Garrett Yamane

Capstone Project 1 Proposal

## Problem Statement

San Francisco crime and safety are concerns for anyone wanting to travel, visit, or live in the city. These concerns highlight the importance of staying educated regarding police reports that have occurred in the past. Are certain areas more dangerous than others? What times of the day do more crimes tend to occur, and do additional factors such as the time of the year (Christmas, Spring Break, etc.) influence the risk of crime? These types of questions help people in San Francisco keep safe and aware of potential danger that could pose a direct threat to them. I want to observe past crime data and police reports to illustrate key features that contribute to common crime patterns in San Francisco. I will use my findings to model San Francisco crime data in order to decipher important factors that contribute to San Francisco crime patterns and ideally predict future potential crime location/classification. The City of San Francisco could use this to better prepare for certain events when crime rates tend to spike and take appropriate precautionary measures, and target areas that are unsafe to keep visitors and residents protected from harm.

## Target Client

The San Francisco Police Department has to be prepared at all times for any potential report of crime. They must act accordingly and it is essential that they are readily on duty at peak crime times and areas. The more informed they are about the patterns and conditions that trigger spikes in crime rates, the better it will be for both the Police Department and public. My client (SF Police Department) can use the models I build to better understand what crimes may happen in the future by targeting the features of the data that I highlight in their crime report dataset. They can prepare for certain events and be aware of certain conditions that tend to lead to higher rates of crime, making sure that police units are readily available and onsite in special cases.

## Data

I will be pulling data from the Kaggle Dataset, "SF Police Calls for Service and Incidents" and the official Crime Data Warehouse for San Francisco. Each report has certain data features, such as:

- Time of day crime was reported reported
- Day of the week crime was reported

- Date reported
- Crime description
- Type of crime (classification)
- Resolution to the crime (book/arrest, nothing)
- Unique Crime ID
- Location of crime
- SF District associated with crime
- Description of crime
- Address type (intersection, premiss address, etc.)

There are over 2.2 million reports available in the dataset, which makes for a very rich sample of data to train and develop a model over. There are two different datasets that I will congregate into one larger dataset:

- https://www.kaggle.com/san-francisco/sf-police-calls-for-service-and-incidents#police-department-calls-for-service.csv
    - This dataset contains all police reports/calls from 03/31/2016 until 12/31/2018.

---

CHANGE:
https://data.sfgov.org/Public-Safety/Police-Department-Incident-Reports-Historical-2003/tmnf-yvry

- This dataset is more similar to the updated dataset listed below. Each row in the first dataset that I listed is a police call for service, whereas the rows in this dataset and the dataset below are crime reports submitted

---

- https://data.sfgov.org/Public-Safety/Police-Department-Incident-Reports-2018-to-Present/wg3w-h783
    - This dataset contains all police reports/calls from 08/2018 to the present. The first dataset stops at the end of 2018 due to the source discontinuing the updates. This website from the Crime Data Warehouse is updated daily and contains reports to the current date.

## Approach

I will first begin by splitting up the data and creating **visualizations** to highlight each feature relating to the crime incident reports. This will make referencing data and storytelling easier by clearly illustrating individual factors contributing to crime reports in San Francisco. I believe a good way to address my problem statement is by treating the

problem as an **unsupervised problem** and performing **clustering** (e.g., through the KMeans clustering algorithm, but I will try others as well)**.** By going through an iterative process to find a good cluster number, I will be able to find a good model that groups similar crimes together so I can find crime patterns through the results. Because there are a lot of different features, graphing all points and visualizing clusters on a single plane will be quite difficult. By performing **dimension reduction** (for instance, using **PCA**), I will be able to focus on the more impactful crime features and clearly graph the data points so the clusters are visible. I can then analyze the clusters, for instance by seeing what the crimes in each cluster have in common (i.e. location, time of the year, time of day, etc.). Time-permitting, I will explore questions related to predicting potentially where, when, or why crimes happen given certain features. I would also like to explore questions like "Which neighborhoods can the SF Police Department target to help increase the safety of SF?" and "Can we use the data to see if certain events such as fairs, concerts, and public speakers cause crime rates to fluctuate?"

## Deliverables

As required, I plan on making a comprehensive Jupyter notebook that is annotated to make reading the code and visualizations as clear as possible. I also plan to make a slide deck that addresses my goals for the project, the steps I took to solve the business problem I presented, and the results that I found from analyzing the data. Lastly, I plan to build a thorough report that documents my findings throughout my project.