Bank Loan Case Study

Project Description: In this project, we will conduct Exploratory Data Analysis (EDA) to tackle a critical challenge faced by our finance company specializing in lending to urban customers. We aim to identify patterns in loan application data and understand which type of people may struggle to repay the loan. We will conduct a thorough analysis.

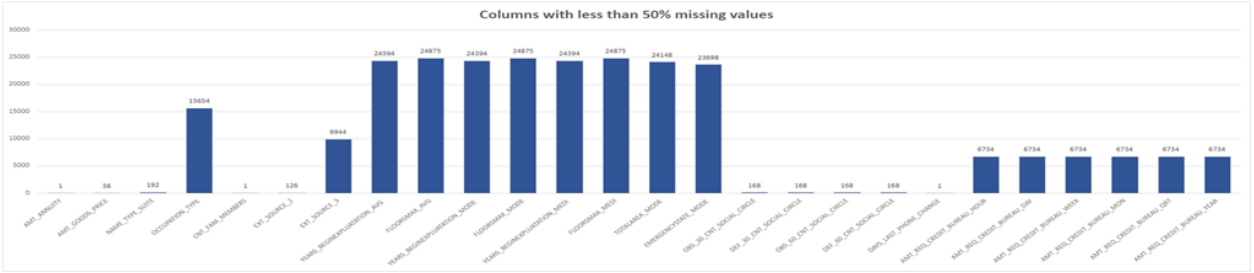
Approach: We have a huge application dataset which we will use to conduct our analysis. We will need to handle missing values and find potential outliers. Once those steps are done, we will start conducting a thorough analysis.

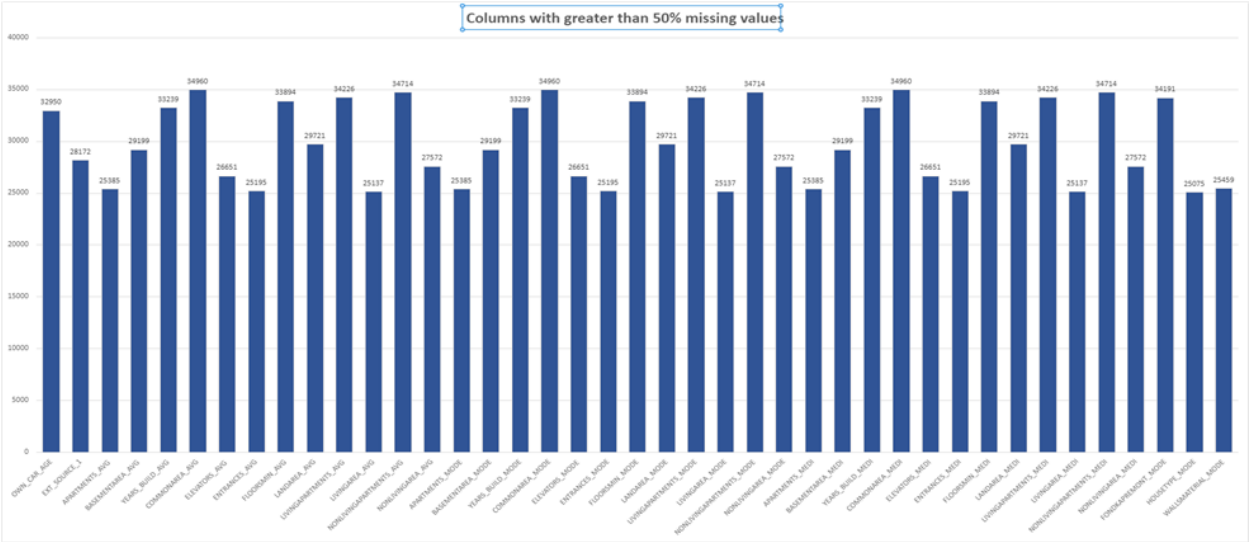
Tech-Stack Used:
Microsoft Excel: This tool served as the backbone of our analysis, offering data cleaning, exploration, visualization, and basic statistical analysis capabilities.

Insights:

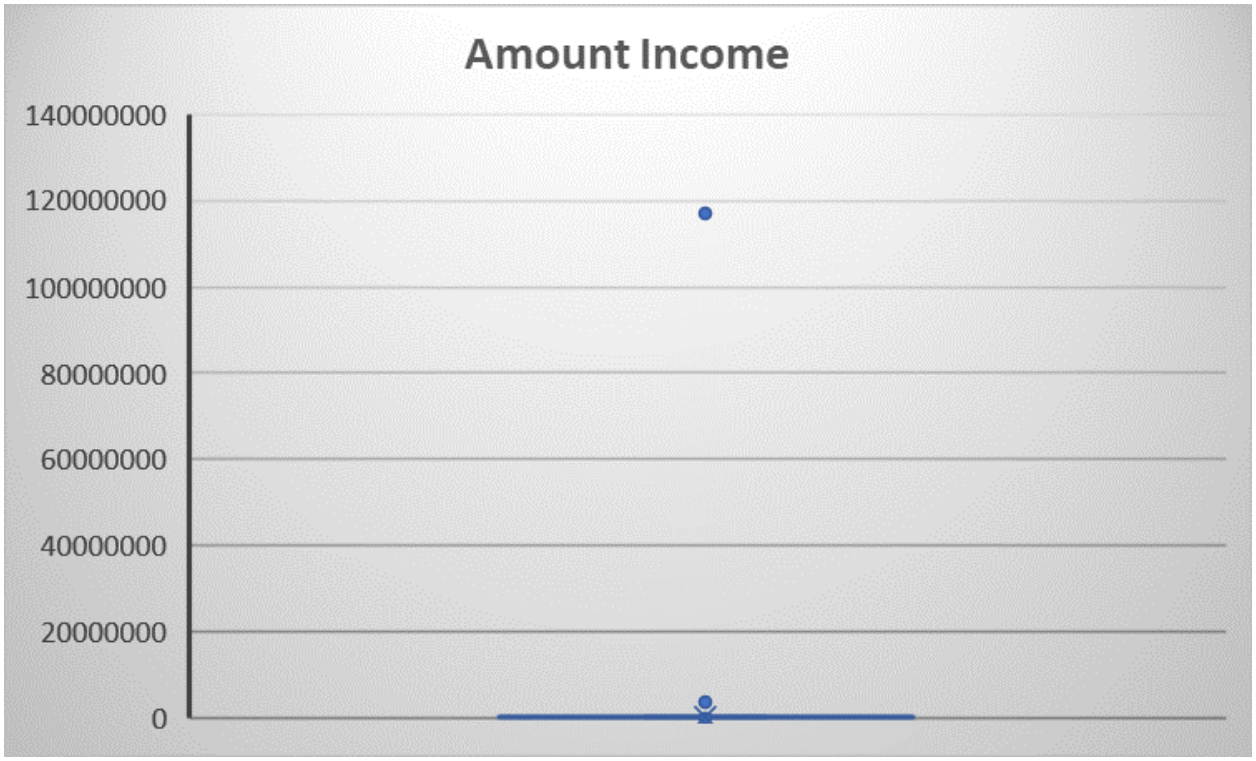
1. Identify Missing Data and Deal with it Appropriately:

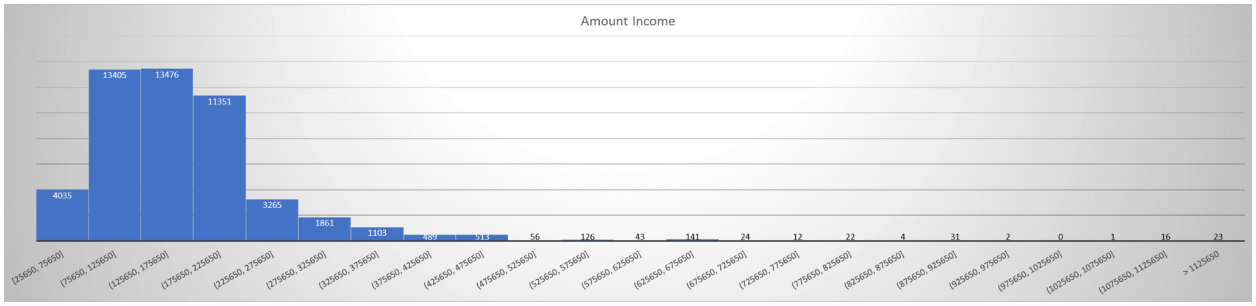
- AMT_ANNUITY, AMT_GOODS_PRICE, AMT_REQ_CREDIT_BUREAU_HOUR,AMT_REQ_CREDIT_BUREAU_DAY, AMT_REQ_CREDIT_BUREAU_WEEK, AMT_REQ_CREDIT_BUREAU_MON, AMT_REQ_CREDIT_BUREAU_QRT, AMT_REQ_CREDIT_BUREAU_YEAR
  The skewness of all this field is more than 1 which means it is right skewed. So median imputation is a better approach for imputation
- In NAME_TYPE_SUITE there are lots of blank values. So wherever it is blank "Unknown" was imputed
- If the applicant does own a car then OWN_CAR_AGE was kept empty. But as he/she doesn't have a car we can put 0 as own_car_age
- All the fields which has loads of data missing were deleted
- The following graph will show the amount of data missing in columns





2. Identify Outliers in the Dataset:
    - AMT_INCOME_TOTAL: As in the box plot we can see that there is wide range of data points for incomes which means there are applicant whose incomes are more compared to median income. But this doesn't give a proper picture of how many outliers we have due to a long spread. So to identify the outlier we will use a histogram to see the distribution.





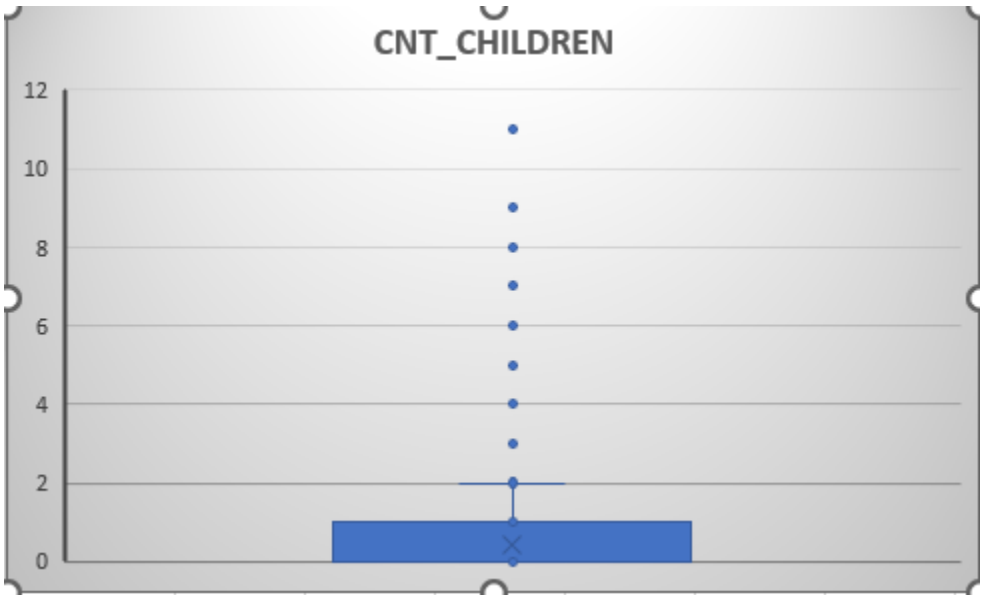
Now we can see the exact distribution of income ranges. Now we can dig deeper by using conditional formatting and find are these income ranges are valid or not.

For example the most of the people who earns more than 1000000 have a Higher education or Secondary / secondary special degree.

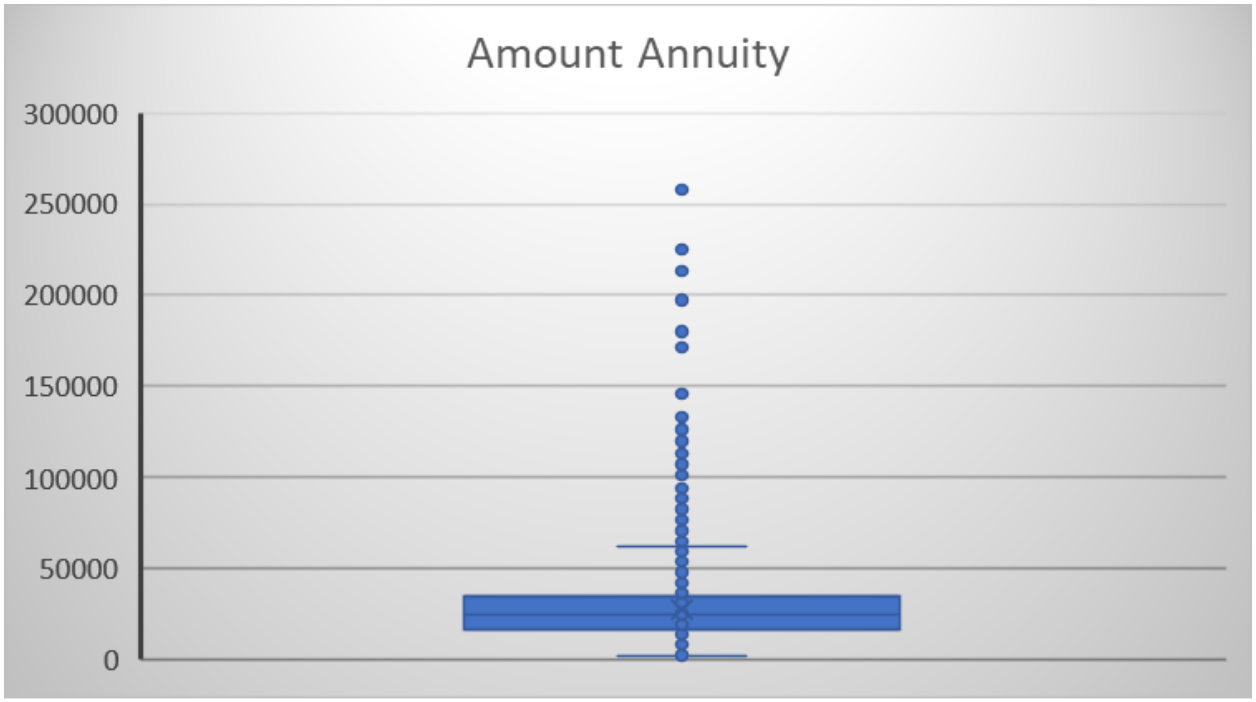
- AMT_CREDIT: If I analyze the top outlier i.e.4050000, so the question arises why the bank has given this big credit to someone. After I digged further I found their annual income was more than 7 lakhs which if we see the above only 12 people have that's why they got this credit. So it's a valid outlier.

- Days Employed in years: here there is outlier which is wrong as no one can work for 1000 years

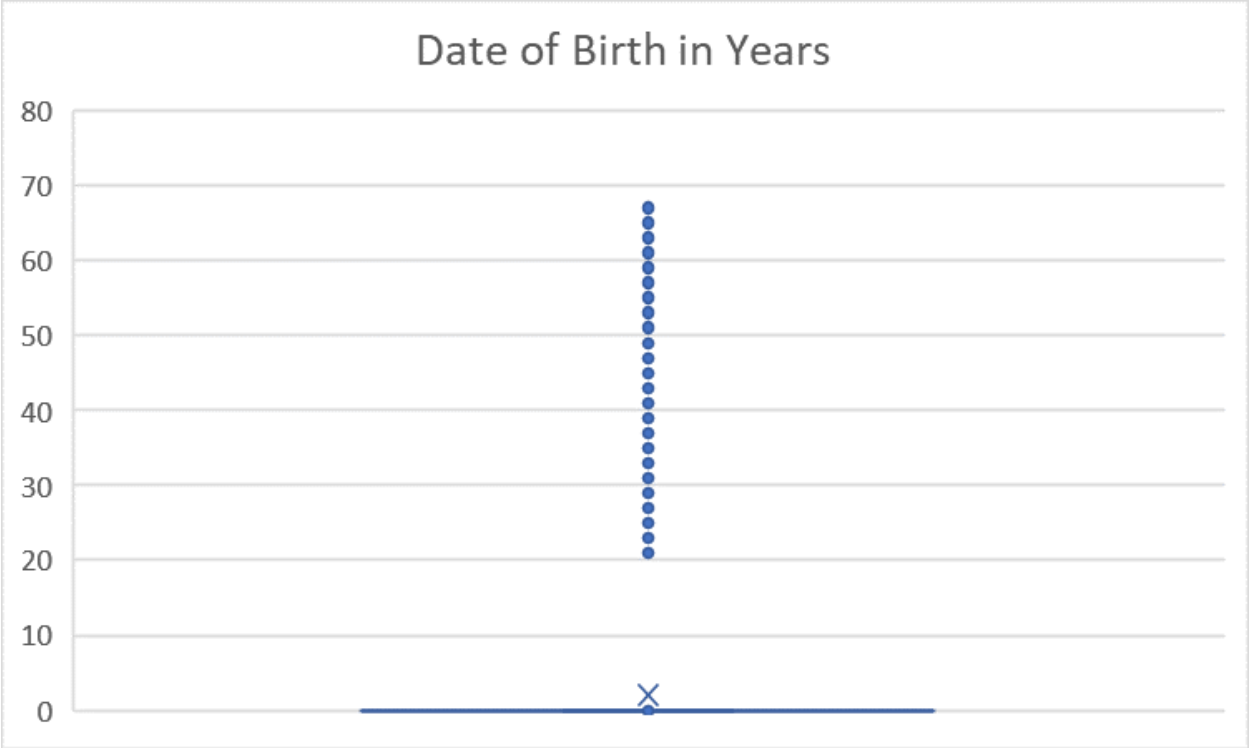
- CNT_CHILDREN: I investigated the person with 11 children. So bank should investigate over here and he is struggling with loan repayment too.



- All other graphs I will attach below

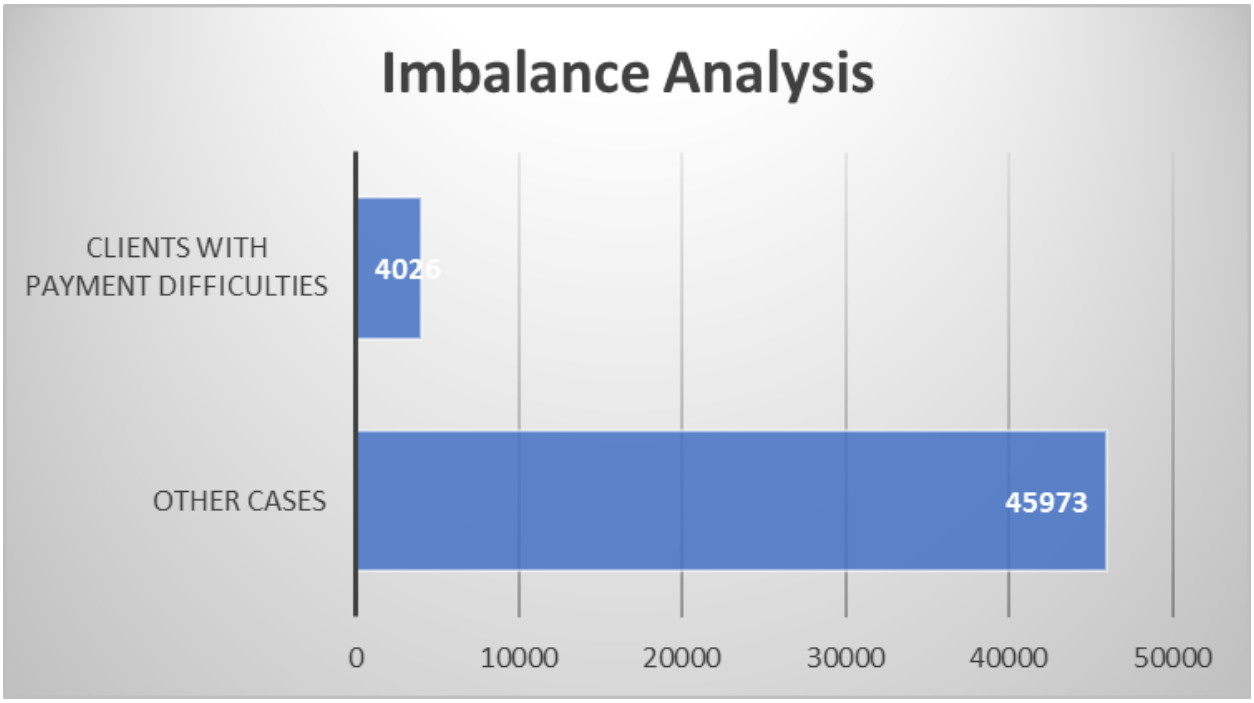

Amount Annuity
300000
250000
200000
150000
100000
50000
0


Amount Good Price
4500000
4000000
3500000
3000000
2500000
2000000
1500000
1000000
500000
0

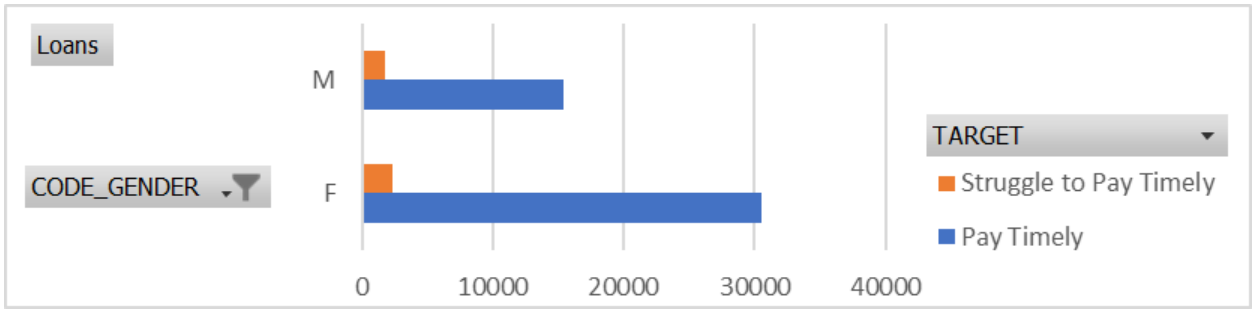


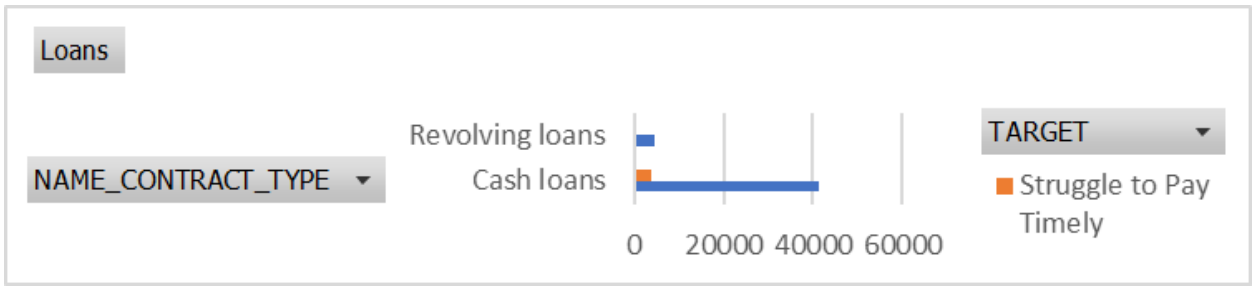
So this says we have outliers but most of them are valid ones.

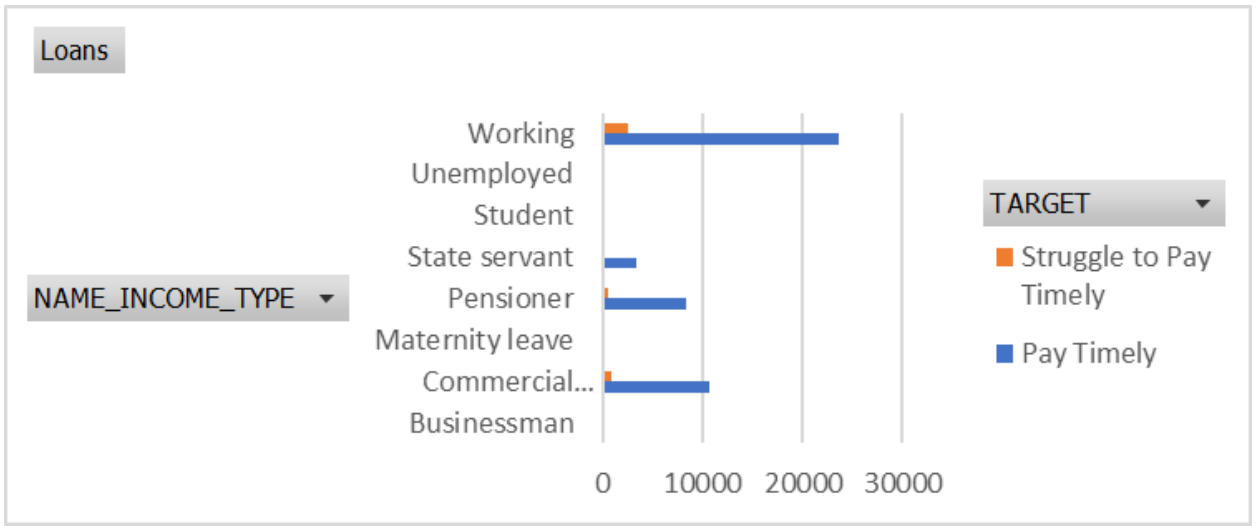
3. Analyze Data Imbalance: The data is highly imbalanced as we have only few data points for clients with payment difficulties.

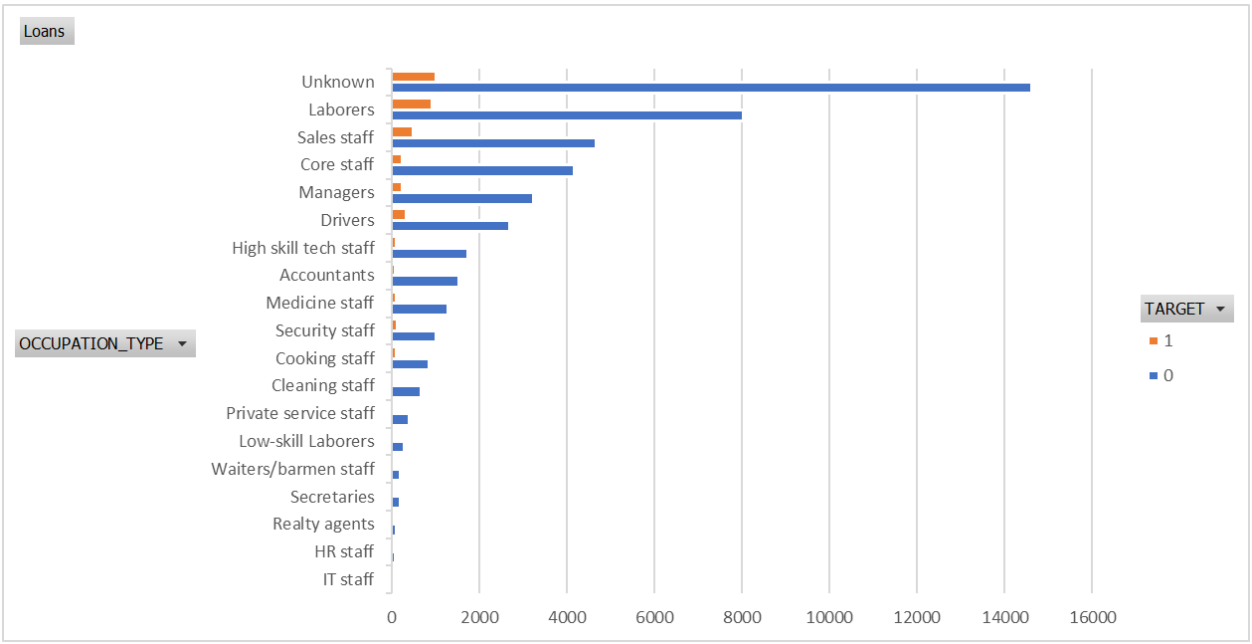


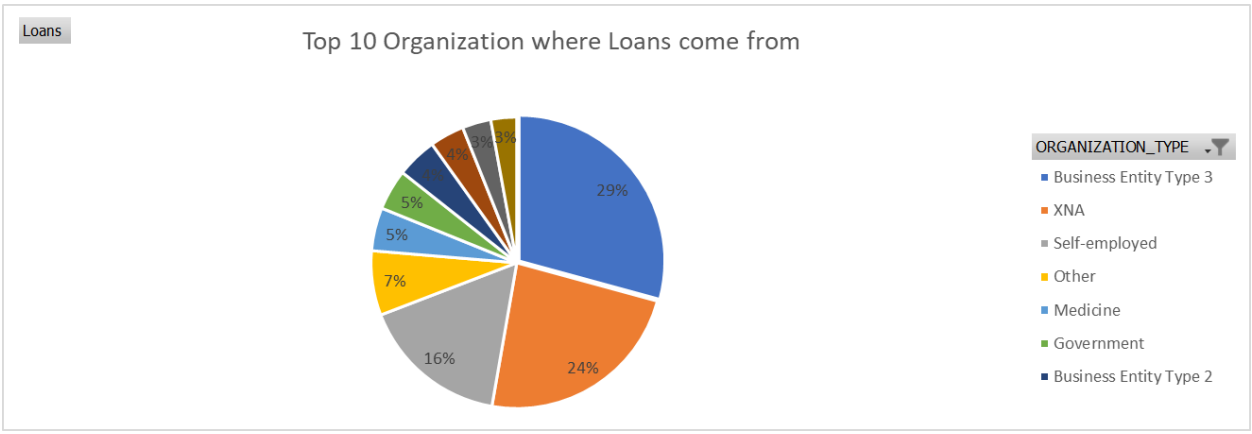
4. Perform Univariate, Segmented Univariate, and Bivariate Analysis:
   The graphs are self explanatory. They tells who should we deal with while lending money,

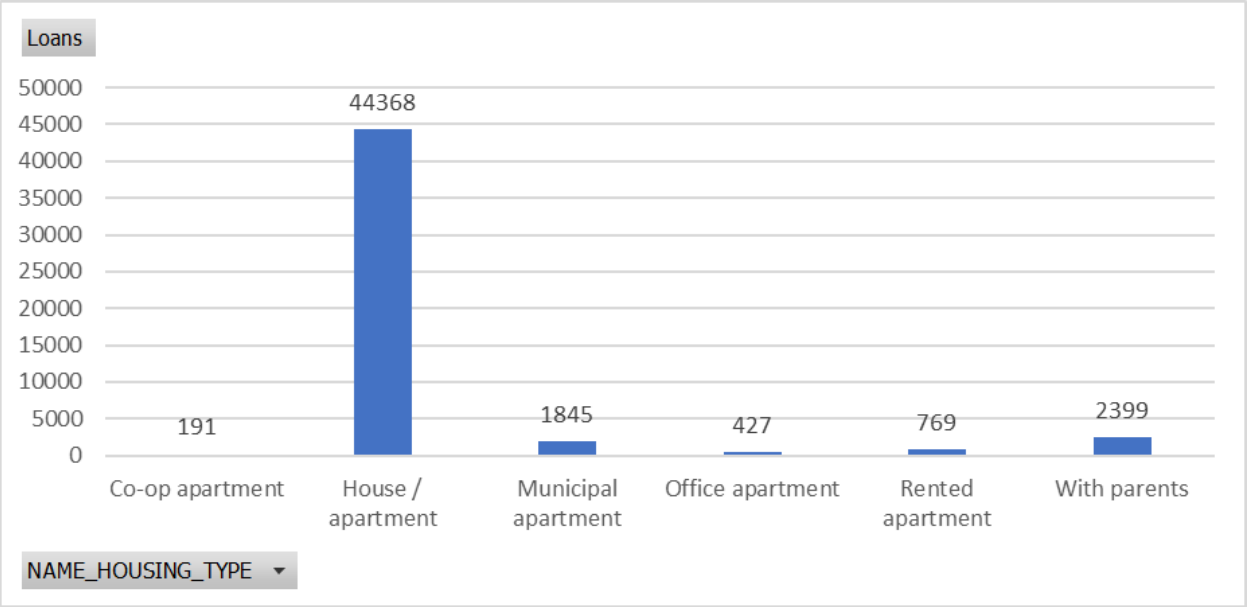


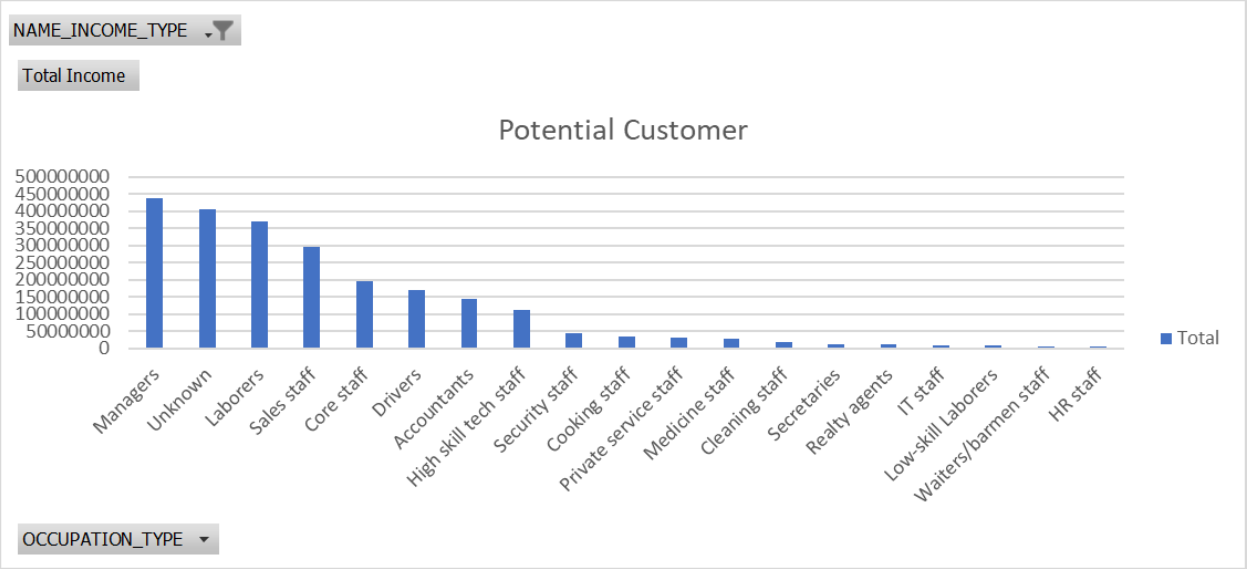





Loans
Unknown
Laborers
Sales staff
Core staff
Managers
Drivers
High skill tech staff
Accountants
Medicine staff
Security staff
Cooking staff
Cleaning staff
Private service staff
Low-skill Laborers
Waiters/barmen staff
Secretaries
Realty agents
HR staff
IT staff
0
2000
4000
6000
8000
10000
12000
14000
16000
OCCUPATION_TYPE
TARGET
1
0


Loans
Top 10 Organization where Loans come from
29%
24%
16%
7%
5%
5%
4%
4%
3%
3%
ORGANIZATION_TYPE
Business Entity Type 3
XNA
Self-employed
Other
Medicine
Government
Business Entity Type 2





Just one suggestion to the bank. There are lots of loans given to people of whom we don't know their occupation type. This could potentially harm the bank in future. Other than that everything is good.

5. Identify Top Correlations for Different Scenarios: Here's a heat for both the targets

Target 0	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	DAYS_BIRTH_YEAR	DAYS_ID_PUBLISH_YEAR	DAYS_REGISTRATION_YEAR	DAYS_EMPLOYED_YEAR
CNT_CHILDREN	1	0.036319722	0.00570546	0.02638396	0.001383048	-0.024912809	-0.335668312	0.032864628	-0.183013906	-0.245517957
AMT_INCOME_TOTAL	0.036319722	1	0.37796575	0.451135167	0.384486402	0.181941261	-0.073713961	-0.032128535	-0.068837498	-0.161680117
AMT_CREDIT	0.005705458	0.377965752	1	0.770772818	0.986879648	0.095539444	0.051061141	0.008706122	-0.007914754	-0.074731255
AMT_ANNUITY	0.02638396	0.451135167	0.77077282	1	0.775888784	0.11727925	-0.009906883	-0.008809879	-0.034444538	-0.111292131
AMT_GOODS_PRICE	0.001383048	0.384486402	0.98687965	0.775888784	1	0.099047191	0.049089272	0.010144545	-0.011073951	-0.07223485
REGION_POPULATION_RELATIVE	-0.024912809	0.181941261	0.09553944	0.11727925	0.099047191	1	0.030551402	0.00273229	0.058515409	-0.006761312
DAYS_BIRTH_YEAR	-0.335668312	-0.073713961	0.05106114	-0.009906883	0.049089272	0.030551402	1	0.268698355	0.334863246	0.623405173
DAYS_ID_PUBLISH_YEAR	0.032864628	-0.032128535	0.00870612	-0.008809879	0.010144545	0.00273229	0.268698355	1	0.103323404	0.273957783
DAYS_REGISTRATION_YEAR	-0.183013906	-0.068837498	-0.00791475	-0.034444538	-0.011073951	0.058515409	0.334863246	0.103323404	1	0.20884097
DAYS_EMPLOYED_YEAR	-0.245517957	-0.161680117	-0.07473126	-0.111292131	-0.07223485	-0.006761312	0.623405173	0.273957783	0.20884097	1

Target 1	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	DAYS_BIRTH_YEAR	DAYS_ID_PUBLISH_YEAR	DAYS_REGISTRATION_YEAR	DAYS_EMPLOYED_YEAR
CNT_CHILDREN	1	0.032396762	0.03501117	0.023326859	0.033510666	-0.020359154	-0.250025578	-0.2496732	-0.189769048	-0.152472925
AMT_INCOME_TOTAL	0.010110177	1	0.01527144	0.018004594	0.013298258	-0.006180303	-0.009117347	-0.009033662	-0.011760562	0.00951005
AMT_CREDIT	0.007601905	0.015271444	1	0.749665201	0.982381964	0.067775624	0.142433858	0.142506035	0.018794834	0.043133743
AMT_ANNUITY	0.029172977	0.018004594	0.7496652	1	0.749904665	0.073123998	0.00862531	0.008751713	-0.078105952	-0.021181972
AMT_GOODS_PRICE	-0.000680095	0.013298258	0.98238196	0.749904665	1	0.077209215	0.141010067	0.14111346	0.023429344	0.043136615
REGION_POPULATION_RELATIVE	-0.020359154	-0.006180303	0.06777562	0.073123998	0.077209215	1	0.01650827	0.016468731	0.007720967	0.04632804
DAYS_BIRTH_YEAR	-0.250025578	-0.009117347	0.14243386	0.00862531	0.141010067	0.01650827	1	0.999680123	0.588259703	0.287855797
DAYS_ID_PUBLISH_YEAR	-0.2496732	-0.009033662	0.14250603	0.008751713	0.14111346	0.016468731	0.999680123	1	0.588227606	0.28784217
DAYS_REGISTRATION_YEAR	-0.189769048	-0.011760562	0.01879483	-0.078105952	0.023429344	0.007720967	0.588259703	0.588227606	1	0.191729307
DAYS_EMPLOYED_YEAR	-0.152472925	0.00951005	0.04313374	-0.021181972	0.043136615	0.04632804	0.287855797	0.28784217	0.191729307	1

Result:

Loan Highly Recommended Groups:

1. Previous application approved clients
2. Senior clients
3. More educated clients
4. Customers with a High Income
5. Females
6. Customers with strong work experience

Loan High Risk Groups:

1. Clients that are unemployed
2. Youth clients
3. Customers whose prior applications were denied
4. Low-income clients
5. Customers with little work experience

Excel link: [application_data.xlsx]

.