Socioeconomic Effects on Health and Well-being Using U.S. County Level Data

Lily Shaw
Angelo State University
Dept of Computer Science
mshaw7@angelo.edu

Steven Womack
Angelo State University
Dept of Computer Science
swomack5@angelo.edu

Minh Le
Angelo State University
Dept of Computer Science
hle2@angelo.edu

ABSTRACT

Using publicly available, county-level data on social demographics, behaviors and health outcomes, we explore select questions and correlations between factors. The questions focus on the possible correlations between socioeconomic factors, race/ethnicity, and health outcomes for counties in which data was provided. Findings are presented with correlational matrices and heatmaps of the US counties for visualizations.

We attempt to answer questions like: Do areas with higher negative economic factors have worse health outcomes than areas with better economic factors? Is there a correlation between educational attainment and health outcomes? How does income inequality, childhood poverty, affect health behaviors such as smoking, alcohol intake, etc.? How do unemployment and lower insurance rates affect mental health versus physical health outcomes? Do areas with higher ratios of primary care physicians and mental health providers have larger amounts of teen births? As unemployment and educational attainment rise or fall, do health outcomes change as a result? How does the physical environment of an area (county) affect the health outcomes of the population?

1. INTRODUCTION

In this paper we will discuss our analysis of the county level health data provided by County Health Rankings. Discussed will be the methodology and criteria for input selection on our exploratory data analysis and predictive modeling. Methodology for our analysis will be included in section 4 of the paper. Most of the work and research for this project was done using Python and the Pandas library on both Google Colab and Jupyter Notebooks. Our notebook will be available for viewing with a link in the citations for transparency.

Findings for exploratory data analysis will be displayed and analyzed through the use of standard Pearson correlations, maximums/minimums, and heatmaps to discuss possible findings. Other attempted methods will be included in a smaller capacity, such as ranked correlations, etc. Findings for the predictive modeling will feature primarily linear regressions and decision trees. Other methods discussed and attempted include polynomial regression, quadratic regression, etc.

2. RELATED WORK

Related works have been done and published using the same county level health data provided by County Health Rankings. Many papers seek to review the status of health and wellbeing in the U.S. and possible relationships between those outcomes and input factors. Nearly all factors have been explored by some group at one point in time.

Due to our relationship with United Health Group, whom we have coordinated with for this project, we can assume that there are multiple projects related to our work and focusing on the same research questions and analysis as our project team.

Papers have focused on analyzing the data and any correlations between factors, but also on the data collection itself. Due to the nature of the data being collected, it is frequently imperfect. Problems of underrepresentation in poorer areas of communities is frequently seen throughout the county health rankings. Many places do not collect certain racial or socioeconomic data. The reasoning for the lack of collection can be speculated on, yet the lack of data still exists. As such, studies have chosen to primarily focus on those categories/factors that have a 90-95 percent completion rate.

3. DATASET

For this study and project, we used County Health Rankings. The data is collected and collated by the University of Wisconsin. The data set is publicly available and has been collected yearly since roughly 2004. It includes all +3000 counties in the United States with the majority reporting on most categories.

The dataset includes statistics on the health, well-being, economics, and social factors of a county. Many of these factors provide racial and ethnic

backgrounds for the factors chosen and collected. The primary driver for the collection of data was to initiate competition between regions and counties about improving the lives of their citizens. This data is very comprehensive in the collection of various types of factors.

Below in Fig 1, we can see how the factors of a county relate to the outcomes. These estimates and percentages show us how various factors can contribute to the overall outcome of health and wellbeing. While these percentages are only estimates provided by the County Health Rankings analysis of data since 2004, they assisted our research group in selecting factors for our project.

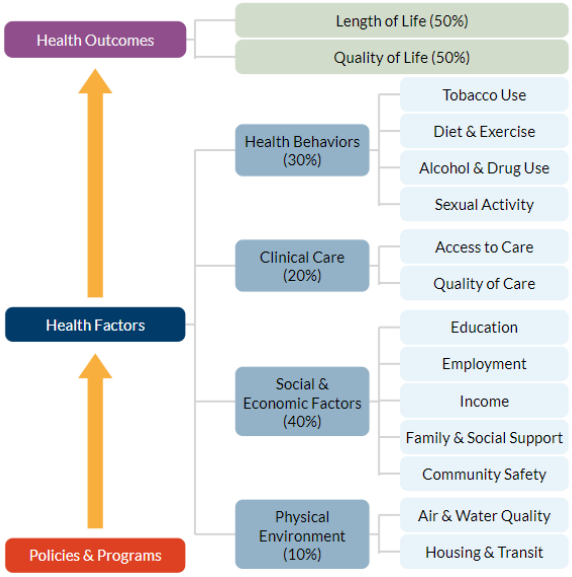

Fig 1 Breakdown of Factors in County Health Rankings Dataset

However, the dataset is limited by the reporting of the county officials in the collection methods. Some counties experience underreporting of various social and health factors due to the collection methods employed. This limit creates a gap in the analysis of the data and certain possible correlations.

4. METHODOLOGY

4.1 FACTOR SELECTION

Factors used in our analysis were chosen based on 3 criteria primarily. Relevance to question, relevance to other factors, and completion of data.

Relevance to the questions was a primary factor in selecting categories from the project to be analyzed and studied. We primarily reviewed the documentation found on the County Health Rankings website to determine which classification a particular category would fall under. Due to our questions, we focused primarily on social determinants such as education and economic determinants such as unemployment.

Completion of data was of utmost concern for our modeling an exploratory data analysis factor selection process. While in exploratory data analysis, incomplete data was not as influential in our results. In our predictive modeling section, incomplete data could result in a very large skew in a particular direction. Due to this, we tried to select only those factors that had a 5-10 percent incompletion rate. This led to certain factors being weighed and relied upon more heavily than others simply due to the high completion rates of the data.

Relevance to other factors was a secondary concern in our selection process for categories. Whilst important, we wanted to make sure each question we looked to answer had some intuitive relationship with the outcomes analyzed. In our research, ensuring that multiple perspectives of the same outcomes were being seen. By not limiting ourselves to only certain factors and being conscious of our use of certain factors we attempted to provide a more solid understanding and picture of our conclusions.

4.2 NOTEBOOK CREATION AND LIBRARIES

For this project we primarily used Python 3 on Google Colab. We chose Python in this format due to the size and breadth of the open source libraries available to us and the ease in sharing the source code we used to process the data. While we could have chosen another language such as R for our analysis, for our needs Python 3 seemed most appropriate. Google Colab offered us an easy way to share our code with team members and our faculty advisor.

Libraries used for data analysis include Pandas and Numpy. Both offered built-in functionality and data processing capabilities, such as finding max/min, outliers, and removing incomplete data. The open source and thorough documentation on these libraries greatly helped our analyses.

Libraries used for visualizations included Seaborne and Altair. Both have built in visualization packages for histograms and charts. The library used varies depending on the needs of the visualization. Many of our box graphs and line charts were done in Seaborne, while the majority of our U.S. heatmaps were done in Altair.

Libraries used for predictive modeling were XGBoost, Numpy, and SciKitLearn. XGBoost was used for our exploration into decision trees using extreme gradient boosting. Numpy was used to rescale our data to minimize any issues with our MSE (Mean Squared Error) and our MAE (Mean Average Error) due to scaling issues. SciKitLearn was used for modeling and creating our linear regression models.

4.3 DATA CLEANING

In our project, data cleaning was achieved using a few different methods.

Due to the highly processed nature of our dataset from County Health Rankings, misentered data was nearly nonexistent in our data. With that in mind, our only obstacle for data cleaning was dropping unentered fields. By handling this possibility in our factor selection, we minimized the impact of a possible skew due to our data dropping practices. In our selection process we made sure that we never had more than a 10 percent incompletion rate for any factor that we selected.

In the predictive modeling section of our project. We used the above methods and boxplots to eliminate outliers. Box plots such as the one pictured below. Eliminating outliers enabled us to prevent an overfit problem. This problem occurs when outliers adversely affect the MSE and the MAE, causing the model to not accurately represent all of the data. Therefore overrepresenting certain outlier data points as opposed to the mean data point in our model.

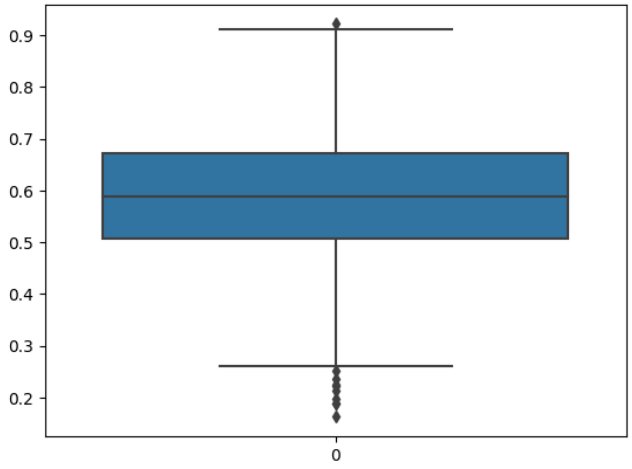

Fig 2. Example of box plot used to eliminate outliers for predictive modeling.

4.4 DATA ANALYSIS METHODS

For our exploratory data analysis, we primarily used basic methods of statistical analysis. We used standard Pearson correlations for most of our numerical correlations, with some attempts made in exploring ranked correlations as an alternative. We looked for maximums and minimums of certain variables and looked for overlaps of any being consistently ranked in the top 10. From there we took deeper looks into what other factors the county had that were outliers.

To identify possible trends visually and test out other methods. Scatterplots were used to visualize our data and allow us to observe any potential trends that would be fruitful in our future exploration of the dataset.

Overall, nationwide trends were viewed and were found to be lacking in numerical correlations. Because of this we used visualizations in the form of U.S. heatmaps to visualize the trends. These heatmaps allowed us to get a regional view of the United States. A view that states and counties were not reflecting accurately.

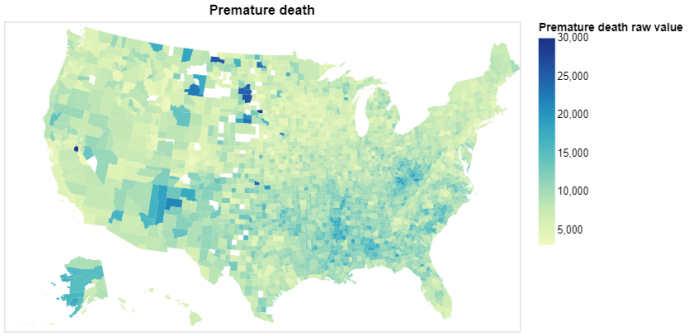

Fig 3. Example of heatmap using the United States Premature Death Rate with county level precision

4.5 PREDICTIVE MODELING METHODS

In the predictive modeling section we primarily used 2 methods in our analysis. We used linear regression models and explored working with XGBoost for gradient boosted decision trees.

For our linear regression models we frequently had to rescale the data, putting it on scales appropriate for predictive modeling. These scaled models can then be applied to the original data set and visualized on a scatterplot for visual confirmation that there is no obvious skew. When it was appropriate we also took the log of both sides to make the models fit a linear regression model better.

During decision trees, we applied most of the same methods listed above in the data cleaning section. However for this many of the functions used were built in functions in the XGBoost library. These built in functions allowed us to quickly and efficiently generate decision trees for certain parts of our data set. Many of these predictive models generated, gave more varied results when compared to the linear regression models.

5 EVALUATION

5.1 EXPLORATORY DATA ANALYSIS OVERALL

For our exploratory analysis of the dataset, we primarily focused on using tried and tested methods of evaluation for our project. In that effort, we began by using scatter plots to visualize the possible correlations and allow us to take deeper dives into certain areas. Our analysis of economic effects on health outcomes is a prime example of our findings.

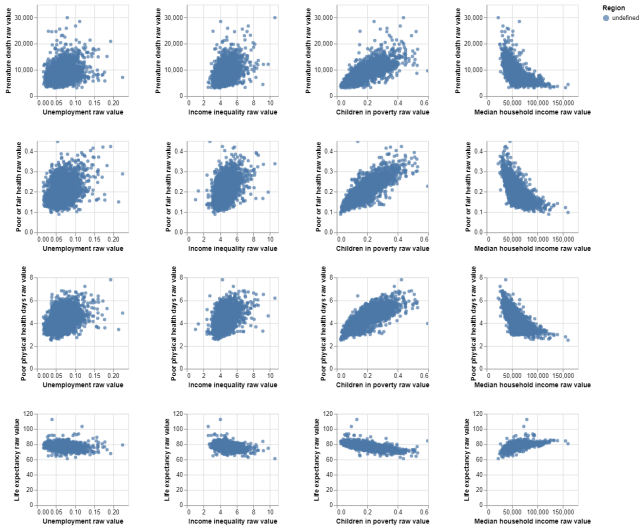

Fig 4. Scatter plots of economic data and health outcomes.

In these scatter plots, we can see some clear trends that seem to fit a correlation between certain factors. While certain factors have little to no correlation, the visualization allowed us to more clearly focus on certain factors for further analysis.

In spite of certain visual correlations we inferred, many of our numerical correlations gave mixed results. We reasoned that the primary reason for these mixed results was simply the size of the dataset. With +3000 data points, the numerical data was misleadingly low for many of the questions we sought to answer. To that end, we continued exploring the dataset using different methods.

Focusing on the same question, we searched for maximums and minimums to see if the same counties routinely appeared in the top. Frequently, in the case of this question, we found that counties appears multiple times in the upper quartile or above of multiple negative economic and health factors. As an example, Campbell County in South Dakota has the highest premature death rate in the nation. It also has the highest income inequality and the lowest median household income.

The finding of these anomalies led us to believe that a more regionalized correlation would be much stronger than any overall, national correlation. To that end, we focused on creating heat maps of the United States. These heatmaps enabled us to see much more highly regionalized correlations that get lost in the overall, national picture. These regionalized correlations permeate state borders, so we kept the county level precision for our project. Further breakdowns inside of counties would probably reveal a neighborhood by neighborhood correlation. As the county level collections are simply too broad for large, heterogeneous counties. Many counties have both an urban and rural divide in the United States, very few are homogenous.

5.2 ECONOMIC EFFECTS ON HEALTH

Overall, the findings for economic effects on health were very mixed in nature. For this portion, we focused on economic factors such as: Unemployment, Income Inequality, Children in Poverty, and Median Household Income. The health factors we selected were: Premature Death rate, Poor or Fair Health, Poor Physical Health Days, and Life Expectancy.

Children in Poverty correlated the highest with the health outcome factors selected on a nationwide, correlation for all counties where data was provided. With there being a significant positive correlation of .81 to .73 with most health outcomes. Children in Poverty also correlated very negatively with the Median Household Income, and moderately with Income Inequality at -.76 and .55 respectively.

While these factors are not correlating directly, we can see that childhood poverty increases adverse health outcomes. Furthermore, childhood poverty increases as economic factors sour. These factors form an indirect correlation between them. While not as strong or statistically significant as a direct correlation, this insight drove us to look at regionalization of these trends.

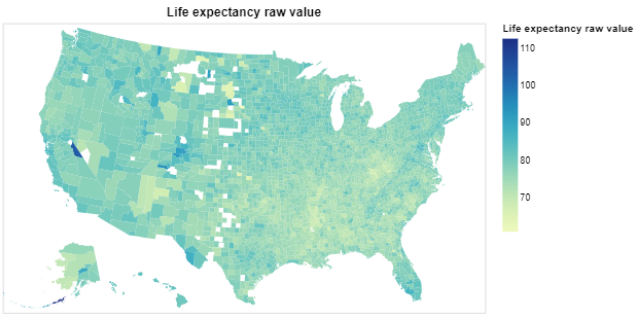

Fig 5. Life expectancy in the United States

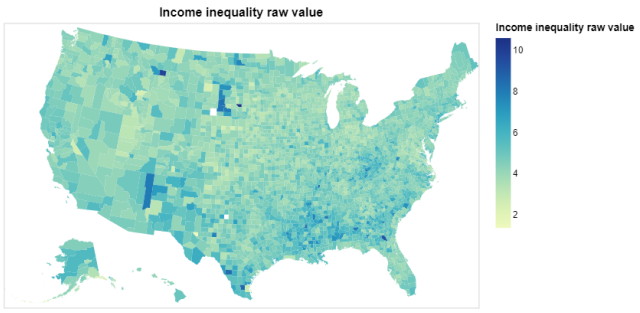

Fig 6. Income inequality in the United States

With these heatmaps above, we can see the regionalization of certain trends. While the overall data did not suggest any correlation between these 2 factors (Income Inequality and Life Expectancy). We can clearly see that in Appalachia and the Deep South, both of these factors are clearly present in combination in these regions. Focusing in on these areas with other heatmaps we can also see further correlations between these factors and other factors with low nation-wide correlations.

5.3 EDUCATIONAL ATTAINMENT AND HEALTH

The overall findings for correlations between educational attainment and health outcomes was very clear in this section of research. Very little ambiguity was left for interpretation of the data in any particular way. Factors selected for educational attainment were: High School Completion Rate, and Some College Attended. Factors for health outcomes were: Premature Death Rate, Poor or Fair Health, Low Birth Weight, and Uninsured Rate.

Completion of high school and some college correlated very strongly in a negative correlation with the rate of poor or fair health in a community. The respective numerical values are -.85 and -.76 for each educational factor. Less strong correlations exist with the educational factors and the premature death rate. Resulting in negative correlations of -.49 for high school completion and -.52 for some college attended.

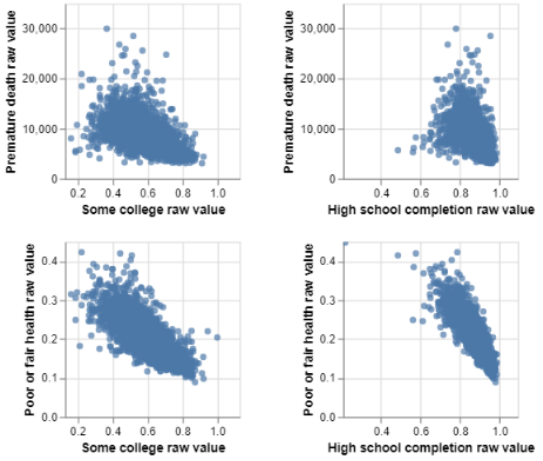

Fig 7. Scatter plot of educational attainment and health outcomes

In this section, regionalized break downs with heatmaps were unnecessary, due to the strong direct correlations found between certain health outcomes and educational attainment. Scatterplots and numerical data helped our team to get a good grasp of the data present and the possible trends between these factors that we found.

An interesting extrema that was found was Kerr County, TX. Kerr County had the lowest educational attainment rates for both high school and college level academics. While this in of itself does not suggest anything, the combination of it having the highest rate of poor and fair health does. With Kerr County representing the extrema for these 3 factors, it drives home the point of

5.4 ECONOMIC FACTORS AND NEGATIVE SOCIAL BEHAVIOR

For the majority of this portion of the project, results seemed to be inconclusive on a national level. Regionalized correlations also seemed to be less fruitful than those discussed in section 5.2 with economic factors and health outcomes. The economic factors selected for this section were income inequality and children in poverty. The negative social behaviors were: adult smoking, adult obesity, and excessive drinking. We selected these as they seemed the most appropriate for viewing negative social behaviors.

The highest correlations found were with childhood poverty having correlations with adult smoking and obesity at .65 and .59 respectively. As discussed in section 5.2, childhood poverty has a direct correlation with negative health outcomes later in life and indirectly related to negative economic factors. Due to the lacking correlation between these factors, we chose to take a look at the regionalized trends to see if any further results could be generated.

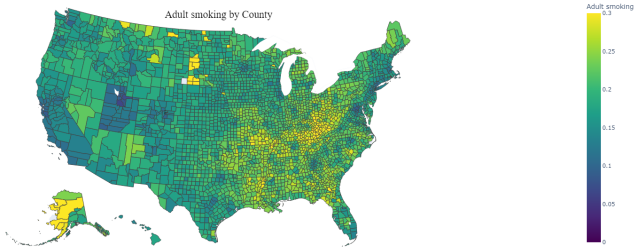

Fig 8. Adult smoking by county map of United States

By comparing Figure 8 with Figure 6, we can see that the same areas that have high income inequality, specifically in the Deep South and Appalachia, tend to have higher rates for this specific factor. However this trend does not hold for the excessive drinking factor and income inequality.

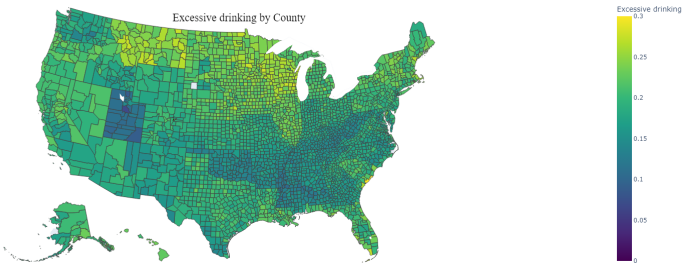

Fig 10. Excessive drinking by county map of United States

In the above heatmap (Figure 10) we can see that the inverse actually appears to happen in the more economically challenged areas. Drinking appears to go down in Appalachia and the Deep South, yet rises as one moves north. This map also shows the effects of state laws moderating negative social behaviors such as excessive drinking.

Overall this question gave mixed results for the national correlations, yet gave some interesting results for regionalized trends.

5.5 INSURANCE RATES AND HEALTH OUTCOMES

In this section we focused on seeing the effects on health outcomes with higher or lower insurance rates. During our exploration of this topic, we selected the most direct factors to compare with the health effects. For insurance rates we selected: overall uninsurance rate, adult uninsurance rate, and childhood uninsurance rate. For our health outcomes we focused on the rate of poor or fair health, the number of poor health days and number of poor mental health days.

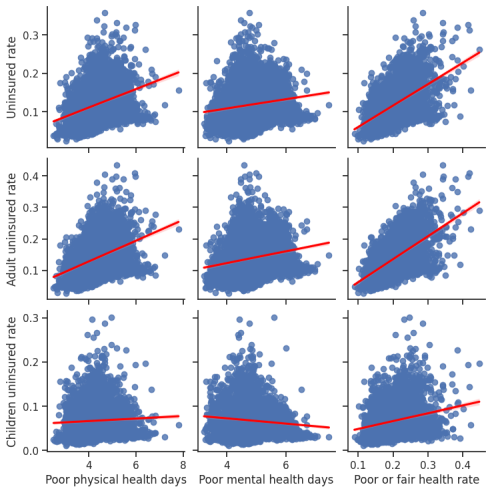

Fig11.Scatterplots of insurance rates vs health effects

Initially we used scatter plots to determine if a regionalized approach was going to be necessary. When doing national correlations with all +3000 counties, many of the correlations were very weak in measure. The clustering of the data points showed us that we needed to take a more in depth, regionalized approach to see if any correlation could be found there.

Using Figures 12 and 13, we can see that there does not appear to even be a regionalized correlation. While certain areas of Texas may reflect a slight regionalized trend, it is not nearly strong enough to be statistically significant. What is interesting however, is the clear state by state divisions with regard to insurance rates. This is most likely due to the way programs like Medicaid are administered on a state by state basis.

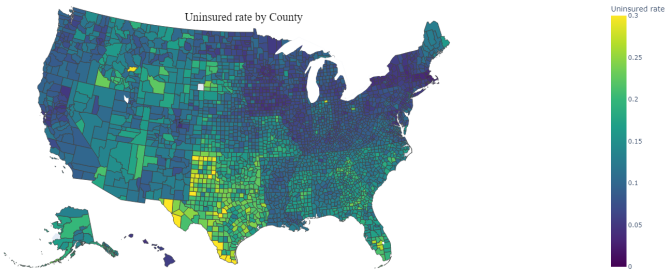

Fig 12. Uninsurance rate by county

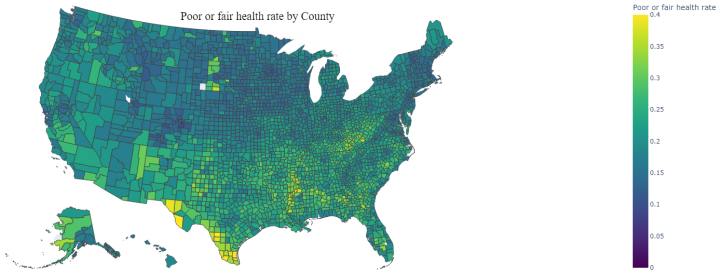

Fig 13. Poor or Fair health rate by county

5.6 PREDICTIVE MODELING OVERALL

For our predictive modeling portion of our research, we focused primarily on analyzing linear regression models using SciKitLearn. Other models were tried with little success such as extreme gradient boosted decision trees and polynomial regressions. Many of these other attempts were unfruitful and only gave inconclusive results.

As discussed in the methodology section of this paper, we used boxplots to eliminate outliers that would otherwise poison the model. Overfit was an issue that our group was having before we eliminated these outliers. By attempting to minimize the MSE and MAE, the outliers held far too much weight in the machine learning algorithm we used for the linear regressions. For both of the predictive modeling sections we did, the drop rate never exceeded 10 percent.

During the development of our predictive models, we attempted many different train/test splits for the data set. The final split we decided on was 50 percent for each section. This balanced the dataset and gave us confidence in the models. In doing this, we ensured that there was not an overfit for the training data which would make the models unreliable.

5.7 USING EDUCATION AND ECONOMICS TO PREDICT HEALTH

In this section, we sought to see if we could generate reliable models to predict health outcomes. This is similar to our exploratory data analysis sections 5.2 and 5.3, but now with machine learning models for linear regressions. The health outcomes used are: poor physical health days, poor or fair health rate, poor mental health days, and the premature death rate. The determinant factors analyzed were: unemployment, high school completion, and the rate of some college attendance.

Overall, some determinants generated better models than others. typically those that saw strong national correlations generated strong models. Such as the one below in Figure 14.

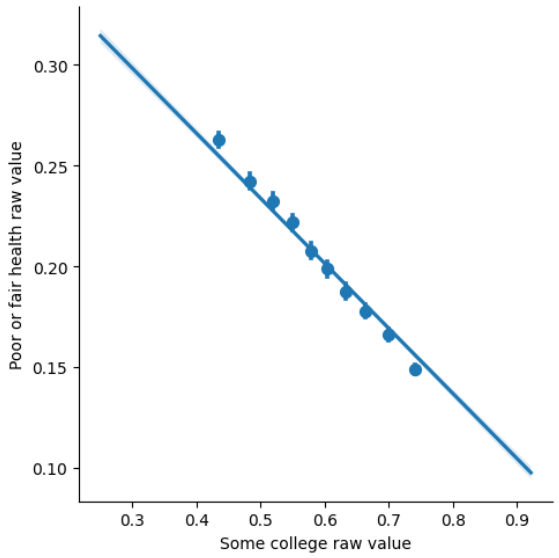

Fig 14. Binned linear regression model for some college determining poor or fair health rate.

In the above figure, we can see that there is little variance from the linear regression line generated by the training section of the dataset. The test data points, while binned for convenience, closely follow the linear regression. The vertical bar from the data points shows the total variance in the data that was binned to form that data point.

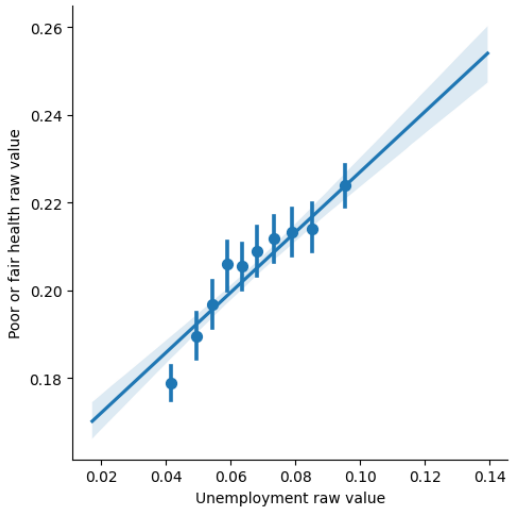

Fig 15. Binned linear regression model for unemployment determining poor or fair health

A model that did not work very well was unemployment and poor health rates. The variance for the model was high as we can see in Figure 15 below.

As we can see in the above model, the variance is much higher than the one found in Figure 14. With a higher MSE and MAE this model is much less reliable than other models generated. The bins as seen above have a much less central, linear distribution. This shows more clustering is occurring in the central part of the graph, causing a reliability issue with the above predictive model in Figure 15.

5.8 USING ENVIRONMENT TO DETERMINE HEALTH OUTCOMES

For this section of our predictive modeling research, we focused on the effects of environmental pollution on health outcomes in the same county. The health outcomes examined were: poor health days, poor health rate, poor mental health days, and the premature death rate. The environmental determinants analyzed were: water drinking violations rate and the air pollution particulate matter.

Overall most models were not of high quality and reliability for this section of our predictive modeling research. Many models had high variance or unacceptably high MSE/MAE values. The most reliable model found was air pollution and the poor or fair health rate, shown on the next page in Figure 16.

In that model we can see the high variance both in the binned data points and the variance in the linear regression. As the model shown in Figure 16 has low reliability, even as the most reliable model, all others in this section showed even lower reliability. Both between the clustering that occurred and the MSE/MAE.

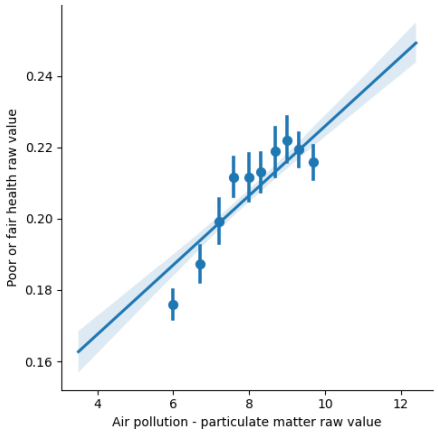

Fig 16. Binned linear regression model for determining the rate of poor or fair health from air pollution values.

With models like these, we determined that an accurate linear regression model was unfeasible. In this case, we did attempt to use decision trees to no avail. Because no model attempted fit the data in this scenario we determined that there was very little way to predict the health outcomes using pollution data.

Some reasons for this could be underreporting of the data in areas where there is a high correlation between the pollution and health outcomes. Another could be the data collection itself. States and counties may underreport the pollution due to ideological bias in the governmental bodies. No matter the possible issues with this data, as it is now we cannot construct a reliable model with what is available.

6. CONCLUSIONS

Overall, the research seemed to be successful. We answered and analyzed all the questions we initially had. Many of the questions were unsurprising in their results, yet some were surprising in their lack of results.

Many of the exploratory analysis questions gave clear cut answers. The ones where we had to take the regionalized approach to the correlations were very surprising though. The lack of nationwide trends emphasized how diverse and large the United States is as a nation, both socioeconomically and in health. Seeing certain areas in the extremes time and time again however, emphasized the difficulty some areas experience with health outcomes that could be caused by a myriad of different factors.

While we will not be continuing study in this area as the same research group, there are many opportunities for future research in this area. As the data is released yearly, it would be interesting to see the effect that the data itself is having on health outcomes over time.

Sources and References:

County Health Rankings. 2022. https://www.countyhealthrankings.org/explore-health-rankings/rankings-data-documentation

Vega Datasets - County FIPS Code and Shape. 2022. https://github.com/vega/vega-datasets

Numpy Documentation. 2022. https://numpy.org/doc/

Pandas Documentation. 2022. https://pandas.pydata.org/docs/

Seaborn Documentation. 2022. https://seaborn.pydata.org/

Altair Documentation. 2022. https://altair-viz.github.io/getting_started/overview.html

XGBoost Documentation. 2022. https://xgboost.readthedocs.io/en/stable/

SKLearn Documentation. 2023. https://scikit-learn.org/stable/

Arndt, Stephan, et al. "How reliable are county and regional health rankings?." *Prevention Science* 14 (2013): 497-502.

Chen, Tianqi, et al. "Xgboost: extreme gradient boosting." *R package version 0.4-2* 1.4 (2015): 1-4.

Govindarajulu, Z. "Rank correlation methods." (1992): 108-108.

Kotsiantis, Sotiris B. "Decision trees: a recent overview." *Artificial Intelligence Review* 39 (2013): 261-283.

McLeod, A. Ian. "Kendall rank correlation and Mann-Kendall trend test." R Package Kendall 602 (2005): 1-10.

Montgomery, Douglas C., Elizabeth A. Peck, and G. Geoffrey Vining. *Introduction to linear regression analysis*. John Wiley & Sons, 2021.

Peppard, Paul E., et al. "Ranking community health status to stimulate discussion of local public health issues: the Wisconsin County Health Rankings." *American Journal of Public Health* 98.2 (2008): 209-212.

Remington, Patrick L., Bridget B. Catlin, and Keith P. Gennuso. "The county health rankings: rationale and methods." *Population health metrics* 13.1 (2015): 1-12.

Google Colab Link With Code

https://colab.research.google.com/drive/1k8H3ueo0VRybdF7jvswDzMMth6tXVNf-?usp=sharing