**INTRODUCTION**

* Business Problem Framing

Build a model which can be used to predict in terms of a probability for each loan transaction, whether the customer will be paying back the loaned amount within 5 days of insurance of loan. In this case, Label ‘1’ indicates that the loan has been paid and Label ‘0’ means he is a defaulter.

* Conceptual Background of the Domain Problem

The main problem is found variables that impact most on the label, also the features which are predictable whether the loan is given to the person or not on basis of previous data we have.

* Review of Literature

Data exploration is the first step in data analysis and typically involves summarizing the main characteristics of a data set, including its size, accuracy, initial patterns in the data, and other attributes. It is commonly conducted by data analysts using visual analytics tools, but it can also be done in more advanced statistical software, Python. Before it can analyze data collected by multiple data sources and stored in data warehouses, an organization must know how many cases are in a data set, what variables are included, how many missing values there are, and what general hypotheses the data is likely to support. An initial exploration of the data set can help answer these questions by familiarizing analysts with the data with which they are working. We divided the data 8:2 for Training and Testing purposes respectively.

* Motivation for the Problem Undertaken

Every problem of Machine learning gives us chance to enhance and develop problem-solving skills. These Problems do’s the same.

When this real-life problem of predicting whether the loan is provided or not to the customer, whether to enter the market or not and with help of A. I technology affordable houses for the future generations is in under development.

As Data scientists it is our role to help companies to understand the market better with older data we have for constructing the houses according to that only and make profitable models.

**Analytical Problem Framing**

* Mathematical/ Analytical Modelling of the Problem

As for any basic model building, we have to understand the type of target variable, the data of the target variable is continued or classified.

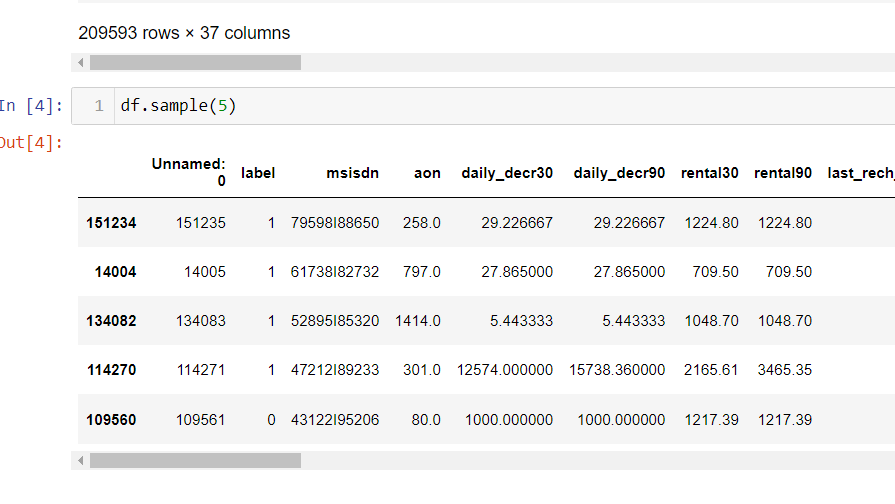
Data Analysis is always the difficult part, for better understanding different kinds of bar plots, distribution plots are created with the target Column for finding the insights of the dataset we have.

Analytical Modelling always starts with the target variable we have, and in that case, our target variable is Label, for that, we create some distribution plots with the target variable to understand which feature columns help to learn the model best and which feature columns reduce the accuracy of the model.

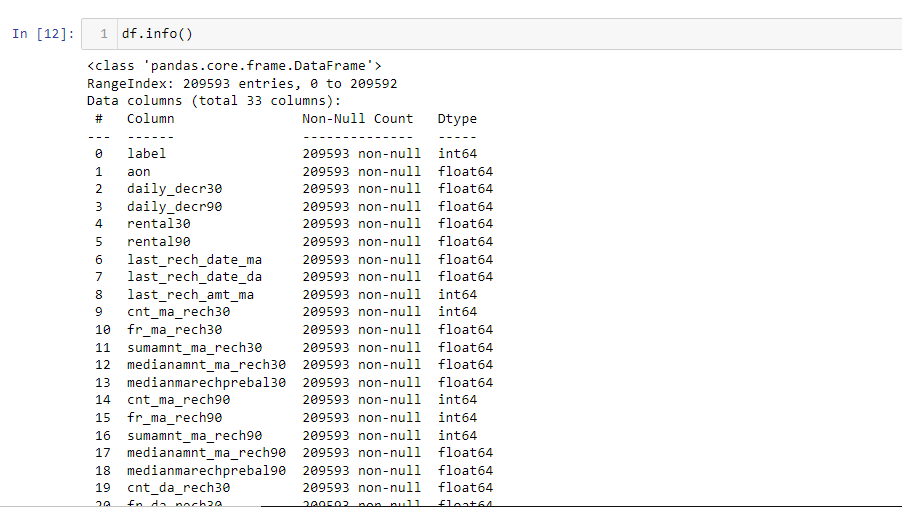
And after finding the relation and correlation with the target variable we choose either Regression Model or Classification Model. Here in this problem, our target feature column is categorical so we build our Machine Learning model on Classification.

* Data Sources and their formats
* We are working with one such client that is in Telecom Industry. They are a fixed wireless telecommunications network provider. They have launched various products and have developed their business and organization based on the budget operator model, offering better products at Lower Prices to all value-conscious customers through a strategy of disruptive innovation that focuses on the subscriber.
* They understand the importance of communication and how it affects a person’s life, thus, focusing on providing their services and products to low-income families and poor customers that can help them in the need of the hour.
* They are collaborating with an MFI to provide micro-credit on mobile balances to be paid back in 5 days. The Consumer is believed to be a defaulter if he deviates from the path of paying back the loaned amount within the time duration of 5 days. For the loan amount of 5 (in Indonesian Rupiah), the payback amount should be6(in Indonesian Rupiah), while, for the loan amount of 10(in Indonesian Rupiah), the payback amount should be 12(in Indonesian Rupiah).
* The sample data is provided to us from our client database. It is hereby given to you for this exercise. To improve the selection of customers for the credit, the client wants some predictions that could help them in further investment and improvement in the selection of customers.

**Dataset looks as follows-**

****

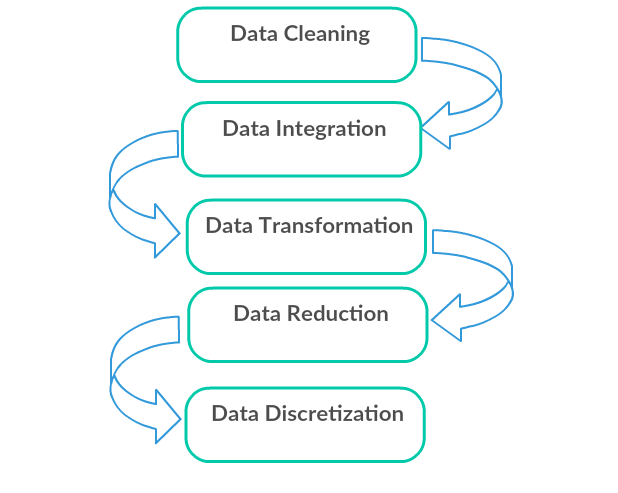
**Dataset Information looks as follows-**

****

Major data is float data type no null values in the data set.

* Data Pre-processing Done

Data pre-processing can refer to the manipulation or dropping of data before it is used to ensure or enhance performance, and is an important step in the data mining process.



1. Data Cleaning: First we clean the data which have no use in prediction like the ID column, then we drop the data which has a high no of missing percentages.

Like we drop unnamed :0, pdate,msidn,pcircle feature columns.

(df = df.drop(['Unnamed: 0','pdate','msisdn','pcircle'], axis=1)

#because there is no unique ness in feature column)

1. Data Integration: then we do some EDA process for finding out the meaning full insights of the data.
2. Data transformation is the process of changing the format, structure, or values of data; we use a labeled encoder for coding the object data into integer data.
3. Data Reduction: it is the process of finding the most correlated columns, and combining them because the machine does not understand which feature columns impact the most on accuracy.
4. Data discretization converts a large number of data values into smaller once, so that data evaluation and data management becomes very easy, using box plots is makes a clear understanding of the data.

* State the set of assumptions (if any) related to the problem under consideration

During Data cleaning we assume that the columns outliers are more than 20% at 3 standard deviations so we remove outliers only from 4.5 standard deviations. Which are 10%.

* Hardware and Software Requirements and Tools Used

**Python**

Python is widely used in scientific and numeric computing:

SciPy is a collection of packages for mathematics, science, and engineering.

Pandas are data analysis and modeling libraries.

Libraries Used for this Project include –

1. Pandas

2. NumPy

3. Matplotlib

4. Seaborn

5. Scikit Learn

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

After analyzing the dataset, I observe that many of the feature columns are object types so first, we have to convert them into integer or float types so that the machine interprets the data and for that we do label encode all the features column.

After label encoding, we find that many feature columns have Nan values so we use mean and median for filling that missing data,

Then find the correlation between the columns with target columns and delete the non-related feature columns.

We observe that the target column is skewed so we remove the skewness of the target column because normal data gives better results when we make the M.L model.

The target column is continuous type so we start work on Regression models building.

* Testing of Identified Approaches (Algorithms)

1. Logistic Regression
2. Regurgitation:

Ridge Classifier

1. Ensemble techniques

Decision Tree Classifier

* Run and Evaluate selected model

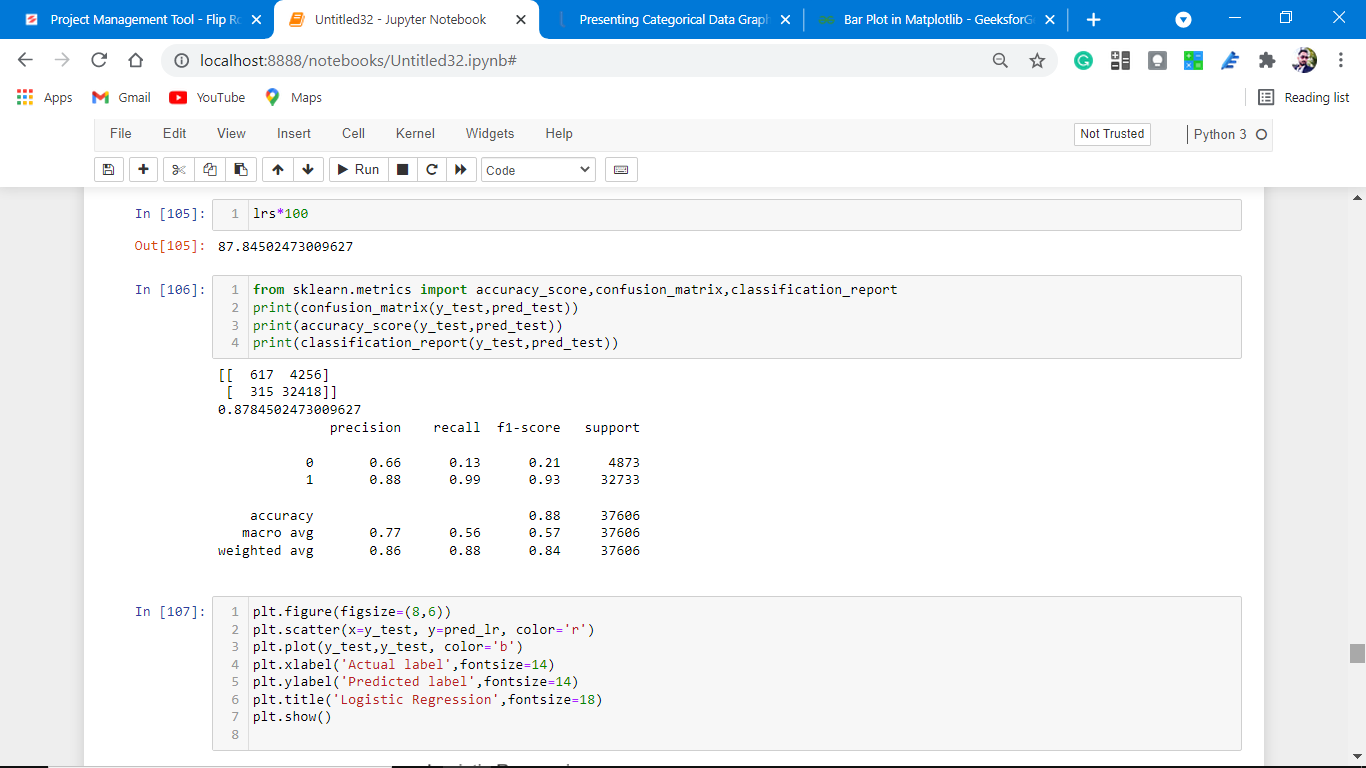
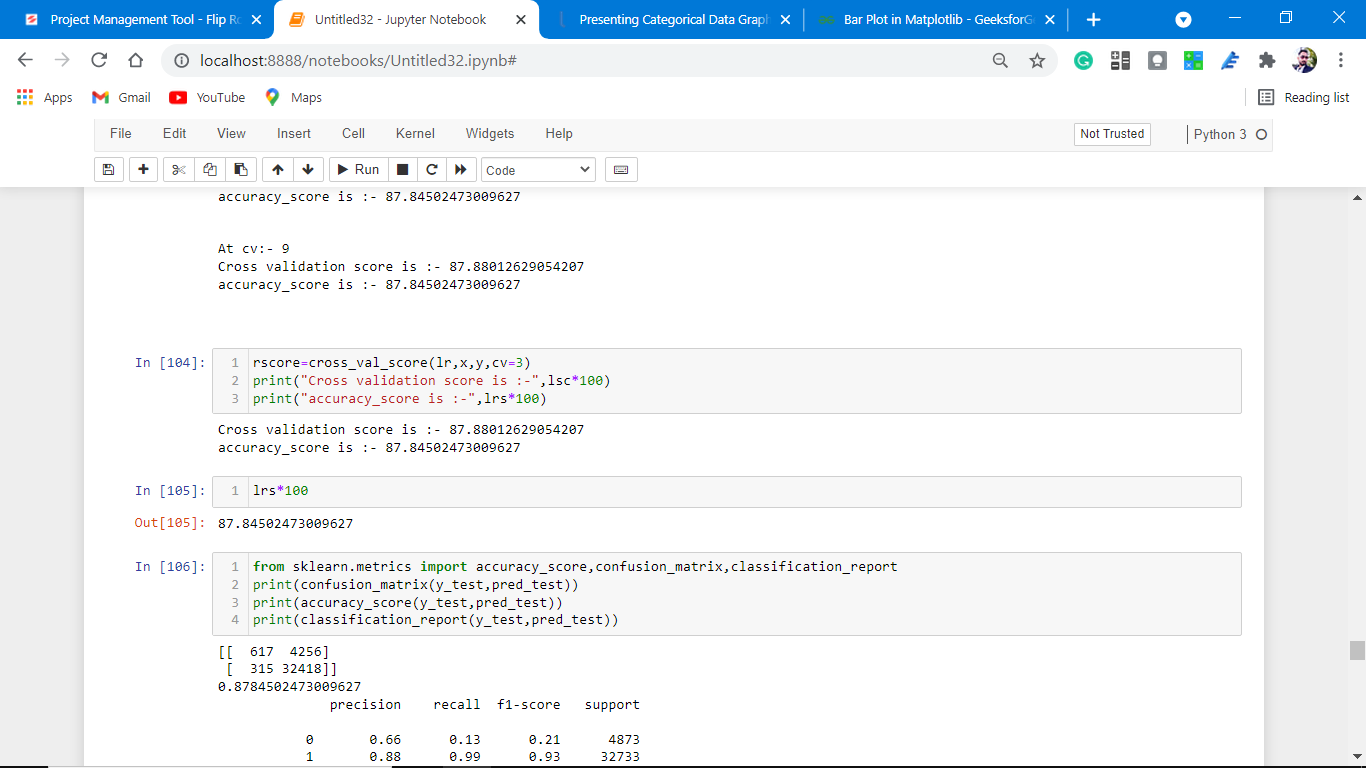
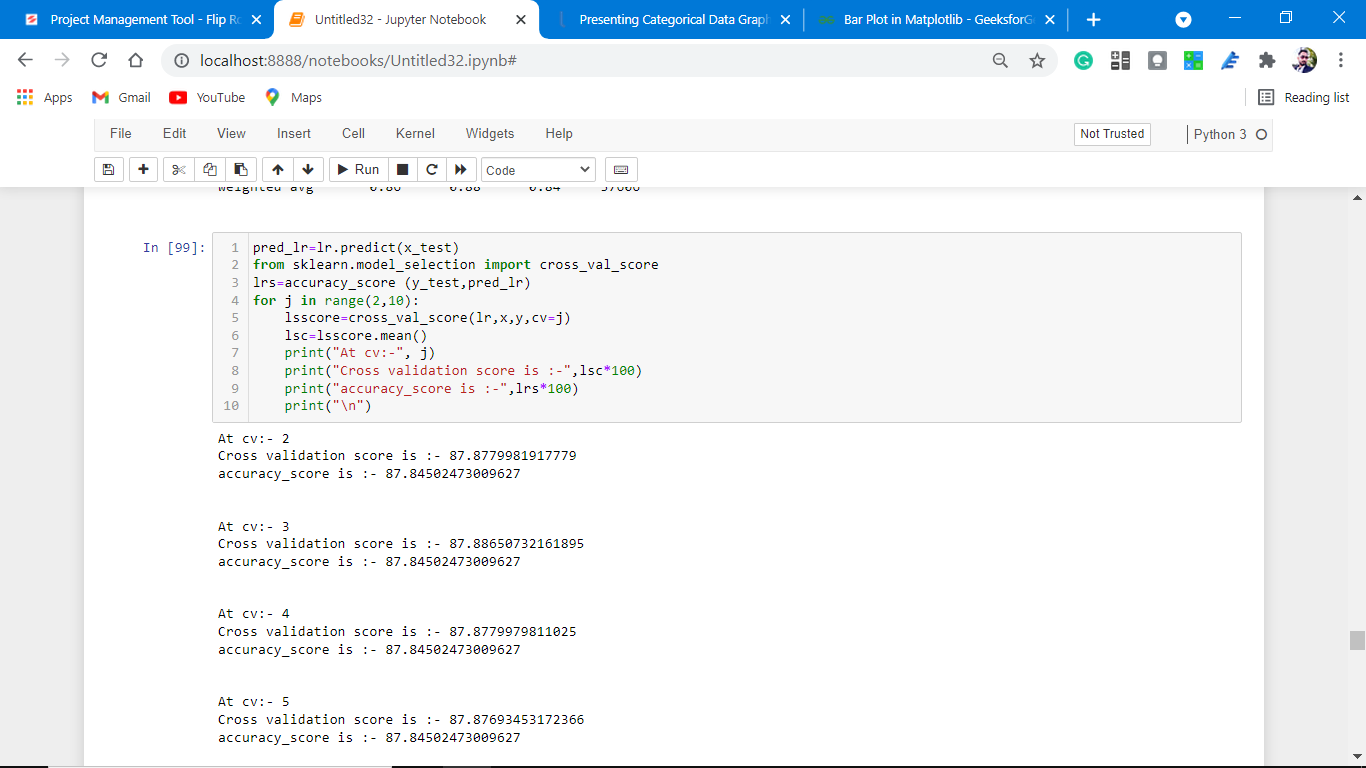
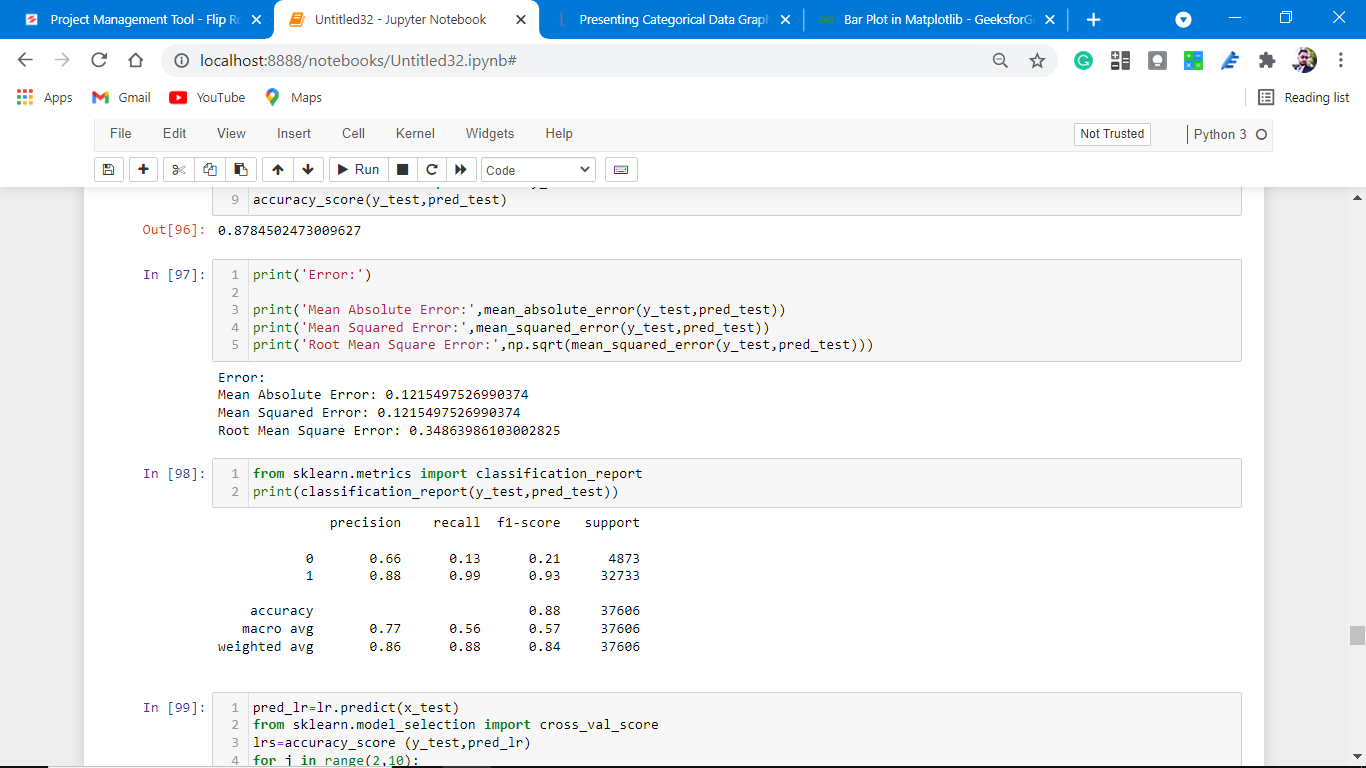
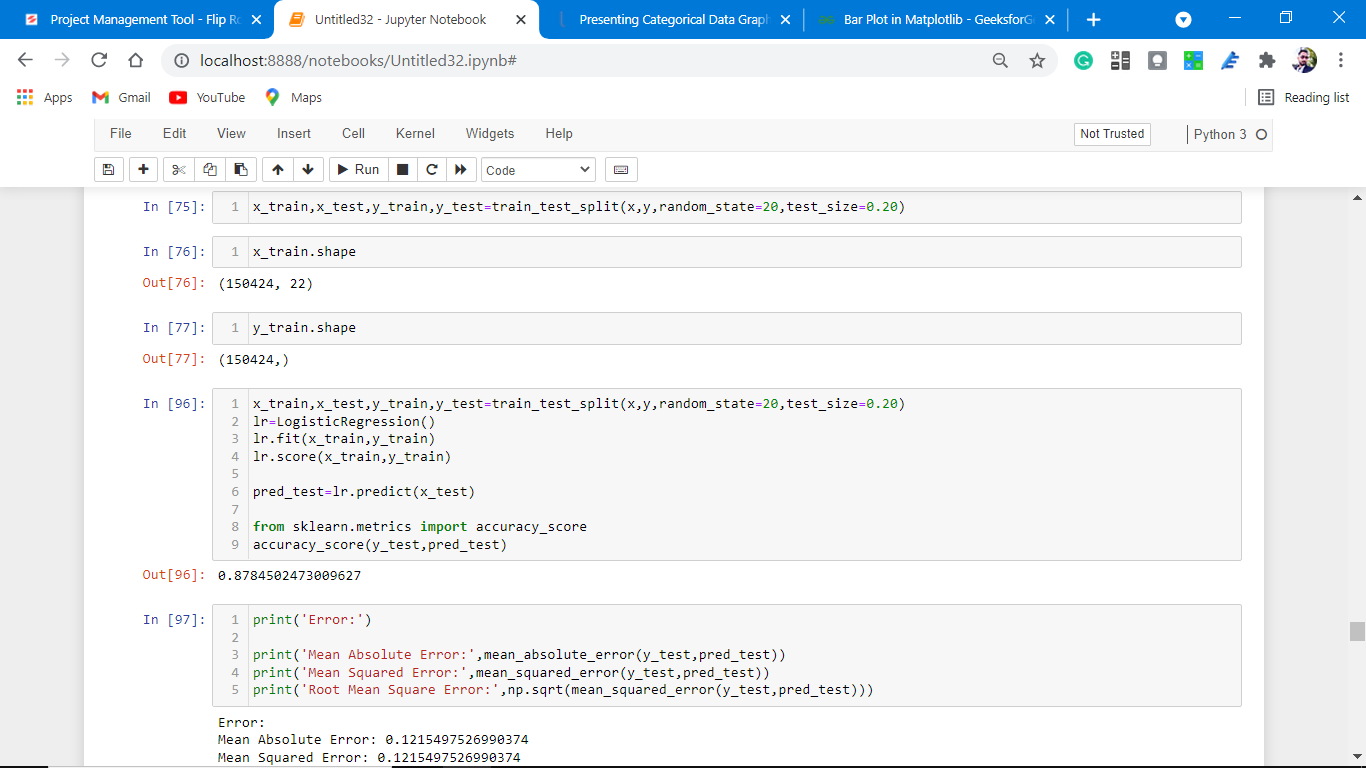
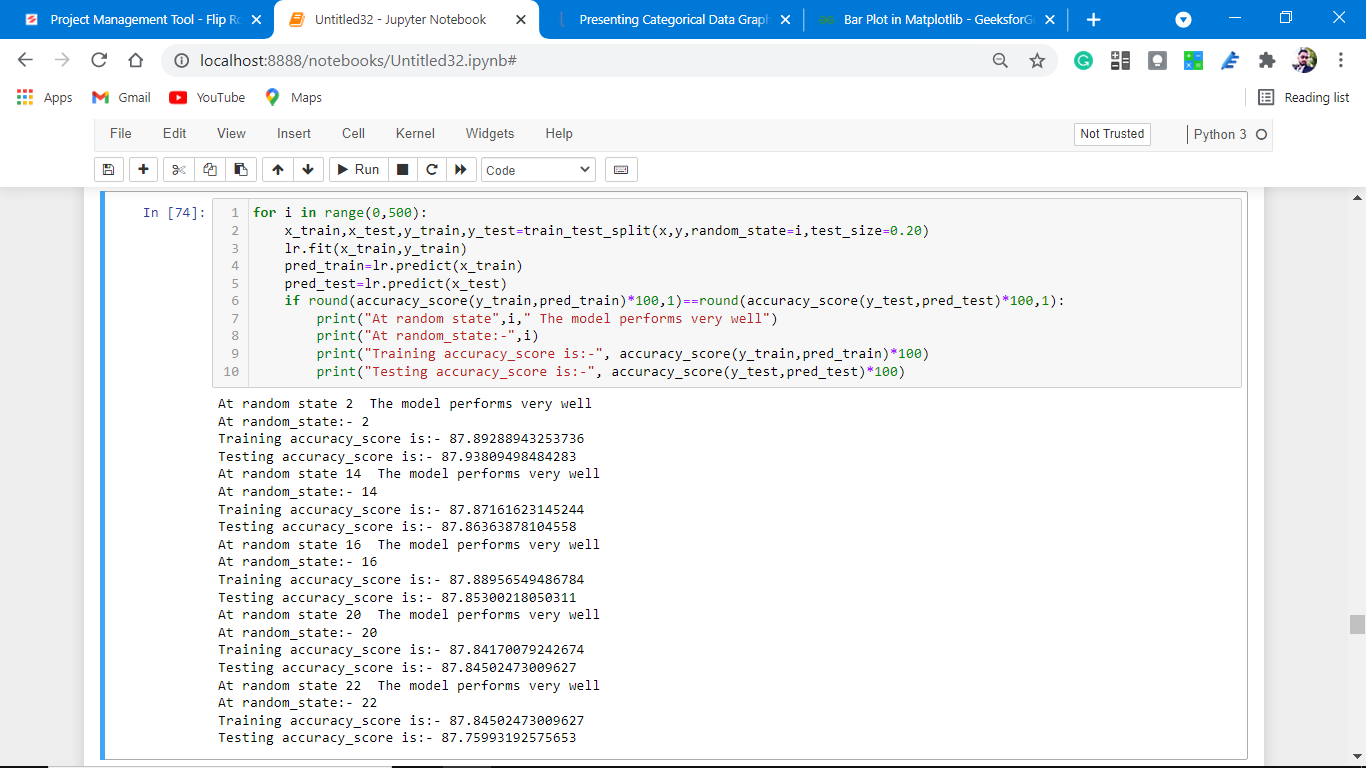
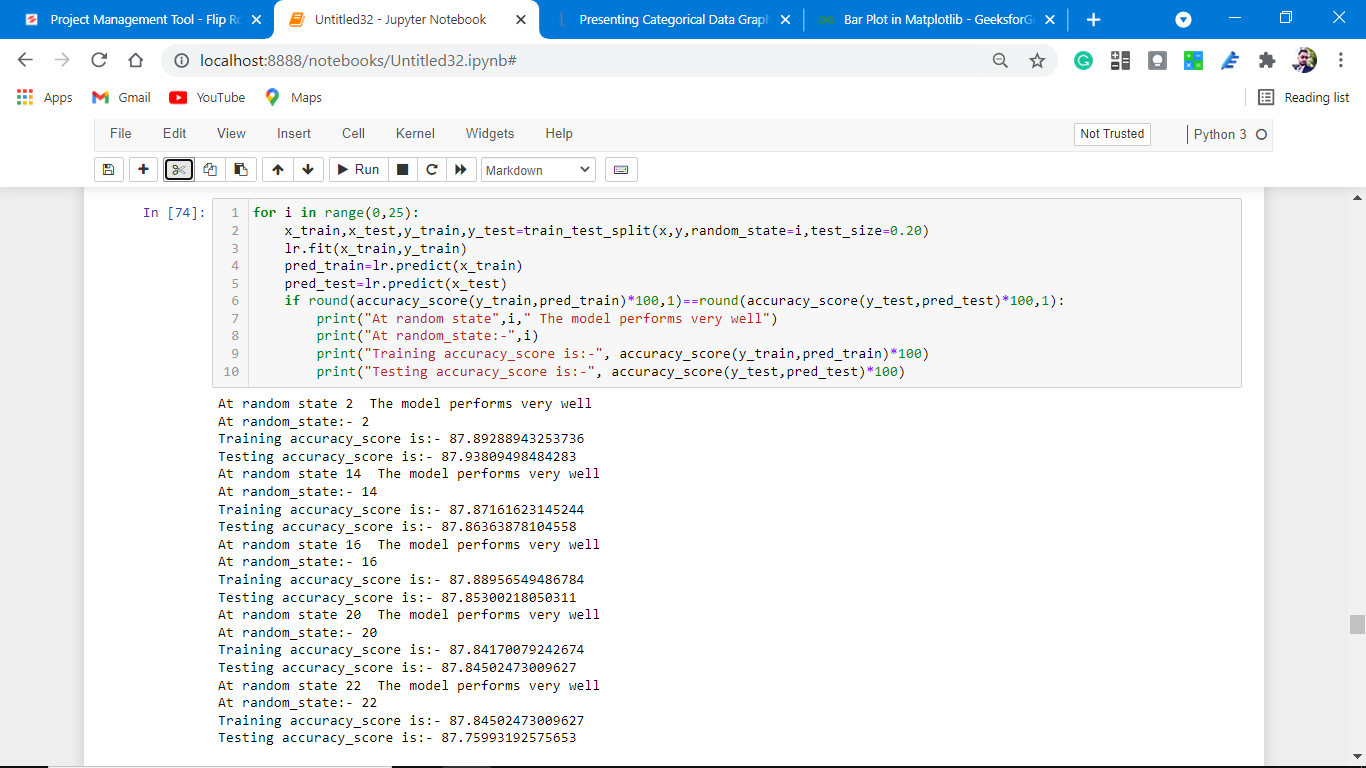
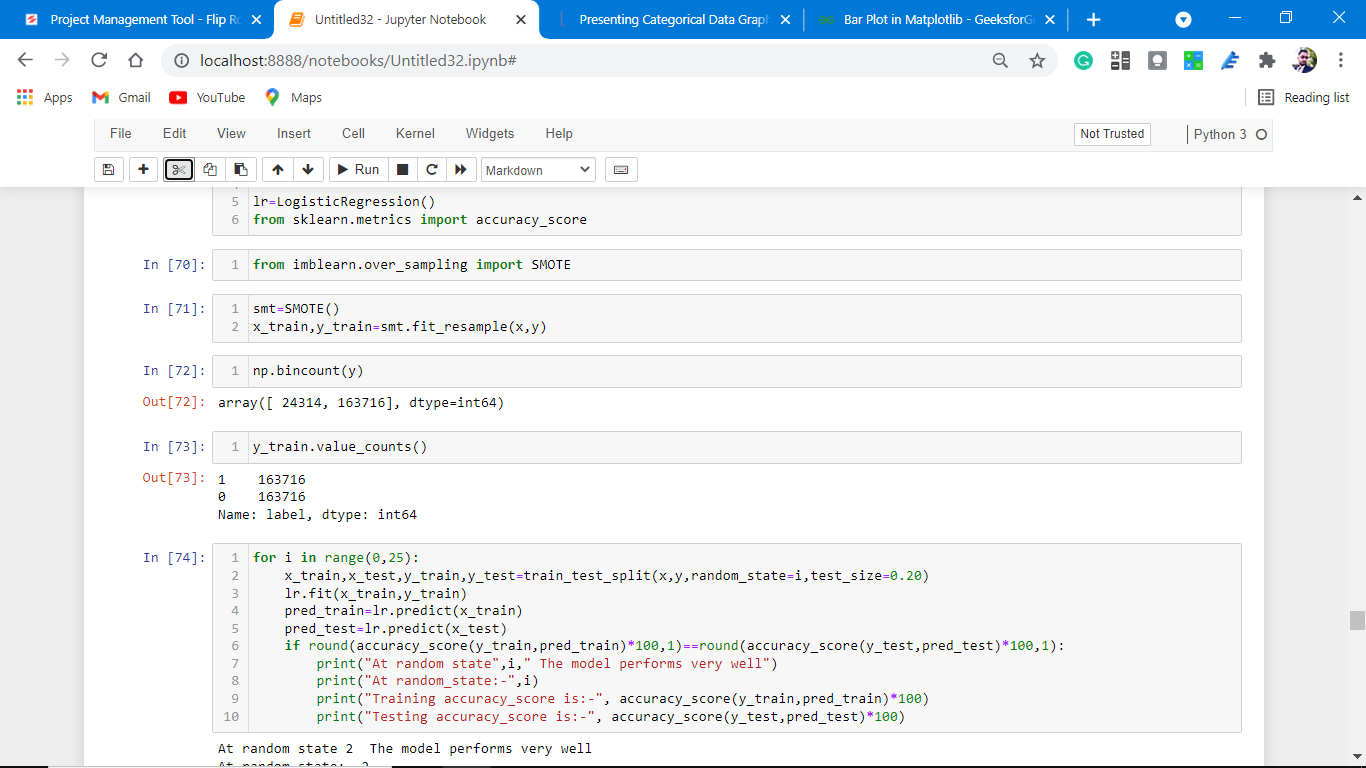
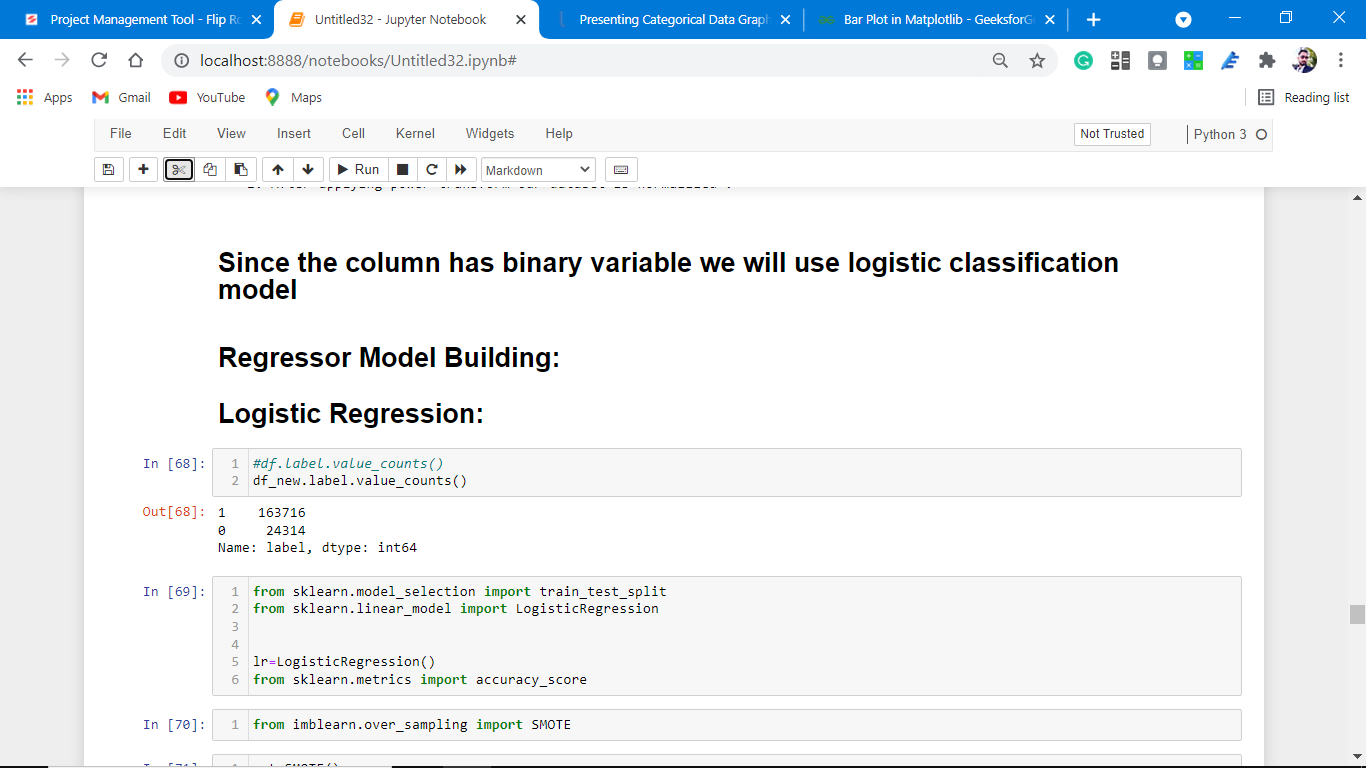
MODELS USED

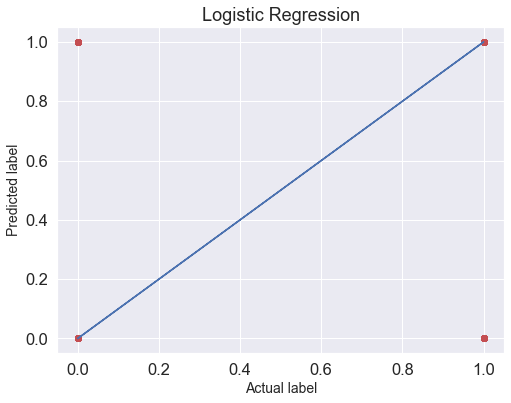
**Logistic Regression Model**

• Logistic Regression is a machine learning algorithm based on supervised learning.

• It performs a regression task. Regression models a target prediction value based on independent variables.

• It is mostly used for finding out the relationship between variables and forecasting.





**Observations:**

1. This Logistic Regression Performs with 88% accuracy for predicting labels.
2. We use the best-fit line.
3. from sklearn.metrics import accuracy\_score,confusion\_matrix,classification\_report
4. [[ 617 4256]
5. [ 315 32418]]
6. 0.8784502473009627
7. precision recall f1-score support
8. 0 0.66 0.13 0.21 4873
9. 1 0.88 0.99 0.93 32733
10. accuracy 0.88 37606
11. macro avg 0.77 0.56 0.57 37606
12. weighted avg 0.86 0.88 0.84 37606

from above we easily find out that

# Precision, recall,fl-score from the above plotting.

# Our model performs well on the initial level,

# Also, the cross-validation score at 3 is 87.88 which is almost the same as the accuracy of the model.

# Regularization

# To correctly fit in our model let's do some regulation.

# The [Ridge](https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Ridge.html#sklearn.linear_model.Ridge) regressor has a classifier variant: [RidgeClassifier](https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.RidgeClassifier.html" \l "sklearn.linear_model.RidgeClassifier" \o "sklearn.linear_model.RidgeClassifier). This classifier first converts binary targets to {-1, 1} and then treats the problem as a regression task, optimizing the same objective as above. The predicted class corresponds to the sign of the regressor’s prediction. For multiclass classification, the problem is treated as multi-output regression, and the predicted class corresponds to the output with the highest value.

# Screenshot (1224).pngScreenshot (1225).pngScreenshot (1226).png

# Screenshot (1227).png

# download (72).png

# Observations:

# 1. This model is non-performing well.

# 2. There is a major difference between cross-validation and accuracy matrices.

# Cross-Validation.

At cv:- 2

R2 Score: -14.462744289375662

Cross Val Score: 86.8590118598096

0.8708982609158114

# print(accuracy\_score(y\_test,pred\_decision))

# print(confusion\_matrix(y\_test,pred\_decision))

# print(classification\_report(y\_test,pred\_decision))

0.8708982609158114

[[ 2514 2359]

[ 2496 30237]]

precision recall f1-score support

0 0.50 0.52 0.51 4873

1 0.93 0.92 0.93 32733

accuracy 0.87 37606

macro avg 0.71 0.72 0.72 37606

weighted avg 0.87 0.87 0.87 37606

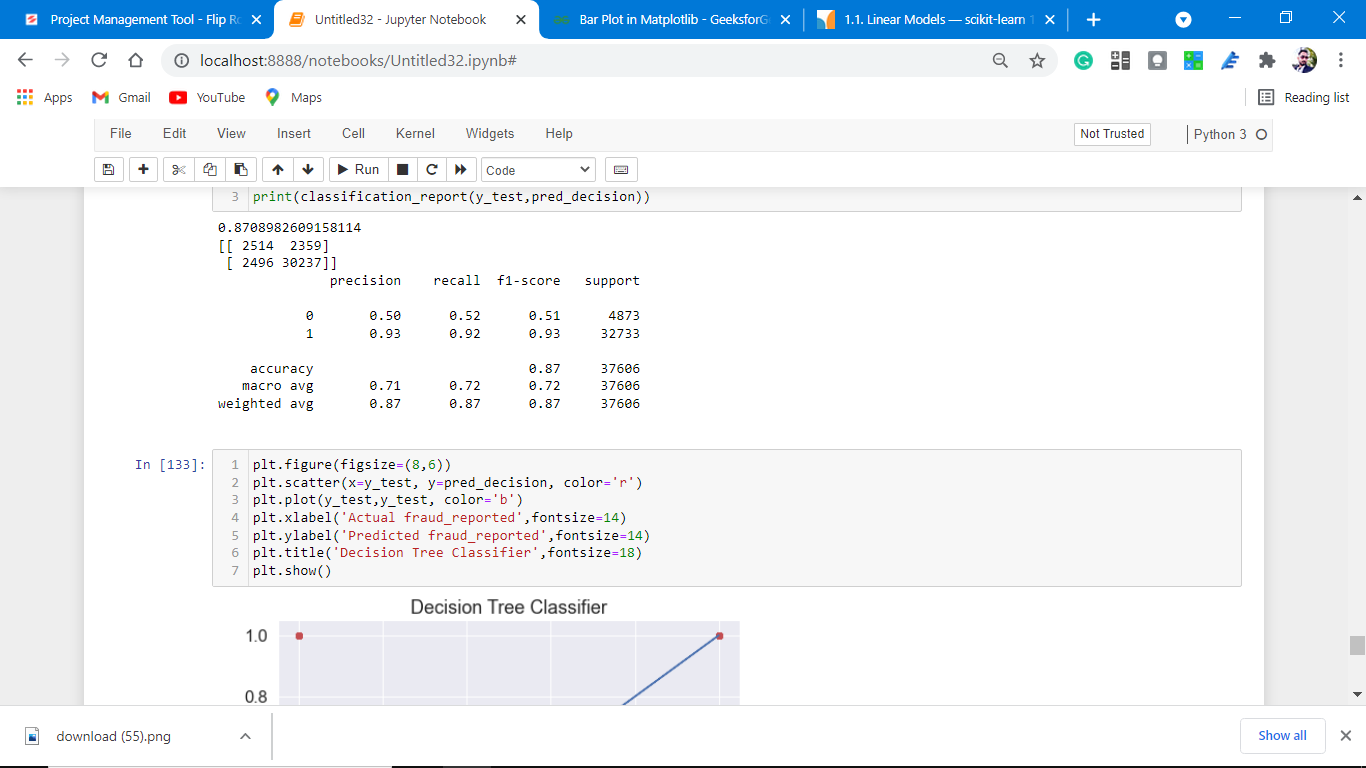
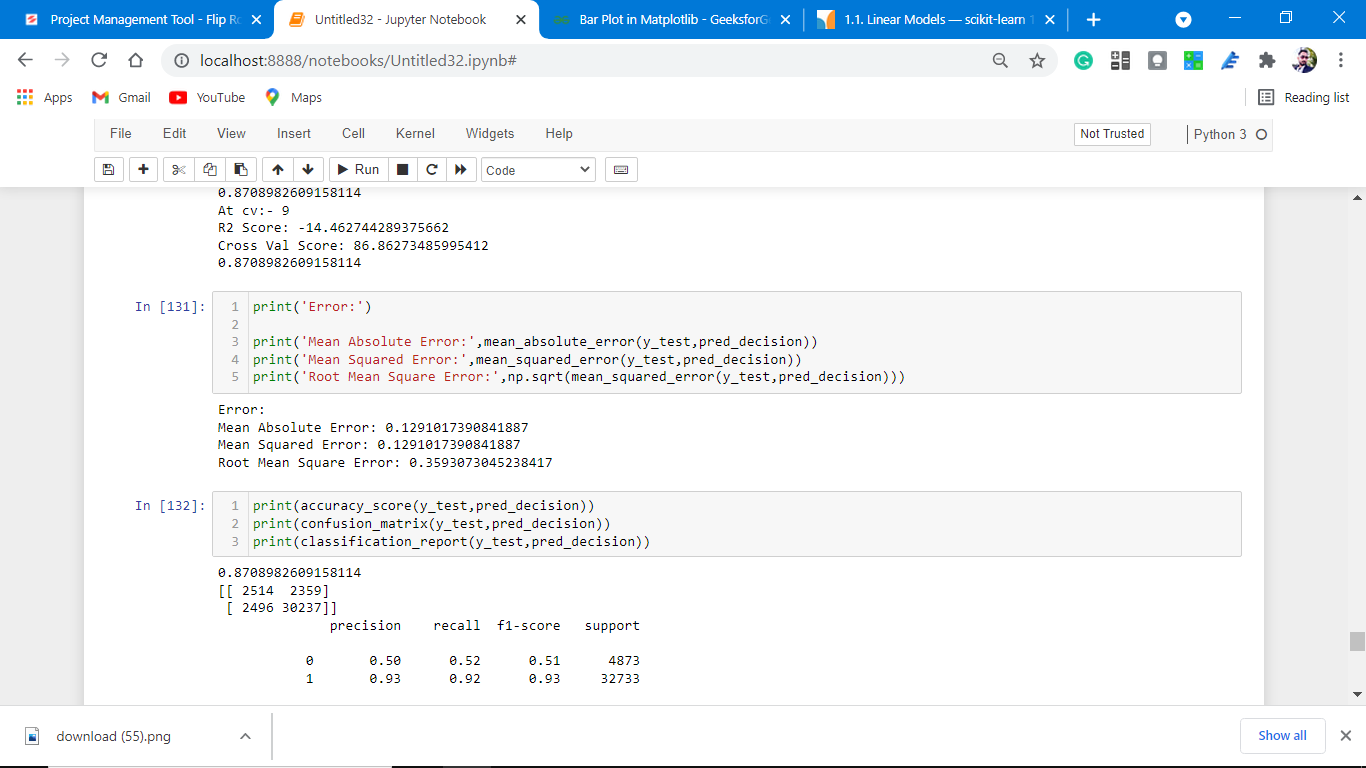
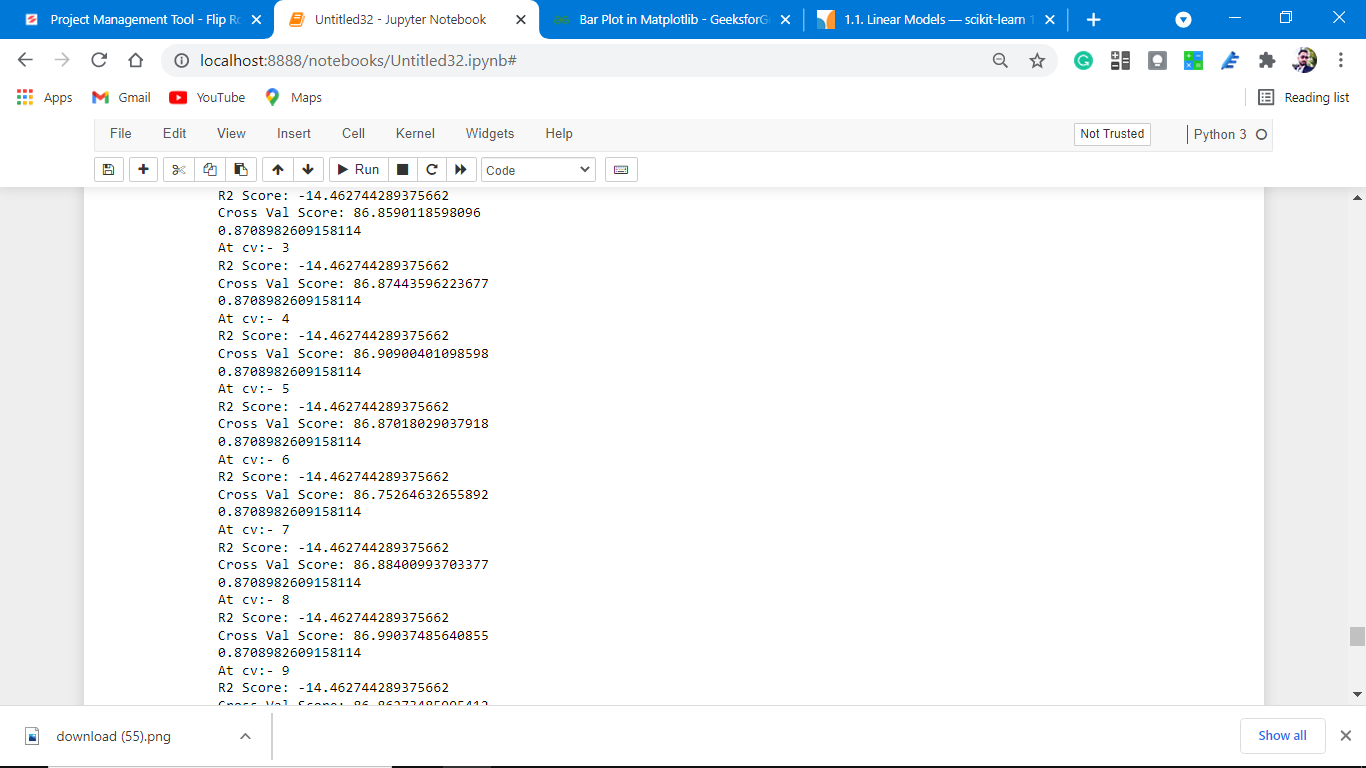
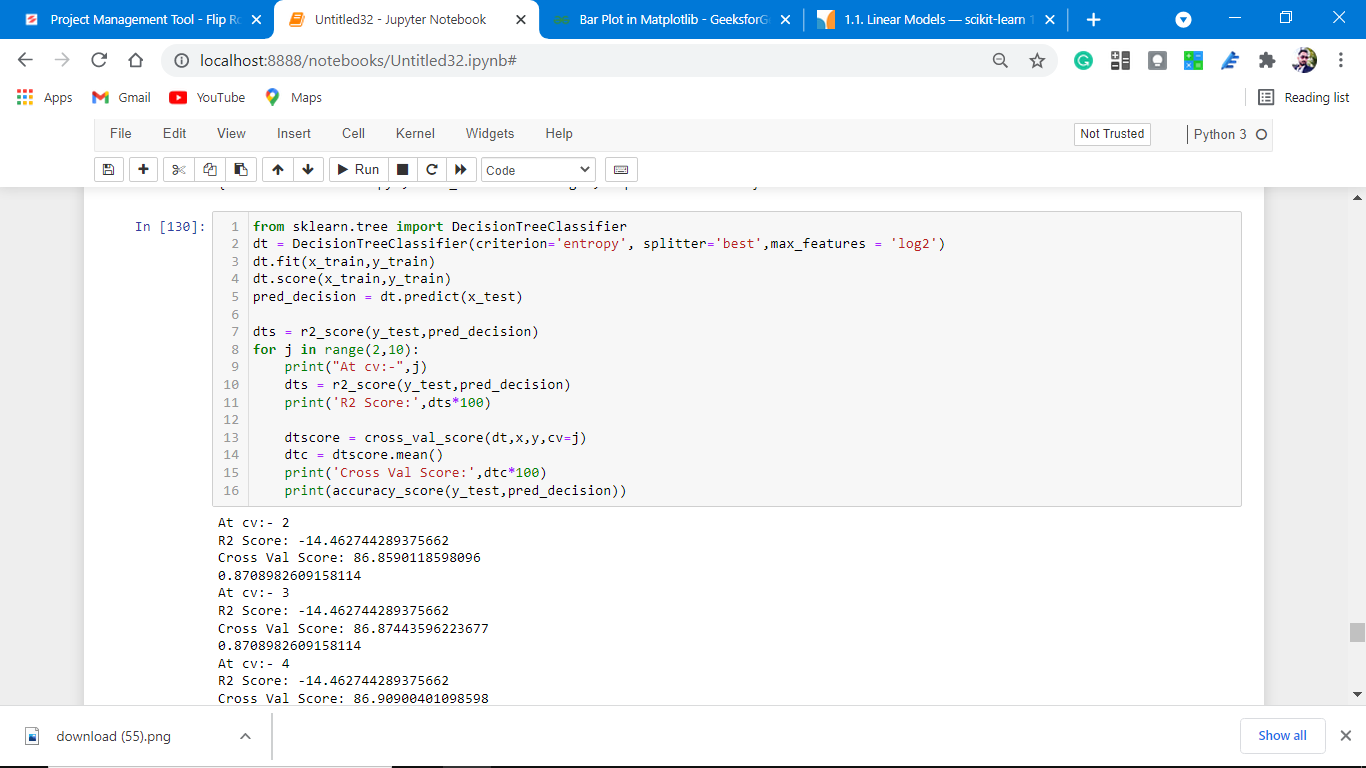
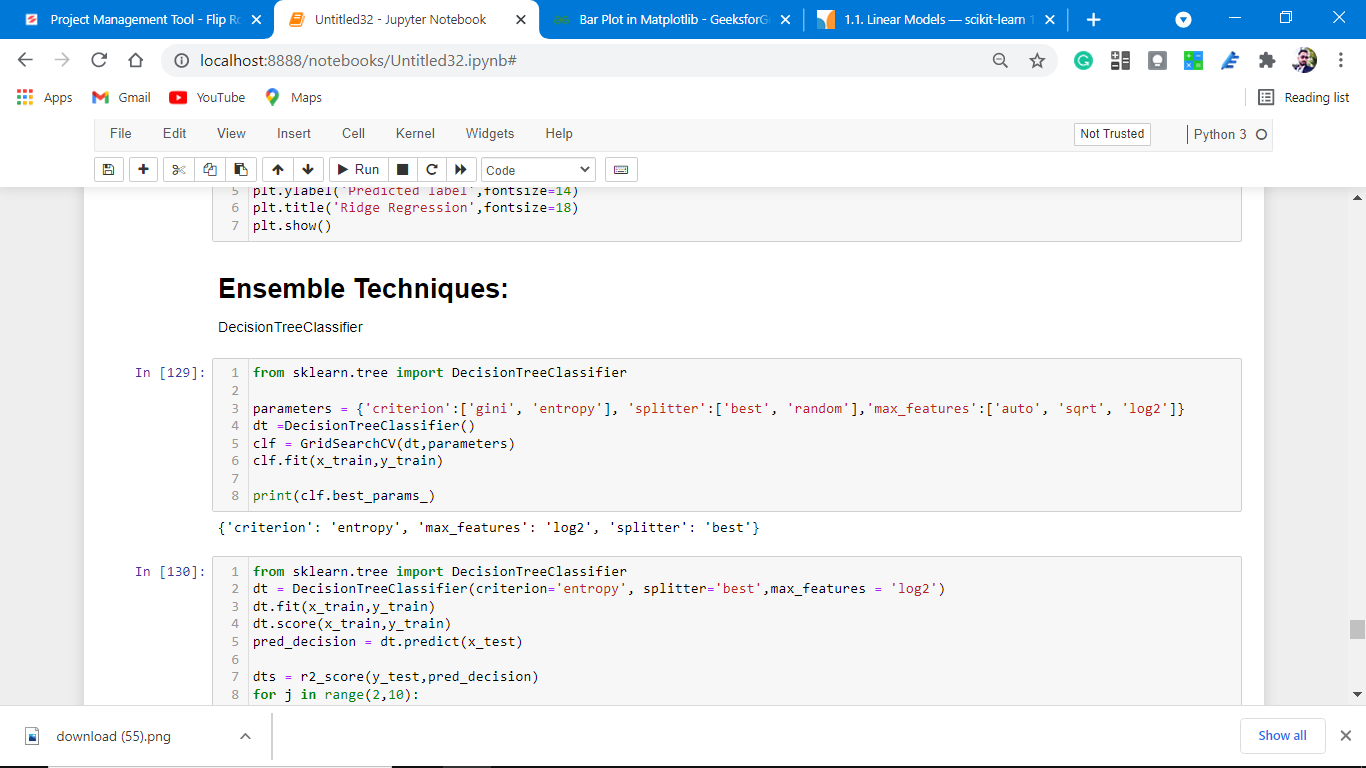
# Accuracy is almost the same but cross-validation is not well for this model.

# Ensemble Techniques:

Decision Tree Classifier

DecisionTreeClassifier is a class capable of performing multi-class classification on a dataset.

As with other classifiers, DecisionTreeClassifier takes as input two arrays: an array X, sparse or dense, of shape (n\_samples, n\_features) holding the training samples, and an array Y of integer values, shape (n\_samples,), holding the class labels for the training samples:





**Observations:**

1. This Decision Tree classifier Performs with 87% accuracy for predicting frauds.
2. After predicting and plotting the predicted data on the best fit line we observe that DT-C is not so accurate.
3. CV is not well. And does not give accurate results.

# Cross-Validation.

At cv:- 2

R2 Score: -14.462744289375662

Cross Val Score: 86.8590118598096

0.8708982609158114

# print(accuracy\_score(y\_test,pred\_decision))

# print(confusion\_matrix(y\_test,pred\_decision))

# print(classification\_report(y\_test,pred\_decision))

0.8708982609158114

[[ 2514 2359]

[ 2496 30237]]

precision recall f1-score support

0 0.50 0.52 0.51 4873

1 0.93 0.92 0.93 32733

accuracy 0.87 37606

macro avg 0.71 0.72 0.72 37606

weighted avg 0.87 0.87 0.87 37606

# Accuracy is almost the same but cross-validation is not well for this model.

* Key Metrics for success in solving a problem under consideration

Confusion matrix, Mean Absolute Error, Mean Squared Error, Root Mean Square Error

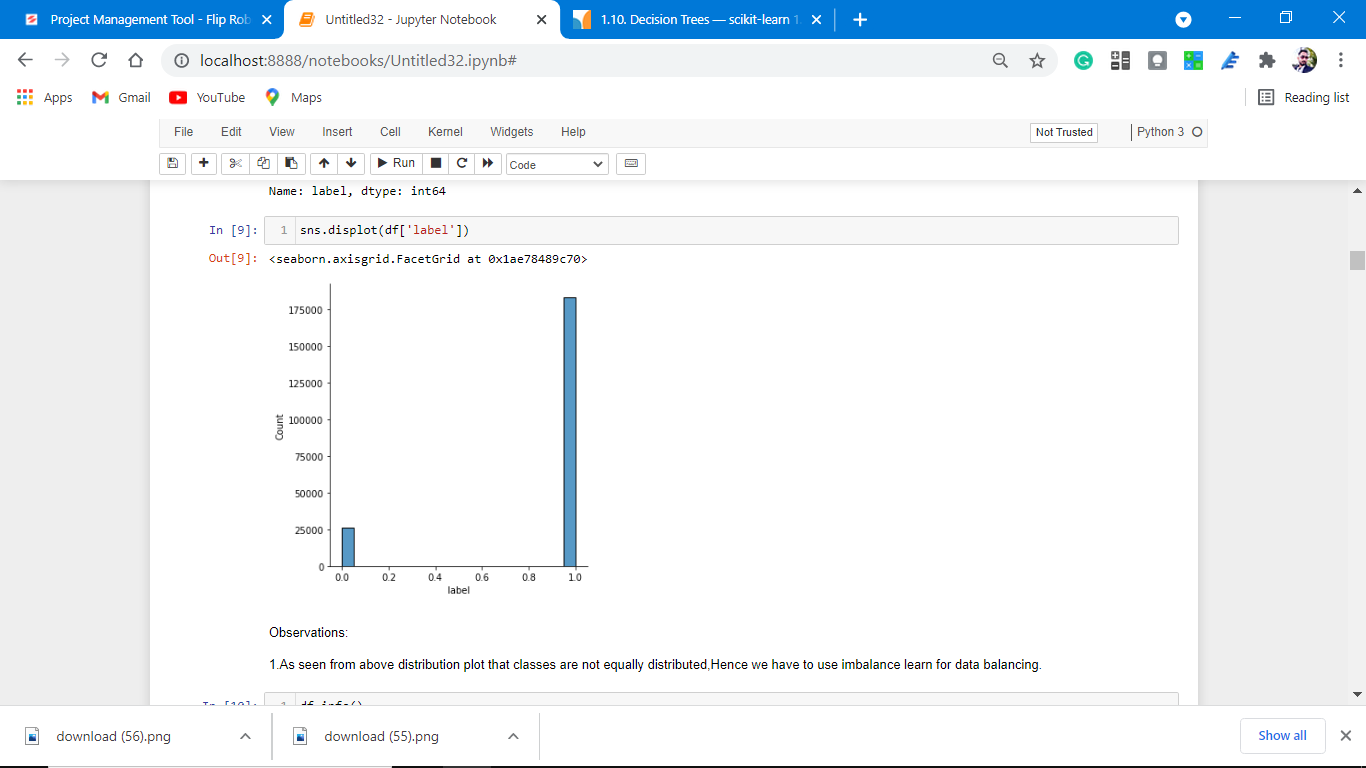
This matrix helps to understand the model more deeply.

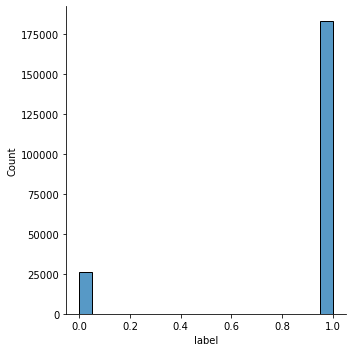
* **Visualizations**

Data visualization is the graphical representation of information and data. By using charts, plots, and graphs data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

In the world of Big Data, data visualization tools and technologies are essential to analyze massive amounts of information and make data-driven decisions.

Starts from label column analysis.(target variable)



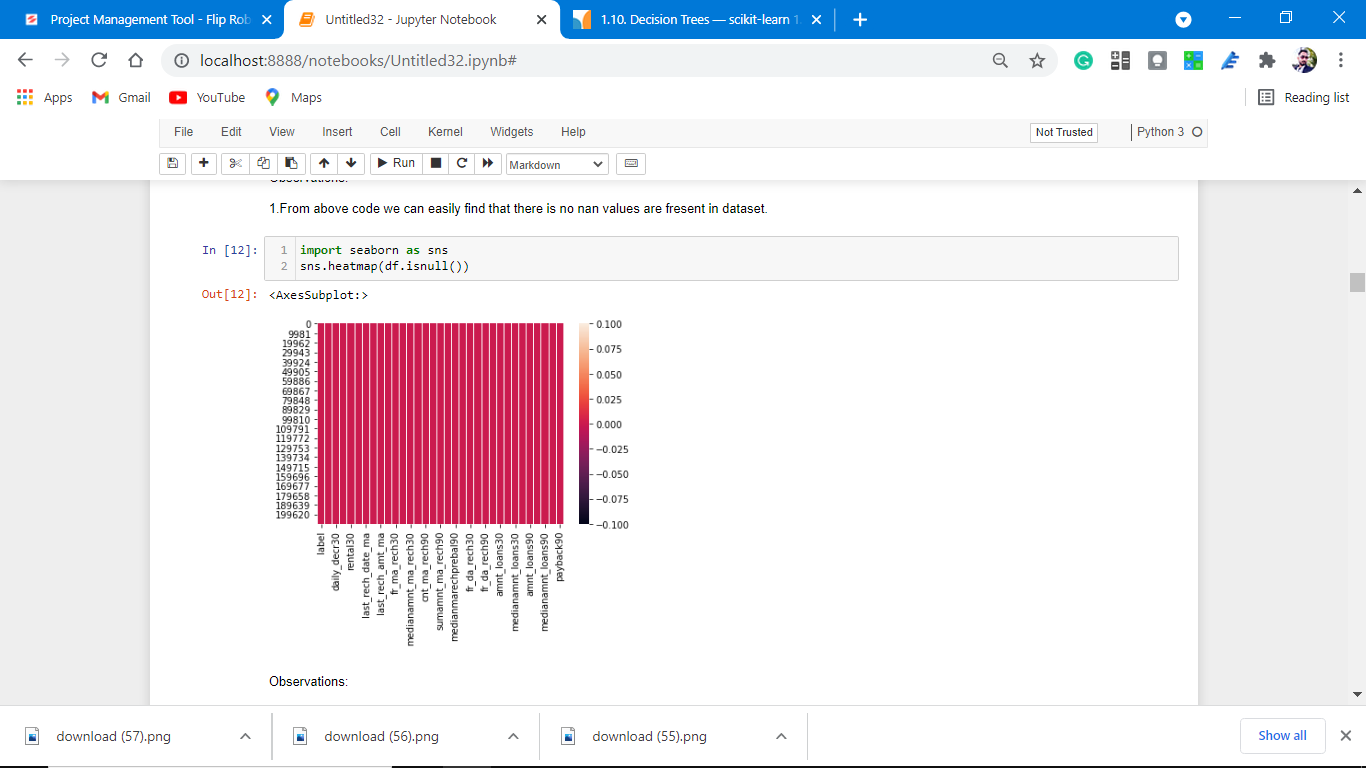


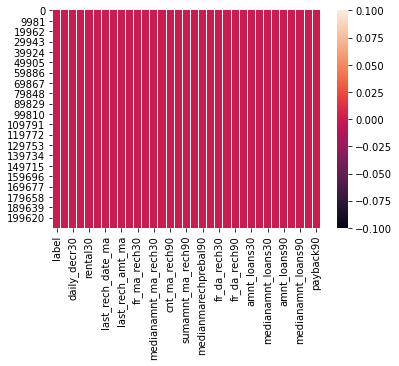
**Observations:**

1. As seen from the above distribution plot that classes are not equally distributed, Hence we have to use imbalance learn for data balancing.

2. The dataset is imbalanced. Label ‘1’ has approximately 87.5% records, while, label ‘0’ has approximately 12.5% records.

**Heat Map Plotting for null values.**

****

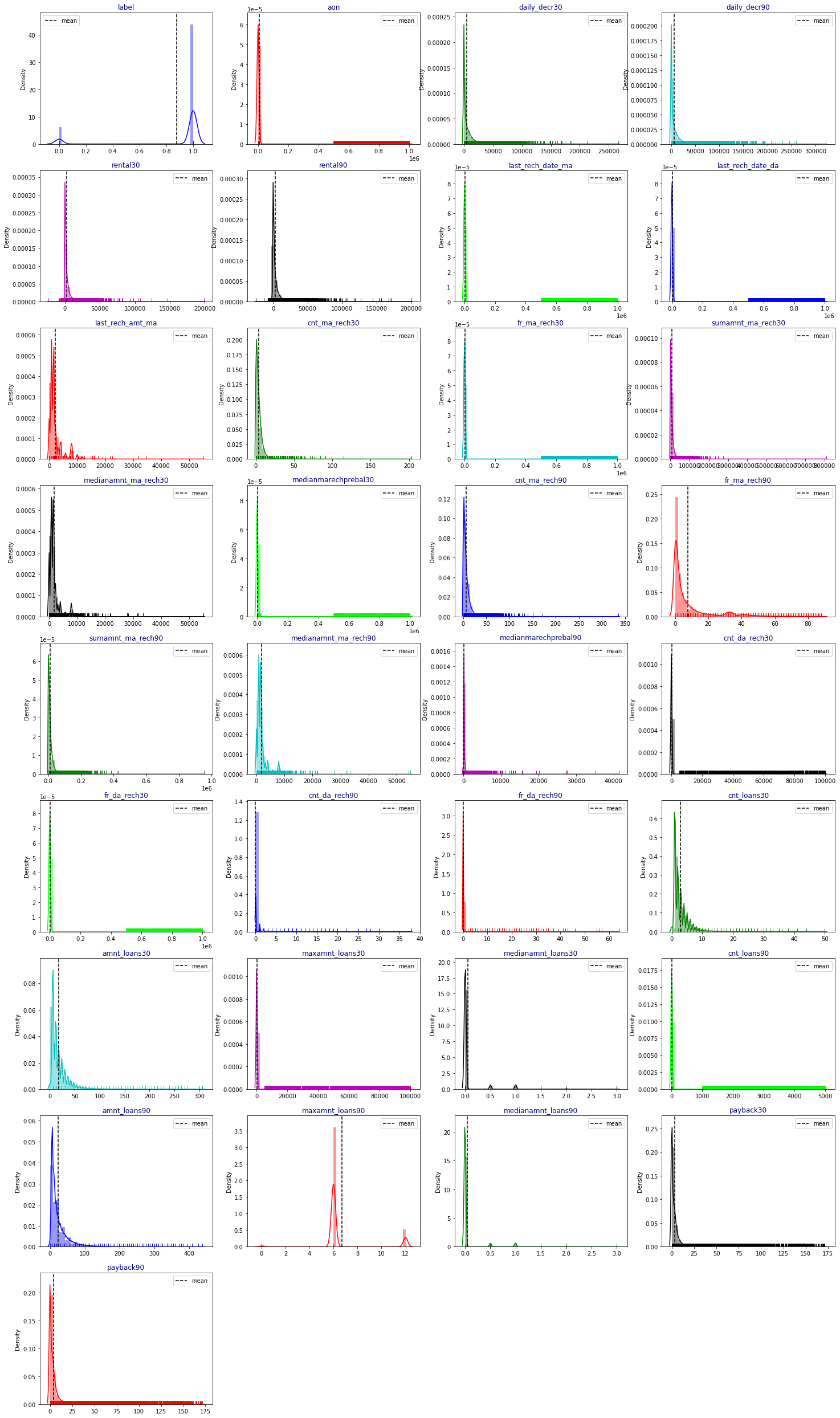
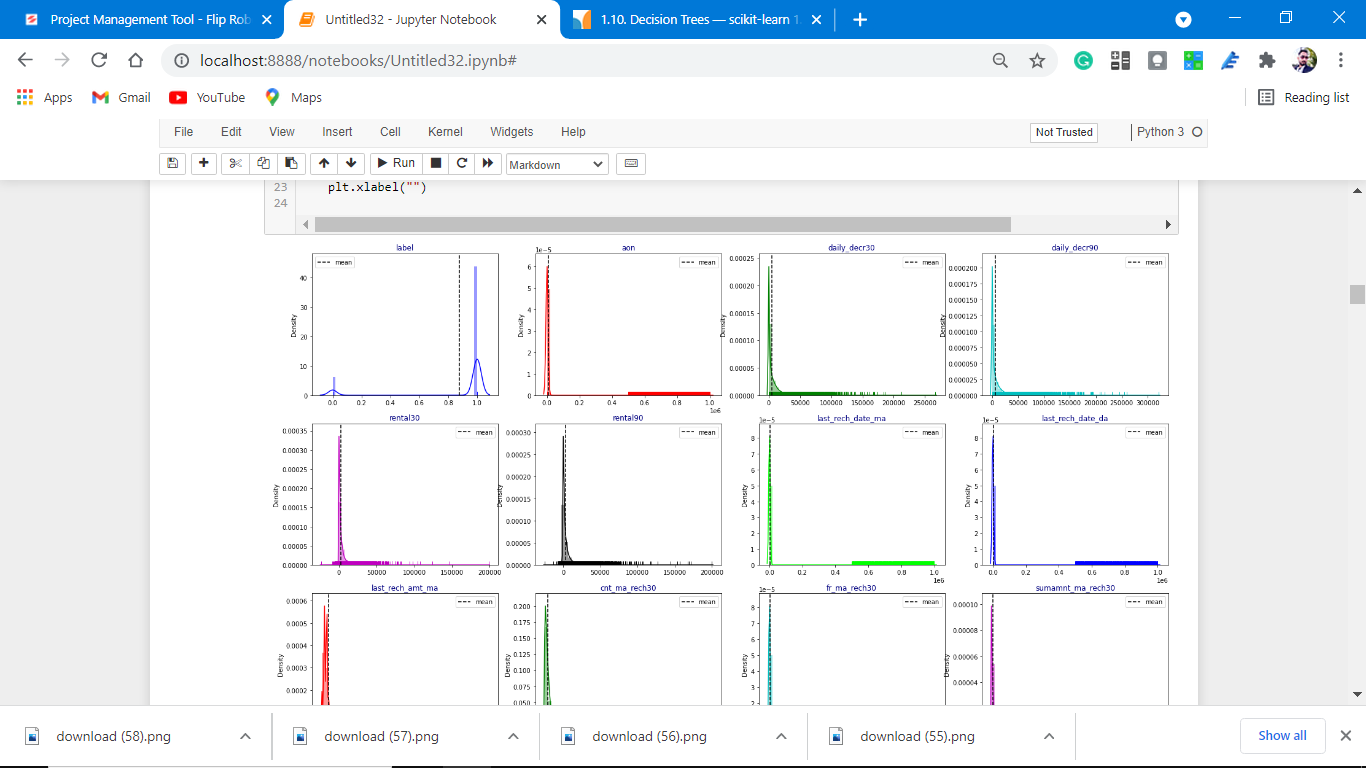
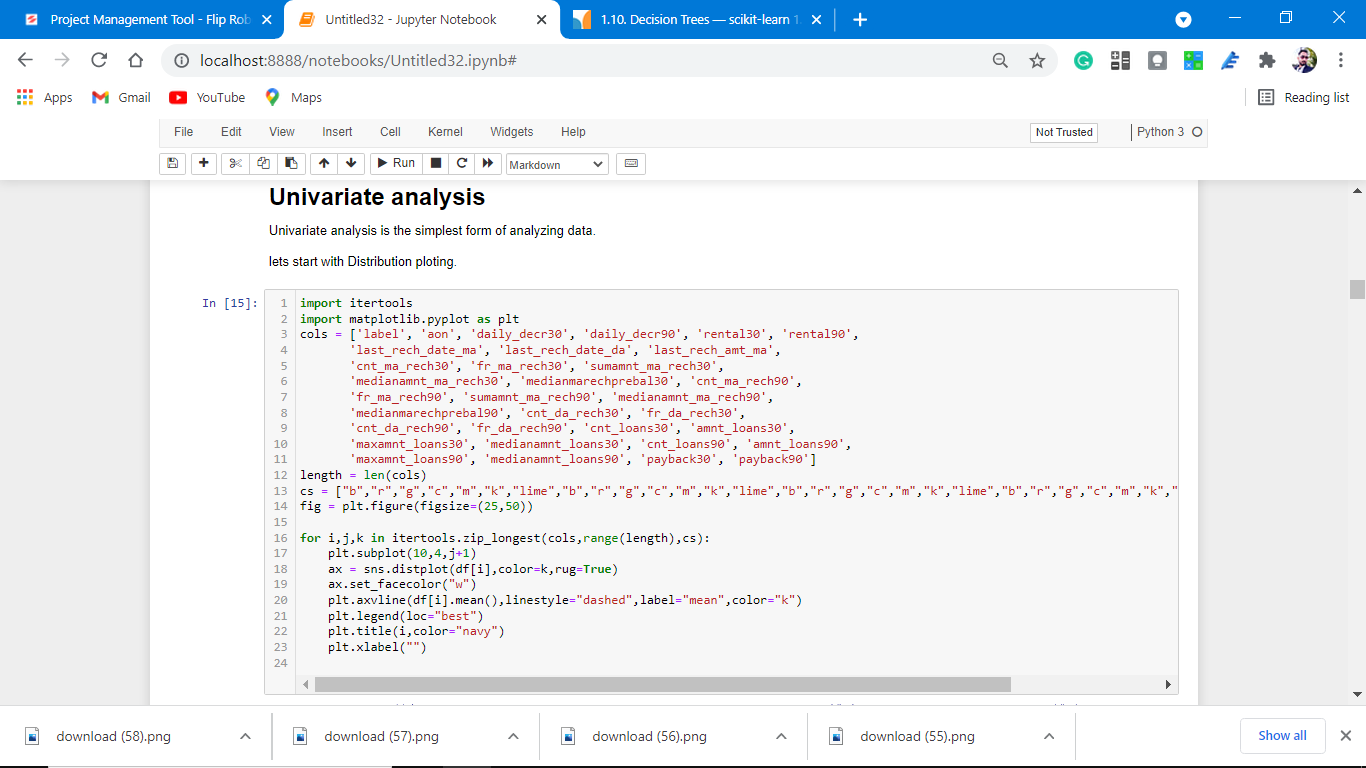
****

Observations:

1. As observed from the above there is no null values count

, we can easily visualize this heat map plotting.

**Univariate Analysis.**



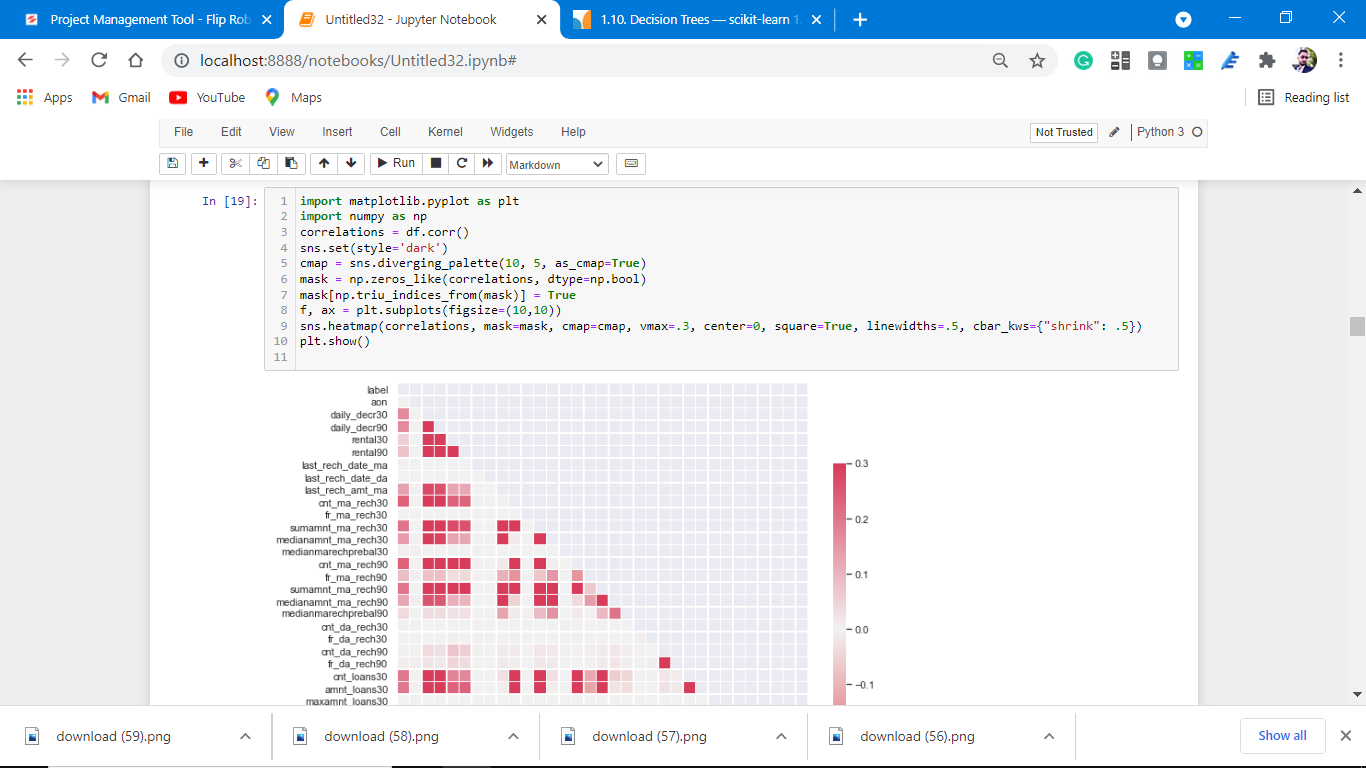
**Observations:**

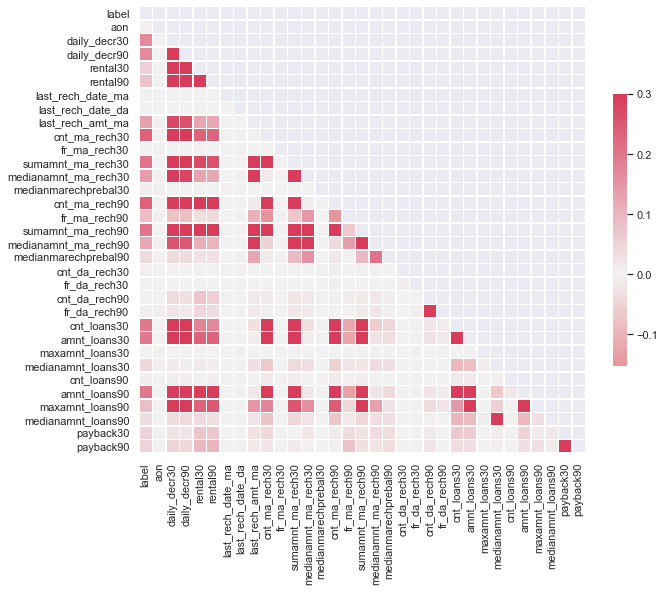
1. From the above plotting of distribution plot we see that some features columns are not normally distributed.

2. some columns are skewed towards the right.

3. Building blocks are out of the normal curve hence outliers are present.

**Correlation with the target variable.**

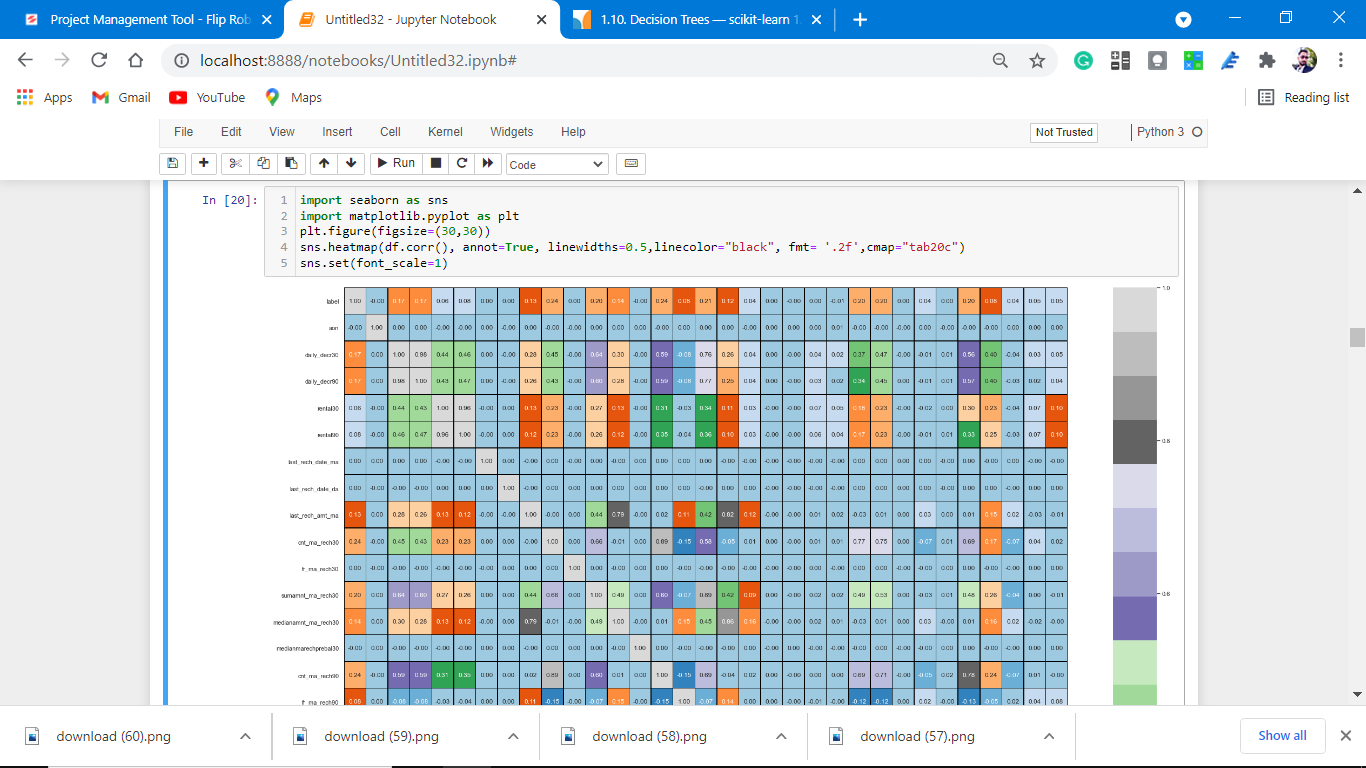
****

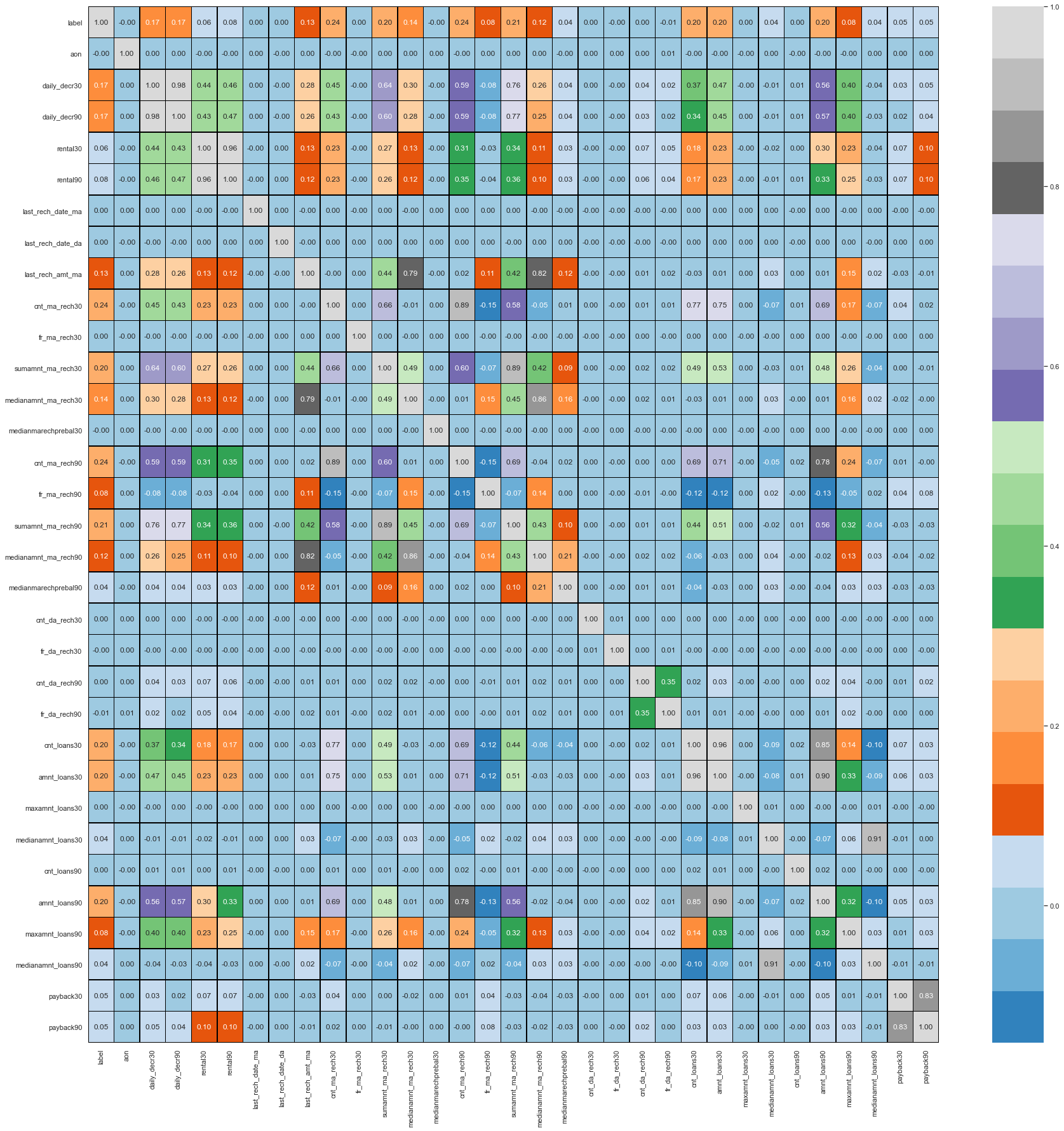
****

observations:

1. As from the above coding multicollinearity is present in the dataset.

2. Color blocks are having similar colors.





Observations:

1. From the above result it is clear that some columns make a positive correlation and some make a negative correlation.

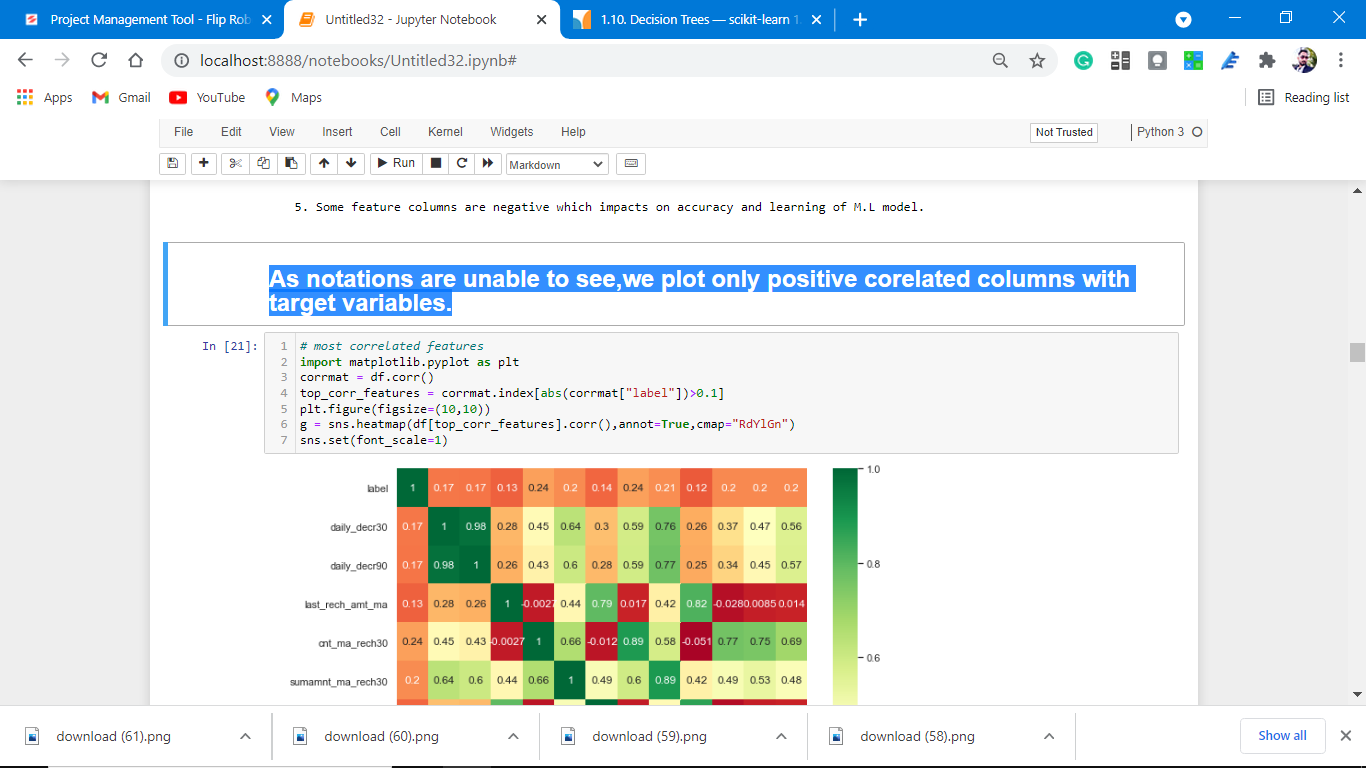
2. The positively correlated columns have a great impact on the target column while the negatively correlated have less or zero impact on the target column.

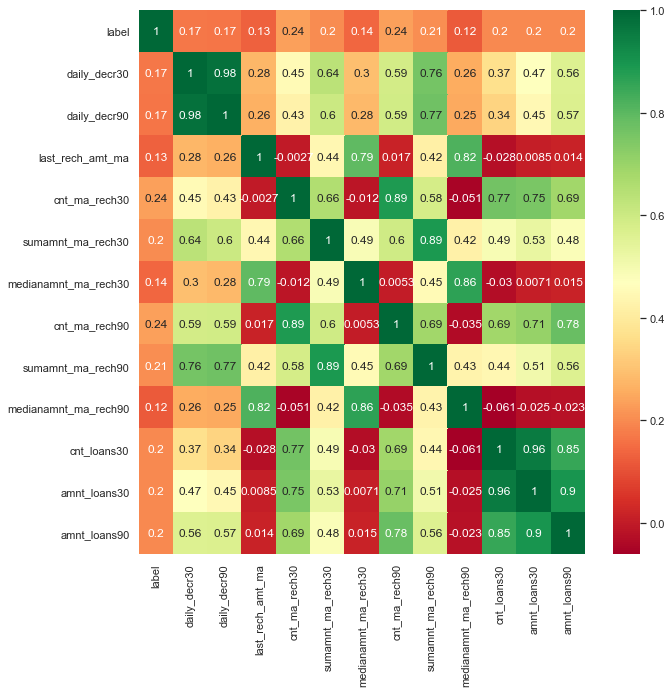
3. when we plot the heat map with notations is true. we can easily observe multicollinearity between the columns.

4. Some feature columns with the same names but with different periods have collinearity.

5. Some feature columns are negative which impacts on accuracy and learning of the M.L model.

**As notations are unable to see, we plot only positive correlated columns with target variables.**

****

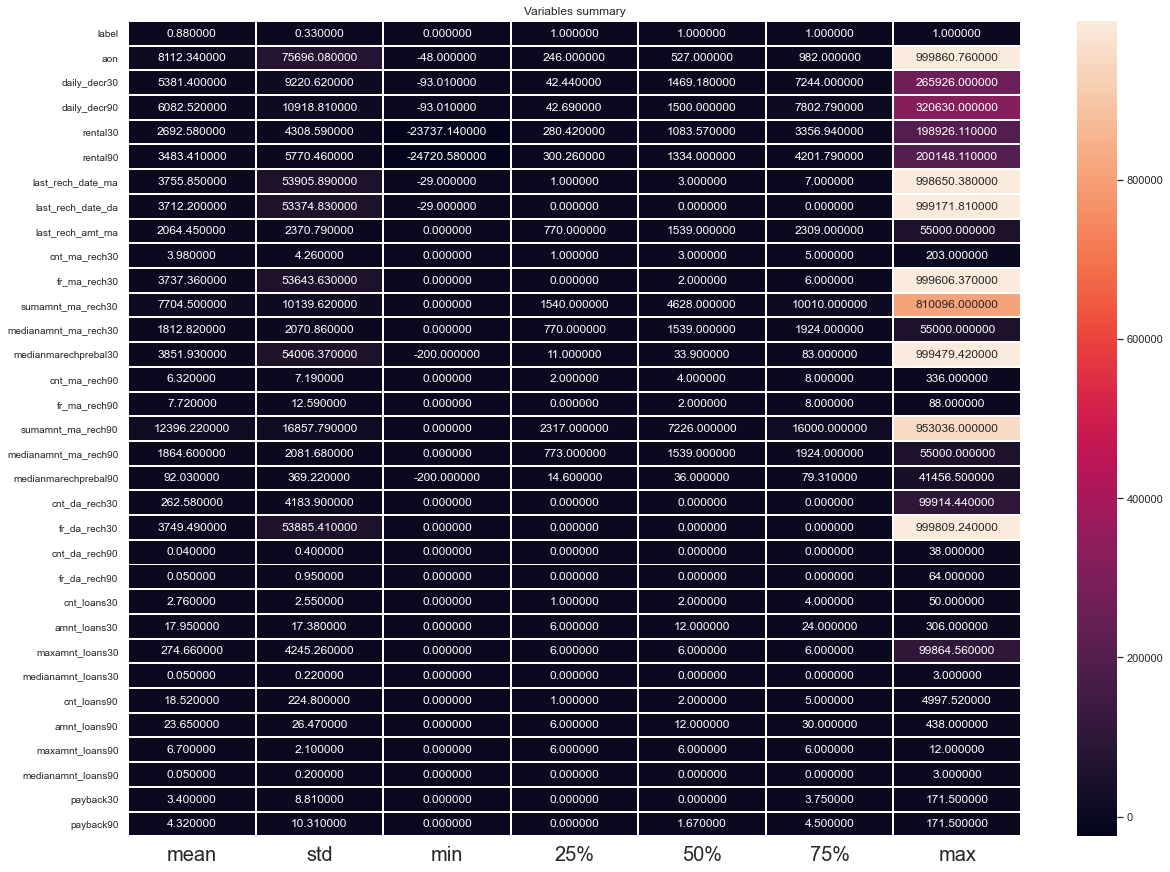
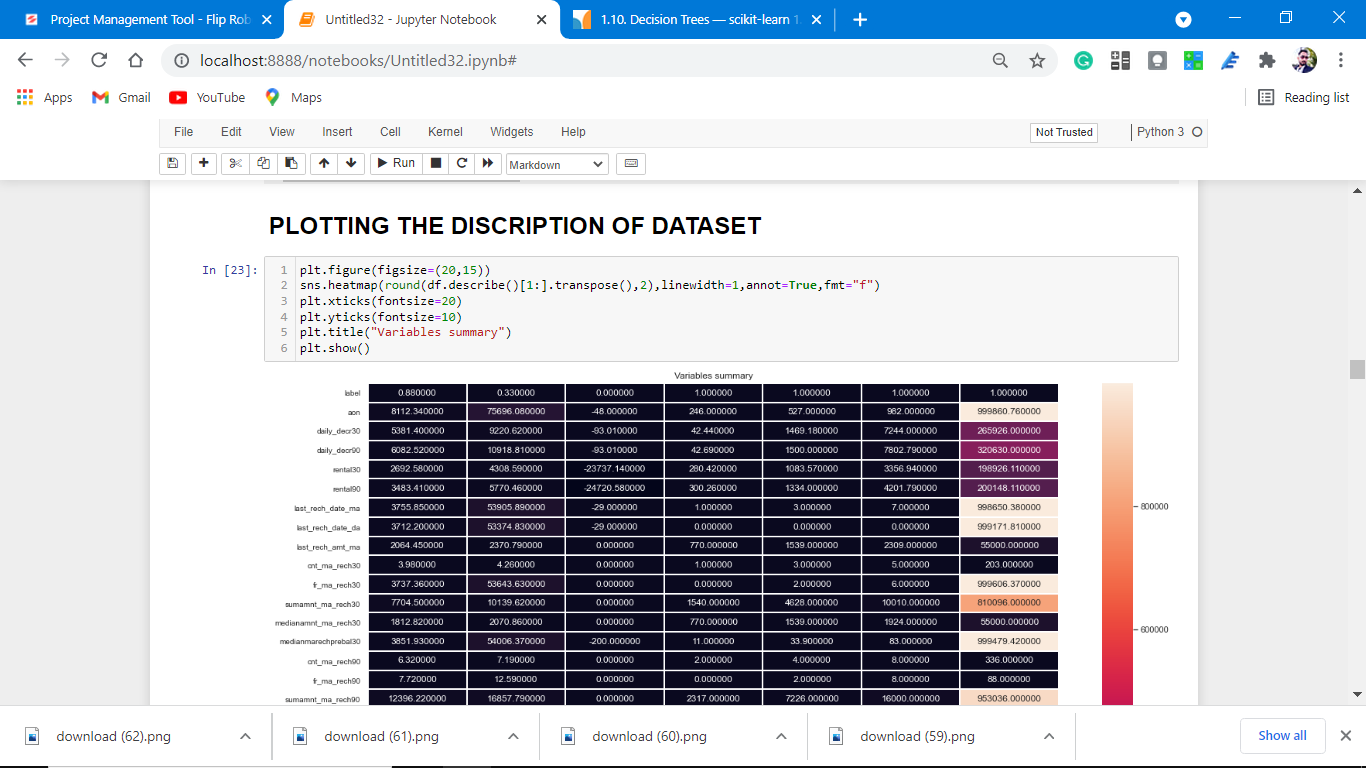
****

Correlation:

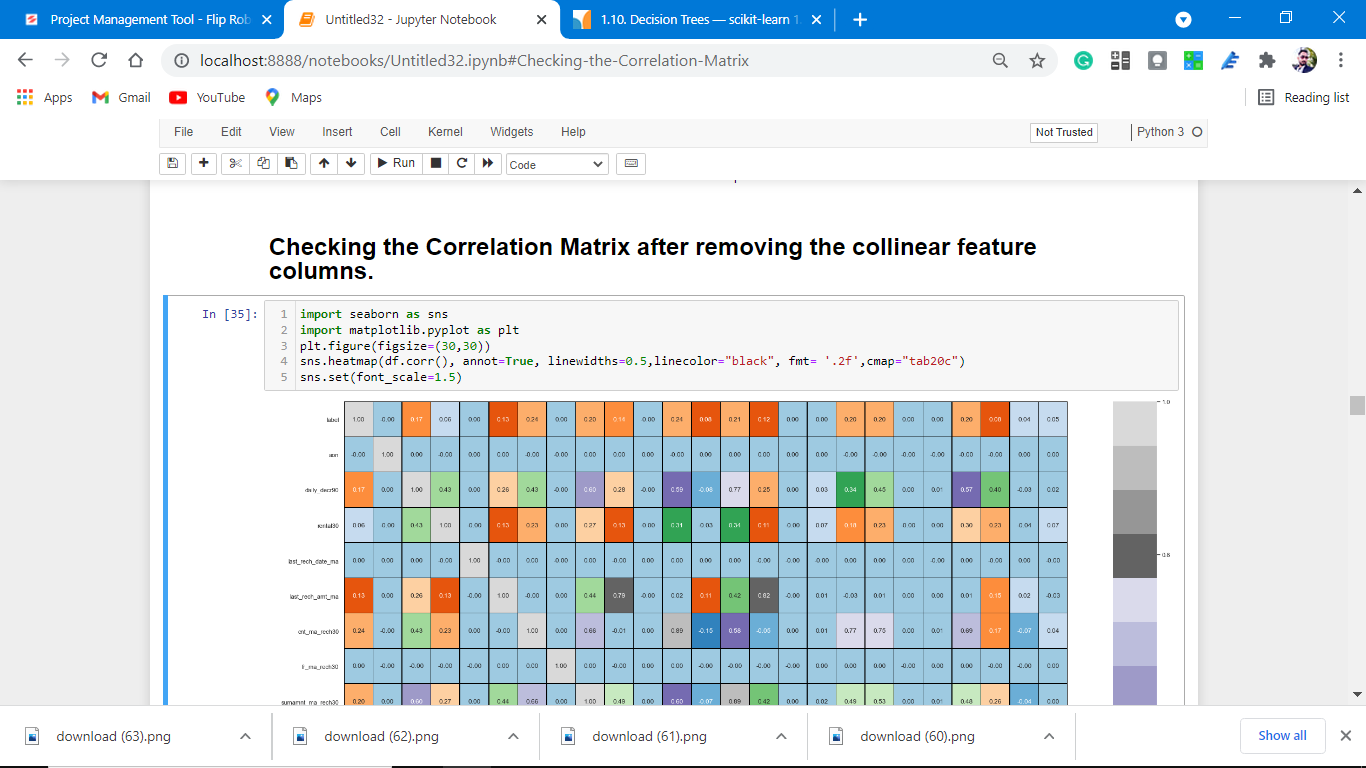
1. These are positive feature columns with target variables.

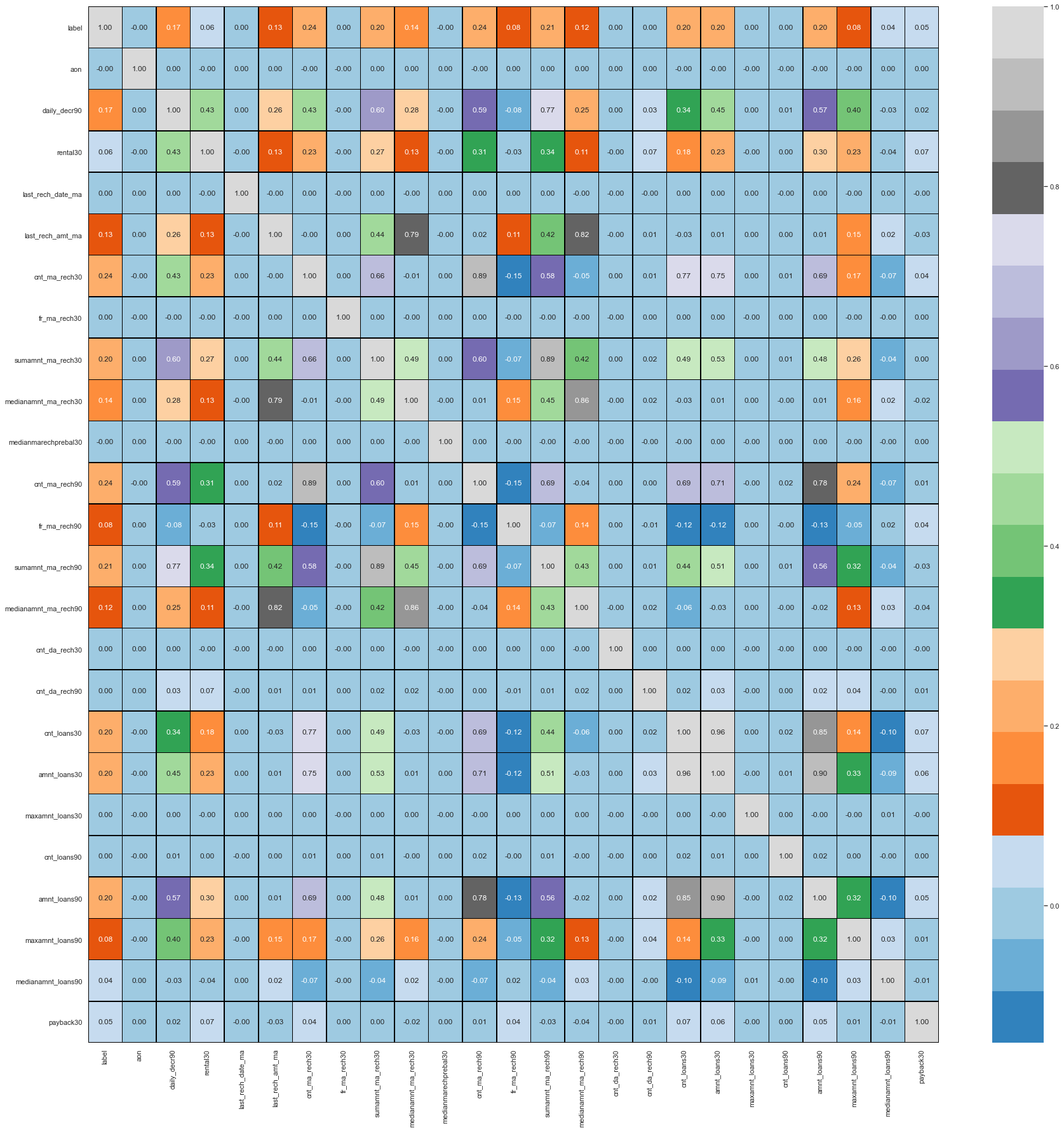
2. These columns make the M.l model more accurate.

**Discription of the dataset.**

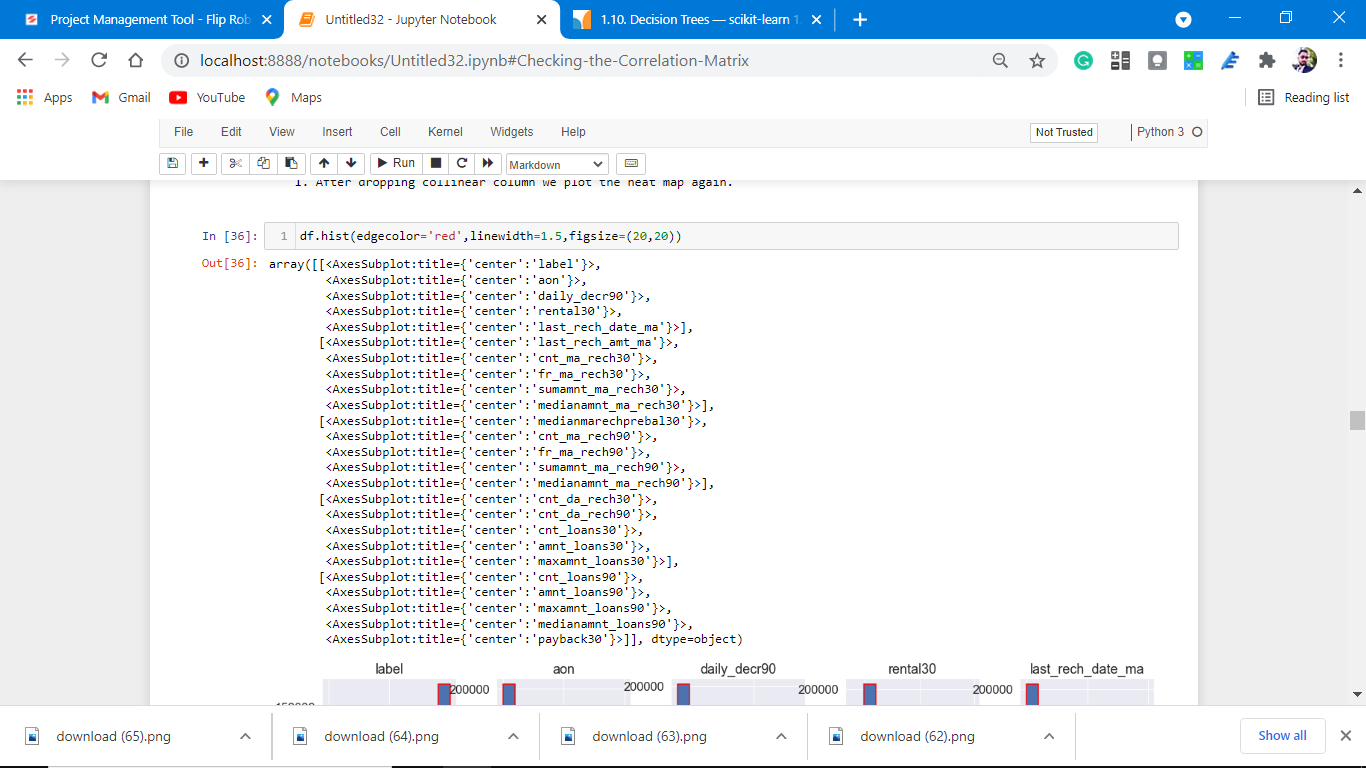
****

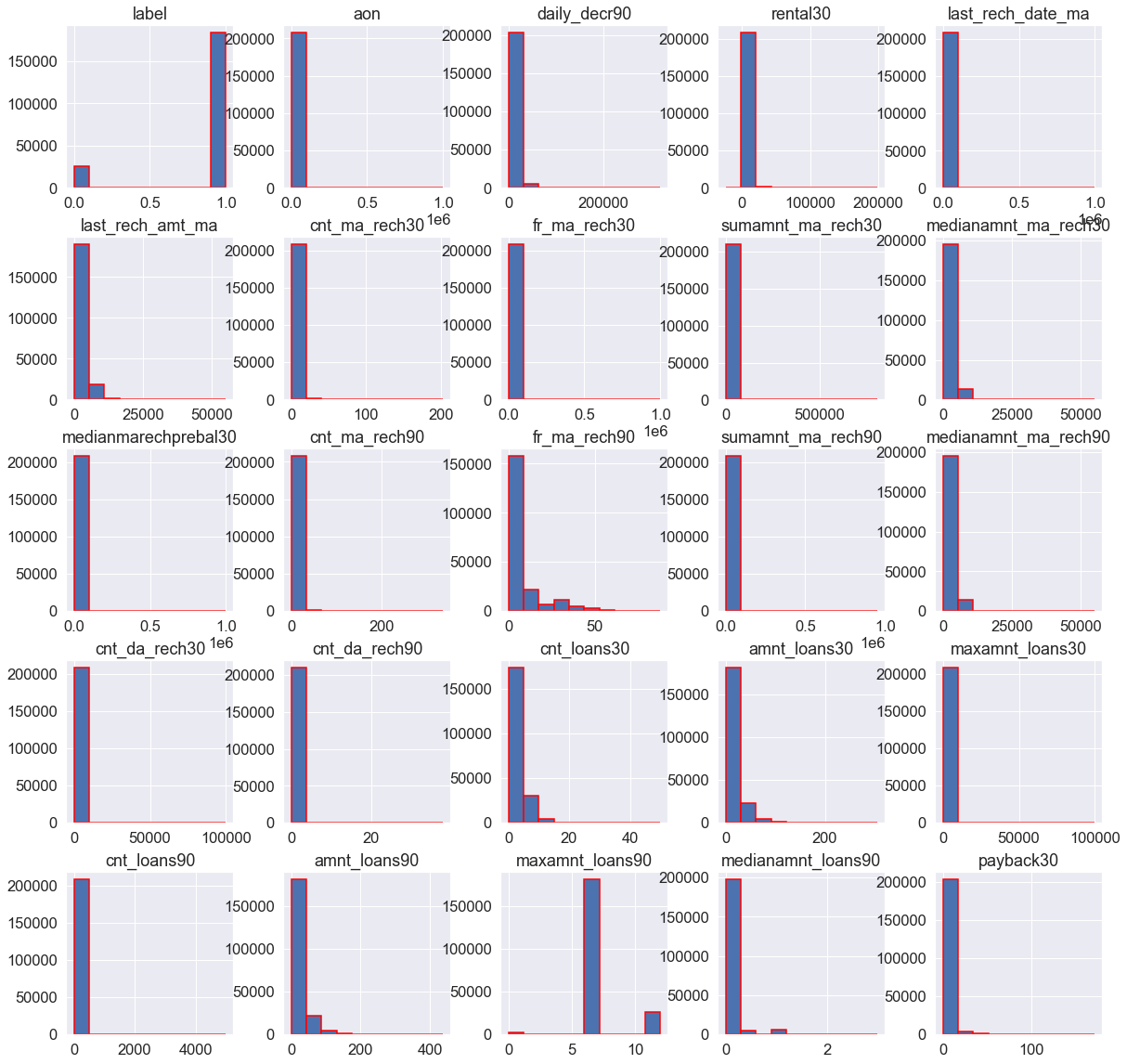
**Checking the Correlation Matrix after removing the collinear feature column.**

****

****

**Histograms of the dataset**

****

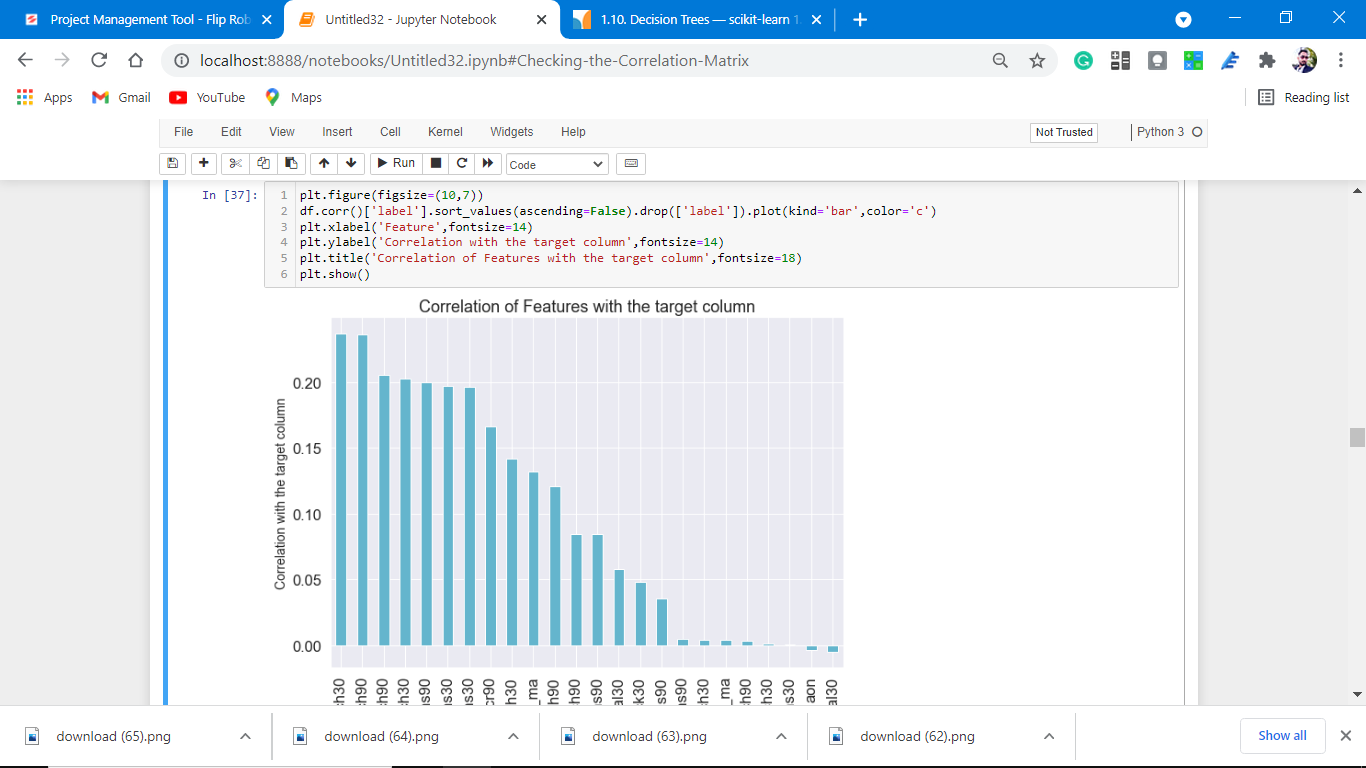
****

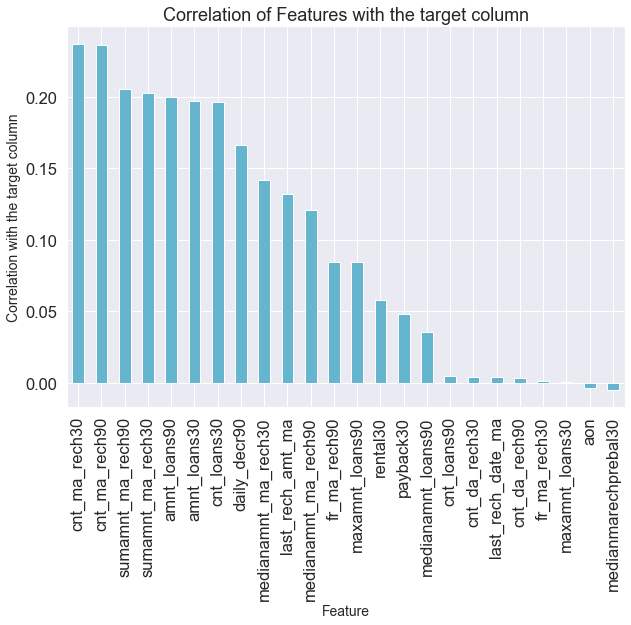
Observations:

1. Data is not distributed; evenly.

**Correlation plotting with target column with Descending order.**

Where maximum impactful column is placed first and negative are plotted in last.





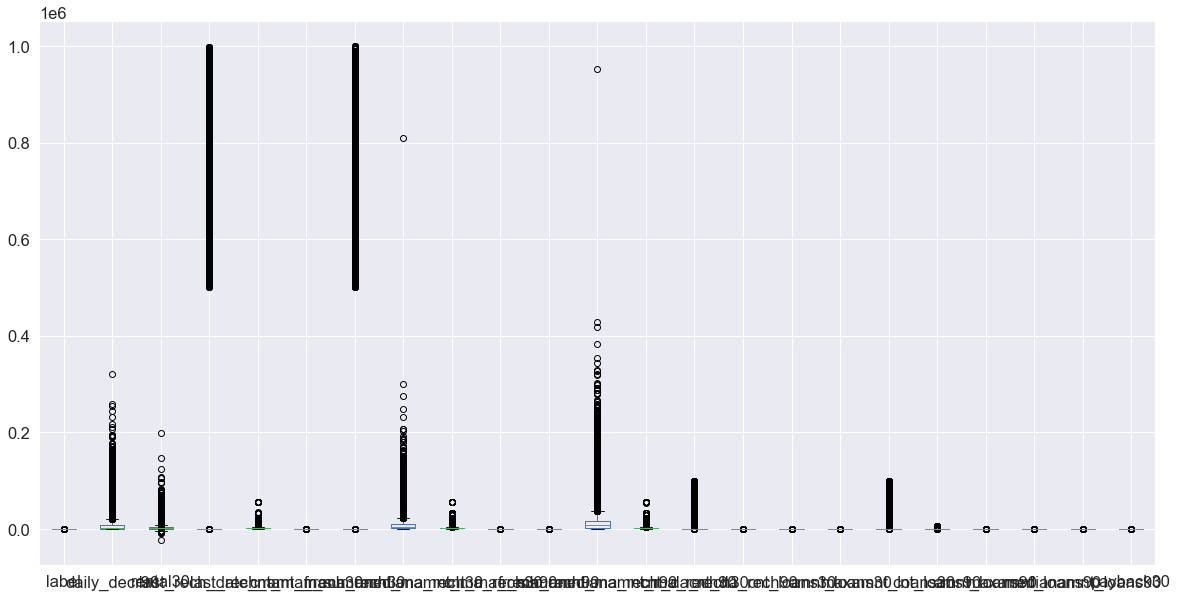
Observations:

1. There is a negative feature column with the target variable.

2. aon and medianmarechprebal30 are negative feature columns with target columns.

**Detecting Outliers**

****

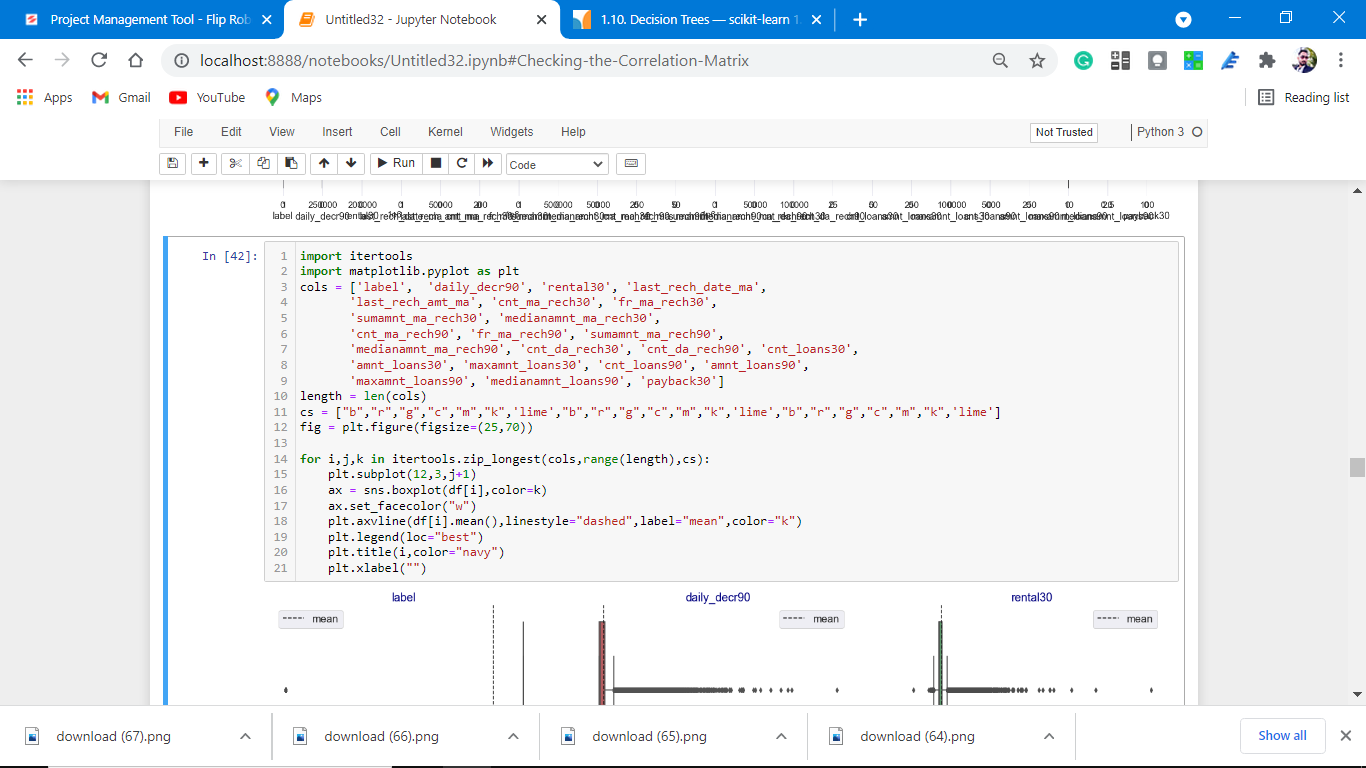
****

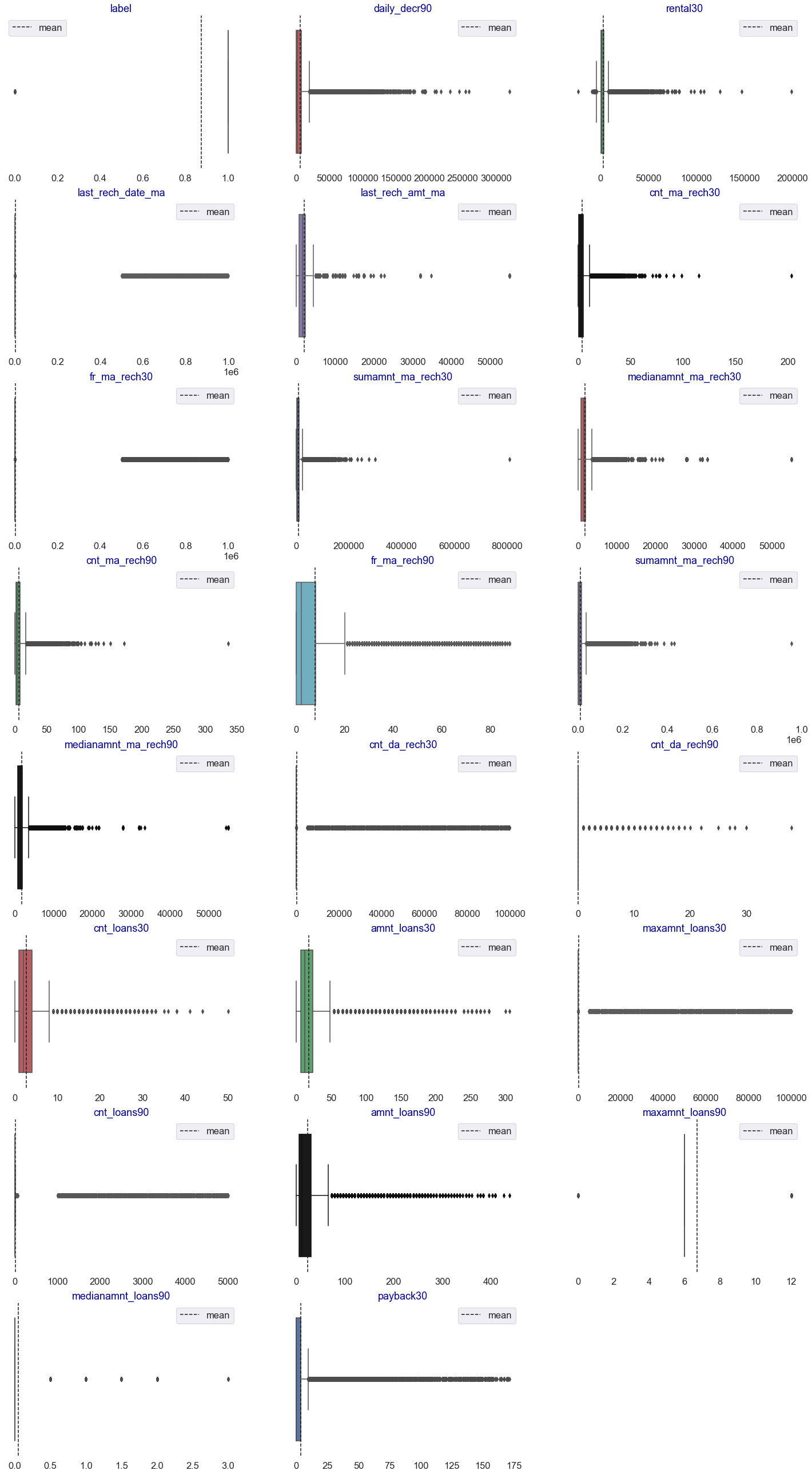
**Observations:**

**1. From the above coding of skewness we predict that outliers are present we are right after plotting the box plot.**

**2. Outliers are the silent killer of the accuracy of the dataset we have to remove them.**

**Plotting in a more detailed form.**

****

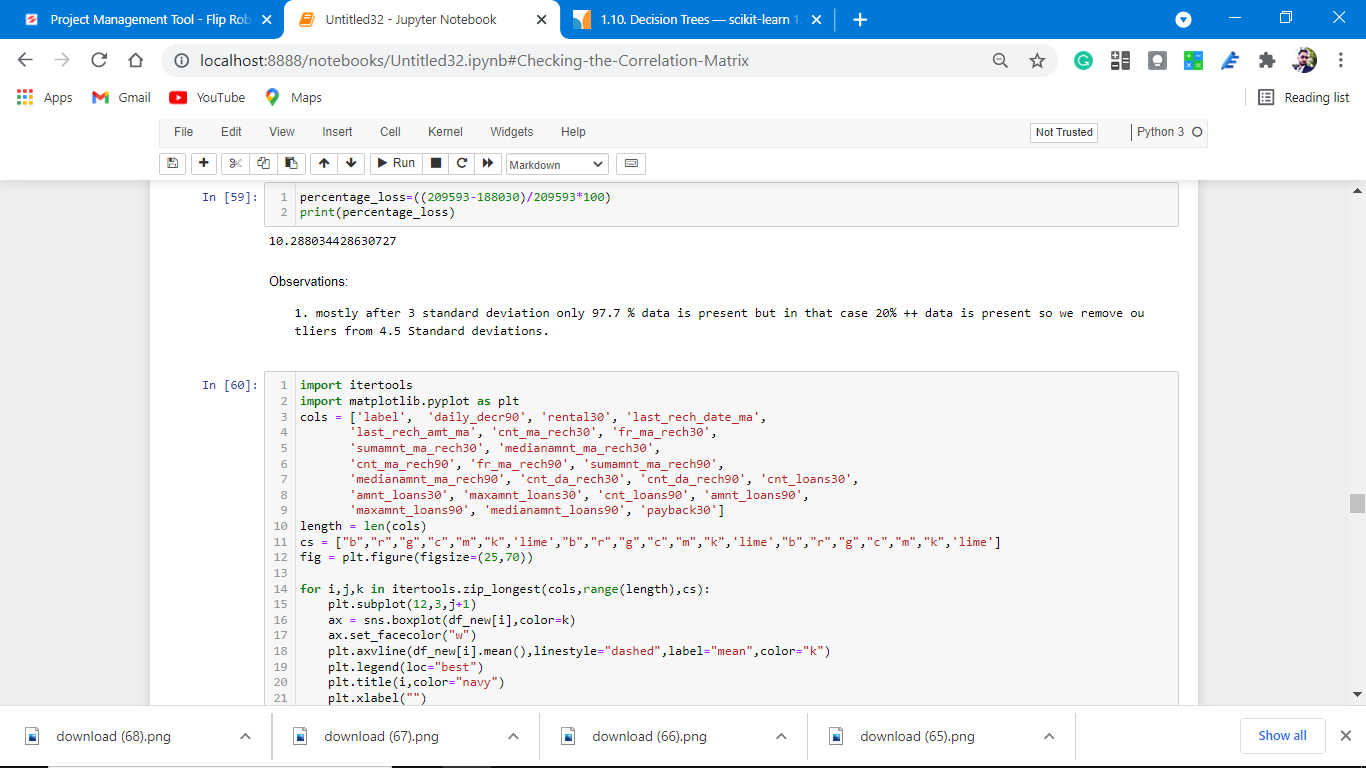
****

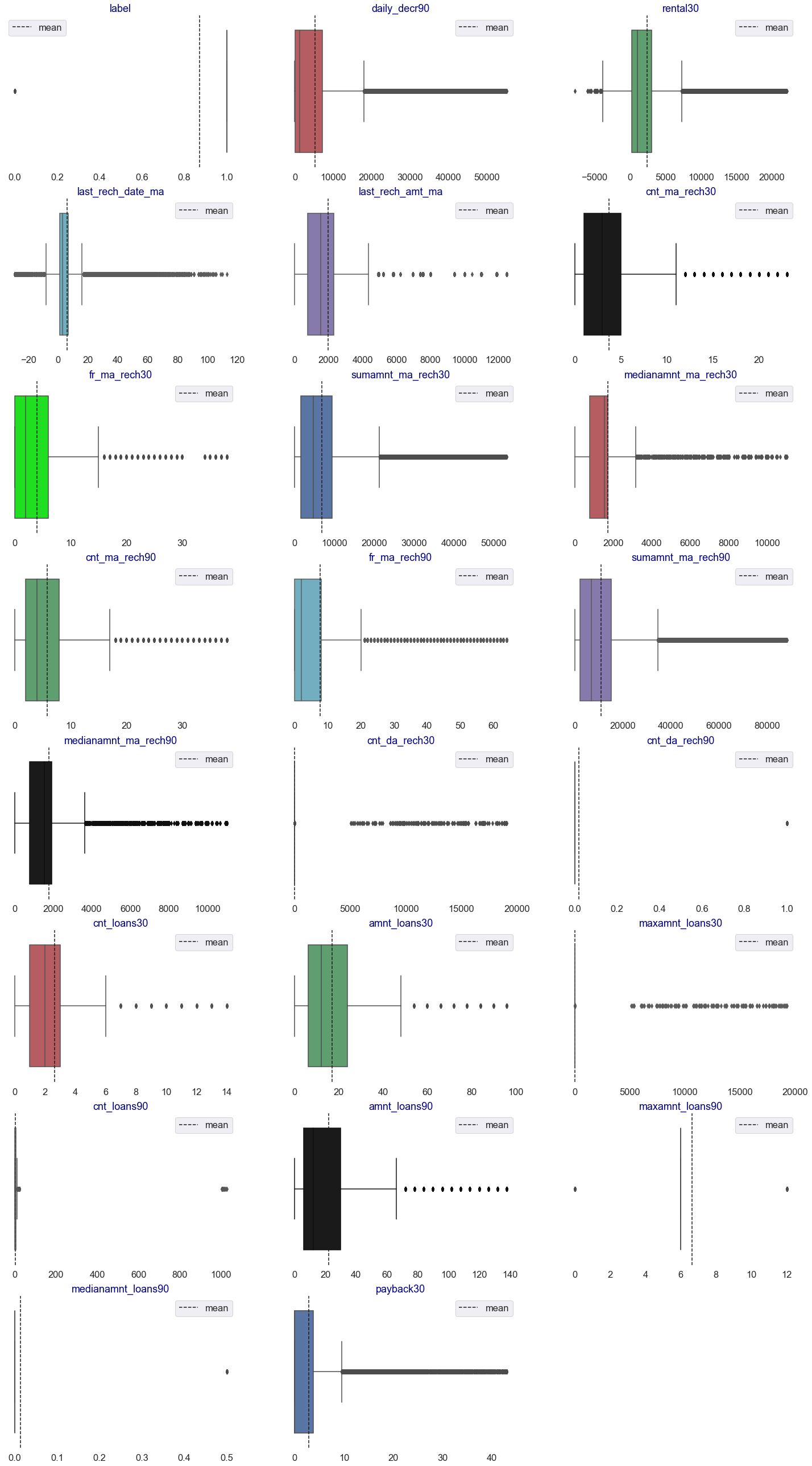
observations:

1. Some feature columns have lots of outliers which we have to remove.

2. payback30,amnt loans, etc having lots of outliers.

**Plotting after removing outliers.**

****

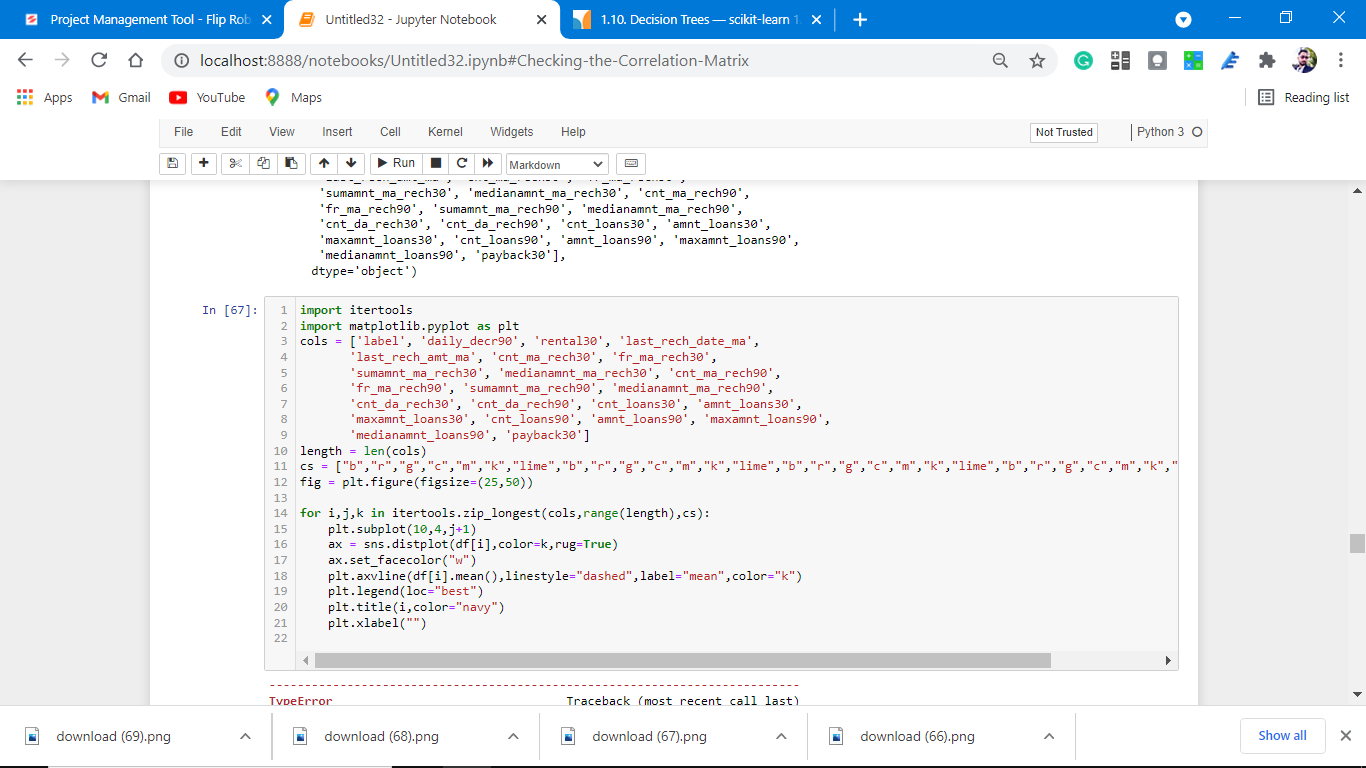


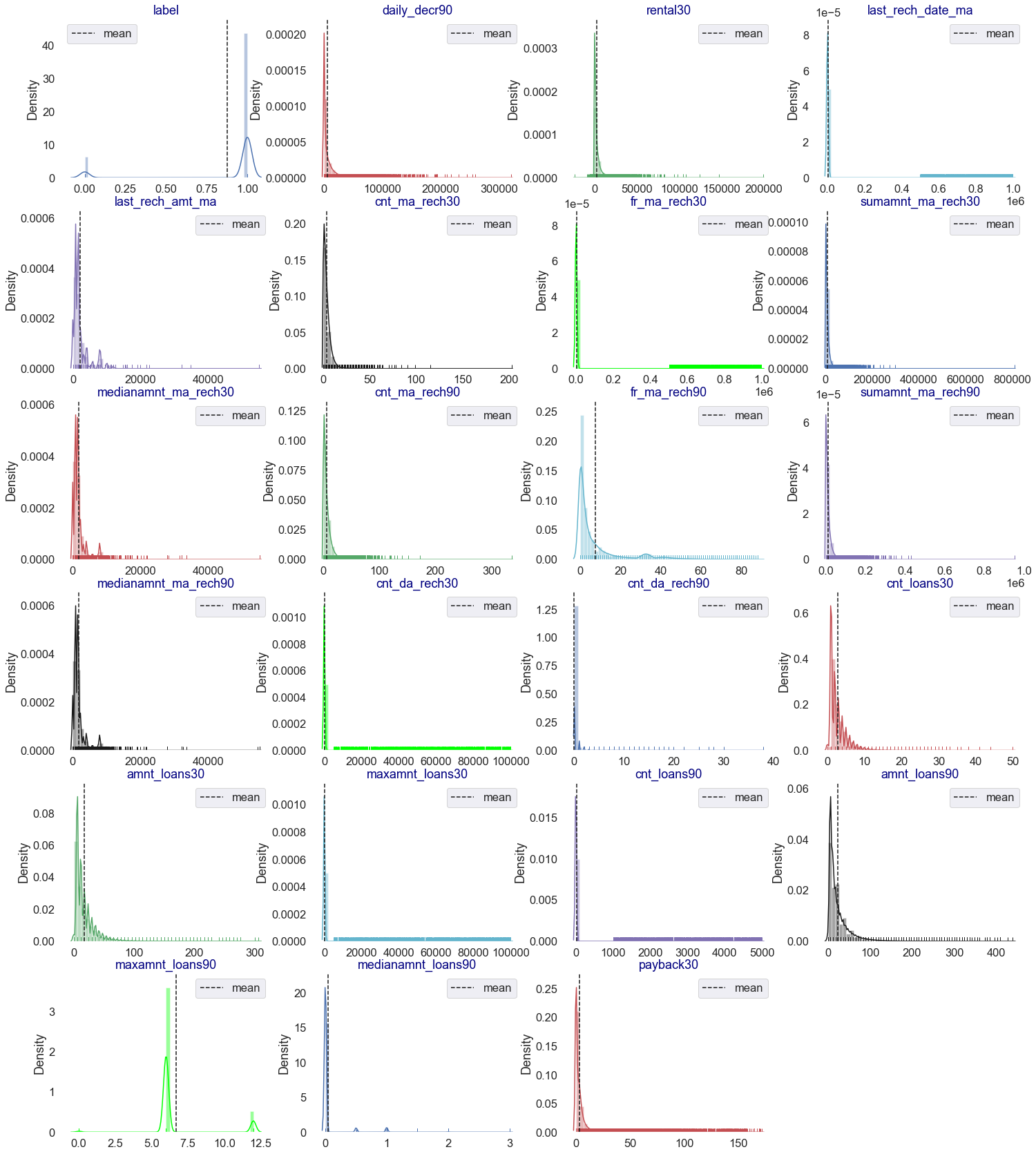
Observations:

1. after removing outliers our dataset looks good for the prediction of the label feature column.

2. Outliers are not completely removed but up to some extent, these are removed from the dataset.

**Plotting of distribution after power trasnsforming the feature columns.**





observations:

1. to remove the skewness from the dataset we removed the skewness by using yeo-johnson(A New Family of Power Transformations to Improve Normality of Symmetry).

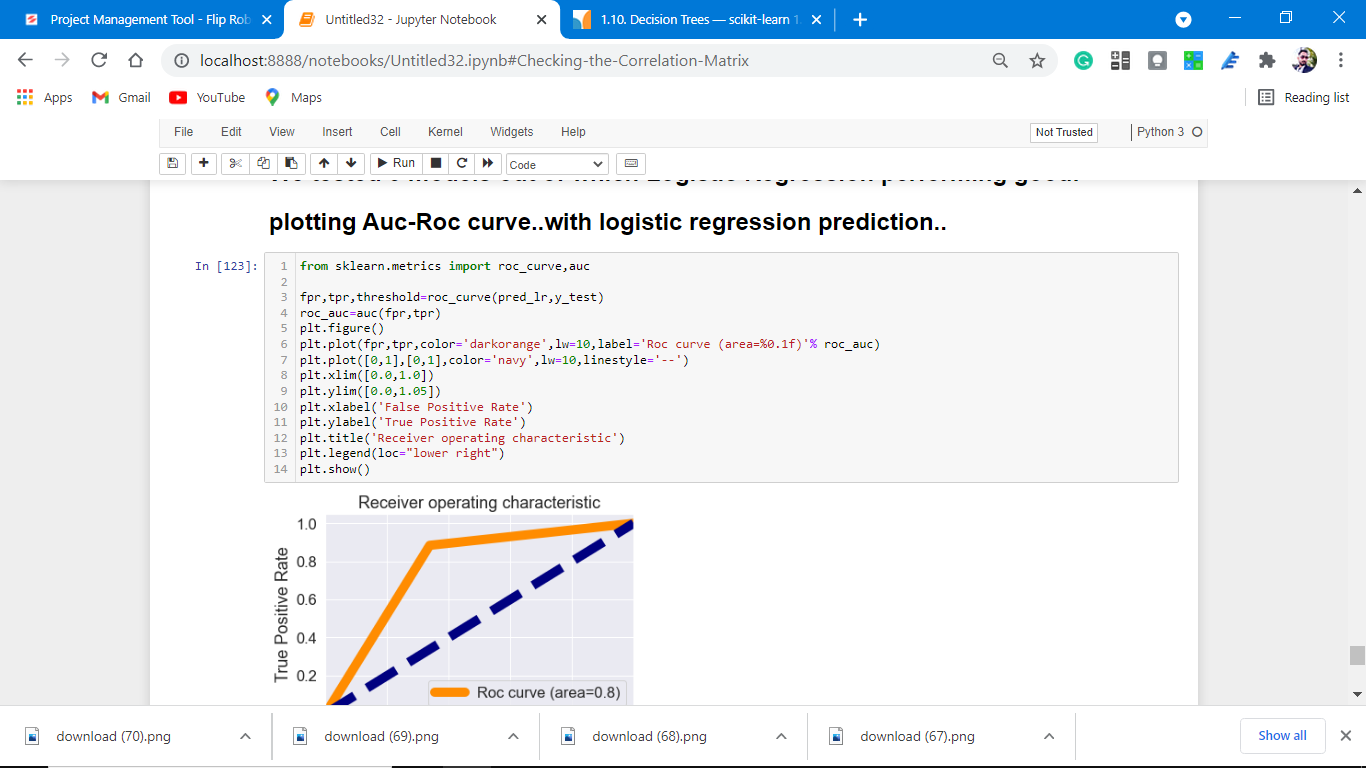
2. After applying power transform our dataset is normalized.

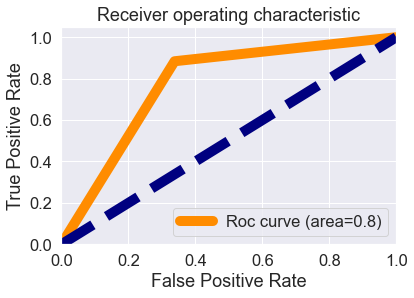
* Interpretation of the Results
* **The best model is Logistic Regression as the CV score is best from the above-tested models.**

At CV: - 7

Cross-validation score is :- 87.84502473009627

Accuracy Score:- 87.84502473009627





observations:

1. Roc curve area is 0.8 which means that our model is distinguished between defaulter label and not defaulter label is 80%.

2. our model understands that label 1 is 80% different than label 0, which is good.

3. It means there is an 80% chance that the model will be able to distinguish between positive class and negative class.

**CONCLUSION**

* Key Findings and Conclusions of the Study

So, our Aim is achieved as we have successfully ticked

all our parameters as mentioned in our Aim Column. It is seen cnt\_ma\_rech30 is the most effective attribute in predicting the label column and that the logistic regression is the most effective model for our Dataset with cv and accuracy is 88%.

* Learning Outcomes of the Study in respect of Data Science

That's it! We reached the end of our exercise.

Throughout this kernel, we put into practice many of the strategies for predicting frauds. We philosophized about the variables, we analyzed 'label' alone and with the most correlated variables, we dealt with missing data and outliers, we tested some of the fundamental statistical assumptions and we even transformed categorical variables into dummy variables. That's a lot of work that Python helped us make easier.

* Limitations of this work and Scope for Future Work

Limitations of this work are as follows:

1. This study works well predicting the frauds loan but is limited when comes to foreign countries,as loan is very small in Asian countries as compared to foreign.

For future work, we need data from foreign markets as well or mixed data globally, which predicts fraud or not whether loan is given or not.

Also, the accuracy of predicting is not 100% so many more models to test and which predict with 100% accuracy.

That’s all from this Project Report.

**Thank you**

Data exploration is the first step in data analysis and typically involves summarizing the main

characteristics of a data set, including its size, accuracy, initial patterns in the data, and other

attributes. It is commonly conducted by data analysts using visual analytics tools, but it can

also be done in more advanced statistical software, Python. Before it can analyze

data collected by multiple data sources and stored in data warehouses, an organization must

know how many cases are in a data set, what variables are included, how many missing

values there are, and what general hypotheses the data is likely to support. An initial

exploration of the data set can help answer these questions by familiarizing analysts with the

data with which they are working.

We divided the data 9:1 for Training and Testing purposes respectively.

Data exploration is the first step in data analysis and typically involves summarizing the main

characteristics of a data set, including its size, accuracy, initial patterns in the data and other

attributes. It is commonly conducted by data analysts using visual analytics tools, but it can

also be done in more advanced statistical software, Python. Before it can conduct analysis on

data collected by multiple data sources and stored in data warehouses, an organization must

know how many cases are in a data set, what variables are included, how many missing

values there are and what general hypotheses the data is likely to support. An initial

exploration of the data set can help answer these questions by familiarizing analysts with the

data with which they are working.

We divided the data 9:1 for Training and Testing purpose respectively.