Dallas Crime Trends Forecasting

using sparse vector autoregressive modeling

TEAM MEMBERS:
ABDEL HOMI,
KAUSHIK PASIKANTI,
MATT BROWN,
JOHN KENNEY
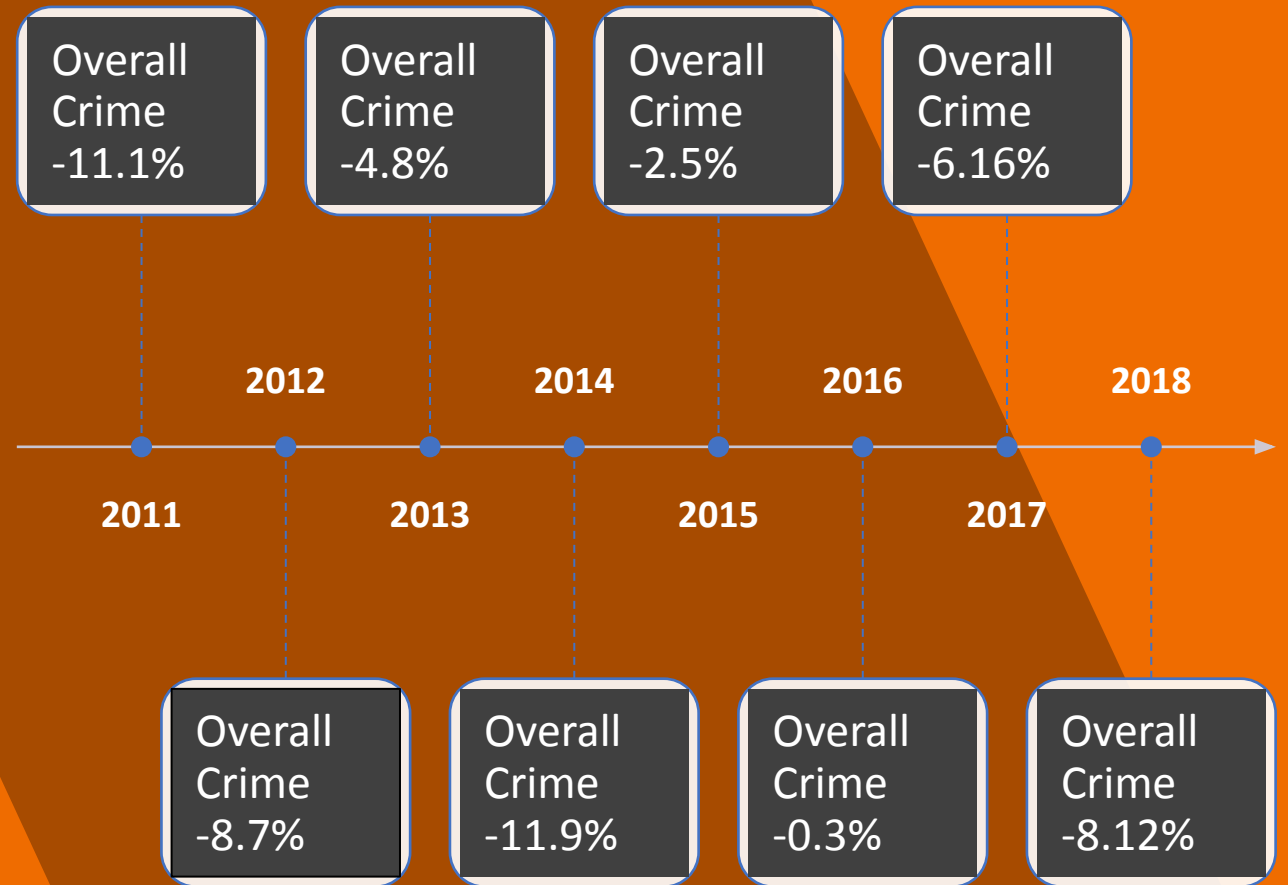
Dr. KEMELLI ESTACIO-HIROMS

Dr. SUNYOUNG SHIN

Subject History

- Started in 2008
- Targeted Action Area Grids (TAAGs) highlight locations with an increased likelihood of criminal activity
- Implemented throughout Dallas County
- More effective use of resources

Dallas TAAG Historical Context

- Targeted Action Area Grids represent:
- 54 distinct TAAG areas currently
- 7% of land area
- ~30% of overall crime

Overall Crime -11.1%	Overall Crime -4.8%	Overall Crime -2.5%	Overall Crime -6.16%

2012   2014   2016   2018

2011   2013   2015   2017

Overall Crime -8.7%	Overall Crime -11.9%	Overall Crime -0.3%	Overall Crime -8.12%

Project Goals

- Compare criminal offenses occurring in TAAGs and non-TAAGs areas

- Forecast crime offenses over time

- Employ a sparse vector auto regression model for analyzing crime time series data

- Perform comprehensive exploratory analysis

- Fully investigate the similarity and dissimilarity between the areas on crime behaviors.

- Model evaluations of Sparse Var w/HLag penalty, regular Var, and AR (1)

Targeted Action Area Grids (TAAGs)

- Crime related information:
  - Arrests
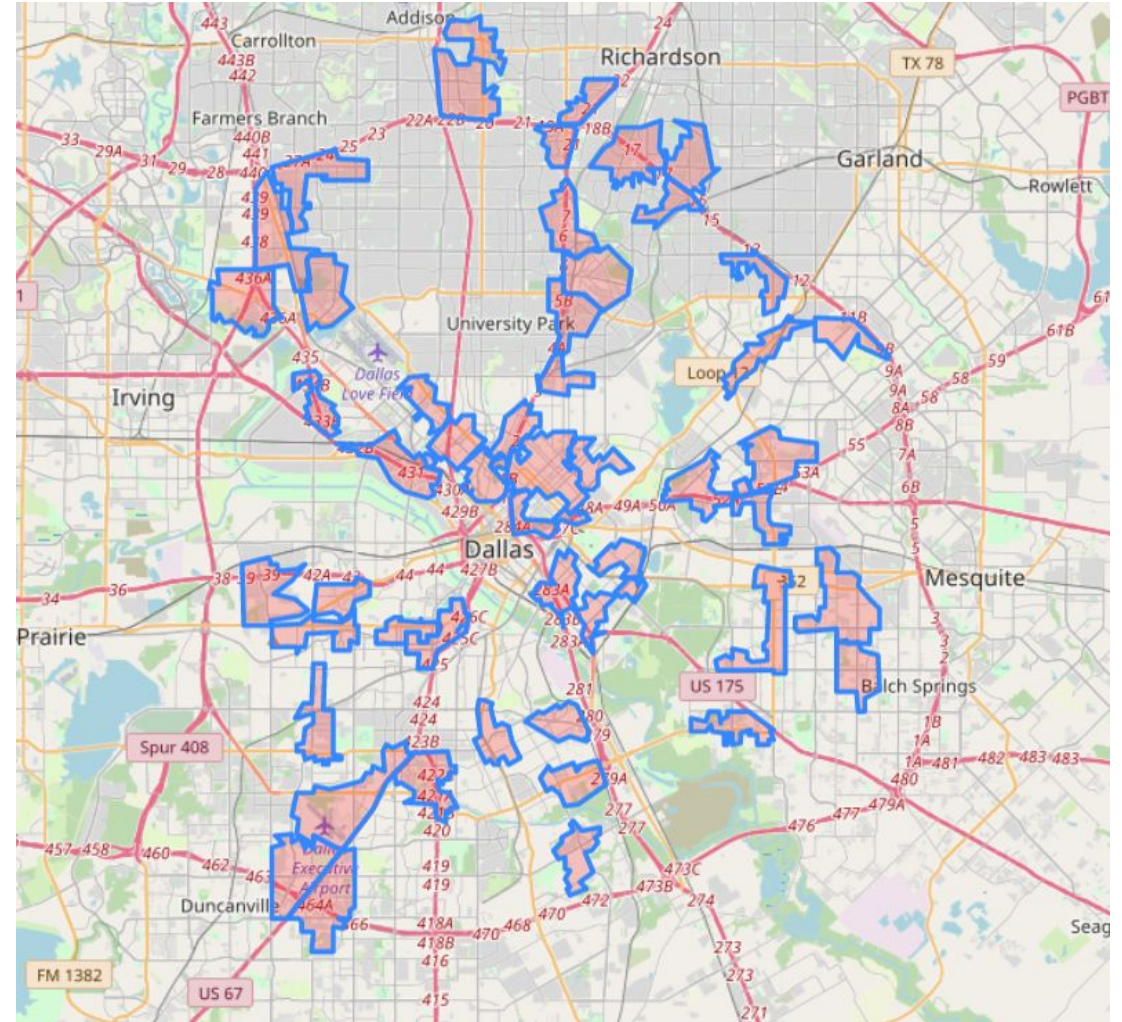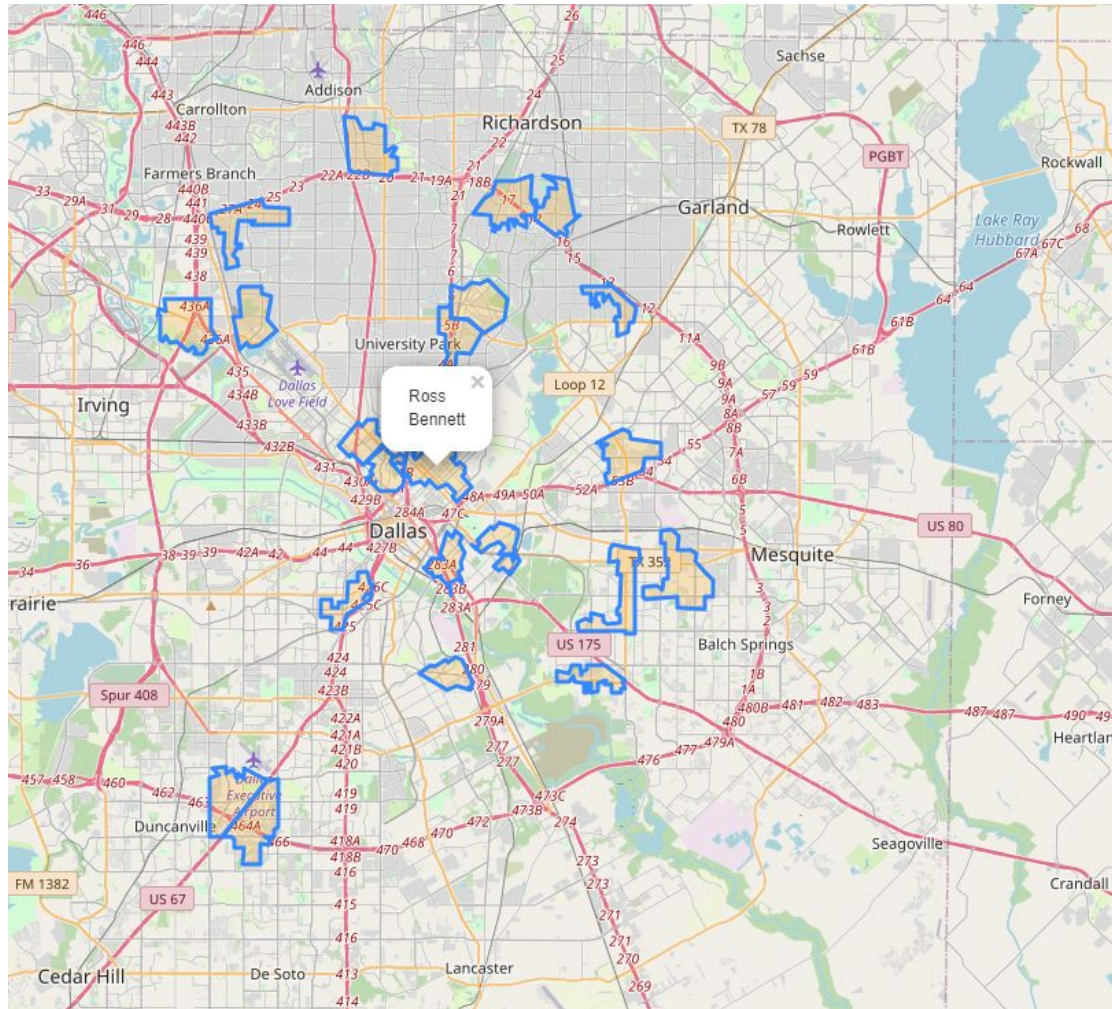 - Calls for service
  - Crime
  - Localize offenders
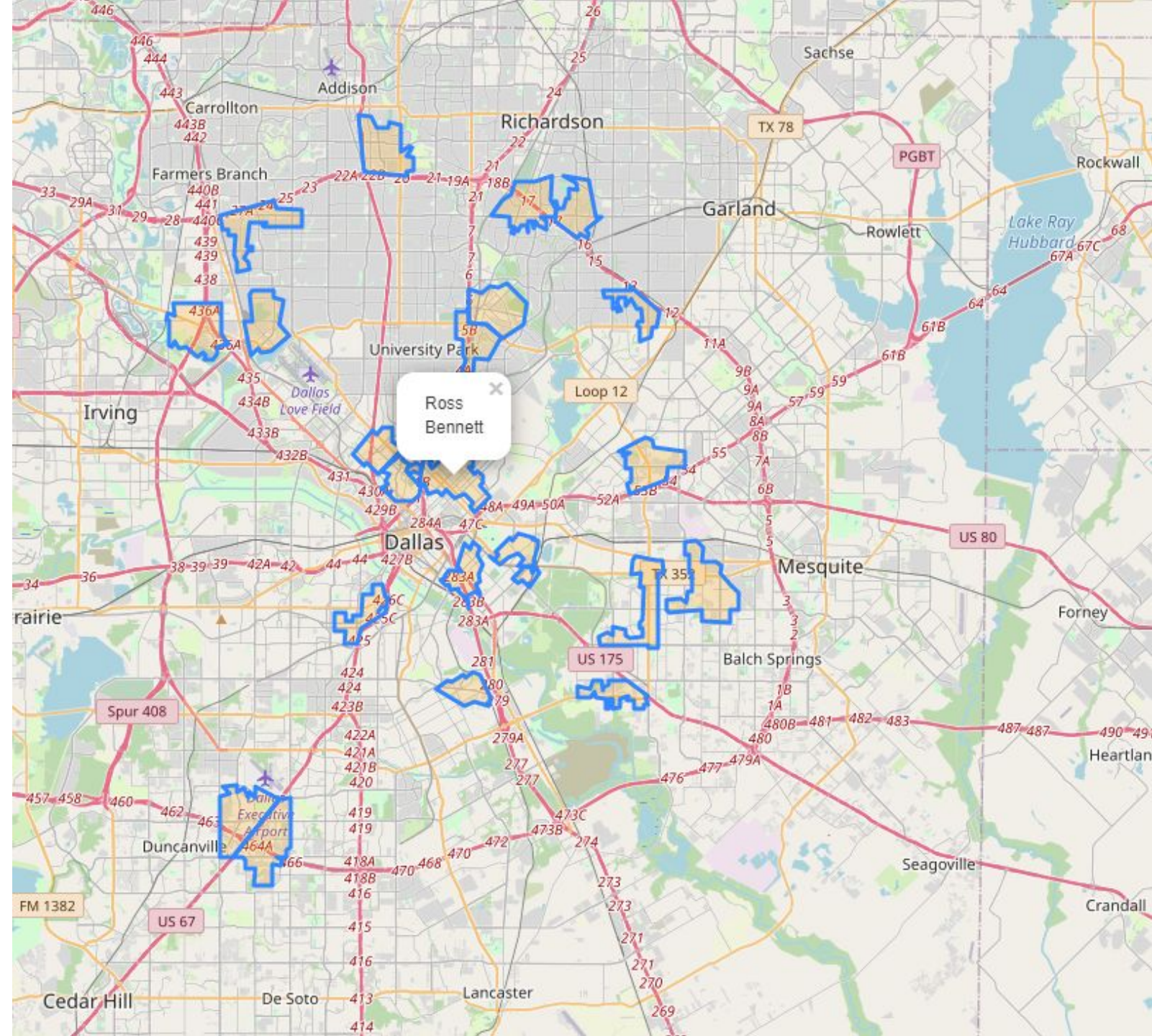
TAAGs: Our First Challenge

```python
print('capstone_initial_taags: ', len(initial_taags)), print('json_updated_taag: ', len(updated_taag))
```

```
capstone_initial_taags:  31
json_updated_taag:  54
```

COMPARING INITIAL AND UPDATED TAAG MAPS

ANOTHER CHALLENGE

- Increase in # of TAAG areas over time as DPD continuously implements this measure

```python
taags_union  = list(set(initial_taags) & set(updated_taag))

taags_union
```

```
['Forest Dennis',
 'Forest Audelia',
 'Ross Bennett',
 'SpringValley Preston',
 'WalnutHill Jupiter',
 'Greenville LBJ',
 'JohnWest Buckner',
 'WebbChapel Timberline',
 'Loop12 JimMiller',
 'Central Southwestern',
 'CampWisdom Chaucer',
 'Hatcher Scyene',
 'McKinney Allen',
 'Five Points',
 'StAugustine Bruton',
 'Wycliff Lemmon',
 'NWHwy WaltonWalker',
 'CampWisdom Westmoreland',
 'Jefferson Corridor',
 'Overton Illinois',
 'JuliusSchepps Central',
 'LakeJune Buckner']
```
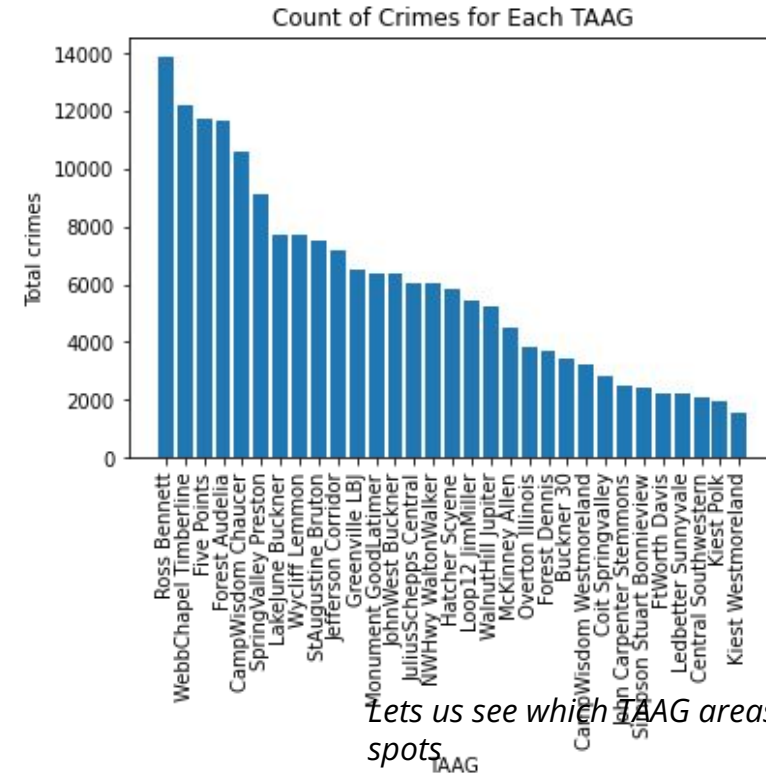
Look at Crime Data Over Time
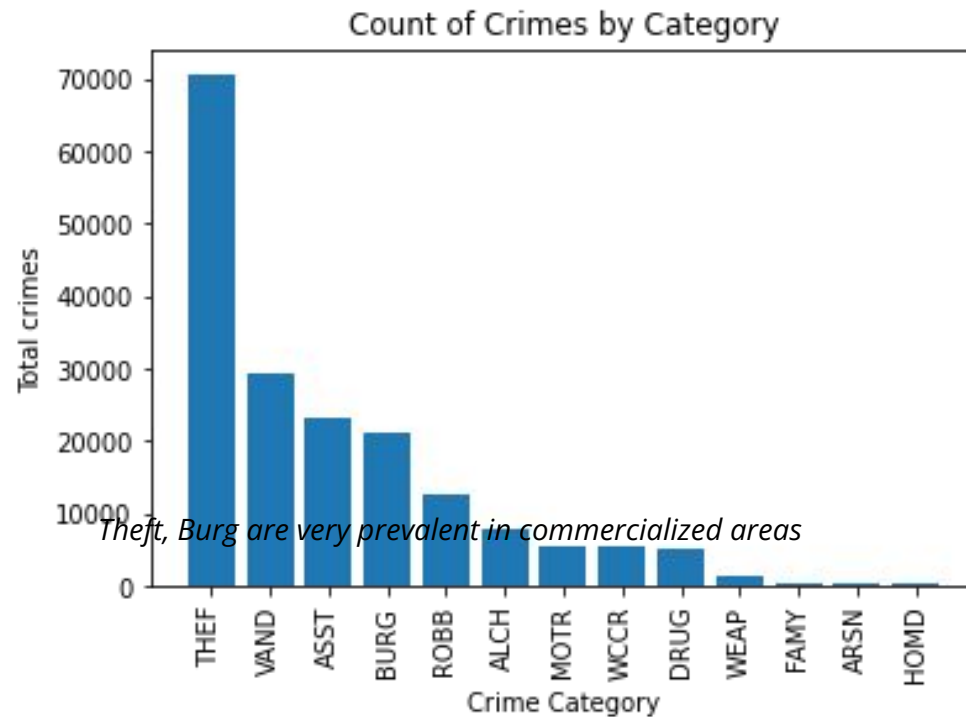
Count of Crimes by Day of Week

Total crimes

Day of the Week
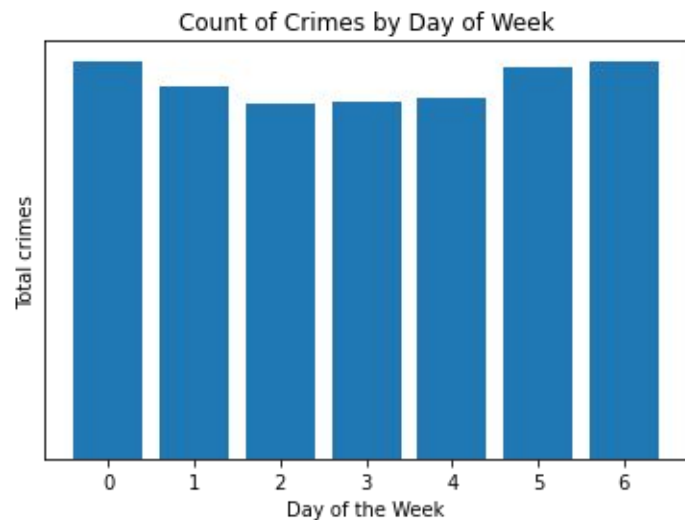
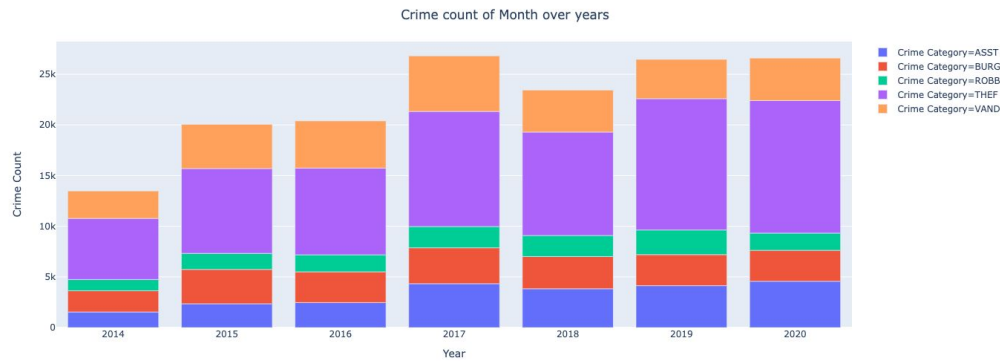*Most crimes occur during Friday-Sunday period*

Month number
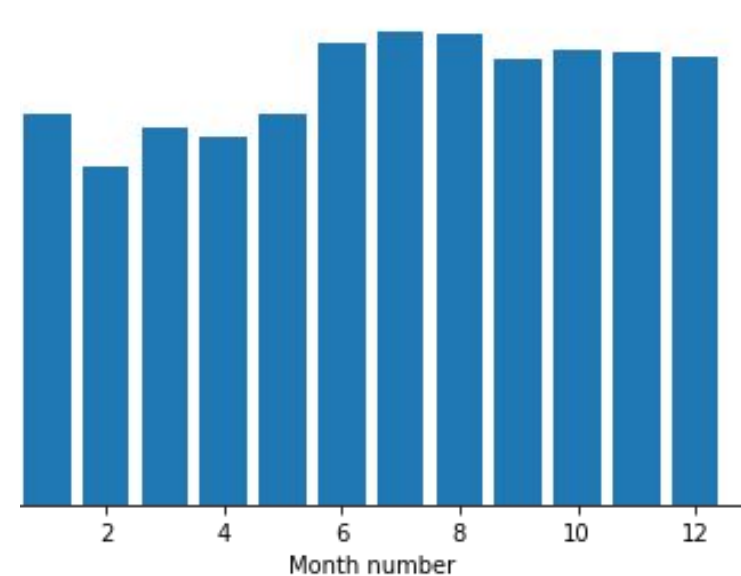
We can see that most crimes occur in the mid calendar year(Summer).

*Can see which TAAG'S are historic hotspots*

Crime Data by Category/TAAG

Count of Crimes by Category

Theft, Burg are very prevalent in commercialized areas

Count of Crimes for Each TAAG

Lets us see which TAAG areas are historical hot spots

Looking at Crime Count at a Granular Level

Crime count of Month over years

Plotting the crime count of the top 5 most popular crime categories over the years

Crime count of Month over years
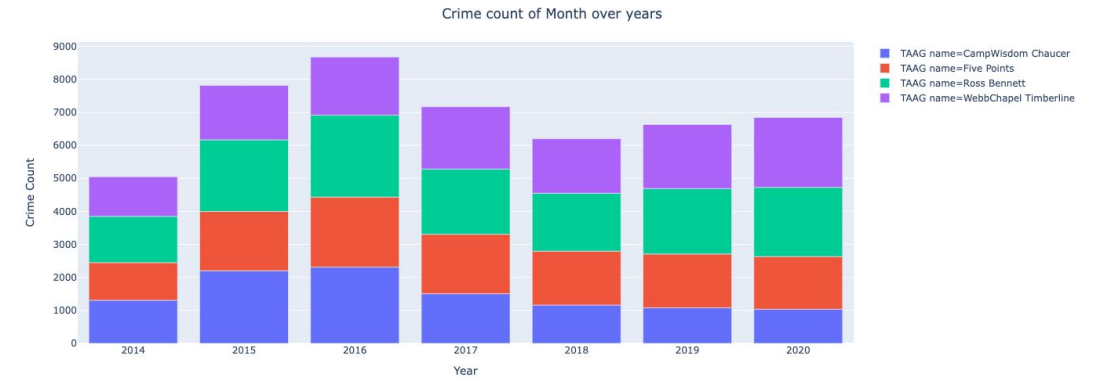
Plotting the top 4 TAAG areas over the years

Crime count of Month over 2020 by all crime categories
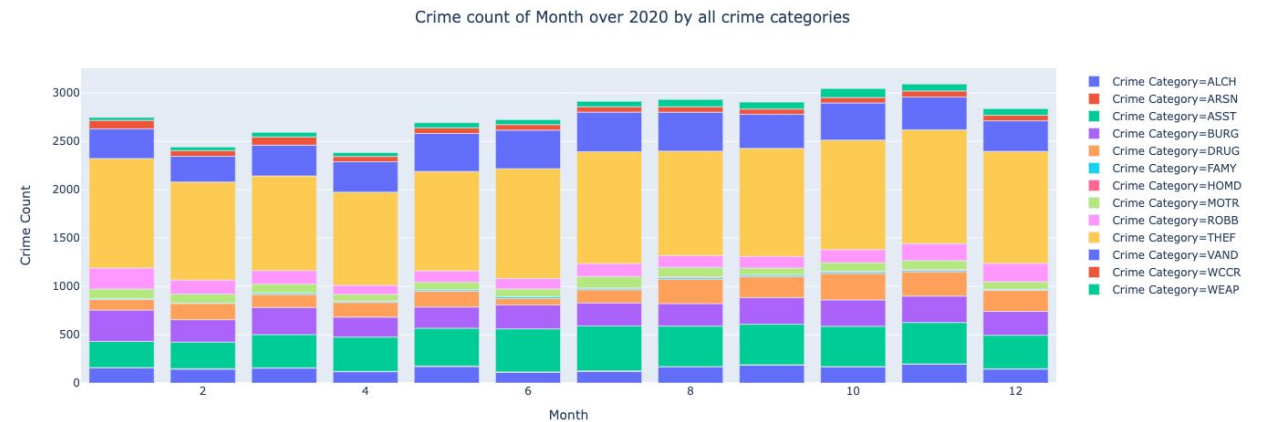
Looking at effects of covid crime count by category

Which Crimes are the Most Common?

Count of Crimes per year, TAAG, Category

Line plots of Crime Categories across TAAGS over all years
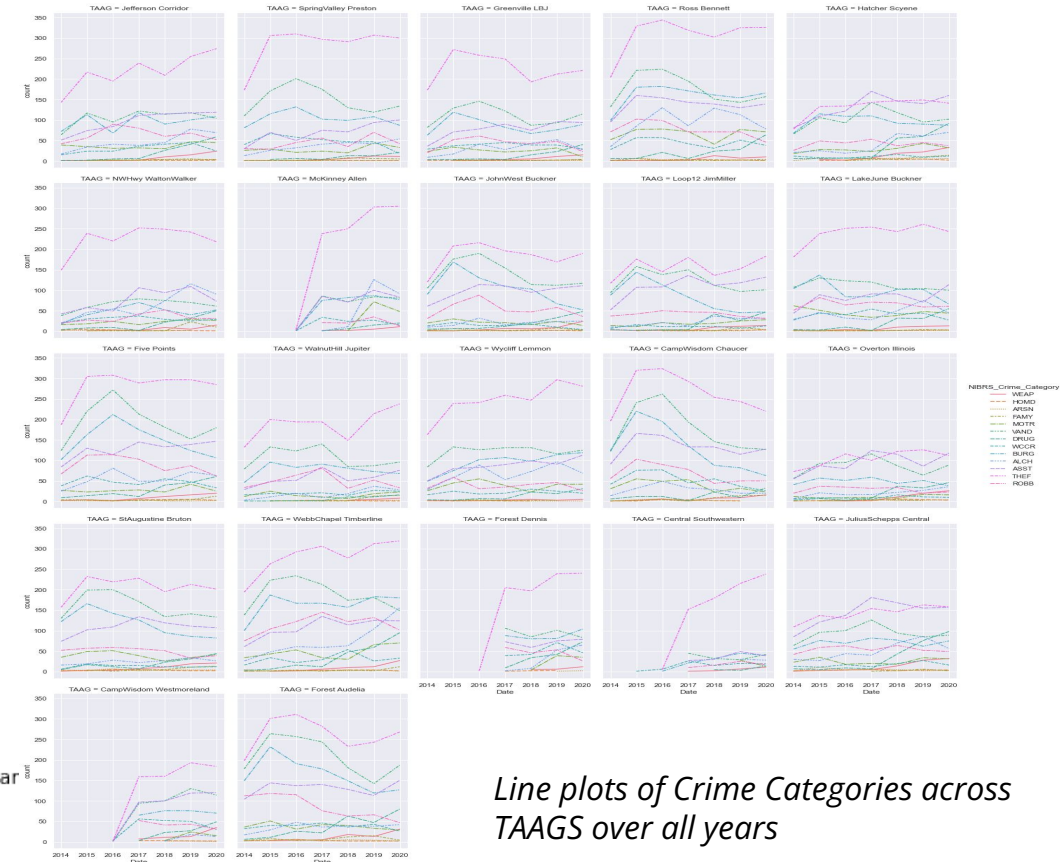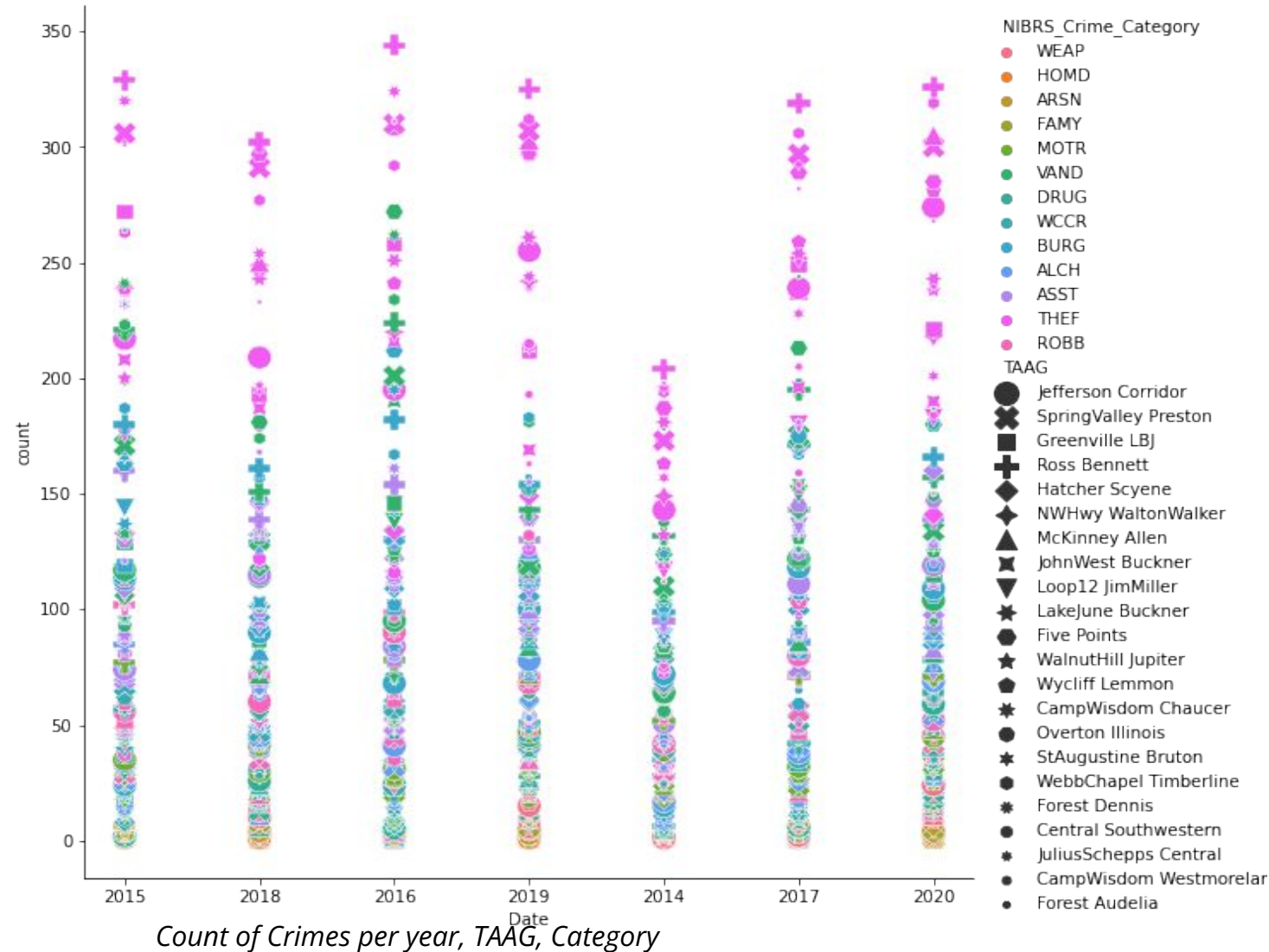
Data Cleaning

- Grouped by incidents of offences into related groups

- Aggregated by weeks

- Collapsed 53rd week into the 52nd week

Untransformed Features plotted across time by weeks for TAAG data
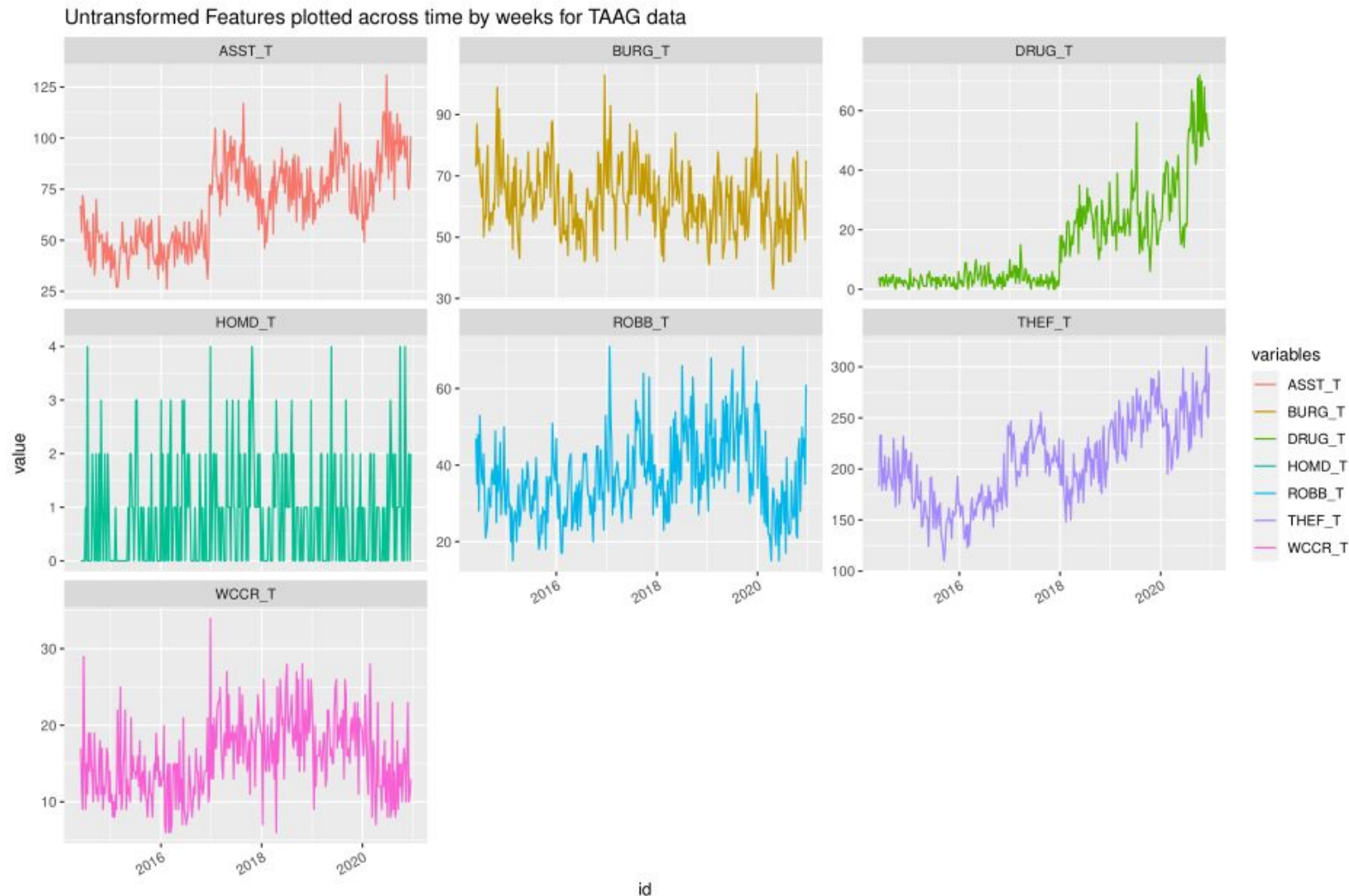
Transforming TAAG data to Stationary

- Applied log transformation if possible and then the difference

- Took the difference of each time series

- We also further scaled our data before modeling

Table 2: Table of Features and the Transformations Applied on Weekly TAAG data

Variable_W	Description_W	Transformation_Applied_W
ASST_T	Assault related Crimes in all of DALLAS	diff(log)
BURG_T	Burglary related Crimes in all of DALLAS	diff(log)
DRUG_T	Drug related Crimes in all of DALLAS	diff()
HOMD_T	Homicide related Crimes in all of DALLAS	diff()
ROBB_T	Robbery related Crimes in all of DALLAS	diff(log)
THEF_T	Theft related Crimes in all of DALLAS	diff(log)
WCCR_T	White Collar related Crimes in all of DALLAS	diff(log)

Transformed Features plotted across time by weeks for TAAG data
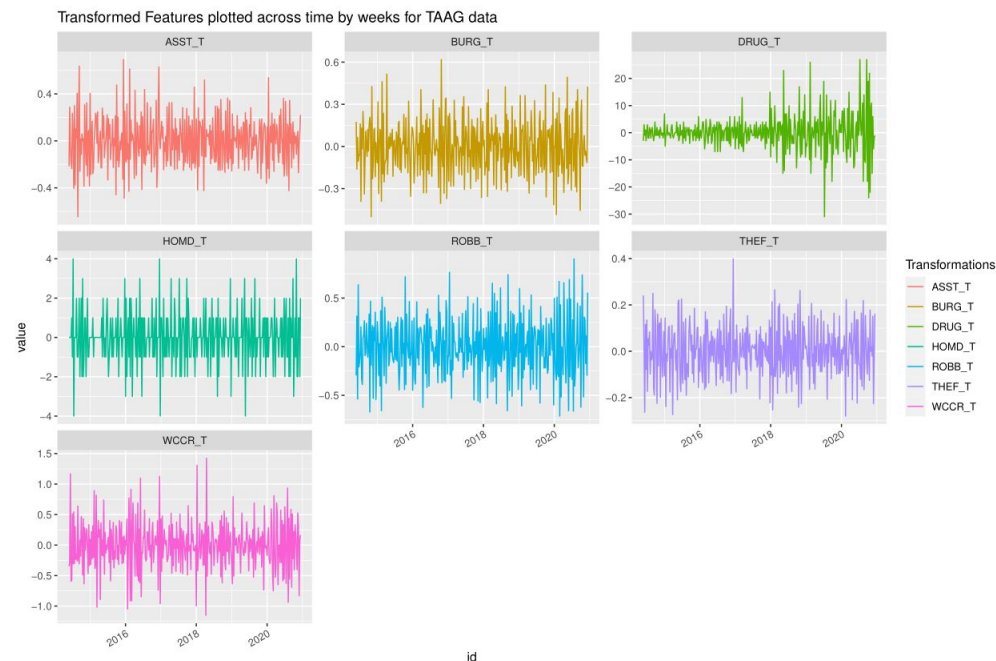
SparseVAR w/ HLag penalty for TAAG and Non-TAAG

- On the right is a visualization of the number of lags for each response that was used

- For TAAG we found relationships with other time series in our data

- We can also see that the Non-TAAG model forms an AR (1) model
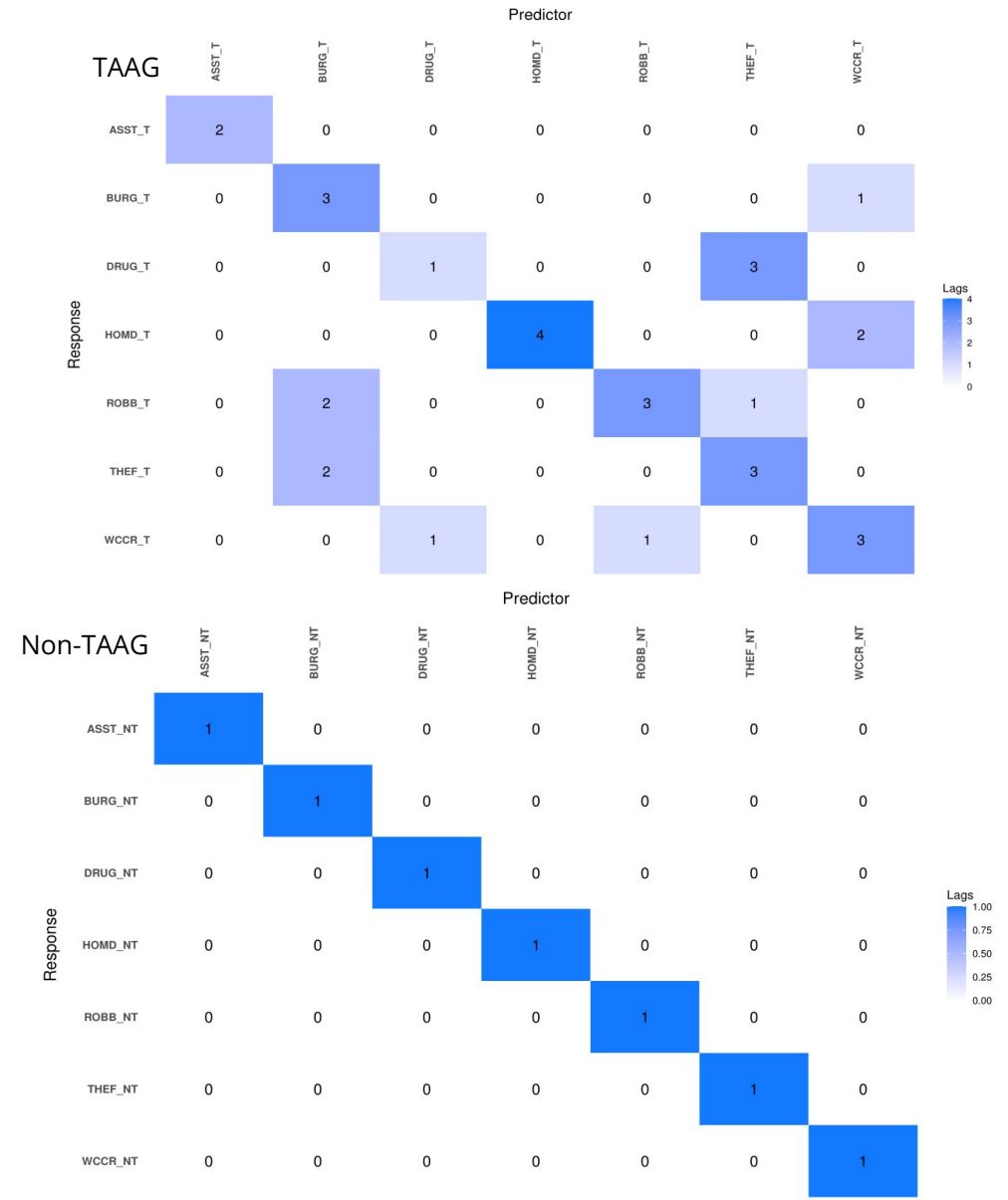
TAAG & Non-TAAG SparseVAR w/ HLag Coefficients

- We can see that Theft has an impact on drugs in TAAG areas

Table 9: TAAG model response variable is Drug

Variables	Coefficients
DRUG_T: Lag 1	-0.3786
THEF_T: Lag 1	-0.0107
THEF_T: Lag 2	0.0032
THEF_T: Lag 3	-0.0001
Intercept	0.0031

Table 16: Non-TAAG model response variable is Drug

Variables	Coefficients
DRUG_NT: Lag 1	-0.2339
Intercept	0.0045

- Also that Robbery is associated with Burglary and Theft in TAAG areas

Table 11: TAAG model response variable is Robbery

Variables	Coefficients
BURG_T: Lag 1	-0.0090
ROBB_T: Lag 1	-0.4813
THEF_T: Lag 1	0.0068
BURG_T: Lag 2	0.0015
ROBB_T: Lag 2	-0.1624
ROBB_T: Lag 3	-0.0376
Intercept	0.0035

Table 18: Non-TAAG model response variable is Robbery

Variables	Coefficients
ROBB_NT: Lag 1	-0.320
Intercept	-0.004

Math Expressions

- On the right tells how we forecasted 1 step ahead by providing new data

- RMSE is the metric we used for evaluating our models

out of sample forecast 1 step ahead equation

$$\begin{pmatrix} \hat{Y}_1(T+1) \\ \hat{Y}_2(T+1) \\ \vdots \\ \hat{Y}_k(T+1) \end{pmatrix} = \Phi_{k,L \times k} \cdot \begin{pmatrix} Y_1(T) \\ Y_2(T) \\ \vdots \\ Y_k(T) \\ \vdots \\ Y_1(T-L+1) \\ Y_2(T-L+1) \\ \vdots \\ Y_k(T-L+1) \end{pmatrix} + \Phi_{0_{k,1}}$$

$RMSE$

$$RMSE = \frac{1}{k \times (T_2 - T_1)} \sum_{i=1}^{k} \sum_{t=T_1}^{T_2-1} \sqrt{(\hat{y}_{i,t+1} - y_{i,t+1})^2}$$

Model Evaluation

- We considered 3 models, sparse Var w/ HLag penalty, regular Var, and AR (1)

- We evaluated our models on 3 datasets TAAG, Non-TAAG, and Overall

- We found that the sparse Var model performed the best for our data

Table 28: Out-Of-Sample model evaluation

	mean(RMSE on TAAG DATA)	mean(RMSE on NON-TAAG DATA)
SparseVar_CV_HLAG	15.68545	25.99907
SparseVar_CV_HLAG_Overall	19.26385	28.05909
VAR_AIC	18.21052	28.76986
VAR_AIC_Overall	19.80248	29.02669
AR(1)	17.06835	26.91332
AR(1)_Overall	19.53949	28.64023

Note:

Train/Test split ratio 0.8

Note:

Prior probabilities of TAAG/Overall for each crime to get the forecast of crimes in TAAG areas from Overall model forecasted values.

It follows that the prior probability for Non-TAAG is 1 - the priors for TAAG/Overall

[1] TAAG Prior for Assault: 0.38

[2] TAAG Prior for Burglary: 0.32

[3] TAAG Prior for Drug: 0.4

[4] TAAG Prior for Homicide: 0.41

[5] TAAG Prior for Robbery: 0.43

[6] TAAG Prior for Theft: 0.3

[7] TAAG Prior for White Collar Crime: 0.32

Future Work

- For the future we would like to see how adding secondary factors affects our model

- We would also like to explore if we get a better overall forecast of crimes by combining the forecasted values of TAAG and Non-TAAG