Capstone Project 1: Submit Your Capstone Project Proposal

Jiacheng(Dylan) Qian
Data Science Career Track
Springboard

After discussing with my mentor, Quan, and also looking over all three datasets I chose before, I have decided to go along with the Chicago crime dataset. Below is the brief introduction of the dataset that is written by Analytic Vidhya, a data science website for both beginners and professionals.

> **Chicago Crime Data Set**
>
> The ability of handle large data sets is expected of every data scientist these days. Companies no longer prefer to work on samples, they use full data. This dataset would provide you much-needed hands-on experience of handling large data sets on your local machines. The problem is easy, but data management is the key! This dataset has 6M observations. It's a multi-classification problem.

I think this dataset is fairly interesting and have many different aspects I can explore. Also, I usually have problems dealing with datasets in large size, and this dataset has more than 6 million observations. Thus, I want to challenge myself to analyze a bigger dataset.

In this project, I want to fully explore this Chicago crime dataset in order to answer the following 3 questions:

1. What is the trend of amount of crimes in Chicago from 2001 to the present?
2. What location and time that the most of crimes happened?
3. How is the crime vs arrest ratio in the past few years?

The main audience or client for this project is the local police and government in Chicago. The answers to those questions can reflect the performance of police and government for the past 17 years, and at the same time, provide the information for them on what preventive action they should perform. For example, they should know from this project that what time and location they should pay more attention in order to minimize the occurrence of crimes.

This dataset is available on City of Chicago's Data Portal, a website that "dedicated to promoting access to government data and encouraging the development of creative tools to engage and serve Chicago's diverse community." The link of this dataset: https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2. This dataset contains all the information from 2001 to the present with 22 variables and 6,516,655 observations.

I will use the techniques I have learned in past two months from Springboard. First, I will extract the variables that I want to explore. Then, I will clean the data and fill/correct the missing and wrong values. Last, I will use exploratory data analysis with many visualizations to examine the dataset in order to answers the questions I want to dig into.

For this project, I will focus on producing a clear and accessible document for the audience who do not have coding experience. This requires me to make understable comments along with the code and a thorough paper as the deliverables.