**Project 6 : By Ramana Bansal**

## Bank Loan Case Study

Colab notebook Link:

<https://colab.research.google.com/drive/1gdARCHcwWReZ1gnJfT9DFgsUbz-9wL8f?usp=sharing>

Video Link:

**https://drive.google.com/file/d/1F2NBCVzVq82tYzOXjDrk0B7p-JP-zbtQ/view?usp=sharing**

**Project Description:** The project deals with risk analytics related to loan applications in a bank. The aim of the project is to use EDA to identify the factors and patterns which may indicate that an applicant might have difficulty in loan payment and use this to identify the applications that should be approved or not, in order to reduce loan defaults.

**Problem Statement: Analyzing provided data to predict whether an applicant might default in loan payment or not.**

**Data Sets:**

* application\_data : Data regarding the current applications and applicants’ details.
* previous\_application\_data : Data regarding the previous applications of applicants.

**Analysis Approach:** The two datasets were initially processed and analyzed separately, and then the data was merged. The following steps were taken:

1. Importing required libraries : numpy, pandas, matplotlib and seaborn.
2. Mounting Google Drive : Since Google Colab allocates fresh RAM for every session, files need to be uploaded for every session. With heavier files, it’s better to simply connect Colab to Drive for easier access to files.
3. Working with application\_data file and previous\_application\_data:

* Understanding data
* Removing columns with high null data
* Removing duplicates
* Checking data imbalance, before and after merging
* Dividing data into Categorical, Discrete and Numerical columns and working on them separately for

1. Working on missing or unknown values
2. Changing datatypes
3. Treating Outliers
4. Univariate Analysis
5. Bivariate Analysis
6. Finding Correlation
7. Visualization

**Tech-Stack Used:** The data was processed and analyzed using Google Colab.

**Learning Insights:** The analysis highlighted various features which might aid in predicting whether an applicant might default or not. It also helped in understanding the type of loan applications and the type of loan applicants a bank gets.

The project gave me an opportunity to revisit Python as well as learn some of its required libraries. It also helped me to understand the work and approach required to work with large amount of data. However, it was a little difficult for me to draw insights from data. The project highlighted the need to work on the same.

## Missing Data

1. **Identify the** missing data **and use appropriate method to deal with it.**

First, we checked the percentage of null values for each column. For application\_data, the columns with null percentage > 40% were dropped. For previous\_application\_data, the columns with null percentage > 50% were dropped.

Then, we checked the columns for XNA values in Categorical columns. If less in number, these were replaced by mode, else, they were replaced by NAN. The percentage of null values was then rechecked. The columns with XNA>50% were also removed.

For numerical columns with few missing values, the outliers were checked. In case of presence of outliers, the null values were imputed with median. If there were no outliers, the null values were replaced by mean. If the number of missing values was high, no imputation was made.

## Outliers

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

The outliers in numerical columns were checked using Box-plot. The values falling above or below the IQR values were considered outliers. There were no outliers below the lower bound. The outliers lying above IQR were capped with 99 percentile values instead of being removed.

For some of DAYS columns, an error value of 365243 (~100 years) was observed. This value was NaN. The DAYS columns were then converted into Years and stored in dataframes.

## Data Imbalance

1. **Identify if there is data imbalance in the data. Find the ratio of data imbalance.**

Since the major aim of study was to differentiate between people with payment difficulties (defaulters) and non-defaulters, the TARGET column was used to check data imbalance. The column had following two values:

0: Applicants with no payment difficulties (Non-Defaulters)

1: Applicants with payment difficulties (Defaulters)

The ratio between the above two values was found to check Data Imbalance.

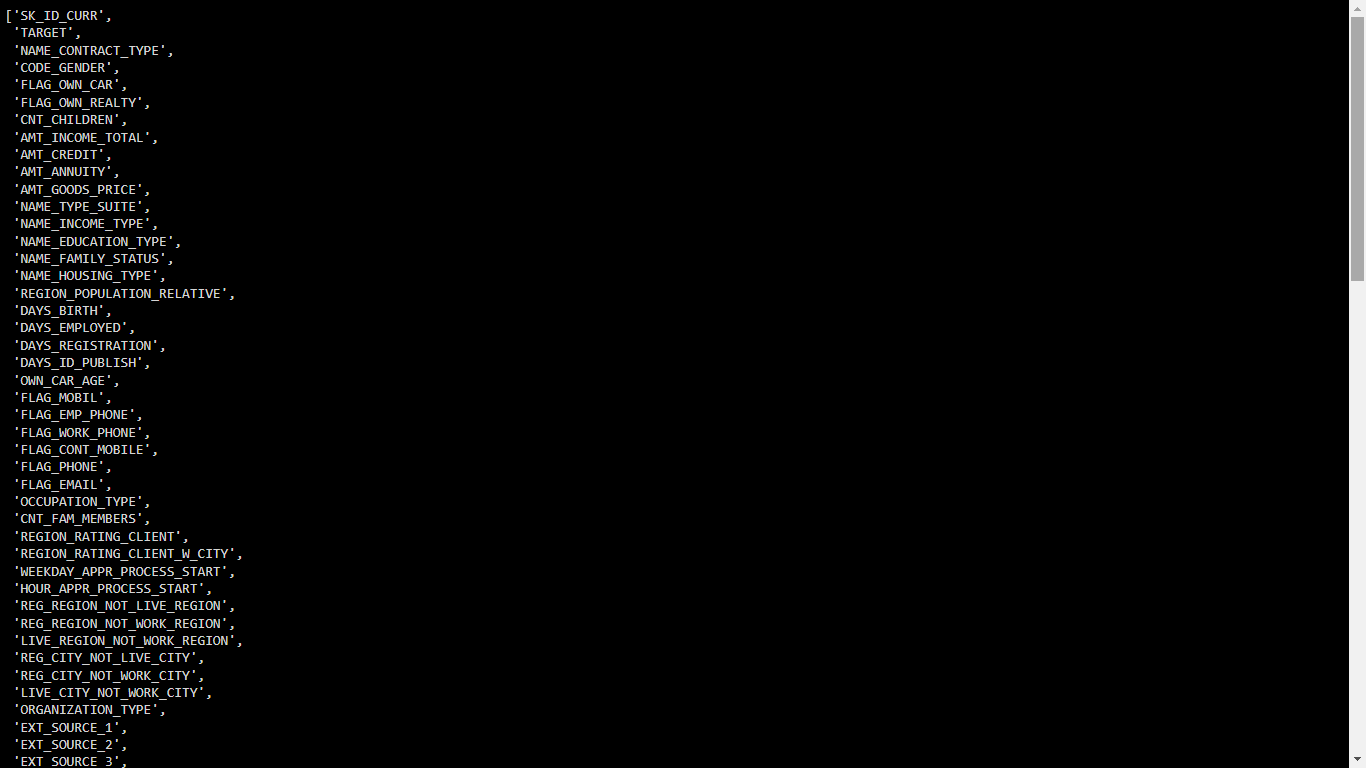
# Working on application\_data

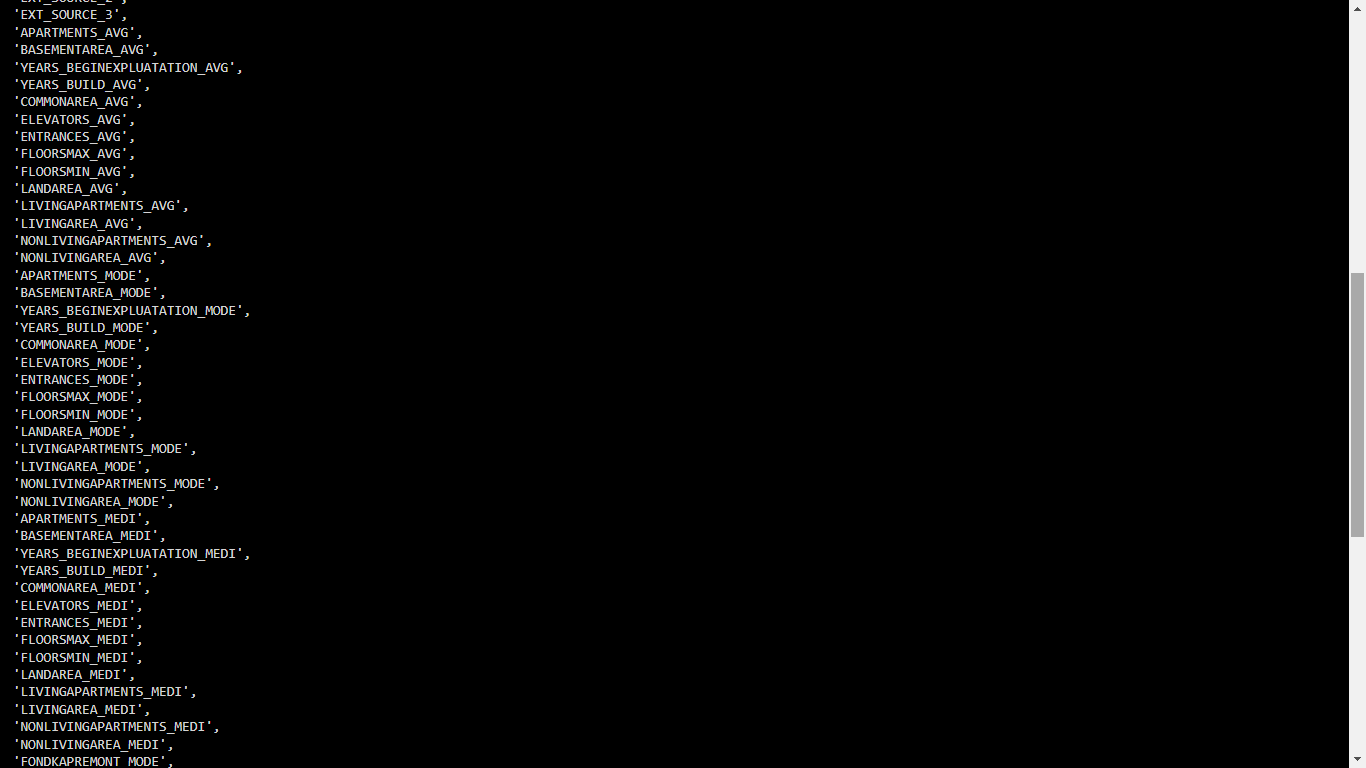
## Description

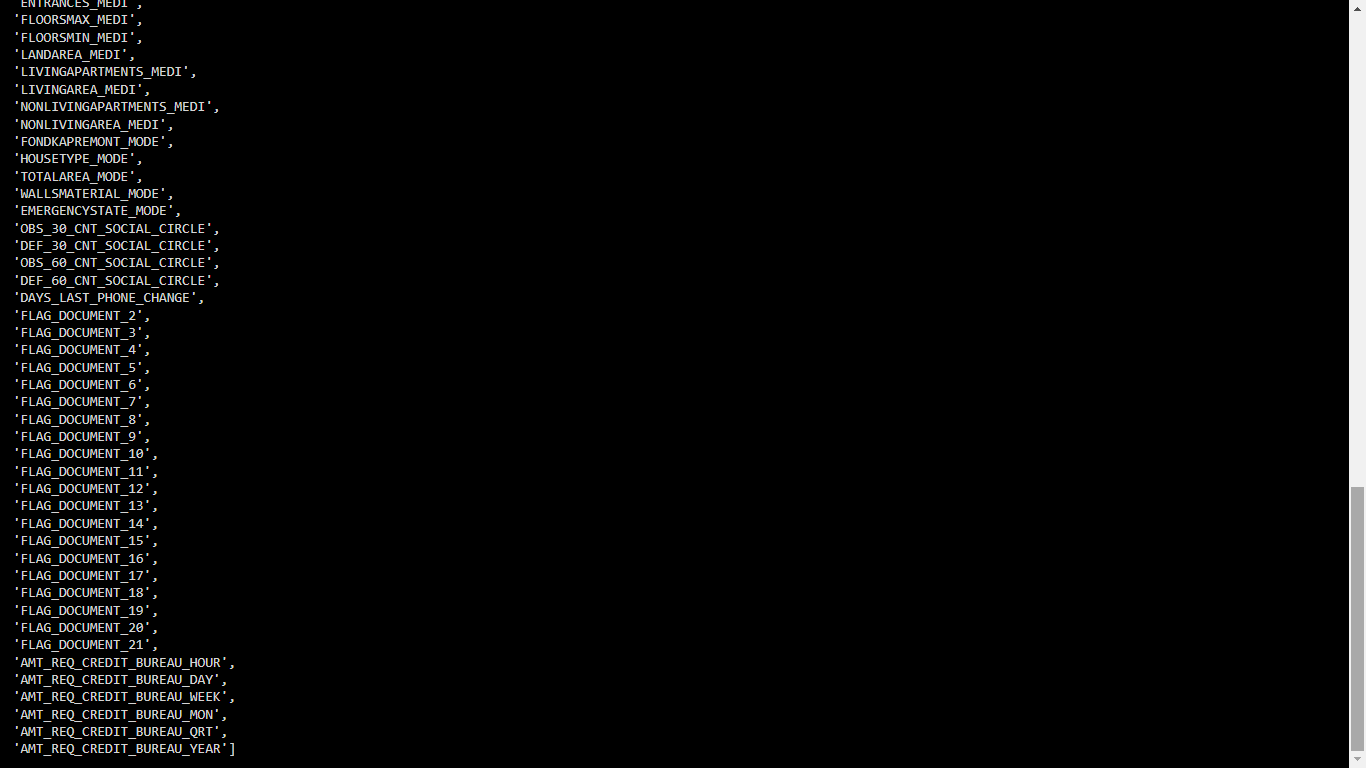
The dataframe app\_data has 122 columns and 307511 rows. There are 65 columns with float datatype, 41 with int and 16 with object datatype.



COLUMNS:







## Irrelevant Columns

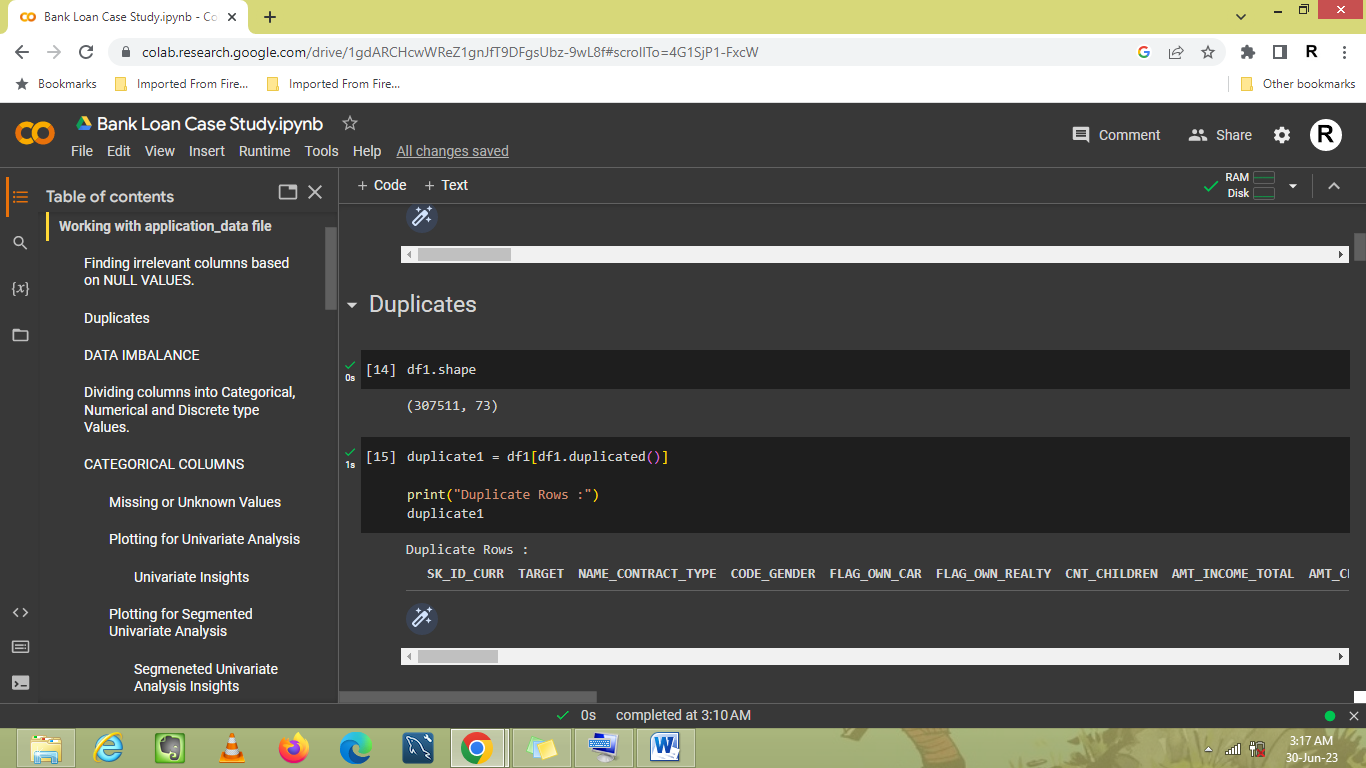
The following **columns with null values > 40%** were removed.



43 columns were removed from app\_data and resulting data was stored in df1. Df1 has 73 columns.

## Duplicates

No duplicates were found in df1.

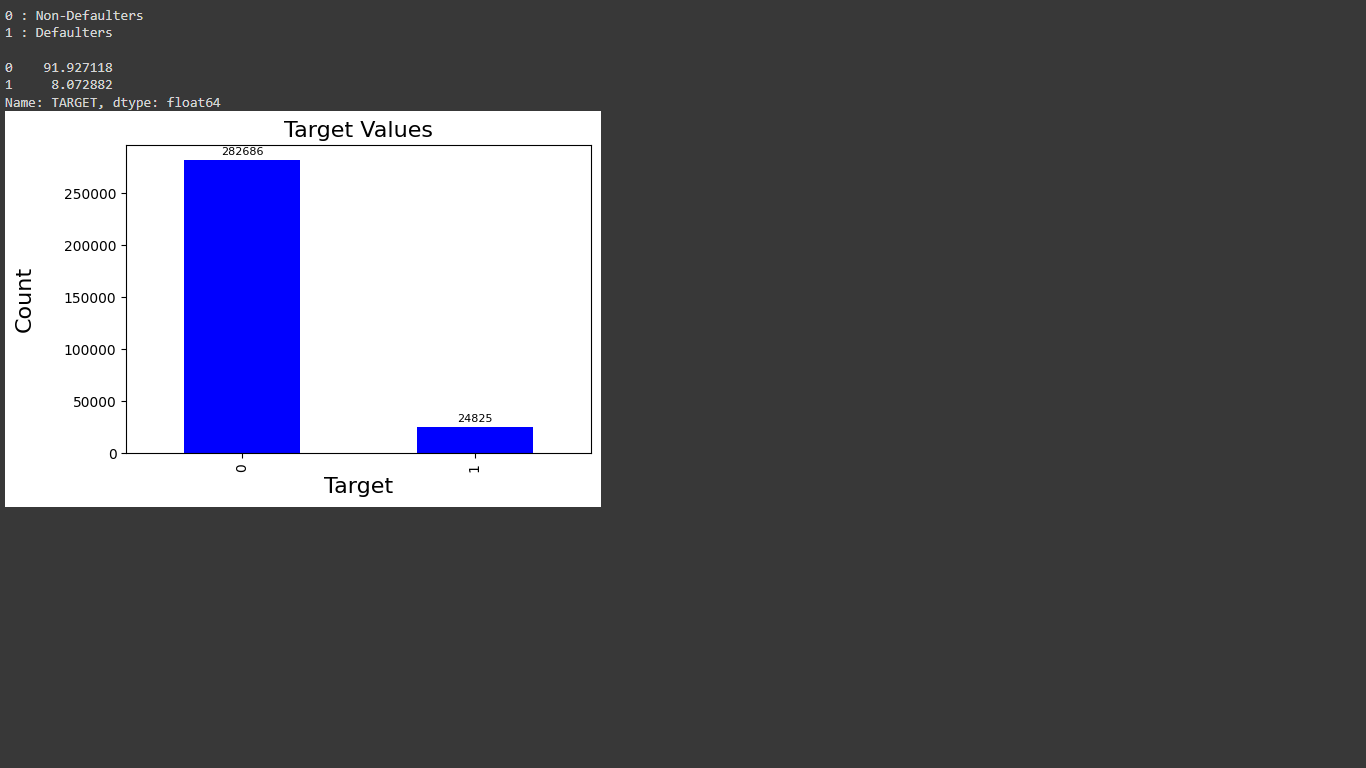


## Data Imbalance

Since the major aim of study is to look into applicants with paying difficulties, the target column will be used to check for data imbalance.

# Value = 0 indicates No Payment Difficulties (Non-Defaulters).

# Value = 1 indicates Payment Difficulties (Defaulters).

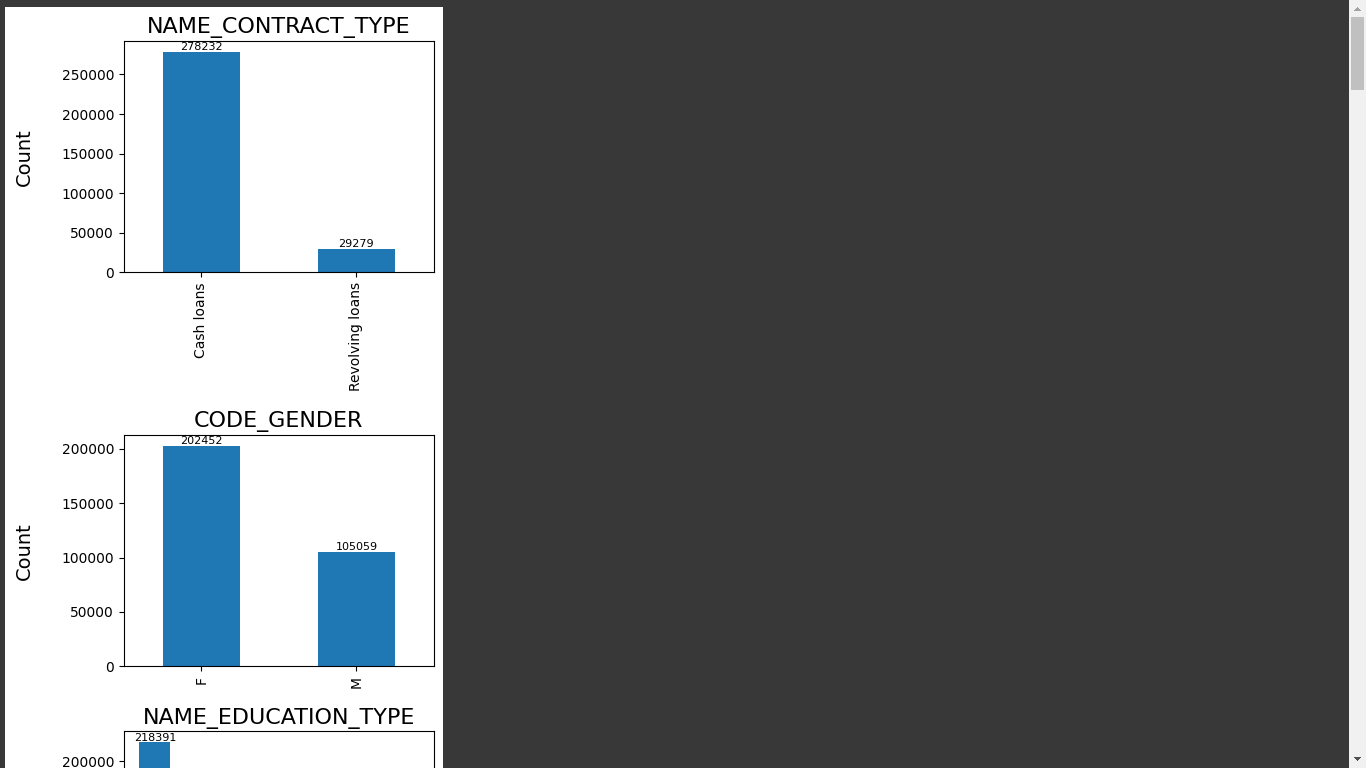


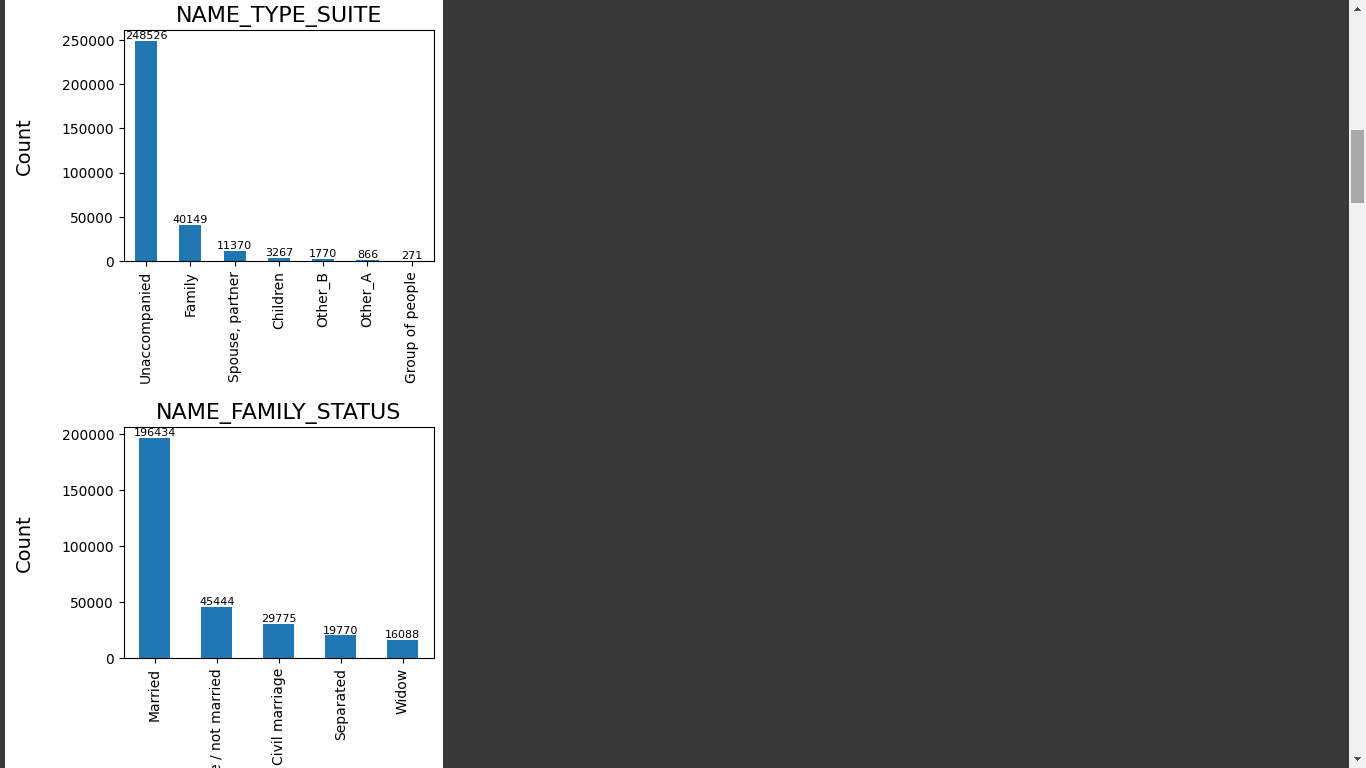
Data Imbalance ratio of 23:2 indicates the number/data of non-defaulters is much higher than that of defaulters.

## Univariate Analysis

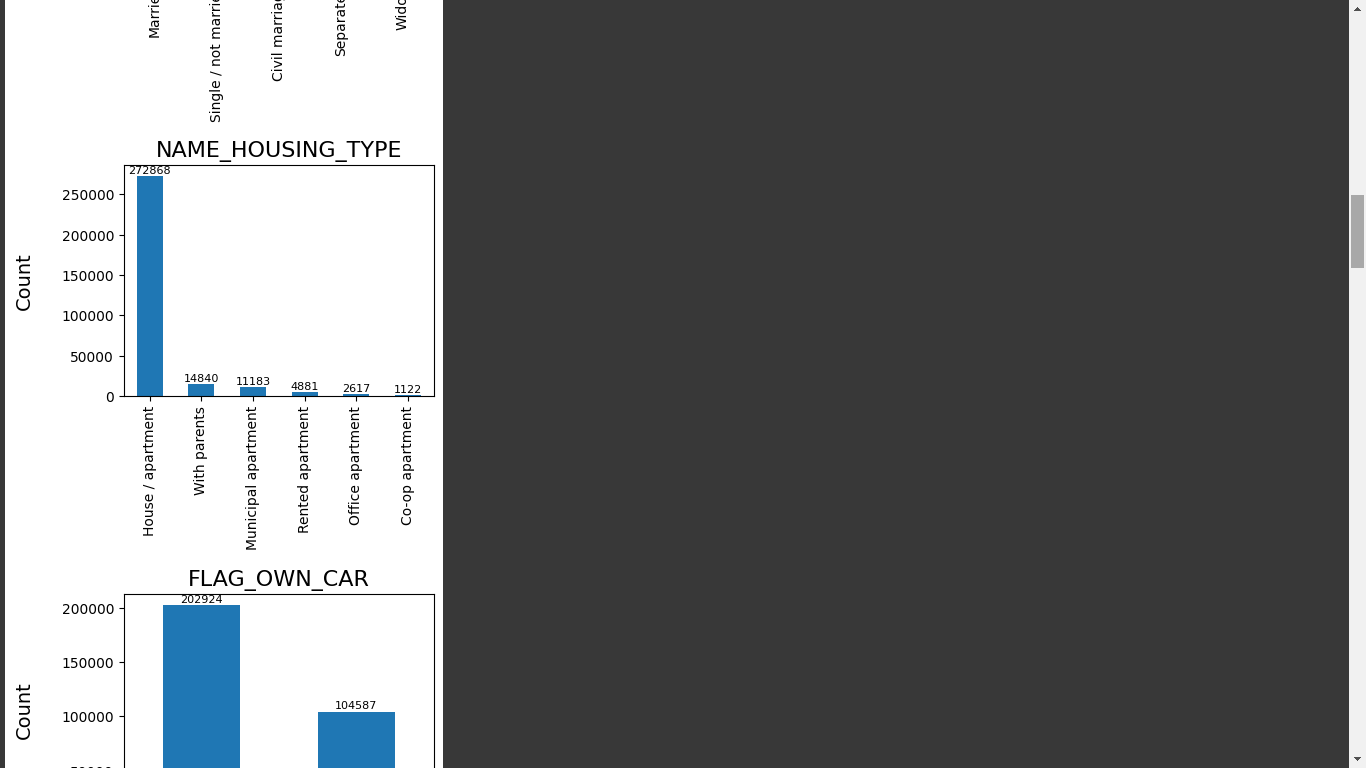
### Categorical Columns

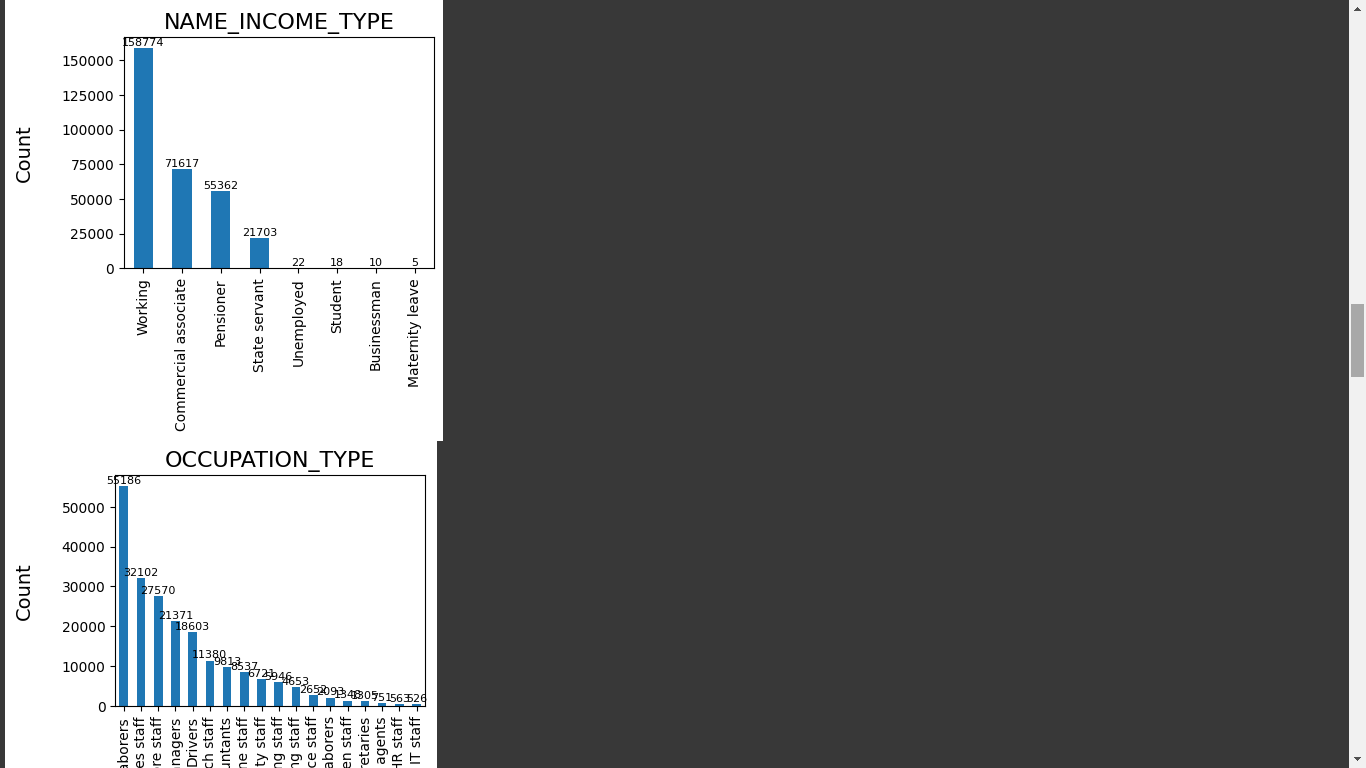
1. About 90.5% of loans were Cash loans while only 9.5% were Revolving loans.
2. The number of female applicants (65%) was almost double of male applicants (35%).
3. Most of the applicants had Secondary or Secondary Special education (71%), followed by Higher education (24%). The least number of applicants were from people with an academic degree.
4. Most of the applicants lived unaccompanied (80%). About 13% lived with their families.
5. 63% of applicants were married, 14% were single, 9.6% had civil marriage, 6.4% were separated and about 5% were widows.
6. 88% of the applicants lived in a house or apartment.
7. About 70% of applicants owned realty while 30% didn’t.
8. About 34% of applicants owned cars while 64% didn’t.
9. Most of the applicants were Working or Commercial associates. Businessmen and people on maternity leave had the least number of applications.
10. The maximum number of applicants was of laborers (17%), followed by sales Staff (10%).
11. People from Business Entity type 3 (22%) applied the most for loan, followed by self-employed people (12%).
12. For most of applicants, registration region was neither work nor live region.
13. Most of the applicants had provided their mobile phone numbers, work phone numbers and email-ids. Moreover, for most of the applicants the number was found to be reachable.
14. Among required documents, only Document 3 was provided by 70% of the applicants, while other documents were not provided by most.

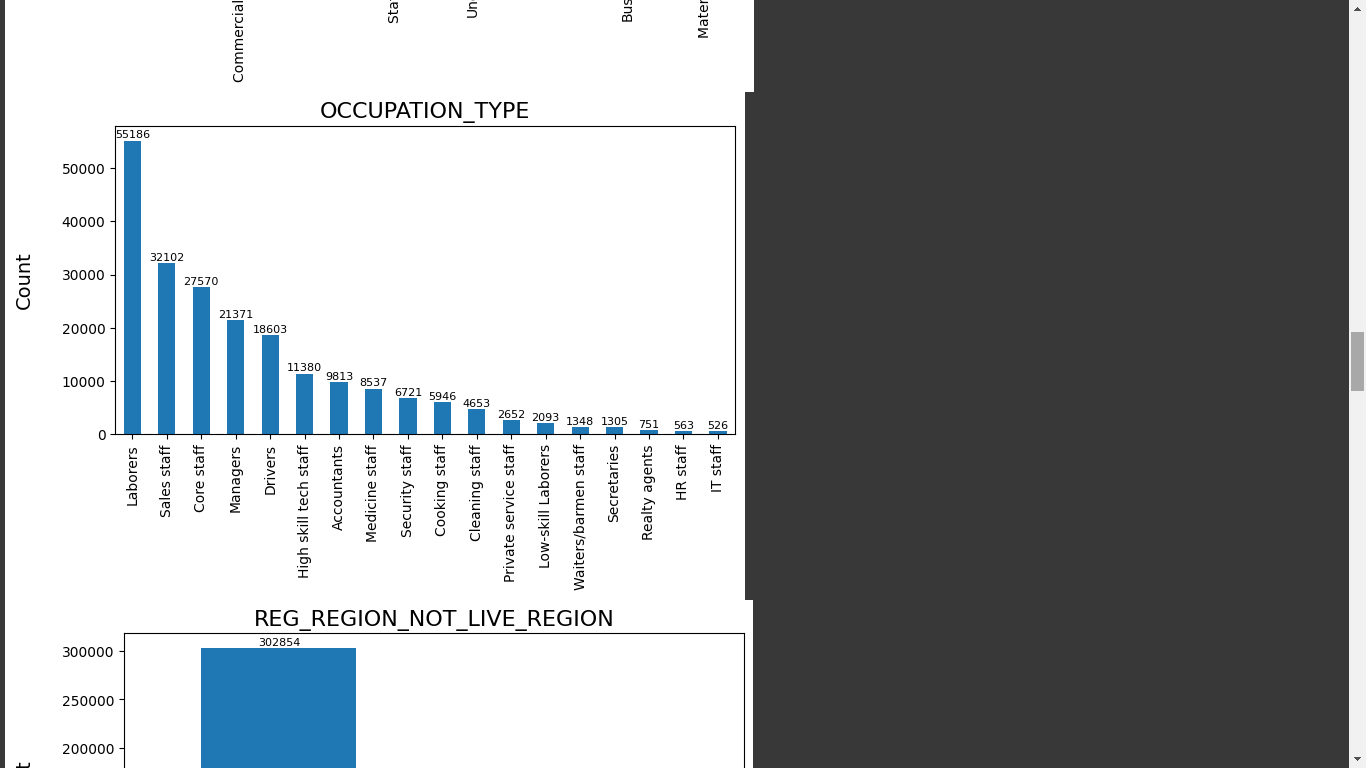






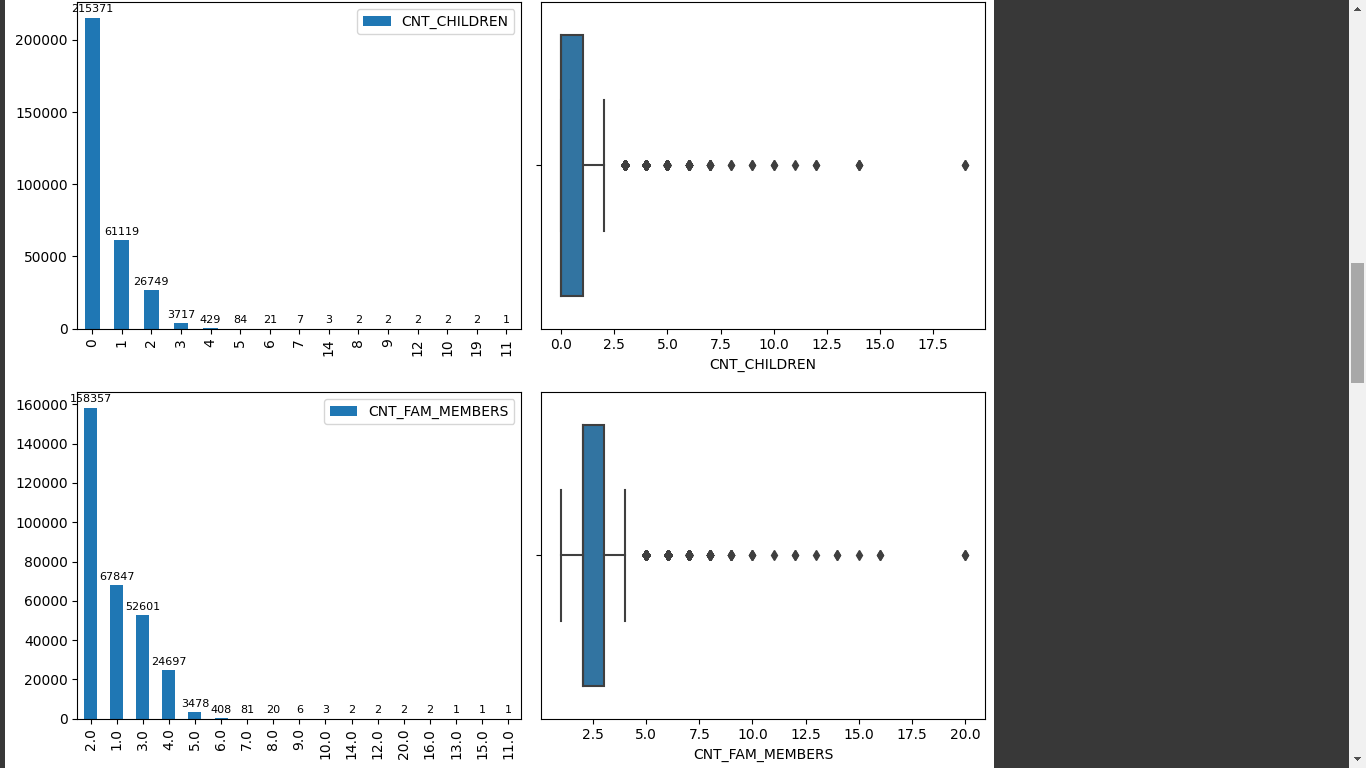


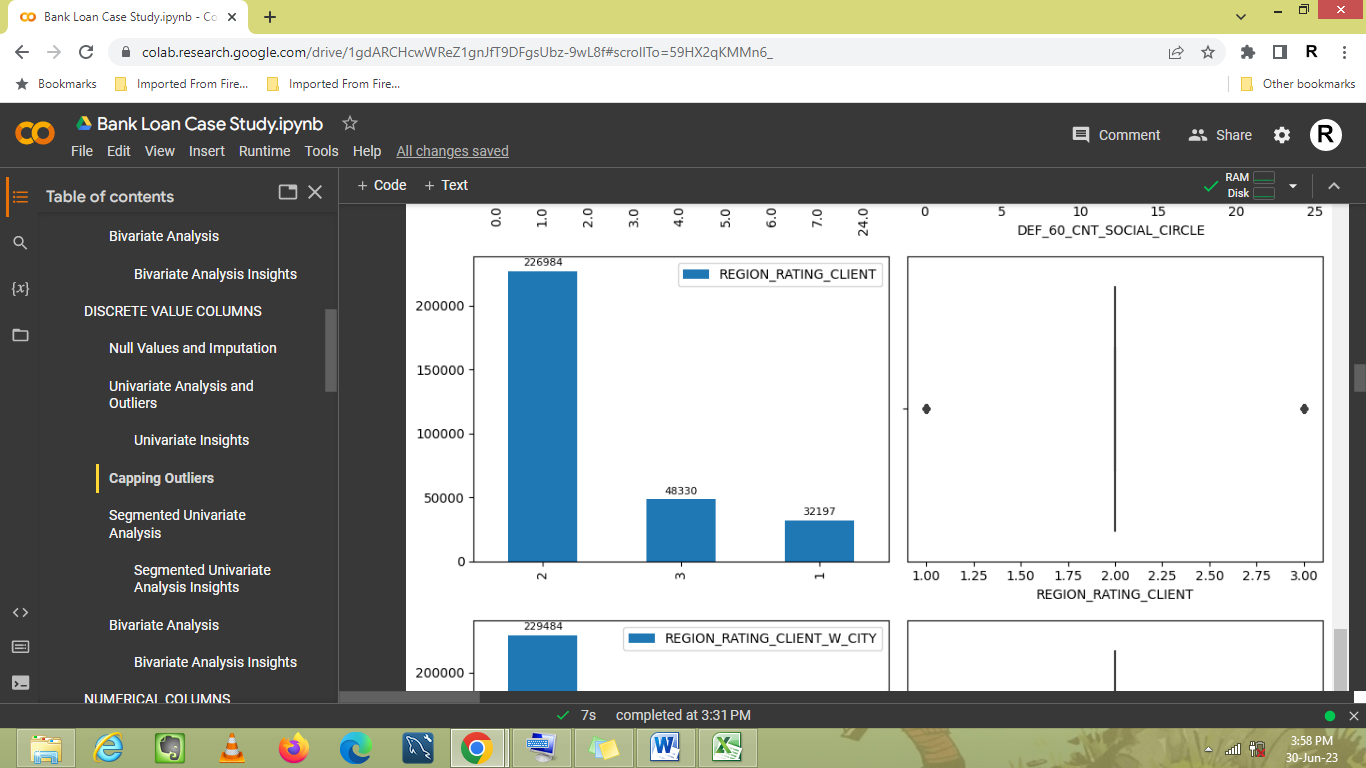




### Discrete Columns

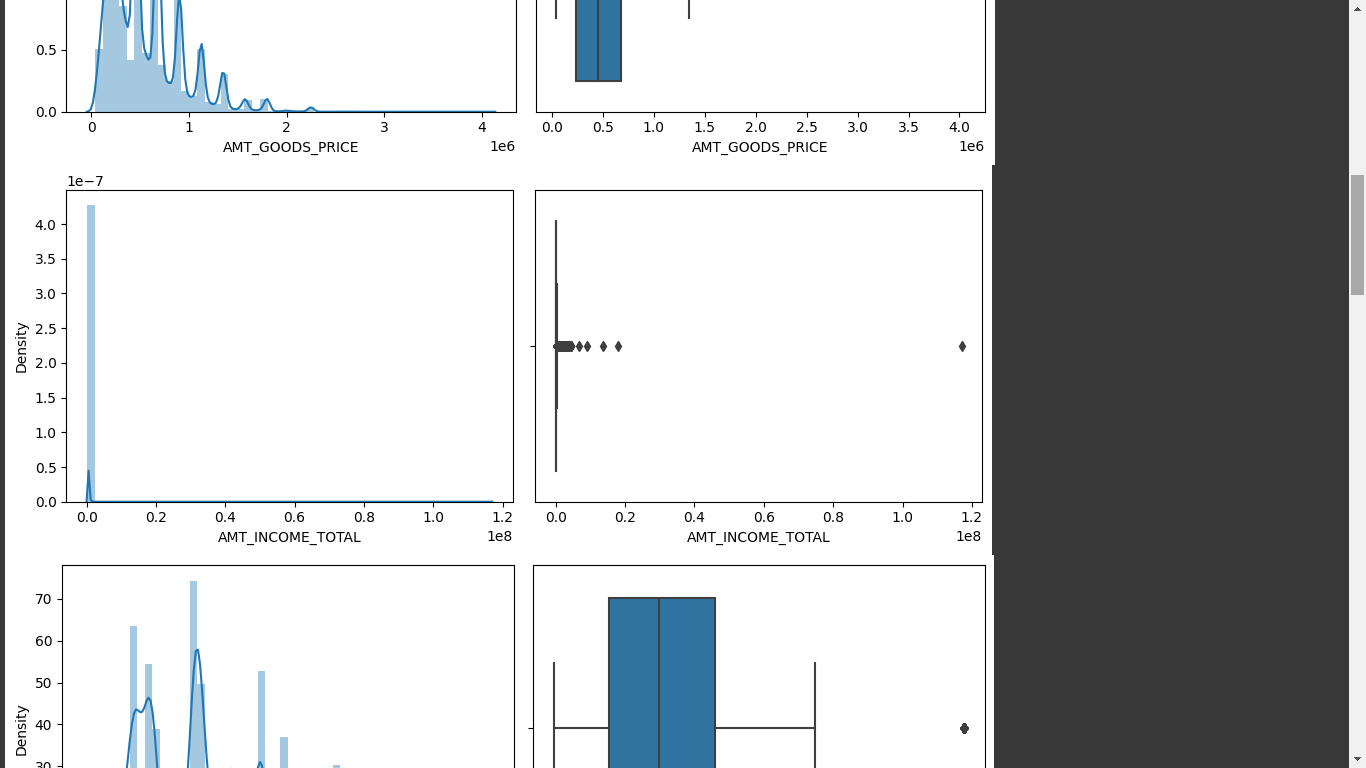
1. 70% of applicants had no children, 19% had 1 child and 8% had 2 children.
2. 51% of applicants had only two family members, 22% had one and 17% had three family members.
3. Most of the applicants were from Region Rating 2.



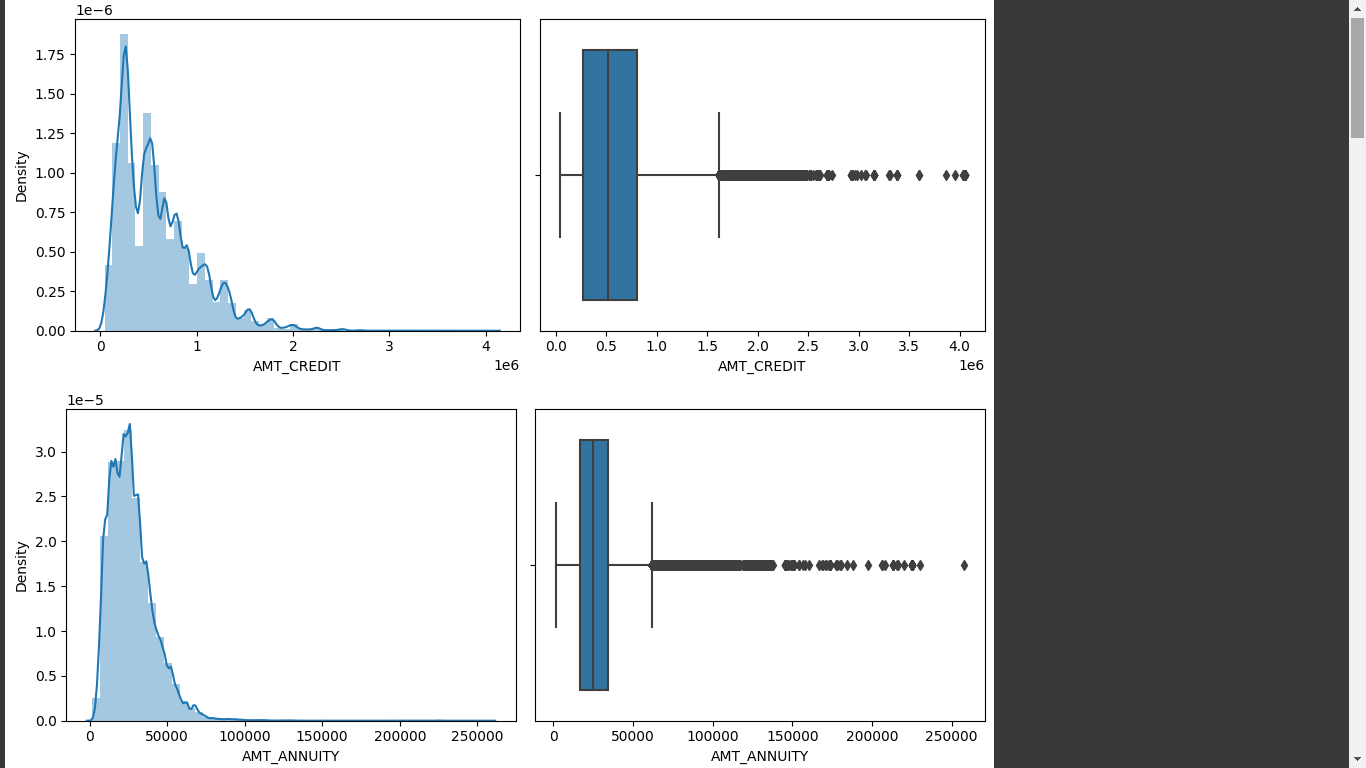
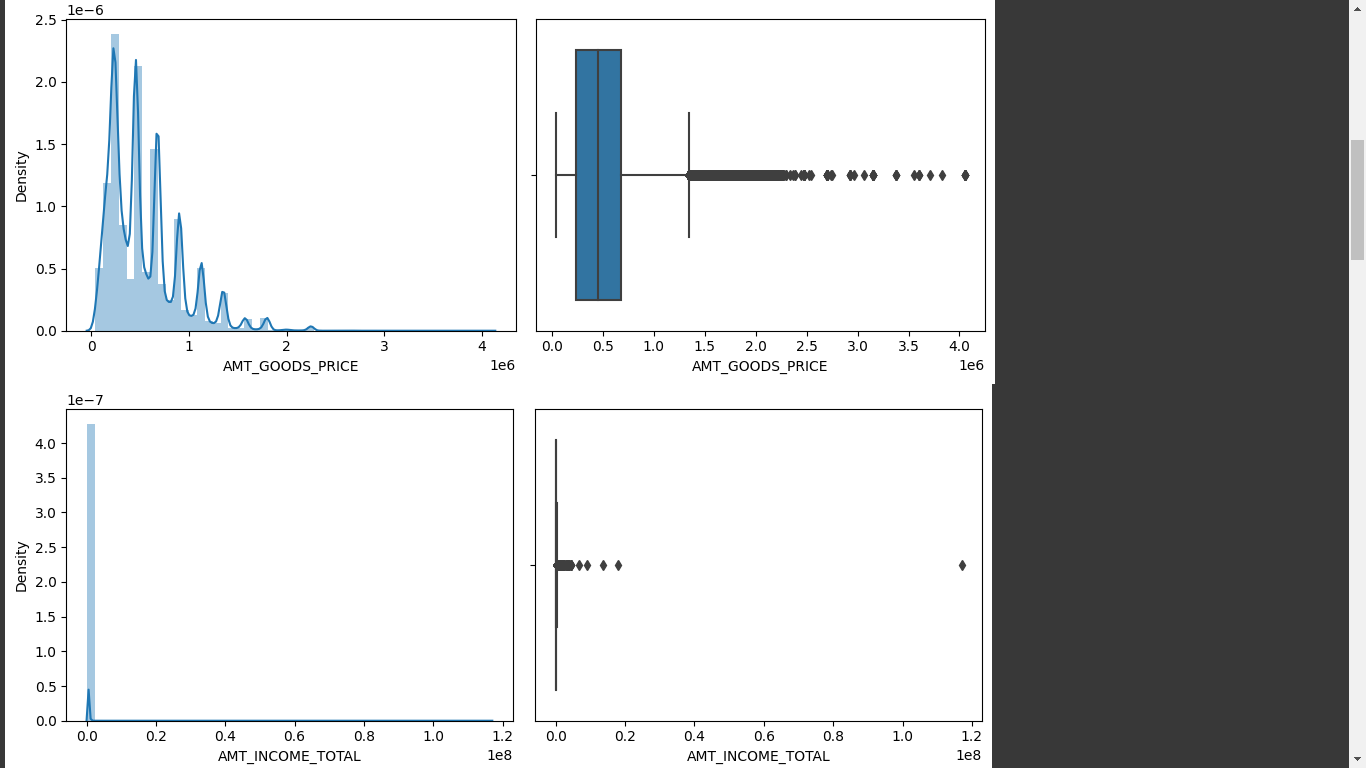


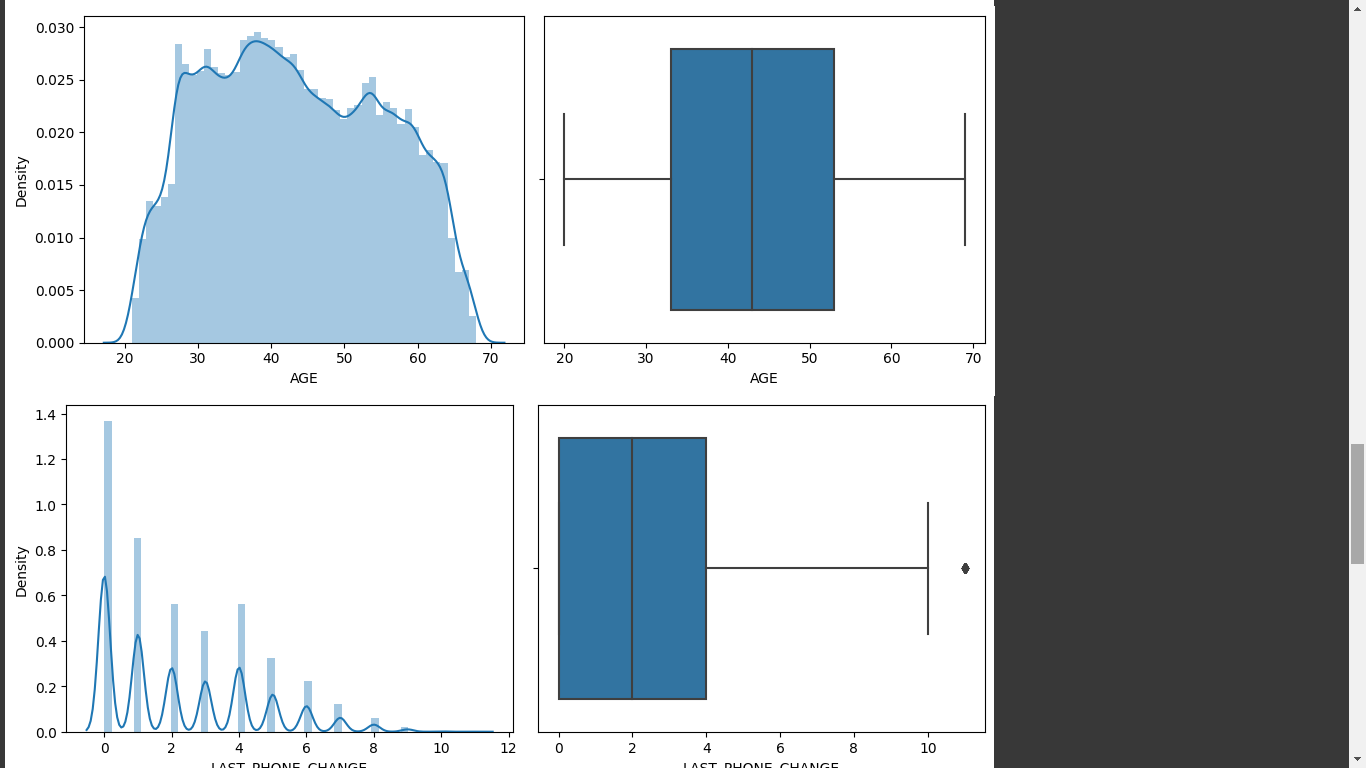
### Numerical Columns

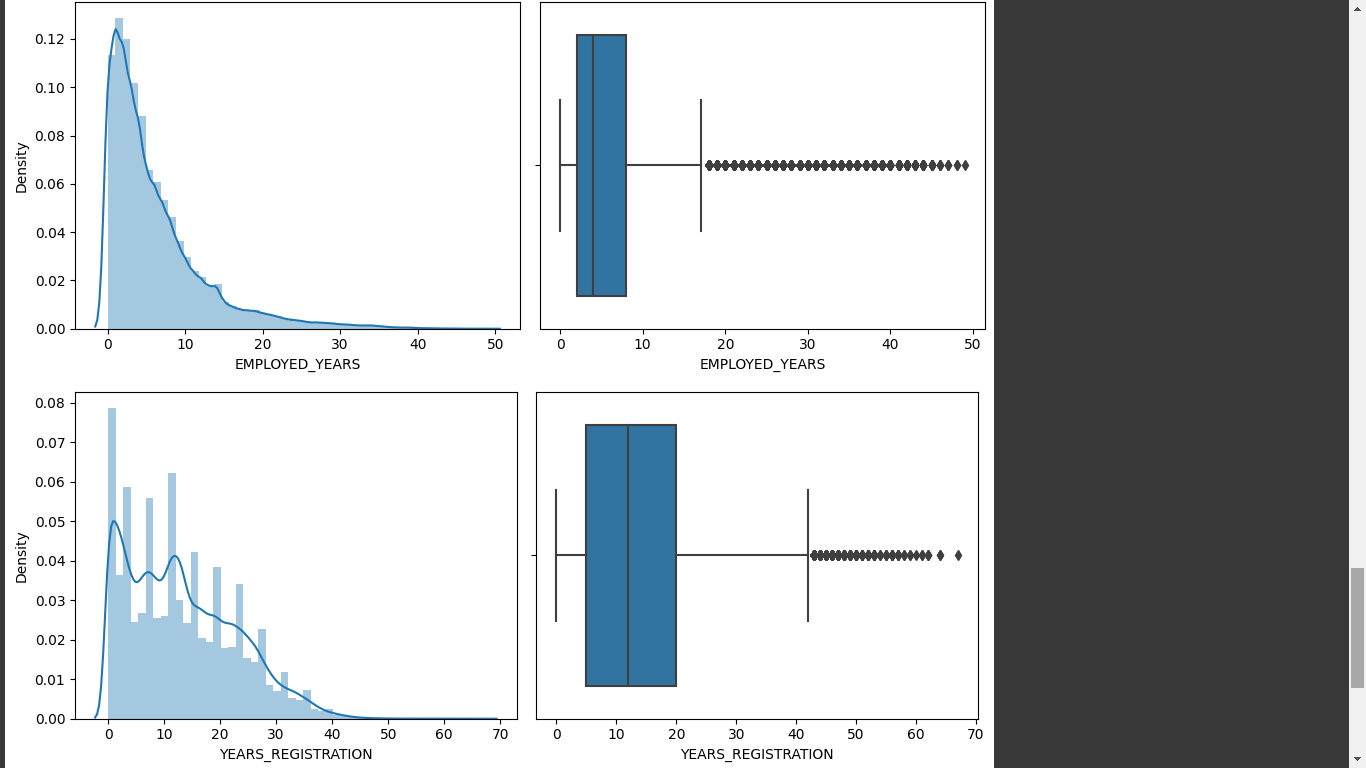
1. 75% of applicants have income up to 2 Lac. The income range with highest number of applicants was between 1Lac and 2 Lac.
2. The minimum income of an applicant is 25,000 while maximum is 11.7 crore. However, 99% of applicants have income below 5 Lac.





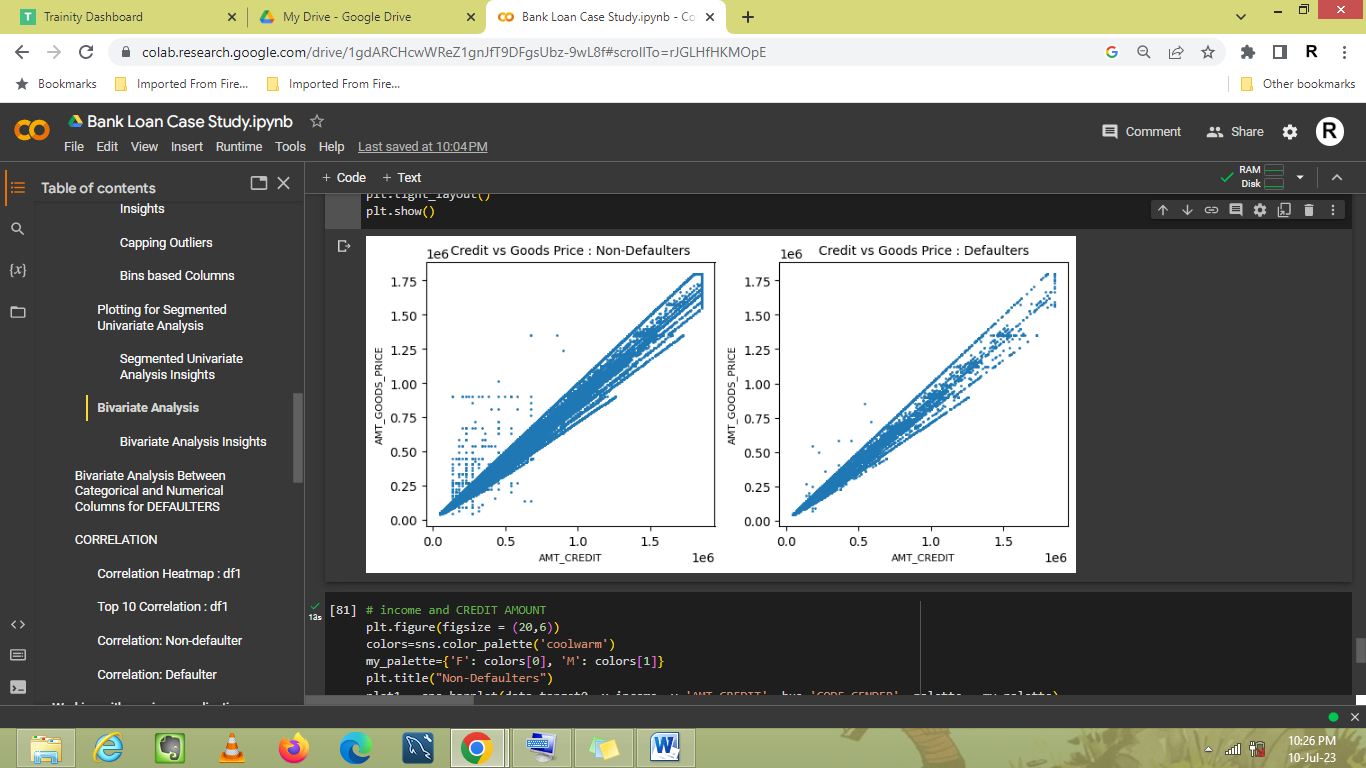
1. The goods price range with maximum number of applicants was from 2 to 4 Lac. 75% of the applicants filed for loan against a goods’ price value under 6.7 Lac. The minimum Goods price was about 40K and maximum was 40 Lac.
2. The range of credit approved for maximum number of applicants was from 2 to 4 Lac. 75% of the applications had Credit approved till the amount of 8 Lac. The maximum amount approved was of 40 Lac.
3. 75% of the applicants paid an Annuity amount below 35K. The maximum Annuity amount was of 2.6 Lac.
4. Most of the applicants were in Age range 33 to 53.
5. Most of the applicants were employed for 2 to 8 years.
6. Most of the applicants had changed their registration in last 5 to 20 years.



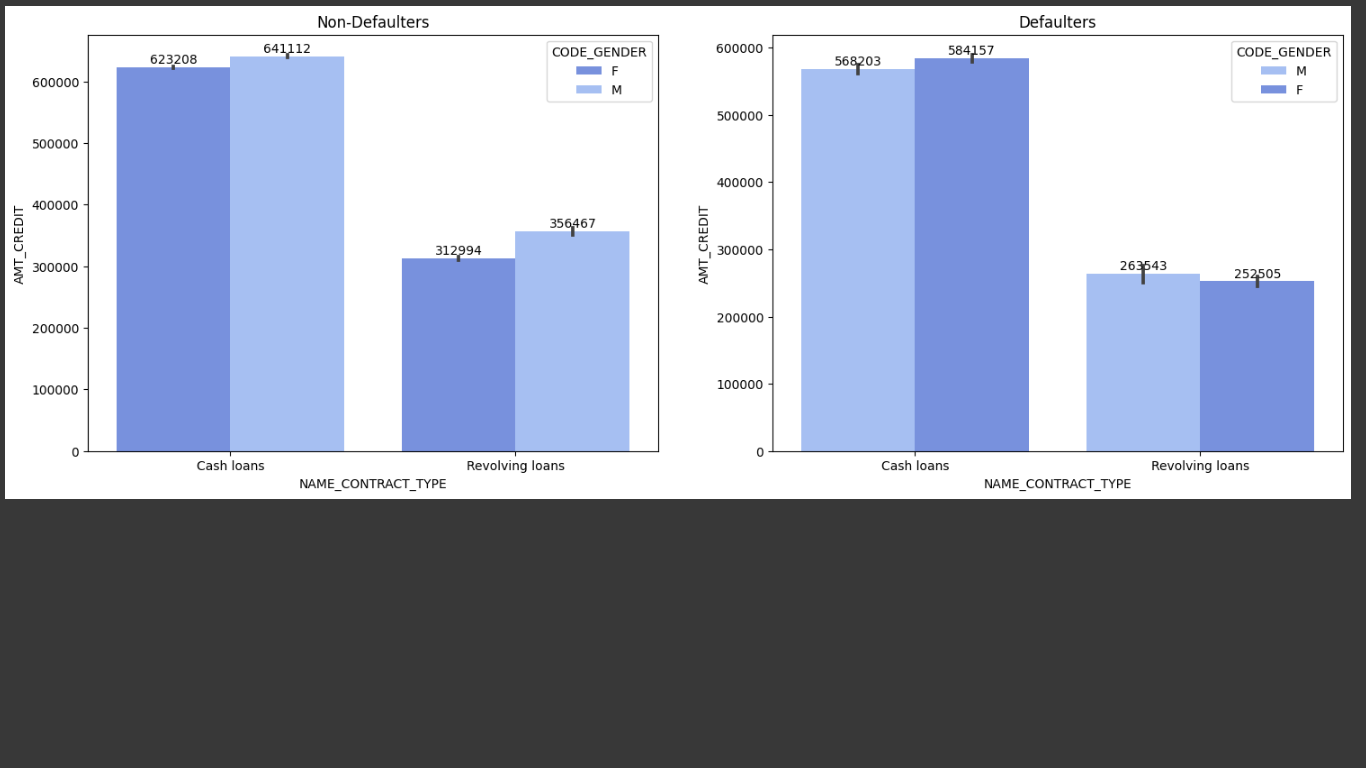


## Bivariate Analysis

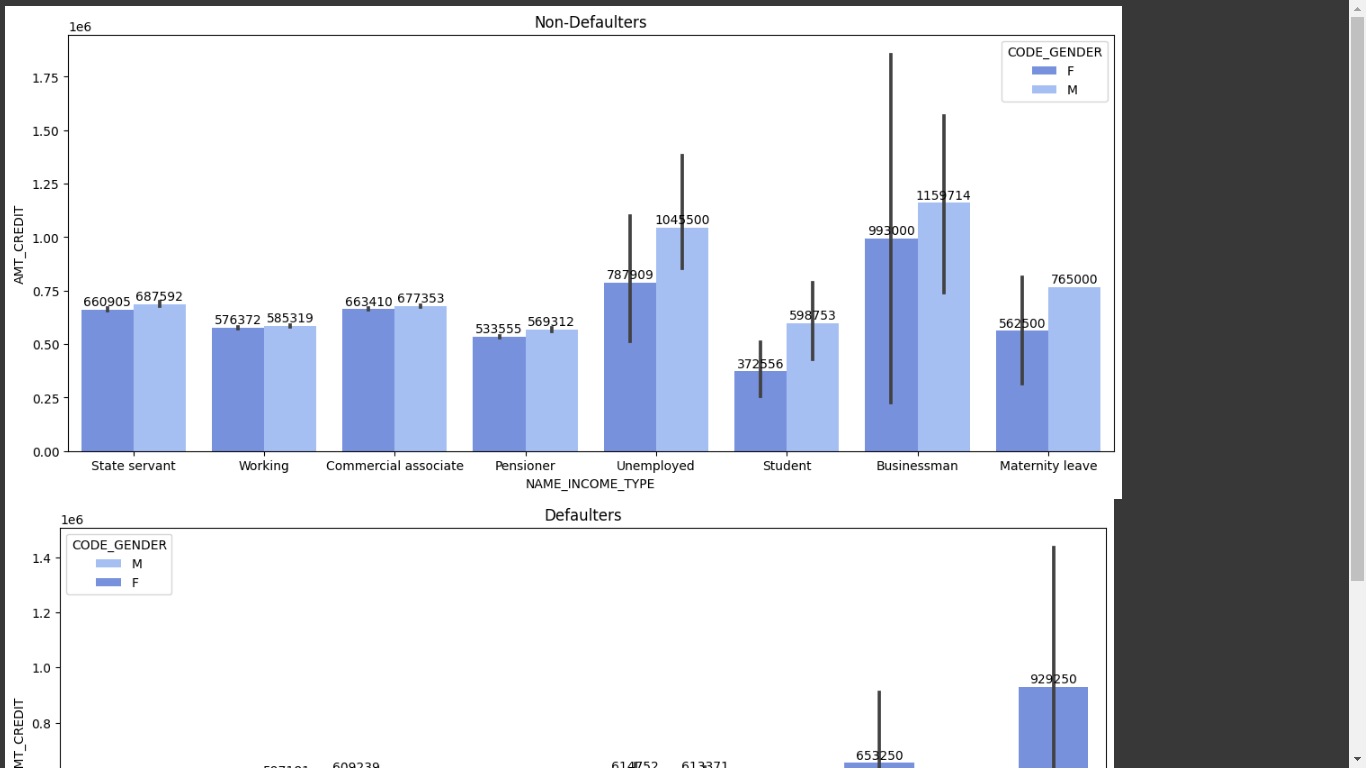
1. Higher correlation between features OBS\_60\_CNT\_SOCIAL\_CIRCLES and OBS\_30\_CNT\_SOCIAL\_CIRCLES was observed.
2. Similarly, higher correlation between features DEF\_60\_CNT\_SOCIAL\_CIRCLES and DEF\_30\_CNT\_SOCIAL\_CIRCLES was observed.
3. There is high correlation between goods’ price and credit amount for both defaulters and non-defaulters.
4. It was observed that with an increase in income, there was an increase in credit amount.

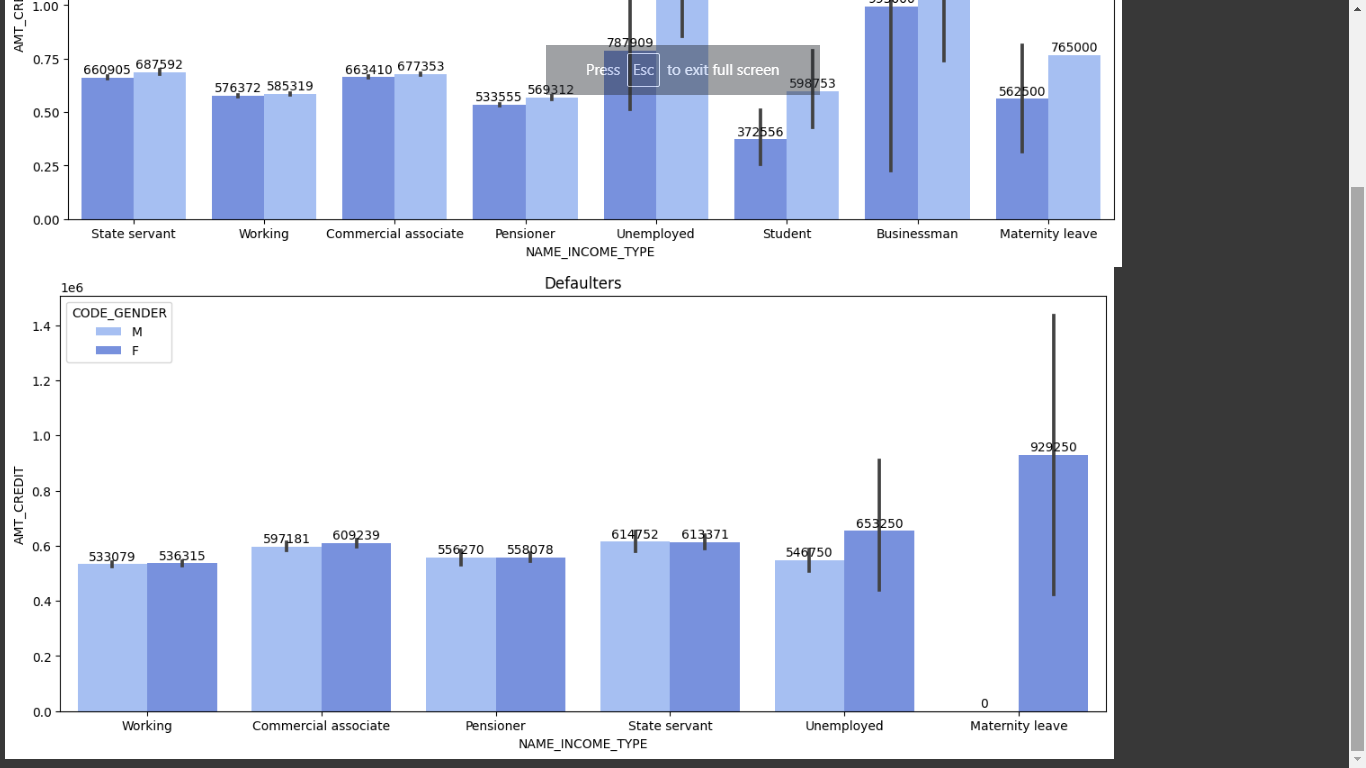


1. There was no correlation between EXT\_SOURCE\_2 and EXT\_SOURCE\_3.
2. Credit amount was higher for Cash loans. Moreover, for non-defaulters, the number of male applicants was higher for both cash as well as revolving loans.



1. The credit amount was highest for the male applicants with an academic degree, followed by male applicants with higher education.
2. Amongst non-defaulters, male businessmen and male unemployed had the highest credit amount. For defaulters, female on maternity leave or unemployed had the highest credit amount.



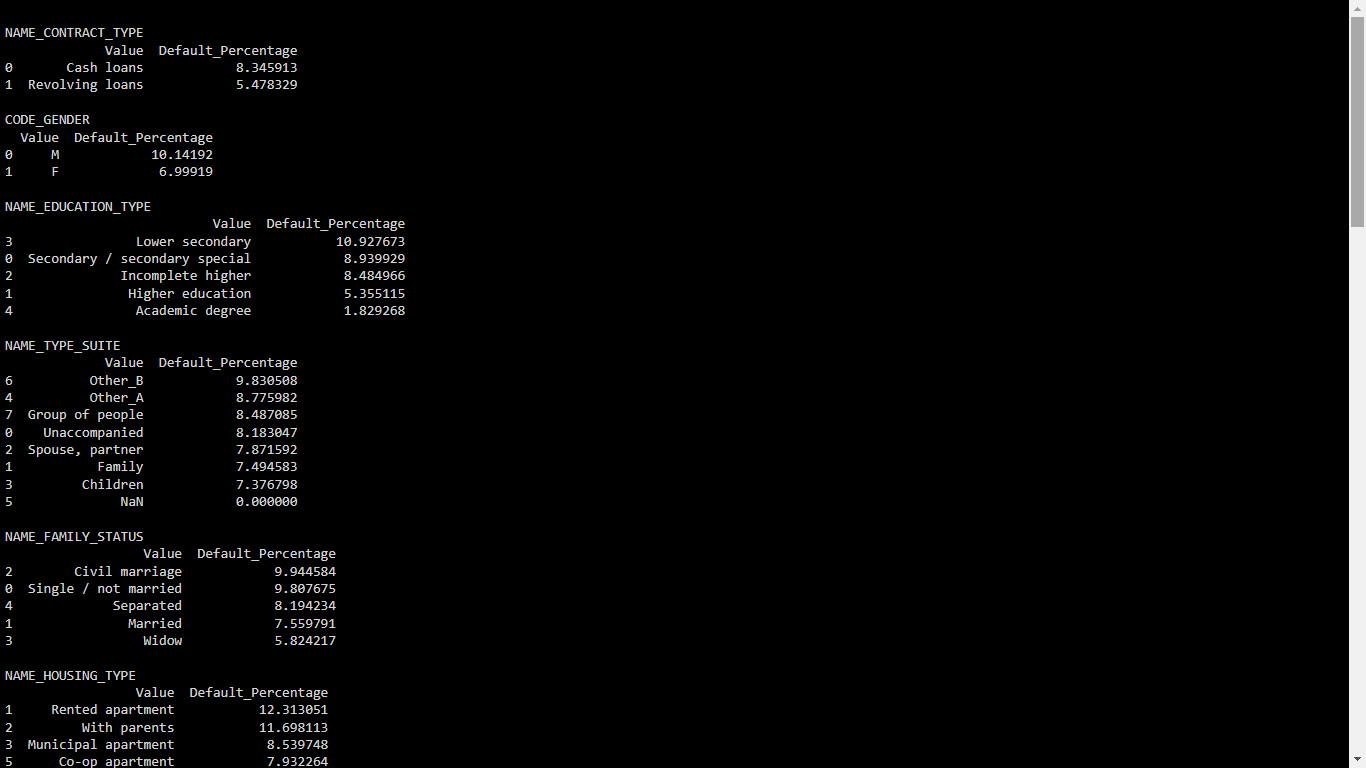


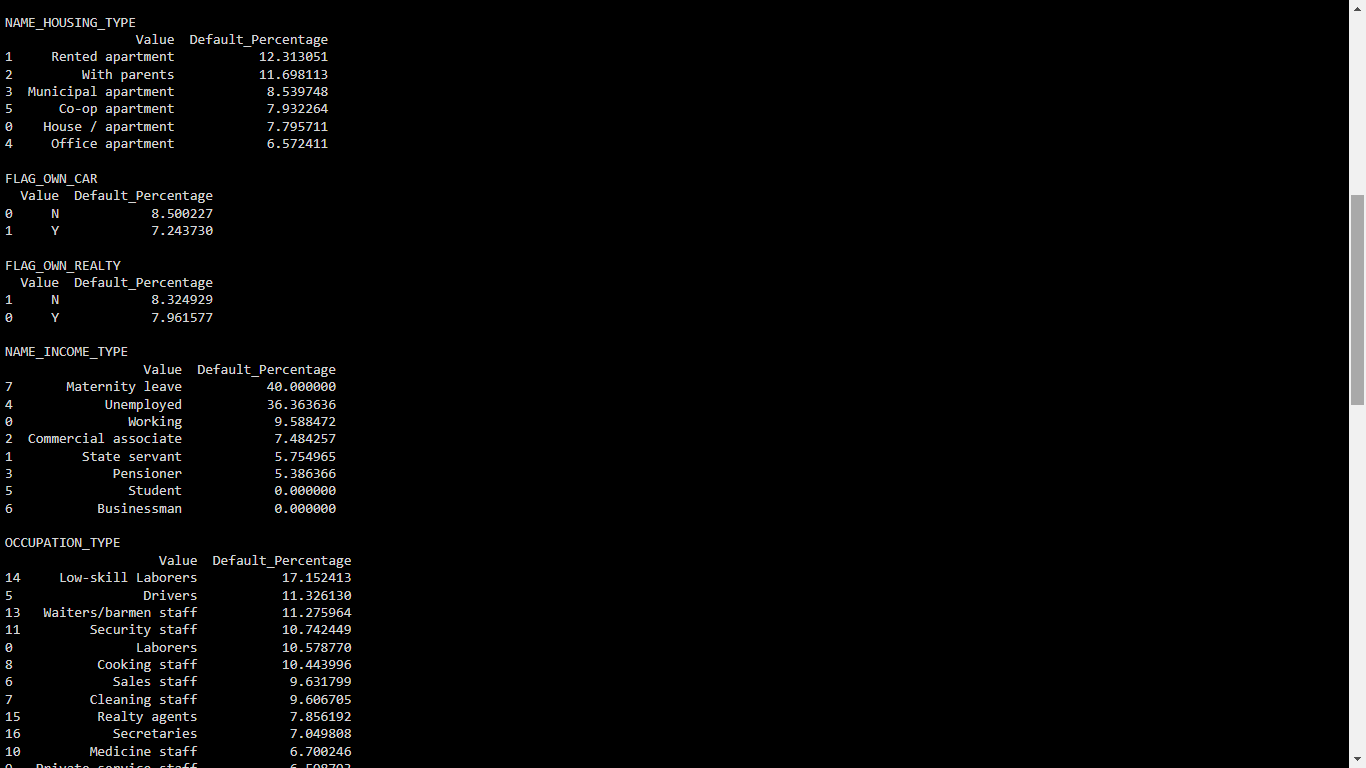
1. For non-defaulters, males earned more irrespective of profession, with an exception of business and student income. For defaulters, the male applicants earned more, with an exception of unemployed and maternity leave.

## Segmented Univariate Analysis

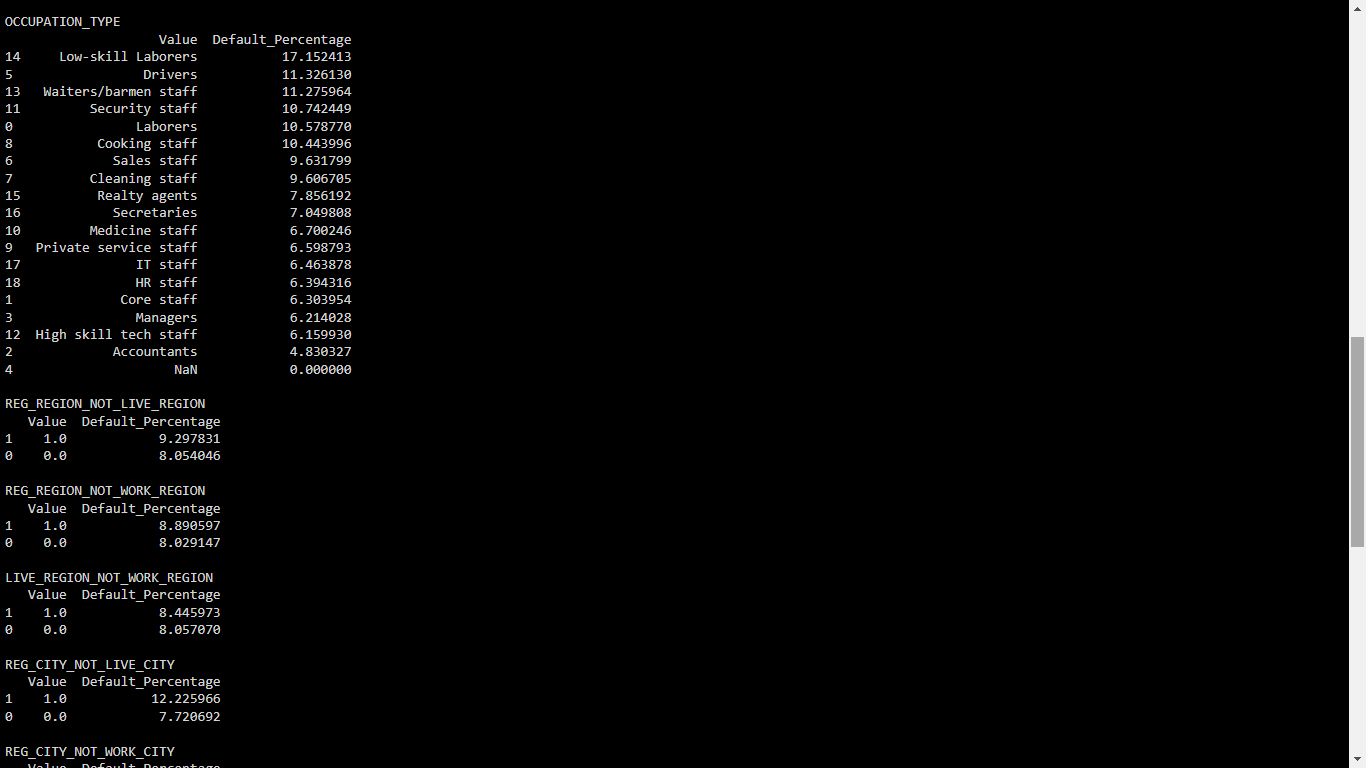
### Categorical Columns

The defaulter percentage for each value of each categorical column is shown in the clips below.





1. The percentage of defaulters was higher in Cash Loans as compared to revolving loans.
2. Males, while being less in number, defaulted more than women.
3. The applicants with lower secondary education, while less in count, defaulted more than other education types. People with academic degrees defaulted the least.
4. The accommodation type Other\_B had the highest percentage of defaulters while people accommodating with family members, especially children, had the smallest default percentage.
5. The applicants with Civil marriage had the most difficulty in repayment, while widows defaulted the least.
6. People living in rented apartments had the highest default percentage while those residing in office apartments had the least difficulty in loan payment.
7. There was negligible difference between people who owned realty/car and people who didn’t, with non-owners defaulting more.
8. The people on maternity leave or unemployed had highest default percentage while students and businessmen had no difficulty in payments.

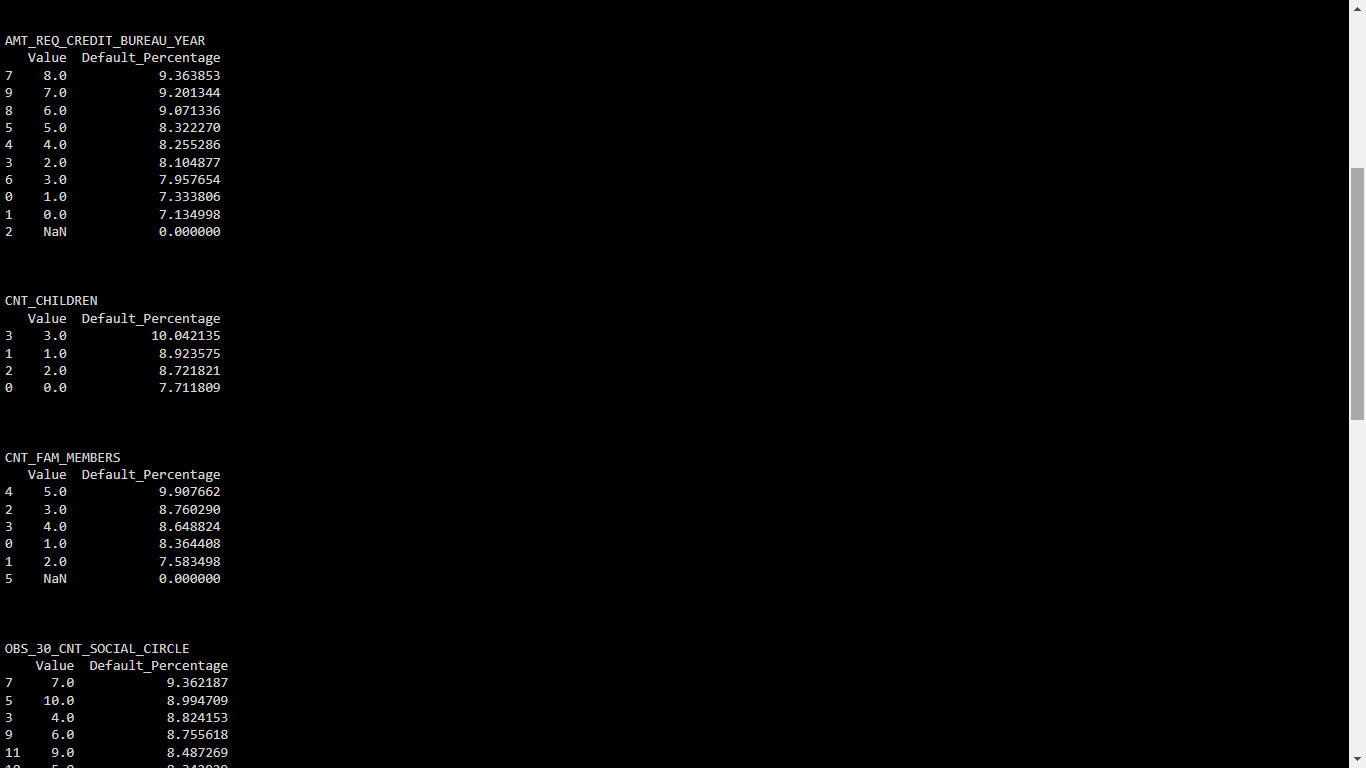


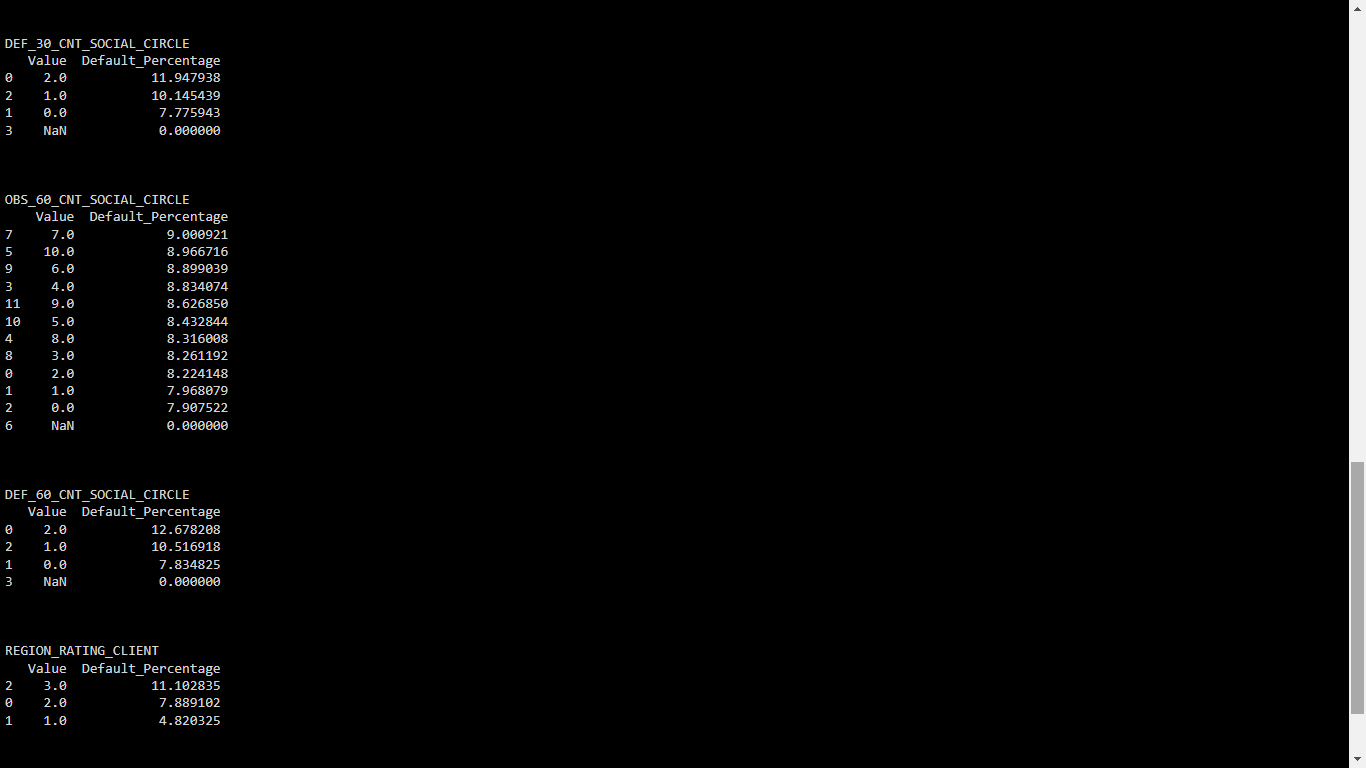


1. Low skill laborers, drivers, waiters had high default percentage while high skill tech staff and accountants had low default percentage.
2. Surprisingly, the people who had provided mobile numbers, work contact, emails, document 3, etc. defaulted more than the ones who didn’t.
3. The people whose contact/work address didn’t match permanent address defaulted more than the ones whose did.

### Discrete Columns

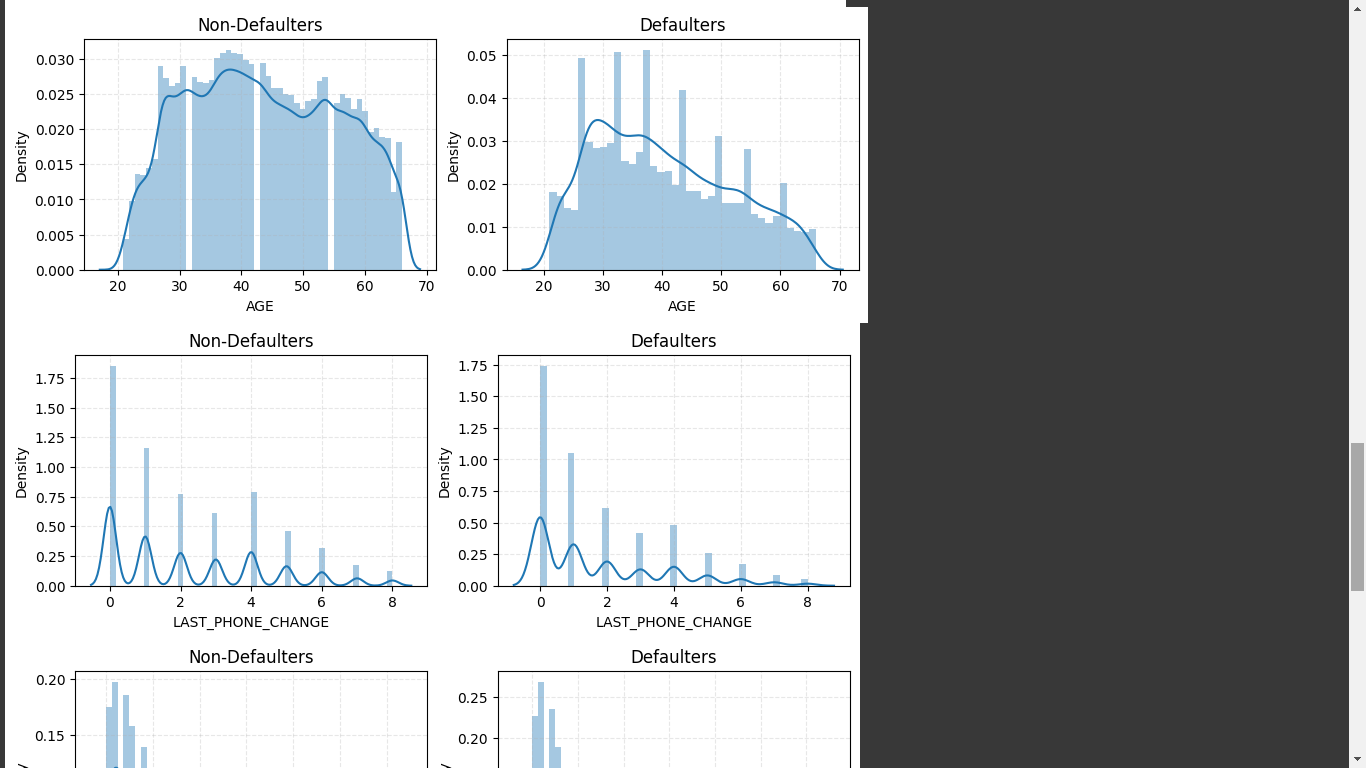
1. It was observed that defaulter percentage increased with an increase in the count of children/family members.
2. Also the Region rating 3 had highest default percentage, followed by Region rating 2. Region rating 1 had the least default percentage.
3. As the observations of client's social surroundings with defaults increased, the default percentage also increased.
4. The clients with higher number of enquiries to Credit Bureau in last one year (excluding last 3 months before application) had higher default percentage.

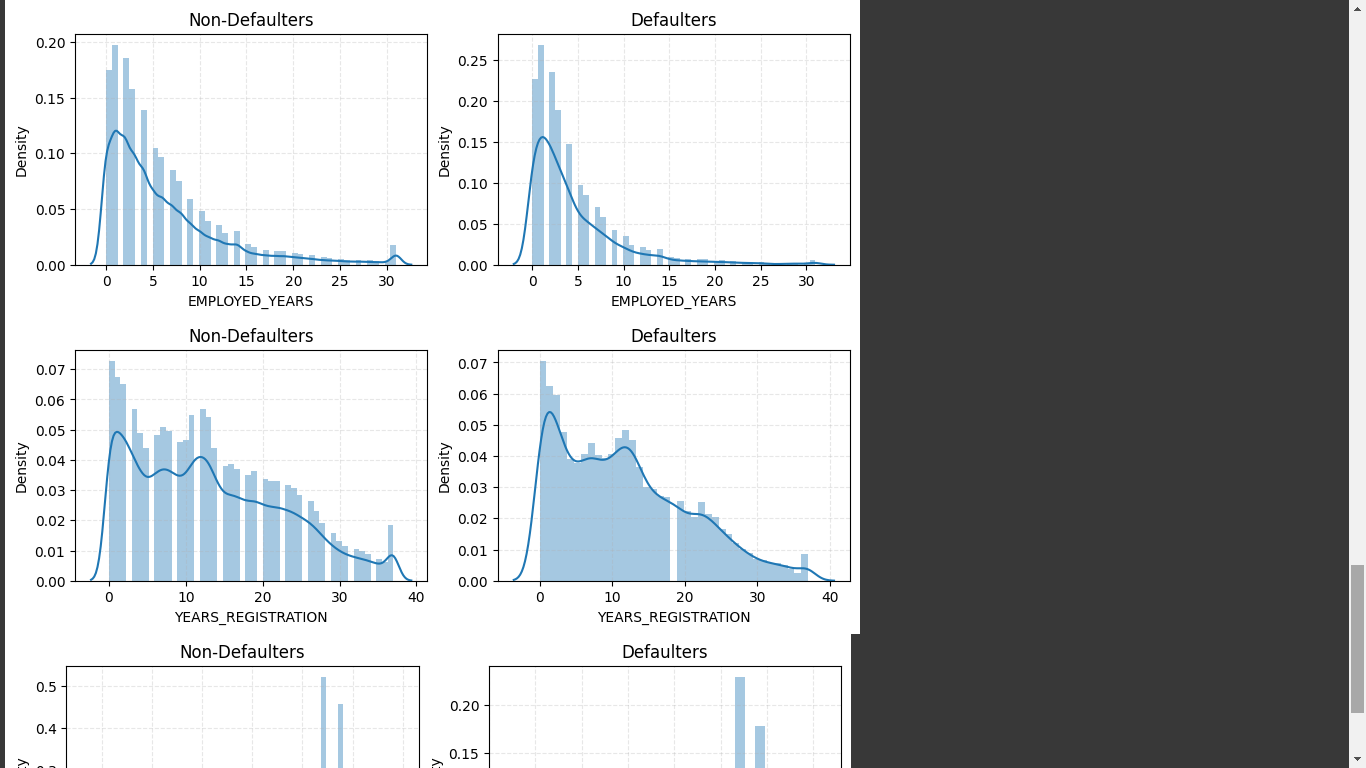




### Numerical Columns

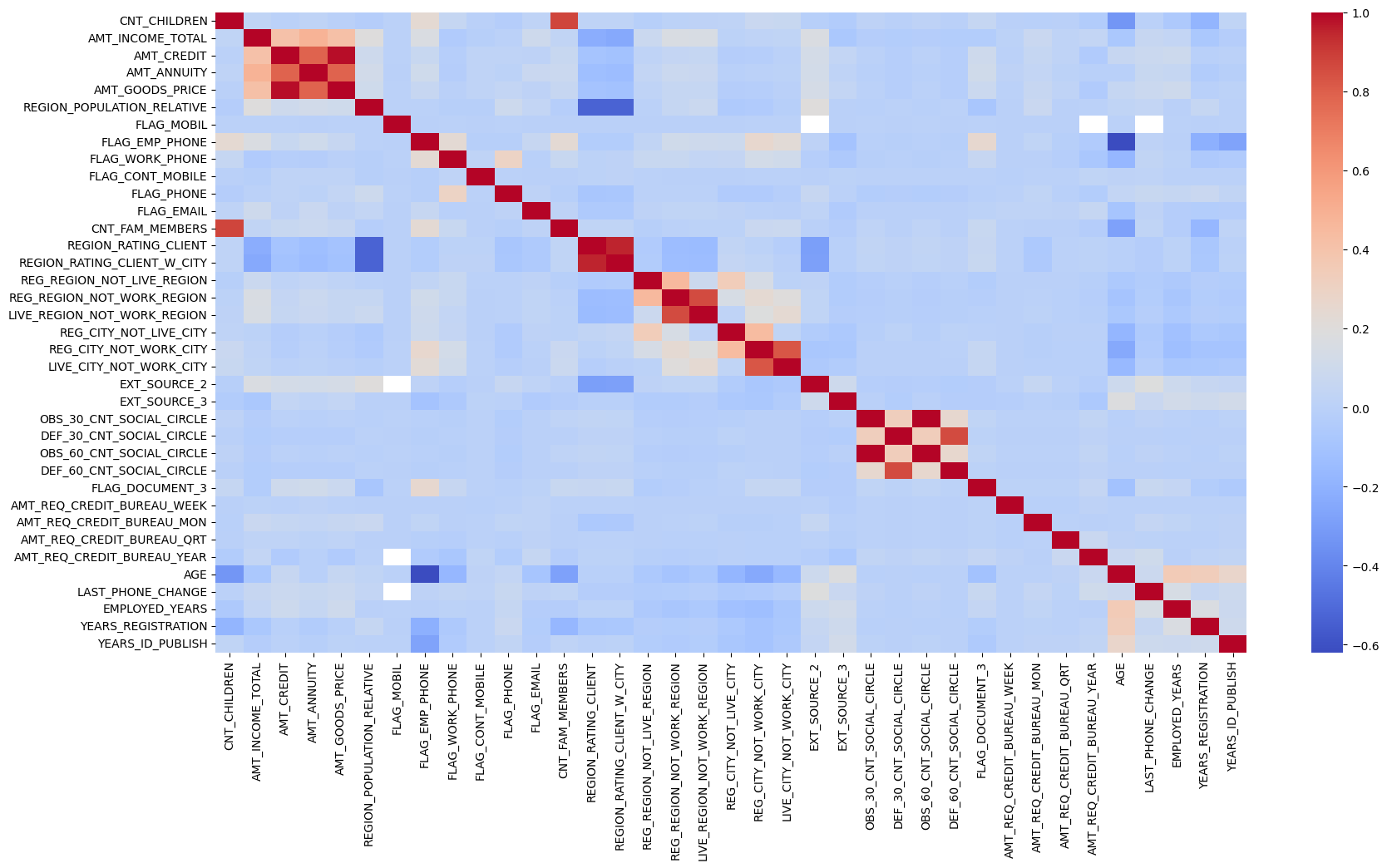
1. The people with age around 28-30 years defaulted the most.
2. The people who had changed their phone number less than a year before also defaulted the most.
3. The people who had been employed for less than 5 years defaulted more than others.
4. The people who had registered less than 5 years before defaulted the most.



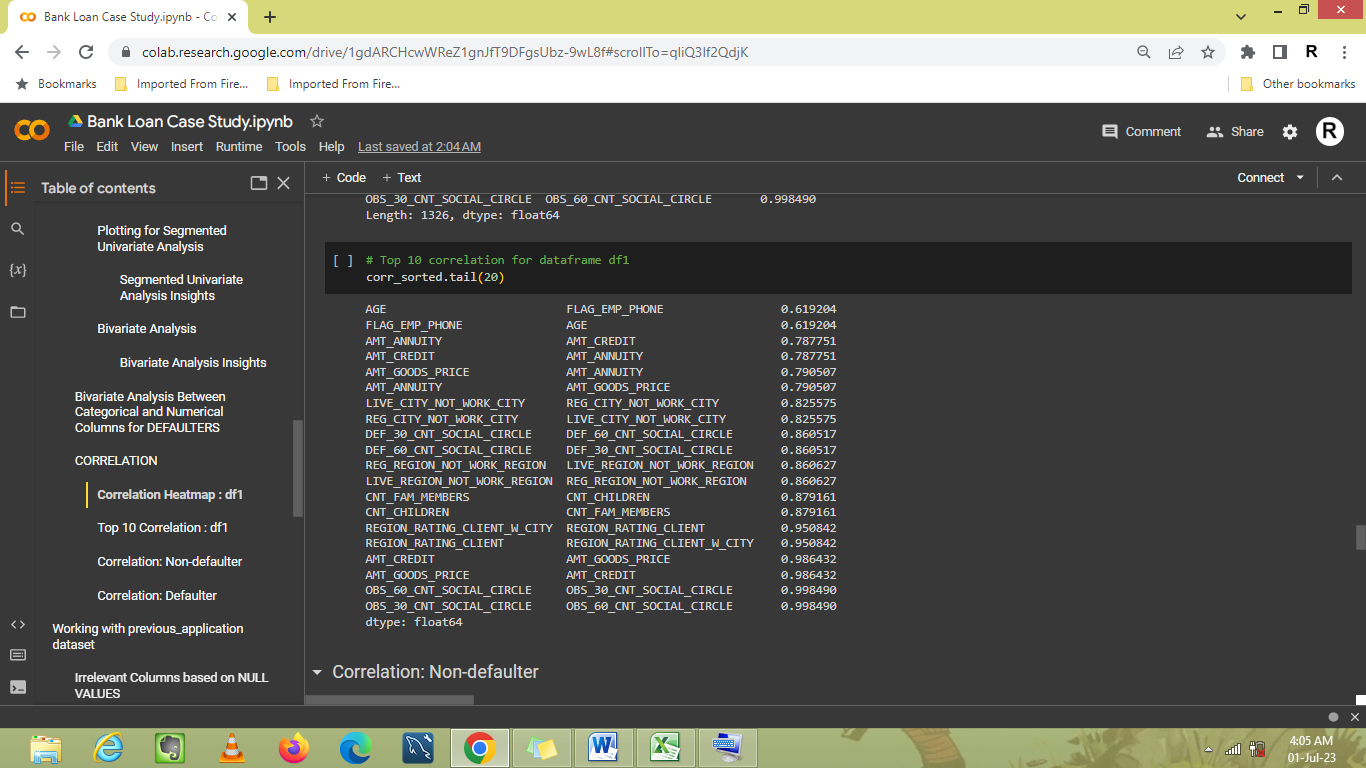


## Correlation

### application\_data

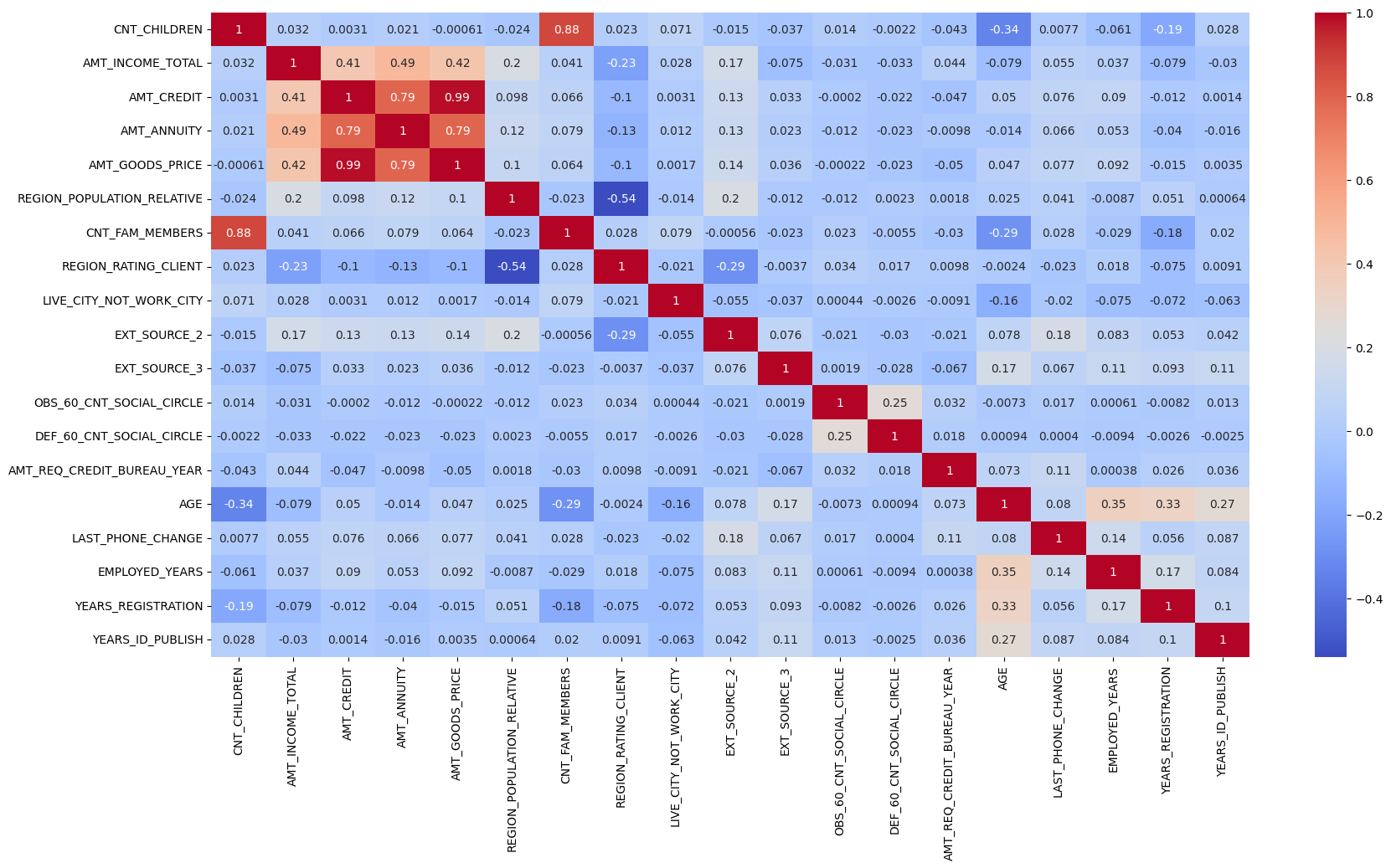


Top 10 correlation for application\_data



### Non-defaulter data

Heatmap for Target = 0

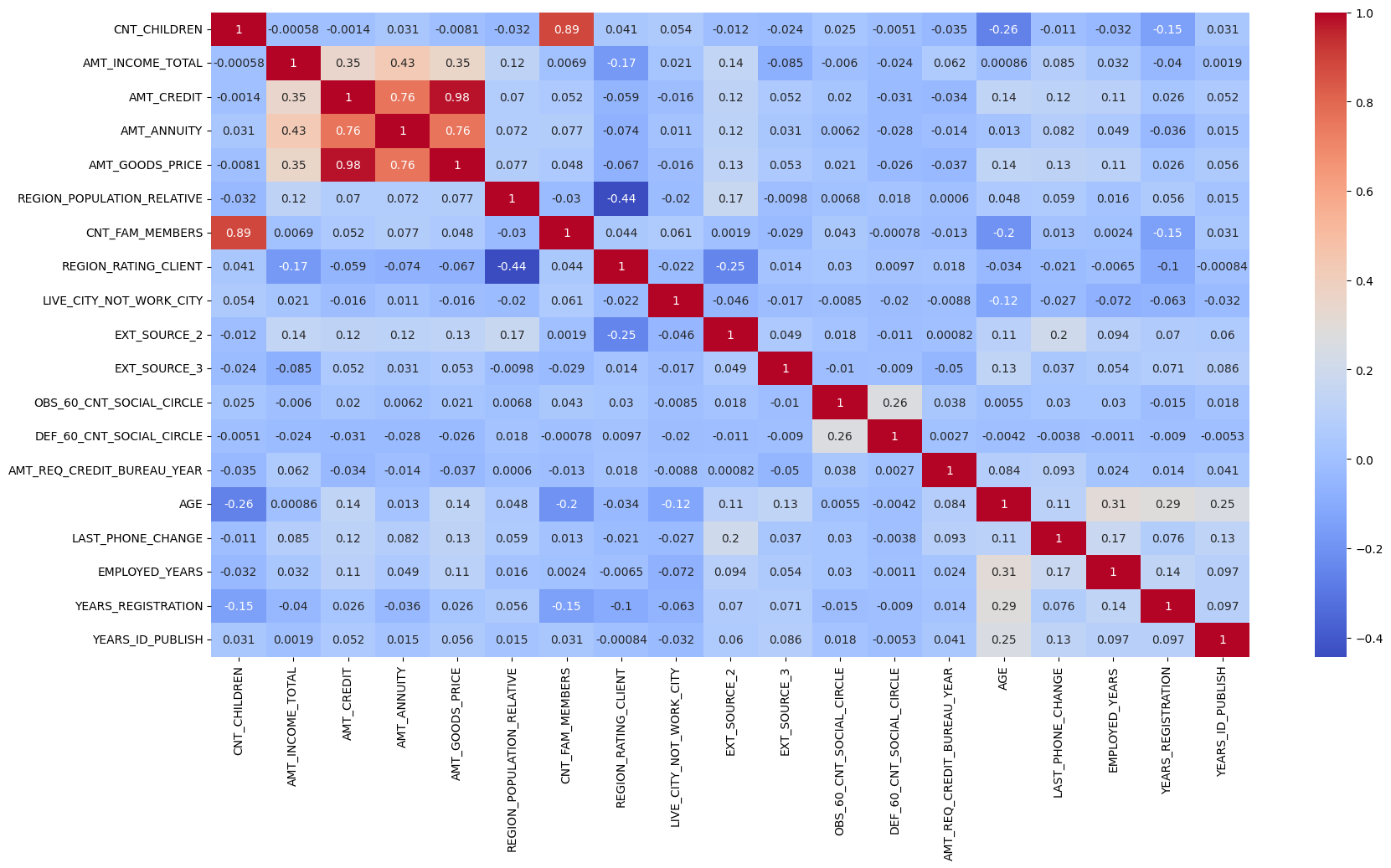


Top 10 Correlation for Non-defaulters

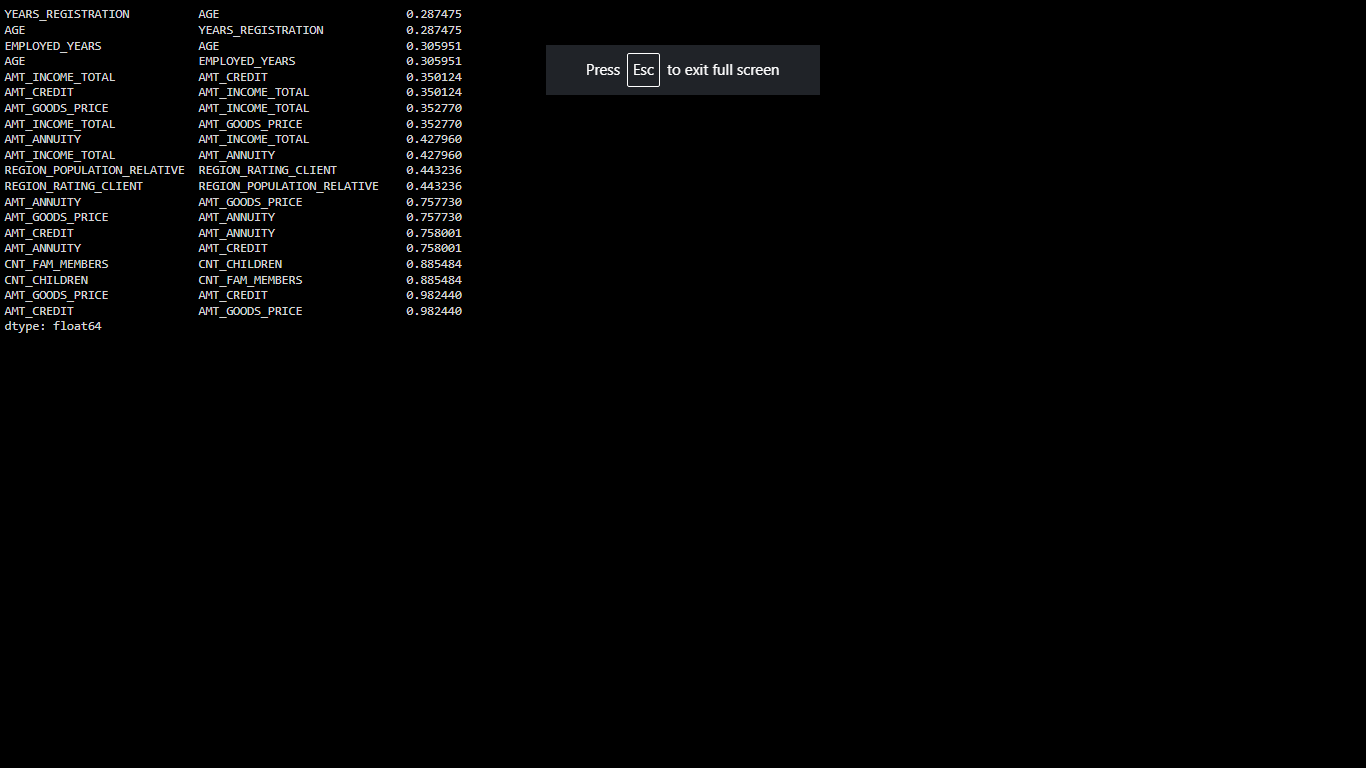


### Defaulter data

Heatmap



Top 10 correlations for Defaulters

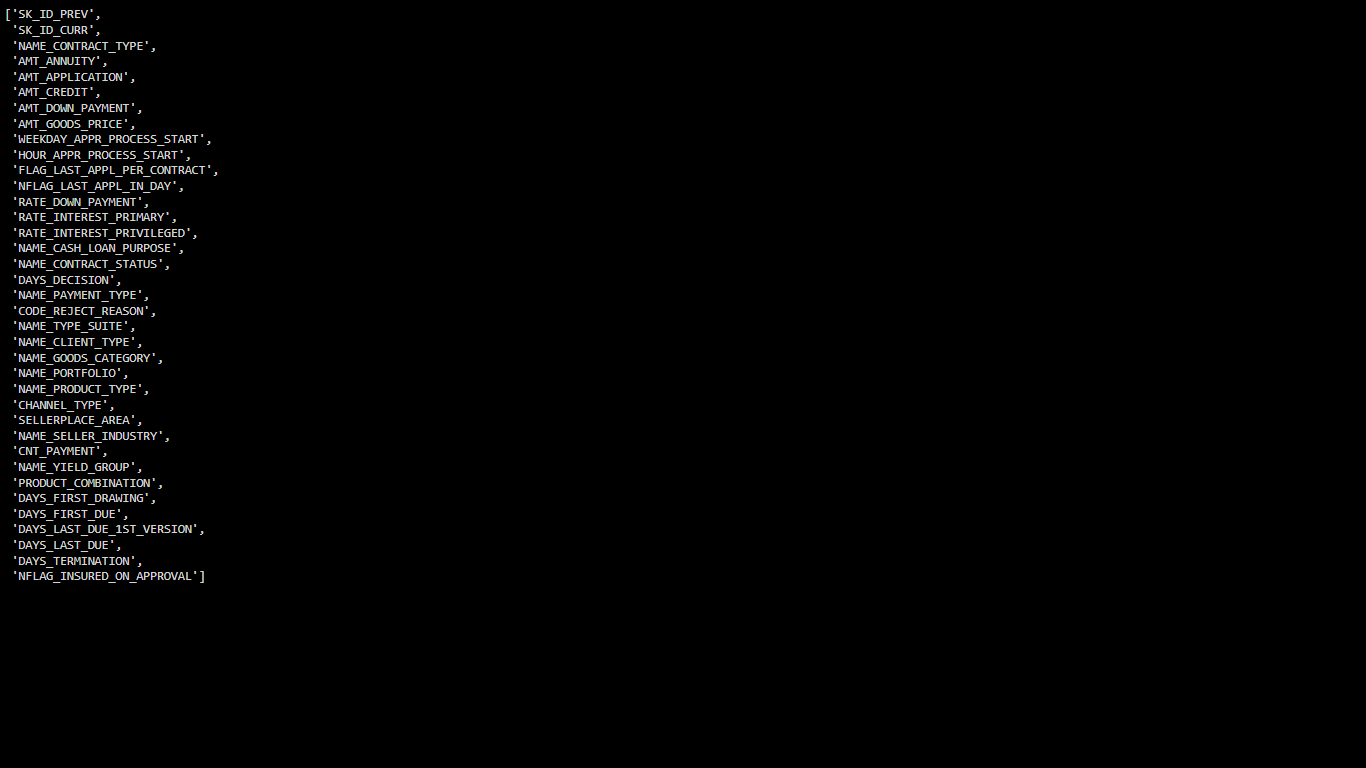


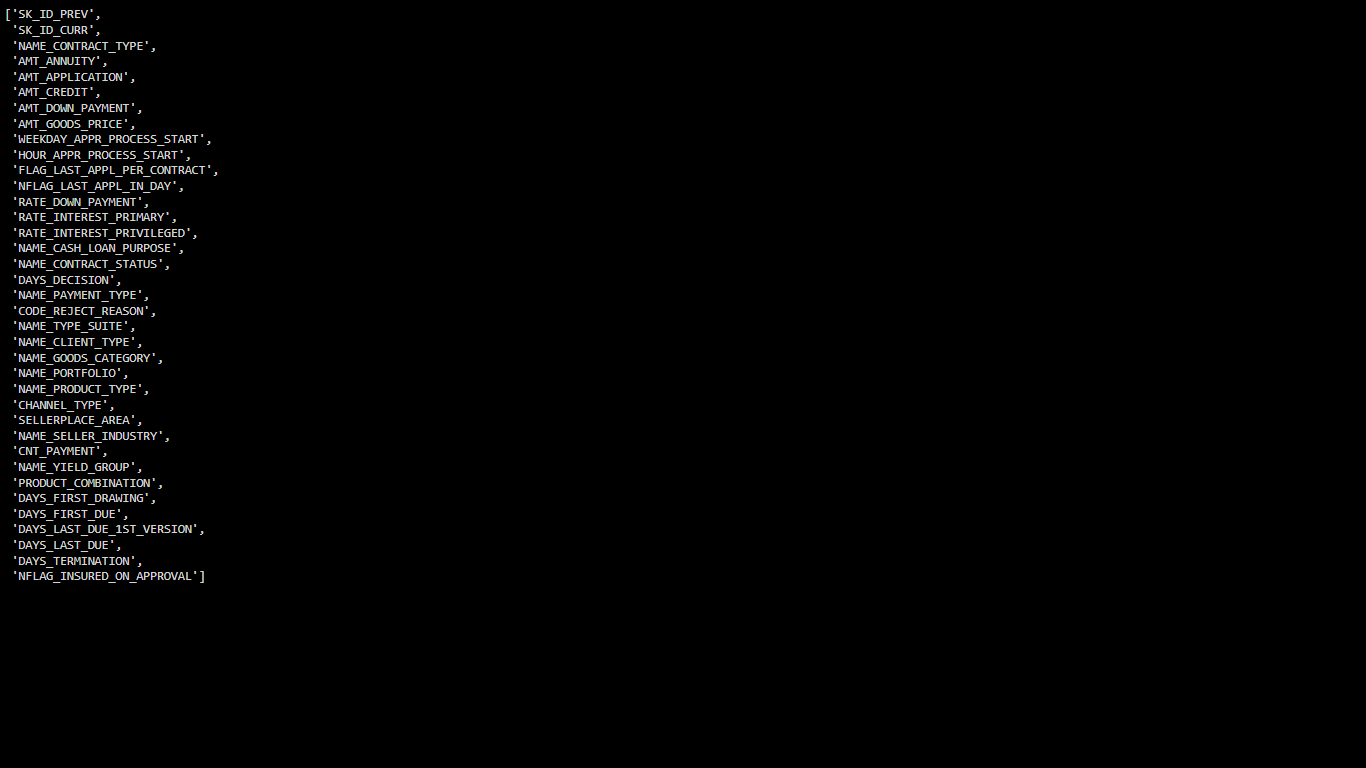
# Working with pervious application data

## Description

The dataframe prev\_app\_data has 37 columns and 1670214 rows. There are 15 columns with float datatype, 6 with integer and 16 with object datatype.

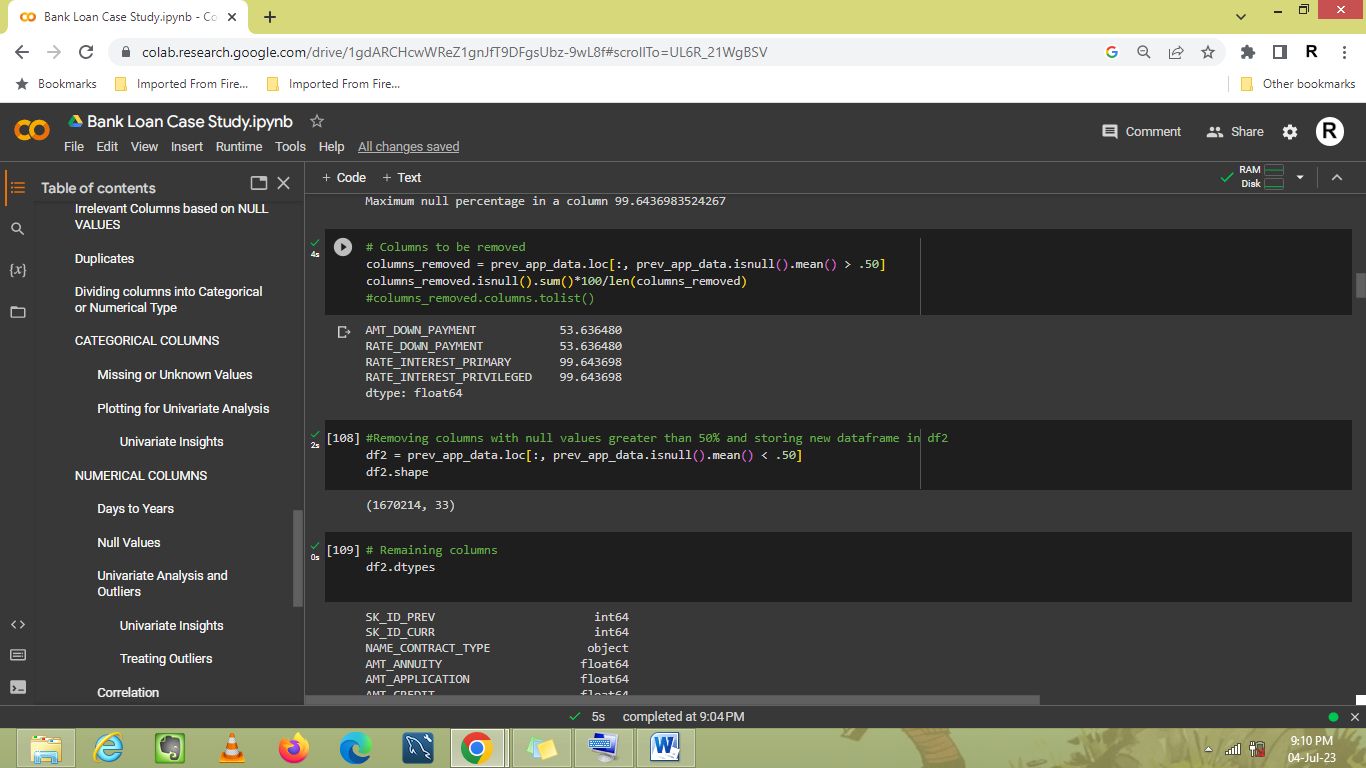
Columns Names:





## Irrelevant Columns

The following columns with Null values > 50% were removed and the data was stored in df2.



Other irrelevant columns that were removed were:

'SK\_ID\_PREV', 'WEEKDAY\_APPR\_PROCESS\_START', 'SELLERPLACE\_AREA', 'HOUR\_APPR\_PROCESS\_START'

The data frame df2 has 1670214 rows and 29 columns.

## Duplicates

There were 74871 duplicate rows in df2.

After removal of these rows, df2 has 1595343 rows and 29 columns.

## Univariate Analysis

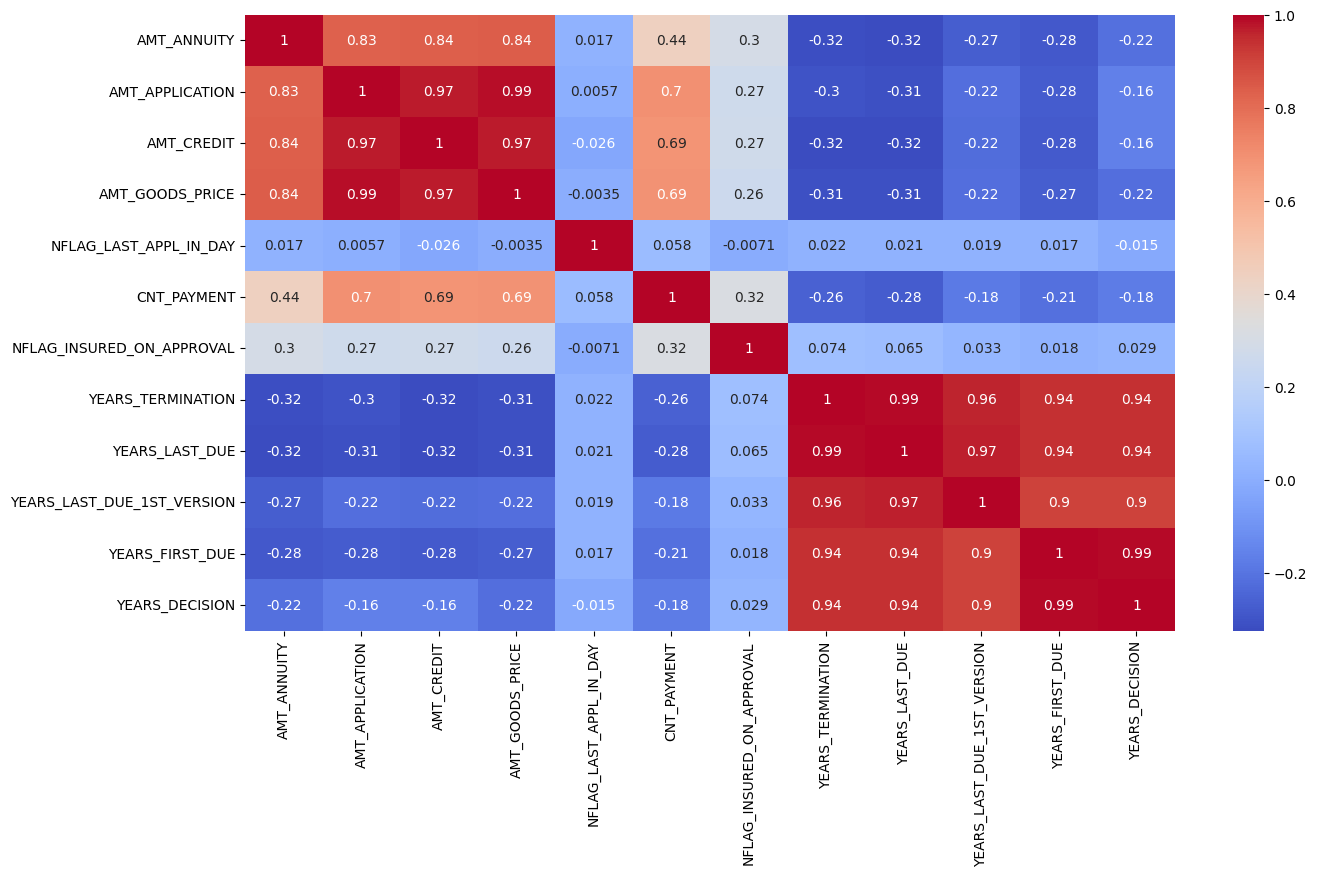
### Categorical Columns

1. Among previous applications, 45% of the application were for consumer loan, 42% for cash loan and about 11% for revolving loans.
2. 65% of the previous applications were approved, 17% were refused, and about 15% were canceled while 1.5% of the offers went unused.
3. At least 64% of the applicants made cash payments through banks.
4. Amongst previous applications, 72% of the applicants were older clients, 18% were new applicants while 8% were refreshed.
5. 43% of applications were made for POS, 28% for cash and about 9% for Cards. Less than 0.1% of applications were made for Cars.
6. 24% of the applications had medium interest rate, 22% had high interest rate, 20% had low normal rate while 5% had low action rate.

### Numerical Columns

1. 75% of the applications were made for a loan amount less than 2lac. The maximum amount for which application was made was of 69 Lac.
2. Most of the applications had credit amount approved up to 2.25 Lac. The maximum credit amount approved was of 69 Lac.
3. Most of the applicants paid an annuity amount between 7000 and 17,000. The maximum annuity amount paid was of 4 Lac. The highest count for an annuity amount was for 10k.
4. For most of the applicants, their last application had terminated 0 to 4 years before.

## Correlation



# Merging datasets

The columns in df2 were renamed with ‘prev\_’ as prefix. The data frames df1 and df2 were then merged into merged\_df data frame.

The new data frame merged\_df has 1351875 rows and 72 columns. The columns in merged data frame are:



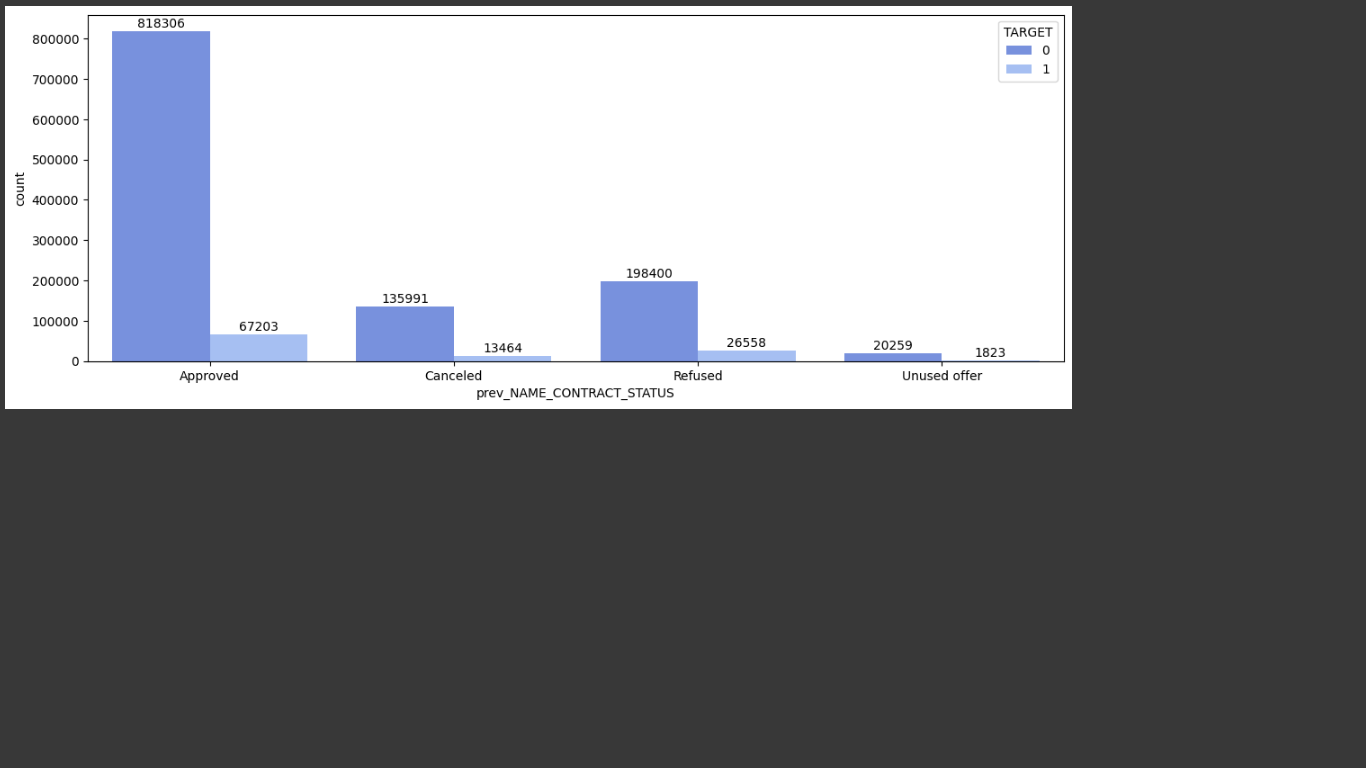
## Duplicates

There were 69871 duplicated rows in 72 columns of merged\_df. After removal of these rows, there were 1282004 rows left for analysis.

## Data Imbalance



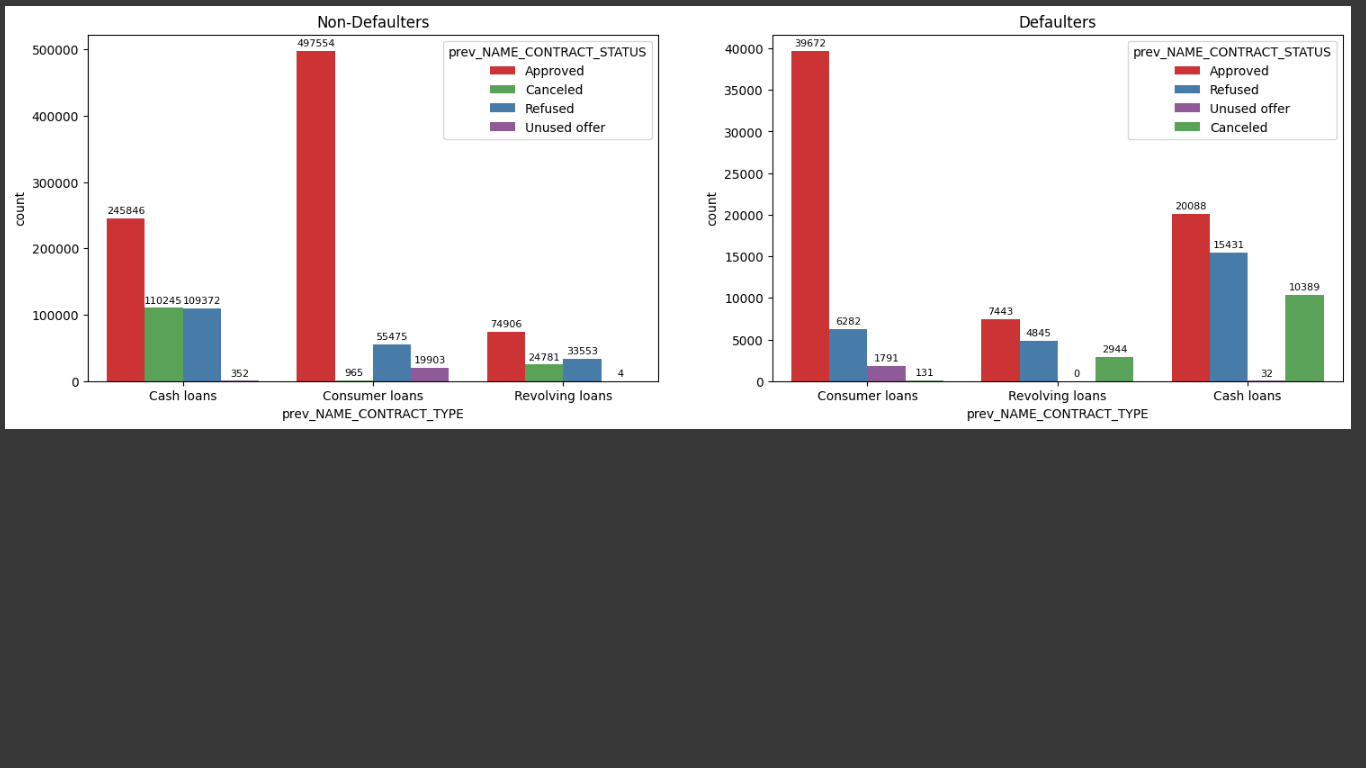
The Non-defaulter to defaulter ratio is 183:17.



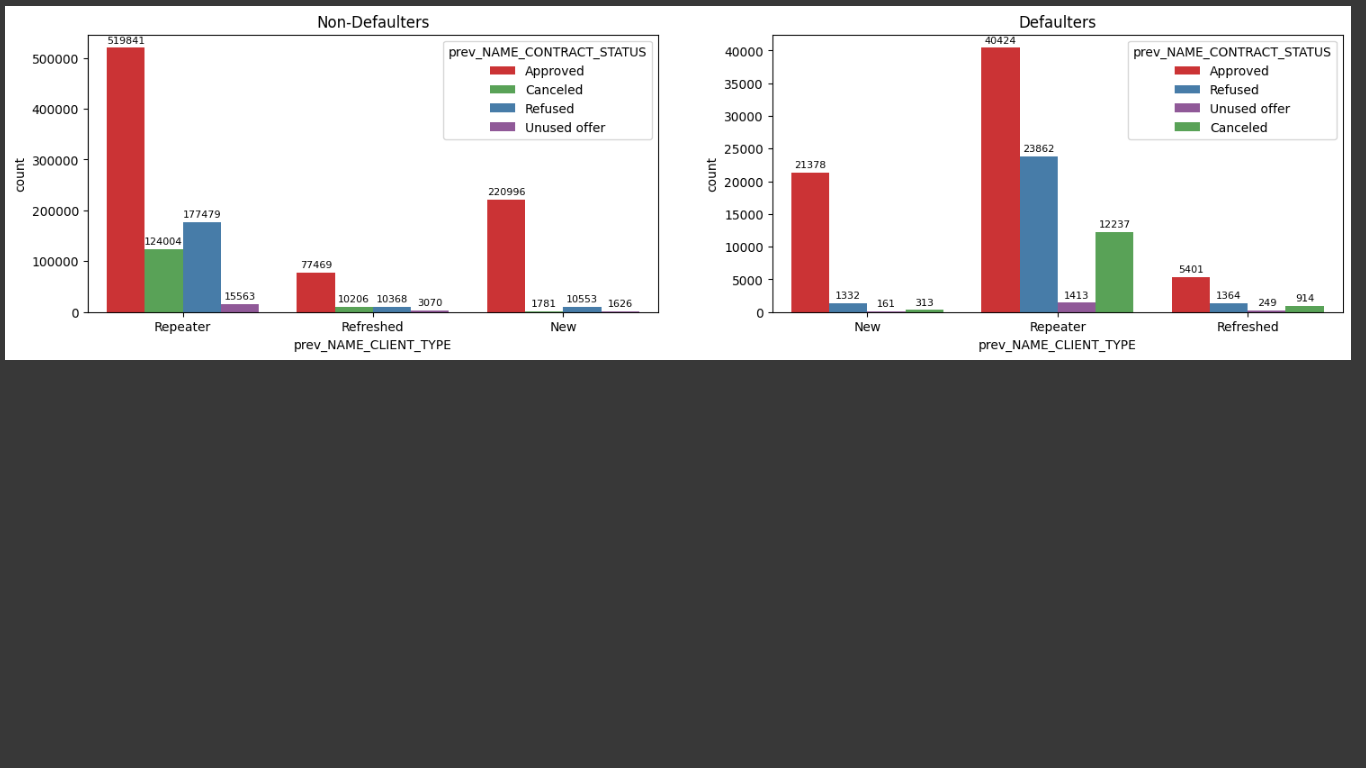
In df2 as well, the amount of approved applications is much higher than other types of applications.

## Bivariate Analysis

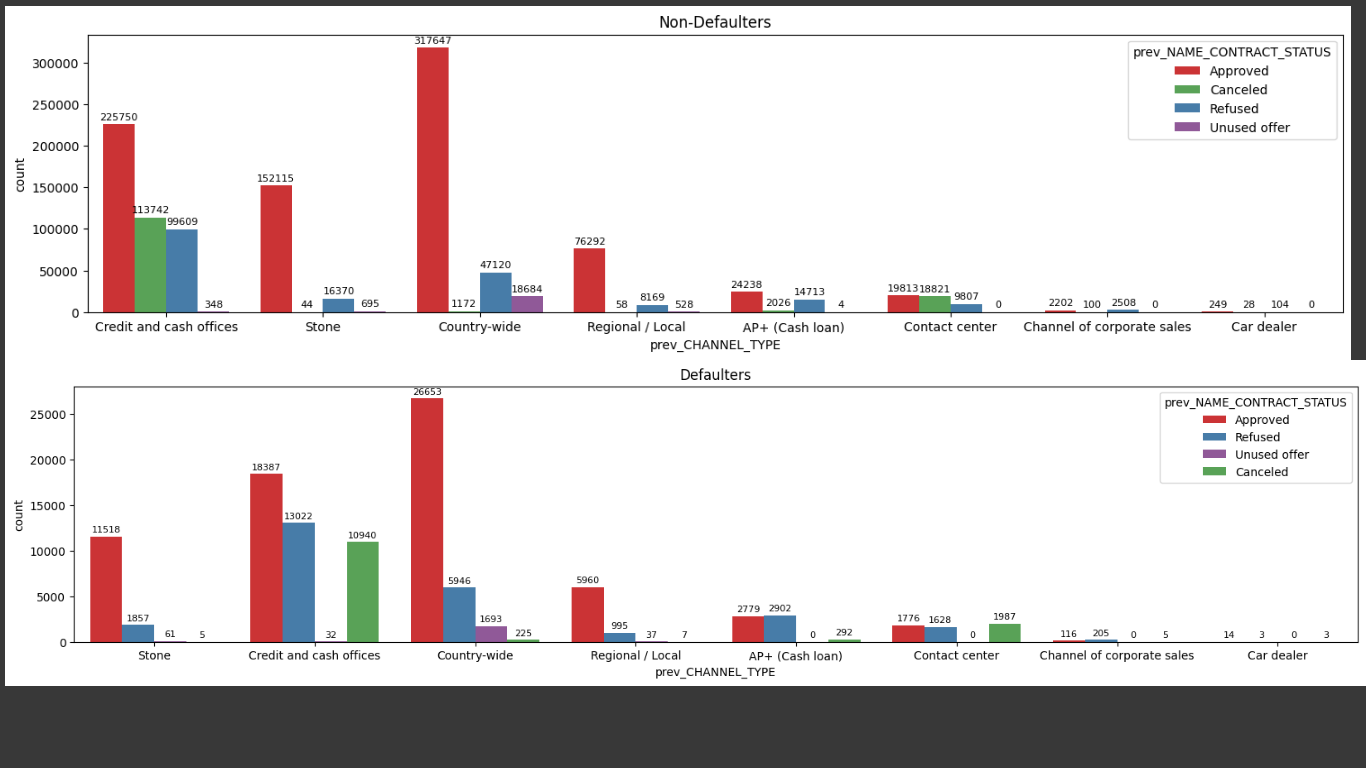
1. The cash loan applications which were refused previously, but approved currently had higher defaulter rate than other applications.



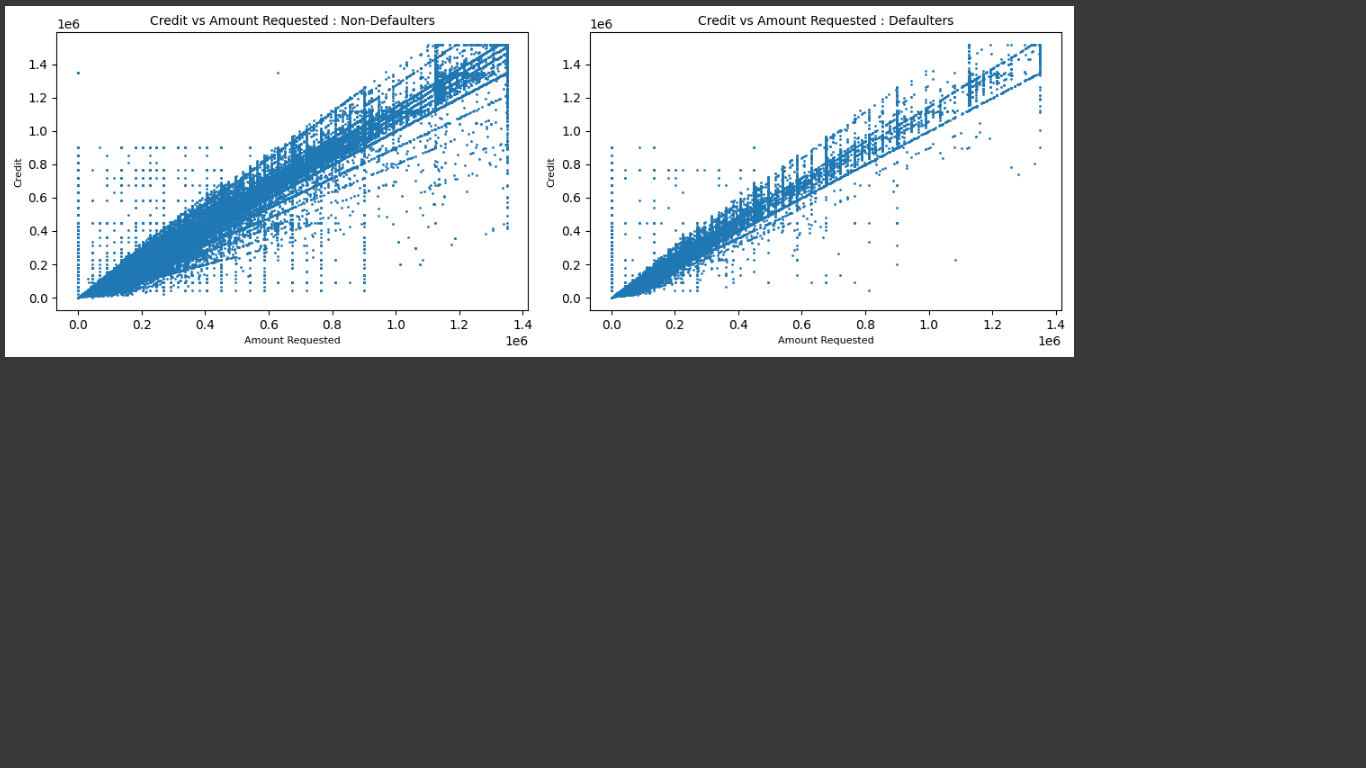
1. Older clients, whose applications had been refused previously, but approved currently, defaulted more than other clients.



1. The number of defaulter clients was higher when acquired through credit and cash offices.



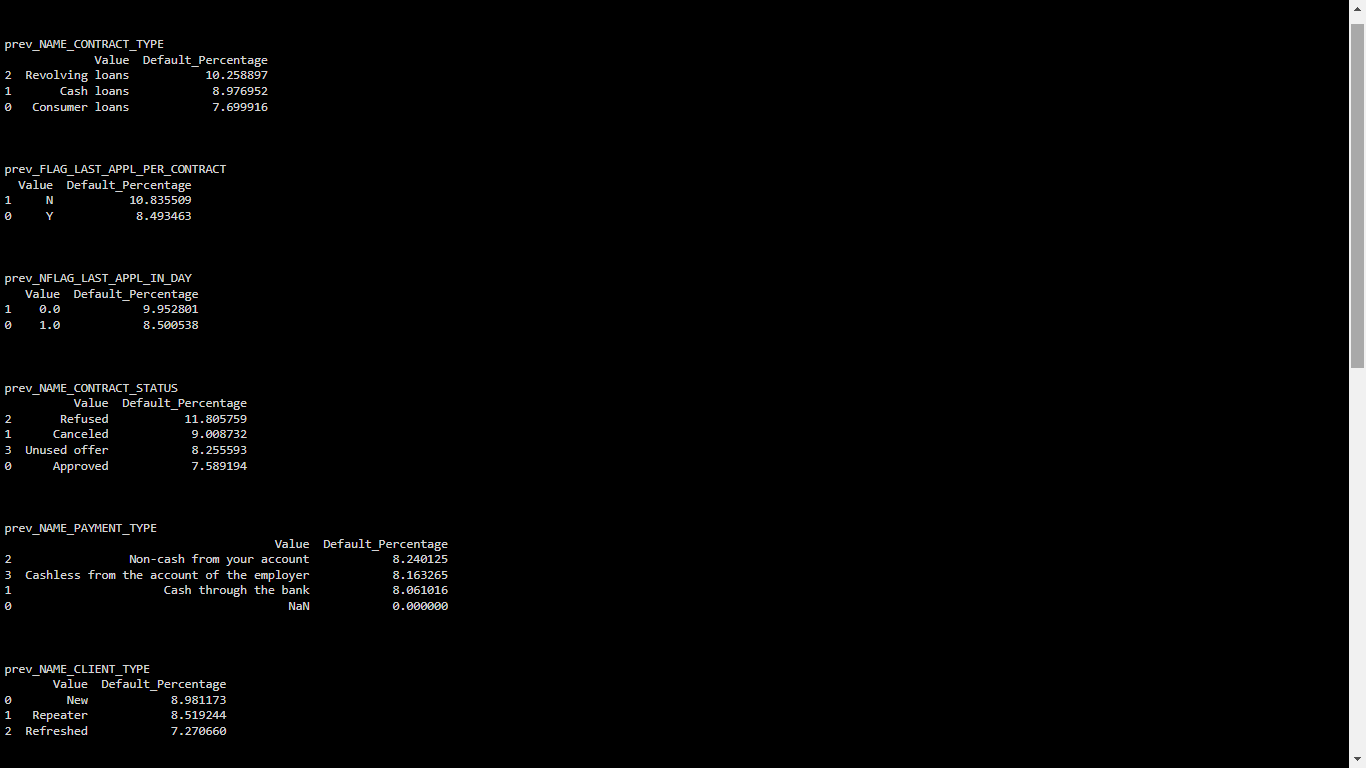
1. Credit approved and amount requested were highly correlated for both defaulters and non-defaulters.

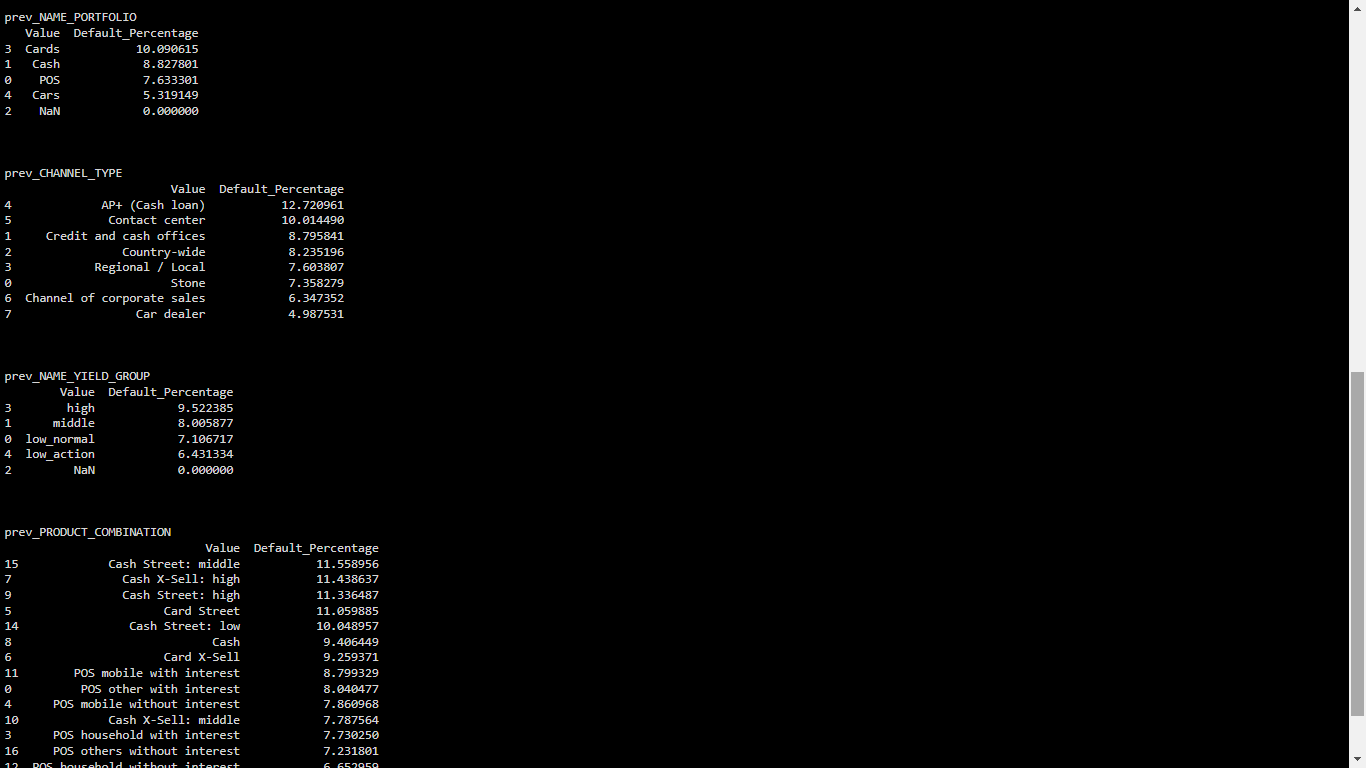


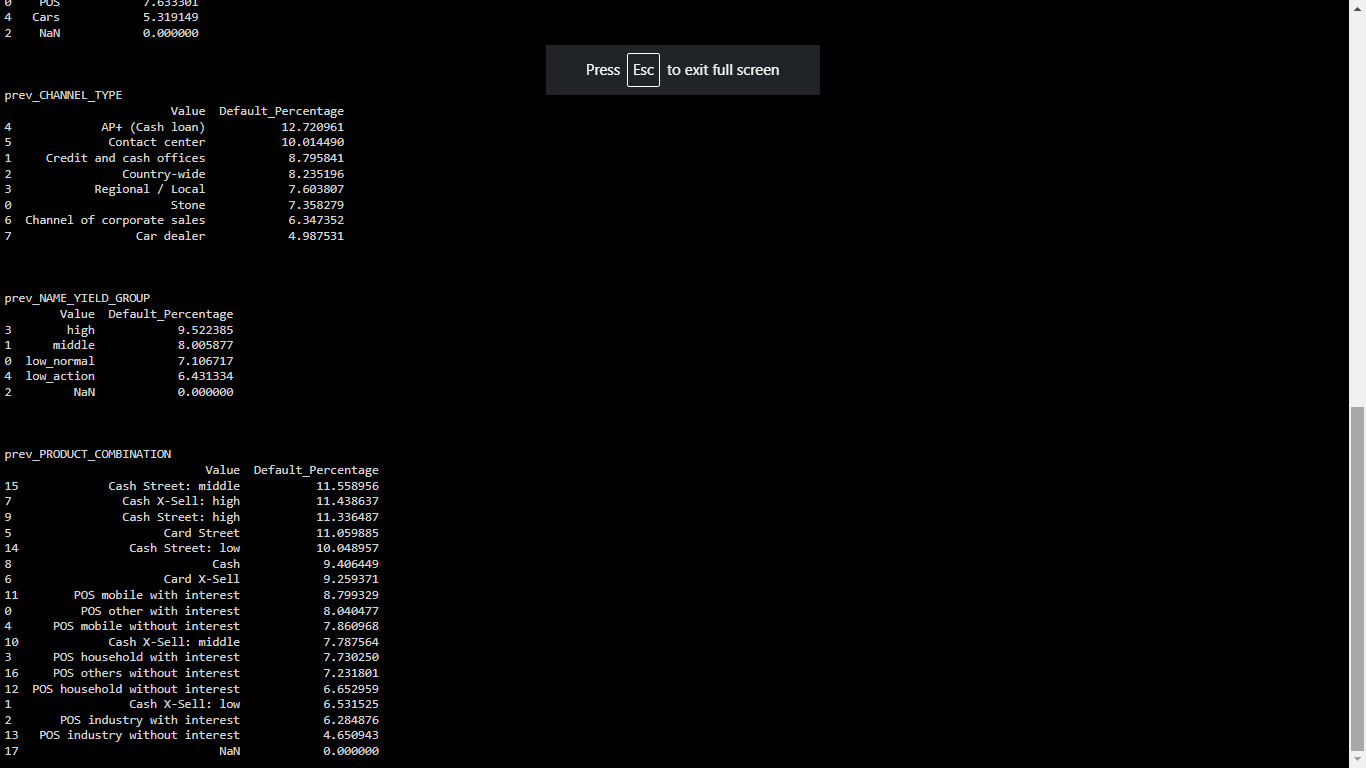
## Segmented Univariate for df2

### Categorical Columns

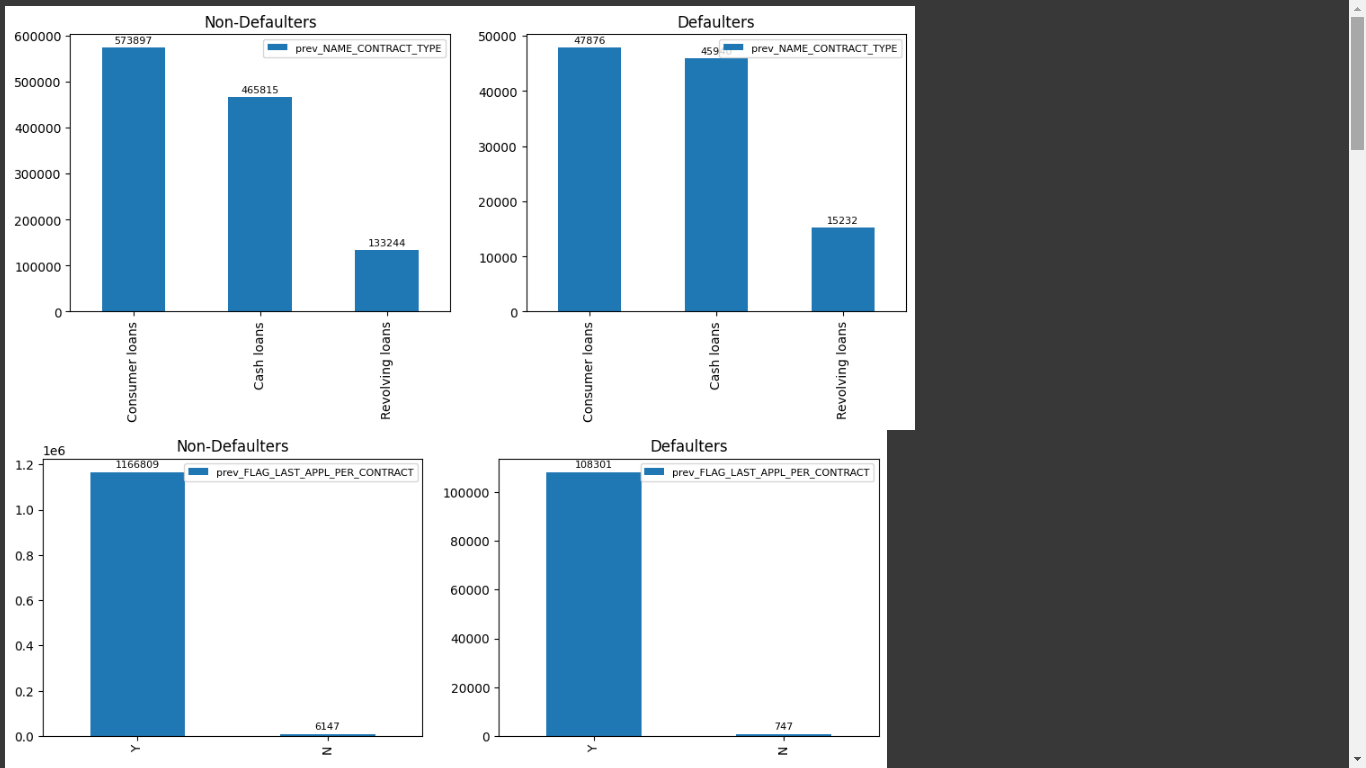
The default percentage for each value in Categorical Columns is:





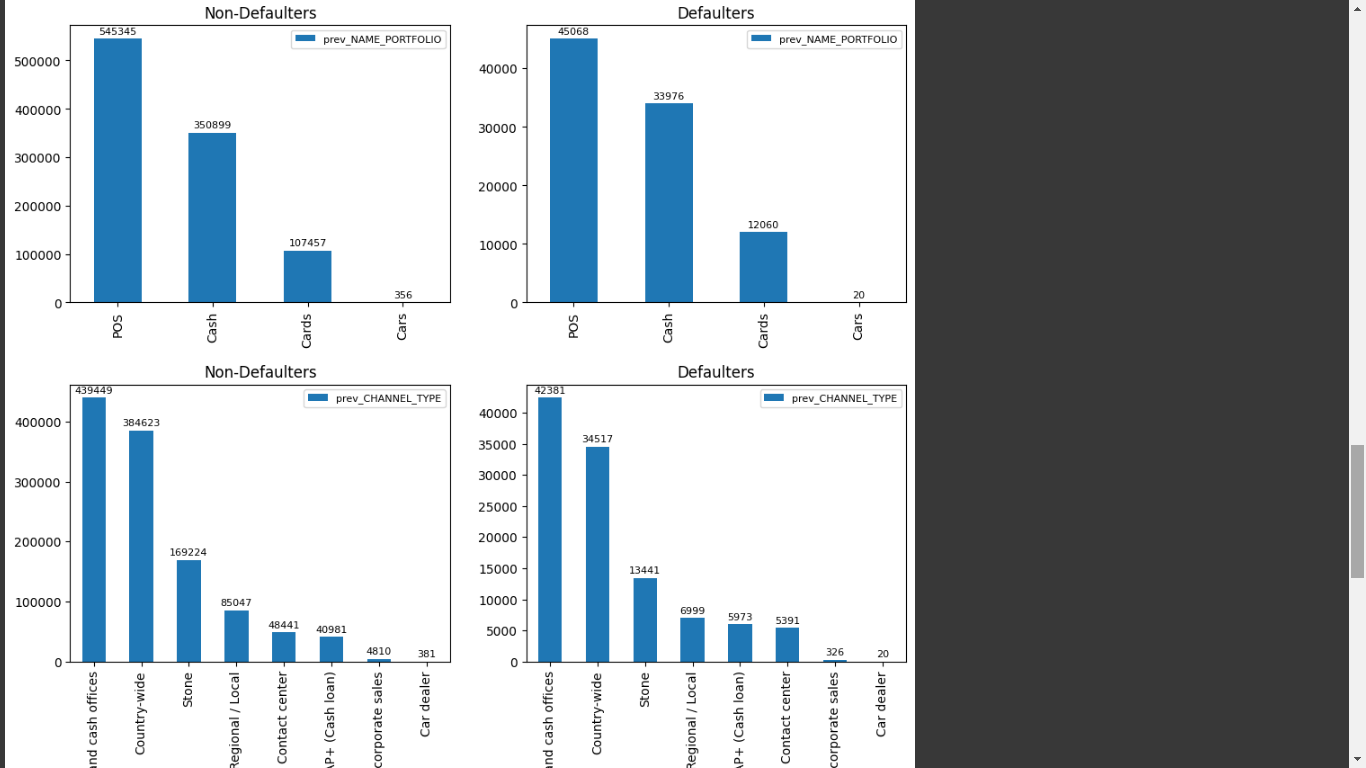


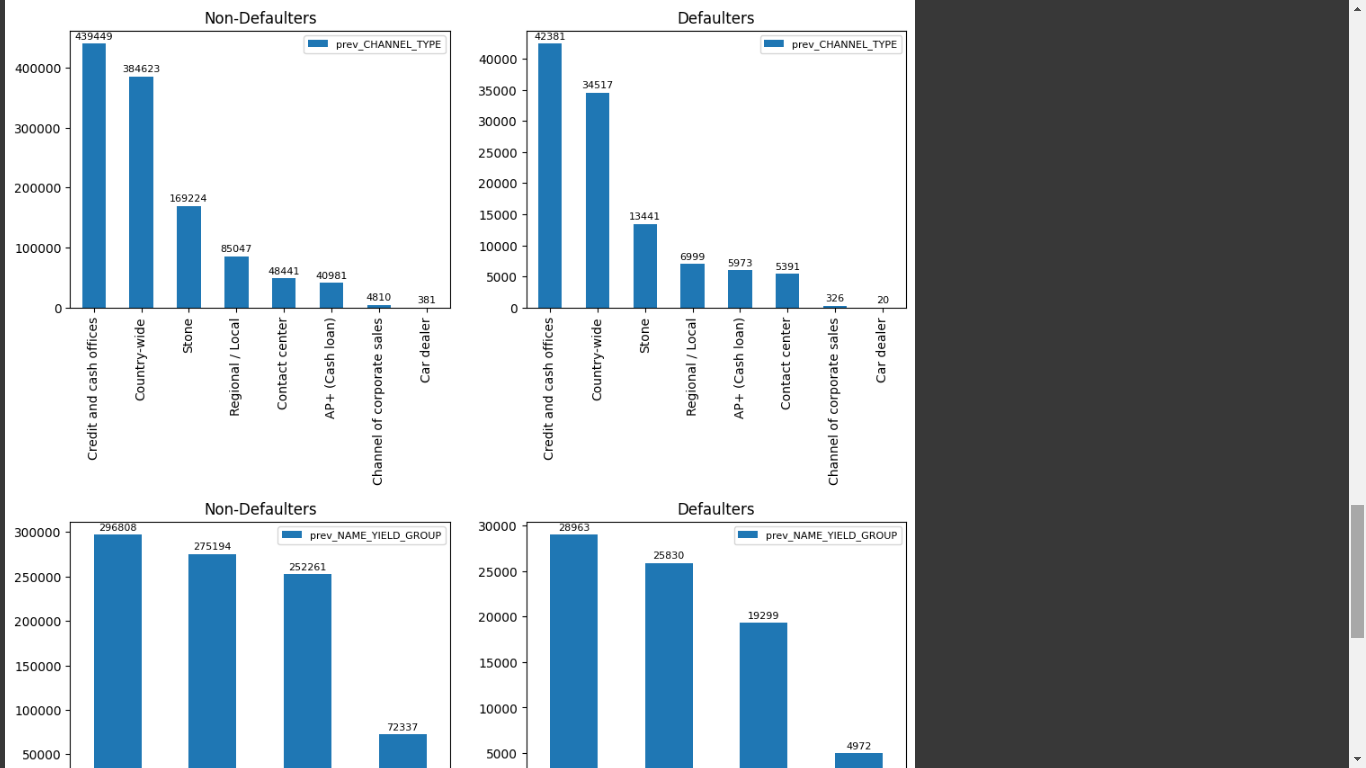
1. Revolving loans have highest default percentage of 10%, followed by cash loans (8.9%) and consumer loans (7.69%).
2. It was observed that the previous applications which were refused (11%), canceled (9%) or went unused (8%) had higher default percentage when approved.
3. The payment type didn’t have any significant relationship with change in default percentage.
4. The new clients had a little higher default percentage than the older ones, but nothing significant.



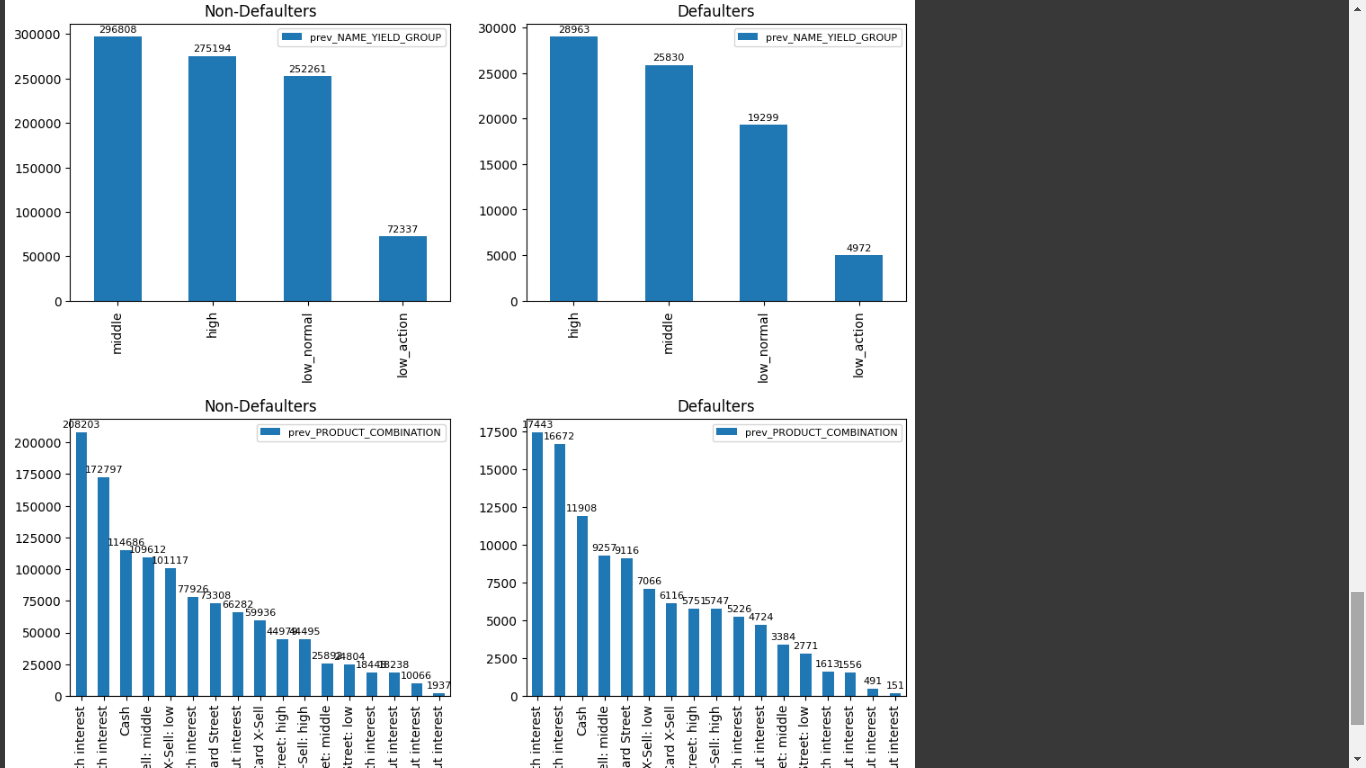


1. The portfolio type Cards had the highest default percentage (10.09%) while Cars had the least (5.3%).
2. The clients acquired through AP+ (Cash loan) and Contact Centre had higher default percentages of 12.7% and 10% respectively.





1. The default percentage increased with the level of yield group. Yield group high had a default percentage of 9.5%.
2. Amongst product combinations, all of Cash Street groups recorded higher default percentage. Cash X-sell high and Card Street also recorded high default percentage.



### Numerical Columns

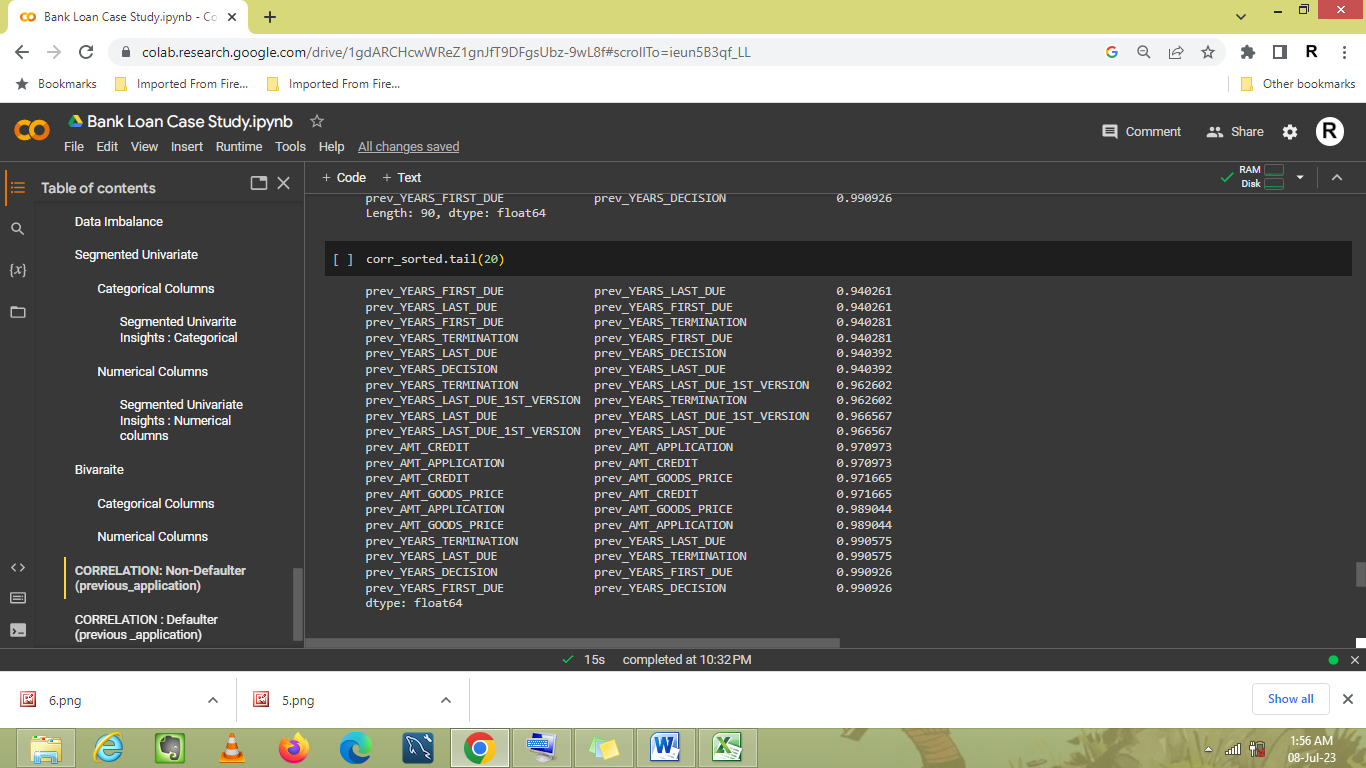
For values such as Application Amount, Credit Amount, Annuity Amount, etc., the graphs for defaulters as well as non-defaulters in previous applications followed similar trends.

## Correlation

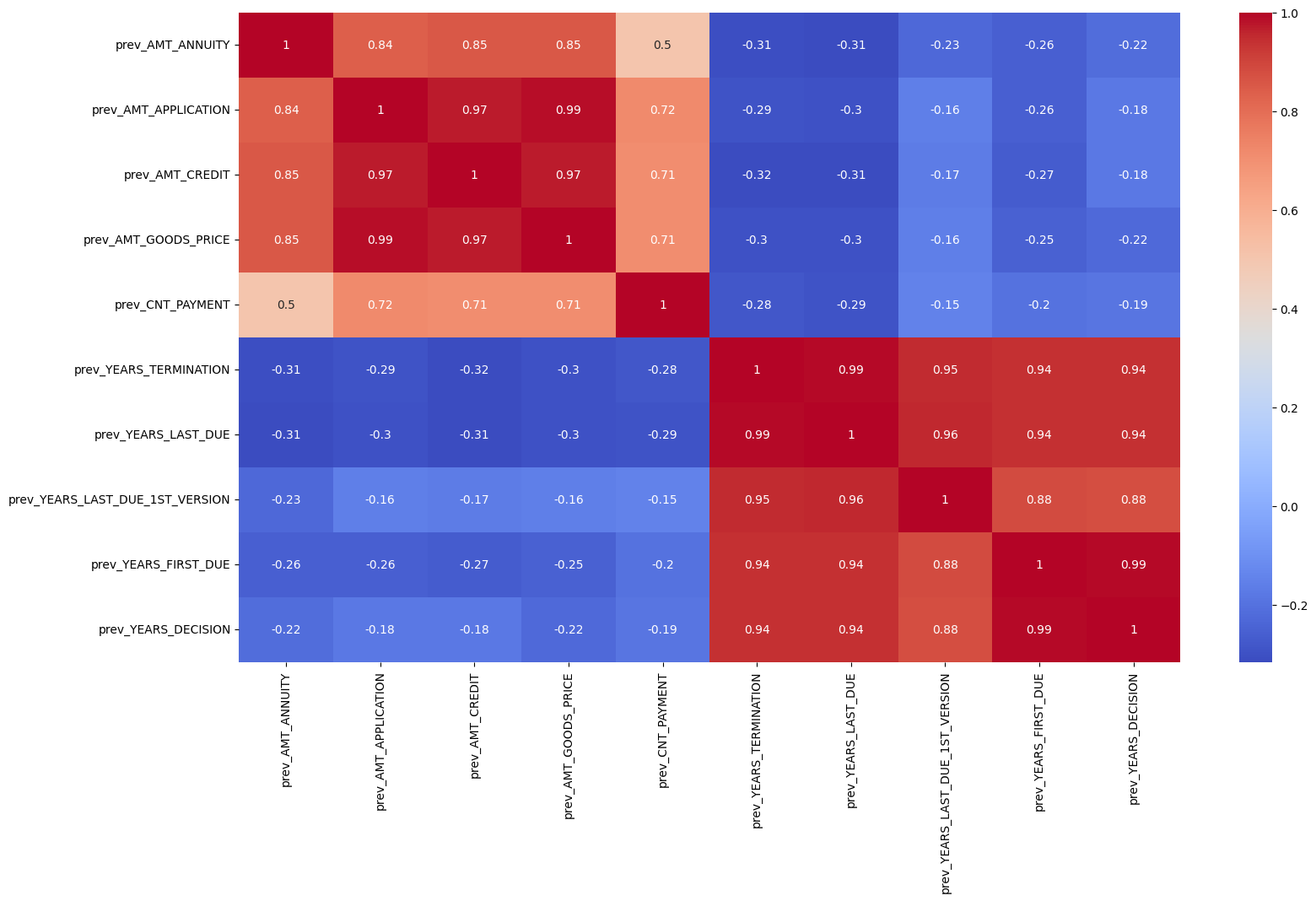
### Non-defaulter data

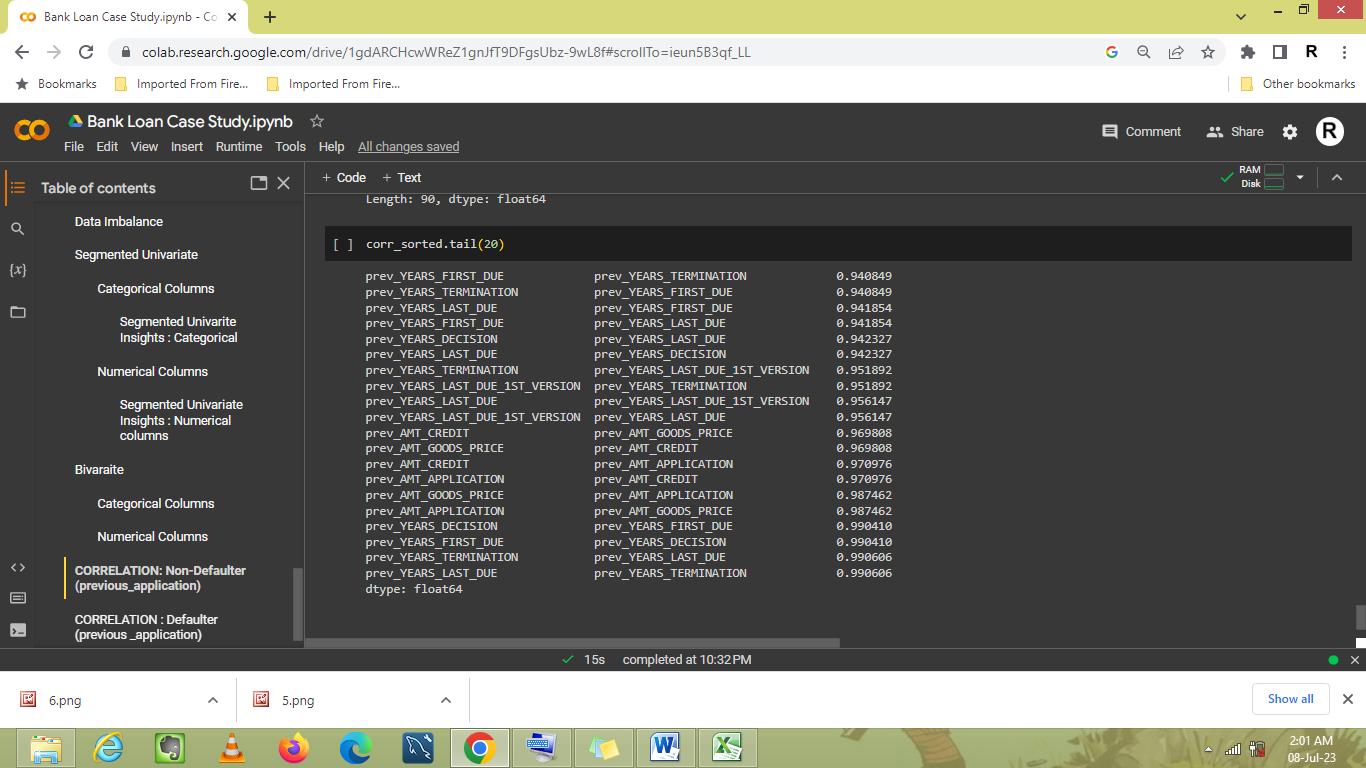


Top 10 Correlations (Non-defaulter)



### Defaulter data

Top 10 Correlations (Defaulters)



# Final Insights:

1. Males, while being less in number, defaulted more than women.
2. The applicants with lower secondary education and people in low skilled labour defaulted more than other types.
3. The accommodation type Other\_B had the higher percentage of defaulters while people accommodating with family members, especially children, had the smaller default percentage. However, it was observed that defaulter percentage increased with an increase in the count of children/family members.
4. People living in rented apartments, on maternity leave or unemployed had the higher default percentage.
5. The people whose contact/work address didn’t match permanent address defaulted more than the ones whose did.
6. Region rating 3 had highest default percentage. Moreover, as the observations of client's social surroundings with defaults increased, the default percentage also increased.
7. The clients with higher number of enquiries to Credit Bureau in last one year (excluding last 3 months before application) had higher default percentage.
8. It was observed that the previous applications which were refused (11%), canceled (9%) or went unused (8%) had higher default percentage when approved.
9. The portfolio type Cards had the highest default percentage (10.09%) while Cars had the least (5.3%).
10. The clients acquired through AP+ (Cash loan) and Contact Centre had higher default percentages of 12.7% and 10% respectively.
11. The default percentage increased with the level of yield group. Yield group high had a default percentage of 9.5%.
12. Amongst product combinations, all of Cash Street groups recorded higher default percentage. Cash X-sell high and Card Street also recorded high default percentage.