Capstone Project - Milestone Project

I. Introduction

The Chicago Data Portal publishes a dataset reflecting reported incidents of crime that occurred in the City of Chicago from 2001 to present.

In this project, I have examined the data, conducted data wrangling, utilized different charts identifying general trends for the crime cases in Chicago, and applied inferential statistics techniques to the data. Further, I would like to do broader correlation analyses between the crime cases and population, eduction, poverty level etc. In the last part of the project, I will apply some machine learning techniques to the data and build a predictive model for different types and trends of crimes in city blocks.

The Chicago city government would be the clients for this project. They can use my analysis report and model to identify the crime trend in each community area, and therefore mobilize resources more efficiently and effectively to reduce crime rate. Meanwhile, they can use my work to understand the impact of eduction, poverty level etc. on crime rate and identify a way to improve the community environment.

II. Data Sources

1. The data is available for download from the Chicago Data Portal as follows:

https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2

2. Chicago community area data covering eduction, poverty indicator etc.

https://datahub.cmap.illinois.gov/dataset/1d2dd970-f0a6-4736-96a1-3caeb431f5e4/resource/8c4e096e-c90c-4bef-9cf1-9028d094296e/download/ReferenceCCA20112015.csv

III. Completed Work

1. Data Wrangling:
   Overall, the data is quite clean. With further exploration, some cleanups are conducted. The Jupyter Notebook is available in Github:
   https://github.com/DavidFlanders/Capstone/blob/master/ChicagoCrimeData%20-%20Data%20Wrangling.ipynb

   i. Remove rows containing at least one null value

The total number of cases is over 6.7 million, and the number of cases that has at least one NULL value is about 0.7 mil. Majority of the NULL values happen for the fields 'Ward' and 'Community Area'. Since the latter is only about 10% of the total cases, they are removed from the whole dataset.

ii. Format date

a. Some column names such as 'Community Area' include a space. I replace space with '_'.
b. The fields of 'District', 'Ward', and 'Community_Area' are float64 data. Actually, they are just numeric codes contain no decimal point. I convert them to the int64 format.
c. I changed the fields of 'IUCR', 'Primarhy_Type', and 'FBI_Code' to the categorical variables.
d. I converted the 'Date' from the string format to the date-time format. Since the column 'Year' comes from the 'Date', at this moment, I remove it.

iii. Remove duplicates

The field 'ID' is unique, but the field 'Case_Number' included 378 duplicates. I remove them from the dataset.

iv. Remove outliers

I use the scatter plot to examine 'X_Coordinate' and 'Y_Coordinate', and find 107 outliers. I remove them from the dataset.

v. Save the clean file

The dataset cleaned up gets saved to a file for data analysis.

2. Data Story Telling:

I utilized different plots to identify patterns of the crime cases. The Jupyter Notebook is available in Github: https://github.com/DavidFlanders/Capstone/blob/master/ChicagoCrimeData%20-%20Data%20Storytelling.ipynb

Below is the list of major trends and patterns identified

i. The total number of crime cases decreased over the years from 2003 to 2015. However, the total number of crime cases stays flat between 2015 and 2018.
ii. The number of crime cases drops to the lowest point around January or February each year, and the number of crime cases rises to the highest point around July each year.
iii. The number of cases is the lowest between 4:00 – 6:00am. The peak hours for the largest number of crime cases varies with primary types of crime cases.
iv. The top three primary types of crime cases are THEFT, BATTERY and CRIMINAL DAMAGE.

v. The Community Area #25 has the largest number of crime cases.

3. Inferential Statistics Techniques:
   The inferential statistics for the crime cases were also conducted by joining the 2006-2010 education data *** for Chicago. The Jupyter Notebook is available in Github: https://github.com/DavidFlanders/Capstone/blob/master/ChicagoCrimeData%20-%20InferentialStatistics.ipynb

   i. the Pearson correlation coefficient between the percentage of people over 25 years old having high school degrees or above and the percentage of the number of crime cases over total 25 years old for a Community Area is - 0.194. There is negative correlation between them.
   ii. the Pearson correlation coefficient between the percentage of people over 25 years old having Bachelor degrees or above and the percentage of the number of crime cases over total 25 years old for a Community Area is - 0.361. There is stronger negative correlation between them.
   iii. It seems that the education plays an important role in the number of crime cases.
   iv. [***] The data was downloaded from the below link. It needs to be validated on if it is valid: http://robparal.com/downloads/CDPH/Education%20by%20Race%20by%20Census%20Tract%20and%20Community%20Area.xlsx

IV. To-do List

1. Correlation analysis
   Since I just found the Chicago community area data covering eduction, poverty indicator etc., I would like to conduct further correlation analysis between these data and the numbers of crime cases.

2. Machine Learning Techniques

V. Deliverables

By completion of this project, I will have the following deliverables:

- Python codes for data analysis and machine learning
- A document describe the approach to be adopted
- A slide deck for the whole project and conclusion

VI. Appendix, Description of Chicago Crime Dataset

Column Name	Description	Type
ID	Unique identifier for the record.	Number
Case Number	The Chicago Police Department RD Number (Records Division Number), which is unique to the incident.	Plain Text
Date	Date when the incident occurred. This is sometimes a best estimate.	Date & Time
Block	The partially redacted address where the incident occurred, placing it on the same block as the actual address.	Plain Text
IUCR	The Illinois Uniform Crime Reporting code. This is directly linked to the Primary Type and Description. See the list of IUCR codes at https://data.cityofchicago.org/d/c7ck-438e.	Plain Text
Primary Type	The primary description of the IUCR code.	Plain Text
Description	The secondary description of the IUCR code, a subcategory of the primary description.	Plain Text
Location Description	Description of the location where the incident occurred.	Plain Text
Arrest	Indicates whether an arrest was made.	Checkbox
Domestic	Indicates whether the incident was domestic-related as defined by the Illinois Domestic Violence Act.	Checkbox
Beat	Indicates the beat where the incident occurred. A beat is the smallest police geographic area – each beat has a dedicated police beat car. Three to five beats make up a police sector, and three sectors make up a police district. The Chicago Police Department	Plain Text

	has 22 police districts. See the beats at https://data.cityofchicago.org/d/aerh-rz74.	
District	Indicates the police district where the incident occurred. See the districts at https://data.cityofchicago.org/d/fthy-xz3r.	Plain Text
Ward	The ward (City Council district) where the incident occurred. See the wards at https://data.cityofchicago.org/d/sp34-6z76.	Number
Community Area	Indicates the community area where the incident occurred. Chicago has 77 community areas. See the community areas at https://data.cityofchicago.org/d/cauq-8yn6.	Plain Text
FBI Code	Indicates the crime classification as outlined in the FBI's National Incident-Based Reporting System (NIBRS). See the Chicago Police Department listing of these classifications at http://gis.chicagopolice.org/clearmap_crime_sums/crime_types.html.	Plain Text
X Coordinate	The x coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection. This location is shifted from the actual location for partial redaction but falls on the same block.	Number
Y Coordinate	The y coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection. This location is shifted from the actual location for partial redaction but falls on the same block.	Number
Year	Year the incident occurred.	Number
Updated On	Date and time the record was last updated.	Date & Time
Latitude	The latitude of the location where the incident occurred. This location is shifted from the actual location for partial redaction but falls on the same block.	Number

Longitude	The longitude of the location where the incident occurred. This location is shifted from the actual location for partial redaction but falls on the same block.	Number
Location	The location where the incident occurred in a format that allows for creation of maps and other geographic operations on this data portal. This location is shifted from the actual location for partial redaction but falls on the same block.	