Insurance Claims- Fraud Detection

**Introduction**

Insurance fraud is any act committed to defraud an insurance process.

This happens when a claimant endeavors to acquire some advantage or benefit they are not qualified for, or when an insurer purposely denies some advantage that is expected.

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 blog, a step-by-step technical approach has been provided on the data described for predicting which has the details of the insurance policy along with the customer details. It also has the details of the accident based on which the claims have been made. Here 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.

**The below is the list of columns provided to predict the insurance fraud claims :**

**months\_as\_customer:** This columns period for which a person was client to the insurance company

**age:** This column defines the age of the person

**policy\_number:** shows the policy number of the customer

**policy\_bind\_date:** shows the bind date of the policy

**policy\_state:** shows the state code in which policy was created

**policy\_csl:** shows the csl of the policy

**policy\_deductable:** shows the policy amount deductible

**policy\_annual\_premium:** shows the annual premium amount

**umbrella\_limit:** shows the premium amount umbrella limit

**insured\_zip:** shows the zip code of the person

**insured\_sex:** shows the gender of the person

**insured\_education\_level:** shows the education level of the person

**insured\_occupation:** Shows the occupation of the person

**insured\_hobbies:** shows the hobbies of an insured person

**insured\_relationship:** shows the relationship status of the insured person

**capital-gains:** shows the capital gains by the person

**capital-loss:** shows the capital loss by the person

**incident\_date:** shows the incident date

**incident\_type:** shows the incident type

**collision\_type:** Shows the collusion type

**incident\_severity:** shows the severity of the incident

**authorities\_contacted:** shows the authorities contacted

**incident\_state:** shows the state of the incident

**incident\_city:** shows the incident city

**incident\_location:** shows the incident location

**incident\_hour\_of\_the\_day:** shows the incident hours of the day

**number\_of\_vehicles\_involved:** Shows the number of vehicles Involved

**property\_damage:** shows the amount of property damage

**bodily\_injuries:** shows the bodily injuries that occurred

**witnesses:** shows the number of witnesses for the incident

**police\_report\_available:** shows the police report Is available or not

**total\_claim\_amount:** show the total claim amount

**injury\_claim:** shows the injury claims

**property\_claim:** shows the property damage claim

**vehicle\_claim:** shows the vehicle damage claim

**auto\_make:** shows the make (brand manufacturer ) of the vehicle

**auto\_model:** shows the model number/name of the vehicle

**auto\_year:** shows the manufacture year of the vehicle

**fraud\_reported:** shows if the fraud is reported or not

**\_c39:** This column has several null values and can be dropped during EDA

**Process Steps**

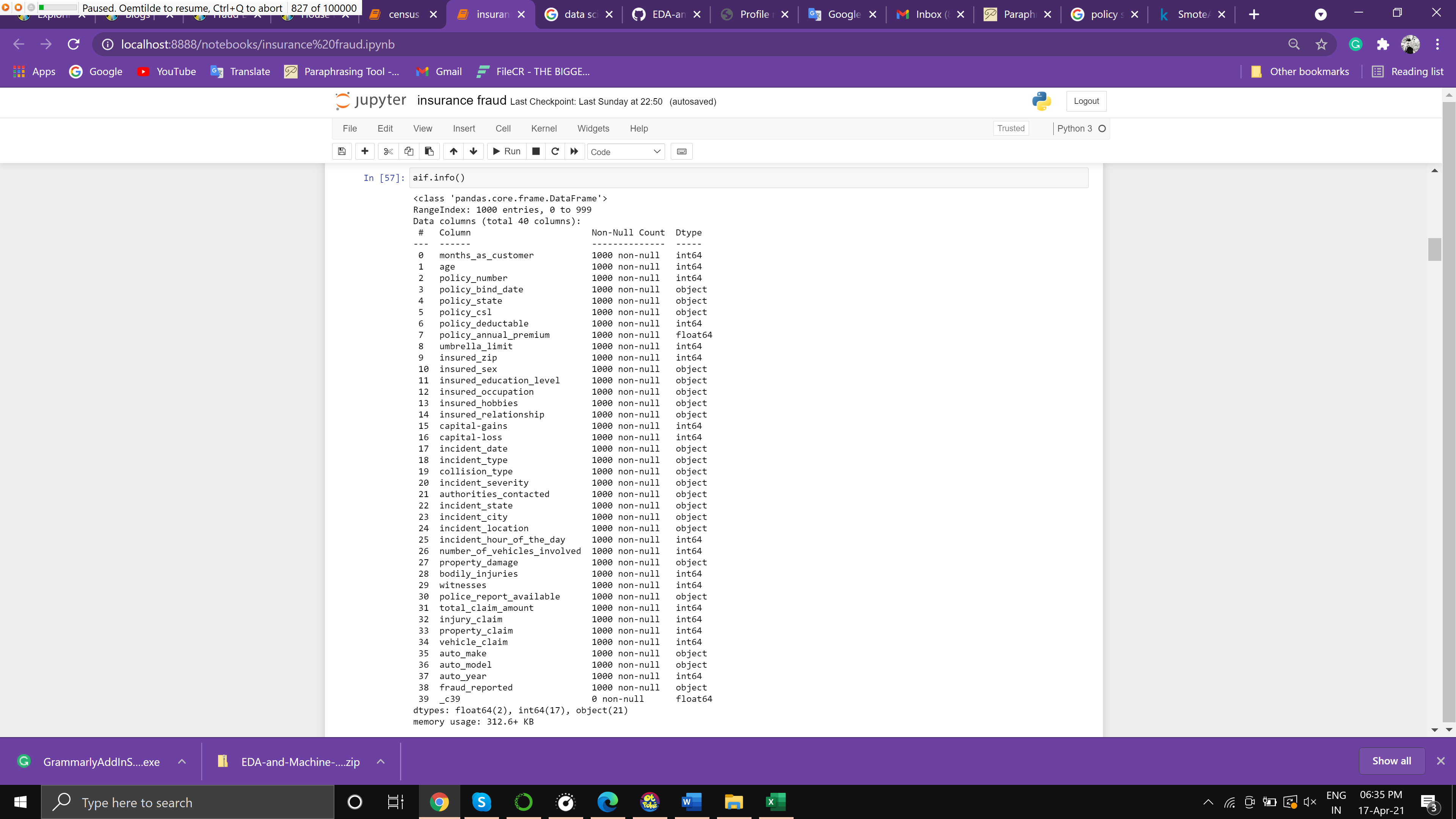
The entire process of machine learning can be divided into 4 main steps to get the desired prediction.

* Getting Data – Data can be collected from the source in this step for exploratory data analysis and visualization for understanding the current/historic data and determine the next step.
* Data Pre-Processing – Pre-Processing aka data wrangling is the technique of cleaning and transforming the raw data (which can be incomplete and inconsistent) to a proper format to be used for modeling.
* Model Evaluation – Different machine learning algorithms can be used and evaluated in this step to measure the accuracy and other confusion metrics.
* Prediction – Depending on the choice of the best model, prediction is done.

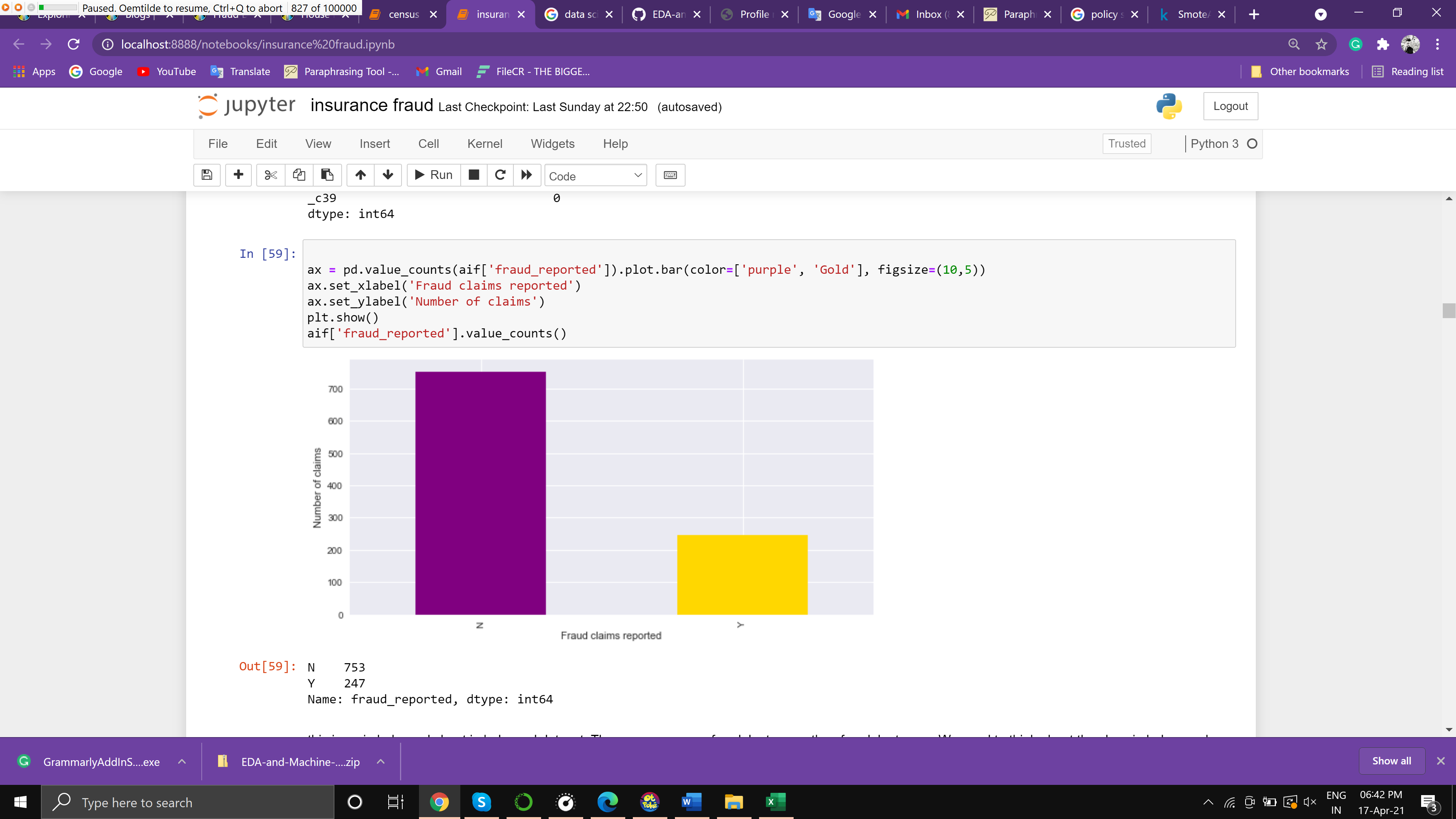
**Exploratory data analysis**

Exploratory Data Analysis refers to the critical process of performing initial investigations on data to discover patterns, spot anomalies, test hypotheses, and check assumptions with the help of summary statistics and graphical representations.

As the data has been imported, the information about the data has been checked.



The above chart shows that there are a total of 40 rows and none of them have Null/ Nan values except for one which will be removed at a later stage as it has all nan values, so dropping is a better option.



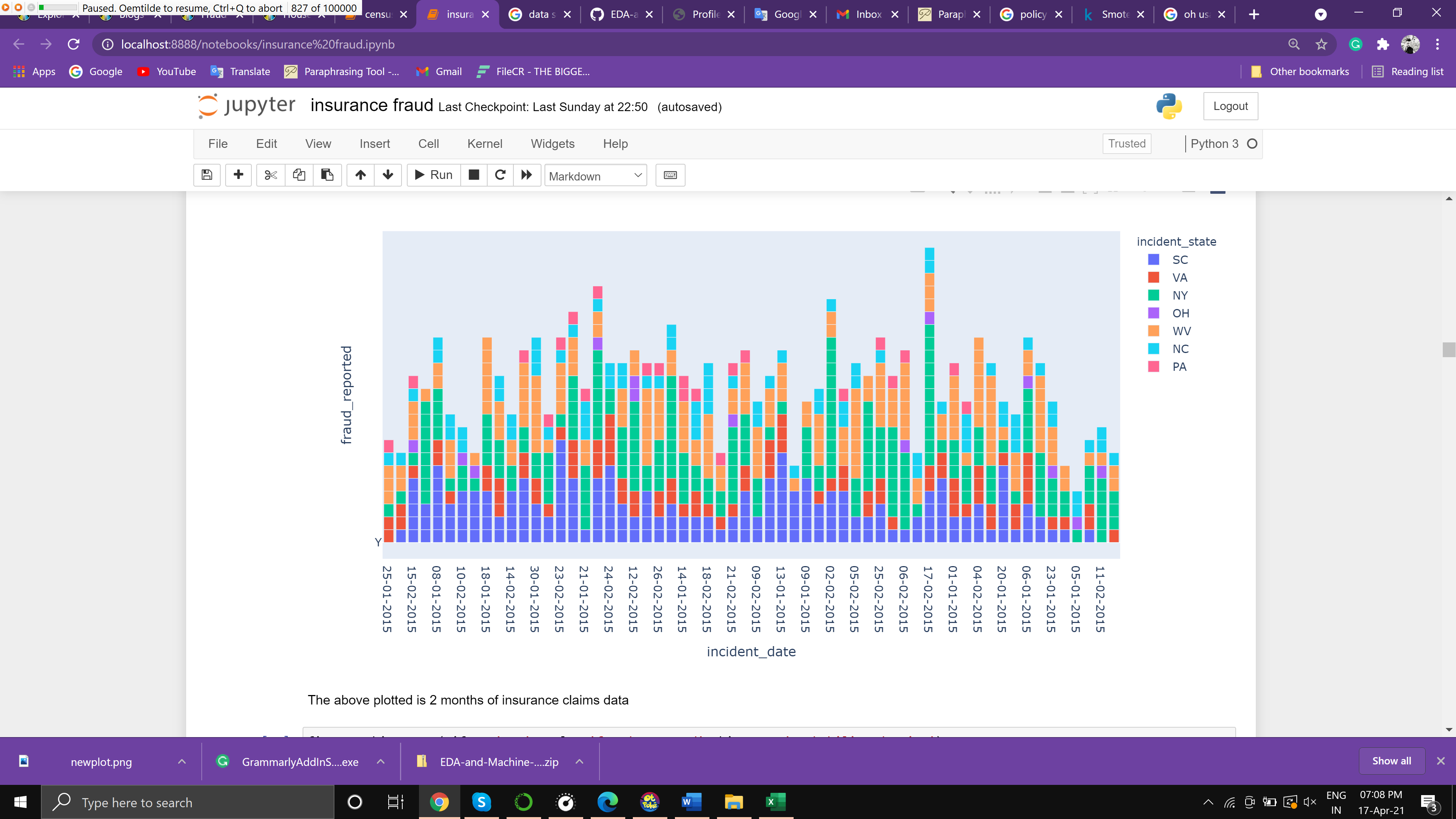
The above figure shows the number of claims and out of those claims the numbers who are not Fraud and the Fraud ones

As per the figure around 250 people are detected as Fraud and others are not Fraud.

That is around 25% of cases are Fraud.



The above image shows the total Claim amount based on state/ City and as per the observation most of the claims are from New York, South Carolina, and West Virginia claiming > 10M $



The above image shows the plotting of 2-month data which has color based representation of the number of claims from each state/city

Where the initials refer to -

SC: South Carolina

VA: Virginia

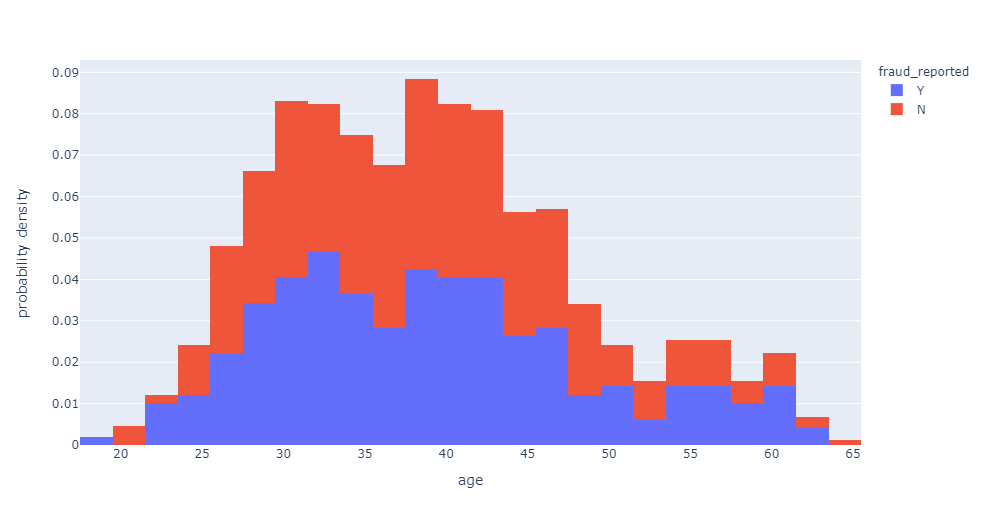
NY: New York

OH: Ohio

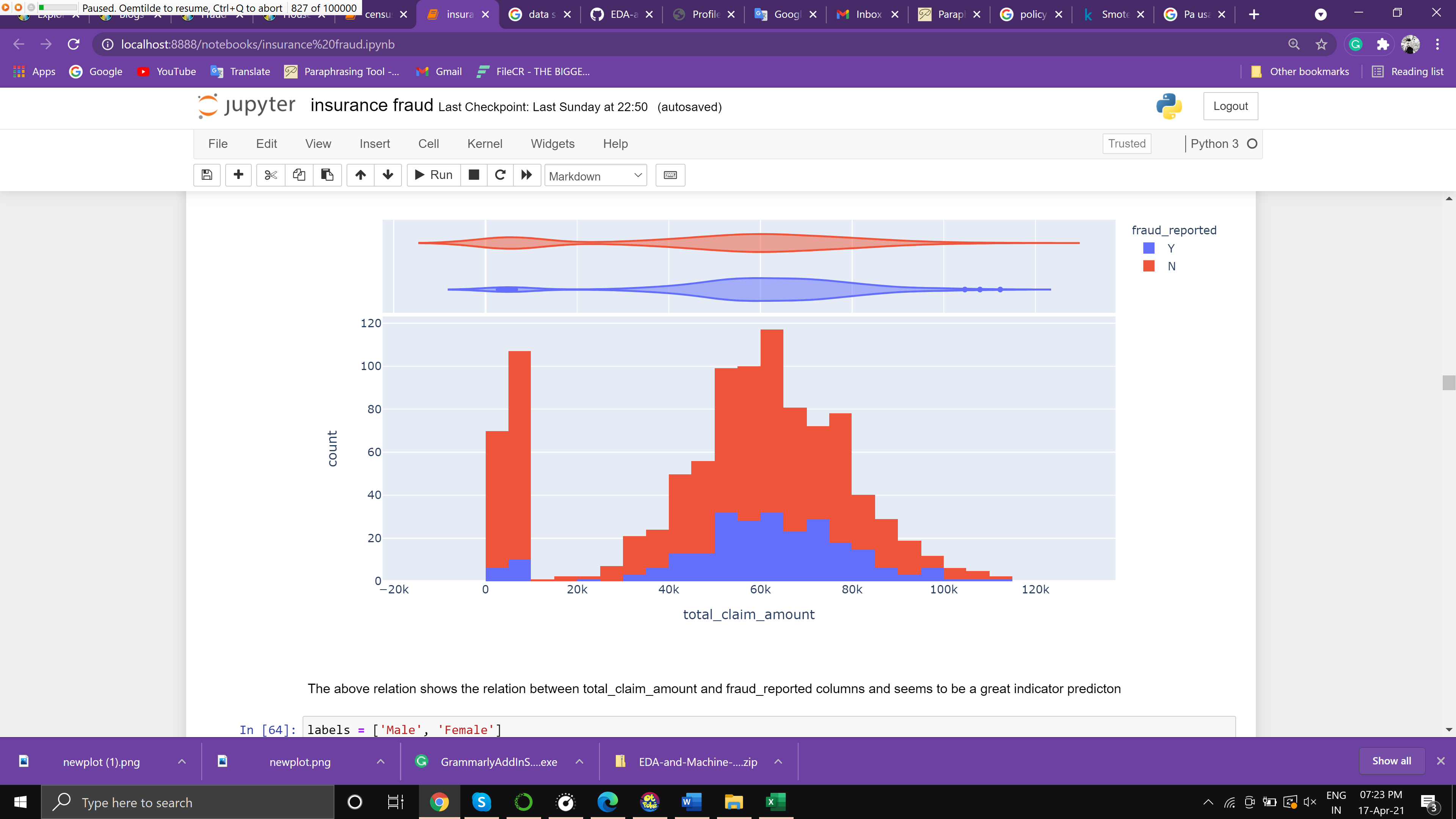
WV: West Virginia

NC: North Carolina

PA: Pennsylvania



The above plot shows the number Probability density of the age whose Fraud was reported. This shows that people from age 25 to 50 have more chances to conduct Fraud claims as compared to other age groups also Age doesn't seem to be a good indicator of fraud detection.

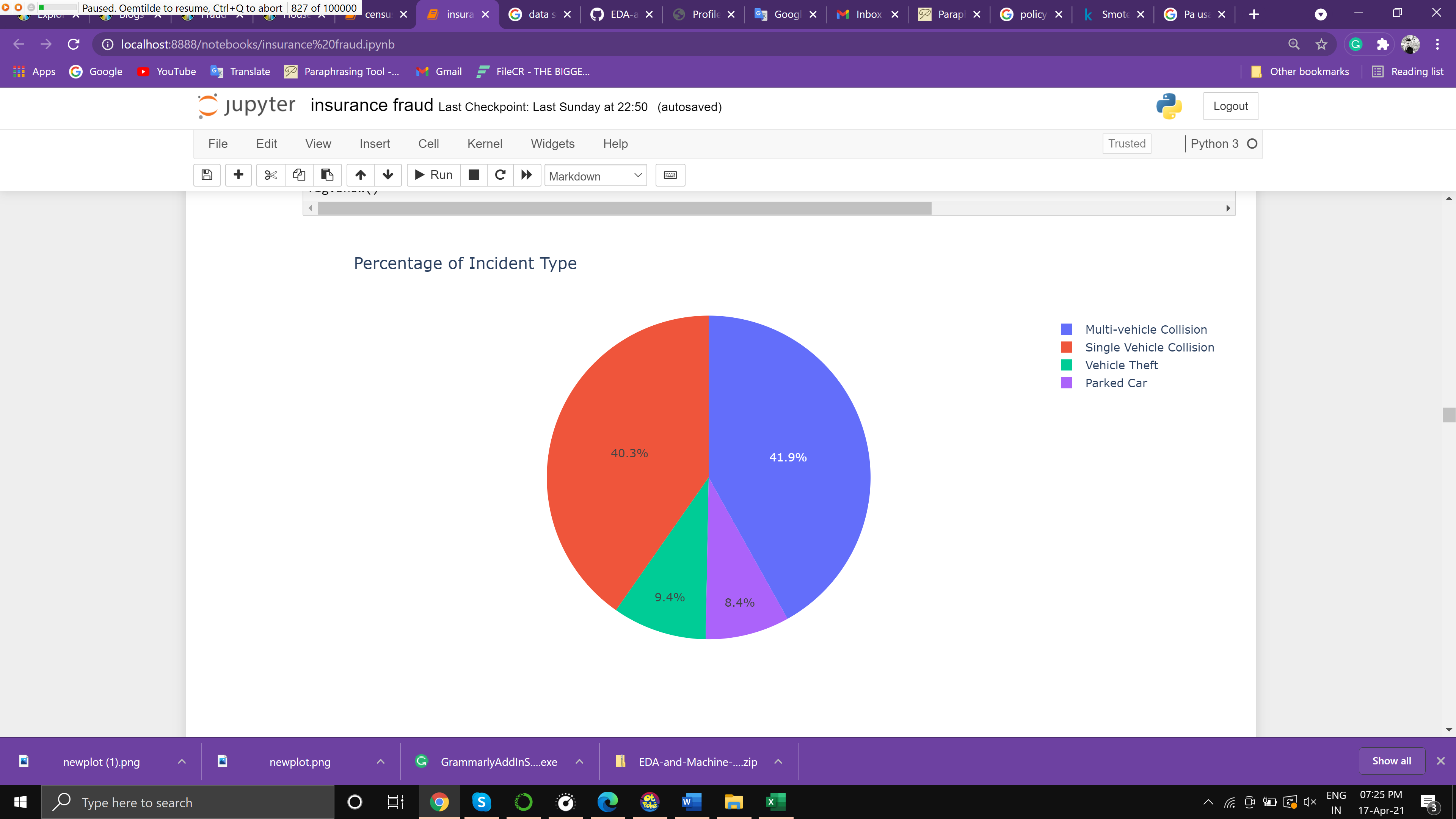


The above figure shows the count of the claims and the amount being claimed.

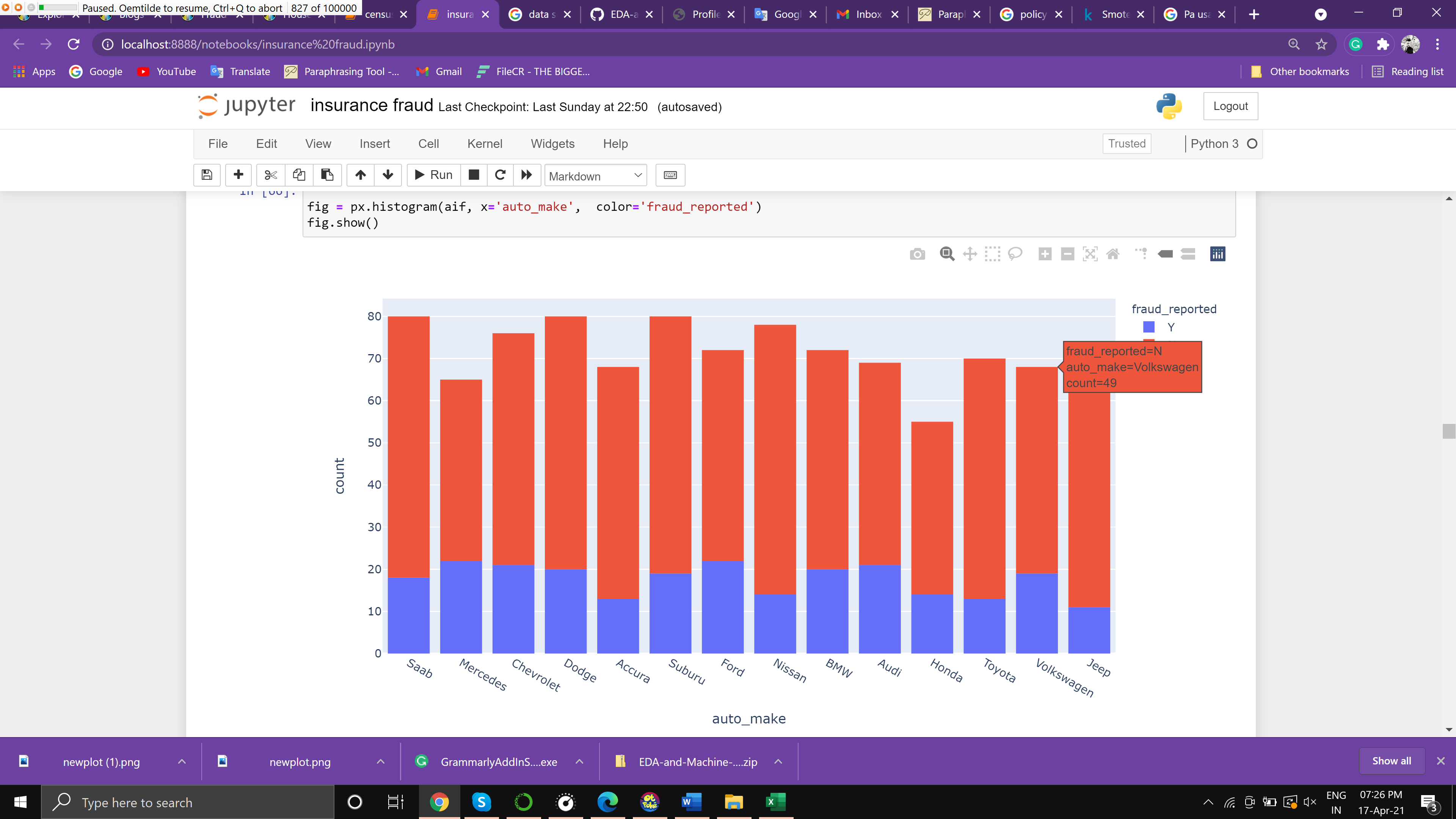
Its been observed that the maximum amount claimed is between 40k to 80k out of which 25 % are fraud reported.



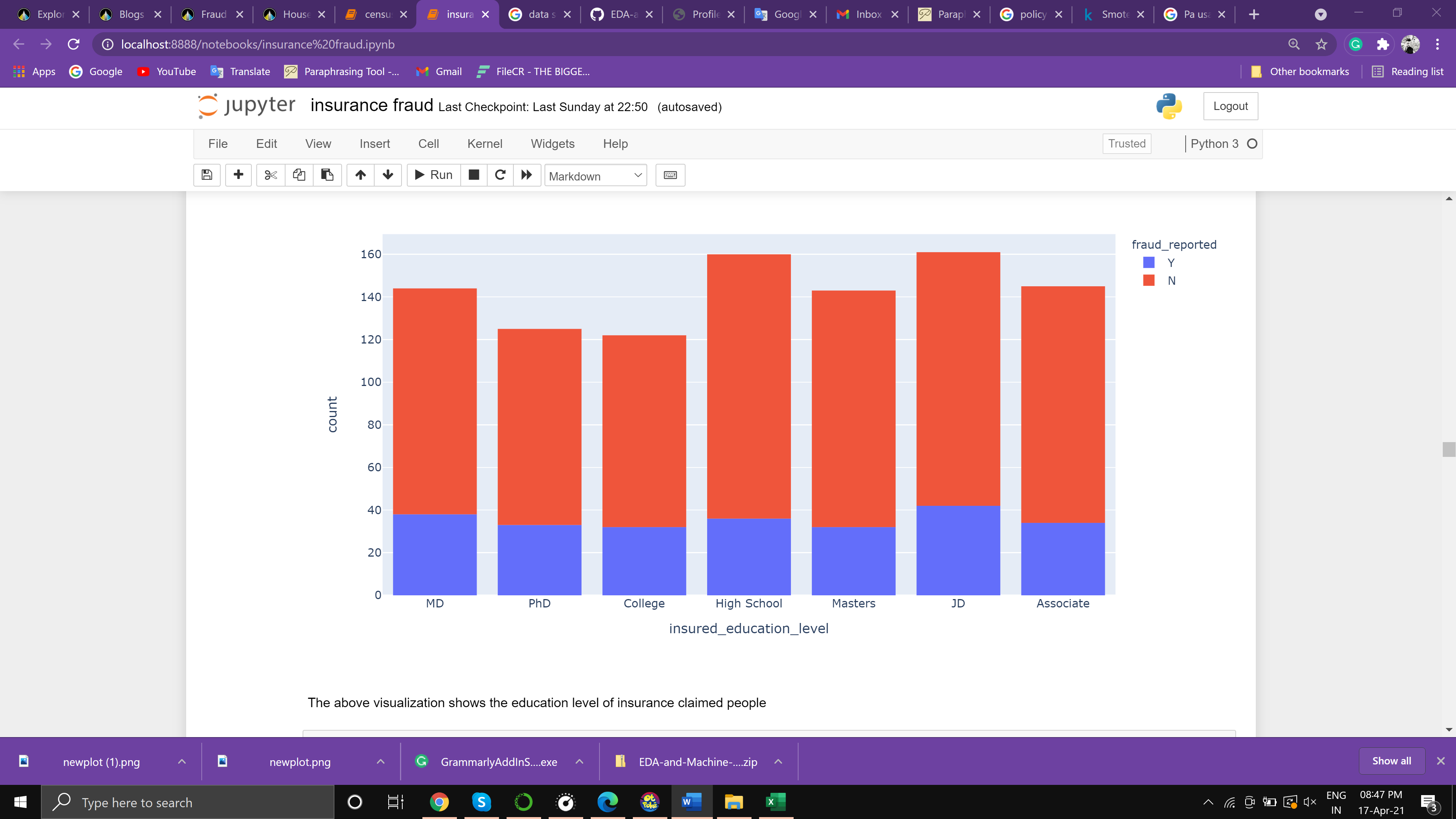
The above figure shows the percentage of males and females in the datasheet, there is 53.7% of males to claim insurance. It is not a great indicator for prediction but we can still use it.



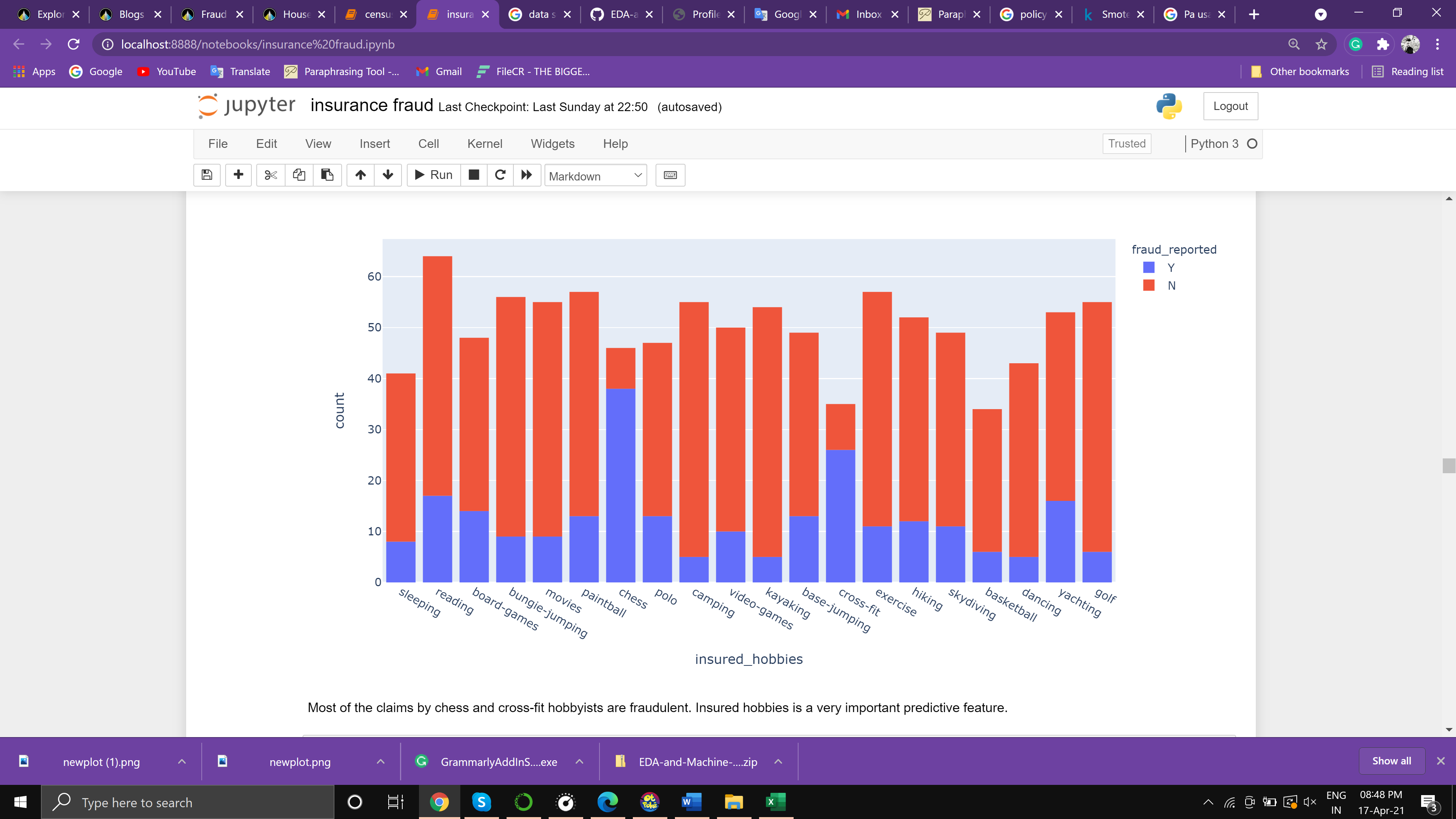
The above figure shows the % of different types of incidents, as the vehicle can meet an accident in many different ways, amongst which the top is Multi-vehicle collision at 41% and single-vehicle collision at 40% this is a great indicator for prediction.



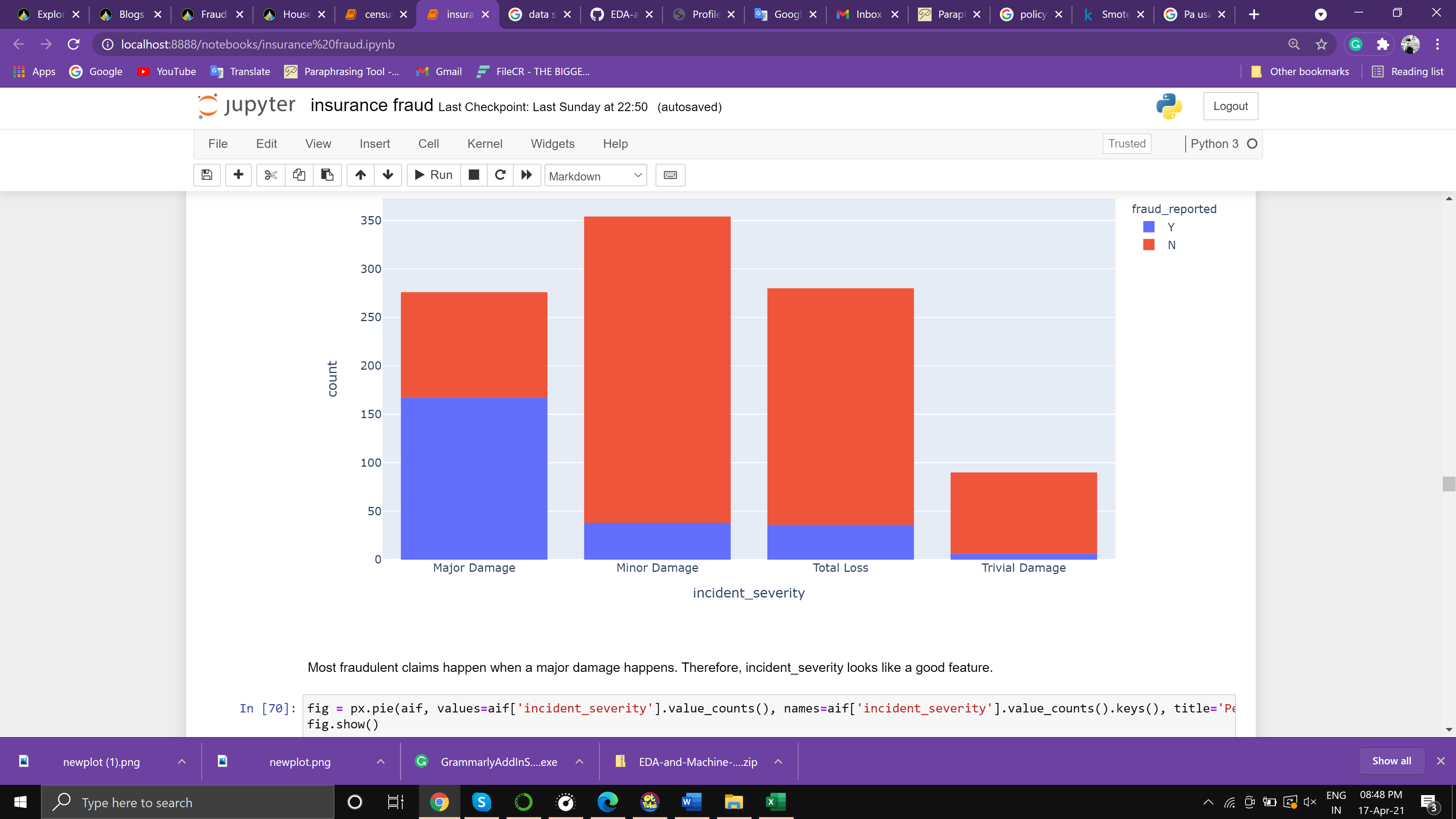
The above figure shows the bar graph distribution of Auto makes based on Fraud detected, unfortunately, this doesn’t prove much helpful to predict as there is not some major evidence



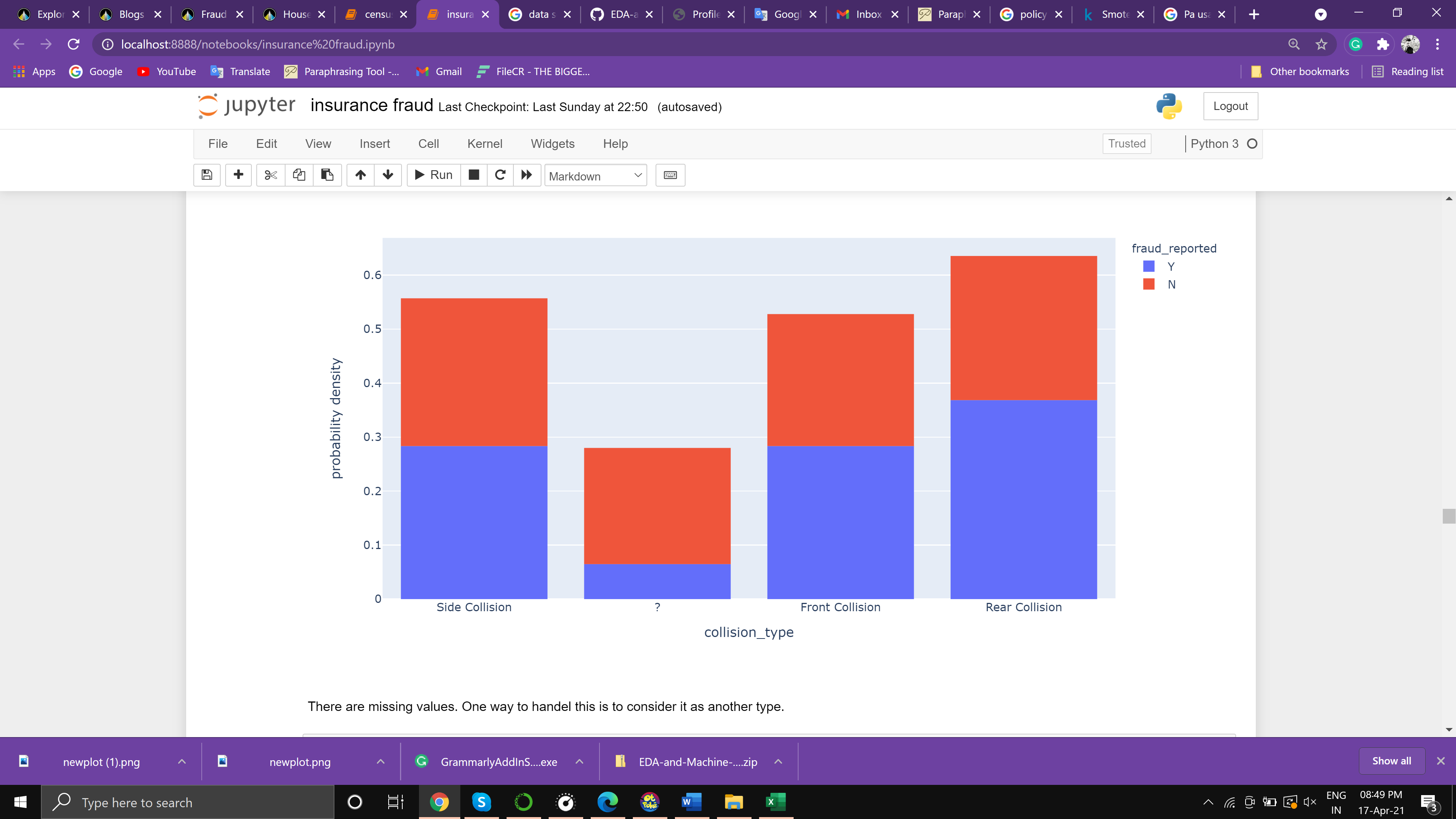
The above Figure shows the bar graph distribution of peoples Fraud detected based on education level, unfortunately, this also doesn’t prove much helpful to predict as there is not some major evidence



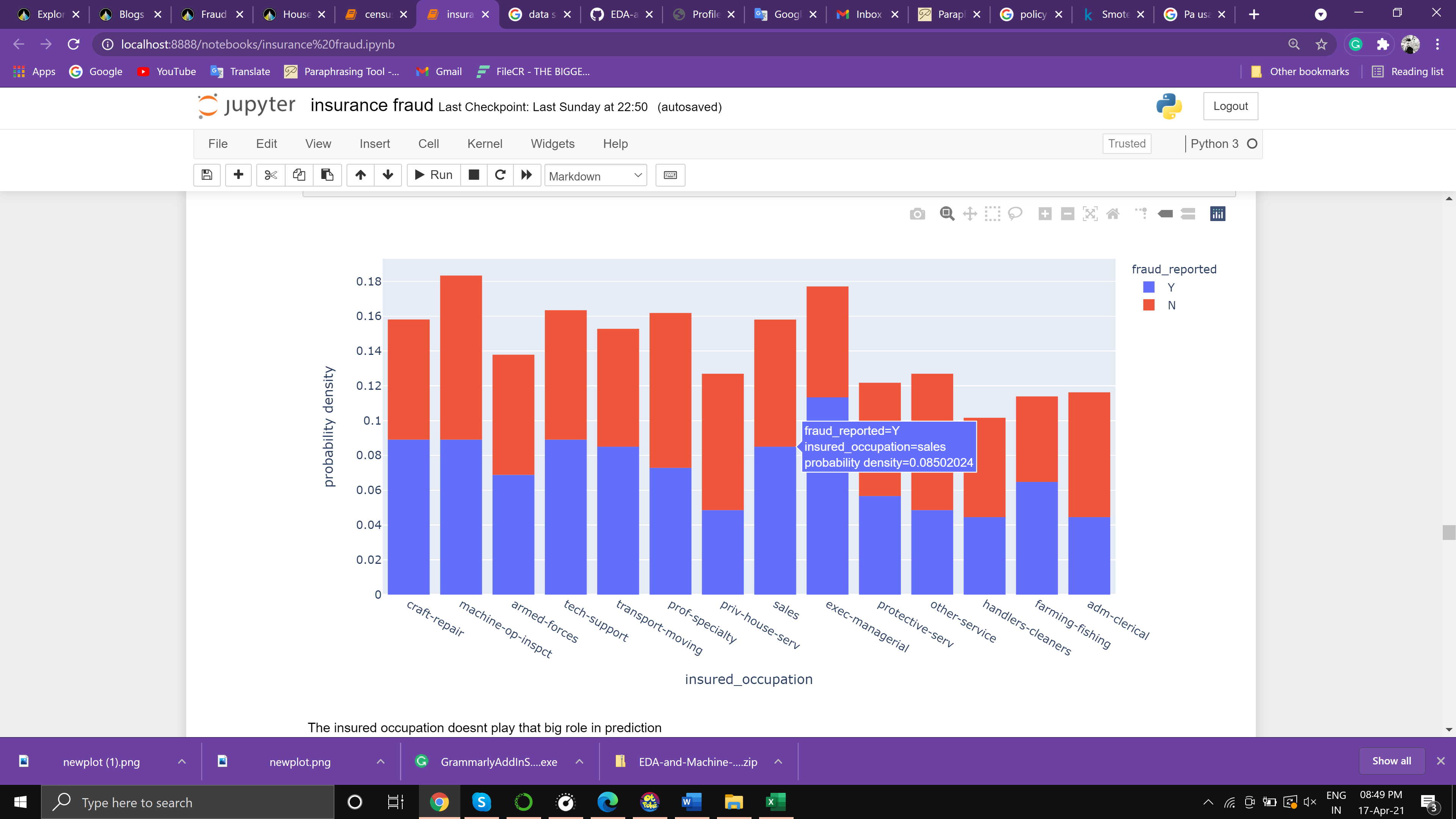
The above Figure shows the bar graph distribution of peoples Fraud detected based on hobbies, the maximum fraud is reported against persons with chess and CrossFit this proves much help to predict as there is some major evidence.



The above figure shows the count of different types of severity as the vehicle can meet an accident in many different ways, amongst which the top is major damage at which is above 150 this is a great indicator for prediction.



The above bar graph shows the collision type also we can see that some values are missing, which indicates that we need to work on data before making a model, the rear collision is the widely used collision type and has a max number of fraud reports too



The above Figure shows the bar graph distribution of peoples Fraud detected based on occupation, unfortunately, this also doesn’t prove much help to predict as there is not some major evidence

The columns "policy\_number", "policy\_bind\_date", "insured\_zip", "incident\_location", "incident\_date" and "\_c39" are been dropped as they are not much helpful for prediction.

Outlier analysis

Outlier Analysis is a process that involves identifying the anomalous observation in the dataset.” In some instances like fraud detection, the outlier indicates a fraudulent activity. Outlier Analysis is a data mining task which is referred to as an “outlier mining”.

Skewness refers to a distortion or asymmetry that deviates from the symmetrical bell curve, or [normal distribution](https://www.investopedia.com/terms/n/normaldistribution.asp), in a set of data. If the curve is shifted to the left or the right, it is said to be skewed. Skewness can be quantified as a representation of the extent to which a given distribution varies from a normal distribution. A normal distribution has a skew of zero, while a [lognormal distribution](https://www.investopedia.com/articles/investing/102014/lognormal-and-normal-distribution.asp)

The outliers are removed based on a threshold of 0.97 and not much skewness has been observed

Model Evaluation

Model selection and evaluation is a critical step of any machine learning project as identifying the pattern and applying the correct algorithm is not a very easy process. Machine Learning provides multiple models to generalize it to the unseen data from the same population and measuring the performance. Along with meeting the business objective, the model should take care of accuracy, execution time, complexity, and scalability as well to be considered as the best model. Sometimes the size of the training data set and numbers of predictor features can be decision-making criteria for model selection. Cross-validation AUC score is the performance metric for this project.

As this is a classification project, classifier algorithms can be used to predict results. we are using for loop to run 4 different algorithms i.e

* Linear Discriminant Analysis
* Decision Tree
* Random forest
* Logistic Regression

The performance will be measured based on a cross-validation AUC score

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Logistic Regression

Best Parameters: {'model\_\_C': 1.5}

Best Score: 0.5941438864166138

ROC AUC: 0.5655

Confusion Matrix:

[[76 75]

[23 26]]

precision\_score: 0.2574

recall\_score: 0.5306

f1\_score: 0.3467

Cross-validation AUC score: 0.5921950317610489

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Decision Tree

Best Parameters: {'model\_\_max\_depth': 3, 'model\_\_min\_samples\_split': 5}

Best Score: 0.8338993960584871

ROC AUC: 0.8241

Confusion Matrix:

[[126 25]

[ 9 40]]

precision\_score: 0.6154

recall\_score: 0.8163

f1\_score: 0.7018

Cross-validation AUC score: 0.8565419128711087

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Random Forest

Best Parameters: {'model\_\_max\_depth': 20, 'model\_\_min\_samples\_split': 5}

Best Score: 0.8543653316380588

ROC AUC: 0.8194

Confusion Matrix:

[[133 18]

[ 25 24]]

precision\_score: 0.5714

recall\_score: 0.4898

f1\_score: 0.5275

Cross-validation AUC score: 0.8421037906023336

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Linear Discriminant Analysis

Best Parameters: {}

Best Score: 0.8782599862258953

ROC AUC: 0.8563

Confusion Matrix:

[[135 16]

[ 13 36]]

precision\_score: 0.6923

recall\_score: 0.7347

f1\_score: 0.7129

Cross-validation AUC score: 0.8704251205117808

As we can observe above, that 4 four different algorithms are been used out of which Linear Discriminant Analysis performs better than all

Cross-validation AUC score: 0.8704251205117808

As the Decision tree performs slightly less than the Linear Discriminant Analysis so we consider Linear Discriminant Analysis for hyperparameter tuning.

Hyperparameter tuning

In machine learning, hyperparameter optimization or tuning is the problem of choosing a set of optimal hyperparameters for a learning algorithm. A hyperparameter is a parameter whose value is used to control the learning process.

It was observed performance of the Linear Discriminant Analysis has been slightly increased after tuning. so we decided to save the model tuning.

Saving model

Model progress can be saved during and after training. This means a model can resume where it left off and avoid long training times. Saving also means you can share your model and others can recreate your work.

The model is been saved with the Object of gridsearchcv after fitting with train data, to reuse it in further prediction.

Conclusion

After cleaning the data, we ran some initial data analysis before we moved on to feature engineering. After we analyzed the dataset, we were able to come to a few conclusions. Most of the claims are from New York, South Carolina, and West Virginia claiming > 10M $.people from age 25 to 50 have more chances to conduct Fraud claims as compared to other age groups. The maximum amount claimed is between 40k to 80k out of which 25 % are fraud reported. Here is 53.7% of males claim insurance. Multi-vehicle collision at 41% and single-vehicle collision at 40% this is a great indicator for prediction. The vehicle can meet an accident in many different ways, amongst which the top is major damage at which is above 150 this is a great indicator for prediction. The rear collision is the widely used collision type and has a max number of fraud reports too. The maximum fraud is reported against persons with chess and CrossFit this proves much help to predict as there is some major evidence. The Linear Discriminant Analysis performs better than other algorithms so we used it for model saving after performing hyperparameter tuning