**REPORT OF DATA EXPLORATION**

**Course Code: BA723**

**Student Name: Charles Ntamack - 301209795**

**Student Name: Samson Ogwuche - 301196569**

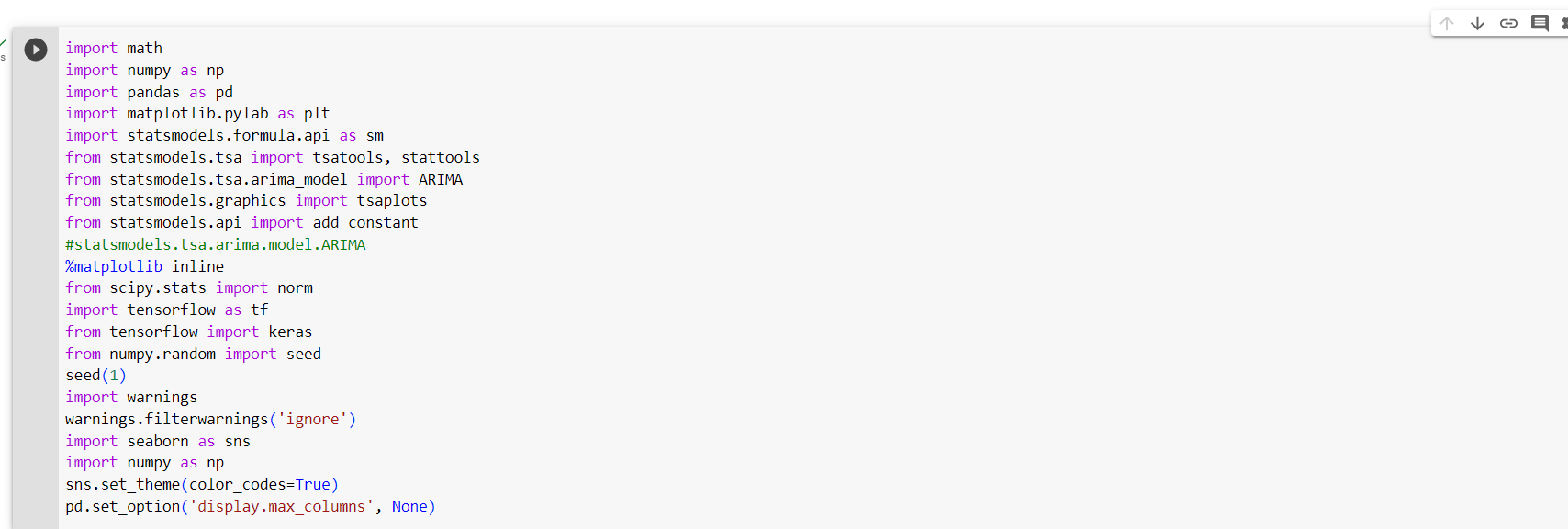
**Data Exploration Steps**

Our project is about predicting credit default by customers.

The main objective of this report is to outline the step-by-step data preprocessing and feature engineering, exploratory data analysis (EDA), and all the stages of exploratory data analysis adopted in this project which are essential in any Capstone project. The data exploration analysis is essentially the next step after data collection. The steps are not limited to our screenshots or visualizations.

1. **Imported the required packages**

The necessary libraries required for the preprocessing have been imported as below.



1. **Data Preprocessing and Feature Engineering**

In our data preprocessing stage, we loaded the dataset, checked the features, and data structure/shape. The dataset has 30,000 observations and 25 variables/features. From the dataset, 6,636 customers defaulted on their credit payment and 23,364 did not default. We checked for missing values, irrelevant/redundant, duplicates variables, and object variables. There were no missing values, but we had four object variables in the dataset. We noticed ID is not relevant to the model and so we dropped ID. We also changed the object variables into categorical variables. We renamed ‘default’ from lowercase to uppercase to align with other variables.

Below is the list of variables in the dataset and the data dictionary. We also have screenshots of information from the dataset and the preprocessing that we carried out.

Definition of variable in our dataset:

**ID:** The key uniquely identifies each customer.

**LIMIT\_BAL:** This is the maximum balance approved for each customer by the bank.

**SEX:** This is the gender of the customer. The sex in the datasets F (female) and M (male).

**EDUCATION:** This is the current level of education of the customer. Educational levels of customer high school, university, graduate, others, and unknown.

**MARRIAGE:** This is the marital status of the customer. Marital categories of customers are married, single, and others.

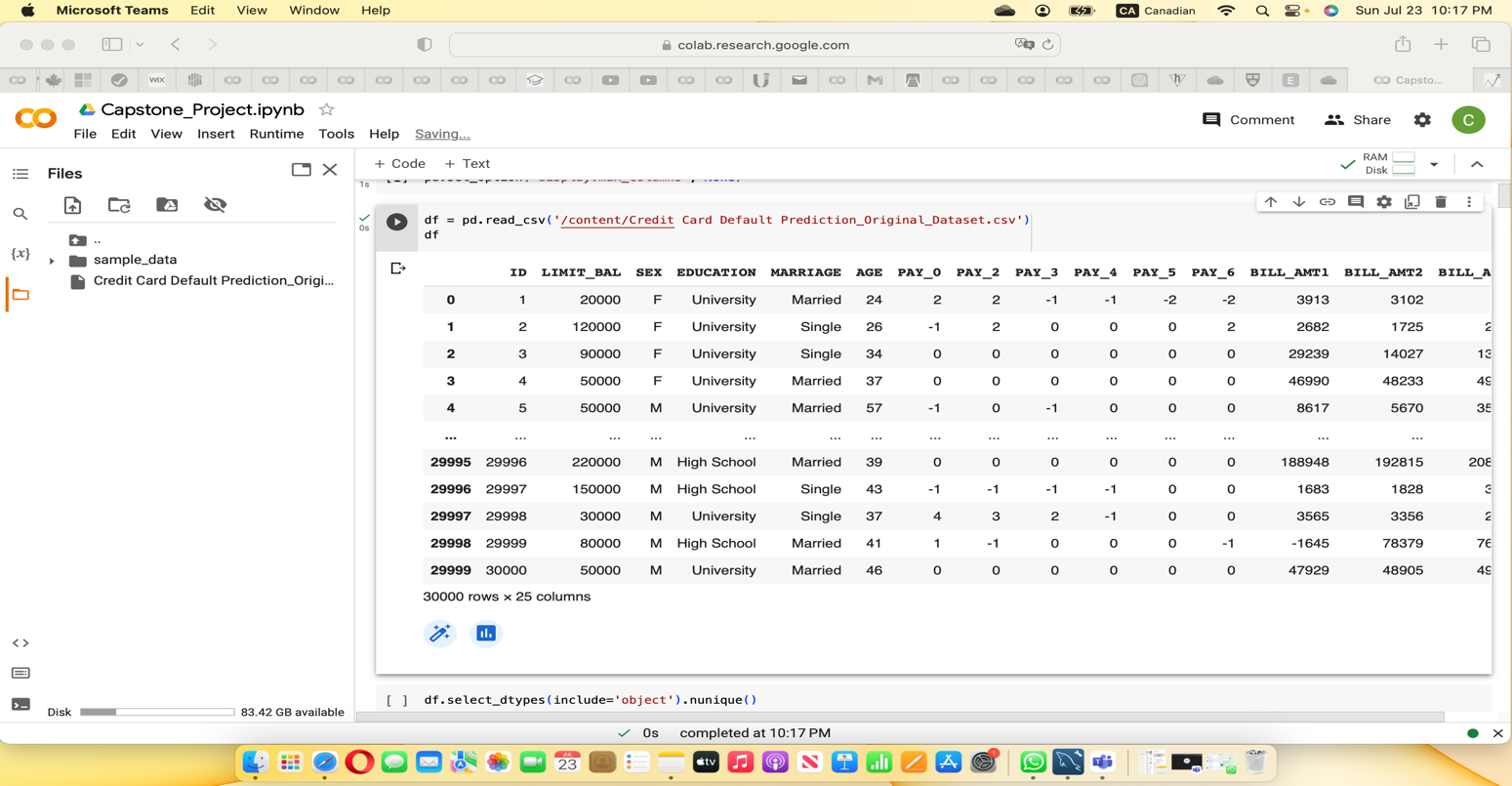
**AGE:** This is the customer. The age of customers is between 21 and 79.

**PAY\_0**: Repayment status in September 2005 (-1=pay duly, 1=payment delay for one month, 2=payment delay for two months,8=payment delay for eight - months, 9=payment delay for nine months and above)  
**PAY\_2**: Repayment status in August 2005 (scale same as above)  
**PAY\_3:** Repayment status in July 2005 (scale same as above)  
**PAY\_4:** Repayment status in June 2005 (scale same as above)  
**PAY\_5:** Repayment status in May 2005 (scale same as above)  
**PAY\_6:** Repayment status in April 2005 (scale same as above)  
**BILL\_AMT1:** Amount of bill statement in September 2005 (NT dollar)  
**BILL\_AMT2:** Amount of bill statement in August 2005 (NT dollar)  
**BILL\_AMT3:** Amount of bill statement in July 2005 (NT dollar)  
**BILL\_AMT4:** Amount of bill statement in June 2005 (NT dollar)  
**BILL\_AMT5:** Amount of bill statement in May 2005 (NT dollar)  
**BILL\_AMT6:** Amount of bill statement in April 2005 (NT dollar)  
**PAY\_AMT1:** Amount of previous payment in September 2005 (NT dollar)  
**PAY\_AMT2:** Amount of previous payment in August 2005 (NT dollar)  
**PAY\_AMT3:** Amount of previous payment in July 2005 (NT dollar)  
**PAY\_AMT4:** Amount of previous payment in June 2005 (NT dollar)  
**PAY\_AMT5:** Amount of previous payment in May 2005 (NT dollar)  
**PAY\_AMT6:** Amount of previous payment in April 2005 (NT dollar)

**DEFAULT:** This is the default status of the customer at the end period represented by Y and N. Y means default while N means not default.

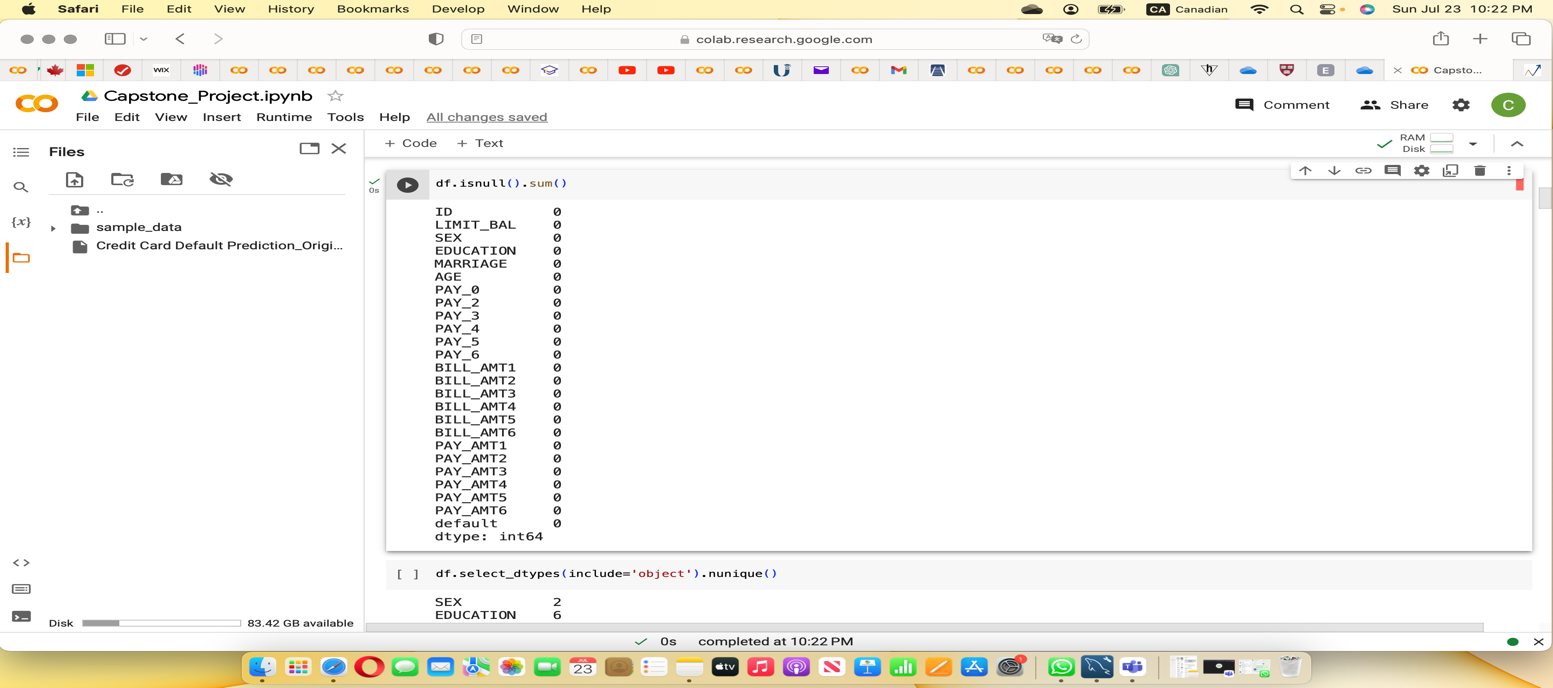
**Data download**

Below is the screenshot of our data download. The data has 30,000 observations and 25 features. The features are made up of 21 numeric features and four object features.



**Checking for Null values in the dataset.**

Below is the screenshot of the dataset features showing that there is null value in the dataset.



**Number unique object variables**

**This screenshot shows the list of the object variables with unique classifications in each of them. The Sex variables has two classifications, the male and female, education variable has six classifications, university, graduate school, high school, others, zero, and unknow. However, the zero class of education has observation. Marriage on the other hand has 4 classifications, married, single, others, and zero. Just in education, the zero class in marriage has no observation. Finally, we have default classified into Y and N representing yes and no.**



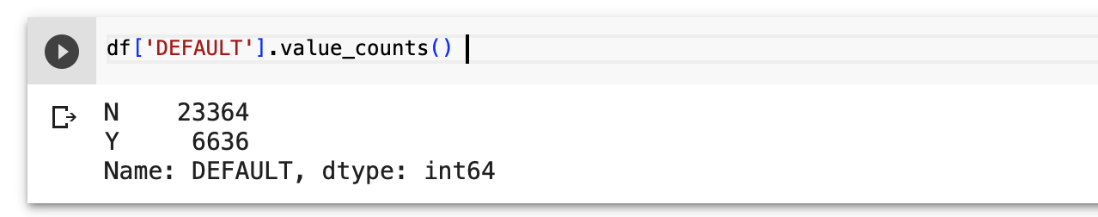
**Screenshot of the renamed variable ‘default’**

As stated earlier, default is the binary variable categorized into yes and no. Yes, means that customer defaulted in credit card payment while no means that customer did not default in the credit card payment.



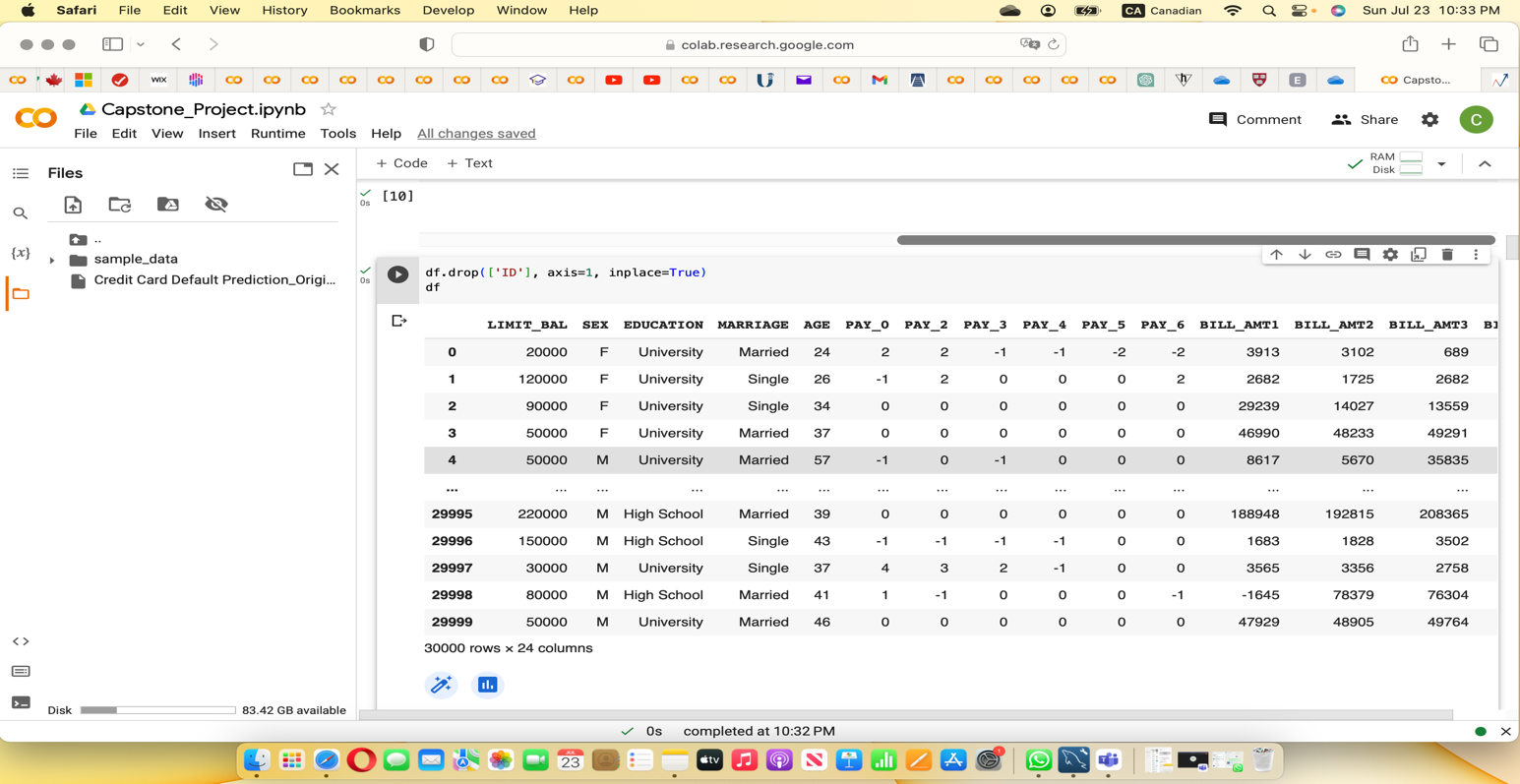
**Count of number of DEFAULT**

In the dataset, there are 6,636 of yes, meaning those customers defaulted while 23, 364 of no meaning those customers did not default. See below the second screenshot of the count of default variable.



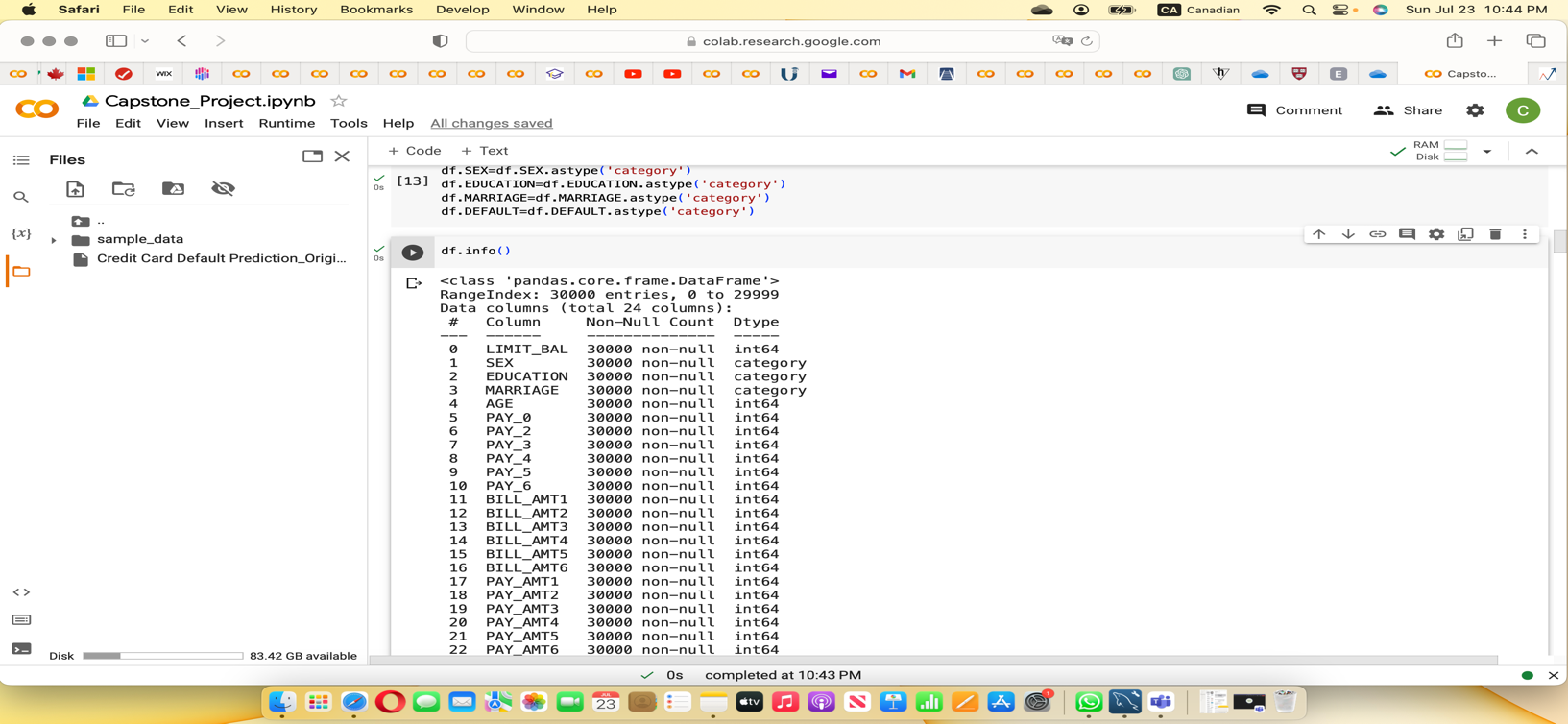
**Reduction of variables (ID removed)**

ID was removed from the dataset because it is not relevant to building our models.



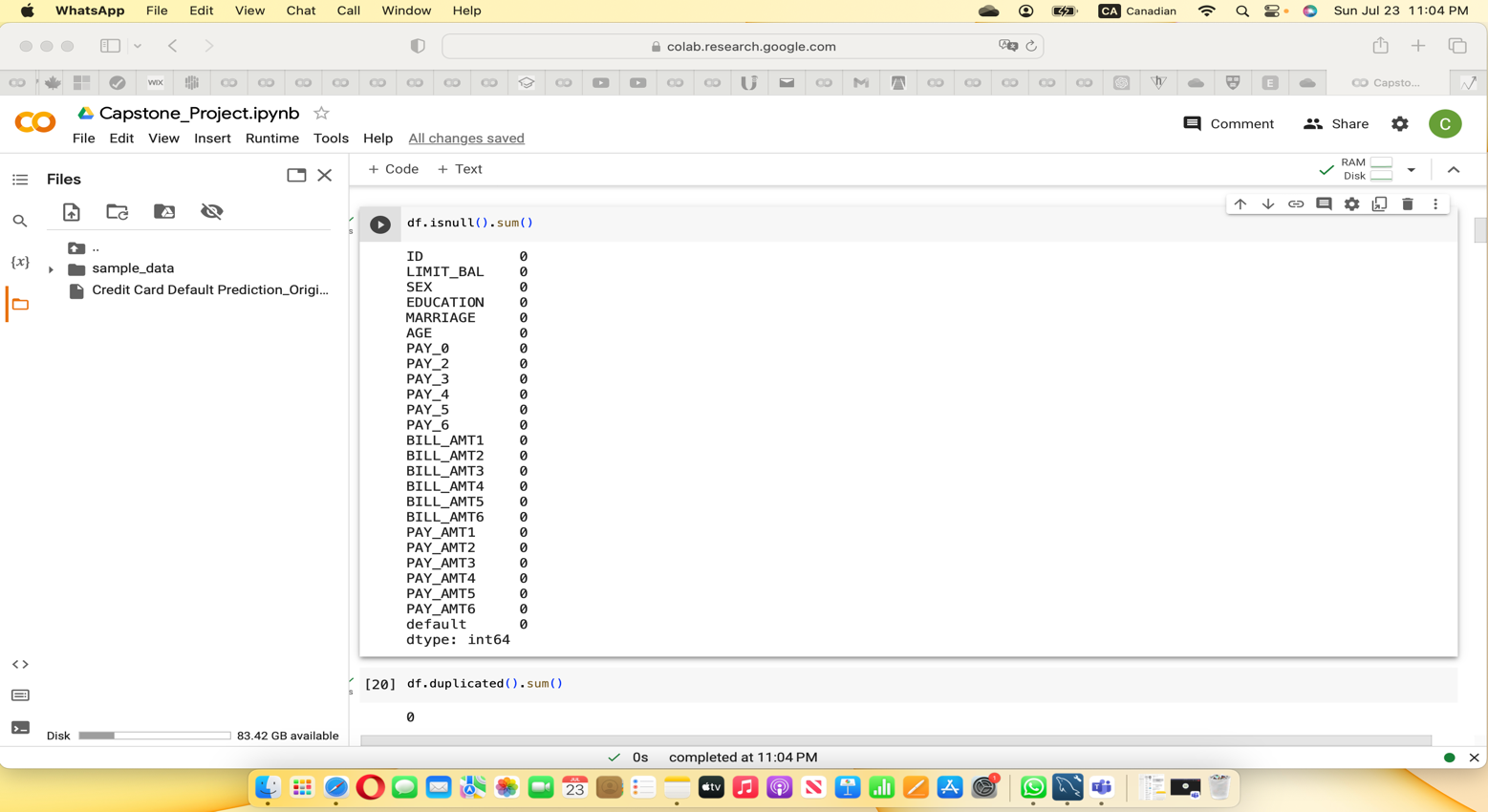
**Data info after transformation of object variables**

**Below is our info after transformation of the object variables into category. The four object variables have been converted into category variables at this stage.**



**Checking for missing value and duplicate**

Here is the screenshot showing no missing values and no duplicate in the dataset.



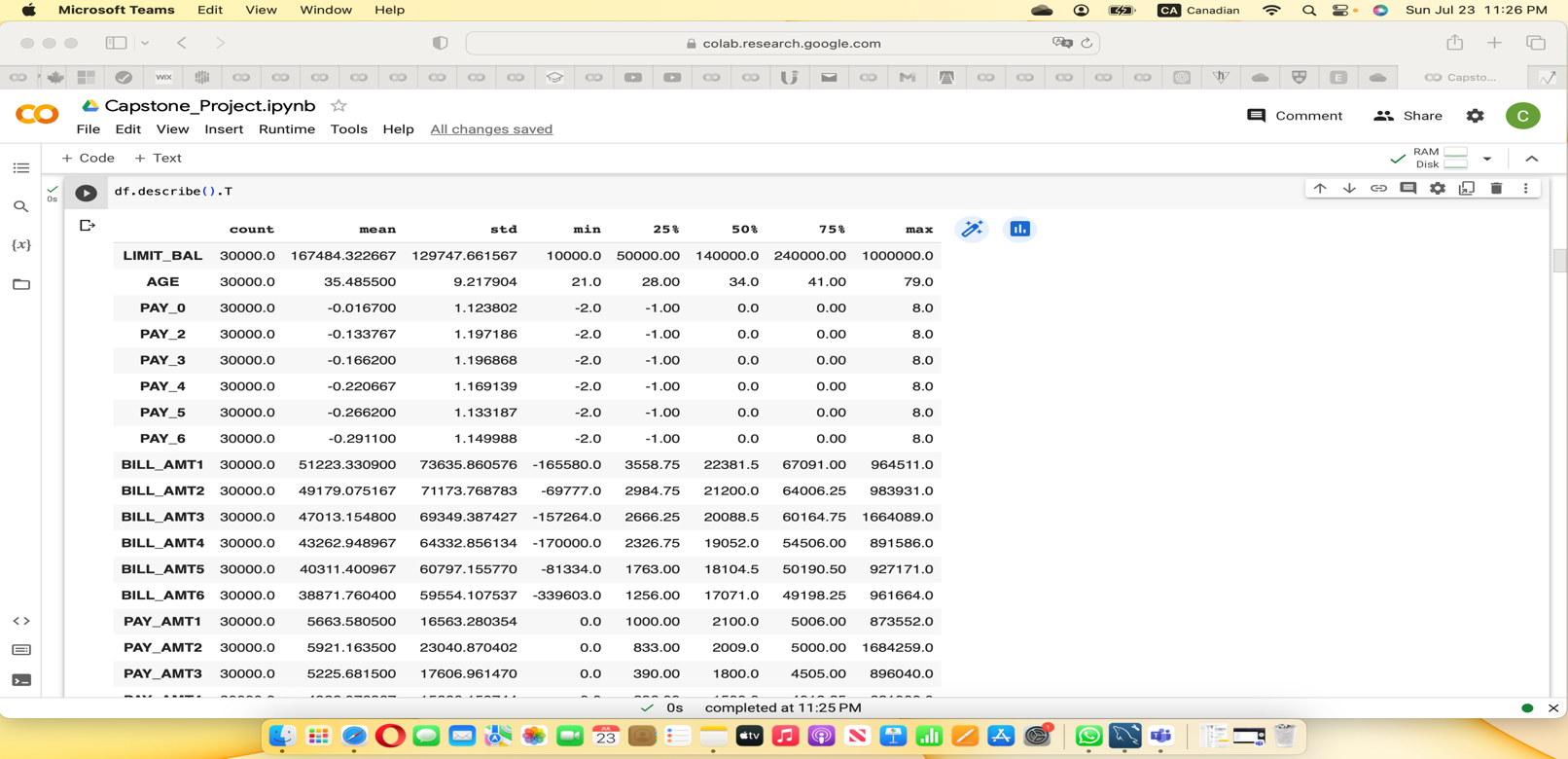
1. Exploratory Data Analysis (EDA)

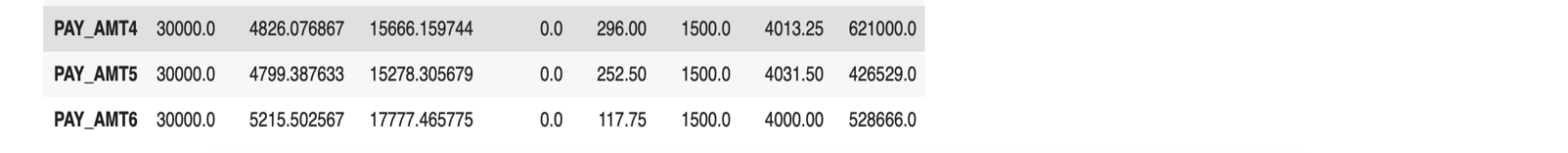
At this stage, we analyzed the data to investigate, discover patterns, and check assumptions using summary statistics and graphs.

The summary statistics provided quick and simple insights into the description of the data.

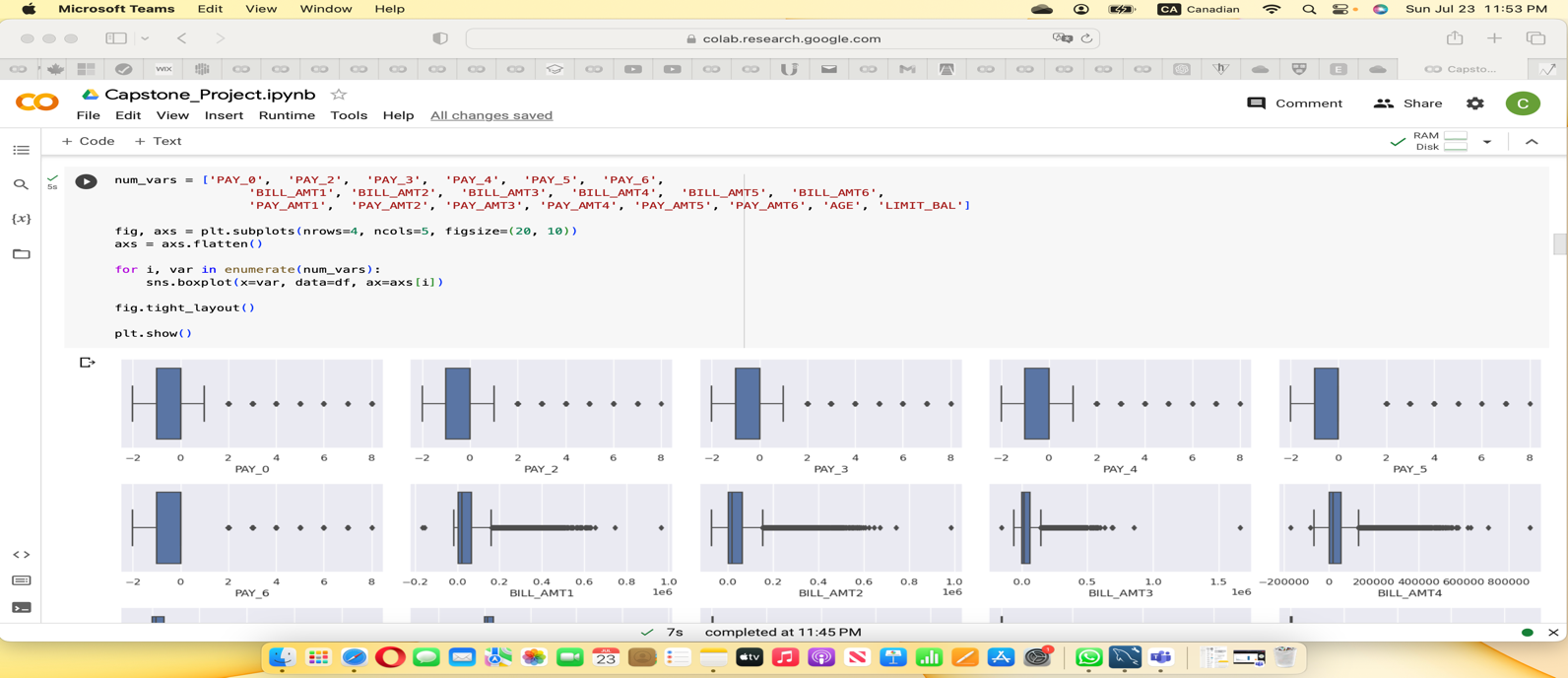
**Summary Statistics**

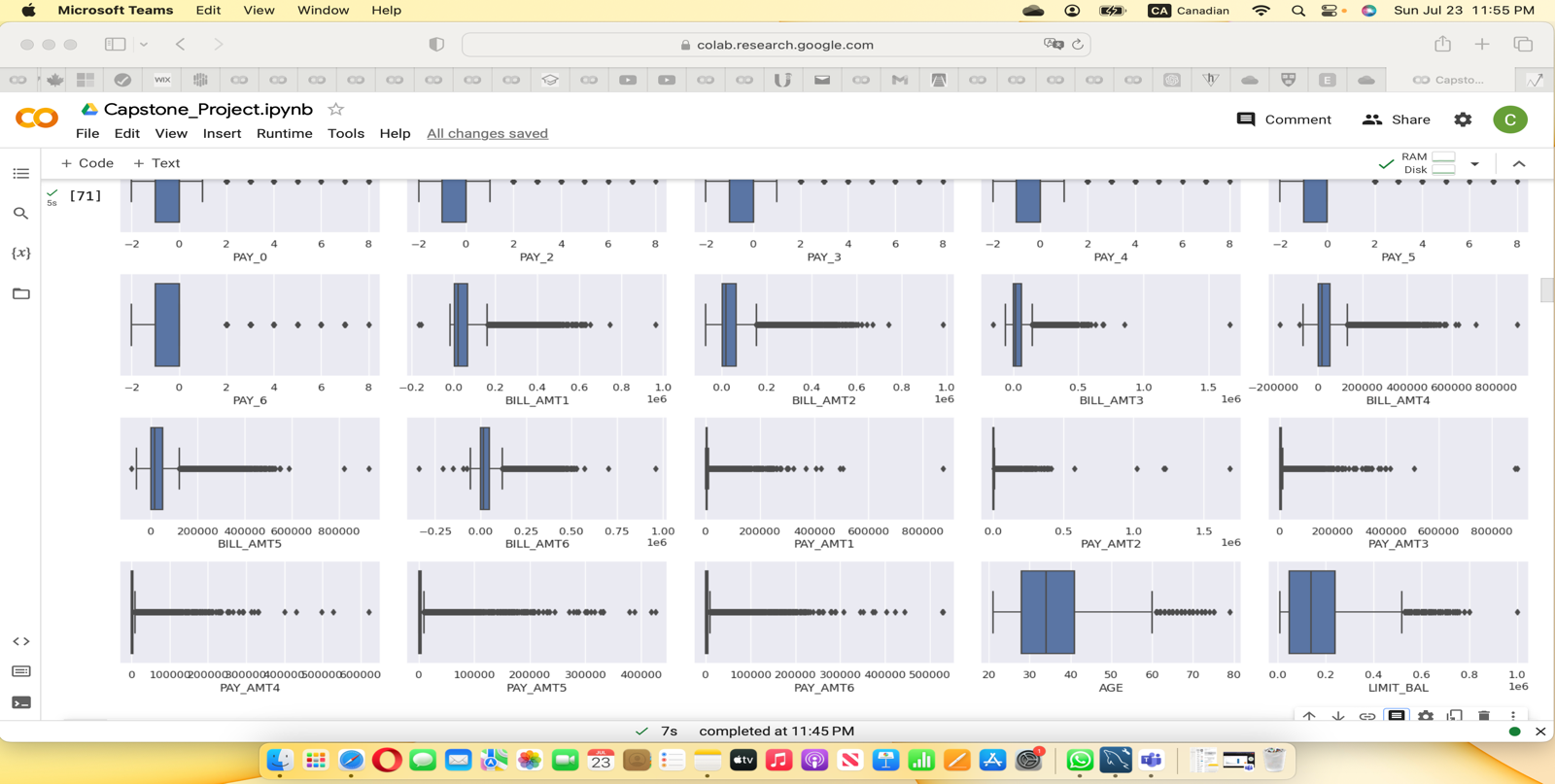
This is the summary statistics showing the statistics of the numerical variables. In the summary, we have the count of the observations, the mean, standard deviation, minimum, 1st quartile, 2nd quartile, 3rd quartile, and maximum values of each of the variables.





**Outliers using Boxplot**



From the summary statistics and the screenshot of the boxplot above, there are big gaps between the mean values and the maximum values of the numerical variables which are abnormal. These are cases of presence of outliers in the dataset. All the numerical variables have outliers. We will treat the dataset of the outliers to remove the outliers. Although all our variables in the dataset have outliers, the following stand out.

**LIMIT\_BAL**: The mean of LIMIT\_BAL of $167, 484.32 compared the maximum LIMIT\_BAL of $1,000,000 is abnormal. There is a huge gap between the minimum value and maximum value which is clearly seen in the boxplot as an outlier. It shows that on average, a credit card customer has a credit limit balance of $167, 484.32 and a maximum limit balance of $1, 000, 000.

**BILL\_AMT3**: Another variable that stands out is the BILL\_AMT3 which has a mean value of $47,013 compared to the maximum value of $1, 664, 089. Again, this is abnormal. This shows that the average bill amount for the 30,000 customers in six months of April to September 2005 is $47, 013 whereas the maximum single bill amount of $1, 664, 089. The $1, 664, 089 is an abnormal bill in the dataset and will be removed along with other outliers when treating the dataset for outliers. The difference between the minimum value and the maximum is so wide.

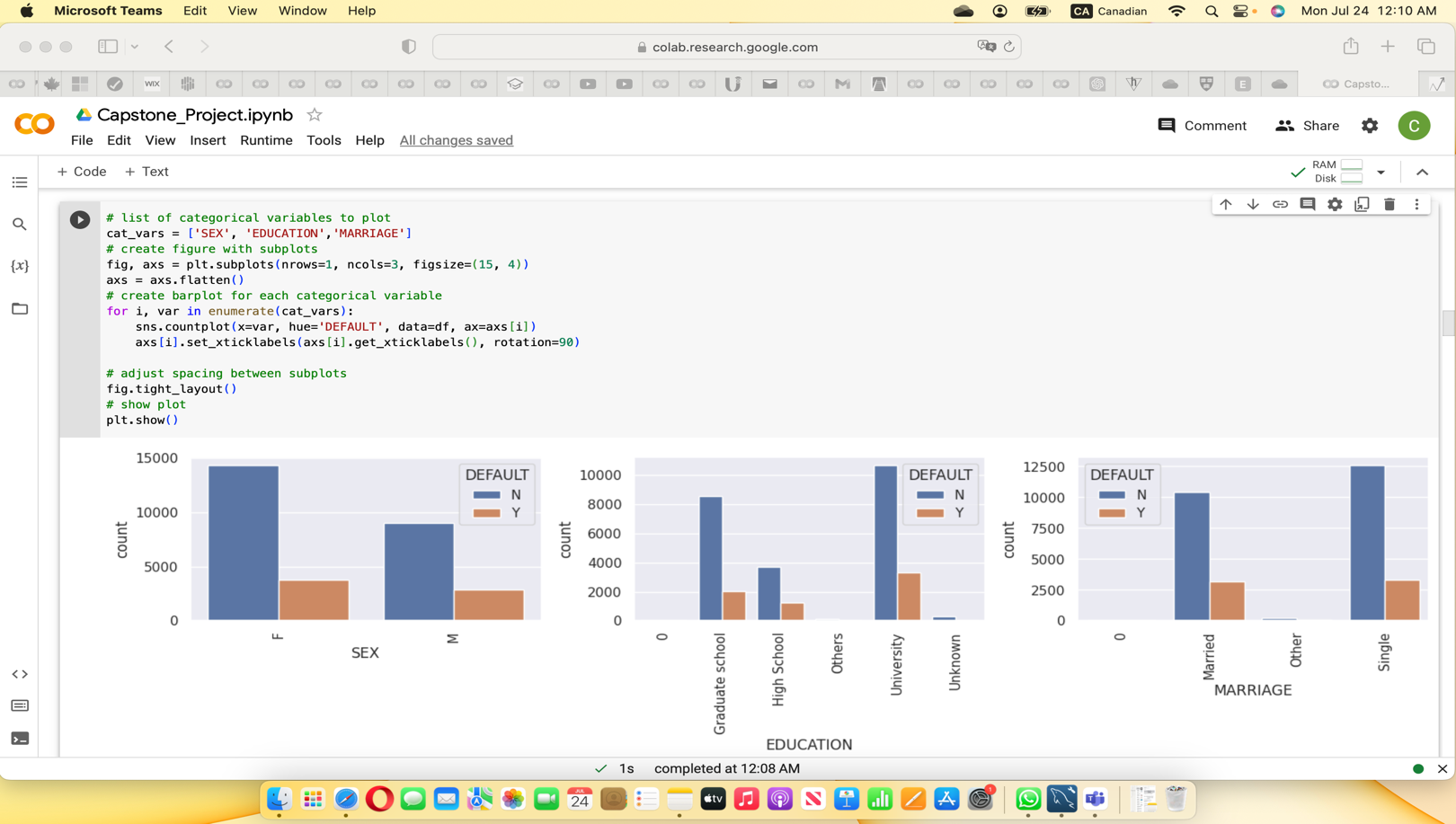
**PAY\_AMT1 to PAY\_AMT6:** These variables represent the payment amounts from April 2005 to September 2005. The standard deviation of these variables shows the variability in payment amount for different months. The gap between the minimum and maximum is also very wide meaning there are outliers.

**AGE:** The mean AGE of 35.48 and the standard deviation is approximately 9.22, indicating some variability in ages show that the ages of customers are more spread out from the mean age. This is evidence of the presence of outliers in the variable.

To corroborate the information from the summary statistics, we plotted the graphs of the numeric variables in a boxplot to detect outliers. It also confirmed that there are outliers in the numeric variables as seen below.

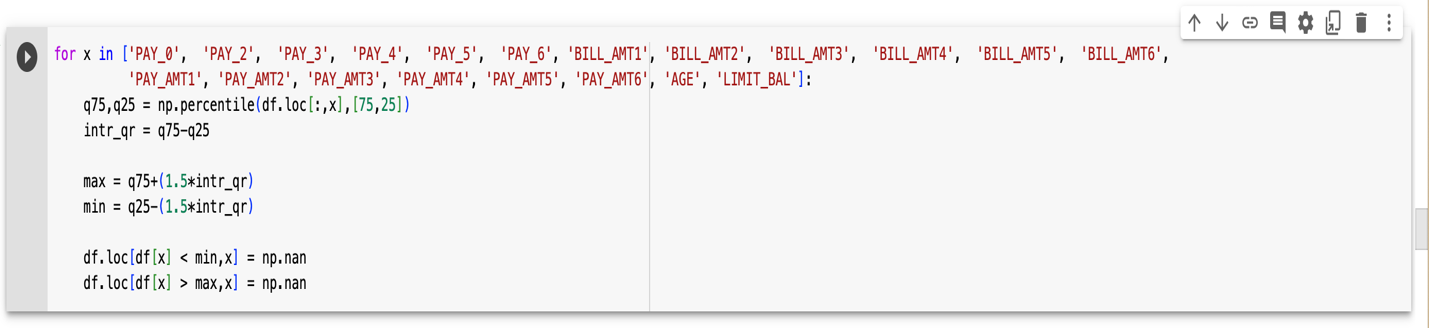
**Graph of categorical variables**

Below are the graphs of categorical variables showing the relationship between the variables and credit card default. From the variable ‘SEX’, it shows that more F (female) gender defaulted in credit card payment than M (male) counterparts. In the variable ‘EDUCATION’ more university students followed by graduate students defaulted in credit card payment than high school students, unknow and others. On the other hand, in the variable ‘MARRIAGE’, more singles defaulted in credit card payment than their married counterparts and others.

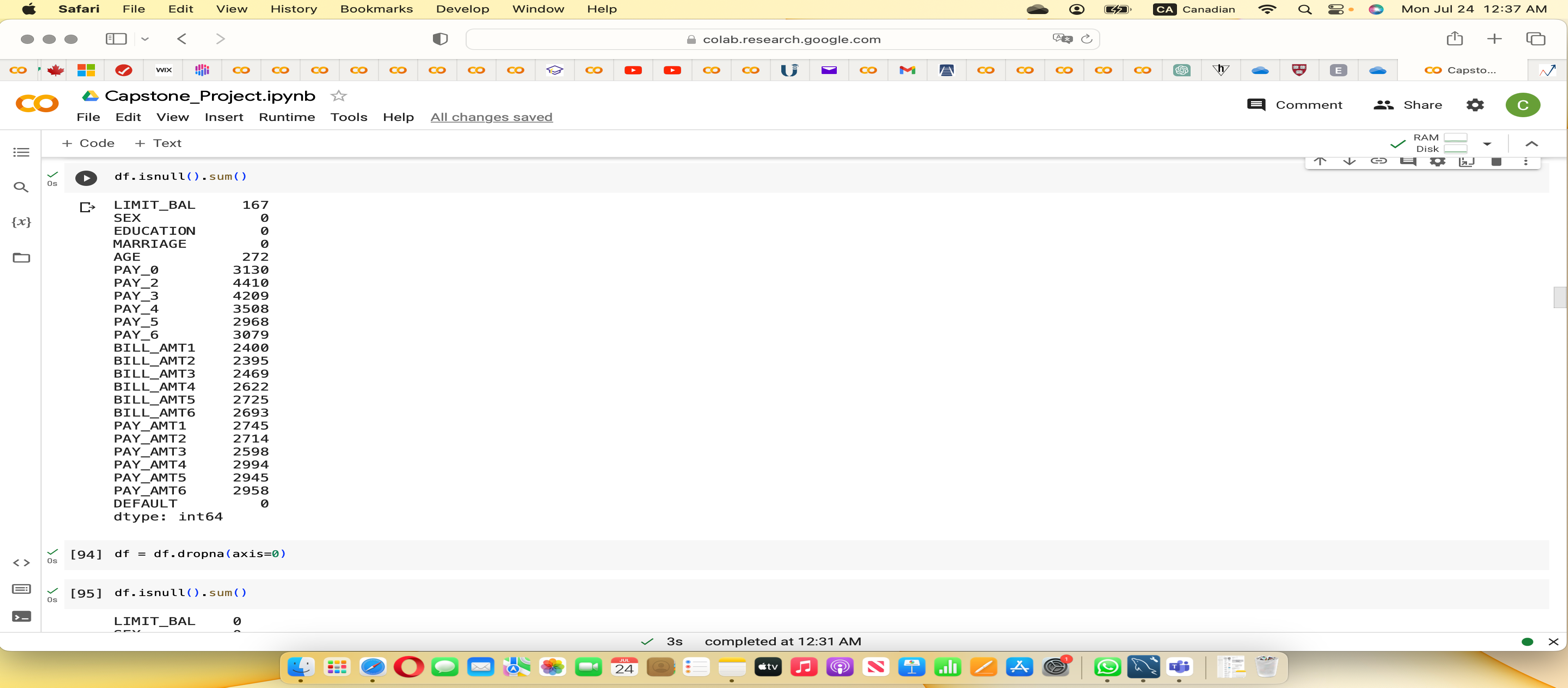


**Code to check for outliers**

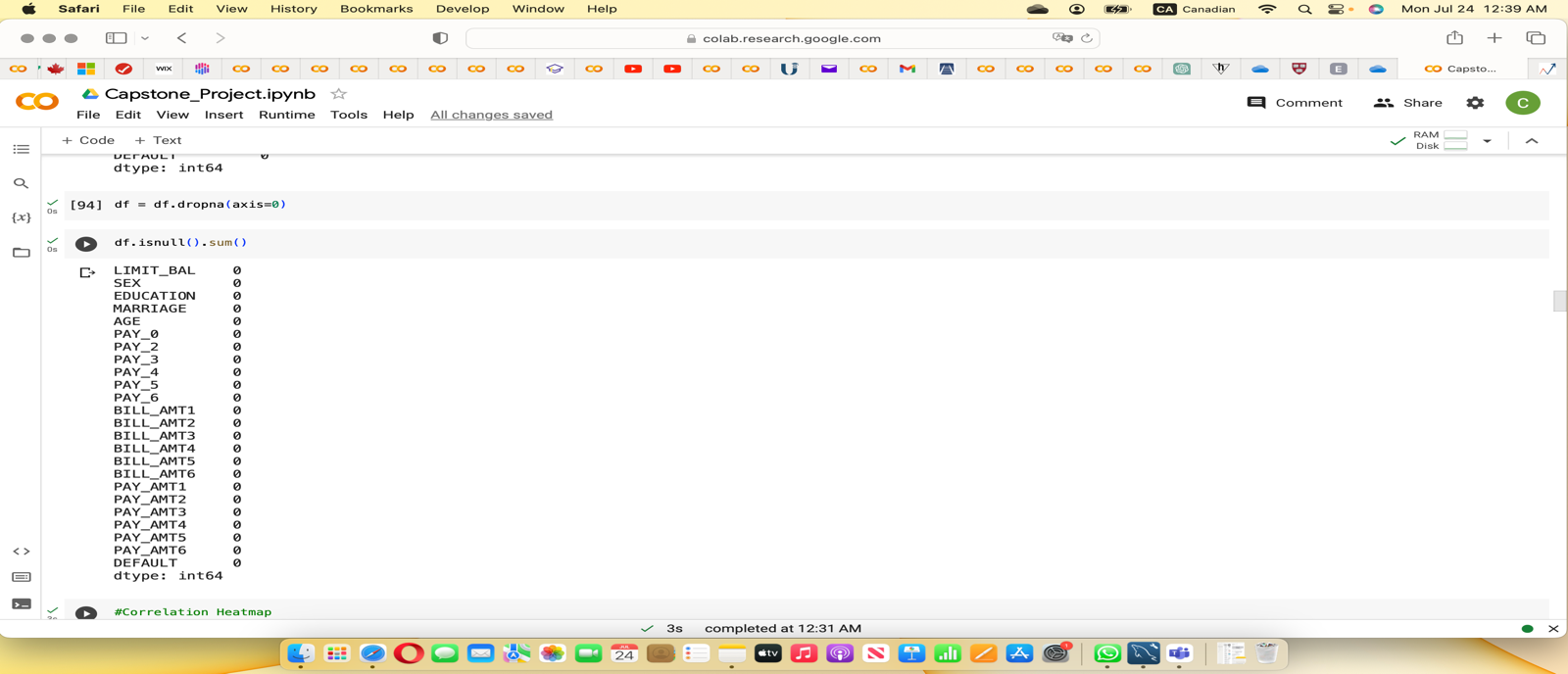
This shows the screenshot of the code used to check the presence of outliers in the dataset.



This is a screenshot showing outliers in numerical variables and the code to remove them.

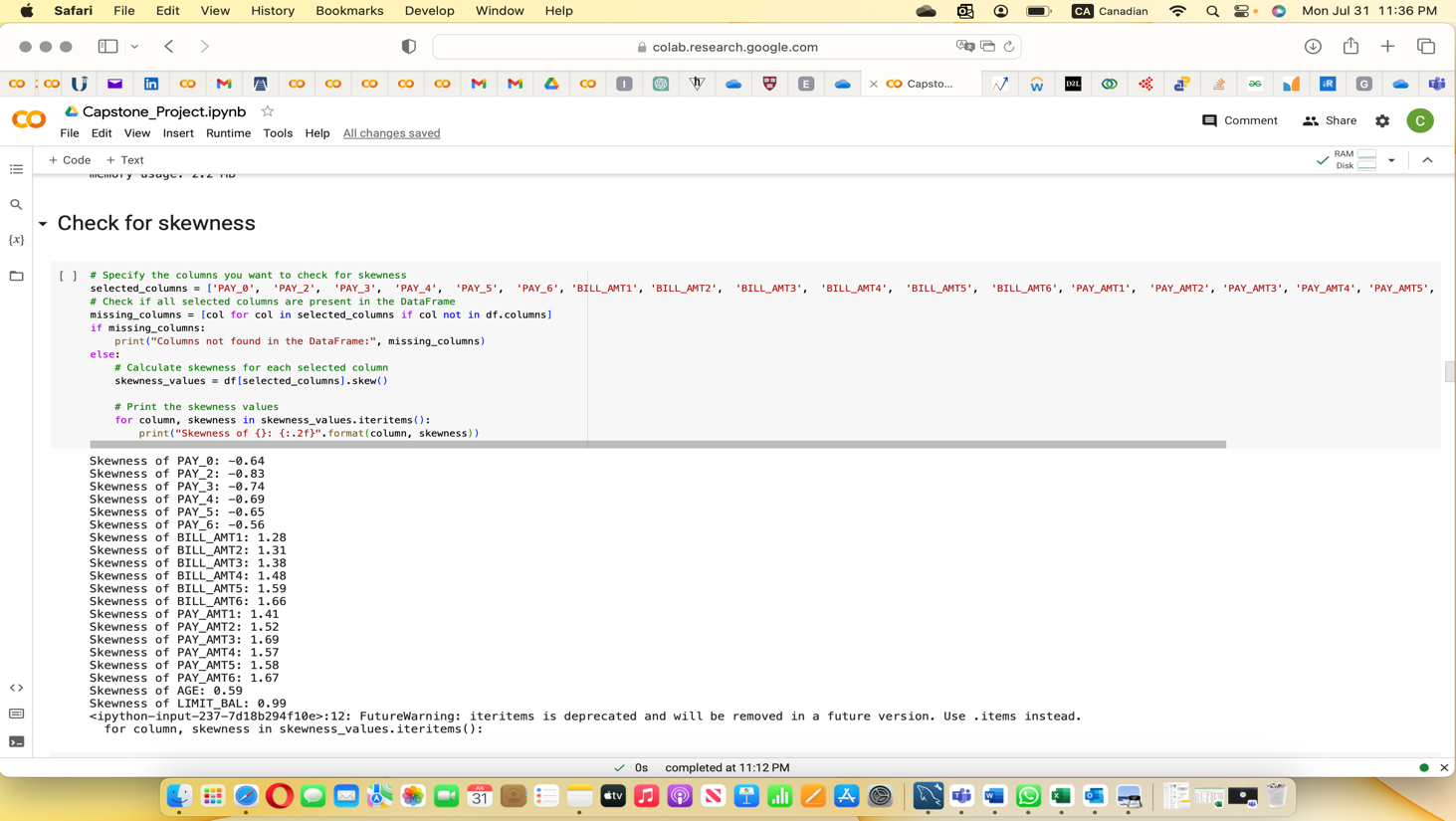


Finally, the screenshot below shows the result that the outliers have been removed from the dataset. We used the code above to remove outliers from the numerical variables. After removing the outliers, our dataset has been reduced from 30,000 to 13, 145 observations with 24 variables. We now have a manageable size of data for our modelling.

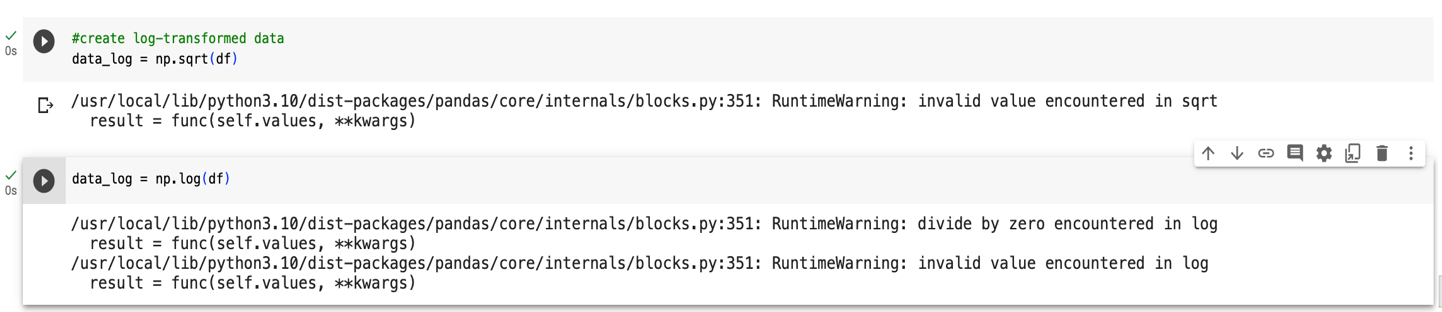


**Checking for Skewness**

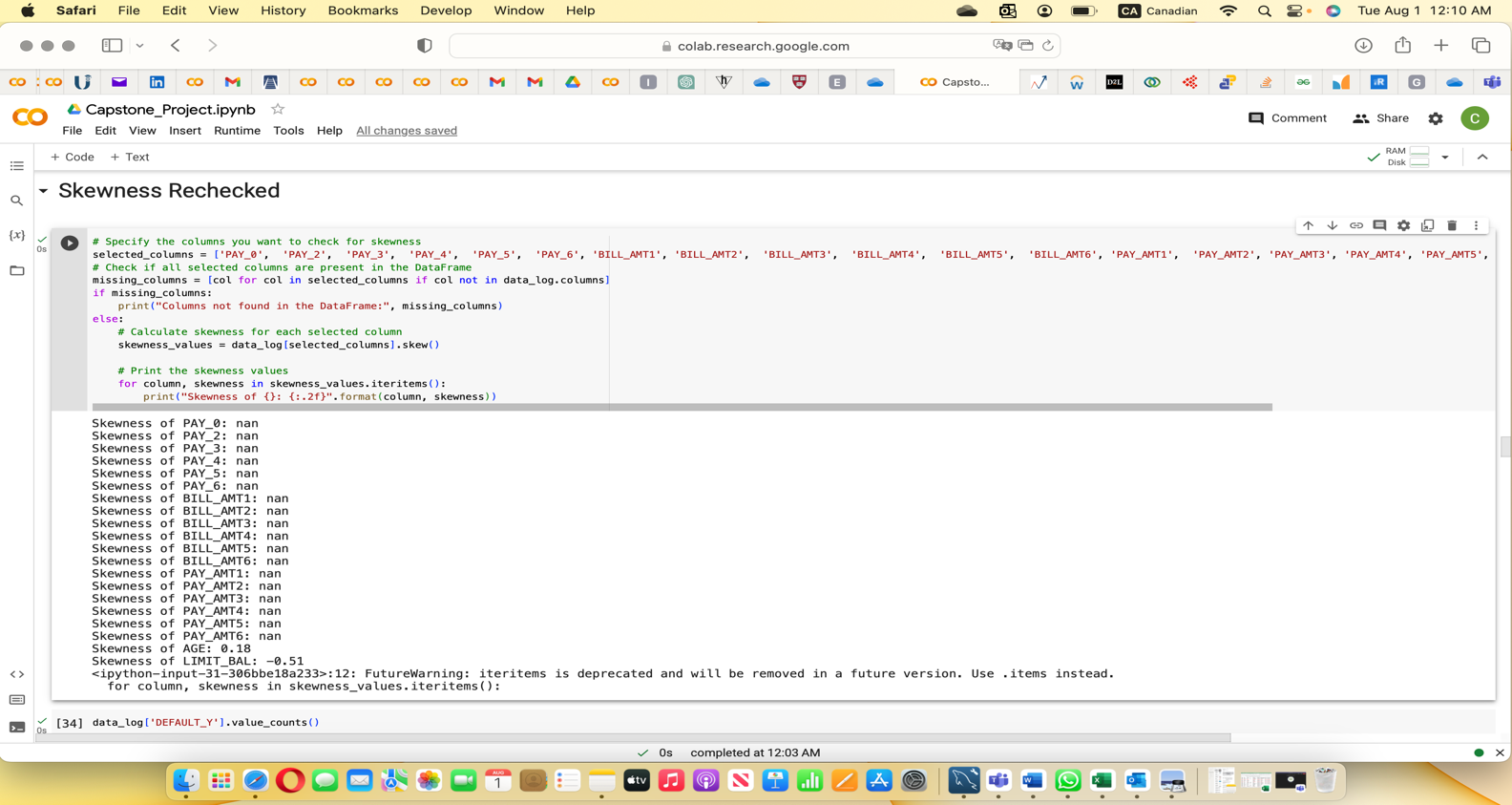
After removing the outliers, we checked the data for skewness. We observed that some of variables are skewed. We applied log transform and other methods to reduce the skewness, however, some are still. We therefore left the dataset with the skewness variables.



Below is the screenshot of code used in trying to reduce the skewness. We tried to reduce skewness from the data but were unable to and so we left them.



Below is the screenshot of the result after trying to reduce the skewness.



**Correlation Heatmap**

We checked for correlated variables in the dataset using heatmap. However, we did not notice any strong correlation between the variables. Although some of the variables are correlated, there is no strong correlation between the variables. We considered correlation coefficient of 0.7 and above as a strong correlation and below –0.5 to 0.6 as normal correlation. The following are the correlated variables in the dataset.

