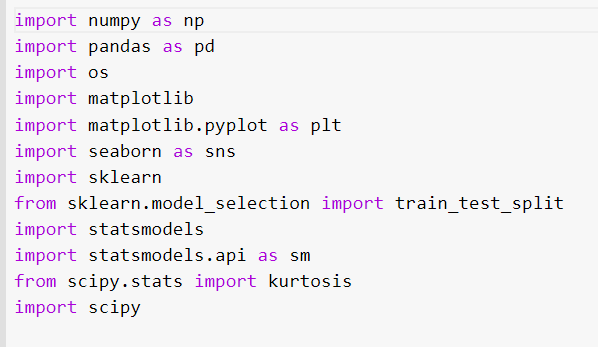
# **Exploratory Data Analysis (EDA)**

# Introduction

# This case study aims to give us an idea of applying EDA in a real business scenario. In this case study, apart from applying the techniques that we have learnt in the EDA module, we will also develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimize the risk of losing money while lending to customers.

## Importing required libraries:

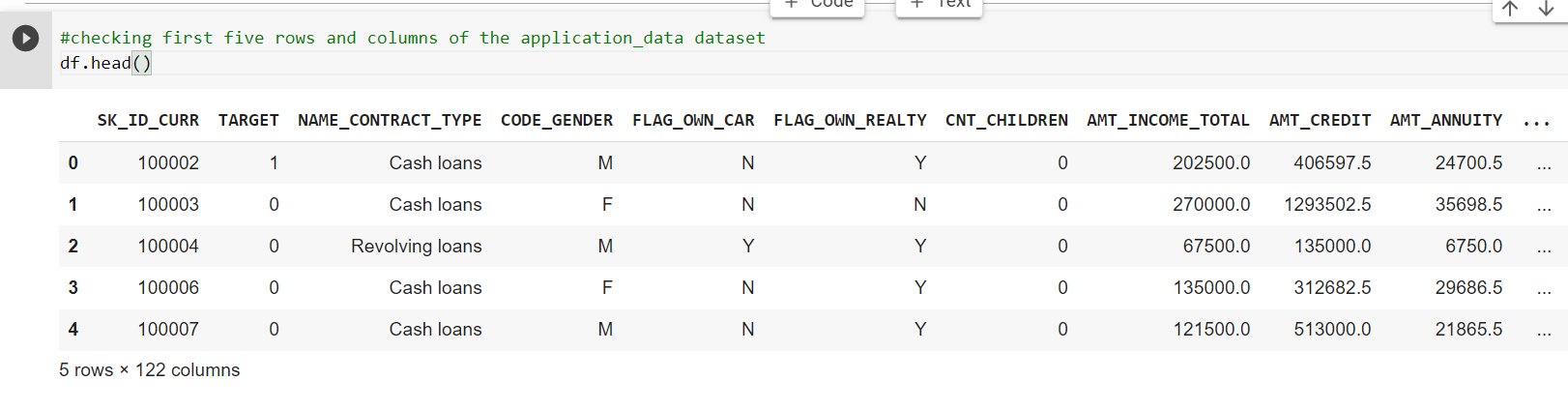


## Importing Data from CSV

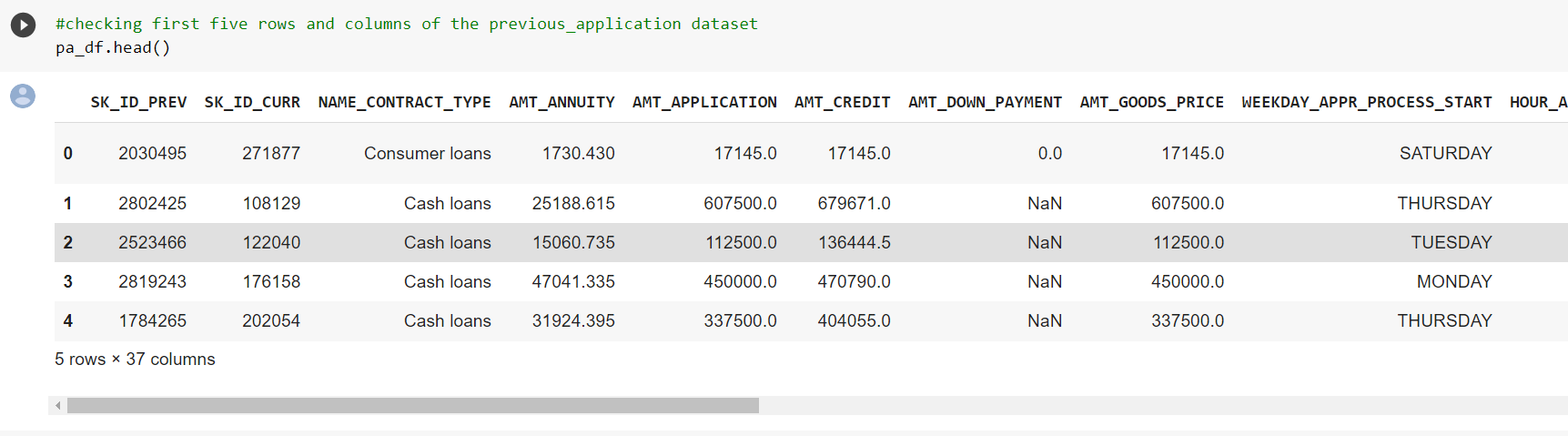


# Viewing the Dataset

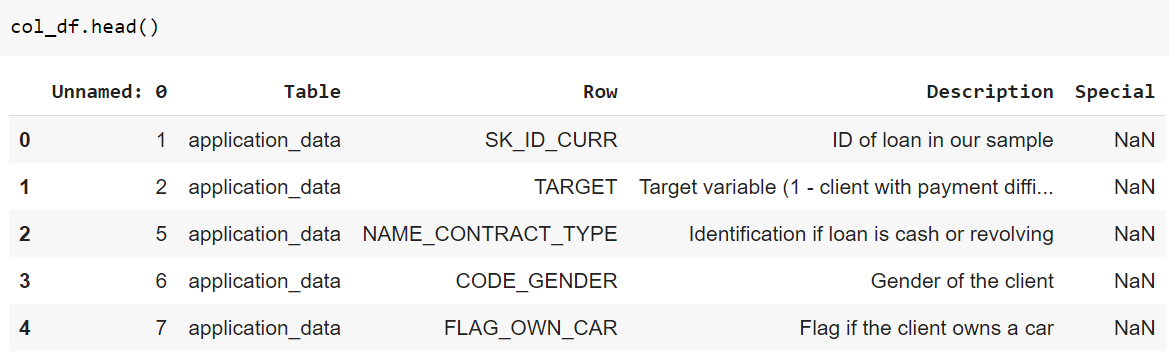
### 1.Viewing the Dataset of application\_data



### 2.Viewing the Dataset of previous\_application



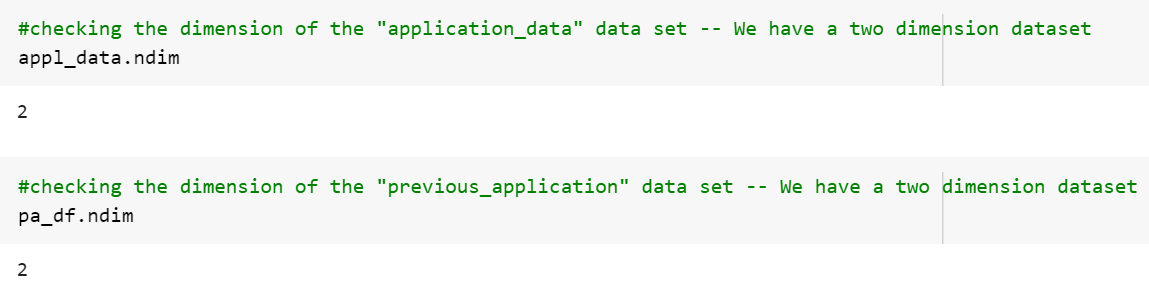
### 3.Viewing the Dataset of Columns\_Description



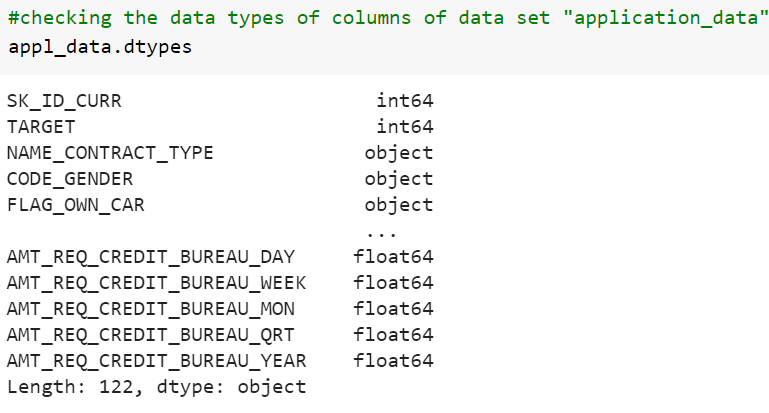
## Checking the shape of dataset

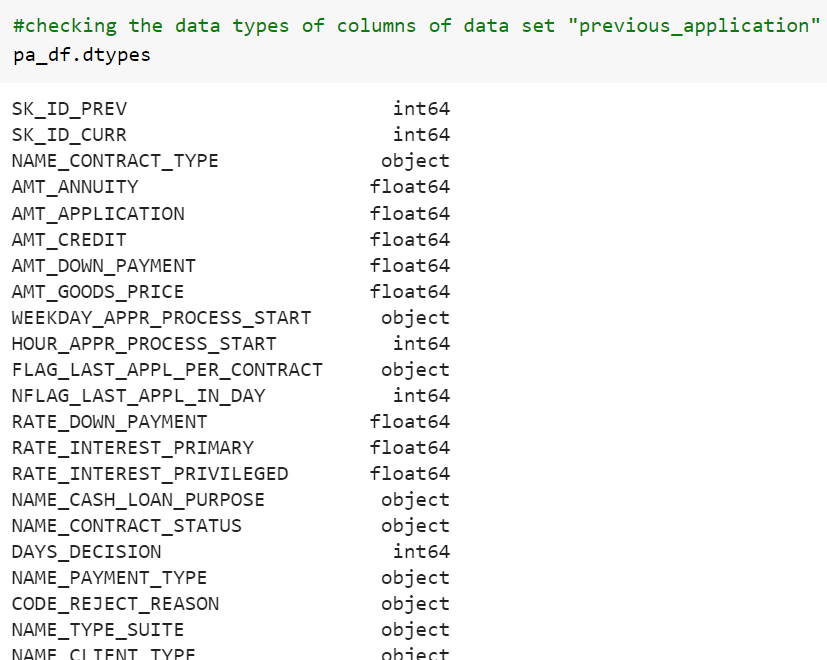
### 

## Checking the dimension of dataset

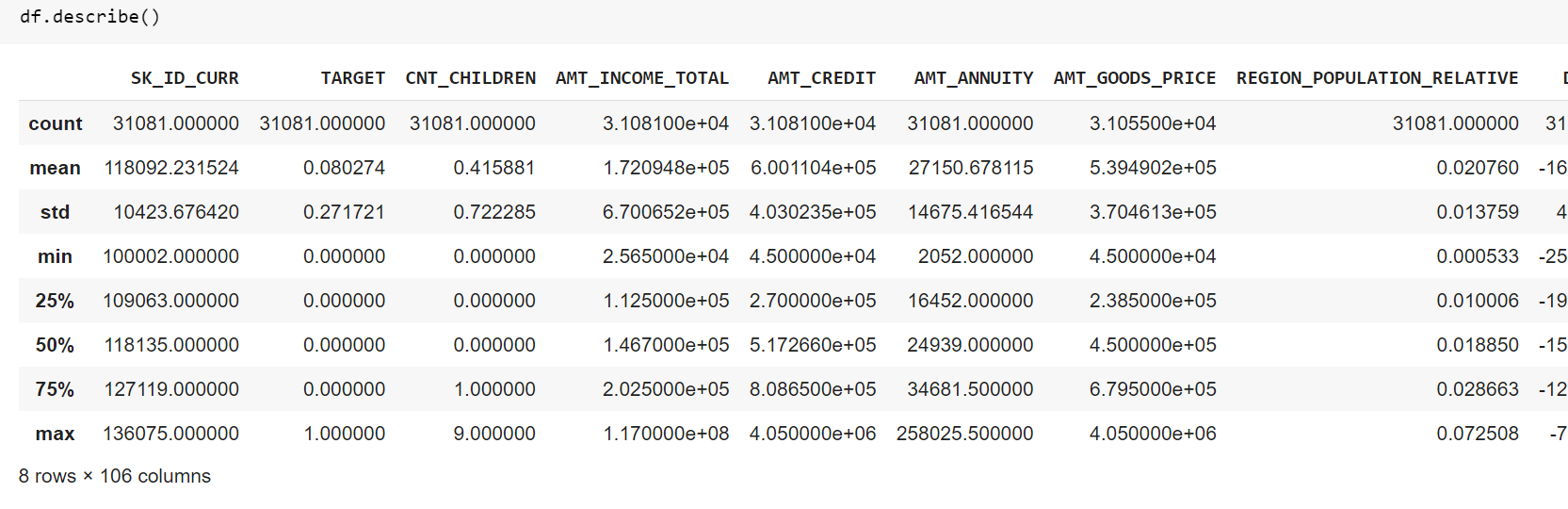


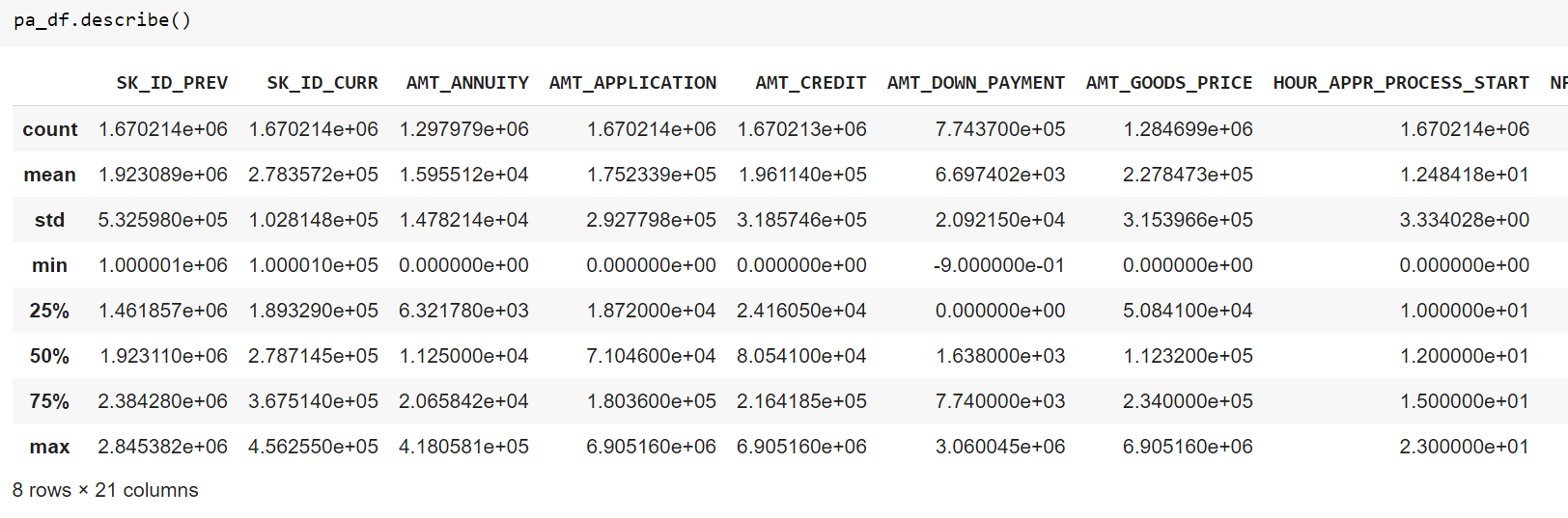
## Checking the datatype of dataset





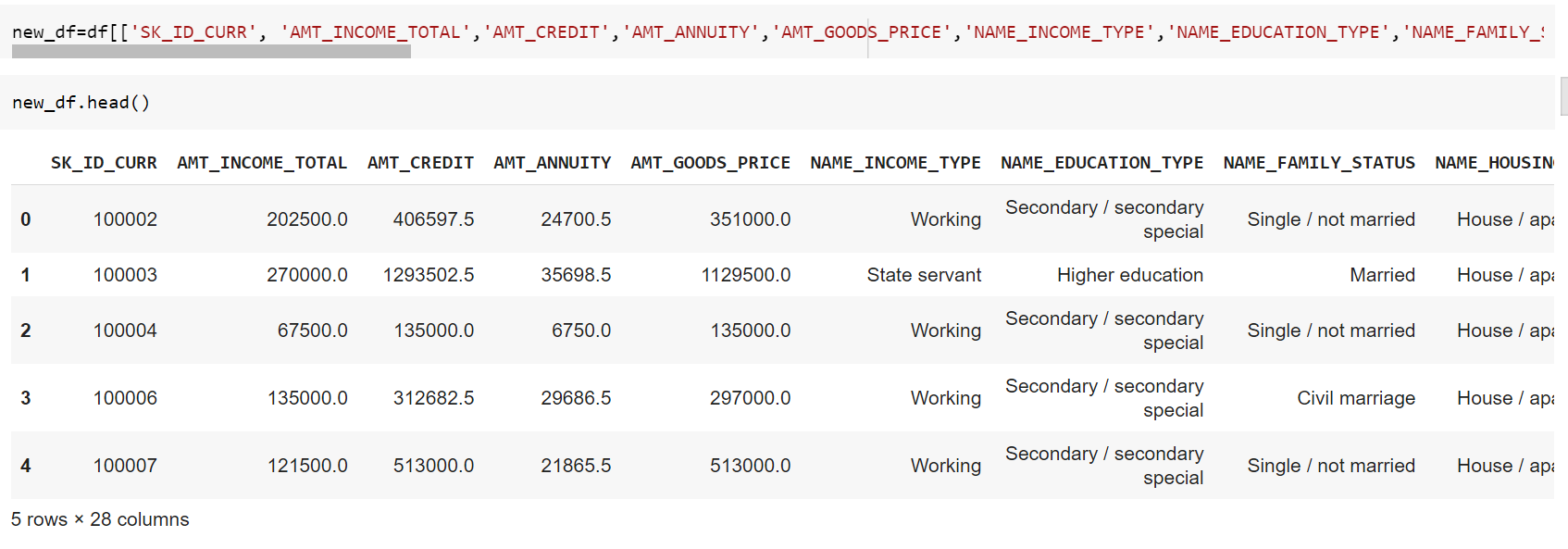
# Statistical summary of all numeric-typed (int, float) columns( from describe command)



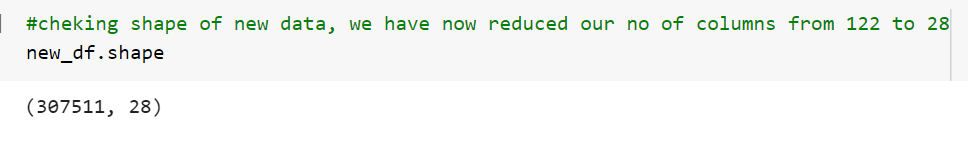


Data Preparation

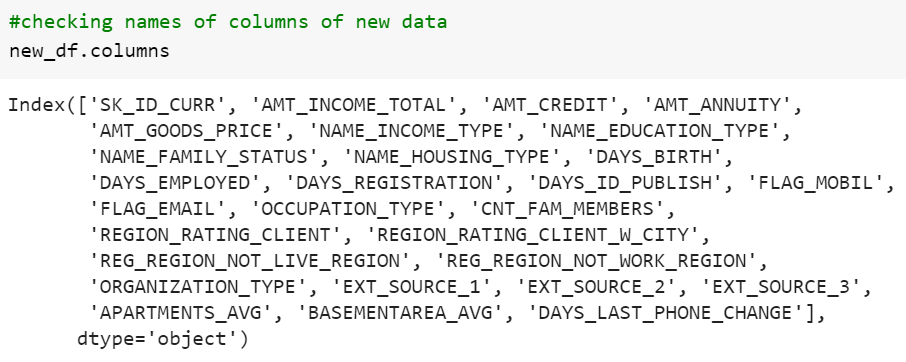
## Creating a new dataframe using the columns of df



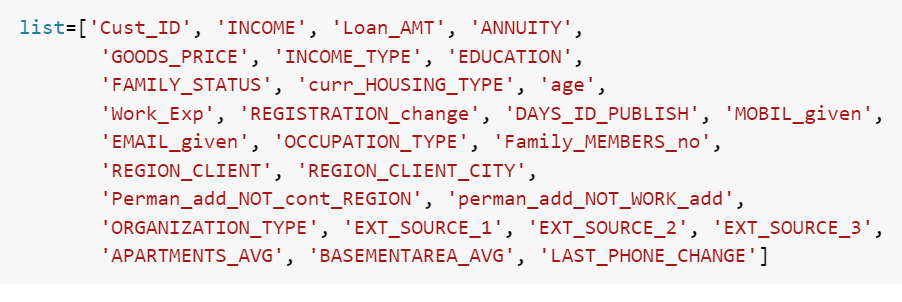
## Checking the shape of new dataframe

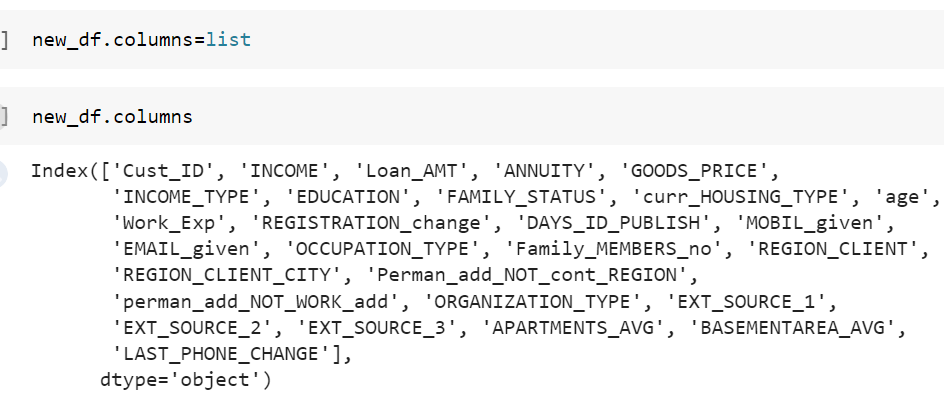


## Checking the names of columns of new dataframe



## Renaming the column names



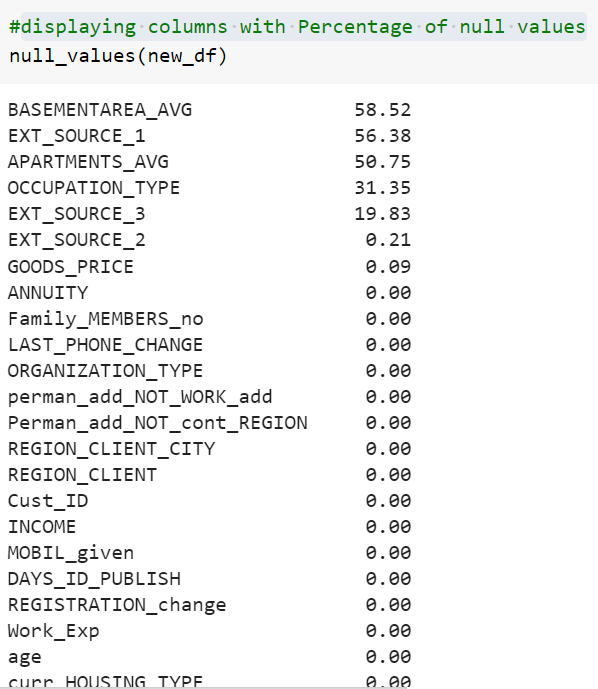


## Checking for columns with null values

Using a for loop in Python, we can quickly figure out the number of missing values in each column. As mentioned above, "True" represents a missing value and "False" means the value is present in the dataset. In the body of the for loop the method ".value\_counts()" counts the number of "True" values.



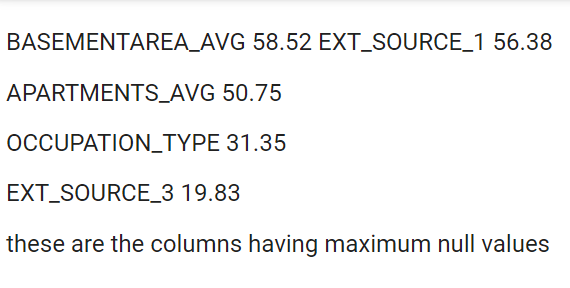
## Displaying columns with Percentage of null values



### 

# Summary:

### Columns having maximum null values



# Replacing null values with NAN

We can deal with missing data by the following ways:

1.Drop data

a. Drop the whole row

b. Drop the whole column

2.Replace data

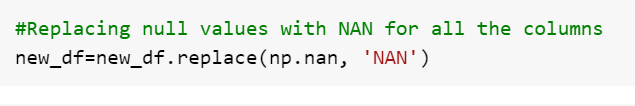
a. Replace it by mean

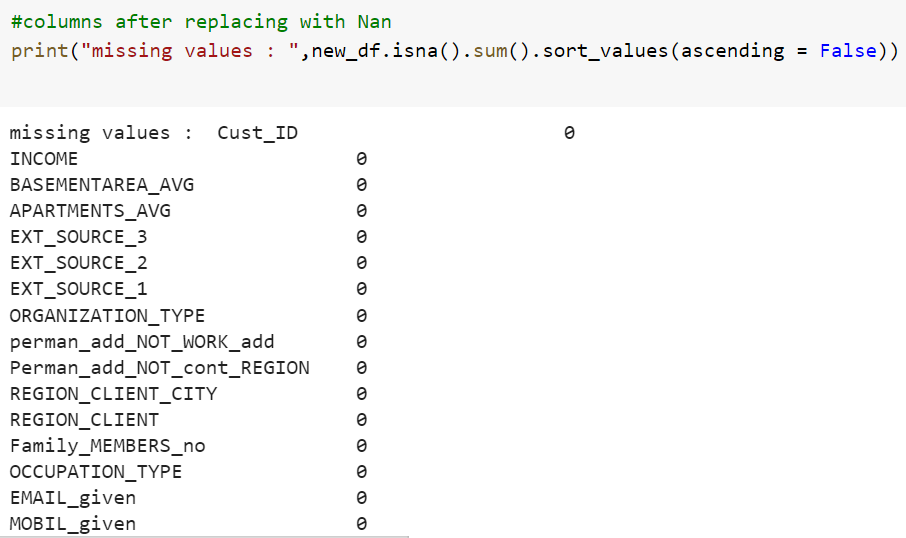
b. Replace it by frequency

c. Replace it based on other functions

Whole columns should be dropped only if most entries in the column are empty. In our dataset, none of the columns are empty enough to drop entirely.

We have some freedom in choosing which method to replace data; however, some methods may seem more reasonable than others.

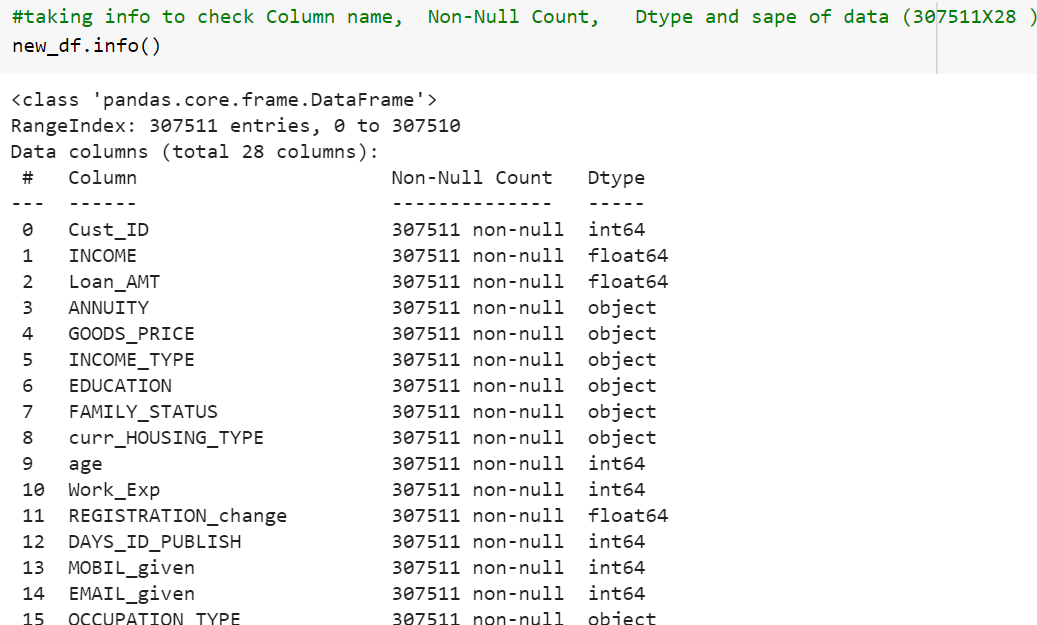




### 

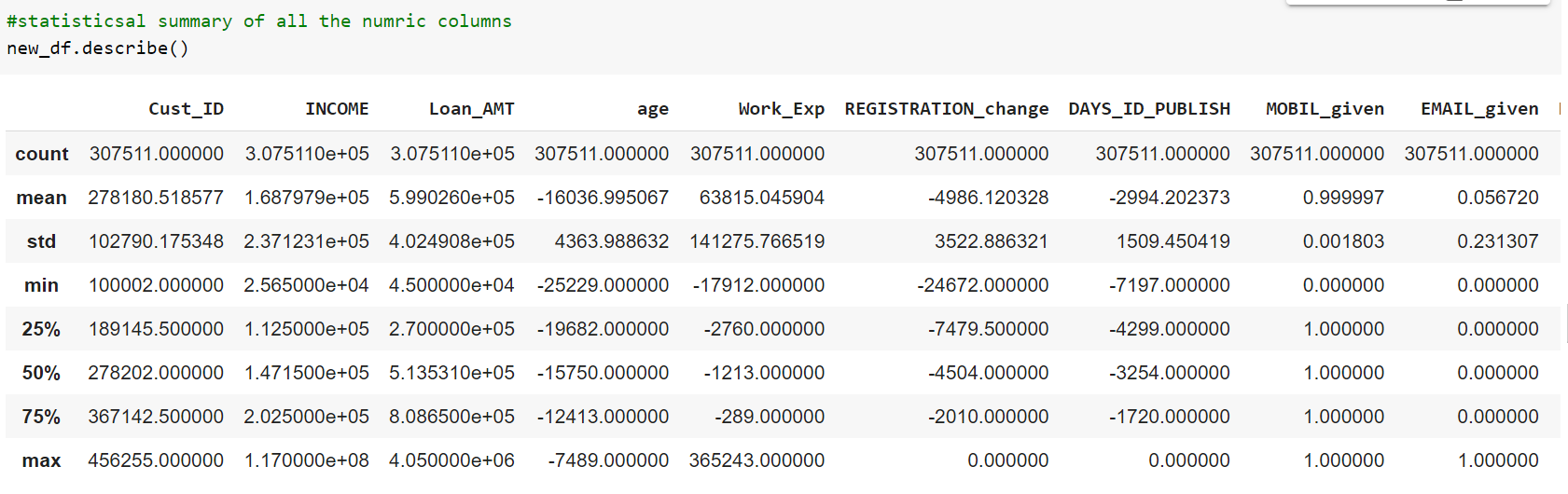
We replaced all empty columns with NAN.

# Understanding of the variables Categorical variables

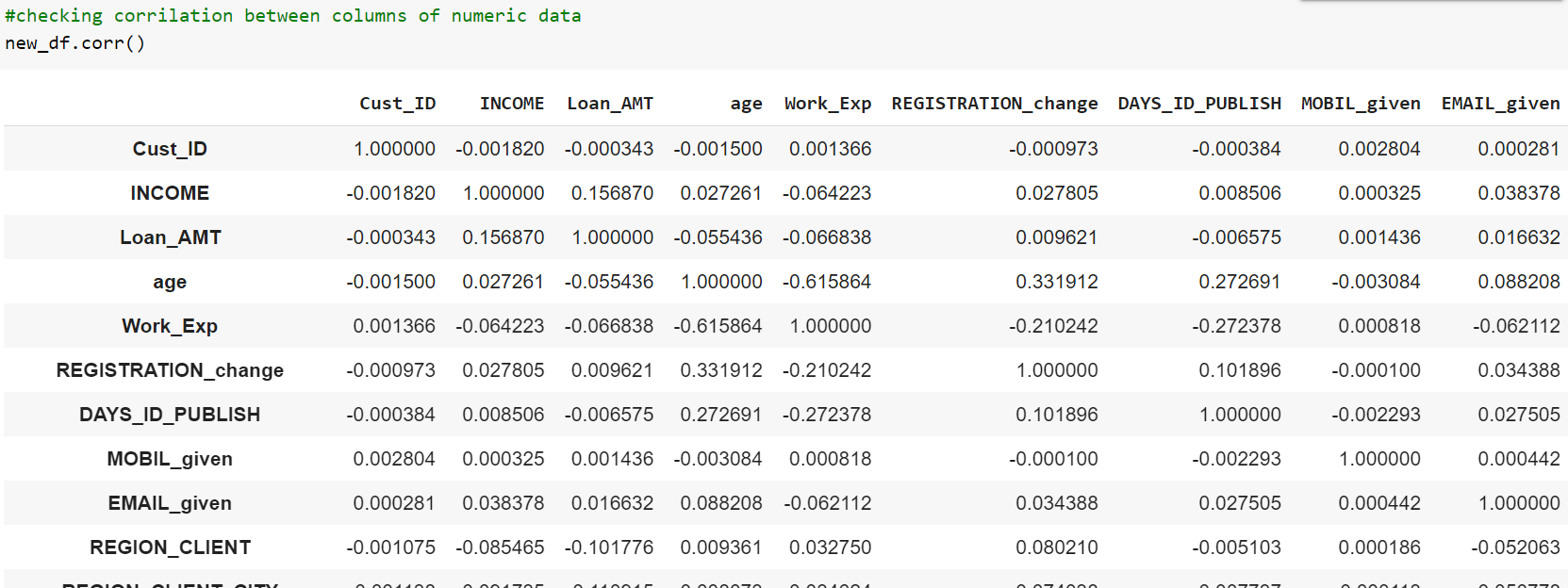


# 

# Checking statistical summary of all the numeric columns

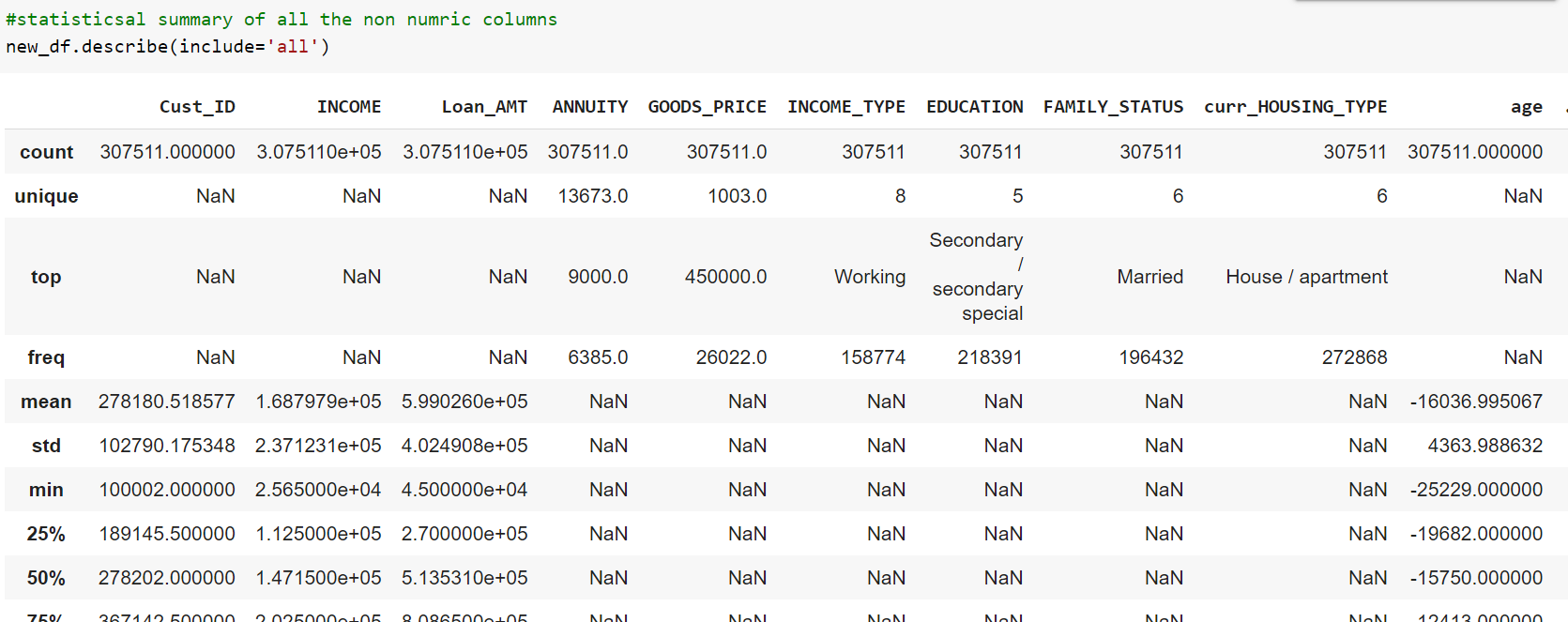


# Checking correlation summary of all the numeric columns



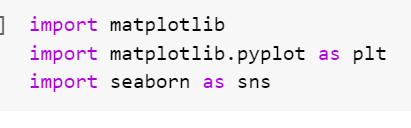
# 

# Checking statistical summary of all the non-numeric columns

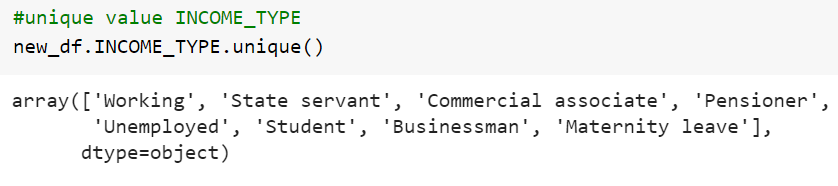


Checking unique values for categorical columns and Visualizing data

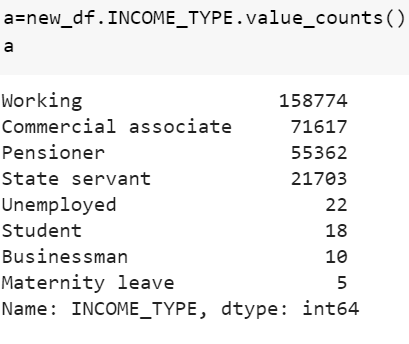
# Importing libraries

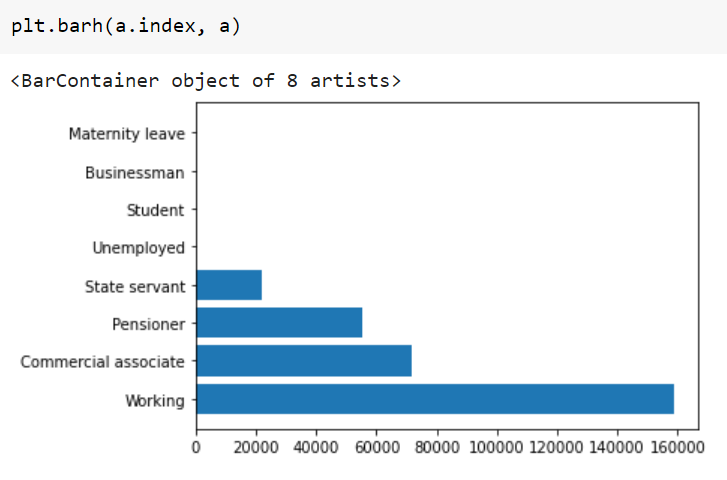


# Unique income type



# Counting and plotting unique income type



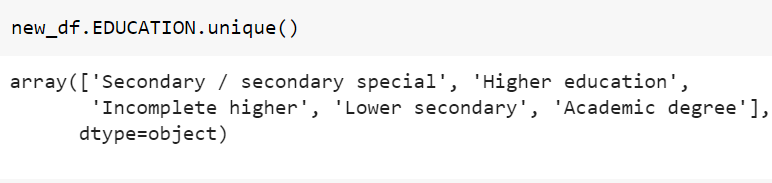


# 

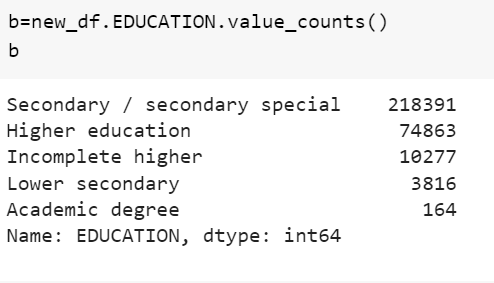
# 

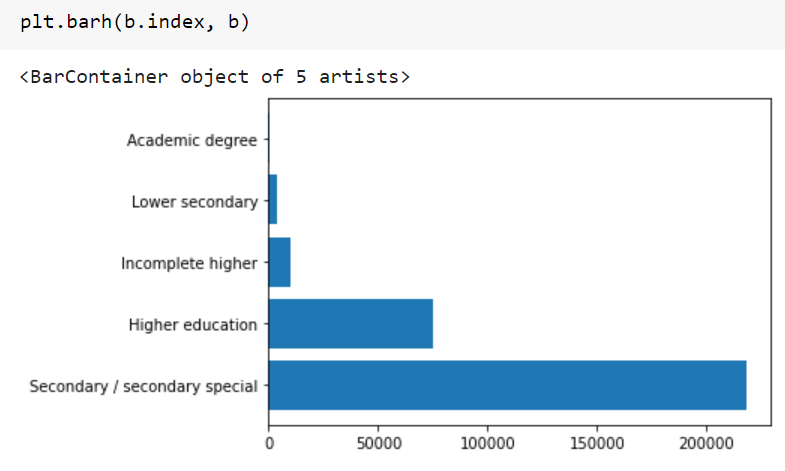
It can be concluded that clients with income type working class are maximum.

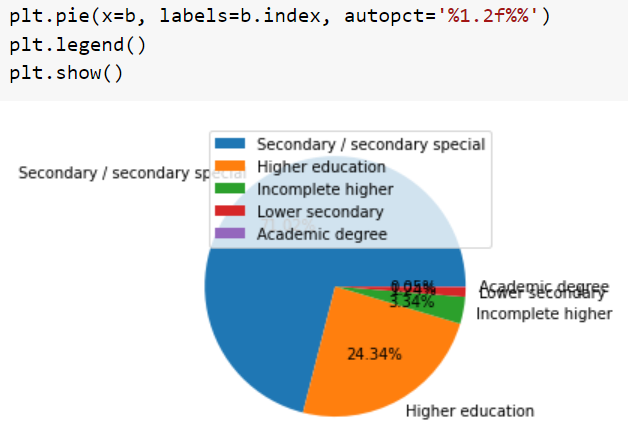
# Unique education type



# Counting and plotting unique education type

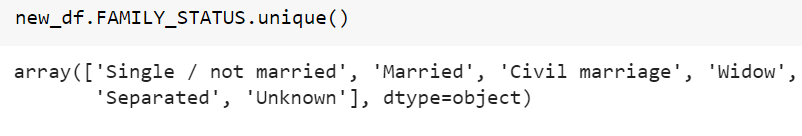




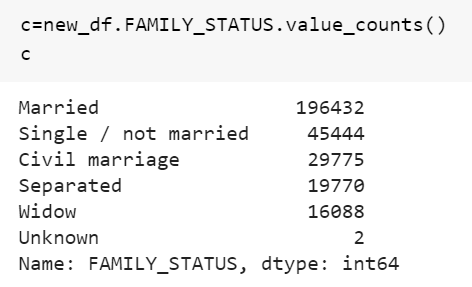


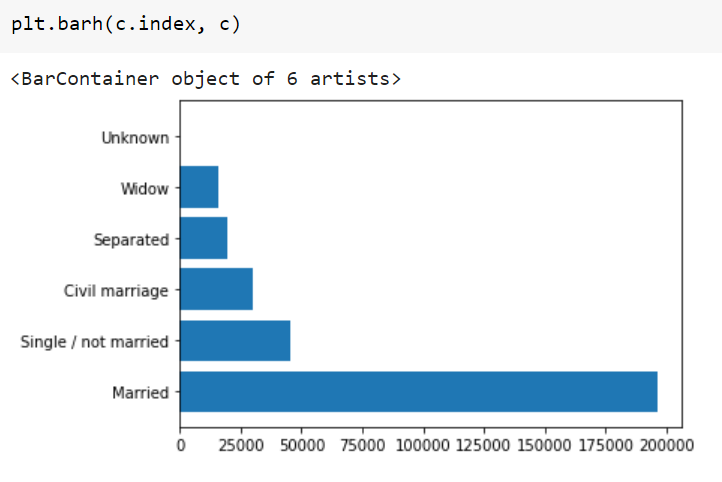
It can be concluded that clients with the education of secondary/secondary special are maximum.

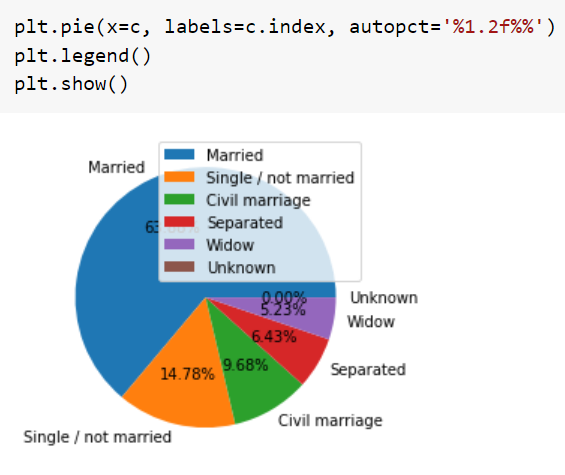
# Unique family status



# Counting and plotting unique family status

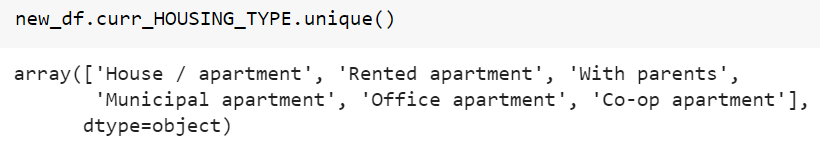




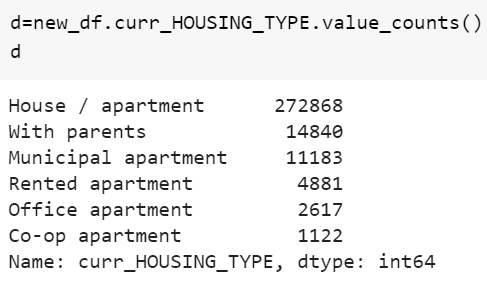


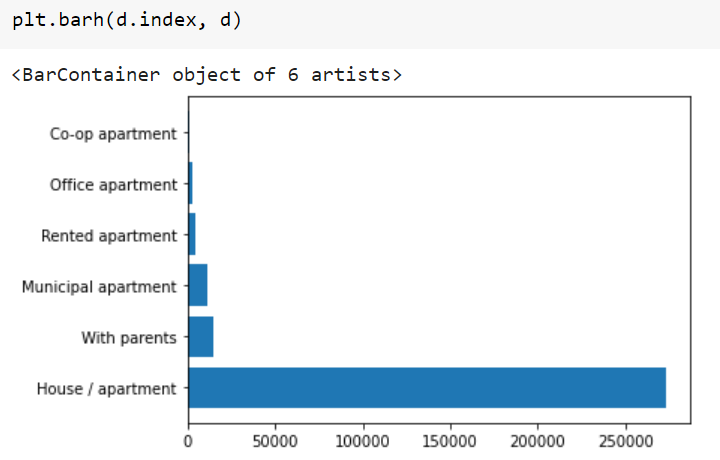
It can be concluded that clients with married family status are maximum.

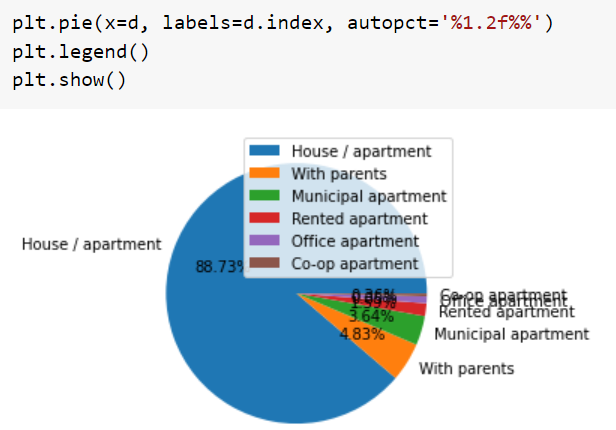
# Unique housing type



# Counting and plotting unique housing type

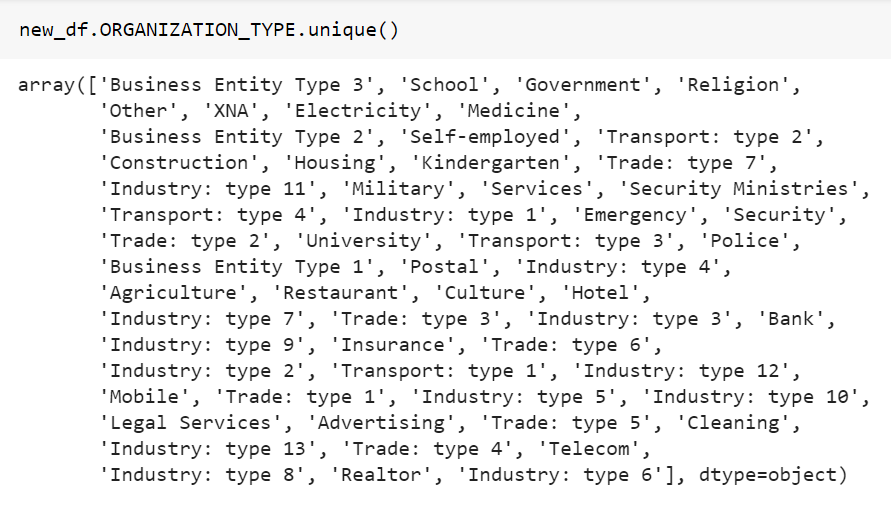




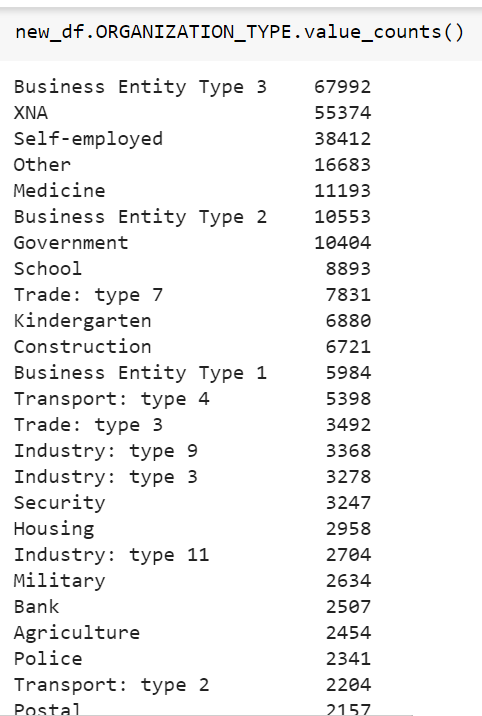


It can be concluded that clients living in housing type of house/apartment are maximum.

# Unique organization type

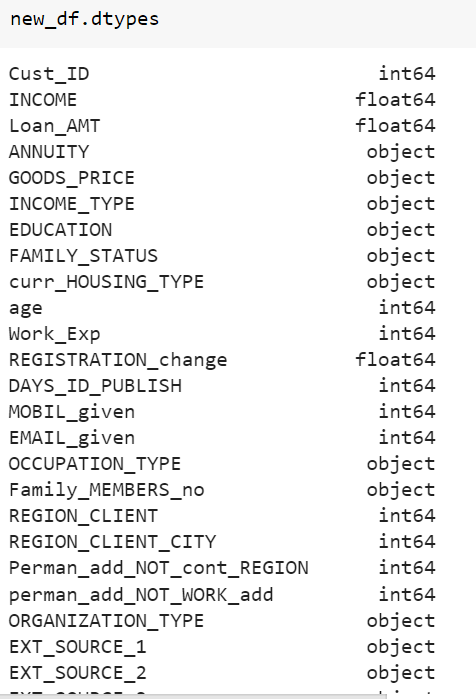


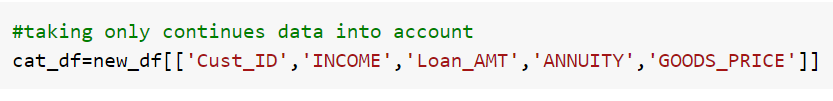
# Counting unique organization type

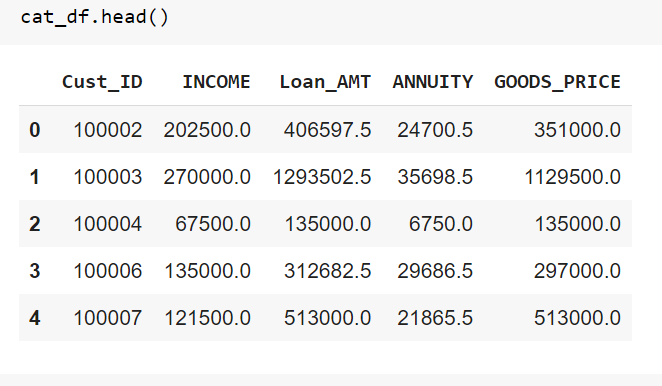


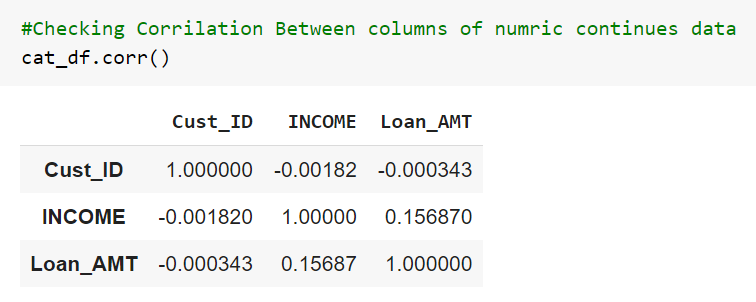
# For categorical data(Analyzing and visualizing)

Checking datatypes of columns for finding continuous data type.



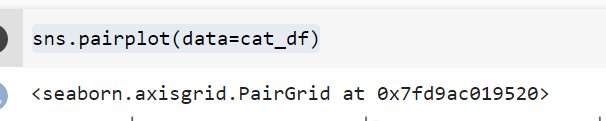


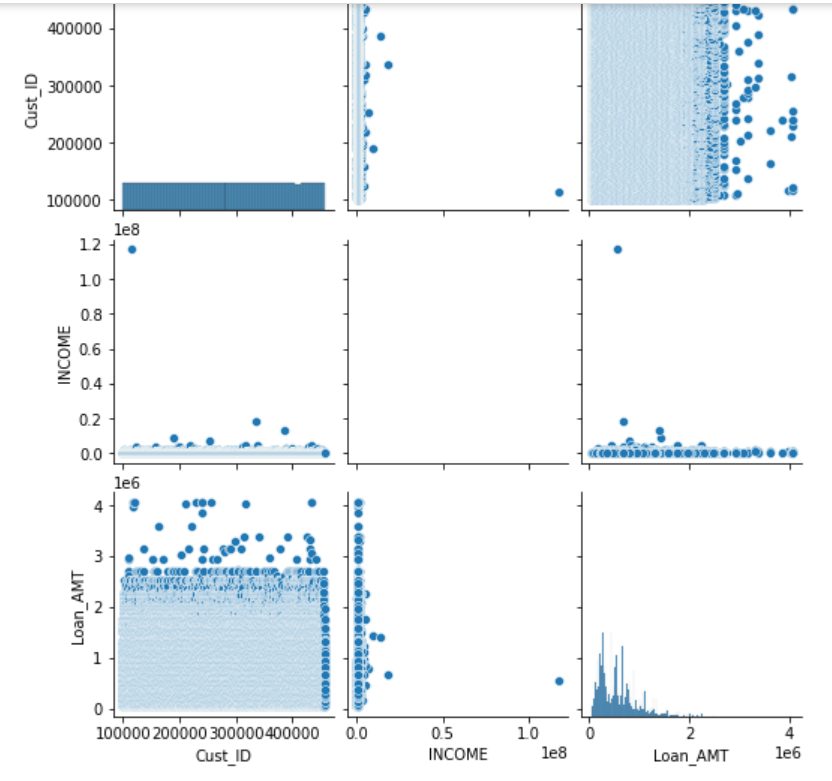




# Plotting pairplot for continuous data to check the relation between columns

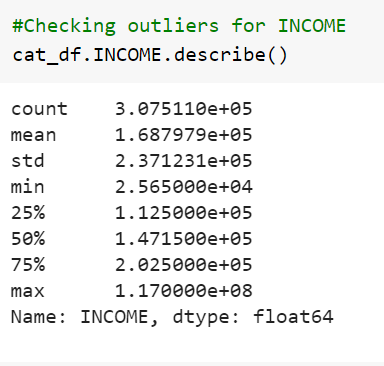
Showing relationship

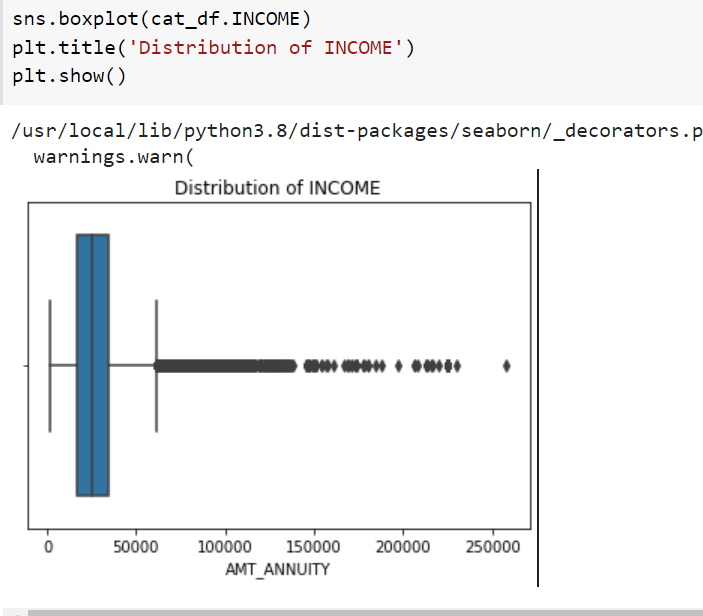


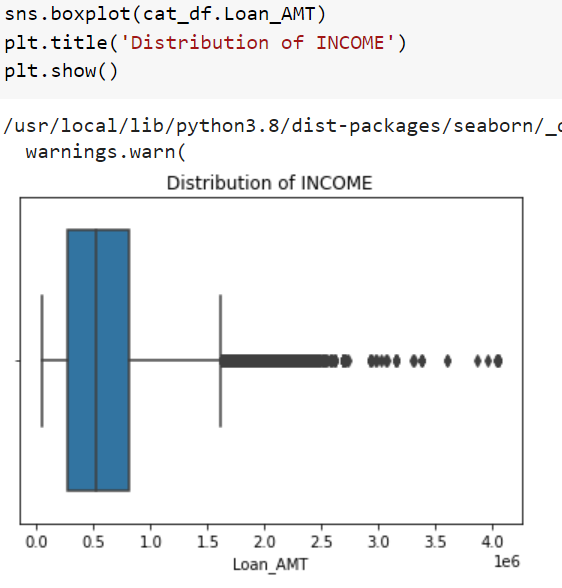
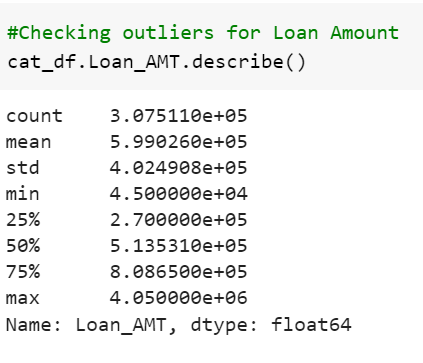


# 

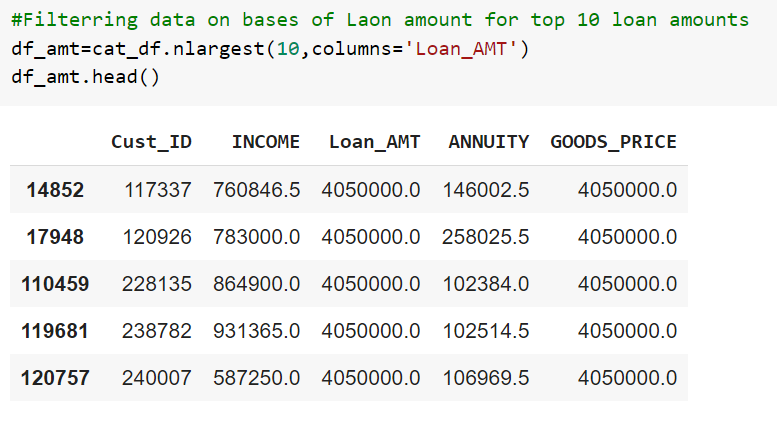
# Handling Outliers



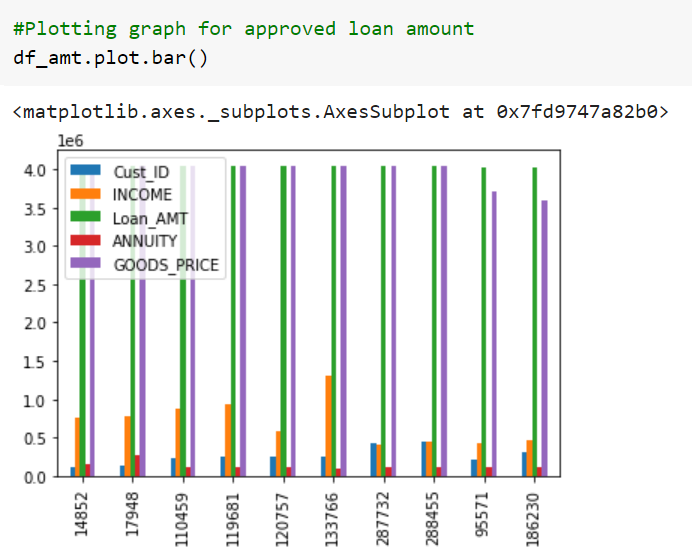




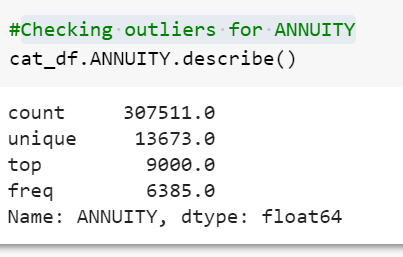
Filtering data on bases of Loan amount for top 10 loan amounts



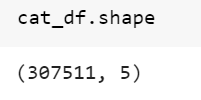
## Plotting graph for approved loan amount(analyzing and visualizing data)



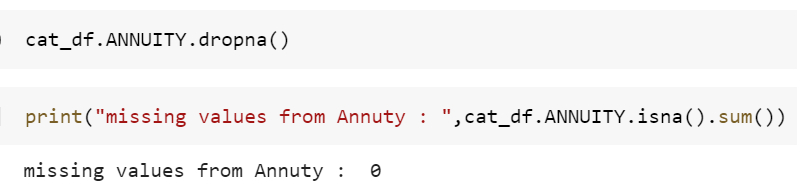
## Checking outliers for ANNUITY



Checking shape



Dropping annuity

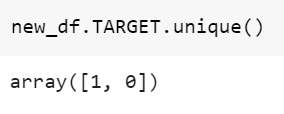


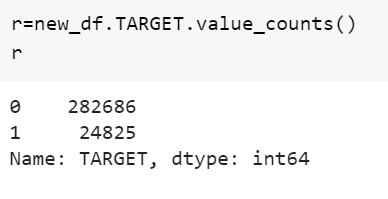
# Checking outliers for GOODS\_PRICE

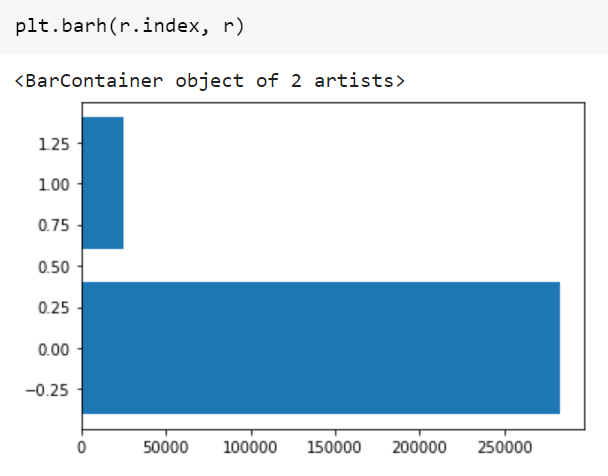
# 

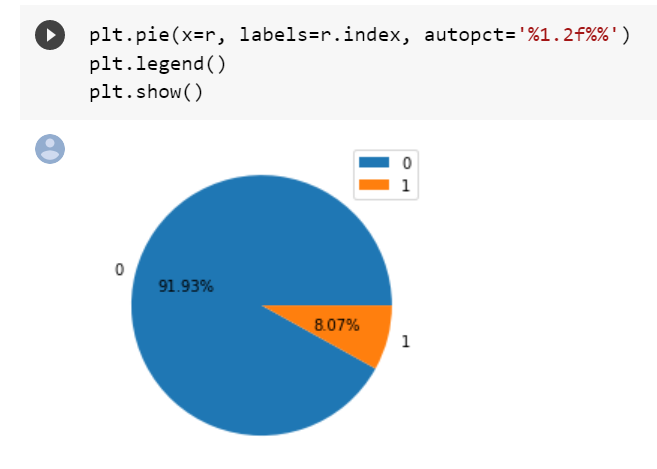
# 

# **Find the top 10 correlation for the Client with payment difficulties and all other cases (Target variable).**



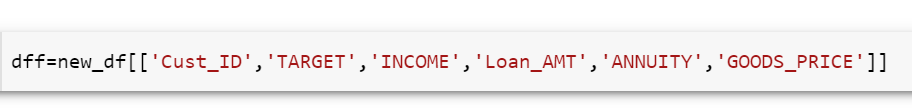


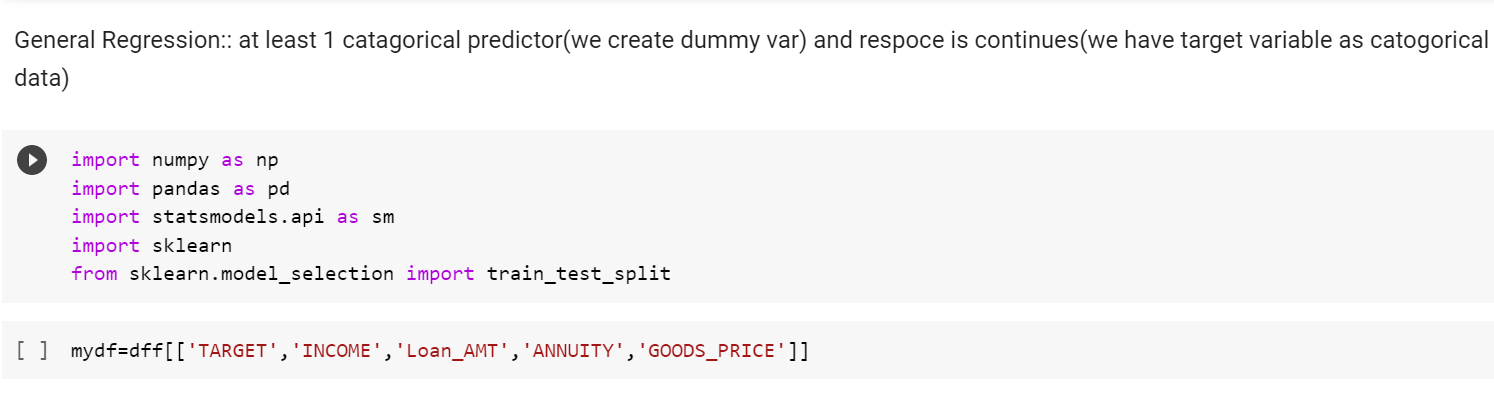




## Conclusion:: 8.07% of people have payment difficulty

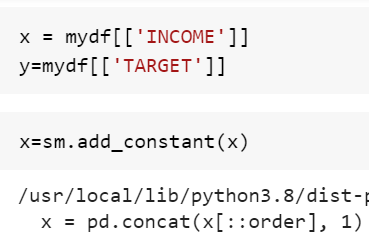
# **Binary Logistic Regression(taking into account Target as response variable)**

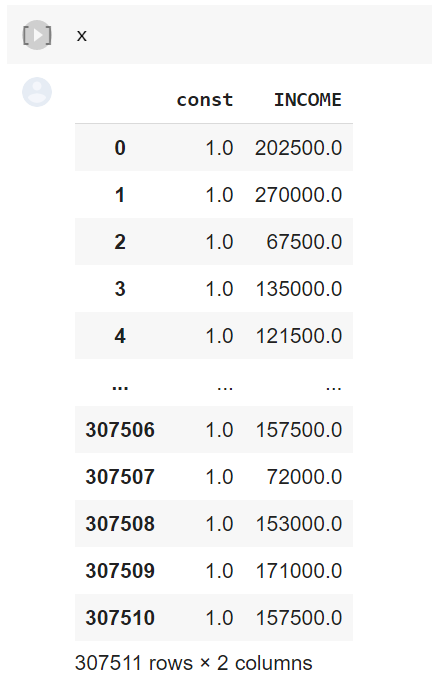


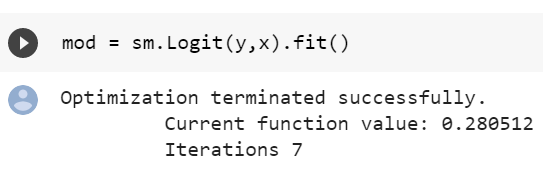


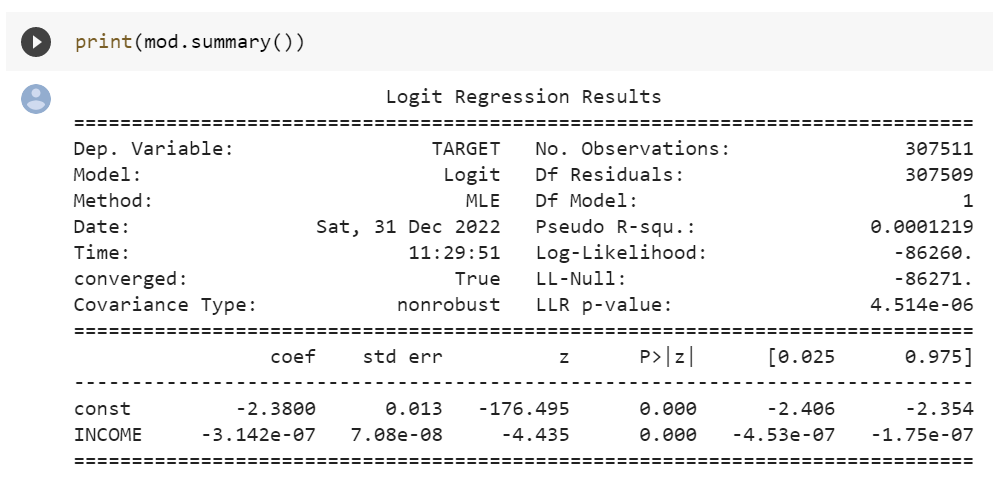


# Checking for Target var vs Income



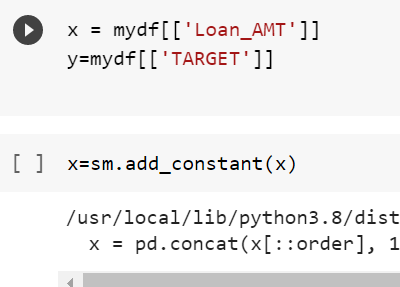


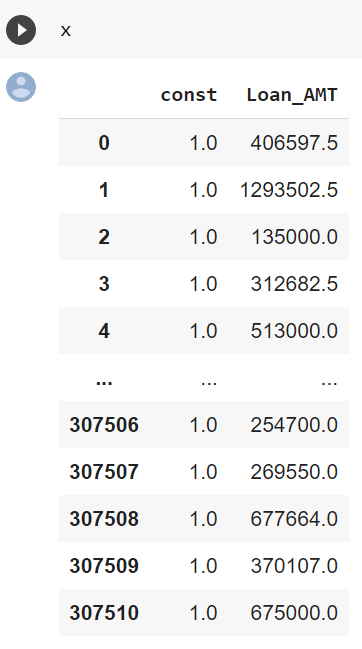


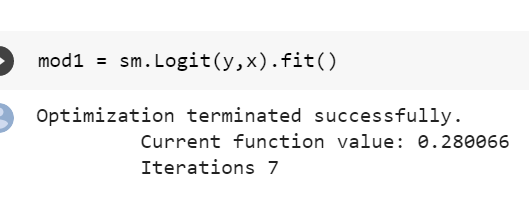


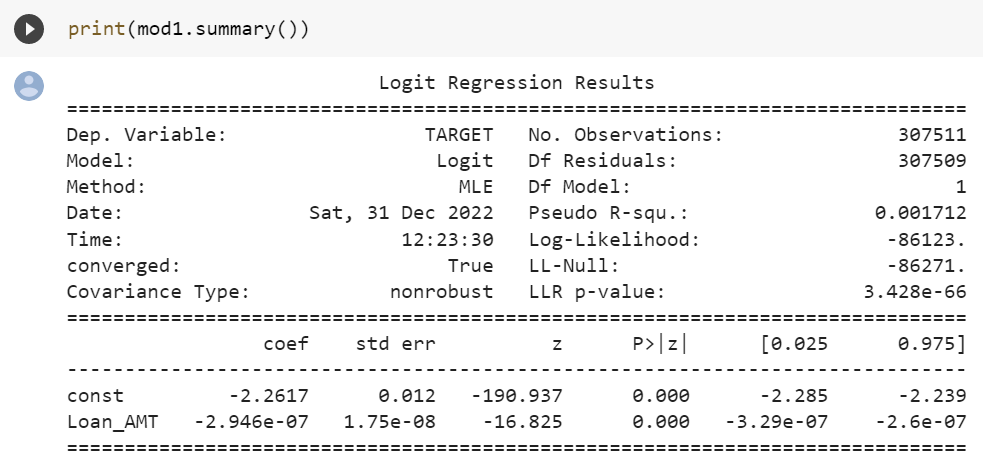
p=value is <0.05 so Ho is rejected and it means that Income affect the person will default or not on loan repayment

# Checking for Target Var vs Loan\_AMT









p=value is <0.05 so Ho is rejected and it means that Income affect the person will default or not on loan repayment