The Wild Blazers Project Proposal

Team Members: Anish Bhardwaj

Major Theme evolved

Investigating correlations between factors on both the officer and the victim's end to see whether there is bias in the officers conduct with victims. To do so, I will be looking at the statistics that take into account the officer's time on the force and the officer's gender with the complaints per officer and on the victims end will look at gender, race, and age of the victim.

As mentioned in the feedback and through the course, my original theme of seeing race on both sides of the equation was something that would have been an issue or a banned theme per se. As a result, I changed the project theme toward looking at other factors that could correlate between the officer and the complaints.

As a result, I have adapted this theme to now look at different factors for the officers (Time on the force and gender) to start with and then look at Gender, race, and age of a victim for the correlation.

The SQL and d3.js checkpoints will allow me to confirm whether there is a correlation between the above mentioned factors following which, working with a regression model will potentially allow me to understand which sets of officers may need sensitivity training at which points in their careers as a result of their complaint rates and time on the force

Relational Analytics:

Questions/Tasks

1. What is the Complaint Rate for police officers in the different districts of Chicago?
    a. For this I plan to calculate a value that represents the Allegations per police officer per capita of a district
2. What is the racial distribution of officers in the districts of Chicago?
    a. For this, I plan to pull the data with regards to the race distribution of the police officers in a district
3. What is the gender distribution of officers in the districts of Chicago?
    a. For this, I plan to pull the data with regards to the gender distribution of the police officers in a district
4. What is the racial distribution of citizens in the districts of Chicago?
5. Combine the demographic distribution of a district correlate with the Officer Complaint Rate into one table
6. Combine the racial distribution of officers in a district alongside the racial distribution of the citizens of a district.

Interactive Visualization:

1. Population to Police Officer Ratio and Allegation Rate per police officer per district (Combined Interactive Visualisation)
2. Population Race Distribution vs Police Race Distribution per district
3. Gender Distribution of Police Officers per District

Machine Learning:

1. Create a Model that determines, based on an officer’s demographics and district information, whether they are more likely to be booked for an allegation by determining a potential risk factor (Scale of 1 to 5) where the high risk suggests that a police officer should potentially undergo sensitivity training to ensure that excess allegations can be avoided. This will take into account things like the Officer’s gender, race, the district they serve and other factors and try to predict a risk level based on the statistics found in previous sections. I plan to use a few different Classification models here (Decision Tree, K-Nearest Neighbour, Logistic Regression). I will create labels for each police officer using the mean and standard deviation of allegations in the district.

2. I would also like to run an Unsupervised learning model (like K means clustering) on officer allegation data. I want to do this to cluster officers who have complaints filed against them by civilians to understand which factors seem to have the most influence on whether a complaint will be filed against an officer or not

   The reason I want to do this is because while my previous task will help me classify the officers and understand whether an officer should be recommended for sensitivity training, there is still an internal bias within me based on the news and stuff that we hear in general that seems to suggest that race of an officer is the only thing that influences allegations or complaints in general. I want to take the officer information and cluster this information after which I will find the optimal number of clusters. For this, I believe I will include the age, gender, race, years served, number of allegations, district and perhaps some more officer info and then cluster to see what factors most influence this grouping just to see whether the bias that exists in the media and within me will be confirmed or not.