Predicting and contextualizing city-level opioid overdose deaths across Massachusetts

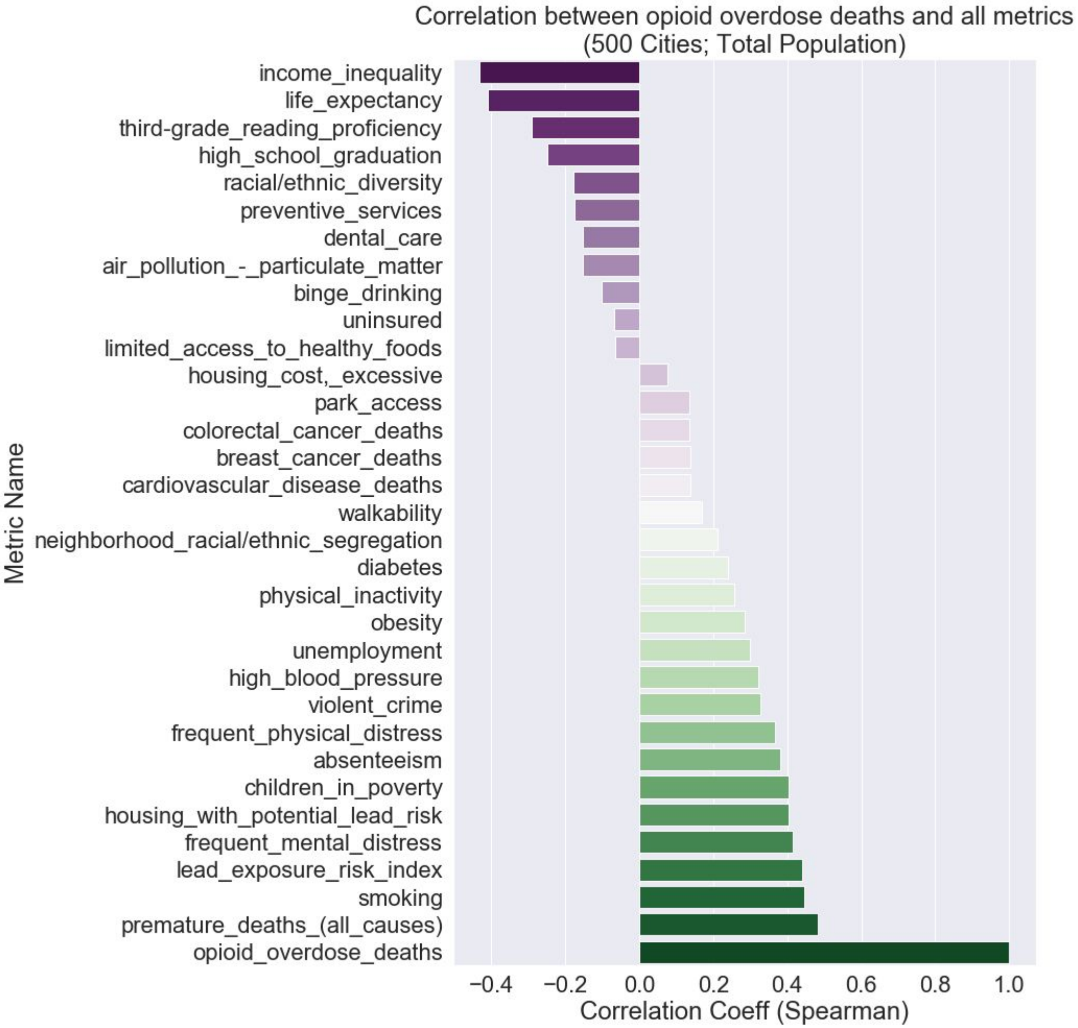
Dasha Akimova
Insight Health Data Science Fellow

Primary goals of the project

- To explore and identify publicly available data sources that could be useful in contextualizing Biobot’s measurements in sewer water
  - Use MA city-level opioid overdose deaths as a proxy for Biobot’s data
  - Build a model to select for meaningful features
- Develop a strategy for merging datasets at different geospatial resolutions

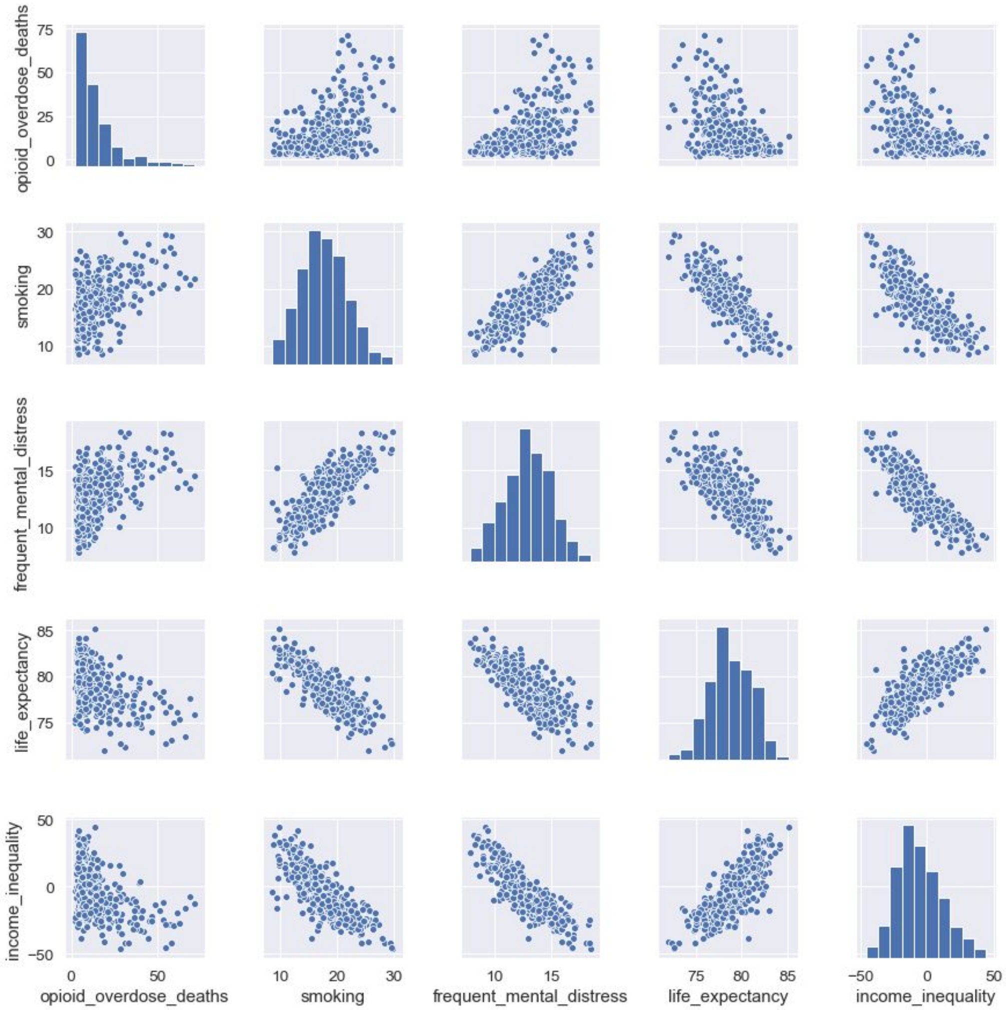
Q: What might be correlated with opioid overdose deaths?

- By recommendation, started with the 500 US city health dashboard dataset
- Opioid overdose deaths only available for general population for one time point
- Potentially interesting features:
  - Income/poverty
  - Education
  - Other drug use (smoking in this dataset, but maybe other drugs?)



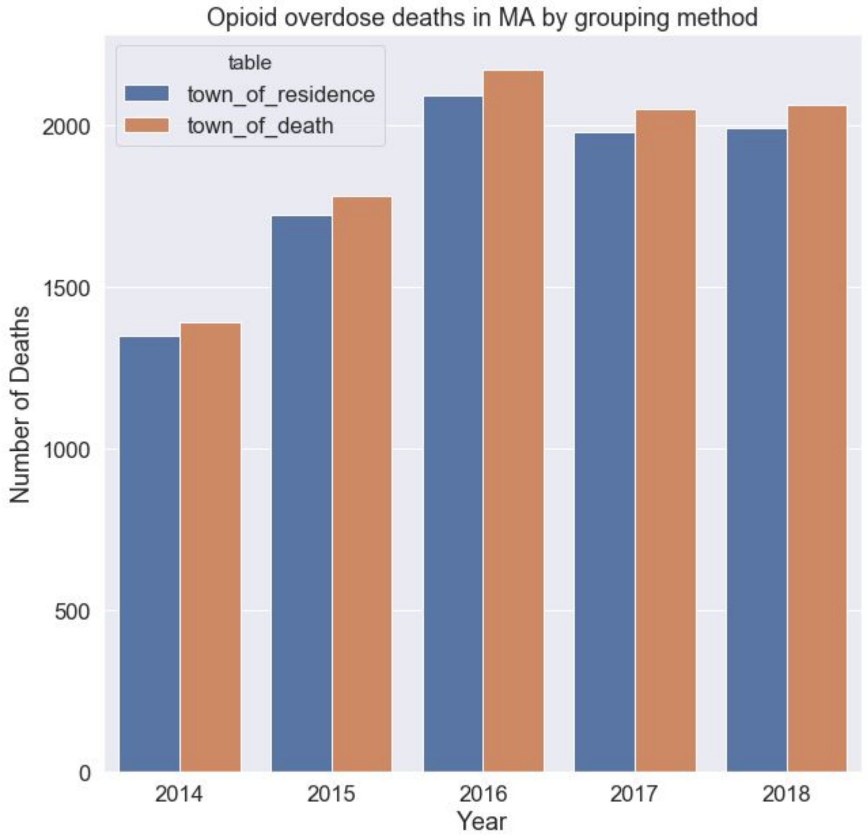
Further EDA on 500 cities dataset

- Many features of interest were strongly correlated with each other - a potential concern for modeling
- Opioid overdose deaths skewed in this dataset (turns out to be very skewed in the MA data also)
- Concern - potential features of interest could just be tied back to poverty?

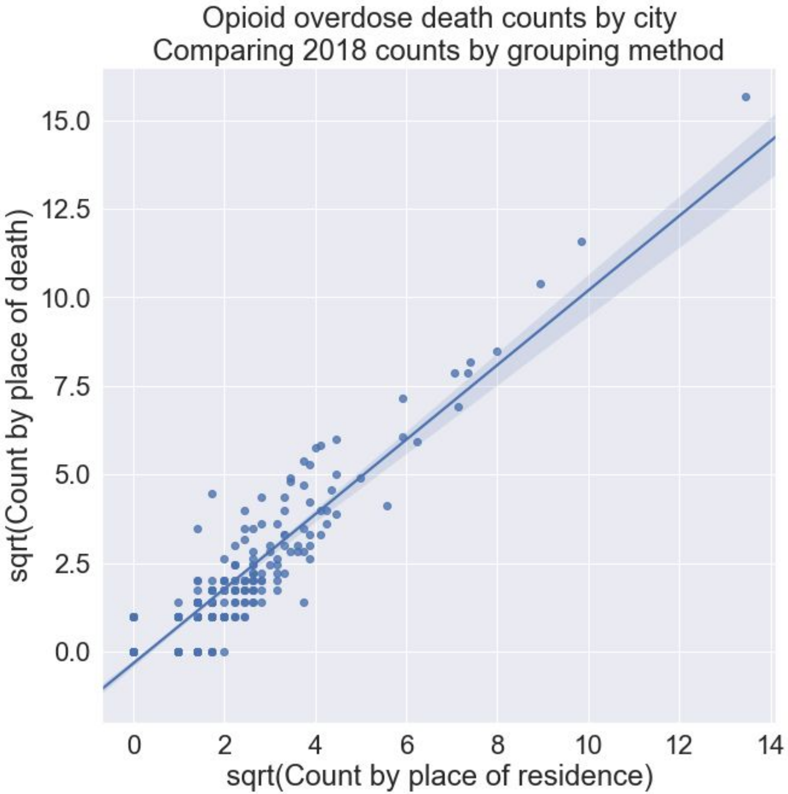


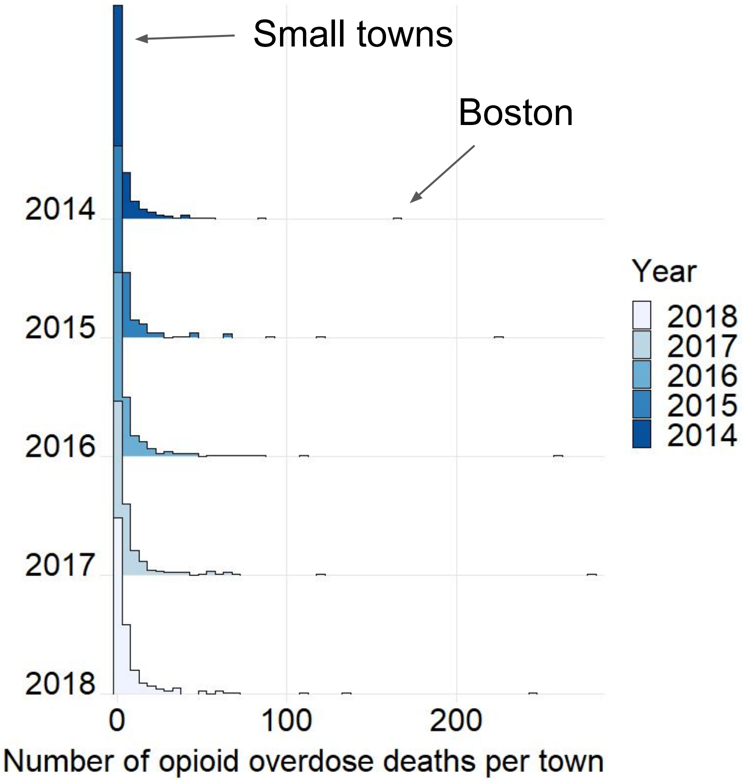
MA opioid overdose death count datasets

- PDF from Mass.gov
- Initially planned to use python to extract data from pdf
- Turned to https://pdftables.com/ instead: pdf > csv
- 2 Tables:
  - MA residents only - by place of residence of decedent
  - By place of death



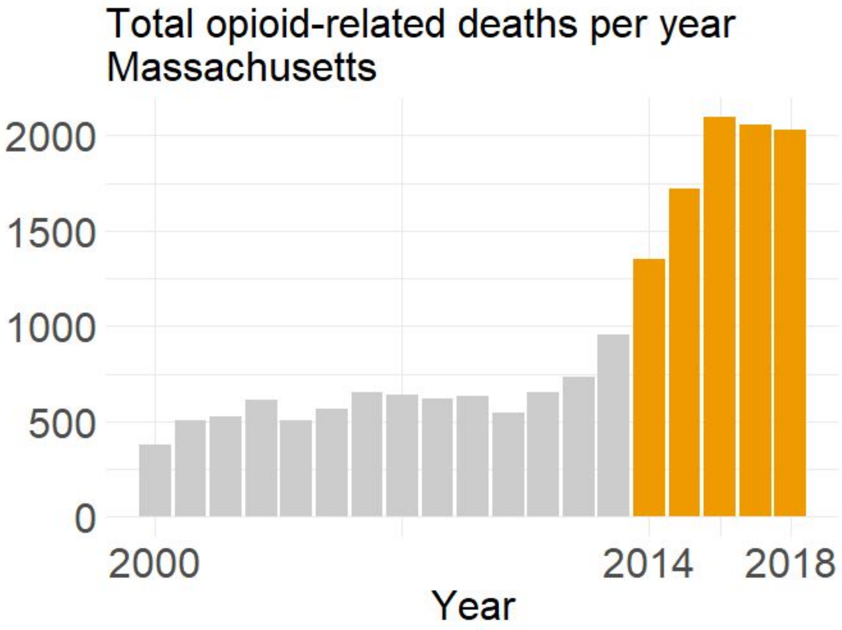
Both tables highly correlated and skewed



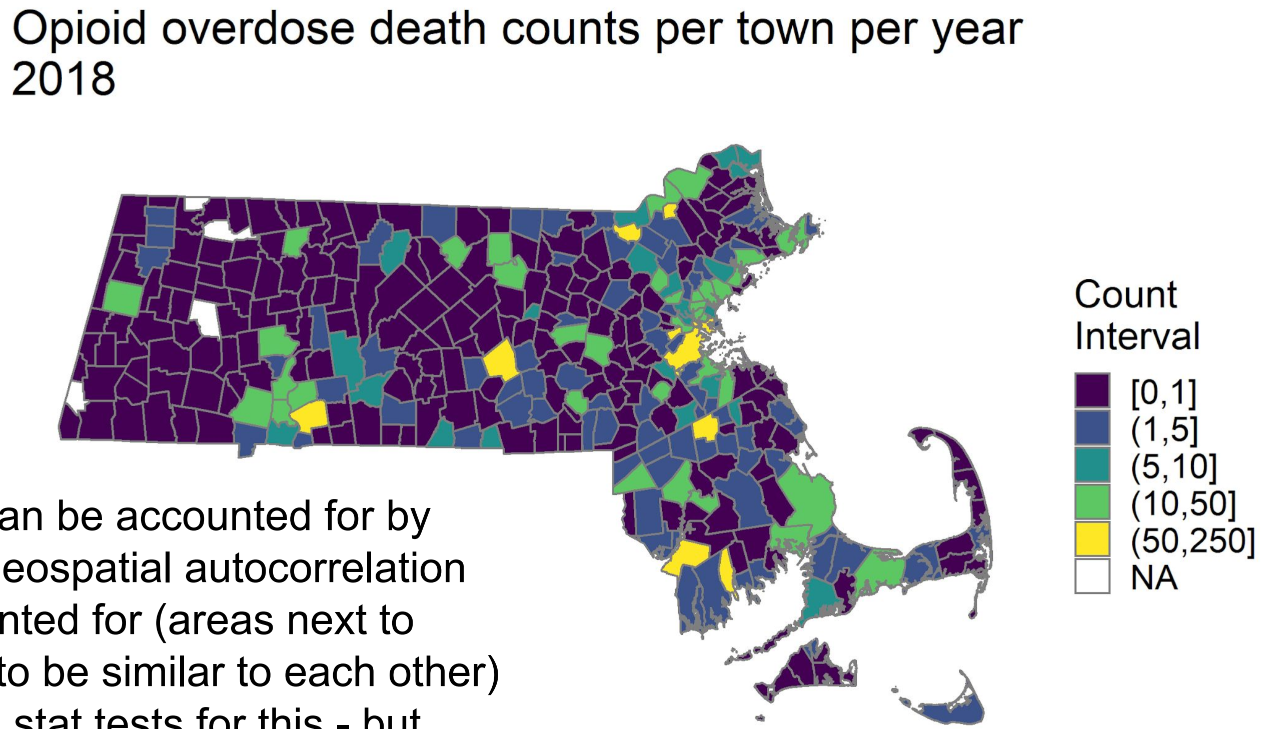


Modeling and data gathering strategies

Why time series?
Wanted to capture year-over-year trend



Modeling and data gathering strategies

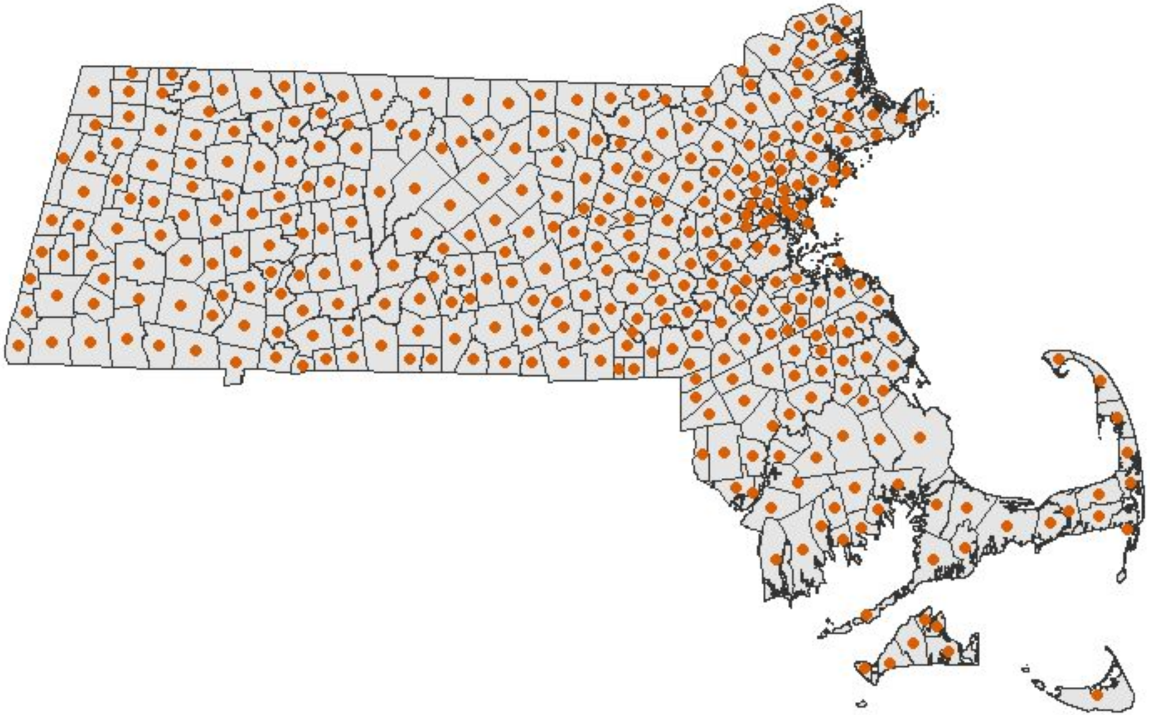


Why geospatial?

- Some patterns can be accounted for by population, but geospatial autocorrelation should be accounted for (areas next to each other tend to be similar to each other)
- There are formal stat tests for this - but figured better safe than sorry

Geospatial component for modeling

- Primary geospatial tools:
  - Python: geopandas - data wrangling/EDA
  - R: sf (choropleth, other figures)
- Luckily Mass.gov shapefile matched all 351 overdose municipalities
- Converted shapefile from polygon geometry to point geometry (centroid)



Dataset building strategy

Need:

- City population counts to normalize opioid overdose death counts

Want (ideas from 500 city dashboard, but also other research):

- Income
- Poverty level estimates
- Education
- Opioid prescriptions
- Other drug use?

Dataset sources

2017 American Community Survey (Claire)

- Population estimates
- Income/Poverty
- Education

Drug prescriptions:

- Considered CDC for opioid - but data only at county level
- Medicare for opioid - zipcode level 2013-2017 datasets
- Medicare also provides data on other drug prescriptions (also 2013-2017)
    - Wanted to pull out data on benzodiazepine prescriptions
- Concern: typical Medicare demographic wrong for opioid overdoses

Data merging strategy

Town Level
Opioid overdose
death counts

geopandas
shapefiles
spatial joins

Census Blocks Level

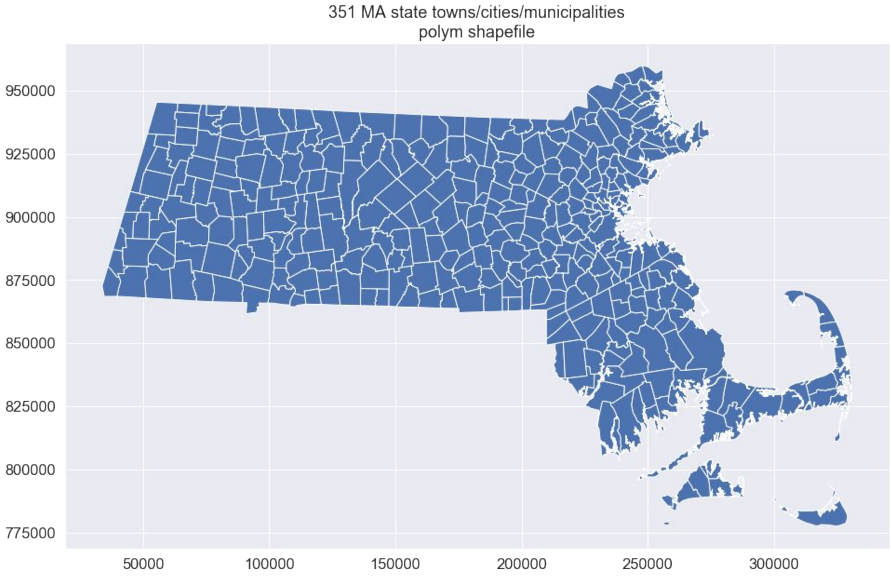
- Population count
- Income / Poverty levels
- Education

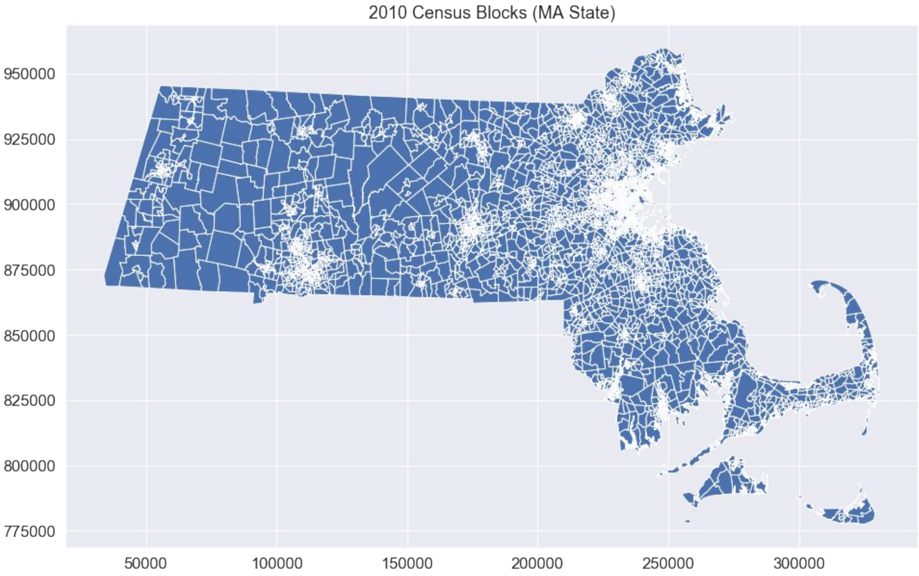
Postal ZIP Code Level
2 datasets:

- Opioid prescription rate
- Benzodiazepine prescription counts

Town - Census block merge

geopandas + Massachusetts shapefiles + spatial joins (joining on overlapping geometries)

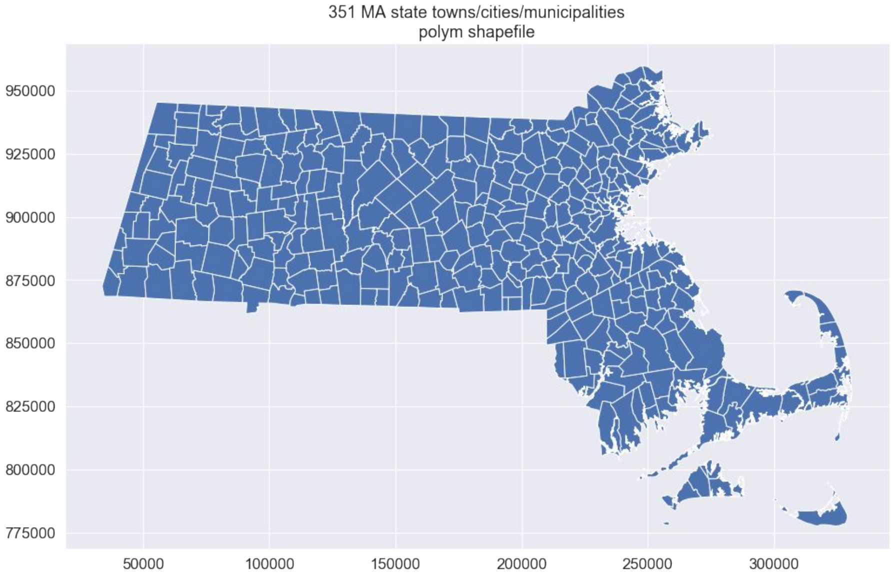


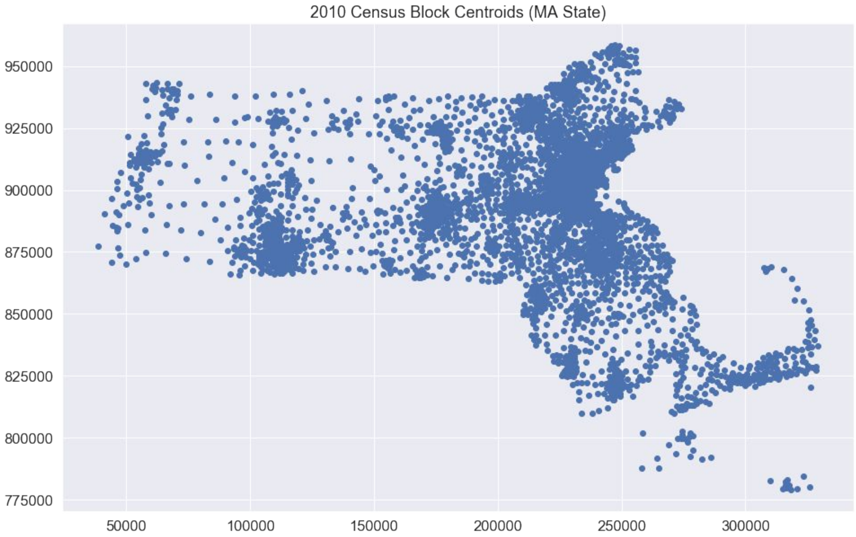


Merging above did not work - either too many or too few associations - lots of errors

Town - Census block merge

geopandas + Massachusetts shapefiles + spatial joins (joining on overlapping geometries)

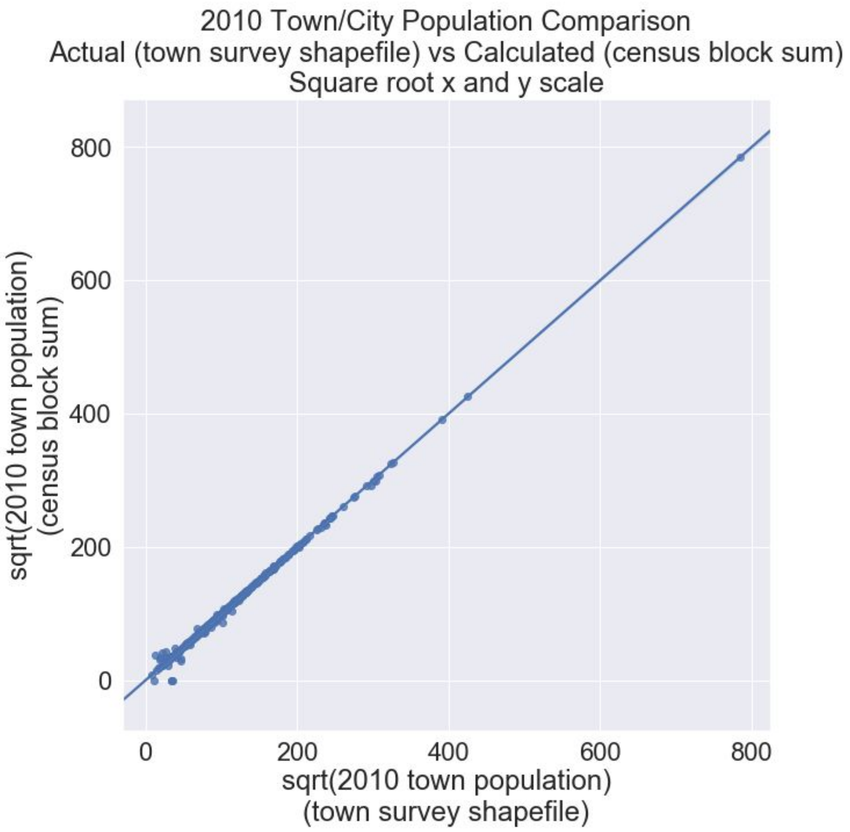




Convert polygons to points (centroids) - worked pretty well for most towns!

Town - Census block merge validation

- Both shapefiles included 2010 Census population counts - compared expected (from towns shapefile) to estimate (from joined census blocks)
- Strategy worked well for most towns:
    - 52 had non-zero error
    - 31 with error > 5%
    - Mostly small towns (100-10k population)

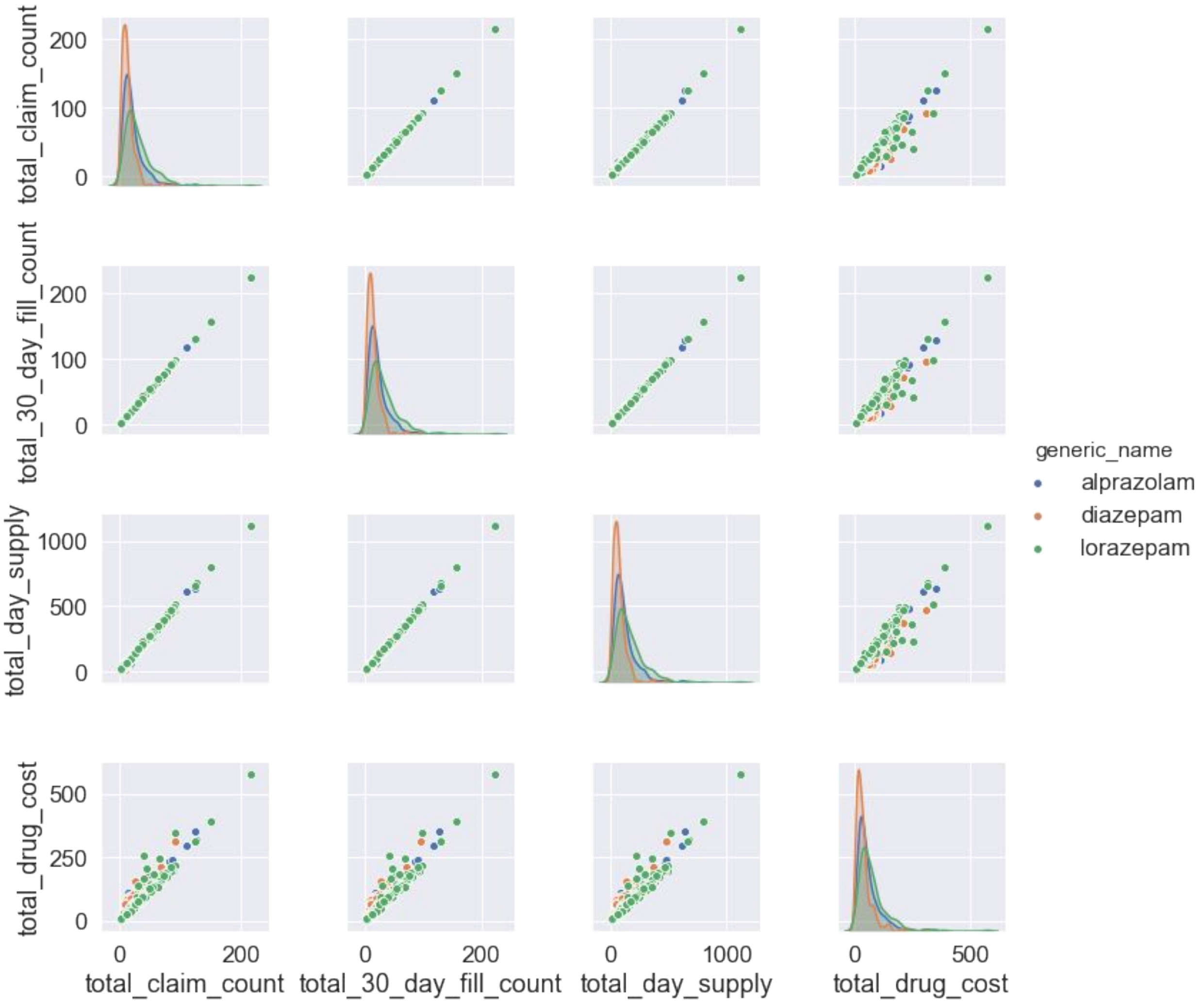


Town - Zipcode merge

- More traditional table joining strategy
- Used shapefile MA postal, but merged on zipcode
- Strategy:
  a. Opioid prescribers came with zipcode - used postal code shapefile to associate them with town that matched the opioid overdose death count towns
  b. Benzodiazepine prescribers came with town, but towns did not match the opioid overdose death count towns
  c. Merged benzo prescribers with opioid prescribers on NPI code - matched to opioid overdose death count town that way
- Problems:
  - Medicare prescriber zipcodes had some errors, lost some data
  - 86 towns with no prescribers were dropped from analysis

Benzodiazepine EDA

- Medicare general prescriber files very rich in data, only pulled out a small chunk
- Most drugs are generics
- Drugs:
    - Alprazolam - Xanax
    - Diazepam - Valium
    - Lorazepam - Ativan



Medicare data normalization

- Concern:
  - Typical recipient of Medicare: aged 65+
  - Typical individual that abuses/overdoses on opioids: 20-30s
- Opioid prescriber dataset - opioid claims already normalized to total Medicare claims
- Benzodiazepine data:
  - Pulled out population count of age 65+ from ACS
  - Divided claim counts by population age 65+

Modeling strategy

- Generalized Additive Models (GAMs)
  - Interpretable(ish)
  - Work with geospatial data
  - Work with time series data
  - Outcomes can be non-Gaussian
- Facebook Prophet is a GAM, but more focused on time series component
- pyGAM - GAMs implemented in Python with similar syntax to scikit learn - limited
- R mgcv package - maintained by one of leaders in GAM field, lots of tutorials, flexible package

Modeling steps: Feature summary

Base
- latitude, longitude
- year
- tot population

Demographics
- avg income
- income / poverty
- town grown / shrunk
- below HS education

Medicare opioid prescriptions:
- avg prescription rate (opioid claims / total claims)

Medicare benzo prescriptions:
- alprazolam / pop age 65+
- lorazepam / pop age 65+
- diazepam / pop age 65 +
- tot benzo / pop age 65 +

1 year lag between features and outcome
- Assumed previous year will influence next year
- But also Medicare data was only from 2013-2017

Modeling steps: Training and validating

Base
- latitude, longitude
- year
- tot population

Demographics
- avg income
- income / poverty
- town grown / shrunk
- below HS education

Medicare opioid prescriptions:
- avg prescription rate (opioid claims / total claims)

Medicare benzo prescriptions:
- alprazolam / pop age 65+
- lorazepam / pop age 65+
- diazepam / pop age 65 +
- tot benzo / pop age 65 +

mgcv

Train: 2014 - 2016
Validation: 2017

Modeling steps: Model variants and errors

Base
- latitude, longitude
- year
- tot population

Demographics
- avg income
- income / poverty
- town grown / shrunk
- below HS education

Medicare opioid prescriptions:
- avg prescription rate (opioid claims / total claims)

Medicare benzo prescriptions:
- alprazolam / pop age 65+
- lorazepam / pop age 65+
- diazepam / pop age 65 +
- tot benzo / pop age 65 +

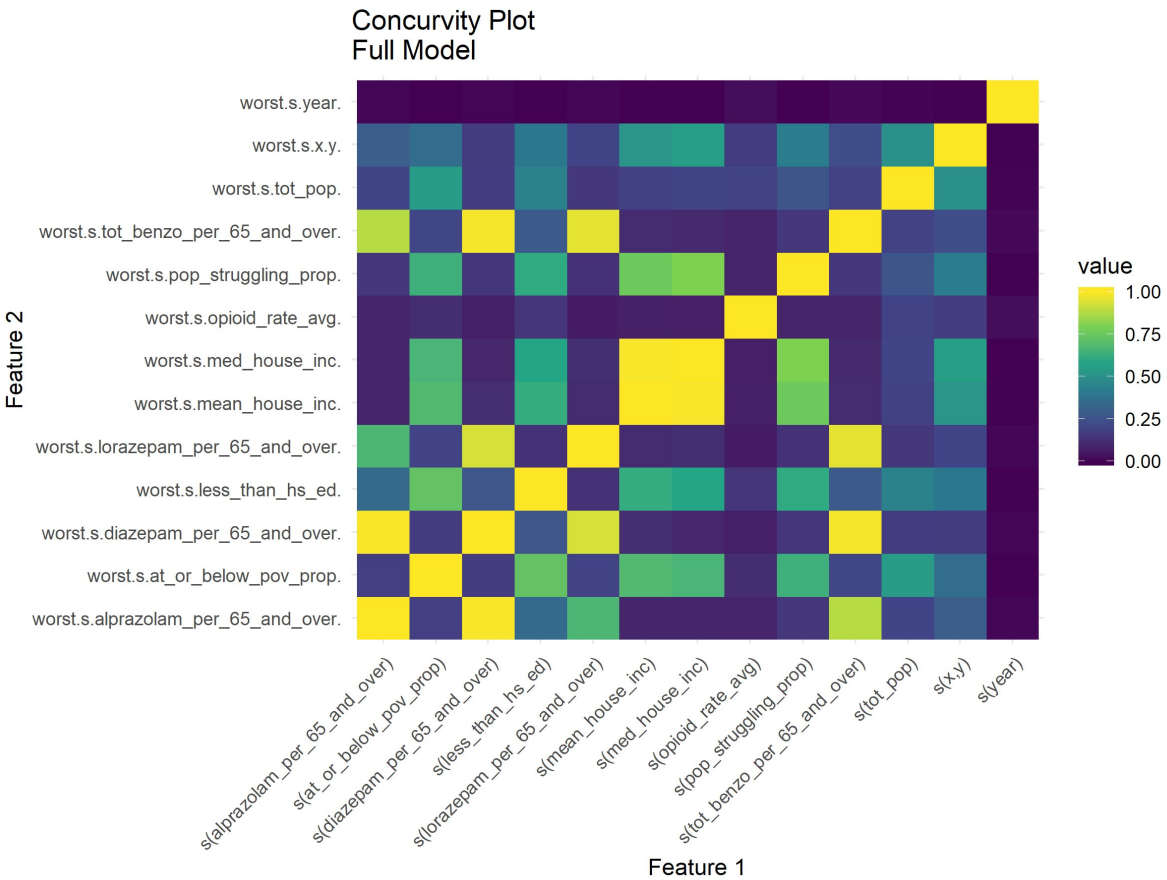
Full RMSE:
- **Train:** 3.78 (vs M 6.67, SD 18.63)
- **Valid:** 8.46
- **Adj R^2** - 0.956 (var explained)

Base RMSE:
- **Train:** 5.40 (vs M 6.67, SD 18.63)
- **Valid:** 8.65
- **Adj R^2** - 0.913

Final model RMSE
(-2 Demo and 2 Benzo features):
- **Train:** 4.41 (vs M 6.67, SD 18.63)
- **Valid:** 7.71
- **Adj R^2** - 0.942 (var explained)

Concurvity

- Generalization of collinearity
- 1 = perfect relationship with another feature
- 0 = no relationship with another feature
- “Rule of thumb” found online = 0.85 for worst estimate



Modeling steps: Training and validating outcome

Base
- latitude, longitude
- year
- tot population

Demographics
- avg income
- income / poverty
- town grown / shrunk
- below HS education

Medicare opioid prescriptions:
- avg prescription rate (opioid claims / total claims)

Medicare benzo prescriptions:
- alprazolam / pop age 65+
- lorazepam / pop age 65+
- diazepam / pop age 65 +
- tot benzo / pop age 65 +

mgcv

Train: 2014 - 2016
Validation: 2017

Best Validation RMSE = 7.7 deaths per town per year

Compared to mean (6.7) and SD (18.6) - model good for high population towns, terrible for small low population towns

Modeling steps: Final error

Base
- latitude, longitude
- year
- tot population

Demographics
- avg income
- income / poverty
- town grown / shrunk
- below HS education

Medicare opioid prescriptions:
- avg prescription rate (opioid claims / total claims)

Medicare benzo prescriptions:
- alprazolam / pop age 65+
- lorazepam / pop age 65+
- diazepam / pop age 65 +
- tot benzo / pop age 65 +

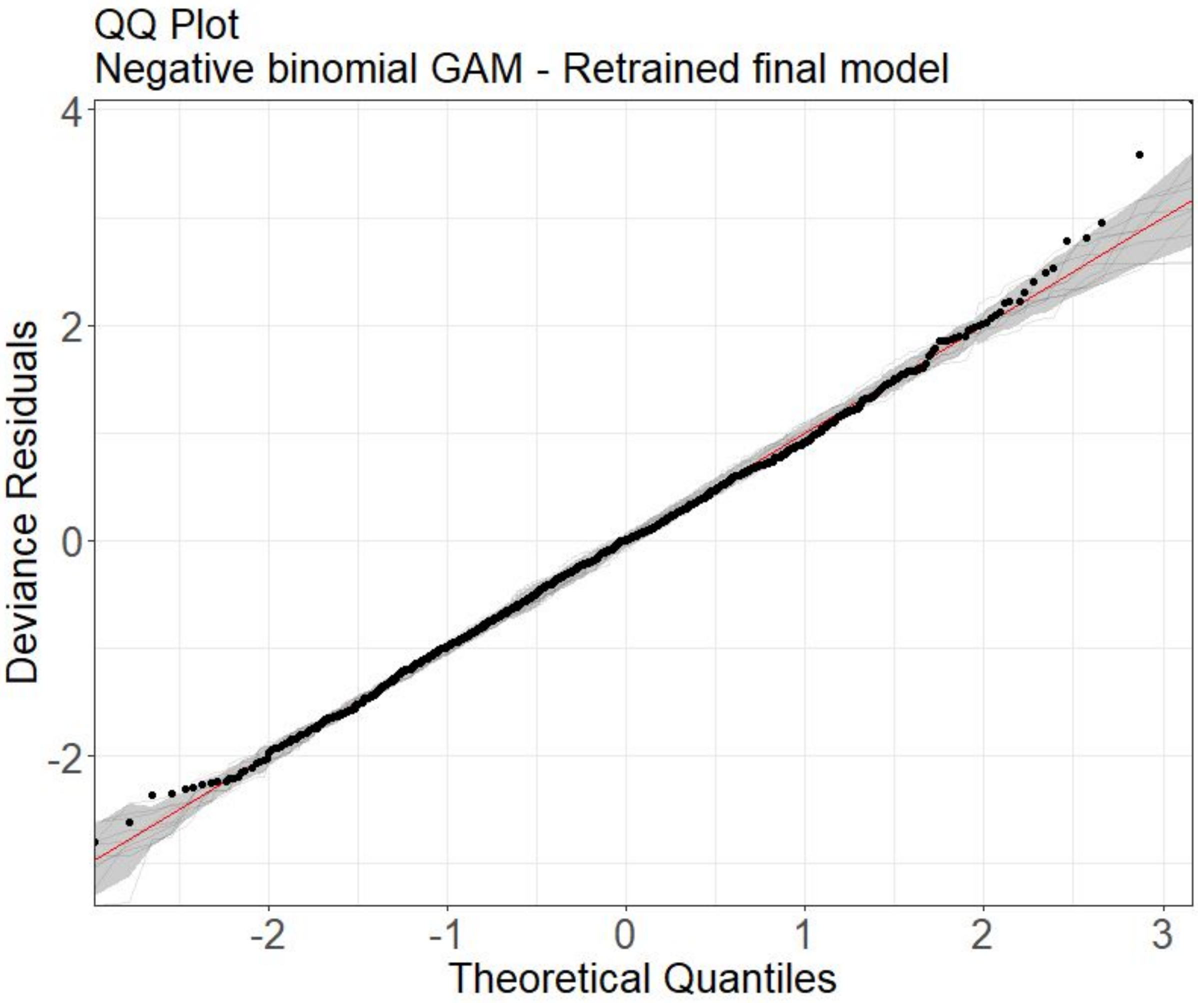
mgcv

Train: 2014 - 2017
Test: 2018

Final Test RMSE = 5.3 deaths per town per year

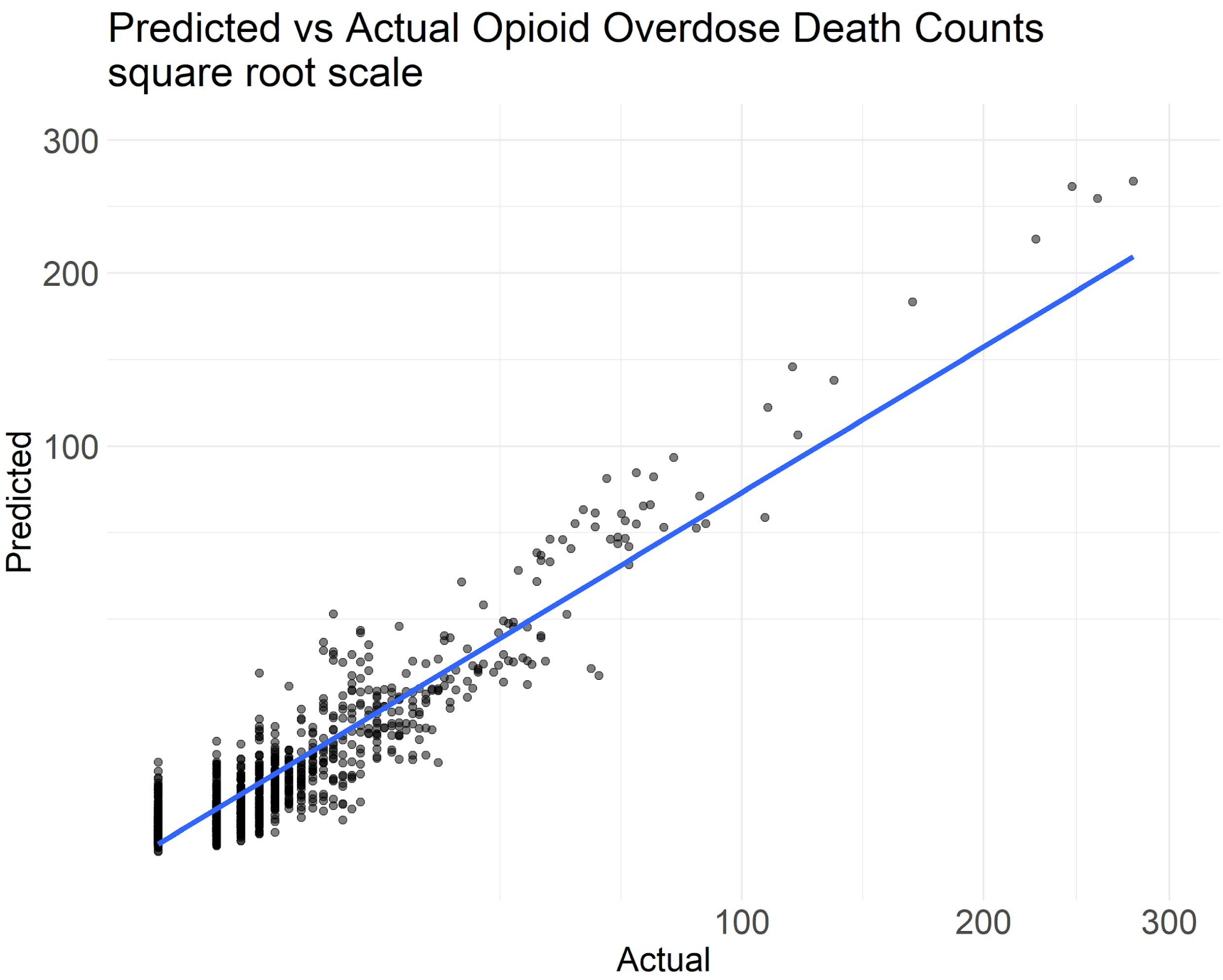
Evaluating fit: QQ Plot

Indicates that model captured the data fairly well, except for the extremes

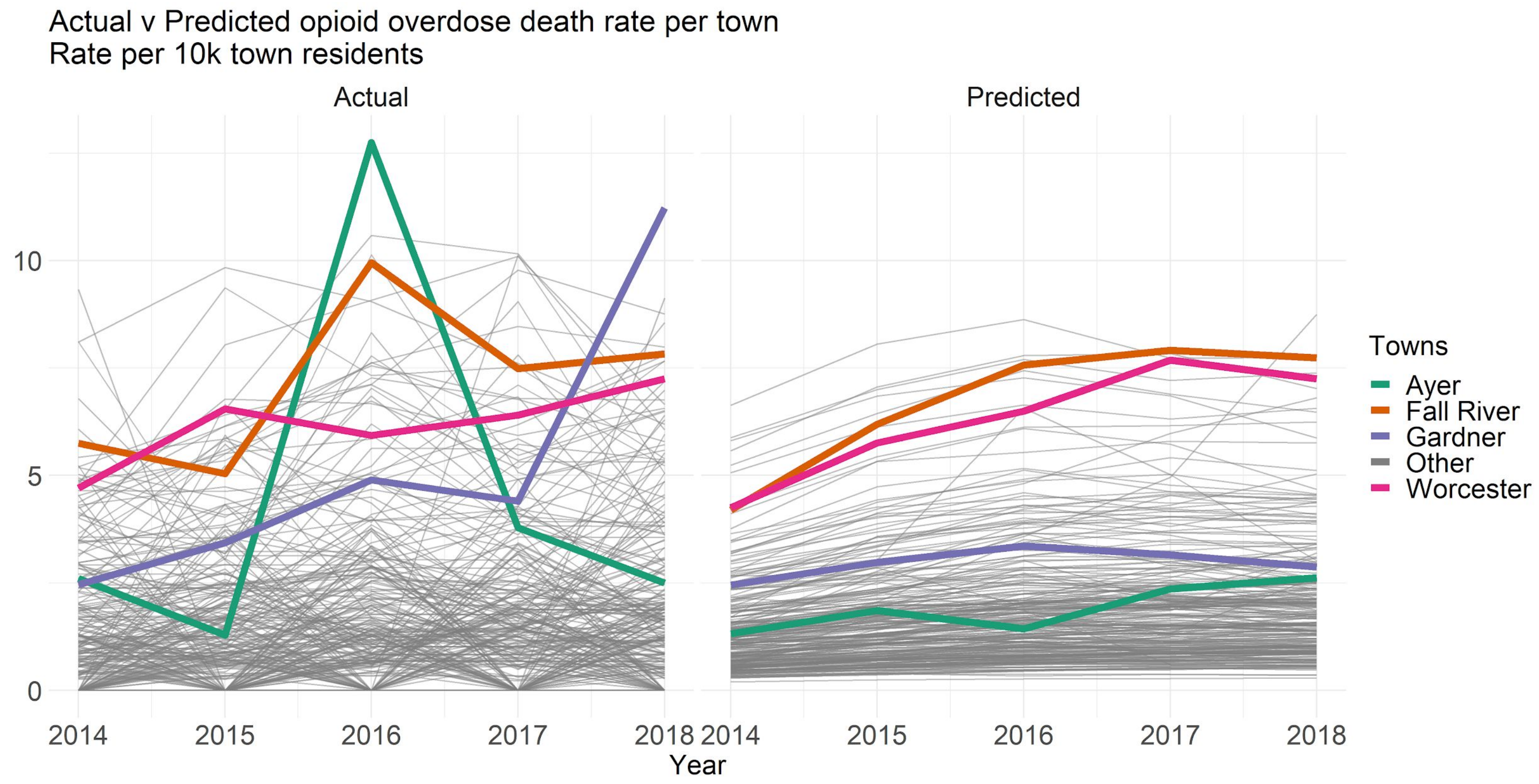


Evaluating fit: Predicted vs Actual (All years)

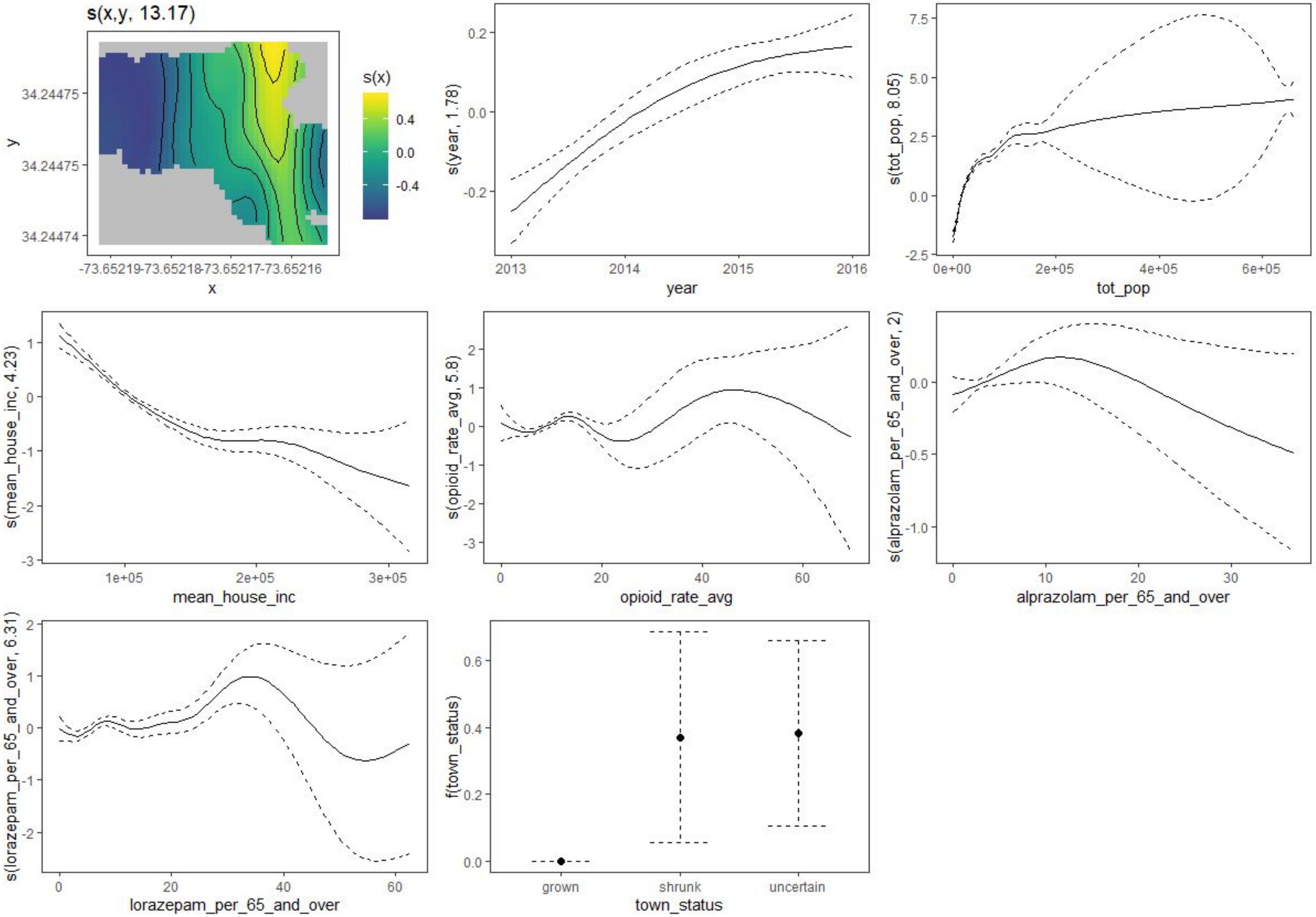
Indicates that model model overestimated very low values



The trained GAM was better at predicting consistent death rates



Interpreting the model: GAM smooths



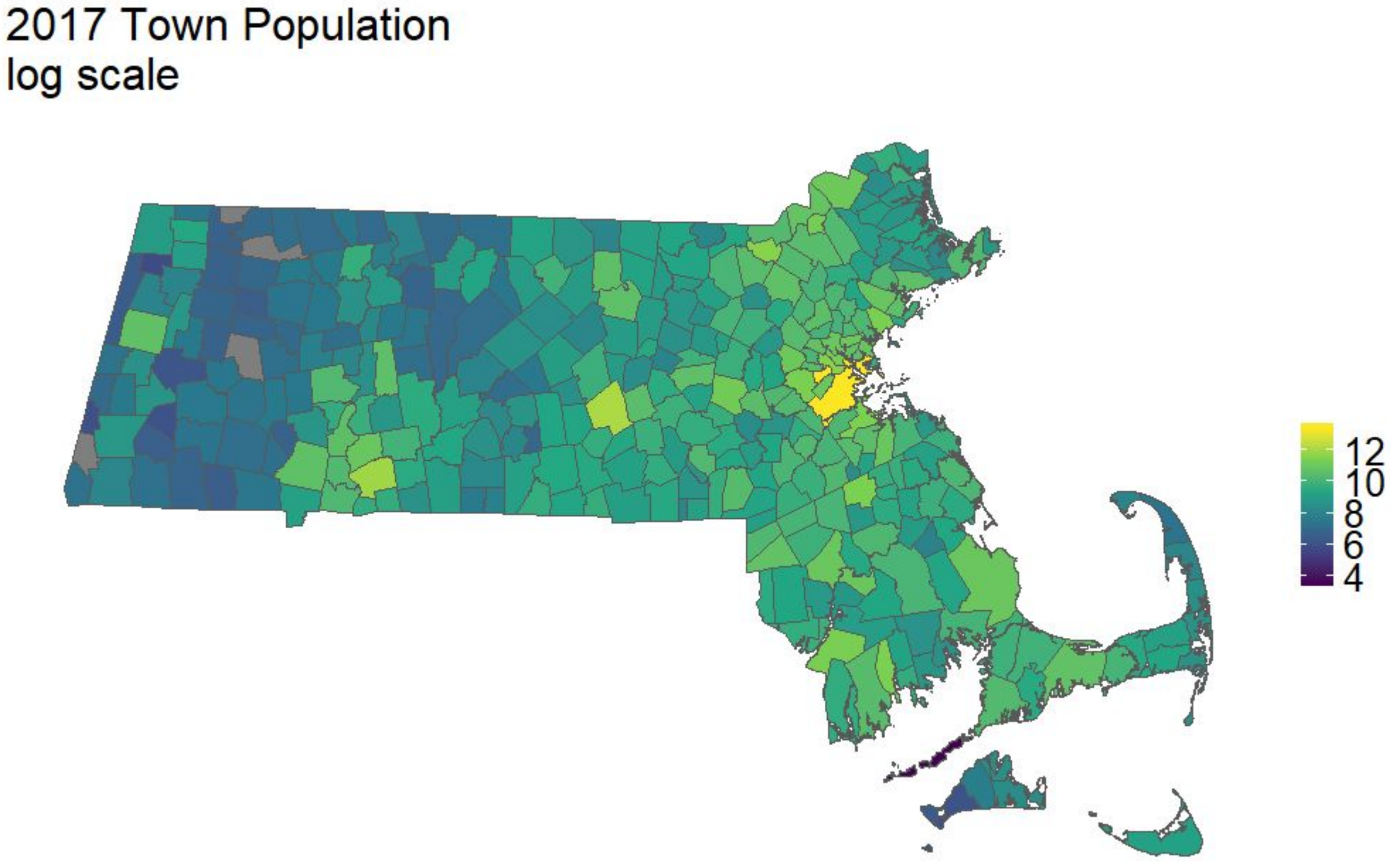
For a GAM, coefficients don't mean much, but can pull out plots of the relationship of the outcome with each individual feature

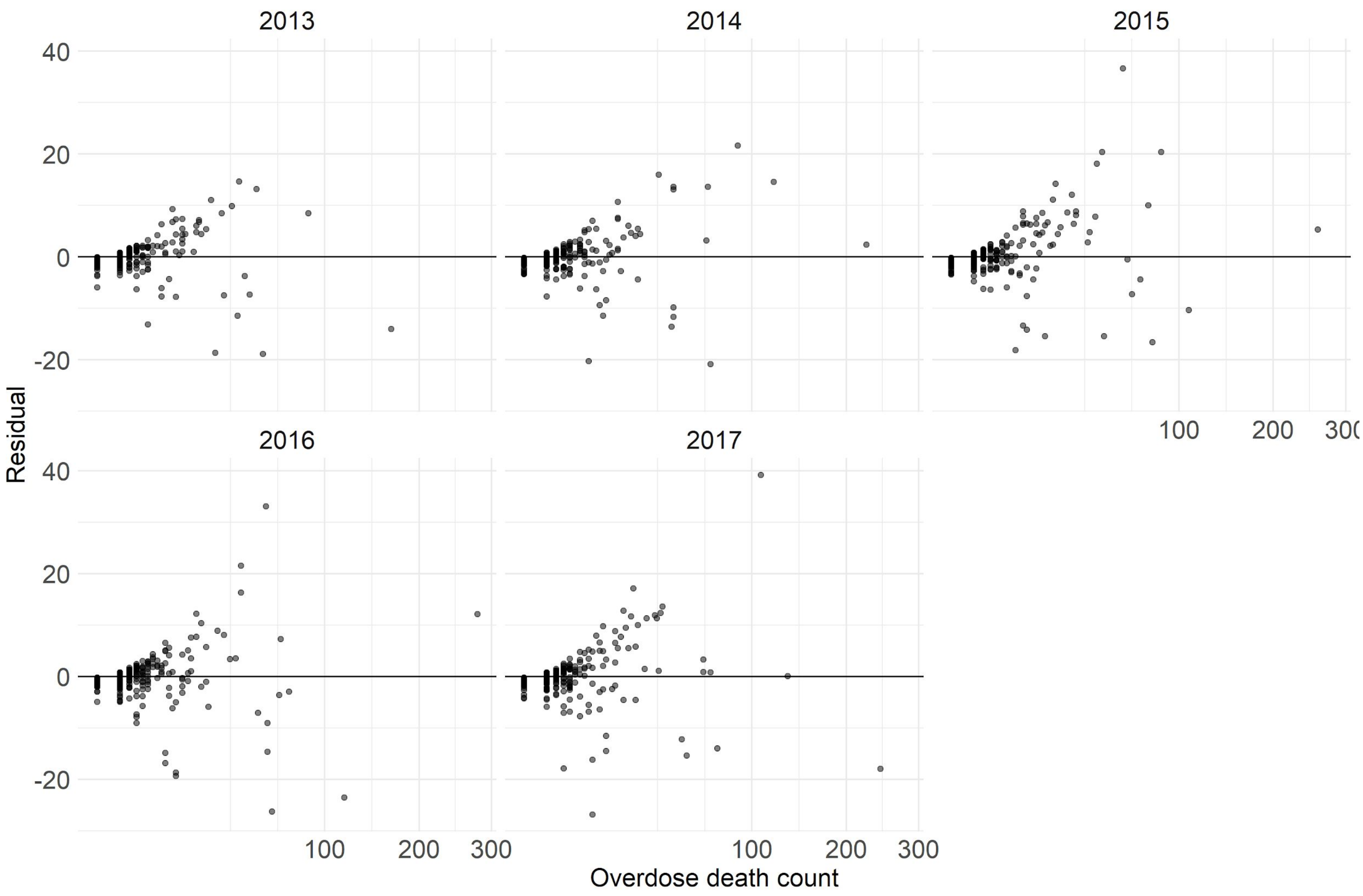
Dashed lines indicate 2 standard error

Future directions/improvements

- Problem - geospatial count and rate data can have a variety of problems. Geospatial data can be aggregated in different units and that can create seemingly interesting patterns where there are none
    - https://mgimond.github.io/Spatial/pitfalls-to-avoid.html
    - Solution: Geospatial smoothing
    - Could also be used to “smooth” drug prescriptions, etc
- Problem - Opioid epidemic is changing from prescription drugs to illegal drugs
    - Medicare features probably won’t be that useful going forward
    - Treatment clinics, crime rate, economic factors could be useful
- Mass.gov also releases a dataset of EMS cases by town that involved naloxone, but didn’t get a chance to incorporate

Thank you! Questions?


2017 Town Population
log scale
12
10
8
6
4


2013
2014
2015
2016
2017
40
20
0
-20
100
200
300
Residual
Overdose death count