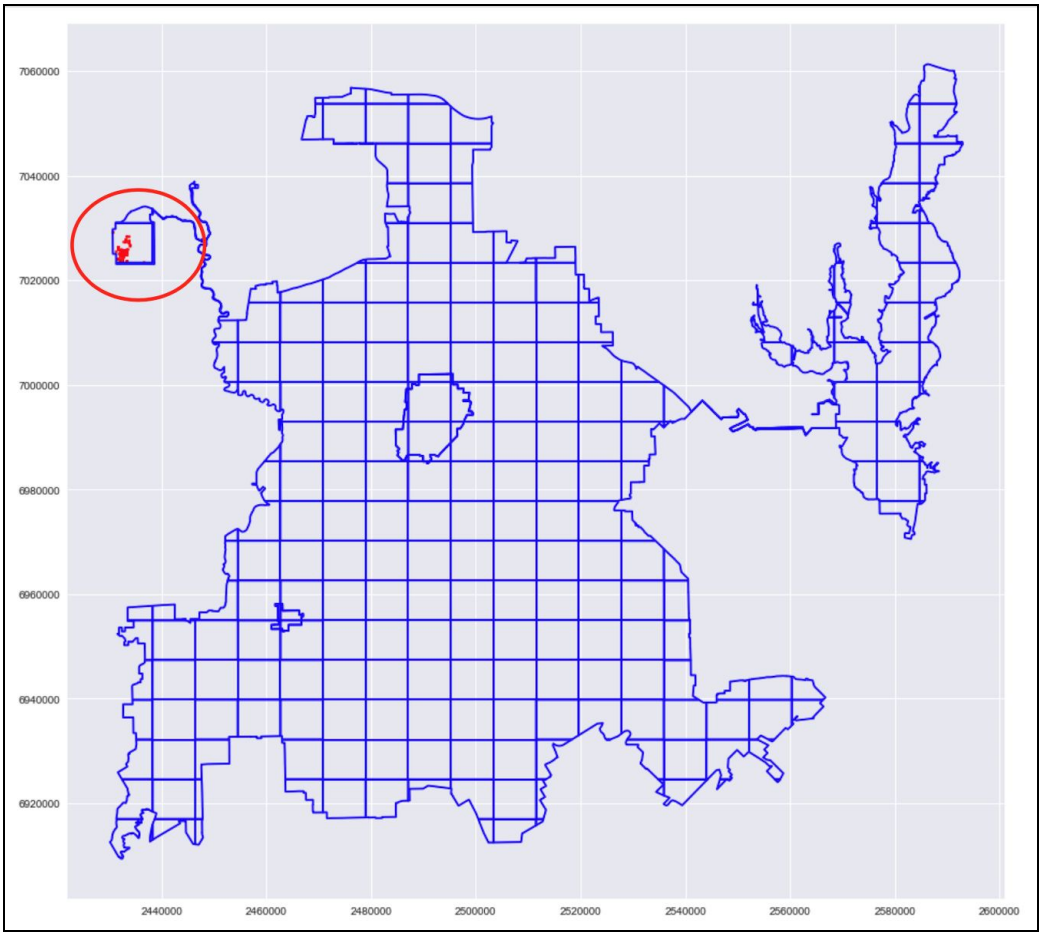
Assignment 2: Initial Findings Report and Executive Summary

MSDS 498 (Spring 2020): Bryan Bruno, Michael Haugan, Ksenia Luu

City Grid

To predict the location of future crimes, we needed to simplify the geospatial crime data. We decided that the best approach would be to split each city into regions using a square grid. All three (San Francisco, Minneapolis, Dallas) datasets have latitude and longitude coordinates for each crime event. A grid will allow us to simplify our location data by classifying each crime latitude/longitude coordinate based on the polygon that it falls within on the city grid. Below is an example of the grid that we created for the city of Dallas. We created a 20x20 square grid covering the full length and width of the city and then trimmed the grid to fit only inside the city boundaries. Using the trimmed-grid polygon coordinates, we tagged all 570,292 crimes to 207 regions based on their latitude and longitude coordinates. We tested our tags by mapping only points that fell within Region 2 against the full city grid to see if they were located inside a single polygon.



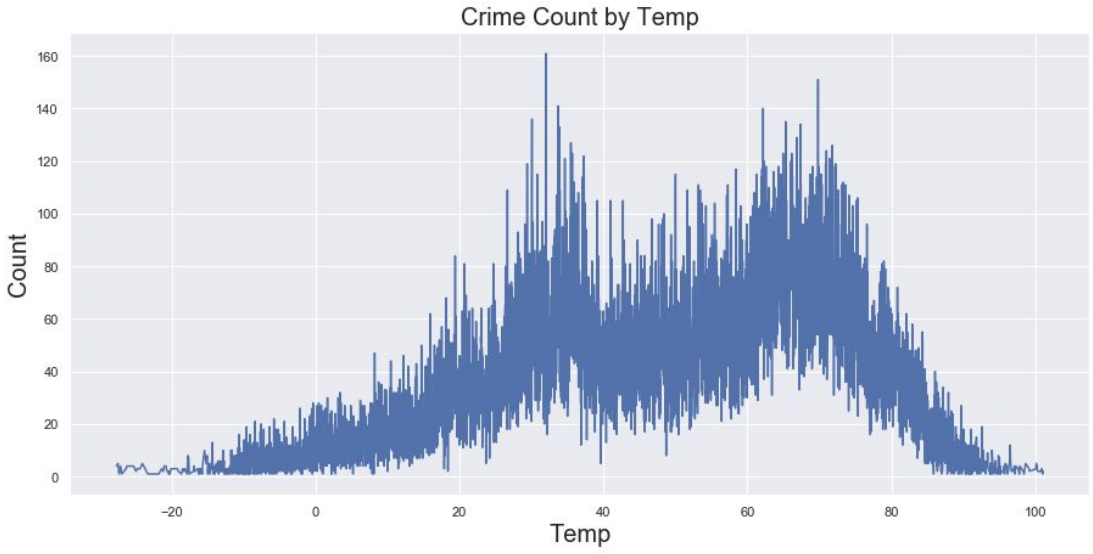
Dallas City Grid

We are now working through applying the same logic to Minneapolis and San Francisco. Our next challenge is to figure out the correct grid size for each city to keep our region areas roughly the same. Dallas city measures at 385.8 square miles, making it much larger than both San Francisco at 46.87 square miles and Minneapolis at 57.49 square miles. We want to make sure that our dimensions stay roughly the same across all three city grids in order to keep all variables the same across all three models.

Weather Data

We believe a huge factor in understanding if a crime may happen is the outside environment conditions. For example, if it is raining or heavily snowing are crimes less likely to occur? Likewise, if it is excessively hot or freezing cold does crime go down? These are relationships we set out to answer throughout our EDA process by bringing in weather data to our crime dataset to validate our assumptions. We also believe weather data could be very valuable information added to our predictive model if our EDA shows some strong relationships among our different weather variables.

We used hourly historical weather data from openweathermap.com Specifically focusing on Minneapolis for the time being as Minneapolis goes through very polarizing seasonal weather. This will give us a really good indication of a potential relationship between crime and weather. Some examples of weather variables we are able to bring in from our weather dataset are: temp, humidity, windspeed, rain within the past hour and snow within the past hour. At a very high level, there is definitely something interesting happening with crime and temperature that seems to support our notion that weather may definitely play a part in the amount of crime that occurs. For example, see the plot below of temperature in Minneapolis when crimes were committed. It appears there is at least some evidence that we have a relevant relationship existing here:



Crime counts by associated temperature

We see crime definitely tail off at the extremes on either tails. As it gets colder crime goes down and as it gets very hot crime also goes down. This supports our prior assumptions, but nonetheless is fascinating to see. This is information that definitely will help our model better predict crime. We will also continue to look into the other weather variables to see if we can find other associations with crime.

Initial Model

Based on our data and it's structure, we have two different planned models to best represent our findings. The aforementioned grid allows us to group aggregated crime types in a concentrated area. This removes the excessively precise longitude and latitude coordinates for location. Additionally, bucketing crimes by certain hourly intervals gives us a larger window to predict crimes. All other features may remain as is.

The first planned model will be a generalized classifier to predict the probability of crimes occurring at all locations on the grid at a certain time and date. The final output will be represented as a percentage. Based on the performance and accuracy of this model, the second may not need to be developed. However, if this model does not function at a high enough confidence level, we will transition to a time series based model. Taking into account the most crime populated areas on the grid, we will create a LSTM RNN for each top-ranking area and output a time and date forecast.