EDA Case Study – Loan Defaults

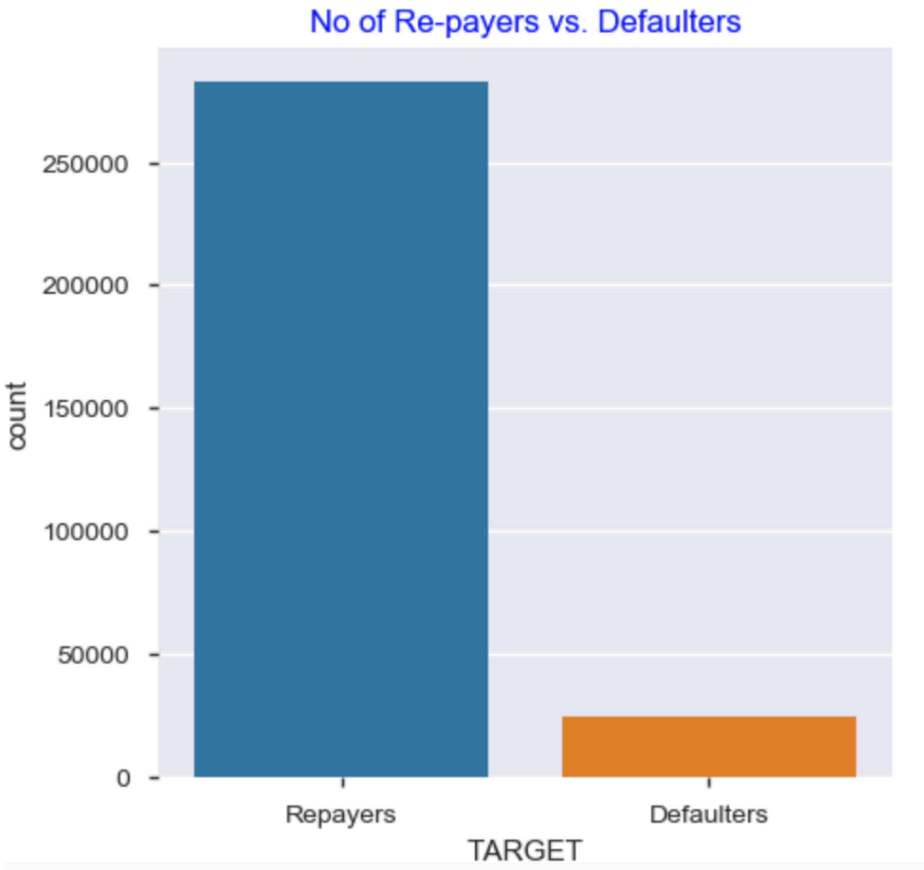
IIITB

Vidhu Jain

Problem Statement

- Data Analysis of Current and Previous Loan applications to check the predictor that can help in loan default detection in order to minimize the risk of Loan defaults

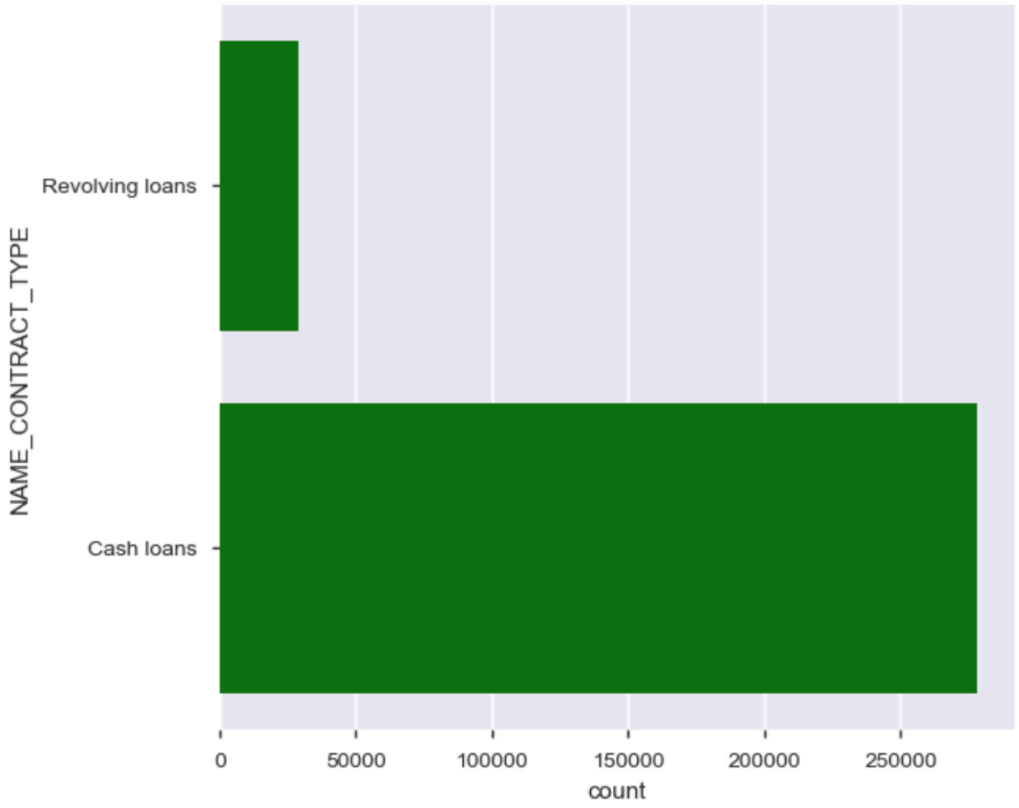
The ratio of loan defaults in comparison to re-payers is always less

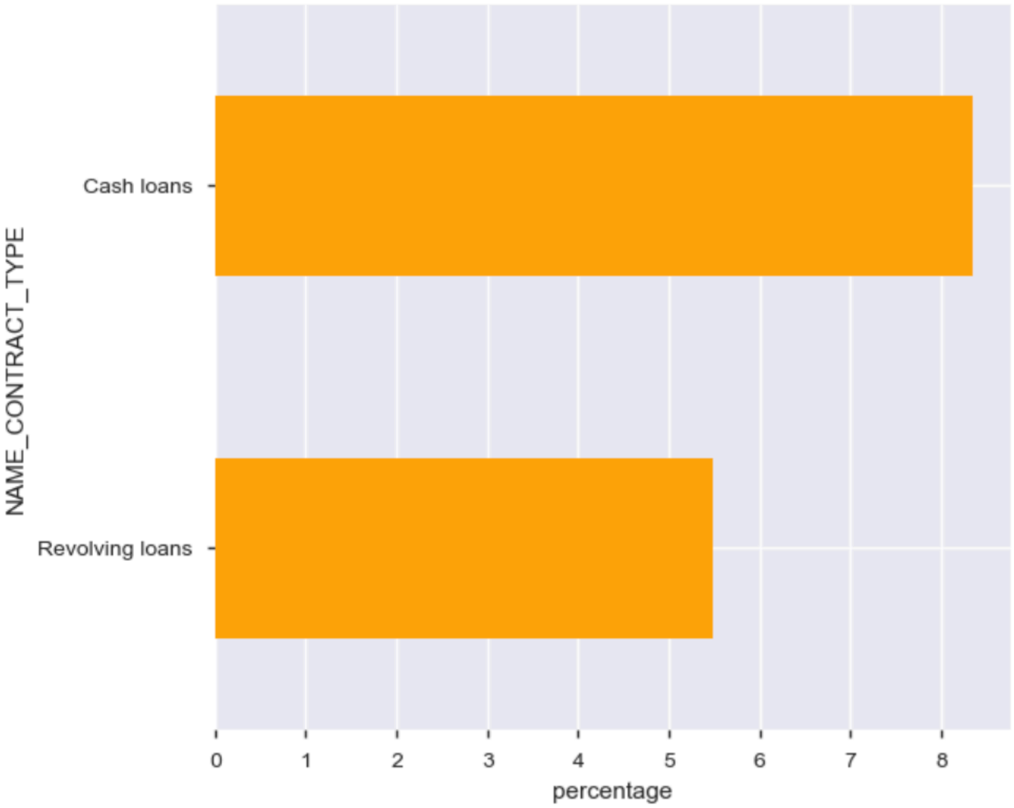


From the data available

- The Target class is highly imbalanced
- 91.93% of observations as "0" - labeled as repayers
- 8.07% of observations as "1" - labeled defaulters

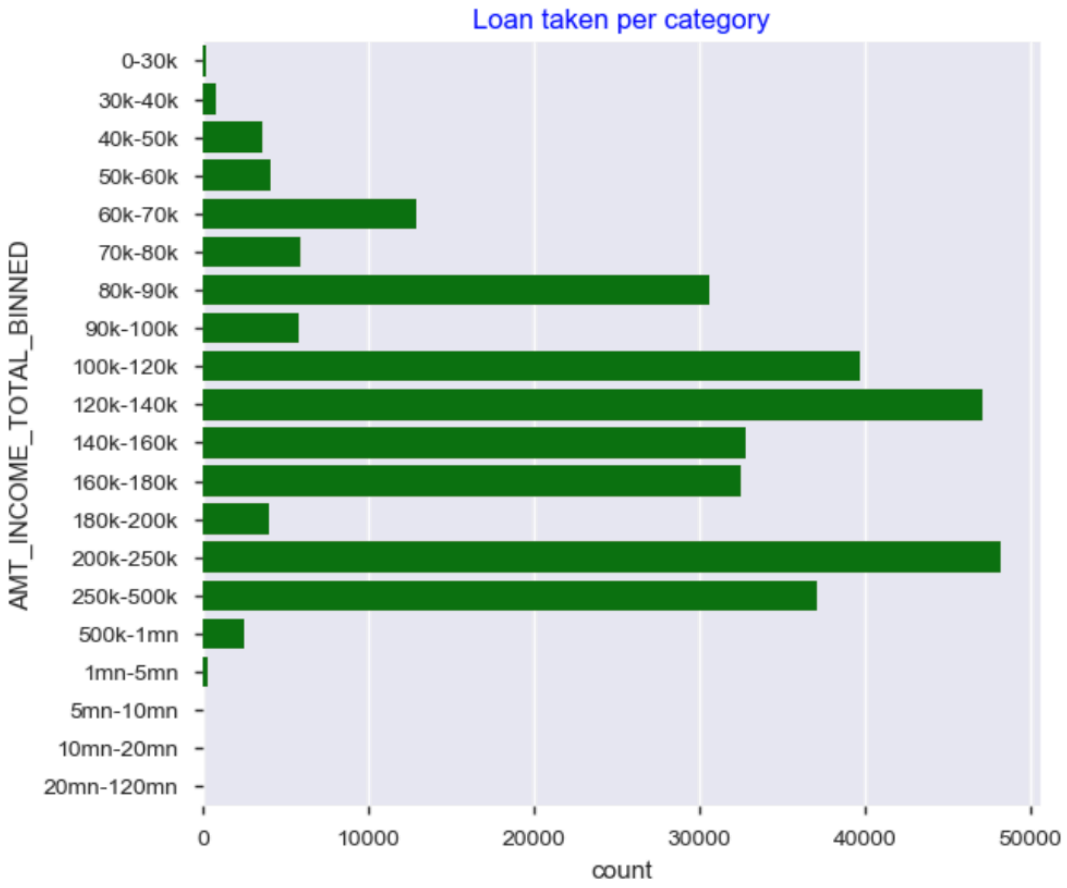
Type of Loans taken and percentage of loan defaults


Loan taken per category

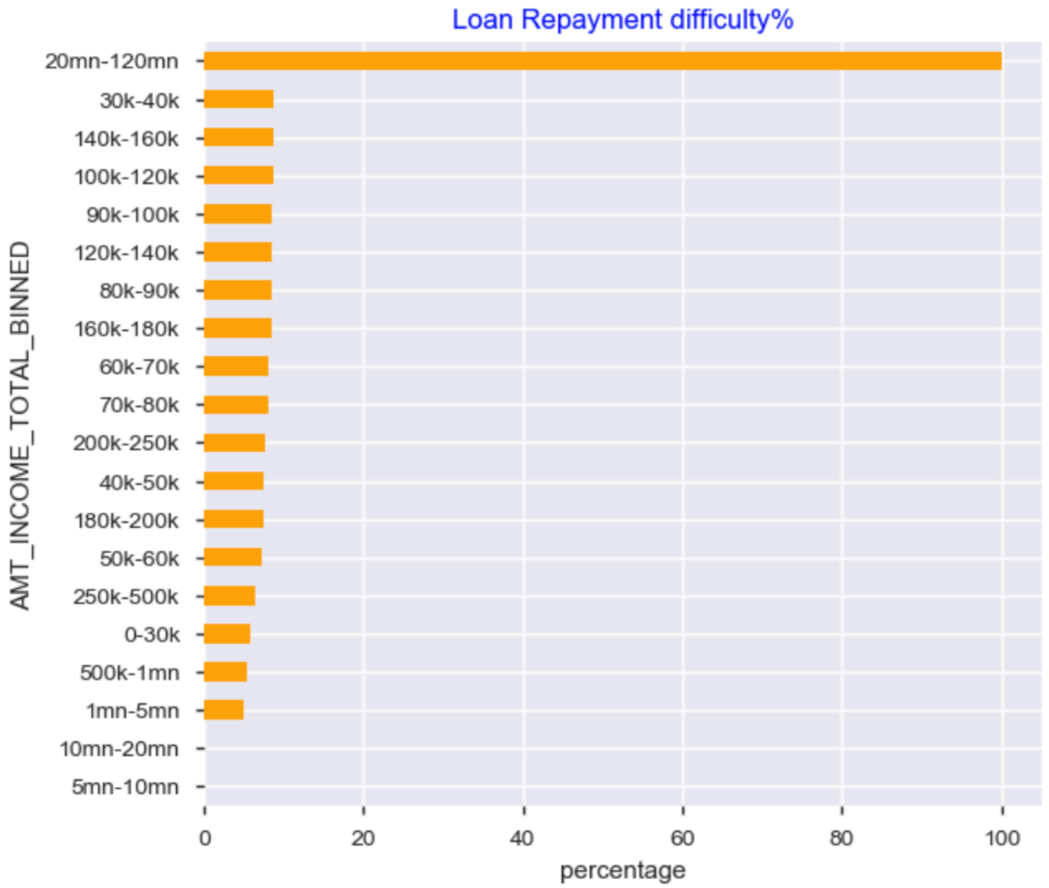

Loan Repayment difficulty%

- Maximum no. of applications are for Cash Loan - 278232
  - Out of these 8.35% Cash Loans have difficulty in repayment of loans
- Revolving Loans are the second type - 29279
  - Out of these 5.48% Revolving Loans have difficulty in repayment of loans

Income range of people and maximum defaults in income range



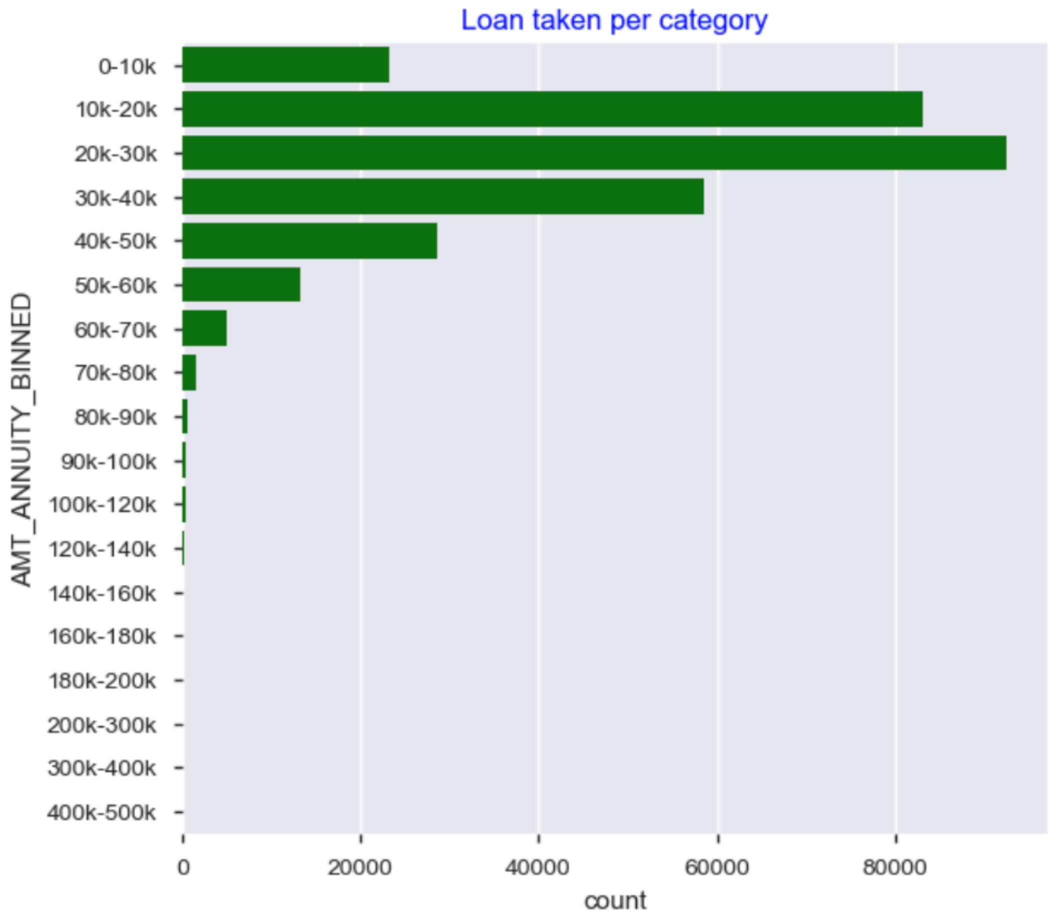
Loan taken per category

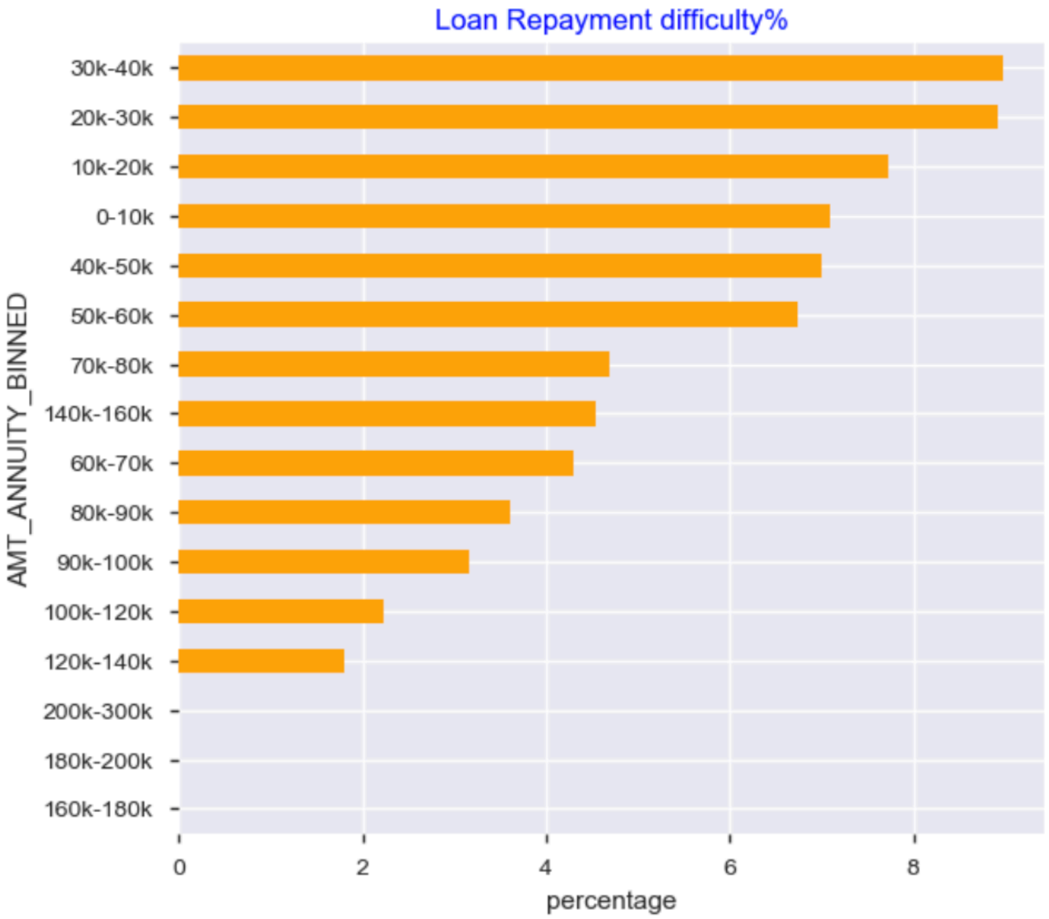


Loan Repayment difficulty%

- Maximum no. of loans are taken by people with income amount in range of [200k – 250k]
- In terms of ratio within each category, people with income amount in range of [20mn – 120mn] face more difficulty in loan repayment (100%) followed by people with income amount in range of [30k – 40k] - 8.76%

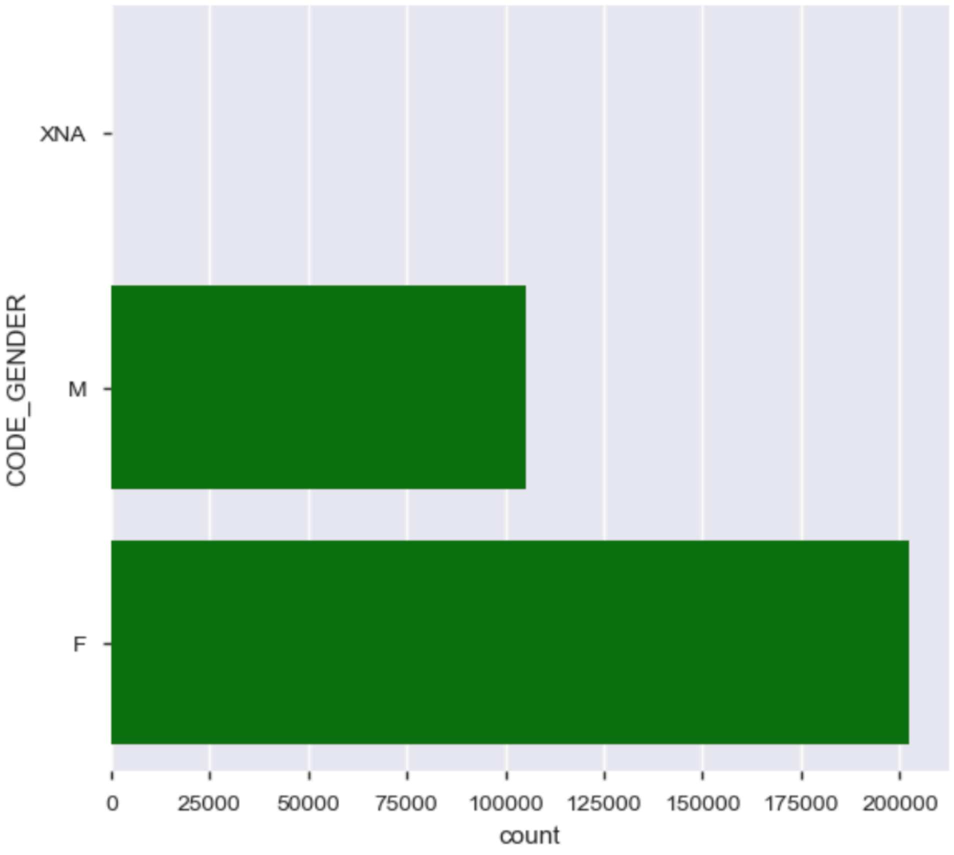
Annuity amount range of people and maximum defaults in annuity amount range

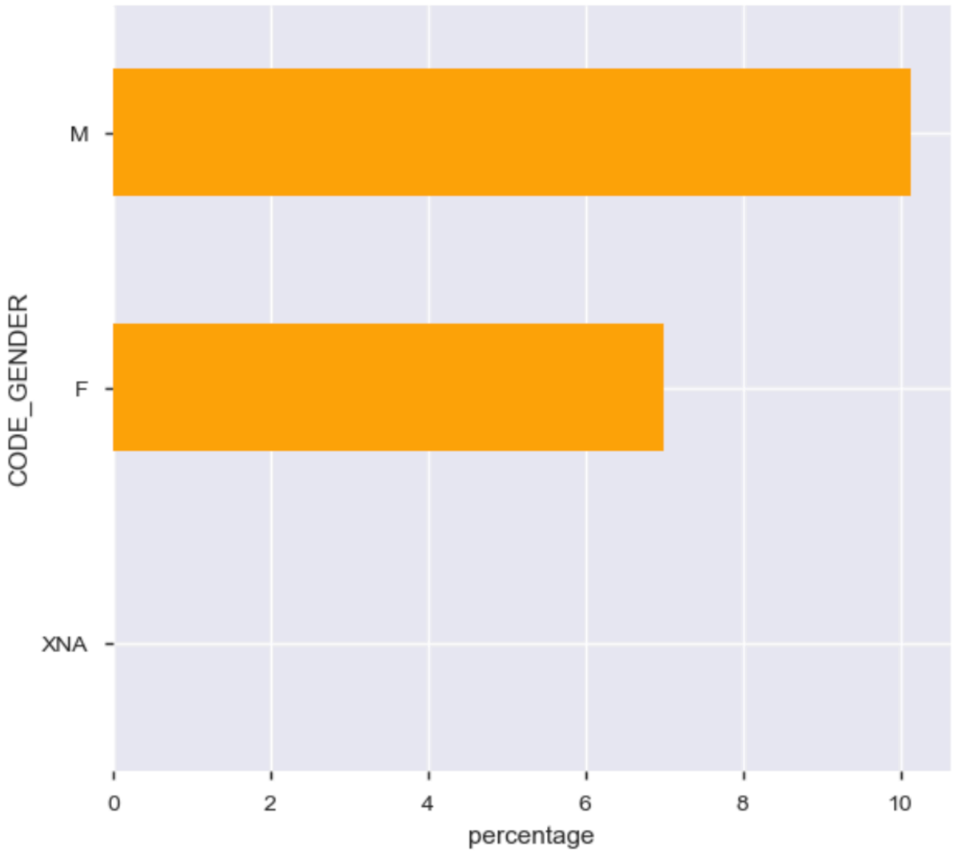




- Maximum number of loans are taken by people with annuity amount in range of [20k – 30k]
- In terms of loan default percentage within each category, people with annuity amount in range of [30k – 40k] face more difficulty in loan repayment (8.97%)

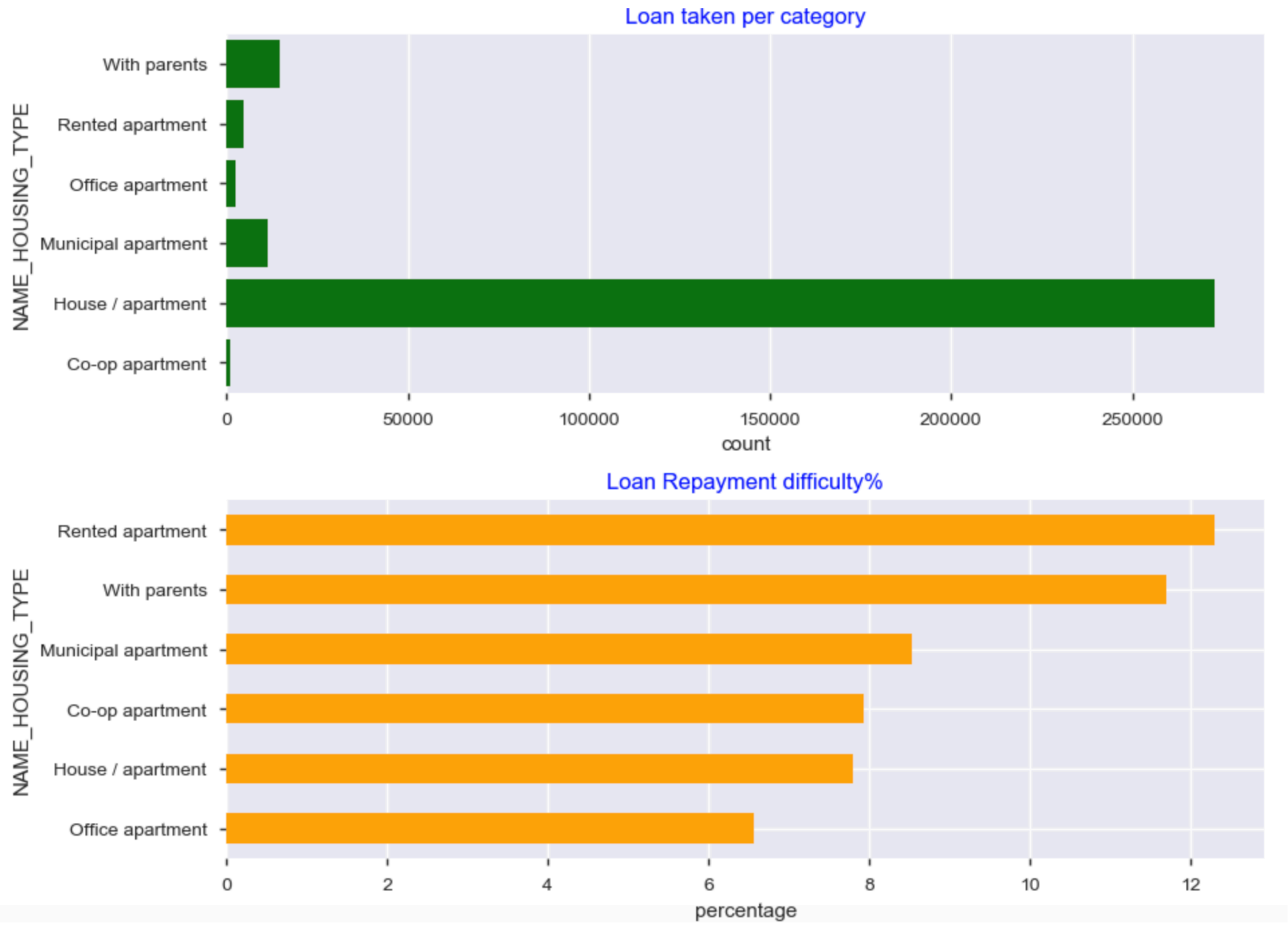
Loans taken per gender and repayment difficulty percentage


Loan taken per category


Loan Repayment difficulty%

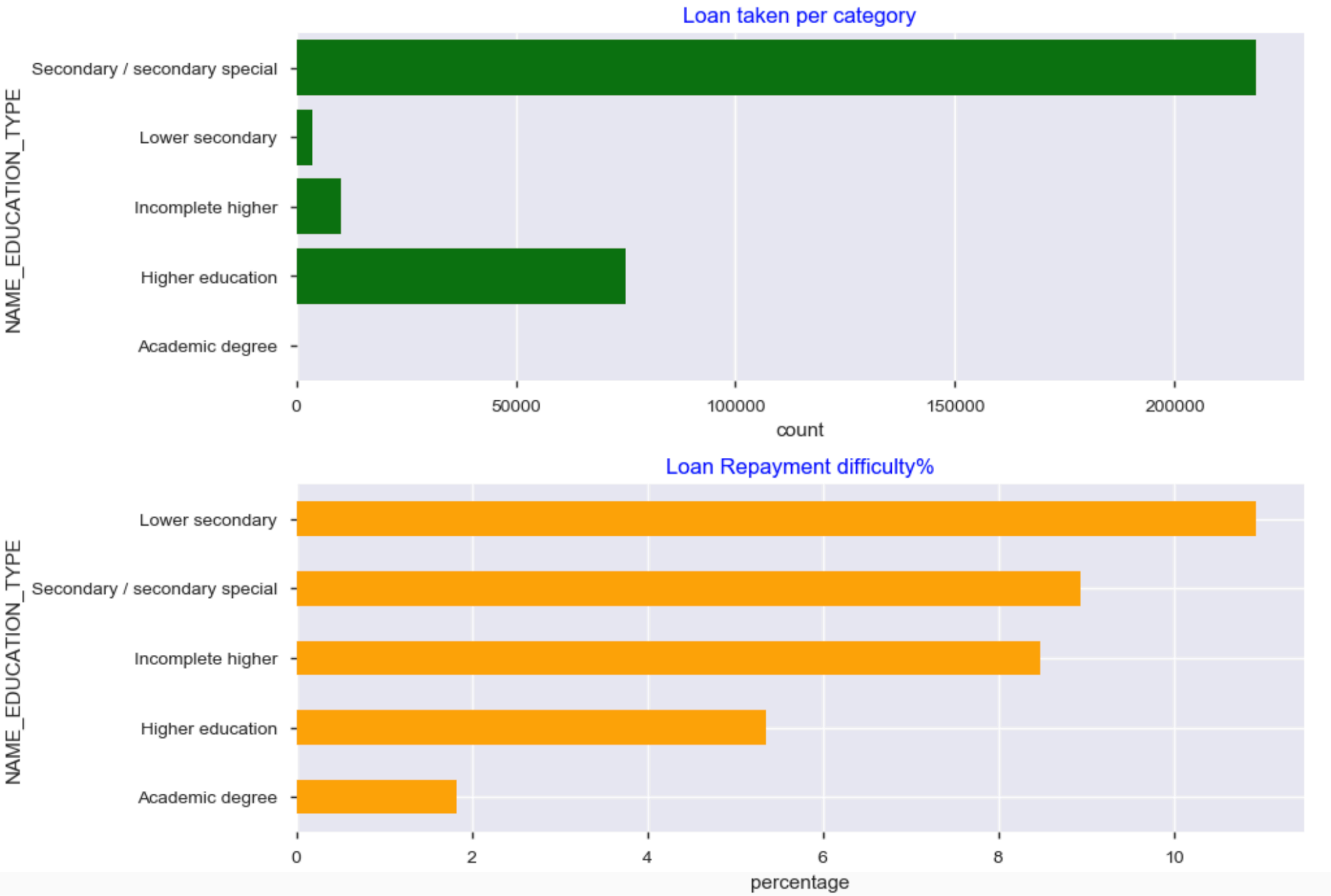
- Maximum number of loans are taken by Female
- In terms of default percentage, Male applicants have highest difficulty in loan repayments (10.14%)

Loans taken per housing type of people and repayment difficulty percentage



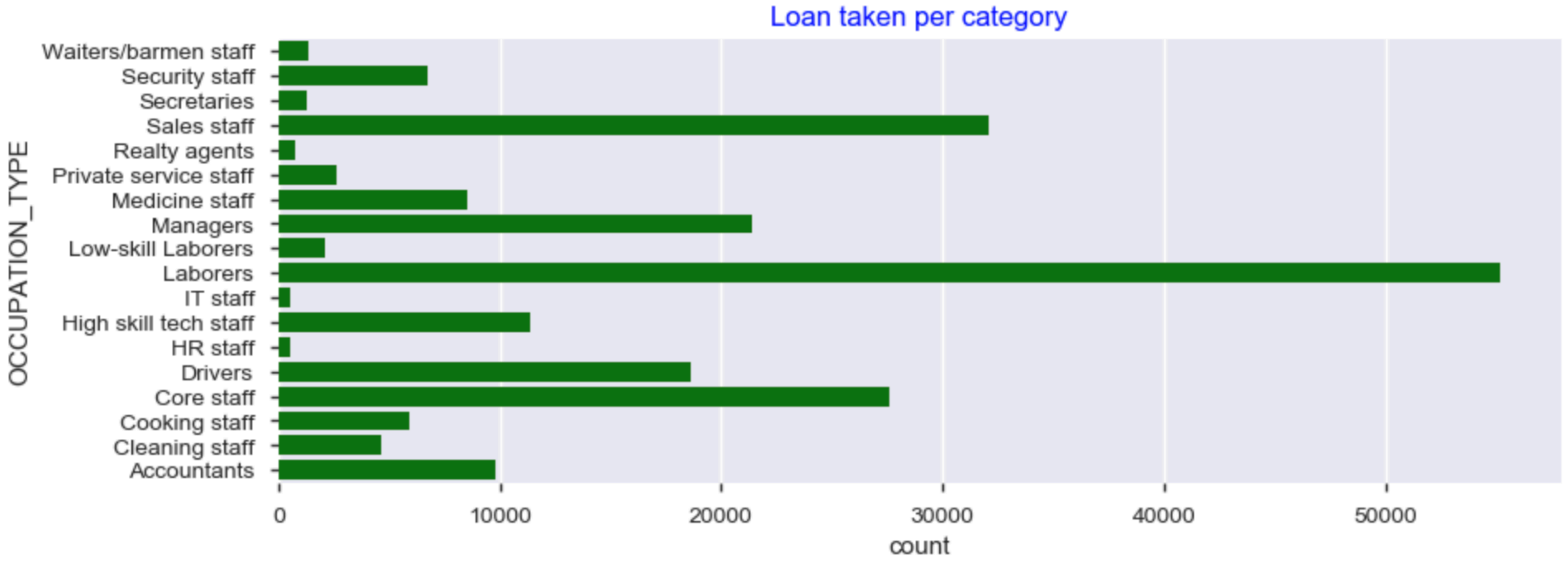
- Maximum number of loans are taken by people who own House/Apartment
- People living in rented apartments have major difficulty in loan repayment (12.31%)

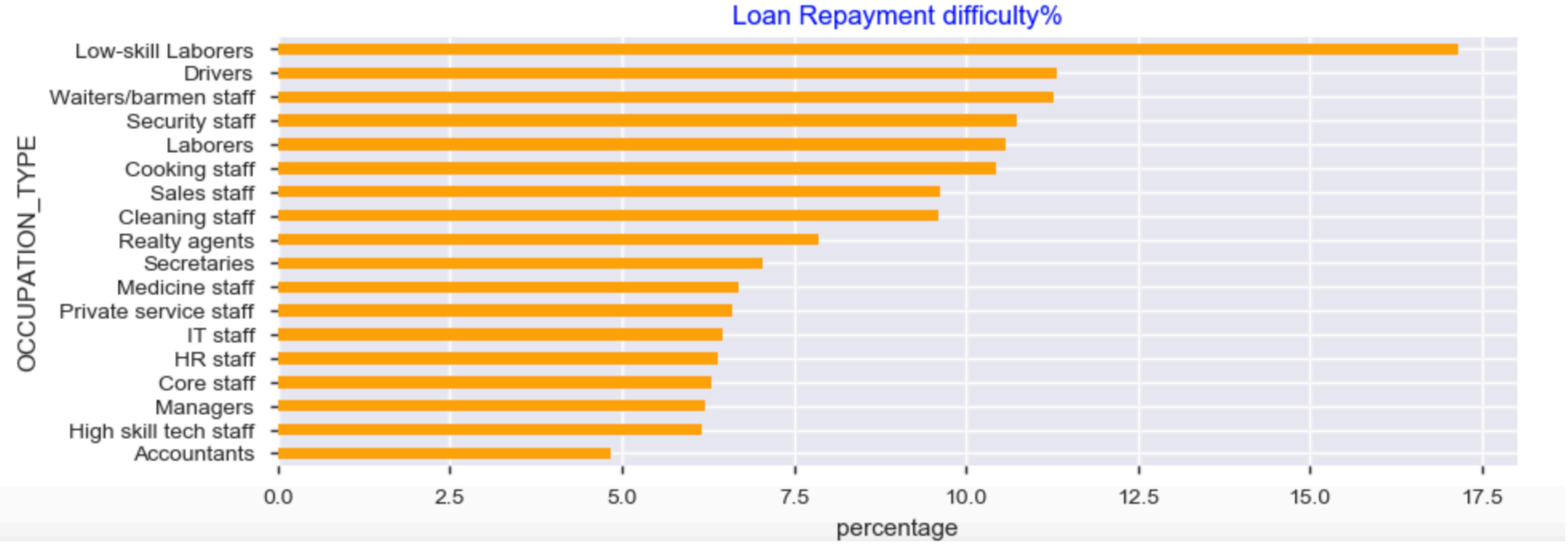
Loans taken per education background of people and repayment difficulty percentage



- Maximum number of loans are taken by people with Secondary / secondary special education background
- People with Lower secondary education background have major difficulty in loan repayment (10.93%)

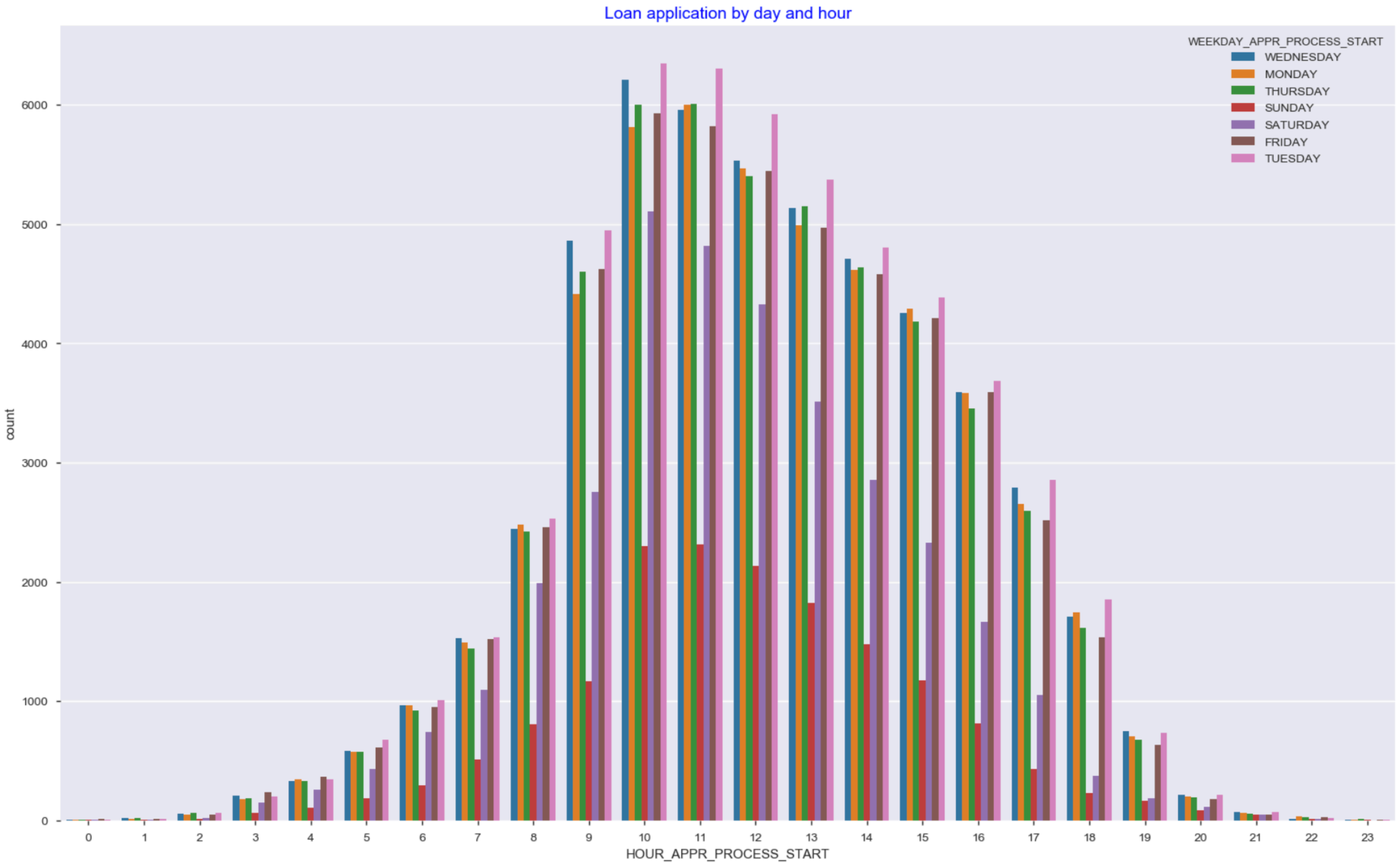
Loans taken per education background of people and repayment difficulty percentage





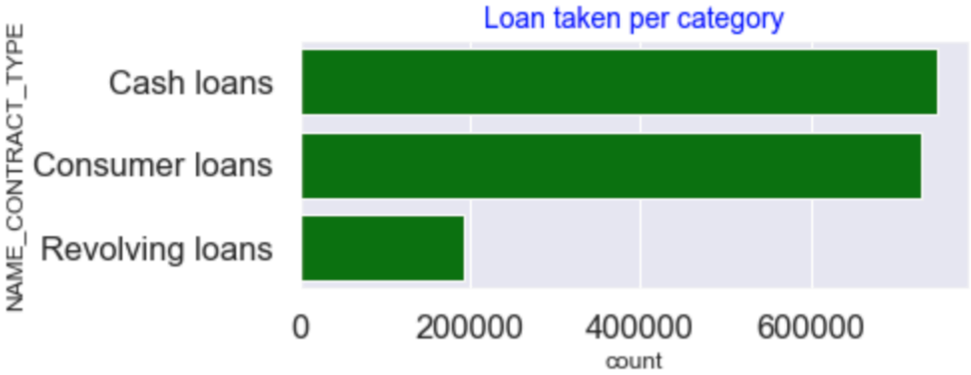
- Maximum number of loan taken is by Laborers
- Low-skill Laborers category has the highest issues of loan repayment (17.15%)

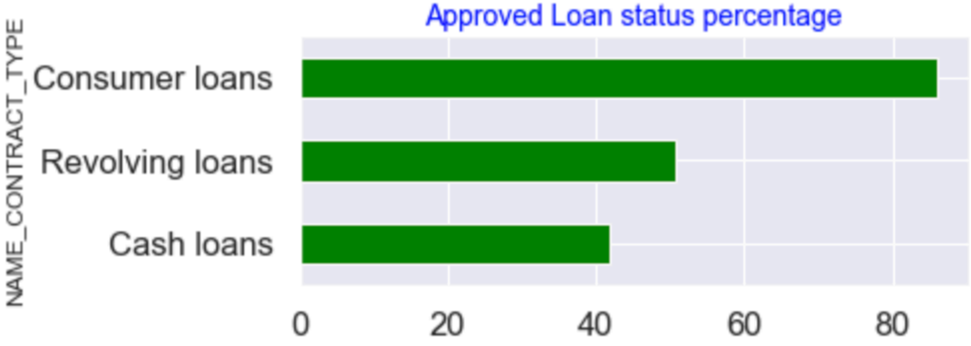
Peak day and hour during loan application

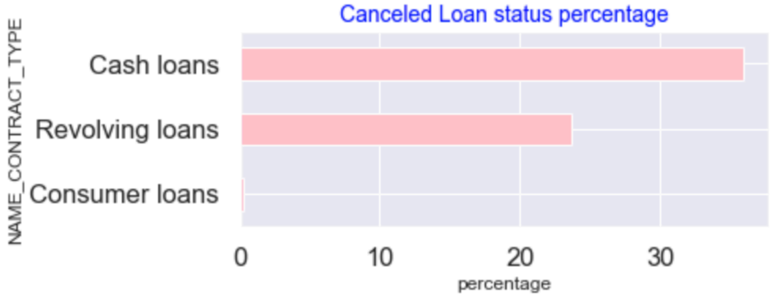


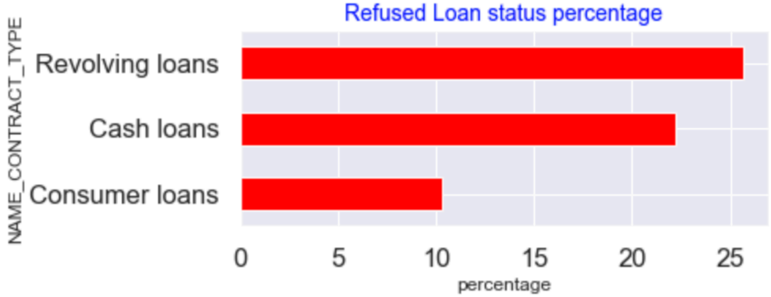
- The loans were applied mainly on Tuesday (17.53%) followed by Wednesday(16.89%)
- 9am - 2pm are peak hours for loan application

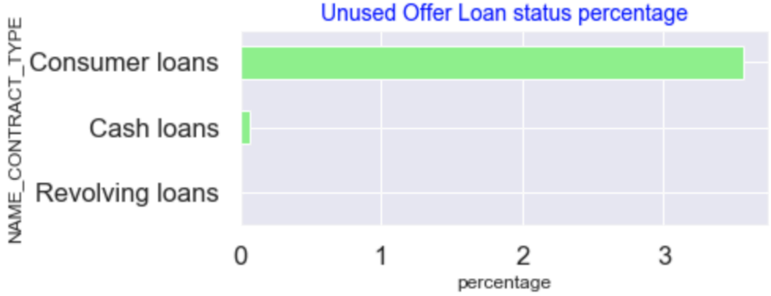
Previous loans applications







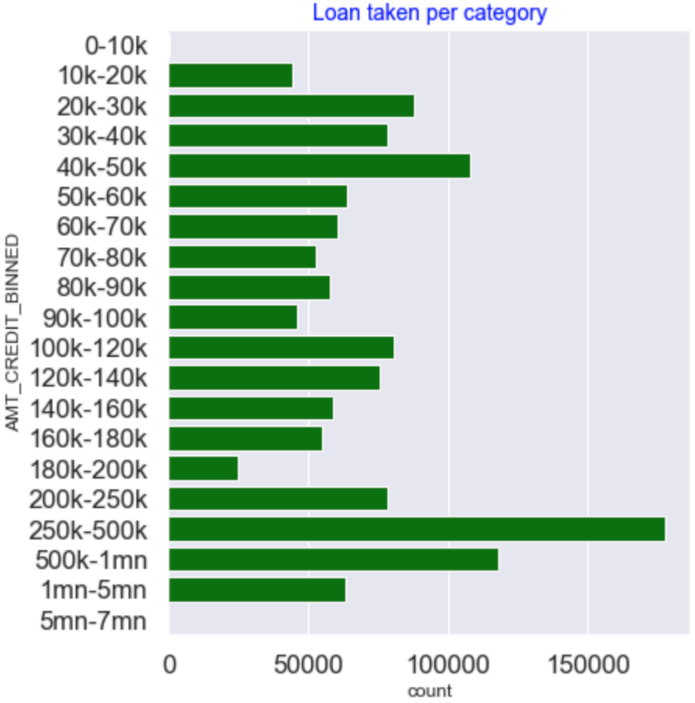


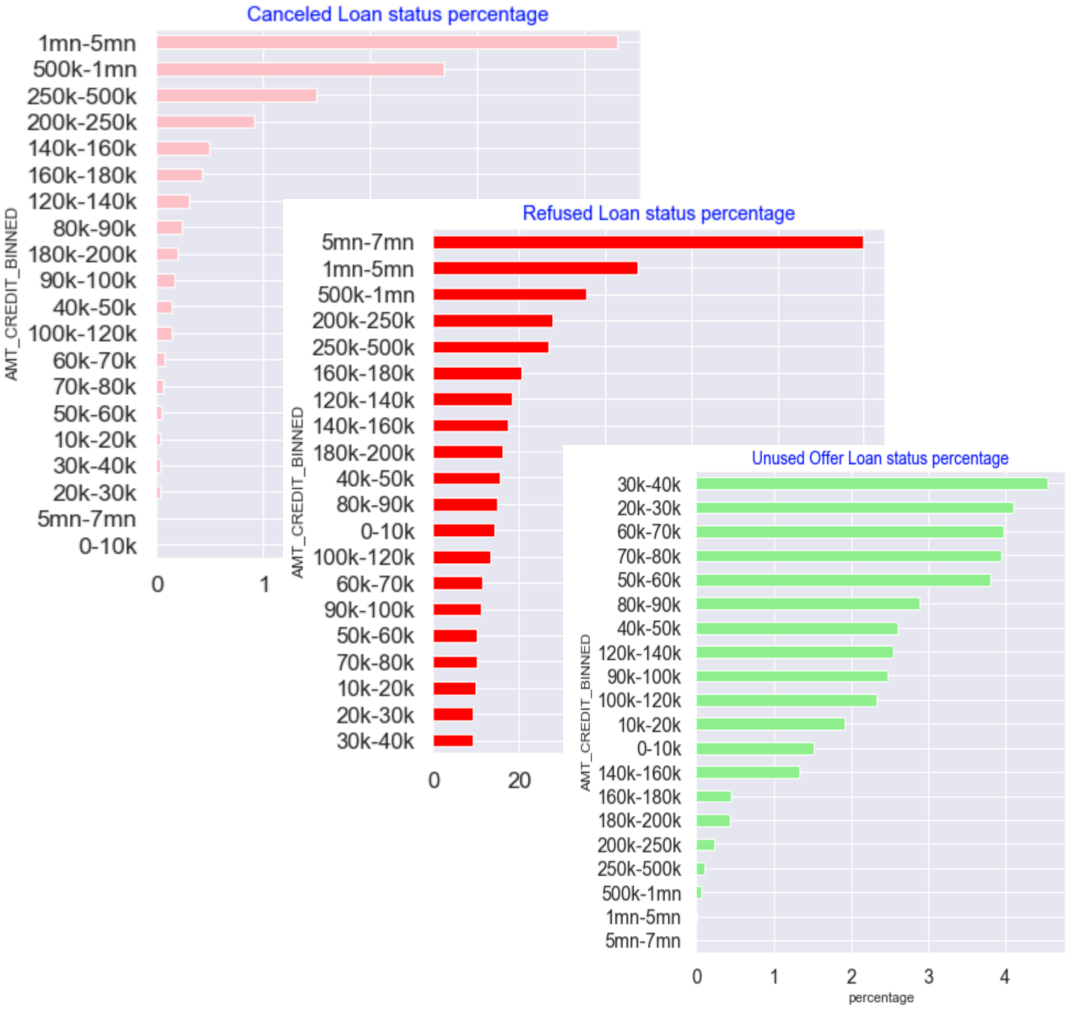


From previous loans applications

- Cash Loans(747553) applications are maximum in number followed by Consumer Loans(729151)
- Consumer Loans percentage of approval is highest (85.92%)
- Cash loan category has maximum percentage of Canceled loans (35.93%)
- Revolving Loans have maximum refusal percentage (25.64%)
- Consumer Loans are the ones that have highest percentage of Unused Offer (3.56%)

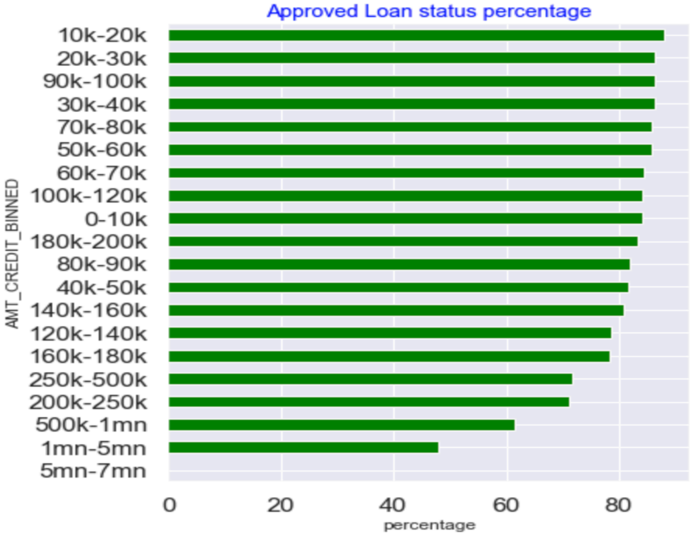
Credit Amount range in Previous loans applications and loan application status



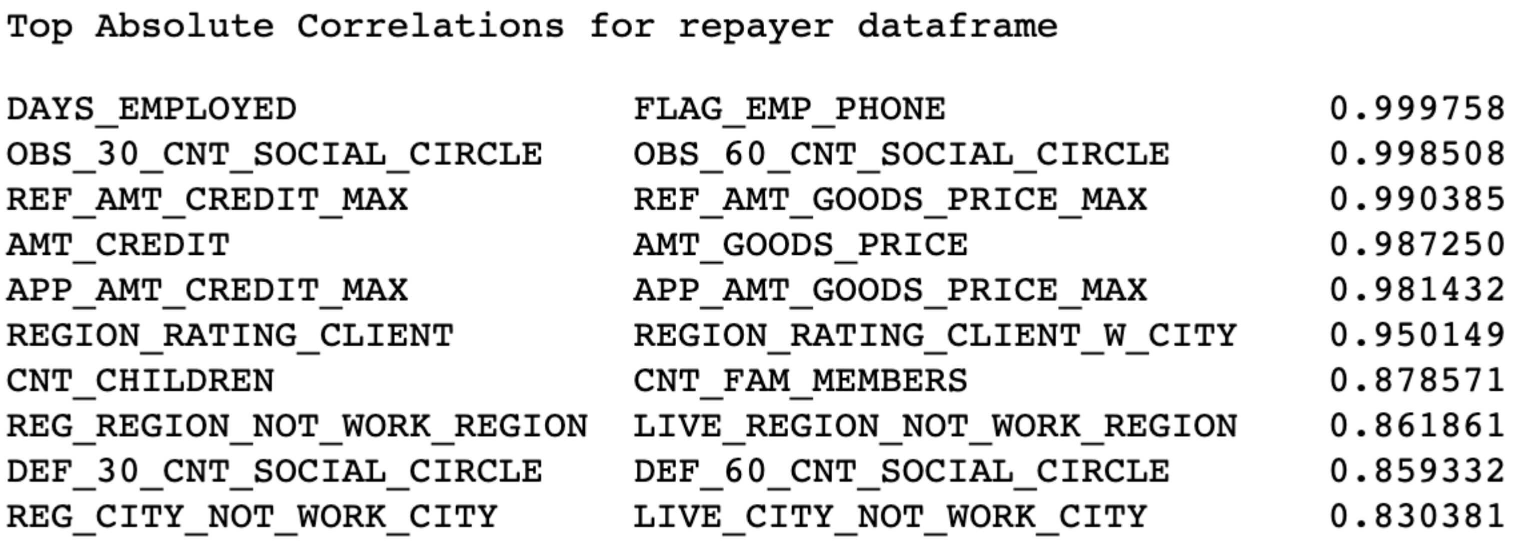


From previous loans applications

- Loan applications are highest for credit amount in range of [250k - 500k] - 177983
- Maximum loans are approved for credit amount in range of [10k - 20k] - 88.11%
- Maximum loans are cancelled for credit amount in range of [1mn - 5mn] - 4.31%
- There is only 1 loan with credit amount in range of [5mn - 7mn] which has been refused as well - 100.0%
- Maximum loans are with unused offer for credit amount in bin 4 i.e. credit amount in range of [30k - 40k] - 4.55%



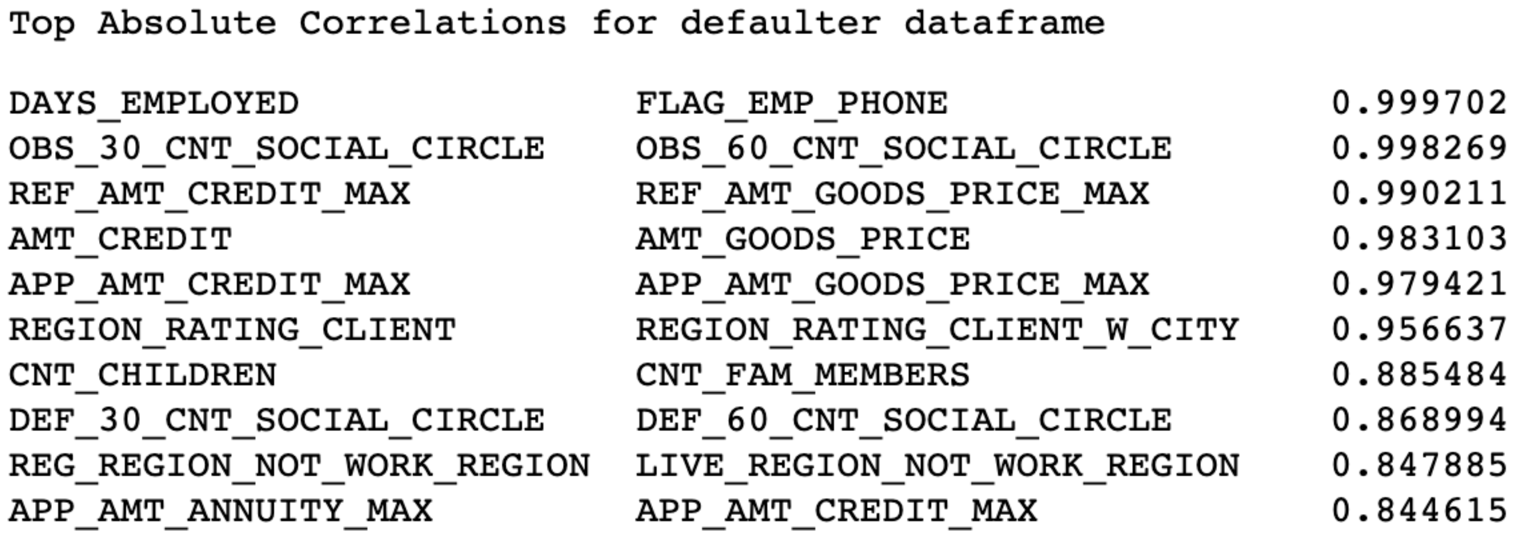
Top 10 correlated features for Target – 0 –Repayer



```
Top Absolute Correlations for repayer dataframe

DAYS_EMPLOYED               FLAG_EMP_PHONE                0.999758
OBS_30_CNT_SOCIAL_CIRCLE    OBS_60_CNT_SOCIAL_CIRCLE      0.998508
REF_AMT_CREDIT_MAX          REF_AMT_GOODS_PRICE_MAX       0.990385
AMT_CREDIT                  AMT_GOODS_PRICE               0.987250
APP_AMT_CREDIT_MAX          APP_AMT_GOODS_PRICE_MAX       0.981432
REGION_RATING_CLIENT        REGION_RATING_CLIENT_W_CITY   0.950149
CNT_CHILDREN                CNT_FAM_MEMBERS               0.878571
REG_REGION_NOT_WORK_REGION  LIVE_REGION_NOT_WORK_REGION   0.861861
DEF_30_CNT_SOCIAL_CIRCLE    DEF_60_CNT_SOCIAL_CIRCLE      0.859332
REG_CITY_NOT_WORK_CITY      LIVE_CITY_NOT_WORK_CITY       0.830381
```

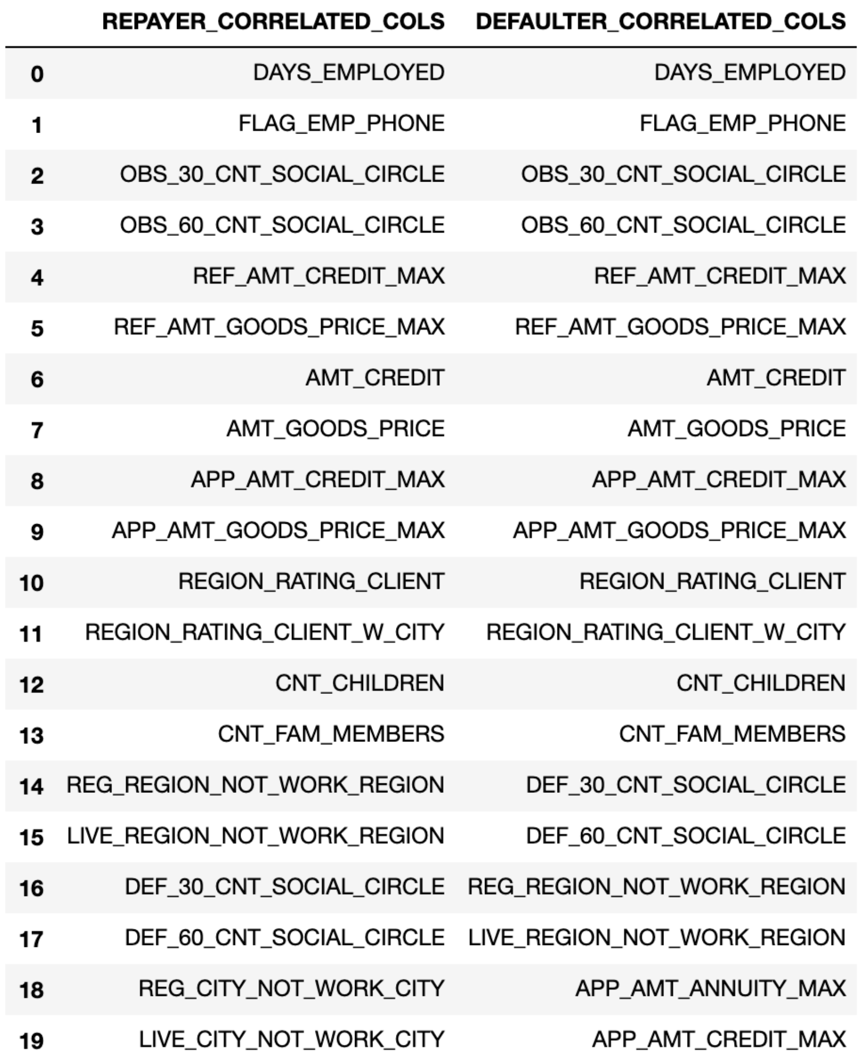
Top 10 correlated features for Target – 1 – Defaulter



```
Top Absolute Correlations for defaulter dataframe

DAYS_EMPLOYED                FLAG_EMP_PHONE                 0.999702
OBS_30_CNT_SOCIAL_CIRCLE     OBS_60_CNT_SOCIAL_CIRCLE       0.998269
REF_AMT_CREDIT_MAX           REF_AMT_GOODS_PRICE_MAX        0.990211
AMT_CREDIT                   AMT_GOODS_PRICE                0.983103
APP_AMT_CREDIT_MAX           APP_AMT_GOODS_PRICE_MAX        0.979421
REGION_RATING_CLIENT         REGION_RATING_CLIENT_W_CITY    0.956637
CNT_CHILDREN                 CNT_FAM_MEMBERS                0.885484
DEF_30_CNT_SOCIAL_CIRCLE     DEF_60_CNT_SOCIAL_CIRCLE       0.868994
REG_REGION_NOT_WORK_REGION   LIVE_REGION_NOT_WORK_REGION    0.847885
APP_AMT_ANNUITY_MAX          APP_AMT_CREDIT_MAX             0.844615
```

Top correlated features list



	REPAYER_CORRELATED_COLS	DEFAULTER_CORRELATED_COLS
0	DAYS_EMPLOYED	DAYS_EMPLOYED
1	FLAG_EMP_PHONE	FLAG_EMP_PHONE
2	OBS_30_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE
3	OBS_60_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE
4	REF_AMT_CREDIT_MAX	REF_AMT_CREDIT_MAX
5	REF_AMT_GOODS_PRICE_MAX	REF_AMT_GOODS_PRICE_MAX
6	AMT_CREDIT	AMT_CREDIT
7	AMT_GOODS_PRICE	AMT_GOODS_PRICE
8	APP_AMT_CREDIT_MAX	APP_AMT_CREDIT_MAX
9	APP_AMT_GOODS_PRICE_MAX	APP_AMT_GOODS_PRICE_MAX
10	REGION_RATING_CLIENT	REGION_RATING_CLIENT
11	REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT_W_CITY
12	CNT_CHILDREN	CNT_CHILDREN
13	CNT_FAM_MEMBERS	CNT_FAM_MEMBERS
14	REG_REGION_NOT_WORK_REGION	DEF_30_CNT_SOCIAL_CIRCLE
15	LIVE_REGION_NOT_WORK_REGION	DEF_60_CNT_SOCIAL_CIRCLE
16	DEF_30_CNT_SOCIAL_CIRCLE	REG_REGION_NOT_WORK_REGION
17	DEF_60_CNT_SOCIAL_CIRCLE	LIVE_REGION_NOT_WORK_REGION
18	REG_CITY_NOT_WORK_CITY	APP_AMT_ANNUITY_MAX
19	LIVE_CITY_NOT_WORK_CITY	APP_AMT_CREDIT_MAX

Almost all highly correlated features are same for Target variable 1 and 0 except below

- In repayer dataframe below features are highly correlated while not in defaulter dataframe

  REG_CITY_NOT_WORK_CITY - LIVE_CITY_NOT_WORK_CITY - 0.830381

  This could be because person living and working in city are highly correlated features and it could be infered that there are less chances of people with good working and living conditions to default loan
- In defaulter dataframe below features are highly correlated while not in repayer dataframe

  APP_AMT_ANNUITY_MAX - APP_AMT_CREDIT_MAX - 0.844615

  This could be because during previous loan application what was the amount credited to customer and what was the annuity are highly correlated features and it could be infered that people with high annuity amount or people with high loan/credit amount can tend to default on loan