Steven Yuan-Yao Chang, Michael Chon, Jianan Su
Group E
Website: https://sites.google.com/view/crimeanalysis501

Introduction

In Sociology, deviance is an unfortunate but inevitable part of all healthy functioning societies. It serves many crucial social functions to constituents but its most important social function is to clarify moral boundary between what is right and wrong. That social boundary is often reinforced by a form of deviance, crime. Crime cannot be avoided and stopped because it is impossible to stop every individual from committing a crime in society. However, big data can suggest numerous ways for police force to handle more efficiently.

Whether it is in big cities or counties, crime has still been a major issue in the United States. Baltimore, Houston, and Pittsburgh are the top 3 dangerous cities in the United States that the combined violent crime rate is 279.39 per 3000, meaning that 1 out 11 will most likely end up being a victim in these major cities[1]. In 2015, approximately, 1,200,000 violent crimes occurred across the nation, which is an increase of 3.9 percent from the 2014 report[2]. In District of Columbia, annual crime totals reported from 2007 to 2014 gradually increase from 35,706 to 40,838, with an increase of 14.37%, including crime types such as homicide, forcible rape, burglary, etc[3]. Throughout the United States, the total number of incidents involving firearms in 2016 is 53,163. Out of 53,163 incidents, there have been 13,743 deaths and 28,302 injuries[4].

In order to handle crime more efficiently, it is important for police force to analyze historical crime data. Police force can use statistical and data science techniques to extract hidden patterns in crime behaviors. The patterns will help police force understand different types of crime and where lots of crime happened in past years. Based on the patterns obtained from analyzing the data, police force can implement prevention work, which will significantly reduce the number of incidents of crime and reduce the harms caused by crime.

Data Science Question

The data science question regarding this matter is “how can police force handle crime more efficiently?” This question is too general to answer because there are many different roles that big data can help police force in handling crime more efficiently. In order to answer this big data science question, there must be smaller questions that can eventually lead to answering

---

[1] https://www.neighborhoodscout.com/neighborhoods/crime-rates/25-most-dangerous-neighborhoods/

[2] https://ucr.fbi.gov/crime-in-the-u.s/2015/crime-in-the-u.s.-2015/offenses-known-to-law-enforcement/violent-crime

[3] http://mpdc.dc.gov/page/crime-statistics-citywide

[4] http://www.gunviolencearchive.org/

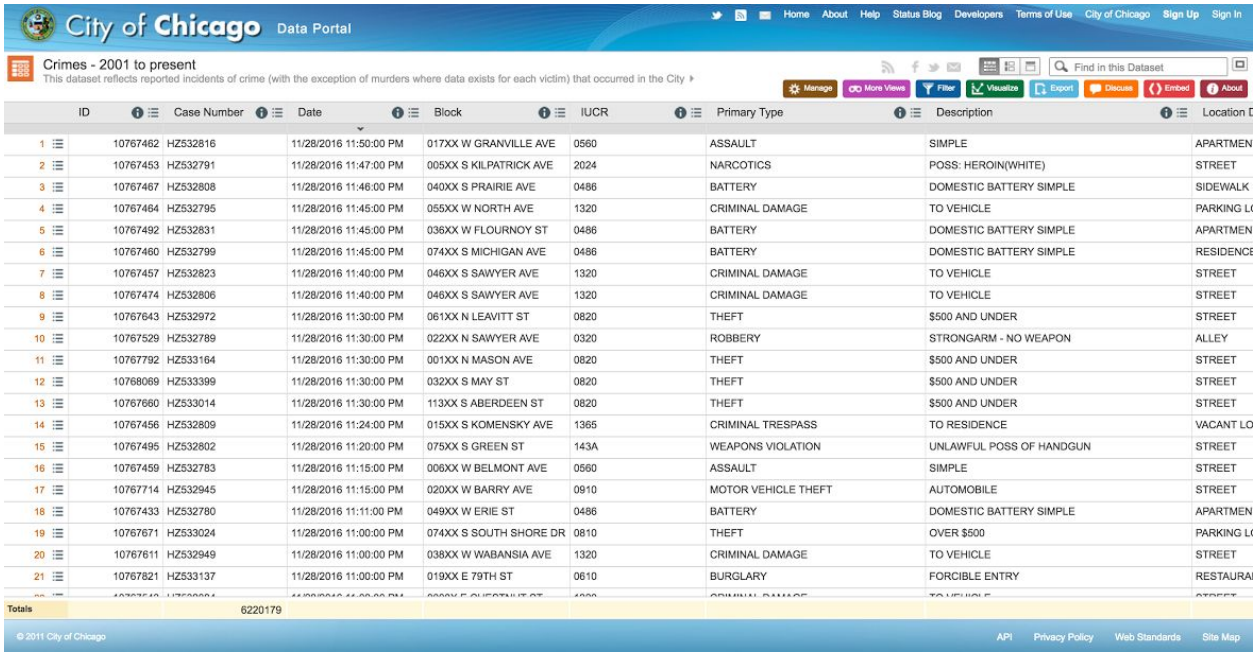
the question. The first smaller question is “in which parts of cities, counties, or states do crime happen the most?” This question will guide to focus specifically on the areas that are most vulnerable to crime. The second smaller question is “on what day of week do crime happen the most in the areas?” This second question is followed by the third smaller question; “in what time period do crime happen the most in the areas?” These two questions will give a general idea of the frequency of crime in the crime hotspots. The fourth smaller question is “what types of crime is prevalent in the areas?” Answering the four smaller questions can eventually lead to answering the data science question.

Datasets

The datasets are from two distinct locations, the city of Chicago of Illinois and Montgomery County of Maryland. The city of Chicago[5] provides a dataset that reflects reported incidents of crime that occurred in the city from 2001 to present minus the most recent seven days. From the dataset, lots of valuable attributes regarding each crime incident are provided. Such attributes are date, primary type (crime type), district, community area, description, location description, latitude, and longitude. The dataset is very well maintained and organized by the city of Chicago; the data does not contain any noise . However, the dataset has a few issues. The first issue is that some attributes such as ID and FBI Code are not necessary and useful for analysis. The second issue is that there are some missing values for x coordinate, y coordinate, latitude, longitude, and location. The dataset does not contain those values for certain crime incident on purpose because of privacy concern for victims. The dataset is

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

double-checked that there is no relationship between the incidents missing latitudes and longitudes and types of crime.



City of Chicago Data Portal

Home About Help Status Blog Developers Terms of Use City of Chicago Sign Up Sign In

Crimes - 2001 to present

This dataset reflects reported incidents of crime (with the exception of murders where data exists for each victim) that occurred in the City

Find in this Dataset

Manage More Views Filter Visualize Export Discuss Embed About

	ID	Case Number	Date	Block	IUCR	Primary Type	Description	Location D
1	10767462	HZ532816	11/28/2016 11:50:00 PM	017XX W GRANVILLE AVE	0560	ASSAULT	SIMPLE	APARTMEN
2	10767453	HZ532791	11/28/2016 11:47:00 PM	005XX S KILPATRICK AVE	2024	NARCOTICS	POSS: HEROIN(WHITE)	STREET
3	10767467	HZ532808	11/28/2016 11:46:00 PM	040XX S PRAIRIE AVE	0486	BATTERY	DOMESTIC BATTERY SIMPLE	SIDEWALK
4	10767464	HZ532795	11/28/2016 11:45:00 PM	055XX W NORTH AVE	1320	CRIMINAL DAMAGE	TO VEHICLE	PARKING L
5	10767492	HZ532831	11/28/2016 11:45:00 PM	036XX W FLOURNOY ST	0486	BATTERY	DOMESTIC BATTERY SIMPLE	APARTMEN
6	10767460	HZ532799	11/28/2016 11:45:00 PM	074XX S MICHIGAN AVE	0486	BATTERY	DOMESTIC BATTERY SIMPLE	RESIDENCE
7	10767457	HZ532823	11/28/2016 11:40:00 PM	046XX S SAWYER AVE	1320	CRIMINAL DAMAGE	TO VEHICLE	STREET
8	10767474	HZ532806	11/28/2016 11:40:00 PM	046XX S SAWYER AVE	1320	CRIMINAL DAMAGE	TO VEHICLE	STREET
9	10767643	HZ532972	11/28/2016 11:30:00 PM	061XX N LEAVITT ST	0820	THEFT	$500 AND UNDER	STREET
10	10767529	HZ532789	11/28/2016 11:30:00 PM	022XX N SAWYER AVE	0320	ROBBERY	STRONGARM - NO WEAPON	ALLEY
11	10767792	HZ533164	11/28/2016 11:30:00 PM	001XX N MASON AVE	0820	THEFT	$500 AND UNDER	STREET
12	10768069	HZ533399	11/28/2016 11:30:00 PM	032XX S MAY ST	0820	THEFT	$500 AND UNDER	STREET
13	10767660	HZ533014	11/28/2016 11:30:00 PM	113XX S ABERDEEN ST	0820	THEFT	$500 AND UNDER	STREET
14	10767456	HZ532809	11/28/2016 11:24:00 PM	015XX S KOMENSKY AVE	1365	CRIMINAL TRESPASS	TO RESIDENCE	VACANT LO
15	10767495	HZ532802	11/28/2016 11:20:00 PM	075XX S GREEN ST	143A	WEAPONS VIOLATION	UNLAWFUL POSS OF HANDGUN	STREET
16	10767459	HZ532783	11/28/2016 11:15:00 PM	006XX W BELMONT AVE	0560	ASSAULT	SIMPLE	STREET
17	10767714	HZ532945	11/28/2016 11:15:00 PM	020XX W BARRY AVE	0910	MOTOR VEHICLE THEFT	AUTOMOBILE	STREET
18	10767433	HZ532780	11/28/2016 11:11:00 PM	049XX W ERIE ST	0486	BATTERY	DOMESTIC BATTERY SIMPLE	APARTMEN
19	10767671	HZ533024	11/28/2016 11:00:00 PM	074XX S SOUTH SHORE DR	0810	THEFT	OVER $500	PARKING L
20	10767611	HZ532949	11/28/2016 11:00:00 PM	038XX W WABANSIA AVE	1320	CRIMINAL DAMAGE	TO VEHICLE	STREET
21	10767821	HZ533137	11/28/2016 11:00:00 PM	019XX E 79TH ST	0610	BURGLARY	FORCIBLE ENTRY	RESTAURA
Totals		6220179						

© 2011 City of Chicago API Privacy Policy Web Standards Site Map

Another dataset is from Montgomery county of Maryland[6]. The county provides the public with direct access to crime statistic databases. The dataset contains the information regarding all founded crimes reported after July 2013. The important attributes of the dataset are class, class description, police district name, block address, city, zip code, place, start date/time, end date/time, and location. The attributes are given per each crime incident happened in Montgomery County. This dataset is also very clean and well maintained by Montgomery county; it does not contain any noise. However, it has a few issues just like the dataset of the city of Chicago has. The first issue is that there are some unnecessary attributes such as Beat, PRA, Incident ID, etc. that are not useful for analysis. The second issue is that there are missing values in zip code (the city of Chicago does not fully disclose zipcode), end date/time, latitude, and longitude. Some end dates/times are missing because Montgomery County police might have not caught criminals yet. Missing latitudes and longitudes are done on purpose to protect the victims' privacy. Also, the dataset is thoroughly examined that there is no relationship between missing latitudes and longitudes and types of crime occurred.

[6] https://data.montgomerycountymd.gov/Public-Safety/Crime/icn6-v9z3/data

dataMontgomery

Search

Data Catalog | Suggest a Dataset | Open Budget | spendingMontgomery | CountyStat | Videos and Resources | About | Sign In

Crime

Updated daily postings on Montgomery County's open data website, dataMontgomery, provide the public with direct access to crime statistic

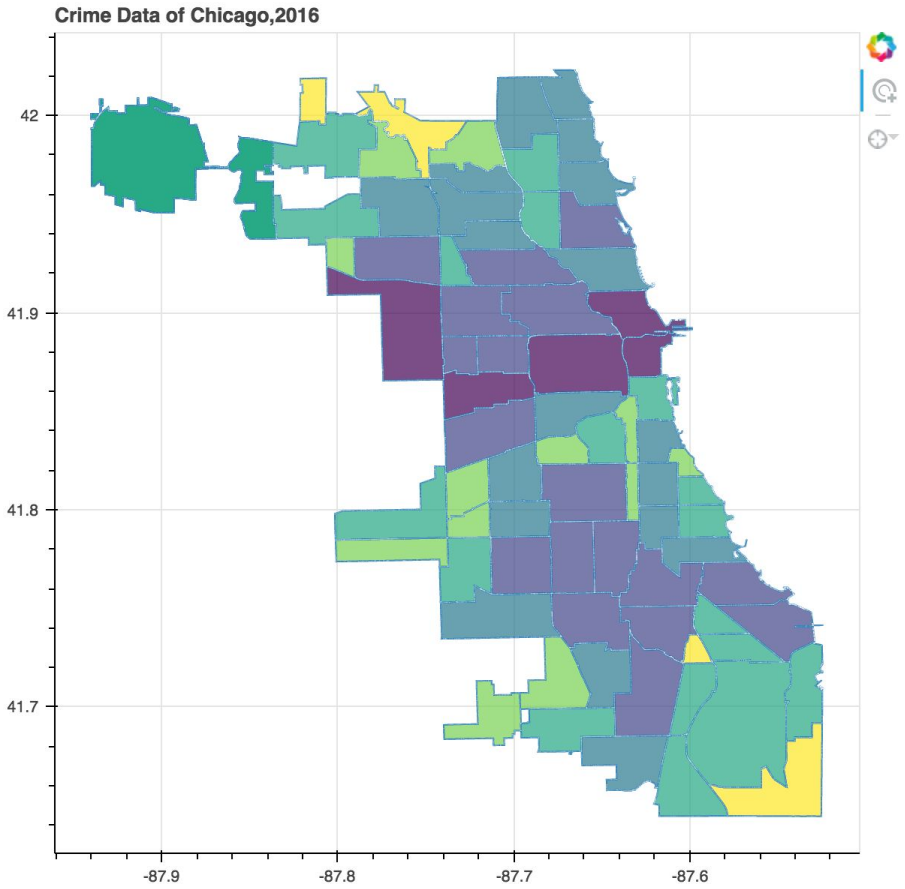
Find in this Dataset

Manage | More Views | Filter | Visualize | Export | Discuss | Embed | About

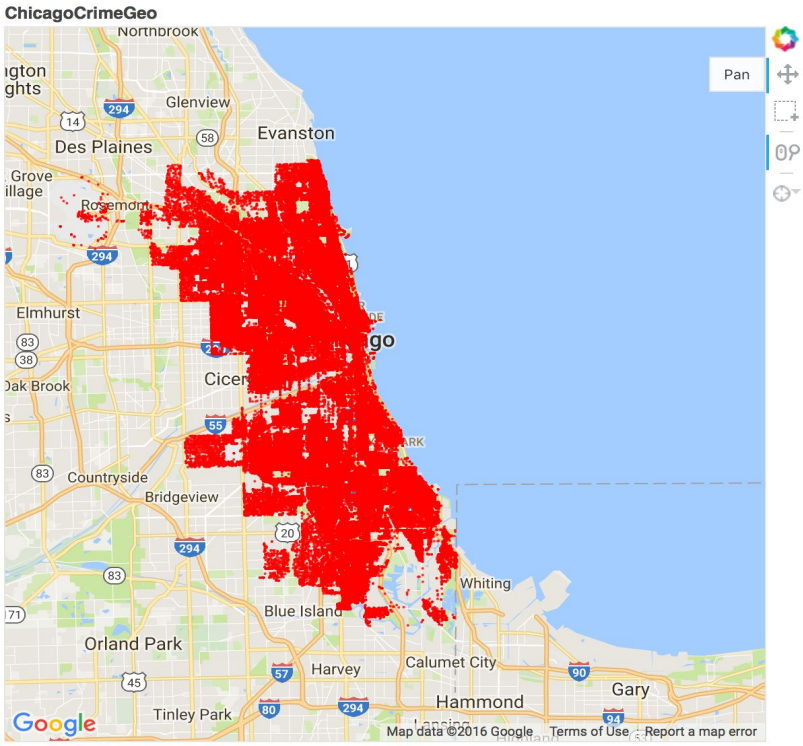
	Incident ID	CR Number	Dispatch Date / Time	Class	Class Description	Police District Name	Block Address	City	State	Zip Code	Agency	Place
1	201110344	16062320	12/04/2016 08:46:18 PM	0813	ASSAULT & BATTERY SPOUSE/PARTNER	GERMANTOWN	13400 CLOVERDALE PL	GERMANTOWN	MD	20874	MCPD	Residence - Single Family
2	201110311	16062308	12/04/2016 06:48:56 PM	2946	RECOVERED PROPERTY/MONT. CO.	GERMANTOWN	20000 AIRCRAFT DR	GERMANTOWN	MD	20874	MCPD	Government Building
3	201110320	16062312	12/04/2016 06:33:49 PM	1412	VANDALISM-MOTOR VEHICLE	GERMANTOWN	19300 CIRCLE GATE DR	GERMANTOWN	MD	20874	MCPD	Parking Lot - Residential
4	201110341	16062325	12/04/2016 06:31:52 PM	0811	ASSAULT & BATTERY - CITIZEN	BETHESDA	5800 NICHOLSON LN	ROCKVILLE	MD	20852	MCPD	Residence - Apartment/Condo
5	201110319	16062306	12/04/2016 06:18:54 PM	0811	ASSAULT & BATTERY - CITIZEN	GERMANTOWN	20000 CENTURY BLVD	GERMANTOWN	MD	20874	MCPD	Street - In vehicle
6	201110316	16062303	12/04/2016 05:47:07 PM	0513	BURG FORCE-RES/TIME UNK	BETHESDA	5700 GROSVENOR LN	BETHESDA	MD	20814	MCPD	Residence - Single Family
7	201110312	16062300	12/04/2016 05:42:45 PM	0619	LARCENY OTHER OVER $200	BETHESDA	10500 WESTLAKE DR	BETHESDA	MD	20817	MCPD	School/College
8	201110328	16062289	12/04/2016 04:21:33 PM	2942	MENTAL TRANSPORT	GERMANTOWN	7600 DAMASCUS RD	GAITHERSBURG	MD	20882	MCPD	Residence - Single Family
9	201110307	16062283	12/04/2016 03:35:58 PM	1834	CDS-POSS MARIJUANA/HASHISH	WHEATON	11500 GEORGIA AVE	SILVER SPRING	MD	20902	MCPD	Street - In vehicle
10	201110317	16062281	12/04/2016 03:17:26 PM	0619	LARCENY OTHER OVER $200	BETHESDA	5400 WESTBARD AVE	BETHESDA	MD	20816	MCPD	Grocery/Supermarket
11	201110300	16062282	12/04/2016 03:04:01 PM	0614	LARCENY FROM AUTO OVER $200	GERMANTOWN	25500 JOY LN	DAMASCUS	MD	20872	MCPD	Residence - Driveway
12	201110298	16062280	12/04/2016 03:01:12 PM	1834	CDS-POSS MARIJUANA/HASHISH	BETHESDA	4900 STRATHMORE AVE	KENSINGTON	MD	20895	MCPD	Street - In vehicle
13	201110295	16062277	12/04/2016 02:33:02 PM	0629	LARCENY OTHER $50 - $199	BETHESDA	6000 COREWOOD LN	BETHESDA	MD	20816	MCPD	Residence - Yard
14	201110302	16062275	12/04/2016 02:26:25 PM	0334	ROB OTHER WPN CONV. STORE	GERMANTOWN	11500 MIDDLEBROOK RD	GERMANTOWN	MD	20876	MCPD	Convenience Store
15	201110294	16062276	12/04/2016 02:07:11 PM	1412	VANDALISM-MOTOR VEHICLE	ROCKVILLE	100 WATKINS POND BLVD	ROCKVILLE	MD	20850	RCPD	Street - Residential
16	201110293	16062273	12/04/2016 01:53:35 PM	2737	TRESPASSING	ROCKVILLE	600 HUNGERFORD DR	ROCKVILLE	MD	20850	RCPD	Grocery/Supermarket
17	201110277	16062260	12/04/2016 12:39:33 PM	0811	ASSAULT & BATTERY - CITIZEN	GERMANTOWN	19900 SWEETGUM CIR	GERMANTOWN	MD	20874	MCPD	Residence - Apartment/Condo
18	201110272	16062257	12/04/2016 12:34:10 PM	0617	LARCENY FROM BUILDING OVER $200	GERMANTOWN	19900 SWEETGUM CIR	GERMANTOWN	MD	20874	MCPD	Residence - Apartment/Condo
19	201110274	16062264	12/04/2016 12:24:36 PM	0634	LARCENY FROM AUTO UNDER $50	MONTGOMERY VILLAGE	17700 MEADOW VISTA WAY	GAITHERSBURG	MD	20877	MCPD	Residence - Driveway
20	201110273	16062253	12/04/2016 12:23:56 PM	1011	FORGERY/CNTRFT-CRDT CARDS	MONTGOMERY VILLAGE	1 FULKS CORNER AVE	GAITHERSBURG	MD	20877	GPD	Retail - Other

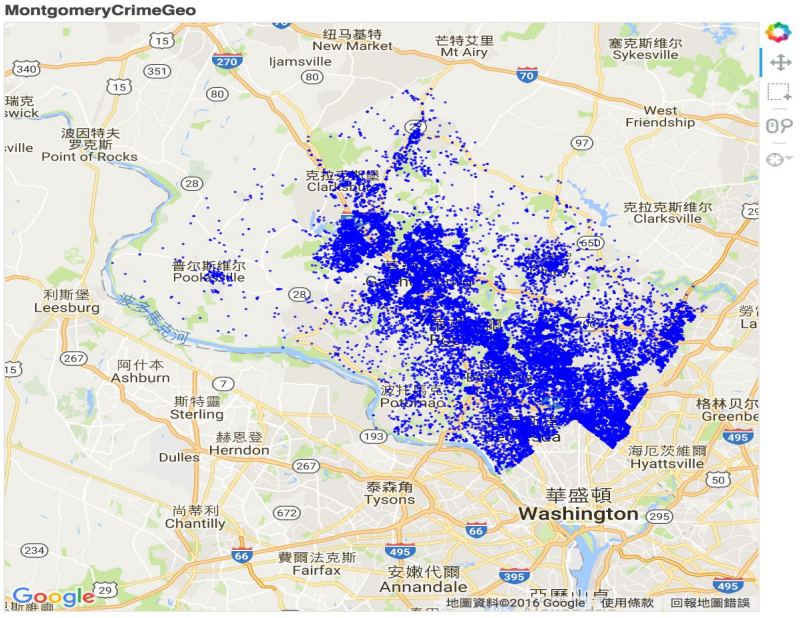
Privacy Policy | User Rights | Disclaimer | Accessibility | Terms of Use | © 2016 Montgomery County Government | Powered By Socrata

Visuals (All the links are on the website)

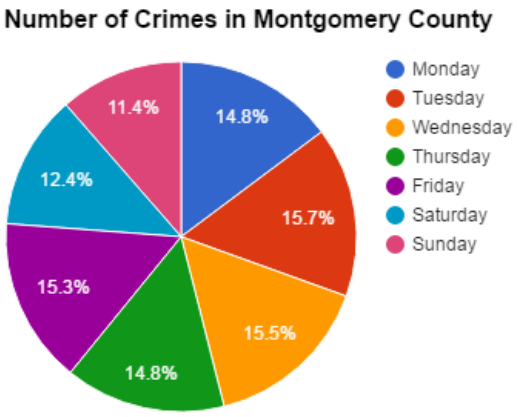


This visual is the map of crime data of Chicago done using bokeh module of python. The shades of color on the visual represents different crime rate. The darker the color in a district is, the higher number of crime is in that district. Based on the visual, the communities 29, 25, 28, 32, and 8 are the most dangerous communities in Chicago. The communities with lower crime rate are 9, 12, 47, and 55.

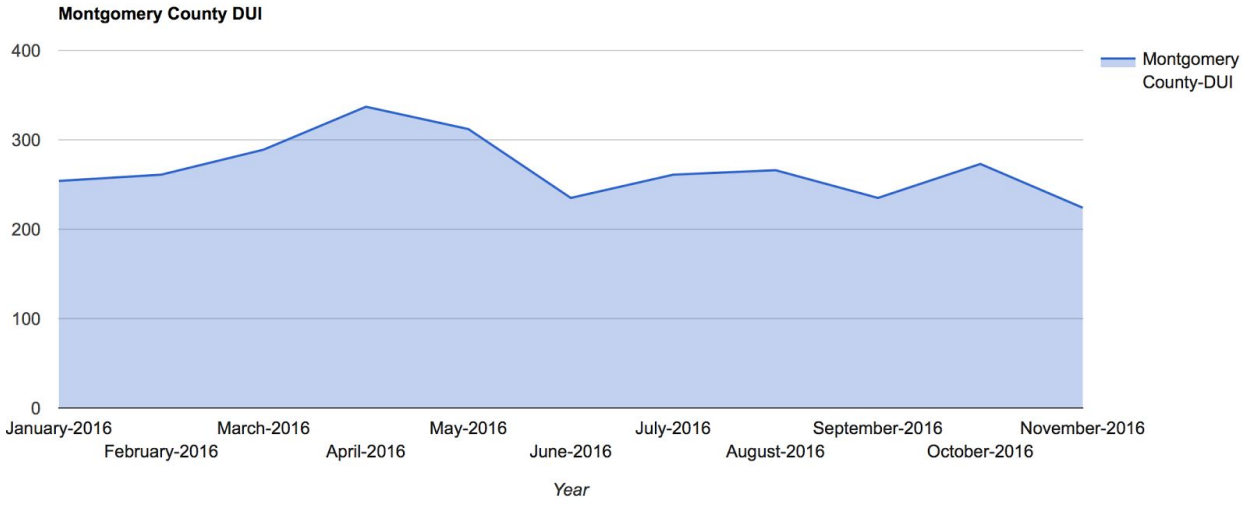




These visuals are done using the same method, bokeh module of python. The visuals show the density of crime happened in each location. They can be enlarged and used to tell which block is safer than the others. Each colored dot in the visuals shows crime type, geolocation, days of week, date, and time.

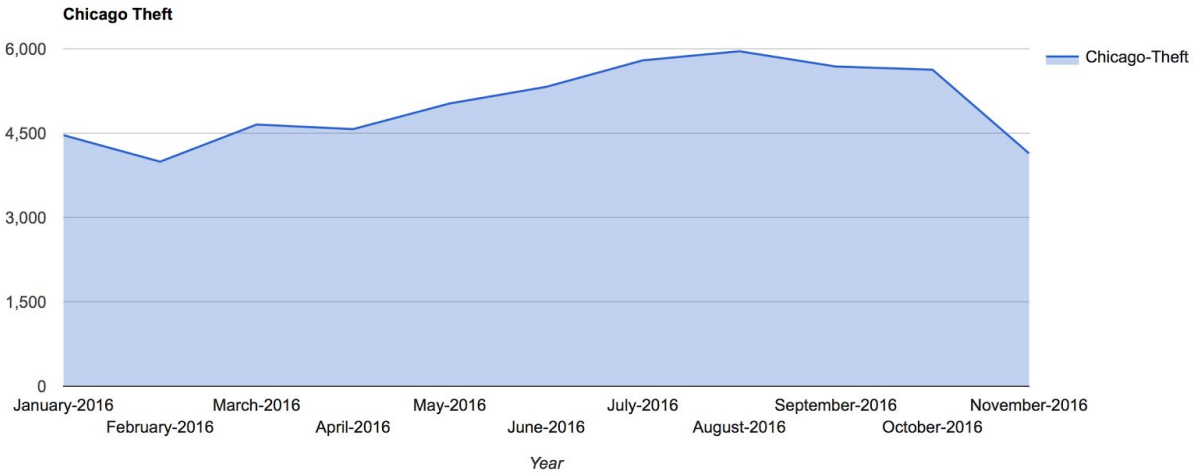


The pie chart is done using Google Chart. It represents the number of crimes occurred on days of week in Montgomery County. Interestingly, the day on which crime happened the most is Tuesday, Not Friday.



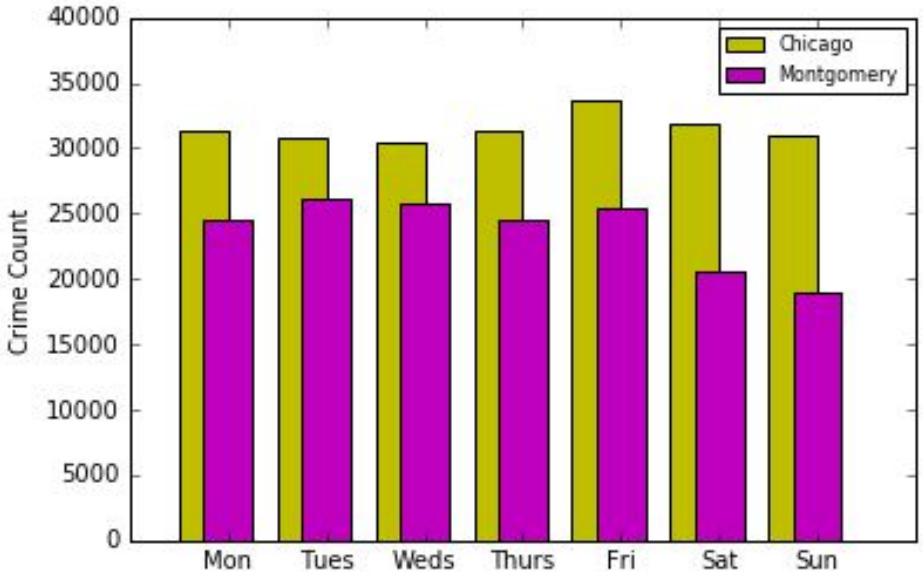
The area chart is done using Google Chart. From analyzing the dataset, the most common type of crime in Montgomery County of Maryland is DUI, driving under the influence.

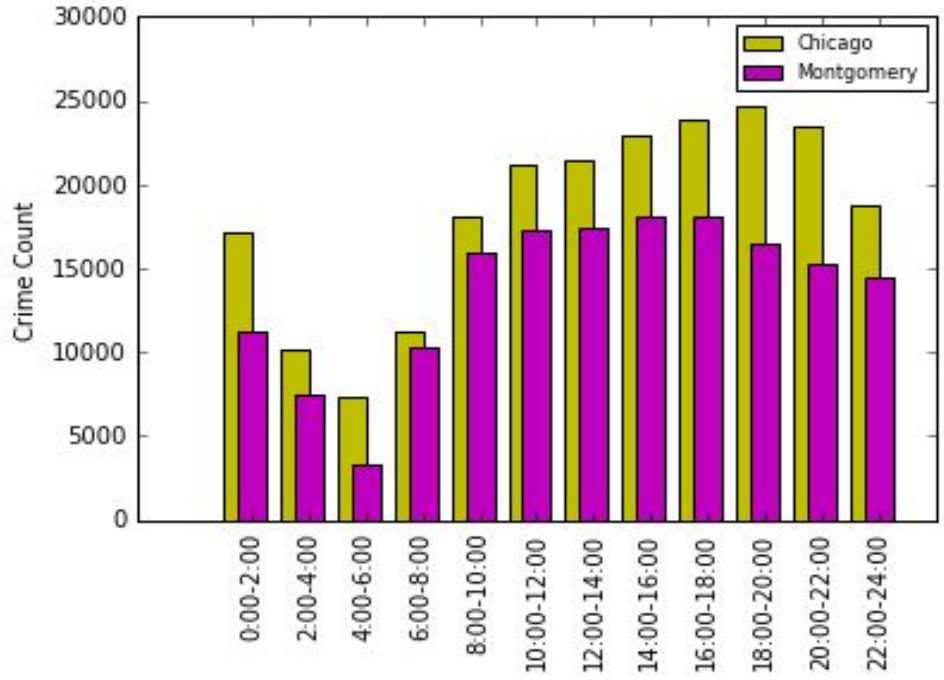
To investigate this a little bit more, the area chart is built on the numbers of occurrence of DUI on each month in 2016. The highest number of occurrence of DUI is on April 2016 because it can be assumed that students are on spring break from school and most likely end up consuming alcoholic beverages and drive without any parents' permission.



This area chart is done by using Google Chart. From analyzing the dataset, the most common type of crime is found to be theft. To understand more about the relationship between time and the number of theft happened, the area chart is used. According to the chart, the highest number of occurrence of theft is August 2016. It is because people tend to be outside when the weather is warm especially the month like August and this fact makes the people outside more vulnerable to theft.

<u>Comparison using visuals</u>





These two histograms are to compare the city of Chicago to Montgomery County. The first histogram is based on the number of crime on days of week for the two places. The second histogram is based on the time period of a day for the two places. The first histogram states that there are more crimes in Chicago than Montgomery County. In Chicago, friday is the most dangerous day than the other weekdays because people in the city have more active nightlife on friday, which make vulnerable to crime. In Montgomery County, Tuesday is the most dangerous day. For the dataset of Montgomery County, the number for Tuesday is not that much higher compared to the numbers for other days.

Based on the second histogram, the majority of crime occurred between 14:00 to 16:00 and 16:00 to 18:00 in Montgomery County. However, in Chicago, the majority of crime occurred from 18:00 to 20:00 followed by 16:00 to 18:00. It is because people who live in the city are more active in the night than the people who lived in the county. What this means there can be more chances of crime happening. The similarity is that a fewer crimes happened between 4:00 and 6:00 followed by 2:00 and 4:00 for the both datasets because during those time periods, people are asleep, which means a fewer crimes happened.

Conclusion

Based on the four smaller questions and the visuals created, there are several ways for police force to handle crime more efficiently. The first way is to increase the number of police patrols on the days/times the crimes happened the most. For example, Chicago Police Department can increase the number of police patrols on Friday, that is found to be the most

dangerous day of a week. For Montgomery County, Montgomery County Police can increase the number of patrols on Tuesday. The second way is to specifically train police officers to handle the crimes that are prevalent in the areas. For Chicago, CPD can train police officers specifically to handle theft. For Montgomery County, police officers are assigned to the areas where lots of bars are located so that they can prevent drunk people from driving. The third way is to install surveillance cameras at crime hotspots. Based on the maps of crime data for the two places, CPD can install surveillance cameras at the corners of streets in Austin, Chicago where the highest number of theft is reported. The Montgomery County Police can install more cameras in Silver Spring, MD, where lots of people are caught DUI. The fourth way is to notify the local residents about the crime that will most likely happen in their neighborhoods. The fifth way is to teach the local residents basic lessons on how to protect themselves from the crime that occurs the most in their neighborhoods. The police force can implement the last two methods to teach the local residents when to be careful and how to protect themselves just in case the police force cannot respond immediately.

Limitations

The hypotheses to predict a next crime in certain areas did not work well. For the first hypothesis “a theft will most likely happen in community 23 from 18:00 to 20:00 on friday in Chicago”, decision tree, KNN, Naive Bayes, SVM, and Random Forest did not work because of the probabilities of fitting these machine learning algorithms to the data were around .17, which is really low and means that these predictive analysis did not work. The second hypothesis “the time that crime most likely happens in chicago is same as Montgomery County” is tested using t.test function in python but the p-value of the test is 1.599e-08, which is very low. This method did not support the hypothesis. The third hypothesis, “driving under influence will most likely happen in Silver Spring from 16:00 to 18:00 on Tuesday in Montgomery County”, is tested by logistical regression and the result is .01037, which is extremely low. That means, the hypothesis was not supported. The fourth hypothesis “the day of week that crime most likely happens in Chicago is same as the Montgomery County” is examined by t.test function and the p-value is 8.0979e^-192, which is extremely low. The fourth hypothesis was not supported by the method.

The research article “to predict and serve?” by Kristina Lum and William Isaac explained bias within police-recorded data. If police focuses on certain races, police records over-represent the certain racial groups. What this means is that crimes that occur in the same locations are more likely to appear in the database. It will lead to bias that only represent the certain ethnic groups in the dataset, not representing crime throughout a place. According to the article, bias in police records is depended on the desired amount of local policing; that is, the types of crime will vary from place to place and from one ethnic group to another. Also, in the article, local police departments uses Tay, which is Microsoft automated chatbot built on machine learning algorithms and using social media such as twitter, to predict a crime. However, it fails to predict because if Tay is bombarded with negative tweets, then Tay uses the negative tweets as data corpus to predict a crime and the crime will not be accurate.

Our hypotheses using machine learning algorithms were to predict where a next crime will happen and the type of the next crime at a certain time period but they did not work. The reason is explained by this quote from the same article, "The data is collected as a by-product of police activity, predictions made on the basis of patterns learned from this data do not pertain to future instances of crime on the whole." This explains that the data collected by police can be full of bias and is not a trustworthy source to be used as to train data and predict a crime. Also, it is impossible to know that these two datasets are bias-free.

References

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

https://data.montgomerycountymd.gov/Public-Safety/Crime/icn6-v9z3/data

http://onlinelibrary.wiley.com/doi/10.1111/j.1740-9713.2016.00960.x/full

https://www.cityofchicago.org/content/dam/city/depts/doit/general/GIS/Chicago_Maps/Citywide_Maps/Community_Areas_W_Numbers.pdf