**Insurance Claims- Fraud Detection**

**Using Python**

Problem Statement:

**Business case:**

Insurance fraud is a huge problem in the industry. It's difficult to identify fraud claims. Machine Learning is in a unique position to help the Auto Insurance industry with this problem.

In this project, you are provided a dataset which has the details of the insurance policy along with the customer details. It also has the details of the accident on the basis of which the claims have been made.

In this example, you will be working with some auto insurance data to demonstrate how you can create a predictive model that predicts if an insurance claim is fraudulent or not.

**Problem description:**

The goal is to detect the fraud which have classes in nature

**Methodology:**

Here, I am going to use 5 simple steps to analyze Employee Attrition

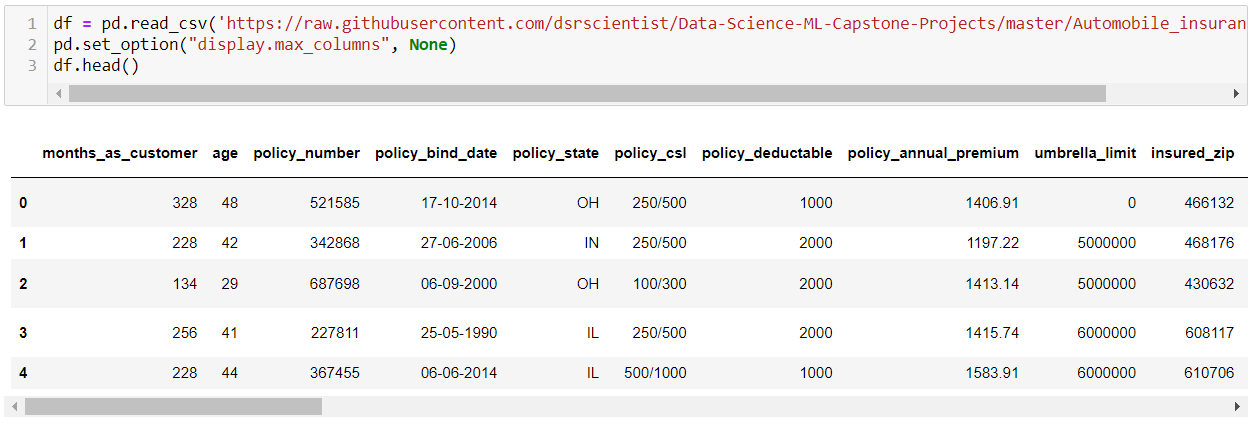
1. **DATA COLLECTION**
2. **DATA PRE PROCESSING**
3. **DIVIDING THE DATA into TWO PARTS “TRAINING” AND “TESTING”**
4. **BUILD UP THE MODEL USING “TRAINING DATA SET”**
5. **DO THE ACCURACY TEST USING “TESTING DATA SET”**

**Importing Library:**



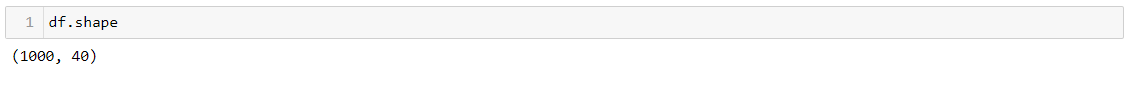
I am importing the all libraries which I required for EDA, Data visualization , Prediction and finding all Matrices . The reason of doing this is that it become easier to use all the import statement at one go and we don’t require to import statement at each point

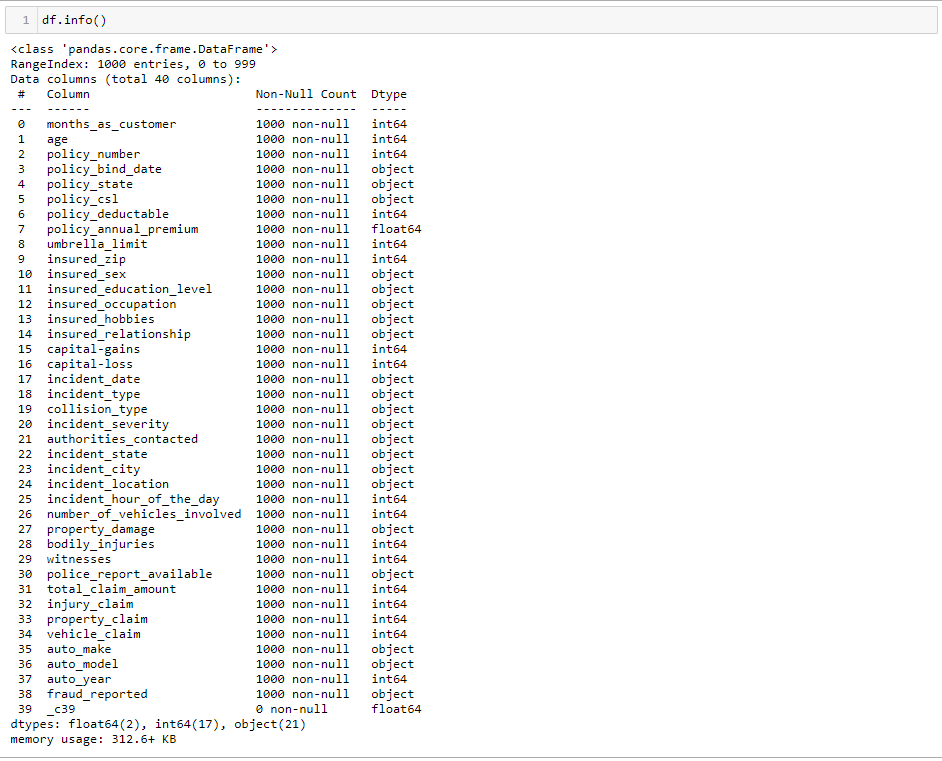
**Loading Dataset:**

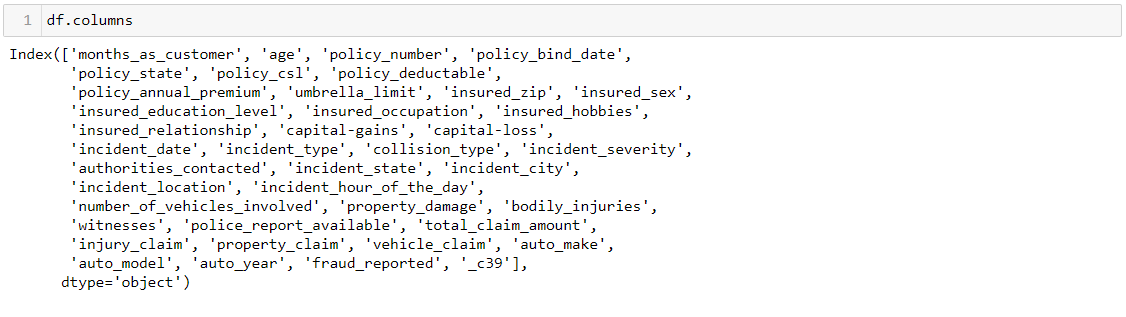


Here I’m loading the dataset into variable i.e ‘ df ’ and processing the first 5 rows and all the data columns with pd.set\_option

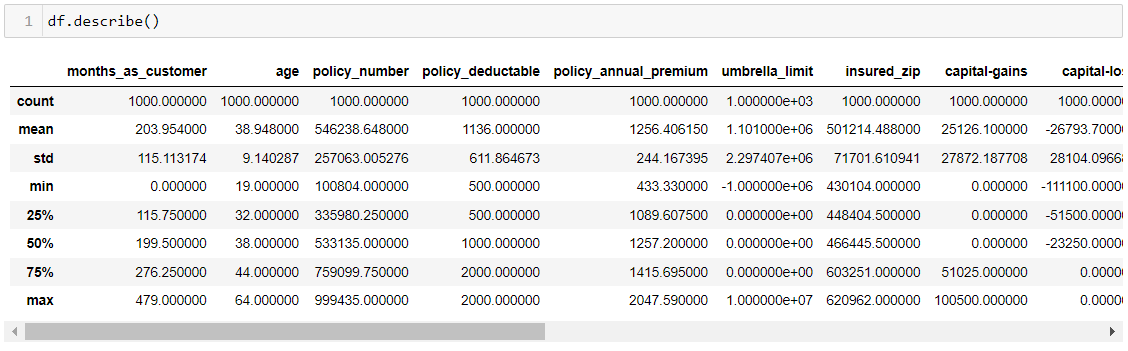
**Exploratory Data Analysis:**

****



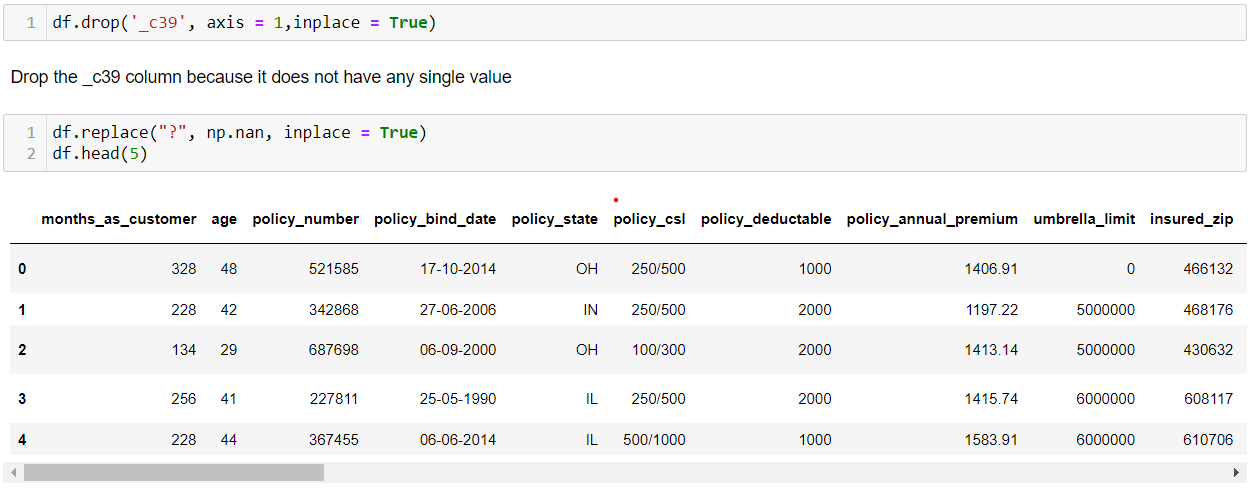


* As seen in the above scenario :
* I checked the shape of the dataset where we get to know there are 1000 rows and 40 columns in this dataset.
* And we can see some of the columns have ‘Int64’ and some of them have ‘object’ dtype , and some them has float dtype column in this dataset.
* And there are nan’s present in this dataset. \_c39

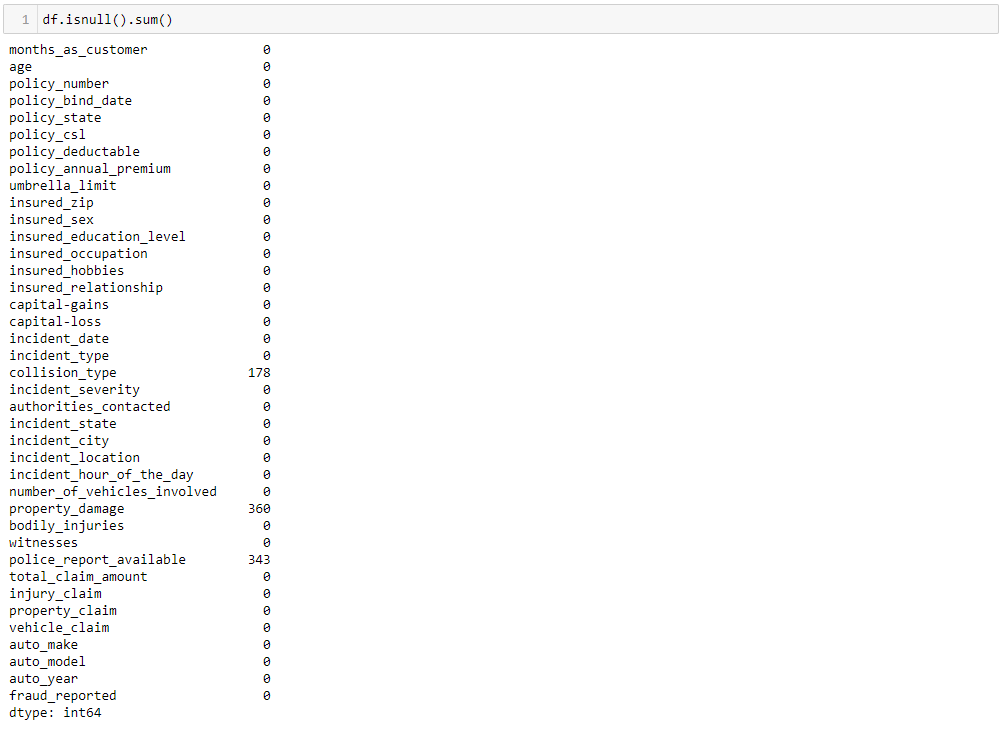


* Above Statistics data shows that their multiple outliers mostly in umbrella limit column
* There is also difference between mean and 50% value in some of the columns which used to get fix for better prediction
* Also, number of rows in each column are same only \_c39 column have 0 value

**Detecting the Missing Values:**

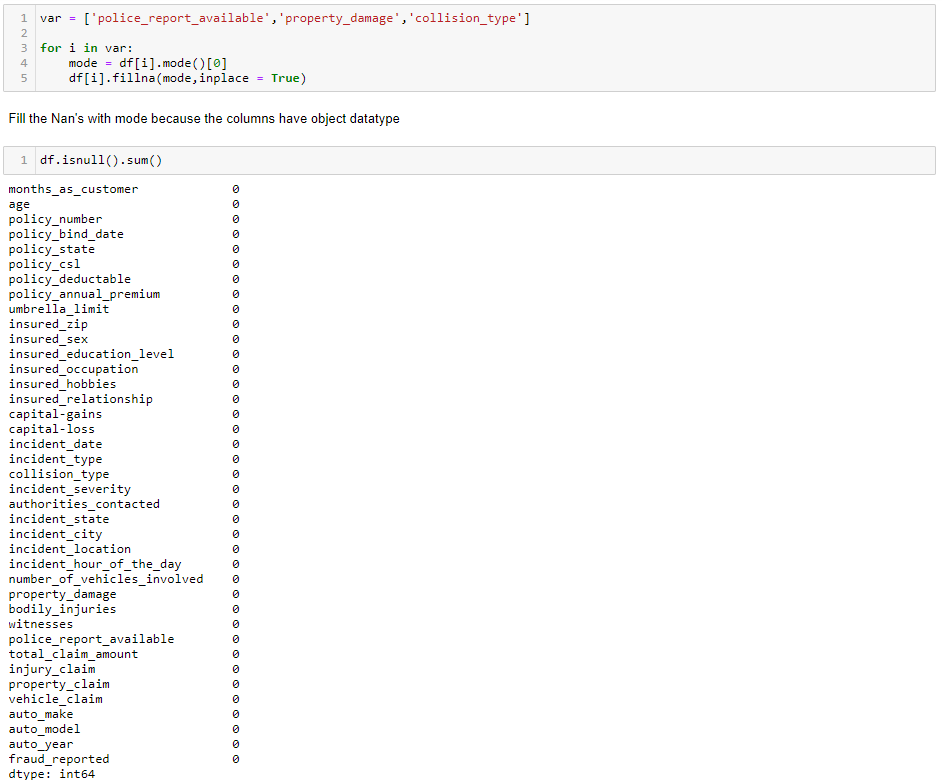


* Droping the \_c39 column because it has 0 values.
* Replace the question marks as Nan.
* As we see earlier there are no null values, But now I changed the question marks to Nan’s
* Now again checking missing values



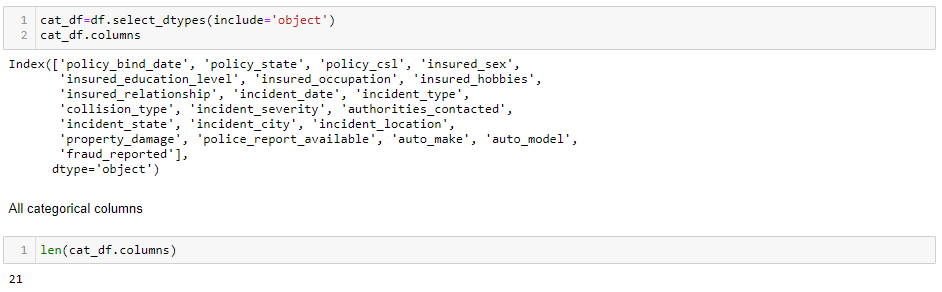
And we can see the Nan’s in the dataset. These are question marks turned into Nan’s.

So we have fill up them using the mean or median method.

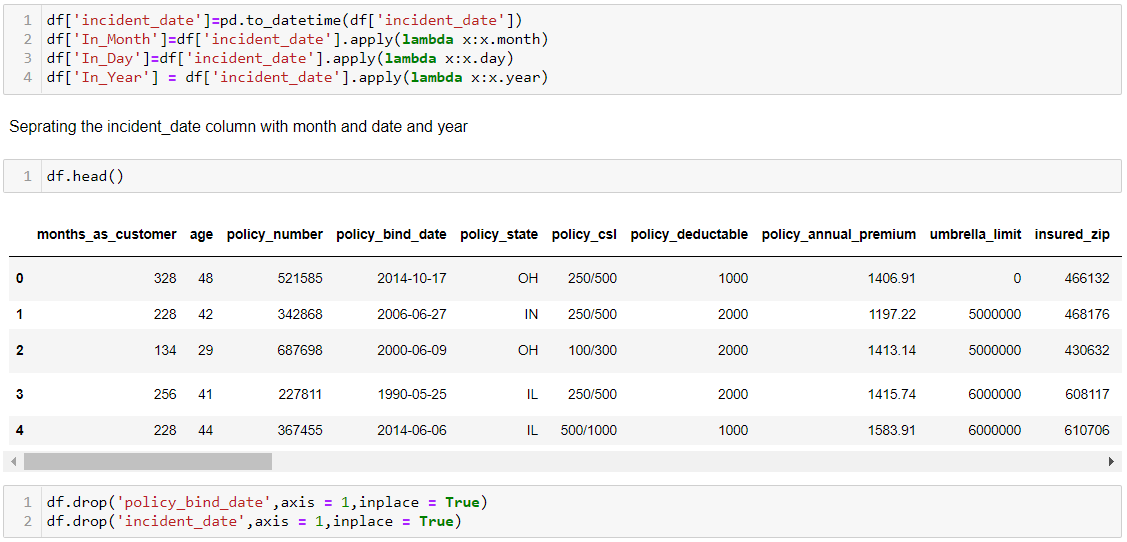


After filling the Nan’s we can see that there are no null values present in the dataset

All the missing values are in categorical columns, so we filled them by using mode method



As we can see policy\_bind\_date & incident\_date has object data type. So we have to convert them into datetime format

****

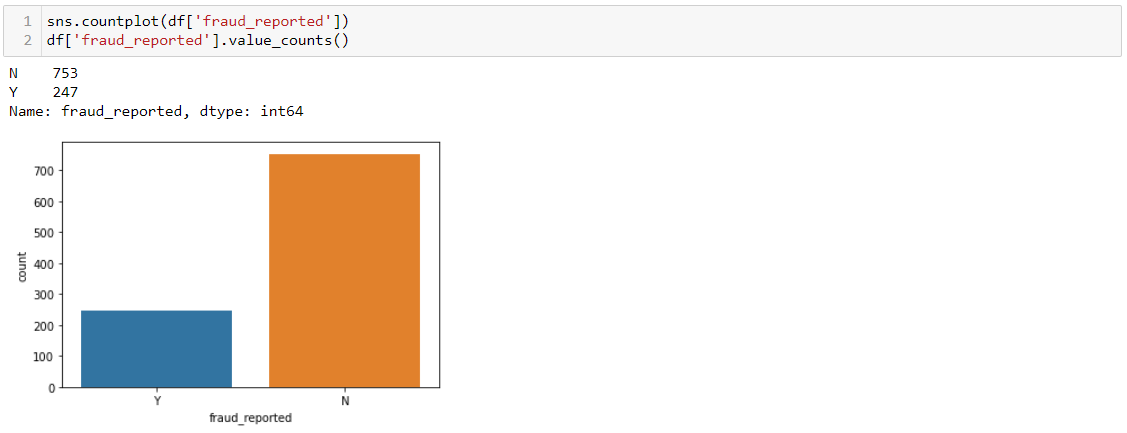
As we can see we converted the policy\_bind\_date and incident\_date from object dat type to datetime format and separate the date, month and year with new column and dropped the policy\_bind\_date & incident\_date because we already have the values in seprate form

**Data Visualization:**

In this portion we plot different graphs and try to visualize the data

We use different graph include:

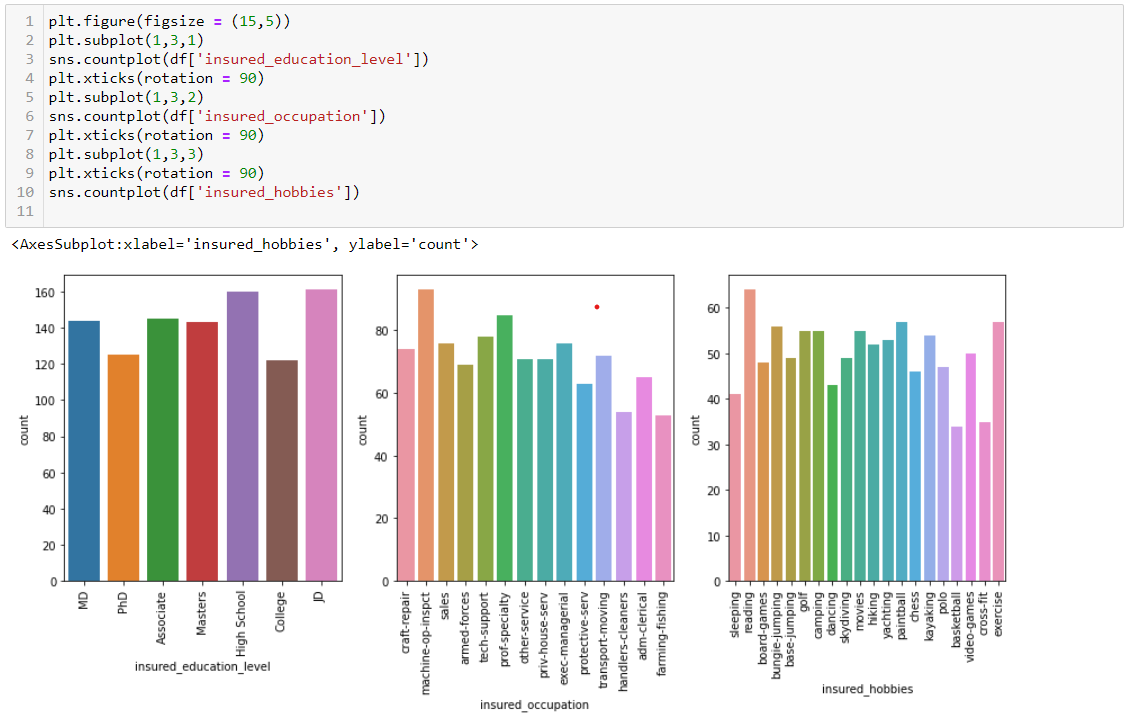
* Bar plot
* Count plot
* Box plot
* Dist plot
* Violin plot
* Cat plot
* Pairplot



As we can see:

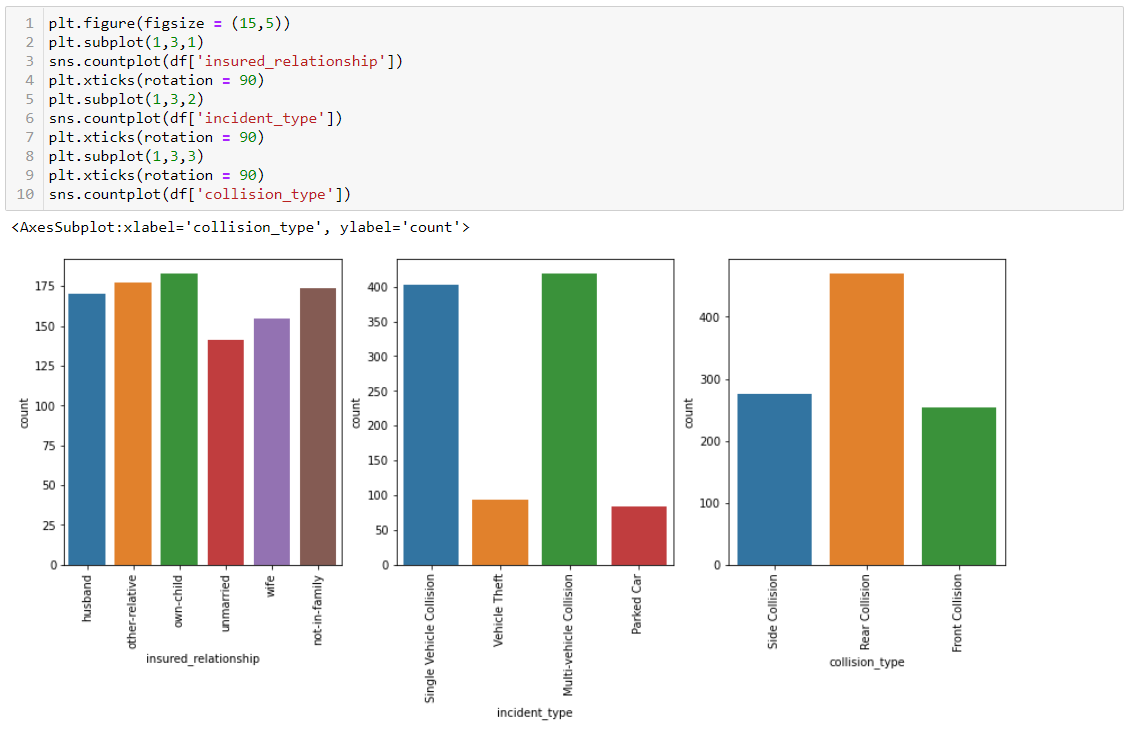
* Fraud\_reported: Yes , Around 18 to 20 % of Total People count along the table.
* Fraud\_reported: No, Around 80 % of Total People count along the table.

Observing the situation we can say that there is class imbalancing problem in the dataset.



From the above observation we can say that:

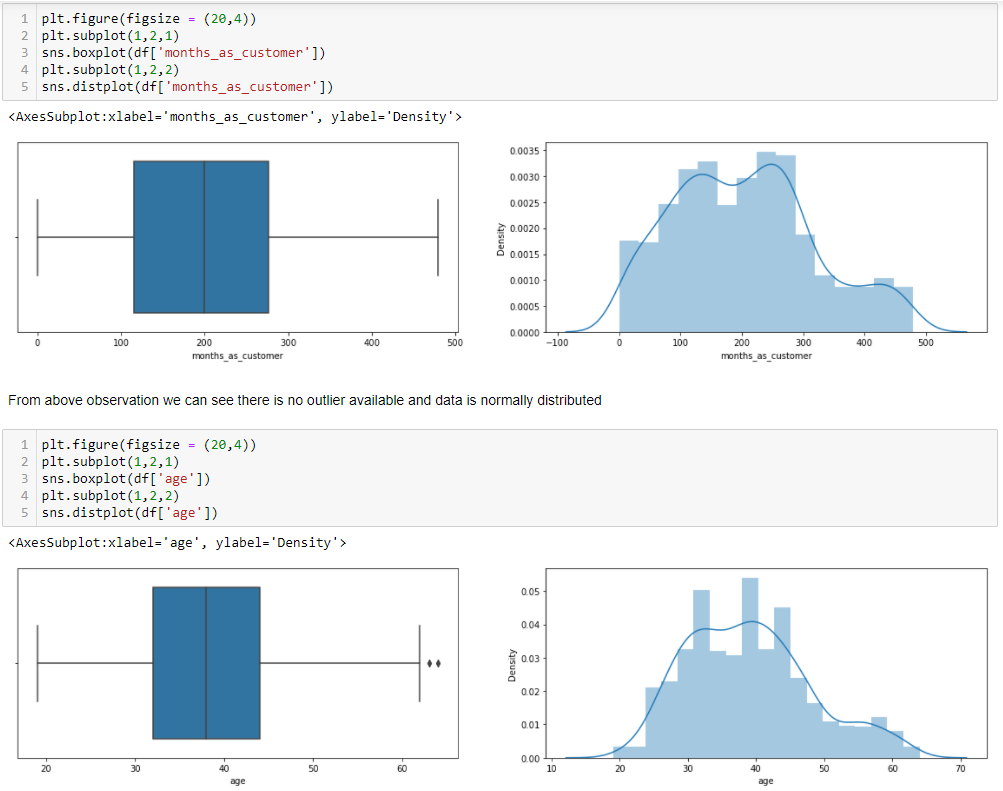
Insured\_education\_level, Insured\_occupation and Insured\_hobbies columns does not have any proper correlation with the fraud reported data



From the above observation we can say that:

* Unmarried people reported less fraud.
* In incident\_type most no. of fraud reported of ‘single vehicle collision’ and ‘multi vehicle collision’ and fraud reported of parked car and vehicle theft are less
* In collision\_type Rear collision is has high no’s in fraud reported than the front and side collision type

Now we will check outliers and skewness using boxplot and distplot.

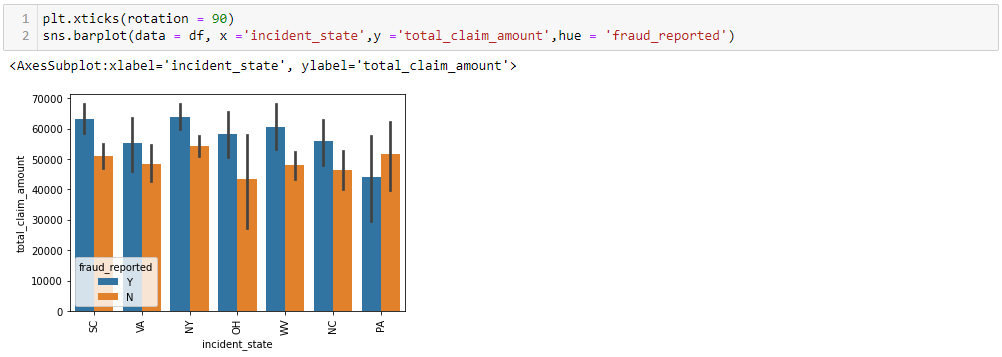


From the above observation we can see that the ‘month\_as\_customer column is clear, the column is normally distributed and don’t have outliers

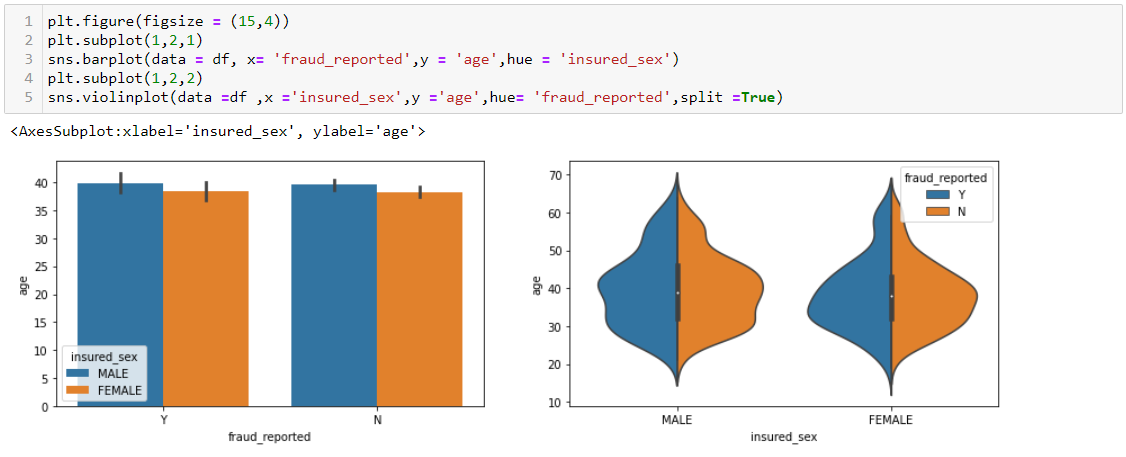
But, We can see that below column means ‘age’ have outliers and data is positively skewed.

As by this code we checked outliers for every column and from that we get couple of columns with the outliers.

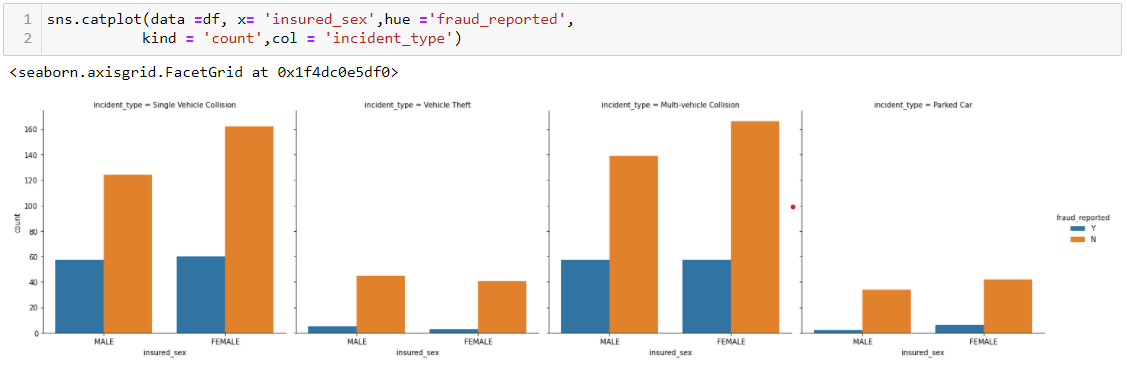
We will removed them in further process.

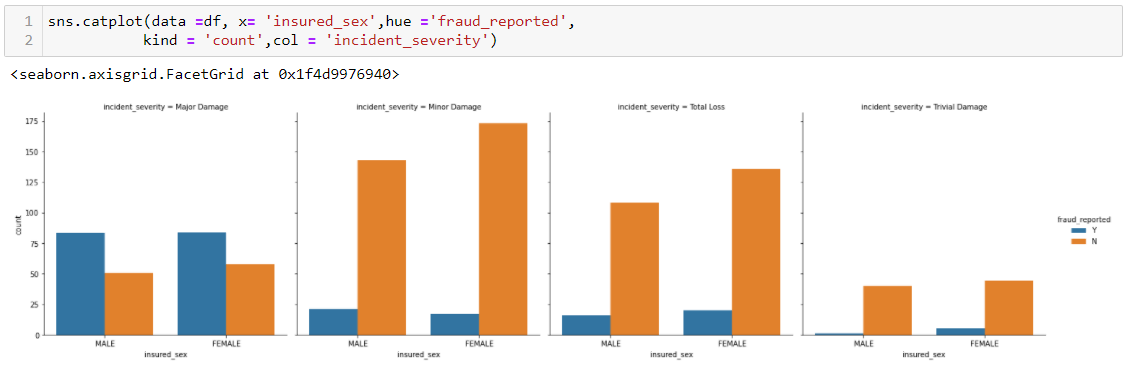


From the above observation we can say that all the states have same no of total\_claim\_amount, less no of the people who did not reported fraud in every state



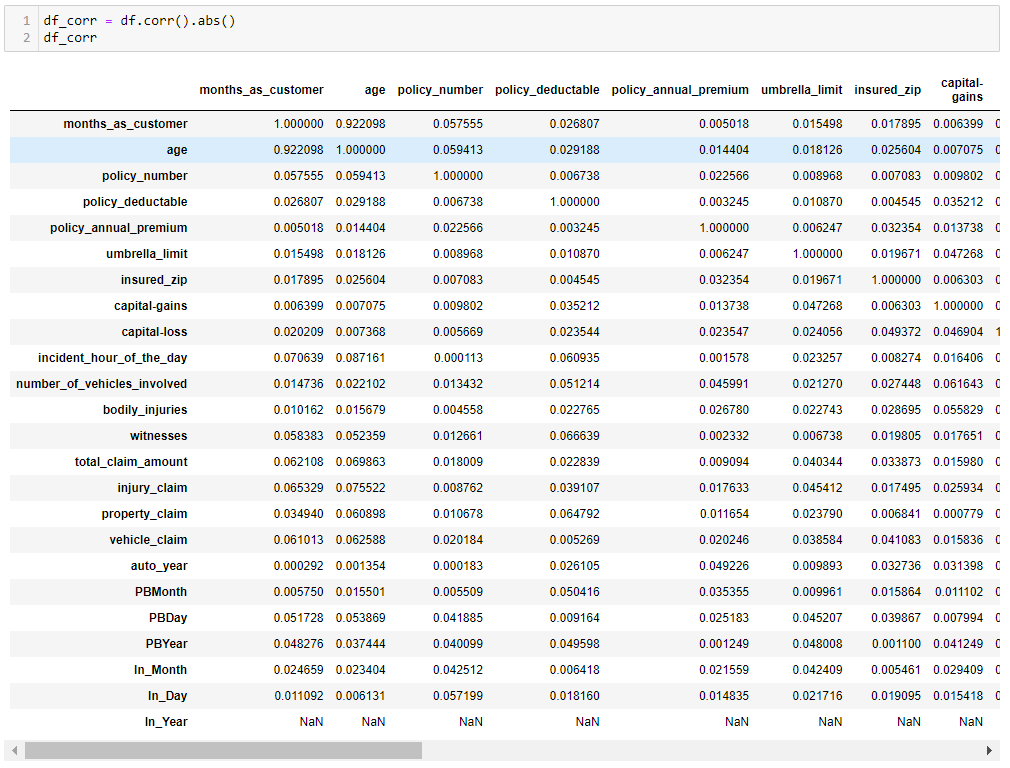
The People of Age between 30 to 45 has reported more fraud’s



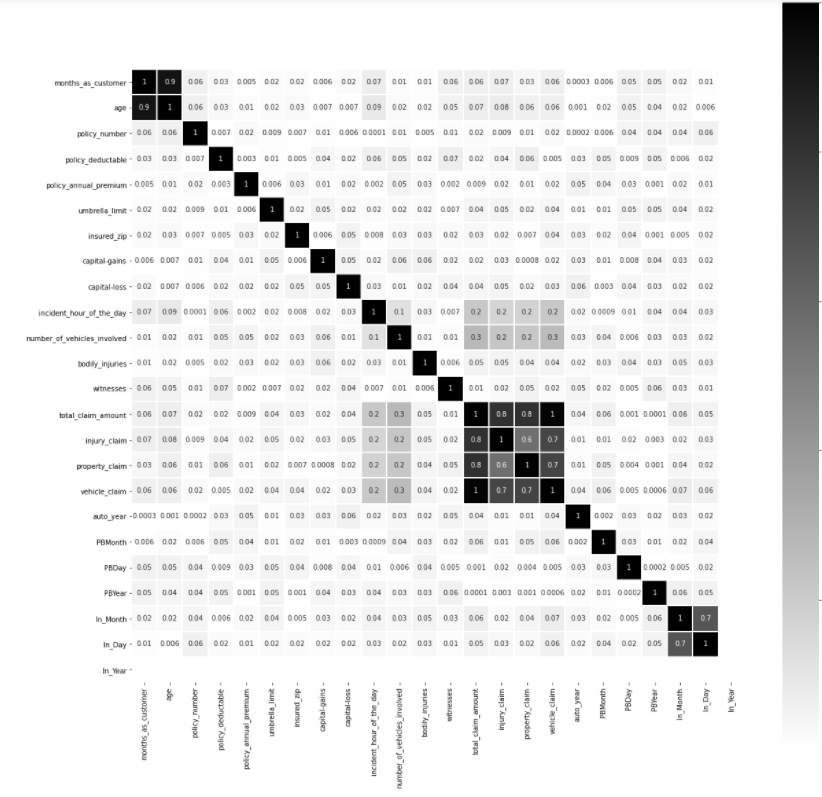


* In incident\_type we can say that Vehicle theft and Parked car have less fraud reported
* And in severity people have major damage have more no of Fraud reported and less no with trival damage.

**Checking correlation of the columns:**

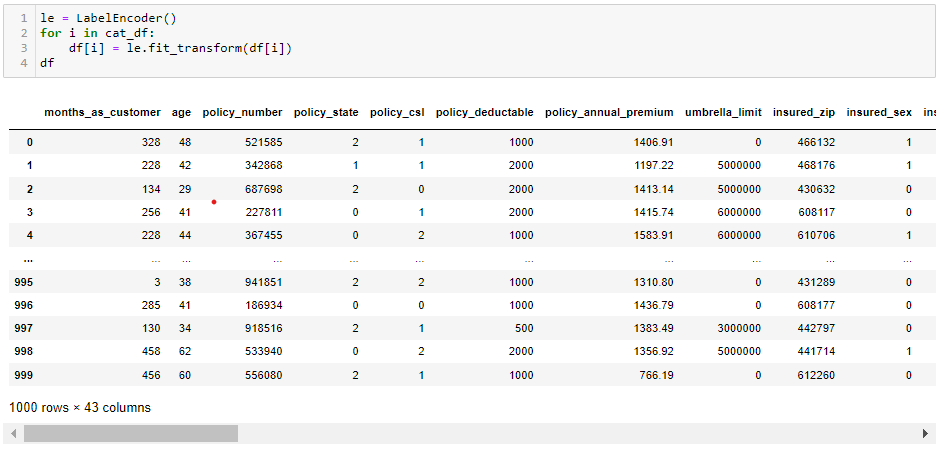


We can see the correlation between the columns, for better visualization we plot the **heatmap:**



As we can see ‘total\_claim’, ‘injury\_claim’, ‘property\_claim’ and vehicle\_claim columns have good correlation with each other

**Data Preprocessing:**



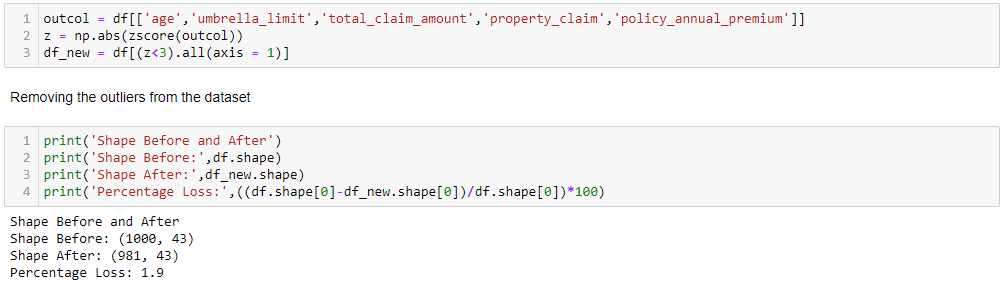
As we can see all categorical values converted into numerical values.

Now we will proceed with the encode data.

**Outliers:**

An outlier is a data point that is distance from all other observations. A data point that lies outside the the overall distribution of the data set.

As we saw earlier some of the columns have outliers and we have to remove them



From the above image we can see that after removing outlier by using zscore method in result we loss 1.9 % of the data.

The rows reduced from 1000 to 981

And that much data loss is acceptable, So we will proceed further with this new data.

**Seprating the Features and Target:**

Now we have seprate the features and target column for the preprocessing

v.png

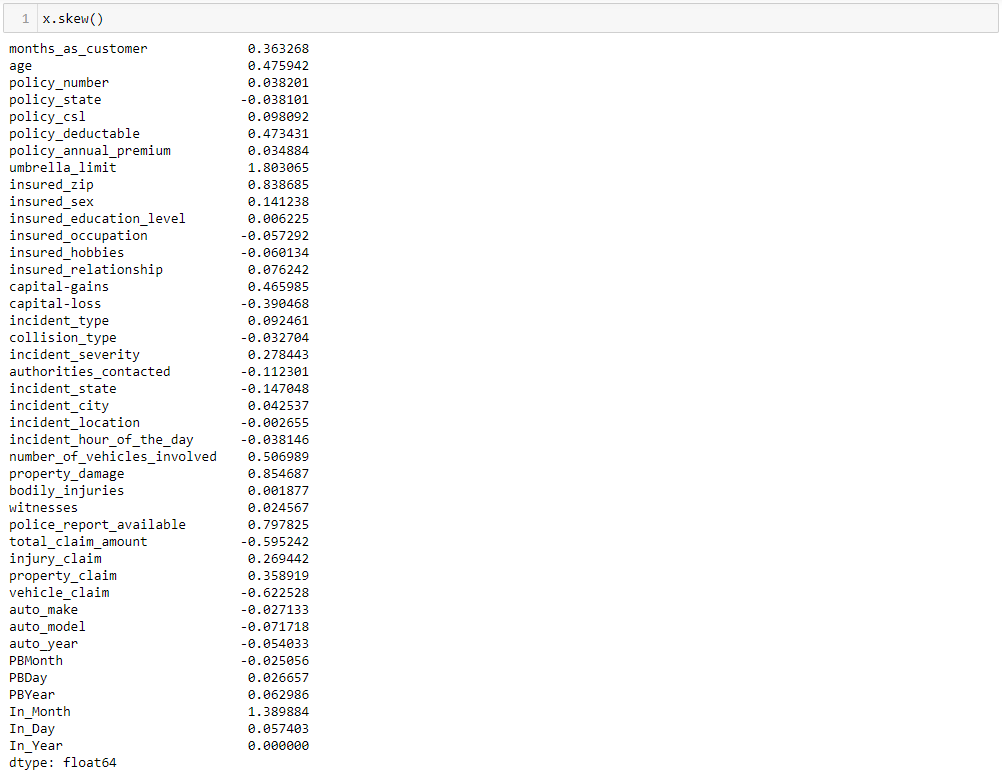
X have all the features and y have target variable.

**Skewness:**

Skewness refers to a distortion or asymmetry that deviates from the symmetrical bell curve, or normal distribution, in a set of data. If the curve is shifted to the left or to the right, it is said to be skewed.

As we saw earlier some of the columns are positively skewed so we have to remove skewness of the columns

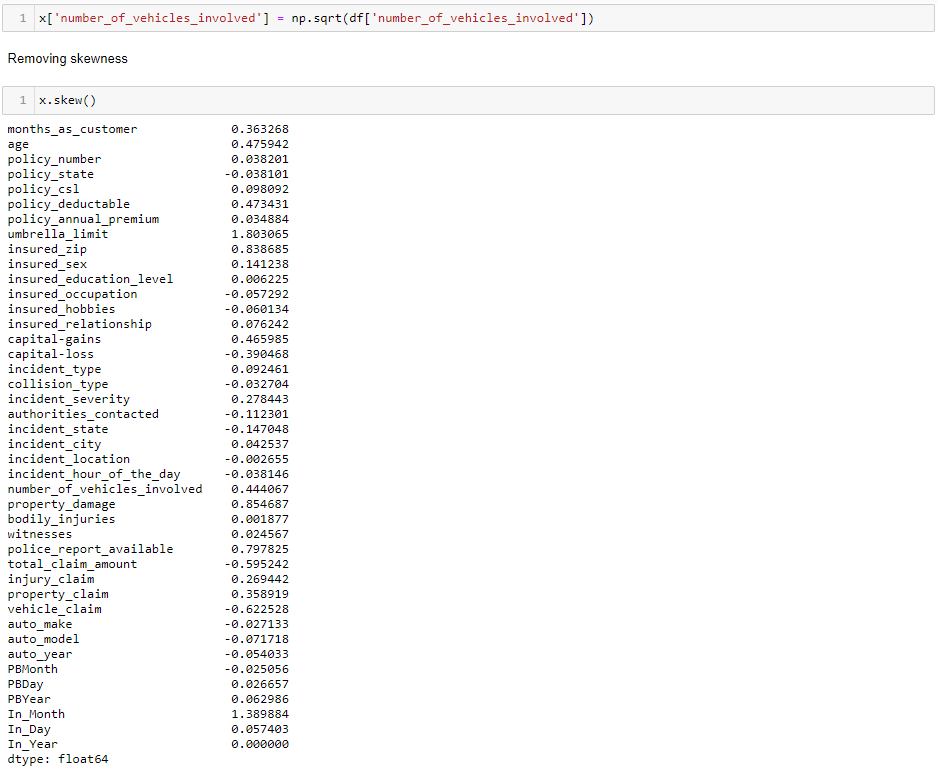
First we have check skewness



From the above image we can see the columns with the skewness.

* The skewness value between 0.5 to -0.5 is acceptable. The data above this value can harm model
* We can only remove skewness of the numerical columns, If there is skewness in categorical ones, not to worry about them.

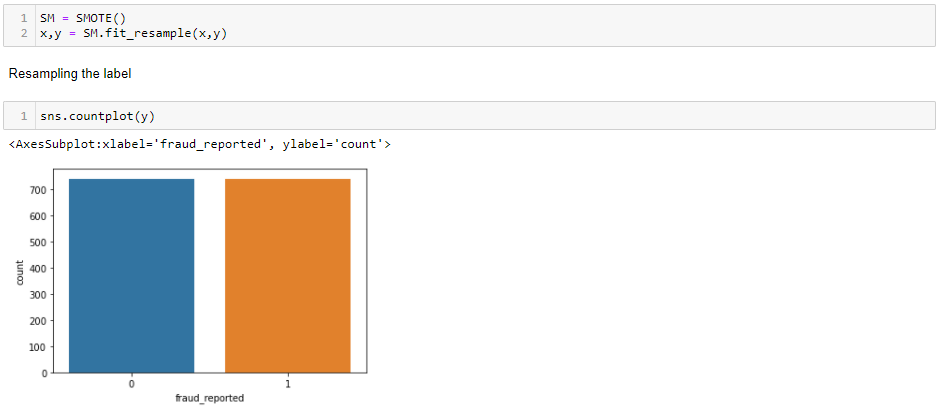
So we have to remove skewness of numerical columns.



So by the above code we removed the skewness using the sqrt method And we can see now there is no skewness available in the data

**Resampling the target variable:**

* As we check above, There were class imbalancing problem so we need to cure that
* If we proceed as it is then it will be problematic for the models and will not give good score.
* We do upscaling of the data using the SMOTE method.



As we can see both 0 (yes) and 1(No) have equal data.

**Standard Scaler:**

The idea behind StandardScaler is that it will transform your data such that its distribution will have a mean value 0 and standard deviation of 1. In case of multivariate data, this is done feature-wise (in other words independently for each column of the data).

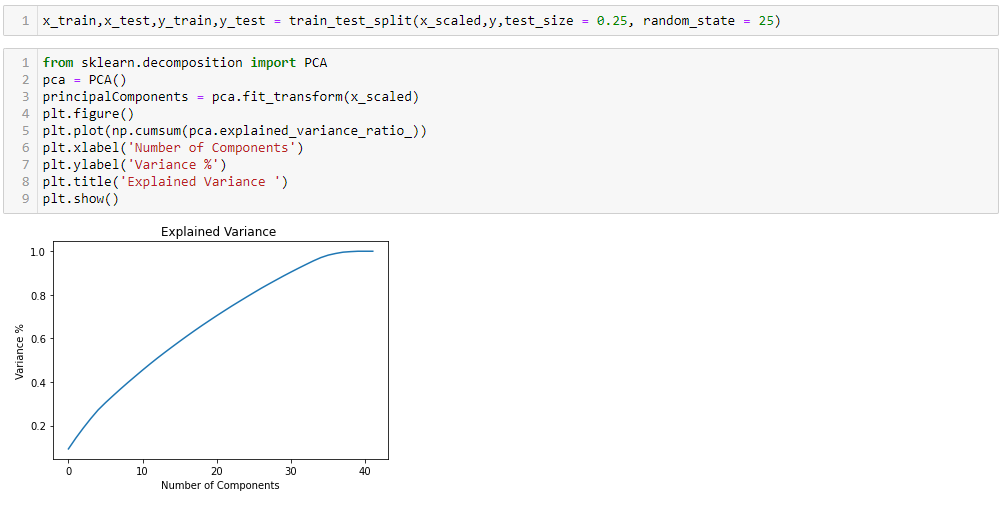
We use standard scaler for scaling the data

z.png

**PCA:**

Principal Component Analysis, or PCA, is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.

So we used PCA for dimensionality reduction because the dataset have total 42 columns and if we proceed further with this data then the model will be overfitted

****

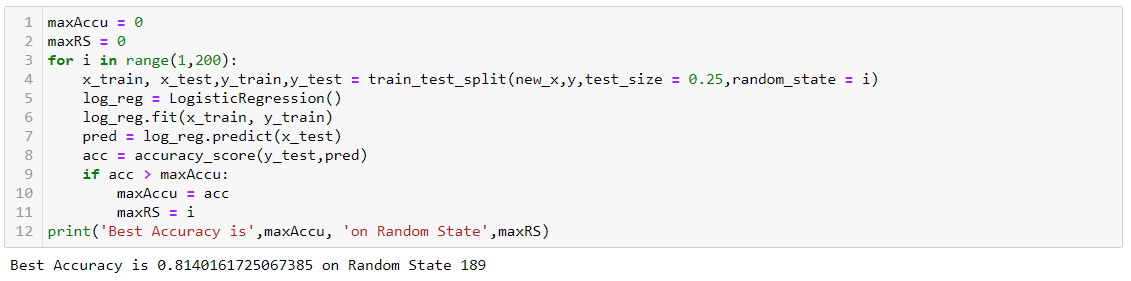
We can see the variance of the dataset if we use 30 components still we get 80 % of variance.

So we reduced the columns from 42 to 30 using PCA



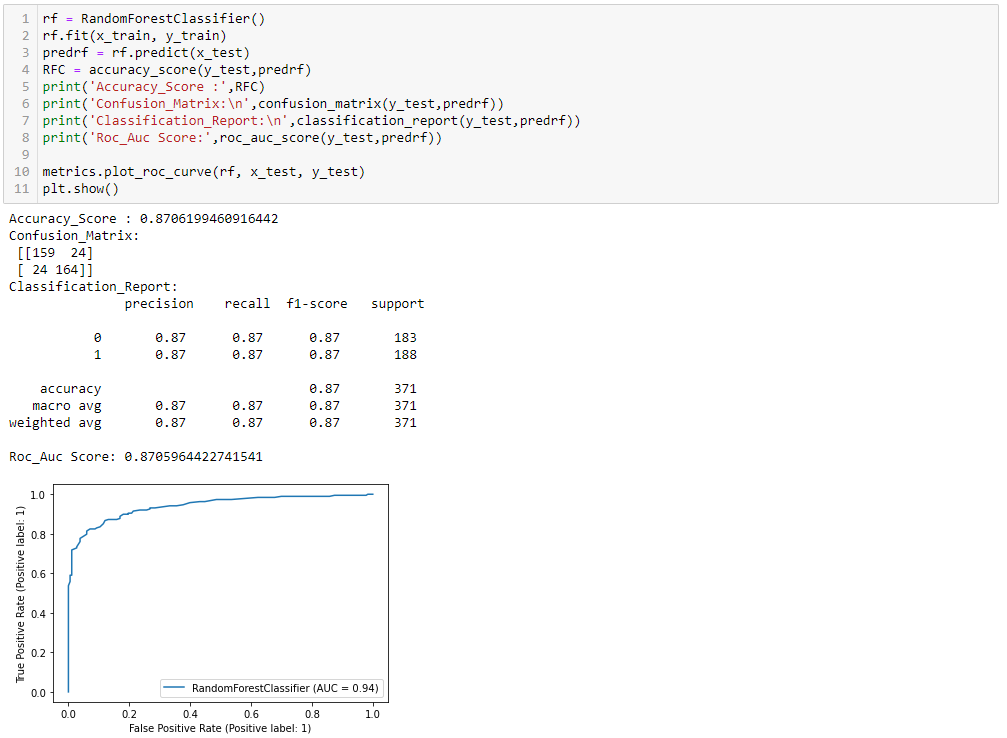
**Prediction:**

Above I am using the for loop which help me to find accuracy score at each random state and for the best state where acc score is maximum is come as output value.



So we get 81.40 accuracy score at 189 Random State

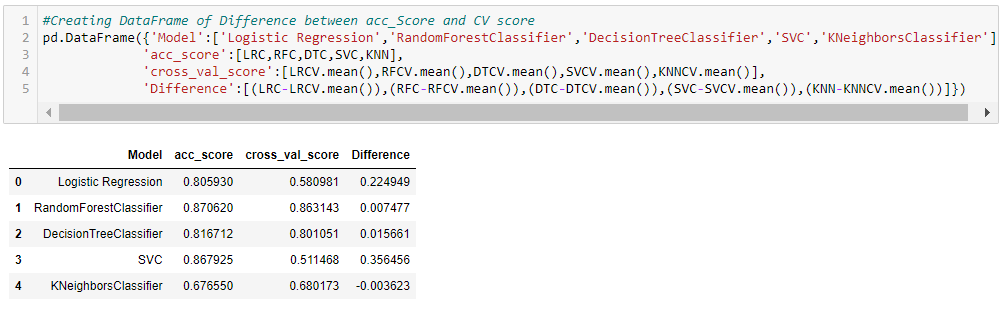
Now we use this random state to train the model



Using RandomForestClassifier we get:

* Accuracy score = 85.08
* F1 score = 87
* Recall = 87

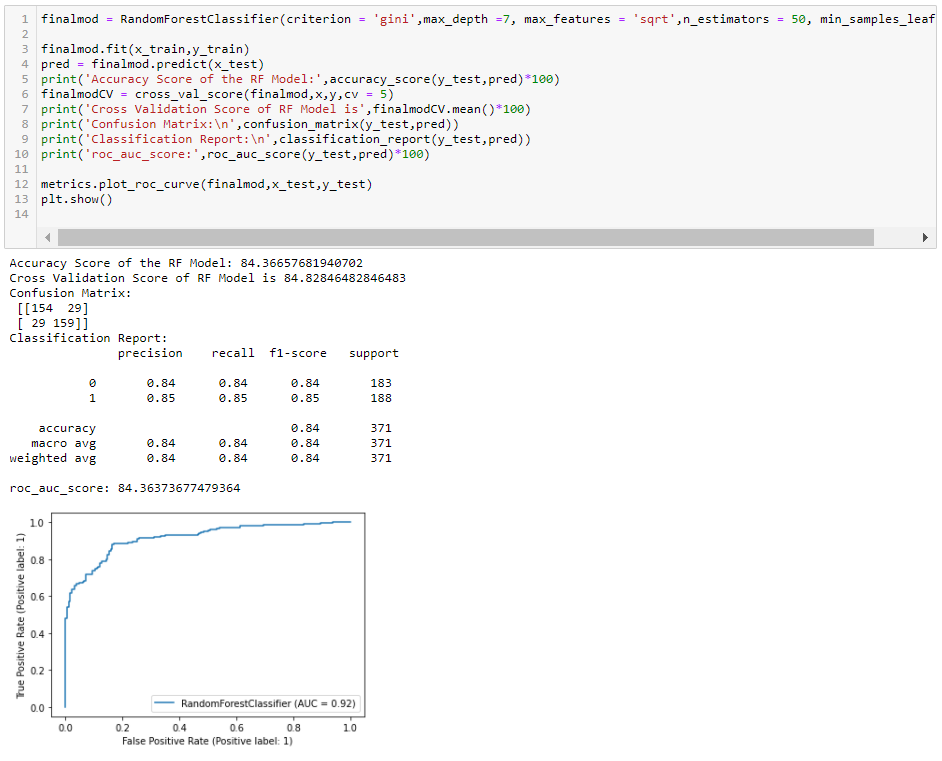
As RandomForest we used 4 more algorithms for model building anf from that we get best acuuracy score from random forest.



We plot the roc\_curve also

But we can’t take it as a final score, We try to get best parameters for RandomForestClassifier by using GridSearchCV





We get best params using the GridSearchCV

As we can see by putting the best params to the model in result we get 84.36Accuracy Score and 84.82 CV score.

By using GridSearchCV the score is reduced, We get best score by the random parameters and it is not mandatory to get best score using the best parameters.

We will save the random parameters model for the further work



By using joblib we saved the model.

**Conclusion**:

The machine learning models that are discussed and applied on the datasets were able to identify most of the fraudulent cases. This enables loss control units to focus on new fraud scenarios and ensuring that the models are adapting to identify them. Certain datasets had severe challenges around data quality, resulting in relatively poor levels of prediction.

Given inherent characteristics of various datasets, it would be impractical to define optimal algorithmic techniques or recommended feature engineering for best performance. However, it would be reasonable to suggest that based on the model performance on back-testing and ability to identify new frauds, the set of models offer a reasonable suite to apply in the area of insurance claims fraud. The models would then be tailored for the specific business context and user priorities.