**Implementation:**

We used a total of 4 datasets. Out of which 3 were from Kaggle and 1 from Washington post.

The first dataset has the below columns:

| **Column Names** | **Description** |
| --- | --- |
| id | Unique ID of each incident |
| name | Name of the person |
| date | Date of incident |
| manner\_of\_death | How were they killed/died |
| armed | Were they armed ? |
| age | Age of victim |
| gender | Gender |
| race | What’s their race ? |
| city | City of incident. |
| state | State in which the incident took place |
| signs\_of\_mental\_illness | Where they mentally ill |
| threat\_level | Were they attacking ? |
| flee | Were they fleeing ? |
| body\_camera | Did the police official have a body cam ? |
| longitude | Longitude of the incident. |
| latitude | Latitude of the incident. |
| is\_geocoding\_exact | Is it the exact location ? |

Table 1: dataset-1 description

**##write about the other 3 data sets.**

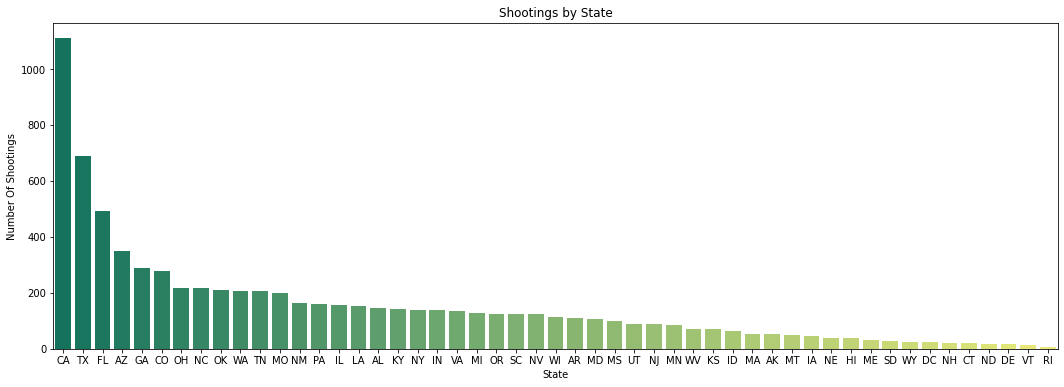
**DATASET VISUALIZATION:**

Fig: shootings by state

The above figure is the plot between the state and the number of shootings that happened in that state. The highest number of incidents happened in CA. The least happened in Rhode Island.

**#ALL OTHER VISUALIZATIONS……**

**CENSUS DATA:**

We requested for the API key from the census and got the key, thus we used the state wise and race wise data from the census and used it to perform various visualizations.

| **url = 'https://api.census.gov/data/2020/acs/acs5?'**  **params = {'get' : 'NAME,B01001\_001E,B02001\_004E,B02001\_005E,B01001H\_001E,B01001B\_001E,B01001I\_001E',**  **'for' : 'state:\*'}**  **r = requests.get(url, params=params)** |
| --- |

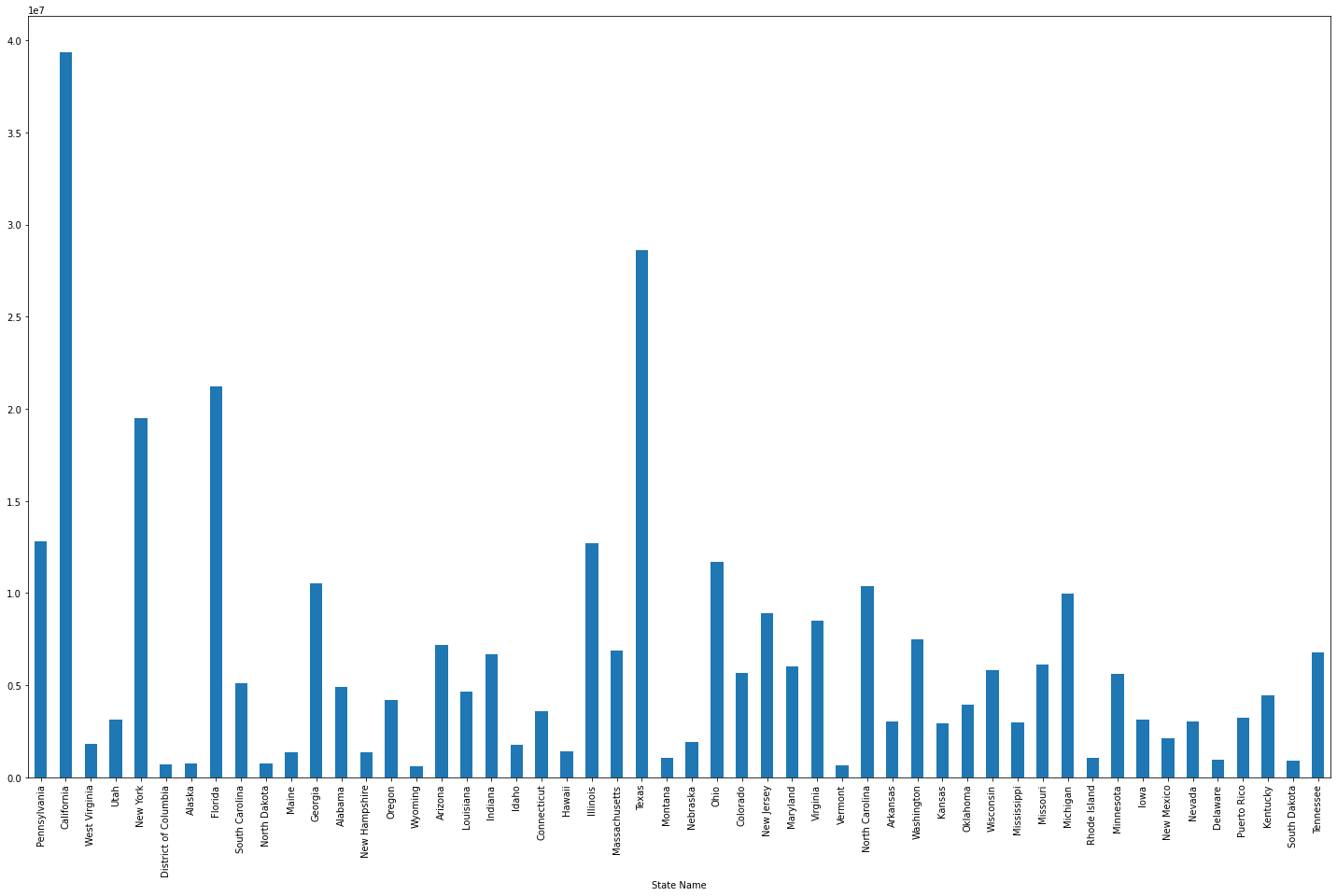
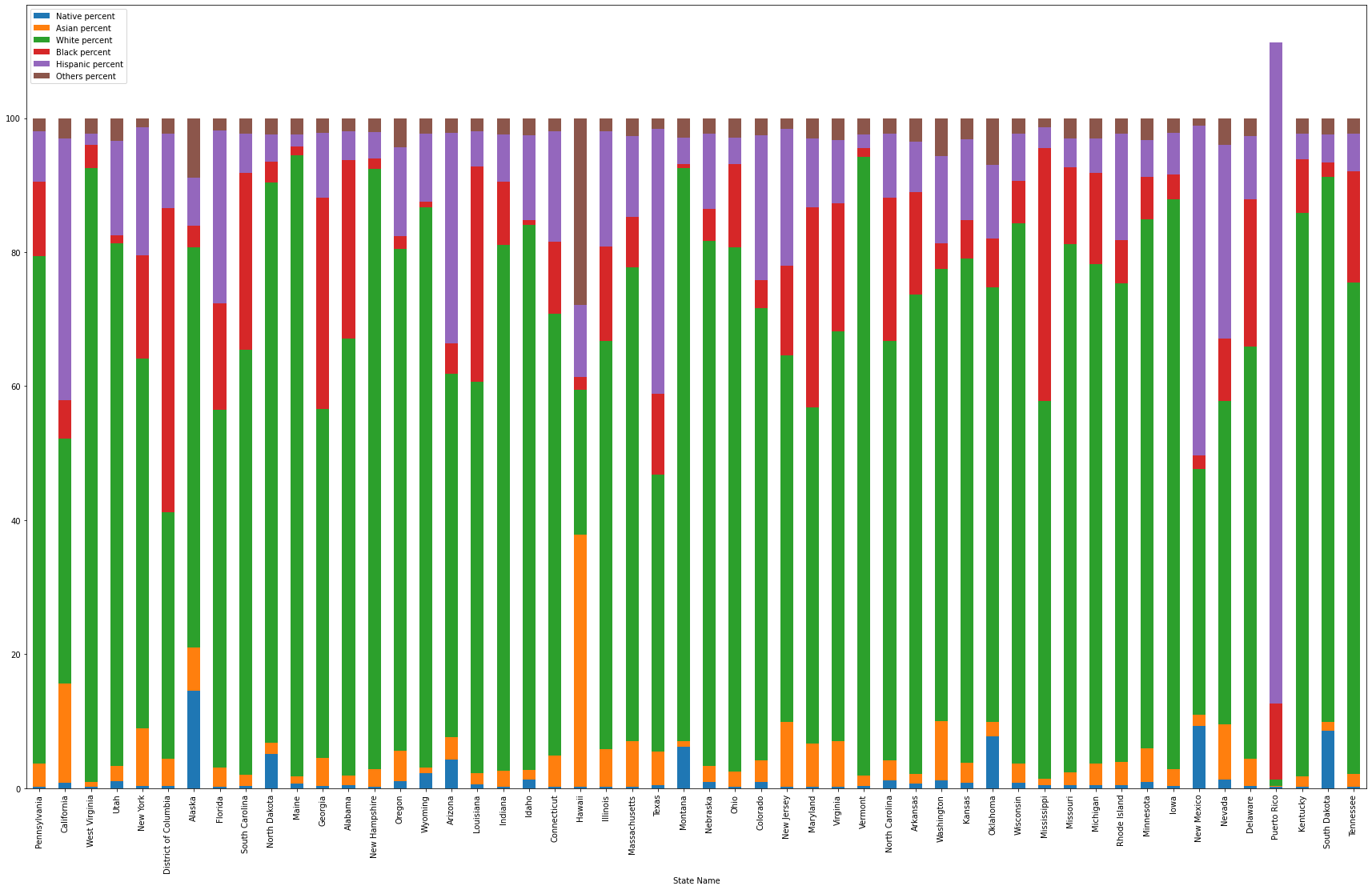
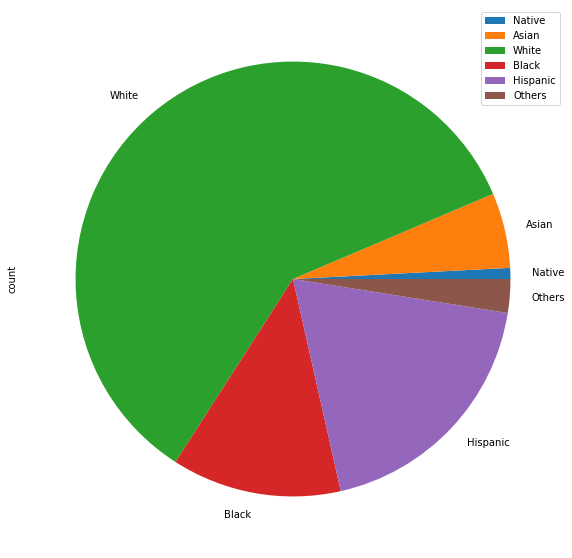
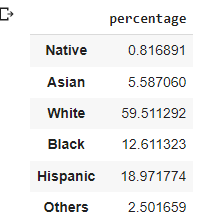
****

Fig: State name vs total population

The above plot is plotted from the data we took from the census. It is clear that the most populated state is California, this explains us the reason behind numbers in the various plots.

Fig: state vs percentage of people race wise

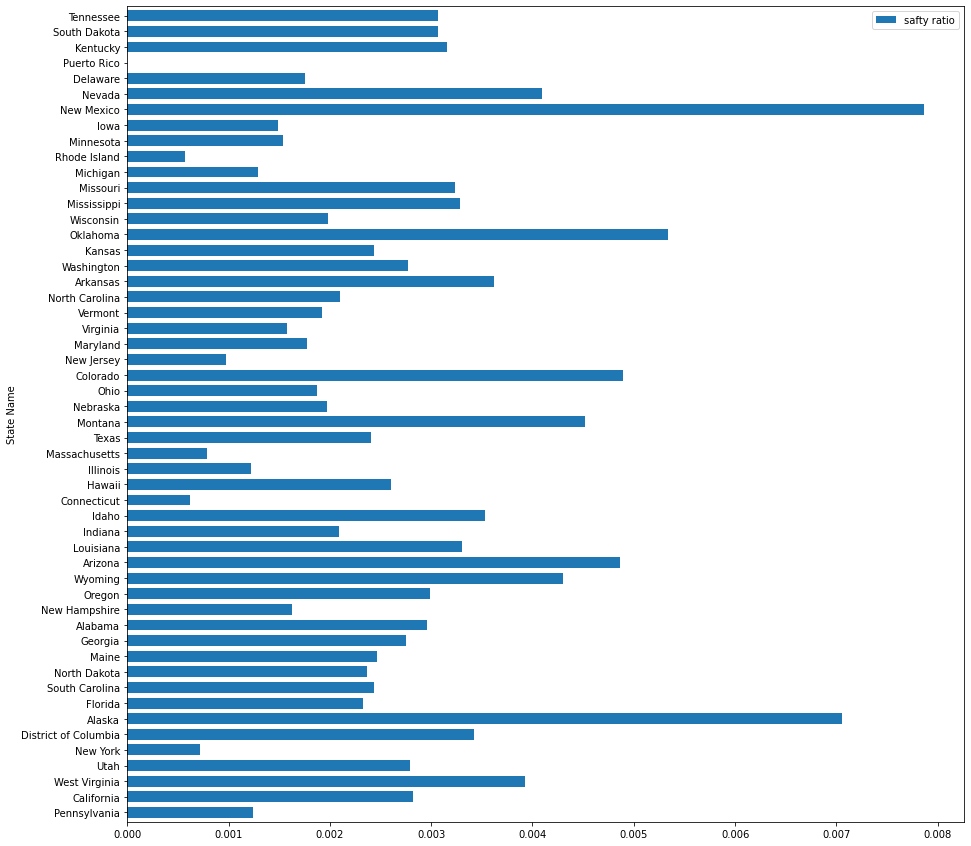
From the above plot it’s clear that most of the states have white people more in number.



Percentage of people by race.

Going further, comparing the number of incidents in each state to the population of the state to know the actual percentage of the population that get involved in the crimes.

For example, the total population of **Pennsylvania is** 12794885. The number of incidents that happened in that state as per our dataset are 159. So, the rate becomes 0.001243.

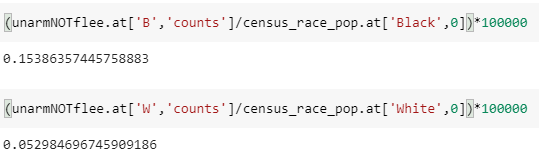


From the above plot it's clear that, even though the number of incidents are more in california the chances for a citizen getting involved is less.

New Mexico is the most unsafe.

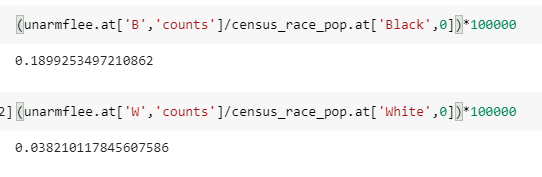
Took our original dataset and compared it with the census race wise population to know the chances of a person from black race getting killed compared to that of white in various scenarios.

**person is not fleeing and unarmed:**

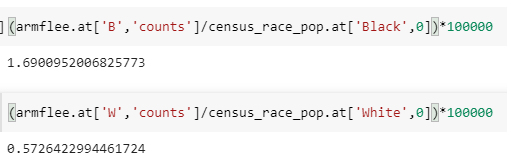
****

This is for the case where the person is not fleeing and unarmed, in this case the percentage of a person from **Black race getting killed is 3 times more than a white person getting killed**.

**Person fleeing and unarmed:**

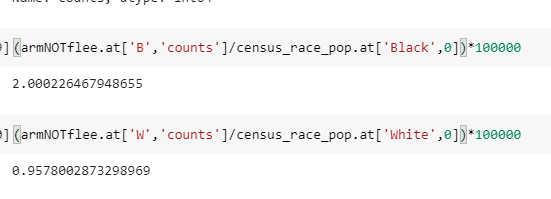
****

The chances of getting killed for a black man if fleeing is **6 times** more than for a white person.

**Person fleeing armed:**

The chances of getting killed for a black guy if fleeing is **more than 3 times** higher compared to a white person.

**Person Not Fleeing and is Armed**



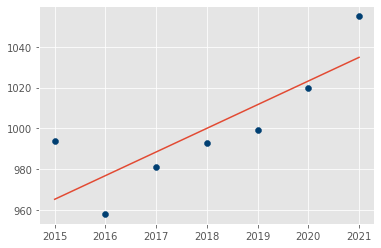
The chances of getting killed for a black guy if he is not fleeing and is armed are **more than 2 times** higher compared to a white person.

It’s very clear that, no matter what the circumstances are. **The chances for a person from black race has high chances of death** compared to any other race.

**Linear Regression:**

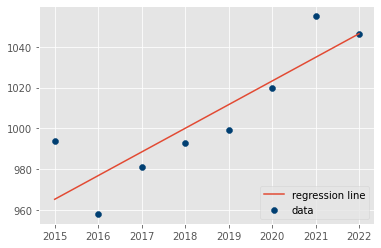
A variable's value can be predicted using linear regression analysis based on the value of another variable. The dependent variable is the one you want to be able to forecast. The independent variable is the one you're using to make a prediction about the value of the other variable. The differences between expected and actual output values are minimized by linear regression by fitting a straight line. We get a straight line/ best fit line by least square method.

In our project, from the original dataset “ US Police shootings in from 2015-22.csv”. I extracted 2 columns namely year and count (number of killings). Year on the X-axis and count on Y-axis.



In the above image, the red coloured straight line is the ‘best fit’ line. With the help of this line we get the slope equation and thus find the dependent number.

‘*The variable you want to predict is called the dependent variable.*’



The predicted value for the year 2022 is 1047.

# **AutoRegressive Integrated Moving Average (ARIMA):**

AutoRegressive Integrated Moving Average, is a set of models that explains a time series using its own previous values given by the lags (AutoRegressive) and lagged errors (Moving Average) while considering stationarity corrected by differencing (opposite of Integration.) In other words, ARIMA assumes that the time series is described by autocorrelations in the data rather than trends and seasonality.

**AR**: Autoregression.

**I**: Integrated.

**MA**: Moving Average.

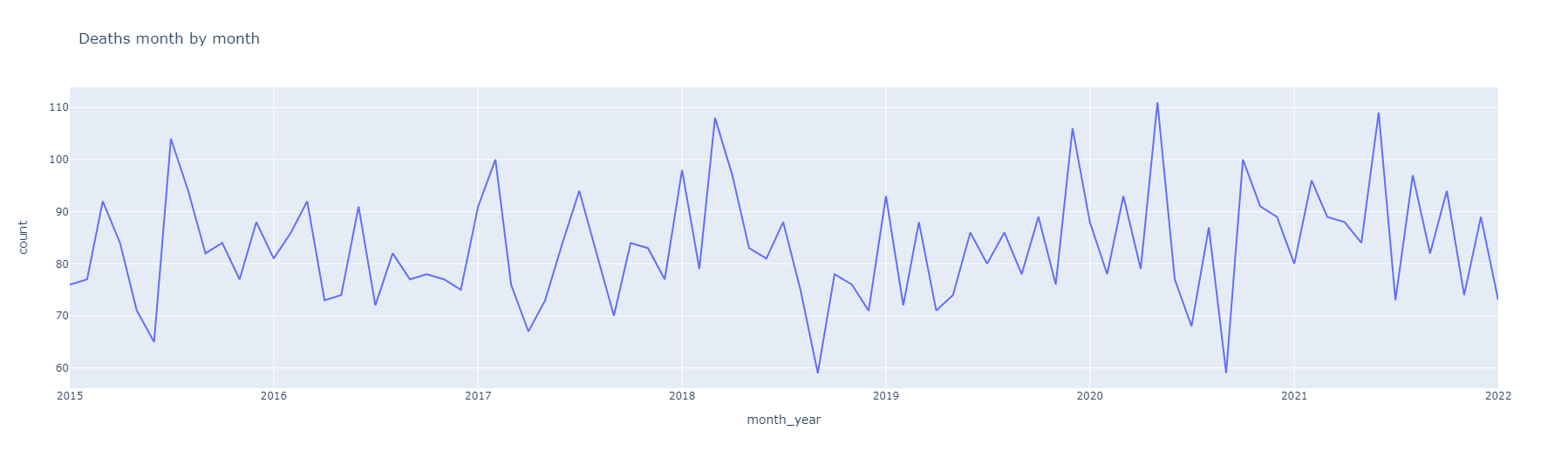
A standard notation is used of ARIMA(p,d,q).

p: Lag order.

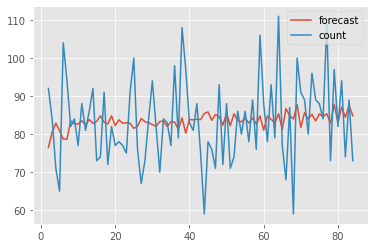
d: Also called the degree of differencing (The number of times that the raw observations are differenced).

q: Also called the order of moving average (The size of the moving average window).

For our dataset we used (3,1,1) as this combination gave the best results.



The above figure is a plot between month\_year column and count (no.of killings).



The red colored waveform in the above plot is the prediction plot we got. When we forecast it for the next 12 months, we are going to predict the increased average mean of no.of killings per month.

