Research Questions

RESEARCH QUESTION 1

Does a state having regulations for community water systems have an effect on tooth loss among 18-64 yos for that state? This question will be investigated for any causal effects between the treatment of community water system regulations and dental health outcomes (no tooth loss percentage among 18-64 year olds).

RESEARCH QUESTION 2

Can we predict the percentage of adults 18-64 who have no tooth loss, which we abstract to be 'good dental health' by looking at state average personal income and health indicators like percentage of adults who visit a dentist once a year, percentage of adults who smoke tobacco, percentage of adults with obesity, etc.

Data Overview

RESEARCH QUESTION 1

COMMUNITY WATER REGULATION STATES

This file contains states that have statutes or regulations requiring community water systems to fluoridate drinking water to a specific concentration or range between the years 2011-2013.

POTENTIAL CONFOUNDERS/COVARIATES

- Socioeconomic Status (SES): Areas with higher socioeconomic status might be more likely to implement fluoridation regulations and could also have better oral health due to access to dental care, education, and healthier lifestyles. poverty rate per state (2010), income inequality (gini coefficient) per state (2010), median household income per state, unemployment rate per state
- Access to Dental Care: Disparities in access to dental services could affect oral health outcomes independent of fluoridation regulations. dentists per capita per state (2007), percentage of people who had dental visit in past year per state (2011-2012) [could not find any data pre 2011, but: "The percentage of adults aged 18–64 with a dental visit in the past year did not change significantly from 2009 to 2013"]
- Dietary Habits: Differences in dietary habits, such as sugar consumption, could impact oral health outcomes and might correlate with fluoridation regulations. Prevalence by State of Sugar-Sweetened Beverage Intake Once Daily or More Among US Adults Aged 18 or Older
- Water Quality: Factors like water contamination or source could influence both the decision to implement community water system policies and oral health outcomes. water quality violations per state (no year specified, used april 2020 state population data to calculate water violations per 100k)
- Healthcare Infrastructure: The availability and quality of healthcare services, including dental care, could affect oral health outcomes. public health funding per capita per state (2010), medicaid spending per enrollee per state (2010)
- Population Demographics: Age distribution, ethnic composition, and population density might influence both the implementation of fluoridation regulations and oral health outcomes. percentage of people living in urban areas per state (2010)
- Smoking and Alcohol Consumption: Smoking and alcohol consumption indirectly influence state policies on water fluoridation through broader socio-economic and cultural factors. Public health priorities, shaped by prevalent health concerns, may lead states to prioritize initiatives such as smoking cessation over water fluoridation. Resource allocation decisions are influenced by competing priorities, with limited resources potentially allocated to address multiple public health issues. Public perception and support for health policies, including water fluoridation, can vary based on attitudes towards government intervention and perceptions of risk and benefit. Health disparities in communities with higher rates of smoking and alcohol

consumption may also affect access to preventive services and influence the prioritization of public health initiatives. [percent smoker per state (2009)], [ethanol consumption per state per capita]

- Educational Attainment: Communities with higher levels of education might be more likely to support or implement fluoridation regulations and could also have better oral health knowledge and practices. [education rate per state, HS and bachelors (2010)]

DENTAL HEALTH OUTCOMES

The outcome for the causality research question is the statistic: percent of tooth loss for 18-64 year olds. I sourced this data from the CDC Disease Indicators dataset provided. To get the specific data I got, I filtered for oral health category and the specific rows for percent of tooth loss for 18-64 year olds, specifically selecting for age-adjusted statistics (different age distributions per state). The available data were the years 2012, 2014, 2016, 2018, and 2020, of which I didn't use year 2012 since the treatments were between 2011-2013.

RESEARCH QUESTION 2

CDC DATASET

Quoting from CDC site, dataset provides "cross-cutting set of 124 indicators that were developed by consensus and that allows states and territories to uniformly define, collect, and report chronic disease data that are important to public health practice and available for states, and territories." Data is collected from various state/federal sponsored censuses, but indicators we use are collected mainly from BRFSS. BRFSS collects data as a disproportionate stratified sample and through phone or in-house interviews. Thus, a possible bias is no sampling for people without an address or phone number.

BEA INCOME DATASET

Bureau of Economic Analysis collects yearly personal income data from the Census Bureau's annual mid year population estimates. This data is collected by census and thus has bias due to sampling from people with a defined address. According to the BEA, "Per capita personal income is calculated as the personal income of the residents of a given area divided by the resident population of that area." This is odd since it means that the average is not calculated by dividing by the population that is employed.

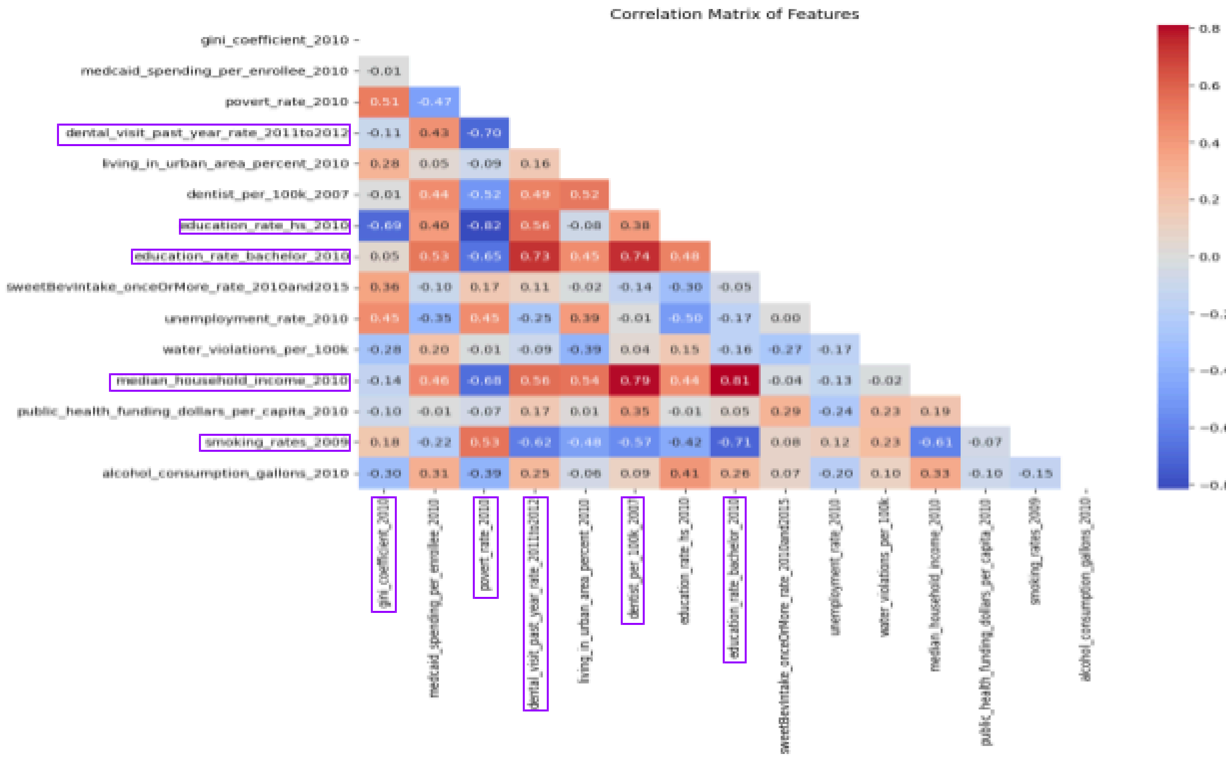
STATE WATER FLUORIDATION DATASET

Data is collected from states and tribes, and thus has variable context and methodologies. 40 states provide their data to the public. The CDC uses this state data to then calculate the percentage of state population that has 'fluoridated' water, which encompasses small and substantial fluoridation.

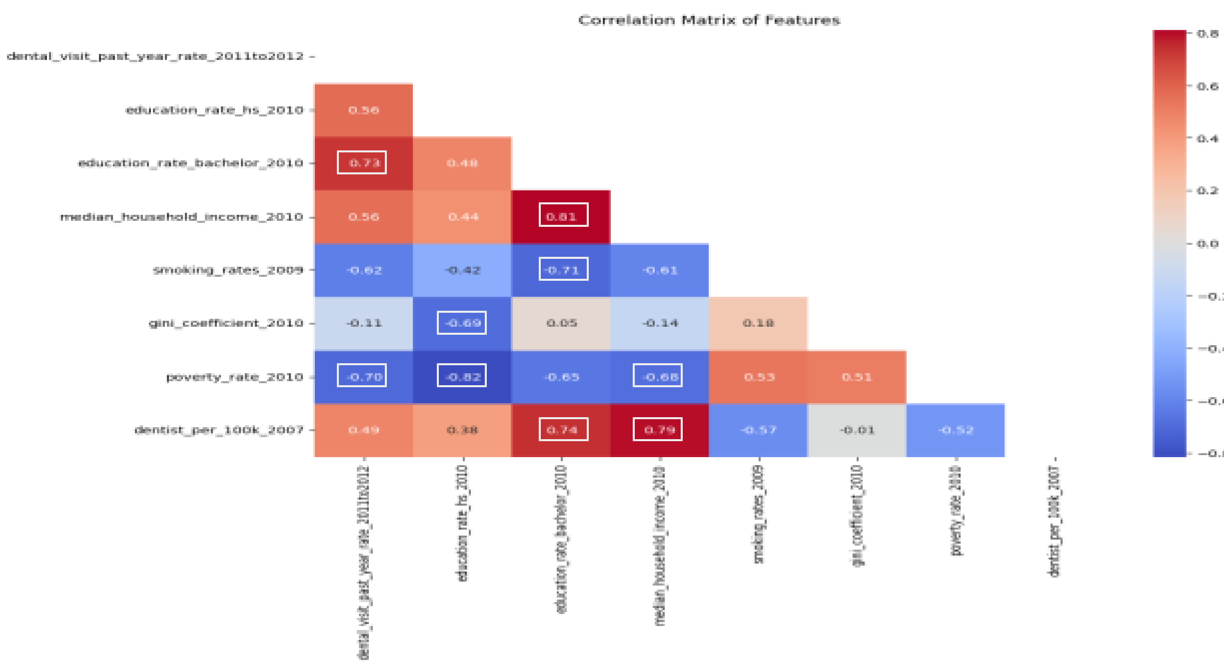
EDA

RESEARCH QUESTION 1

1. correlation matrix/heatmap to identify multicollinearity between confounding variables
2. The relevant regulation of the 13 states seems to be between 2011-2013, so I will use pre-2011 data for confounders so that the confounding variable data precedes and cannot be affected by the treatment itself.



I identified the covariates with the highest correlations, then mapped those covariates below.



I then boxed the correlations that were around absolute value of 0.70 or higher.

covariate1	covariate2	correlation
~~poverty_rate_2010~~	~~education_rate_hs_2010~~	~~-0.82~~
~~education_rate_bachelor_2010~~	~~median_household_income_2010~~	~~0.81~~
dentist_per_100k_2007	median_household_income_2010	0.79
~~education_rate_bachelor_2010~~	~~dentist_per_100k_2007~~	~~0.74~~
~~education_rate_bachelor_2010~~	~~dentist_visit_past_year_rate_2011to2012~~	~~0.73~~
~~education_rate_bachelor_2010~~	~~smoking_rates_2009~~	~~-0.71~~
~~poverty_rate_2010~~	~~dentist_visit_past_year_rate_2011to2012~~	~~-0.70~~
~~gini_coefficient_2010~~	~~education_rate_hs_2010~~	~~-0.69~~
~~poverty_rate_2010~~	~~median_household_income_2010~~	~~-0.68~~

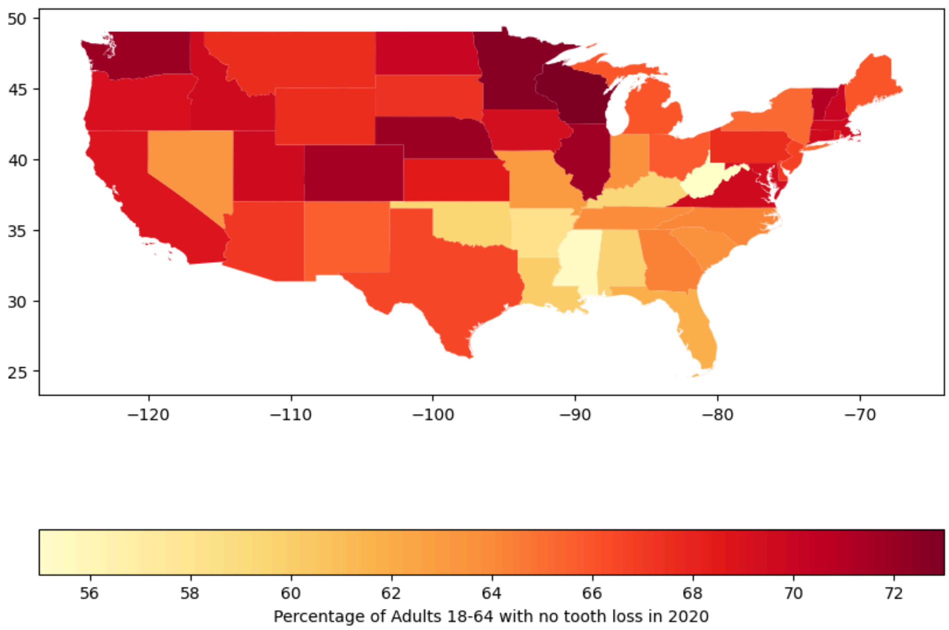
Ranking of the most frequent and highest average absolute correlation covariates:

Covariate	Times Shown	Chi-square Value	Avg Absolute Correlation
~~education_rate_bachelor_2010~~	~~4~~	~~44.80~~	~~0.7475~~
median_household_income_2010	3	49.99	0.7600
~~poverty_rate_2010~~	~~3~~	~~41.33~~	~~0.7333~~
dentist_per_100k_2007	2	49.99	0.7650
education_rate_hs_2010	2	39.60	0.7550
dentist_visit_past_year_rate_2011to2012	2	49.99	0.7150
smoking_rates_2009	1	34.40	0.7100
~~gini_coefficient_2010~~	~~1~~	~~7.974~~	~~0.6900~~

Since gini_coefficient_2010 has the smallest chi-square value by far, I will drop this feature. I will also drop poverty_rate_2010 since it has a smaller chi-squared value than median_household_income_2010, and they are both in the socioeconomic category. I will drop education_rate_bachelor_2010 over education_rate_hs_2010 even though it has a higher chi-square value because after gini_coefficient_2010 and poverty_rate_2010 are dropped, there are no more high correlation pairs for education_rate_hs_2010; dropping education_rate_bachelor_2010 will get rid of 4 high correlation pairs. Now we are left with one high correlation pair between median_household_income_2010 and dentist_per_100k_2007. I plan to keep both of these features since they are both strongly associated with the target variable, and are in separate context categories. To reduce multicollinearity between median_household_income_2010 and dentist_per_100k_2007, I plan to replace the two original features with a new feature, the ratio dentist_per_100k_2007/median_household_income_2010, prioritizing dentist_per_100k_2007 as the dominant feature.

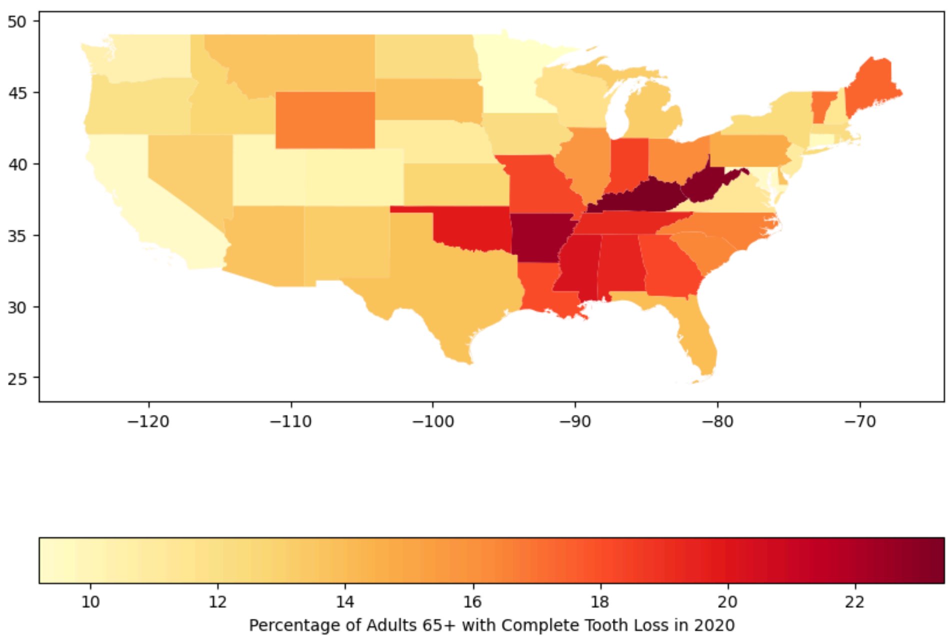
RESEARCH QUESTION 2

For this EDA, we will only be looking at health data collected from the CDC's dataset, further data from the BEA and other sources will be used in the GLM/nonparametric predictions. We will try to predict two statewide variables: percentage of adults 18-64 with no tooth loss and percentage of adults 65+ with complete tooth loss. These variables are indicators for 'oral health'. Our features are statewide percentage of adults who smoke, percentage of adults with obesity, percentage of adults who have visited a dentist in the past year, and percentage of adults who have short sleep durations.



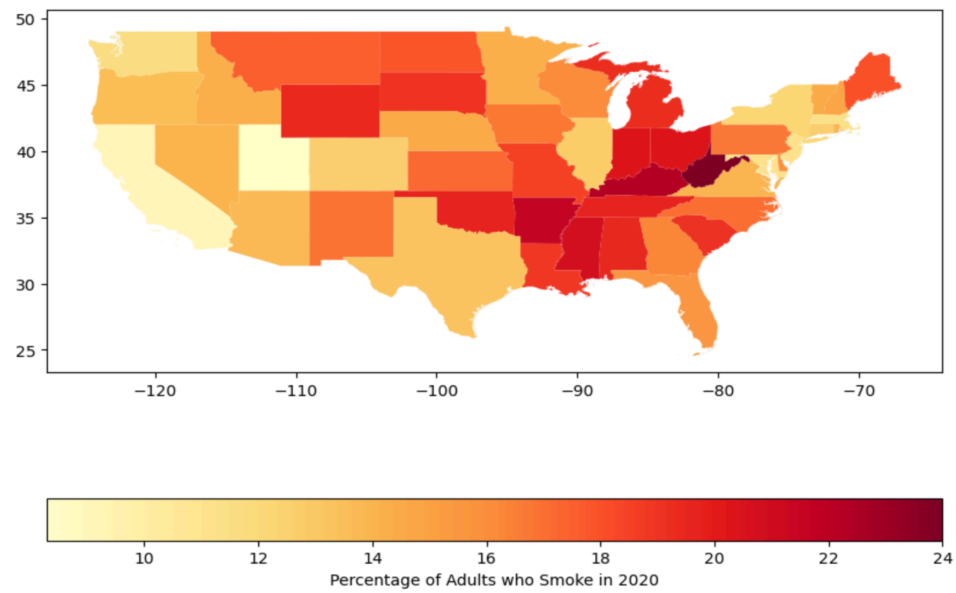
Age Adjusted Prevalence of Adults 18-64 with No Tooth Loss

From the graph, it seems the deep south has the lowest oral health overall, and the northeast has significant variations of oral health per state. Nevada has bad oral health compared to other southwest states.



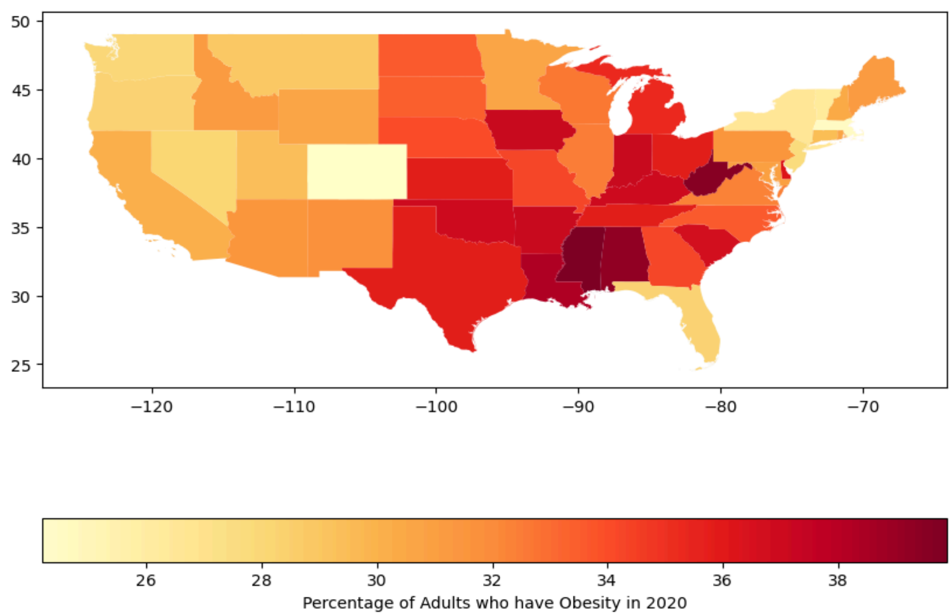
Age Adjusted Prevalence of Adults 65+ with Complete Tooth Loss

This graph is almost the inverse of the first graph. Higher values here mean worse overall oral health. The deep south has higher percentages of bad oral health. States with good oral health in the first graph are the same states in this graph that have low complete tooth loss.



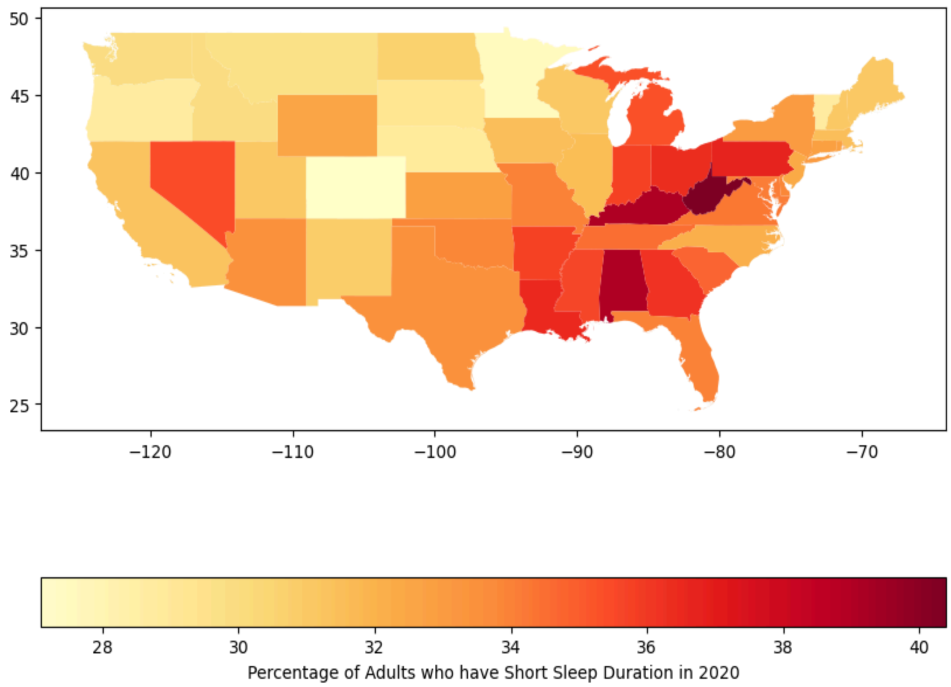
Age Adjusted Prevalence of Adults who Smoke

This graph shows the upper midwest and the deep south have higher rates of smokers. Smoking is a significant predictor of oral health because it is tied to higher risks of oral cancer and gum disease.



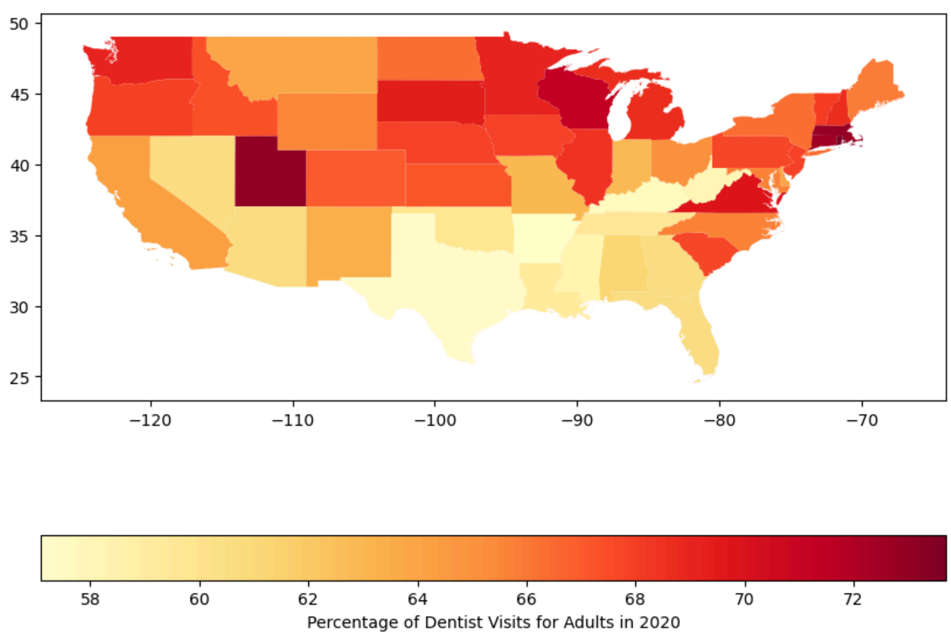
Age Adjusted Prevalence of Obesity Among Adults

This graph shows the deep south and the eastern midwest have the highest prevalence of obesity. Obesity is linked to diabetes, cardiovascular disease and other diseases that can affect overall health, and thus oral health of an individual.



Age Adjusted Prevalence of Adults with Short Sleep Duration

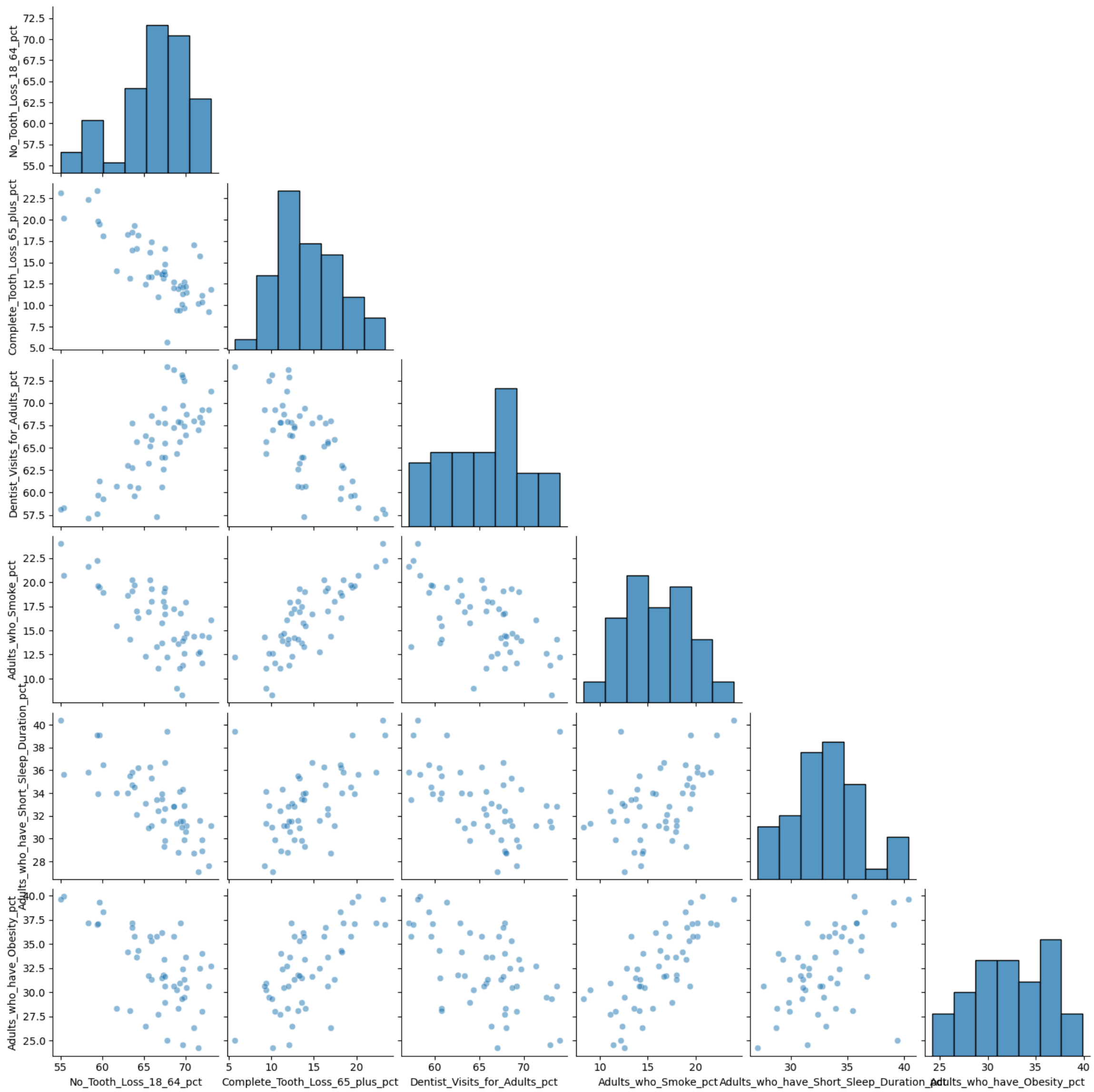
This map shows that the southern parts of the US have higher prevalence of short sleep durations. Specifically, the South and the lower Northwest have significantly higher prevalence. Sleep is linked to overall health, having short sleep durations thus may lead to overall lower health including oral health.



Age Adjusted Prevalence of Adults who have Visited a Dentist in the Past Year

This graph shows the southern states have a population that is less likely to have visited a dentist in the past year. This is crucial to oral health because yearly dental visits are good for treating cavities, gum disease and other oral diseases.

Comparison of Correlation Between Data Features for the Year 2020



The first two columns are most relevant to us since these compare our features to the values we want to predict. In summary, there seems to be positive correlation for yearly dentist visit rates and good ‘oral health’. Higher prevalence of smokers is negatively proportional to good ‘oral health’. Short sleep duration prevalences are also negatively proportional to good ‘oral health’. There seems to be less significant but still negative correlation between obesity and good ‘oral health’.

The other three columns describe correlation between our X values used in prediction. There seems to be correlation between all these variables. Notably, higher propensity of yearly dentist visits correlates with lower high risk behavior (Smoking, Short Sleep Duration, and Obesity). Smoking, Short Sleep Duration and Obesity are all positively correlated with each other.

Question Report

RESEARCH QUESTION 1

Question

Does a state having regulations for community water systems have an effect on tooth loss among 18-64 yos for that state? This question will be investigated for any causal effects between the treatment of community water system regulations and dental health outcomes (no tooth loss percentage among 18-64 year olds).

Algorithm/Method

Using an ensemble of logistic regression models in order to model and assign propensity scores to each unit of data, where the output of the propensity score model is the treatment and the features being trained on are the covariates selected during the EDA process. Then, the propensity scores are used to calculate average treatment effect.

Assumptions

1. The covariates selected were relevant and enough to correctly model the propensity score model.
2. The number of models used in the ensemble were enough to make the model useful.
3. 0.7 is a good accuracy threshold that doesn't cause the ensemble to overfit the data but also captures enough relevant information.

Implementation: data preparation, accuracy thresholding, and model training

After conducting feature selection, I scaled the features to each have a mean of 0 and a standard deviation of 1 to ensure that they have similar influences on the model. For the train/test sets, I did a 0.8/0.2 split or 40/10 rows. To introduce cross-validation, I went with 20 runs, where each run pseudorandomly determines the train/test split based on the seed being set to i, where i is the ith run. Within each run, I trained 5000 models (a total of 100,000 models tested), where models and their outputs were added to the ensemble model based on the test set's accuracy. I decided to use bagging (bootstrapped aggregation) since I only have 40 rows in the training set, so for each of the 1000 models per run, I bootstrapped the training set from the determined 40 rows based on the run. One issue with bootstrapping the treatment values is that rarely, a model would have only 1's or only 0's, so I wrote code to catch that error. I then tested the model performance based on the 10 rows of the training set (which is unchanged). I decided to set an accuracy threshold of 0.7 (meaning at least 7 of the 10 rows must have their treatments predicted based on the covariates). 0.7 is a moderate level of accuracy, compared to 0.5 being entirely random and 0.8/0.9 overfitting. I also calculated the metric of the overall rate at which models were being rejected, and a threshold of 0.7 rejected around half of the models (a threshold of 0.5 rejected ~10% and a threshold of 0.9 rejected 90%+). To summarize, I had 20 runs with 5000 models per run, using an accuracy threshold of 0.7 on the test set to decide whether to add or reject the model from the ensemble; all of the good models were compiled to calculate the propensity scores per state which were then subsequently used to calculate treatment effects.

Statement/Interpretation of Results: ATE calculations and significance

I calculated the treatment effect using dental health outcomes for no tooth loss for 18-64 year olds (age-adjusted) for even years between 2014-2020, since the treatments are between 2011-2013 and I want outcomes to be after the treatment chronologically. Using inverse propensity weighting, I got an average treatment effect of approximately -15 for each year. However, using augmented inverse propensity weighting led me to get non-significant treatment effects, where the p value is above 0.9 for 2014, 2018, and 2020, though 2016 had an average treatment effect of 1.07 with a p-value of 0.10. This means that overall, there is no significant treatment effect of the CWS treatment.

RESEARCH QUESTION 2

Question:

Can we predict the statewide percentage of adults age 18-64 who have no tooth loss using statewide data such as: percentage of people who visited a dentist in the past year (**Dentist Visits**), percentage of people who report having sufficient sleep duration (**Good Sleep**), percentage of adults who smoke tobacco (**Smoke**), percentage of adults who have obesity (**Obesity**), average personal income for each state (**Income**), and percentage of population whose public water is fluoridated (**Fluoridation**).

Extra details:

1. Each percentage is age adjusted to only consider people age 18-64
2. Sufficient sleep duration is defined by the CDC as >7 hours per night

Design Choices:

All data is not modified in any way (no logging, or squaring of columns). This is because when comparing correlation between features, all correlations seemed to be linear (See EDA graph above)

Model Choices:

Dentist Visits are crucial to maintaining good dental health. Thus it is arguable that a higher percentage of people who visit the dentist at least yearly can be good for predicting how orally healthy the population is. Good sleep is crucial to maintaining good overall health. Like dentist visits, it can be a good indicator for dental health. Obesity is linked to overall bad health habits, thus might be a good indicator of bad dental health. Smoking is linked to oral cancer and other oral illnesses like gum disease. Higher average income might mean a population is more likely to afford better healthcare, thus might be a good indicator. Lastly, water fluoridation is a good source of fluoride, which helps fortify enamel.

Assumptions:

State data is independent of year and state when conditioned on the features we list. This also means we assume that good dental health is independent and identically distributed for all states given our features.

RESEARCH QUESTION 2A NONPARAMETRIC IMPLEMENTATION

An analysis was conducted to predict statewide tooth health using statewide demographic data through nonparametric methods. During this process, data sourced from the CDC and BEA were utilized to build two models: a decision tree and a random forest. These models employed variables such as the percentage of people who visited a dentist in the past year and the percentage of adults who smoke to predict the rate of tooth loss. The performance of the models was evaluated using training and testing data, and their generalization abilities were verified through cross-validation. The decision tree model exhibited very high performance on the training data but showed a significant drop in performance on the test data, indicating an issue with overfitting. In contrast, the random forest model maintained relatively high performance on the test data, and the results from cross-validation were consistently positive. Based on these outcomes, the random forest model is deemed more suitable for predicting statewide tooth health based on the complex structure of statewide demographic data, with a lower risk of overfitting. Future research could focus on tuning the hyperparameters of the model or utilizing additional data to further improve performance.

RESEARCH QUESTION 2A NONPARAMETRIC RESULTS

Decision Tree Training RMSE: 0.18708286933869708 Training R2: 0.9986365995187398
Decision Tree Test RMSE: 3.9766820340580415 Test R2: 0.27333596048682707
Decision Tree Cross-validation RMSE: 3.8532856110078306
Random Forest Training RMSE: 1.0919727521274931 Training R2: 0.9535507183595311
Random Forest Test RMSE: 3.0840837131808514 Test R2: 0.5629367910771783
Random Forest Cross-validation RMSE: 3.15988942400152

RESEARCH QUESTION 2B PARAMETRIC (OLS) IMPLEMENTATION

The Ordinary Least Squares Regression (OLS) model was utilized to describe the relationship between independent quantitative variables, including "Complete_Tooth_Loss_65_plus_pct," "Dentist_Visits_for_Adults_pct," "Adults_who_Smoke_pct," "Adults_who_have_Short_Sleep_Duration_pct," and "Adults_who_have_Obesity_pct," and a dependent variable which in this case is "No_Tooth_Loss_18_64_pct." The application of the Ordinary Least Squares Regression (OLS) model also included analyses of single parameters, with "No_Tooth_Loss_18_64_pct" as the dependent variable, and analyses of two parameters, with "No_Tooth_Loss_18_64_pct" as the dependent variable. The analyses of the variables are as follows.

RESEARCH QUESTION 2B PARAMETRIC (OLS) RESULTS

Single Parameter (With "No_Tooth_Loss_18_64_pct"):

1) "No_Tooth_Loss_18_64_pct" vs. "Complete_Tooth_Loss_65_plus_pct"

```
==============================================================================
Dep. Variable:          No_Tooth_Loss_18_64_pct   R-squared (uncentered):          0.903
Model:                  OLS   Adj. R-squared (uncentered):          0.901
Method:                 Least Squares   F-statistic:                454.8
Date:                   Mon, 06 May 2024   Prob (F-statistic):            1.91e-26
Time:                   13:02:03   Log-Likelihood:              -222.60
No. Observations:                 50   AIC:                                447.2
Df Residuals:           49   BIC:                                449.1
Df Model:               1
Covariance Type:        nonrobust
======================================================================================
                          coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------
Complete_Tooth_Loss_65_plus_pct   4.2650   0.200   21.326   0.000   3.863   4.667
==============================================================================
Omnibus:                3.793   Durbin-Watson:              1.638
Prob(Omnibus):          0.150   Jarque-Bera (JB):           3.541
Skew:                   -0.643   Prob(JB):                  0.170
Kurtosis:               2.785   Cond. No.                   1.00
==============================================================================
```

2) "No_Tooth_Loss_18_64_pct" vs. "Dentist_Visits_for_Adults_pct"

```
==============================================================================
Dep. Variable:          No_Tooth_Loss_18_64_pct   R-squared (uncentered):          0.998
Model:                  OLS   Adj. R-squared (uncentered):          0.998
Method:                 Least Squares   F-statistic:                2.272e+04
Date:                   Mon, 06 May 2024   Prob (F-statistic):            5.12e-67
Time:                   13:02:03   Log-Likelihood:              -127.32
No. Observations:                 50   AIC:                                256.6
Df Residuals:           49   BIC:                                258.6
Df Model:               1
Covariance Type:        nonrobust
======================================================================================
                          coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------
Dentist_Visits_for_Adults_pct   1.0136   0.007   150.738   0.000   1.000   1.027
==============================================================================
Omnibus:                0.343   Durbin-Watson:              2.051
Prob(Omnibus):          0.842   Jarque-Bera (JB):           0.096
Skew:                   -0.106   Prob(JB):                  0.953
Kurtosis:               3.034   Cond. No.                   1.00
==============================================================================
```

RESEARCH QUESTION 2B PARAMETRIC (OLS) RESULTS

3) "No_Tooth_Loss_18_64_pct" vs. "Adults_who_Smoke_pct"

Dep. Variable:	No_Tooth_Loss_18_64_pct	R-squared (uncentered):	0.933
Model:	OLS	Adj. R-squared (uncentered):	0.931
Method:	Least Squares	F-statistic:	679.2
Date:	Mon, 06 May 2024	Prob (F-statistic):	2.26e-30
Time:	13:02:03	Log-Likelihood:	-213.39
No. Observations:	50	AIC:	428.8
Df Residuals:	49	BIC:	430.7
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Adults_who_Smoke_pct	3.9192	0.150	26.061	0.000	3.617	4.221

Omnibus:	1.201	Durbin-Watson:	1.374
Prob(Omnibus):	0.549	Jarque-Bera (JB):	1.226
Skew:	-0.295	Prob(JB):	0.542
Kurtosis:	2.509	Cond. No.	1.00

4) "No_Tooth_Loss_18_64_pct" vs. "Adults_who_have_Short_Sleep_Duration_pct"

Dep. Variable:	No_Tooth_Loss_18_64_pct	R-squared (uncentered):	0.979
Model:	OLS	Adj. R-squared (uncentered):	0.979
Method:	Least Squares	F-statistic:	2283.
Date:	Mon, 06 May 2024	Prob (F-statistic):	9.06e-43
Time:	13:02:03	Log-Likelihood:	-184.29
No. Observations:	50	AIC:	370.6
Df Residuals:	49	BIC:	372.5
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Adults_who_have_Short_Sleep_Duration_pct	1.9827	0.041	47.786	0.000	1.899	2.066

Omnibus:	2.152	Durbin-Watson:	1.838
Prob(Omnibus):	0.341	Jarque-Bera (JB):	1.932
Skew:	-0.471	Prob(JB):	0.381
Kurtosis:	2.803	Cond. No.	1.00

5) "No_Tooth_Loss_18_64_pct" vs. "Adults_who_have_Obesity_pct"

Dep. Variable:	No_Tooth_Loss_18_64_pct	R-squared (uncentered):	0.971
Model:	OLS	Adj. R-squared (uncentered):	0.971
Method:	Least Squares	F-statistic:	1653.
Date:	Mon, 06 May 2024	Prob (F-statistic):	2.07e-39
Time:	13:02:03	Log-Likelihood:	-192.17
No. Observations:	50	AIC:	386.3
Df Residuals:	49	BIC:	388.3
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Adults_who_have_Obesity_pct	2.0067	0.049	40.651	0.000	1.908	2.106

Omnibus:	2.018	Durbin-Watson:	1.842
Prob(Omnibus):	0.365	Jarque-Bera (JB):	1.928
Skew:	-0.450	Prob(JB):	0.381
Kurtosis:	2.661	Cond. No.	1.00

RESEARCH QUESTION 2B PARAMETRIC (OLS) RESULTS

Two Parameters (With "No_Tooth_Loss_18_64_pct"):

With "Complete_Tooth_Loss_65_plus_pct":

1) "No_Tooth_Loss_18_64_pct" vs. ["Complete_Tooth_Loss_65_plus_pct", "Dentist_Visits_for_Adults_pct"]

Dep. Variable:	No_Tooth_Loss_18_64_pct	R-squared (uncentered):	0.998
Model:	OLS	Adj. R-squared (uncentered):	0.998
Method:	Least Squares	F-statistic:	1.119e+04
Date:	Mon, 06 May 2024	Prob (F-statistic):	8.62e-65
Time:	13:02:03	Log-Likelihood:	-127.20
No. Observations:	50	AIC:	258.4
Df Residuals:	48	BIC:	262.2
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Complete_Tooth_Loss_65_plus_pct	0.0472	0.096	0.491	0.626	-0.146	0.240
Dentist_Visits_for_Adults_pct	1.0035	0.022	46.177	0.000	0.960	1.047

Omnibus:	0.061	Durbin-Watson:	2.045
Prob(Omnibus):	0.970	Jarque-Bera (JB):	0.047
Skew:	-0.042	Prob(JB):	0.977
Kurtosis:	2.876	Cond. No.	14.8

2) "No_Tooth_Loss_18_64_pct" vs. ["Complete_Tooth_Loss_65_plus_pct", "Adults_who_Smoke_pct"]

Dep. Variable:	No_Tooth_Loss_18_64_pct	R-squared (uncentered):	0.934
Model:	OLS	Adj. R-squared (uncentered):	0.931
Method:	Least Squares	F-statistic:	337.5
Date:	Mon, 06 May 2024	Prob (F-statistic):	5.36e-29
Time:	13:02:03	Log-Likelihood:	-213.05
No. Observations:	50	AIC:	430.1
Df Residuals:	48	BIC:	433.9
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Complete_Tooth_Loss_65_plus_pct	-0.8952	1.105	-0.810	0.422	-3.116	1.326
Adults_who_Smoke_pct	4.7192	0.999	4.725	0.000	2.711	6.727

Omnibus:	0.704	Durbin-Watson:	1.332
Prob(Omnibus):	0.703	Jarque-Bera (JB):	0.768
Skew:	-0.117	Prob(JB):	0.681
Kurtosis:	2.440	Cond. No.	13.2

3) "No_Tooth_Loss_18_64_pct" vs. ["Complete_Tooth_Loss_65_plus_pct", "Adults_who_have_Short_Sleep_Duration_pct"]

Dep. Variable:	No_Tooth_Loss_18_64_pct	R-squared (uncentered):	0.982
Model:	OLS	Adj. R-squared (uncentered):	0.982
Method:	Least Squares	F-statistic:	1343.
Date:	Mon, 06 May 2024	Prob (F-statistic):	7.33e-43
Time:	13:02:03	Log-Likelihood:	-179.80
No. Observations:	50	AIC:	363.6
Df Residuals:	48	BIC:	367.4
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Complete_Tooth_Loss_65_plus_pct	-1.1596	0.377	-3.073	0.003	-1.918	-0.401
Adults_who_have_Short_Sleep_Duration_pct	2.4868	0.168	14.763	0.000	2.148	2.826

Omnibus:	1.373	Durbin-Watson:	2.019
Prob(Omnibus):	0.503	Jarque-Bera (JB):	0.887
Skew:	-0.323	Prob(JB):	0.642
Kurtosis:	3.095	Cond. No.	11.7

RESEARCH QUESTION 2B PARAMETRIC (OLS) RESULTS

4) "No_Tooth_Loss_18_64_pct" vs. ["Complete_Tooth_Loss_65_plus_pct", "Adults_who_have_Obesity_pct"]

Dep. Variable:	No_Tooth_Loss_18_64_pct	R-squared (uncentered):	0.978
Model:	OLS	Adj. R-squared (uncentered):	0.977
Method:	Least Squares	F-statistic:	1043.
Date:	Mon, 06 May 2024	Prob (F-statistic):	2.83e-40
Time:	13:02:03	Log-Likelihood:	-186.00
No. Observations:	50	AIC:	376.0
Df Residuals:	48	BIC:	379.8
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Complete_Tooth_Loss_65_plus_pct	-1.7945	0.490	-3.666	0.001	-2.779	-0.810
Adults_who_have_Obesity_pct	2.8046	0.222	12.629	0.000	2.358	3.251

Omnibus:	1.104	Durbin-Watson:	2.177
Prob(Omnibus):	0.576	Jarque-Bera (JB):	0.612
Skew:	0.262	Prob(JB):	0.737
Kurtosis:	3.139	Cond. No.	13.3

With "Dentist_Visits_for_Adults_pct":

1) "No_Tooth_Loss_18_64_pct" vs. ["Dentist_Visits_for_Adults_pct", "Adults_who_Smoke_pct"]

Dep. Variable:	No_Tooth_Loss_18_64_pct	R-squared (uncentered):	0.998
Model:	OLS	Adj. R-squared (uncentered):	0.998
Method:	Least Squares	F-statistic:	1.114e+04
Date:	Mon, 06 May 2024	Prob (F-statistic):	9.58e-65
Time:	13:02:03	Log-Likelihood:	-127.31
No. Observations:	50	AIC:	258.6
Df Residuals:	48	BIC:	262.4
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Dentist_Visits_for_Adults_pct	1.0093	0.026	38.130	0.000	0.956	1.062
Adults_who_Smoke_pct	0.0179	0.106	0.169	0.866	-0.195	0.231

Omnibus:	0.247	Durbin-Watson:	2.041
Prob(Omnibus):	0.884	Jarque-Bera (JB):	0.054
Skew:	-0.081	Prob(JB):	0.973
Kurtosis:	3.001	Cond. No.	16.5

2) "No_Tooth_Loss_18_64_pct" vs. ["Dentist_Visits_for_Adults_pct", "Adults_who_have_Short_Sleep_Duration_pct"]

Dep. Variable:	No_Tooth_Loss_18_64_pct	R-squared (uncentered):	0.998
Model:	OLS	Adj. R-squared (uncentered):	0.998
Method:	Least Squares	F-statistic:	1.113e+04
Date:	Mon, 06 May 2024	Prob (F-statistic):	9.69e-65
Time:	13:02:03	Log-Likelihood:	-127.32
No. Observations:	50	AIC:	258.6
Df Residuals:	48	BIC:	262.5
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Dentist_Visits_for_Adults_pct	1.0170	0.050	20.510	0.000	0.917	1.117
Adults_who_have_Short_Sleep_Duration_pct	-0.0068	0.098	-0.070	0.945	-0.204	0.190

Omnibus:	0.370	Durbin-Watson:	2.050
Prob(Omnibus):	0.831	Jarque-Bera (JB):	0.096
Skew:	-0.104	Prob(JB):	0.953
Kurtosis:	3.054	Cond. No.	18.1

RESEARCH QUESTION 2B PARAMETRIC (OLS) RESULTS

3) "No_Tooth_Loss_18_64_pct" vs. ["Dentist_Visits_for_Adults_pct", "Adults_who_have_Obesity_pct"]

Dep. Variable:	No_Tooth_Loss_18_64_pct	R-squared (uncentered):	0.998
Model:	OLS	Adj. R-squared (uncentered):	0.998
Method:	Least Squares	F-statistic:	1.149e+04
Date:	Mon, 06 May 2024	Prob (F-statistic):	4.52e-65
Time:	13:02:03	Log-Likelihood:	-126.52
No. Observations:	50	AIC:	257.0
Df Residuals:	48	BIC:	260.9
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Dentist_Visits_for_Adults_pct	0.9658	0.039	24.803	0.000	0.887	1.044
Adults_who_have_Obesity_pct	0.0975	0.078	1.247	0.218	-0.060	0.255

Omnibus:	0.195	Durbin-Watson:	2.014
Prob(Omnibus):	0.907	Jarque-Bera (JB):	0.393
Skew:	-0.070	Prob(JB):	0.822
Kurtosis:	2.589	Cond. No.	14.5

With "Adults_who_Smoke_pct":

1) "No_Tooth_Loss_18_64_pct" vs. ["Adults_who_Smoke_pct", "Adults_who_have_Short_Sleep_Duration_pct"]

Dep. Variable:	No_Tooth_Loss_18_64_pct	R-squared (uncentered):	0.980
Model:	OLS	Adj. R-squared (uncentered):	0.979
Method:	Least Squares	F-statistic:	1183.
Date:	Mon, 06 May 2024	Prob (F-statistic):	1.45e-41
Time:	13:02:03	Log-Likelihood:	-182.90
No. Observations:	50	AIC:	369.8
Df Residuals:	48	BIC:	373.6
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Adults_who_Smoke_pct	-0.7302	0.442	-1.651	0.105	-1.620	0.159
Adults_who_have_Short_Sleep_Duration_pct	2.3370	0.218	10.700	0.000	1.898	2.776

Omnibus:	0.863	Durbin-Watson:	1.939
Prob(Omnibus):	0.650	Jarque-Bera (JB):	0.946
Skew:	-0.243	Prob(JB):	0.623
Kurtosis:	2.534	Cond. No.	13.4

2) "No_Tooth_Loss_18_64_pct" vs. ["Adults_who_Smoke_pct", "Adults_who_have_Obesity_pct"]

Dep. Variable:	No_Tooth_Loss_18_64_pct	R-squared (uncentered):	0.975
Model:	OLS	Adj. R-squared (uncentered):	0.974
Method:	Least Squares	F-statistic:	942.9
Date:	Mon, 06 May 2024	Prob (F-statistic):	2.99e-39
Time:	13:02:03	Log-Likelihood:	-188.46
No. Observations:	50	AIC:	380.9
Df Residuals:	48	BIC:	384.7
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Adults_who_Smoke_pct	-1.7550	0.633	-2.773	0.008	-3.028	-0.482
Adults_who_have_Obesity_pct	2.8780	0.318	9.062	0.000	2.239	3.517

Omnibus:	0.257	Durbin-Watson:	2.220
Prob(Omnibus):	0.879	Jarque-Bera (JB):	0.401
Skew:	-0.146	Prob(JB):	0.818
Kurtosis:	2.673	Cond. No.	17.1

RESEARCH QUESTION 2B PARAMETRIC (OLS) RESULTS

With "Adults_who_have_Short_Sleep_Duration_pct":

1)"No_Tooth_Loss_18_64_pct" vs. ["Adults_who_have_Short_Sleep_Duration_pct", "Adults_who_have_Obesity_pct"]

Dep. Variable:	No_Tooth_Loss_18_64_pct	R-squared (uncentered):	0.979
Model:	OLS	Adj. R-squared (uncentered):	0.978
Method:	Least Squares	F-statistic:	1137.
Date:	Mon, 06 May 2024	Prob (F-statistic):	3.69e-41
Time:	13:02:03	Log-Likelihood:	-183.88
No. Observations:	50	AIC:	371.8
Df Residuals:	48	BIC:	375.6
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Adults_who_have_Short_Sleep_Duration_pct	1.6476	0.379	4.346	0.000	0.885	2.410
Adults_who_have_Obesity_pct	0.3426	0.385	0.889	0.378	-0.432	1.117

Omnibus:	2.854	Durbin-Watson:	1.838
Prob(Omnibus):	0.240	Jarque-Bera (JB):	2.440
Skew:	-0.540	Prob(JB):	0.295
Kurtosis:	2.929	Cond. No.	18.2

RESEARCH QUESTION 2C PARAMETRIC (GLM) IMPLEMENTATION

Used statsmodels GLM library to train generalized linear models based on data wrangled from multiple sources(CDC, BEA) . Designed multiple simple models (3 or less features) and complex models (3+ features) and compared AIC. Each model had three GLM's with likelihood functions being either Poisson, Negative Binomial, or Gamma. Justification for these choices in link functions is the following : Poisson might be a great fit if the variance of our results are 'tight' and not too different from the mean, Negative Binomial might be a great fit in the case that our variance is larger, Gamma is a more appropriate distribution since our results are real continuous variables and might be a better fit if our results have a skew in their distribution. Poisson and Negative Binomial likelihood functions are more appropriate for count data, but for simplicity we can abstract the percentages to be 'counts'.

Data collected from various sources and filtered to include only age adjusted(when appropriate) data. Data was then wrangled together by outer joins on year and state names, so that some rows had null values. This is so when selecting certain columns to use in a model, we can get all possible data and then filter only rows with no null values, thus ensuring we get all possible data for this model.

Used statsmodels GLM library to perform fitting of GLMs. Selected what features are in linear model and filtered data so it contains only selected feature columns and has no null datavalues. Split data into training and test data, then fit the model to training data specifying the function family to either be Poisson, Negative Binomial or Gamma. Then predicted results of test data and calculated difference between predicted and real results and plotted bar graph of distribution of difference. To compare different models we used AIC, with models that have the lower AIC being a better choice. We also made sure that the chi2 of a model was reasonable.

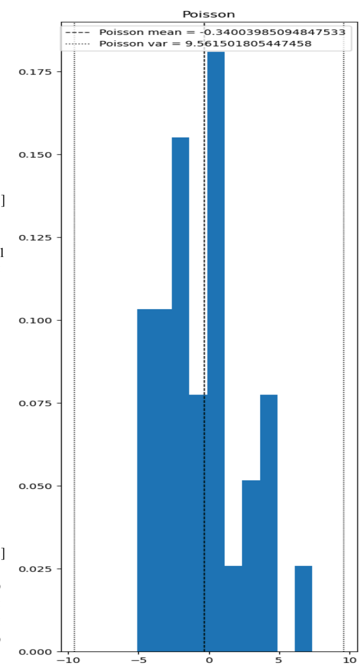
RESEARCH QUESTION 2C PARAMETRIC (GLM) RESULTS

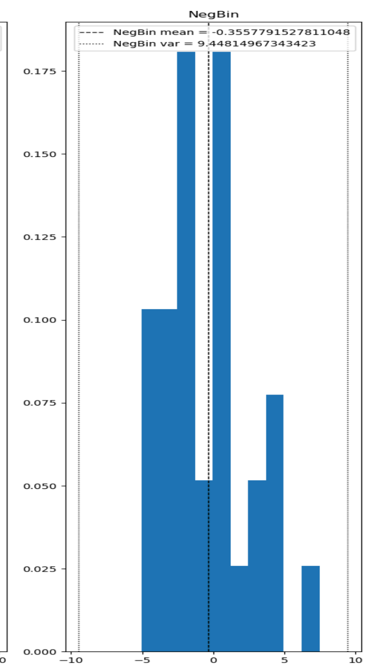
These are the results of our best fitting models. On the left is a summary of our results on training the GLM on our test data for each likelihood function. The AIC is below each summary. On the right are histograms of the difference between our predicted value from test data and the actual test data value for each likelihood function. Mean and variance of these differences are plotted as well.

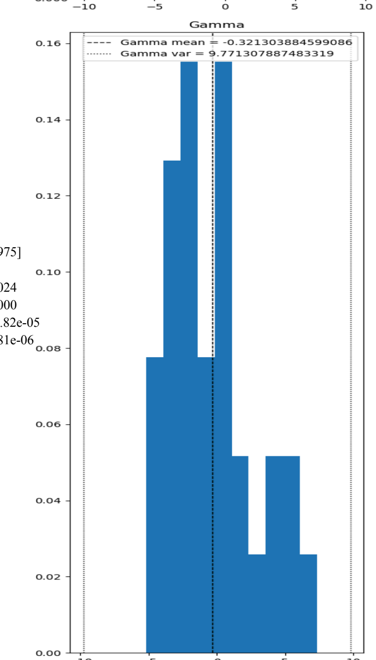
Second Simple Model:

*No Tooth loss = b1 + b2 * Obesity pct + b3 * Good Sleep Duration pct + b4 * Fluoridated Water pct*

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:     No_Tooth_Loss_18_64_pct   No. Observations:       123
Model:                   GLM                 Df Residuals:           119
Model Family:            Poisson             Df Model:               3
Link Function:           Log                 Scale:                  1.0000
Method:                  IRLS                Log-Likelihood:         -382.01
Date:              Wed, 01 May 2024          Deviance:               26.244
Time:                    06:05:56            Pearson chi2:           26.0
No. Iterations:          3                   Pseudo R-squ. (CS):     0.1193
Covariance Type:         nonrobust
==============================================================================
                         coef     std err      z        P>|z|    [0.025    0.975]
------------------------------------------------------------------------------
const                    3.8422   0.270        14.235   0.000    3.313     4.371
Obesity_pct              -0.0079  0.003        -2.330   0.020    -0.015    -0.001
Sleep_Duration_pct       0.0084   0.004        2.352    0.019    0.001     0.015
Fluoridated_Water_pct    0.0003   0.001        0.608    0.543    -0.001    0.001
==============================================================================
Model with Poisson AIC :772.0107173071916
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:     No_Tooth_Loss_18_64_pct   No. Observations:       123
Model:                   GLM                 Df Residuals:           119
Model Family:            NegativeBinomial    Df Model:               3
Link Function:           Log                 Scale:                  1.0000
Method:                  IRLS                Log-Likelihood:         -635.55
Date:              Wed, 01 May 2024          Deviance:               0.41673
Time:                    06:05:56            Pearson chi2:           0.407
No. Iterations:          3                   Pseudo R-squ. (CS):     0.001990
Covariance Type:         nonrobust
==============================================================================
                         coef     std err      z        P>|z|    [0.025    0.975]
------------------------------------------------------------------------------
const                    3.8515   2.172        1.773    0.076    -0.406    8.109
Obesity_pct              -0.0082  0.027        -0.303   0.762    -0.062    0.045
Sleep_Duration_pct       0.0084   0.029        0.291    0.771    -0.048    0.065
Fluoridated_Water_pct    0.0004   0.004        0.078    0.938    -0.008    0.009
==============================================================================
Model with NegBin AIC :1279.1084666602637
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:     No_Tooth_Loss_18_64_pct   No. Observations:       123
Model:                   GLM                 Df Residuals:           119
Model Family:            Gamma               Df Model:               3
Link Function:           InversePower        Scale:                  0.0035178
Method:                  IRLS                Log-Likelihood:         -337.97
Date:              Wed, 01 May 2024          Deviance:               0.42912
Time:                    06:05:56            Pearson chi2:           0.419
No. Iterations:          5                   Pseudo R-squ. (CS):     0.4301
Covariance Type:         nonrobust
==============================================================================
                         coef     std err      z        P>|z|    [0.025    0.975]
------------------------------------------------------------------------------
const                    0.0205   0.002        10.238   0.000    0.017     0.024
Obesity_pct              0.0001   2.52e-05     4.895    0.000    7.4e-05   0.000
Sleep_Duration_pct       -0.0001  2.64e-05     -4.925   0.000    -0.000    -7.82e-05
Fluoridated_Water_pct    -5.27e-06  4.12e-06   -1.279   0.201    -1.33e-05 2.81e-06
==============================================================================
Model with Gamma AIC :683.9485911410311
```







RESEARCH QUESTION 2C PARAMETRIC (GLM) RESULTS

First Complex Model:

*No Tooth loss = b1 + b2 * Dentists Visits pct + b3 * Smoke pct + b4 * Obesity pct + b5*Good Sleep Duration*

Generalized Linear Model Regression Results

Dep. Variable:	No_Tooth_Loss_18_64_pct	No. Observations:	163
Model:	GLM	Df Residuals:	158
Model Family:	Poisson	Df Model:	4
Link Function:	Log	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-495.70
Date:	Wed, 01 May 2024	Deviance:	12.137
Time:	06:05:58	Pearson chi2:	12.0
No. Iterations:	3	Pseudo R-squ. (CS):	0.2309
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	3.1876	0.322	9.910	0.000	2.557	3.818
Dentist_Visits_pct	0.0077	0.003	2.924	0.003	0.003	0.013
Smoke_pct	-0.0065	0.003	-1.872	0.061	-0.013	0.000
Obesity_pct	0.0024	0.003	0.746	0.456	-0.004	0.009
Sleep_Duration_pct	0.0081	0.003	2.708	0.007	0.002	0.014

Model with Poisson AIC :1001.4053892180461

Generalized Linear Model Regression Results

Dep. Variable:	No_Tooth_Loss_18_64_pct	No. Observations:	163
Model:	GLM	Df Residuals:	158
Model Family:	NegativeBinomial	Df Model:	4
Link Function:	Log	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-843.63
Date:	Wed, 01 May 2024	Deviance:	0.19046
Time:	06:05:58	Pearson chi2:	0.186
No. Iterations:	4	Pseudo R-squ. (CS):	0.004084
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	3.1851	2.626	1.213	0.225	-1.963	8.333
Dentist_Visits_pct	0.0078	0.021	0.366	0.714	-0.034	0.050
Smoke_pct	-0.0066	0.028	-0.235	0.814	-0.062	0.048
Obesity_pct	0.0023	0.026	0.089	0.929	-0.049	0.053
Sleep_Duration_pct	0.0081	0.024	0.332	0.740	-0.040	0.056

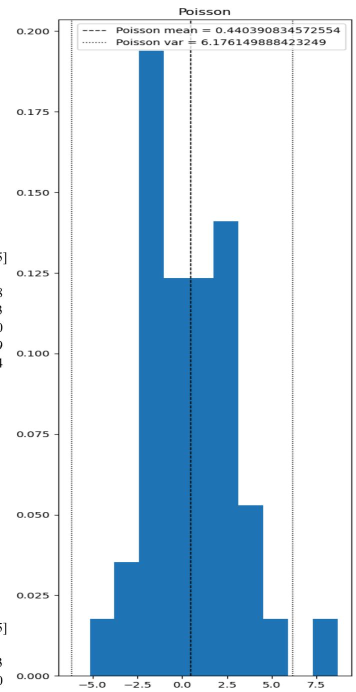
Model with NegBin AIC :1697.2551909699564

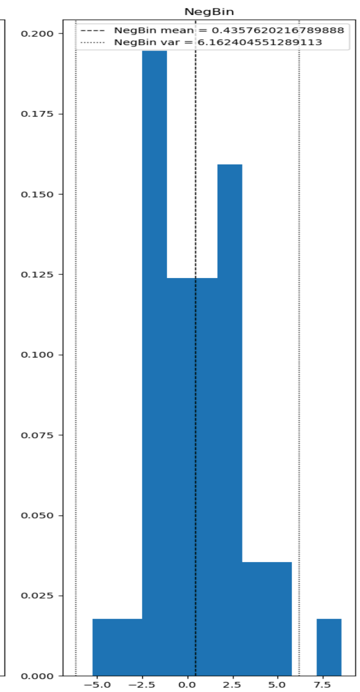
Generalized Linear Model Regression Results

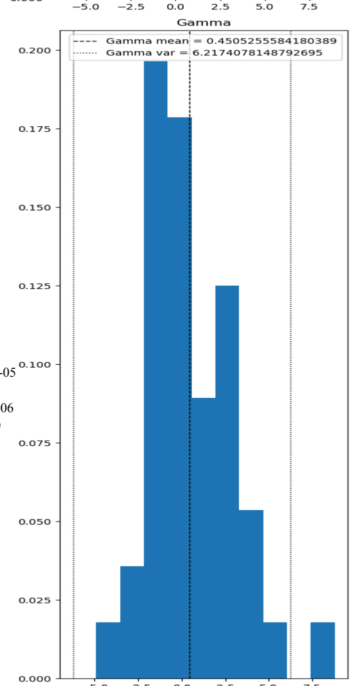
Dep. Variable:	No_Tooth_Loss_18_64_pct	No. Observations:	163
Model:	GLM	Df Residuals:	158
Model Family:	Gamma	Df Model:	4
Link Function:	InversePower	Scale:	0.0012676
Method:	IRLS	Log-Likelihood:	-366.45
Date:	Wed, 01 May 2024	Deviance:	0.20543
Time:	06:05:58	Pearson chi2:	0.200
No. Iterations:	5	Pseudo R-squ. (CS):	0.9601
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.0306	0.001	21.686	0.000	0.028	0.033
Dentist_Visits_pct	-0.0001	1.16e-05	-10.260	0.000	-0.000	-9.62e-05
Smoke_pct	0.0001	1.52e-05	6.626	0.000	7.09e-05	0.000
Obesity_pct	-3.46e-05	1.41e-05	-2.452	0.014	-6.23e-05	-6.94e-06
Sleep_Duration_pct	-0.0001	1.32e-05	-9.546	0.000	-0.000	-0.000

Model with Gamma AIC :742.9040730183242







RESEARCH QUESTION 2C PARAMETRIC (GLM) RESULTS

Second Complex Model:

$$No\ Tooth\ loss = b1 + b2 * Dentists\ Visits\ pct + b3 * Smoke\ pct + b4 * Obesity\ pct$$
$$+ b5 * Good\ Sleep\ Duration + b6 * Fluoridated\ Water\ pct$$
$$+ b7 * Average\ Personal\ Income$$

Generalized Linear Model Regression Results

Dep. Variable:	No_Tooth_Loss_18_64_pct	No. Observations:	123
Model:	GLM	Df Residuals:	116
Model Family:	Poisson	Df Model:	6
Link Function:	Log	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-373.64
Date:	Wed, 01 May 2024	Deviance:	10.180
Time:	06:06:00	Pearson chi2:	10.1
No. Iterations:	3	Pseudo R-squ. (CS):	0.2364
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	3.2990	0.404	8.162	0.000	2.507	4.091
Dentist_Visits_pct	0.0079	0.003	2.527	0.011	0.002	0.014
Smoke_pct	-0.0058	0.004	-1.322	0.186	-0.014	0.003
Obesity_pct	0.0003	0.005	0.072	0.942	-0.009	0.009
Sleep_Duration_pct	0.0071	0.004	1.926	0.054	-0.000	0.014
Fluoridated_Water_pct	0.0002	0.001	0.380	0.704	-0.001	0.001
Avg_Personal_Income	-7.134e-07	2.21e-06	-0.322	0.747	-5.05e-06	3.63e-06

Model with Poisson AIC :761.2833072781766

Generalized Linear Model Regression Results

Dep. Variable:	No_Tooth_Loss_18_64_pc	No. Observations:	123
Model:	GLM	Df Residuals:	116
Model Family:	NegativeBinomial	Df Model:	6
Link Function:	Log	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-634.77
Date:	Wed, 01 May 2024	Deviance:	0.16364
Time:	06:06:00	Pearson chi2:	0.159
No. Iterations:	3	Pseudo R-squ. (CS):	0.004245
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	3.3065	3.284	1.007	0.314	-3.129	9.742
Dentist_Visits_pct	0.0081	0.025	0.317	0.751	-0.042	0.058
Smoke_pct	-0.0060	0.035	-0.168	0.866	-0.075	0.063
Obesity_pct	0.0002	0.037	0.006	0.995	-0.072	0.072
Sleep_Duration_pct	0.0070	0.030	0.233	0.816	-0.052	0.066
Fluoridated_Water_pct	0.0002	0.004	0.052	0.959	-0.009	0.009
Avg_Personal_Income	-7.708e-07	1.8e-05	-0.043	0.966	-3.61e-05	3.45e-05

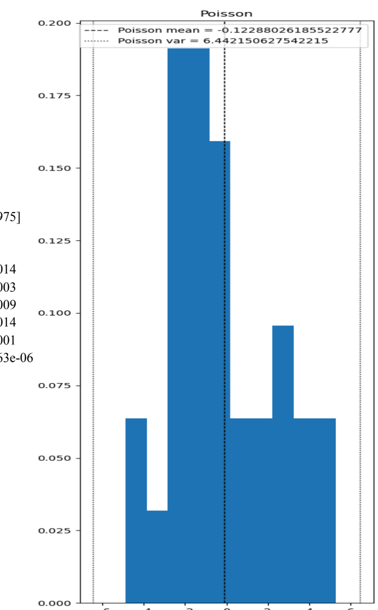
Model with NegBin AIC :1283.535706491311

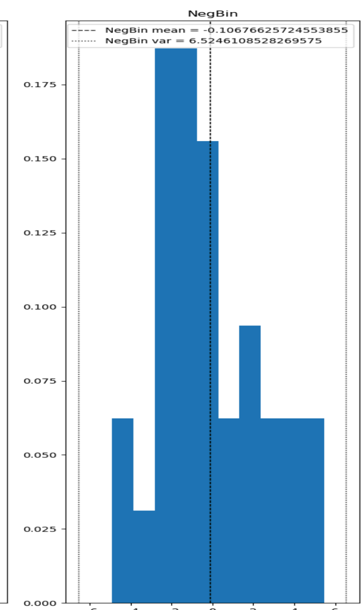
Generalized Linear Model Regression Results

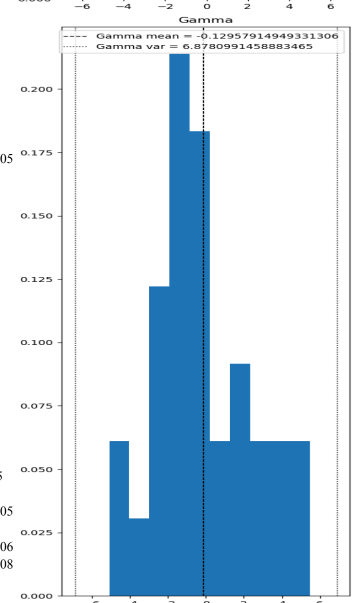
Dep. Variable:	No_Tooth_Loss_18_64_pct	No. Observations:	123
Model:	GLM	Df Residuals:	116
Model Family:	Gamma	Df Model:	6
Link Function:	InversePower	Scale:	0.0014644
Method:	IRLS	Log-Likelihood:	-282.08
Date:	Wed, 01 May 2024	Deviance:	0.17480
Time:	06:06:00	Pearson chi2:	0.170
No. Iterations:	5	Pseudo R-squ. (CS):	0.9452
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.0290	0.002	15.125	0.000	0.025	0.033
Dentist_Visits_pct	-0.0001	1.5e-05	-8.299	0.000	-0.000	-9.48e-05
Smoke_pct	9.095e-05	2.08e-05	4.367	0.000	5.01e-05	0.000
Obesity_pct	-1.761e-06	2.18e-05	-0.081	0.936	-4.44e-05	4.09e-05
Sleep_Duration_pct	-0.0001	1.76e-05	-6.355	0.000	-0.000	-7.73e-05
Fluoridated_Water_pct	-3.224e-06	2.67e-06	-1.208	0.227	-8.45e-06	2.01e-06
Avg_Personal_Income	1.08e-08	1.05e-08	1.033	0.302	-9.7e-09	3.13e-08

Model with Gamma AIC :578.1610512323341







RESEARCH QUESTION 2C PARAMETRIC (GLM) INTERPRETATION OF RESULTS

The top 3 models are : Second Complex Model with Gamma function (AIC 578), Second Simple Model with Gamma function (AIC 683) and the First Complex Model with Gamma Function (AIC 742). The Complex models seem to have the best predictions(both having prediction variance under 10). From these models, we can see that dentist visits and good sleep duration contribute to better statewide tooth health.

Our Gamma GLM can be modeled using the following Equations:

(1) $E[Y|X] = g^{-1}(XB)$

(2) $g(u) = 1/u = g^{-1}(u)$

Given these equations, if we increase a parameter by one:

$E[Y|X(x_i = x_i + 1)] = g^{-1}(XB + b_i)$

=> $\frac{1}{XB + b_i}$

If we choose our best model, our coefficients are:

Intercept = 0.0290, Dentist = -0.0001, Smoke = 9.095e-05, Obesity = -1.761e-06
Sleep = -0.0001, Fluoridation = -3.224e-06, Income = 1.08e-08

Thus, choosing if we have an arbitrary starting value for no tooth loss percentage P, a +1% increase in the percentage of yearly dentist visits increases the percentage of no tooth loss in adults by P/(P - 0.0001). The same can be said for sleep duration. Other variables with this property are Obesity and Water Fluoridation. Obesity has a best fitting coefficient of -1.761e-06 with 95% confidence it falls between [-4.44e-05, 4.09e-05], so evidence points to it being a predictor of good dental health but can be ‘negligible’. Water Fluoridation has a best fitting coefficient of -3.224e-06 with 95% confidence it falls between [-8.45e-06, 2.01e-06], so evidence points to it being a predictor of good dental health. Unlike Obesity, the ‘true’ parameter value is more likely to be negative since the confidence interval is skewed to the left.

Smoke has a best fitting coefficient of 9.095e-05 with a 95% confidence interval between [0, 5.01e-05], thus it is likely to be positive and an indicator of bad dental health following our
P/(P + b) rule. Higher Average Personal Income seems to be a indicator of bad dental health, but with a best fitting coefficient of 1.08e-08 and a 95% confidence interval of [-9.7e-09, 3.13e-08] it is ‘negligible’ like obesity due to the confidence interval.

Conclusions

RESEARCH QUESTION 1

The null result observed while assessing causal effects can be attributed to several factors. Firstly, inaccuracies in the propensity score model may have arisen from incorrect covariate selection or insufficient training data. If essential variables are overlooked or irrelevant ones included, the model may fail to adequately balance treatment and control groups, compromising the validity of the results. In this case, it is quite likely covariates were missed, since the research question highly depends on domain-specific knowledge that we don't have, and there are a huge amount of covariates (of which we only considered <15). Additionally, the absence of a logical reasoning for expecting a causal relationship between the treatment and outcomes can lead to null findings; we thought that the proposed treatment and outcomes may have causal effect, but are in no means domain experts so maybe in reality the treatment/outcome pair chosen was unwise. Unmeasured confounding variables, which are associated with both treatment assignment and outcomes, can introduce bias, obscuring true causal effects. For insufficient training data, the propensity score model was looking at the data from a state level, which may not have been enough data to train the model properly. Overall, addressing these potential sources of bias requires careful consideration of model design, variable selection, and specific domain knowledge to ensure robust and reliable conclusions about causal relationships.

RESEARCH QUESTION 2

The application of non-parametric models, namely decision trees and random forests, provided additional insights into the factors influencing statewide dental health. These models pinpointed critical variables such as the frequency of dentist visits and smoking habits as significant predictors of dental health outcomes. Decision trees, while offering granular insights into the data, were prone to overfitting, affecting their generalizability. In contrast, random forests demonstrated substantial robustness and were more adept at managing the inherent complexities of the demographic data. The stability and reliability of the random forest model, affirmed through extensive cross-validation, indicate its suitability for integration into state and federal health initiatives. This approach not only corroborates the impact of well-recognized health behaviors on dental outcomes but also highlights the necessity for targeted public health interventions in communities with lower access to dental care and higher prevalence of risky health behaviors.

In terms of the application of the Ordinary Least Squares Regression (OLS) model, the coefficient values calculated by the analyses of the single parameters, with "No_Tooth_Loss_18_64_pct" as the dependent variable, range from 1.0136 to 4.2650. As one can observe, the signs of the coefficient values calculated by the analyses of the single parameters are positive which implies that "No_Tooth_Loss_18_64_pct" has a positive relationship with all of the independent variables. The coefficient values calculated by the analyses of the double parameters, with "No_Tooth_Loss_18_64_pct" as the dependent variable, range from -1.7945 to 4.719. As one can observe, the signs of the coefficient values calculated by the analyses of the double parameters which implies that there is a variety of negative relationships and positive relationships between the dependent variable and all of the independent variables.

A complex Gamma GLM can be used to reasonably predict the statewide percent of 'good dental health' (defined as percentage of adults 18-64 who have no tooth loss) with an error of around ∓6. By interpreting our model, we see that features like the percentage of adults who have yearly dentist visits and percentage of adults who have sufficient sleep duration are positively correlated with higher good dental health. Percentage of adults who use tobacco is negatively correlated with statewide good dental health. Water fluoridation is likely to be positively correlated with good dental health, since the 95% confidence interval has both positive and negative coefficients for this feature. Average State Income and percentage of adults with obesity have inconclusive influence in prediction. This prediction was not very granular, an analysis on the county level would probably produce better results. Regardless, it is clear that yearly dentist visits and good sleep duration predict higher good dental health. Thus, state and federal programs should focus on areas with low values of these features to provide aid to and study if they want to improve the population's oral health.