BANK LOAN CASE STUDY

FILE LINKS:

This file link: Click Here!

Working file links:

application_data.xlsx Click Here!

previous_application.xlsx Click Here!

merged.xlsx Click Here!

SUBMITTED BY:
PRINCE KUMAR
prince22495@gmail.com

Project Description:

Title: Exploratory Data Analysis (EDA) for Loan Default Prediction

Introduction:

In this project, we will perform Exploratory Data Analysis (EDA) on a dataset containing information about loan applicants and their previous loan applications. The goal of this analysis is to gain insights into the factors that are significant in determining a customer's likelihood of defaulting on a loan. We will utilize various data analysis techniques and visualizations to better understand the patterns and relationships in the data.

Dataset Description:

The dataset consists of two CSV files: "application_data.csv" and "previous_application.csv." The "application_data.csv" file contains information about current loan applicants, including their demographics, financial attributes, and loan status. On the other hand, the "previous_application.csv" file contains data related to previous loan applications made by these same applicants.

Objective:

Our main objective is to identify key variables and attributes that play a crucial role in determining whether a loan should be provided to a customer or not. Through EDA, we aim to uncover meaningful patterns and correlations in the data that can help us make informed decisions and create a predictive model for loan default prediction.

Steps for EDA:

1. Data Cleaning and Preprocessing:
    - Remove irrelevant columns from both datasets.
    - Handle missing values appropriately by imputation or removal.
    - Convert categorical variables into a suitable format for analysis.
2. **Univariate Analysis:**
    - Explore the distribution of each variable to understand its range and spread.
    - Use visualizations such as histograms, box plots, and bar charts to examine data characteristics.
3. **Bivariate Analysis:**
    - Analyze the relationship between loan attributes and customer attributes.
    - Identify correlations and patterns between variables that can impact loan approval.

4. **Data Visualization:**
    - Create various visualizations such as scatter plots, heatmaps, and correlation matrices to uncover trends and insights.
    - Visualize the distribution of loan amounts, income levels, and other key variables.
5. **Outlier Detection:**
    - Use quartiles and Interquartile Range (IQR) to identify potential outliers in the data.
    - Visualize outliers through box plots and assess their impact on loan approval.
6. **Merge Datasets:**
    - Merge the two datasets based on the common identifier "SK_ID_CURR" to combine information about previous loan applications with current loan applicants.
7. **Correlation Analysis:**
    - Calculate correlation coefficients between various variables to determine their interdependence.
    - Focus on variables that have a strong correlation with loan default.
8. **Feature Importance:**
    - Rank variables based on their significance in predicting loan default using correlation analysis results.

Conclusion:

Through thorough Exploratory Data Analysis, we will gain valuable insights into the factors that influence loan approval and default. This analysis will serve as the foundation for further predictive modeling to develop a robust loan default prediction system. The results and recommendations from this project will be valuable for financial institutions in making informed decisions about loan approvals and mitigating the risk of default.

Approach for Exploratory Data Analysis (EDA):

1. **Data Understanding:**
    - Initially, I acquired two datasets: "application_data.csv" containing information about current loan applicants and "previous_application.csv" with data related to their previous loan applications.
    - I carefully examined the data dictionary and understood the meaning of each variable and its significance in the loan approval process.
2. **Data Cleaning:**
    - I started by checking for missing values, duplicates, and irrelevant columns in both datasets.
    - Missing values were either imputed or handled using appropriate techniques based on the nature of the data.
    - I removed columns that were not relevant to the loan default prediction task to streamline the analysis.
3. **Univariate Analysis:**
    - To understand the distribution of individual variables, I used various descriptive statistics and visualizations.
    - I plotted histograms, box plots, and bar charts to examine the spread and characteristics of key variables such as income, loan amount, and annuity.
4. **Bivariate Analysis:**
    - Next, I explored the relationships between loan attributes and customer attributes.
    - I used scatter plots and heatmaps to visualize correlations between continuous variables.
    - For categorical variables, I created grouped bar charts to observe patterns and trends.
5. **Outlier Detection:**
    - I utilized the QUARTILE.INC formula in Excel to calculate quartiles and Interquartile Range (IQR) for identifying potential outliers.
    - Outliers were visualized using box plots to understand their impact on loan approval.

6. **Data Visualization:**
    - I created a range of visualizations to present the analysis in an easily interpretable format.
    - Visualizations included correlation matrices, scatter plots, bar charts, and histograms.
    - The use of color-coding and labeling enhanced the clarity of the findings.
7. **Merging Datasets:**
    - To enrich the analysis, I merged both datasets based on the common identifier "SK_ID_CURR."
    - This consolidation provided a comprehensive view of customer attributes and their previous loan history, aiding in the loan default prediction.
8. **Correlation Analysis:**
    - I calculated correlation coefficients between relevant variables to determine their strength of association with loan default.
    - Variables with a high correlation were deemed crucial for predicting loan outcomes.
9. **Feature Importance:**
    - I ranked the variables based on their significance in predicting loan default, as identified through correlation analysis.
    - This helped identify the most influential features to prioritize in a predictive model.
10. **Conclusion:**
    - The EDA process culminated in a detailed understanding of the dataset and its potential insights for loan default prediction.
    - The project resulted in a clear understanding of significant variables, correlations, and patterns that influence loan approvals and defaults.
    - The findings and recommendations from the EDA will form a strong foundation for further modeling and building a robust loan default prediction system.
11. **Documentation and Reporting:**
    - All steps, findings, and visualizations were documented in a systematic manner to facilitate clear communication and reporting.

- I presented the results in a concise yet comprehensive manner, providing actionable insights for stakeholders and decision-makers in financial institutions.

Tech Stack Used for Exploratory Data Analysis (EDA):

Microsoft Excel (2021)

- Microsoft Excel is the primary tool used for data analysis in this project.
- Excel provides a familiar and user-friendly interface to manipulate and visualize data, making it accessible to a wide range of users, including non-technical individuals.
- Its powerful formulas, functions, and features enable data cleaning, transformation, and computation, making it suitable for conducting EDA.
- Excel's extensive charting and graphing capabilities allow for the creation of various visualizations to understand the data distribution and relationships between variables.

Data Import:

- Excel allows seamless import of data from CSV files like "application_data.csv" and "previous_application.csv," making it easy to work with large datasets.
- The "Data" tab in Excel facilitates data import from different sources, ensuring efficient data handling.

Data Cleaning and Transformation:

- Excel's functions and formulas were utilized to clean and preprocess the data, addressing missing values and duplicates.
- Conditional formatting and filtering helped identify outliers and anomalies in the data.
- Columns were renamed and irrelevant variables were removed to streamline the analysis.

Data Visualization:

- Excel's charting tools were employed to create a variety of visualizations, such as histograms, bar charts, scatter plots, and box plots.
- Conditional formatting allowed for color-coding data to highlight patterns and relationships. The "Data Analysis" add-in in Excel was utilized to generate a correlation matrix, providing insights into variable relationships.

Merging Datasets:

- Excel's IFERROR, INDEX and MATCH functions enabled the merging of the two datasets based on a common identifier, "SK_ID_CURR."
- This process allowed for combining customer attributes from "application_data.csv" with their respective previous loan information from "previous_application.csv."

Data Analysis:

- Excel's formula capabilities, including functions like "IF," "IFERROR," and "SUM," were leveraged to perform data analysis tasks.
- QUARTILE.INC function helped identify outliers, while other statistical functions supported data exploration and decision-making.

In summary, Excel served as a powerful and versatile tool for conducting Exploratory Data Analysis in this project. Its data manipulation, visualization, and analysis capabilities facilitated a comprehensive understanding of the datasets, enabling the extraction of meaningful insights to inform decision-making processes related to loan default prediction.

EDA of application data:

Tasks to be performed for the EDA analysis are:

- Task A: Identify Missing Data and Deal with it Appropriately
- Task B: Identify Outliers in the Dataset
- Task C: Analyze Data Imbalance
- Task D: Perform Univariate, Segmented Univariate, and Bivariate Analysis
- Task E: Identify Top Correlations for Different Scenarios

Task A: Identify Missing Data and Deal with it Appropriately:

1. The initial dataset contains 122 columns and 50,000 row.

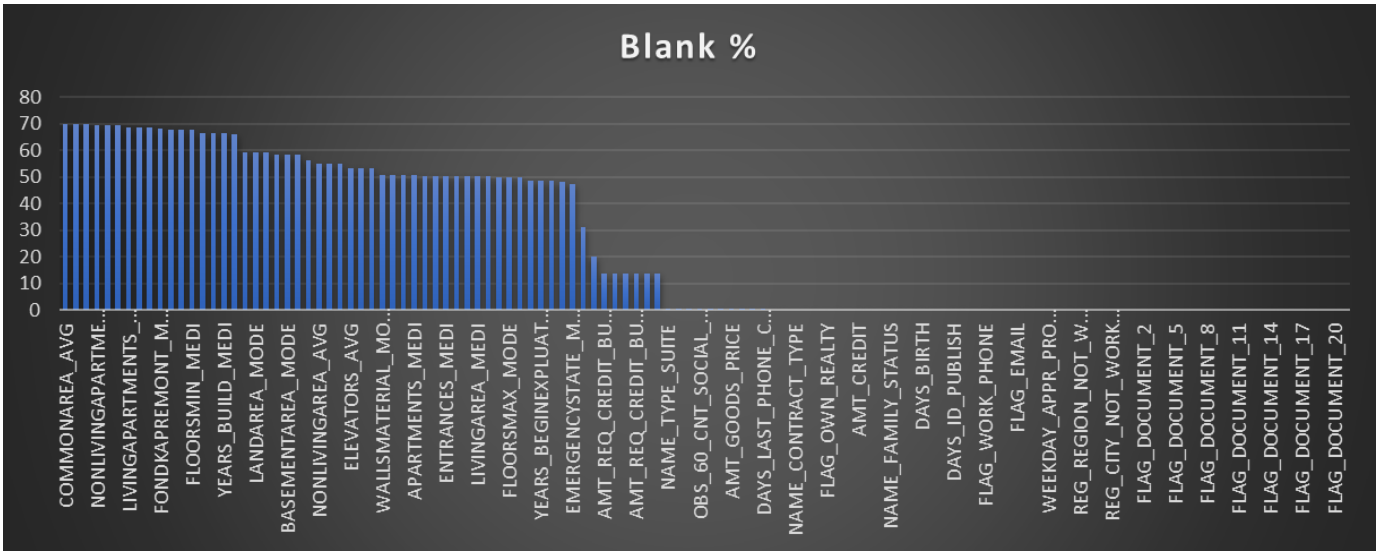
SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN
100002	1	Cash loans	M	N	Y	0
100003	0	Cash loans	F	N	N	0
100004	0	Revolving loans	M	Y	Y	0
100006	0	Cash loans	F	N	Y	0
100007	0	Cash loans	M	N	Y	0
100008	0	Cash loans	M	N	Y	0
100009	0	Cash loans	F	Y	Y	1
100010	0	Cash loans	M	Y	Y	0
100011	0	Cash loans	F	N	Y	0
100012	0	Revolving loans	M	N	Y	0
100014	0	Cash loans	F	N	Y	1
100015	0	Cash loans	F	N	Y	0
100016	0	Cash loans	F	N	Y	0
100017	0	Cash loans	M	Y	N	1
100018	0	Cash loans	F	N	Y	0
100019	0	Cash loans	M	Y	Y	0
100020	0	Cash loans	M	N	N	0
100021	0	Revolving loans	F	N	Y	1

After cleaning and removing irrelevant columns, data was left with 24 columns. The columns having blank percentage greater than 40% were removed from the dataset file.

Excel formula used to find blank values in column: =COUNTBLANK($A$2:$A$50000). Then conditional formatting was used to highlight values of blank percentage greater than 40%. These identified columns were then removed to create a clean dataset.

Columns Identified to be removed as they have blank % greater than 40%			
Columns	***Blanks Coun***	***Total Rows***	***Blank %***
SK_ID_CURR	0	49999	0
TARGET	0	49999	0
NAME_CONTRACT_TYPE	0	49999	0
CODE_GENDER	0	49999	0
FLAG_OWN_CAR	0	49999	0
FLAG_OWN_REALTY	0	49999	0
CNT_CHILDREN	0	49999	0
AMT_INCOME_TOTAL	0	49999	0
AMT_CREDIT	0	49999	0
AMT_ANNUITY	1	49999	0.00200004
AMT_GOODS_PRICE	38	49999	0.07600152
NAME_TYPE_SUITE	192	49999	0.38400768
NAME_INCOME_TYPE	0	49999	0
NAME_EDUCATION_TYPE	0	49999	0
NAME_FAMILY_STATUS	0	49999	0
NAME_HOUSING_TYPE	0	49999	0
REGION_POPULATION_RELATIVE	0	49999	0
DAYS_BIRTH	0	49999	0
DAYS_EMPLOYED	0	49999	0
DAYS_REGISTRATION	0	49999	0
DAYS_ID_PUBLISH	0	49999	0
OWN_CAR_AGE	32950	49999	65.901318
FLAG_MOBIL	0	49999	0

Graph of Blank percentage per column:



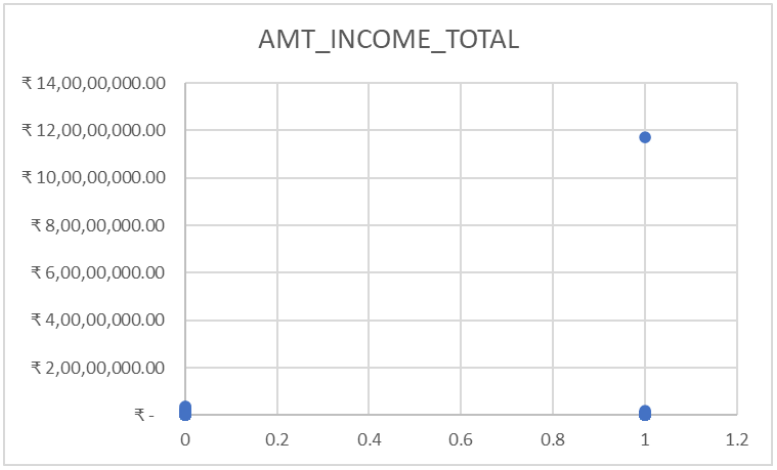
One blank in AMT_ANNUITY column was imputed with the 5% value of credit amount as per the observed trend in other cases. 5% of 4,50,000 is 22,500.

	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	INCOME CLASS	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	NAME_INCOME_T
47535	N	0	₹ 1,80,000.00	175000 - 225000	₹ 4,50,000.00		₹ 4,50,000.00	Commercial assoc
50003								

Task B: Identify Outliers in the Dataset

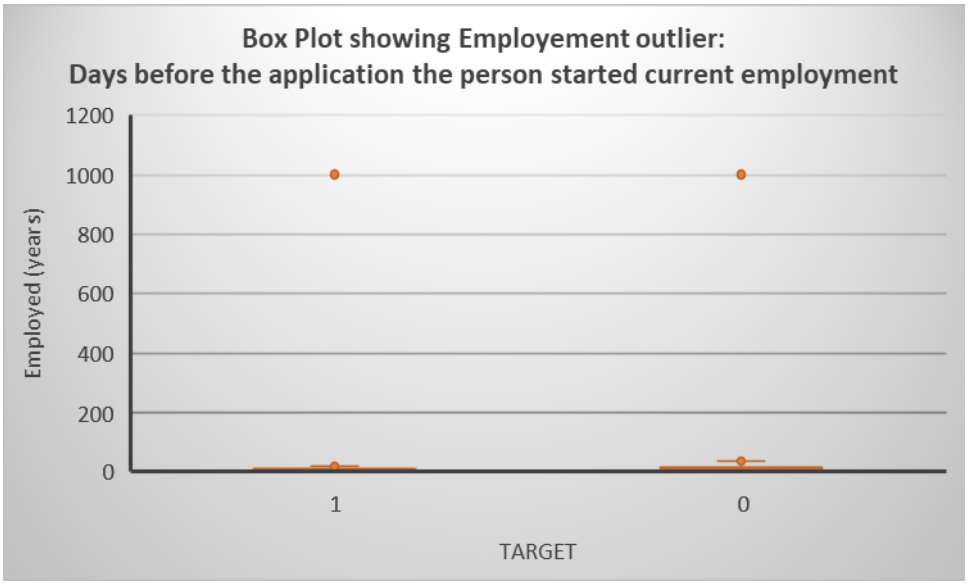
Income outlier - 117000000

1. **Income Outliers:**



QUARTILE 1	112500
QUARTILE 3	202500
Interquartile Range (IQR).	90000

2. Employment Outlier:

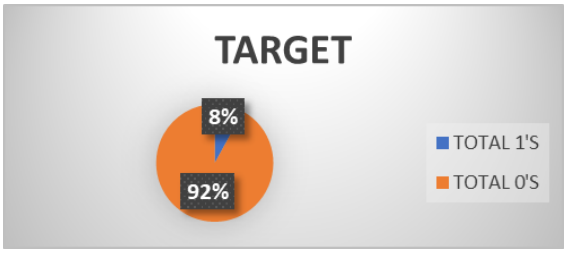


QUARTILE 1	2.6
QUARTILE 3	15.7
Interquartile Range (IQR).	13.1

TARGET	Employed (years)	Outlier
1	1.7	Not an Outlier
0	3.3	Not an Outlier
0	0.6	Not an Outlier
0	8.3	Not an Outlier
0	8.3	Not an Outlier
0	4.4	Not an Outlier
0	8.6	Not an Outlier
0	1.2	Not an Outlier
0	1000.7	Outlier
0	5.5	Not an Outlier
0	1.9	Not an Outlier
0	1000.7	Outlier
0	7.4	Not an Outlier
0	8.3	Not an Outlier
0	0.6	Not an Outlier
0	3.2	Not an Outlier
0	3.6	Not an Outlier
0	0.5	Not an Outlier
0	21.4	Not an Outlier

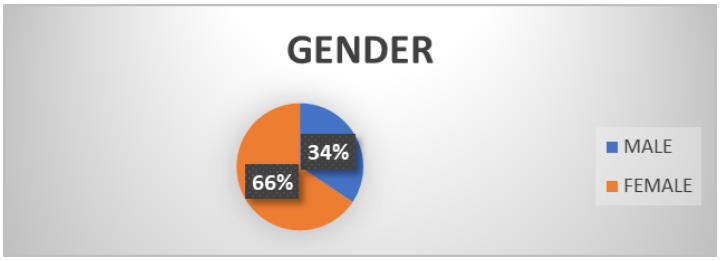
Task C: Analyze Data Imbalance

DATA IMBALANCE	
TARGET	
TOTAL 1'S	4026
TOTAL 0'S	45973
GRAND TOTAL	49999



The ration of data imbalance in the TARGET variable is 7:9

GENDER	
MALE	17174
FEMALE	32823
TOTAL	49997

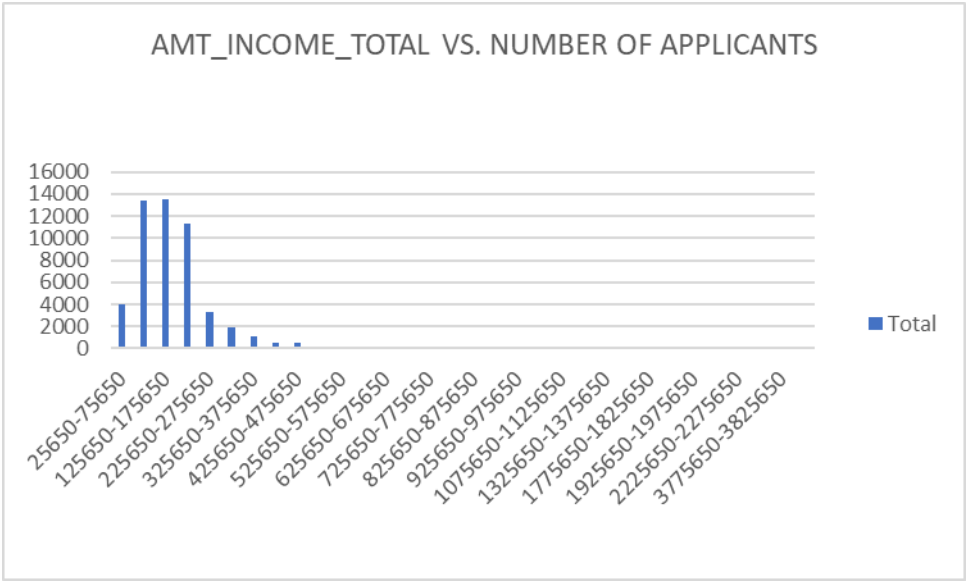


The ratio of gender imbalance is 7:11.

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

1. Income vs. Number of applicants

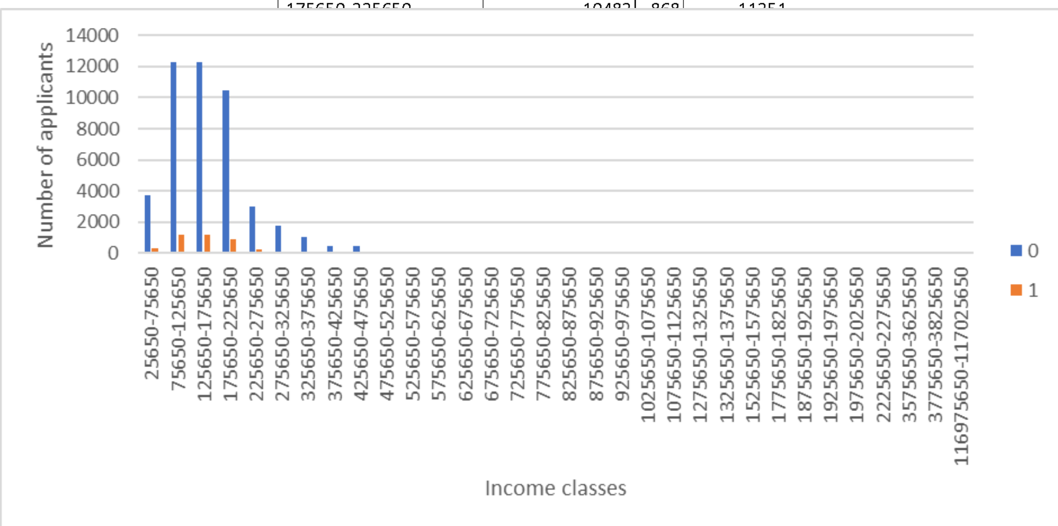
Row Labels	Count of SK_ID_CURR
25650-75650	4035
75650-125650	13405
125650-175650	13476
175650-225650	11351
225650-275650	3265
275650-325650	1861
325650-375650	1103
375650-425650	489
425650-475650	513
475650-525650	56
525650-575650	126
575650-625650	43
625650-675650	141
675650-725650	24
725650-775650	12
775650-825650	22
825650-875650	4
875650-925650	31
925650-975650	2
1025650-1075650	1
1075650-1125650	16
1275650-1325650	1
1325650-1375650	10
1525650-1575650	1
1775650-1825650	2
1875650-1925650	1
1925650-1975650	1
1975650-2025650	2
2225650-2275650	2
3575650-3625650	1
3775650-3825650	1
116975650-117025650	1
Grand Total	**49999**



2. **SEGMENTED UNIVARIATE ANALYSIS:**

Income classes, TARGET and number of applicants:

Count of SK_ID_CURR	Column Labels		
Row Labels	0	1	Grand Total
25650-75650	3726	309	4035
75650-125650	12249	1156	13405
125650-175650	12285	1191	13476
175650-225650	10482	868	11351
1775650-1825650	2		2
1875650-1925650		1	1
1925650-1975650	1		1
1975650-2025650	2		2
2225650-2275650	2		2
3575650-3625650	1		1
3775650-3825650	1		1
116975650-117025650		1	1
Grand Total	**45973**	**4026**	**49999**



Most loan applicants fall in the income class of 12650 – 17650 range i.e 13476. Majority of this income class also pay their instalments on time without facing any payment difficulty. Major number of applicants fall in the income class of less than 5 lakhs i.e more than 90%.

More loan default happens with less income groups.

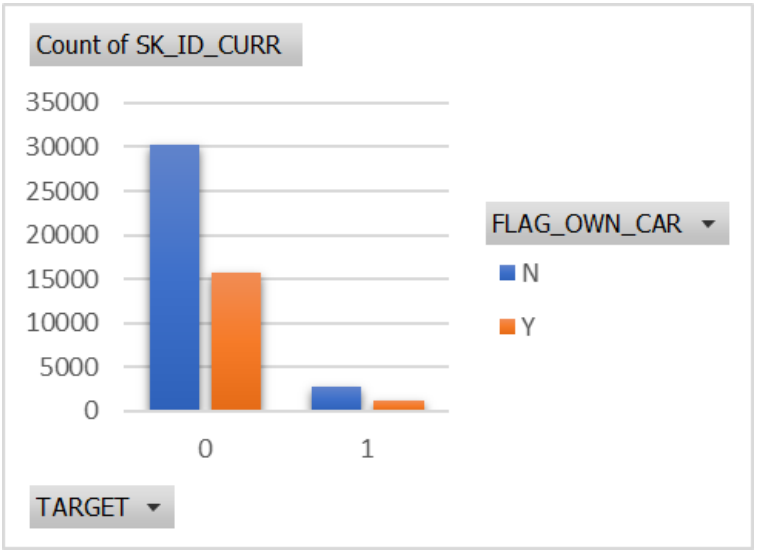
3. BIVARIATE ANALYSIS:

a) If client owns a car

Count of SK_ID_CURR	Column Labels		
Row Labels	N	Y	Grand Total
0	30176	15797	45973
1	2773	1253	4026
Grand Total	**32949**	**17050**	**49999**

Values in percentage of the row total for better understanding:

Count of SK_ID_CURR	Column Labels		
Row Labels	N	Y	Grand Total
0	65.64%	34.36%	100.00%
1	68.88%	31.12%	100.00%
Grand Total	**65.90%**	**34.10%**	**100.00%**



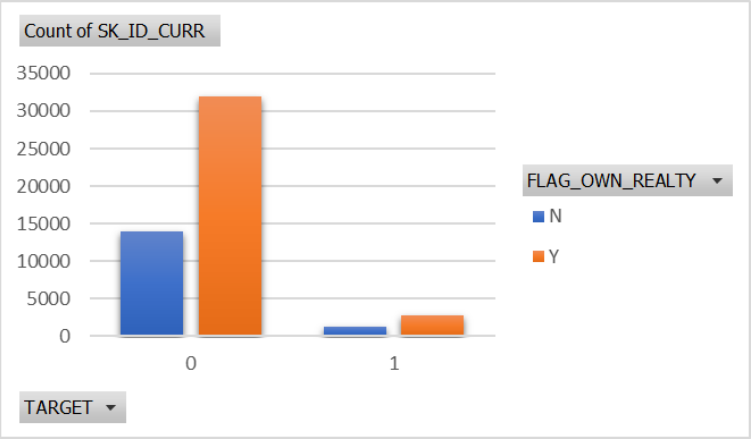
Most loan applicants do not own a car.

b) If client owns a house or a flat:

Count of SK_ID_CURR	Column Labels		
Row Labels	N	Y	Grand Total
0	14034	31939	45973
1	1274	2752	4026
Grand Total	**15308**	**34691**	**49999**

Values in percentage of the row total for better understanding:

Count of SK_ID_CURR	Column Labels		
Row Labels	N	Y	Grand Total
0	30.53%	69.47%	100.00%
1	31.64%	68.36%	100.00%
Grand Total	**30.62%**	**69.38%**	**100.00%**



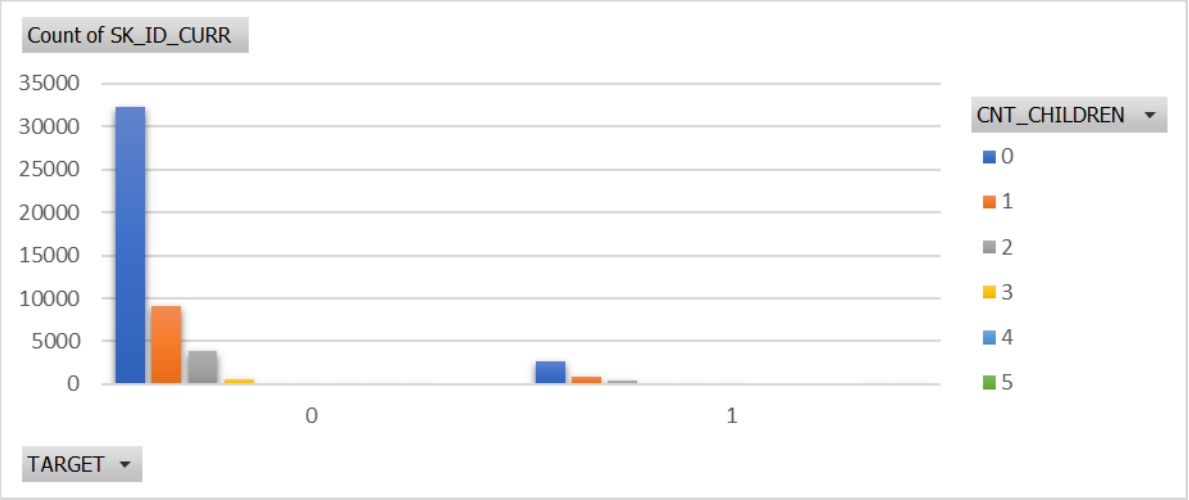
There is a large number of loan applicants who owns a house or flat. They pay their instalments on time.

c) Number of children the client has

Count of SK_ID_CURR	Column Labels											
Row Labels	0	1	2	3	4	5	6	7	8	9	11	Grand Total
0	32272	9118	3935	570	59	10	6	2	1			45973
1	2644	923	384	56	14	3				1	1	4026
Grand Total	**34916**	**10041**	**4319**	**626**	**73**	**13**	**6**	**2**	**1**	**1**	**1**	**49999**

Values in percentage of the row total for better understanding:

Count of SK_ID_CURR	Column Labels											
Row Labels	0	1	2	3	4	5	6	7	8	9	11	Grand Total
0	70.20%	19.83%	8.56%	1.24%	0.13%	0.02%	0.01%	0.00%	0.00%	0.00%	0.00%	100.00%
1	65.67%	22.93%	9.54%	1.39%	0.35%	0.07%	0.00%	0.00%	0.00%	0.02%	0.02%	100.00%
Grand Total	**69.83%**	**20.08%**	**8.64%**	**1.25%**	**0.15%**	**0.03%**	**0.01%**	**0.00%**	**0.00%**	**0.00%**	**0.00%**	**100.00%**



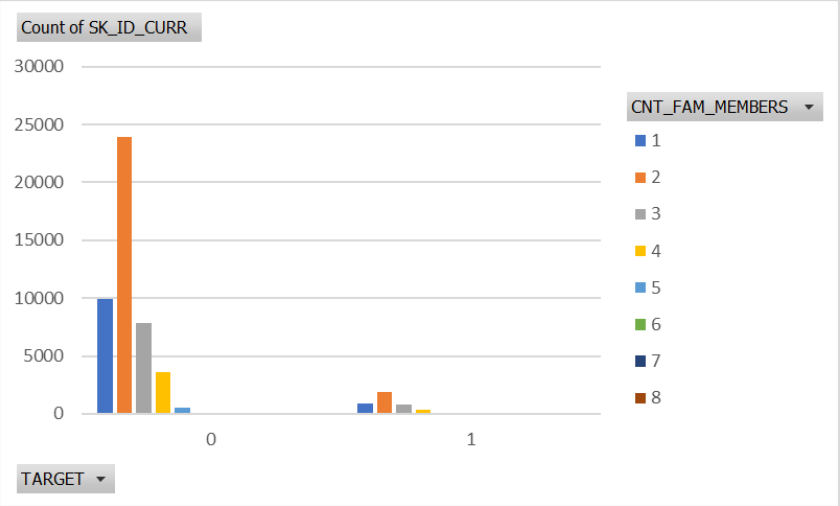
Most loans applicants do not have any children and they also do not face any loan payment difficulties, means they pay their loan annuity on time.

d) Client family members

Count of SK_ID_CURR	Column Labels											
Row Labels	1	2	3	4	5	6	7	8	9	10	13 (blank)	Grand Total
0	9951	23901	7858	3651	538	55	9	6	2	1	1	45973
1	922	1906	777	349	54	13	3			1	1	4026
Grand Total	**10873**	**25807**	**8635**	**4000**	**592**	**68**	**12**	**6**	**2**	**2**	**1**	**49999**

Values in percentage of the row total for better understanding:

Count of SK_ID_CURR	Column Labels											
Row Labels	1	2	3	4	5	6	7	8	9	10	13 (blank)	Grand Total
0	21.65%	51.99%	17.09%	7.94%	1.17%	0.12%	0.02%	0.01%	0.00%	0.00%	0.00%	100.00%
1	22.90%	47.34%	19.30%	8.67%	1.34%	0.32%	0.07%	0.00%	0.00%	0.02%	0.00%	100.00%
Grand Total	**21.75%**	**51.62%**	**17.27%**	**8.00%**	**1.18%**	**0.14%**	**0.02%**	**0.01%**	**0.00%**	**0.00%**	**0.00%**	**100.00%**



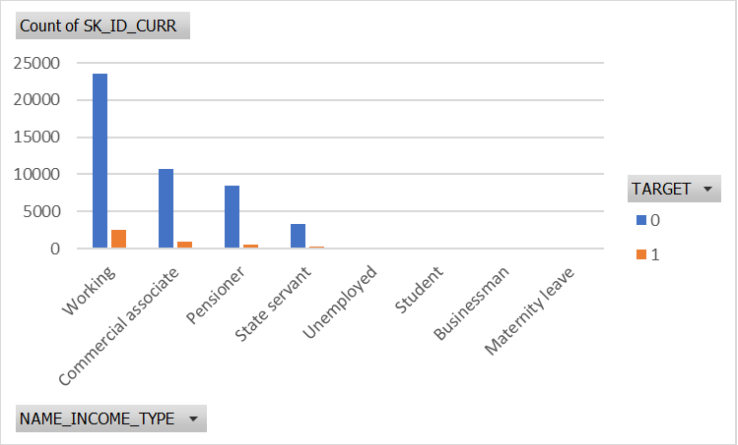
Most loans applicants have only **2** family members and most of them do not face any payment difficulty.

e) NAME_INCOME_TYPE:

Count of SK_ID_CURR	Column Labels		
Row Labels	**0**	**1**	**Grand Total**
Working	23549	2461	26010
Commercial associate	10679	864	11543
Pensioner	8419	501	8920
State servant	3314	198	3512
Unemployed	4	2	6
Student	5		5
Businessman	2		2
Maternity leave	1		1
Grand Total	**45973**	**4026**	**49999**

Values in percentage of the row total for better understanding:

Count of SK_ID_CURR	Column Labels		
Row Labels	**0**	**1**	**Grand Total**
Working	90.54%	9.46%	100.00%
Commercial associate	92.51%	7.49%	100.00%
Pensioner	94.38%	5.62%	100.00%
State servant	94.36%	5.64%	100.00%
Unemployed	66.67%	33.33%	100.00%
Student	100.00%	0.00%	100.00%
Businessman	100.00%	0.00%	100.00%
Maternity leave	100.00%	0.00%	100.00%
Grand Total	**91.95%**	**8.05%**	**100.00%**



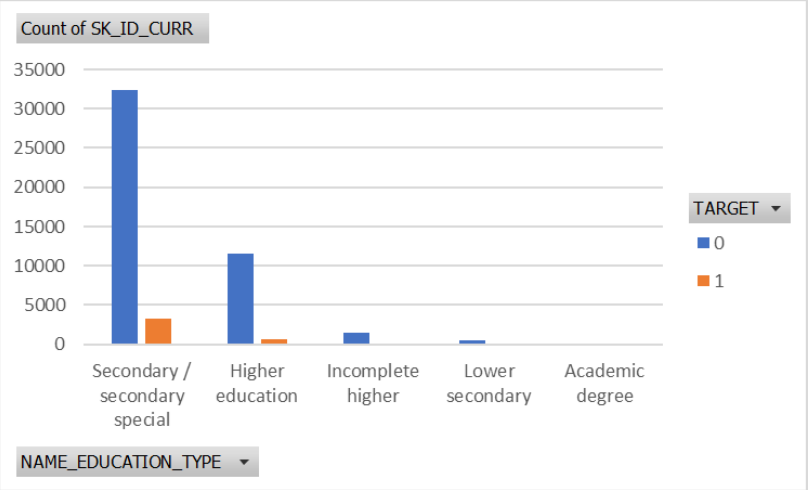
More loans have been applied by the **working people** and most of them do not face any payment difficulty.

f) NAME_EDUCATION_TYPE

Count of SK_ID_CURR	Column Labels		
Row Labels	**0**	**1**	**Grand Total**
Secondary / secondary special	32363	3209	35572
Higher education	11561	606	12167
Incomplete higher	1482	138	1620
Lower secondary	547	73	620
Academic degree	20		20
Grand Total	**45973**	**4026**	**49999**

Values in percentage of the row total for better understanding:

Count of SK_ID_CURR	Column Labels		
Row Labels	**0**	**1**	**Grand Total**
Secondary / secondary special	90.98%	9.02%	100.00%
Higher education	95.02%	4.98%	100.00%
Incomplete higher	91.48%	8.52%	100.00%
Lower secondary	88.23%	11.77%	100.00%
Academic degree	100.00%	0.00%	100.00%
Grand Total	**91.95%**	**8.05%**	**100.00%**



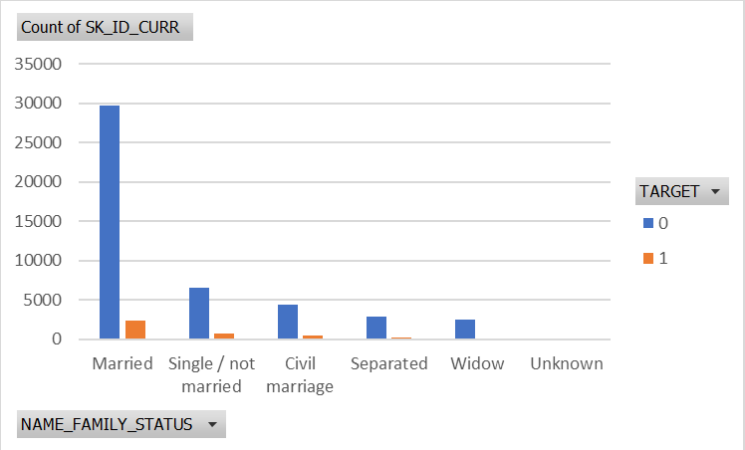
Applicants with **secondary education** seem to be making their payments on time without any delay. Thus, secondary education type applicants have more chance of getting their loan approved. This also means that largest number of loan applicants possess secondary education.

g) NAME_FAMILY_STATUS:

Count of SK_ID_CURR	Column Labels		
Row Labels	**0**	**1**	**Grand Total**
Married	29699	2395	32094
Single / not married	6577	729	7306
Civil marriage	4377	482	4859
Separated	2870	272	3142
Widow	2449	148	2597
Unknown	1		1
Grand Total	**45973**	**4026**	**49999**

Values in percentage of the row total for better understanding:

Count of SK_ID_CURR	Column Labels		
Row Labels	**0**	**1**	**Grand Total**
Married	92.54%	7.46%	100.00%
Single / not married	90.02%	9.98%	100.00%
Civil marriage	90.08%	9.92%	100.00%
Separated	91.34%	8.66%	100.00%
Widow	94.30%	5.70%	100.00%
Unknown	100.00%	0.00%	100.00%
Grand Total	**91.95%**	**8.05%**	**100.00%**



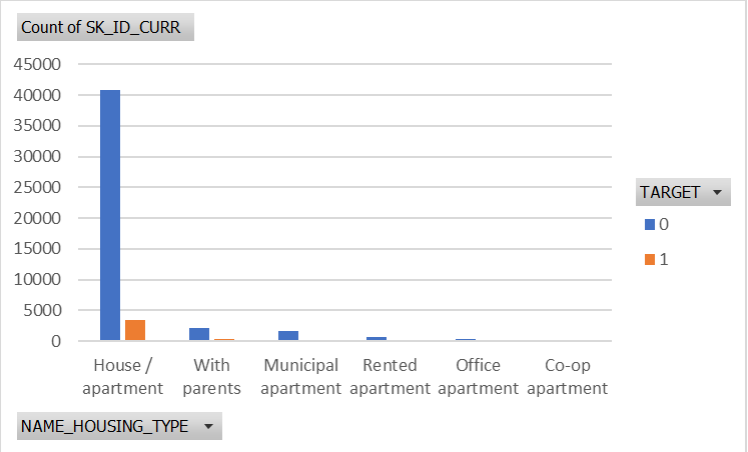
Married applicants were found to be the one without any payment difficulties, so they are more likely to have their loan approved. Also, they applied for most loans.

h) NAME_HOUSING_TYPE:

Count of SK_ID_CURR	Column Labels		
Row Labels	**0**	**1**	**Grand Total**
House / apartment	40895	3473	44368
With parents	2122	277	2399
Municipal apartment	1700	145	1845
Rented apartment	682	87	769
Office apartment	398	29	427
Co-op apartment	176	15	191
Grand Total	**45973**	**4026**	**49999**

Values in percentage of the row total for better understanding:

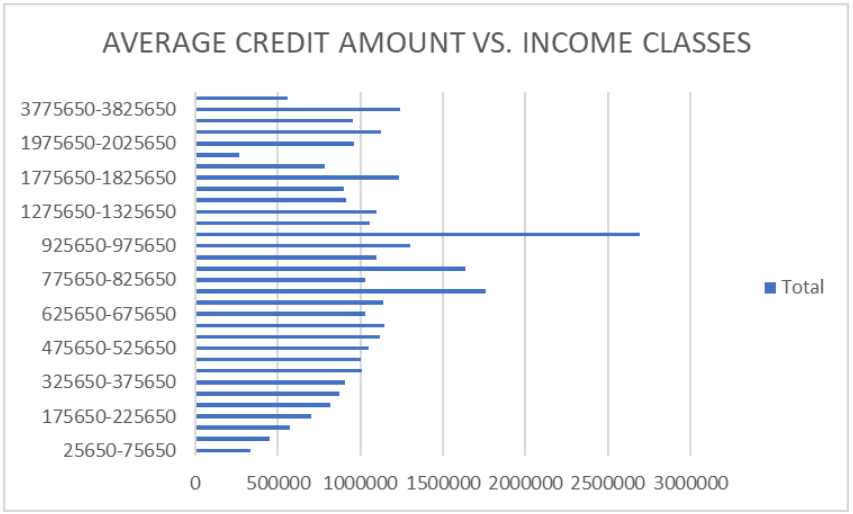
Count of SK_ID_CURR	Column Labels		
Row Labels	**0**	**1**	**Grand Total**
House / apartment	92.17%	7.83%	100.00%
With parents	88.45%	11.55%	100.00%
Municipal apartment	92.14%	7.86%	100.00%
Rented apartment	88.69%	11.31%	100.00%
Office apartment	93.21%	6.79%	100.00%
Co-op apartment	92.15%	7.85%	100.00%
Grand Total	**91.95%**	**8.05%**	**100.00%**



Applicants living in their own **house / apartment** did not see any payment difficulties, therefore they are more likely to get their loan approved. Also, they applied for more loans.

i) Average credit amount Vs. Income classes

Row Labels	Average of AMT_CREDIT
25650-75650	335998
75650-125650	452088
125650-175650	573628
175650-225650	703003
225650-275650	820492
275650-325650	872960
325650-375650	904707
375650-425650	1006409
425650-475650	1004628
475650-525650	1049978
525650-575650	1115100
575650-625650	1148572
625650-675650	1029972
675650-725650	1141550
725650-775650	1762535
775650-825650	1027900
825650-875650	1638944
875650-925650	1099720
925650-975650	1303200
1025650-1075650	2695500
1075650-1125650	1057505
1275650-1325650	1095111
1325650-1375650	914911
1525650-1575650	900000
1775650-1825650	1237500
1875650-1925650	781920
1925650-1975650	269550
1975650-2025650	961828
2225650-2275650	1125000
3575650-3625650	953460
3775650-3825650	1241024
116975650-117025650	562491
Grand Total	**599701**



Highest average credit amount is approved for income class 10,25650 - 10,75650 i.e., 26,95500.

Task E: Identify Top Correlations for Different Scenarios

CORRELATION WHEN T=0 (Client without payment difficulties)

	CAR	*REALTY*	*CHILDREN*	*INCOME*	*CREDIT*	*ANNUITY*	*REGION POPULATION*	*BIRTH (Years)*	*EMPLOYMENT (Years)*	*REGISTRATION (Years)*	*FAMILY MEMBERS*
CAR	1.000	0.004	0.112	0.206	0.109	0.136	0.038	-0.132	-0.158	-0.085	0.159
REALTY	0.004	1.000	-0.002	-0.001	-0.043	-0.004	0.008	0.118	0.065	0.022	0.008
CHILDREN	0.112	-0.002	1.000	0.036	0.006	0.026	-0.025	-0.336	-0.246	-0.183	0.879
INCOME	0.206	-0.001	0.036	1.000	0.378	0.451	0.182	-0.074	-0.162	-0.069	0.042
CREDIT	0.109	-0.043	0.006	0.378	1.000	0.771	0.096	0.051	-0.075	-0.008	0.065
ANNUITY	0.136	-0.004	0.026	0.451	0.771	1.000	0.117	-0.010	-0.111	-0.035	0.078
REGION POPULATION	0.038	0.008	-0.025	0.182	0.096	0.117	1.000	0.030	-0.007	0.058	-0.023
BIRTH (Years)	-0.132	0.118	-0.336	-0.074	0.051	-0.010	0.030	1.000	0.623	0.335	-0.284
EMPLOYMENT (Years)	-0.158	0.065	-0.246	-0.162	-0.075	-0.111	-0.007	0.623	1.000	0.209	-0.235
REGISTRATION (Years)	-0.085	0.022	-0.183	-0.069	-0.008	-0.035	0.058	0.335	0.209	1.000	-0.171
FAMILY MEMBERS	0.159	0.008	0.879	0.042	0.065	0.078	-0.023	-0.284	-0.235	-0.171	1.000

INTERPRETATION:

The correlation matrix reveals the relationships between different attributes in the dataset.

1. AMT_INCOME_TOTAL, AMT_CREDIT, and AMT_ANNUITY are positively correlated, suggesting that clients with higher income tend to apply for larger credits and have higher annuities.
2. AMT_GOODS_PRICE has a positive correlation with AMT_CREDIT and AMT_ANNUITY, indicating that clients who request higher goods prices also apply for larger credits and have higher annuities.
3. AMT_INCOME_TOTAL, AMT_CREDIT, and AMT_ANNUITY have a moderate positive correlation with FAMILY MEMBERS, suggesting that clients with more family members tend to have higher income, request larger credits, and have higher annuities.
4. CHILDREN and FAMILY MEMBERS have a strong positive correlation, indicating that clients with more children tend to have larger families.
5. BIRTH (Years) and EMPLOYMENT (Years) have a strong positive correlation, suggesting that older clients have been employed for longer durations.
6. AMT_CREDIT and AMT_ANNUITY are positively correlated, suggesting that clients with larger credits also have higher annuities.
7. REGION POPULATION has a weak positive correlation with AMT_CREDIT and AMT_ANNUITY, indicating that clients from regions with higher population density may apply for slightly larger credits and have higher annuities.

CORRELATION WHEN T=1 (Client with payment difficulties)

	CAR	*REALTY*	*CHILDREN*	*INCOME*	*CREDIT*	*ANNUITY*	*REGION POPULATION*	*BIRTH (Years)*	*EMPLOYMENT (Years)*	*REGISTRATION (Years)*	*FAMILY MEMBERS*
CAR	1.000	0.019	0.054	-0.001	0.085	0.143	0.028	-0.072	-0.105	-0.046	0.093
REALTY	0.019	1.000	0.001	0.012	-0.021	0.005	0.019	0.104	0.046	0.003	0.009
CHILDREN	0.054	0.001	1.000	0.010	0.008	0.029	-0.020	-0.250	-0.190	-0.152	0.893
INCOME	-0.001	0.012	0.010	1.000	0.015	0.018	-0.006	-0.009	-0.012	0.010	0.013
CREDIT	0.085	-0.021	0.008	0.015	1.000	0.750	0.068	0.143	0.019	0.043	0.061
ANNUITY	0.143	0.005	0.029	0.018	0.750	1.000	0.073	0.009	-0.078	-0.022	0.076
REGION POPULATION	0.028	0.019	-0.020	-0.006	0.068	0.073	1.000	0.016	0.008	0.046	-0.017
BIRTH (Years)	-0.072	0.104	-0.250	-0.009	0.143	0.009	0.016	1.000	0.588	0.289	-0.199
EMPLOYMENT (Years)	-0.105	0.046	-0.190	-0.012	0.019	-0.078	0.008	0.588	1.000	0.192	-0.183
REGISTRATION (Years)	-0.046	0.003	-0.152	0.010	0.043	-0.022	0.046	0.289	0.192	1.000	-0.152
FAMILY MEMBERS	0.093	0.009	0.893	0.013	0.061	0.076	-0.017	-0.199	-0.183	-0.152	1.000

INTERPRETATION:

The correlation matrix shows the relationships between various attributes in the dataset. AMT_ANNUITY, AMT_APPLICATION, AMT_CREDIT, and AMT_GOODS_PRICE are positively correlated, suggesting that clients who apply for larger loan amounts also request higher annuity and goods prices.

1. REGION POPULATION has a weak positive correlation with AMT_CREDIT and AMT_ANNUITY, indicating that clients from regions with higher population density may tend to apply for slightly larger loans.
2. FAMILY MEMBERS has a moderate positive correlation with AMT_APPLICATION and CHILDREN, suggesting that clients with more family members may apply for larger loan amounts and may have more children.
3. BIRTH (Years) and EMPLOYMENT (Years) have a moderate positive correlation, implying that older clients may have been employed for longer durations.
4. CHILDREN has a moderate positive correlation with FAMILY MEMBERS, indicating that clients with more children tend to have larger families.
5. CAR and REALTY have very weak correlations with other variables, suggesting they have little impact on the loan-related attributes.

Here are the variables in order of their importance based on the absolute correlation values:

1. FAMILY MEMBERS (0.893)
2. CREDIT (0.750)
3. ANNUITY (0.750)
4. BIRTH (Years) (0.588)
5. CHILDREN (0.250)
6. EMPLOYMENT (Years) (0.192)

CONCLUSION

1. Most loan applicants fall in the income class of 12650 – 17650 range i.e., 13476. Majority of this income class also pay their instalments on time without facing any payment difficulty. Major number of applicants fall in the income class of less than 5 lakhs i.e., more than 90%.
2. Most loan applicants do not own a car.
3. There is a large number of loan applicants who owns a house or flat. They pay their instalments on time.
4. Most loans applicants do not have any children and they also do not face any loan payment difficulties, means they pay their loan annuity on time.
5. Most loans applicants have only **2** family members and most of them do not face any payment difficulty.
6. More loans have been applied by the **working** people and most of them do not face any payment difficulty.
7. Applicants with **secondary education** seem to be making their payments on time without any delay. Thus, secondary education type applicants have more chance of getting their loan approved. This also means that largest number of loan applicants possess secondary education.
8. **Married** applicants were found to be the one without any payment difficulties, so they are more likely to have their loan approved. Also, they applied for most loans.
9. Applicants living in their own **house / apartment** did not see any payment difficulties, therefore they are more likely to get their loan approved. Also, they applied for more loans.
10. Highest average credit amount is approved for income class 10,25650 - 10,75650 i.e., 26,95500.

Applicants facing payment difficulties are as follows:

1. If the client doesn’t own a car, more chances of default.
2. If the client doesn't own a house. There are more than 30% chance of defaulting.
3. Children impact the loan payment i.e., 35%. More the children, more the chances of default.
4. More family members more chances of default.
5. Unemployed income type faces more difficulty in loan payment i.e., 33.33%.
6. Lower secondary faces more cases of payment difficulty. 11.77%.
7. Civil marriage saw comparatively more payment difficulties then single / unmarried
8. Rented apartment applicants saw more payment difficulties, followed by municipal and co-op apartment.

PART 2. EDA OF PREVIOUS APPLICATION DATA

Tasks to be performed for the EDA analysis are:

- Task A: Identify Missing Data and Deal with it Appropriately
- Task B: Identify Outliers in the Dataset
- Task C: Analyze Data Imbalance
- Task D: Perform Univariate, Segmented Univariate, and Bivariate Analysis
- Task E: Identify Top Correlations for Different Scenarios

Task A: Identify Missing Data and Deal with it Appropriately:

Raw data: It consisted of 37 columns and 50,000 rows.

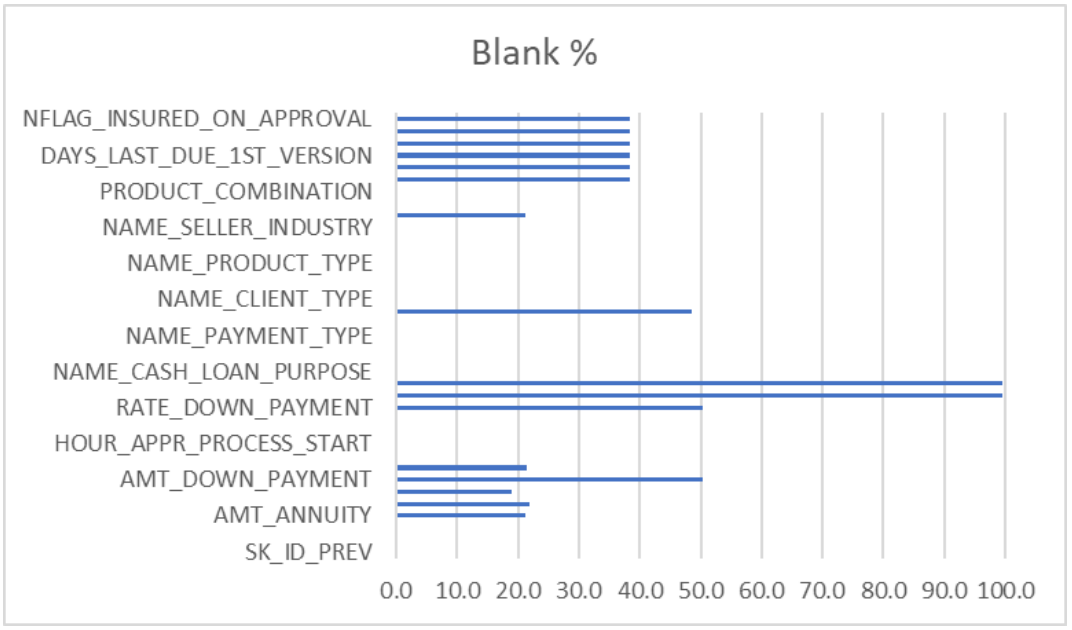
SK_ID_PREV	SK_ID_CURR	NAME_CONTRACT_TYPE	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_DOWN_PAYMENT	AMT_GOODS_PRICE
2030495	271877	Consumer loans	1730.43	17145	17145	0	17145
2802425	108129	Cash loans	25188.615	607500	679671		607500
2523466	122040	Cash loans	15060.735	112500	136444.5		112500
2819243	176158	Cash loans	47041.335	450000	470790		450000
1784265	202054	Cash loans	31924.395	337500	404055		337500
1383531	199383	Cash loans	23703.93	315000	340573.5		315000
2315218	175704	Cash loans		0	0		
1656711	296299	Cash loans		0	0		
2367563	342292	Cash loans		0	0		
2579447	334349	Cash loans		0	0		
1715995	447712	Cash loans	11368.62	270000	335754		270000
2257824	161140	Cash loans	13832.775	211500	246397.5		211500
2330894	258628	Cash loans	12165.21	148500	174361.5		148500
1397919	321676	Consumer loans	7654.86	53779.5	57564	0	53779.5
2273188	270658	Consumer loans	9644.22	26550	27252	0	26550
1232483	151612	Consumer loans	21307.455	126490.5	119853	12649.5	126490.5
2163253	154602	Consumer loans	4187.34	26955	27297	1350	26955
1285768	142748	Revolving loans	9000	180000	180000		180000
2393109	396305	Cash loans	10181.7	180000	180000		180000
1173070	199178	Cash loans	4666.5	45000	49455		45000

Cleaned data:

After cleaning, 26 rows were left and columns having blank percentage greater than 35% were removed in this process.

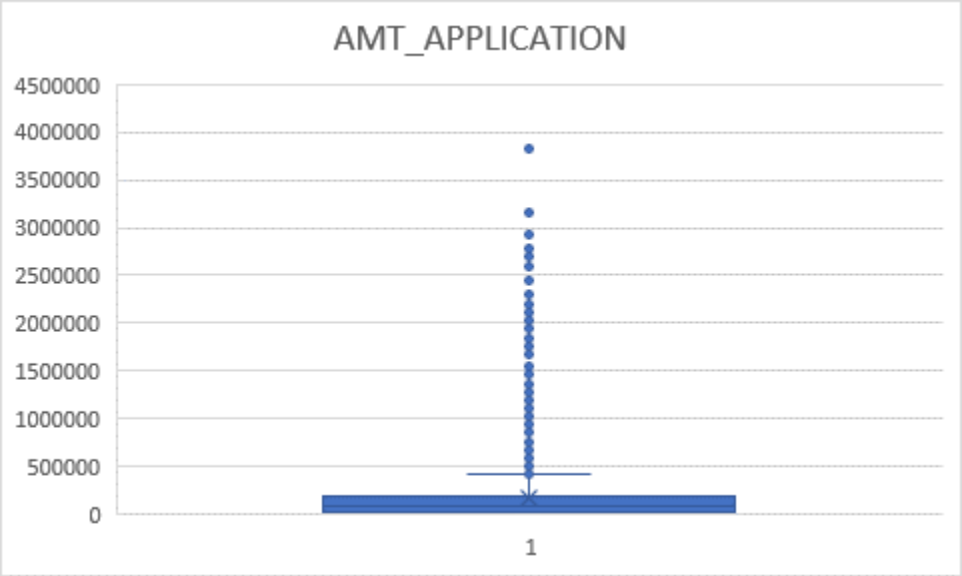
SK_ID_PREV	SK_ID_CURR	NAME_CONTRACT_TYPE	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_GOODS_PRICE	WEEKDAY_APPR_PROCESS_START
2030495	271877	Consumer loans	1730.43	17145	17145	17145	SATURDAY
2802425	108129	Cash loans	25188.615	607500	679671	607500	THURSDAY
2523466	122040	Cash loans	15060.735	112500	136444.5	112500	TUESDAY
2819243	176158	Cash loans	47041.335	450000	470790	450000	MONDAY
1784265	202054	Cash loans	31924.395	337500	404055	337500	THURSDAY
1383531	199383	Cash loans	23703.93	315000	340573.5	315000	SATURDAY
2315218	175704	Cash loans		0	0		TUESDAY
1656711	296299	Cash loans		0	0		MONDAY
2367563	342292	Cash loans		0	0		MONDAY
2579447	334349	Cash loans		0	0		SATURDAY
1715995	447712	Cash loans	11368.62	270000	335754	270000	FRIDAY
2257824	161140	Cash loans	13832.775	211500	246397.5	211500	FRIDAY
2330894	258628	Cash loans	12165.21	148500	174361.5	148500	TUESDAY
1397919	321676	Consumer loans	7654.86	53779.5	57564	53779.5	SUNDAY
2273188	270658	Consumer loans	9644.22	26550	27252	26550	SATURDAY
1232483	151612	Consumer loans	21307.455	126490.5	119853	126490.5	TUESDAY
2163253	154602	Consumer loans	4187.34	26955	27297	26955	SATURDAY
1285768	142748	Revolving loans	9000	180000	180000	180000	FRIDAY
2393109	396305	Cash loans	10181.7	180000	180000	180000	THURSDAY
1173070	199178	Cash loans	4666.5	45000	49455	45000	SATURDAY
1506815	166490	Cash loans	25454.025	450000	491580	450000	MONDAY
1182516	267782	Cash loans	20361.6	405000	451777.5	405000	SATURDAY

Blanks:



Columns	Blank %	Blank count	To be removed (Blank% > 35)
SK_ID_PREV	0.0	0	No
SK_ID_CURR	0.0	0	No
NAME_CONTRACT_TYPE	0.0	0	No
AMT_ANNUITY	21.2	10592	No
AMT_APPLICATION	21.9	10938	No
AMT_CREDIT	18.9	9435	No
AMT_DOWN_PAYMENT	50.4	25198	Yes
AMT_GOODS_PRICE	21.5	10744	No
WEEKDAY_APPR_PROCESS_START	0.0	0	No
HOUR_APPR_PROCESS_START	0.0	0	No
FLAG_LAST_APPL_PER_CONTRACT	0.0	0	No
NFLAG_LAST_APPL_IN_DAY	0.0	0	No
RATE_DOWN_PAYMENT	50.4	25198	Yes
RATE_INTEREST_PRIMARY	99.7	49834	Yes
RATE_INTEREST_PRIVILEGED	99.7	49834	Yes
NAME_CASH_LOAN_PURPOSE	0.0	0	No
NAME_CONTRACT_STATUS	0.0	0	No
DAYS_DECISION	0.0	0	No
NAME_PAYMENT_TYPE	0.0	0	No
CODE_REJECT_REASON	0.0	0	No
NAME_TYPE_SUITE	48.5	24243	Yes
NAME_CLIENT_TYPE	0.0	0	No
NAME_GOODS_CATEGORY	0.0	0	No
NAME_PORTFOLIO	0.0	0	No
NAME_PRODUCT_TYPE	0.0	0	No
CHANNEL_TYPE	0.0	0	No
SELLERPLACE_AREA	0.0	0	No
NAME_SELLER_INDUSTRY	0.0	0	No
CNT_PAYMENT	21.2	10592	No
NAME_YIELD_GROUP	0.0	0	No
PRODUCT_COMBINATION	0.0	8	No
DAYS_FIRST_DRAWING	38.3	19160	Yes
DAYS_FIRST_DUE	38.3	19160	Yes
DAYS_LAST_DUE_1ST_VERSION	38.3	19160	Yes
DAYS_LAST_DUE	38.3	19160	Yes
DAYS_TERMINATION	38.3	19160	Yes
NFLAG_INSURED_ON_APPROVAL	38.3	19160	Yes

Task B: Identify Outliers in the Dataset

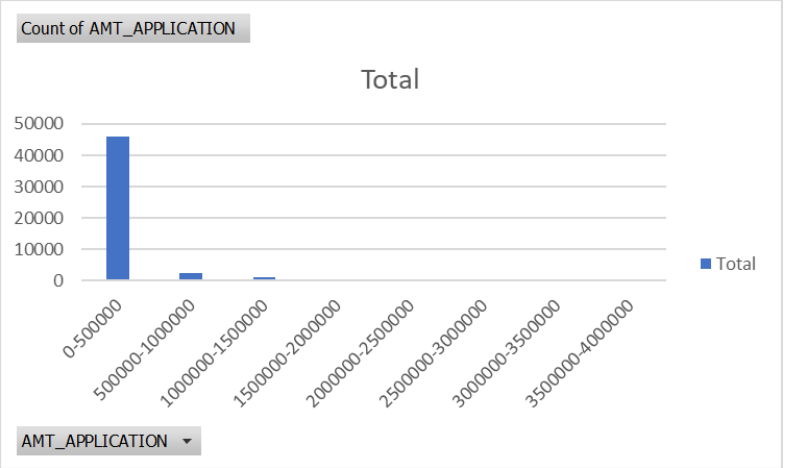


AMT_APPLICATION outlier is 3826373.

Task C: Analyze Data Imbalance

1. **AMT_APPLICATION**

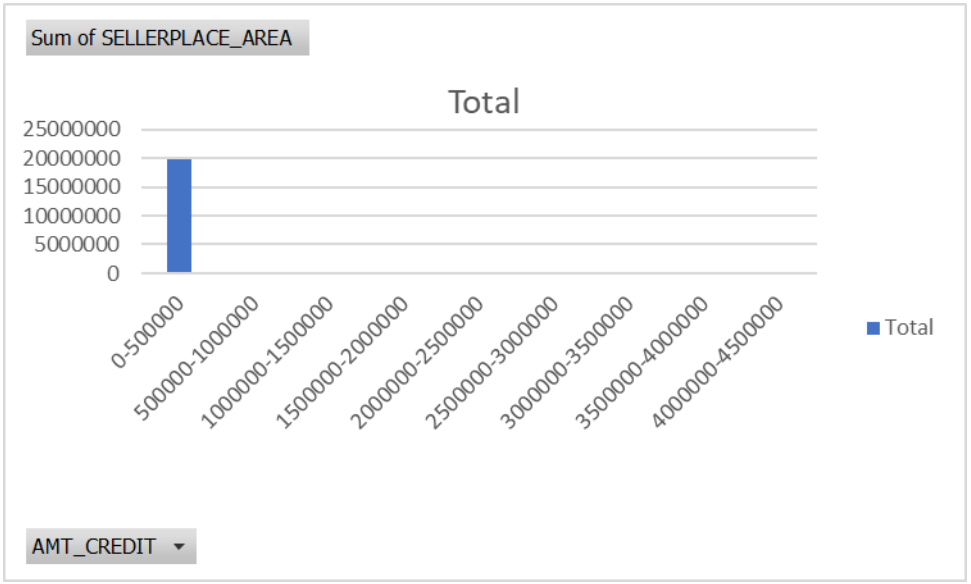
Row Labels	Count of AMT_APPLICATION
0-500000	45950
500000-1000000	2588
1000000-1500000	1149
1500000-2000000	171
2000000-2500000	123
2500000-3000000	9
3000000-3500000	8
3500000-4000000	1
Grand Total	**49999**



Since most of the loan applications are applying in the range of 0 – 5 lakhs, it shows data imbalance.

2. SELLER PLACE

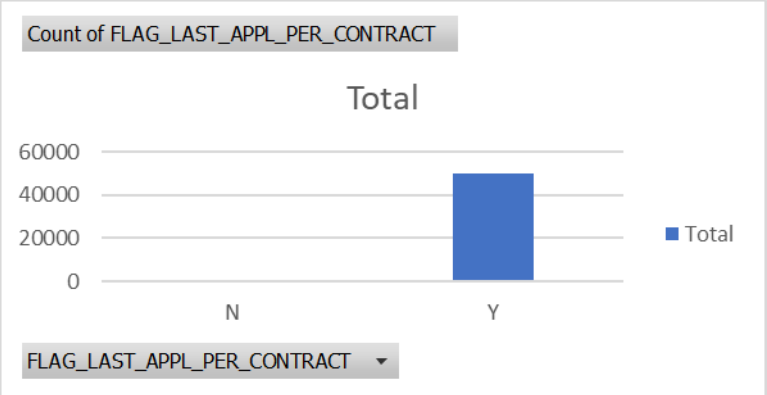
Row Labels	Sum of SELLERPLACE_AREA
0-500000	19801545
500000-1000000	210802
1000000-1500000	61389
1500000-2000000	2200
2000000-2500000	264
2500000-3000000	5992
3000000-3500000	194
3500000-4000000	-2
4000000-4500000	4
Grand Total	**20082388**



Greatest seller place area falls in the range of 0-5 lakhs of credit amount, which shows significant data imbalance.

3. LAST_APPL_PER_CONTRACT

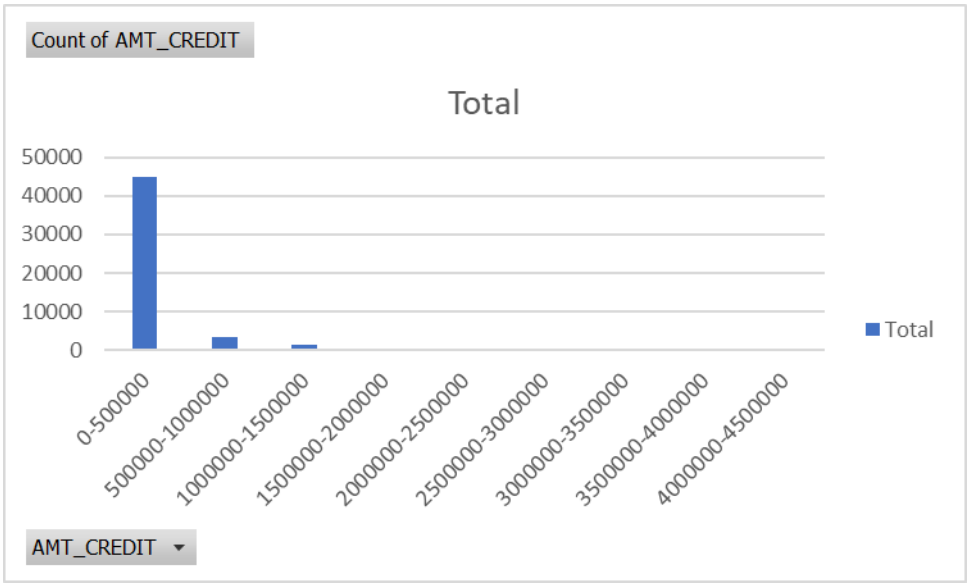
Row Labels	Count of FLAG_LAST_APPL_PER_CONTRACT
N	252
Y	49747
Grand Total	**49999**



Since most applications fall in the category of last application per contract. It shows data imbalance.

4. AMT_CREDIT

Row Labels	Count of AMT_CREDIT
0-500000	44945
500000-1000000	3289
1000000-1500000	1281
1500000-2000000	298
2000000-2500000	137
2500000-3000000	36
3000000-3500000	10
3500000-4000000	2
4000000-4500000	1
Grand Total	**49999**

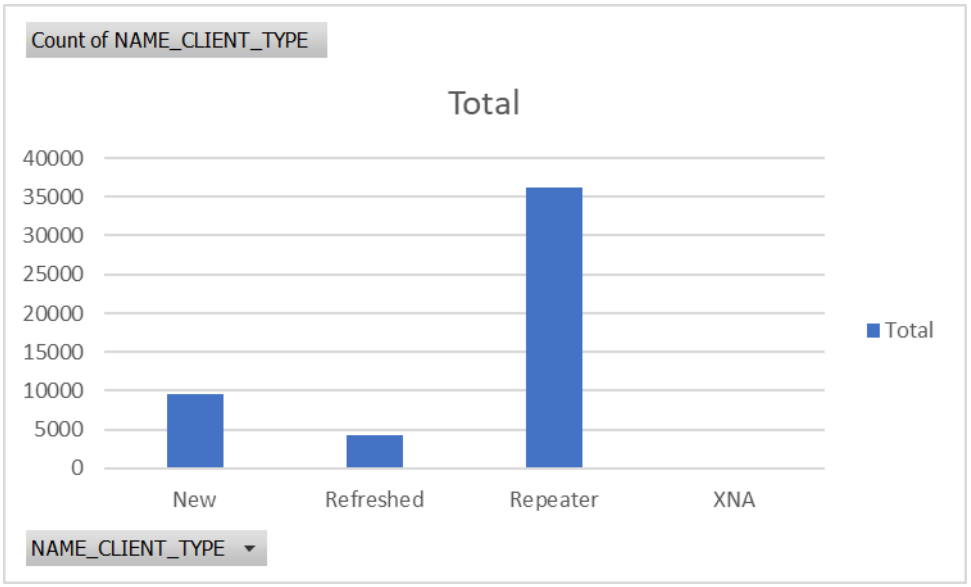


There is a data imbalance in the credit amount issued by the bank since majority of the credit amount falls under the range of **0 – 5 lakhs**.

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

Univariate analysis of CLIENT_TYPE:

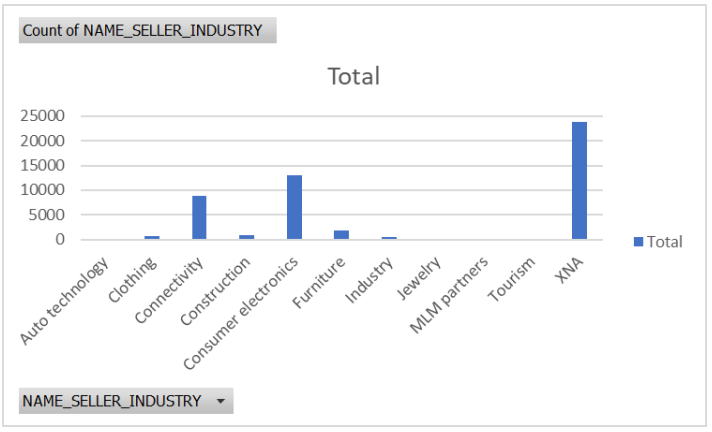
Row Labels	Count of NAME_CLIENT_TYPE
New	9548
Refreshed	4227
Repeater	36167
XNA	57
Grand Total	**49999**



Most loan applicants were found to be from **Repeater** client Type.

UNIVARIATE ANALYSIS OF SELLER INDUSTRY:

Row Labels	Count of NAME_SELLER_INDUSTRY
Auto technology	165
Clothing	774
Connectivity	8783
Construction	988
Consumer electronics	12942
Furniture	1854
Industry	595
Jewelry	83
MLM partners	32
Tourism	12
XNA	23771
Grand Total	**49999**



Under seller industry, **XNA** and **Consumer electronics** industry has the greatest number of applicants.

Bivariate analysis:

AMT_INCOME VS. NAME_CONTRACT_TYPE

Count of NAME_CONTRACT_STATUS	Column Labels				
Row Labels	Approved	Canceled	Refused	Unused offer	Grand Total
0-500000	29548	8558	6987	857	45950
500000-1000000	1597	14	975	2	2588
1000000-1500000	642	16	491		1149
1500000-2000000	54	4	113		171
2000000-2500000	43	3	77		123
2500000-3000000	1		8		9
3000000-3500000			8		8
3500000-4000000			1		1
Grand Total	**31885**	**8595**	**8660**	**859**	**49999**

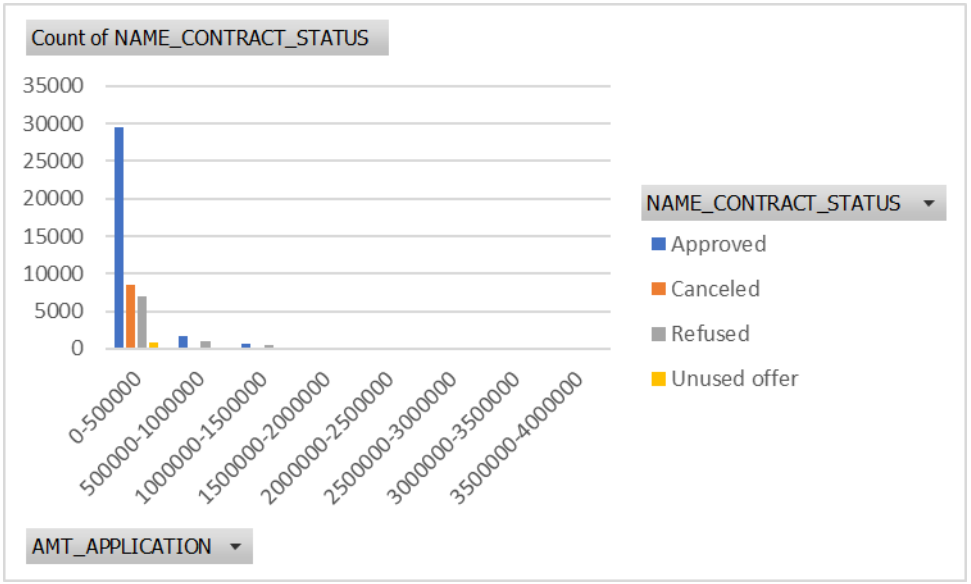
Values in percentage of the row total for better understanding:

Count of NAME_CONTRACT_STATUS	Column Labels				
Row Labels	Approved	Canceled	Refused	Unused offer	Grand Total
0-500000	64.30%	18.62%	15.21%	1.87%	100.00%
500000-1000000	61.71%	0.54%	37.67%	0.08%	100.00%
1000000-1500000	55.87%	1.39%	42.73%	0.00%	100.00%
1500000-2000000	31.58%	2.34%	66.08%	0.00%	100.00%
2000000-2500000	34.96%	2.44%	62.60%	0.00%	100.00%
2500000-3000000	11.11%	0.00%	88.89%	0.00%	100.00%
3000000-3500000	0.00%	0.00%	100.00%	0.00%	100.00%
3500000-4000000	0.00%	0.00%	100.00%	0.00%	100.00%
Grand Total	**63.77%**	**17.19%**	**17.32%**	**1.72%**	**100.00%**

Most Refused loans belong to **2500000 – 3000000** income range in terms of percentage wise comparison i.e., **88.89%**

The Least Refused loans belong to 0 – 5 lakhs income range with only 15.21% refused loans.

This shows that as the income class increases approval rate drops.



Most loan applications were approved for the amount in the range of 0 – 5 lakhs.

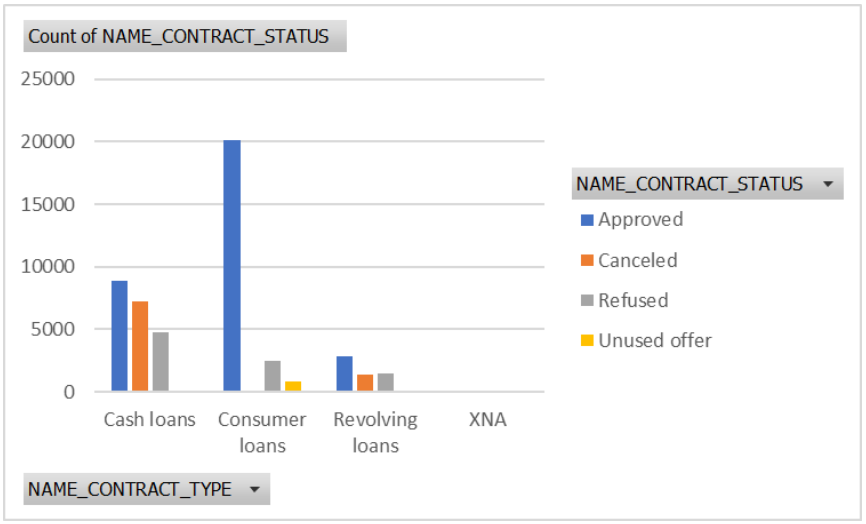
CONTRACT_TYPE VS. CONTRACT_STATUS

Count of NAME_CONTRACT_STATUS	Column Labels				
Row Labels	**Approved**	**Canceled**	**Refused**	**Unused offer**	**Grand Total**
Cash loans	8899	7199	4741	17	20856
Consumer loans	20149	51	2468	842	23510
Revolving loans	2837	1337	1451		5625
XNA		8			8
Grand Total	**31885**	**8595**	**8660**	**859**	**49999**

Values in percentage of the row total for better understanding:

Count of NAME_CONTRACT_STATUS	Column Labels				
Row Labels	**Approved**	**Canceled**	**Refused**	**Unused offer**	**Grand Total**
Cash loans	42.67%	34.52%	22.73%	0.08%	100.00%
Consumer loans	85.70%	0.22%	10.50%	3.58%	100.00%
Revolving loans	50.44%	23.77%	25.80%	0.00%	100.00%
XNA	0.00%	100.00%	0.00%	0.00%	100.00%
Grand Total	**63.77%**	**17.19%**	**17.32%**	**1.72%**	**100.00%**

Revolving loans and **XNA** get refused more compared to other loans.



Under contract type, consumer loans were the most approved loans by the bank.

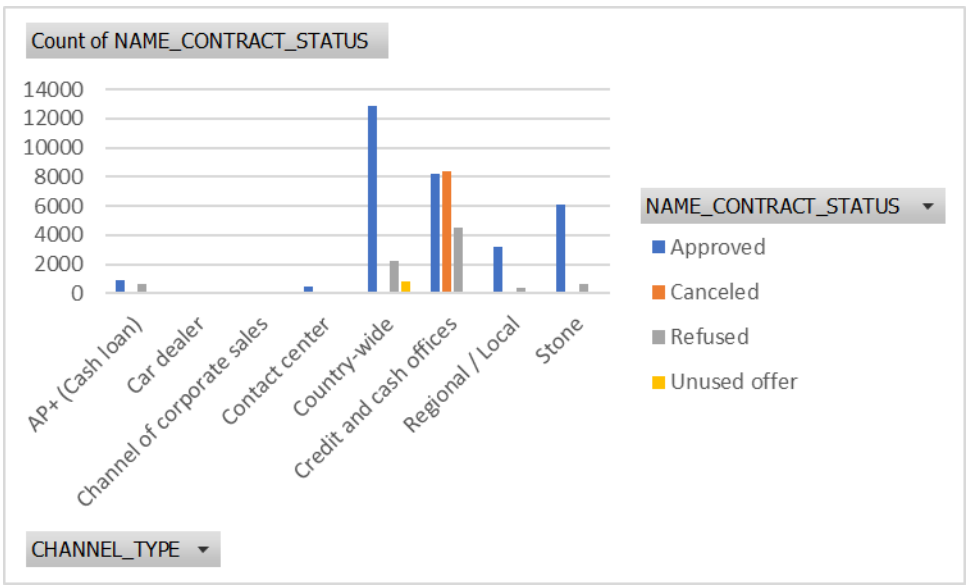
CHANNEL_TYPE VS. CONTRACT STATUS

Count of NAME_CONTRACT_STATUS	Column Labels				
Row Labels	**Approved**	**Canceled**	**Refused**	**Unused offer**	**Grand Total**
AP+ (Cash loan)	901	95	642		1638
Car dealer	13	1	3		17
Channel of corporate sales	112	6	158		276
Contact center	524	90	61		675
Country-wide	12871	45	2192	792	15900
Credit and cash offices	8197	8352	4553	17	21119
Regional / Local	3165	3	369	26	3563
Stone	6102	3	682	24	6811
Grand Total	**31885**	**8595**	**8660**	**859**	**49999**

Values in percentage of the row total for better understanding:

Count of NAME_CONTRACT_STATUS	Column Labels				
Row Labels	**Approved**	**Canceled**	**Refused**	**Unused offer**	**Grand Total**
AP+ (Cash loan)	55.01%	5.80%	39.19%	0.00%	100.00%
Car dealer	76.47%	5.88%	17.65%	0.00%	100.00%
Channel of corporate sales	40.58%	2.17%	57.25%	0.00%	100.00%
Contact center	77.63%	13.33%	9.04%	0.00%	100.00%
Country-wide	80.95%	0.28%	13.79%	4.98%	100.00%
Credit and cash offices	38.81%	39.55%	21.56%	0.08%	100.00%
Regional / Local	88.83%	0.08%	10.36%	0.73%	100.00%
Stone	89.59%	0.04%	10.01%	0.35%	100.00%
Grand Total	**63.77%**	**17.19%**	**17.32%**	**1.72%**	**100.00%**

Channel of corporate sales see more loan refusal (i.e., 57.25%) than approval.



Channel through which most clients were acquired by the bank are **country-wide, credit and cash offices, stone** and **regional/local** channels.

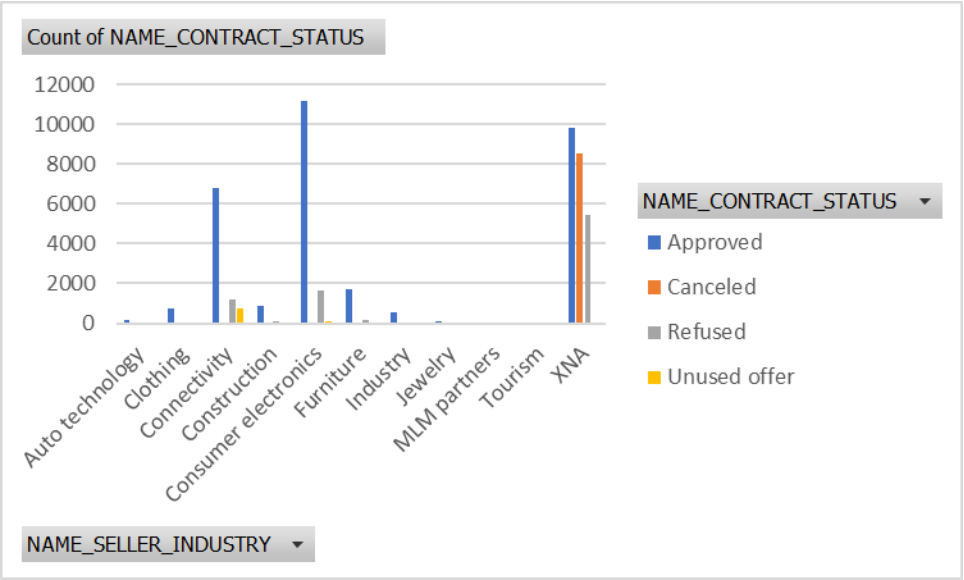
SELLER_INDUSTRY VS. CONTRACT_STATUS

Count of NAME_CONTRACT_STATUS	Column Labels				
Row Labels	**Approved**	**Canceled**	**Refused**	**Unused offer**	**Grand Total**
Auto technology	151		14		165
Clothing	720		53	1	774
Connectivity	6813	40	1213	717	8783
Construction	881		102	5	988
Consumer electronics	11184	9	1635	114	12942
Furniture	1684	2	165	3	1854
Industry	544	1	48	2	595
Jewelry	76		7		83
MLM partners	28		4		32
Tourism	11		1		12
XNA	9793	8543	5418	17	23771
Grand Total	**31885**	**8595**	**8660**	**859**	**49999**

Values in percentage of the row total for better understanding:

Row Labels	**Approved**	**Canceled**	**Refused**	**Unused offer**	**Grand Total**
Auto technology	91.52%	0.00%	8.48%	0.00%	100.00%
Clothing	93.02%	0.00%	6.85%	0.13%	100.00%
Connectivity	77.57%	0.46%	13.81%	8.16%	100.00%
Construction	89.17%	0.00%	10.32%	0.51%	100.00%
Consumer electronics	86.42%	0.07%	12.63%	0.88%	100.00%
Furniture	90.83%	0.11%	8.90%	0.16%	100.00%
Industry	91.43%	0.17%	8.07%	0.34%	100.00%
Jewelry	91.57%	0.00%	8.43%	0.00%	100.00%
MLM partners	87.50%	0.00%	12.50%	0.00%	100.00%
Tourism	91.67%	0.00%	8.33%	0.00%	100.00%
XNA	41.20%	35.94%	22.79%	0.07%	100.00%
Grand Total	**63.77%**	**17.19%**	**17.32%**	**1.72%**	**100.00%**

XNA and **MLM partners** seller industry see more refusal on seeing their percentage of the total.



Loans from the consumer electronics, XNA and connectivity seller industries were readily approved compared to other industries, indicating more importance of these industries.

Task E: Identify Top Correlations for Different Scenarios

	AMT_ANNUITY	*AMT_APPLICATION*	*AMT_CREDIT*	*AMT_GOODS_PRICE*	*SELLERPLACE_AREA*	*CNT_PAYMENT*
AMT_ANNUITY	1	0.812970626	0.819841718	0.825564271	-0.010295138	0.385875334
AMT_APPLICATION	0.812970626	1	0.975771049	0.999902539	-0.003965725	0.671340368
AMT_CREDIT	0.819841718	0.975771049	1	0.993495986	-0.004949463	0.666621175
AMT_GOODS_PRICE	0.825564271	0.999902539	0.993495986	1	-0.007737876	0.663684765
SELLERPLACE_AREA	-0.010295138	-0.003965725	-0.004949463	-0.007737876	1	-0.001015924
CNT_PAYMENT	0.385875334	0.671340368	0.666621175	0.663684765	-0.001015924	1

VARIABLES:

Based on the absolute values of the correlation coefficients, the variables can be listed in decreasing order of importance:

1. AMT_GOODS_PRICE (0.825564271)
2. AMT_APPLICATION (0.812970626)
3. AMT_CREDIT (0.819841718)
4. AMT_ANNUITY (0.825564271)
5. CNT_PAYMENT (0.671340368)
6. SELLERPLACE_AREA (0.010295138)

INTERPRETATION:

The variables with the highest correlation are AMT_GOODS_PRICE, AMT_CREDIT, and AMT_APPLICATION, indicating strong relationships among them. CNT_PAYMENT and AMT_ANNUITY also exhibit moderate correlations with the other variables. SELLERPLACE_AREA shows very weak correlations with the rest of the variables, suggesting little or no relationship with them.

CONCLUSION:

1. Most loan applicants were found to be from **Repeater** client Type.
2. Under seller industry, **XNA** and **Consumer electronics** industry has the greatest number of applicants.
3. Most loan applications were approved for the amount in the range of 0 – 5 lakhs.
4. Under contract type, consumer loans were the most approved loans by the bank.
5. Channel through which most clients were acquired by the bank are **country-wide, credit and cash offices, stone** and **regional/local** channels.
6. Loans from the consumer electronics, XNA and connectivity seller industries were readily approved compared to other industries, indicating more importance of these industries.

Patterns of loan refusal:

1. Most Refused loans belong to **2500000 – 3000000** income range in terms of percentage wise comparison i.e., **88.89%**
   The Least Refused loans belong to 0 – 5 lakhs income range with only 15.21% refused loans.
   This shows that as the income class increases approval rate drops.
2. **Revolving loans** and **XNA** get refused more compared to other loans.
3. **Channel of corporate sales** see more loan refusal (i.e., 57.25%) than approval.
4. **XNA** and **MLM partners** seller industry see more refusal on seeing their percentage of the total.

PART 3: MERGING SHEETS ON COMMON COLUMN:

To combine the datasets application_data.csv and previous_application.csv in Excel, I used the IFERROR, INDEX and MATCH functions.

Common column in both dataset: "SK_ID_CURR"

Formula used: =IFERROR(INDEX([Sheet2]cleaned_previous_data!$M:$M, MATCH($A2, [Sheet2]cleaned_previous_data!$B:$B, 0)), "No Match Found")

SK_ID_CURR	TARGET	CONTRACT_STATUS	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN
100002	1	No Match Found	Cash loans	M	N	Y	0
100003	0	No Match Found	Cash loans	F	N	N	0
100004	0	No Match Found	Revolving loans	M	Y	Y	0
100006	0	No Match Found	Cash loans	F	N	Y	0
100007	0	Approved	Cash loans	M	N	Y	0
100008	0	No Match Found	Cash loans	M	N	Y	0
100009	0	Approved	Cash loans	F	Y	Y	1
100010	0	No Match Found	Cash loans	M	Y	Y	0
100011	0	No Match Found	Cash loans	F	N	Y	0
100012	0	Approved	Revolving loans	M	N	Y	0
100014	0	No Match Found	Cash loans	F	N	Y	1
100015	0	No Match Found	Cash loans	F	N	Y	0
100016	0	No Match Found	Cash loans	F	N	Y	0
100017	0	No Match Found	Cash loans	M	Y	N	1
100018	0	No Match Found	Cash loans	F	N	Y	0
100019	0	No Match Found	Cash loans	M	Y	Y	0
100020	0	No Match Found	Cash loans	M	N	N	0
100021	0	No Match Found	Revolving loans	F	N	Y	1
100022	0	No Match Found	Revolving loans	F	N	Y	0
100023	0	No Match Found	Cash loans	F	N	Y	1
100024	0	No Match Found	Revolving loans	M	Y	Y	0
100025	0	No Match Found	Cash loans	F	Y	Y	1

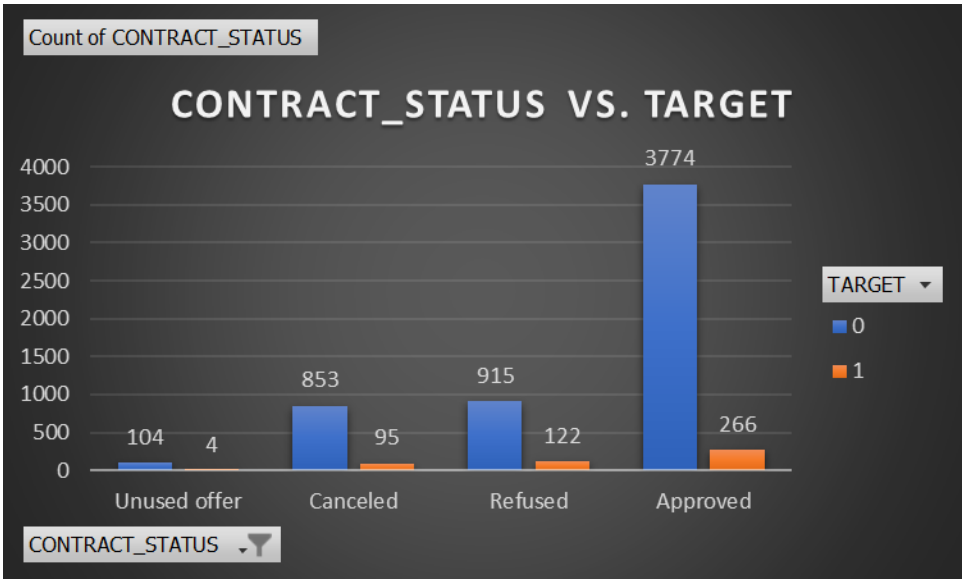
I brought CONTRACT_STATUS next to TARGET column using the above excel function for easier analysis using merged data from both the datasets.

CONTRACT STATUS VS. TARGET

Count of CONTRACT_STATUS	Column Labels		
Row Labels	**0**	**1**	**Grand Total**
Unused offer	104	4	108
Canceled	853	95	948
Refused	915	122	1037
Approved	3774	266	4040
Grand Total	**5646**	**487**	**6133**

Values in percentage of the row total for better understanding:

Count of CONTRACT_STATUS	Column Labels		
Row Labels	**0**	**1**	**Grand Total**
Unused offer	96.30%	3.70%	100.00%
Canceled	89.98%	10.02%	100.00%
Refused	88.24%	11.76%	100.00%
Approved	93.42%	6.58%	100.00%
Grand Total	**92.06%**	**7.94%**	**100.00%**



This shows that clients having no payment difficulties where easily approved for the loan since their numbers are high, whereas those clients with payment difficulties were not easily approved for loan.

Final conclusion:

Variables based on their significance in predicting loan default:

- Less income groups have more chances of loan default.
- People living in a rented accommodation and those who do not own a house have more chances of loan default.
- More the children, more are the chances of loan default.
- Unemployed income groups have more chances of loan default.
- Less educated applicants have more chances of loan default.
- More family members more chances of default. i.e., more than 2.