Assignment 2 Report

Latera Tesfaye Olana

21 February, 2023

Plot-1: Fatal encounter with police in US from 2013 and 2022: A time series view

Background:

The issue of fatal police violence being a significant public health concern in the United States is pressing. There is a growing amount of evidence that indicates that individuals from specific racial and ethnic groups are disproportionately affected by deaths caused by police officers, indicating a problem with systemic racism within policing. Measuring the trend of police brutality is crucial in understanding and addressing the issue of excessive force used by law enforcement. It allows for the identification of patterns and areas where improvements need to be made. The first plot provides the trends in fatal police encounter from 2013 to 2022. The dataset is collected by *Campaign Zero*. They used, official data sources from local, state government agencies and publicly accessible media sources. The steps in their data collection was: record potential news media mentions of police violence, researchers review articles to determine if the events reported qualify as incidents of police violence, researchers create a draft entry of the incident and review, researchers perform periodic validation of incident information, and researchers perform periodic review to fill in potentially missing data.

The Visualization

My objective in creating this time trend plot is:

1. To show how how the trends in police brutality has been changing over the years and how community of color has been affected by it.
2. To deliver the most accurate visual representation of police brutality data, without any distortion.
3. To create visually appealing and informative representations.
4. To provide clear explanations of the data and its significance to help prevent misinterpretation.

By appropriately selecting design elements, such as color and typography, simplifying the visualization by removing unnecessary data and highlighting only the important trends, and using as many annotations as possible, the objective can be easily attained. In addition, these methods will help us to circumvent the inevitable cognitive issues that humans face. Proper data transformation and modeling also play a key role, but I will not delve into this step in this report.

In order to make peculiar and important patterns noticeable, it is important to make them unique in our visual presentation. Accordingly, in this trend plot, I have annotated points where considerably high fatal shootings occurred. Another important factor is scaling. Although not presented here, the first order modeling suggests an increasing cumulative fatal police encounter over the years. Not scaling our axes and starting from a y-value of zero will hide this fact. Once the starting y-value is scaled, one concern would be what seems like a rapid increase or decrease in the trend for relatively small deaths (compared to the cumulative incidence in one year). Here we are talking about what should have been a rare outcome, and it is essential to echo each number through our visual representation.

Another critical aspect is annotation and labeling. In this visual presentation, the trend line is well-labeled, making it easier to obtain information from the graph. Annotation is also helpful in adding emphasis to the parts of the trend that we want our readers to give more attention. Sometimes differences take a conscious effort to distinguish, and nudging readers towards a certain realization is essential. In addition, this helps the reader to easily remember the years where unexpected increases happened and the magnitude of these increases. With proper shading and annotation, I have also added that the data for 2022 is incomplete. This helps readers to preattentively identify the rapid change in the slope for this particular year and quickly look for an explanation within the graph. I included 2022 to focus the reader on the number of deaths in 2022, rather than the trend or line between 2021 and 2022. One key aspect of labeling that I took advantage of is to include a brief description of my key findings. This facilitates the process of gaining knowledge and insights for the readers.

Adding more information without complicating the visual representation is one of the things I thought through in this assignment. Finding a trade-off between simplicity and the amount of information that needs to be conveyed is crucial in designing our visual tools. For this particular visual graph, I wanted to focus on simplicity. However, in the physical world, finding a boundary between these two requires a deeper assessment of our objective, our audience, and many more. One of my objectives in designing this graph was to show the portion or amount of deaths attributed to each racial group. Rather than making small multiples (which can sometimes be distracting and difficult to find a story that links our multiple graphs), I decided to overlay a horizontal point-plot on the existing trend line plot. Unlike pie charts (which were my first choice), a point plot provides a precise reading of magnitude (with the help of a few labeling and annotations). Coming back to the idea of using pie-charts, each slice of a pie chart can be annotated as well, but if one cannot guess the difference between two pie slices without looking at the annotation or labeling, then it just becomes reading, and what good is using visualization?

I used *ibm* from families of **IBM Plex Sans Condensed**, which is part of Google fonts (if you are planning to run this code, make sure you are connected to the internet). Apart from being a personal choice, I have to say, this font is really smooth and easy on the eye.

In this work, I have utilized ColorBrewer and mostly manually-defined color palettes. Colors are used in the first trend line visualization to enhance the graph's appearance and draw attention to important messages. A light green color (node #2FC1D3) is used at each year to guide readers through the trend and aligns well the line plot and background color. Although green is challenging for color-blind individuals to discern, since there are no meaningful qualitative interpretations related to this color choice, the line trend and annotated information provide sufficient context. For the line plot, I employed light blue (#076FA1), which pairs well with a white background. Overall, the graph's elements were positioned to prevent overlapping colors and shapes.

Regarding aesthetics, the graph features a clean theme with minimal grid lines and higher semi-transparent (alpha = 0.1) grid lines. The horizontal point plot graphical element's color, position, and labeling were designed to minimize overlaps. Sensible labels, such as legible legend titles, axes titles, and annotations, are provided.

In conclusion, the selection of colors, shapes, and graphical elements aims to maximize the information extracted from the graph by emphasizing simplicity and minimizing unnecessary distractions from the comparisons at hand.

Key findings of this work are already indicated in the graph it self.

Plot-2: Studying variables impacting Forced Expiratory Volume (FEV) through visualization

Background

A cohort study of adults aged 65 years and older was conducted to observe the incidence of cardiovascular disease (especially heart attacks and congestive heart failure) and cerebrovascular disease (especially strokes) in the elderly over an 11 year period, and to relate the incidence of those diseases to various risk factors measured in the population on a regular basis. This is of particular importance, because there is increasing evidence that some of the associations observed between cardiovascular or cerebrovascular disease and various risk factors in middle aged adults are not observed in older adults. In this study, elderly, generally healthy, adults were randomly selected from Medicare rolls. Agreement to participate was high, and thus the sample can be regarded as a fairly accurate representation of healthy older Americans. At the time of study enrollment, and on annual visits over the length of the study, the participants' data regarding various behavioral (e.g., smoking, alcohol consumption), functional (e.g., ability to perform routine tasks), and clinical (e.g., blood pressure, laboratory tests, forced exhaled volume) measures are recorded. For this analysis and visualization I will focus on the measure of forced expiratory volume (FEV). A measure (in liters per second) of forced expiratory volume in the participant at the time of MRI. FEV measures the volume of air that can be forcibly exhaled within 1 second. Normal FEV measurements depend upon the size of the lungs, which in turn is usually proportional to body size. In addition, FEV is highly impacted by behavioral characteristics such as smoking.

In this work I will visually explore the magnitude and type of association between FEV and body size approximated by height (centimeters) and and how they change when considering other variables such as, smoking behavior and birth assigned gender.

The Visualization

The rationale for choosing colors, shapes, and graphical alignments to enhance aesthetics and circumvent cognitive biases, as indicated earlier, also applies here. Since both our primary outcome (FEV) and predictor of interest (height) are continuous variables, I will use a scatter plot. A first-order linear fit has also been added to provide information on how the slope and intercept change for different sets of data dis-aggregation. The linear fit models are custom-fit models (FEV ~ Height).

To show and correlate multivariate variables, using multiple or multifunctional elements such as color, size, and shape is imperative. In this work, I will make use of shapes and colors. My rationale for selecting the shapes and colors used in the plot was based on providing a quick, correct, and accurate reading of the distribution of the data and preattentively providing a view of how the association or linear trend between FEV and height changes while considering smoking behavior and birth-assigned gender. The axes are also scaled and transformed to accentuate the relationship between the variables. Additionally, scaling and transformation facilitate accurate and quick reading of the graph. Unlike my first plot, here, I will use less labeling and annotations as my primary objective is to show the association between FEV and height.

On the axes, rug plots were used. This is an important method to study the concentration (homoscedasticity) and variability (heteroscedasticity) of the data distribution. Additionally, it provides a hint about which parts group our data are mostly skewed towards (i.e., are there more short study participants compared to tall participants). This complements our expectation in which region of the linear fit we should expect higher or lower uncertainties. A transparency has been added to each rug plot to avoid distracting clustered black areas at the axes.

I have created two small multiple plots. The first one shows the general relationship between FEV and height while considering birth-assigned gender. As a minimal set of colors and shapes were used, we can simply imply what kind of relationship would exist between FEV and height without dis-aggregation by gender. Two sets of colors, one for each gender, were used (from color-brewer Set1). The colors work well with the white background, and it is easy to tell them apart (opponent color theory implemented within color-brewer). The second smaller plots were challenging. Here, I wanted to show four different data dis-aggregations (male smokers, male non-smokers, female smokers, and female non-smokers) and their fitted linear trend on the original data distribution. More small multiples could have been created here, but as indicated in the first plot, small multiples can be distracting, and they may impact comparisons across groups of interest. I made an ambitious choice of including everything in one plot. The first thing I did was to remove the confidence interval from the four line fits to avoid an overcrowded and overlapping plot. I used two sets of shapes (rectangle and triangle to distinguish smoking behavior) and four sets of colors (color-brewer, Set1, from 2 to 5) to distinguish male smokers, male non-smokers, female smokers, and female non-smokers. I admit it is kind of overkill to use all these combinations of colors and shapes. However, I wanted to draw the readers' attention towards the distribution of smoking behavior across the dataset and its impact on FEV across changing height while analyzing the interaction and the impact it has with birth-assigned gender. Moreover, assigning identical variables to different pre-attentive features can aid in highlighting differences and separating distributions, even in cases where data points are highly clustered. The two selected glyphs or shapes are the most effective means for encoding smoking behavior. Even if they overlap, it is easy to distinguish between them preattentively. In addition, they are easy to remember, which makes comparison easier. Using transparency (alpha = 0.45) also improves the visibility of each glyph, even when there are many overlapping and crowded points.

The plot size has been set to improve readability and clarity. The graphical elements are neither too small nor too large. The legends are provided within the plot itself, which ensures full use of the available area and makes reading of the plot much easier. I have removed the horizontal and vertical grid lines, as we are not interested in reading the values (FEV and height) of a single point on the graph. The axes are equally spaced.

Key messages

The following are key points I expect the reader to gain from the plot:

1. From the first small plot: the first linear trend of suggestive an increasing FEV with increasing height
2. From the first small plot: the change in slope and intercept in the two fits, when considering the gender, clearly implies that gender is an effect modifier for the relationship between FEV and height. Therefore, it is necessary to use modeling methods that incorporate an interaction term.
3. From the first small plot: male participants tend to be taller than female participants.
4. From rug plot: the variability in the relationship of FEV and height is high for short and tall study participants.
5. From the second small plot: there are more male smokers as compered to female smokers.
6. From the second small plot: the estimated mean difference between two female study groups who varies by one centimeter of height is less than the estimated mean difference between two male study groups who varies by the same height.
7. From the second small plot: regardless of gender and height smokers tend to have smaller FEV.
8. From the second small plot: looking at female data distributions, the difference in the slope of female smokers and non-smokers is very small. This implies, smoking is more associated to FEV rather height. This means smoking behavior explains more of the relationship between FEV and height making it a precision variable. This seems true for male study groups as well. However, I will take greater caution from concluding as such, since the slope for smoker and non-smoker males seems to be different (p-value will help in deciding this).

The realization of these points from the given graph depends on who my target audiences are, however, for those with some analaytical and statistical background it has to be a walk in the park. If not, then, I made a really terrible graph :)