MSCA31006

CHICAGO CRIME ANALYSIS

USING TWO DECADES OF OFFICIAL DATA FROM THE CHICAGO POLICE DEPARTMENT

Moushumi Pardesi

PROBLEM STATEMENT

- The dataset contains reported crimes occurred in the City of Chicago from 2001 to 2023
- Sourced from Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system
- Similar forecasts have been made in the past, however, most use limited period data. Using a longer duration of data (such as 2 decades in this analysis) will help build a robust model to understand trends.
- The dataset used here spans 21 years from 2001 to 2022 and has 7800490 observations and 30 different identifiers. This is substantial data to predict crime.

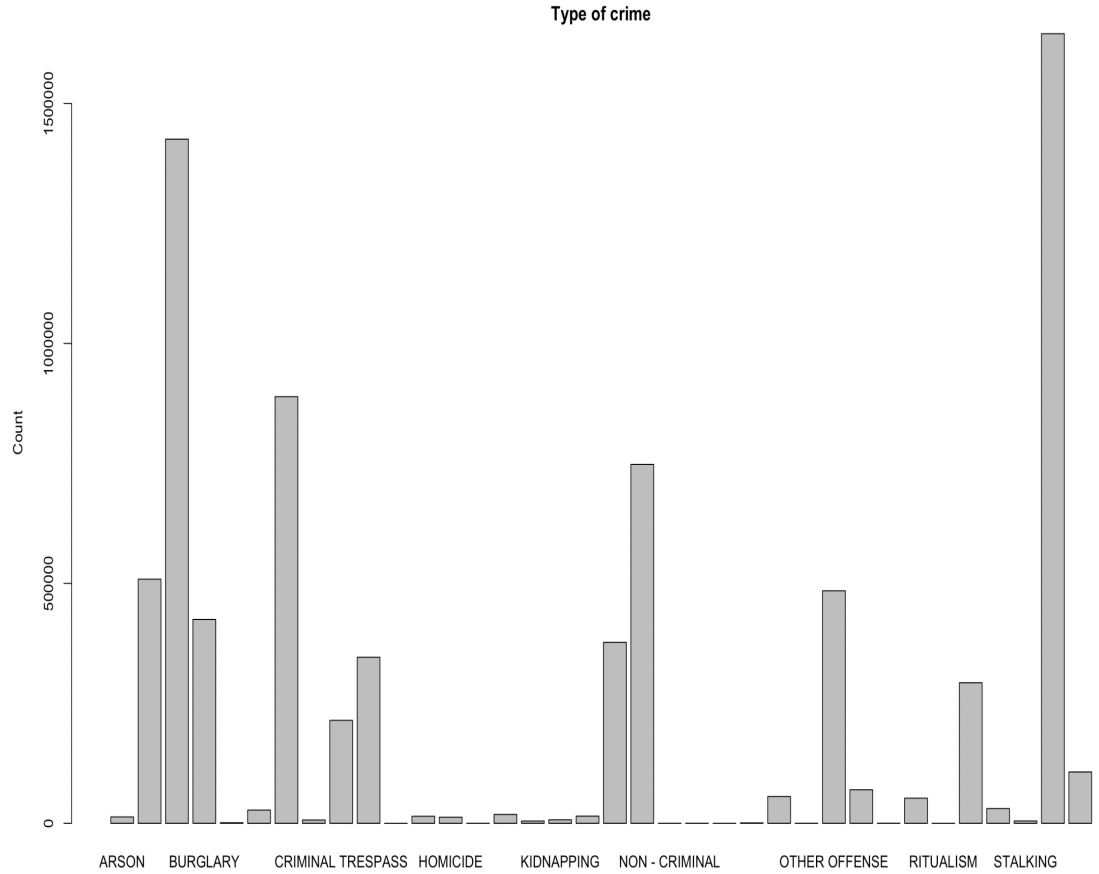
THE DATA

- Features in the data (not all are useful for prediction):
  - ID: Unique identifier for the record.
  - Case Number: The Chicago Police Department RD Number (Records Division Number), which is unique to the incident.
  - Date: Date when the incident occurred.
  - Block: address where the incident occurred
  - IUCR: The Illinois Uniform Crime Reporting code.
  - Primary Type: The primary description of the IUCR code.
  - Description: The secondary description of the IUCR code, a subcategory of the primary description.
  - Location Description: Description of the location where the incident occurred.
  - Arrest: Indicates whether an arrest was made.
  - Domestic: Indicates whether the incident was domestic-related as defined by the Illinois Domestic Violence Act.
  - Beat: Indicates the beat where the incident occurred. A beat is the smallest police geographic area – each beat has a dedicated police beat car.
  - District: Indicates the police district where the incident occurred.
  - Ward: The ward (City Council district) where the incident occurred.
  - Community Area: Indicates the community area where the incident occurred. Chicago has 77 community areas.
  - FBI Code: Indicates the crime classification as outlined in the FBI's National Incident-Based Reporting System (NIBRS).
  - X Coordinate: The x coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection.
  - Y Coordinate: The y coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection.
  - Year: Year the incident occurred.
  - Updated On: Date and time the record was last updated.
  - 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.
  - 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.
  - 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.

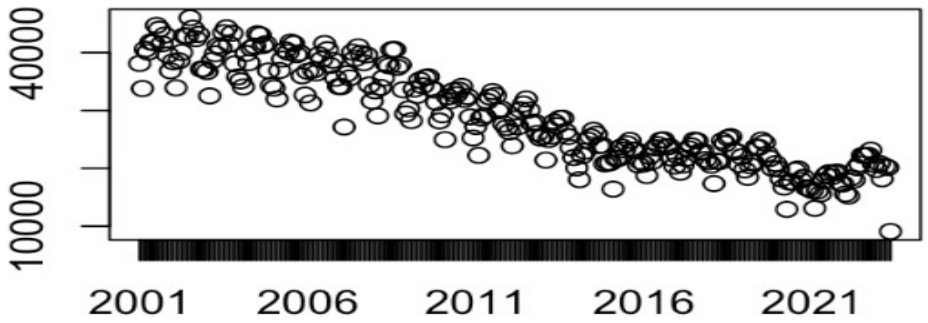
ASSUMPTIONS / HYPOTHESIS

- For simplicity, this model excludes the location data altogether (including police beats, wards, longitude / latitude location pin, ward, community areas, historical classification of location, district, etc.) and broadly classifies all crime data as that occurred within Chicago
- The model is concerned about the time series dataset of crime and for simplicity, converts the data from multiples times in a day (non-uniform frequency) to monthly crime data for 22 years.
- As per the Augmented Dickey-Fuller Test, there is strong evidence to assume that the crime data is stationary

VISUALIZING THE DATA

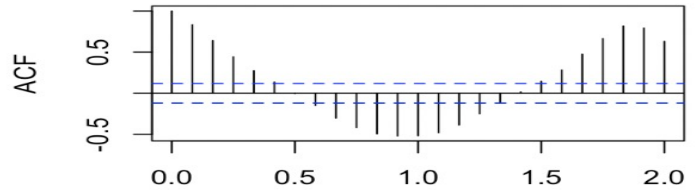


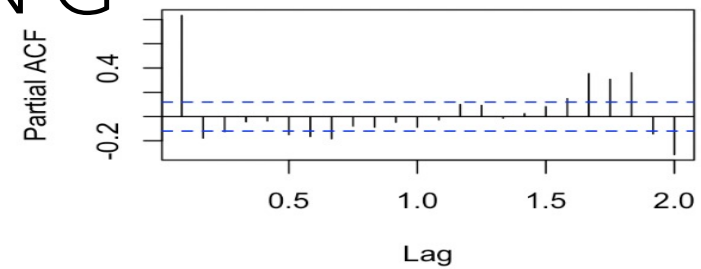
- Assessing the two-decade long data proves that Chicago is more prone to certain types of crimes than others
- There is a steady fall in crime over the two decades

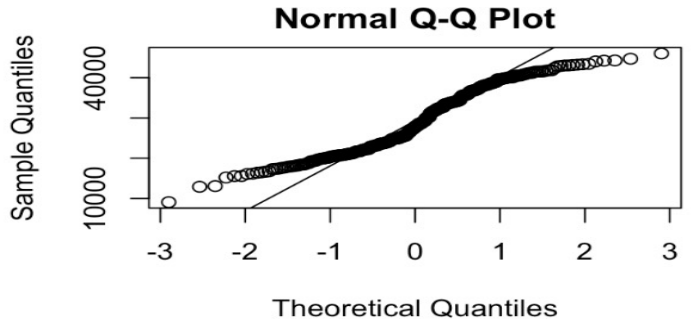


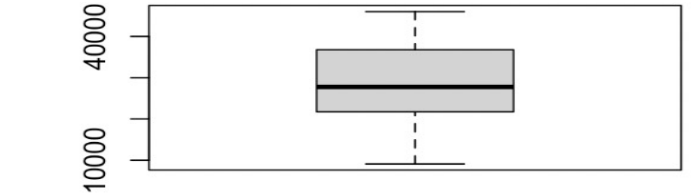
EXPLORING AND CLEANING

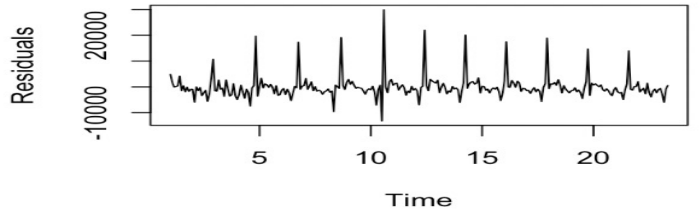
- The ACF, PACF tests suggest that the data has autocorrelation of a high degree
- The QQ plot suggests that the data is symmetrical, but deviates from the normal distribution
- The boxplot of the data shows that most data is concentrated within the mid 50 percentile (25%ile to 75%ile), however, there are significant outliers in the data
- ARIMA residual analysis suggests that the residuals from the ARIMA model are not normally distributed.



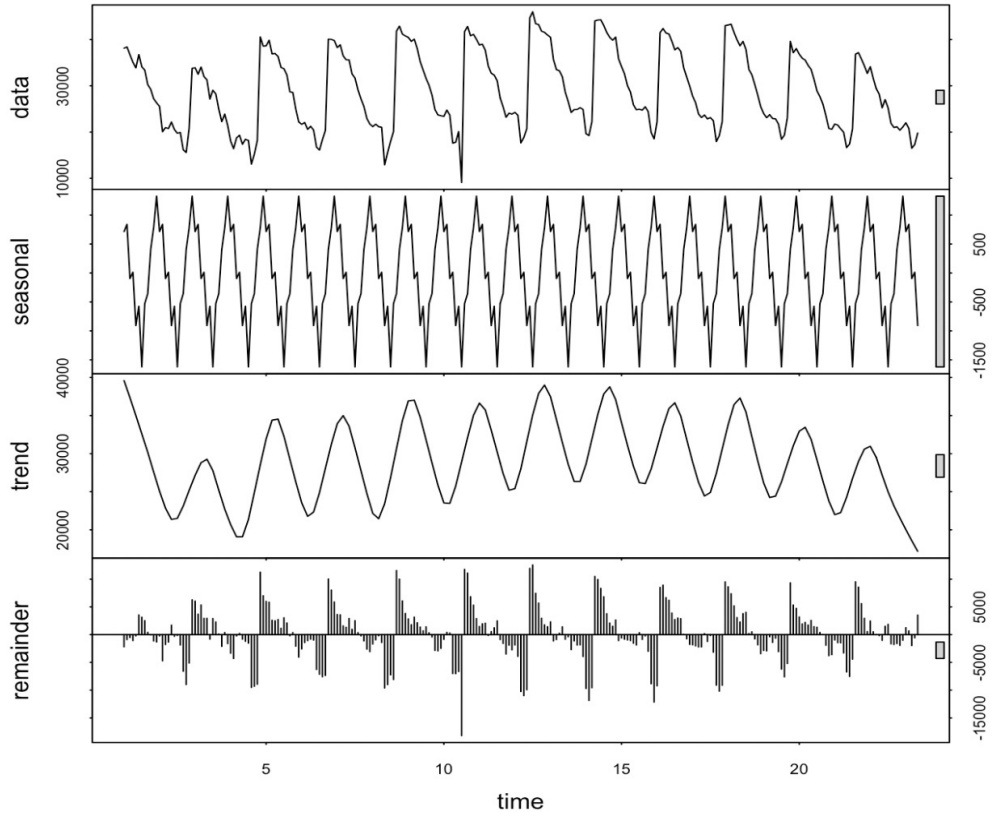








FEATURE ENGINEERING



- I created lagged variants of the crime data to capture temporal dependencies
- Also used Fourier transform to decompose the time series into different frequency components
- Finally, I used STL decomposition to to extract the trend and seasonal components from the time series

PROPOSED APPROACHES

- ARIMA model can capture temporal dependencies, trend, and seasonality in the data. It can be effective for modeling crime data with stationary or stationary-differenced components.
- SARIMA extend ARIMA models to incorporate seasonal patterns in the data. They are suitable for time series data with both temporal dependencies and seasonal fluctuations which are apparent in this Chicago Crime dataset.
- Exponential smoothing methods, such as Holt-Winters' Exponential Smoothing are commonly used for forecasting time series with trend and seasonality and can be suitable for the crime data.
- Prophet, developed by Facebook's Core Data Science team, is specifically designed to handle time series data with trend changes, seasonality, and holiday effects. Prophet utilizes additive regression models and incorporates a customizable modeling approach. It offers automatic detection of changepoints and flexibility in modeling various components of time series data.

CHOOSING PROPHET

Other than the fact that the classroom discussion and reading material got me intrigued into the mechanics and applications of Prophet, the following are my practical considerations for choosing it.

- Handling Complex Patterns: Prophet can handle the crime data with complex patterns, including nonlinear trends, multiple seasonality, and holiday effects.
- Flexibility: Prophet allows customization of modeling components, such as trend flexibility, seasonality parameters, and inclusion of additional regressors, providing flexibility in capturing specific patterns in crime data.
- Automatic Changepoint Detection: Prophet automatically detects and incorporates changepoints, identifying shifts or changes in the underlying patterns of the time series data.
- Interpretability: Prophet provides transparent and interpretable outputs, including visualizations of trend, seasonality, and forecast components, making it easier to understand and explain the model results.

RESULTS

- The model can handle the cyclical as well as the seasonal patterns in the data. It also considers the overall decreasing trend in crime in Chicago.
- The model detects 25 change points in the time series across two decades.
- It also detects seasonality in the dataset.
- The model makes forecast for the nest 12 periods, i.e. one year.

CONCLUSION – FUTURE WORK

- Crime in Chicago has several types seasonality, for example, the crime is higher in the summer compared to winter, it is higher on weekends compared to weekdays, it is higher in the night compared to the day. In this analysis, we consider monthly data which overlooks daily/weekly/hourly seasonality. This is an area that can be explored further.
- More sophisticated models such as LTSM (long short-term memory networks), Gaussian Processes, Hidden Markov Models, BSTS, Ensemble Models would be able to capture and model the crime data in ways Prophet may have missed. We can explore those as a next step and see if there is anything that can enhance the way we are projecting crime for the next few periods.

GITHUB LINK TO CODE

https://github.com/MoushumiP/MSCA_final_2023/blob/main/MSCA_final_Project.Rmd

CITIZEN LAW ENFORCEMENT ANALYSIS AND REPORTING LINK TO DATA

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