

**Project Description:** This case study aims to give you an idea of applying EDA in a real business scenario. In this case study, apart from applying the techniques that you have learnt in the EDA module, you will also develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimize the risk of losing money while lending to customers.

**Tech stack used:** I have used google colab to run my code and display the charts.

## Solutions:

Here I am attaching the link of code file

[https://drive.google.com/file/d/1JJtNsQ7Lhew-5iGMYJzLTzmlkpPQ3FSI/view?usp=share\\_link](https://drive.google.com/file/d/1JJtNsQ7Lhew-5iGMYJzLTzmlkpPQ3FSI/view?usp=share_link)

Can open the file in google colab in drive.

**Task-1:** Present the overall approach of the analysis. Mention the problem statement and the analysis approach briefly?

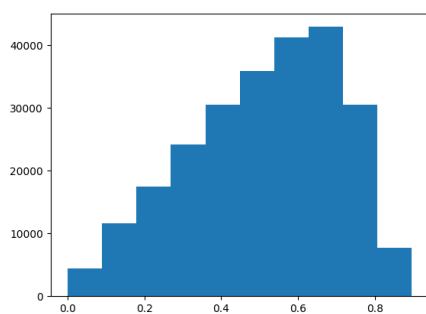
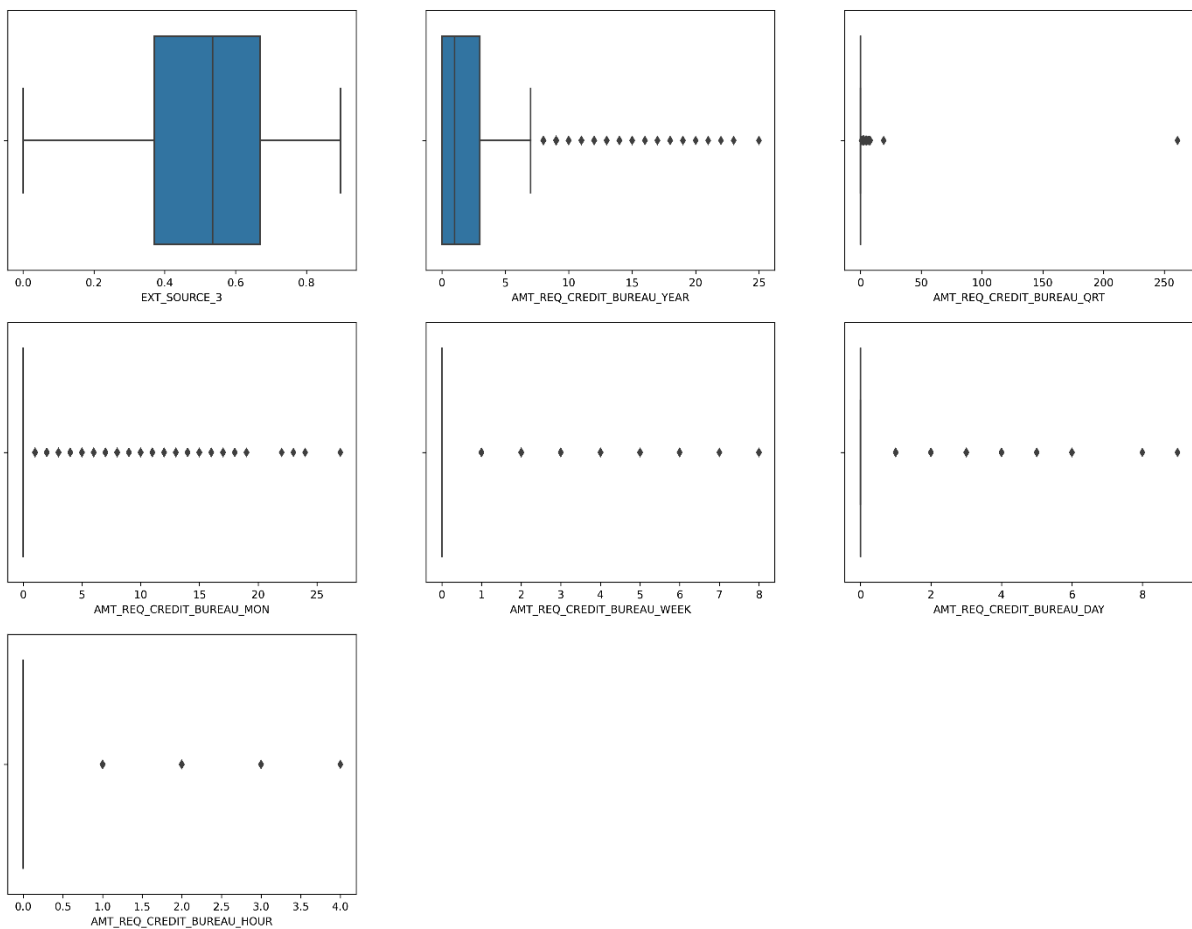
Problem statement: It aims to identify patterns which indicate if a client has difficulty paying their installments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. This will ensure that the consumers capable of repaying the loan are not rejected. Identification of such applicants using EDA is the aim of this case study. In other words, the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.

**Approach:** First step is to get an idea about the structure and summary of the given 2 application dataset and previous application dataset. Second step is handling the null values here we use different thresholds and use different approaches to handle the columns within each threshold. Third step is identifying the outliers here we use statistical techniques as well as visualizations to find the outliers. Fourth step is identifying the data imbalance and finding the ratio of imbalance. Fifth step is we are using various analytical methods and plots to conduct univariate and bi variate analysis and derive some valuable insights from them by identifying the driving variables behind loan default. Final step is to do correlation analysis by finding the top 10 correlated variables in the dataset.

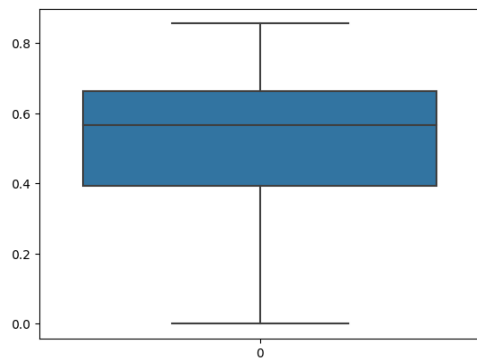
**Task-2:** Identify the missing data and use appropriate method to deal with it. (Remove columns/or replace it with an appropriate value)

**Hint:** Note that in EDA, since it is not necessary to replace the missing value, but if you have to replace the missing value, what should be the approach. Clearly mention the approach.

**Approach:** Check the percentage of null values in each columns. First step is we fix a threshold of 40 percent and remove those columns above 40 percent of null values. We have removed 49 columns using this method. For the remaining columns we categorise our approach based on whether the column is numerical or categorical. For categorical columns if the count of missing values is high add a new category called 'Unknown/Missing' else replace them with mode(most frequent category). For numerical columns we do mean imputation for columns with no outliers and those follows almost a symmetric distribution. If it has outliers and follows non symmetric distribution then do median imputation. We check outliers using boxplots and symmetry of distribution using distribution plot.



Symmetric distribution of EXT\_SOURCE\_3



Box plot of EXT\_SOURCE\_2

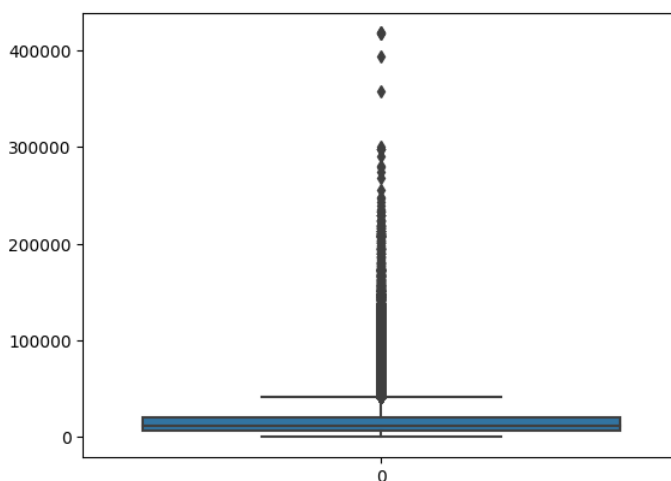
	ORGANIZATION_TYPE	NAME_INCOME_TYPE
0	Business Entity Type 3	Working
1	School	State servant
2	Government	Working
3	Business Entity Type 3	Working
4	Religion	Working
5	Other	State servant
6	Business Entity Type 3	Commercial associate
7	Other	State servant
8	XNA	Pensioner
9	Electricity	Working
10	Medicine	Working
11	XNA	Pensioner
12	Business Entity Type 2	Working
13	Self-employed	Working
14	Transport: type 2	Working

XNA values in ORGANIZATION\_TYPE matches with Pensioner values in NAME\_INCOME\_TYPE

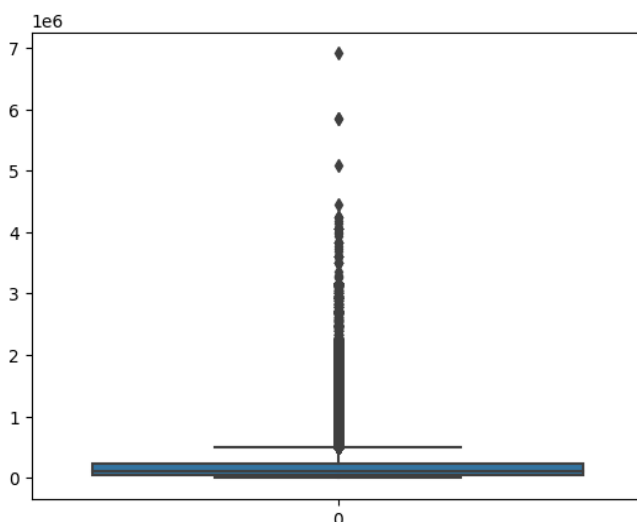
**Insights:** For OCCUPATION\_TYPE column, count of missing values are very huge and many categories are present so its safe to handle null values by adding a category called 'Unknown'. For NAME\_TYPE\_SUITE column we do mode imputation as we can see that the count of missing values is not huge. Now for numerical columns Only EXT\_SOURCE\_3 column has no outliers and almost symmetric distribution so do mean imputation and all other numerical columns will be imputed by median.

We need to handle a column called ORGANISATION\_TYPE. Here we have a category called XNA which is a representation of missing values but the count is very high. From columns description we came to know NAME\_INCOME\_TYPE and ORGANISATION\_TYPE are related columns. So we check whether pattern of missingness exists in this column also. To check whether the missing values in the ORGANISATION\_TYPE column are missing at random or not, we can compare the distribution of NAME\_INCOME\_TYPE between the missing and non-missing values in the ORGANISATION\_TYPE column. If the p-value from the chi-squared test of independence is less than the significance level (usually 0.05), we reject the null hypothesis and conclude that there is evidence of an association between the two variables. If the p-value is greater than the significance level, we fail to reject the null hypothesis and conclude that there is no evidence of an association between the two variables.

From the chisquare test we can confirm both columns are related. After checking the values of both the columns, we can infer that XNA values are present in ORGANIZATION\_TYPE column for pensioner in NAME\_INCOME\_TYPE column. Also check the count of pensioner in NAME\_INCOME\_TYPE. Here almost same counts in XNA values and pensioner counts so we can impute XNA values with pensioner. This is one special case of handling missing data by investigating the values.



Box plot of AMT\_ANNUIITY In Previous application Dataset



Box plot of AMT\_GOODS\_PRICE

**Task-3:** Identify if there are outliers in the dataset. Also, mention why do you think it is an outlier. Again, remember that for this exercise, it is not necessary to remove any data points.

### Approach:

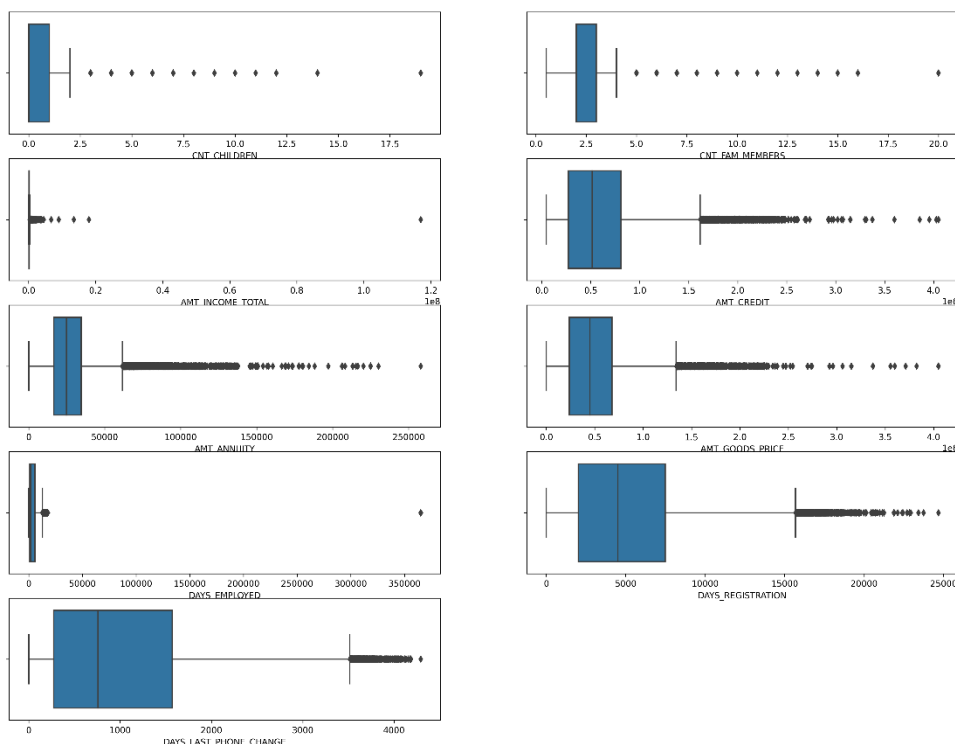
Using describe we will get columns that have huge difference between max and 75 percent. Also the columns having too much extreme positive values in max. To justify the identified points are outliers we just used iqr rule to confirm the outliers. Any observations that are more than 1.5 IQR below Q1 or more than 1.5 IQR above Q3 are considered outliers. From this we got the outlier columns and the count of outliers.

```
{'CNT_CHILDREN': 4272, 'CNT_FAM_MEMBERS': 4007, 'AMT_INCOME_TOTAL': 14035, 'AMT_CREDIT': 6562, 'AMT_ANNUITY': 7504, 'AMT_GOODS_PRICE': 14728, 'DAYS_EMPLOYED': 56357, 'DAYS_REGISTRATION': 659, 'DAYS_LAST_PHONE_CHANGE': 435}
```

```
{'AMT_ANNUITY': 157960, 'AMT_APPLICATION': 201864, 'AMT_CREDIT': 174913, 'AMT_GOODS_PRICE': 229147, 'DAYS_DECISION': 16106, 'CNT_PAYMENT': 335065}
```

Then use those columns to check outliers. Then using a box plot we identify the presence of outliers for those selected columns.

### Application\_dataset



### Insights:

1. For AMT\_INCOME\_TOTAL, the majority of the data is concentrated in a narrow range (IQR), while there are many outliers present. This indicates that a small number of customers have extremely high or low income, which can potentially impact their ability to repay the loan.

2. AMT\_CREDIT has a wide range of values, with the majority of the loan amounts falling in the third quartile. This suggests that most customers are taking out larger loans, and there are many outliers present indicating some customers are taking out very large loans.

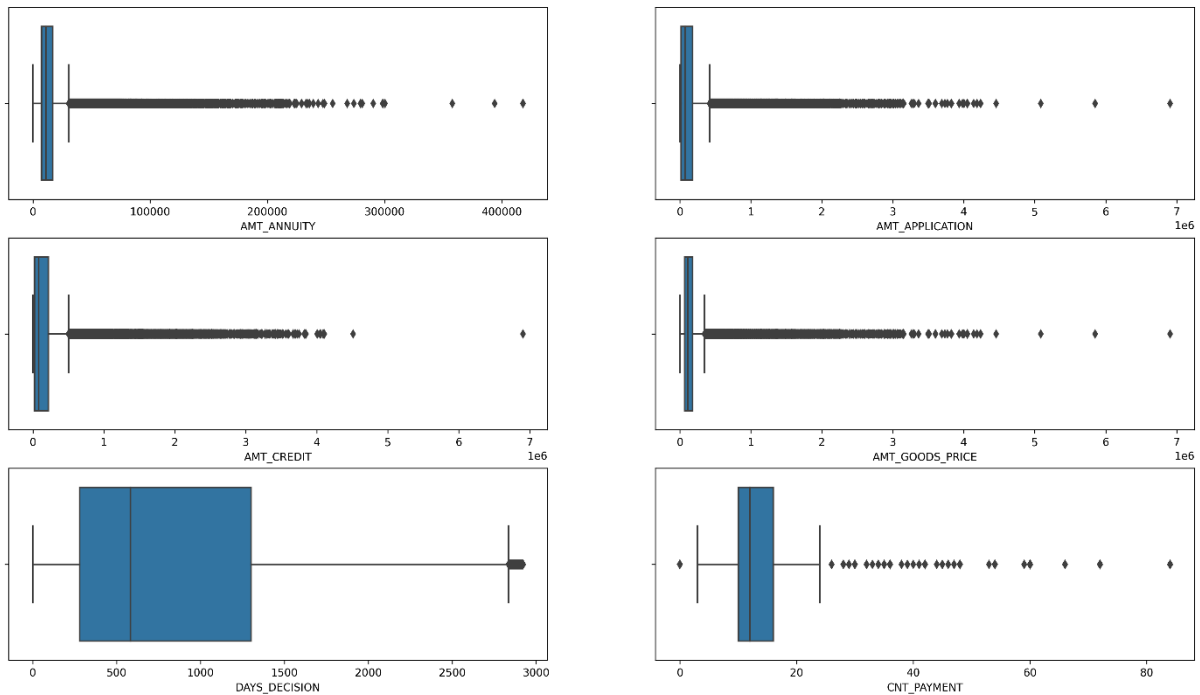
3. The AMT\_ANNUITY box plot also shows a wide range of values, with the majority of the loan amounts falling in the third quartile, and a large number of outliers indicating some customers have high annuities.

4. The box plots for AMT\_GOODS\_PRICE, DAYS\_REGISTRATION, and DAYS\_LAST\_PHONE\_CHANGE all show a wide range of values with the majority of data in the third quartile, indicating that most customers are buying expensive goods, registering and changing phones after a certain time, and there are many outliers present indicating that some customers are buying exceptionally expensive goods, registering and changing phones after a very long time.

5. The box plot for DAYS\_EMPLOYED shows a slim IQR, with most outliers present below 25000, which could indicate that most customers have stable employment history, while an outlier at 375000 suggests a possible data entry error.

6. CNT\_FAM\_MEMBERS has a narrow IQR, and most clients have 4 family members. Some outliers indicate that a few customers have significantly more or fewer family members, which could potentially impact their ability to repay the loan.

Previous Application dataset.:



1.CNT\_PAYMENT has few outlier values: This suggests that most of the values for this variable are clustered closely together, with only a few observations that fall outside this range.

2.DAYS\_DECISION has little number of outliers: This implies that the majority of the observations for this variable are relatively close to the median value, with only a few observations that lie far away from the bulk of the data. This could suggest that most of the previous applications were decided relatively recently, with only a few that were processed a long time ago.

3.AMT\_ANNUIITY, AMT\_APPLICATION, AMT\_CREDIT, AMT\_GOODS\_PRICE have significant large number of outliers with narrow iqr range. These four variables are all related to previous credit applications made by the client. So significant outliers in these variables could indicate several things: The client may have applied for a credit amount that is significantly higher or lower than their usual borrowing pattern, possibly due to a change in their financial situation or a change in their borrowing needs. The client may have been approved for a credit amount that is significantly higher or lower than their initial application, which could be due to a variety of factors such as their credit score, income, or the lender's policies. The client may have applied for a credit amount that is significantly higher or lower than the value of the goods they intended to purchase, indicating a possible discrepancy in their application or intended use of the credit.

**Task-4:** Identify if there is data imbalance in the data. Find the ratio of data imbalance.

**Hint:** Since there are a lot of columns, you can run your analysis in loops for the appropriate columns and find the insights

**Approach:** Data imbalance occurs when the classes in a classification problem have significantly different numbers of instances. In other words, one class may have a much higher or lower number of examples compared to the other classes. To identify data imbalance, you can calculate the distribution of the target classes in your dataset. You can do this by counting the number of instances for each class and plotting the distribution. Use `value_counts` to count the number of instances of each class and plot it using `count plot` of `seaborn` library. Since majority is `target0` (non default) and minority is `Target1` we get the imbalance ratio by dividing `len(Target0)/len(Target1)`. We have also calculated gender imbalance ratio to find any imbalance in Gender.

```
1 class_distribution
```

```
0      282686
```

```
1       24825
```

```
Name: TARGET, dtype: int64
```

```
No. of defaulters: 282686
```

```
No. of non-defaulters: 24825
```

```
Percentage of defaulters: 8.072881945686495
```

```
Percentage of non-defaulters: 91.92711805431351
```

---

Imbalance Ratio: 11.39

The ratio of gender imbalance is 1.93

## Insights:

1. Imbalanced Data: The data set is imbalanced, with a large majority of clients likely to repay the loan (91%), and a relatively small percentage of clients likely to default on payments (8.1%). This indicates that the data is skewed towards one class, making it difficult to build a reliable predictive model.

2. Sampling techniques: As a result of the data imbalance, it is important to consider sampling techniques such as oversampling or undersampling to address the class imbalance issue. Oversampling involves increasing the number of samples in the minority class, while under sampling involves reducing the number of samples in the majority class. By balancing the data, the predictive model can learn from equal samples of both classes, leading to better results.

3. Real-world scenario: The insight that the number of people defaulting on loans is way less than those repaying loans on time is a reflection of the real-world scenario. This highlights the importance of identifying the few clients who are likely to default early, so that



appropriate measures can be taken to mitigate the risk of non-payment. This can be achieved through effective predictive modelling and risk assessment techniques.

4.The dataset is imbalanced in terms of gender with gender imbalance ratio of 1.93. The number of female applicants are almost double that of the male applicants.

**Task-5:** Explain the **results of univariate, segmented univariate, bivariate analysis, etc.** in business terms.

**Approach:** First since AMT\_INCOME\_TOTAL,AMT\_CREDIT are continuous data, we will categorize them into bins for better analysis and plotting.

**Univariate Analysis:** It is the simplest form of analysing data. “Uni” means “one”, so in other words your data has only one variable. It doesn’t deal with causes or relationships (unlike regression) and it’s major purpose is to describe; It takes data, summarizes that data and finds patterns in the data. We will be using count plots, Box plots, distribution plots for Univariate analysis.

**Bivariate Analysis:** It is an analysis of two variables to determine the relationships between them. We will be using scatter plots, pivot table, chi-square test etc here. Apart from this, we will be also plotting using multiple variables.

Application\_dataset plots :

count plot: It shows the counts of observations in each categorical bin using bars.

Insights:

1.People who dont have own car are high chances of defaulters. People who dont have own house/flat are high chances of defaulters.

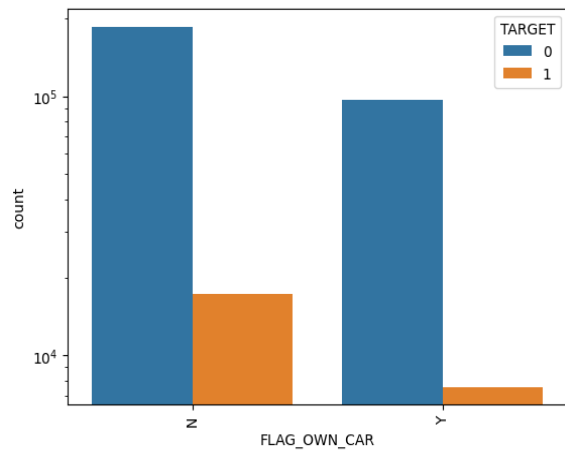
2.People with income range (<100k,100k-200k,200k-300k) are high defaulters. People with credit range (500k and above ,200k-300k) are high defaulters.

3. Education with Secondary/Secondary special customers face difficulty in paying loans compare with other level of educated people. People living in OFFICE APARTMENT, CO-OP APARTMENT face less difficulty in paying loans Where as people living in House/apartment faces most difficulty in paying loans.

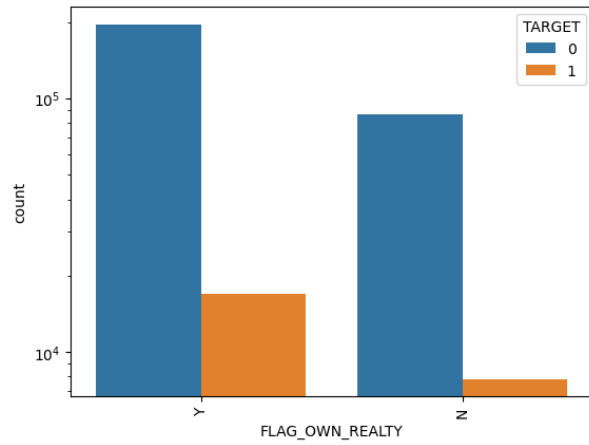
4.Cash loans have difficulty in paying than revolving loans.Clients with income type as Working/Commercial associate/Penisoners have difficulties in Paying Loan.Students,businessman do not have any pending loans.

5.Married people and females have more difficulties in paying loans.

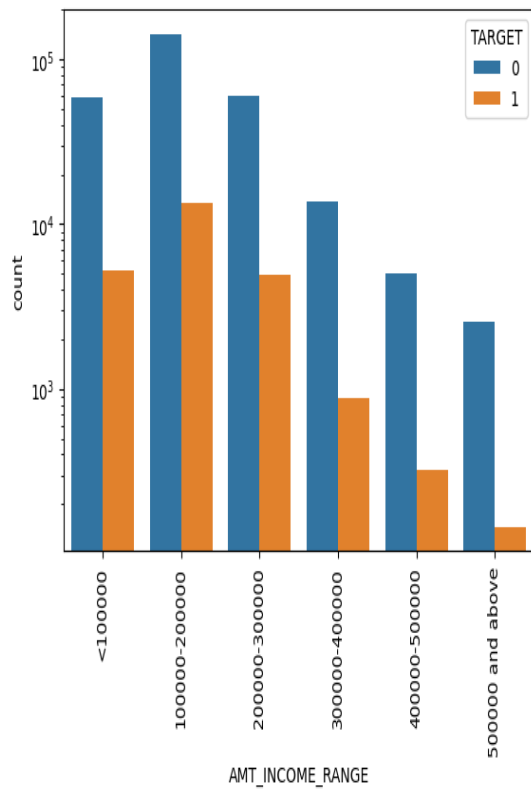
TARGET VS OWN\_CAR\_FLAG



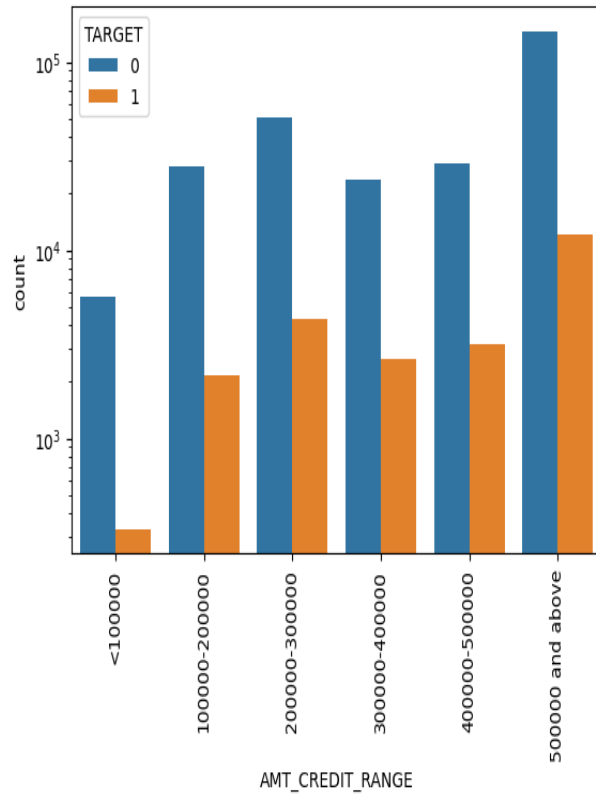
TARGET VS OWN\_HOUSE\_FLAG

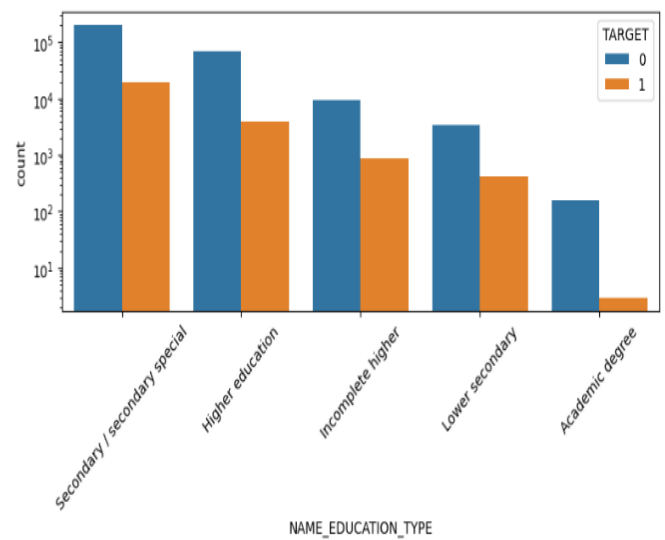
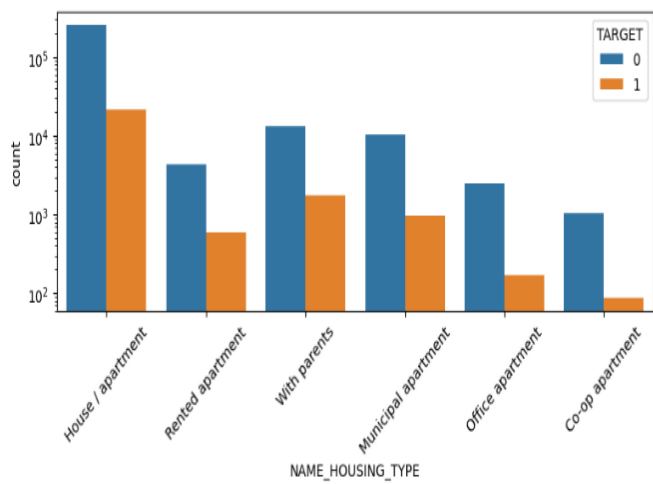
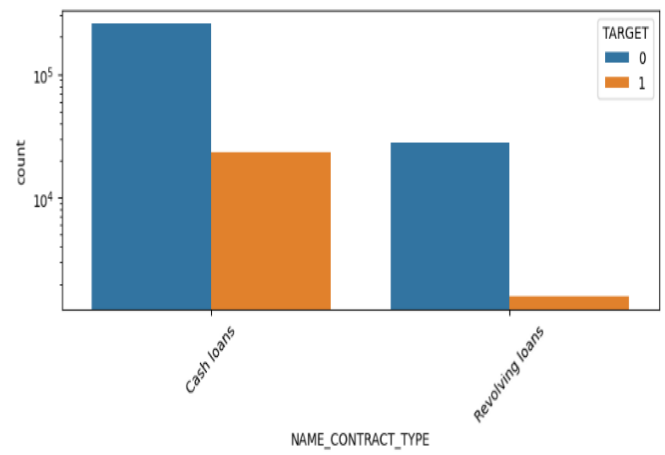
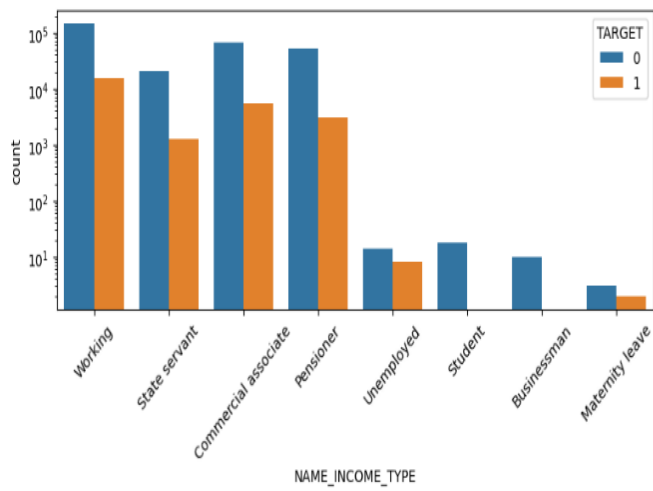
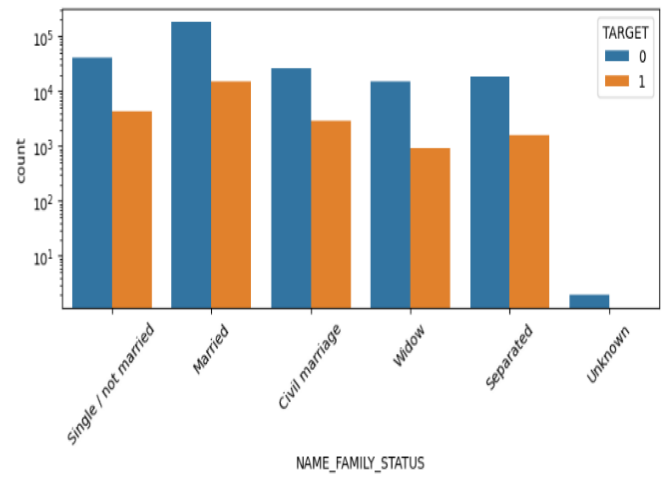
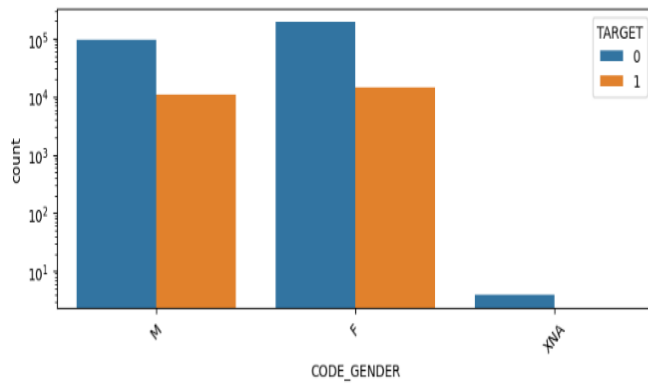


TARGET VS INCOME RANGE



TARGET VS CREDIT RANGE





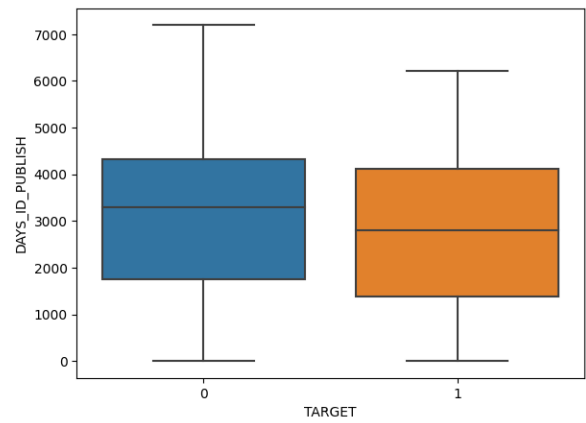
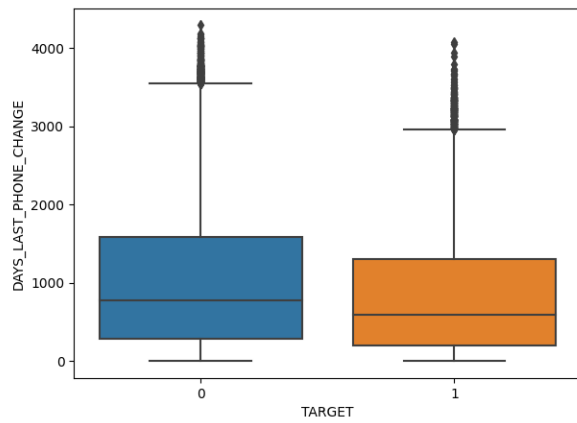
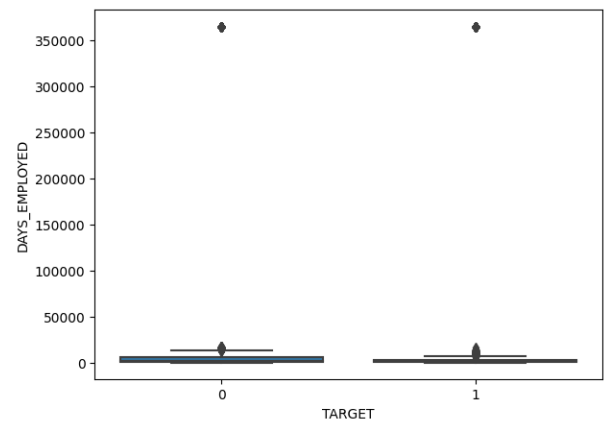
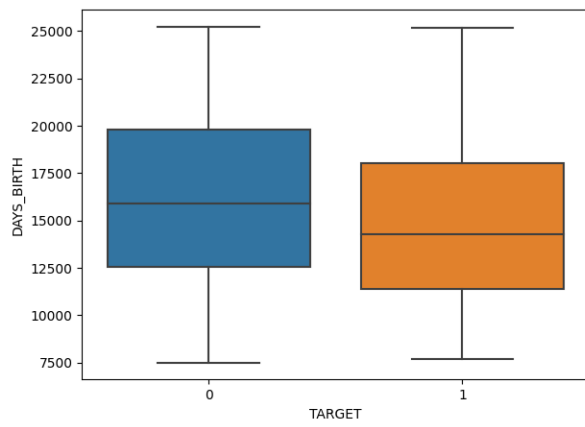
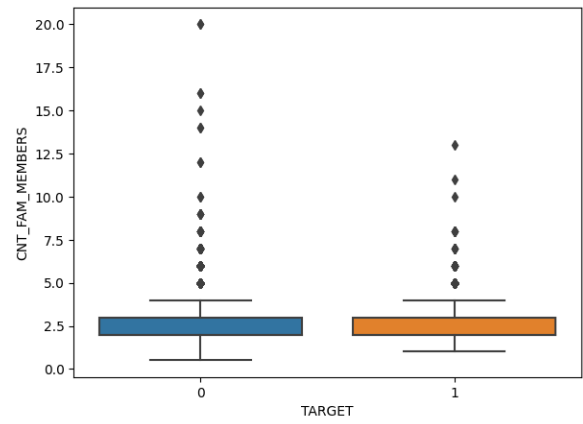
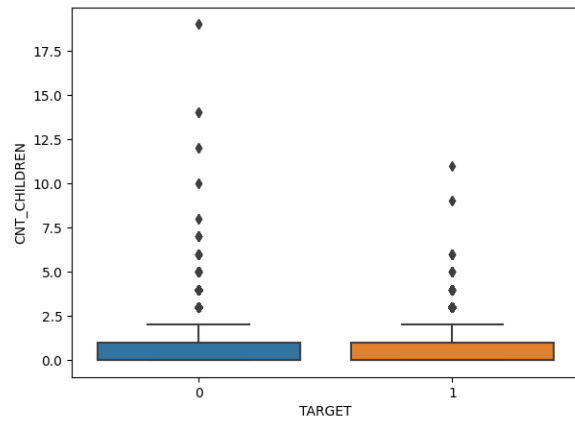
Boxplots: They are a great chart to use when showing the distribution of data points across a selected measure. These charts display ranges within variables measured. This includes the outliers, the median, the mode, and where the majority of the data points lie in the “box”.

Insights:

1.Number of children or Number of family members of client have or days employed no role in separating among defaulters and non defaulters due to same level range in Boxplot.

2.Non defaulters tend to have a slightly higher median value for days\_last\_phone\_change compared to defaulters. This could indicate that clients who change their phones less frequently are less likely to default on their loans.The fact that there is some overlap between the ranges of days\_last\_phone\_change for defaulters and non defaulters suggests that this variable alone may not be a strong predictor of loan default.

3.Clients who defaulted on their loans tended to change their identity documents closer to the loan application date than clients who did not default. The interquartile range (IQR) for non-defaulters is slightly larger than that of defaulters. This suggests that the distribution of days between identity document changes and loan applications for non-defaulters is more spread out than for defaulters.

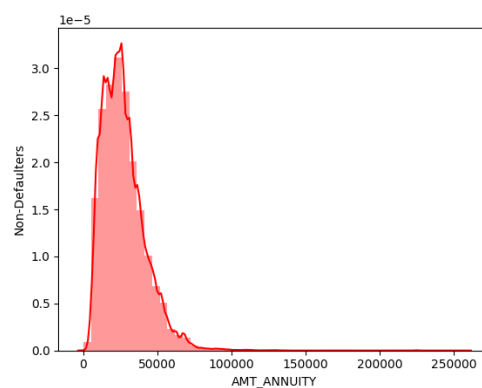
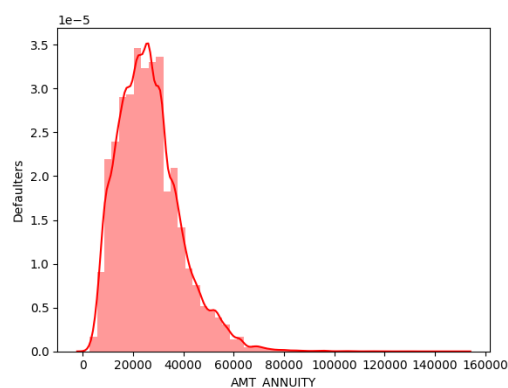


Distribution plots: Distribution plots represents the overall distribution of variable. They help us detect outliers and skewness, or get an overview of the measures of central tendency (mean, median, and mode).

Insight:

1. The annuity amount for both defaulters and non-defaulters is right-skewed, indicating that most customers have lower annuity amounts and a few have higher annuity amounts.

2. The mean annuity amount for defaulters is slightly lower than for non-defaulters (26481.74 vs 27162.47), but the standard deviation is higher for defaulters (12450.68 vs 14659.06), indicating greater variability in annuity amounts among defaulters. Overall, the annuity amount seems to be a relevant variable in predicting default, but further investigation is needed to understand the relationship between annuity amount and default.



```
1 amt_annuity_target1
```

count	24825.000000
mean	26481.744290
std	12450.676999
min	2722.500000
25%	17361.000000
50%	25263.000000
75%	32976.000000
85%	38331.000000
90%	42642.000000
100%	149211.000000
max	149211.000000

Name: AMT\_ANNUITY, dtype: object

```
1 amt_goods_target1
```

count	24825.000000
mean	488558.780737
std	311828.862782
min	0.514393
25%	238500.000000
50%	450000.000000
75%	675000.000000
85%	774000.000000
90%	900000.000000
100%	3600000.000000
max	3600000.000000

Name: AMT\_GOODS\_PRICE, dtype: object

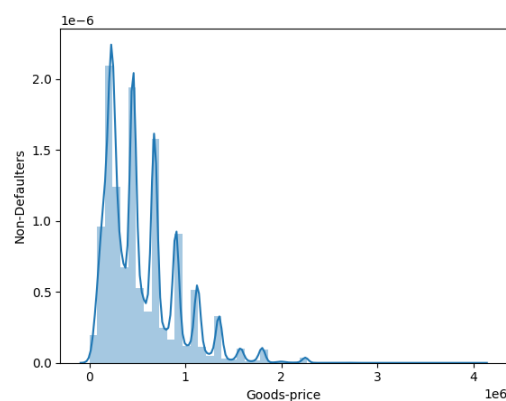
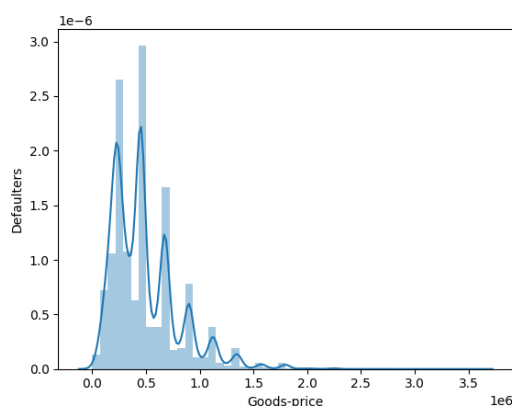
```
9] 1 amt_annuity_target0
```

count	282686.000000
mean	27162.470277
std	14659.064364
min	0.514393
25%	16456.500000
50%	24876.000000
75%	34749.000000
85%	41274.000000
90%	46134.000000
100%	258025.500000
max	258025.500000

```
1 amt_goods_target0
```

count	282686.000000
mean	542243.373949
std	373973.433881
min	0.514393
25%	238500.000000
50%	450000.000000
75%	685002.375000
85%	900000.000000
90%	1125000.000000
100%	4050000.000000
max	4050000.000000

Name: AMT\_GOODS\_PRICE, dtype: object



## Insights:

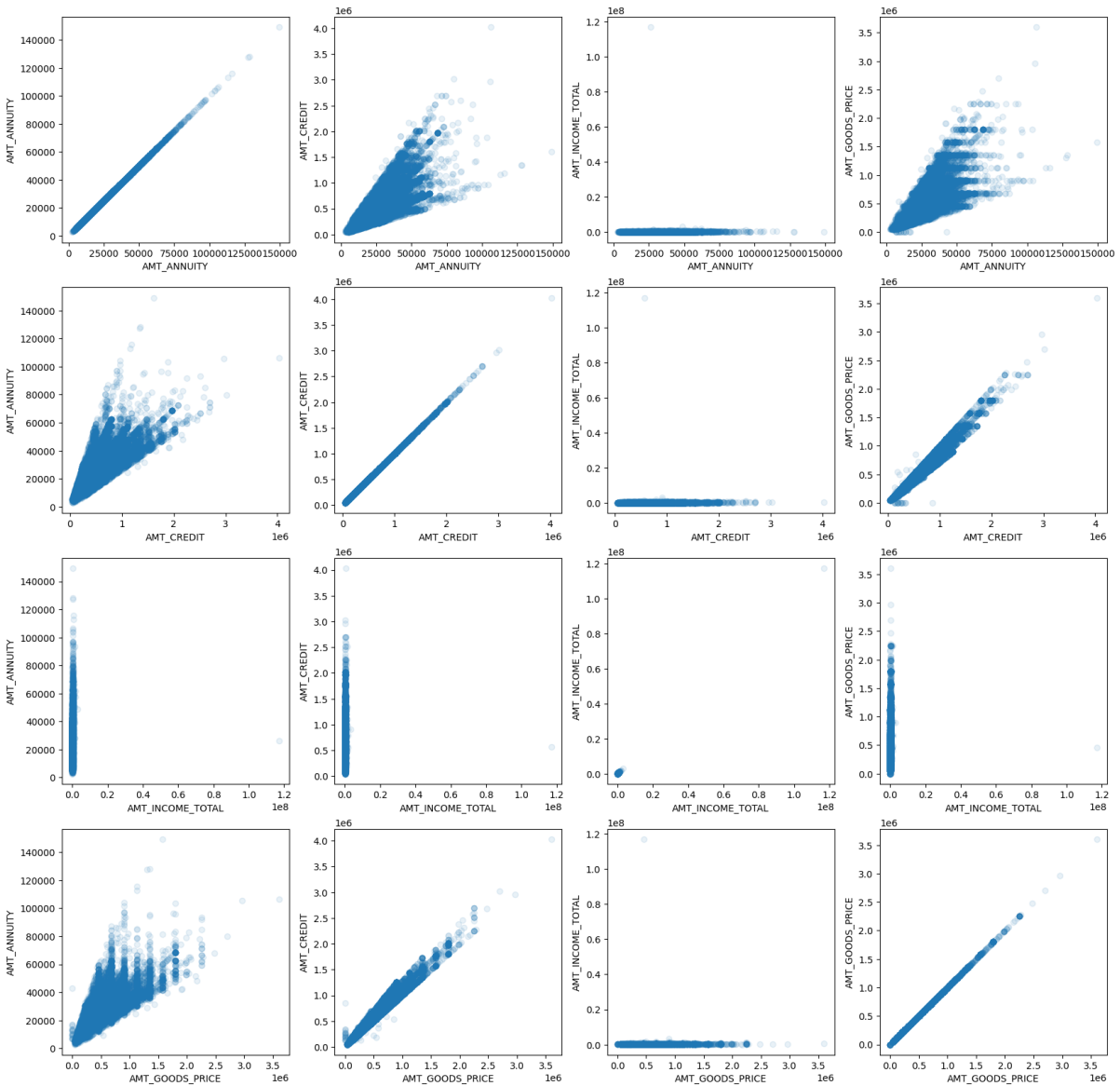
1.From the distribution plots, we can see that both the plots are right-skewed, which indicates that there are more loans with lower amounts and fewer loans with higher amounts.The mean of Amount\_goods\_price is higher for target 1 (defaulters) compared to target 0 (non-defaulters), which suggests that borrowers who default tend to take loans for more expensive goods.

2.Overall, the insights suggest that borrowers who take loans for expensive goods have a higher probability of defaulting on their loans. The standard deviation and higher percentiles of Amount\_goods\_price are higher for defaulters, which suggests that defaulters have more variation in the amount of loans taken and are more likely to take loans for expensive goods.

**Bivariate Analysis:** Bivariate analysis is an analysis of two variables to determine the relationships between them.

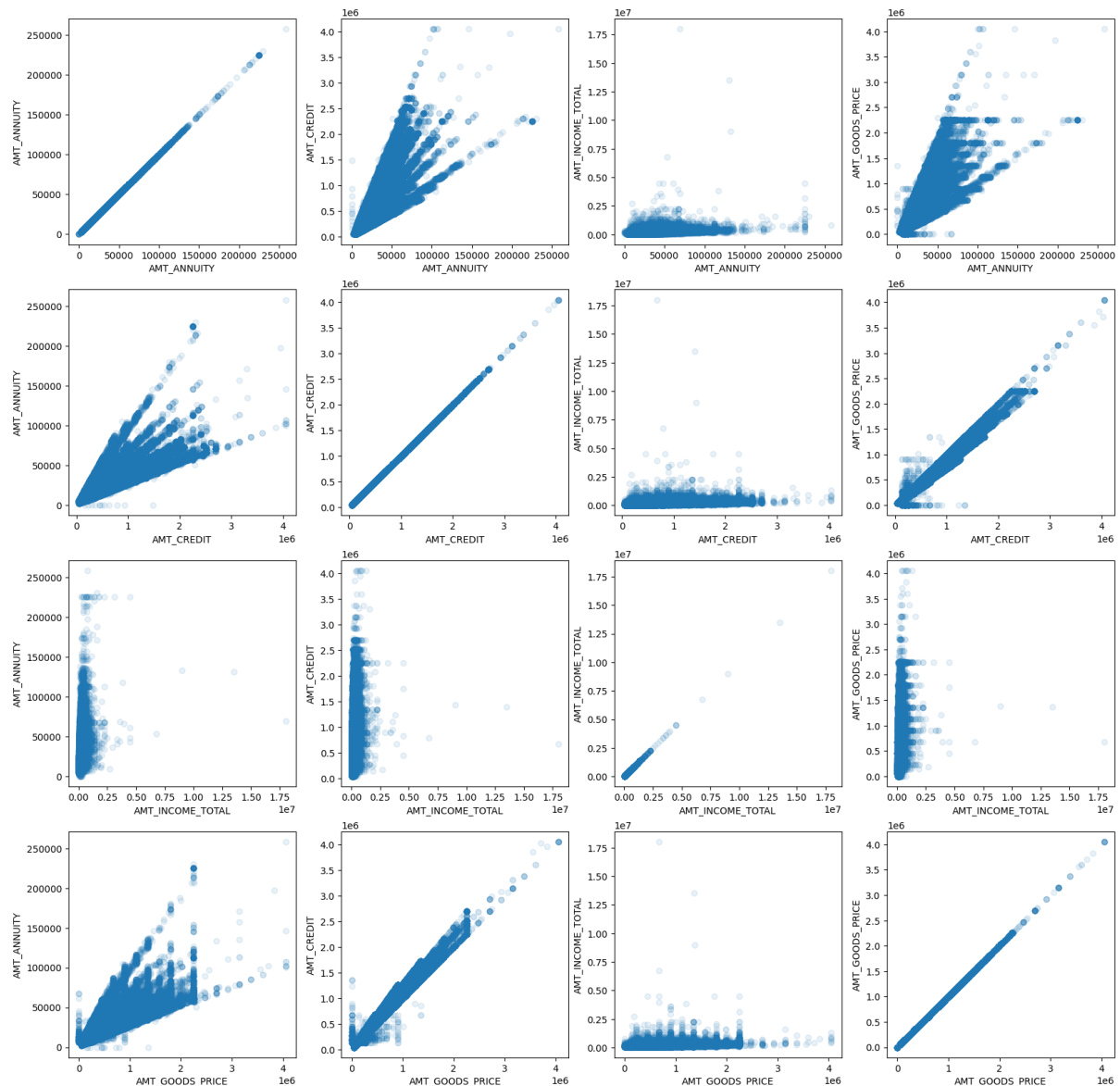
Scatter plots: The purpose of the scatter plot is to display what happens to one variable when another variable is changed.

This plot for defaulters(Target=1).



This plot for defaulters(Target=1).





## Insights:

- 1.AMT\_CREDIT and AMT\_GOODS\_PRICE have a positive linear relationship: This means that as the amount of credit increases, so does the amount of goods price. This relationship holds true for both defaulters and non-defaulters. This could indicate that people who take out larger loans also tend to purchase more expensive goods.
- 2.AMT\_CREDIT and AMT\_ANNUIITY have a positive linear relationship: This means that as the amount of credit increases, so does the amount of the EMI. This relationship holds true for both defaulters and non-defaulters. This could indicate that people who take out larger loans also tend to have higher monthly payments.
- 3.Overall, these insights suggest that there is a strong relationship between the amount of credit and other variables such as goods price and EMI. However, we should note that correlation does not necessarily imply causation, and further

analysis may be needed to fully understand these relationships and their implications.

Chisquare Test: The chi-square test of independence is used to test whether two categorical variables are related to each other.

```
NAME_CONTRACT_TYPE: chi-square=293.15, p-value=0.0000
CODE_GENDER: chi-square=920.79, p-value=0.0000
OCCUPATION_TYPE: chi-square=1975.08, p-value=0.0000
ORGANIZATION_TYPE: chi-square=1609.24, p-value=0.0000
NAME_TYPE_SUITE: chi-square=31.95, p-value=0.0000
AMT_INCOME_RANGE: chi-square=212.14, p-value=0.0000
AMT_CREDIT_RANGE: chi-square=385.01, p-value=0.0000
NAME_INCOME_TYPE: chi-square=1253.47, p-value=0.0000
NAME_EDUCATION_TYPE: chi-square=1019.21, p-value=0.0000
NAME_FAMILY_STATUS: chi-square=504.69, p-value=0.0000
NAME_HOUSING_TYPE: chi-square=420.56, p-value=0.0000
```

Insight:

The chi-square tests results indicate that there are significant associations between the target variable (default/non-default) and all of the categorical variables you tested. The p-values are all less than 0.05, indicating that the null hypothesis of independence between the categorical variable and the target variable can be rejected. So there may be differences in default rates

a.between cash loans and revolving loans


b.between male and female borrowers.

c.between borrowers of different occupation types/working in different types of organizations/with different income ranges/ different credit ranges

d.The income type/education level/familystatus of the borrower may be associated with default rates.

However, it is important to note that these are only associations and do not necessarily imply causation. Further analysis and modeling may be needed to fully understand the relationships between these variables and default rates.

Pivot table: Pivot tables is a powerful tool for bivariate analysis as they allow you to quickly summarize and compare data across multiple dimensions.

NAME_CONTRACT_TYPE	Cash loans	Revolving loans	
CODE_GENDER			
F	623137.316111	310206.382329	
M	637214.564407	352175.132440	
XNA	0.000000	399375.000000	

Insights: The mean credit value for cash loans is higher than revolving loans for both males and females.

NAME_FAMILY_STATUS	Civil marriage	Married	Separated	Single / not married	Unknown	Widow
NAME_EDUCATION_TYPE						
Academic degree	958458.38	727986.95	742390.31	634809.10	0	706914.00
Higher education	638638.64	733012.28	649704.80	583684.76	585000	583193.45
Incomplete higher	523530.67	627796.36	553278.00	467899.01	0	508372.41
Lower secondary	430803.42	541880.01	476315.81	402614.07	675000	400033.51
Secondary / secondary special	517703.94	613690.78	517591.43	477258.69	0	477028.70

Insights: Clients with academic degrees have the highest mean credit amount across all family status categories except Married and unknown catogeries. Married clients have highest mean credit amounts across the Higher education type categories. Clients with lower secondary degrees have lowest mean credit amounts across all family status categories except Unknown category.

NAME_HOUSING_TYPE	Co-op apartment	House / apartment	Municipal apartment	Office apartment	Rented apartment	With parents
TARGET						
0	167143.98	169396.90	168744.42	189573.02	168688.25	159665.29
1	173539.92	167226.74	160061.65	164217.38	158714.83	150995.33

Insights: Clients who live with parents have the lowest income among all housing types for both default and non-default cases. Clients who live in Co-op apartments have the highesh income among all housing types for default cases. Clients who live in office apartments have the highest income among all housing types for non default cases.

CODE_GENDER	F	M	XNA
TARGET			
0	85465.56	38166.79	4090.5
1	59619.34	26926.25	0.0

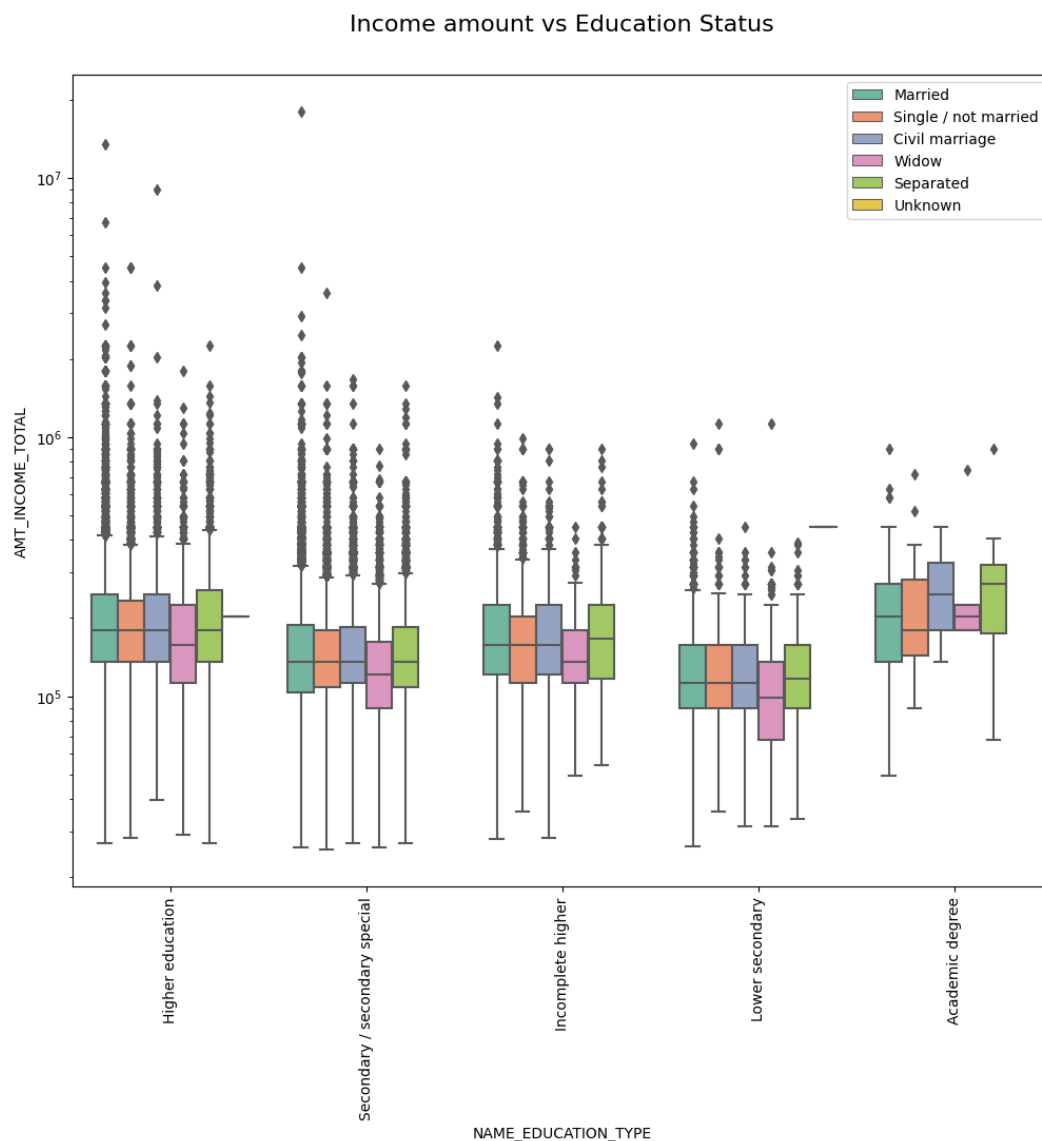
Insights: On average, individuals who default on their loans have been employed for fewer days before making their loan applications compared to those who do not default.

Among non-defaulters, females tend to have longer employment tenures compared to males. For instance, the average days employed for females in Target0 is 85,465.56, while for males, it is 38,166.79. Among defaulters, females still tend to have longer employment tenures compared to males, although the difference is not as pronounced as it is among non-defaulters. For instance, the average days employed for females in Target1 is 59,619.34, while for males, it is 26,926.25. There is a category in the gender column labeled XNa, which could stand for "Not applicable." The average days employed for this category are the lowest across all targets. However, we cannot make any meaningful conclusions about this category without additional information.

Overall, we can conclude that employment tenure is a significant factor in loan default rates, and females tend to have longer employment tenures compared to males.

Multiple variables plots:

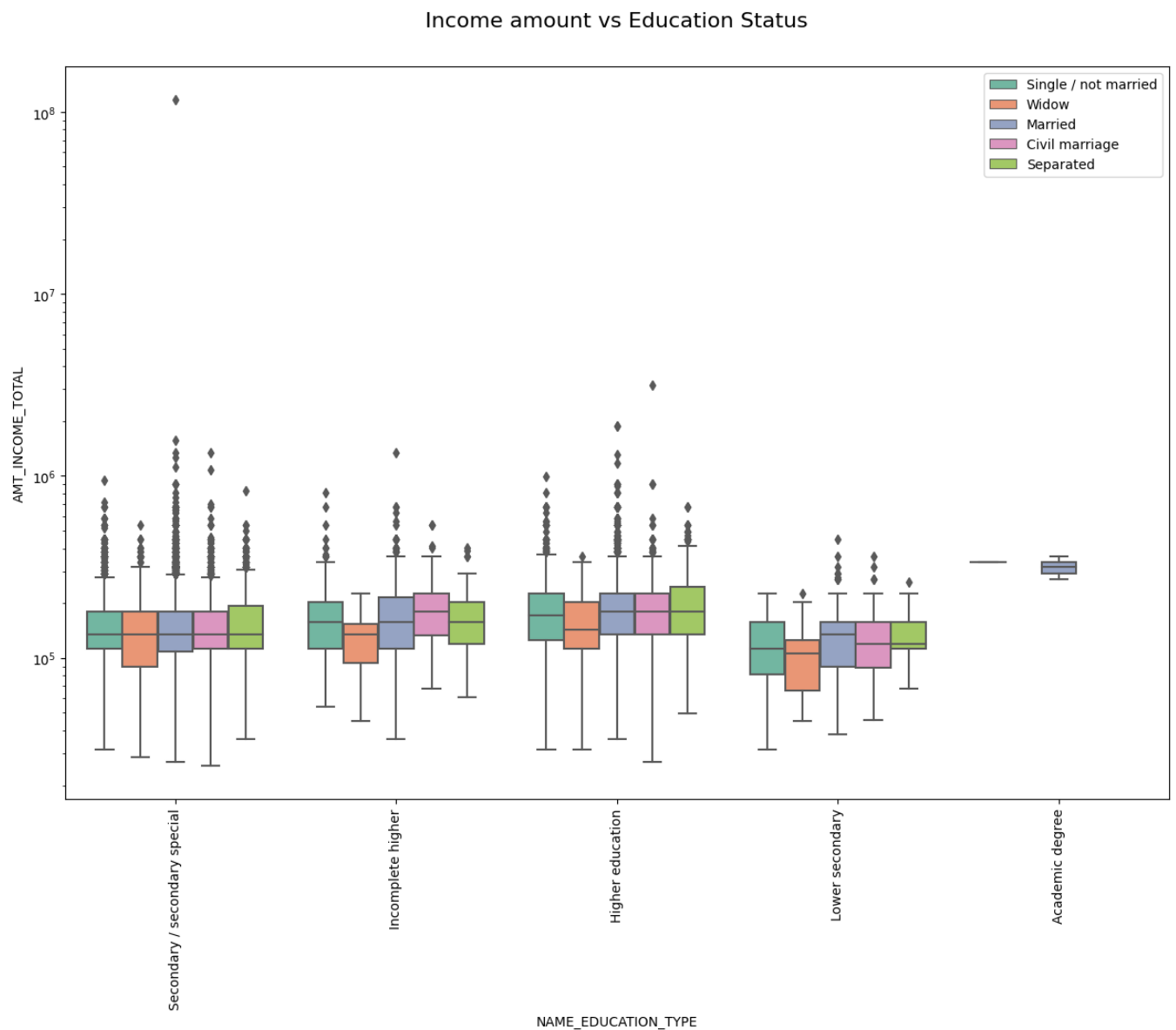
Income Amount VS Education status Vs Family status for Non - Defaulters(Target 0)



## Insights

1. Clients with all types of family status having academic degrees have very less outliers as compared to other types of education.
2. Clients having Higher Education, Incomplete Higher Education, Lower Secondary Education and Secondary/Secondary Special have a higher number of outliers.
3. From the above figure, we can say that some of the clients having Higher Education tend to have the highest income compared to others.
4. For all education categories except academic degree, income level of clients with married, single and civil marriage categories are almost the same as median of each are at same level.

## Income Amount VS Education status Vs Family status for Defaulters(Target 1)

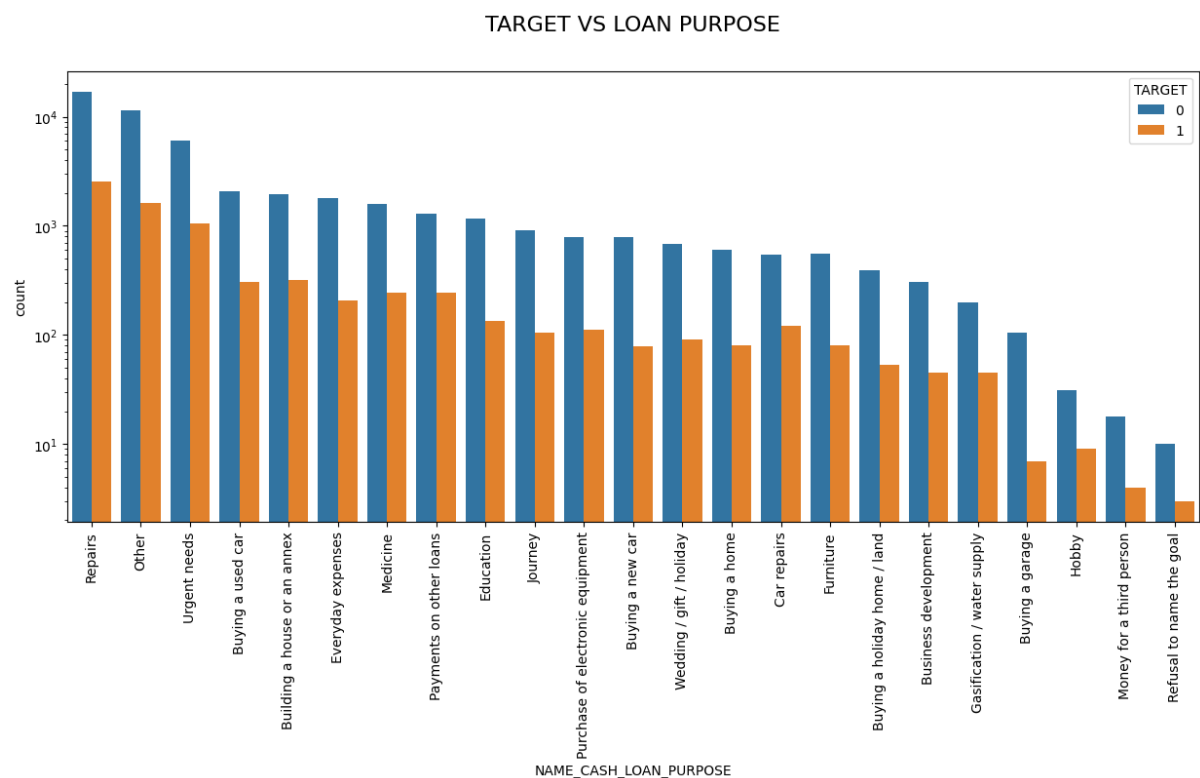


## Insights:

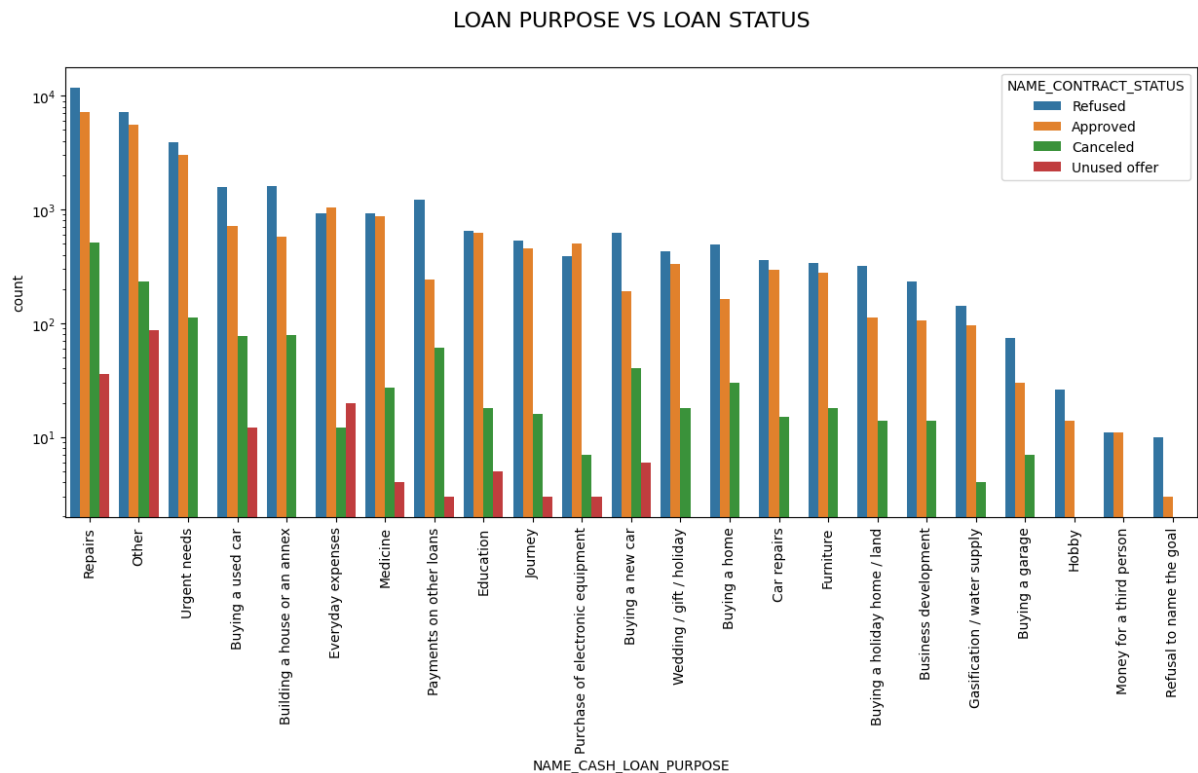
- 1.Defaulter clients have less income values than non defaulter clients.
- 2.Large number of outliers in secondary/secondary special and Higher education categories.
- 3.Married clients having high variability in income levels in all education categories.
- 4.Separated clients tend to have highest income in higher education category.
- 5.Clients with civil marriage tend to have highest income in Incomplete Higher and Secondary/secondary special education categories.

Previous application dataset:

Next we have to analyse previous application dataset also for that we can merge the both datasets.

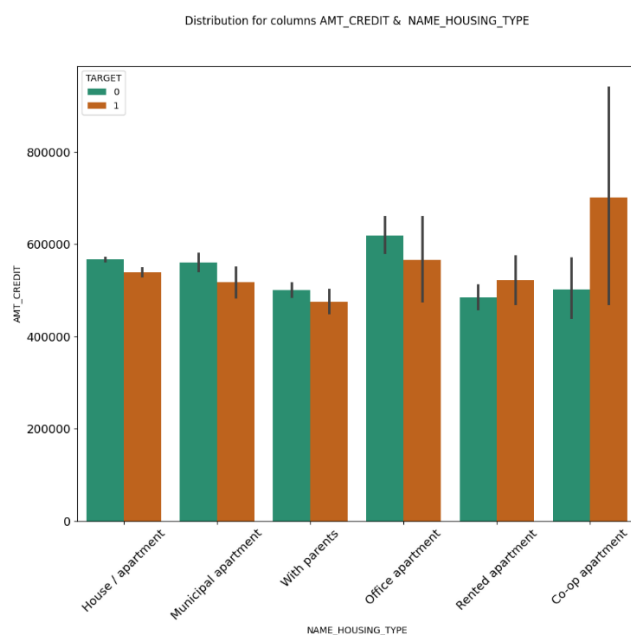


Insight: Repairs are facing difficulty in repaying loan.

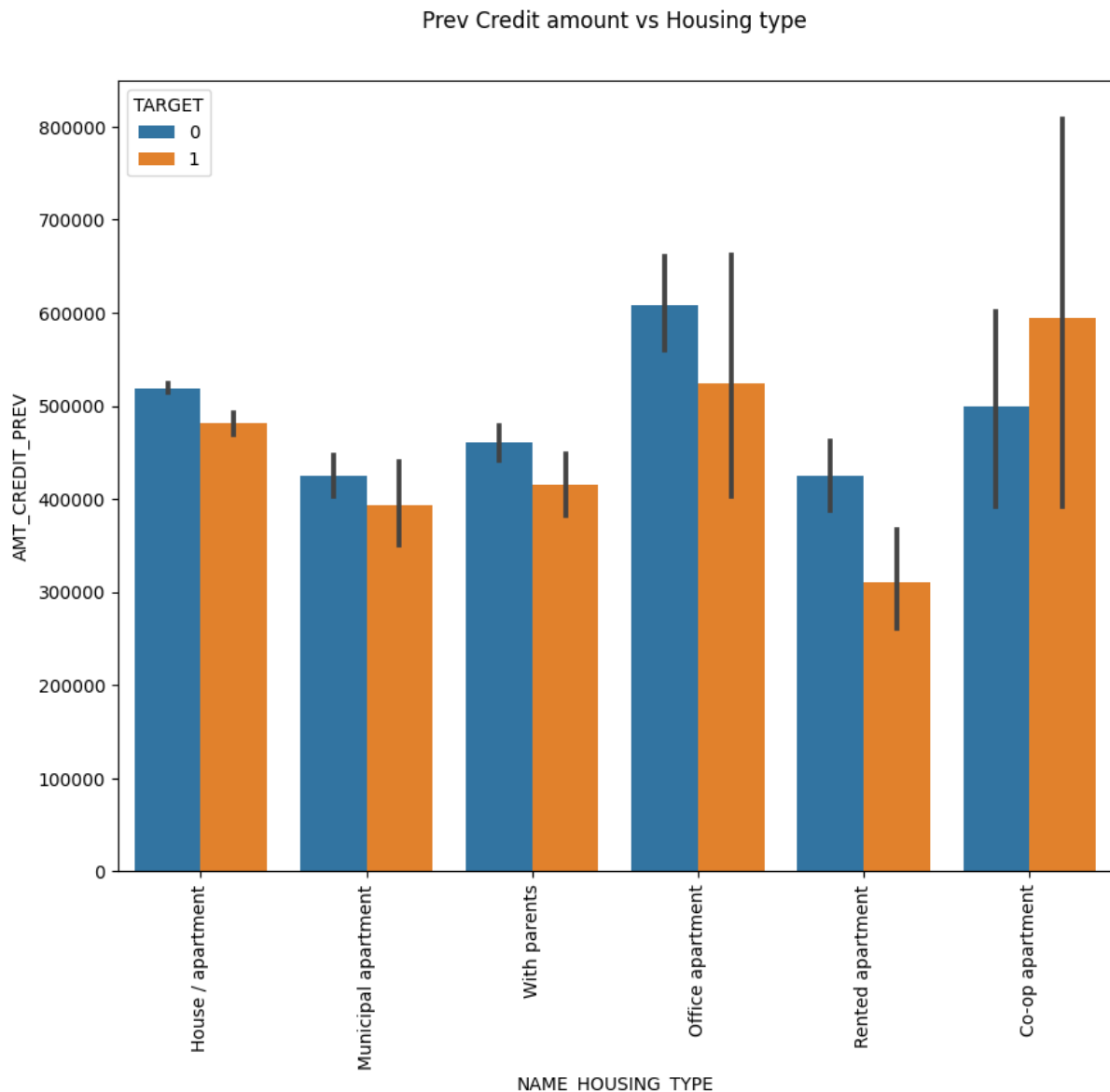


#### Insights:

- 1.Highest approvals in Repairs, Others and Urgent needs categories of loan purpose. Most rejection of loans came from purpose "Repairs".
- 2.For "Education" & "Medicine" purposes we have equal number of approvals and rejection.
- 3."Payments on other loans" and "Buying a new car' is having significant higher rejection than approvals.



Insights: Here for Housing type, office apartment is having higher credit of target 0 and co-op apartment is having higher credit of target 1.



Insight : Here for Housing type, office apartment is having higher credit of target 0 and co-op apartment is having higher credit of target 1.

**Task-5:** Find the top 10 **correlation** for the Client with payment difficulties and all other cases (Target variable). Note that you have to find the top correlation by segmenting the data frame w.r.t to the target variable and then find the top correlation for each of the segmented data and find if any insight is there. Say, there are 5+1(target) variables in a dataset: Var1, Var2, Var3, Var4, Var5, Target. And if you have to find top 3 correlation, it can be: Var1 & Var2, Var2 & Var3, Var1 & Var3. Target variable will not feature in this correlation as it is a categorical variable and not a continuous variable which is increasing or decreasing



**Approach:** We will calculate the correlation matrix for dataframes with target1 and target 2. Get the top 10 correlated variables in both lists. . Also we calculate correlation between other cols apart from the top 10 correlation list using a heatmap and try to find any patterns for negative correlation.

Correlation for Target=1/Default:

```
] 1 top_corr_target
```

OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998270
AMT_CREDIT	AMT_GOODS_PRICE	0.982854
REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	0.956637
CNT_FAM_MEMBERS	CNT_CHILDREN	0.885484
DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.869016
REG_REGION_NOT_WORK_REGION	LIVE_REGION_NOT_WORK_REGION	0.847885
LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.778540
AMT_ANNUITY	AMT_GOODS_PRICE	0.752891
	AMT_CREDIT	0.752195
FLAG_DOCUMENT_6	DAYS_EMPLOYED	0.617646

dtype: float64

Correlation for Target=0/Non-Default:

```
[268] 1 top_corr_non_target
```

OBS_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	0.998510
AMT_GOODS_PRICE	AMT_CREDIT	0.986966
REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.950149
CNT_FAM_MEMBERS	CNT_CHILDREN	0.878568
REG_REGION_NOT_WORK_REGION	LIVE_REGION_NOT_WORK_REGION	0.861861
DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.859371
REG_CITY_NOT_WORK_CITY	LIVE_CITY_NOT_WORK_CITY	0.830381
AMT_GOODS_PRICE	AMT_ANNUITY	0.776624
AMT_ANNUITY	AMT_CREDIT	0.771248
DAYS_EMPLOYED	DAYS_BIRTH	0.626114

dtype: float64

## Insights:

1.Comparing the two correlation lists, we can see that many of the same variables are present in both lists,

2.For the clients with payment difficulties (target=1):

a.There is a very high correlation between the number of observation of social circles in the last 60 and 30 days, which suggests that clients with payment difficulties may have a consistent social circle.

b.The correlation between the amount of credit and the amount of goods price is very high, indicating that clients with payment difficulties may have a tendency to overspend.

c.The rating of the region where the client lives and the rating of the region where the client works are highly correlated, indicating that clients with payment difficulties may not be very mobile and may not move around much for work.

d.The number of family members and the number of children have a high correlation, which is expected as having more children typically means having a larger family size.

3. For the clients with no payment difficulties (target=0):

a.There is a very high correlation between the number of observation of social circles in the last 60 and 30 days, similar to the clients with payment difficulties.

b.The correlation between the amount of goods price and the amount of credit is very high, indicating that clients with no payment difficulties may also have a tendency to overspend.

c.The rating of the region where the client lives and the rating of the region where the client works are highly correlated, similar to the clients with payment difficulties.

d.The number of family members and the number of children have a high correlation, which is also similar to the clients with payment difficulties.

4.Overall, these correlations suggest that the client's social and economic context, as well as their credit history, are important factors in predicting loan defaults.

Negative strong correlation: The correlation coefficient of -0.99(for both default/nondefault) suggests a very strong negative linear relationship between the variables 'days employed' and 'flag\_emp\_phone'. Since the value of the flag\_emp\_phone variable indicates whether the client provided a work phone or not, it is likely that clients who do not provide a work phone tend to have longer employment tenures with the same employer, while clients who provide a work phone tend to have shorter employment tenures.

Negative Correlations:

	Var1	Var2	Correlation
0	DAYS_EMPLOYED	FLAG_EMP_PHONE	-0.999756

Correlation for other cols of interest like

CNT\_CHILDREN', 'AMT\_INCOME\_TOTAL', 'AMT\_CREDIT', 'AMT\_ANNUITY', 'AMT\_GOODS\_PRICE', 'DAYS\_EMPLOYED', 'CNT\_FAM\_MEMBERS', 'DAYS\_BIRTH'

	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	DAYS_EMPLOYED	CNT_FAM_MEMBERS	DAYS_BIRTH
CNT_CHILDREN	1.000000	0.027397	0.003081	0.020949	-0.000723	-0.245174	0.878568	-0.336966
AMT_INCOME_TOTAL	0.027397	1.000000	0.342799	0.418906	0.349473	-0.140392	0.034237	-0.062609
AMT_CREDIT	0.003081	0.342799	1.000000	0.771248	0.986966	-0.070104	0.064534	0.047378
AMT_ANNUITY	0.020949	0.418906	0.771248	1.000000	0.776624	-0.104933	0.075811	-0.012233
AMT_GOODS_PRICE	-0.000723	0.349473	0.986966	0.776624	1.000000	-0.068160	0.062649	0.045135
DAYS_EMPLOYED	-0.245174	-0.140392	-0.070104	-0.104933	-0.068160	1.000000	-0.238292	0.626114
CNT_FAM_MEMBERS	0.878568	0.034237	0.064534	0.075811	0.062649	-0.238292	1.000000	-0.285811
DAYS_BIRTH	-0.336966	-0.062609	0.047378	-0.012233	0.045135	0.626114	-0.285811	1.000000

### Insights:(Non-default):

1.Income amount has a moderate positive relation with credit and annuity prices.This suggests that individuals with higher incomes are able to afford larger loan payments, and thus are able to take out larger loans.

2.The correlation coefficient between AMT\_INCOME\_TOTAL and CNT\_FAM\_MEMBERS is positive but weak (0.034237). This suggests that there is a slight tendency for individuals with higher incomes to have larger families, but the effect is not strong.

3.The correlation coefficient between AMT\_INCOME\_TOTAL and DAYS\_BIRTH is negative but weak (-0.062609). This suggests that there is a slight tendency for individuals with higher incomes to be younger, but the effect is not strong.

4.Credit amount is having a weak negative corelation with the DAYS\_EMPLOYED.This means that as the number of days an applicant has been employed increases, the amount of credit they request may decrease slightly.One possible explanation for this relationship could be that individuals who have been employed for a longer time may have built up more savings or have a more stable income, which may allow them to request smaller amounts of credit. On the other hand, individuals who have been employed for a shorter time may have less financial stability and may need to request larger amounts of credit to meet their financial needs.

5.The correlation coefficient of 0.064534 between the credit amount and the number of family members of the client suggests a weak positive correlation between these two variables. This means that as the number of family members increases, there is a slightly higher likelihood that the client may request a larger amount of credit.

6.a weak positive correlation between credit amount and days of birth. This means that, on average, as the days of birth of the client increase, the credit amount also tends to increase slightly.

	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	DAYS_EMPLOYED	CNT_FAM_MEMBERS	DAYS_BIRTH
CNT_CHILDREN	1.000000	0.004796	-0.001675	0.031257	-0.008122	-0.192864	0.885484	-0.259109
AMT_INCOME_TOTAL	0.004796	1.000000	0.038131	0.046421	0.037647	-0.014977	0.006654	-0.003096
AMT_CREDIT	-0.001675	0.038131	1.000000	0.752195	0.982854	0.001930	0.051224	0.135316
AMT_ANNUITY	0.031257	0.046421	0.752195	1.000000	0.752891	-0.081207	0.075711	0.014303
AMT_GOODS_PRICE	-0.008122	0.037647	0.982854	0.752891	1.000000	0.006734	0.047245	0.135516
DAYS_EMPLOYED	-0.192864	-0.014977	0.001930	-0.081207	0.006734	1.000000	-0.186515	0.582185
CNT_FAM_MEMBERS	0.885484	0.006654	0.051224	0.075711	0.047245	-0.186515	1.000000	-0.203267
DAYS_BIRTH	-0.259109	-0.003096	0.135316	0.014303	0.135516	0.582185	-0.203267	1.000000

Insights:(Default):

1.Credit amount is inversely proportional to number of children client have, means credit amount is higher for less children count client have and vice-versa.

2.Income amount is inversely proportional to days of date of birth , which means income amount is higher for younger clients and vice versa.

**Task-6: Include visualizations and summarize** the most important results in the presentation. You are free to choose the graphs which explain the numerical/categorical variables. Insights should explain why the variable is important for differentiating the clients with payment difficulties with all other cases.

## FINAL CONCLUSIONS:

1.Bank should focus on acquiring more clients which are students, pensioners, widows or businessmen, with housing type “CO-OP Apartment” as there are very less no defaulters or late payments from them .Bank can focus mostly on housing type with parents or House\apartment or municipal apartment for successful payments.

2.Bank should restrict loan amount for people earning less than 300k as the no. of defaulters is high.

3.Bank should focus less on “Working” type people who are living in house/apartment as they have the highest number of defaulters.

4.Bank should approve less loans for “repair” , “urgent”, “others” purposes as these are having highest defaulters.

5.Banks should focus less on clients who change their phones more frequently and change their identity documents closer to application date as there are high chances of them being defaulters.

6. Bank should restrict loan amount on people who take loans for expensive goods and those clients who are employed for a few days before making their loan applications.

**Results:** I have learnt a lot about the EDA steps and how to find the variables that drives the loan default and what are the significant areas for the banks to focus on and the categories where bank needs to restrict loan amount. Also understood the steps of Univariate and Bivariate analysis and how to derive insights from each of the plots.

**THANK YOU**