2022 County Health Findings

1st Eduardo Palacios
Intro. to Data Science (CS-4330-010)
Angelo State University
epalasios5@angelo.edu

2nd Mitchell Martin
Intro. to Data Science (CS-4330-010)
Angelo State University
mmartin46@angelo.edu

3rd Trevor Smith
Intro. to Data Science (CS-4330-010)
Angelo State University
tsmith102@angelo.edu

I. Abstract

A. Introduction

The objective of the research project is to analyze data set given by United Health Group(UHG)[2] in order to develop different predictions that have impacted the country based on race, social and demographic factors. Through the semester, the students in the Introduction To Data Science course have been working with the County Health Rankings & Roadmaps 2022 dataset[1]. The dataset, a grouping of information, given involves statistics grouped by ethnicity which consists of various physical factors that have an effect on our environment. Moreover, the task for the students is to attempt to find correlated factors within the dataset that affect the environment to a higher extent. As a result, our group has decided to make use of a computer language called Python in order to use various visualization tools (graphs, charts, geocharts) for our analysis of the dataset. Furthermore, the visualization tools which we have used for our dataset are Pandas[3], Matplotlib[10], Altair[8], Shap[6], and Seaborn[9]. Moreover, we used the tools within sections in the dataset such as “Alcohol Related Deaths”, “YPLL (Years of Potential Life Lost)”, “High-School Dropout Rates”, and we will eventually use machine learning tools, to see if we can find future predictions within our dataset if certain trends continue.

Keywords: dataset, data-visualization, Pandas, Altair, Seaborn

II. Dataset(s)

A. Data Cleanup

The **County Findings Dataset** was high in storage, so it was necessary to utilize different forms of application software in order to manipulate and visualize the data in order to minimize the time needed to work on the questions for this dataset (e.g. Pandas, Matplotlib, Altair). Moreover, the team eventually moved to the data cleanup process as many of the columns received from the dataset were ”unnamed”. [5] The algorithm utilized sequentially iterates through each column and if the column shows as ”unnamed”, it will be changed to a blank column (see Figure A).[5]

Identify applicable funding agency here. If none, delete this.

The written psuedocode for the algorithm used is given in the figure below along with the python code.

Algorithm 1 Data Cleanup Algorithm

Require: $dataset$
  for c_i, $columns$ in $enumerate(dataset.columns.levels)$ **do**
    $new\_columns$ = $columns.tolist()$
    for r_j, row in $enumerate(new\_columns)$ **do**
      if $'Unamed:'$ in row:
        $new_columns\_j$ = ”
    $dataset_i$ = $new\_columns_i$
  end for
  return dataset

```
[ ] def remove_unnamed(df):
      """Solution found on https://stackoverflow.com/questions/40839609/rename-unnamed-multiindex-columns-in-pandas-dataframe"""
      for i, columns in enumerate(df.columns.levels):
        new_columns = columns.tolist()
        for j, row in enumerate(new_columns):
          if "Unnamed: " in row:
            new_columns[j] = ""
        if pd.__version__ < "0.21.0":
          df.columns.set_levels(new_columns, level=i, inplace=True)
        else:
          df = df.rename(columns=dict(zip(columns.tolist(), new_columns)),
                         level=i)
      return df
```

Figure A

Continuing the discussion regarding the data cleanup process, there were plenty of rows that were had no values, which concluded to more problems. Moreover, plenty of rows within the given dataset had values that were left as missing for numerous countries. As a result, initializing all the values seemed like an optimal option but led to a numerous amount of results be initialized to zero, so unfortunately when predictive modeling because a strong topic in our research, we had to completely remove many of the columns in order to get a more accurate predictive model for our dataset.

Additionally, a major problem we had was trying to understand how to match our dataset with the geomap visualization. Within the dataset some of countries within our dataset had to be removed due to the software’s understood countries for the United States. As a result, this may have lead to some missing information within the data visualizations. In order to attempt to solve this problem was to manually put in the each state’s ID manually, which as a result gave us fuller data visualizations.

The team used the Ranked, Sub-Ranking, and the Additional Measure statistics within the dataset, as these were the only areas that had any social, physical, and demographic factors that the data visualization tools could make accurate predictions with.

978-1-7281-4387-3/19/$31.00 © 2019 IEEE

III. Results

A. Predictive Model

As we consider what happens in both physical and social environments by splitting our dataset by 80 / 20. We trained the model and found that in both poor physical and social environments, they both lead to a higher amount of poor physical health days. (See Figure A.1)
As the group continued working on the dataset, we began utilizing the County Health Rankings 2022 Model [1] found within the website. Due to the late notice, the group mainly used the other model for predictive modeling questions regarding the dataset.
As a result, more detail is explained regarding the situation within the **Evaluation** portion of the conference paper.

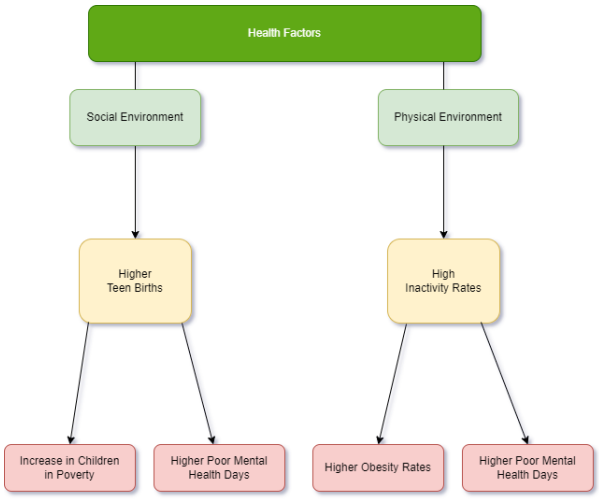


Figure A.1

B. Social & Economic Findings

The % of teen births with all races shows a higher relationship of an increase of poor mental health days among Indian, Asian, and Black Americans, while ethnicities such as White and Hispanic Americans tend to have a lower amount of poor mental health days that happened to have children as teenagers (See Figure B).
Within the software tool, Sckit-Learn [6], we utilized two different scores in order classify which factors impacted different ethnicites to a greater extent. The $R\hat{2}$ is used to evaluate the accuracy of a model given the model utilizes linear regression.
Moreover, The mean squared error evaluates how unreliable is our model compared to the actual dataset.

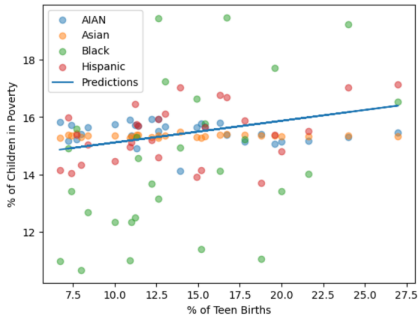


Figure B

Scores for Figure B

AIAN:
R2 Score: 0.2705412056967329
Mean Squared Error: -0.133554445471193
Asian:
R2 Score: 0.25190290095938483
Mean Squared Error: -0.05546085844566995
Black:
R2 Score: 0.20318076632332094
Mean Squared Error: 0.1486825073211927
Hispanic:
R2 Score: 0.2462485436291499
Mean Squared Error: -0.031769377247958364

The % of children that people that have children as teenagers and have children in poverty are usually were mainly minorities, while White Americans tend to have the lowest amount of children in poverty (See Figure C).

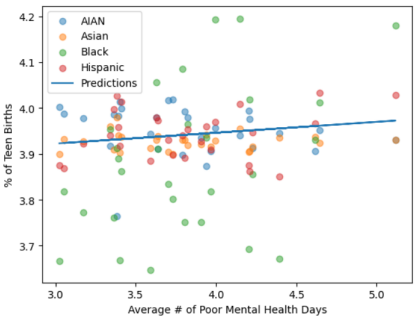


Figure C

One of the problems with the given visualizations is that are the data tends to be all over the place for each race and we had to rely on the mean squared error (MSR) in order to check which ethnicites were effected more on different social factors.

Scores for Figure C

AIAN:
R2 Score: 29.811832312501757
Mean Squared Error: -0.13331146914344338

Asian:
R2 Score: 28.399316526412086
Mean Squared Error: -0.0796139867498411
Black:
R2 Score: 22.016977132912768
Mean Squared Error: 0.16301378462620797
Hispanic:
R2 Score: 25.592778314358796 Mean Squared Error: 0.027077943855695508

Unfortunately, end-users will not be able to understand exactly what these scores so utilitizing a different data visualization seemed like an optimal approach. Using a heat-map as an way of visualizing our dataset, we tend to find that
- Quality of Life has a relationship with health behaviours.
- Quality of Life also has a relationship with Social Economic factors.
- Social Economic Factors also tend to have a strong effect on health behaviors.
(See Figure D)

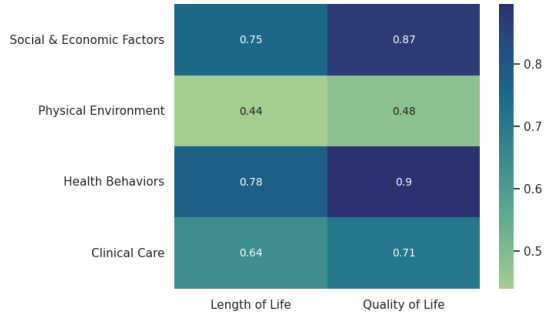


Figure D

C. Project Objectives

1. Do income inequality, unemployment and high school completion rates affect the number of premature deaths of certain racial groups?

Asian: Highschool Completion Rate
Black: Income Inequality
White: Highschool Completion Rate
Hispanic: Highschool Completion Rate (See Figure 1)

The first step to finding how different social factors affected the numbers in premature death among different races was to clean out the data set to only look at the columns that were needed. After importing the data, it was cleaned to only keep the values in the columns that involved Premature Death YLL by Race, Income Percentile Ratio, Unemployment Rate and High School Completion Rate. These factors were then analyzed in relation to premature death, which was used as the primary health outcome.

The dataset was grouped by race and ethnicity, including **Asians, Whites, Blacks, American Indians and Alaska Natives, and Hispanics**. With the dataset grouped, I used the correlations to chart out the data and look for patterns. Some of the Races showed promising results however, I decided to go into a deeper view by keeping the state column in the data set to only have a focused view on the correlations in the state of Texas. With this focal view, there were less outliers and stronger correlations were brought out.

Heat maps were then created using Python's Seaborn library, a statistical data visualization library built on top of Matplotlib. Seaborn's heatmap function was used to visualize the correlations between the selected social factors and premature death among the different racial and ethnic groups. The heat maps revealed that High School Completion Rate had the highest correlation to premature death among Asians and Whites. This suggests that improving educational attainment in these populations could have a significant impact on reducing premature death rates. On the other hand, Income Inequality was found to have the highest correlation to premature death among Blacks.

This indicates that addressing income disparities could potentially improve health outcomes in this population. Interestingly, no correlation was found between social factors and premature death among American Indians and Alaska Natives. Further research may be needed to identify other factors that could contribute to health disparities in this group since the native population is much smaller in comparison to others in the data set.

Hispanics were found to have a similar correlation between Income Inequality and High School Completion Rate with respect to premature death. This suggests that both factors may be important in addressing health disparities within the Hispanic population.

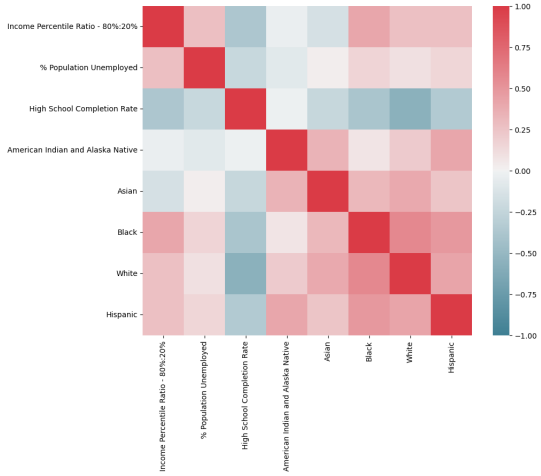


Figure 1

2. Which 5 States have the counties with lowest mental health days, and do excessive drinking ratings lead to larger amounts of people having poor mental health days?

In Question 2, we tried to find out which 5 states hold the counties with the lowest mental health days, and tried to see if excessive drinking has any correlation to it. After gathering the data and organizing it, we found out that the 5 states in question were West Virginia, Arkansas, Alabama, Tennessee and Louisiana. We then graphed the counties together against the excessive Drinking averages and found that there was a correlation of almost 60%, where as Excessive drinking increases, poor mental health days drops, this could mean that people are trying to escape their problems into drinking. The visualization shows that as poor mental health lowers, we tend to see the amount of excessive drinking increases

States:
West Virginia, Arkansas, Alabama, Tennessee, Louisiana

The graph above shows that as poor mental health lowers, the amount of excessive Drinking increases.
Escaping poor mental health by drowning oneself in alcohol? (See Figure 2)

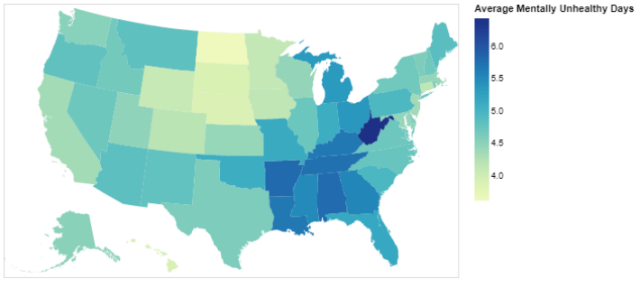


Figure 2

3. Which counties and states have shown to have the highest average number of alcohol related deaths reported within the health rankings and does this correlate with driving alone to work?

The "correlation" legend represents the counties where the Percentage of the - Driving Deaths with Alcohol Involvement - Lone Drive to Work
are near idententical.
Our findings show that there is a higher corrleation in the northwestern part of the United States (e.g. Alaska, North Dakota, Wyoming) than in other counties within the United States. (See Figure 3.A)

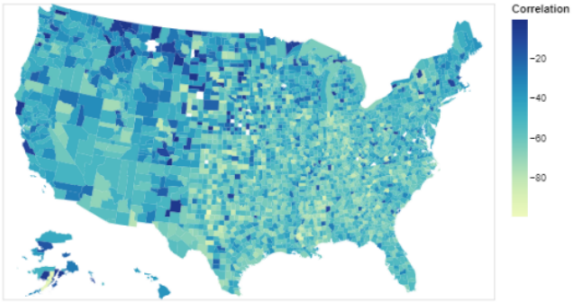


Figure 3.A

Our findings also show that there is a correlation with alcohol related driving deaths with people that drive alone to work. The average is ranges between (8% - 50%) of people dying while drinking and driving with around (70% - 90%) of people driving alone. This county findings show us that while it isn't a suprise that many
individuals drive alone to work, a suprisingly high number of counties have an unecessarily high amount of driving deaths.

Could it be possible that this isolated driving tends to lead to unecessary driving deaths? (See Figure 3.B)

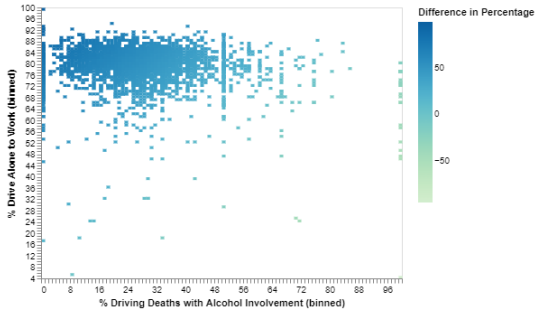


Figure 3.B

Using the correlation formula, we found that we receive a value of 0.5446131445201996. Which shows that there is a relationship between the two factors if the value is closer to 1 than 0.

$$Pearson'sR = \frac{N\Sigma XY - (\Sigma X \Sigma Y)}{\sqrt{[N\Sigma x^2 - (\Sigma x)^2][N\Sigma y^2 - (\Sigma y)^2]}}$$

Pearson's correlation use for the given dataset is used as a measure of the linear association between two factors for our given examples.

4. Do high physical inactivity rates found within the rankings tend to correlate with high adult obesity rates?

We find that Obesity and Physical Activity Rates tend to lead to a YPLL rate of 5,000 to 20,000. What does this data mean?
Our findings show us that both Obesity and Physical Activity Rates tend to have a similar effect of shortening the length of American lives. Another noticeable factor is that individuals that tend to have more poor mental health days within the week tend to also be more
- Physically inactive
- Obese
- Potential years of life lost
(See Figure 4.A)

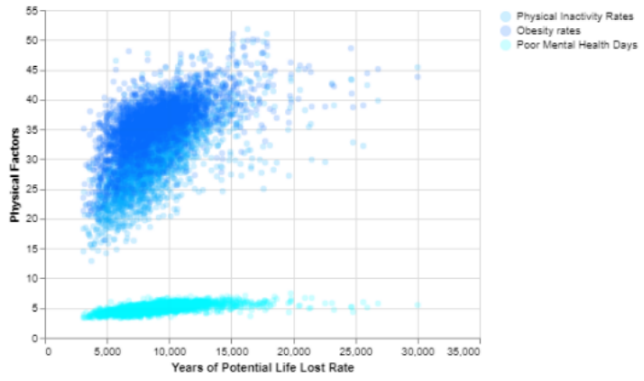


Figure 4.A

We can also compare the amount of poor mental health days for the population within the week with the amount of physical inactivity rates and also see that the coorelation tends to be similar to the obesity rates.
(We can find most of the population has 20-40% of obesity rates and 4-6 days within the week of poor mental health.)

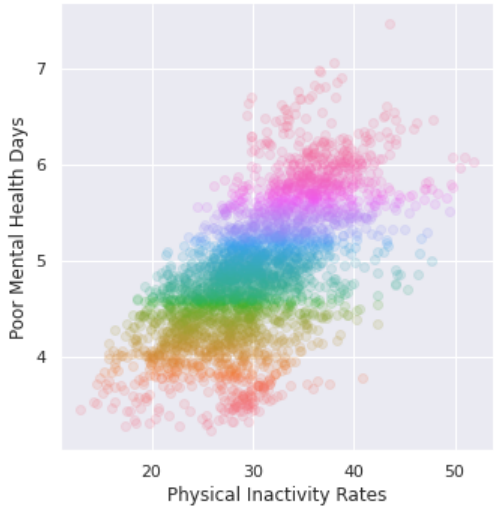


Figure 4.B

As we look more at the correlation between the obesity rates, we tend to find that counties with (30-40%) of obesity rates tend to have in between (4 - 6) days of poor mental health days.
* Pink - 6-7 poor mental health days within the week
* Red - $\leq$ 3 poor mental health days within the week
Even within the outskirts of our data, counties that have obesity rates above 40% within their population tend to have poor mental health days throughout the week.

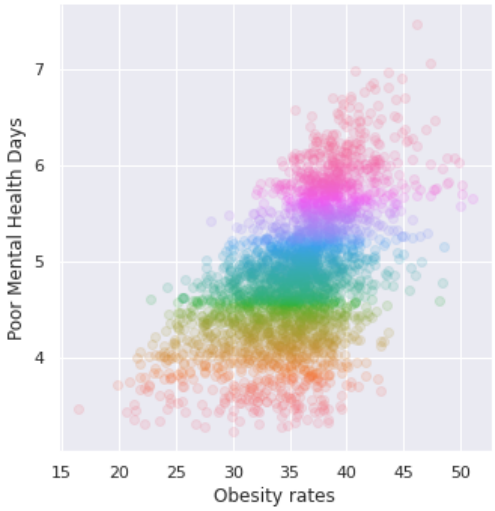


Figure 4.C

5. What are the top 20 counties with the lowest high school completion rates grouped by each state and does this correlate to the number of teen births?

In Question 5, we tried to find the top 20 countries with the lowest high school completion rates and tried to see if there was a correlation between it and Teen births. Among the top 20 counties with the lowest high school completion rates, 14 of the 20 were all located in Texas, and among all counties with low high school completion rates, most are all located in the south. We also found a correlation between it and Teen birth rates reaching almost 70%, this could show that teens who are having children are also uneducated. As for the low high school completion rates being mostly in Texas, this could be due to the fact that it is next to the border and the data taking into account immigrants who haven't had the luxury of pursuing an education even up until high school.

Top 20 counties with the lowest high school completion rates, grouped by state:
Texas: Kenedy, Presidio, Hudspeth, Starr, Maverick, Gaines, Culberson, Frio, Garza, La Salle, Moore, Zapata, Brooks, Hidalgo
Ohio: Holmes
Indiana: LaGrange
Idaho: Clark
Mississippi: Issaquena
Kentucky: Clay
Georgia: Atkinson

Texas holds almost 3/4's of the lowest 20 Highschool completion States, Additionally, possible reasoning could be its location next to the border?
The graph above shows that as the High School Completion rate drops, the Teen Birth rate steadaly rises. There is a correlation of almost 70%. (See Figure 5)

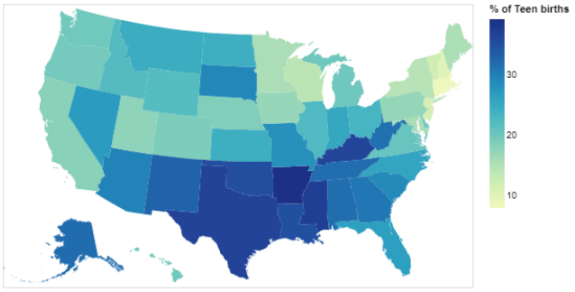


Figure 5

D. Linear Regression

A problem the group was having regarding predictive modeling was that the **mean squared error**, which shows how far off our predictive analysis is compared to the actual data, was usually high. We turned out to use a more efficient method of minimizing the mean squared error by utilizing the MinMaxScaler from the Sckit-learn library within Python.[6]

$$X\_std = \frac{x - x\_min}{x\_max - x\_min} \tag{1}$$

$$X\_scaled = X\_std * (max - min) + min \tag{2}$$

Min-max scaling works by transforming each feature within the dataset individually so that in the resulting range will be between zero and one.[16]

Using this method allowed us the minimize the mean squared error as well as being provided a more accurate looking data visualization for our predictive models.

As stated in the introduction, one of the problems that the group confronted was the missing data in the given dataset. Each time we had to develop a predictive model we always had missing data. Our solution, while not the most accurate, was to remove all the rows that did not have a value.

Are there demographic and social factors that are predictors of drug overdose, alcohol-related driving incidents?

Using **linear regression**, we are able predict that with the percentage of alcohol deaths being factored within each county, we can predict that a higher percentage of alcohol deaths will gradually decrease the social ranking of a county as well as their general health behavior.

$$\boldsymbol{Y_i = \beta_0 + \beta_1 X_i} \tag{3}$$

Linear regression is used within machine learning to predict a variable, our dependent variable, considering a value of a different variable, our independent variable.

With the percentage of drug-overdoses considered there surprisingly turns out to be a remedial increase of both the social ranking and the health behavior within each county.

Blue Line - Social Ranking and Health Behaviors comparison

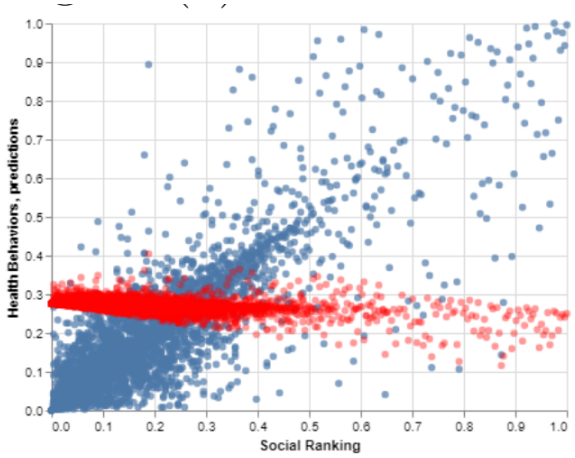


Figure 6.A

Red Line - Social Ranking and Health Behaviors comparison **(% of Alcohol Deaths considered)**

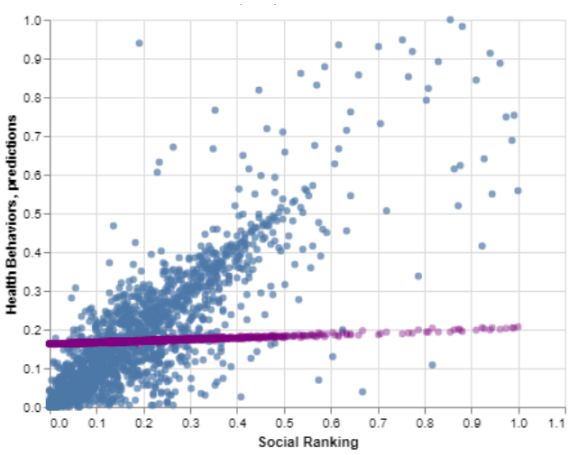


Figure 6.B

Purple Line - Social Ranking and Health Behaviors **(% of Drug-Overdoses considered)**

In general, if we look at a visualization of all races, we tend to find that the relationship between Poor Physical Health Days and the of Preventable Hospital rates tend to be similar, with whites and blacks having a slight edge in poor physical health days.

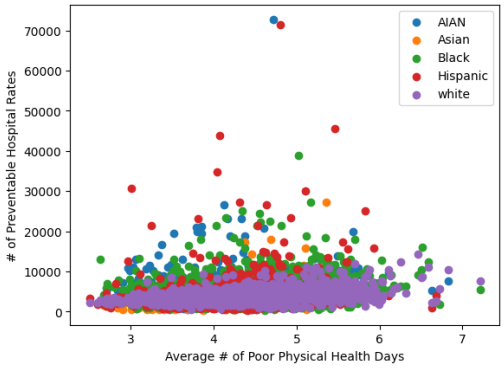


Figure 6.C

IV. Additional Research

A. Objectives

After completing the project objectives for the given dataset, most of the additional research will encompass a machine learning approach. Utilizing different predictive modeling techniques, it seems ideal to see if we can find different social, economic, and educational factors that could impact how well a country may be in the future regarding social & economic rankings. The main data structuring libraries used within the additional research will mainly involve Pandas, Sckit-learn for automated machine learning algorithms and using Matplotlib for data visualization techniques.

B. Findings

An increase in college students is predicted to cause an increase of Calmedia Cases.

The data visualization shows us that as the of Chlamydia cases increases, the premature death rate also tends to increase. Using Kernel Ridge Regression, the algorithm shows us that if we consider the more people take courses in college, it predicts that we may have an increase of chlamydia cases and death rates the more college students we have. (See Figure Z.1)

$$\hat{\beta} = (X^T X + \lambda I)^{-}1 X^T y \tag{4}$$

Kernel Ridge Regression is a combination of Ridge Regression, which imposes a penalty on the size of the coefficients, Ridge Classification, the linear least squares with l2-norm regularization, and the Kernel Trick, which utilizes for non-linear transformations.

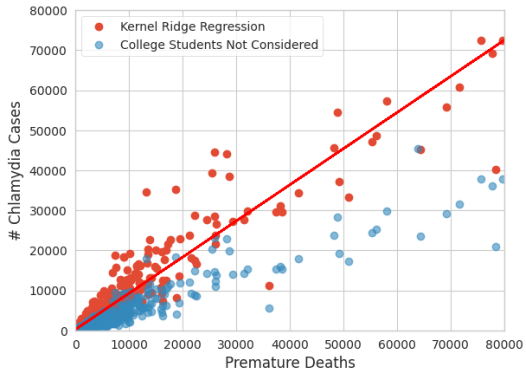


Figure Z.1

Utilizing a pipeline[20], the chained operations with a Standard Scaler, Linear Regression, and Principal Components Regression, another machine learning approach, we are able to see that as reading scores decrease, the same will be for the social ranking. Moreover, we are are actually able to predict that a higher gender pay gap may slightly increase the reading scores for most counties.

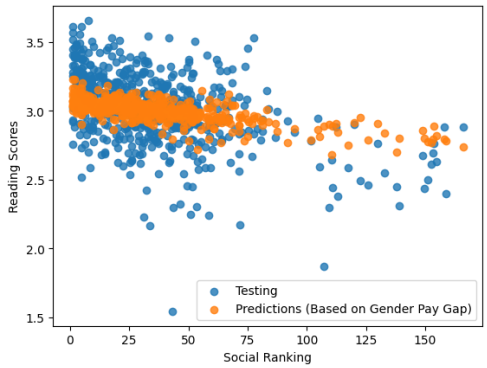


Figure Z.2

$$RSS = \Sigma(y_i - \hat{y}_i)^2 (5)$$

Principal Components Regression

This form of regression uses the least squares to fit a linear regression model by using the principal components as the predictors.

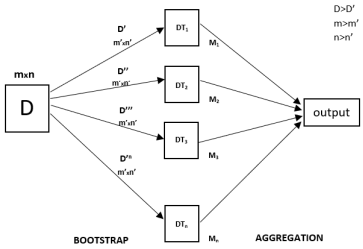
$$z = \frac{(x-u)}{s} \tag{6}$$

Standard Scaler

According to Sckit-Learn documentation, Standard Scaler is used by removing the mean and scaling to the unit variance.

Using another predictive analysis option, Random Forest Regression, shows us that individuals that had higher math scores usually are expected to have a longer life expectancy than those who have lower math scores. The pink represents individuals who tend to have expectancy over 80 years while the individuals in blue represent a life expectancy of less than 74 years. Using a different software tool, SHAP[7],

allows us to see each counties typical math score within the visualization.



Random Forest Regression

Random Forest Regression is a collection of multiple random decision trees, conditions that recursively split until we are left with a non-splitable condition. The Random Forest Regression ensures that the same data isn't being used on each tree, this "method" is called **bootstrapping**.

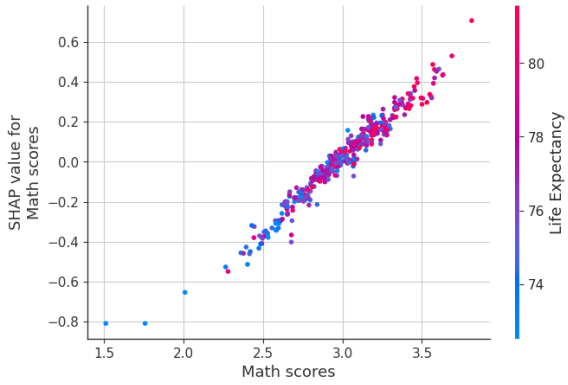


Figure Z.3

Using Principal Component Analysis, Linear Regression, and the Standard Scaler, we are to predict that Chlamydia Rate will be expected to increase as we have a higher rate of excessive drinking within the given counties. On the other hand, when exercise opportunities are factored for Principal Component Analysis, we can see that the Chlamydia Rate actually tends to remedially decrease. (See Figure Z.4)

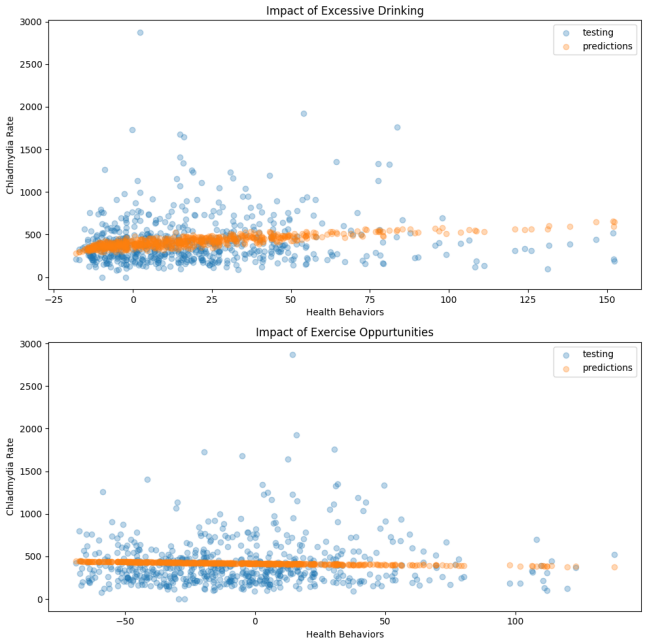


Figure Z.4

V. Evaluation

After learning about our dataset and the given County Health Rankings Model, there seemed to be more of a misunderstanding as to what outcomes needed to be evaluated within our dataset. In the Health Rankings Model, utilizing the Health Factors subcategories were supposed to be the inputs as to what to use to make predictions on the subcategories of the Health Outcomes. If the group had more time, separating the data set's factors and outcomes could have possibly saved the team a greater amount of time as to understanding what was supposed to be understood within the dataset. On the positive side, the process of understanding what our dataset was trying to show the team allowed us to utilize a variety of different software tools to find results that were not even expected to be found within the dataset. At the beginning of the project the main visualization software that was used was Pandas[3] and Altair[8], which allowed us to manipulate and visualize the dataset. Overtime, utilizing different data visualization libraries such as Matplotlib[5] and Seaborn[10] allowed the group to find different ways to visualize our dataset. Moreover, utilizing Sckit-learn's machine learning algorithms[6] and utilizing the County Health Rankings Model would have made for interesting project objectives that could have replaced a majority of the questions that were used to find the results of the given dataset. Continuing the discussion of the County Health Rankings Model, before the group came to a stronger awareness of how the model worked, just looking through the excel sheets and checking which factors had the strongest correlation was used in order to obtain a decent amount of the results within the document.

VI. Conclusions & Future Work

Our analysis revealed that we were able to find different situations in health and social factors based on different ethnicites. As in our findings for premature death, we found that Asian, White, and Hispanic Americans will tend to have premature deaths more frequently if they haven't completed high school, and for Black Americans we found that Income Inequality turned out to be a more significant factor. Moreover, utilizing different visualizations of our data, were able to find how factors such as Quality of Life tend to closely depend on how high a county has ranked within Social & Economic Behaviours and Health Behaviors, as an example. Additionally, using the University of Wisconsin's model for the specific dataset[1], we were able to use a variety of factors such as educational, clinical, physical, and social factors in order to come up with more accurate results for our health outcomes, allowing our findings to gradually improve as the progress on working on the dataset continued. The group as a result, began using different machine learning approaches in order to predict different outcomes for certain factors of our dataset. (e.g. Linear Regression, Kernel Ridge Regression, Random Forest Regression, etc.) using a prediction tool known as SK Learn[6]. Using these machine learning approaches we were able to find how different educational and social factors may have an impact on health behaviours within different counties.

Future work within the given dataset would more than likely involve learning more about different machine learning approaches to given more accurate predictive models for the given dataset. Additionally, utilizing neural networks have been a strong consideration as to making our data visualizations possibly more accurate. Moreover, another consideration is to utilize the 2023 County Health Findings dataset and possibly use a machine learning oriented approach regarding project objectives.

References

[1] County Health Ranking Datasets "How Healthy Is Your County?: County Health Rankings.", University of Wisconsin Population Health Institute, County Health Rankings amp; Roadmaps, https://www.countyhealthrankings.org/about-us

[2] UHG - Project Partner of the EA Ranked Data File CSV Dimensions: (3192 rows, 249 columns)

[3] Pandas McKinney, Wes. "Pandas." Pandas, https://pandas.pydata.org/.

[4] K. Elissa, "Title of paper if known," unpublished.

[5] Rename unnamed multiindex columns in Pandas DataFrame. (2016, November 28). Stack Overflow. https://stackoverflow.com/questions/40839609/rename-unnamed-multiindex-columns-in-pandas-dataframe

[6] Scikit-learn: Machine Learning in Python — Scikit-learn 1.2.2 Documentation. scikit-learn.org/stable.

[7] Welcome to the SHAP Documentation — SHAP Latest Documentation. shap.readthedocs.io/en/latest/index.html.

[8] altair.Chart — Vega-Altair 5.0.0dev Documentation. altair-viz.github.io/user_guide/generated/toplevel/altair.Chart.html.

[9] Linear Regression With Marginal Distributions — Seaborn 0.12.2 Documentation. seaborn.pydata.org/examples/regression_marginals.html.

[10] —. matplotlib.org/stable/index.html.

[11] GeeksforGeeks. "Random Forest Regression in Python." GeeksforGeeks, Apr. 2023, www.geeksforgeeks.org/random-forest-regression-in-python/.

[12] Simple Linear Regression and Pearson Correlation - StatsDirect. www.statsdirect.com/help/regression_and_correlation/simple_linear.htm: :text=Pearsoń%20product%20moment%20correlation%20coefficient, regression%20of%20Y%20on%20x.

[13] About Linear Regression — IBM. www.ibm.com/topics/linear-regression: :text=Resources-,What%20is%20linear%20regression%3F,is%20called%20the%20independent%20variable.

[14] Visually Explained. "The Kernel Trick in Support Vector Machine (SVM)." YouTube, 9 May 2022, www.youtube.com/watch?v=Q7vT0–5VII.

[15] GeeksforGeeks. "Python Coefficient of Determination R2 Score." GeeksforGeeks, Jan. 2023, www.geeksforgeeks.org/python-coefficient-of-determination-r2-score.

[16] "Sklearn.Preprocessing.MinMaxScaler." Scikit-learn, scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html.

[17] Normalized Nerd. "Random Forest Algorithm Clearly Explained!" YouTube, 21 Apr. 2021, www.youtube.com/watch?v=v6VJ2RO66Ag.

[18] Zach. "Principal Components Regression in Python (Step-by-Step)." Statology, Nov. 2020, www.statology.org/principal-components-regression-in-python.

[19] Scribbr. "Scribbr - Your Path to Academic Success." Scribbr, 25 Jan. 2023, www.scribbr.com.

[20] Greg Hogg. "Scikit-Learn Model Pipeline Tutorial." YouTube, 22 Oct. 2021, www.youtube.com/watch?v=xIqX1dqcNbY.