**Data Visualization Process Book**

Mingqian Wu, Felix Lyu

**Overview**

This is the CS 3891 Data Visualization final project for the group of Ming and Felix. This is a two-part visualization that builds on the data from the NYPD on the city violent crime rate to present the component of violent crimes for each year from 2000-2018 and the trend of change in each category of violent crime.

**Question**

Our initial project proposal aims to map out the correlations between a selected number of factors with the crime rate over some selected regions. However, as we found serious problems with homoscedasticity in our project design, we changed our project plan and revised the questions to be answered from our design. In the finalized version, there are two questions to be answered. The first one is what is the breakdown of violent crime rates in New York City for each year. The second one is what is the trend of each category of violent crime over the past 18 years.

**Related Work**

The official website of NYPD provides us not only the raw statistical data for our project, but also has a built-in visualization system at <https://compstat.nypdonline.org/> that maps out the crime occurrences with the geographical areas. While the visualization design is not directly related to our project, it provides some background knowledge of what to expect on the overall crime rate in NYC. It also inspired us to dig further into the time factor that contributed to the change of crime rates.

**Data**

Ideally we want to gather as more years of crime data as possible for our project design to better understand the historical trend. However, the official website of NYPD only provides open source data for violent crimes from 2000-2018. Though we could find scattered datasets on third-party websites, we decided not to use them, as the NYPD site is more reliable, and the focus of this assignment is on the visual designs. Our dataset is a two-dimensional array that contains the yearly violent crime rates by category in NYC. It contains the major 7 categories of crimes, a yearly total number of violent crimes, and 19 years of data. So for the data preprocessing, we only extracted the table into more machine-readable files or data structures that associate key-value pairs for better code access.

**Exploratory Data Analysis**

Initially, we were looking at small multiples and parallel coordinates to display our data since we needed to allow our visualization scale into a high density without clustering. Specifically, we wanted a series of scatterplots that show correlations between all the factors and the violent crime rate. For each scatterplot, the x-axis represents the specific factor, and the y-axis was the violent crime rate. To compare the relationship between factors, we planned to use a parallel coordinates graph, each mapped to the line using a linear scale, and the cities would be encoded categorically using colors. We wanted interactions to select specific cities, factors or years.

**Design Evolution**

In our proposal, we were overly ambitious regarding the scope of the project. We hope to conduct data visualization across 5 major cities across 15 years to look at hundreds if not thousands of data points using small multiples and parallel coordinates to find the relationship between crimes and income. However, it turns out the data gathering is difficult. We may not get the data of a certain city or year, and they may not merge with each other smoothly. Sometimes even when we get the data, we need to clean it up, which turns out to be too much work. As a result, we have to switch our project from the massive scope toward a much smaller scope that is the crime distribution of New York City across 15 years at a much later stage of the semester. In the end, we decided to use pie chart to look at the distribution for a single year and line graph to depict trends. I wrote the pie chart part, and Felix wrote the line part.

For the pie chart, although in class we talked about how it is a bad visualization technique, I would still like to make a case for it here. I think it being so popular allows people to intuitively view the data without much cognitive process. The circular shape not only efficiently conveys to the proportion of each crime, but also saves space when compared to a cartesian visualization. To counteract its inability to display the exact proportion or quantity, unlike that of bar chart, I added some interaction to the pie chart to allow it to show a tooltip displaying the exact proportion and count of each crime when hover over the pie chart. In addition, a drop down button is added to select the different years.

For the line chart, we used it as a media to communicate the trend of different categories of violent crime, as well as to compare the absolute values between different categories. One major problem in the design is the scale. As different categories of violent crimes can have really significant difference in occurrences, we have two lines that are obtuse in terms of change, which are Murder and Rape. Also we have some degree of cluster in the middle part of the graph, as we have a total of 7 lines, which definitely would cross with each other. Therefore, we decided to add an interaction feature that display the year’s values for each type of crime when hovering the mouse over the graph. In this way, it is easier to access the actual values when neede.

**Analysis**

Using the pie chart, it is fast to interpret the contribution of a certain crime to the overall number of crimes in a certain year. To many people’s surprise, we can see that deadly crimes like Murder and Rape only contribute a portion of the overall crimes. By projecting our data onto a line graph, we clearly see that the overall trend of the violent crimes are decreasing in NYC, while the decreasing factor varies drastically across different categories of crimes. For example, the occurrences of Murders and Rapes only fluctuates at a certain level without significant changes, whereas the occurrences of crimes like Burglary decreases dramatically over the past 18 years.

**Future Work**

For future work, we hope to allow comparison of crime rates across different cities and not just New York. For the line chart, we plan to do that via selecting certain cities to give it more opacity to highlight the lines of the cities to prevent occlusion and crowding. For Pie chart, we plan to make the pies smaller and transform it into small multiples across different years.

**Peer Assessment**

Mingqian Wu

Self-assessment: I think I did a good job in coding and coming up with the preliminary idea of our project. I helped in our code a lot. Specifically, we were struggling on how to fix the tooltip box, and I fixed it in the end by appending it to a wrapper div instead of the svg.

Peer-assessment: I am glad Felix pointed out to me that our initial proposal is beyond our reach and we should change it mid-semester. Although in the beginning, I am uncertain whether this is a good idea. On hindsight, I think it turned out well. I also think he is very hardworking in the coding and helped me clean up my code from raw html into a js file.

Felix Lyu

Self-assessment: I personally think I contributed what I’m supposed to in this team project. I came up with the actual topic we are doing our research on, and gathered the data for our visualization. I also worked hard to search examples on interactions to aid our coding, and devoted a lot of time in realizing them.

Peer-assessment: Ming is a great partner to work with. He is resourceful and highly efficient. For our project proposal, he contributed the majority of the contents while I was doing the backend jobs like the data gathering. He is also good at D3 coding, and helped me a lot with bugs in the line graph. Overall, we made a really good team.