Crime Analysis in Austin, Tx

By: Anita Durg
07/13/2023

Objective

- Observe the crime trend in downtown area of Austin, Tx

- Forecast the crimes for the next 14 days and help APD sectors

Data Science Pipeline

- Data acquisition
- Data wrangling
- Exploratory Data Analysis(EDA)
- Baseline Modeling(Preprocessing and Training)
- Modeling (Prophet and Auto Arima(pmdarima))

Data Source

- https://data.austintexas.gov/Public-Safety/Crime-Reports/fdj4-gpfu

Glimpse of the Dataset - crime_report

|	incident_report_number	crime_type	ucr_code	family_violence	occ_date_time	occ_date	occ_time	rep_date_time	rep_date	rep_time
0	20121171927	RECKLESS DAMAGE	1401	N	2012-04-26T22:34:00.000	2012-04-26T00:00:00.000	2234	2012-04-26T23:04:00.000	2012-04-26T00:00:00.000	2304
1	20072871892	FAMILY DISTURBANCE	3400	N	2007-10-14T20:01:00.000	2007-10-14T00:00:00.000	2001	2007-10-14T20:01:00.000	2007-10-14T00:00:00.000	2001
2	2006471156	FAMILY DISTURBANCE	3400	N	2006-02-16T14:25:00.000	2006-02-16T00:00:00.000	1425	2006-02-16T14:25:00.000	2006-02-16T00:00:00.000	1425
3	20045017276	IDENTITY THEFT	4022	N	2003-07-31T12:00:00.000	2003-07-31T00:00:00.000	1200	2004-04-14T09:45:00.000	2004-04-14T00:00:00.000	945
4	20173300229	FAMILY DISTURBANCE	3400	N	2017-11-26T07:43:00.000	2017-11-26T00:00:00.000	743	2017-11-26T07:43:00.000	2017-11-26T00:00:00.000	743

5 rows × 33 columns

Dataset Information

- Using API request, downloaded 2 million records.
- # of rows = 2000000
- # columns = 33
- ['incident_report_number', 'crime_type', 'ucr_code', 'family_violence',
- 'location_type', 'address', 'clearance_date', 'clearance_status',
- 'zip_code', 'council_district', 'sector', 'district', 'pra',
- 'census_tract', 'ucr_category', 'category_description', 'x_coordinate',
- 'y_coordinate', 'latitude', 'longitude', 'location',
- 'occurred_date_time', 'reported_date_time'],

Missing Columns and Clean up Data
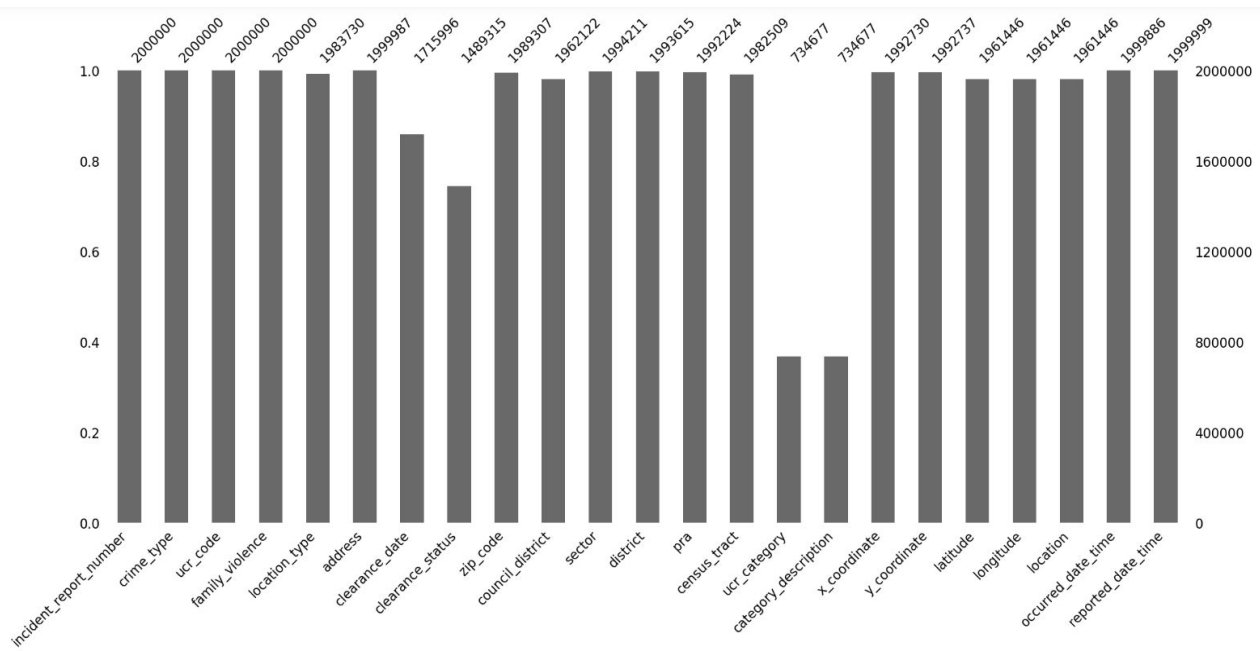
Dimensions associated

Determined the dimensions associated with this project:

- Council District
- Sectors(Austin Police Department) sectors
- Zip Codes
- Crime Types

Identify the target and its type (categorical/non-categorical)

- Ucr_code
- Crime_type
- Family_violance

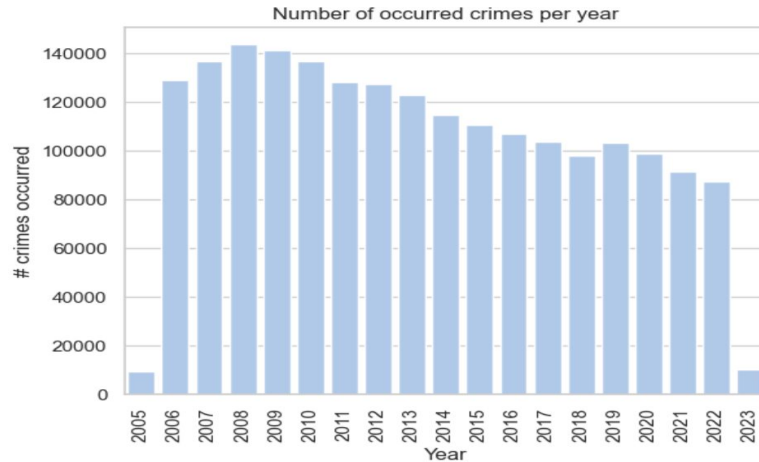# Identify the resolution of Time Series

- What should be the resolution of the Times series?
- What is the forecast horizon?
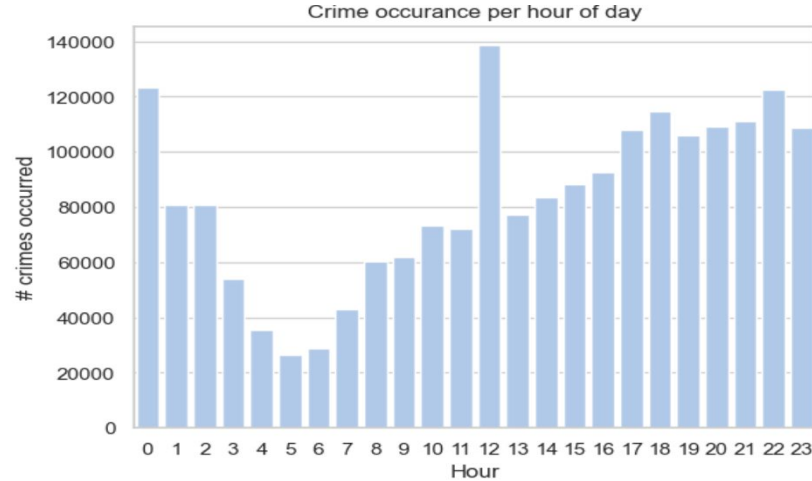- Any Exogenous Variables noticed?

EDA Process

https://nbviewer.org/github/anidurg/Springboard_2022/blob/main/Capstone3/Notebook/CP3_EDA.ipynb

Crimes Trend over the years (2006-2022)

Number of occurred crimes per year

A decrease in the number of crimes.

Hourly Crime pattern

We observe that crimes seem to occur more in the afternoon and midnight

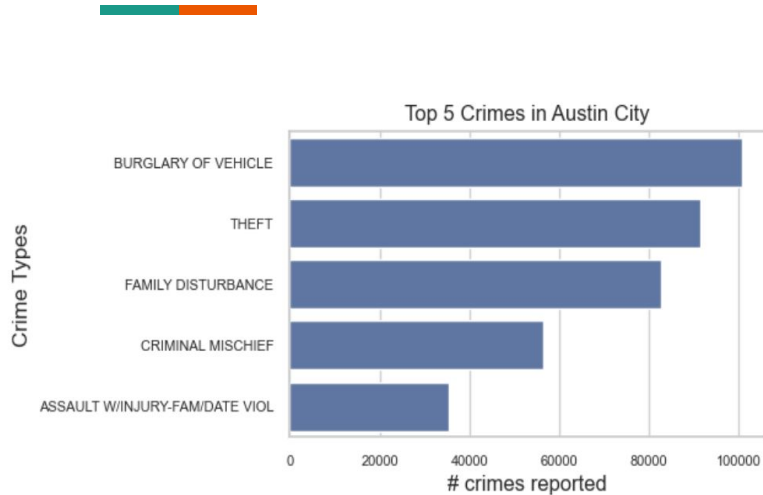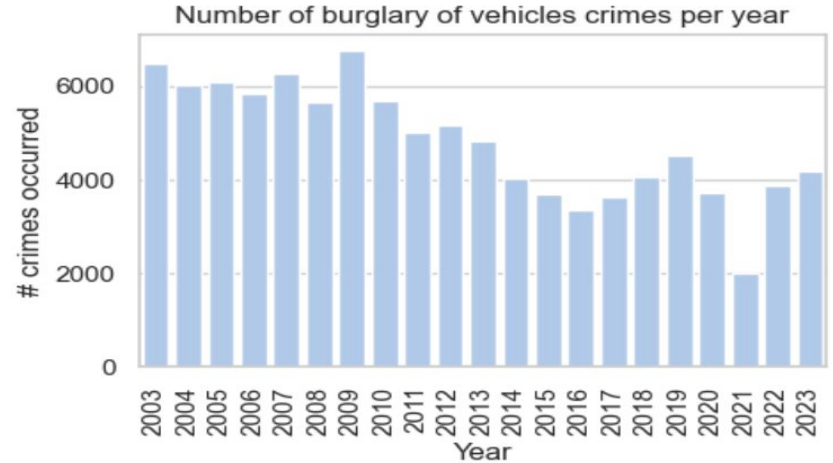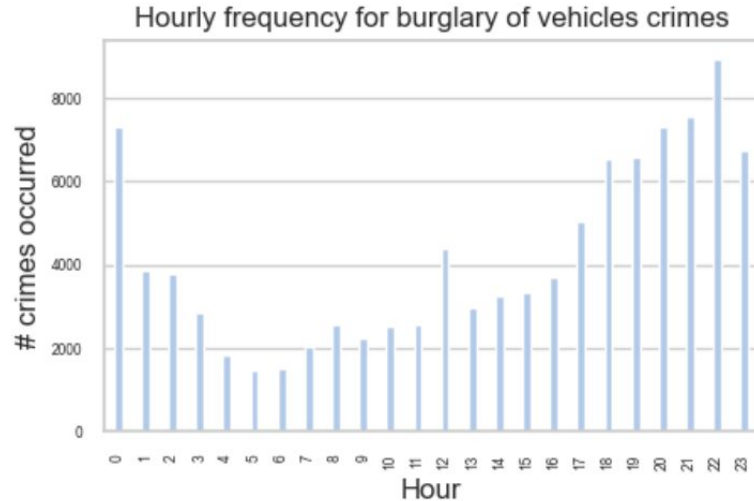# Analysis for Vehicle Burglary Crime by Year

Fig 1

Fig 2

Burglary of vehicles is the top most crime in Austin. Fig 2. Shows the Year 2009 reported the most burglary of vehicles crime. 2021 was the lowest. Could it be due to Covid? From 2021 to 2022, we see a big jump for that crime.

Hour-wise Crime Counts

Highest crime occurs 10 pm. Noon (12 pm) the crime counts are significantly high. Most of the crimes happens after 5pm till midnight.
Lowest number of crime reported is at 5am

Analysis for Vehicle Burglary Crime by Month

Monthly frequency for burglary of vehicles crimes
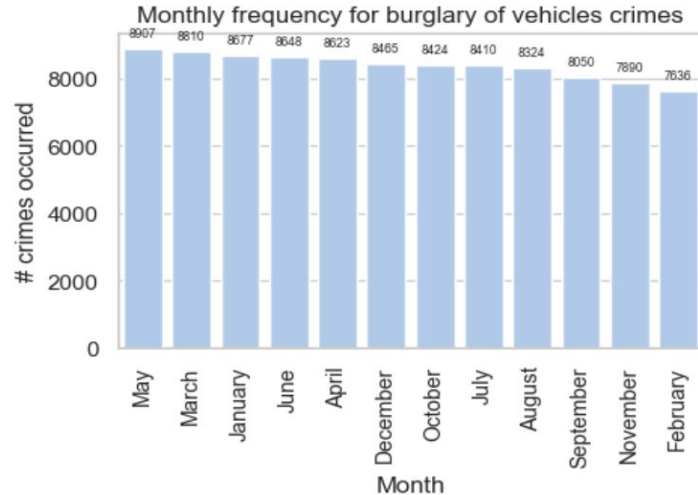
In 2009, since vehicle burglary crime rate is so high, wonder which month produced high vehicle burglary? Let us find out

Monthly frequency(in 2009) for burglary of vehicles crimes
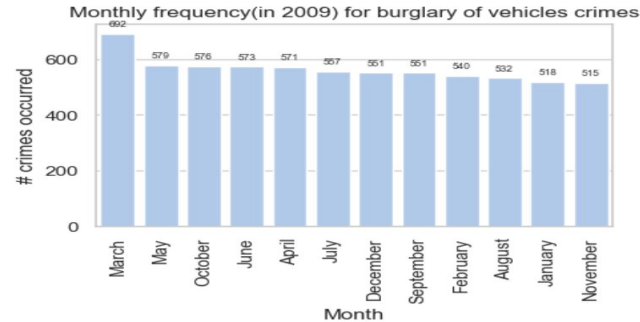
May month seems to be very active for crimes February is low.

March 2009 had the highest Vehicle Burglaries

Visualization of Crime Count based on date features and per district and sector
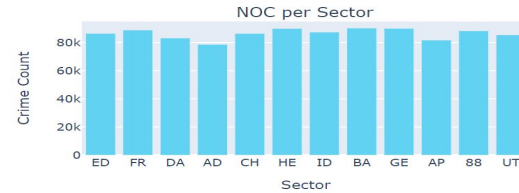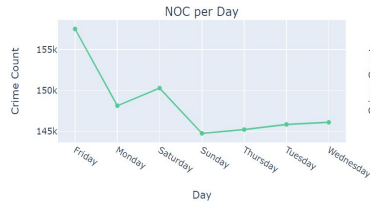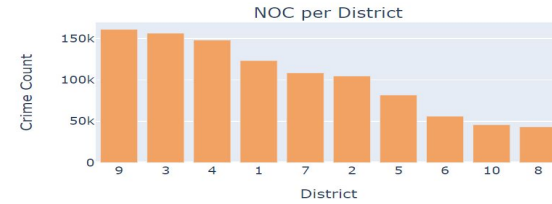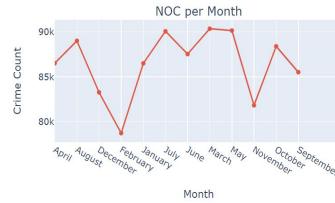
Which Council district and sector have high crime rate?

https://www.austintexas.gov/edims/document.cfm?id=168703
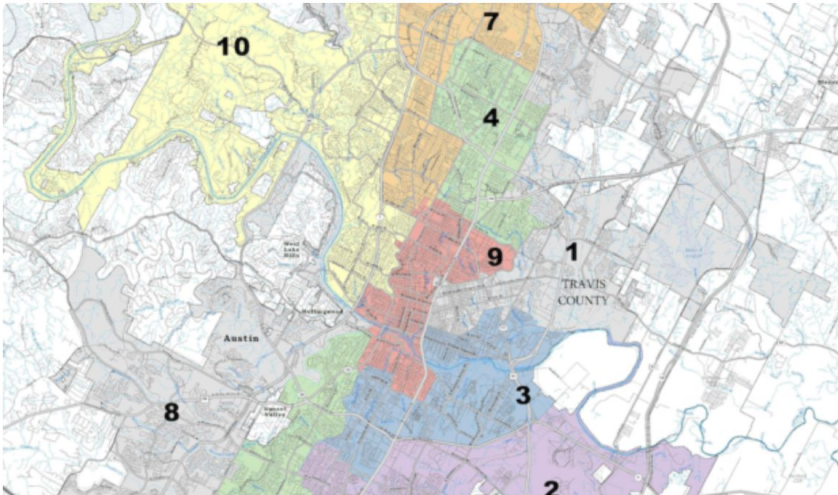
Top 10 Crime types and APD sectors

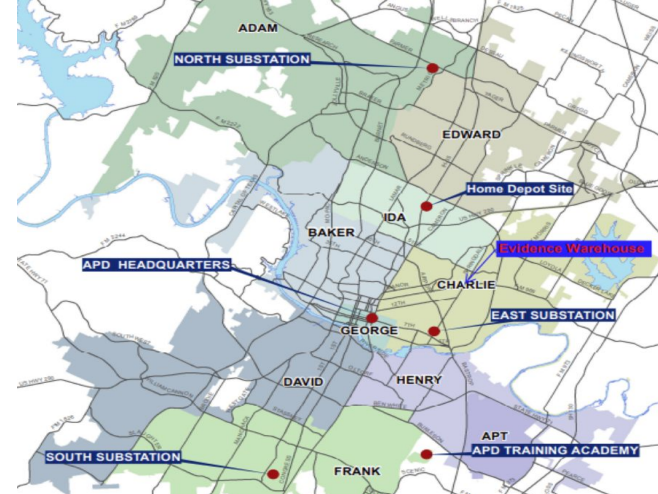|	
THEFT	16409
BURGLARY OF VEHICLE	13109
ASSAULT BY CONTACT	9010
CRIMINAL MISCHIEF	7685
DWI	7443
ASSAULT WITH INJURY	6965
THEFT OF BICYCLE	6048
FAMILY DISTURBANCE	5909
WARRANT ARREST NON TRAFFIC	5199
DISTURBANCE - OTHER	4332

GE	63827
BA	62136
DA	13813
HE	12781
CH	5937
ID	2290
UT	219
HENRY	1

Name: sector, dtype: int64

EDA on Zip codes - Used Tableau
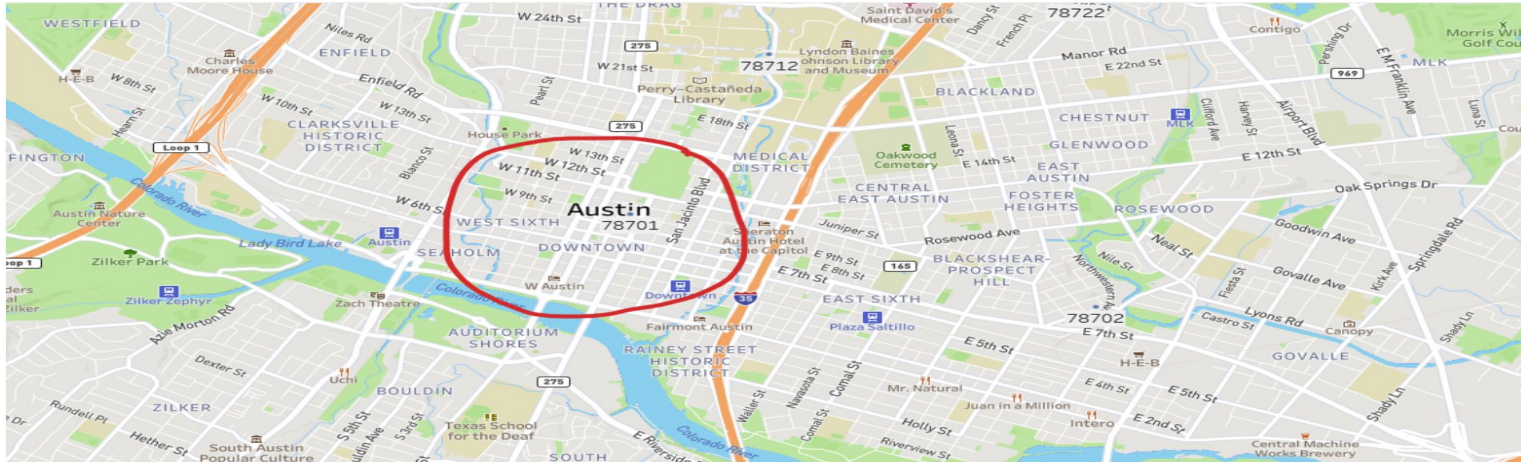
Pre-processing and Training

https://nbviewer.org/github/anidurg/Springboard_2022/blob/main/Capstone3/Notebook/CP3%20DailyTime%20Series%20Preprocessing%20and%20Baseline%20Modeling-Final.ipynb

- Goal is to build one Time Series Forecast Model
- Read the Forecasting Principles and Practice book by Hyndman-Athanasoupoulos

Analyze one Time Series

- Prepared the TS for top ten crimes for APD sectors GEORGE and BAKER
- I subset the data to keep only observations for the year to be greater than 2021
-

occurred_date	crime_count
2022-03-11	13
2022-03-12	8
2022-03-13	23
2022-03-14	16
2022-03-15	16

Tested for stationarity using KPSS and ADF Test and made it stationary by differencing and log transformation. Did not observe any seasonality

ACF and PACF Plots
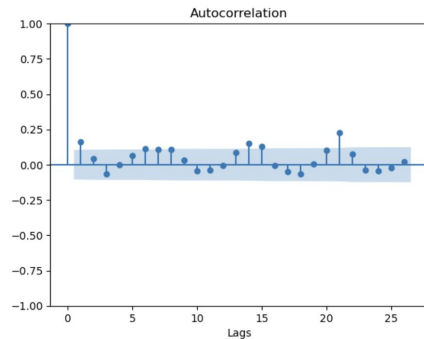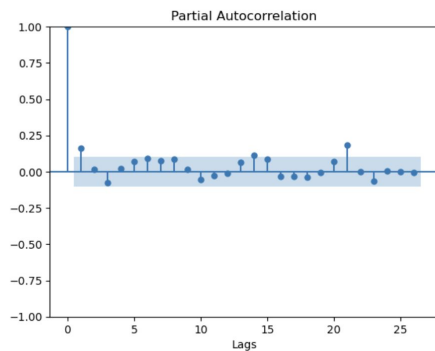
Best model: ARIMA(3,1,1)(0,0,0)[0]
Total fit time: 4.145 seconds

Simple Forecast Methods

- **Naive Method** - considers what happened in the previous period and predicts the same thing will happen again.
- Simple Average
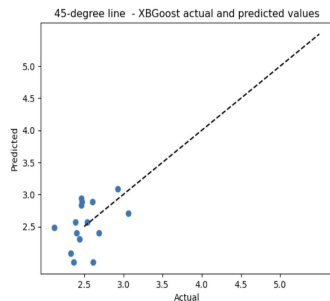- Forecasting with Holt Winter's method
- Uni-step
- XGBoost

45-degree line - XGBoost actual and predicted values

	var1(t-2)	var1(t-1)	var1(t)	var1(t+1)
occurred_date				
2022-03-13	2.564949	2.079442	3.135494	2.772589
2022-03-14	2.079442	3.135494	2.772589	2.772589
2022-03-15	3.135494	2.772589	2.772589	2.564949
2022-03-16	2.772589	2.772589	2.564949	2.564949
2022-03-17	2.772589	2.564949	2.564949	3.091042
...	...	...	...	...
2023-02-19	2.995732	3.218876	3.218876	3.258097
2023-02-20	3.218876	3.218876	3.258097	2.302585
2023-02-21	3.218876	3.258097	2.302585	2.772589
2023-02-22	3.258097	2.302585	2.772589	2.079442
2023-02-23	2.302585	2.772589	2.079442	2.995732

	RMSE	MAPE
Naive	0.58	20.97%
Simple Average	0.40	14.17%
Holt Winter's	0.43	15.46%
Arima	0.43	15.7%
uni-step	0.43	13.3%
XGBoost	0.33	12.0%

Modeling(All Time Series)

https://nbviewer.org/github/anidurg/Springboard_2022/blob/main/Capstone3/Notebook/CP3%20Modeling-Prophet%20and%20PMDARIMA_v_2.ipynb
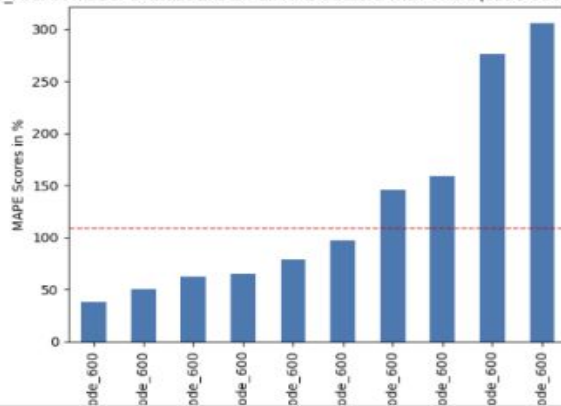
- AUTO_ARIMA
- PROPHET

Twenty Time series -10 APD Sectors and 2 Most common Crimes

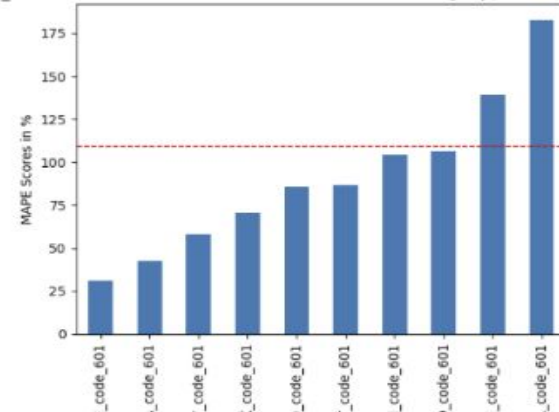Ucr_codes for two most common crimes - 600 and 601

Auto_Arima

PMDARIMA MAPE evaluation for both the crime types (Theft and Burglary) :

Auto_ARIMA model evaluations on all APD sectors with Theft (code 600) Time series
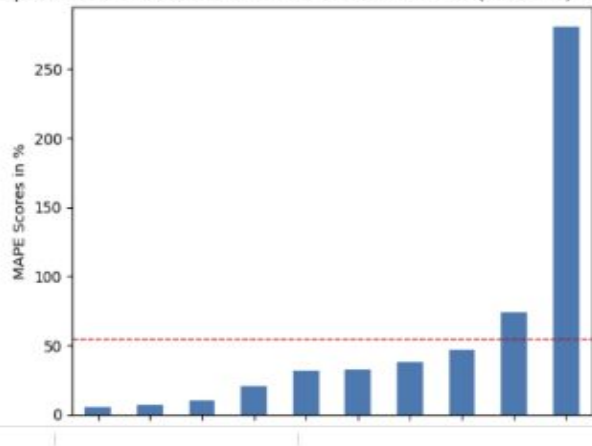
Auto_ARIMA model evaluations on all APD sectors with Burglary(code 601) Time series
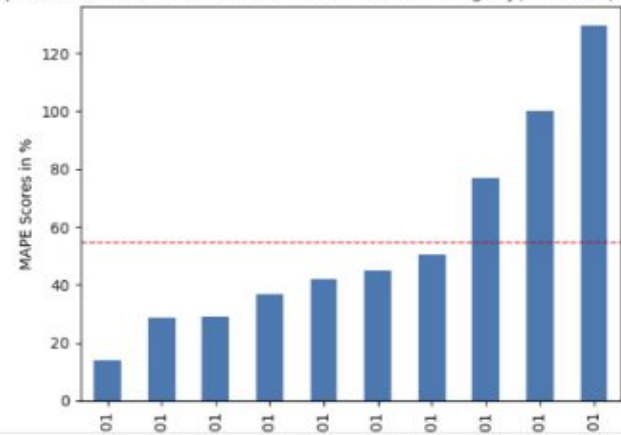
Prophet

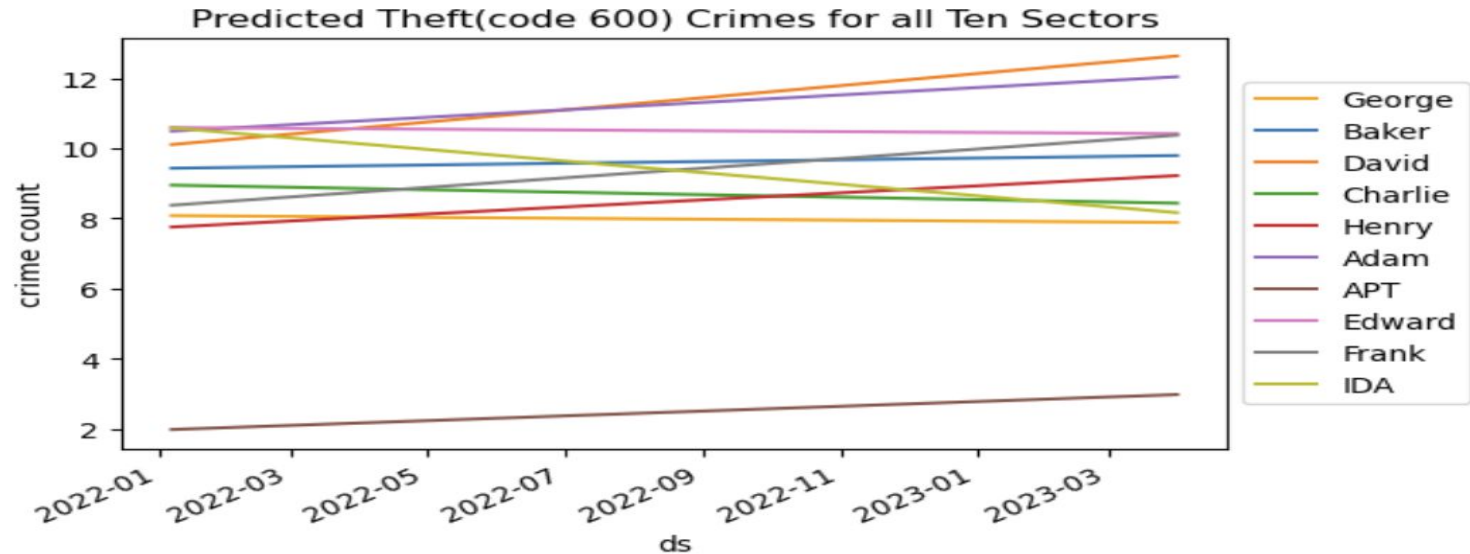PROPHET Evaluations:

Prophet model evaluations on all APD sectors with Theft(code 600) Time series

Prophet model evaluations on all APD sectors with Burglary(code 601) Time series

Prophet Forecast for Thefts for all ten sectors

Predicted Theft(code 600) Crimes for all Ten Sectors

Prophet Forecast for Burglary for all ten sectors

Predicted Burglary(code 601) Crimes for all Ten Sectors
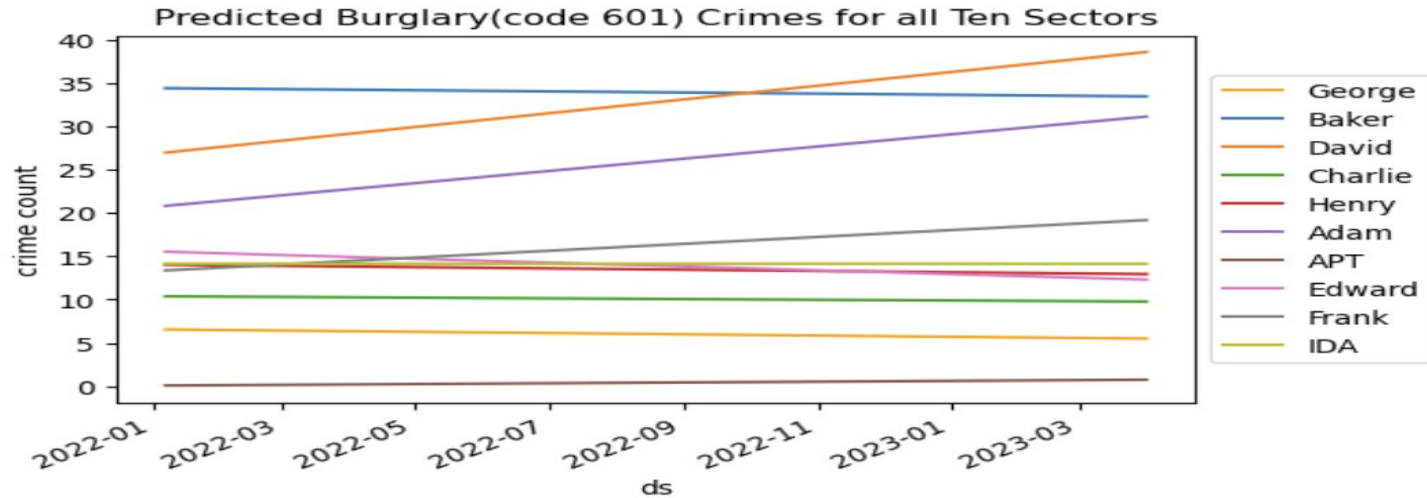
Summary

- Modeling is implemented on twenty time series (ten Austin Police Department Sectors and two major crimes with code of 600 and 601
- code 600 is for Thefts and code 601 for Burglary
- Ten sectors are George, Henry, Baker, David, Charlie,Adam, APT, Edward,Frank, IDA, APT
- Dates will be from 2022 and above
- I did downsampling to make it a weekly time series
- I implemented Pivot method so each time series is a separate column
- I am using two methods for modeling this multiple time series, viz.,Auto_Arima method and Prophet method
- I split the data where cut off date for train set is 2022-12-30
- Forecast for both the methods were evaluated based on MAPE values.
- Auto_arima gave almost the mean of MAPE which is twice the mean of MAPE from Prophet.
- I attempted to display the forecast for all the time series for each crime respectively. Graph shows straigt lines instead of time series wavy lines. Not sure if it is due to the very negligible differences in the series for yhat values.
- I had applied only one time series with XGBoost algorithm. Given more time, I would apply for all the time series and evaluate its performance.
- There is a lot of scope to imporve this project. Given more time, I would have tried to test with more powerful regression algorithms like RandomForestRegressor, LightGBM ensemble algorithm.

Future Work

- I need to work thoroughly on pre-processing and building a baseline model with one of the time series which I implemented along with the remaining 19 time series.
- Would like to attempt a few regression algorithms and see which will forecast the crimes accurately
- I was reading about K-Means Time clustering. I would like to implement that method on this data.
- Deploy the model utilizing Flask.

Acknowledgements

- My mentor A J Sanchez
- Stackoverflow.com
- Austintexas.gov
- YouTube videos