NEIGHBORHOOD-LEVEL CRIME FORECASTING IN CHICAGO

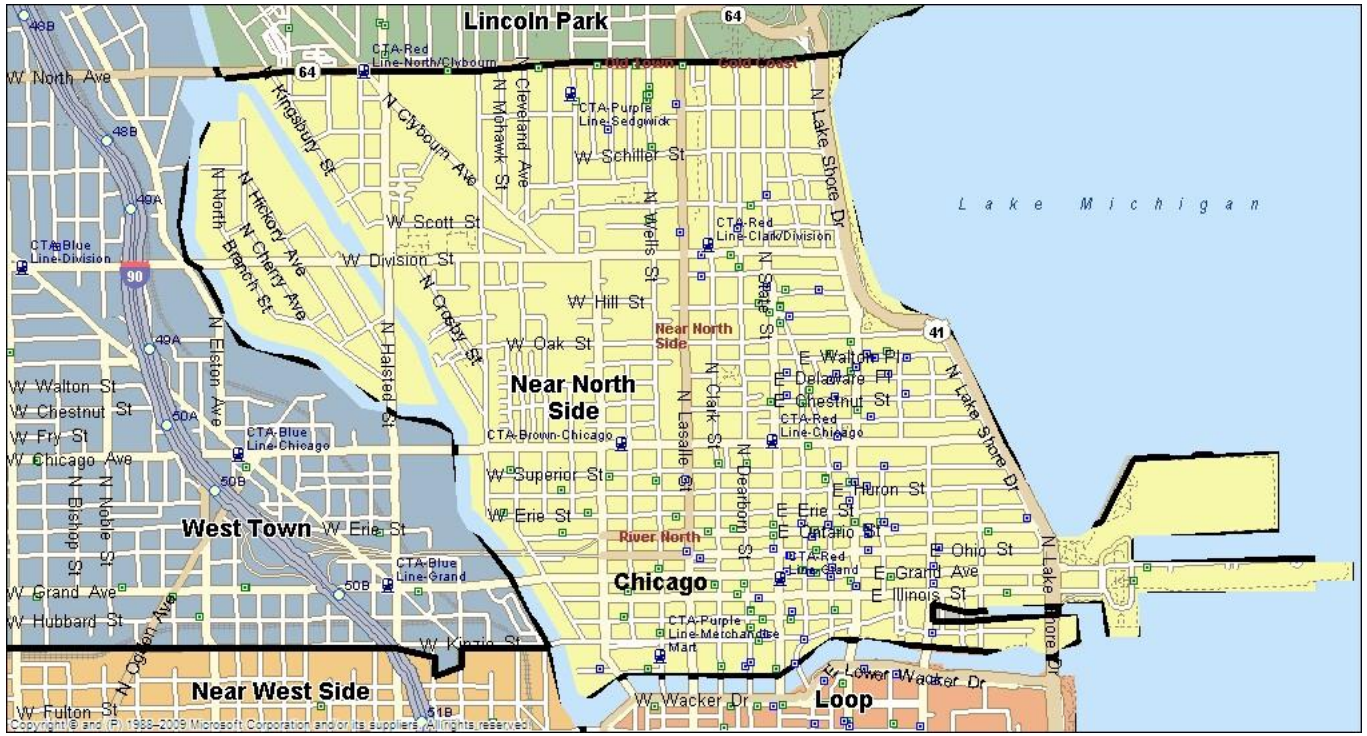
IAN VOGT

MSCA 31006 IP02 (SPRING 2023) TIME SERIES ANALYSIS AND FORECASTING

AGENDA

- Problem Statement
- Data Properties
- Data Processing
- Exploratory Data Analysis
- Proposed Approaches
- Assumptions
- Feature Engineering
- Results
- Future Work

WHAT IS THE PROBLEM?



- Chicago's "Near North Side" neighborhood includes the location of the University of Chicago M.S. in Applied Data Science program, Booth School, and residential locations for many students

- Problem Statement: When do members of the University of Chicago Booth School and M.S. in Applied Data Science programs need to be most vigilant regarding local crime?

DIVING INTO THE DATA

- Source: [Chicago Data Portal Crimes (January 2001 – April 2023) Dataset]
- Description: Timestamped crime log of every arrest in Chicago
- Important Property: Non-Uniform time-series in default state

ID	Case ...	Date	Block
13073148	JG260938	05/15/2023 05:15:00 AM	057XX W ...
13073396	JG261279	05/15/2023 05:00:00 AM	106XX S ...
13073311	JG261125	05/15/2023 05:00:00 AM	087XX S ...
13074790	JG261989	05/15/2023 05:00:00 AM	043XX S ...
13073146	JG260911	05/15/2023 04:57:00 AM	070XX S ...
13073160	JG260917	05/15/2023 04:55:00 AM	006XX W ...
13073175	JG260909	05/15/2023 04:48:00 AM	019XX W ...
13073354	JG260924	05/15/2023 04:30:00 AM	041XX W ...

NARROWING THE SCOPE

Filtering:

- Window of analysis limited to the last 5 years (January 2018 – April 2023)
- Filter out any crimes that did not occur in Community Area 8 (Near North Neighborhood)

Aggregation:

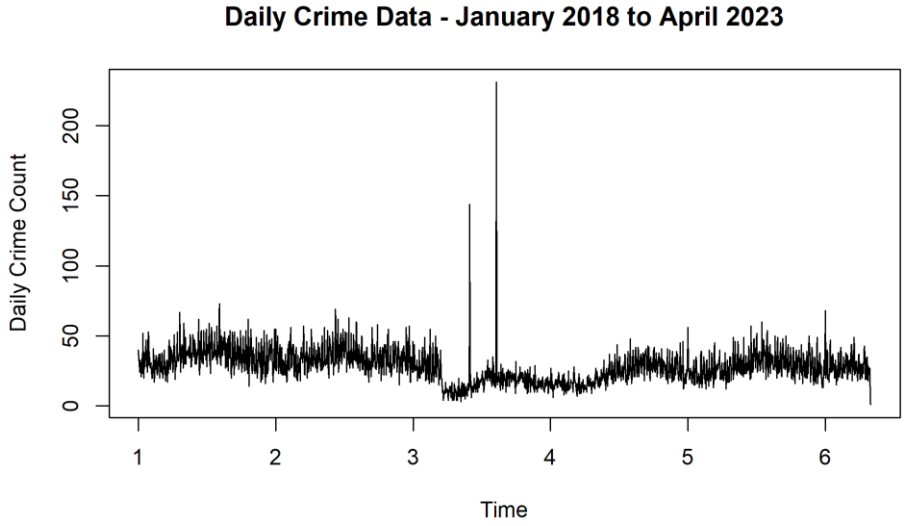
- Aggregate the data to daily and hourly levels to consider multiple seasonality's
- This transforms our data into a uniform time-series

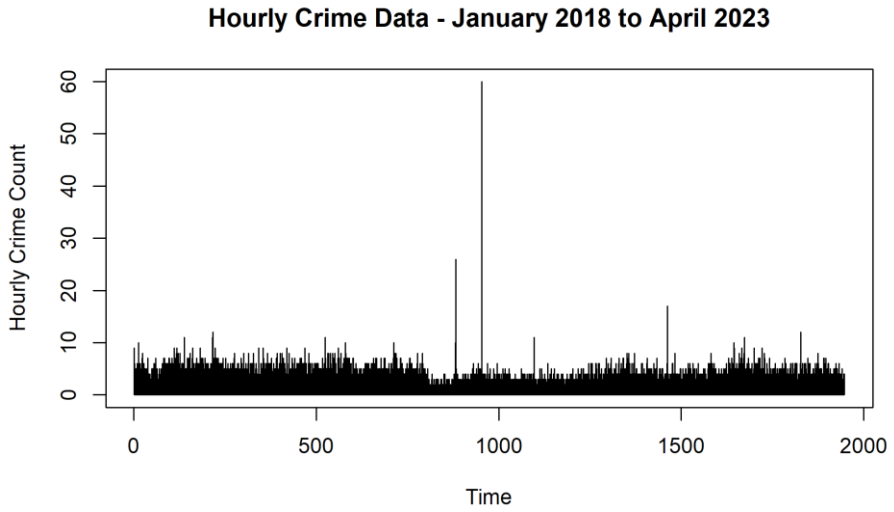
	hour	count
1	2018-01-01 00:00:00	8
2	2018-01-01 01:00:00	9
3	2018-01-01 02:00:00	3
4	2018-01-01 03:00:00	4
5	2018-01-01 04:00:00	1
6	2018-01-01 05:00:00	1
7	2018-01-01 07:00:00	1

	hour	count
1	2018-01-01 00:00:00	8
2	2018-01-01 01:00:00	9
3	2018-01-01 02:00:00	3
4	2018-01-01 03:00:00	4
5	2018-01-01 04:00:00	1
6	2018-01-01 05:00:00	1
7	2018-01-01 06:00:00	0
8	2018-01-01 07:00:00	1

Problem: Missing Data for 2018-01-01 06:00:00

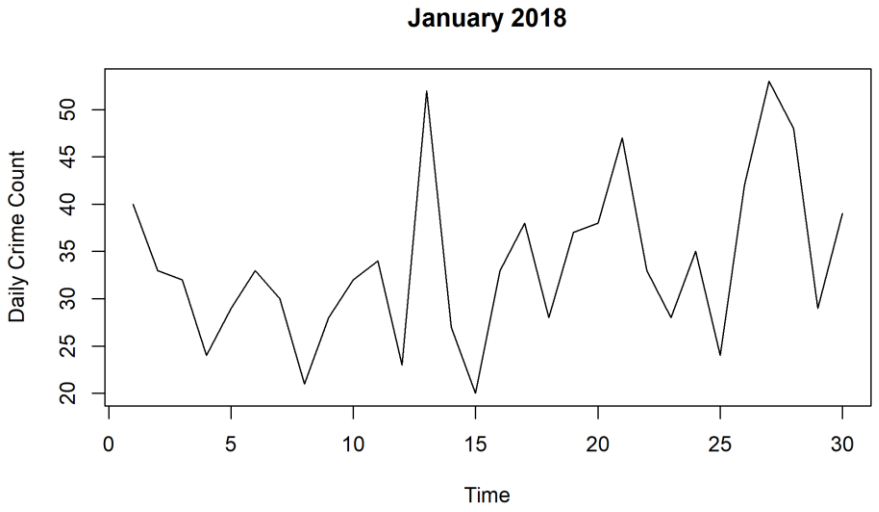
Solution: Create a new data frame with all hourly timestamps from 2018-01-01 00:00:00 to 2023-04-30 23:00:00, left join with the incomplete data, and replace leftover NAs with 0's

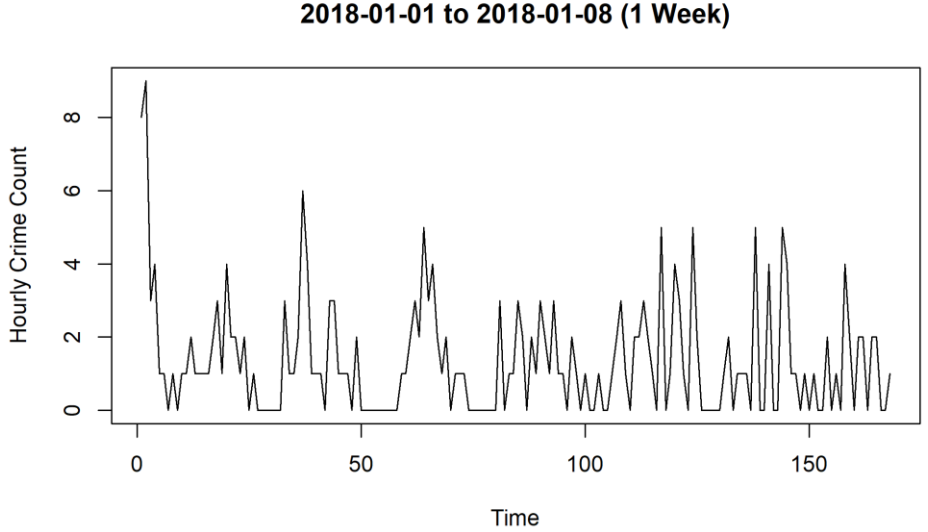




- No distinct trend, but year-over-year seasonality is observed
- The time-series is not smooth. Some anomalies are seen.

- The hourly time-series is a bit smoother but also contains the anomalies observed in the daily series

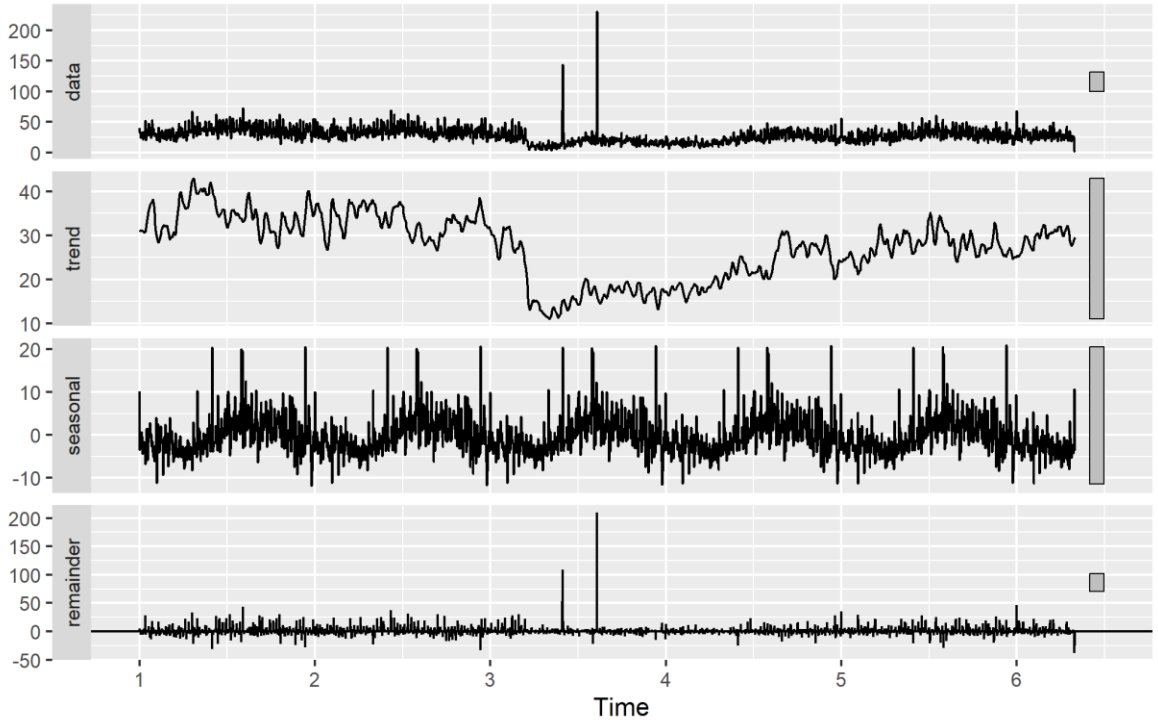




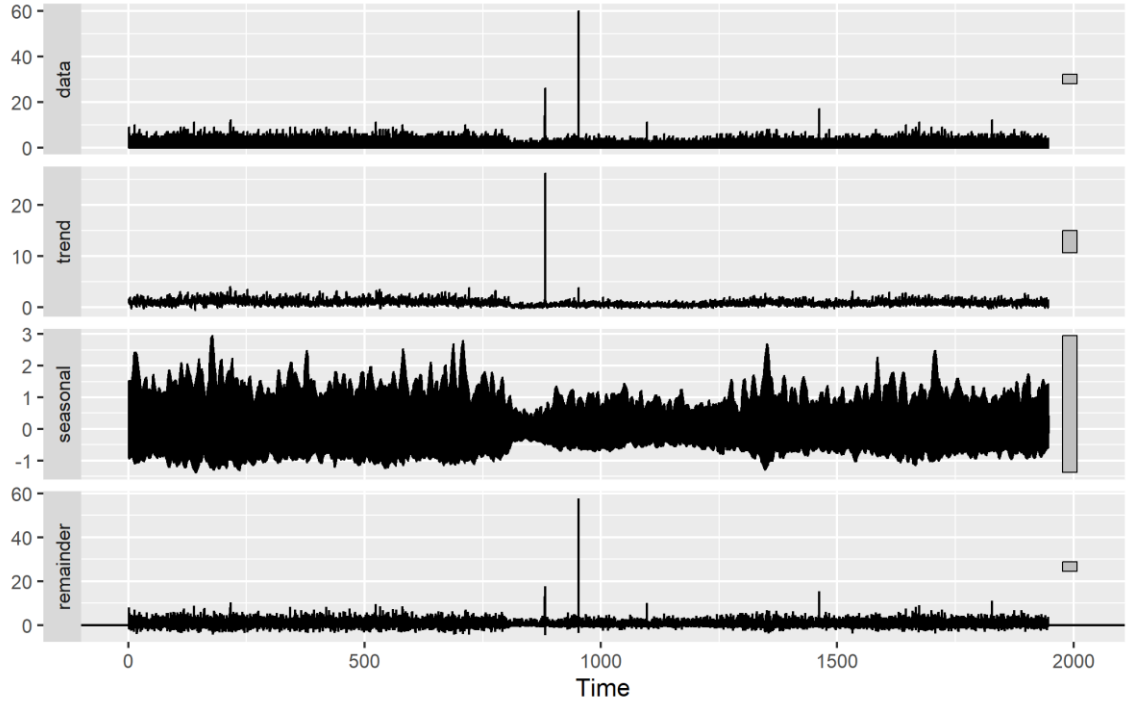
- Closer look reveals some day-of-the-week seasonality

- There appear to be two spikes per day in the hourly data

Daily



Hourly



- The STL decomposition of both daily and hourly time-series' reveal a more distinct trend. Crime dropped off precipitously around 2020 and has been steadily rising to pre-pandemic levels since.
- We can see again how the hourly data is smoother.

DEVELOPING AN APPROACH

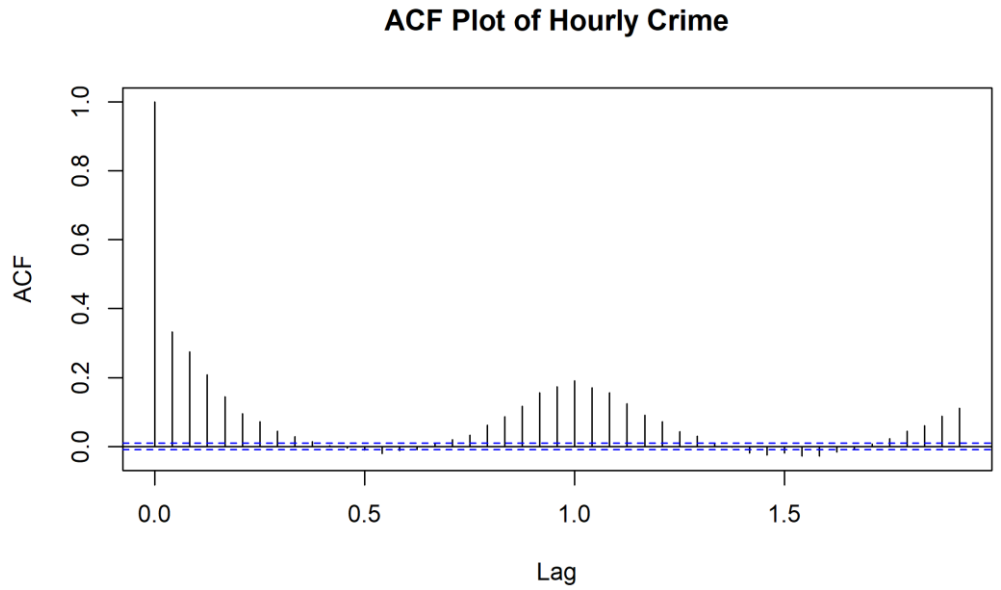
Proposal #1: Seasonal ARIMA Modeling

- ARIMA and Seasonal ARIMA modeling can be done on hourly data with some assumption validation.
- These models combine autoregressive, moving average, differencing, and seasonal components to produce forecasts, but cannot incorporate multiple seasonality's.

Proposal #2: Prophet Modeling

- Prophet from Meta can handle time series data with multiple seasonality components like simultaneous daily, weekly, and yearly patterns, combining trend estimation, seasonality modeling, and holiday effects to generate forecasts.

VALIDATING ASSUMPTIONS OF SEASONAL ARIMA



- ACF plot exhibits autoregression and seasonality on daily level. We will need to lag the data with d variable in the ARIMA models.

- Extreme values from the TS plots are considered anomalies given circumstances around protests and activism during Covid-19 Pandemic. I assume these to be a one-off case. Outliers ☑
- ADF Test P-Value < 0.01. Stationarity ☑

ENGINEERING AND SELECTING MODELS

- I train four models for forecasting hourly crime in Near North Chicago
- First I try ARIMA, then two Seasonal ARIMA models, and lastly I use Prophet
- I use auto.arima in R to engineer ARIMA/SARIMA parameters

MODELS

ARIMA w/ No Seasonality	Day-of-the-Week Seasonal ARIMA	Hour-of-the-Day Seasonal ARIMA	Prophet
P = 0 Q = 1 D = 1	P = 5 Q = 1 D = 0	P = 0 SAR = 0 Q = 1 SI = 0 D = 1 SMA = 2	Fitted curve with no features

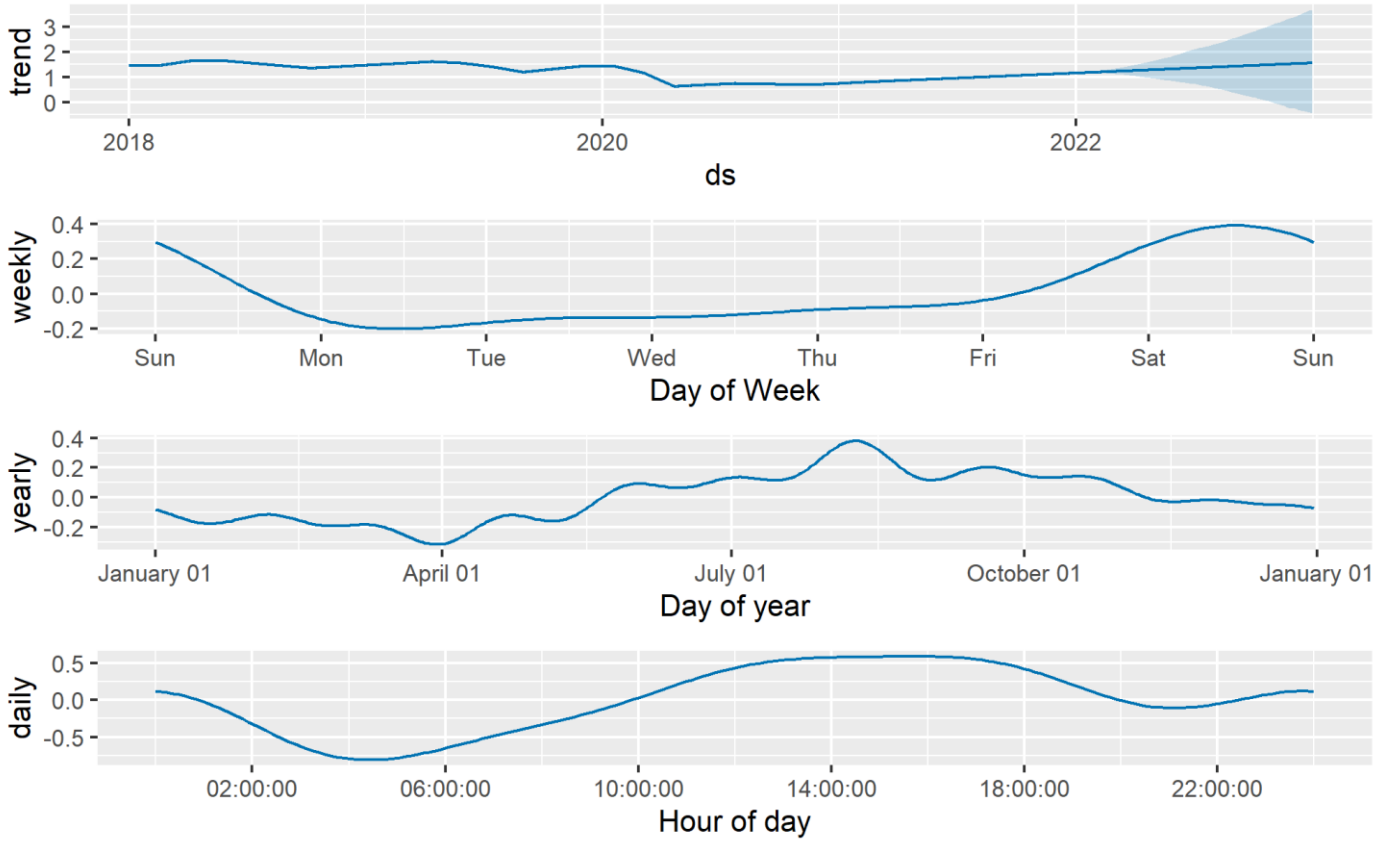
- Train Window: 01-01-2018 to 12-31-2021. Forecast Window: 01-01-2022 to 04-30-2023

REVIEWING MODEL RESULTS

Model	MAE	RMSE	sMAPE
Non-Seasonal Arima	1.30	1.55	103.21
Day-of-the-Week Seasonal Arima	1.98	2.21	113.24
Hour-of-the-Day Seasonal Arima	1.31	1.56	103.40
Prophet	**1.01**	**1.29**	**101.46**

- Adding singular day-of-the-week or hour-of-the-day seasonality did not improve the non-Seasonal ARIMA model
- Prophet outperforms the other three in all testing metrics.

DIVING INTO PROPHET



- We can see Prophet is able to capture multiple seasonality's, as advertised.
- Crime tends to peak on the weekends, in the summer, and twice in a day (afternoon and overnight)
- Prophet also better handled the prevalence of zero values in the time-series and appropriately predicts lower values.

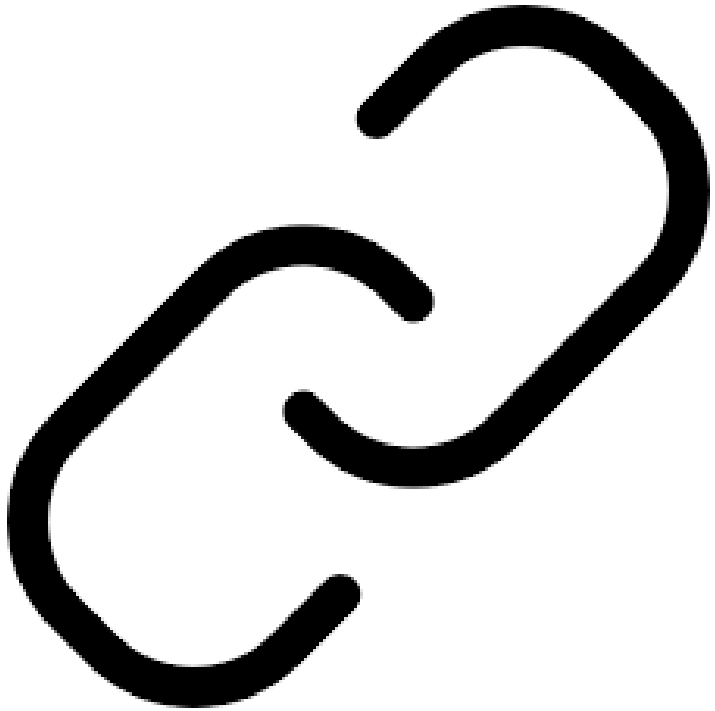
CONCLUSIONS AND FUTURE CONSIDERATIONS

- Univariate time-series data with multiple seasonality's and a prevalence of zero values presents an apparent use-case for Prophet.
- Many approaches can be taken with this dataset, but comparing results to daily data with more neighborhoods included or with data aggregated at the weekly level would be noteworthy to see how Prophet holds up.
- Computational limitations and data size made tinkering with ARIMA models difficult and prevented inclusion of data from before 2018
- Crime activity in 2020 was very unique due to the Covid-19 pandemic. This year of data could be worth a closer look to see how it clouds the 2022 and 2023 forecasts. Did crime activity normalize in 2022 and 2023?

APPENDIX: LINKS

LINKS:

- DATA (Crimes 2018 Version)
- CODE (GITHUB)

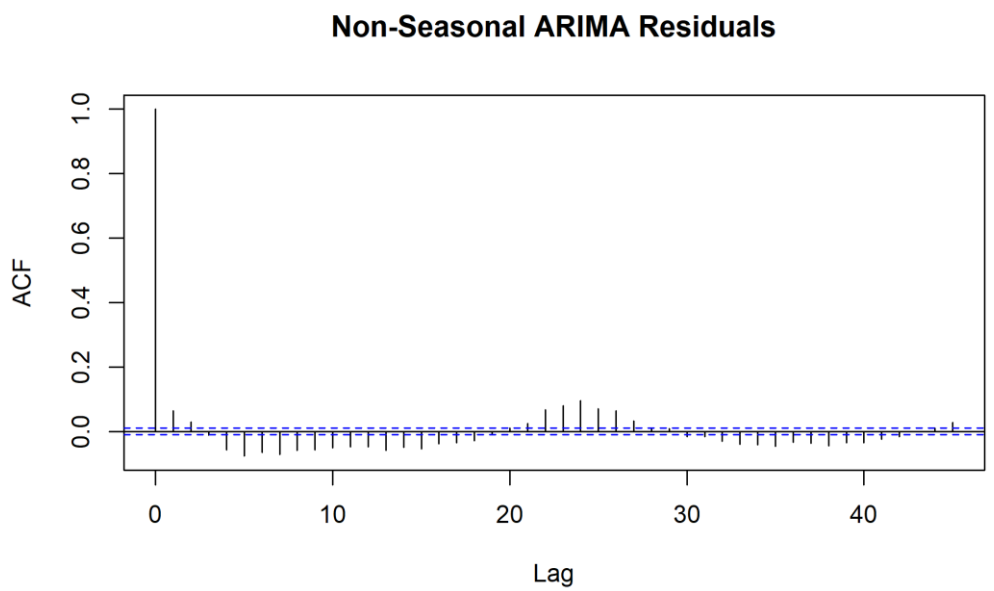


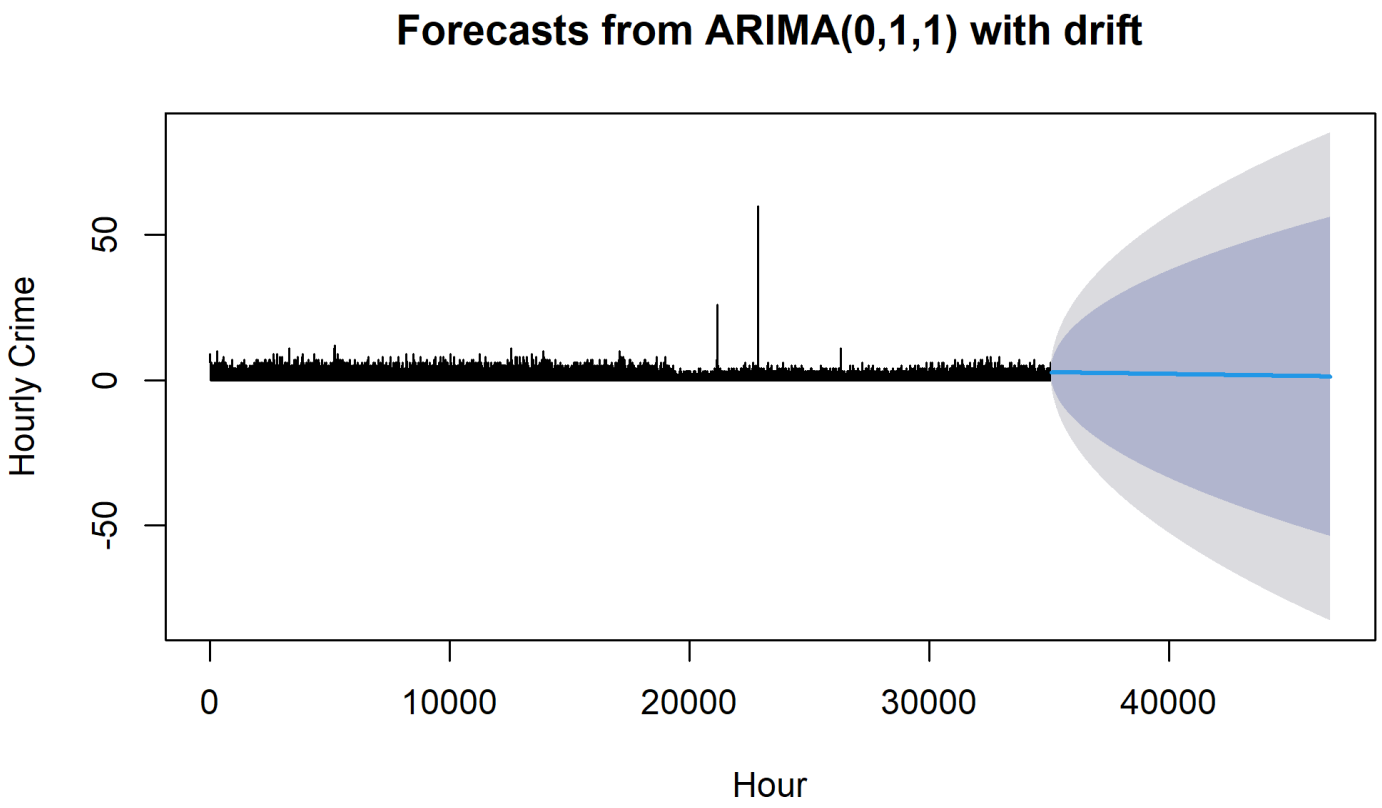
APPENDIX: NON-SEASONAL ARIMA RESULTS

```
Series: ch_train_ts
ARIMA(0,1,1) with drift

Coefficients:
          ma1    drift
      -0.7107  -0.0001
s.e.   0.0056   0.0021

sigma^2 = 1.89:  log likelihood = -60909.54
AIC=121825.1   AICc=121825.1   BIC=121850.5
```



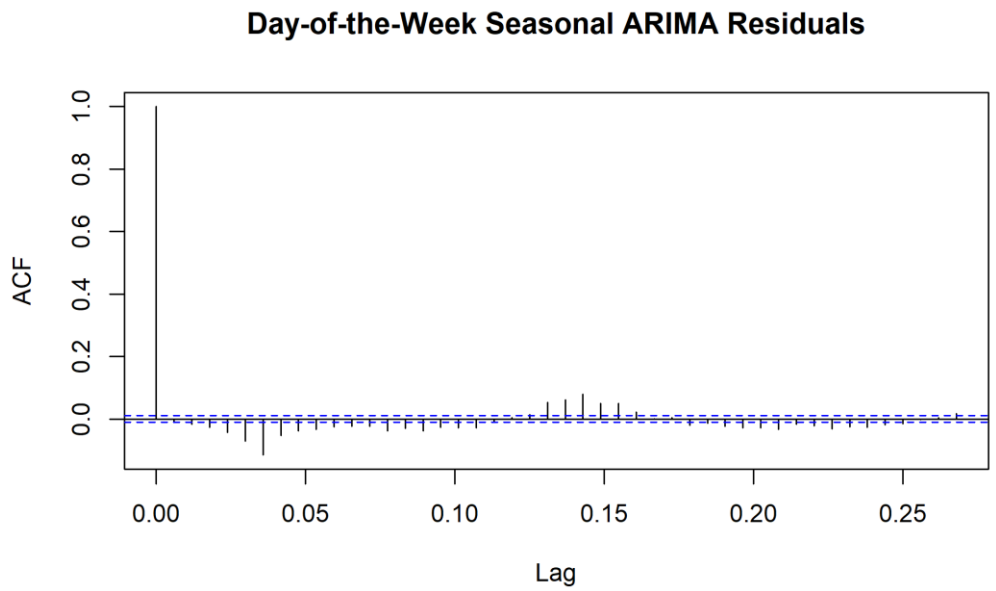


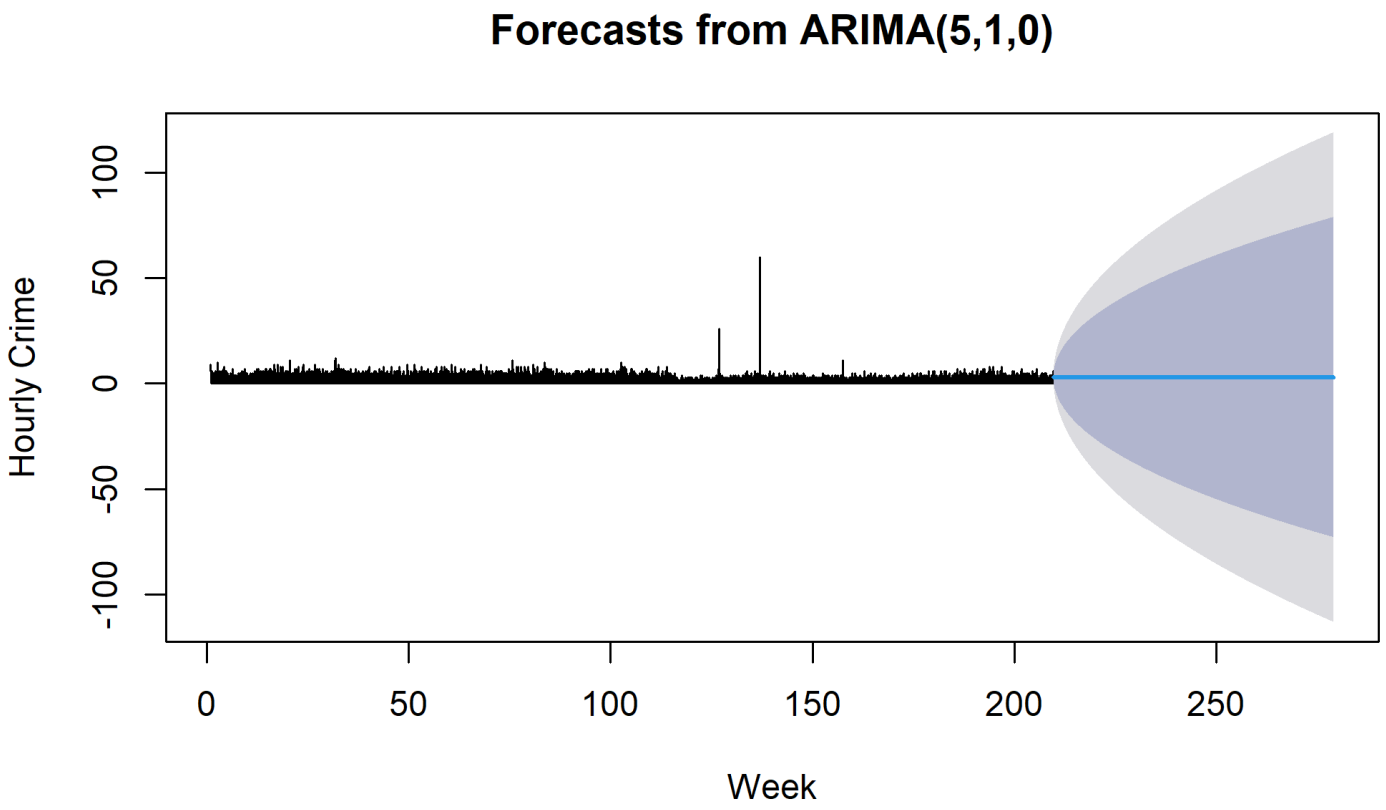
APPENDIX: DAY-OF-THE-WEEK SEASONAL ARIMA RESULTS

```
Series: ch_train_ts_weekly
ARIMA(5,1,0)

Coefficients:
          ar1      ar2      ar3      ar4      ar5
      -0.6216  -0.3823  -0.2430  -0.1699  -0.1052
s.e.   0.0053   0.0062   0.0064   0.0062   0.0053

sigma^2 = 1.92:  log likelihood = -61187.5
AIC=122387   AICc=122387   BIC=122437.8
```



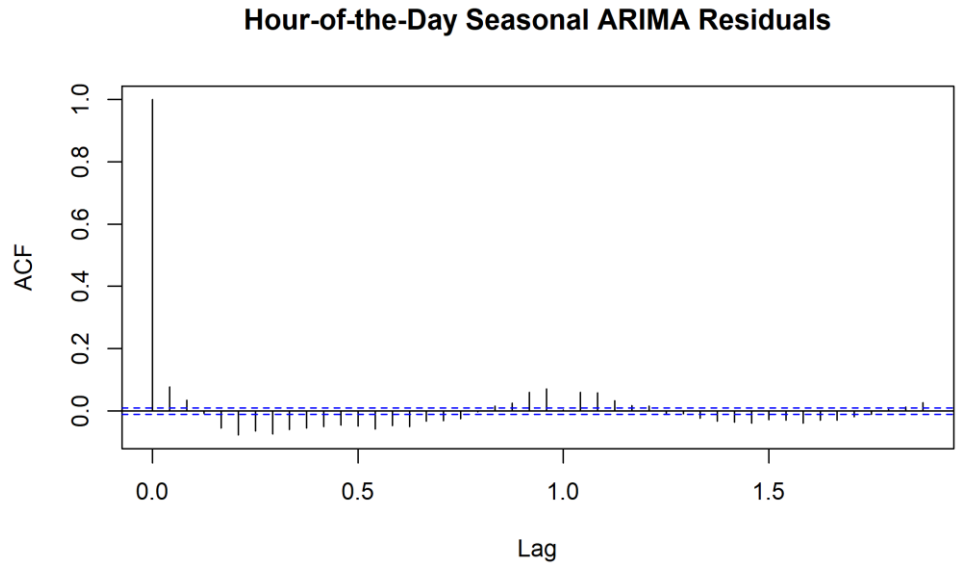


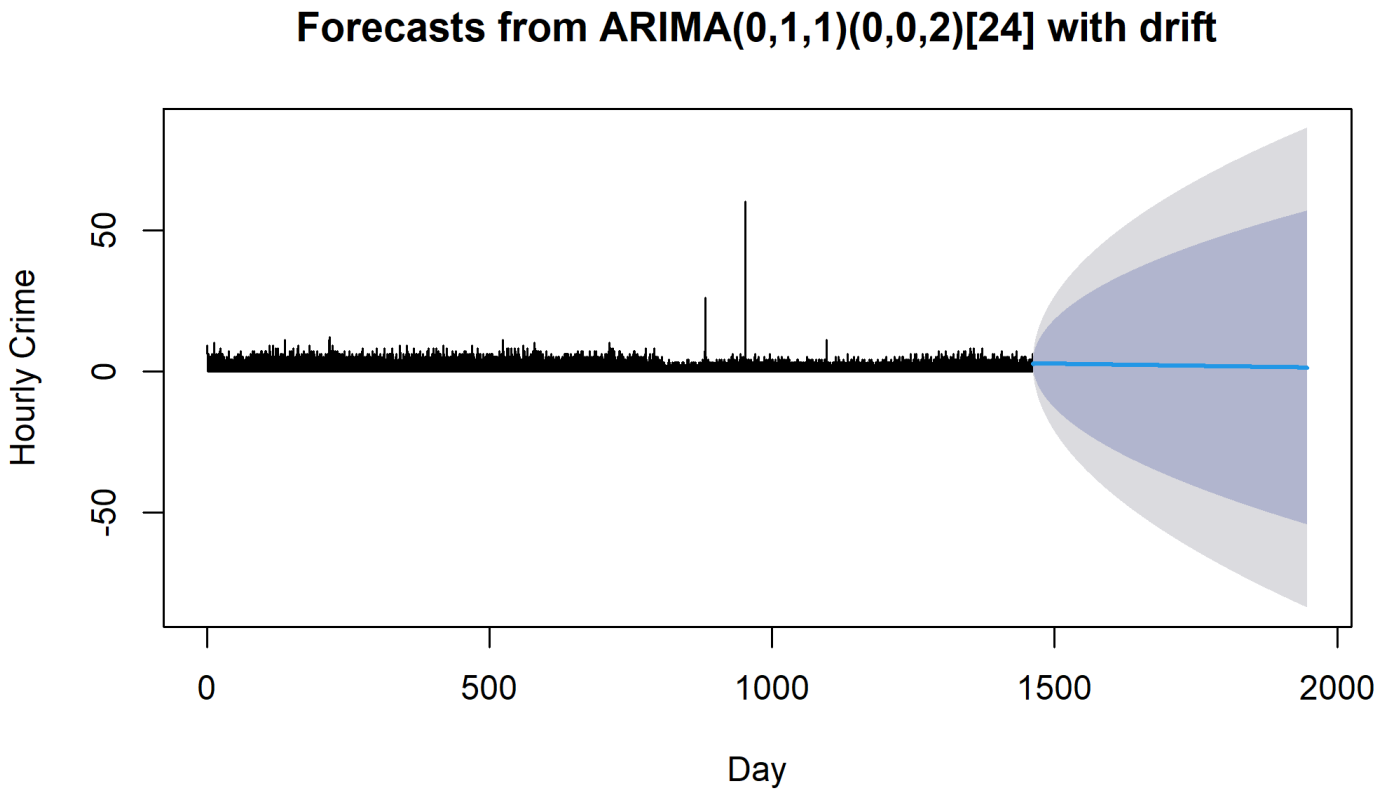
APPENDIX: HOUR-OF-THE-DAY SEASONAL ARIMA RESULTS

```
Series: ch_train_ts_daily
ARIMA(0,1,1)(0,0,2)[24] with drift

Coefficients:
          ma1    sma1    sma2    drift
      -0.7455  0.0904  0.0662  -0.0001
s.e.   0.0060  0.0055  0.0052   0.0021

sigma^2 = 1.865:  log likelihood = -60678.84
AIC=121367.7   AICc=121367.7   BIC=121410
```





APPENDIX: PROPHET RESULTS