Foundations of Data Science

Crimes in NYC

Himanshu Payal - hp2427
Rahul Ramaswamy - rr4059
Rajat Raghuwanshi - rrr9293

1. Problem, motivations and outcome :

We aim to study the crime in New York to gauge the safety of various neighborhoods and how it evolves with time.

Our problem statement is to analyze and use data for complaints made to the NYPD by employing models to forecast crime rates in New York City as well as apply an unsupervised clustering algorithm to identify hotspots of criminal activity.

Motivation :

- Guiding tourists/visitors throughout the city:

  People who want to explore the city can use the app to understand which places should be relatively safe according to the crime rate at that location and time.

- Providing insights to the law enforcement agencies :

  Help law and enforcement agencies deploy forces based on the results. This can be used as a tool to find out which neighborhood is vulnerable to a crime and can be used to assist in crime prevention

2. Background :

We referred various articles and blogs related to the crime rate in NYC, and also referred to currently existing similar apps on mobile like 'Citizen' which show the recent events categorized into threat levels for a neighborhood. We also referred various publications on this topics including :

"Crime Hot Spots: A Study of New York City Streets in 2010, 2015, and 2020"

"Crime and Enforcement Activity Reports"

"Crime forecasting: a machine learning and computer vision approach to crime prediction and prevention"

"Crime hotspot prediction based on dynamic spatial analysis".

3. Data :

1. NYPD Complaint Data Current (Year To Date) (NYC OpenData): This primary dataset includes all complaints received by the New York City Police Department (NYPD) for the current year (2022).
2. Precinct Data: This dataset contains all precincts and their addresses in New York City.
3. New York Zip Codes by Population: This dataset contains all zip codes in New York City with their population. The population data here are from the 2020 American Community Survey.

Our primary dataset - NYPD Complaint Data has 397,000 rows and following 36 columns:

Column Name	Column Description	Data Type
CMPLNT_NUM	Randomly generated persistent ID for each complaint	**Text**
ADDR_PCT_CD	The precinct in which the incident occurred	**Number**
BORO	The name of the borough in which the incident occurred	**Text**
CMPLNT_FR_DT	Exact date of occurrence for the reported event (or starting date of occurrence, if CMPLNT_TO_DT exists)	**DateTime**
CMPLNT_FR_TM	Exact time of occurrence for the reported event (or starting time of occurrence, if CMPLNT_TO_TM exists)	**Text**
CMPLNT_TO_DT	Ending date of occurrence for the reported event, if exact time of occurrence is unknown	**DateTime**
CMPLNT_TO_TM	Ending time of occurrence for the reported event, if exact time of occurrence is unknown	**Text**
CRM_ATPT_CPTD_CD	Indicator of whether crime was successfully completed or attempted, but failed or was interrupted prematurely	**Text**
HADEVELOPT	Name of NYCHA housing development of occurrence, if applicable	**Text**
HOUSING_PSA	Development Level Code	**Number**

JURISDICTION_CODE	Jurisdiction responsible for incident. Either internal, like Police(0), Transit(1), and Housing(2); or external(3), like Correction, Port Authority, etc.	**Number**
JURIS_DESC	Description of the jurisdiction code	**Text**
KY_CD	Three digit offense classification code	**Number**
LAW_CAT_CD	Level of offense: felony, misdemeanor, violation	**Text**
LOC_OF_OCCUR_DESC	Specific location of occurrence in or around the premises; inside, opposite of, front of, rear of	**Text**
OFNS_DESC	Description of offense corresponding with key code	**Text**
PARKS_NM	Name of NYC park, playground or greenspace of occurrence, if applicable (state parks are not included)	**Text**
PATROL_BORO	The name of the patrol borough in which the incident occurred	**Text**
PD_CD	Three digit internal classification code (more granular than Key Code)	**Number**
PD_DESC	Description of internal classification corresponding with PD code (more granular than Offense Description)	**Text**
PREM_TYP_DESC	Specific description of premises; grocery store, residence, street, etc.	**Text**
RPT_DT	Date event was reported to police	**DateTime**
STATION_NAME	Transit station name	**Text**
SUSP_AGE_GROUP	Suspect's Age Group	**Text**
SUSP_RACE	Suspect's Race Description	**Text**
SUSP_SEX	Suspect's Sex Description	**Text**
TRANSIT_DISTRICT	Transit district in which the offense occurred.	**Number**

VIC_AGE_GROUP	Victim's Age Group	**Text**
VIC_RACE	Victim's Race Description	**Text**
VIC_SEX	Victim's Sex Description	**Text**
X_COORD_CD	X-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD 83, units feet (FIPS 3104)	**Number**
Y_COORD_CD	Y-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD 83, units feet (FIPS 3104)	**Number**
Latitude	Midblock Latitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)	**Number**
Longitude	Midblock Longitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)	**Number**
Lat_Lon	Latitude and Longitude combined	**Location**
New Georeferenced Column	Location on earth as WGS84 Latitude and Longitude	**Point**

These features are a mix of :

– Quantitative Continuous (Time of occurrence, Latitude/Longitude),
– Qualitative Ordinal (Age groups for suspect and victim),
– Qualitative Nominal (Precinct code, type of offense)

with Numerical, Categorical, and Textual values.

4. Existing predictions?

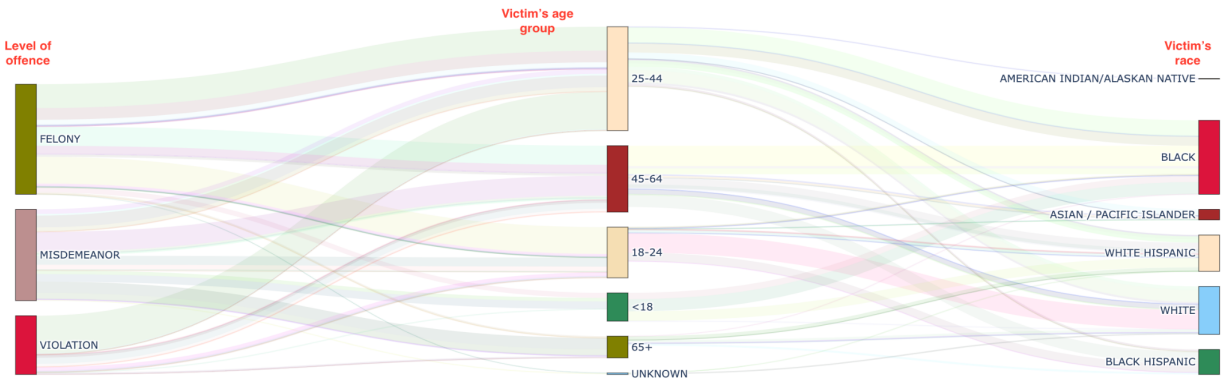
There are crowd sourced applications which give only the real time incidents but our model takes historical data of each location (zip code) into account to assign it a cluster and predict how its assigned cluster would change in near future. We could only find static data analysis and no Kaggle competitions or applications on the topic.

5. Preprocessing and EDA:

We first performed various **preprocessing tasks** for our data cleaning:

- Replaced appropriate values of '(null)' with null, 'Unk'/'U' with 'Unknown' etc.
- Removed features which had most of the values as null and carry little to no information about the type of crime and its location such as 'HADEVELOPT','HOUSING_PSA','PARKS_NM'
- Dropped null rows for features which had less than 1% null values in the entire dataset.
- All rows with complaint time before 1st Jan 2022 were discarded, since the dataset claims to have complaints registered in the current year.
- Replaced the row having multiple dates of crime with their mean.
- Split timestamps to date and time for day wise predictions.
- Latitude/Longitudes were used to reverse lookup the zip code for each complaint
- Precinct codes were converted to addresses and later to precinct latitude and longitude.
- Merged additional data – precinct code and zip codes with the original dataset.
- We divided all offenses into one of these 5 categories – Loss of Life/Violence, Sexual, Weapons and Drugs, Theft and others. Out of these five categories, we assumed that region safety is more influenced by the first four.
- Zip code level features (Counts for offenses related to Loss of Life/Violence, Sexual, Weapons and Drugs, Theft) were calculated for each day by aggregating the complaints in original dataset.

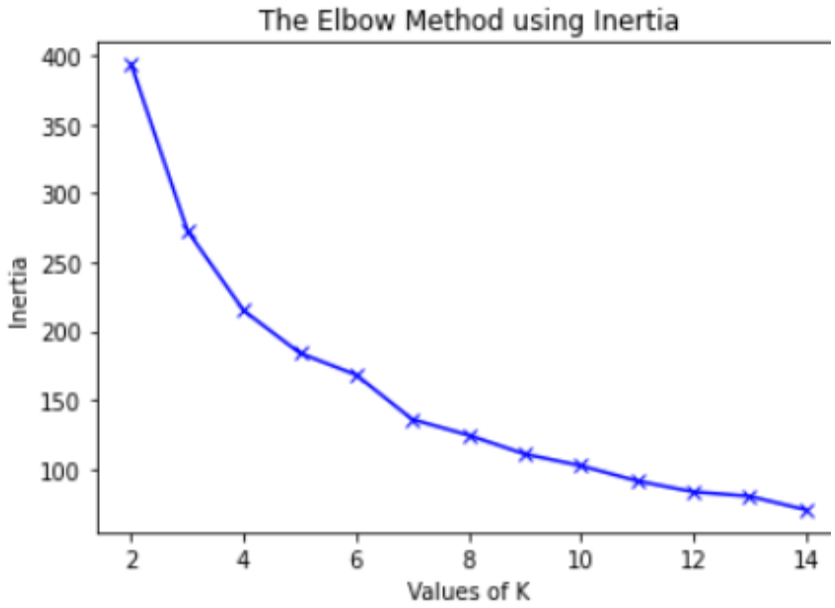
We also performed **Exploratory Data Analysis** on our primary dataset after the cleaning. We studied how the data is distributed over different boroughs, distribution of each crime level in different districts. We drew a basic Sankey Diagram to study the distribution of victim's race vs suspect's race. We also estimated the population by borough to count the crime rate.



6. Model and Evaluation

There were no known labels in our dataset for each zip code, so we used **unsupervised clustering algorithms** – k means and MeanShift, for clustering similar zip codes together using zip-level features.

Since the number of samples i.e number of zip codes is low (196), we preferred non-hierarchical algorithms which allowed unequal clusters – k-means with 5-7 clusters appeared to be a good fit. We also used the Elbow method with inertia to select a good value for the number of clusters (k = 7).

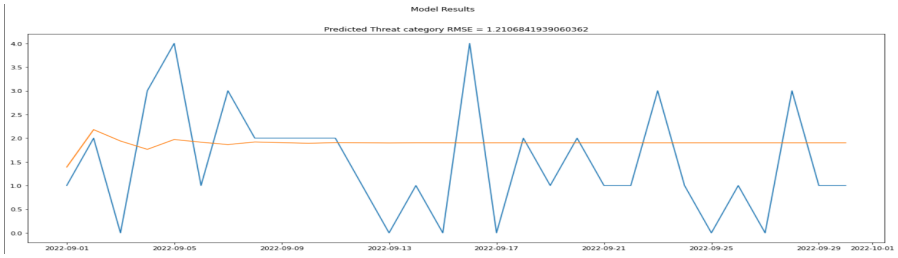


Existing seasonality in the data indicated that time series models such as ARIMA and AutoARIMA can be used to predict the feature values for a zip code and the predicted features can in turn be used to assign the zip to one of the now known clusters.

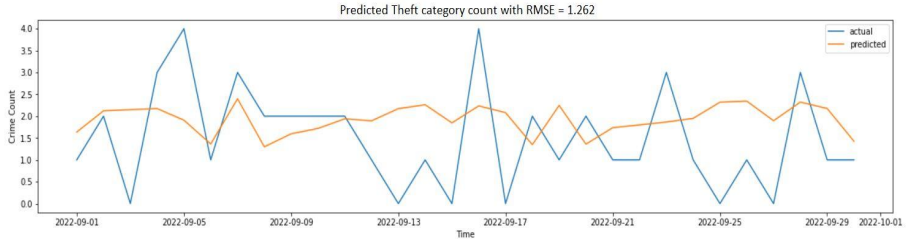
We explored both ARIMA and AutoARIMA models and decided to use the **AutoARIMA** which allowed us to use different model parameters for each zip code to account for different time series patterns in the zip codes' features. Even though the ARIMA model had a marginally better RMSE score as shown below, its output was almost a constant value with no seasonality.

We avoided linear regression models because our target variable is categorical (k clusters) which is better suited for classification models.

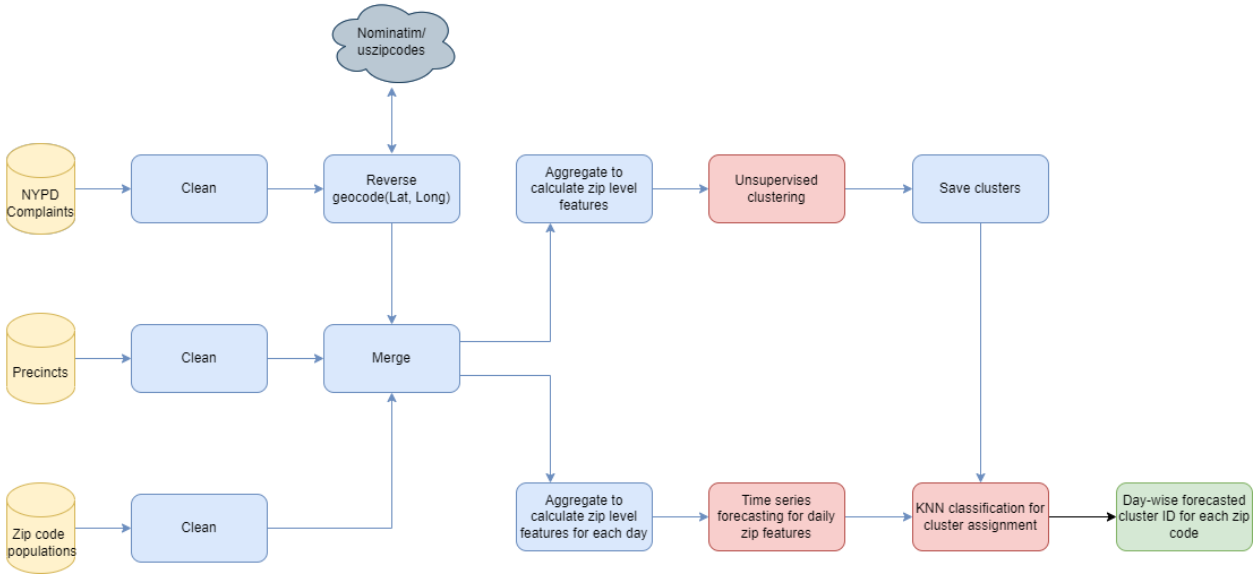
Almost constant predictions by ARIMA



Changes captured by AutoARIMA:



Model Pipeline



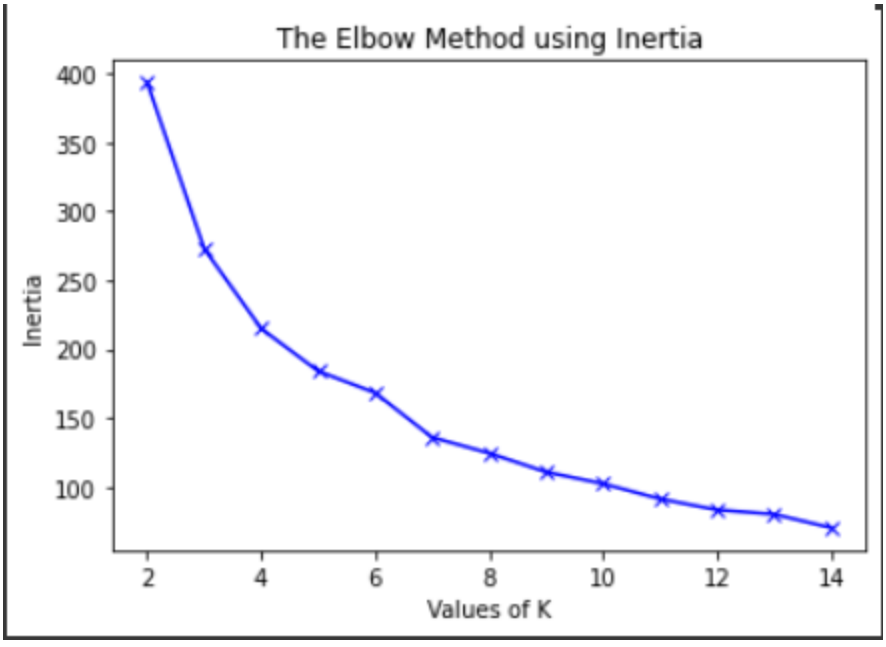
Evaluation and Visualization :

- For evaluating time series predictions, we used **RMSE** and **MFE** errors.

Arima	**RMSE (Mean across zip codes)**	**MFE (Mean across zip codes)**
Category 1	1.55	0.75
Category 2	0.68	0.19
Category 3	1.26	0.26
Category 4	4.68	1.48

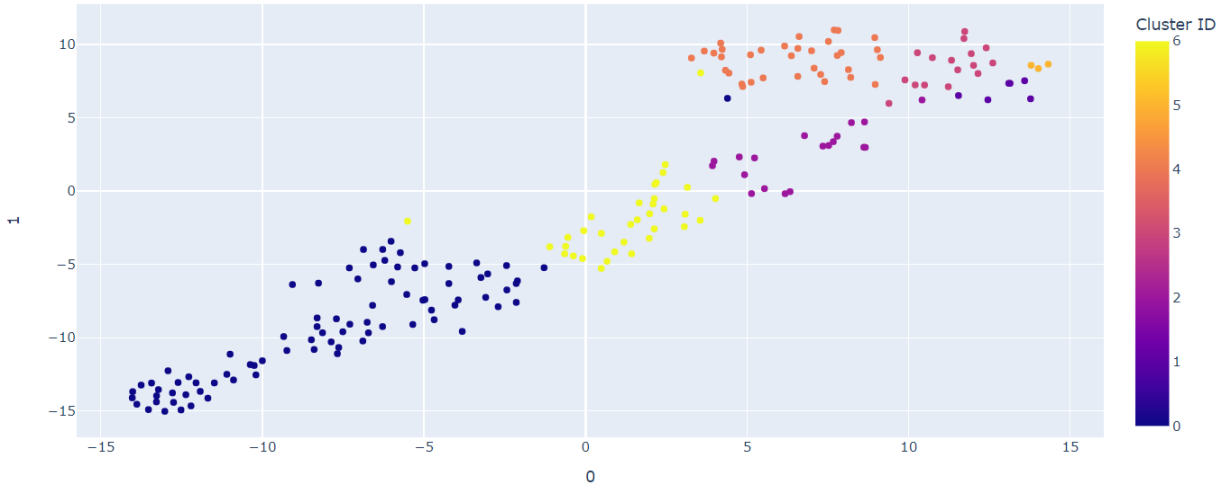
AutoArima	**RMSE (Mean across zip codes)**	**MFE (Mean across zip codes)**
Category 1	1.59	0.74
Category 2	0.71	0.17
Category 3	1.30	0.29
Category 4	4.69	1.41

- For selecting the number of clusters (k = 7) in unsupervised clustering, we used the **Elbow-method with inertia.**

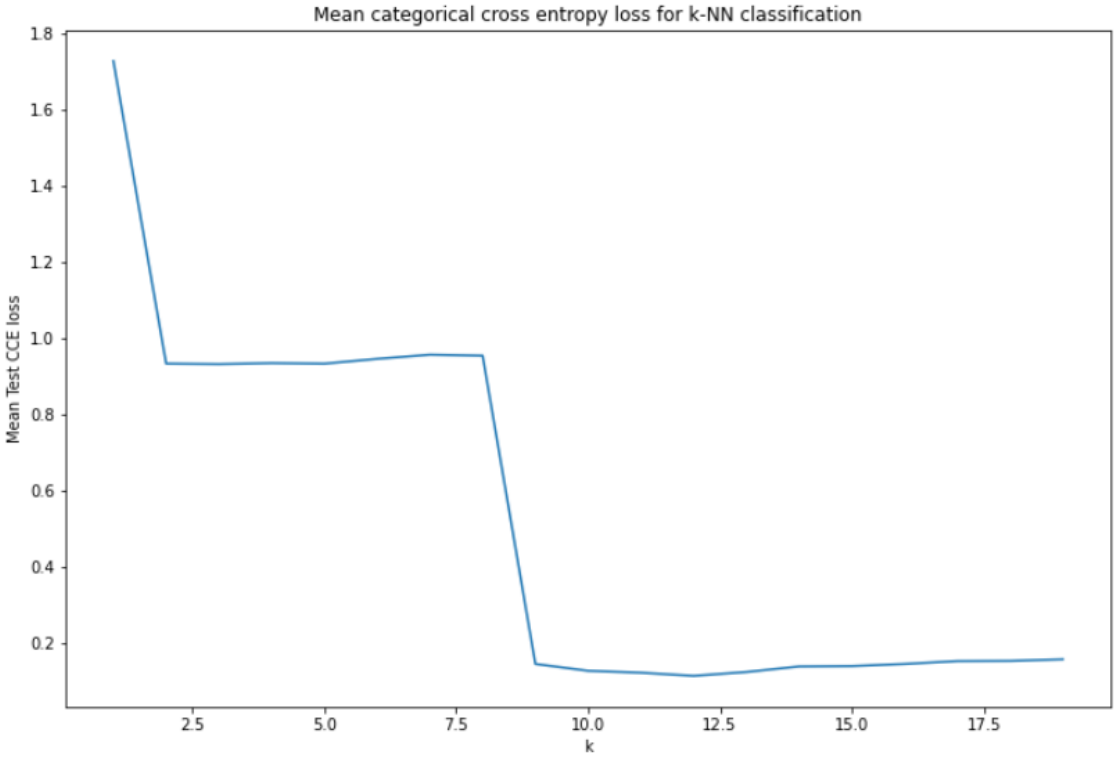


- **t-SNE algorithm** was used to visualize the separation of assigned 4-D clusters in 2 dimensions. Each point represents one zip code and its assigned cluster. Cluster centers scaled :

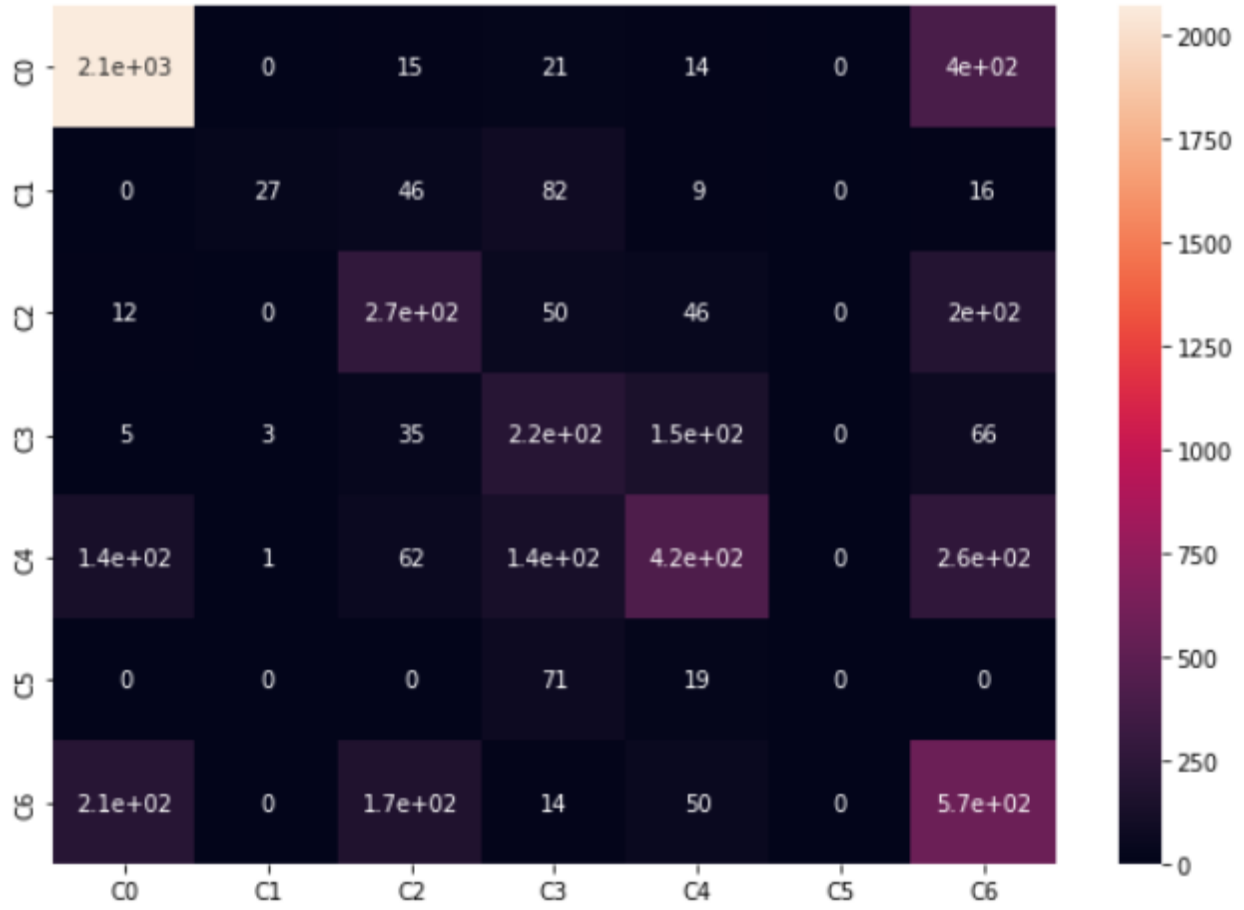
Cluster ID	Category1 (Loss of life/Felony)	Category2 (Sexual)	Category3 (Drugs/ Weapons)	Category4 (Theft)	Interpretation
0	-0.57	-0.64	-0.67	-0.85	Fewest crimes across all 4 categories; Safest zip codes
1	2.00	1.69	3.37	1.44	High crime rates across all 4 categories; second worst cluster of zip codes
2	-0.10	0.00	0.79	1.87	Large number of thefts and weapon related complaints; Relatively few violent crimes
3	1.64	1.96	1.04	0.75	High crime rates across first 3 categories excluding thefts; Should be avoided
4	0.49	0.73	-0.04	0.12	Average crime rates
5	4.57	2.83	2.24	1.23	Most crimes across all 4 categories; worst zip codes
6	-0.56	-0.61	-0.03	0.21	Large number of thefts; Relatively few violent crimes



- Finally for classification on new features, we used k-NN classification with probabilistic interpretation and selected the k value which minimized **Categorical Cross Entropy** loss on a validation dataset(80–20 split) i.e. k = 9.



The variation of cluster assignment in September 2022 against the ground truth cluster for each zip code has been captured by the following confusion matrix where the rows denote the assumed ground truth cluster for each zip code, and the columns denote the assigned cluster by k-NN model on each day of the month :



This shows that the ‘safe’ zip codes (C0 and C6) are consistently safe. The ‘unsafe’ zipcodes are consistently unsafe (C5, C1, C3). The variation in the assigned cluster ID captures the temporal nature of criminal activities for the zip codes.

7. Assumptions and Limitations:

We have made the following assumptions for our model :

- Data is representative of the population.
- All committed crimes are being reported to the NYPD.
- Number of crimes committed under each of the four categories – Loss of Life/Violence, Sexual, Weapons and Drugs, Theft determine relative safety of the zip codes.
- We are using a subset of Offense Descriptions which are more severe than other crimes to calculate the zip code features e.g. 'NEW YORK CITY HEALTH CODE','LOITERING/GAMBLING(CARDS, DIC', DISRUPTION OF A RELIGIOUS SERV', 'FRAUDULENT ACCOSTING' crimes have been discarded while calculating zip level features. It is a reasonable assumption because financial and fraud related crimes do not directly impact the safety of an individual in a region
- We assumed that the crime location was accurate. However the crime location was set to the precinct location for some of the severe crimes ('MURDER & NON-NEGL. MANSLAUGHTER','RAPE','SEX CRIMES','FELONY SEX CRIMES', 'KIDNAPPING','FELONY ASSAULT'). We have assumed that the precinct to which the crime is reported lies in the same zip code where the crime occurred. This is a reasonable assumption because crimes are reported to the nearest precinct and aggregating the statistics on a zip code level captures the spatial variance of criminal activity
- For crimes where the exact time of crime was not known, we assumed it occurred exactly between the reported ‘from’ and ‘to’ times. It only affects the daily crime statistics if the from and to times cross over into the next day
- For assigning cluster ID, we extrapolate daily crime statistics for each zip code to match the ground truth cluster assignment (which was done using data aggregated for 272 days - 1st Jan 2022 to 30th Sept 2022)

These are the following limitation of our approach:

- Lack of known classification for the zipcodes makes it difficult to evaluate the classification model.
- The predicted outcomes are also limited by the time horizon i.e. we cannot predict the features beyond n days (30)
- Extrapolation of daily statistics to a longer time period can lead to high variation in the predicted results

Next Steps :

- More features can be incorporated for each zip code to create richer predictions
- Historical data can be used to produce better time series models by incorporating monthly trends
- A deployed solution such as web app needs to use online learning by consuming live data periodically