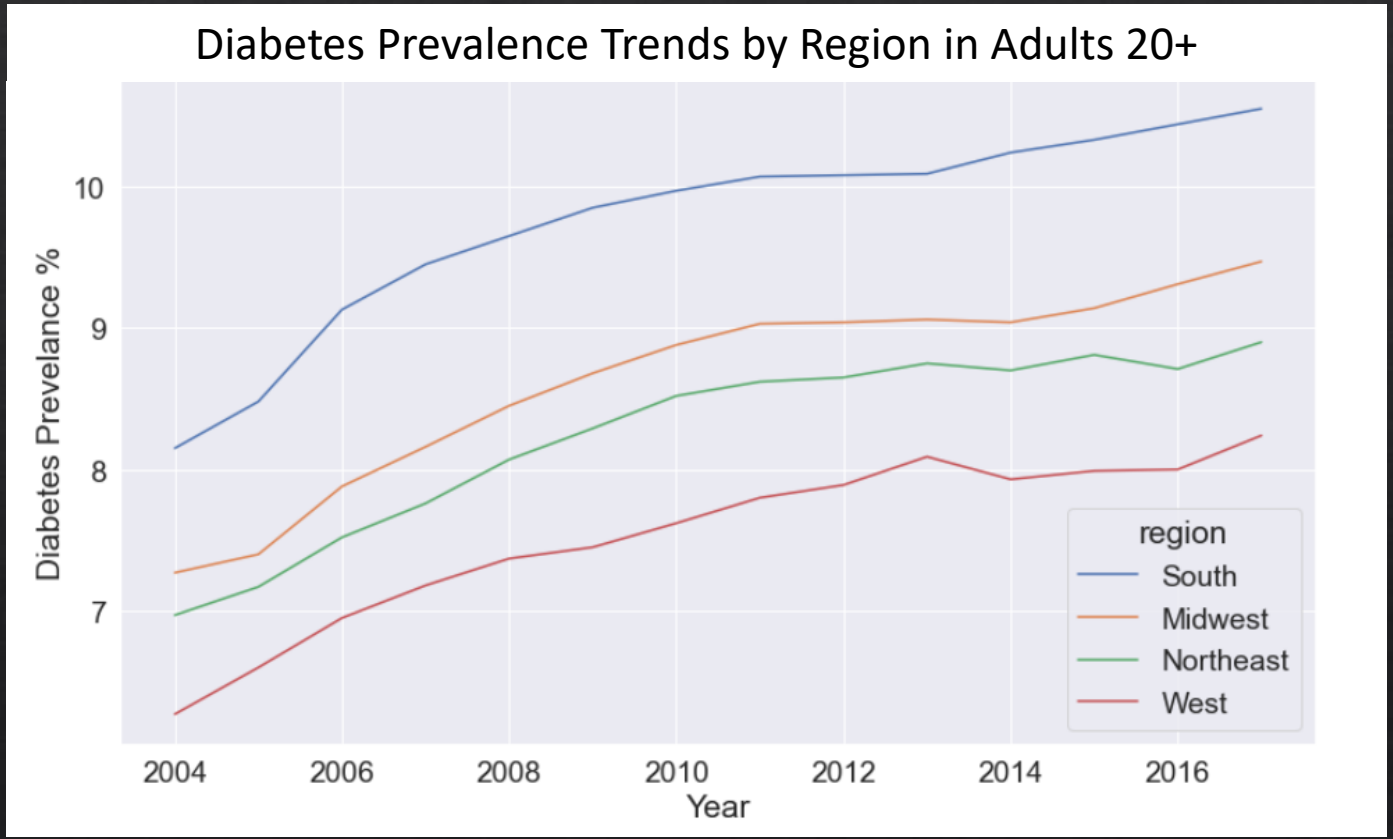
Predicting US Diabetes Prevalence

Can changes in demographics predict changes in diabetes prevalence?

By Aisling Casey
Capstone Presentation
Springboard School of Data
July 2021

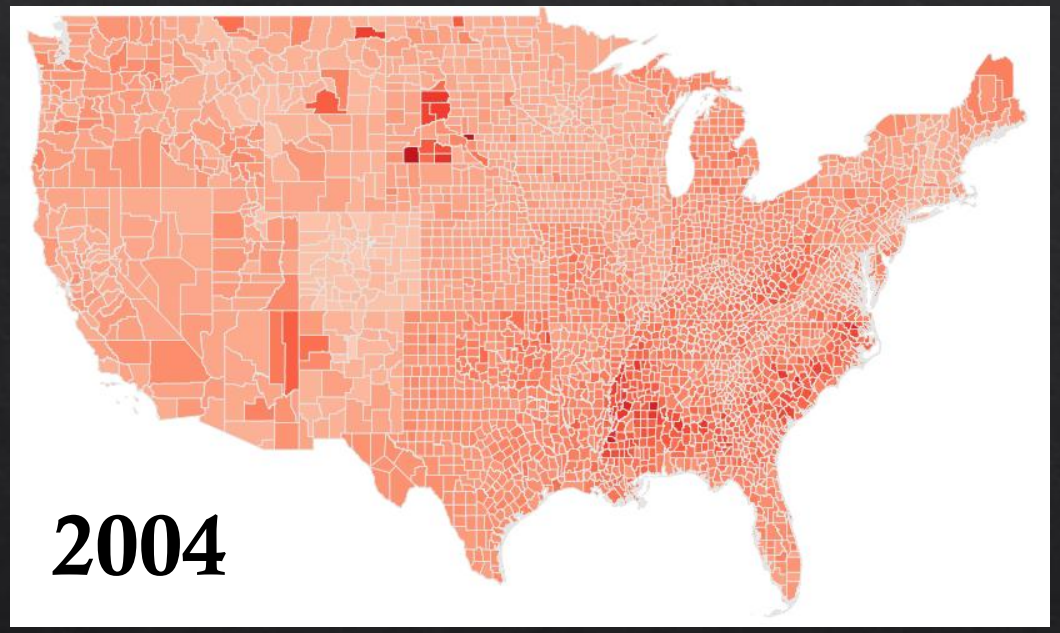
Diabetes in America

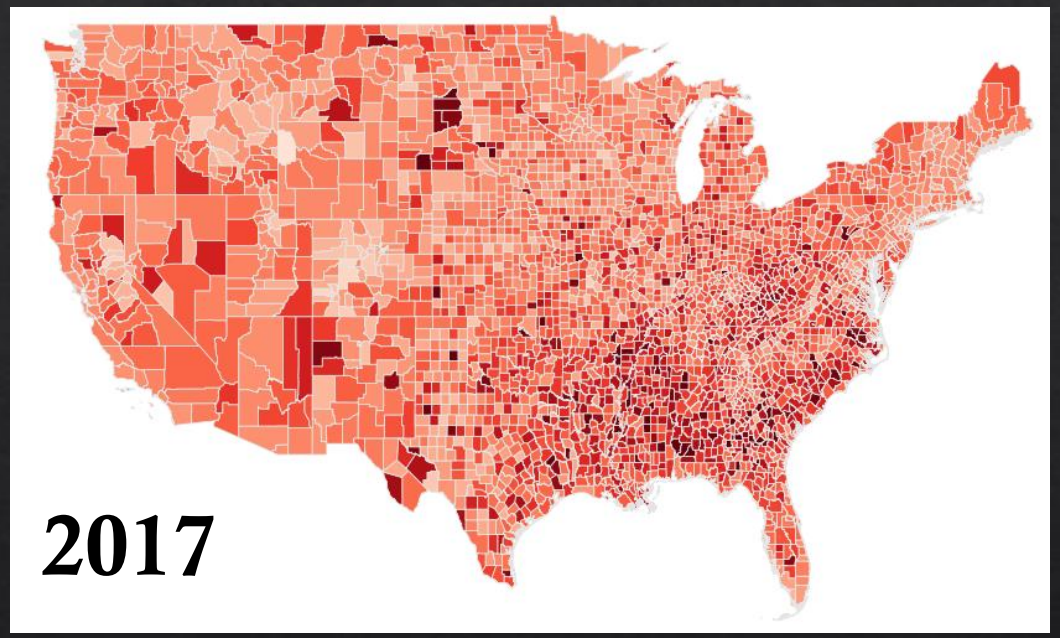
- In 2017, about 1 in 7 healthcare dollars spent on Diabetes and its complications [1]
- From 2000 to 2018, Diagnosed Diabetes rate in American adults has increased more than 50% [2]

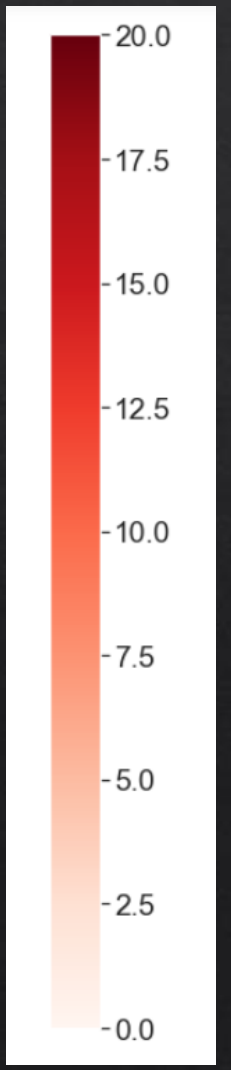


[1]: American Diabetes Association
[2]: Centers for Disease Control

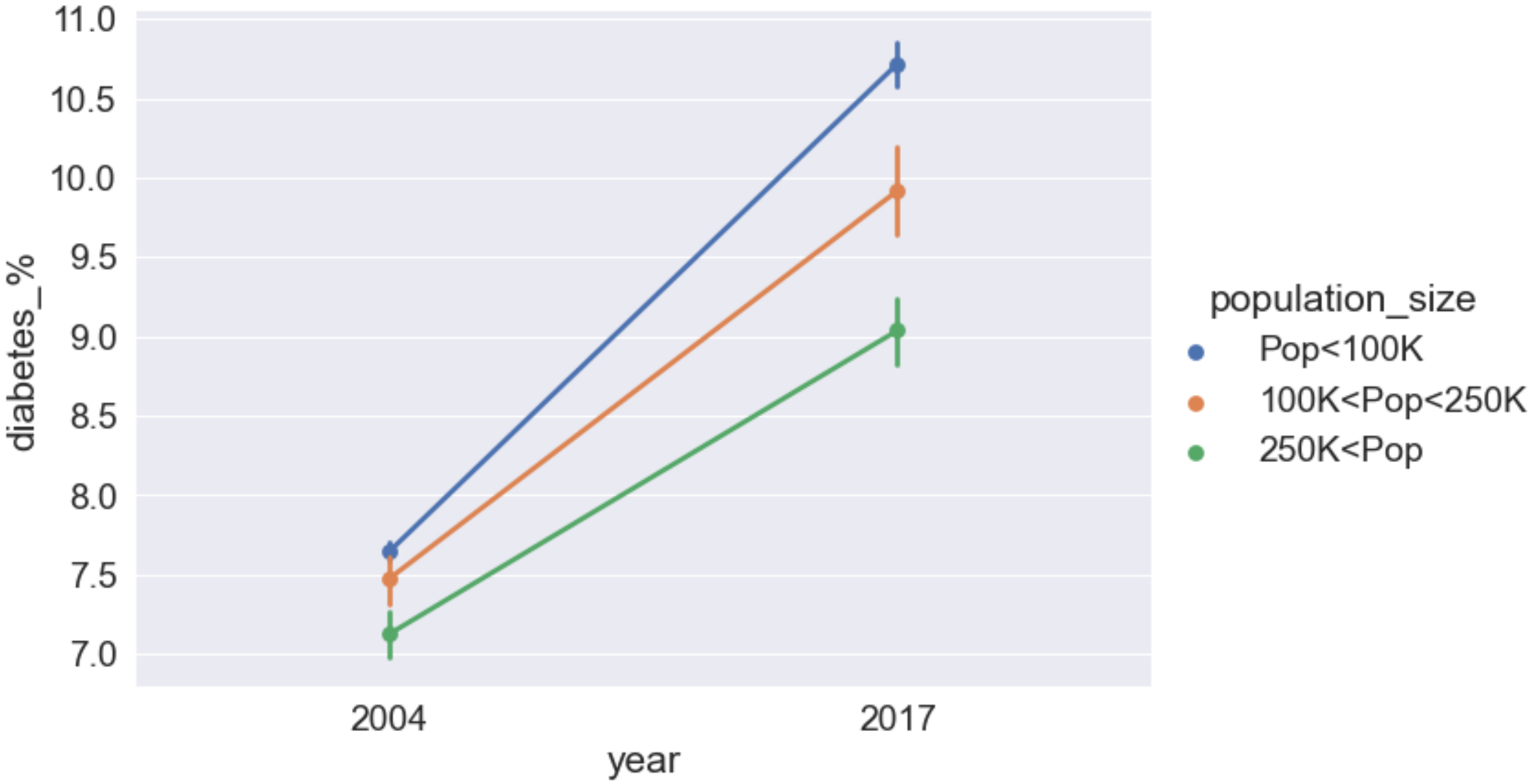
Diabetes Trends – County Level







Change in Diabetes%, Large vs. Small Counties



Problem Identification

Context

County level Diabetes prevalence data not yet available for 2018 & 2019

Solution Space

Use changes in demographic data to predict county level diabetes prevalence

Success Criteria

For each county, a smaller prediction error than using previous year's value as a prediction

Data Sources

-US Census: America Community Survey
-Centers for Disease Control (CDC)

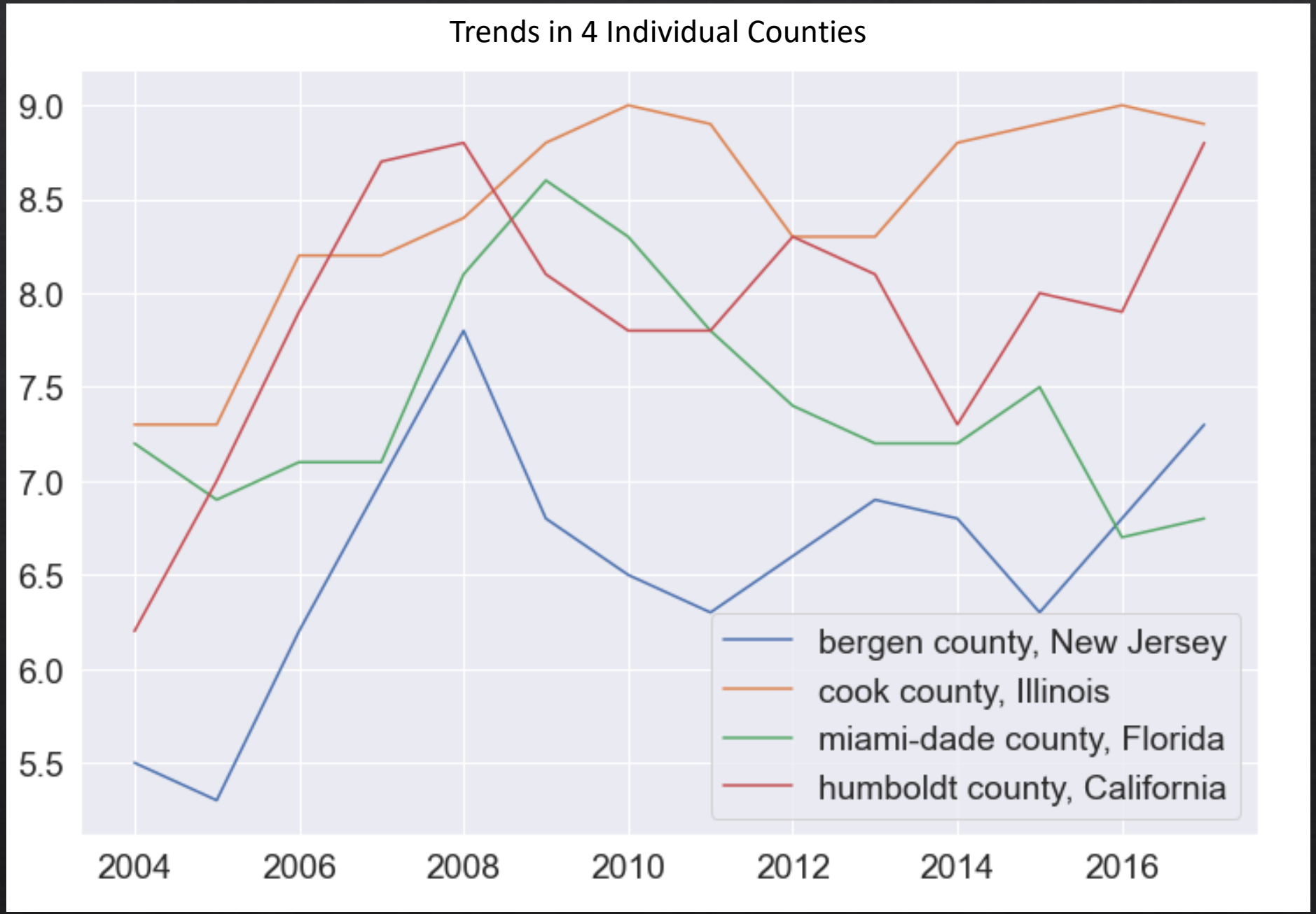
Problem Statement:

Can changes in demographic data be used to predict county-level prevalence of diabetes among adults in the United States in 2018 and 2019?

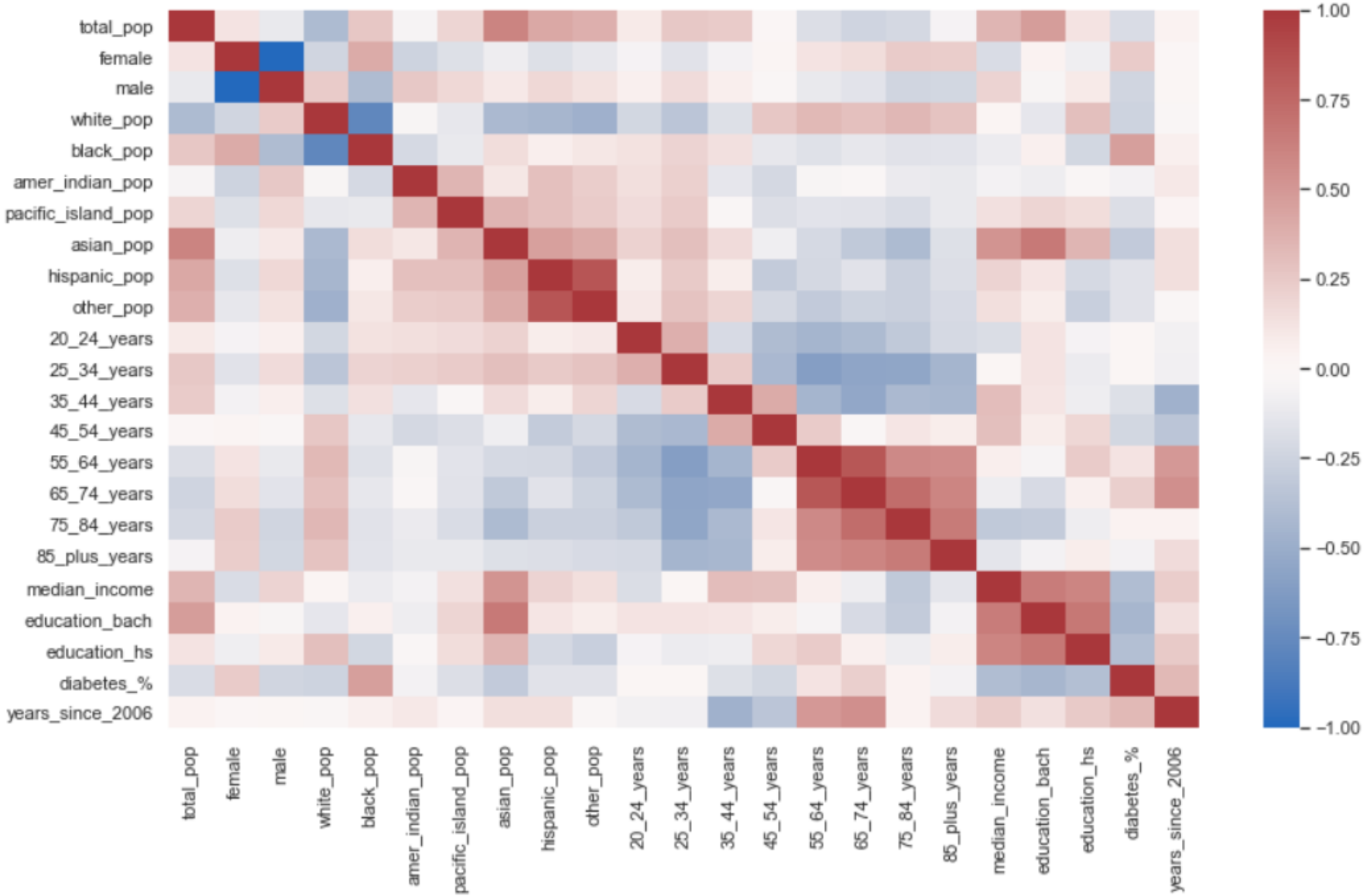
Data Details

Feature Category	Detail	Type	Source	Availability
Diabetes	% Prevalence 20+ years	Target	CDC	-2004-2017 -All Counties
Population	# 20+ years	Explanatory	Census	-2006-2019 -Counties w/ population >65,000
Sex	# Total	Explanatory	Census	
Race	# Total (x6)	Explanatory	Census	
Age	# in 10 year bins	Explanatory	Census	
Education I	% Highschool or >	Explanatory	Census	
Education II	% Bachelor's or >	Explanatory	Census	
Income	$ Median household income	Explanatory	Census	

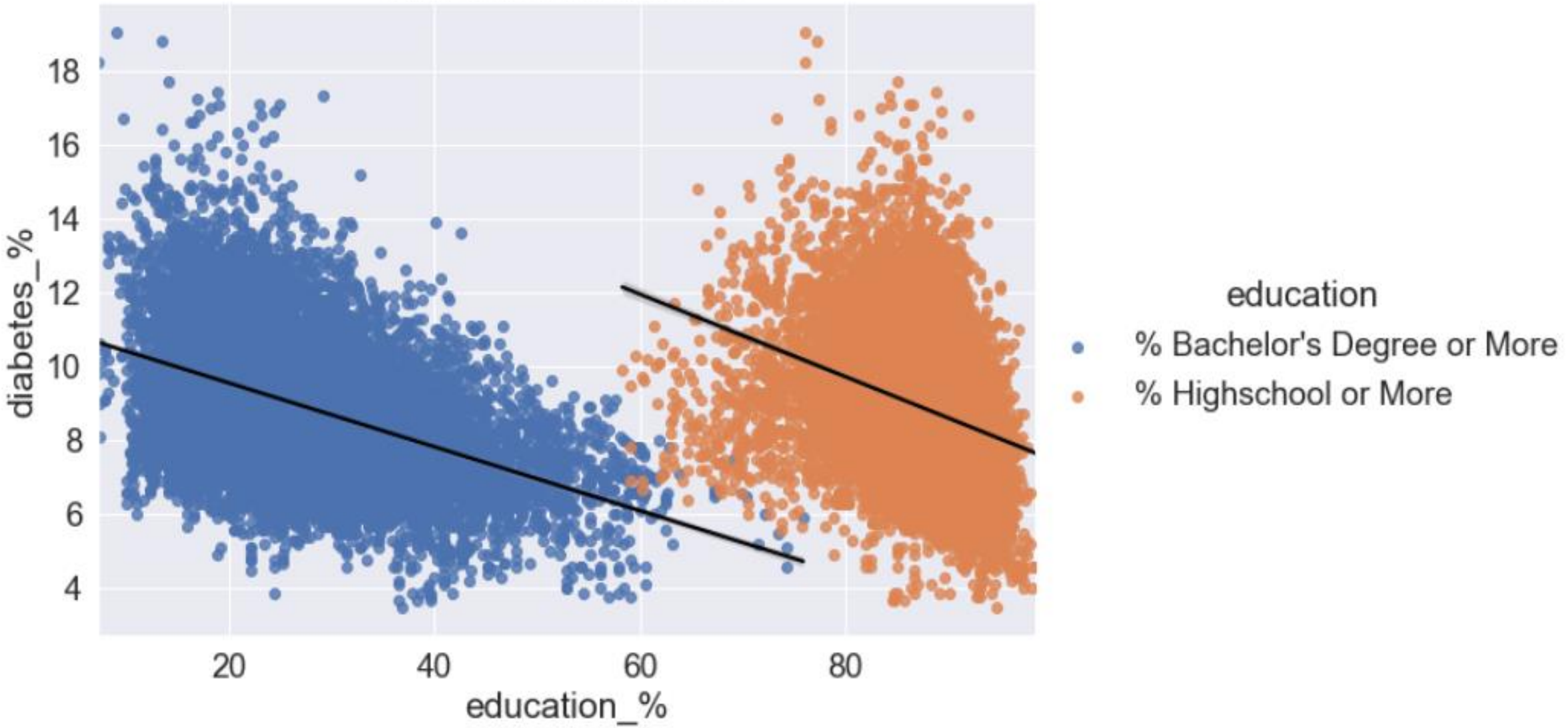
Trends in 4 Individual Counties



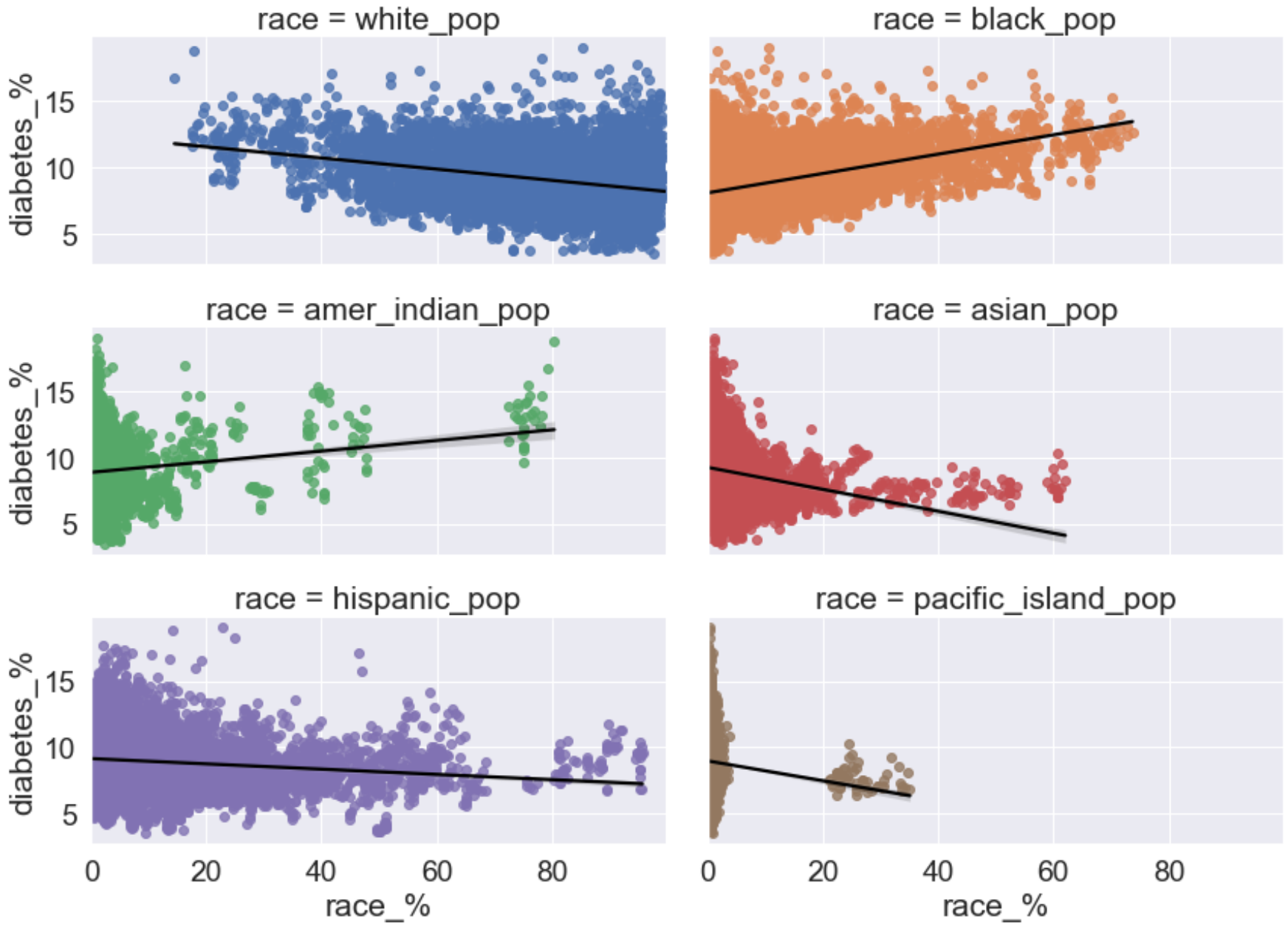
Spearman's Correlation: Relative % Features



Education vs. Diabetes



Race vs. Diabetes



2018 Prevalence by Race [3]	
White	7.5%
Black	11.7%
Amer. Indian	14.7%
Asian	9.2%
Hispanic	12.5%

[3]: American Diabetes Association

Modeling

Considerations

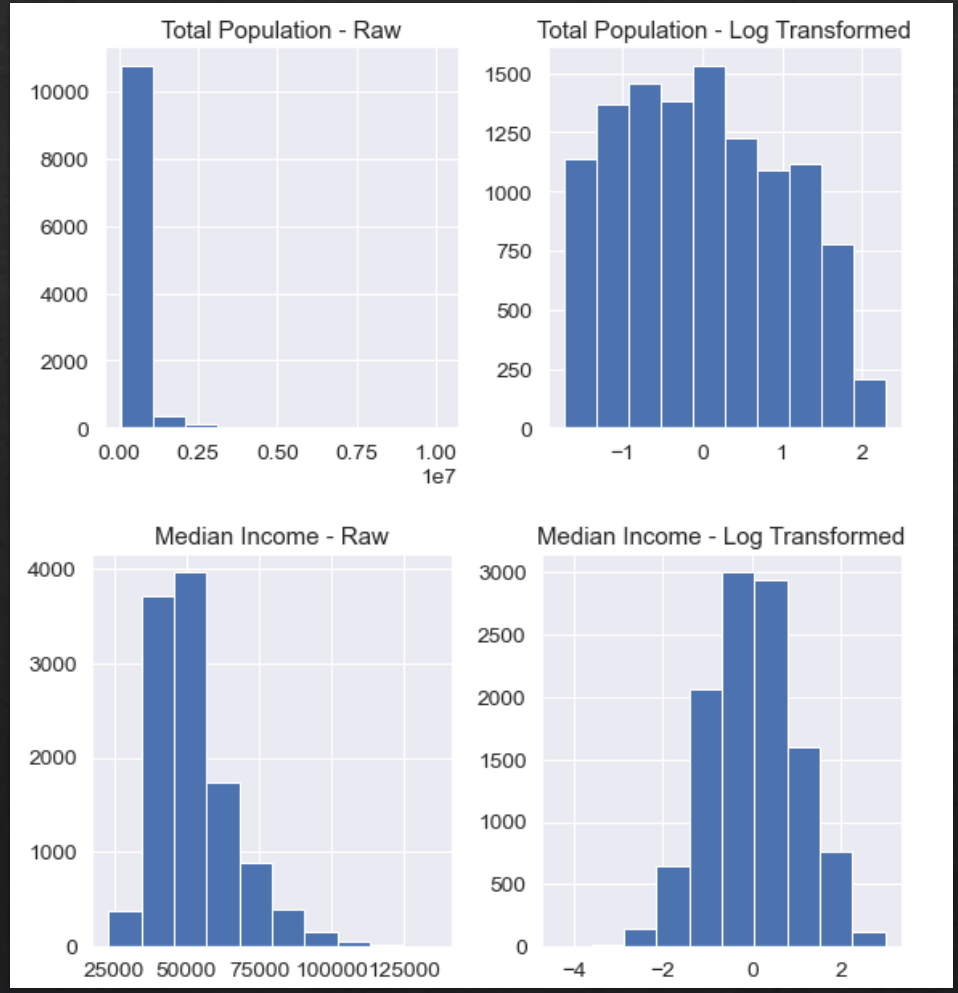
- Capture County-Level Trends
- Out of Sample Predictions
- Understand Features

Choices

- Linear Regression
- Feedforward Neural Net
- Multivariable LSTM Network

General Preprocessing Steps

- Drop counties with population <65K for >3 years
  - Back & forward fill remaining values
- Years-> ‘years_since_2006’
- Dummy-encode region
- Log transform
- Min & max scaling



Accounting for Time

- Linear regression & feedforward net:
  - Added diabetes % from 1, 2 & 3 years ago
- LSTM network:
  - Experimented with different sequence lengths
- Test/Train Split:
  - Testing Data: Up to 2016
  - Training Data: 2017

Years Since 2006	Diabetes %	Diabetes % (t-1)	Diabetes % (t-2)	Diabetes % (t-3)
1	8.1	7.8	7.2	7.2
2	8.6	8.1	7.8	7.2
3	9.9	8.6	8.1	7.8
4	10.2	9.9	8.6	8.1
5	9.8	10.2	9.9	8.6

Sample Data, Baldwin County Alabama

Model Comparison

Predicting 2017's Value
Error: 0.876

Linear Regression
Error: 0.804

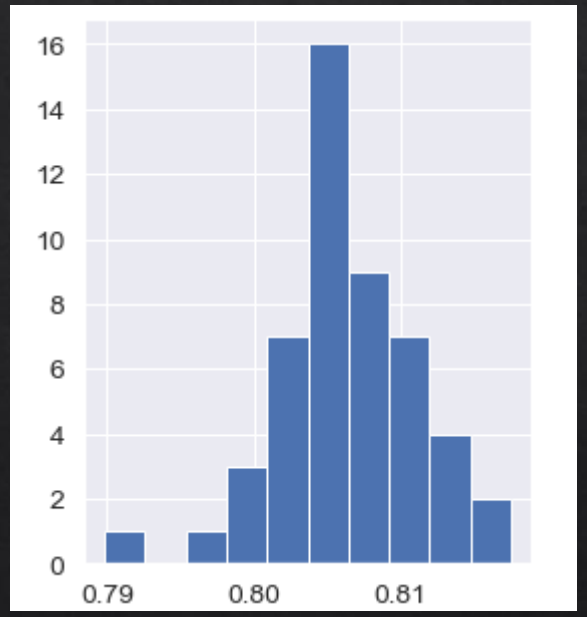
Feedforward Neural Net
Mean Error: 0.806

LSTM Network (n=3)
Mean Error: 0.786

Error = Mean Absolute Error (MAE)

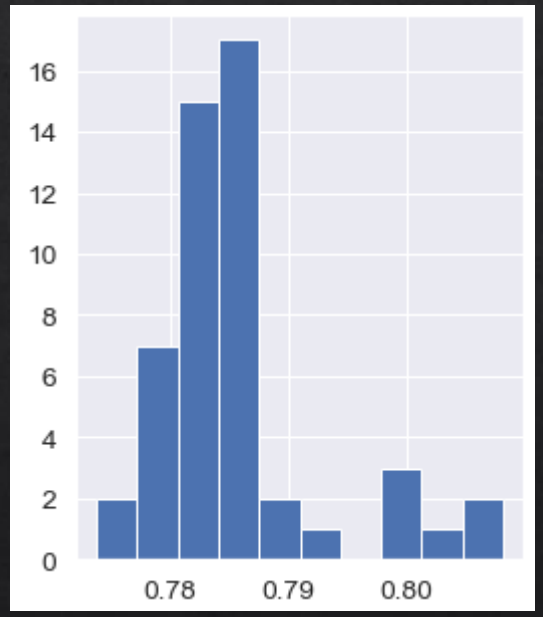
Error ‘Bootstrapping’: 50 Iterations

Feedforward Neural Net



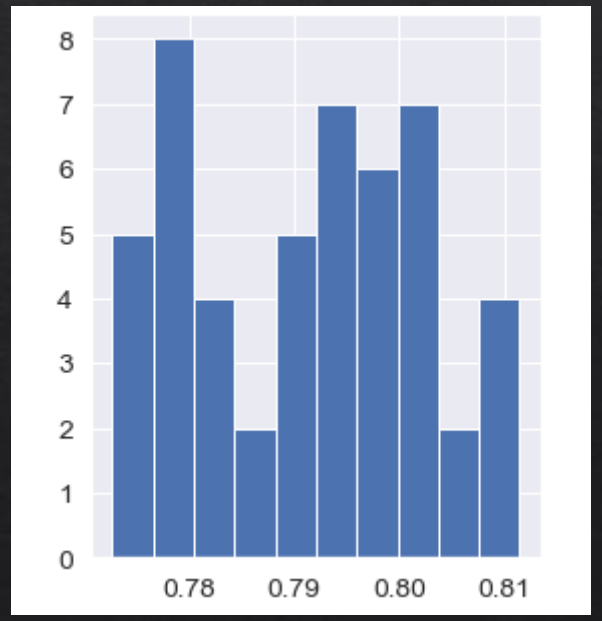
Mean: 0.806
Standard Dev: 0.00486

LSTM Network n=3



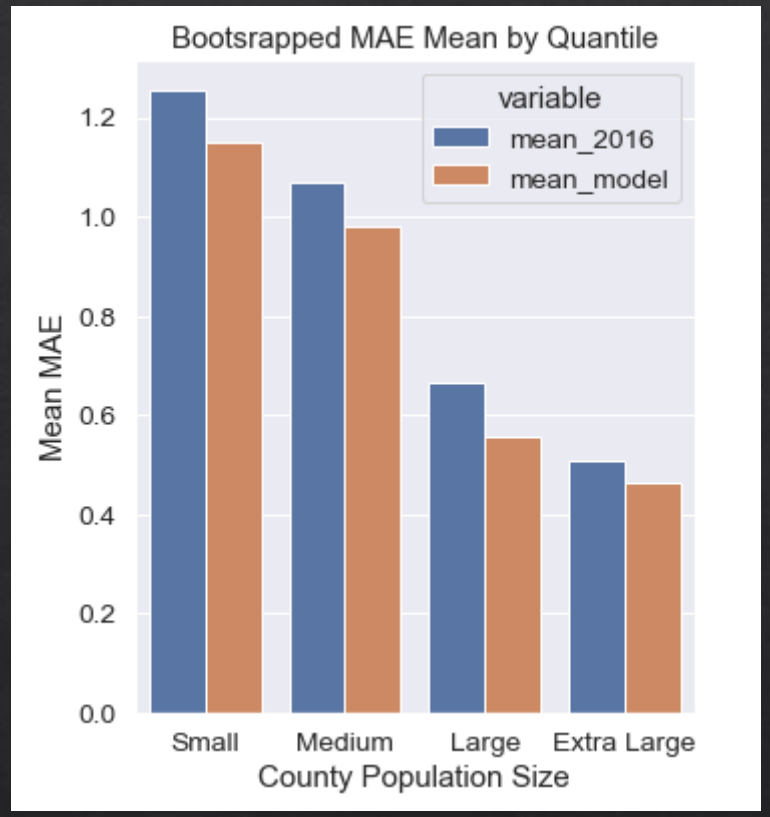
Mean: 0.786
Standard Dev: 0.00709

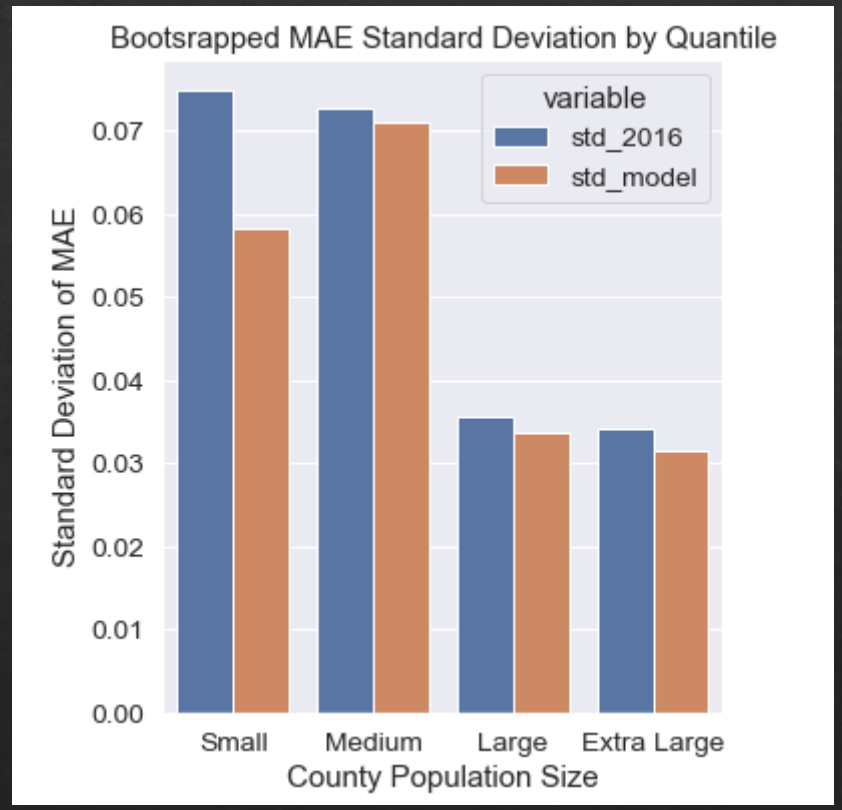
LSTM Network n=4



Mean: 0.791
Standard Dev: 0.0113

Error by Population Quantile for Final Model





Final model vs. 2016 predictions by population quantile

Feature Importance

Linear Regression

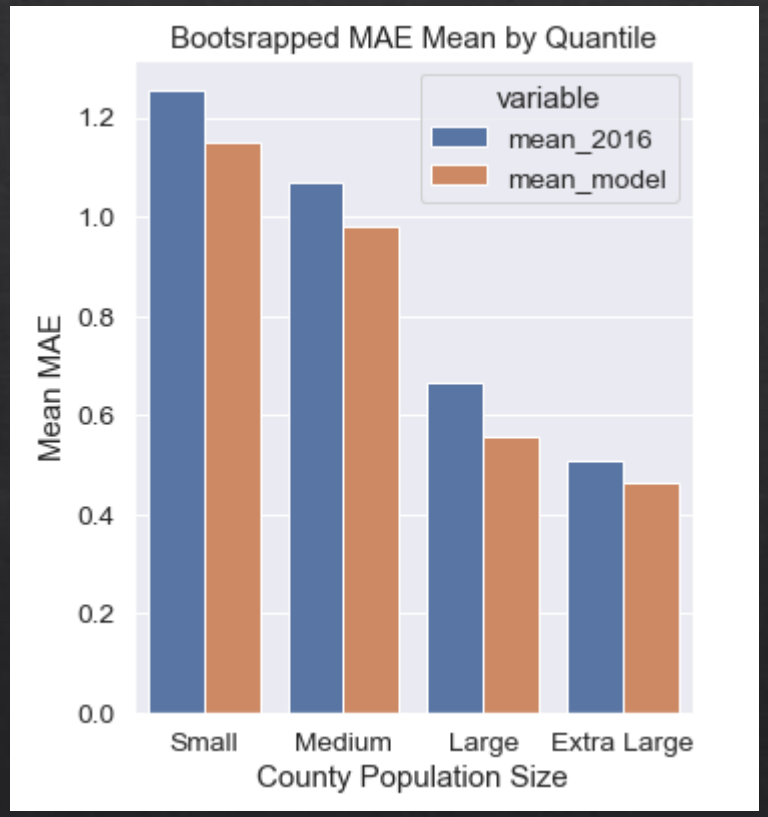
	p	coef	2.5%	97.5%
diabetes_1_year_past	0	14.0418	13.645	14.439
const	0	5.1329	4.029	6.237
diabetes_3_years_past	0	-1.4072	-1.728	-1.087
education_bach	0	-1.2907	-1.502	-1.079
hispanic_pop	0	-0.8132	-1.026	-0.601
median_income	0	-0.6391	-0.873	-0.405
years_since_2006	0	0.478	0.353	0.603
region_West	0	-0.184	-0.244	-0.124
region_South	0	0.0861	0.039	0.133
asian_pop	0.002	1.1804	0.438	1.923
amer_indian_pop	0.004	1.3698	0.446	2.294
85_plus_years	0.005	-0.3147	-0.533	-0.097
black_pop	0.006	1.2269	0.354	2.1
education_hs	0.008	-0.3059	-0.532	-0.079
25_34_years	0.009	-0.3573	-0.625	-0.09
35_44_years	0.016	0.3837	0.07	0.697
45_54_years	0.017	-0.3565	-0.65	-0.063
pacific_island_pop	0.021	-0.4929	-0.911	-0.075
20_24_years	0.028	-0.3611	-0.684	-0.039
other_pop	0.038	0.5574	0.031	1.084
white_pop	0.122	0.7927	-0.213	1.799
55_64_years	0.153	-0.1924	-0.456	0.071
region_Northeast	0.218	0.0309	-0.018	0.08
65_74_years	0.267	-0.346	-0.957	0.265
female	0.547	0.0817	-0.184	0.348
75_84_years	0.584	-0.1126	-0.516	0.291
total_pop	0.639	-0.0198	-0.102	0.063

Conclusions

Is the model useful?

Marginally, yes

On the whole, Race & Socioeconomic status more indicative than age in predicting diabetes



References

- [1] “The Cost of Diabetes.” The Cost of Diabetes | ADA. Accessed June 19, 2021. https://www.diabetes.org/resources/statistics/cost-diabetes.
- [2] “U.S. Diabetes Surveillance System.” Centers for Disease Control and Prevention. Centers for Disease Control and Prevention. Accessed June 19, 2021. https://gis.cdc.gov/grasp/diabetes/DiabetesAtlas.html.
- [3] “Statistics about Diabetes.” Statistics About Diabetes | ADA. Accessed July 27, 2021. https://www.diabetes.org/resources/statistics/statistics-about-diabetes.