**Factors Driving Public Health**

**CAPSTONE FINAL REPORT - MAY 08, 2023**

**BY**

Healthy Bytes: Mir Abdullah, Rubina Joshi, Nahom Kifetew, Connor Pepin, Gaurav Shah

**FOR**

Sam Walmsley - Virginia Department of Health

**1 Summary**

In collaboration with Sam Walmsley, the Assistant Chief of Staff at the Virginia Department of Health (VDH), our project seeks to investigate the potential correlation between various factors related to restaurants, food availability, and public health outcomes.

In summary, despite our initial hypothesis that there would be a significant relationship between restaurant types and obesity rates, our analysis revealed that the data did not support this assumption. While there was a slight correlation between certain restaurant types and obesity rates, this relationship was not significant enough to make any definitive conclusions. Moreover, we discovered that poverty rates and limited access to stores and nutritional assistance programs were stronger predictors of obesity rates than restaurant types.

Although our analysis did not yield the expected results, we believe that our research provides valuable insights into the complex relationship between restaurant permits, food availability, and public health outcomes. Our use of statistical techniques such as lasso regression and gradient boosting allowed us to identify key factors that contribute to obesity rates and predict health outcomes with a high degree of accuracy. Furthermore, the tools and methods we used in our research can be applied to other health concerns beyond obesity, enabling policymakers to better understand and address public health challenges.

Despite our efforts to account for the limitations of our data set and statistical methods, we recognize that our research has some limitations due to the limited resources and data available. However, we believe that our findings can help guide future research efforts and inform public health policies aimed at promoting healthy eating habits and addressing the obesity epidemic.

In conclusion, while our analysis did not reveal a significant relationship between restaurant permits and obesity rates, it highlights the need to focus on other factors such as poverty rates and limited access to stores and nutritional assistance programs in addressing the public health challenges associated with poor diet. We hope that our research will contribute to the ongoing efforts to improve public health outcomes and inform policy decisions that promote healthy eating habits and prevent obesity and related health concerns.

**1 Problem Statement**

The issue of public health has been a topic of discussion for many years, with particular emphasis on the impact of diet on health outcomes. Poor diet is often linked to the prevalence of obesity and diabetes, which are significant public health challenges. To address these challenges, researchers have explored various factors related to food availability and restaurants, as they play a crucial role in shaping dietary choices.

Despite these efforts, there is still a lack of understanding of the complex relationship between restaurant permits, food availability, and public health outcomes. Our research project aims to address this gap by investigating the potential correlations between various factors related to restaurants, food availability, and public health outcomes. Specifically, we aim to identify the variables that have the most significant impact on health outcomes and explore the reasons behind this correlation. We will also investigate the relationship between restaurant permits and obesity and diabetes rates and create a predictive model to analyze potential future health outcomes in counties.

However, our research project is not without challenges. One of the major challenges we encountered was the high correlation between some of our variables, which affected our initial plan to perform a multiple linear regression. Additionally, we faced data quality issues, with a lot of null values and messy data that took longer than anticipated to clean and prepare. To address these challenges, we sought expert advice from Professor Inyoung Kim from Virginia Tech's statistics department. She provided valuable insights into the statistical analysis techniques and tools that could be used to address the technical challenges we faced. Through her expertise, we were introduced to the glmnet library in R, which uses the lasso regression method on a Poisson distribution for variable selection and imputation methods for null values.

The findings of this research could provide valuable insights to the VDH in identifying areas or demographics with limited access to healthy food options and in improving the availability of such options in those regions.

**2 Ethical Considerations**

We had two big ethical considerations when doing this analysis:

* First, a lot of the data we have is from public health surveys. While these surveys are very helpful to gain insight, there can be a lot of flaws with collecting data this way. There can be segments of the population that are not represented and therefore masked by those that are, and this can lead to incorrect conclusions about the results of certain public programs. We also need to take into account that some of the data we get comes from hospitals. Hospitals, like every institution, have flaws in their data reporting that can affect the people they serve. Richer hospitals can have better reporting than those with less funding. In some cases, richer hospitals have been found to even over-report some statistics.
* The second topic we have to consider when looking at health is the applications that a model will have on the public. This model allows for the chance to be misused and consequently hurt public health rather than help it. The biggest thing that can be misused is how some of the results can be presented without taking into account the full picture. For example, a basic regression shows that SNAP and WIC are positively correlated with obesity. This would allow bad actors to believe and persuade others to think that these benefits need to be cut in order to solve the obesity problem. Obviously a big problem, because these food supplement programs are a massive source of aid for many families and the only thing keeping them from starvation or spiraling deeper into poverty.

To prevent these from happening, we can take measures to ensure that no one uses the report or model maliciously. We are very clear about where we get our data from so no one extrapolates to other populations that our data does not cover. We also make it very clear that there are lots of factors at play when it comes to statewide public health. Therefore multiple agencies should be considered when making decisions that will affect many Virginians' lives.

**3 Literature Review**

Since our project was doing an analysis of public health, we considered the following pieces of literature as a base guideline for our project.

**3.1 Fast food, race/ethnicity, and income: a geographic analysis**

[1] The first piece of literature that we looked at was a study that investigates the relationship between the number of fast-food restaurants with race/ethnicity, median home values, commercial activity, and the presence of major highways. This was related to our project in the sense that it conducts the analysis at the level of a census tract, which is also what we were seeking to do. This was done by geocoding the restaurants and importing them onto a census tract map via the MapInfo software. This could be applicable in our analysis as well since the data originally given to us does not categorize the restaurants at the level of the census tract. The researchers in this study did not merely overlay the fast food restaurants over the census tract map, they included buffer areas as well as they “...provide a more realistic representation of geographic exposure than census tracts alone because people often have to travel outside of their census tract to purchase goods.” We could also create similar buffer zones as this would take into account the fact that people in one census tract may travel to another census tract for food. This study also used multiple linear regression to determine the correlation between the regressor, which in this case was the fast food restaurant density, and the predictors. They did this by sequentially adding variables to the model to determine their effect on variance. We could perform a similar regression analysis between the number of restaurants and the rates of obesity, diabetes, and other possible public health outcomes. And finally, the study conducted a bivariate analysis using Spearman’s rank coefficient. This was used in determining the correlation between the number of restaurants and another independent variable.

**3.2 Food Insecurity and Health Outcomes**

[2] The second piece of literature we looked at was a study about how food insecurity is negatively associated with health. This is related to our project, as we are looking to find a relationship between food permits and health. The food permits can help us determine areas of food insecurity, as well as types of food that may or may not be available to the population of a certain area. In the study, they showed that children who are more food insecure are more likely to report health problems than children who are not food insecure. These health problems were not just obesity but also asthma, depression, and many others. Food insecurity could also lead to birth defects to some degree. These problems are not only in children but also adults and seniors. This study helped us narrow down the area of health to concentrate our efforts on in order to have a more measurable metric to do analytical problem-solving for our project.

**3.3 Comparing Forms of Neighborhood Instability as Predictors of Violence in Richmond, VA**

[3] The third piece of literature we looked at was a scientific journal that examined a relationship between different forms of neighborhood instability and violence in Richmond, VA. This is related to our project in the sense of exploring the relationship between various factors and public health outcomes. Overall, our project/research is interested in understanding the impact that different environmental factors have on public health outcomes. This can provide valuable insights into the underlying mechanisms that contribute to health disparities and help inform the development of effective strategies for improving public health. Therefore, the study by West would help and provide us with an example of how data can be used to explore the relationship between environmental factors and public health outcomes.

**4 Project Criteria**

**The Criteria established for the project are as follows:**

1. **Identify variables related to restaurants and food availability:** To investigate the potential correlation between various factors related to restaurants, food availability, and public health outcomes, the project aimed to identify relevant variables that could affect health outcomes. These variables include the number of restaurants, restaurant permits, food accessibility, and food insecurity.
2. **Quantify public health outcomes:** Our project sought to quantify public health outcomes, specifically obesity and diabetes rates, as these are significant public health challenges associated with poor diet. Data on these outcomes were obtained from the Virginia Department of Health.
3. **Analyze potential correlations:** Our project aimed to identify the variables that have the most significant impact on health outcomes and explore the reasons behind this correlation. To achieve this, statistical analysis techniques were used, including lasso regression on a Poisson distribution and imputation methods for null values.
4. **Create a predictive model:** The project aimed to create a predictive model to analyze potential future health outcomes in counties. The model would enable the VDH to identify areas of demographics with limited access to healthy food options and improve the availability of such options in those regions.

The quantification of each criterion was achieved through the collection and analysis of relevant data, statistical analysis techniques, and the creation of a predictive model. The identified variables related to restaurants and food availability were quantified using data on the number of restaurants, restaurant permits, food accessibility, and food insecurity. Public health outcomes, specifically obesity and diabetes rates, were quantified using data obtained from the Virginia Department of Health. Statistical analysis techniques, including lasso regression on a Poisson distribution and imputation methods for null values, were used to analyze potential correlations and identify the variables that have the most significant impact on health outcomes. Finally, a predictive model was created to analyze potential future health outcomes in counties, enabling the VDH to identify areas or demographics with limited access to healthy food options and improve the availability of such options in those regions.

**5 Selected Solutions**

Some of the techniques we used for our analysis are the following:

**5.1 Imputing Missing Data with Predictive Mean Matching(PMM)**

When dealing with missing data, it is important to choose an appropriate imputation method that can minimize bias and accurately estimate the true values of the missing data. Therefore, one of the methods that we used to impute missing data is PMM (Predictive Mean Matching). PMM is a multiple imputation method that is based on matching the observed values of a variable to the imputed values generated from a model.

There are several reasons why we chose PMM over other imputation methods:

* The flexible method that can handle different types of data distributions and variable types
* Can reduce bias in the imputed data by matching the observed values to the imputed values using a regression model
* Can also preserve the variability and distribution of the data, which can be important for subsequent statistical analysis
* Is widely used and supported in many statistical software packages, such as the Mice library in R. Overall, PMM is a robust and widely used method for imputing missing data in a variety of data analysis contexts hence we chose it.

**5.2 Standardizing Predictor Variables to Address Multicollinearity**

Knowing that in almost all cases, no matter what model we decide to implement, cases of multicollinearity are a prominent issue. This can cause problems with the interpretation of regression coefficients and can lead to unstable and unreliable estimates of model parameters.

One way to address multicollinearity is to standardize the predictor variables by subtracting the mean and dividing by the standard deviation. Standardizing the variables can help to reduce the correlation between them and improve the stability and reliability of the regression estimates.

Standardization can also help to make the interpretation of the regression coefficients more straightforward, as they will represent the change in the response variable associated with a one-unit change in the predictor variable, measured in standard deviation units.

So, we standardized our predictor variables to combat high multicollinearity in our dataset, and to help us improve the accuracy and interpretability of our results.

**5.3 Choosing LASSO between Ridge or ElasticNet for Feature Selection**

When dealing with high-dimensional datasets, such as our dataset with many predictor variables, it is often necessary to perform feature selection to identify the most important variables that are associated with the response variable. One popular method for feature selection is LASSO (Least Absolute Shrinkage and Selection Operator).

LASSO is a type of regularized regression that adds a penalty term to the regression coefficients, which shrinks the coefficients toward zero and effectively removes the least important variables from the model. LASSO can be particularly useful in our situation where there are many predictor variables, and where some of the variables are highly correlated with each other.

Several reasons why we chose LASSO over Ridge or ElasticNet for feature selection:

* Can perform the variable selection by setting some of the regression coefficients to zero, which can result in a simpler and more interpretable model.
* Can help to address multicollinearity by identifying the most important variables and effectively removing the less important variables from the model by setting the coefficient to 0.
* Can be computationally efficient, particularly when dealing with the high-dimensional datasets we have.

Therefore, we chose Lasso because it can help to identify the most important variables that are associated with the response variable while removing less important variables from the model.

**5.4 Training and testing set**

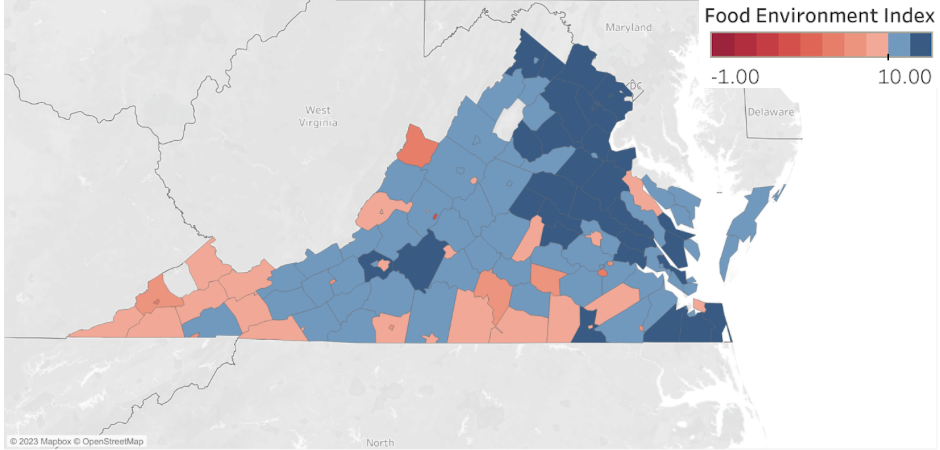
It is necessary to split the data into training and testing sets because we need to evaluate the performance of our predictive model on new data that it has not seen before. If we train and evaluate our model on the same data, it is likely to overfit the data and not generalize well to new data. Splitting the data into a 70/30 ratio for training and testing, respectively, is a common practice in machine learning so we were able to use the 70/30 ratio for training. The larger training set allows our model to learn more patterns and relationships in the data, which can improve its performance. The smaller testing set serves as a check on how well our model can generalize to new data. This evaluation helps us determine if our model is overfitting or underfitting and if we need to adjust our model to improve its accuracy.

**6 Results**

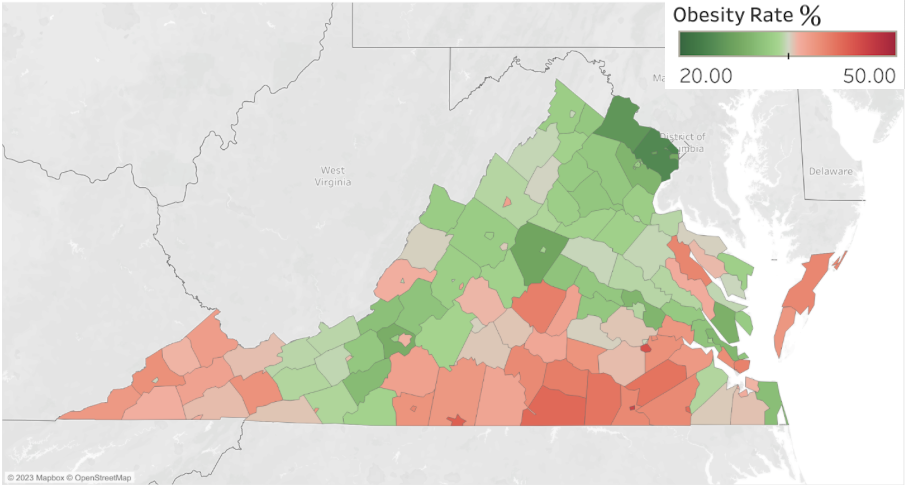
**6.1 Primary Analysis: Lasso Covariate Selection and Regression**

Initially, the data that we received from the VDH was roughly 90,000 rows long where each row corresponds with a specific permit. This ‘permit data,’ as we call it, contained information regarding the food establishment. The information included the name of the food establishment, the type of permit that was issued (if it was a restaurant, grocery store, etc.), the location of the establishment (which Virginia county it was in), FIPS code (an identifiable code for each county), city, zip code, facility address, and application date. Since this permit data was the only data given to us, we looked at other indexes or metrics that we could investigate further to help us visualize a relationship between obesity rates and other variables that we could include in our analysis. For example, the number of fast-food restaurants per capita may seem significant if it is the only predictor; however, when other factors, such as access to grocery stores or poverty rates are considered, this may change. Therefore, to obtain this additional information, we gathered data from the food environment atlas. The food environment atlas is data from the United States Department of Agriculture. It includes variables such as store/restaurant proximity, food prices, food and nutrition assistance programs, and community characteristics. With the help of our sponsor, we came across the Robert Wood Johnson Foundation website. On this, we found the food environment index, which is an index of “factors that contribute to a healthy food environment, from 0 (worst) to 10 (best).” (1) The food environment index data that we found online was derived from all the variables in the food environment atlas to create a single measure indicative of a healthy food environment in a county.

After gathering the data on the food environment index, we plotted the indexes and obesity rates by Virginia County in Tableau.



*Figure 1: Food Environment Index by Virginia County. The index ranks Virginia counties from 0 - 10, where 0 is the least healthy food environment and 10 is the healthiest food environment.*



*Figure 2: Obesity Rate by Virginia County.*

At the onset of this statistical analysis, we conducted a stepwise regression on a subset of the variables that we believed to be a good representation of the variables that we aimed to include in our final model.

|  | Estimated Coefficients | P-value | Variance Inflation Factor |
| --- | --- | --- | --- |
| Intercept | 33.9205 | 0.0591 |  |
| Median Household Income 2015 | -0.0000546 | 0.2954 | 5.641907 |
| Fast Food 2016 | 0.0260559 | 0.2003 | 20.884387 |
| Snap Benefits 2017 | 0.0972906 | 0.6153 | 8.067405 |
| Vegetable Farms 2012 | 0.0275705 | 0.4894 | 1.323850 |
| Farmers Markets 2018 | -0.2947068 | 0.4833 | 2.950595 |
| WicPerCapita | 0.3488668 | 0.2178 | 3.846052 |
| Grocery Stores | -0.1655296 | 0.1757 | 20.987279 |
| Poverty Rate | 0.0175537 | 0.8302 | 7.010703 |

*Table 1: Multiple linear regression on the subset of data.*

Upon analyzing the AIC, these were the most significant variables in the model: WIC (Women, Infants, and Children) benefits redeemed per Capita, SNAP (Supplemental Nutrition Assistance Program) benefits redeemed per Capita, Number of Vegetable Farms, and Number of Farmers Markets. Although this was a rough regression model, it informed us of some issues that would arise later in our analysis when we would include other explanatory variables. These issues were primarily multicollinearity among the predictors (evidenced by high VIF factors), null values, and instances of nonlinearity.

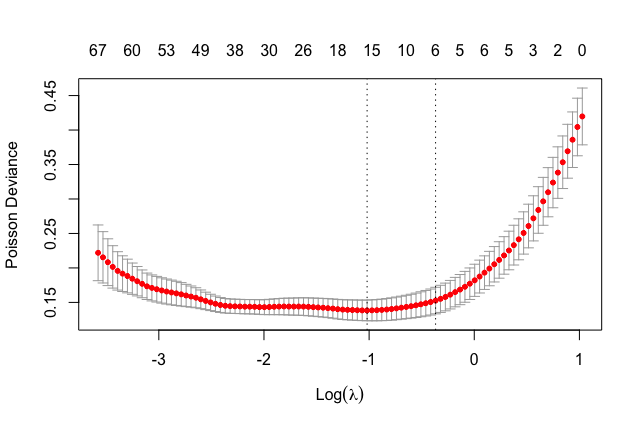
|  | Estimated Coefficients |
| --- | --- |
| Intercept | 29.548 |
| Snap Benefits 2017 | 0.198 |
| Vegetable Farms 2012 | 0.023 |
| Farmers Markets 2018 | -0.573 |
| WicPerCapita | 0.374 |

*Table 2: Most significant coefficients after stepwise regression that analyzed AIC.*

Before proceeding any further in our analysis with regards to the permit data, we needed to organize the restaurant permits by county. We aimed at extracting the number of restaurants, by permit type, for each county in Virginia. After combining many of the same permit types into more general categories, such as `Fast Food Restaurant,` and `Fast Food` to `Fast Food Restaurants,` we used the *groupby* function in Python to obtain the number of restaurants by permit type for each county and converted it to an excel file. We also took the restaurant permit types and calculated them per capita. Then, we combined this restaurant data with 133 columns of data from the food environment atlas. Information ranged from the level of access that low-income families had to stores, the prevalence of nutrition assistance programs, and the number of farmers' markets in a county. After finalizing our dataset, we met with GTA Sam Myren to discuss the challenges we faced with our data. Upon his suggestion, we met with Prof. Inyoung Kim the following week.

At the meeting with Prof. Inyoung Kim from the statistics department, we discussed our data, the scope of our project, our objectives (to understand the relationship between restaurants and obesity), and the issues with multicollinearity and null values. She provided us with a comprehensive lasso regression technique using the *glmnet* library in R that allowed us to select the most important variables from the dataset.

To do this, we first ran a cross-validation function of the lasso regression using *glmnet*. This provided us with the optimal lambda tuning parameter to use for the lasso variable selection. The lambda is a penalty term that minimizes the cross-validation prediction error rate.

****

*Figure 3: The optimal lambda tuning parameter is the minimum lambda*

After running the *glmnet* function with the minimum lambda as a tuning parameter, we obtained the following variables. The highlighted ones are the most significant.

|  | Estimate | Std. Error | t - value | p - value |
| --- | --- | --- | --- | --- |
| (Intercept) | 3.550500 | 0.004526 | 784.470 | <2e-16 |
| Median Household Income 2015 | -0.013763 | 0.011236 | -1.225 | 0.2231 |
| Poverty Rate | 0.024255 | 0.009856 | 2.461 | 0.0153 |
| Black, low access to store (%), 2015 | 0.011473 | 0.005856 | 1.959 | 0.0525 |
| Hawaiian or Pacific Islander, low access to store (%), 2015 | 0.010116 | 0.004916 | 2.058 | 0.0419 |
| Students eligible for free lunch (%), 2010 | 0.015232 | 0.009293 | 1.639 | 0.1039 |
| Farms with direct sales (%), 2012 | -0.013494 | 0.005501 | -2.453 | 0.0157 |
| Direct farm sales per capita, 2012 | -0.002350 | 0.005806 | -0.405 | 0.6865 |
| Farmers' markets that report accepting SNAP, 2018 | -0.010218 | 0.006871 | -1.487 | 0.1397 |
| Farmers' markets that report accepting credit cards, 2018 | -0.016266 | 0.007778 | -2.091 | 0.0387 |
| Orchard farms, 2007 | -0.006300 | 0.006176 | -1.020 | 0.3098 |
| Agritourism receipts, 2012 | -0.005610 | 0.006299 | -0.891 | 0.3749 |
| Adult diabetes rate, 2008 | 0.002106 | 0.009897 | 0.213 | 0.8318 |
| Adult diabetes rate, 2013 | 0.023209 | 0.009205 | 2.521 | 0.0130 |
| Caterer Per Capita | -0.011463 | 0.005634 | -2.035 | 0.0442 |
| Summer. Food Dispensing Per Capita | 0.013226 | 0.005989 | 2.209 | 0.0292 |

*Table 3: The initial lasso covariate selection*

After having discovered these coefficients we met with our sponsor to do a “common sense check” to ensure that we are not blindly following our output, rather that we are taking into account any additional information that would lead us to better understand these results. Our sponsor, Ms. Walmsley, recommended that we further clean our data to only include those restaurant permits that are currently ‘permitted’, meaning that their permits are still active. This was done as that data initially contained permits that had expired. After cleaning the data and running the lasso we obtained the following covariates.

|  | Estimate | Std. Error | t - value | p - value |
| --- | --- | --- | --- | --- |
| (Intercept) | 3.5505 | 0.0045 | 782.499 | <2e-16 |
| Median household income, 2015 | -0.0100 | 0.0114 | -0.885 | 0.3780 |
| Poverty rate | 0.0289 | 0.0102 | 2.820 | 0.0056 |
| Black, low access to store (%), 2015 | 0.0138 | 0.0058 | 2.373 | 0.0192 |
| Hawaiian or Pacific Islander, low access to store (%), 2015 | 0.01108 | 0.0049 | 2.260 | 0.0257 |
| SNAP participants (% pop), 2012 | 0.0091 | 0.0048 | 1.887 | 0.0617 |
| Students eligible for free lunch (%), 2010 | 0.0124 | 0.0142 | 0.868 | 0.3869 |
| Students eligible for free lunch (%), 2015 | 0.0026 | 0.0147 | 0.175 | 0.8616 |
| Farms with direct sales (%), 2012 | -0.0114 | 0.0054 | -2.099 | 0.0380 |
| Direct farm sales per capita, 2012 | -0.0071 | 0.0051 | -1.386 | 0.1686 |
| Farmers' markets that report accepting SNAP, 2018 | -0.010908 | 0.006979 | -1.563 | 0.12078 |
| Farmers' markets that report accepting credit cards, 2018 | -0.0153 | 0.0078 | -1.963 | 0.0521 |
| Orchard farms, 2007 | -0.0018 | 0.0060 | -0.291 | 0.7717 |
| Agritourism receipts, 2012 | -0.0079 | 0.0063 | -1.261 | 0.2097 |
| Adult diabetes rate, 2008 | 0.0066 | 0.0099 | 0.666 | 0.5069 |
| Adult diabetes rate, 2013 | 0.0249 | 0.0092 | 2.713 | 0.0076 |
| Grocery Store Food Service/1,000 pop | -0.0090 | 0.0046 | -1.944 | 0.0543 |

*Table 4: After running lasso on the permitted data.*

As can be seen here, out of all the restaurant data, only grocery stores are a slight predictor of obesity.

Furthermore, upon the advice of our client, we ran a separate lasso covariate selection on a subset of our data that included all local farm variables. For example, variables such as ‘Orchard farms, 2007’ and ‘Farms with Direct Sales (%) 2012’ are relevant in the model. From this, we selected only the most significant farm variables to include in the future. After running regression on all of the local farm variables (65 variables), 6 were deemed to have non-zero coefficients. They are listed below:

* Berry acres, 2012
* Berry acres/1,000 pop, 2012
* Farmers' markets that report selling fruit & vegetables (%), 2018
* Farmers' markets/1,000 pop, 2013
* Farmers' markets/1,000 pop, 2018
* Farmers' markets, 2013.

|  | Estimate | Std. Error | t - value | p - value |
| --- | --- | --- | --- | --- |
| (Intercept) | 3.550500 | 0.004702 | 755.177 | < 2e-16 |
| Farmers' markets, 2013 | -0.005844 | 0.008739 | -0.669 | 0.50494 |
| MedianHouseholdIncome2015 | -0.019228 | 0.011300 | -1.702 | 0.09149 |
| PovertyRate | 0.027604 | 0.010196 | 2.707 | 0.00780 |
| Black, low access to store (%), 2015 | 0.014752 | 0.006075 | 2.429 | 0.01668 |
| SNAP redemptions/SNAP authorized stores, 2012 | 0.010621 | 0.005003 | 2.123 | 0.03587 |
| Students eligible for free lunch (%), 2010 | 0.012593 | 0.009615 | 1.310 | 0.19285 |
| Agritourism receipts, 2012 | -0.006321 | 0.005491 | -1.151 | 0.25202 |
| Farm to school program, 2013 | -0.003898 | 0.004998 | -0.780 | 0.43704 |
| Adult diabetes rate, 2008 | 0.001272 | 0.010355 | 0.123 | 0.90241 |
| Adult diabetes rate, 2013 | 0.025814 | 0.009326 | 2.768 | 0.00656 |
| Recreation & fitness facilities, 2016 | -0.014449 | 0.009181 | -1.574 | 0.11823 |
| Caterer per Capita | -0.013625 | 0.005076 | -2.684 | 0.00832 |
| Summer Food. Dispensing. Per Capita | 0.008791 | 0.006123 | 1.436 | 0.15377 |
| Hawaiian or Pacific Islander, low access to store (%), 2015 | 0.010984 | 0.005134 | 2.140 | 0.03445 |

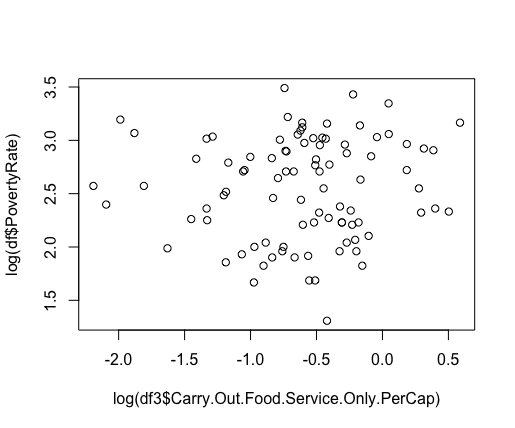
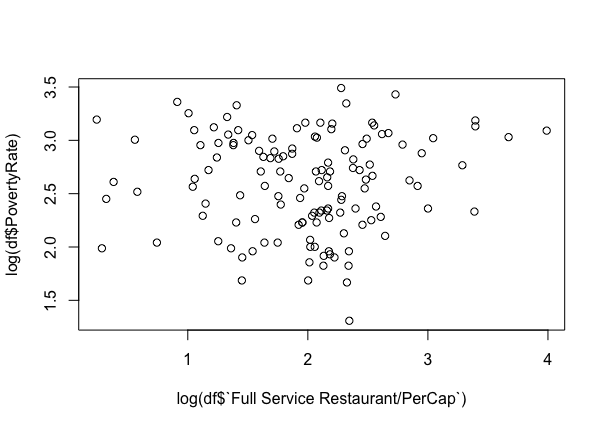
*Table 5: Covariates after running lasso with the reduced farm variables*

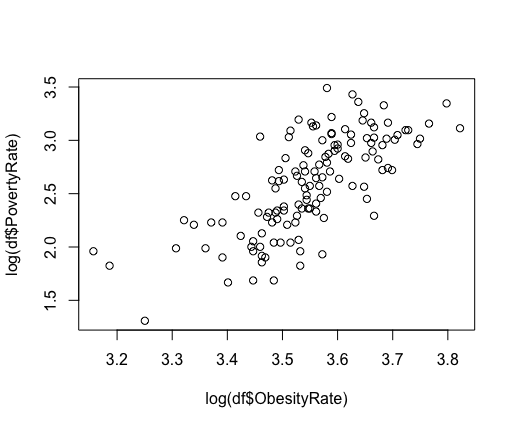
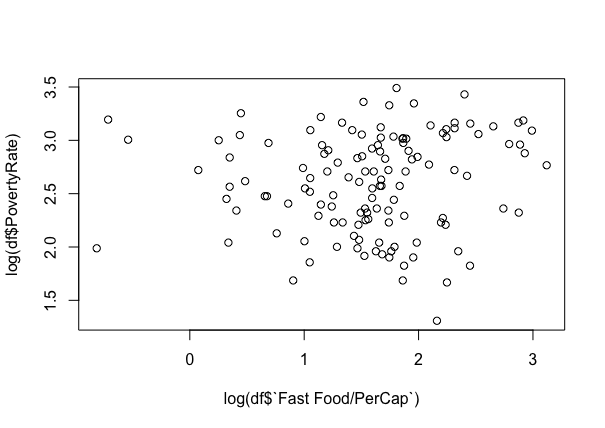
As can be seen once again, there is only one restaurant permit category that is in the model. Summer food dispensing, as these are the permits given to those that serve meals to people who are impoverished.

Through all these analyses, the poverty rate seems to consistently be one of the most significant factors in obesity. Due to this, we ran a final lasso on poverty as the regressor. From this, a few more restaurant permit types were deemed as having a relationship with poverty. They were:

* Adult Care Home Food Service/1,000 pop
* Carry Out Food Service Only/1,000 pop
* Full Service Restaurant/1,000 pop
* Fast Food/1,000 pop
* Grocery Store Food Service/1,000 pop

Although there is a clear relationship between poverty and obesity, none of these variables were significant enough to show a clear relationship when plotted.

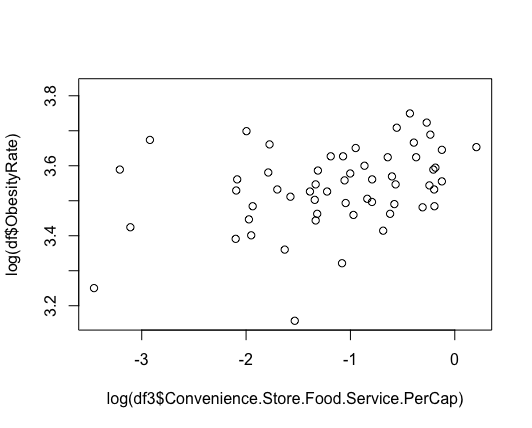
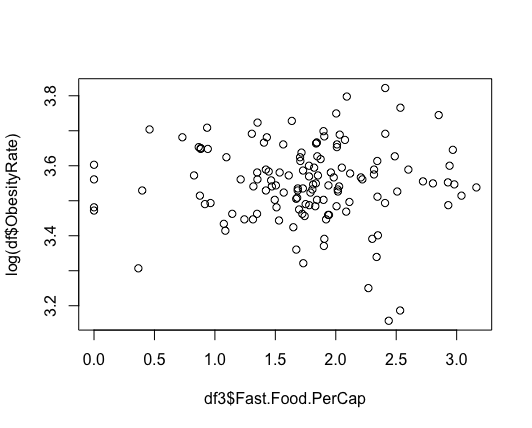




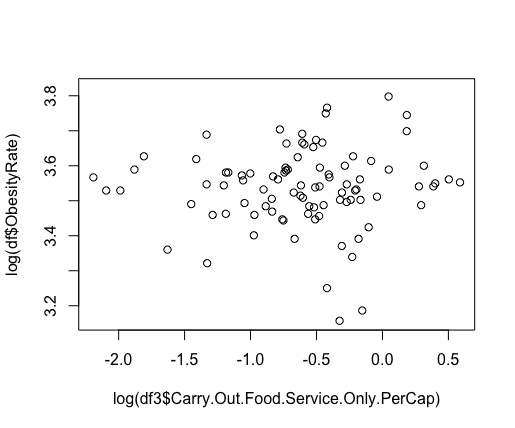
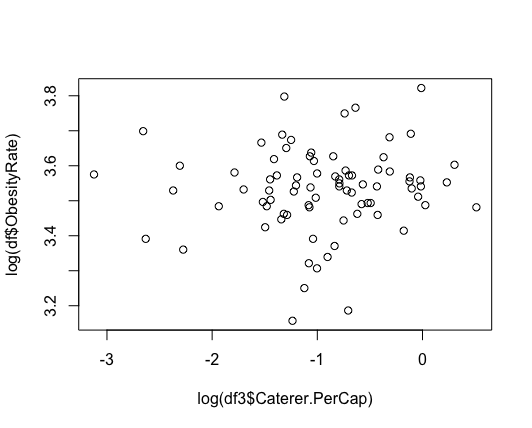
*Graph 1: Poverty Rate vs Carry Out Permit Graph 2: Poverty Rate vs Full Service Permit*

*Graph 3: Poverty Rate vs Fast Food Permit Graph 4: Poverty Rate vs Obesity Rate*

Furthermore, the relationship between restaurant permit types and obesity is also quite poor.



*Graph 5: Obesity Rate vs Convenience Store Permit Graph 6: Obesity Rate vs Fast Food Permit*



*Graph 7: Obesity Rate vs Carry Out Food Permit Graph 8: Obesity Rate vs Carry Out Food Permit*

After multiple efforts to try and evaluate the relationship between restaurant types and obesity rate, we cannot conclusively say that restaurant permits in Virginia counties have an effect on rates of obesity and poverty. The link is weak at best with the data that we used.

**6.2 Second Analysis: Gradient Boost**  
As the primary analysis for our project was heavily focused on Lasso Regression, we identified key factors that impact obesity rates, including poverty rates, access to healthy food options, and diabetes prevalence. We leveraged the selected features obtained from it to develop a Gradient boosting model to create a predictive model and improve the generalization and performance. The model was trained on significant features that Lasso Regression identified as strongly linked to obesity. Following a thorough evaluation of the model's performance, we then proceeded to pinpoint the variables that exhibited a robust association with obesity rates.

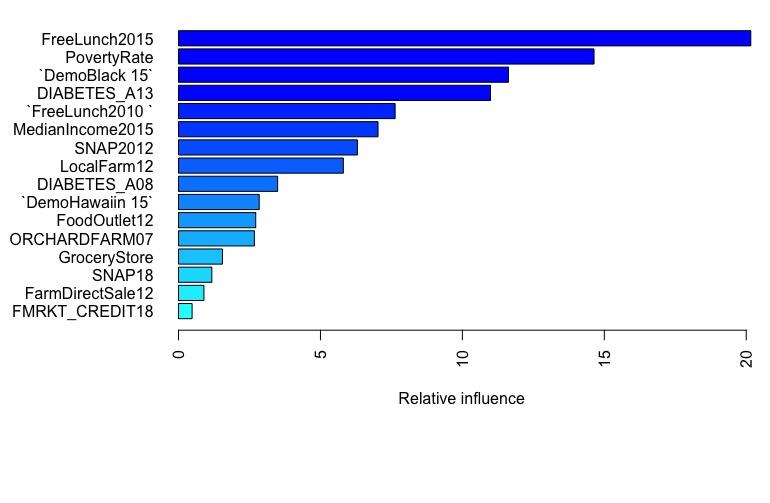
| ***Variables*** | ***Relative influence*** |
| --- | --- |
| *Free Lunch 2015* | 20.1605848 |
| *Poverty Rate* | 14.6380 |
| *Access to Store Blacks((Demographics))* | 11.6238314 |
| *Diabetes Adults 2013* | 10.9888570 |
| *MedianHouseholdIncome2015* | 7.0283649 |
| *SNAP 2012* | 6.3025350 |
| *Local Farm 2012* | 5.8056659 |
| *Adult Diabetes 2008* | 3.4907059 |
| *Hawaiian or Pacific Islander (Demographics)* | 2.8420482 |
| *Orchard Farms (Farm and Crops)* | 2.6731144 |
| *Grocery Store Food Service* | 1.5481816 |
| *\_SNAP Year 18* | 1.1738 |
| *Direct Sales Farm* | 0.8955 |
| *Farmer’s Market* | 0.478 |

*Table 6: Gradient Boost Analysis Results*

This result shows the relative influence of each variable in a Gradient Boosting model that was trained to predict the model. The relative influence column indicates the importance of each variable in the predicted outcomes, where a higher value indicates a more substantial impact on the response variable. In contrast, a lower value shows a weaker impact.

From Table 6, based on the relative influence, access to free lunch in 2015 has the strongest relative influence, which accounts for 20.16% of the variance in obesity rates. It is followed by poverty rates with 14.64% and access to stores for black demographics with 11.6238314 %. This suggests that these variables were the most influenceable predictors of the outcome in the model.

Additionally, we also observe that variables such as “Diabetes Adults 2013”, “MedianHouseholdIncome2015”, “SNAP 2012”, and “Local Farm 2012” have moderate relative importance, indicating that they may have some impact on obesity rates. The remaining variables had even less relative influence on obesity rates.

****

*Figure 4: Visualizing Variable Importance*

Figure 4 is the visual representation of the relative influence of each variable. The graphs represent the importance of each variable in predicting obesity rates and help identify the variables that should be prioritized while addressing Obesity.

| **Model** | **Root Mean Square Error (RMSE)** |
| --- | --- |
| *Gradient Boosting* | 2.329 |

*Table 7: RMSE Table for Gradient Boost*

Table 7 shows that the selected variables are highly relevant in explaining the outcome and can further improve the model’s performance. We have computed the root mean square error (RMSE) value of 2.329, indicating that the model had an average error of 2.33 in predicting obesity rates.

**6.3 Transformations**

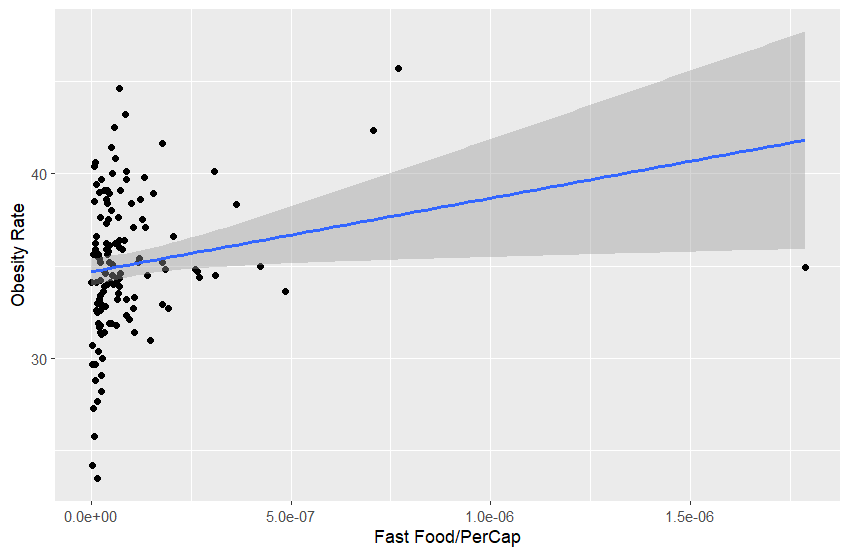
When fitting models to our data, we found that the data was mostly nonlinear. This meant that the models we were using were not accurately capturing the relationship between the variables. To address this issue, we tried several transformations to make the data more linear. These included log, square root, power, and Generalized Additive Model (GAM) transformations. We looked at these transformations before and after our lasso regression.

Pre-Lasso:

After trying multiple different transformations, we found that the log transformation appeared to be the most effective. We determined this by looking at the R-squared value. The R-squared value measures how well the model fits the data.

However, we noticed that even with the log transformation, the R-squared value was often less than 0.5. This raised concerns about the accuracy of our model. Despite this, visual inspection showed that the data was better with the transformation than without it. The transformed data appeared to follow a more linear relationship, which made it easier to model.

When we created a model without applying any transformation to the data, We have the following figure:



*Figure 5: Relationship between Fast Food and Obesity Rate Pre-Transformation*

From Figure 5, we can see that most of the values are clustered at the beginning, with some outliers. This means that the majority of the data points are concentrated in a small range, while a few data points are far away from the others. This can make it challenging to fit a good linear model to the data because the outliers can have a large influence on the model – as you can see from the model line fit.

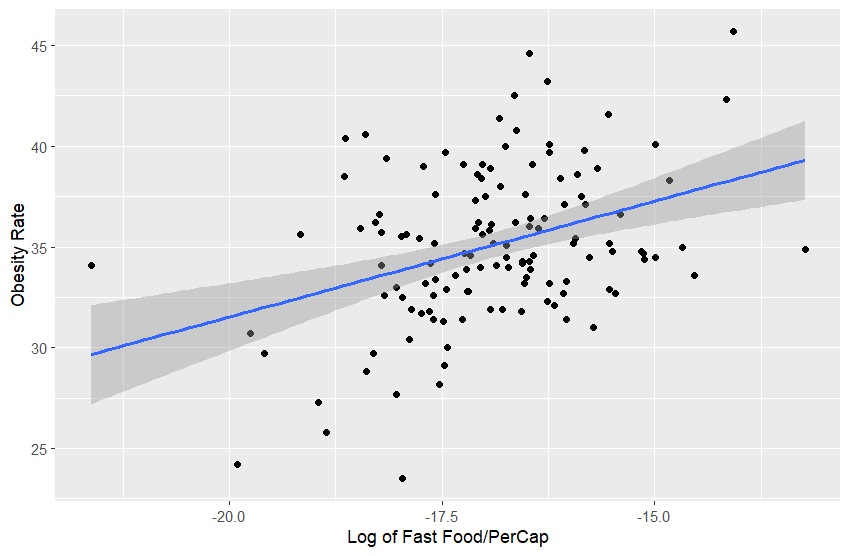
The model summary provides information about the model we have fitted to the data. It includes statistics such as the R-squared value and the p-values for the coefficients. These statistics can help us assess how well the model fits the data and whether the relationships between the variables are statistically significant. In this case, we focus on the R-squared value to evaluate the model’s fit.

| Multiple R - squared: | Adjusted |
| --- | --- |
| 0.1179 | 0.1095 |

*Table 8: Model Summary from Pre-Transformation Model*

From the model summary Table 8, we can see that the R-squared value is 0.11, which means that only 11% of the variation in the data can be explained by the model. This is a very low value, indicating that the linear model does not fit the data well. It suggests that a linear model may not be the best choice for this data.

Now let’s take a look at the model after applying a transformation. We’ll start by examining the graph that shows the effect of taking the log of the Fast Food/Per Cap variable.



*Figure 6: Relationship between Fast Food and Obesity Rate Post-Transformation*

From Figure 6, we can see that the points are more spread out and follow a slightly positive trend. This indicates that the transformation has improved the linearity of the data. As a result, we can fit a better linear model to the data. Compared to the previous graph, where the points were clustered at the beginning, and there was no clear trend, this graph shows a much clearer relationship between the variables.

After fitting a linear model, from the model summary, we can see the R-squared value, which measures how well the model fits the data:

| Multiple R - squared: | Adjusted |
| --- | --- |
| 0.1375 | 0.1307 |

*Table 9: Model Summary from Post-Transformation Model*

We see from Table 9 that there was a small increase in the R-squared value from 0.11 (in Table 8) to 0.13 (Table 9); however, both values are still relatively low, indicating that the linear model does not explain a large proportion of the variation in the data.

Despite the relatively low R-squared value, it is still better to apply the transformation because it makes the relationship between the variables more linear. This can be seen from the graph, where the points are more spread out and follow a clearer trend. Overall, we see a much better graph and data after applying the transformations. We then applied these transformations when we created our Lasso model, which gave us much better results.

Post-Lasso:

We tried multiple transformations again for this case as well, and we ended up seeing the highest R-squared value with the log transformation. This new R-squared gave us way better results as seen in Table 10.

| Multiple R - squared: | Adjusted |
| --- | --- |
| 0.7949 | 0.7723 |

*Table 10: Model Summary from Post-Transformation Model*

The high R-squared value of 0.7949 in Table 10 indicates that the Lasso model with the log transformation explains a large proportion of the variation in the data. The Adjusted R-squared value of 0.7723 also confirms that the model fits the data well, even after accounting for the number of variables in the model.

Overall, the combination of Lasso regression and the log transformation has significantly improved the accuracy of the model and provided valuable insights into the relationship between the variables.

**6.4 Exploration of New Response Variable**

Although obesity is a good response to measure when looking at public health outcomes, we wanted to also look into another response variable to make sure there wasn’t something being missed. We decided to look into diabetes as the new response. Diabetes is a disease where your body can no longer process insulin and glucose correctly. Type two diabetes and obesity often occur together, both as a result of more fat being stored than needed. Therefore, we thought diabetes would be a good metric to use because of its relation to obesity.

In order to explore this new response variable, we removed obesity from the dataset and moved the percent diabetes from the explanatory to the response. Once the dataset was now updated, we did the same analysis as before: use glmnet() to fit a lasso model, get the lambda values, run it again, and analyze the fit. The results from the new response were underwhelming. The new results indicated that the total number of restaurants and the number of grocery stores are the main factors that affect the percentage of diabetes in a county. Although these results make sense, there was not enough statistical significance to use them outright in our study. In conclusion, we will continue to use the obesity rate as our response variable and main determination for health in a county.

**7 Limitations on Deliverables**

Due to the weak relationship between restaurant permits and obesity rates, we cannot conclusively state that restaurants have an effect on obesity rates at the level of Virginia counties. This may be due to many missing values that were imputed and generally a low range of restaurant permit types. We used the data that we had under the assumption that it is representative of the complete set of permits in each Virginia county. However, after reviewing the data, we saw very low numbers of permits in some permit categories and believe that a more comprehensive set of permits structured by county could be used in the future. Furthermore, the data also had roughly 9000 rows of permits that were not assigned in a county or city. This also could be a factor that led to distorted results and low numbers. Much of the data that we gathered from the food environment atlas was full, and therefore represented a more complete sample than the restaurant data we had.

**8 Interpretation of Results for Clients**

Even after determining that log transforms were the most efficient, we plotted the permitted restaurant data that was revealed as the most significant from the lasso regression. From plotting these restaurant permit types against obesity, we are able to clearly see that there is not a significant relationship between restaurant type and obesity. This is likely due to the low number of observations that we had and due to many missing values in our data set. From this, we are also able to determine that other socioeconomic factors are more predictive of rates of obesity across Virginia counties, such as having low access to stores, nutritional assistance programs, and poverty rates. The VDH could focus more efforts on areas where these factors are prevalent and introduce measures that could promote healthier eating while taking into account the financial status of these areas. As the poverty rate seems to be a leading factor of obesity and given our knowledge of this domain, we conducted a lasso regression on Poverty Rate as the response. The new covariates changed slightly. However, when plotting the log graphs for permit types with obesity and poverty, we see a very weak relationship.

Due to multiple weak relationships between restaurant types and obesity rate. We cannot conclusively determine a relationship between obesity and poverty. More investigation on poverty as a whole rather than restaurants should be conducted in the future.

**9 Team Roles**

**Mir Abdullah**

Technical Contribution: Worked on doing data transformations. Cleaned data from NaN, Inf, and non-positive values. Transformed non-linear data using log, square roots, GAM, and power transformations.

Non-Technical Contribution: Contributed to Tech Memos, Midterm Report, Tools and Techniques, and Final Report. Presented midterm presentation and planned on presenting on the final presentation.

**Connor Pepin**

Technical Contribution: Connor helped in the final stages of data cleaning, programmed the initial Lasso regression, and explored the use of diabetes as a response variable.

Non-Technical Contribution: Connor contributed to Tech Memos, Midterm Presentation, and the Final Report. Attended a meeting with Dr. Kim from the Stat department, and presented the Midterm Presentation.

**Nahom Kifetew**

Technical Contribution: Nahom worked on finding an effective way for replacing NA values, splitting the data into training and testing sets to perform Lasso, Ridge, and ElasticNet regression in R then comparing the performance of these three models. Found extra data and helped with cleaning and preparing the data.

Non-Technical Contribution: Nahom contributed to all 4 Tech memos, the Midterm presentation, and the Final report. Additionally, discussed problems and solutions with Dr. Kim from the Statistics department.

**Gaurav Shah**

Technical Contribution: Gathered external data for regression analysis. Cleaned and prepared the data into excel files. Explored principal component analysis, conducted lasso covariate selection, stepwise regression, linear regression, and merged all external data with VDH permit data.

Non-Technical Contribution: Scheduled meetings with Dr. Kim from the statistics department and Sam Myren to discuss issues that arose during the data analysis phase and for help running the lasso covariate selection technique. Contributed to all tech memos, midterm presentations, tools and techniques, and elevator speeches. Presented in the tools and techniques.

**Rubina Joshi**

​​Technical Contribution: Rubina made significant technical contributions by preparing and cleaning data in Python for analysis. She performed gradient boost analysis on selective features and trained and tested the data. She also explored covariate selection techniques and multiple regression.

Non-Technical: Rubina also contributed to all tech memos, contributed slides for the midterm presentation, presented and contributed to the Tools and Techniques Presentation, and helped with the final report. She attended a meeting with Samuel Myren to discuss and brainstorm project ideas. She also ensured that all deadlines and project scope were met.

**10 Conclusions and Future Work**

In conclusion, our analysis showed that permitted restaurants and obesity rates in Virginia counties have no significant relationship between the restaurant types and health outcomes such as obesity. However, poverty rates were found to be a significant factor in the increment in obesity rates. Therefore, VDH or stakeholders should focus more on the areas with high poverty rates and invest in programs to promote nutritious food in these areas.

With the help of statistical approaches such as lasso regression, we identified the key factors that contributed to health outcomes such as obesity rate; in the future, our model can be utilized to develop a customizable predictive model for specific geographical locations. The statistical approach used in our research, such as the gradient boost and lasso regression, has the potential to predict severe health implications, which includes chronic disease such as obesity, and can serve as a powerful tool in society. Although we acknowledge that our research has some limitations due to limited resources, the findings of this research can offer valuable information to prevent and develop solutions to these health concerns. For instance, if certain regions have high obesity rates, with the help of research like this, policymakers can invest by providing permits for healthy food options in those areas. Similarly, areas with inadequate food options such as food deserts should be provided with more nutritious food. In addition, insights gained from the predictive model can help educate the communities about the importance of healthy food and promote the well-being of preventive health concerns such as obesity and diabetes.

Works Cited

[1] Block, Jason P., Richard A. Scribner, and Karen B. DeSalvo. "Fast food, race/ethnicity, and income: a geographic analysis." American journal of preventive medicine 27.3 (2004): 211-217.

[2] Gunderson, Craig, and James P Ziliak. “Food Insecurity and Health Outcomes | Health Affairs.” Health Affairs, https://www.healthaffairs.org/doi/10.1377/hlthaff.2015.0645.

[3] West, Samuel J., et al. “Comparing Forms of Neighborhood Instability as Predictors of Violence in Richmond, VA.” PLOS ONE, Public Library of Science, https://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0273718.

[4] *Food environment index*. County Health Rankings & Roadmaps. (n.d.). Retrieved April 13, 2023, from/https://www.countyhealthrankings.org/explore-health-rankings/county-health-rankings-model/health-factors/health-behaviors/diet-and-exercise/food-environment-index?year=2023

VIRGINIA TECH HONOR CODE

“We have neither given nor received unauthorized assistance on this assignment.”

