BANK LOAN CASE STUDY

Project Description

The main aim of this project is to identify patterns that indicate if a customer will have difficulty paying their instalments. 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.

Approach and Tech used

I have performed EDA on the Bank loan data set. It includes understanding the data, identifying and handling the missing data and outliers in the data, then I checked the data imbalance in the data set. I have also done univariate analysis, bivariate and multivariate analysis. And lastly I have checked the correlation between some features of the data set.

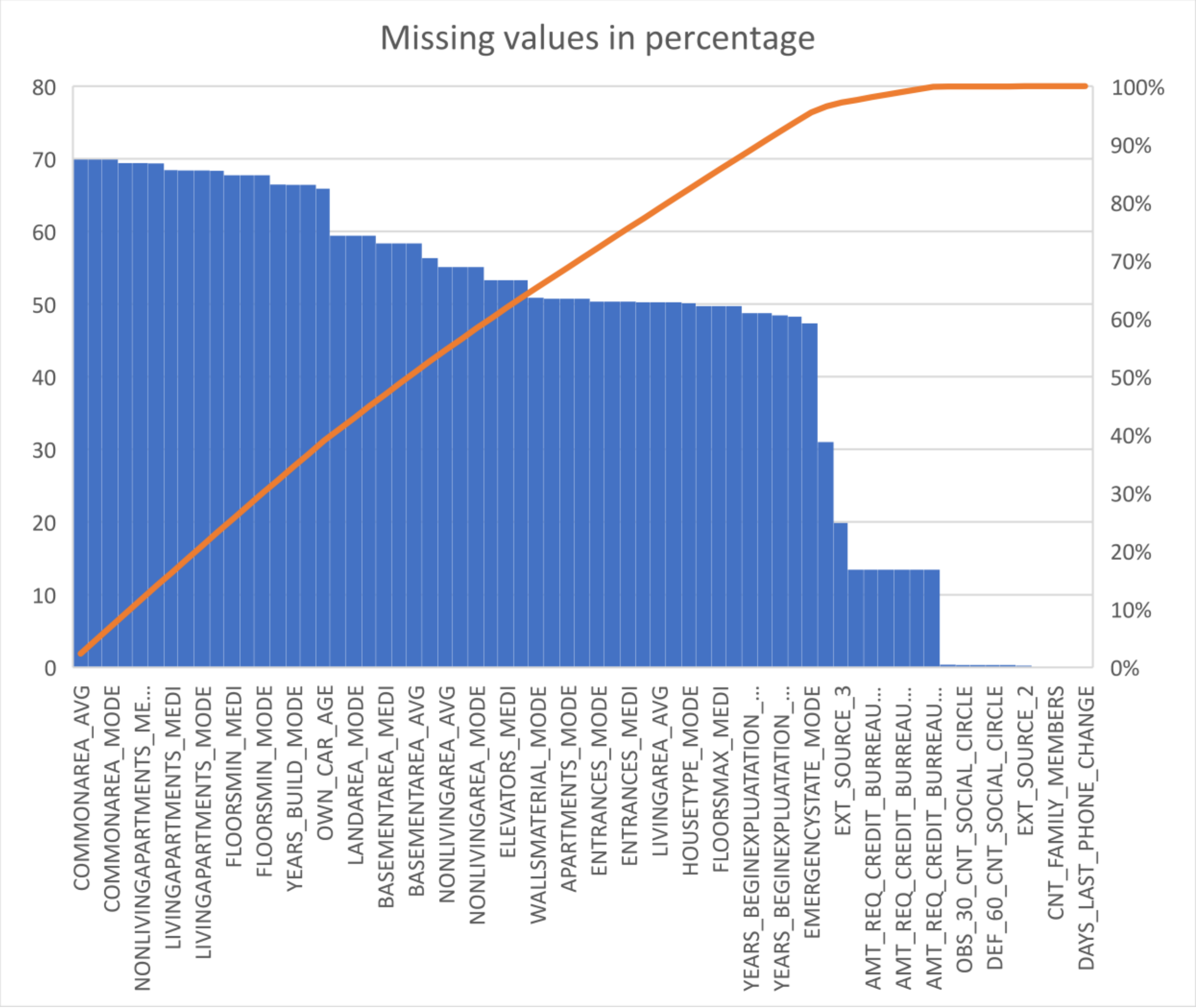
The tech used in this project is Microsoft excel. Excel is a versatile tool for data analysis due to its user-friendly interface and widespread accessibility. Its features allows quick data manipulation, facilitating easy organization and visualization. For straightforward analyses or initial data exploration, Excel's pivot tables, charts, and formulas provide a solid foundation. Excel remains an efficient and practical tool for basic data analysis and quick insights.

Insights

Application Data

1) Missing value Analysis-
    - It is essential to handle missing data effectively to ensure the accuracy of the analysis.
    - Common area average, common area mode and nonliving apartments median have around 89% missing values which means more than half of the data is missing.

- Dropping the columns with more than 25% missing data. 50 columns dropped.
- Missing values in Occupation type, code gender and EXT source were replaced by mode.
- The other columns containing missing data less than 1%, the null value rows were dropped from those columns.



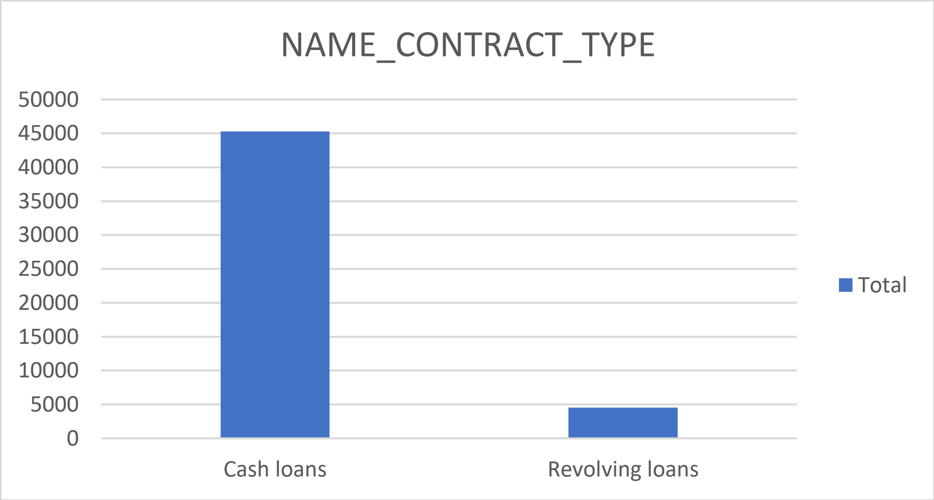
2) Outlier Analysis –

- Outliers can significantly impact the analysis and distort the results.
- I have plotted boxplots to identify the outliers.
- AMT_CREDIT, AMT_ANNUITY, AND CNT_FAM_MEMEBERS are observed to have outliers.

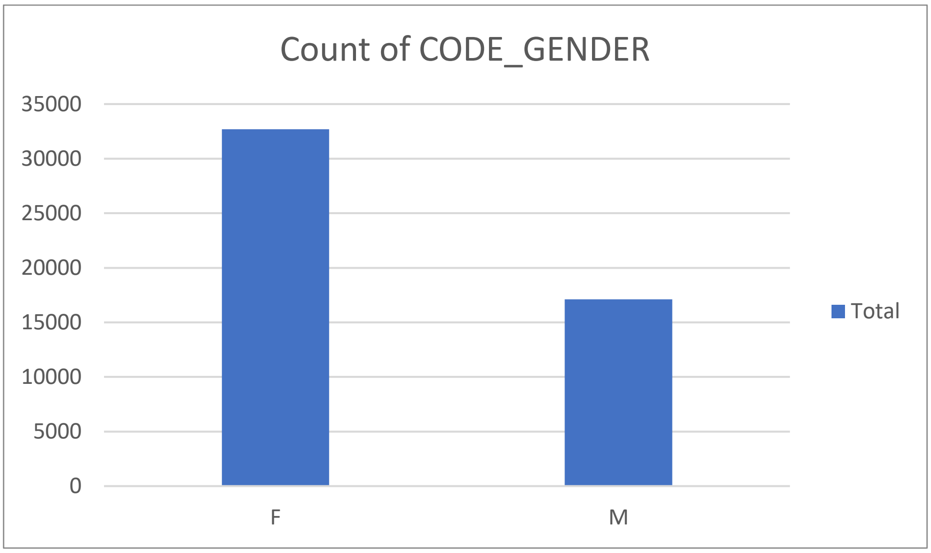
- But outliers in CNT_FAM_MEMBERS are kept as it is because the number of family members is an important feature for the bank.
- As for the outliers in AMT_GOODS_PRICE were dropped from the dataset.
- Since the outliers in other columns mentioned earlier are in continuous format and not away from the rest of the data hence I have kept them as it is.

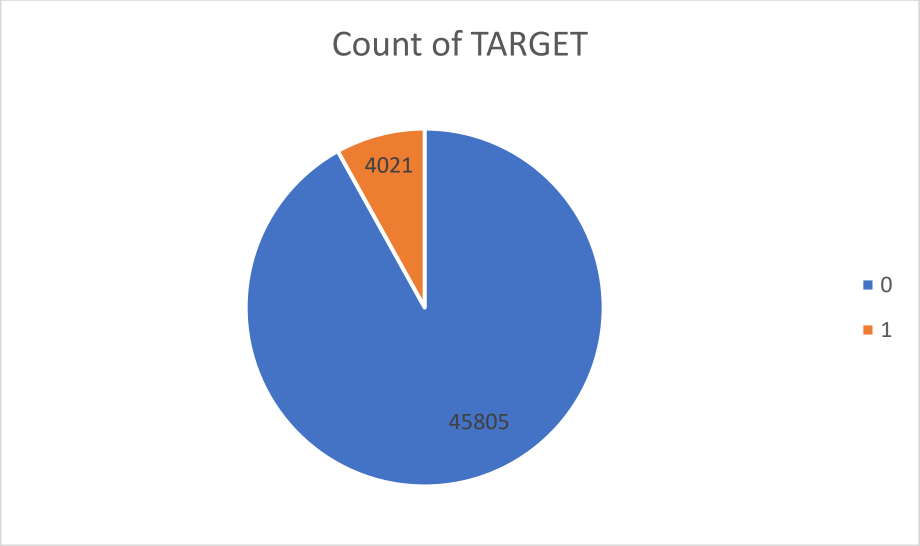
3) Data Imbalance

- Data imbalance can affect the accuracy of the analysis, especially for binary classification problems. Understanding the data distribution is crucial for building reliable models.
- NAME_CONTRACT_TYPE has huge imbalance in data. Cash loans has a total of 45000 count and Revolving loans has around 5000 count.
- Most customers prefer taking cash loans instead of Revolving loans.



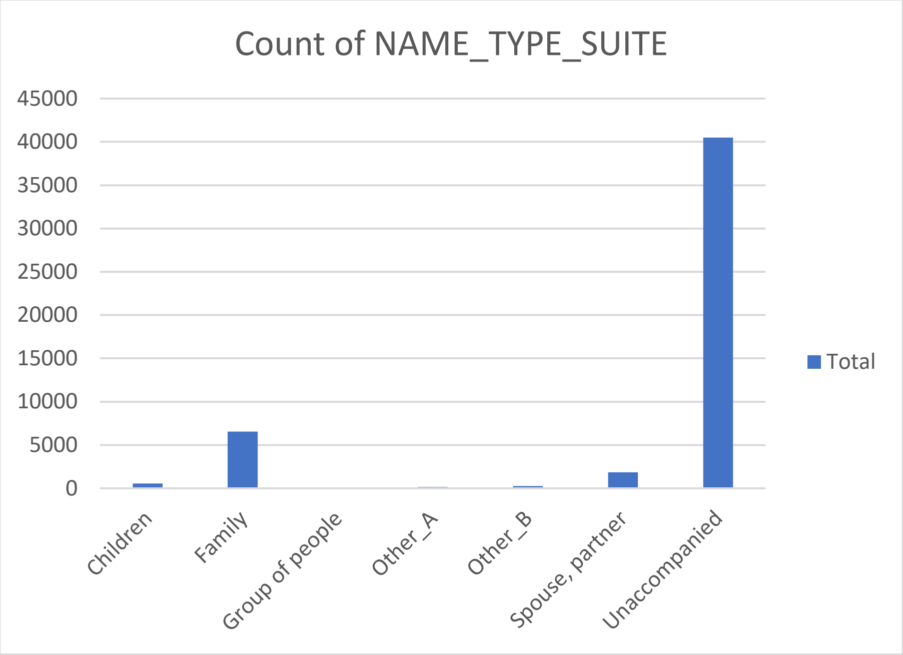
- CODE_GENDER also has data imbalance where count of females taking loans is almost twice the count of males from the bank.
- In Target column most of the records recorded are 0, where 1 - client with payment difficulties: he/she had late payment more than X days on at least one of the first Y instalments of the loan in our sample and 0 - all other cases.



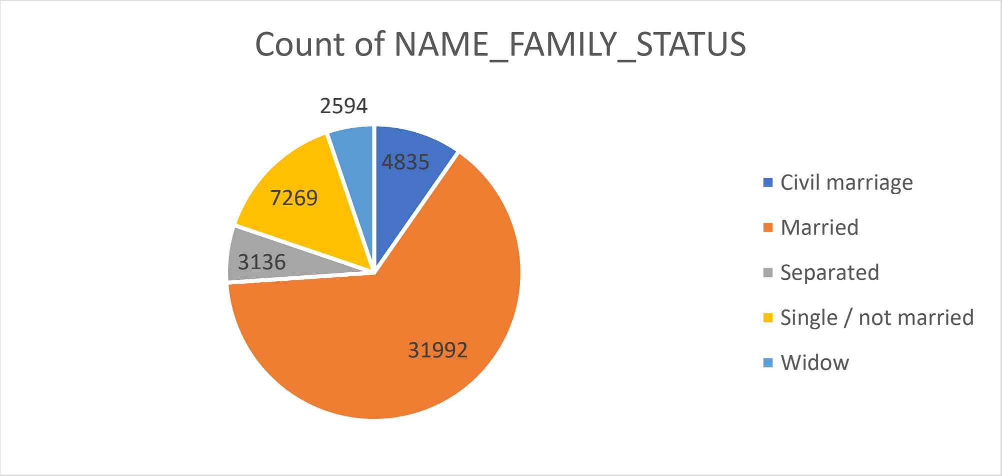


4) Univariate Analysis

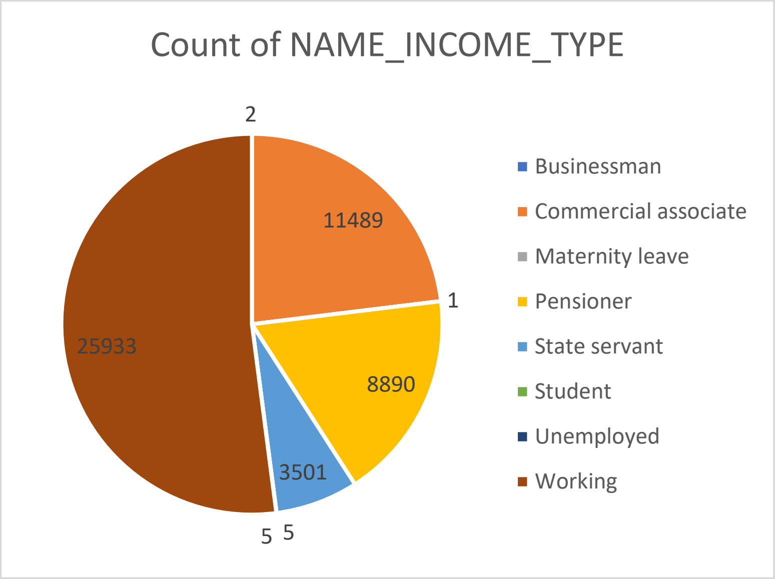
- Many customers came accompanied when applying for the loan.
- Some customers came with family and very few customers came with their partner.
- Most of the customers prefer coming alone to take the loan.



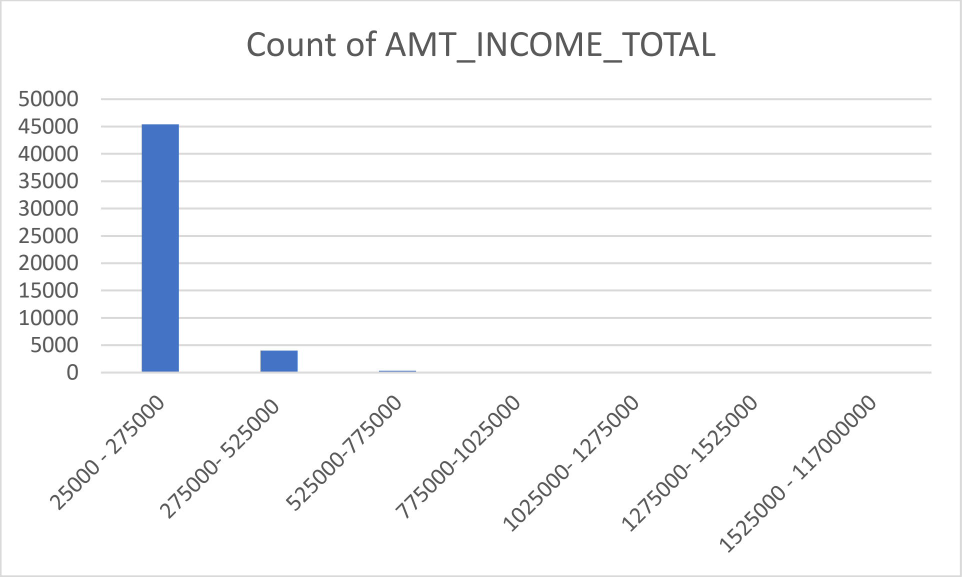
- From the pie chart below we infer that the customers who apply for the loan are married.
- Followed by single/not married customers with count of 7269.
- Count of widows applying for the loan is very less.



- The customers who apply for loan are working professionals.
- Commercial associates also apply for loan followed by Pensioner.
- Very few students and unemployed people apply for loan.



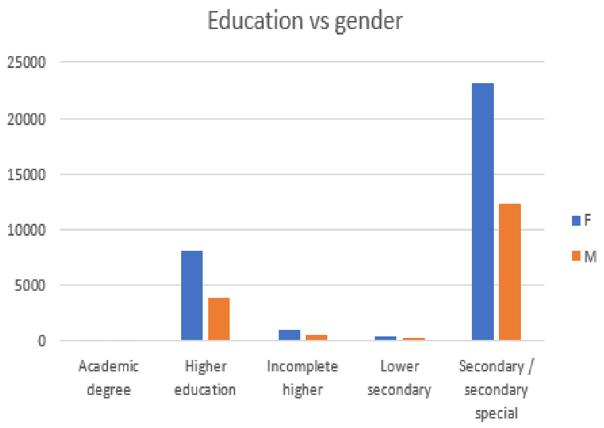
- Most of the customers of the bank have income between the range 25000-275000.
- Around 5000 customers have income range between 275000-525000.

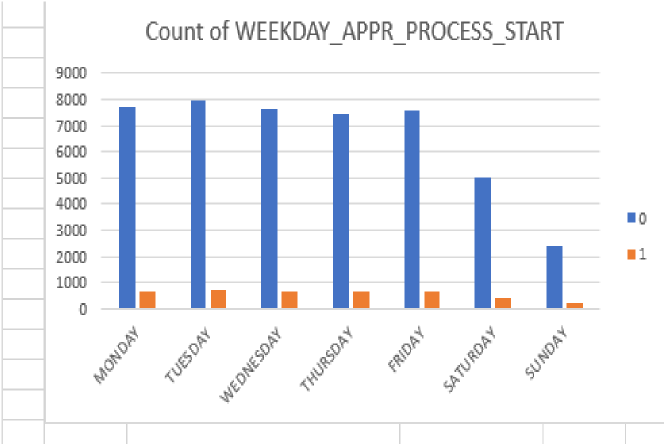


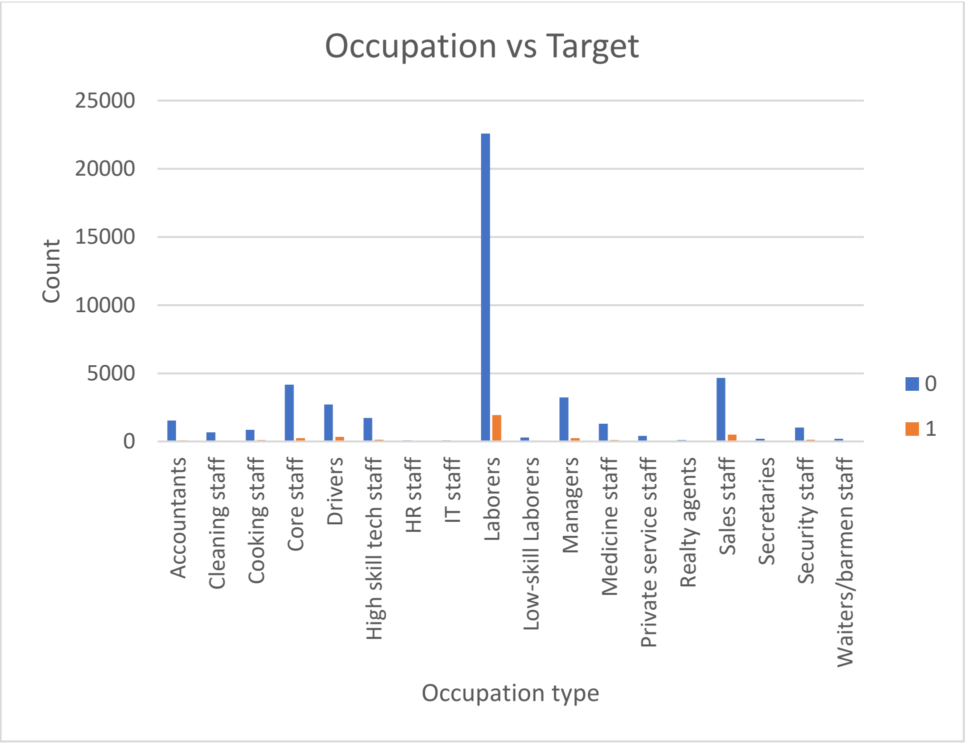
5) Bivariate Analysis

- Females having secondary / secondary special education apply for loans more.
- Females and males with higher education also initiate taking loans.

- Sunday has very less customers having payment difficulties.
- Tuesday has highest count of 0 as target variable indicates that the payment was made on time followed by Friday.



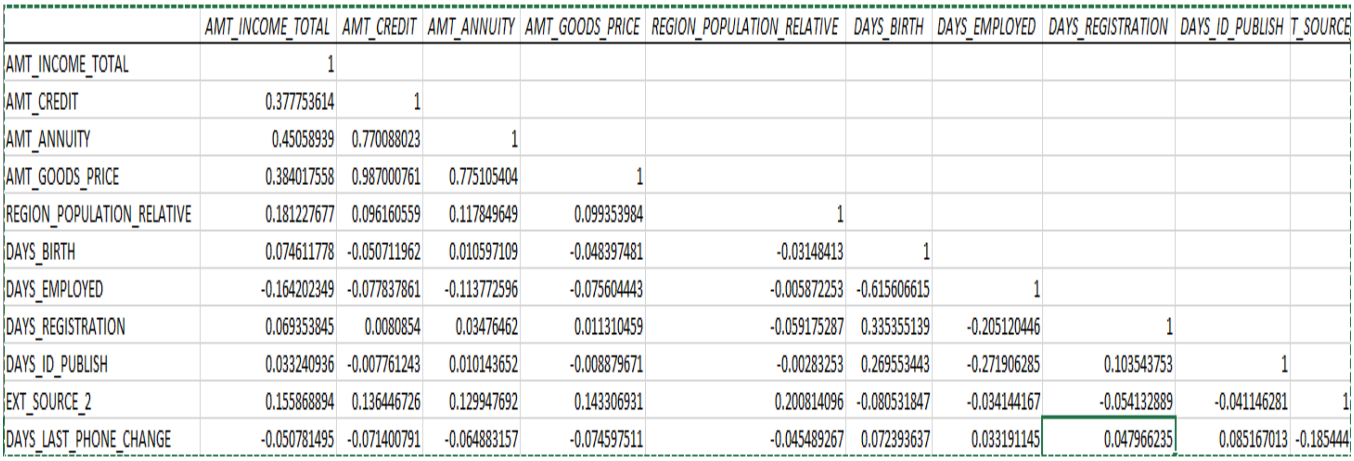




- Highest number of customers are labourers who pay the loan back without difficulties and on time.
- Core staff and Sales staff also repay the loan on time.

- Married customers repay the loan in time and are also the ones who take high number of loans.

6) Correlation



	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_ID_PUBLISH	T_SOURCE
AMT_INCOME_TOTAL	1									
AMT_CREDIT	0.377753614	1								
AMT_ANNUITY	0.45058939	0.770088023	1							
AMT_GOODS_PRICE	0.384017558	0.987000761	0.775105404	1						
REGION_POPULATION_RELATIVE	0.181227677	0.096160559	0.117849649	0.099353984	1					
DAYS_BIRTH	0.074611778	-0.050711962	0.010597109	-0.048397481	-0.03148413	1				
DAYS_EMPLOYED	-0.164202349	-0.077837861	-0.113772596	-0.075604443	-0.005872253	-0.615606615	1			
DAYS_REGISTRATION	0.069353845	0.0080854	0.03476462	0.011310459	-0.059175287	0.335355139	-0.205120446	1		
DAYS_ID_PUBLISH	0.033240936	-0.007761243	0.010143652	-0.008879671	-0.00283253	0.269553443	-0.271906285	0.103543753	1	
EXT_SOURCE_2	0.155868894	0.136446726	0.129947692	0.143306931	0.200814096	-0.080531847	-0.034144167	-0.054132889	-0.041146281	1
DAYS_LAST_PHONE_CHANGE	-0.050781495	-0.071400791	-0.064883157	-0.074597511	-0.045489267	0.072393637	0.033191145	0.047966235	0.085167013	-0.185444

- GOODS_PRICE and AMT_CREDIT have very high positive correlation the value being 0.987.
- AMT_CREDIT and AMT_ANNUITY also have high positive correlation as 0.77.
- DAYS_EMPLOYED and REGION have very low negative correlation.
- DAYS_EMPLOYED and DAYS_BIRTH has high negative correlation.

Results

- Common area average, common area mode and nonliving apartments median have around 89% missing values which means more than half of the data is missing.
- AMT_CREDIT, AMT_ANNUITY, AND CNT_FAM_MEMEBERS are observed to have outliers. The outliers were dealt with as outliers can significantly impact the analysis and distort the result
- NAME_CONTRACT_TYPE, CODE_GENDER and TARGET are the columns which have data imbalance.
- Many customers came accompanied when applying for the loan.
- Most loyal customer who apply for the loan are married and they have high count of repaying the loan in time.

- Hence the married customers can be given loan benefits for being loyal customers.
- Very few students and unemployed people apply for loan, to attract more of these customers the bank should make new policies which will benefit them as well and enable the bank in getting more customers.
- Most of the customers of the bank have income between the range 25000-275000.
- Females having secondary / secondary special education apply for loans more.
- Sunday has very less customers having payment difficulties.
- Tuesday has highest count of 0 as target variable indicates that the payment was made on time followed by Friday.
- Highest number of customers are labourers who pay the loan back without difficulties and on time.
- Highest number of customers are labourers who pay the loan back without difficulties and on time.
- GOODS_PRICE and AMT_CREDIT have very high positive correlation the value being 0.987 and DAYS_EMPLOYED and REGION have very low negative correlation.

The excel sheets is attached here LINK