Bank Loan Case Study

- By Sagir Mehmood

Kindly download my notebook: https://drive.google.com/file/d/1-q8PJEHd9PKAI0_ebu2pQU1QY4hNFm4x/view?usp=sharing

Q1. Present the overall approach of the analysis. Mention the problem statement and the analysis approach briefly.

Problem Statement:

Objectives:

It aims to identify patterns that 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 the 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.

Data & descriptions :

- Here I am provided with 3 datasets, they are previous loan application data, application data, and column descriptions.
- The application data have 122 coloumns,application_data.csv contains all the information of the client at the time of application. The data is about whether a client has payment difficulties.
- The previous loan application data have 37 columns, previous_application.csv contains information about the client's previous loan data. It contains the data on whether the previous application had been approved, Canceled, Refused, or Unused offer.
- columns_descrption.csv is a data dictionary that describes the meaning of the variables.

<u>Analysis approach:</u>

Here I am explaining my analysis approach below steps:

- STEP 1: Data importing and understanding the data 1st, I imported the two datasets(application_data.csv & previous_application.csv, then I thoroughly read the columns_descrption.csv to understand every column of both the datasets.
- STEP 2: Data cleaning
    - i. 1st I dropped all the columns which are not necessary for this case study
    - ii. Then, I looked for data which are wrongly labeled and corrected it
    - iii. Then, I looked for null values and replaced them with appropriate measurements.
    - iv. Checked outliers. NOTE: Here most of the time I have used Q1 as the 20th percentile and Q3 as the 80th percentile of the data.
- STEP 3: EDA In the EDA process I performed some univariate analysis while performing bivariate analysis mainly I analyzed which segment of applicants faced payment difficulties, and which loan applications are likely to be approved, Canceled, Refused, or Unused offers.

Tech-Stack Used: MS Excel & Jupyter Notebook.

Insights & Result:

- application_data.csv data:
    - Shape(307511,122)
    - 106 numeric columns
    - 16 categorical columns
    - Target column name: "TARGET"

- previous_application.csv data :
    - Shape: (1048575,37)
    - 21 numeric columns
    - 16 categorical columns
    - Target column name: "NAME_CONTRACT_STATUS"

Data cleaning:

- In both datasets, it is observable that some columns contain data about days and some of the days can be seen in a negative form, I converted them to a positive form using the MOD function.

- I can see that in both datasets few columns are labeled as numeric, such as in the application data, the TARGET column is labeled as numeric, whereas the column descriptions file says that "Target variable (1 - client with payment difficulties: he/she had late payment more than X days on at least one of the first Y installments of the loan in our sample, 0 - all other cases)", after reading the description one can easily understand that this column must be stored in object form. There are a few more columns that I converted to objects from integers, they are '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' from the application data and NFLAG_LAST_APPL_IN_DAY from the previous application data.

- In both datasets exist a few columns which are not necessary for solving this case study, so I dropped them all from the datasets.

Q2. Identify the missing data and use the appropriate method to deal with it. (Remove columns/or replace them with an appropriate value).

Null percentage = Number of missing values / Number of total rows

application_data.csv :

1. In the application data, 50 columns have more than 20% null values, I dropped them all.
2. After dropping these 50 columns the shape of the application data is (307511,41).
3. The remaining 41 columns also contain some (less than 20%) null values.

4. Among these 41 columns, I replaced the null values of categoric columns with mode. For the numeric column, which columns have outliers (Q1: 25^{th} percentile, Q3: 75^{th} percentile) I replaced their null values with median, as the median does not get affected by outliers, and the rest of the numeric columns which do not contain any outliers replaced their null values with mean.

previous_application.csv :

1. In the previous application data, 14 columns have more than 20% null values, I dropped them all.
2. After dropping these 14 columns the shape of the application data is (1048575,21).
3. Of the remaining 21 columns, only the PRODUCT_COMBINATION column contains 0.02% missing data. This column is an object-type column, so I replaced the missing data with mode.

Q3. Identify if there are outliers in the dataset. Also, mention why you think it is an outlier.

Here I have used the quartile method to find the outliers,

Denoted:

- 20^{th} percentile as Q1
- 80^{th} percentile as Q3
- IQR = Q3-Q1
- Upper limit: Q3+(1.5*IQR)
- Lower limit: Q1-(1.5*IQR)

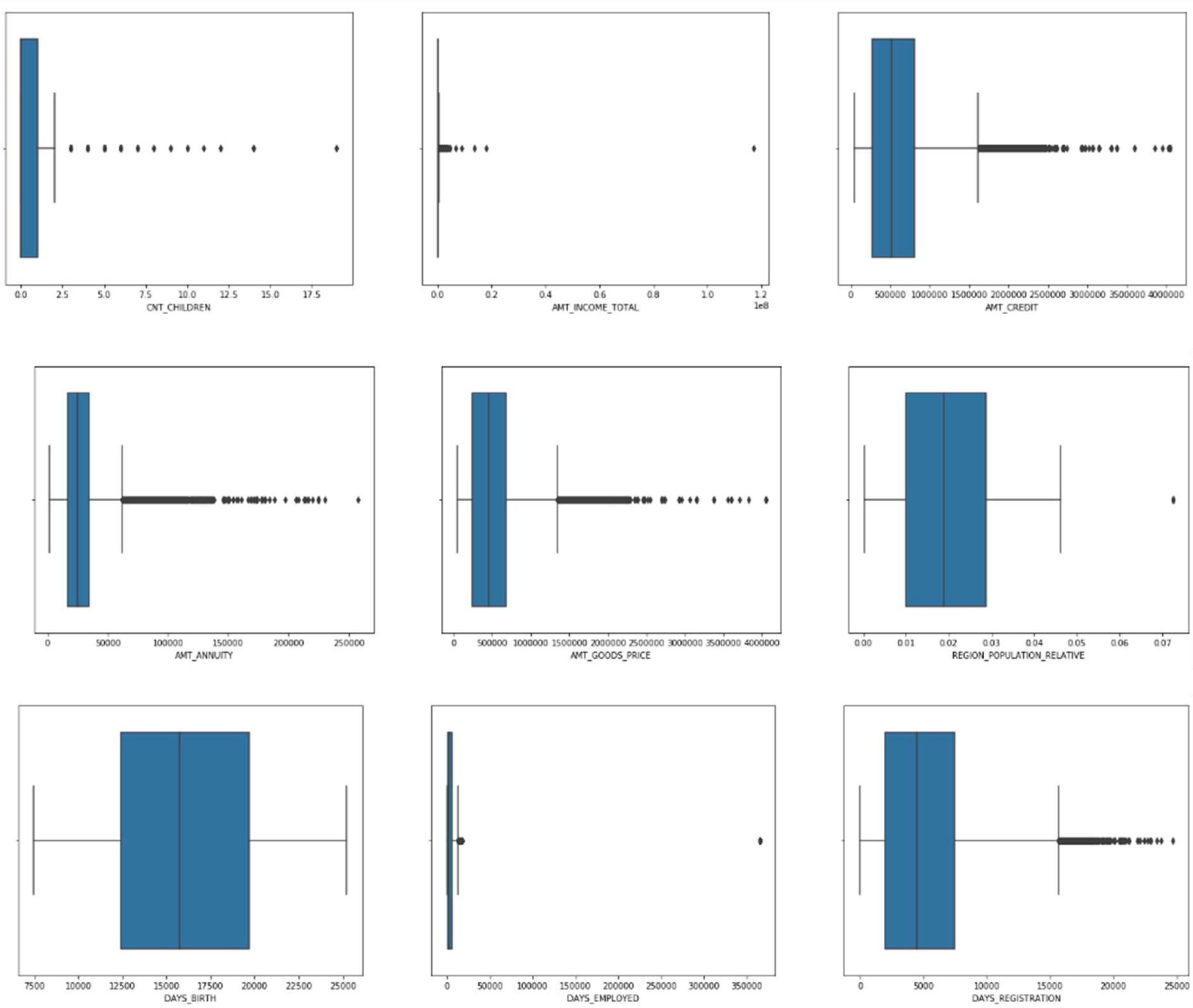
Those values greater than the upper limit or less than the lower limit are outliers. If any column's maximum value > Upper limit or column's minimum value < Lower limit, we can conclude that the column contains outliers.

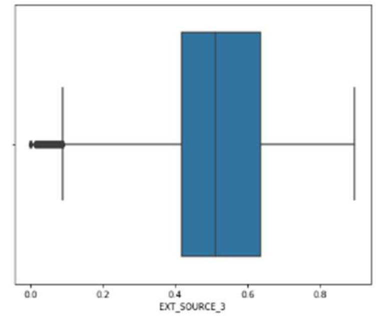
And to visualize them I used a boxplot.

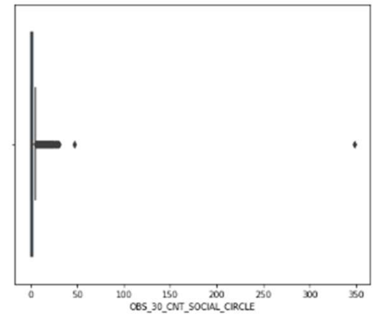
application_data.csv :

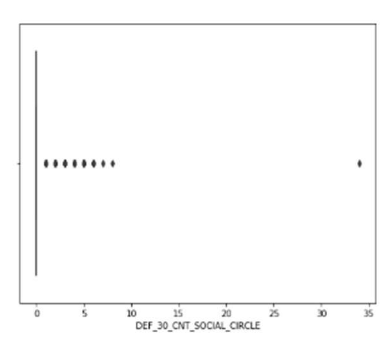
Out[98]:

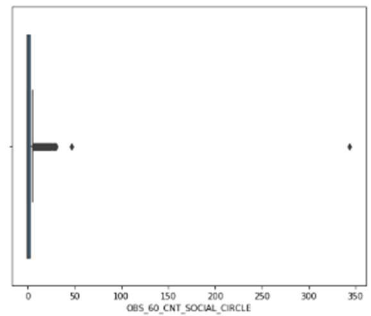
Column name	Min	Lower limit	Upper Limit	Max	Mean	Median	Outliers
CNT_CHILDREN	0.	-1.5	2.5	19.	0.417052	0.000000	YES
AMT_INCOME_TOTAL	25650.	-90000.	414000.	117000000.	168797.919297	147150.000000	YES
REQ_CREDIT_BUREAU_QRT	0.	0.	0.	261.	0.229631	0.000000	YES
REQ_CREDIT_BUREAU_MON	0.	0.	0.	27.	0.231293	0.000000	YES
EQ_CREDIT_BUREAU_WEEK	0.	0.	0.	8.	0.029723	0.000000	YES
REQ_CREDIT_BUREAU_DAY	0.	0.	0.	9.	0.006055	0.000000	YES
EQ_CREDIT_BUREAU_HOUR	0.	0.	0.	4.	0.005538	0.000000	YES
EF_60_CNT_SOCIAL_CIRCLE	0.	0.	0.	24.	0.099717	0.000000	YES
S_60_CNT_SOCIAL_CIRCLE	0.	-4.5	7.5	344.	1.400626	0.000000	YES
EF_30_CNT_SOCIAL_CIRCLE	0.	0.	0.	34.	0.142944	0.000000	YES
S_30_CNT_SOCIAL_CIRCLE	0.	-4.5	7.5	348.	1.417523	0.000000	YES
EQ_CREDIT_BUREAU_YEAR	0.	-4.5	7.5	25.	1.778463	1.000000	YES
DAYS_REGISTRATION	0.	-8617.5	18338.5	24672.	4986.120328	4504.000000	YES
DAYS_EMPLOYED	0.	-11909.5	21846.5	365243.	67724.742149	2219.000000	YES
ON_POPULATION_RELATIVE	0.00029	-0.023967500000000003	0.0635885	0.072508	0.020868	0.018850	YES
AMT_GOODS_PRICE	40500.	-659250.	1698750.	4050000.	538316.294367	450000.000000	YES
AMT_ANNUITY	1615.5	-19521.	71739.	258025.5	27108.573909	24903.000000	YES
AMT_CREDIT	45000.	-713250.	1867950.	4050000.	599025.999706	513531.000000	YES

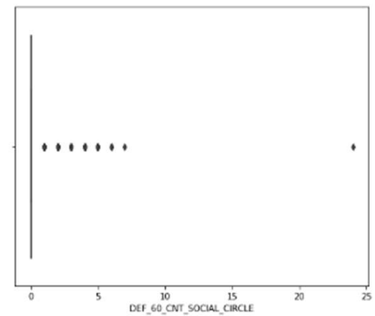


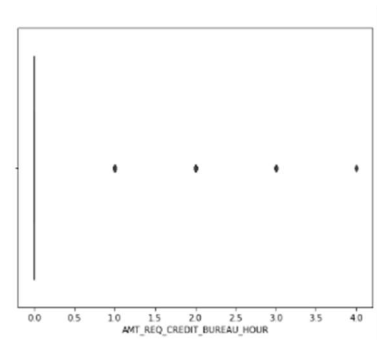


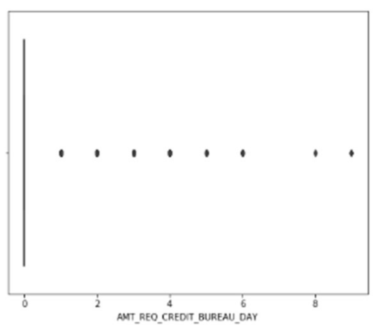


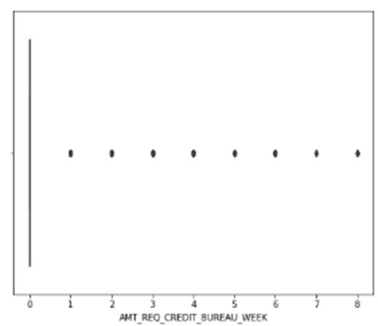


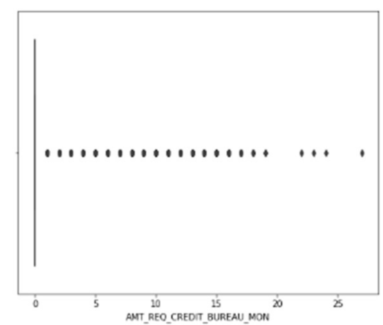


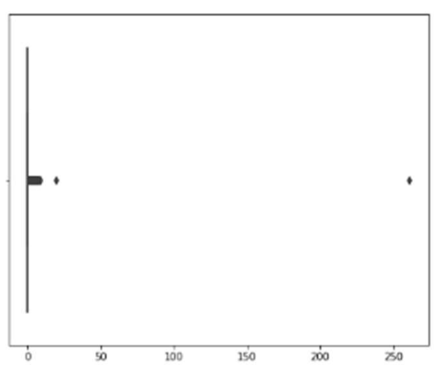


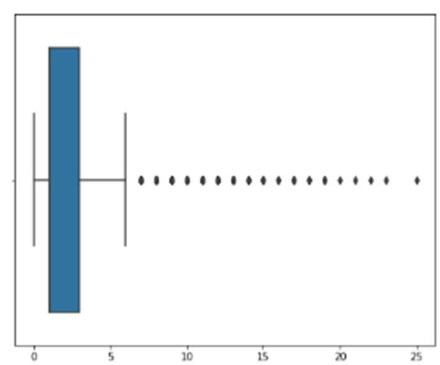












- DAYS_BIRTH, DAYS_ID_PUBLISH, and EXT_SOURCE_2 have no outliers.
- The rest of the columns have outliers.

NOTE:

DAYS_EMPLOYED has data which are more than 350000 days, which is nearly 1000 years.

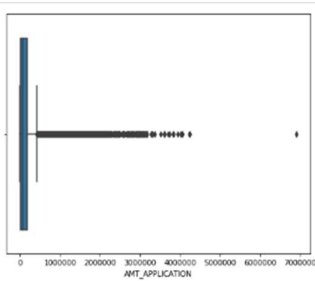
Here 18% of data are wrongly inputted as 365243, as per my understanding these might be missing data, which were wrongly entered as 365243.

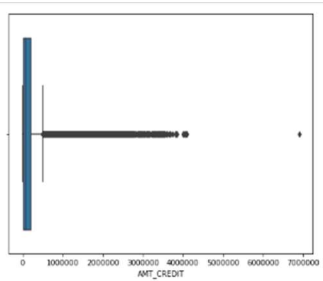
The "DAYS_EMPLOYED" column is highly skewed as 55374 numbers of data are wrongly entered as 365243, so the calculated mean value of the column would give a biased value. Thus I am replacing the wrong value with the median of the column.

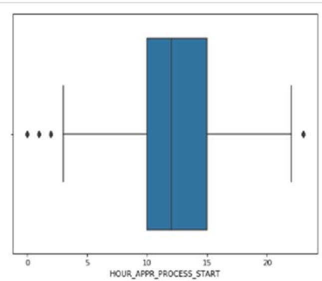
previous_application.csv :

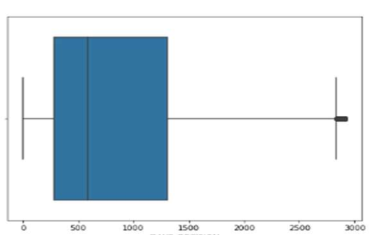
]:

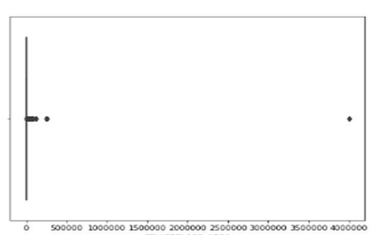
	Column name	Min	Lower limit	Upper Limit	Max	Mean	Median	Outliers
0	AMT_APPLICATION	0.	-337500.	562500.	6905160.	174269.769421	70816.5	YES
1	AMT_CREDIT	0.	-405000.	675000.	6905160.	195000.011725	80253.0	YES
2	HOUR_APPR_PROCESS_START	0.	1.	25.	23.	12.484856	12.0	YES
4	SELLERPLACE_AREA	-1.	-215.5	356.5	4000000.	318.390420	4.0	YES
3	DAYS_DECISION	2.	-1735.5	3532.5	2922.	882.038059	583.0	NO











In the previous application data, except DAYS_DECISION all other columns have outliers.

Q4. Identify if there is a data imbalance in the data. Find the ratio of data imbalance.

A dataset with a skewed class or data is called data imbalance.

For numeric data here checked the skewness.

For categorical data, I divided them into two parts, one is a binary class (Those columns, which have only two classes, such as yes/no, male/female), and the other is multiclass (Those columns, which have more than two class, such as approved/rejected/un-used). Here I counted the number of every class of the columns, if any class's count value is highly skewed than other classes of the column then I denoted that column as an imbalanced column.

application_data.csv :

Numeric data:

	Features	Skewness	Mean	Median	Mean-Median
1	AMT_INCOME_TOTAL	391.559654	168797.919297	147150.00000	21647.919296984503
21	AMT_REQ_CREDIT_BUREAU_QRT	141.400915	0.229631	0.00000	0.22963080995476584
18	AMT_REQ_CREDIT_BUREAU_DAY	29.081577	0.006055	0.00000	0.006055067948788235
17	AMT_REQ_CREDIT_BUREAU_HOUR	15.641990	0.005538	0.00000	0.005538013274321894
13	OBS_30_CNT_SOCIAL_CIRCLE	12.143796	1.417523	0.00000	1.4175232755901415
15	OBS_60_CNT_SOCIAL_CIRCLE	12.075153	1.400626	0.00000	1.4006263190585053
19	AMT_REQ_CREDIT_BUREAU_WEEK	10.008033	0.029723	0.00000	0.029722513991369416
20	AMT_REQ_CREDIT_BUREAU_MON	8.371505	0.231293	0.00000	0.23129253912868158
16	DEF_60_CNT_SOCIAL_CIRCLE	5.287339	0.099717	0.00000	0.09971675809971026
14	DEF_30_CNT_SOCIAL_CIRCLE	5.192572	0.142944	0.00000	0.14294448003486054
7	DAYS_EMPLOYED	2.212846	2354.427019	2219.00000	135.4270188708697
0	CNT_CHILDREN	1.974604	0.417052	0.00000	0.4170517477423572
3	AMT_ANNUITY	1.579777	27108.573909	24903.00000	2205.573909183444
5	REGION_POPULATION_RELATIVE	1.488009	0.020868	0.01885	0.0020181120577894
22	AMT_REQ_CREDIT_BUREAU_YEAR	1.465643	1.778463	1.00000	0.7784632094461661
4	AMT_GOODS_PRICE	1.350143	538316.294367	450000.00000	88316.29436670558
2	AMT_CREDIT	1.234778	599025.999706	513531.00000	85494.9997057016
8	DAYS_REGISTRATION	0.590872	4986.120328	4504.00000	482.1203275384187

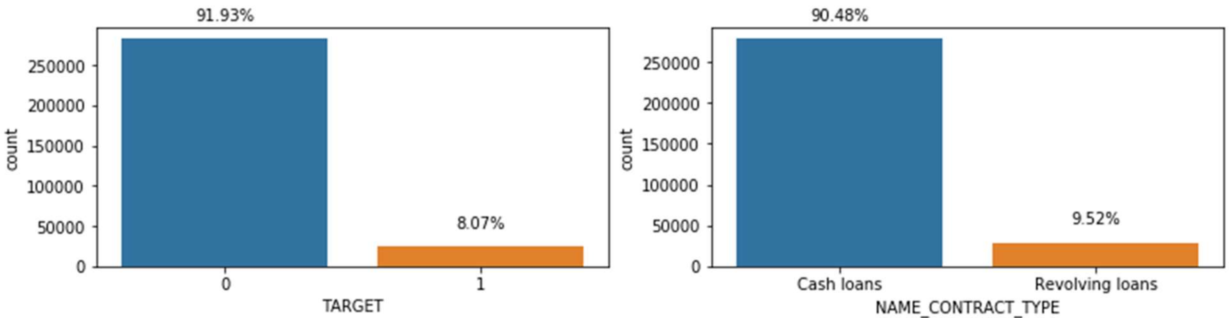
↑↑ Above data shows the data which are Positive Skewed or Right-Skewed (Mean > Median)

	Features	Skewness	Mean	Median	Mean-Median
11	EXT_SOURCE_2	-0.794429	0.514393	0.565467	-0.05107448534840242

⮅ Above data shows the data which are Negative Skewed or Left-Skewed (Mean < Median)

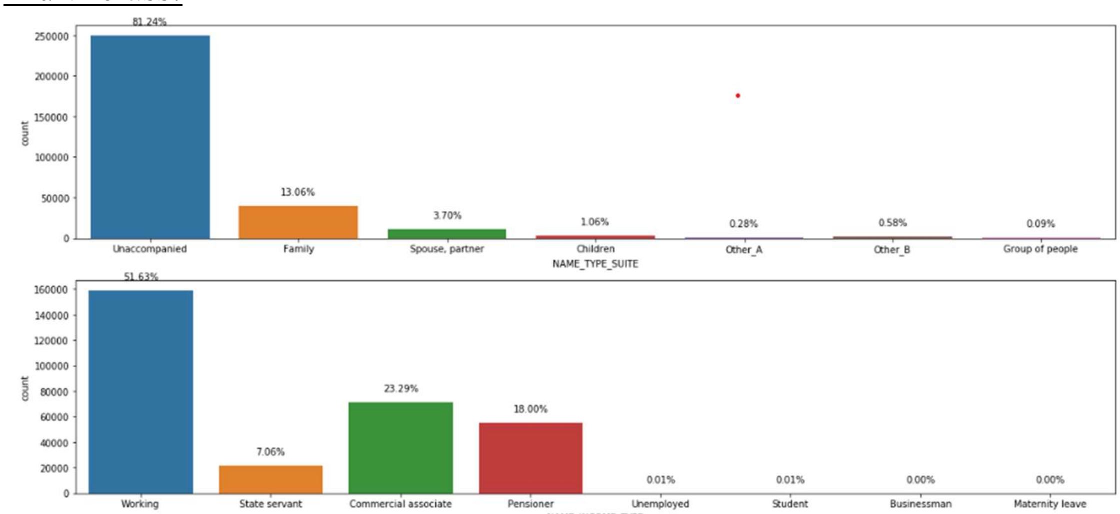
Categoric data:

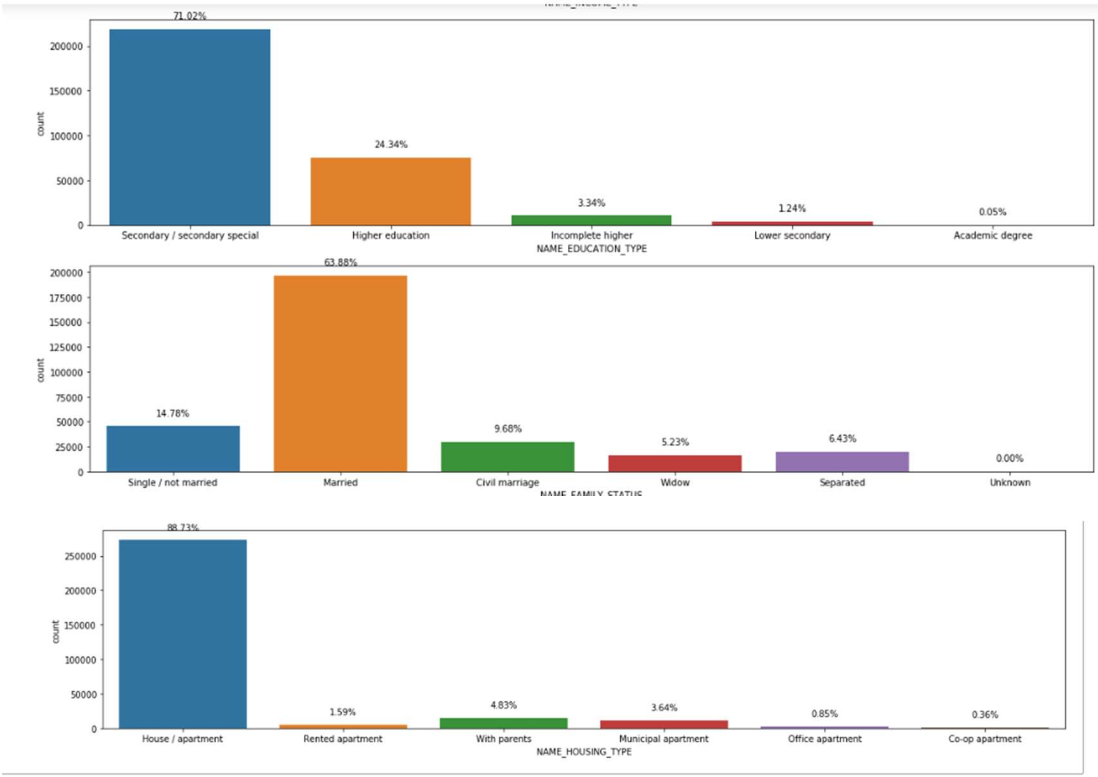
Binary-class:



In the Binary class segment 'TARGET', the 'NAME_CONTRACT_TYPE' columns are highly imbalanced.

Multi-class:





In the multiclass segment 'NAME_TYPE_SUITE', 'NAME_INCOME_TYPE', 'NAME_EDUCATION_TYPE', 'NAME_FAMILY_STATUS', and 'NAME_HOUSING_TYPE' columns are imbalanced.

previous_application.csv :

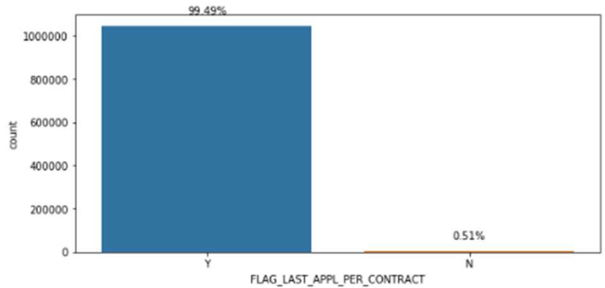
Numeric data:

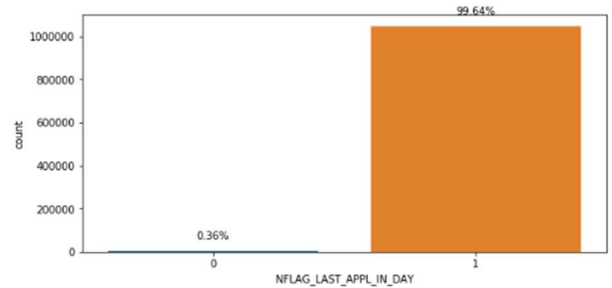
	Features	Skewness	Mean	Median	Mean-Median
4	SELLERPLACE_AREA	477.768999	318.390420	4.0	314.3904203323558
0	AMT_APPLICATION	3.390787	174269.769421	70816.5	103453.26942099651
1	AMT_CREDIT	3.255049	195000.011725	80253.0	114747.01172482444
3	DAYS_DECISION	1.050799	882.038059	583.0	299.03805927091526

⬆⬆ Above data shows the data which are Positive Skewed or Right-Skewed (Mean > Median)

Categoric data:

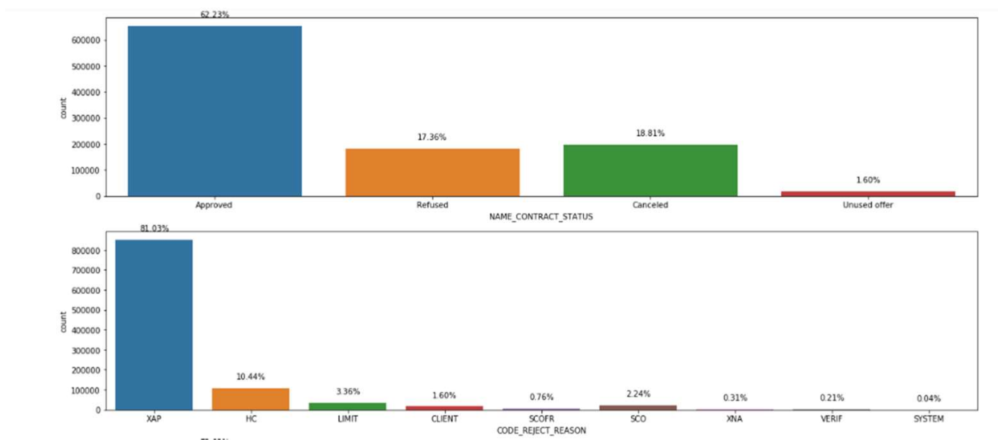
<u>Binary-class:</u>

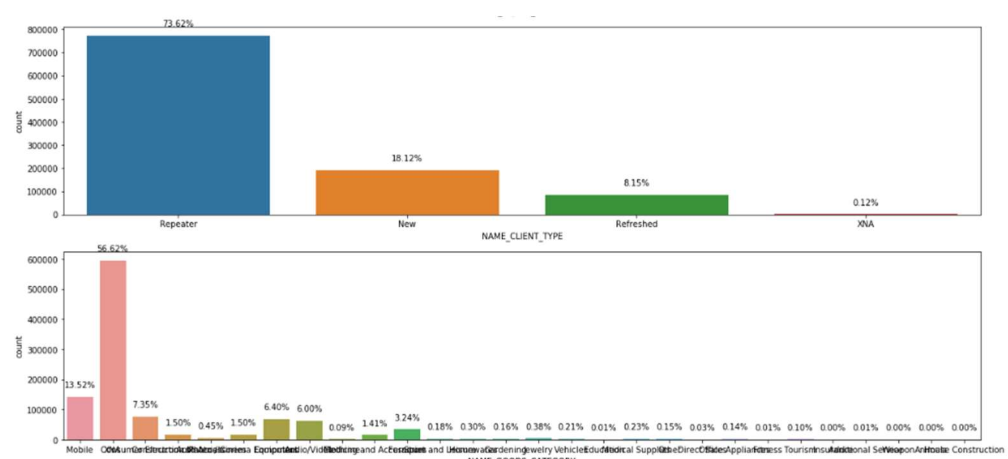


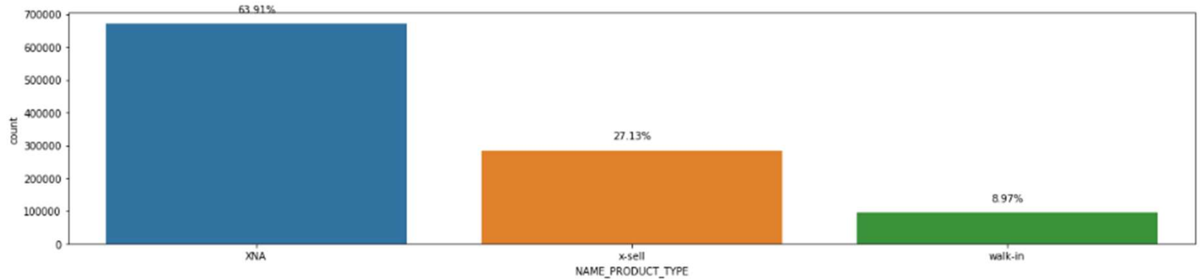


In the Binary class segment 'FLAG_LAST_APPL_PER_CONTRACT', and 'NFLAG_LAST_APPL_IN_DAY' columns are highly imbalanced.

<u>Multi-class:</u>







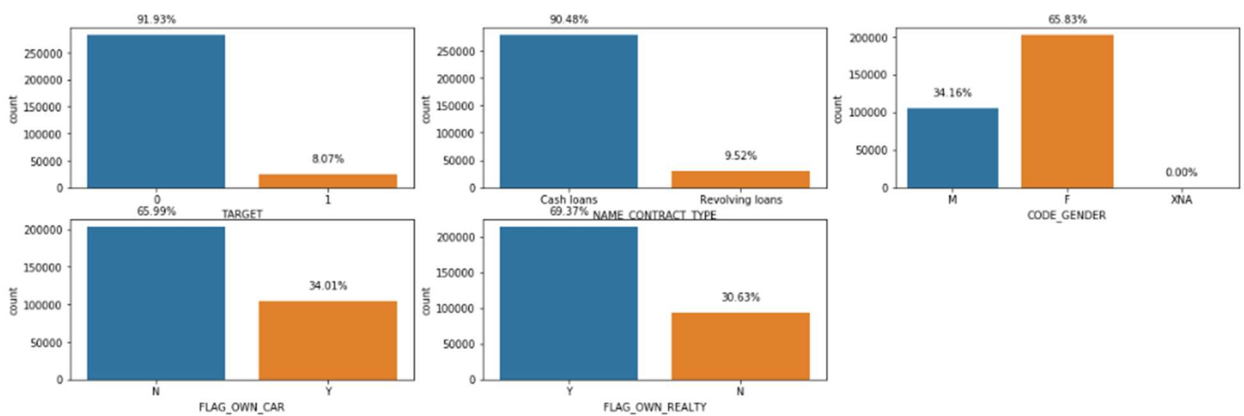
In the multiclass segment 'FLAG_LAST_APPL_PER_CONTRACT', 'NFLAG_LAST_APPL_IN_DAY', 'NAME_CONTRACT_STATUS', 'CODE_REJECT_REASON', 'NAME_CLIENT_TYPE', 'NAME_GOODS_CATEGORY', 'NAME_PRODUCT_TYPE' columns are imbalanced.

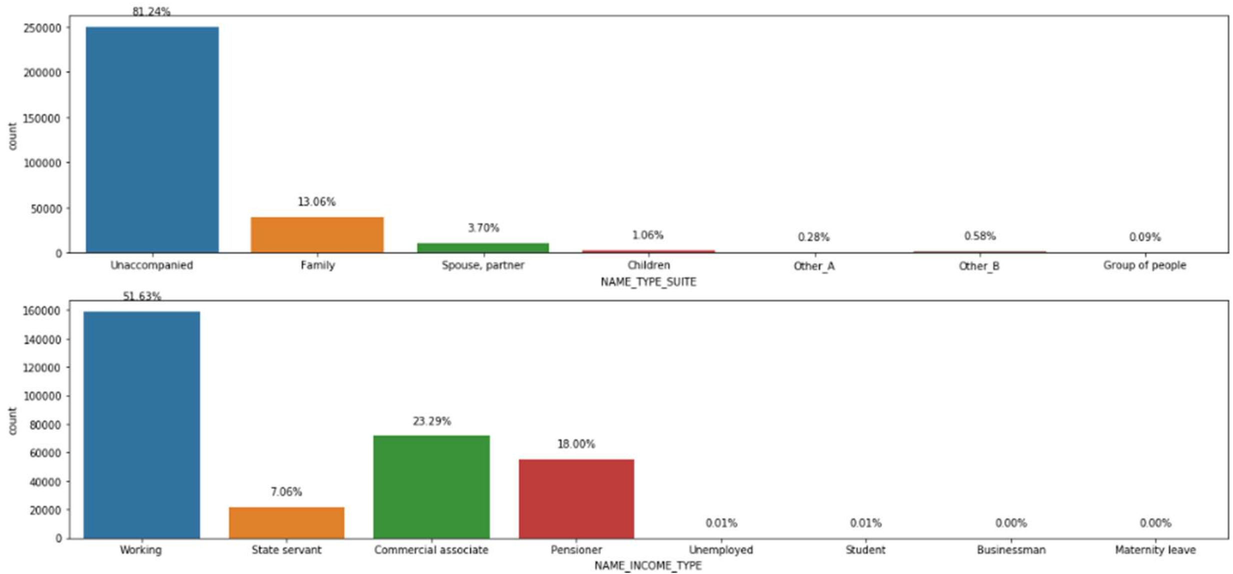
Q5. Explain the results of univariate, segmented univariate, bivariate analysis, etc. in business terms.

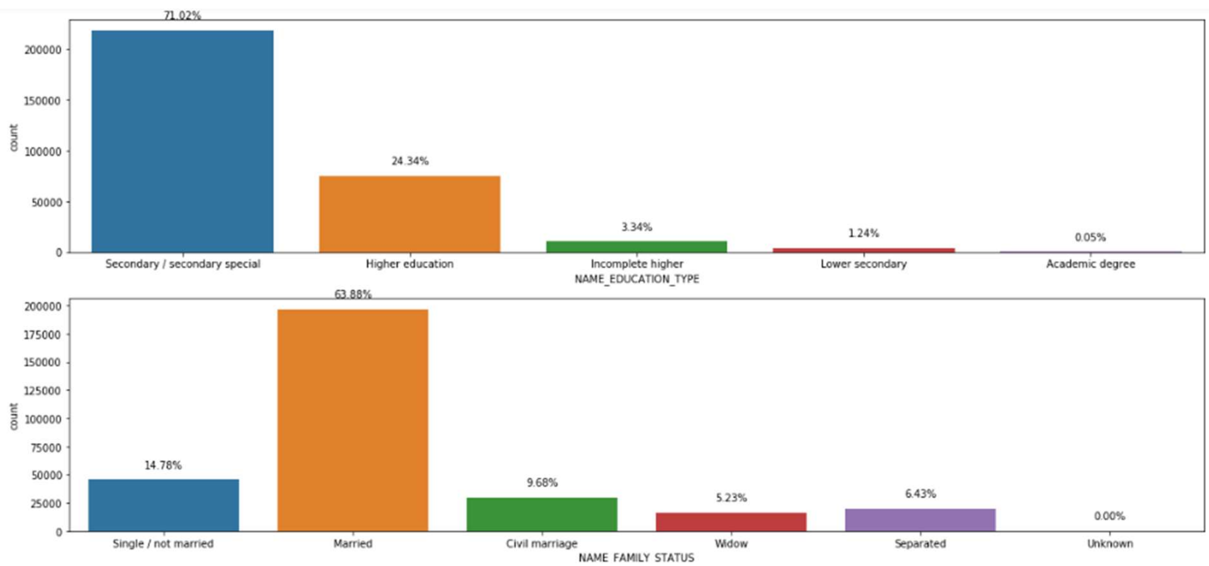
application_data.csv :

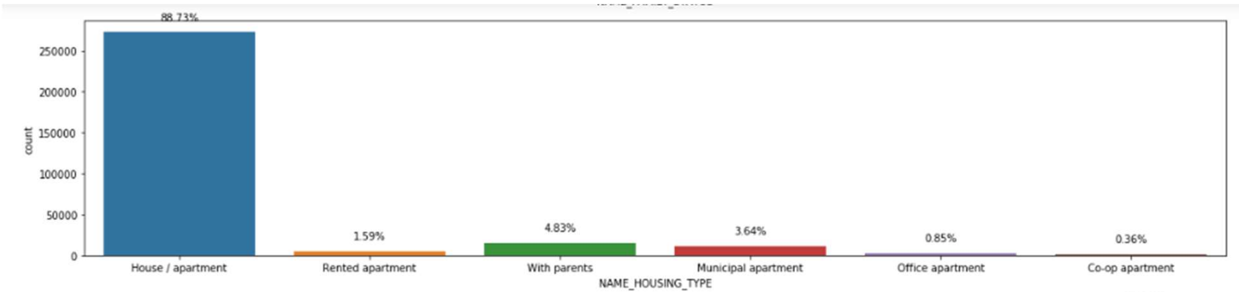
Observations :

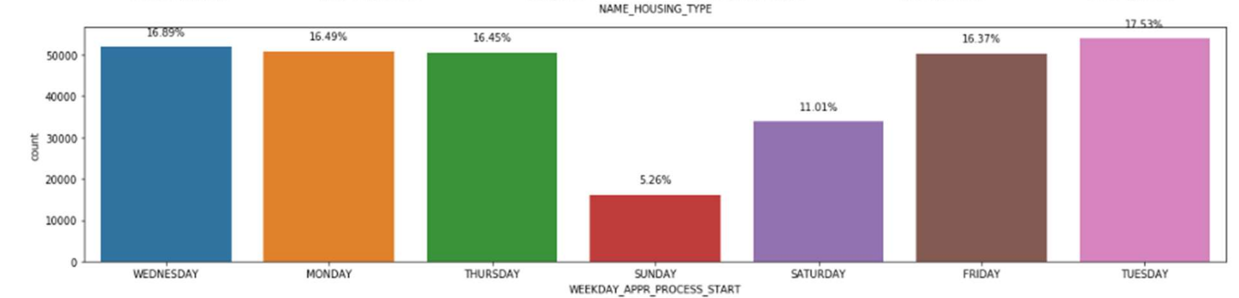
1.




81.24%
13.06%
3.70%
1.06%
0.28%
0.58%
0.09%
250000
200000
150000
100000
50000
0
count
Unaccompanied
Family
Spouse, partner
Children
Other_A
Other_B
Group of people
NAME_TYPE_SUITE
51.63%
23.29%
18.00%
7.06%
0.01%
0.01%
0.00%
0.00%
160000
140000
120000
100000
80000
60000
40000
20000
0
count
Working
State servant
Commercial associate
Pensioner
Unemployed
Student
Businessman
Maternity leave
NAME_INCOME_TYPE


71.02%
24.34%
3.34%
1.24%
0.05%
200000
150000
100000
50000
0
count
Secondary / secondary special
Higher education
Incomplete higher
Lower secondary
Academic degree
NAME_EDUCATION_TYPE
63.88%
14.78%
9.68%
5.23%
6.43%
0.00%
200000
175000
150000
125000
100000
75000
50000
25000
0
count
Single / not married
Married
Civil marriage
Widow
Separated
Unknown
NAME_FAMILY_STATUS





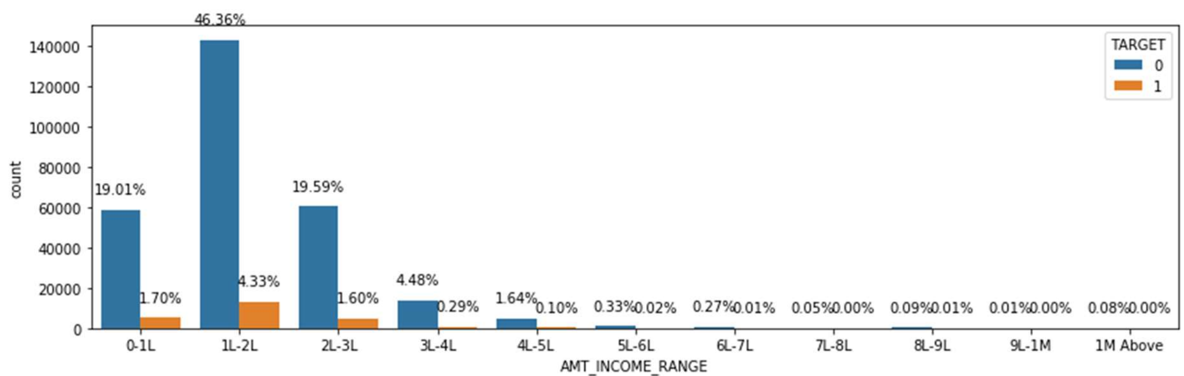
- Here most of the applicants (91.93%) don't have any payment difficulties.
- 90.48% applied for a Cash loan.
- 65.83% of the applicants are female.
- Nearly 66% of the applicants own a car.
- Nearly 70% of the client owns a house or flat.
- More than 80 of the clients came solo for their loan applications.
- More than 50 of the clients are working.
- Nearly 72% of the applicants have completed their secondary education.
- Nearly 90% of the applicants are staying in their own houses.
- Surprisingly here I can see that more than 16% (Saturday: 11.01% & Sunday: 5.26%) of the application proceeded on the weekends.

2.

```
1L-2L        50.70
2L-3L        21.19
0-1L         20.71
3L-4L         4.77
4L-5L         1.74
5L-6L         0.36
6L-7L         0.28
8L-9L         0.10
1M Above      0.08
7L-8L         0.05
9L-1M         0.01
Name: AMT_INCOME_RANGE
```

AMT_INCOME_RANGE v/s % TARGET

	0	1
1L-2L	91.452103	8.547897
2L-3L	92.449675	7.550325
0-1L	91.797231	8.202769
3L-4L	93.969747	6.030253
4L-5L	93.993658	6.006342
5L-6L	93.698630	6.301370
6L-7L	95.166858	4.833142
8L-9L	94.295302	5.704698
1M Above	94.800000	5.200000
7L-8L	98.148148	1.851852
9L-1M	92.857143	7.142857



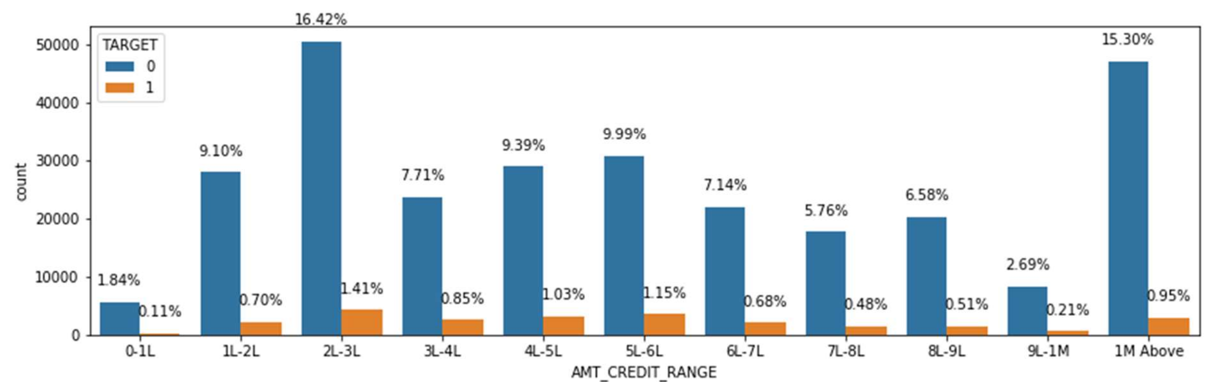
- More than 50% of people who applied for the loan have an income range between 1-2 Lakh.
- Among the applicants with a 7 lakh to 8 lakh income range, more than 98% of people don't face any payment difficulties.

3.

AMT_CREDIT_RANGE v/s % TARGET

```
2L-3L        17.824728
1M Above     16.254703
5L-6L        11.131960
4L-5L        10.418489
1L-2L         9.801275
3L-4L         8.564897
6L-7L         7.820533
8L-9L         7.086576
7L-8L         6.241403
9L-1M         2.902986
0-1L          1.952450
Name: AMT_CREDIT_RANGE,
```

	0	1
2L-3L	92.116834	7.883166
1M Above	94.134240	5.865760
5L-6L	89.708460	10.291540
4L-5L	90.102378	9.897622
1L-2L	92.836762	7.163238
3L-4L	90.041005	9.958995
6L-7L	91.280303	8.719697
8L-9L	92.864354	7.135646
7L-8L	92.361799	7.638201
9L-1M	92.752324	7.247676
0-1L	94.487009	5.512991



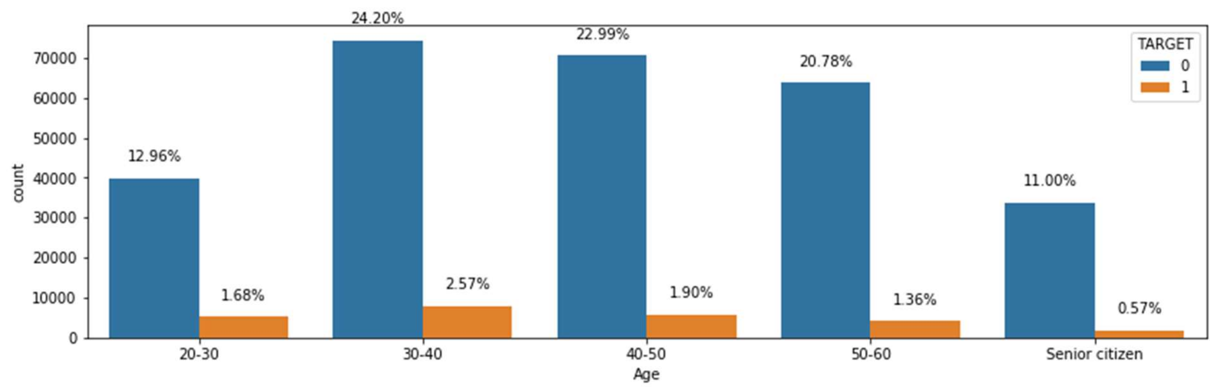
- Major loan applications are for loan amounts between 2-3 lakh.
- There are significant numbers (16.25%) of applicants who have applied for loan amount more than 1 million, among them nearly 95% don't face any payment difficulties

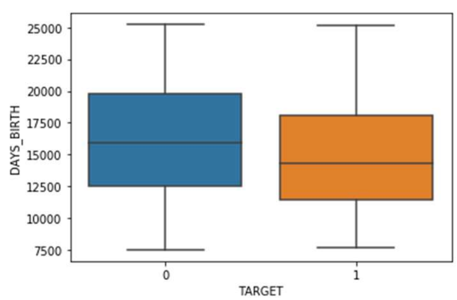
4.

```
7]: 30-40              26.77
    40-50              24.89
    50-60              22.13
    20-30              14.64
    Senior citizen     11.57
    Name: Age, dtype: float64
```

Age v/s % TARGET

	0	1
30-40	90.416484	9.583516
40-50	92.349198	7.650802
50-60	93.870295	6.129705
20-30	88.543124	11.456876
Senior citizen	95.078558	4.921442





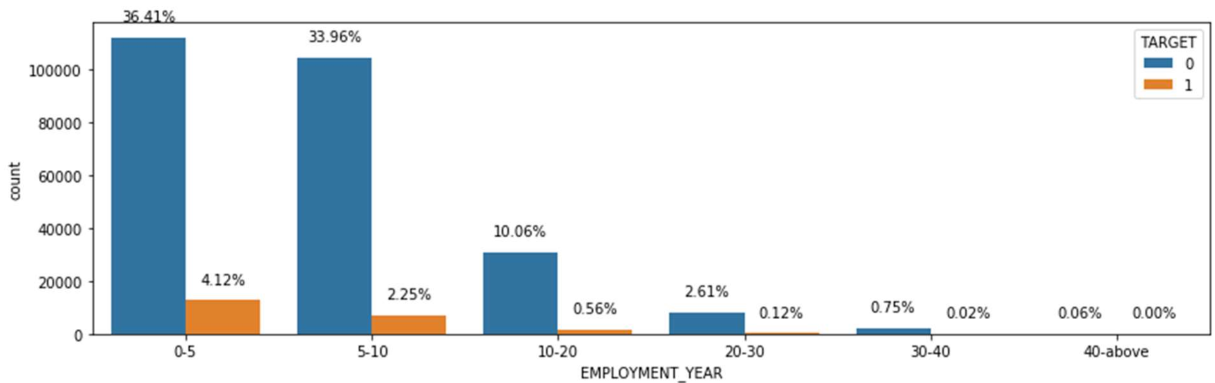
- Those who have applied for a loan, among those, aged between 30-40 people are the most. Here We can see that more than 11% of people who are senior citizens (Aged above 60) applied for a loan and overall more than 30% of the applicant are aged more than 30.
- Older people face fewer payment difficulties than younger people, As we can see that among senior citizens less than 5% of applicants face payment difficulties.

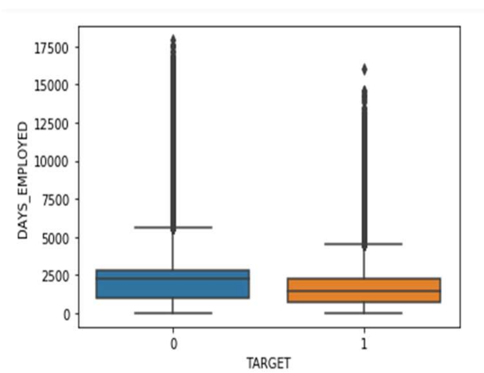
5.

```
0-5          40.53
5-10         36.21
10-20        10.62
20-30         2.73
30-40         0.77
40-above      0.06
Name: EMPLOYMENT_YEAR,
```

EMPLOYMENT_YEAR v/s % TARGET

	0	1
0-5	89.838246	10.161754
5-10	93.785752	6.214248
10-20	94.745545	5.254455
20-30	95.433464	4.566536
30-40	96.882898	3.117102
40-above	99.428571	0.571429





- People are likely to apply for a loan in their early career, here more than 40% of applicants' employment year is between 0-5 years and more than 36% of applicants' employment year is between 5-10 years.
- Applicants who have applied for a loan in their early careers face difficulties in loan repayment. And who have worked for more than 40 years, among them less the 1% of applicants face payment difficulties.

6.

NAME_TYPE_SUITE v/s % TARGET

	0	1
Unaccompanied	91.831253	8.168747
Family	92.505417	7.494583
Spouse, partner	92.128408	7.871592
Children	92.623202	7.376798
Other_B	90.169492	9.830508
Other_A	91.224018	8.775982
Group of people	91.512915	8.487085

NAME_INCOME_TYPE v/s % TARGET

	0	1
Working	90.411528	9.588472
Commercial associate	92.515743	7.484257
Pensioner	94.613634	5.386366
State servant	94.245035	5.754965
Student	81.818182	22.222222
Unemployed	77.777778	18.181818
Businessman	100.000000	0.000000
Maternity leave	60.000000	40.000000

- Those who are on Maternity leave and who are students face difficulties in payment while repaying the loan.
- Out of all applicants, only 10 people are businessmen, and all of them don't face any payment difficulties.

7.

NAME_EDUCATION_TYPE v/s % TARGET

	0	1
Secondary / secondary special	91.060071	8.939929
Higher education	94.644885	5.355115
Incomplete higher	91.515034	8.484966
Lower secondary	89.072327	10.927673
Academic degree	98.170732	1.829268

- Those who have completed their Academic degree, among them less than 2% of applicants face payment difficulties.

8.

Industry: type 7	91.966335	8.033665
Transport: type 3	84.245998	15.754002
Industry: type 1	88.931665	11.068335
[illegible]	[illegible]	[illegible]

Trade: type 5	93.877551	6.122449
Industry: type 8	87.500000	12.500000

- Those who are working in the "Transport: type 3" and "Industry: type 8" organizations, among them 15.75% and 12.50% of applicants faced payment difficulties respectively.

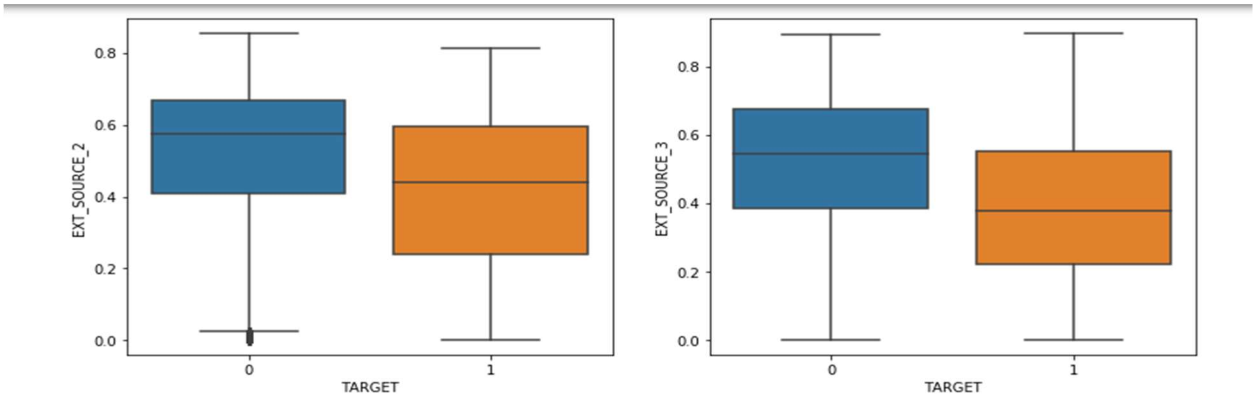
9. Correlation between TARGET and other numeric columns:

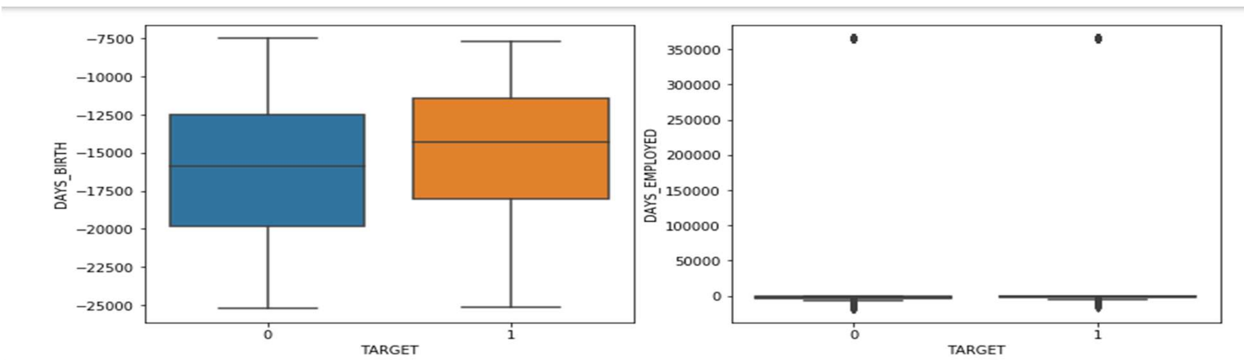
Bi-serial Correlation Coefficient:

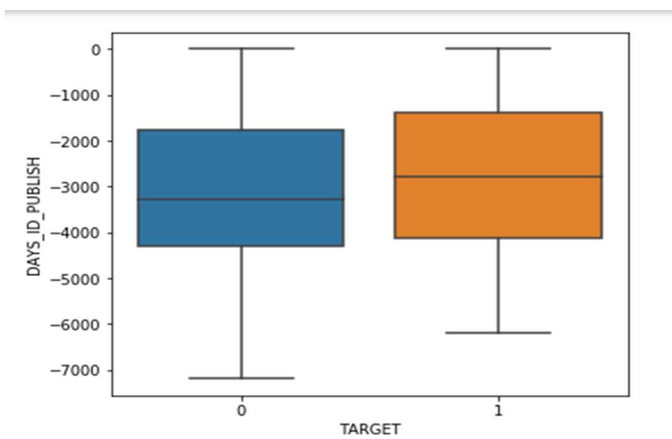
	Variable	Bi-serial Corr
11	EXT_SOURCE_2	0.160303
12	EXT_SOURCE_3	0.157396
6	DAYS_BIRTH	0.078239
7	DAYS_EMPLOYED	0.068665
9	DAYS_ID_PUBLISH	0.051457

Here I can observe some +ve correlation between TARGET and 'EXT_SOURCE_2','EXT_SOURCE_3','DAYS_BIRTH','DAYS_EMPLOYED', and 'DAYS_ID_PUBLISH'

Visualization:



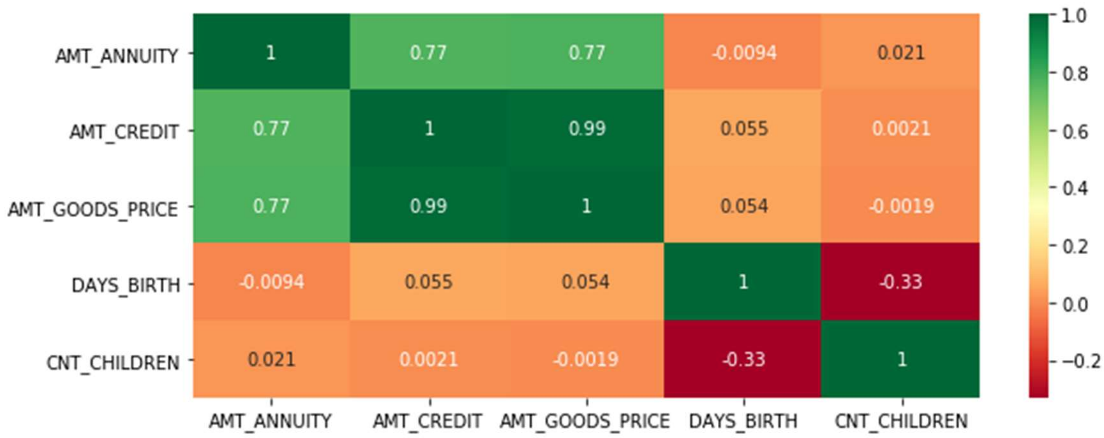




In the above visualization, one can easily see the relation.

10. Top 5 correlations within the numeric data.

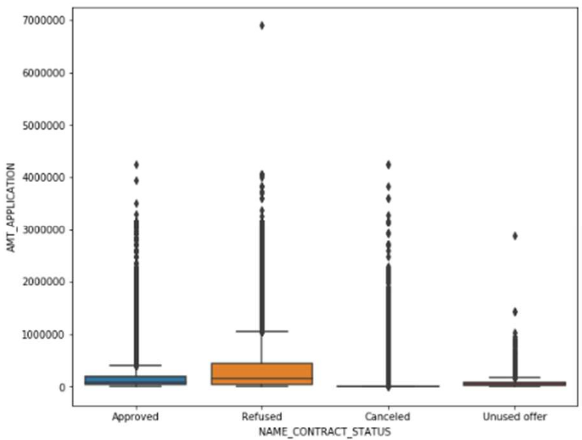
```
OBS_30_CNT_SOCIAL_CIRCLE  OBS_60_CNT_SOCIAL_CIRCLE    0.998491
AMT_CREDIT                AMT_GOODS_PRICE             0.986734
DEF_30_CNT_SOCIAL_CIRCLE  DEF_60_CNT_SOCIAL_CIRCLE    0.860556
AMT_ANNUITY               AMT_GOODS_PRICE             0.774848
AMT_CREDIT                AMT_ANNUITY                 0.770138
dtype: float64
```

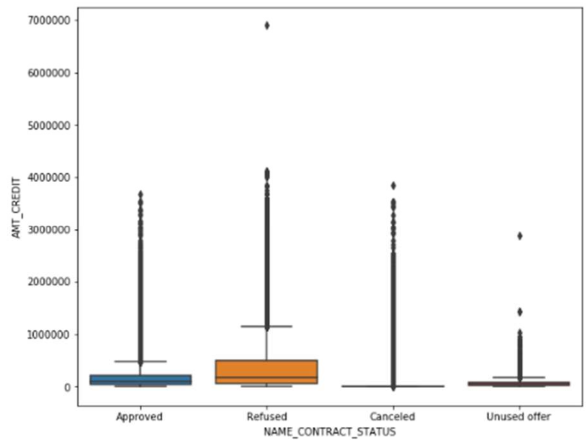


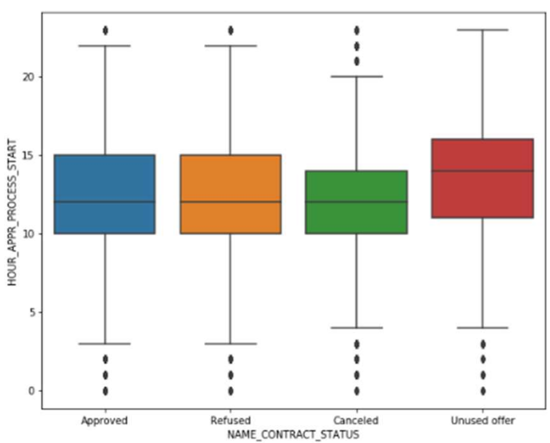
- For higher credit amounts the annuity is also higher
- If the goods price is higher the applicant is likely to get a higher credit amount
- Surprisingly, among the applicant older people have fewer children than younger people.

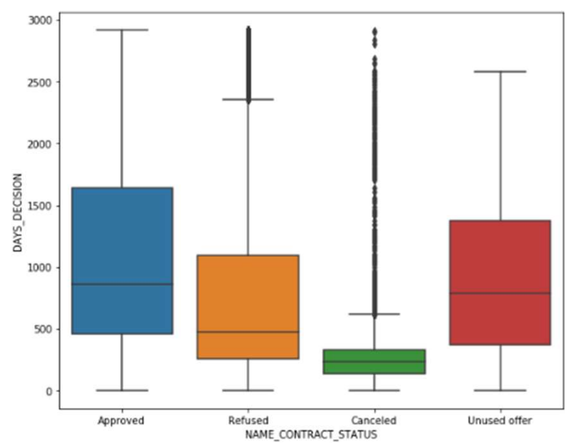
previous_application.csv :

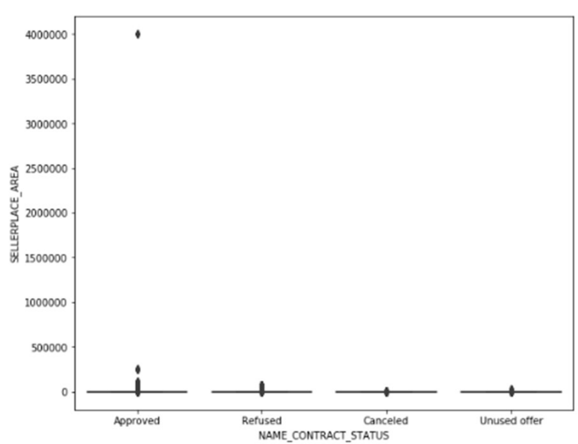
1.





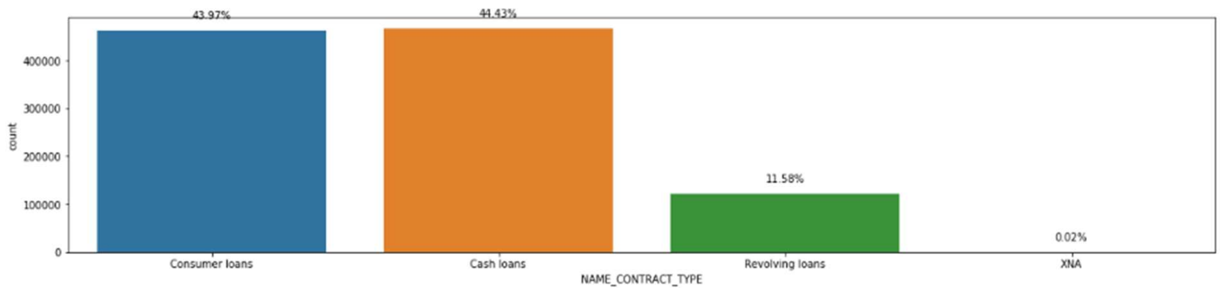


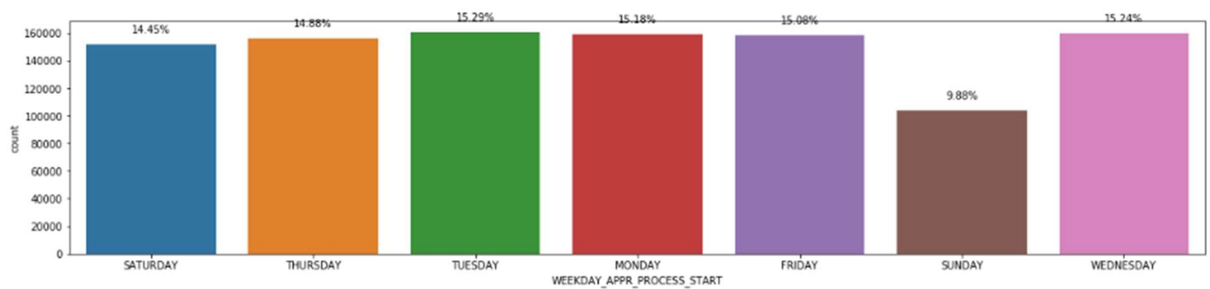


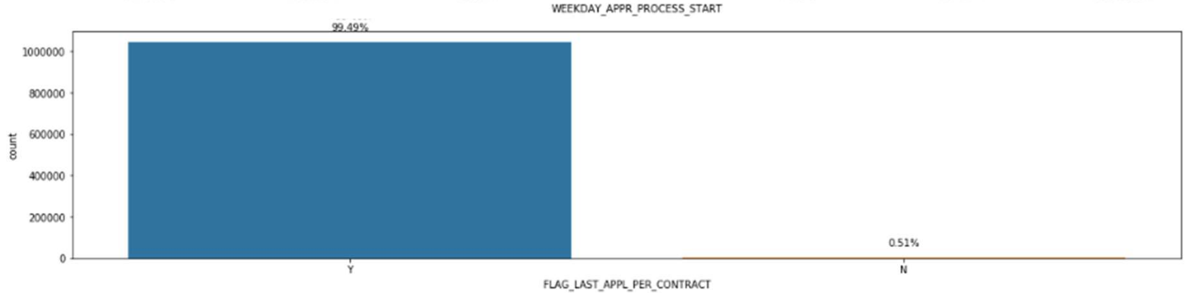


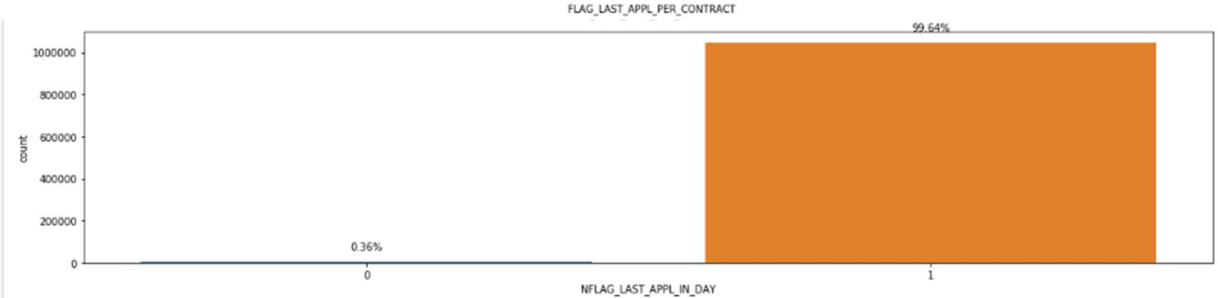
- Here I can see that the average applied loan amount is higher for refused cases.
- Also, the average revised loan amount by a bank (AMT_CREDIT) is higher for refused cases.
- Unused offers were mostly processed at the 15th hour of the day.

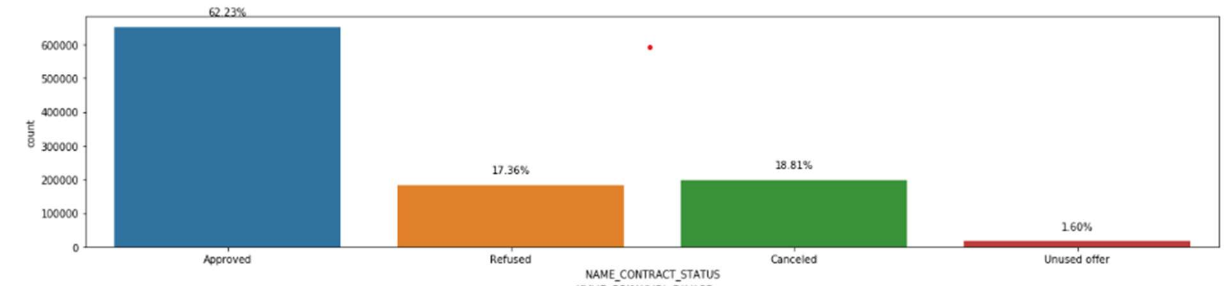
2.

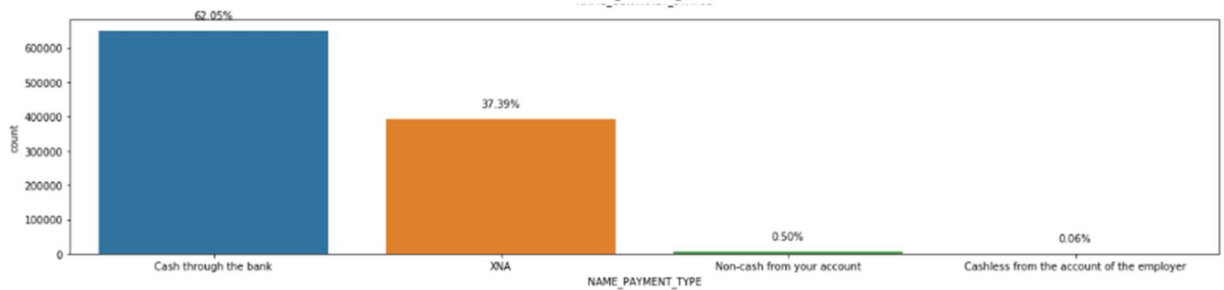


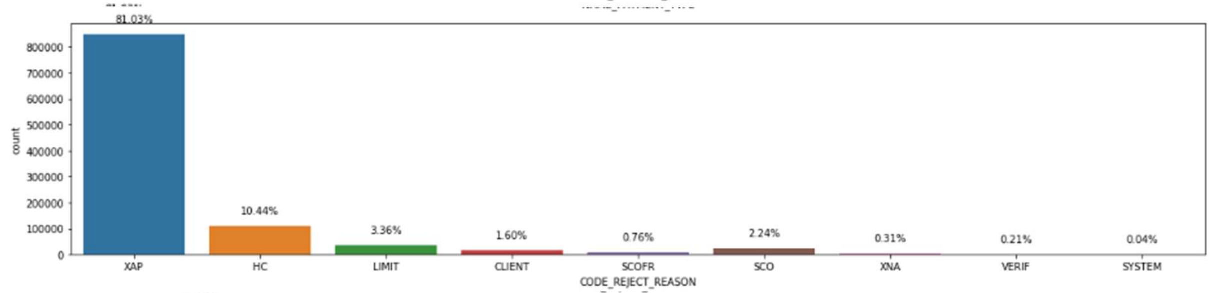




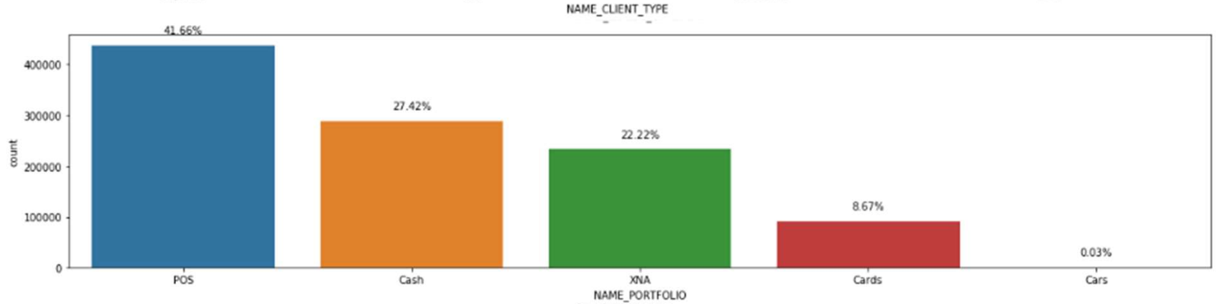


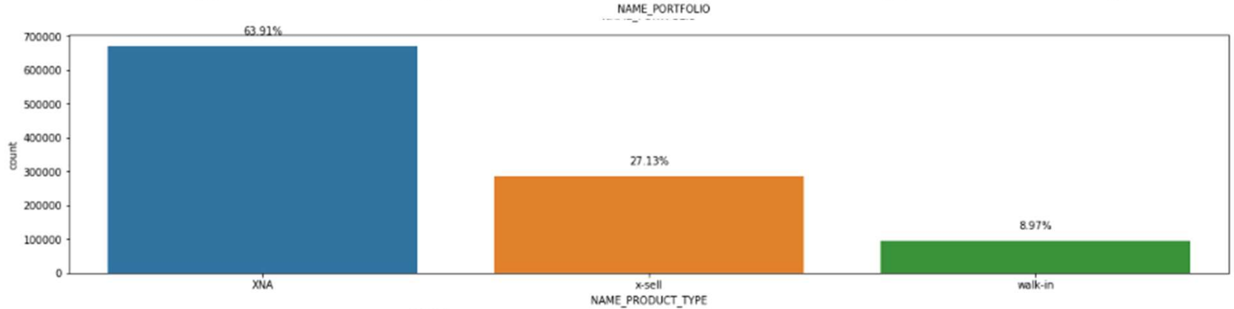

62.23%
17.36%
18.81%
1.60%
count
Approved
Refused
Canceled
Unused offer
NAME_CONTRACT_STATUS

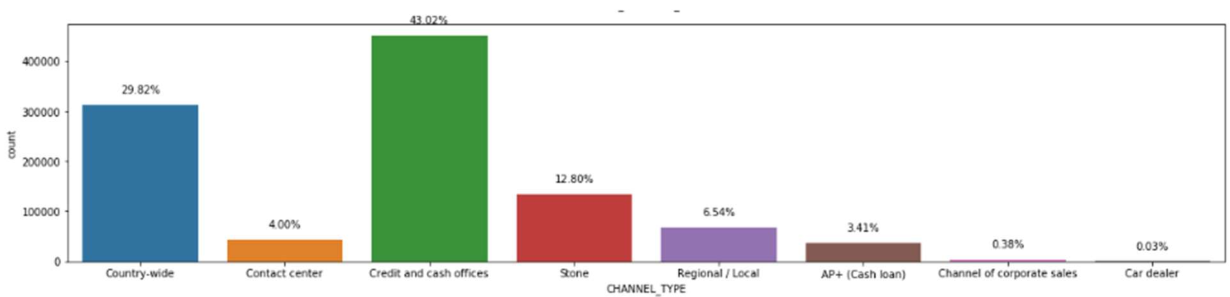

62.05%
37.39%
0.50%
0.06%
count
Cash through the bank
XNA
Non-cash from your account
Cashless from the account of the employer
NAME_PAYMENT_TYPE


81.03%
10.44%
3.36%
1.60%
0.76%
2.24%
0.31%
0.21%
0.04%
count
XAP
HC
LIMIT
CLIENT
SCOFR
SCO
XNA
VERIF
SYSTEM
CODE_REJECT_REASON


73.62%
18.12%
8.15%
0.12%
count
Repeater
New
Refreshed
XNA
NAME_CLIENT_TYPE


41.66%
27.42%
22.22%
8.67%
0.03%
count
POS
Cash
XNA
Cards
Cars
NAME_PORTFOLIO


63.91%
27.13%
8.97%
count
XNA
x-sell
walk-in
NAME_PRODUCT_TYPE



- Nearly 44% of applicants applied for consumer loans and nearly 45% applied for cash loans.
- More than 25% of the loan processing starts on weekends.
- Over 60% of loans got approval
- Over 60% of payment types are cash through the bank.
- In more than 80% of the case, the code rejects the reason is XAP.
- Over 70% of applicants are a repeater and 18% are new clients.
- Over 60% of portfolios are POS.

NAME_CONTRACT_STATUS v/s Others:

- In 80% of cases, consumer loans get approved.
- In 22% of cases, cash loans get refused.
- Those loan applications were processed on Sunday, and in 71 cases they got approved.
- The new client has more approval rate than other clients.
- In 90% of cases, the POS portfolio got approved.

Q6. Find the top 10 correlations for the Client with payment difficulties and all other cases (Target variable).

Here TARGET=1

	Var1	Var2	Correlation
0	OBS_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	0.998491
1	AMT_CREDIT	AMT_GOODS_PRICE	0.986734
2	DEF_30_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE	0.860556
3	AMT_ANNUITY	AMT_GOODS_PRICE	0.774848
4	AMT_CREDIT	AMT_ANNUITY	0.770138
5	DEF_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	0.331951
6	DAYS_BIRTH	DAYS_REGISTRATION	0.331912
7	CNT_CHILDREN	DAYS_BIRTH	0.330938
8	OBS_30_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.329721
9	DAYS_BIRTH	DAYS_ID_PUBLISH	0.272691