Bank Loan Case Study

Excel File Link

Project by Mayur Rajput

Project's Agenda

1. Project description
2. Approach & Tech-Stack Used
3. Data Cleaning
4. Data Analysis
5. Interactive Dashboard
6. Insights & Results

Excel File Link

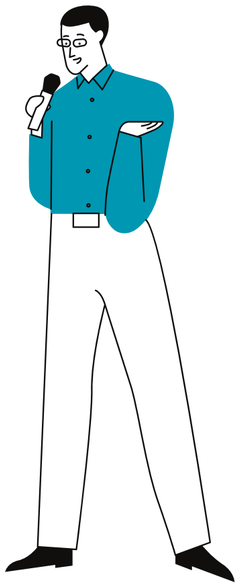
Project Description

- The goal of this project is to identify patterns that indicate if a customer will have difficulty in paying their installments. This information can be used to make decisions such as denying the loan, reducing the amount of loan, or lending at a higher interest rate to risky applicants. The company wants to understand the key factors behind loan default so it can make better decisions about loan approval.

- The goal of this project is to use Exploratory Data Analysis (EDA) to analyze patterns in the data and ensure that capable applicants are not rejected.

Approach

1. Identifying Missing Data and Handling it
2. Identifying Outliers
3. Analyzing Data Imbalance
4. Performing Univariate, Segmented Univariate, and Bivariate Analysis
5. Identifying Top Correlations



Tech Stack Used

Microsoft Excel

Microsoft Excel is used to identify and handle missing values, outliers, Analyzing data imbalance, Performing Univariate, Segmented Univariate, and Bivariate Analysis, Identify Top Correlations and use Exploratory Data Analysis (EDA) to analyze patterns in the data and ensure that capable applicants are not rejected.

Canva

Canva is used to prepare this presentation

Task 1 : Identifying and Handling missing values in Dataset

a) Identifying Missing values

Finding Missing Values					Total Rows =	49999	Total Column =	122					
0	0	0	0	0	0	0	0	0	1	38	192	0	
0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.1%	0.4%	0.0%	
SK_ID_CUI	TARG	NAME_CONTRACT_TYI	CODE_GENDI	FLAG_OWN_CA	FLAG_OWN_REALT	CNT_CHILDRE	AMT_INCOME_TOTA	AMT_CRED	AMT_ANNUI	AMT_GOODS_PRI	NAME_TYPE_SUI	NAME_INCOME_TYI	NAME_E
100002	1	Cash loans	M	N	Y	0	202500	406597.5	24700.5	351000	Unaccompanied	Working	Seconda
100003	0	Cash loans	F	N	N	0	270000	1293502.5	35698.5	1129500	Family	State servant	Higher e
100004	0	Revolving loans	M	Y	Y	0	67500	135000	6750	135000	Unaccompanied	Working	Seconda
100006	0	Cash loans	F	N	Y	0	135000	312682.5	29686.5	297000	Unaccompanied	Working	Seconda
100007	0	Cash loans	M	N	Y	0	121500	513000	21865.5	513000	Unaccompanied	Working	Seconda
100008	0	Cash loans	M	N	Y	0	99000	490495.5	27517.5	454500	Spouse, partner	State servant	Seconda
100009	0	Cash loans	F	Y	Y	1	171000	1560726	41301	1395000	Unaccompanied	Commercial associate	Higher e
100010	0	Cash loans	M	Y	Y	0	360000	1530000	42075	1530000	Unaccompanied	State servant	Higher e
100011	0	Cash loans	F	N	Y	0	112500	1019610	33826.5	913500	Children	Pensioner	Seconda
100012	0	Revolving loans	M	N	Y	0	135000	405000	20250	405000	Unaccompanied	Working	Seconda
100014	0	Cash loans	F	N	Y	1	112500	652500	21177	652500	Unaccompanied	Working	Higher e
100015	0	Cash loans	F	N	Y	0	38419.155	148365	10678.5	135000	Children	Pensioner	Seconda
100016	0	Cash loans	F	N	Y	0	67500	80865	5881.5	67500	Unaccompanied	Working	Second
100017	0	Cash loans	M	Y	N	1	225000	918468	28966.5	697500	Unaccompanied	Working	Second
100018	0	Cash loans	F	N	Y	0	189000	773680.5	32778	679500	Unaccompanied	Working	Second
100019	0	Cash loans	M	Y	Y	0	157500	299772	20160	247500	Family	Working	Second
100020	0	Cash loans	M	N	N	0	108000	509602.5	26149.5	387000	Unaccompanied	Working	Second
100021	0	Revolving loans	F	N	Y	1	81000	270000	13500	270000	Unaccompanied	Working	Seconda
100022	0	Revolving loans	F	N	Y	0	112500	157500	7875	157500	Other_A	Working	Seconda
100023	0	Cash loans	F	N	Y	1	90000	544491	17563.5	454500	Unaccompanied	State servant	Higher e
100024	0	Revolving loans	M	Y	Y	0	135000	427500	21375	427500	Unaccompanied	Working	Seconda
100025	0	Cash loans	F	Y	Y	1	202500	1132573.5	37561.5	927000	Unaccompanied	Commercial associate	Seconda
100026	0	Cash loans	F	N	N	1	450000	497520	32521.5	450000	Unaccompanied	Working	Seconda

Task 1 : Identifying and Handling missing values in Dataset

b) Data Cleaning

- Removed 49 columns having missing values greater than 40%.
- 28 columns which are not required for further analysis are removed such as FLAG_MOBIL, FLAG_EMAIL, EXT_SOURCE_1,EXT_SOURCE_3, FLAG_DOCUMENT_2, FLAG_DOCUMENT_21
- 45 Columns left after performing data cleaning
- In DAYS_BIRTH, DAYS_EMPLOYED, DAYS_REGISTRATION, DAYS_ID_PUBLISH, DAYS_LAST_PHONE_CHANGE columns Negative values are replaced with positive values.

Finding Missing Values			Total Rows =	49999	Total Column =	45						
0	0	0	0	0	0	0	1	38	192	0	0	0
0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.1%	0.4%	0.0%	0.0%	0.0%
NAME_CONTRACT_TYP	CODE_GENDE	FLAG_OWN_CA	FLAG_OWN_REAL	CNT_CHILDRE	AMT_INCOME_TOTA	AMT_CRED	AMT_ANNUI	AMT_GOODS_PRIC	NAME_TYPE_SUI	NAME_INCOME_TYP	NAME_EDUCATION_TYPE	NAME_FAMILY_STATU
Cash loans	M	N	Y	0	202500	406597.5	24700.5	351000	Unaccompanied	Working	Secondary / secondary special	Single / not married
Cash loans	F	N	N	0	270000	1293502.5	35698.5	1129500	Family	State servant	Higher education	Married
Revolving loans	M	Y	Y	0	67500	135000	6750	135000	Unaccompanied	Working	Secondary / secondary special	Single / not married
Cash loans	F	N	Y	0	135000	312682.5	29686.5	297000	Unaccompanied	Working	Secondary / secondary special	Civil marriage
Cash loans	M	N	Y	0	121500	513000	21865.5	513000	Unaccompanied	Working	Secondary / secondary special	Single / not married
Cash loans	M	N	Y	0	99000	490495.5	27517.5	454500	Spouse, partner	State servant	Secondary / secondary special	Married
Cash loans	F	Y	Y	1	171000	1560726	41301	1395000	Unaccompanied	Commercial associate	Higher education	Married
Cash loans	M	Y	Y	0	360000	1530000	42075	1530000	Unaccompanied	State servant	Higher education	Married
Cash loans	F	N	Y	0	112500	1019610	33826.5	913500	Children	Pensioner	Secondary / secondary special	Married
Revolving loans	M	N	Y	0	135000	405000	20250	405000	Unaccompanied	Working	Secondary / secondary special	Single / not married
Cash loans	F	N	Y	1	112500	652500	21177	652500	Unaccompanied	Working	Higher education	Married
Cash loans	F	N	Y	0	38419.155	148365	10678.5	135000	Children	Pensioner	Secondary / secondary special	Married
Cash loans	F	N	Y	0	67500	80865	5881.5	67500	Unaccompanied	Working	Secondary / secondary special	Married
Cash loans	M	Y	N	1	225000	918468	28966.5	697500	Unaccompanied	Working	Secondary / secondary special	Married
Cash loans	F	N	Y	0	189000	773680.5	32778	679500	Unaccompanied	Working	Secondary / secondary special	Married
Cash loans	M	Y	Y	0	157500	299772	20160	247500	Family	Working	Secondary / secondary special	Single / not married
Cash loans	M	N	N	0	108000	509602.5	26149.5	387000	Unaccompanied	Working	Secondary / secondary special	Married
Revolving loans	F	N	Y	1	81000	270000	13500	270000	Unaccompanied	Working	Secondary / secondary special	Married
Revolving loans	F	N	Y	0	112500	157500	7875	157500	Other_A	Working	Secondary / secondary special	Widow
Cash loans	F	N	Y	1	90000	544491	17563.5	454500	Unaccompanied	State servant	Higher education	Single / not married
Revolving loans	M	Y	Y	0	135000	427500	21375	427500	Unaccompanied	Working	Secondary / secondary special	Married
Cash loans	F	Y	Y	1	202500	1132573.5	37561.5	927000	Unaccompanied	Commercial associate	Secondary / secondary special	Married
Cash loans	F	N	N	1	450000	497520	32521.5	450000	Unaccompanied	Working	Secondary / secondary special	Married
Cash loans	F	N	Y	0	83250	239850	23850	225000	Unaccompanied	Pensioner	Secondary / secondary special	Married

Task 1 : Identifying and Handling missing values in Dataset

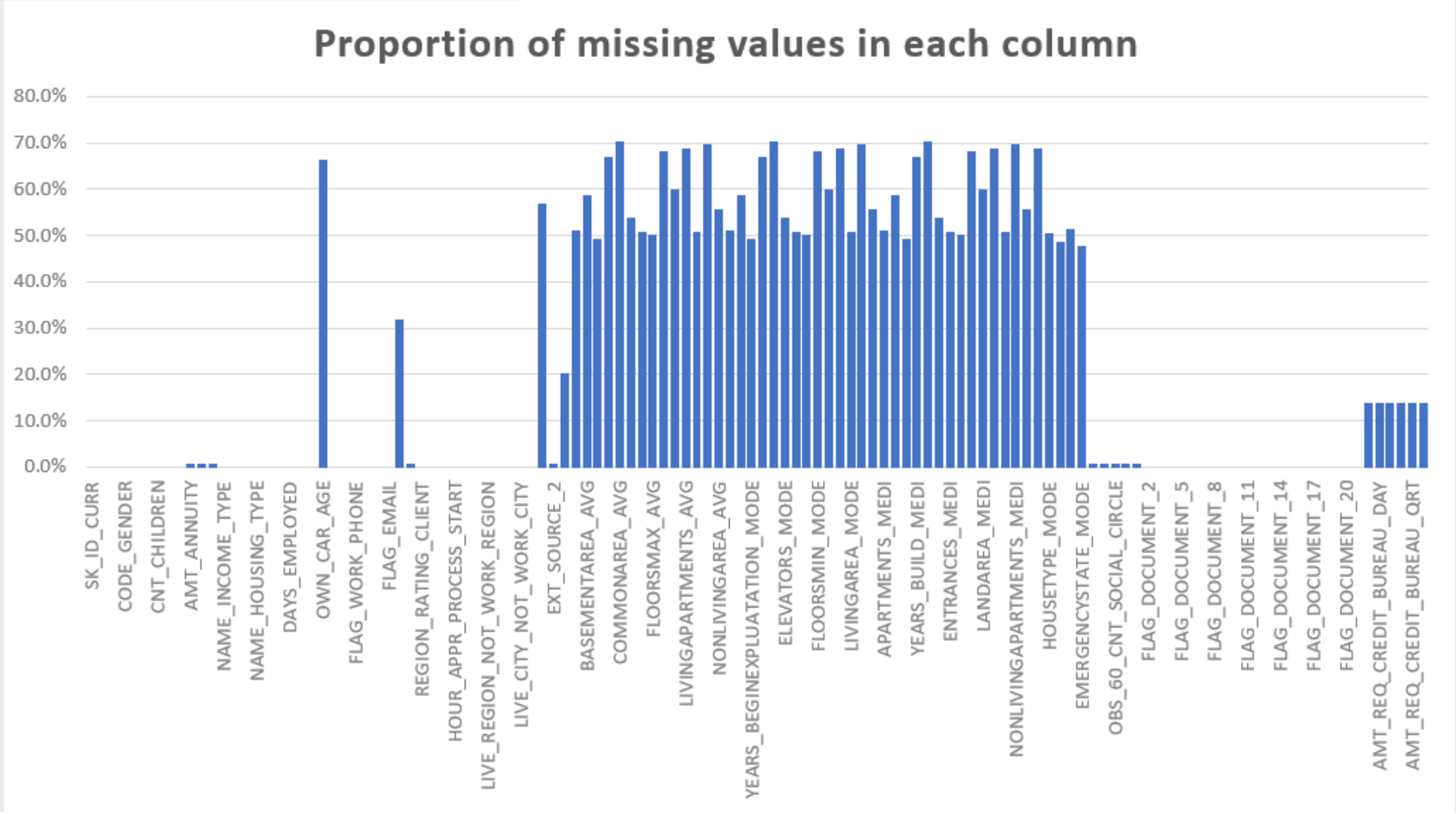
c) Handling Missing Data

- Average() function used to replace missing values in following columns : AMT_ANNUITY, DAYS_LAST_PHONE_CHANGE, AMT_GOODS_PRICE,
- Mode() function used to replace missing values in following columns : NAME_TYPE_SUITE, OCCUPATION_TYPE, CNT_FAM_MEMBERS, OBS_30_CNT_SOCIAL_CIRCLE, DEF_30_CNT_SOCIAL_CIRCLE, OBS_60_CNT_SOCIAL_CIRCLE, DEF_60_CNT_SOCIAL_CIRCLE, AMT_REQ_CREDIT_BUREAU_HOUR, AMT_REQ_CREDIT_BUREAU_DAY, AMT_REQ_CREDIT_BUREAU_WEEK, AMT_REQ_CREDIT_BUREAU_MON, AMT_REQ_CREDIT_BUREAU_QRT, AMT_REQ_CREDIT_BUREAU_YEAR

Handling Missing Data					Total Rows =	49999	Total Column =	45				
0	0	0	0	0	0	0	0	0	0	0	0	0
0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
									27107	539060	Unaccompanied	
									Avg() = 27107	Avg() = 539060	Mode () = Unaccompanied	
SK_ID_CUR	TARG	NAME_CONTRACT_TYP	CODE_GENDE	FLAG_OWN_CA	FLAG_OWN_REAL	CNT_CHILDRE	AMT_INCOME_TOTA	AMT_CRED	AMT_ANNUI	AMT_GOODS_PRI	NAME_TYPE_SUI	NAME_INCOME_TYP	NAME_E
100002	1	Cash loans	M	N	Y	0	202500	406597.5	24700.5	351000	Unaccompanied	Working	Seconda
100003	0	Cash loans	F	N	N	0	270000	1293502.5	35698.5	1129500	Family	State servant	Higher e
100004	0	Revolving loans	M	Y	Y	0	67500	135000	6750	135000	Unaccompanied	Working	Seconda
100006	0	Cash loans	F	N	Y	0	135000	312682.5	29686.5	297000	Unaccompanied	Working	Seconda
100007	0	Cash loans	M	N	Y	0	121500	513000	21865.5	513000	Unaccompanied	Working	Seconda
100008	0	Cash loans	M	N	Y	0	99000	490495.5	27517.5	454500	Spouse, partner	State servant	Seconda
100009	0	Cash loans	F	Y	Y	1	171000	1560726	41301	1395000	Unaccompanied	Commercial associate	Higher e
100010	0	Cash loans	M	Y	Y	0	360000	1530000	42075	1530000	Unaccompanied	State servant	Higher e
100011	0	Cash loans	F	N	Y	0	112500	1019610	33826.5	913500	Children	Pensioner	Seconda
100012	0	Revolving loans	M	N	Y	0	135000	405000	20250	405000	Unaccompanied	Working	Seconda
100014	0	Cash loans	F	N	Y	1	112500	652500	21177	652500	Unaccompanied	Working	Higher e
100015	0	Cash loans	F	N	Y	0	38419.155	148365	10678.5	135000	Children	Pensioner	Second
100016	0	Cash loans	F	N	Y	0	67500	80865	5881.5	67500	Unaccompanied	Working	Second
100017	0	Cash loans	M	Y	N	1	225000	918468	28966.5	697500	Unaccompanied	Working	Second
100018	0	Cash loans	F	N	Y	0	189000	773680.5	32778	679500	Unaccompanied	Working	Second

Task 1 : Identifying and Handling missing values in Dataset

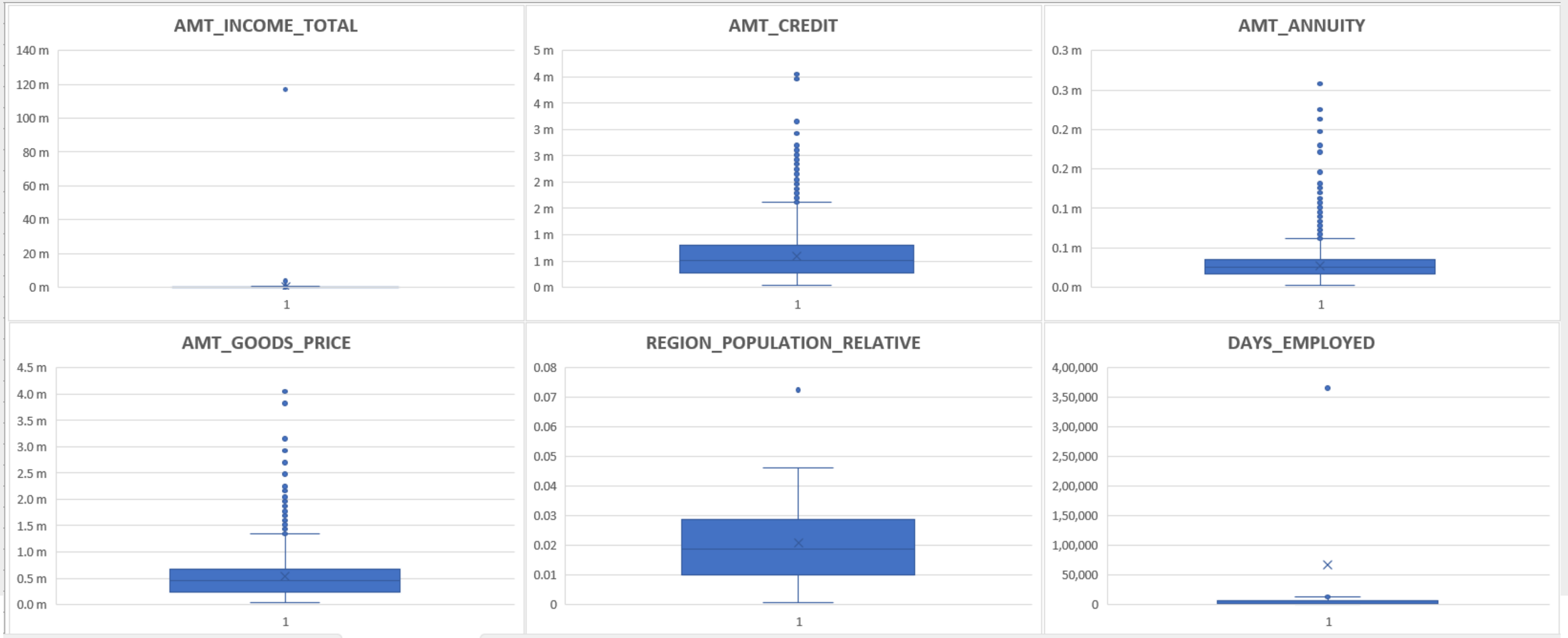
d) Proportion of missing values for each column



Task 2 : Identify Outliers in the Dataset

a) Identify outliers in numerical variables

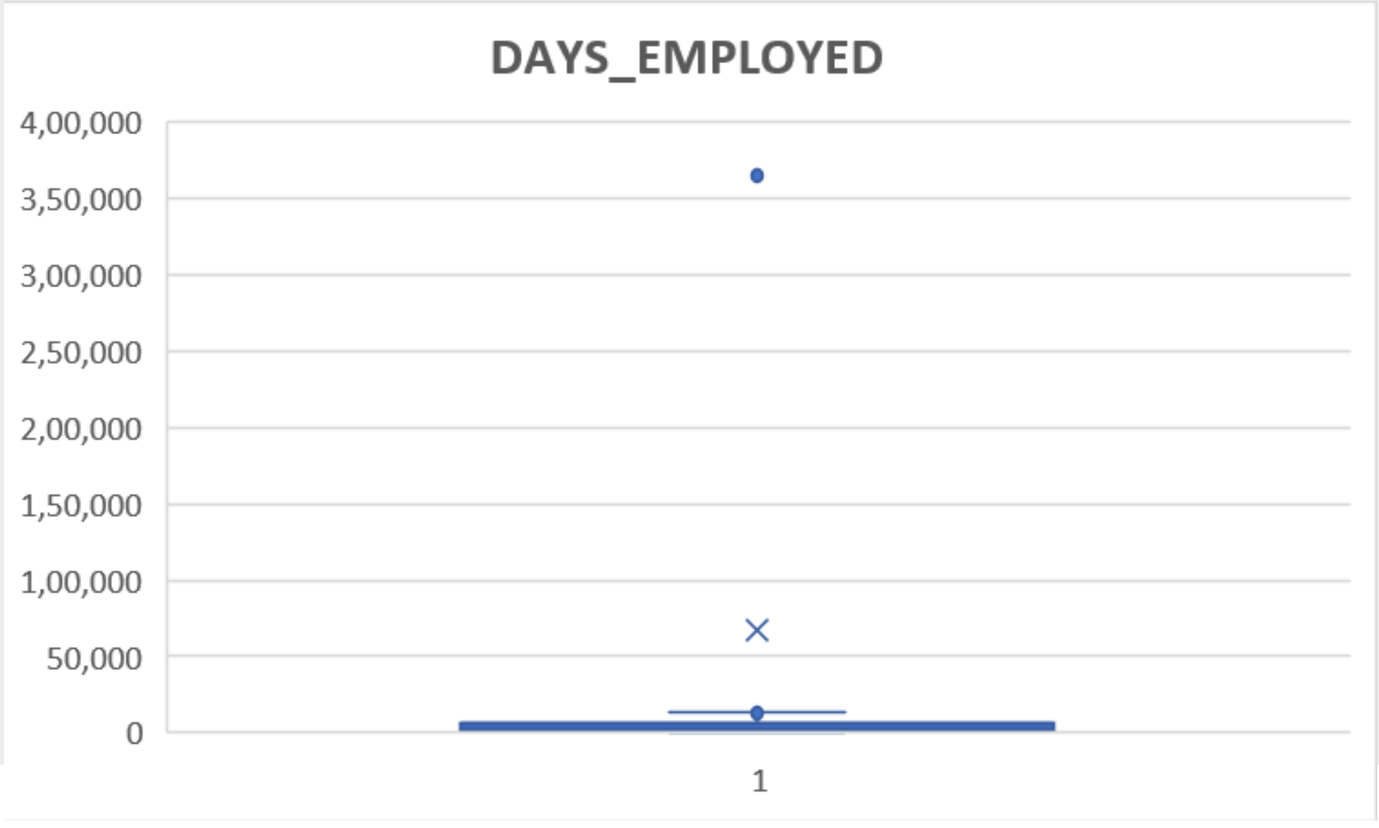
- Outliers found in following numerical columns : AMT_INCOME_TOTAL, AMT_CREDIT, AMT_ANNUITY, AMT_GOODS_PRICE, REGION_POPULATION_RELATIVE, DAYS_EMPLOYED, DAYS_REGISTRATION,DAYS_LAST_PHONE_CHANGE



Task 2 : Identify Outliers in the Dataset

b) Identify if the outliers are valid data points or require further investigation

- Outliers found in following columns could be valid data points: AMT_INCOME_TOTAL, AMT_CREDIT, AMT_ANNUITY, AMT_GOODS_PRICE, REGION_POPULATION_RELATIVE,DAYS_REGISTRATION,DAYS_LAST_PHONE_CHANGE
- Outliers found in following columns may require further investigation: DAYS_EMPLOYED

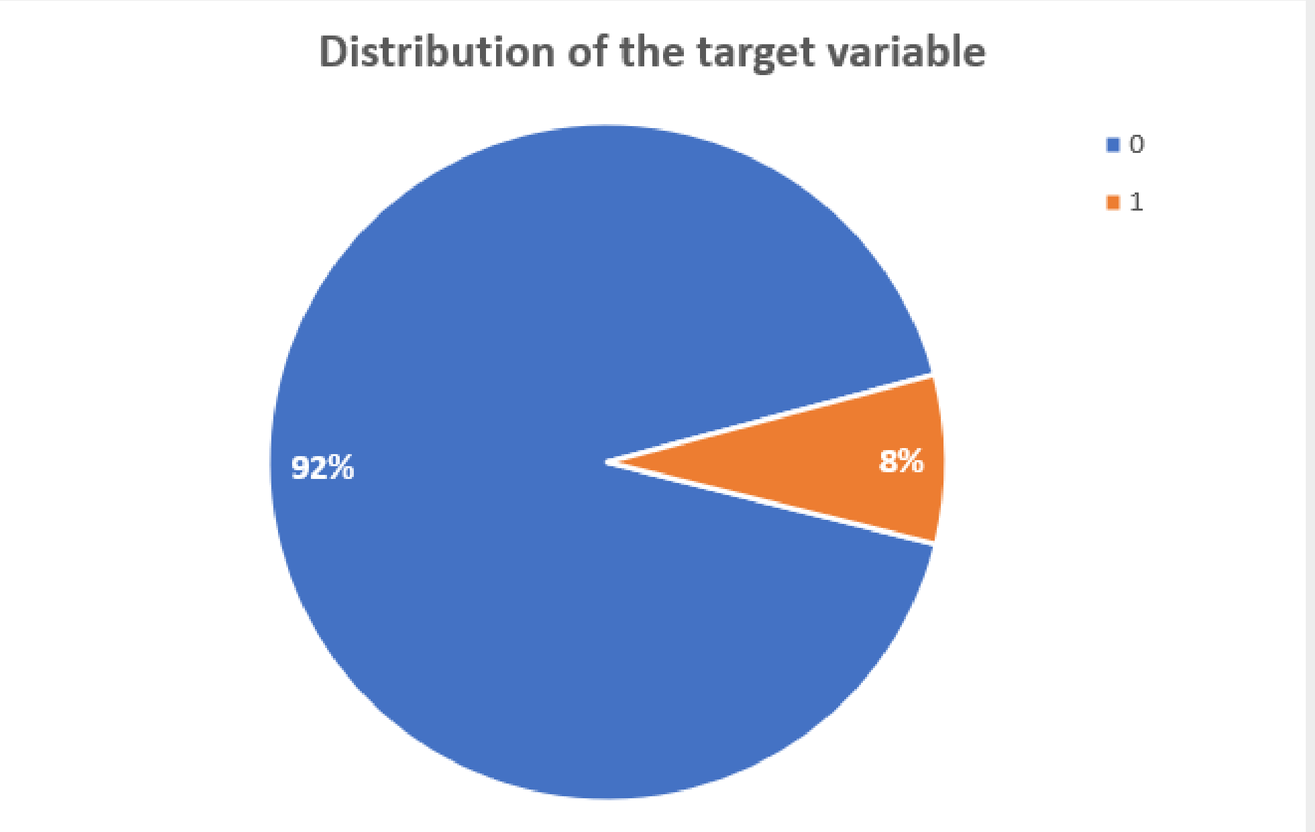


Task 3 : Analyze Data Imbalance

a) Determine data imbalance and calculate the ratio of data imbalance

TARGET	Count of TARGET
0	45973
1	4026
Grand Total	**49999**
Data Imbalance Ratio =	**11.4**

b) Distribution of the target variable

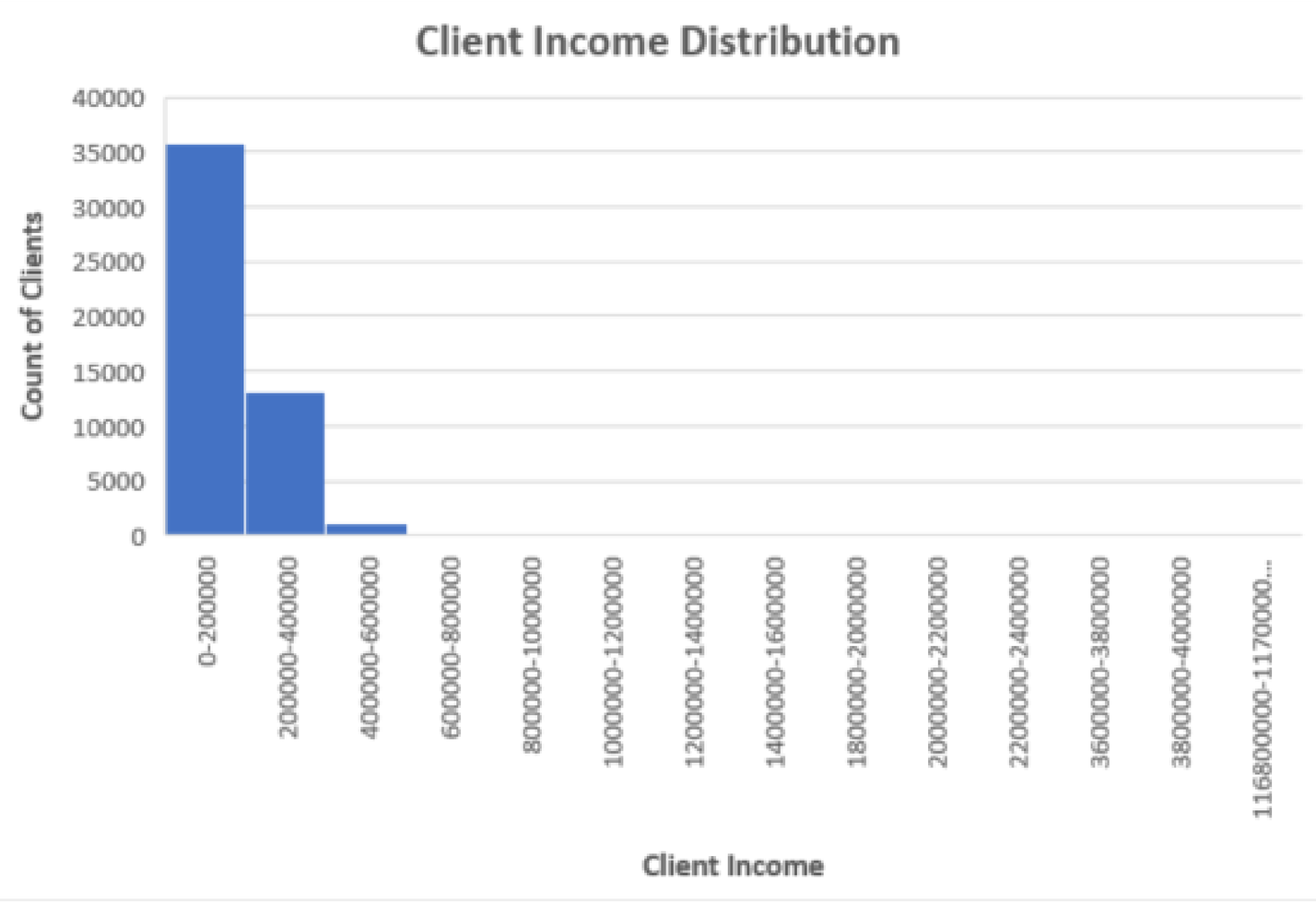


- Distribution of the target variable shows high data imbalance since 8% clients have payment difficulties and 66% clients are Female

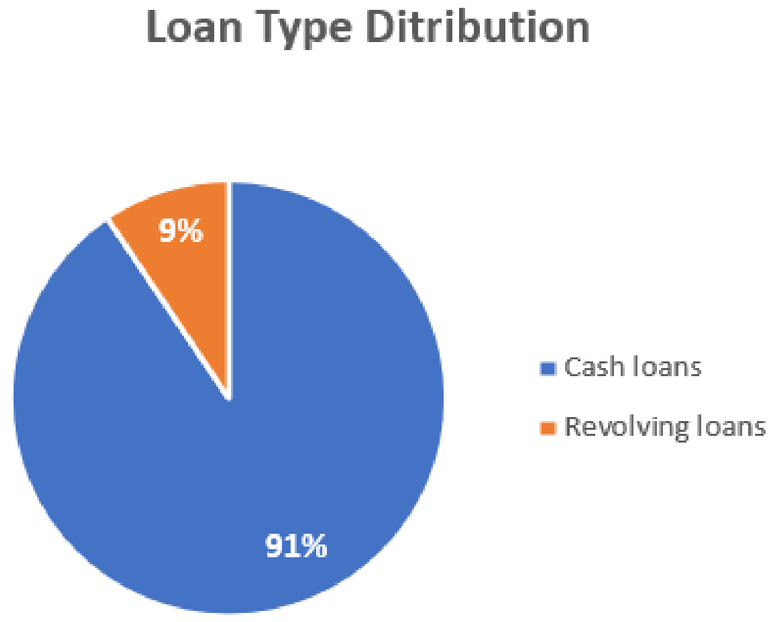
Task 4 : Perform Univariate, Segmented Univariate, and Bivariate Analysis

a) Univariate analysis

Perform univariate analysis to understand the distribution of individual variables

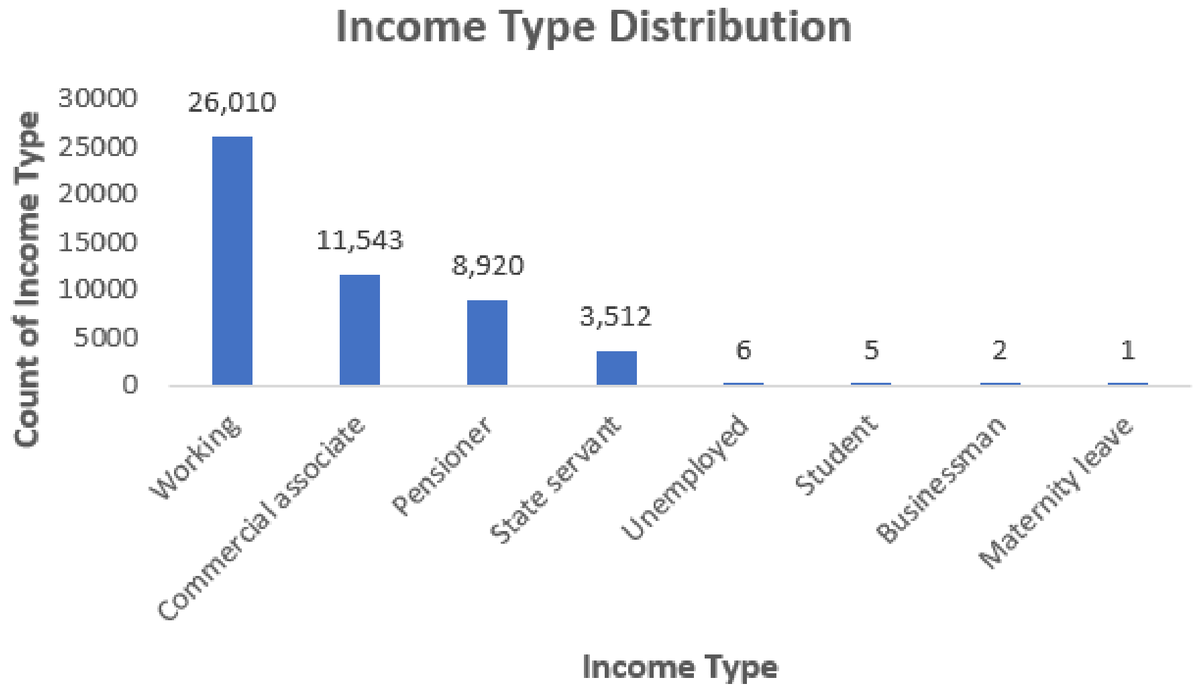


Type of Loan	Count of NAME_CONTRACT_TYPE
Cash loans	45276
Revolving loans	4723
Grand Total	**49999**



Task 4 : Perform Univariate, Segmented Univariate, and Bivariate Analysis

a) Univariate analysis

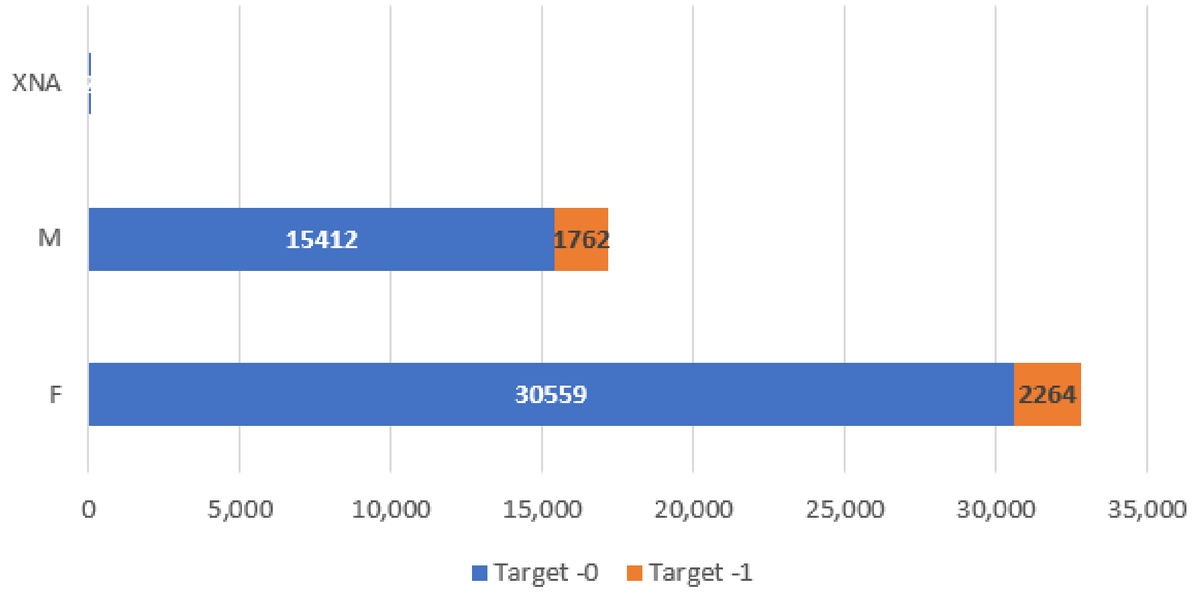


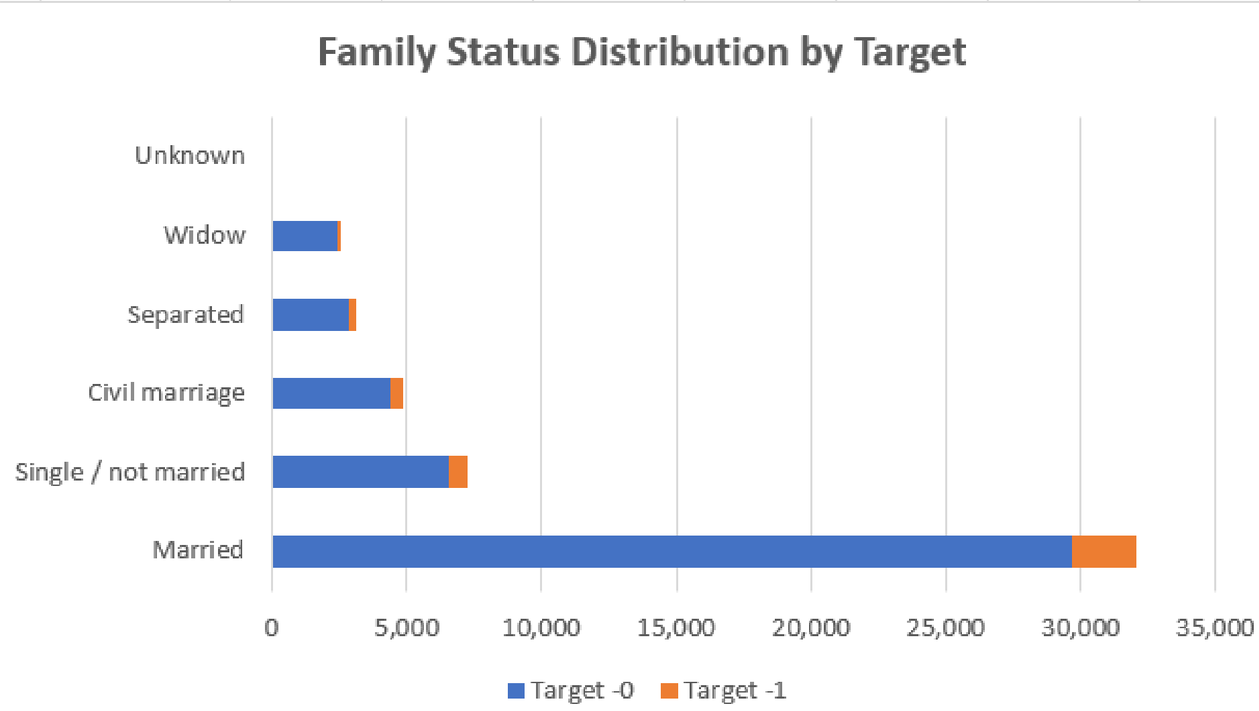
- 8% of clients have payment difficulties.
- 66% of loan applications are from Female clients.
- Most of Loan applications are done from Monday to Friday.
- Most of Loan applications are done by Working Clients.
- Most of loan applications are done by clients having income upto 3 lac.
- Most of loan applications are done for credit amount upto 75 lac.
- 91% of loans are Cash Loans and 69% of Clients own realty.
- Most of loans have annuity upto 50,000.

Task 4 : Perform Univariate, Segmented Univariate, and Bivariate Analysis

b) Segmented Univariate analysis

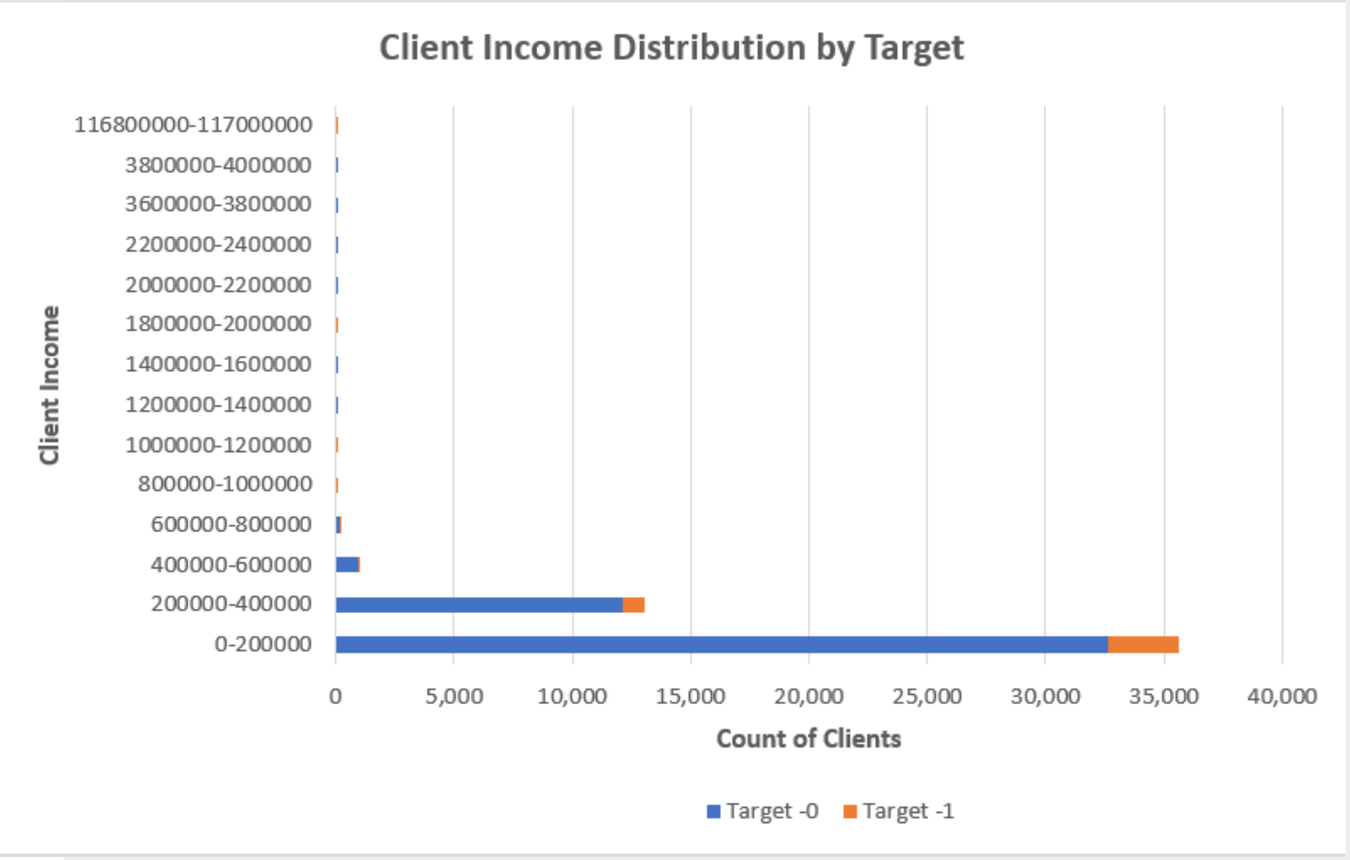
Perform segmented univariate analysis to compare variable distributions for different scenarios.





Task 4 : Perform Univariate, Segmented Univariate, and Bivariate Analysis

b) Segmented Univariate analysis

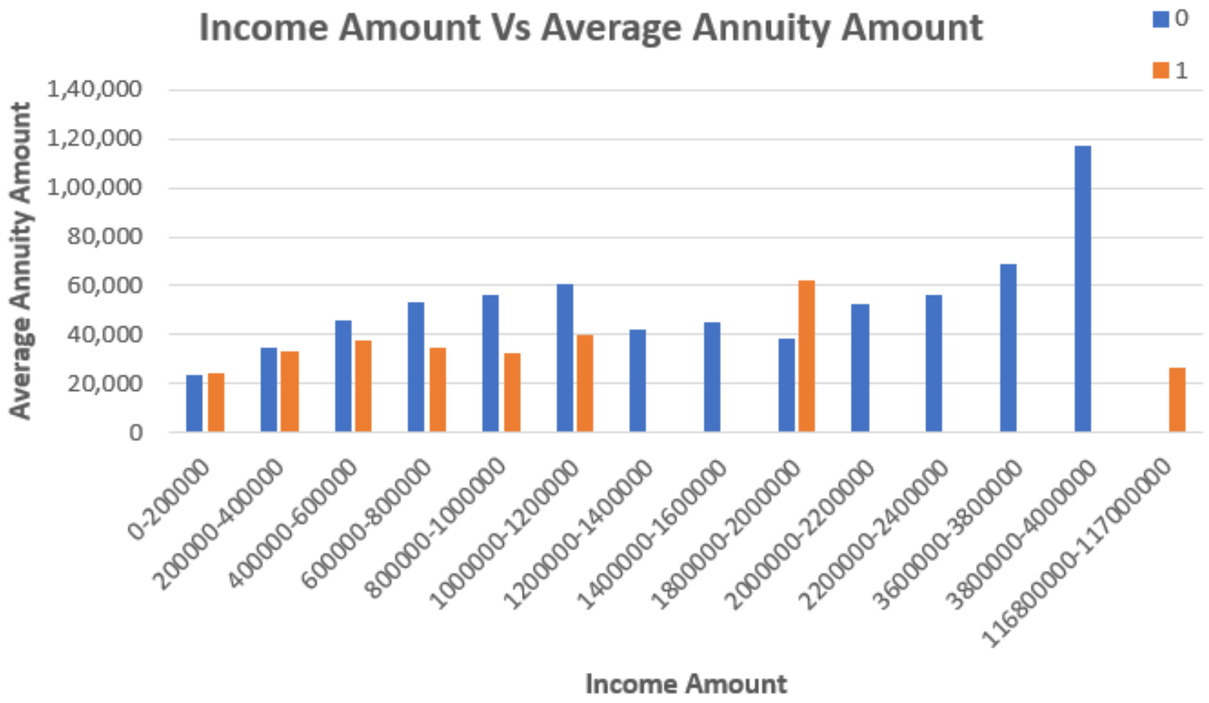


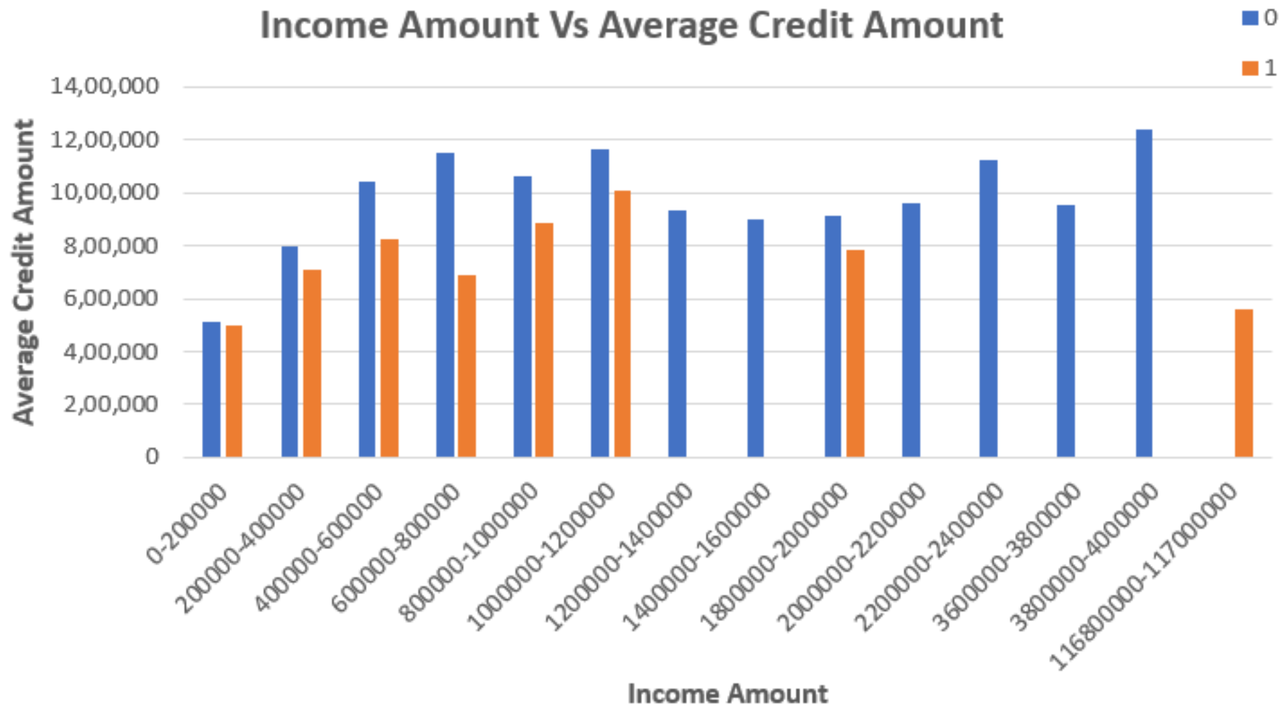
- Male clients tend to have higher chances of payment difficulties than Female clients.
- 93% of clients have Annuity upto 50,000 while 8% of them have payment difficulties.
- 64% of clients are married clients, while only 5% of them have payment difficulties.
- Most of loan applications are done by clients with secondary education and they tend to higher payment difficulties than client with Higher education.

Task 4 : Perform Univariate, Segmented Univariate, and Bivariate Analysis

c) Bivariate analysis

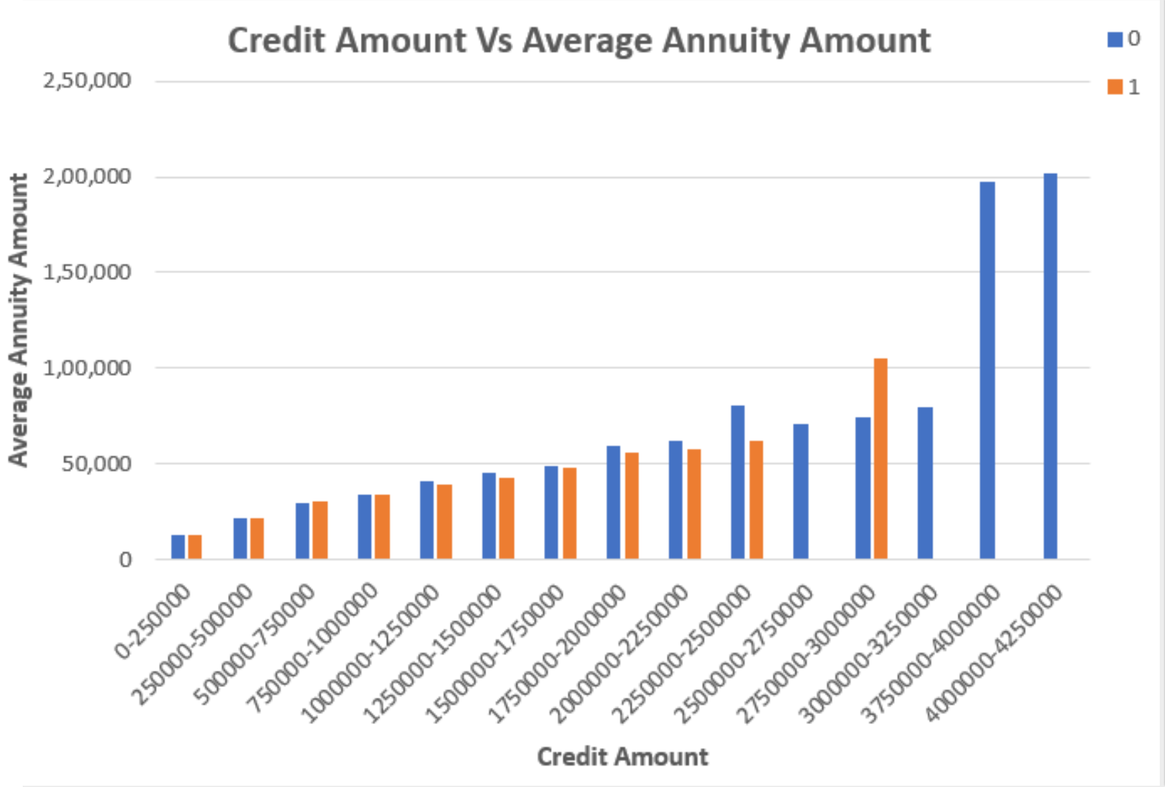
Perform bivariate analysis to explore relationships between variables and the target variable

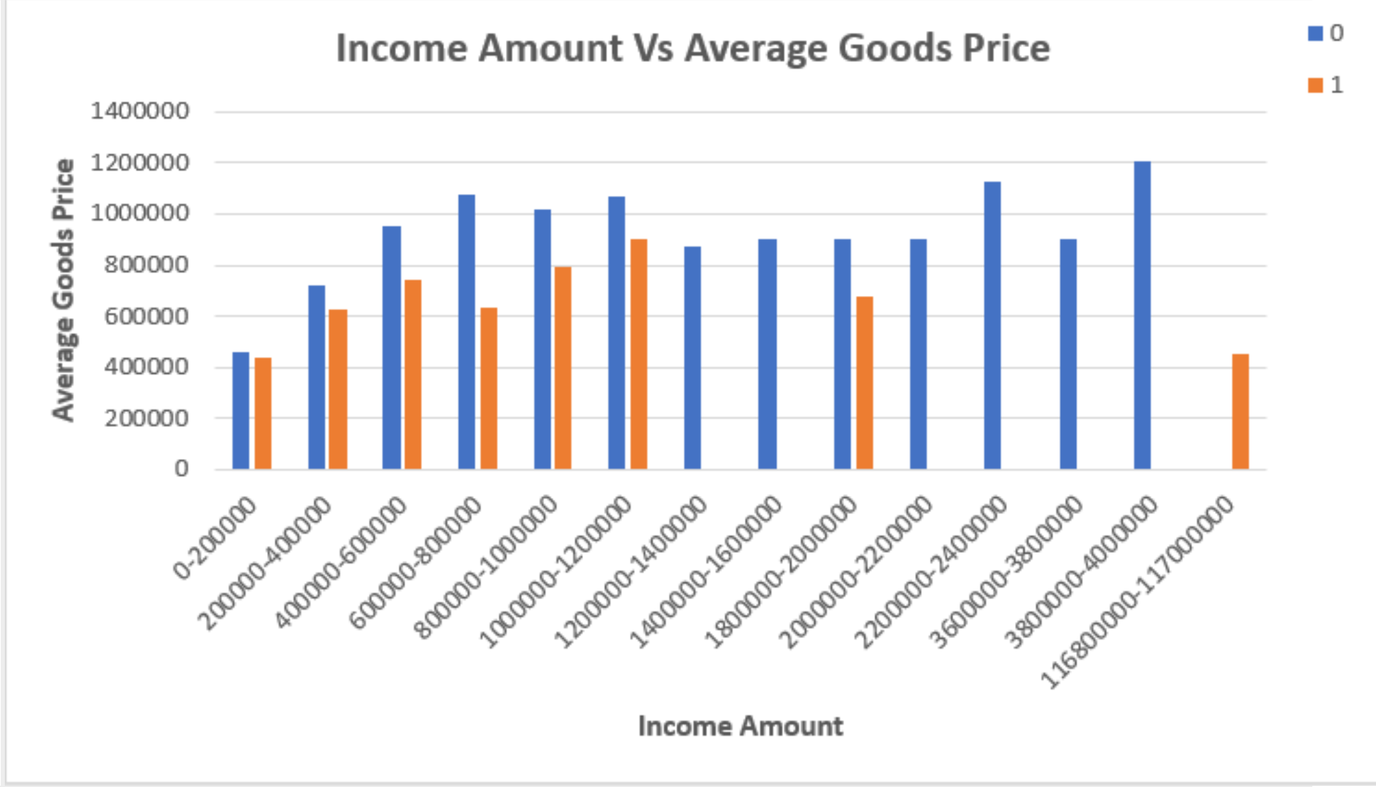




Task 4 : Perform Univariate, Segmented Univariate, and Bivariate Analysis

c) Bivariate analysis





- Client with higher income tend to higher amount of Goods price, Annuity and credit.

Task 5 : Identify Top Correlations for Different Scenarios

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

a) Top correlation between variable with Target 0

Correlation coefficient between variables for Target 0

CNT_CHILDREN	1.000	0.036	0.006	0.026	0.002	-0.025	-0.336	-0.246
AMT_INCOME_TOTAL	0.036	1.000	0.378	0.451	0.385	0.182	-0.074	-0.162
AMT_CREDIT	0.006	0.378	1.000	0.771	0.987	0.096	0.051	-0.075
AMT_ANNUITY	0.026	0.451	0.771	1.000	0.776	0.117	-0.010	-0.111
AMT_GOODS_PRICE	0.002	0.385	0.987	0.776	1.000	0.099	0.049	-0.072
REGION_POPULATION_RELATIVE	-0.025	0.182	0.096	0.117	0.099	1.000	0.030	-0.007
DAYS_BIRTH	-0.336	-0.074	0.051	-0.010	0.049	0.030	1.000	0.623
DAYS_EMPLOYED	-0.246	-0.162	-0.075	-0.111	-0.072	-0.007	0.623	1.000
	CNT_FAM_MEM	AMT_INCOME_T	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PI	REGION_POPUL	DAYS_BIRTH	DAYS_EMPLOYED

- AMT_CREDIT and AMT_GOODS_PRICE have very strong positive correlation. This indicates that there is a strong tendency for both variables to increase together and decrease together.
- AMT_GOODS_PRICE and AMT_ANNUITY have strong positive correlation. This indicates that there is a notable tendency for both variables to increase together and decrease together.
- AMT_CREDIT and AMT_ANNUITY have strong positive correlation. This indicates that there is a notable tendency for both variables to increase together and decrease together.
- DAYS_BIRTH and DAYS_EMPLOYEED have moderate positive correlation.

Task 5 : Identify Top Correlations for Different Scenarios

b) Top correlation between variable with Target 1

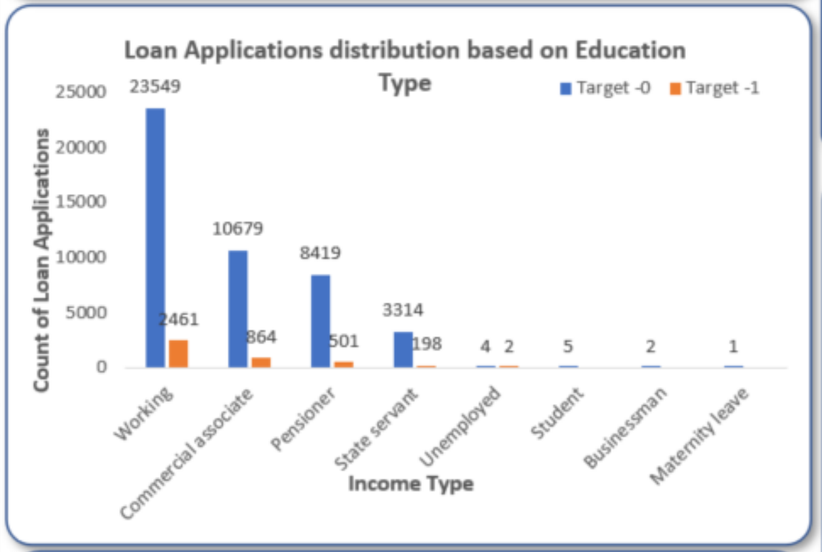
Correlation coefficient between variables for Target 1								
CNT_CHILDREN	1.000	0.010	0.008	0.029	-0.001	-0.020	-0.250	-0.190
AMT_INCOME_TOTAL	0.010	1.000	0.015	0.018	0.013	-0.006	-0.009	-0.012
AMT_CREDIT	0.008	0.015	1.000	0.750	0.982	0.068	0.143	0.019
AMT_ANNUITY	0.029	0.018	0.750	1.000	0.749	0.073	0.009	-0.078
AMT_GOODS_PRICE	-0.001	0.013	0.982	0.749	1.000	0.077	0.141	0.023
REGION_POPULATION_RELATIVE	-0.020	-0.006	0.068	0.073	0.077	1.000	0.016	0.008
DAYS_BIRTH	-0.250	-0.009	0.143	0.009	0.141	0.016	1.000	0.588
DAYS_EMPLOYED	-0.190	-0.012	0.019	-0.078	0.023	0.008	0.588	1.000
	CNT_CHILDREN	AMT_INCOMI	AMT_CREDIT	AMT_ANNUIT	AMT_GOODS_	REGION_POPU	DAYS_BIRTH	DAYS_EMPLOYED

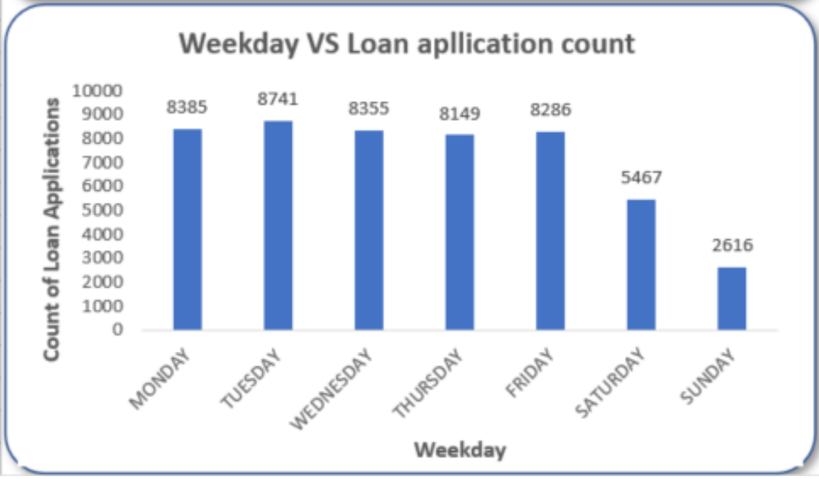
- AMT_CREDIT and AMT_GOODS_PRICE have very strong positive correlation. This indicates that there is a strong tendency for both variables to increase together and decrease together.
- AMT_GOODS_PRICE and AMT_ANNUITY have strong positive correlation. This indicates that there is a notable tendency for both variables to increase together and decrease together.
- AMT_CREDIT and AMT_ANNUITY have strong positive correlation. This indicates that there is a notable tendency for both variables to increase together and decrease together.
- DAYS_BIRTH and DAYS_EMPLOYEED have moderate positive correlation.

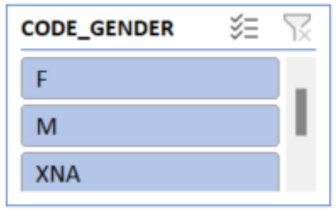
Interactive Dashboard:

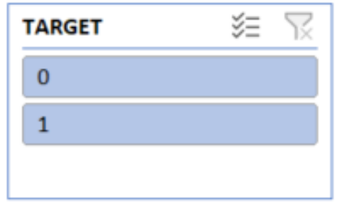
Bank Loan Case Study

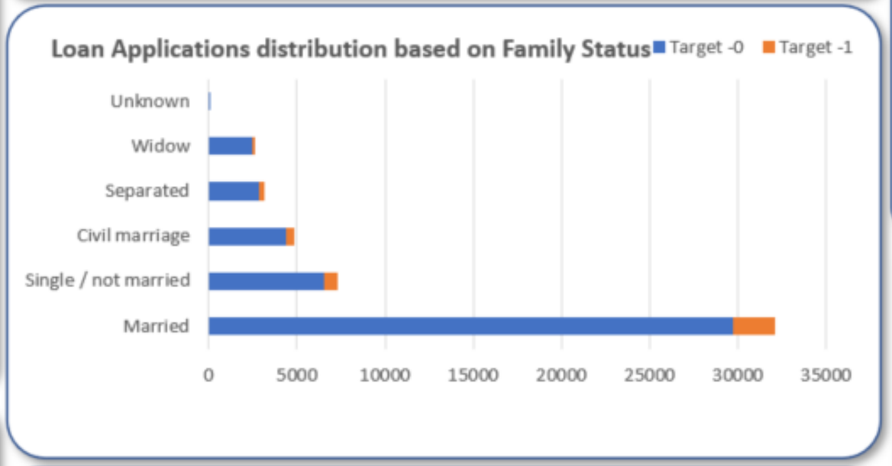
Project By: Mayur Rajput

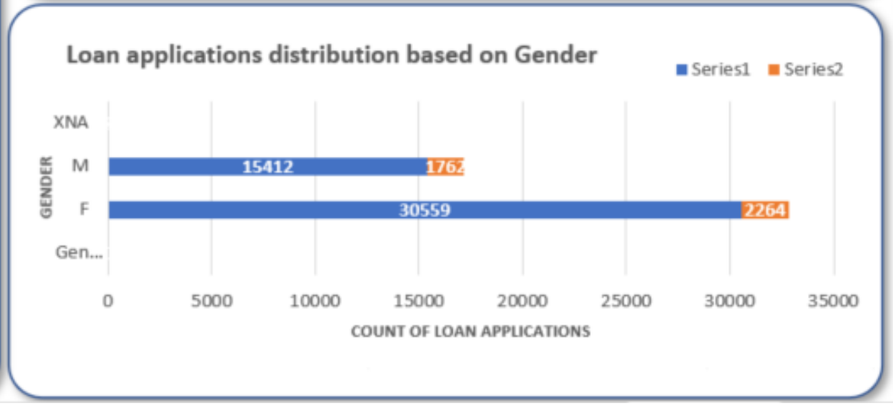


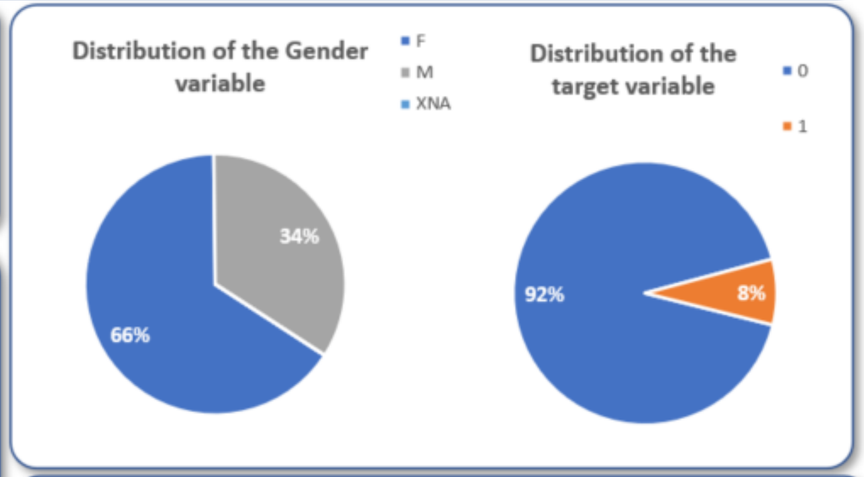


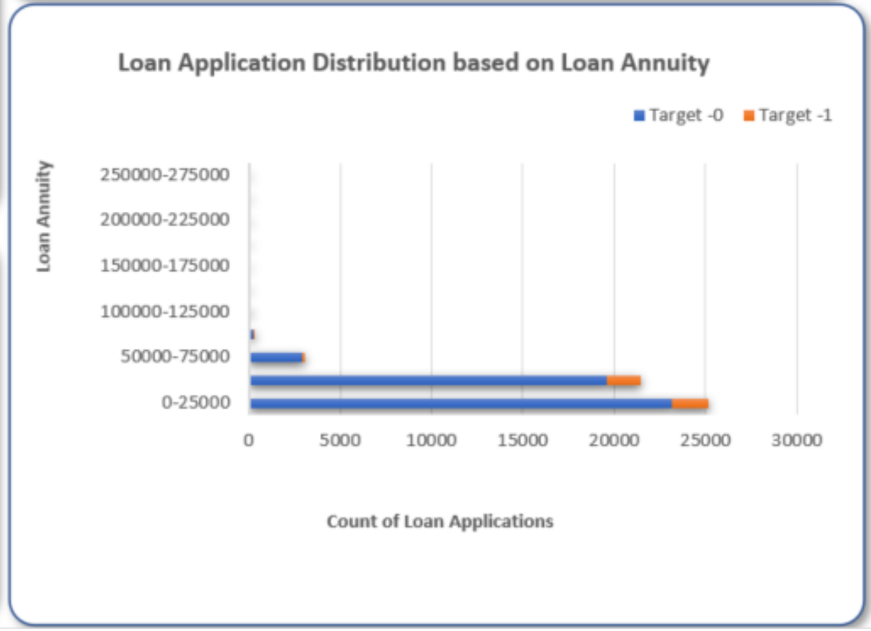












Insights :

1. 66% of loan applications are from Female clients and Male clients tend to have higher chances of payment difficulties than Female clients.
2. 93% of clients have loan Annuity upto 50,000 while 8% of them have payment difficulties.
3. Most of Loan applications are done from Monday to Friday.
4. Most of Loan applications are done by Working Clients.
5. 64% of clients are married clients, while only 5% of them have payment difficulties.
6. Most of loan applications are done by clients having income upto 4 lac.
7. Most of loan applications are done for credit amount upto 75 lac.
8. 91% of loans are Cash Loans and 69% of Clients own realty.
9. Most of loan applications are done by clients with secondary education and they tend to higher payment difficulties than client with Higher education.
10. Client with higher income tend to higher amount of Goods price, Annuity and credit.

Results :

This project helped me to advance Excel skills and problem solving ability. Through this project I learned how to handle missing data, outliers, Perform EDA which enabled me to give better representation of output in form of charts.