RISK ANALYSIS IN BANKING

Table of contents

Introduction

- This project focuses on uncovering patterns that signal potential challenges for clients in meeting their installment payments. The insights gained can inform decisions such as loan approval, adjusting loan amounts, or offering loans to riskier applicants at higher interest rates. By leveraging exploratory data analysis (EDA), the goal is to identify applicants who are likely to repay their loans successfully. Essentially, the aim is to understand the key factors driving loan defaults – the variables that strongly predict default. Armed with this understanding, the company can optimize its portfolio management and risk assessment strategies.
- Technology Stack Utilized: Jupyter Notebook – employed for data cleaning, analysis, and visualization, leveraging Python libraries including NumPy, Pandas, Matplotlib, and Seaborn.

Data description

1st Dataset - PREVIOUS APPLICATION

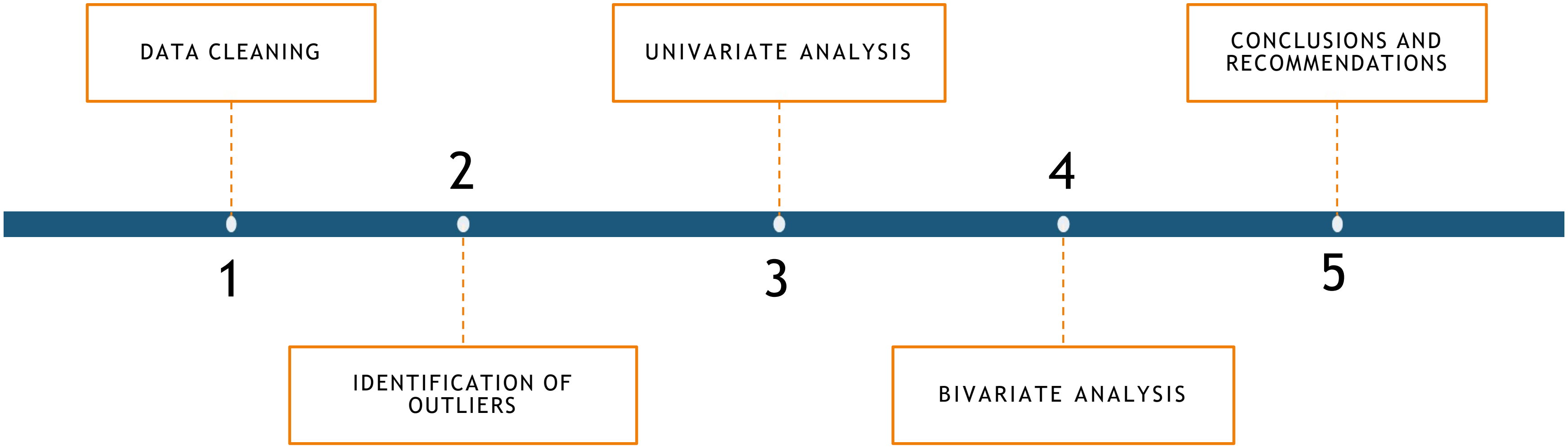
The client's prior loan information is contained in "previous_application.csv" (37 columns).
The Target variable in "previous_application.csv" is "NAME_CONTRACT_STATUS" which contains: Approved, Refused, Cancel, Unused offer

- Approved: The Company has approved loan application.
- Refused: The Company has rejected the loan application.
- Cancel: The client cancelled the application during approval.
- Unused offer: Loan has been cancelled by the client but on different stages of the process.

2nd Dataset - APPLICATION DATA

The client's complete information at the time of application is contained in "application_data.csv" (122 columns).
The Target variable in "application_data.csv" is "TARGET" which contains:

- 1: means all the client has difficulties in payment.
- 0: all other cases. Refused: The Company has rejected the loan application.


1 DATA CLEANING

2 IDENTIFICATION OF OUTLIERS

3 UNIVARIATE ANALYSIS

4 BIVARIATE ANALYSIS

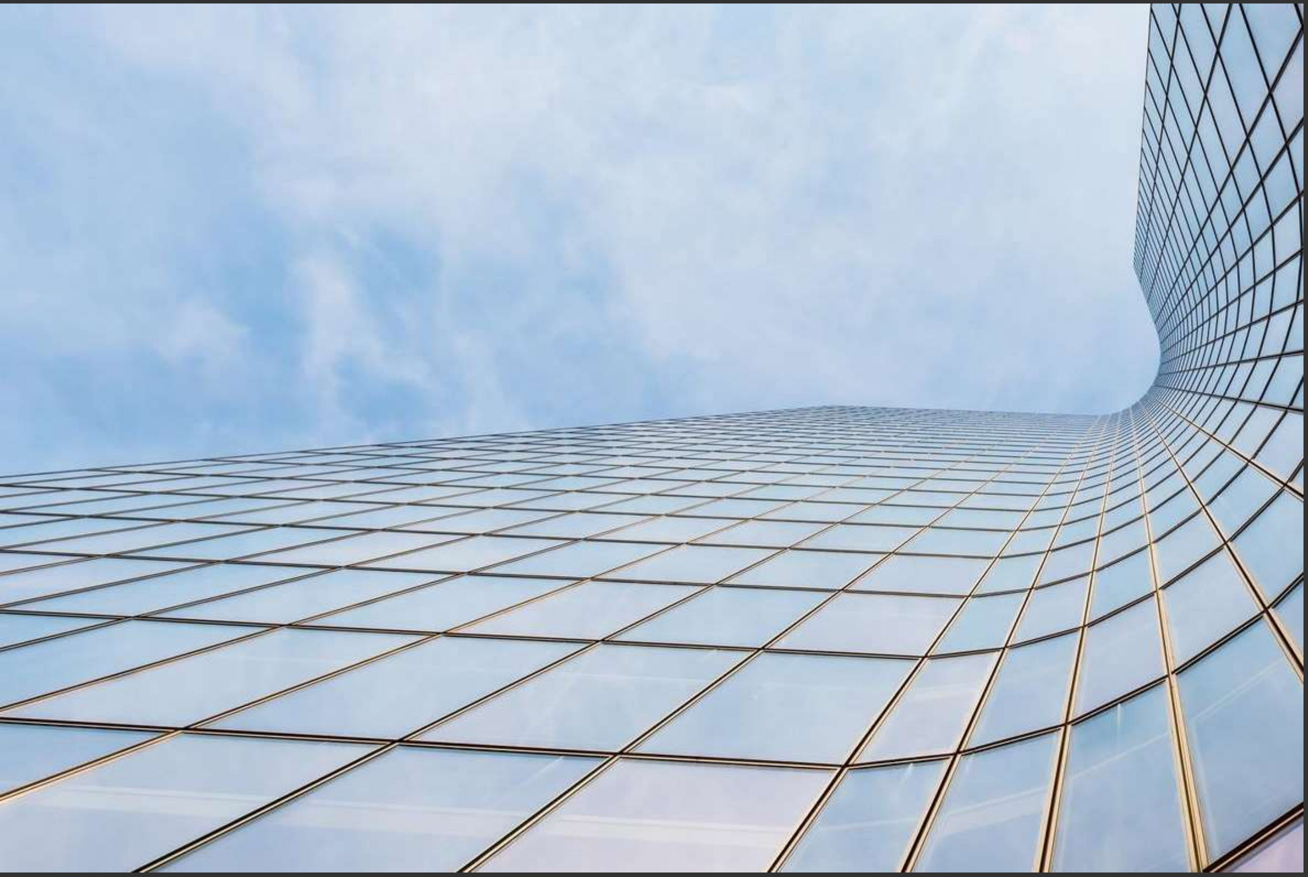
5 CONCLUSIONS AND RECOMMENDATIONS

Project timeline

Data cleaning

PREVIOUS APPLICATION and APPLICATION DATA

- Checked columns for null values and removed columns having null values greater than 40% for application data and removed columns having null values greater than 30% for previous application data and the columns which are not necessary for analysis from both the datasets.
- Imputing null values with suitable values (mean or median)
- Binning certain columns into categorical columns which will be used for analysis.
- Checked the datatype of variables and converted them into suitable datatypes.
- Checked the dataset for duplicates and found that there are no duplicate values.

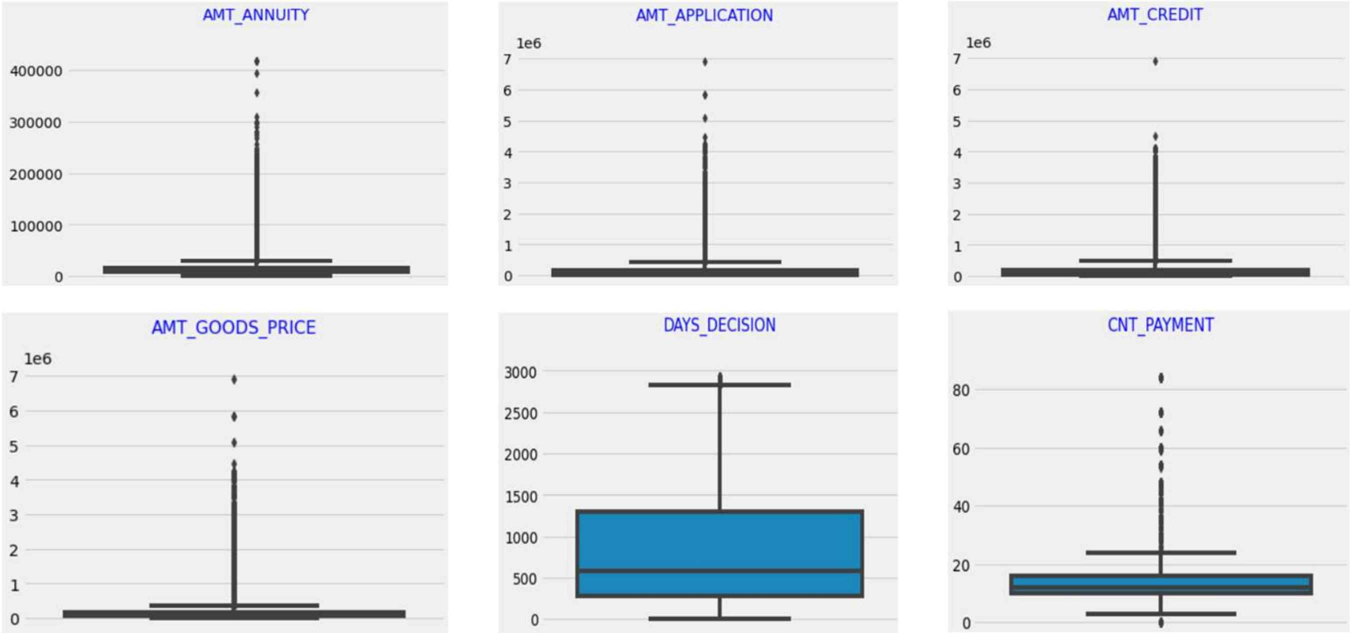


Identification of outliers

Identification of outliers

Previous application data

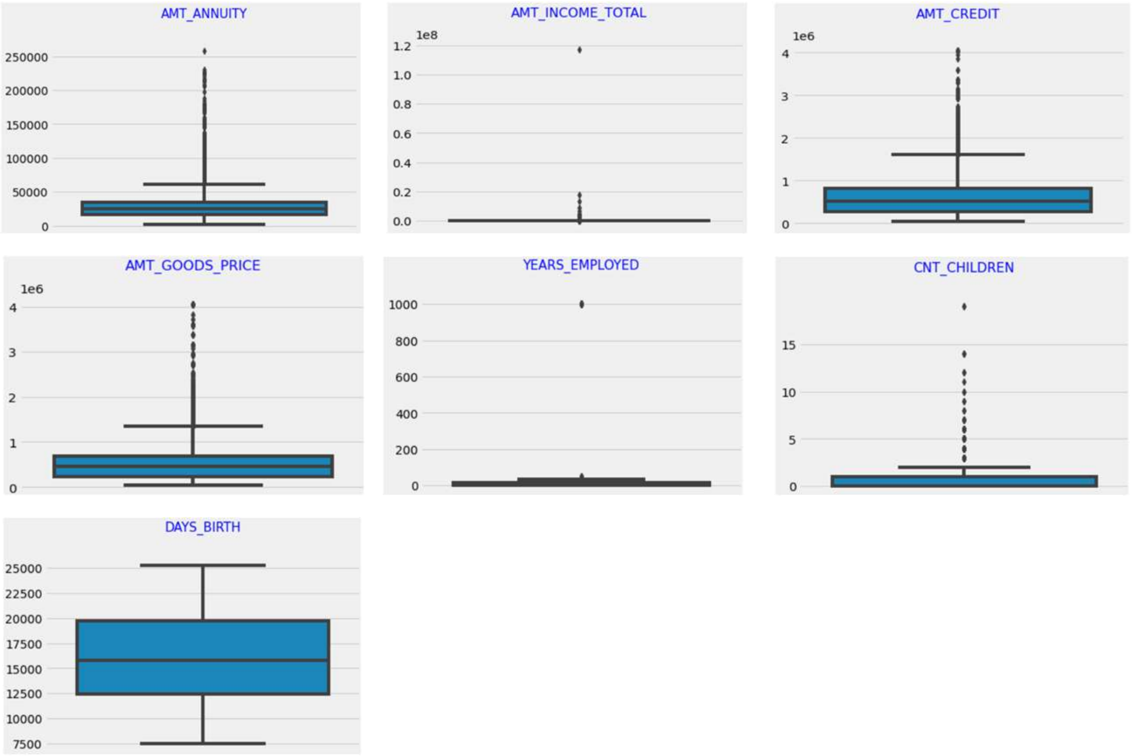
By using both IQR (Interquartile Range) method and by visualizing boxplots we can see that the columns 'AMT_ANNUITY', 'AMT_APPLICATION', 'AMT_CREDIT', 'AMT_GOODS_PRICE', 'DAYS_DECISION', 'CNT_PAYMENT' have outliers.

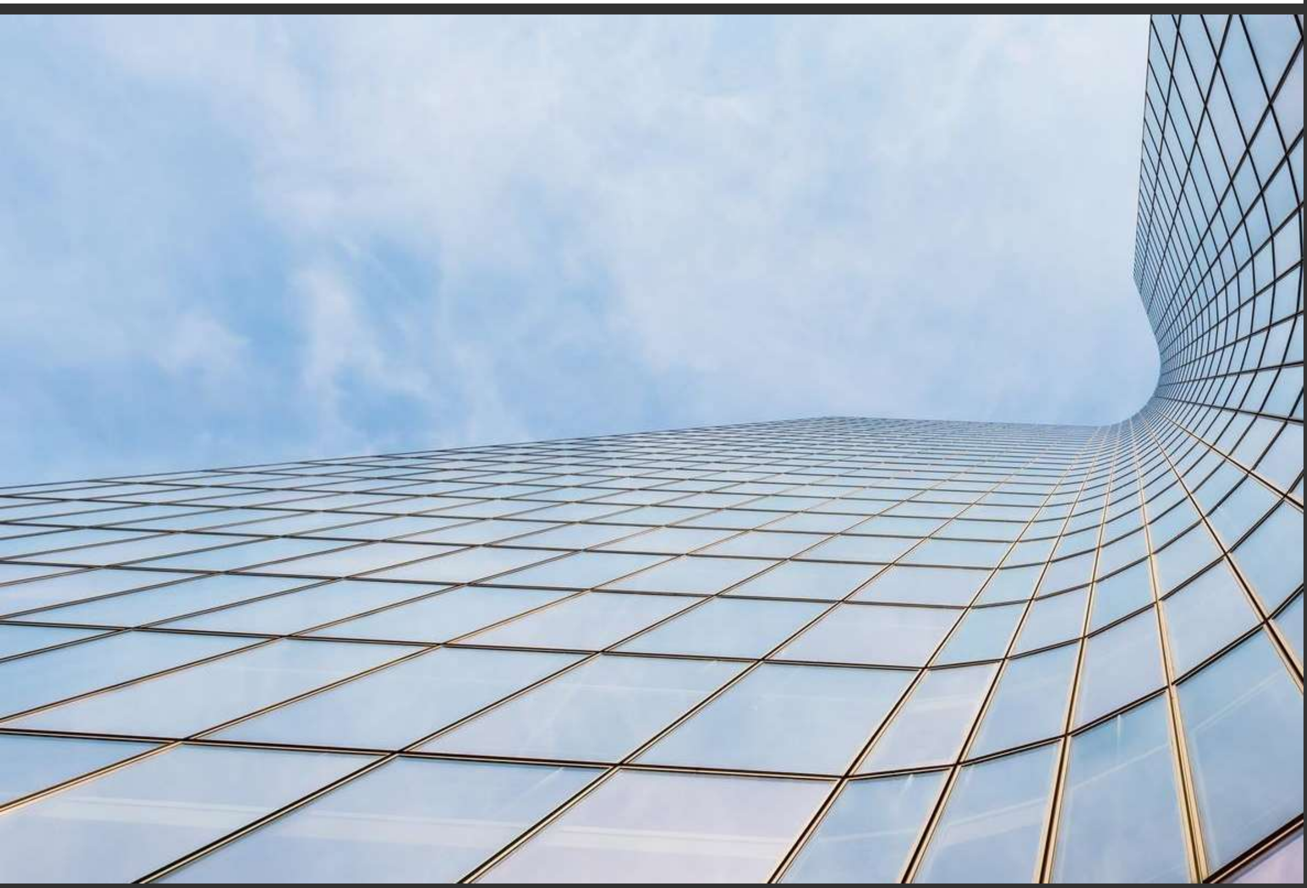


Identification of outliers

Application data

- By using both IQR (Interquartile Range) method and by visualizing boxplots we can see that the columns 'AMT_ANNUITY’, 'AMT_INCOME_TOTAL’, 'AMT_CREDIT’, 'AMT_GOODS_PRICE’, 'YEARS_EMPLOYED’, 'CNT_CHILDREN' have outliers.
- We can see that the maximum outlier value for the column CNT_CHILDREN is 19 which is highly unlikely, and the maximum outlier value for YEARS_EMPLOYED is 1000 which is not possible.
- The column DAYS_BIRTH does not have any outliers.



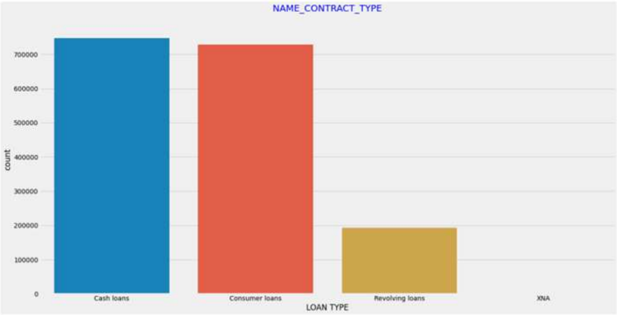


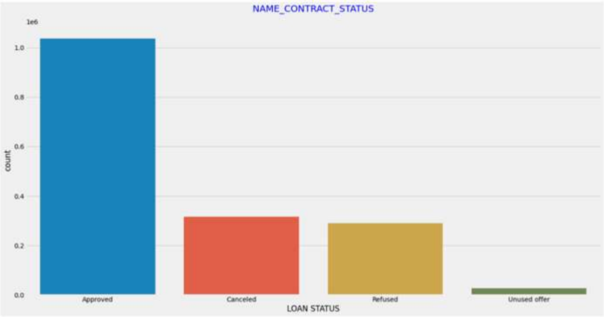
Univariate analysis

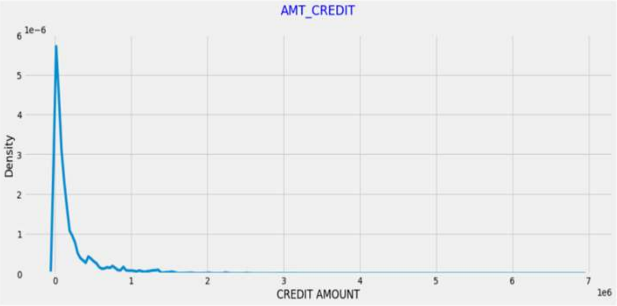
Univariate analysis

Previous application data

- Majority of the previous loans are either cash loans or consumer loans.
- Majority of the loan applications were approved.
- Most of the previous loans have low credit amount.







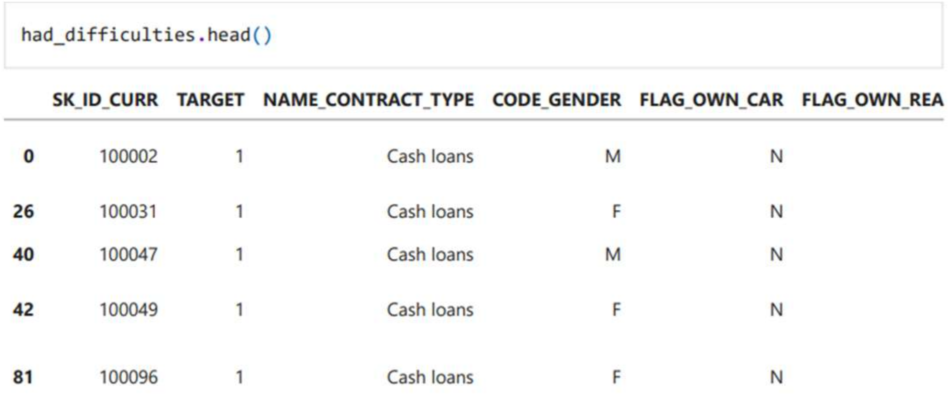
Univariate analysis

Application data

- The data have been divided into two data frames based on the TARGET variable for comparative analysis:

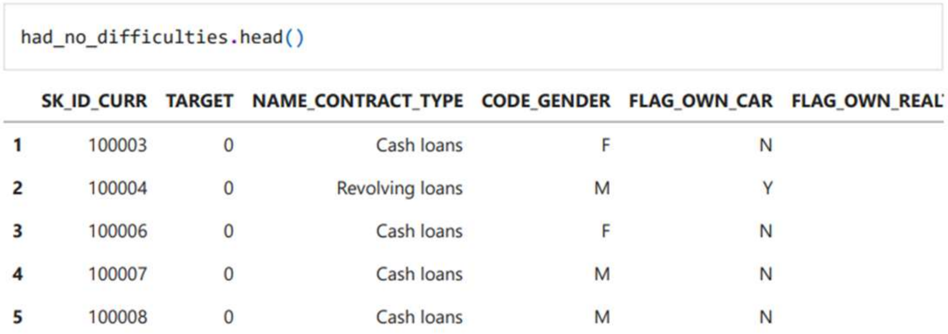
  1) **DEEFAULTERS** – had_difficulties

  2) **NON DEEFAULTERS** – had_no_difficulties



```
had_difficulties.head()
```

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REA
0	100002	1	Cash loans	M	N	
26	100031	1	Cash loans	F	N	
40	100047	1	Cash loans	M	N	
42	100049	1	Cash loans	F	N	
81	100096	1	Cash loans	F	N	



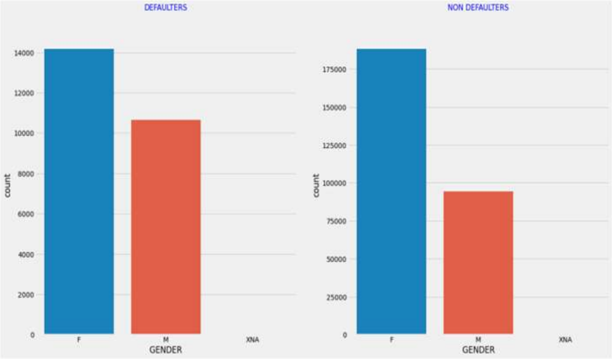
```
had_no_difficulties.head()
```

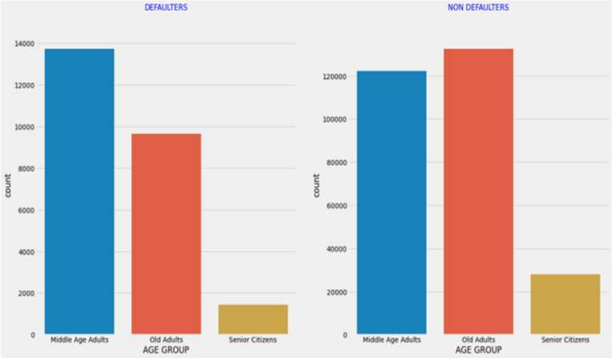
	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REAL
1	100003	0	Cash loans	F	N	
2	100004	0	Revolving loans	M	Y	
3	100006	0	Cash loans	F	N	
4	100007	0	Cash loans	M	N	
5	100008	0	Cash loans	M	N	

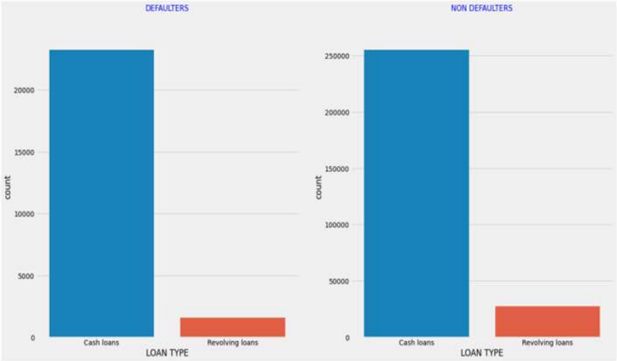
Univariate analysis

Application data

- **How is repayment of loan affected by gender?**
  The number of female applicants is greater than male applicants in both defaulters and non defaulters list
- **How is repayment of loan is affected by age?**
  Middle-aged adults are likely to default more, and Senior citizens are the least to default.
- **How is repayment of loan affected by loan type?**
  Both cases seem to follow the same pattern where Cash loans is greater than Revolving loans.



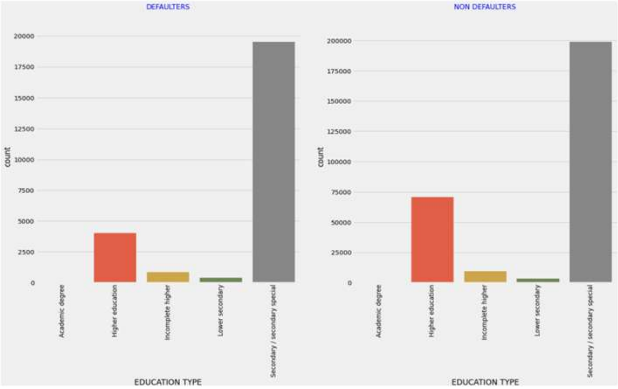


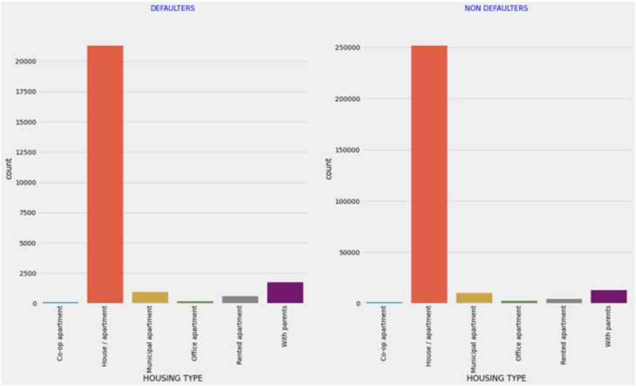


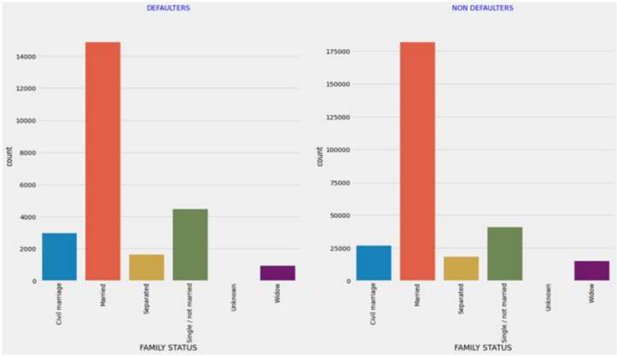
Univariate analysis

Application data

- **How is repayment of loan affected by education type?**
  Both cases seem to follow the same pattern where applicants who have completed Secondary/Secondary special have highest count followed by applicants with Higher education.
- **How is repayment of loan affected by housing type?**
  Both cases seem to follow the same pattern where most applicants live in House/apartment.
- **How is repayment of loan affected by family status?**
  Both cases seem to follow the same pattern where most applicants are married.



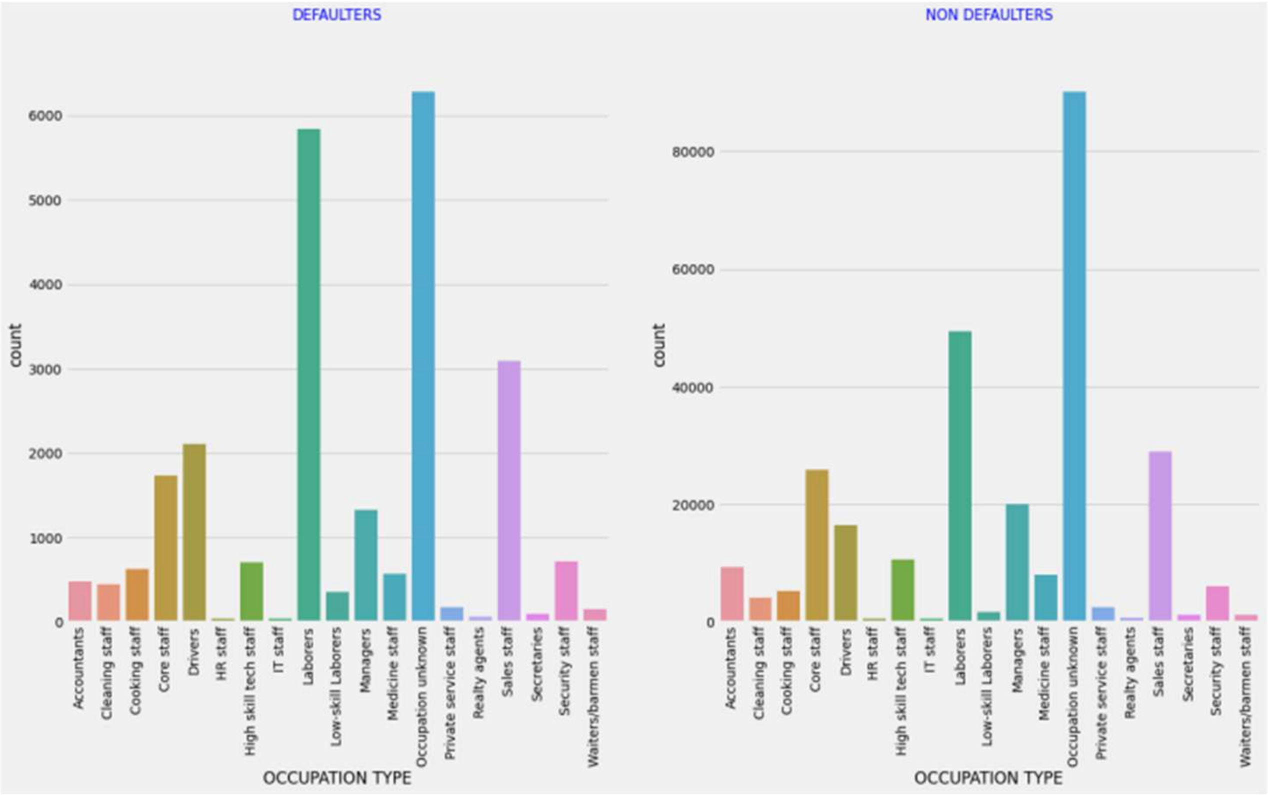




Univariate analysis

Application data

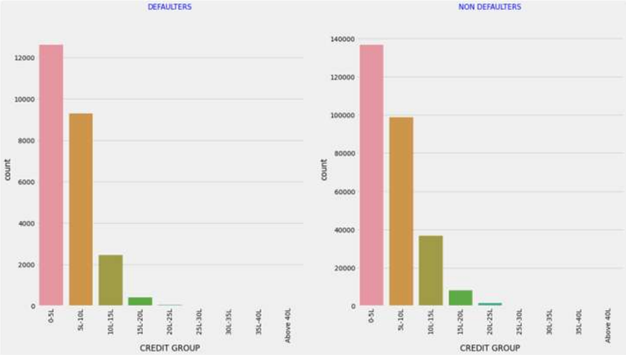
- **How is repayment of loan affected by occupation type?**
  In both Defaulters and Non defaulters the occupation is unknown for a huge percentage of applicants. For the applicants who have mentioned their occupation, we find that Laborers are most likely to default followed by Sales staff.
- **How is repayment of loan is affected by income type?**
  Both cases seem to follow the same pattern where most applicants who are working have higher count in both Defaulters and Non defaulters.

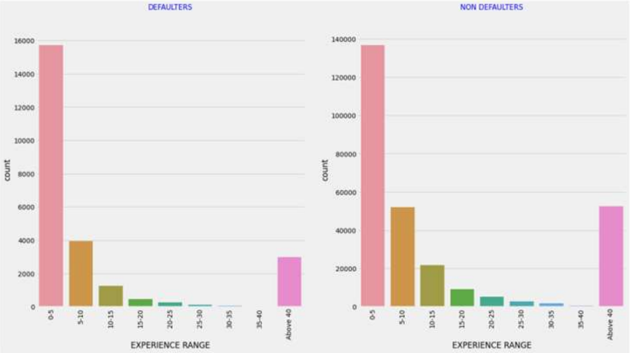


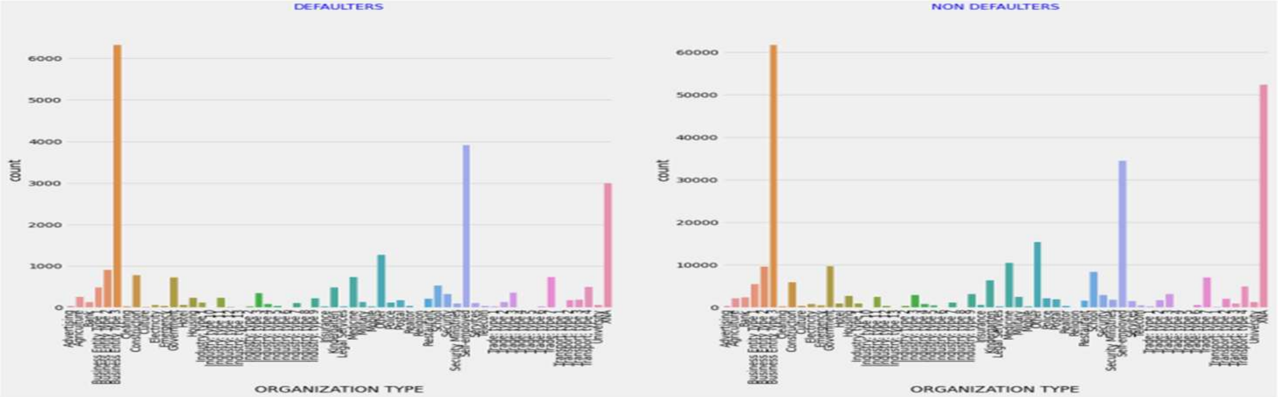
Univariate analysis

Application data

- **How is repayment of loan affected by credit amount group?**
  Applicants with credit amount 0-5 lakhs are the most to default.
- **How is repayment of loan affected by Experience Range?**
  Applicants who have less work experience in the range 0-5 years default more than applicants with more experience.
- **How is repayment of loan is affected by organization type?**
  We find that applicants who are most likely to default are those who work in Business_Entity_Type 3 followed by those who are self-employed.





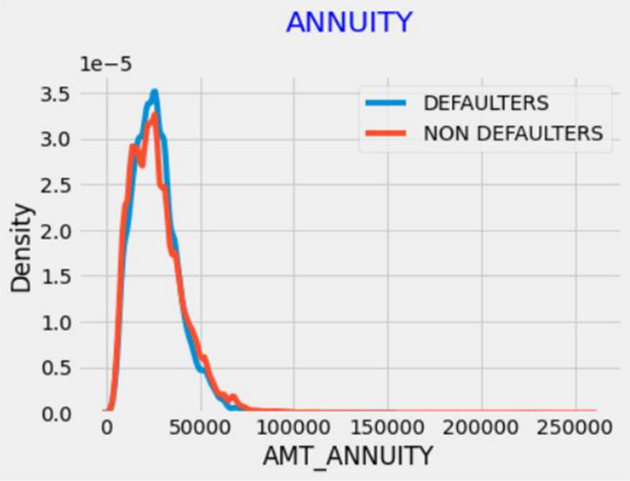


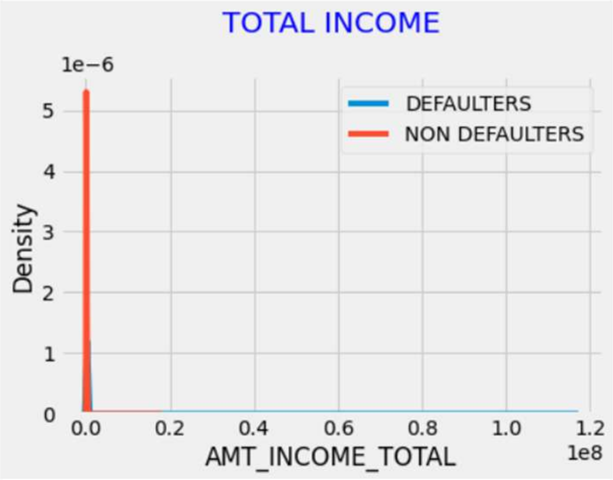
Univariate analysis

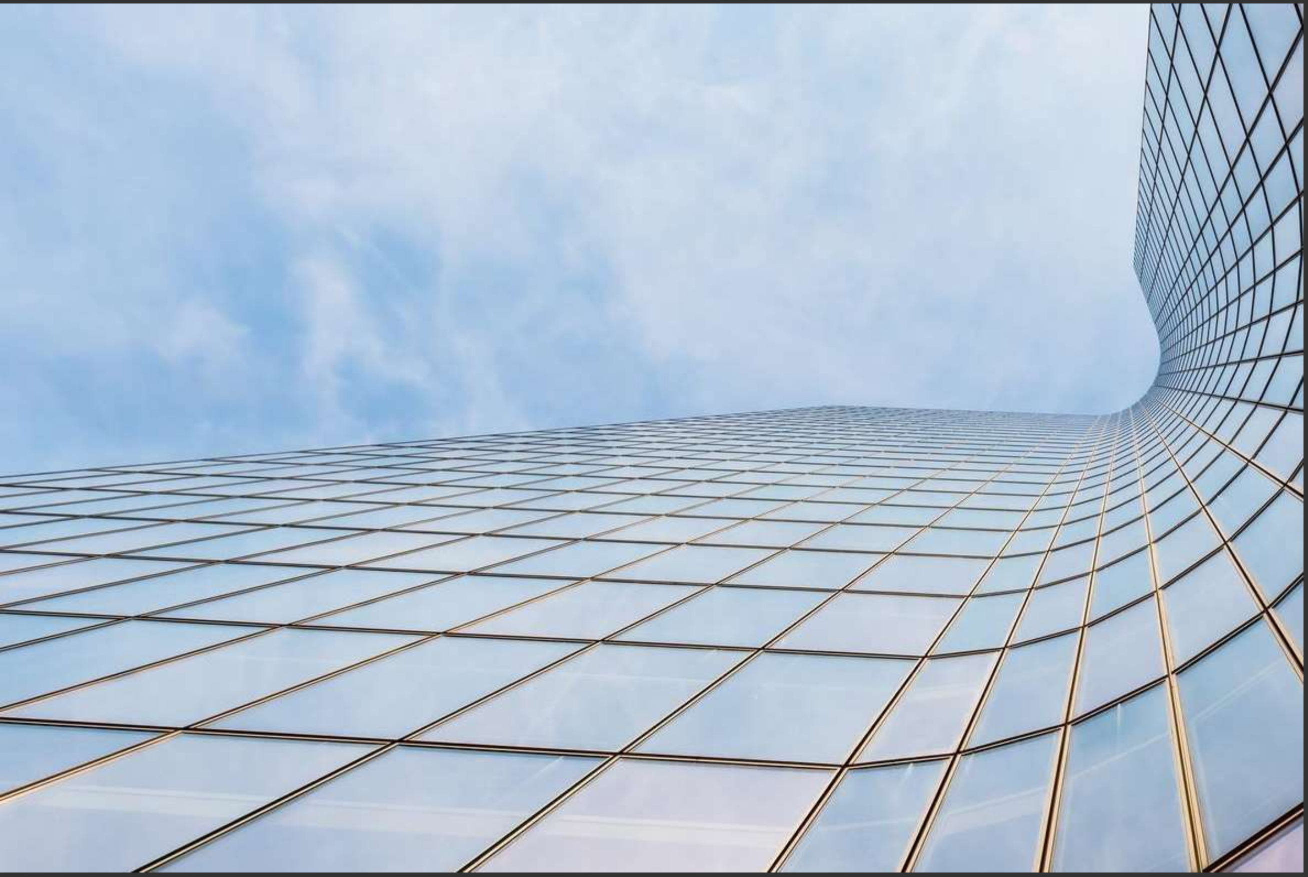
Application data

Univariate Analysis on numerical variables

- Majority of applicants annuity amount is below 50000
- Most of the applicants are in the low-income range.





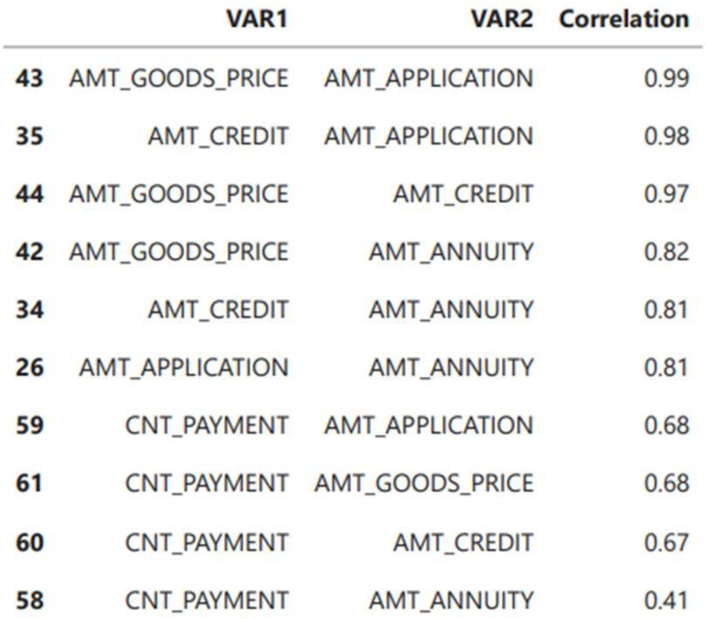


Bivariate analysis

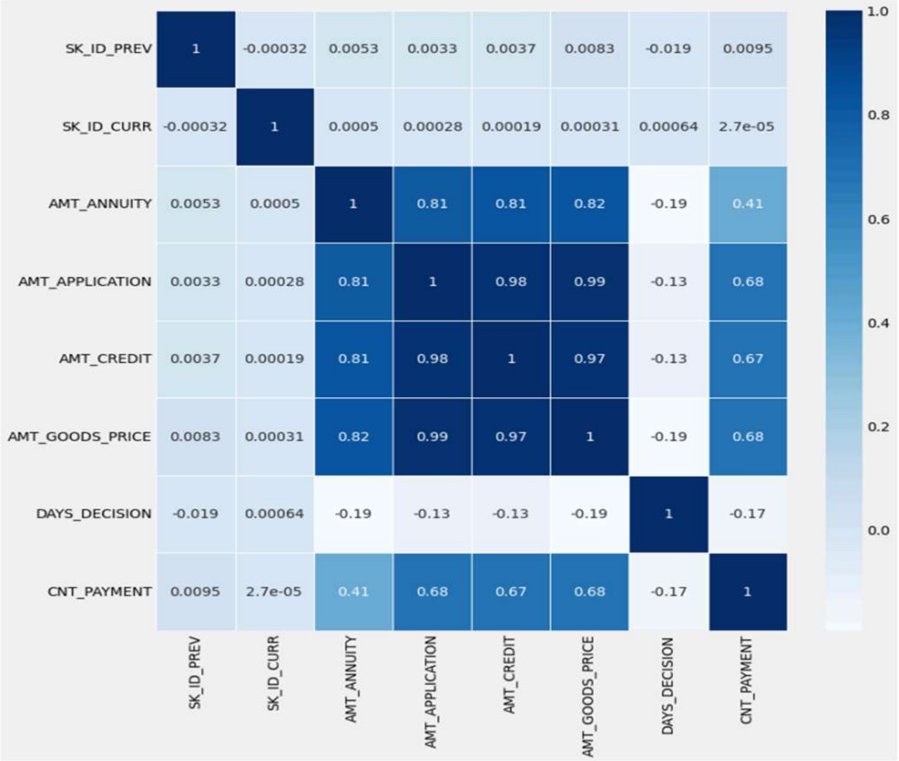
Bivariate analysis

Previous application data

- **Correlation**
  Top 10 correlation variables



	VAR1	VAR2	Correlation
43	AMT_GOODS_PRICE	AMT_APPLICATION	0.99
35	AMT_CREDIT	AMT_APPLICATION	0.98
44	AMT_GOODS_PRICE	AMT_CREDIT	0.97
42	AMT_GOODS_PRICE	AMT_ANNUITY	0.82
34	AMT_CREDIT	AMT_ANNUITY	0.81
26	AMT_APPLICATION	AMT_ANNUITY	0.81
59	CNT_PAYMENT	AMT_APPLICATION	0.68
61	CNT_PAYMENT	AMT_GOODS_PRICE	0.68
60	CNT_PAYMENT	AMT_CREDIT	0.67
58	CNT_PAYMENT	AMT_ANNUITY	0.41

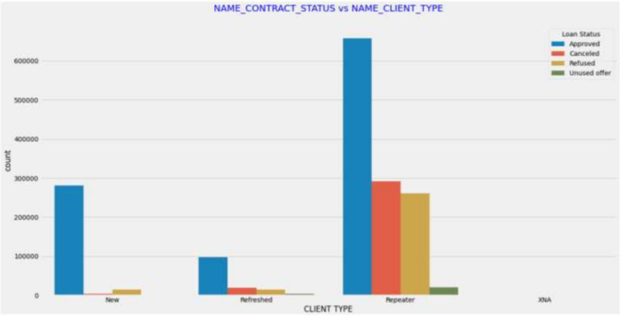


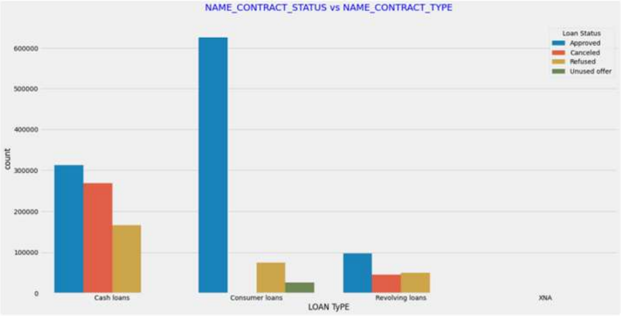
	SK_ID_PREV	SK_ID_CURR	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_GOODS_PRICE	DAYS_DECISION	CNT_PAYMENT
SK_ID_PREV	1	-0.00032	0.0053	0.0033	0.0037	0.0083	-0.019	0.0095
SK_ID_CURR	-0.00032	1	0.0005	0.00028	0.00019	0.00031	0.00064	2.7e-05
AMT_ANNUITY	0.0053	0.0005	1	0.81	0.81	0.82	-0.19	0.41
AMT_APPLICATION	0.0033	0.00028	0.81	1	0.98	0.99	-0.13	0.68
AMT_CREDIT	0.0037	0.00019	0.81	0.98	1	0.97	-0.13	0.67
AMT_GOODS_PRICE	0.0083	0.00031	0.82	0.99	0.97	1	-0.19	0.68
DAYS_DECISION	-0.019	0.00064	-0.19	-0.13	-0.13	-0.19	1	-0.17
CNT_PAYMENT	0.0095	2.7e-05	0.41	0.68	0.67	0.68	-0.17	1

Bivariate analysis

Previous application data

- **NAME_CONTRACT_STATUS vs NAME_CONTRACT_TYPE**
  Consumer loans have the highest approval rate followed by cash loans and then revolving loans. Majority of the loans which were refused are cash loans.
- **NAME_CONTRACT_STATUS vs NAME_CLIENT_TYPE**
  Majority of the applications which were approved are from repeating applicants, followed by new applicants.

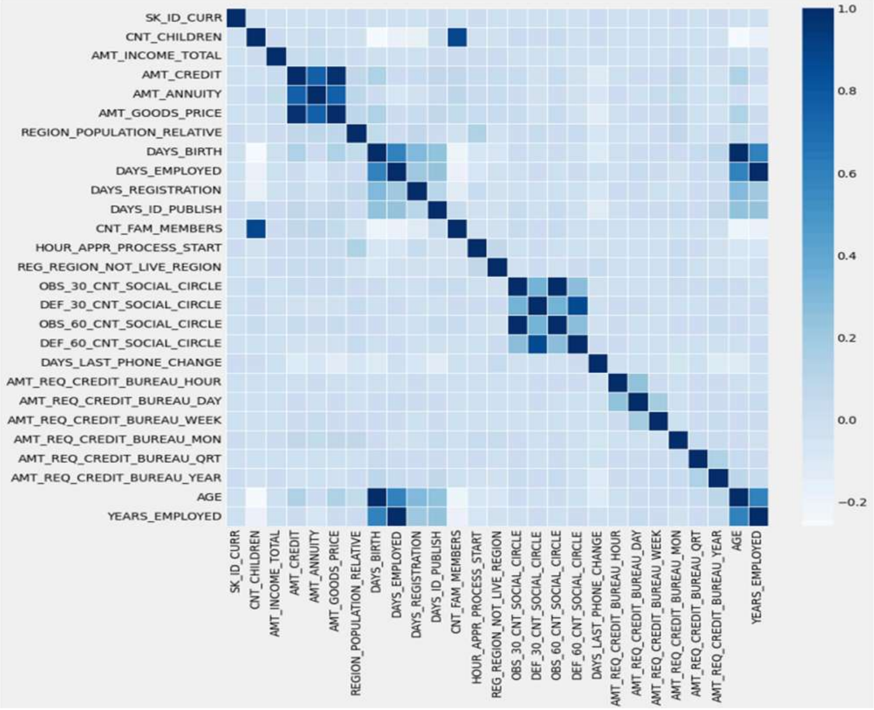




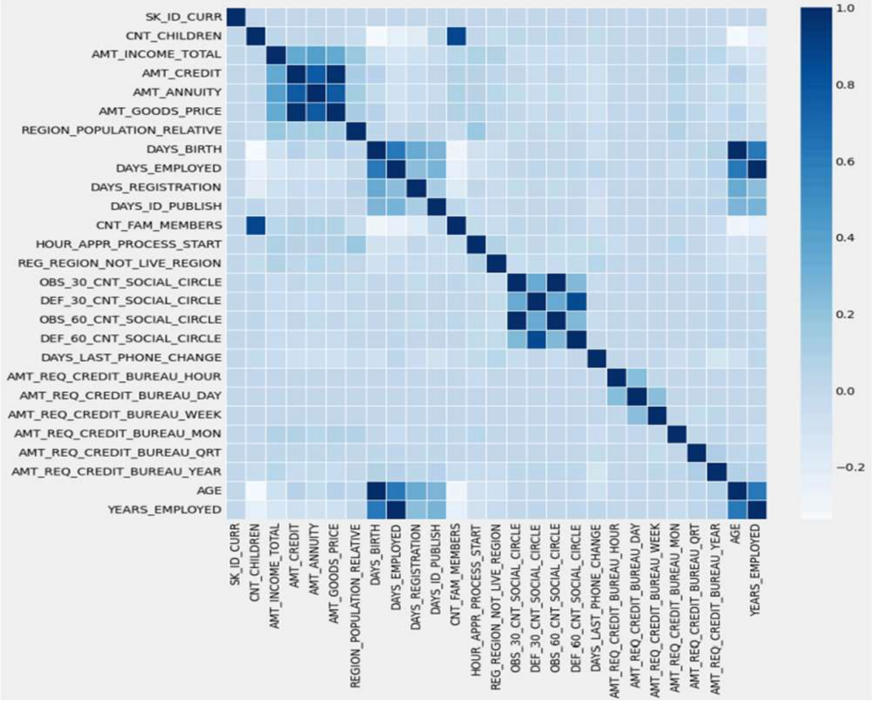
Bivariate analysis

Application data

- top 10 correlation variables for Defaulters



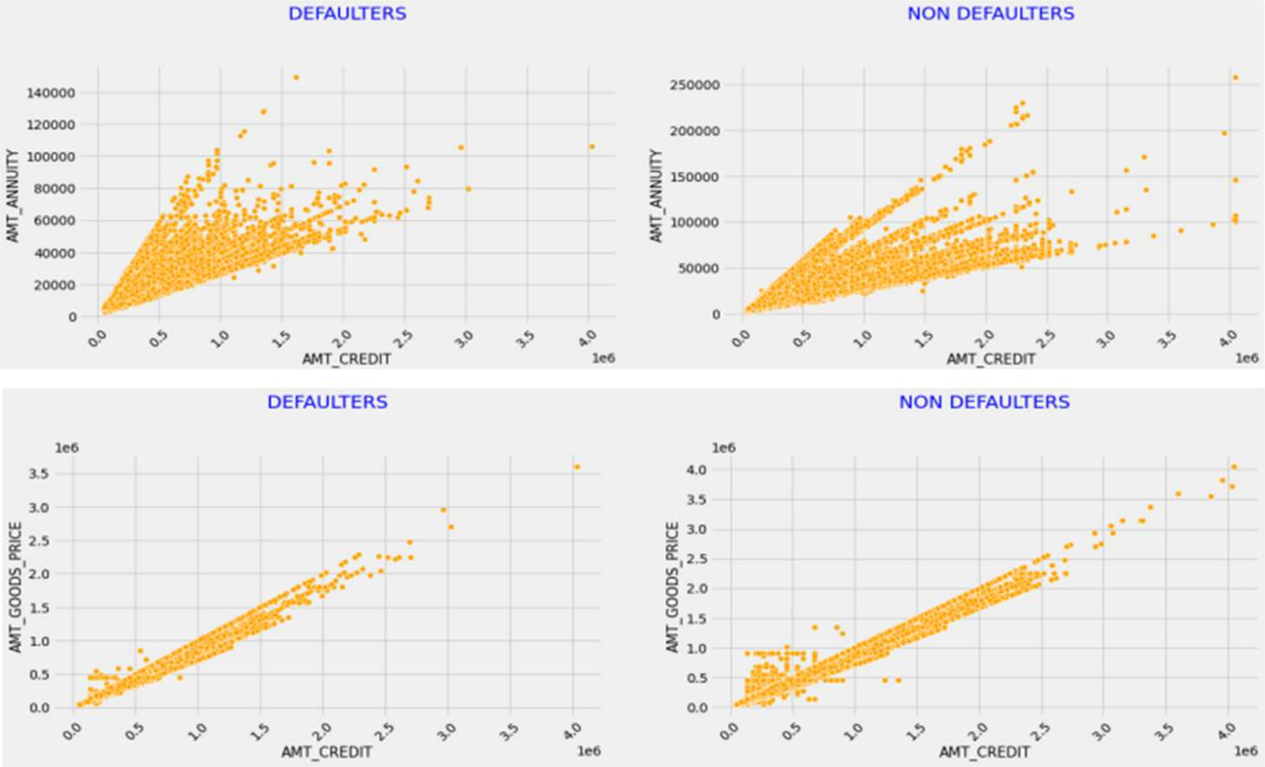
- top 10 correlation variables for Non defaulters



Bivariate analysis

Application data

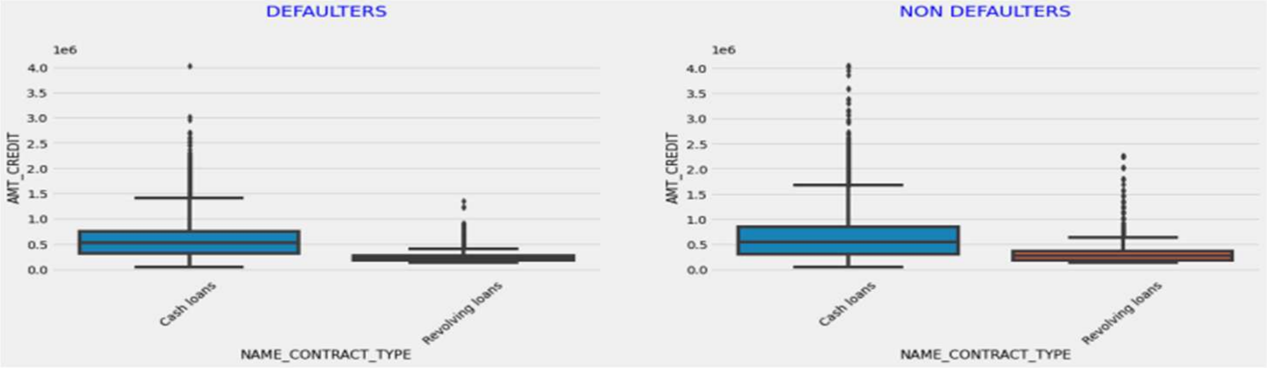
- **AMT_CREDIT VS AMT_ANNUITY**
  Higher the loan amount credited, higher will be the annuity. Therefore AMT_CREDIT and AMT_ANNUITY are positively correlated.
- **AMT_CREDIT VS AMT_GOODS_PRICE**
  Higher the price of the goods, higher will be the loan amount. Therefore AMT_CREDIT and AMT_GOODS_PRICE are positively correlated.

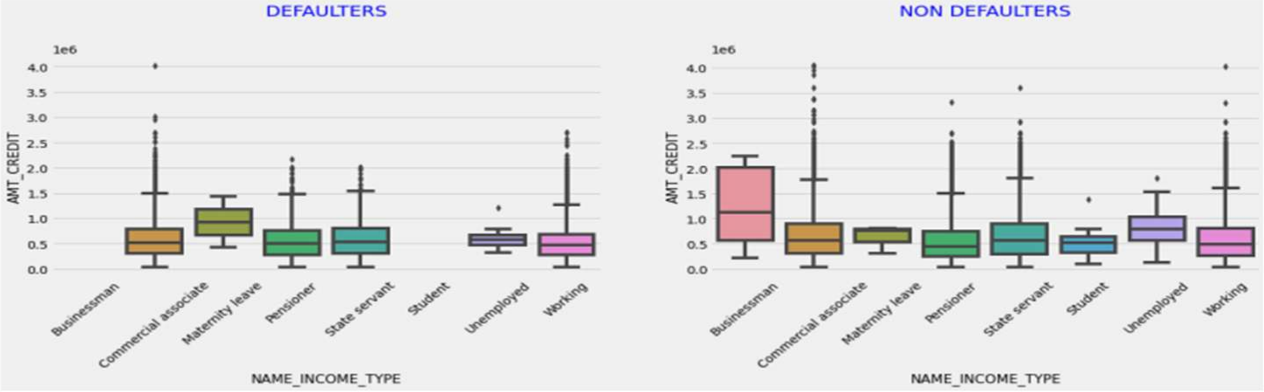


Bivariate analysis

Application data

- **AMT_CREDIT VS NAME_CONTRACT_TYPE** Cash loans have higher amount credited when compared to revolving loans.
- **AMT_CREDIT VS NAME_INCOME_TYPE** Businessman have taken higher loan amounts and they do not seem to default the loan along with students.

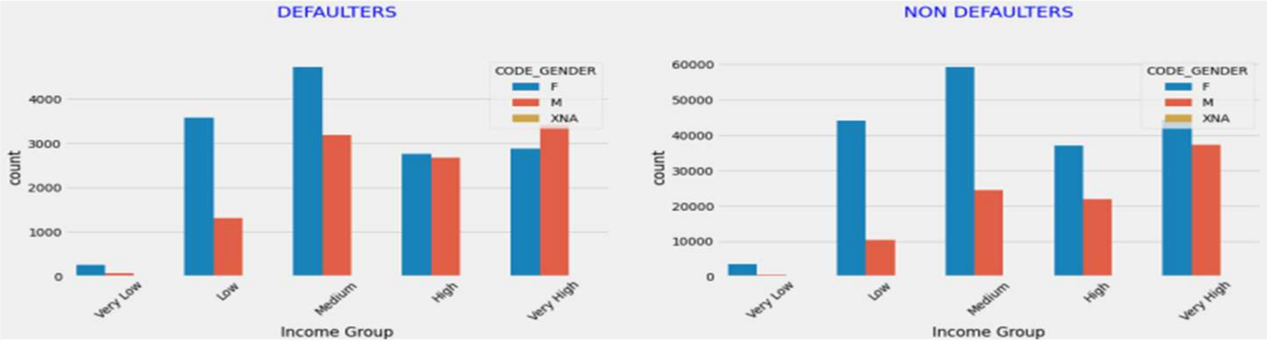


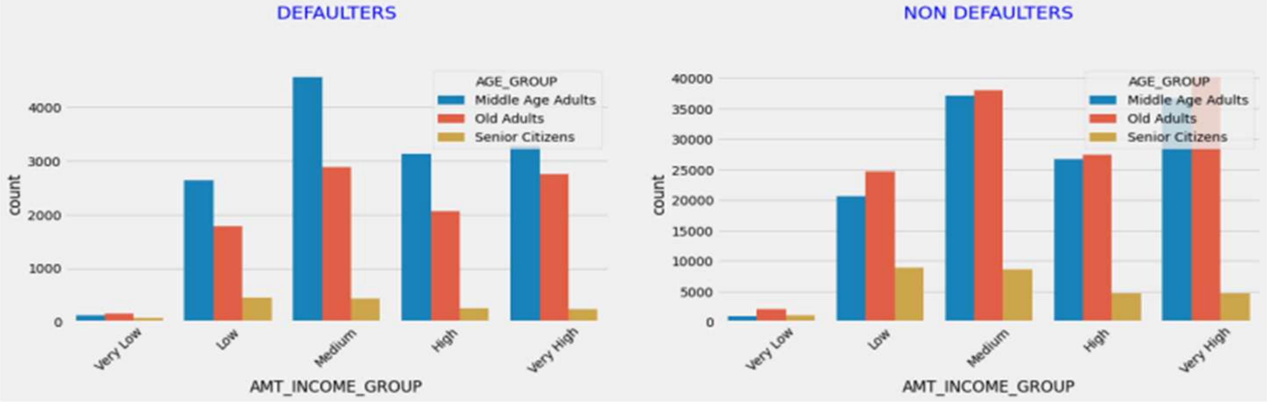


Bivariate analysis

Application data

- **NAME_CONTRACT_TYPE vs CODE_GENDER** Majority of the applicants are female and have taken cash loans.
- **AMT_INCOME_GROUP vs CODE_GENDER** Male applicants with very high income have defaulted more than female
- **AMT_INCOME_GROUP vs AGE_GROUP** Middle-aged adults belonging to medium income group is most defaulted.



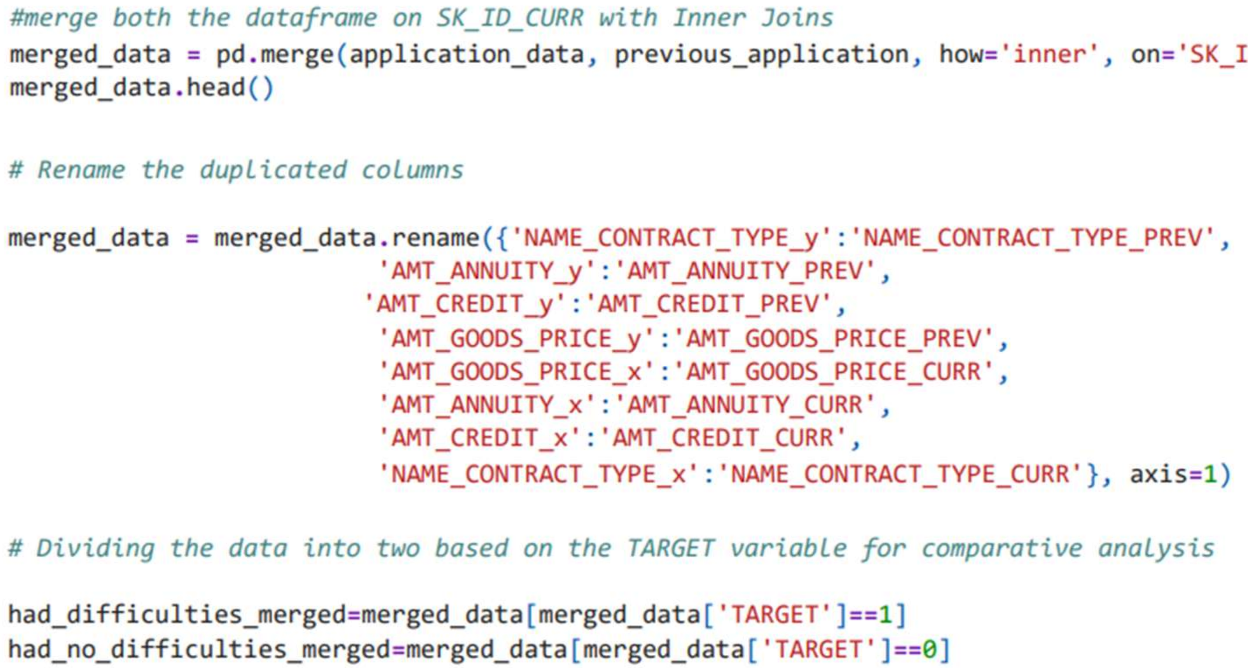


Bivariate analysis

Merged data

The process of merging data included:

- Merging and previous_application and application_data dataframes on SK_ID_CURR with Inner Joins
- Renaming the duplicated columns
- Dividing the data into two based on the TARGET variable for comparative analysis

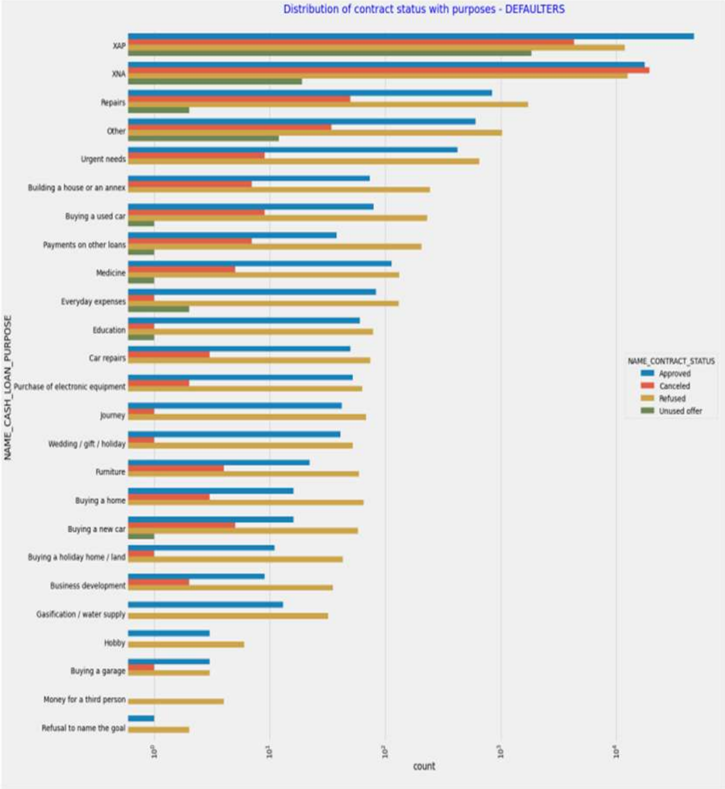


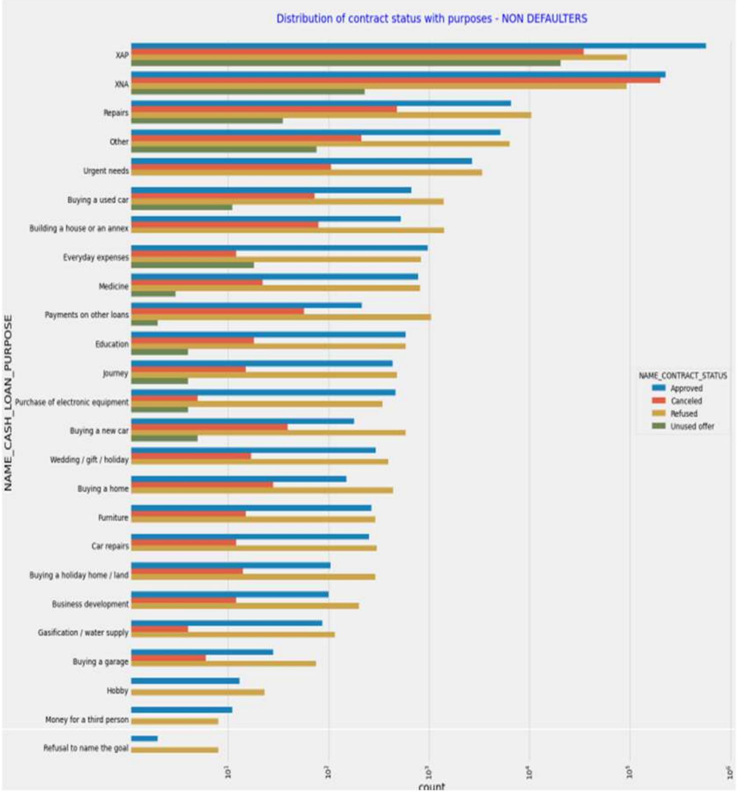
```
#merge both the dataframe on SK_ID_CURR with Inner Joins
merged_data = pd.merge(application_data, previous_application, how='inner', on='SK_I
merged_data.head()

# Rename the duplicated columns

merged_data = merged_data.rename({'NAME_CONTRACT_TYPE_y':'NAME_CONTRACT_TYPE_PREV',
                         'AMT_ANNUITY_y':'AMT_ANNUITY_PREV',
                        'AMT_CREDIT_y':'AMT_CREDIT_PREV',
                         'AMT_GOODS_PRICE_y':'AMT_GOODS_PRICE_PREV',
                         'AMT_GOODS_PRICE_x':'AMT_GOODS_PRICE_CURR',
                         'AMT_ANNUITY_x':'AMT_ANNUITY_CURR',
                         'AMT_CREDIT_x':'AMT_CREDIT_CURR',
                         'NAME_CONTRACT_TYPE_x':'NAME_CONTRACT_TYPE_CURR'}, axis=1)

# Dividing the data into two based on the TARGET variable for comparative analysis

had_difficulties_merged=merged_data[merged_data['TARGET']==1]
had_no_difficulties_merged=merged_data[merged_data['TARGET']==0]
```

Bivariate analysis

Merged data

- **NAME_CASH_LOAN_PURPOSE and NAME_CONTRACT_STATUS with respect to TARGET** Loan purpose has high number of unknown values (XAP, XNA). Cash loans for repairs and urgent needs are more likely to default





Conclusion

- **Loan Type and Approval Rates:** focus on assessing risk associated with cash loans, which have a higher likelihood of refusal compared to consumer loans.
- **Applicant Demographics:** middle-aged adults are more likely to default, while senior citizens default less. Consider age as a risk factor. Male applicants with very high-income default more than females and tailor risk assessment accordingly.
- **Income and Credit Amount:** medium income earners are prone to default, particularly those with credit amounts ranging from 0-5 lakhs.
- **Loan Purpose and Amount:** repairs and urgent needs loans have a higher default rate. Adjust loan terms or offer financial education for better management.
- **Correlations and Financial Indicators:** use the positive correlation between loan amounts, annuity, and income levels to assess loan affordability accurately.
- **Experience and Organization Type:** applicants with less experience and certain organization types are more likely to default. Incorporate this into risk assessment models.
- **Continuous Improvement:** continuously update risk assessment strategies based on evolving market conditions and borrower behavior.

By applying these insights, the company can minimize defaults and optimize its loan portfolio

THANK YOU