Bank loan Case study

**Project description**:

This case study aims to illustrate how EDA can be used in a practical corporate setting. We will discover the fundamentals of risk analytics in banking and financial services in this case study, as well as how data is used to lower the risk of losing money when lending to consumers, in addition to using the techniques you acquired in the EDA module.

Three datasets with information about bank loans are provided to us for our final project. Due to their weak or non-existent credit histories, banks that provide loans find it difficult to grant loans to individuals. Due to this, some customers take advantage of it by getting insolvent. Let's say we work for a consumer finance business that specializes in providing urban customers with different kinds of loans. To examine the patterns found in the data, we must use EDA. By doing this, it will be ensured that only those applicants who can repay the loan will be accepted.

Bank wants to understand the driving factors (or driver variables) behind loan default, i.e. the

variables which are strong indicators of default. The bank can utilize this knowledge for its

portfolio and risk assessment.

There are four possible actions the business could take in response to a loan application from a client:

1. Approved: The loan application has been accepted by the business.

2. Cancelled: The client canceled the application during the approval process. Either the customer changed their mind about the loan, or in other situations; a larger risk resulted in the customer receiving an unfavorable price.

3. Rejected: The business turned down the loan (because the client didn't comply with their conditions, etc.).

4. Unused Offer: The client has canceled the loan, albeit it is still in the approval process.

**Datasets:**

1. ‘application\_data.csv' contains all the information of the client at the time of application.

The data is about whether an applicant has payment difficulties.

2. ‘previous\_application.csv’ contains information about the applicant's previous loan data. It

contains the data on whether the previous application had been **Approved, Cancelled, Refused**

**or Unused** offer.

3. columns\_description.csv’ is a data dictionary that elaborates the meaning of the variables.

Approach:

The current and prior applications are two big data sets used in this case study. Each contained many blank data fields as well as several unnecessary columns that were useless for risk evaluations. I cleaned up first. I cleaned the data, found some outliers, and eliminated them before using pivot tables and charts to perform univariate and bivariate analyses to examine this massive data set.

The technology stack used:

The following technology stack was used:

* MySQL Workbench 8.0 CE,
* Microsoft Excel 2010
* Python Programming language
* Google Collab

## Comment:

* prev\_ap\_df (previous\_application.csv’ )contains 37 features and 1670214 rows

(Out of which 15 features are float64, 6 features are integers and 16 features are object datatype)

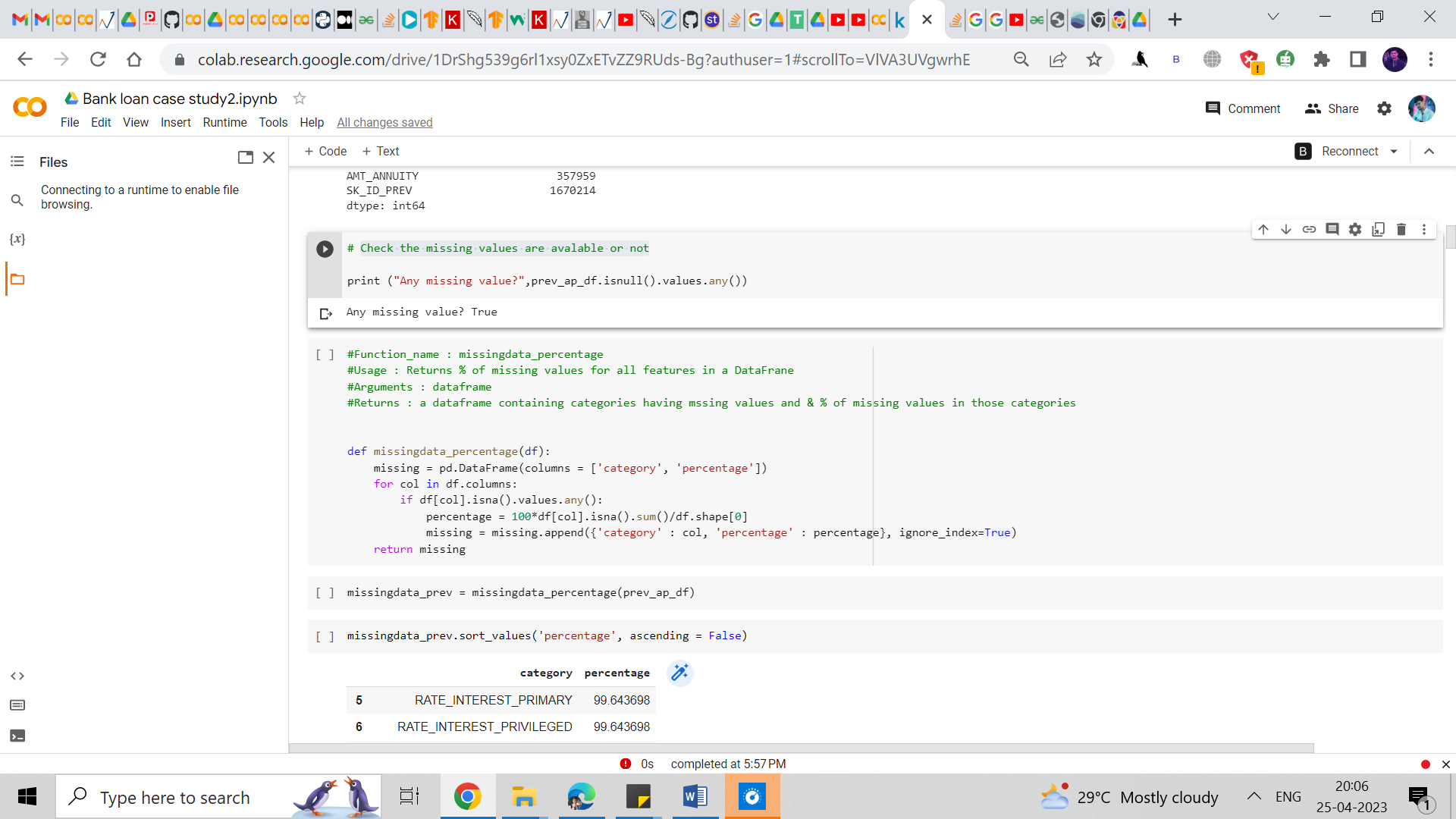
* applications\_df (application\_data.csv) contains 121 features, 1 target variable, and 307511 rows

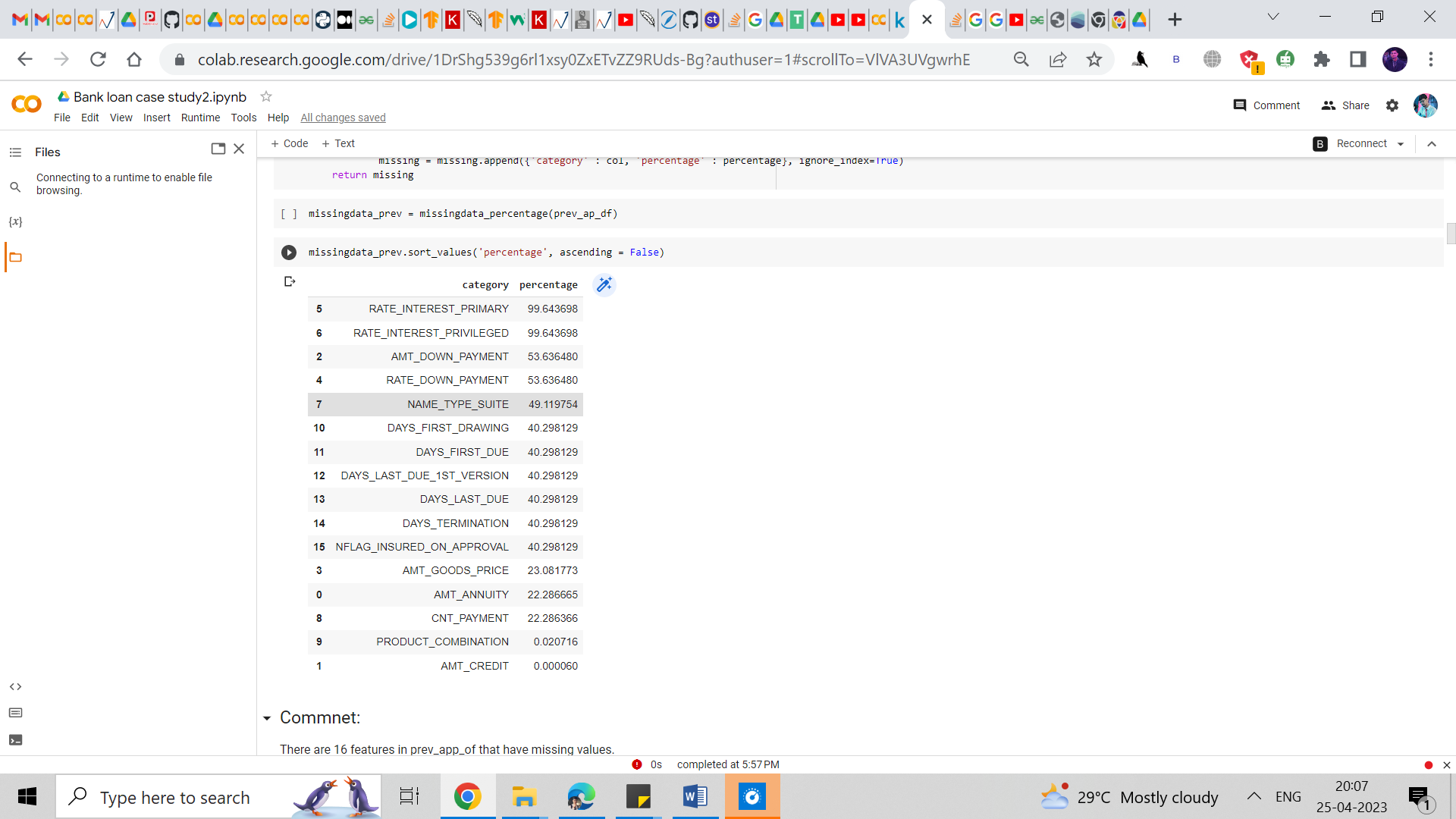
(Out of which 65 features are float64, 41 features are integers and 16 features are object datatype)

* SK\_ID\_CURR" is a unique identifier, which can be used to merge the relevant columns of 2 dataframes.

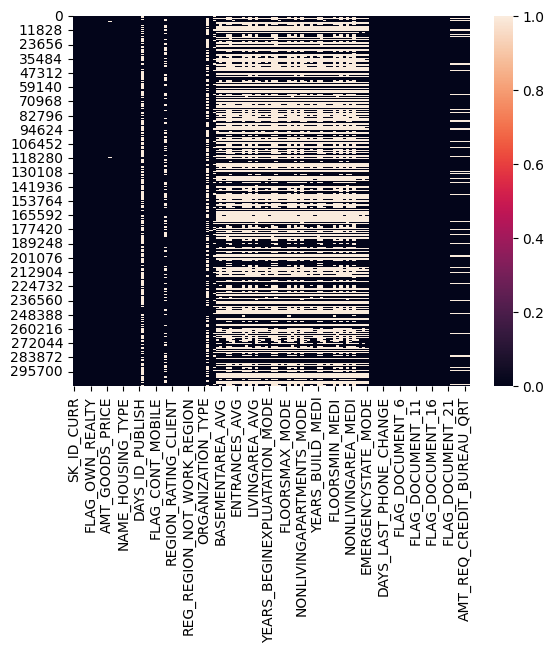
# **Cleaning the data**

First, I Checked whether there are missing values available or not.





Then I checked the percentage of missing values in the given dataset.



## 

## Then I checked the info of the dataset given using the. Info command.

## Comments:

There are 16 features in prev\_app\_of that have missing values.

* Permanently dropping the features (RATE\_INTEREST\_PRIMARY and RATE\_INTEREST\_PRIVILEGED) as 99% data is missing.
* Dropping rows containing missing values for the features (AMT\_CREDIT and PRODUCT\_COMBINATION) for very low % of missing data. Dropping entries would not cause impact the analysis as the percentage of missing values is very low (-2%).

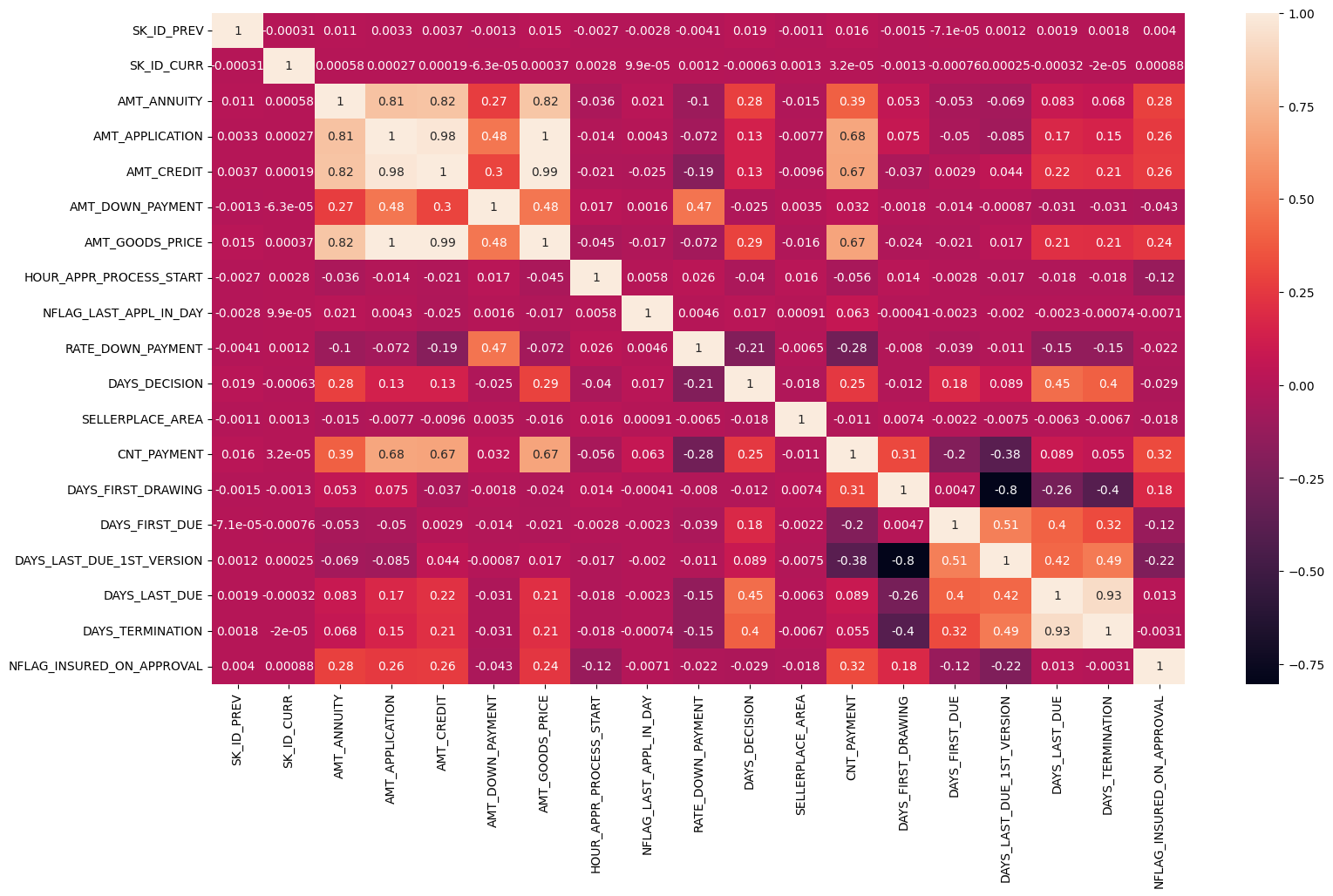
Then I performed correlation studies of the given columns.

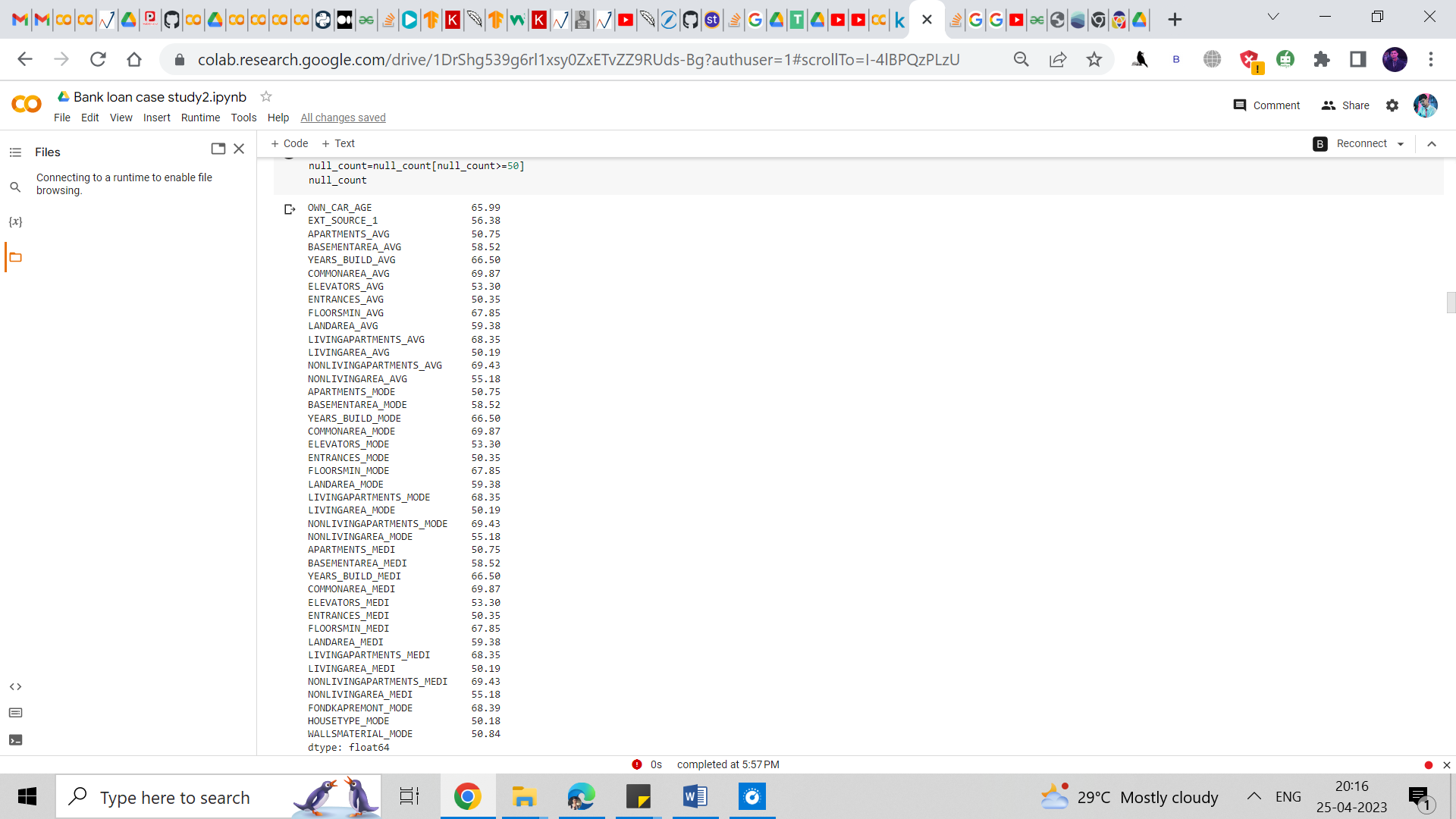
Comment:

* DAYS\_LAST\_DUE and ‘DAYS\_TERMINATION are highty correlated
* DAYS\_FIRST\_DRAWING and DAYS\_LAST\_DUE\_tst\_VERSION have high negative correlation
* 'AMT\_ANNUITY', 'AMT\_APPICATION, 'AMT\_CREDIT', 'AMT\_GOODS\_PRICE' are highly correlated

The features can be removed before modeling ths data. 25 they would cause colinearty "DAYS\_TERMINATION DAYS\_LAST\_DUE\_1st\_VERSION", "AMT\_APPLICATION AMT\_CREDIT","AMT\_GOODS\_PRICE" For EDA purpose we are not removing them.

* SK ID\_PREV columns are not required for analysis.





Then I dropped columns having missing values more than 50%.

Comments:

Notice that there are columns having almost 48%, 13%, 19% etc. missing values. When dealing with columns, you have two simple choices - either delete or retain the column. If you retain the column, you'll have to treat (i.e. delete or impute) the rows having missing values.

If you delete the missing rows, you lose data. If you impute, you introduce bias.

Apart from the number of missing values, the decision to delete or retain a variable depends on various other factors, such as:

The analysis task at hand, The usefulness of the variable (based on your understanding of the problem), The total size of available data (if you have enough, you can afford to throw away some of it) etc. Thus, for this exercise, let's remove the columns having more than missing values and which are not necessary for our analysis.

We Drop unnecessary columns from the dataset

## Handling Outliers

Major approaches to the treat outliers:

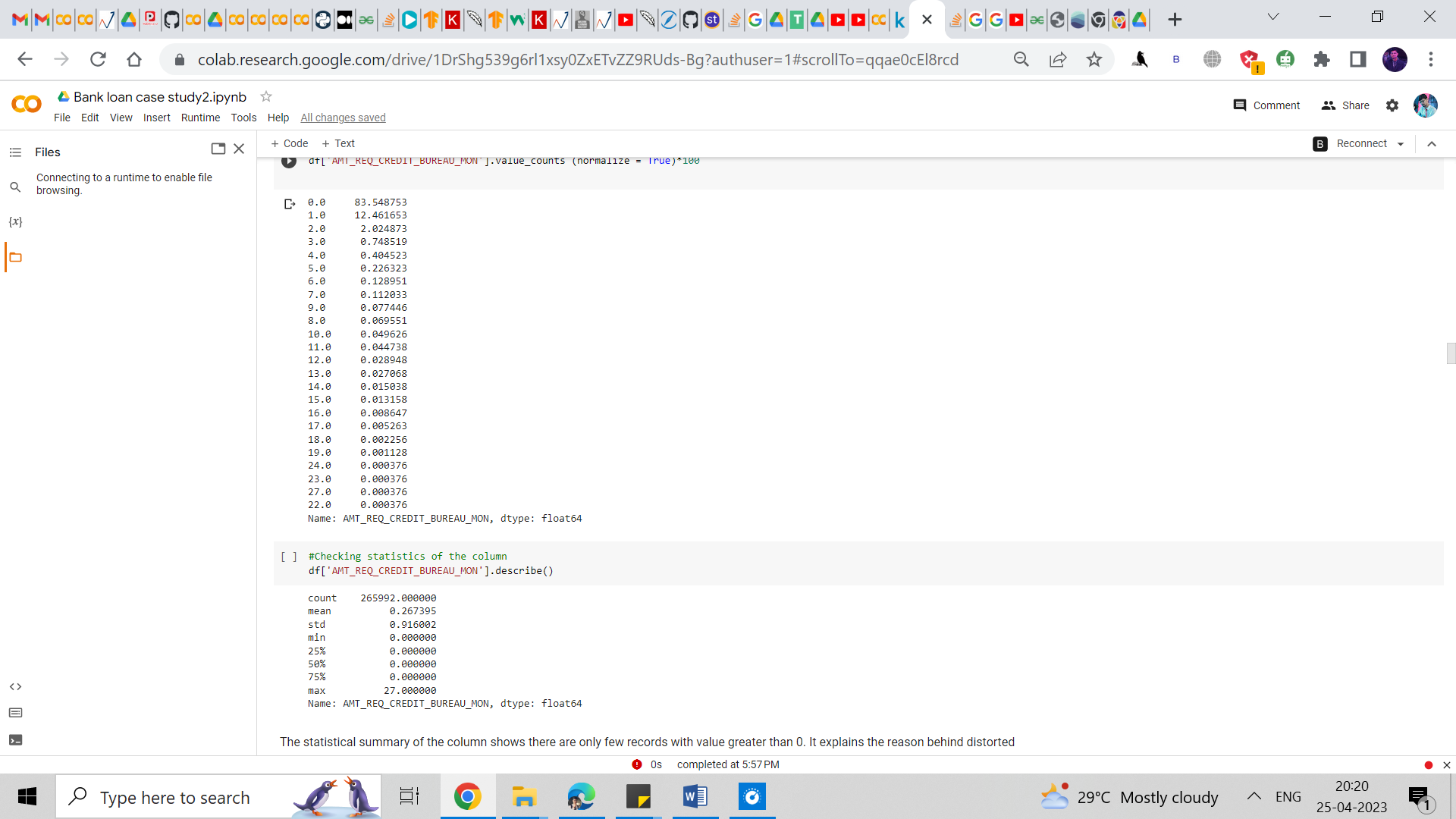
* **Imputation**
* **Deletion of outliers**
* **Binning of values**
* **Cap the outlier**

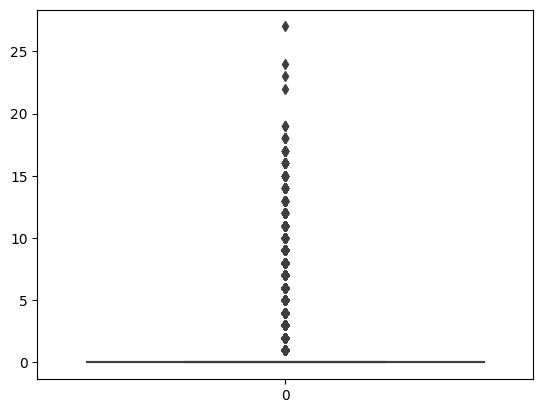
## Imputation technique for with ~ 13% of missing values

The target columns for this analysis will be

* AMT\_REQ\_CREDIT\_BUREAU\_MON
* AMT\_REQ\_CREDIT\_BUREAU\_WEEK
* AMT\_REQ\_CREDIT\_BUREAU\_DAY
* AMT\_REQ\_CREDIT\_BUREAU\_HOUR
* AMT\_REQ\_CREDIT\_BUREAU\_QRT
* AMT\_REQ\_CREDIT\_BUREAU\_YEAR

Identifying imputation technique for AMT\_REQ\_CREDIT\_BUREAU\_MON



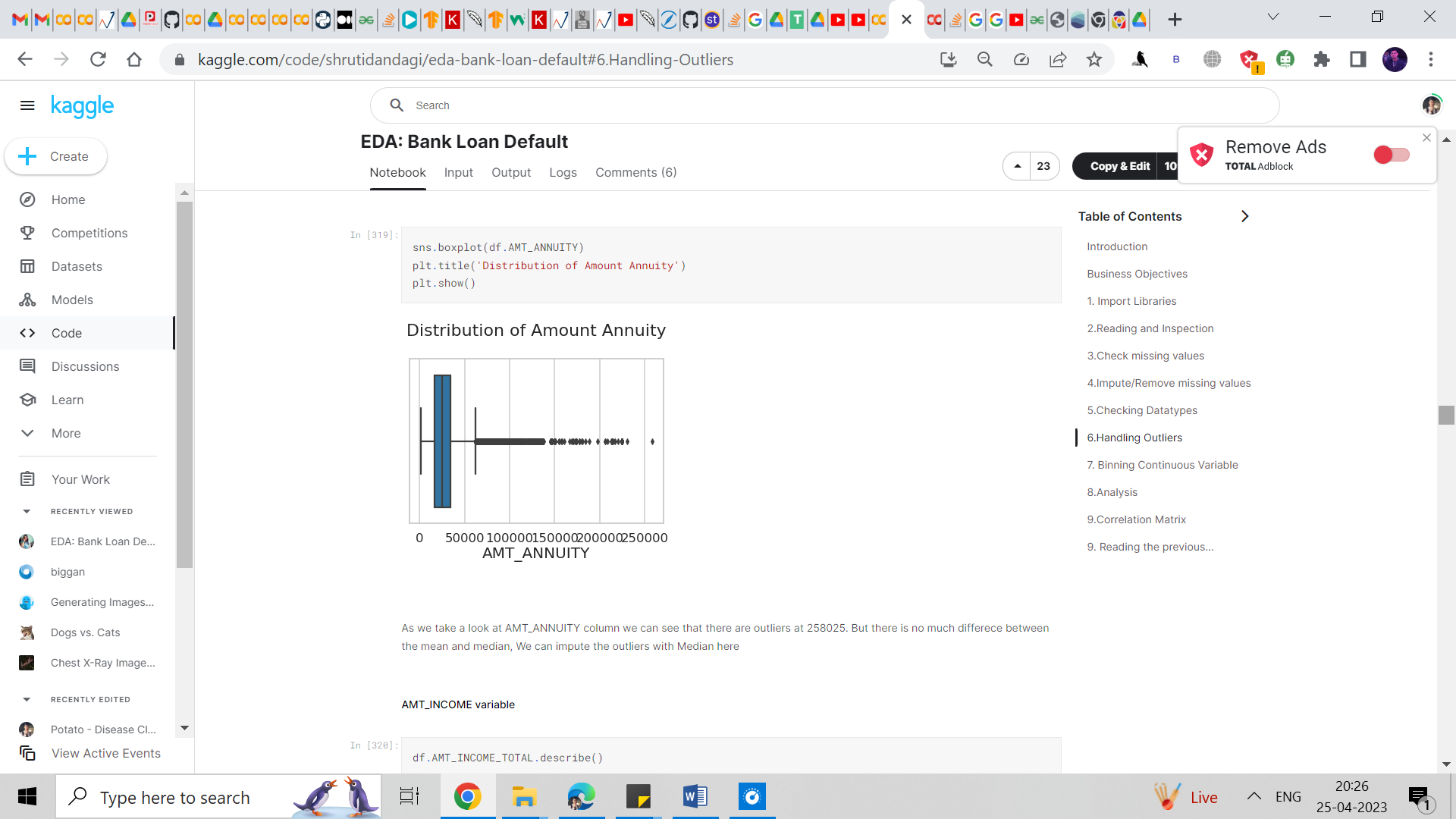


Comments:

The statistical summary of the column shows there are only few records with value greater than 0. It explains the reason behind distorted boxplot.

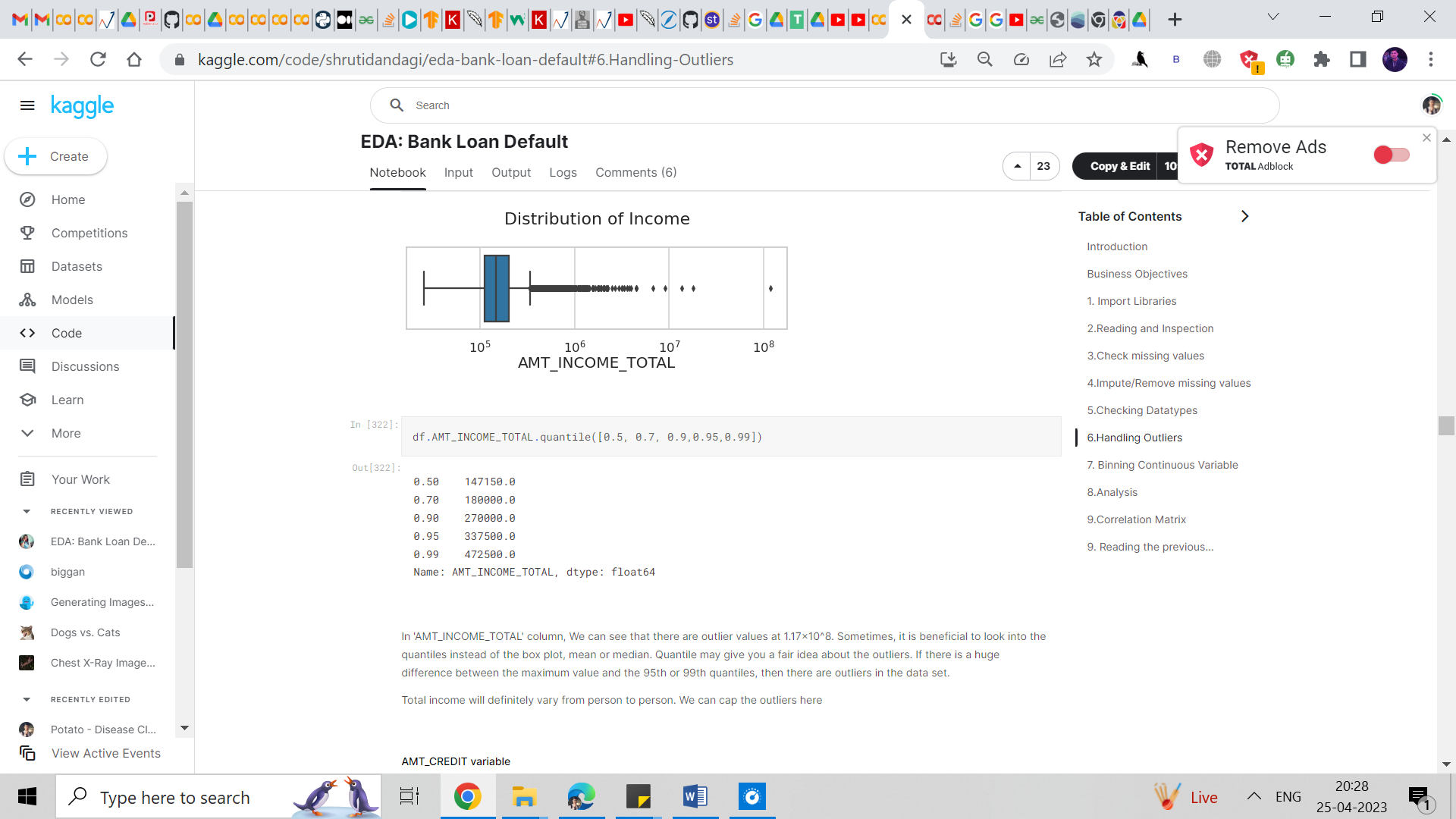
For column AMT\_REQ\_CREDIT\_BUREAU\_MON, we have two approaches either exclude missing values or impute the column with value 0 which is present in more than 83% of the rows. Hence, the recommended imputation technique is replacing null by the mode which is 0.

For the rest of the columns as well we followed the same technique to impute or restore values as they showed the same pattern as above.



As we take a look at the AMT\_ANNUITY column we can see that there are outliers at 258025. But there is not much difference between the mean and median, We can impute the outliers with Median here.

#### **AMT\_INCOME variable**

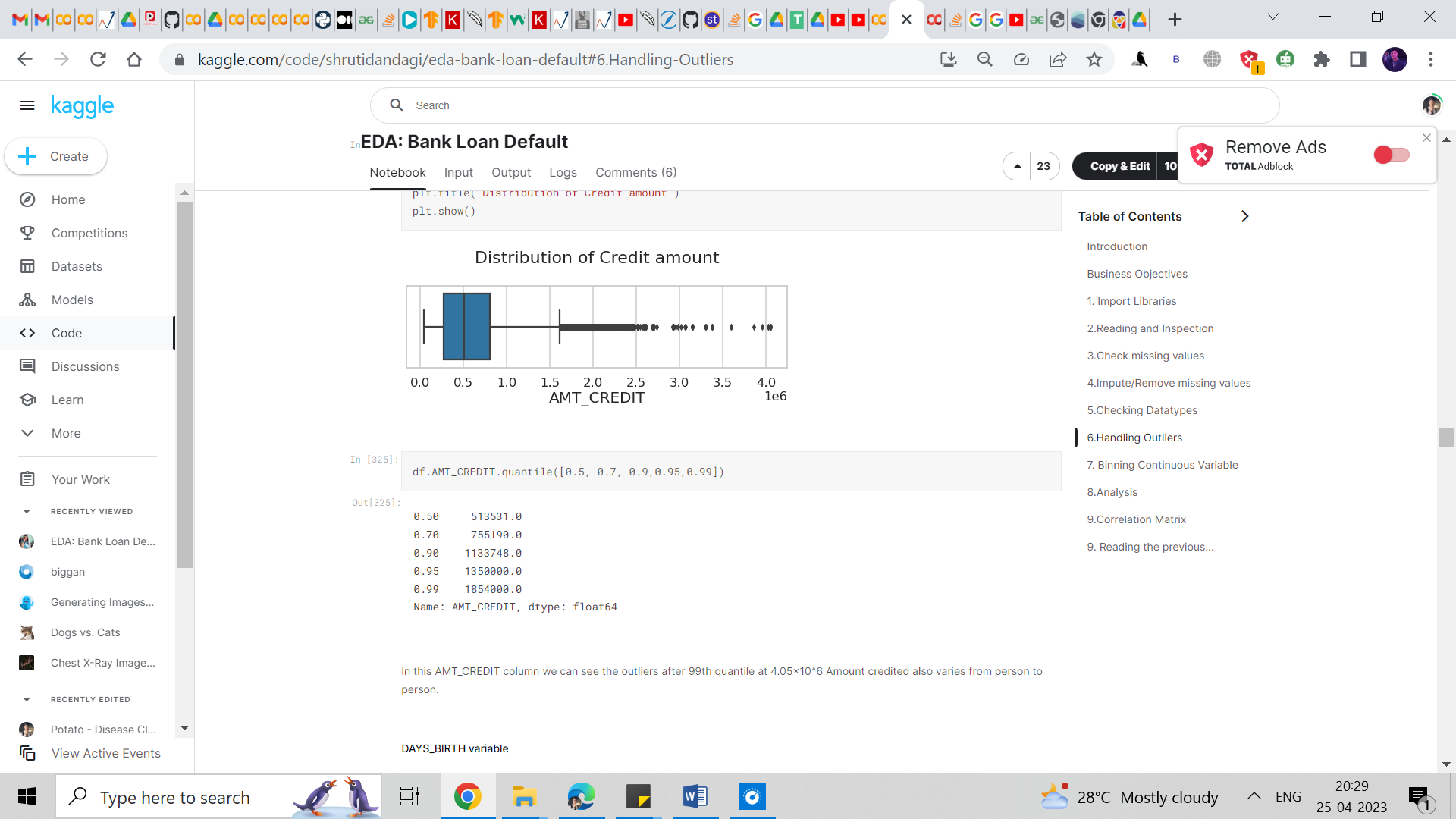


Comment:

In the 'AMT\_INCOME\_TOTAL' column, We can see that there are outlier values at 1.17x10^8. Sometimes, it is beneficial to look into the quantiles instead of the box plot, mean, or median. Quantile may give you a fair idea about the outliers. If there is a huge difference between the maximum value and the 95th or 99th quantiles, then there are outliers in the data set.

Total income will vary from person to person. We can cap the outliers here

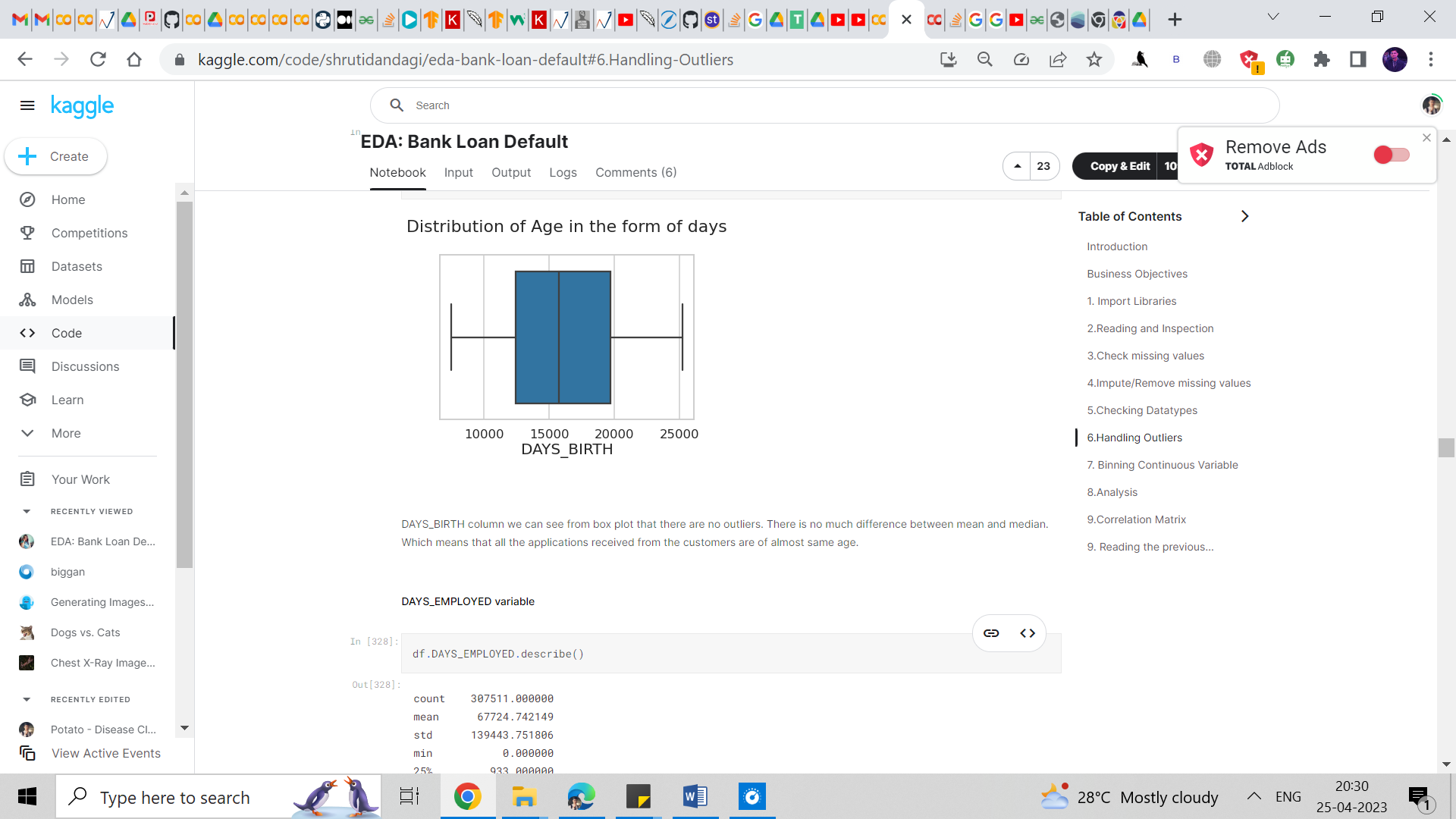
#### **AMT\_CREDIT variable**



Comment:

In this AMT\_CREDIT column we can see the outliers after 99th quantile at 4.05x10^6 Amount credited also varies from person to person.

#### **DAYS\_BIRTH variable**



DAYS\_BIRTH column we can see from the box plot that there are no outliers. There is not much difference between mean and median. This means that all the applications received from the customers are of almost the same age.

#### **DAYS\_EMPLOYED variable**

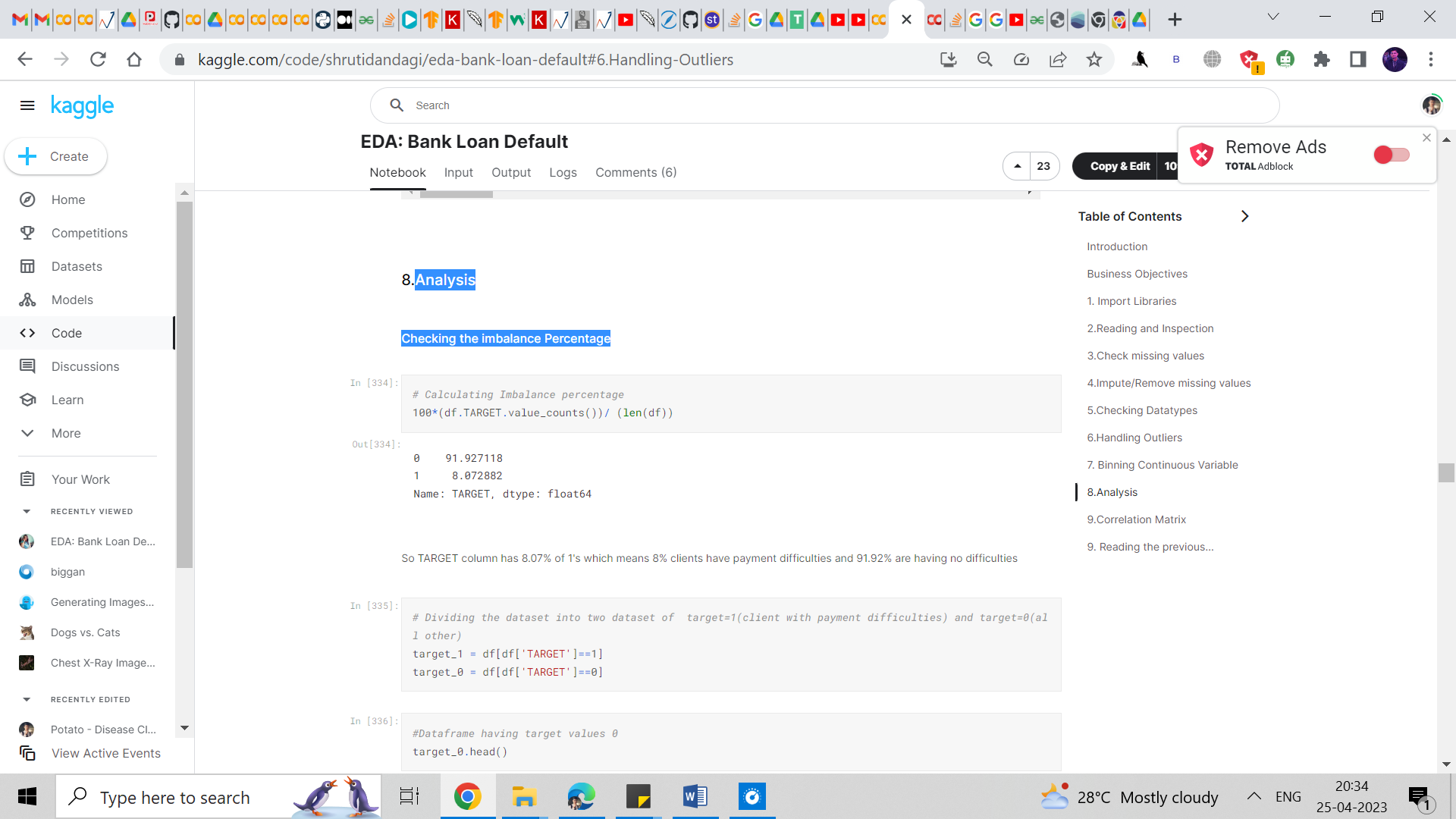


Comment:

DAYS\_EMPLOYED column has outliers at 365243. The number of days the person was employed varies from person to person

## Analysis

### **Checking the imbalance Percentage**



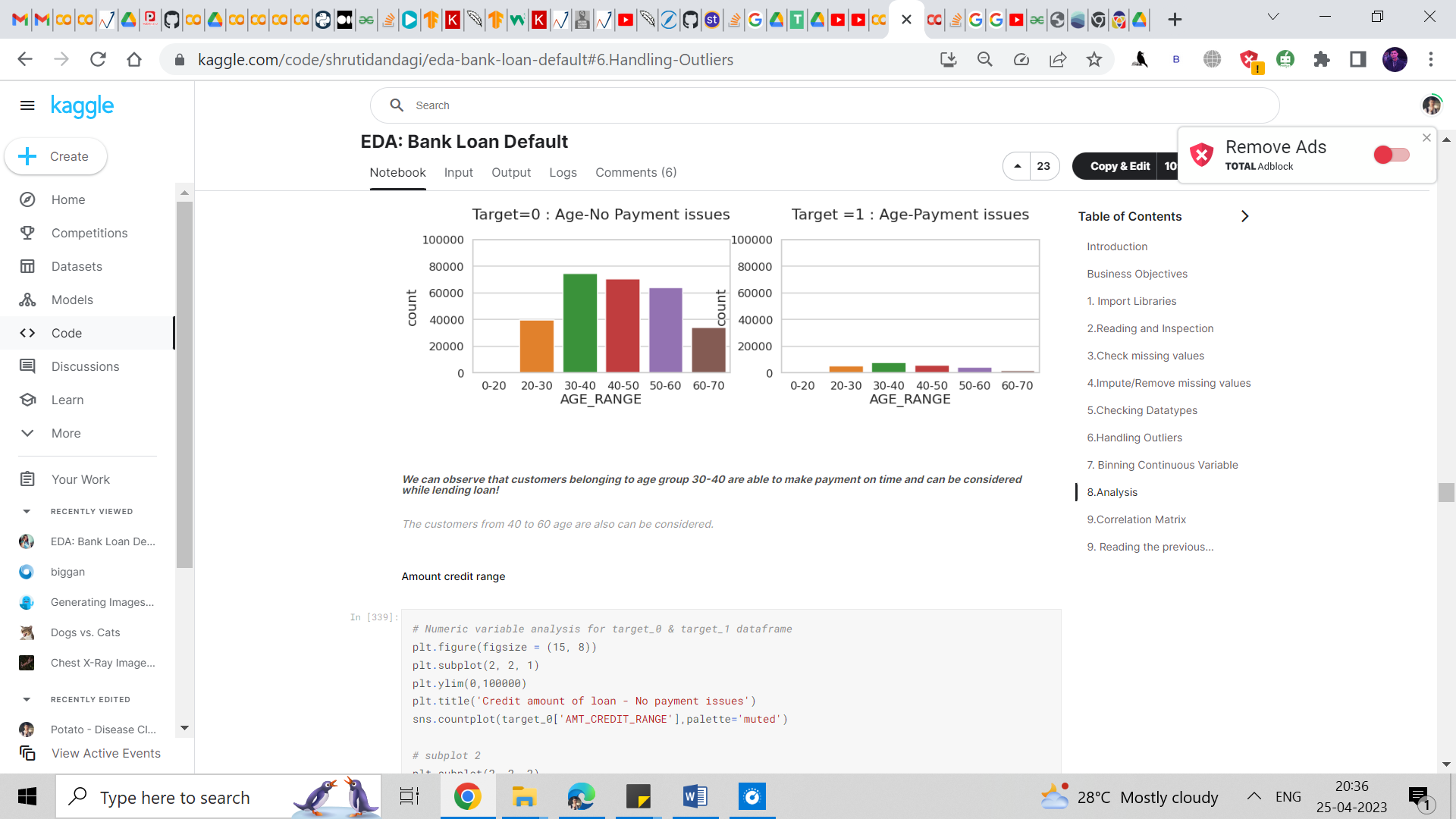
Comment:

So the TARGET column has 8.07% of 1's which means 8% of clients have payment difficulties and 91.92% are having no difficulties

### **Univariate Analysis for target =0 and target=1**

### **Numeric variable**

#### **Age**

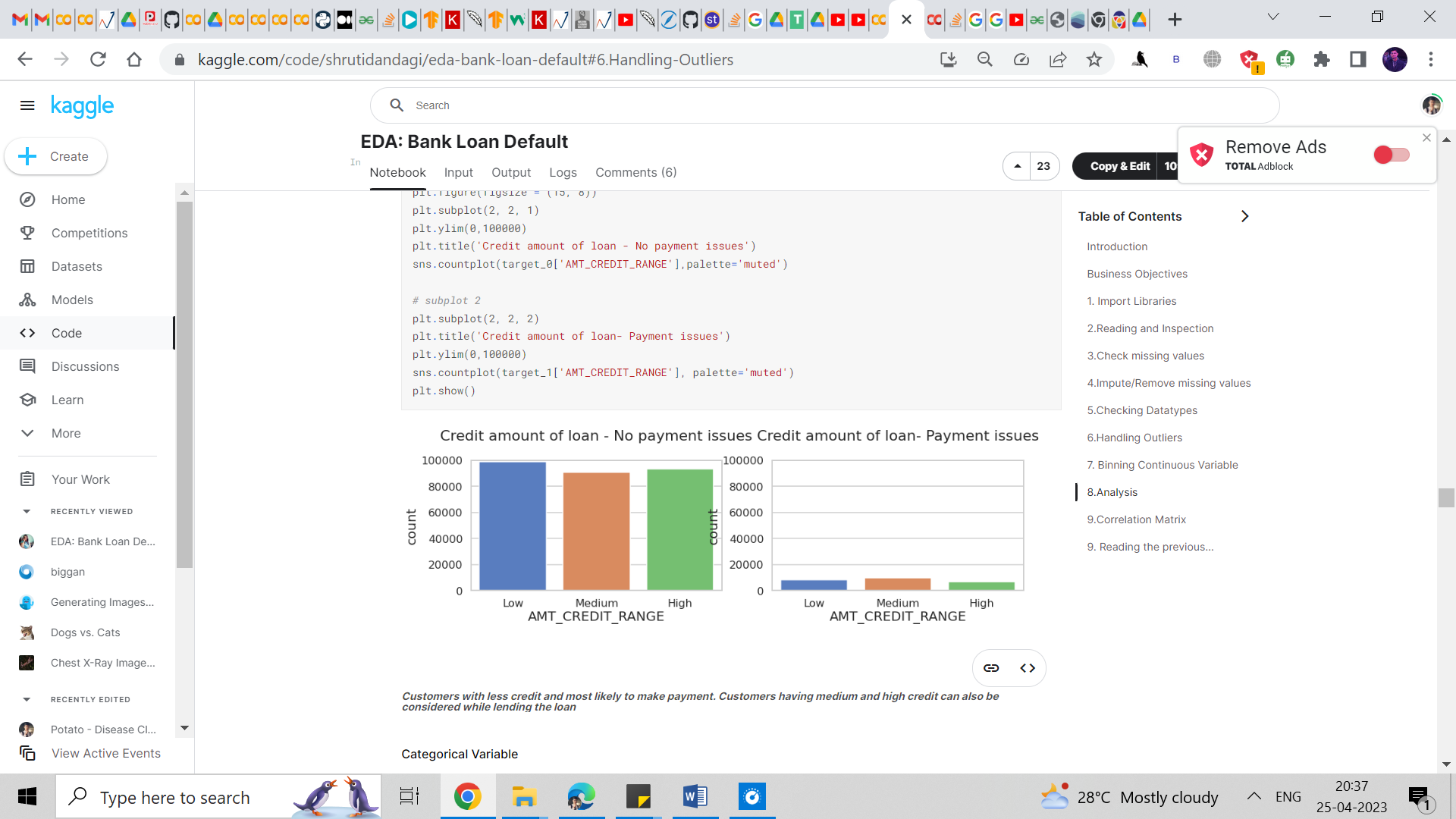


Comment:

###### *We can observe that customers belonging to the age group 30-40 are able to make payment on time and can be considered while lending a loan!*

##### ***Customers from 40 to 60 age are also can be considered.***

#### **Amount credit range**

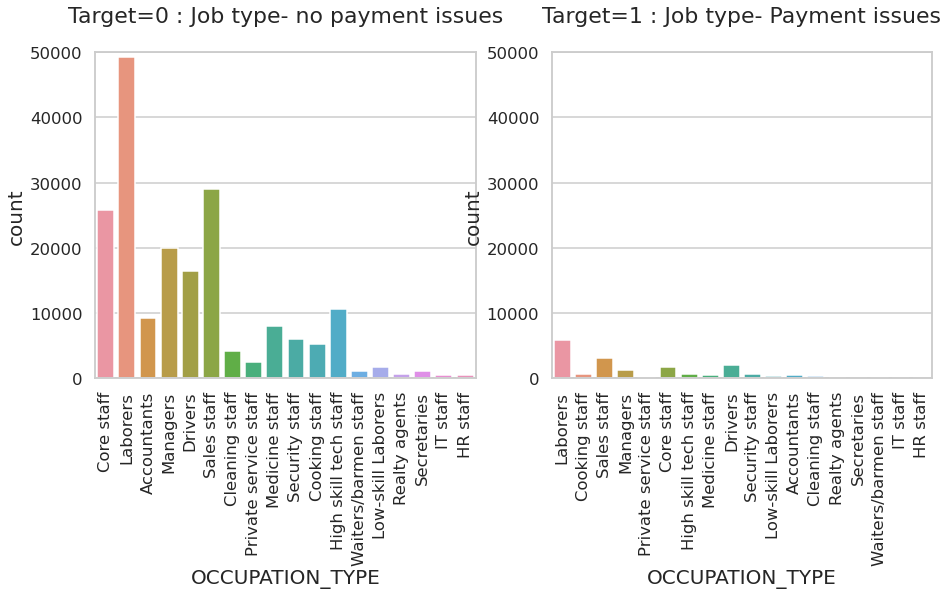


Comment:

***Customers with less credit and most likely to make payments. Customers having medium and high credit can also be considered while lending the loan***.

### **Categorical Variable**

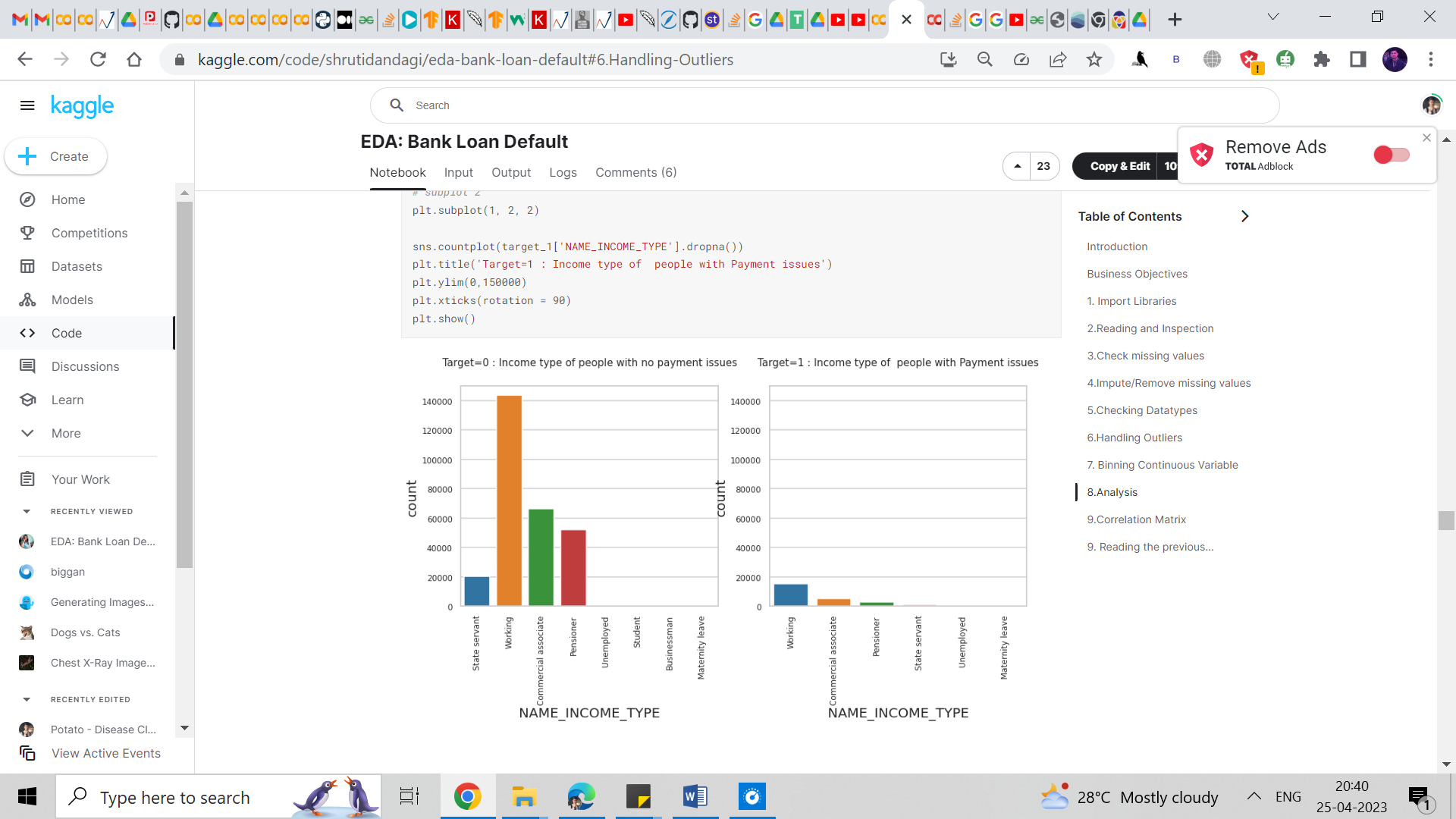
#### **Occupation\_type**



Comment:

##### ***The plot clearly shows that labourers are most likely to make payments on time whereas HR staff are less likely to make payments on time***

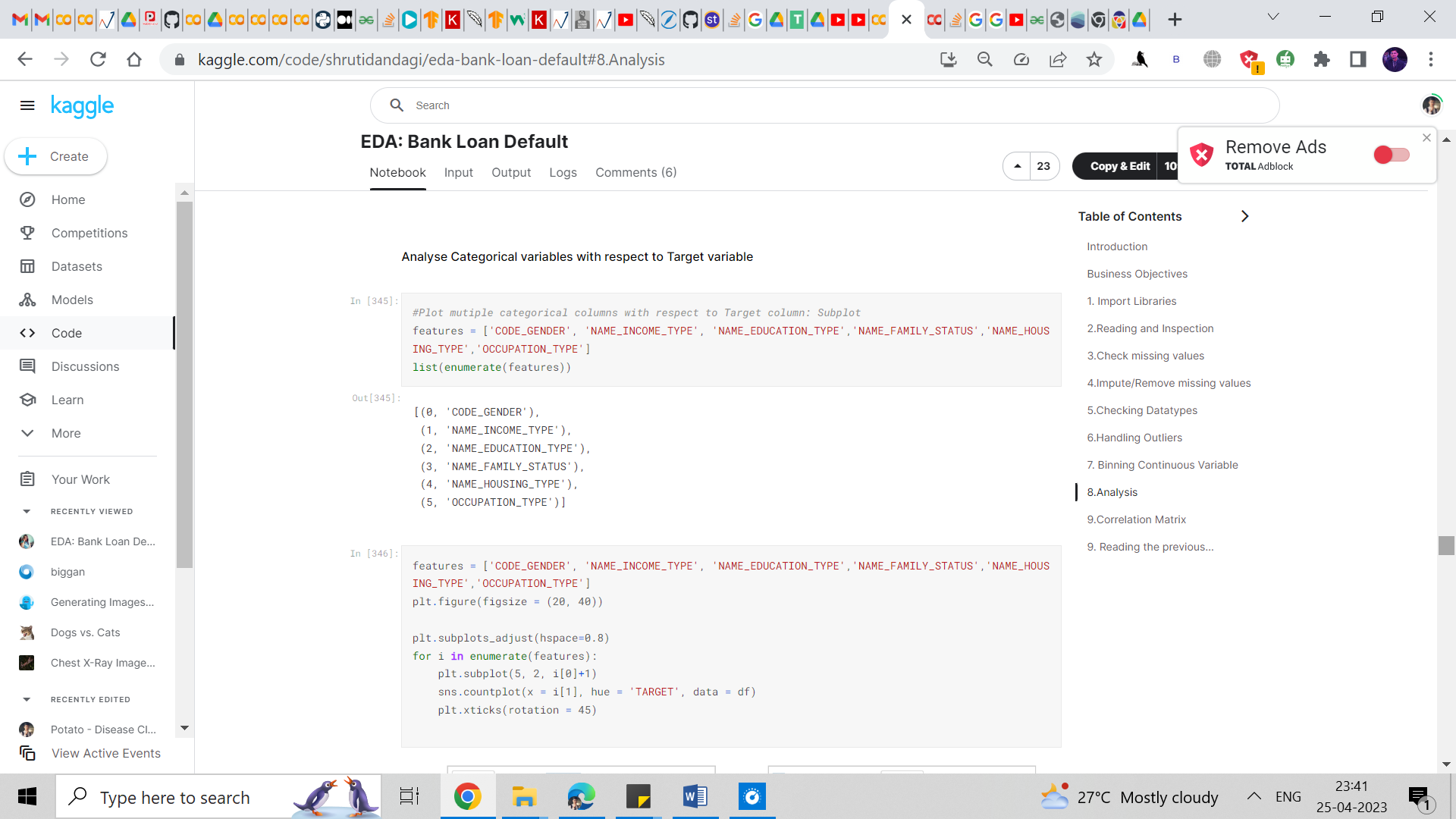
#### **Name \_Income \_Type**



Comment:

##### ***The plot clearly shows that labourers are most likely to make payment on time whereas HR staff are less likely to make payment on time.***

### **Analyse Categorical variables with respect to the Target variable**

We tried to plot multiple categorical columns with respect to Target Column: Subplot

##### ***Comment:***

##### ***Comment:***

##### ***From the above plot, we can see that,***

##### ***Female customers pay loan amounts on time and banks can target more female customers for lending loans.***

##### ***Working customers can be targeted to lend loans as they have a higher percentage of making payments on time.***

##### ***Customers with secondary education are most likely to make payments when compared to customers with academic degrees.***

##### ***Married customers have paid loan amounts on time when compared to widows.***

##### ***Customers owning houses/apartments are most likely to make payments on time compared to those living in CO-OP apartments.***

##### ***Laborers have a high repayment percentage. Hence banks can think of lending small amounts of loans to them.***

## Correlation Matrix

To get rid of the repeated correlation values between two variables we perform the following steps.



##### ***We saw that for the Target\_0 data frame, Social circle for 30 days and 60 days are most correlated, and Goods price and Loan amount credit are highly correlated. Then we have Goods price and amount annuity in 4th place***

Comment:

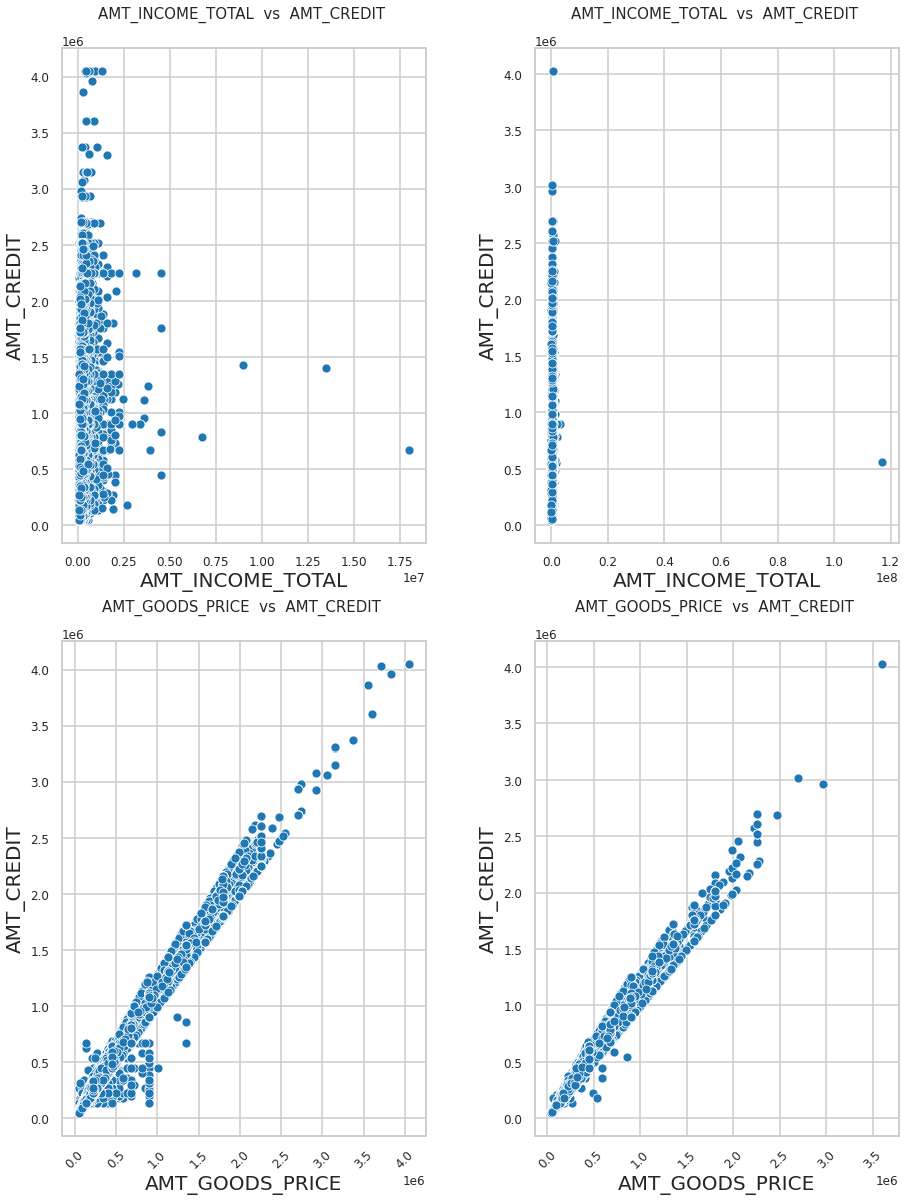
***From the observation above we can say that for the target\_1 data frame Goods price and loan credit amount are most correlated next to social circle observations for different days. So, the variables correlated in the target\_0 data frame and target\_1 data frame are same with slightly varying correlation values***

### **Bivariate Analysis for Target 0 and Target 1**

##### ***Numeric-Numeric Analysis***

There are three ways to analyze the numeric-numeric data types simultaneously. Scatter plot: describes the pattern that how one variable is varying with another variable. Correlation matrix: to describe the linearity of two numeric variables. Pair plot: a group of scatter plots of all numeric variables in the data frame

### **Income vs Credit, Goods price vs Credit**



Comment:

***Those who have paid the loan amount on/within time are more likely to get higher credits than those who didn’t pay/did late payments.***

***People who have higher goods prices and have made payments on time have higher credits than those with higher goods price but didn’t pay loan.***

### **Numerical categorical analysis**

### **Income range- Gender**



Comment:

##### ***We can see that Females with low income don’t have any payment issues.***

#### **Credit amount vs Education Status**

##### ***Comment:***

##### ***From the above plot, we can see that***

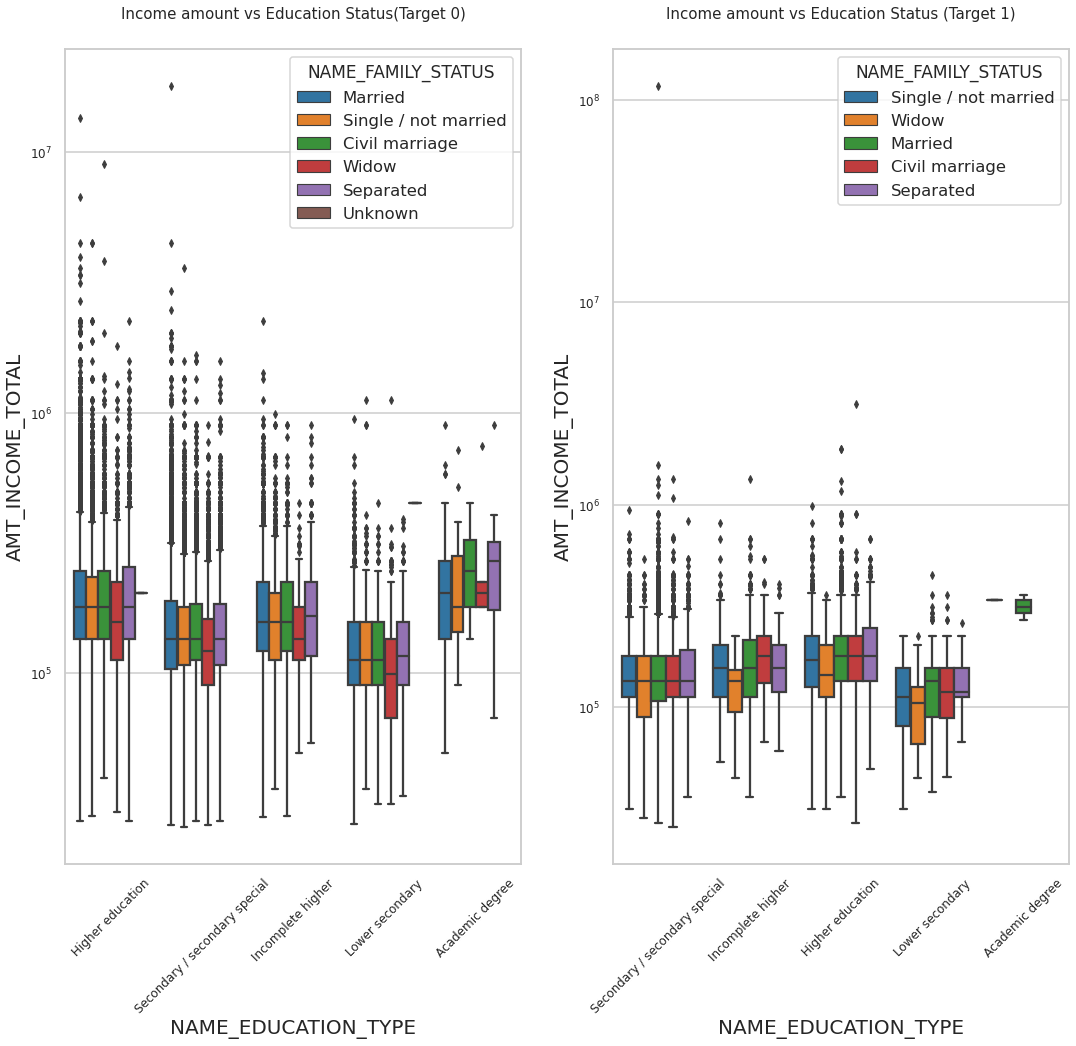
##### ***1. Some of the highly educated, married person is having credits higher than those who have done lower secondary education.***

##### ***2. Those with higher education have higher credits and are more likely to make payments on time.***

##### ***3. More outliers are seen in higher education.***

##### ***4. people with secondary and secondary special education are less likely to make payments on time.***

### **Income vs Education Status**



Comment:

##### ***From the above plots,***

##### ***1. we can see that Higher education has many outliers.***

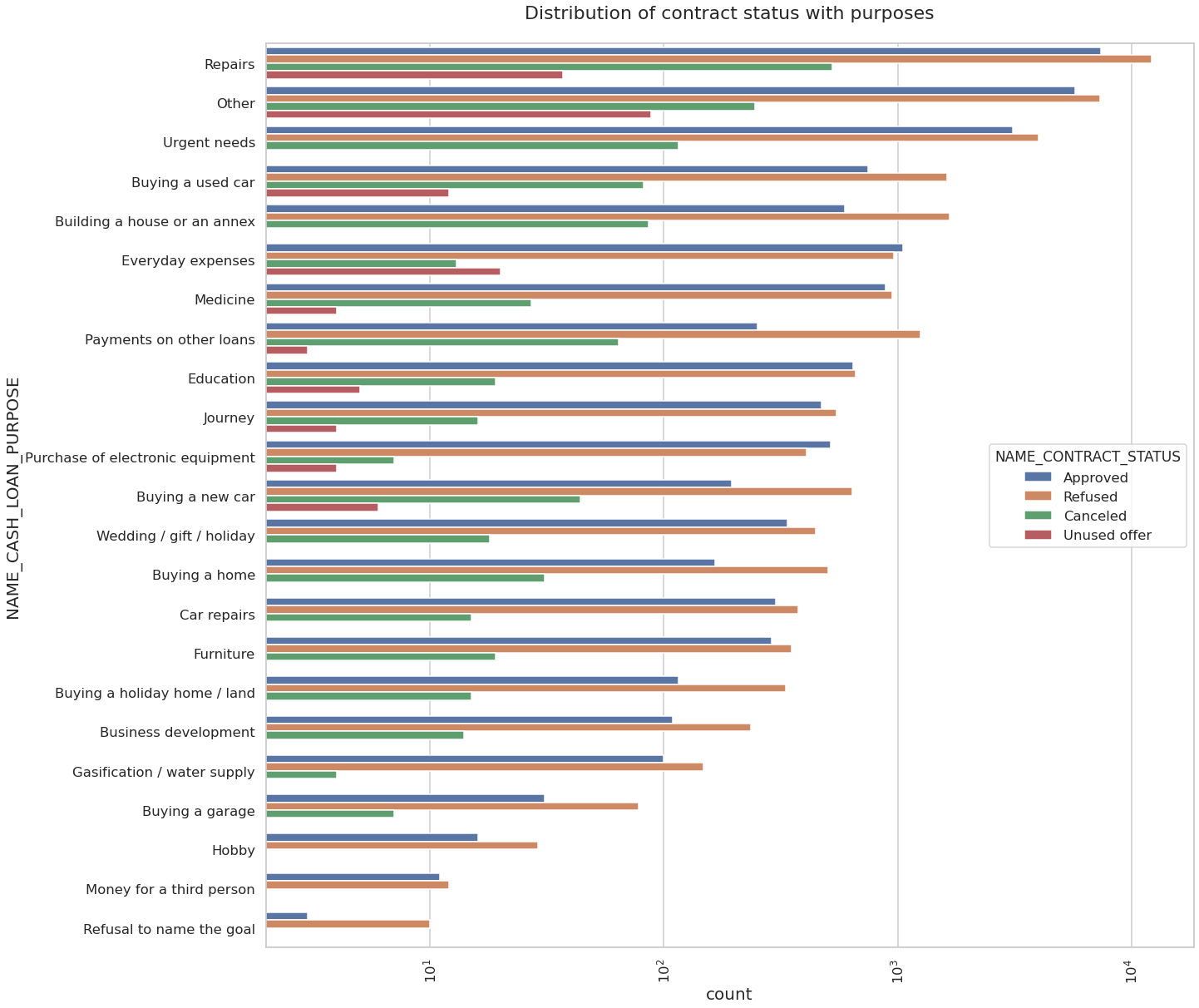
##### ***2. People with higher education have higher incomes and don’t have difficulties in making loan payments.***

##### ***3. People with higher education who have lesser income are unable to pay the loan.***

##### ***Hence, we can conclude that people with Higher incomes are most likely to make payments.***

Univariate Analysis:

*Distribution of contract status in a logarithmic scale*

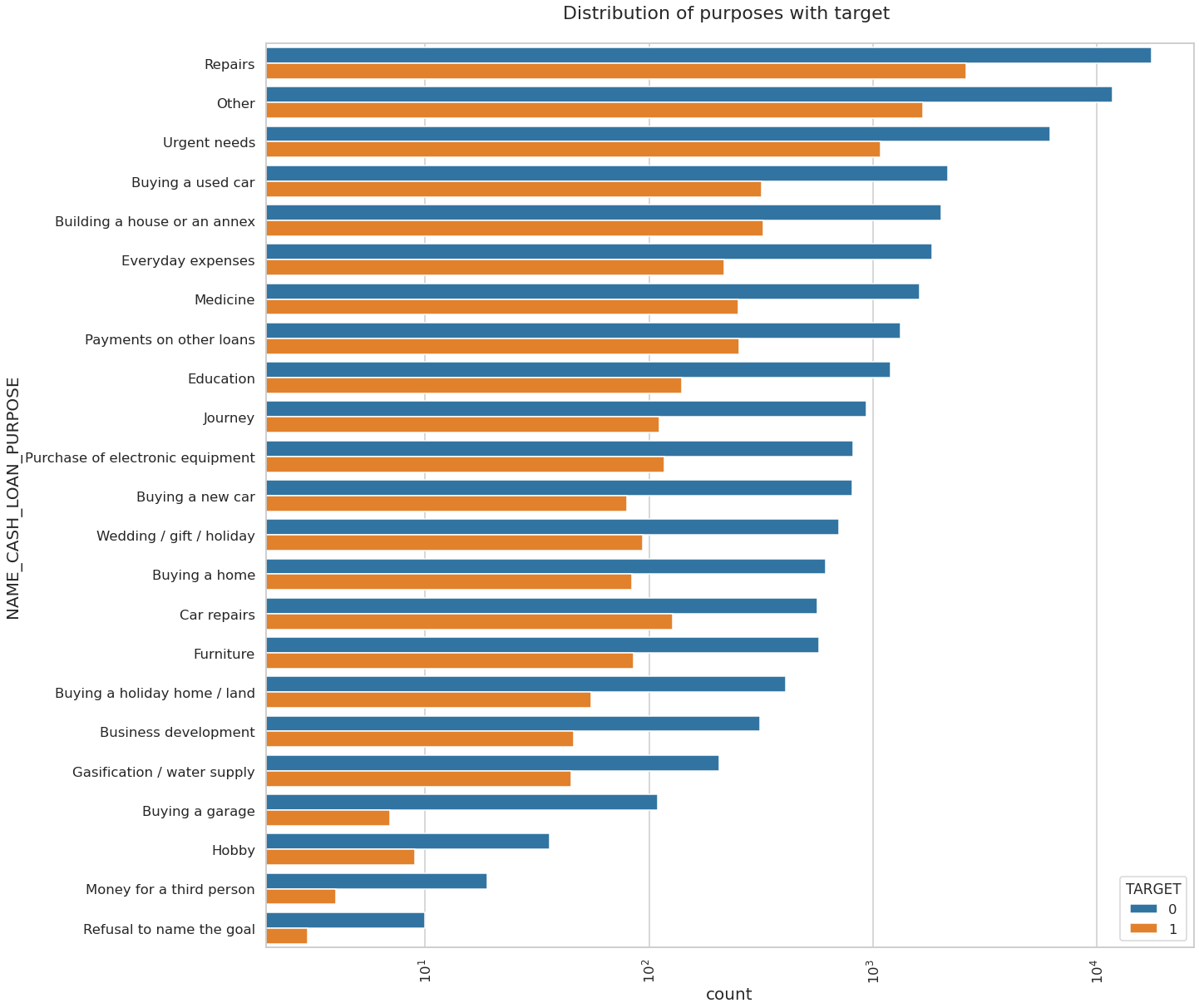


Comments:

###### *Points to be concluded from the above plot:*

Most rejections of loans came from the purpose of 'Repairs'. For education purposes, we have an equal number of approved and rejection Paying’s other loans and buying a new car is having significant higher rejections than approves.

*Distribution of contract status*



Comments:

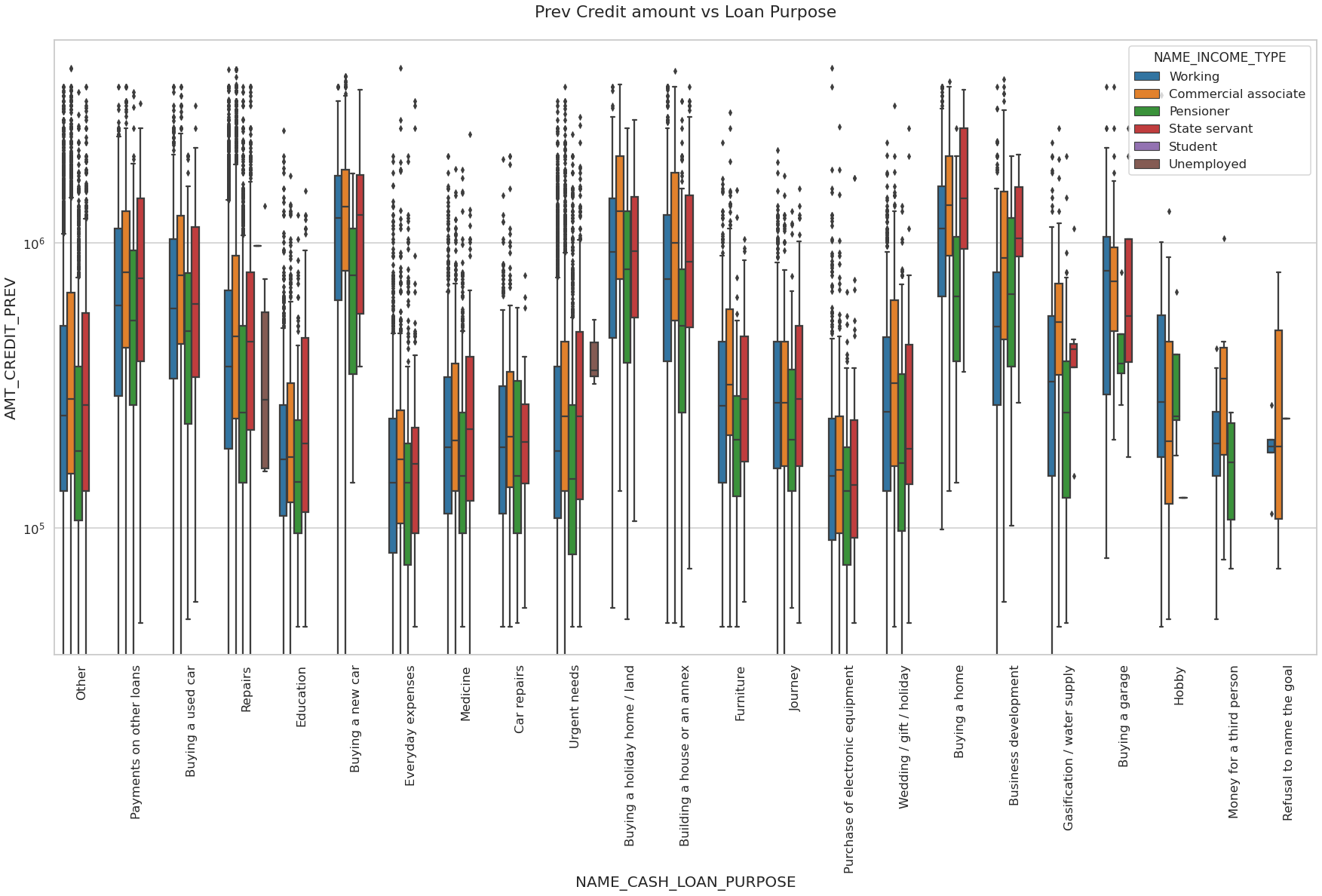
###### *Few points we can conclude from the above plot:*

Loan purposes with 'Repairs' are facing more difficulties in payment on time. There are few places where loan payment is significantly higher than facing difficulties. They are 'Buying a garage', 'Business development', 'Buying land', 'Buying a new car’, and 'Education' Hence we can focus on these purposes for which the client is having minimal payment difficulties

### **Bivariate Analysis**

#### **Prev Credit amount vs Loan Purpose**

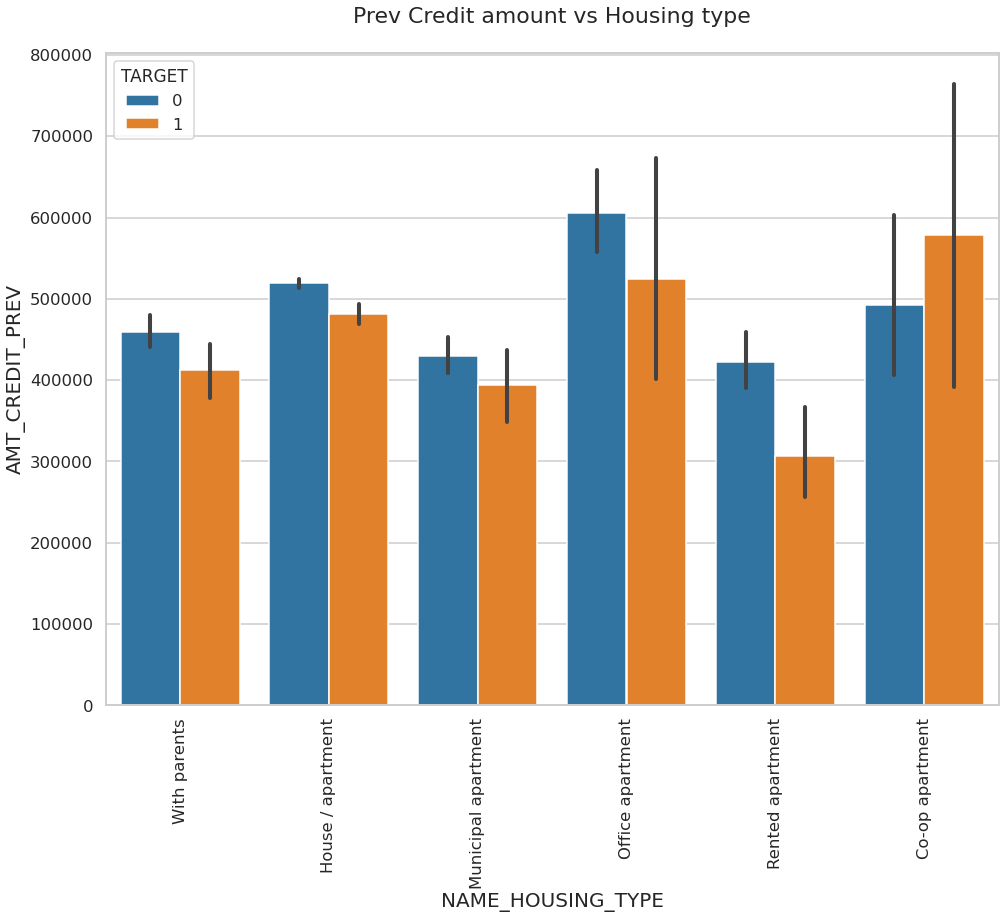
*Box plotting for Credit amount in logarithmic scale*



Comments:

From the above we can conclude some points- The credit amount for Loan purposes like 'Buying a home',' Buying land',' Buying a new car', and building a house' is higher. The income type of state servants has a significant amount of credit applied Money for the third person or a Hobby is having fewer credits applied.

*Box plotting for Credit amount prev vs Housing type in logarithmic scale*



Comments:

##### ***Here for Housing type, office apartment is having higher credit of target 0 and co-op apartment is having higher credit of target 1. So, we can conclude that bank should avoid giving loans to the housing type of co-op apartment as they are having difficulties in payment. Bank can focus mostly on housing type with parents or House\apartment or municipal apartment for successful payments.***

#### **CONCLUSION**

##### ***1. Banks should focus more on contract types ‘Student’,’ pensioner’, and ‘Businessman’ with housing ‘types other than ‘Co-op apartment’ for successful payments.***

##### ***2. Banks should focus less on income type ‘Working’ as they are having the most number of unsuccessful payments.***

##### ***3. Also with loan purposes ‘Repair’ is having a higher number of unsuccessful payments on time.***

##### ***4. Get as many clients from the housing type ‘With parents’ as they are having the least number of unsuccessful payments.***

**Results:**

After running all the formulas in Microsoft Excel and plotting the charts we analyzed all the data sets of this Project.

In the making of this report, we used our Microsoft Excel knowledge and Python as a real-world examples.

**Drive Link**

[**https://drive.google.com/drive/folders/1ppGHWEgDT\_QgoCvrFqdCa6\_BPZLL7UNu?usp=sharing**](https://drive.google.com/drive/folders/1ppGHWEgDT_QgoCvrFqdCa6_BPZLL7UNu?usp=sharing)