Team 1

Aaby, Avalos, Bergquist, Betts

As a group, we chose to use two separate datasets and relate them to each other. The first dataset we chose was from the Comprehensive Hospital Abstract Reporting System (CHARS). CHARS is a department of health system which records down patient information and hospital stays. The second dataset we found from data.seattle.gov. This dataset consists of crime that occurred in Seattle. This dataset is from the Seattle Police Department. Both datasets are extremely high dimensional. The first dataset is 50X1,148,009 and our second dataset is 18X702952. As you can see, both datasets are extremely large and will need to be cleaned. We will be storing our data in our GitHub repository. We will also be joining our data using python. The way we will do this is by using the “pd.merge” function. This function will allow us to merge both datasets and see where each dataset has commonalities on dates that people were admitted to the hospital and dates people committed crimes. We will also be able to see commonalities among the locations that people were arrested and admitted to the hospital. If there are no similarities between both datasets then the input will be expressed as NULL. The dataset seems to be numerical and categorical. For example we have zip codes in our data which is considered categorical. It seems like we also have some missing values in one of our datasets. Luckily we aren’t missing any data from the actual data we plan on using. This will be a part of our cleaning process.