02806 Social Data Analysis and Visualization

Final Project Assignment B

Motivation

What is your dataset?

For this project, we chose two datasets to analyze. The first dataset chosen is the Stop, Question and Frisk Dataset (http://www1.nyc.gov/site/nypd/stats/reports-analysis/stopfrisk.page). This dataset was found in the New York Police Department (NYPD) section of The Official Website of the City of New York under 'statistics'. The stop, question and frisk program is a practice established by the New York City Police Department (NYPD) that involves temporarily detaining, questioning, and searching civilians they deem a threat on the street for weapons and other contraband. The second dataset chosen is titled Demographic Statistics by Zip Code (https://data.cityofnewyork.us/City-Government/Demographic-Statistics-By-Zip-Code/kku6-nxdu). This dataset was found on NYC Open Data.

Why did you choose this/these particular dataset(s)?

The implementation of the stop, question and frisk procedure by the Police Department in New York has always been incredibly controversial. There are some that claim this procedure has made many neighborhoods in NYC much safer, however there are others that accuse the NYPD of targeting individuals of color and other minorities. This controversy is something that has always been of interest to us, so we wanted to take the analysis a step further and see how effective the practice really is in catching criminals. By comparing the stop, question and frisk dataset with the dataset of demographic Statistics by Zip Codes, we will be able to get a better look at if an individual's race plays a role in the NYPD's enforcement of stop, question and frisk.

What was your goal for the end user's experience?

The end goal is for the user to easily see whether or not race is a factor in a police officer's decision to stop, question and frisk a person in each postal code in New York City. Our visualizations allow the user to click on any district in New York City to see both a scatter plot of stop and frisk encounters classified by race and two donut graphs. One graph shows the area's breakdown of demographic of the overall population while the other shows the area's breakdown of demographic of stop and frisk encounters. These visualizations will make it easier to see if there is a large difference between who lives in New York and who is getting stopped in New York.

Basic Stats

Write about your choices in data cleaning and preprocessing

There were many of aspects of the stop and frisk data set that we did not need or needed to change. Data like whether the suspect was cuffed, their height and whether they were verbally aggressive was not needed, and those columns were removed. The data set gave us the position of the frisking using NYC's data geocoding system. This corresponds to state plane coordinates which we could not use in visualizations. In order to use these numbers we had to pass the numbers through a function we made that translated the coordinates into longitudes and latitudes which we could then project onto a map in d3. We also had to reformat some of the dataset to remove white space and drop duplicate data.

Write a short section that discusses the dataset stats, containing key points/plots from your exploratory data analysis.

Our dataset started 7.2 MB with 22561 rows. After cleaning and preprocessing we cut it to 1.6 MB with 21075 rows. We found that the breakdown of the demographic of stop and frisk was as follows: 54% Black, 29% Latino, 11% White, 9% other. The overall population of New York City was 45% White, 25% Black, 27% Latino, 3% other. There is a huge discrepancy, especially in Black residents of New York City, in the percentage of a given race in the city and the frequency that that race got stopped by officers. While only about a quarter of the population is Black, they represent over half of all people who are frisked.

Genre

We are using a partitioned poster for this visualization. There are a few reasons for this. Firstly, we do not believe that it is necessary to look at our data in a certain order. The initial map shows an overarching view of the five boroughs of New York City. It is possible to then click on a certain race and see that population throughout the five boroughs. The viewer can look at whichever one interests them first, and then look to the other to give their opinions more context. This is perfect for a partitioned poster genre, and we believe it helps tell the story most completely.

Which tools did you use from each of the 3 categories of Visual Narrative (Figure 7 in Segal and Heer). Why?

From the 3 categories of Visual Narrative, the tools we found useful were close ups, feature distinction, and zooming. All three of these tools were from the highlighting category. With the use of these features, the user can have a much better understanding of how to navigate through the information on the visualization. These features also keep the data clean and organized by only covering one piece of information at a time, which will make the user more inclined to dig deeper into the data.

Which tools did you use from each of the 3 categories of Narrative Structure (Figure 7 in Segal and Heer). Why?

For Narrative Structure, the tools we utilized in our visualization were hover highlighting / details, selection, navigation buttons, explicit instruction, captions / headlines and accompanying article. The tools that came from the interactivity category were hover highlighting / details, selection, navigation buttons and explicit instruction. When creating a successful visualization, it is very important to make sure the user can get involved in learning about the information provided and to make sure the story of the data is told in an enticing and easy-to-interpret way. These tools were useful because it allowed the user to have a hands-on experience in learning about the data set while also making it easy for the user to navigate through the visualization. The tools that came from the messaging category were captions / headlines and and accompanying article. These tools were utilized in order to benefit the user because it guides the user through the material and makes it easier to figure out where certain information can be found.

Visualizations

Explain the visualizations you've chosen.

The visualizations we have chosen consist of a scatter plot of stop and frisk encounters classified by race. The scatter plots makes it possible for the user to click on any district in New York City to see both a scatter plot of stop and frisk encounters classified by race. The donut graphs have specific representations as well. One graph shows the area's breakdown of demographic of the overall population while the other shows the area's breakdown of demographic of stop and frisk encounters.

Why are they right for the story you want to tell?

We want our users to get an understanding about the correlation between who lives in New York and who is getting stopped in New York. The use of a scatter plot will allow the user to see what areas are more concentrated with certain races without having to do any reading. The donut graphs provide further information to the reader again in a very clean and basic way that does leave room for interpretation. We feel the visualizations that we have chosen of utilizing a scatter plot as well as donut graphs will keep the user engaged in learning about the information presented as well as make the data organized and appealing to the user.

Discussion

What went well?

There are many things that went well as a result of constructing this visualization. Firstly, we were very excited to create the visualization and then use the viz to see the connection between the demographic in the boroughs of NYC and races that were involved in the stop, question and frisk practice by the NYPD. In our findings, we noticed that a substantial amount of stop and frisks were blacks, showing an obvious imbalance in the system. Additionally, we were happy with how the visualization itself turned out. We worked very hard to make sure it was easy to comprehend by users and that people would be engaged in the information included on the visualization.

What is still missing? What could be improved?, Why?

Just like with any project, there are always ways to improve on the work the at has been done. Although the data processing went well, there was some missing data on the stop and frisk data set that would have been incredibly helpful in further simplifying and specifying our data. Because there were no zip codes listed in the stop and frisk data set, it was necessary to do reverse geocoding to get the zip codes. This definitely set our original ideas back a little bit considering our they were based on having that information at hand, however we worked hard with the information that was provided to us and were able to draw some other conclusions that we did not originally think of. Additionally, our visualization could have had a bit more annotation that elaborated on certain statistics about stop and frisk with each race to provide more information to the reader.

Contributions

Our group did a really good job at making sure everyone was involved in all aspects of the project. We wanted to make sure each of us had a solid understanding of every component of the project, however we split up main responsibilities for each section. Emily was responsible for finding the datasets and doing background research on the information within the data to get a better understanding of how we can analyze it. She was also responsible for the motivation and genre sections of the report. Kyle was responsible for getting the data ready for the visualizations. He cleaned it up and manipulated it so it only reflected the data needed for the project. He also completed the basic stats and discussion section of the report. Adam was responsible for creating the visualization to its entirety and making sure it utilized all of the necessary tools to make it as effective as possible for users. He was also responsible for the visualization section of the report.