# **Technical Exposition**

# **Data Exploration and Preprocessing**

Next, we needed statewide data that gave us info on commitment to healthier foods. We decided to use the percentage of health-related laws that were about food safety. From the NCSL website, we scraped a list of all health laws passed, then extracted the total number of health and food laws passed per state per year using sentiment analysis, and the total number of food laws passed per state per year. This gave a measurement of how much health and food considerations were a priority in the eyes of lawmakers and politicians **(??? yes or no).**

In this analysis, we investigated the correlation between obesity rates and food safety laws across various U.S. states using data from 2018. After loading the data, we performed interpolation on numerical columns to handle missing values.

We filtered the dataset to focus on the percentage of adults aged 18 years and older with obesity, specifically for 2018. To address the issue of multiple data values for one state per year due to stratification by gender and race, we calculated a weighted average. We introduced a new column, a weighted value, which is the product of the data value and sample size. We then grouped the data by state to sum the weighted values and sample sizes, ultimately calculating the weighted average for each state.

Simultaneously, we processed another dataset containing information on food safety laws by state and year. We transposed this data frame, set appropriate column names, and filtered it to focus on 2018. After resetting the index and renaming columns, we merged this DataFrame with the obesity data on the state abbreviation. Finally, we filtered the merged DataFrame to include only valid state abbreviations.

Since our data proved to be heteroskedastic, we used a log-log transformation. To prepare the data, we first replaced all of the 0 values with 1 to ensure compatibility with the logarithm function. We then applied the natural logarithm to both the percentage of laws passed and the percentage of identified obese people. Additionally, we winsorized the top and bottom 1% of the log-transformed data to reduce the impact of outliers.

For the statistical analysis, we performed an Ordinary Least Squares (OLS) **WHY** regression to explore the relationship between the variables. The model was defined as: log⁡(WeightedAverage)=β0+β1log⁡(Percentage)\log(\text{WeightedAverage}) = \beta\_0 + \beta\_1 \log(\text{Percentage})log(WeightedAverage)=β0​+β1​log(Percentage).

Exploratory Data Analysis

**Modeling**

# State-level analysis:

After conducting a country-level analysis, we conduct a state-level analysis. Based on our findings in the country-level analysis, we investigate the following question at the state level: *is there a statistically significant relationship between the percentage of health-related laws passed and the obesity rate for that state?*

To investigate this data, we.. insert preprocessing (we did already)

# Results

The regression analysis yielded an intercept (β0\beta\_0β0​) of 3.4302 with a p-value less than 0.001 and a slope (β1\beta\_1β1​) of -0.0006 with a p-value of 0.843. The R-squared value was 0.001, indicating that the model explains only 0.1% of the variance in the log-transformed weighted average obesity rate. The F-statistic was 0.03956 with a p-value of 0.843.

The regression results suggest that the slope (β1\beta\_1β1​) is not statistically significant, with a p-value of 0.843, indicating that the percentage of health-related laws does not have a meaningful effect on the weighted average obesity rates across states. The R-squared value is very close to zero, implying that the model does not explain much of the variability in obesity rates.

The model has several shortcomings, including its simplicity, which may not capture more complex relationships, and the potential omission of other influential factors. Winsorization, while reducing the effect of outliers, may also exclude valuable information. The overall performance of the model was weak, with an R-squared value of 0.001, indicating that it explains only 0.1% of the variance in obesity rates.

# Conclusion

The analysis reveals no significant relationship between the percentage of laws related to obesity and health and the weighted average number of obese people in the states. The weak correlation and non-significant regression results suggest that other factors might be more influential in determining obesity rates. Further research is needed to explore additional variables and more complex models to gain a better understanding of the factors influencing obesity.

**Citations**

Cobb, Laura K., et al. “The Relationship of the Local Food Environment with Obesity: A Systematic Review of Methods, Study Quality, and Results.” Obesity, vol. 23, no. 7, 12 June 2015, pp. 1331–1344, https://doi.org/10.1002/oby.21118.

Edwards, John S. A., et al. “Overweight, Obesity and the Food Service Industry.” Food Service Technology, vol. 5, no. 2-4, June 2005, pp. 85–94, https://doi.org/10.1111/j.1471-5740.2005.00115.x.

CITE THE SCRAPING STUFF WE USED