Bank Loan Analysis

Objective: Analyze patterns in loan application data to identify key factors influencing loan defaults, thereby improving loan approval decisions and minimizing financial losses.

Project Description

Background: As a data analyst at a finance company, we face the challenge of applicants with insufficient credit history defaulting on loans. This analysis aims to address two risks:

- Losing business from capable applicants who are rejected.
- Financial losses from approving loans to applicants who default.

Dataset:

- Customers with Payment Difficulties: Late payment of more than X days on at least one of the first Y installments.
- All Other Cases: Timely payments.

Loan application outcomes include:

- Approved: The company has approved the loan application.
- Cancelled: The customer canceled the application during the approval process.
- Refused: The company rejected the loan.
- Unused Offer: The loan was approved, but the customer did not use it.

Project Description

Goals:

- Identify patterns indicating higher risk of default.
- Improve loan approval decisions based on EDA insights.

Methodology:

- **Data Collection and Cleaning**: Prepare the dataset by handling missing values and outliers.
- **Exploratory Data Analysis**: Analyze customer and loan attributes, visualize data to uncover patterns.
- **Risk Assessment**: Evaluate risk for each loan application.
- **Recommendations**: Provide actionable insights to optimize loan approval processes.

Business Impact:

- Reduce financial losses by identifying high-risk applicants.
- Increase business by ensuring capable applicants are approved.
- Optimize loan terms and interest rates to manage risk better.

Approach

We’ll begin by loading the dataset into the excel. Then, we will thoroughly analyze the dataset. This includes understanding the structure, examining the variables, and ensuring the data is clean and ready for detailed analysis. We perform this initial step to ensure accuracy and reliability in our subsequent analysis.

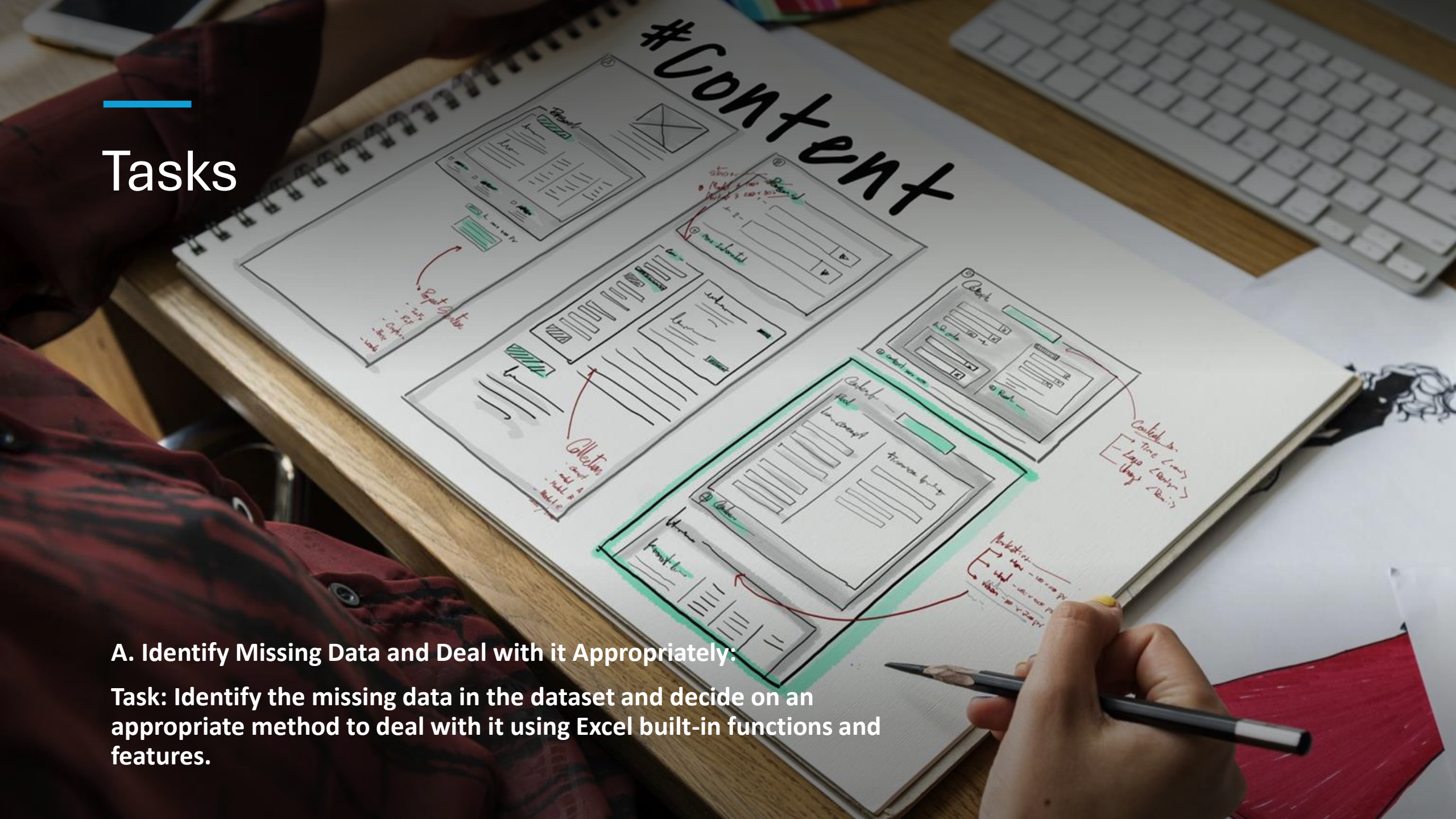
Tech-Stack : Excel

Why Excel?

Excel is a versatile tool widely used for data analysis due to its user-friendly interface, powerful data manipulation capabilities, and extensive range of built-in functions and visualization options. Here's how Excel will be utilized for the Bank Loan Default Analysis project:

- Data Import: Import and consolidate data
- Data Cleaning: Handle missing values, remove duplicates, and format data types.
- Exploratory Analysis: Calculate statistics, create charts (e.g., histograms, scatter plots), and use pivot tables.
- Visualization: Generate visual insights to understand data distributions and relationships effectively.

Tasks



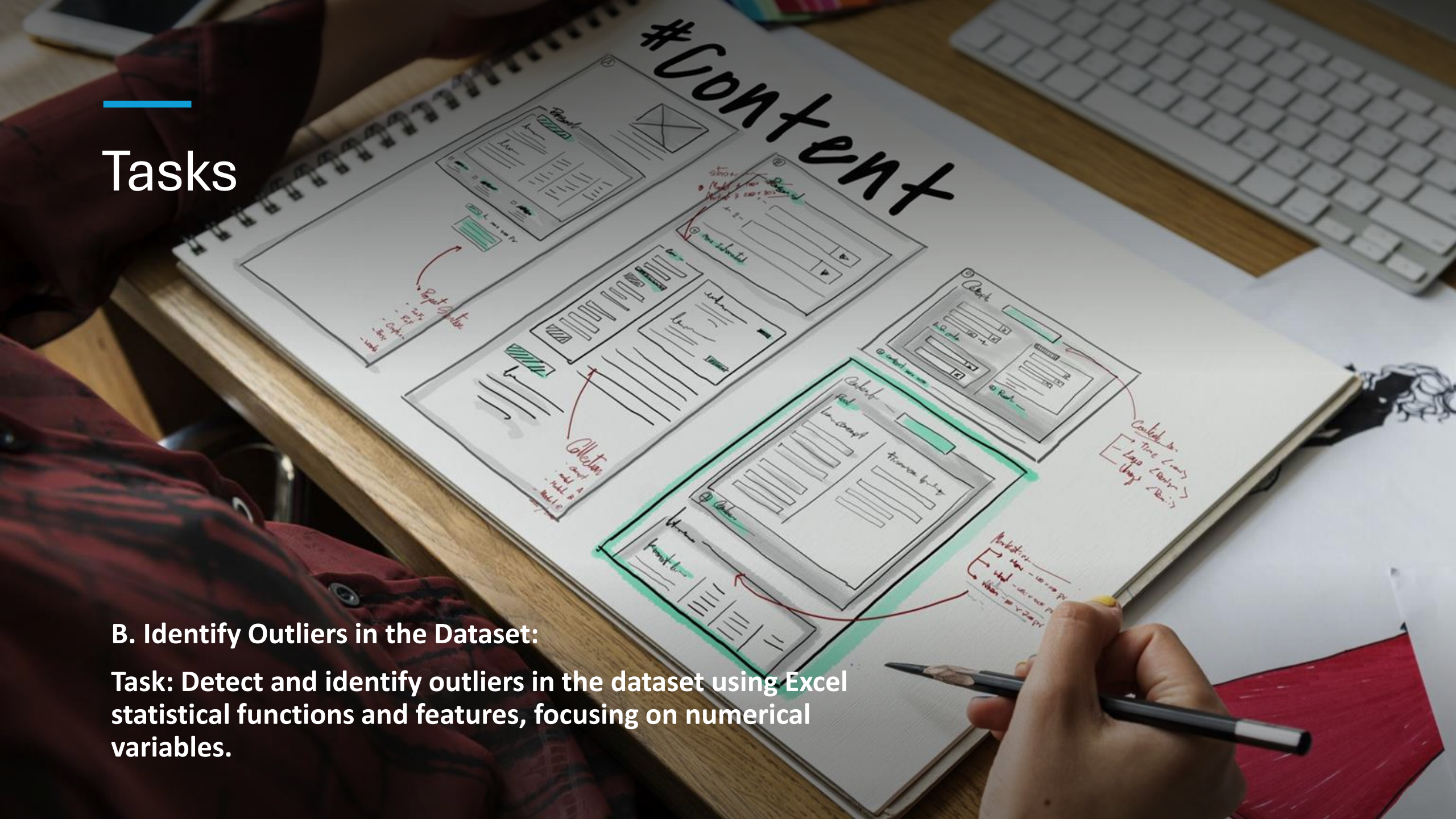
A. Identify Missing Data and Deal with it Appropriately:

Task: Identify the missing data in the dataset and decide on an appropriate method to deal with it using Excel built-in functions and features.

Insight

- The data had 50k rows and 122 columns and many of which had missing values.
- We calculated the missing values with the use of Function CountBlank
- We then calculated the percentage of the missing value and discarded the columns which had more than 30% missing values.
- We performed imputation using Average function for the numerical values in the columns.

Tasks

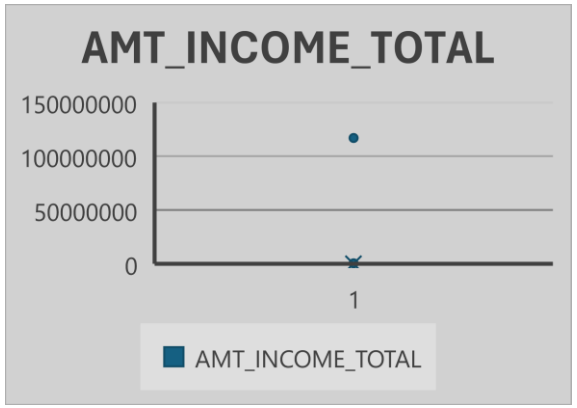


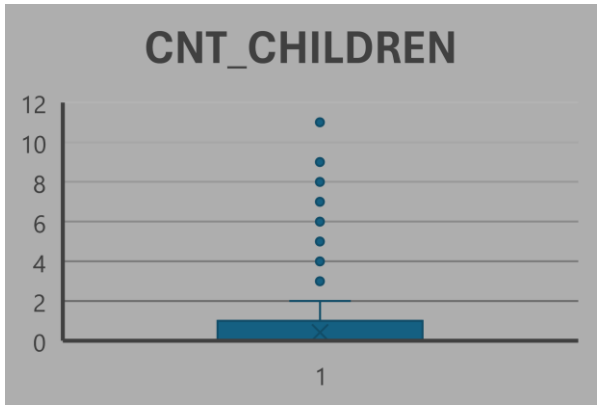
B. Identify Outliers in the Dataset:

Task: Detect and identify outliers in the dataset using Excel statistical functions and features, focusing on numerical variables.

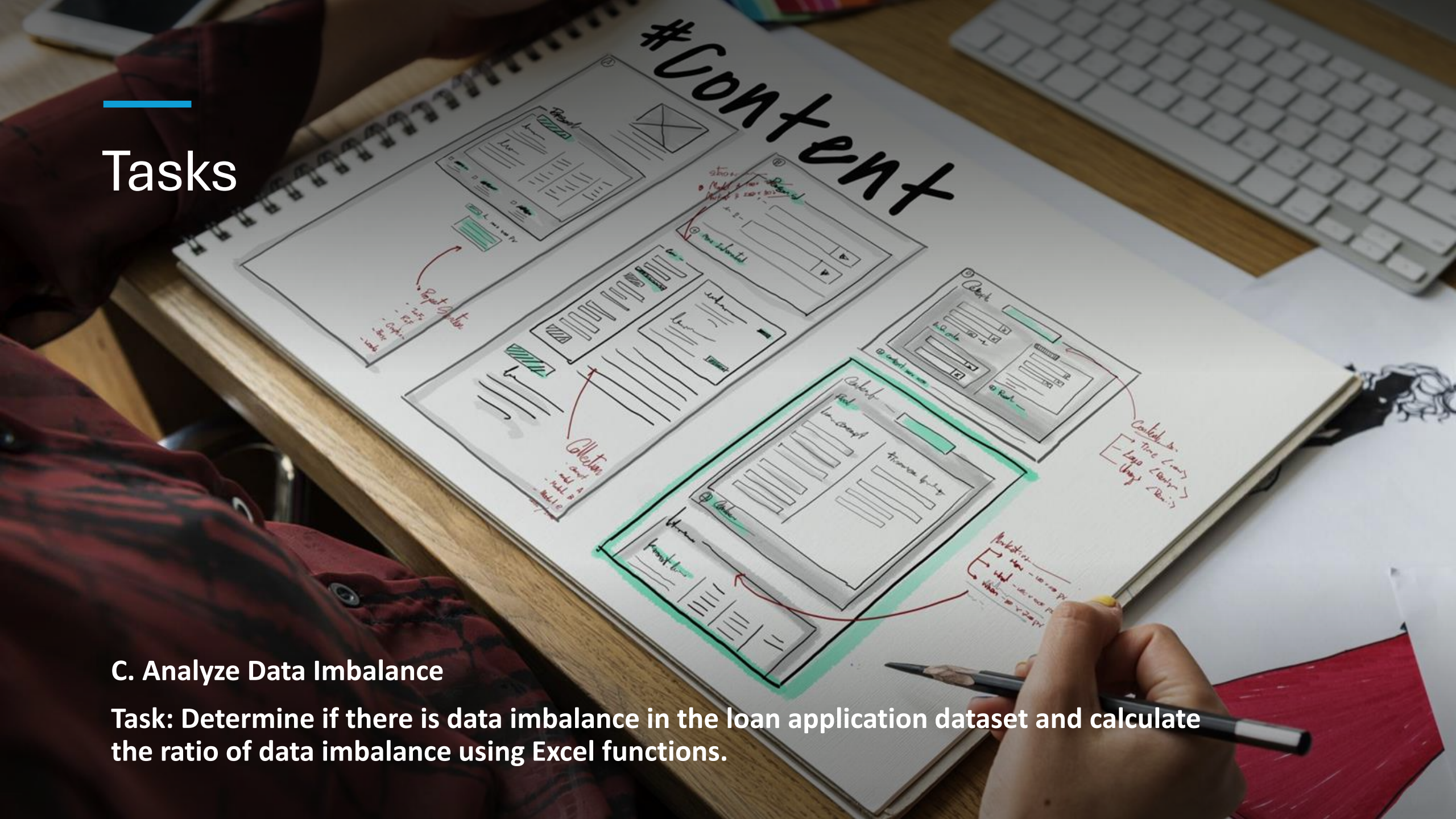
Insight

We identified outliers in several columns; however, some outliers were found to be implausible. We have highlighted these anomalies below:





Tasks

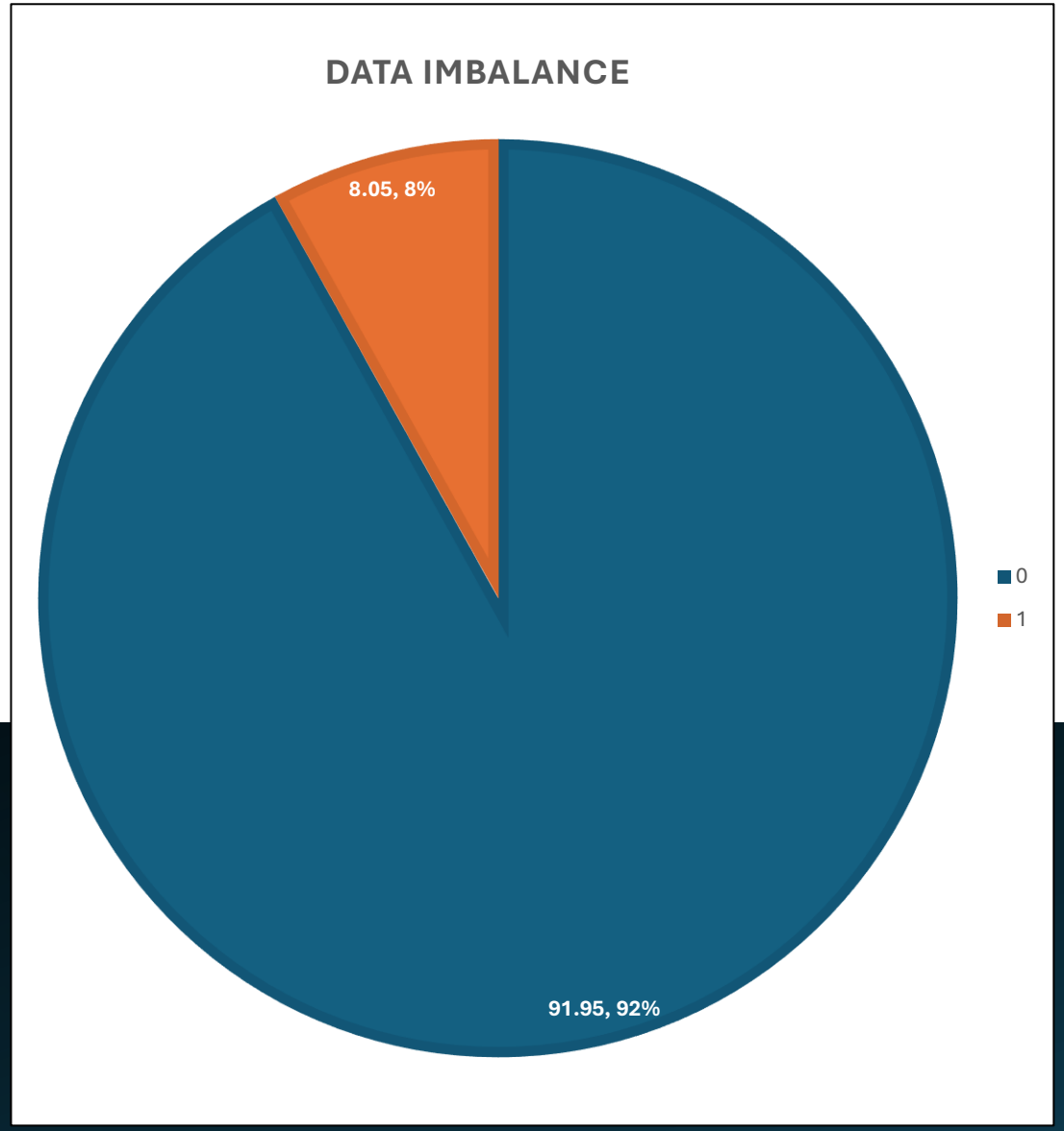


C. Analyze Data Imbalance

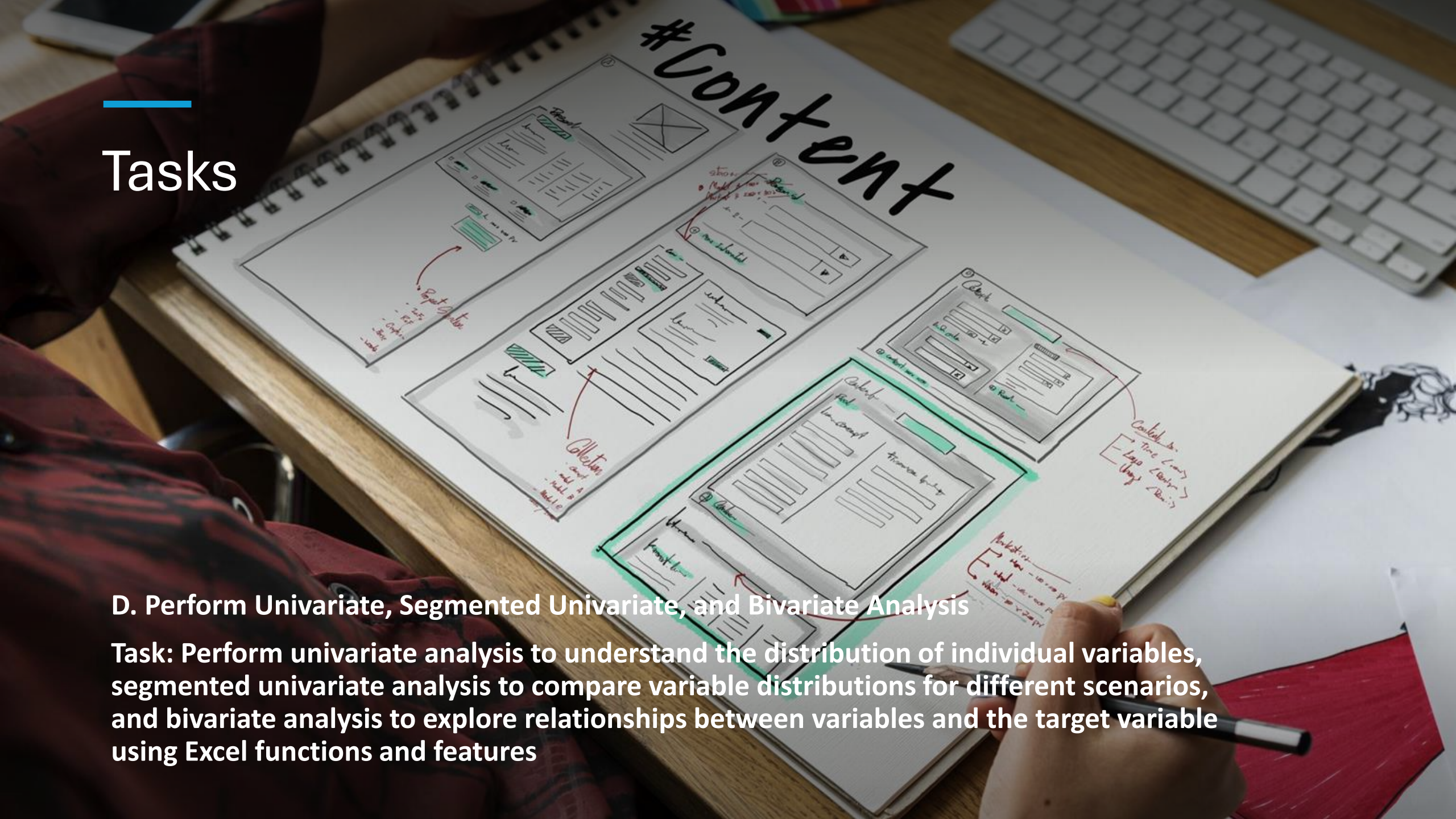
Task: Determine if there is data imbalance in the loan application dataset and calculate the ratio of data imbalance using Excel functions.

Insight

The Data Imbalance ratio was found out to be 11.419



Tasks



D. Perform Univariate, Segmented Univariate, and Bivariate Analysis

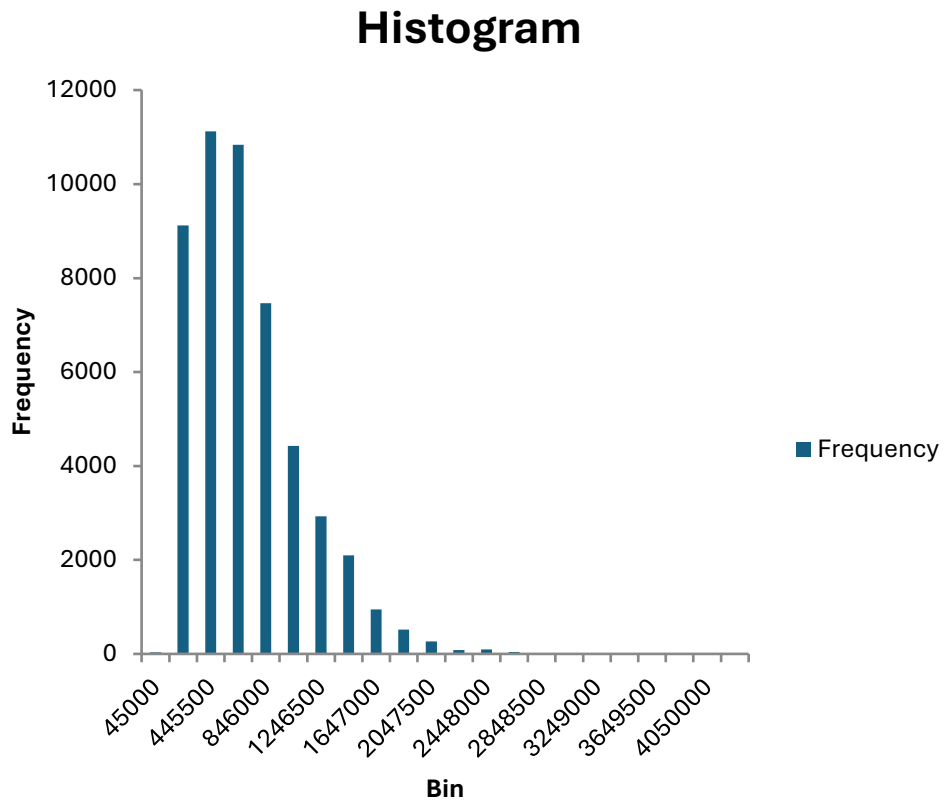
Task: Perform univariate analysis to understand the distribution of individual variables, segmented univariate analysis to compare variable distributions for different scenarios, and bivariate analysis to explore relationships between variables and the target variable using Excel functions and features

Insight

By conducting the univariate, segmented and the bivariate analysis, we achieved the following:

- Clear understanding of how individual variables are distributed within the dataset.
- Identification of differences in variable distributions across different segments or scenarios.
- Insights into how variables relate to each other and their potential impact on the target variable (e.g., loan default).

Finally, we used Charts and graphs to visually convey findings, aiding in clear communication and decision-making.

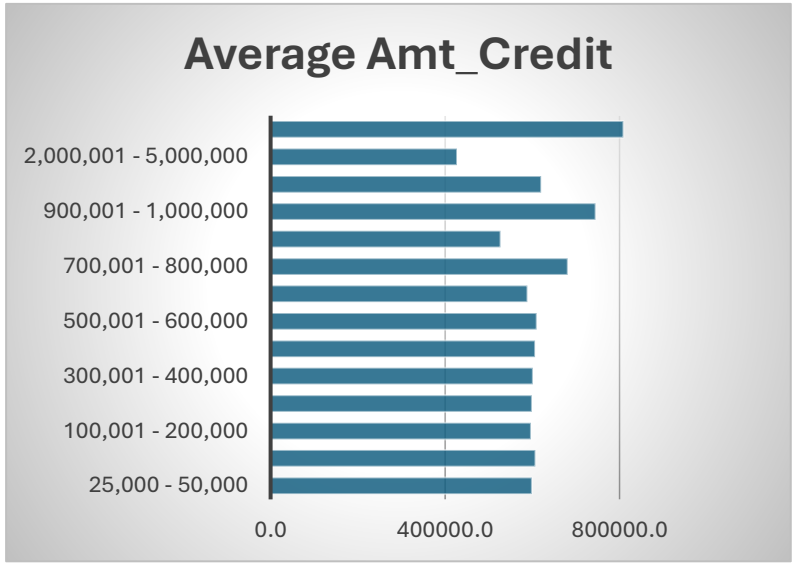


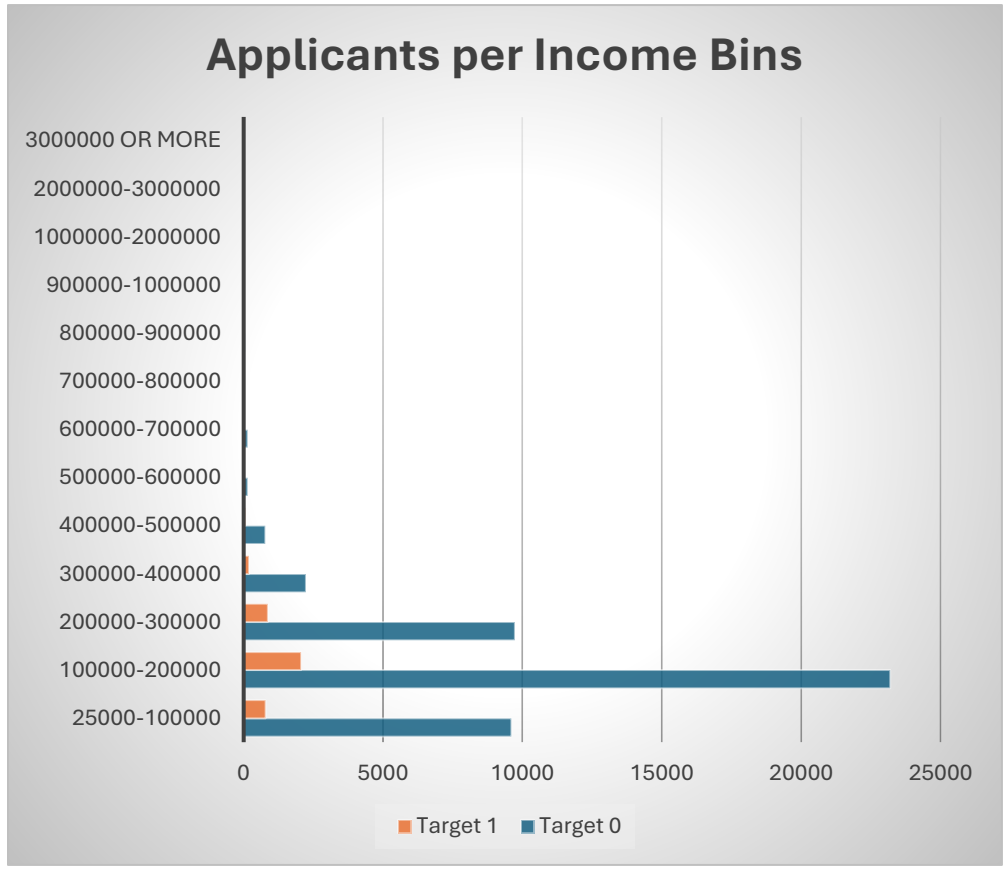
Univariate Analysis

Analysis of individual variables.

Bivariate Analysis

Analysis of variable relationships.

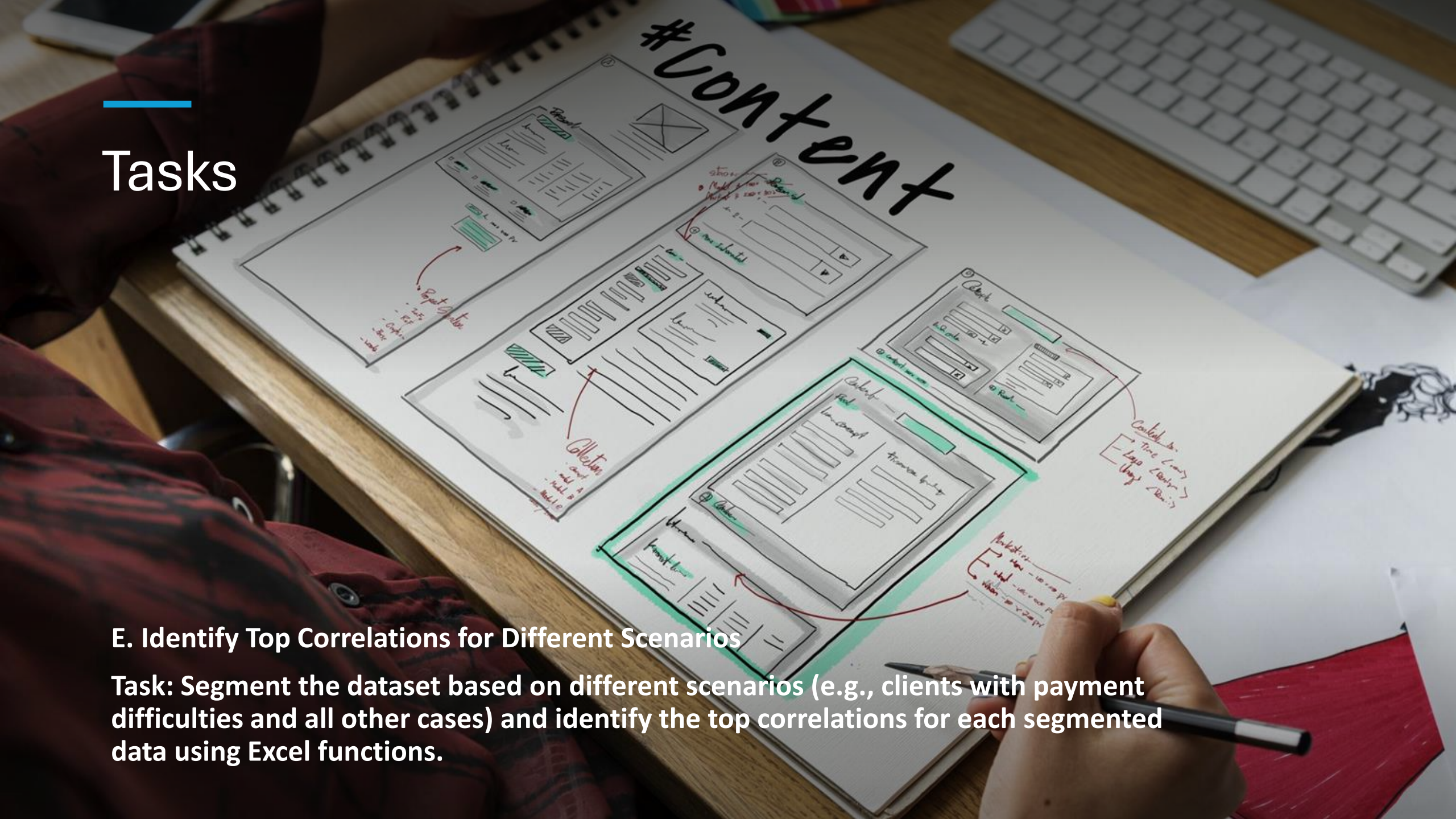




Segmented Univariate Analysis

Detailed analysis of segmented variables.

Tasks



E. Identify Top Correlations for Different Scenarios

Task: Segment the dataset based on different scenarios (e.g., clients with payment difficulties and all other cases) and identify the top correlations for each segmented data using Excel functions.

Insight

Correlation for Target 1

	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	REGION_POPULATION_RELATIVE	DAYS_BIRTH (Years)	DAYS_EMPLOYED (Years)	DAYS_REGISTRATION4 (Years)	DAYS_ID_PUBLISH5	REGION_RATING_CLIENT_W_CITY
CNT_CHILDREN	1	0.009588558	0.00497156	-0.025555665	-0.329086361	-0.241544855	-0.180890946	0.032216887	0.022777663
AMT_INCOME_TOTAL	0.009588558	1	0.069315897	0.029841469	-0.015823731	-0.031504237	-0.009741098	-0.003455339	-0.040719164
AMT_CREDIT	0.00497156	0.069315897	1	0.095111221	0.059486879	-0.067738113	-0.003287576	0.011966006	-0.109486833
REGION_POPULATION_RELATIVE	-0.025555665	0.029841469	0.095111221	1	0.032471459	-0.004163683	0.059193718	0.004430163	-0.530438555
DAYS_BIRTH (Years)	-0.329086361	-0.015823731	0.059486879	0.032471459	1	0.621489139	0.333246909	0.270959903	-0.014659507
DAYS_EMPLOYED (Years)	-0.241544855	-0.031504237	-0.067738113	-0.004163683	0.621489139	1	0.208933695	0.271888778	0.036976884
DAYS_REGISTRATION (Years)	-0.180890946	-0.009741098	-0.003287576	0.059193718	0.333246909	0.208933695	1	0.104526727	-0.079656448
DAYS_ID_PUBLISH	0.032181658	-0.003457803	0.011959309	0.004345687	0.270894566	0.271908031	0.104574138	1	0.006705227
REGION_RATING_CLIENT_W_CITY	0.022618239	-0.040712493	-0.109476208	-0.53043311	-0.014574138	0.036982629	-0.079693992	0.006705227	1

Correlation for Target 0

	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	REGION_POPULATION_RELATIVE	DAYS_BIRTH (Years)	DAYS_EMPLOYED (Years)	DAYS_REGISTRATION4 (Years)	DAYS_ID_PUBLISH5	REGION_RATING_CLIENT_W_CITY
CNT_CHILDREN	1	0.009588558	0.00497156	-0.025555665	-0.329086361	-0.241544855	-0.180890946	0.032216887	0.022777663
AMT_INCOME_TOTAL	0.00497156	1	0.069315897	0.029841469	-0.015823731	-0.031504237	-0.009741098	-0.003455339	-0.040719164
AMT_CREDIT	0.00497156	0.069315897	1	0.095111221	0.059486879	-0.067738113	-0.003287576	0.011966006	-0.109486833
REGION_POPULATION_RELATIVE	-0.025555665	0.029841469	0.095111221	1	0.032471459	-0.004163683	0.059193718	0.004430163	-0.530438555
DAYS_BIRTH (Years)	-0.329086361	-0.015823731	0.059486879	0.032471459	1	0.621489139	0.333246909	0.270959903	-0.014659507
DAYS_EMPLOYED (Years)	-0.241544855	-0.031504237	-0.067738113	-0.004163683	0.621489139	1	0.208933695	0.271888778	0.036976884
DAYS_REGISTRATION (Years)	-0.180890946	-0.009741098	-0.003287576	0.059193718	0.333246909	0.208933695	1	0.104526727	-0.079656448
DAYS_ID_PUBLISH	0.032216887	-0.003455339	0.011966006	0.004430163	0.270959903	0.271888778	0.104526727	1	0.006711396
REGION_RATING_CLIENT_W_CITY	0.022777663	-0.040719164	-0.109486833	-0.530438555	-0.014659507	0.036976884	-0.079656448	0.006711396	1

Result

Through these tasks, we gained a comprehensive understanding of the loan application dataset, identified issues such as missing data and outliers, assessed data distribution and imbalance, explored relationships between variables and loan default, and identified key correlations influencing loan outcomes. This hands-on experience with Excel provided practical skills in data cleaning, exploratory analysis, and visualization, crucial for informed decision-making in loan management and risk assessment within the financial services sector