Crimes in Philadelphia

Team 3

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

Date:

10/20/2020

Team Members:

Name: Hong Son

Name: Raj Patel

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.

**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 dataset is available as a CSV file while the weather dataset is an XLS file. We don’t anticipate having to write code to acquire any data from outside sources. Very little data cleansing had to be done as there was very little missing values and invalid data.

The weather data file will have to be cleansed to include the month and year which are indicated by “Month-Year” on the headers. We will utilize a left join the data using the date field.

**Issues:**

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

During the weather data cleansing, we are missing a couple of dates that were cut off. Hong has reformat the PDF files so that the dates aren’t cut off and has a new file for the weather data.

We are on track to continue working on the next steps in the project.

**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.]

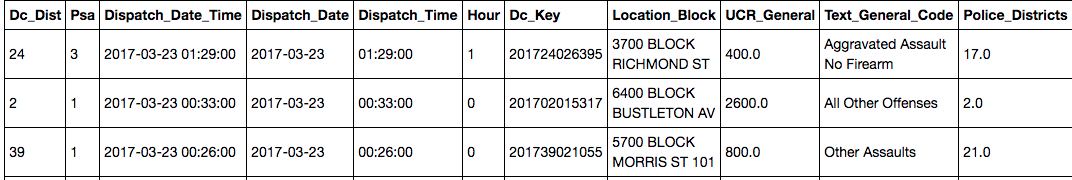
There was missing data in the crime dataset, albeit not a large amount. We removed the rows with missing values. To find locations in North Philadelphia in our data set we first found the Geo location of the town hall ([39.952479, -75.163668]) which is right at the center of philadelphia and then we found the rows in our dataset where the latitude is more than the latitude of the center of Philadelphia. Similarly, for locations in South Philadelphia we found all the rows in our dataset where the latitude is less than the latitude of the center of Philadelphia. To find locations in West Philadelphia we found all the rows where the longitude is less than the longitude of the center of Philadelphia. In this case the rows we found were a combination of locations in north and south philadelphia.

For the weather data cleansing, we didn’t notice any incorrect values. We were only missing a couple dates when we formatted it into PDF files. Hong has reformat the PDF files so that the dates aren’t cut off and has a new file for the weather data.

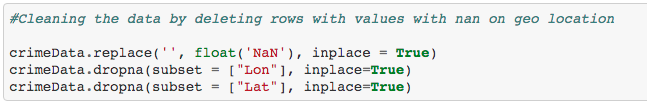
**Appendix**

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

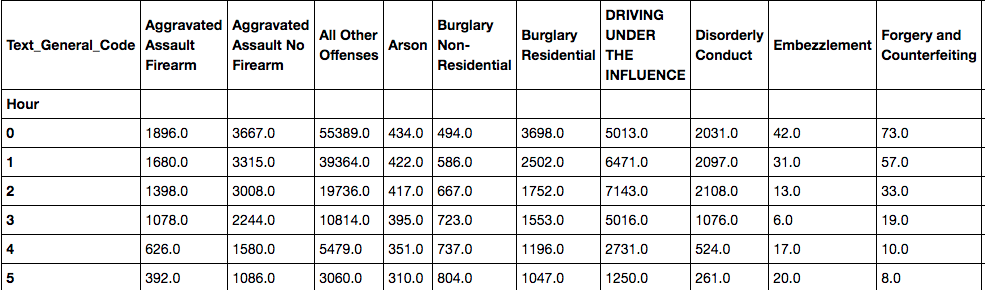
Crime



Top 3 rows in crime data [sorted by Dispatch date]



Remove rows with missing values in locations



Creating a dataframe with hours and number of crimes

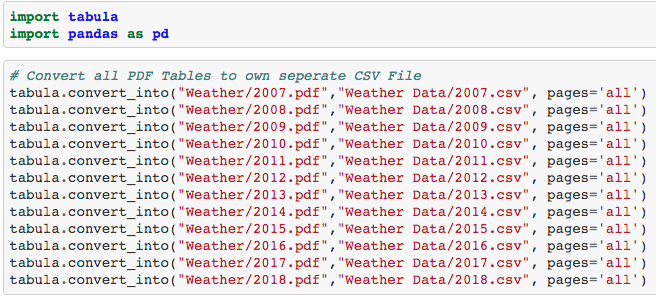


Looking at regions [North, South, West]

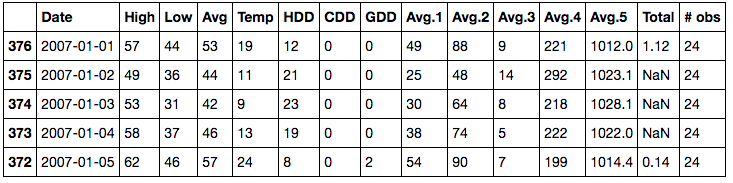
Weather

Weather Data Dictionary -

<https://github.com/hongson1218/Crime-in-Philadelphia/blob/master/Weather%20Data%20Dictionary.docx>



Using tabula to convert pdf to csv files



Looking at top 5 rows

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.