**INSURANCE CLAIM – FRAUD DETECTION BLOG ARTICLE**



**Insrance Claim -Fraud Detection Status Prediction**

****

**Introduction:**

Insurance Claim is made by the requestor to the policy provider. Insurance can be made for Home, Property, Land, Accident, Car, Health, Auto etc. We can claim if our home is damage from Earth quake or got into an accident. At that moment we can do insurance claim by submitting the form. So, that we don’t give money from our pocket. For example, if we have Health Insurance, suddenly we get serious health issues, we no need to worry of money, whether we can afford the medical expenses or not. We can claim the health insurance by submitting the insurance claim form. All the medical expenses can be done in the health insurance claim. In this article we see the Auto insurance claim, to predict where the insurance claim is fraudulent or not.

**1)Problem Statement:**

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

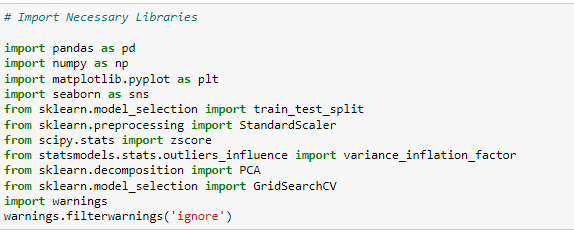
**2)Data Analysis:**

Data Analysis is the process of cleaning, transforming, pre-processing, modelling data to get a useful information and make a prediction.

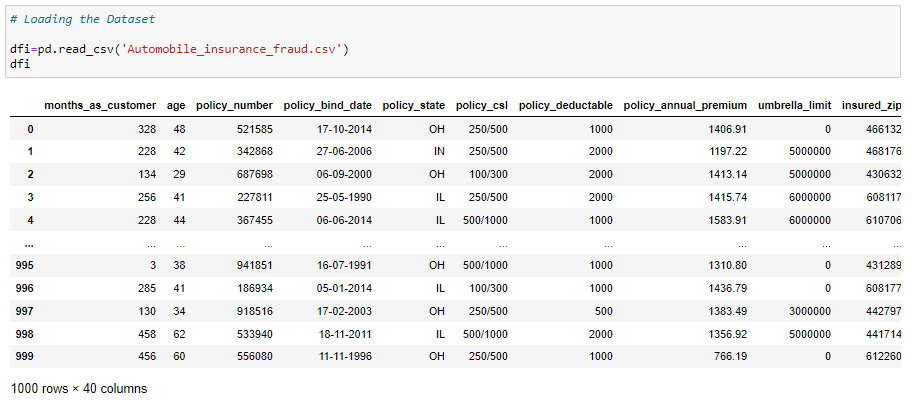
Data Analysis can consist of Text, Statistical analysis, Description, Diagnostic, Predictive Analysis.

Data Analysis plays a very important role, before building the model, we do the analysis of data, we can replace any missing value present in data, we can remove unnecessary column from dataset. We can convert categorical data to numerical data using Label Encoder/Original Encoder. Data Analysis play an important role in Decision making, improve accuracy and help in scientific approach to give a good insight using a visualization technique.

**Importing Necessary Libraries:**



**Loading the dataset:**



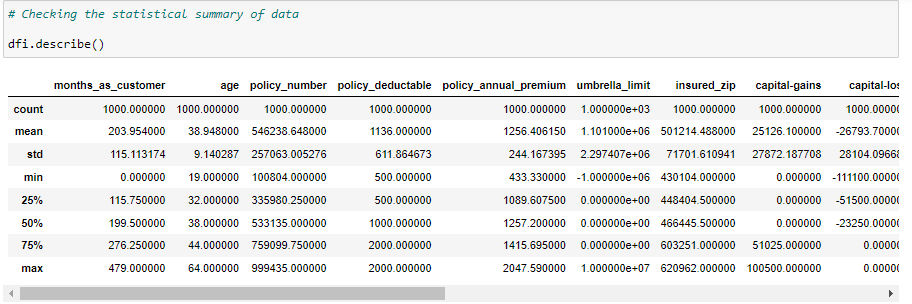
There are 1000 rows and 40 columns present in Insurance claim dataset. It contains both categorical and Numerical data.

**Data Pre-processing and Data Cleaning:**

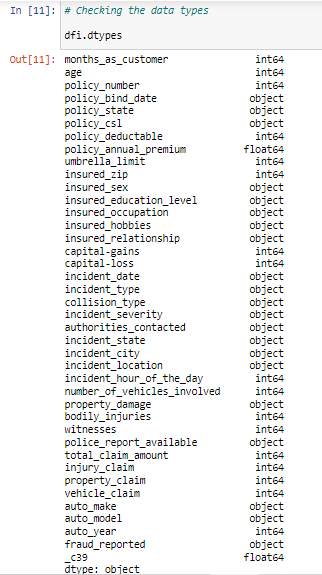
In data pre-processing we check the statistical summary, data info, Unique value, column name, shape of data, value count.

In Data cleaning, we drop the unnecessary column, fill the missing value with mean/median if it is numerical data. If it is categorical data, we can replace with mode.

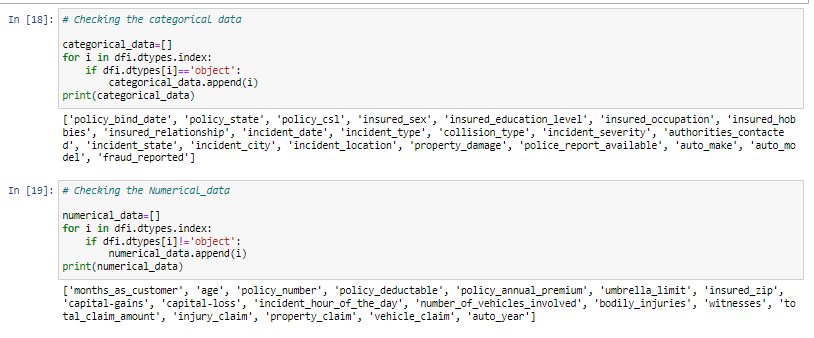
**Statistical summary of data:**



Using the describe method, we check the statistics of data, their count value, mean, standard deviation, Minimum value, Maximum value, 1st Quartile, 2nd Quartile, 3th Quartile. Here mean value is greater than median.



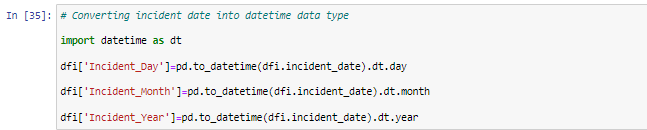
The data contains float, integer and Object data type.



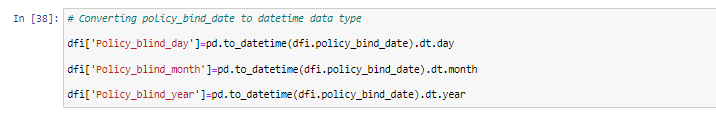
Separated the data into categorical and numerical data to understand it more clearly. We can visualize it separately for categorical and numerical data.



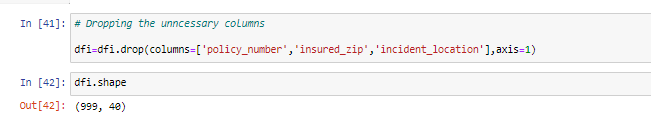
In dataset most of the places contain ‘?’, So, we have replaced with ‘No Info’



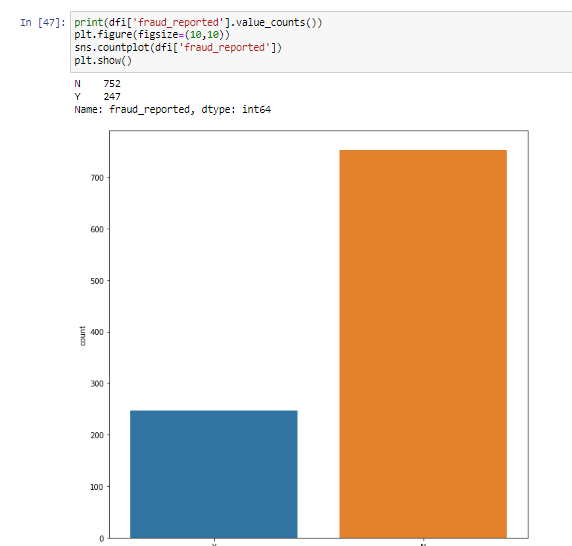
In the Incident Date column, it is considered as Object data type. Which is not correct. So, we have converted date time data type and we have separated into Day, Month and Year.



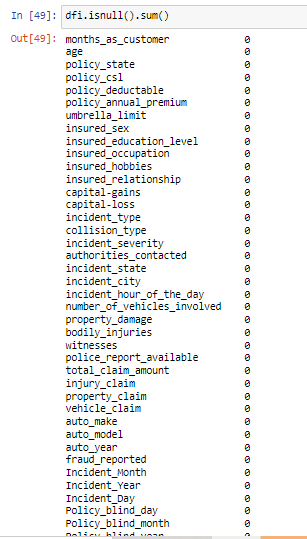
In Policy Bind date, the same way we have converted into datetime data type and separated into Day, Month and Year. Then we can drop the Policy bind date column.



We can drop the unnecessary column, policy number, insured zip, incident location. After dropping the data contains 999 rows and 40 columns.



We can see that a greater number of people reported No for Fraud Claim. No-752. Yes-247. Still few people reported there a fraud claim.

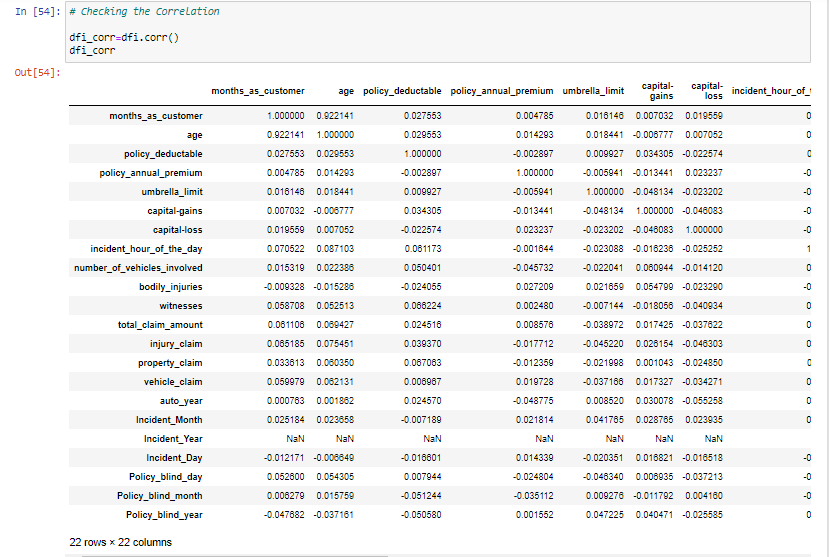


We can see there is no null value present in data.

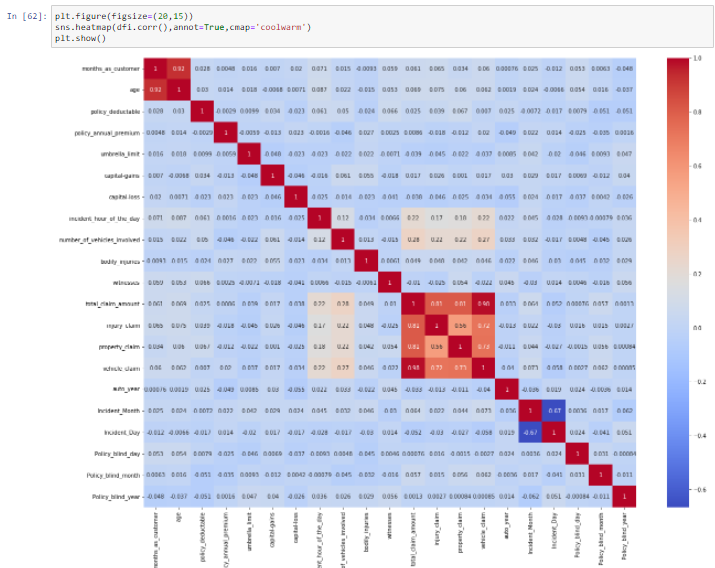


There is no null value present in the data. We have visualized using the heatmap.

**Correlation:**



We checked the correlation of data, there is positive correlation, in few columns there is negative correlation. There is NaN value in Incident Year. So, we can drop the Incident Year column.

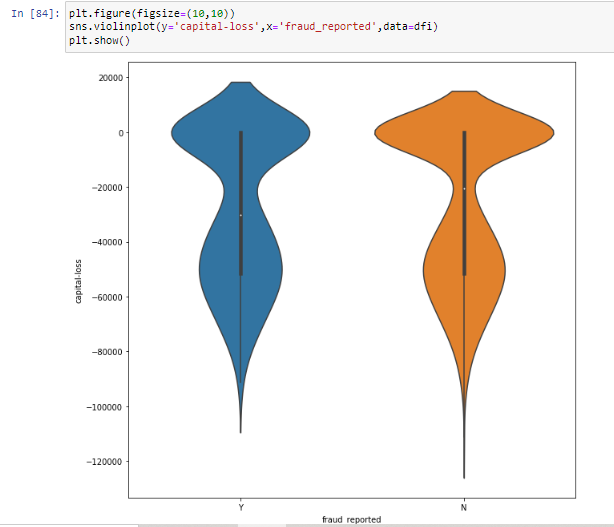


Visualizing the correlation using the heatmap. In total\_claim\_amount, injury\_claim, grocery\_claim, vehicle\_claim there is a high correlation present, other 's columns are less corelated.

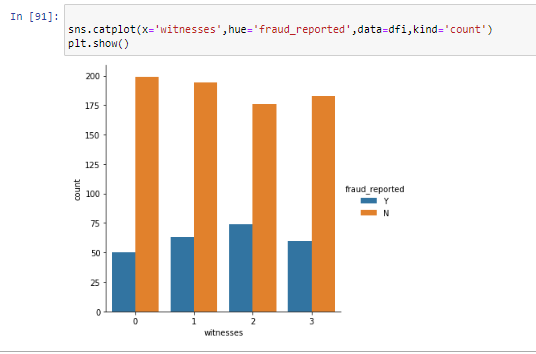




Using the count plot, we have checked the number of vehicles involved with fraud\_reported. Those who have only 1 vehicle, the count is high that there is a no fraud claim. Those who are having more than 1 vehicle, there is high change of fraud claim.



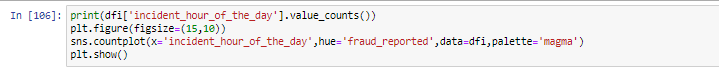
Using the violin plot we checked the Capital loss with fraud reported. We clearly see that the fraud reported in the capital loss ranges between -40000 to -60000

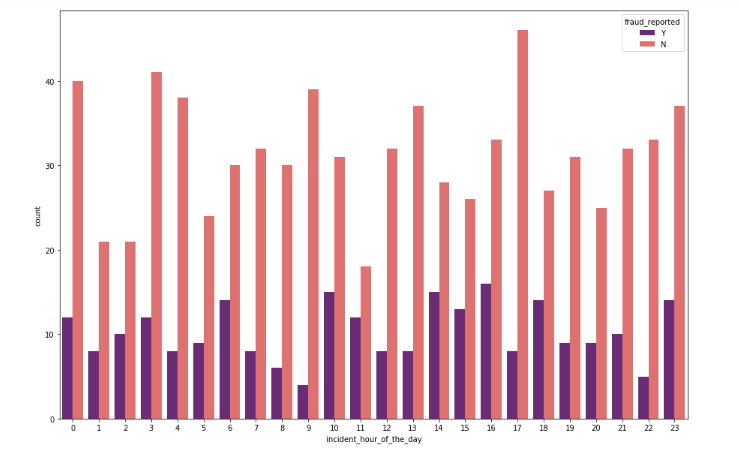


We have more than 1 witness reported the fraud claim. High count is, there is no fraud claim, few people has reported the fraud claim.

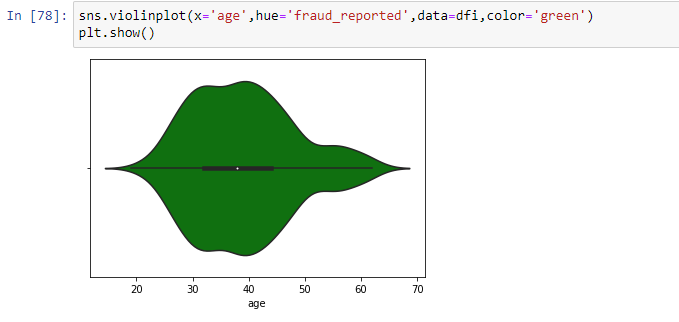


We have compared the Policy deductable and fraud reported. We see that policy deductable count 1000 has more than others.

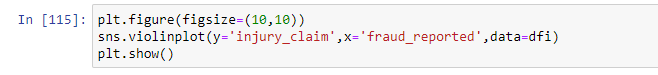


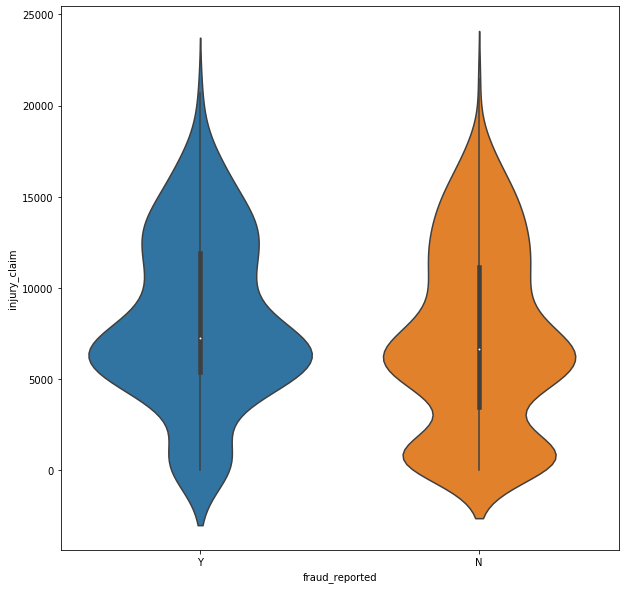


We have checked the incident hour of day; in which hour more fraud claim is happening. In the peak hour there is high fraud claim.

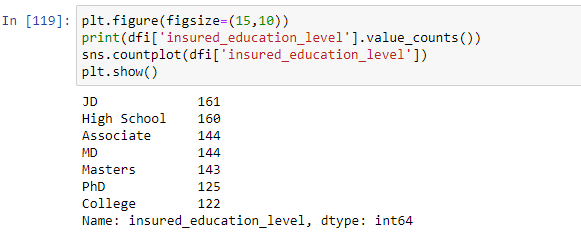


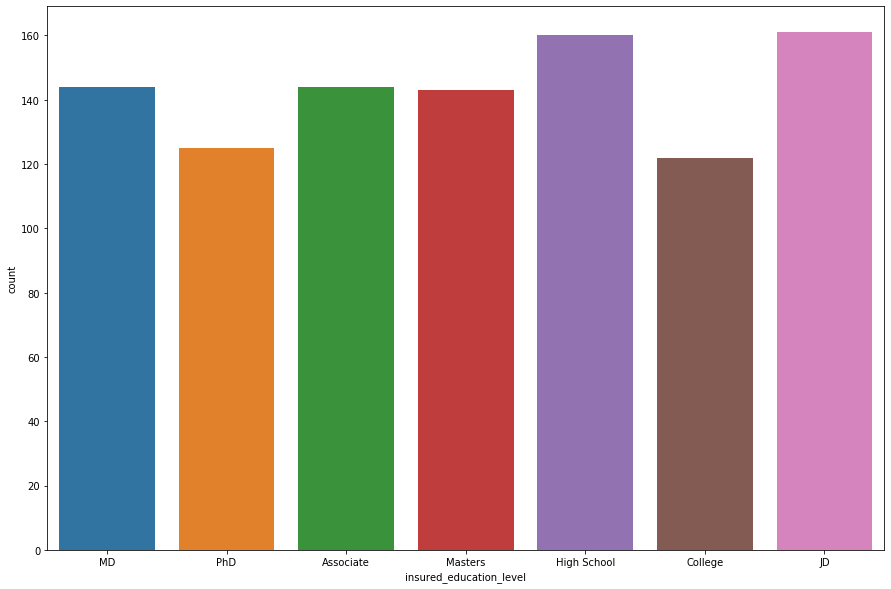
We compare the age with fraud reported, we clearly see that in the range of 30-40 years there is high number of frauds reported.



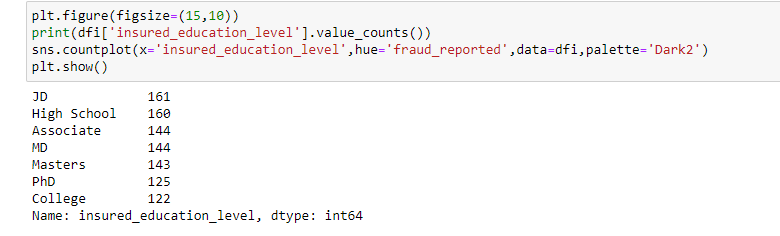


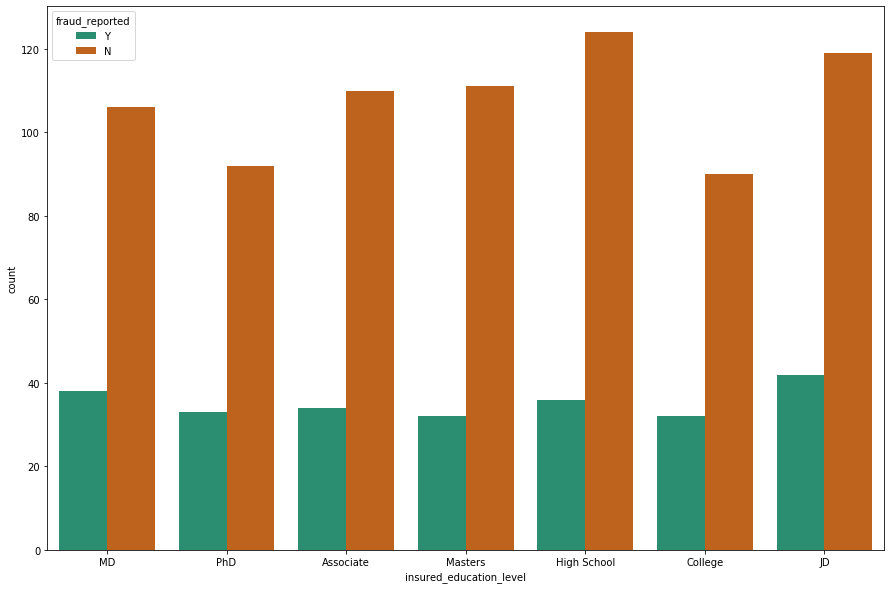
In the injury claim between 5000-10000 there is high number of frauds reported.



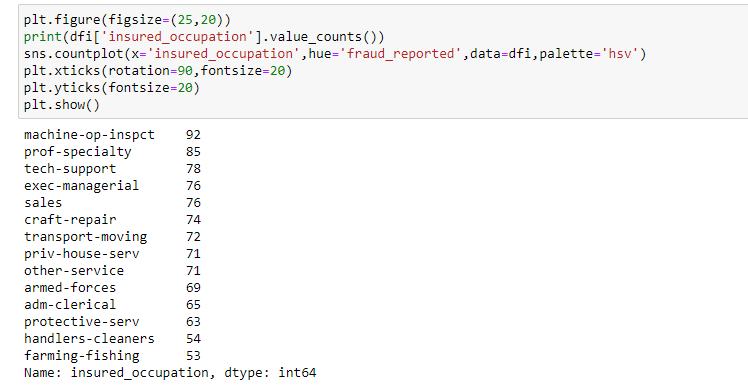


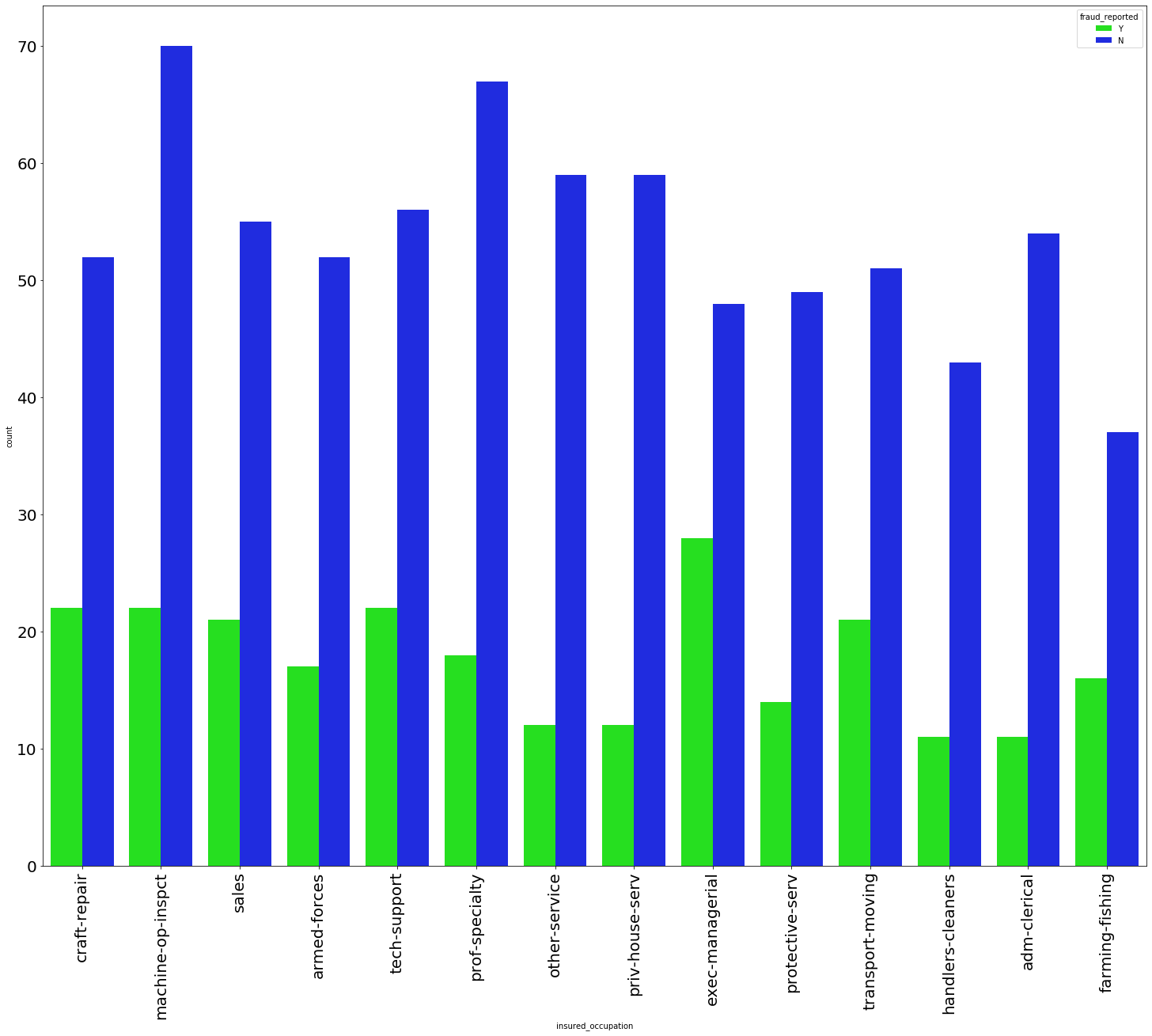
If we see the education level of insured person who has taken insurance from the company. Most of the insured person has completed JD and High School.



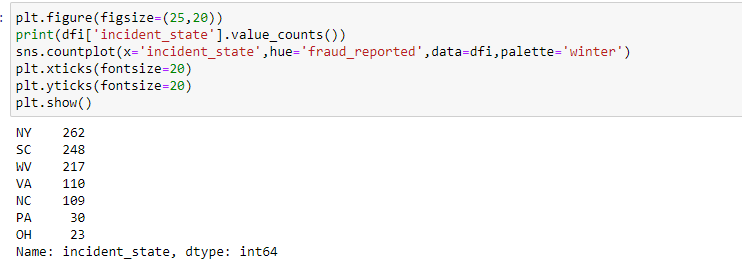


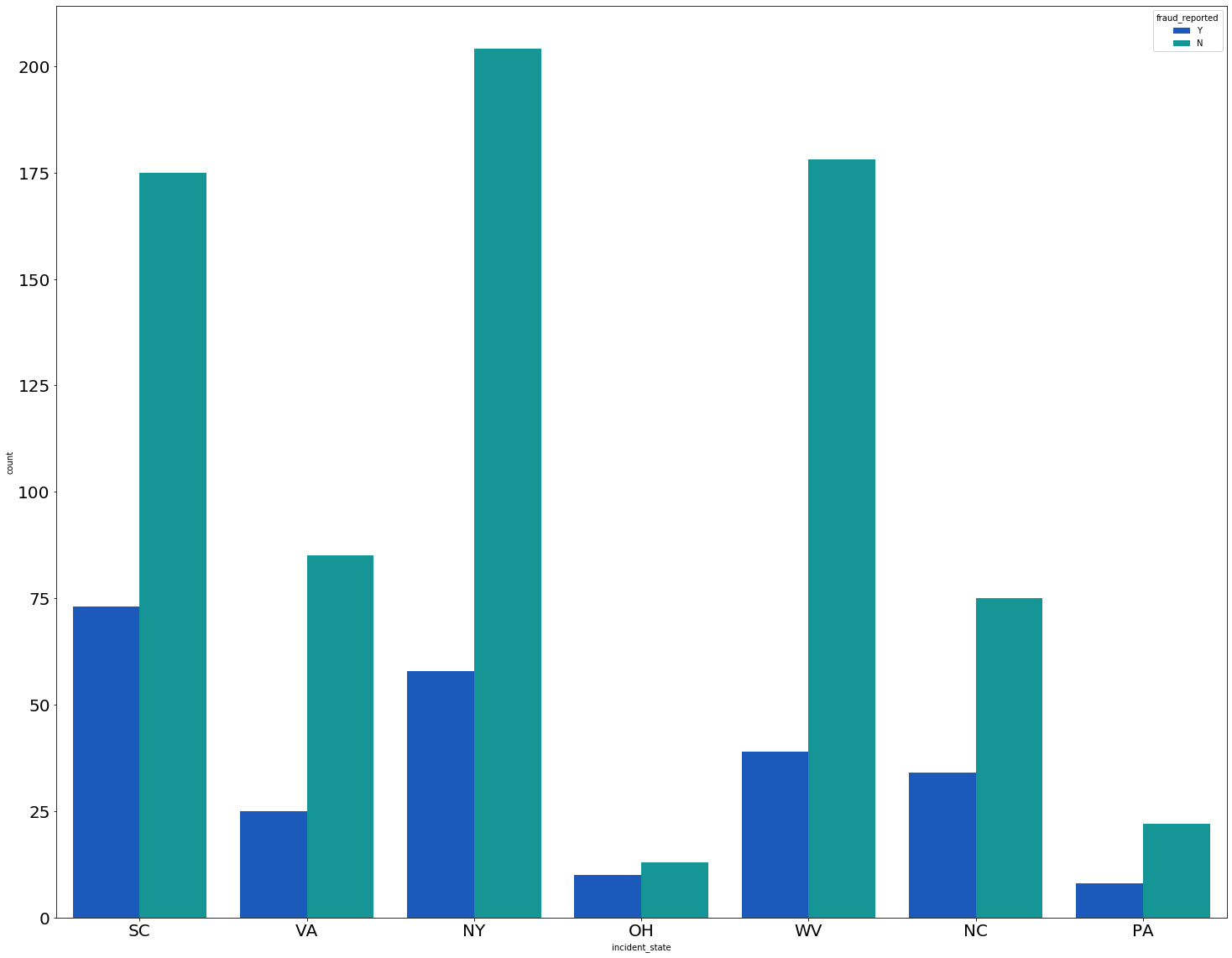
If we compare the education level with fraud reported, those have completed JD has higher chance of fraud claim.



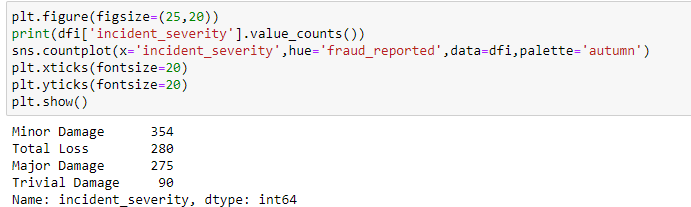


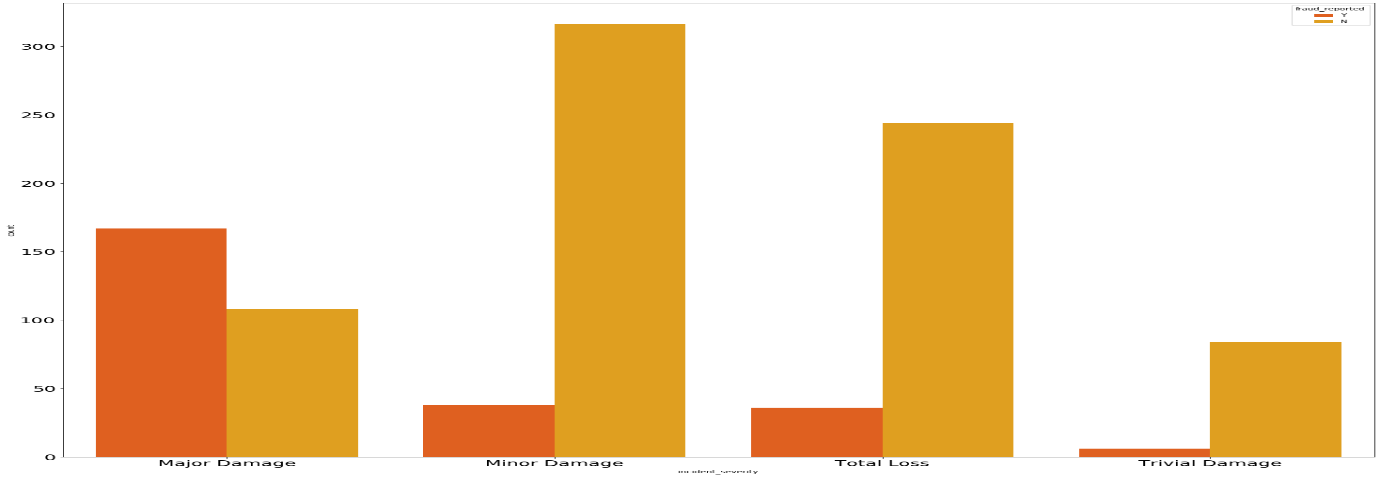
Majority of insured person occupation is Machine-on-inspect, when we checked the insured person occupation and fraud reported. The diagram shows the person who is in executive-managerial has high chance of fraud claim.



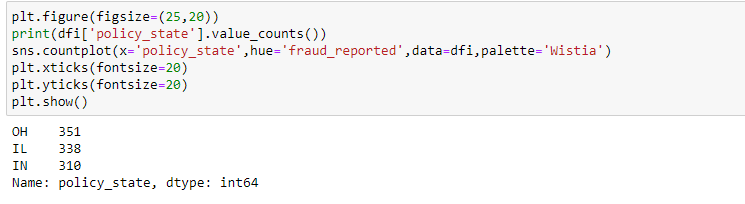


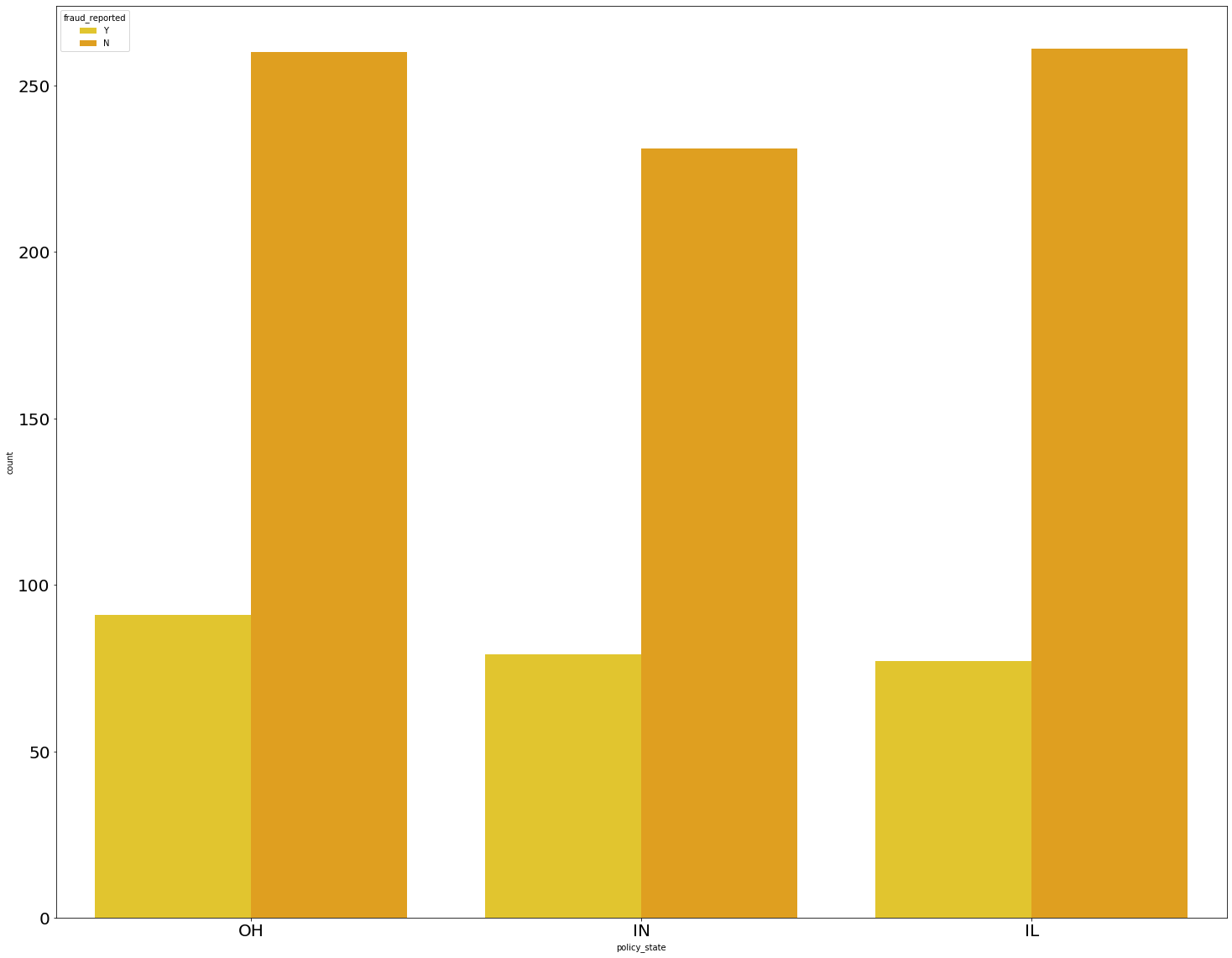
Majority of insured person belongs to NY state. The person who belongs to SC State there is high chance of fraud claim. The PA state have very less fraud claim, that is because there are only few insured people belongs to PA state.



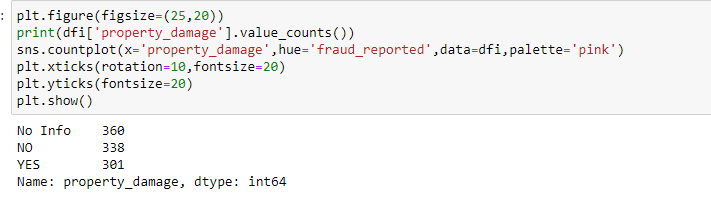


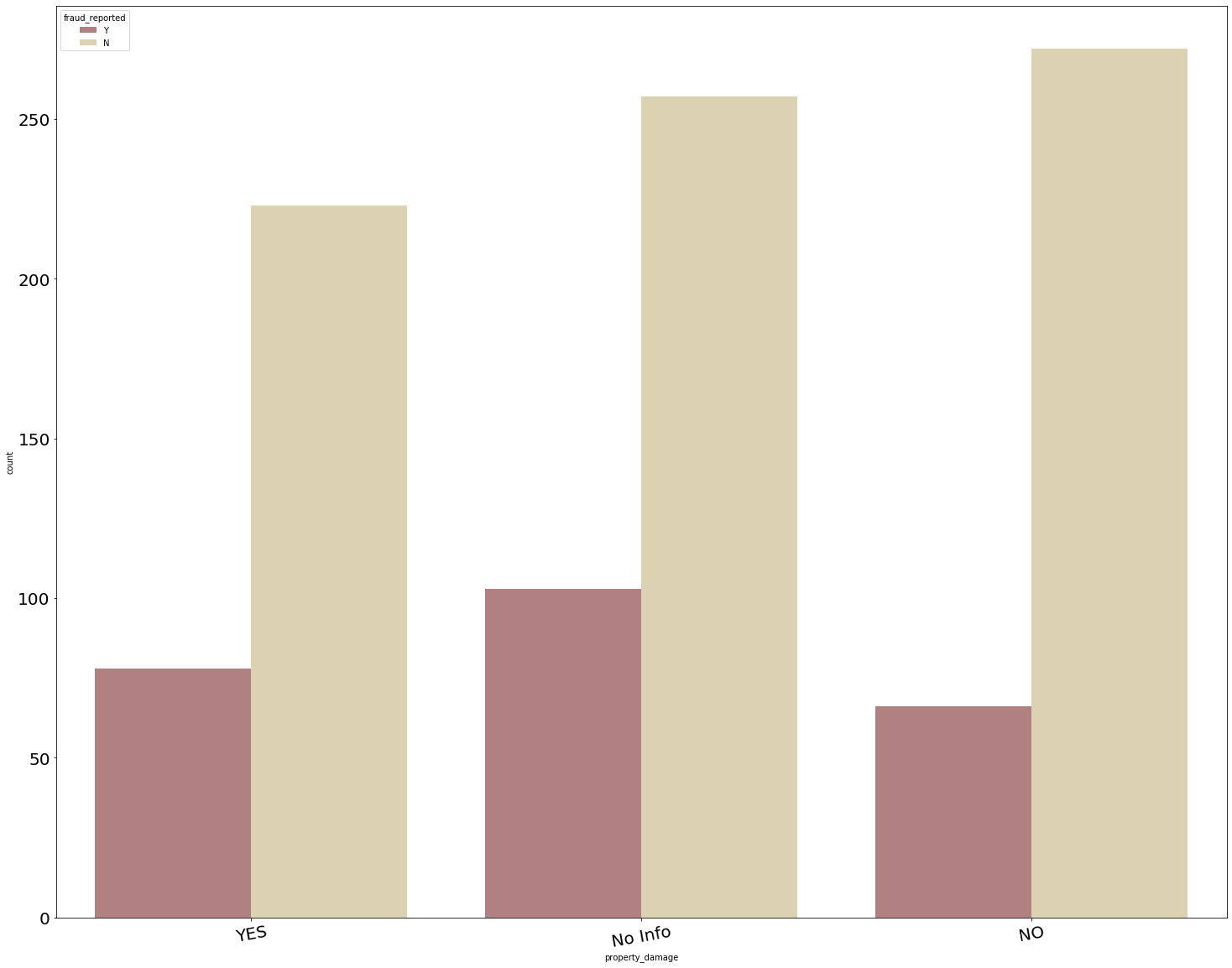
There is a high count in Major Damage and also there is high chance of fraud reported in the Major Damage category. If there is light damage happen to vehicle, they claim huge amount from insurance company for Major Damage. Sometimes they even damage their own vehicle and submitted the insurance claim form. For that purpose, we can build the Machine learning model to predict which is fraudulent claim or not.



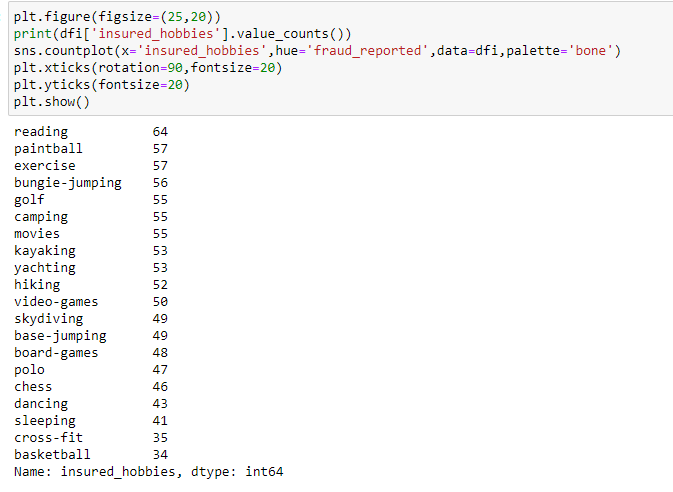


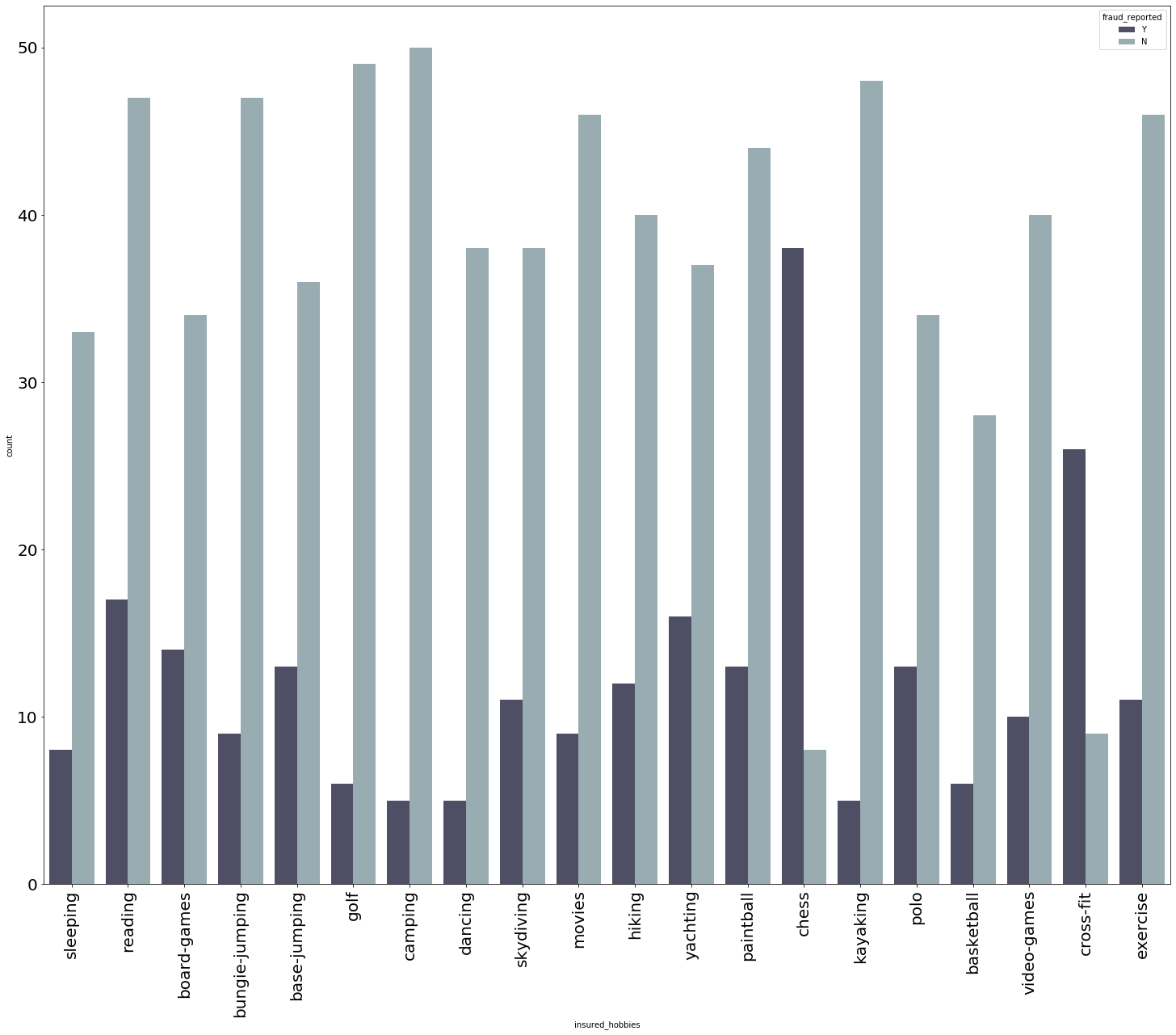
There is almost equal fraud reported in all the policy state. There is a little bit high fraud reported in OH policy state.



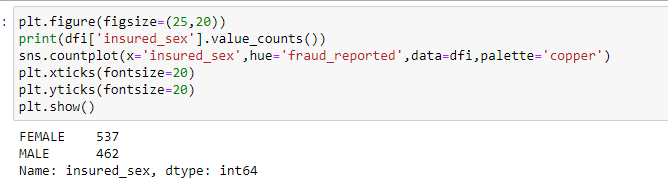


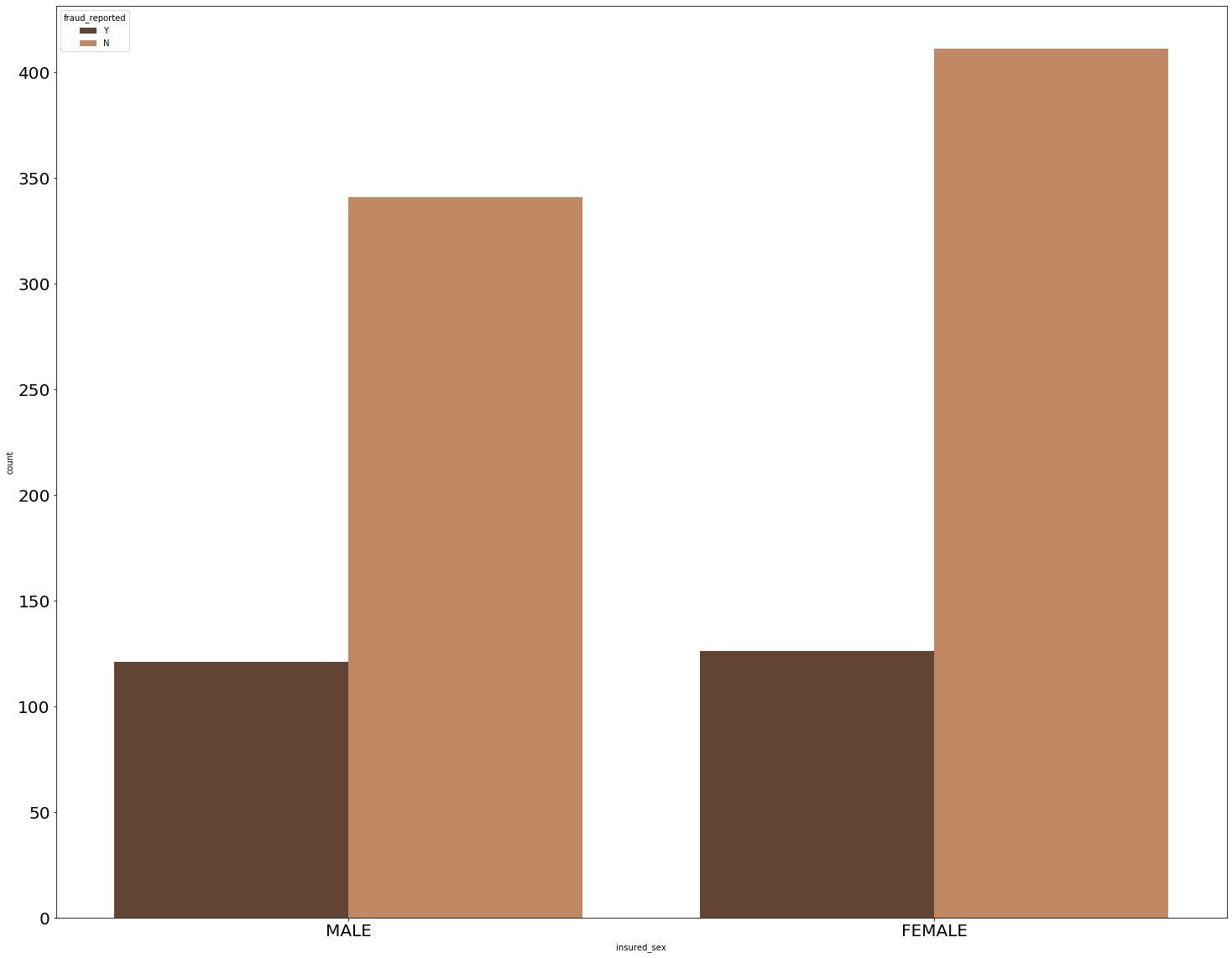
In the Property damage column, there are many rows that don’t have any information. 301 Insured persons has reported the property damage and 338 insured persons has not reported any property damage.





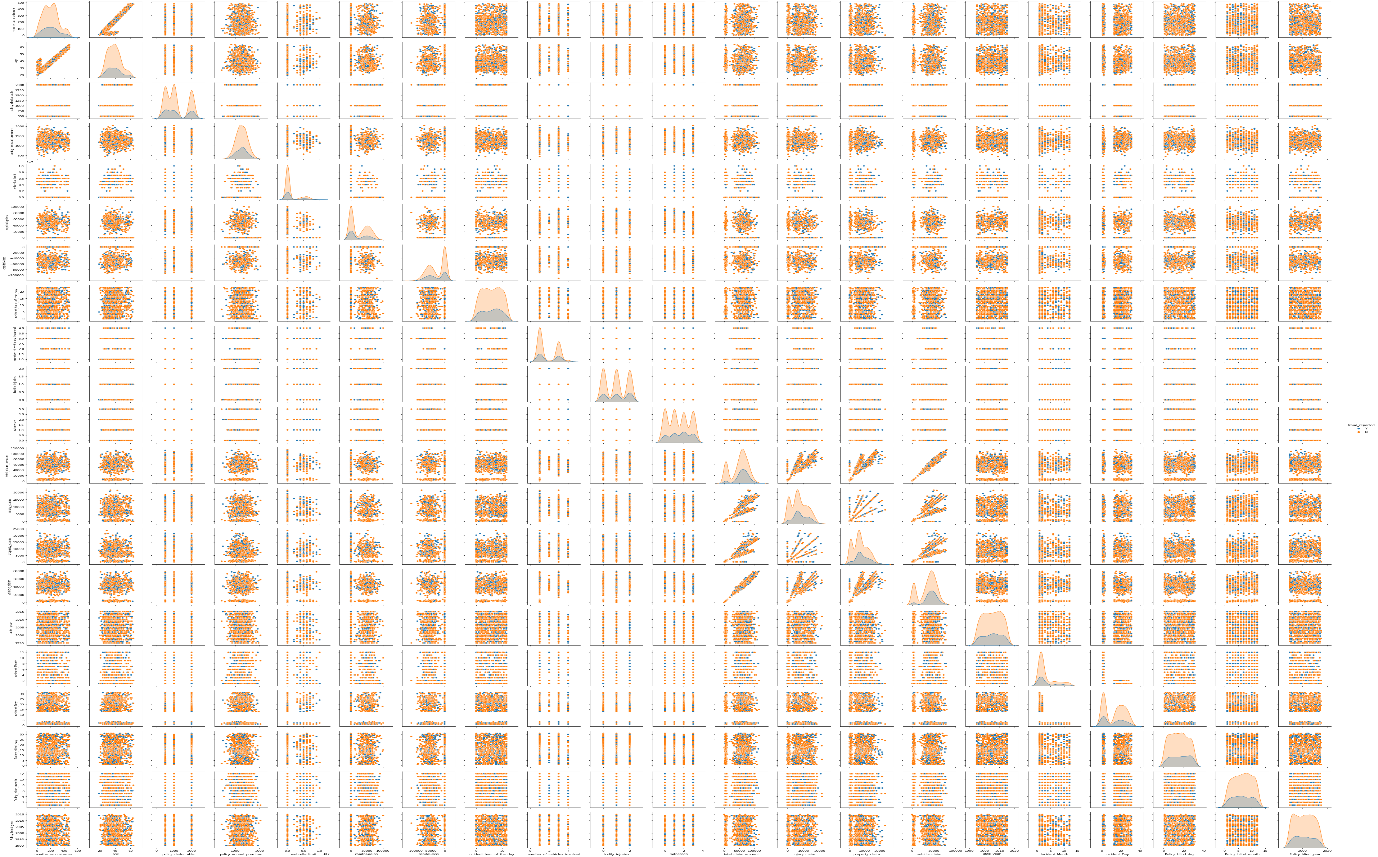
Let’s see what are the hobbies insured person have. Majority of the insured person have the hobbies of reading books. Then comes next Paintball and exercise. Only few insured persons are interested in playing basketball. If we checked the fraud reported with Hobbies. We see in the diagram, those who are interested in playing chess there is a high chance of fraud reported. Obviously, chess is mind game, we have to think each move carefully. So, the fraud claims also need to do carefully, only those people who strategically and logically strong they can do the fraud claim.



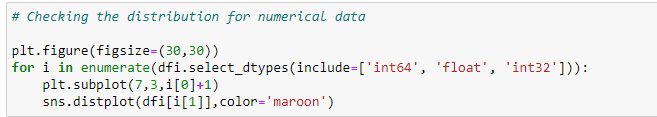


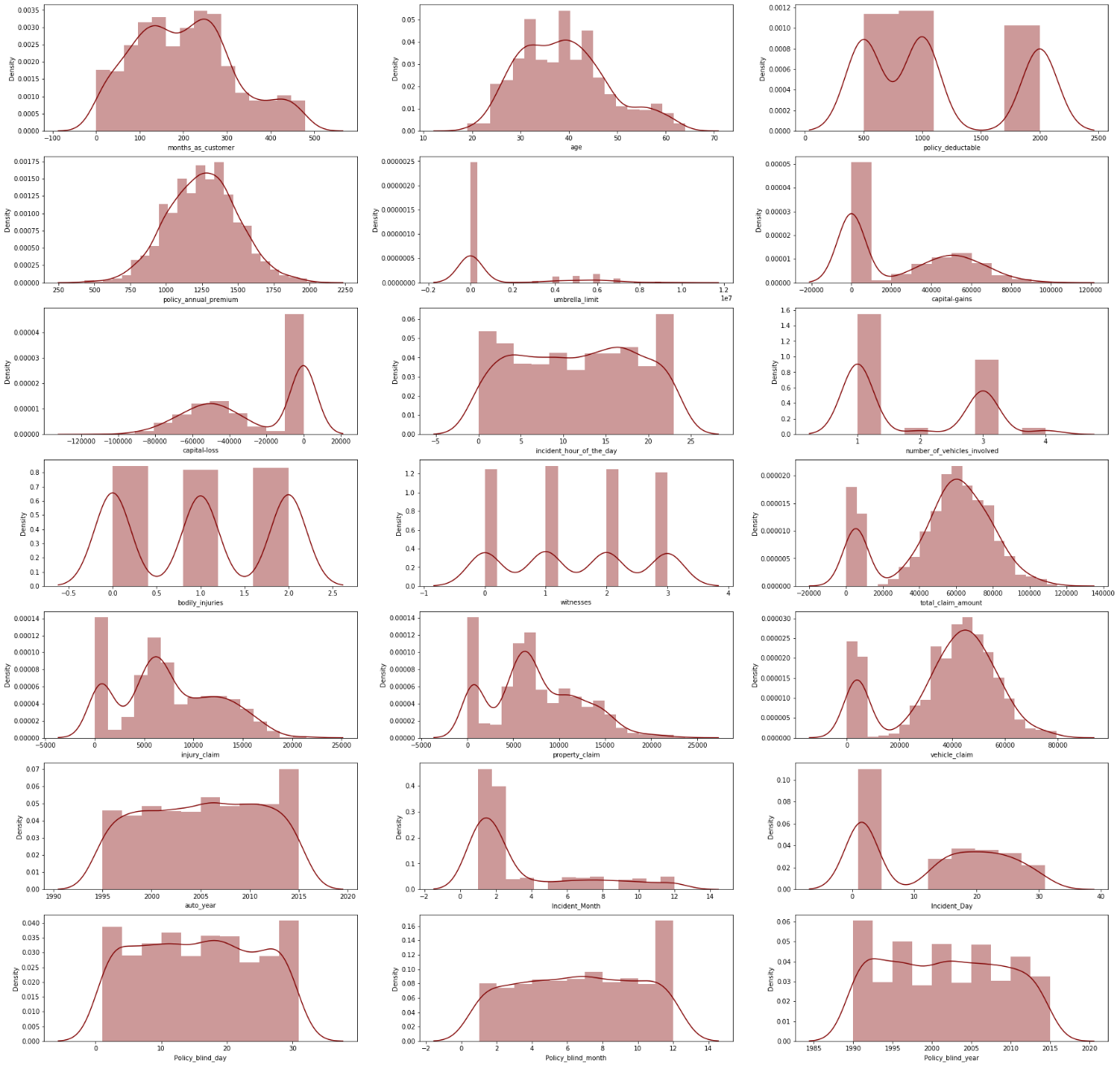
There is equal chance of fraud claim in both the gender.





In the pair plot we see each feature with fraud reported. In the diagram the diagonal it is normally distributed. The data are scatter.

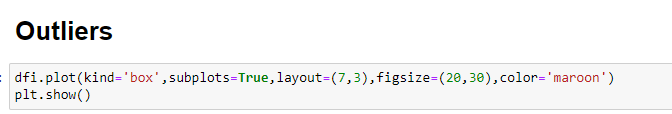


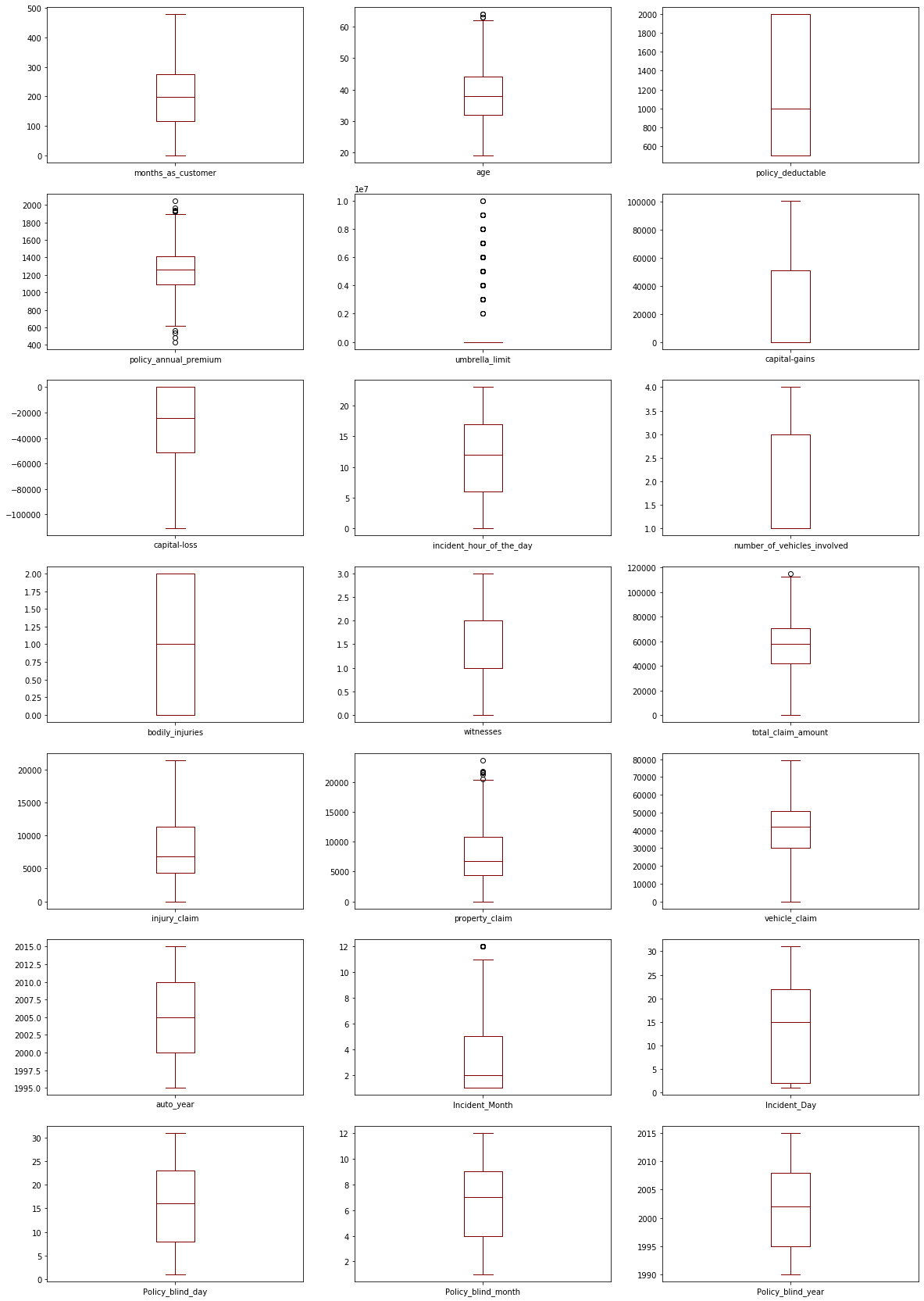


In the distribution plot we see that data are normally distributed and skewness also present. In insurance claim, property claim, injury claim, that data is slightly goes high and the diagram looks like a wave. We can remove the skewness to make the good accuracy.

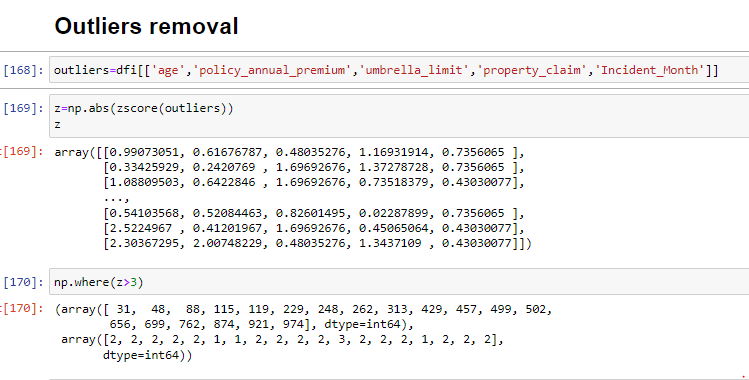
**3. EDA Concluding Remark**

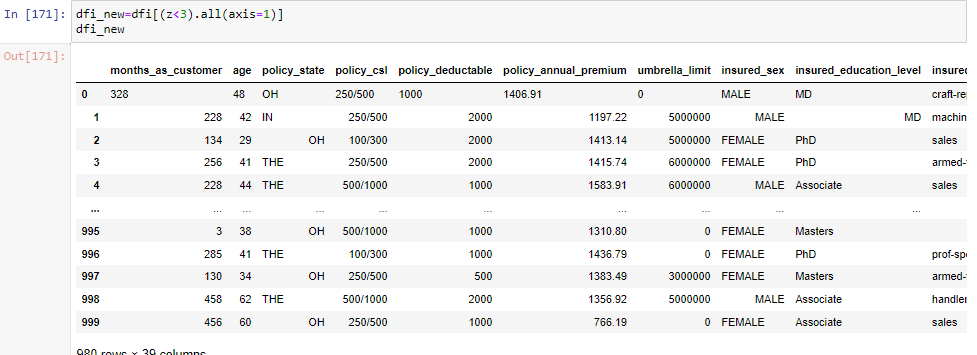
* Checked the null value using the heatmap, we found there is no null value present in data.
* We have converted the date datatype from object to datatime datatype and we have split into day, month and years.
* We have converted the categorical data to numerical data using the Label Encoder.
* We dropped the irrelevant columns from our data set.
* We checked the unique value, data info, shape of data, column name, statistical summary of data using describe method.
* We have checked the count of each feature and visualize them using bar chart, violin plot, count plot.
* We have visualized each feature with fraud reported and analysis in which category there is high chance of fraud reported.
* We have checked the correlation of data.
* In some column there is a ‘?’ present we have replaced it with ‘No info’
* Checked the normal distribution of feature to find whether the data contained skewness or not.
* Using the pair plot we have checked the relations of each feature with fraud reported.



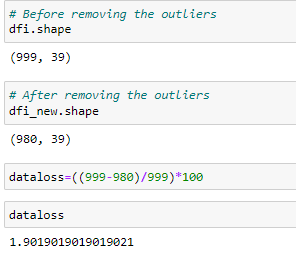


To check Outliers, I have used a boxplot. Outliers is present in age, policy\_annual\_premium, umbrella\_limit, property\_claim, incident\_month.





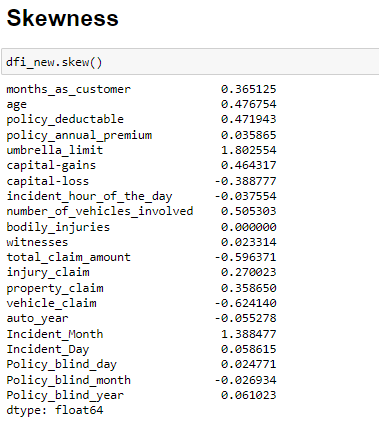
To remove Outliers, I have used a ZScore method. Outliers that are above the expected value. So, it is unnecessary to remove outliers from our dataset to give a good accuracy. After Outliers are removed, we make a data frame and store it in the new variable df\_new.



We checked both the data before removing the Outliers and after removing Outliers.

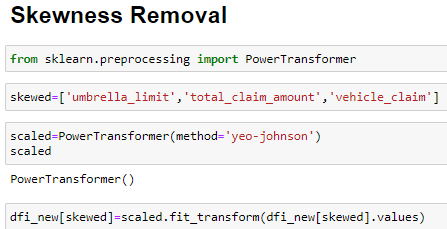
Before removing the Outliers, the shape is 999 rows and 39 columns. After removing the Outliers, the shape is 980 rows and 39 columns.

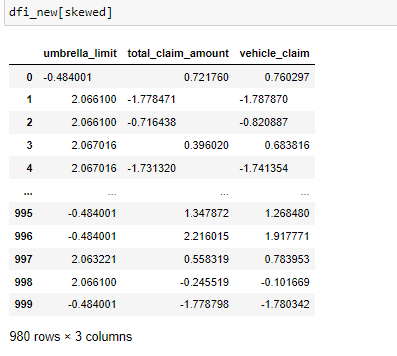
We checked the data loss how much data removed in the Outliers process. It has removed **1.9%** of data in Outliers.

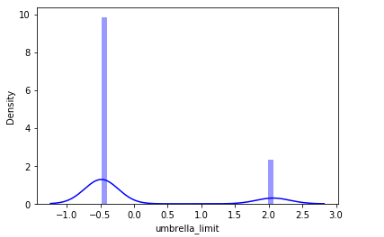


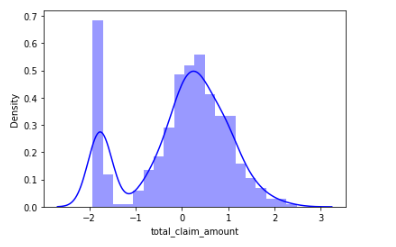
Skewness is the distortion of the curve of normal distribution. Skewness can be of three types, positive, negative or zero skewness. If the tail of the normal curve is on the right side, then it is positive skewness. If the tail of the normal curve is on the left side, then it is a negative skewness.

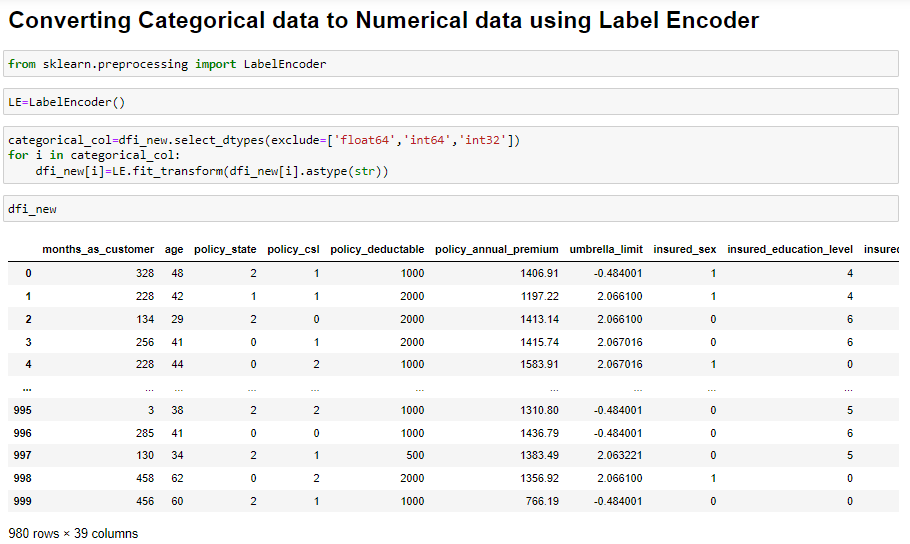
In our dataset the skewness is present in Umbrella limit, total\_claim\_amount, vehicle\_claim. We can remove the skewness using Power Transformer technique.



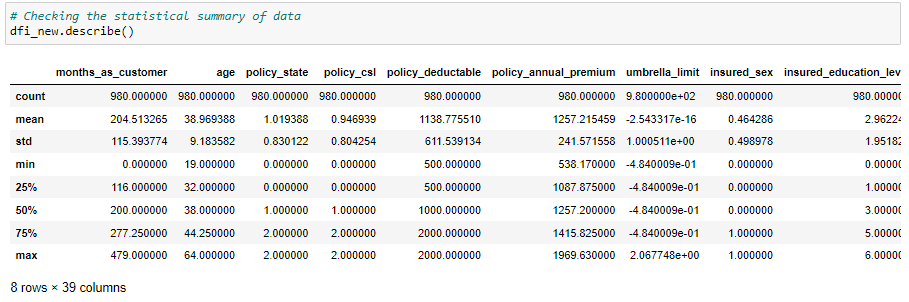


We have removed the skewness using **yeo-johnson** method and transform the skewed data to scaled data.

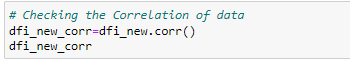
  
After removing the skewness, the normally distribution curve look like above.



We can’t make the model using a categorical data. So, we need to convert categorical data to numerical data. I have used **Label Encoder** to convert my categorical data to numerical data.

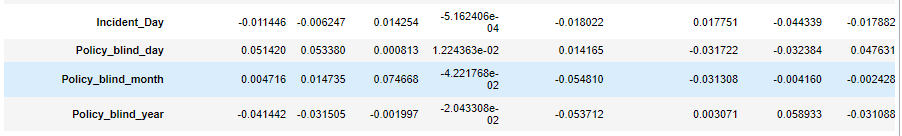


We have checked the Statistical summary of data after the conversion of Categorical data. Here is mean is greater than median in all the columns. Count is equal 980, so there is no missing value. In month\_as\_customer the minimum value is 0 and maximum value is 479. In most of the columns minimum value is 0. In some column there is negative value present. There is 1st quartile, 2nd quartile and 3rd quartile value are mentioned for all the columns and their Standard Deviation.



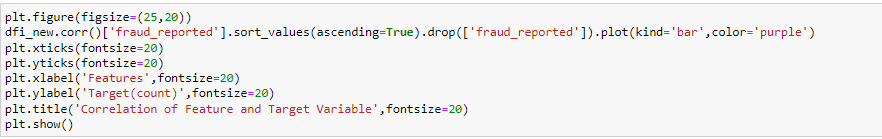


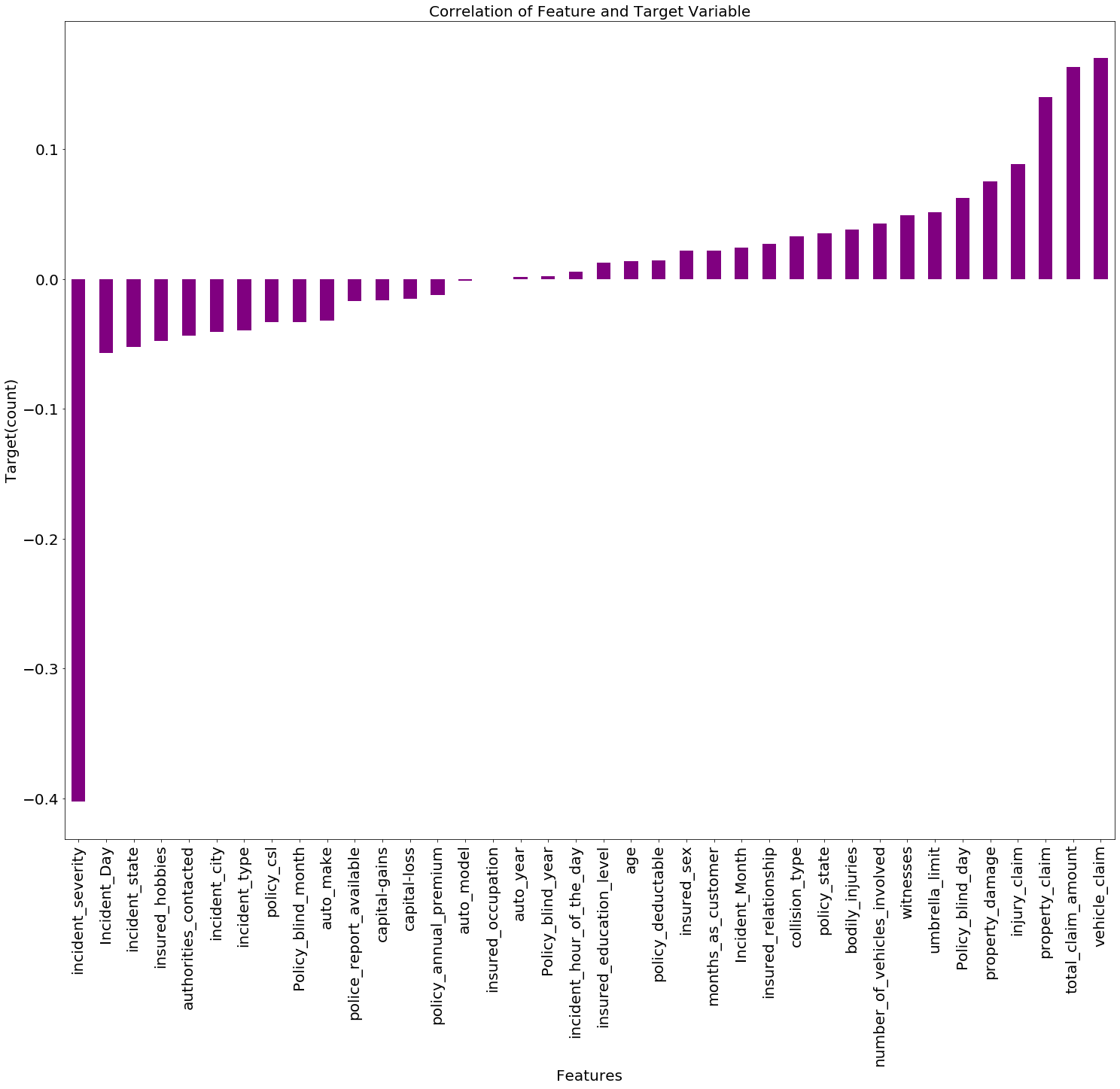




We checked the correlation for all the features after the conversion of data.

There is negative as well as positive correlation present.



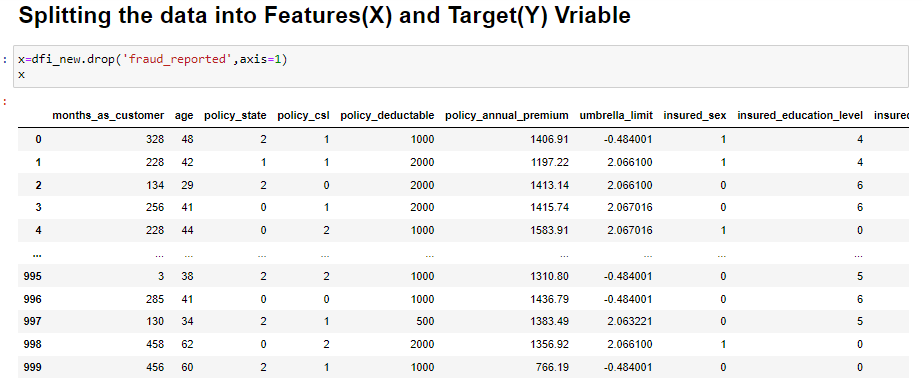


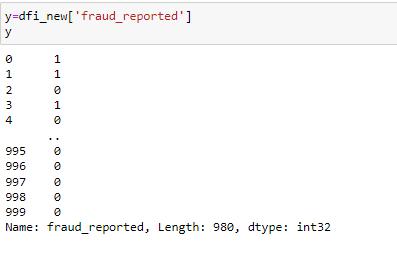
Visualizing the correlation of features and Target using the bar plot. Those who are positive correlated they are facing upward. Those who are negatively correlated, they are facing the downwards.

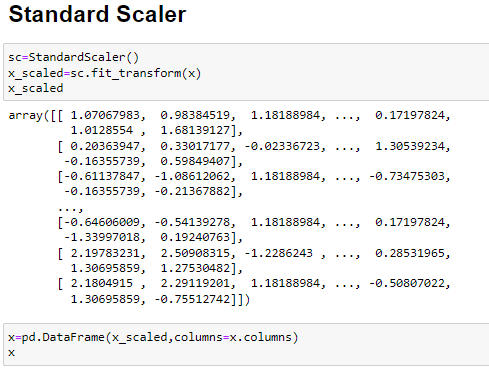
Property\_claim, total\_claim\_amount and vehicle claim are highly positive correlation. Incident\_severity is highly negative correlation.

We can split the data into X and Y variable for building the model.

**4. Pre-Processing Pipeline.**



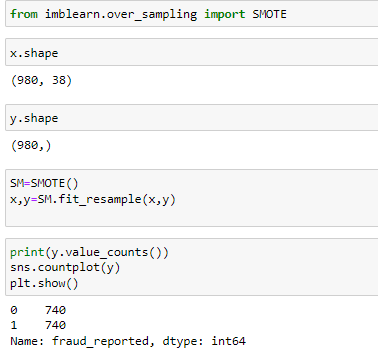




After data splitted into x and y, we can scaled the data. So, our data looks equal in size and store the data in a data frame.



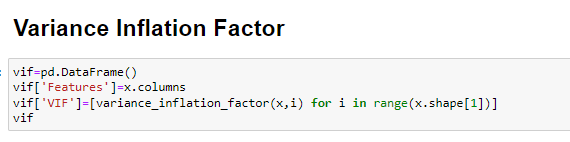
The target variable fraud\_reported is not equal. We have balanced the data. So, we have used the over sampling technique.

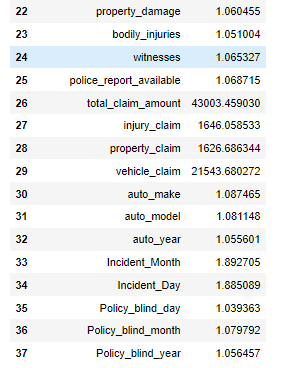


Now the data is balanced, our target value is equal 740.



We see the data is balanced using the count plot.



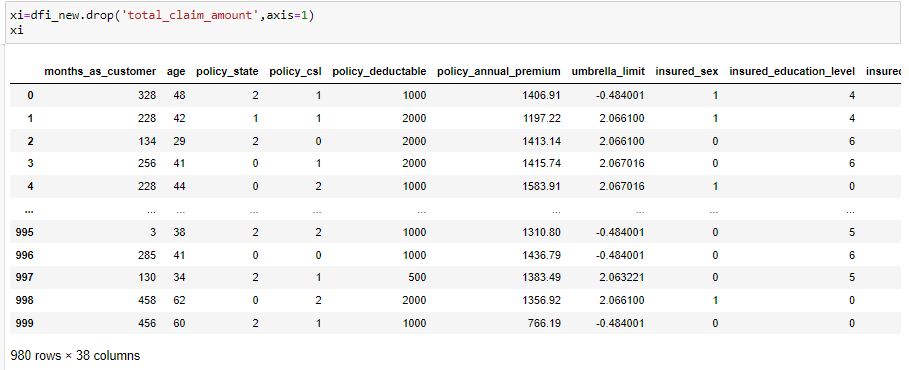


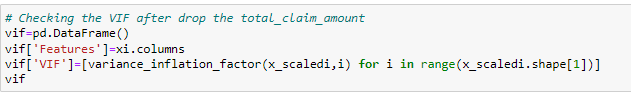
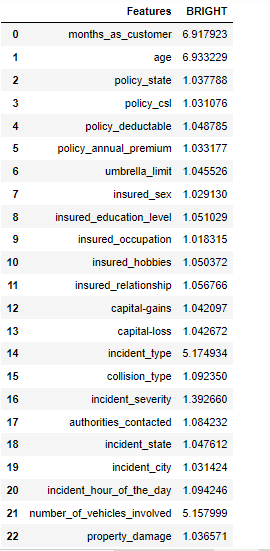
Variation Inflation Factor measures the severity of multi-collinearity. If there is high variation inflation present in data then there is result of collinearity.

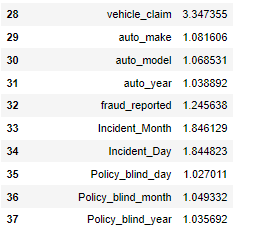
We have considered the high inflation factor which is more than 10 for our data.

High variance inflation is present in total\_claim\_amount, injury\_claim, property\_claim, vehicle\_claim.

So, we can drop the total\_claim\_amount, its VIF is 43003.459.



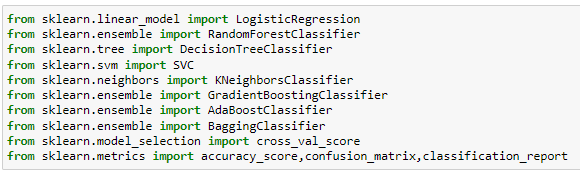


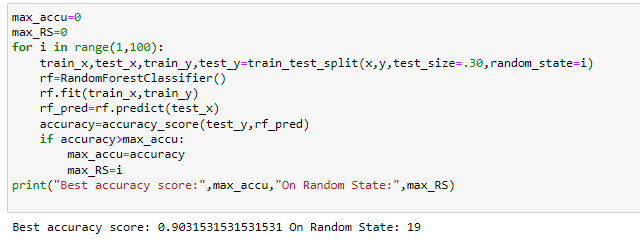


After dropping the total\_claim\_amount, we removed all the high inflation factor from our data set. There is no multi-collinearity exist.

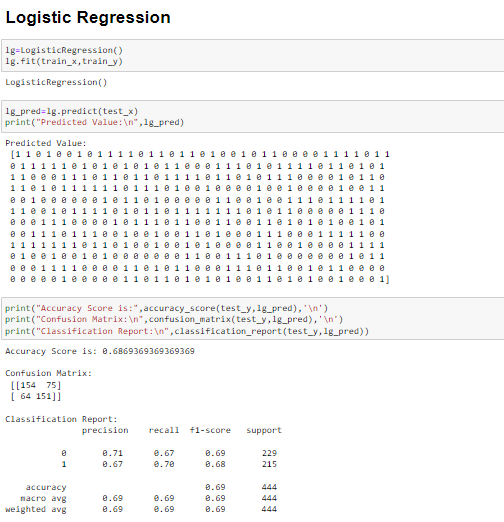
**5. Building Machine Learning Models**

Importing the necessary libraries for our model



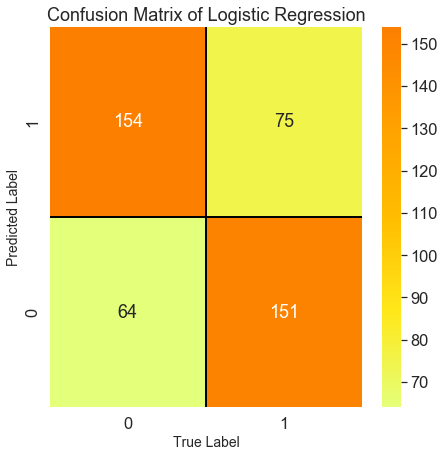


Checking the best random state for our model. Using Random Forest Classifier, the best accuracy score is 90% on random state is 19.

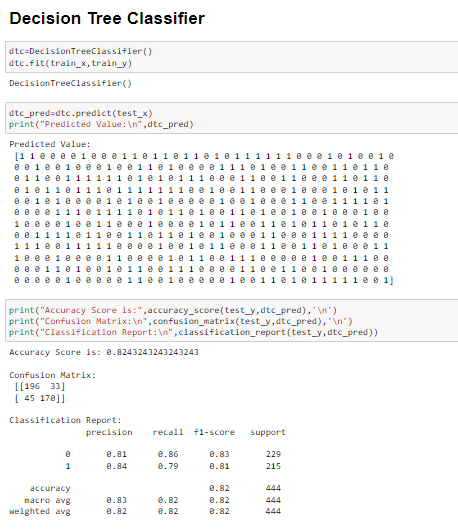


Using the Logistic Regression Model, our accuracy score is 68%. We can use different model to check accuracy score for our data and pick the one model which gives a good accuracy.



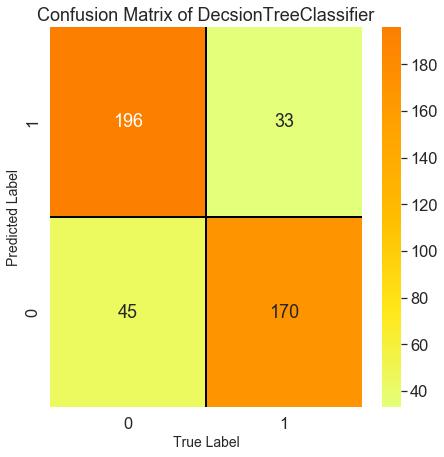


Visualizing the confusion matrix for Logistic Regression using the heatmap. We can observe the true positive rate, false positive rate, true negative rate, false negative rate. The predicted data is plotted against the actual value.

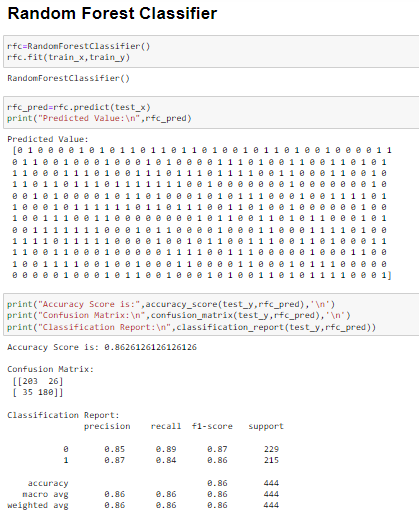


In Decision Tree model, our accuracy score is 82%.



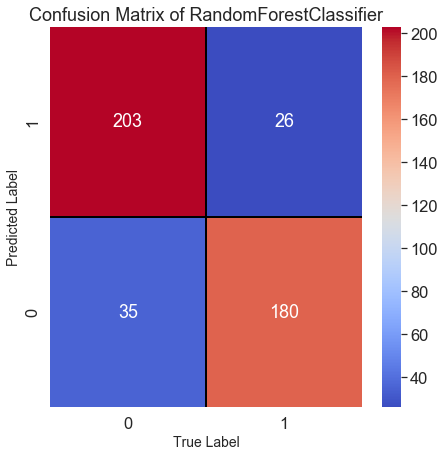


Visualizing the confusion matrix for Decision Tree Classifier using the heatmap. We can observe the true positive rate, false positive rate, true negative rate, false negative rate. The predicted data is plotted against the actual value.

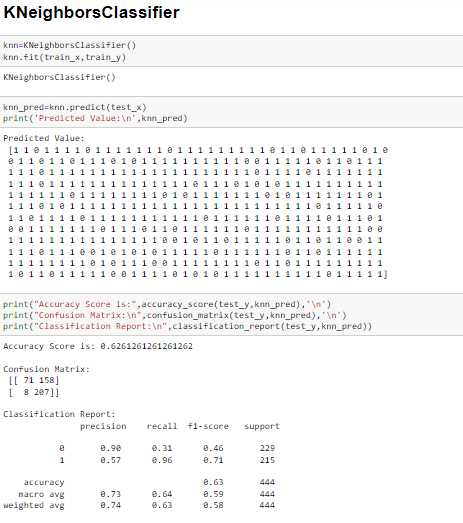


In Random Forest Classifier, our accuracy score is 86%. Which is a good score.



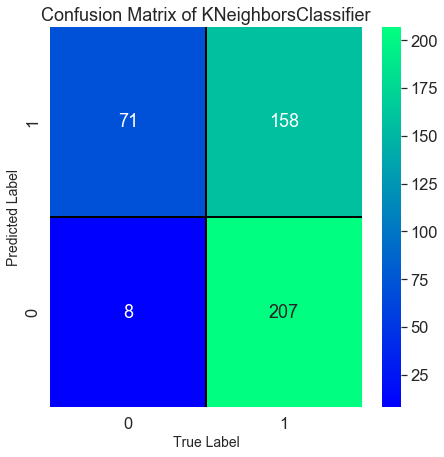


Visualizing the confusion matrix for Random Forest Classifier using the heatmap. We can observe the true positive rate, false positive rate, true negative rate, false negative rate. The predicted data is plotted against the actual value.

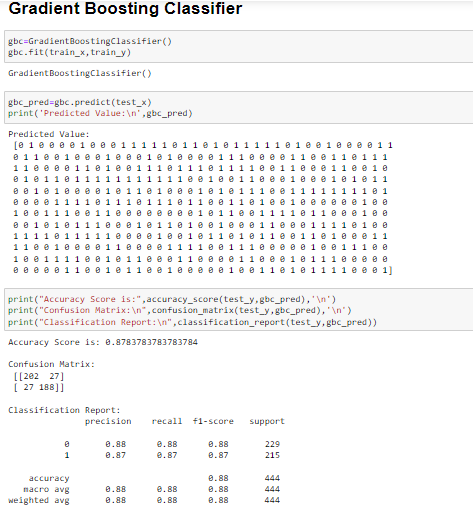


In KNeighbors Classifier, our accuracy score is 62%.



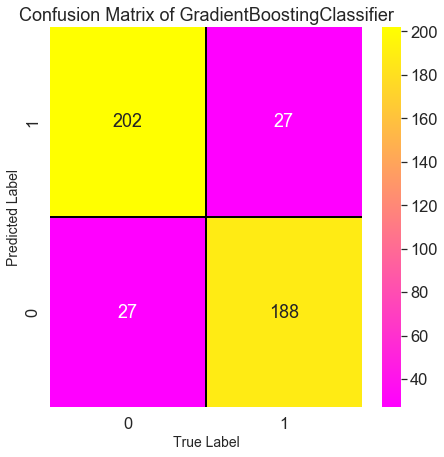


Visualizing the confusion matrix for KNeighbors Classifier using the heatmap. We can observe the true positive rate, false positive rate, true negative rate, false negative rate. The predicted data is plotted against the actual value.

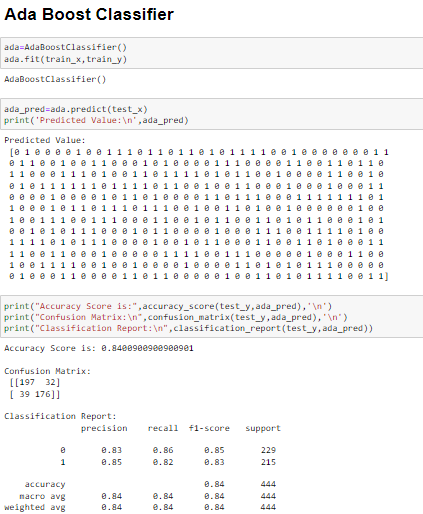


In Gradient Boosting Classifier, our accuracy score is 87%.



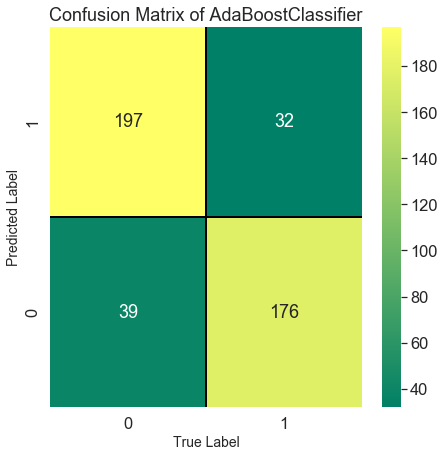


Visualizing the confusion matrix for Gradient Boosting Classifier using the heatmap. We can observe the true positive rate, false positive rate, true negative rate, false negative rate. The predicted data is plotted against the actual value.



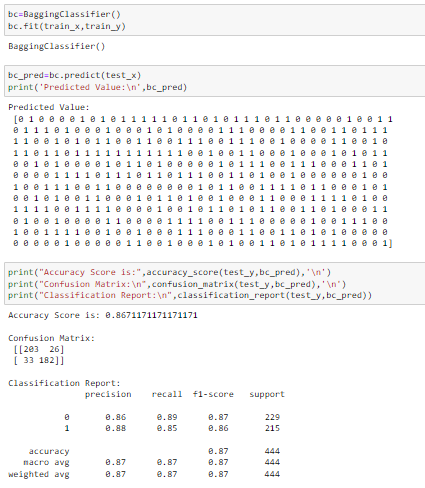
In Ada Boost Classifier, our accuracy score is 84%





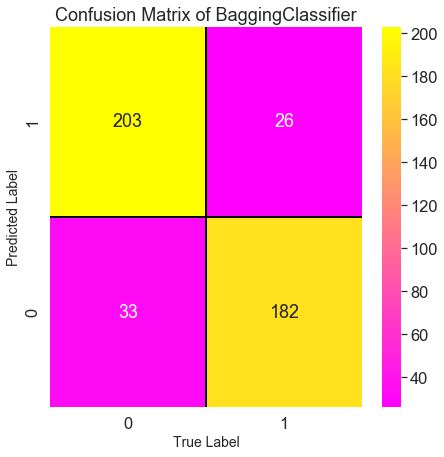
Visualizing the confusion matrix for Ada Boost Classifier using the heatmap. We can observe the true positive rate, false positive rate, true negative rate, false negative rate. The predicted data is plotted against the actual value.

**Bagging Classifier:**

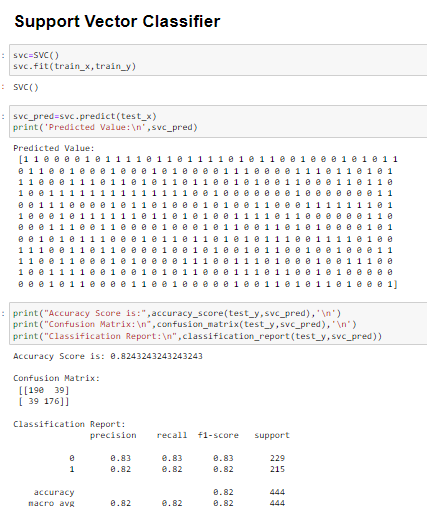


In Bagging Classifier, our accuracy score is 86%



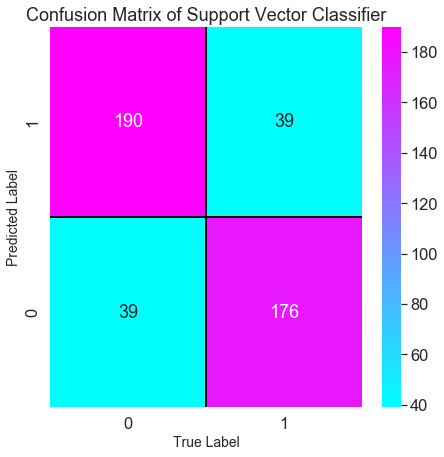


Visualizing the confusion matrix for Bagging Classifier using the heatmap. We can observe the true positive rate, false positive rate, true negative rate, false negative rate. The predicted data is plotted against the actual value.

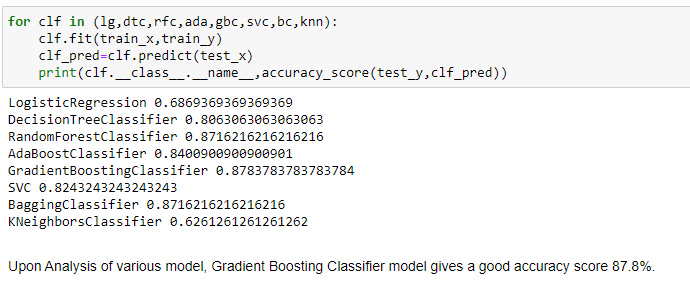


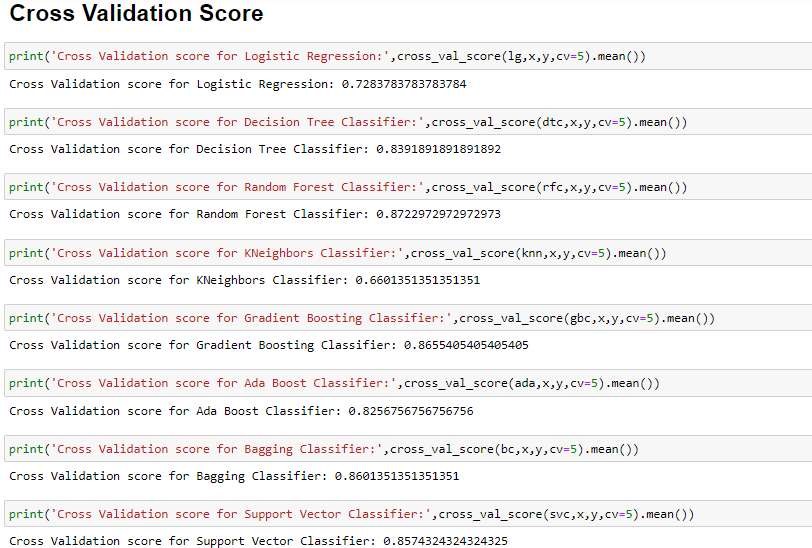
In Support Vector Classifier, our accuracy score is 82%



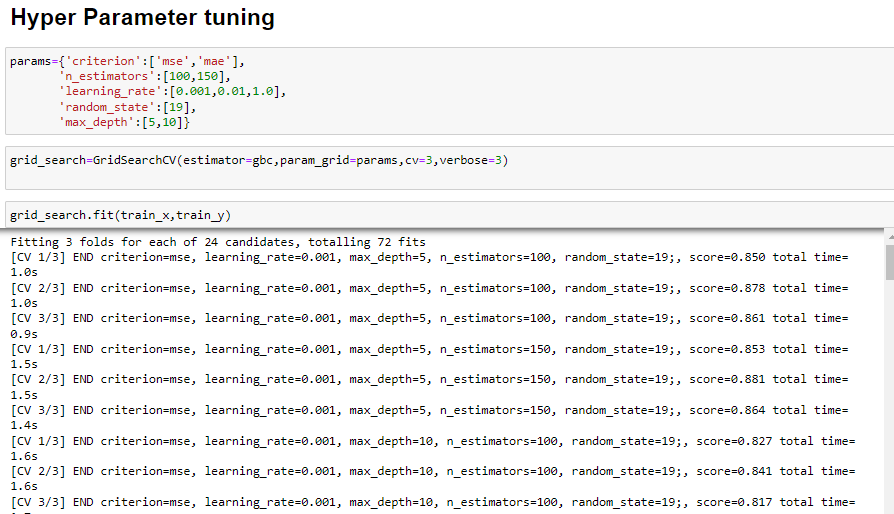


Visualizing the confusion matrix for Support Vector Classifier using the heatmap. We can observe the true positive rate, false positive rate, true negative rate, false negative rate. The predicted data is plotted against the actual value.

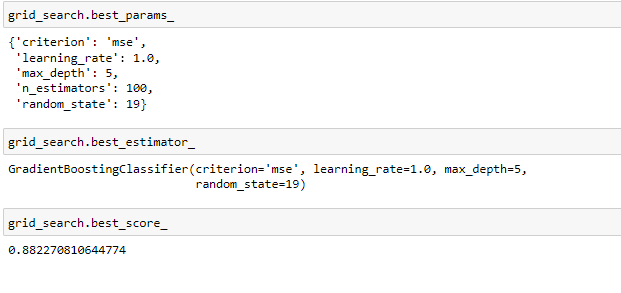




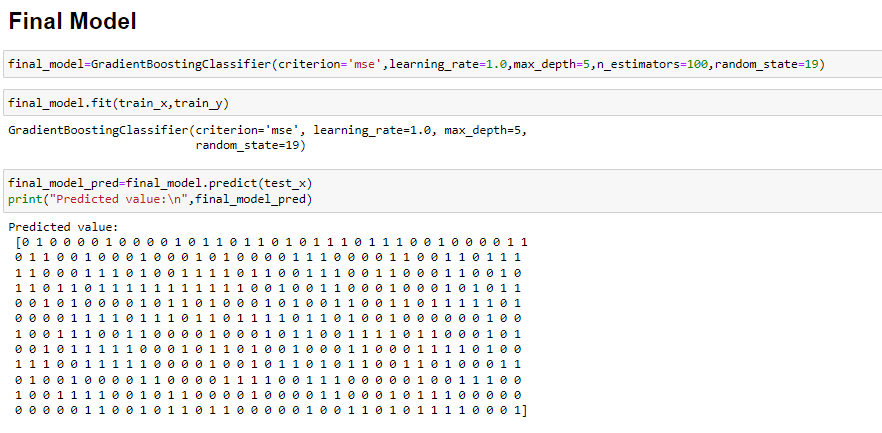
We checked cross validation score of various models, Gradient Boosting Classifier gives good accuracy score and cross validation score. So, we will consider Gradient Boosting Classifier as our final model. Let’s do the Hyper parameter tuning to check if we can increase the accuracy of our model.

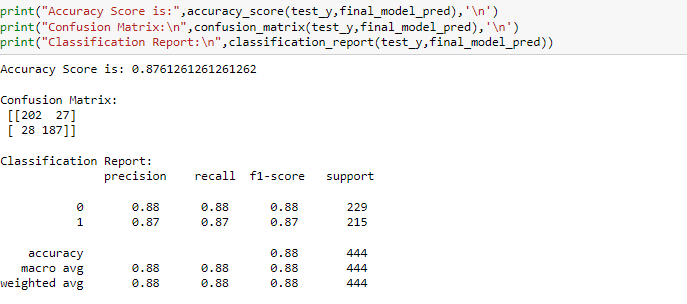


This are the best parameters and best estimators that are present in Gradient Boosting Classifier



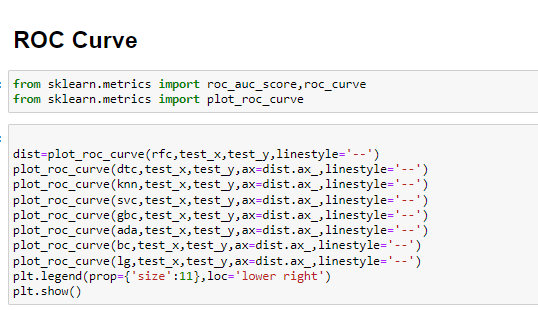
We have done the hyper parameter tuning using grid search CV and passed some parameter and we trained the model. In Grid Search CV the best score is 88%. We will pass this estimator in our final model to check the accuracy.

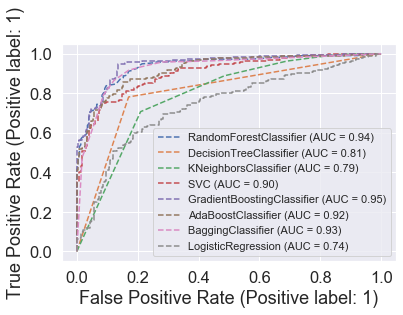




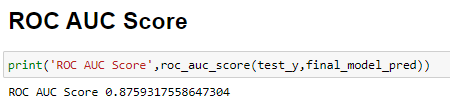
After tuning the parameter in our final model, the accuracy score is 87.6%. Which is a good score.

ROC Curve shows the relationship between sensitivity and specificity for every possible cut off.

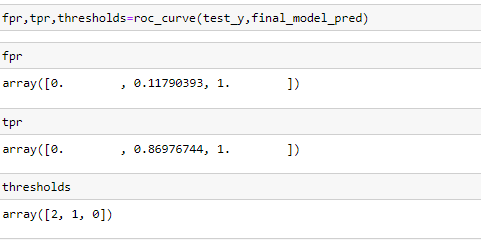




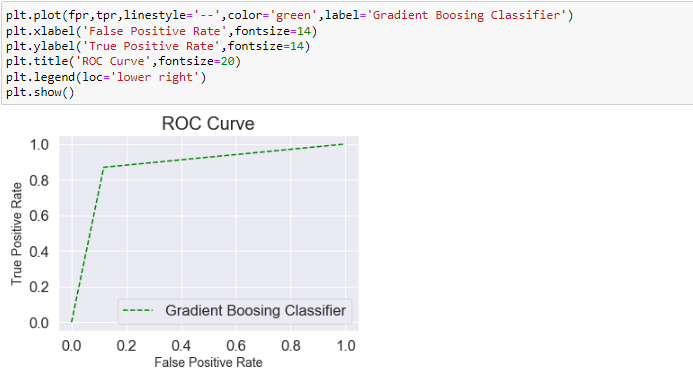
In the above roc\_curve, we see that Gradient Boosting Classifier gives AUC Score 95%



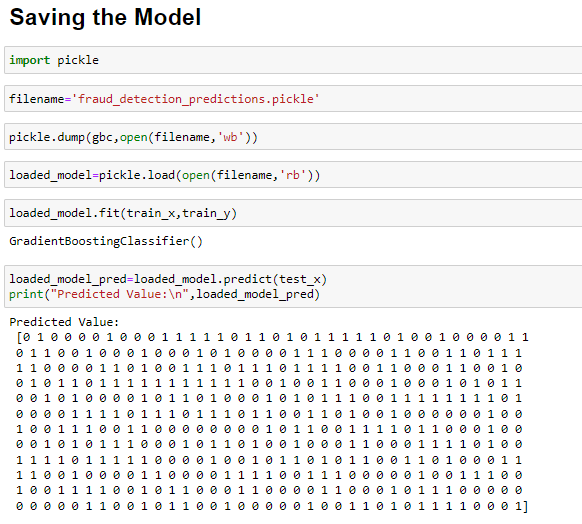
The ROC AUC Score for our final model is 87%.



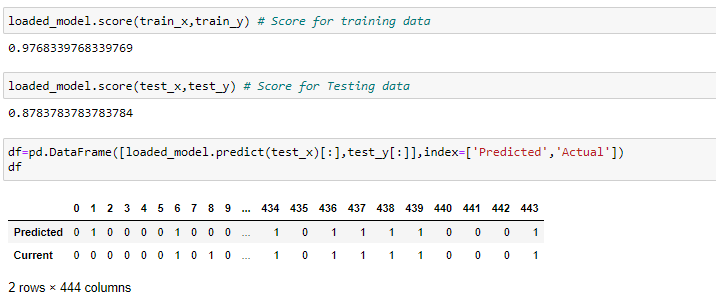
This is our false positive rate, true positive rate and thresholds of our model.



We have plotted the ROC Curve of our final model Gradient Boosting Classifier.

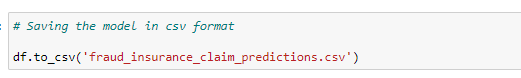


We have saved the model using Pickle library. This is the predicted value of our final model.



The score for train data is 97.6% and score for test data is 87.8. We make a data frame and store the predicted and actual value. Predicted and actual value almost look same.

We then saved the predicted and actual value in csv format.



**6. Concluding Remarks**

In this Insurance Claim – Fraud Detection project we have analysis the data using various plot in Visualization. We have gone through Feature Engineering, Pre-processing the data. We have handled the imbalance data. We have converted the categorical data to numerical data using a Label Encoder.

We have checked the distribution of data and removed the skewness present. We also checked the Outliers and removed it using the zscore method. We have checked the correlation of data and the variation inflation factor. Then we have scaled the data.

We have splitted the data for training and testing. We have given 70% of data for training and 30% of data for testing.

We have dropped the column which is having the high variation inflation factor to avoid multi-collinearity issue. We have built various classification model and checked their accuracy score, confusion matrix and classification report.

We have checked the cross-validation score of each model and compare with other models. Picked the best model which gives a good score. In our project Gradient Boosting Classifier gives a good score. We have done the hyper parameter tuning to increase accuracy. In our scenario the accuracy score didn’t increase.

We have seen that predicted and actual value almost similar; this means our model is working well. Based on this prediction we can predict which insurance claim is fraudulent or not.

The insurance company can check the insured person features and predict the results whether to accept the insurance claim or reject it. This will help to decide and avoid the major loss to the company. Machine Learning model plays very important role in predicting the fraudulent claim.