**Milestone 1**

**Data Source**:

The three data sources for my project are from Kaggle, New York City Government website and OpenWeatherMap. Links for the dataset from New York City government and OpenWeatherMap are attached in reference below and for the dataset from Kaggle is located in GitHub repository.

**Excel file** - Kaggle is an excellent source of getting datasets from. I will be using dataset from Kaggle which includes all valid felony, misdemeanor and violation crimes reported to the New York City Police Department from 2006 to 2019. The file contains list of location in New York City where the crime occurred, borough, zip code, time/hour of the day when the crime occurred and so on.

**Website** - Second source of data is from New York city open source website. The website is from New York City agencies which provides free public data. The dataset contains details of the crash events in New York city. Each row represents a crash event. Columns included in the dataset are crash date – occurrence date of collision, borough – Borough where collision occurred, zip code – Postal code of incident occurrence, crash time – Occurrence time of collision more.

**API**: It contains the location of city which is categorized by zip code. Since, I will be pulling historical weather data, other information I would have city id, country code, hour of the location, start time and end time of the weather.

**Relationship and Accomplishment:**

The above dataset is from New York City which has location as the connecting point between all three datasets which are borough, zip code, longitude and latitude. The relationship I would like to work on is what difference in crime exist between 5 boroughs, how weather plays a role in increase or decrease of crime occurrences, where does the most crime occur and what during what time of the day. All of the above data sources are connected by location such as zip code. The website has 1 to many relationships with excel file location and has one to many relationships with the API data by zip code as well.

To accomplish above task first, I would need to clean the datasets. In addition to that I would need to analyze the data, combine all 3 data sources if necessary, extract meaning from each analysis, construct question in each milestone which could lead into deeper analysis and finally present the data with expressive visualization to tell the story.

**References:**

<https://www.kaggle.com/adamschroeder/crimes-new-york-city>  
<https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95>