**Insurance Claims- Fraud Detection**

**Problem Definition:-**

What is an insurance claim? An insurance claim is a formal request to your insurance provider for reimbursement against losses covered under your insurance policy.

The major problem faced by this industry-

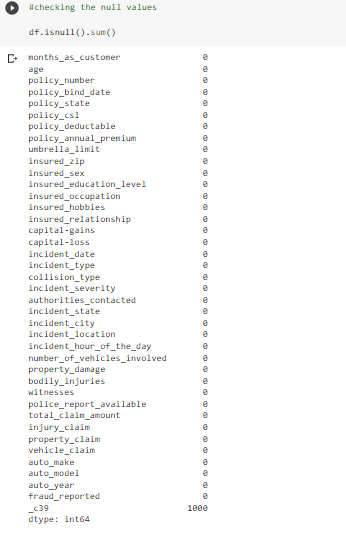
Insurance fraud is a huge problem in the industry. It's difficult to identify fraud claims.

Insurance fraud is any act committed to [defraud](https://en.wikipedia.org/wiki/Fraud) an insurance process. It occurs when a claimant attempts to obtain some benefit or advantage they are not entitled to, or when an insurer knowingly denies some benefit that is due. According to the United States [Federal Bureau of Investigation](https://en.wikipedia.org/wiki/Federal_Bureau_of_Investigation), the most common schemes include premium diversion, fee churning, asset diversion, and workers compensation fraud. Machine Learning is in a unique position to help the Auto Insurance industry with this problem.

In this project, we 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.

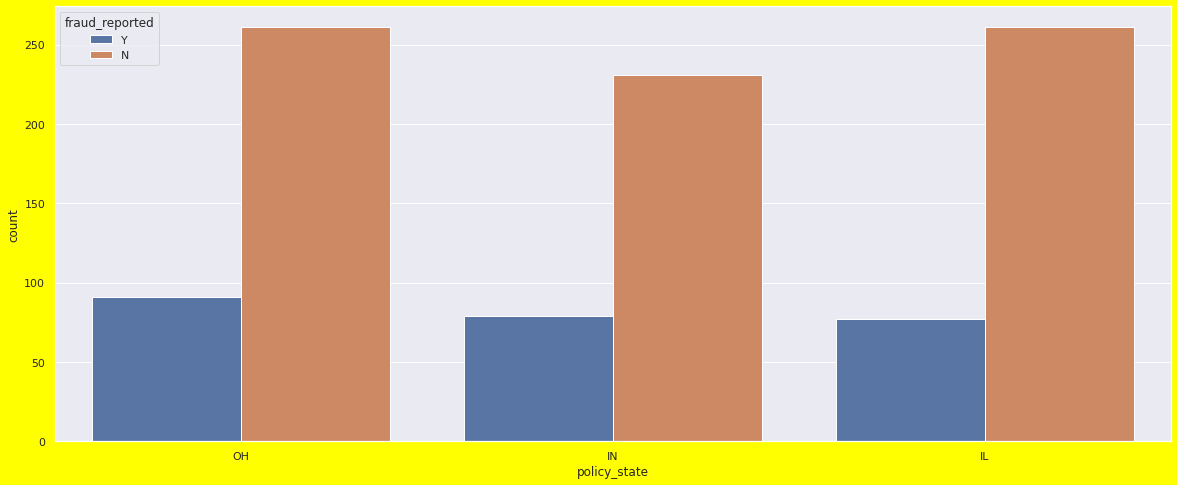
**Data Analysis:-**

The dataset contains 1000 rows and 40 columns.

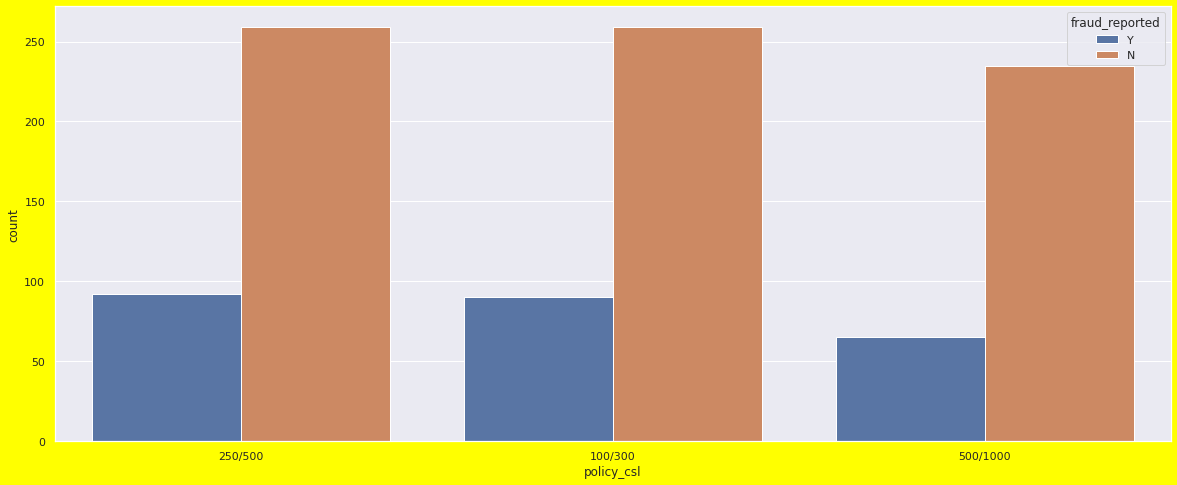
Columns are as follows(the columns are self explanatory)

As per the above code , we come to a conclusion that there are no null values in the dataset except \_c39 which has all the null values so we will drop that column as it has not relation with the problem statement.

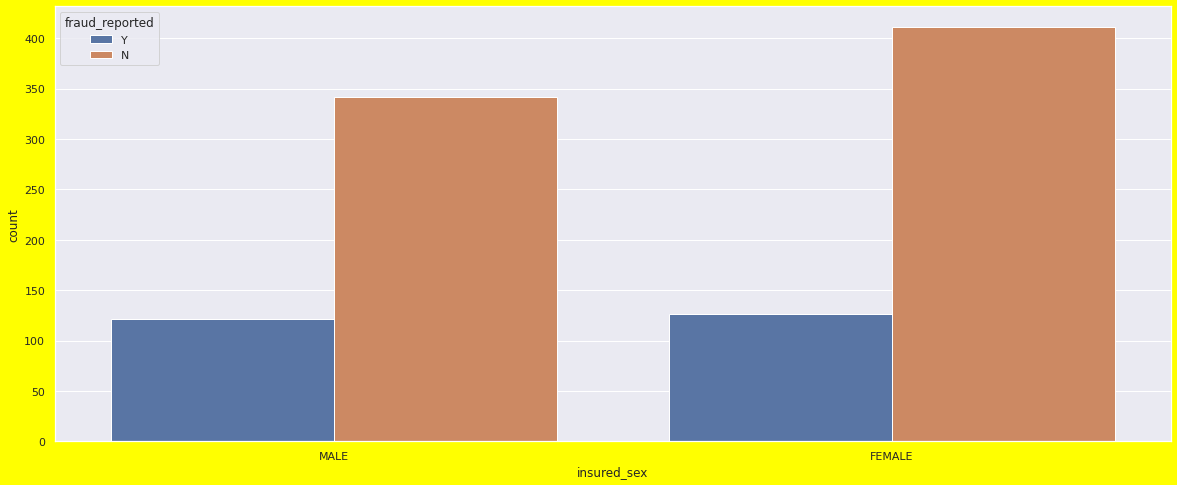
**EDA(EXPOLATORY DATA ANALYSIS):-**

CATEGORICAL DATA ANALYSIS WITH TARGET VARIABLE

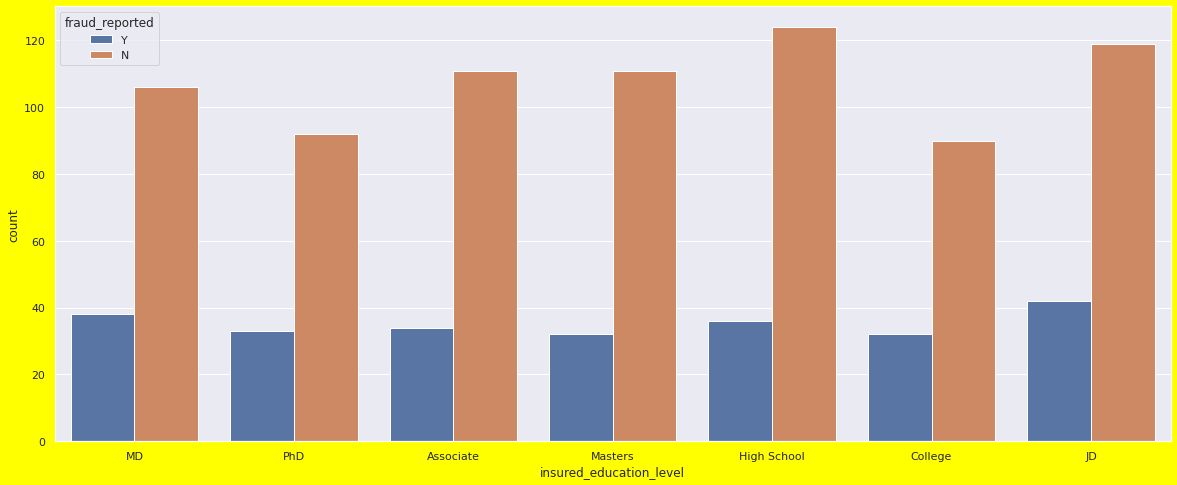
Policy state is place where policy has been taken. All the states are showing some what same result on the fraud reported. This feature has no effect on the target variable.



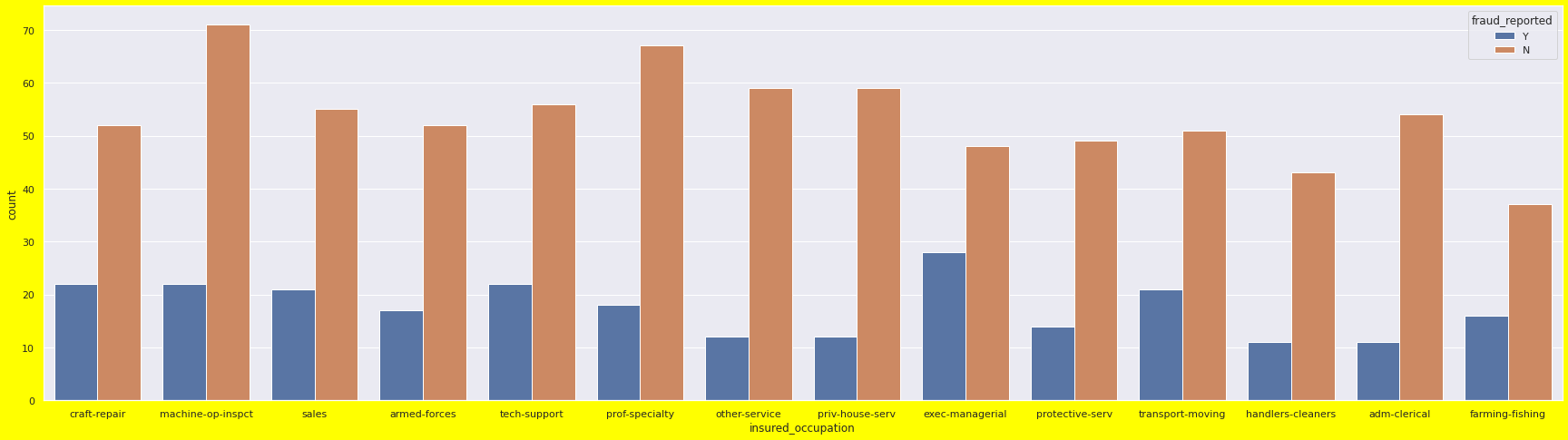
**Policy\_csl :Combined single limits** are a provision of an insurance policy that limits the coverage for all components of a claim to a single dollar amount. A combined single limit policy has a maximum dollar amount that covers any combination of injuries or property damage in an incident.

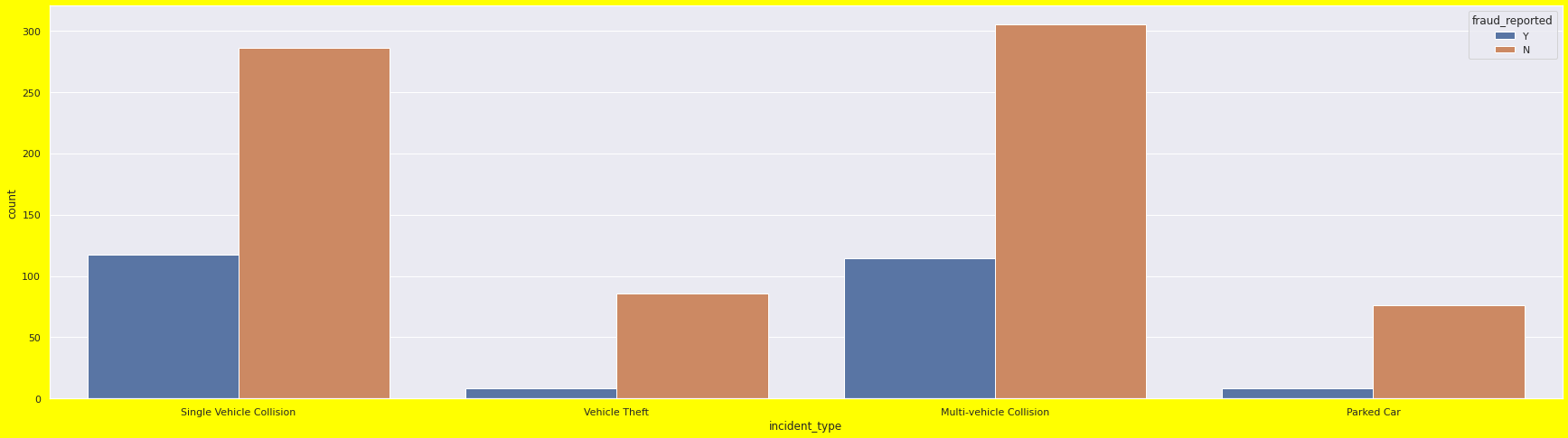


The fraud reported is almost same in all the cases.

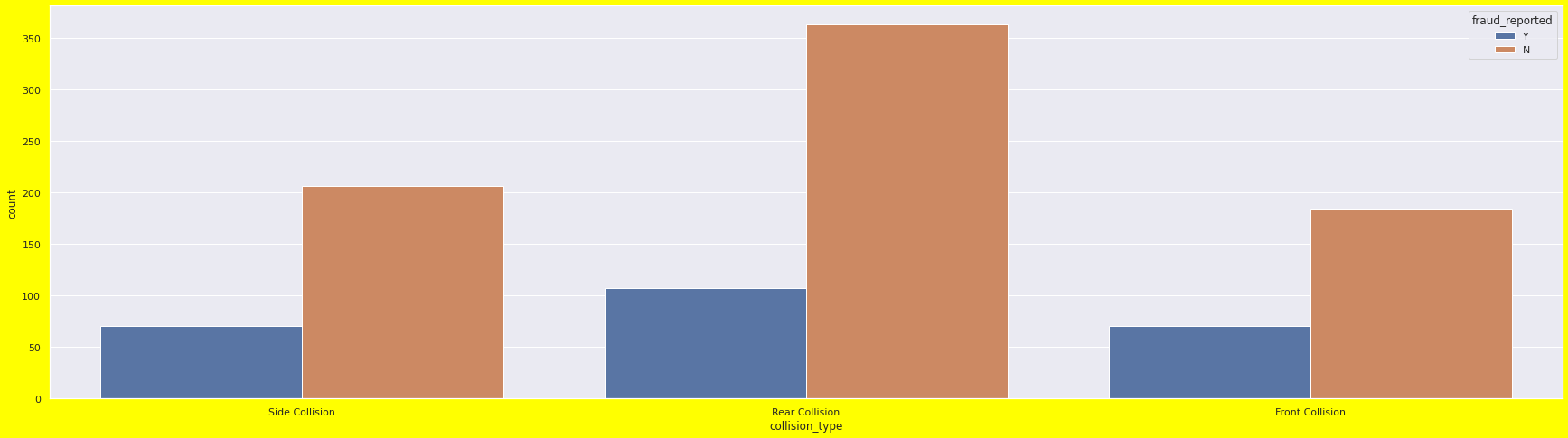


The above graph shows that there are more females in the insured than males. Fraud reported in females is 32% whereas in males it is 34%.

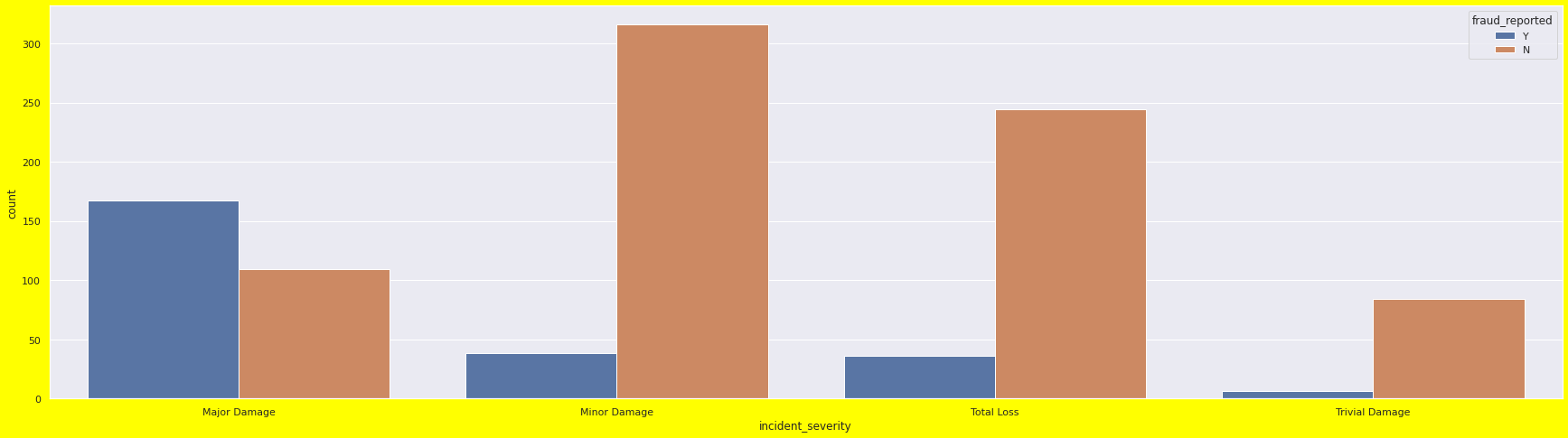
The above graph shows the education level of the insured people and the fraud reported. Most of the frauds are reported in JD education level. JD means doctor of law. An inference can be drawn that people who are into law know how to manipulate it to their advantage.



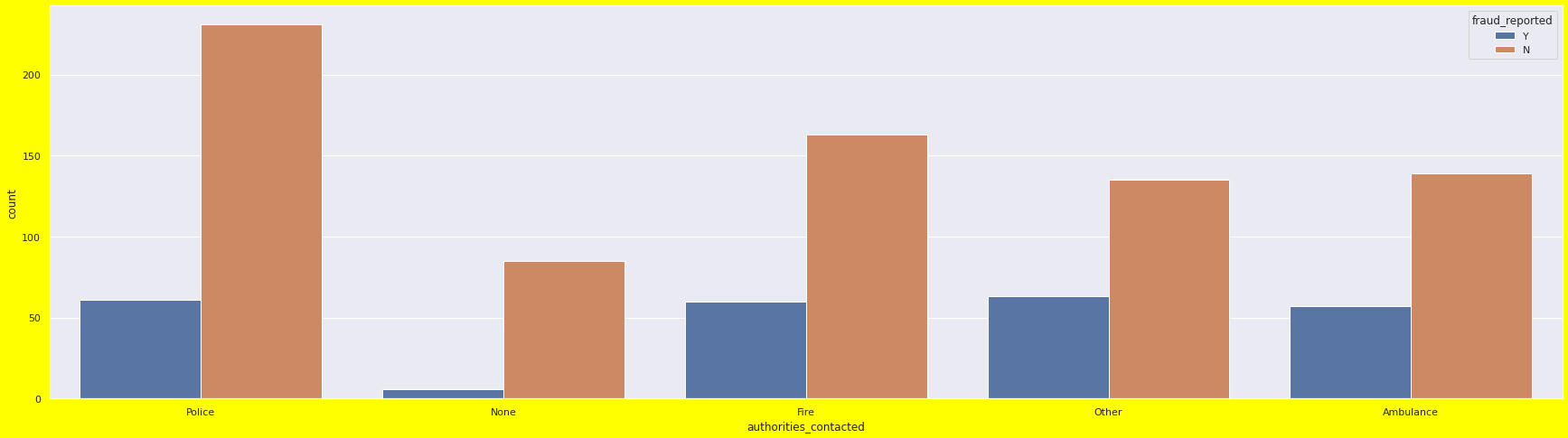
The above graph shows the relationship b/w various occupation and fraud reported. People with tech knowledge, transport or are at managerial level have more fraud cases reported .



The above two graphs shows the incident type with respect to fraud reported. There are more cases of single and multi vehicle collision and both have almost equal number of fraud reported. Rear collision cases are more in number.

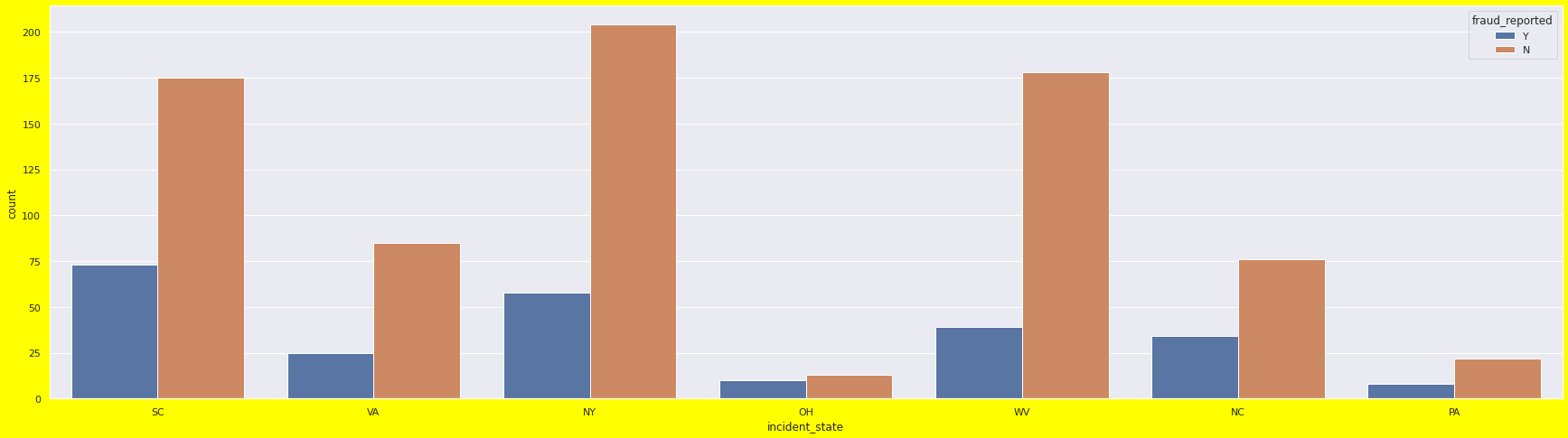


Where the damages are major fraud cases are more in number and there are negligible fraud cases where trivial damage is reported.

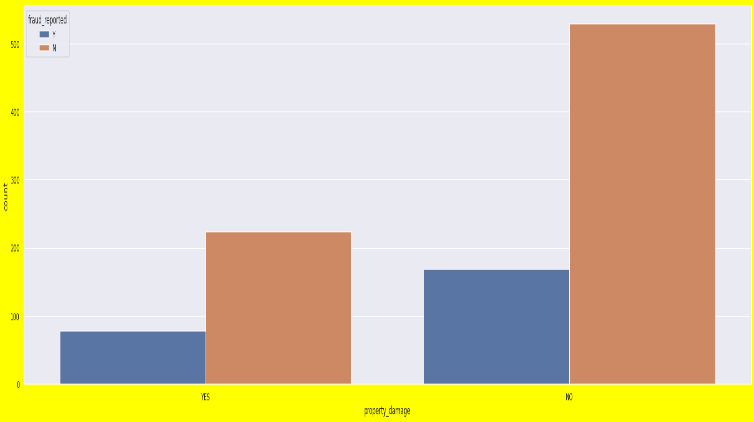
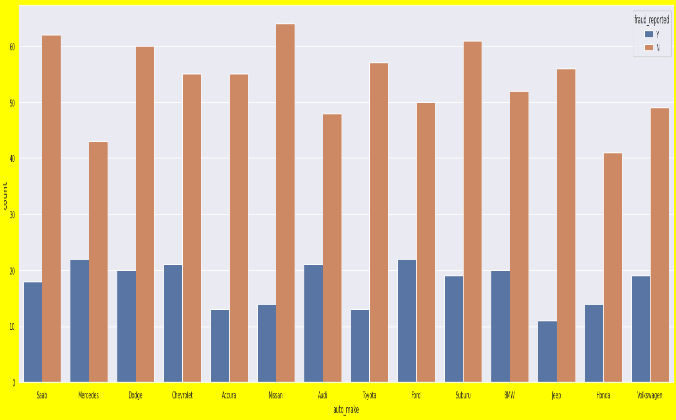


There are least fraud cases where the accident is not reported. Maximum fraud cases are reported where the authorities contacted are other than police, fire or ambulance.

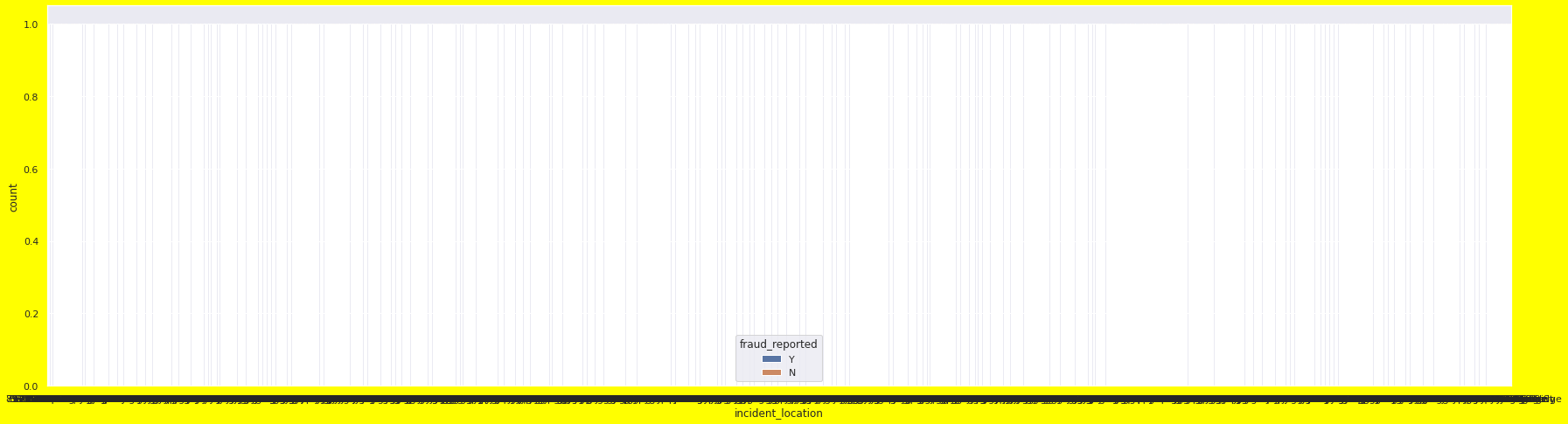
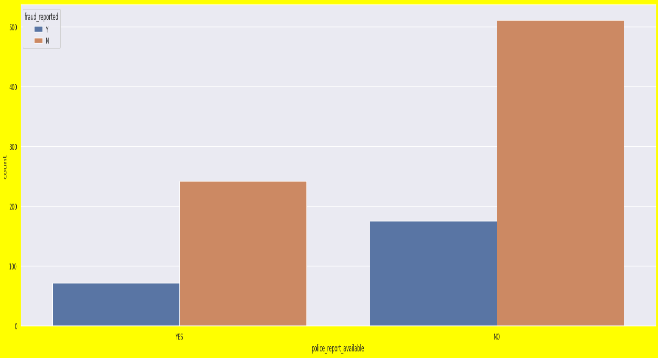
These are some more graphs which shows that categories of that column has similar effects on the fraud reported.

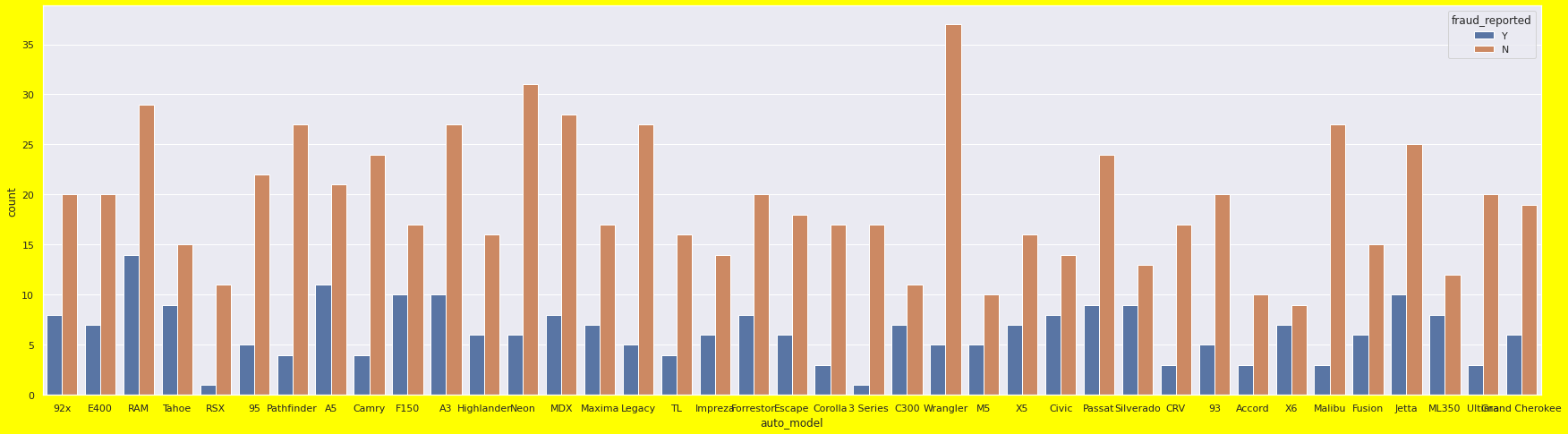


Checking the correlation of all the columns



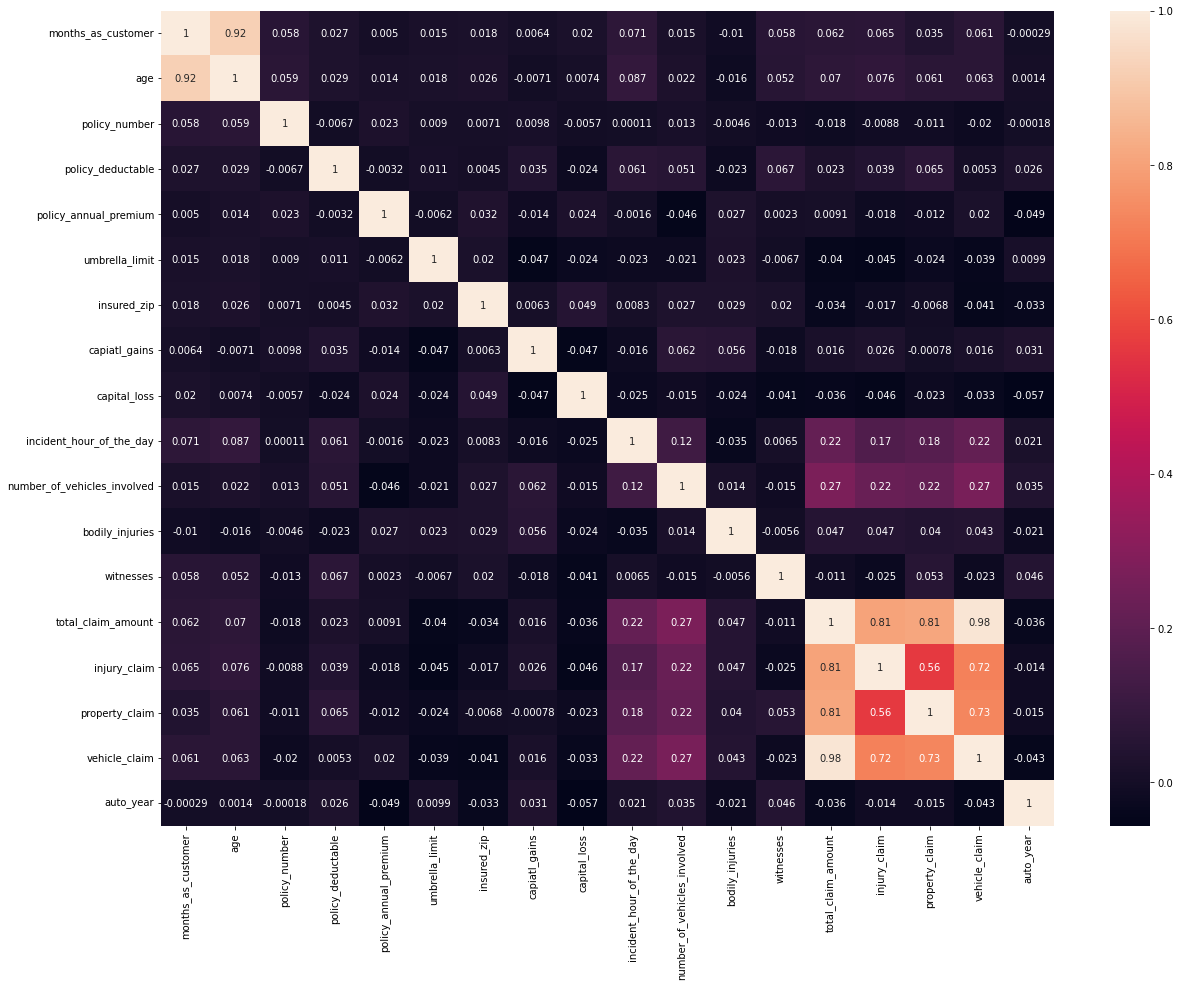
.



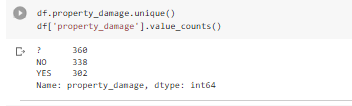
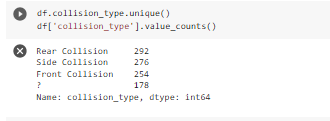


The following heatmap shows that there is correlation of age and months as customer.

Total claim injury has high correlation with injury\_claim, property\_claim , vehicle\_claim.



**Pre-Processing Pipeline:-**

After the analysis of null values, let us check the unique values of the columns. While checking the null values these observations were seen.

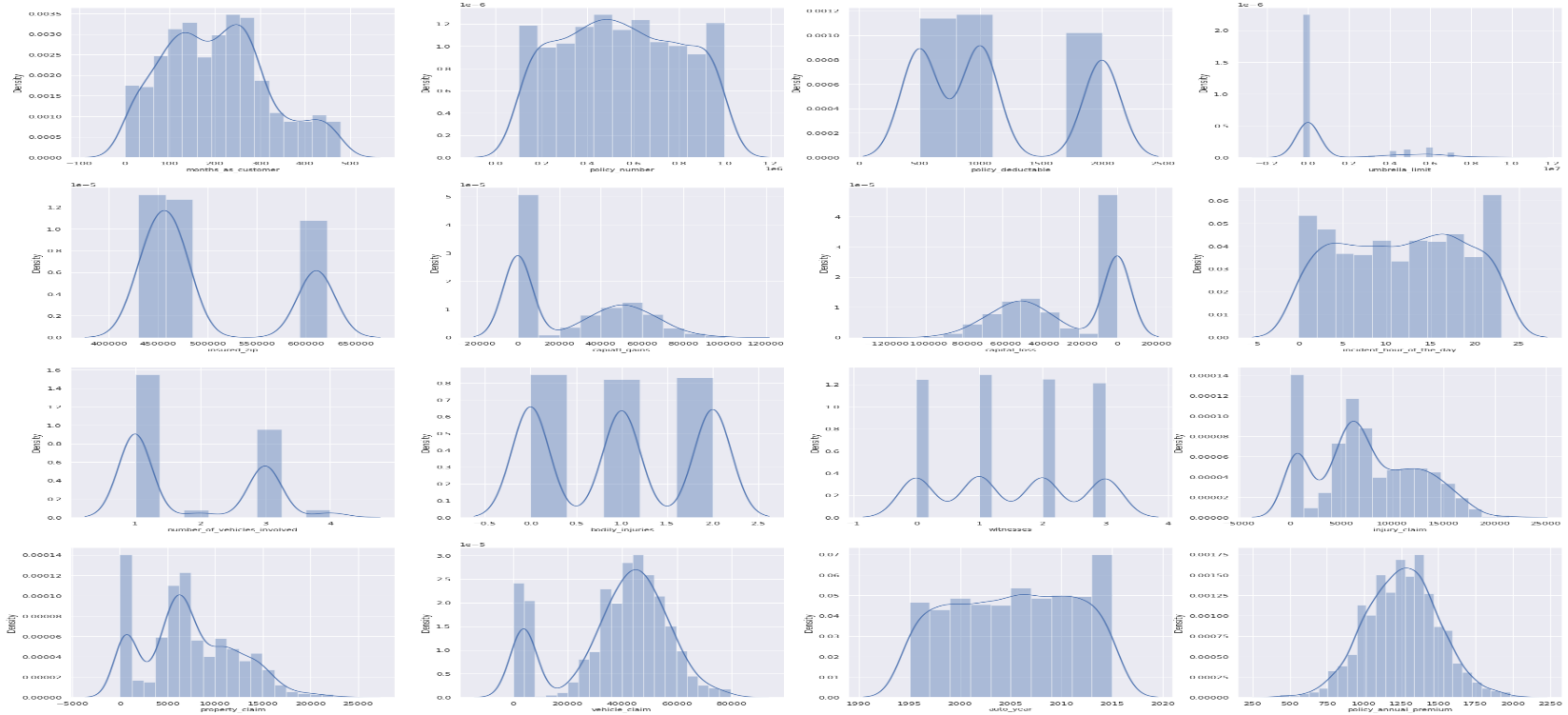
Though the data set showed null values , we could unique value “?” in the these three columns.It wasn’t shown in the previous code as these columns are strin datatypes and ? is treated as a value.

To treat the “?” following code was done:



Some columns were renamed as well in proper format.



Now let us check the numerical columns distribution.

Some columns have categorical data in numbers like bodily injuries, number of witnesses, vehicles involved.

1.Months as customer is normally distributed.

2.Policy number is normally distributed.

3.Policy deductible with less amount has more density.

4.Umbrella\_limit as 0 values in maximum that is why the data point are more in number in 0.

5.Insured zip is more between 450000 to 500000 as compared to higher amount.

6.Capital gains and capital loss have skewness towards zero values.

7.Incident hour of the day is normally distributed.

8.1 vehicle involved is more in number followed by 3.

9.Bodily injuries have categorical data and are evenly distributed in all categories.

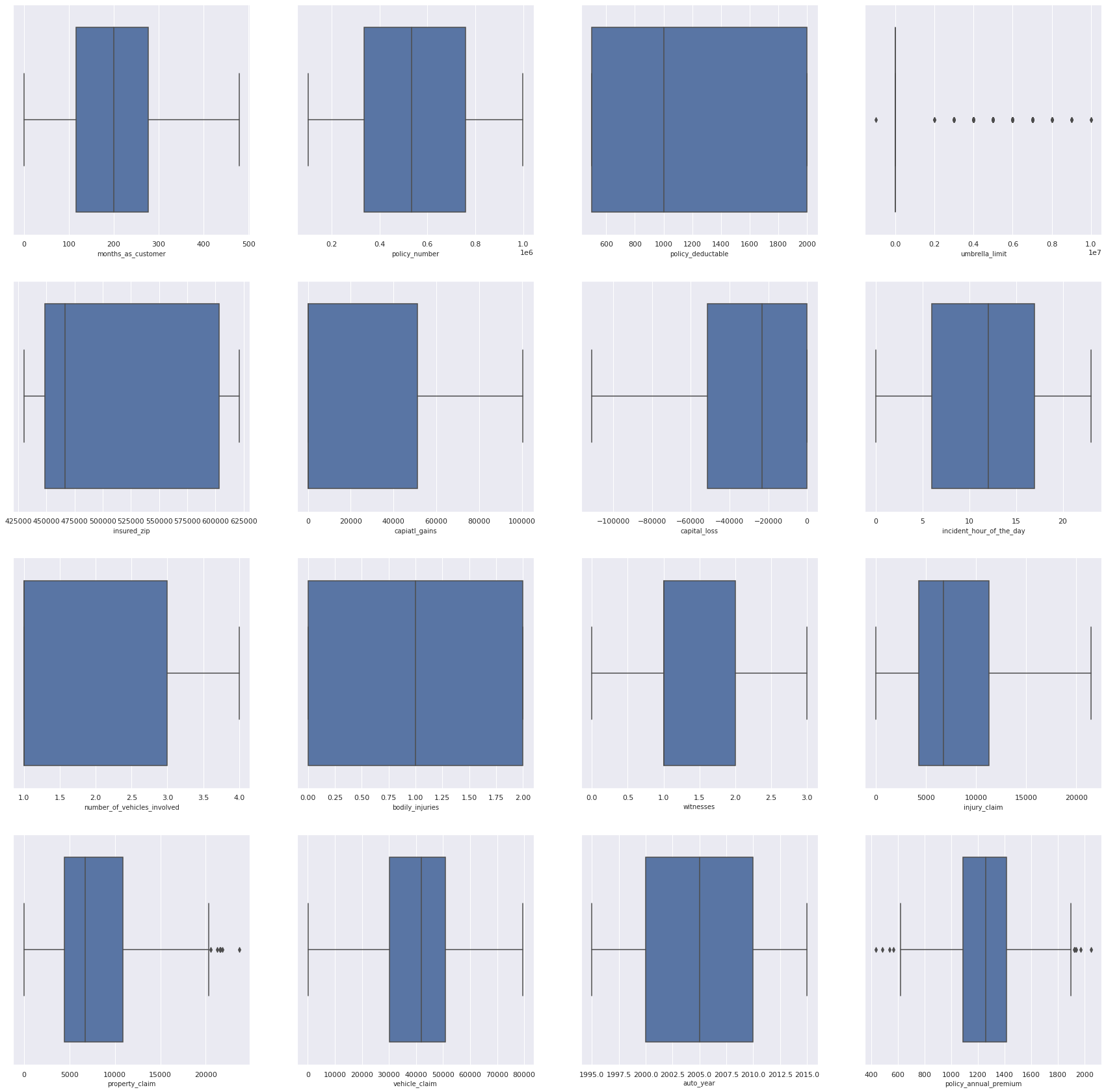
10.Witness injuries have categorical data and are evenly distributed in all categories.

11.Insurance claims are left skews and have data accumulated near 0 value points.

12.Prpoerty and vehicle claims are normally distributed except most of the data points are accumulated in 0. It can be because of zero claims in this category.

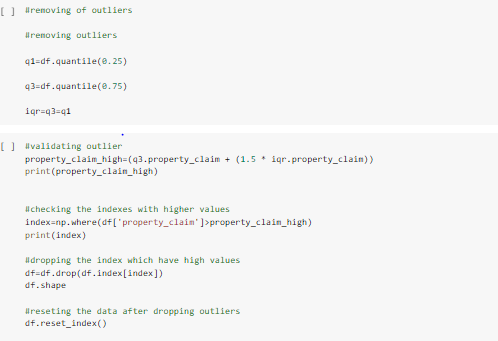
13.Auto year is normally distributed.

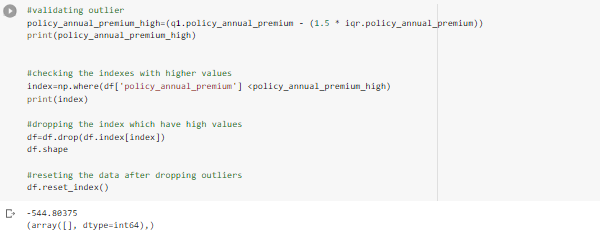
Lets check outliers:

**Umbrella limit and property claims have outliers** .

Lets treat the outliers. We have opted IQR(INTER-QUARTILE-RANGE) approach to remove outliers.

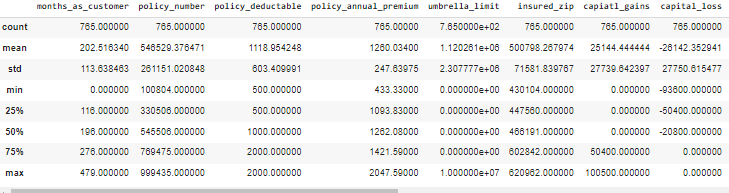
the interquartile range tells you the spread of the middle half of your distribution. Quartiles segment any distribution that's ordered from low to high into four equal parts. The interquartile range (IQR) contains the second and third quartiles, or the middle half of your data set. The interquartile range is the best measure of variability for skewed distributions or data sets with outliers.





After removing outliers the dataset shape is 765,38 which was 1000,48.

Lets go through the statistics of the data after removing outliers(a snapshot of few columns)



After the observation of statistics following things can be concluded:

1.Months\_as\_customer is number of days person has been associated with the policy. Minimum days is 0 and the least days are 120 and maximum us 479 days.

2.Minimum annual premium is $ 433 and maximum is 2027.

3Umbrella limit has zero values.

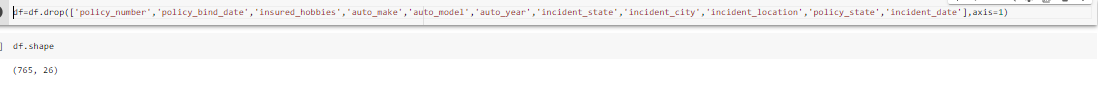
4.Number of minimum vehicles involved is 1 and maximum is 4.

5.Body injuries is minimum zero which implies that accident must have not caused any damage.

6.Maximum insurance claim is $21450.

7.Claims have zero values because that specific customer must have not taken the claims.

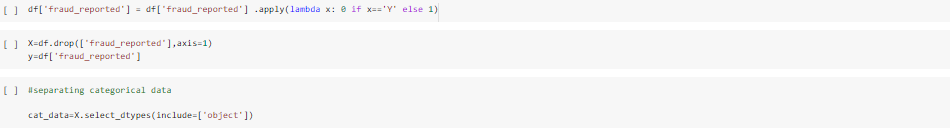
8.Vehicles are manufactured between the years 1995 to 2015.

Next we will be dropping columns which have zero relation with the fraud reported as seen in the above graphs.

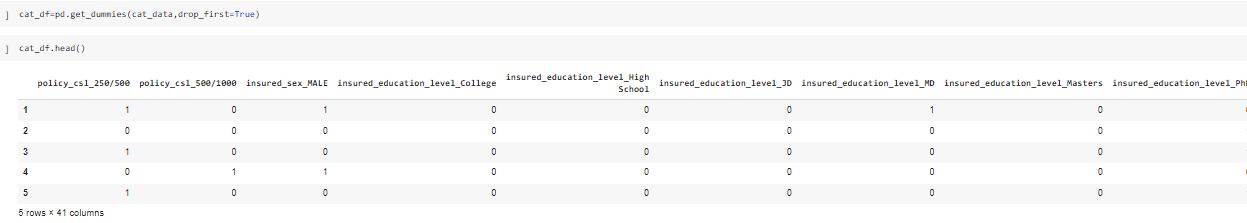
The next step after removing outliers is encoding the data . Encoding refers to converting the labels into a numeric form so as to convert them into the machine-readable form. Machine learning algorithms can then decide in a better way how those labels must be operated. It is an important pre-processing step for the structured dataset in supervised learning.

We will opt pandas get\_dummies technique, **pandas.get\_dummies()** is used for data manipulation. It converts categorical data into dummy or indicator variables.

For that we will divide categorical and numerical columns.

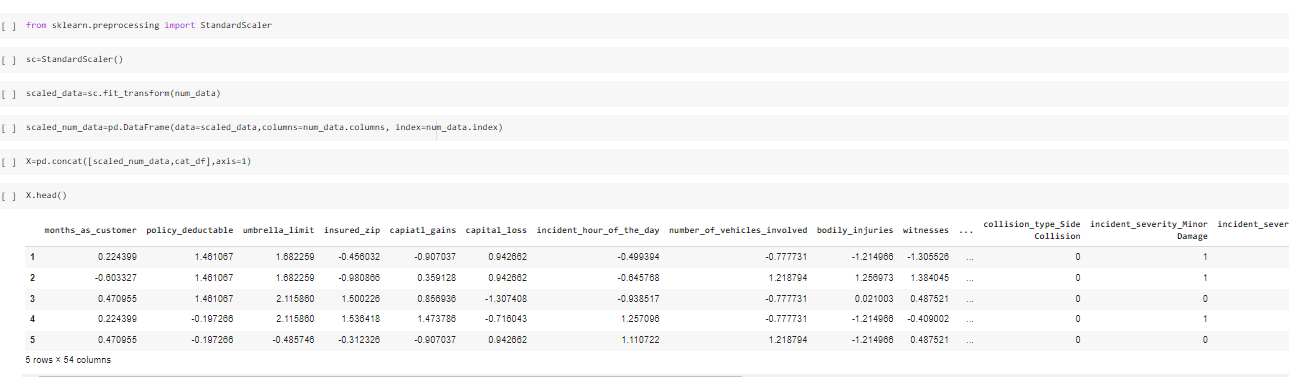




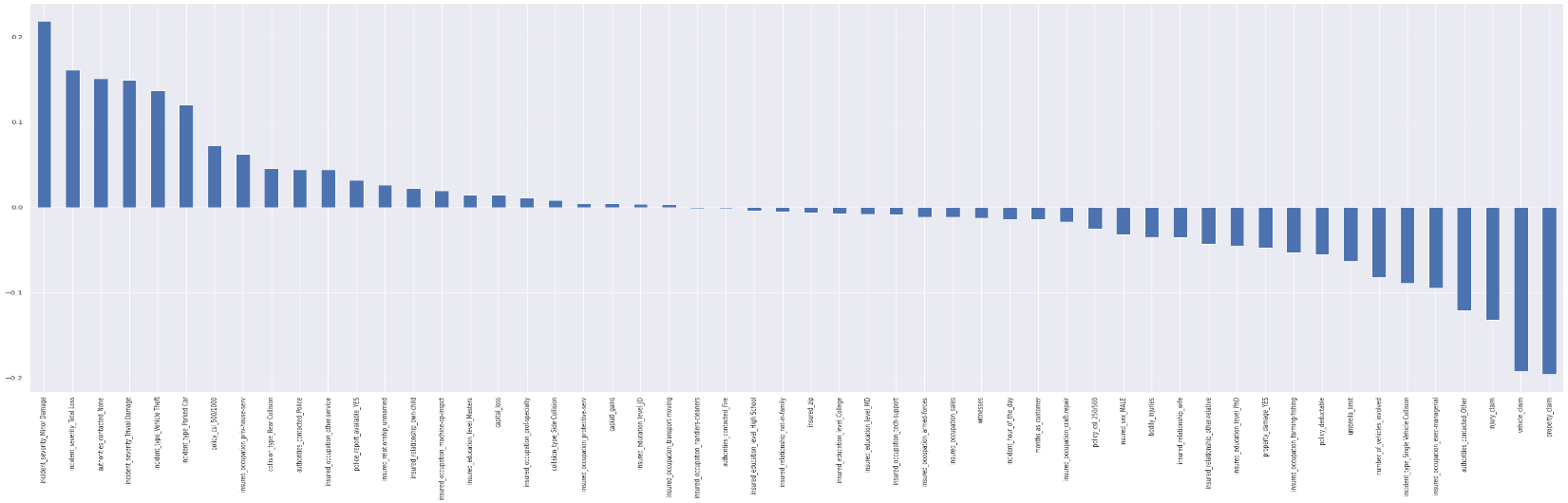


Now after encoding the data we will standardize the data of other numerical columns. Asthe characteristics of the input dataset differ greatly between their ranges, some in 0 and 1 and others with high values.

Standard scalar standardizes features of the data set by scaling to unit variance and removing the mean (optionally) using column summary statistics on the samples in the training set. This process is a very common pre-processing step. Standardization improves the convergence rate during the optimization process.



Checking the correlation with target variable.



Though some columns have very less relation with the target still we will keep them as we removed zero relation columns already.

**Building Machine Learning Models:-**

After pre-processing techniques, the data is cleaned to use.

Sklearn provide two model building category that is regression and classification.

Our dataset is training model to predict the fraud reported. Our target value is classified into Yes or No which makes it a classification problem.

Importing necessary libraries.

from sklearn.model\_selection import train\_test\_split

from sklearn.linear\_model import LogisticRegression

from sklearn.metrics import accuracy\_score

from sklearn.metrics import classification\_report

from sklearn.model\_selection import cross\_val\_score

from sklearn.tree import DecisionTreeClassifier

from sklearn.ensemble import RandomForestClassifier

from sklearn.neighbors import KNeighborsClassifier

knn=KNeighborsClassifier()

rf=RandomForestClassifier()

dt=DecisionTreeClassifier()

lr=LogisticRegression()

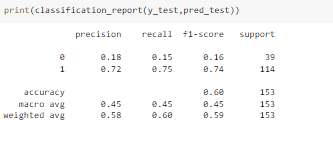
After importing we will divide our data into train\_test\_split. The train-test split procedure is used to estimate the performance of machine learning algorithms when they are used to make predictions on data not used to train the model.It is a fast and easy procedure to perform, the results of which allow you to compare the performance of machine learning algorithms for your predictive modeling problem.

We will train in for range(0,1000) random state to find out the best accuracy score and then at random state we train all the models.

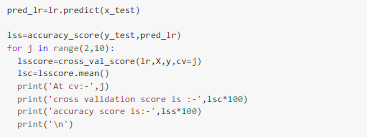


The best random\_state is 27.

Classification report of Logistic Regression model



After the classification report we generate cv score. In cross-validation, we run our modeling process on different subsets of the data to get multiple measures of model quality.

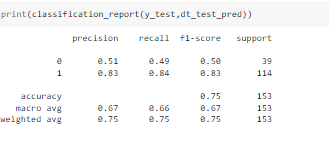


At cv= 8 we get the least difference between cv and accuracy score.

The cv score is 81.04714912280701

the accuracy score is 85.62091503267973

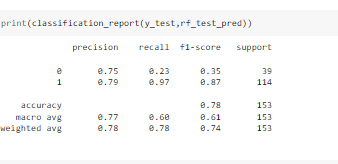
Decision tree classifier



The cv score is 72.02614379084967

the accuracy score is 75.16339869281046

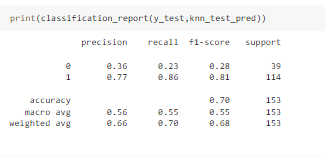
Random Forest Classifier



The cv score is 78.29679319371728

the accuracy score is 78.43137254901961

KNeighbors classifier



The cv score is 72.54581151832461

the accuracy score is 69.93464052287581

As we have two categories in our classification we will take f1-Score for model evaluation.

The **F1-score** combines the precision and recall of a classifier into a single metric by taking their harmonic mean. It is primarily used to compare the performance of two classifiers.

F1 score= 2(P\*R) / P +R

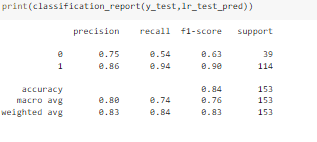
P=PRECISION

R=RECALL

The best f1 score is of logistic regression model and random forest classifier.

So, to improve their performance well will tune the data through hyper parameter tuning techniques. Hyperparameters are parameters whose values control the learning process and determine the values of model parameters that a learning algorithm ends up learning. It helps us to solve the problem of over-fitting/under-fitting.





Now we will again take out cross\_val\_csore with new parameters model. At cv=9

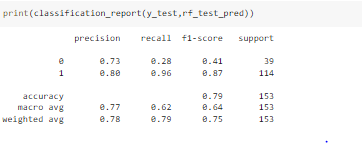
The cv score is 81.43790849673202

the accuracy score is 83.66013071895425

As we can observe the accuracy score went down to 83% from 85% after tuning. This means the logistic model was working better before tuning.

Now let us tune another model.





At cv= 2

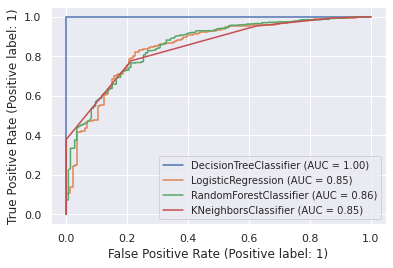
The cv score is 74.64082129235985

the accuracy score is 79.08496732026144

The model has improved performance by 1 % after tuning.

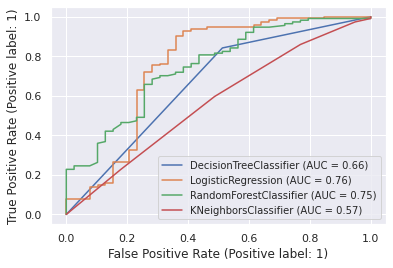
Lets plot ROC\_AUC\_CURVE to check the best model

Training model curve



Decision tree classifier is working best on training data with AUC(AREA UNDER CURVE) =100%

The Area Under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve. The higher the AUC, the better the performance of the model at distinguishing between the positive and negative classes.

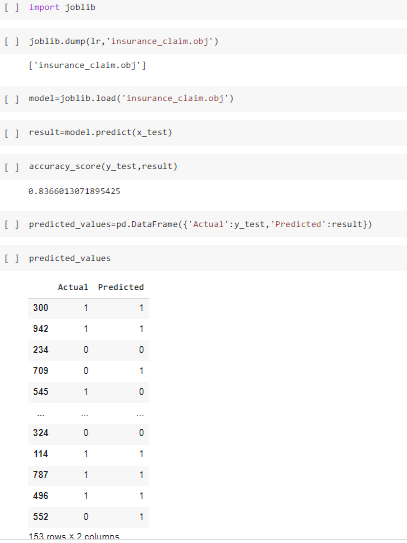


On test data Logistic regression model is working the best with AUC=76%.

There is lot of difference between test and train data for Decision tree classifier.

We will take Logistic regression as our final model as it is working the best with testing data.

Now let us save the model and check how well our model is able to predict.



**Conclusion:-**

Logistic regression model works the best with the dataset with accuracy score of 83% and ROC\_AUC= 76%.

With 6 true positives out of 39 positives and 86 true negatives out of 114 total negatives.

If class imbalance is removed the model efficiency can be improved. Feature selection techniques like Spercentile can help to have best features which in turn reduce the burden on the model and improve its efficiency

This dataset comprises data of three states so it is biased to it. Larger dataset with more states will help the model to learn better.