San Jose State University

Project report on,

Forecasting Police Killings

Under the guidance of,
Prof. Jorjeta Jetcheva

CMPE 255 Data Mining
Fall 2021

Team 10
Jayanth Reddy Sheri (015242565)
Saketh Gali (015504853)
Priyank Bardolia (015742532)

Github Link: https://github.com/Preeyank/Data_Mining/tree/main/Group_project_team_10

1. Introduction:

Motivation:

With the rise of the Information Age, practically everyone now has a camera phone, allowing for an unprecedented view into the lives of everyone. Unfortunately for the police officers who are interested in power abuses, this has exposed their wrongdoings. In 2013, the acquittal of George Zimmerman began the Black Lives Matter (BLM) movement, and the killings of Eric Garner and Michael Brown in 2014 solidified police violence as a front-burner issue in the minds of BLM supporters. This went on to the next dimension with the George Floyd assassination in 2020. BLM has expanded over time as more people have been aware of the concerns and have been exposed to horrifying footage of police brutality.

Objective:

To explore and analyze the dataset and statistics of police violence in the United States. Also to predict the number of killings for a particular month which we have derived from the given dataset.

Approach:

Data mining and machine learning approaches could be used to solve this category of problems, according to a deeper examination of the problem statement. The problem essentially involves cleaning the data, visualizing and doing pre-processing for the Time series analysis, detecting whether a particular dataset is stationary in terms of time. Then training the derived dataset on a model based on stationarity and plotting the AutoCorrelation and Partial Correlation for the hyperparameters.

Literature/Market review:

- Time Series Analysis is the most explored area in the field of machine learning.
- It is not about the accuracy that we get in the Time Series Analysis, it is about the precision. Even Though the model seems to be having an accuracy of 98%. It might seem to a layman that the accuracy is too good but when this same model is applied to stock market prediction then these smaller changes on 2% can cause billions in the loss.
- In order to achieve precision in a Time series analysis. Before moving forward with prediction, a person needs to understand the data based on his/her own intuition. For eg. by plotting the results a person needs to figure out whether the data seems to be stationary or non-stationary by performing some test like adfuller and then move forward with choosing the best-suited model.
- Even after this, there are many factors that might result in the worst precision for a model. So for this, we just need to do hyperparameter tuning using the order term in Time series analysis.

2. System Design and Implementation:

Algorithms:

1) ARIMA Model:

- The ARIMA model predicts a given time series data (which is in DateTime format) based on the past values which we have(in our case the number of kills)
- For non-stationary data, the ARIMA model is often the best choice.

- Arima is the abbreviated version of Automated Regression Integrated Moving Average.

```
monthlydf_value_counts['forecast']=model_fit.predict(start=230,end=247,dynamic=True)
monthlydf_value_counts[['Kills','forecast']].plot(figsize=(12,8))
```

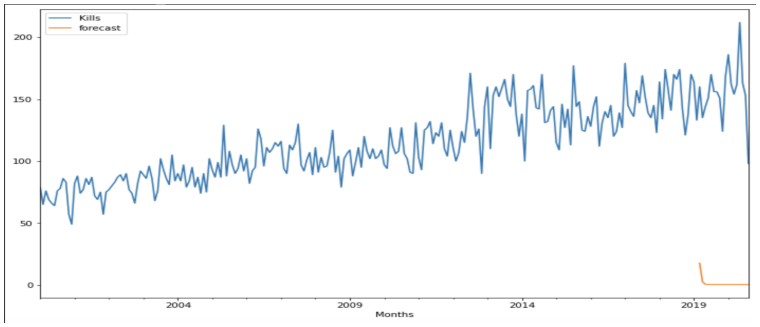


Fig 2.1: Resultant Graph for Arima Model. The orange line is the result of the prediction

2) SARIMAX Model:

- The most important difference between ARIMA and SARIMAX is that in SARIMAX the model is considered to be stationary and based on that we have to do the prediction.
- But actually, our mode was non-stationary which we concluded from the Augmented Dickey Fuller test which we performed and from that we obtained the p-value less than 0.05. So the data was non-stationary.
- We converted the non-stationary data to stationary using the seasonal order of 12.

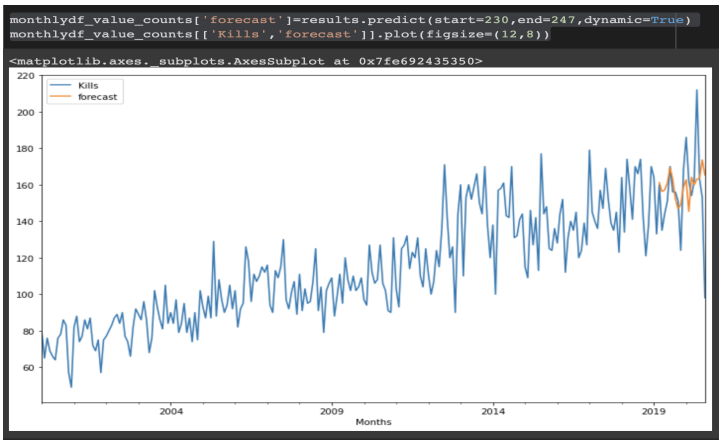


Fig 2.2. a: Resultant Graph for Sarimax Model. The orange line is the result of the prediction

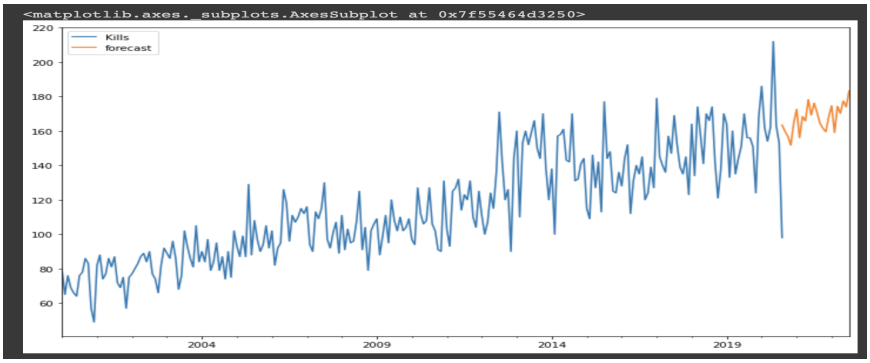


Fig 2.2. b: Future prediction of Police Killings for the next 24 months.

3) Imputing different values of the seasonal term:

After trying the following seasonal terms:

- 3 - the peaks were not detected and the predictions were similar every three months
- 4 - almost the same as the seasonal term of 3 with just some variations
- 6 - Peaks were detected but still the results were not that appealing
- 12 - In this, the peaks were detected and the prediction graph was changing with the actual values. Even Though the values were not exact, it was still way better than any other seasonal term.

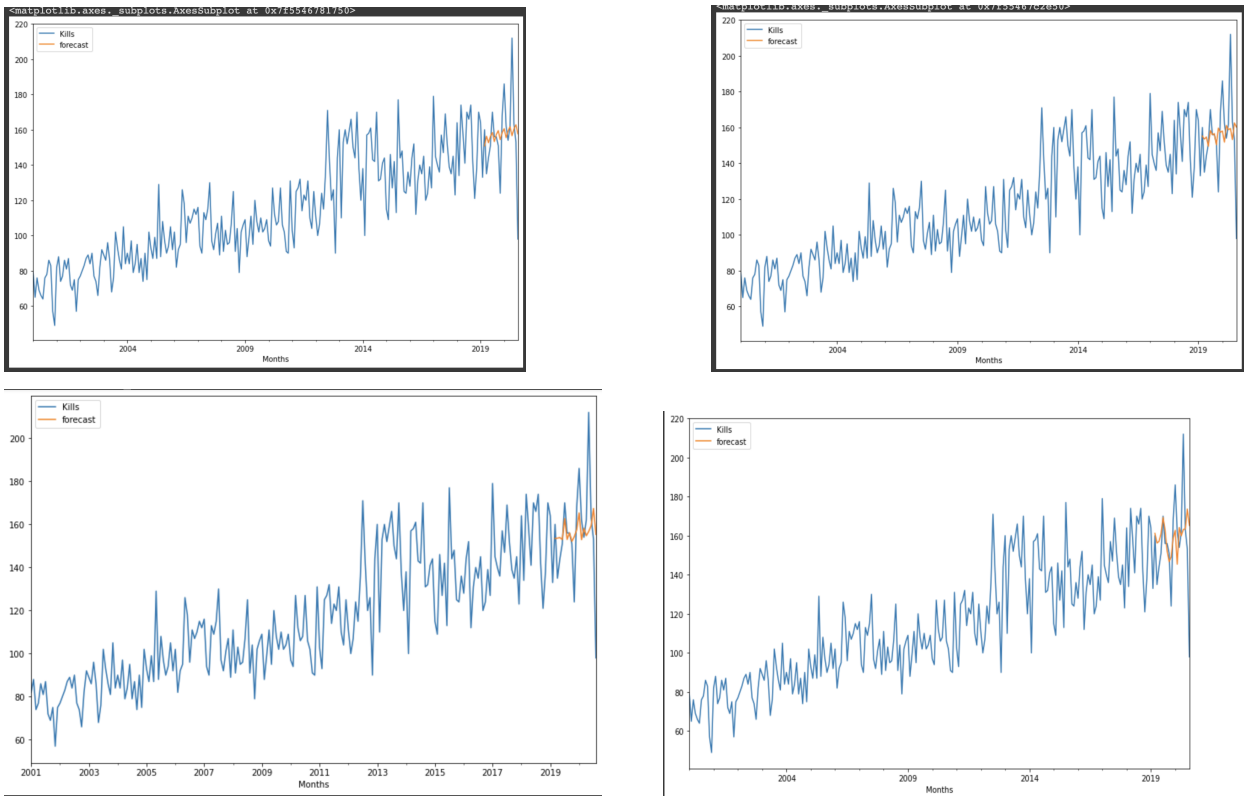


Fig 2.3 : Seasonal term values for 3(Top left), 4(Top right) ,6(Bottom left) and 12(Bottom right)

4) Getting the value of p and q from AutoCorrelation and Partial Correlation plot:

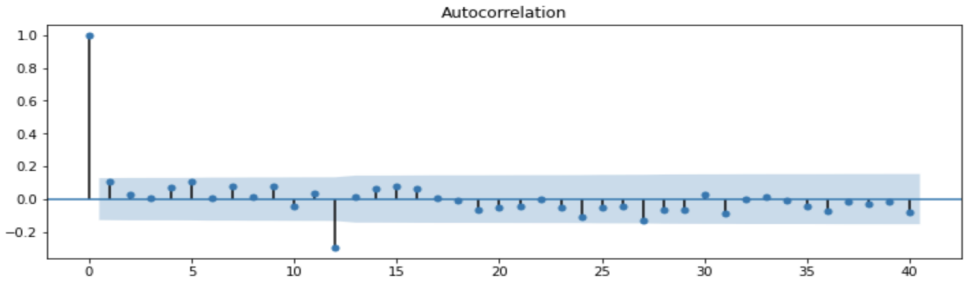


Fig 2.4 : AutoCorrelation Plot

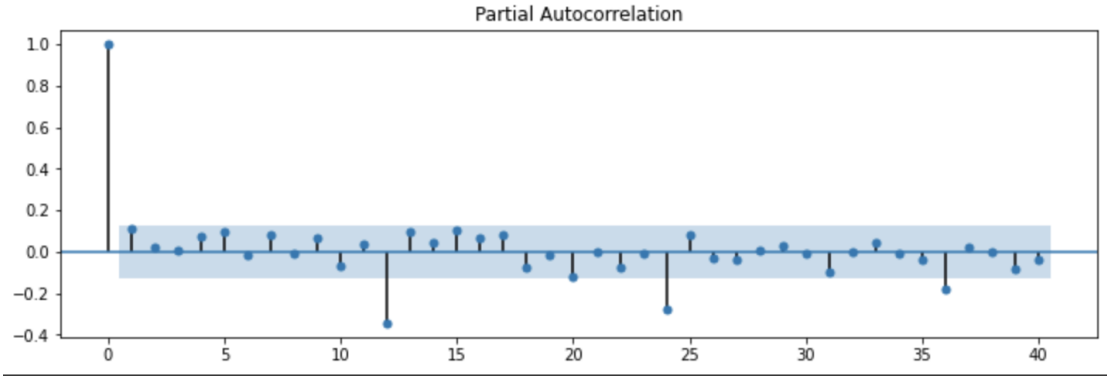


Fig 2. 5: Partial Correlation Plot

5) Imputing different values of d:

The values of d we used are as follows:

- d = 0 : is in the range of the actual results but not able to identify the peaks properly
- d = 1 : is able to identify peaks when compared to d = 0
- d = 2 : the peaks are identified but the values of predictions are way higher than the actual values.

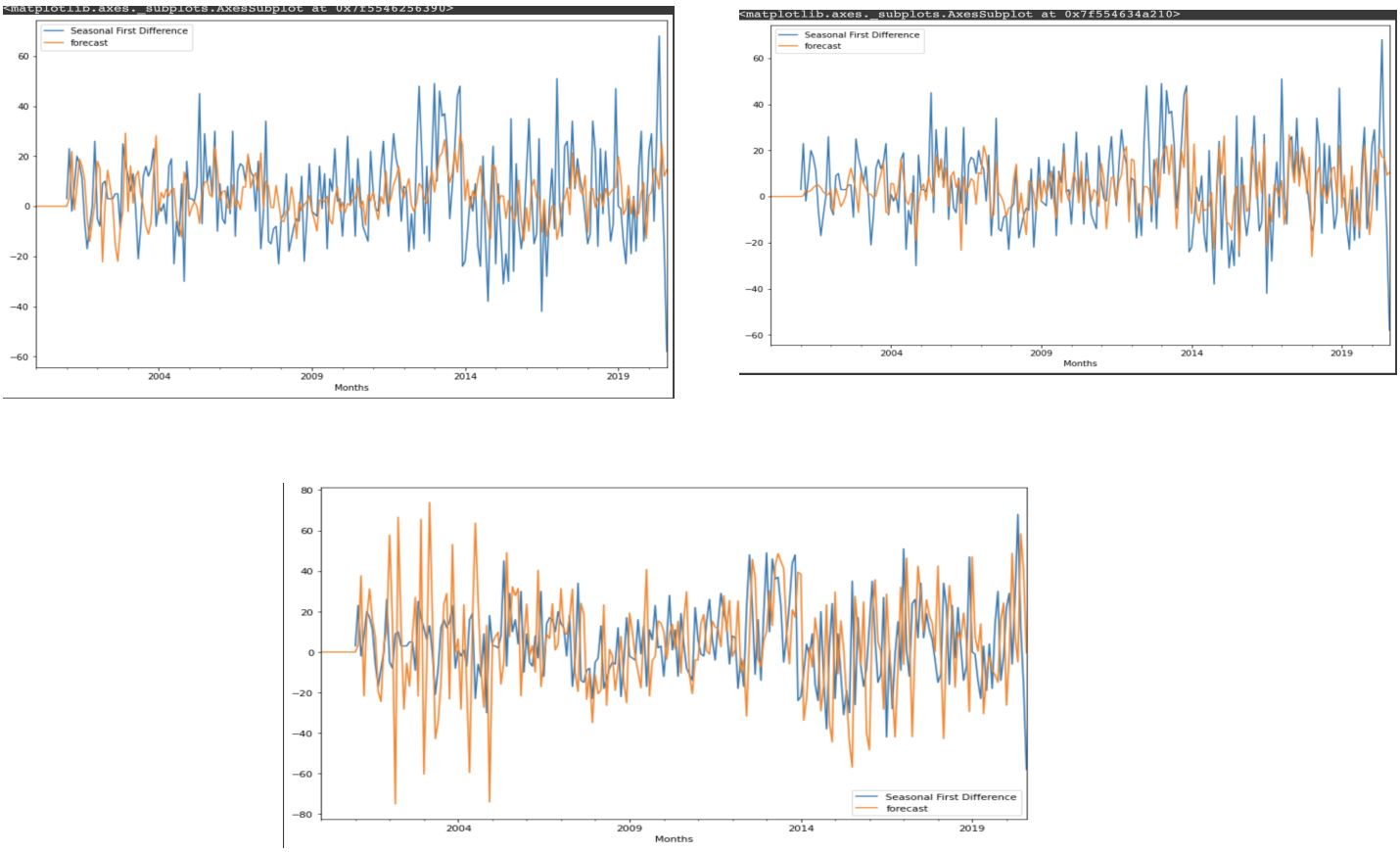


Fig 2 .6: The above graphs are plotted with d values 0(Top Left),1(Top Right), and 2(Middle) respectively.

So to conclude this, it seems that the value of d = 1 is giving better results than d=0 and d=2. So we will be moving forward with d = 1 values.
Note: Even Though points 3,4 and 5 are not an algorithm, it was the most important part while performing Time Series Analysis.

Technologies/Tools used:

- Google Colab
- Libraries used: Numpy,Pandas,Seaborn,Matplotlib,Statsmodel,Scipy,Sci kit learn, Folium ,Plotly, Side Table, Datetime,adfuller,ARIMA,sm, plot_acf,plot_pacf, DateOffset

Data Flow Diagram:

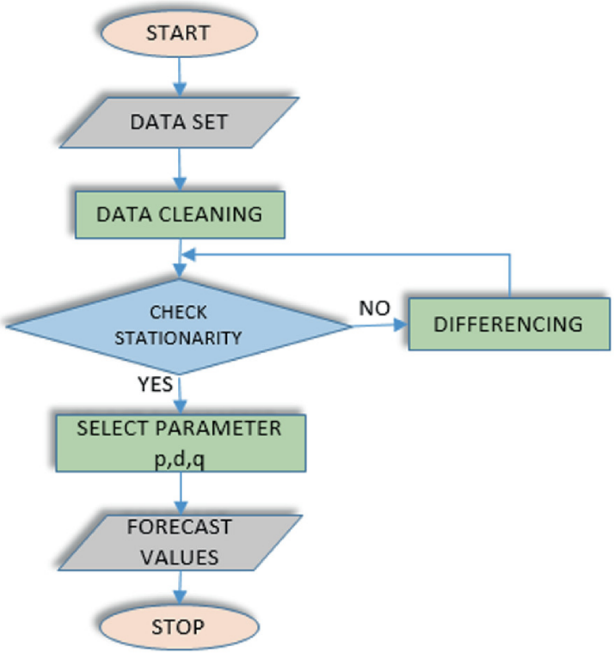


Fig 2.7: Data Flow Diagram

3. Experiments / Proof of Concept Evaluation:

Datasets Used:

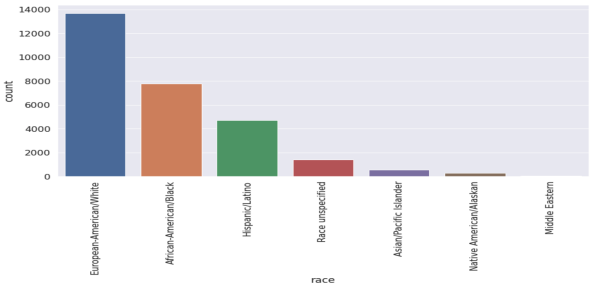
- Fatal_Encounters - **28600 rows x 29 columns** (https://www.kaggle.com/jpmiller/police-violence-in-the-us?select=fatal_encounters_dot_org.csv)
- Police_Killings - **8427 rows x 67 columns** (https://www.kaggle.com/jpmiller/police-violence-in-the-us?select=police_killings_MPV.csv)

Methodologies:

- Obtained the datasets from the Kaggle site.
- Removed all the unessential columns and also dropped the sample where the null values were more.
- Removed the noisy data that includes incorrect date format or age format in absolute numbers.
- Imputed the null values in the age column with the mean replacement strategy.
- Converted the data types of the column from objects to their appropriate ones.
- Data Preprocessing for the Time series analysis.
- Checking for the stationarity of the time duration obtained from the derived Dataset.

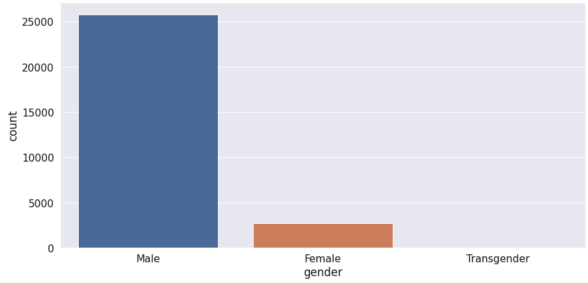
- Training the specific model based on the Stationarity.
- Imputing different values for seasonal order
- Plotting Autocorrelation and Partial Correlation for getting values of p and q
- Imputing different values of d based on intuition and results.
- Predicting the future values and checking for accuracy.

Graphs & Analysis:

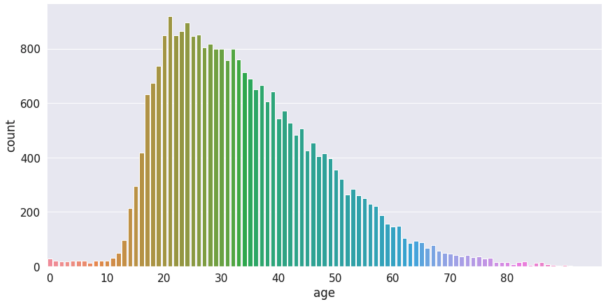


Graph 3.1 Number of kills Race wise

White individuals accounted for nearly 50 percent of those slain, followed by Black people with 29 percent. In terms of absolute statistics, white individuals have been killed by cops in greater numbers than any other race, with 14000 fatalities since 2000. When the frequency of fatal police shootings is normalized by demographics, Black individuals are nearly twice as likely as white people to be killed by police. Police fatally shot around one out of every 100,000 African Americans, compared to 0.2 out of every 100,000 white persons.

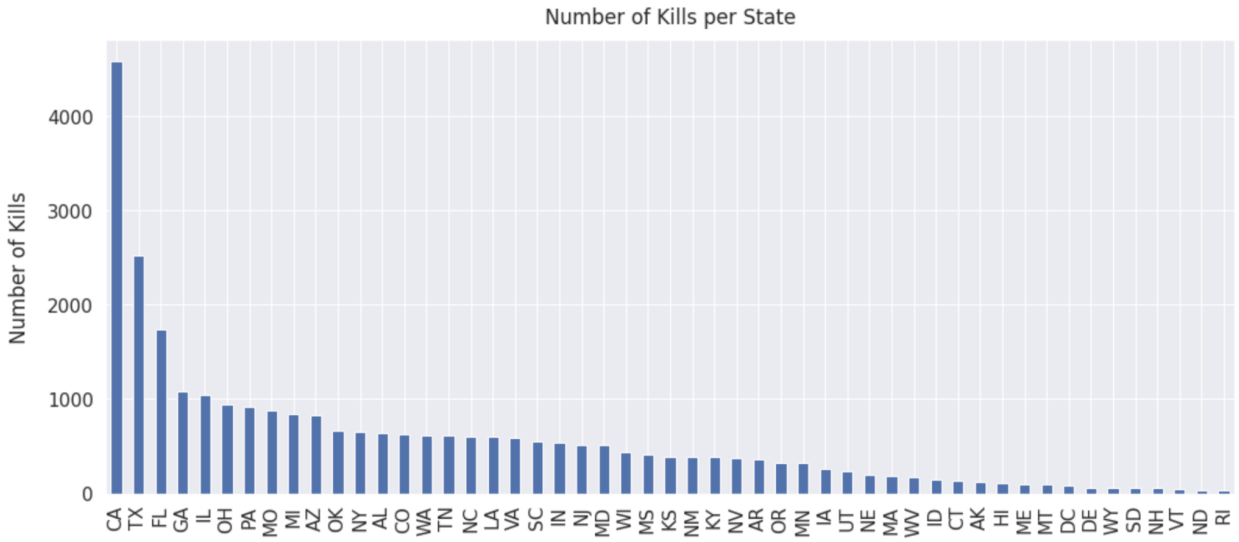


Graph 3.2.a Number of Kills Gender wise



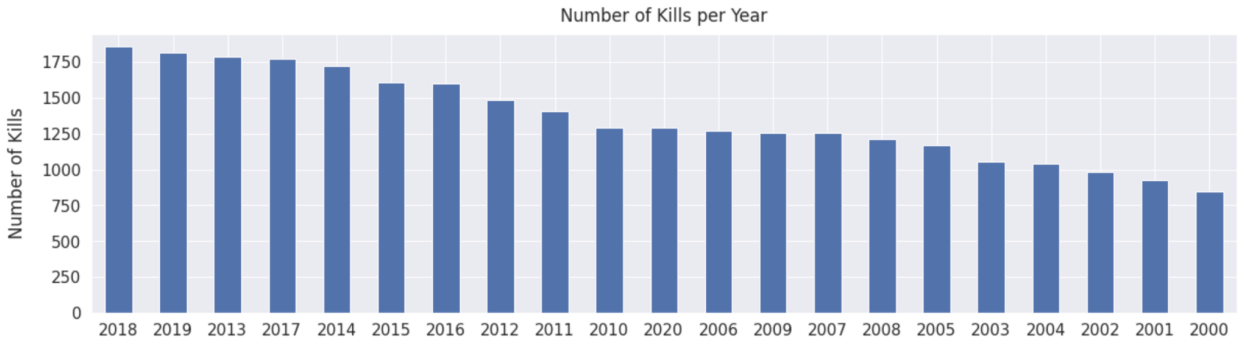
Graph 3.2.b Number of kills Age wise

In terms of Gender, more than 92% of the kills were of males while the rest are of females. This Dataset contains ages ranging between 1 and 107 and most people killed with ages between 20 to 40 with a mean of 35.



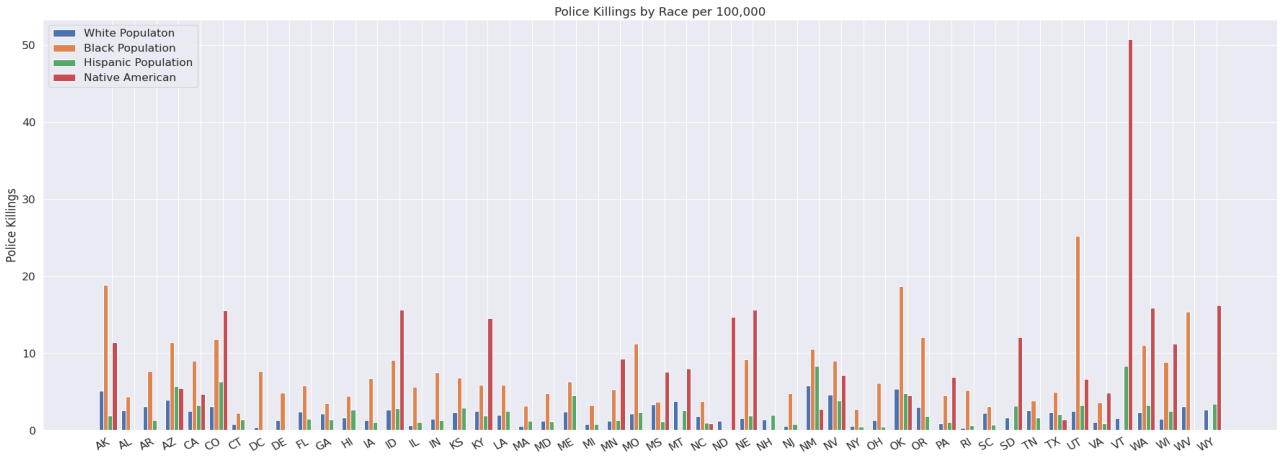
Graph 3.3 Number of kills state wise

Graph 3.3 illustrates the total number of kills in each state. From the graph, we can see that California, Texas, and Florida together account for almost one-third of all fatal police shootings. The fatal shootings in California are double that of Texas. This is larger than all differences between consecutive states in the above plot. While the least being Rhode Island in the Number of kills.

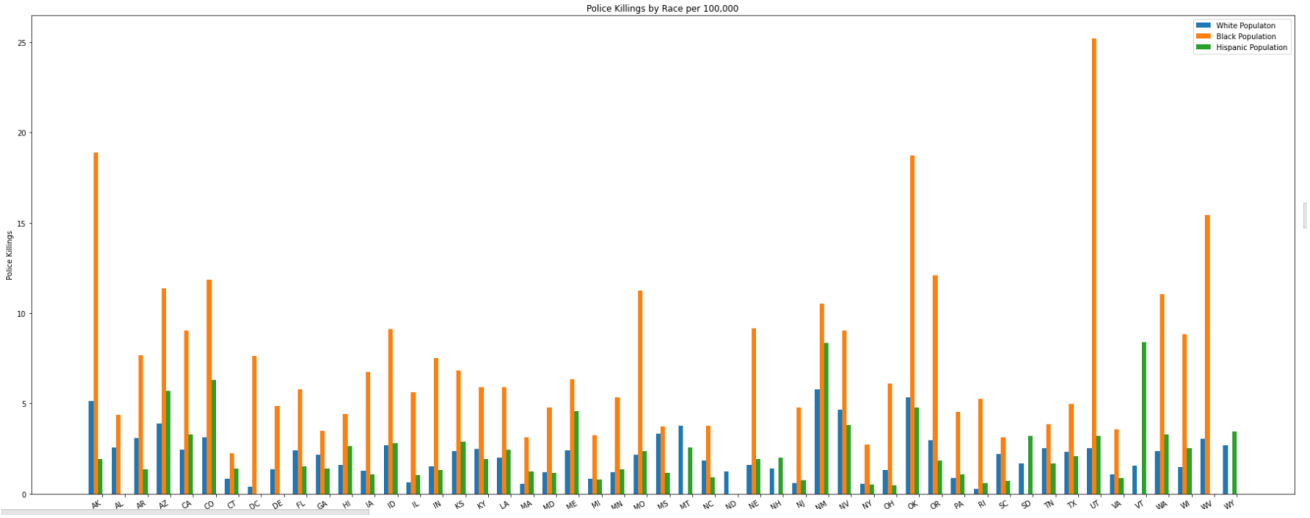


Graph 3.4 Number of kills Year wise

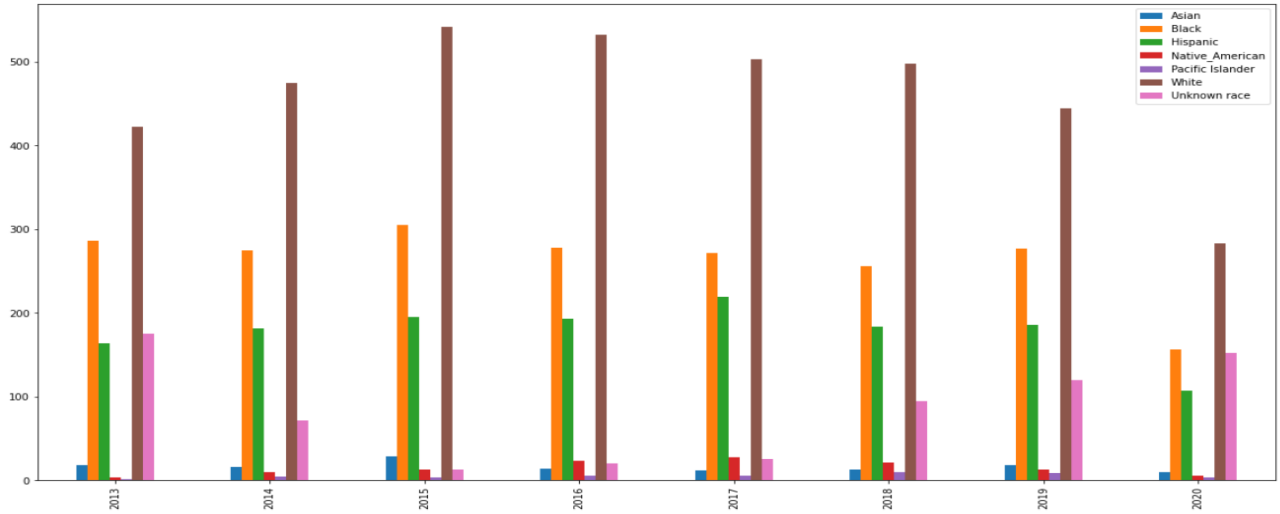
Graph 3.4 depicts the total number of kills in each year from 2000 to 2020. We can clearly see that in 2018 there were more deaths than any other year. The year 2019 and 2013 also had a similar range of deaths while in 2000 there was the least number of deaths.



Graph 3.5.a Comparing Police killings per 100,000 individuals for each population by state in United States



Graph 3.5.b Modified graph for better analysis



Graph 3.6 Number of Kills based on race and year wise

4. Discussion & Conclusions:

Decisions Made:

- Removing the unnecessary columns in the datasets which were not useful for visualization as well as prediction.
- Removing the outliers in the age column and also replacing it with mean.
- Checking for correlation between the attributes to check whether any kind of prediction is valid or not.
- Checking if the attribute values in the dataset were evenly distributed or not.
- Finalizing that our dataset will be well suited for forecasting the police killing through time series analysis.
- Creating an entirely new data frame for time series analysis from the existing dataset.
- Converting the data frame to DateTime format suitable for Time Series Analysis and resetting the index values to date
- Tuning the hyperparameters like order and seasonal order for the Arima and Seasonal Arima model.

Difficulties Encountered:

- Figuring out the attributes to remove.
- Figuring out the regular expression to use for removing the outliers in the age column.
- Figuring out what to predict and what kind of algorithm to use.
- Converting the dataset into Time Series Format.
- Figuring out the values p(Autoregressive Terms),q(Number of lagged forecast errors), and d(Number of non-seasonal differences).

Things that worked well:

- Removing the outliers.
- Visualizing the dataset based on race, year, gender, state, county and population density.
- Creating the data frame suitable for Time Series Analysis.
- Converting the dates available in the actual dataset to the number of kills month-wise.
- Forecasting the future police kills monthly.

Things that didn't worked well:

- The model is not able to predict when the number of kills in a particular month is very high compared to other months and vice versa.
- Converting the available dates to a weekly format.
- Making the data frame for forecasting weekly kills.

Future work:

- Instead of forecasting monthly kills, we can forecast weekly kills.
- Instead of predicting the number of kills in the United States. What if we can predict it in a specific State and after that maybe a county.
- Training the model in such a way that the model considers Non-stationary data as Non-stationary data only instead of Stationary data.

Conclusion:

- **SARIMAX MODEL** with order(1,1,1) and Seasonal order(1,1,1,12) - **77.85%**
- We successfully converted **non-stationary** data to **stationary** data for better predictions

- We tried forecasting data for the **next 24 months** and the predictions were found to be in the range of the actual dataset.
- We even tried ARIMA but Arima wasn't performing well because our dataset was nonstationary

5. Task Distribution:

- **Jayanth Reddy Sheri -** Visualization on fatal_encounters dataset, Implemented replacement strategies, converting the data to stationary to check for seasonal order.
- **Saketh Gali -** Visualization on Police Killings dataset, Data Cleaning for Visualization, Tuning Hyperparameters for order and seasonal order.
- **Priyank Bardolia -** Creating the data frame suitable for Time Series Analysis, Prediction for monthly kills using ARIMA and SARIMAX, Forecasting on the next 2 years of dates made available by DateOffset.