COMP6235 Referral Coursework : Gun Violence in U.S

Yuan Lou
29363373
Data Science
yl2m17@soton.ac.uk

ABSTRACT

We often see news of American shootings on television, especially the vicious shootings that took place in Las Vegas in 2017. (BBC. 2017)As a tourist, when we want to travel to the United States, the security issue is our first consideration: Is the US shooting frequent? Which area has the most shootings? When will a shooting incident occur? Therefore, I tried to analyze whether the United States is safe by analyzing the datasets of the annual shootings in the United States. And try to predict the occurrence of future shootings. At the same time, I also established a data science application to visually display the situation of the US shootings. I hope that I can help users who want to go to the United States to visit relevant data and have a faster understanding of the shooting situation in the United States.

Keywords

U.S, shooting, analysis, predict,

1.INTRODUCTION

Is it safe for the United States to be a matter of great concern to me because we often see news of American shootings on television. I tried to analyze whether the United States is safe by analyzing the data of the annual shootings in the United States. And to predict the occurrence of future shootings. My US shooting datasets come from kaggle, which is obtained from two places, and because I want to analyze the relationship between the time of the crime and the holidays, I get also the dataset of the US holiday. At the same time, in order to compare the deaths and shootings of the United States and other countries, I get the data from other paper. The steps from data cleansing to data analysis and forecasting are described in the article in order. First, the original data is cleaned up, and then the EDA is analyzed by category, such as the age of the attacker, the place of the crime, the number of deaths, etc. Finally, the prediction using the time series includes the use of the Prophet tool and the ARMA method. Then it introduces the functions, advantages and disadvantages of the data science WEB application. Finally, by comparing with the population of other countries, it is concluded that the US shootings are very high in developed countries.

2.Data and Tools Preparation

We import data from different data sets and observe the data types. The amount of data is actually not large, but there are a lot of missing data, which will be processed later. The next step is the data cleaning. firstly, processing dates and the features unification of Race, Gender, and Mental Health Issues are important. Because in the latter analysis, it was found that the data in 2013 was less, so removing the data in 2013. I standardized all of Gender as "Male", "Female", "Unknown". On the Race question, according to the original data, it can be divided into white, black, and Asian, Latino, Native American and others. At the same time, in order to facilitate the statistics of the frequency of shootings in each city and each state, we processed the city and state through the Location variable. In addition, in order to judge whether the shooting was happening on holidays, the data set of the US holiday was merged.

I chose pyhton for data processing and analysis. The reason I chose Python is because it has a lot of packages to use. Some packages need to be imported. Basemap is a library that uses maps in Python. Plotly is a convenient drawing tool for js, python, R, DB, etc. Seaborn is a python data visualization library based on matplotlib, providing a higher level API, it is obviously convenient and quick to use. Fbprophet is Facebook's open source time series prediction framework 'prophet', currently supports R language and Python language

3.DATA ANALYSIS

3.1.Number of deaths and injuries per year

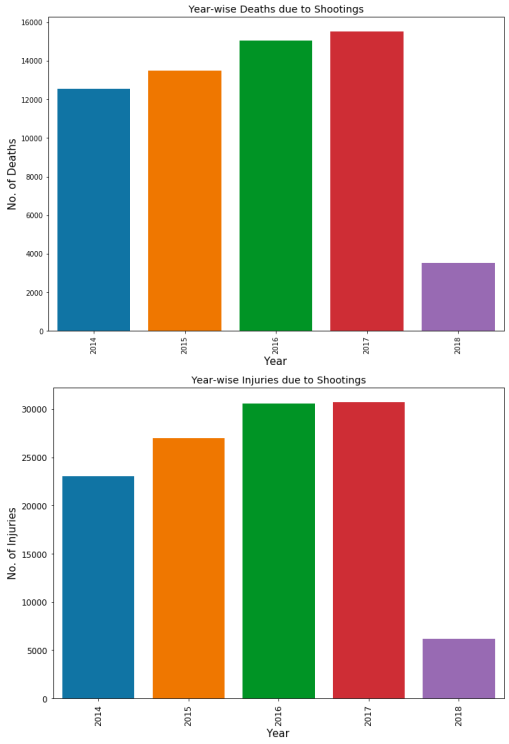


Table 1. Year-wise deaths and injuries due to shooting

According to the above two figures, the number of deaths and injuries per year is on the rise.

3.2.Regional distribution of deaths and injuries

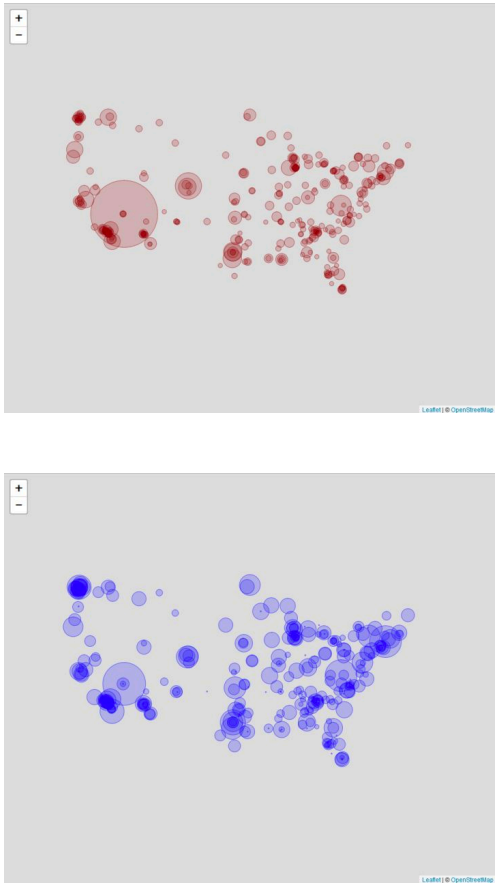


Table 2.regional distribution of deaths and injuries

The graphical results are obvious. There are shooting incidents in all parts of the United States, but the number of shootings in the western region is obviously less than that in the eastern region. The blank area belongs to the American Rocky Mountains, and the area is sparsely populated so there is no shooting.

3.3.Type of criminal

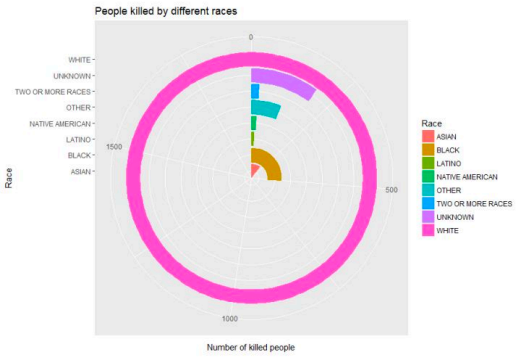


Table 3. Type of criminal

An analysis of the types of criminals reveals that men are the most important, and blacks and whites seem to be the protagonists of the shootings.

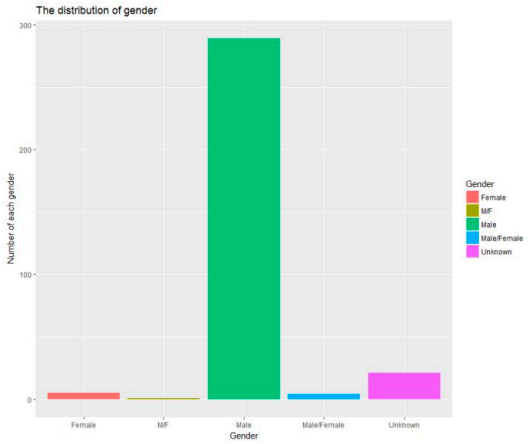


Table 4. Type of sex

4.TIME SERIES PREDICTION

Time series prediction refers to a sequence in which the values of the same statistical indicator are arranged in chronological order in which they occur. (Imdadullah.2014)The main purpose of time series analysis is to predict the future based on existing historical data. Can we use time series to predict how many gun violence will occur in the future? Print a time series of 2018 existing data:

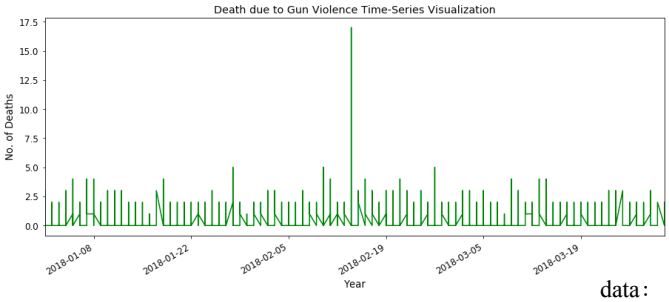


Table 5.2018 time series

4.1.fbprophet

I use Prophet in the open source fbprophet framework of facebook to perform time series analysis. This framework can also display detailed information such as the weekly effect of the analysis, which is very easy to use. The principle of Prophet is to analyze various time series features: periodicity, trend, holiday effects, and some outliers. In terms of trends, it supports the inclusion of

change points for piecewise linear fit. In terms of cycle, it uses the Fourier series to establish the periodic model (sin+cos). In terms of holidays and emergencies, the user can specify the holidays and the related N days by means of tables. Prophet can be seen as an integrated solution for time series.

The specific steps to use Prophet are: fill in the training data according to the format requirements, fill in holiday data, fill in specify the time period to be predicted, and then train. In addition to predicting specific values, Prophet also splits the predictions into components such as trend, year, season, week, etc., and provides the upper and lower bounds of the prediction intervals for each component. Not only a predictive tool, but also a good statistical analysis tool. Of course, Prophet also has its weaknesses, such as few adjustable parameters, no support for its timing features, etc., but these can also be solved by predictive processing and model fusion.

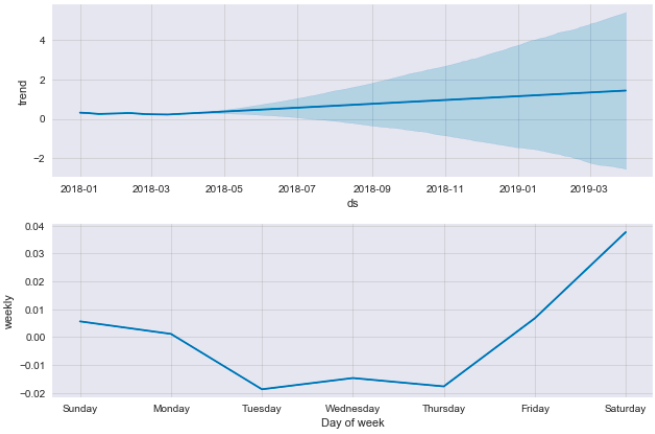


Table 6. Prediction of 2018

The results show that the number of shooting incidents in 2018 is still on the rise, and there is a strong upward trend on Friday and Saturday.

4.2.ARMA

The most basic time series models in time series: AR, MA, and ARMA. Which is the best model and what criteria are used to evaluate the "optimal". Generally speaking, akaike information criterion (AIC) and bayesian information criterion (BIC) are two indicators for evaluating models. These two evaluation indicators are not only applicable to event sequence models, but also widely used in other mathematical models:

$AIC = -2 \ln(L) + 2 k$

$BIC = -2 \ln(L) + \ln(n)*k$

The method of the AR model is very simple. The model executes that the current time point can be predicted by the linear combination of time series of past time points plus white noise, which is a simple extension of the random move. The MA model is similar to the AR, which is not a linear combination of historical time series values but a linear combination of historical white noise. The biggest difference from AR is that the effect of historical white noise in the AR model is indirectly affecting the current predicted value (by affecting historical time series values). The ARMA model is obtained by mixing the AR and MA models, and AR(p) and MA(q) together constitute ARMA(p, q).

I tried to use the ARMA time series model to predict. If a time series has both AR and MA parts and is stationary, it constitutes an ARMA model. Data from 17 years and January and February of 18 years were selected as training, and data for March 1, March 2, and March 3 was predicted. We use the following conditions to determine whether to use the ARMA model:

4.3.Autocorrelation

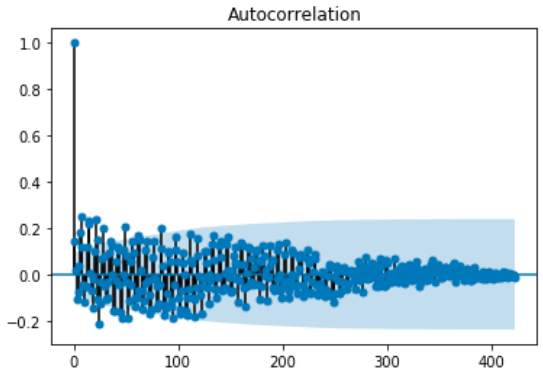


Table 7.autocorreclation

The results on the graph indicate that there is a significant first-order smearing effect, so there is no correlation in the judgment sequence.

4.4.Stationarity judgment (ADF detection)

The ADF test results of the original sequence are:

(-3.0443905852697957, 0.03095172945803373, 14, 408, {'1%': -3.446479704252724, '5%': -2.8686500930967354, '10%': -2.5705574627547096}, 2874.7803552969235)

The ADF test results showed that p=0.03<0.05, indicating that there was no significant autocorrelation.

4.5.Partial autocorrelation

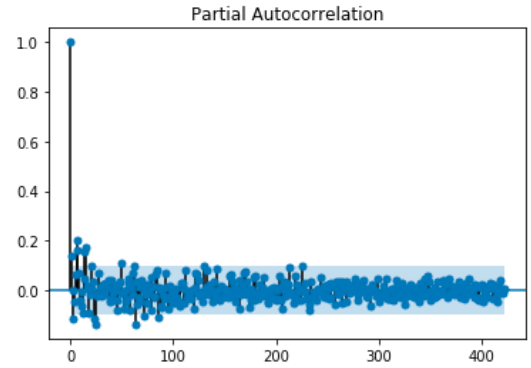


Table 8.partial autocorrelation

It can be clearly seen from the figure that the partial autocorrelation plot of the sequence has a significant first-order smearing effect.

4.6.White noise

After detection, the white noise test result of the differential sequence is:

(array([8.36275837]), array([0.00382989]))

p=0.001<0.00, indicating that the sequence is a stable sequence of non-white noise, which can be analyzed in the next step.

4.7.ARMA forecast

According to the above several conditions, the autocorrelation and partial autocorrelation plots of the sequence are both 1st order and are stationary non-white noise sequences, so I can determine to use the ARMA model, and the model parameters p=1, q =1,

So I built the model, made a five-day forecast, and returned the predicted results, standard error, and confidence intervals. The results are as follows:

(array([40.01428989, 42.05286687, 42.33799725]),

Array([8.95947756, 9.05342063, 9.05524871]),

Array([[22.45403655, 57.57454324], [24.3084885, 59.79724523], [24.59003592, 60.08595858]]))

It is known from the results that the true value falls within the confidence interval (5%), and it can be inferred that the number of people we predicted to die due to the shooting incident is reasonable, but the shooting is still a sudden uncertainty event. As time goes by, the accuracy of prediction will drop significantly.

5.APPLICATION

5.1.Functions

Through these analyses, I implemented a data application. The application is based on WEB. The main technology is js and html. I use some frameworks such as JQuery and bootstrap. D3.js and Echart.js were used to implement data visualization. Ajax and VUE.js are used for data interaction.

In this application, users can understand the shooting situation in the United States through visual operations, such as viewing by category, year, and region. At the same time, it can be compared with other countries.

5.2.Strong Point

In this application, I want users to have the easily and flexibility to view the data they need, such as by swiping and clicking. These can be implemented through frameworks such as D3.js, and bootstrap also has front-end interactive buttons, which are very beautiful and simple. For fast interaction of data, ajax is the best choice. The biggest advantage of AJAX is that it can communicate with the background without refreshing the entire page. This allows web applications to respond more quickly to user interactions, and avoids sending unchanging information on the network, reducing user wait times and delivering a very good user experience.

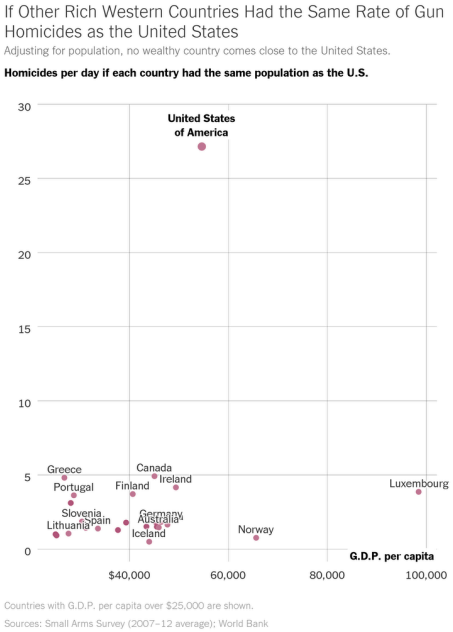
5.3.Weak Point

At the same time, I didn't make my content display on mobile devices very well. Some handheld devices (such as mobile phones, PDAs, etc.) still don't support Ajax very well. For example, we open it on the browser of the mobile phone. When it comes to Ajax technology's website, it is likely to be very slow. The responsive layout of the application should be improved to support mobile devices.

6.Comparison with other countries

In Germany, the rare degree of murder and death caused by guns is as good as death from the fall of the United States. About two out of every one million people will become victims of gun homicides. In several other European countries, including the Netherlands and Austria, gun-killing cases are equally rare. In the United States, the same probability is almost the fatality of hypothermia or plane crash.（kevin.Q,2016）

In Poland and the United Kingdom, about one in every one million people die every year from gun homicides – almost the same as the fatality rate of agricultural accidents or falling ladders in the United States. In Japan, where gun homicides are even rarer, the probability of being shot is the same as the death of a lightning strike in the United States—about one in a million. In the United States, the death rate from gun killings is about 31 parts per million - equivalent to 27 people being shot every day of the year. The chart below assumes that the population of each country is the same as the US：



SOURCES: SMALL ARMS SURVEY (2007–12 AVERAGE); WORLD BANK

The incidence of gun violence in the United States is not the highest in the world. In parts of Central America, Africa and the Middle East, gun-related deaths are even higher – close to heart disease and lung cancer mortality in the United States. In Mexico, a neighboring country ravaged by the drug war, 122 people per 100,000 people died in gun-killing cases - slightly higher than the pancreatic cancer death rate in the United States. But these countries, where gun violence is extremely serious, are very different from the United States in terms of GDP, life expectancy, and education. In developed democracies, the United States is a unique one.

7.Conclusions

Overall, using the US shooting data set, I analyzed the annual shootings in the United States and predicted the occurrence of future shootings. I also obtained the deaths and shootings of other countries from the data, comparing the United States with other countries. The article first describes how to clean up the original data, and then analyze the EDA by category, such as the age of the attacker, the place of the crime, the number of deaths, etc. Finally, the prediction using the time series includes the use of the Prophet tool and the ARMA method. The data application also demonstrates the visual capabilities to help users better understand the US shooting situation.

Github SSH: git@github.com:louyuanyuan/refferl.git

8.REFERENCES

1. https://www.kaggle.com/jameslko/gun-violence-data
2. shivam bansal,2018 https://www.kaggle.com/shivamb/deep-exploration-of-gun-violence-in-us
3. Kevin Quealy, 2016 www.nytimes.com
4. Imdadullah. "Time Series Analysis". *Basic Statistics and Data Analysis*. itfeature.com. Retrieved 2 January 2014.
5. Margot Sanger-Katz 2016 www.nytimes.com