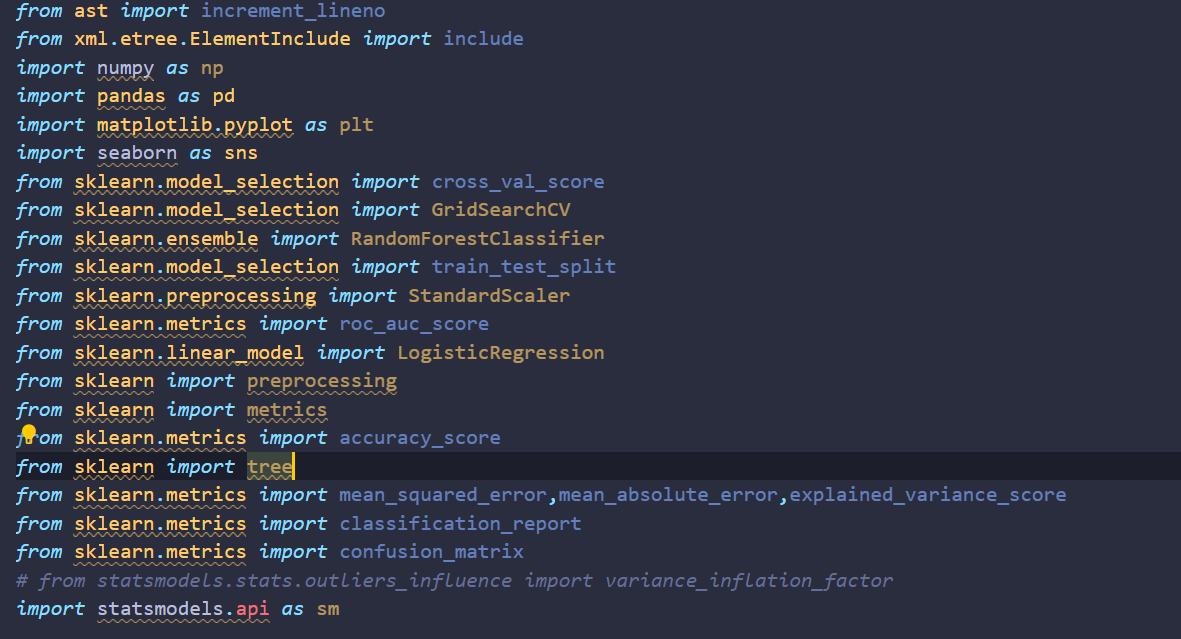
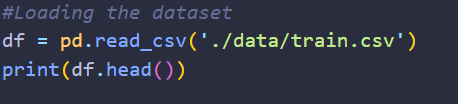
1. Install and import all the required libraries in Python



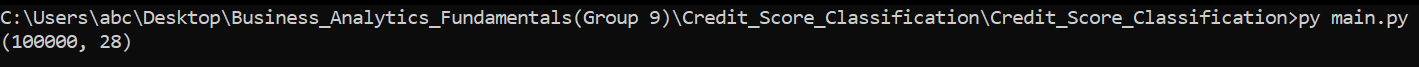
1. Loading the dataset



1. Check the data quality

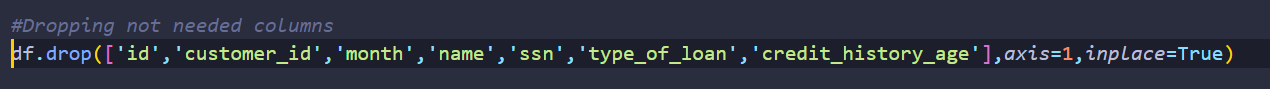


Output:



It says we have 100,000 rows and 28 columns in our dataset.

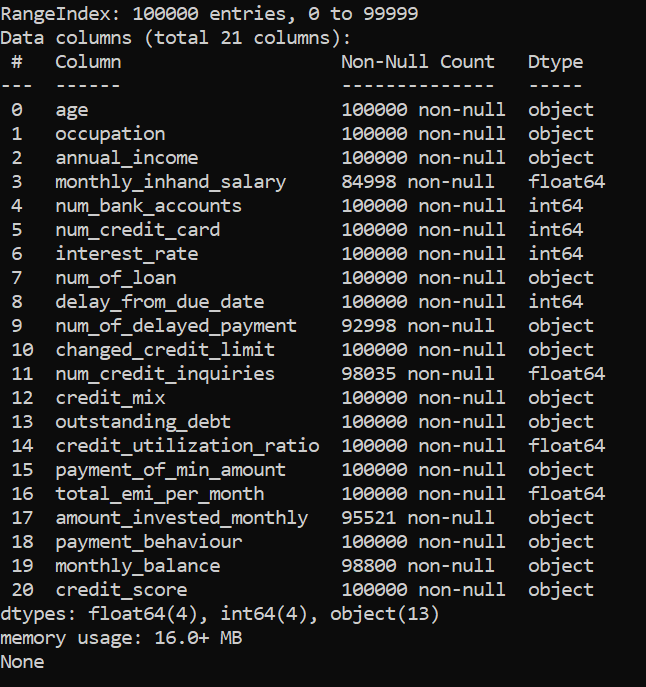
1. We would then drop all the columns that are not useful for analysis



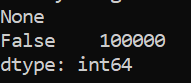
1. A useful command to get some details on the features



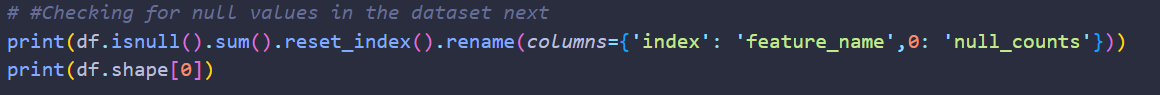
Output:



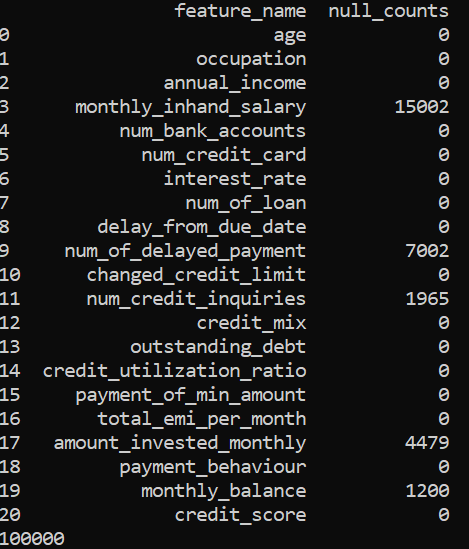
1. Check for duplicate Values



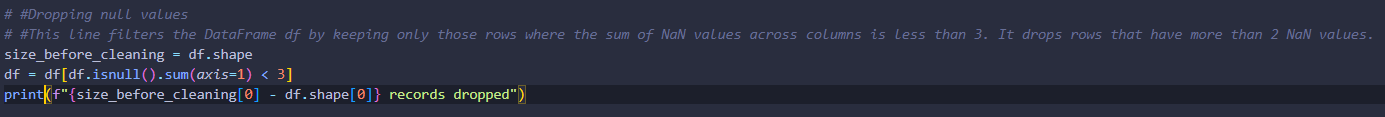
1. Check for null values in the df



Output:

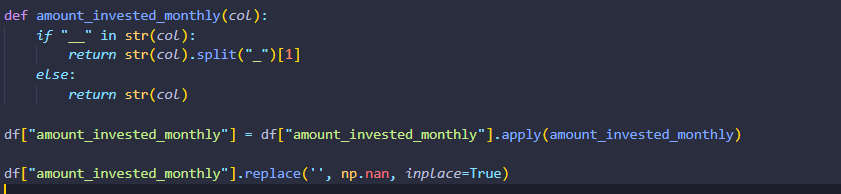


1. Check across columns in a particular row if there are NaN values and see if that count is 3 or more than 3.We will remove those rows from our dataset.

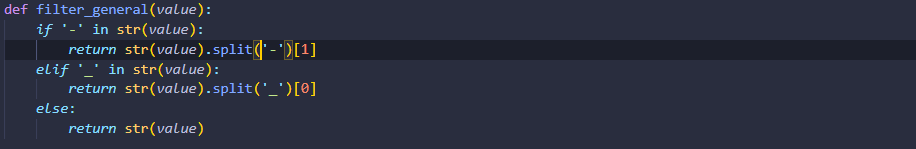


Data Cleaning

1. We will create new functions that will help us clean features, some of the functions would be feature specific like the one below:

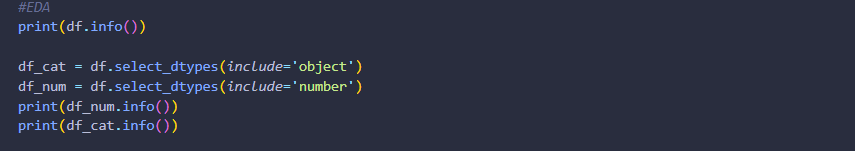


1. Others would be common for some of the features:



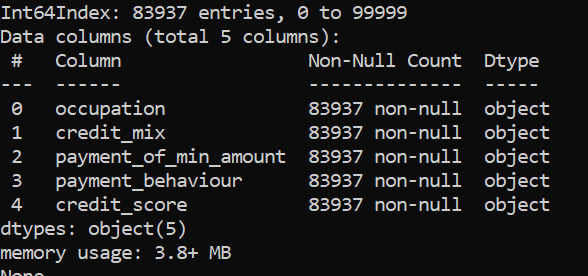
EDA

1. Its time for Exploratory Data Analysis now. We will first split the dataframe into 2 separate dataframes.One for numerical data and the other one for categorical data.

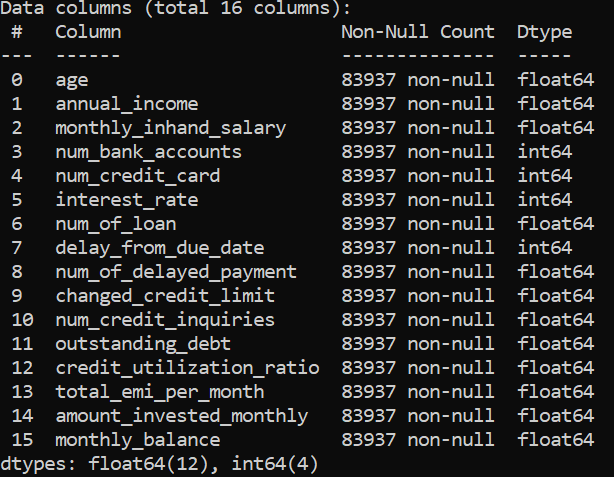


Output:

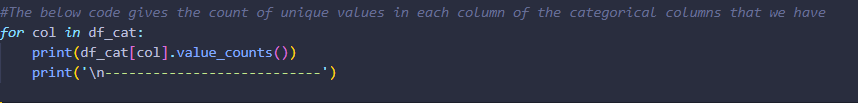
1. Categorical Data



1. Numerical Data

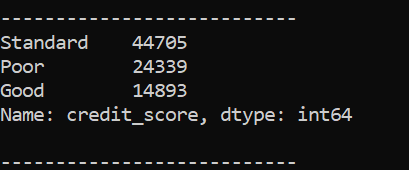


1. For categorical data, we will count all the unique values and plot it with matplotlib library.

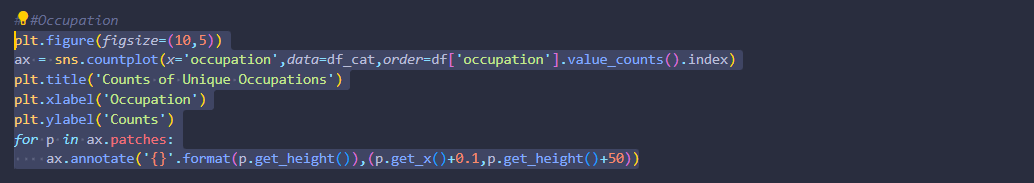


Output:

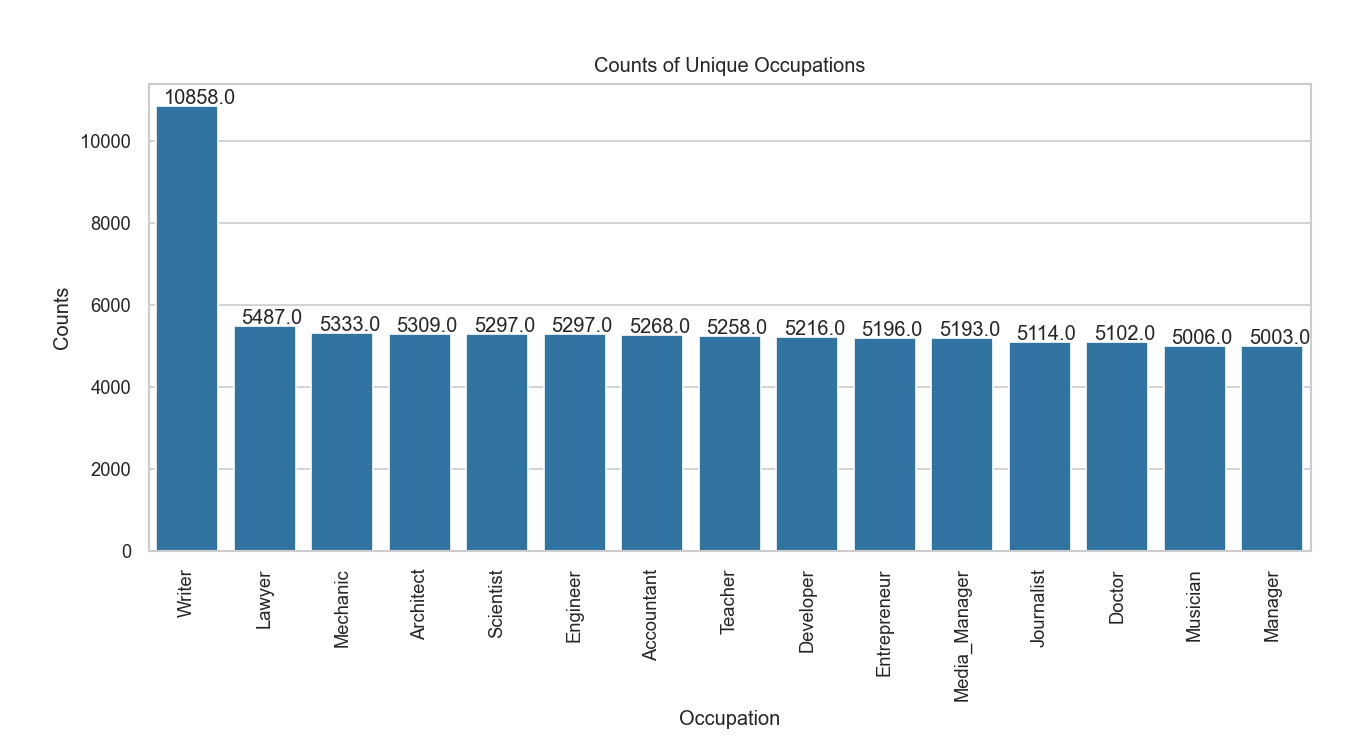
We get 5 more tables similar to the one below as the output. The graphs plotted with matplotlib do a better job of analysing the results.



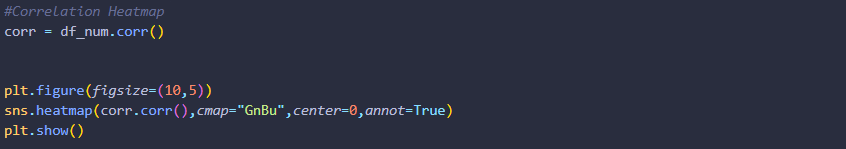
1. We will use the categorical data with the below code. We will do this for all the 5 features which are categorical to analyse the data.



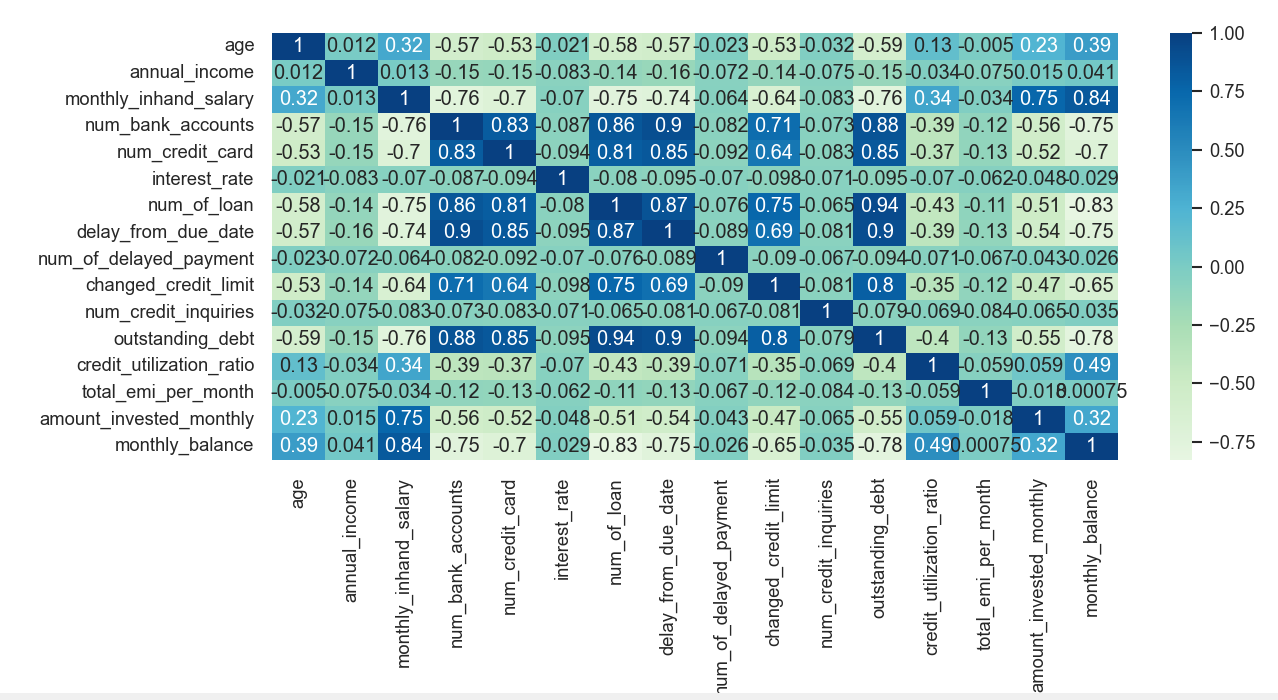
Output:



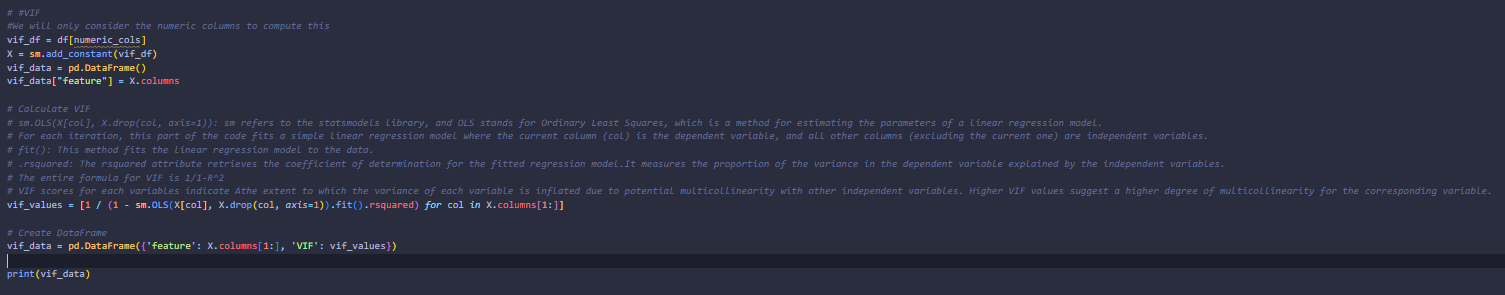
1. We will then plot the correlation for the numerical data with the seaborn library using a heatmap.



Output:

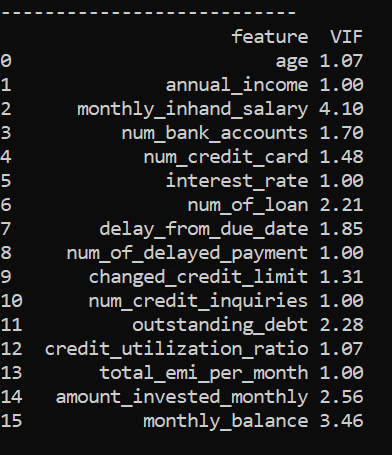


1. Calculating the VIF to negate multicollinearity issues. We will now remove the features with ViF greater than 2 as values above 2 are seen as a signal that there may be multicollinearity issues. The formula for VIF calculation is 1/1-R^2.We have kept 2 as the threshold for our data. Data analysts/scientists/researchers select this threshold based on the data available for analysis. We get rid of the features that have VIF > 2.





Output:



Data Preparation