It Takes a Village

ELI WINTON, EMILY K. SANDERS, RADHA MOHANTY

A Presentation for the *Coalition for Unprecedented Transformative Education for California Teachers and Students*, 2024 Meeting

Problem Statement

- Dropping out of high school impairs a child's opportunities for the rest of their life.
- You all want to help!
- We used government data to predict dropout rates from social conditions at the county level in California.
- Goals: Plan community actions, set political goals

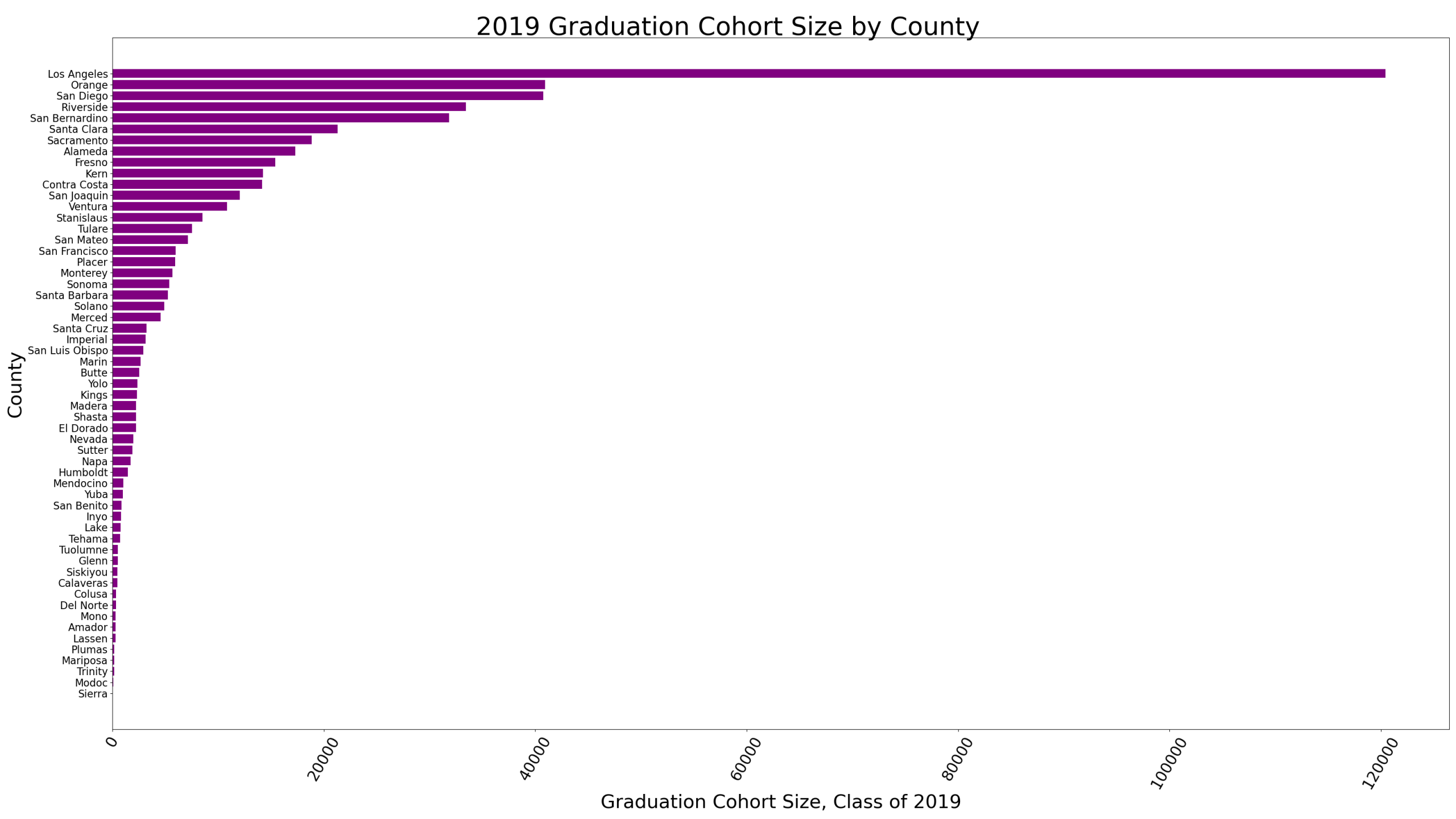
Definitions

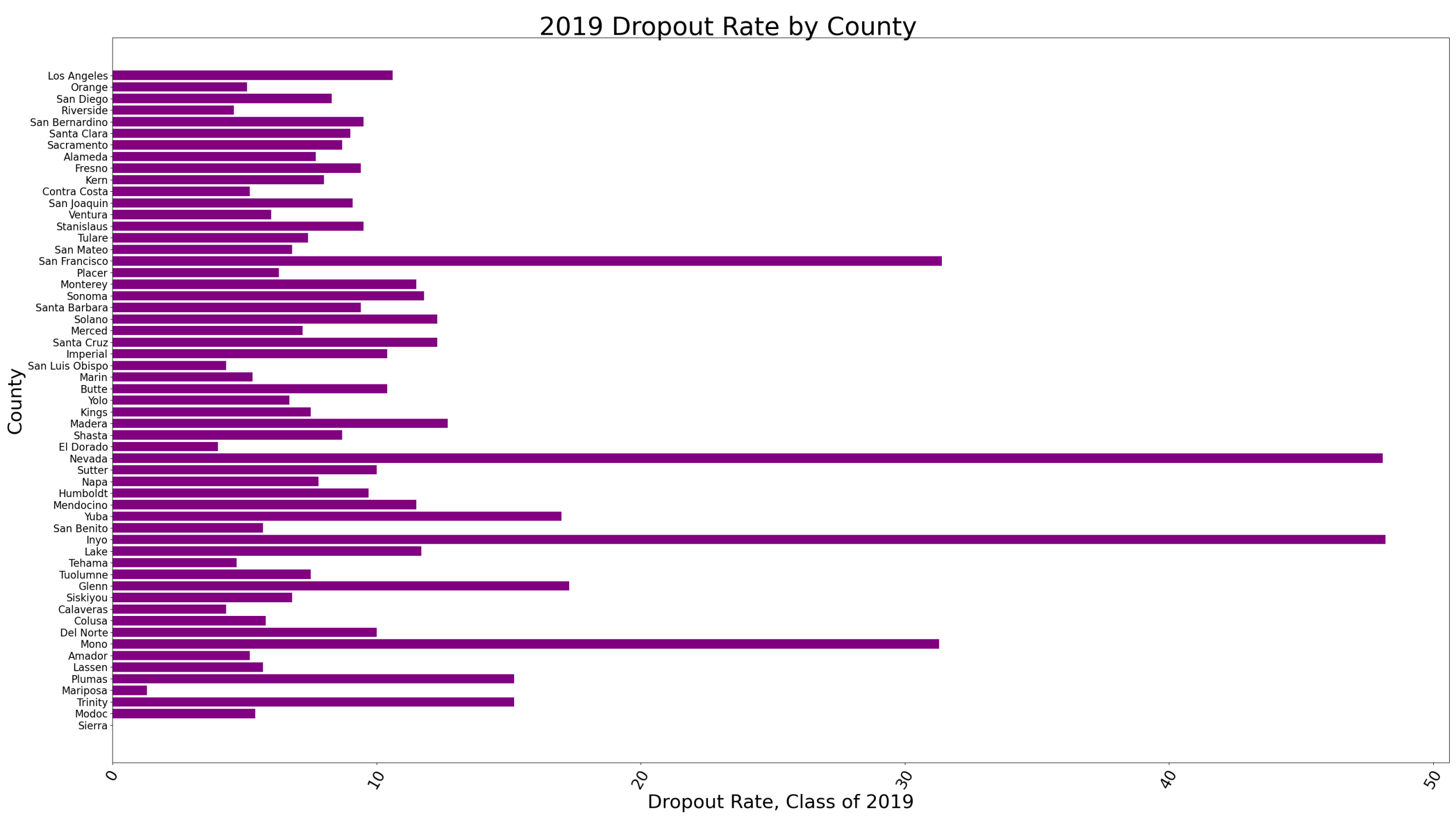
Graduation Cohort

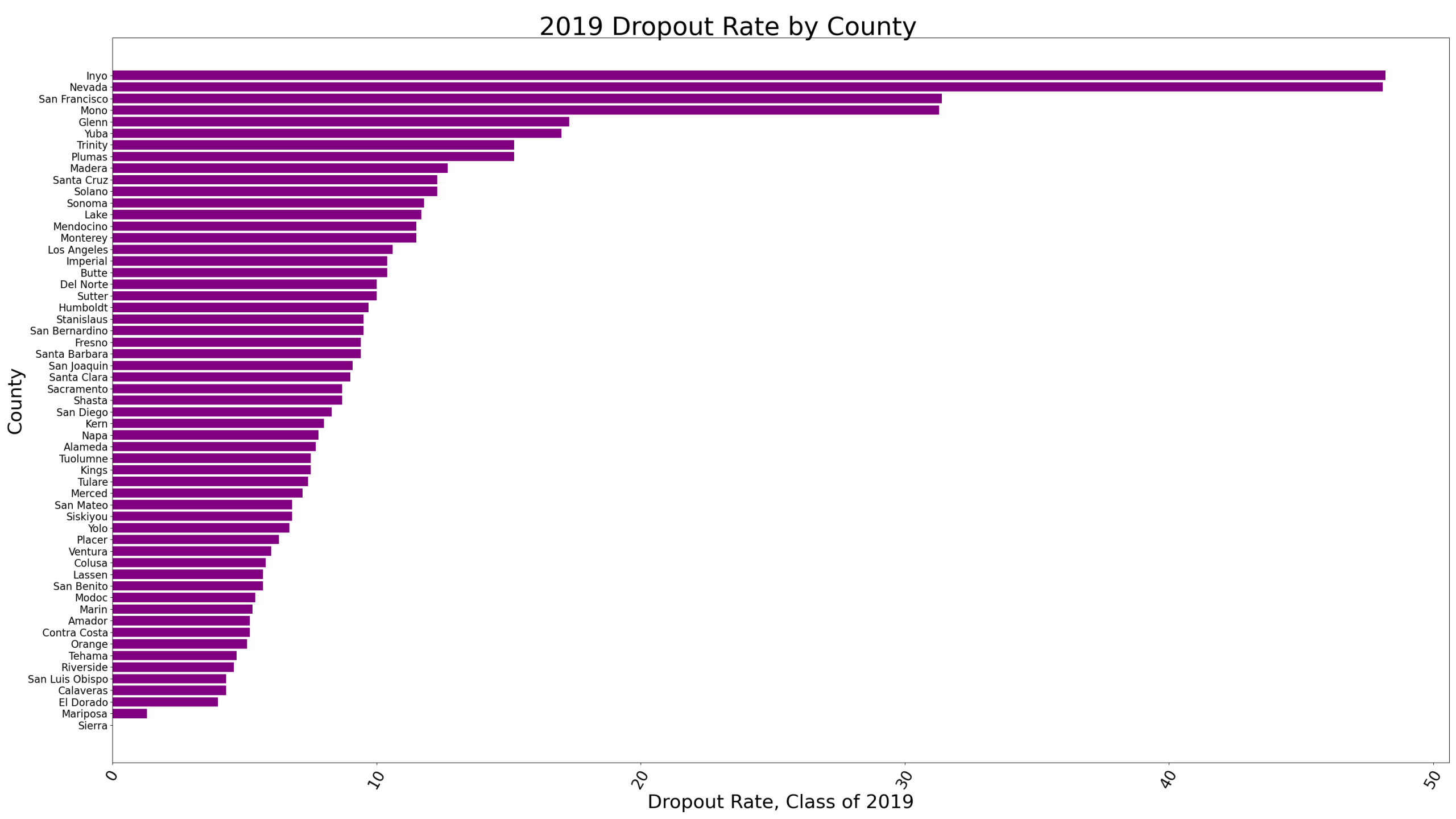
- All students who enter 9th grade at the same time
- Plus/minus students who leave or join the school
- Public school only

Dropout Rate

- Percentage of the graduation cohort who:
- Do not complete in any way
- Are not still enrolled


2019 Graduation Cohort Size by County
County
Los Angeles
Orange
San Diego
Riverside
San Bernardino
Santa Clara
Sacramento
Alameda
Fresno
Kern
Contra Costa
San Joaquin
Ventura
Stanislaus
Tulare
San Mateo
San Francisco
Placer
Monterey
Sonoma
Santa Barbara
Solano
Merced
Santa Cruz
Imperial
San Luis Obispo
Marin
Butte
Yolo
Kings
Madera
Shasta
El Dorado
Nevada
Sutter
Napa
Humboldt
Mendocino
Yuba
San Benito
Inyo
Lake
Tehama
Tuolumne
Glenn
Siskiyou
Calaveras
Colusa
Del Norte
Mono
Amador
Lassen
Plumas
Mariposa
Trinity
Modoc
Sierra
0
20000
40000
60000
80000
100000
120000
Graduation Cohort Size, Class of 2019


2019 Dropout Rate by County
County
Los Angeles
Orange
San Diego
Riverside
San Bernardino
Santa Clara
Sacramento
Alameda
Fresno
Kern
Contra Costa
San Joaquin
Ventura
Stanislaus
Tulare
San Mateo
San Francisco
Placer
Monterey
Sonoma
Santa Barbara
Solano
Merced
Santa Cruz
Imperial
San Luis Obispo
Marin
Butte
Yolo
Kings
Madera
Shasta
El Dorado
Nevada
Sutter
Napa
Humboldt
Mendocino
Yuba
San Benito
Inyo
Lake
Tehama
Tuolumne
Glenn
Siskiyou
Calaveras
Colusa
Del Norte
Mono
Amador
Lassen
Plumas
Mariposa
Trinity
Modoc
Sierra
0
10
20
30
40
50
Dropout Rate, Class of 2019

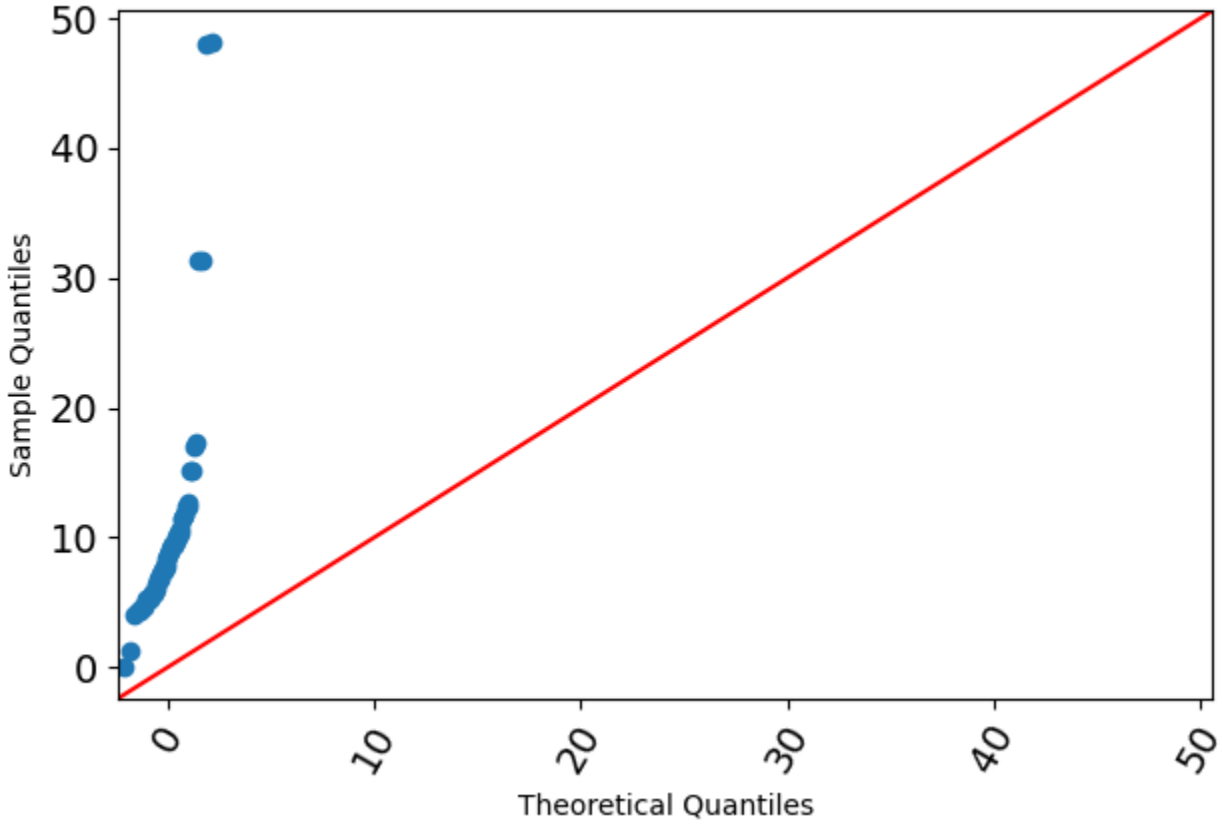

2019 Dropout Rate by County
County
Inyo
Nevada
San Francisco
Mono
Glenn
Yuba
Trinity
Plumas
Madera
Santa Cruz
Solano
Sonoma
Lake
Mendocino
Monterey
Los Angeles
Imperial
Butte
Del Norte
Sutter
Humboldt
Stanislaus
San Bernardino
Fresno
Santa Barbara
San Joaquin
Santa Clara
Sacramento
Shasta
San Diego
Kern
Napa
Alameda
Tuolumne
Kings
Tulare
Merced
San Mateo
Siskiyou
Yolo
Placer
Ventura
Colusa
Lassen
San Benito
Modoc
Marin
Amador
Contra Costa
Orange
Tehama
Riverside
San Luis Obispo
Calaveras
El Dorado
Mariposa
Sierra
0
10
20
30
40
50
Dropout Rate, Class of 2019

Method

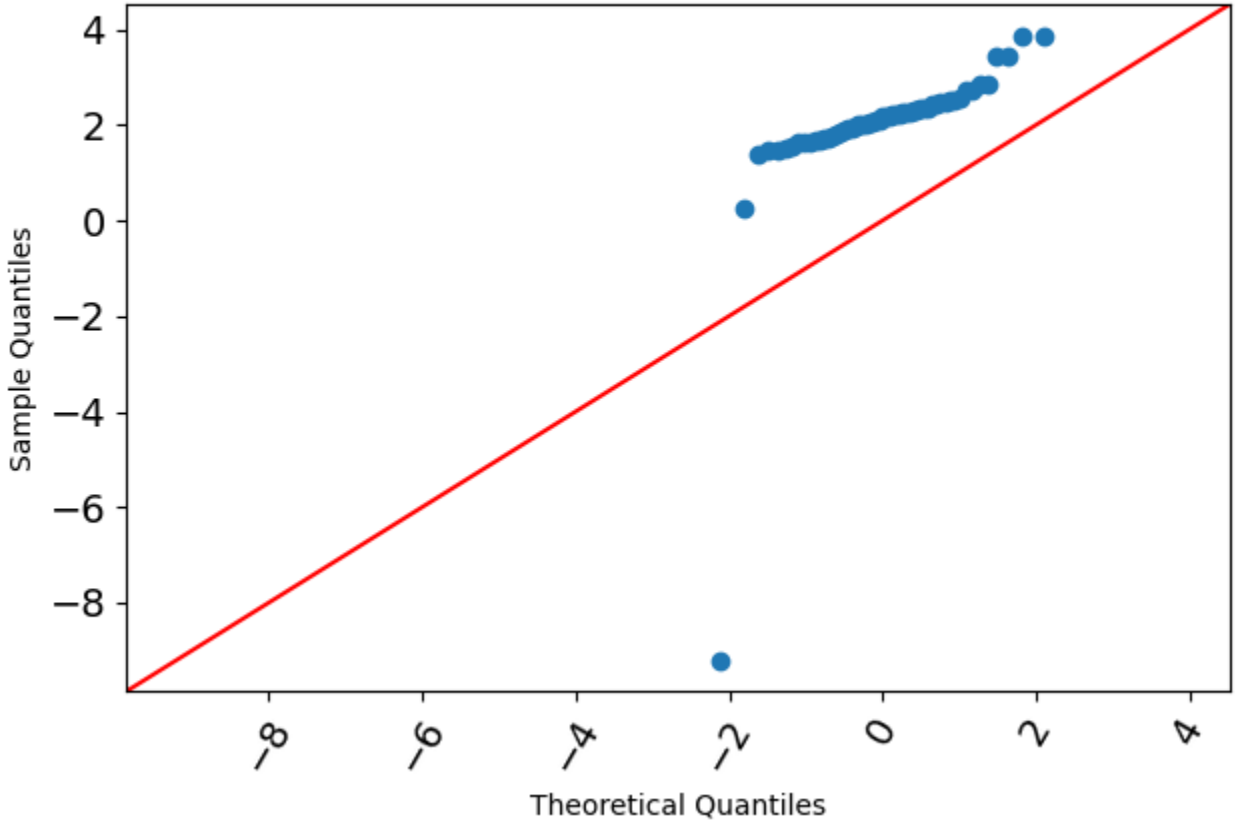
- 14 government datasets
- 1 row per county, 100s of features (at first)
- All predictors came from ~2015, dropout rate from 2019
- Feature engineering: condense, impute, polynomialize
- Feature selection: correlations, LASSO

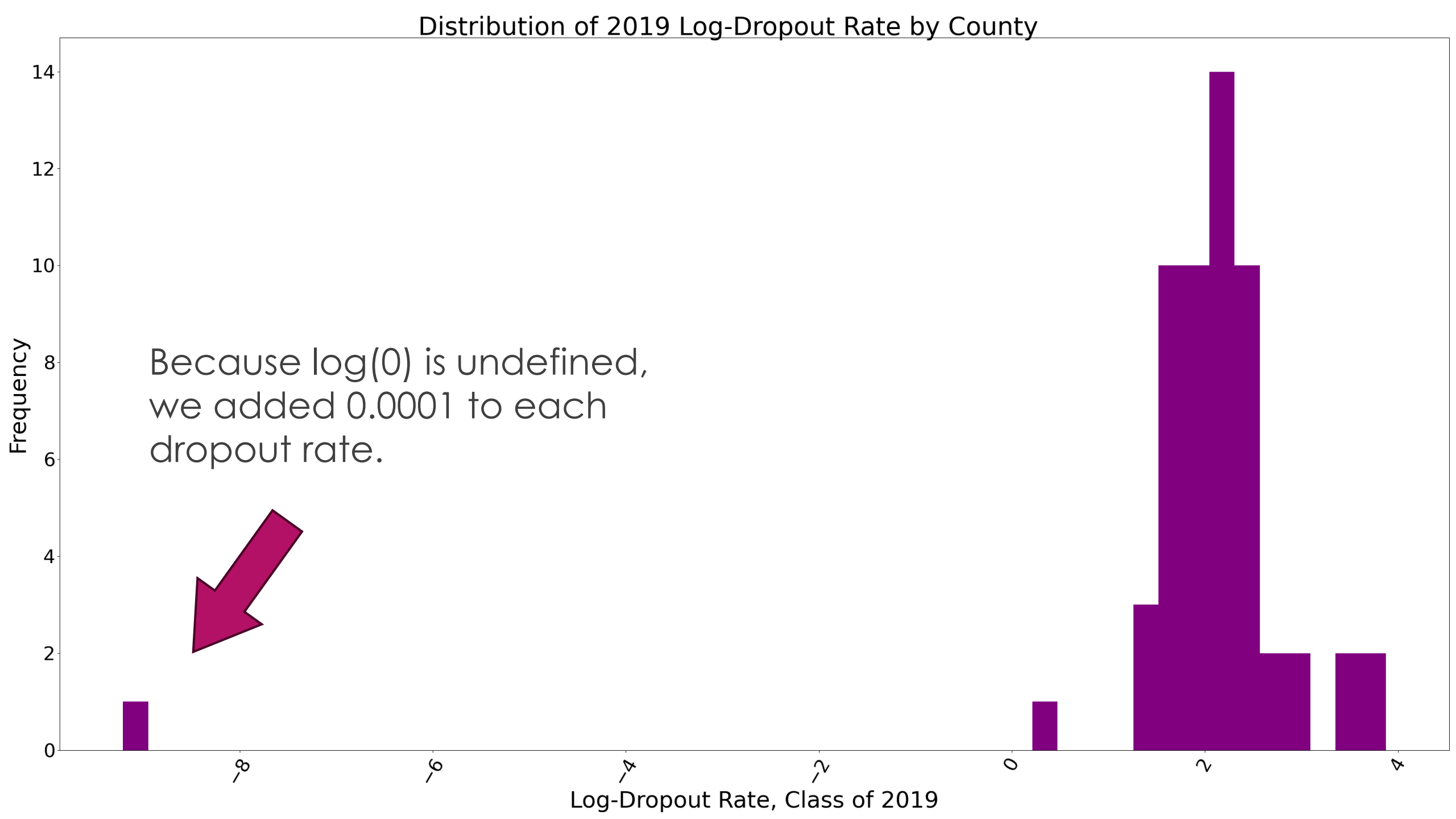
Distribution Woes

QQ-Plot of 2019 Dropout Rate by County



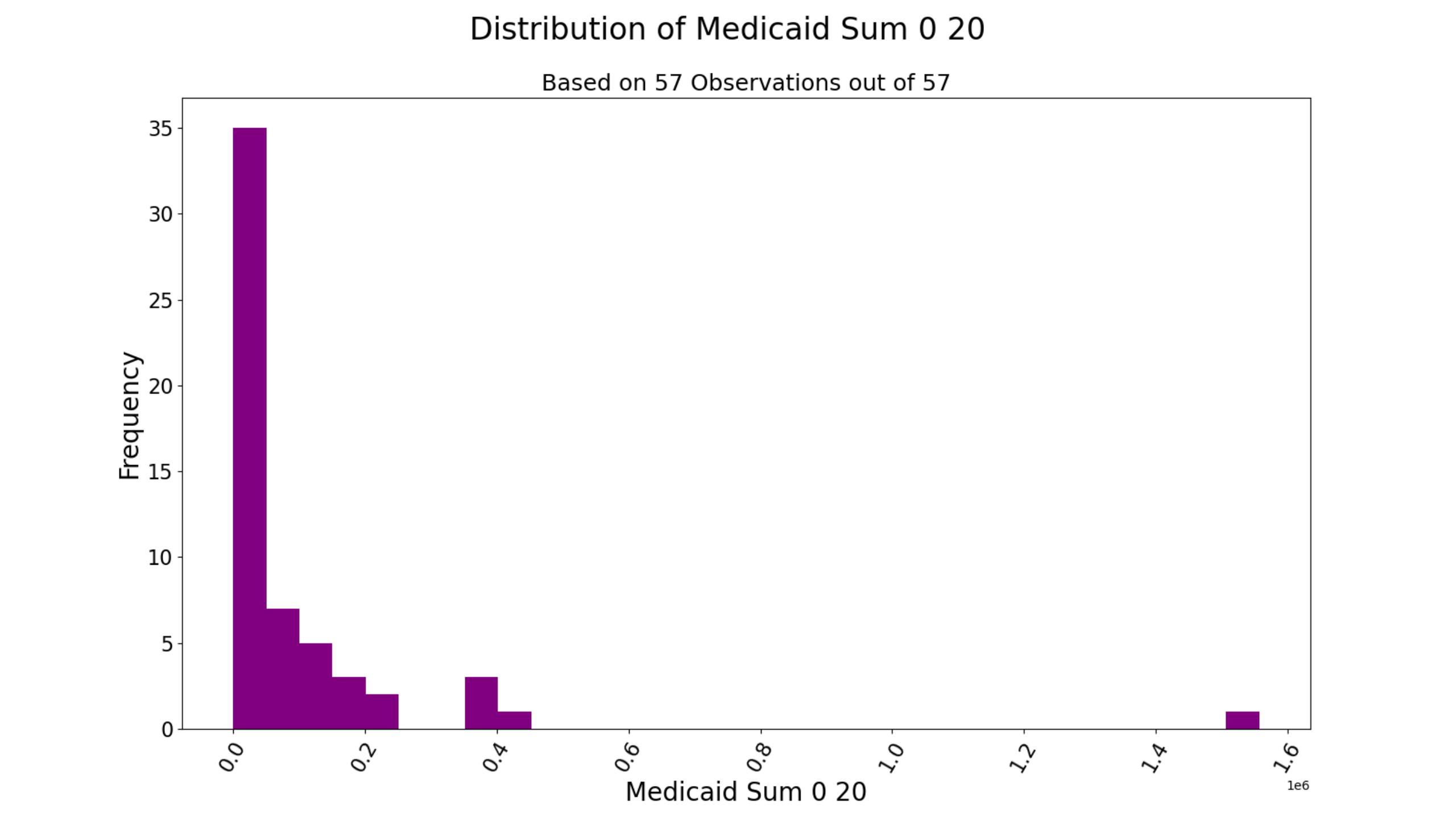
QQ-Plot of 2019 Log Dropout Rate by County

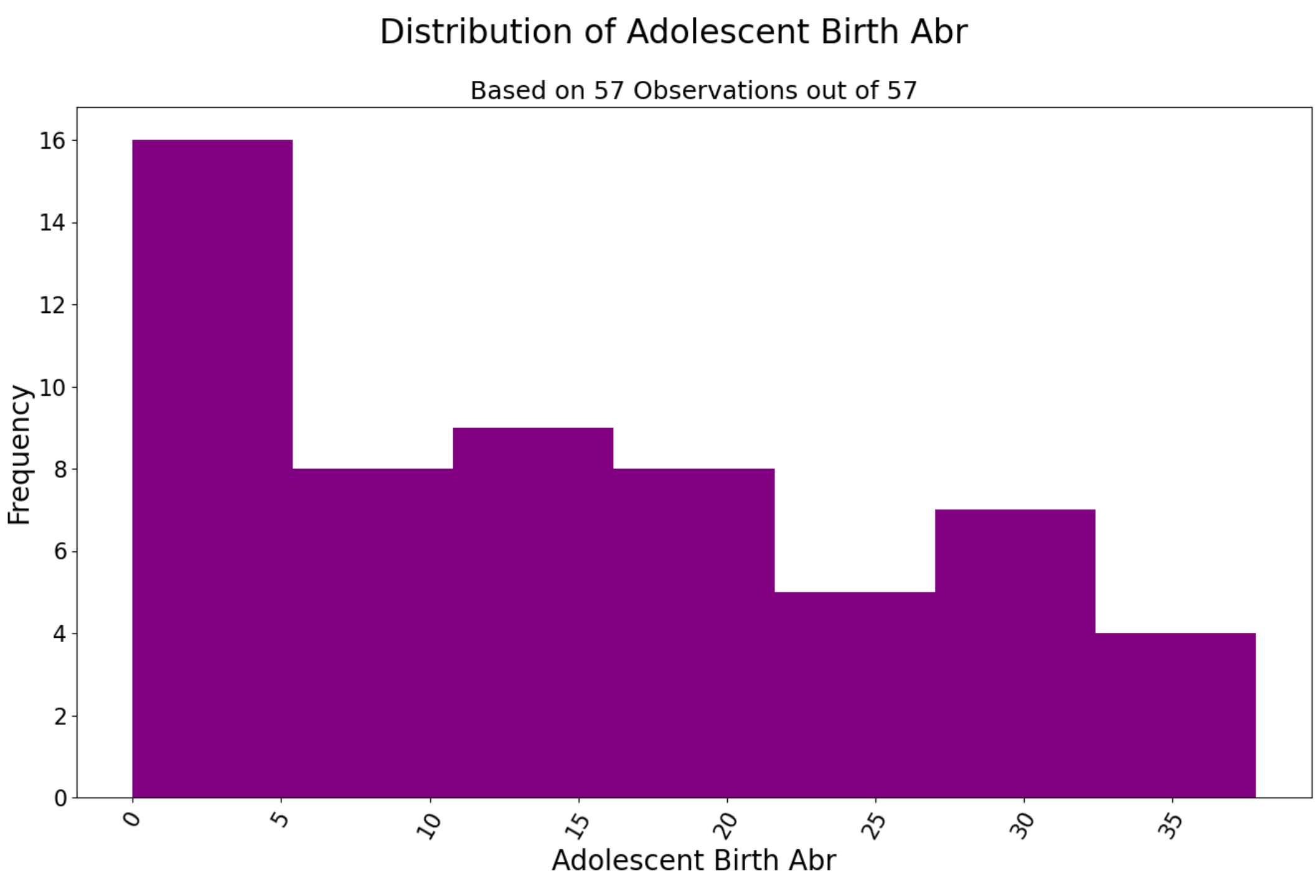


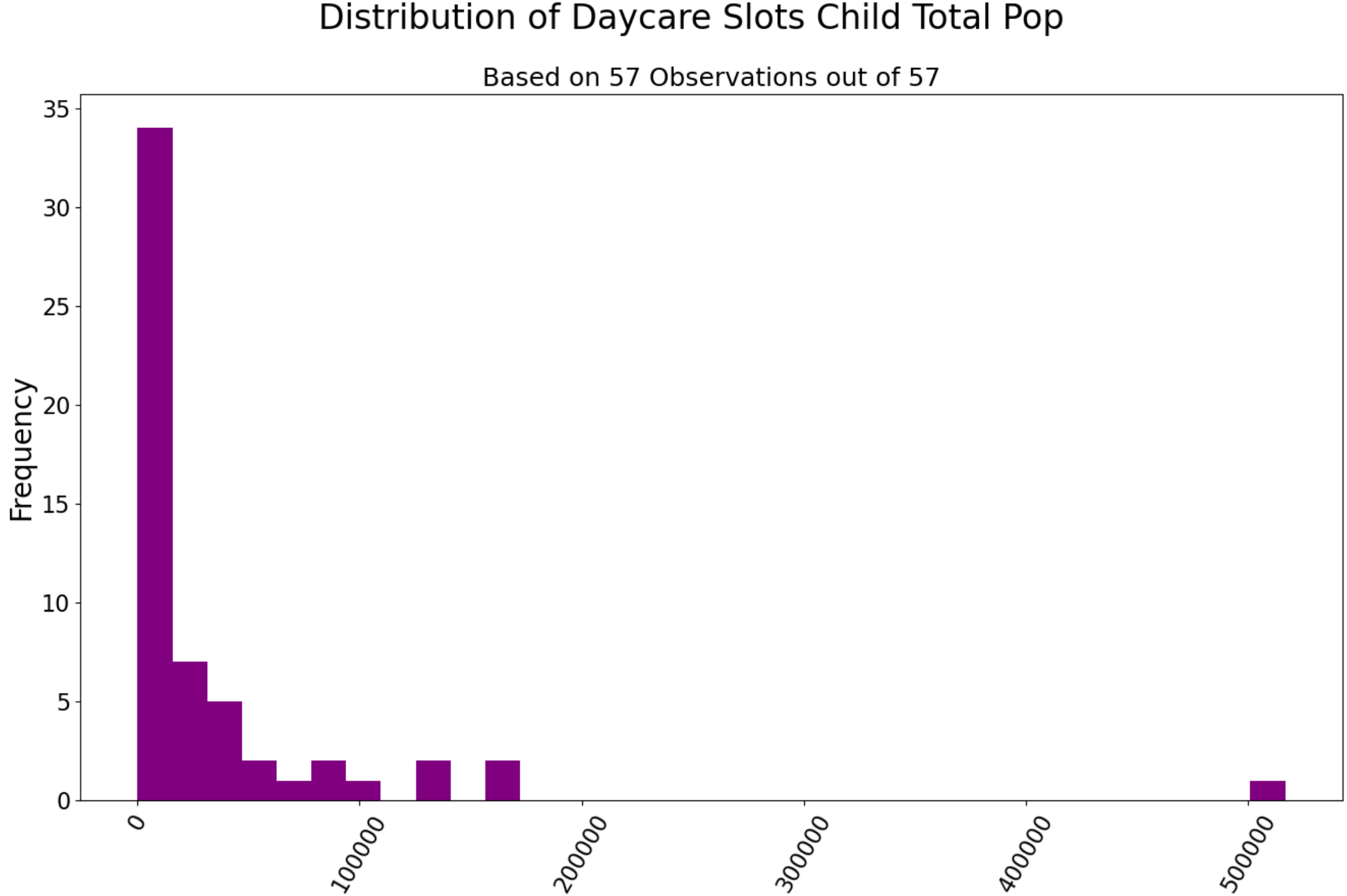

Distribution of 2019 Log-Dropout Rate by County
Frequency
0
2
4
6
8
10
12
14
Because log(0) is undefined, we added 0.0001 to each dropout rate.
−8
−6
−4
−2
0
2
4
Log-Dropout Rate, Class of 2019

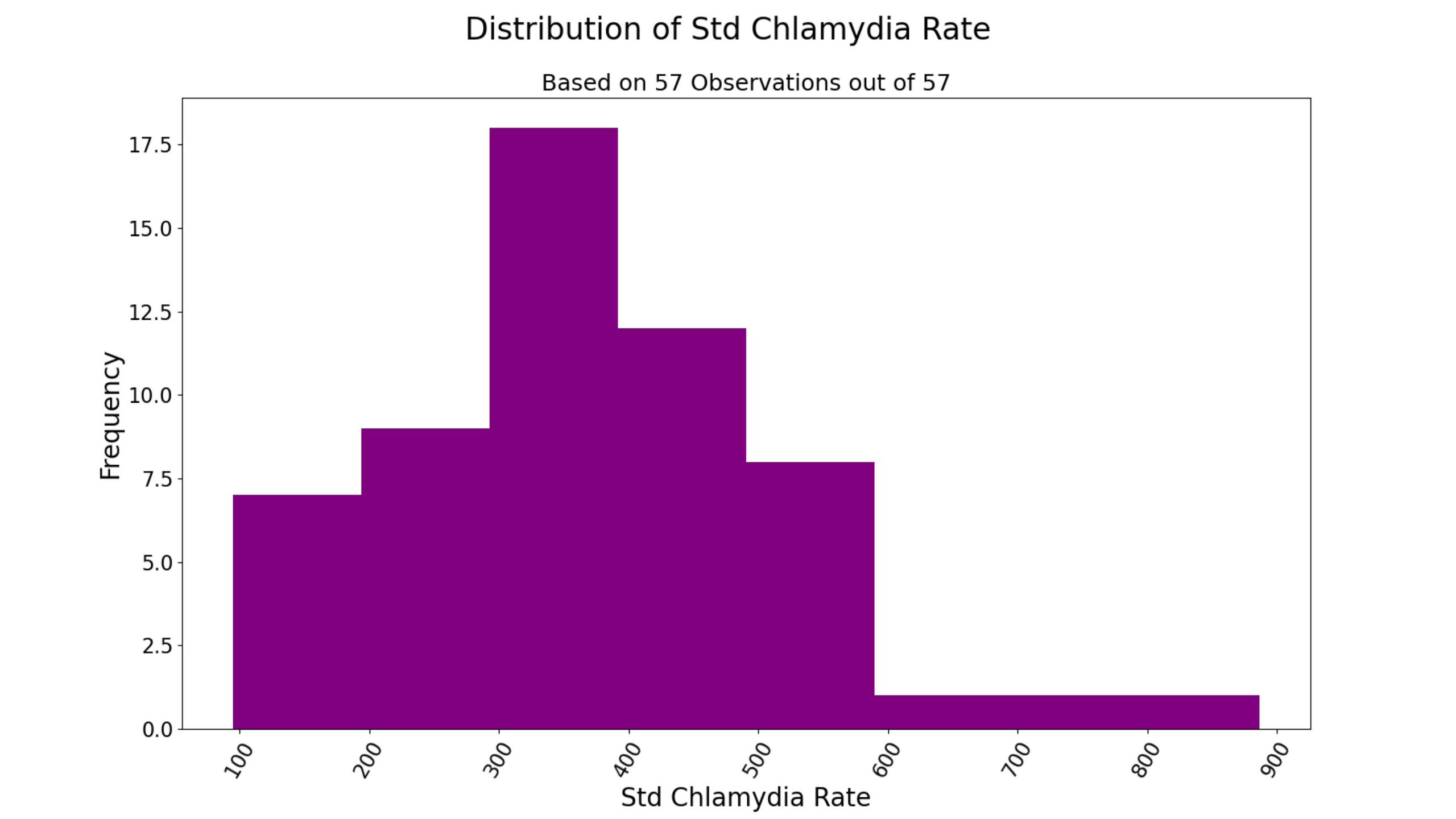
Features

- Population & cohort size
- Poverty & income inequality
- Unemployment rates
- Medicaid enrollees <21yo
- Racial demographics <6yo
- Teenage birth rate
- Availability of ecigarettes
- STD rates
- Suicide, asthma death rates
- Infant mortality
- Medicaid abortion funding
- Daycare availability


Distribution of Medicaid Sum 0 20
Based on 57 Observations out of 57
Frequency
0
5
10
15
20
25
30
35
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1e6
Medicaid Sum 0 20


Distribution of Adolescent Birth Abr
Based on 57 Observations out of 57
Frequency
0
2
4
6
8
10
12
14
16
0
5
10
15
20
25
30
35
Adolescent Birth Abr


Distribution of Daycare Slots Child Total Pop
Based on 57 Observations out of 57
Frequency
0
5
10
15
20
25
30
35
0
100000
200000
300000
400000
500000


Distribution of Std Chlamydia Rate
Based on 57 Observations out of 57
Frequency
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
100
200
300
400
500
600
700
800
900
Std Chlamydia Rate

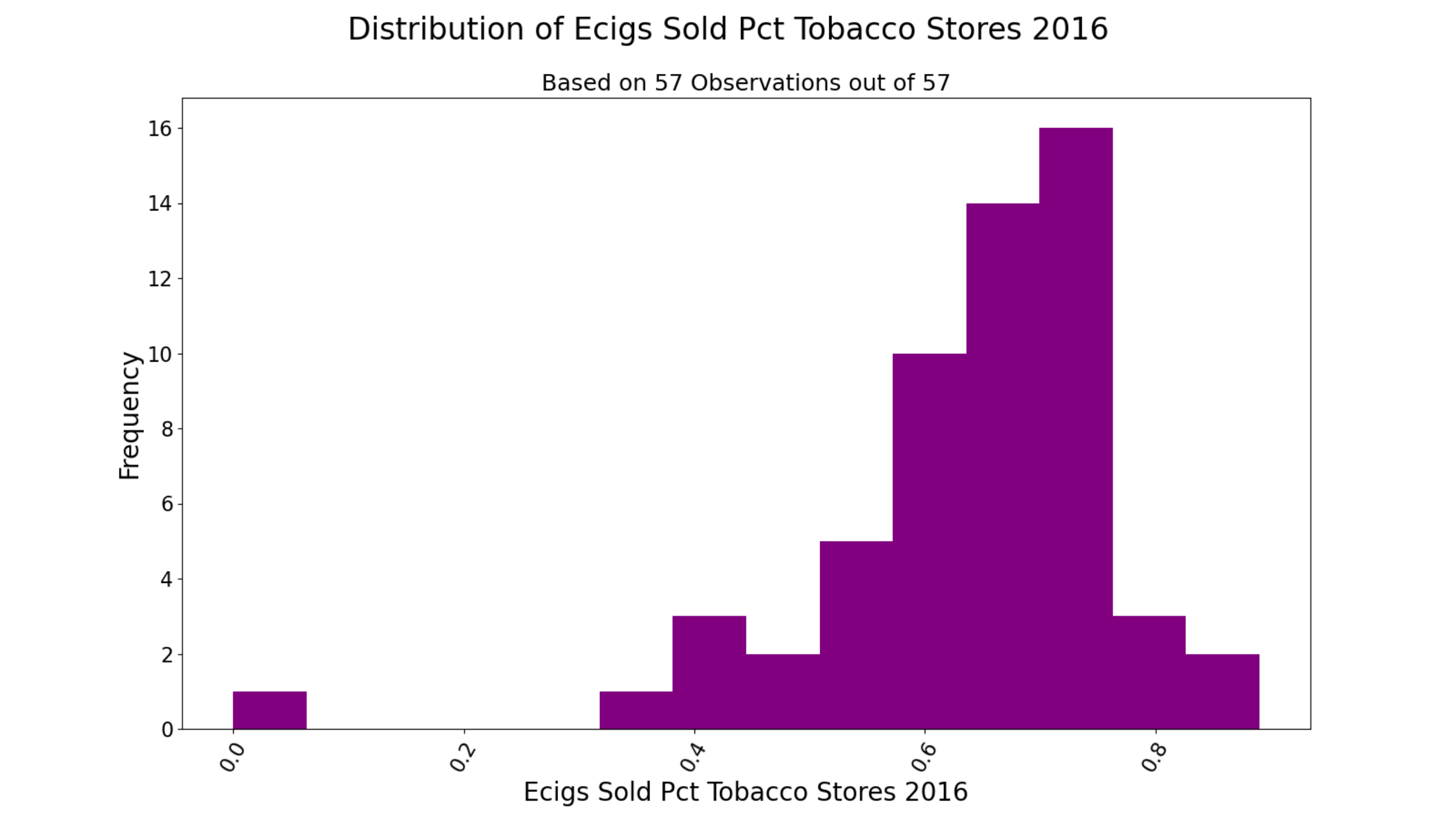
Results

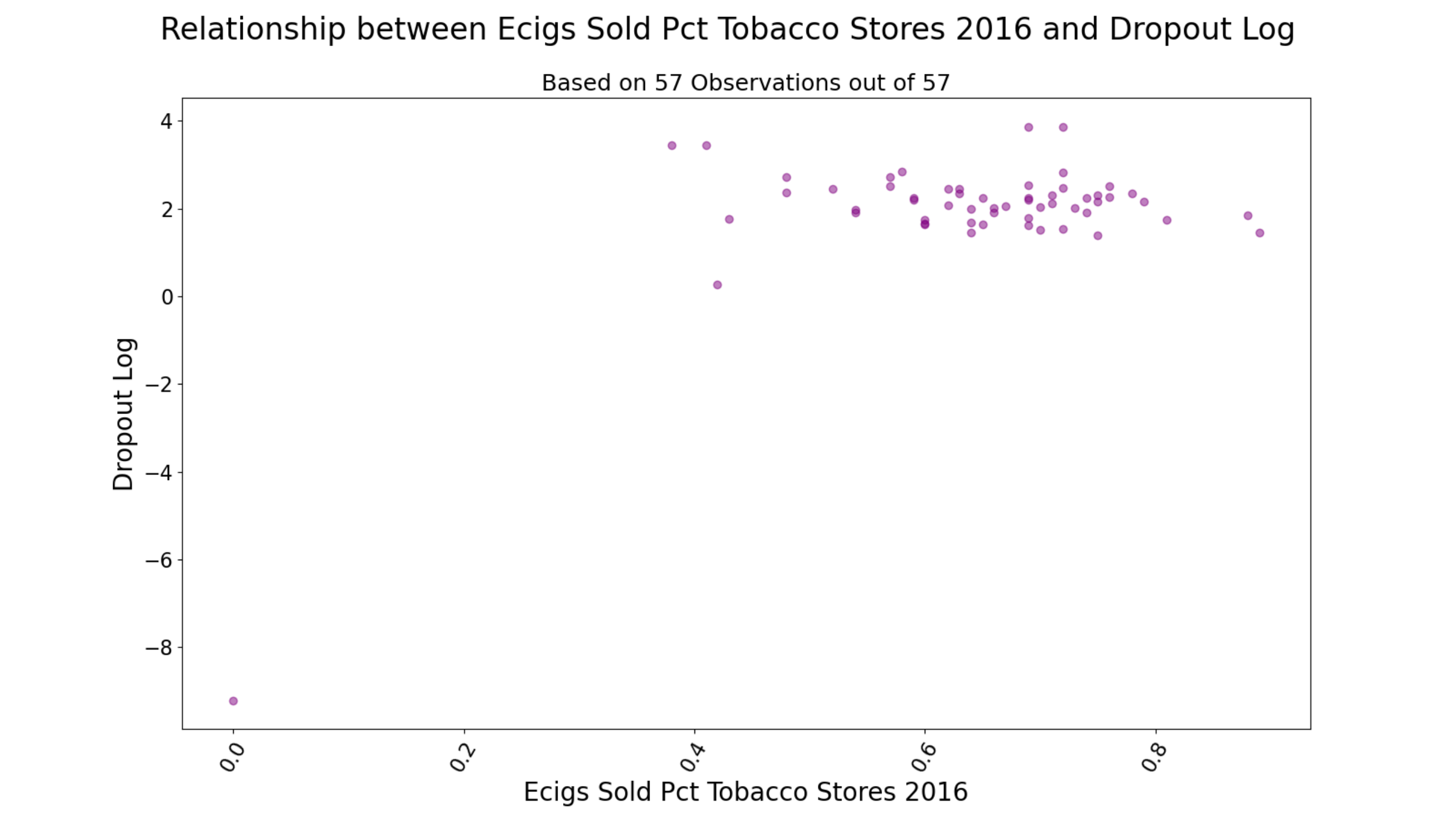
- 10 predictor variables (scaled), 3 types of models
- Regression: LASSO and Random Forest
  - Training: max R2 = 0.83
  - Testing: max R2 = 0.13
- Clustering: Kmeans
  - 13 clusters, poorly defined (silhouette = 0.24)

Results

- Correlations do not imply causation, but…
- Availability of ecigarettes seems bad
- Moreso in combination with other factors than alone

Photo credit: Vaping360 , CC BY 2.0


Distribution of Ecigs Sold Pct Tobacco Stores 2016
Based on 57 Observations out of 57
Frequency
0
2
4
6
8
10
12
14
16
0.0
0.2
0.4
0.6
0.8
Ecigs Sold Pct Tobacco Stores 2016


Relationship between Ecigs Sold Pct Tobacco Stores 2016 and Dropout Log
Based on 57 Observations out of 57
Dropout Log
4
2
0
−2
−4
−6
−8
0.0
0.2
0.4
0.6
0.8
Ecigs Sold Pct Tobacco Stores 2016

Limitations

- Irregular data leads to faulty models
- Have to choose: exclude counties, or limit performance
- “Obvious” predictors not readily available (e.g., race)
- Many predictors available as counts, not rates
- Data is old and pre-pandemic

Recommendations

For us

- Control for population better
- Evaluate outliers separately
- Refine, drop, add features
- Fresher data (Covid!)
- Change level of analysis?

For you

- The problem is real, so keep using your expertise!
- Ponder combined effects
- Regulate ecigarette sales
- Evaluate capacity at city / school level

Conclusion

- The goal was to understand and reduce dropout rates.
- Our first outing was not a huge success.
- Try to reduce the availability of ecigarettes.
  - Legal regulation
  - Cultural popularity
- It'll take more work, but we're in if you're in.

Thank you!

We will now take questions.