TRAINITY

BANK LOAN CASE STUDY

Final Project-2

Shahnawaz Akhtar
shahnawazakhtar2@gmail.com

CONTENTS

- ❖Project Description
- ❖Approach
- ❖Tech-Stack Used
- ❖Insights
- ❖Result

SHAHNAWAZ AKHTAR
Shahnawazakhtar2@gmail.com

Project Description

This case study focuses on analyzing data patterns using Exploratory Data Analysis (EDA) to ensure that qualified applicants are not unfairly rejected. The primary goal is to detect trends that may suggest whether a customer could face challenges in repaying their installments. These insights can then guide decisions, such as rejecting a loan application, offering a reduced loan amount, or providing a loan at a higher interest rate to applicants who pose a higher risk.

SHAHNAWAZ AKHTAR
Shahnawazakhtar2@gmail.com

Approach

- **Downloading the Data:** This is very first step to start while doing data analysis, which is downloading or converting the data in suitable format.
- **Understanding the data:** After successfully downloading or converting the data, one need to understand the labels given in the dataset. Also needs to understand problem statement.
- **Cleaning the data:** This is first step of analysis where we need to identify null, duplicate values from the data & clean it with proper process. Also it involves to delete unnecessary column which will not be used during analysis
- **Analyzing the data:** After the cleaning next is to explore the dataset and find answers to the question.
- **Representation:** This is last but important step which will make analysis representative and easy to understand.

SHAHNAWAZ AKHTAR
Shahnawazakhtar2@gmail.com

Tech Stack Used

- Microsoft Excel 2019 (for working, analysis purpose)
- Presentation (for presentation purpose)

SHAHNAWAZ AKHTAR
Shahnawazakhtar2@gmail.com

Data Description

The dataset provided for analysis includes loan application data from a finance company with two types of scenarios:

1. Customers with payment difficulties: Those who were late by more than 'x' days on at least one of the first 'y' installments.
2. Other cases: Where payments were made on time. The dataset includes three files:
    - previous_application.csv: Contains information about previous loan applications.
    - application_data.csv: Provides details about the current loan applications.
    - columns_description.csv: Describes the columns present in the other datasets, explaining what each column represents.

SHAHNAWAZ AKHTAR
Shahnawazakhtar2@gmail.com

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

The initial dataset had 123 columns and 50,000 rows. To handle the missing data, I first calculated the percentage of null values for each column. Based on this, I removed columns with more than 30% missing data. Additionally, I deleted some columns that were not useful. For the remaining columns with missing values, I used the median method to fill in gaps for numerical data.
To improve analysis, I also converted the DAYS_BIRTH as DAYS_BIRTH(YEAR) DAYS_EMPLOYED as DAYS_EMPLOYED(YEAR), DAYS_REGISTRATION as DAYS_REGISTRATION (YEAR), and DAYS_ID_PUBLISH as DAYS_ID_PUBLISH(YEAR) columns from days to years.

SHAHNAWAZ AKHTAR
Shahnawazakhtar2@gmail.com

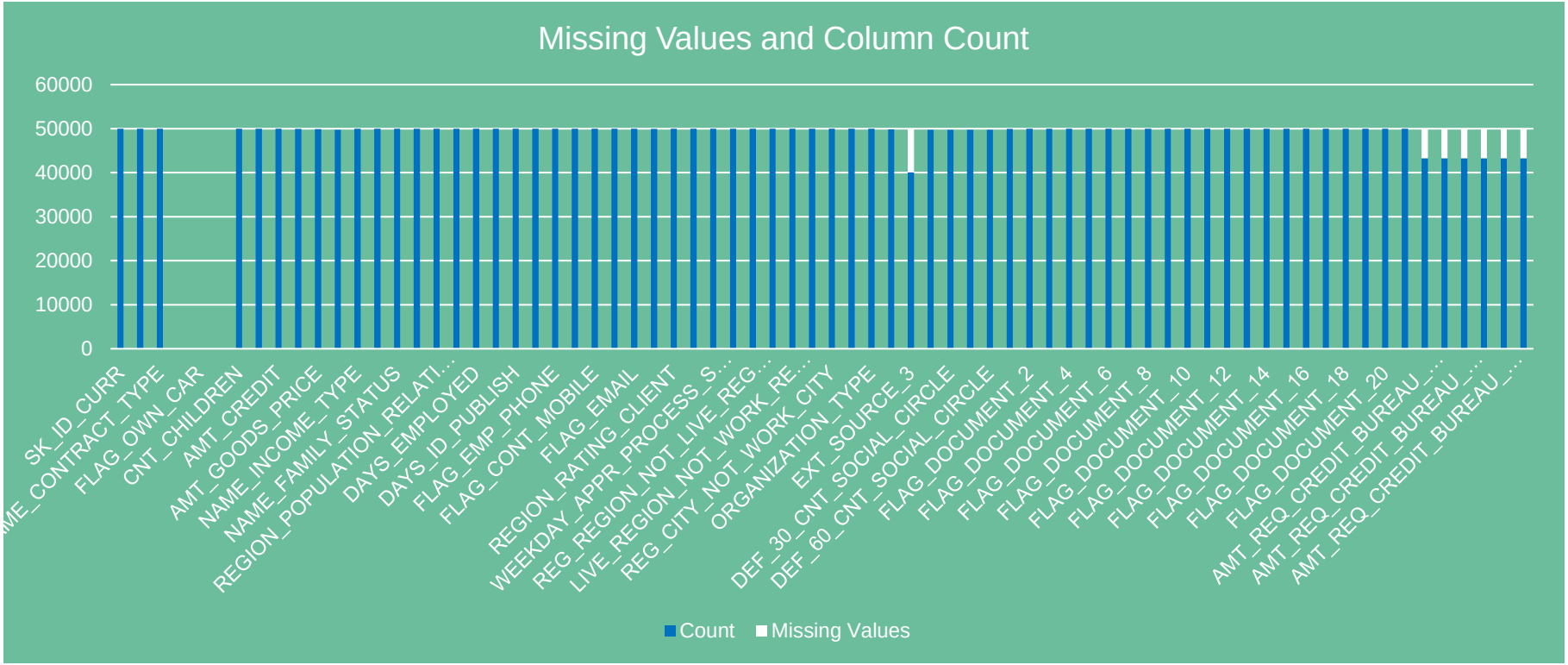
Columns	Count	Missing Values
SK_ID_CURR	49999	0
TARGET	49999	0
NAME_CONTRACT_TYPE	49999	0
CODE_GENDER	0	0
FLAG_OWN_CAR	0	0
FLAG_OWN_REALTY	0	0
CNT_CHILDREN	49999	0
AMT_INCOME_TOTAL	49999	0
AMT_CREDIT	49999	0
AMT_ANNUITY	49998	1
AMT_GOODS_PRICE	49961	38
NAME_TYPE_SUITE	49807	192
NAME_INCOME_TYPE	49999	0
NAME_EDUCATION_TYPE	49999	0
NAME_FAMILY_STATUS	49999	0
NAME_HOUSING_TYPE	49999	0
REGION_POPULATION_RELATIVE	49999	0
DAYS_BIRTH	49999	0
DAYS_EMPLOYED	49999	0

Columns	Count	Missing Values
DAYS_REGISTRATION	49999	0
DAYS_ID_PUBLISH	49999	0
FLAG_MOBIL	49999	0
FLAG_EMP_PHONE	49999	0
FLAG_WORK_PHONE	49999	0
FLAG_CONT_MOBILE	49999	0
FLAG_PHONE	49999	0
FLAG_EMAIL	49999	0
CNT_FAM_MEMBERS	49998	1
REGION_RATING_CLIENT	49999	0
REGION_RATING_CLIENT_W_CITY	49999	0
WEEKDAY_APPR_PROCESS_START	49999	0
HOUR_APPR_PROCESS_START	49999	0
REG_REGION_NOT_LIVE_REGION	49999	0
REG_REGION_NOT_WORK_REGION	49999	0
LIVE_REGION_NOT_WORK_REGION	49999	0
REG_CITY_NOT_LIVE_CITY	49999	0
REG_CITY_NOT_WORK_CITY	49999	0
LIVE_CITY_NOT_WORK_CITY	49999	0
ORGANIZATION_TYPE	49999	0
EXT_SOURCE_2	49873	126
EXT_SOURCE_3	40055	9944
OBS_30_CNT_SOCIAL_CIRCLE	49831	168
DEF_30_CNT_SOCIAL_CIRCLE	49831	168
OBS_60_CNT_SOCIAL_CIRCLE	49831	168
DEF_60_CNT_SOCIAL_CIRCLE	49831	168
DAYS_LAST_PHONE_CHANGE	49998	1
FLAG_DOCUMENT_2	49999	0
FLAG_DOCUMENT_3	49999	0
FLAG_DOCUMENT_4	49999	0

Columns	Count	Missing Values
FLAG_DOCUMENT_5	49999	0
FLAG_DOCUMENT_6	49999	0
FLAG_DOCUMENT_7	49999	0
FLAG_DOCUMENT_8	49999	0
FLAG_DOCUMENT_9	49999	0
FLAG_DOCUMENT_10	49999	0
FLAG_DOCUMENT_11	49999	0
FLAG_DOCUMENT_12	49999	0
FLAG_DOCUMENT_13	49999	0
FLAG_DOCUMENT_14	49999	0
FLAG_DOCUMENT_15	49999	0
FLAG_DOCUMENT_16	49999	0
FLAG_DOCUMENT_17	49999	0
FLAG_DOCUMENT_18	49999	0
FLAG_DOCUMENT_19	49999	0
FLAG_DOCUMENT_20	49999	0
FLAG_DOCUMENT_21	49999	0
AMT_REQ_CREDIT_BUREAU_HOUR	43265	6734
AMT_REQ_CREDIT_BUREAU_DAY	43265	6734
AMT_REQ_CREDIT_BUREAU_WEEK	43265	6734
AMT_REQ_CREDIT_BUREAU_MON	43265	6734
AMT_REQ_CREDIT_BUREAU_QRT	43265	6734
AMT_REQ_CREDIT_BUREAU_YEAR	43265	6734

SHAHNAWAZ AKHTAR
Shahnawazakhtar2@gmail.com

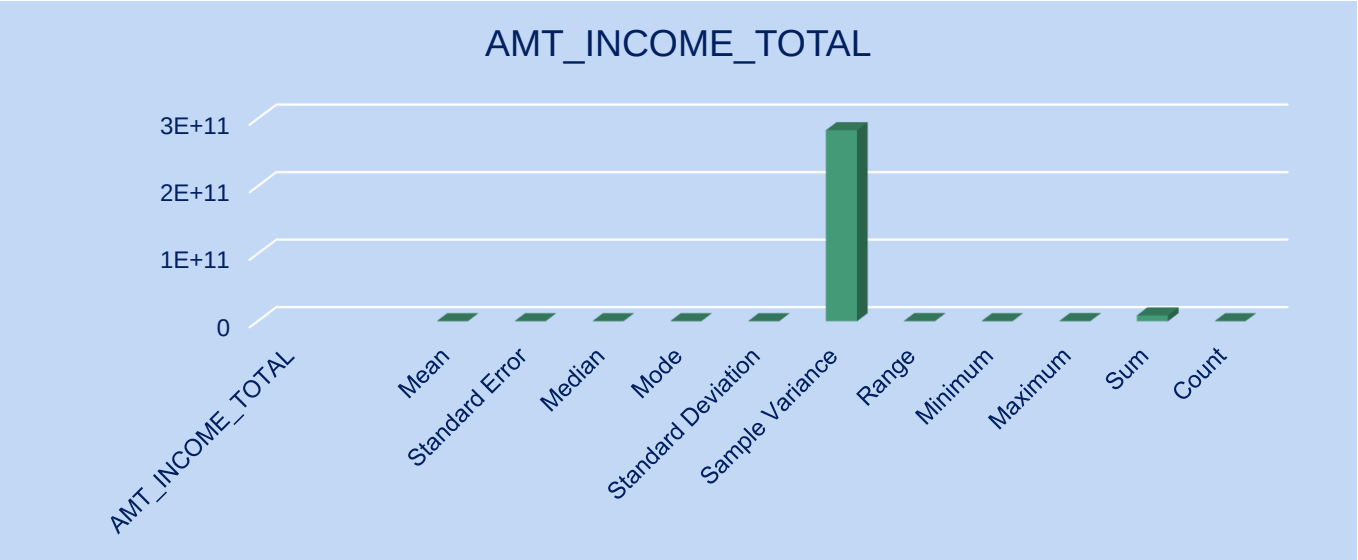
Graphical Representation



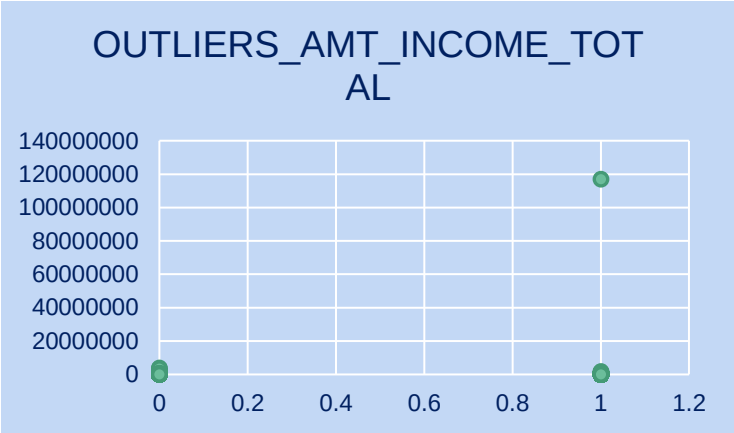
SHAHNAWAZ AKHTAR
Shahnawazakhtar2@gmail.com

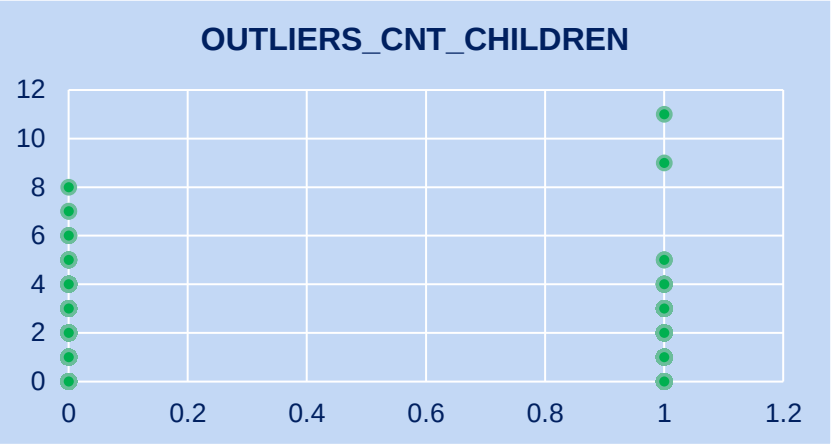
Task-B: Identify Outliers in the Dataset: Outliers can significantly impact the analysis and distort the results. You need to identify outliers in the loan application dataset.

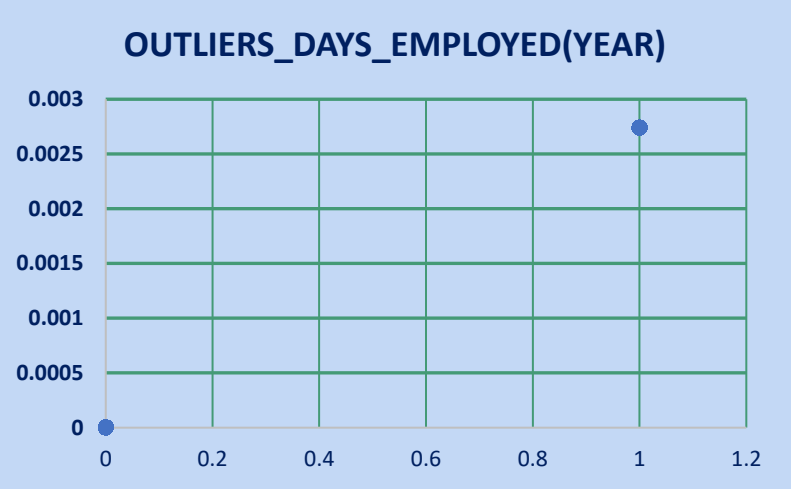
Quartile 1	Quartile 3	Inter Quartile Range	Upper Limit	Upper Limit
112500	202500	90000	337500	-22500



SHAHNAWAZ AKHTAR
Shahnawazakhtar2@gmail.com

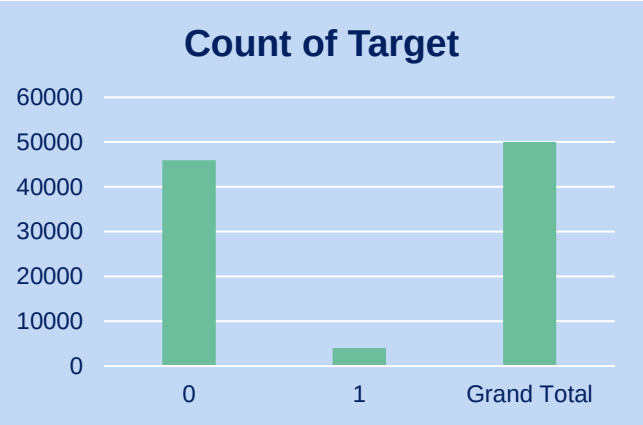


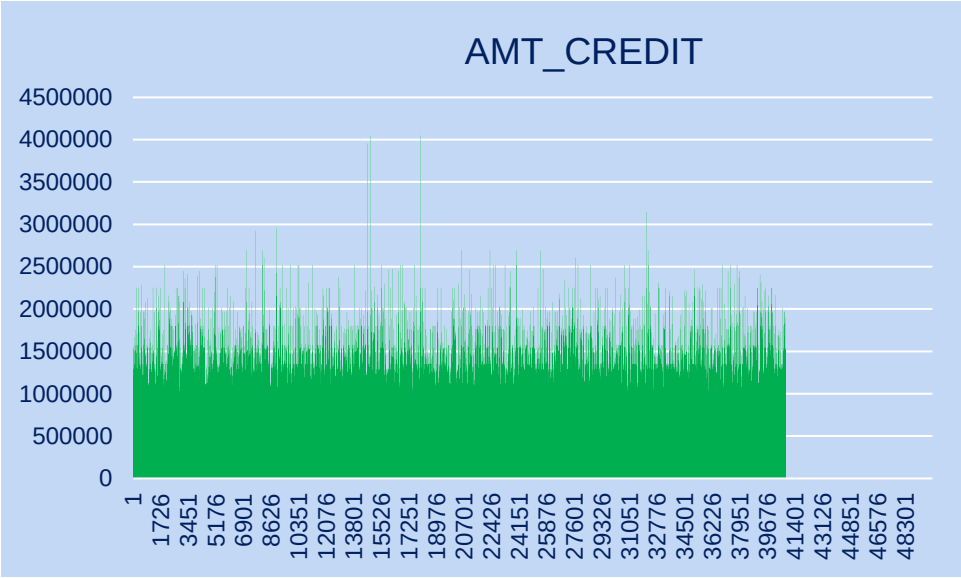




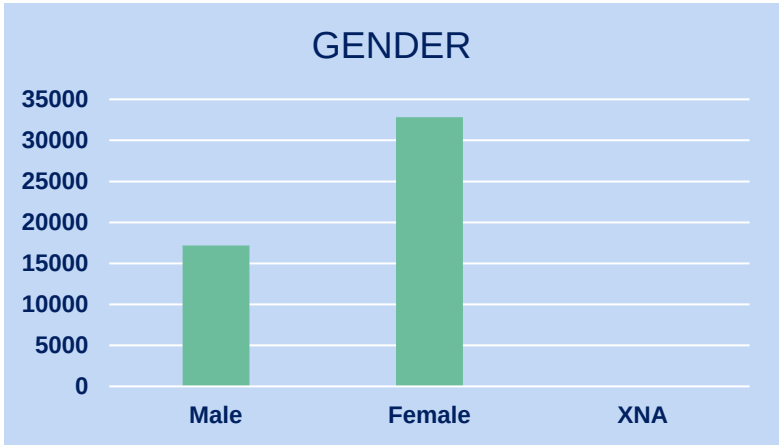
SHAHNAWAZ AKHTAR
Shahnawazakhtar2@gmail.com

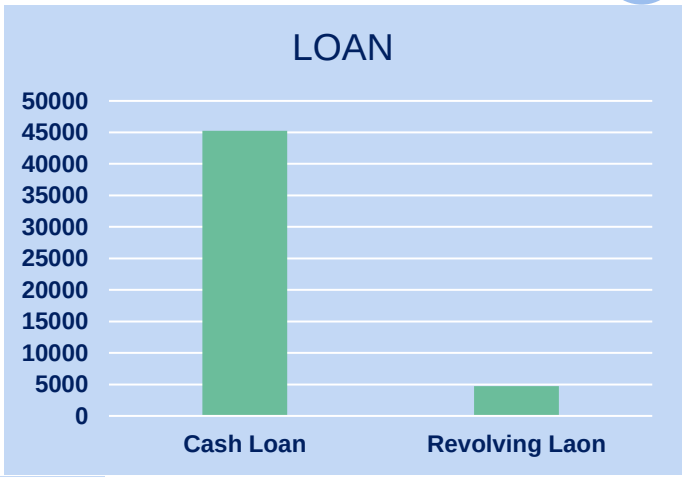
Task-C. Analyze Data Imbalance: Data imbalance can affect the accuracy of the analysis, especially for binary classification problems. Understanding the data distribution is crucial for building reliable models.

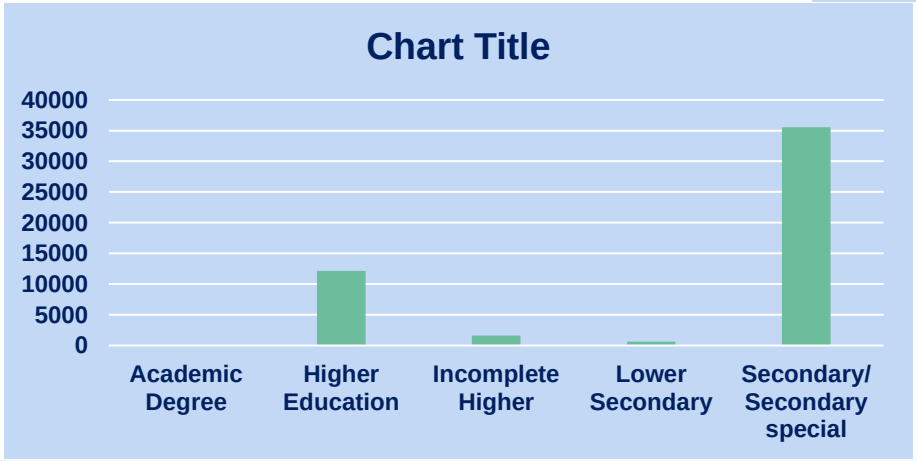




SHAHNAWAZ AKHTAR
Shahnawazakhtar2@gmail.com





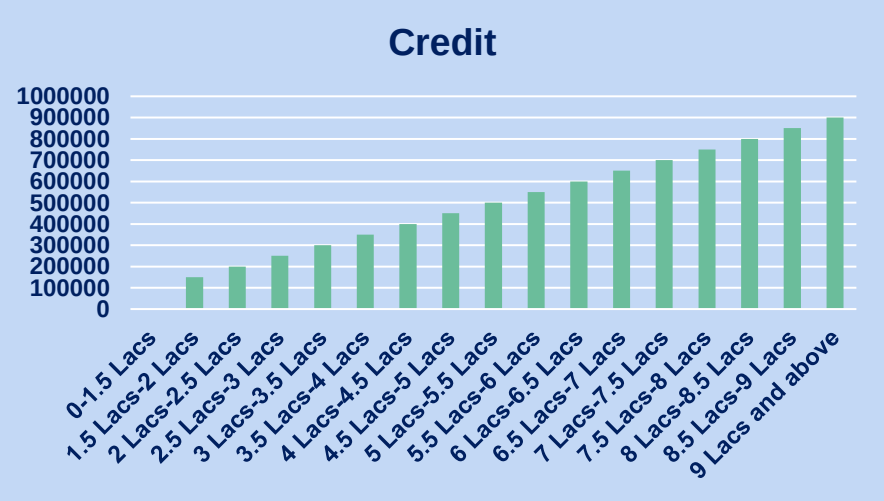


SHAHNAWAZ AKHTAR
Shahnawazakhtar2@gmail.com

Task-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.

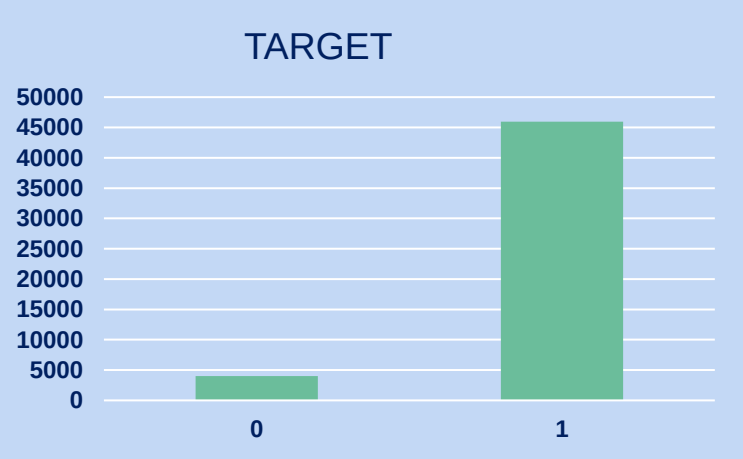
Univariate Analysis: People with target 1 has largely staggered income as compared to target 0

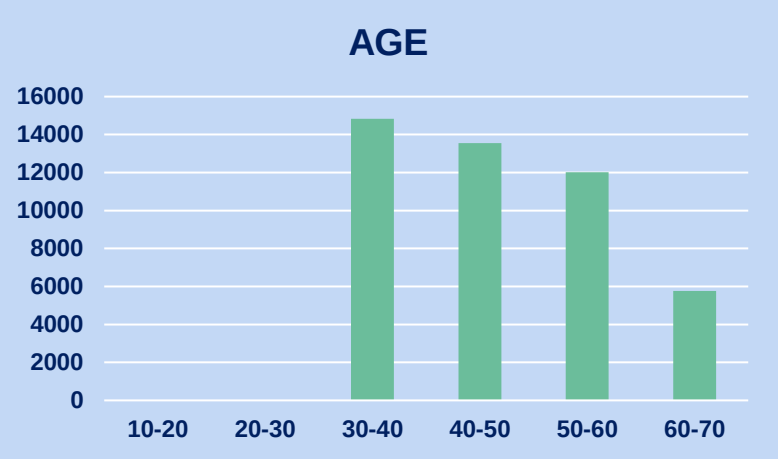
- Plots highlights that people having difficulties to pay back with respect to their income ,loan amount and cost of goods against which loan and annuity is prepared.



SHAHNAWAZ AKHTAR
Shahnawazakhtar2@gmail.com

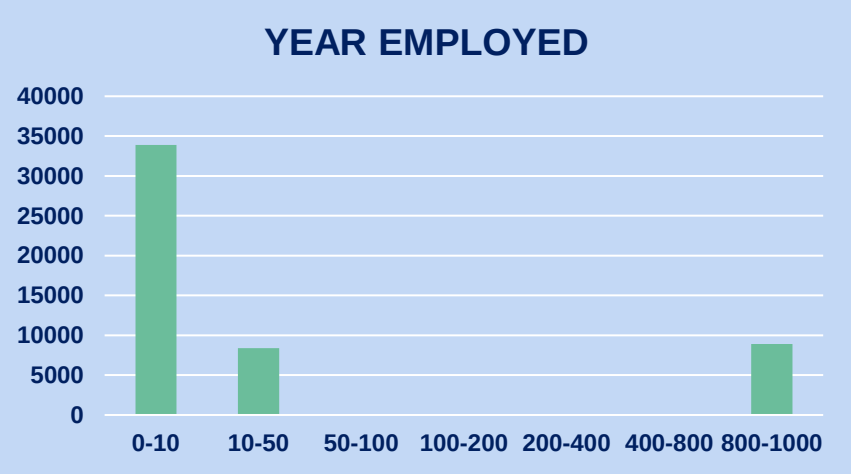
Univariate Analysis

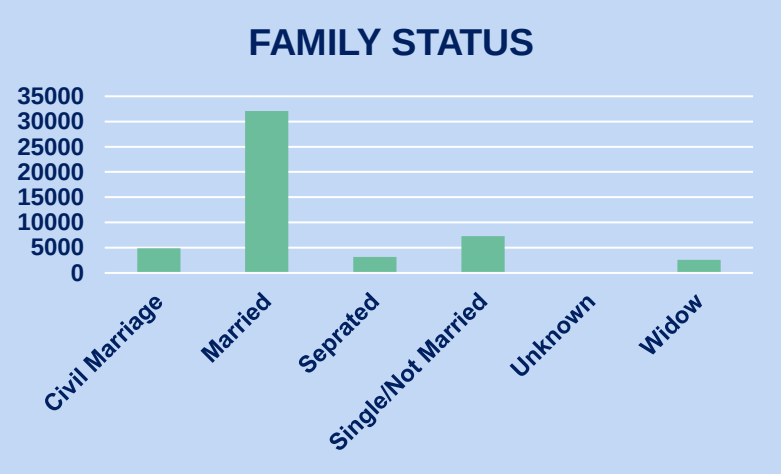




SHAHNAWAZ AKHTAR
Shahnawazakhtar2@gmail.com

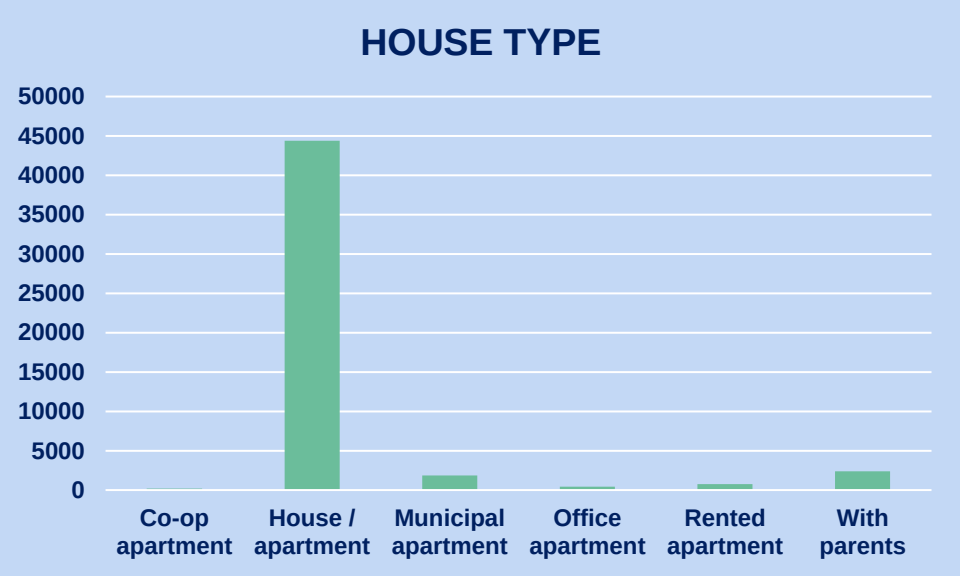
Univariate Analysis





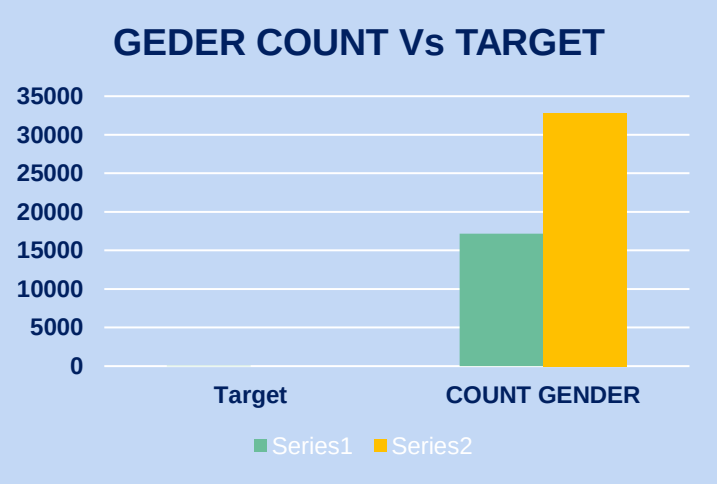
SHAHNAWAZ AKHTAR
Shahnawazakhtar2@gmail.com

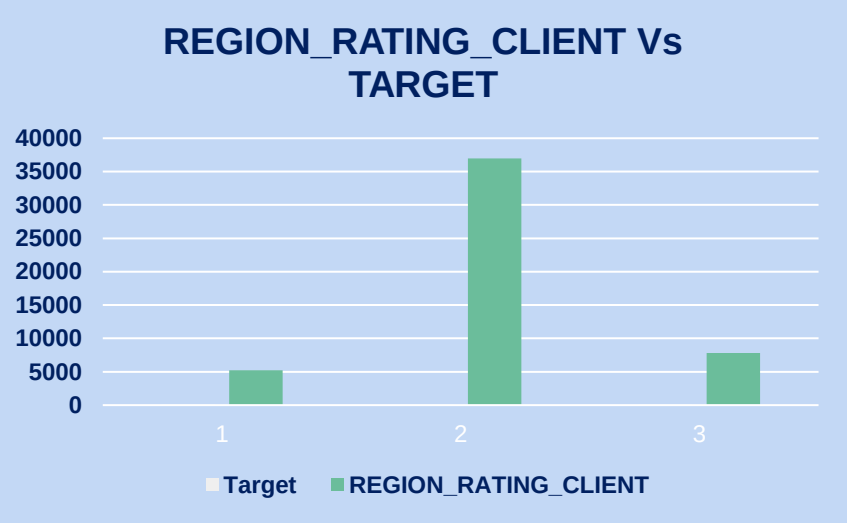
Univariate Analysis



SHAHNAWAZ AKHTAR
Shahnawazakhtar2@gmail.com

Bivariate Analysis





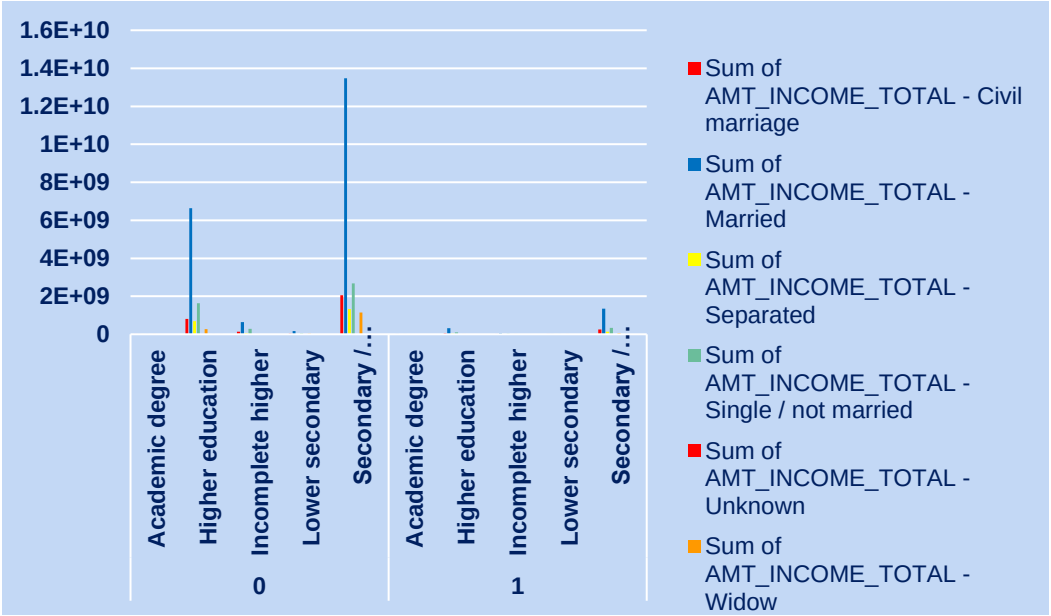
SHAHNAWAZ AKHTAR
Shahnawazakhtar2@gmail.com

Bivariate Analysis

BIVARIATE ANALYSIS: NUMERICAL & CATEGORICAL TARGET VARIABLES:

Widow client with academic degree have few outliers and don't have first and third quartile for target 0.

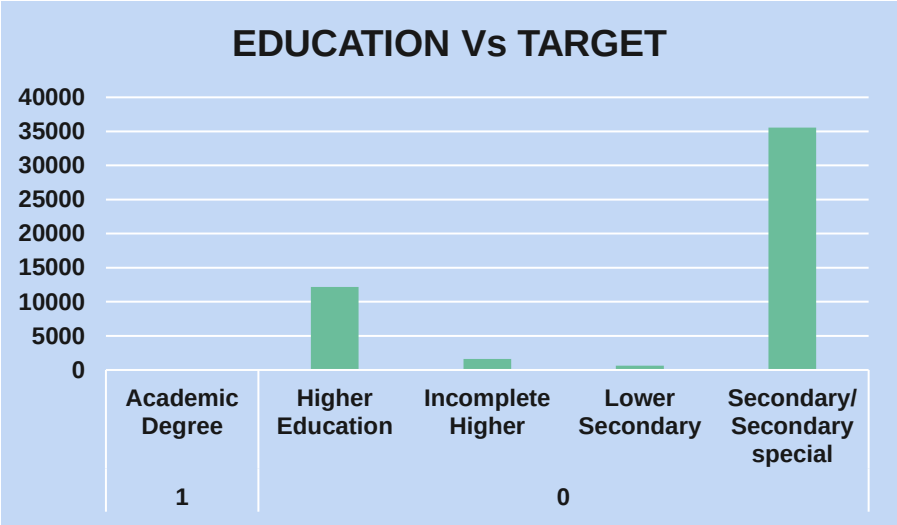
For AMT_INCOME_TOTAL & TARGET 0,1: Income of clients with all types of family status having rest of the education type l below 25%.



1. Clients who haven't completed their degree are more likely to have higher income.
2. Some clients with Secondary and Secondary special to have higher incomes.

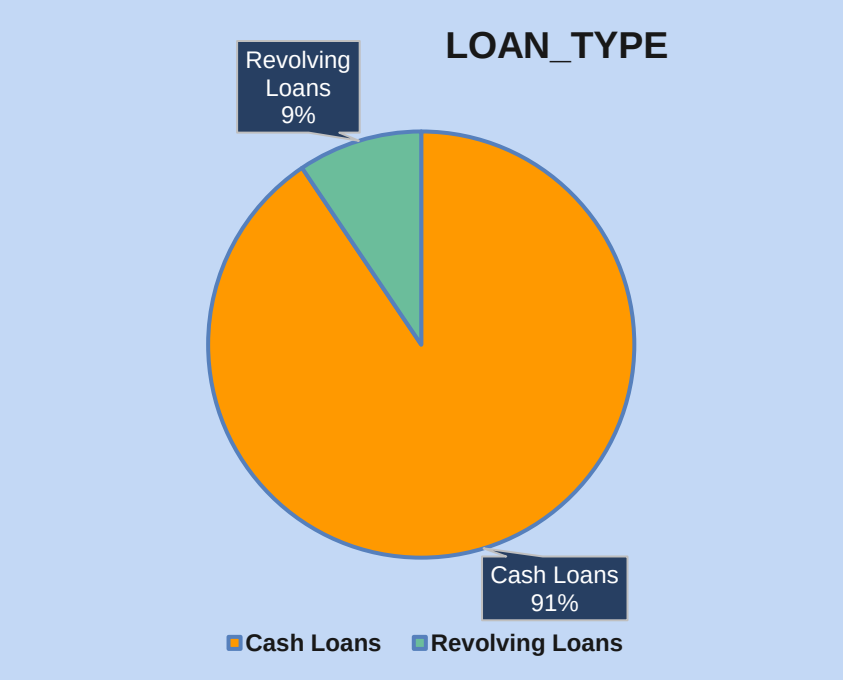
SHAHNAWAZ AKHTAR
Shahnawazakhtar2@gmail.com

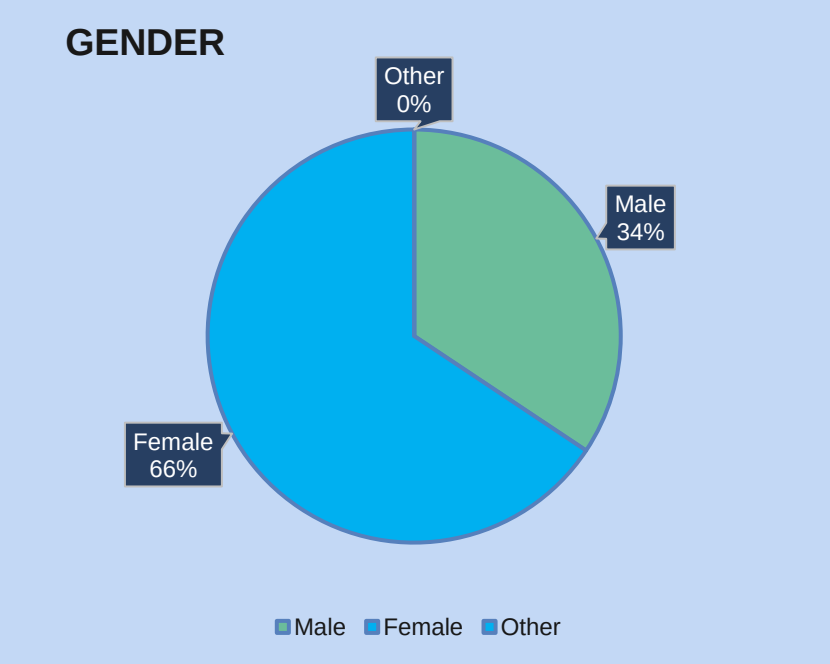
Bivariate Analysis



SHAHNAWAZ AKHTAR
Shahnawazakhtar2@gmail.com

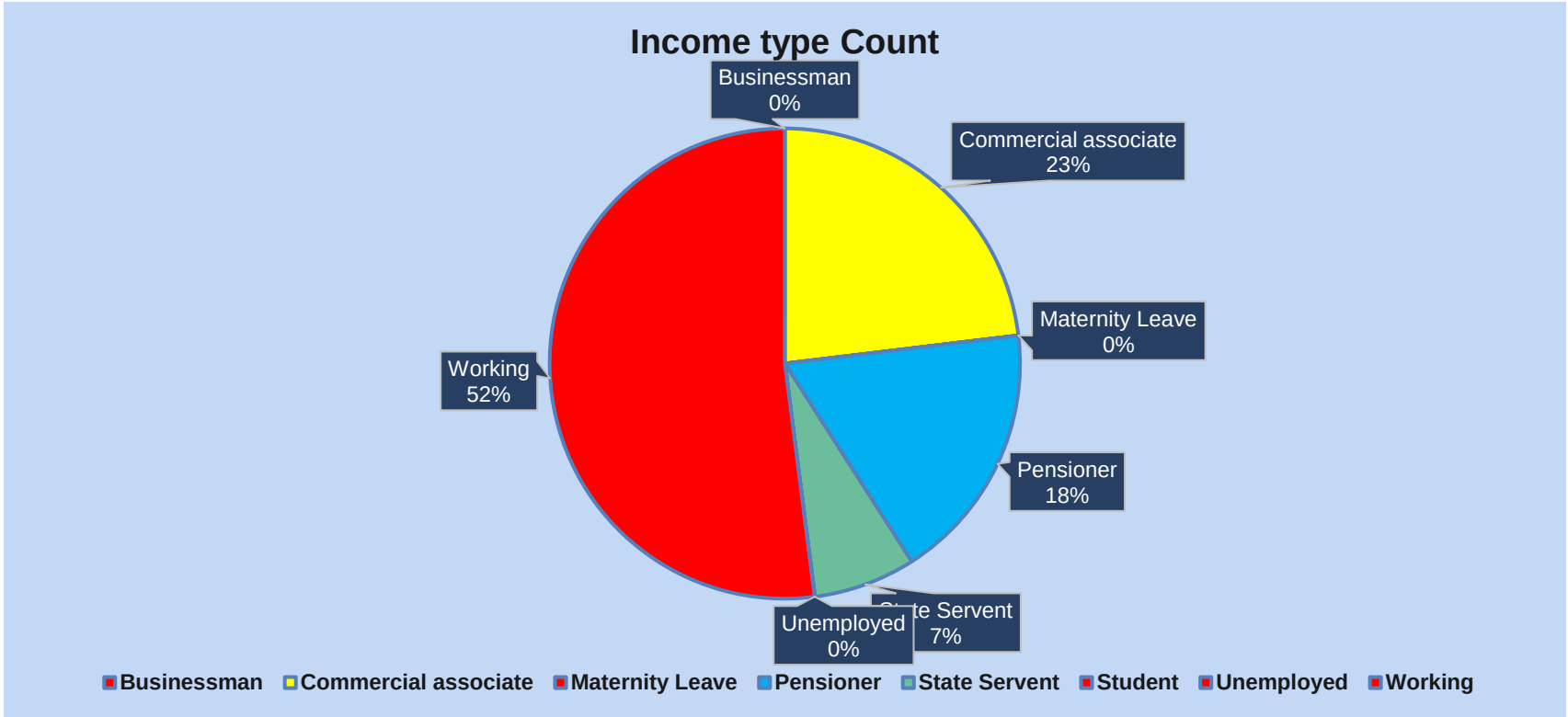
Segmented Univariate Analysis:





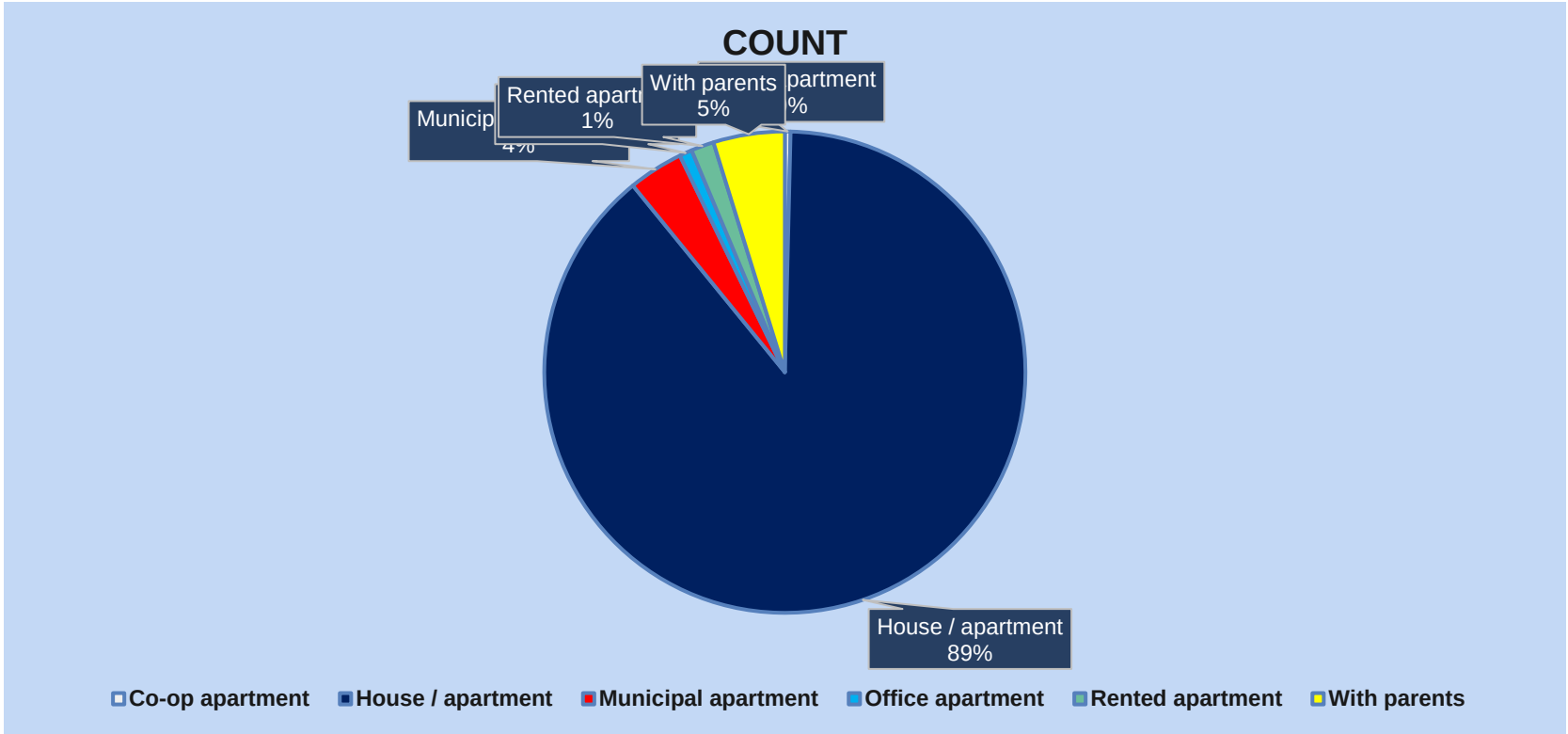
SHAHNAWAZ AKHTAR
Shahnawazakhtar2@gmail.com

Segmented Univariate Analysis:



SHAHNAWAZ AKHTAR
Shahnawazakhtar2@gmail.com

Segmented Univariate Analysis:



SHAHNAWAZ AKHTAR
Shahnawazakhtar2@gmail.com

Task-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.

- Target-0 correlations with other cases

	CORRELATION FOR APPLICANTS WITH PAYMENT MADE ON TIME							
CNT_CHILDREN	1	0.036319722	0.005705458	-0.024912809	-0.335876269	-0.245521512	0.032537221	-0.004878357
AMT_INCOME_TOTAL	0.036319722	1	0.377965752	0.181941261	-0.073769425	-0.161680938	-0.032286356	-0.009966134
AMT_CREDIT	0.005705458	0.377965752	1	0.095539444	0.051084182	-0.074733443	0.008290189	0.003871232
REGION_POPULATION_RELATIVE	-0.024912809	0.181941261	0.095539444	1	0.030435419	-0.006767142	0.002236288	0.002424316
DAYS_BIRTH(YEARS)	-0.335876269	-0.073769425	0.051084182	0.030435419	1	0.623474675	0.270073313	0.005605472
DAYS_EMPLOYED (YEARS)	-0.245521512	-0.161680938	-0.07473344	-0.006767142	0.623474675	1	0.274516224	0.01067213
DAYS_ID_PUBLISH (YEARS)	0.032537221	-0.032286356	0.008290189	0.002236288	0.270073313	0.274516224	1	-0.00338445
REGION_RATING_CLIENT	-0.004878357	-0.009966134	0.003871232	0.002424316	0.005605472	0.01067213	-0.00338445	1
	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	REGION_POPULATION_RELATIVE	DAYS_BIRTH(YEARS)	DAYS_EMPLOYED (YEARS)	DAYS_ID_PUBLISH (YEARS)	REGION_RATING_CLIENT

SHAHNAWAZ AKHTAR
Shahnawazakhtar2@gmail.com

Task-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.

- Target-1 correlations with other cases

	CORRELATION FOR APPLICANTS WITH PAYMENT MADE ON TIME							
CNT_CHILDREN	1	0.010110177	0.007601905	-0.020359154	-0.2496732	-0.189773227	0.042360717	0.055515557
AMT_INCOME_TOTAL	0.010110177	1	0.015271444	-0.006180303	-0.009033662	-0.011758681	0.009122006	-0.012846697
AMT_CREDIT	0.007601905	0.015271444	1	0.067775624	0.142506035	0.018782223	0.043771901	-0.045024534
REGION_POPULATION_RELATIVE	-0.020359154	-0.006180303	0.067775624	1	0.016468731	0.007710059	0.005118563	-0.430032303
DAYS_BIRTH(YEARS)	-0.2496732	-0.009033662	0.142506035	0.016468731	1	0.588242824	0.247896571	-0.045027112
DAYS_EMPLOYED (YEARS)	-0.189773227	-0.011758681	0.018782223	0.007710059	0.588242824	1	0.232661912	-0.009237108
DAYS_ID_PUBLISH (YEARS)	0.042360717	0.009122006	0.043771901	0.005118563	0.247896571	0.232661912	1	-0.025335227
REGION_RATING_CLIENT	0.055515557	-0.012846697	-0.04502453	-0.430032303	-0.045027112	-0.009237108	-0.025335227	1
	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	REGION_POPULATION_RELATIVE	DAYS_BIRTH(YEARS)	DAYS_EMPLOYED (YEARS)	DAYS_ID_PUBLISH (YEARS)	REGION_RATING_CLIENT

SHAHNAWAZ AKHTAR
Shahnawazakhtar2@gmail.com

KEY INSIGHTS

With this analysis, we can identify several key factors that will help us assess whether a client is likely to default or not. These factors are outlined below:

- As Age and Years of Experience increase, the likelihood of defaulting decreases. Therefore, the bank should prioritize older and more experienced clients.
- Clients with higher education levels are less likely to default compared to those with only a Lower Secondary or Secondary education.
- Male clients have a higher tendency to default than female clients.
- Corporate clients are a safer bet compared to labor-class clients.
- People from Region Rating 3 have the highest percentage of defaulters. The bank could consider implementing stricter loan policies for clients from this region. Clients from Region 1 are the safest option.
- We also notice that as clients age, the loan amounts they borrow tend to be considerably higher. Since the default rate is lower among older clients, they should be viewed as low-risk and highly profitable for the bank.

SHAHNAWAZ AKHTAR
Shahnawazakhtar2@gmail.com

RESULT

This project involved extensive use of Excel. The major challenge was working with such huge data. This project helped me understand how to work with huge datasets. This helped me understand how 2 datasets are merged to analyze the details. The dataset involved a lot of missing data and outliers, handling them was a task and this project helped me understand what to how and why of handling the outliers and Null values. The project also helped me discover new add-ins such as data analyze.

SHAHNAWAZ AKHTAR
Shahnawazakhtar2@gmail.com

THANKS!

SHAHNAWAZ AKHTAR

shahnawazakhtar2@gmail.com
+91 8299591611