showed a gradual growth in the account balance.

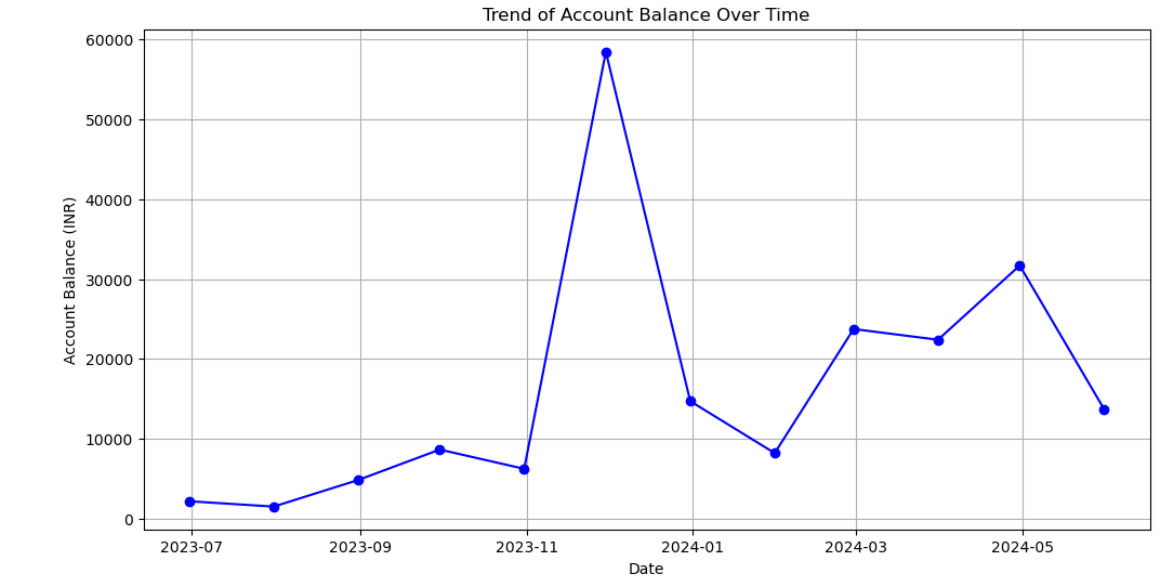
****

Figure 1: Trend of Account Balance over Time

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

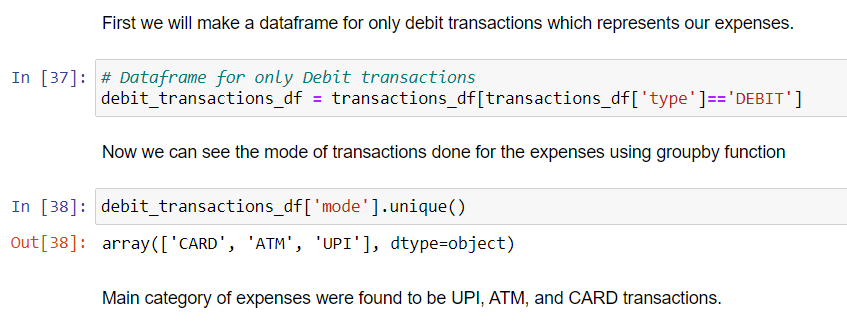
##### The account balance graph showed that that there was significant rise in account balance in the month of December as compared to other months. However, it dipped right drastically in the following month in January showing unusual transactions.

##### Again in the month of May there was a significant dip in the current balance showing a big debit transaction was made from the account.

**Spending Patterns:**

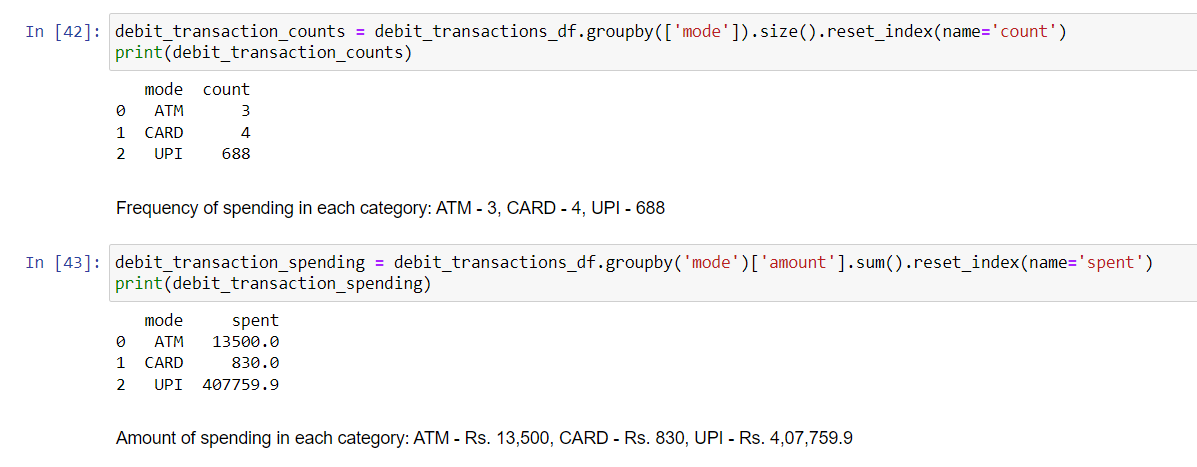
##### Main Categories of Expenses

Expenses were analysed by categorizing transactions into different types such as fuel, e-commerce, food, shopping, ATM withdrawals, and UPI transactions. This was achieved by filtering debit transactions and using:



##### Frequency and Amount of Spending in Each Category

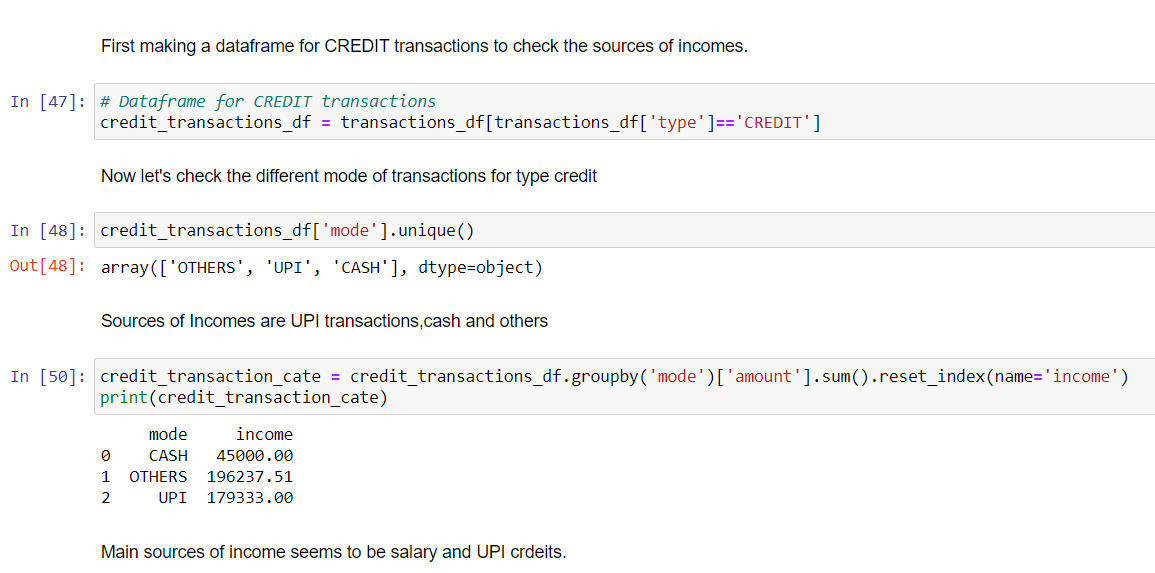
Each expense category was analysed to determine the frequency and total amount spent. This provided insights into spending behaviour and highlighted the most significant areas of expenditure.



**Income Analysis:**

##### Main Sources of Income

Income sources were identified by filtering credit transactions and using:



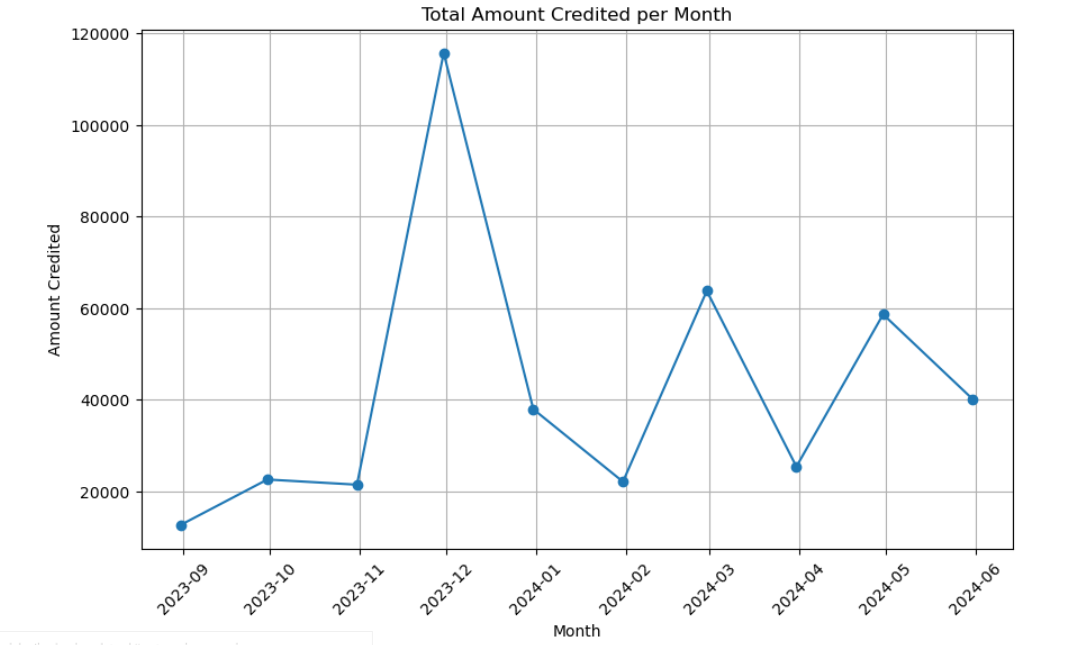
Main sources of income were found to be UPI transactions, cash and other payments (most likely salary transaction).

This analysis revealed the primary sources of income, such as salary and UPI credits.

**Patterns in the timing and amount of income received.**

This shows the monthly generated income from various sources over the period of 10 months.

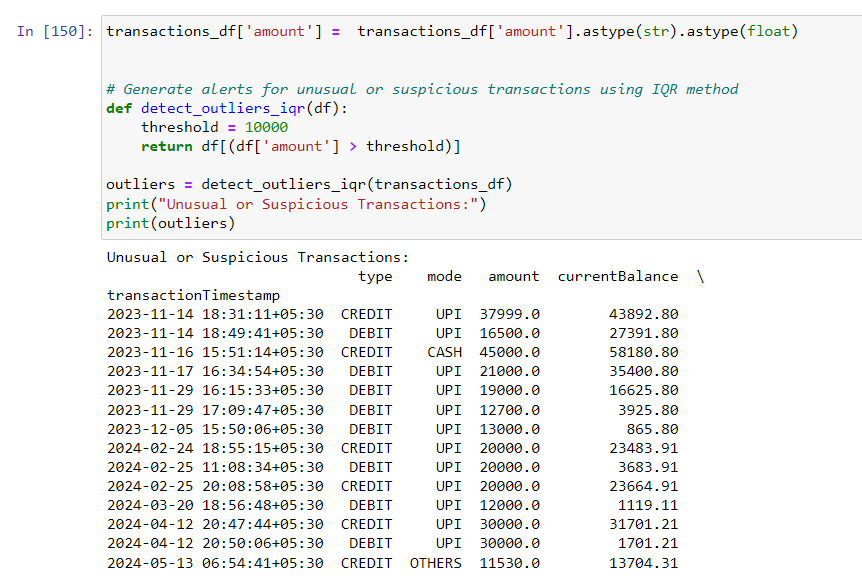
Highest income generated or amount credited to the account was in the month of December where more than 1Lakh was credited to the bank account.



**5. Alert Generation**

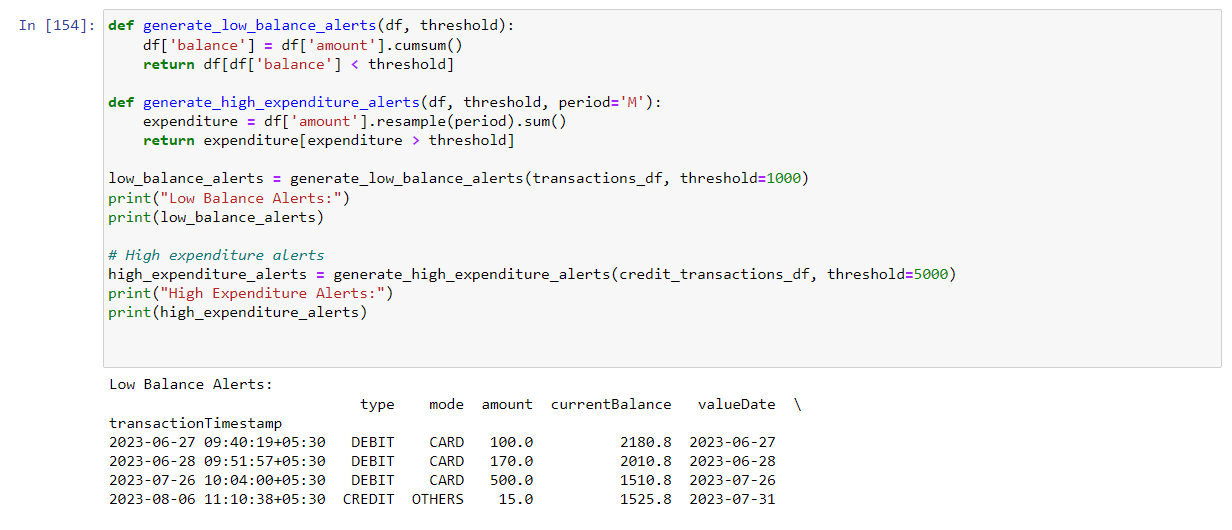
##### Unusual or Suspicious Transactions

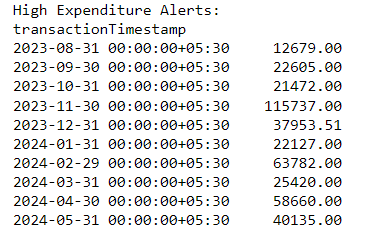
Unusual transactions were identified using outliers in transaction amounts. Transactions that significantly deviated from the average were flagged as potentially suspicious.



##### Alerts for Low Balance or High Expenditure Periods

Alerts were generated for periods with low account balance or high expenditure. Thresholds were set for low balance and high expenditure, and any periods crossing these thresholds were marked for alerts.





**Office Supplies Data**

**Sales Analysis:**

##### Total sales for each product category

##### 

##### Total sales in each category is represented in the above figure.

##### Product category with the highest sales

##### Binder had the highest sale across all products with all region combined.

##### There are only 5 products in the dataset naming ‘Blinder’, ‘Desk’, ‘Pen’, ‘Pen Set’, and ‘Pencil’.

#### **Customer Analysis:**

**Top 10 customers by sales**

##### 

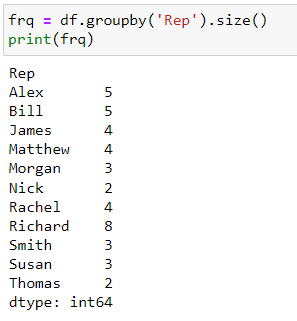
##### Top 10 customers by sales is represented by the figure. Matthew had the highest sale with 3109.44 units spending to buy various products.

**Unique Customers:**

##### 

##### There 11 unique customers in our dataset.

**Customer Purchase Frequency**

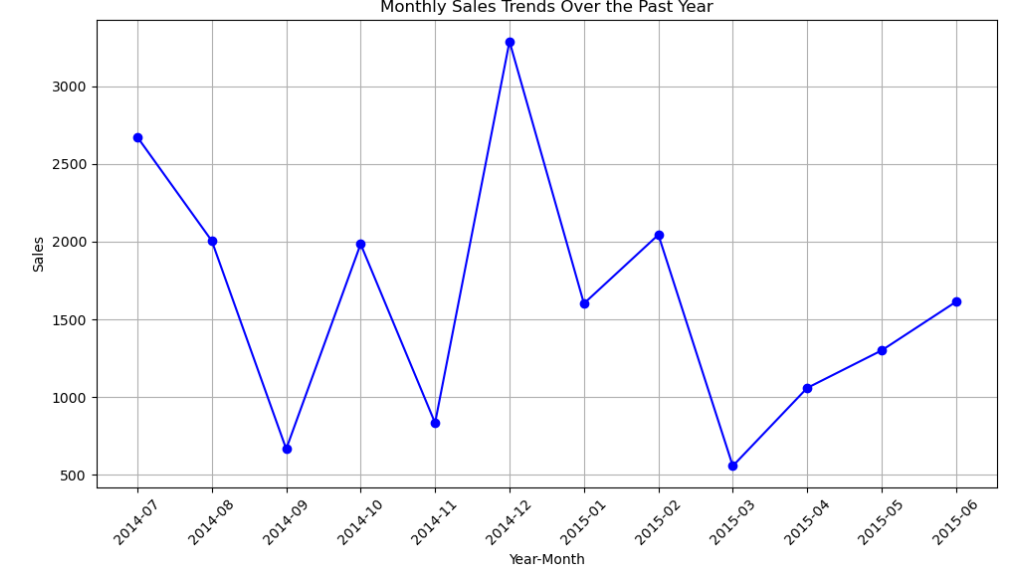
****

The figure shows the frequency of each customer purchasing any product.

Richard had the highest number of purchases i.e. 8 times, while Thomas and Nick had the lowest both making a purchase for only 2 times.

#### **Time Series Analysis:**

**Monthly sales trends over the past year/ seasonal patterns**



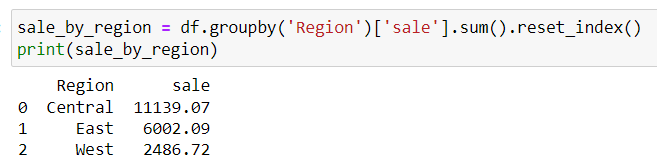
The figure above shows the monthly sales trends over the past year.

We can say that in the month of December the sales hit all time high followed by in the month of June, February and October which.

March had the lowest sales as per the data.

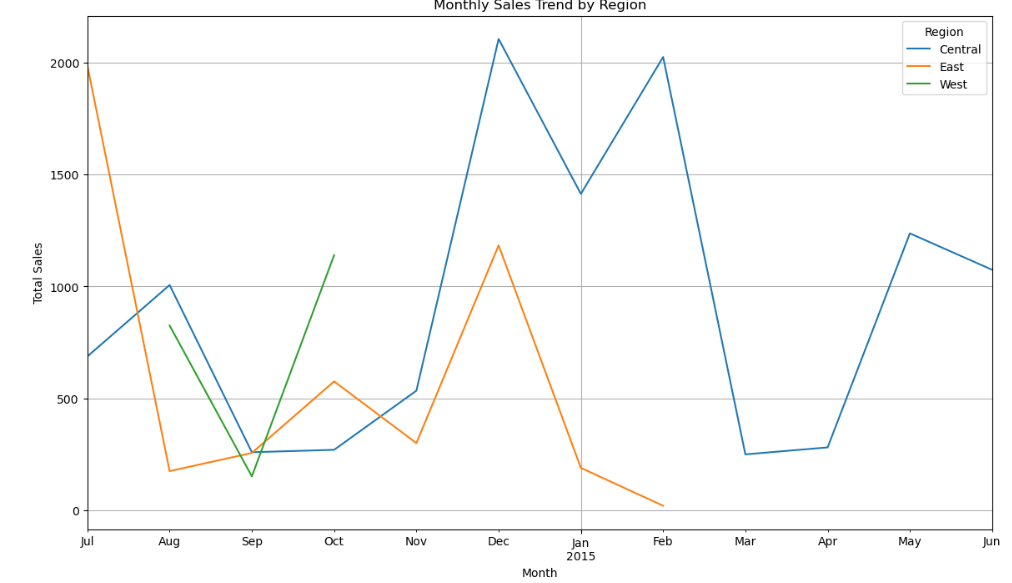
#### **Geographical Analysis:**

**Region generating the most sales**



Central region made the highest sales.

**Sales trends across different regions**



The above figure shows the monthly sales trend across different regions.

Central region had continuous sale over the year and accounts to highest sales made within the year.

West region made the lowest sale which were only seen in the month of August, September and October only.

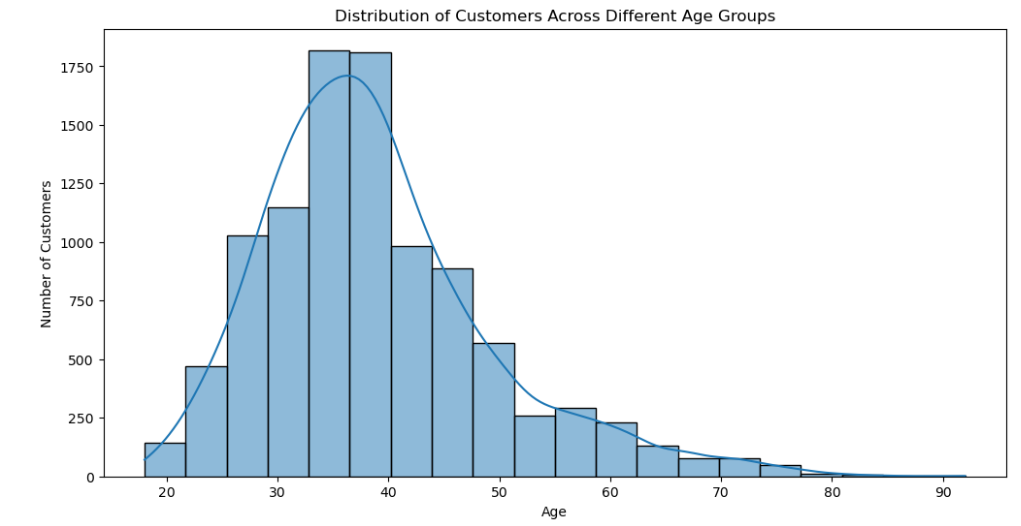
As for the East region it seems the sales gradually dropped and eventually no sales were made after February 2015.

**There was no data for selling price of any item to calculate the profits generated for the products or do the analysis for profits.**

**Churn Modelling**

#### **Customer Demographics:**

**Distribution of customers across different age groups**

****

|  |  |
| --- | --- |
| **Age (Years)** | **No. of Customers** |
| 0-20 | 89 |
| 20-40 | 6330 |
| 40-60 | 3117 |
| 60-80 | 452 |
| 80-100 | 12 |

This shows the customers distribution in different age groups. Highest customer purchases were made among the adults between the ages 35 – 40.

**Gender distribution of customers**

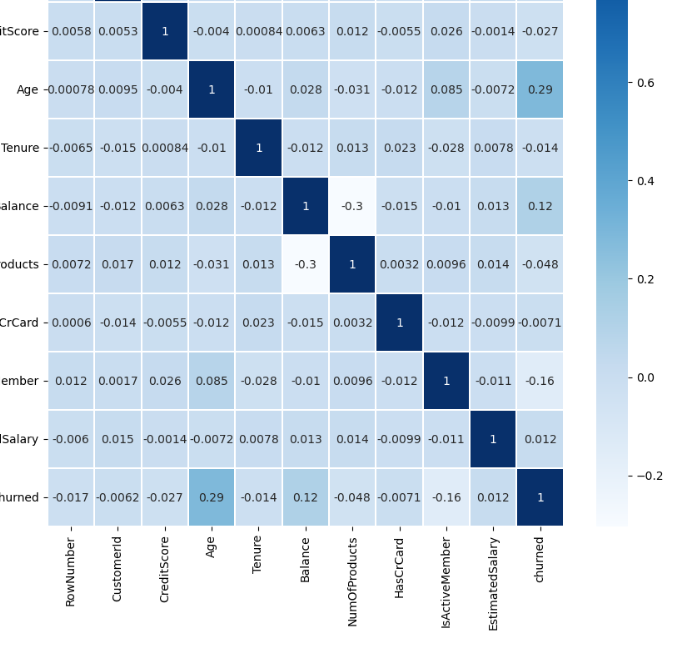
There are more Male customers than female. The distribution over male and female came out to be 55% and 45% respectively.

**Churn Analysis:**

**Percentage of customers who have churned**

80% of customers of the customers were churned which accounts to 7963 of the total 10000 customers were churned.

**Reasons for customer churn**

****

As per the correlation between the variables:

Age has a moderate positive correlation with churn.

Active membership has a negative correlation with churn, suggesting that active members are less likely to churn.

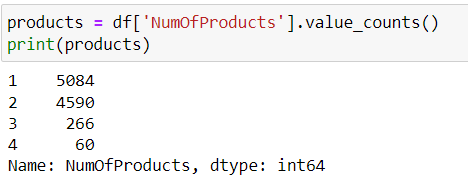
Balance and number of products purchased or used show low correlation with churn, indicating these might be less influential.

Estimated salary of Churned customers was found out to be ‘101465.68 INR’.

Active members are less likely to churn as per the data. 2037 active customers have been churned which accounts to 20% of the total customers.

**Product Usage:**

**Most commonly used products or services**

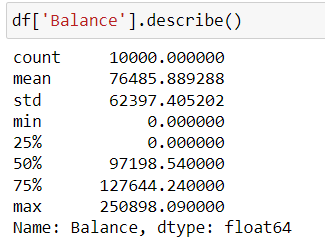
****

**While the product names are not given in the dataset.**

However we can identify that the mostly only one or two products are used by the customers.

**Financial Analysis:**

Average account balance of customers came out to be 76485.89 INR.



**Financial characteristics of churned vs. non-churned customers**

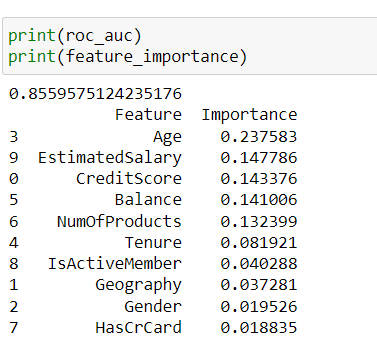
|  |  |  |
| --- | --- | --- |
| **Characteristics** | **Churned** | **Non-Churned** |
| **Average Estimated Salary (INR)** | 101465.68 | 99738.39 |
| **Average Account Balance (INR)** | 91108.54 | 72745.29 |

**Predictive Modelling:**

**Most significant predictors of customer churn:**

Here we have done the feature importance of the factors affecting the prediction of a customer being churned or not.

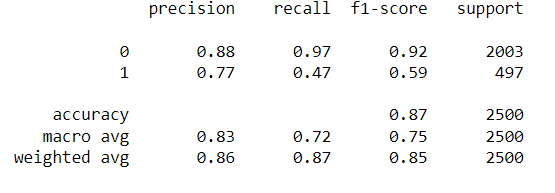
This was done by using feature\_importances method from the sklearn library.

****

* Age is the most significant predictor of customer churn.
* Other important features include Estimated Salary, Credit Score, Balance, and Number of Products.

**Predictive model to identify at-risk customers (Churned):**

Here we go with the Random Forest Classification as we only want to predict if the customer will be at risk or not depending on the financial and other factors.



This was the result of our RFC model giving us the accuracy of 87%.

For the model we did a split of 75% and 25% for training and testing our datasets.

A higher f1 score for predicting non-churned customers shows that model predicts correctly if the customer is not at risk.

While the Precision in predicting customers at risk has also better outcome.