**Loan Default Analysis**

Overview: Loan Default Analysis:

The below mentioned dataset comprises loan application data from a financial institution, where applicants have applied for various types of loans. The key focus of this analysis is to understand the factors influencing loan defaults, identify patterns among defaulters, and explore the relationships between various demographic, financial, and behavioral attributes.

The dataset includes information such as the applicant's income, credit amount, annuities, ownership status (e.g., car, real estate), and more. By applying statistical methods, you will investigate how these variables relate to the likelihood of loan default, identify significant predictors, and assess the overall financial health of the applicants.

Dataset:

Dataset Columns Description: This sheet provides a detailed explanation for

each column in the dataset, describing the purpose and meaning of each variable (e.g., "ID of loan in our sample," "Target variable") to help understand the structure and content of the data.

Objective:

As an analyst, your task is to use statistical techniques to derive insights from the dataset. This will involve:

1. Descriptive Analysis: Understanding the basic statistical properties of the data, such as central tendency and variability.

2. Probability and Sampling: Exploring the probabilities of loan default among

different segments and testing the robustness of these findings using various sampling techniques.

3. Hypothesis Testing: Conducting hypothesis tests to determine the significance of various factors, such as income levels, gender, and ownership status, on loan defaults.

4. Distribution Analysis: Examining the distribution of key financial variables to understand their spread and shape, which will guide further statistical analysis.

5. Advanced Statistical Testing: Using more sophisticated techniques such as

ANOVA and chi-square tests to uncover relationships between categorical and numerical variables.

Expected Outcomes: By the end of this analysis, you should be able to:

Identify which demographic and financial factors are most strongly associated with loan defaults.

Determine if there are significant differences between various segments of the population (e.g., based on gender or contract type).

Validate or refute claims regarding the average values of key financial variables (e.g., income, annuities).

Provide actionable insights that the financial institution can use to improve its loan approval process, better assess risk, and potentially reduce the rate of defaults.

This comprehensive statistical analysis will provide a robust foundation for understanding the drivers of loan defaults and guide the financial institution in making data-driven decisions.

Tasks to be performed:

1. Load the data into the python by reading the file in a data frame and check the head of dataframe.

2. For each categorical column (with unique values up to 20), we'll compute the following statistics:

1. Top 1: The most frequent value.

2. Top 2: The second most frequent value.

3. Top 3: The third most frequent value.

4. Mode: The value that appears most frequently.

5. Unique Count: The number of unique values.

3. For each numerical column, we'll calculate the following statistics:

1. 1st Quartile (25th percentile)

2. 2nd Quartile (Median or 50th percentile)

3. 3rd Quartile (75th percentile)

4. Mean

5. Median

6. Skewness

7. Kurtosis

8. Standard Deviation

9. Variance

4. Let's analyze the probability of loan default based on the gender of the applicant. You will:

1. Calculate the probability that a loan application results in default (TARGET = 1) given that the applicant is female.

2. Calculate the probability that a loan application results in default (TARGET = 1) given that the applicant is male.

5. Building on the previous task, you'll now calculate:

1. The probability that an applicant is female and the loan application results in default.

2. The probability that an applicant is male and the loan application results in default.

6. Finally, let's reverse the perspective using Bayes' Theorem:

1. Calculate the probability that an applicant is female given that the loan application resulted in default.

7. Implement stratified sampling based on the CODE\_GENDER column to ensure that your sample represents both genders proportionally. Perform the same analysis as in Task 2 & 3. Analyze the difference in the statistics.

1. Implement Stratified Sampling based on gender, ensuring proportional representation.

2. Repeat the descriptive statistics and probability analysis with this stratified sample.

3. Compare the findings with both the full dataset and the Simple Random Sample.

8. Treating the outliers

1. Visualize the presence of outliers in the 'AMT\_INCOME\_TOTAL', 'AMT\_CREDIT', 'AMT\_ANNUITY" columns using box plots. (Create a function to visualize outliers)

2. Identify which columns have significant outliers.

9. Removing the outliers (write a function to remove outliers)

1. After visualizing, remove the outliers from the identified columns (AMT\_INCOME\_TOTAL, AMT\_CREDIT, AMT\_ANNUITY) using the IQR

(Interquartile Range) method.

2. Re-visualize the distributions of the cleaned columns to confirm the outliers have been removed. (using the function created in the above question)

10. Evaluate the Impact of Sample size on overall mean (create a function to do this)

1. Draw samples of different sizes from the AMT\_INCOME\_TOTAL column.

2. Investigate how the distribution of sample means changes as you increase the sample size.

3. Visualize the results

11. Evaluate the Impact of number of samples

1. With a fixed sample size, draw different numbers of samples.

2. Explore how the distribution of sample means evolves as the number of samples increases.

3. Visualize the results

12. Is the average AMT\_INCOME\_TOTAL of loan applicants significantly different from a hypothesized population mean of $160,000? (One-sample t-test)

13. What are the 95% confidence intervals for the mean AMT\_CREDIT for all loan applicants?

14. Is there a significant difference in AMT\_ANNUITY across different types of contract (NAME\_CONTRACT\_TYPE)? (ANOVA)

15. Is there an association between gender (CODE\_GENDER) and loan default (TARGET)? (Chi-square test)

16. Which categorical variables are not significantly associated with loan default? (Chi-Square Test)

1. Categorical Variables:

1. CODE\_GENDER

2. FLAG\_OWN\_CAR

3. FLAG\_OWN\_REALTY

4. NAME\_CONTRACT\_TYPE

17. Is the average loan amount (AMT\_CREDIT) for female applicants within the 95% confidence interval of the overall average loan amount for all applicants?

18. Does the 95% confidence interval for the mean income (AMT\_INCOME\_TOTAL) of applicants who own a car overlap with those who do not?

19. Consider the average AMT\_ANNUITY for all applicants. Suppose a financial analyst claims that the mean annuity amount should be $30,000 based on industry standards. Conduct a one-sample t-test to evaluate this claim. Discuss your approach, assumptions, and the implications of your findings.

20. The dataset includes the AMT\_INCOME\_TOTAL column, representing the total income of applicants. Assume that the government has set a minimum income threshold of $180,000 for a specific loan program. Using a one-sample t-test, determine whether the average income of the applicants in the dataset significantly differs from this threshold. Explain your methodology, interpretation, and any potential limitations of your analysis.