**Data Analysis: Factors of Gun Violence in the United States**

DATA 201: Thinking with Data

Prof. Sydney Pratte

December 9, 2021

**Group Members:**

Kusumpreet Kaur Heer (30114618)

Aaren Abigail Atmadja (30102099)

Nishan Soni (30147280)

Phuong Thao Nguyen (30118157)

Sarah Vinh Han Chung (30092553)

**Section A: Introduction**

In recent years, mass shootings have occurred with an increasing number in frequency and corresponding severity within the United States. According to the Gun Violence Archive (GVA, 2021), there are a total of 650 mass shootings in the U.S as of November 30th, 2021; from gun violence overall, the total number of injuries stands at 37,314, and the death toll stands at 1,986. Given the frequency and severity of gun violence occurring in the U.S, our goal is to study data on mass shootings to highlight the relationship between gun violence and factors to gun violence – gun laws (at the time of the shooting), mental health, race, sex, age, and relation to the shooting place – and what this relationship has on the implications on public safety and rights in the United States.By understanding the connection between the number of shooting incidents and the other factors pertaining to gun violence, primarily relating to mental health, the U.S government can form policies to prevent gun violence and the number of casualties and injuries.

Before reading our project, the reader should ask themselves, “how do mass shootings and mental illness relate to one another in terms of existing data and research done on gun violence in America?”. Both gun violence and mental health in the U.S is heavily researched and discussed on social media and traditional media. Additionally, the two are inexplicably linked to each other as something that is a status-quo in Western society, or in other words, something that only requires common sense to think of. Normally when you see a case with a mass shooting, there is likely internet and media buzz bringing the mental illness of the shooter or victim into discourse, whether true or false. So long as the reader understands how gun violence and mental illness are cemented as part of an unspoken status-quo, they would additionally need to have a basic understanding of American gun laws, how mass shootings are defined, and what mass shootings entail.

The GVA defines mass shootings as “any incident in which at least four people are shot, excluding the shooter.” (Silverstein, 2020). The definition of mental illness is noted to be “health conditions involving changes in emotions, thinking, or behaviour (or a combination of those)” by the American Psychiatric Association (“What is Mental Illness?”). In our project, the term “mental illness” is an overall blanket statement for any sort of mental health concerns (e.g. depression, anxiety, PSTD). According to Jonathan Masters for the Council on Foreign Relations, some notable gun laws include the Assault Weapons Ban of 2013, which was proposed (and rejected by the Senate), the Gun Control Act of 1968, and the Brady Handgun Violence Prevention Act (“global comparisons”, 2021).

For our project, we mined data from two sources: The Mass Shootings in America Database by The Stanford Geospatial Center, and the Gun Violence Archive. The Stanford Geospatial Center's database includes data from 1982 to 2017, while the data from the Gun Violence Archive is updated daily from 2013 to present day and onwards.

**Section B: Obtaining Data**

To choose reasonable data we had three different options from The Mass Shootings in America Database. Maintained by The Stanford Geospatial Center (Bernet et al., 2018), first being [MSA database](https://github.com/StanfordGeospatialCenter/MSA), the second was [Mother Jones database](https://www.motherjones.com/politics/2012/12/mass-shootings-mother-jones-full-data/) and the third was [Gun Violence Archive database](https://www.gunviolencearchive.org/mass-shooting). From these databases we used the MSA database for four of our visualisations, because the attributes provided in the MSA database such as “History of Mental Illness”, detailed “latitude” and “longitude”, “Total Number Of Victims”, “Date”, “Average Shooter Age”, “School Related”, etc. The data is dated from 1966 to 2016. While on the other hand Mother Jones data was quite limited when it came to row values. Most of the data values for the attribute “prior\_signs\_of\_mental\_issues” are either unknown, TBD or Unclear, which is not only pretty challenging to clean but even if it was cleaned properly only a few tuples would be left at the end of the cleaning. So due to these reasons we had to reject the Mother Jones database. The Gun Violence Archive Database has attributes such as, “state”, “incident id”, etc . This database has 2000 rows of data, which was used for plotting the graph to a much greater accuracy. This data is only dated from January 12 2018 to 21 November 2021. So we chose this database to have a more indepth visualisation of mass shootings, as the more tuples we have the more accurate our visualisation would be.

The Gun Violence Archive Database has attributes such as “state”. From these attributes the variable attribute “state” was used as a variable to plot the map and it was also used as a derived attribute for COUNT(state) while calculating the number of mass shootings per state. This visualisation was used to display states with highest mass shootings in the past 3-4 years.

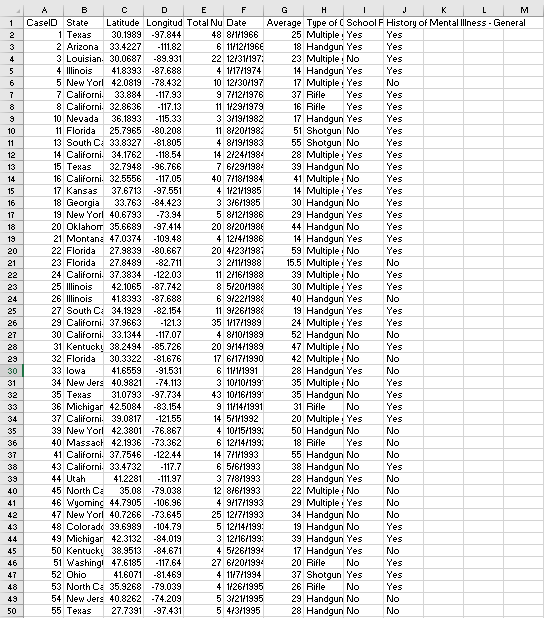
The MSA dataset has attributes such as “History of Mental Illness”, detailed “latitude” and “longitude”, “Total Number Of Victims”, “Date”, “Average Shooter Age”, “School Related”, etc. For the second visualization the attributes “Average Shooter Age” and “History of Mental illness”. Which is used to visualize mental illness related mass shootings. The third visualization uses the attributes, “state”, “CaseID”, it reports the state with the highest number of victims of gun violence shootings from 1966 to 2016. The fourth visualisation uses the attributes, “date” and “Mental Health related Shooting”, it compares the number of mental health related shootings and other shootings over time and if there is a correlation between mental health related shootings and time. The fifth visualisation uses the attributes, “Total number of victims” and “School related” which is used to visualize how likely it is that a school shooter with mental illness would commit a school shooting rather than someone with no signs of mental illness.

Links where the datasets were downloaded from:

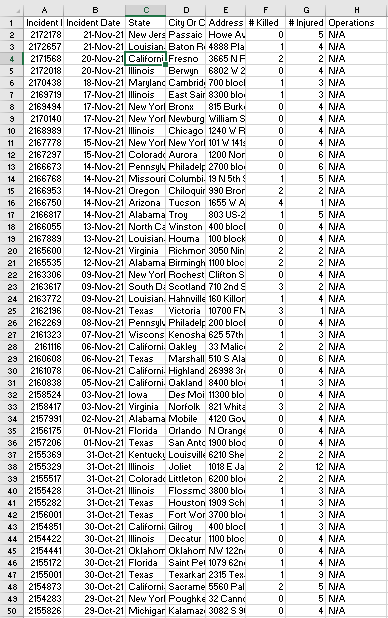
[Stanford Mass Shootings in America(MSA), courtesy of the Stanford Geospatial Center and Stanford Libraries](https://github.com/StanfordGeospatialCenter/MSA)

[Gun Violence Archive](https://www.gunviolencearchive.org/)

**Stanford Mass Shootings in America (MSA)**

****

**Gun Violence Archive**



**Section C: Analysis Questions**

**Gun Violence Data Set**

1. Which state has the most mass shootings?

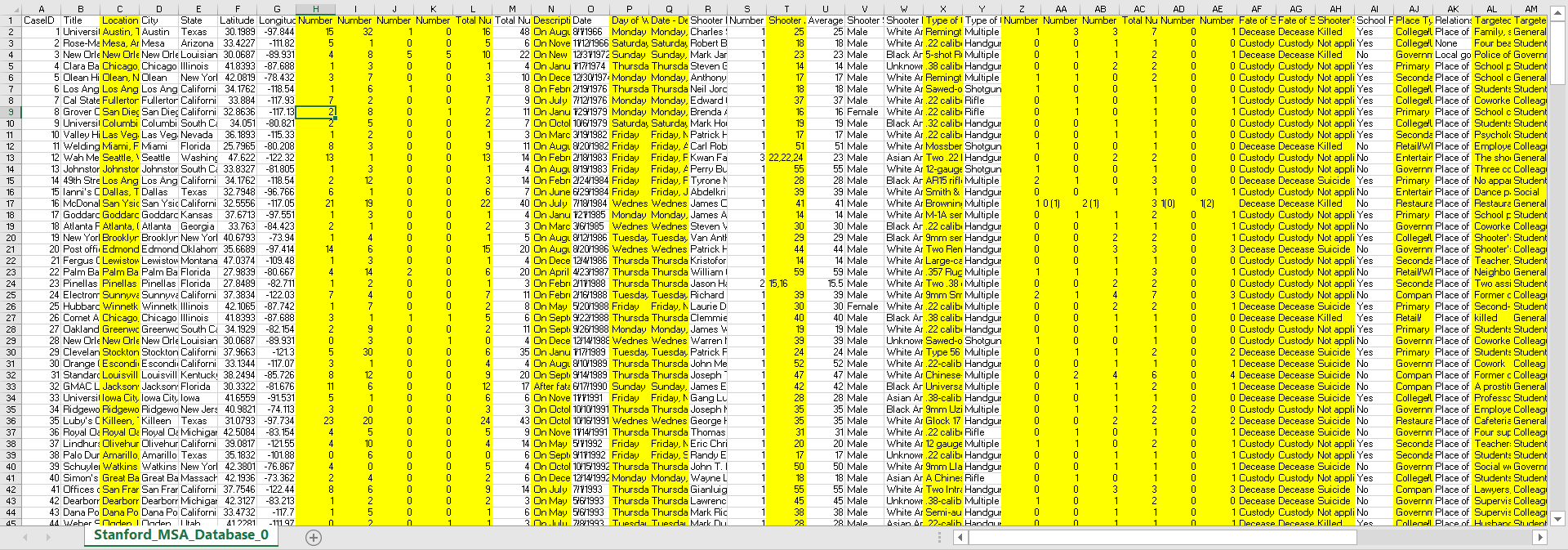
**Stanford MSA Database**

1. Which states have the highest number of victims of mass shootings?
2. What is the age of the shooter where there is a maximum number of mental-illness-related shootings?
3. What is the trend of mental health-related shootings over time?
4. Are there more mental-illness school related incidents than non mental illness school related incidents?

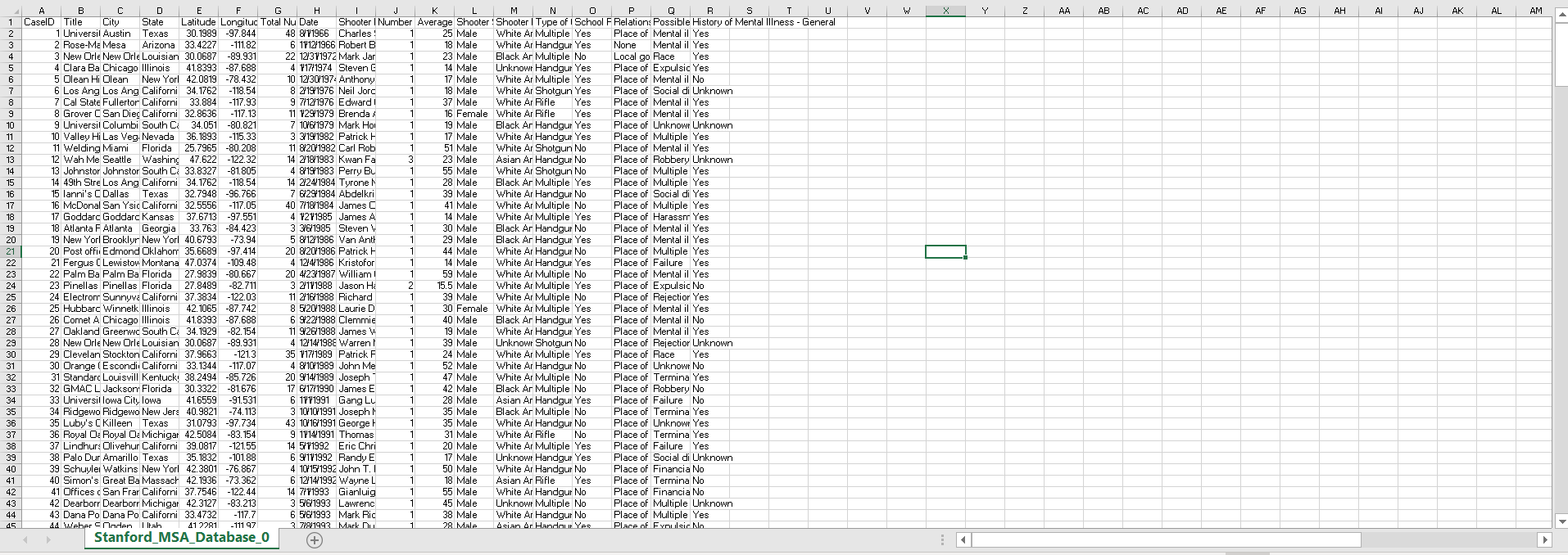
**Section D: Cleaning Data**

1. **Removed Irrelevant Columns**

Before:

****

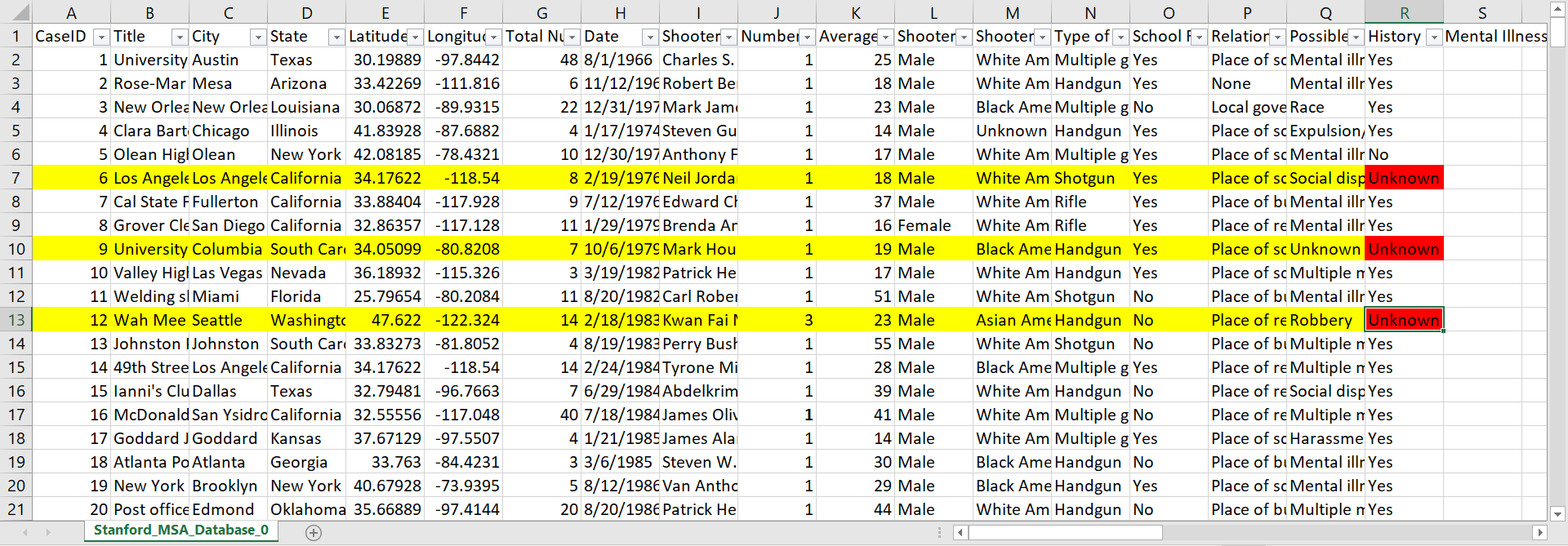
After:

****

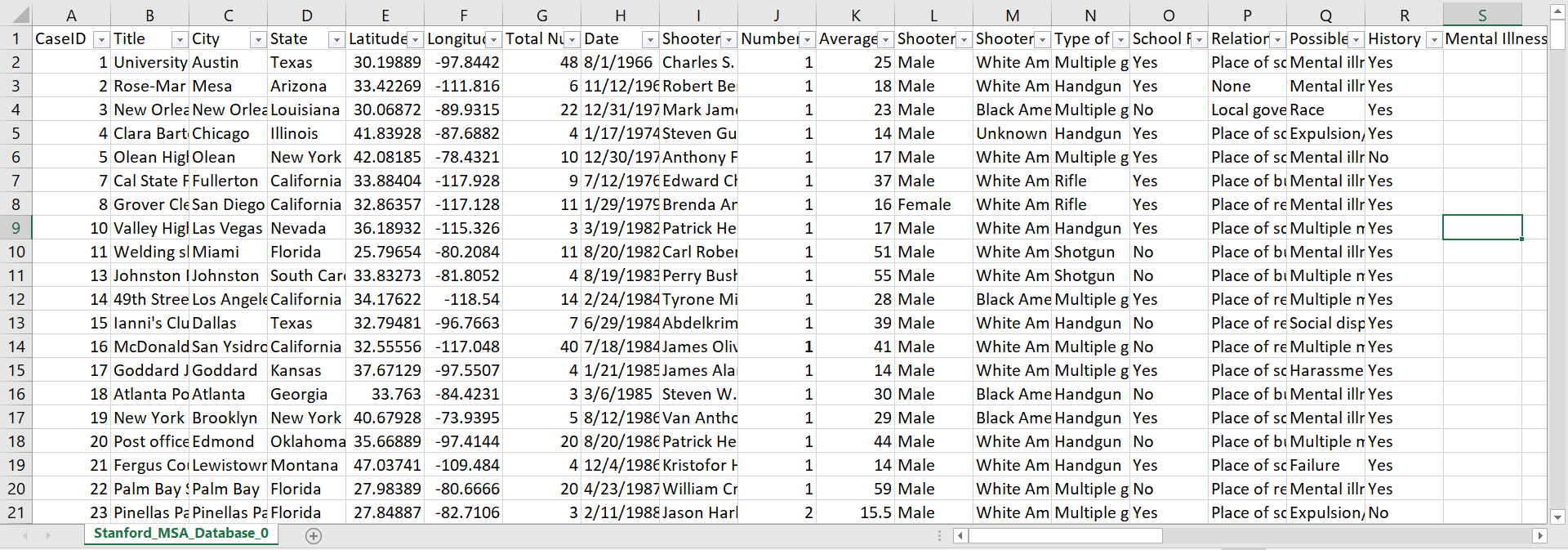
The screenshots show that there are several columns that have been deleted. The columns deleted do not have any relevant values that we deemed to be helpful for our analysis of gun violence and shooting incidents in relation to mental health background, therefore there is no reason to keep them in the dataset. Our analysis questions are tailored to include the characteristics of the shooter, victims, and relations to mental-health backgrounds to aid our evaluation process. Therefore, keeping columns that are irrelevant to our analysis thesis and objectives will only take up unnecessary space, and will only confuse the process of reading data if left uncleaned. It is important to clean these empty columns to reduce distracting spaces in the dataset to prevent errors during data analysis and visualization processes. Once these columns are removed, the dataset will appear to be more cohesive and organized. It will also be more accessible and readable to the reader. In the case of data analysis, removing empty records will allow for more concise results. It will also reduce the chances of misreading and referencing the wrong values in relation to other columns. Having too many columns with data that is unrelated to the analysis or evaluation will be In the case of data analysis, removing empty records will allow for more concise results.

1. **Removed Rows containing “Unknown” in the History of Mental Illness – General column.**

Before:

****

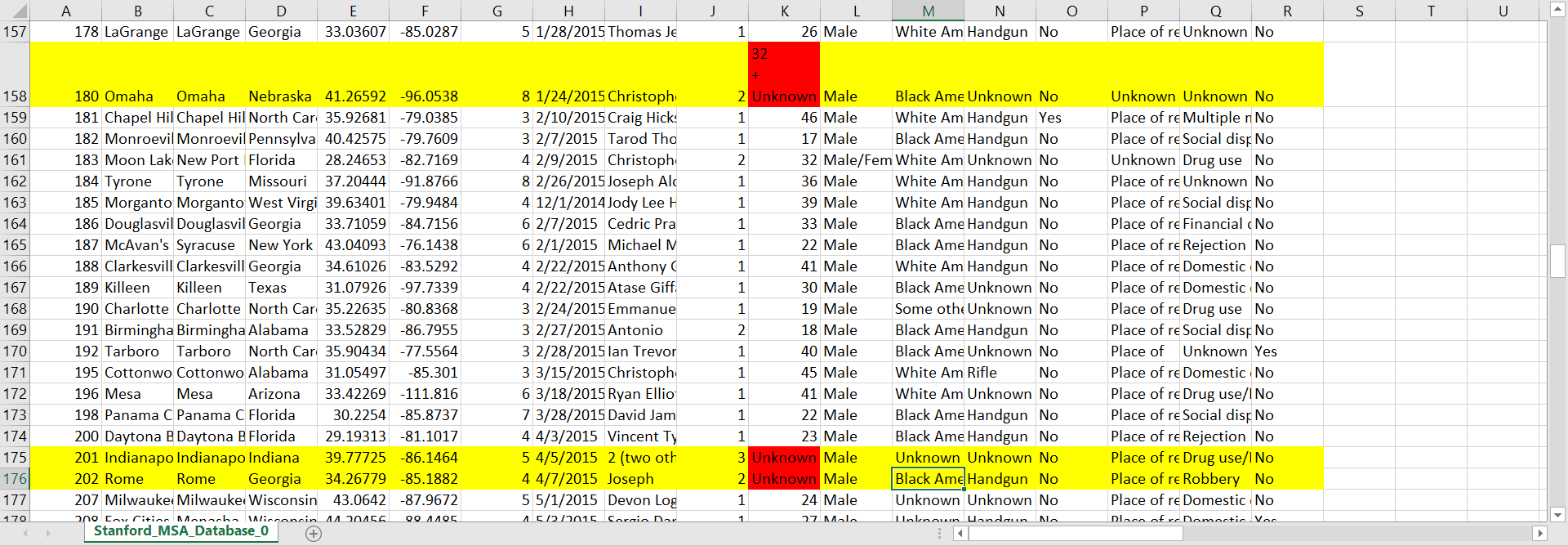
After:

****

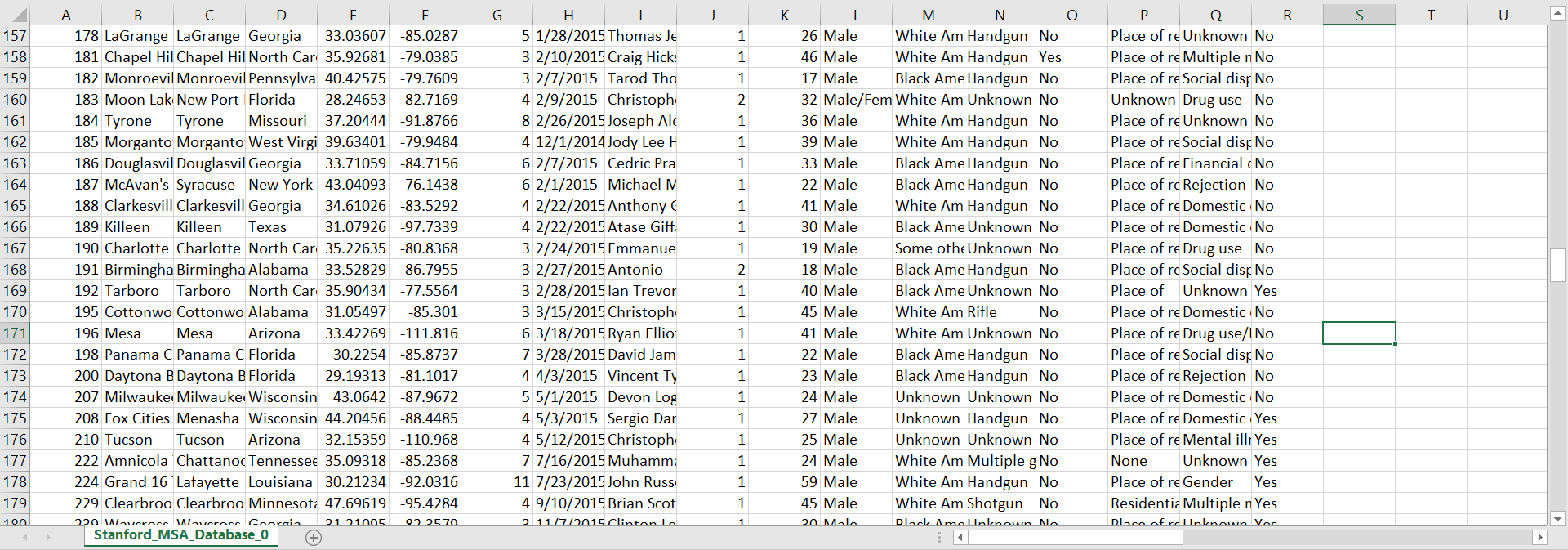
The screenshots above show that several rows containing unknown values in the history of mental illness column have been deleted from the dataset. Since our analysis and evaluation heavily depends on data that clearly distinguishes shooting incidents that are mental-illness related, and those that are not, it is not relevant to include data values where the mental health history of the shooter is unknown. To draw distinct correlations in our visualizations and data analysis, we need to make sure that the finalized dataset can reflect and represent reality accurately. To do this, removing shooting incidents where the mental illness history of the shooter is unknown seems to be the most reasonable thing to do. It is crucial to state however, that there is no clear line that we can draw to connect the history of mental health of the shooter to the shooting incident itself, nor can we draw a connection between the two attributes as a cause or effect of each other. Deleting rows that hold unknown values for the shooter’s mental health history is relevant to our analysis and the integrity of the data we are using as the basis of our analysis. Removing unknown values can only help prove the validity of our visualizations and evaluations and further deduce a more reliable and cohesive conclusion to our analysis.

1. **Removed rows containing “Unknown” in the Average Shooter Age Column**

Before:

****

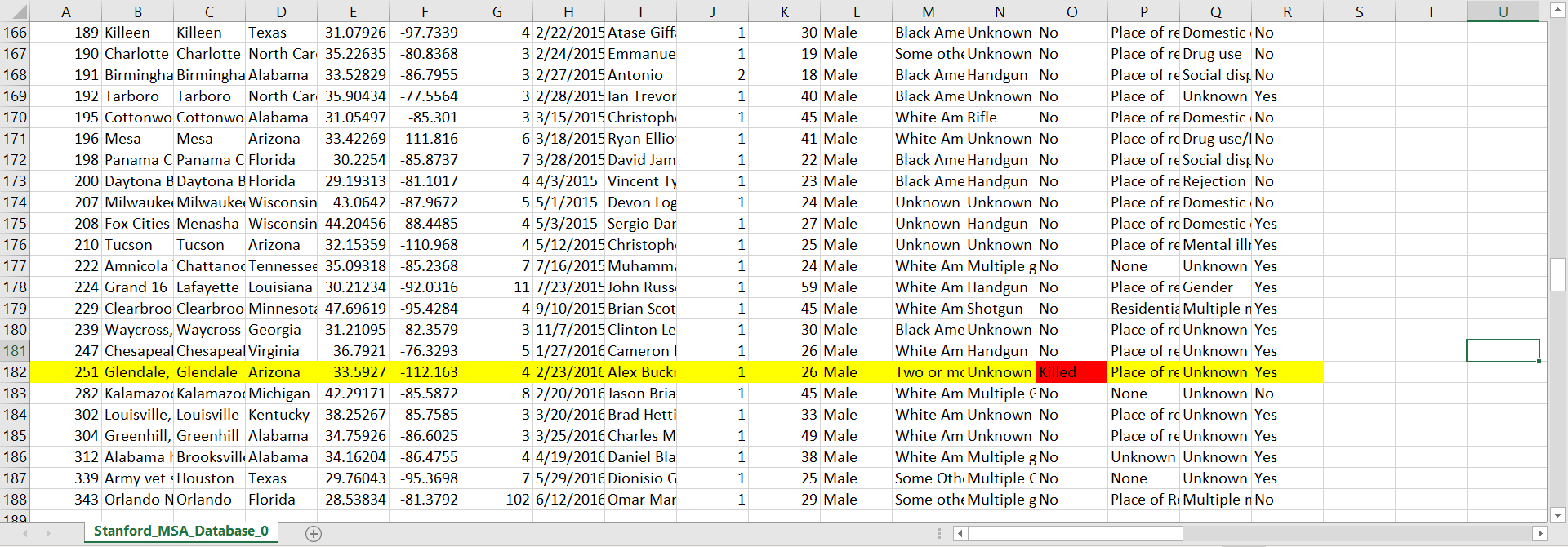
After:

****

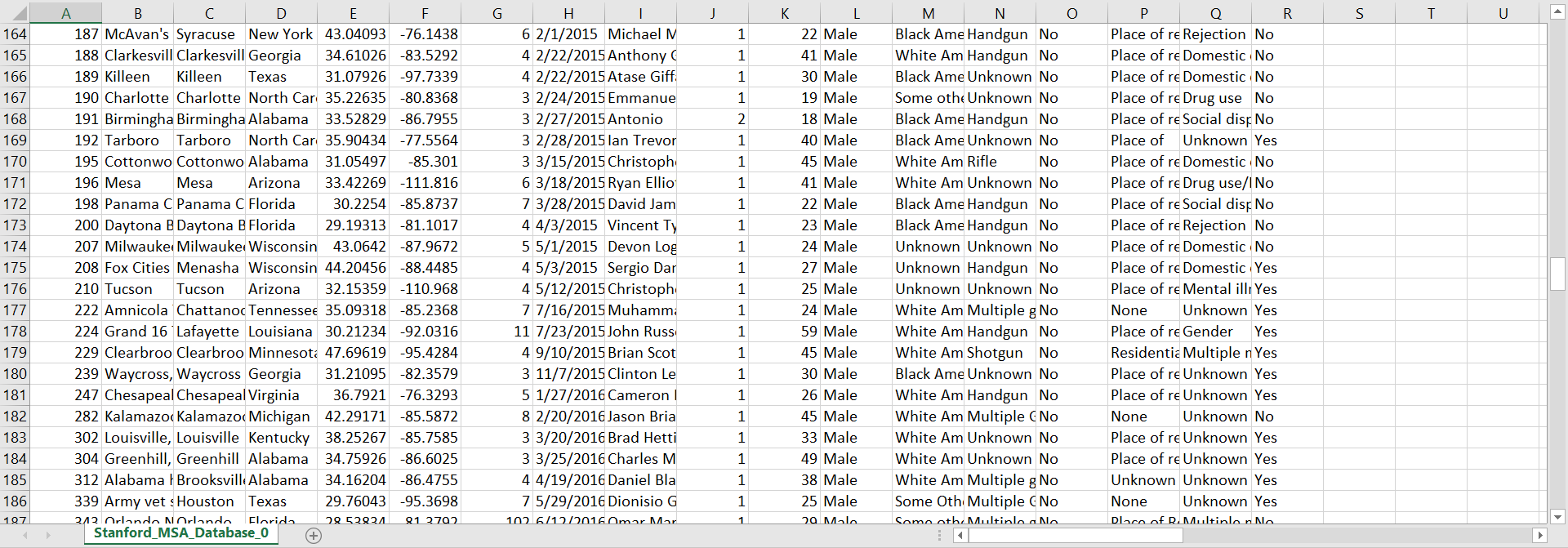
The Average Shooter Age column needed to be cleaned because it contained incorrect data and unknown values. Row 158 included the values “32 + Unknown” in the Average Shooter Age column resulting in an incorrect input. This input most likely implies the sum of two ages instead of an average. Other rows contained “Unknown” in the Average Shooter Age column, which represents N/A or blank values. Removing these rows was beneficial because the “Unknown” values would serve no purpose for the data analysis. This is because analysis regarding the Average Shooter Age would not be performed on those rows. Although the Average Shooter Age column is only used for one analysis question, keeping the rows would result in other visualizations having more data points than the visualization that uses this column as an attribute. This would make the chart misrepresentative of the data set as it contains fewer data points than the other charts. Furthermore, the Average Shooter Age column contains numbers, so non-numerical values (such as the “32 + Unknown”) can cause errors during the analysis, possibly producing inaccurate results. Ultimately, removing these rows will result in a more accurate analysis that is representative of the data set.

1. **Removed Rows containing “killed” in the School Related column**

Before:

****

After:

****

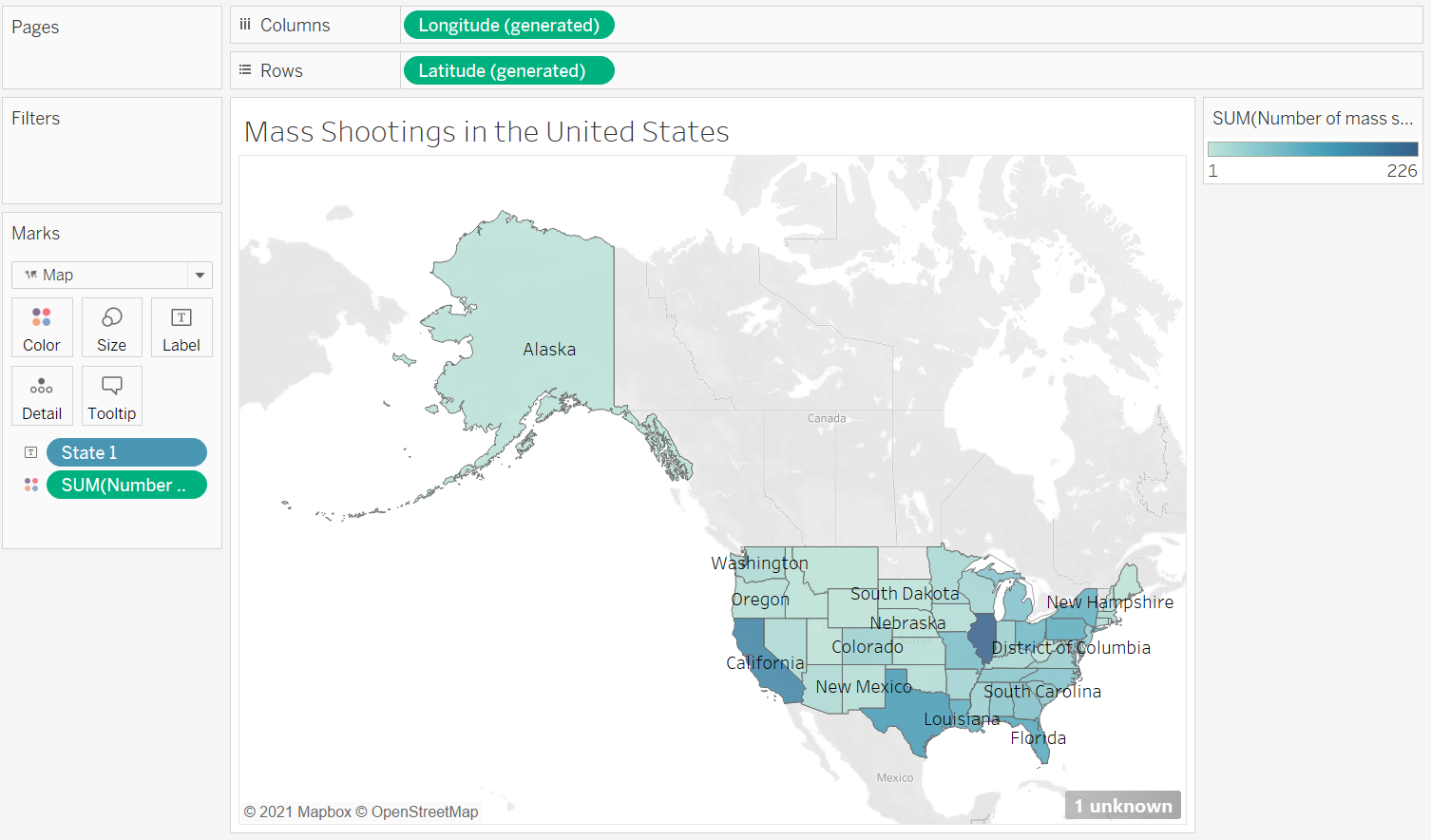
The row holding the "Killed" value in the school-related column (row 182) has been deleted as can be seen from the screenshots above. Rows with the value "Killed" in the "School Related" column will cause disturbances in the data that make it difficult to visualize because the two values that are suitable to appear in this column are "Yes" or "No". Eliminating shooting instances where the school-related value is "Killed" is the most appropriate way for us to create visualizations that correspond to the data being analyzed. Removing the "Killed" value in this column helps us to most accurately reflect the actual data because the ratios are not biased due to the presence of confounding factors. The clean data after deleting rows with the value "Killed" in the school-related column is ready for analysis on the relationship between mass shootings and schools because it is no longer affected by redundant factors leading to inaccurate visualization and statistics. Deleting rows that include "Killed" values for the school-related column is important to our study and the integrity of the data we're working with.

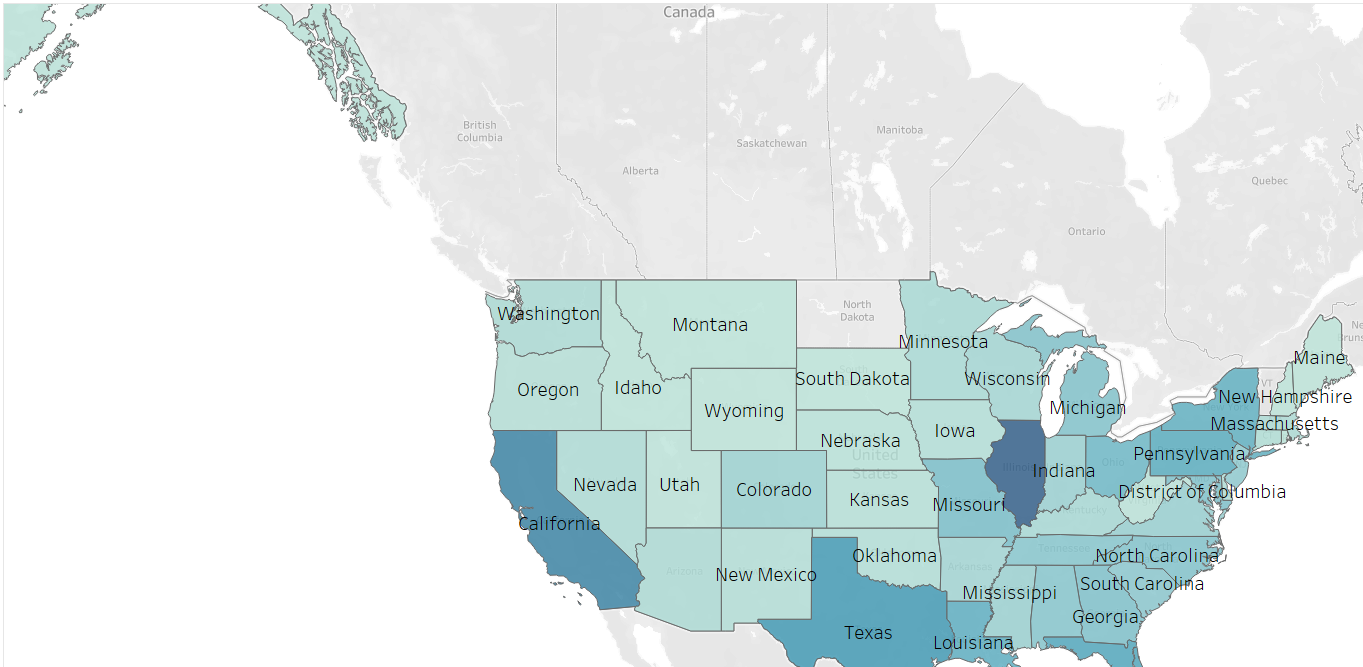
**Section E: Visualizations and Findings**

**Gun Violence Data Set**

1. **Which state has the most mass shootings?**

The visualization created by Tableau to demonstrate the number of mass shootings in the United States from 2018 to 2021:



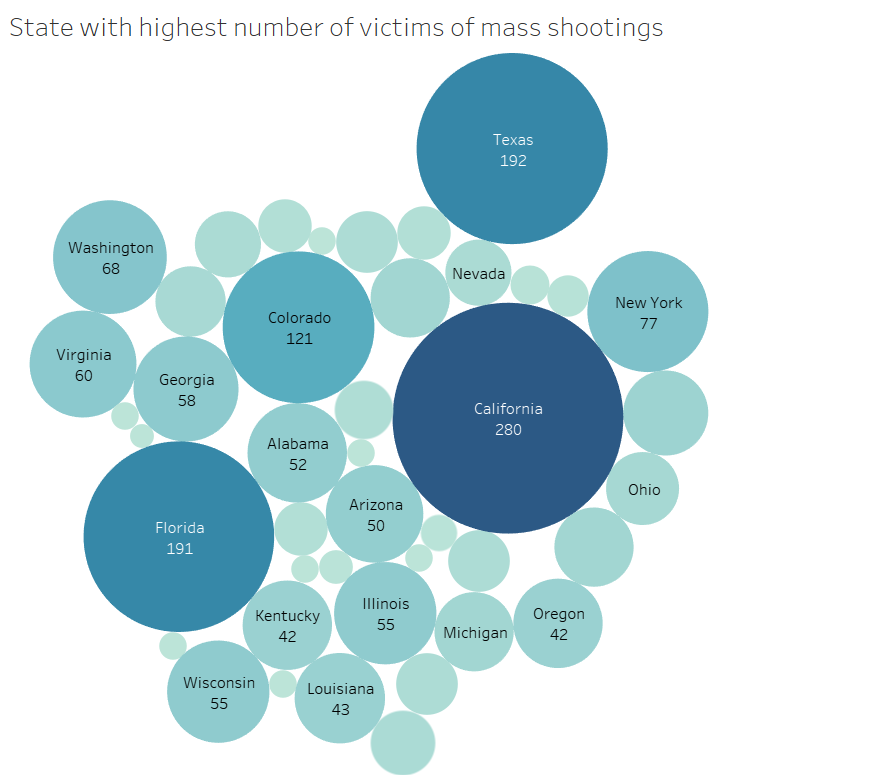


The question requires us to show the number of mass shootings in all states in the US from 2018 to 2021 and from there make a comparison between the data to find the state with the highest number of mass shootings. To show the quantitative values ​​for individual locations, a symbol map is the best. The given symbol map represents the data based on the geographical map of the United States, the number of mass shootings is expressed from the lightest to the darkest of blue, corresponding to the number of mass shootings from 2 to 399. Symbol maps use colors to represent data so it is easy to understand and evaluate. Not only that, representing the data of the states using a map makes it possible to see the distribution of mass shootings across the US and easily identify the areas with the most mass shootings. Using the symbol map we can mine the data and answer the question posed, so this is the right visualization for this part.

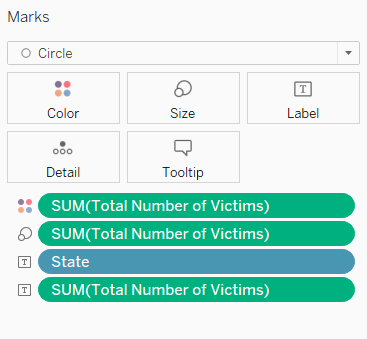
The visualization shows the number of mass shootings in all states of the United States. The lightness of blue shows the number of mass shootings, the darker the color, the higher the number of mass shootings in that area. The locations and names of the states are clearly indicated because the graph used is a symbol map. From this map, we can easily see that Illinois is the state with the highest number of mass shootings because it is the area with the darkest blue color. Followed by California and ranked 3rd in terms of mass shootings is Texas. The remaining states do not differ too much in the number of mass shootings because the blue colors of these states are quite similar. Based on the distribution of mass shootings in the states, we can see that the number of mass shootings is not greatly affected by geographical location because the states with high mass shootings are not concentrated in a specific region. Analyzing the attached symbol map, we were able to answer the question posed: the state with the highest number of mass shootings is Illinois.

**Stanford MSA Database**

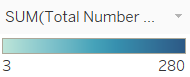
**2. Which state has the highest number of victims of mass shootings?**

****

This visualization was created using Tableau. The reason for this visualisation for the analysis question is that size and color combined helps us determine the maximum value very easily. As we can see from the visualisation, California has the darkest shade of blue with the largest circle and it says “280” victims (which is the highest). The attributes of the data base are as follows:

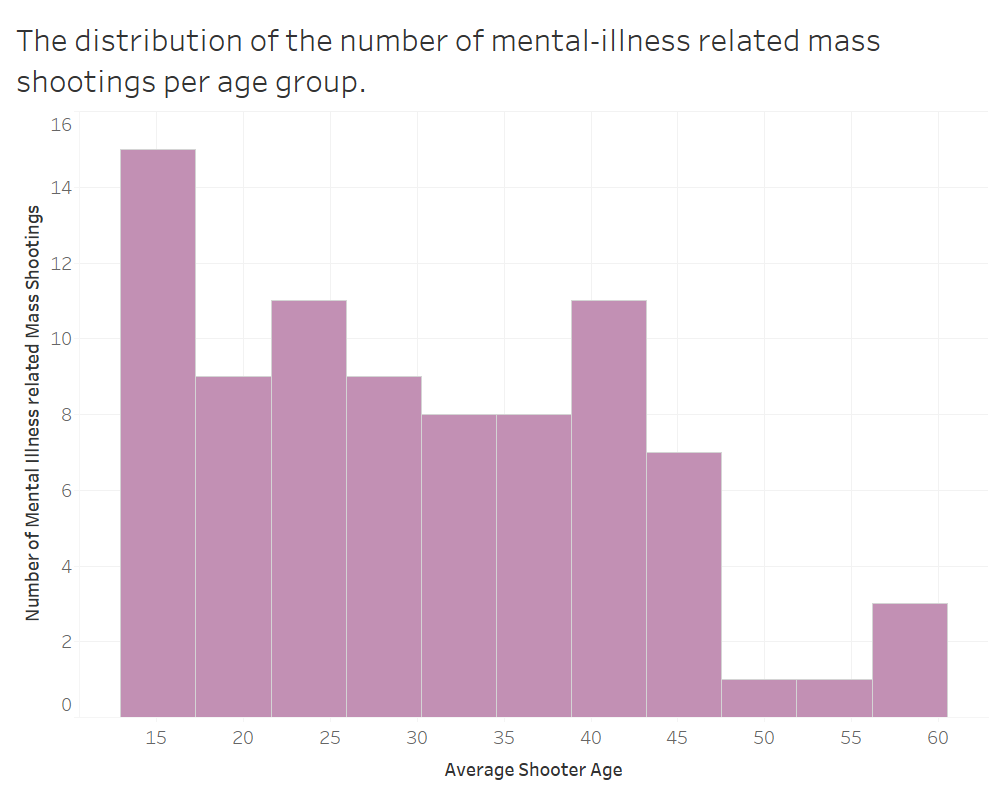


The attribute “State” is used as a label for each circle and the attribute “Total Number of Victims” is evaluated for each state using the SUM(Total Number of Victims). Further the calculated attribute SUM(Total Number of Victims) accounts for the size and the color of each state value shown as follows:



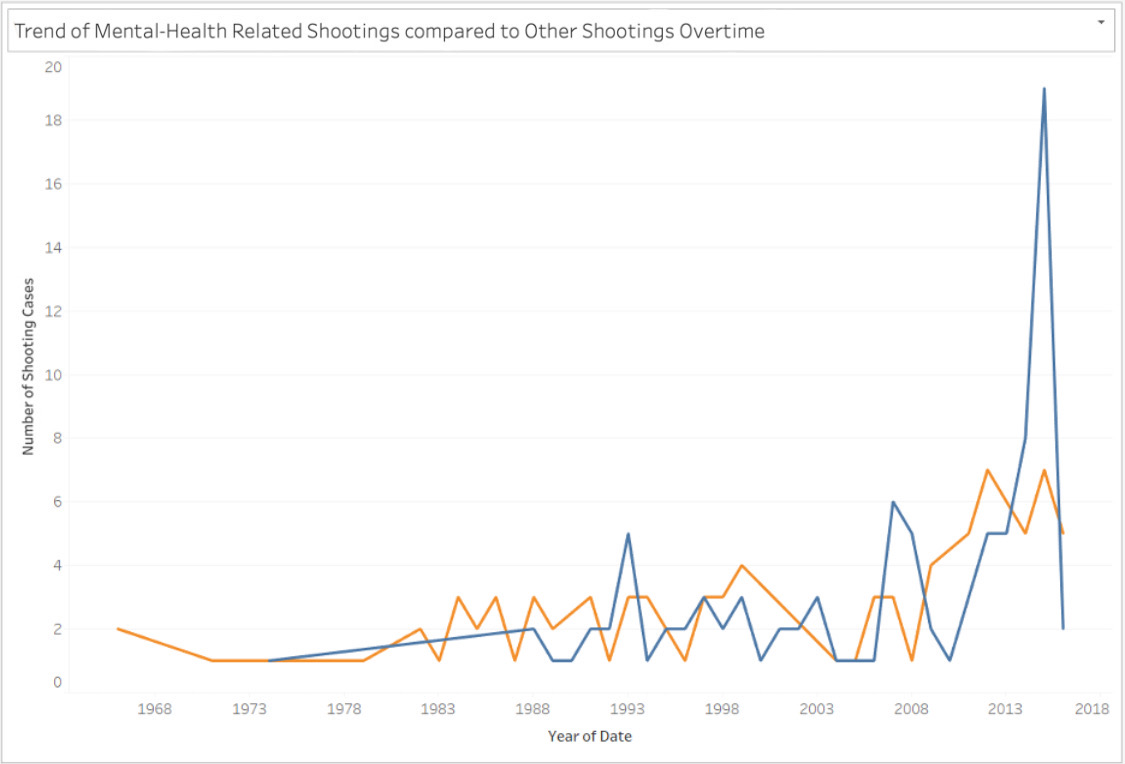
From the visualisation we can see that California has the deepest shade of blue and has the biggest circle therefore, California has the highest number of victims in mass shootings (280). As seen from the visualisation the states with highest number of mass shooting victims in descending order are California (280) > Texas (192) > Florida (191) > Colorado (121) > Washington (68).

**3. What is the age of the shooter where there is a maximum number of mental-illness related shootings?**



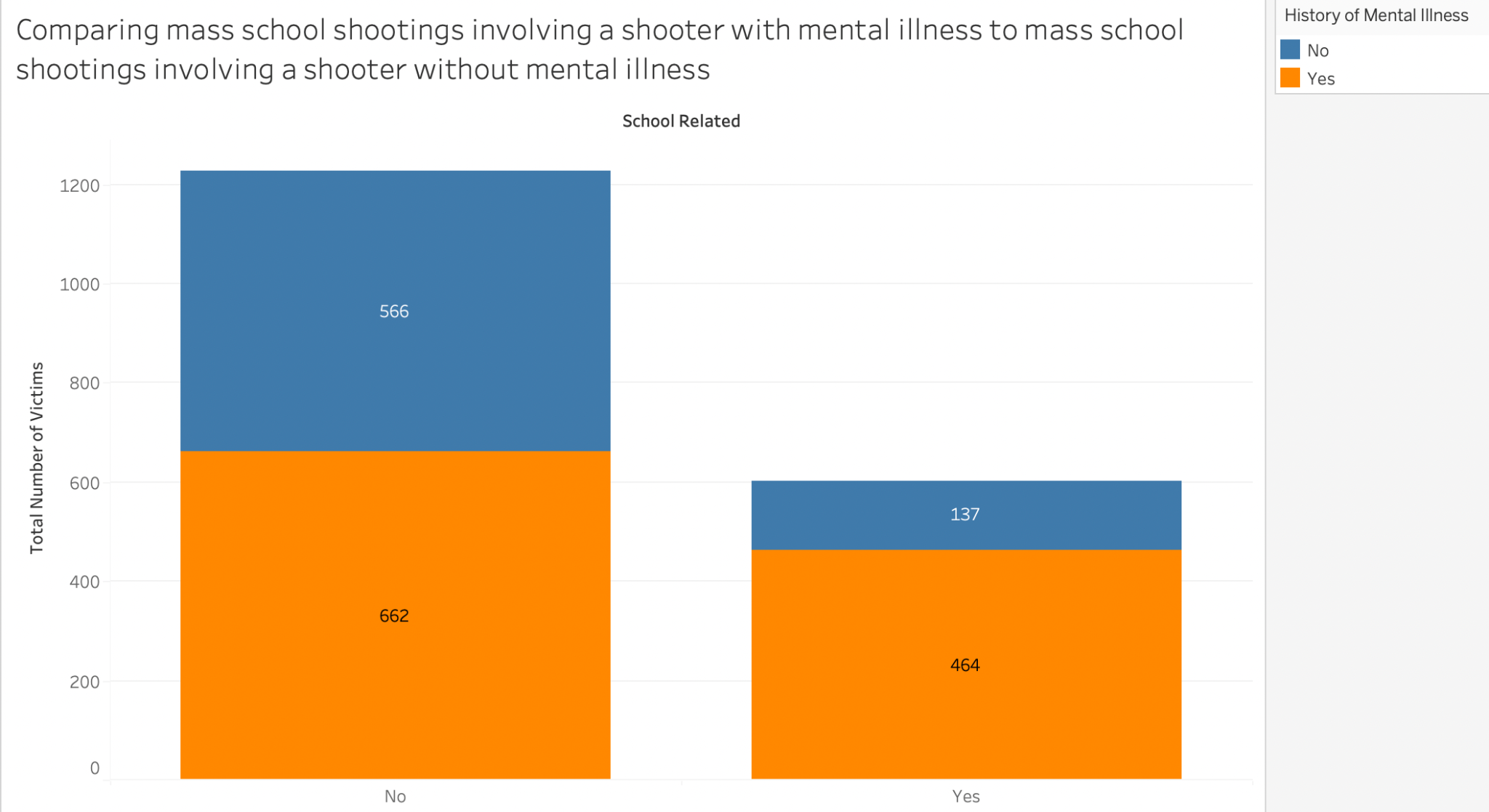
The visualization shown above is a histogram displaying the distribution of the number of mental illness-related mass shootings for each age group. The x-axis shows the average shooter age, and the y-axis represents the number of mental-illness-related mass shootings. This visualization was chosen because histograms display the distribution of values in a dataset, which will aid in finding the number of shootings per age group. The visual variable of length helps in accurately identifying the age groups that have the most/least number of mass shootings. Furthermore, this chart can also indicate how the number of cases vary as the age increases. The visualization indicates that the least number of mental-illness-related mass shootings were committed by individuals in their late 40s to late 50s. The average number of mental-illness-related mass shootings was from the ages of 20 to the mid-40s. The most mental illness-related mass shootings were carried out by individuals ages 12 to 17 (Teenagers/Young Adults), with 15 mental illness-related cases in the past three years. From this visualization, one can conclude that younger individuals with mental illness may be more likely to act in violence or extreme matters than older individuals. A possible explanation for this can be that teenagers are still in the process of development and may lack the judgement skills necessary to understand the consequences of their actions.

**4. What is the trend of mental health-related shootings in comparison to other shootings over time?**



The visualization shows a line chart that showcases two lines representing the comparison of mental health related shootings to the total number of shootings overtime documented in the Stanford MSA Database. The blue line represents the trend of shootings that has happened overtime, while the orange line represents mental-health related shootings. Both lines are showcased in one visualization side-by-side to identify the varying trends of mental-health related shootings against the total number of shootings regardless of mental-health background. A comparative line graph was selected to showcase this comparison because we are comparing two sets of data on the same axis. The x-axis is the year of date, and the y-axis is the number of shooting cases. The visual variable being utilized here is color - with the use of different colors, the trend becomes readable, organized, and more accessible. From both lines, we can see how both trends upwards. Therefore, we can deduce that regardless of mental-health background, the trend of shootings has gone upward and increased overtime. Additionally, the area below the orange line is significantly smaller than the area below the blue line - this tells us that mental-health related shootings make up less than half of the total number of shootings that are documented in the Stanford MSA Database.

**5. Are mass shootings in a school setting more likely to have a shooter with a history of mental illness than mass shootings in other settings?**

****

This visualization compares school-related mass shooting incidents involving a shooter *with a history of mental illness* and school-related mass shooting incidents involving a shooter *without a history of mental illness.* A stacked bar graph was selected to demonstrate the findings/values from the data set since this question necessitates a part-to-whole comparison of numeric values over a single categorical property. There are three attributes: the first attribute, school-related incidents, is represented on the X-axis. The second attribute, the total number of victims from mass shootings in general, is represented on the Y-axis. The third attribute, the shooter’s mental illness status, is represented by two different colors: orange for yes and blue for no. The use of contrasting colors is intentional since the Stanford MSA database generalizes all mental illnesses into a blanket statement. The contrasting colors, therefore, helps to differentiate between which values signify the involvement of a shooter with mental illness (orange) and which values signify the involvement of a shooter without mental illness (blue). From this visualization, we can see that the number of victims in a school shooting where the shooter has a history of mental illness sits at 464, and the number of victims in a school shooting where the shooter does not have a history of mental illness sits at 137. On the other hand, the number of victims in a non-school-related shooting where the shooter has a history of mental illness sits at 662, and the number of victims in a non-school-related shooting where the shooter does not have a history of mental illness sits at 566. We can conclude that mental illness is not a causal attribute to mass shootings in general, however, the likelihood of a school shooter having a history of mental illness is greater than in any other setting.

**Section F: Contributions**

**Part I: Topic Approval**

* Sarah Vinh Han Chung, 30092553
* Kusumpreet Kaur Heer, 30114618

**Part II: Written Report**

**Section A**

* Phuong Thao Nguyen, 30118157
* Sarah Vinh Han Chung, 30092553

**Section B**

* Kusumpreet Kaur Heer, 30114618

**Section C**

* Aaren Abigail Atmadja. 30102099
* Nishan Soni, 30147280

**Section D**

Data Cleaning:

* Nishan Soni, 30147280
* Kusumpreet Kaur Heer, 30114618

Descriptions:

* Phuong Thao Nguyen, 30118157
* Aaren Abigail Atmadja. 30102099

**Section E** (1 Visualization Each)

* Sarah Vinh Han Chung, 30092553
* Kusumpreet Kaur Heer, 30114618
* Phuong Thao Nguyen, 30118157
* Aaren Abigail Atmadja, 30102099
* Nishan Soni, 30147280

**Part III: Presentation Slides**

* Aaren Abigail Atmadja. 30102099
* Sarah Vinh Han Chung, 30092553

**Section G: Reflection**

**Aaren Abigail Atmadja, 30102099**

**Assignment 2 (Obtaining data):**

In Assignment 2, my biggest mistake was not double-checking after exporting my word document to PDF - it resulted in some of the questions I screenshotted from Qualtrics overlapping with each other, making some of the questions not visible in the final document. With this final project, we did not choose to obtain primary data as the basis of our data analysis. However, I found that searching for existing secondary data seems to follow a similar thought process as Assignment 2. We had to find the dataset that was compatible with our objective in this analysis. Since we chose to focus our analysis on determining the correlation of different factors, though primarily mental health backgrounds to gun violence and shooting incidents over a period of time, we had to scour the internet to find datasets that consider these attributes so we can decipher its possible connection in our evaluation. While evaluating each dataset we came across to choose the suitable dataset for our analysis, I had to keep in mind the objective of the entire analysis and align those goals to the datasets I came across. Just like Assignment 2, we had to think about what attributes we needed for our analysis and whether or not they were necessary to produce a coherent conclusion. Since the dataset is the basis of our findings, we had to be extremely cautious in choosing which dataset to use in this project. For instance, we chose the Stanford MSA database over the Mother Jones database because the values that indicate the history of mental illness in the Mother Jones database were minimal and inconsistent. I felt more confident in choosing the MSA database because it had explicit attributes that were crucial to our analysis. Being able to identify this discrepancy is a skill I have improved from the initial Assignment 2.

**Assignment 3 (Cleaning Data):**

In the initial Assignment 3, I felt that I did a great job explaining why some columns/rows/attributes needed to be cleaned. I definitely could have improved my skills in using OpenRefine from my Assignment 3 submission because I was initially very wary about using an unfamiliar cleaning tool. Although I did not directly have a hand in cleaning the dataset in this project as it was assigned to other group members, I was involved in writing why they were cleaned. I think I descriptively and persuasively conveyed why some attributes needed to be cleaned from the original dataset we found online. Additionally, I found that I was more confident in utilizing OpenRefine and Excel in this project than I was in Assignment 3. I was able to run the dataset through OpenRefine and Excel to double-check if my group members had correctly cleaned the data while I was writing the descriptions. In this project, I was also better at identifying what types of errors were present in the dataset we found and which cleaning tool is best to clean specific errors. In writing the descriptions for the screenshots of the data being cleaned, I was urged to look at the bigger picture and think of the feasibility of the dataset during data analysis processes. I was able to determine whether or not analysis is feasible if some columns/rows/attributes were left uncleaned. This thought process helped me write my line of reasoning as to why some data needed to be cleaned. I also kept referring to the attributes being cleaned and whether or not the deletion of some values would positively or negatively impact the integrity of the dataset and the main objectives of the analysis as a whole. I felt that overall, my familiarity with OpenRefine has improved from Assignment 3, and this project is a manifestation of that.

**Assignment 4 (Digital Visualisation):**

In Assignment 4, I had several points taken off because the questions I came up with were not in the format of an actual question. Otherwise, I thought my Assignment 4 submission was well visualized, and its descriptions were well-argued. I was very involved in forming the basis of our analysis with the questions in this project for Section C, so I made sure to format them into questions. I thought that my digital visualizations in Assignment 4 were solid, and I feel that I showcased that same strength in forming the comparative line graph for Visualization 4 in our Analysis and Evaluation section of the project. I paid great attention to detail in Assignment 4, making sure to label all graphs in my bar charts and points in my scatter plot and label the axis and the title of each visualization. I paid the same amount of attention to the digital visualization I created for this project as well. I was also able to use the visual variable of color that I could not utilize in Assignment 4. With the comparative line graph that I visualized for this project, I was able to plot two lines in the same axis that represented mental health-related shootings (orange line) and the total number of shootings (blue line) and differentiate it using two colors. While I cannot draw a direct cause and effect relationship between mental-health background and shooting incidents, I could draw some interesting correlations between different attributes that I feel are essential to deducing a summative conclusion of our analysis. Additionally, I was more confident in using Tableau for this project than in Assignment 4, and this familiarity was very evident in this project. Overall, I adhered to the question format for each digital visualization in this project and utilized visual variables I could not in Assignment 4.

**Kusumpreet Kaur Heer, 30114618**

**Assignment 2 (Obtaining Data):**

In the initial Assignment 2, I provided the wrong level of measurement for ordinal scale for the survey question. For question three, four, seven, the survey question was pertaining to the ordinal scale but instead of stating the level of measurement as ordinal, dimension I listed it as interval, measure. Ordinal scale is an ordinal value because the scale establishes a degree of variation between them. “ordinal” indicates “order”.

In this final project, while making the graphs for each visualisation, the level of measure of each variable was handled according to the requirement. This helped us choose proper visualisation, as the level of measurement is a contributing factor towards the type of graph most appropriate for the analysis.

**Assignment 3 (Cleaning Data):**

In the initial Assignment 3, I could have improved assignment 3 by supporting my cleaning steps with more explanation. For the first question my description for the cleaning was lacking along with the reason behind the cleaning and how my cleaning would help the analysis. For the second question I should have provided much more information about the data quality issue being cleaned. Again for the fifth question I could have given a more reasonable description to the data cleaning process I was doing.

In this project I paid attention to adding a reasonable amount of description for everything I contributed to in this project.

**Assignment 4 (Digital Visualisation):**

In the initial Assignment 4, I could have improved my assignment 4 by being more creative with my analysis questions and visualisations. I felt like the data provided to us was quite sophisticated, and I could have explored more mediums representing a much wide variety of graphs and plots.

This assignment was quite more creative when it comes to the analysis questions, a lot of attributes were used to serve our goals and purposes. The database had latitude and longitude values, which were used to create a map. This sort of visualisation was quite not possible for assignment 4, as the latitude and longitude of the places were not given/ limited.

**Phuong Thao Nguyen, 30118157**

**Assignment 2 (Obtaining data):**

In my second assignment, I did not fully understand the level of measurement for each attribute and whenever it is a measure. Because of that, I confused the concepts and made the wrong conclusions. After a thorough revision, I have mastered these theories so that I will not make the same mistake later. In this assignment, the level of measurement is not used so I haven't had a chance to show how I fixed this error. Except for the above error, in assignment 2, I had no other errors.

**Assignment 3 (Cleaning data):**

Assignment 3 is my worst assignment. I have not analyzed the data quality issues in detail and clarity. My analysis is still general and does not specifically show why those entities need to be cleaned and how that will help with data analysis. Because of that error, I was not able to provide valuable analysis and I am not convinced about cleaning data. In this assignment, I have changed a lot in the way of data analysis. My analyzes have become more detailed and specific to changes in metrics and visualization. I clearly feel that my analysis is more convincing and valuable.

**Assignment 4 (Digital Visualization):**

After improving from previous assignments, my assignment 4 has become much better than before and has the maximum result. However, one thing I still don't really like is that my questions are still simple and safe. Because of that, my visualizations are still at a basic level without any uniqueness. Choosing the safe option like that helps me ensure the maximum score, but it has not shown my creativity. In this group project, I used a more complex visualization that I had never used before as a symbol map for my question. Although my question is still pretty basic, with this particular type of visualization, my visualization was much more interesting than previous visualizations.

**Nishan Soni, 30147280**

**Assignment 2 (Obtaining data):**

From my assignment 2 submission, one thing that I could improve on is getting a better understanding on the difference between a dimension and measure as it led to me giving wrong labels to some attributes. After revising, I now have a better understanding of the difference. I think that since I was able to understand the difference, it was easier for my group and I to create analysis questions as we knew what attributes were measures or dimensions.

**Assignment 3 (Cleaning data):**

I think one thing that I needed to improve on for the data cleaning was understanding why I had to clean certain parts of the data set and if it would really help in the analysis process. Consequently, I was able to improve on this with the final project as I really had to think about what cleaning methods would make the data analysis process easier. For example, when removing unnecessary columns or rows I was able to better understand why they helped in the data analysis process and how it would ultimately make our analysis more accurate.

**Assignment 4 (Digital Visualization):**

One thing that I needed to improve on for the data visualization process was figuring out what chart I had to use and how to create them more easily on Tableau. For example, I may have incorrectly used a line chart to answer a question or I just had trouble figuring out how to use Tableau. After figuring out how to use Tableau better, I was able to make charts much quicker and after reviewing the lectures I had a better understanding of what attributes I needed to use for certain charts. Since I had a better understanding of these concepts, creating the histogram for analysis question 3 in the final project was much easier.

**Sarah Vinh Han Chung, 30092553**

**Obtaining data (Assignment 2):**

Overall, I was very satisfied with assignment 2 from its inception to its completion. Something from my own assignment that can be improved is the number of qualitative questions; not that there is any official scale of measurement that caps the amount of qualitative and quantitative questions asked in a survey, and sometimes a survey topic is best suited for one or the other, but I think it would’ve been more interesting and challenging to incorporate a few more quantitative questions.

While our group project did not entail the creation and deployment of our own survey, I found that several processes for how we obtained data were quite similar to what I’ve learned during the unit. I realized that there must be a lot of time and effort put into the question-making process before actualizing what your research topic is; the question-making process was not necessarily hard, since we all had so many ideas when brainstorming, but it was difficult to filter through every idea for a relevant question with attributes that would help support our topic. I think that I did well on my solo assignment because it was centered around a topic that I’m passionate about and know from the top of my head, so coming up with questions and whether those questions have sustainable and relevant attributes for data analysis in a group setting was a learning curve for me to prioritize forming specific questions with relevant attributes to the research topic over coming up with random filler questions. Additionally, I realized that there is no room to worry about the “righteousness” (offensiveness, political correctness) of certain attributes, and it’s better to simply focus on what the data will show us post-analysis.

**Cleaning data (Assignment 3):**

Overall, I have always (and still do) find data cleaning the most intimidating aspect of the course. In fact, I didn’t complete this assignment and was unable to hand it in because of how I was just, so so so extremely intimidated by the material and grew frustrated at how I didn’t instantly understand anything. If anything, the process of trying to complete assignment 3 and the outcome shows how I don’t even know what I don’t know. In short, I would’ve greatly benefitted from asking for help when I needed it.

Moving on from my personal woes, and being too honest for any evaluation I’ve done in this term, I did not necessarily put in the heavy work for cleaning either the Stanford MSA or Mother Jones data for our group project. Despite this, it was still a great learning experience seeing how my group members were able to clean the data set plus the breakdown of their data cleaning processes. The collaborative aspect of a group project allows me to witness and learn from the experiences of those around me. The way we organized tasks for each section allocated each member to doing a specific section, which resulted in me not having to clean any data set whatsoever. However, I’m glad to see and understand the data cleaning processes from the members who put in that work since it made me realize the importance of data cleaning, which is that if nobody was cleaning the data sets, the rest of the project (visualizations and analysis) would not be possible.

**Digital visualization (Assignment 4):**

Overall, the most outstanding improvement that could’ve elevated my digital visualizations is how I go about choosing the proper visualization to represent my data and findings to their greatest potential. My default choices are always either a histogram, pie chart, or bar chart, and I need to learn how to be comfortable with using different visualization styles. Doing so would not only add variation in data representation but would have also reduced the redundancy and represented some data better. I also did not add a legend for one of my visualizations, which can confuse viewers and/or make them do mental gymnastics trying to figure out what is even being represented in the visualization.

For our group project, I did visualization 5 which asked for a visual demonstration of any possible correlation with a shooter with mental illness in a school setting. I remembered to include the legend this time as it would’ve been extremely confusing if there was not one included given the simple yes or no nature of the Stanford MSA data set and the outcomes for this specific question. Although I wish I could’ve been more adventurous with my visualization choices, I’m glad I went with a stacked bar graph since I think it’s the best way to convey this kind of part-of-a-whole data in a way that’s visually pleasing and comprehensible.

**Reference List**

Number of Deaths in 2021 Interactive Map*. Gun Violence Archive.* (2021). Retrieved

November 30, 2021 from <https://www.gunviolencearchive.org/>.

Silverstein, J. (2020, January 2). “There were more mass shootings than days in 2019”*. CBS*

*News*. Retrieved November 30, 2021 from

<https://www.cbsnews.com/news/mass-shootings-2019-more-than-days-365/>.

Parekh, R, M.D, M.P.H. (2018, August). “What is mental illness?”. *American Pyschiatric*

*Association*. Retrieved November 30, 2021 from <https://www.psychiatry.org/patients-families/what-is-mental-illness>.

Masters, J. (2021, April 12). “U.S gun policy: global comparisons”. *Council on Foreign*

*Relations*. Retrieved from

<https://www.cfr.org/backgrounder/us-gun-policy-global-comparisons>.

Bernet, R., Maples, S., & dmede. (2018, March 26). “Mass shootings in America”. *Stanford*

*Geospatial Center/MSA.* Retrieved from GitHub, October 29, 2021, <https://github.com/StanfordGeospatialCenter/MSA>.