**Automobile Insurance claim**

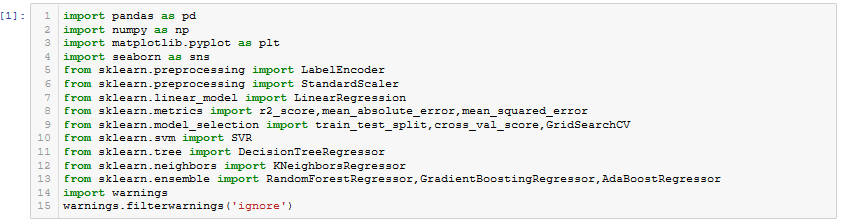
**Problem Definition:**

The purpose of an Insurance is to provide protection against the risk of any financial loss. Insurance is a form of risk management in which an insurer agrees to take the risk of the insured entity against future events, uncertain loss due to Tsunami, earthquake or damage against the vehicle or personal property. Here we are provided with Automobile insurance claim dataset.

We have to predict the claim amount in the Automobile insurance dataset depending on the features.

As this is related to price prediction so it is a Regression Problem so first of all we load the libraries required by Regression Problem and the libraries required for Data Analysis and loading.

**Step-1**



1. Pandas is for Loading, Saving and Tabular Representation of Datasets.

2. Numpy is for numerical operations, numpy array.

3. matplotlib.pyplot is for plotting.

4. seaborn is for graphical representation.

5. LabelEncoder is for Pre-processing like encoding categorical variables into their numerical representation.

6. StandardScaler is for Pre-processing like Scaling/normalization of the dataset.

7. r2\_score is for checking the accuracy score, mean\_absolute\_error and mean\_squared\_error for checking error percentage.

8. train\_test\_split for splitting the dataset into train and test dataset, cross\_val\_score for checking model fitting and GridSearchCV for detecting best score and best model.

9. LinearRegression, SVR, DecisionTreeRegressor, KNeighborsRegressor are the Machine Learning algorithims

10. RandomForestRegressor, GradientBoostingRegressor and AdaBoostRegressor are the ensembling techniques for Machine Learning.

11. During the process may warnings are come so we are suppressing the warnings by filtering.

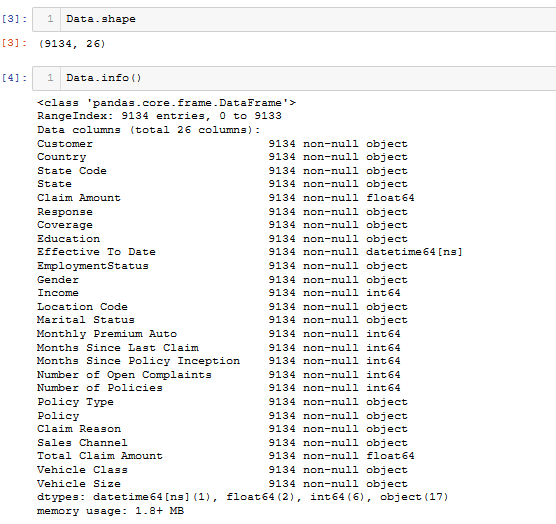
**Step-2 (Loading the Dataset)**:



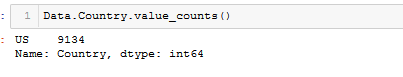
**Step-3**

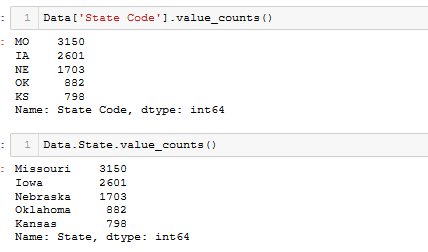
**Exploratory Data Analysis**:-

For understanding the data always we have to do EDA. What is the shape, What are the Features, What are the dtypes of the features, if there any null value present in the dataset and what are the relationship among the features, checking is there any Outlier present in the features and remove them is included in EDA. In short we explore data, find patterns and tells insights.

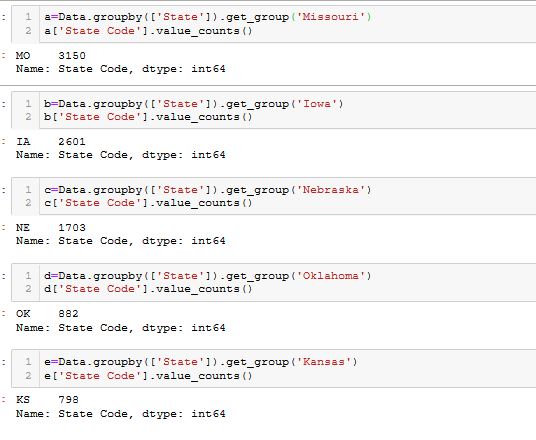


Dataset consist of 9134 observations with 26 features and there is no null value as the observations of all the features are 9134.

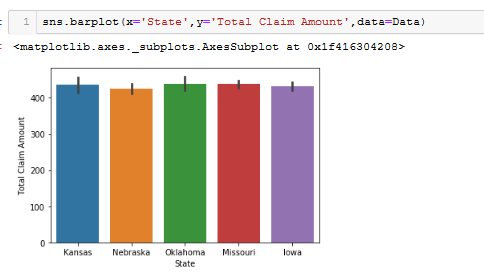


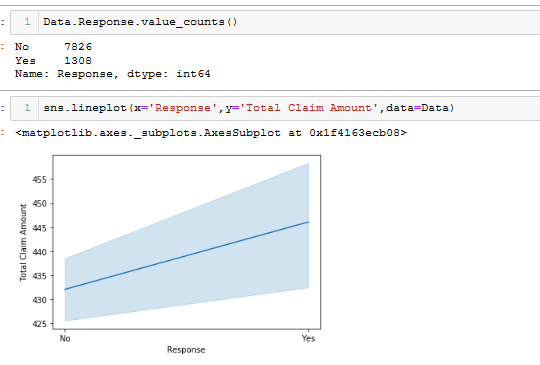


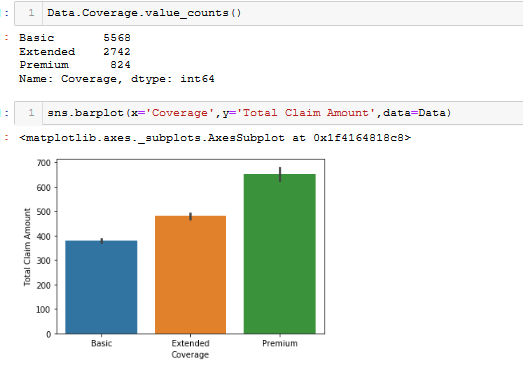
Here State Code and State have same number of observations. We check are the states are Coded as State code.

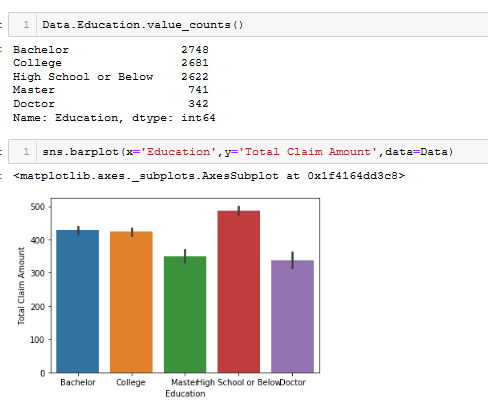


Yes, the state are coded. There is no need of present same feature 2 times so we drop state code among them.

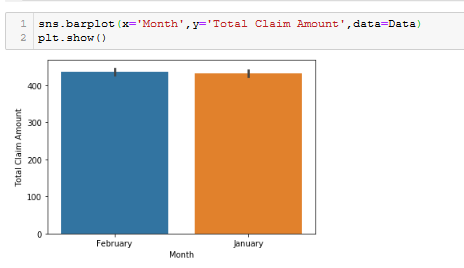


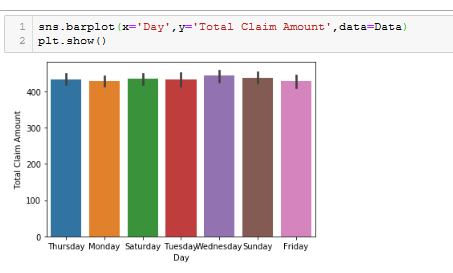


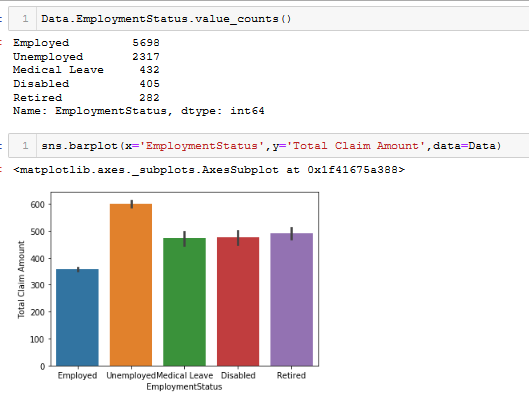


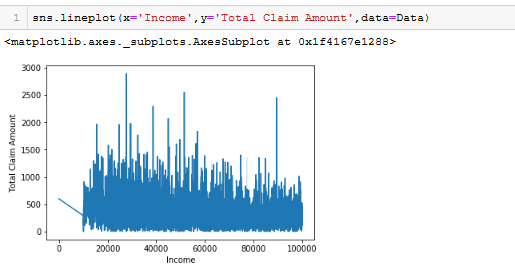


From ‘Effective To Date’ we generate Month and Day column for better Analysis and drop ‘Effective To Date'

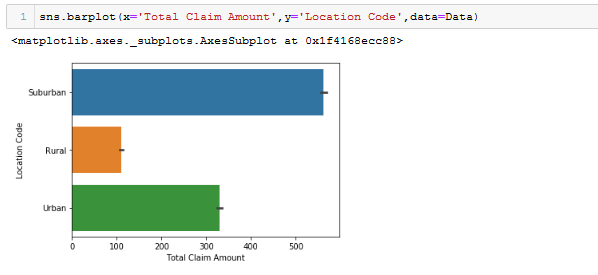


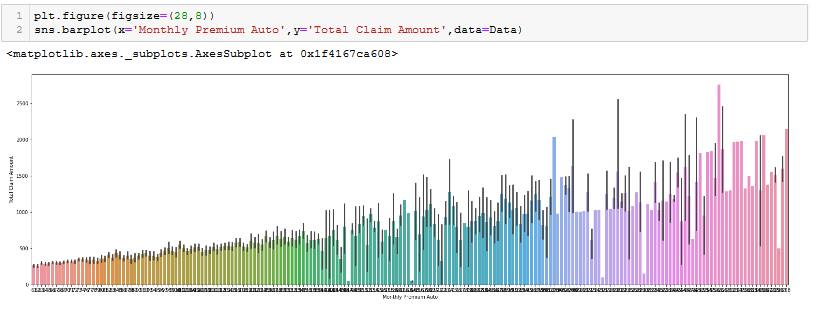


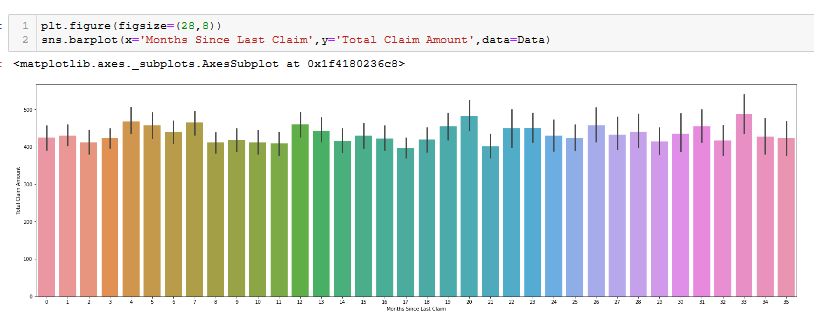


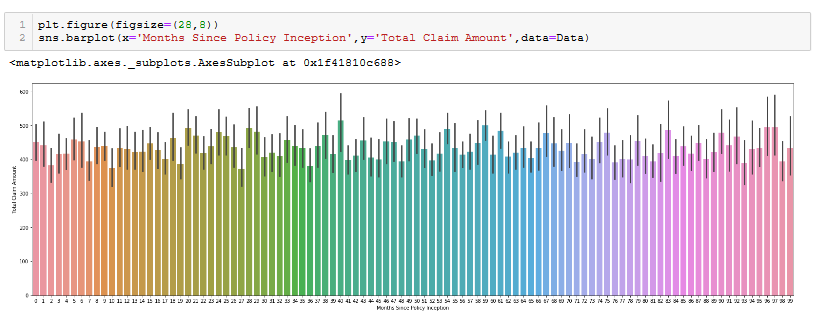


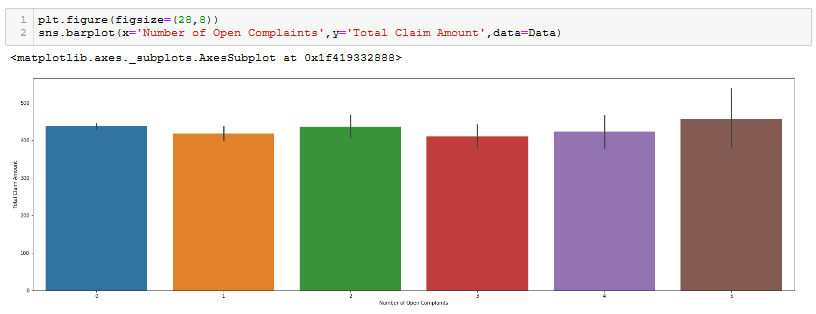


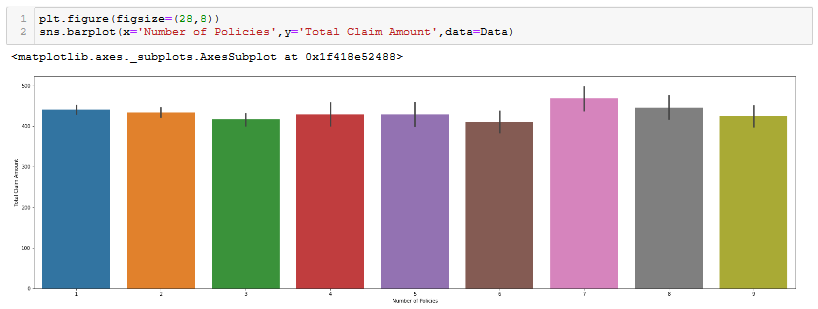


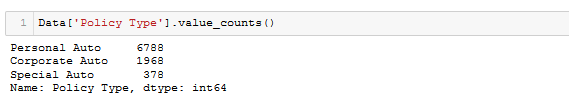


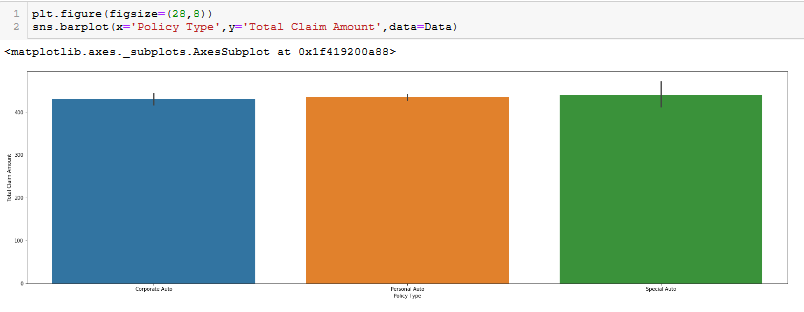


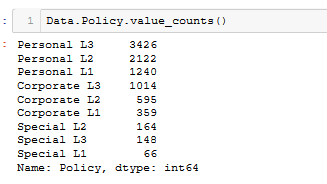


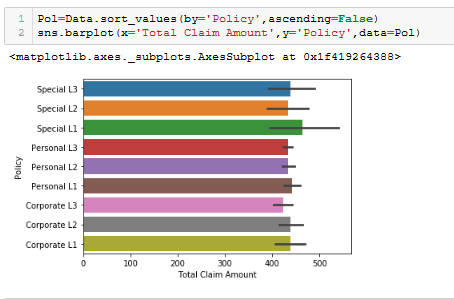


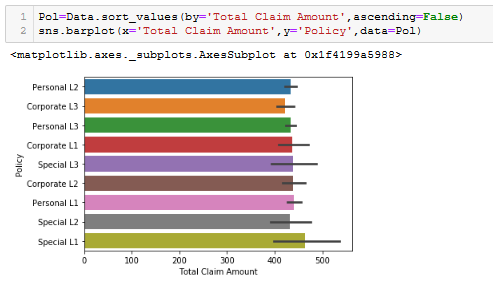


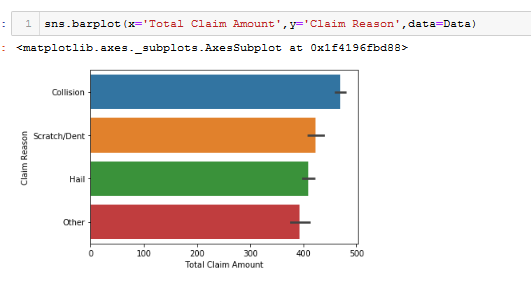


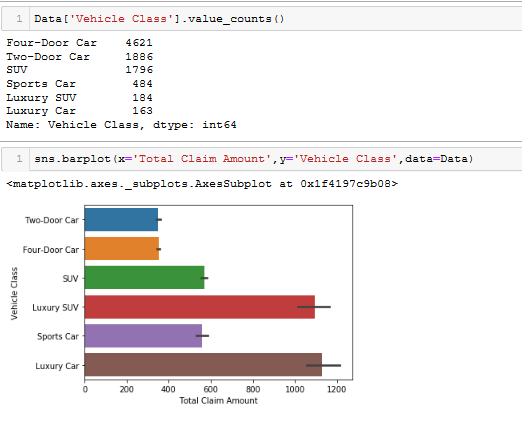


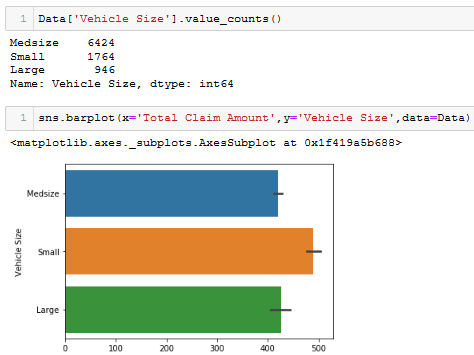












**Dropping The Redundant columns which are not required for machine learning.**



**Conclusion**

1. This is the Auto insurance data of USA only. As it is the dataset of only one country so we can drop the Country column.
2. Out of 5 States Missouri has highest number of insurance claims i.e. 3150 but Oklahoma has the Highest Total Claim Amount.
3. In the Insurance Company there are three types of Insurance Coverages are available in which Basic Coverage is very popular followed by Extended and Premium Coverage , out of 3 types of Coverages Premium type has highest Total Claim Amount followed by Extended Coverage and Basic type has lowest Total Claim Amount.
4. If we see insurance Effect to month wise or day wise Total claim amount, the total claims are averagely same and on Wednesday it is slightly peaked.
5. Peoples Insured under the company are different types of Education levels among them the peoples with low education like School or Below level have highest Total Claim Amount and highly Educated Peoples like Doctors have lowest Total Claim Amount.
6. From all the Employment types Unemployment people have highest Total Claim Amount compared to all employment type people.
7. Females have higher number of claims but Males Total Claim Amount is higher than Females.
8. If the Location is suburban Total claim amount is highest and in Rural Location it is lowest.
9. With increase in ‘Monthly Premium Auto’ ‘Total Claim amount’ increases.
10. Total Claim not depends on Months Since last Claim and Months since Policy Inception.
11. When the Number of Open Complaints are 5 Total claim Amount is highest.
12. Most of the peoples have taken Personal Auto Policy, followed by Corporate and very less people have taken Special Auto Policy. But Total Claim Amount is somehow more in the cases of Special Auto.
13. Three types of policy available in every Policy type and among them Special L1 has the highest Total Claim Amount.
14. Among the Reasons for insurance claim ‘Collision’ has highest Total Claim Amount in compared to other Reasons.
15. Total Claim amount is highest in case of Sales Channel is Agent.
16. From the all vehicle classes Luxury Car has the highest Total Claim Amount followed by Luxury SUV, SUV and Sports Car.
17. Small Vehicles has highest Total Claim Amount.

**Step-4**

**Pre-Processing**

Before feeding our data for Machine Learning it is also very very important to Pre-Process the data. Like

1. Handling Categorical variables.
2. Removing Outliers
3. Normalization.
4. Standardisation.

1.**Handling Categorical variables**- Machine can only understand numerical values not the categorical variables present in the dataset so converting categorical variables into its numerical representation is needed else we cannot perform Machine Learning. Now we encode the categorical values into their numerical representations by LabelEncoder

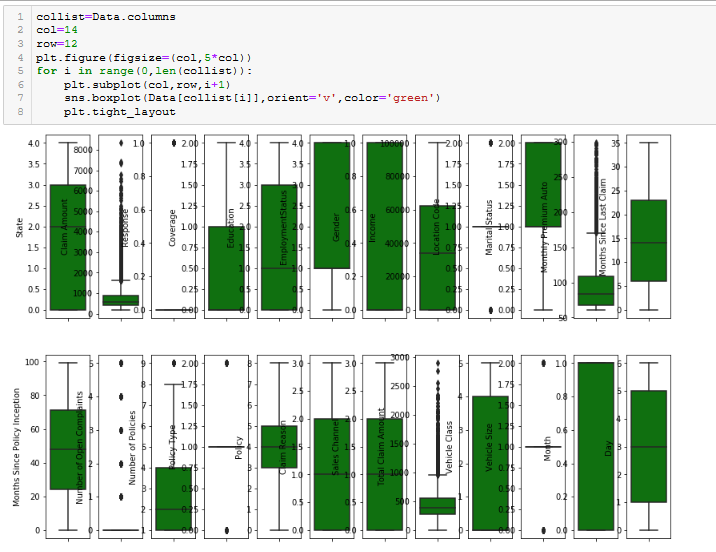
Let’s encode categorical features

list1=['State','Response','Coverage','Education','EmploymentStatus','Gender','Location Code','Marital Status','Policy Type','Policy','Claim Reason','Sales Channel','Vehicle Class','Vehicle Size','Month','Day']



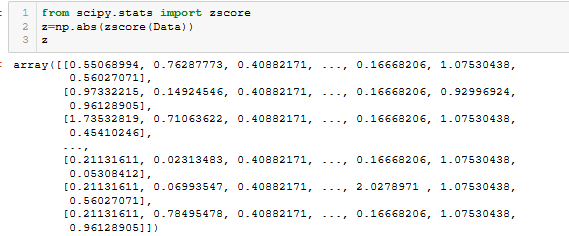
**2.Remove the outliers-** Outliers are the unusual values in the dataset with high variance and standard deviation. Presence of outliers can decrease the percentage of accuracy score of the prediction. So all the time removing outlier should be done.

Let’s do



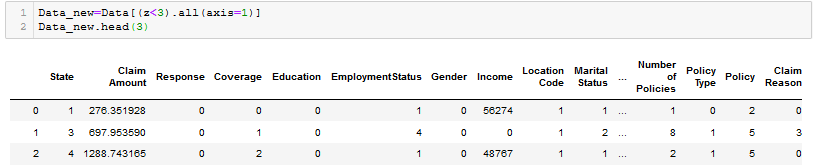
Yes, Outliers are present in ‘Response’, ‘Monthly Premium Auto’ and ‘Total Claim amount’

Removal of outlier can be done by zscore .So we import zscore from scipy

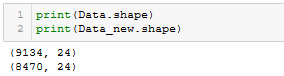


A positive z-score says the data point is above mean. A negative z-score says the data point is below mean. A z-score close to 0 says the data point is close to average. A data point can be considered unusual if its z-score is above 3 or below −3. Because almost full data i.e. 99.7 % of data falls under 3 standard deviation.

So we make threshold=3 and get Outlier free Data

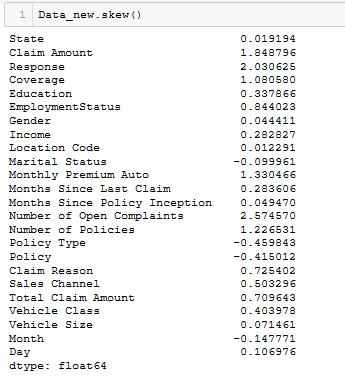


Check the shape and find what are the number of outliers are removed.



664 numbers of outlier observations are removed.

**3. Normalization-** First we check the distribution then if needed we do Normalization.



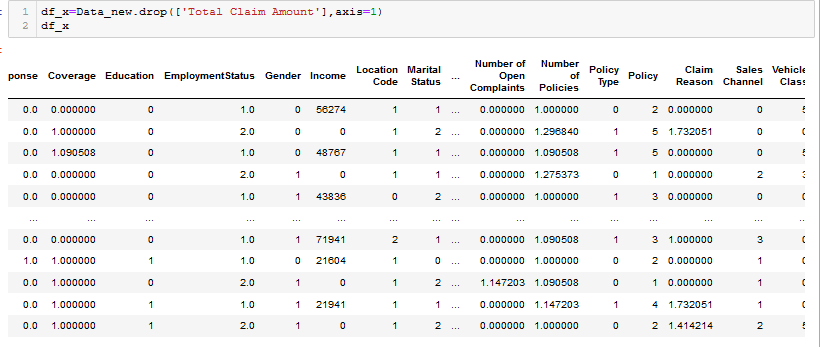
'Claim Amount', 'Response', 'Coverage', 'EmploymentStatus', 'Monthly Premium Auto', 'Number of Open Complaints', 'Number of Policies', 'Claim Reason', 'Total Claim Amount'] are right skewed , these features should be Normalized. There are three processes for making the distribution normal they are boxcox, log and square root transform (sqrt). Here we use sqrt transform for making the distribution normal.

list2=['Claim Amount','Response','Coverage','EmploymentStatus','Monthly Premium Auto','Number of Open Complaints','Number of Policies','Claim Reason','Total Claim Amount']

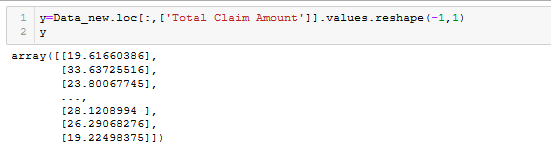


Now we split the dataset into Independent and dependent variable datasets and Standardise. Here the target is 'Total Claim Amount' means it is the dependent variable and other features are independent variables.

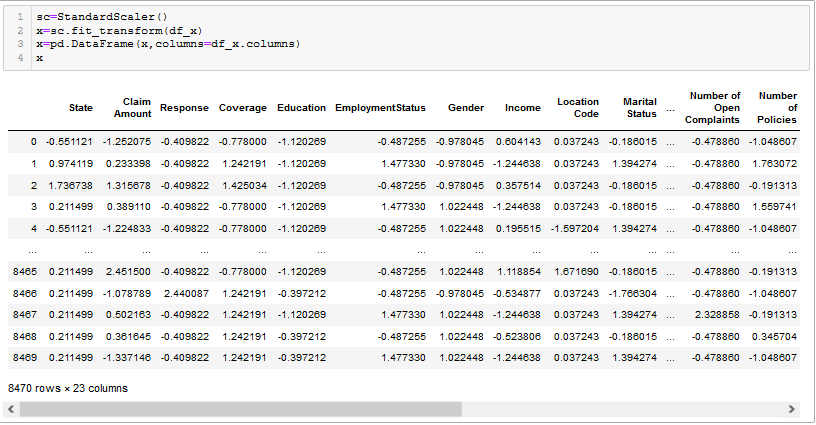
**Independent variable dataset**



**Dependent Variable**



4**. Standardisation-** It is the part of Pre-Processing in which features of dataset is standardised. If the features have mean of 0 and standard deviation 1, then they are standard else these data should to be standardised before Machine Learning. Let’s Standardise/scale the independent dataset by StandardScaler.

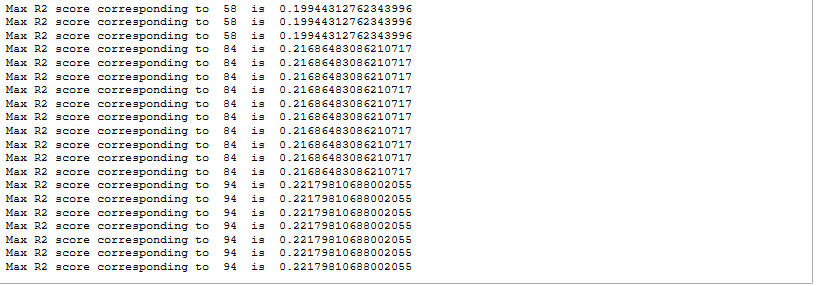
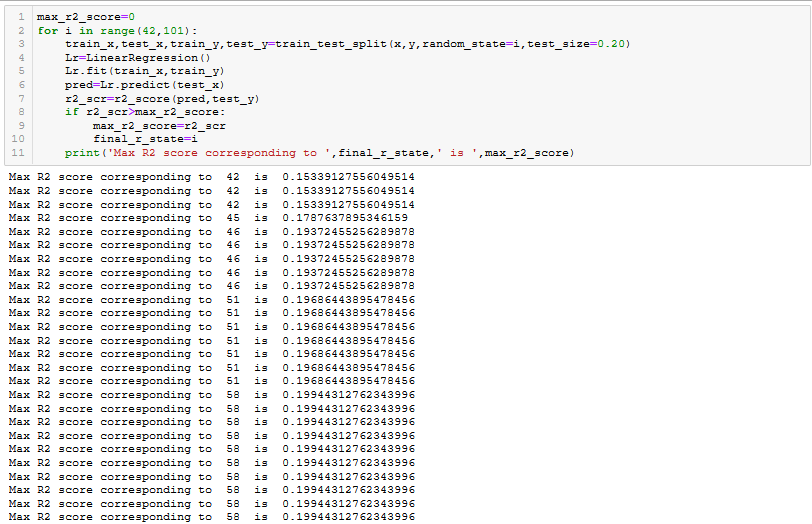


Now the independent and dependent datasets are ready to learn the Machine

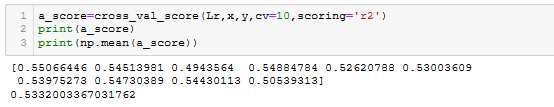
**Machine Learning**

It is the process to learn the machine so that machine should predict.

Now we will split the x and y datasets taking test size=20% and see which random\_state performs best result so that we take the best random state for final model.

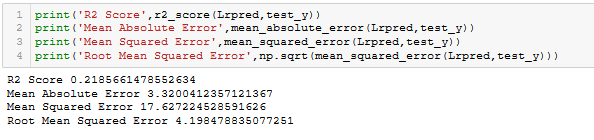
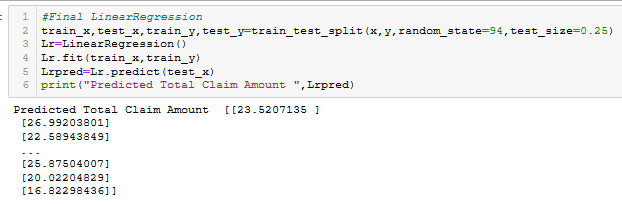


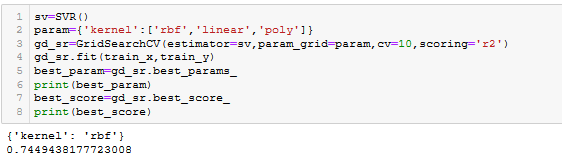
**Check if the model is over or underfitted**

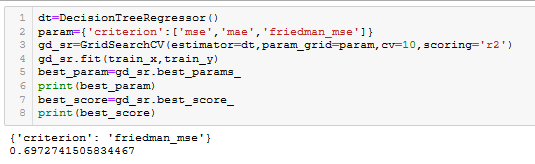


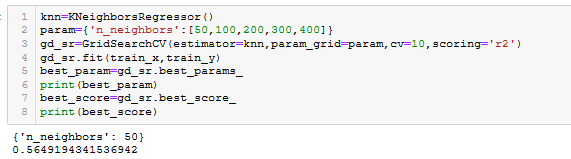
The upper model is underfitted.

Let’s perform Machine Learning Splitting the dataset again giving seed value 94 as the r2 score is highest when the r\_state is 94 through the possible Machine Learning algorithms and use the best model for Production.

**SupportVectorRegressor**

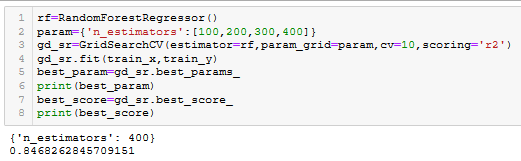
**DecisionTreeRegressor**

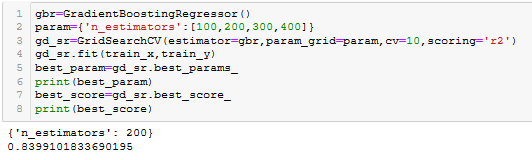
**KNeighborsRegressor**

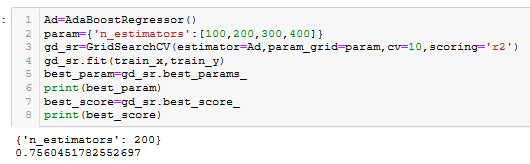


**Using Ensemble Techniques**

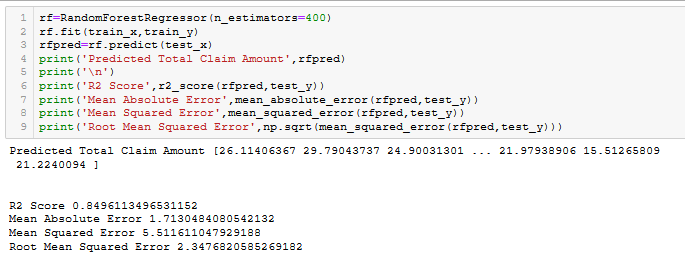
**RandomForestRegressor**

**GradientBoostingRegressor**

**AdaBoostRegressor**



**Final Model:**



**Save the Prediction:** Saving the prediction in a csv file can be done by df.to\_csv method.



**Save the model:** For production we can save our best model by 2 methods by pickle and joblib. Let’s save the model by importing joblib from sklearn.externals



**Conclusion:**

With RandomForestRegressor we have achieved highest r2\_score of 0.8468262845709151 (84.62%) as compared to other algorithms it is the best r2\_score when the n\_estimators=400. We know that with increase in accuracy Score of a prediction the percentage of errors are decreased so we make this model as final and save the prediction for future use and save the model for production by which more accurate Total Claim Amount can be predicted.

Thanks for Reading