Domain:

Open data from the city of SF. I will use 311 call requests to clean up human waste to predict when and where people will poop on the street. This data will be geographic (latitude and longitude) and time series (granularity down to the hour).

I have limited experience in this domain. I have worked on some projects in the past where we wanted to measure the impact of sales on stores when they were in a certain proximity to certain store types. I worked with time series data in the past as well (Weekly Sales Data)

Data:

There are 104000 incidents of human waste in SF since 2008. Depending on how I structure the problem, there could be a total of 100,000,000 rows of data. The smallest unit of measurement I am considering is one block and I can build up from there if it is too granular.

I will pull in census data, weather data, and yelp data to enrich the dataset if time permits. Otherwise the MVP will consist of time of day, day of week, season, and neighborhood variables.

Known Unknowns:

There is a very low incidence rate of human waste (less than 0.001%) which could prove difficult in trying to predict when and where this might happen. Bringing in outside datasets could be difficult because geolocation data adds another layer of complexity. There might not be any patterns present and it could be truly random when and where people poop on the street.