INFM600 Information Environment

Team Exceptional

Data Cleaning and Documentation

November 2, 2016

Document Overview

The purpose of the document is to keep a track of the progress of the project. This phase includes (supporting) data extraction, data validation as well as data cleaning. By the end of this phase, we will have clean data required to answer our research questions.

Project Description

The main dataset that we are working on for our project is the DC Metro Crime Rate. The dataset was obtained from *DC.gov* website which is an initiative by the government of District of Columbia to make governance transparent. (“Crime Incidents - 2011,” 2016, “Crime Incidents - 2012,” 2015, “Crime Incidents - 2013,” 2015, “Crime Incidents - 2014,” 2016, “Crime Incidents - 2015,” 2015) The dataset contains a subset of locations and attributes of incidents reported in the ASAP (Analytical Services Application) crime report database by the District of Columbia Metropolitan Police Department (MPD).

As per the Terms and Conditions of Use for District Data, the District allows the users to copy, modify, distribute, and perform the data, even for commercial purposes, all without asking permission (“Terms and Conditions of Use for District Data,” n.d.) The DC Metro Crime dataset is a part of the District Data.

The metadata is very well-defined. Each and every entity and attribute is explained along with details like access constraints, use constraints, point of contact as well as credits for the dataset (“The District of Columbia,” 2007).

Following are a few questions that we aim to answer with the abovementioned dataset and a few other relevant datasets:

1. Does weather (temperature change) have any effect on the crime rate?
2. Is there a relation between number of nightclubs with the crime rate in a particular geographic unit (cluster)?
3. Is there a spike in the crime rate during the major public holidays?

For each research question, we drill down to the rate of each type of crime (arson, robbery, etc.).The following section describes the supporting datasets that we plan to use to along with the DC Metro Crime Rate dataset, to answer the abovementioned questions.

Supporting Datasets

Following are the datasets we plan to use:

* **Historical Weather Data (for DC)**

The historical weather data of D.C area has been extracted from *NOAA (National Oceanic and Atmospheric)* website. The desired range is from January 1, 2011 to December 31, 2015. Our original weather dataset contains 1,826 rows and 20 columns. The original dataset contains the following information: Date (Year, Month, and Day), Temperature in degree Fahrenheit (Max, Minimum and Observation), Precipitation, Evaporation, Soil Temperature etc.

The Terms and Conditions of Use for the *NOAA Climate Data* claims: “The data available through CDO is available at no charge and can be viewed online or ordered and delivered to your email inbox.” (“NOAA,” 2016)

All weather data was collected by weather stations. For weather metadata, National Centers for Environment Information (NCEI) provides a dataset which names “Climate Divisional Dataset”, the dataset contains temporally and spatially complete information on 344 climate divisions in the contiguous United States. These data are used to generate historical climate analyses for drought, temperature, precipitation, and heating/cooling degree data (“National Centers for Environmental Information,” 2016).

* **Night Clubs Data (for DC)**

The night clubs dataset is available on *DC.gov* website (“Night Club,” 2013). The dataset details like location of the clubs, the geographic co-ordinates, etc. It also contains the MAR (Master Address Repository) ID. The MAR is a database of building addresses, blocks, intersections, place names and other location identifiers in the District (“Master Address Repository,” n.d.)

As per the Terms and Conditions of Use for District Data (of which the nightclubs data is a part), the District allows the users to copy, modify, distribute, and perform analysis on the data, even for commercial purposes, all without asking permission (“Terms and Conditions of Use for District Data,” n.d.)

The metadata for this dataset is very well-defined. There is a separate page dedicated to metadata. It includes all possible information like the description of the dataset, the time period of the content, Access and use constraints, point of contact, data quality information, entity and attribute description of the dataset, etc. (“DCGIS.NIGHTCLUBPT,” 2013)

* **Address Points Dataset**

This data has been retrieved from the *DC.gov* website (“Address Points,” 2004). This dataset contains locations and attributes of Address points, created as part of the Master Address Repository (MAR). It contains the addresses in the District of Columbia which are typically placed on the buildings. This dataset contains information (mapping between the clusters and the unique address-ids given in MAR) that is useful for merging the crime dataset with the nightclubs dataset.

As per the Terms and Conditions of Use for District Data (Address Points data is included in this), the District allows the users to copy, modify, distribute, and perform the data, even for commercial purposes, all without asking permission (“Terms and Conditions of Use for District Data,” n.d.)

The metadata for this dataset too, is very well-defined. It contains information like the description of the dataset, the time period of the content, Access and use constraints, point of contact, data quality information, entity and attribute description of the dataset, etc. (“Address points - MAR (AddressPt),” 2004)

* **List of Public Holidays (for United States of America)**

The data has been retrieved from the *DC.gov* website (“Holiday Schedules,” n.d.). The original dataset contains information about the list of Federal and local holidays in Washington DC area and is organized in the following columns: Date, Year, Holiday, Comments.

The Terms of Use for this dataset state that this dataset is openly available to the public. Under Section 3 of their Open Government and Open Data Directive, it is mentioned that “Datasets under paragraph (4) shall be made available in accordance with technical standards published by OCTO not later than November 1, 2014 that ensure that data is published in a format that is machine readable, and fully accessible to the broadest range of users, for varying purposes. Datasets shall be made available to the public on an open license basis. An open license on a dataset signifies there are no restrictions on copying, publishing, further distributing, modifying or using the data for a non-commercial or commercial purpose.” (“Transparency, Open Government and Open Data Directive,” n.d.)

As far as the metadata is concerned, the dataset is seemingly too small to demand a metadata and all the column names are self-explanatory.

Data Cleaning: Issues and Remediation Steps

The DC Metro Crime dataset is common for all research questions and need to be merged with the supporting datasets on some common attribute. This attribute differs for each research question and hence, the data cleaning for all three questions yields different crime datasets: Weather\_Crime\_Incidents2011-15.csv, NightClubs\_Cluster.csv and Holidays\_Data.csv. The steps for data cleaning process for each question are mentioned under the relevant supporting dataset’s data cleaning process.

* **Historical Weather data (for DC)**

The original dataset was in the pdf format and each record (corresponding to a day) was reported as follows:

”2011-1-8,34,25,30,0.00………”.

The original pdf file was directly transformed into a comma-separated values format for the convenience of subsequent data cleaning and analysis.

The question what we are concerned about is “Does weather (temperature change) have any effect on the crime rate?” The original weather dataset contains a lot of information, including the temperature and other factors. However, we only need to pay attention to date and the change of corresponding temperatures, and not just for “max temperature”. We also might consider other weather-related factors like “the range of temperature in a day”. We want to analyze the effect of different weather-related factors that influenced the crime rate, rather than “fishing” the results what we are expected.

The weather dataset is quite “clean”. Every data format is standard.

Regarding the integration of weather dataset with DC Metro Crime dataset (crime dataset), every observation has the information of a crime incident’s starting date like “2015-01-03T22:30:00.000Z” in the crime dataset. The reason why we only focus on the “start date” because in some cases, the “report date” in the crime dataset was a few days later than the day when the case occurred. Therefore, the following steps were taken to merge the two datasets:

1. To arrange cases according to start date in ascending order and only retain the case rows of start date from January 1, 2011 to December 31, 2015, all other irrelevant rows were omitted.
2. To change format of start date in weather dataset, from “1-1-2016” to “1/1/2016”, the “Find/Replace” function of Microsoft Excel was used.
3. To change format of start date in crime dataset, from “2015-01-03T22:30:00.000Z” to “1/1/2016”, the “Text to Column” function of Microsoft Excel was used.
4. To merge the two aforementioned datasets, *VLOOKUP* equation function of Microsoft Excel was used.
5. For the final Weather\_Crime\_Incidents2011-15 file, there are 7 columns (ReportDate,Shift,Offense, Start\_Date,MaxTemp,MinTemp,RangeTemp) and 179400 rows in the dataset.

* **Night Clubs Data (for DC)**

The crime dataset contains location details in various forms: wards, districts, blocks, clusters, voting precincts, etc. Determining the attribute to merge this dataset with the crime dataset was challenging. Cluster number was chosen since the zip code is not mentioned in the crime dataset. The District of Columbia is divided into 39 clusters and hence, cluster number was chosen as the basis for merging the two datasets.

Another issue faced was that the nightclubs dataset does not have the cluster number as a field. Only the street address (not even the zip code) and MAR (Master Address Repository) id are mentioned. Following steps were taken to merge the nightclubs and crime datasets:

1. Following R script merges the two datasets- nightclubs and Address points (contains MAR id and cluster mapping):

*# loading data.table package for setnames function*

*> library(data.table)*

*> MAR\_Cluster<- read\_csv(choose.files())*

*>Night\_Club<- read\_csv(choose.files())*

*# changing the name of a column to match that in the other dataset, for merging*

*> setnames(MAR\_Cluster, "ADDR\_ID", "ADDRID")*

*# merging the two datasets by common attribute ‘ADDRID’ i.e. MAR id*

*> final= merge(MAR\_Cluster,Night\_Club,by="ADDRID")*

*#Saving the data frame as a CSV file*

*> write.csv(final,"NightClubs\_DC.csv",row.names = FALSE)*

The NightClubs\_DC file contains 57 columns and 96 rows.

1. There were several extra columns like the coordinates, entrance types, etc. in the Address Points dataset, which were quite irrelevant for our purpose which got merged in the NightClubs\_DC dataset. The only columns retained were AddressID, Name of the nightclub, Address and Cluster. The following Excel VBA code deletes the irrelevant columns:

*Sub sbVBS\_To\_Delete\_Multiple\_Columns()*

*Columns(“C:AZ”).EntireColumn.Delete*

*End Sub*

Following this step, the NightClubs\_DC dataset contains 4 columns and 96 rows. We have not ascertained as to what statistical measure we will use to relate the nightclubs with crime in a particular cluster. We plan to compare the number of crimes in a particular cluster in a year to the number of nightclubs in that particular cluster. In order to do so, we need to compute the count of nightclubs, clusterwise. The following R-snippet does the required computation and stores it into a csv file named “NightClubs\_Cluster.csv” which contains 2 columns and 14 rows:

*# Loading readr library to read csv file*

*> library(readr)*

*# choosing the required dataset from the file directory*

*> NightClubs\_DC <- read\_csv(choose.files())*

*# computing the count of nightclubs in each neighbourhood cluster*

*>clubs.cluster=table(NightClubs\_DC$Cluster)*

*# writing the results down in a csv file*

*> write.csv(clubs.cluster,"NightClubs\_Cluster.csv", row.names = FALSE)*

The crime dataset had few crime records where the crime location in terms of cluster was missing. Around 9% percent of the total records (2.6k out of 2770k records) of the crime dataset have been removed using the Filter feature in Microsoft Excel.

Finally, the irrelevant columns like ward, geographical coordinates, voting precincts, etc. have been removed using the above mentioned VBA code.

* **Holidays Dataset**

An issue faced while cleaning this data was that the columns Date and Year were mentioned separately having the format *“Month-DD”* and *“YYYY”* respectively. The following R script merges both the columns into a single column with the name “Date” with proper date format as *YYYY-MM-DD*. This made manipulation on the column “Date” much easier than before. The Holidays dataset now contains 61 rows and 5 columns.

*#installing tidyr package*

*> library(tidyr)*

*#merging the columns Date and Year into newDate2 and removing the old columns*

*> unite(HolidaysRefined, newDate2, c(Date, Year), sep = "-", remove = TRUE)*

*#Formatting the new columns into proper date format*

*> HolidaysRefined$Date <- as.Date(HolidaysRefined$newDate2, format = "%b-%d-%Y")*

*> View(HolidaysRefined)*

*#Saving the data frame as a CSV file.*

*> write.csv(HolidaysRefined, file = "Holidays Data.csv", row.names = FALSE)*

**Word count: 2072**

References

[1] *NOAA.* (2016). Historical Weather Data [Data file]. Retrieved November 10, 2016, https://www.ncei.noaa.gov/orders/cdo/839080.pdf

[2] *National Centers for Environmental Information.* (2016). Station Metadata. Retrieved October 31, 2016, <https://www.ncdc.noaa.gov/data-access/land-based-station-data/station-metadata>

[3] *NOAA.* (2016). Climate Data Online Help. Retrieved November 10, 2016,

<https://www.ncdc.noaa.gov/cdo-web/faq>

[4] *Night Club*. (2013, October 04). Retrieved October 22, 2016, from DC.gov, <http://opendata.dc.gov/datasets/4589d3e500404dc5a648dcdf3bc2732e_29>

[5] *Master Address Repository.* Retrieved October 28, 2016, from DC.gov, <http://octo.dc.gov/service/master-address-repository>

[6] *Terms and Conditions of Use for District Data*. Retrieved October 30, 2016, from DC.gov, <http://dc.gov/page/terms-and-conditions-use-district-data>

[7] *DCGIS.NIGHTCLUBPT.* (2013, October 04). Retrieved October 30, 2016, from Dc.gov, <https://www.arcgis.com/sharing/rest/content/items/4589d3e500404dc5a648dcdf3bc2732e/info/metadata/metadata.xml?format=default&output=html>

[8] *Holiday Schedules*. Retrieved November 1, 2016, from DC.gov, <http://dchr.dc.gov/page/holiday-schedules>

[9] *Transparency, Open Government and Open Data Directive*. Retrieved November 1, 2016, from DC.gov, <http://dc.gov/node/871012>

[10] *Crime Incidents - 2011.* (2016, August 23). Retrieved October 31, 2016, from DC.gov, <http://opendata.dc.gov/datasets/9d5485ffae914c5f97047a7dd86e115b_35>

[11] *Crime Incidents - 2012.* (2015, April 29). Retrieved October 31, 2016, from DC.gov, <http://opendata.dc.gov/datasets/010ac88c55b1409bb67c9270c8fc18b5_11>

[12] *Crime Incidents - 2013*. (2015, April 29). Retrieved October 31, 2016, from DC.gov, <http://opendata.dc.gov/datasets/5fa2e43557f7484d89aac9e1e76158c9_10>

[13] *Crime Incidents - 2014*. (2016, April 29). Retrieved October 31, 2016, from DC.gov, <http://opendata.dc.gov/datasets/6eaf3e9713de44d3aa103622d51053b5_9>

[14] *Crime Incidents - 2015.* (2015, December 17). Retrieved October 31, 2016, from DC.gov, <http://opendata.dc.gov/datasets/35034fcb3b36499c84c94c069ab1a966_27>

[15] *The District of Columbia.* (2007, June 26). Retrieved November 1, 2016, from DC.gov, <http://data.octo.dc.gov/Metadata.aspx?id=3>

[16] *Address Points.* (2004, January 01). Retrieved November 1, 2016, from DC.gov, <http://opendata.dc.gov/datasets/aa514416aaf74fdc94748f1e56e7cc8a_0>

[17] *Address points - MAR (AddressPt).* (2004, January 01). Retrieved November 1, 2016, from DC.gov, <https://www.arcgis.com/sharing/rest/content/items/aa514416aaf74fdc94748f1e56e7cc8a/info/metadata/metadata.xml?format=default&output=html>

Latest Update: 12/13/2016 10:51:33 PM