****

**MSc Big Data Analytics**

CMM704 – Data Mining

**IIT** Student ID: **20200361**

**RGU** Student ID: **2016629**

Don Manula Ransika Udugahapattuwa

Table of Contents

[(a) 2](#_Toc67923023)

[I. Handle missing values and outliers 2](#_Toc67923024)

[II. Produce Q-Q Plots and Histograms of the features, and apply the transformations if required 10](#_Toc67923025)

[III. Feature Coding 16](#_Toc67923026)

[IV. Scaling data 18](#_Toc67923027)

[V. Feature discretization 18](#_Toc67923028)

[(b) 19](#_Toc67923029)

[I. SVD (Singular Value Decomposition) for feature reduction. 19](#_Toc67923030)

[II. Significant and independent features. 21](#_Toc67923031)

[(c). Applying Logistic Regression and Support Vector Machine techniques 22](#_Toc67923032)

[(d) applicability of SVM and LR 25](#_Toc67923033)

[(e) How significant your findings are 28](#_Toc67923034)

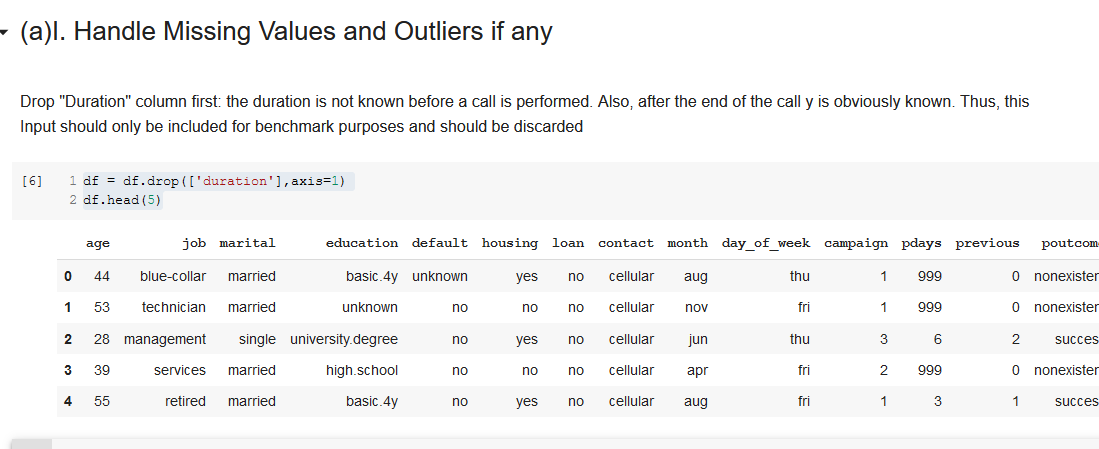
# (a)

We can start the project by setting up the prerequisites and introducing the data set.

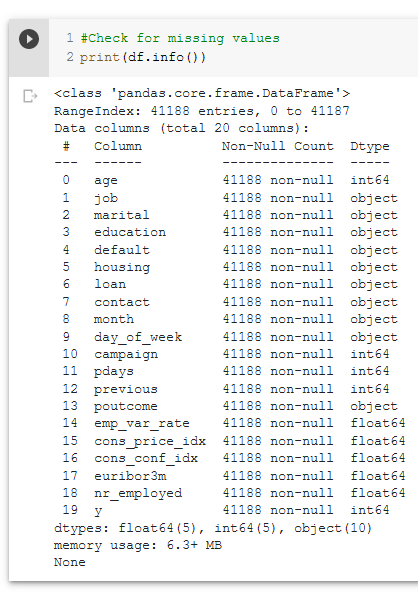


## Handle missing values and outliers

Drop "Duration" column first: the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this Input should only be included for benchmark purposes and should be discarded.

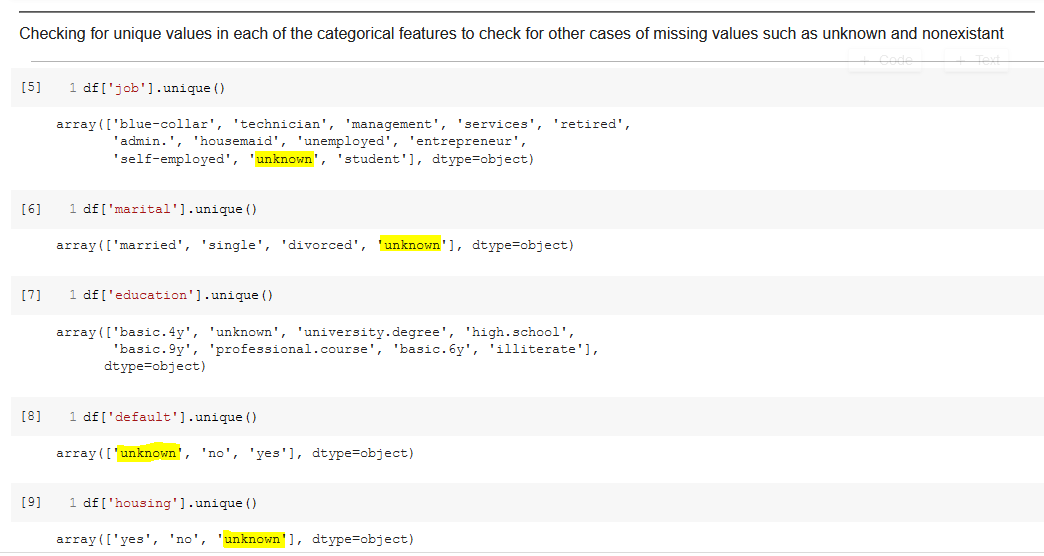


Then we can proceed with checking for missing values:



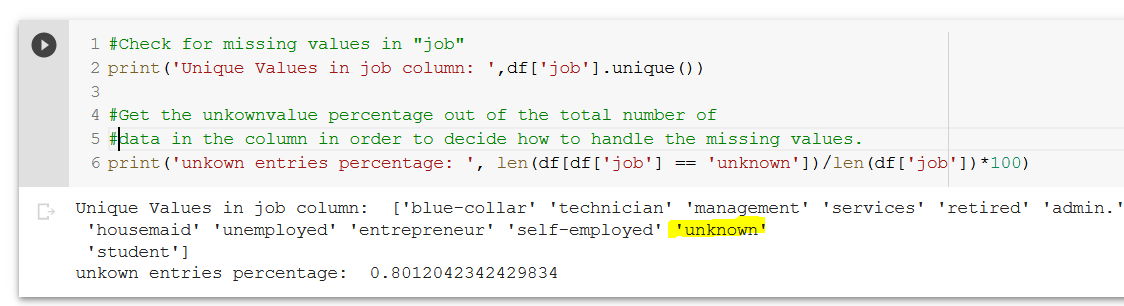
Even though there are no NULL values, there might be other custom values like "unknown" that suggest a missing values.

* To check this, we can get the unique values of the columns.

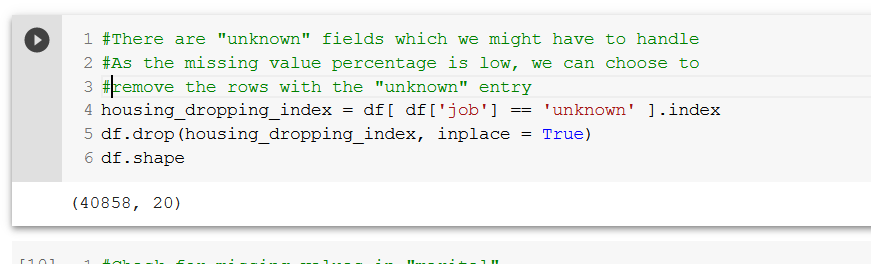




* Then we can check the missing value percentage in order to get an idea on handling the data.

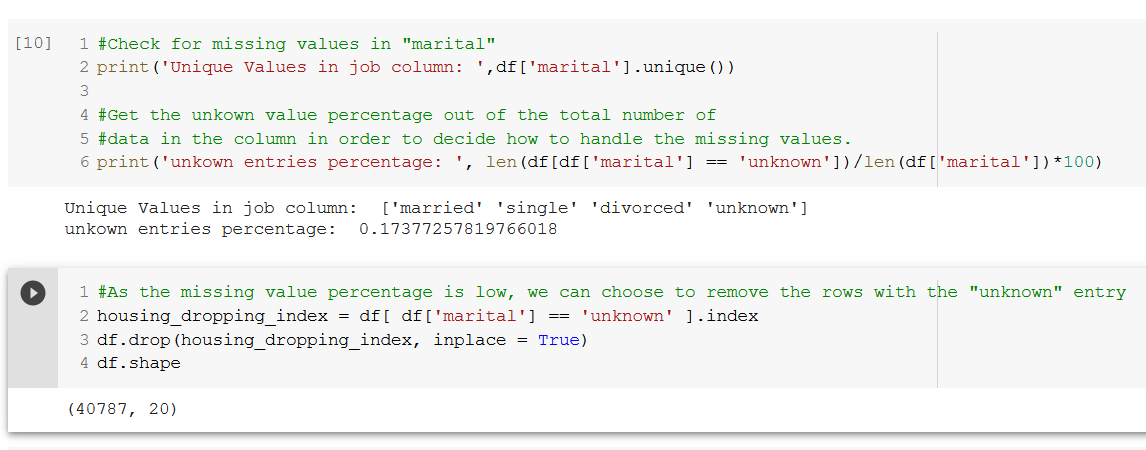


As the missing value percentage is low in the feature “job” (0.8%), we can choose to remove the rows with the "unknown" entry.

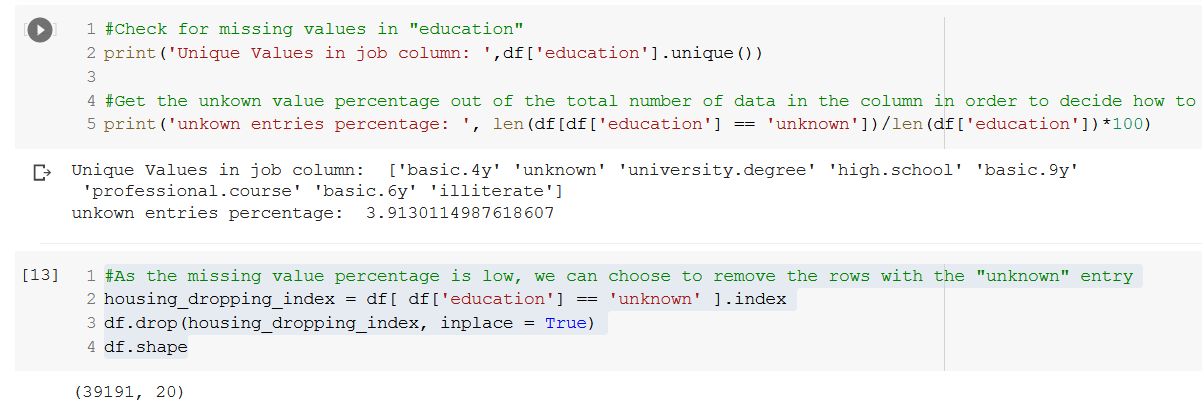


It can be done for all the columns which show a lower percentage of missing value count.

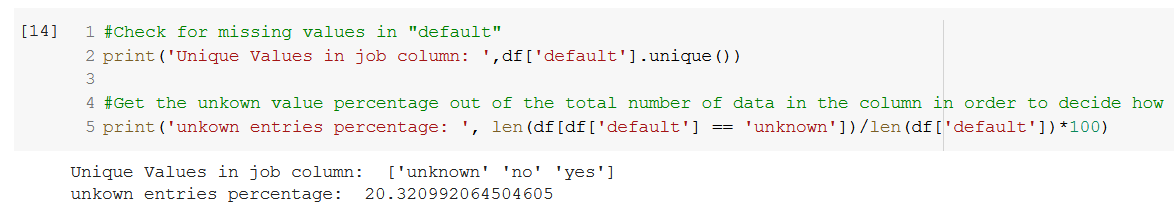
Marital:0.1%

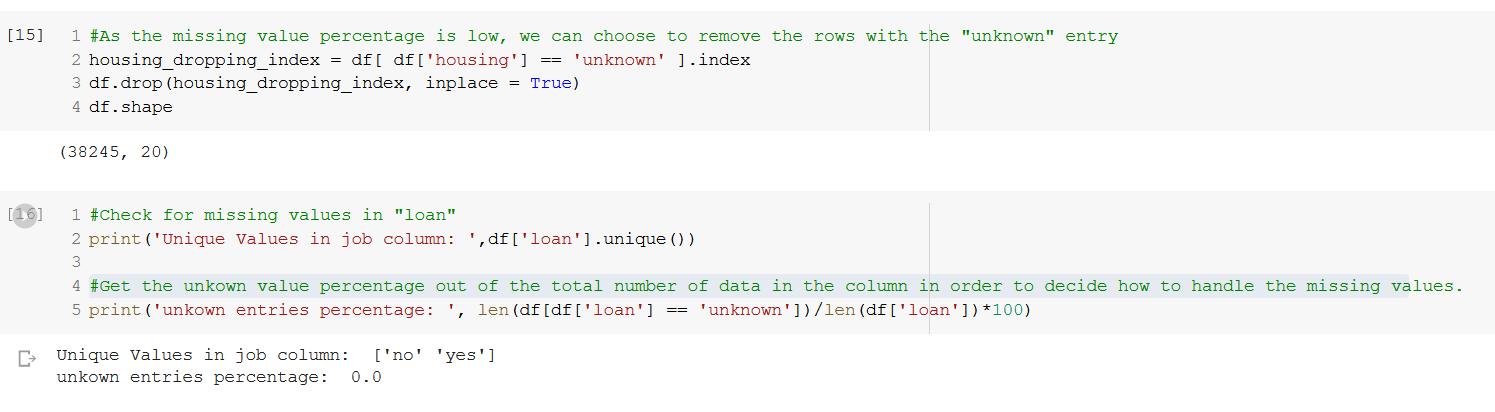


Education: 3.9%



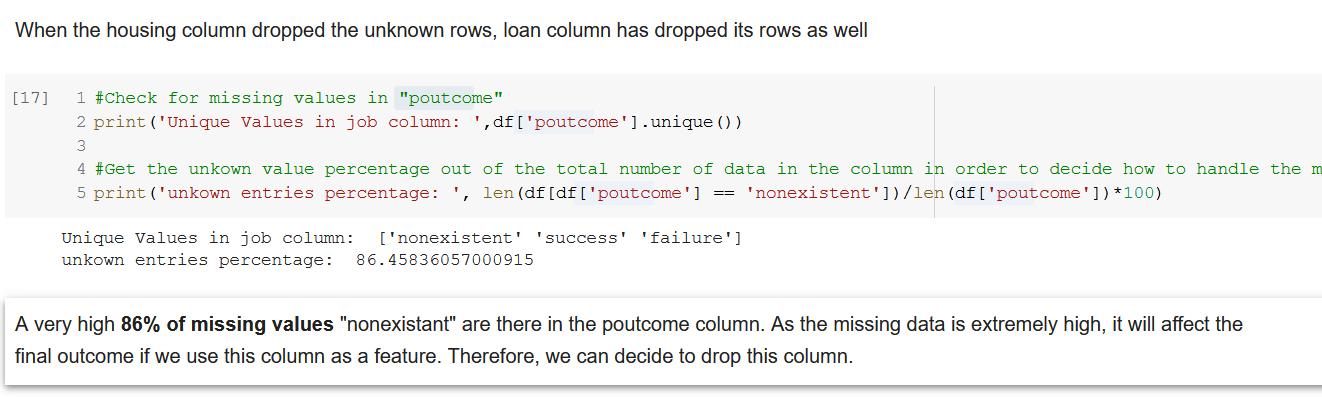
Loan: 20%

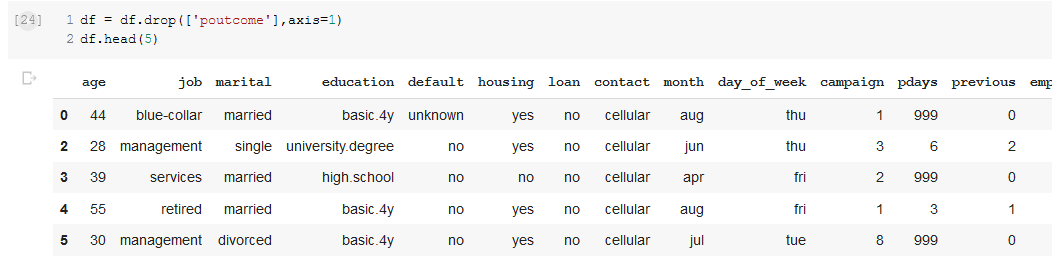
20% is a considerable "unknown" percentage. Still, the percentage is not high enough to fully drop the column. How ever, removing the rows would result in a considerable effect on the final predictions. Thus, decided to keep the "default" column as it is.



Poutcome: 86%

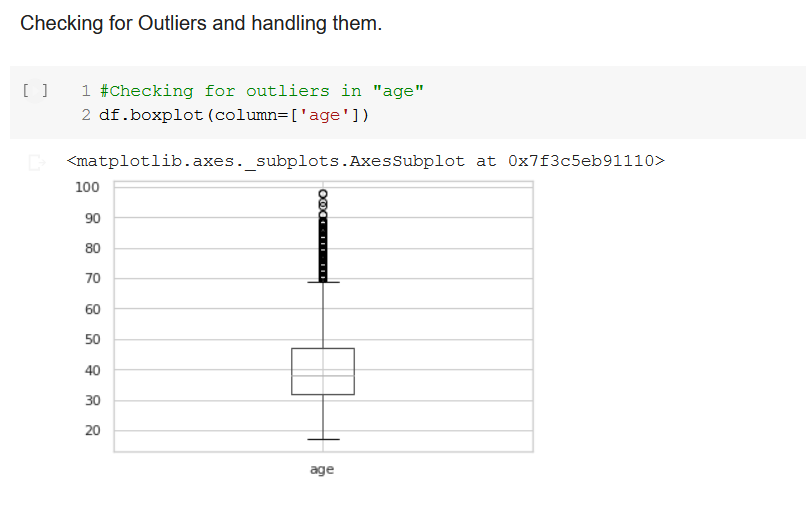
A very high \*86% of missing values \* "nonexistant" are there in the poutcome column. As the missing data is extremely high, it will affect the final outcome if we use this column as a feature. Therefore, we can decide to **drop** this column.





**Checking and handling Outliers**

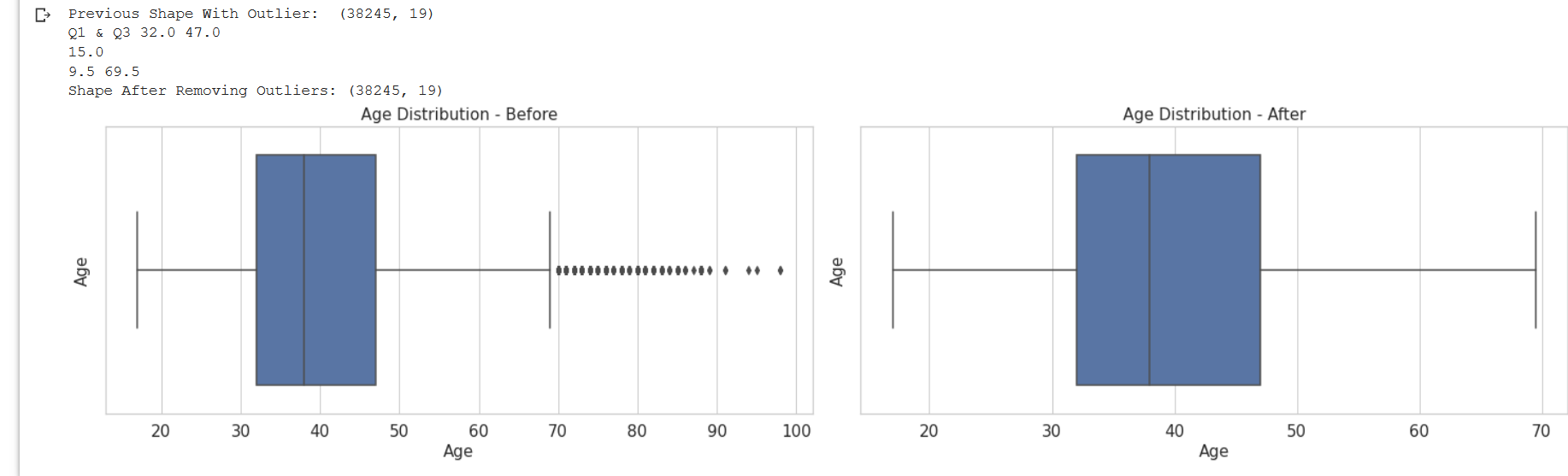
Boxplots can identify the outliers efficiently.

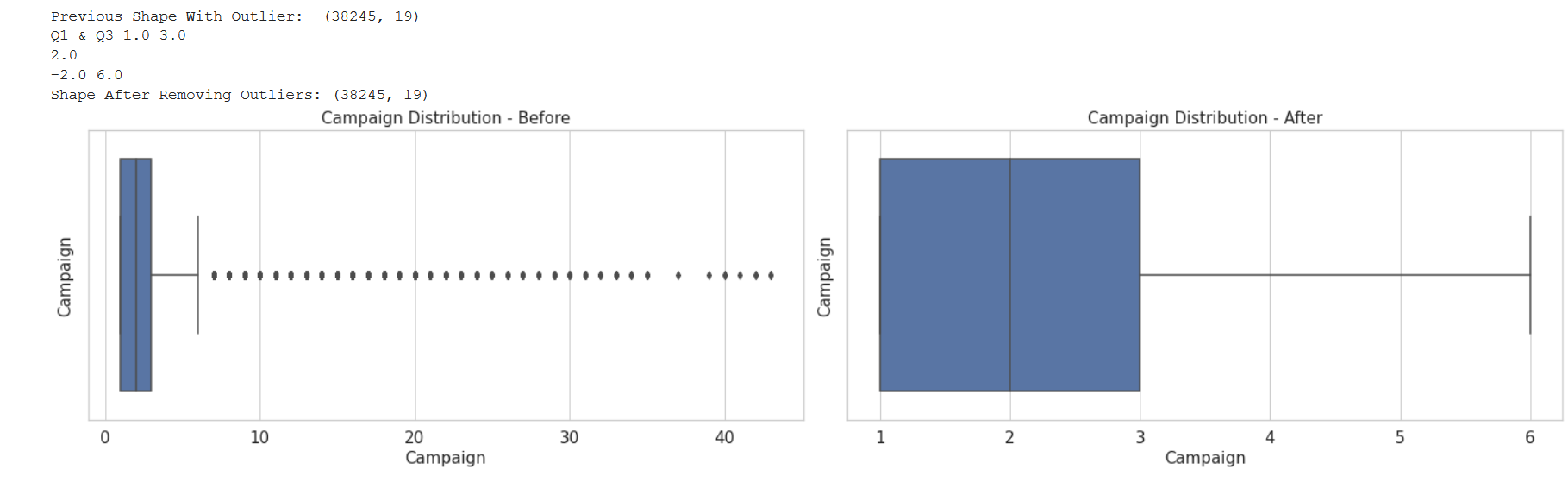


Removing the outliers in age would affect the overall result as the outlier count is almost half the total dataset. Thus, we can decide to replace the missing values using the Inter Quartile Range.



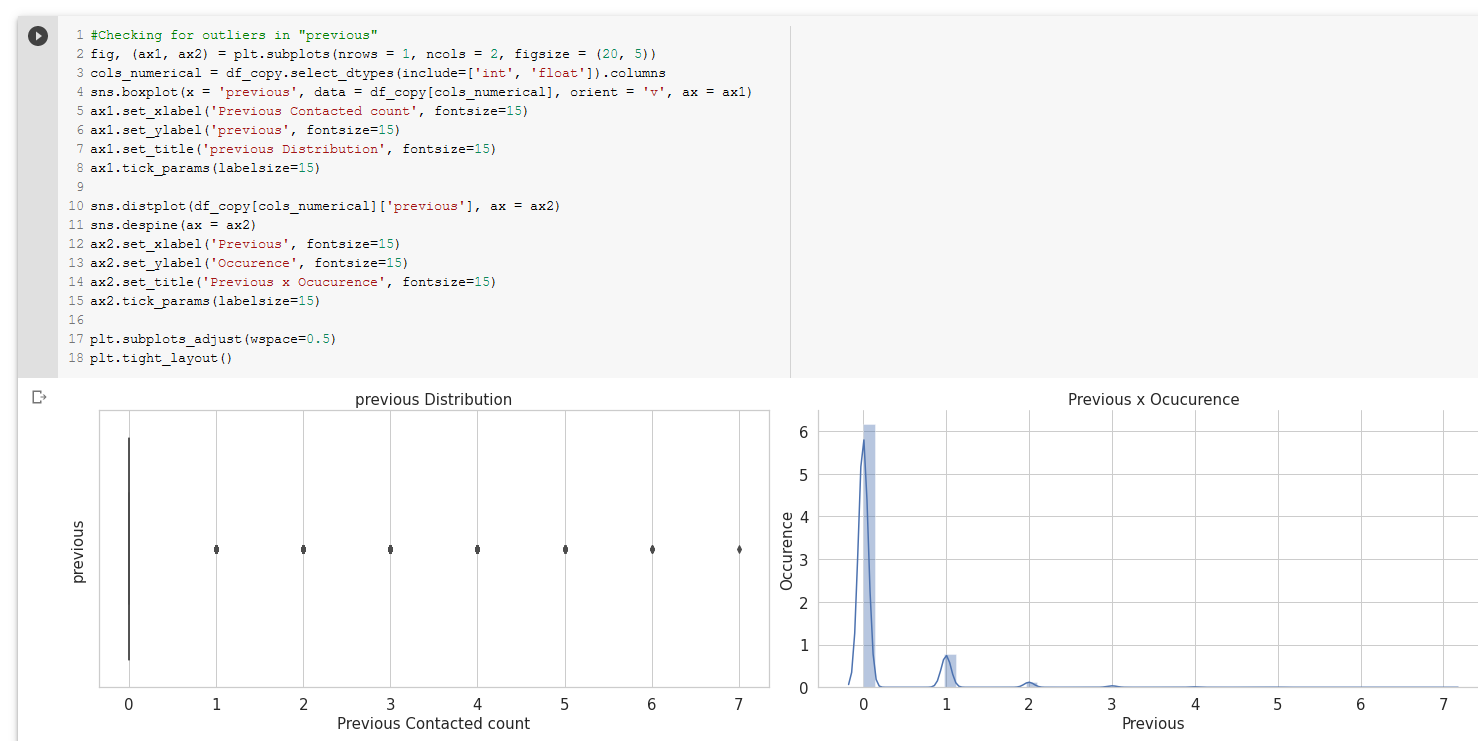
This is how the Before and after boxplots look like.



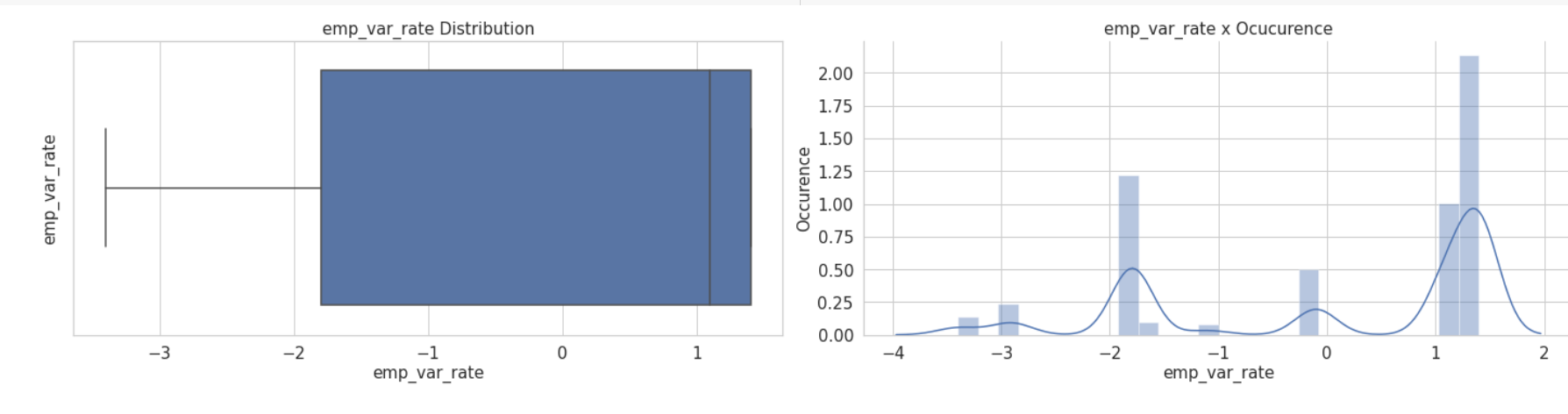


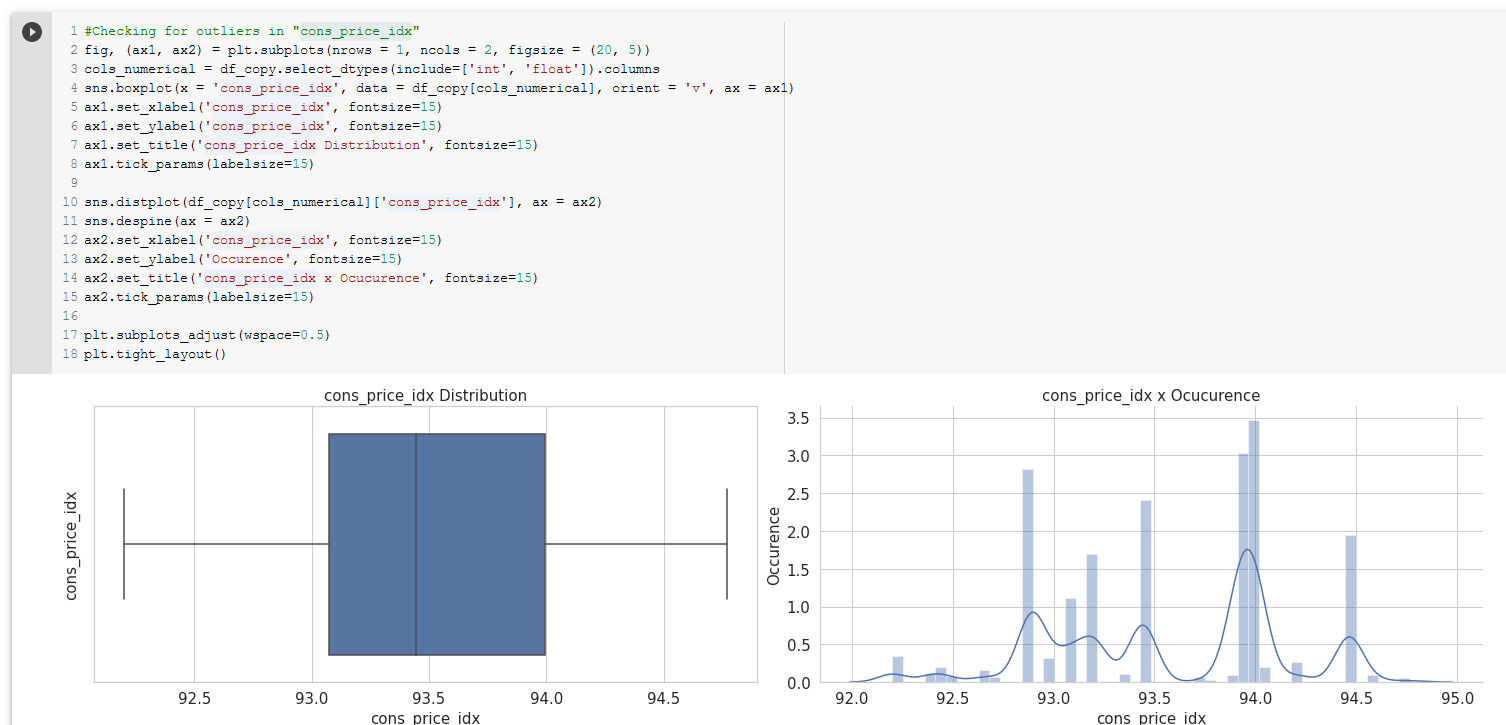
Number of previous contacts are distributed between 0 times and 7 times. Number of previous contacts are distributed between 0 times and 7 times.

This would not be manipulated as that might make it loose almost all the data in previous column.

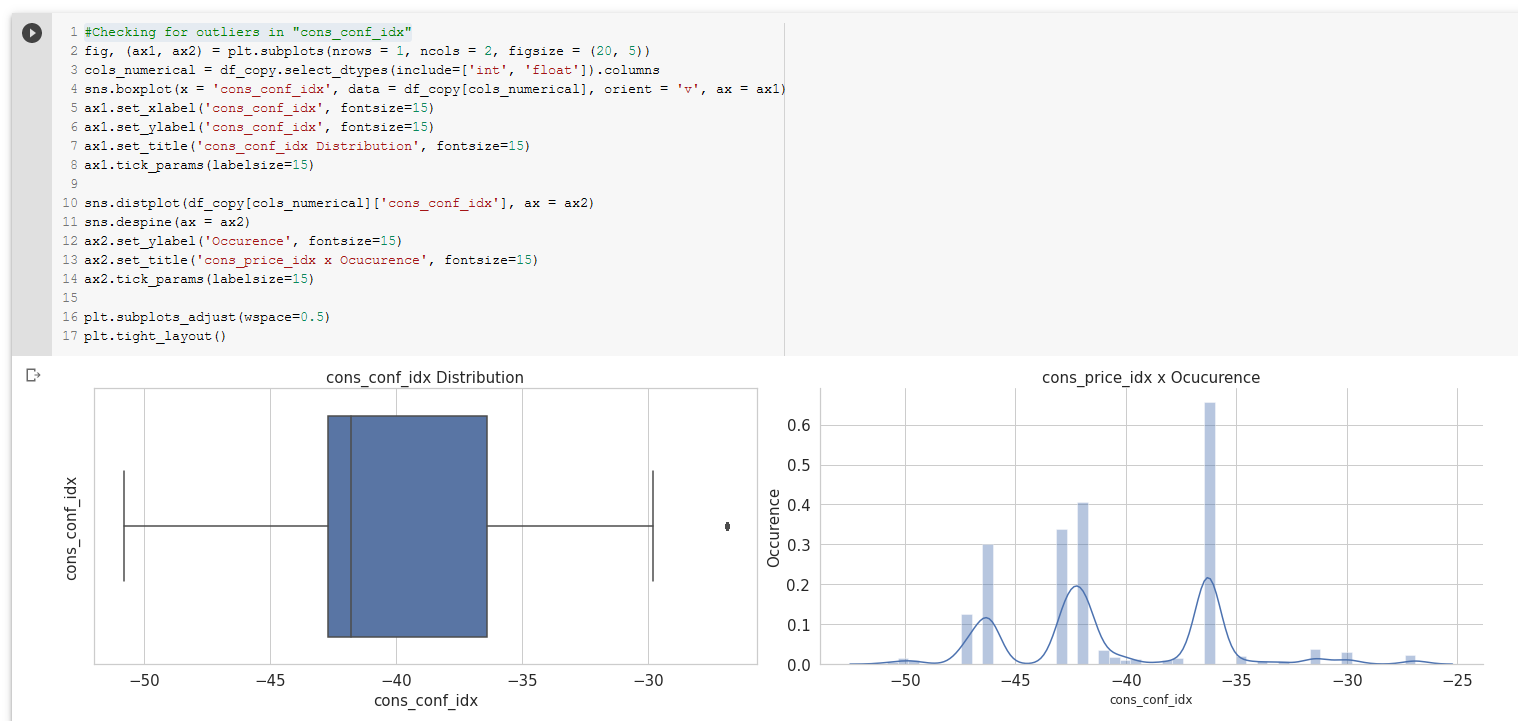


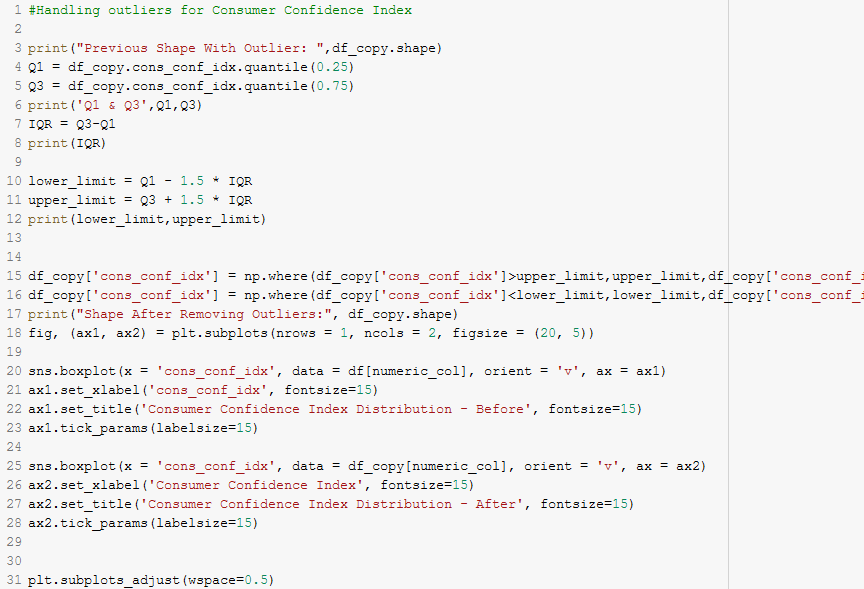
emp\_var\_rate and cons\_price\_idx has no visible outliers to be removed:

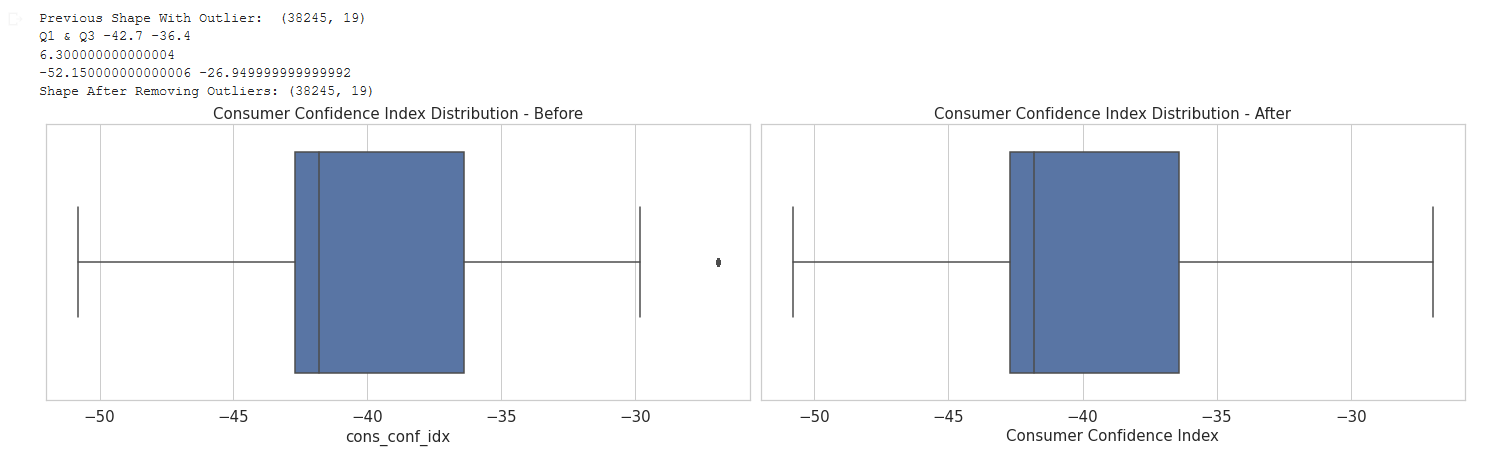




Checking for outliers in "cons\_conf\_idx" ehich will then be replaced using the IQR

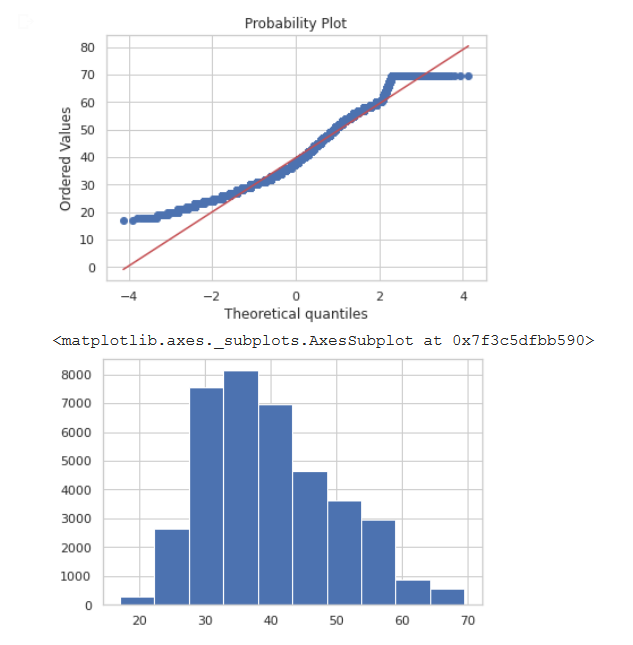




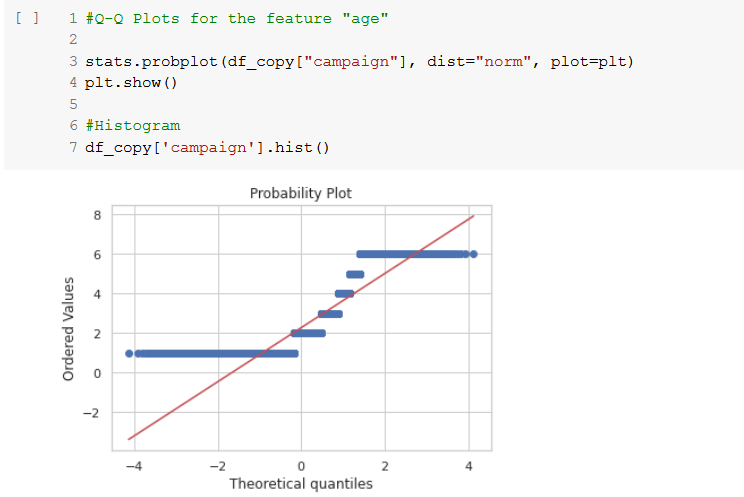


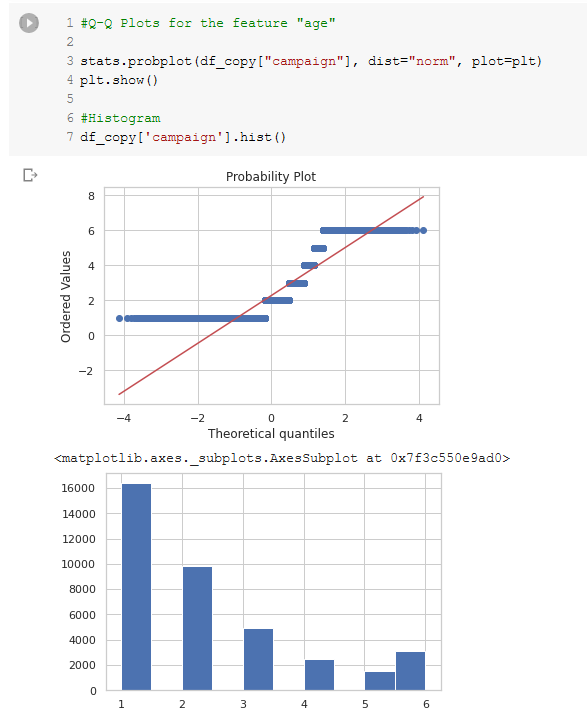
## Produce Q-Q Plots and Histograms of the features, and apply the transformations if required



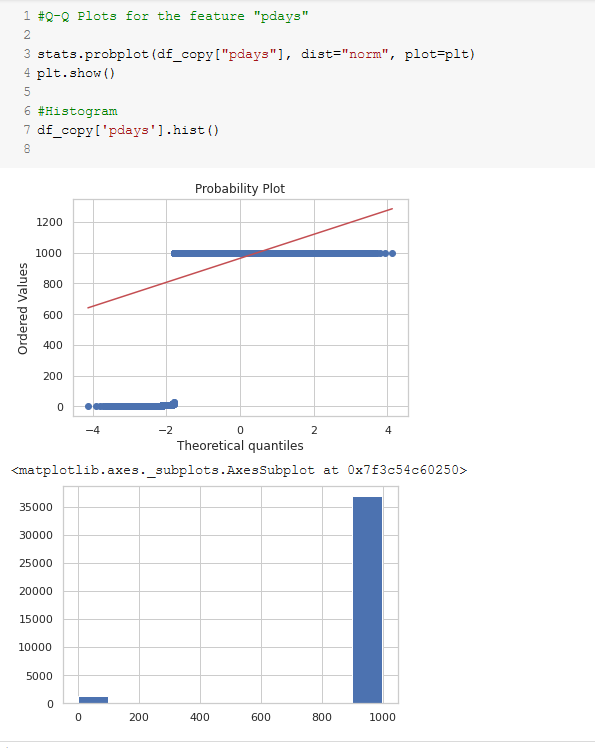


The feature "age" is right skewed when we take a look at the histogram. Thus, we can apply the technique of Transforming **Right Skewed Data using logarithm**

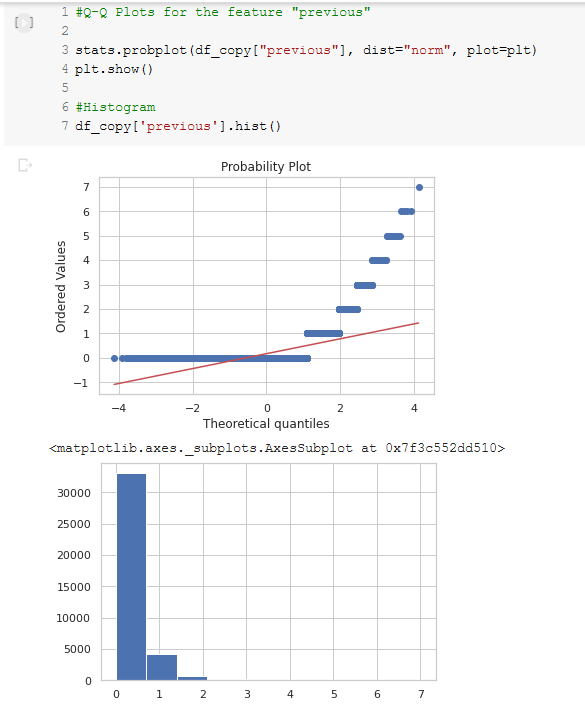


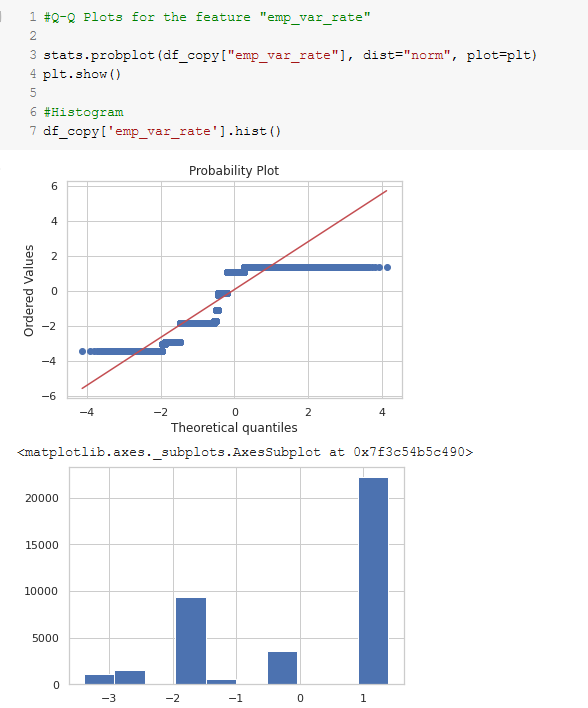


The feature "campaign" is also right skewed when we take a look at the histogram. Thus, we can apply the technique of **Transforming** Right Skewed Data using logarithm



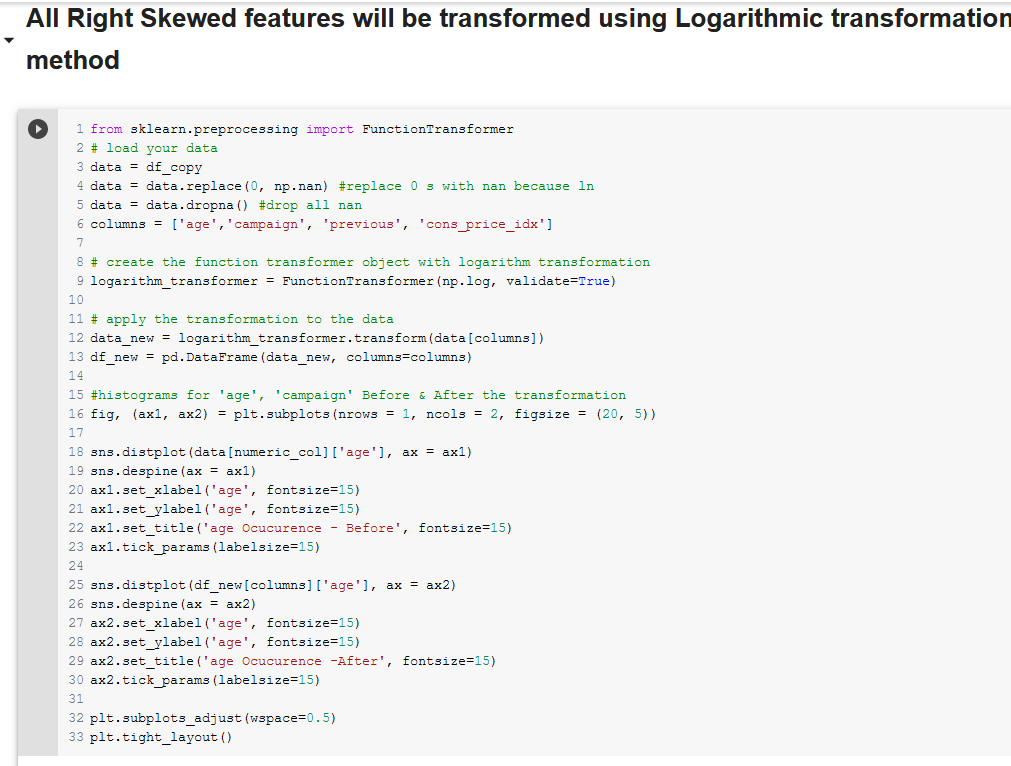
\*Pdays\* is heavily left skewed. Exponential transformation can be used to transform for left skewed data.

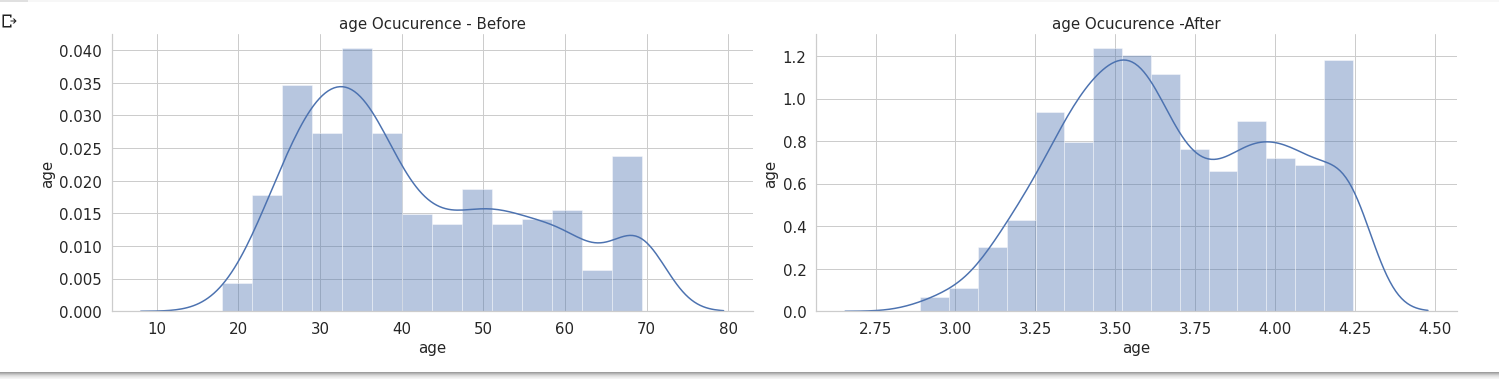


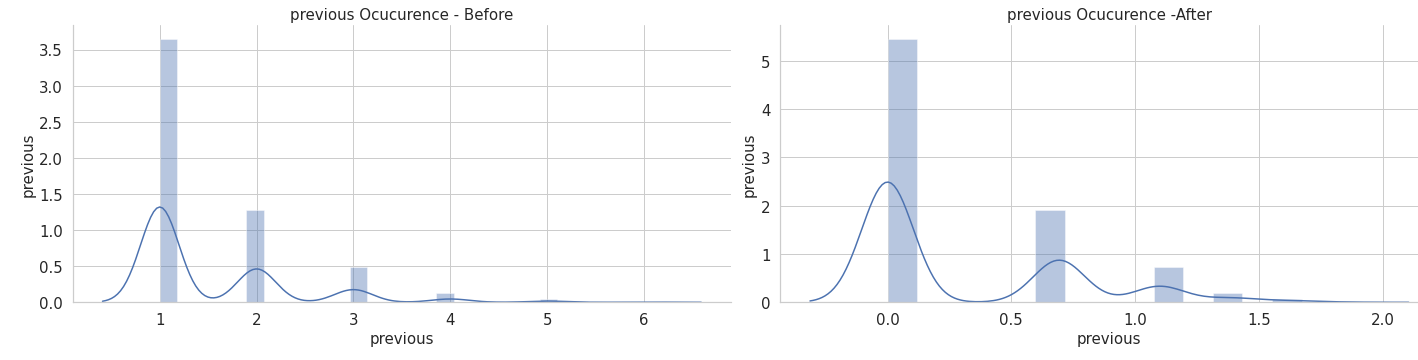


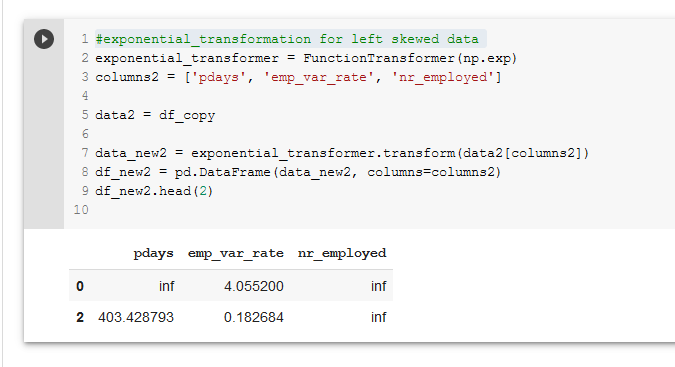
emp\_var\_rate is left skewed. Exponential transformation can be used to transform for left skewed data.

* emp\_var\_rate is left skewed. Exponential transformation can be used to transform for left skewed data.
* exponential\_transformation can be applied for left skewed data

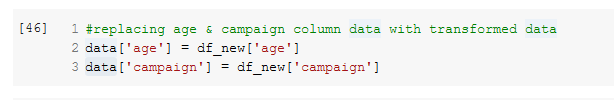


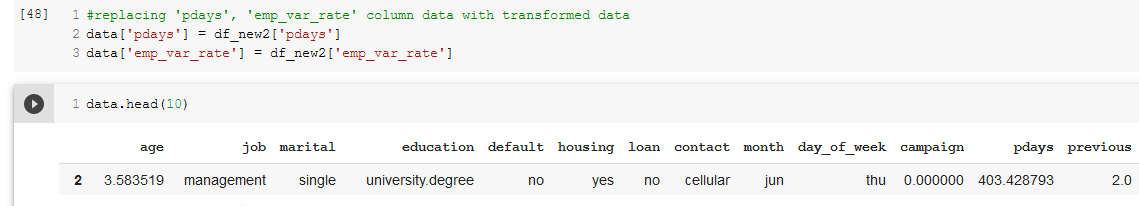






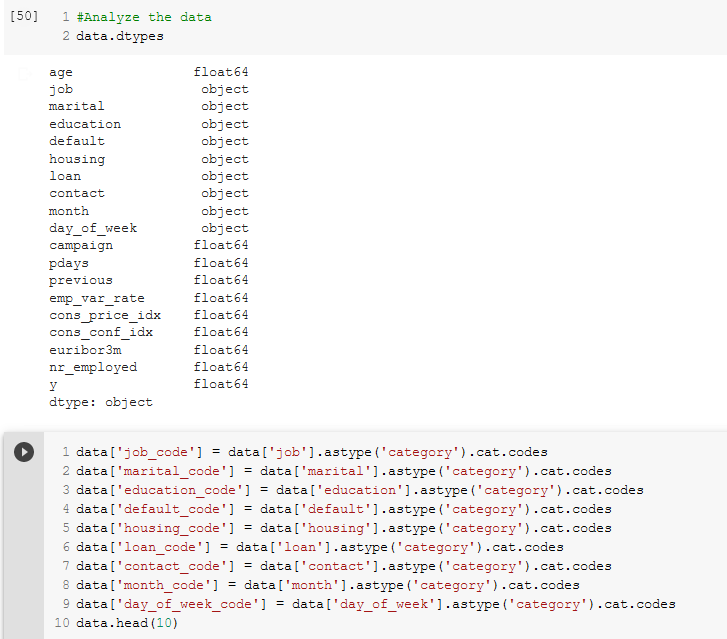
Now we can combine the transformed features with the complete dataset we have:



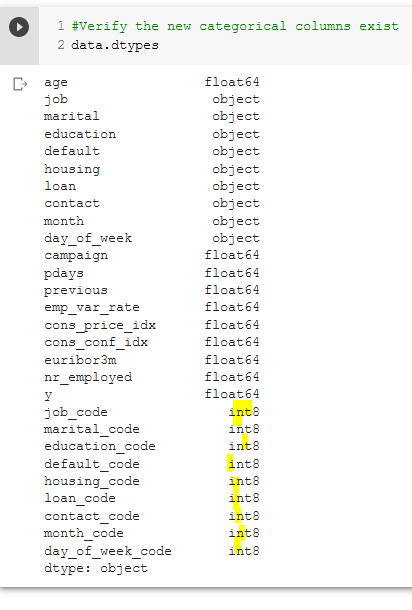


## Feature Coding

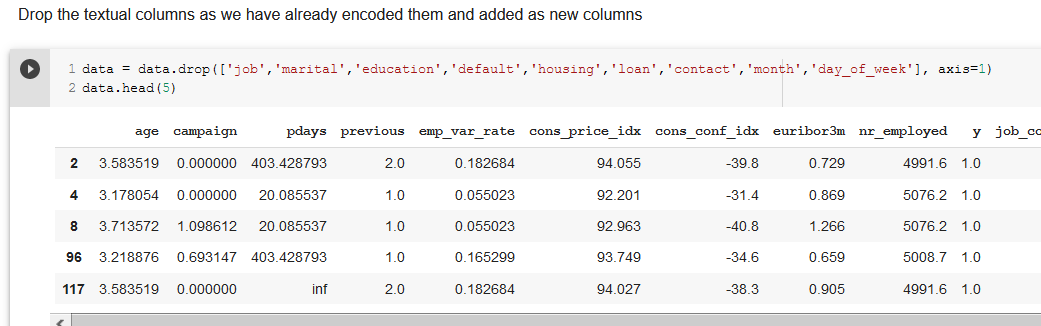
Integer (Label) Encoding is used as it will not add new columns and does not expand feature space



After data encoding:



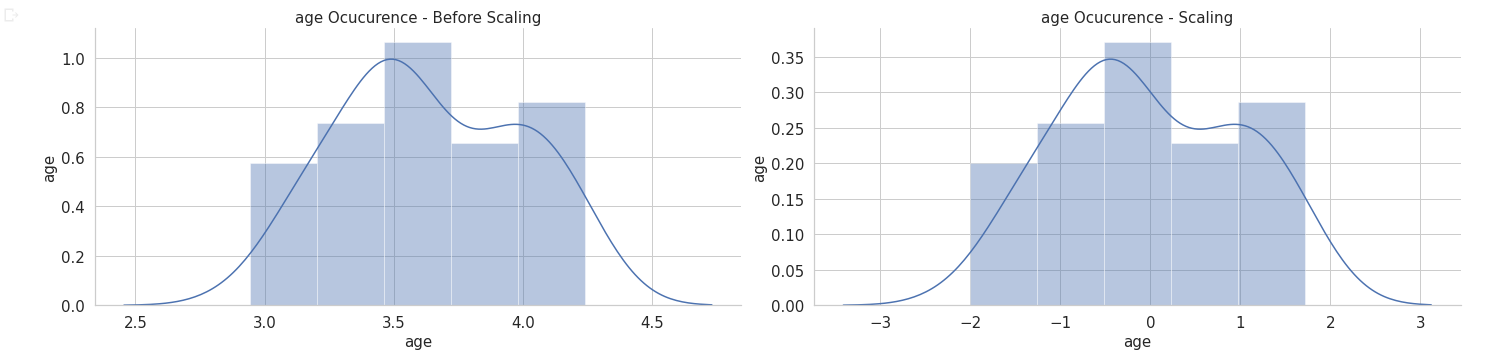
Now we can drop all the unencoded textual columns from the dataframe:

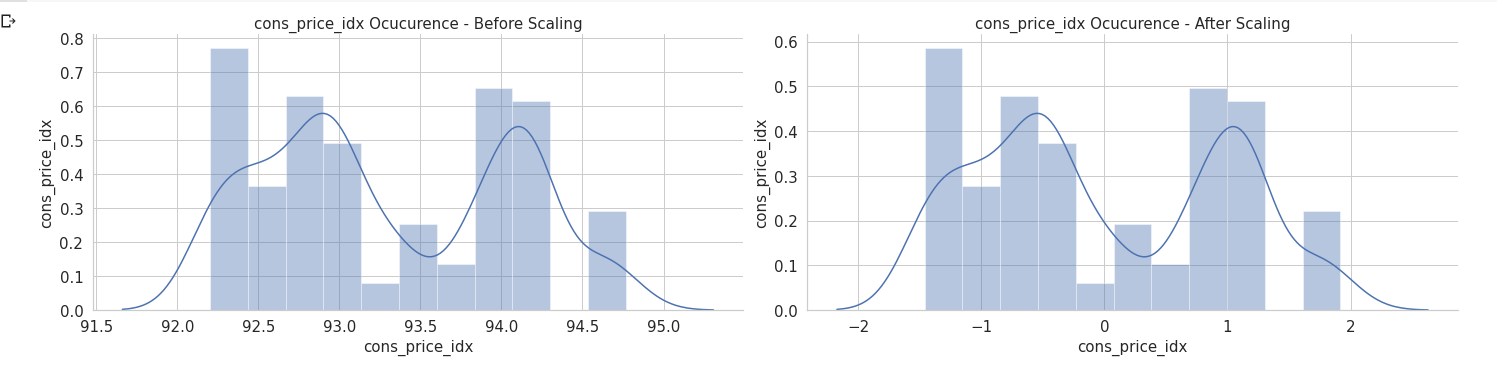


## Scaling data

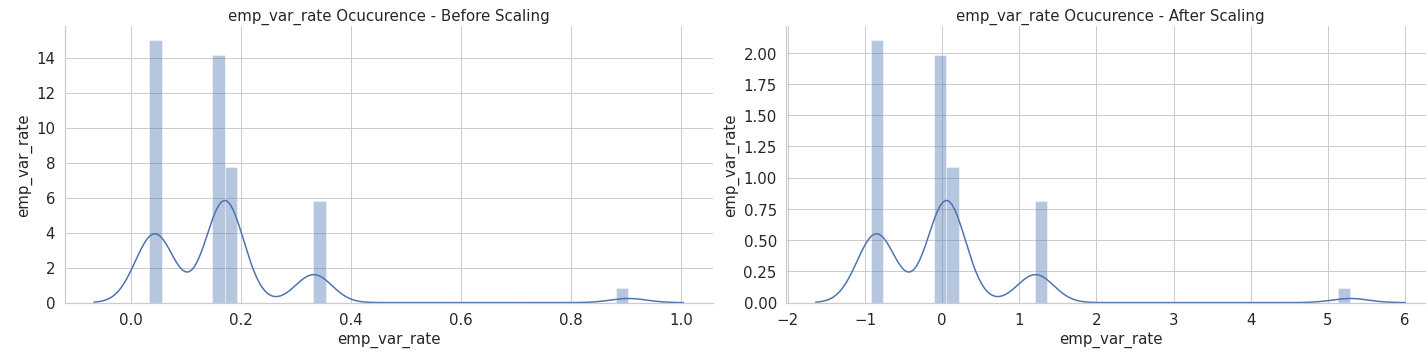


This is how the scaled age column looks like:



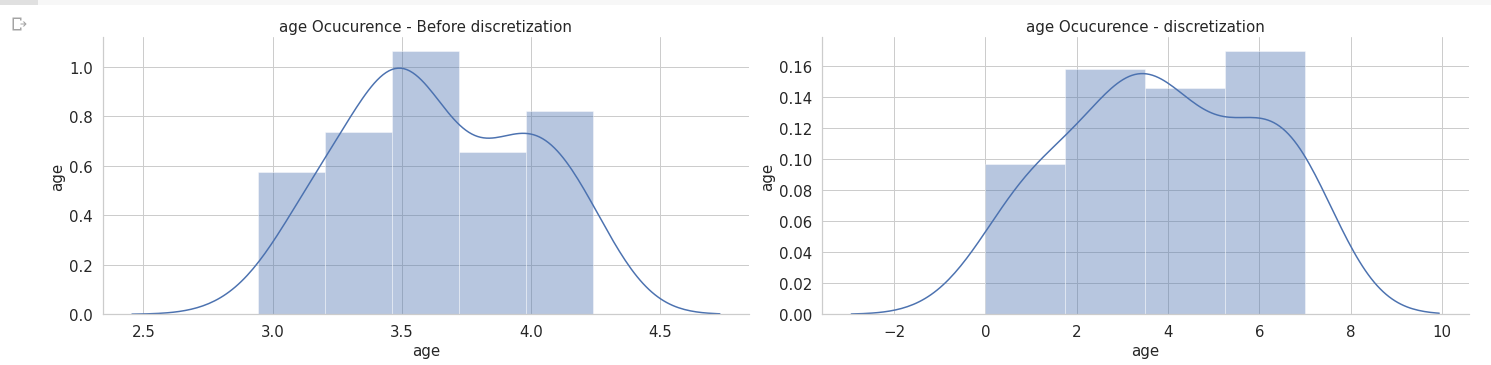


## Feature discretization



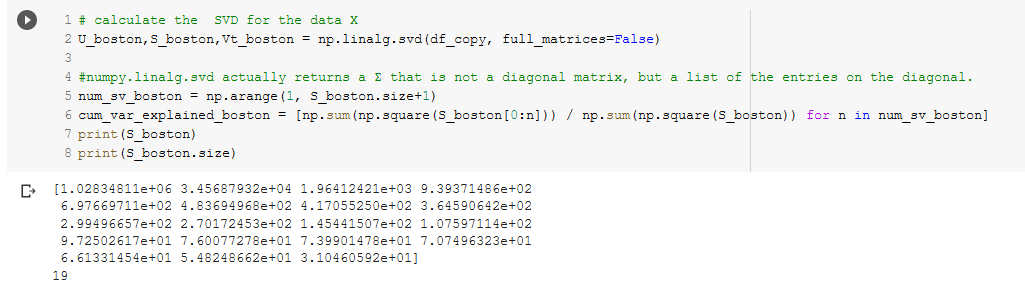
KBinsDiscretizer was used on age to discretize the feature:





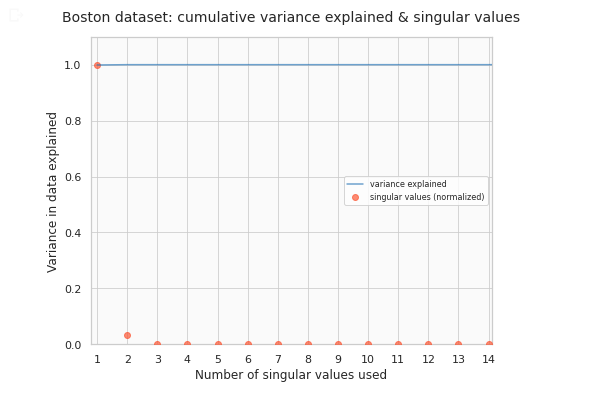
# (b)

## SVD (Singular Value Decomposition) for feature reduction.

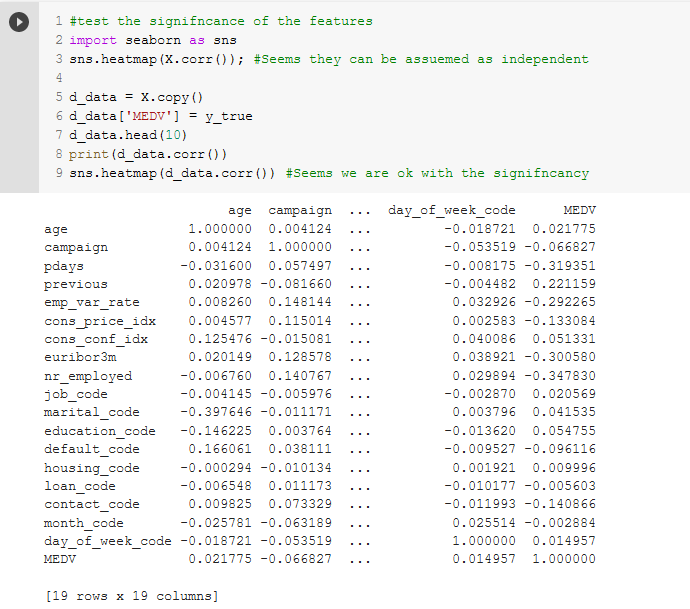


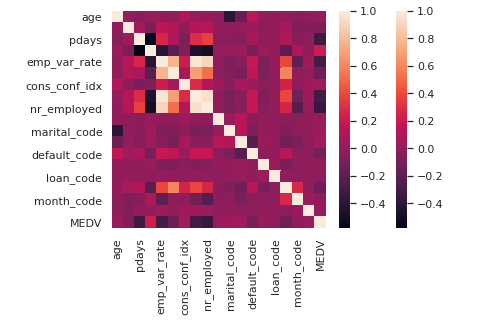
SVD gives a weighted value for each feature so that we can choose the most suitable ones.





## Significant and independent features.

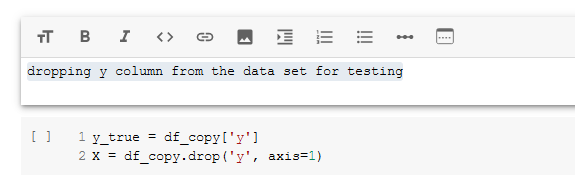




Darker the color the higher the correlation between the two variables.

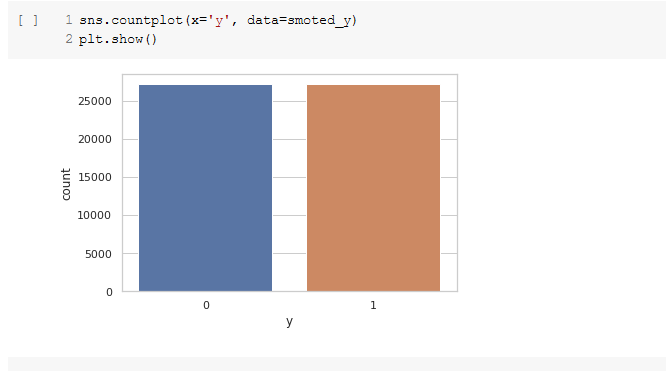
# (c). Applying Logistic Regression and Support Vector Machine techniques

First drop the y column from the data set for testing:

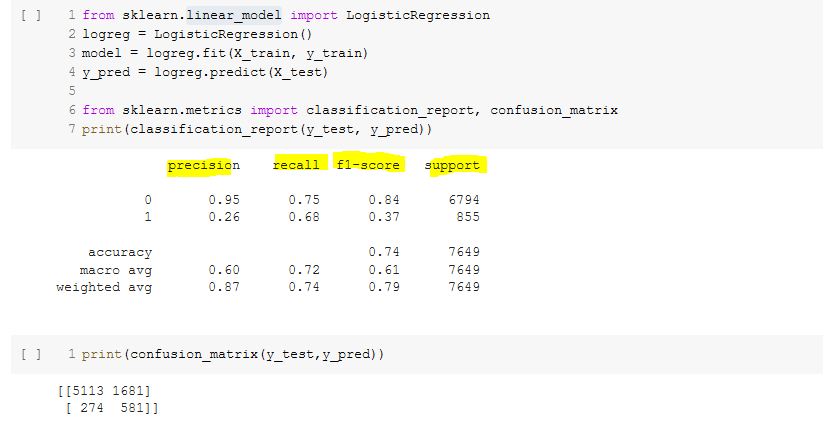


Split the dataset to **80% and 20%**

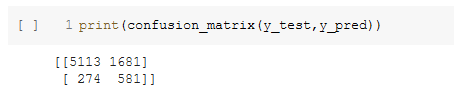




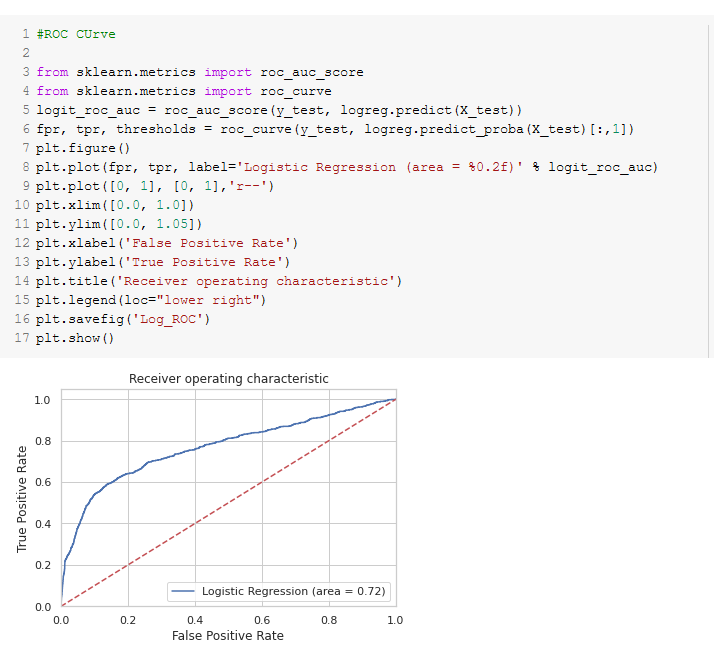
Training the dataset **with Logistic Regression**



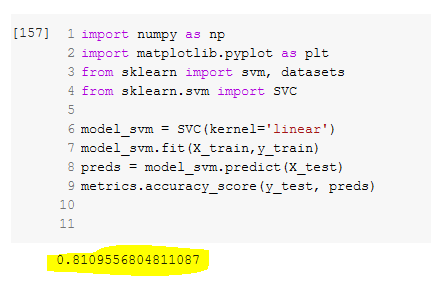
Confusion matrix is as follows:

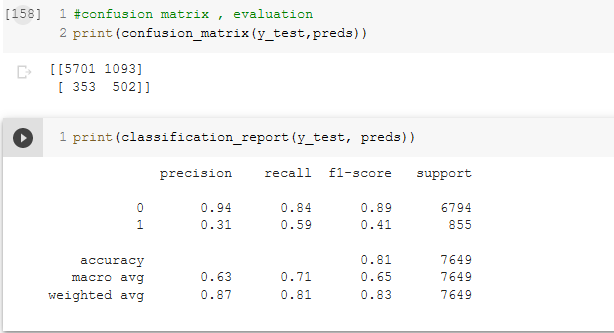


ROC Curve is used to identify the diagnostic capability of a model.



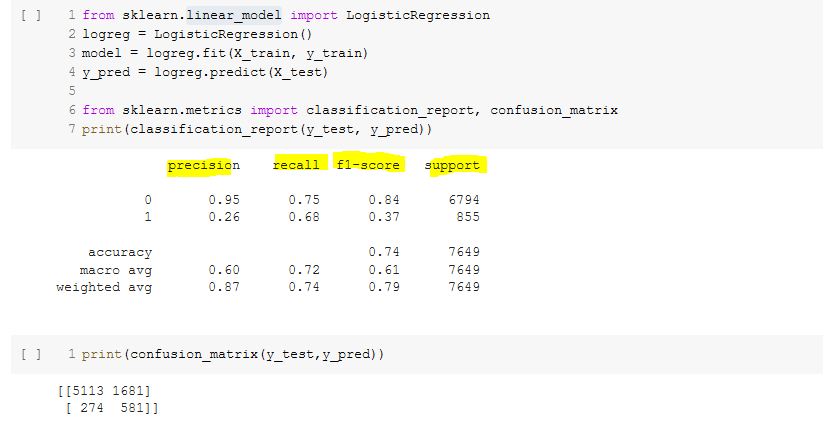
Training the data **with SVM**

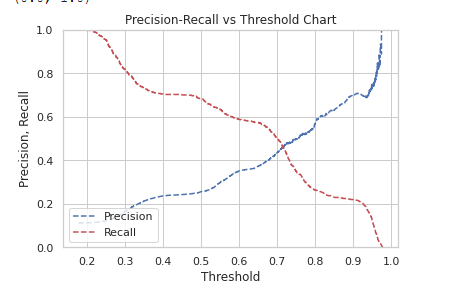
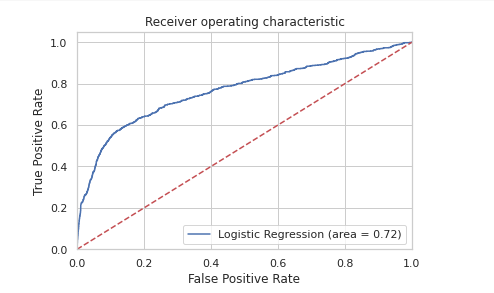




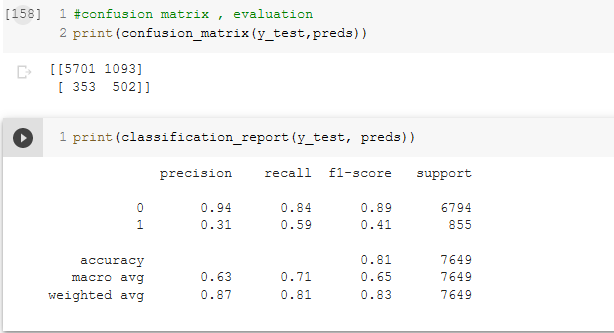
# (d) applicability of SVM and LR

**Logistic Regression:**

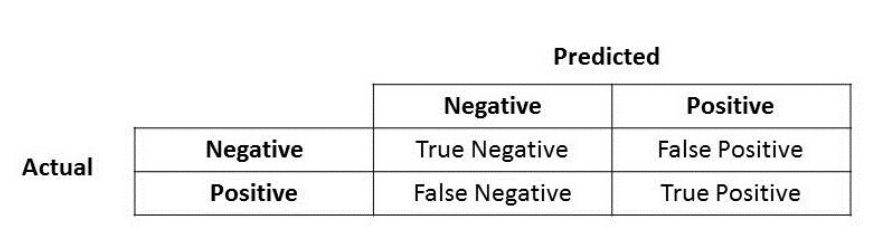




**Support Vector Machine**



In Logistic Regression, the confusion matrix shows that negative-negative (true negative) count is 5113 while positive-positive count is 502. In SVM, negative-negative is 5701 and positive-positive (true-positive) count is 502 which is equal.

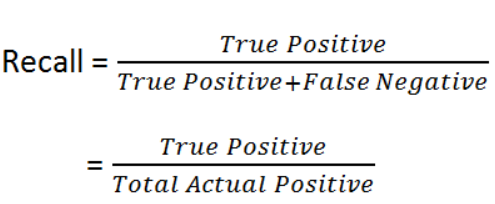


According to the confusion matrix, SVM seems to be a better fit for our data set.

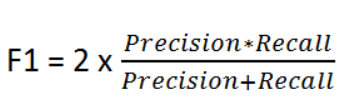
When comparing the 2 classification reports, LR shows an accuracy of 0.75 while SVM shows 0.84 which is a better score than LR.

Recall can be stated as the best metric to evaluate algorithms. This is because it takes the value of true positives over the total actual positives which gives a clear picture on the correctly predicted positive values.

In LR , recall value is 0.75 while in SVM it’s 0.84, which suggests that SVM works better with the banking dataset.



Even in F1-score which is a good measure to check the balance between the two values, recall and precision, SVM has served our data set better.



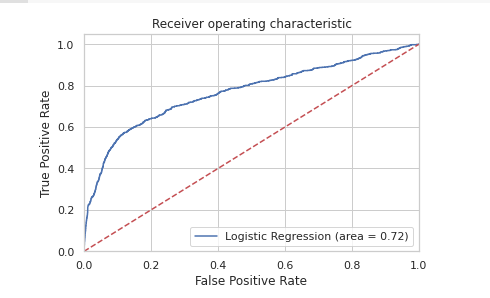
Finally, we can conclude the fact that **SVM works best** with the banking dataset

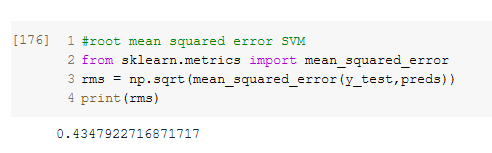
# (e) How significant your findings are

|  |  |
| --- | --- |
| Logistic Regression |  |
| Support Vector Machine |  |

The accuracy for SVM is 81% which is a considerable finding.

ROC Curve of svm is as follows.





RMSE shows a value of 0.43. An RMSE value lower than 0.5 reflects a good prediction rate of the model.

In Logistic Regression, recall value is 0.75 while in SVM it’s 0.84. Both the values are considerably good scores which signifies the findings.

\*\*\* END OF REPORT \*\*\*