Credit Exploratory Data Analysis Case Study

PG DIPLOMA IN DATA SCIENCE, IIIT-B & UPGRAD

BY SHREYAS DUBEY & NITIN DINESAN

SUBMITTED ON 5 APRIL 2021

DS C27 BATCH

Objective:

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.

A thorough analysis of the bank's application and previous application data sets were conducted using exploratory data analysis, and several comparisons and correlations were performed among various categories and target variables.

This is a high-level management presentation covering the analysis and insights of selected key categories and variables, that would be useful to the management to assess potential risks, risk mitigation and to make sound business decisions.

Univariate Analysis

Categorical Variable Analysis

Analysis of Name_Contract_Type column

```
Default % according to category for NAME_CONTRACT_TYPE column is :
Cash loans           8.345913
Revolving loans      5.478703
Name: NAME_CONTRACT_TYPE, dtype: float64
```


Key Insights:
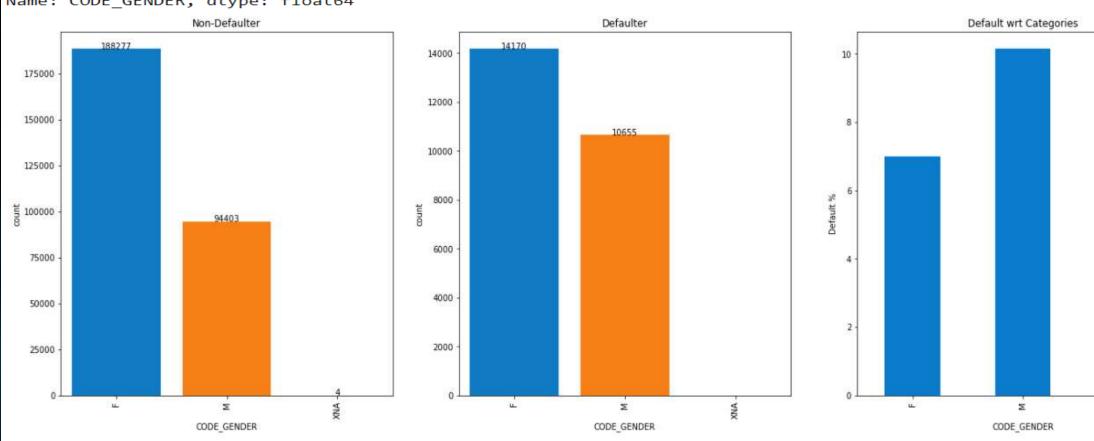
❖ Both defaulter and non-defaulter avail more cash loans compared to revolving loans.
❖ Default % is higher in "cash loan" category versus "revolving loan" category.

Analysis of CODE_GENDER column

```
Default % according to category for CODE_GENDER column is :
M          10.142017
F           6.999363
XNA         0.000000
Name: CODE_GENDER, dtype: float64
```
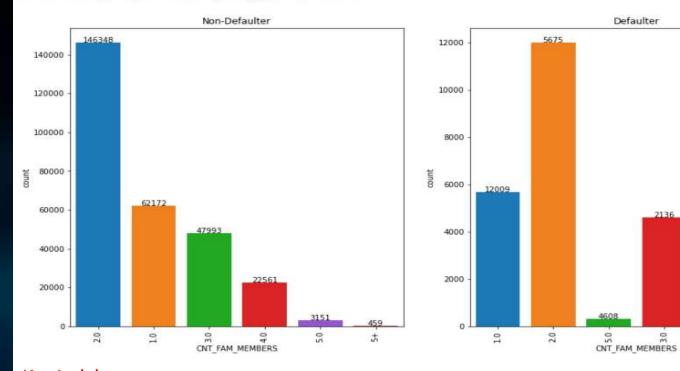

Key Insights:
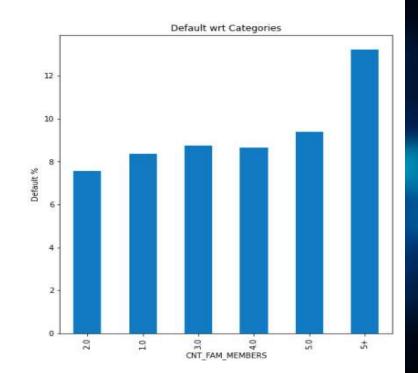
- ❖ We can observe that "Female" category have taken a greater number of loans compared to "Male" category.
- ❖ Default % of Male category is higher at 10%, as compared to Female category at 7%.

Analysis of CNT_FAM_MEMBERS column

```
Default % according to category for CNT_FAM_MEMBERS column is :
5+        13.232514
5.0        9.401955
3.0        8.760290
4.0        8.648824
1.0        8.364408
2.0        7.583498
Name: CNT_FAM_MEMBERS, dtype: float64
```


Key Insights:

❖ We can observe that maximum number of loans are availed by people with two family members, and they have the least default %.

❖ Whereas the highest defaults are made by people who have more than 5 family members in their family.

Analysis of NAME_INCOME_TYPE column

```
Default % according to category for NAME_INCOME_TYPE column is :
Others                 18.181818
Working                 9.588472
Commercial associate    7.484466
State servant           5.754965
Pensioner               5.386366
Name: NAME_INCOME_TYPE, dtype: float64
```
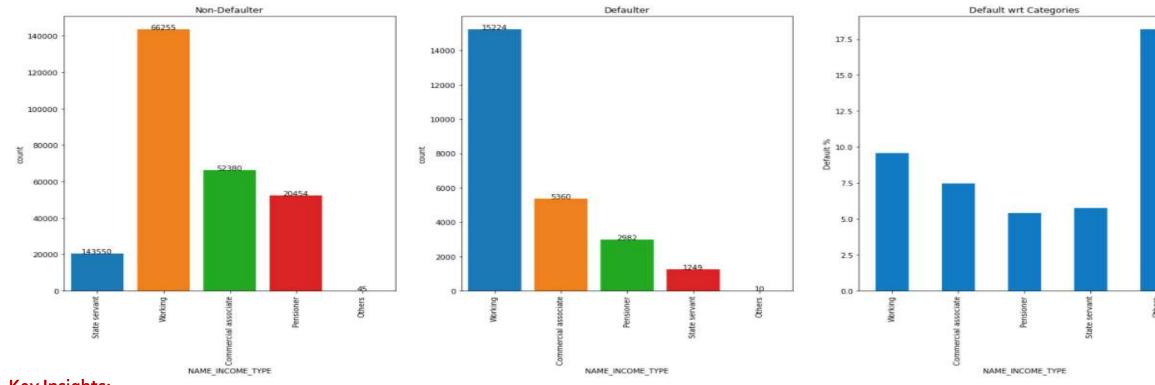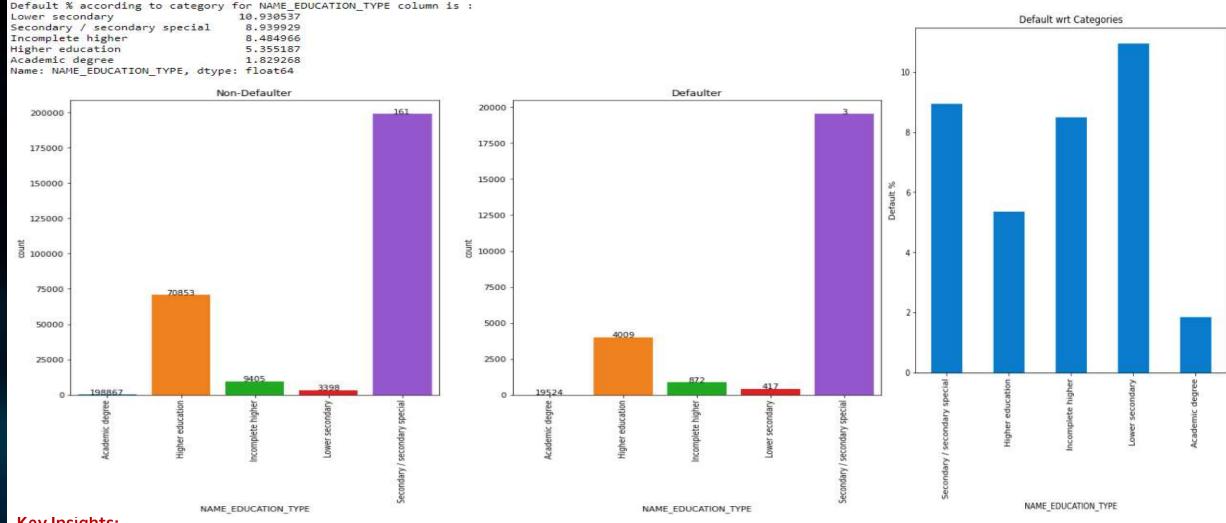

Key Insights:

❖ Most number of Loans were taken by "Working" category people.

❖ "Pensioner" category is the safest group to grant loans to.

❖ "Others" category which includes (Maternity leave, Businessman, Student, Unemployed) have the highest default rate of 18% and bank must be vigilant prior to granting loans to this category.

❖ Bank is providing a greater number of loans to the "Working" category, which has the second highest default rate.
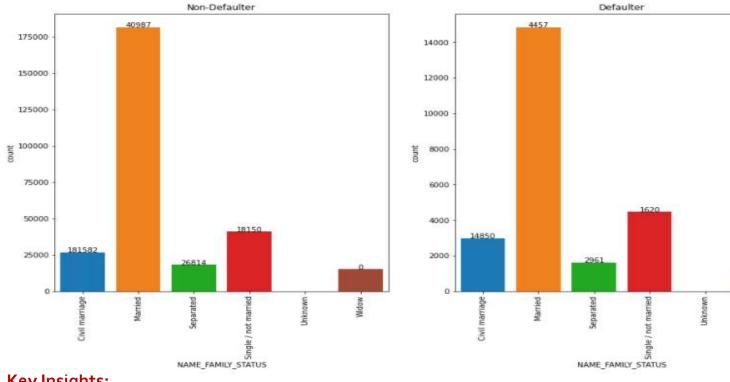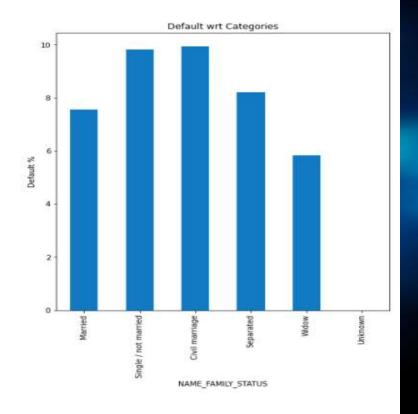
Analysis of NAME_EDUCATION_TYPE column


```
Default % according to category for NAME_EDUCATION_TYPE column is :
Lower secondary                10.930537
Secondary / secondary special   8.939929
Incomplete higher               8.484966
Higher education                5.355187
Academic degree                 1.829268
Name: NAME_EDUCATION_TYPE, dtype: float64
```

Key Insights:

- ❖ There is a trend which reflects that education level and default % are inversely related to each other.
- ❖ Most people availing loans are from the "Secondary / Secondary special" category, which is the second highest default making category. It is recommended that the bank should target more towards "Academic degree" category people, as it has the lowest default %.
- ❖ "Lower secondary" category has the highest default %, i.e., nearly 11%.

Analysis of NAME_FAMILY_STATUS column

```
Default % according to category for NAME_FAMILY_STATUS column is :
Civil marriage          9.944584
Single / not married    9.807675
Separated               8.194234
Married                 7.559868
Widow                   5.824217
Unknown                      NaN
Name: NAME_FAMILY_STATUS, dtype: float64
```
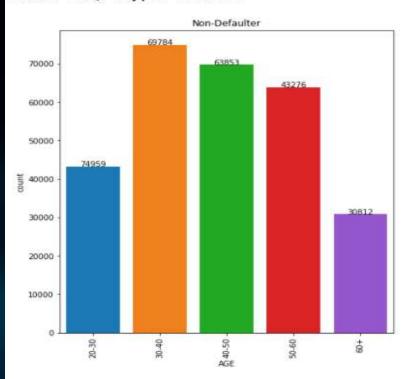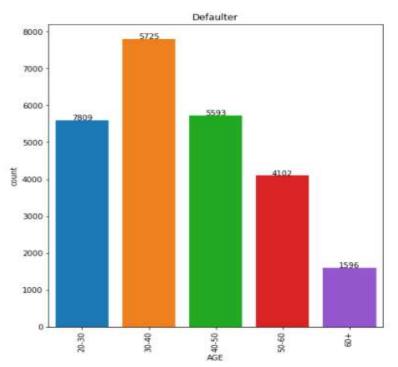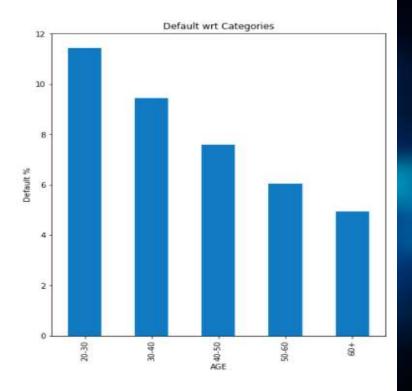

Key Insights:

❖ It is evident that maximum number of loans are availed by "Married" people category.

❖ Highest defaulting categories are "Single / not married" and "Civil marriage".

Analysis of AGE column

```
Default % according to category for AGE column is :
20-30      11.444883
30-40       9.434806
40-50       7.581878
50-60       6.036348
60+         4.924710
Name: AGE, dtype: float64
```
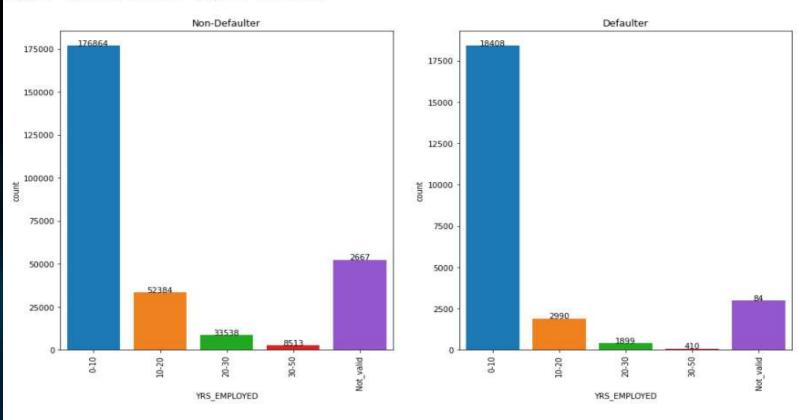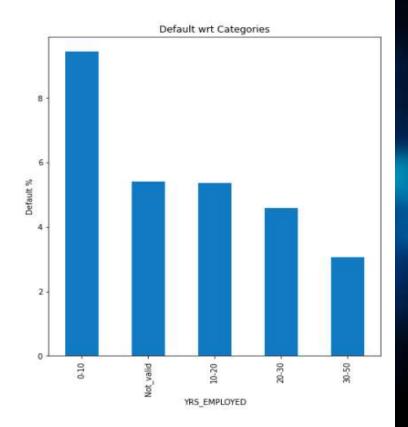

Key Insights:

❖ There is a clear relation which says that higher the age, lower is the chance of making a default.

❖ Most number of loans are given to "20-30" and "30-40" categories and they hold the highest default % of 11.4% and 9.4% respectively. It may be a good approach for bank to consider providing loans at a higher interest rate to these categories, in order to control/reduce the number of defaults.

Analysis of YRS_EMPLOYED column

```
Default % according to category for YRS_EMPLOYED column is :
0-10         9.426851
Not_valid    5.399646
10-20        5.358806
20-30        4.594867
30-50        3.053435
Name: YRS_EMPLOYED, dtype: float64
```
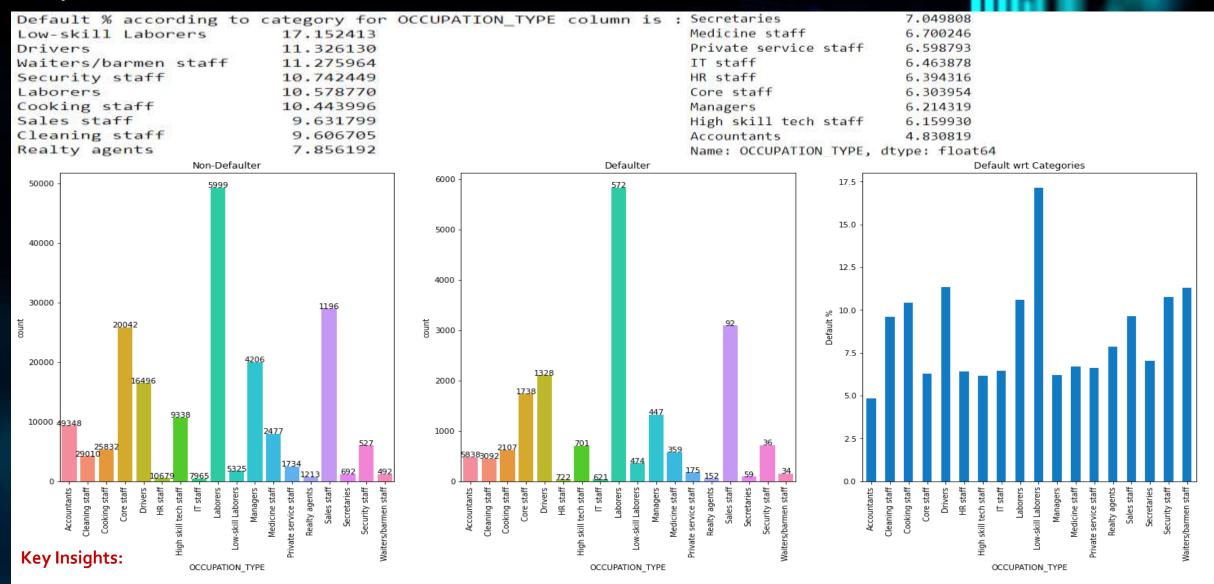

Key Insights:

❖ Higher the experience level, lower chance to default.

Analysis of OCCUPATION_TYPE column


```
Default % according to category for OCCUPATION_TYPE column is :   Secretaries                7.049808
Low-skill Laborers        17.152413                                Medicine staff             6.700246
Drivers                   11.326130                                Private service staff      6.598793
Waiters/barmen staff      11.275964                                IT staff                   6.463878
Security staff            10.742449                                HR staff                   6.394316
Laborers                  10.578770                                Core staff                 6.303954
Cooking staff             10.443996                                Managers                   6.214319
Sales staff                9.631799                                High skill tech staff      6.159930
Cleaning staff             9.606705                                Accountants                4.830819
Realty agents              7.856192                                Name: OCCUPATION_TYPE, dtype: float64
```

Key Insights:

❖ We can observe that highest loans are taken by "Laborers" category and has a default rate of around 11%. The bank must be vigilant when granting loans to this this category.

❖ Highest default is made by "Low-skill Laborers" category, i.e., around 17% and the lowest default % is "Accountants" category, i.e. around 5%.
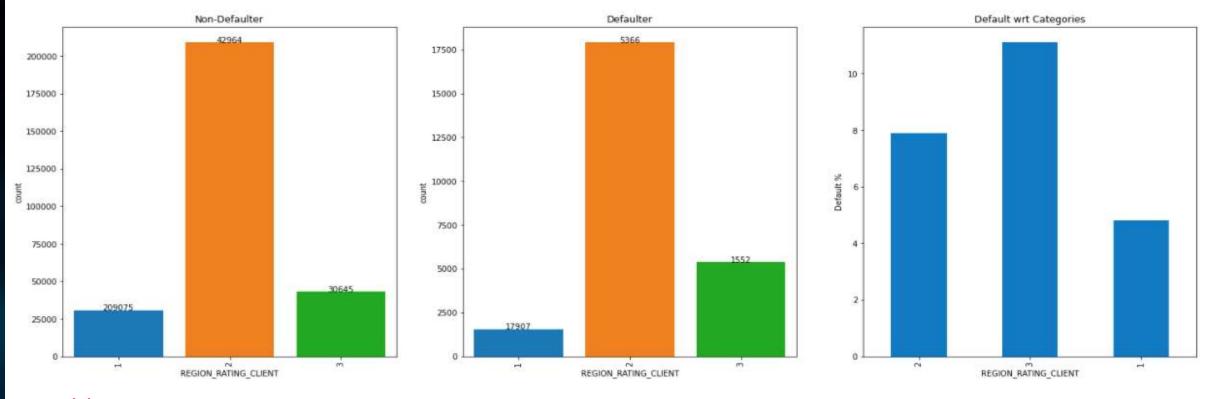
Analysis of REGION_RATING_CLIENT column

```
Default % according to category for REGION_RATING_CLIENT column is :
3    11.102835
2     7.889172
1     4.820325
Name: REGION_RATING_CLIENT, dtype: float64
```
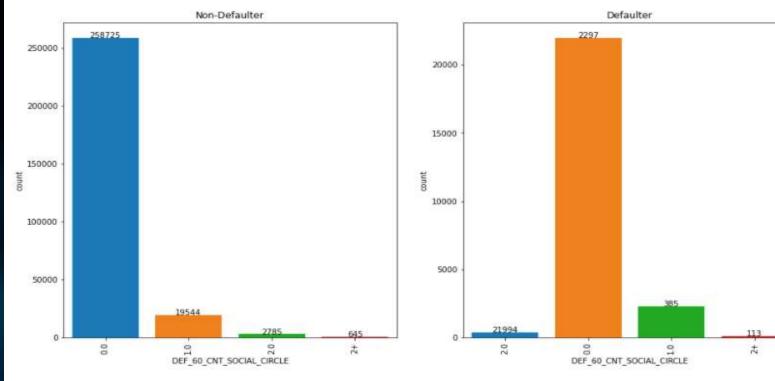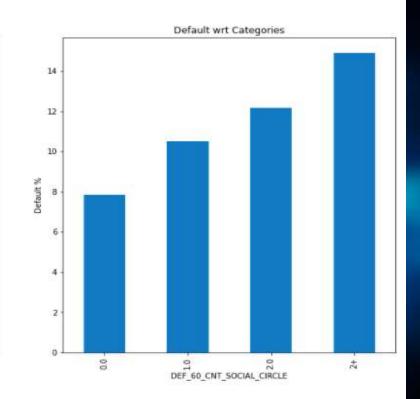

Key Insights:

❖ "2nd class region" category is availing highest number of loans and has a default rate of around 8%.

❖ Bank is providing a lot of loans to the people living in "3rd class region" category, this could be risky as they have the highest default rate of 11%. Bank should consider providing more loans to "1st class region" category, as they have low default rate of around 5%.
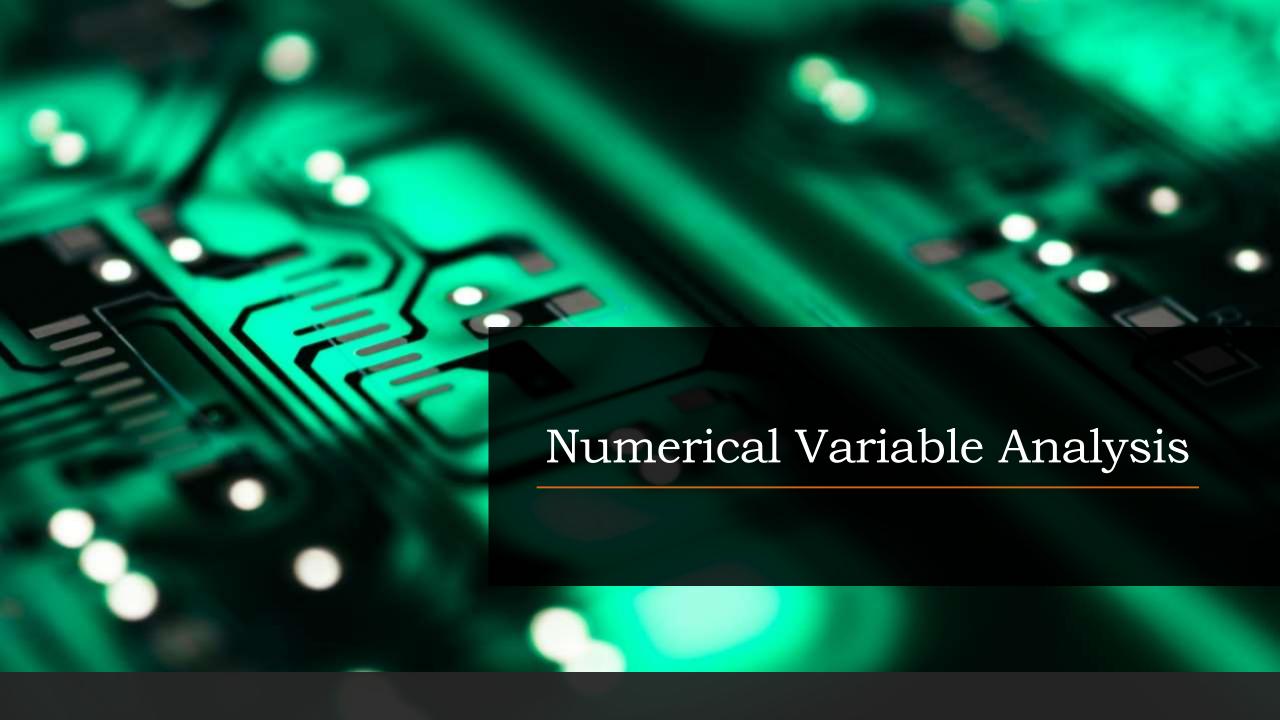
Analysis of DEF_6o_CNT_SOCIAL_CIRCLE column

```
Default % according to category for DEF_60_CNT_SOCIAL_CIRCLE column is :
2+      14.907652
2.0     12.145110
1.0     10.516918
0.0      7.834881
Name: DEF_60_CNT_SOCIAL_CIRCLE, dtype: float64
```
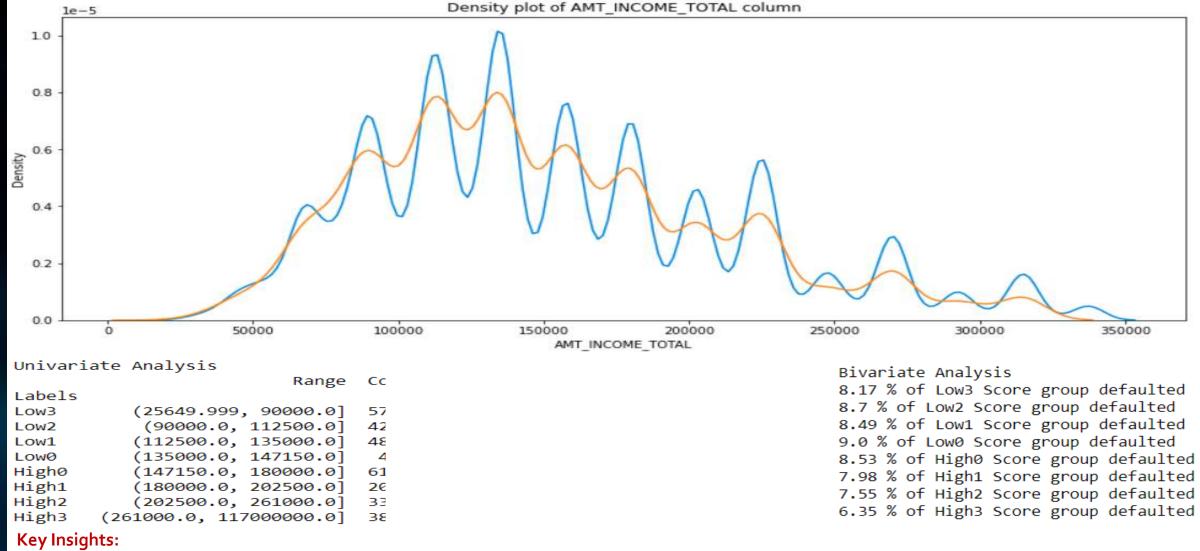

Key Insights:

❖ Most people who availed loans belong to "o" category.
❖ A positive note, bank is aware of this correlation and is granting very few loans to people of "2 and 2+" categories.
❖ There is a clear trend which says, higher the default in last 6o days or more, higher is the chance for default of loan.

Numerical Variable Analysis
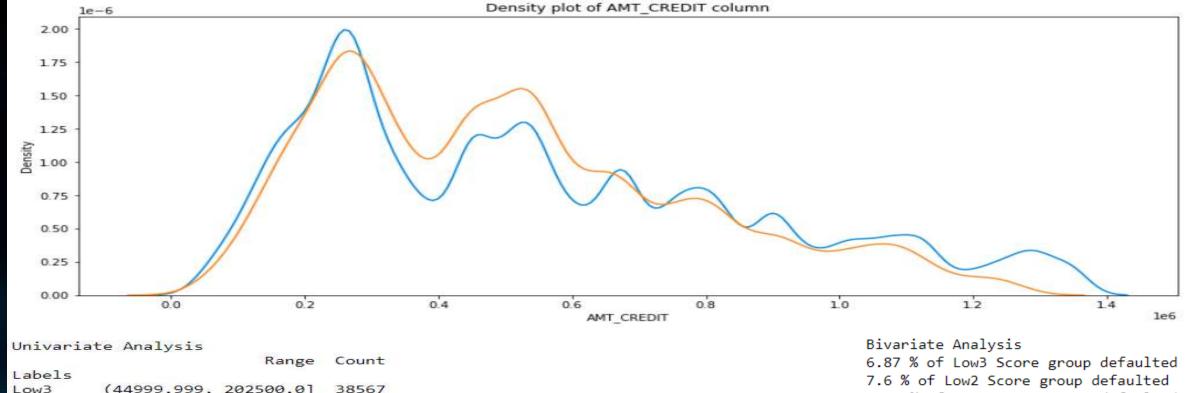
Analysis of AMT_INCOME_TOTAL column

Density plot of AMT_INCOME_TOTAL column

Univariate Analysis

Labels	Range	Cc
Low3	(25649.999, 90000.0]	57
Low2	(90000.0, 112500.0]	42
Low1	(112500.0, 135000.0]	48
Low0	(135000.0, 147150.0]	4
High0	(147150.0, 180000.0]	61
High1	(180000.0, 202500.0]	20
High2	(202500.0, 261000.0]	33
High3	(261000.0, 117000000.0]	38

Bivariate Analysis

8.17 % of Low3 Score group defaulted
8.7 % of Low2 Score group defaulted
8.49 % of Low1 Score group defaulted
9.0 % of Low0 Score group defaulted
8.53 % of High0 Score group defaulted
7.98 % of High1 Score group defaulted
7.55 % of High2 Score group defaulted
6.35 % of High3 Score group defaulted

Key Insights:

❖ There is a trend here, people with low and high salary are making less defaults. High defaults are made by people in mid salary range.
❖ Highest default is made by people who earn salary in range of "135000.0 to 147150.0".

Analysis of AMT_CREDIT column

Density plot of AMT_CREDIT column

Univariate Analysis
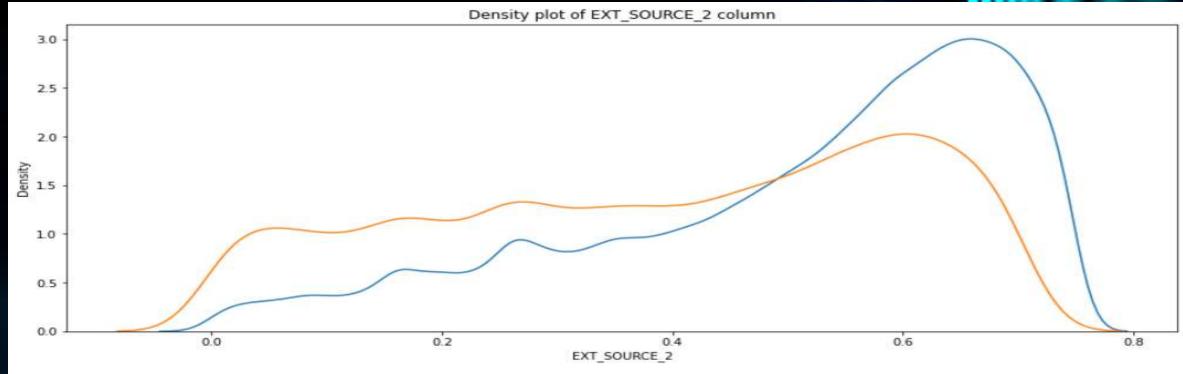
| Labels	Range	Count
Low3	(44999.999, 202500.0]	38567
Low2	(202500.0, 270000.0]	39854
Low1	(270000.0, 385164.0]	36969
Low0	(385164.0, 513531.0]	38459
High0	(513531.0, 646920.0]	38368
High1	(646920.0, 808650.0]	39416
High2	(808650.0, 1078200.0]	37749
High3	(1078200.0, 4050000.0]	38127

Bivariate Analysis

6.87 % of Low3 Score group defaulted
7.6 % of Low2 Score group defaulted
9.66 % of Low1 Score group defaulted
9.7 % of Low0 Score group defaulted
10.65 % of High0 Score group defaulted
7.71 % of High1 Score group defaulted
7.05 % of High2 Score group defaulted
5.39 % of High3 Score group defaulted

Key Insights:

❖ There is a trend here, people with low and high "AMT_CREDIT" are making less defaults. High defaults are made by people in mid credit range.
❖ There has been a high default rate between the loan range of "270000 to 800000".

Analysis of EXT_SOURCE_2 column

Density plot of EXT_SOURCE_2 column

Univariate Analysis

	Range	Count
Labels		
Low3	(-0.0009999183, 0.254]	38371
Low2	(0.254, 0.392]	38345
Low1	(0.392, 0.496]	38356
Low0	(0.496, 0.566]	38354
High0	(0.566, 0.618]	38356
High1	(0.618, 0.664]	38356
High2	(0.664, 0.711]	38355
High3	(0.711, 0.855]	38356

Bivariate Analysis
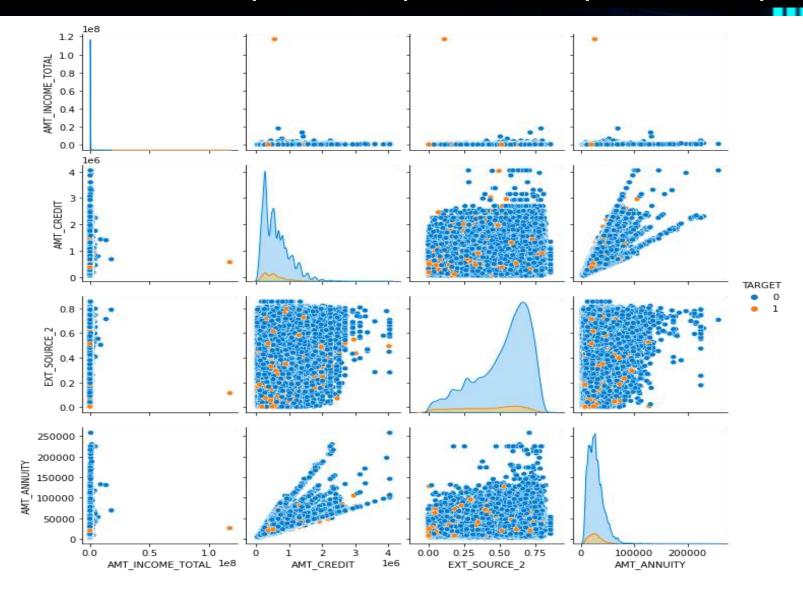
17.4 % of Low3 Score group defaulted
11.18 % of Low2 Score group defaulted
8.83 % of Low1 Score group defaulted
7.52 % of Low0 Score group defaulted
6.61 % of High0 Score group defaulted
5.43 % of High1 Score group defaulted
4.47 % of High2 Score group defaulted
3.14 % of High3 Score group defaulted

Key Insights:

❖ There is clear trend that lower the score, more is the chance to default.
❖ "Low3 Score" group has the highest default rate of 17.4% and "High3 Score" group has the lowest default rate of 3%.

Bivariate Analysis

Numerical Variable Analysis

Seaborn pair plot of AMT_INCOME_TOTAL, AMT_CREDIT, EXT_SOURCE_2, AMT_ANNUITY, TARGET columns

Key Insights:
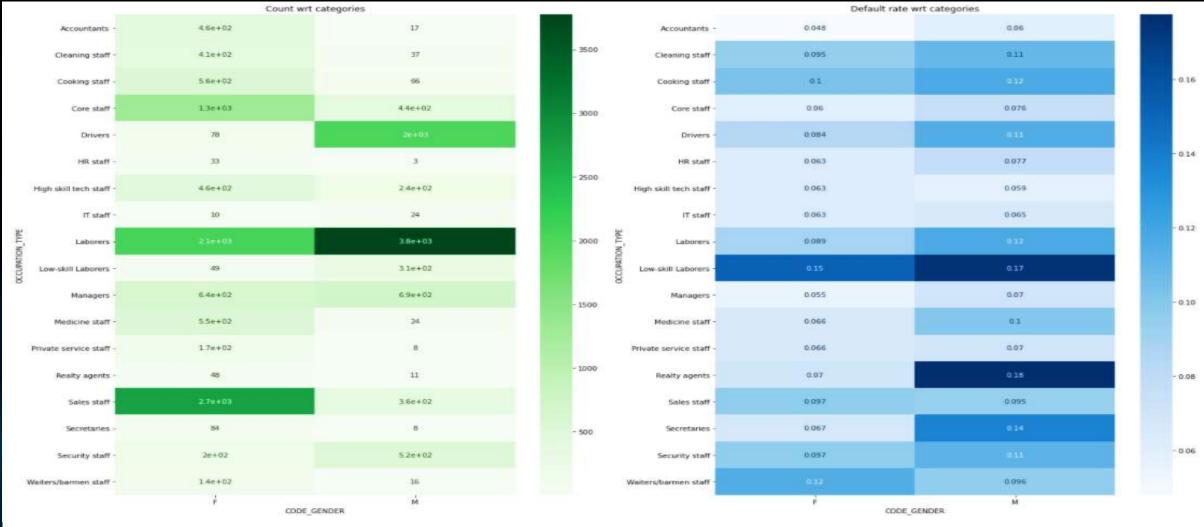
❖ We can observe that higher amount of credits are majorly seen with non-defaulter category, which is a positive news for the bank.

Categorical Variable Analysis

Analysis of CODE_GENDER','OCCUPATION_TYPE' wrt default
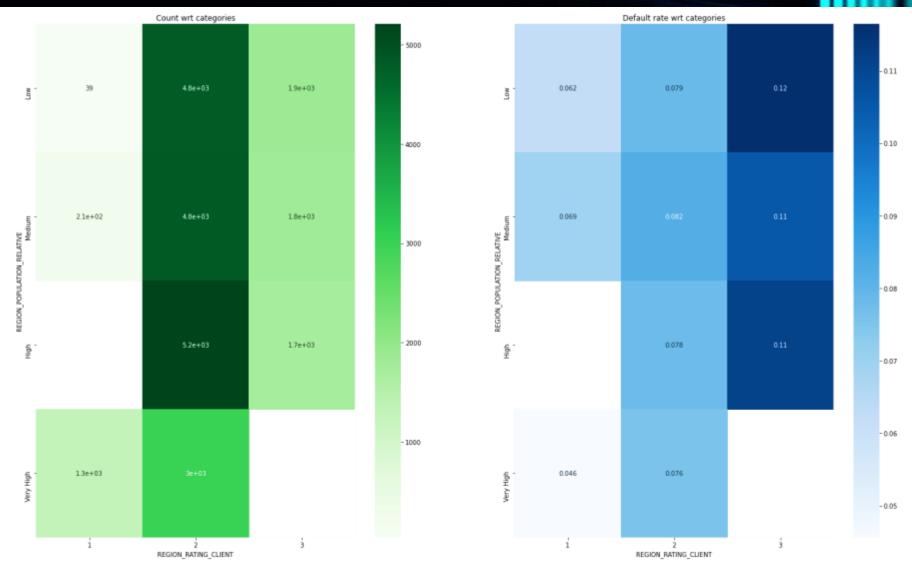
Key Insights:

Why do men default more than women?

❖ As we can see here, there are more men in the job roles like "Laborers, Low-skill Laborers and Driver" categories compared to women. These categories make up the highest default making categories. Therefore, men have more default rates than women.

❖ Women are majorly at the roles of "Sales staff, core-staff and laborers" which have low default rates.

❖ Bank has provided a lot of loans to "Laborers and Low-skill Laborers" in male category, these categories have one of the highest default rates. The bank must be more vigilant and mindful about this observation.

❖ Highest default rate is in the "Realty agents" male category, i.e., around 18%.
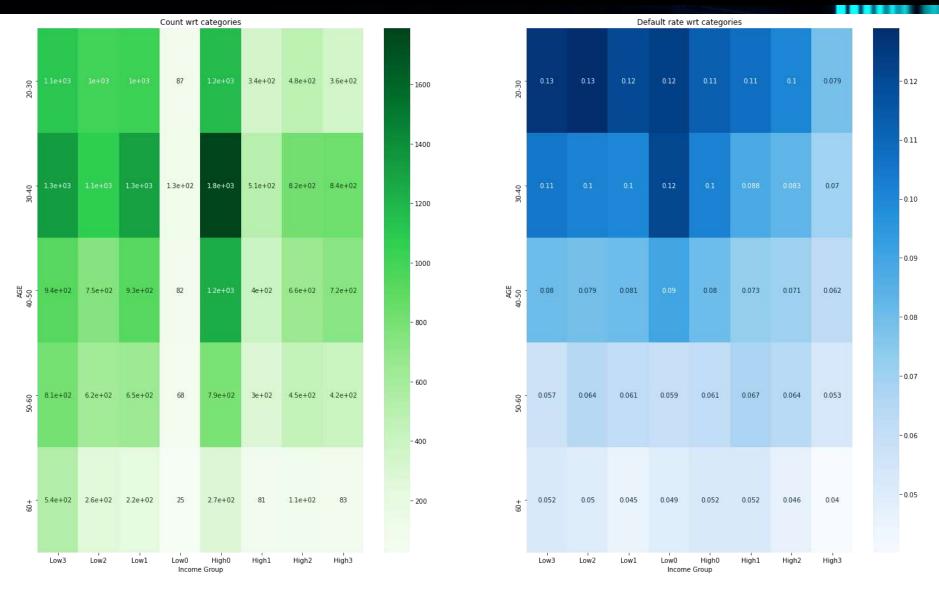
Key Insights:

❖ There is a clear pattern which indicates that, if a person belongs to "1st class" region and is having very high population, the default rate is very low, i.e., 4.6% and if a person who belongs to "3rd class" region and having low population, have high default rate of 12%. Hence, population and default rate are inversely proportional.

❖ There is trend across population which indicates that, higher the population, lesser the default rate. This can also be interpreted as high population regions are majorly cities, where more educated people live. Whereas, low population places denote villages and less literacy, hence more default rate.
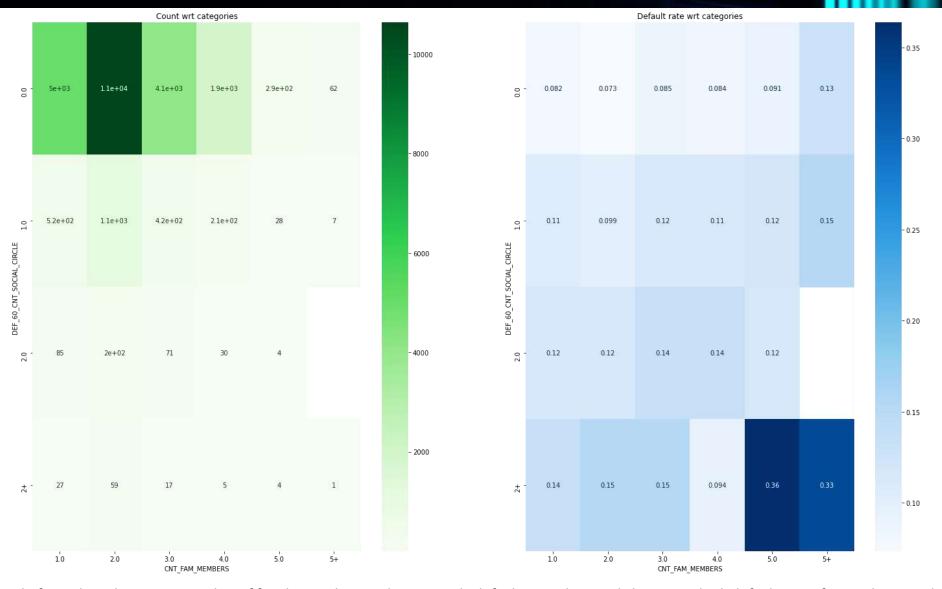
Analysis of 'INCOME GROUP','AGE' wrt default

Key Insights:
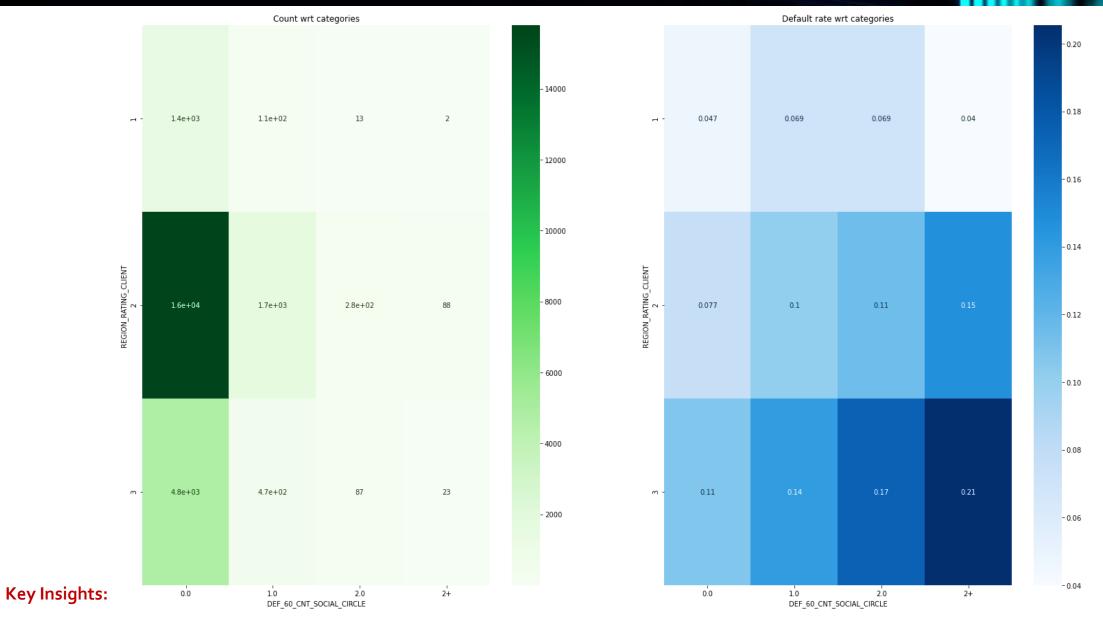
❖ There is clear trend which indicates that, people of category "60+" with very high income is a good category to provide loans to, as it has only 4% default rate. People aged 20-30 and having low income, have high default rate of 13%. However, banks must still exercise caution when giving loans to 60+ category, as this class of people are older in age and there is a risk of recovery of loan in case of death.

Analysis of CNT_FAM_MEMBERS,DEF_60_CNT_SOCIAL_CIRCLE columns

Key Insights:
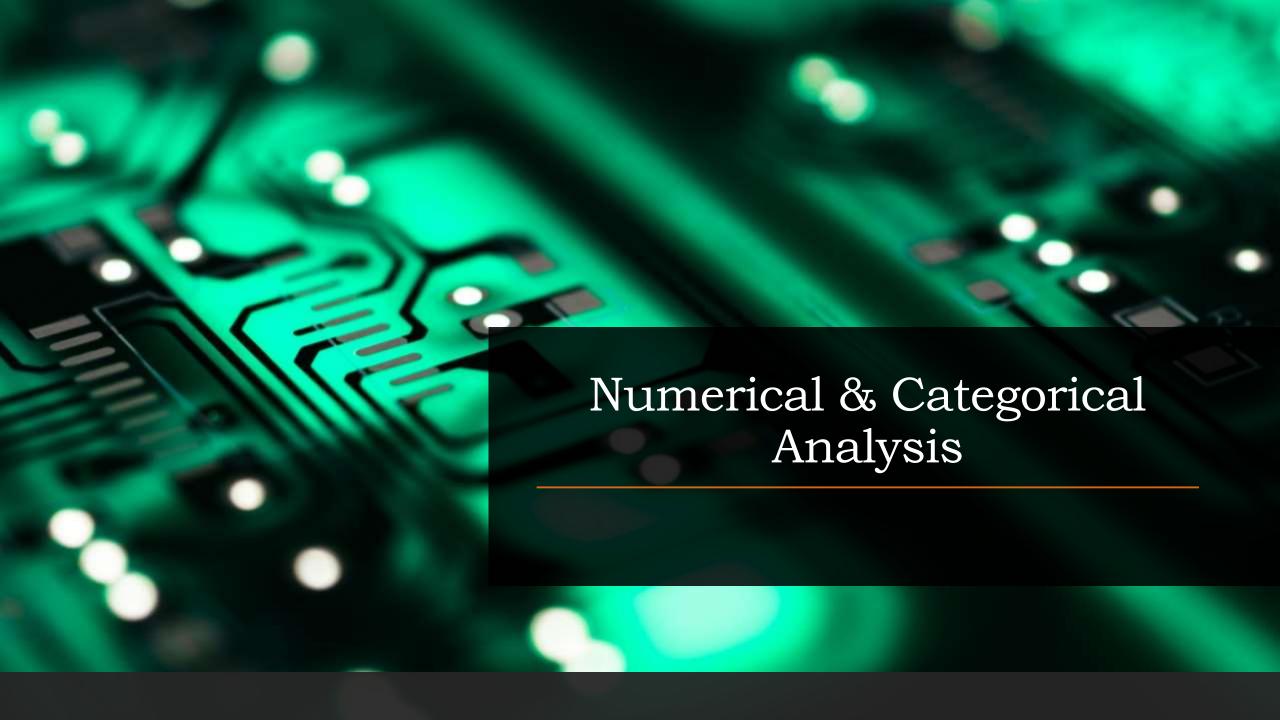
❖ There is a clear trend of people with "5 or 5+" number of family members and "2+" people defaulting in their circle have very high default rate of around 36% and people with "2" family members and no people defaulting in their circle have low default rate of 7.3%.

❖ As the number of members in the family increases and number of default in their social circle increases, the default rate also increases significantly.
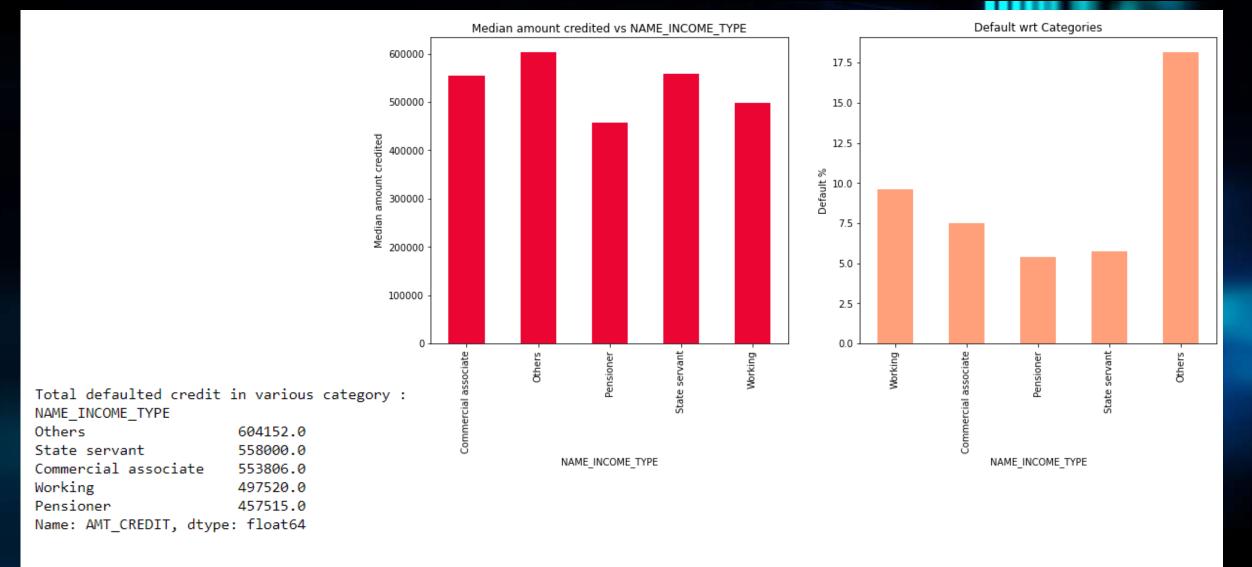
Analysis of DEF_6o_CNT_SOCIAL_CIRCLE,REGION_RATING_CLIENT columns

Key Insights:

❖ There is a clear trend that as the "REGION_RATING_CLIENT" and "DEF_6o_CNT_SOCIAL_CIRCLE" decreases, the default rate also decreases and vice versa.
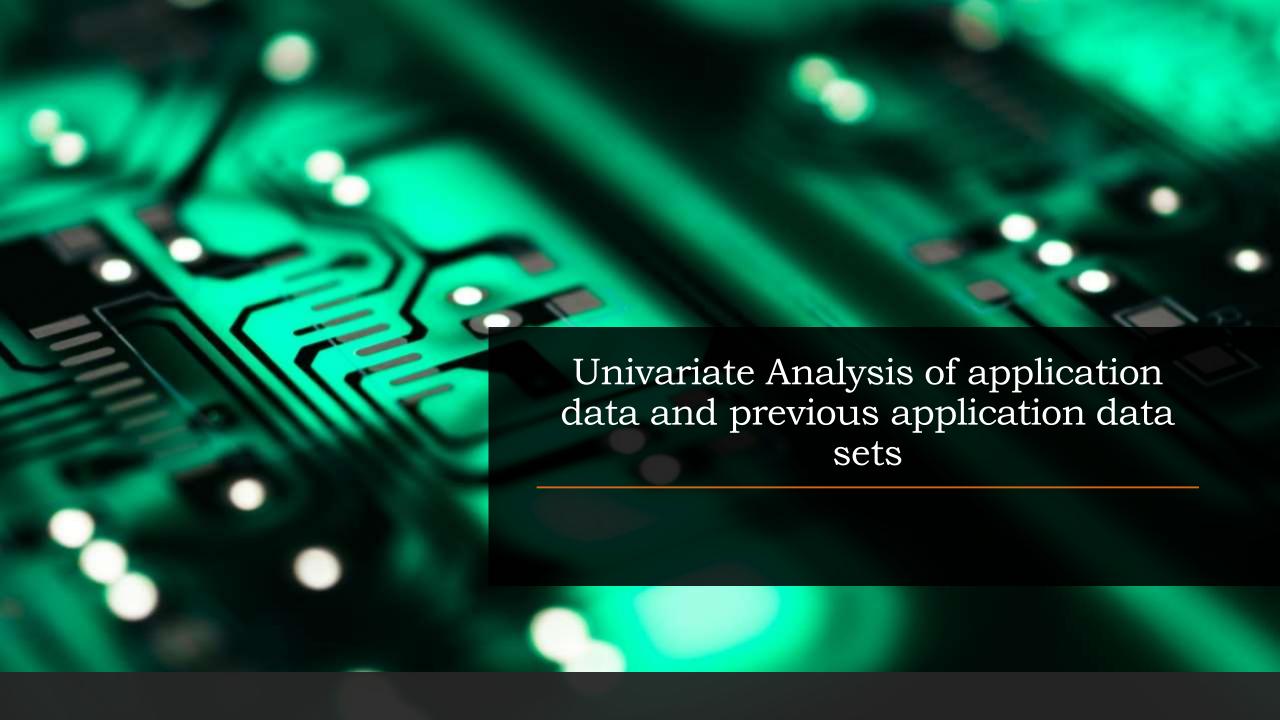
Numerical & Categorical Analysis

Analysis of NAME_INCOME_TYPE column wrt default

Median amount credited vs NAME_INCOME_TYPE

Default wrt Categories

```
Total defaulted credit in various category :
NAME_INCOME_TYPE
Others                   604152.0
State servant            558000.0
Commercial associate     553806.0
Working                  497520.0
Pensioner                457515.0
Name: AMT_CREDIT, dtype: float64
```

Key Insights:
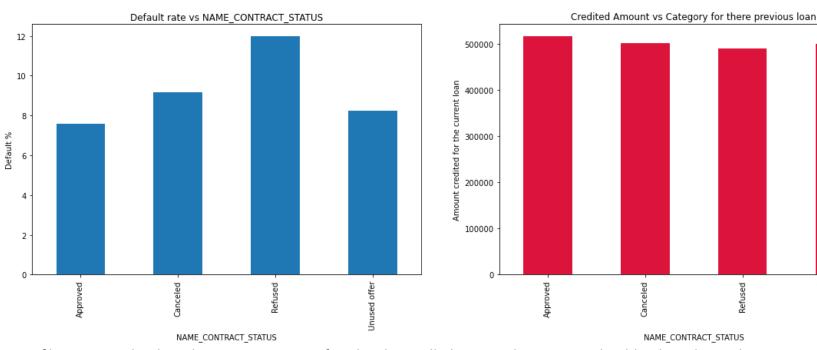
❖ Bank is granting high amount of loans to "Other category" which includes "Maternity leave, Businessman, Student, Unemployed", which has the highest default rate of around 18%. Therefore, bank should shift to "Pensioner" or "State servant" category for granting higher amount of loans, as it has low default rate of 5%.

Univariate Analysis of application data and previous application data sets

Analysis of NAME_CONTRACT_STATUS column

```
Median amount given to category of NAME_CONTRACT_STATUS column is:
NAME_CONTRACT_STATUS
Approved        517500.0
Canceled        501435.0
Unused offer    500211.0
Refused         490495.5
Name: AMT_CREDIT_x, dtype: float64
Default percentage for the categories of NAME_CONTRACT_STATUS column is:
NAME_CONTRACT_STATUS
Refused         11.996414
Canceled         9.173569
Unused offer     8.251724
Approved         7.588655
Name: TARGET, dtype: float64
```


Key Insights:
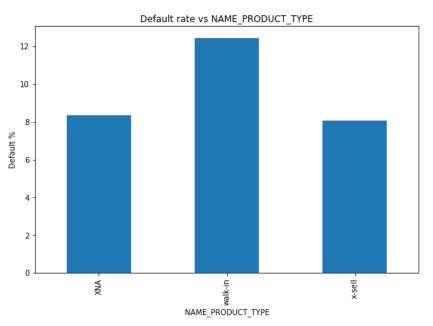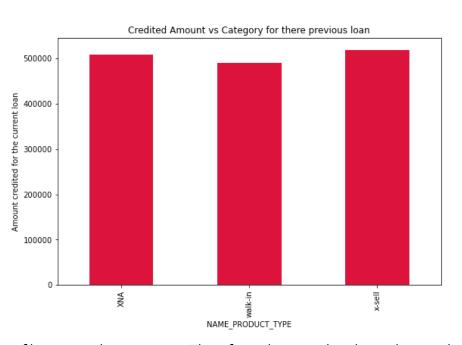
❖ Bank is granting high amounts (Rs. 490495.5) of loan to people whose loan contract was refused and cancelled previously. Hence, it should reduce the credit amount as these categories have high default rate of 12% and 9% respectively.

Analysis of NAME_PRODUCT_TYPE column

```
Median amount given to category of NAME_PRODUCT_TYPE column is:
NAME_PRODUCT_TYPE
x-sell      519633.0
XNA         508495.5
walk-in     490536.0
Name: AMT_CREDIT_x, dtype: float64
Default percentage for the categories of NAME_PRODUCT_TYPE column is:
NAME_PRODUCT_TYPE
walk-in     12.469007
XNA          8.354841
x-sell       8.087827
Name: TARGET, dtype: float64
```
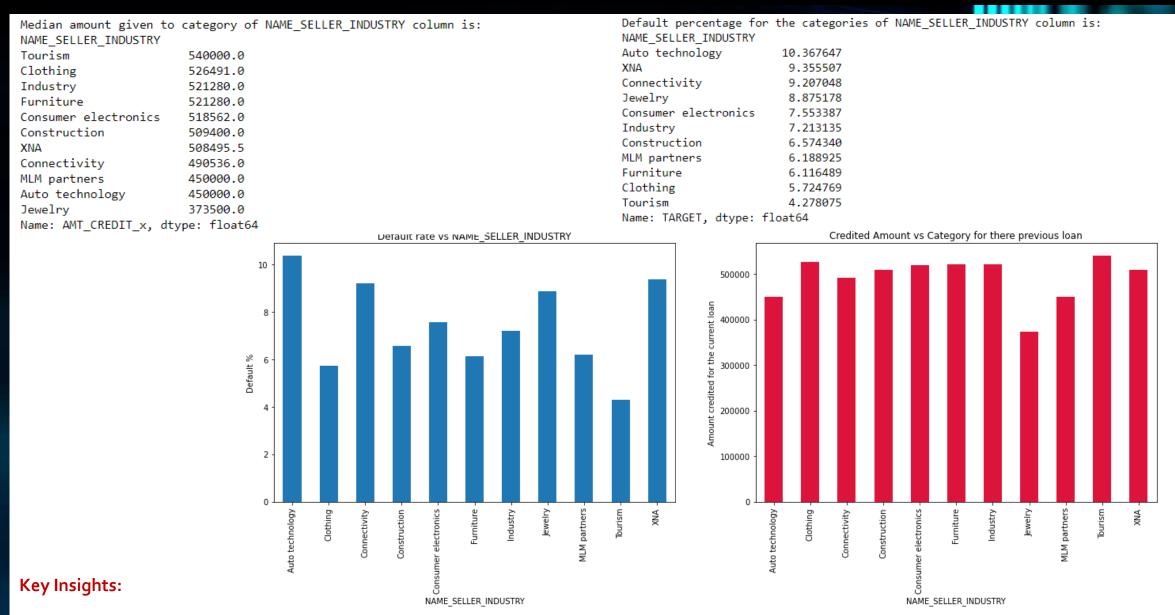

Default rate vs NAME_PRODUCT_TYPE

Credited Amount vs Category for there previous loan

Key Insights:

❖ "Walk-In" category is making highest default, i.e., 12% but bank is granting them almost as high amount of loans as other category. Therefore, this must be planned properly to avoid or reduce any potential risk.
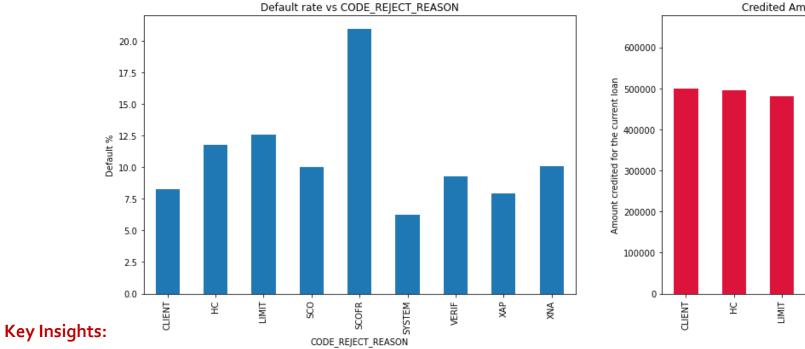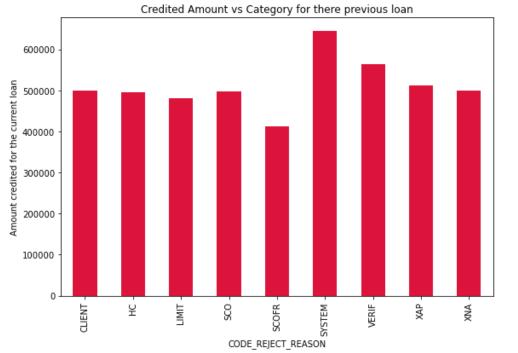
Analysis of NAME_SELLER_INDUSTRY column

```
Median amount given to category of NAME_SELLER_INDUSTRY column is:
NAME_SELLER_INDUSTRY
Tourism                  540000.0
Clothing                 526491.0
Industry                 521280.0
Furniture                521280.0
Consumer electronics     518562.0
Construction             509400.0
XNA                      508495.5
Connectivity             490536.0
MLM partners             450000.0
Auto technology          450000.0
Jewelry                  373500.0
Name: AMT_CREDIT_x, dtype: float64
```

```
Default percentage for the categories of NAME_SELLER_INDUSTRY column is:
NAME_SELLER_INDUSTRY
Auto technology          10.367647
XNA                       9.355507
Connectivity              9.207048
Jewelry                   8.875178
Consumer electronics      7.553387
Industry                  7.213135
Construction              6.574340
MLM partners              6.188925
Furniture                 6.116489
Clothing                  5.724769
Tourism                   4.278075
Name: TARGET, dtype: float64
```


Key Insights:

❖ Highest loan default is made by "Auto technology" category, i.e., 10% and lowest default is made by "Tourism" category, i.e., 4%. Bank has taken this into consideration and has distributed the loans keeping these factors in mind.

Analysis of CODE_REJECT_REASON column

```
Median amount given to category of CODE_REJECT_REASON column is:
CODE_REJECT_REASON
SYSTEM    645138.0
VERIF     563269.5
XAP       512064.0
CLIENT    500211.0
XNA       499500.0
SCO       497520.0
HC        495000.0
LIMIT     479974.5
SCOFR     411813.0
Name: AMT_CREDIT_x, dtype: float64
```

```
Default percentage for the categories of CODE_REJECT_REASON column is:
CODE_REJECT_REASON
SCOFR     20.928736
LIMIT     12.559395
HC        11.734163
XNA       10.095934
SCO        9.998162
VERIF      9.256252
CLIENT     8.251724
XAP        7.947654
SYSTEM     6.250000
Name: TARGET, dtype: float64
```


Key Insights:

❖ Highest default is made by "SCOFR" category, i.e., is nearly 21% and lowest default is made by "SYSTEM" category, i.e., 6%. Bank has distributed their line of credit keeping this factor in mind.

Thank you