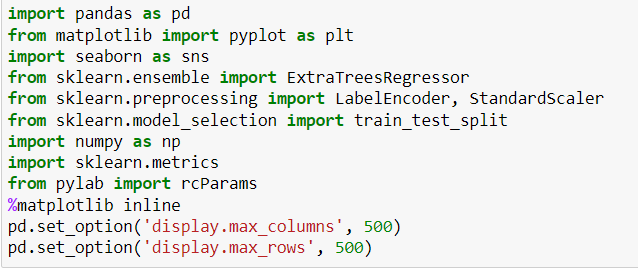
***Insurance Claims- Fraud Detection***

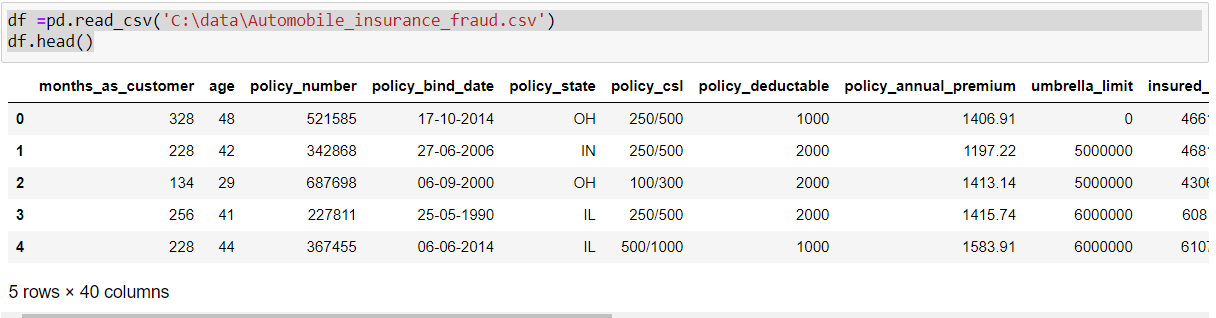
Insurance fraud is a huge problem in the industry. It's difficult to identify fraud claims. Machine Learning is in a unique position to help the Auto Insurance industry with this problem.



Importing all Libraries

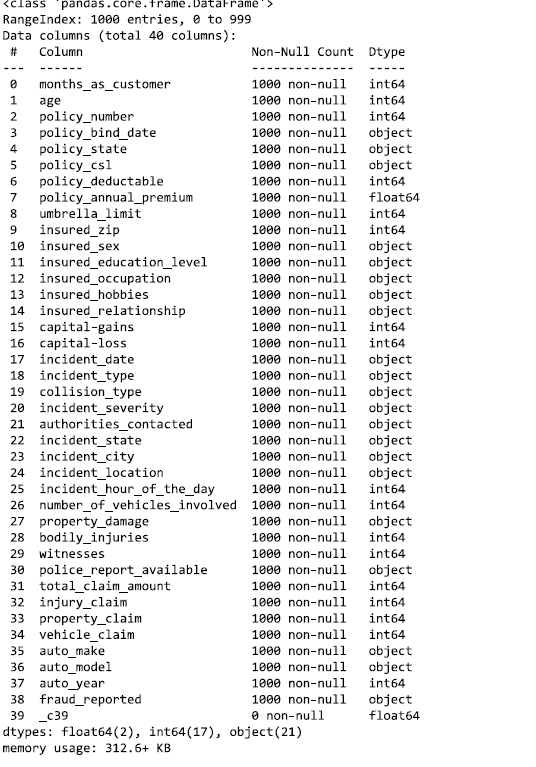


Upload the dataset through CSV file



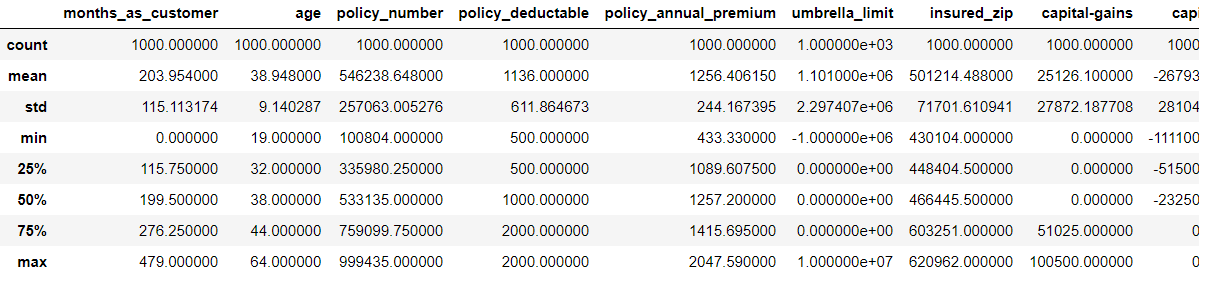
# Data Exploration/Analysis





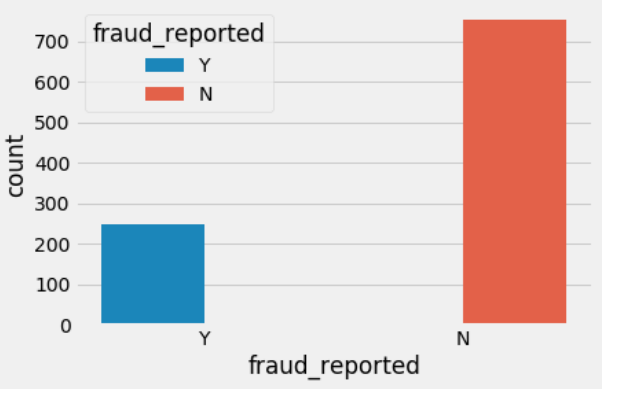
**The training-set has 1000 examples and 38 features** 2 of the features are floats, 17 are integers and 21 are objects. Below I have listed the features with a short description:





Some variables such as 'policy\_bind\_date', 'incident\_date', 'incident\_location' and 'insured\_zip' contain very high number of level. We will remove these columns for our purposes.

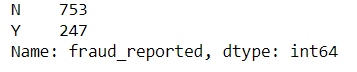




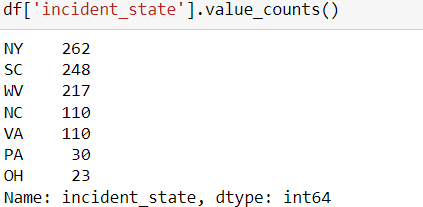
##### From above plot, like most fraud datasets, the label distribution is skewed.

***Count number of frauds vs non-frauds***

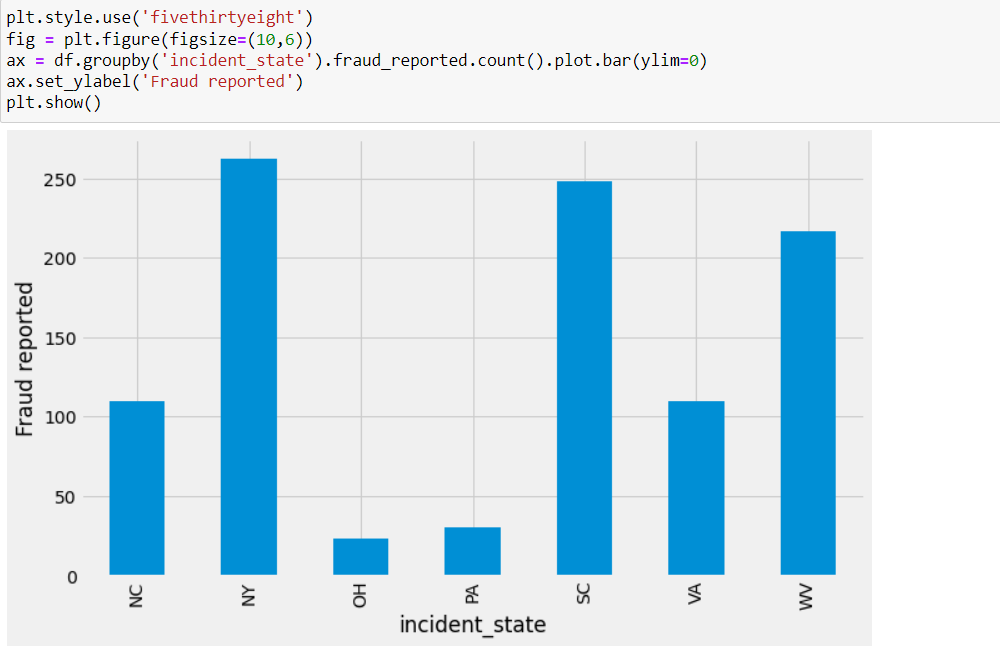




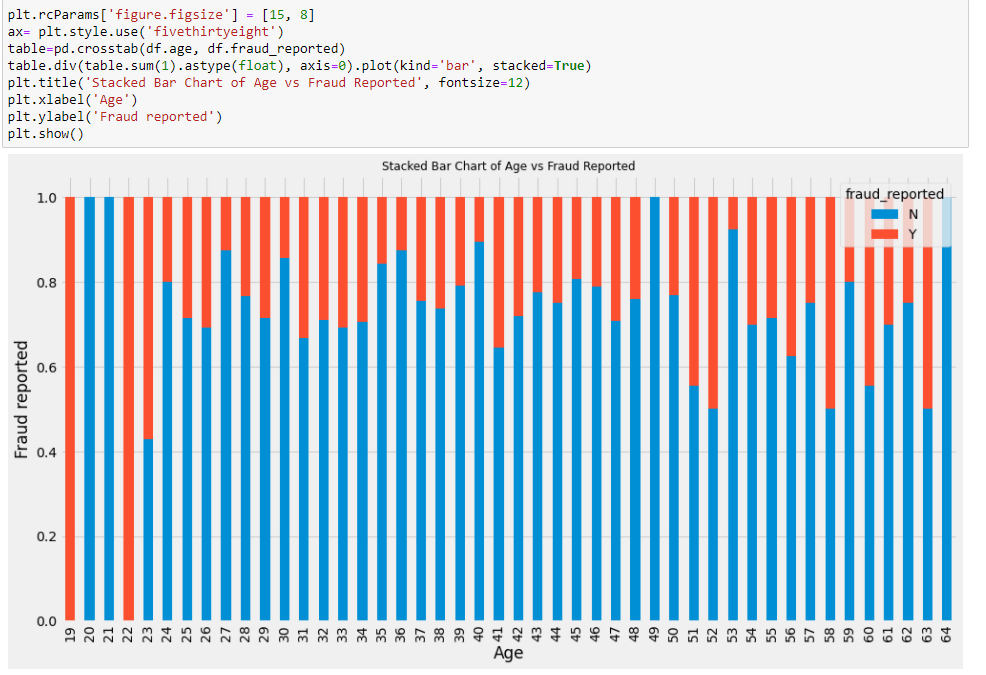
*You can see here in our data base Non Frauds reported count is high.*



We can see state wise Incident\_state count from our dataset.

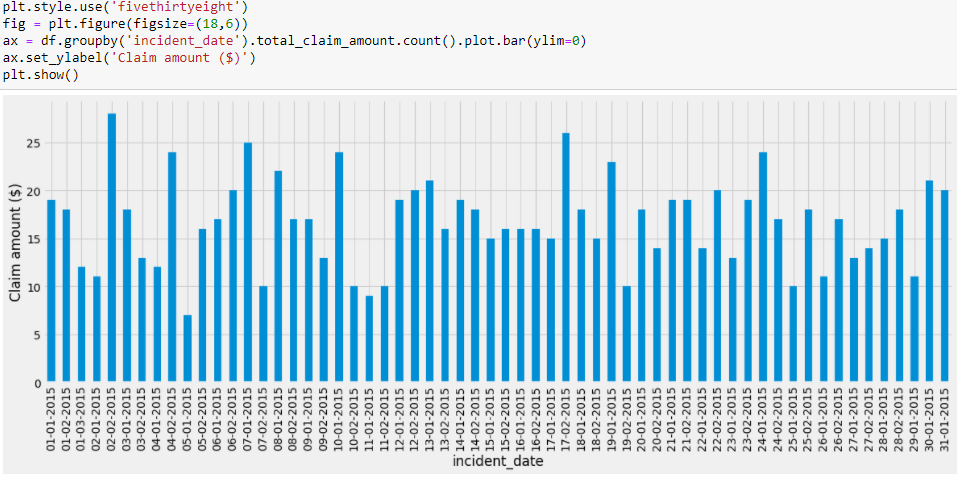


Bar plot between incident state and Fraud reported.



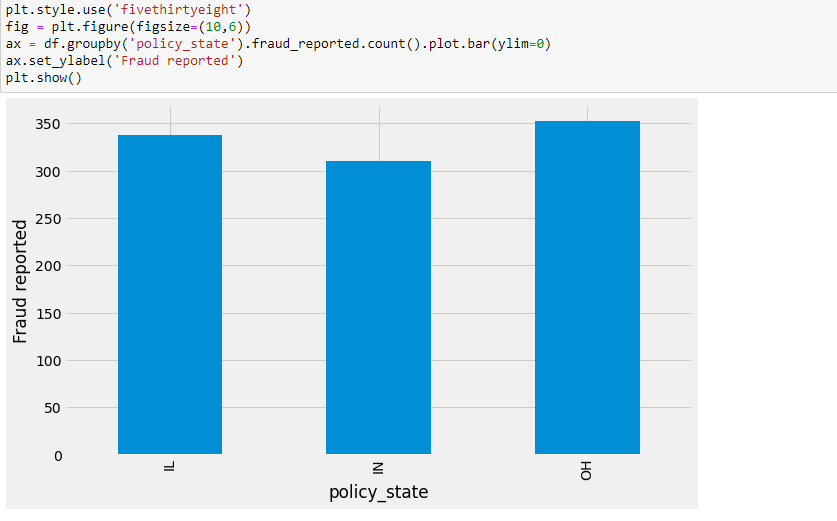
Bar plot between Age and Fraud reported.

You see here some of age people are report against fraud report and many age people not report against fraud.

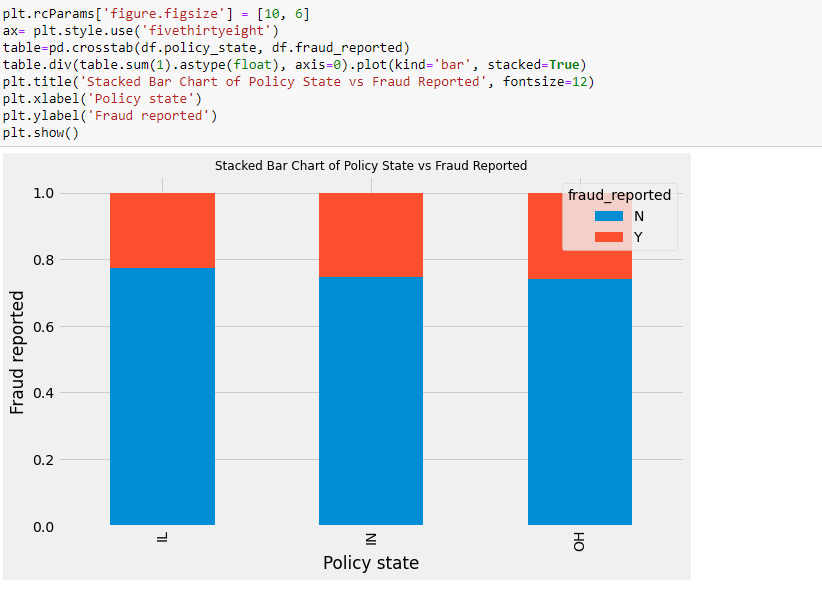


Bar plot between Incident\_date and Claim Amount

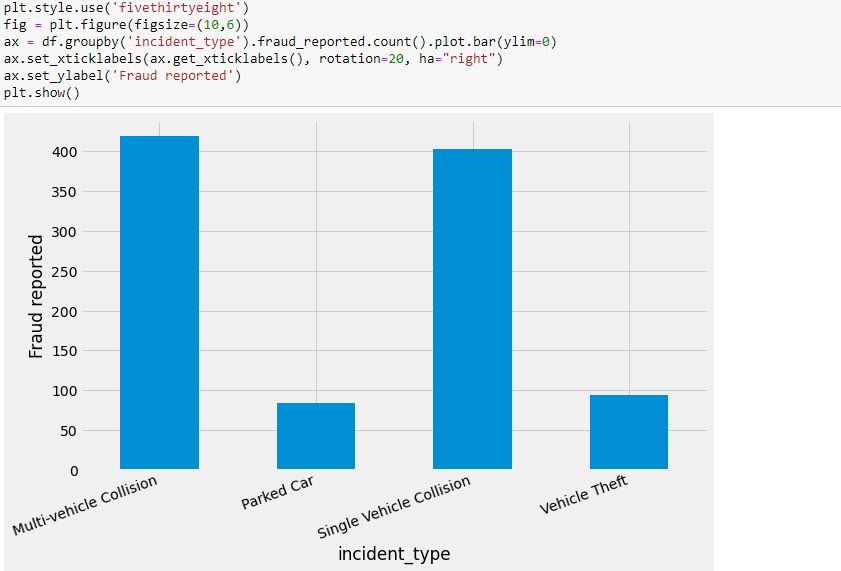
##### You can see here that all the cases in above plot are for the months of January and February 2015



Bar plot between Policy\_sate & Fraud reported.

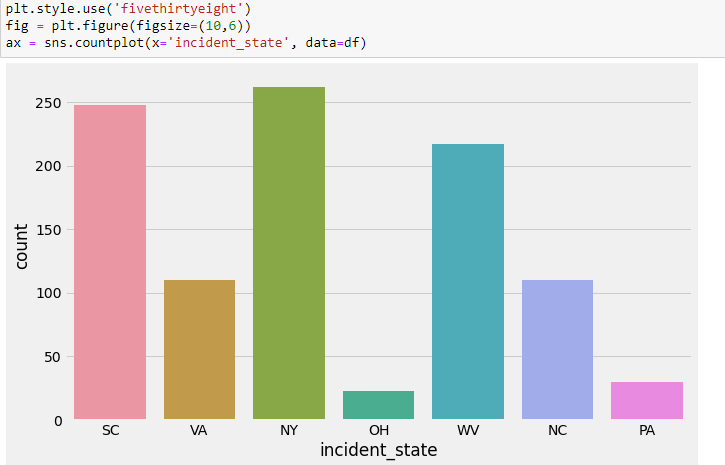


Stacked bar plot between policy state and fraud reported.



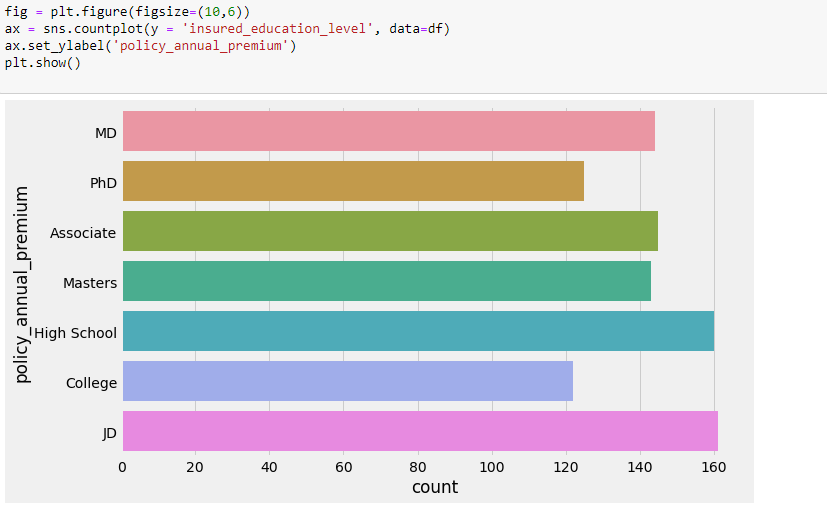
Bar plot between Incident type and Fraud reported.

Here we see many case is fraud reported against Multi vehicle Collision.

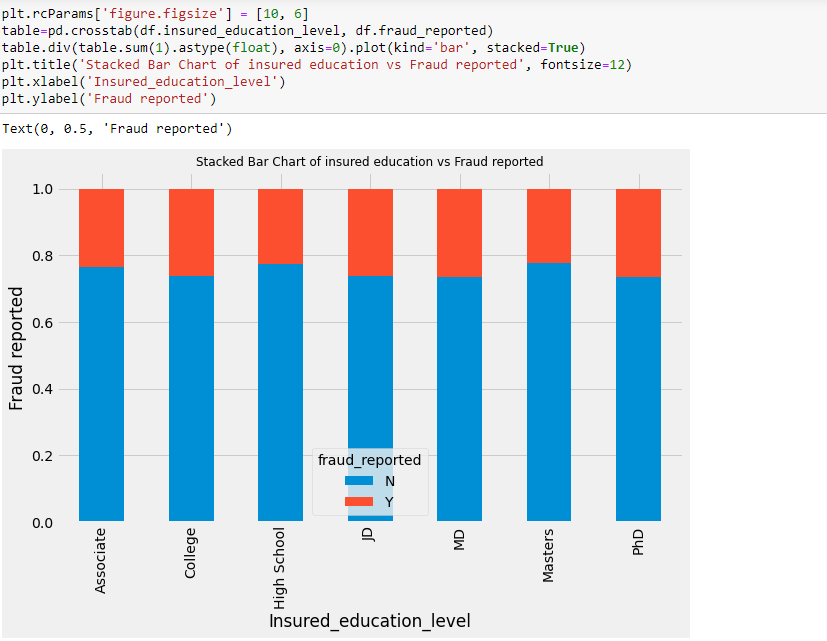


Bar Plot between Incident state and Count

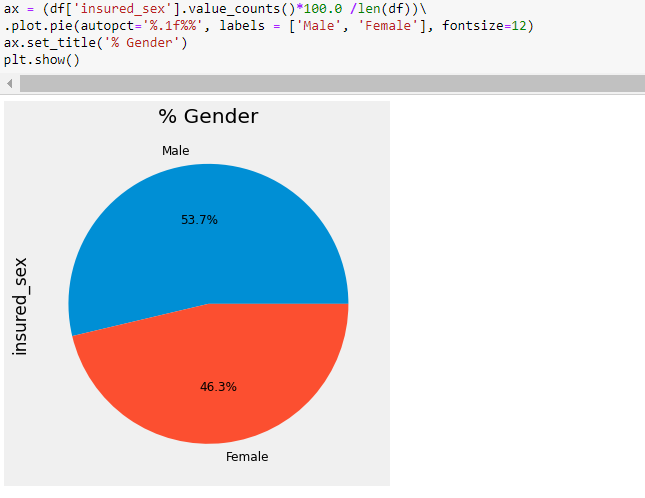
Here we see some state is count is high against incident



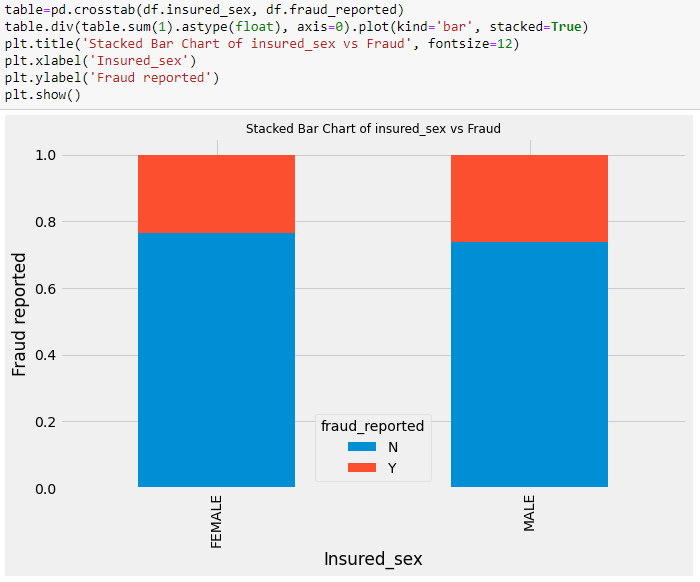
*Breakdown of Average Vehicle claim by insured's education level, grouped by fraud reported*



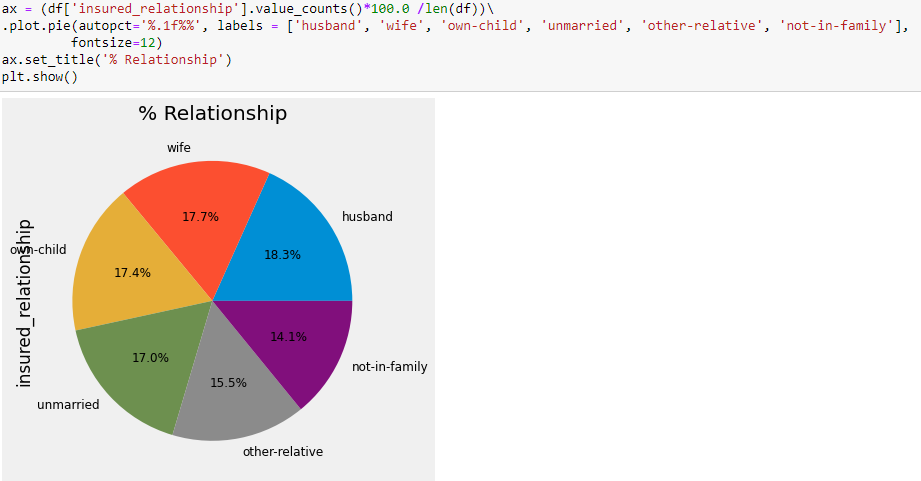
Stacked bar chart between Insured education and Fraud reported.



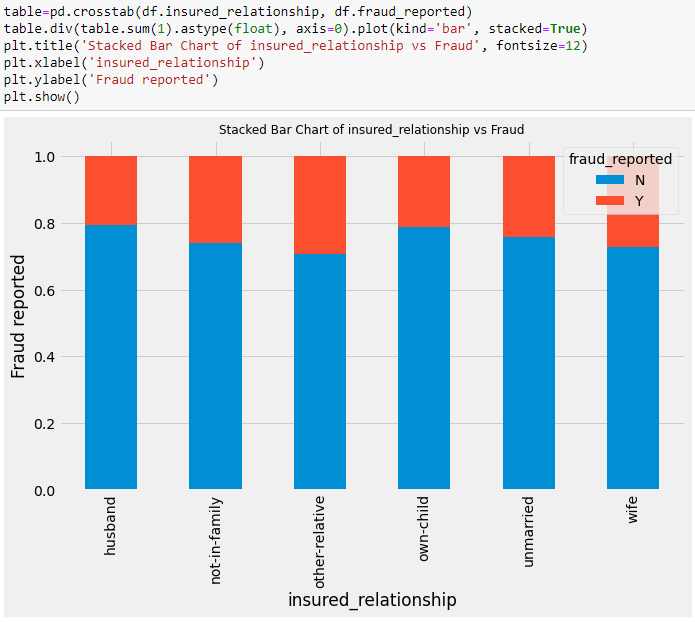
Here in pie chart we see that Male % is high in our dataset.



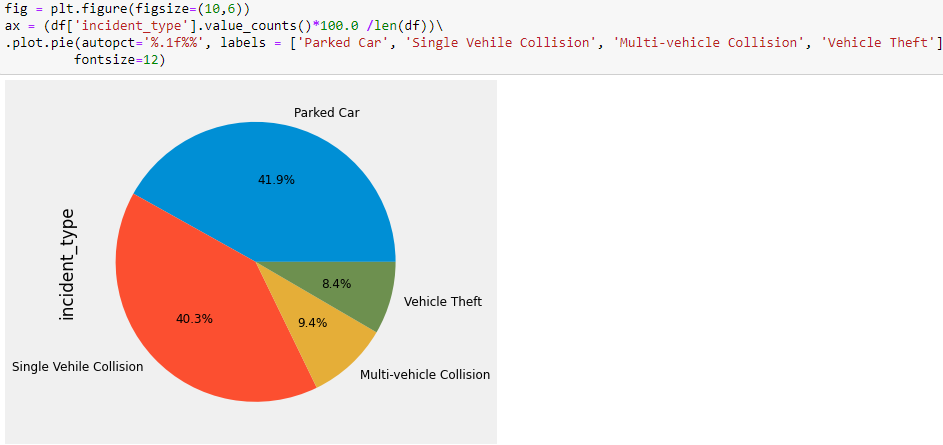
Here we see Both male and female claim in same against Fraud reported



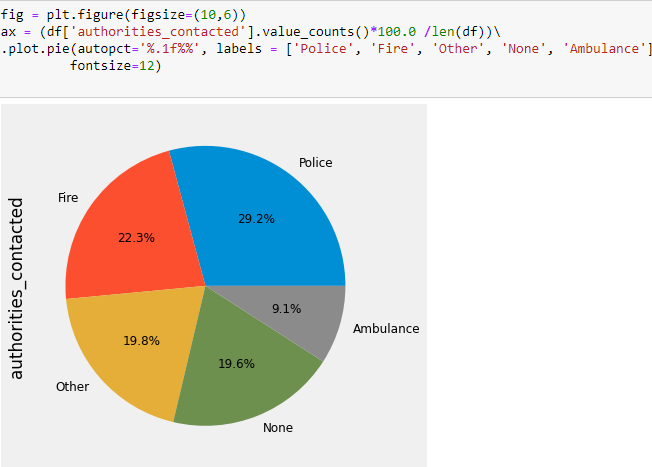
Here we see percentage of relationship against insurance policy.



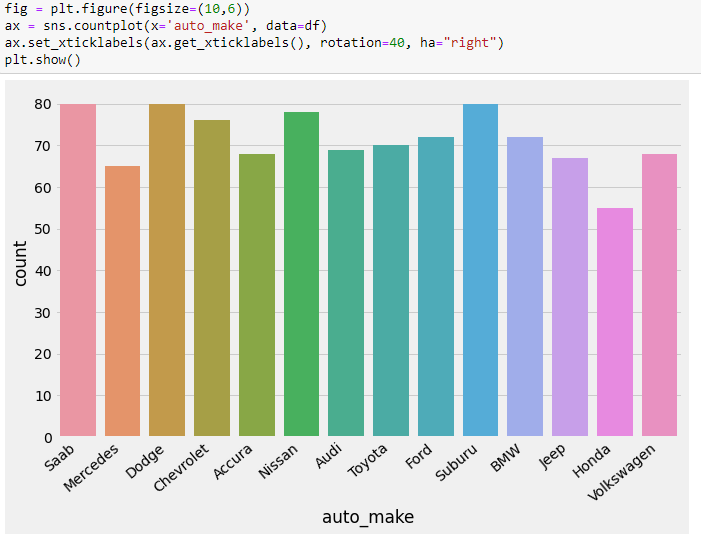
Here we see relationship claim fraud reported against the policy.



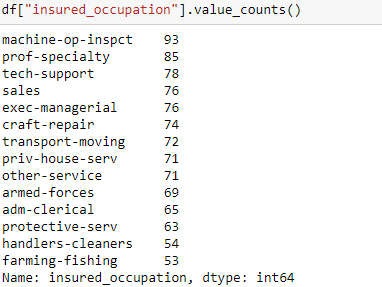
Here we see 41.9 % parked car claim against policy fraud reported.

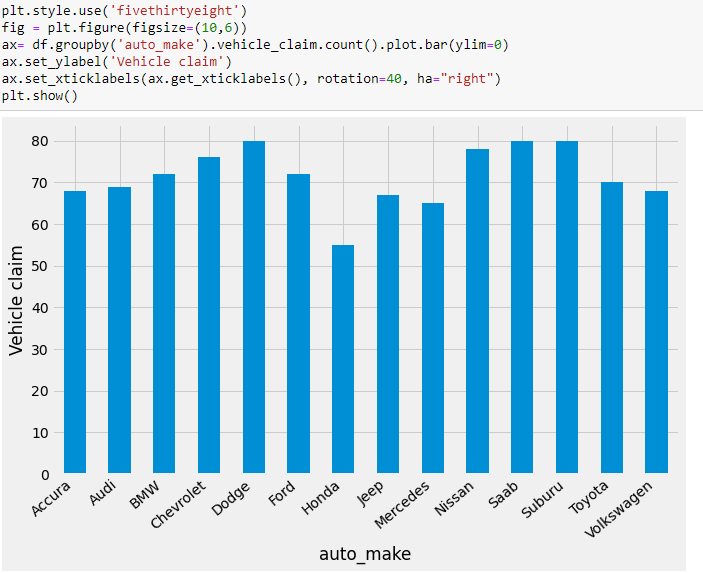


Here we see 29% of Police case involve against authorities contracted.



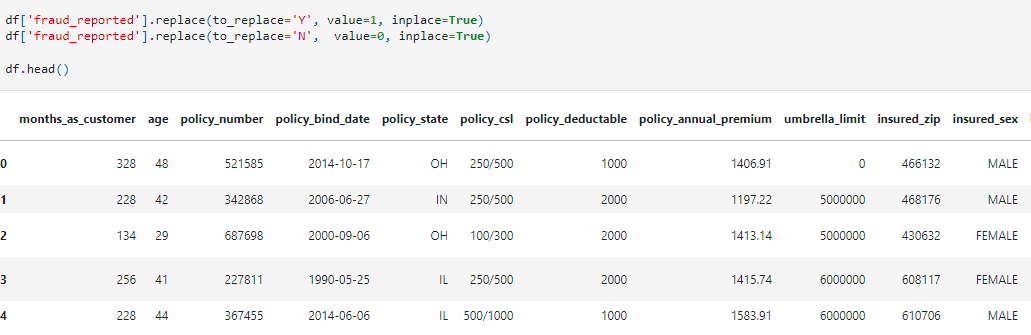
Here we see count of insurance taken by vehicle.



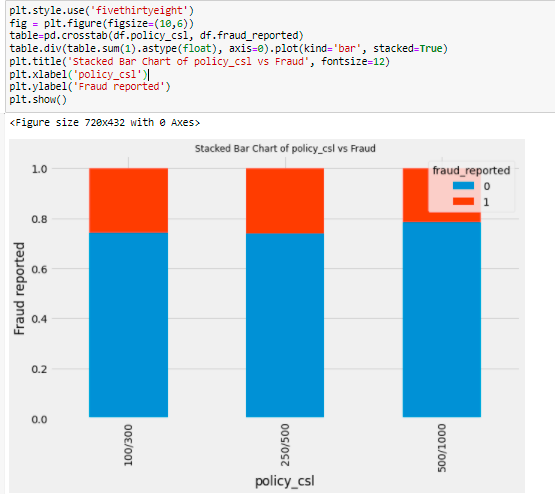


Here we count to see how many auto claim against vehicle.

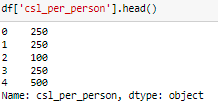
# Data Processing



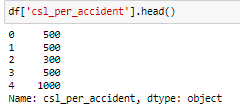
Here we Cleaning up the data and prepare it for machine learning model.



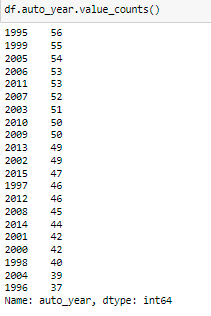
Here we check Bar plot between Policy csl and Fraud reported



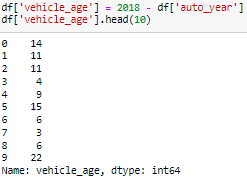
Here we Count Policy csl per person



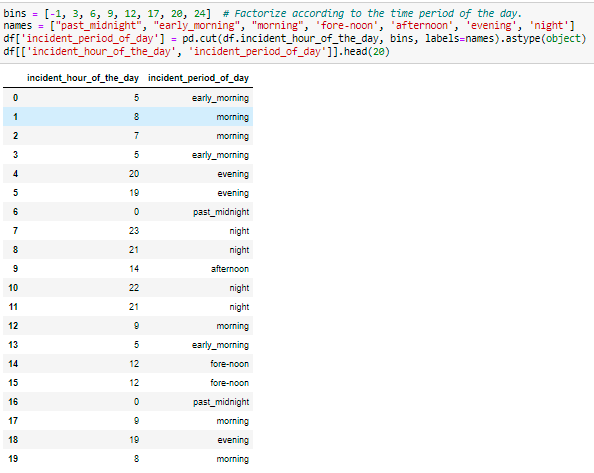
Here we Count Policy csl per Accident



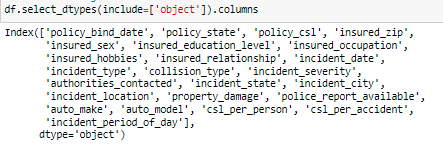
Here we see auto year has 21 levels, and the number of records for each of the levels are quite significant considering datasize is not so large. We will do some feature engineering using this variable considering, the year of manufacturing of automobile indicates the age of the vehicle and may contain valuable information for insurance premium or fraud is concerned.



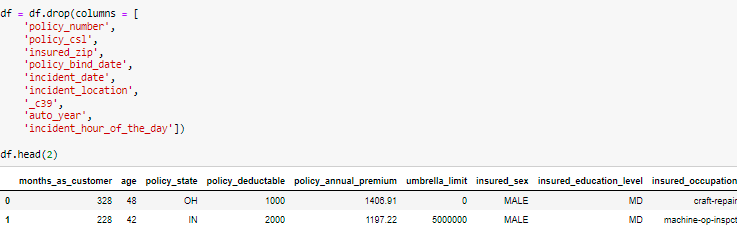
*Here we Deriving the age of the vehicle based on the year value*



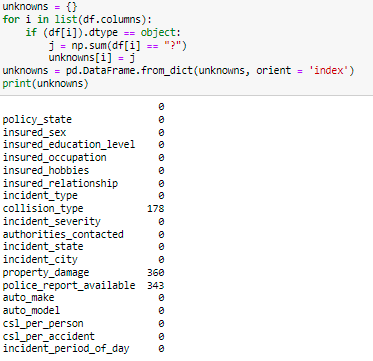
*Here we see Factorize according to the time period of the day.*



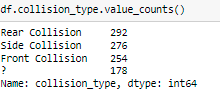
*Here we checking categorical columns in our dataset.*



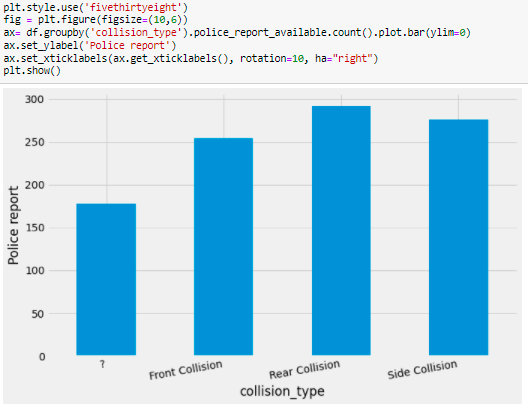
*Here we dropping unimportant columns from our dataset.*



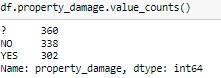
*Here we see identify variables with their values*



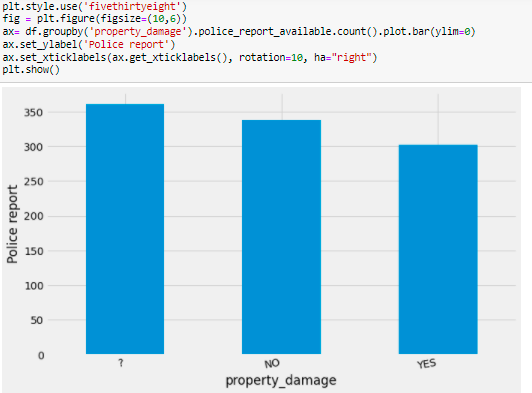
Here we see there collision covered in our dataset against car insurance



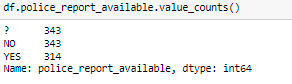
Here we see barplot against Car insurance collision type.



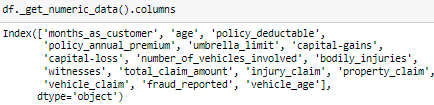
Here we see how many property damage against insurance as per our dataset.



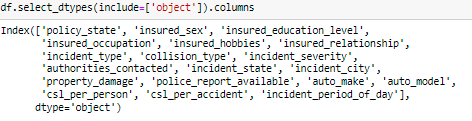
Here we see bar plot wise how many property damage against insurance as per our dataset.



Here we see How many police report available in our dataset.

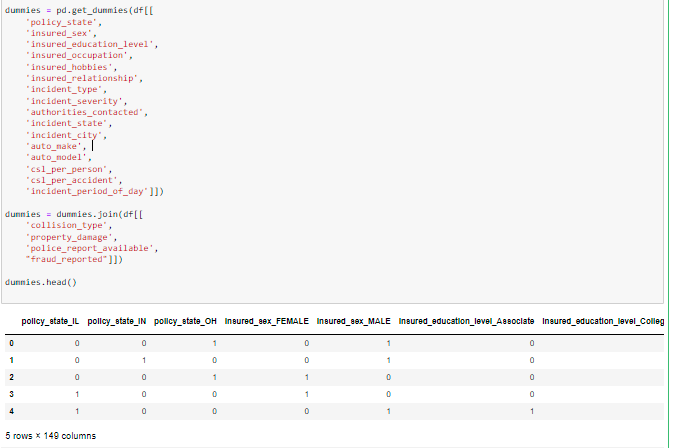


Here we check how many numerical data column in our dataset.

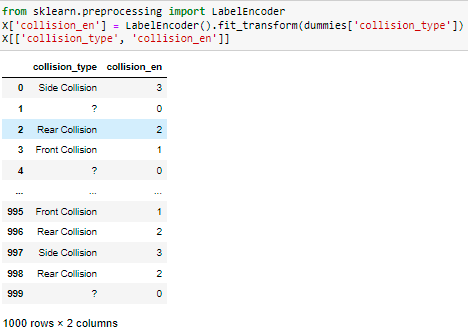


Here we see how many Object type column data in our database.

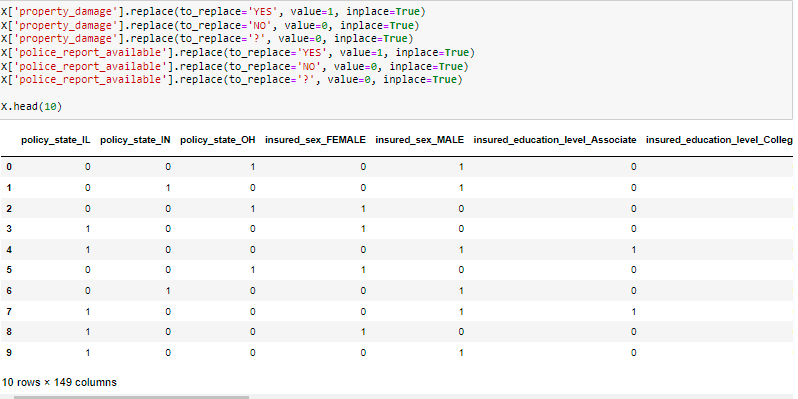
# Applying one-hot encoding to convert all categorical variables except out target variables

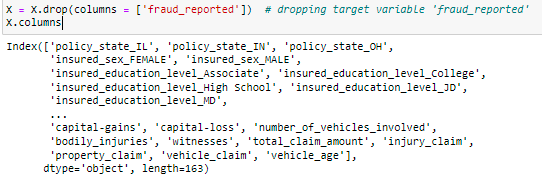


# Label encoding

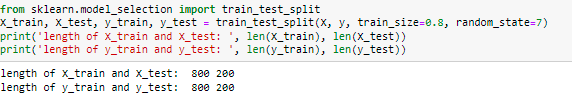


Here we see Label Encoding refers **to converting the labels into a numeric form so as to convert them into the machine-readable form**. Machine learning algorithms can then decide in a better way how those labels must be operated. It is an important pre-processing step for the structured dataset in supervised learning.



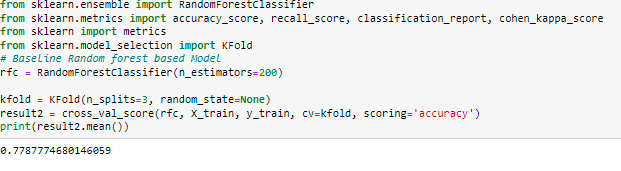


##### Here We now have a dataset that we could use to evaluate an algorithm sensitive to missing values like LDA.

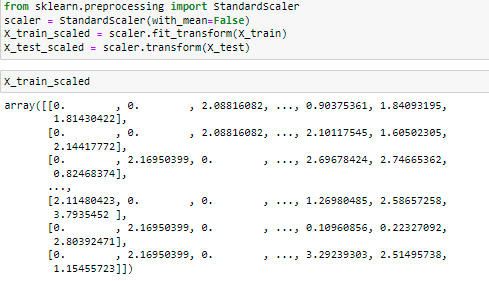


Here we check our x,y length in our dataset.

# Random Forest Classification



Here we see after using random forest our score is 77%



# Here we see Standardizing the data and recheck the data distribution.

# 

# 

# 

* Above a list of each algorithm, the mean accuracy and the standard deviation accuracy and a box & whisker plot showing the spread of the accuracy scores across each cross validation fold for each algorithm.
* It is clear that the LR or LDA is good enough for both feature selection (81% and 84% accuracy with 100 features) as well as model selection.
* I will analyse both both logistic regression and linear discriminate analysis further on this problem.