Visualization Project

Ignacio Regaña, Antón Aba, Akos Tánczos

Link to the published app https://my1staccount.shinyapps.io/myapp/

Problem characterization

The U.S. has a well known public problem with shootings at educational centres related to the gun control controversy. As someone who has lived there for a year, I know that one of the most common drills is the preparation for these events, where students lock themselves out inside the classroom avoiding being visible for possible shooters inside and outside of the building.

As much preparation as there is and the periodic reports from the media when one of these tragedies occur, there is a lack of general knowledge about the problem. How many shootings has there been? How many fatalities do we have? Does it occur more frequently in certain places?

The list final list of questions that we have come up with is the following trying to cover the following all the questions that may arise about the location and the numbers:

1. How is the distribution of shootings, fatalities and wounded spread across the U.S.?
2. How have the mentioned variables evolved over time? And their relations?
3. How can the school type and the location affect each variable?

In order to do so we will also need to gather the information required from one or more public and trustworthy datasets.

Data and Task abstractions

Data

The dataset https://www.kaggle.com/ecodan/us-school-shootings-dataset which is the result of the combined data from both Wikipedia and Pah table datasets. For every recorded case of a shooting at a U.S. educational centre the file contains the date, city and state of the incident, its area and school type; plus the number of fatalities and wounded people.

This information can be used in order to answer most parts of the questions written above. Additionally it states its source and whether it is a duplication of the other dataset. We will use this information to remove the entries which were duplicates.

The missing requirements to properly answer the first question would be the coordinates of the incidents. These can be found at the next link: https://github.com/akshay7424/Data1050-Startup-Data-Final-Project/tree/main/simplemaps_uscities_basicv1.72, which provided us with a geometry dataset containing, among other variables the name, state, longitude and latitude of almost all us settlements.

We joined to the shootings dataset the coordinates of each shooting using the name of the city and the state. The values of the few missing values were added manually, in order to have a final version in which each entry can be used for any idiom. The only exception were the coordinates of the non-continental U.S. territories, since there are considerably less cases.

The final result is a dataset that combines both table and geometry aspects, in which each shooting is an item composed by the following data types:

- **Date**: date of incident, ordinal attribute, it ranges from 1990-2020.
- **City**: location of incident, categorical attribute which has a lower hierarchy than state.
- **State**: location of incident, categorical attribute the upper class of City.
- **Area type**: urban or suburban, categorical attribute .
- **School**: C = college, HS = high school, MS = middle school, ES = elementary school, categorical attribute.
- **Fatalities**: number of people killed, quantitative numerical attribute.
- **Wounded**: number of people wounded, quantitative numerical attribute.
- **Latitude**: latitude of incident, position.
- **Longitude**: longitude of incident, position.

Tasks

It is clear that we want to **consume** the information gathered in order to visualize, understand and communicate what others have compiled. The most important aspect is that we want to **present** information to third parties, the public or any interested organization on the matter.

This is reflected in the Problem characterization segment where before stating our questions we clearly express that our interest is to **educate** people on the severity and the real numbers of a well known problem. Precisely, the 1st question is simply about plotting the information without any transformations applied to it.

Furthermore, we want to explore our dataset in order to **discover** new knowledge, not really to validate hypotheses regarding how different factors affect. Since we have not read any previous studies on the matter, we do not know which aspects remain uncertain in the academic world.

Moreover, we want to **generate** new hypotheses, all of the questions show how each target is connected to different explanatory variables. Using the geographical location, the time, the area type and the school type we may find interesting relations or tendencies in our dataset.

Idiom 1: U.S. map

It aims to answer Question #1: How is the distribution of shootings, fatalities and number of wounded across the U.S. from 1990-2020? The data to be used here is out of every entry its city, state, coordinates and number of fatalities and wounded, the number of shootings is clearly one per entry.

Interaction & Visual encoding

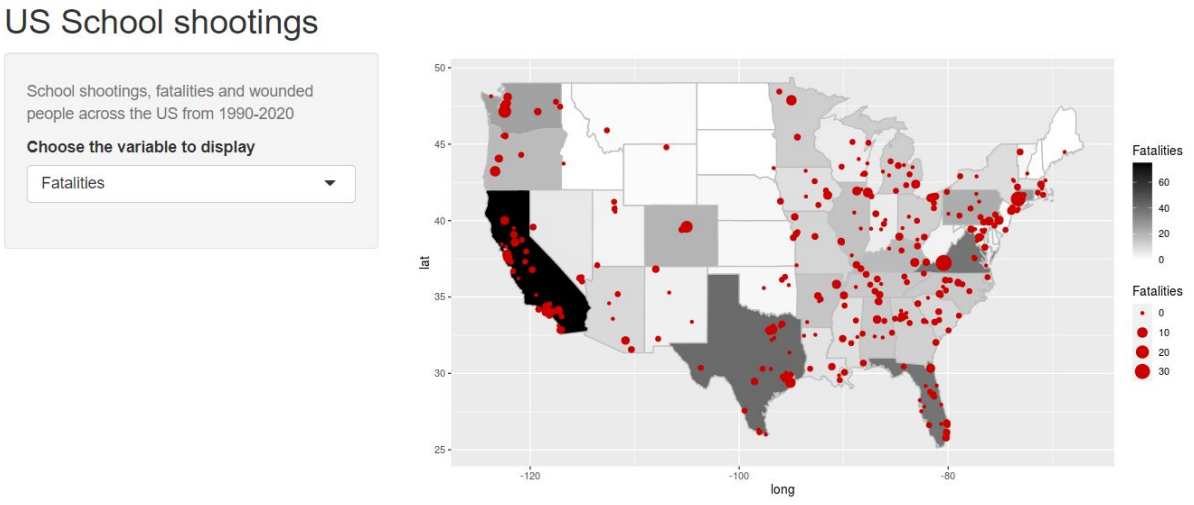
It allows the user to choose between the different variables to display: the number of shootings, fatalities or wounded people each one with different colors. This was done because plotting the 3 variables on the map simultaneously would not be feasible or intuitive.

The map shows the cumulative value of each variable on each state allowing for a discrete visualization using a scale of colors to represent the value. On top of that, for every population nucleus it groups all of its incidents and uses the coordinates to display as a dot with the size of the resulting value of the target variable.

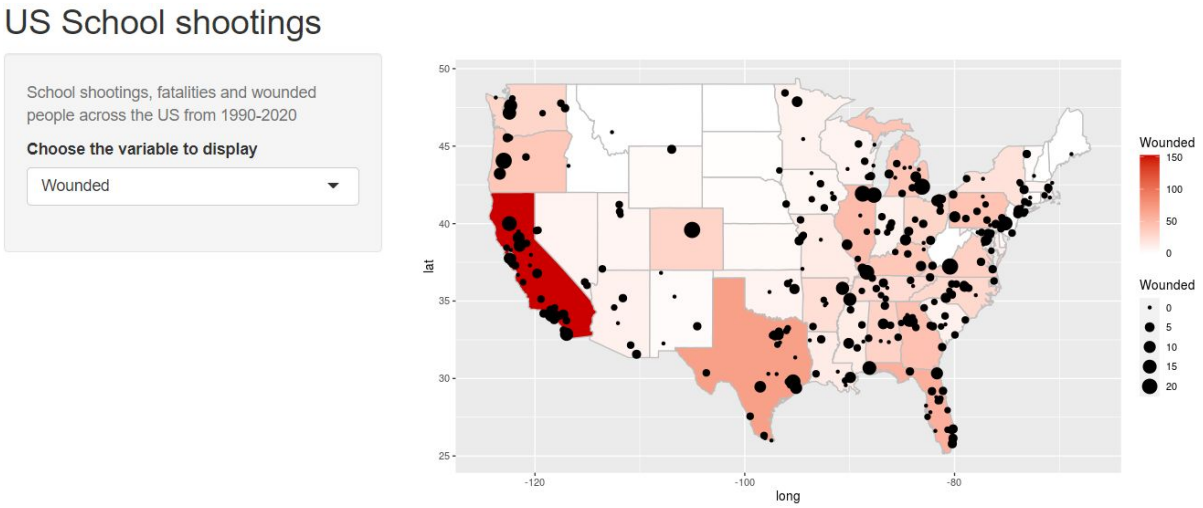
So here we have 2 **marks**: the points, with dimension 0 and the map of the US by states, areas, with dimension 2. Both have an identity **channel** of communication, the color Hue for distinguishing between variables. To represent the value's magnitude, the points have the size, and the map has the saturation for each state.

This allows the user to quickly and accurately visualize the overall value of each region (discriminately) and the actual spread of all the incidents across the map and its associated value, all in one glance. Since every variable has its own assigned set of colors for the state and the dots, it allows for the user to distinguish between them more clearly.

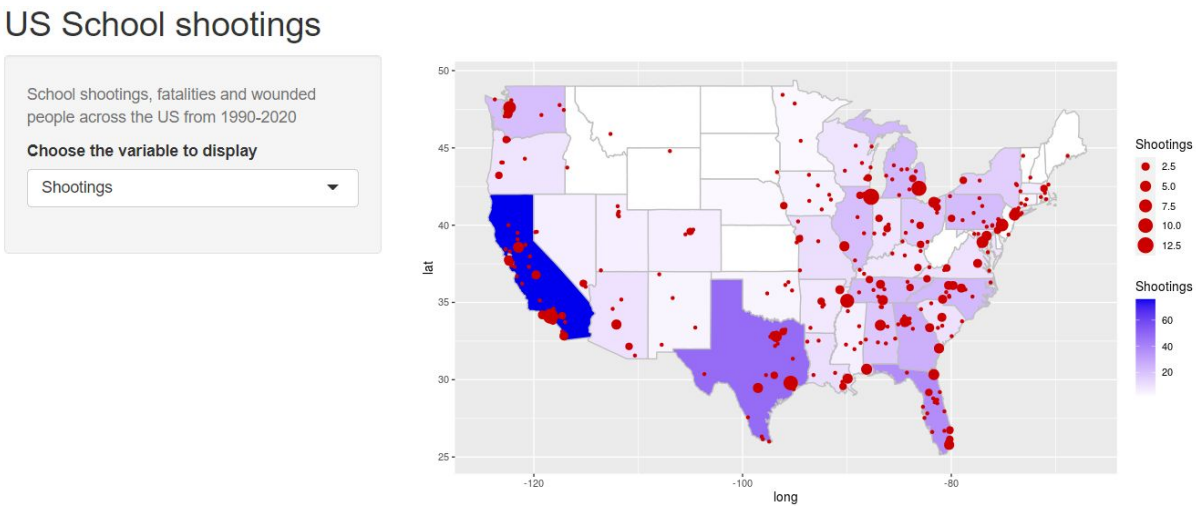
When plotting the number of fatalities, the idiom uses a black shade for the states and dark red points, the colors of mourning and blood. This reinforces the impression of the lives lost due to this national problem, as the red dots resemble blood drops, which gets the mind of the user to think about the missing lives.



When displaying the number of wounded, we use the opposite, a dark red shade for the states and black points for the number of injured at each location. This reminds the viewer of the blood spread on the U.S. land and that there places (black dots) where it is not safe, giving the impression of the danger of this problem



Finally, for plotting the number of shootings we use a dark blue shade for the states and red points. Combined they look toxic, like an infectious illness, a place where it is better not to go. This represents the national health and societal problem that the U.S. has.



Algorithmic implementation

The loaded dataset gets first summarized by the city where entries from the same place get reduced to one which consist of the coordinates, the number of shootings (number of entries), the sum of the fatalities and of wounded.

A similar procedure is done but grouping by state and omitting the coordinates. Then the program uses the map_data function from the library ggplot2 to load a map of the states of the U.S. It then merges both data frames by the state name, saving all the info into one.

This allows us to keep all the possible 3 outputs in memory in order to optimize the use of resources. The minimum allocation of this data and the initial computation that the program does is insignificant compared to the savings on computation.

Now, every time the user selects a new input it does not need to recompute all of these operations every time as the result is already in memory at a very low cost. Instead, the render plot which produces the map requests that input and based on it displays the one corresponding to the selected variable.

On the user interface we created a SidebarLayout for this idiom, which includes a SidebarPanel where the user can choose the desired plot and the mainPanel where the chart is visualized. Both of them can be seen in the attached images.

Idiom 2: Linear evolution

We try to answer the question how the shootings evolved over time. This is done with a line chart including more lines. The plot shows how the number of shootings, and how many people died and were wounded as a function of time. To do so, we use the year, and the number of fatalities and wounded from every entry.

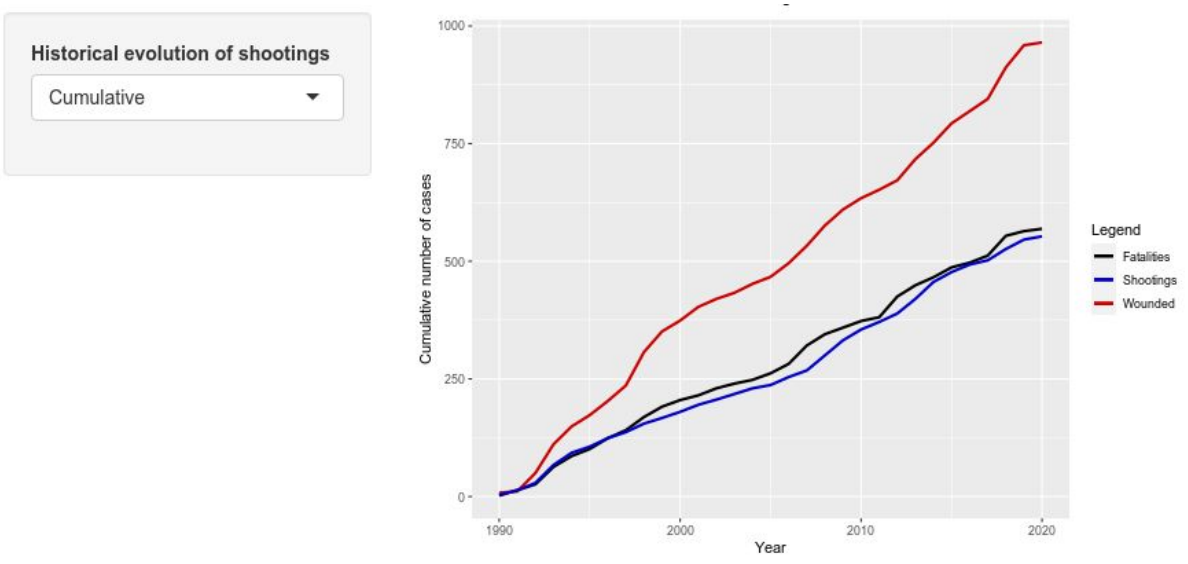
Interaction & Visual encoding

The user can use the interaction to set his preference whether he wants to have the cumulative data (at year X, all the values of the variables from the shootings that happened before X will be summed), or just the actual cases of every year.

From a visualization point of view, it contains lots of information: one should use the cumulative chart if he/she wants to know how many shootings have happened until a specific year. On the other hand, the non-cumulative chart is better if somebody wants to make inferences about whether the number of fatalities/wounded/shootings has any trend or not. The number of cases per year may be cyclic, increasing, stagnating or decreasing).

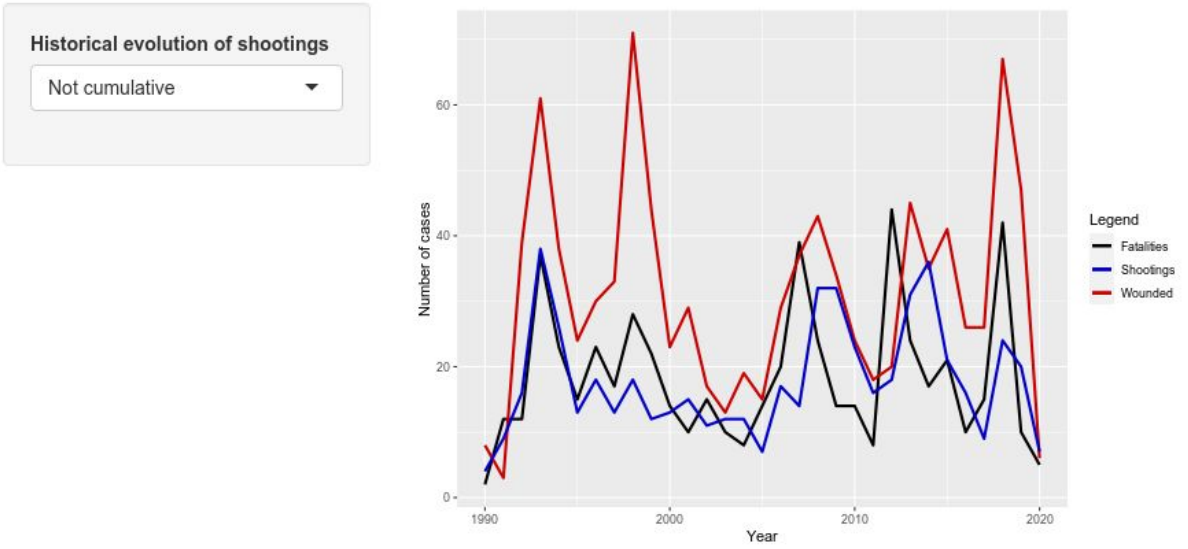
This data is on the cumulative chart too, for instance, if the curve is more or less a linear function it means there is no significant trend. If the curve looks to be speeding up or slowing down we have detected a trend, which may not be captured in the previous option. Furthermore it is better for visualizing the overall trend of the whole range of years.

In this idiom, we use one type of **mark**, the lines with dimension 1. About **channels**, there are two spatial positions (one of them for time while the other for the number of each case) and Hue color to make a difference between Fatalities/Wounded/Shootings. Since position is the most effective visualization channel, it is recommended to use it whenever it's possible.



Using the cumulative chart also has the benefit to provide the user with the feeling about how many casualties one shooting involves on average (1 fatality and nearly 2 injury per one shooting).

We used the same colors as in the previous idiom to remain consistent. But since in that one, red and black had different meanings based on user interaction, here we needed to decide about which one to stick to. In this idiom, black means the fatalities representing mourning, red reflects the danger the blood and blue creates a great contrast with the others.



Algorithmic implementation

For data preparation, we keep the Date, School, Fatalities and Wounded variables, then we extract the year from the date. This is done by separating the date into three different variables, and keeping the year from them. Now, we are able to group by the year variable and summarize the values.

Also, we need to summarize the data for the cumulative chart too, this can be done with the use of the cumsum*()* function. To avoid unnecessary computations, we compute the values of the two charts right after the start of the program. This way, in the reactive part the program doesn't need to run the computations, only choose which one to display between the two dataframe based on the argument given by the user input.

On the user interface we create a SidebarLayout for this idiom, which includes a SidebarPanel where the user can choose the desired plot and the mainPanel where visualize the chart. The ggplot with the geom_lines is put in an attribute of the output, which is reactive in order to the user being able to interact.

Idiom 3: Chart

In this section we have developed the answer to how school type and location can affect each variable. This is addressed by the use of two different types of charts, pie and barplot, depending on the group to analyze. It is necessary to specify the time interval, which can be selected from 1990 to 2020, in order to obtain the desired results about school shootings in the US. Every chart shows different data, we decided to make this distinction because, depending on the type of representation, the information can be more or less intuitive.

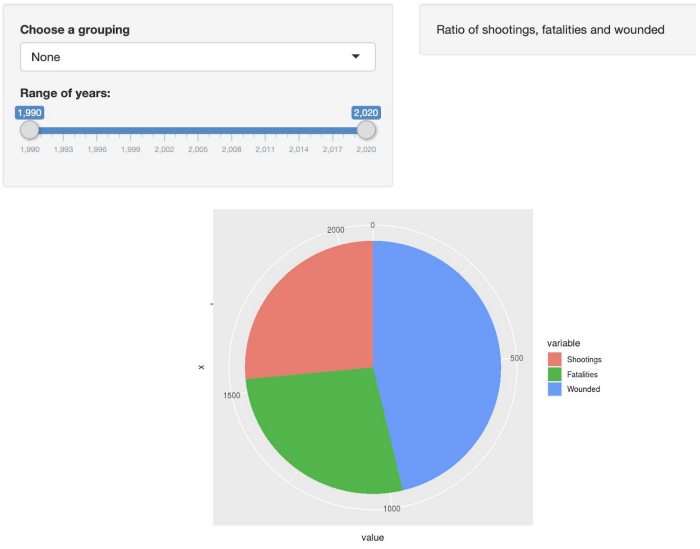
Interaction & Visual encoding

The user can interact with two features to obtain its results. The first one allows the user to choose the group of interest, which could be none, filtered by school type or area type. If the none tag is selected we would be able to see the pie chart, the reason why it's plotted instead of a barchart is because each sector represents a proportion of the whole data analyzed, which makes possible to see at a glance the impact of these serious incidents that happen every year in the United States.

If the other groups are chosen we would be able to see a barchart with the corresponding subcategories to the specified class. We have decided to choose the barchart because it is an intuitive option to display the relation between a continuous and a categorical variable. There would be three columns per each subcategory that represent shootings, fatalities and wounded people. The vertical axis in these charts show us the total number of cases, and the x axis shows us the subcategories of every group of columns.

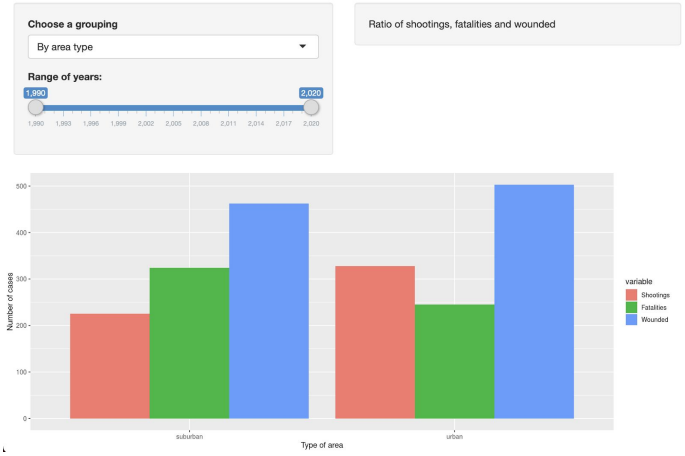
We can make use of the time range functionality, which allows us to specify the segment that we could consider more useful for our analysis without worrying if the interval is too wide or narrow.

On one hand, the pie chart we don't use any kind of **mark**, each sector composes the whole circle. On the other hand the bar charts use columns, which have dimension 1. Respecting the **channel** we count with the color hue again to identify the variable of each column or each sector.

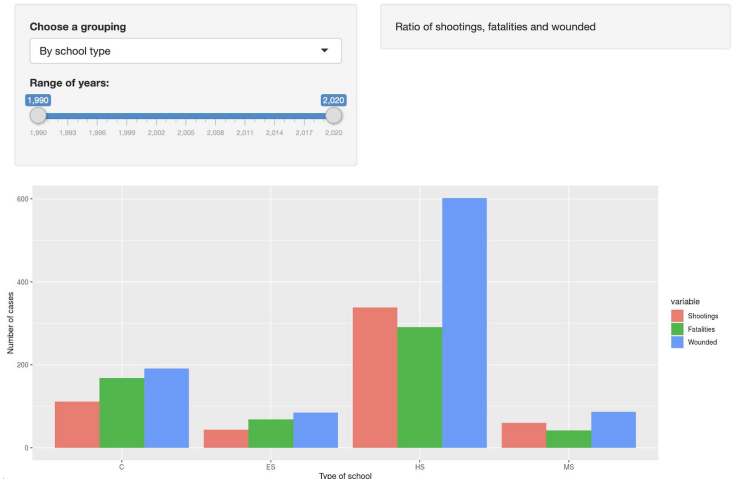


As we can see in this idiom, some colours do not represent any attribute of the data. The red is the only one that we can take into the account as we have already explained in the other idioms.

However, we can make a small analysis of the information shown in the graphs. In general terms the amount school shootings haven't changed a lot in thirty years, also the proportion of fatalities, wounded people and shootings is barely the same. This could mean that the US institutions are not putting enough resources in order to solve this issue.



By the area type barchart, we look into the size of the columns between urban and suburban along different periods of time. In general, the amount of cases that used to happen in the urban areas seems to be moving to suburban areas. This could mean that institutions are investing more in the security of the school cities rather than in schools of towns or small cities.



Finally, for the barchart grouped by type of school, we have seen that the biggest amount of entries are gathered in high schools and the smallest in middle schools and elementary schools in general terms. This could be related with the fact that people in high schools start to have enough age in order to have legal access to a gun license and some kinds of fire weapons.

Algorithmic implementation

We have chosen the ggplot2 package as is one of the best options to build grouped or stacked barcharts. We input into the data frame the categorical variables which were passed to the x and filled with arguments of the aes() function. Through the position argument we made possible to establish the columns by groups.

For the pie chart there is no specific geom to build pie charts with ggplot2. So we decided to build it with a barplot and use coord_polar to make it circular.

In order to make the time interval feature it was necessary to develop a SliderInput function by range, which we thought could be more useful than showing just a plot per year. For the menu of the charts we developed another SliderInput, indicating the possible options to choose.

This procedure works thanks to the school_types() function, which depends on the type of area chosen in order to perform the computation, it keeps the date, the fatalities, the wounded variable and the type of school or the type of area. This function is called everytime we change the kind of plot.

We decided to do it this way because it is not efficient to compute and keep all the possible outputs in memory since that would be the possible 3 variables and every possible range of years. That would be 3 * 496, the cumulative sum of 31 down to 1, this would result in a total of 1488 different dataframes.

Conclusions

Accomplishments

All the questions which emerged during the problem characterization part have been adequately addressed. Furthermore, the tasks have been successfully implemented and properly optimized. The visuals obtained provide a powerful message to the general public and could even be used to elaborate new hypotheses.

Moreover, the level of interaction that the user has has been determined to be more than sufficient, providing 1 level of interaction for each of the first 2 plots and 2 levels for the last one. Finally, the app has been published and can be found online which makes us feel that our work could be used by anyone interested in the matter and actually reach our target.

Future improvements

For a next step in the project, the first idiom could include the possibility to make it an animated gif where each image shows each year in the dataset. A range for selecting the year interval could be added, which would allow for more interaction from the user side.

The second idiom can be further improved to provide more interaction. This could be accomplished by allowing the user to detect seasonalities, beside the overall trends. In order to do so, we could add the possibility of making different time groupings.

The years can be divided by weekdays, weeks and months; on top of the year division which may be combined if desired. For instance, the overall distribution of the variables on each weekday every year, each week, month, etc.