ANALYSIS OF INCOMES AND LOANS OF INDIVIDUALS ALONG WITH THEIR AGES

This Notebook is a comprehensive data analysis of a bank loan dataset. The dataset includes various attributes such as age, education level, years of employment, address, income, debt-to-income ratio, credit debt, other debt, and a default indicator. The primary goal of the Notebook is to understand the characteristics of the individuals who default on their loans and identify any patterns or correlations that might exist within the data.

The Notebook begins by installing the necessary Python libraries for the analysis, including pandas for data manipulation, matplotlib for basic plotting, and seaborn for more advanced statistical visualizations. It then reads the data from a CSV file into a pandas DataFrame, a two-dimensional, size-mutable, and heterogeneous tabular data structure that allows for flexible data manipulation.

Next, the Notebook performs data cleaning to ensure the quality and reliability of the analysis. This includes removing duplicate rows and rows with missing values, changing the data types of some columns for appropriate analysis, and recoding the 'default' column to binary values to simplify the interpretation of results.

The Notebook then describes the data by providing descriptive statistics for the 'age,' 'income,' and 'default' columns. It calculates the mean and median 'age' and 'income' for each 'default' group, providing a clear picture of these variables' central tendency and dispersion. It also calculates the correlation between 'default' and 'income,' indicating the linear relationship between these two variables.

THE NOTEBOOK CULMINATES IN THE CREATION OF THREE VISUALIZATIONS TO ANALYZE DATA:

Histogram of Average Income and Count of Defaults for Different Age Groups: This chart visually represents the average income and count of defaults across different age groups. The age groups are created by dividing the 'age' column into five equal intervals. The chart uses a bar plot to show the average income in each age group and a line plot to show the count of defaults. This allows for comparing income levels and default rates across different age groups, which can reveal whether age impacts the likelihood of defaulting.

Boxplot of Age Distribution for Default and Non-Default Cases: This chart uses a boxplot to compare the age distribution for default and non-default cases. A boxplot is a standardized way of displaying the distribution of data based on a five-

number summary (minimum, first quartile (Q1), median, third quartile (Q3), and maximum). It can provide valuable insights about age variability in default and non-default cases and whether age is a significant factor in loan default.

Bar Chart of Mean Default Rate and Count for Different Income Bins: This chart shows the mean default rate and count for different income bins. The income bins are created by dividing the 'income' column into five equal intervals. The chart uses a bar plot to show the mean default rate in each income bin and a line plot to show the count of individuals. This visualization can help identify if there's a pattern between income levels and default rates.

In conclusion, this Notebook thoroughly analyzes a bank loan dataset, exploring the characteristics of individuals who default on their loans. The combination of descriptive statistics and visualizations offers a comprehensive understanding of the data, revealing patterns and correlations that could be instrumental in predicting loan defaults.