

**DATA SCIENCE TOOLBOX PYTHON PROGRAMMING**

**PROJECT REPORT**

(Project Semester January-April 2025)

**Customer Churn Analysis**

**Submitted by**

**NAME:** B Devendranadh Venkat

**REGISTRATION NO:** 12301888

**SECTION:** K23GN62

**ROLL NO:** 62

**COURSE CODE** **:** INT375

Under the Guidance of

**MRS. AASHIMA (UID: 28968)**

**Discipline of CSE/IT**

**Lovely School of Computer Science**

**Lovely Professional University, Phagwara**

**CERTIFICATE**

This is to certify that **B Devendranadh Venkat** bearing Registration no. **12301888** has completed **INT375** project titled, **“Customer Churn Analysis”** under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

**DECLARATION**

I, B Devendranadh Venkat student of CSE (Program name) under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: 11-04-2025

Registration No: 12301888 Name of the student: B Devendranadh Venkat

**Table of Contents**

|  |
| --- |
| 1. **Introduction** |
| 1. **What is EDA?** |
| 1. **Why EDA is Important for Churn Analysis** |
| 1. **Source of Dataset** |
| 1. **Step-by-Step EDA Process** |
| 1. **Dataset Preprocessing** |
| 1. **Univariate Analysis** |
| 1. **Bivariate Analysis** |
| 1. **Correlation Analysis** |
| 1. **Analysis on Dataset** |
| 1. **Conclusion** |
| 1. **Future Scope** |
| 1. **References** |

**1. Introduction**

In today’s competitive business environment, customer retention is a critical success factor. One of the major challenges faced by service-oriented businesses, especially in the banking sector, is customer churn. Customer churn refers to when a customer stops doing business with a company. High churn rates can indicate dissatisfaction with services and lead to financial losses.

This project focuses on analyzing a bank customer dataset to identify the key factors that contribute to customer churn. Using exploratory data analysis (EDA), we investigate demographic, behavioral, and financial variables to discover patterns and gain actionable insights. The objective is to help the bank better understand its customers and develop data-driven strategies to reduce churn.

**2. What is EDA?**

Exploratory Data Analysis (EDA) is the process of analyzing data sets by summarizing their main characteristics. It involves visual methods and statistical techniques that help in discovering trends, patterns, relationships, and anomalies in the dataset. EDA is a crucial step in the data analysis process because it allows analysts to:

* Better understand the structure of the dataset
* Detect missing or incorrect data
* Identify outliers
* Find relationships between variables

EDA provides the foundation for model building and helps in making informed decisions.

**3. Why EDA is Important for Churn Analysis**

Understanding why customers leave is essential for developing effective retention strategies. EDA helps break down the raw data to reveal patterns related to churn behavior. In this project, EDA allows us to:

* Quantify the churn rate among customers
* Detect customer segments with high churn probability
* Assess the impact of variables like age, balance, credit score, etc.
* Use visualizations to make complex data more interpretable

By conducting EDA, we can move beyond assumptions and base our decisions on real evidence.

**4. Source of Dataset**

The dataset used in this analysis was sourced from a public dataset often used in churn prediction projects. The data simulates a real-world scenario in a European bank. Each row represents a customer with several features such as credit score, geography, tenure, and balance.

* File Name: Bank\_Churn.csv
* Format: CSV (Comma Separated Values)
* Features:
  + CreditScore
  + Geography
  + Gender
  + Age
  + Tenure
  + Balance
  + NumOfProducts
  + HasCrCard
  + IsActiveMember
  + EstimatedSalary
  + Exited (target variable: 1 = churned, 0 = stayed)

**5. Step-by-Step EDA Process**

1. Import Libraries (Pandas, NumPy, Matplotlib, Seaborn, Scipy)
2. Load the dataset
3. Inspect structure and check for missing values
4. Clean and preprocess the data
5. Visualize each feature with respect to churn
6. Analyze correlations

**6. Dataset Preprocessing**

Before diving into the analysis, we cleaned and prepared the dataset. The key preprocessing steps included:

* Checking for missing values using df.isnull().sum() which showed none
* Ensuring column names were formatted properly (no trailing spaces)
* Converting data types where required (e.g., integers to floats if needed)
* Removing duplicate entries if found (none in this case)
* Understanding the balance between churned and retained customers

**7. Univariate Analysis**

Univariate analysis involves examining one variable at a time. For our dataset, we started with:

* **Churn Rate (Exited column):** We plotted the number of customers who churned vs. who stayed. A pie chart also provided a percentage view.
* **Gender:** We checked the gender distribution and how it relates to churn.
* **Geography:** We looked at churn rates across different countries.
* **Age:** Age distribution helped us understand which age groups are most prone to churn.

**8. Bivariate Analysis**

In bivariate analysis, we explore the relationship between two variables. For churn prediction, the most useful bivariate comparisons were:

* **Tenure vs. Churn:** Box plots helped us see if long-term customers were more loyal.
* **Number of Products vs. Churn:** Customers with fewer products had a higher chance of leaving.
* **IsActiveMember vs. Churn:** Inactive members were more likely to churn.

**9. Correlation Analysis**

Correlation analysis helps us understand how numerical features move in relation to each other. We used:

* **Heatmap:** Shows correlation coefficients for multiple variables at once
* **Pearson Correlation:** Specifically used to check the relationship between Credit Score and Churn

The heatmap revealed that age and balance had some correlation with churn, while credit score and estimated salary had weaker relationships.

# Correlation Heatmap

corr=df[['CreditScore','Age','Tenure','Balance','NumOfProducts','EstimatedSalary','Exited']].corr()

sns.heatmap(corr, annot=True, cmap='coolwarm')

plt.title("Correlation Matrix")

plt.show()

**Heatmap:**

**A diagram of a heatmap

AI-generated content may be incorrect.**

**10. Analysis on Dataset**

**Objective 1: Customer Churn Overview**

• Most customers stayed with the bank (Exited = 0).

• Pie chart shows approx. 80% stayed, 20% churned.

**Code:**

# Bar chart

sns.countplot(x='Exited', data=df,color='brown')

plt.title('Customer Churn Distribution')

plt.show()

# Pie chart

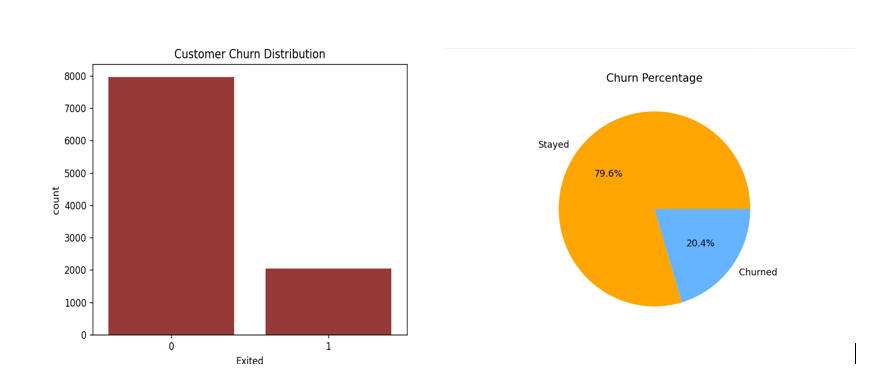
churn\_counts = df['Exited'].value\_counts()

plt.pie(churn\_counts,labels=['Stayed','Churned'], autopct='%1.1f%%',colors=['orange','#66b3ff'])

plt.title('Churn Percentage')

plt.show()

**Graph:**

****

**Objective 2: Demographic Impact Analysis**

• Gender: Slightly more males churned

• Geography: Customers from Germany had a higher churn rate

• Age: Older customers (40+) churn more

**Code:**

sns.countplot(x='Gender', hue='Exited', data=df)

plt.title("Gender vs Churn")

plt.show()

sns.countplot(x='Geography', hue='Exited', data=df)

plt.title("Geography vs Churn")

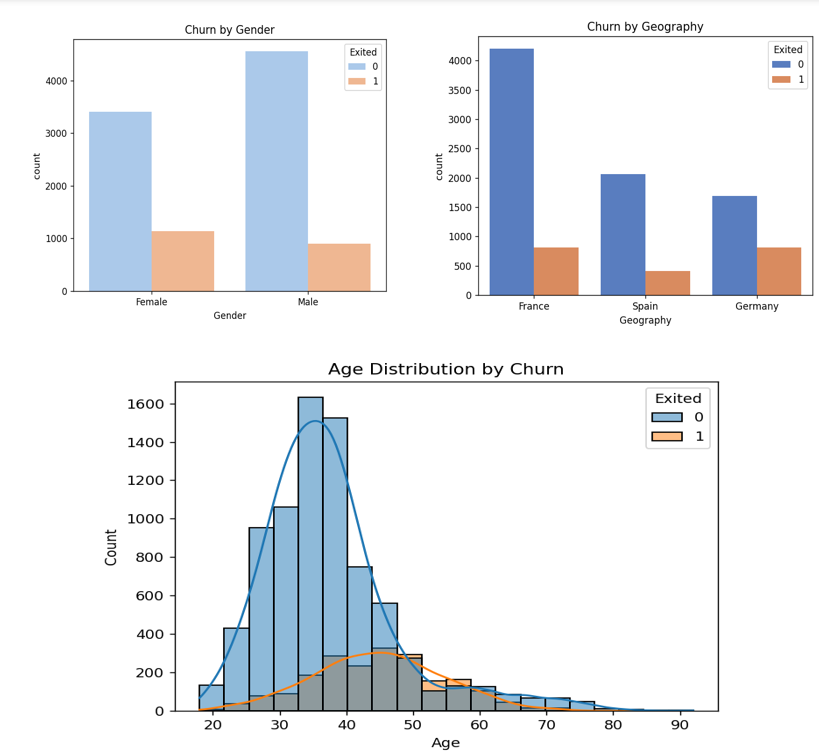
plt.show()

sns.histplot(data=df, x='Age', hue='Exited', bins=20, kde=True)

plt.title("Age Distribution by Churn")

plt.show()

**Graph:**



**Objective 3: Customer Behavior Insights**

• Lower Tenure and fewer Products increase churn risk

• Inactive members churn more

**Code:**

sns.boxplot(x='Exited', y='Tenure', data=df)

plt.title("Tenure vs Churn")

plt.show()

sns.countplot(x='NumOfProducts', hue='Exited', data=df)

plt.title("Number of Products vs Churn")

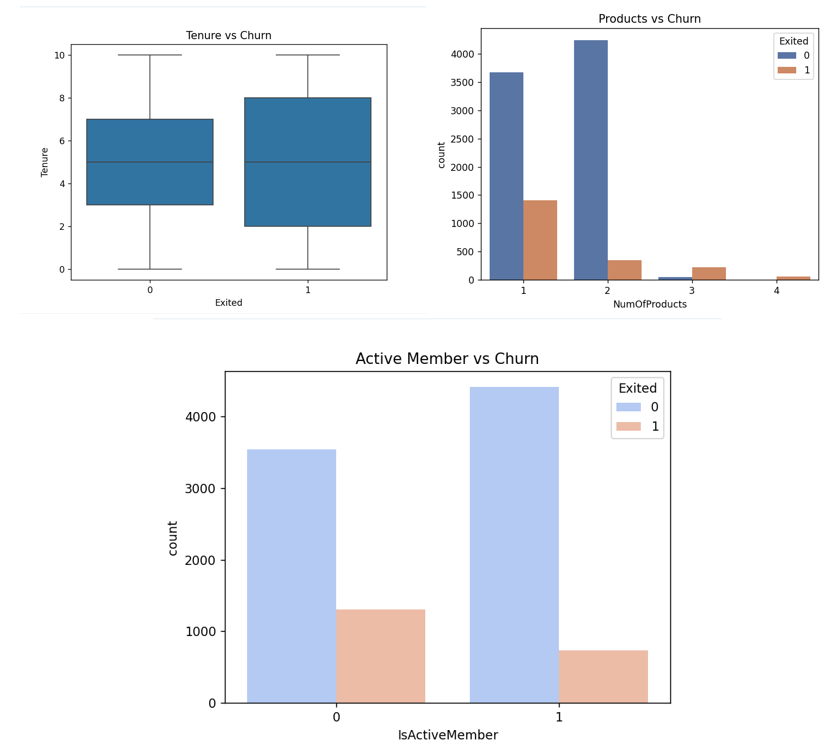
plt.show()

sns.countplot(x='IsActiveMember', hue='Exited', data=df)

plt.title("Activity Status vs Churn")

plt.show()

**Graph:**



**Objective 4: Credit Score and Churn**

• Weak negative correlation found between credit score and churn

Pearson Correlation: -0.16

**Code:**

sns.boxplot(x='Exited', y='CreditScore', data=df)

plt.title("Credit Score vs Churn")

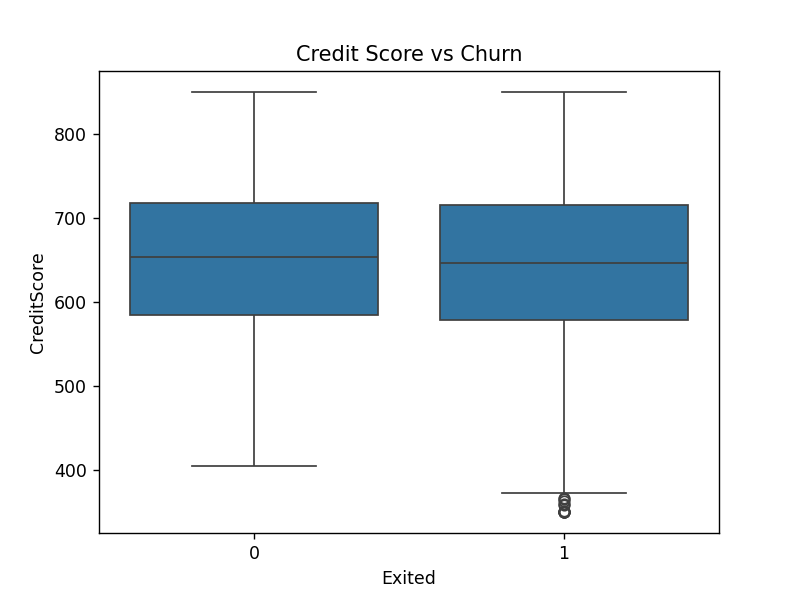
plt.show()

from scipy.stats import pearsonr

credit\_corr, \_ = pearsonr(df['CreditScore'], df['Exited'])

print(f"Pearson Correlation between Credit Score and Churn: {credit\_corr:.2f}")

**Graph:**



**Objective 5: Income & Churn Relationship**

• Salary shows little impact on churn

• Balance is higher in churned customers

**Code:**

sns.boxplot(x='Exited', y='EstimatedSalary', data=df)

plt.title("Estimated Salary vs Churn")

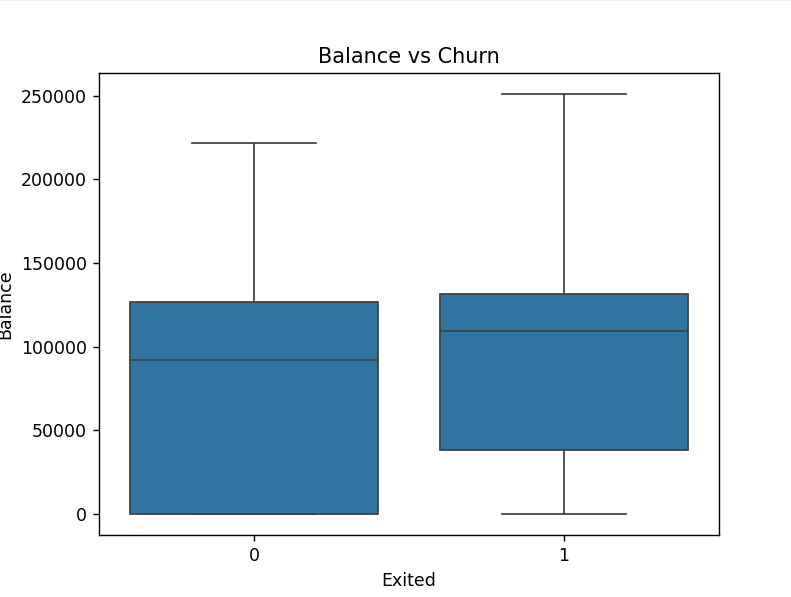
plt.show()

sns.boxplot(x='Exited', y='Balance', data=df)

plt.title("Balance vs Churn")

plt.show()

**Graph:**

****

A chart with blue squares

AI-generated content may be incorrect.

**11. Conclusion**

Through exploratory data analysis, we identified various factors that influence customer churn in the banking sector. Key findings include:

* High churn among customers aged 40+
* More churn in Germany than other regions
* Behavior-based indicators like tenure, number of products, and activity are strong churn predictors

By understanding these insights, banks can build customer retention strategies, target vulnerable segments, and improve service quality.

**12. Future Scope**

While EDA provides valuable insights, future steps can enhance this analysis:

* Implement machine learning models (e.g., logistic regression, decision trees) for churn prediction
* Introduce new features like transaction frequency or customer feedback
* Design targeted campaigns to reduce churn and improve satisfaction
* Use clustering to segment customers into risk categories

**13. References**

* Dataset: Bank\_Churn.csv
* Tools Used: Python, Pandas, NumPy, Seaborn, Matplotlib, Scipy
* Concepts: Exploratory Data Analysis, Pearson Correlation, Visualization Techniques