Crime In Chicago

Adrian Aguada (aaguad3@uic.edu) Github: aaguad3
Xuhui Wu (xwu83@uic.edu) Github: xwu83
Aditya Guda (aguda4@uic.edu) Github: aguda4

https://github.com/orgs/CS418/teams/data-cyclists

The Idea

- We are analyzing Chicago crime data to correlate them with certain attributes or patterns
- We feel this is important because of the national perception of the city being seen as a “dangerous, crime filled” city. We want to confirm if this notion is true, true in a different light, or completely false.
- We collectively decided upon this question because UIC’s location is relevant to the question here.
- Our hypothesis is “We believe that the crime in Chicago is overstated for how it is perceived”.

The Data

- We intend to use datas of Crime descriptions, Demographic, level of economic inequality.
- We can get those data from Chicago police department, census bureau and Survey of Income and Program Participation(SIPP).
- We will collect about 200 lines of data for each type of data, which we think is suitable for code analysis and not too large.
- We plan on using the data Crimes_-_1_year_prior_to_present and Crimes_-_2021.csv that we found on data.cityofchicago, data.world to view crime descriptions in Chicago.

Crimes - One year prior to present:

https://data.cityofchicago.org/Public-Safety/Crimes-One-year-prior-to-present/x2n5-8w5q

Crimes - 2022:

https://data.cityofchicago.org/Public-Safety/Crimes-2022/9hwr-2zxp

Export those two, 2021 one is too big so I change to 2022 one, but still have 40mb. And I thought we just need one of those two.

Data Cleaning

show clearly how you cleaned your data. Show the original data size and the final size. For numerical data, use boxplots to identify and remove outliers.

Our data was cleaned by removing irrelevant columns to the data we wanted to visualize and hypothesis we wanted to prove.

We had 23 columns from the original data set we downloaded. The original filesize was 14 MB.

1		ID	Case Num	Date	Block	IUCR	Primary Ty	Descriptic	Location D	Arrest	Domestic	Beat	District	Ward	Communi	FBI Code	X Coordin	Y Coordin	Year	Updated (	Latitude	Longitude	Location
2	4506608	9878952	HX529642	########	010XX E 47	497	BATTERY	AGGRAVA	APARTME	FALSE	TRUE	222	2	4	39	04B	1183896	1874058	2014	########	41.8096	-87.601	(41.809597, -87.601016)

After cleaning, we reduced it to only 11 columns.

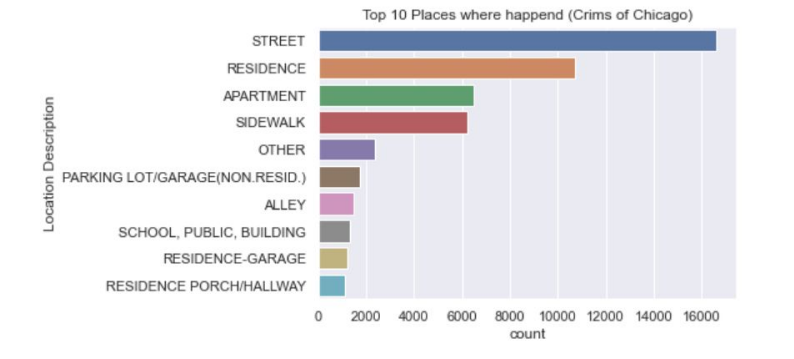
Out[16]:

	ID	Case Number	Date	Block	Primary Type	Description	Location Description	Arrest	Domestic	Beat	Year

Exploratory Data Analysis

You should try to find correlations between the different attributes. Include any interesting issues or preliminary conclusions you have about your data. Use visualizations.

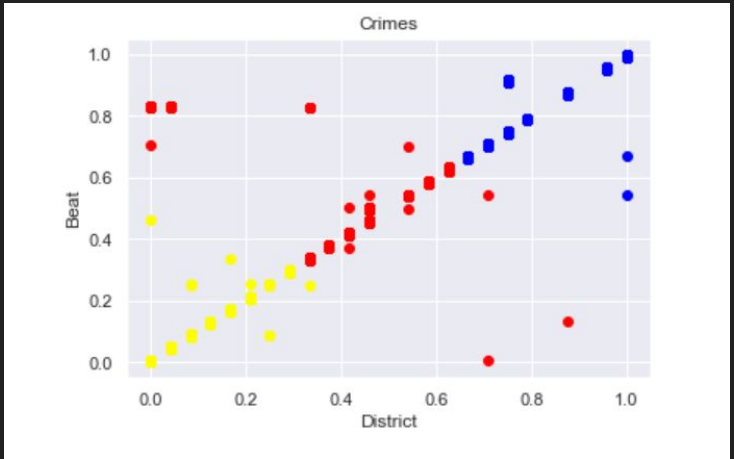
We explored the relationship between crimes and crime location. What we found was most crimes happened on the street. We believed crime happened in public more than in private places like homes. This shows us that domestic crimes are not the most common type of crime. These are crimes where there is no relationship between the victim and assaulter.



Linear Regression

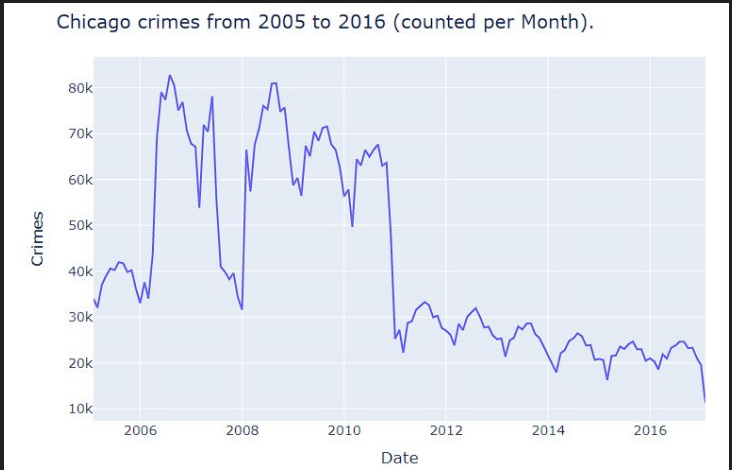
In one slide, List all the data science questions that were you trying to answer and the corresponding Machine Learning technique that you used to answer such questions. For each ML technique, explain what you did, and talk about the inferences you uncovered. An evaluation that shows whether your solution worked well or not. Use visualizations. (see the requirements in the Final Project section)

We used a linear regression model on a 1:1 scale to map the crimes from Districts to Beats. Police Districts are areas of Chicago that are under control by Policing. Beats are the specific patrolling locations for police officers on duty. We analyzed the relationship between the two to predict if heavier patrolling was done by District count. The regression line confirms this below



Time Series Analysis

Last, we have a time series analysis of the crimes by date. We compared the years column of the dataframe to the amount of crimes and we see seasonal trends from 2005-2016 by month. There was always a dip midway through the year. This would be around the summer time in Chicago where people are out more, therefore policing is heavier in the cities.



K means Clustering

Introduction: I used a data that is used in the group project to implement a ML method that is K-means Clustering and I plotted a scatter plot by using the dataset of 2 columns and those are “District” and ‘Beat”.

Explanation: First I imported a few libraries form Sklearn and pandas. I read the csv file by using file and print it by using pandas. Them I used the KMeans function as well as implemented the number of clusters which creates the number of clusters and objects. Then I use the fit_predict function the dataset that is excluding the columns which contains of strings and float data points and this function gives an array. Now I added the array column to the data set. Now to group the clusters or all the values in one cluster, then I used the MinMaxScaler function to implement the 2 columns which I used to to plot the graph. The columns are district and beat(the number of officers that are patrolled). I assigned a column named plotter to the dataset. Then after modifying the values then we have to follow the same process as we did before clustering. I used the clustere_centers_ function to find the centers or the centroids of the clustered values then at the end i labeled the graph and plotted the graph.

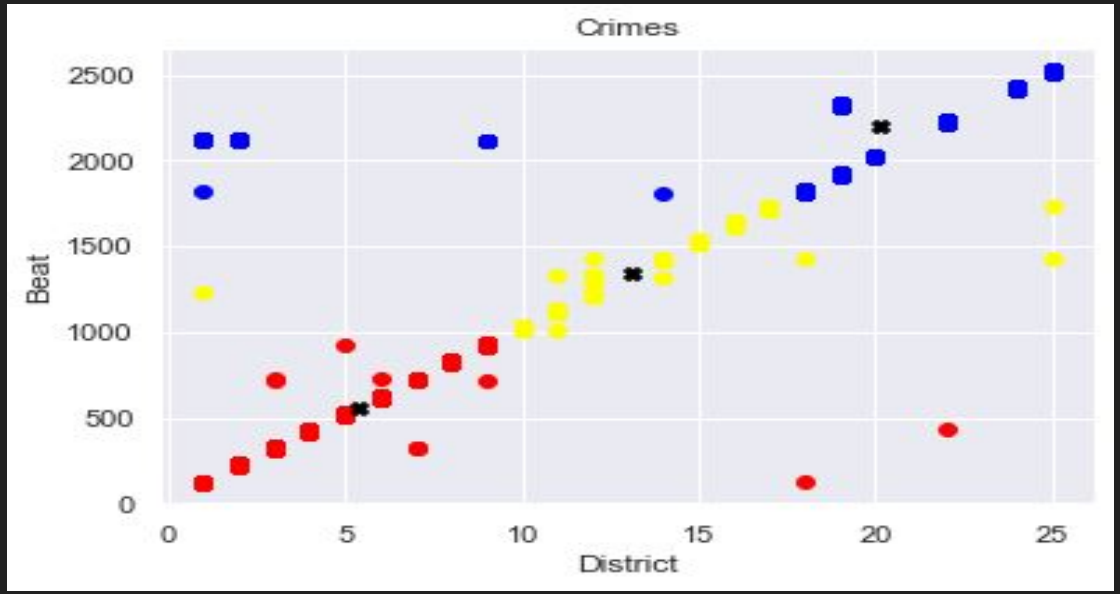
Inferences: After doing this ML method, I realized that the k-means clustering algorithm does not work with very high and low values, string or float data points. So I have to plot the the graph by using the columns sets such as “District” amd “Beats” and there are few columns anyone can choose from the data set. Also, I found that the values of the data is gathered based on the maximum and the minimum order of the value.

Evaluation: I believe that I’m able to do the k-means algorithm graph so I would rate my program good and I can proof it that my code is working since on the next slide there is a plot which is based on this method.

Visualization: From the scatter plot I can analyze a visualization where the scatter plot represents the number of police that are patrolling on a particular distinct and this is useful since this graph gives people a understanding of where the police officers are assigned to a particular spot.

K-means Clustering plot

K-means Clustering plot based on the dataset of the columns “District” and “Beat”.



Data science questions based on K-means Clustering

Data science questions:

Q1. Is there any other method to do the k-mean clustering algorithm other than using the K-means function and the predict function? How to use the predict and fit functions on python?

Q2. How are the centered points determined from the dataset by using the km.cluster_centers_ function? Is there a function to check the centered points?

Q3 . Are there any evaluation methods that can be used to compute the K-means algorithm to find the k-means of a particular column from the data set?

Q4. How to use the predict function to print the values of a single column in a particular column to make sure that all the values remains in the graph?

Lesson Learned

Discuss the main takeaways from your project

Our project taught us many things about Chicago crime. What we first believed was that Chicago was one of the more dangerous cities in the world. We wanted to see if there was truth to that notion. By the ML techniques we used, we saw that volume and policing was a key factor because Chicago is very big, and also selective with where and when they task their police force.

We analyzed the districts/beats such as in our Linear Regression model to show there was a direct correlation where we could predict the two.

We looked at Time-Series to show when crimes were committed.

We also looked at K-Means Clustering to visualize the most common places of crime.

Teamwork

We assigned a weekly “scrum master” for our group to take charge of the project and assign tasks for the other team members. The scrum master would usually go to office hours and ask questions the group had. They would also coordinate the group members’ schedules to make sure we met and talked about what progress we made.

We divided the work into each of us getting an individual ML technique to work on. We all cleaned the data and worked on the exploratory analysis together.

References

List all the resources you have used so far.

https://towardsdatascience.com/machine-learning-algorithms-part-9-k-means-example-in-python-f2ad05ed5203

https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a