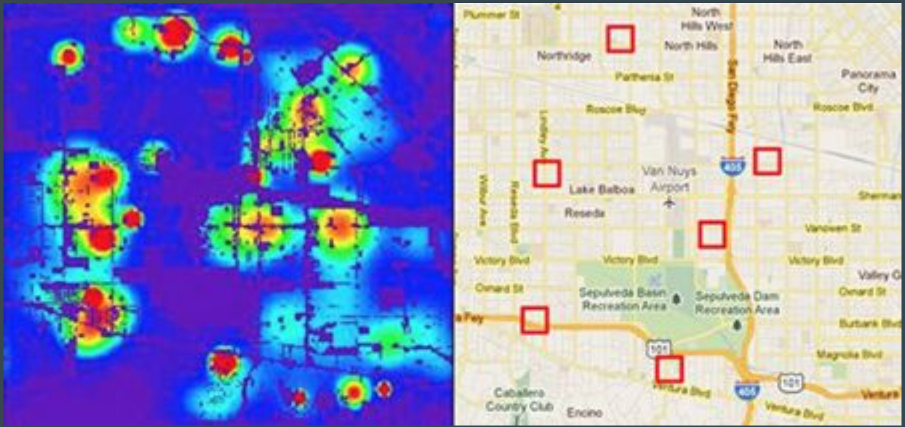
Forecasting Crime in Chicago Communities

Authors: Camilla Bendetti and Alexander Foster
Capstone Advisor: Brianna Heggeseth
Presentation Coach: Lori Ziegelmeier

“Can we model future crime in Chicago by analyzing historical crime data?”



PredPol Predictions in LA

Correlated Data
Stat 494

vs.

Intro to Stats
Stat 155

Assuming independence is not always accurate or reasonable.

Correlated Data

Longitudinal Data:

- A type of temporal data
- Repeated measurements on many units
- Looking at correlation within cases of crime

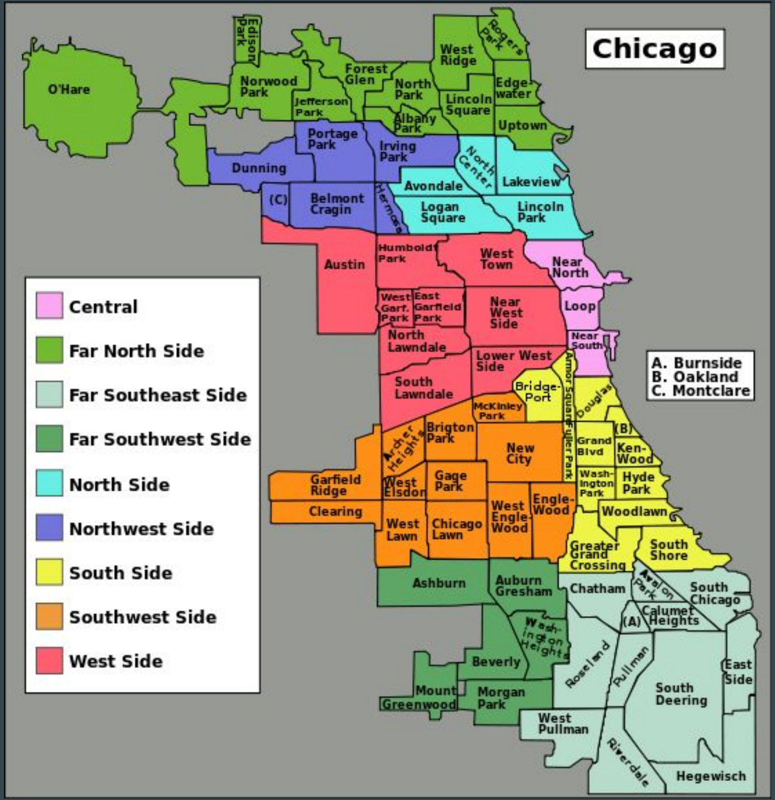
Battery Crime in Chicago

Correlated Data

Spatial Data:

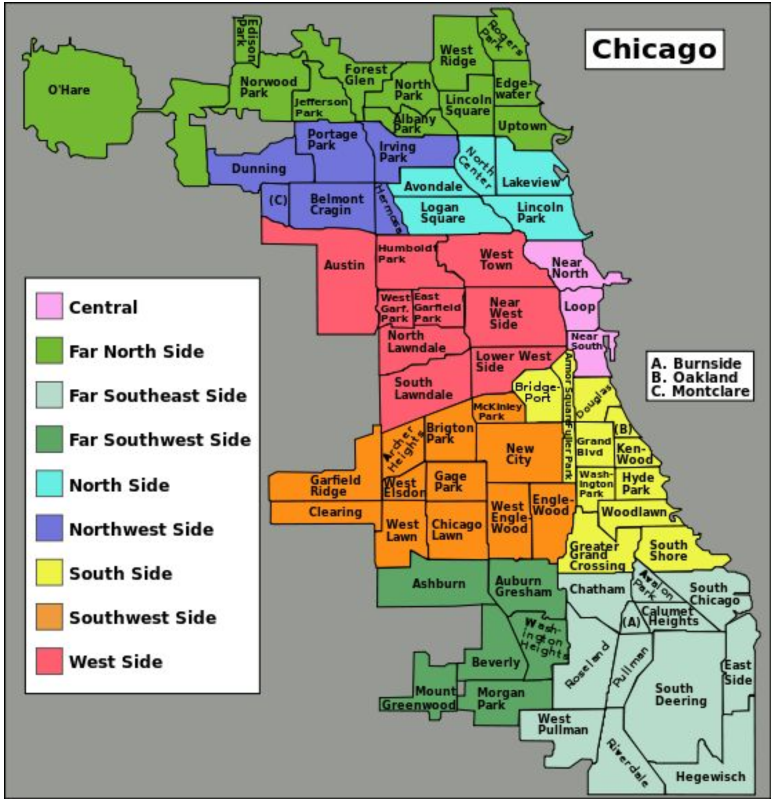
- Areal units
- Polygons rather than coordinates
- Looking for correlation between community areas

Community Areas of Chicago

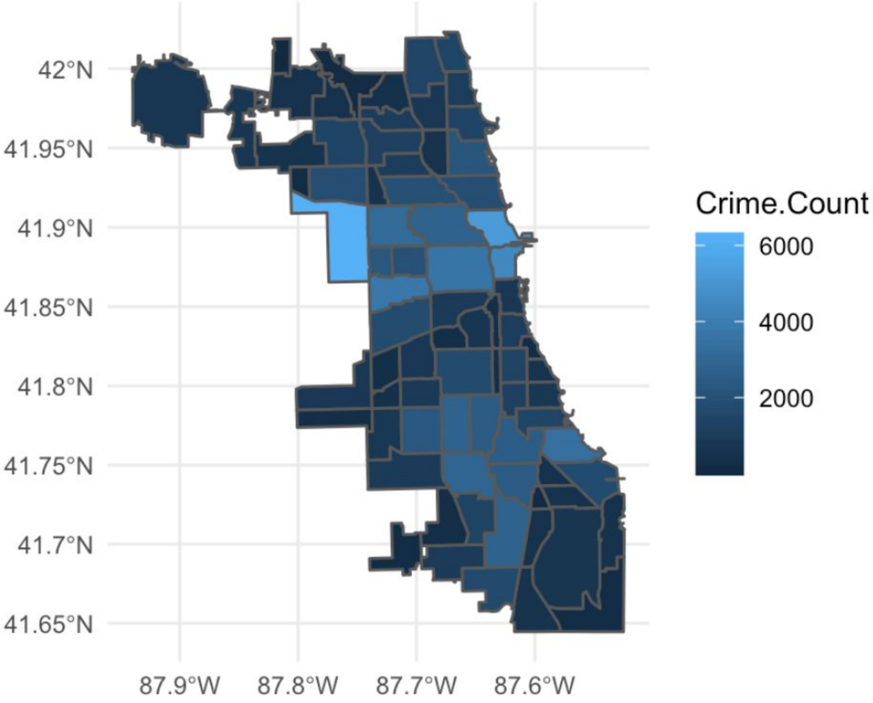


Chicago Data Portal

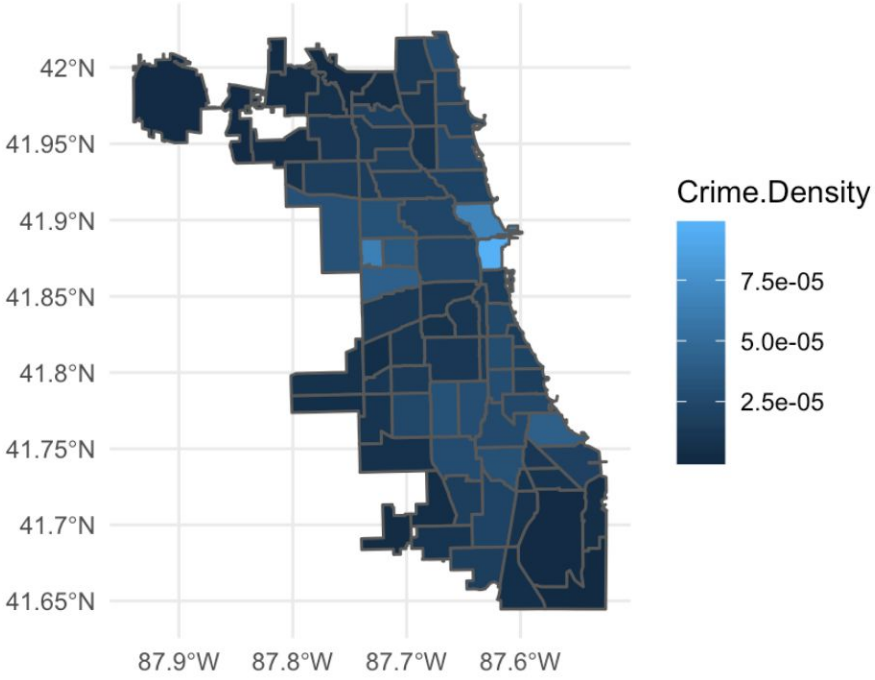
77 Community Areas



Shape File

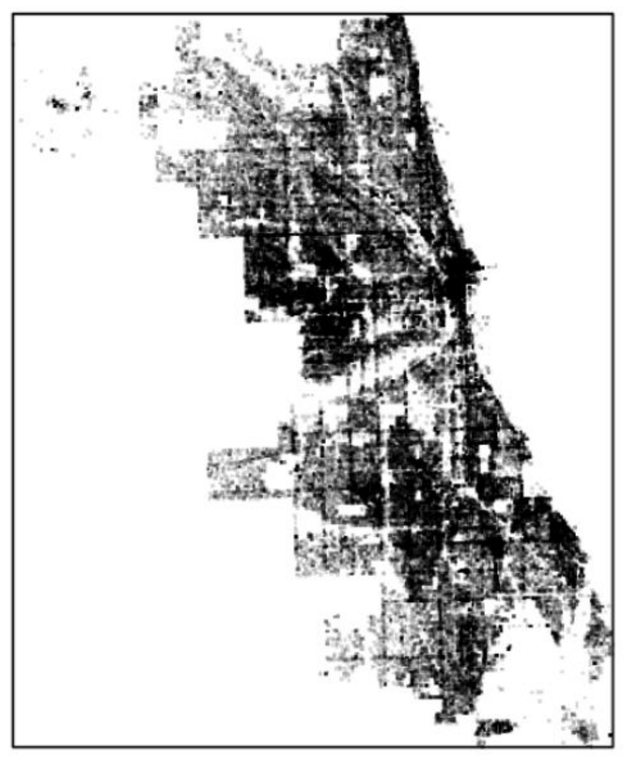


Crime Count by Community Area

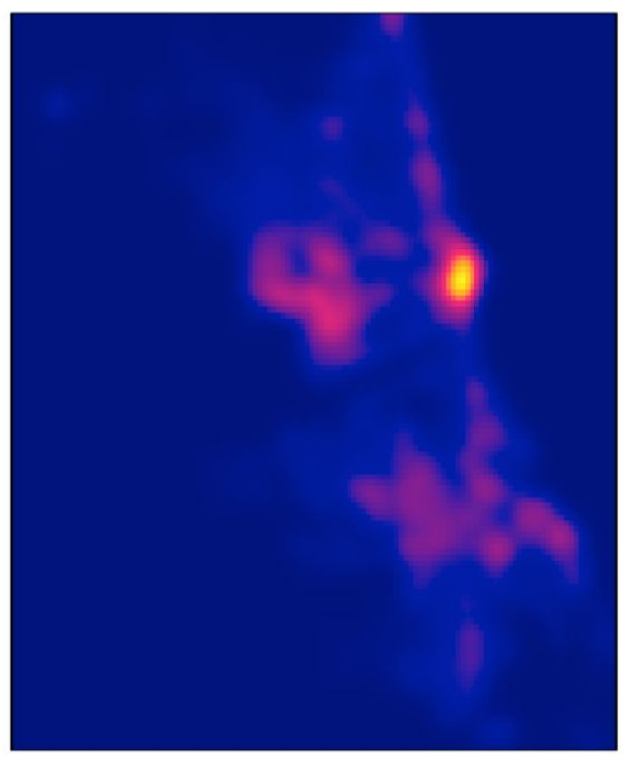


Crime Density by Community Area

Point Processes of all crime from 2018



Density plot of all crime from 2018



Picking Variables

ID, Case Number, IUCR, Description, Location Description, Arrest, Domestic, Beat, District, Ward, Block, FBI Code, X Coordinate, Y Coordinate, Year, Updated, Latitude, Longitude, Location, Date, Community Area, Primary Type

Date

Sample: 1/2/18 9:00

Separated into Day of Week

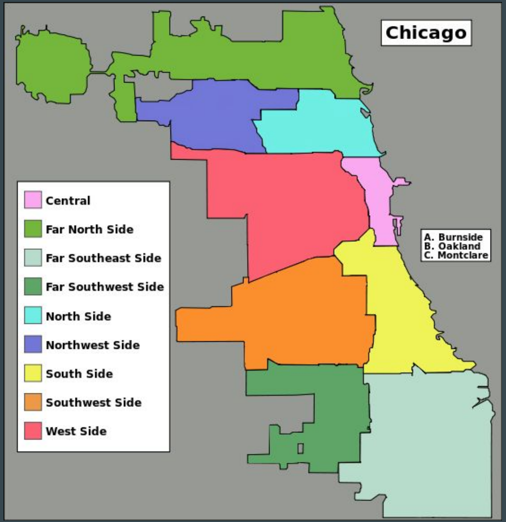
- Monday - Sunday

Separated into Time of Day

- Morning
- Afternoon
- Night

Community Area

Region

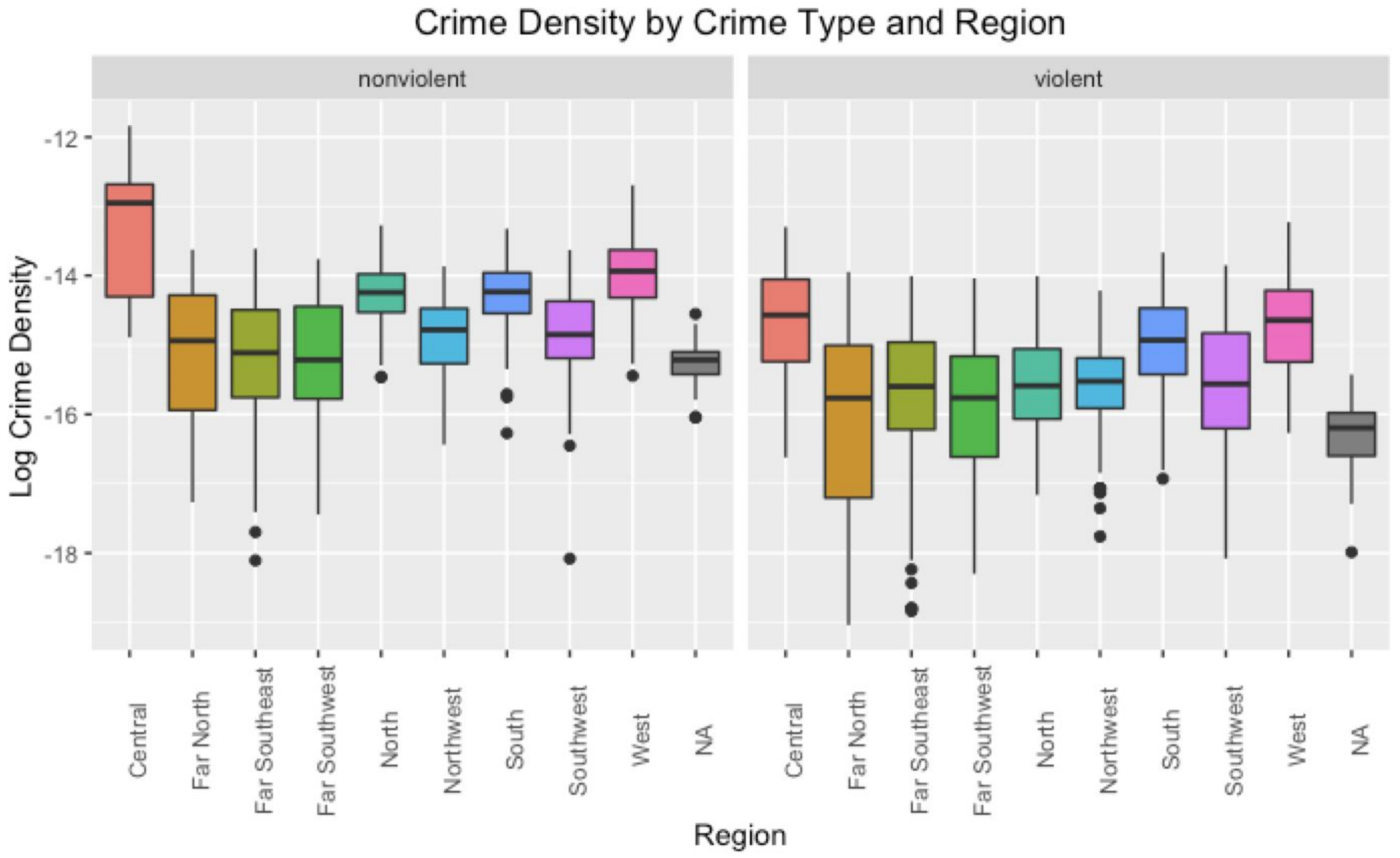


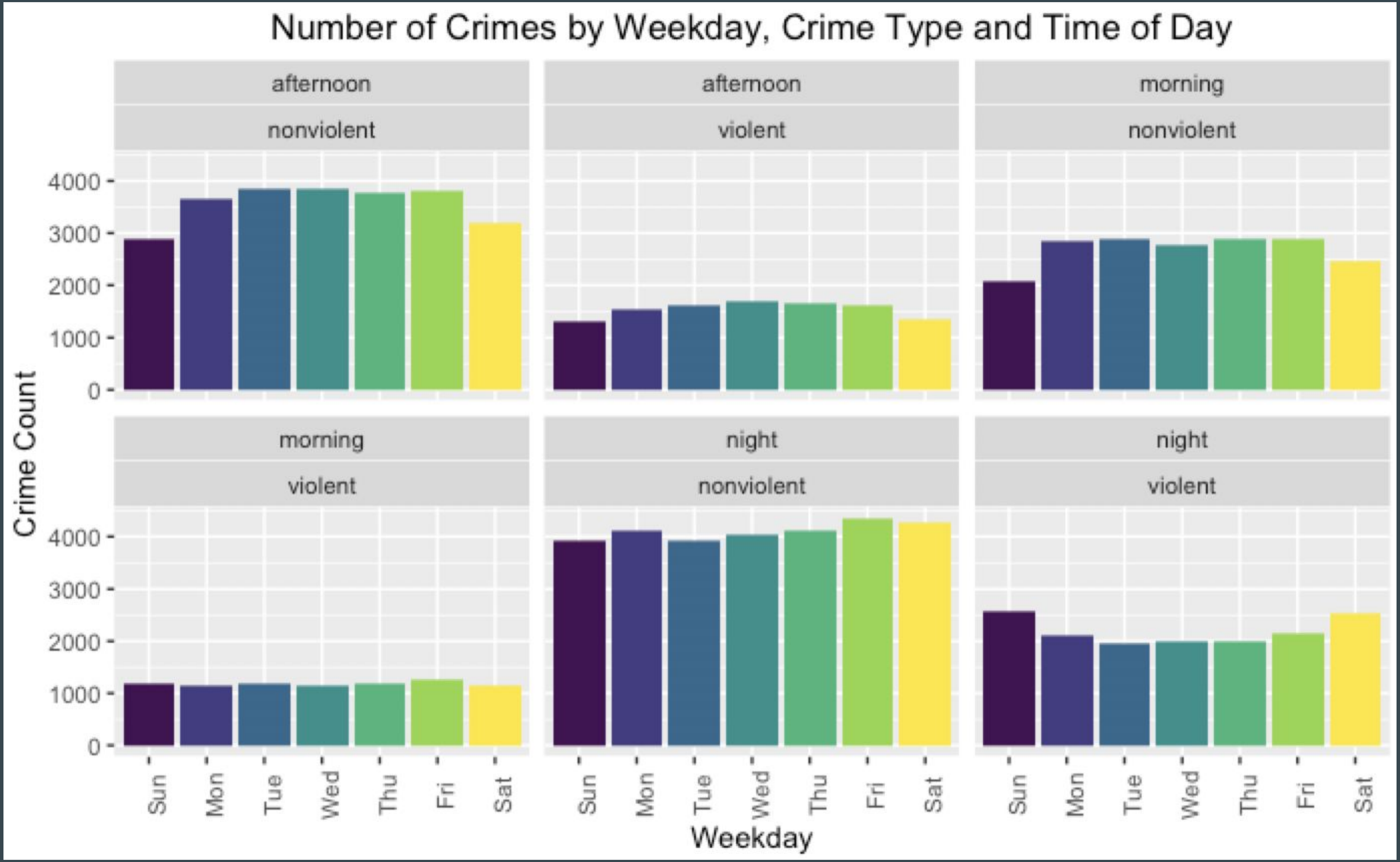
Primary Type

Crime.Type:

Violent - Homicide, Kidnapping

Non-Violent - Gambling, Interference with a public officer


Crime Density by Crime Type and Region
nonviolent
violent
Log Crime Density
-12
-14
-16
-18
Central
Far North
Far Southeast
Far Southwest
North
Northwest
South
Southwest
West
NA
Region


Number of Crimes by Weekday, Crime Type and Time of Day
afternoon
nonviolent
afternoon
violent
morning
nonviolent
morning
violent
night
nonviolent
night
violent
Crime Count
4000
3000
2000
1000
0
Sun
Mon
Tue
Wed
Thu
Fri
Sat
Weekday

Our Model

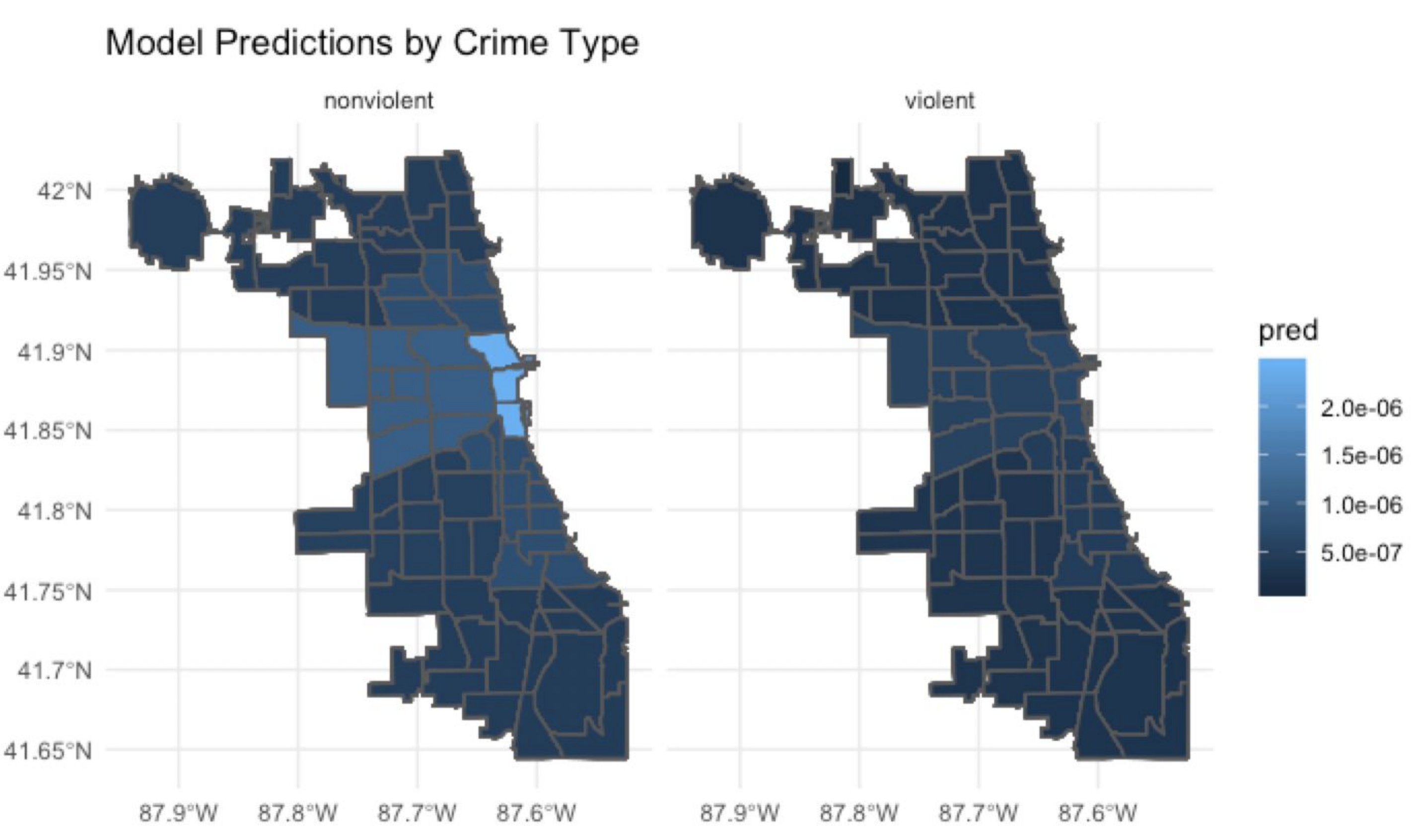
$$(\text{Crime.Type} * \text{ToD}) + (\text{Crime.Type} * \text{Region})$$

- We included the type of crime (Crime.Type), Time of Day (ToD), and Region in our final model
- We added interactions between Crime.Type and ToD as well as Crime.Type and Region.
  - Multiplying Crime.Type by ToD and Crime.Type by Region
  - If we did not add an interaction term, we would be making the assumption that the crime density is the same throughout regions for the time of day or the type of crime. We can see from our initial visualizations that this is not the case.

Marginal Model

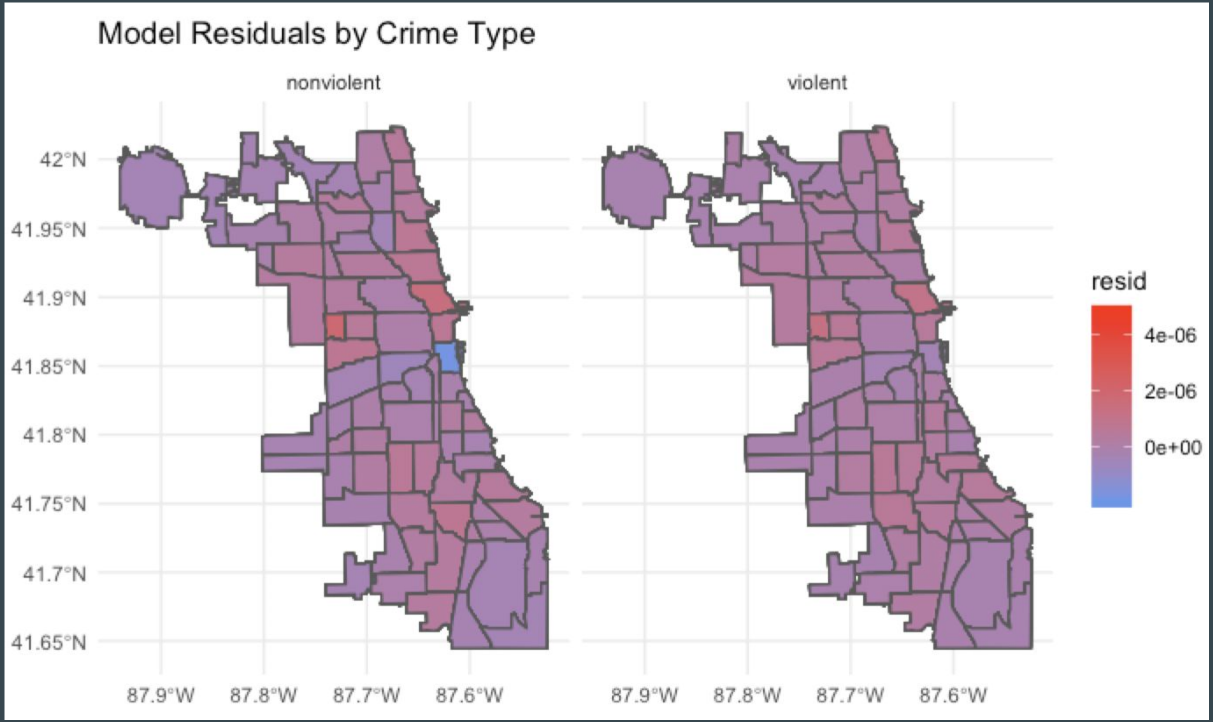
- Used a marginal model because our data has repeated measures
  - Type of linear model that can explain repeated measures
  - Within our data, there are two different kinds of correlation
    - The repeated measures within regions are correlated
    - Nearby regions are spatially correlated.

Model Forecasts



Residuals

- Represent the error in our model
- Residuals help us to see how well our model performs
- Overall low residuals
- Variability in the central region

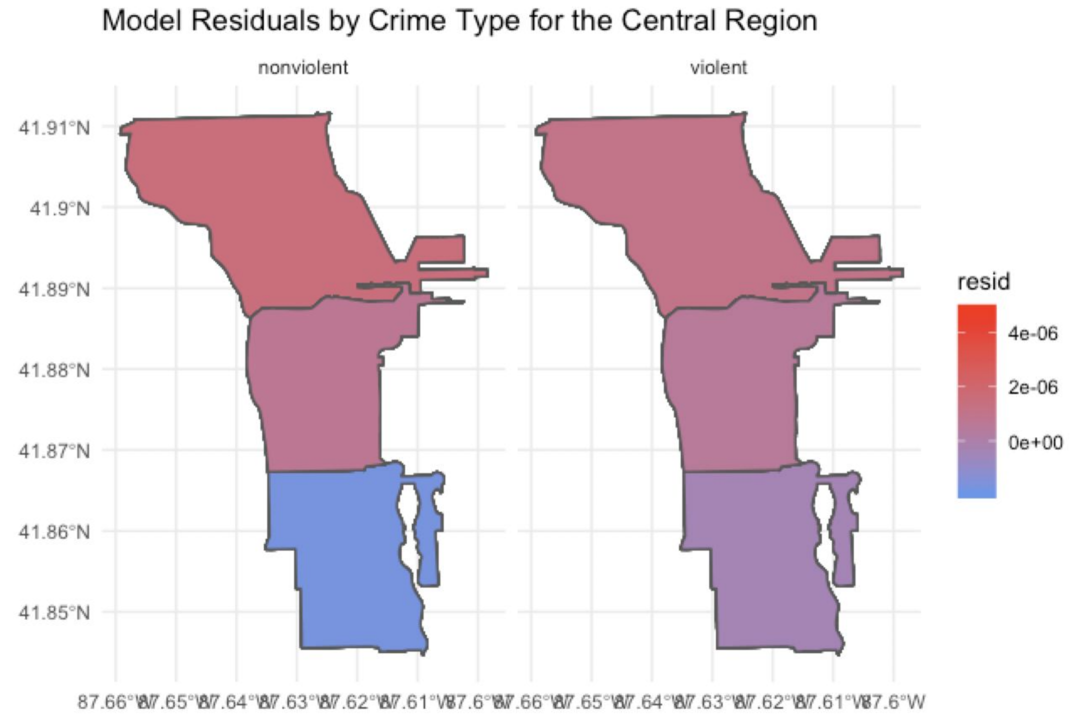


Central Region Residuals

- The central region has a significantly higher density of crime than all other regions in Chicago.

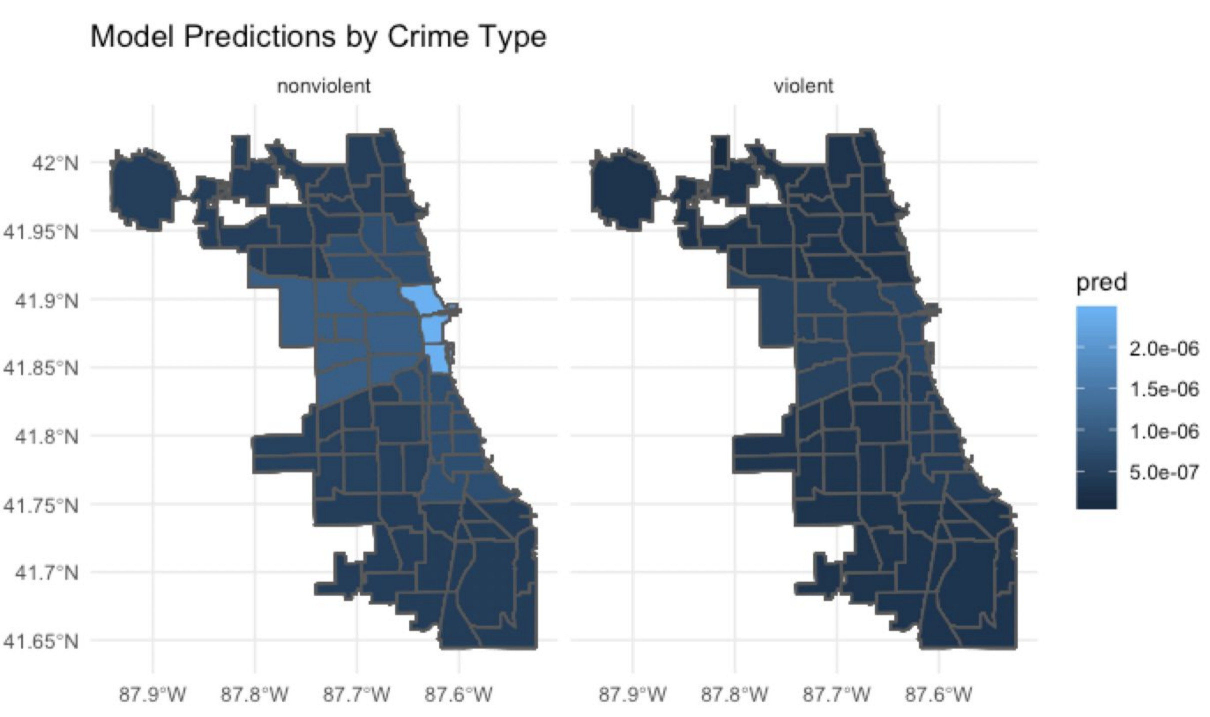
- Higher residuals because the central region is an outlier.

- Our model is not specific enough.



Conclusion

We found that most crimes were centered in Central, South, West and North regions and that non-violent crime was concentrated in the Central region.



Limitations and Future Directions

- Used data only from 2018
    - Use data from multiple years

- Predictions are by region not community area
    - Create models for individual regions or even community areas

- Classified crimes as violent vs non-violent
    - Classify crimes more specifically

- Arrests vs. Convictions

Community Costs

- Regardless of the number of data points, our predictions will be biased because of racial profiling amongst other things.

- Predictive policing algorithms are “designed to learn and reproduce patterns in data.” [@lum2016predict]

- However, if we are using already biased data to train these models, our output will create feedback loops that reproduce these biases in the future.

Thank you!