Bank loan case study

Project description

This primary objective of this project is to analyse a dataset containing details of loan applicants to identify patterns and factors influencing loan default. The dataset includes two types of scenarios.

1. customers with payment difficulties: these are customers who experienced a late payment of more than x days on at least one of the first y instalments of the loan
2. All other cases: customers who made payment on time

In this project I utilised Explanatory Data Analysis (EDA) to clean the dataset and analyse patterns to ensure that eligible applicants are not unjustly rejected. The key analytical task performed as a part of the EDA include:

- Identifying and appropriately addressing missing data
- Identifying outliers in the dataset
- Analysing data imbalance in the dataset
- Performing univariate, segment univariate and bivariate analysis to explore relationship and patterns
- Identifying top correlations for different scenario

These tasks will clean the data, identify patterns in loan defaults, and ensure accurate insights for informed loan approval, ultimately reducing financial risk

Approach

Firstly, I imported application dataset into excel and conducted Explanatory Data Analytics to identify patterns and factors that affecting loan default. I used COUNTBLANK function to determine the total number of missing entries. Columns with more than 40% missing data were deleted, while those with less than 40% missing data were filled with median and categorical columns were filled with the most frequent value. I created a bar chart to visualise the proportion of missing values, distinguish between columns with more or less than 40% missing data. Outliers were detected using the Inter Quartile Range (IQR) method, and I highlighted these outliers in each column using conditional formatting. I used pivot table to show imbalance of target variable. Further, I Performed univariate analysis, univariate segment analysis and bivariate analysis using pivot table and bar chart to gain deeper insights. Finally, I calculated top correlation for defaulters and non-defaulters using the CORREL function, providing a comprehensive understanding of the factors influencing loan defaults.

Tech-stack used

I used Microsoft excel 2021 for this project which is part of Microsoft office home and student 2021 suite. it was particularly useful for identifying missing values, creating pivot table for analysis, applying conditional formatting to highlight outliers, and creating bar charts for visualisation.

Project insights

Below are the analytical task done in this project

1.identify missing data and deal with it appropriately

Determine the missing values in the dataset

Explanation

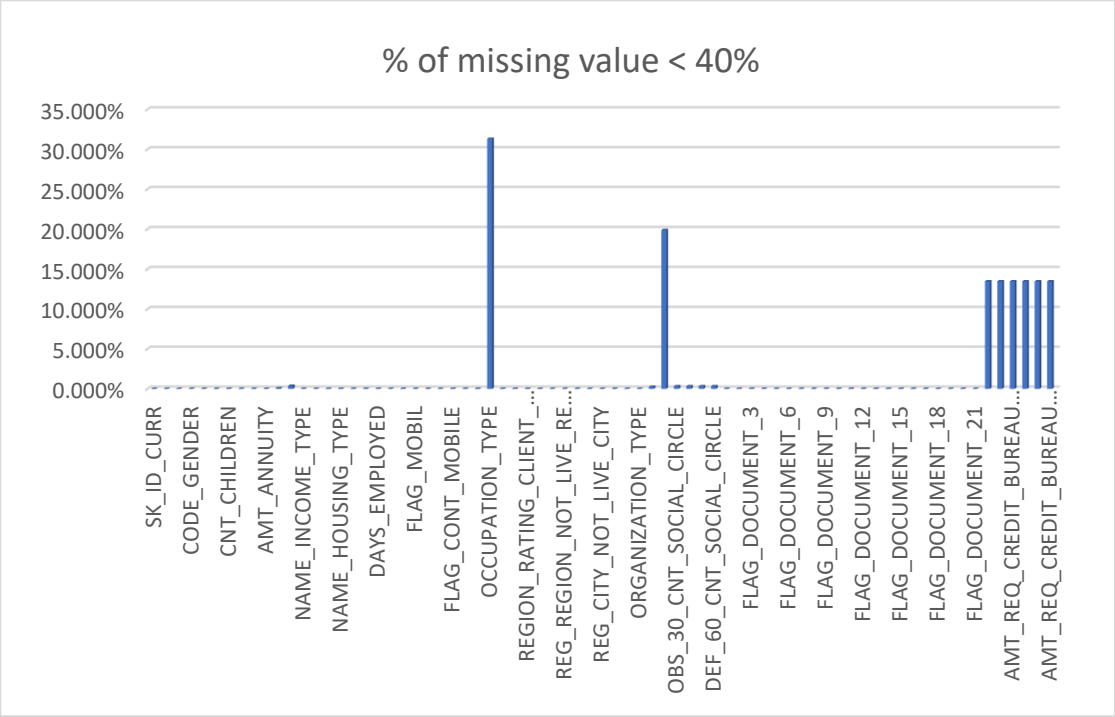
For this task first I divided the dataset into 2 types. That is

a) Percentage of missing value< 40%

% of missing value < 40 %	
column	% of missing value
SK_ID_CURR	0.000%
TARGET	0.000%
NAME_CONTRACT_TYPE	0.000%
CODE_GENDER	0.000%
FLAG_OWN_CAR	0.000%
FLAG_OWN_REALTY	0.000%
CNT_CHILDREN	0.000%
AMT_INCOME_TOTAL	0.000%
AMT_CREDIT	0.000%
AMT_ANNUITY	0.002%
AMT_GOODS_PRICE	0.076%
NAME_TYPE_SUITE	0.384%
NAME_INCOME_TYPE	0.000%
NAME_EDUCATION_TYPE	0.000%
NAME_FAMILY_STATUS	0.000%
NAME_HOUSING_TYPE	0.000%
REGION_POPULATION_RELATIVE	0.000%
DAYS_BIRTH	0.000%
DAYS_EMPLOYED	0.000%
DAYS_REGISTRATION	0.000%
DAYS_ID_PUBLISH	0.000%
FLAG_MOBIL	0.000%
FLAG_EMP_PHONE	0.000%
FLAG_WORK_PHONE	0.000%
FLAG_CONT_MOBILE	0.000%
FLAG_PHONE	0.000%
FLAG_EMAIL	0.000%
OCCUPATION_TYPE	31.309%
CNT_FAM_MEMBERS	0.002%
REGION_RATING_CLIENT	0.000%
REGION_RATING_CLIENT_W_CITY	0.000%
WEEKDAY_APPR_PROCESS_START	0.000%
HOUR_APPR_PROCESS_START	0.000%
REG_REGION_NOT_LIVE_REGION	0.000%
REG_REGION_NOT_WORK_REGION	0.000%
LIVE_REGION_NOT_WORK_REGION	0.000%
REG_CITY_NOT_LIVE_CITY	0.000%
REG_CITY_NOT_WORK_CITY	0.000%
LIVE_CITY_NOT_WORK_CITY	0.000%
ORGANIZATION_TYPE	0.000%
EXT_SOURCE_2	0.252%
EXT_SOURCE_3	19.888%
OBS_30_CNT_SOCIAL_CIRCLE	0.336%
DEF_30_CNT_SOCIAL_CIRCLE	0.336%
OBS_60_CNT_SOCIAL_CIRCLE	0.336%
DEF_60_CNT_SOCIAL_CIRCLE	0.336%
DAYS_LAST_PHONE_CHANGE	0.002%
FLAG_DOCUMENT_2	0.000%
FLAG_DOCUMENT_3	0.000%
FLAG_DOCUMENT_4	0.000%
FLAG_DOCUMENT_5	0.000%

FLAG_DOCUMENT_6	0.000%
FLAG_DOCUMENT_7	0.000%
FLAG_DOCUMENT_8	0.000%
FLAG_DOCUMENT_9	0.000%
FLAG_DOCUMENT_10	0.000%
FLAG_DOCUMENT_11	0.000%
FLAG_DOCUMENT_12	0.000%
FLAG_DOCUMENT_13	0.000%
FLAG_DOCUMENT_14	0.000%
FLAG_DOCUMENT_15	0.000%
FLAG_DOCUMENT_16	0.000%
FLAG_DOCUMENT_17	0.000%
FLAG_DOCUMENT_18	0.000%
FLAG_DOCUMENT_19	0.000%
FLAG_DOCUMENT_20	0.000%
FLAG_DOCUMENT_21	0.000%
AMT_REQ_CREDIT_BUREAU_HOUR	13.468%
AMT_REQ_CREDIT_BUREAU_DAY	13.468%
AMT_REQ_CREDIT_BUREAU_WEEK	13.468%
AMT_REQ_CREDIT_BUREAU_MON	13.468%
AMT_REQ_CREDIT_BUREAU_QRT	13.468%
AMT_REQ_CREDIT_BUREAU_YEAR	13.468%

The given above are the list columns whose missing value percentage is less than 40 %.

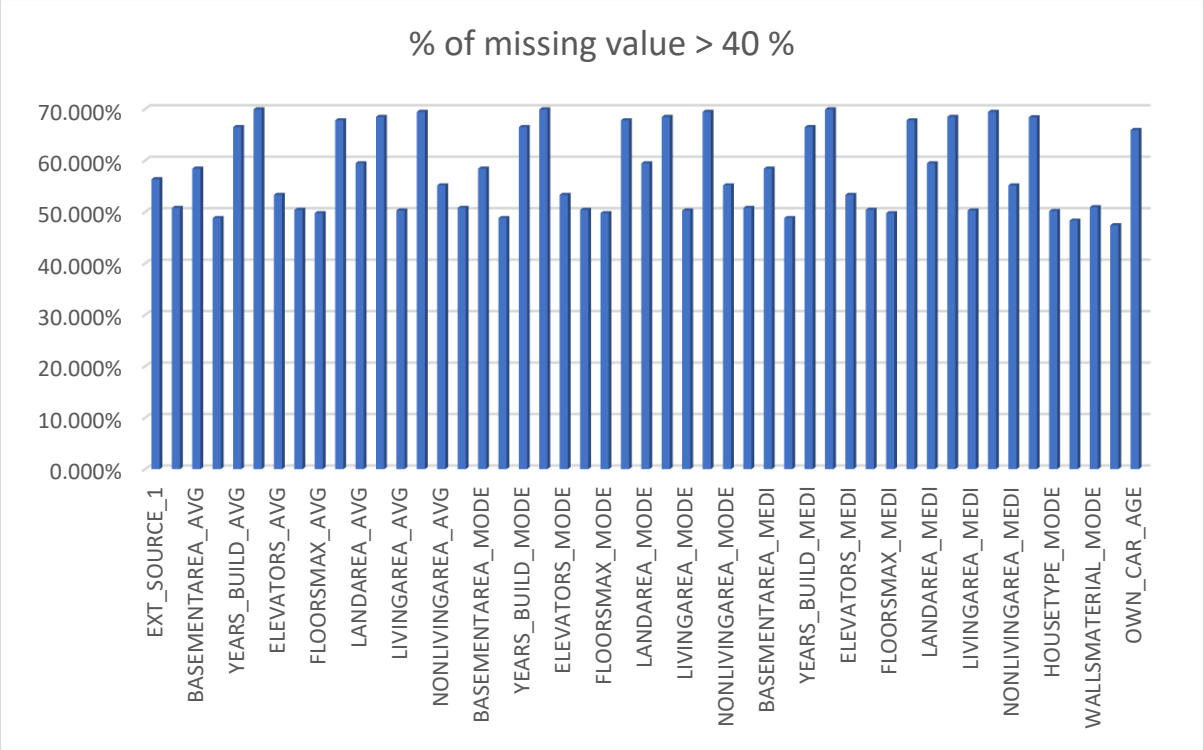


Conclusion

- From this table and bar chart given above, we observed that 55 columns had no missing values, while the remaining 18 columns contained missing values.
- Numerical columns with missing values were filled using their median, while categorical columns were imputed with the most frequently occurring words in each column.
- For instance, missing values in the column NAME_TYPE_SUIT were replaced with ‘unaccompanied’, the most common entry and those in the column OCCUPATION_TYPE were filled with ‘labourers.

b) Percentage missing value above 40%

% of missing value > 40%	
column	% of missing value
EXT_SOURCE_1	56.345%
APARTMENTS_AVG	50.771%
BASEMENTAREA_AVG	58.399%
YEARS_BEGINEXPLUATATION_AVG	48.789%
YEARS_BUILD_AVG	66.479%
COMMONAREA_AVG	69.921%
ELEVATORS_AVG	53.303%
ENTRANCES_AVG	50.391%
FLOORSMAX_AVG	49.751%
FLOORSMIN_AVG	67.789%
LANDAREA_AVG	59.443%
LIVINGAPARTMENTS_AVG	68.453%
LIVINGAREA_AVG	50.275%
NONLIVINGAPARTMENTS_AVG	69.429%
NONLIVINGAREA_AVG	55.145%
APARTMENTS_MODE	50.771%
BASEMENTAREA_MODE	58.399%
YEARS_BEGINEXPLUATATION_MODE	48.789%
YEARS_BUILD_MODE	66.479%
COMMONAREA_MODE	69.921%
ELEVATORS_MODE	53.303%
ENTRANCES_MODE	50.391%
FLOORSMAX_MODE	49.751%
FLOORSMIN_MODE	67.789%
LANDAREA_MODE	59.443%
LIVINGAPARTMENTS_MODE	68.453%
LIVINGAREA_MODE	50.275%
NONLIVINGAPARTMENTS_MODE	69.429%
NONLIVINGAREA_MODE	55.145%
APARTMENTS_MEDI	50.771%
BASEMENTAREA_MEDI	58.399%
YEARS_BEGINEXPLUATATION_MEDI	48.789%
YEARS_BUILD_MEDI	66.479%
COMMONAREA_MEDI	69.921%
ELEVATORS_MEDI	53.303%
ENTRANCES_MEDI	50.391%
FLOORSMAX_MEDI	49.751%
FLOORSMIN_MEDI	67.789%
LANDAREA_MEDI	59.443%
LIVINGAPARTMENTS_MEDI	68.453%
LIVINGAREA_MEDI	50.275%
NONLIVINGAPARTMENTS_MEDI	69.429%
NONLIVINGAREA_MEDI	55.145%
FONDKAPREMONT_MODE	68.383%
HOUSETYPE_MODE	50.151%
TOTALAREA_MODE	48.297%
WALLSMATERIAL_MODE	50.919%
EMERGENCYSTATE_MODE	47.397%
OWN_CAR_AGE	65.901%

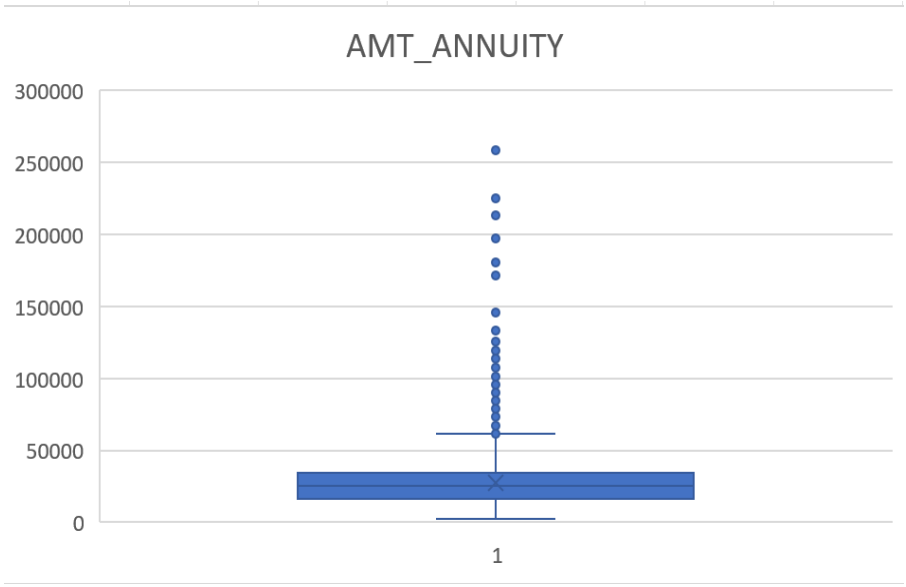


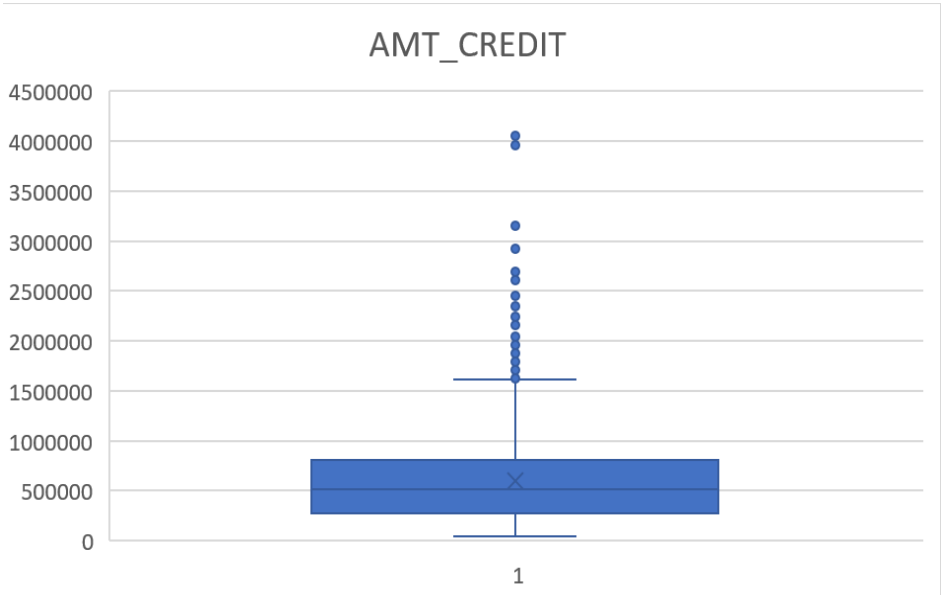
Conclusion

There are 49 columns with more than 40% missing values, and I deleted all of them

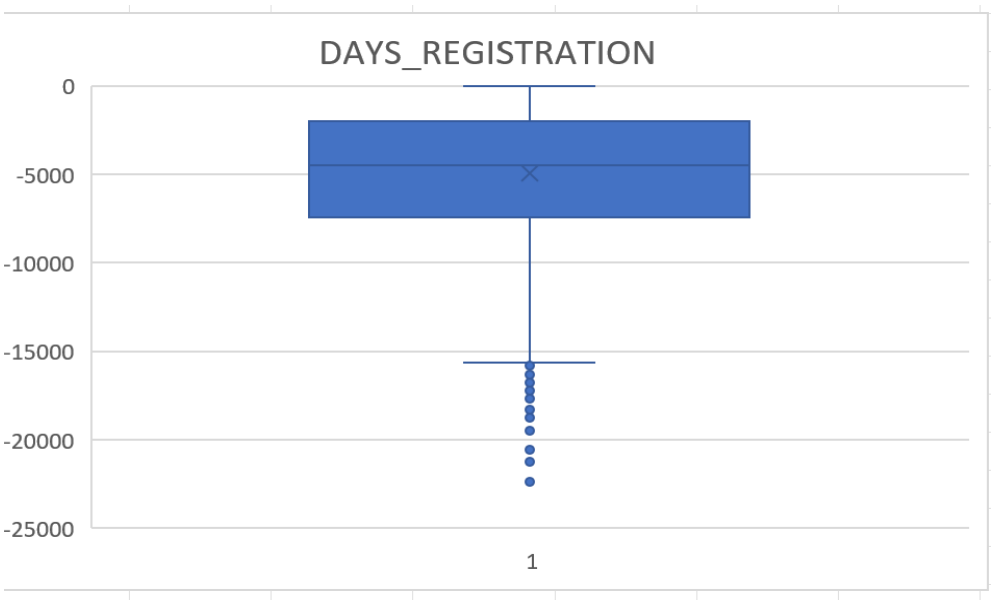
2 Identify outliers in the dataset

In the dataset 18 columns contains outliers. The sample of some boxplot visualisation of columns are given below





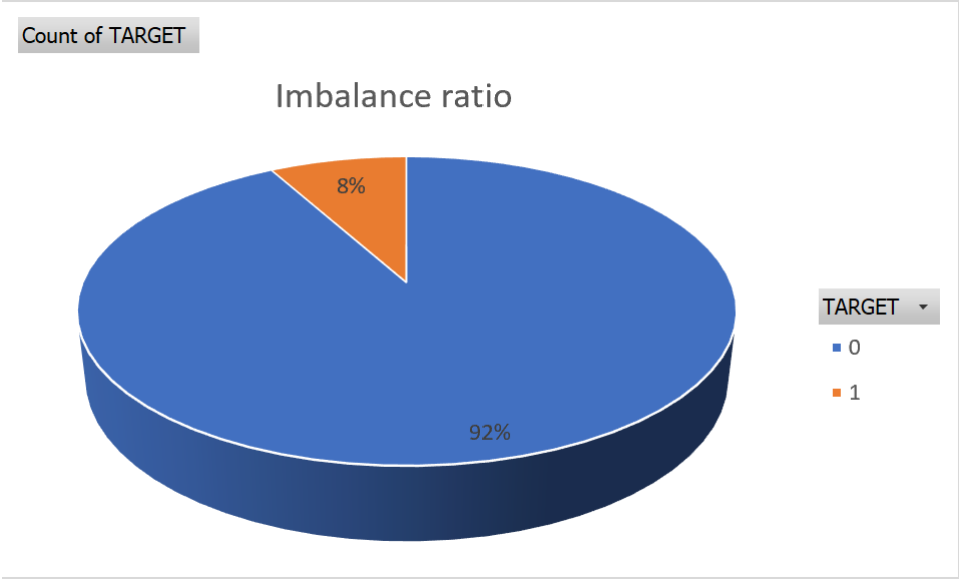

AMT_GOODS_PRICE
4500000
4000000
3500000
3000000
2500000
2000000
1500000
1000000
500000
0
1


DAYS_REGISTRATION
0
-5000
-10000
-15000
-20000
-25000
1

3.Analyse Data imbalance

Determine the data imbalance in the target column and the ratio of data imbalance using excel functions

Result



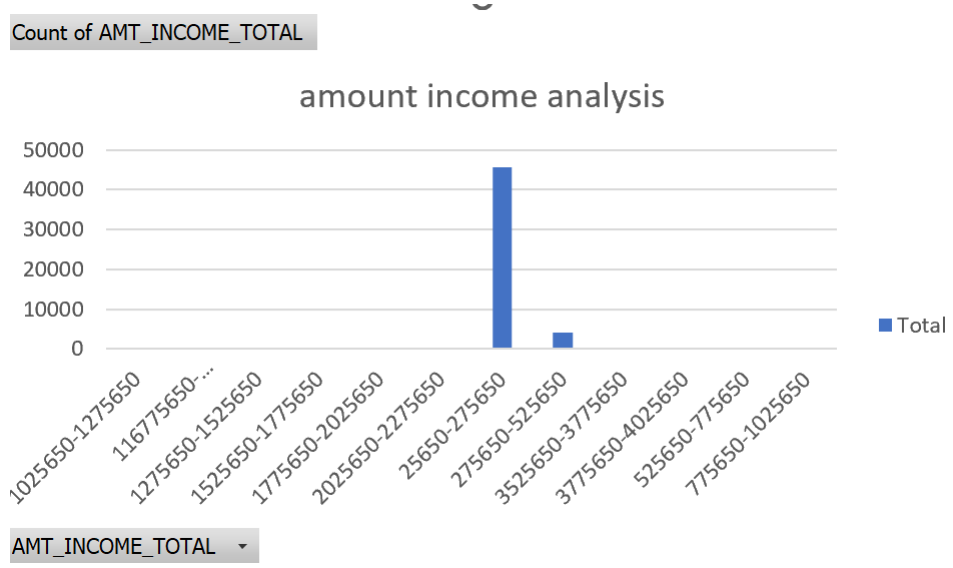
The dataset is highly imbalanced, with 92% of the target variable representing loan non-defaulters (target 0) and only 8% representing loan non-defaulters (target 1). This imbalance makes accurately predicting target 1 more challenging due to its minority representation.

4.Perform univariate analysis, Segmented univariate and Bivariate analysis

Performe univariate analysis to understand the distribution of individual variables, segmented univariate analysis to compare variable distributions for different scenarios, and bivariate analysis to explore the relationship between variables and target variables.

Univariate analysis

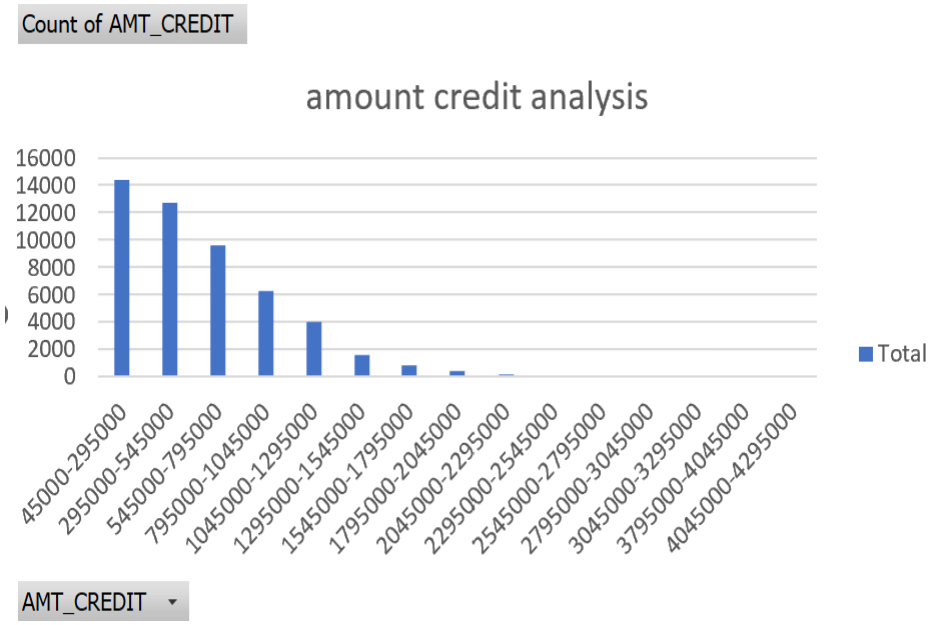
AMT_INCOME_TOTAL Analysis



The majority of applicants have an income in the range of

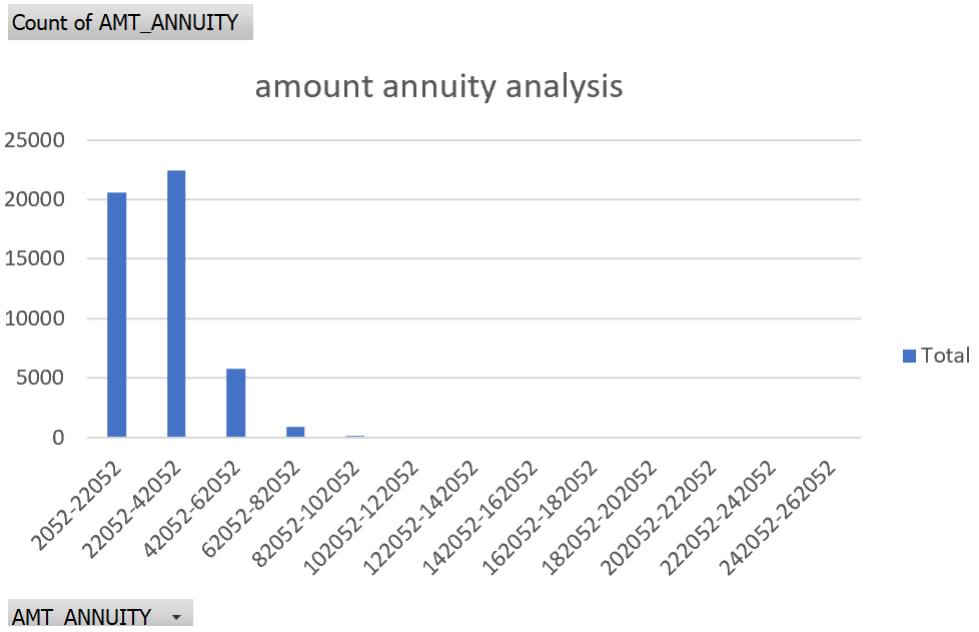
25650 –275650 rupees. The median of income is 130500 rupees and the average income is 130500 rupees.

AMT_CREDIT Analysis



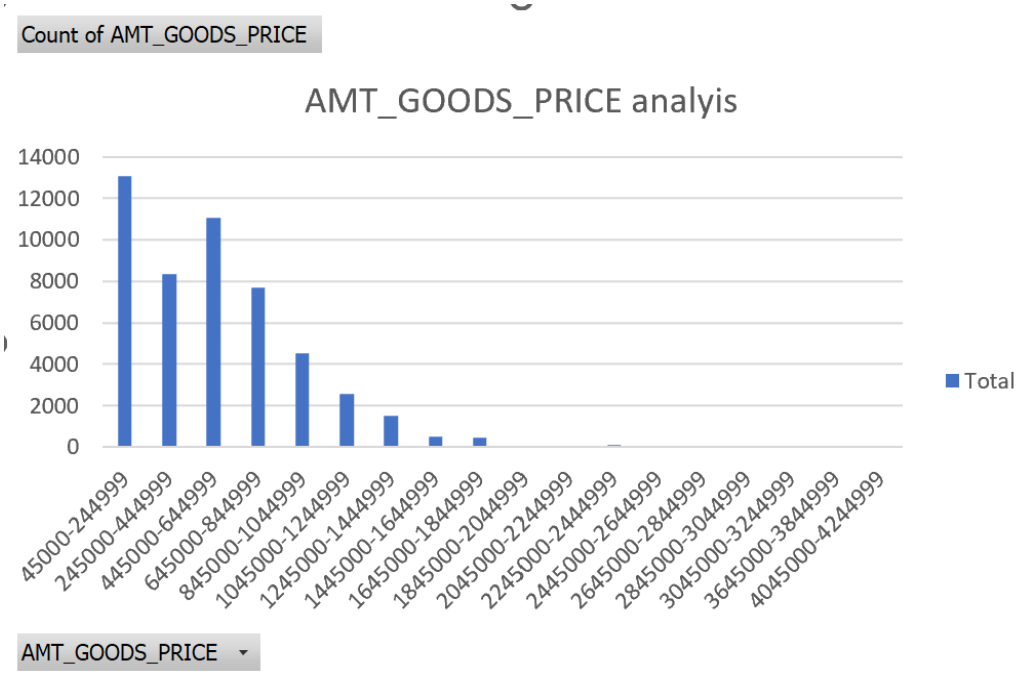
The majority of applicants have credit amount of the loan in the range of 45000-295000 rupees. The average credit amount is 521020.4 rupees and the median value is 514777.5 rupees

AMT_ANNUITY Analysis



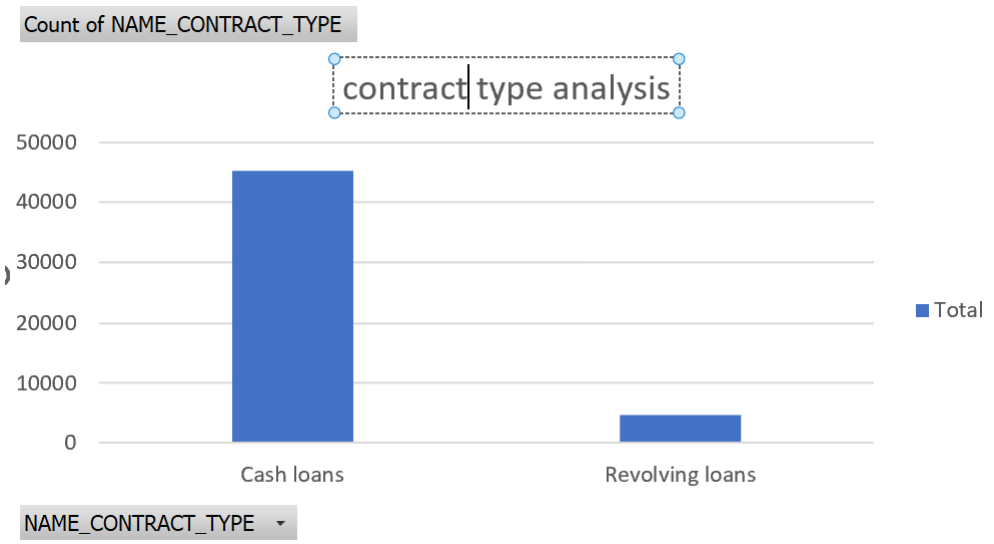
The majority of applicants have a loan annuity in the range of 22052-42052 rupees. The average of loan annuity is 27107.33 rupees and the median is 24939 rupees.

AMT_GOODS_PRICE Analysis



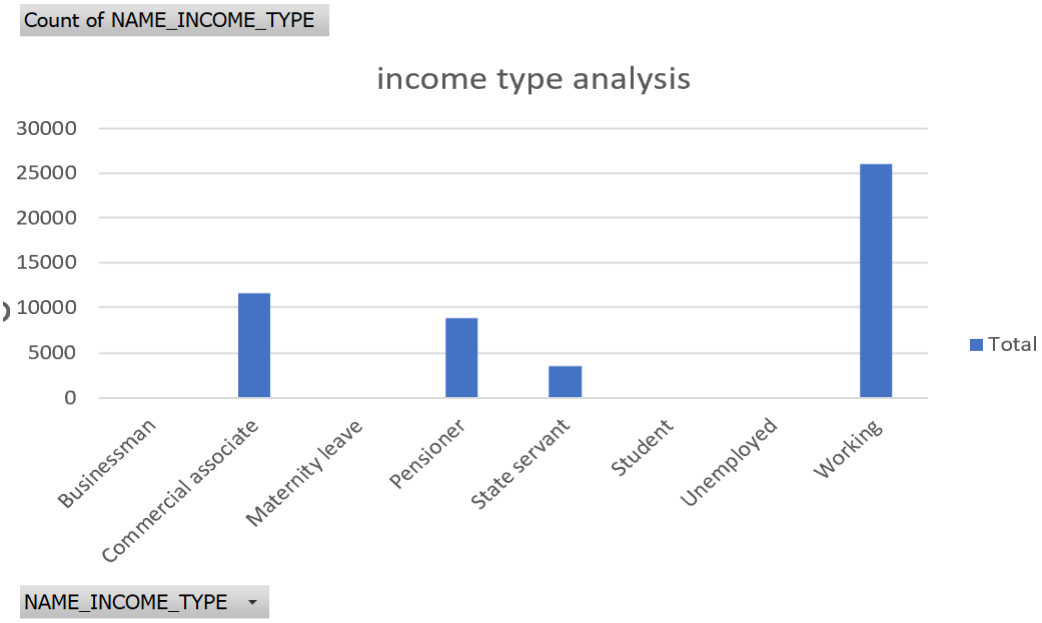
The majority of applicants have goods price in the range of 45000-244999 rupees. The average of goods price is 538992.3 rupees and the median is 450000 rupees.

NAME_CONTRACT_TYPE Analysis



Majority of the applicants are using cash loans.

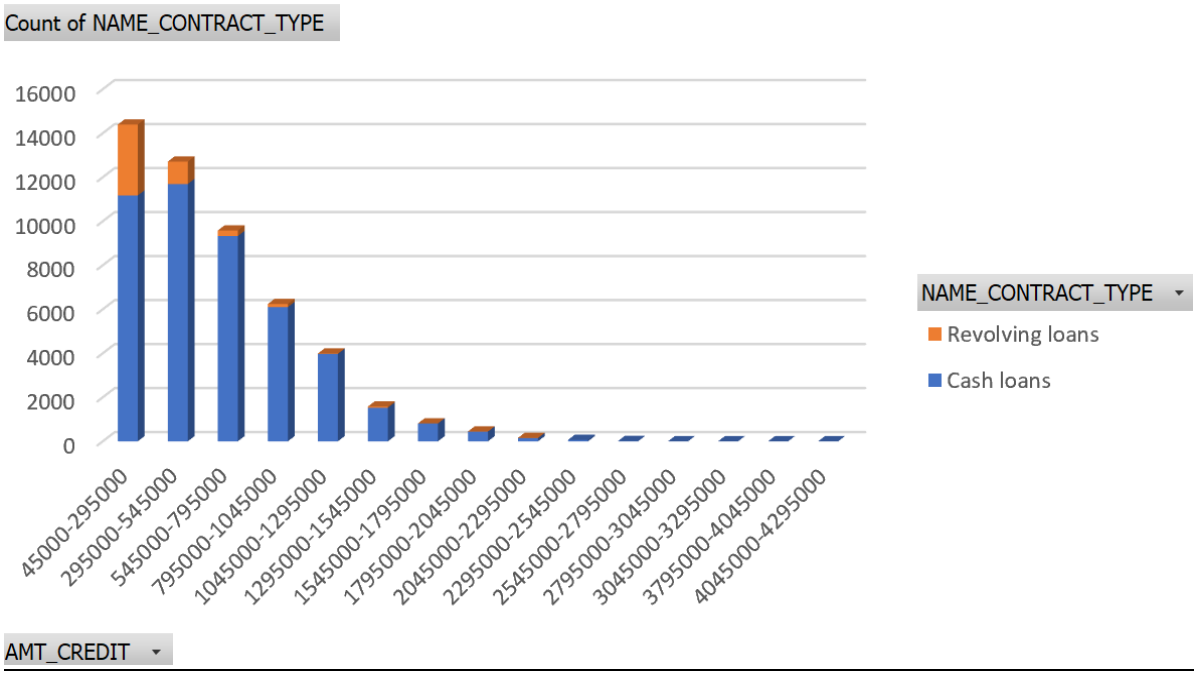
NAME_INCOME_TYPE Analysis



Most of the applicants are working

Univariate segment analysis

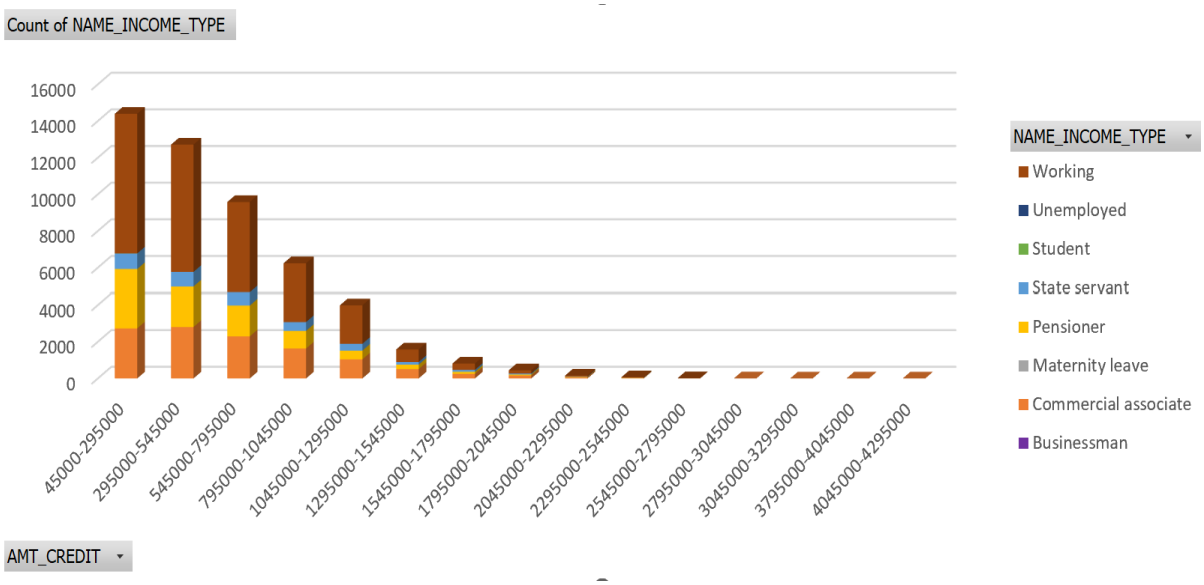
Contract type VS Amount credit analysis



By analysing this chart, we get that

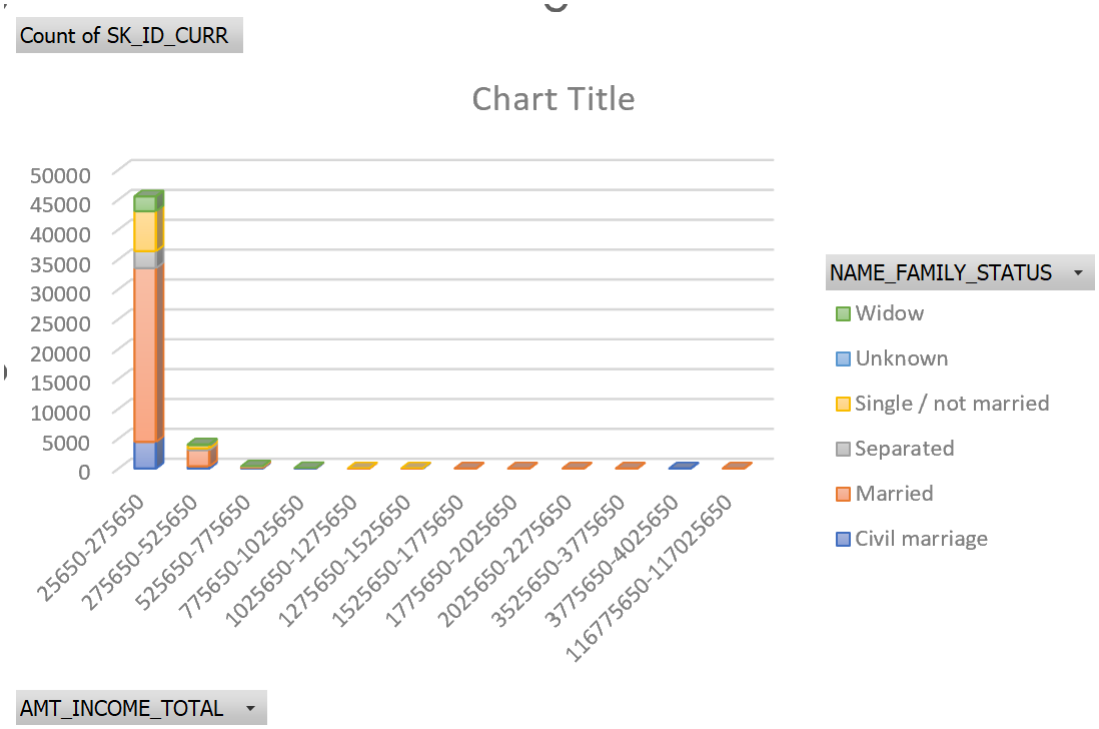
- The largest number of loans fall within the credit rage of 45000 – 545000 rupees.
- The majority of loans in all credit ranges are cash loans.
- Revolving loans are far fewer compared to cash loans,especially as the credit amount increases.

Income type VS Amount credit analysis



- The majority of credit amount falls within the range 45000-295000 rupees.
- As the credit amount increases the count of each income type decreases
- Higher credit amounts are less common and linked with fewer people in all income groups.

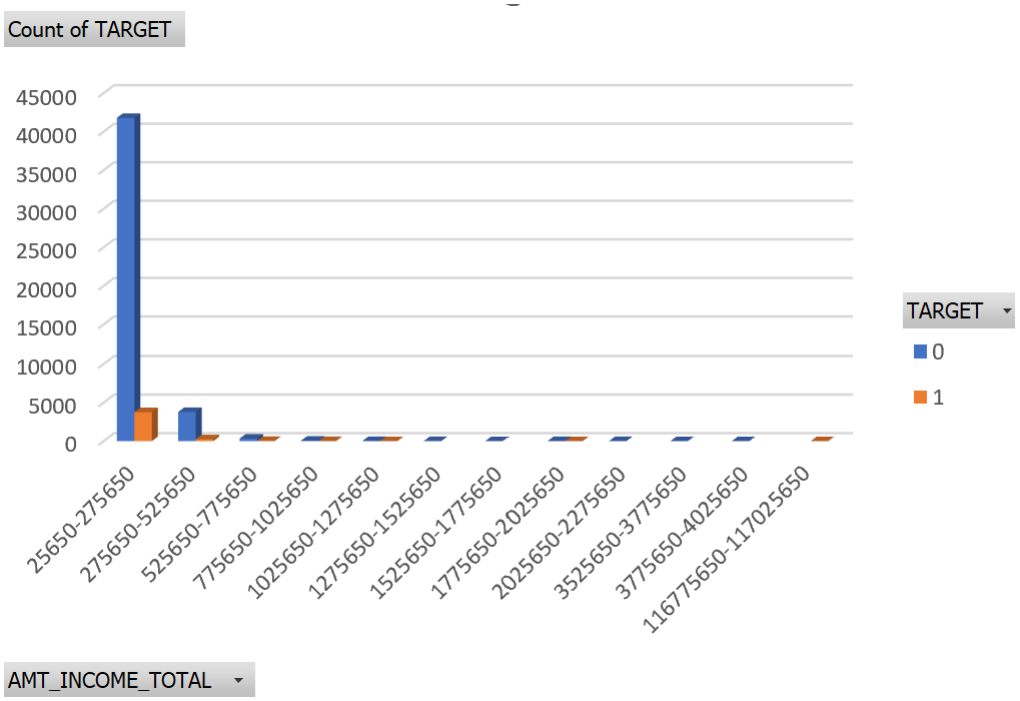
Income amount vs family status



- Most applicants are married especially in the income range 25650 – 275650
- Single applicants are the next largest group, mainly in lower income ranges.
- Higher incomes are rare across all groups, with married applicants being the majority at every level.

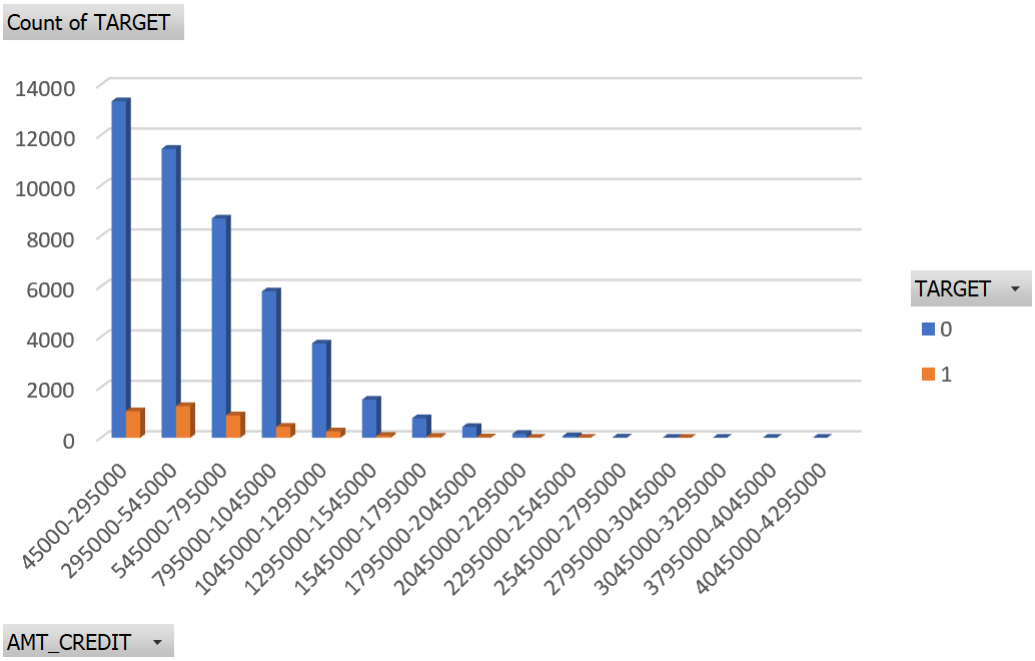
Bivariate Analysis

Income VS target



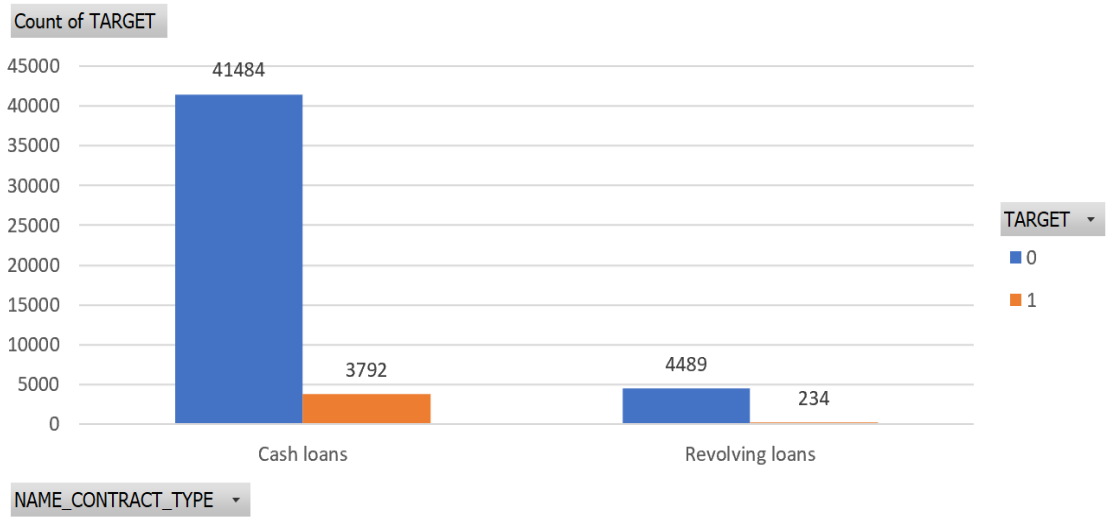
- Both Target 0(non- defaulters) and target 1(defaulters) are high in the range of 25650-275650 rupees
- Non-defaulters are exceeding defaulters across all income ranges
- As income range increases both defaulters and non-defaulters decrease.

Credit amount VS Target



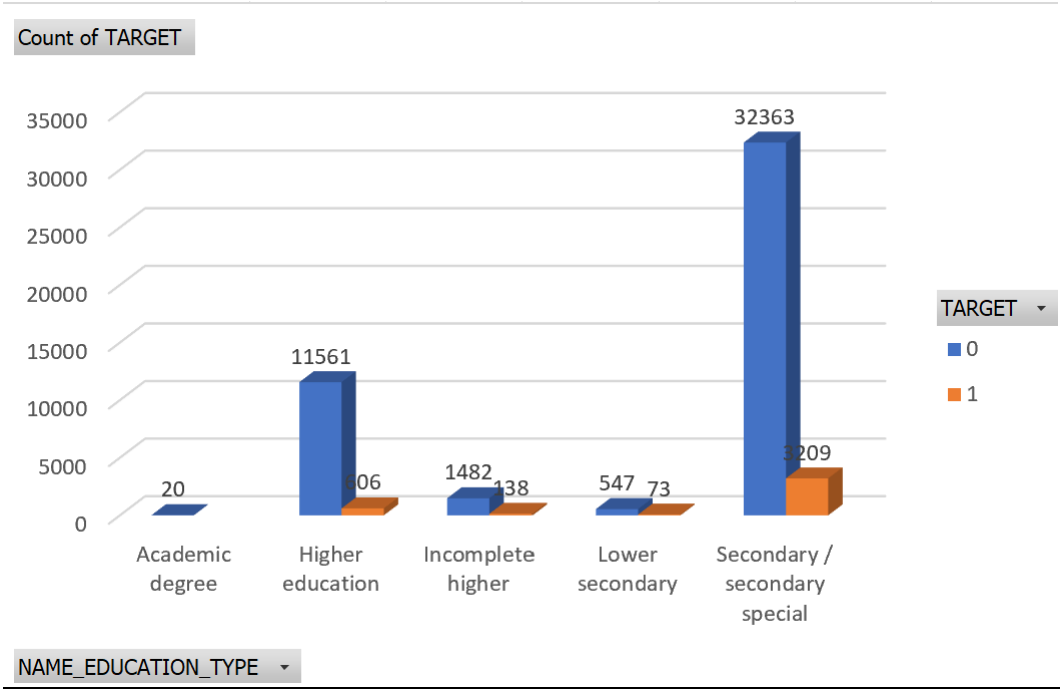
- Both target 0(non-defaulters) and target 1(defaulters) are high in the range of 45000-295000 rupees
- Non-defaulters exceed defaulters across all credit ranges.
- As credit amount increase both defaulters and non-defaulters decrease.

Contract type vs Target



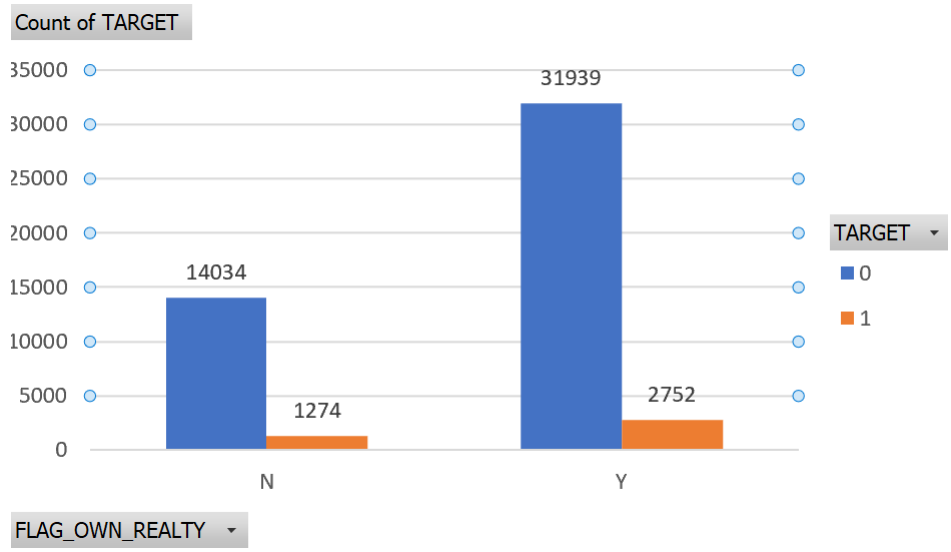
- Most loans are cash loans, with non-defaulters(target 0) being significantly higher compared to defaulters
- Revolving loans are much less common, with non-defaulters still being higher than defaulters

Education type vs target



- Most applicants belong to the secondary/secondary special education type, with non-defaulters being much higher than defaulters.
- Applicants with higher education are next largest group with non-defaulters being much higher than defaulters
- Loan defaulters are not present in Academic degree
- Other education types like Incomplete higher and lower secondary have fewer applicants, with non-defaulters consistently being higher than defaulters across all groups

Flag_own_realty VS Target



- Most applicants own a house or flat with non-defaulters being much higher than defaulters

- Applicants who do not own a house or property are fewer, with non-defaulters still higher than defaulters

5. Identify top correlation for different scenario

Identify top correlation for loan defaulters (clients with payment difficulties) and loan non- defaulters (other cases)

Result

Top 10 correlation of Non-defaulters

Top Correlation of Non-defaulters		
column1	column2	correlation
OBS_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	0.998332596
AMT_CREDIT	AMT_GOODS_PRICE	0.986703851
REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	0.950710198
CNT_CHILDREN	CNT_FAM_MEMBERS	0.880458195
REG_REGION_NOT_WORK_REGION	LIVE_REGION_NOT_WORK_REGION	0.857141422
DEF_30_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE	0.85620785
REG_CITY_NOT_WORK_CITY	LIVE_CITY_NOT_WORK_CITY	0.82156989
AMT_ANNUITY	AMT_GOODS_PRICE	0.774125332
AMT_CREDIT	AMT_ANNUITY	0.769489445
DAYS_BIRTH	FLAG_EMP_PHONE	0.617713731

Top 10 correlation of Defaulters

Top 10 correlation of defaulters		
column1	column2	correlation
CNT_CHILDREN	CNT_FAM_MEMBERS	0.880443487
AMT_CREDIT	AMT_ANNUITY	0.769465786
AMT_CREDIT	AMT_GOODS_PRICE	0.986704632
AMT_ANNUITY	AMT_GOODS_PRICE	0.774107154
DAYS_BIRTH	FLAG_EMP_PHONE	0.617796751
REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	0.950702583
REG_REGION_NOT_WORK_REGION	LIVE_REGION_NOT_WORK_REGION	0.857140022
REG_CITY_NOT_WORK_CITY	LIVE_CITY_NOT_WORK_CITY	0.821536039
OBS_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	0.998332304
DEF_30_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE	0.856259333

Conclusion

Through this project I gained deeper understanding of the importance of Explanatory Data Analytics (EDA) in data analysis. I learned how EDA helps to identify patterns and factors that affect loan defaults. I also developed skills in handling missing data, identifying outliers and addressing data imbalances. By performing univariate analysis, univariate segment analysis and bivariate analysis, I was able to explore the dataset thoroughly. Additionally, I identified top correlations between various features and loan default, which provide valuable insights into the behaviour of loan defaulters and non-defaulters

Click here to open excel sheets