Trainity project 6

Bank Loan Case Study

Project Description:

Imagine you're a data analyst at a finance company that specializes in lending various types of loans to urban customers. Your company faces a challenge: some customers who don't have a sufficient credit history take advantage of this and default on their loans. Your task is to use Exploratory Data Analysis (EDA) to analyze patterns in the data and ensure that capable applicants are not rejected.

When a customer applies for a loan, your company faces two risks:

1. If the applicant can repay the loan but is not approved, the company loses business.
2. If the applicant cannot repay the loan and is approved, the company faces a financial loss.

The dataset you'll be working with contains information about loan applications. It includes two types of scenarios:

1. Customers with payment difficulties: These are customers who had a late payment of more than X days on at least one of the first Y instalments of the loan.
2. All other cases: These are cases where the payment was made on time.

When a customer applies for a loan, there are four possible outcomes:

1. Approved: The company has approved the loan application.
2. Cancelled: The customer cancelled the application during the approval process.
3. Refused: The company rejected the loan.
4. Unused Offer: The loan was approved but the customer did not use it.

Your goal in this project is to use EDA to understand how customer attributes and loan attributes influence the likelihood of default.

Approach

Firstly I downloaded the dataset given and imported to the excel workbook and cleaned the given data by identifying blanks in the columns and removing columns containing more than 35% blanks and started doing the given tasks by using excel functions, pivot table, charts etc.

Tech-Stack Used

I used Microsoft Excel 2021 for this project for its various features like functions, pivot tables etc to perform tasks easily.

Insights

Applicants with payment difficulties often have higher financial burdens,lower incomes,struggle to repay loan and weaker credit scores

Loan amounts are often proportional to income

Data imbalance in the TARGET variable.

Result

This project provides actionable information for the finance company to refine its loan process, improve risk assessment , and reduce financial losses.

My excel workbook: https://docs.google.com/spreadsheets/d/1x45K680eVVN-vAcKn2zvMtFc8U6XNaVh/edit?usp=sharing&ouid=110428113670481159623&rtpof=true&sd=true

Data Analytics Tasks:

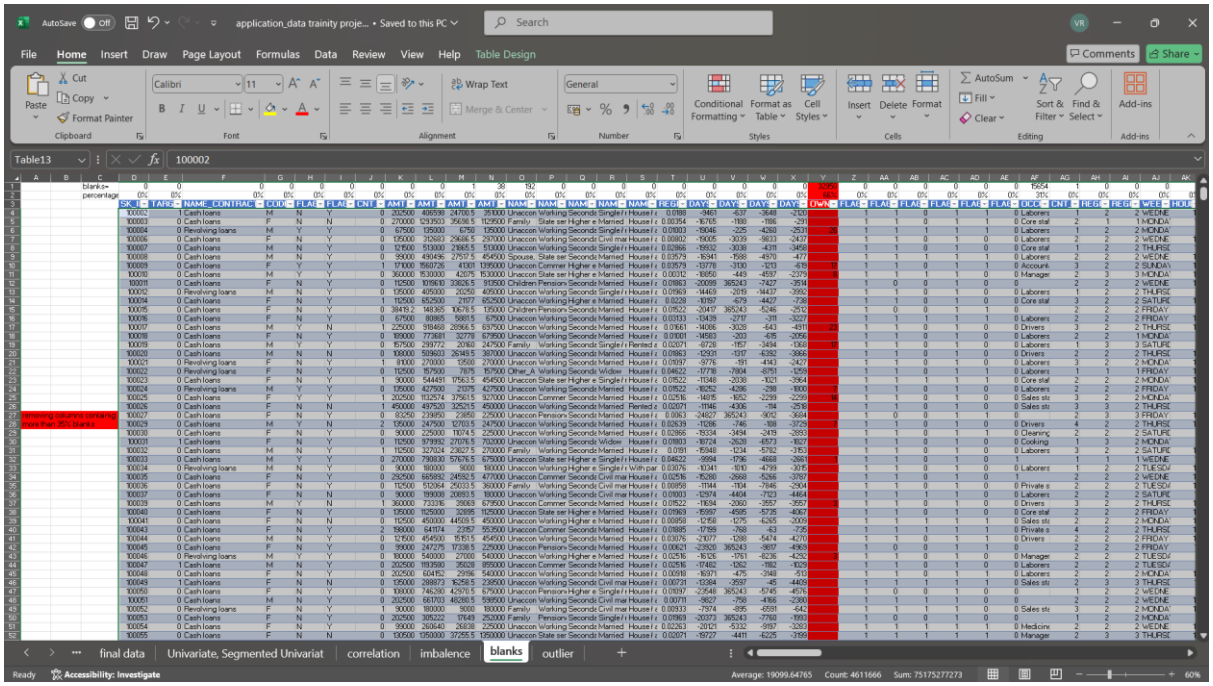
A. **Identify Missing Data and Deal with it Appropriately:** As a data analyst, you come across missing data in the loan application dataset. It is essential to handle missing data effectively to ensure the accuracy of the analysis.

- **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.
- **Hint:** Utilize Excel functions like COUNT, ISBLANK, and IF to identify missing data. Consider using functions like AVERAGE or MEDIAN for imputation or other appropriate methods available in Excel.
- **Graph suggestion:** Create a bar chart or column chart to visualize the proportion of missing values for each variable.

Insights:

High Proportion of Missing Data in Some Variables

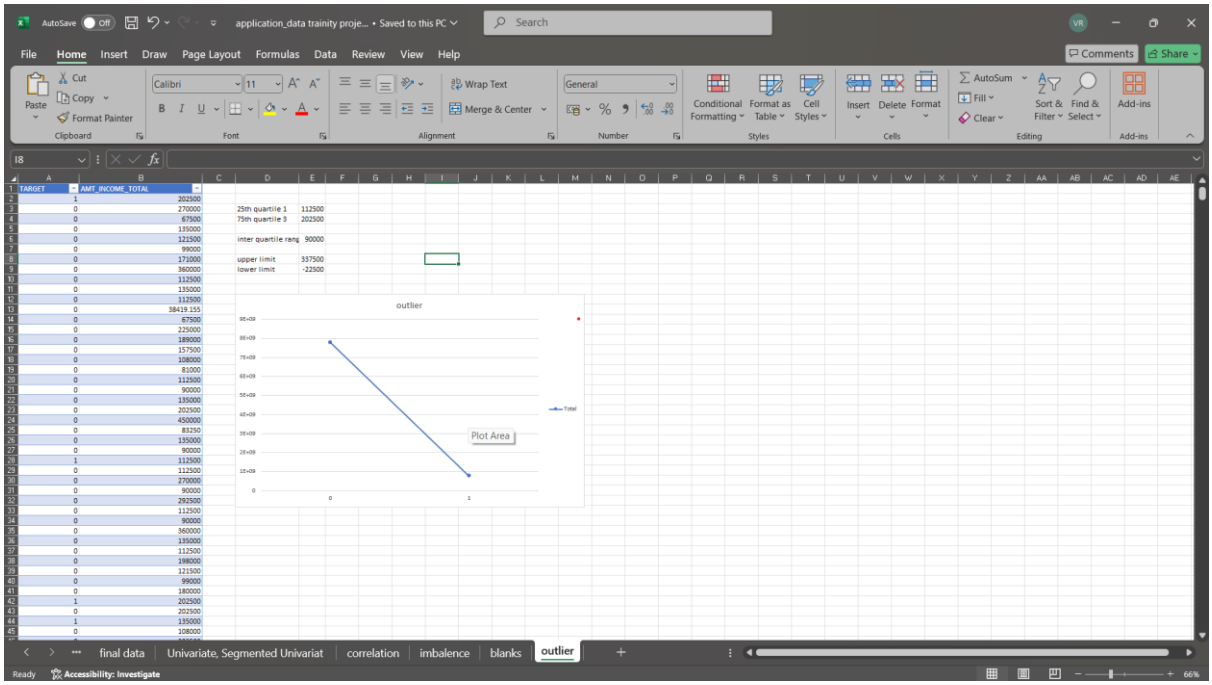
Columns like COMMONAREA_MEDI, NONLIVINGAPARTMENTS_MEDI, and LIVINGAPARTMENTS_MODE have over 68% missing data. Such high levels of missingness make these variables unreliable for analysis and were recommended for exclusion.



B. Identify Outliers in the Dataset: Outliers can significantly impact the analysis and distort the results. You need to identify outliers in the loan application dataset.

- **Task:** Detect and identify outliers in the dataset using Excel statistical functions and features, focusing on numerical variables.
- **Hint:** Utilize Excel functions like QUARTILE, IQR, and conditional formatting to identify potential outliers. Consider applying thresholds or business rules to determine if the outliers are valid data points or require further investigation.
- **Graph suggestion:** Create box plots or scatter plots to visualize the distribution of numerical variables and highlight the outliers.

Variables showed significant outliers. These likely represent extreme cases, such as applicants with very high credit limits or incomes.
The Interquartile Range (IQR) method was applied to detect outliers



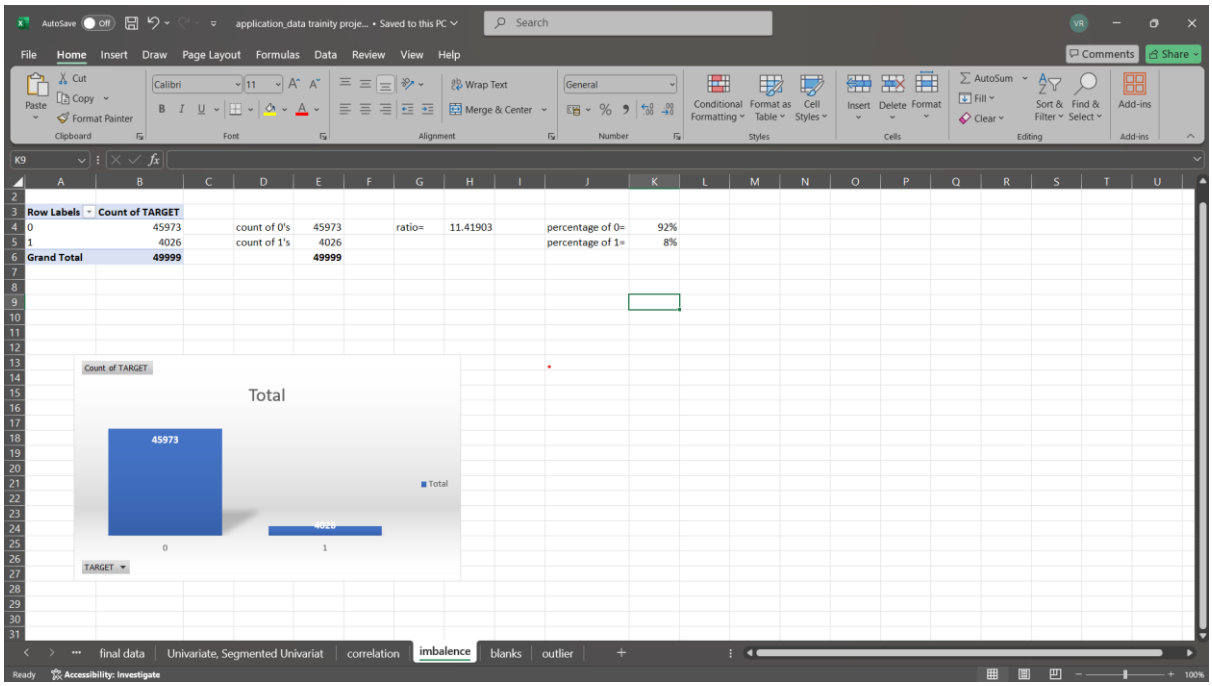
C. Analyze Data Imbalance: Data imbalance can affect the accuracy of the analysis, especially for binary classification problems. Understanding the data distribution is crucial for building reliable models.

- **Task:** Determine if there is data imbalance in the loan application dataset and calculate the ratio of data imbalance using Excel functions.
- **Hint:** Utilize Excel functions like COUNTIF and SUM to calculate the proportions of each class. Compare the class frequencies to assess data imbalance.
- **Graph suggestion:** Create a pie chart or bar chart to visualize the distribution of the target variable and highlight the class imbalance.

Insights:

Imbalanced Target Variable (TARGET):

- The dataset is highly imbalanced, with a majority of applicants (approximately 91%) not facing payment difficulties (TARGET=0) and a small proportion (about 9%) facing difficulties (TARGET=1).



D. Perform Univariate, Segmented Univariate, and Bivariate Analysis: To gain insights into the driving factors of loan default, it is important to conduct various analyses on consumer and loan attributes.

- **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.
- **Hint:** Utilize Excel functions like COUNT, AVERAGE, MEDIAN, and statistical functions for descriptive analysis. Utilize Excel features like filters, sorting, and pivot tables for segmented and bivariate analysis.
- **Graph suggestion:** Create histograms, bar charts, or box plots to visualize the distributions of variables. Create stacked bar charts or grouped bar charts to compare variable distributions across different scenarios. Create scatter plots or heatmaps to visualize the relationships between variables and the target variable.

Insights:

Univariate:

Amt_credit: Most loans fall between 50K and 1L, with a few larger loans exceeding 20L.

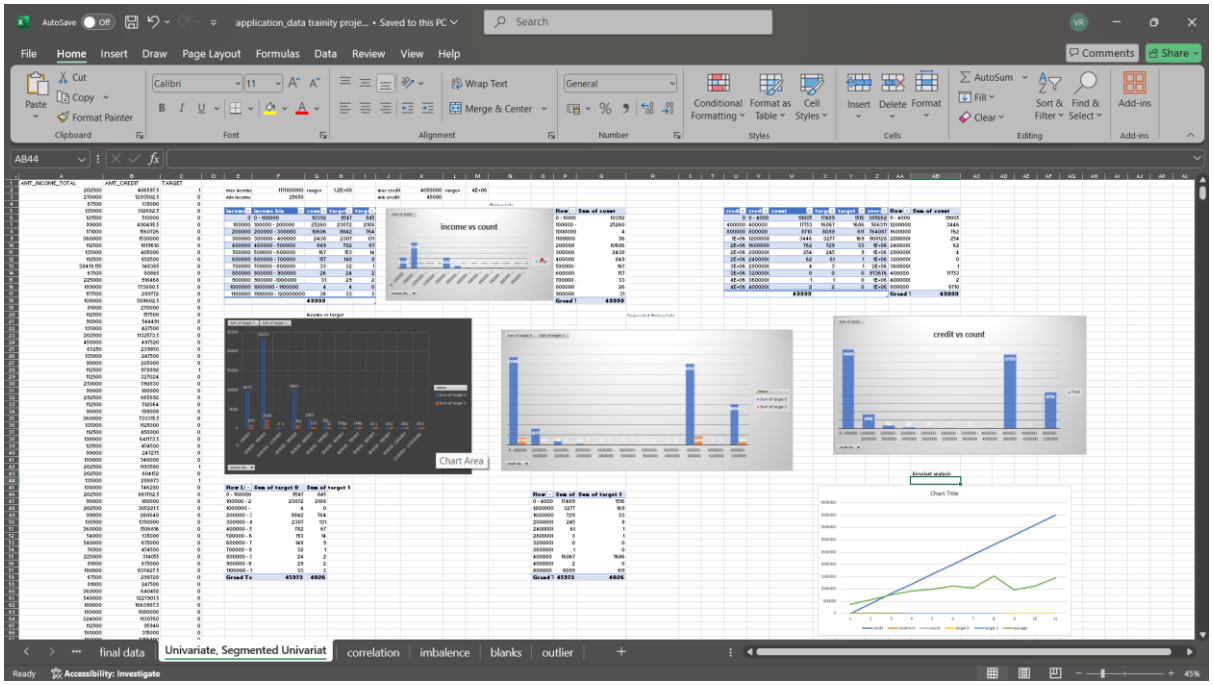
Amt_income_total: Income is within a large proportion earning below 200K annually.

Segmented univariate:

Loan Defaults vs. Non-Defaults ,Applicants with payment difficulties (TARGET=1) tend to have Lower AMT_INCOME_TOTAL values on average.

Bivariate:

A positive correlation between AMT_INCOME_TOTAL and AMT_CREDIT suggests that higher-income applicants tend to apply for higher loan amounts.



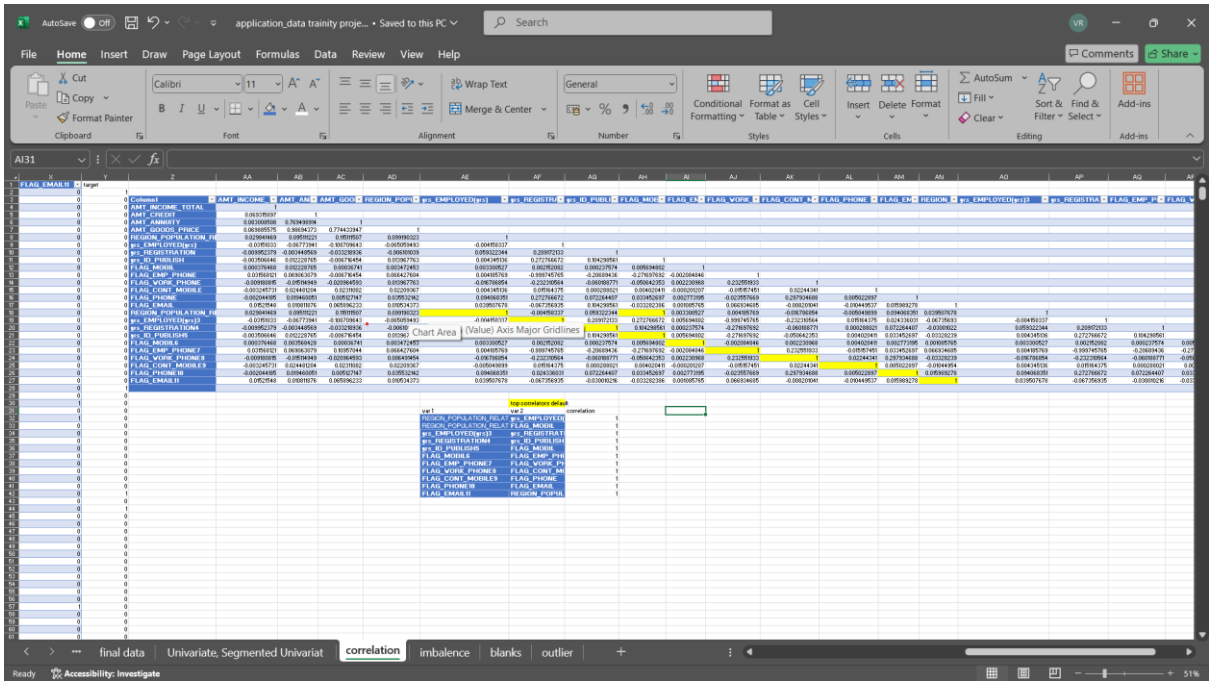
E. **Identify Top Correlations for Different Scenarios:** Understanding the correlation between variables and the target variable can provide insights into strong indicators of loan default.

- **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.
- **Hint:** Utilize Excel functions like CORREL to calculate correlation coefficients between variables and the target variable within each segment. Rank the correlations to identify the top indicators of loan default for each scenario.
- **Graph suggestion:** Create correlation matrices or heatmaps to visualize the correlations between variables within each segment. Highlight the top correlated variables for each scenario using different colors or shading.

Insights:

Correlations :

Many columns have a correlation of 1 which means they are having same values.


application_data trainity proje... • Saved to this PC
Search
File Home Insert Draw Page Layout Formulas Data Review View Help
Comments Share
Cut Copy Format Painter
Calibri 11
Wrap Text
Merge & Center
General
Conditional Formatting
Format as Table
Cell Styles
Insert Delete Format
AutoSum Fill Clear
Sort & Filter
Find & Select
Add-ins
Clipboard Font Alignment Number Styles Cells Editing Add-ins
AI31
FLAG_EMAIL11
target
Column1
AMT_INCOME_TOTAL
AMT_CREDIT
AMT_ANNUITY
AMT_GOODS_PRICE
REGION_POPULATION_RE
gs_EMPLOYED[grp3]
gs_REGISTRATION
gs_ID_PUBLISH
FLAG_MOBIL
FLAG_EMP_PHONE
FLAG_WORK_PHONE
FLAG_CONT_MOBILE
FLAG_PHONE
FLAG_EMAIL
REGION_POPULATION_RE
gs_EMPLOYED[grp3]
gs_REGISTRATION4
gs_ID_PUBLISH5
FLAG_MOBIL6
FLAG_EMP_PHONE7
FLAG_WORK_PHONE8
FLAG_CONT_MOBILE9
FLAG_PHONE10
FLAG_EMAIL11
Chart Area
(Value) Axis Major Gridlines
top correlators default
var 1
var 2
correlation
final data
Univariate, Segmented Univariat
correlation
imbalence
blanks
outlier
Ready
Accessibility: Investigate
51%

Thank You