MTH516: Project Plan Report

Group 15

Name	Roll Number
Divya Gupta	210353
Kumar Kanishk Singh	210544
Kundan Kumar	210547
Soni Verma	211051
Vala Yash	211142

1 Introduction

This provides an overview of the dataset under analysis, with a detailed explanation of the variables involved. Additionally, we outline the non-parametric tests we plan to use in the upcoming analysis to derive meaningful inferences from the dataset. Non-parametric tests are chosen as they do not assume any specific distribution for the data and are robust for handling non-normally distributed datasets.

2 Dataset Overview

The dataset can be accessed using the following link: **Dataset Link**

The dataset contains information on customers of a financial institution. It includes demographic, financial, and lifestyle data. The aim of the analysis is to investigate patterns and relationships between these variables, with a focus on credit status, income, and other socio-economic factors. The key variables in the dataset are as follows:

- `ID`: A unique identifier for each customer.
- `CREDITSTATUS`: Credit score status, indicating the creditworthiness of a customer (e.g., "0", "C", "X").
- `GENDER`: Gender of the customer (e.g., "M" for Male, "F" for Female).
- `OWNCAR`: Whether the customer owns a car (Y/N).
- `OWNPROPERTY`: Whether the customer owns property (Y/N).
- `CHILDRENCOUNT`: Number of children the customer has.
- `INCOMETOTAL`: The total income of the customer in the specified currency (e.g., Rupees).
- `INCOMETYPE`: The type of income (e.g., "Working", "Commercial associate", "Pensioner").

- `EDUCATIONLEVEL`: The highest level of education attained (e.g., "Higher education", "Secondary education").
- `MARITALSTATUS`: The marital status of the customer (e.g., "Married", "Single").
- `HOUSINGTYPE`: Type of housing the customer lives in (e.g., "House/apartment").
- `MOBILE`: Whether the customer owns a mobile phone (1 = Yes, 0 = No).
- `EMAIL`: Whether the customer owns an email account (1 = Yes, 0 = No).
- `OCCUPATION`: The occupation of the customer (e.g., "Laborers", "Managers").
- `FAMSIZE`: The size of the customer's family.

3 Descriptive Analysis on Data

- **Summary Statistics:** Provide mean, median, standard deviation for continuous variables like `INCOMETOTAL`, `CHILDRENCOUNT`.
- **Categorical Variable Distribution:** Show frequency distribution for variables like `GENDER`, `CREDITSTATUS`, `OWNCAR` using bar/pie charts.
- **Missing Data Analysis:** Identify missing values, their patterns, and how to handle them (e.g., imputation or exclusion).
- **Outlier Detection:** Use box plots to detect outliers in continuous variables like `INCOMETOTAL`.
- **Bivariate Analysis:** Explore relationships between key variables (e.g., correlation heatmap for continuous variables, cross-tabulation for categorical ones).
- **Visualizations:** Include histograms, box plots, and bar charts to visualize data distributions and group comparisons.

4 Non-parametric Tests

The following are some of non-parametric tests which will be used to analyze the dataset

4.1 Run Test

The Run Test will be used to determine if a sequence is random or if there is a pattern in the order of the values. For instance, we will use this test to check whether the sequence of customer credit statuses is randomly ordered or shows a non-random pattern.

- **Hypothesis**:
  - Null Hypothesis (H0): The sequence is random.
  - Alternative Hypothesis (H1): The sequence is not random.

This test can also be applied to check randomness in the order of binary variables like OWNCAR (Y/N), or sequences of credit status over time.

4.2 Chi-square Test of Independence

The Chi-Square Goodness of Fit test will be used to determine whether the observed distribution of a categorical variable significantly differs from an expected distribution.

- **Hypothesis**:
  - Null Hypothesis (H_0): The observed distribution matches the expected distribution
  - Alternative Hypothesis (H_1): The observed distribution does not match the expected distribution

This test can be applied to check if the distribution of categories such as `GENDER`, `CREDITSTATUS`, or `MARITALSTATUS` follows an expected proportion.

4.3 Wilcoxon Signed-Rank Test

The Wilcoxon Signed-Rank Test will be used to compare two related samples or repeated measurements on a single sample, for example, comparing customer income before and after a specific policy change or intervention. It is a non-parametric alternative to the paired t-test.

- **Hypothesis:**
  - Null Hypothesis (H0):There is no difference between the paired samples.
  - Alternative Hypothesis (H1): There is a difference between the paired samples.

4.4 Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov (K-S) test will be used to compare the observed distribution of a numeric variable against a reference distribution, or to compare two empirical distributions.

- **Hypothesis:**
  - Null Hypothesis (H0): The sample distribution follows the reference distribution (or the two sample distributions are the same).
  - Alternative Hypothesis (H1): The sample distribution does not follow the reference distribution (or the two sample distributions are different).

This test is useful for determining if income distributions align with expected financial models or if customer behavior varies significantly across different income groups.

4.5 Mann-Whitney U Test

The Mann-Whitney U test will be used to compare the income distributions (`INCOMETOTAL`) between two groups, such as customers who own a car (`OWNCAR` = Y) and those who do not (`OWNCAR` = N).

- **Hypothesis**:

- Null Hypothesis (H_0): The income distributions between the two groups are the same.
- Alternative Hypothesis (H_1): The income distributions between the two groups are different.

5 Conclusion

This provides a structured overview of the dataset and the variables involved. The non-parametric tests outlined will be performed in the subsequent analysis to uncover meaningful insights regarding relationships between customer demographics, financial standing, and credit status.