DS 41
Credit EDA Case Study - Bank Loan

- Venkata Kailash Chollangi

Problem Statement:

This case study 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.

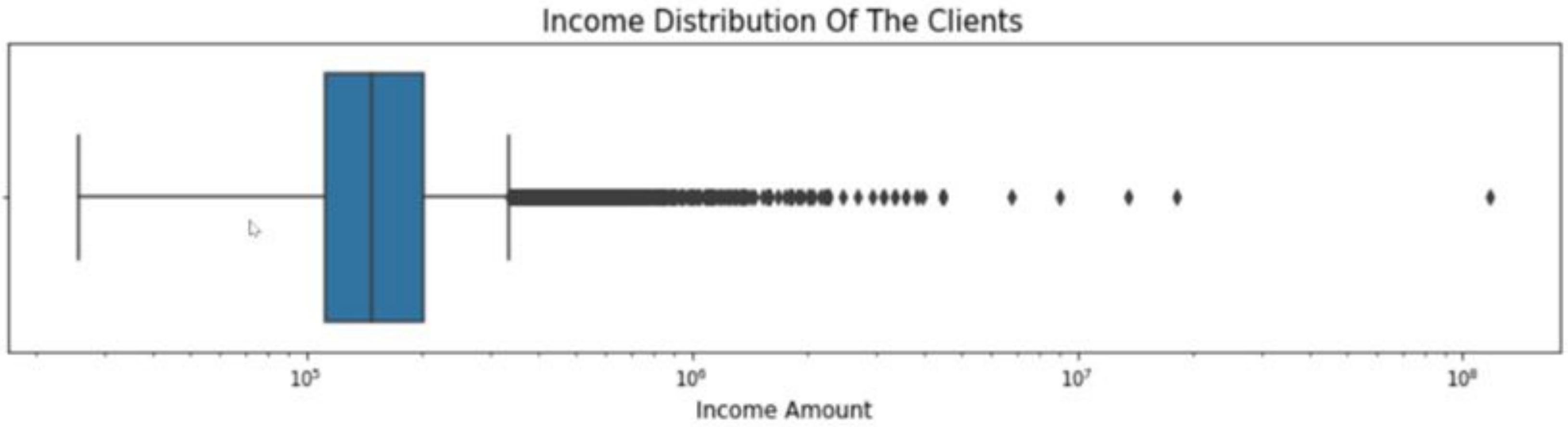
Analysis:

- Data Quality check and missing values
- Checking Imbalance percentage
- Analysis done with respect to Target Variables as
  Target 1: Customer with payment difficulties in the past
  Target 0: Customers who have paid on time

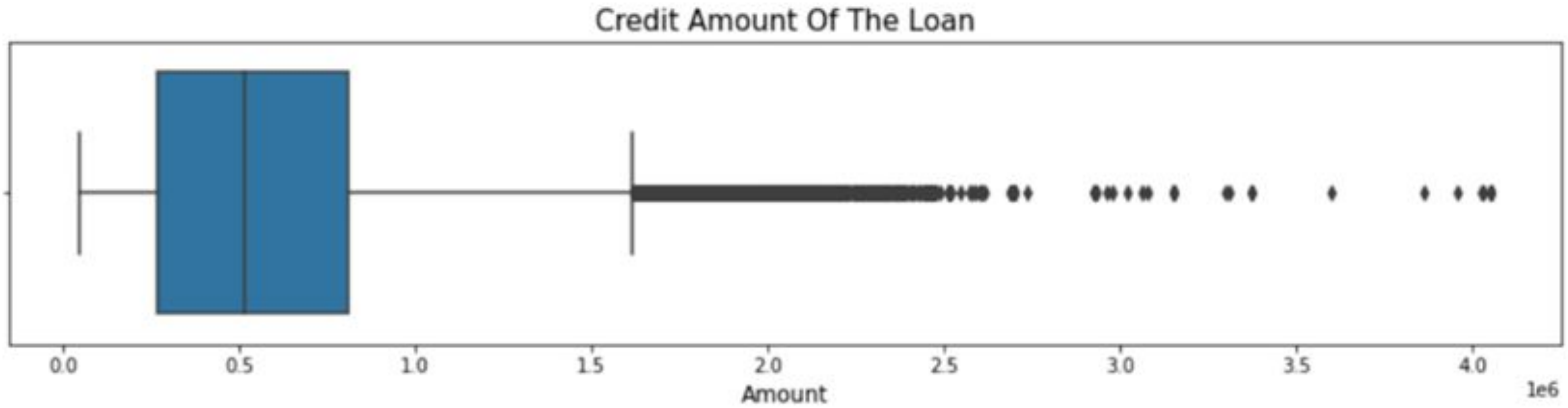
Types of Analysis

- Univariate on Application Data and Previous Application Data
- Bivariate on Application Data and Previous Application Data
- Multivariate analysis

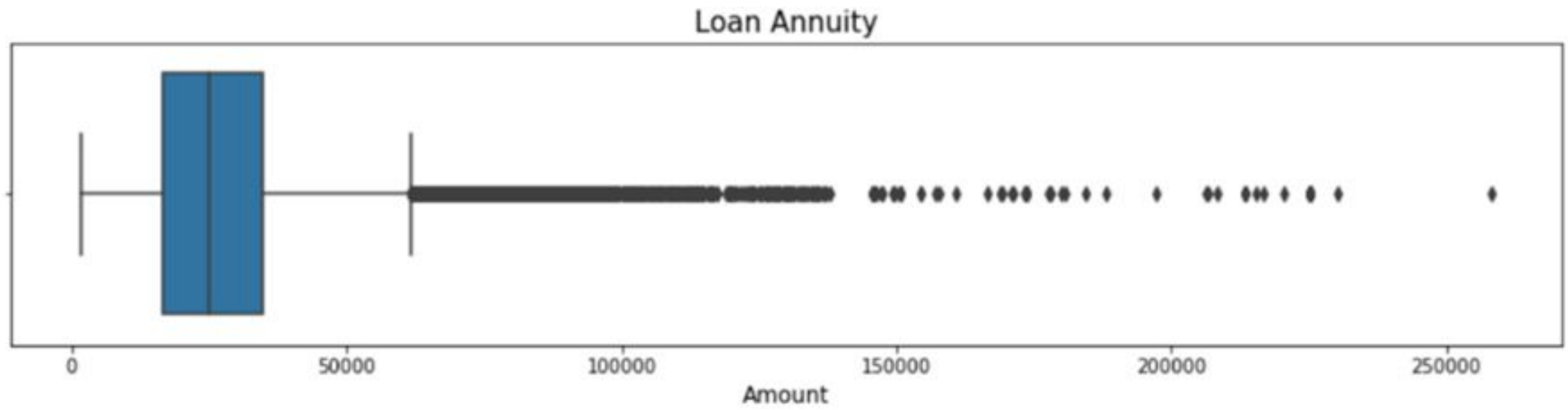
Univariate Analysis on Application Data



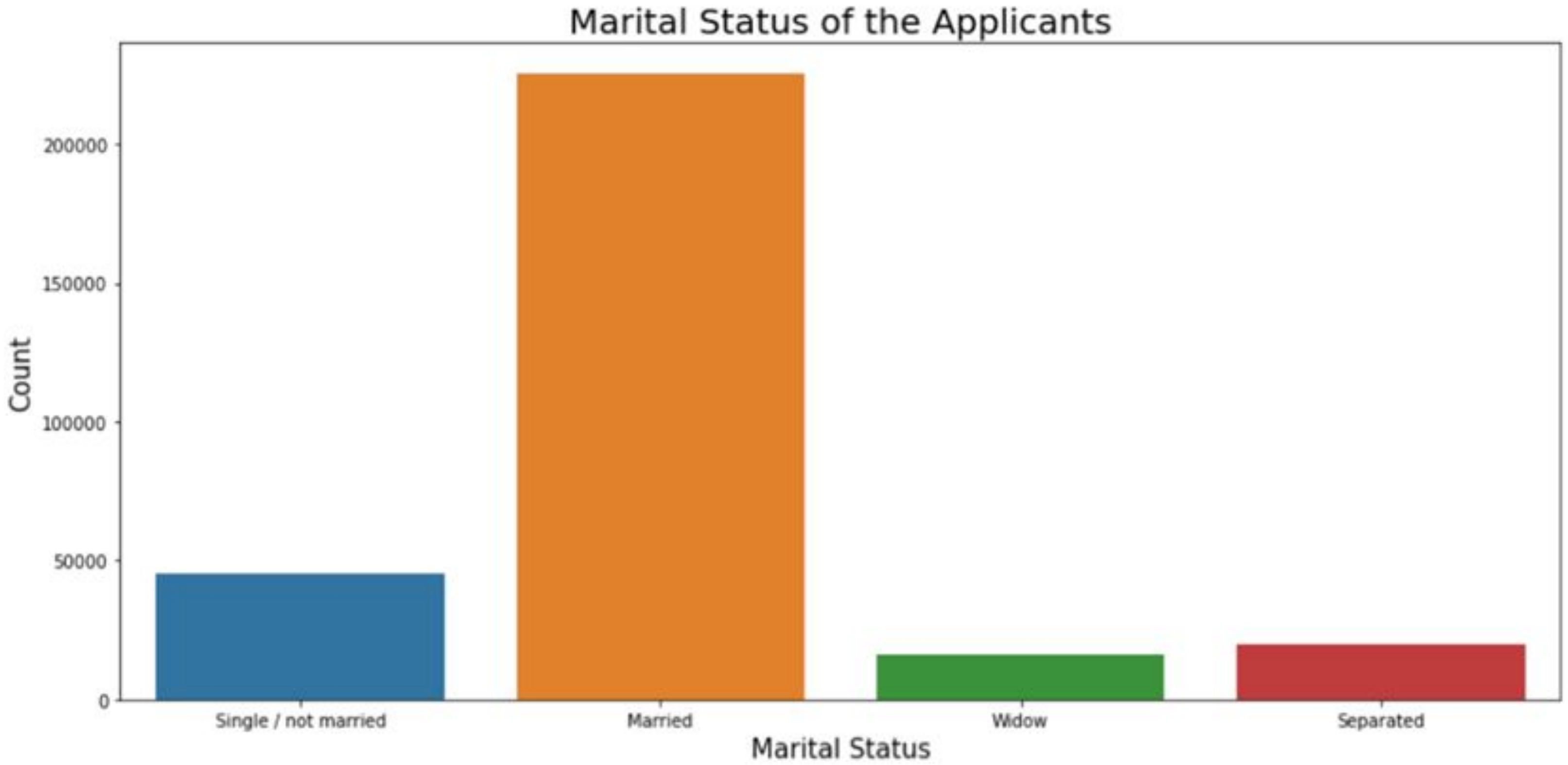
- The income distribution of clients has some extreme outlier values beyond 99th percentile



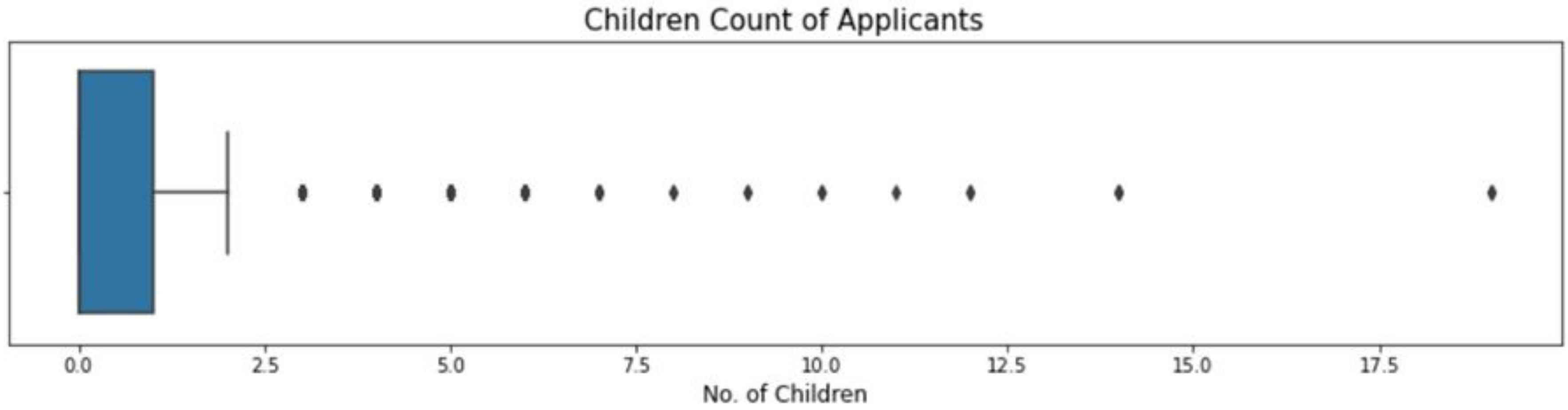
- The median almost divides the IQR equally, but the max whisker is much longer than the min whisker.
- There are many outliers in credit amount as per boxplot, most significant ones being above 2.5.



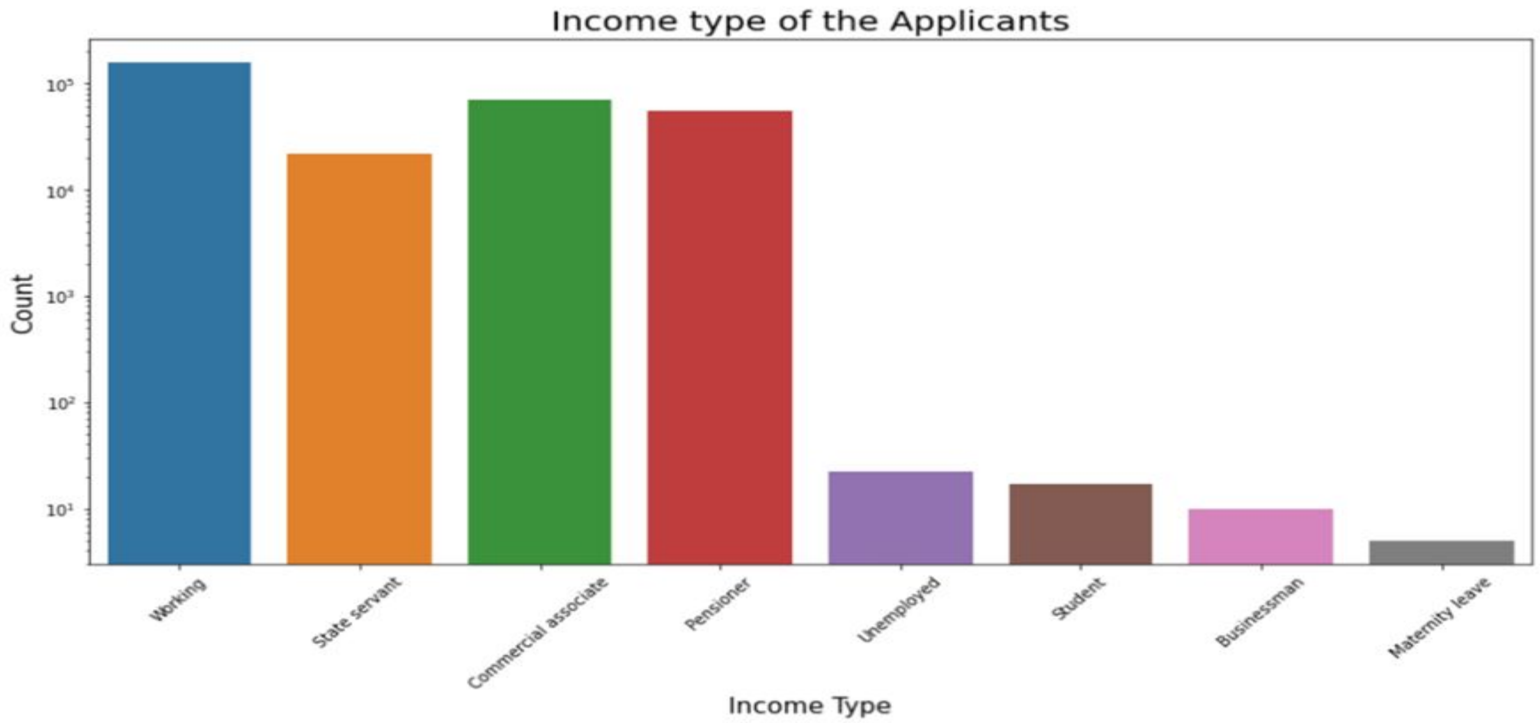
- The median almost divides the IQR equally , but the max whisker is notably higher than the minimum whisker
- There are many outliers in the loan annuity as per boxplot



- Maximum loan application come from married people
- Minimal loan application comes from widows

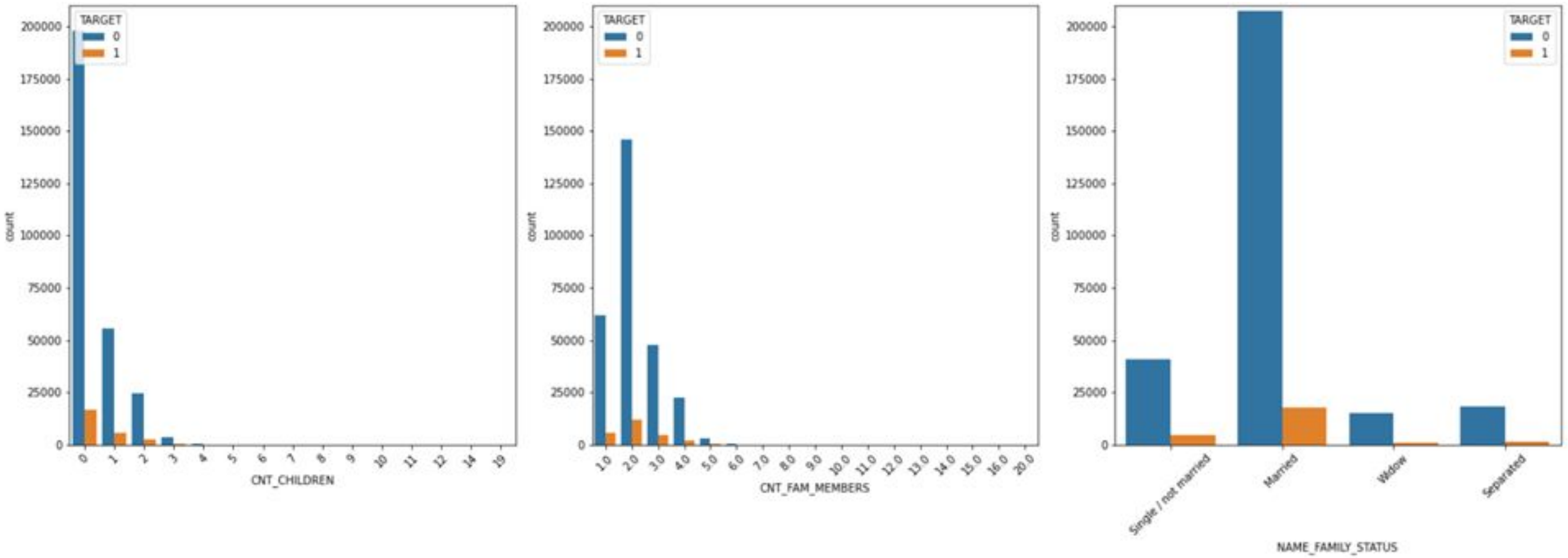


- Out of 306207 only 553 applicants has more than 3 kids which we can consider as our outlier of number of kids.



Inference: Most application we got from 3 income type Working, State Servant and Commercial associate.

- Most application we got from 3 income type Working, State Servant and Commercial associate.



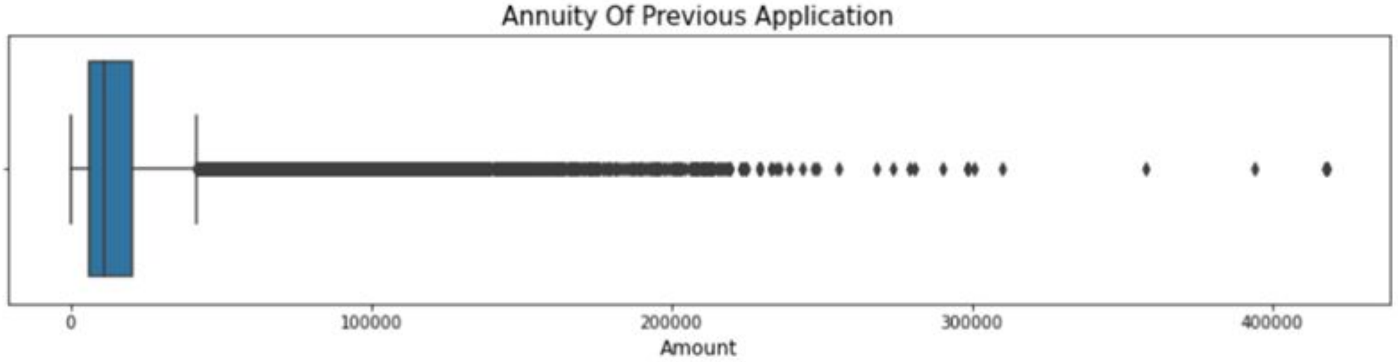
- Most of the defaulters are from Married and Single category

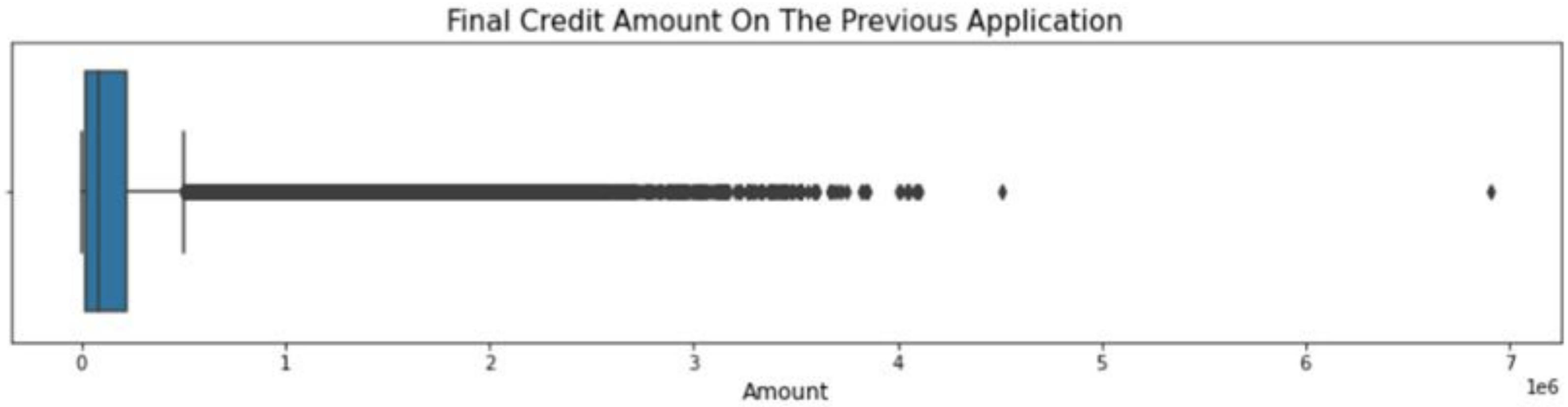
Checking outliers and perfoming univariate analysis of Previous application dataframe

Checking outliers and perfoming univariate analysis of Previous application dataframe

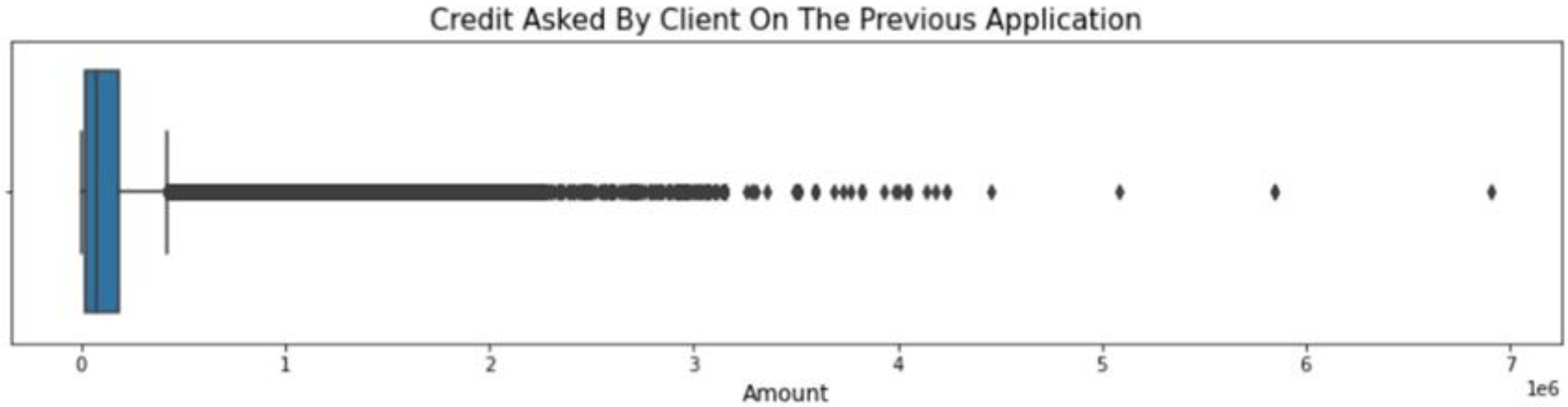
```
# plot the boxplot of Amount Annuity

plt.figure(figsize=[15,3])
sns.boxplot(prevdata.AMT_ANNUITY)
plt.title('Annuity Of Previous Application', fontsize=15)
plt.xlabel('Amount', fontsize=12)
plt.show()
```





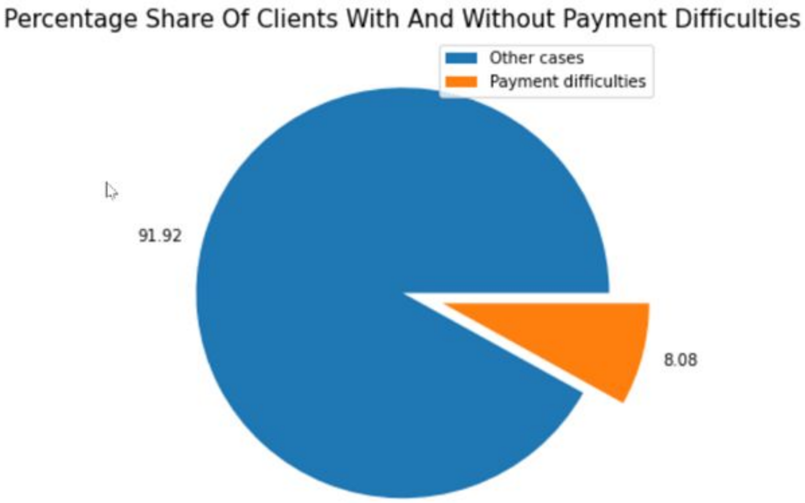
- There are 16703 outliers in Amount Credit as per boxplot, most significant ones being above 4000000.

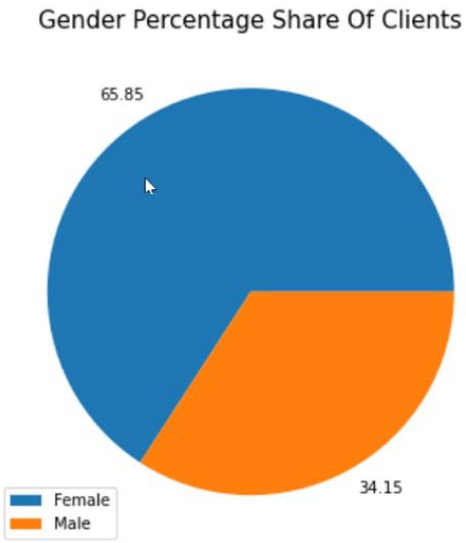


There are 15952 outliers in Amount Application as per boxplot but most significant ones being above 4000000.

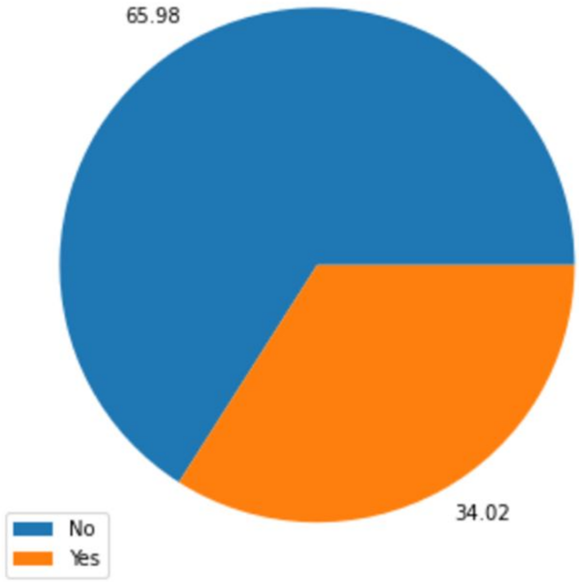
Data imbalance. Finding the ratios of data imbalance.

- **Bivariate and Multivariate Analysis on application_data**

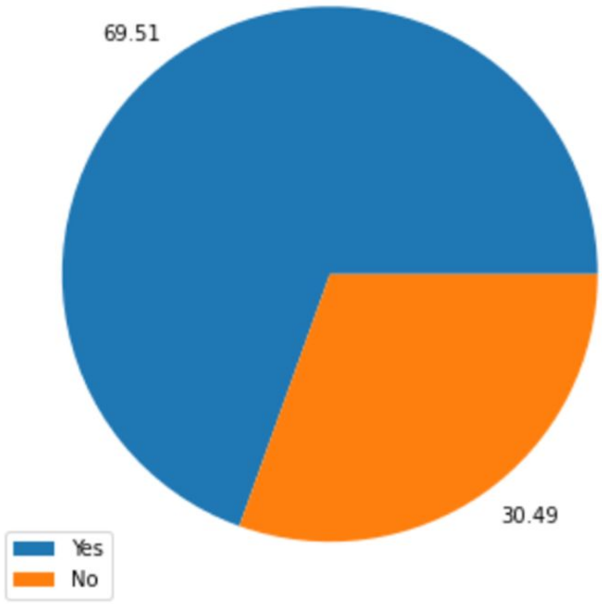


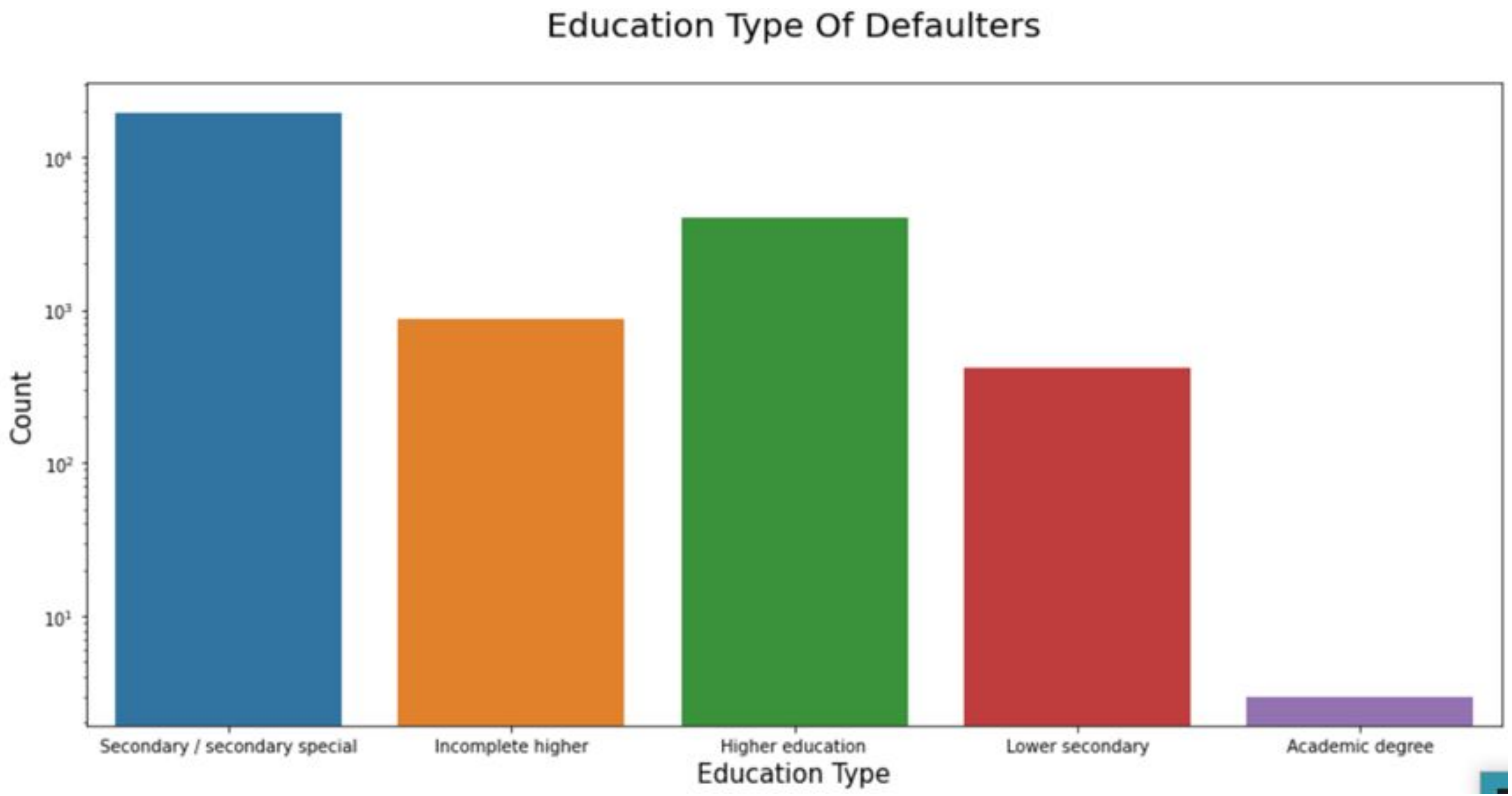


Car Ownership Percentage Share Of Clients

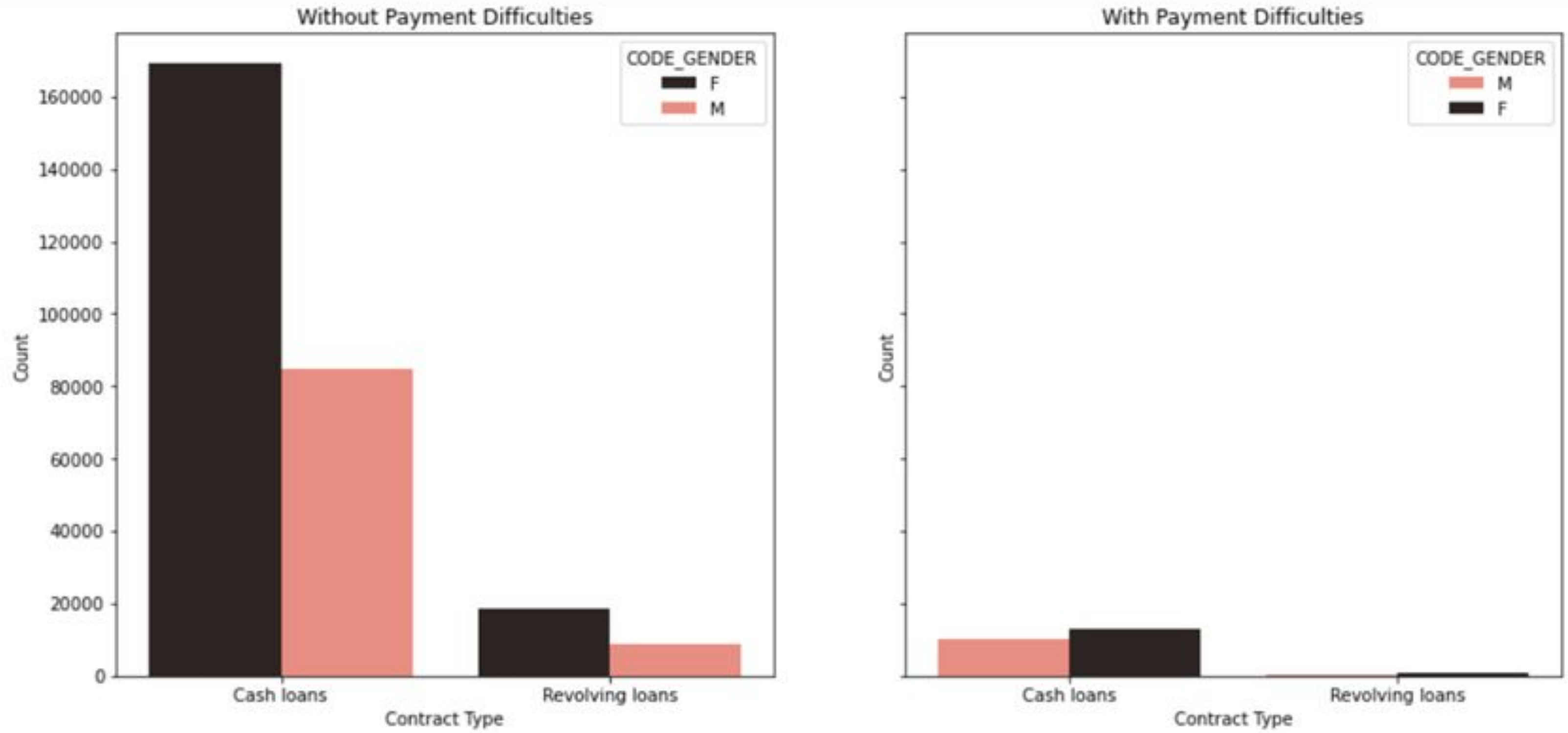


Realty Ownership Share Of Clients

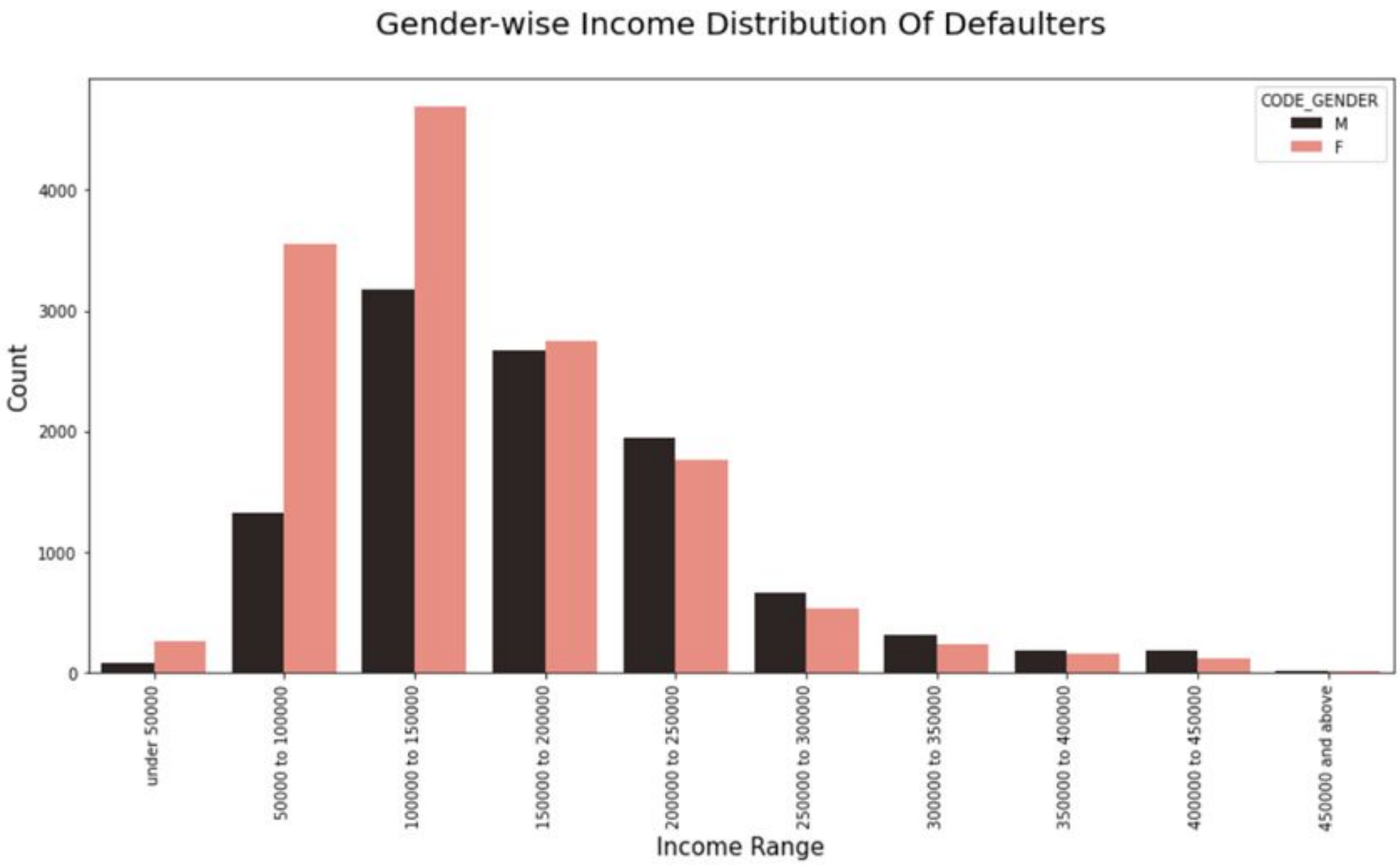




- Most defaulters came from Secondary and Higher education background.
- Least defaulter came from Academic degree background

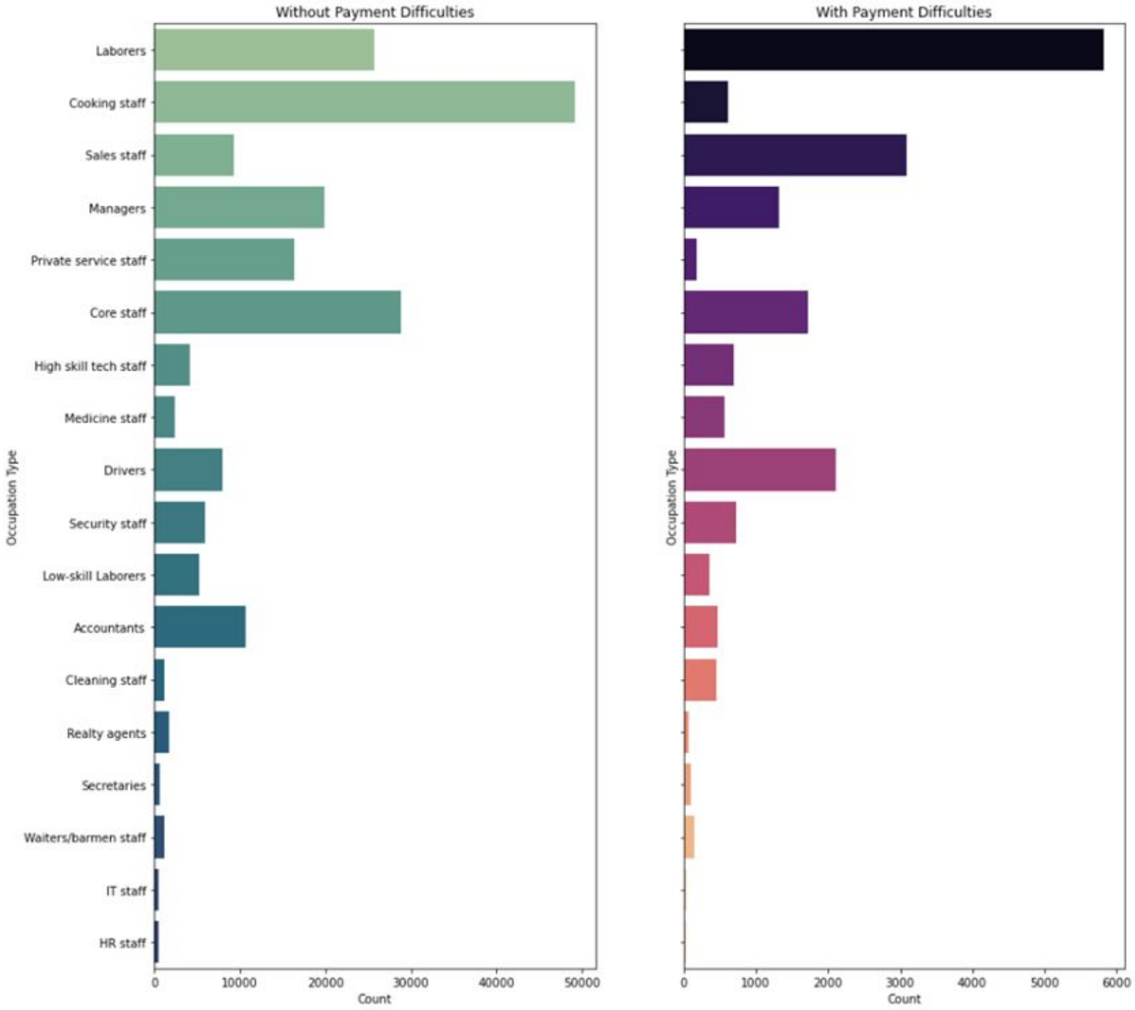


- Demand for cash loans is significantly higher than revolving loans.
- Demand for both types of loan contracts are almost twice that of males. However, the default rate is almost equal.



Gender-wise Income Distribution Of Defaulters
CODE_GENDER
M
F
Count
4000
3000
2000
1000
0
under 500000
50000 to 100000
100000 to 150000
150000 to 200000
200000 to 250000
250000 to 300000
300000 to 350000
350000 to 400000
400000 to 450000
450000 and above
Income Range

- Most defaulters for both Male and Female come from the 100000 to 150000 income range.
- There are more female defaulters in income range 250000 and below.
- There are more male defaulters in income range 250000 and above.

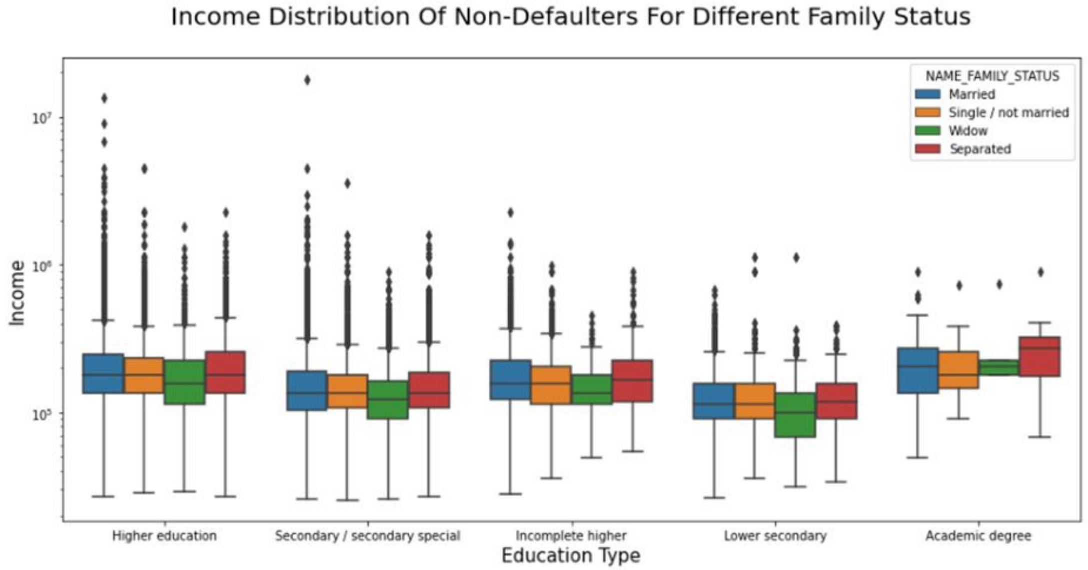


Among different occupations,

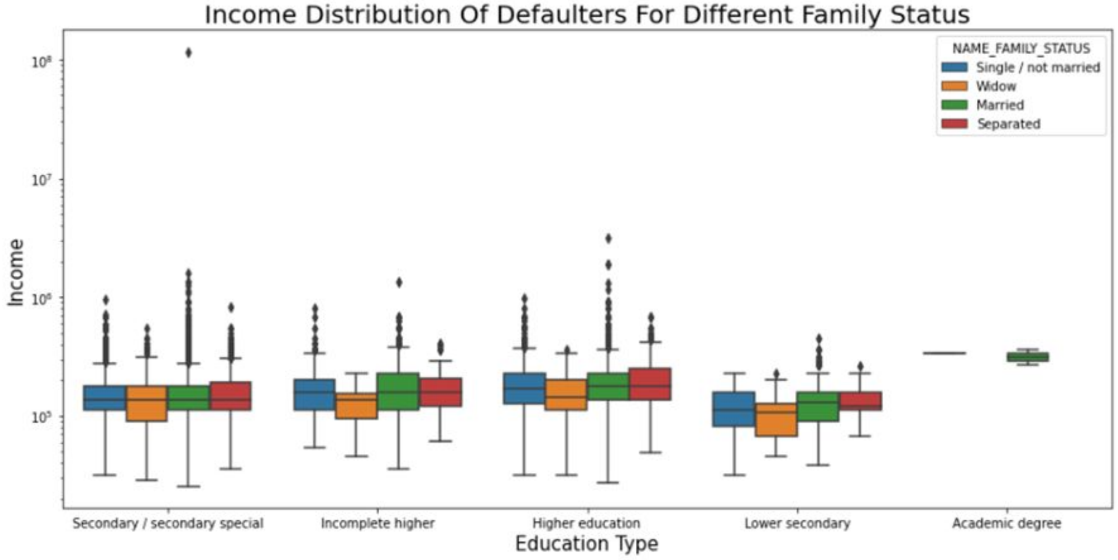
- Cooking staff had the least count of clients with difficulty in payment of loan.
- Labourers had the highest count of clients with difficulty in payment of loan.

Numerical - Categorical Correlation Analysis

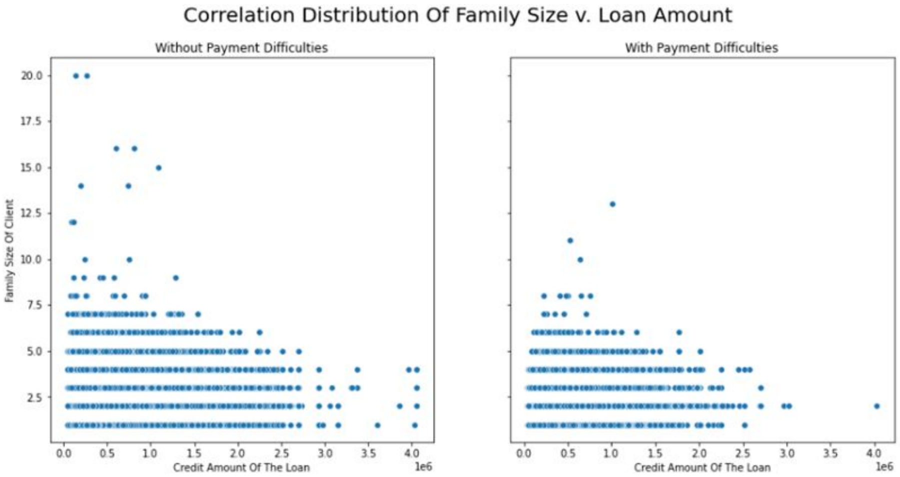
- **Top 10 correlation for the Client with payment difficulties and all other cases (Target variable)**



- Non-married clients with academic degrees have a much higher minimum whisker than all other categories.
- Married clients with higher education or secondary/secondary special education have significant outliers on the higher side.
- There are no lower outliers for any category.

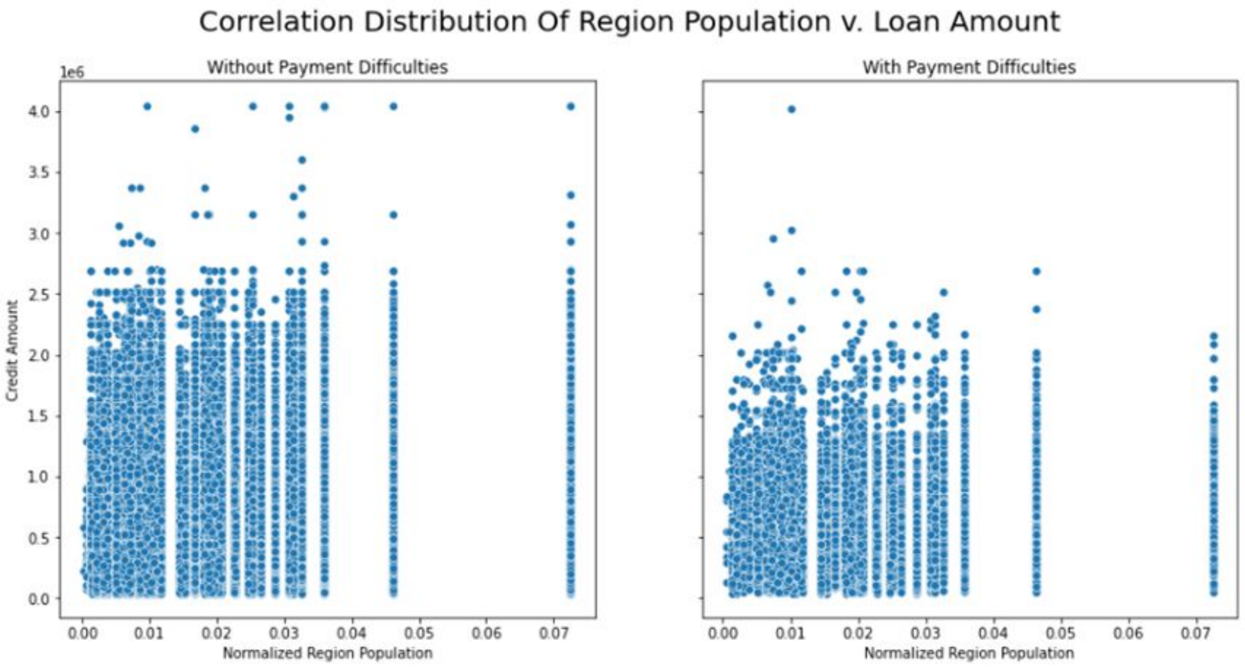


- For majority of defaulting clients, across all education types, the income is comparatively on the lower side compared to non-defaulters.
- Exceptions to this are outliers in married clients with higher education or secondary/secondary, who are defaulting despite higher income.
- Widows and seperated clients with academic degree, appear to be facing the least payment difficulties

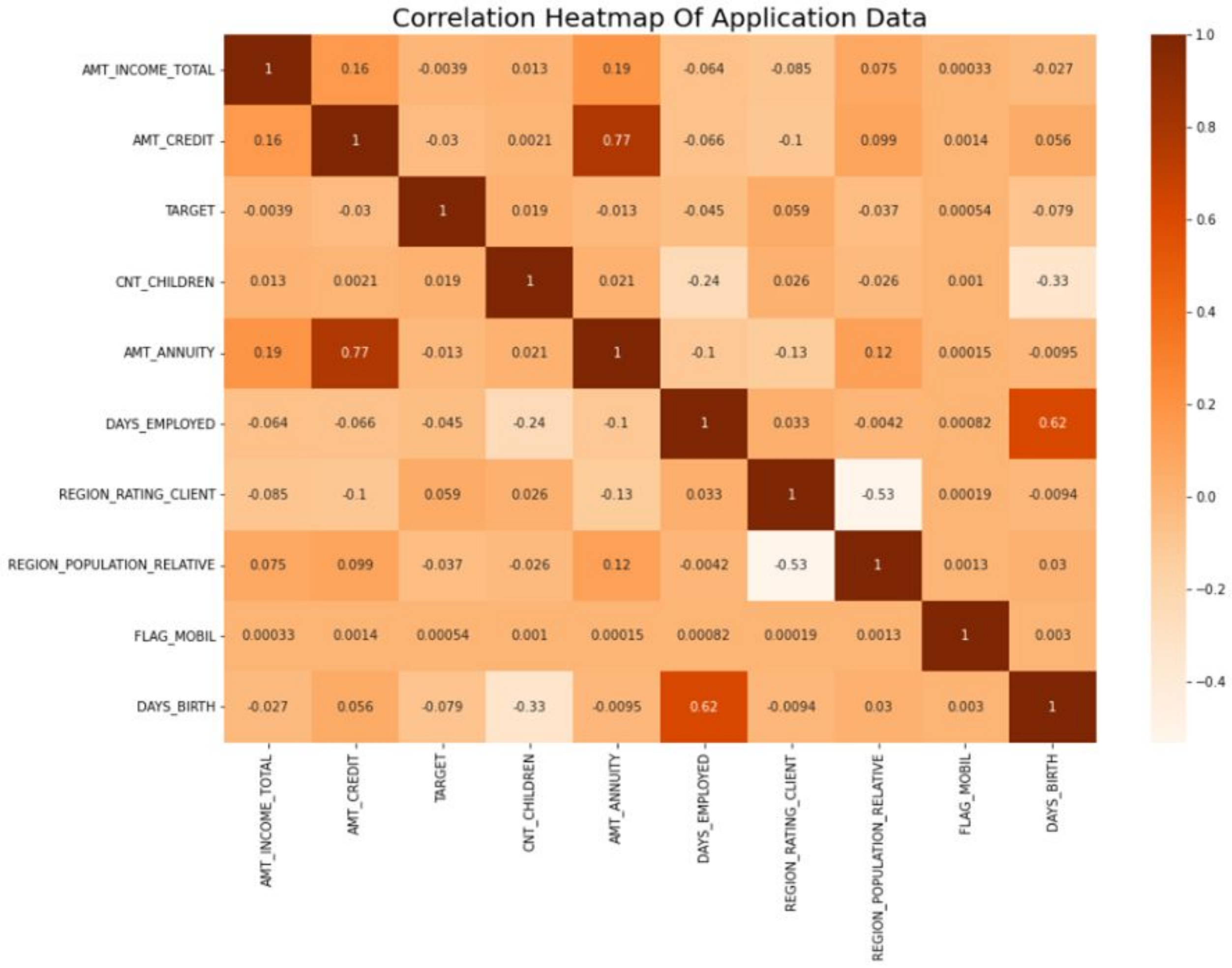


- There is no correlation between family size of client and the credit amount, for both defaulters and non-defaulters.
- Infact, clients with extremely large families haven't had payment difficulties.

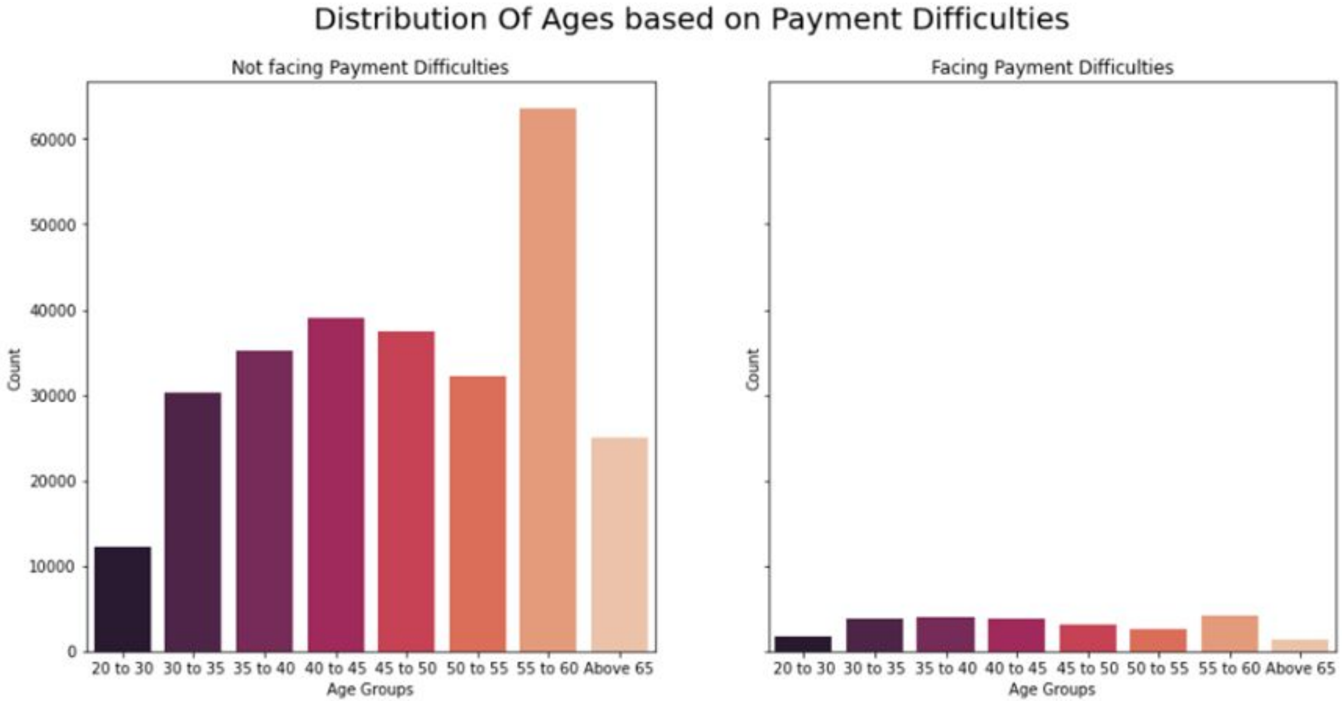
- There is a linear correlation between credit amount of loan and price of goods for which the loan is given.
- This shows that when the price the goods increases, the credit amount of loan also increases.



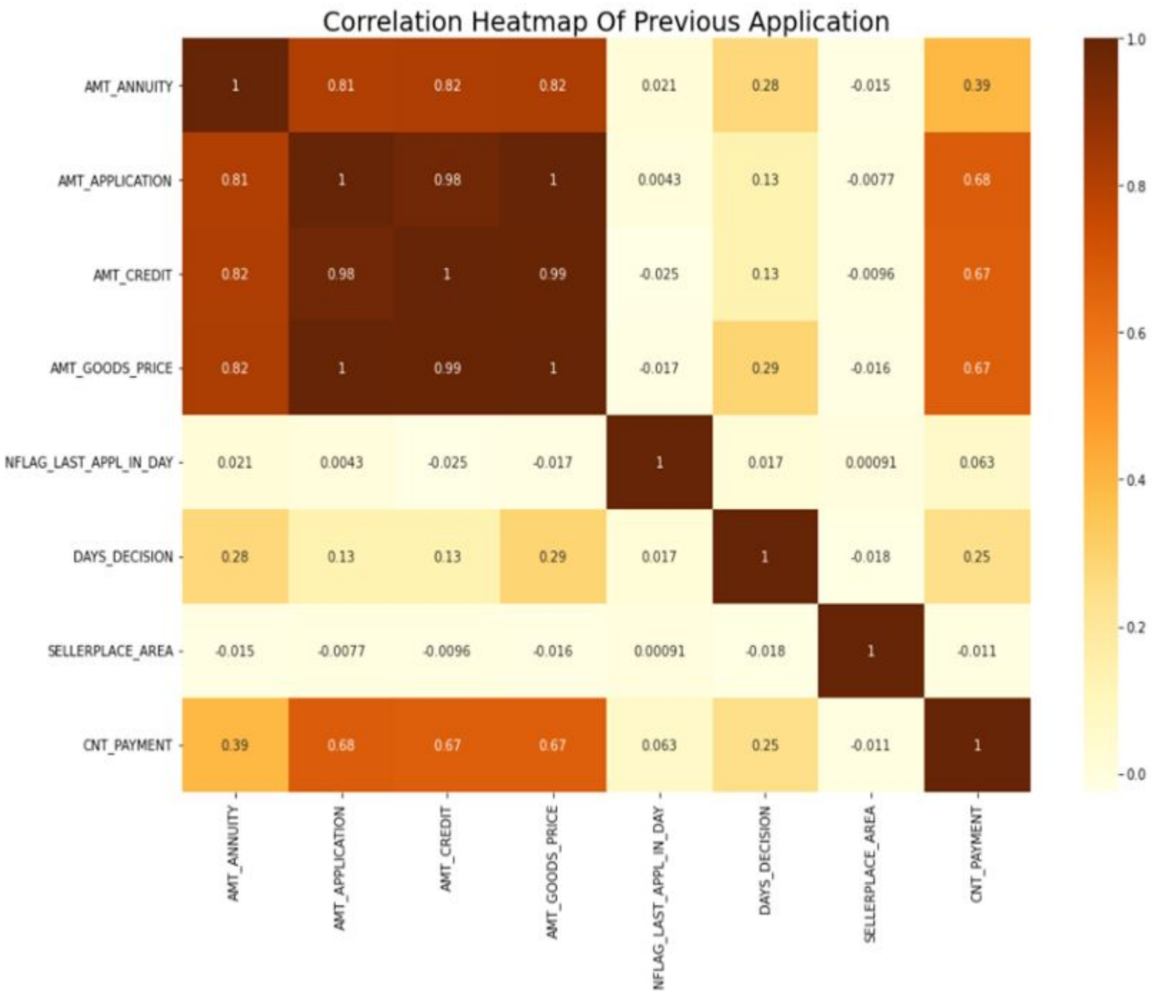
- For clients without payment difficulties, there is no visible correlation between client's region population and credit amount.
- For clients with payment difficulties, clients with higher credit amount and very low region population have noticable correlation outliers, compared to clients in regions with higher population.



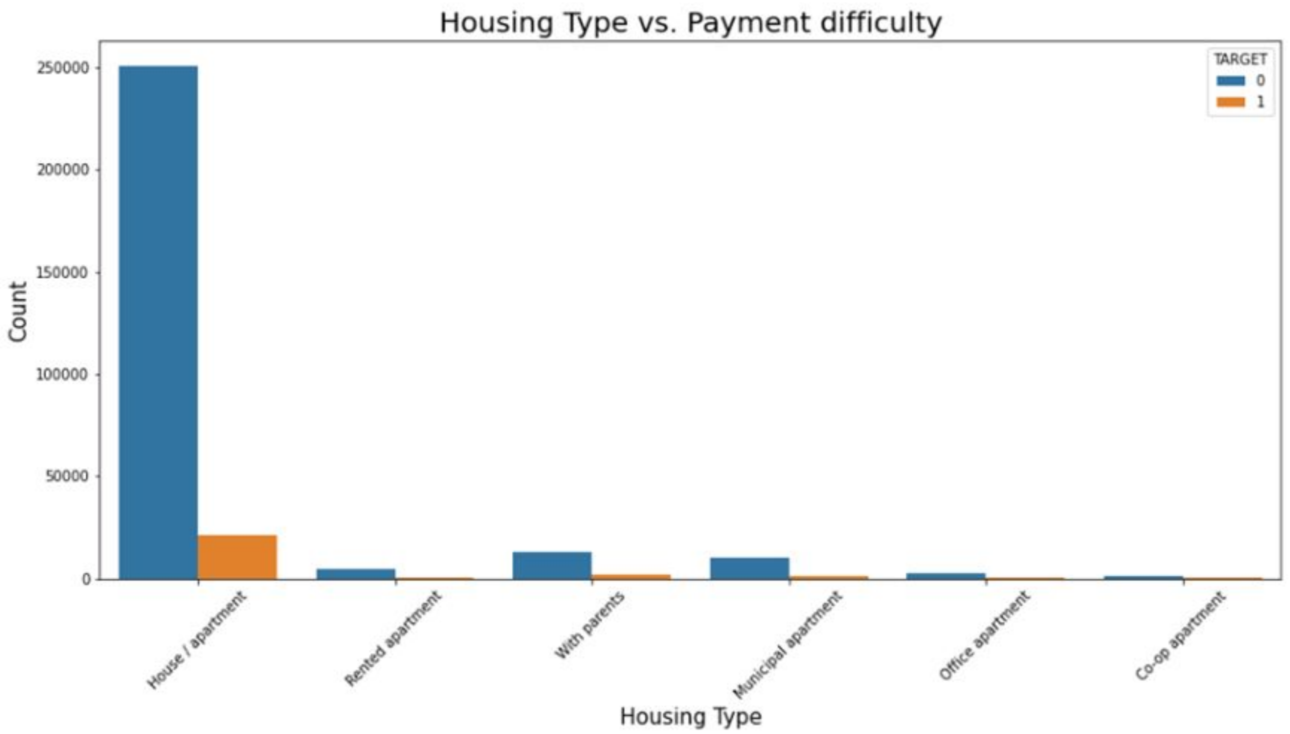
1. High income people take larger credit amount loan along and pay larger loan annuity as well.
2. High population density area pay higher loan annuities.
3. The chances of having payment difficulty is very low with high income people.
4. If target lives in high population area and has high number of kids, there are higher chances of facing payment difficulty.



- While applicants of age group 55 to 60 face most difficulty in payments, they are also the group with least difficulty in payment.
- 20 to 30 group and above 65 group face least difficulty in payment.



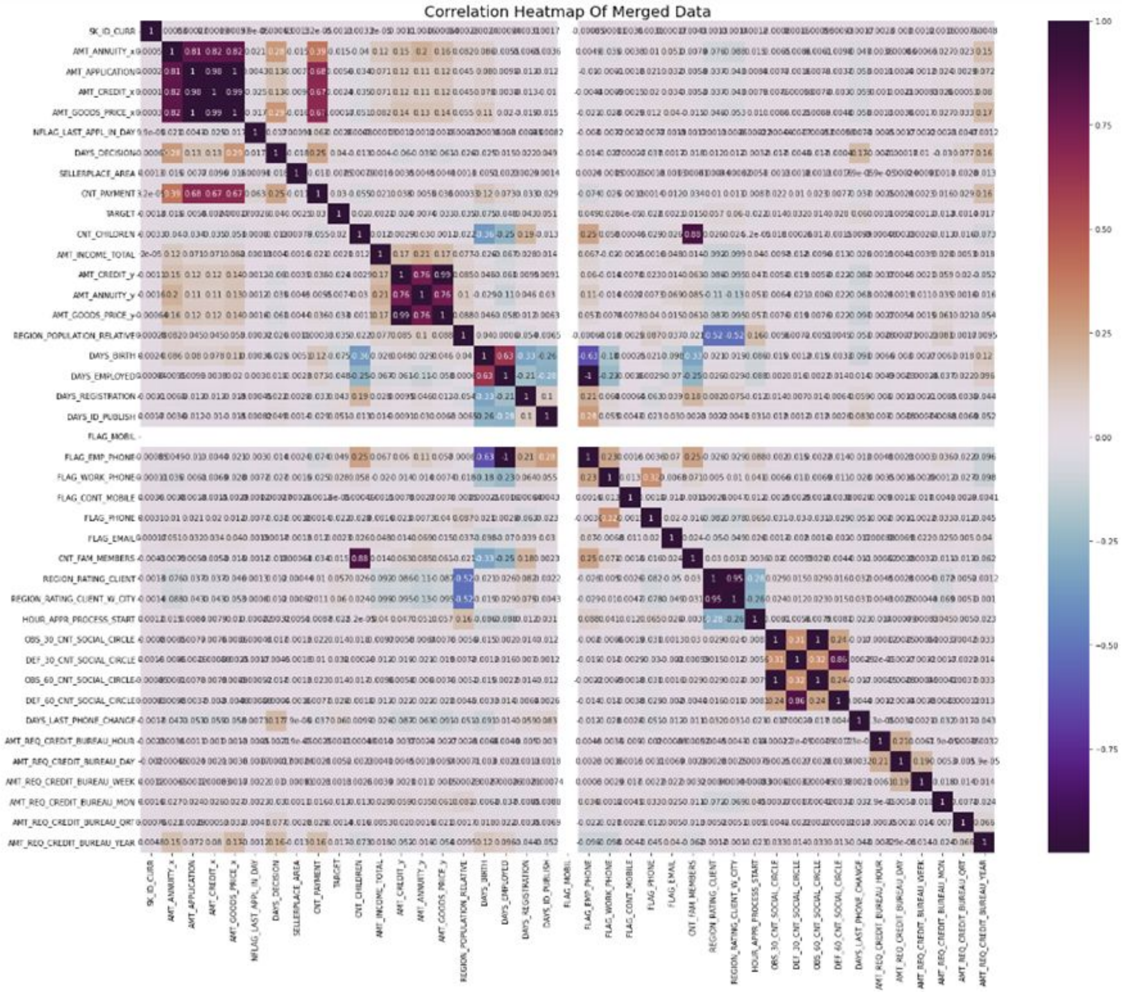
1. Highest Correlations:
   - Credit amount and Application Amount: Suggests that most of the loans amounts sanctioned were as per application of the client.
   - Goods Price and Application Amount: Suggests that goods price has a possible correlation with loan amount applied.
   - Credit amount and Goods Price: Above two observations naturally suggest correlation between these two. Same is proven in the heatmap.
2. Lowest Correlations:
   - Last application in day flag and sellerplace area appear to have to no correlation with other columns.



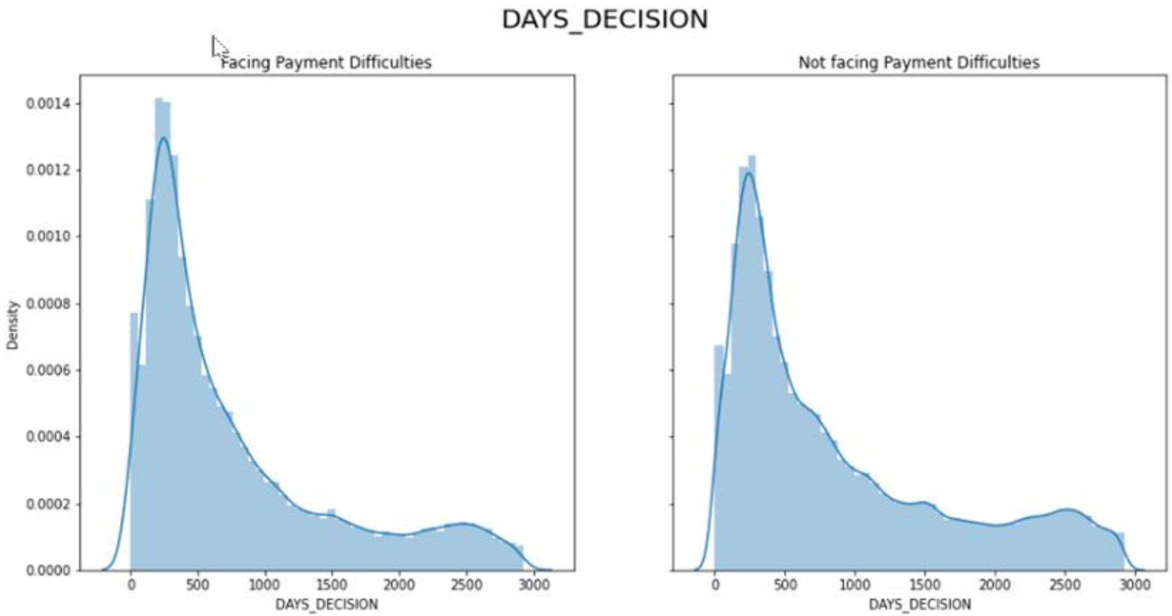
- Most of the applicants live in house or apartment however those living with parents or living on rented house have more percentage of payment difficulty compared to those that don't, when you compare target 0 with target 1.
- Therefore, along with House/apartment, we can consider these two housing types as our defaulter factors as well.

Merging application data with previous application data

- 'Number of Children' is highly correlated with 'Loan Annuity', 'Previous applicant credit amount' and 'Goods price', which means more applications are receive from applicants with higher number of kids.
- Based on the diagram we found that the attributes below are highly correlated with Target attribute:
    - DAYS_DECISION - 0.04
    - DAYS_REGISTRATION - 0.043
    - DAYS_ID_RUBLISH - 0.051
    - FLAG_EMP_PHONE - 0.049
    - REGION_RATING_CLIENT- 0.057
    - REGION_RATING_CLIENT_W-CITY - 0.06
    - DAYS_LAST_PHONE_CHANGE - 0.06

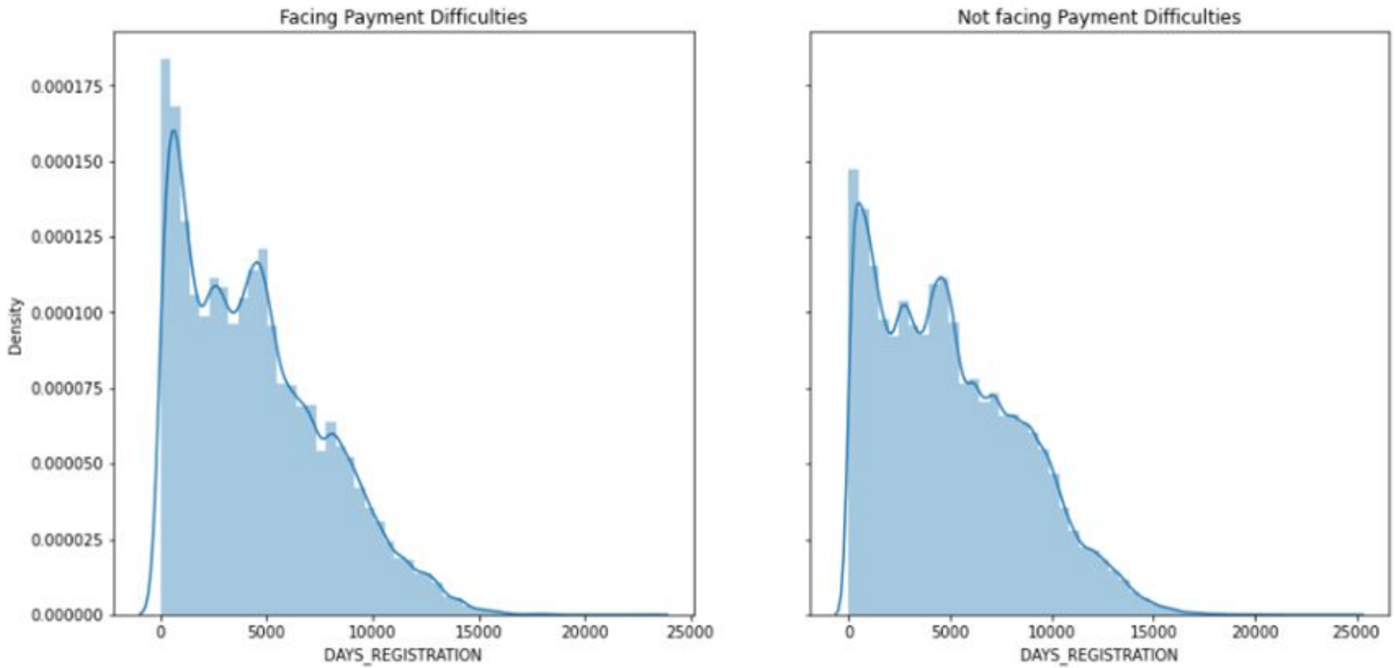

Correlation Heatmap Of Merged Data
SK_ID_CURR
AMT_ANNUITY_x
AMT_APPLICATION
AMT_CREDIT_x
AMT_GOODS_PRICE_x
NFLAG_LAST_APPL_IN_DAY
DAYS_DECISION
SELLERPLACE_AREA
CNT_PAYMENT
TARGET
CNT_CHILDREN
AMT_INCOME_TOTAL
AMT_CREDIT_y
AMT_ANNUITY_y
AMT_GOODS_PRICE_y
REGION_POPULATION_RELATIVE
DAYS_BIRTH
DAYS_EMPLOYED
DAYS_REGISTRATION
DAYS_ID_PUBLISH
FLAG_MOBIL
FLAG_EMP_PHONE
FLAG_WORK_PHONE
FLAG_CONT_MOBILE
FLAG_PHONE
FLAG_EMAIL
CNT_FAM_MEMBERS
REGION_RATING_CLIENT
REGION_RATING_CLIENT_W_CITY
HOUR_APPR_PROCESS_START
OBS_30_CNT_SOCIAL_CIRCLE
DEF_30_CNT_SOCIAL_CIRCLE
OBS_60_CNT_SOCIAL_CIRCLE
DEF_60_CNT_SOCIAL_CIRCLE
DAYS_LAST_PHONE_CHANGE
AMT_REQ_CREDIT_BUREAU_HOUR
AMT_REQ_CREDIT_BUREAU_DAY
AMT_REQ_CREDIT_BUREAU_WEEK
AMT_REQ_CREDIT_BUREAU_MON
AMT_REQ_CREDIT_BUREAU_QRT
AMT_REQ_CREDIT_BUREAU_YEAR
1.00
0.75
0.50
0.25
0.00
-0.25
-0.50
-0.75

- **Breaking down the merged data for easier understanding and further examination of the relevant attributes.**



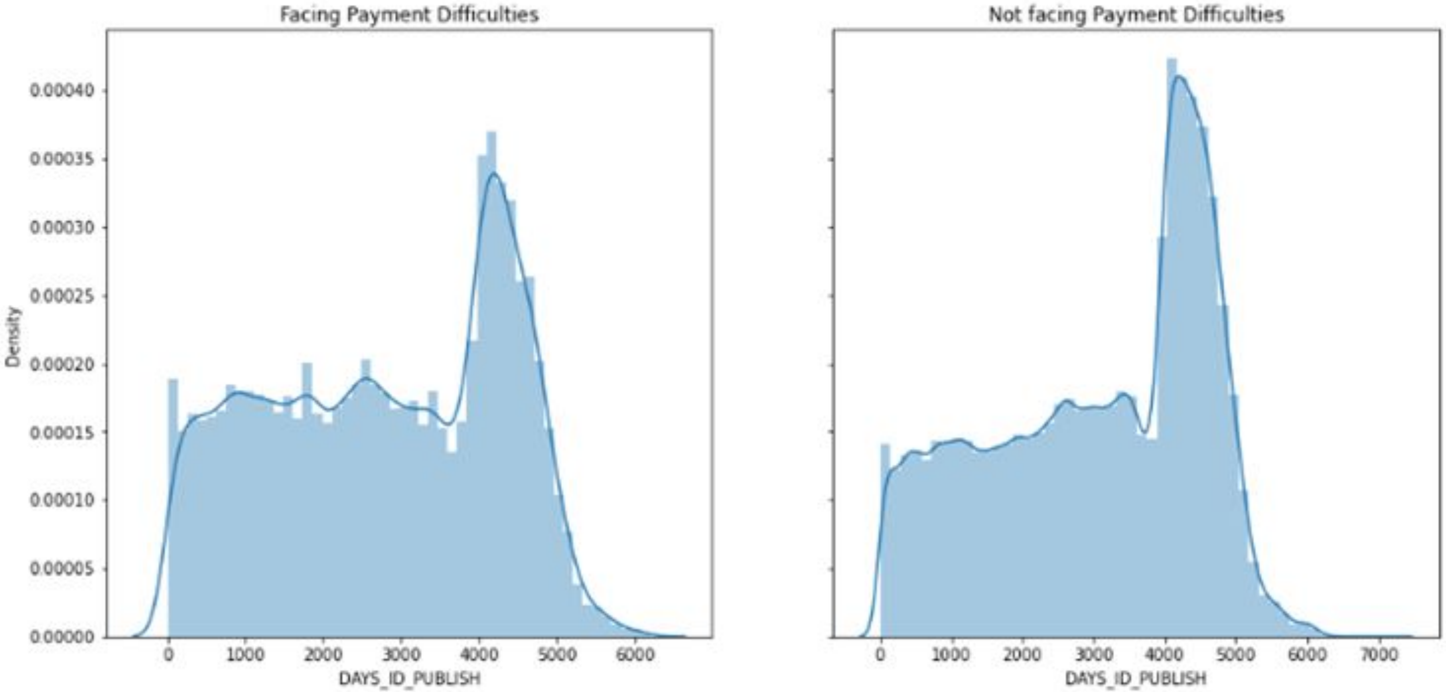
- Both categories, defaulter and non defaulter, are showing similar kind of structure so we can ignore this attribute.

DAYS_REGISTRATION



- Both categories defaulter and non defaulter are showing similar kind of structure so we can ignore this attribute.

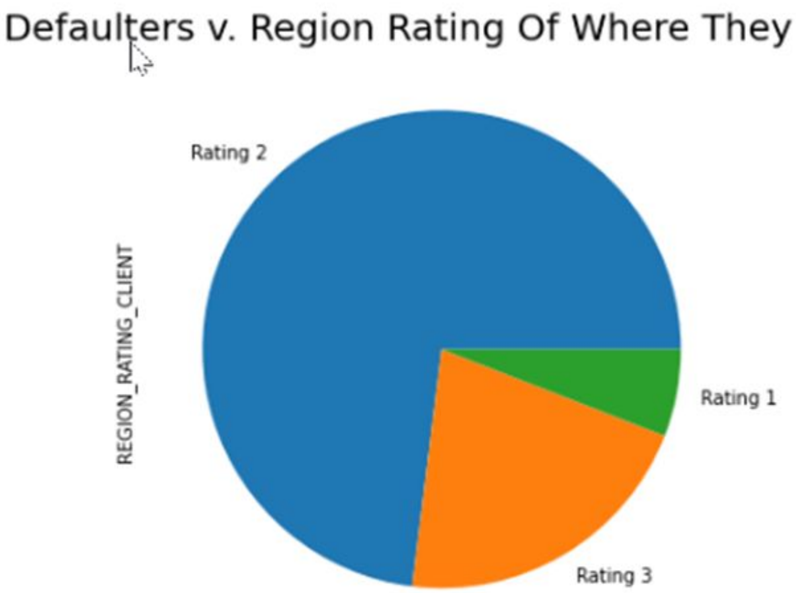
How many days before the application did client change the identity document with which he applied for the loan



- Both categories defaulter clients provide phone number

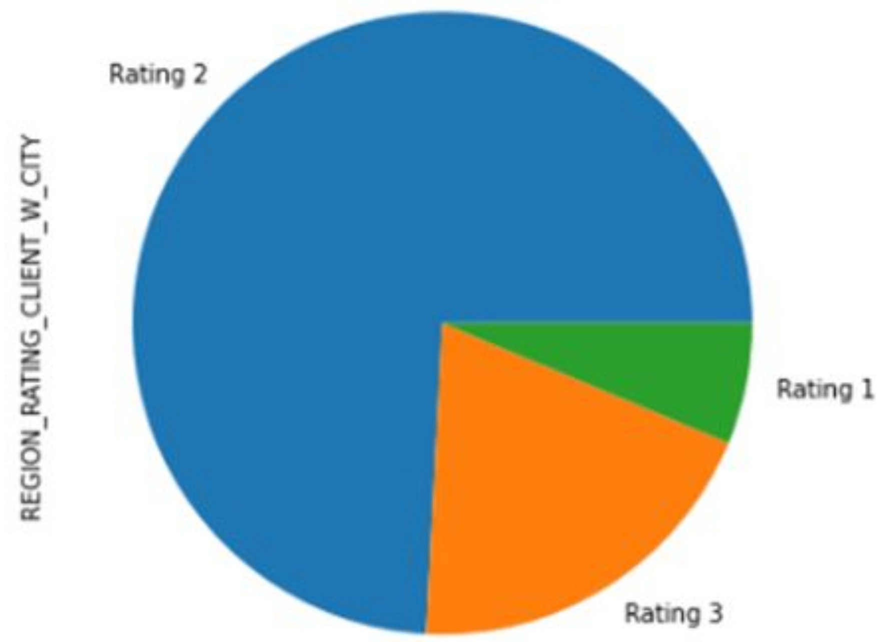
Defaulter are also providing their contact details so we can not infer anything from this attribute.

Chart to check defaulter clients vs. region rating of their residency

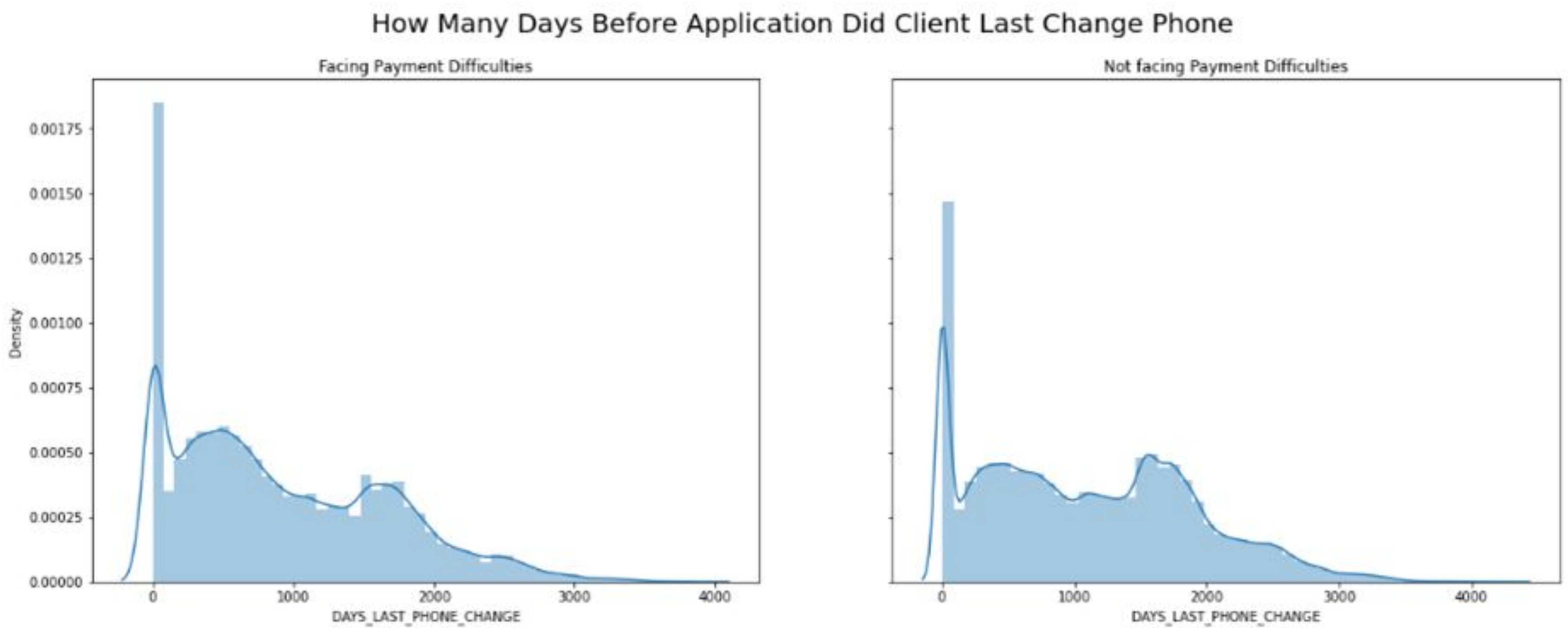


- The clients who live in 2 rated regions are more likely to have payment difficulty.

Defaulters v. Region And City Rating Of Where They Live



- The clients who live in 2 rated cities and regions are more likely to have payment difficulty.



- The clients who last changed the phone within few days of applying are more likely to default

Top 10 positive correlations for Defaulters

Top 10 positive correlations for Defaulters

In [214]:
```
#Top 10 correlation
top10_merge = merge_currappdata_prevdata.corr().unstack().sort_values(ascending=False).drop_duplicates()

#Starting index from 1 because SK_ID_CURR was used to merge
top10_merge[1:11]
```

Out[214]:
```
AMT_GOODS_PRICE_x           AMT_APPLICATION             0.999884
OBS_30_CNT_SOCIAL_CIRCLE    OBS_60_CNT_SOCIAL_CIRCLE    0.998563
AMT_CREDIT_x                AMT_GOODS_PRICE_x           0.993087
AMT_CREDIT_y                AMT_GOODS_PRICE_y           0.986341
AMT_APPLICATION             AMT_CREDIT_x                0.975822
REGION_RATING_CLIENT_W_CITY REGION_RATING_CLIENT        0.945503
CNT_FAM_MEMBERS             CNT_CHILDREN                0.879213
DEF_60_CNT_SOCIAL_CIRCLE    DEF_30_CNT_SOCIAL_CIRCLE    0.862736
AMT_GOODS_PRICE_x           AMT_ANNUITY_x               0.820895
AMT_CREDIT_x                AMT_ANNUITY_x               0.816429
dtype: float64
```

In [215]:
```
# Assigning dataframe as per target 0 and target 1 variable
target0_merge_currappdata_prevdata  = merge_currappdata_prevdata[merge_currappdata_prevdata["TARGET"] == 0]
target1_merge_currappdata_prevdata  = merge_currappdata_prevdata[merge_currappdata_prevdata["TARGET"] == 1]
```

In [216]:
```
#Top 10 correlation of applicants who did not face problems with payment
top10_merge_target0 = target0_merge_currappdata_prevdata.corr().unstack().sort_values(ascending=False).drop_duplicates()

#Starting index from 1 because SK_ID_CURR was used to merge
top10_merge_target0[1:11]
```

Out[216]:
```
AMT_APPLICATION           AMT_GOODS_PRICE_x              0.999888
OBS_30_CNT_SOCIAL_CIRCLE  OBS_60_CNT_SOCIAL_CIRCLE       0.998579
AMT_GOODS_PRICE_x         AMT_CREDIT_x                   0.993297
AMT_CREDIT_y              AMT_GOODS_PRICE_y              0.986625
AMT_APPLICATION           AMT_CREDIT_x                   0.975764
REGION_RATING_CLIENT      REGION_RATING_CLIENT_W_CITY    0.944260
CNT_CHILDREN              CNT_FAM_MEMBERS                0.878467
DEF_30_CNT_SOCIAL_CIRCLE  DEF_60_CNT_SOCIAL_CIRCLE       0.863173
AMT_ANNUITY_x             AMT_GOODS_PRICE_x              0.821057
AMT_CREDIT_x              AMT_ANNUITY_x                  0.816580
dtype: float64
```

In [217]:
```
#Top 10 correlation of applicants who faced problems with payment
top10_merge_target1 = target1_merge_currappdata_prevdata.corr().unstack().sort_values(ascending=False).drop_duplicates()

#Starting index from 1 because SK_ID_CURR was used to merge
top10_merge_target1[1:11]
```

Out[217]:
```
AMT_GOODS_PRICE_x           AMT_APPLICATION             0.999675
OBS_60_CNT_SOCIAL_CIRCLE    OBS_30_CNT_SOCIAL_CIRCLE    0.998391
AMT_CREDIT_x                AMT_GOODS_PRICE_x           0.992292
AMT_GOODS_PRICE_y           AMT_CREDIT_y                0.982936
AMT_CREDIT_x                AMT_APPLICATION             0.975686
REGION_RATING_CLIENT_W_CITY REGION_RATING_CLIENT        0.956395
CNT_FAM_MEMBERS             CNT_CHILDREN                0.886265
DEF_60_CNT_SOCIAL_CIRCLE    DEF_30_CNT_SOCIAL_CIRCLE    0.858279
AMT_CREDIT_x                AMT_ANNUITY_x               0.840375
AMT_ANNUITY_x               AMT_GOODS_PRICE_x           0.840052
dtype: float64
```

Final Observations

Important Columns For The Bank To Watchout Against Defaults:

- AMT_INCOME_TOTAL *As a result, clients with higher incomes are less likely to experience payment difficulties, so low-income groups are more likely to default.

    - The majority of defaulters are both male and female in the income range of 100000 to 150000.
        - In the income range of 250000 and below, there are more female defaulters.
        - The income range 250000 and above has more male defaulters.

- AMT_CREDIT

    - Customers with the highest credit amount and the lowest region population have significant correlation outliers, compared to clients in regions with higher populations.

- NAME_FAMILY_STATUS

    - Clients with academic degrees who are not married have a much higher minimum whisker than all other categories.
    - Outliers on the high side are those with higher education or special education in secondary or secondary education.

- CNT_CHILDREN

    - People who live in high-density areas and have a large number of kids are more likely to have difficulties paying.

- NAME_EDUCATION_TYPE
    - Defaulters were primarily from secondary and higher education backgrounds.
    - Academic degrees were the background of the least defaulters

- OCCUPATION_TYPE
    - Cooking staff had the least count of clients with difficulty in payment of loan.
    - Labourers had the highest count of clients with difficulty in payment of loan.

- NAME_HOUSING_TYPE
    - When comparing target 0 with target 1, applicants living with parents or renting a house have more difficulty paying compared to applicants who don't.
    - These two types of housing are thus also defaulters in our analysis along with the House/Apartment.

- Apart from the above, following are few more attributes that can also help us to identify defaulters:
    - DAYS_LAST_PHONE_CHANGE
        - Default rates are higher for clients who change their phone within a few days of applying.
    - REGION_RATING_CLIENT and REGION_RATING_CLIENT_W_CITY
        - Most clients who live in a tier2 city or region are at risk of defaulting.

Important Columns For The Bank To Increase Revenue And Clients:

- DAYS_BIRTH
  - Numbers of non-defaulted credit cards are highest in the 55-60 age bracket.
  - Compared to others, they do not have a high default rate, so they might be a target for growth.
- NAME_FAMILY_STATUS
  - It would be beneficial to put more focus on widows and separated clients, since they are observed to take good numbers of loans with a much lower default rate than married and single clients.
- NAME_EDUCATION_TYPE
  - Non-married clients with higher education or secondary/secondary special education are key outliers to tap as less risky clients on the upperside.
  - All people with academic degree also have high income overall.
- OCCUPATION_TYPE
  - Cooking staff and private service staff drive provide a very good volume as well as very low chance of payment issues.
- REGION_POPULATION_RELATIVE
  - Medium to high density population cities have very low default rate for loan amounts. They can be focused on for bigger loans to increase revenue.