**Mathematics For Computer Science**

**UE23MA242A**

**Teaching Assistants: Shubhika Verma, Abdullah Khan, Shruti Singh**

Bank Customer Churn Analysis Case Study

**Introduction**

The sections below, **Background** and **Case Study**, provide context for this data science assignment. This exercise will allow you to explore a real-world dataset using Python and analyze various features through descriptive statistics, tables, and visualizations. Happy coding!

**Background**

The provided dataset contains detailed information about bank customers, including demographic data, credit card usage patterns, and account information. It also indicates whether a customer has churned (i.e., closed their account) or not. By analyzing this data, financial analysts and bank managers can gain insights into customer behavior, identify factors contributing to churn, and develop strategies to improve customer retention.

**Case Study**

A researcher has acquired a dataset of bank customers and seeks to perform statistical analysis to explore various factors that influence customer churn. The goal is to uncover relationships among demographic variables, credit card usage patterns, and churn status.

**Dataset Description**

The dataset contains the following variables:

1. **CLIENTNUM**: Client number. Unique identifier for the customer holding the account.
2. **Attrition\_Flag**: Whether the customer has churned ('Attrited Customer') or is an existing customer ('Existing Customer').
3. **Customer\_Age**: Age in years of the customer.
4. **Gender**: Gender of the customer (M = Male, F = Female).
5. **Dependent\_count**: Number of dependents.
6. **Education\_Level**: Educational qualification of the customer (e.g., High School, Graduate, Uneducated).
7. **Marital\_Status**: Marital status of the customer (e.g., Married, Single, Divorced).
8. **Income\_Category**: Annual income category of the customer (e.g., Less than $40K, $40K - $60K, $60K - $80K).
9. **Card\_Category**: Type of credit card (e.g., Blue, Silver, Gold, Platinum).
10. **Months\_on\_book**: Total months the customer has been with the bank.
11. **Total\_Relationship\_Count**: Total number of products held by the customer.
12. **Months\_Inactive\_12\_mon**: Number of months inactive in the last 12 months.
13. **Contacts\_Count\_12\_mon**: Number of contacts in the last 12 months.
14. **Credit\_Limit**: Credit limit on the credit card.
15. **Total\_Revolving\_Bal**: Total revolving balance on the credit card.
16. **Avg\_Open\_To\_Buy**: Open to buy credit line (average of last 12 months).
17. **Total\_Amt\_Chng\_Q4\_Q1**: Change in transaction amount (Q4 over Q1).
18. **Total\_Trans\_Amt**: Total transaction amount in the last 12 months.
19. **Total\_Trans\_Ct**: Total transaction count in the last 12 months.
20. **Total\_Ct\_Chng\_Q4\_Q1**: Change in transaction count (Q4 over Q1).
21. **Avg\_Utilization\_Ratio**: Average card utilization ratio.

**Problem Set**

Solve the below given questions. Bonus questions, if any can be solved if you have additional time left but are not mandatory. Questions that aren’t applicable to the dataset provided, or questions that yield irregular output maybe omitted/the irregular output will be accepted.

**Unit 1: Data Understanding and Preprocessing**

1. **Feature Classification**:
   * Divide the features in the BankChurners dataset into appropriate data types (ordinal, nominal, interval, or ratio) and provide justification for each classification choice.
2. **Data Quality Issues**:
   * Identify and describe potential data issues, inconsistencies, or anomalies within the dataset. Describe your approach to cleaning and preprocessing the data to ensure accuracy and readiness for analysis.
3. **Summary Statistics**:
   * Provide a summary of each feature’s central tendency and dispersion. Describe the best measures of central tendency for each attribute, and calculate standard deviation and range for all numerical columns.
4. **Visualizations**:
   * Plot both histograms and box plots for **'Customer\_Age'** and **'Credit\_Limit'**.
     + **i)** Identify the type of distribution each variable appears to have.
       - *Hint: Scale the visualizations to clearly depict the data distribution.*
     + **ii)** Determine how many outliers exist for each variable.
     + **iii)** *Bonus*: After identifying the distribution types, make adjustments and re-plot **'Credit\_Limit'** if necessary to better capture observed patterns.
5. **Outlier Treatment**:
   * Describe the steps you would take to adjust for outliers. Display the modified data with new box plots and histograms.
6. **Normal Probability Plot (Q-Q Plot)**:
   * Create a Q-Q plot for **'Credit\_Limit'** and describe what the plot reveals about the distribution.
7. **Correlation Analysis**:
   * Calculate the correlation of **'Customer\_Age'** with other numeric variables and identify the variable most strongly correlated with **'Credit\_Limit'**.
8. **Pair Plot Analysis**:
   * Randomly sample 5,000 records and create a pair plot for **'Customer\_Age'** and **'Credit\_Limit'** using **'Attrition\_Flag'** as the hue.

**Unit 2: Hypothesis Testing**

1. **Hypothesis Testing**:
   * Develop null and alternative hypotheses to test whether the median **'Credit\_Limit'** differs significantly between customers who have churned and those who haven’t. Use the Mann-Whitney U test to evaluate this difference.
2. **Margin of Error Calculation**:
   * Calculate the margin of error to assess the precision of the hypothesis test conducted earlier.

**Unit 3: Prediction and Feature Engineering**

1. **Linear Regression Analysis**:
   * Use **'Customer\_Age'** and **'Total\_Trans\_Amt'** to build a linear regression model predicting **'Credit\_Limit'**.
2. **Feature Engineering**:
   * Given **'Customer\_Age'**, **'Credit\_Limit'**, and **'Total\_Trans\_Amt'**, propose two additional features that could improve the prediction of **'Credit\_Limit'**.