Forecasting Robbery in Chicago

Team #5: Irsa, Ryoya, Taka, Sarah

Time Series Analysis and Forecasting

Our Team

Ryoya Hashimoto

Irsa Ashraf

Taka Nitta

Sarah Lueling

Today's Agenda

Introduction

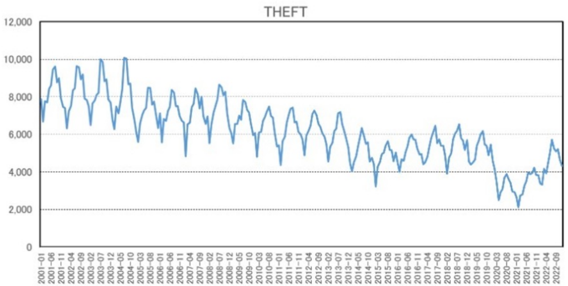
EDA

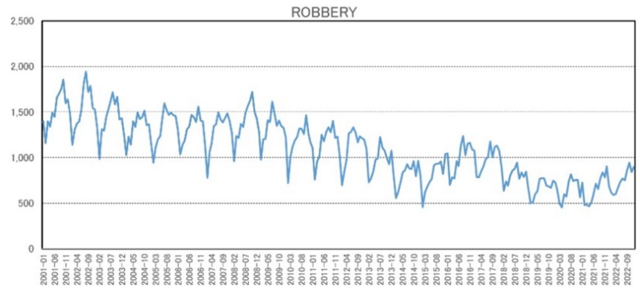
Models and Comparison

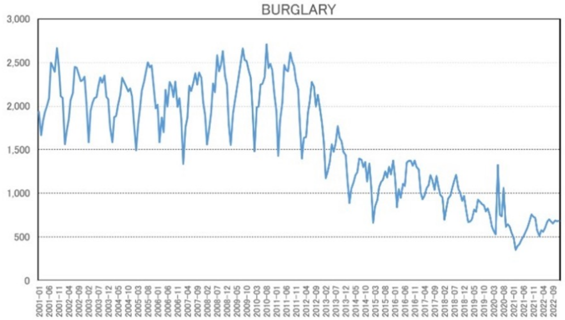
- ARIMA
- Regression with ARIMA Errors
- VAR
- RNN

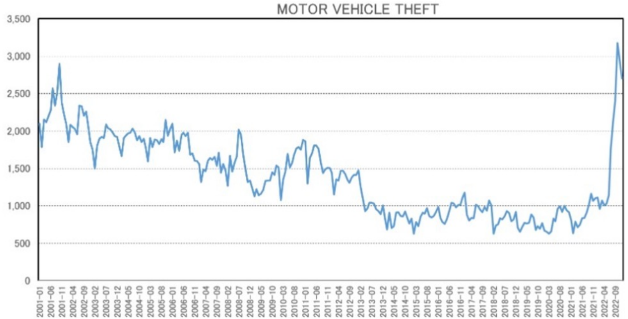
Future steps

Why did we select robbery?









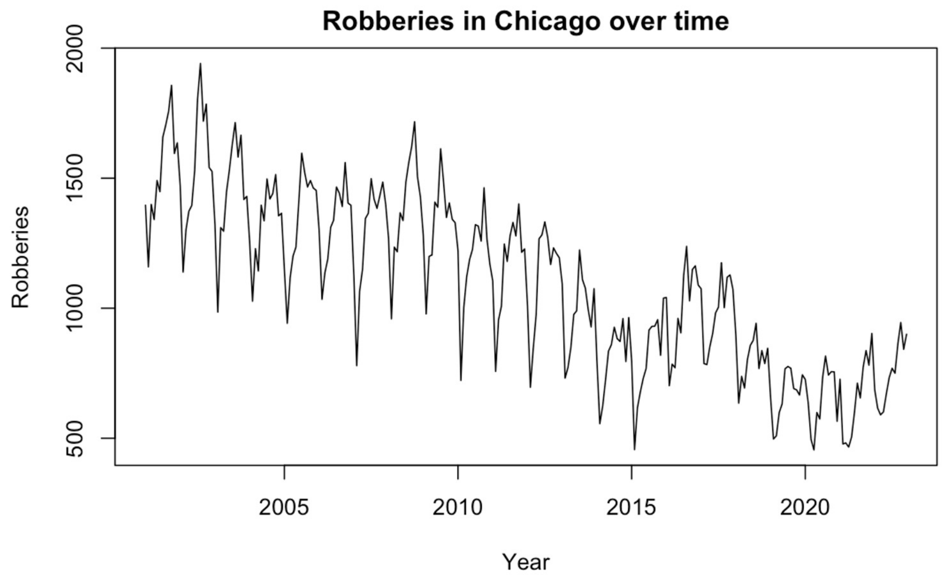
- **Theft/Burglary:** Dramatic change in seasonal pattern since the pandemic
- **Motor vehicle theft:** Huge spike in 2022
- **Robbery:** Relatively stable and unbiased

Source: Chicago Data Portal

Dataset

Data	Source	Note
Robbery cases	Chicago Data Portal	-
Unemployment rate	Bureau of Labor Statistics	Chicago-Naperville-Arlington Heights IL Metropolitan Division
Average temperature	National Oceanic and Atmospheric Administration	-
Average precipitation		-
Average snowfall		-
Rainy days		Count days of > 0.1 inches
Snowy days		Count days of > 0.1 inches
Period		
Training Set: 2001 Jan - 2021 Dec Test Set: 2022 Jan - 2022 Dec Frequency: Monthly		

Analyzing Dataset



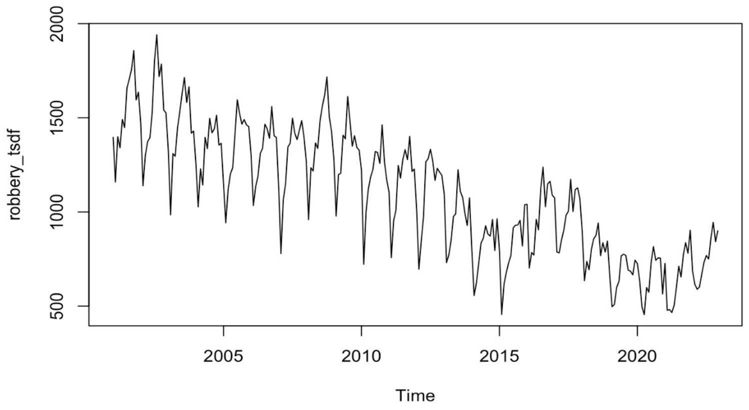
- Downwards trend
- Annual seasonality (additive)
- Non-stationary data

Data Transformation

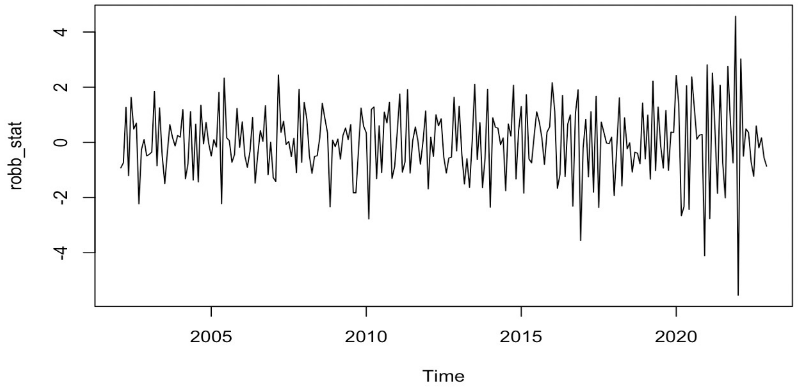
To transform to stationary data, we need to apply the following:

1. Box-cox transformation (lambda=0.3580306)
2. First order and seasonal differencing

Original Dataset



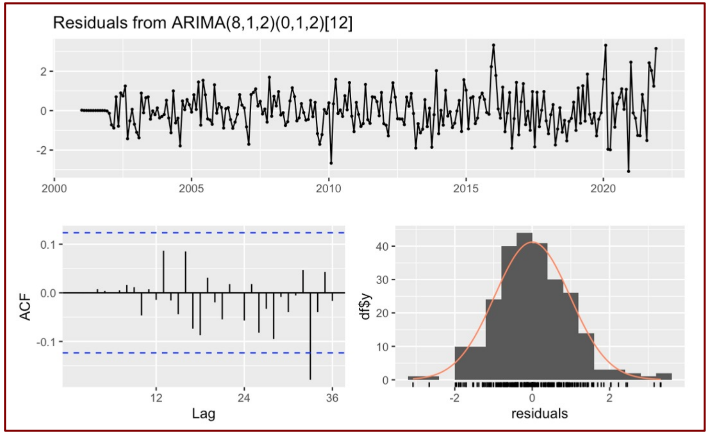
Stationary Dataset



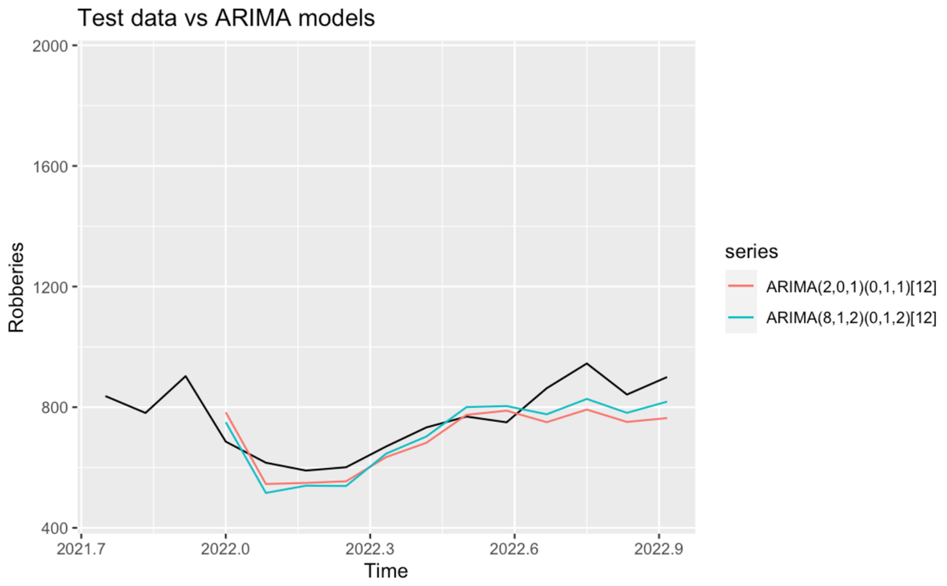
sArima Model

Best manual model:
ARIMA(4,1,2)(2,1,3)[12]

Ljung-Box p=0.20



sARIMA Forecast



ARIMA (2,0,1)(0,1,1)
RMSE = 84.85595,
MAE = 73.17060

ARIMA (8,1,2)(0,1,2)
RMSE = 74.37062,
MAE = 67.13100

Takeaway

- Best **ARIMA (8,1,2)(0,1,2)**
- Manual model performs better
- Auto.generated model not always best model

Regression with ARIMA Errors

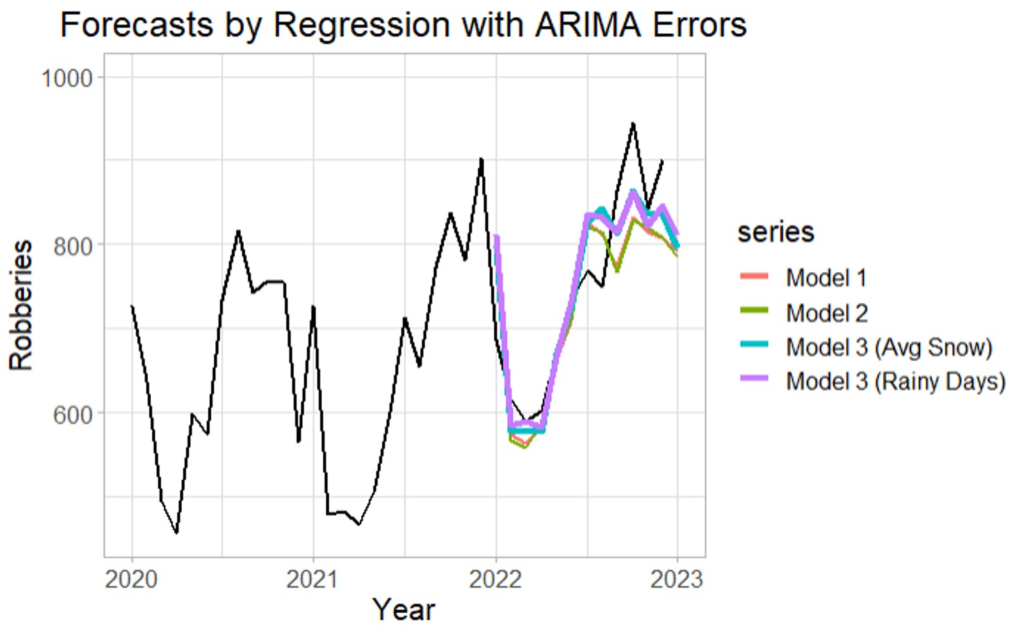
$$y_t = \beta_0 + \beta_1 x_{1,t} + \cdots + \beta_k x_{k,t} + n_t,$$

Where: n_t follows an ARIMA model (i.e., $\phi(B)n_t = \theta(B)e_t$)

Source: Lecture 7 Slide 14

- Model : ARIMA(1,0,1)(0,1,1)[12] (Selected by auto.arima)
- Different Combinations of independent variable(s)
- Box-cox transformed all dependent and independent variables

Regression with ARIMA Errors



Source: Chicago Data Portal, Bureau of Labor Statistics, NOAA

Model 1: Average Temperature, Rainy Days, Snowy days, Unemployment Rate

RMSE = 66.54 MAE = 55.96

Model 2: Average Temperature, Precipitation, Snowfall, Unemployment Rate

RMSE = 66.89 MAE = 55.59

Model 3: Pick up One Variable

Best RMSE = 57.76 (Avg Snowfall)

Best MAE = 44.66 (Rainy Days)

VAR Model

- **Variable selection by checking correlations**

	ROBBERY	avg_temp	snowy_days	rainy_days	unemp
ROBBERY	1.00000000	0.22921873	-0.23893068	-0.0863701092	0.0755475288
avg_temp	0.22921873	1.00000000	-0.79305332	0.2542184083	0.0473275317
snowy_days	-0.23893068	-0.79305332	1.00000000	-0.1212782183	0.0128097485
rainy_days	-0.08637011	0.25421841	-0.12127822	1.0000000000	0.0007460362
unemp	0.07554753	0.04732753	0.01280975	0.0007460362	1.0000000000

VAR: Models Implemented

1. **Model 1**: Robbery and Average Temperature
2. **Model 2**: Robbery and Snowy Days
3. **Model 3**: Robbery and Unemployment
4. **Model 4**: Robbery, Average Temperature and Unemployment
5. **Model 5:** Robbery, Snowy Days and Unemployment

Model Results

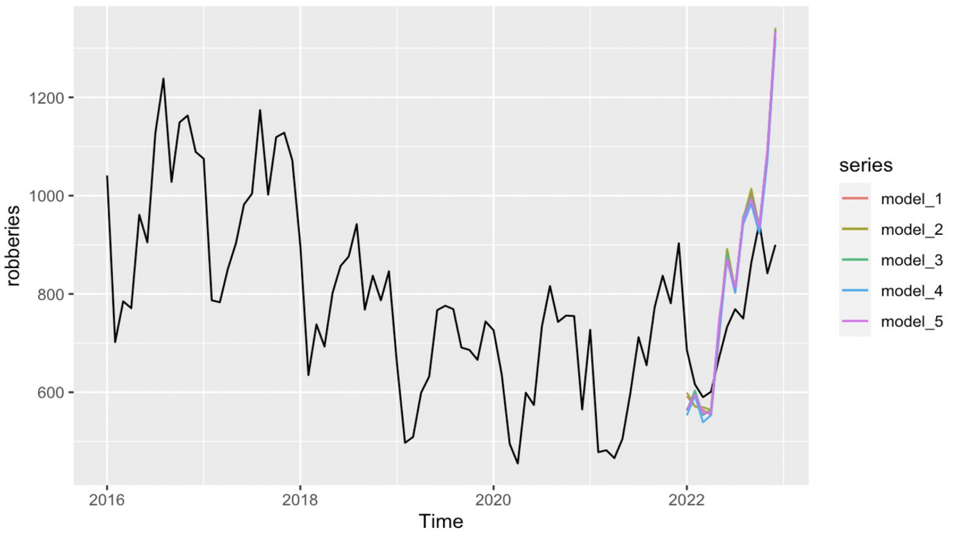
Model 1: Robbery, Avg temp
Rmse: 167.3
Mae: 121.9

Model 3: Robbery, Unemployment
Rmse: 172.3
Mae: 123.8

Model 2: Robbery, Snowy Days
Rmse: 173.9
Mae: 125.0

Model 4: Robbery, Avg temp, Unemployment
Rmse: 164.8
Mae: 121.8

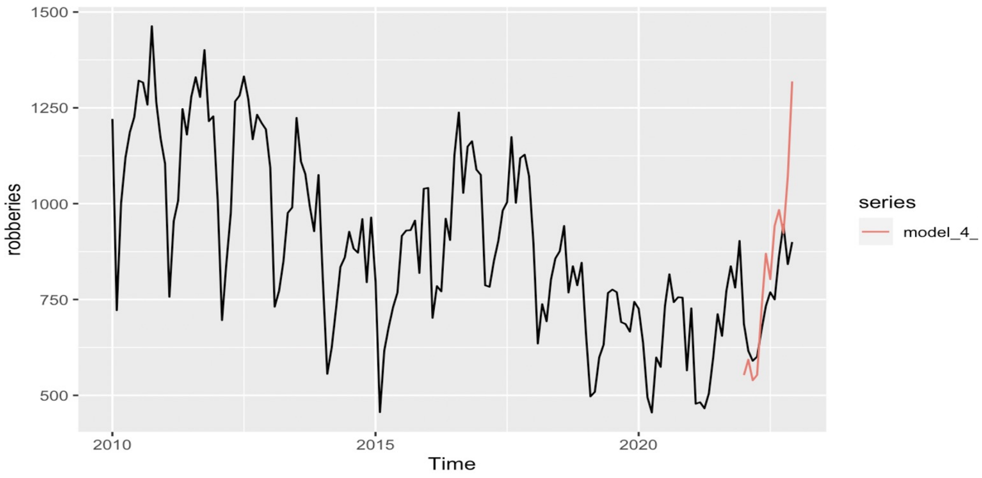
Model 5: Robbery, Snowy Days, Unemployment
Rmse: 171.8
Mae: 125.1

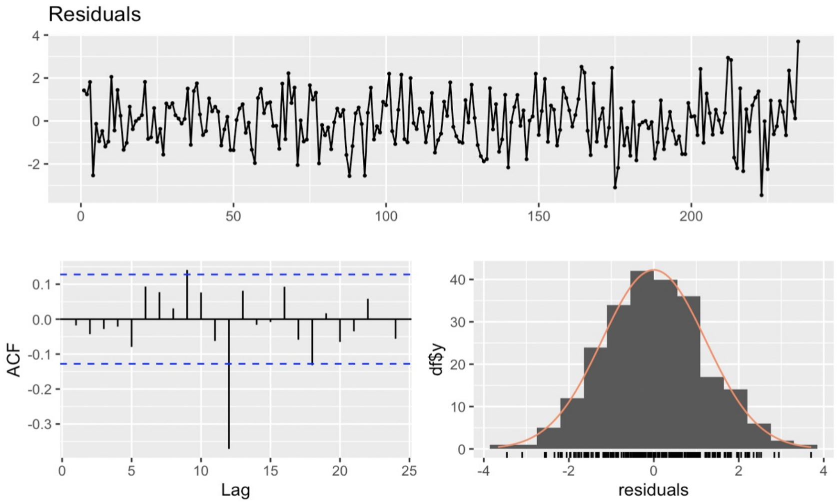


VAR: Best Model

Best Model: Model 4
Variables: Robberies, Average Temperature, Unemployment
p = 4

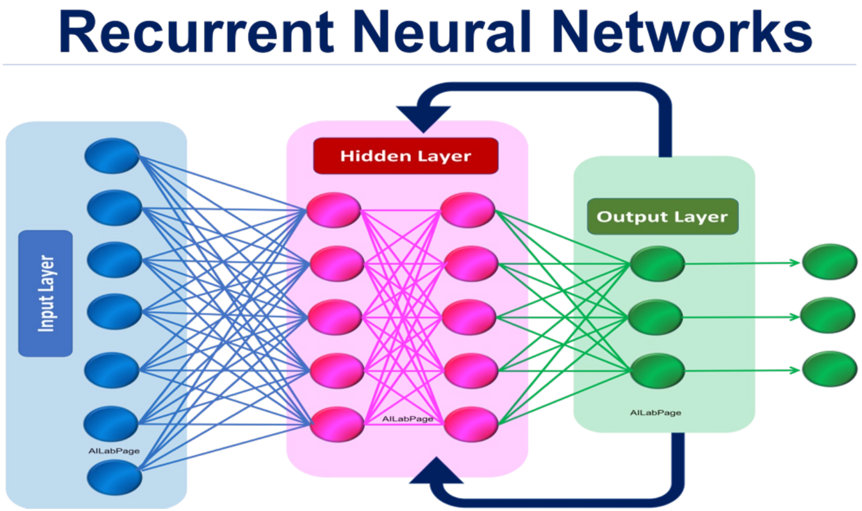
RMSE
Robberies: 164.8
Avg Temp: 9.9
Unemployment: 1.3



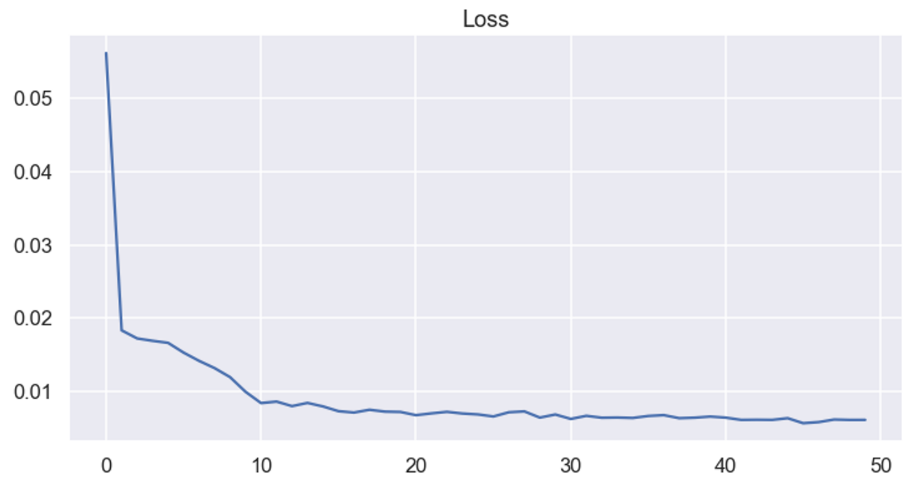


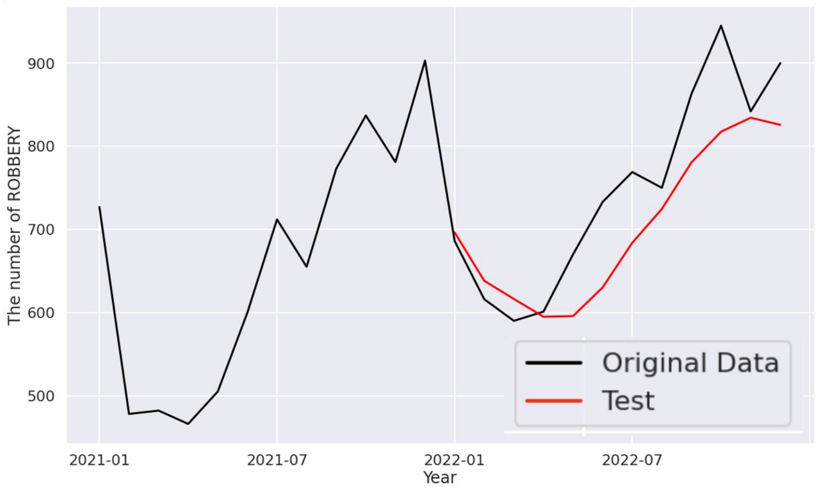
Recurrent Neural Network (LSTM)

- RNN produce predictive results in sequential data.
- Compared to conventional RNN, LSTM is able to retain past information even longer.



Recurrent Neural Network (LSTM)





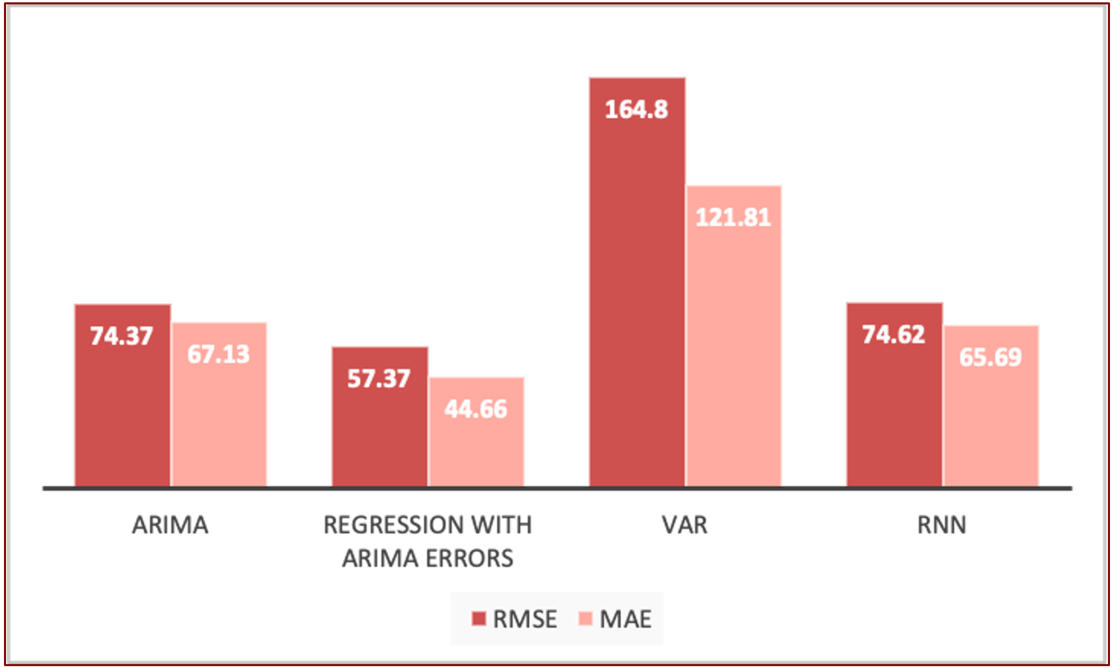
Epoch Size: 50
Look Back Period: 12

RMSE = 74.600, MAE = 65.692

Comparison of each model

- Lowest RMSE: Regression with ARIMA errors - we include relevant independent variables

- Similar RMSE for exponential smoothing, sARIMA and RNN



Conclusion & Next Steps

- **Conclusion: Simpler is better**
  - Regression with ARIMA Errors have better score
  - Complex models (VAR and RNN) have larger errors

- **Next Steps**
  - Intervention analysis with shocks such as Covid
  - Include other criminal data such as Burglary
  - Try Expanding Window or Sliding Window

Questions?