Gabrielle Salamanca
September 10, 2024
Math 748: Project Proposal

I. Project Description

There can be multiple reasons why a police officer may stop someone on the road: a traffic violation, the possibility of being part of a crime, a warning about a situation, etc. In this project, my objective is to find out if there's some sort of pattern for those who were stopped, if there's any common characteristics on whether or not someone received a citiaton. And it can be furthered to questions like: was there a higher percentage of a certain race being stopped, being given a citation; do certain police precincts or neighborhoods have a significant amount of incidents?

II. Data Description

Retrieved from the datasets of Vincent Arel Bundock's Github, I have chosen the dataset called Minneapolis Police Department 2017 Stop Data, which was obtained from http://opendata.minneapolismn.gov/datasets/police-stop-data. Bundock's version of Open Data Menneapolis' is much smaller than the original. The original has 231,448 records and 18 columns that spans from August 1, 2017 to Septmeber 9, 2024; while Bundock's is within 2017 and has 51,857 records and 14 variables.

The 14 variables are: idNum, date, problem, citationIssued, personSearch, vehicleSearch, preRace, race, gender, lat, long, policePrecinct, neighborhood, and MDC. I may omit lat and long, which are the latitude and longitude of the location of the incident, somewhat rounded. This is only because I'm not interested in exact coordinates, and I believe the neighborhood variable, which provides the name of the Minneapolis neighborhood of the incident, will suffice. I may also omit MDC, because I don't believe how the data was collected is related to the objective I want to explore, but it could be something to use for fun, just to see the spread of in-vehicle vs not in a vehicle like horseback or on foot.

III. Supervised vs Unsupervised

Because I want to find patterns in the data, this project will be using unsupervised learning. I can cluster the data in various ways such as doing it by neighborhood, race, gender, etc. However, there is the possibility of doing some supervised learning, using citationIssued as the response variable.

IV. Comments/Concerns

I do have concerns about empty values, because I know paperwork can be terribly tedious for the police. There's always a possibility of the police forgetting details, but considering there's 51,857 cases, I don't think I have to worry about a lack of samples. I also wonder if this could be a case of semi-supervised learning, but I don't know much about it like supervised and unsupervised.