**TEAM 15: ADVANCED ANALYTICAL SKILLS**

**Problem Statement**

This analysis is done for the loan providing companies who find it hard to give loans to the people due to their insufficient or non-existent credit history. Basic data provide gave us a lot of information about how the consumers use their credit history to their advantage by becoming a defaulter. This analysis will ensure that the applicants capable of repaying the loan are not rejected.

**To achieve the model following methods were followed:**

1. **Cleaning the Data**:

This was observed that the size of the data was huge and many of the columns were null value. So this was mandatory to clean the data and columns with duplicate data shall be handled.

There are feature columns in the dataset that are highly correlated to each other. Which means both will have similar impact on the target value. Those features can be removed before feeding this data to a model to avoid collinearity.

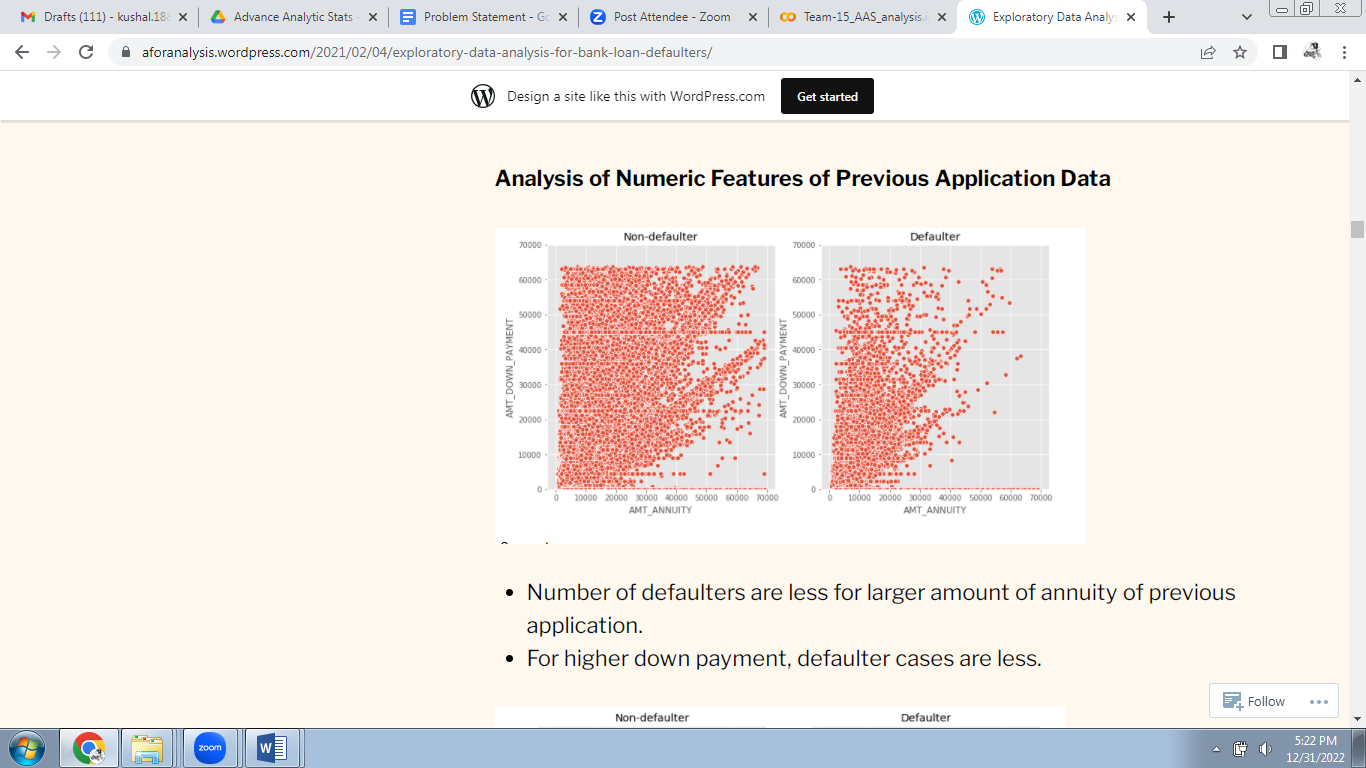
Feature columns with 50% or more missing data can be dropped. Following columns should be converted to integer. DAYS\_FIRST\_DRAWING float64 DAYS\_FIRST\_DUE float64 DAYS\_LAST\_DUE\_1ST\_VERSION float64 DAYS\_LAST\_DUE float64 DAYS\_TERMINATION float64.

This categorical column has only 0 and 1 and hence can be converted into integer column. NFLAG\_INSURED\_ON\_APPROVAL float64.

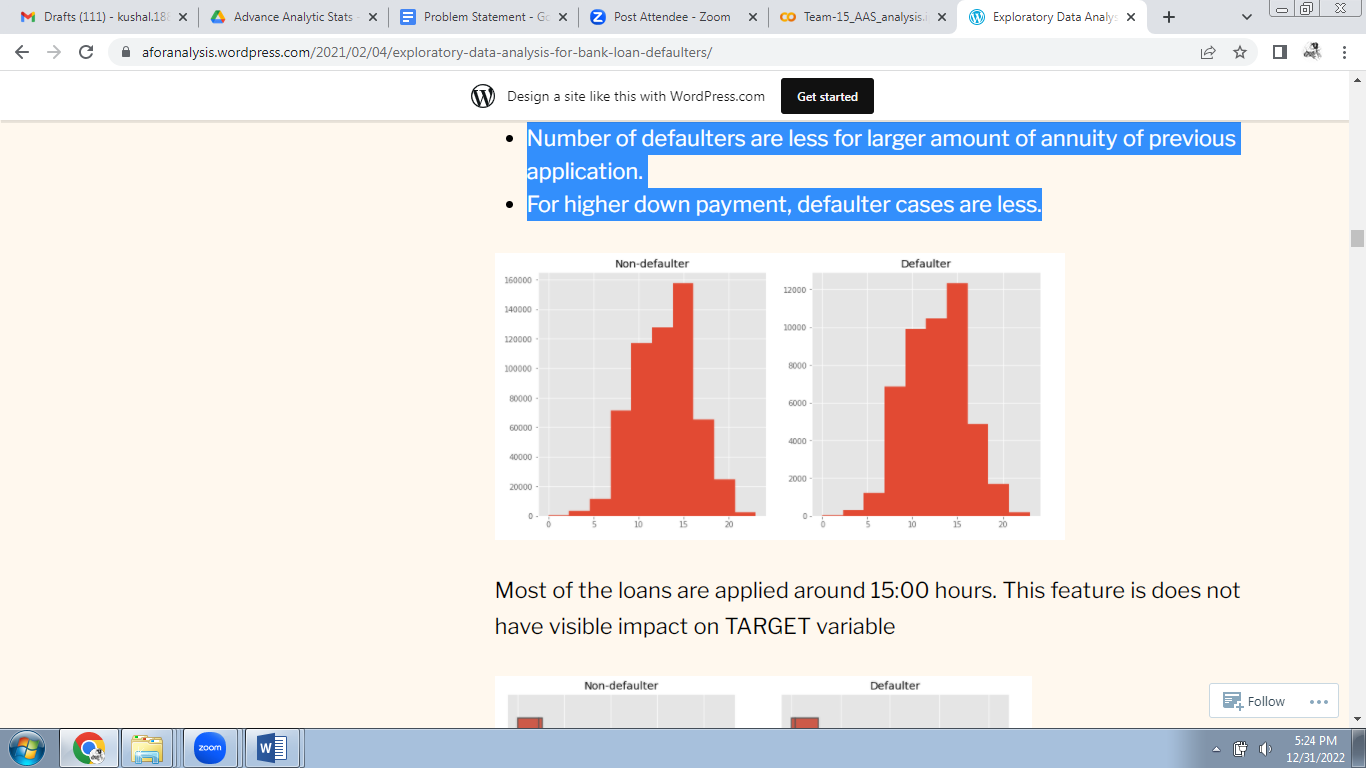
2. **EDA**:

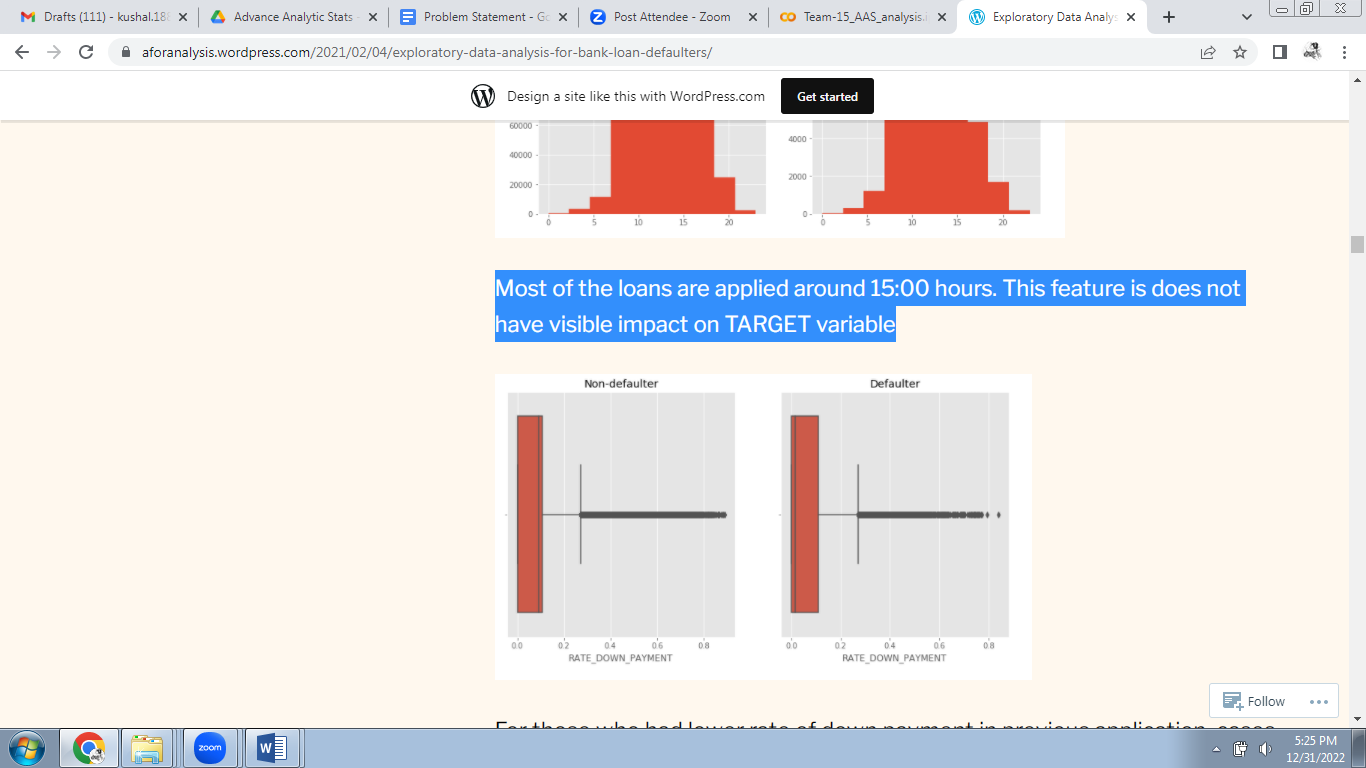
Exploratory Data Analysis was done to check categorical variables. We can find that some numerical variables consisted of very high values as compared to their respective means.

1. **Analysis of Numeric Features of Previous Application Data**



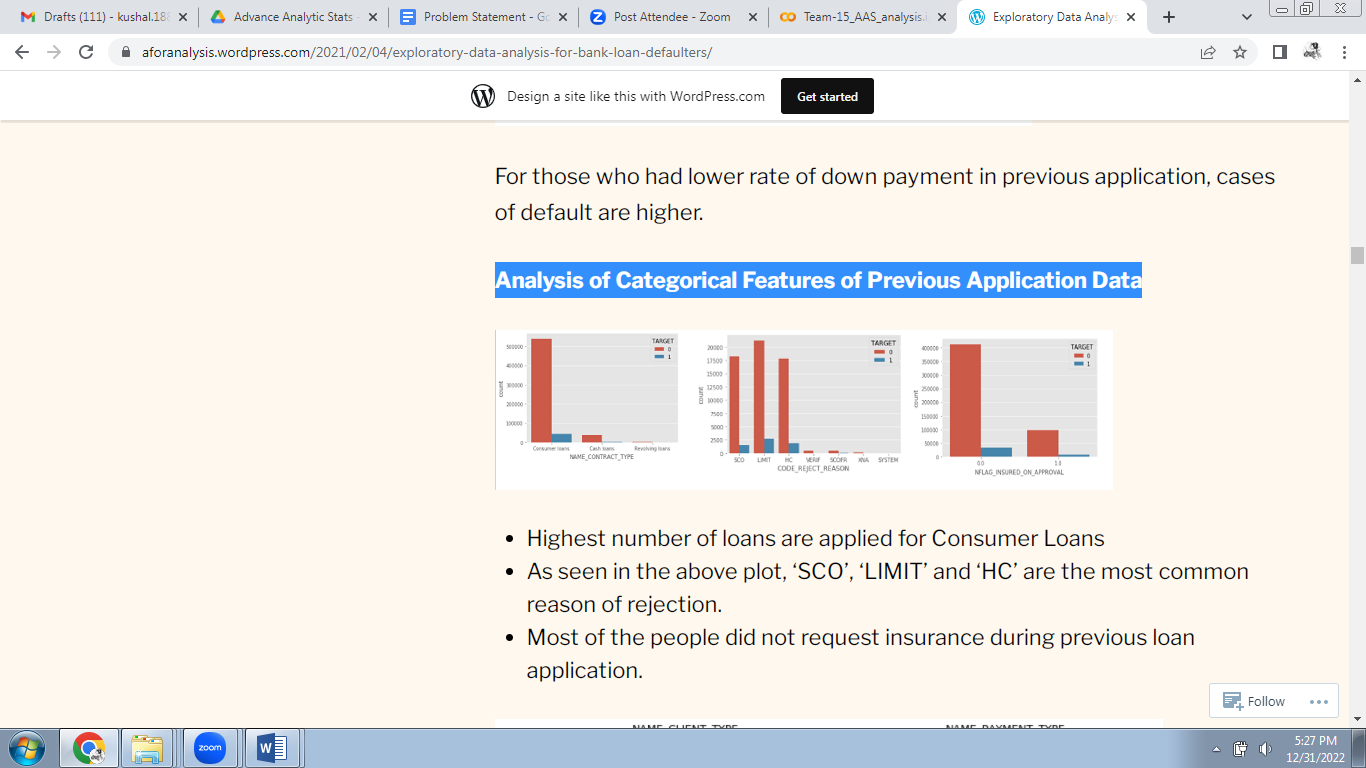
* Number of defaulters are less for larger amount of annuity of previous application.
* For higher down payment, defaulter cases are less.

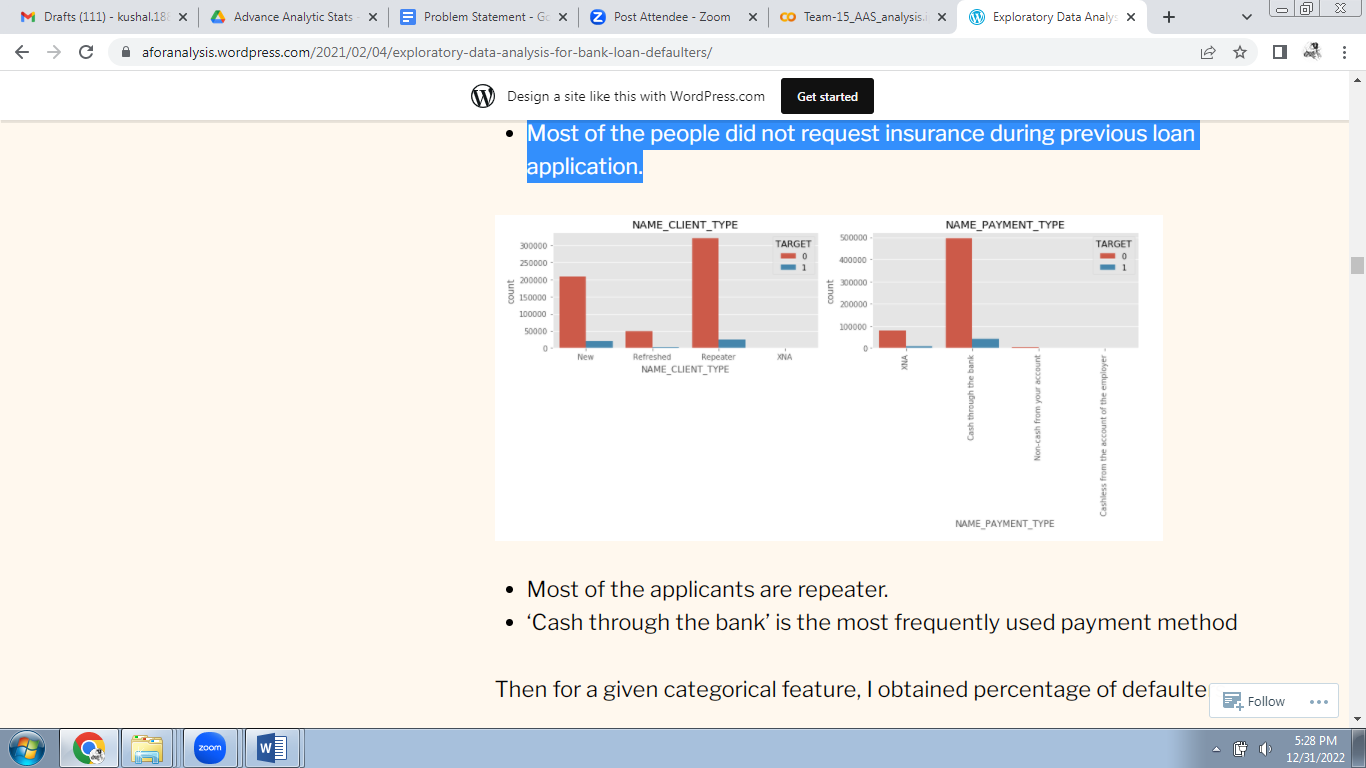


* Most of the loans are applied around 15:00 hours.
* This feature is does not have visible impact on TARGET variable.

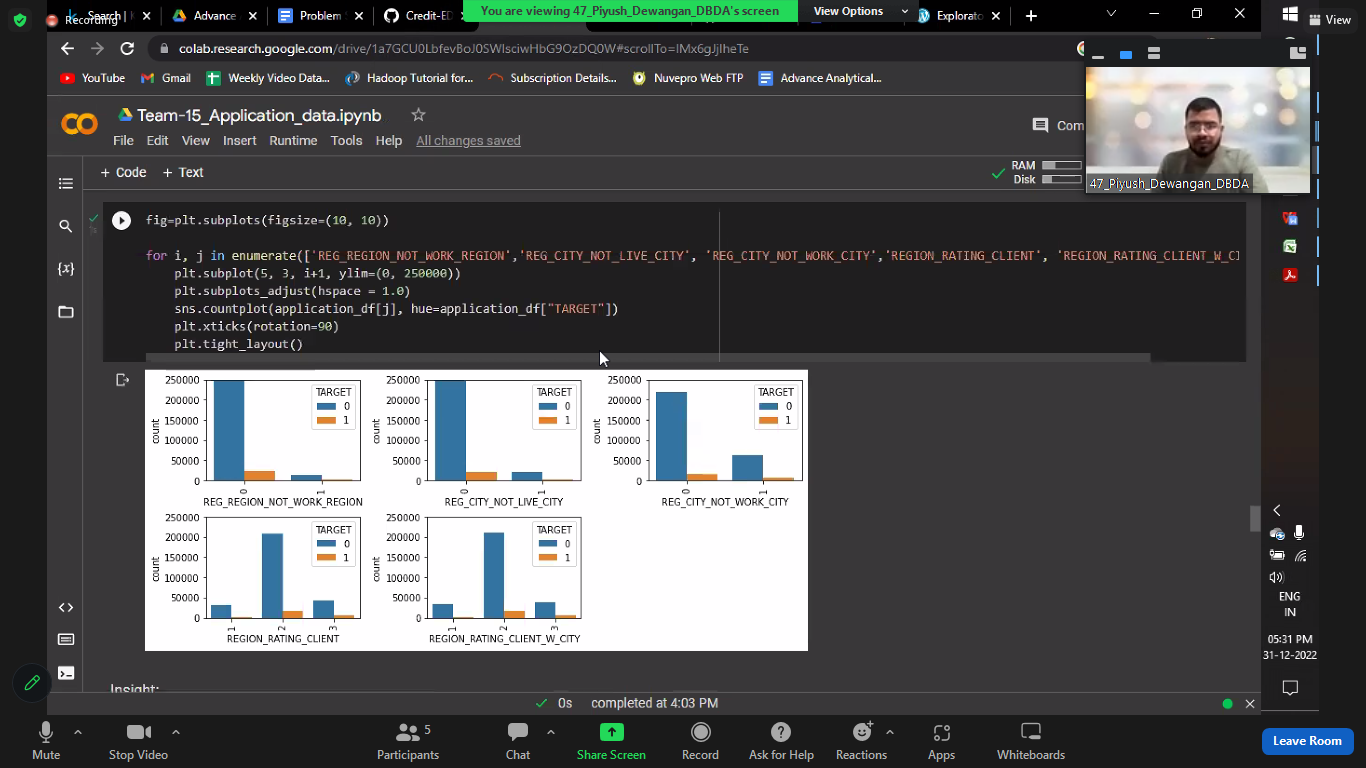
For those who had lower rate of down payment in previous application, cases of default are higher.

1. **Analysis of Categorical Features of Previous Application Data**

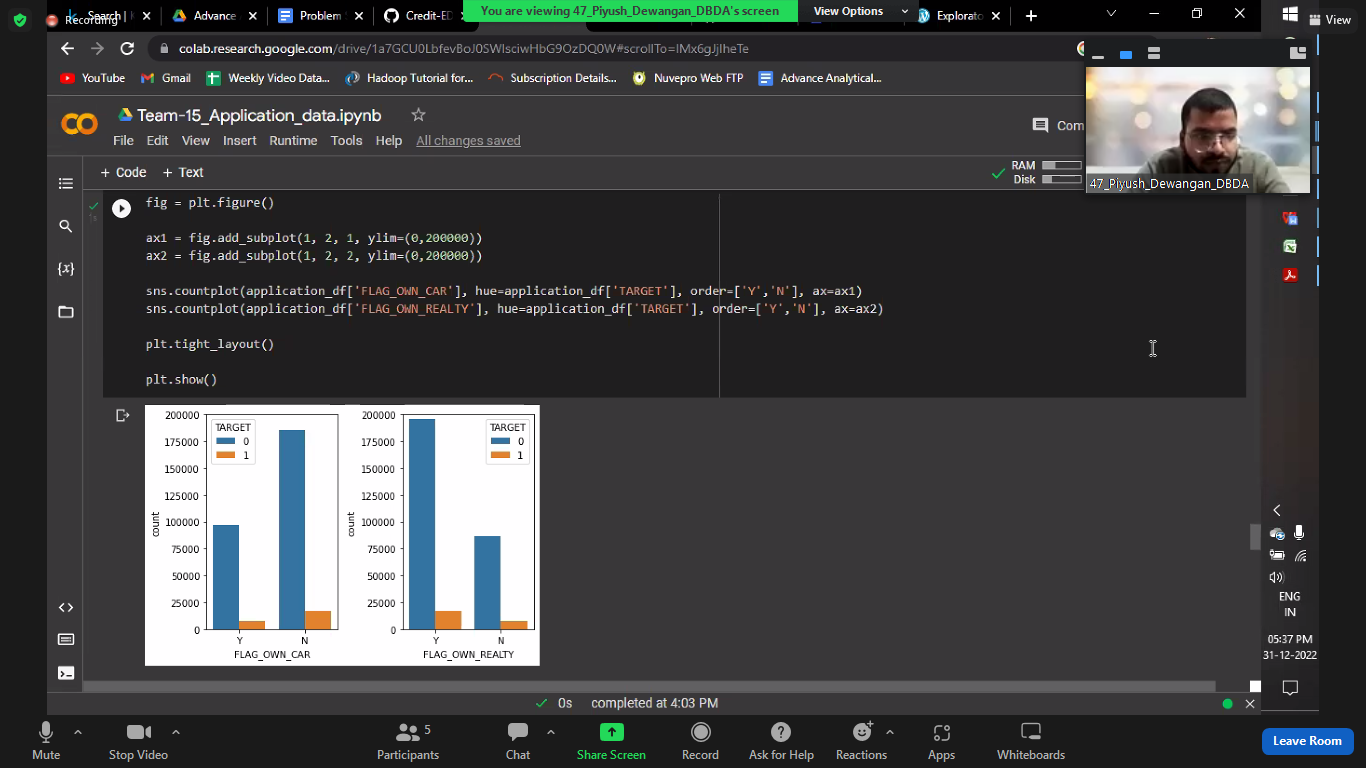


* Highest number of loans are applied for Consumer Loans
* As seen in the above plot, ‘SCO’, ‘LIMIT’ and ‘HC’ are the most common reason of rejection.
* Most of the people did not request insurance during previous loan application.
* Most of the applicants are repeater.
* ‘Cash through the bank’ is the most frequently used payment method

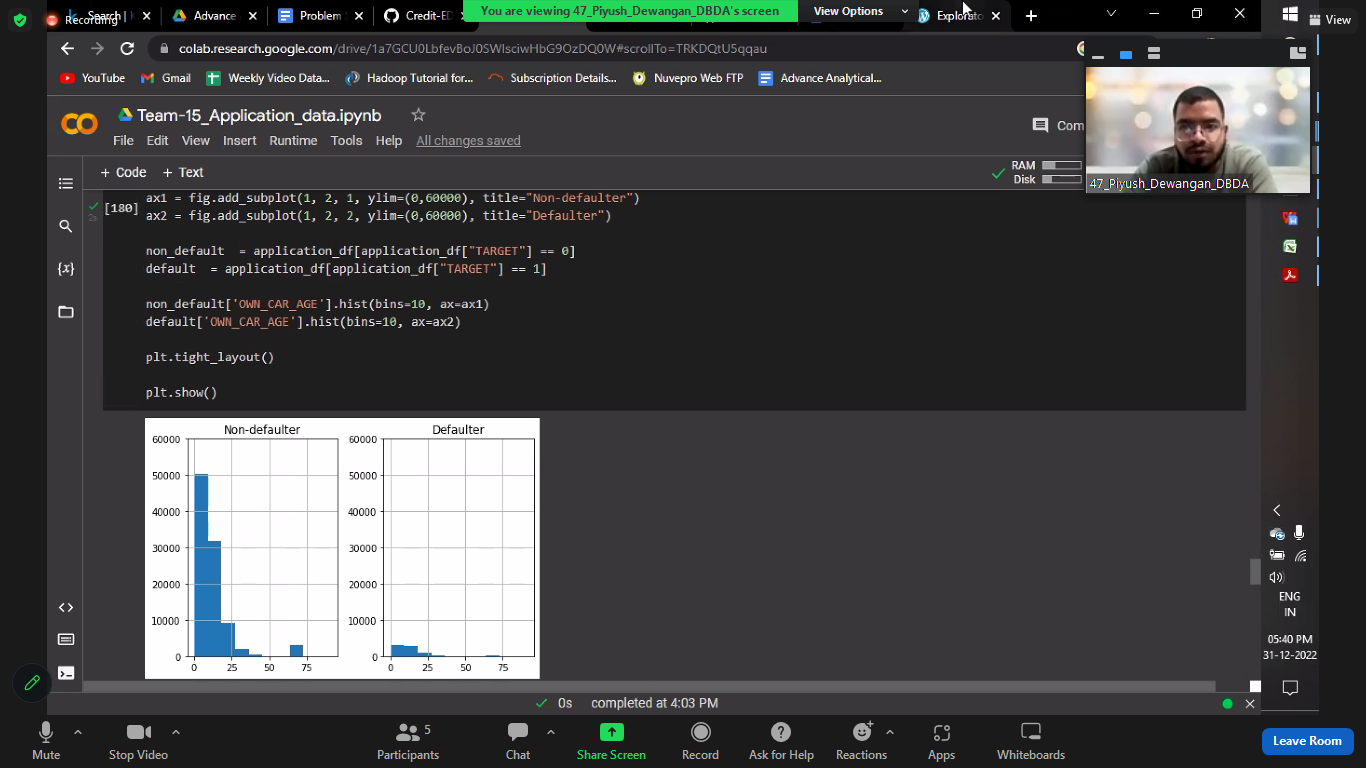
1. **Region related analysis of Current Application Data**



* Defaulter rate is highest when REG\_REGION\_NOT\_WORK\_REGION=0 i.e. permanent address and working address is same
* Highest Applicants have Region rating of 2

1. **Asset related analysis of Current Application data**

* Most of the applicants own realty
* Most of the applicants do not own cars
* People not owning reality and car and have a slightly higher default rate than the people who own reality and car

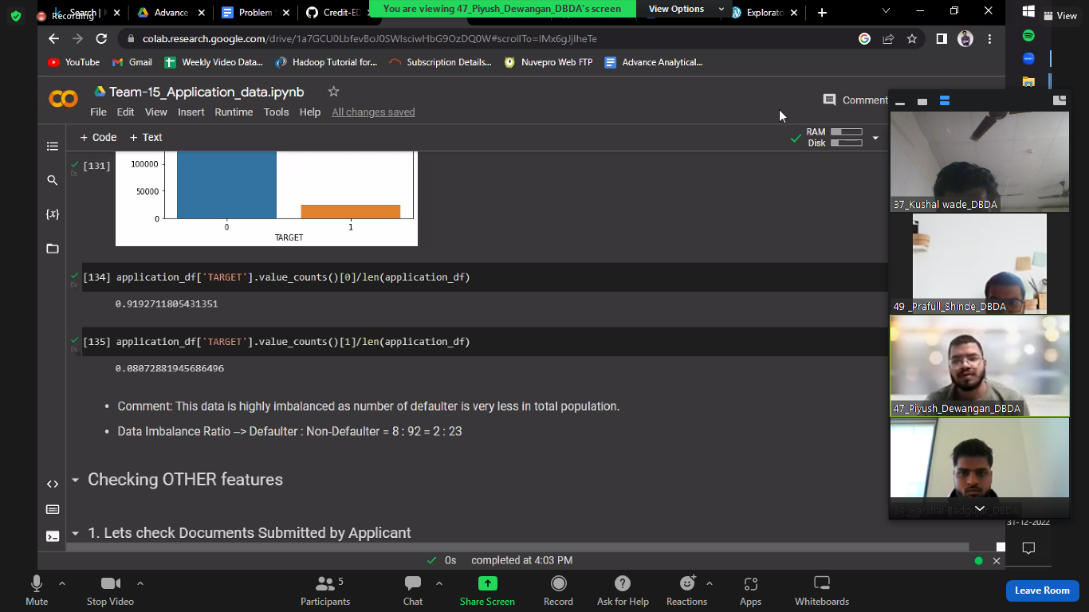


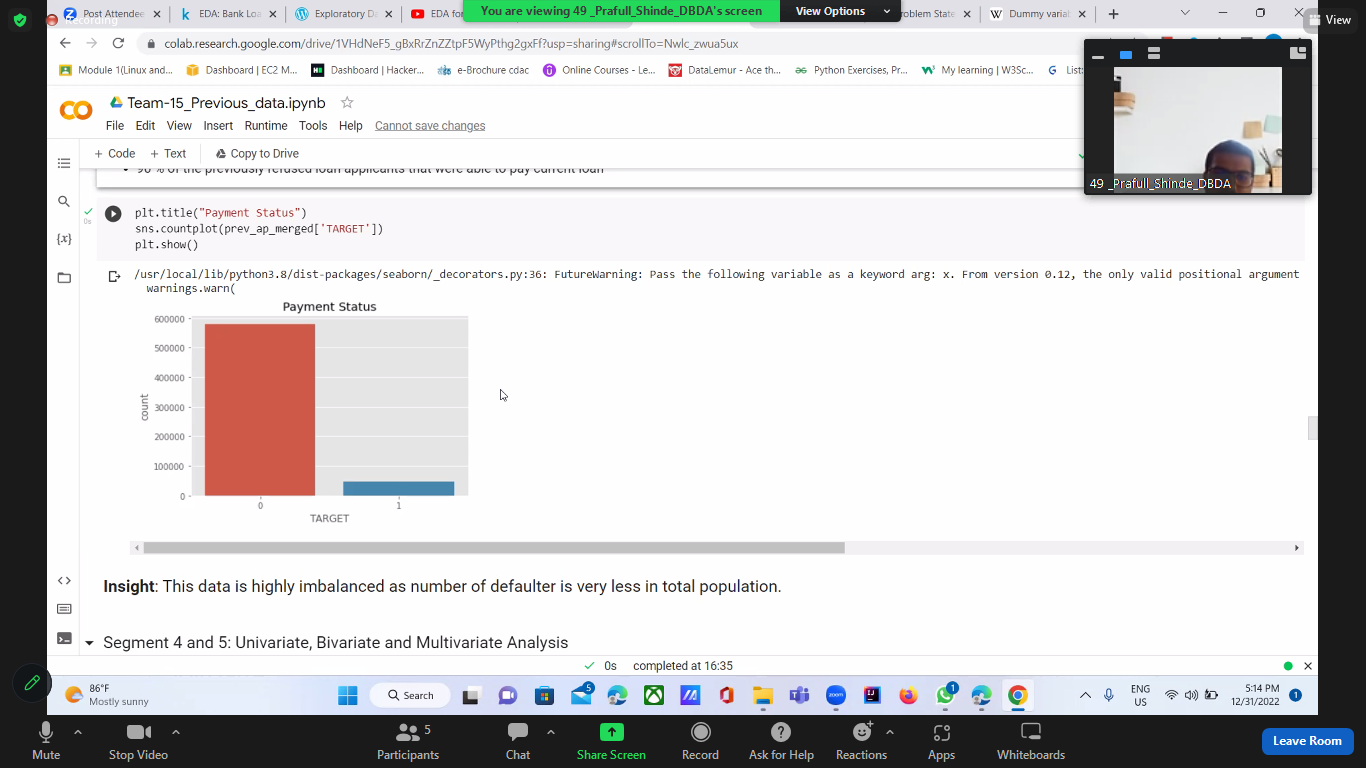
* Defaulter or not, most applicants have car age between 0-25 years.
* Since for both target value, trend is similar, this feature can dropped.

**E) Further analysis of current application data can be seen on notebook.**

**3. Handling Imbalanced Data:**

* This dataset is highly imbalanced
* The applicants whose previous loans were approved are more likely to pay current loan in time, than the applicants whose previous loans were rejected. NAME\_CONTRACT\_STATUS is an important feature.
* 7% of the previously approved loan applicants that defaulted in current loan. 90 % of the previously refused loan applicants that were able to pay current loan. ‘SCO’, ‘LIMIT’ and ‘HC’ are the most common reason of rejection.
* Most of the people did not request insurance during previous loan application. For “Cards” defaulter percentage is highest (17%). ‘NAME\_PORTFOLIO’ is an important feature for analyzing ‘TARGET’ variable.



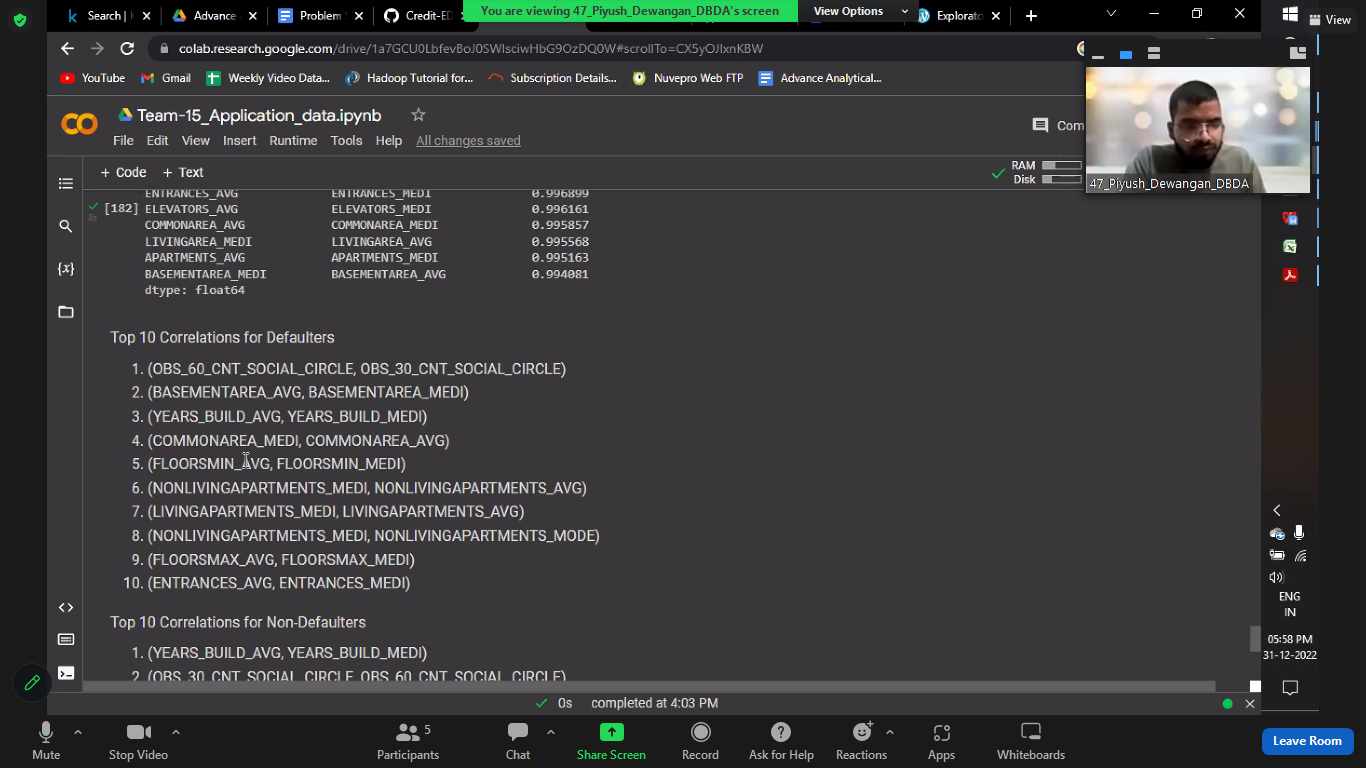


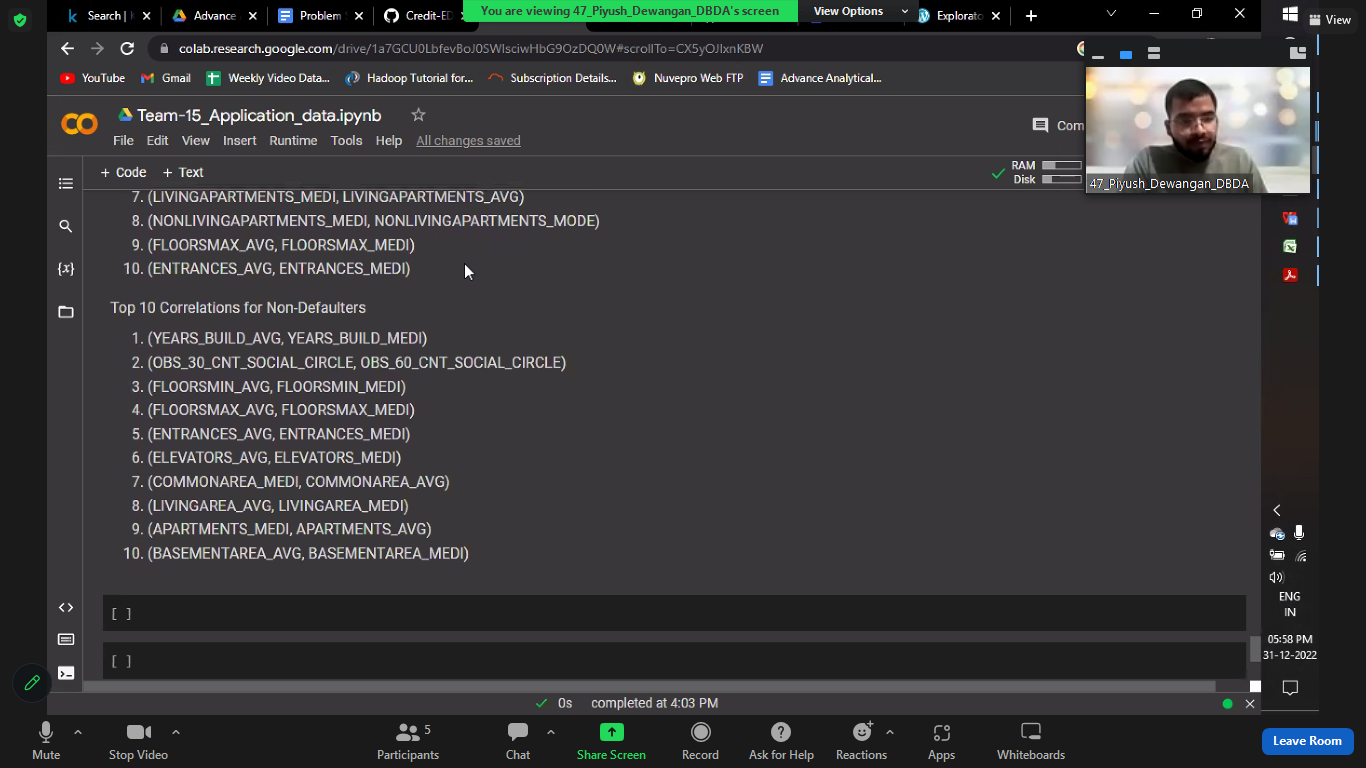
**4. Univariate Analysis, Bivariate and Multivariate Analysis:**

First, I extracted the numerical variables in the dataset and checked out the correlation coefficients with the help of a heatmap.

* ‘DAYS\_LAST\_DUE’ and ‘DAYS\_TERMINATION’ are highly correlated
* ‘DAYS\_FIRST\_DRAWING’ and ‘DAYS\_LAST\_DUE\_1st\_VERSION’ have high negative correlation
* ‘AMT\_ANNUITY’,’AMT\_APPLICATION’,’AMT\_CREDIT’,’AMT\_GOODS\_PRICE’ are highly correlated

**5. Top Ten Correlations:**

**A) Defaulters**

**B) Non-Defaulters**

**5**. **Model Building**:

* When a client applies for a loan, there are four types of decisions that could be taken by the client/company: Approved, Cancelled, Refused and Unused Offer.
* As our target column is a categorical column and our dataset contains different types of categorical and numerical features, so we can choose any type of supervised machine learning algorithm because it is a classification type of a problem statement.
* Random forest classifier and/or Decision tree classifier would work best for this problem statement.

**6. SUMMARY:**

This data is highly imbalanced as number of defaulter is very less in total population.

‘CNT\_FAM\_MEMBERS’, ‘CNT\_CHILDREN’,’NAME\_INCOME\_TYPE’, ‘OCCUPATION\_TYPE’,CODE\_GENDER, ‘EXT\_SOURCE\_1’ and ‘EXT\_SOURCE\_3’ are some of the important driving factors.

1. Documents: Considered features ‘FLAG\_DOCUMENT\_2′,’FLAG\_DOCUMENT\_3′,…,’FLAG\_DOCUMENT\_21’ for this segment. Majority of the applicants did not submit any documents apart from DOCUMENT\_3. FLAG\_DOCUMENT\_3 has similar impact on defaulters and non-defaulters. Hence these columns can be dropped.
2. Housing: All of the features considered have very high (47-70%) missing data percentage. Hence all these features can be dropped. Plot of ‘NAME\_HOUSING\_TYPE’ vs ‘TARGET’ shows that

* Most of the applicants live in House/Apartment.
* Applicants living with their parents or in rented apartment have higher rate of default.

1. Social Circle Info: The features show similar trend for defaulters and non-defaulters, can be dropped.
2. Regional Info: Defaulter rate is highest when REG\_REGION\_NOT\_WORK\_REGION=0 i.e. permanent address and working address is same
3. Contact Info: Considered ‘FLAG\_MOBIL’,’FLAG\_EMP\_PHONE’ etc. for this segment. No impact on Target, features can be dropped.
4. Asset Info: Most of the applicants own realty. Most of the applicants do not own cars iii. People not owning reality and car and have a slightly higher default rate than the people who own reality and car