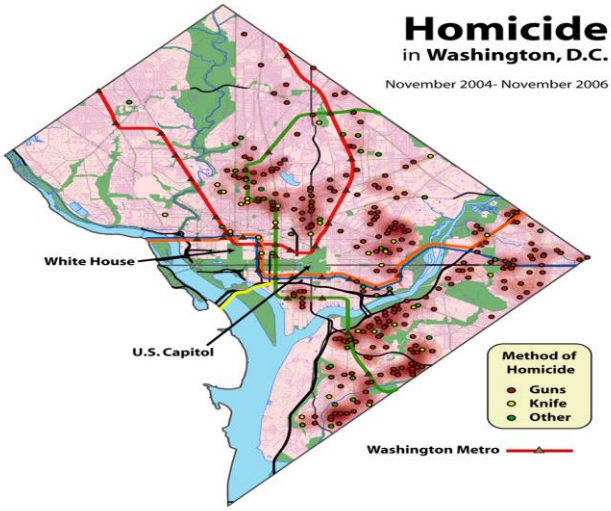
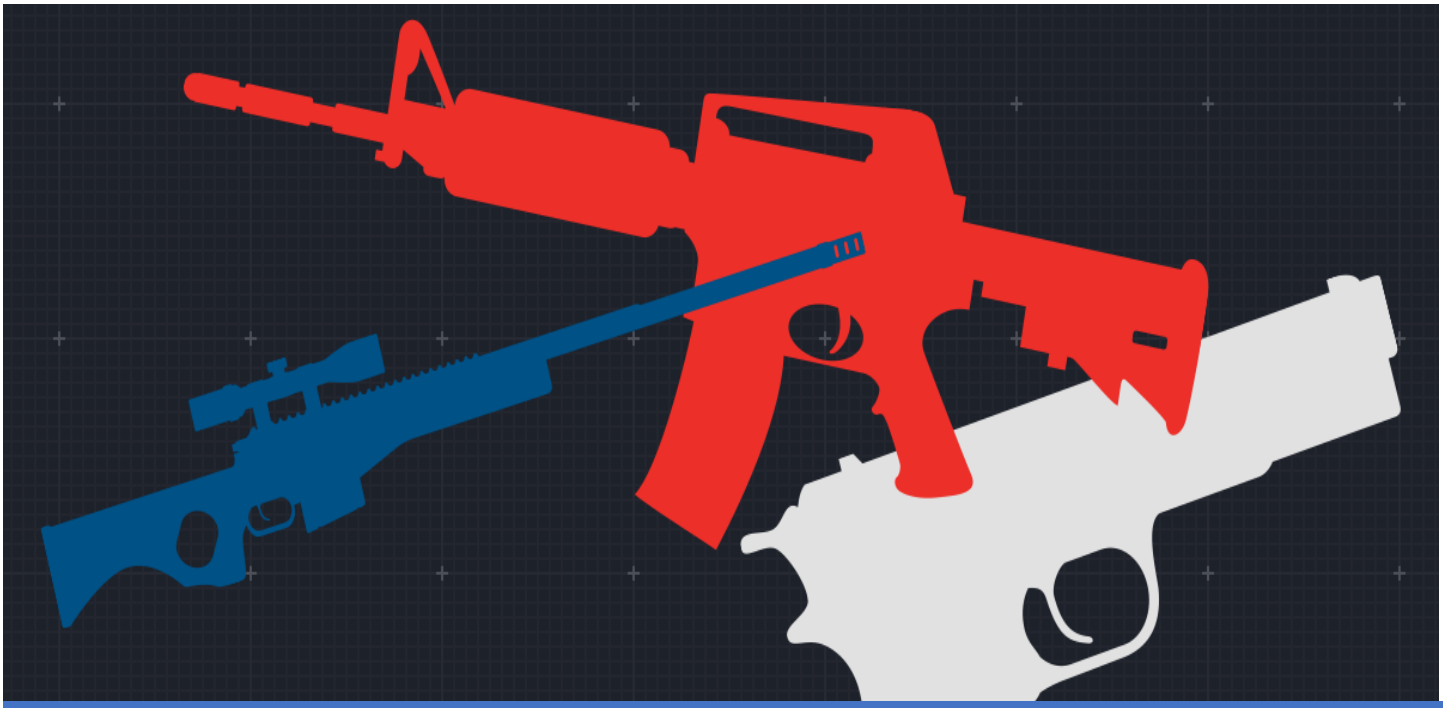




Data Exploration Project on:

VIOLENT MASS SHOOTINGS IN THE UNITED STATES FROM 1966-2017

28682157 5/1/18 FIT 5147

Contents

1.Introduction

Project Background:

The people of the United States have a historically strong relationship with guns and the right to bear arms. It is a part of the nation's constitution. There is a lot of opposition to gun ownership in the United States because of many violent acts related to the use of guns in the last few decades. Opposition to the freedom to own guns has increased greatly because of the high frequency of mass shootings in recent years.

My project is focussed on analysing Mass Shooting Violence that has occurred in the United States from 1966-2017.

Motivation:

I am inspired to analyse this topic because I am interested to know the possible factors that might influence or create an environment that produces Mass Shooters and Gun Violence.

Disclaimer: It is difficult to find relevant datasets/information on gun ownership statistics as it is actively suppressed by the National Rifle Association. For further reference on this topic please visit:

https://www.npr.org/2018/04/05/599773911/how-the-nra-worked-to-stifle-gun-violence-research

Objectives:

This project will analyse some key factors to determine if there is a relationship between them and incidents of violent Mass Shootings.

Key Questions:

1. Do the **economic conditions** of a state have an impact on the number of Mass Shooting incidents that occur?
2. Does the number of **Mental Health Professionals** available in a state impact the number of Mass Shootings that occur?
3. Does the existing overall **violent crime rate** within a state have an impact on the number of Mass Shootings that occur?
4. Are there patterns or trends that can be observed within these Mass Shooting Incidents related to **race, gender, mental health, injuries** and **fatalities**?

Final Aim:

From this analysis, we can determine if there are other important factors besides gun ownership that can influence the number of mass shooting incidents that occur in the United States.

2.About the Data

Data Set Challenges

Although this topic is very current, there is very limited datasets with robust information on the topic.

This project required me to find data from many sources and use special calculations and methods to insert missing values. The timeline for these shootings has made it a challenge to find the relevant data in a clean format and I have had to merge various files to get data that covered the whole timeline from 1966-2017.

Errors

All the datasets had certain types of formatting that had to be streamlined in one way or another. Usually the dates would be an issue. I also had to frequently change value types from integer to string and vice versa. I also had missing values in many columns and I filled them in by imputation.

Mass Shooting Data

The Mass Shootings information was acquired from The Gun Violence Database that compiles gun violence statistics and data from over 2000 sources. The information is consolidated from credible news sources.

Source:

http://www.gunviolencearchive.org/

GDP Data

I acquired 2 datasets from the Bureau of Economic Analysis USA to get the Gross Domestic Product (GDP in millions) by state over the timeline. I had to merge the 2 files. I had to then get the values for 2017 from a different website because the agency did not have this information yet.

Source:

https://www.bea.gov/iTable/index_nipa.cfm

www.usgovernmentspending.com

Crime Data

I acquired data on violent crime statistics by year and state from the FBI uniform crime reporting initiative that consolidates data from agencies across the country. The values for 2017 were not provided and I had to use an additional statistic. I generated values using the 0.8% percentage reduction in violent crimes indicated by the FBI website and applied then on 2016 data to get the 2017 values.

Additional Statistic:

January to June
Percent Change
for Consecutive Years 2013–2017

Data Declaration Download Excel

Years	Violent crime	Murder	Rape[1]	Robbery	Aggravated assault	Property crime	Burglary	Larceny-theft	Motor vehicle theft	Arson
2014/2013	-4.6	-6.0	-10.1	-10.3	-1.6	-7.5	-14.0	-5.6	-5.7	-6.5
2015/2014	+1.7	+6.2	+1.1	+0.3	+2.3	-4.2	-9.8	-3.2	+1.0	-5.4
2016/2015	+5.3	+5.2	+3.5	+3.2	+6.5	-0.6	-3.4	-0.8	+6.6	-1.1
2017/2016	-0.8	+1.5	-2.4	-2.2	-0.1	-2.9	-6.1	-3.0	+4.1	-3.5

Source:

https://ucr.fbi.gov/crime-in-the-u.s/2017/preliminary-report/home

Mental Health Data

I used an interesting factor known as the Mental Health Care Health Professional Shortage Areas (HPSAs) factor to determine how much of the shortage in mental health professional services had been fulfilled by each state. The higher then number indicated that the state had covered that percentage of the total number of mental health care providers needed by that state. This data was extracted from the Henry J Kaiser Family Foundation.

Source:

https://www.kff.org/other/state-indicator/mental-health-care-health-professional-shortage-areas-hpsas

Final Data Set

Information about the final data set has been generated using R Studio and describes the variables and columns as it is easier to display then showing the excel sheet columns:

```
Classes 'tbl_df', 'tbl' and 'data.frame':     323 obs. of  27 variables:
 $ incident_num         : int  1 2 3 4 5 6 7 8 9 10 ...
 $ title                : chr  "Texas church mass shooting" "Walmart shooti
ng in suburban Denver" "Edgewood businees park shooting" "Las Vegas Strip
mass shooting" ...
 $ location             : chr  "Sutherland Springs, TX" "Thornton, CO" "Edg
ewood, MD" "Las Vegas, NV" ...
 $ state                : chr  "Texas" "Colorado" "Maryland" "Nevada" ...
 $ date                 : chr  "05/11/2017" "01/11/2017" "18/10/2017" "01/1
0/2017" ...
 $ incident_area        : chr  "Church" "Wal-Mart" "Remodeling Store" "Las
Vegas Strip Concert outside Mandala Bay" ...
 $ open/close_location : chr  "Close" "Open" "Close" "Open" ...
 $ target               : chr  "random" "random" "coworkers" "random" ...
 $ cause                : chr  "unknown" "unknown" "unknown" "unknown" ...
 $ summary              : chr  "Devin Patrick Kelley, 26, an ex-air force o
fficer, shot and killed 26 people and wounded 20 at a church in Texa"| __t
runcated__ "Scott Allen Ostrem, 47, walked into a Walmart in a suburb goer
s for n"| __truncated__ ...
 $ fatalities           : int  26 3 3 59 3 3 5 3 3 5 ...
 $ injured              : int  20 0 3 527 2 0 0 0 0 6 ...
 $ total_casualties     : int  46 3 6 585 5 3 5 3 3 11 ...
 $ policeman_killed     : int  0 0 0 1 0 NA 1 NA ...
 $ age                  : int  26 47 37 64 38 24 45 43 39 26 ...
 $ employed             : int  NA 1 1 1 1 NA ...
 $ employed_at          : chr  NA "Advance Granite Store" NA ...
 $ mental_health_issues: chr  "no" "no" "no" "unknown" ...
 $ race                 : chr  "white" "white" "black" "white" ...
 $ gender               : chr  "male" "male" "male" "male" ...
 $ latitude             : num  29.3 39.9 39.4 36.2 37.8 ...
 $ longitude            : num  -98.1 -105 -76.3 -115.1 -122.4 ...
 $ year                 : int  2017 2017 2017 2017 2017 2017 2017 2017 2017
2017 ...
 $ GDP                  : int  1703100 350300 409300 161500 2847600 777500
1018000 675400 2847600 1018000 ...
 $ population           : int  28304596 5607154 6052177 2998039 39536653 12
805537 20984400 11658609 39536653 20984400 ...
 $ violent crime rate   : num  430 339 467 671 441 ...
 $ hp_percentage        : num  0.44 0.282 0.525 0.437 0.34 ...
```

Additional Information:

- **GDP** is shown in Millions of USD
- **violent crime rate** indicates No. of Violent Crimes committed per 100,000 residents of a state
- **hp_percentage** is the healthcare psychiatrist % that has been fulfilled of the total shortage that state is facing with regards to mental healthcare workers.

3.Data Wrangling & Cleaning

This project required me to use Python/Panda's extensively to get my datasets in order and combined effectively. The key activities done using Python were as follows:

Geocoding

1) Geocoding: I used Google's reverse Geocoding to get the state names for all the incident locations. This process was also problematic because Google's API does not allow you to harvest the names using a FOR LOOP easily. It was able to convert 10 rows of latitude and longitude data at a time. I looked up the few longitude and latitude coordinates that were

missing manually and inserted them first before doing the Geocoding to get states. There were about 15 missing coordinate pairs only.

In [14]:
```
#for 0-10 outputs worked at 5.25pm
#have to extract each tuple out of the list
#converting in tuple into a mini-list
#b should be the list of states
b= []
for i in state_ID[0:4]:
    mini_list = (list(i))
#print(mini_list[1])
    results = Geocoder.reverse_geocode(mini_list[0], mini_list[1])
    a = results.administrative_area_level_1
    b.append(a)
print(b)
```

```
['Texas', 'Colorado', 'Maryland', 'Nevada']
```

Merging & Melting Data Sets

2) Merging and Melting Data frames: To search and execute joins between data frames I had to use the melt function to change the shape of the files. I merged different data sets as well. I had to merge data frames to get the GDP data from a few of their csv. files split between different years. I had to use the melt function to align information for join commands to my main Mass Shooting data frame for a GDP, Violent Crimes and Mental Health information.

In [1]:
```
import pandas as pd  #pandas will read everything as numeric
import numpy as np
df1 = pd.read_csv('download1997-2016.csv', encoding='utf-8', error_bad_lines=False)
df2 = pd.read_csv('download1963-1997.csv', encoding='utf-8')
df2
# we merged two data frames based on two columns Fips and Area
#how? with inner join (like inner join of two tables)
#( there also left and right join ...se the theory))
df_result = pd.merge(df2, df1, how='inner', on=['Fips', 'Area'])
df_result = df_result.drop(['Fips'], axis=1)
# we rearrange our data frame from horizontal view to vertical view
# it means column names became fields values
df_result = pd.melt(df_result, id_vars=['Area'], var_name='year')
df_result.head()
```

Out[1]:

	Area	year	value
0	Alabama	1963	7343
1	Alaska	1963	1083
2	Arizona	1963	4482
3	Arkansas	1963	3791
4	California	1963	67809

I had to use some lamda functions/regex methods to remove whitespaces to ensure that the searches would work. It required a lot of tests to make sure the correct values were being displayed and there were no 'Not a Number' issues. The same testing had to be done for all the datasets when it was being combined with the master mass shooting data set.

In [17]:
```
# we are getting rid of all white spaces from front and back of the field to make states look okay
df_base['state'] = [str(x).strip() for x in df_base['state']]
```

In [ ]:
```
```

In [18]:
```
df_lastCombined = pd.merge(df_base, df_result, how='left', on=['year', 'state'])
```

In [19]:
```
#df_lastCombined.info
```

In [ ]:
```
```

In [20]:
```
# 2016 is an int here( we just check if there such fields exist)
#df_lastCombined[((df_lastCombined['state']=='Texas') & (df_lastCombined['year']==2016))]
```

In [21]:
```
#df_lastCombined
```

In [22]:
```
#Len(df_lastCombined[df_lastCombined['GDP'].isnull())]
```

In [23]:
```
#df_lastCombined[df_lastCombined['GDP'].isnull())][-4:]
```

Cleaning and Processing Datasets

3) The formatting for my main dataset was very bad. There were many values that were inserted in a varying number of formats. I had to streamline these values to be able to group data and generate visualizations. Some of the variables that were not inserted correctly were gender and mental health and race. Column names had to be streamlined to be able to merge and join tables easily.

```
cols_renamed = [col.lower().replace(' ', '_') for col in cols]
cols = zip(cols, cols_renamed)
mydict = {}
for col in cols:
    mydict[col[0]] = col[1]
df.rename(columns=mydict, inplace=True)
#df
```

In [5]:
```
df['date'] = pd.to_datetime(df['date']) #changing the date format to generate a year column
```

In [6]:
```
df['year'] = df.loc[:, 'date'].apply(lambda x: x.year) #adding a year column
```

In [7]:
```
df.rename(columns={'s#':'incident_num',
                   'total_victims': 'total_casualties',
                   'employeed_(y/n)':'employed'
                  }, inplace=True)
df.age.fillna('0', inplace=True)
```

In [8]:
```
df.head(3)
```

Out[8]:

	incident_num	title	location	state	date	incident_area	open/close_location	target	cause	summary	...	policeman_killed
0	1	Texas church mass shooting	Sutherland Springs, TX	Texas	2017-11-05	Church	Close	random	unknown	Devin Patrick Kelley, 26, an ex-air force offi...	...	0.0
1	2	Walmart shooting in suburban Denver	Thornton, CO	Colorado	2017-11-01	Wal-Mart	Open	random	unknown	Scott Allen Ostrem, 47, walked into a Walmart ...	...	0.0
2	3	Edgewood businees park shooting	Edgewood, MD	Maryland	2017-10-18	Remodeling Store	Close	coworkers	unknown	Radee Labeeb Prince, 37, fatally shot three pe...	...	0.0

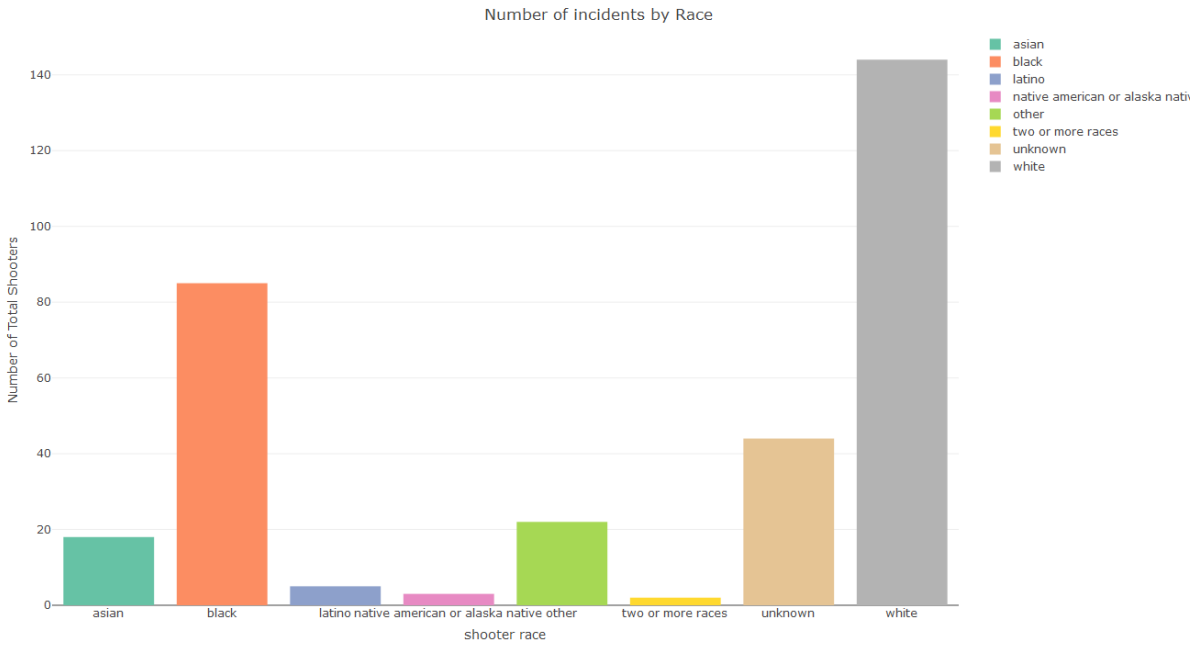
3 rows x 23 columns

4.Data Exploration & Findings

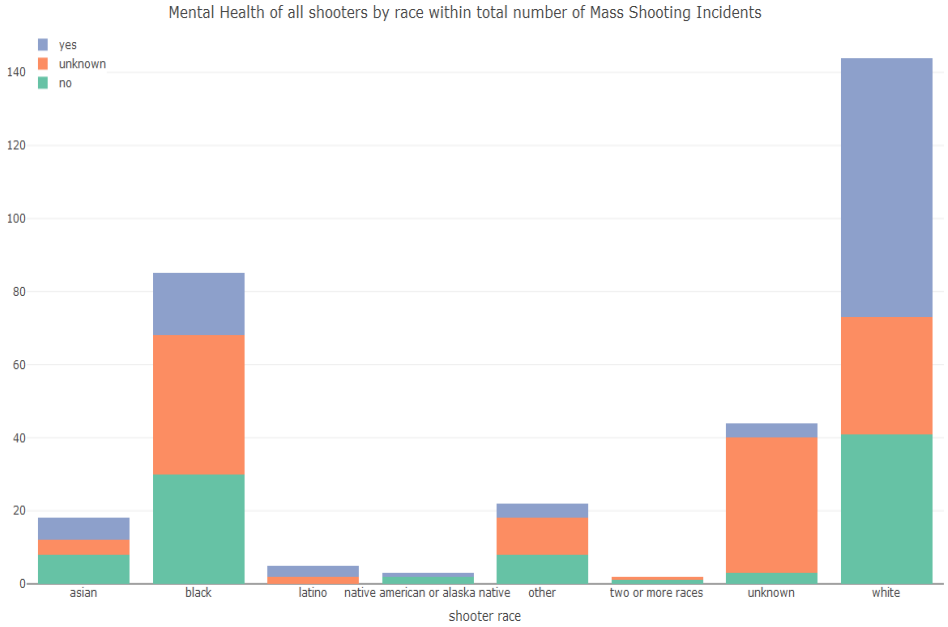
Exploration of Key Features in Dataset

I have used Rstudio in this part of the exercise to create histograms and a map to provide some visual identification of the trends happening with these Mass Shooting Incidents.

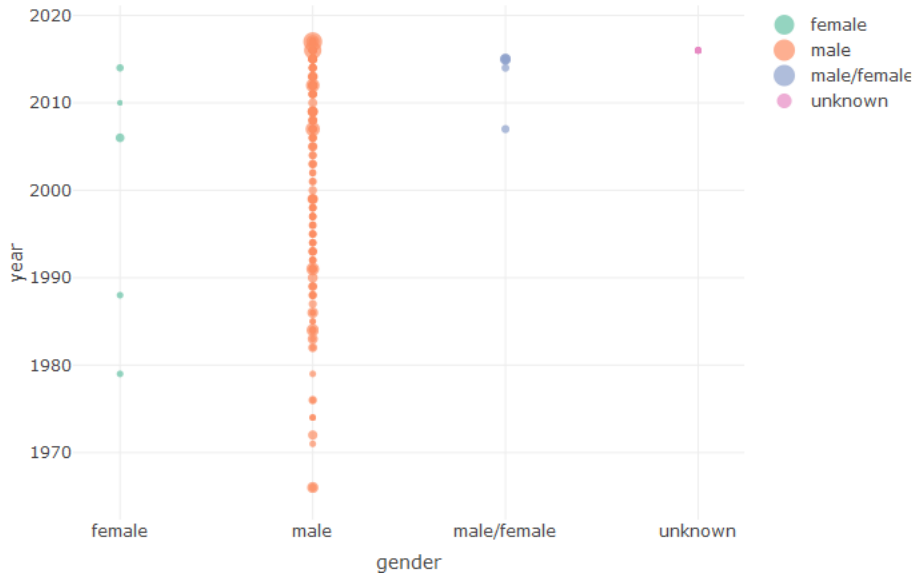
1) In the histogram below we can see that most mass shooters were ethnically white and in second position they black. Out of the total 323 incidents of mass shootings, more then 50% were of these 2 races of people.



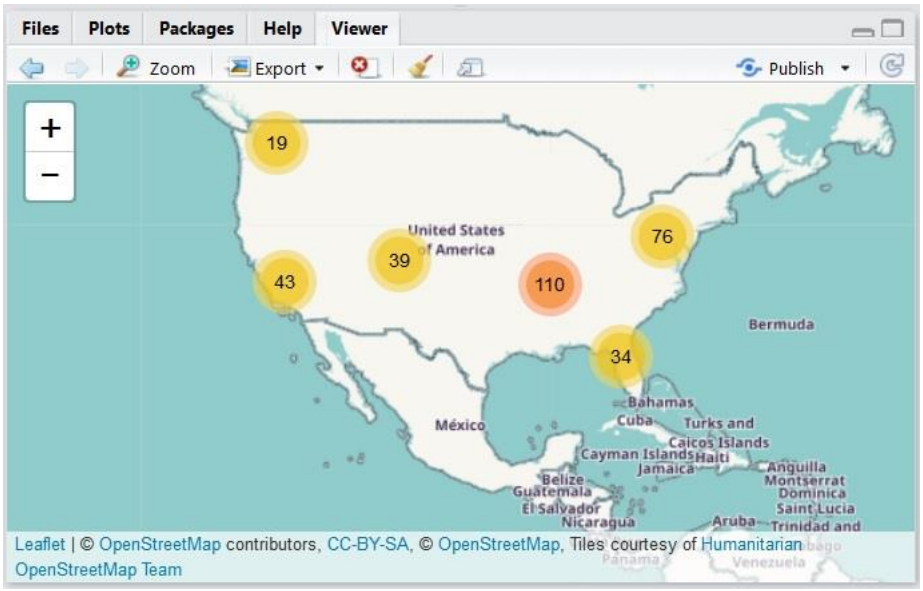
2) In this RStudio histogram below we can see that more than 50% of white mass shooters had mental problems. It is significantly higher among them then other ethnic backgrounds. We have not reasons for this.



3) In the scatter plot below, we can see that a clear majority of the shooters are male. The male shooters are highlighted in orange.



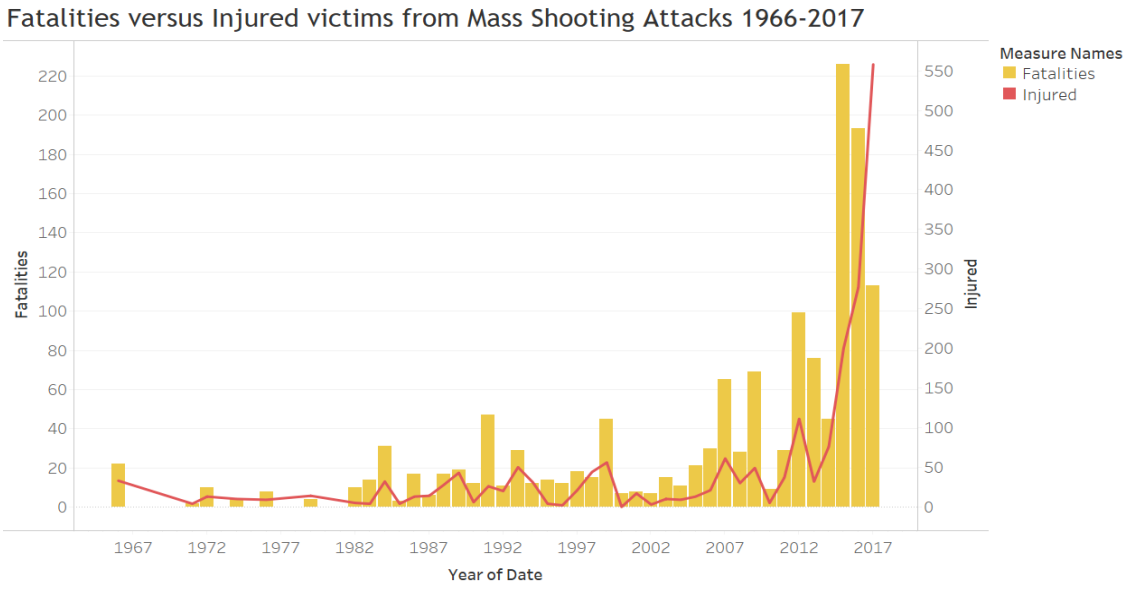
4) The map below shows the clusters of where all the shootings happen by region within the United States. The key spots could be clustered at the upper and lower East Coast, West Coast and Midwest States from visual observation.



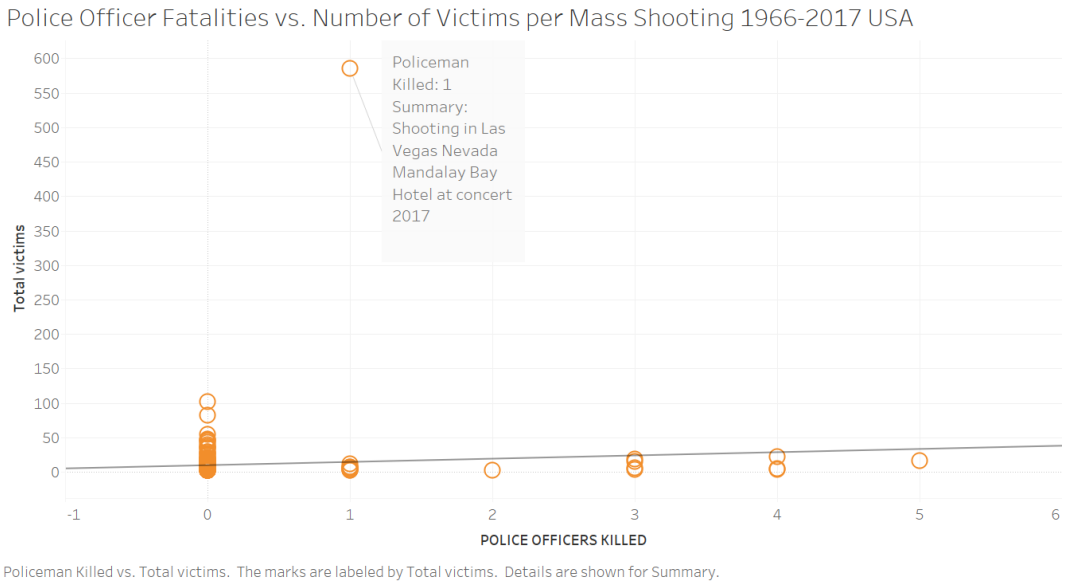
Code that was used in Rstudio for the map above is here (All Rstudio and Python code can be provided if required):

```
> KnownLocations %>%
+   leaflet() %>%
+   addTiles() %>%
+   addProviderTiles("OpenStreetMap.HOT")%>%
+   #addCircleMarkers()%>%
+   addMarkers(lng = ~longitude, lat = ~latitude,clusterOptions = ma
rkerClusterOptions(freezeAtZoom = 3)
+              ,label=~location)
```

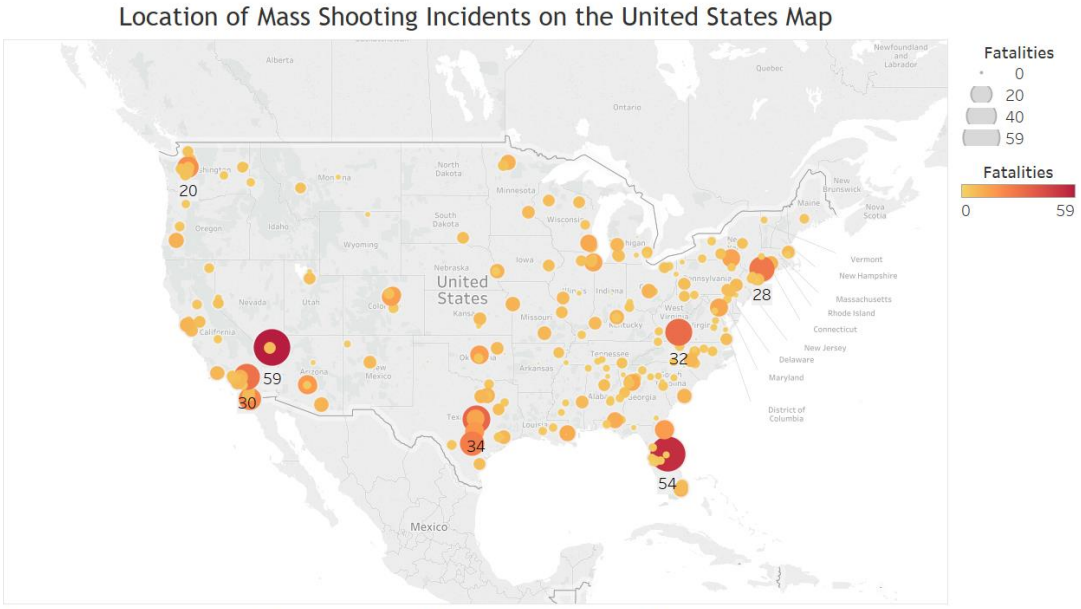
5) The following trend line/histogram/maps and charts were produced with tableau. The below chart shows how the number of fatalities is dropping but the number of injuries is increasing with mass shooting occurrences.



6) The following correlation line drawn in the plot below indicates that there is a low positive correlation between police officer deaths and the number of casualties from mass shooting incidents. The correlation value is only 0.007. With more incidents/deaths, there are more police officer fatalities, but it is not increasing at a significant rate.

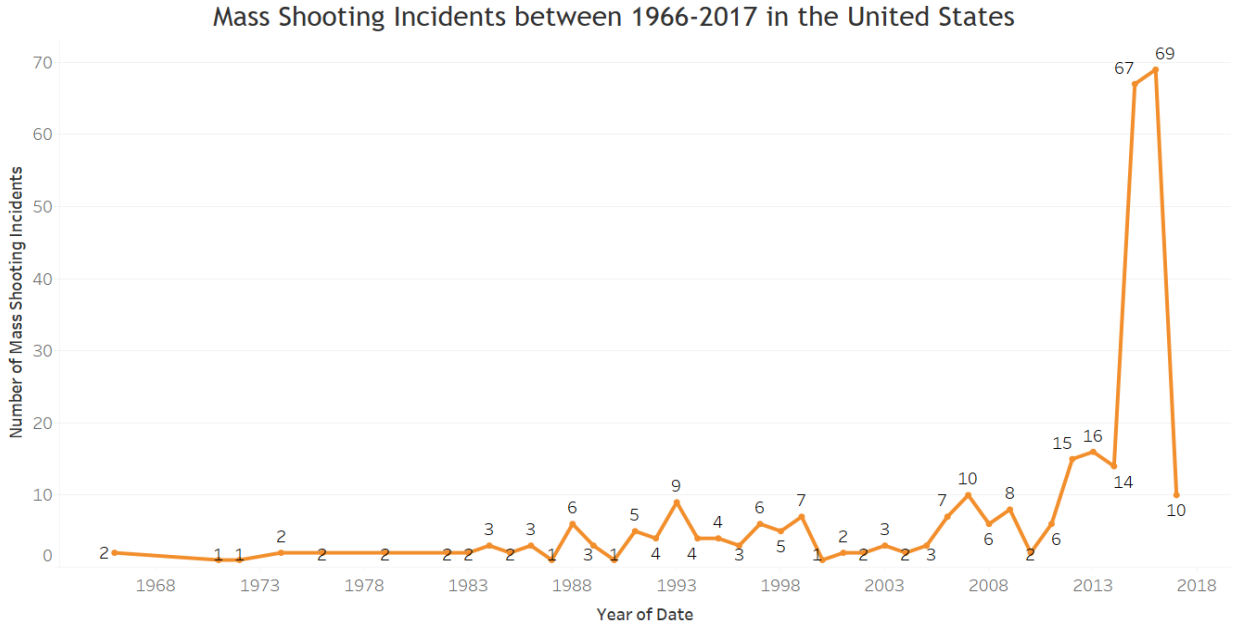


7) The following map indicates where the most violent mass shooting incidents have occurred. Each circle is an incident. The colour darkness indicates severity of the event by number of casualties.



Map based on Longitude and Latitude. Color shows sum of Fatalities. Size shows sum of Fatalities.

8) The map below shows the number of mass shooting incidents that have occurred and indicates a significant spike in mass shooting incidents between 2015-2017.

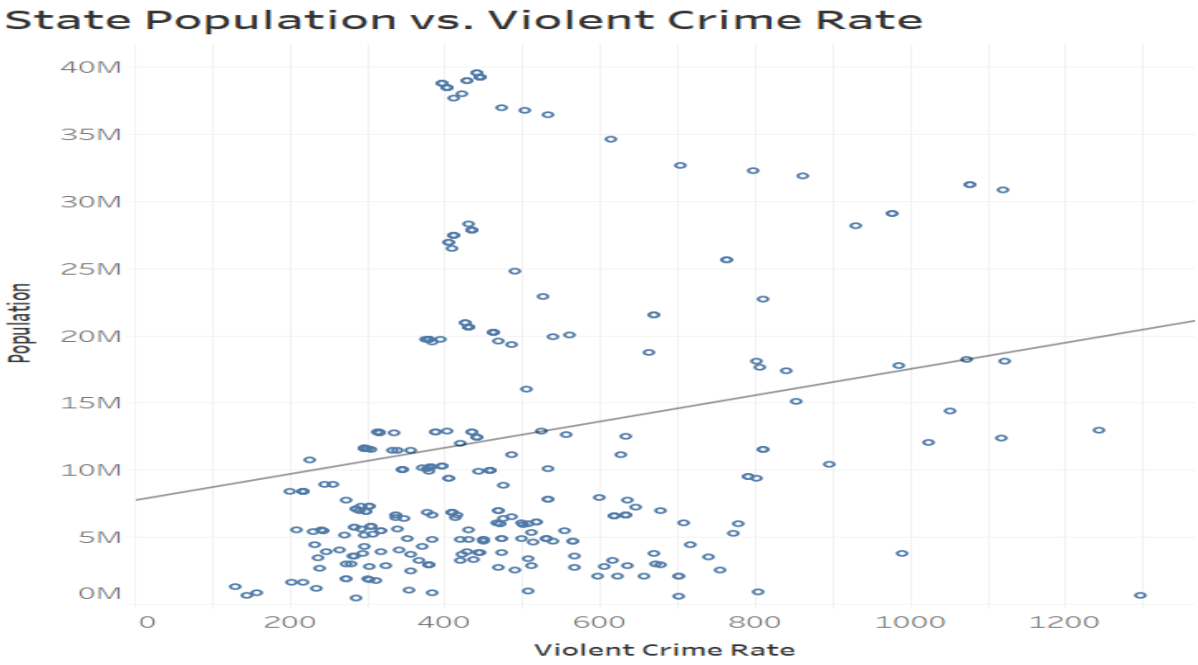


The trend of sum of Number of Records for Date Year.

5.Impact of Human Development Factors on Mass Shootings

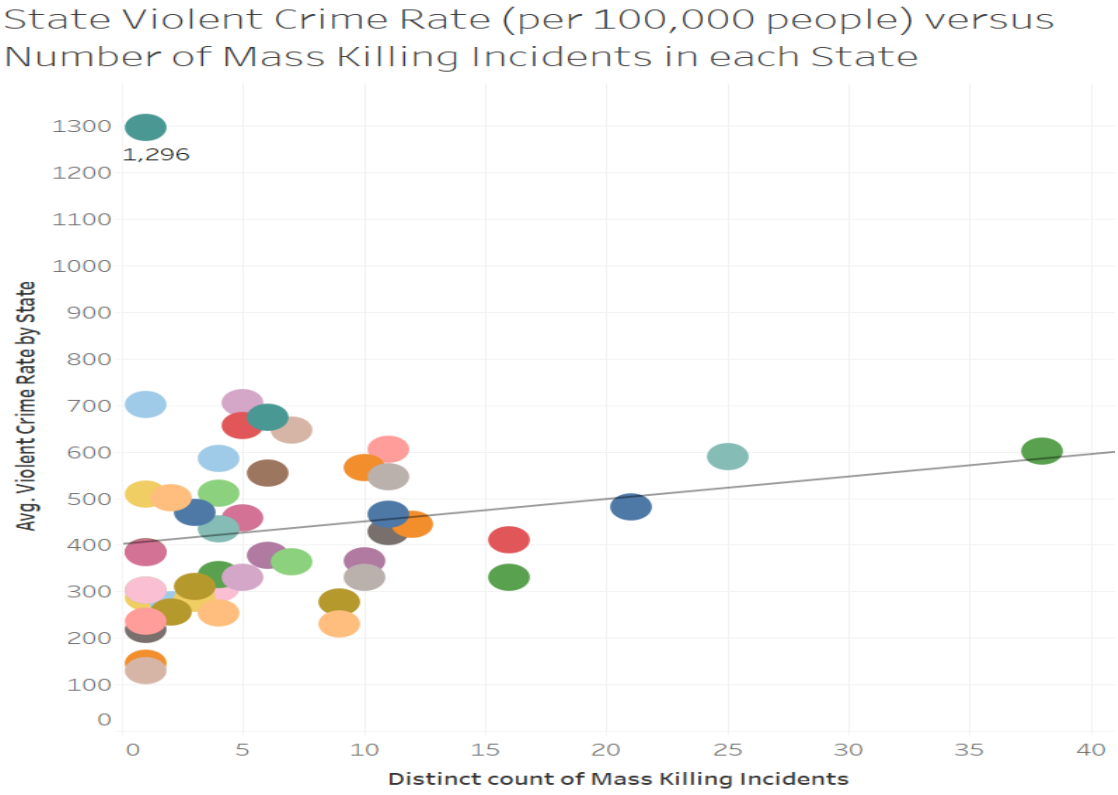
Impact of State Level Violent Crime Rate on Mass Shooting Incidents

The scatter plot below indicates that as the state population increases the number of violent crimes increase as well but the relationship is not strong because the correlation coefficient is R-Squared: 0.0342646. The trend line has a positive slope. All plots are using data between 1966-2017.



Sum of Violent Crime Rate vs. sum of Population. Details are shown for Incident Num.

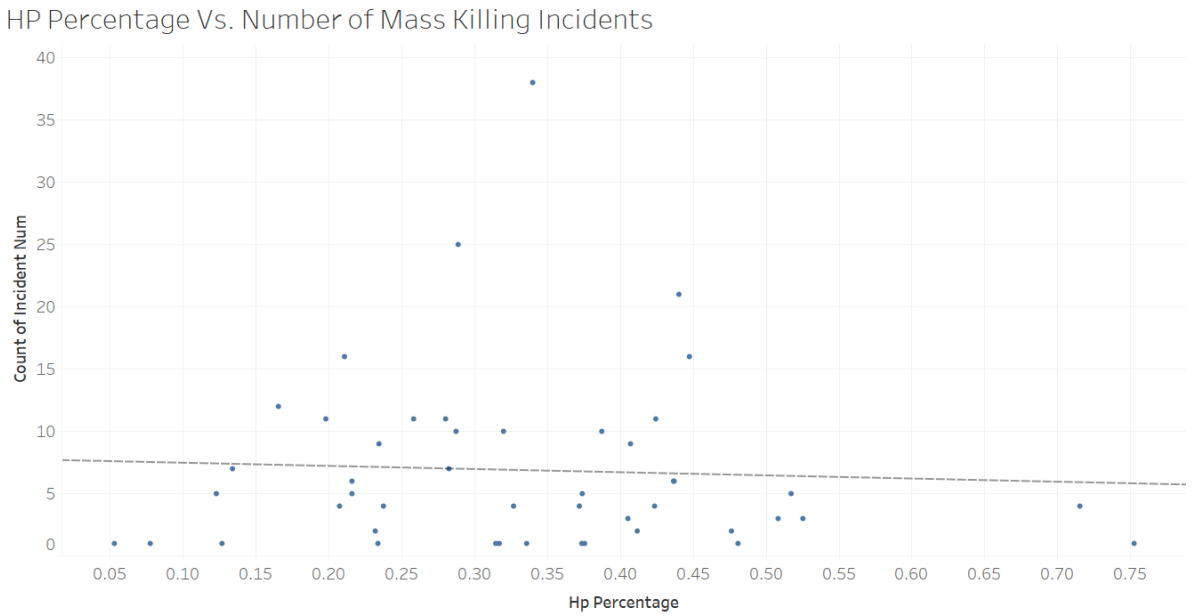
The scatter plot below indicates that as the rate of violent crime increases; the number of incidents of mass shootings also increase. Each bubble is a state and the legend has been removed because of space related considerations. The correlation coeffecient is 0.03 and it is very similar to the population versus violent crime plot.



Count of Incident Num vs. average of Violent Crime Rate. Color shows details about State.

Impact of Mental Health Specialist Shortage Rate on Mass Shooting Incidents

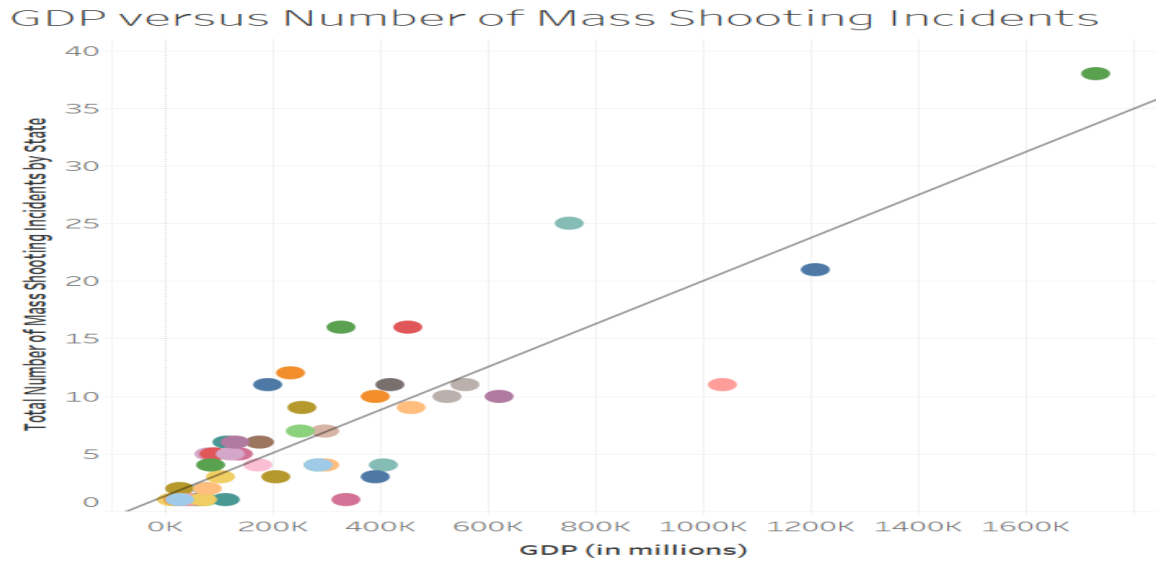
The HP_percentage is a percentage used to describe the % of mental health specialist shortage that has been fulfilled by the state. This means if 10 psychiatrists joined a state with 100 openings. The HP_pecentage is 10%. In the following scatter plot below, we can see that as a state increases its number of mental health workers required to manage the state's needs; we observe a decreasing trend in the number of mass incident shootings. There seems to be a correlation here. As we had observed earlier in the histogram about mental illness and race; approximately 30% of mass shooters and over 50% of white mass shooters were mentally unstable. The correlation coefficient here is 0.00267.



The trend of count of Incident Num for Hp Percentage. Details are shown for State.

Impact of GDP of state on Mass Shooting Incidents

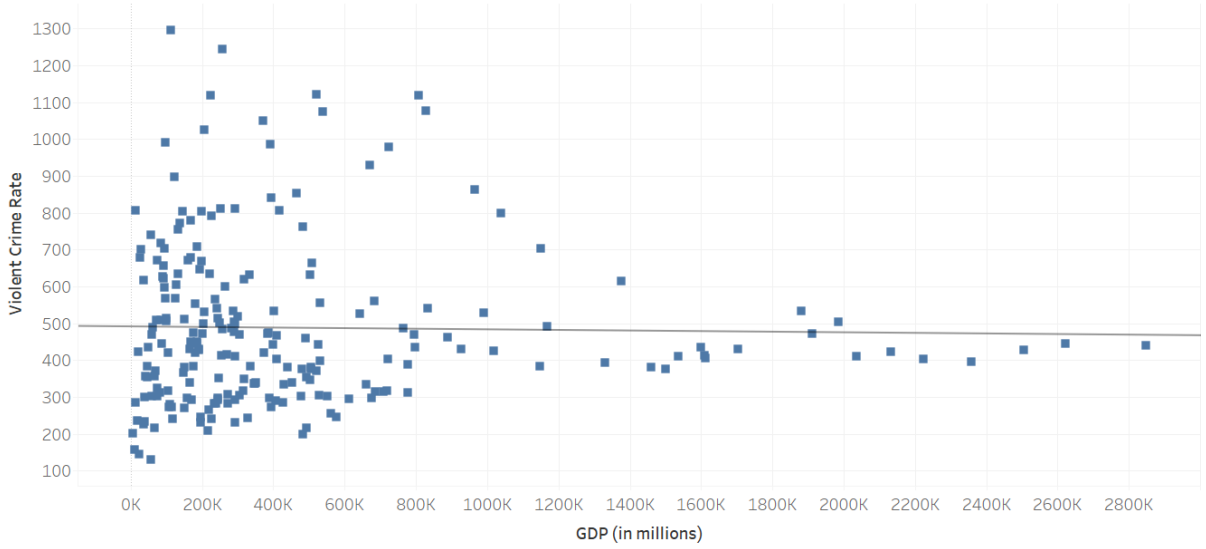
In the graph below we have plotted the total number of shootings that occur against the GDP of different states. It seems to be the case that as GDP increases; the number of mass shootings increase in each state. GDP is usually proportional to population of states as well.



Average of GDP vs. count of Incident Num. Color shows details about State.

In this following graph the correlation coefficient is 0.0003 with a negative slope and this is very low. However, it could still be significant as the relationship between both factors are strong. GDP does not seem to have a very strong correlation to violent crime.

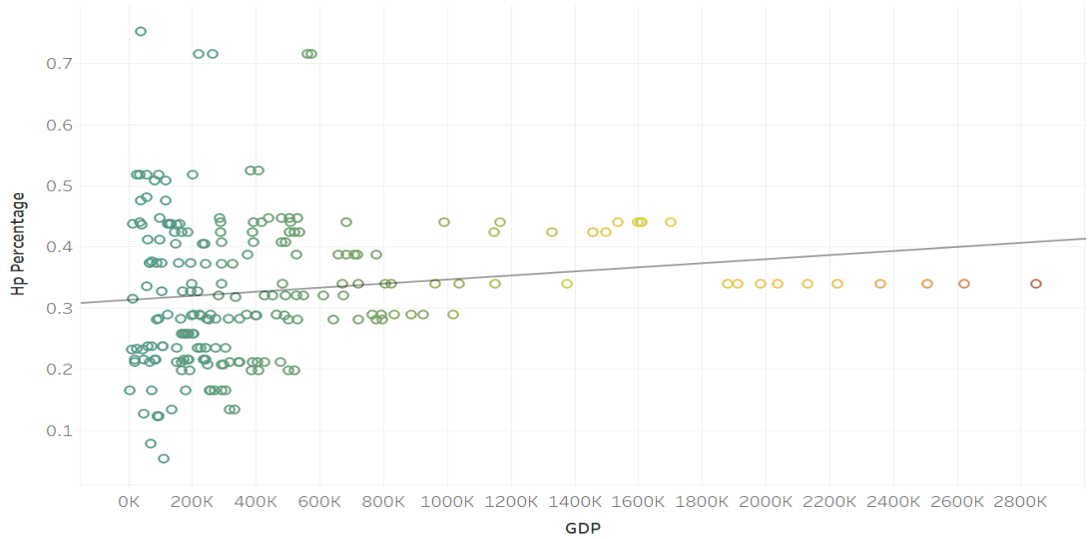
Affect of GDP and Population on Violent Crime Rates across different states (Violent Crime incidents per 100,000 persons in the population)



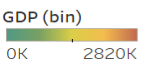
GDP vs. Violent Crime Rate. Details are shown for State.

In this following graph we can see that GDP has a direct impact on the ability of the state to hire more mental healthcare professionals to fill empty positions. The correlation coeffecient is 0.02 and the factors are strongly related to each other. Higher state income means more money to hire more people.

State GDP vs. HP_Percentage



GDP vs. Hp Percentage. Color shows details about GDP (bin).



6.Summary of Observations and Conclusion

(Note: From the analysis of the above factors, we can identify that a high percentage of the shooters had mental issues.)

1.Do the economic conditions of a state have an impact on the number of Mass Shooting incidents that occur? As the GDP of a state is bigger the number of mass shootings also increase. The number of violent crimes also increase, and the population is larger.

2.Does the number of Mental Health Professionals available in a state impact the number of Mass Shootings that occur? Based on the observations above there seems to be indication that with an increase in mental health professionals, the number of mass shootings decline. So if states focus on filling empty positions in mental health professions; there could be a decline in the number of mass shootings.

3.Does the existing overall violent crime rate within a state have an impact on the number of Mass Shootings that occur? There is a strong correlation indicating that a state culture of violence is proportional to rates of mass shootings. An overall curb in violence may have an impact on minimizing mass shootings if this correlation can be further tested.

4.Are there patterns or trends that can be observed within these Mass Shooting Incidents related to race, gender, mental health, injuries and fatalities? Most shooters are white, male, and have mental problems. Although the number of deaths has declined from shootings, the number of injured is only increasing at a much higher rate since 2016.

Final Summary:

Based on the visual observations and statistical correlations, it can be inferred that if a state increases the contribution of its GDP towards increasing the number of mental healthcare professionals needed by the state, more mental health patients would be treated and this could have an impact on the number of mass shootings being perpetrated by mentally unstable people in the United States. A decline in mass shootings could take place.

-END-

7.Additional References:

1. Chrisalbon.com. (2018). *Geocoding And Reverse Geocoding*. [online] Available at: https://chrisalbon.com/python/data_wrangling/geocoding_and_reverse_geocoding/ [Accessed 3 May 2018].

2. Follman, M., Aronsen, G. and Pan, D. (2018). *US mass shootings, 1982-2018: Data from Mother Jones' investigation*. [online] Mother Jones. Available at: https://www.motherjones.com/politics/2012/12/mass-shootings-mother-jones-full-data/ [Accessed 3 May 2018].

3. Swanson, J., McGinty, E., Fazel, S. and Mays, V. (2015). Mental illness and reduction of gun violence and suicide: bringing epidemiologic research to policy. *Annals of Epidemiology*, 25(5), pp.366-376.

4. Cover Page Image credit: https://everytownresearch.org/mass-shootings/