Ronin Chasan (rcc263), Abu Qader (abq2), Jacob Wise (jmw555)
INFO3300 - Project 1 Written Description

Data Description

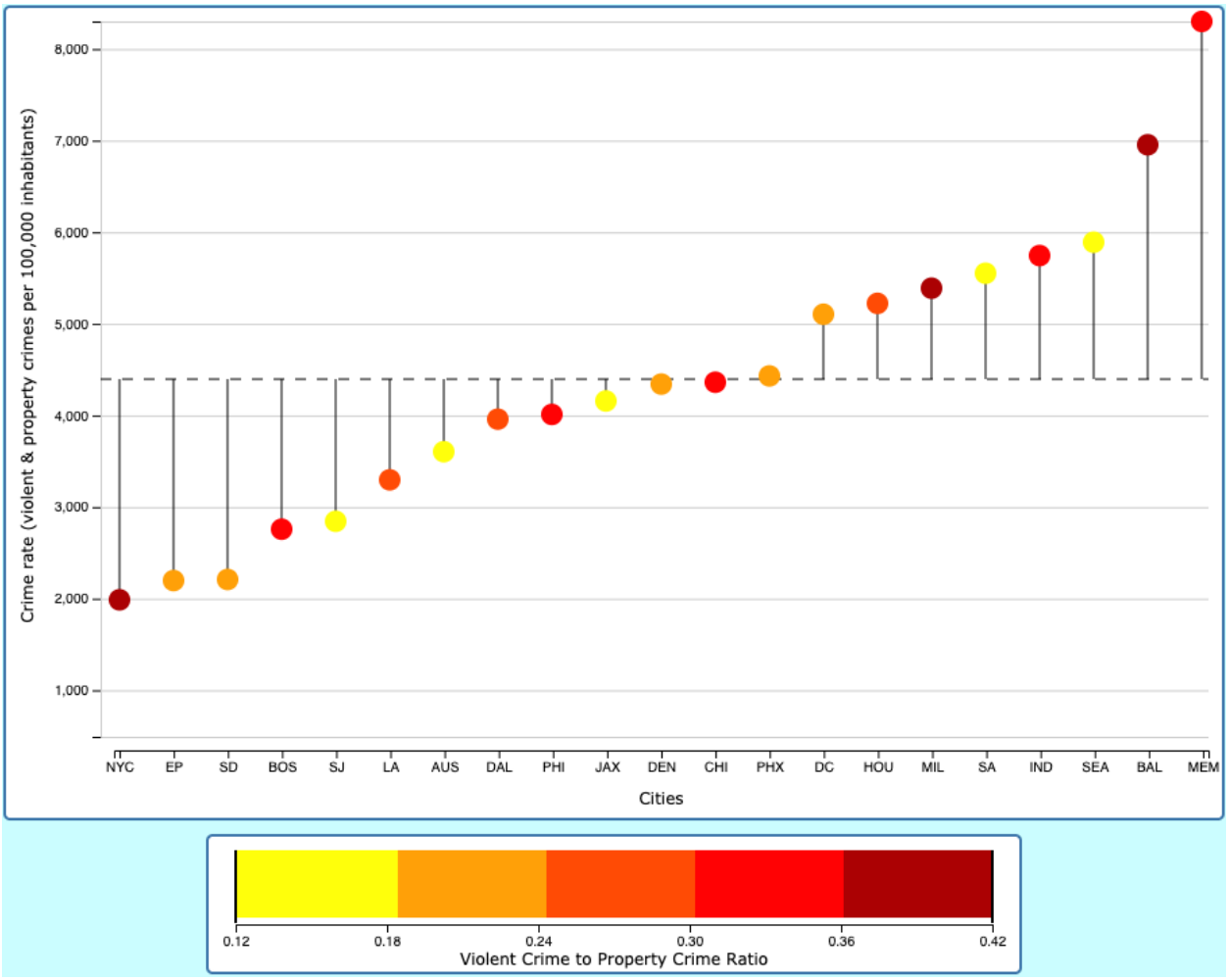
To collect our data, we had to crosslist several tables from several sources: one for crime rates, education scores, and unemployment. Luckily, we were able to find many reputable sources with data from 2010. We found crime rate and population data directly from the FBI database[1], city-by-city unemployment rates from the US Bureau of Labor Statistics[2], and municipal education scores from WalletHub[3]. However, each table had a different file format (including HTML tables and PDF files), some more complicated than others with subcategories for some rows and columns. For example, the FBI crime data was organized into separate pages by state, each with data for every city in that state, something that would've been needlessly inefficient to filter through JavaScript. Some also had a few missing values; for example, we were able to find unemployment and education data for Charlotte, but the total crime rate was missing. We therefore thought it best to go down the list of cities by population, transferring complete values over to an Excel spreadsheet until we had 21 or so cities, a number we decided was a usable sample size while also being low enough to not overwhelm users with data. We then converted this spreadsheet to a CSV file for use.

Even after combining and importing the data, we had a few filtering/cleaning tasks to take care of. We looped through each data value, converting unemployment rates from whole number percentages to decimals out of 1, and used these decimals while formatting the x-axis on the unemployment scatterplot. We also made a new column, 'Crime Ratio,' containing the ratio of violent to property crime for each city. After this editing, our dataset was ready for use in building visualizations. The resulting variables for each city were: city name, city abbreviation, total crime rate, violent crime rate, property crime rate, crime ratio, unemployment rate, education score, and population (all crime rates are per 100,000 people).

1. https://ucr.fbi.gov/crime-in-the-u.s/2010/crime-in-the-u.s.-2010/tables/10tbl08.xls/view
2. https://socds.huduser.gov/Bls_laus/emplstat.pdf
3. https://wallethub.com/edu/e/most-and-least-educated-cities/6656#cassandra-guarino

Ronin Chasan (rcc263), Abu Qader (abq2), Jacob Wise (jmw555)
INFO3300 - Project 1 Written Description

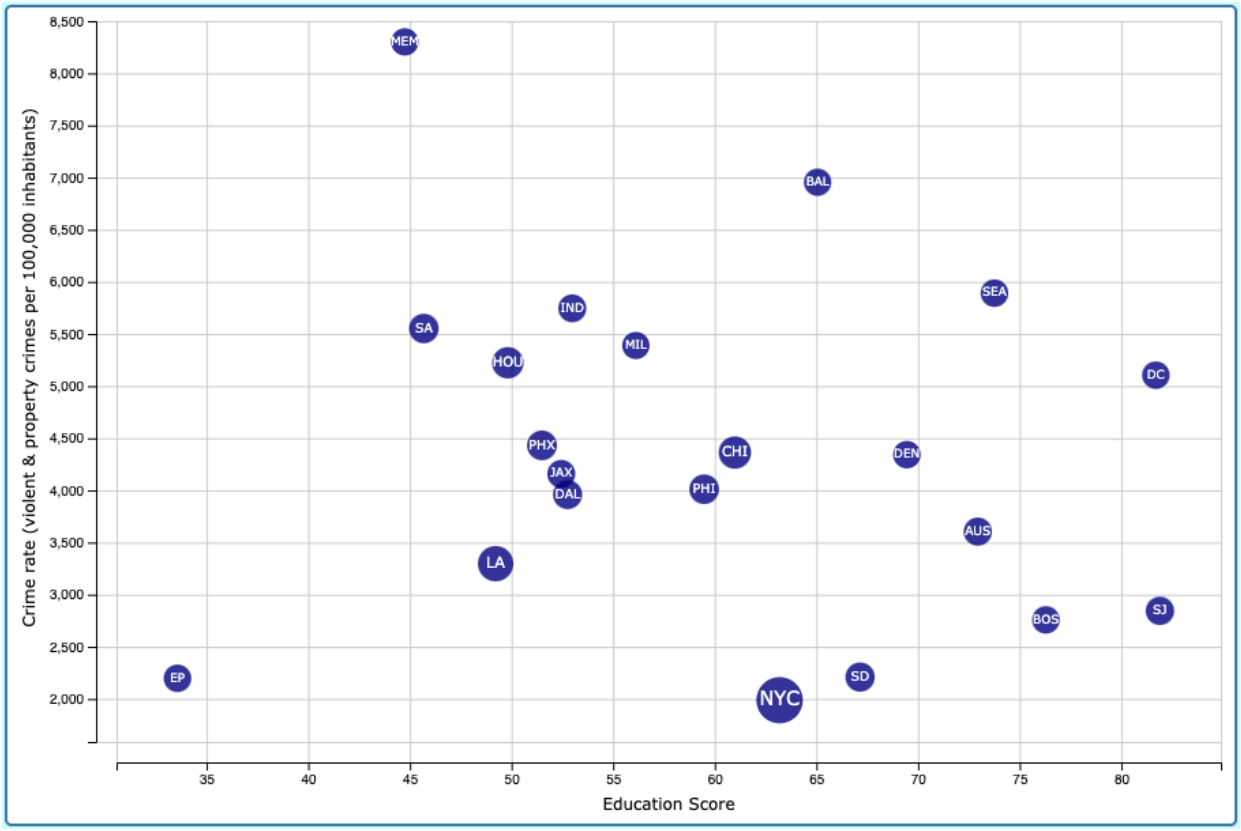
Design Rationale



Our first visualization was a lollipop chart, aimed at portraying how the crime rates of some of the most populated cities in America compared to the average. We believed a lollipop chart was the best way to visualize each city's difference from the average, while allowing us to focus on another variable by changing the color, size, or luminosity of the circles as well. Thus, we used circle color to represent the ratio of violent crime to property crime, so users can identify any trends between having an above average crime rate and having particularly violent or nonviolent crime occur. The marks are the colored "lollipops," which are in reality made up of lines and circles, sorted along the x-axis by ascending crime rate. The channels are aligned y-position of the circle (or length of the lollipop stick) and the color hue of the lollipop circles. The major design decision came when choosing circle colors. Red, a color often associated with danger and violence, seemed practical to represent the cities with the highest violent/property crime ratios. However, many color scales we attempted were unsuccessful; for example, a linear scale from white to red ends up making most of the points a medium-orange color, with a few being more red. The result made it difficult to distinguish violence levels between cities.

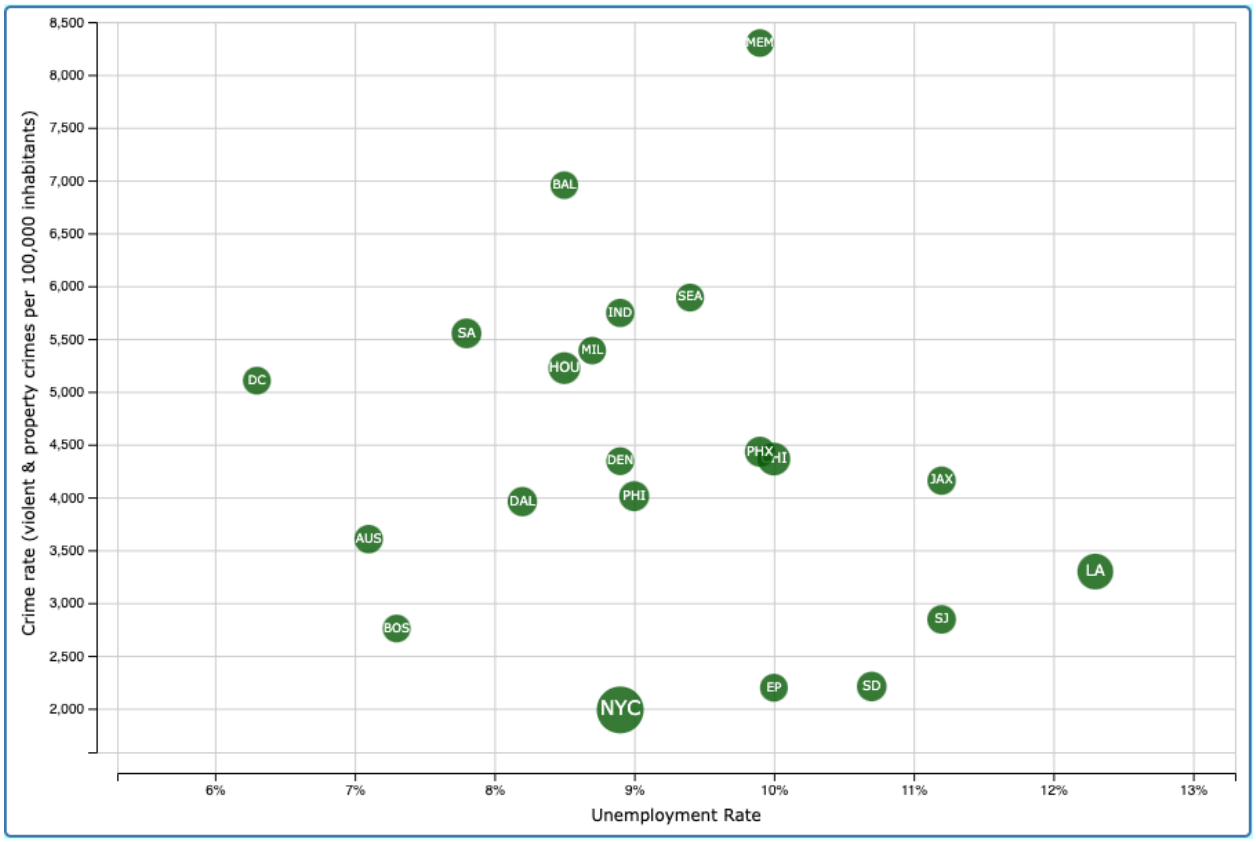
Ronin Chasan (rcc263), Abu Qader (abq2), Jacob Wise (jmw555)
INFO3300 - Project 1 Written Description

Ultimately, we decided to use a quantile scale, ranging color hues from yellow (low violence) to orange to red (high violence). Since the violence ratios are well-distributed, about the same amount of cities fall into each quantile, as can be seen by the width of each color band on our legend (credit to Professor Rz. for inspiration for the legend!). With these colors and quantiles, the scale is easy to follow, shows the distribution of violence ratios accurately, and is easy to map to real-world properties (i.e. red is more violent, yellow is less). Like the scatterplots, we used city abbreviations to label each lollipop along the x-axis - full city names were simply too long, and the tradeoff in user comprehension is small.



In our second visualization, we created a scatterplot visualizing crime rates in our selected cities as compared to the education score for that city. Seeking to identify a potential relationship between two numerical variables, we believed a scatterplot would be the best way to display this data and allow the user to draw their own conclusions about any correlation. The marks in this graph include dark blue circles, while the channels include the aligned x-axis position, aligned y-axis position, and radius/area of the circles, all of which were mapped linearly directly from the data. We elected to use dark blue circles to represent the data as circles are easily identifiable as data points and this shade of blue is a dark, neutral color in this context, with minimal cultural connotations, while also offering a clean, aesthetically pleasing background for light text. We also decided to marginally vary the size of the circles by the population of each city. This was a decision made with the intent to give the user both a feature that would make each point a little more easily identifiable as a particular city (e.g. Americans will likely know the largest one is New York City even out of the corner of their eye) as well as to allow the user to

make their own judgements about population versus either of the two mapped variables. Lastly, we decided to overlay the city abbreviations on top of the data points themselves. Although this may make the text smaller and thus more difficult to read, we believed this tradeoff was worth reducing the clutter of having text next to the points or using an ordinal color scale to identify each city through a legend. This labelling allows users to easily identify each city and relate to the data more, without taking the focus off the values and greater relationship itself.



Our last visualization is a scatterplot depicting unemployment rates across the country and crime rates. Similar to the last graph, we chose a scatterplot as the best way to represent the data. Since we want to uncover a correlation between certain variables and crime rates, a scatter plot helps us do this best. The marks on this scatterplot are green circles. One channel on this visualization is the size of the circle. Our intent here is to reflect the size of the city (population). We decided that population is a variable that is well-encoded as size because there's an intuitive relationship between the size of something (the circle here) and the size of a city. We also encode the city on the circle. We needed a way to represent what city each dot was and we found placing the text on the circle worked best. The circle represents a specific city and applying a text marker on it is intuitive. Other channels include the aligned x and y axis'. The axis is linearly scaled as the data is best represented linearly. Logarithmic scales are not as intuitive to understand and are especially complex when the distribution of data doesn't scale logarithmically.

Ronin Chasan (rcc263), Abu Qader (abq2), Jacob Wise (jmw555)
INFO3300 - Project 1 Written Description

The Story

Oftentimes we see lowering crime rate used as a sort of measure of the "success" of any given city or neighborhood in politics and in the news, however, we are so rarely actually shown real data on what types of municipal investments actually lower crime and make residents safer. We sought to visualize crime rates across the country against a couple of different variables to try and find a correlation that may help guide city investments.

With our first visualization, we tried to get a sense of the distribution of our data. Given our relatively small sample size of data, we wanted to see if there was a wide range of crime rates in our selected cities, and how much the types of crimes committed in these cities differed. As it turned out, we did end up with the broad range of crime rate levels in these cities and found that the ratios of violent to property crimes were well distributed and not really correlated with the total number of crimes. With this information confirmed, we were able to proceed to the next stages of our visualization.

The first of the two analyses we wanted to make was the comparison of crime rates to education in each of these cities. We elected to make use of the pre-calculated education score because of the wide range of factors and inequalities that it accounted for, making it the most comprehensive education data we could find. When plotting this against crime rates, we weren't particularly surprised to find out that our hypothesis had some merit to it: education scores and crime rates in cities appear to have a negative linear correlation, albeit not a particularly strong one, with the notable exception of just one outlier (El Paso, Texas). We believe that the way we visualized this on a scatterplot makes this relationship relatively easy to discern for users.

The second and final visual analysis takes a look through a similar lens as the previous one, but this time with an eye on unemployment rates in the selected cities rather than education. Somewhat to our surprise, there was essentially no correlation of any kind between the two statistics. However, we were glad to see the results of this analysis in the sense that it helped formulate our overall takeaway.

After viewing each of our visualizations, we wanted the user to come away with an idea of what politicians and figures in the media are talking about with regard to these statistics to enable them to make more educated decisions both in the voting booth and when talking with others. We hoped to make the point that there is reason to believe that investments in a city's education system may be able to help lower crime rates and create a safer environment for all inhabitants of that city.

Ronin Chasan (rcc263), Abu Qader (abq2), Jacob Wise (jmw555)
INFO3300 - Project 1 Written Description

Team Member Contributions:

- **Abu**
  - Create GitHub repository
  - Locate and format unemployment dataset for visualization use
  - General cleanup at end to make sure we met all of the project requirements (file naming, etc)
  - Implement SVG visualization for education scores vs. crime rates
  - Write description in HTML design rationale for unemployment rates vs. crime rates visualization

- **Ronin**
  - Locate and format education score dataset for visualization use
  - Implement SVG visualization for education scores vs. crime rates
  - Write description in HTML and design rationale for education scores vs. crime rates visualization
  - Write “story” section of written report
  - Collaborated on ideation of data description section

- **Jacob**
  - Get dataset for 4-year average education attainment rates
  - Locate and format crime rate dataset for visualization use
  - Implement SVG visualization for crime rate distribution and ratios
  - Write description in HTML design rationale for crime rate distribution and ratios visualization
  - Write data description section of written report
  - Collaborate together to outline team contributions to the project (see more in P1 doc)

Time Spent

Overall, we would estimate that this project took us about fifteen hours total between the three of us. We believe the most time-intensive part of this project was not the coding, but rather the formulation of ideas and researching for usable datasets. This was likely due to the extensive time we’ve spent practicing coding throughout this semester which reduced the amount of time we needed to code, however locating datasets with quality data, overlapping entries, and that would fit with our research question.