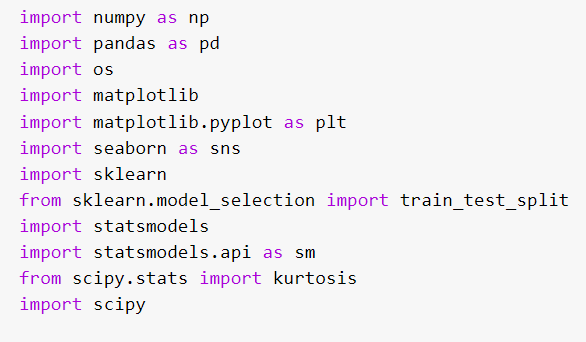
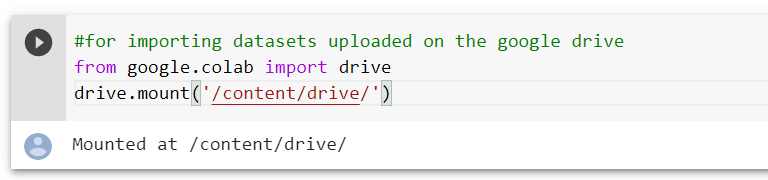
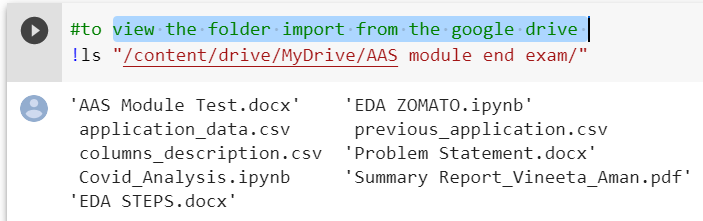
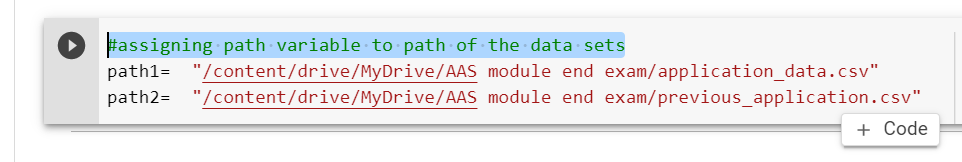
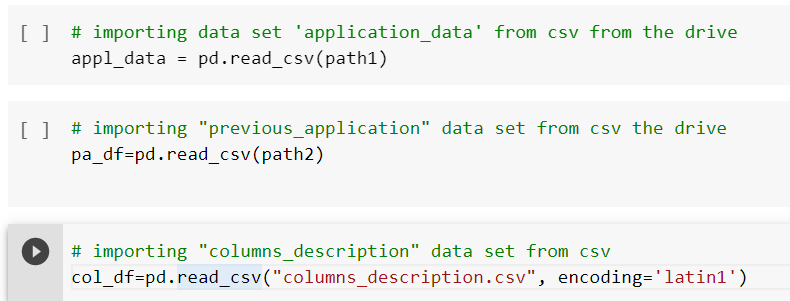
Importing required data

****

**importing datasets uploaded on the google driveview the folder import from the google drive**

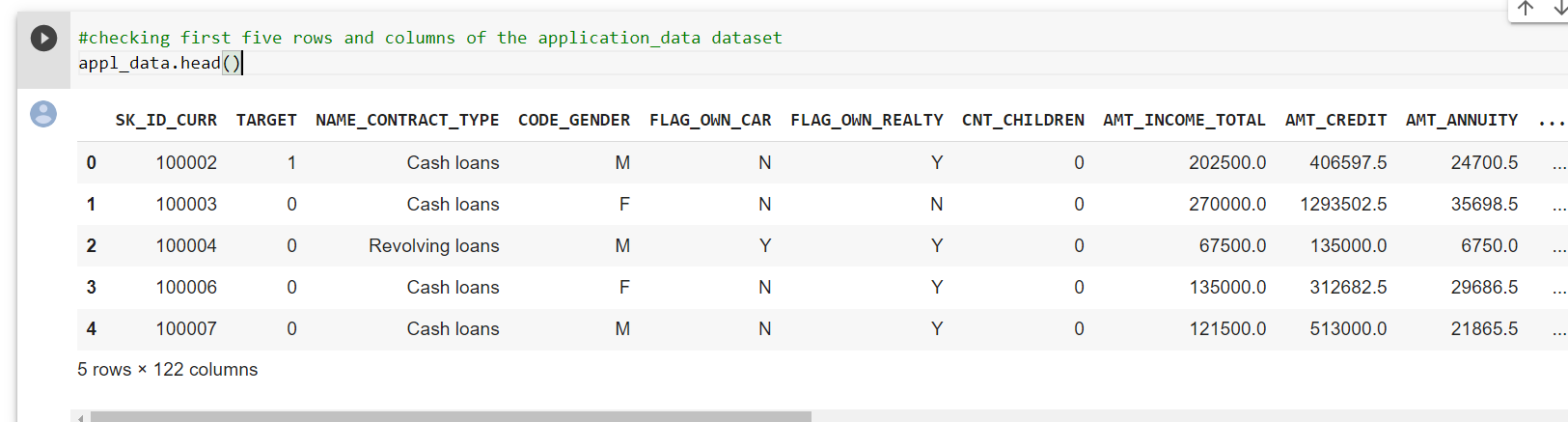
**assigning path variable to path of the data sets**

**Importing Data from CSV**

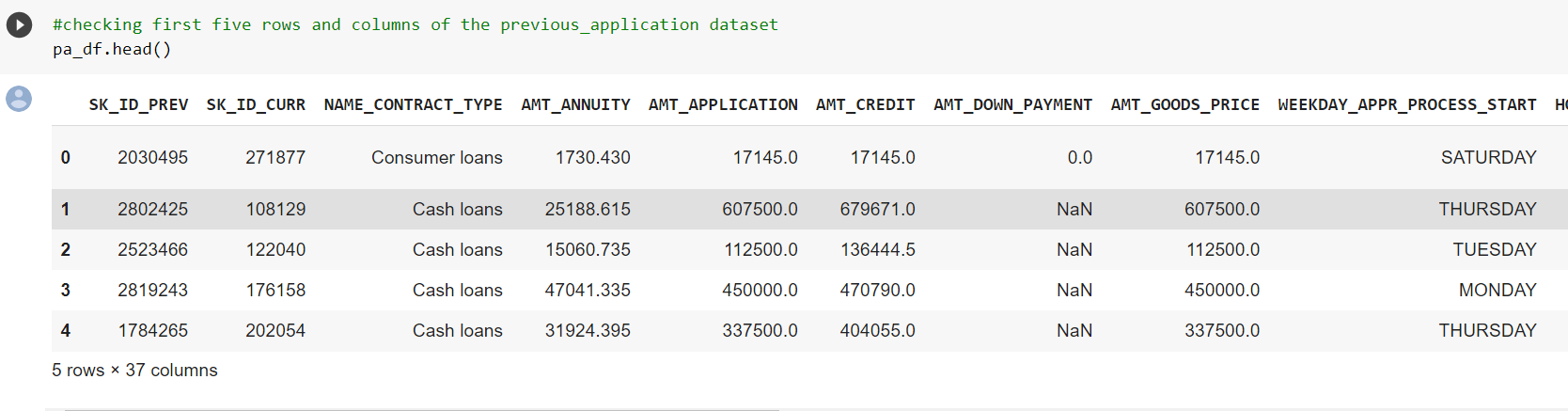
****

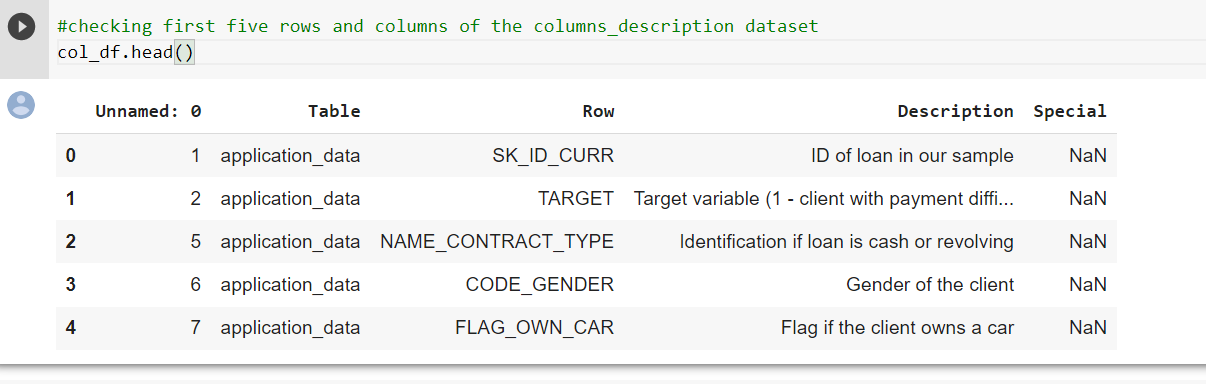
# Viewing the Dataset

checking first five rows and columns of the application\_data dataset

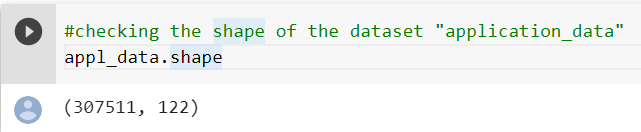
****

checking first five rows and columns of the previous\_application dataset

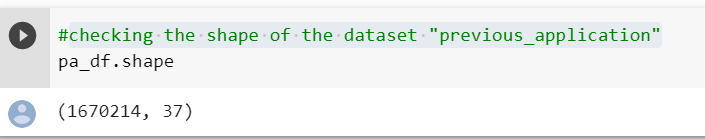
****

checking first five rows and columns of the columns\_description dataset

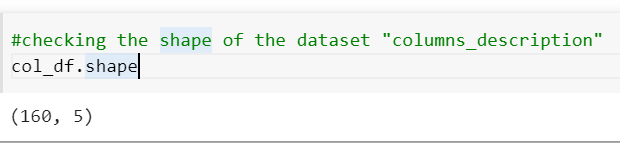
checking the shape of the dataset "application\_data"

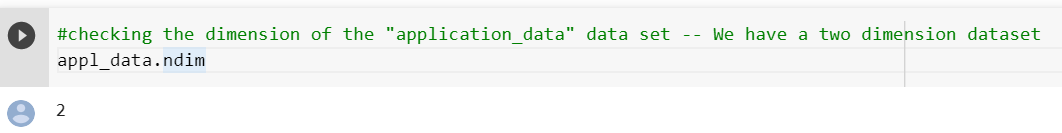


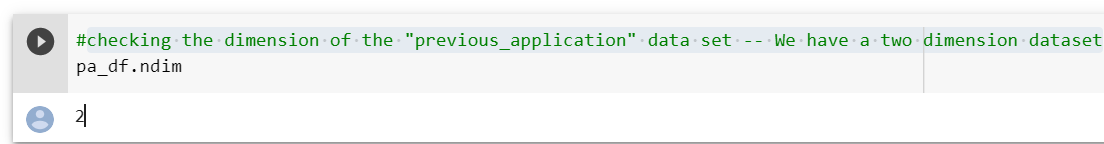
checking the shape of the dataset "previous\_application"

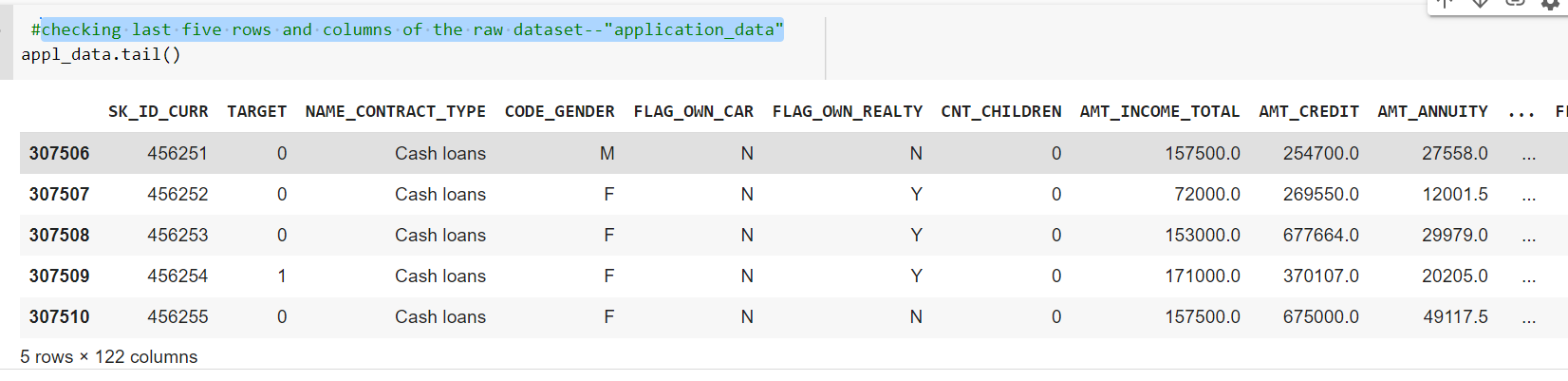


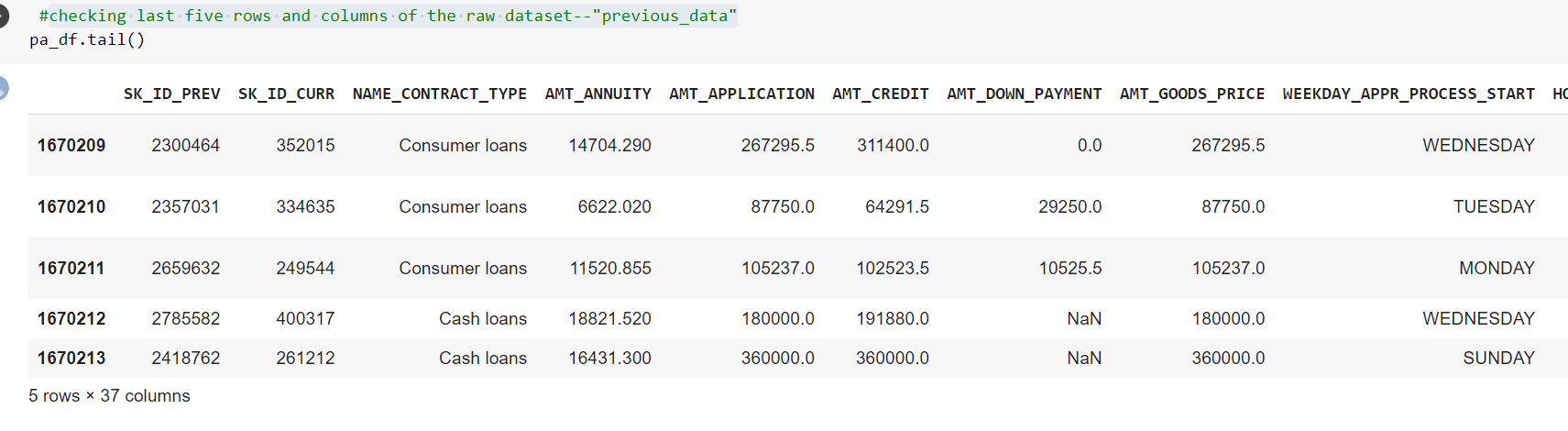
checking the shape of the dataset "columns\_description"



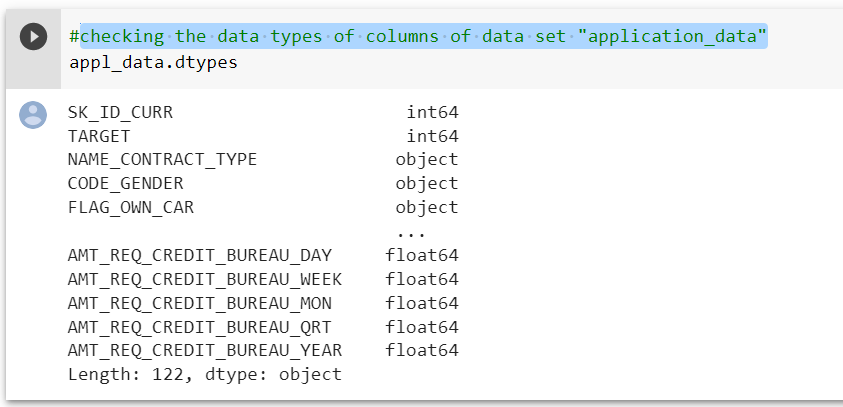
checking the dimension of the "application\_data" data set -- We have a two dimension dataset

checking the dimension of the "previous\_application" data set -- We have a two dimension dataset

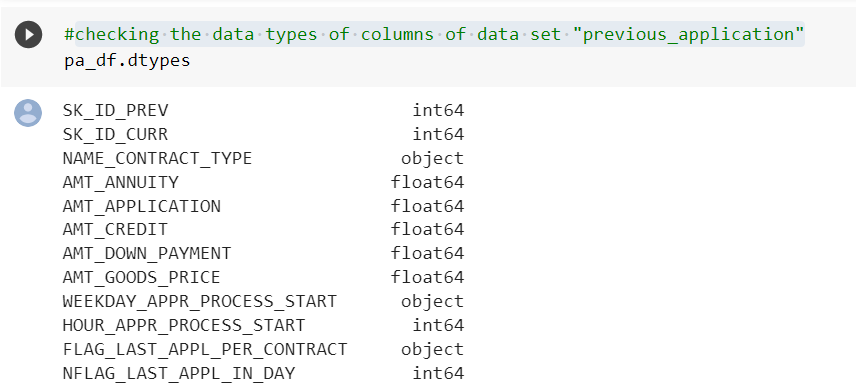
checking last five rows and columns of the raw dataset--"application\_data"checking last five rows and columns of the raw dataset--"previous\_data"



checking the data types of columns of data set "application\_data"



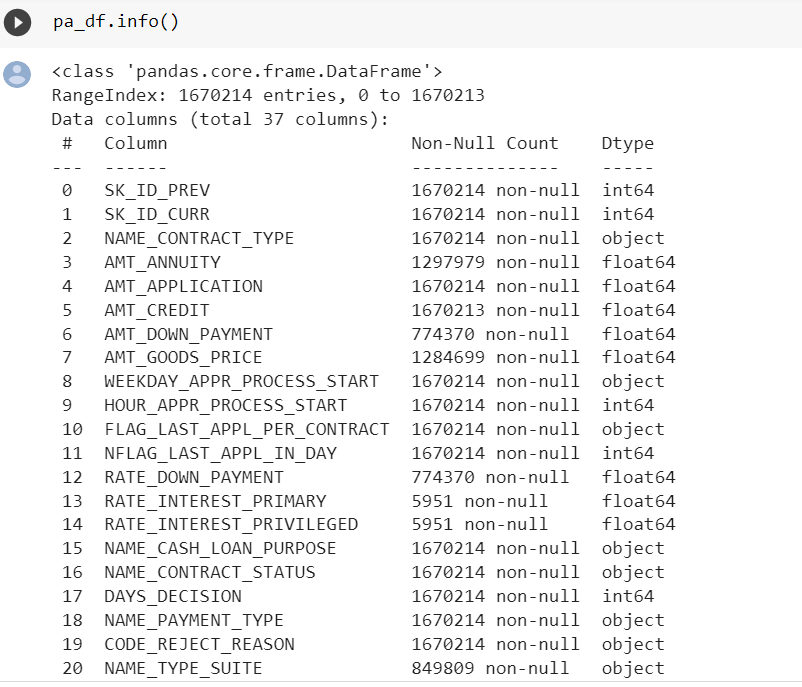
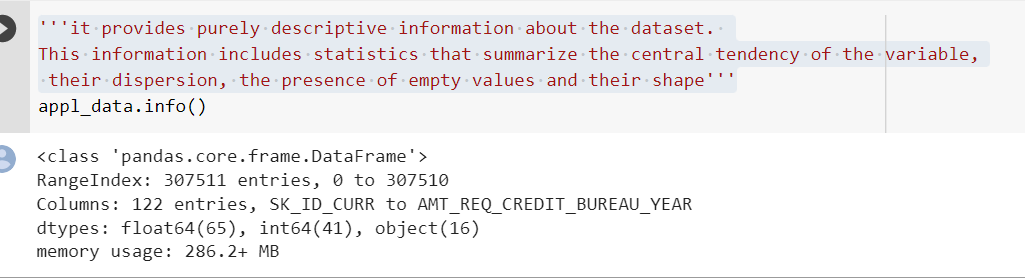
checking the data types of columns of data set "previous\_application"

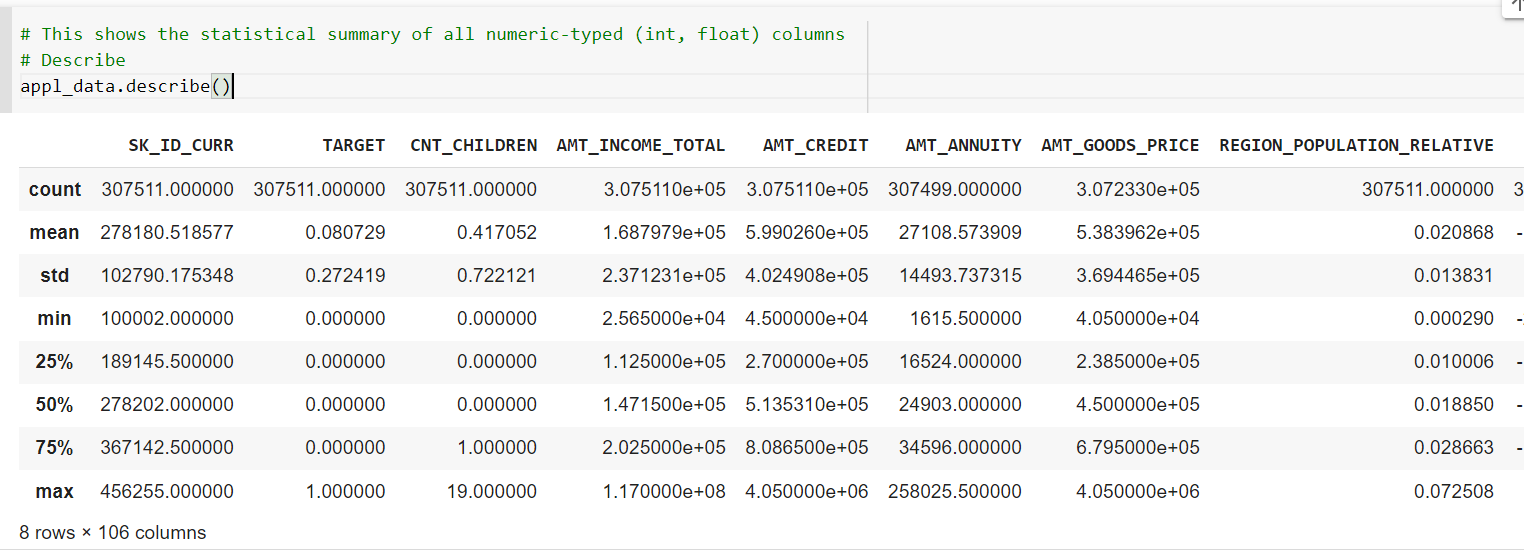


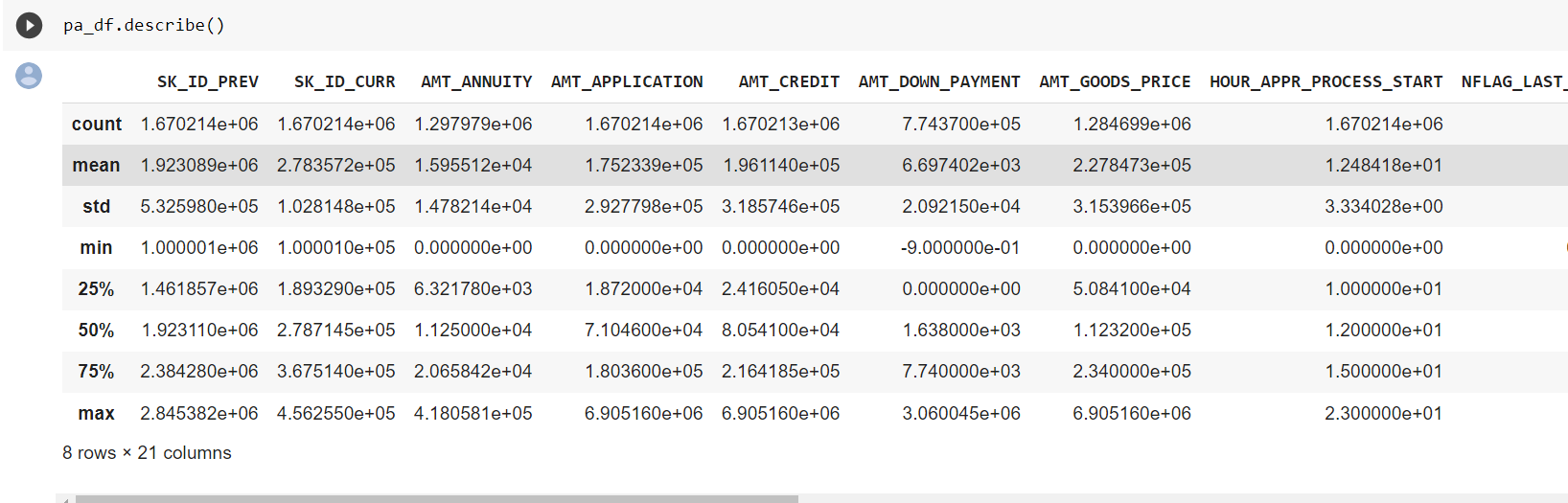
'''it provides purely descriptive information about the dataset.

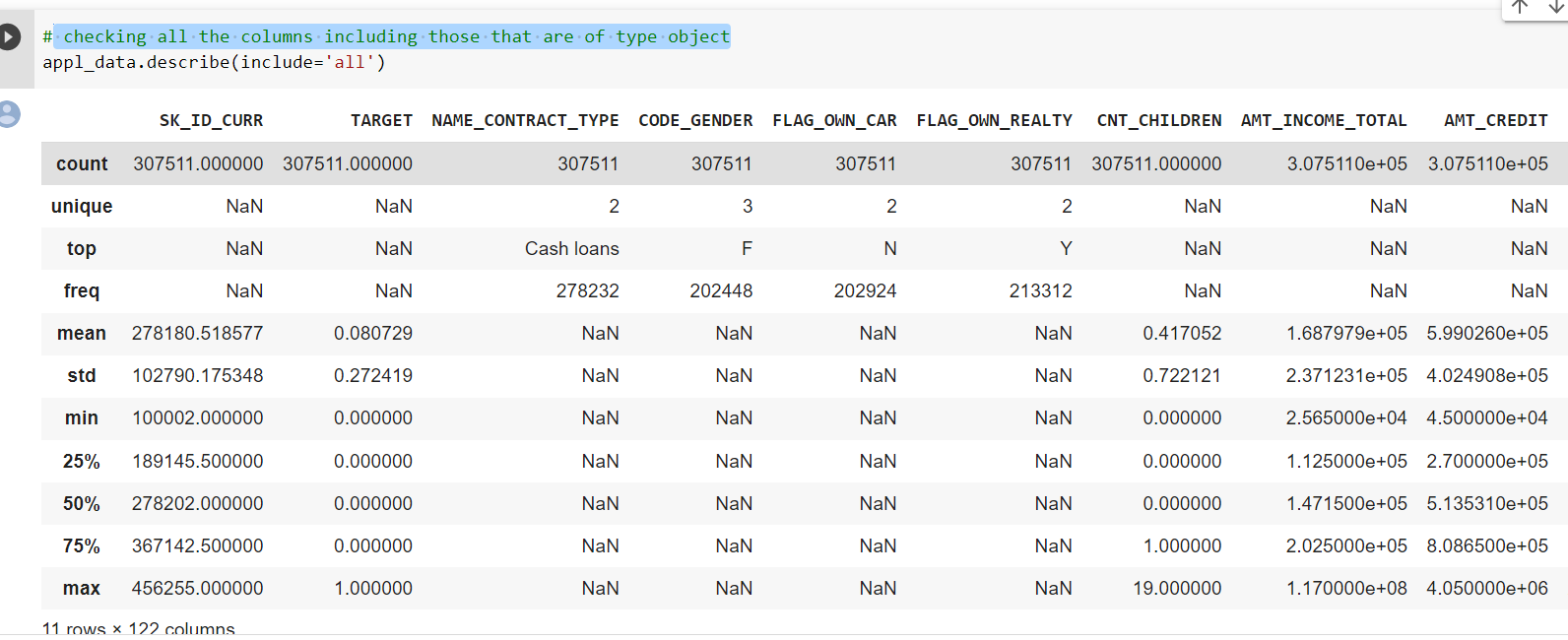
This information includes statistics that summarize the central tendency of the variable,

their dispersion, the presence of empty values and their shape'''

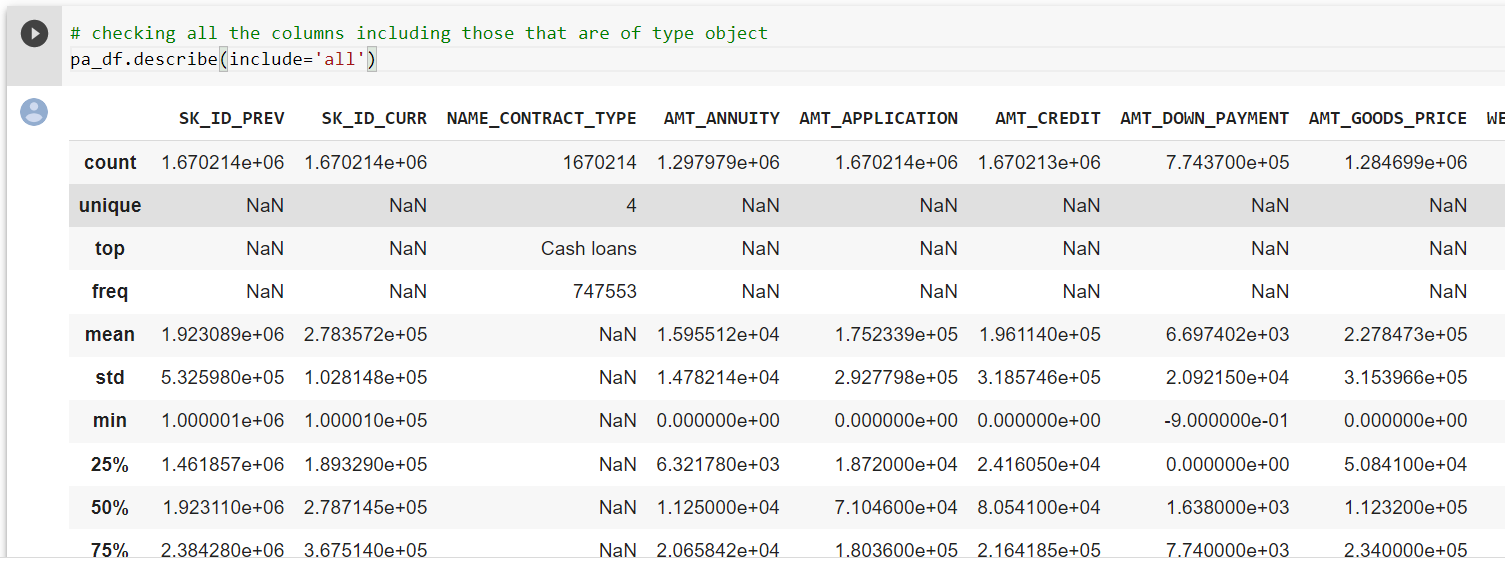
This shows the statistical summary of all numeric-typed (int, float) columns



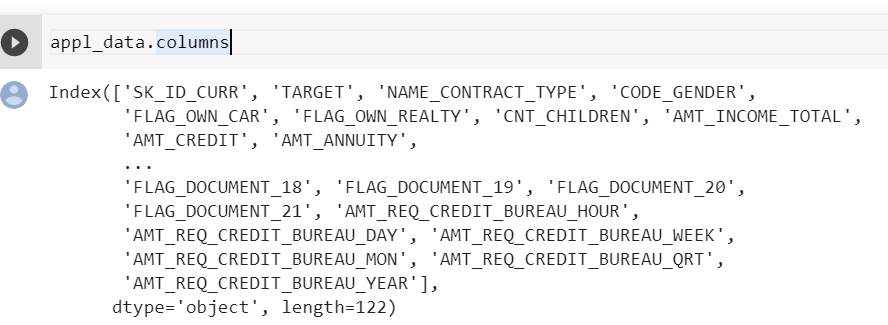
 checking all the columns including those that are of type object



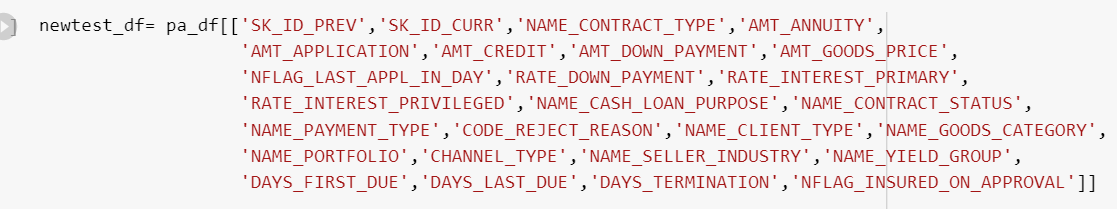
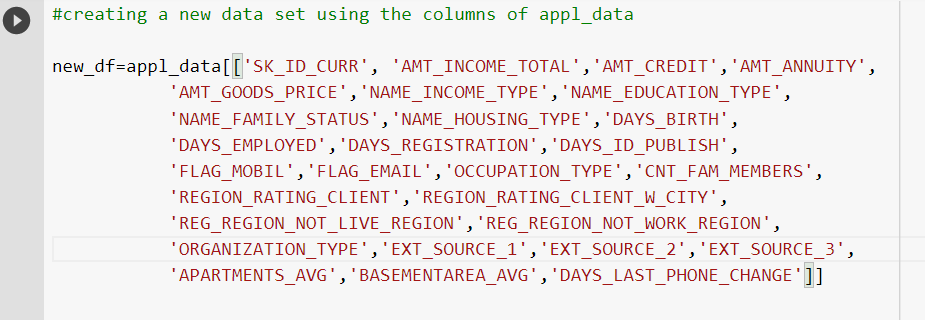
checking all the columns including those that are of type object

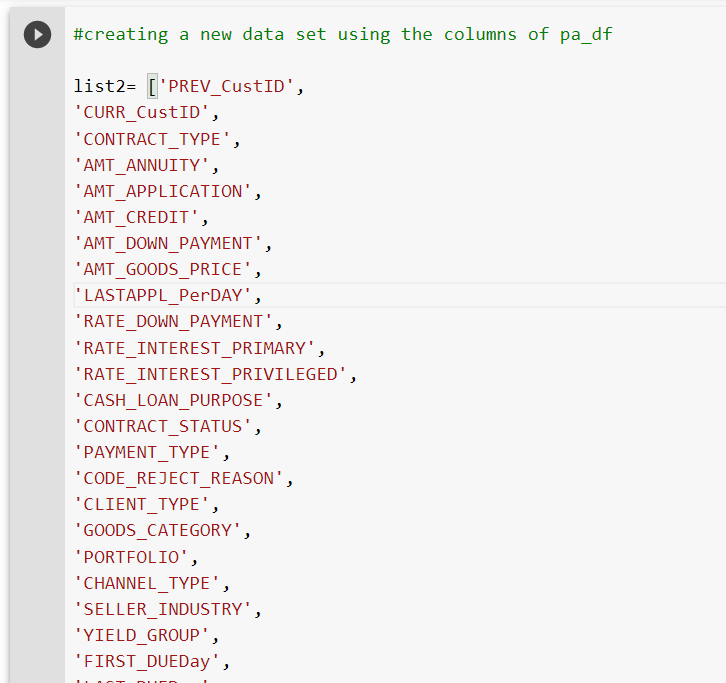


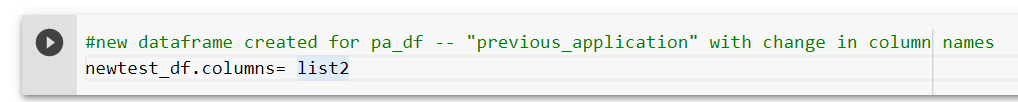
# Data Preparation

As data is having huge number of lines, we are considering applicable columns only by looking at data set

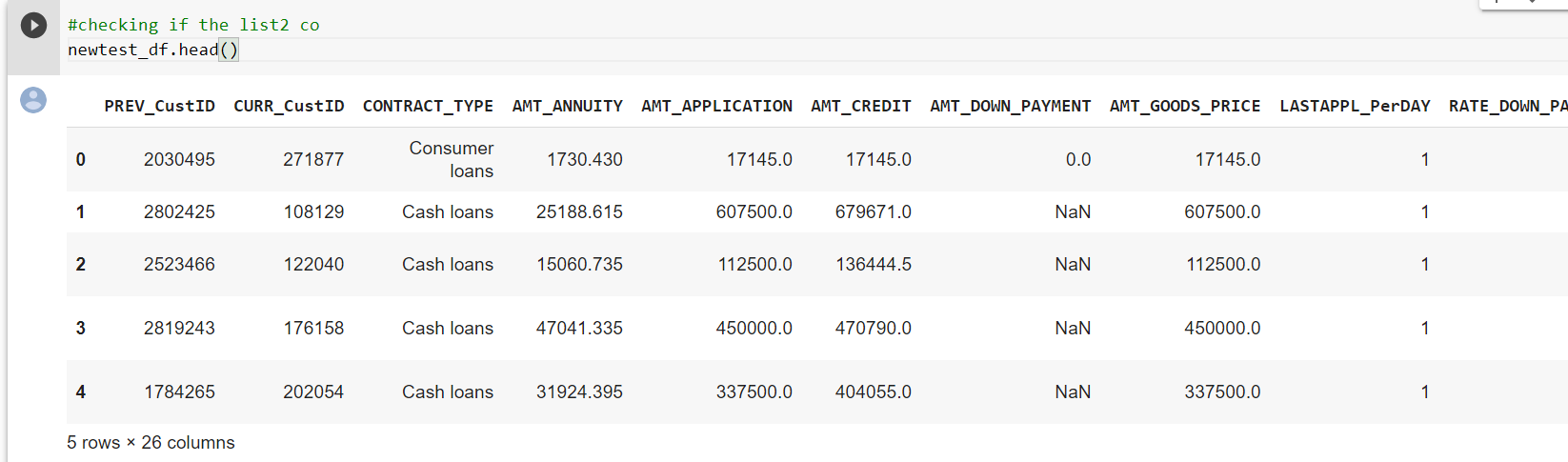
creating a new data set using the columns of appl\_data

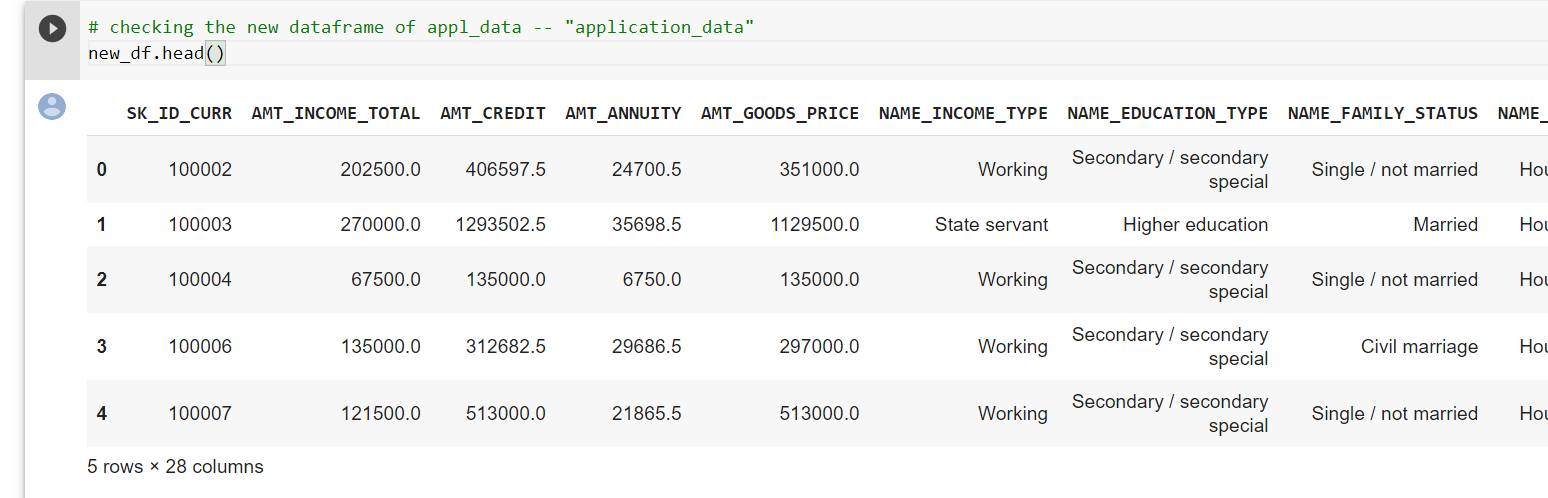


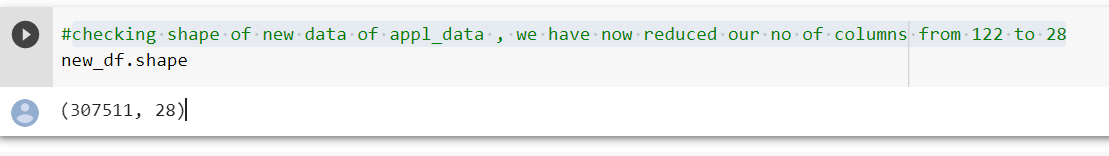
creating a new data set using the columns of pa\_df

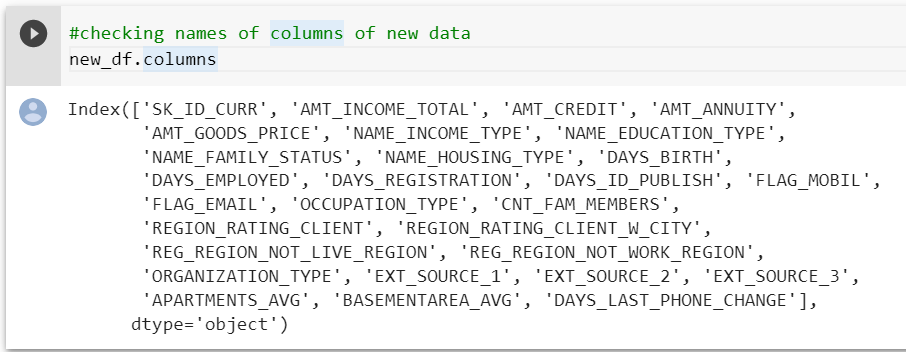
new dataframe created for pa\_df -- "previous\_application" with change in column names 

checking if the list2 co

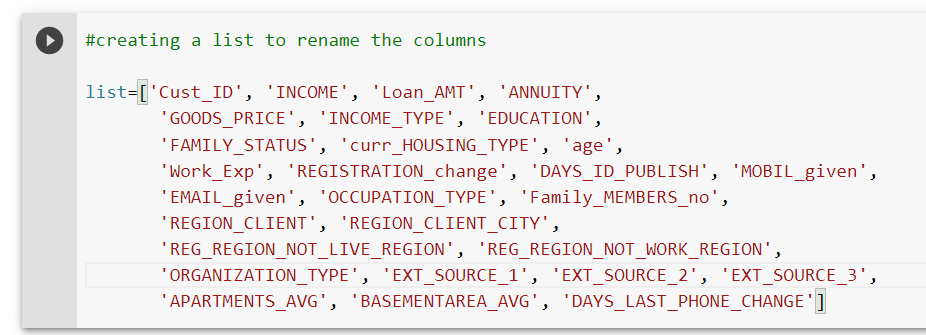
checking the new dataframe of appl\_data -- "application\_data"

checking shape of new data of appl\_data , we have now reduced our no of columns from 122 to 28

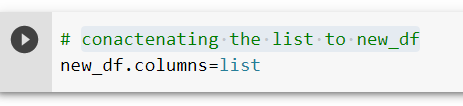


checking names of columns of new datarenaming the column names

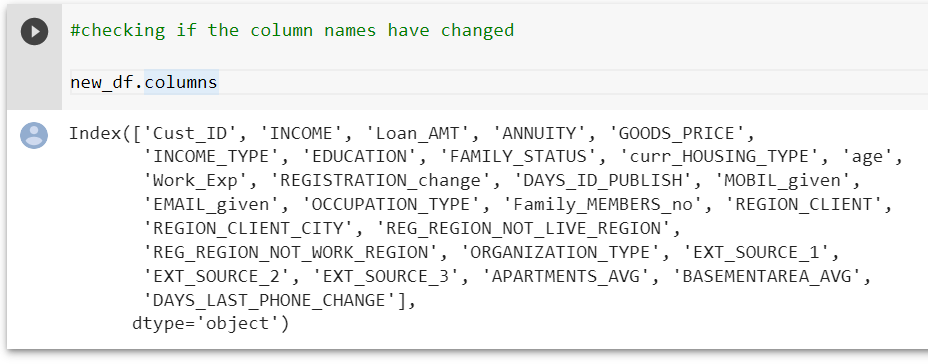
creating a list to rename the columns



conactenating the list to new\_df



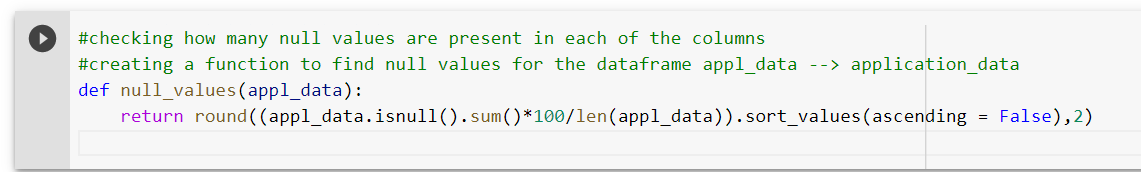
checking if the column names have changed



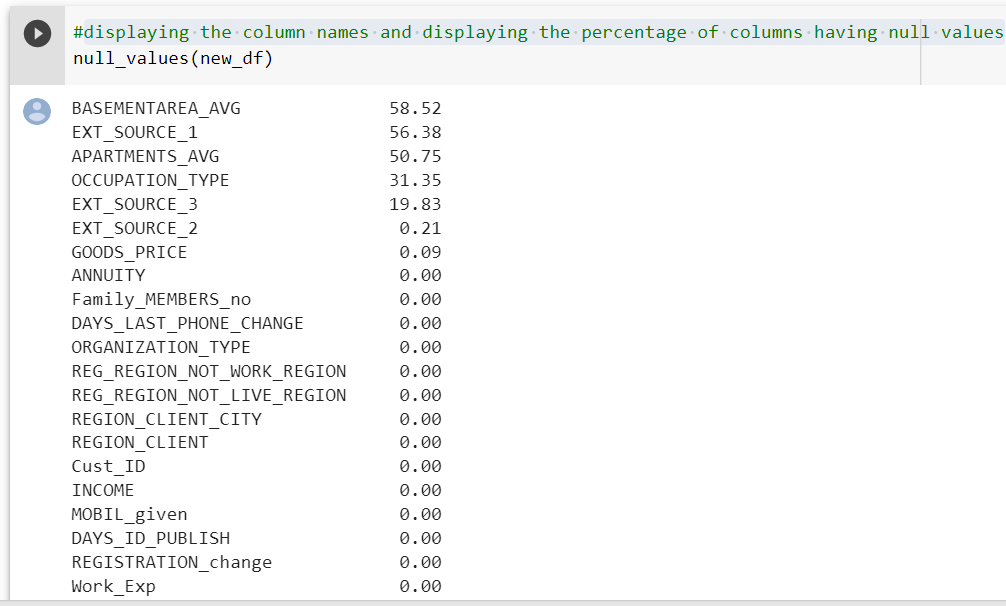
checking for columns with null values

checking how many null values are present in each of the columns

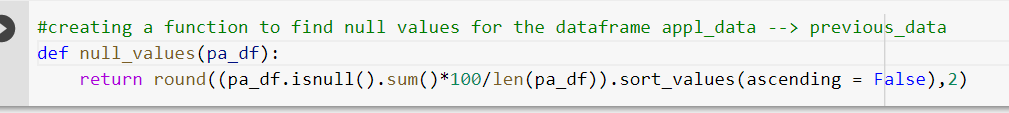
creating a function to find null values for the dataframe appl\_data --> application\_data



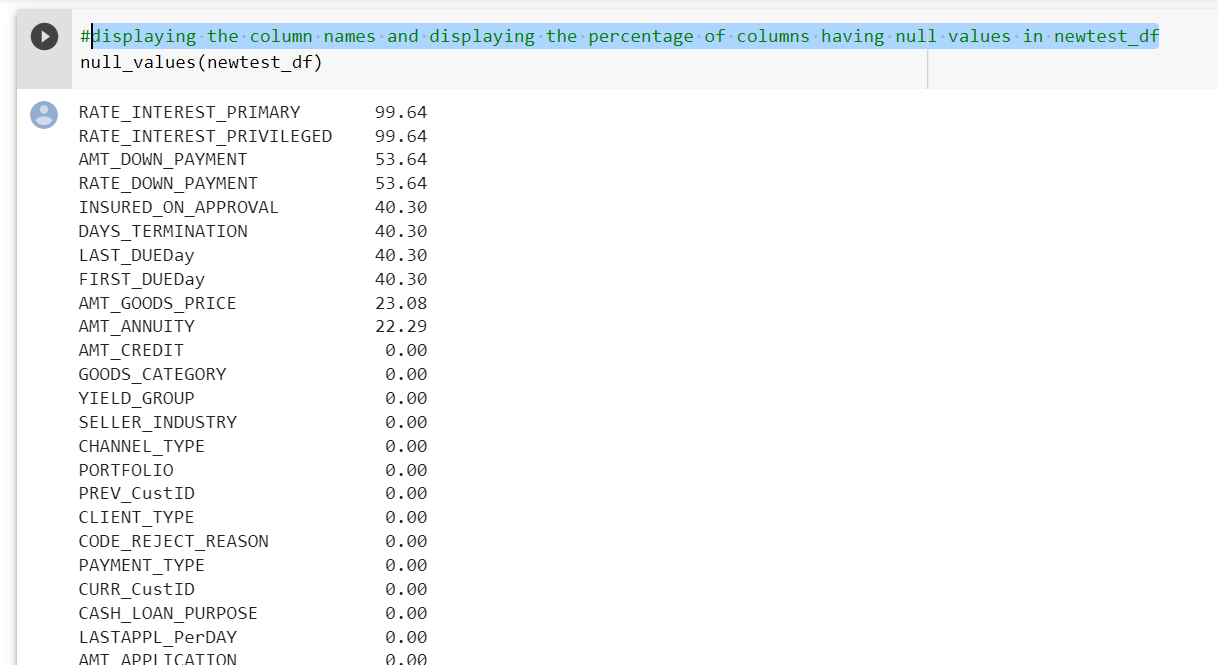
displaying the column names and displaying the percentage of columns having null values



creating a function to find null values for the dataframe appl\_data --> previous\_data



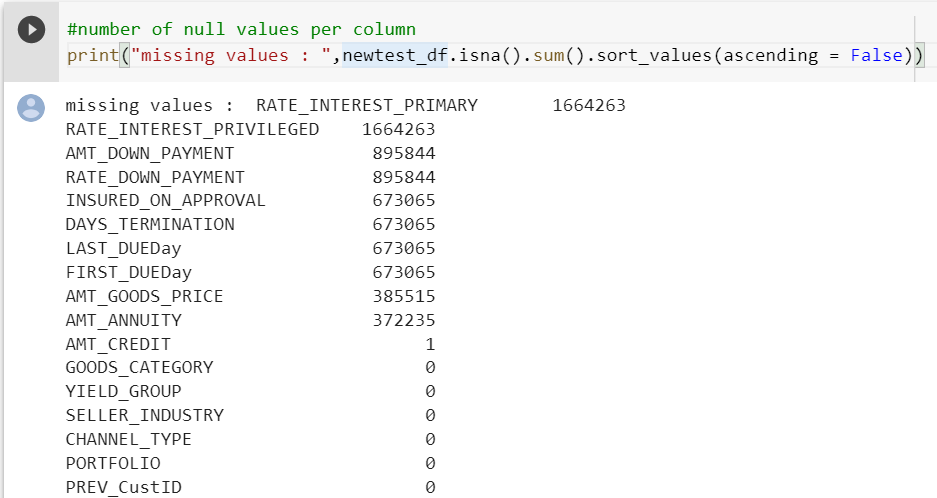
displaying the column names and displaying the percentage of columns having null values in newtest\_df

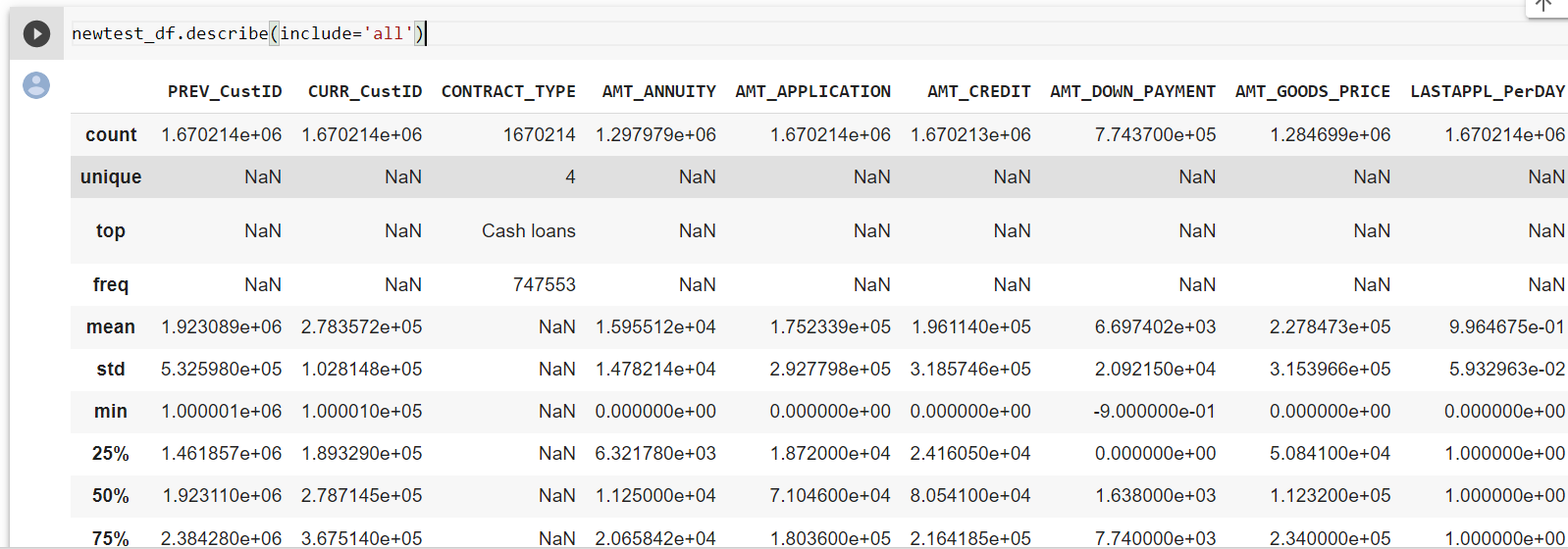


these are the columns having maximum null values

1. RATE\_INTEREST\_PRIMARY 99.64
2. RATE\_INTEREST\_PRIVILEGED 99.64
3. AMT\_DOWN\_PAYMENT 53.64
4. RATE\_DOWN\_PAYMENT 53.64
5. INSURED\_ON\_APPROVAL 40.30
6. DAYS\_TERMINATION 40.30
7. LAST\_DUEDay 40.30
8. FIRST\_DUEDay 40.30
9. AMT\_GOODS\_PRICE 23.08
10. AMT\_ANNUITY 22.29

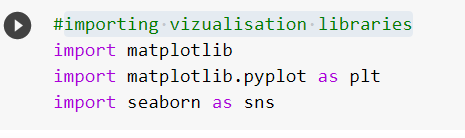
number of null values per column



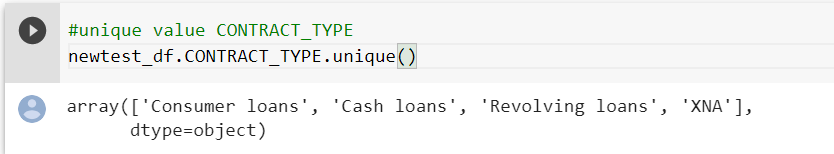


# checking unique values for categorical columns and visualizing data

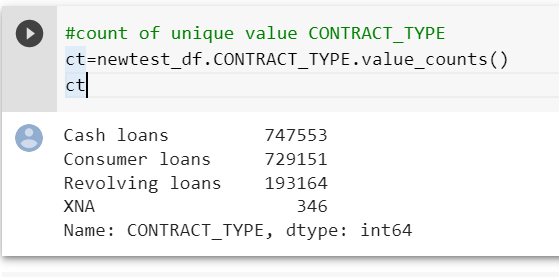
importing visualization libraries



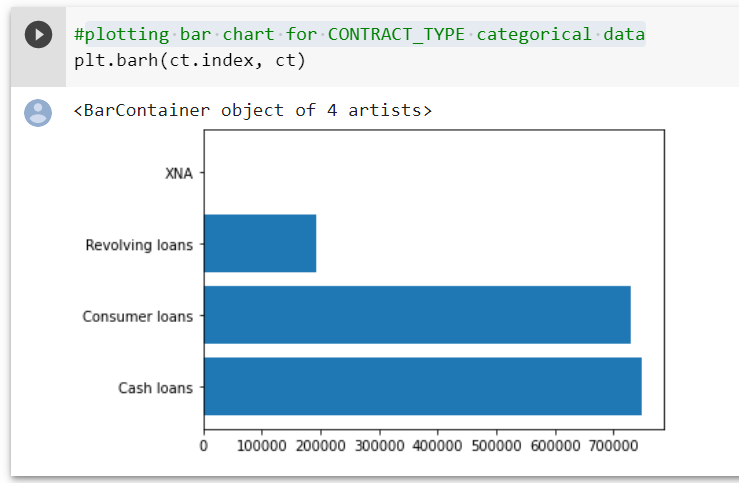
unique value CONTRACT\_TYPE



count of unique value CONTRACT\_TYPE



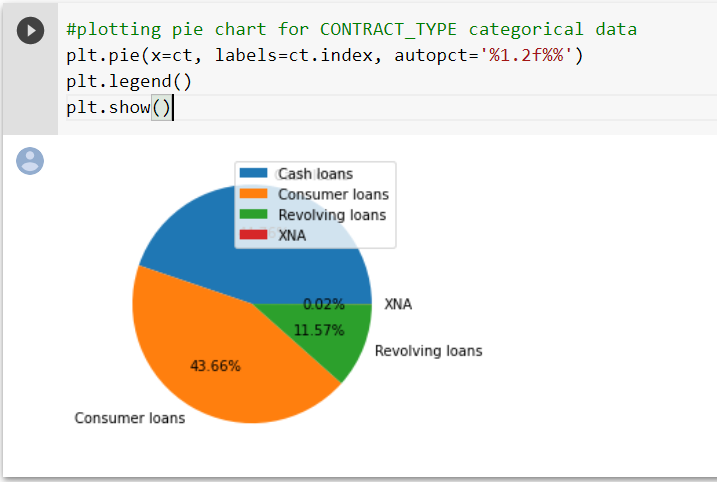
plotting bar chart for CONTRACT\_TYPE categorical data



Conclusion:

plot clearly shows that Customer taking cash loans and customer loans are more than compared to those taking revolving loans

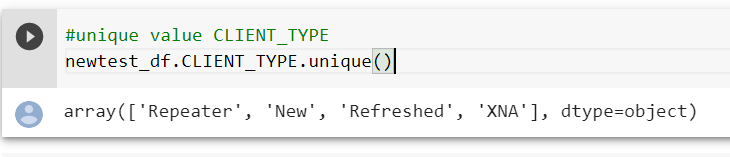
plotting pie chart for CONTRACT\_TYPE categorical data



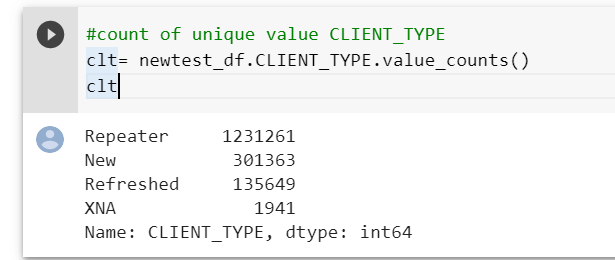
Conclusion:

Client taking cash loan more

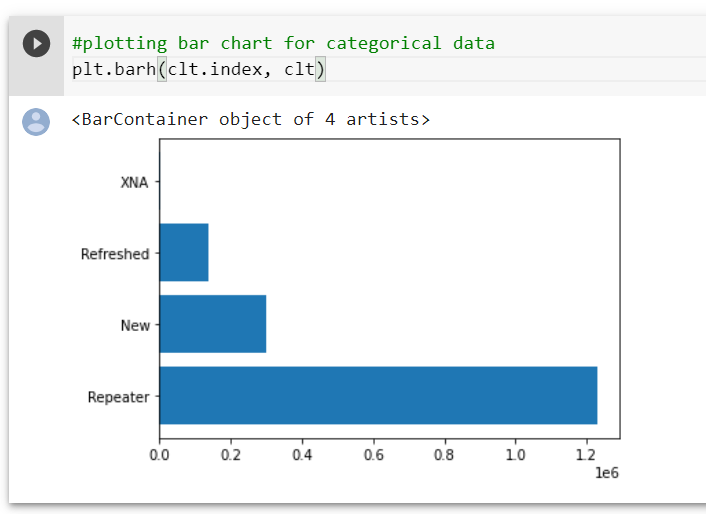
unique value CLIENT\_TYPE



count of unique value CLIENT\_TYPE

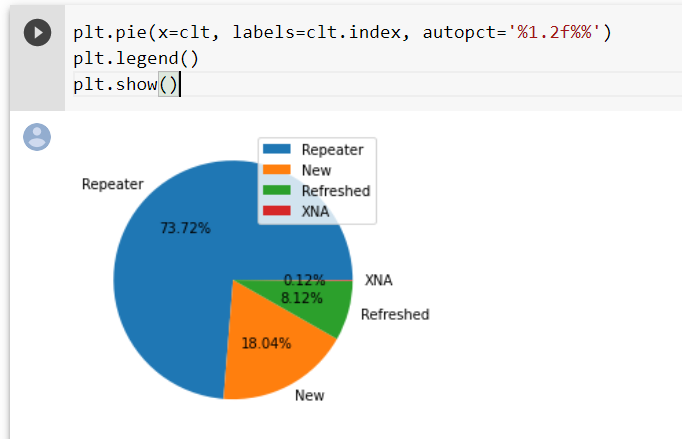


plotting bar chart for categorical data



Conclusion:

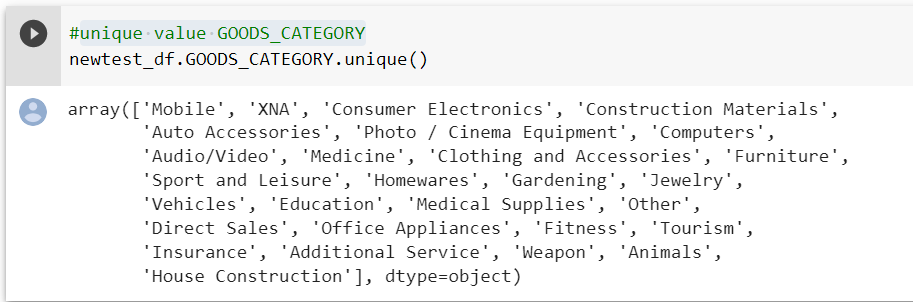
From bar graph is concluded that clients who are repeater are maximum

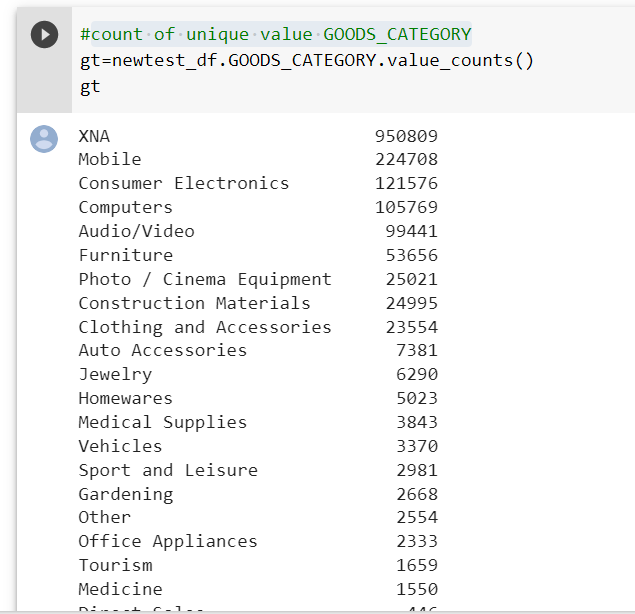


Conclusion:

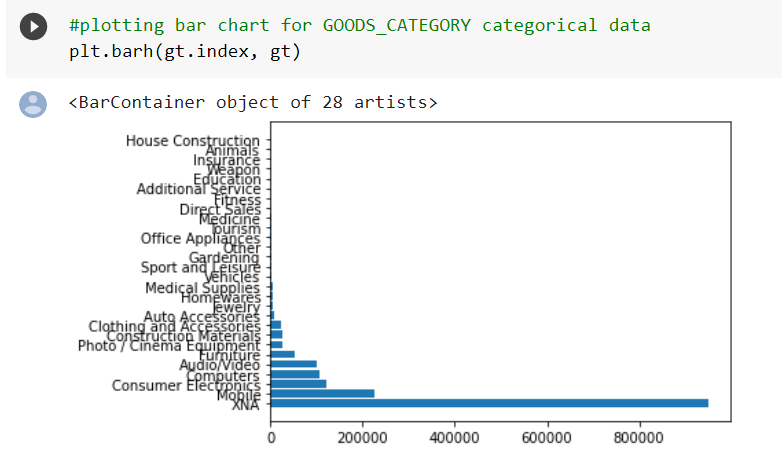
From pie graph is concluded that clients who are repeater are maximum

unique value GOODS\_CATEGORY

count of unique value GOODS\_CATEGORY



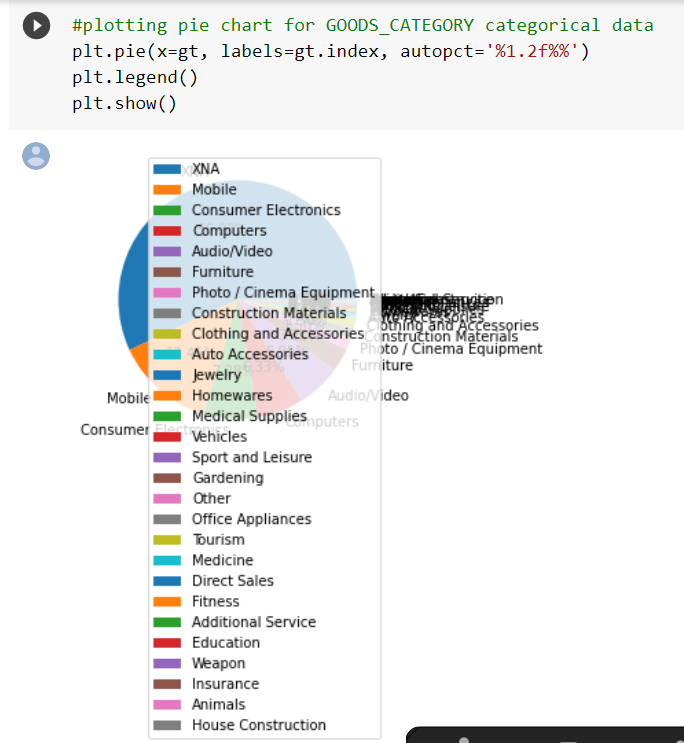
plotting bar chart for GOODS\_CATEGORY categorical data



Conclusion:

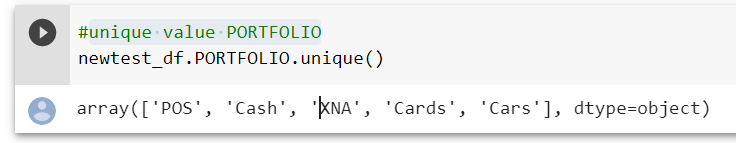
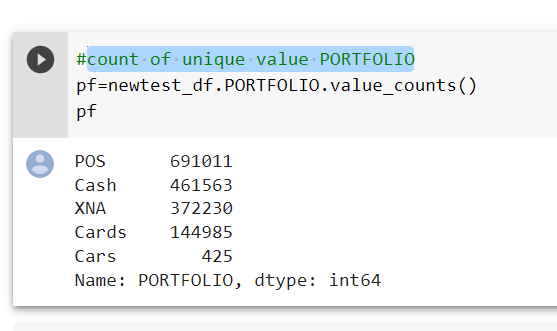
It can be concluded that clients who are taking loan for goods category of XNA are maximum

plotting pie chart for GOODS\_CATEGORY categorical data

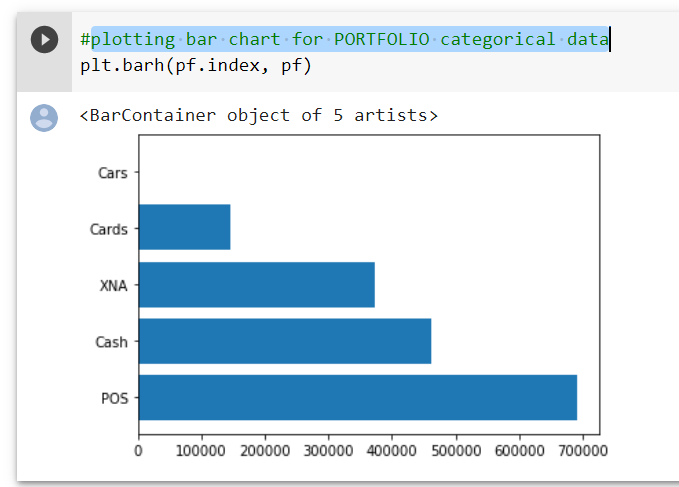


Conclusion:

People taking loans for electronics equipment are more as compared to people taking loans for house construction or insurance.

unique value PORTFOLIOcount of unique value PORTFOLIO

plotting bar chart for PORTFOLIO categorical data

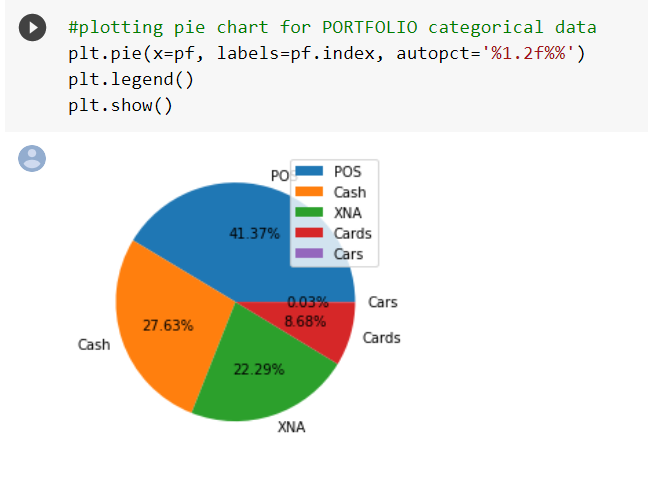


Conclusion:

It is concluded that POS (point-of-sale) type loan category are maximum

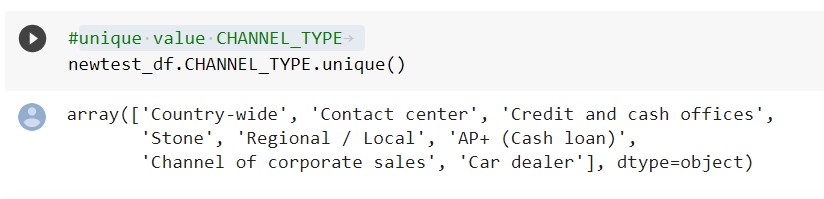
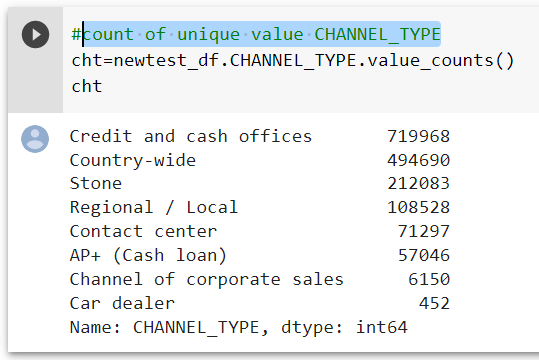
POS financing is **a broad term that describes methods for giving shoppers flexible, pay-over-time installment options**.

plotting pie chart for PORTFOLIO categorical data

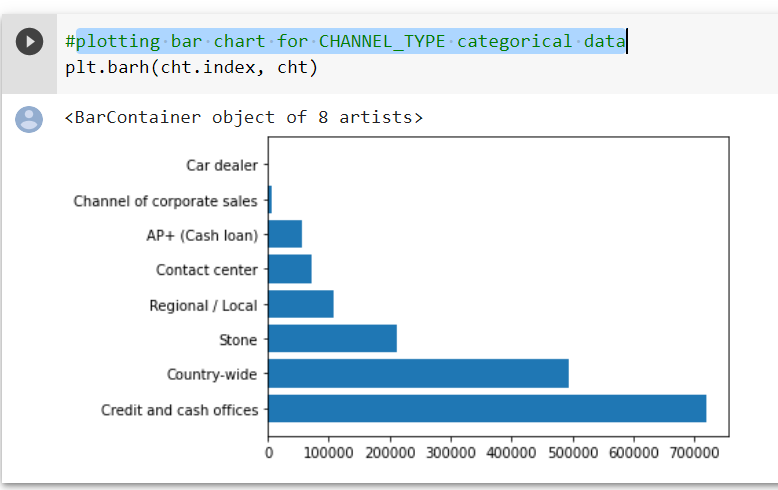


Conclusion:

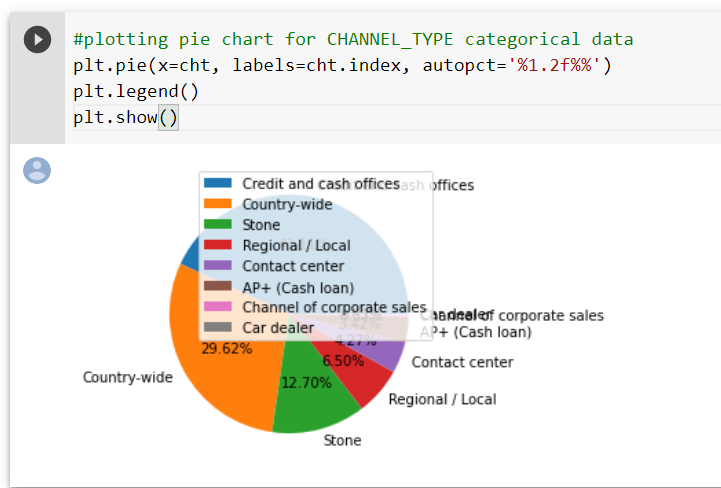
unique value CHANNEL\_TYPE

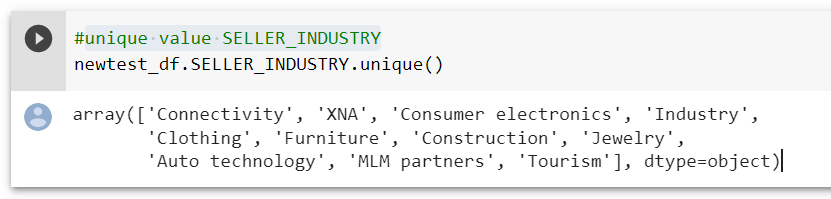
count of unique value CHANNEL\_TYPE 

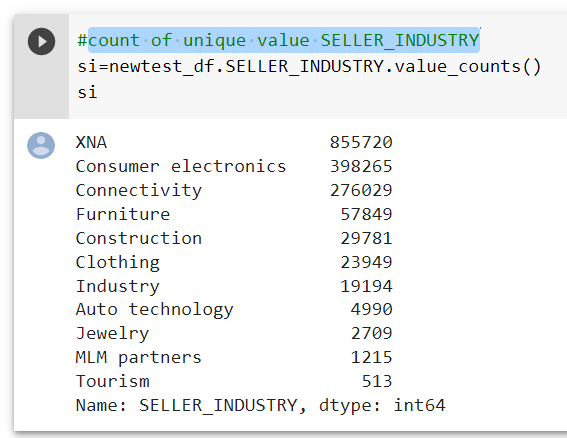
plotting bar chart for CHANNEL\_TYPE categorical data



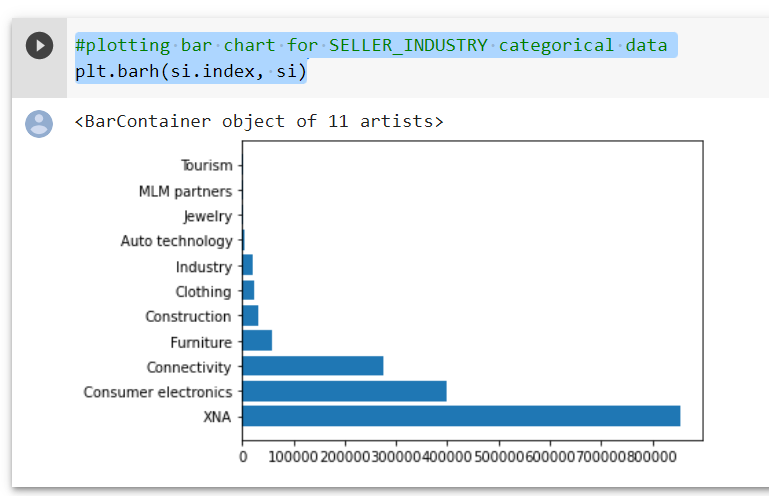
Conclusion:

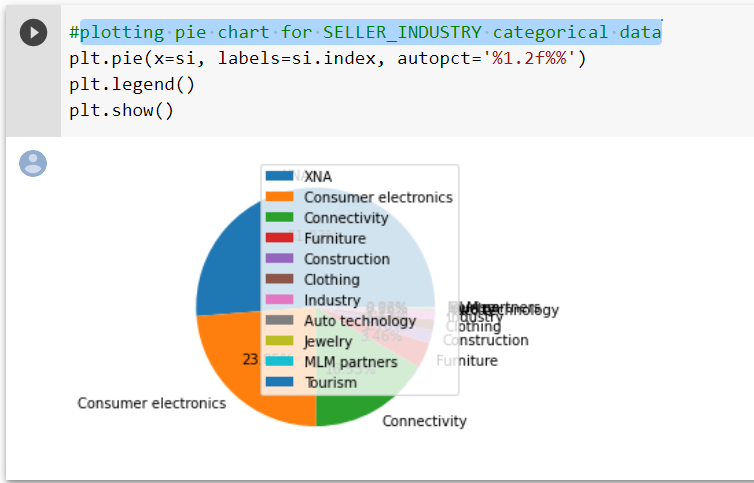
plotting pie chart for CHANNEL\_TYPE categorical data

unique value SELLER\_INDUSTRY

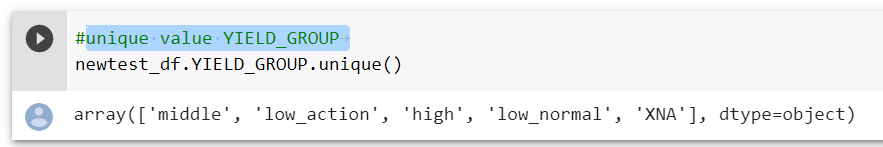
count of unique value SELLER\_INDUSTRY

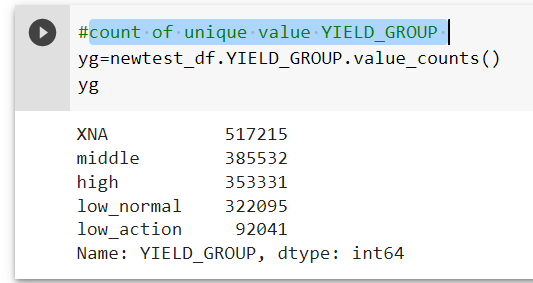
plotting bar chart for SELLER\_INDUSTRY categorical data



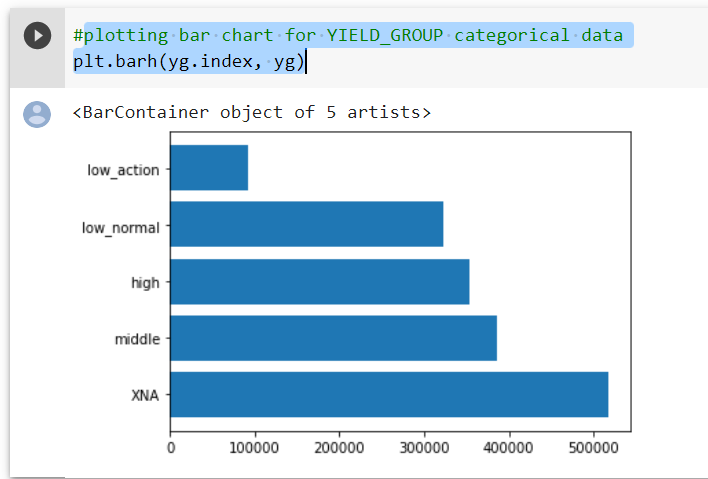
plotting pie chart for SELLER\_INDUSTRY categorical data

Conclusion:

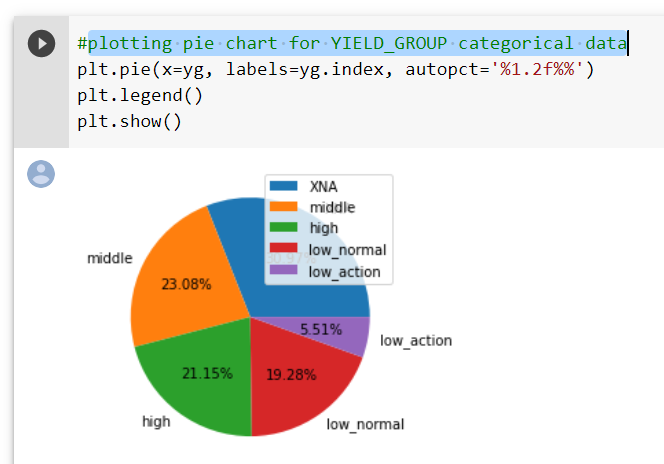
unique value YIELD\_GROUP

count of unique value YIELD\_GROUP 

plotting bar chart for YIELD\_GROUP categorical data

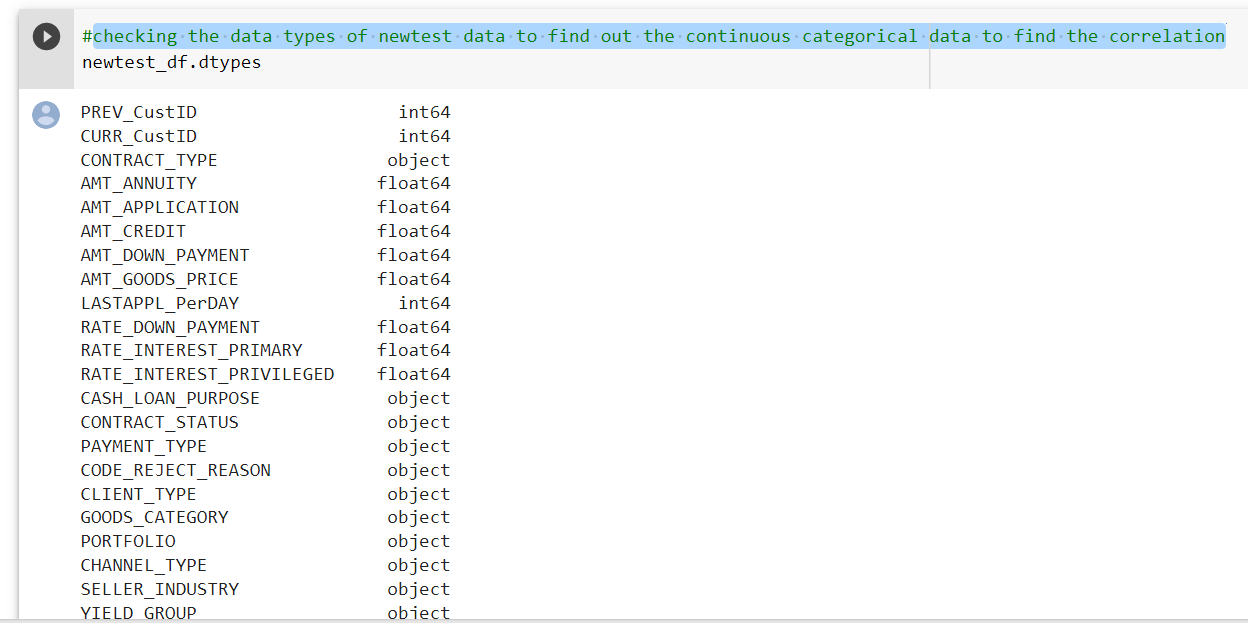


plotting pie chart for YIELD\_GROUP categorical data

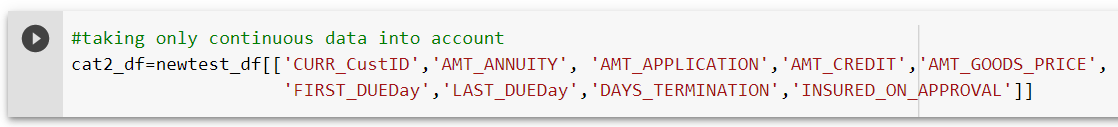


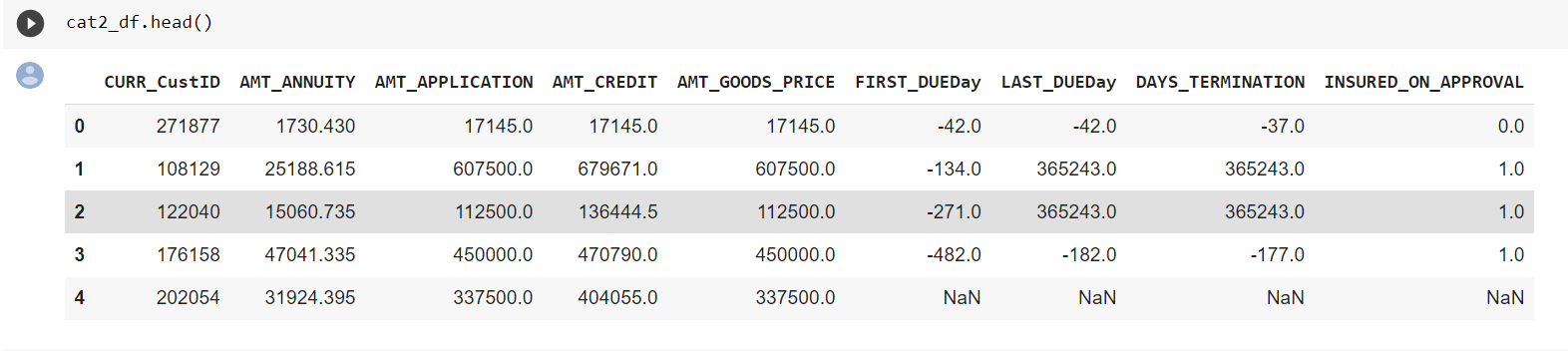
Conclusion:

--> for categorical data

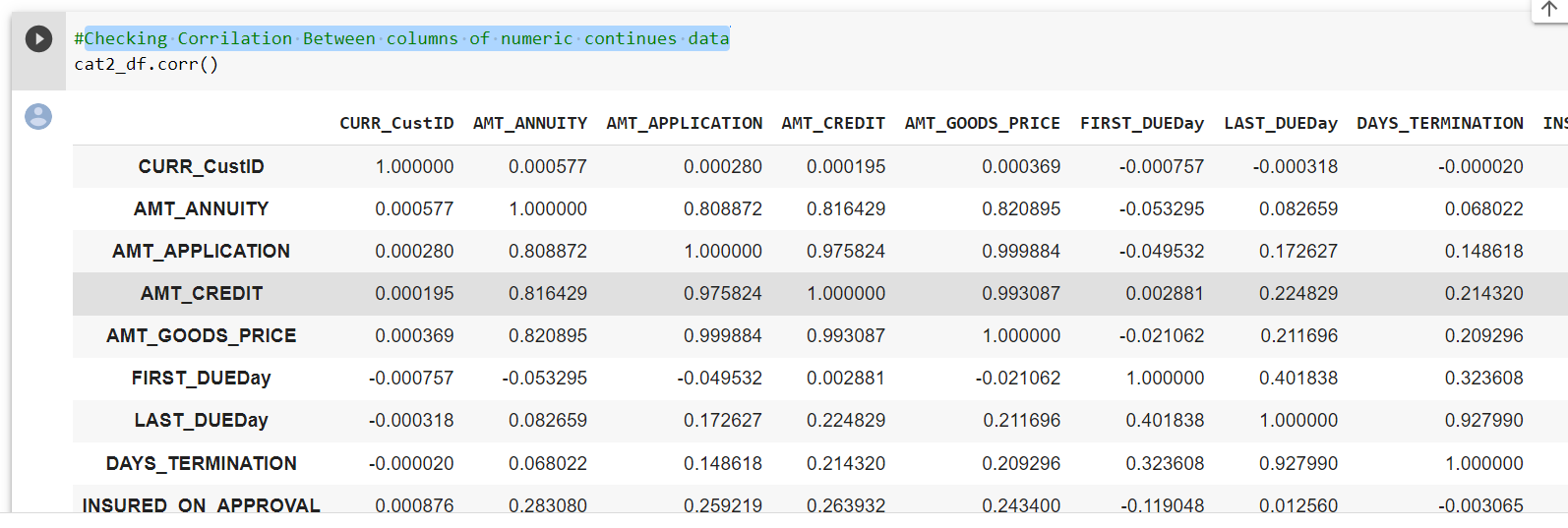
checking the data types of new test data to find out the continuous categorical data to find the correlation

taking only continuous data into account

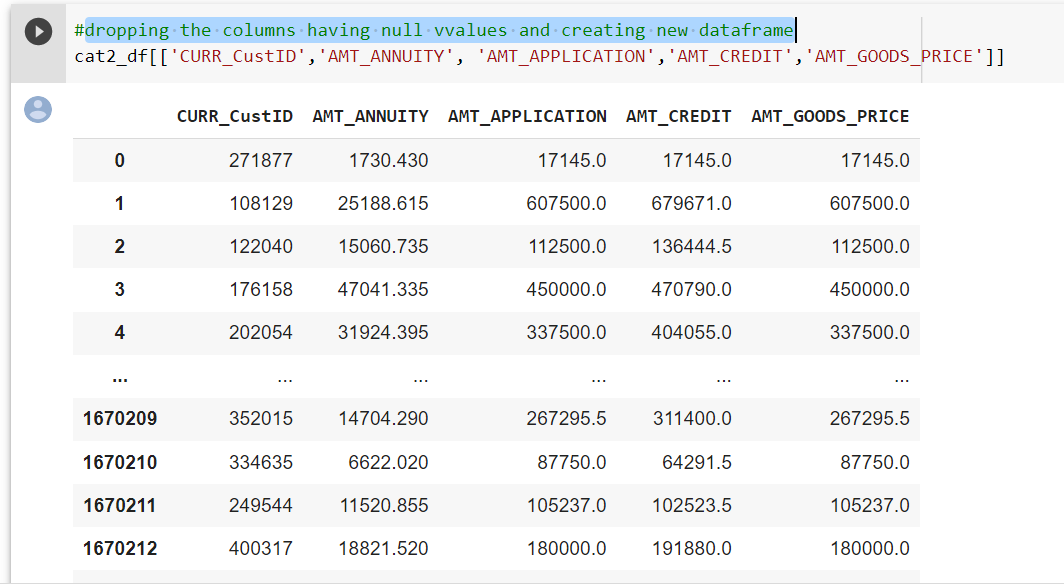


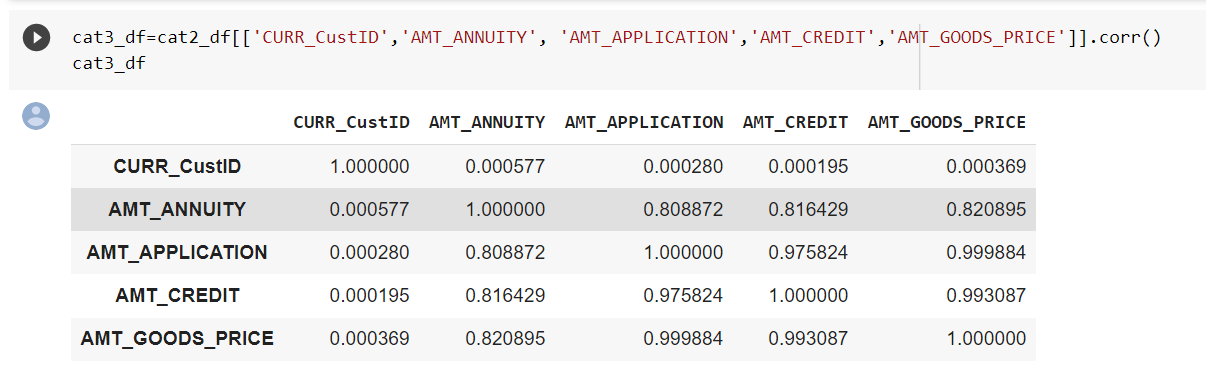


Checking Corrilation Between columns of numeric continues data



dropping the columns having null vvalues and creating new dataframe





Conclusion-->

1. AMT\_ANNUITY is strongly correlated with

* AMT\_APPLICATION by corelation coefficient 0.808872
* AMT\_CREDIT by corelation coefficient 0.816429
* AMT\_GOODS\_PRICE by corelation coefficient 0.820895

1. AMT\_APPLICATION is strongly correlated with

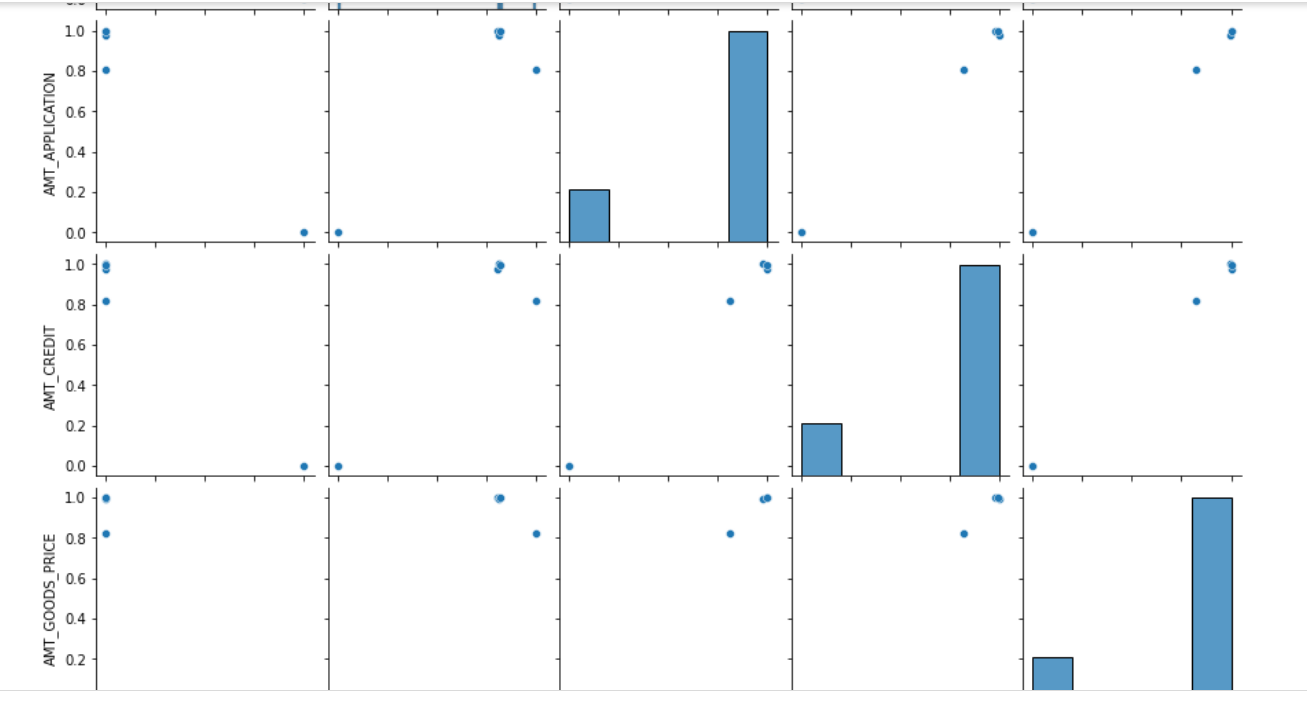
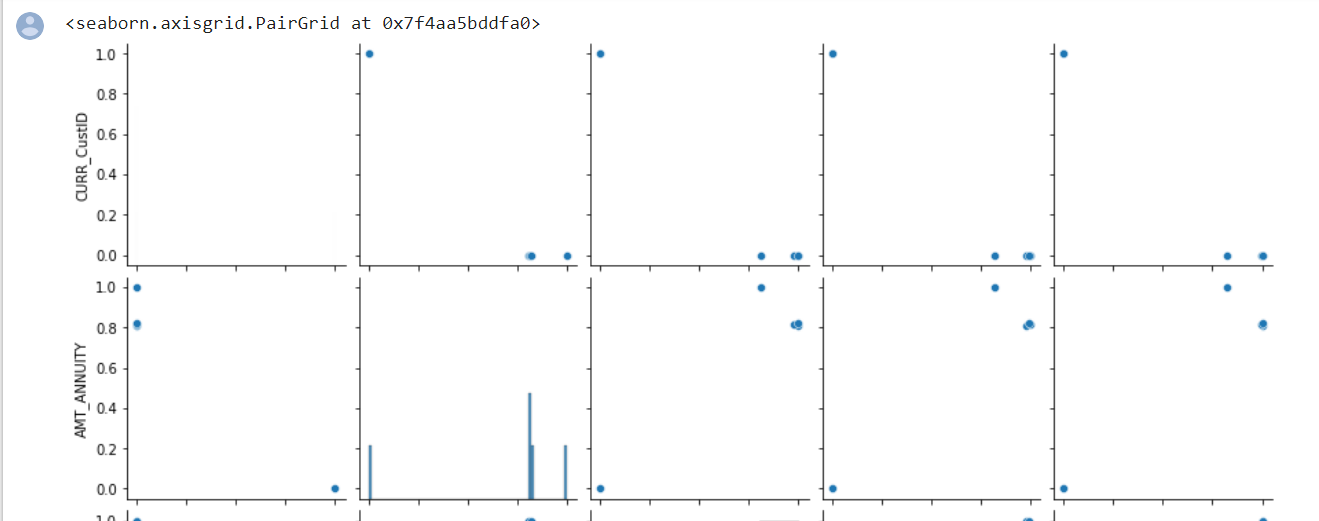
* AMT\_CREDIT by corelation coefficient 0.975824
* AMT\_GOODS\_PRICE by corelation coefficient 0.999884

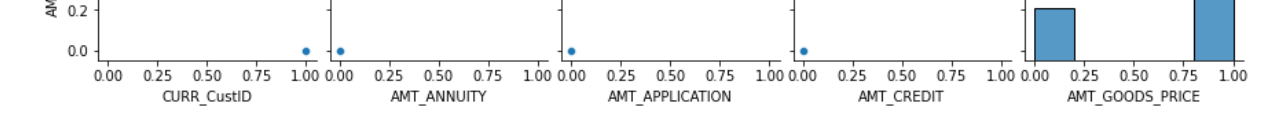
1. AMT\_CREDIT is strongly correlated with

* AMT\_GOODS\_PRICE by corelation coefficient 0.993087

# ploting pairplot for continues data to check the retaion between columns

creating pairplot graph

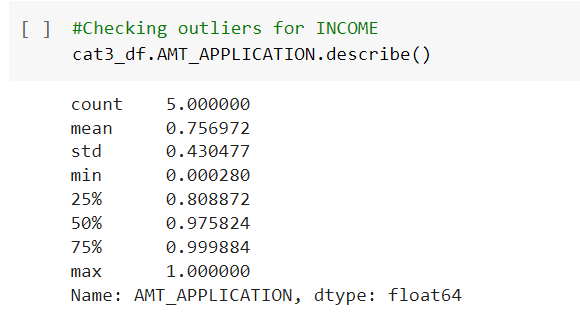


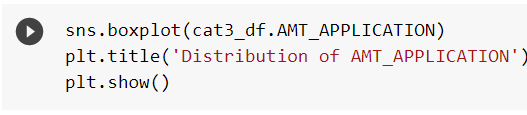


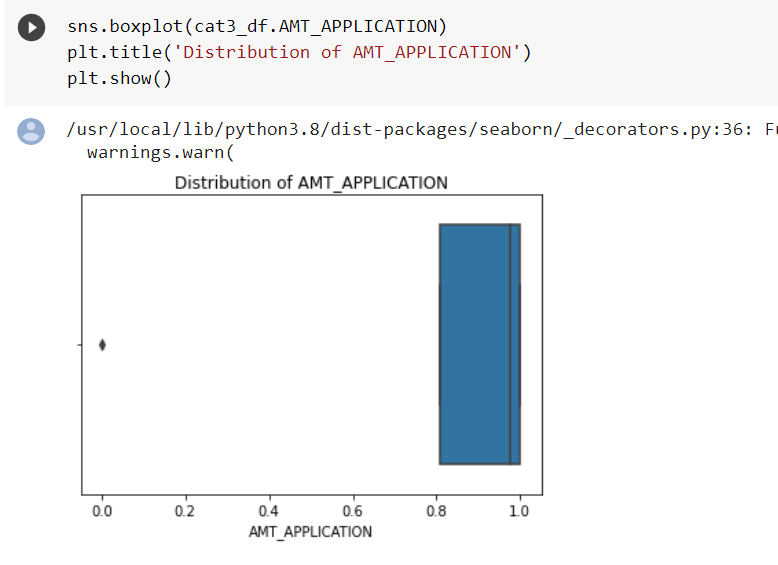
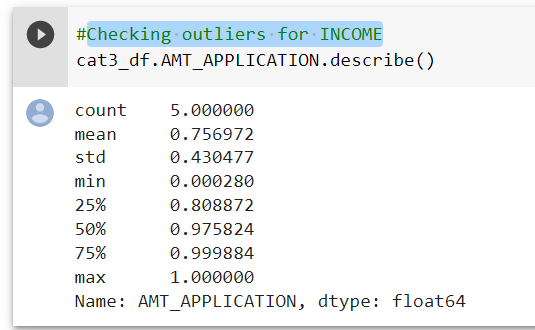
Conclusion:

# Handling Outliers

Checking outliers for INCOME



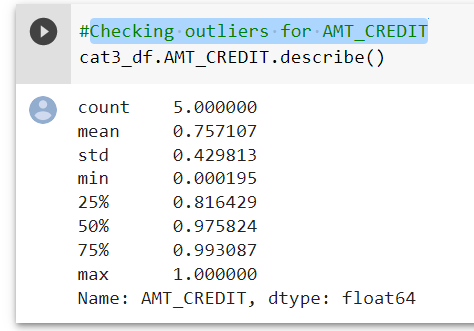


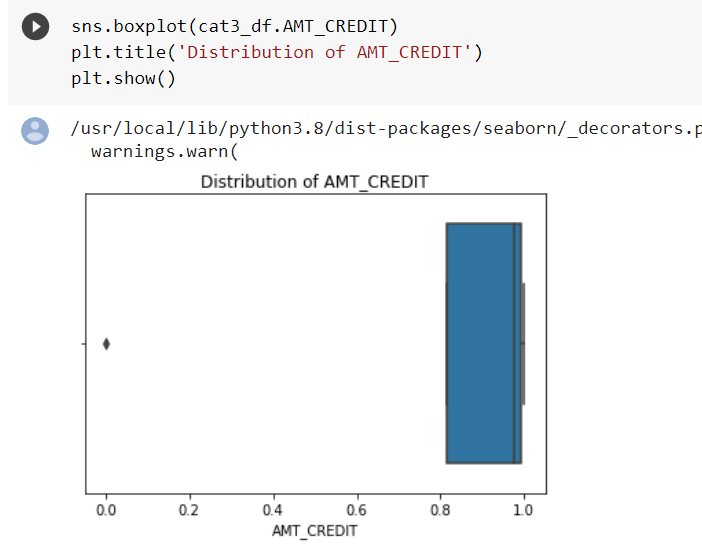


Conclusion:

AMT\_CREDIT

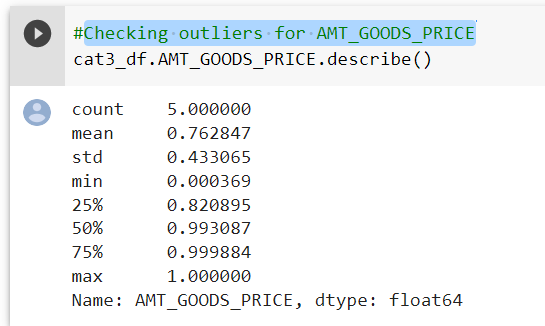
Checking outliers for AMT\_CREDIT

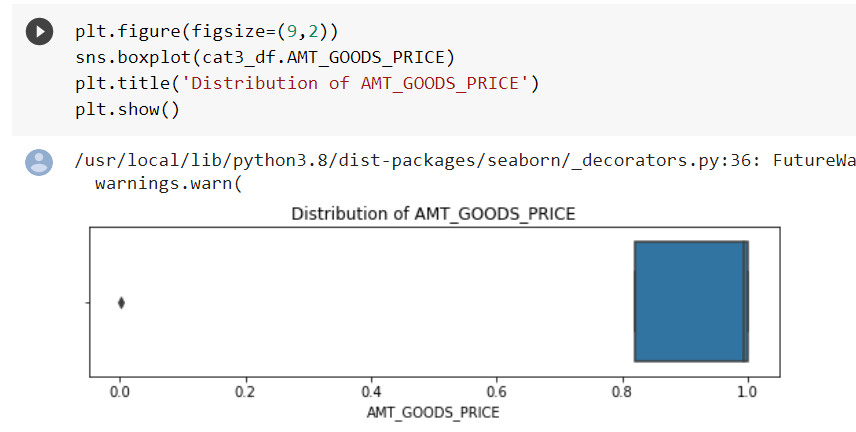




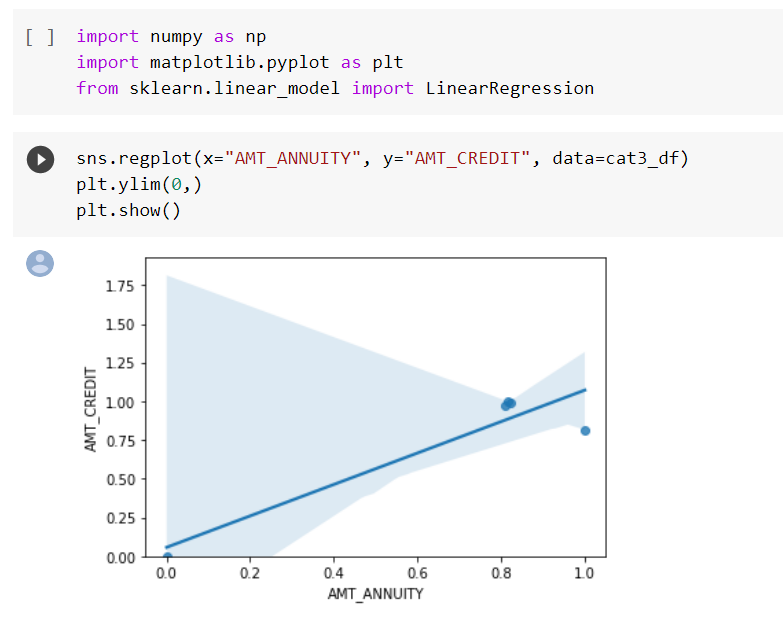
Conclusion:

AMT\_GOODS\_PRICE

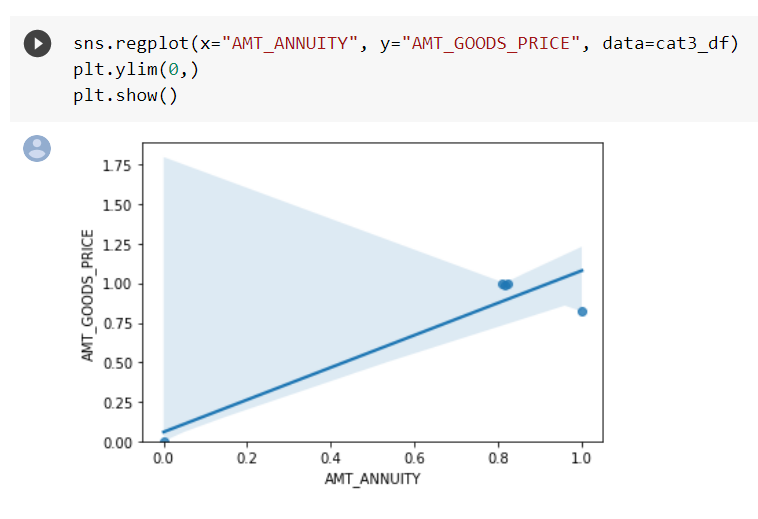
Checking outliers for AMT\_GOODS\_PRICE

****

**Conclusion:**

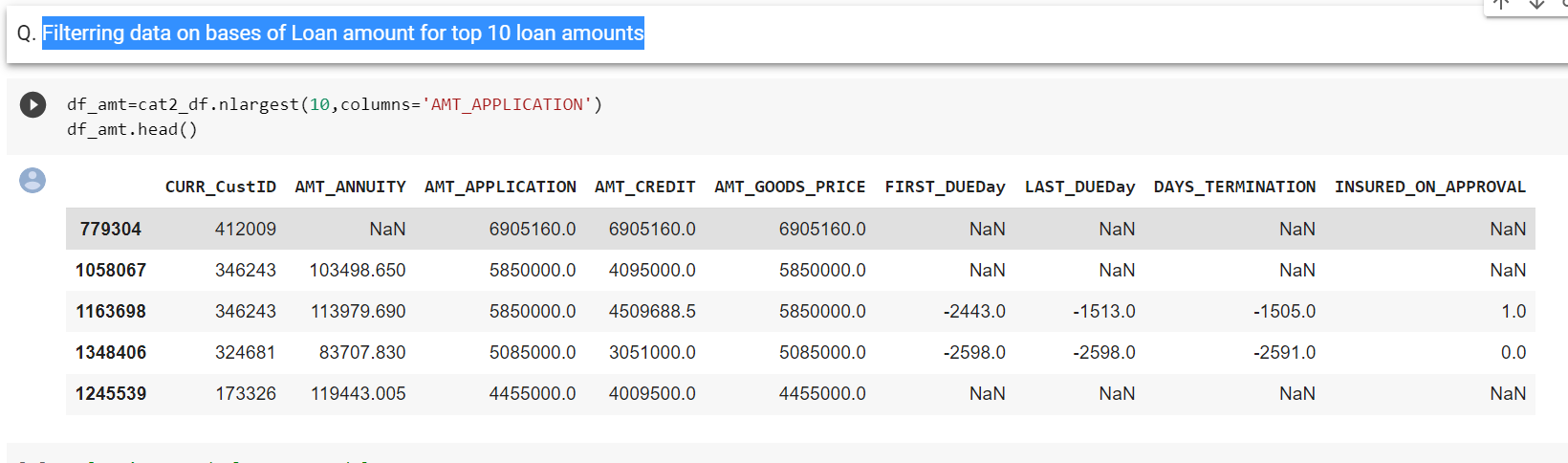
****

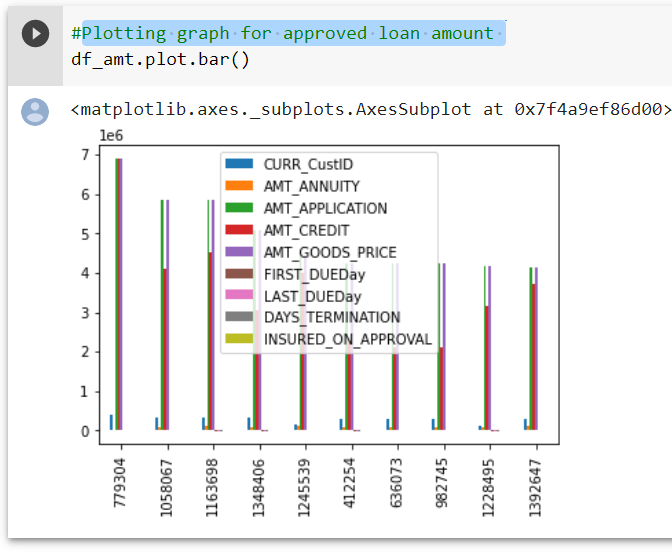
**Conclusion:**

****

**Conclusion:**

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



Plotting graph for approved loan amount 

conclusion: