Bank loan case study

Introduction:

There is a finance company that specializes in lending various types of loans to urban customers. The company faces a challenge: some customers who don't have a sufficient credit history take advantage of this and default on their loans.

Task:

Using Exploratory Data Analysis (EDA) to analyse patterns in the data and ensure that capable applicants are not rejected.

Issues:

When a customer applies for a loan, your company faces two risks:

1. If the applicant can repay the loan but is not approved, the company loses business.
2. If the applicant cannot repay the loan and is approved, the company faces a financial loss.

When a customer applies for a loan, there are four possible outcomes:

1. Approved: The company has approved the loan application.
2. Cancelled: The customer cancelled the application during the approval process.
3. Refused: The company rejected the loan.
4. Unused Offer: The loan was approved but the customer did not use it.

Missing values:

In application_data csv:

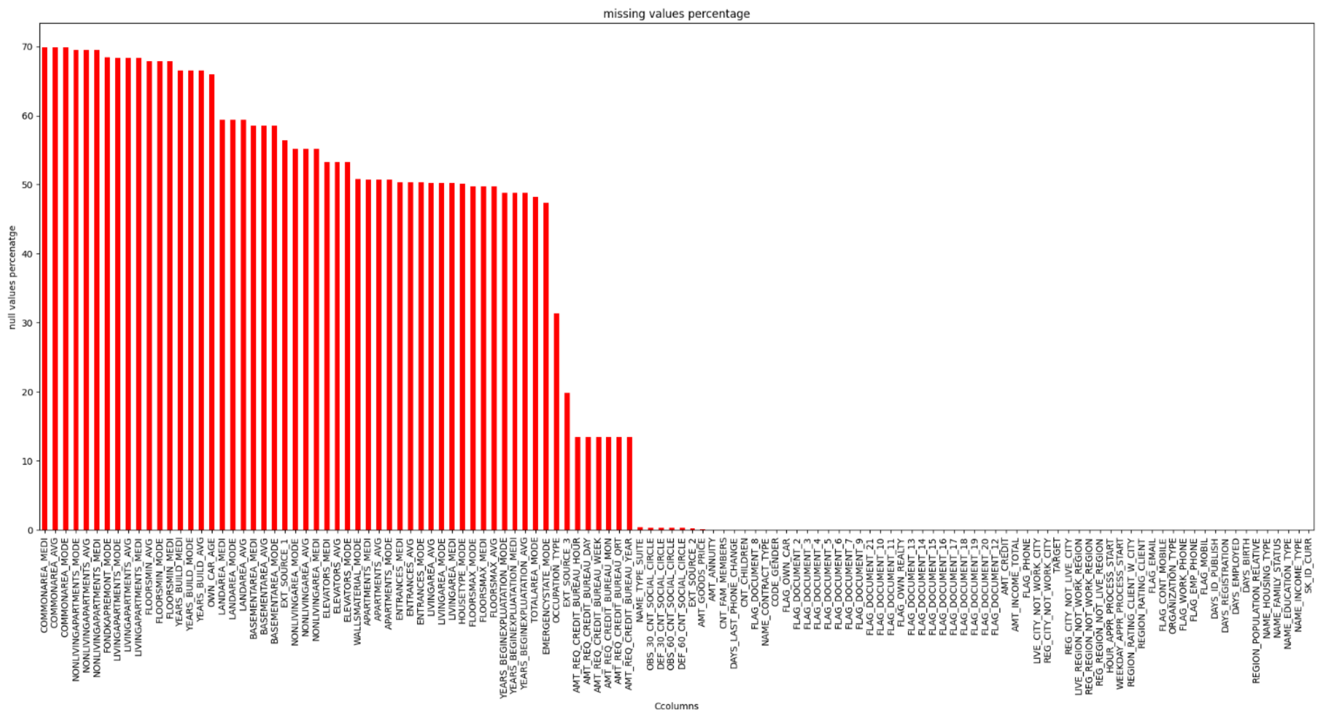
Percentage of missing values in each column:

```
COMMONAREA_MEDI                 69.87230
COMMONAREA_AVG                  69.87230
COMMONAREA_MODE                 69.87230
NONLIVINGAPARTMENTS_MODE        69.43296
NONLIVINGAPARTMENTS_AVG         69.43296
NONLIVINGAPARTMENTS_MEDI        69.43296
FONDKAPREMONT_MODE              68.38617
LIVINGAPARTMENTS_MODE           68.35495
LIVINGAPARTMENTS_AVG            68.35495
LIVINGAPARTMENTS_MEDI           68.35495
FLOORSMIN_AVG                   67.84863
FLOORSMIN_MODE                  67.84863
```

FLOORSMIN_MEDI	67.84863
YEARS_BUILD_MEDI	66.49778
YEARS_BUILD_MODE	66.49778
YEARS_BUILD_AVG	66.49778
OWN_CAR_AGE	65.99081
LANDAREA_MEDI	59.37674
LANDAREA_MODE	59.37674
LANDAREA_AVG	59.37674
BASEMENTAREA_MEDI	58.51596
BASEMENTAREA_AVG	58.51596
BASEMENTAREA_MODE	58.51596
EXT_SOURCE_1	56.38107
NONLIVINGAREA_MODE	55.17916
NONLIVINGAREA_AVG	55.17916
NONLIVINGAREA_MEDI	55.17916
ELEVATORS_MEDI	53.29598
ELEVATORS_AVG	53.29598
ELEVATORS_MODE	53.29598
WALLSMATERIAL_MODE	50.84078
APARTMENTS_MEDI	50.74973
APARTMENTS_AVG	50.74973
APARTMENTS_MODE	50.74973
ENTRANCES_MEDI	50.34877
ENTRANCES_AVG	50.34877
ENTRANCES_MODE	50.34877
LIVINGAREA_AVG	50.19333
LIVINGAREA_MODE	50.19333
LIVINGAREA_MEDI	50.19333
HOUSETYPE_MODE	50.17609
FLOORSMAX_MODE	49.76082
FLOORSMAX_MEDI	49.76082
FLOORSMAX_AVG	49.76082
YEARS_BEGINEXPLUATATION_MODE	48.78102
YEARS_BEGINEXPLUATATION_MEDI	48.78102
YEARS_BEGINEXPLUATATION_AVG	48.78102
TOTALAREA_MODE	48.26852
EMERGENCYSTATE_MODE	47.39830
OCCUPATION_TYPE	31.34555
EXT_SOURCE_3	19.82531
AMT_REQ_CREDIT_BUREAU_HOUR	13.50163
AMT_REQ_CREDIT_BUREAU_DAY	13.50163
AMT_REQ_CREDIT_BUREAU_WEEK	13.50163
AMT_REQ_CREDIT_BUREAU_MON	13.50163
AMT_REQ_CREDIT_BUREAU_QRT	13.50163
AMT_REQ_CREDIT_BUREAU_YEAR	13.50163
NAME_TYPE_SUITE	0.42015
OBS_30_CNT_SOCIAL_CIRCLE	0.33202
DEF_30_CNT_SOCIAL_CIRCLE	0.33202
OBS_60_CNT_SOCIAL_CIRCLE	0.33202
DEF_60_CNT_SOCIAL_CIRCLE	0.33202
EXT_SOURCE_2	0.21463
AMT_GOODS_PRICE	0.09040
AMT_ANNUITY	0.00390
CNT_FAM_MEMBERS	0.00065
DAYS_LAST_PHONE_CHANGE	0.00033
CNT_CHILDREN	0.00000
FLAG_DOCUMENT_8	0.00000
NAME_CONTRACT_TYPE	0.00000

```
CODE_GENDER                    0.00000
FLAG_OWN_CAR                   0.00000
FLAG_DOCUMENT_2                0.00000
FLAG_DOCUMENT_3                0.00000
FLAG_DOCUMENT_4                0.00000
FLAG_DOCUMENT_5                0.00000
FLAG_DOCUMENT_6                0.00000
FLAG_DOCUMENT_7                0.00000
FLAG_DOCUMENT_9                0.00000
FLAG_DOCUMENT_21               0.00000
FLAG_DOCUMENT_10               0.00000
FLAG_DOCUMENT_11               0.00000
FLAG_OWN_REALTY                0.00000
FLAG_DOCUMENT_13               0.00000
FLAG_DOCUMENT_14               0.00000
FLAG_DOCUMENT_15               0.00000
FLAG_DOCUMENT_16               0.00000
FLAG_DOCUMENT_17               0.00000
FLAG_DOCUMENT_18               0.00000
FLAG_DOCUMENT_19               0.00000
FLAG_DOCUMENT_20               0.00000
FLAG_DOCUMENT_12               0.00000
AMT_CREDIT                     0.00000
AMT_INCOME_TOTAL               0.00000
FLAG_PHONE                     0.00000
LIVE_CITY_NOT_WORK_CITY        0.00000
REG_CITY_NOT_WORK_CITY         0.00000
TARGET                         0.00000
REG_CITY_NOT_LIVE_CITY         0.00000
LIVE_REGION_NOT_WORK_REGION    0.00000
REG_REGION_NOT_WORK_REGION     0.00000
REG_REGION_NOT_LIVE_REGION     0.00000
HOUR_APPR_PROCESS_START        0.00000
WEEKDAY_APPR_PROCESS_START     0.00000
REGION_RATING_CLIENT_W_CITY    0.00000
REGION_RATING_CLIENT           0.00000
FLAG_EMAIL                     0.00000
FLAG_CONT_MOBILE               0.00000
ORGANIZATION_TYPE              0.00000
FLAG_WORK_PHONE                0.00000
FLAG_EMP_PHONE                 0.00000
FLAG_MOBIL                     0.00000
DAYS_ID_PUBLISH                0.00000
DAYS_REGISTRATION              0.00000
DAYS_EMPLOYED                  0.00000
DAYS_BIRTH                     0.00000
REGION_POPULATION_RELATIVE     0.00000
NAME_HOUSING_TYPE              0.00000
NAME_FAMILY_STATUS             0.00000
NAME_EDUCATION_TYPE            0.00000
NAME_INCOME_TYPE               0.00000
```

```
SK_ID_CURR                     0.00000
```




```
app.drop(['EXT_SOURCE_3','FLAG_OWN_REALTY','FLAG_OWN_CAR','EXT_SOURCE_2
','FLAG_DOCUMENT_7','FLAG_DOCUMENT_2','FLAG_DOCUMENT_3',
             'FLAG_DOCUMENT_4','FLAG_DOCUMENT_5' ,'FLAG_DOCUMENT_6'
,'FLAG_DOCUMENT_11','FLAG_DOCUMENT_8','FLAG_DOCUMENT_9',
             'FLAG_DOCUMENT_10','FLAG_DOCUMENT_12','FLAG_DOCUMENT_13','F
LAG_DOCUMENT_14','FLAG_DOCUMENT_15','FLAG_DOCUMENT_16',
             'FLAG_DOCUMENT_17','FLAG_DOCUMENT_18','FLAG_DOCUMENT_19','F
LAG_DOCUMENT_20','FLAG_DOCUMENT_21','FLAG_MOBIL',
             'FLAG_EMP_PHONE','FLAG_WORK_PHONE','FLAG_CONT_MOBILE','FLAG
_PHONE', 'FLAG_EMAIL','DAYS_LAST_PHONE_CHANGE','NAME_TYPE_SUITE'],
axis=1, inplace=True)
```

I dropped the above columns as I felt there not related to the analysis.

I removed columns with more than 35% as they cannot be used for analysis.

```
OCCUPATION_TYPE            96391
AMT_REQ_CREDIT_BUREAU_YEAR    41519
AMT_REQ_CREDIT_BUREAU_QRT     41519
AMT_REQ_CREDIT_BUREAU_MON     41519
AMT_REQ_CREDIT_BUREAU_WEEK    41519
AMT_REQ_CREDIT_BUREAU_DAY     41519
AMT_REQ_CREDIT_BUREAU_HOUR    41519
OBS_60_CNT_SOCIAL_CIRCLE      1021
OBS_30_CNT_SOCIAL_CIRCLE      1021
DEF_30_CNT_SOCIAL_CIRCLE      1021
DEF_60_CNT_SOCIAL_CIRCLE      1021
AMT_GOODS_PRICE            278
AMT_ANNUITY             12
CNT_FAM_MEMBERS              2
HOUR_APPR_PROCESS_START        0
```

```
ORGANIZATION_TYPE               0
LIVE_CITY_NOT_WORK_CITY         0
REG_CITY_NOT_WORK_CITY          0
REG_CITY_NOT_LIVE_CITY          0
LIVE_REGION_NOT_WORK_REGION     0
REG_REGION_NOT_WORK_REGION      0
REG_REGION_NOT_LIVE_REGION      0
SK_ID_CURR                      0
WEEKDAY_APPR_PROCESS_START      0
REGION_RATING_CLIENT_W_CITY     0
NAME_CONTRACT_TYPE              0
CODE_GENDER                     0
CNT_CHILDREN                    0
AMT_INCOME_TOTAL                0
AMT_CREDIT                      0
NAME_INCOME_TYPE                0
NAME_EDUCATION_TYPE             0
NAME_FAMILY_STATUS              0
NAME_HOUSING_TYPE               0
REGION_POPULATION_RELATIVE      0
DAYS_BIRTH                      0
DAYS_EMPLOYED                   0
DAYS_REGISTRATION               0
DAYS_ID_PUBLISH                 0
TARGET                          0
REGION_RATING_CLIENT            0
dtype: int64
```

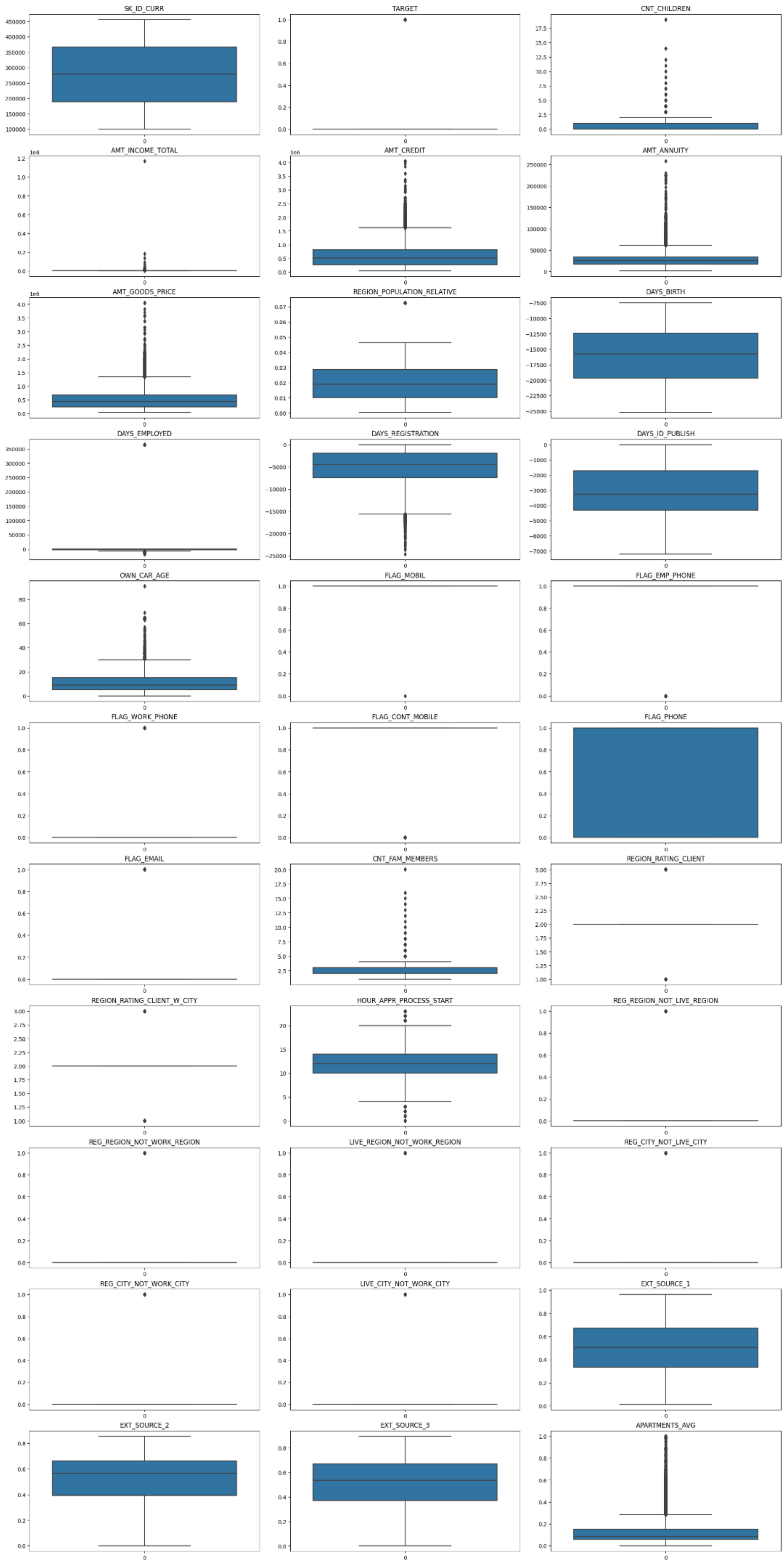
Outliers:

After analyzing the outliers I have decided to use mode for

```
OBS_60_CNT_SOCIAL_CIRCLE      1021
OBS_30_CNT_SOCIAL_CIRCLE      1021
DEF_30_CNT_SOCIAL_CIRCLE      1021
DEF_60_CNT_SOCIAL_CIRCLE      1021
```

And median for the following

```
AMT_REQ_CREDIT_BUREAU_YEAR    41519
AMT_REQ_CREDIT_BUREAU_QRT     41519
AMT_REQ_CREDIT_BUREAU_MON     41519
AMT_REQ_CREDIT_BUREAU_WEEK    41519
AMT_REQ_CREDIT_BUREAU_DAY     41519
AMT_REQ_CREDIT_BUREAU_HOUR    41519
AMT_ANNUITY                   12
CNT_FAM_MEMBERS               2
```

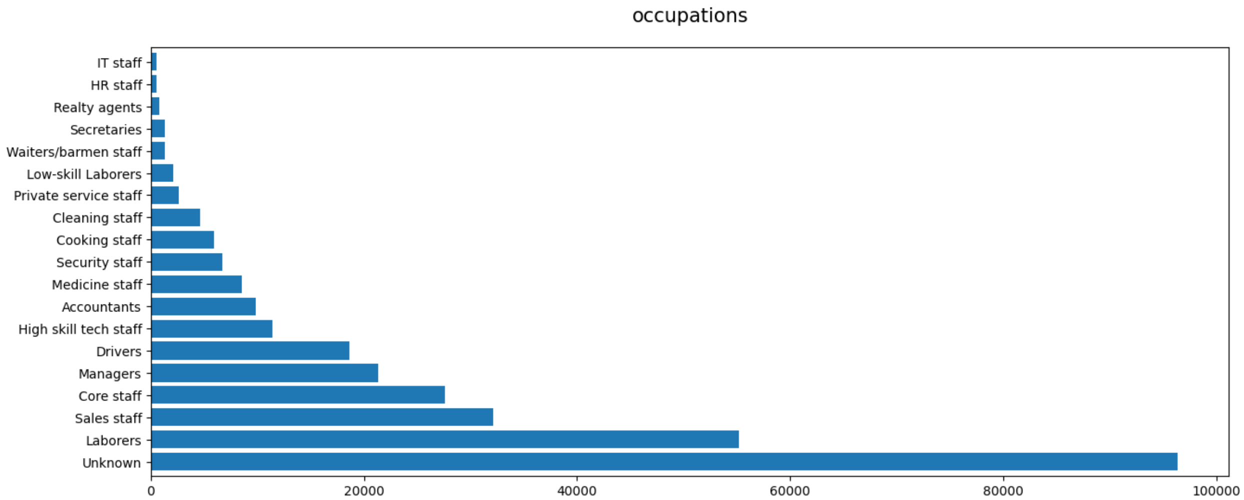

SK_ID_CURR
TARGET
CNT_CHILDREN
AMT_INCOME_TOTAL
AMT_CREDIT
AMT_ANNUITY
AMT_GOODS_PRICE
REGION_POPULATION_RELATIVE
DAYS_BIRTH
DAYS_EMPLOYED
DAYS_REGISTRATION
DAYS_ID_PUBLISH
OWN_CAR_AGE
FLAG_MOBIL
FLAG_EMP_PHONE
FLAG_WORK_PHONE
FLAG_CONT_MOBILE
FLAG_PHONE
FLAG_EMAIL
CNT_FAM_MEMBERS
REGION_RATING_CLIENT
REGION_RATING_CLIENT_W_CITY
HOUR_APPR_PROCESS_START
REG_REGION_NOT_LIVE_REGION
REG_REGION_NOT_WORK_REGION
LIVE_REGION_NOT_WORK_REGION
REG_CITY_NOT_LIVE_CITY
REG_CITY_NOT_WORK_CITY
LIVE_CITY_NOT_WORK_CITY
EXT_SOURCE_1
EXT_SOURCE_2
EXT_SOURCE_3
APARTMENTS_AVG

As AMT_GOODS_PRICE will be closer to AMT_CREDIT I replaced the empty values with the corresponding ones.

occupation type had the highest number of missing values

I used 'unkown' to fill the empty cells.

The graph shows the different occupations distribution.



And also converted the below values from negative to their absolute values

```
'DAYS_BIRTH','DAYS_EMPLOYED','DAYS_REGISTRATION','DAYS_ID_PUBLISH'
```

In the gender column as there are only 4 rows with XNA entries and most of the entries are females I replaced those 4 rows with female entry.

I aggregated the INCOME_GROUP and CREDIT_GROUP into low, medium, high, very high groups for better analysis.

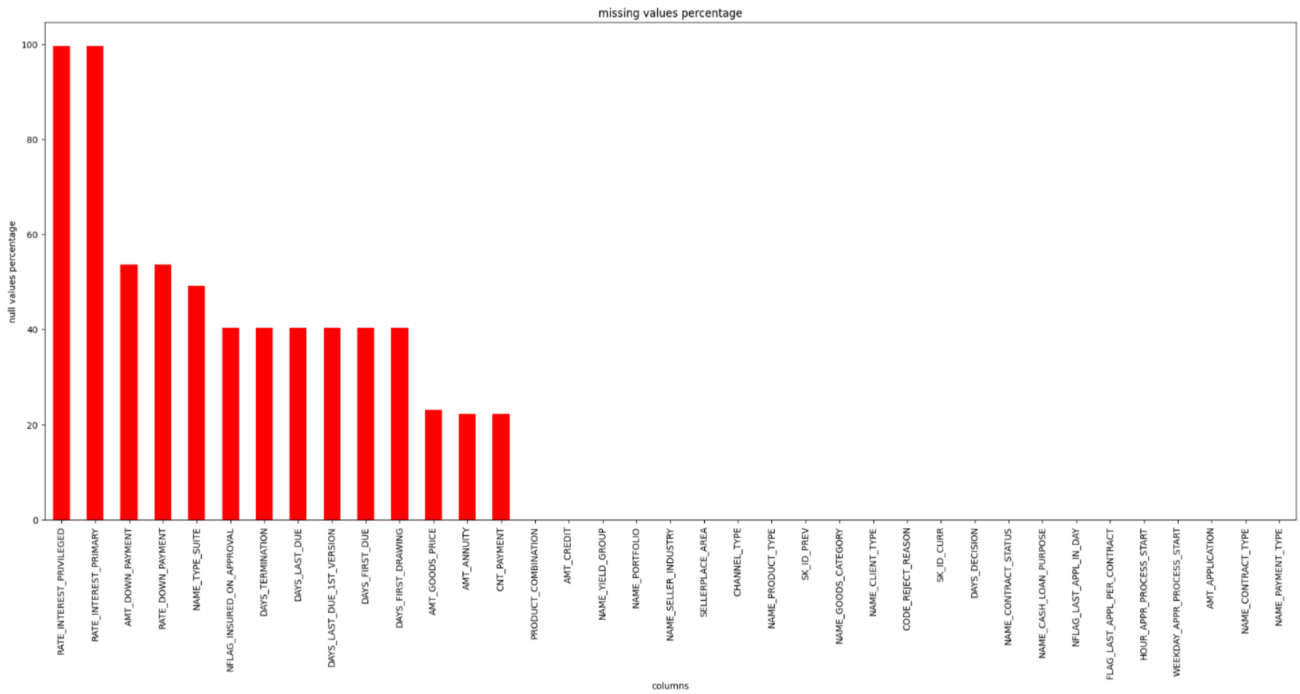
I also aggregated the AGE_GROUP for better analysis.

I did the similar data preprocessing for the previous_data.csv as well

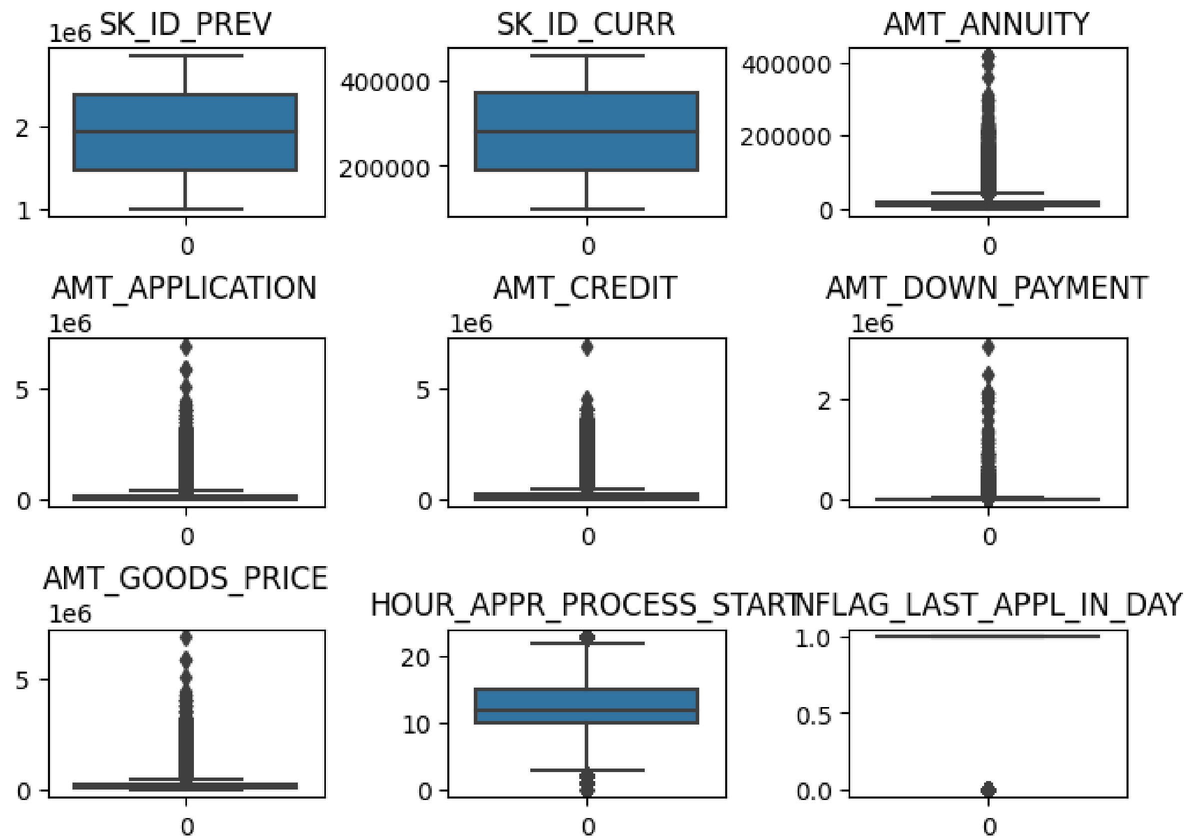
It had percentage of missing values in each columns as shown

```
RATE_INTEREST_PRIVILEGED        99.64370
RATE_INTEREST_PRIMARY           99.64370
AMT_DOWN_PAYMENT                53.63648
RATE_DOWN_PAYMENT               53.63648
NAME_TYPE_SUITE                 49.11975
NFLAG_INSURED_ON_APPROVAL       40.29813
DAYS_TERMINATION                40.29813
DAYS_LAST_DUE                   40.29813
DAYS_LAST_DUE_1ST_VERSION       40.29813
DAYS_FIRST_DUE                  40.29813
DAYS_FIRST_DRAWING              40.29813
AMT_GOODS_PRICE                 23.08177
AMT_ANNUITY                     22.28667
CNT_PAYMENT                     22.28637
PRODUCT_COMBINATION              0.02072
```

```
AMT_CREDIT                          0.00006
NAME_YIELD_GROUP                    0.00000
NAME_PORTFOLIO                      0.00000
NAME_SELLER_INDUSTRY                0.00000
SELLERPLACE_AREA                    0.00000
CHANNEL_TYPE                        0.00000
NAME_PRODUCT_TYPE                   0.00000
SK_ID_PREV                          0.00000
NAME_GOODS_CATEGORY                 0.00000
NAME_CLIENT_TYPE                    0.00000
CODE_REJECT_REASON                  0.00000
SK_ID_CURR                          0.00000
DAYS_DECISION                       0.00000
NAME_CONTRACT_STATUS                0.00000
NAME_CASH_LOAN_PURPOSE              0.00000
NFLAG_LAST_APPL_IN_DAY              0.00000
FLAG_LAST_APPL_PER_CONTRACT         0.00000
HOUR_APPR_PROCESS_START             0.00000
WEEKDAY_APPR_PROCESS_START          0.00000
AMT_APPLICATION                     0.00000
NAME_CONTRACT_TYPE                  0.00000
NAME_PAYMENT_TYPE                   0.00000
```



I removed the HOUR_APPR_PROCESS_START and NAME_TYPE_SUITE as I felt they weren't relevant for the analysis.

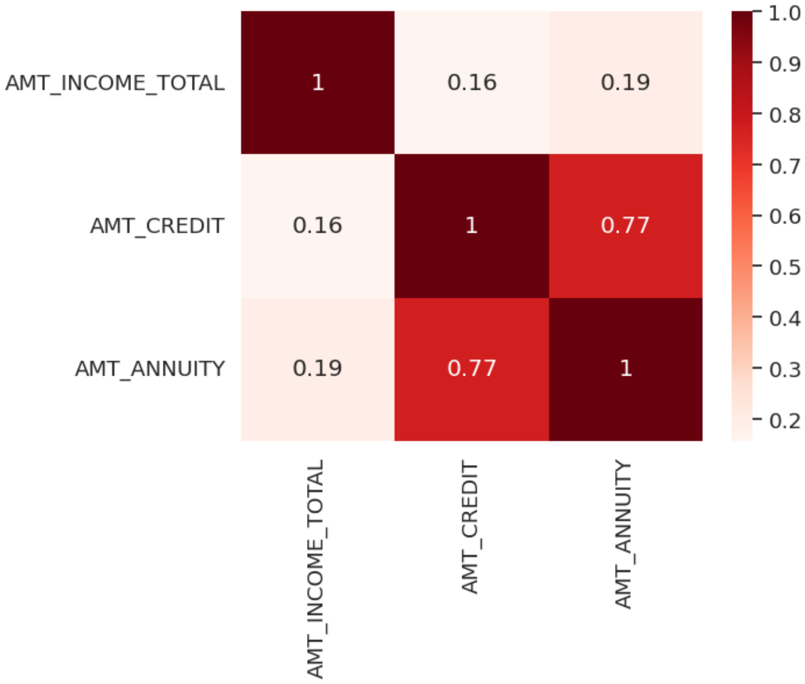


Similar to the application_data.csv. I used median to fill the empty cells in AMT_ANNUITY, CNT_PAYMENT, PRODUCT_COMBINATION, AMT_CREDIT

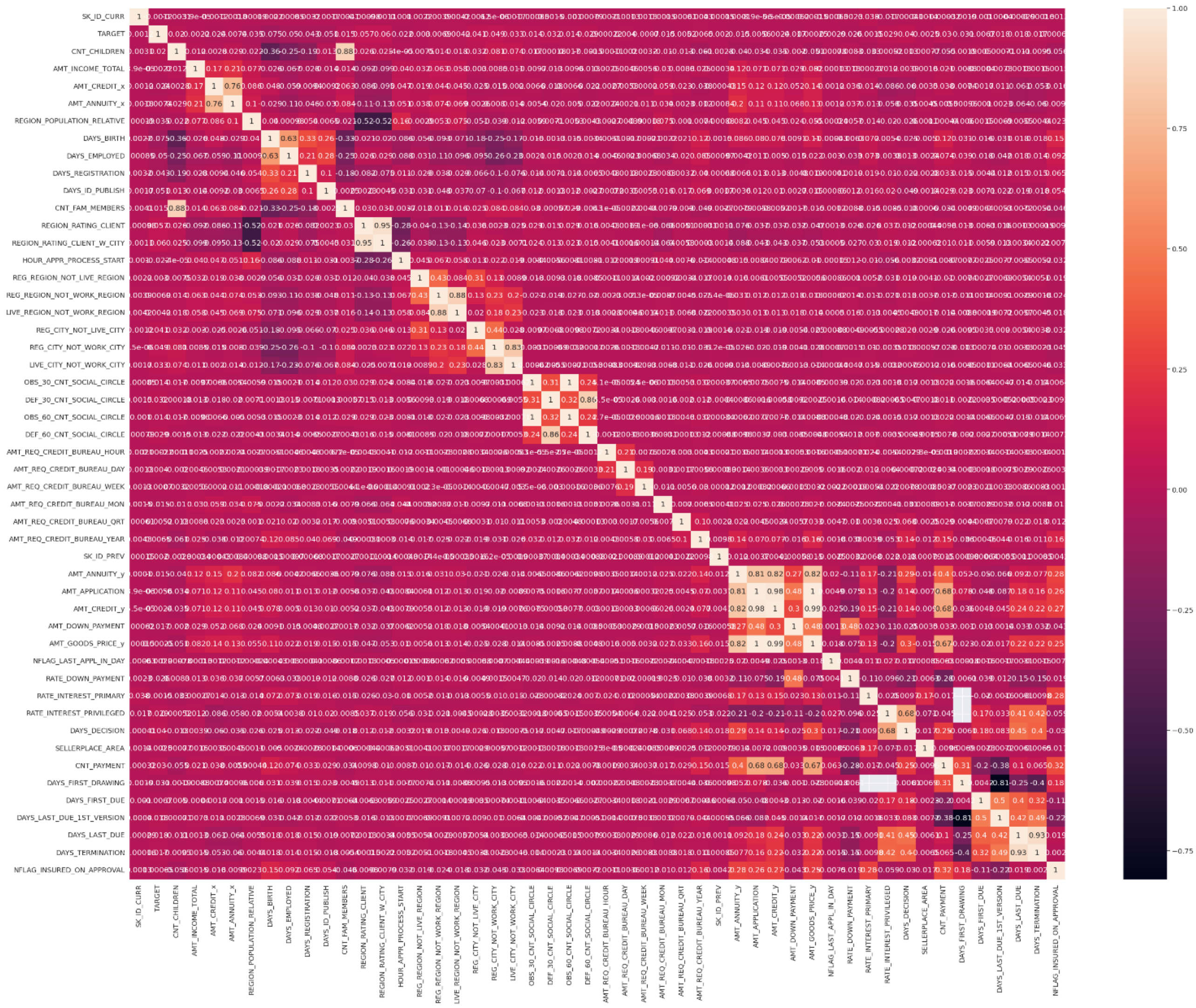
I used AMT_CREDIT values to fill the empty values in AMT_GOOD_PRICE

Heatmaps:

I used heatmaps to find the correlation between the columns in the separate datasets and the merged dataset.

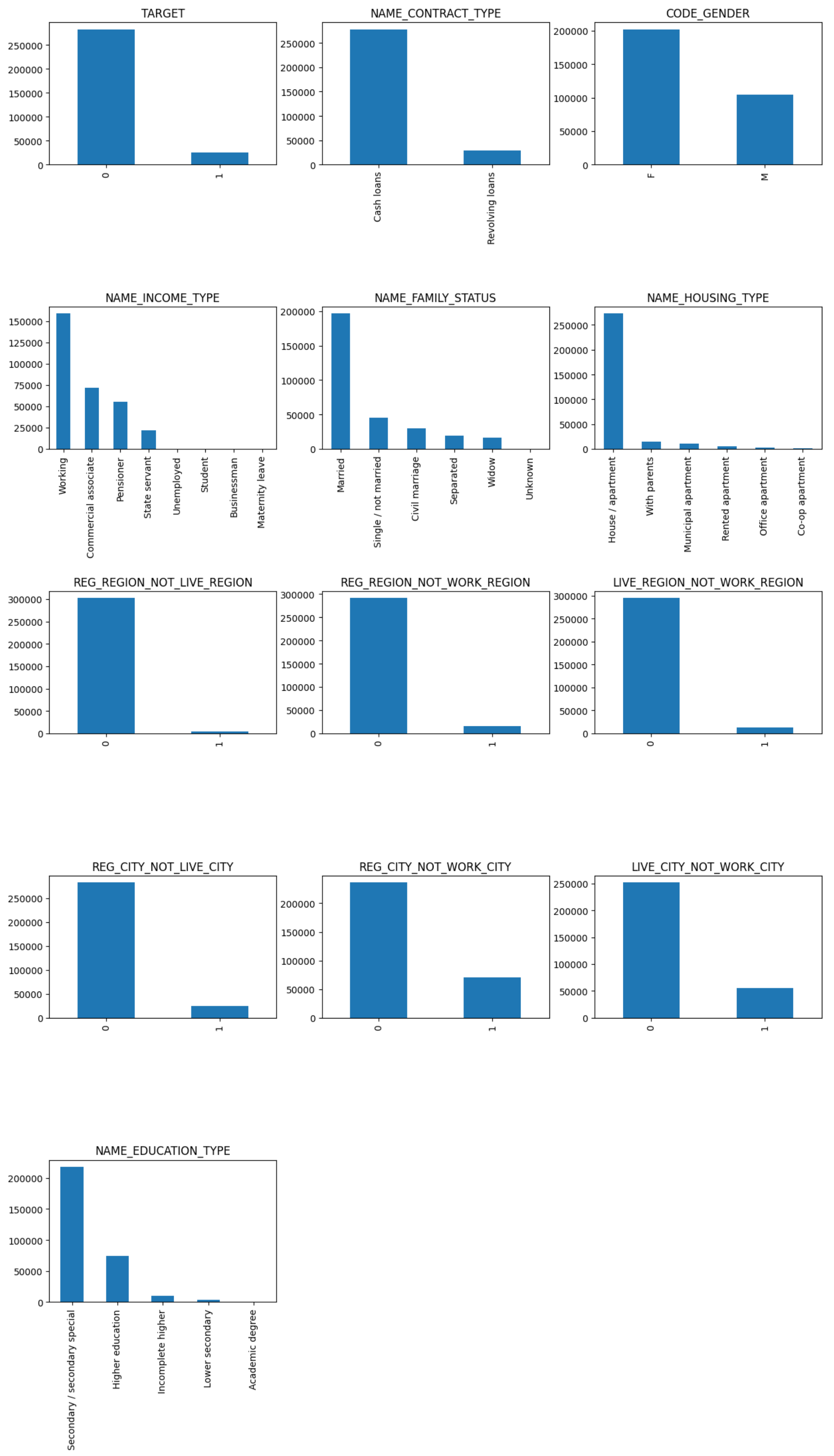

AMT_INCOME_TOTAL
AMT_CREDIT
AMT_ANNUITY
1
0.16
0.19
0.16
1
0.77
0.19
0.77
1
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2


AMT_CREDIT
AMT_ANNUITY
AMT_APPLICATION
1
0.81
0.98
0.81
1
0.81
0.98
0.81
1
1.000
0.975
0.950
0.925
0.900
0.875
0.850
0.825



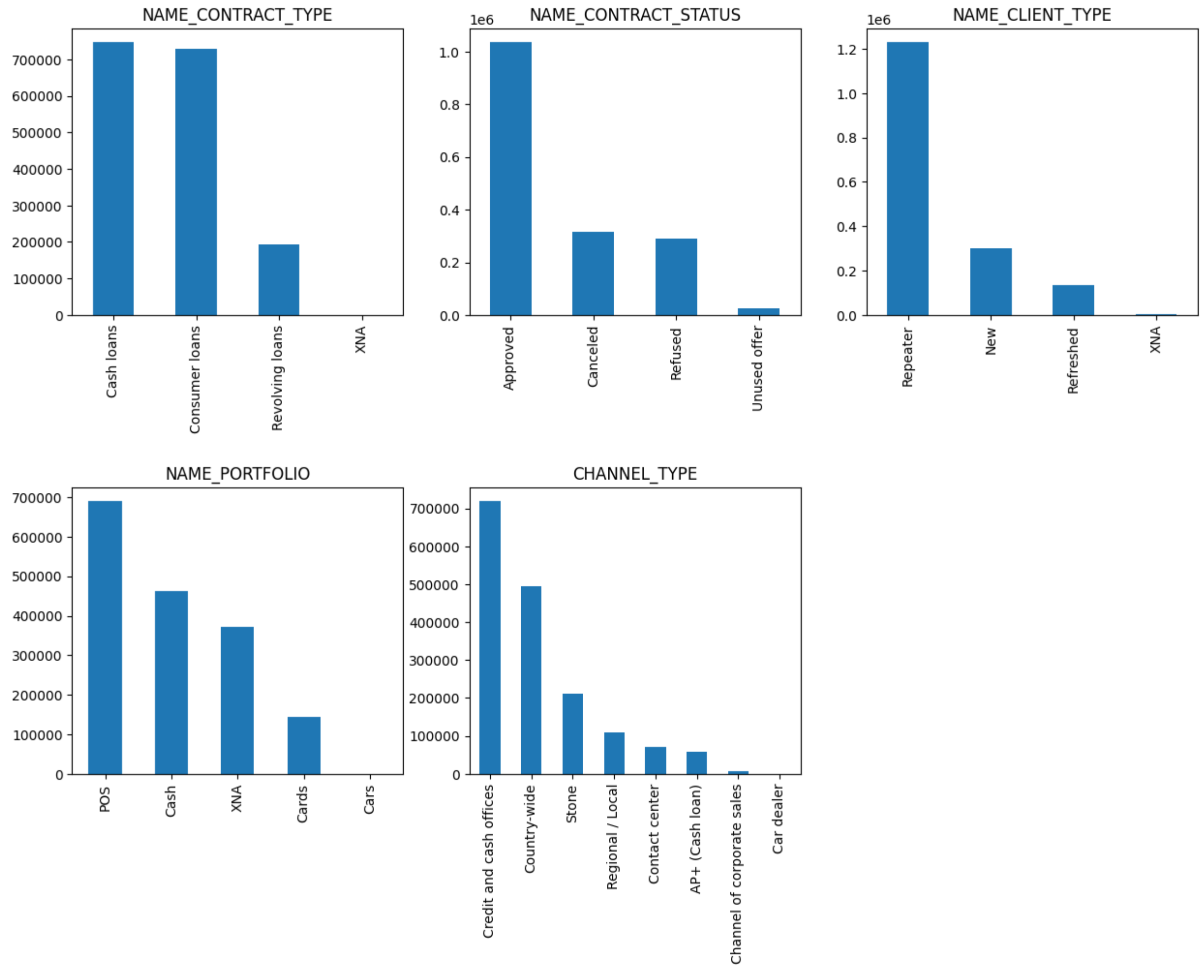
Data imbalance:

Data imbalance in application_data.csv



TARGET
0
1
NAME_CONTRACT_TYPE
Cash loans
Revolving loans
CODE_GENDER
F
M
NAME_INCOME_TYPE
Working
Commercial associate
Pensioner
State servant
Unemployed
Student
Businessman
Maternity leave
NAME_FAMILY_STATUS
Married
Single / not married
Civil marriage
Separated
Widow
Unknown
NAME_HOUSING_TYPE
House / apartment
With parents
Municipal apartment
Rented apartment
Office apartment
Co-op apartment
REG_REGION_NOT_LIVE_REGION
REG_REGION_NOT_WORK_REGION
LIVE_REGION_NOT_WORK_REGION
REG_CITY_NOT_LIVE_CITY
REG_CITY_NOT_WORK_CITY
LIVE_CITY_NOT_WORK_CITY
NAME_EDUCATION_TYPE
Secondary / secondary special
Higher education
Incomplete higher
Lower secondary
Academic degree

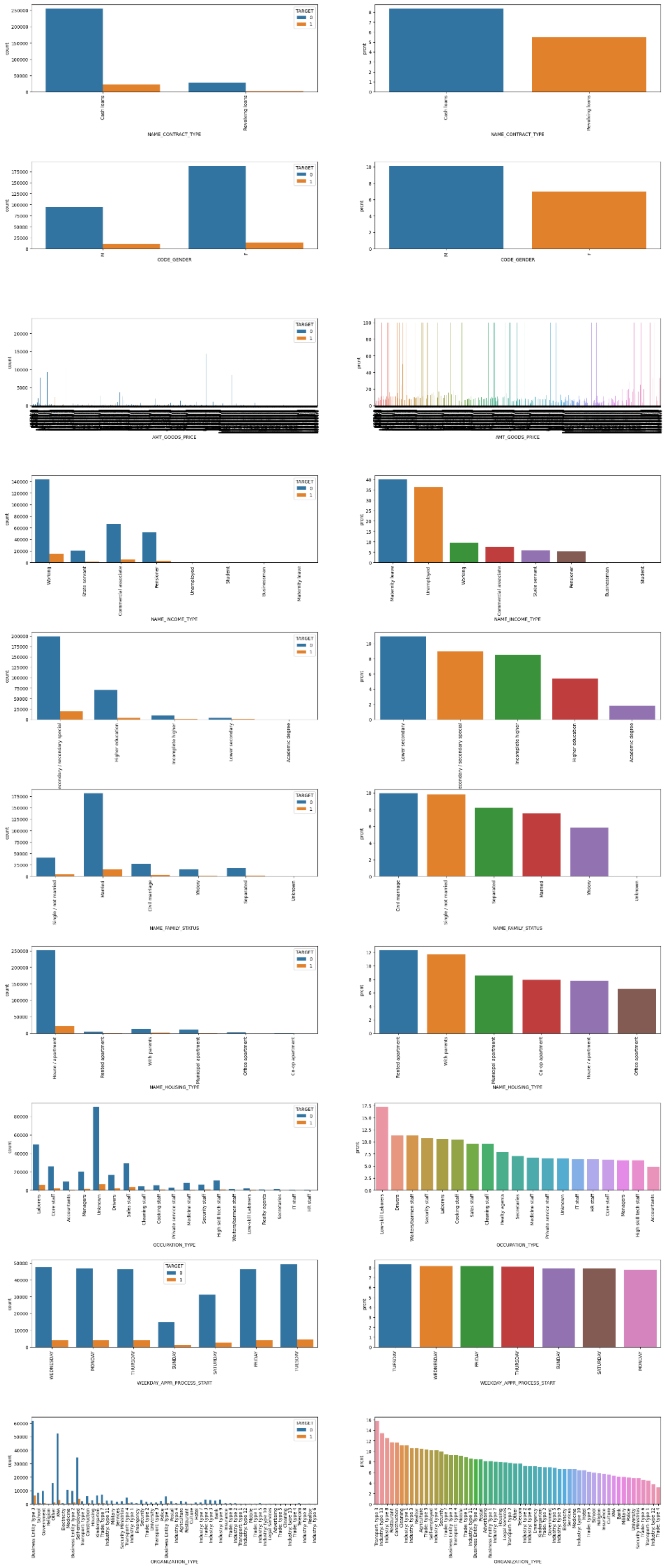
Data imbalance in previous_application_data.csv



Univariate analysis:

Univariate analysis involves the examination of a single variable at a time. It focuses on describing and summarizing the distribution of values within that variable.

The main goal is to understand the characteristics of a single variable, such as central tendency, dispersion, and shape of the distribution.



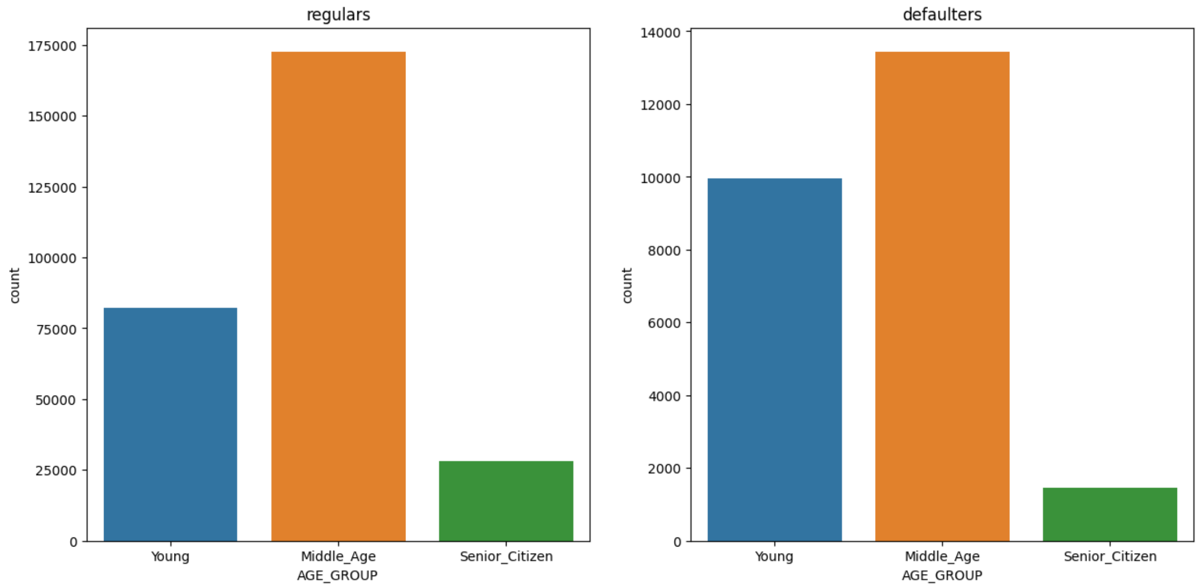
TARGET
count
prcnt
NAME_CONTRACT_TYPE
Cash loans
Revolving loans
CODE_GENDER
AMT_GOODS_PRICE
NAME_INCOME_TYPE
Working
State servant
Commercial associate
Pensioner
Unemployed
Student
Businessman
Maternity leave
Lower secondary
Secondary / secondary special
Higher education
Incomplete higher
Academic degree
NAME_FAMILY_STATUS
Single / not married
Married
Civil marriage
Widow
Separated
Unknown
NAME_HOUSING_TYPE
House / apartment
Rented apartment
With parents
Municipal apartment
Office apartment
Co-op apartment
OCCUPATION_TYPE
WEEKDAY_APPR_PROCESS_START
WEDNESDAY
MONDAY
THURSDAY
SUNDAY
SATURDAY
FRIDAY
TUESDAY
ORGANIZATION_TYPE

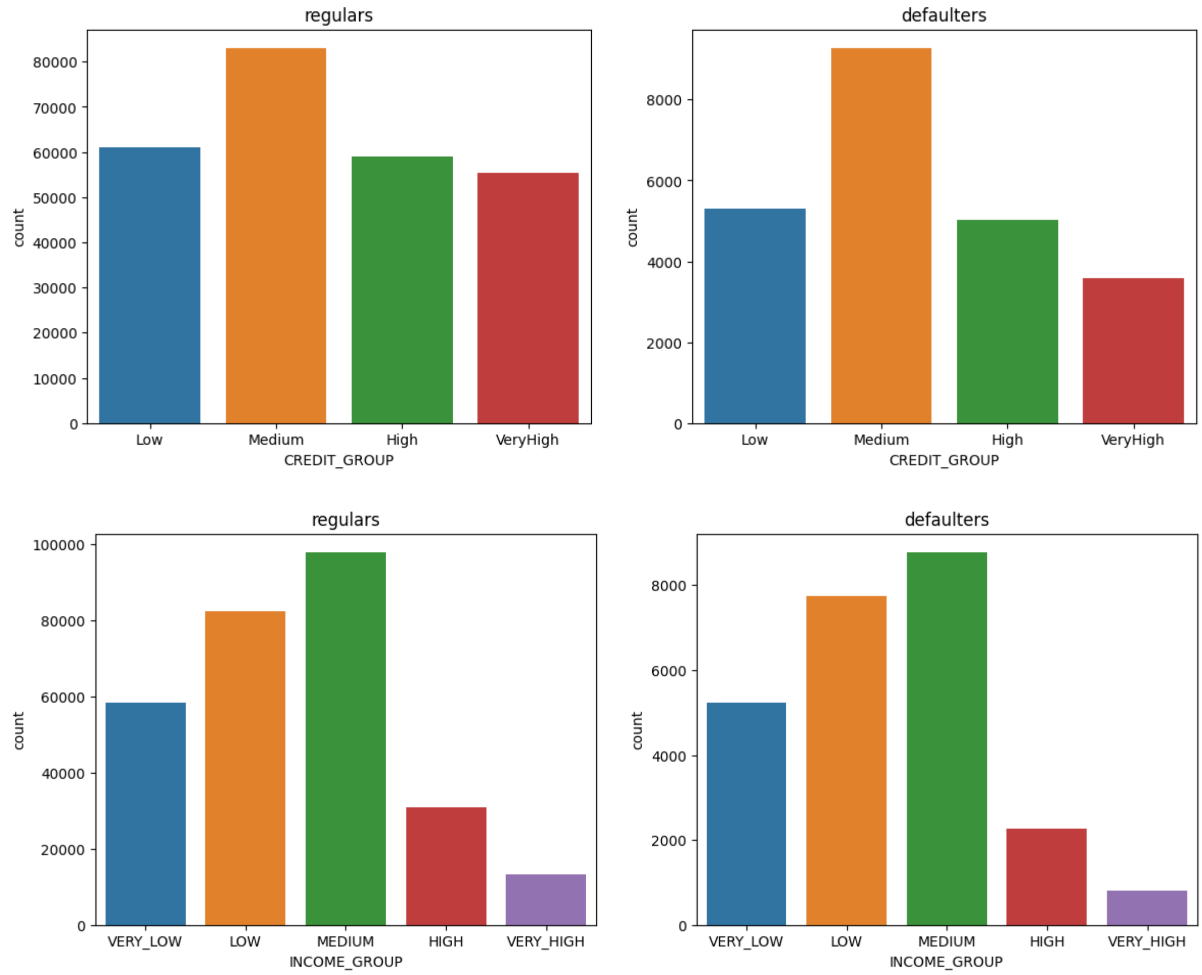
- People who have cash loans are less likely to be defaulters.
- Females are less likely to be defaulters compared to males as when compared many females are regulars.
- Working professionals are very less likely to default compared to others as we have more data on them and the graphs indicate.
- Commercial Associate, pensioner and state servant are less likely to be defaulters.
- Secondary/Secondary Special are very less likely to be defaulters.
- People with higher education are less likely to be defaulters.
- Married people are very less likely to default. And they are the ones who took the most loans compared to others.
- People who own a house or an apartment are very less likely to be defaulters.
- People in business entity type 3 and XNA are very less likely to be defaulters.
- People who are self employed are less likely to be defaulters.

Segmented univariate analysis:

Segmented univariate analysis is an extension of univariate analysis where the data is divided into subgroups or segments based on another variable, and then univariate analysis is performed on each segment.

This type of analysis helps in understanding how the distribution of a variable varies across different segments. It allows for a more nuanced exploration of the data, revealing potential patterns or differences that may not be apparent in a global univariate analysis.



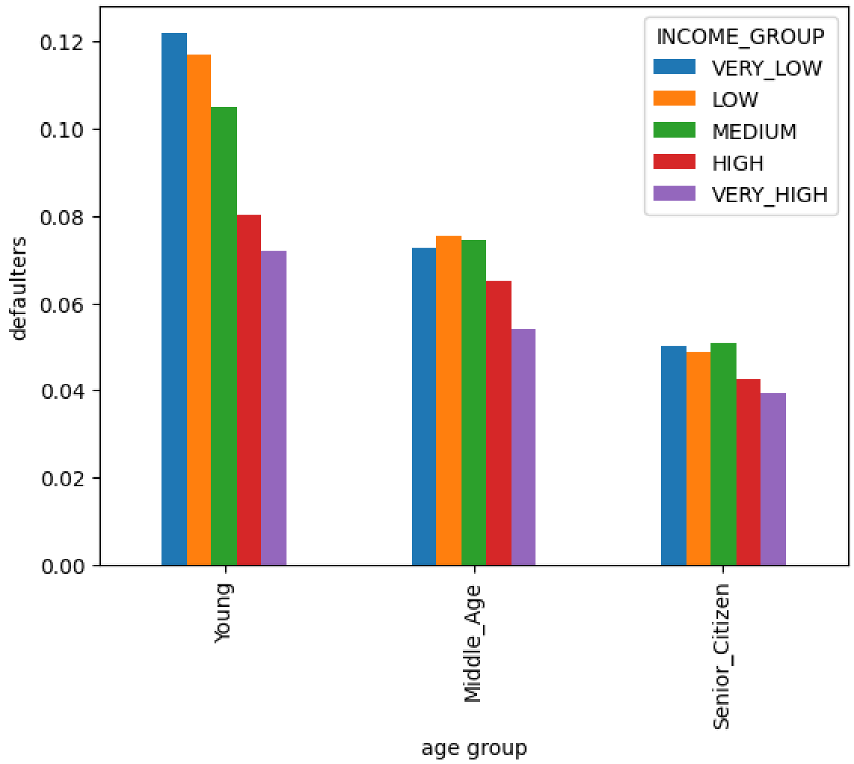


- Middle aged people and senior citizens are less likely to be defaulters compared to young people.
- In case of CREDIT_GROUP it is hard to predict as all groups are almost same in comparison of defaulters and regulars. But we can say people with very high income are less likely to be defaulters.
- In case of INCOME_GROUP also it is hard to predict as the ratios are similar.

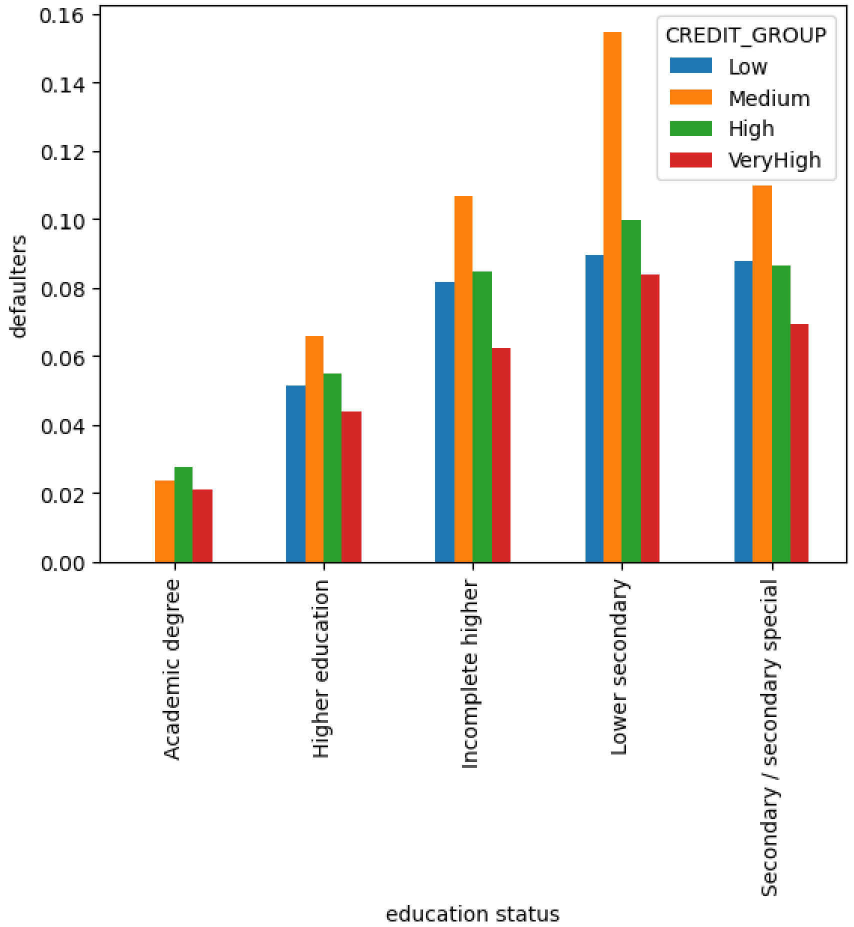
Bivariate analysis:

Bivariate analysis involves the simultaneous analysis of two variables to determine if there is a relationship or association between them.

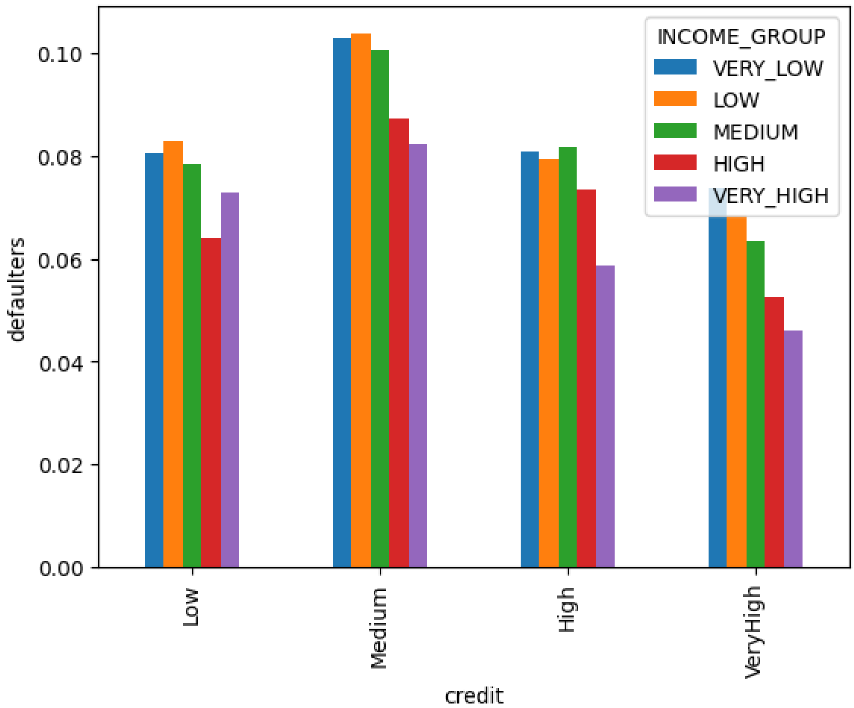
The main goal is to explore how changes in one variable are related to changes in another.

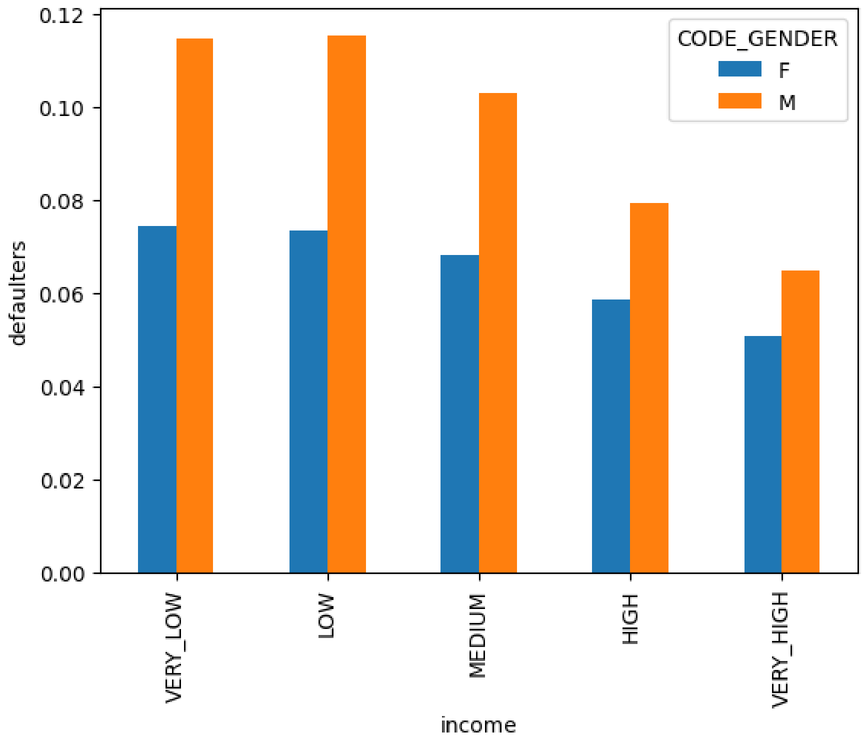


INCOME_GROUP
VERY_LOW
LOW
MEDIUM
HIGH
VERY_HIGH
defaulters
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Young
Middle_Age
Senior_Citizen
age group



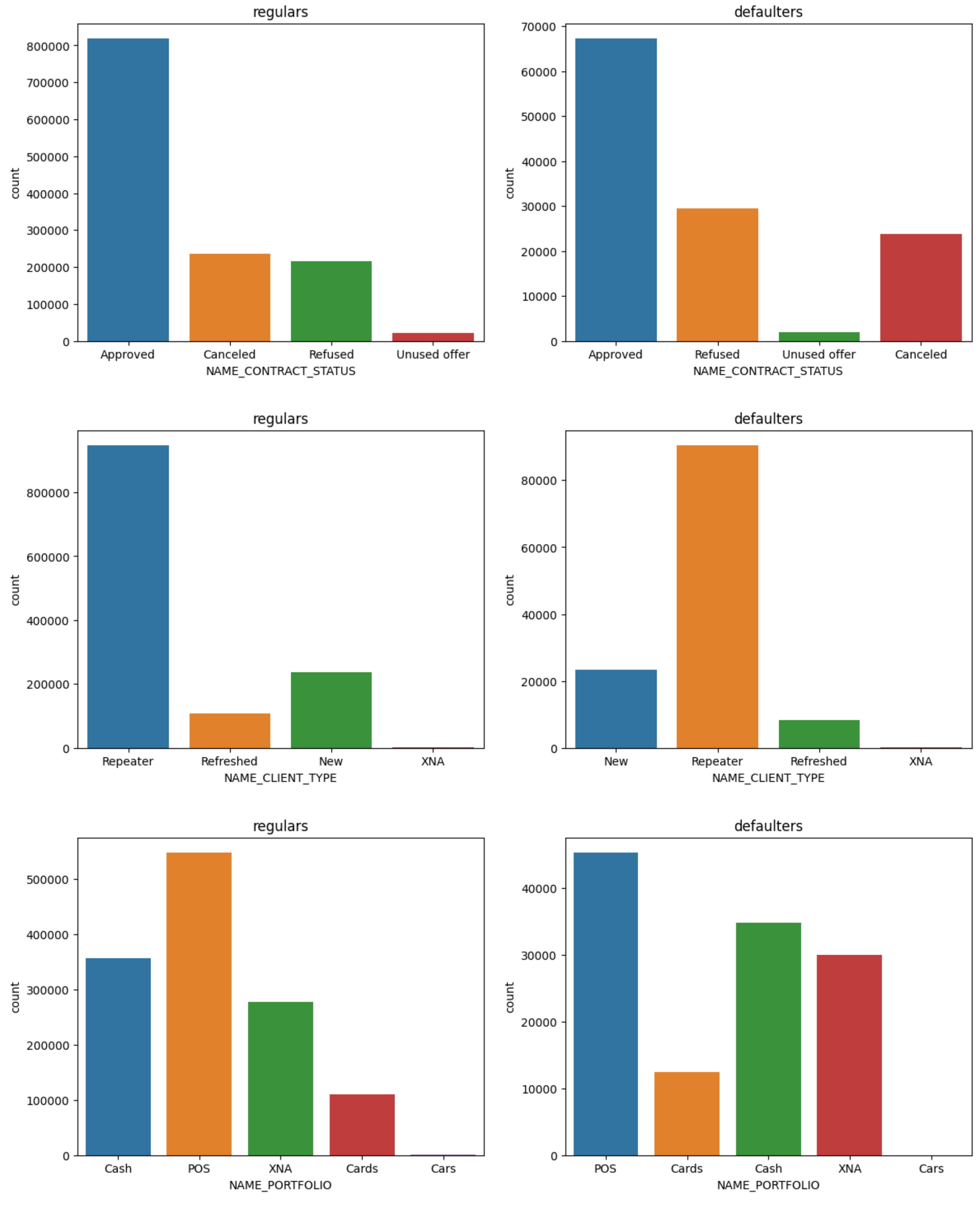
CREDIT_GROUP
Low
Medium
High
VeryHigh
defaulters
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Academic degree
Higher education
Incomplete higher
Lower secondary
Secondary / secondary special
education status





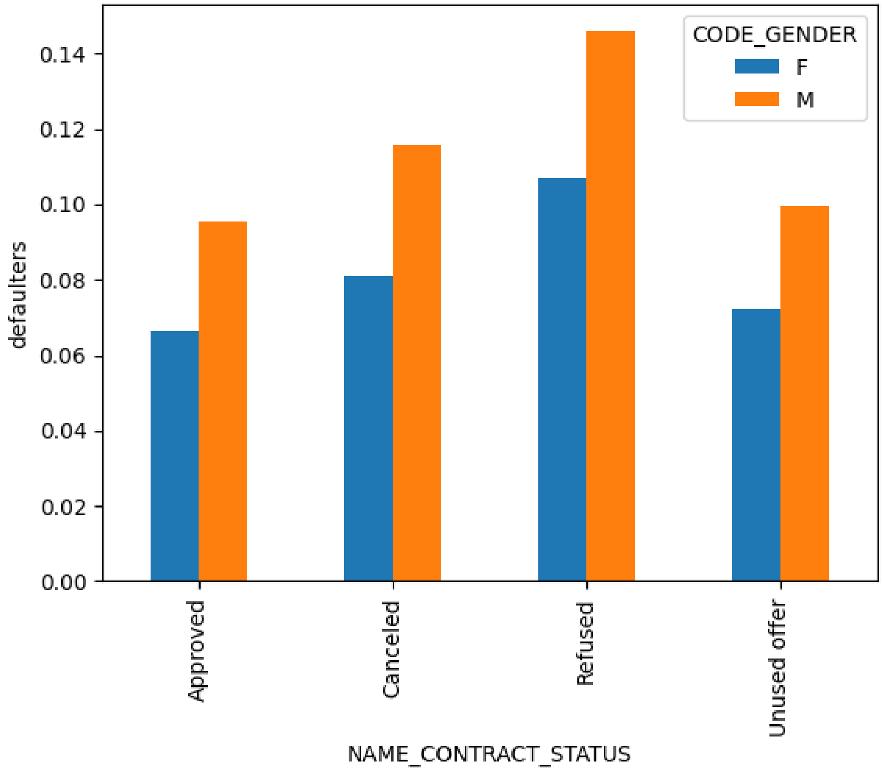
- People who are young and have low income are more likely to be defaulters.
- People who have lower secondary education and medium credit are likely to be defaulters.
- People who have incomplete higher education or secondary or secondary special education and in medium credit group are also likely to be defaulters.
- People who are medium credit group and medium income are a bit likely to be defaulters. – it is hard to predict from the graph though.
- Males with low and very low income are more likely to be defaulters.

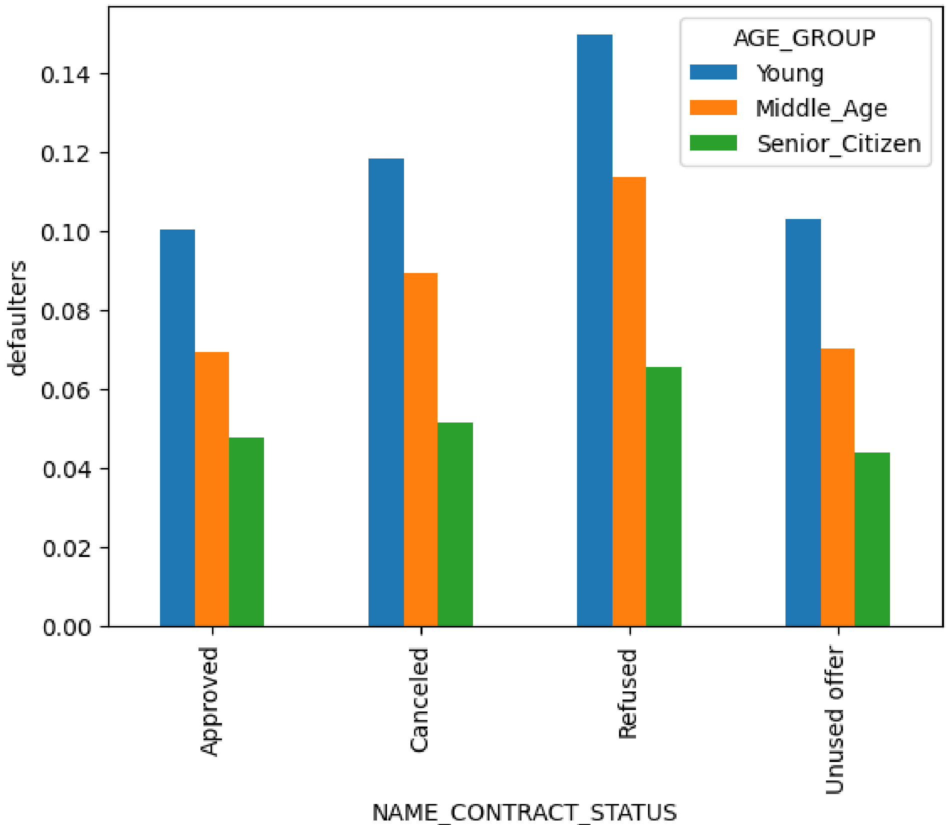
Univariate analysis of merged dataset

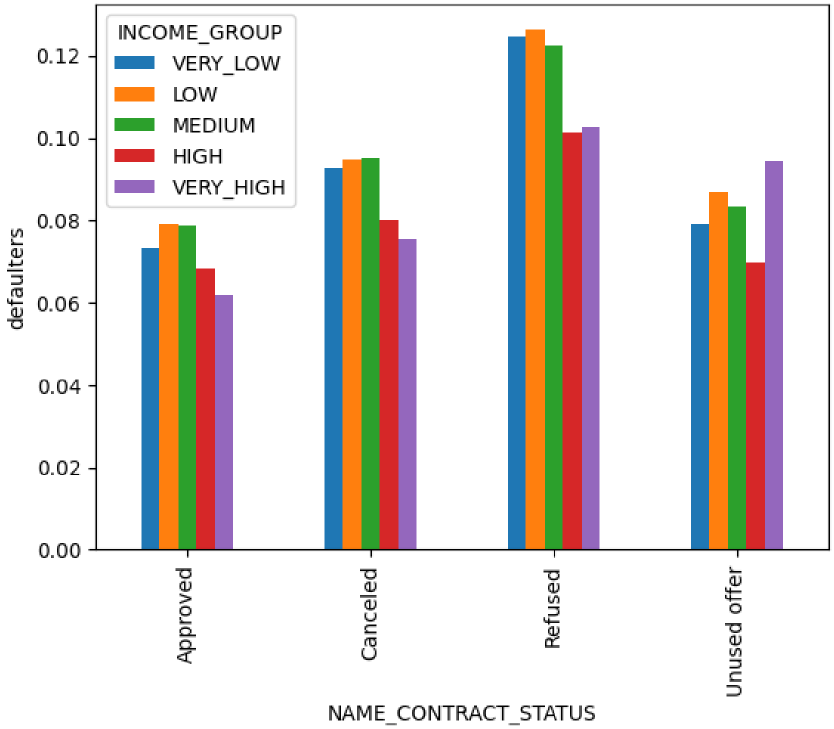


- People with approved contract status are high in both defaulters and non-defaulters. – it is hard to predict from the graph.
- Repeaters are more likely to be defaulters.
- People with POS as the name portfolio are more likely to be defaulters.

Bivariate analysis of merged dataframe







- Males are more likely to be defaulters and to be refused.
- Young people are more likely to be defaulters and to be refused.
- People with low and very low income are likely to be defaulters and to be refused.

Remarks:

The bank needs to work on the missing data as there were many columns with very high missing percentage of entries.

The bank is giving loans to mostly middle-aged group, medium income group, and people with POS name portfolio.

LINK:

https://drive.google.com/file/d/1STInucbVbOPNikFFccflKU6uL3M5R_fh/view?usp=sharing

https://colab.research.google.com/drive/1vPzgIlHkD9Nw7v60D8h8Jz8YFXtoTOdM?usp=sharing