BANK LOAN CASE STUDY

FINAL PROJECT - 2
BY NAZIREEN SANIA

INTRODUCTION

INTRODUCTION

We have three datasets :

- application_data - It contains current loan applications with customer attributes. It has 122 columns, providing detailed financial and demographic information for pattern analysis.
- previous_application - Contains information about previous loan applications made by customers, including contract details, status, and repayment history.
- columns_description - description about each column in the application_data.

DATA CLEANING

We have three datasets :

- application_data - It contains current loan applications with customer attributes. It has 122 columns, providing detailed financial and demographic information for pattern analysis.
- previous_application - Contains information about previous loan applications made by customers, including contract details, status, and repayment history.
- columns_description - description about each column in the application_data.

Out of the 3 datasets, we use application_data for our analysis.

HANDLING MISSING VALUES

If there are too many missing values in a column that is really important for our analysis, removing the rows might not be a feasible option. Hence we can fill these vaues using :

- Mean (if data is normally distributed)
- Median (if data has outliers or is skewed)
- Mode (if certain values dominate, like "0" for missing loan values)
- For categorical data, fill missing values with the most frequent category (mode).

To check if data is normally distributed, we use skewness. Skewness (should be close to 0 or ideally less than -1 for normal distribution).

- Positive skew: Tail is longer on the right.
- Negative skew: Tail is longer on the left.

If data is normally distributed, we use mean for imputation, and is data is not normally distributed, we use median to avoid outliers affecting results.

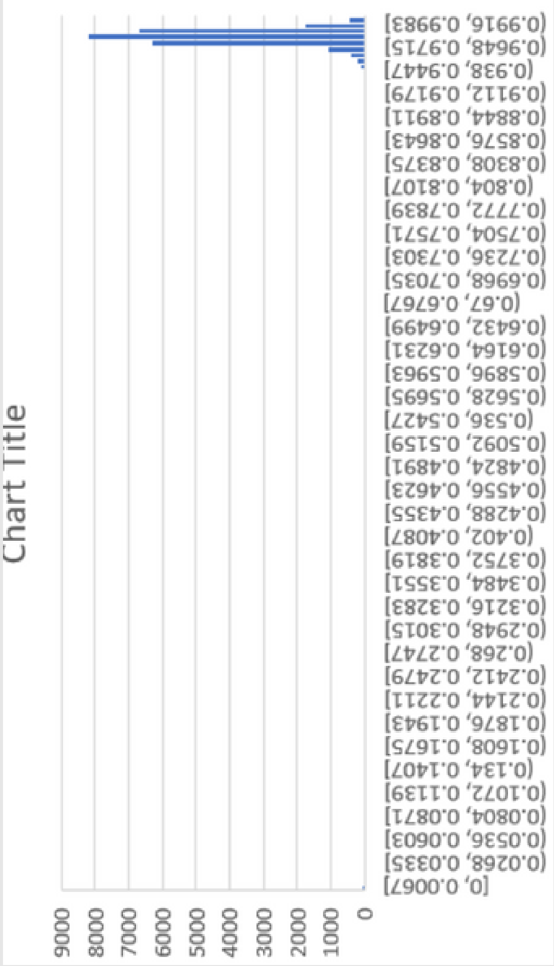
iIf skewness is not less than -1, but still close to 0, it is better to use median instead of mean for imputation since the mean would be influenced by extreme lower values, while the median is more robust.

DATA ANALYTICS TASKS

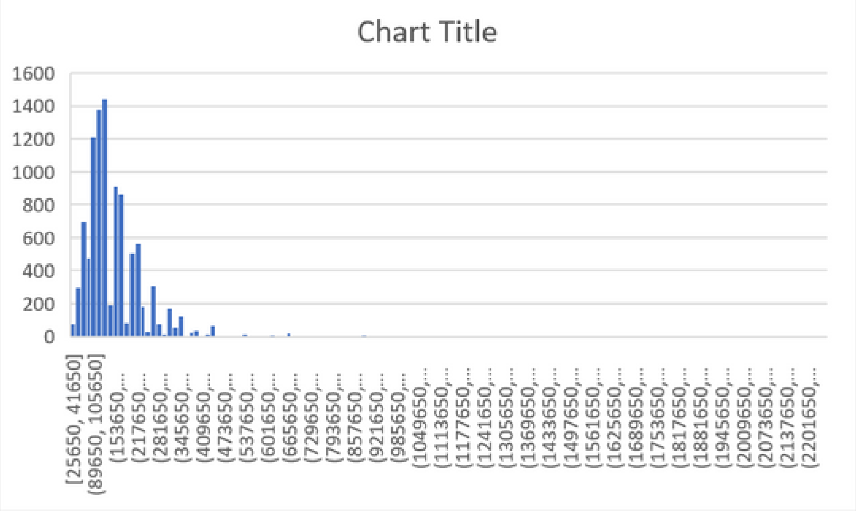
MISSING DATA

- The dataset contains columns that are not necessary for the analysis so we remove any columns that fall under this category.
- There are few columns which have more than half empty values and dealing with such columns by simply deleting rows will skew the analsis.
- First we remove duplicate rows by using the 'Remove Duplicates' option under the Data tab in Excel.
- For filing the empty values, we use the procedure discussed under 'Handling Missing Values' in Page 1.
-

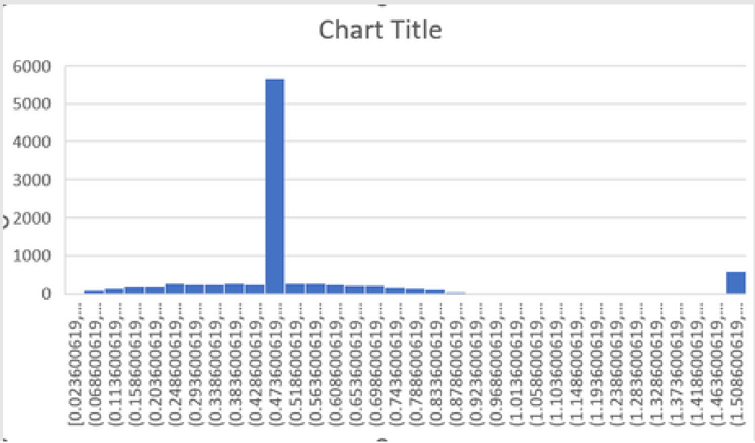
For example, the column EXT_Source_2 has a skewness -0.79 which is close to 0 but not less than 1, hence we use median instead of mean to imputate values.



Histograms *with many missing values hence not a normal curve*



Histograms *with no missing values and inmputated values, normal bell-shaped curve*



DATA ANALYTICS TASKS

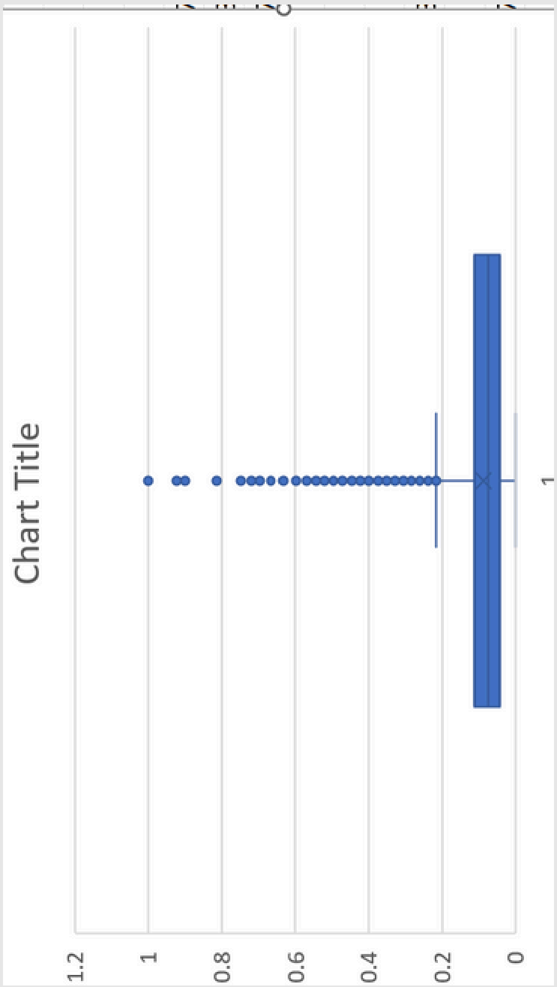
OUTLIERS

Outliers are extreme values that differ significantly from other observations. They can distort analysis and affect decision-making. Few ways to find outliers include :

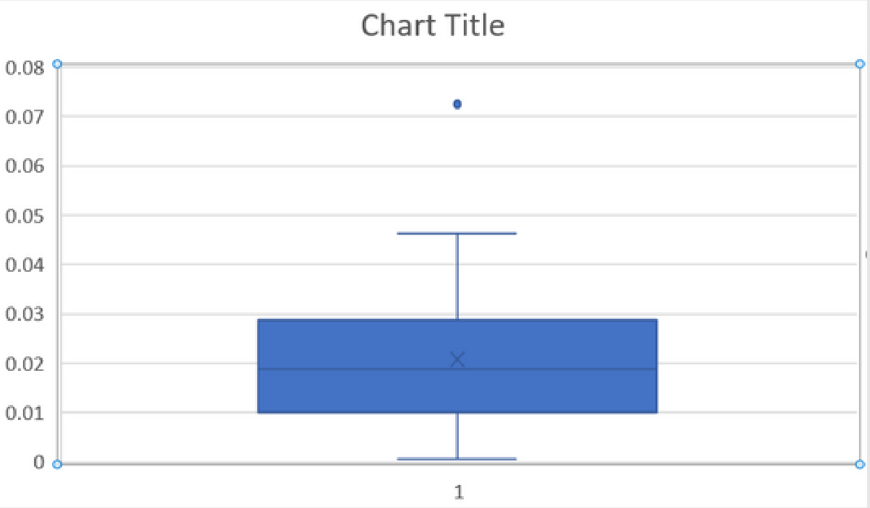
- Box plot
- Z-score
- Interquartile Range

Using IQR, we can find the 3rd and the 1st quartile and subtract IQR = Q3 - Q1.
Lower Bound: =Q1 - 1.5*IQR
Upper Bound: =Q3 + 1.5*IQR

We use to filter out values that fall into these conditions to detect outliers.



Box plot for a column with many missing values, hence many outliers



Box plot for a column with no missing values, hence only one or no outlier

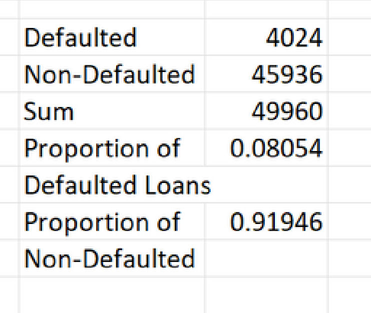
DATA ANALYTICS TASKS

DATA IMBALANCE

o analyze data imbalance in the loan application dataset, we follow these steps:

(1) IDENTIFY TARGET VARIABLE
Here, we use the column Target where

- 1 = Loan default
- 0 = No default (paid on time)



Defaulted	4024
Non-Defaulted	45936
Sum	49960
Proportion of Defaulted Loans	0.08054
Proportion of Non-Defaulted	0.91946

(2) CALCULATE CLASS DISTRIBUTION

- Using COUNTIF

- Defaulted Loans: =COUNTIF(TARGET,1)
- Non-Defaulted Loans: =COUNTIF(TARGET,0)
- Total Records: =SUM(deafulted+non-defaulted)
- Proportion of Defaulted Loans: =defaulted/sum
- Proportion of Non-Defaulted Loans: = non-defaulted/sum

---> Interpreting the Results

- If the proportion of one class is much smaller (e.g., 10% default, 90% no default), the dataset is imbalanced.
- If both classes are similar (e.g., 50-50), the dataset is balanced.

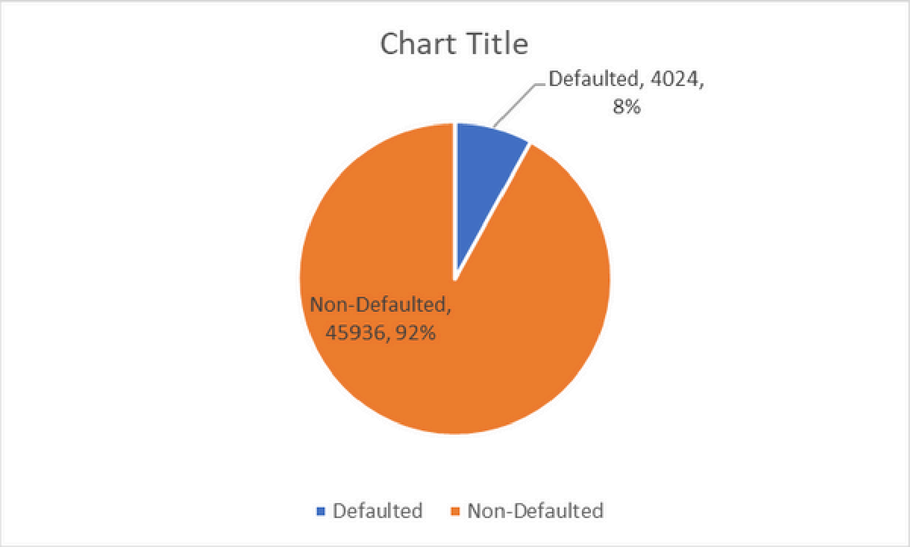
We can see the proportion of Non-defaulted Loans is much higher than the proportion of Defaulted Loans.

- Defaulted Loans (TARGET = 1): 8.05%
- Non-Defaulted Loans (TARGET = 0): 91.95%

- The majority class (Non-Defaulted Loans: 91.95%) dominates the dataset.
- The minority class (Defaulted Loans: 8.05%) is significantly underrepresented.
- This suggests a class imbalance problem, which can affect analysis and predictive models.

(3) VISULAIZING THE DATA IMBALANCE

- Using a Pie Chart
- Using a Bar Chart



Why is this a Concern?

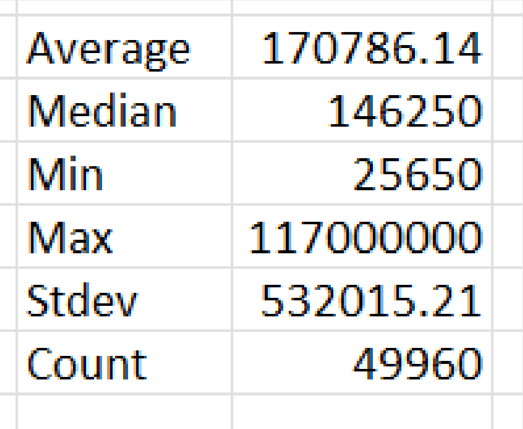
- Bias in Decision-Making: A model trained on this dataset may favor the majority class and fail to accurately predict defaults.
- Skewed Insights: Analysis may not properly capture factors influencing default, as there's less data on defaulters.
- Risk in Banking: Identifying default risks is critical. If the model struggles with defaulters, it may approve risky loans.

DATA ANALYTICS TASKS

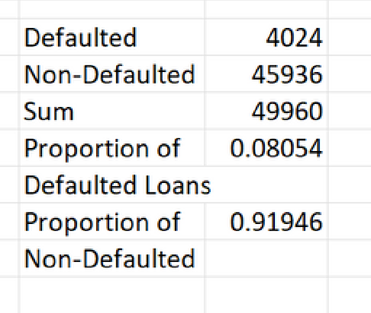
UNIVARIATE ANALYSIS

Univariate Avalysis is exploring individual variables' distributions.

Let us consider the Income column for this analysis.



Average	170786.14
Median	146250
Min	25650
Max	117000000
Stdev	532015.21
Count	49960



Defaulted	4024
Non-Defaulted	45936
Sum	49960
Proportion of Defaulted Loans	0.08054
Proportion of Non-Defaulted	0.91946

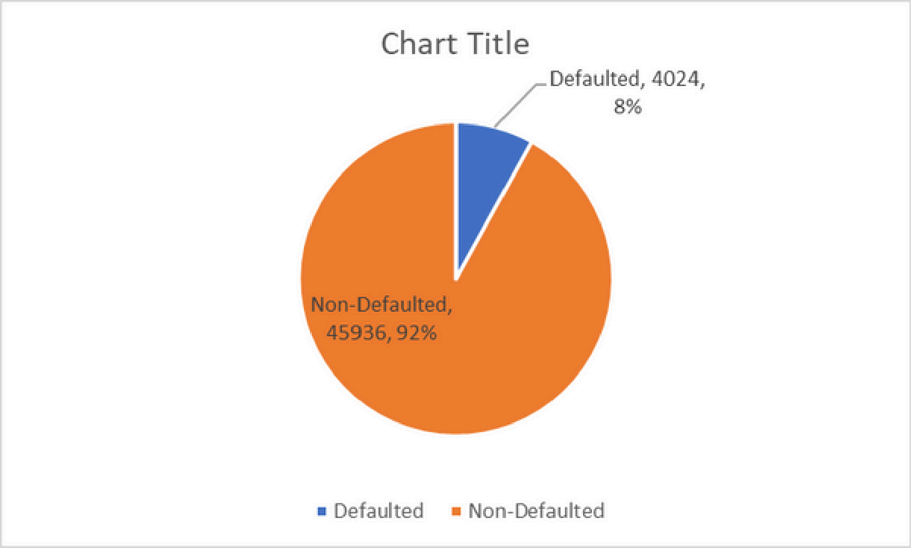
We can see the proportion of Non-defaulted Loans is much higher than the proportion of Defaulted Loans.

- Defaulted Loans (TARGET = 1): 8.05%
- Non-Defaulted Loans (TARGET = 0): 91.95%

- The majority class (Non-Defaulted Loans: 91.95%) dominates the dataset.
- The minority class (Defaulted Loans: 8.05%) is significantly underrepresented.
- This suggests a class imbalance problem, which can affect analysis and predictive models.

Why is this a Concern?

- Bias in Decision-Making: A model trained on this dataset may favor the majority class and fail to accurately predict defaults.
- Skewed Insights: Analysis may not properly capture factors influencing default, as there's less data on defaulters.
- Risk in Banking: Identifying default risks is critical. If the model struggles with defaulters, it may approve risky loans.

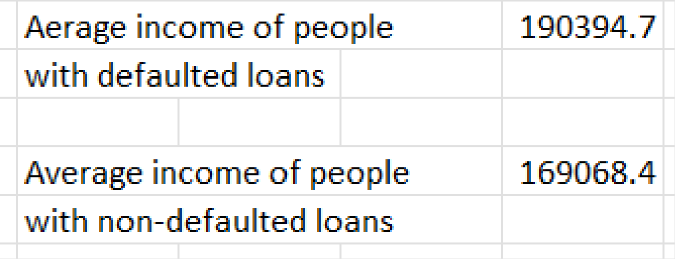


DATA ANALYTICS TASKS

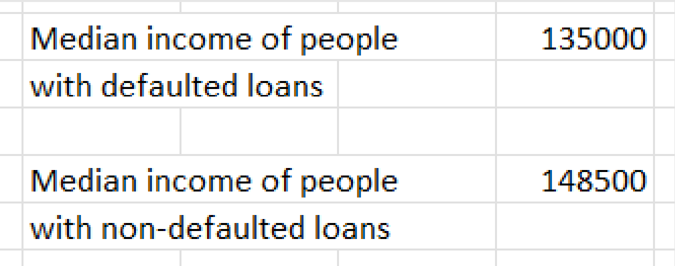
SEGMENTED UNIVARIATE ANALYSIS

For this analysis, we can compare variable distributions for defaulted vs non-defaulted cases.

Using the filter option, we filter out incomes of people with defaulted and non-defaulted loans seperately and find the averages and medians of those incomes respectively

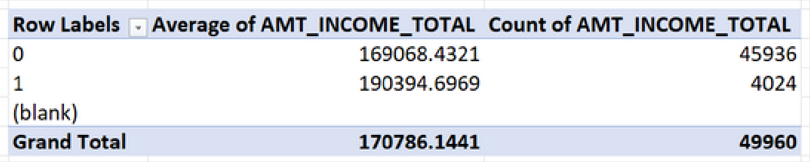


Aerage income of people with defaulted loans	190394.7
Average income of people with non-defaulted loans	169068.4



Median income of people with defaulted loans	135000
Median income of people with non-defaulted loans	148500

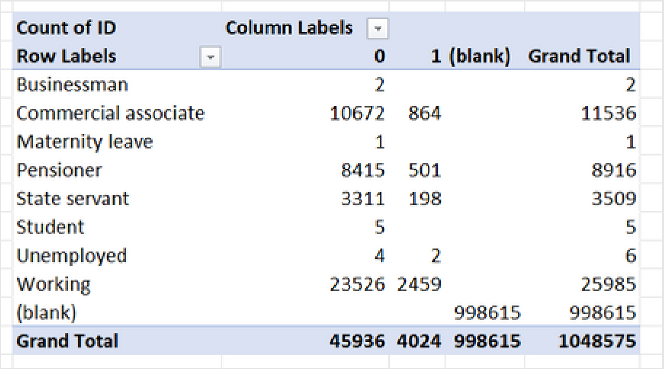
To sum it all up, we use a pivot table



Row Labels	Average of AMT_INCOME_TOTAL	Count of AMT_INCOME_TOTAL
0	169068.4321	45936
1	190394.6969	4024
(blank)		
Grand Total	**170786.1441**	**49960**

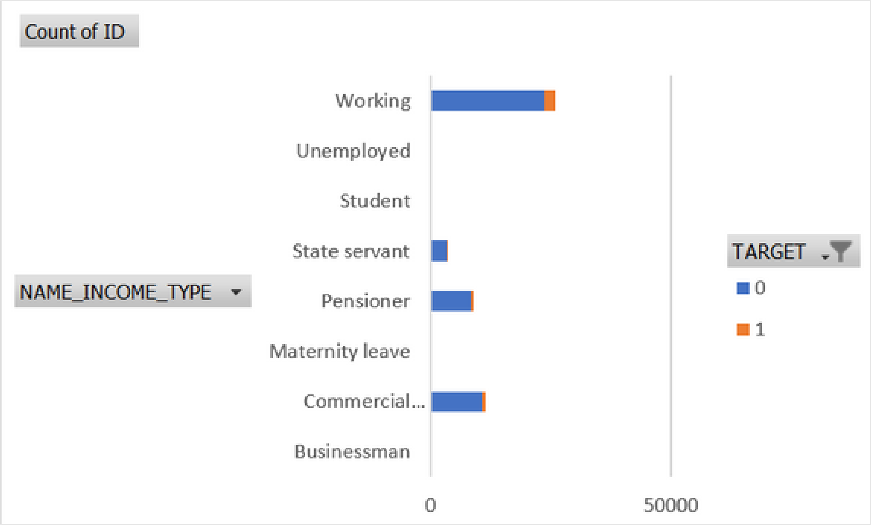
In another instance to compare variable distributions for different scenarios, we use the columns Target, Name_Income_Type and count of ID, which we created to make our anaylysis simpler. The ID column is just serial numbers.

In the pivot column, we put Target under Columns, Name_Income_Type under Rows and Count of ID under Values.



Count of ID	Column Labels			
Row Labels	0	1	(blank)	Grand Total
Businessman	2			2
Commercial associate	10672	864		11536
Maternity leave	1			1
Pensioner	8415	501		8916
State servant	3311	198		3509
Student	5			5
Unemployed	4	2		6
Working	23526	2459		25985
(blank)			998615	998615
Grand Total	**45936**	**4024**	**998615**	**1048575**

This helps us in understanding the number of deafulted and non-defaulted loans for people belonging to different professions. In pending applications, this makes it easier to know which segment of people are more likely to defualt a loan.



We can visualise the data in pivot table using a stacked bar chart. Shows if certain income groups (e.g., Pensioners, Business Owners, Employees) have higher default rates.

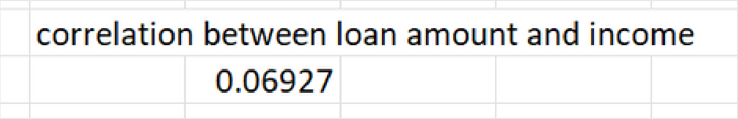
DATA ANALYTICS TASKS

TOP CORRELATIONS

For this analysis, we can explore correlations between variables and loan default.

(1) Finding Correlation between income and loan amount to see if higher income lead to larger loan amounts.

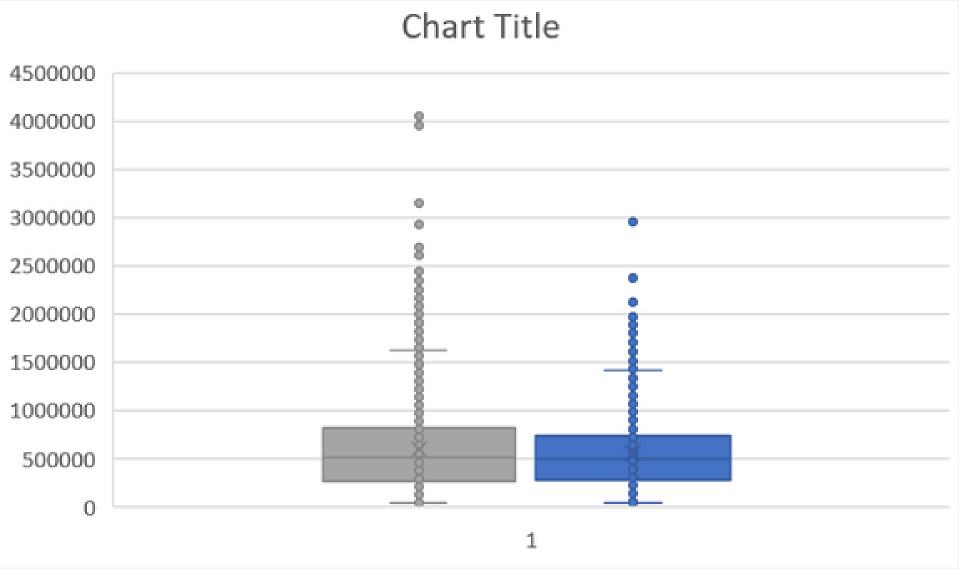
The correlation value 0.06927 is very close to 0, meaning there is almost no linear relationship between the loan amount and whether a person defaults.



correlation between loan amount and income

0.06927

- A low correlation suggests that loan amount alone is not a strong predictor of default.
- Other factors (like income, employment type, credit history) might have a stronger impact on loan default.
- Even if there is a relationship, it may be non-linear (not well captured by Pearson correlation).



Loan distribution for defaulted vs. non-defaulted loans