Crimes in Philadelphia

Team 3

**Data Science Capstone Project   
Data Acquisition and Pre-Processing Report**

Date:

02/12/2021

Team Members:

Name: Raj Patel

Name: Hong Son

Name: Kunal Sharma

[The purpose of this report is to describe the data of your project. It includes three major sections: Data Sources, Data-Processing, and Appendix]

**Identifying Data**

**Data Sources:**

[Identify the data sources of your project. It may have more than one data source. Describe each of them and explain why you select the data sources.]

Philadelphia Crime Data:

<https://www.kaggle.com/mchirico/philadelphiacrimedata>

For our project, we would like to conduct an analysis on a Philadelphia crime dataset published by OpenPhillyData. The dataset dates from 12-31-05 to 03-22-17 (majority of the dates are from 2006-2016). There are 14 total columns in this dataset. They include the district number, sector, dispatch date, dispatch time, location, general code, descriptions of the crime, and the district. This dataset was published to the public so people can assess the crimes happening in Philadelphia.

Weather Data:

<http://www.climate.psu.edu/data/city_information/index.php?city=phl&page=dwa&type=big7>

To assist in getting more out of our analysis, we will be scraping Philadelphia weather data for Philadelphia from 2007 to 2018. The dataset includes the temperatures in Fahrenheit, the max and min temperature, environment information, wind, and precipitation. We hope to look into the seasonal forecast and see if there are weather trends in crimes.

Parking Violations Data:

<https://www.opendataphilly.org/dataset/parking-violations>

We will be using the latitude and longitude points to map the zip codes in this dataset to the crime dataset using a key pair (lat, long) with zip code.

**Acquisition Process:**

[Describe the data acquisition process. Is the dataset ready for download? How do you download the data? Do you need to write your own code to acquire the data from a public or private source? Describe how you do it. Are there multiple data sources? How do you integrate the data from multiple sources? Any other process involved in the acquisition process?]

The crime and weather dataset has already been cleansed and merged from the previous term. For the parking violations dataset, data cleansing would include removing rows with null property values to have zip-codes in the dataset and then merging with the crime-weather dataset.

**Issues:**

[Are there any potential issues in data acquisition that have not been solved yet?]

No issues were found yet with the data acquisition process.

**Data-Processing:**

[Examine the data you have acquired and understand the data properties. Is there any pre-processing you need to do before you can start analyzing the data? For example, missing data, sparsity, noise, veracity, ambiguity, interoperability, etc. Describe each data issue in a sub-section and explain how you clean up the data.]

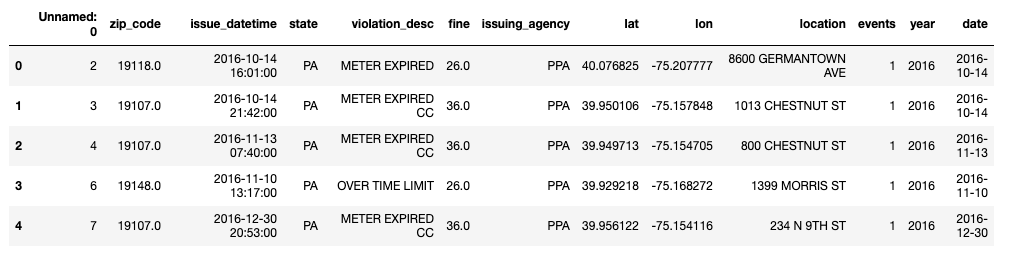
For the parking violation dataset, we looked at the columns and determined that we wanted to keep the latitude, longitude, and zip code so that we could fill in the zip codes for the crime dataset. We created a key that contains the min and max value [lat, long] for each zip code and then used that to map onto the crime dataset.

**Appendix**

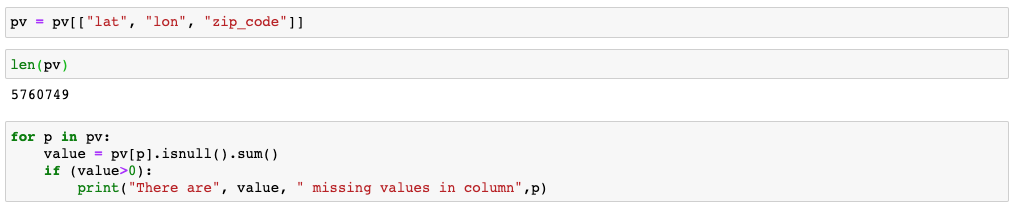
[Provide the code or pseudo code, data definition, sample data, and any other information in the appendix here.]

Parking Violations Data Dictionary -

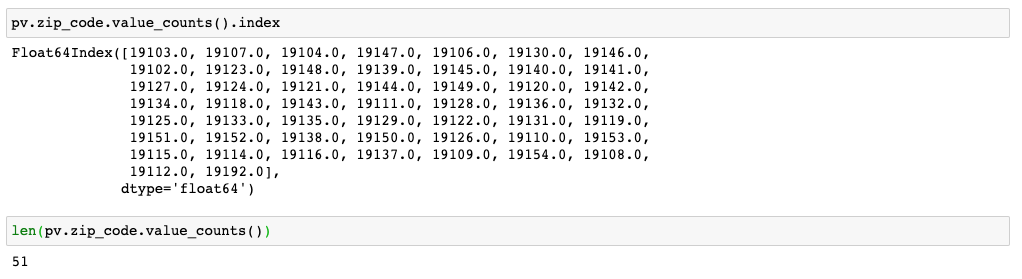
<https://metadata.phila.gov/#home/datasetdetails/5543865c20583086178c4eda/>



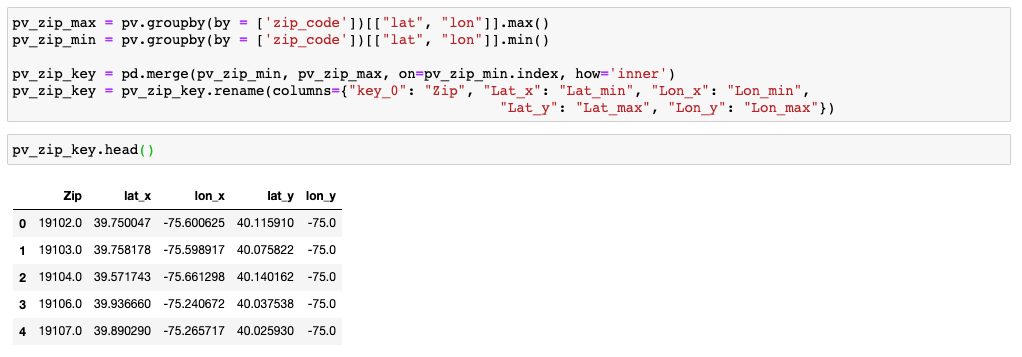
Top 5 rows of Parking Violations Data



Keeping columns we need, looking to see if there are any missing values



We have 51 zip codes we can use to map the data



Creating the key pairs

Table of Contributions

The table below identifies contributors to various sections of this document.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Section** | **Writing** | **Editing** |
| **1** | **Data Sources** | **Hong** | **Raj, Kunal** |
| **2** | **Data Pre-Processing** | **Hong** | **Raj, Kunal** |
| **3** | **Appendix** | **Hong** | **Raj, Kunal** |

**Grading**

The grade is given on the basis of quality, clarity, presentation, completeness, and writing of each section in the report. This is the grade of the group. Individual grades will be assigned at the end of the term when peer reviews are collected.