County-Level Predictive Modeling of Gastric Cancer Mortality: The Role of Demographics, SDOH, and Health Equity Factors*

Shailja Somani†
Applied Data Science
Master's Program
Shiley Marcos School of
Engineering / University of
San Diego
ssomani@sandiego.edu

Yicong Qiu†
Applied Data Science
Master's Program
Shiley Marcos School of
Engineering / University of
San Diego
yqiu@sandiego.edu

ABSTRACT

Stomach cancer is an ongoing public health issue in the United States, with approximately 11,000 Americans expected to die of the disease in 2024 and approximately 130,263 people estimated to be living with the disease in 2021. Past research has shown that oncology patient outcomes are greatly influenced by demographic, social determinants of health (SDOH), and health equity factors, but past research has often only looked at a subset of factors rather than considering a wide cross-section. The goal of this work was to consider 353 such factors to: (a) predict counties with high expected stomach cancer mortality rates to allocate higher medical resources to and (b) determine which factors are most strongly correlated with mortality rates to invest further research and program funding into.

Various models and hyperparameters were evaluated, with the optimal model being a Light Gradient Boosting Machine Regressor trained on all demographic and health equity factors, as well as 90 components derived from Principal Components Analysis (PCA) on the SDOH features. The final model achieved a Root Mean Squared Error (RMSE) of 0.0000355 (to offer context, that is 36.8% of the mean of the target variable) and adjusted R^2 value of 0.9607. Given that the model's performance declined at higher mortality rates, it is recommended that public health officials set a cutoff based on resources available (for example, top 100 county and age/race combinations, as the target data is separated by age and race) and then allocate additional resources to all populations above that cutoff, as determined by our model. To improve model explainability and gain the ability to deep dive into how features used impact a specific prediction, SHapley Additive exPlanations (SHAP) values were used. Top predictive factors in the model include age, race, preventative care (such as annual checkups and mammograms), and comorbidities (such as binge drinking and current asthma). It is recommended that public health officials and hospital networks invest in research, programs, and case manager support in these areas.

* This research was conducted as the authors' capstone project for the MS in Applied Data Science program at the Shiley-Marcos School of Engineering at the University of San Diego.

† Shailja Somani, Roger Qiu, Shiley-Marcos School of Engineering, University of San Diego.

KEYWORDS
machine learning, health disparities, cancer, mortality rates, patients, health outcomes, social determinants of health, stomach cancer, gastric cancer, oncology, predictive modeling, county demographics, SDOH, data science, health tech, racial disparities, health equity

1 Introduction

Stomach cancer remains a large public health concern in the United States, with an estimated 27,000 Americans expected to be diagnosed with the disease in 2024, and approximately 11,000 of these cases anticipated to be fatal (American Cancer Society, 2024). This accounts for 1.5% of new cancers each year (Katella, 2024). To further quantify the magnitude of this impact, in 2021, there were an estimated 130,263 people living with stomach cancer in the United States (Surveillance, Epidemiology, and End Results, 2021). Because of this, understanding the factors that cause stomach cancer mortality rates is critical for creating targeted interventions and policies that can help at reducing these rates and address health disparities. The National Center for Health Statistics (NCHS) provides a comprehensive data set on stomach cancer cases at the county level, aggregated by demographic factors. By joining that dataset to county-level socioeconomic and health equity metrics, there is an opportunity to analyze the distribution of this disease across multiple different influences.

The goal of this research is to compare, analyze and predict the effects of various demographic, socioeconomic, and health disparity metrics on stomach cancer mortality rates across U.S. counties in 2019. In doing so, the importance of these factors in the health outcomes of individuals can be discovered, providing evidence-based recommendations for public health interventions. The working hypothesis is that disparities in stomach cancer mortality rates are influenced by these metrics, and that machine learning modeling techniques can be used to better understand these relationships, identify high risk counties and demographics, and create more effective public health interventions.

By leveraging advanced data science methods and machine learning models, this project aims to provide actionable insights that can drive policy changes and effective health programs. These efforts are important in addressing the different underlying comorbidities associated with stomach cancer and improving overall health outcomes. Through collaboration with public health officials and policymakers, it is intended that the findings from this study will be used in creating strategies to mitigate stomach cancer mortality and create better health outcomes across populations."

2 Background

Stomach cancer, or gastric cancer, is an important health issue in the United States, impacting thousands of individuals every year. With variations in mortality rates across different demographic groups and geographical regions, this disease poses a substantial challenge to public health. Data from the National Center for Health Statistics (NCHS) reveals county-level disparities in stomach cancer outcomes, creating the opportunity for use of data science and machine learning methods to

analyze the contributing factors. Furthermore, by integrating NCHS data with additional datasets containing social determinants of health (SDOH) and detailed health measures, deeper insights can be uncovered on the socio-economic and healthcare-related factors that influence stomach cancer mortality rates. This comprehensive, national data driven approach is crucial for creating effective public health strategies and interventions.

2.1 **Problem Identification and Motivation**

The main problems addressedthis research is the high incidence of new cases and fatalities caused by stomach cancer in the United States as well as the lack of comprehensive, large-scale research delving into the socioeconomic factors impacting the mortality rate. Stomach cancer affects thousands of individuals and places a large burden on healthcare systems. The disparities in mortality rates across different ethnic groups and geographic regions indicate underlying socio-economic and other factors that need to be understood and addressed. Existing literature has largely focused on only a few SDOH factors or small subsets of the national population (Tucker-Seeley et al., 2024). To contribute to the field, this research examines a large cross-section of factors across the country. This problem is crucial not only for the affected individuals and their families but also for public health officials, policymakers, and healthcare providers who are responsible for developing effective strategies to address this disease. By identifying and understanding these contributing factors through data science and machine learning methods, valuable insights can be provided that could lead to preventive measures and policy interventions. This research is motivated by the potential to greatly improve health outcomes and reduce mortality rates associated with stomach cancer, as well as improve upon health equity in our country, thus enhancing the overall wellbeing of communities across the United States.

2.2 **Definition of Objectives**

The objective of this research is to utilize data science methods and machine learning techniques to discover important patterns and correlations in stomach cancer and socioeconomic metrics data to make predictions based on these patterns and correlations. Specifically, the study aims to identify the key demographic, socio-economic, and health disparity metrics that contribute to variations in stomach cancer mortality rates across U.S. counties. By creating and validating machine learning models, mortality rates can be predicted based on these contributing factors and validate the models on unseen data to ensure their robustness and accuracy. The expected outcome is that our models will effectively identify key factors and provide actionable insights that can inform public health strategies and policy decisions. Furthermore, the resulting model will be able to identify communities at high risk for high stomach cancer mortality rates. If our hypothesis is true, the findings could lead to the development of targeted interventions, increased cancer screenings, and other preventive measures to reduce stomach cancer mortality rates. On the other hand, if the hypotheses are disproved, the research will still contribute valuable knowledge by highlighting areas where additional investigation is needed. The ultimate goal is to contribute to a comprehensive understanding of stomach cancer mortality and to support efforts aimed at improving oncology outcomes and reducing disparities in cancer care.

3 Literature Review

The purpose of this literature review is to position this research project in relation to existing research on stomach cancer mortality rates. By reviewing recent studies, gaps can be identified in the literature and demonstrate the importance of our own analysis. Although many papers acknowledge the effect of SDOH factors and health inequity on stomach cancer patients' outcomes (Tucker-Seeley et al., 2024) and a need to study such effects further, there is a lack of a comprehensive study that considers a variety of demographic, SDOH, and health equity factors. This review highlights the importance of integrating a myriad of such factors to better understand and address the disparities in stomach cancer outcomes across different populations in the United States.

3.1 Social Determinants of Health and Cancer Care: An ASCO Policy Statement

In 2024, the American Society of Clinical Oncology (ASCO) ("the national organization representing nearly 50,000 physicians and other health care professionals specializing in cancer treatment") published a policy statement addressing SDOH within cancer care (Tucker-Seeley et al., 2024). The statement first highlights how SDOH can affect "as much as 50% of health outcomes'' and health-related social needs (HRSNs) are "negatively associated with outcomes across the cancer continuum" (Tucker-Seeley et al., 2024). Specific SDOH and HRSNs called out are housing stability and quality, access to cancer screening, and financial ability to access precision cancer therapies. ASCO calls for health care providers to integrate SDOH needs assessment and intervention into their cancer treatment plans. Providers are told to consider "SDOH as the equivalent of risk mutations across patient populations" and develop "tailored interventions" for them as such (Tucker-Seeley et al., 2024).

However, ASCO repeatedly highlights the lack of systematic data collection and cross-population analysis as a large issue impeding SDOH analysis and interventions. The statement notes that "there is neither a single perfect measure of SDOH nor a validated social risk score in oncology," "a lack of systematic data collection" and "highly variable measures that limit comparison across populations" (Tucker-Seeley et al., 2024). Furthermore, given the current financial incentive structure within healthcare, providers are unable "to find sufficient time during a brief clinic encounter to ask detailed SDOH questions and address multiple social risks," although ASCO hopes this will change with "future value-based care models" (Tucker-Seeley et al., 2024).

Tucker-Seeley et al., 2024 repeatedly mentions a gap in the current literature for SDOH impact on oncology outcomes. Specifically, "existing literature for behavioral and social health interventions tend to focus on specific patient populations, clinical settings, and pilot programs. This results in a relative lack of knowledge related to addressing structural and organizational barriers to SDOH intervention" (Tucker-Seeley et al., 2024). Our work specifically aims to start to fill this gap by looking at aggregate county-level SDOH, health metrics, and gastric cancer mortality data from across the United States. The aggregated measures are not ideal, but patient-level measures are hard to gather during short appointments (as ASCO notes) and are often gathered too late (after a patient is already symptomatic and thus visiting a doctor). Our goal in using aggregated measures

to model gastric cancer risk across a cross-section of demographic, SDOH, and health equity factors is to provide a large-scale analysis that can be used to provide a high-level view of SDOH needs across the country, spur further research, and incentivize health plans to further invest in value-based care models that reimburse providers for SDOH assessment and intervention (as suggested by ASCO).

3.2 **Socioeconomic disparities in gastric cancer and identification of a single SES variable for predicting risk**

Before diving into our own research, various types of existing SDOH studies were reviewed across the ASCO policy statement mentioned, which were often focused on specific SDOH factors and/or specific populations. One such study is that, in 2021, researchers at the Montefiore Medical Center and Albert Einstein College of Medicine evaluated the impact of education, income, and occupation (as proxy measures for socioeconomic status) on risk of a patient developing gastric cancer (Sarker et al., 2021). When evaluating each variable's impact, age, sex, and race were controlled for. Logistic regression models to predict gastric cancer risk were built using each variable and then "compared using AIC [Akaike information criterion], c-statistics, and pseudo-R square to determine the model that had the highest risk predictive ability" (Sarkar et al., 2021). The article found that all three variables were correlated with gastric cancer risk, but "education contributed the most to model variability" and thus "can be employed as an ideal single indicator of" gastric cancer risk when multiple socioeconomic "factors cannot be obtained" (Sarker et al., 2021).

There is no doubt that Sarkar et al.'s (2021) work is valuable in better understanding which factors contribute to gastric cancer risk. However, there are two key areas upon which we intend to add upon in our research. First, we plan to investigate how the combination of multiple SDOH factors and health metrics can lend to more robust predictive models than investigating one factor at a time. Second, individual-level data is often hard to obtain until it is too late: aggregate county-level data can be easily obtained for preventative measures, but individual-level data is often only obtained after patients already visit doctors with symptoms. Thus, in our work, we use available county-level data rather than individual-level data in the hope of identifying profiles of high-risk individuals prior to them even visiting a hospital. The goal is for healthcare agencies to be able to preventatively conduct outreach to such individuals to schedule cancer screenings, annual primary care visits, case worker consults if needed, and more.

3.3 **The Impact of Racial Disparities and SDOH on Esophageal and Gastric Cancer Outcomes**

In 2024, researchers in the Department of Surgery at the University of Michigan investigated the impact of race on gastric cancer outcomes (Bonner & Edwards, 2024). Their research found that minorities were disadvantaged throughout their gastric cancer journeys for four key reasons. First, there is a "higher incidence of disease" within marginalized groups (Bonner & Edwards, 2024). Furthermore, "marginalized groups are significantly less likely to receive guideline concordant care and high-quality oncologic treatment" (Bonner & Edwards, 2024). On top of not consistently receiving "guideline concordant therapy, marginalized groups face larger financial toxicity related

to their cancer" (Bonner & Edwards, 2024). Finally, marginalized groups are also underrepresented in clinical trials (Bonner & Edwards, 2024). Given how genomics impacts cancer spread and response to treatment, failing to adequately research certain racial groups could lead to failing to adequately research a large subset of potential genomic markers patients have that may impact their response to treatment.

Bonner and Edwards highlight how crucial it is to consider patient race when studying gastric cancer patient outcomes. Thus, the county-level gastric cancer mortality measures are separated out by patient race rather than just looking at the fully aggregated county-level metrics across all patients. However, as previously mentioned, our goal is to also understand how race, age, and gender intersect with other SDOH and health equity metrics. Not only will that lend more predictive power to algorithms predicting gastric cancer mortality, but it will also allow us to investigate if certain racial minority groups are more likely to face certain SDOH challenges than the average patient, thus allowing us to reach out to minority patients with programs for how best to meet their specific needs.

3.4 **Housing Insecurity Among Patients with Cancer**

In 2021, "the National Cancer Policy Forum of the National Academies of Science, Engineering, and Medicine sponsored a series of webinars addressing social determinants of health, including food, housing, and transportation insecurity, and their associations with cancer care and patient outcomes" (Fan et al., 2022). Fan et al. summarizes the presentations and discussions regarding housing insecurity. Housing insecurity impacts cancer patients' care in four key ways. First, a lack of stable housing can result in patients having to move or becoming homeless, both of which may lead to gaps in their cancer treatment (Fan et al., 2022). Second, low quality housing may result in environmental exposures that contribute to cancer risk or exacerbate existing diagnoses. Third, high housing costs leave patients with less money to be able to spend on healthcare costs. Fourth, "neighborhood context can shape health via its proximity to health systems and health-promoting resources, such as access to specialty health-care providers, recreational spaces, stores selling fresh food, and social capital" (Fan et al., 2022). Cancer care is also uniquely intertwined with housing insecurity as "housing can have dynamic, bidirectional associations with access to cancer" because paying for cancer treatment can cause patients to struggle to pay for housing (Fan et al., 2022).

Fan et al.'s (2022) work clearly demonstrates a need to include housing metrics when discussing SDOH impacts on oncology patients. However, it is clear how intertwined housing insecurity is with general financial insecurity, unemployment, and other such socioeconomic factors. Furthermore, Fan et al. (2022) also calls out that "racially discriminatory housing policies," such as how banks have historically provided loans, redlining, and gentrification "intersect with other forms of structural racism such as employment discrimination and mass incarceration to further limit housing access." Thus, as previously mentioned, our work seeks to evaluate the intersection of SDOH housing factors with a variety of other socioeconomic and demographic factors, including race and employment, to conduct a comprehensive analysis of how such factors impact gastric cancer mortality rates.

3.5 The Role of SDOH in Gastrointestinal Cancer Outcomes in the United States Context: A Systematic Review

To reaffirm our understanding of the existing gaps in the literature, we turned to a 2024 article published in the Journal of the National Comprehensive Cancer Network that performed "a systematic review in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) statement" of US studies "studies reporting the effect of SDOH on GI cancer survival outcomes, published on or after January 1, 2002" (Santellano et al., 2024). The quantitative results of the meta-analysis were that healthcare system domain, insurance status, economic stability, and education level were amongst the most evaluated SDOH categories. Privately insured patients had higher survival rates, Medicaid patients had lower survival, and higher education levels were significantly associated with higher survival rates (Santellano et al., 2024).

The meta-analysis concludes that "there is a clear relationship between SDOH and survival for patients with GI cancer. However, there is great variability on how SDOH are measured and reported preventing a holistic analysis of SDOH in this context. A better understanding of the whole spectrum of SDOH is needed to design strategies to improve outcomes for GI cancer patients" (Santellano et al., 2024). That is the very gap in literature that our research attempts to start to fill. Rather than examining only a few SDOH factors in isolation or considering only a small subset of the US population, correlations and models were investigated using a large set of SDOH and health metrics across all US counties.

4 Methodology

Data was extracted from numerous sources online, including a target table with data regarding stomach cancer mortality, socioeconomic and health equity features from two other tables, and a crosswalk table to be able to join the aforementioned. Numerous transformations and aggregations were required before the dataset was ready for exploratory data analysis (EDA) and other statistical analysis. Analyses conducted include Univariate Non-Graphical Analysis, Univariate Graphical Analysis, and Multivariate Non-Graphical and Graphical Analysis. Then, distribution analysis and other visualizations were created to help further understand the dataset. Finally, additional pre-processing is conducted to prepare for modeling and hyperparameter tuning.

4.1 Data Acquisition and Aggregation

The raw data was extracted from four websites in the form of either CSV files or Excel files. The four datasets used are as follows:

1. United States Stomach Cancer Mortality Rates by County, Race, and Ethnicity 2000-2019: Obtained from the Institute for Health Metrics and Evaluation Global Health Data Exchange (2024). This data was obtained in the format of two CSV files: one for male mortality and one for female mortality resulting from stomach cancer in 2019. The CSVs were 54 MB each.
2. Social Determinants of Health (SDOH) Data by U.S. County: Obtained from the Agency of Healthcare Research and Quality (2023). AHRQ "is the lead federal agency charged with

improving the quality and safety of America's health-care system." (Kronick, 2016). Thus, the AHRQ SDOH database is widely used and accepted within the healthcare research realm. The 2019 SDOH dataset was used to remain aligned with our 2019 stomach cancer mortality dataset. The dataset is 13.4MB.

3. Centers for Disease Control and Prevention (CDC) PLACES Local Data for Better Health: This dataset obtained from the CDC contains Zip Code Tabulation Area (ZCTA) estimates of 36 health equity measures. Those 36 are split into "13 for health outcomes, 9 for preventive services use, 4 for chronic disease-related health risk behaviors, 7 for disabilities, and 3 for health status" (CDC, 2023). The dataset is 250.6MB.
4. Federal Information Processing Standard (FIPS) to Zip Code Tabulation Area (ZCTA) Crosswalk: This well-formatted dataset is obtained from DataWorld, but adapted from publicly-available data from the United States Census Bureau (Rippner, 2024; US Census Bureau, 2021). FIPS and ZCTA codes are two distinct codes used to delineate geographical boundaries. Because some datasets of the three are using are metrics aggregated at the FIPS-level and others at the ZCTA-level, this crosswalk is used to join the aforementioned datasets. This crosswalk is a 3MB CSV.

Because all datasets used in our work are publicly-available, aggregated health and/or census datasets (no individual-level health data is used), there are no Health Insurance Portability and Accountability Act (HIPAA) concerns. Furthermore, stomach cancer mortality rates are reported by a variety of age groups, five race categories, and by male and female. By considering the various cross-sections of these demographic groups as independent observations (rather than looking at aggregated data that may be skewed towards certain racial or age majorities), the goal is to minimize bias in our research.

4.1.1 Preparation and Consolidation

To prepare all the raw data from their sources and to combine them together, numerous methods were used during the preparation phase. The Pandas library was used to read each file in and inspect it. The stomach cancer mortality rates data came in both male and female datasets so the two were unioned together to create one dataset with the data for both genders. This target dataset has over 800,000 rows and 19 fields. Out of these, only seven fields were kept as the others did not provide any useful information. This table was then filtered based on numerous conditions to remove records that were either at an aggregated level (for example, state-level combining data from multiple counties) or counties for which no mortality rate was reported. The resulting table has 277,894 records and seven columns. It reports mortality rate by county, race category, sex, and age group.

Next, the SDOH dataset was read in. County identification fields (including the FIPS field) were kept, as well as all socioeconomic fields resulting from the American Community Survey (ACS), which obtains a large variety of "social, economic, housing, and demographic data" for each U.S. county (US Census Bureau, 2024). There are 304 American Community Survey (ACS) SDOH metrics in the table. This dataset is joined to the stomach cancer mortality dataset on the County FIPS field.

Following that, the FIPS to ZCTA crosswalk dataset is read in initially as a text file, so delimiters, skiprows and encoding was required to manually be defined before the file could be properly read in. After that, only two fields were kept: the county FIPS code and the associated ZCTA code for each area for a total of 44,139 records and two fields. This crosswalk will be used shortly to join the CDC Health Measures table to the rest of our data. ZCTA areas are generally smaller than FIPS areas, so there are multiple ZCTAs that map to each FIPS.

Finally, the CDC Health Measures dataset is read in and inspected. The raw table is structured such that each row has the ZCTA code for one location, measure name for that row, and value for that measure in that location (amongst a few other metadata columns). To make it easier to work with, the table was aggregated at the ZCTA level so that each row is a ZCTA region, and each column name is one health measure name, with the column values being the value for that measure within that row's ZCTA region. For example, a record will indicate the ZCTA location such as 1001 and then a column would be Binge Drinking, with the value being 15.6, indicating that 15.6% of adults aged 18 or older in that ZCTA area binge drink. The higher the value for any given measure, the more prevalent that issue is within that ZCTA.

To join the CDC Health Measures dataset to our other two datasets, the FIPS to ZCTA crosswalk is joined to the CDC table (after aggregation at the ZCTA level). Then, because there are multiple ZCTA areas per FIPS area, the CDC health measures are aggregated again at the FIPS (county) level by taking the means of the values for each measure across all rows for each given FIPS code. The resulting table is then finally joined back to our main dataset on the FIPS code column. The consolidated dataset has 277,894 rows and 354 columns.

4.2 **Preliminary Feature Engineering**

The first iteration of feature engineering was used to convert categorical fields to numerical ones for the machine learning models in the next part. To begin, the categorical fields that were present were inspected. The first was race_name, which contained values such as White or Latino. Then, the age_name field contains ranges such as <1 year or 25 to 29. Finally, the age_group_id field contains numerical values such as 28 or 160.

The process begins by performing one-hot encoding on the race_name field to result in a column for each race, with 0 or 1's as values to indicate the race for that record. The first of the five fields produced by one-hot encoding is dropped to reduce multicollinearity, as this extra field will not provide additional information. If four of the existing race fields are zero for a record, then it is known that the record must belong to the 5th option.

The sex column is one-hot encoded in the same fashion. The result is just one field: sex_Male. When a record is male, the column value will be 1. When a record is female, the column value will be 0.

For the age field, this field indicates the age range each record belongs to. For this, an ordinal mapping was created to map the categorical range values to a numerical value. For example, <1 year

is mapped to the value of 0, 1 to 4 is mapped to a value of 1, 5 to 9 is mapped to a value of 2, and so forth. It is important to note that these values are kept as ordinal as opposed to nominal because it is important to maintain the relationship that the values increase as the age group does. Thus, the field can remain in one column rather than being mapped to multiple columns, as in the one-hot encoding. The resulting field is added to the table as age_category, which contains values from 0 to 18, 0 being the youngest age group (<1 year old) and 18 being the oldest (85 years old and older).

Following the above transformations, the "val" field from the stomach cancer mortality dataset is renamed to "TARGET_mortality_rate" to better describe the field. Additionally, a few counties and ZCTA fields that were used for the join are dropped. These join keys are no longer needed for the rest of the project. After this preliminary feature engineering, the dataset contains 277,894 records and 354 columns: 1 location name column (kept in for labeling data in the below EDA portion), 1 target variable column (stomach cancer mortality rates), and 352 feature columns. The dataset is written out as a CSV file within the folder location where the Data Preparation notebook is being run. This is done so that the remainder of the project can be performed by simply loading in the consolidated data CSV file and eliminating the need to repeat any of the data preparation steps.

4.3 **Exploratory Data Analysis (EDA)**

The EDA conducted on the fully consolidated dataset includes numerous areas of checks and methods. Univariate Non-Graphical Analysis, Univariate Graphical Analysis, Multivariate Non-Graphical and Graphical Analysis were conducted as part of the statistical analysis. Distributions of a myriad of variables were analyzed and visualized using a variety of plots and charts. Finally, aggregation was used to check the prevalence of a few illnesses across different counties to see which counties had the highest prevalence for select illnesses.

4.3.1 Statistical Analysis

The analysis began with a basic statistical analysis of the stomach cancer mortality rates across different counties. The maximum rate observed in the dataset was 0.00254, and the minimum rate was 0. The mean rate was found to be 0.00009633, and the median rate was found to be 0.0000147, showing that half of the counties had rates below this value.

To further understand the distribution, skewness and kurtosis of the stomach cancer rates were calculated. Skewness, which measures the asymmetry of the distribution, was 2.92, showing a right-skewed distribution with a tail extending towards higher values. Kurtosis, measuring the peakness of the distribution, was 11.14, showing a leptokurtic distribution with a sharper peak than a normal distribution.

For univariate graphical analysis, boxplots and histograms were used to visualize the distribution of stomach cancer mortality rates by county populations. The boxplots showed a large number of high-value outliers in comparison to the median, indicating variability and the number of counties with notably higher rates of stomach cancer. The histograms also illustrated the distribution, revealing the skewness and the presence of outliers. Figure 4.1 below is one of the visualizations

that allowed for visualizing the distribution of stomach cancer mortality rates across US counties. While outliers are noted during this EDA section to better understand the data distribution, they are left in for all further work (rather than being removed from the dataset) to avoid the loss of any crucial data regarding county populations with especially high stomach cancer mortality rates. By keeping the outliers, the full range of data is retained, allowing for a more comprehensive analysis of potential risk factors and correlations. Additionally, the outliers are evenly distributed, suggesting that they are not due to random errors or anomalies but rather represent real variation in the dataset that is important for understanding the full context of stomach cancer mortality.

Figure 4.1

Boxplot of Stomach Cancer Mortality Rates (Separated out by Race, Sex, and Age Group) Across U.S. Counties.

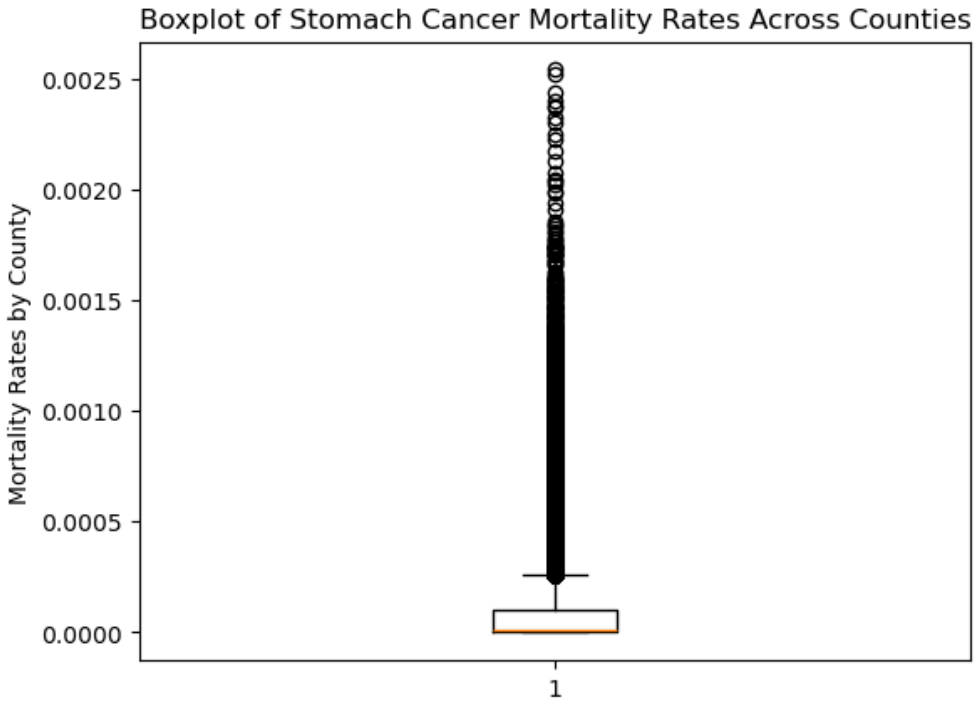
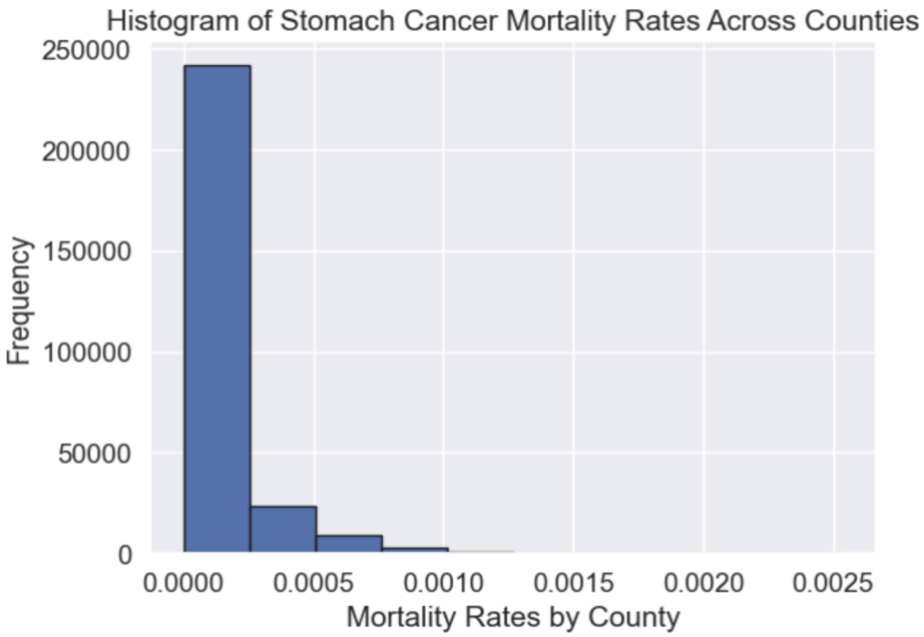


Figure 4.2

Histogram of Stomach Cancer Mortality Rates (Separated out by Race, Sex, and Age Group) Across U.S. Counties.

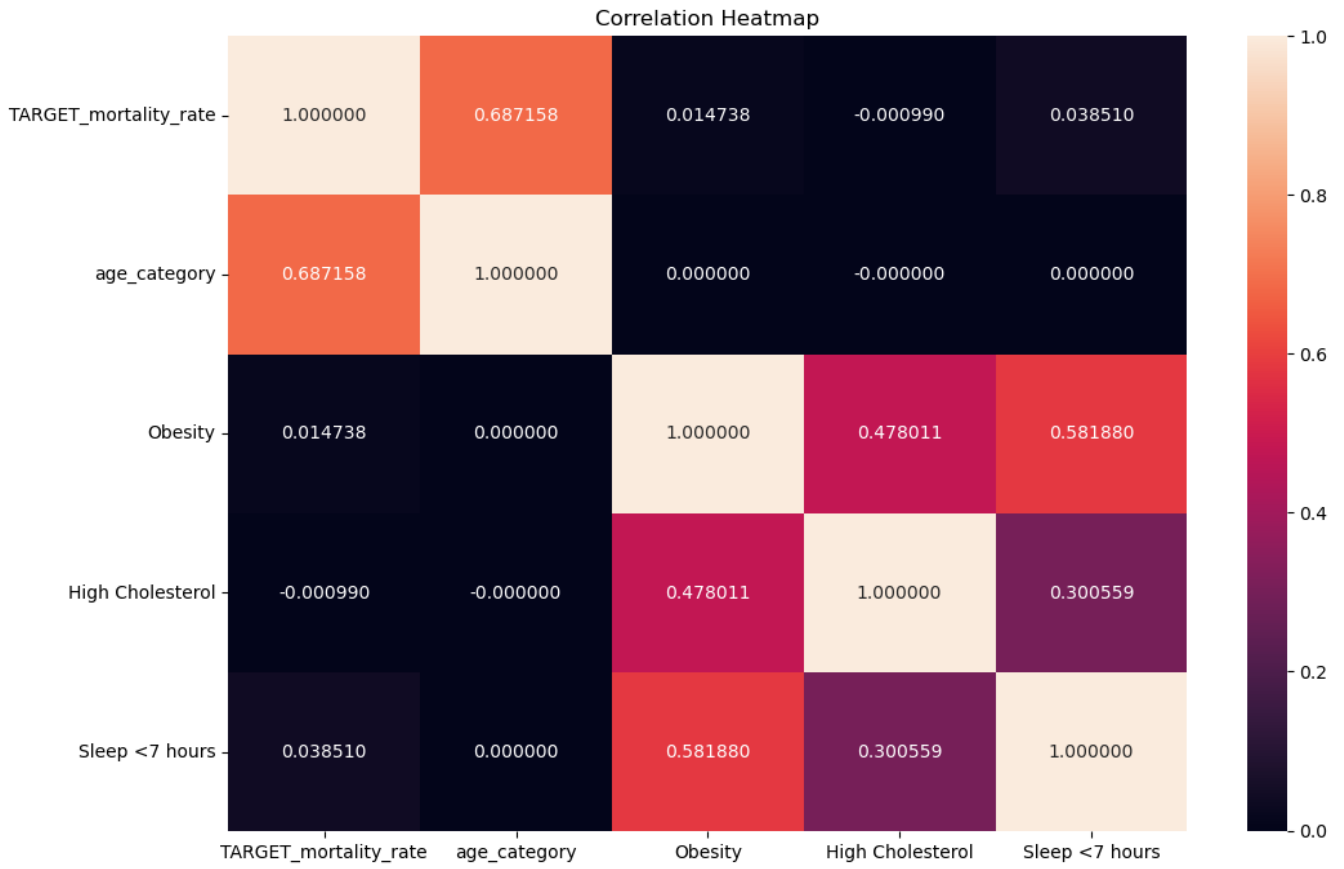


In our multivariate non-graphical and graphical analysis, the relationship between stomach cancer mortality rates and a select few other variables are analyzed. The results were displayed both in a table and through a correlation heatmap created using the Seaborn library. The heatmap provided a visual representation of the correlations, making it easier to see if there are relationships.

One easily notable finding from the heatmap was the strong positive correlation (0.69) between stomach cancer mortality rates and the age group within a county. This shows that age is a significant factor that is associated with stomach cancer mortality rates, which helps to highlight the importance of demographic factors like age in understanding the distribution of health outcomes. Obesity and high cholesterol were also moderately correlated with stomach cancer mortality rates, highlighting the importance of considering all aspects of patients' health history and correlated metrics, not simply their oncology care. Figure 4.2 depicts the correlation heatmap of the target variable (stomach cancer mortality rates) and select features.

Figure 4.3

Correlation Heatmap of Target Variable (Stomach Cancer Mortality Rate) and Select County-Level Demographic and CDC Health Measures Features

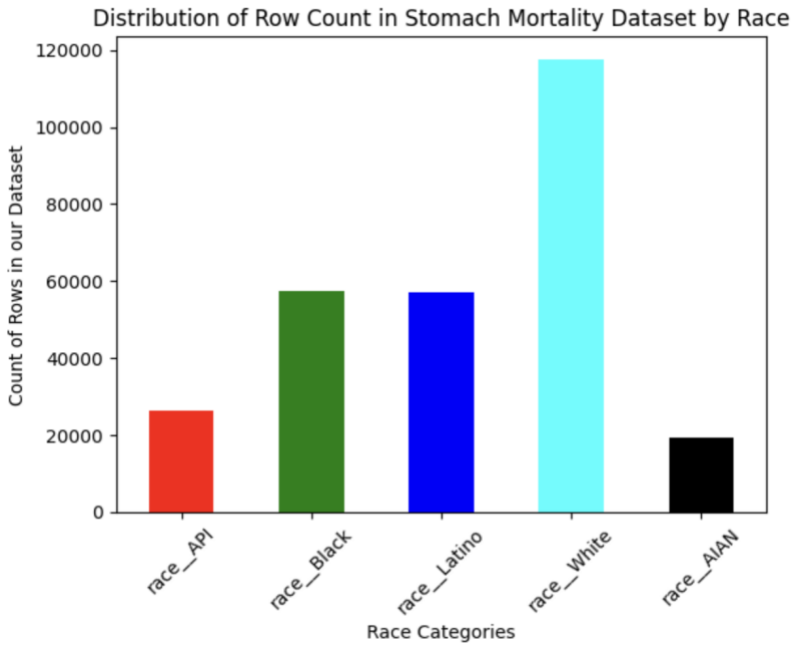


4.3.3 Distribution Analysis

In the distribution analysis, the availability of data for various racial groups within the dataset was examined. The dataset is split into stomach cancer mortality rates by county, age, sex, and race. However, certain counties may not report mortality rates for all five available race categories if their counties do not have a sufficient population or data for certain race groups. Thus, the row count by race in the dataset was graphed. The results (Figure 4.3 below) showed that the highest row count was for mortality rates in White populations. This was followed by Black and Latino populations, who had very similar counts. The next highest row count was Asian Pacific, and the smallest row count was American Indians. It is worth noting this disparity in data availability between racial groups.

Figure 4.4

Distribution of Row Count/Observations in the Stomach Cancer Mortality Dataset by Racial Categories



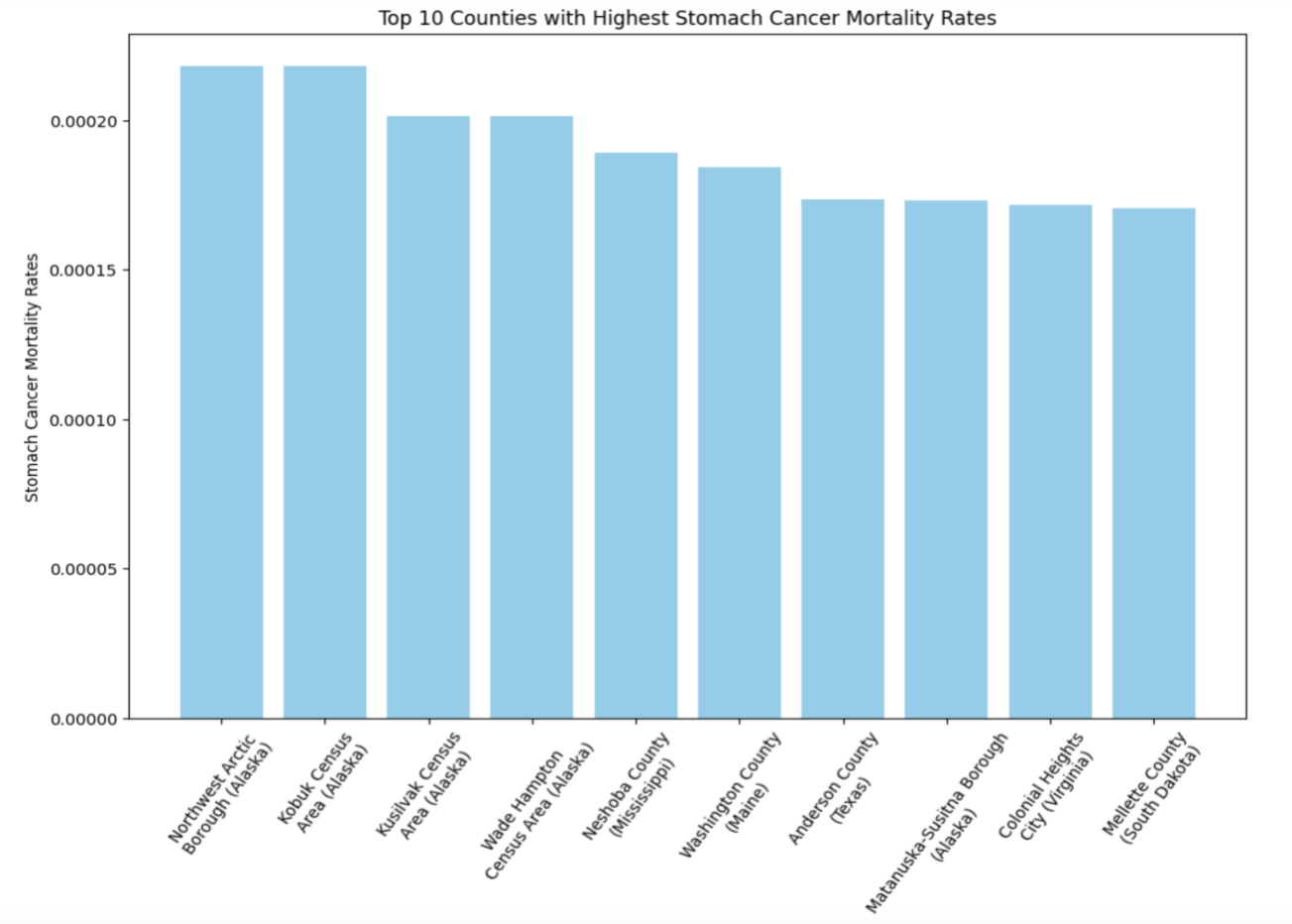
Next, the availability of data by sex and age group are checked. The distributions for these variables were completely even. These evenly distributed variables are from the dataset being provided to us at a county level, with an even split across the age and sex demographic variables.

4.3.4 County level Analysis

Our county-level analysis was conducted to find which counties had the highest prevalence of specific adverse health conditions or metrics, starting with our target variable: stomach cancer mortality rates. Of the top five counties with the highest stomach cancer mortality rates, four were in Alaska (with the Northwest Arctic Borough at the top) and the fifth was Neshoba County in Mississippi. Figure 4.4 below depicts the top 10 counties with highest stomach cancer mortality rates in 2019 across the United States.

Figure 4.5

Top 10 Counties with Highest Stomach Cancer Mortality Rates in 2019 Across the U.S.



In addition to stomach cancer rates, other key health metrics (obtained from the CDC dataset) were investigated, including obesity, high cholesterol, binge drinking, depression, and sleep duration of less than seven hours. This detailed analysis allowed the identification of counties with significant health challenges across multiple areas.

4.4 **Data Preprocessing**

Prior to modeling, additional data transformations were required in the consolidated dataset. These transformations were reserved for after exploratory data analysis so that any columns dropped or nulls imputed would not interfere with the analysis of the data values or availability as-is.

4.4.1 Modeling Approaches

In the consolidated dataset, there are 352 total features, with 309 of those being SDOH features. To make the dataset more manageable to work with and interpret for public health leaders or subject matter experts prior to deployment, two modeling approaches will be conducted:

1. In addition to the age, gender, sex, and 36 CDC health measures features, only a subset of the SDOH features will be selected for the first modeling feature set. This subset was chosen by selecting features past research has shown to be especially impactful on patient outcomes. However, as discussed in the Literature Review section, past research has often only considered a few of these features per study whereas the first modeling approach will consider 15 key SDOH features.
2. The dimensionality reduction technique Principal Components Analysis (PCA) will be used on all SDOH columns without significant nulls. The resulting columns will be combined with the age, gender, sex, and 36 CDC health metrics features for the feature set for the second modeling approach.

4.4.2 Preprocessing for Modeling Approach 1

4.4.2.1 Feature Selection

For the first modeling approach, the preprocessing was relatively straightforward. First, the location_name column was dropped as it was just a string column kept in for labels for the graphs in the aforementioned exploratory data analysis section. Following that, the target column, age, gender, race, CDC health measures columns, and a subset of the SDOH columns were selected. This subset of county-level SDOH features includes: total population, rate of poverty, percentage disabled, percentage of limited English-speaking households, percentage veterans, percentage single-parent households with a child, percentage of households with no computing device, percentage of households without internet, percentage of population 16+ that is unemployed, Gini index of income inequality, median household income, percentage of population 25+ with some college/associate's degree, percentage of households with no vehicle, percentage of population with Medicaid, and percentage of population that is uninsured.

After selecting the features, it was investigated if there were nulls in any rows, which would pose an issue in regression modeling. Of the 277,894 observations in the dataset, only 3.82% have nulls in any of the columns. Given how low that number is and how many observations will remain, the decision was made to simply drop those rows. The resulting feature set for Modeling Approach 1 has 267,292 observations and 58 columns.

4.4.2.2 Feature Scaling

After the features and rows are selected for Modeling Approach 1, a 75% training and 25% test split is performed for the data. This allocates 75% of the dataset to be used for training models that will be built and 25% for testing those models. Following the train-test split, the feature set was scaled using scikit-learn's StandardScaler package. This package transforms each feature by "removing the mean and scaling to unit variance," transforming the data so that each feature has a mean of zero and a standard deviation of one (Scikit Learn Guide, n.d.). The formula applied is shown in Equation 1, where Xi represents each individual observation's feature value, Xmean represents the mean for

that given feature's distribution, and Xstd represents the sample standard deviation for that given feature (using the dataset as the sample).

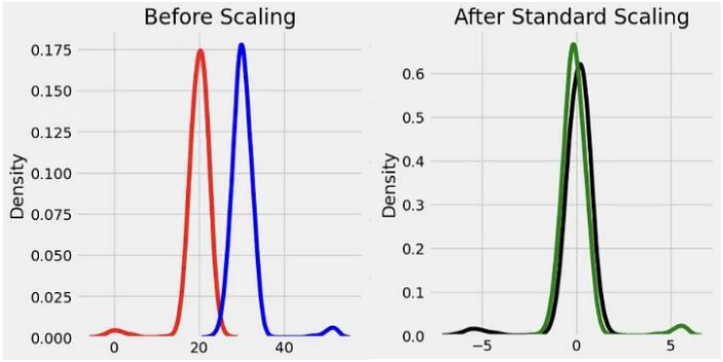
$$X_i' = \frac{X_i - \mu}{\sigma} = \frac{X_i - X_{\text{mean}}}{X_{std}} \quad (1)$$

Equation Image Obtained From: Cosgun, 2023

This scaling ensures that features with larger scales do not dominate the model, leading to a more stable and effective regression model. Given that a regression model will be built with the goal of predicting stomach cancer mortality rates, scaling is critical to ensure features of larger scales do not dominate the model, the model remains more stable overall, and future potential regularization does not disproportionately impact certain features. Figure 4.5 below demonstrates how the StandardScaler package scales all features, so they have approximately the same mean (zero) and standard deviation.

Figure 4.6

Data Normalization and Standardization Visualized on a 2D plot (Images Below from Cosgun, 2023)



When scaling our features, it is crucial to note that the StandardScaler transformer is fit only on the features for the training dataset. Once it is fitted, it is then used to transform the features for both the training and test dataset. The StandardScaler transformer is never fit on the test dataset to avoid any data leakage, which would artificially improve model performance. Following the train-test split and scaling, the training feature dataset is 200,469 rows and 58 columns and the test feature dataset is 66,823 rows and 58 columns.

4.4.3 Preprocessing for Modeling Approach 2

4.4.3.1 Feature Selection

The goal of this preprocessing approach is to use dimensionality reduction to make the sheer number of SDOH columns easier to work with. However, Principal Components Analysis (PCA) does not work on a dataset with null values, so these must first be handled. Of the 277,894 total observations in the data, 3.76% have no data for any SDOH columns or CDC Health Measures

columns because data for those counties was present in the stomach cancer mortality dataset, but not in either of the two other datasets. These rows are dropped as there is very little that can be done with them given an almost entirely missing feature set. The resulting feature set has 267,444 rows and 352 columns.

Following the removal of those 3.76% of observations, an investigation was conducted into which SDOH columns have remaining null values. Many of those columns are race-related columns (for example, the percentage of the population in a county that is Asian). The stomach cancer mortality dataset already captures race data, so these columns can be comfortably dropped. The few other columns with many nulls are: (a) four columns regarding how long the commute workers in a county must take to work on public transportation and (b) two columns regarding the median income of grandparent-led households. Given the number of nulls in these two column categories and that there are many other well-populated columns within the SDOH dataset regarding general median household income and how many workers do not have access to cars, these columns are also dropped. After these drops, the resulting feature set has 267,444 rows and 317 columns.

4.4.3.2 Feature Scaling

Following feature selection, feature scaling is done in the same fashion that it was done for Modeling Approach 1. A 75% train and 25% test split is done on the data, then a new instance of the StandardScaler transformer is fit on only the training data. Following that, both the training and test datasets for Modeling Approach 2 are transformed using that StandardScaler transformer.

4.4.3.3 Imputation of Remaining Nulls

After the aforementioned steps, there are still a few SDOH columns with null values, but the maximum null values in any one column is 950, which is only 0.36% of the total observations. Given how low that number is, the decision was made not to drop those rows (as that would result in the loss of valuable data), but rather to use the scikit-learn package KNNImputer to impute those nulls based on closely related observations within the SDOH dataset. Because many SDOH values are closely related, using a k-nearest neighbors (KNN) based modeling approach to impute values is logical. Within the KNNImputer, the hyperparameter for the number of neighbors considered during imputation is set to five to avoid the work becoming too computationally expensive, especially considering the number of features being worked with. Similarly to scaling, the KNNImputer is fit only on the training data and then used to transform both the training and test datasets to avoid any data leakage.

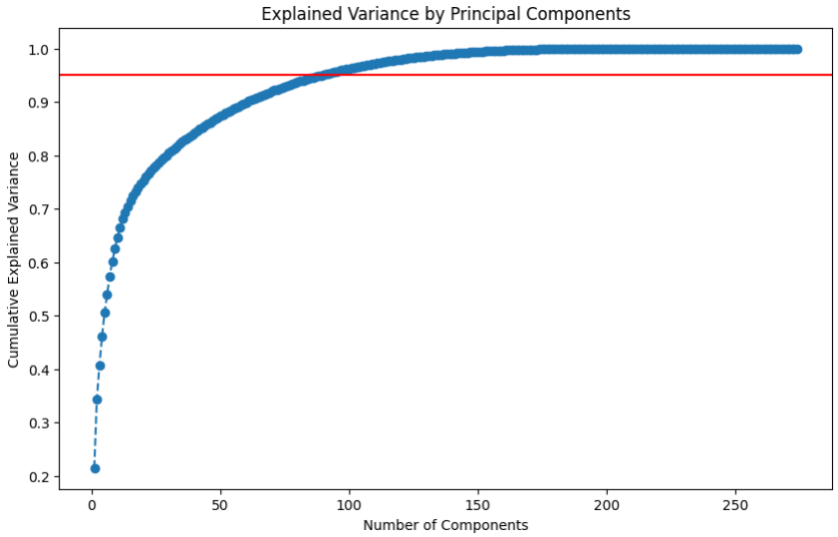
4.4.3.4 Conducting PCA

Following the handling of all null values in the dataset, the aim is to keep all the age, gender, race, and 36 CDC Health Measures features, but reduce the dimensionality of the SDOH features using Principal Components Analysis (PCA). The original SDOH dataset had 309 features, but that number fell to 274 after certain columns were removed to handle null values in the preceding section. Starting with those 274 columns in the training dataset, PCA is performed with the goal of reducing

down to only the number of components necessary to explain 95% of the variance in the 274 SDOH columns. To determine the components necessary for that, the PCA transformer is fit on just the 274 training dataset SDOH columns and the cumulative explained variance ratio is plotted against the number of components. The resulting graph (Figure 4.6 below) and numerical calculations indicate that 90 components are required to explain 95% of the variance among the 274 SDOH columns.

Figure 4.7

Cumulative explained variance by number of principal components when conducting PCA on 274 SDOH columns



Although 90 may still sound like quite a few components, (a) it is less than a third of the 274 SDOH columns prior to PCA and (b) because there are 200,583 training observations, it is possible to have many features in the feature set while still being well above the well-established guideline of maintaining at least 10 observations for every 1 feature in a modeling dataset (Peduzzi et al., 1996). Thus, PCA is conducted on the 274 SDOH columns to bring it down to 90 components.

Similarly to the scaling and imputation, the PCA transformer is fit on just 274 SDOH training data columns and then applied to both those and the 274 SDOH test data columns. The resulting 90 components for both are concatenated back to the age, gender, race, and CDC health measure training and test features. Following all transformations, the training feature dataset is 200,583 rows and 133 columns and the test feature dataset is 66,861 rows and 133 columns.

4.4 **Modeling**

At this point, the data is already fully preprocessed and split into training and test datasets for Modeling Approach 1 and Modeling Approach 2. GridSearchCV is then used to initialize a series of models and hyperparameters to loop through to identify the optimal model for each modeling approach. Table 4.1 below lists all models and hyperparameters considered. All of the models in consideration that take a random state parameter are initialized with random_state=42 to allow for consistent, reproducible results.

Table 4.1
All Models and Hyperparameters Tested in our Modeling, along with if We Set a Random State for Each

Model	**Hyperpara-meters Considered**	**Random State?**
Linear Regression	N/A	No
Ridge	alpha: [0.1, 1.0, 10.0]	Yes, 42
Lasso	alpha: [0.1, 1.0, 10.0]	Yes, 42
ElasticNet	alpha: [0.1, 1.0, 10.0], l1_ratio: [0.1, 0.5, 0.9]	Yes, 42
Extreme Gradient Boosting Regressor (XGBRegressor)	n_estimators: [50, 100, 200], learning_rate: [0.01, 0.1], max_depth: [10, 20, 30]	Yes, 42
Light Gradient Boosting Machine Regressor (LightGBM Regressor),	n_estimators: [50, 100, 200], learning_rate: [0.01, 0.1], max_depth: [10, 20, 30]	Yes, 42
Support Vector Regression (SVR)	C: [0.1, 1, 10], epsilon: [0.01, 0.1, 1]	No

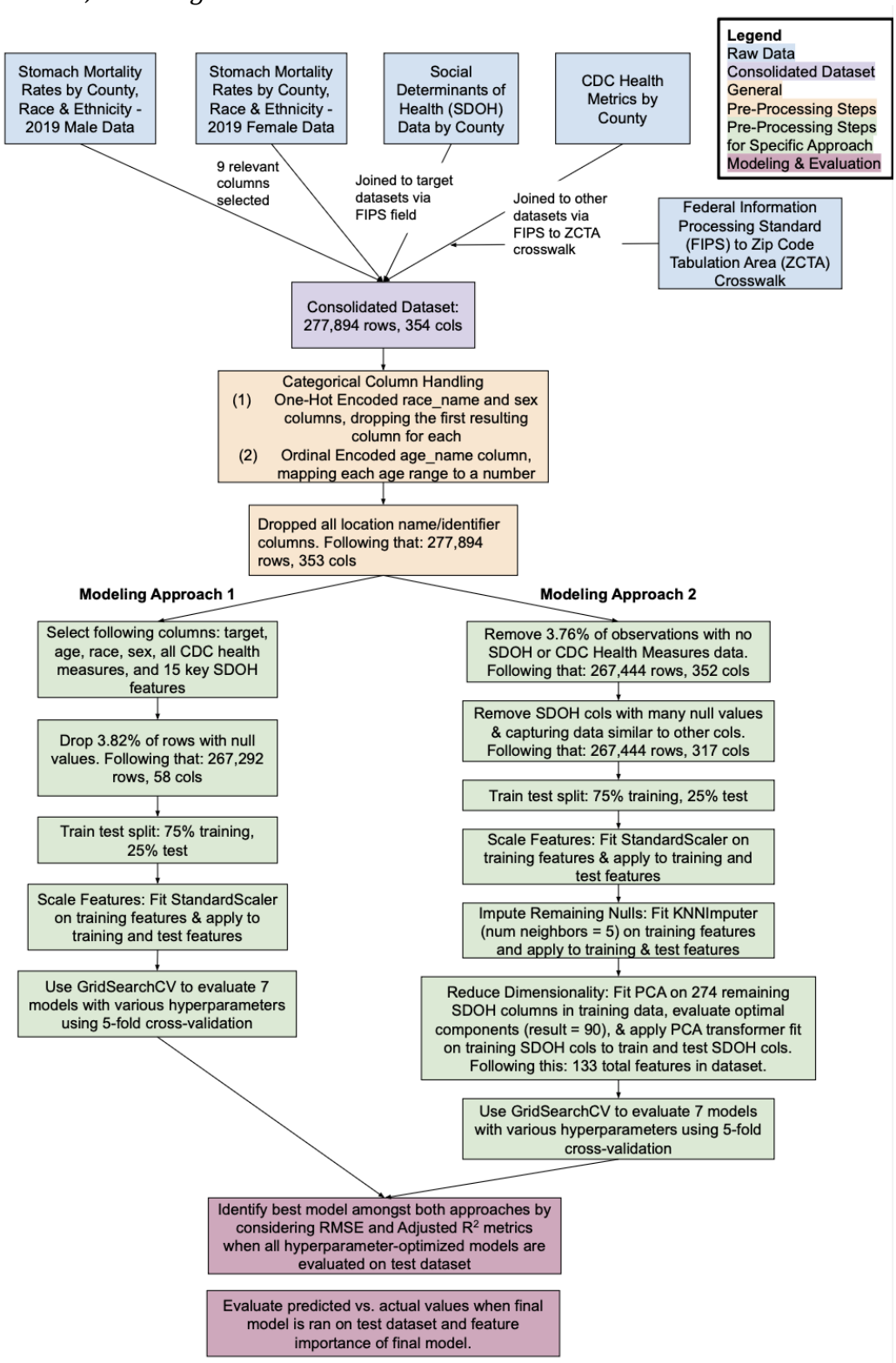
For each model type (listed in the furthest left column in Table 4.1), all combinations of hyperparameter options for that model were tested using the training datasets for both Modeling Approach 1 and Modeling Approach 2. To determine the best hyperparameters for each model type (within each modeling approach), five-fold cross-validation was used with root mean squared error (RMSE) as the evaluation metric. RMSE calculates the variance between actual observed data values and values the model predicted for those observations. RMSE was selected as the evaluation metric

because it is easier to interpret than other metrics due to having the same units as the target variable and because it penalizes larger errors more than smaller errors. In the context of predicting stomach cancer mortality, attention must be paid to the magnitude of a prediction error, as a larger error would mean more resources are erroneously allocated (whether in excess or lacking).

After the optimal hyperparameters were identified for each model type (within each modeling approach) using cross-validation within the training dataset, the RMSE values were evaluated for each optimized model when applied to their respective test sets. The modeling approach, model, and hyperparameters that result in the lowest RMSE were chosen as the final approach and model. To summarize all of the methodology discussed in this section, Figure 4.7 is provided below.

Figure 4.8

Summary of Methodology: Joining Datasets, Handling Nulls, Scaling Features, Dimensionality Reduction, Modeling and Evaluation



5 Results and Findings

5.1 Model Performance: RMSE

All models mentioned in Table 4.1 were trained twice, once on the training dataset generated from Modeling Approach 1 and once on the training dataset generated from Modeling Approach 2. Five-fold cross-validation within each training dataset was used to identify the optimal hyperparameters for each model within each modeling approach, then the RMSE was calculated on the training and test dataset. Table 5.1 below shows the resulting optimized hyperparameters and RMSEs for both modeling approaches.

Table 5.1
Results from Grid Search through Hyperparameters using Five-Fold Cross-Validation. For each Modeling Approach's Training Dataset, the Optimal Hyperparameters and RMSE Values are Shown.

Model	Approach 1 Optimal Hyperpara-meters	Approach 1 Training RMSE	Approach 1 Test RMSE	Approach 2 Optimal Hyperpara-meters	Approach 2 Training RMSE	Approach 2 Test RMSE
Linear Regression	N/A	0.0001238	0.0001238	N/A	0.0001237	0.0001238
Ridge	alpha: 10.0	0.0001238	0.0001238	alpha: 10.0	0.0001237	0.0001238
Lasso	alpha: 0.1	0.0001778	0.0001782	alpha: 0.1	0.0001779	0.0001783
ElasticNet	alpha: 0.1, l1_ratio: 0.1	0.0001778	0.0001782	alpha: 0.1, l1_ratio: 0.1	0.0001779	0.0001783
XGB Regressor	learning_rate: 0.1, max_depth: 20, n_estimators: 20	0.0000480	0.0000484	learning_rate: 0.1, max_depth: 20, n_estimators: 20	0.0000478	0.0000491
LightGBM Regressor	learning_rate: 0.1, max_depth: 20, n_estimators: 20	0.0000311	0.0000359	learning_rate: 0.1, max_depth: 20, n_estimators: 20	0.0000300	0.0000355

Support Vector	C: 0.1, epsilon: 1	0.0011862	0.0011862	C: 0.1, epsilon: 1	0.0011862	0.0011863

When interpreting the results in Table 5.1 above, it is key to recall that RMSE is in the same units as the target variable, so the RMSE shows, on average, how far off the model's predictions are from the observed values. For example, an RMSE of 0.0001238 means that, on average, that model's predictions deviate from the actual observed stomach cancer mortality rates by about 0.0001238. Although this number sounds very small, it is important to recall that the mean stomach cancer mortality rate observed in our entire dataset (prior to dropping any rows or the train-test splits) was 0.00009633, so even very small RMSEs are quite impactful. Given that context, none of our models performed very well.

Overall, model performance did not change drastically between the two modeling approaches used. Additionally, for both modeling approaches, the Light Gradient Boosting Machine (LightGBM) with a learning rate of 0.1, maximum depth of 20, and number of estimators equal to 20 performed best. When trained on Training Dataset 1 and tested on Test Dataset 1 (both from Modeling Approach 1), the resulting RMSE was 0.0000359. When trained on Training Dataset 2 and tested on Test Dataset 2 (both from Modeling Approach 2), the resulting RMSE was 0.0000355, which was just slightly better than the resulting from Approach 1.

5.2 Model Performance: Adjusted R^2

In addition to comparing the RMSE of the optimal model from both approaches, the adjusted R2 metric of both optimized LightGBM models was also evaluated. The standard, non-adjusted R^2 metric of a model explains how well a model's features explain the variability in the target variable (Chugh, 2020). Although this is useful, one issue is the non-adjusted R^2 always increases or remains the same as the number of features in the model increases because adding more features will never decrease the amount of variation explained in the target variable by those features (Chugh, 2020). Thus, the standard R^2 metric is not a good one by which to adjudicate if additional features are useful in a model or not.

In contrast to the standard R^2 metric, the adjusted R^2 metric takes into account the number of features that a model has. The formula to calculate the adjusted R^2 is included below (Equation 2) where *n* is the number of observations in the test dataset and *k* is the number of features in the dataset.

$$R^2_{adj} = 1 - \left[\frac{(1 - R^2)(n-1)}{n-k-1} \right] \quad (2)$$

Equation Image Obtained From: Chugh, 2020

Because *k* is in the denominator, there is a penalty for adding additional features in if they do not significantly improve the R^2 value (Chugh, 2020). Thus, the adjusted R^2 metric is an optimal choice for comparing which of two models with a different number of features is a better performing model.

Given that our two modeling approaches have a significant difference in the number of features (58 features for Modeling Approach 1 versus 133 features for Modeling Approach 2), it is useful to use the adjusted R^2 metric in addition to RMSE to compare the performance of the LightGBM models from both approaches. The adjusted R^2 metric for the LightGBM model from Modeling Approach 1 is 0.9595. The adjusted R^2 metric for the LightGBM model from Modeling Approach 2 is 0.9607. It is good to see that the adjusted R^2 metric for Modeling Approach 2 is, in fact, higher because it reaffirms the value in including all the SDOH columns via PCA in our dataset of 133 features.

5.3 **Model Feature Importance**

Because different features were used for Modeling Approach 1 versus 2 and both resulted in strong LightGBM regression models, the feature importance of the LightGBM models for both was analyzed. Figure 5.1 below shows the top 15 features for the LightGBM model trained on the feature set for Modeling Approach 1, whereas Figure 5.2 shows the top 15 features for the LightGBM model trained on the feature set for Modeling Approach 2.

Figure 5.1
Top 15 features for LightGBM Regression Model Trained on Training Dataset for Model Approach 1

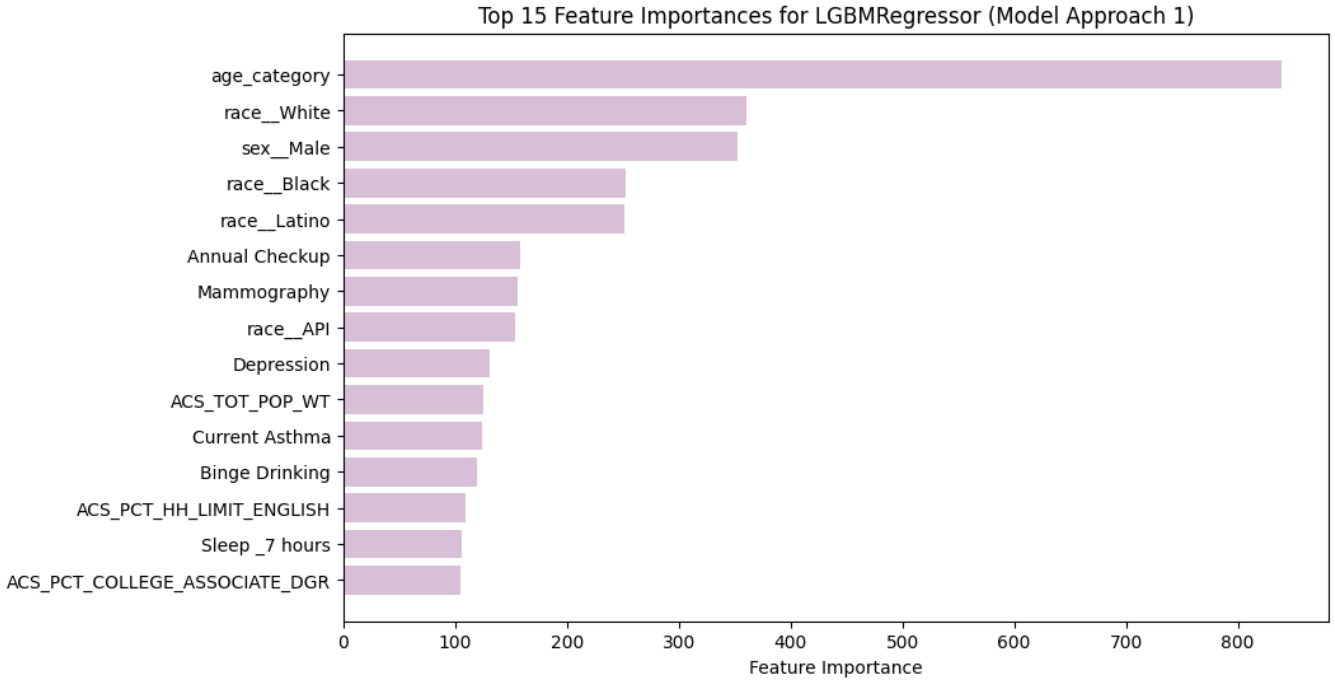
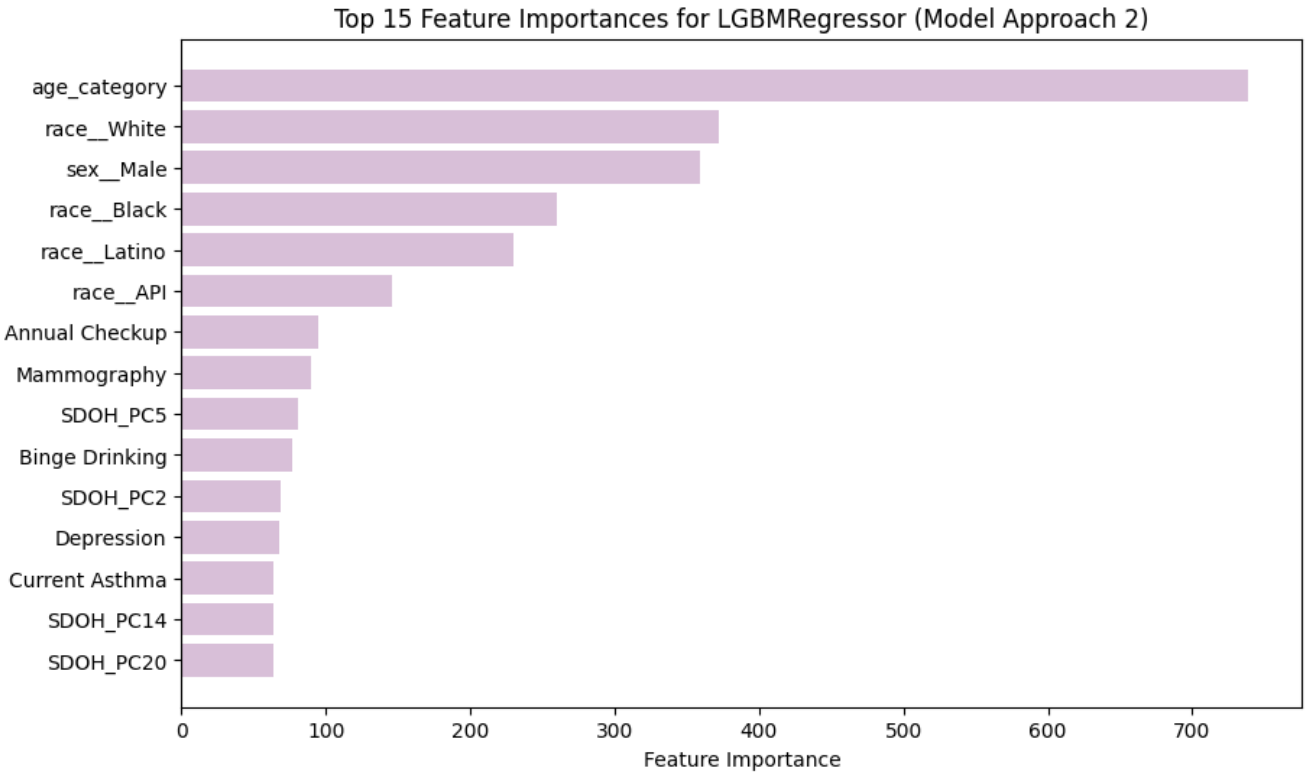


Figure 5.2
Top 15 Features for LightGBM Regression Model Trained on Training Dataset for Model Approach 2



Based on the two above graphs, it is clear that age category and racial group are the two most predictive features regarding stomach cancer mortality rates. This makes sense not only because they are generally predictive healthcare features, but also because they are the only features in our dataset that came with the target data. Thus, the mortality rate for each row is associated specifically with the age group and race of that row. On the other hand, the other features report general aggregated measures and metrics for the county at hand that are not specific to stomach cancer patients or their mortality rates. Thus, it makes sense that these general, county-level features will be less predictive.

Following the age group and race features, Annual Checkup and Mammography (which indicate the percentage of a county's population that has had an annual checkup or mammography within an appropriate time period) are important features in both models. This makes sense because preventative care is often critical for catching cancers early and avoiding mortality, so counties with higher rates of preventative care most likely have lower rates of stomach cancer mortality. Although this graph does not show us if a feature has a positive or negative impact on the prediction value, our hypothesis is that preventative care decreases stomach cancer mortality rates. On the other hand, Binge Drinking, Depression, and Current Asthma (which indicate the percentage of a county's population that faces those issues) are also shared amongst both models, but likely indicators of

higher rates of mortality due to the fact that they are comorbidities that may add complications to cancer treatment.

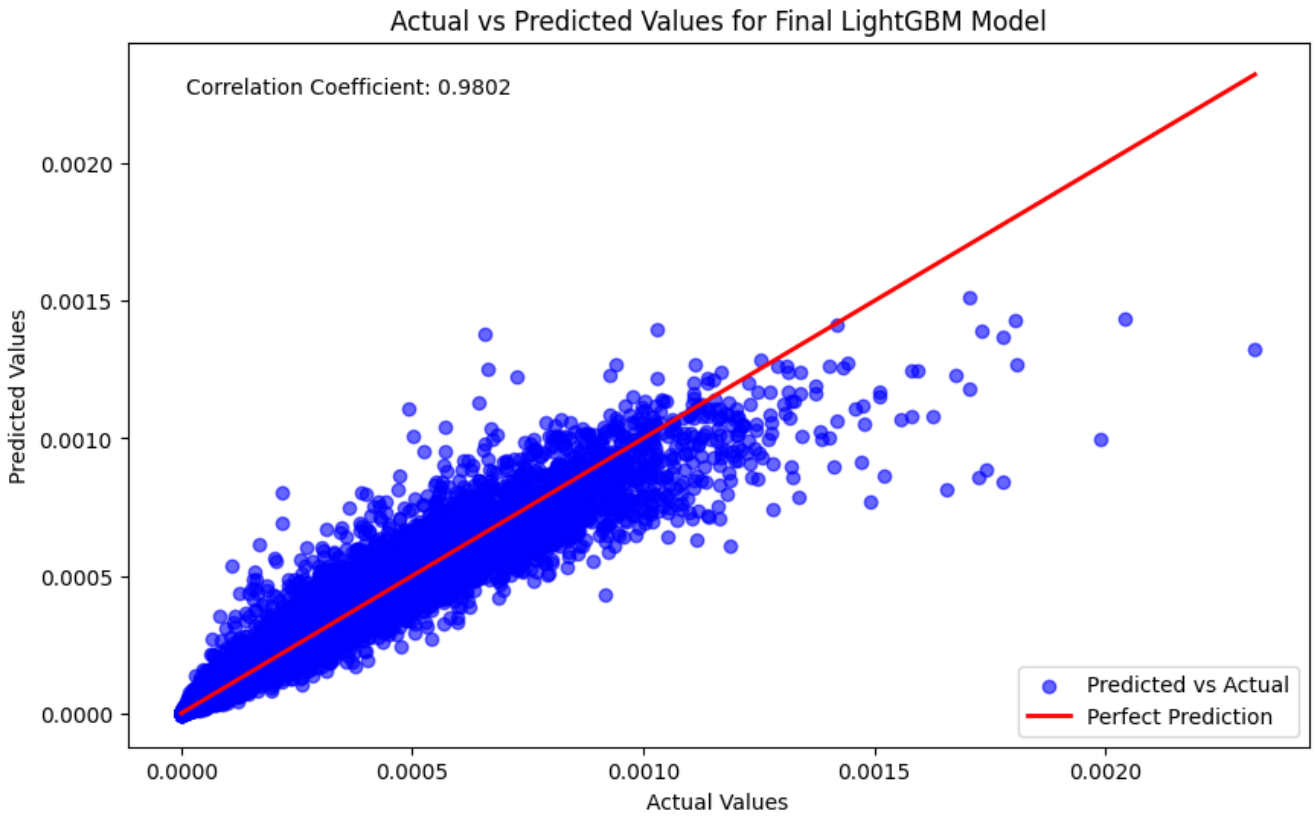
In terms of crucial SDOH factors, the feature importance for Modeling Approach 2 will not be helpful because there is no visibility into which underlying SDOH features are within each SDOH PCA component. However, from the feature importance graph for Modeling Approach 1, it is evident that the top three most predictive SDOH factors were the total weighted population of a county, the percentage of a county's population that speaks limited English, and the percentage of a county's population with a college degree (including an associate's degree). These are all logical key factors as county population impacts the availability of resources for individuals, English ability impacts accessibility of healthcare, and college degrees are often associated with increased understanding of how to navigate the healthcare system and increased wealth to be able to do so.

Although both feature importance graphs were insightful, the decision was made to hereafter use only the LightGBM model trained on the data from Modeling Approach 2 as the final model, due to its slightly lower test set RMSE and slightly higher test set adjusted R2 value.

5.4 **Final Model: Comparison of Predicted vs. Actual Values**

For the final model (the LightGBM model trained on data from Modeling Approach 2), an investigation was conducted into how the predicted values for the test dataset compared to the actual, observed values for the test dataset. Figure 5.3 below shows the actual values in the test dataset on the x-axis versus the values predicted by our LightGBM model on the y-axis. The red line denotes what a perfect prediction result would have looked like with a slope of exactly one as every predicted value would have been the exact same as the actual value. The correlation coefficient of the predicted versus actual values is included in the top left (0.9802, which is quite high).

Figure 5.3
Comparing the Actual versus Predicted Values in the Test Dataset for the Final LightGBM Model



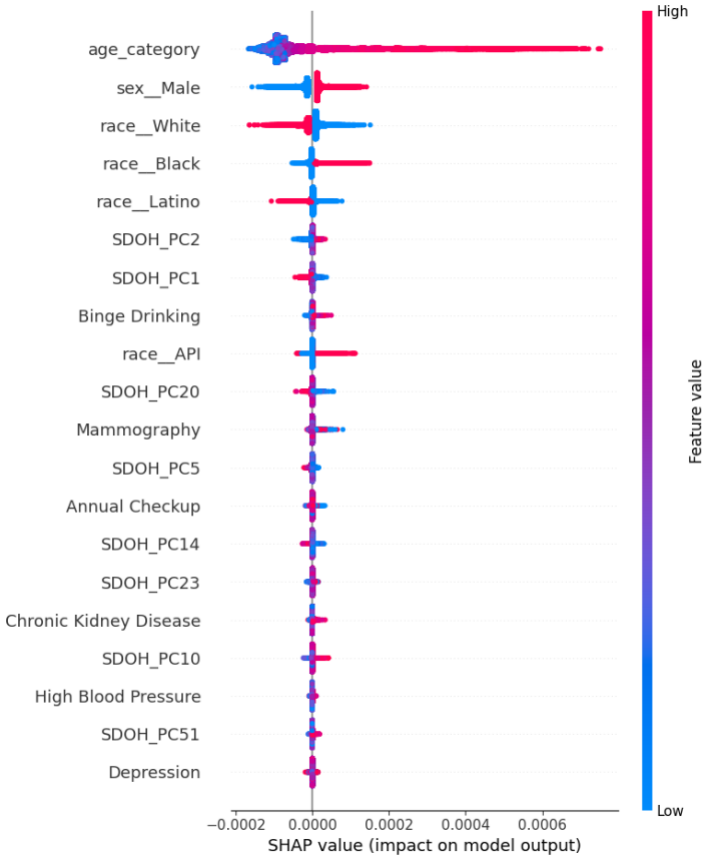
In the plot above, it is evident that the model does better at predicting lower stomach mortality rates but, as the actual values get higher, the model begins to predict a bit lower than the actual values. However, it is crucial to note that most of those values are high outliers whereas the crux of the data is in the area where the model predicted fairly well.

5.4 **Final Model: Delving into Shapley Additive Explanations Values**

In order to further investigate how our model operates and provide increased model explainability, Shapley Additive ExPlanations (SHAP) values were used. "The core idea behind Shapley value-based explanations of machine learning models is to use fair allocation results from cooperative game theory to allocate credit for a model's output *f(x)* among its input features" (Lundberg, 2018). SHAP values use game theory to explain the impact of each individual feature on a model's predictions, either globally across all features or locally on individual predictions. When looking at an individual prediction made by the model, SHAP values are additive so "SHAP values of all the input features will always sum up to the difference between baseline (expected) model output and the current model output for the prediction being explained" (Lundberg, 2018). While SHAP values can be very useful in interpreting a model, its creators warn us not to use it to assign numerical causality as confounding variables that are unaccounted for can always be present in our models. However, SHAP values still remain a great tool for taking a look under the hood of our model to estimate each feature's contribution.

LightGBM is a tree-based machine learning model, so we used the TreeExplainer method within the SHAP Python package to calculate the SHAP values for our final model. The TreeExplainer "is a fast and exact method to estimate SHAP values for tree models and ensembles of trees, under several different possible assumptions about feature dependence" (Lundberg, 2018, TreeExplainer). Following the calculation of our final LightGBM model's SHAP values using TreeExplainer, Figure 5.4 below was generated to show the global impact of the model's features on predictions.

Figure 5.4
Global SHAP Values for Final LightGBM Model



While the y-axis of Figure 5.4 looks quite similar to Figure 5.2 (our feature importance graph), it has the added benefit of showing us not only the magnitude of feature impact, but also the direction. The color gradient from blue to red represents the feature values, where blue indicates lower values and red indicates higher values. Thus, the higher age categories (represented as redder points) are associated with a larger positive impact on the model prediction, meaning older patients are predicted to have higher stomach cancer mortality rates. Similarly, being male is associated with an increase in predicted mortality rate whereas being White (in terms of race) is associated with a decrease in predicted mortality rate.

SHAP values also allow for an analysis of how features interact with each other. Figure 5.5 below demonstrates the interaction between age and whether patients are Black. The x-axis shows the

age category feature, and the y-axis represents SHAP values, indicating that SHAP values (representing impact on predicted mortality rates) non-linearly increase as patient age increases. The color gradient for race adds another dimension. On the right side of the plot, where the largest SHAP values are observed (in older patients, indicating the greatest risk), being Black further increases the SHAP values, as the red points are clustered at the top of the bars on the right side of the plot. Thus, being Black amplifies the predicted stomach cancer mortality rate in older patients, demonstrating an insightful interaction between age and race in the model.

To provide a contrasting plot, Figure 5.6 below shows the relationship between age and whether counties have high rates of binge drinking. Unlike in Figure 5.5, Figure 5.6 shows no clear color differentiation on the bars anywhere in the plot. Therefore, age and county levels of binge drinking do not interact with each other in the model, meaning whether a county has a high percentage of binge drinkers does not significantly alter the effect of age on the predicted stomach cancer mortality rate in that county.

Figure 5.5
Dependence Plot Showing How Age Category and Race (Black) Features Impact SHAP Values

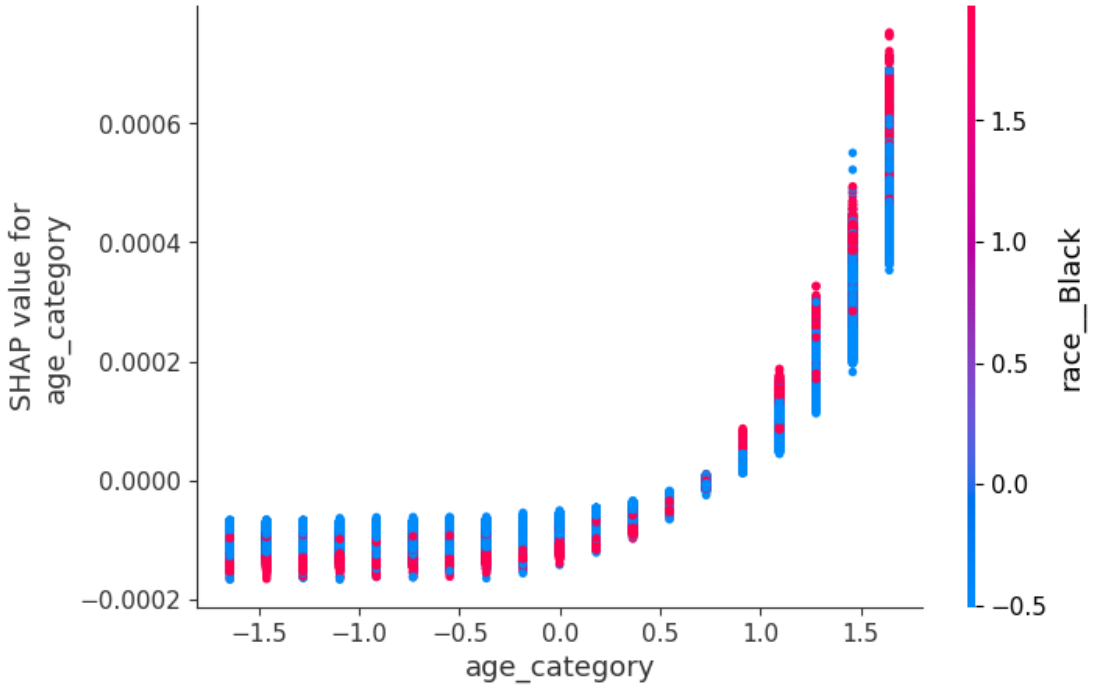
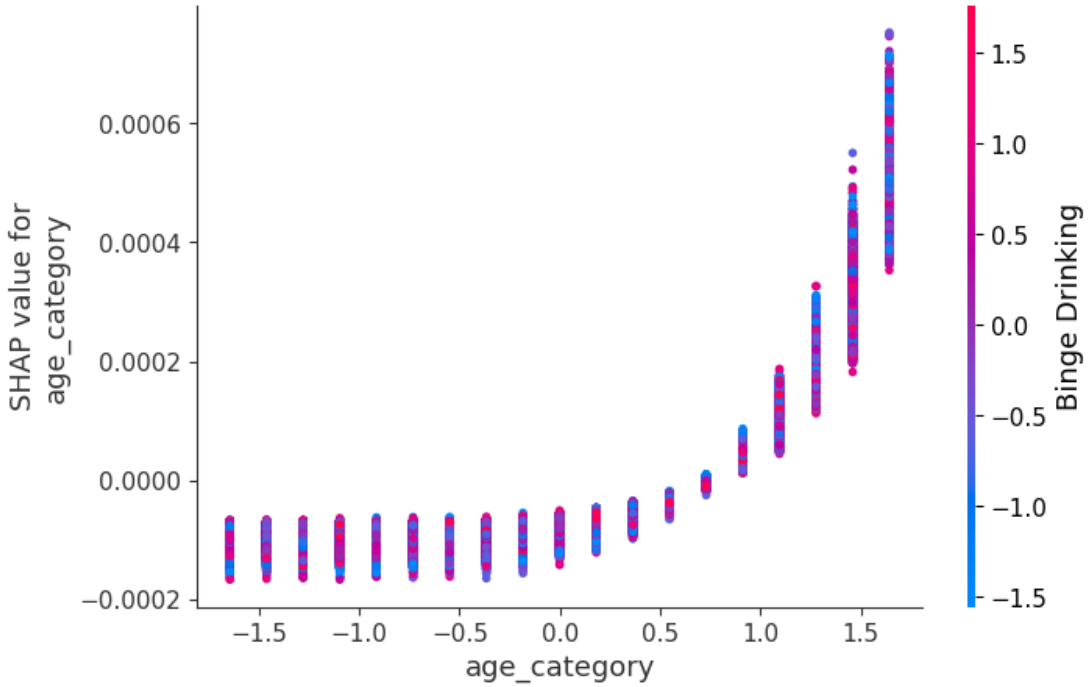


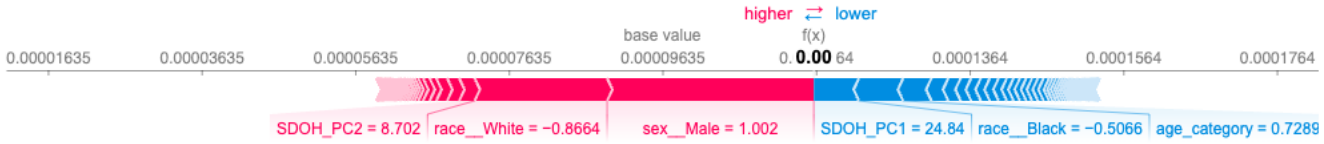
Figure 5.6
Dependence Plot Showing How Age Category and Binge Drinking Rates in a County Features Impact SHAP Values



Finally, one of the greatest strengths of SHAP values is the ability to dive into feature impact on individual predictions. When using our model to inform public health decisions in the future, being able to investigate the exact features impacting a prediction will be useful for public health officials to validate the prediction and determine how best to intervene to decrease mortality rates. Figure 5.7 below demonstrates the features with the largest SHAP value contributions to the prediction for a randomly selected observation within our test dataset to demonstrate this capability to drill down into local feature importance.

Figure 5.7

SHAP Values for Observation Number 100 in the Test Dataset



On the left side of Figure 5.7, in the red, features and values that increased the predicted stomach cancer mortality rate for the given county population are shown. Those features include: a positive value for sex_Male, a negative value for race_White, and a positive value for SDOH_PC2. On the right side, in blue, features and values that decreased the predicted stomach cancer mortality rate for the given county population are shown. Those include: a positive value for SDOH_PC1, a negative value for race_Black, and a positive (but very low) value for age_category. This means that, for this specific population within this specific county, patients being male, not being White, and the county having

a high SDOH_PC2 value pushed the predicted stomach cancer mortality rate higher, whereas a high county value for SDOH_PC1, patients not being Black, and patients being young pushed the predicted mortality rate lower. The actual predicted value for this specific population within this specific county was 0.000116, which is denoted by *f(x)* in Figure 5.7. Such analyses can be done for any given prediction our model makes, boosting its explainability and functionality.

6 Discussion

6.1 Conclusion

The final model obtained from our research was a Light Gradient Boosting Machine (LightGBM) Regressor model trained on 133 features. Those 133 features included all demographic and health equity features in the original feature set as well as 90 components obtained from conducting PCA on all of the SDOH features (except the few columns removed for comprising of a high proportion of null values). The final model had an RMSE of 0.0000355. Given the context that the mean of the target variable was 0.00009633, the RMSE is acceptable, but not exceptional. The RMSE is approximately 36.8% of the mean of the target variable (a comparison can be made as they are both in the same units), meaning the average prediction error on the test set was approximately 36.8% of the mean above or below the actual, observed values. Although that is acceptable, it is not particularly good in terms of model performance.

However, RMSE is just one metric measuring model performance and it is crucial to note that the model had an excellent adjusted R^2 score when tested on the test set, that being 0.9607. That indicates that approximately 96.07% of the variability in the target variable can be explained by the features in the model. This incredibly high adjusted R^2 value indicates that the model is performing quite well and the features used in the model are highly relevant in predicting the target variable. Furthermore, the use of the adjusted R^2 score rather than the standard R^2 score ensures that the model does not have excess features, but rather that the features included are, in fact, important in predicting stomach cancer mortality rates.

Furthermore, Figure 5.3 highlights that the correlation coefficient for the actual versus predicted values for the test set is 0.9802, with a lot of the poorer predictions happening due to high outliers. This extremely high correlation coefficient further reaffirms that the model performs well when predicting stomach cancer mortality rates by county populations.

Finally, it is crucial to mention that, even if the model only performs decently rather than exceptionally, its feature importance is very insightful. As highlighted in our introduction and literature review, similar past studies have often only been able to assess the impact of a subset of SDOH, demographic and/or health equity factors on oncology outcomes. The goal of our work was to conduct a much broader study evaluating a large cohort of factors to investigate which are most impactful on stomach cancer mortality rates. Upon investigation of the feature importance of our LightGBM model, it is found that key predictive factors include age, race, percentage of a county's

population that received preventative care (such as annual checkups and mammograms), and percentage of a county's population with comorbidities (such as binge drinking, depression, and current asthma). Regardless of exact model performance metrics, the model's feature importance (considering both magnitude and direction, as provided by the global SHAP values) offers crucial insights on areas public health officials may want to focus to decrease stomach cancer mortality rates. Furthermore, using SHAP values allows us to drill down into each specific prediction made, offering insight into the specific features that are driving up the stomach cancer mortality rates in populations with high predicted rates. That prediction-level insight allows public health officials and hospital networks to further tailor their programmatic interventions to the specific issues driving up stomach cancer mortality rates within populations in the counties they serve.

6.2 **Recommended Action Items**

The goal of this work is to improve healthcare resource allocation, foster the growth of public health programs, and decrease health inequity. To that end, a model was built that can be used to predict counties likely to have high stomach cancer mortality rates. As mentioned, while the model may not be perfect, it generally performs well with the correlation coefficient for the actual versus predicted values for the test set being 0.9802. Furthermore, it is critical to note that the model performs quite well at lower values but underestimates stomach cancer mortality rates a bit at very high values. To combat that, the model need not be used to predict exact stomach cancer mortality rates, but rather could just be used to predict the top 10-15% of counties likely to have mortality rates so that resources could be poured into those counties. For example, in Figure 5.3, although model performance drops a bit above approximately 0.0012 (for the actual, observed value), even if all counties with a predicted stomach cancer mortality rate above 0.0012 were segmented out and more resources were directed to them (in comparison to other counties), that would be beneficial.

Thus, it is recommended that public health officials set a cutoff based on resources available (for example, the top 100 county and age/race combinations, as the target data is separated by age and race) and then allocate additional resources to all populations above that cutoff, as determined by the model. Furthermore, it is recommended that public health officials and hospital networks allocate more research, funding, and case manager support to the top features shown in the model. For the preventative care aspect, this may look like conducting proactive outreach calling patients reminding them to come in for an annual checkup. For the comorbidities, programs to consider may include subsidized alcohol rehabilitation programs and increasing access to mental healthcare. Collaboration with public health officials and hospital networks is anticipated to determine how the model's outputs can be used to improve stomach cancer patient outcomes.

6.3 **Recommended Future Studies**

While this research is promising, certain limitations are acknowledged, and it is hoped future research can overcome them to improve upon this field. First, aggregated county-level data regarding SDOH and health equity metrics had to be used. Although this allows for making useful

conclusions about how best to allocate medical resources across counties, it does not offer the most information regarding the impact on the individual patient-level of SDOH and health equity barriers. To better understand the impact of these issues on individual patients and their stomach cancer mortality rates, it is recommended that a subset of future work be done at the patient-level where all features are representative of one individual rather than at an aggregate level. This would most likely need to be done on a much smaller scale than nationally but would still provide valuable insight.

An additional area of improvement is data quality, both in terms of the quality of data obtained, as well as how nulls are handled. In this research, certain nulls were handled by dropping entire rows (but keeping that percentage low to avoid serious data loss or a lack of representation in the dataset). Other nulls (specifically in the SDOH columns) were handled using KNN Imputation with the number of neighbors considered during imputation set to five to avoid the work becoming too computationally expensive. Due to the constraint of working on a local machine, future work should consider more robust imputation techniques to investigate if that improves model results.

Beyond the constraints on imputation methods, computational limitations also restricted the range of modeling techniques explored in this study. Conducting the research locally with limited computational resources presented challenges that may not exist for well-funded research organizations or companies. Future work should consider leveraging advanced computational resources to explore deep learning methods and additional hyperparameter tuning to enhance model robustness. It is anticipated that subsequent research in the public health field will build upon these findings, advancing efforts to improve stomach cancer patient outcomes and promote health equity.

References

Agency for Healthcare Research and Quality (AHRQ). (2023, Jun). *Social Determinants of Health Database.* Retrieved Jun 30, 2024, from https://www.ahrq.gov/sdoh/data-analytics/sdoh-data.html#download.

American Cancer Society. (2024, Jan 19). *Key statistics for stomach cancer.* Retrieved June 30, 2024, from https://www.cancer.org/cancer/types/stomach-cancer/about/key-statistics.html.

Bonner, S. N., & Edwards, M. A. (2024). *The impact of racial disparities and the social determinants of health on esophageal and gastric cancer outcomes.* Surgical Oncology Clinics of North America, 33(3), 595–604. https://doi.org/10.1016/j.soc.2023.12.015

Centers for Disease Control and Prevention (CDC). (2023, Aug 25). PLACES: Local Data for Better Health, ZCTA Data 2023 release. Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Division of Population Health. Retrieved Jun 30, 2024, from https://data.cdc.gov/500-Cities-Places/PLACES-Local-Data-for-Better-Health-ZCTA-Data-2023/qnzd-25i4/about_data.

Chaiyachati, K. H., Hubbard, R. A., Yeager, A., Mugo, B., Shea, J. A., Rosin, R., & Grande, D. (2018, Jan 29). *Rideshare-Based Medical Transportation for Medicaid Patients and Primary Care Show Rates:*

A Difference-in-Difference Analysis of a Pilot Program. Journal of General Internal Medicine, 33(6), 863–868. Retrieved Jul 1, 2024, from https://doi.org/10.1007/s11606-018-4306-0.

Chugh, A. (2020, Dec 8). *MAE, MSE, RMSE, Coefficient of Determination, Adjusted R Squared — Which Metric is Better?* Analytics Vidhya. Retrieved Aug 5, 2024, from https://medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e.

Cosgun, H.H. (2023, Aug 5). *Which data scaling technique should I use?* Medium. Retrieved Jul 29, 2024, from https://medium.com/@hhuseyincosgun/which-data-scaling-technique-should-i-use-a1615292061e.

Fan, Q., Keene, D. E., Banegas, M. P., Gehlert, S., Gottlieb, L. M., Yabroff, K. R., & Pollack, C. E. (2022). *Housing Insecurity Among Patients With Cancer.* Journal of the National Cancer Institute, 114(12), 1584–1592. Retrieved Jul 7, 2024, from https://doi.org/10.1093/jnci/djac136.

Ilic, M., & Ilic, I. (2022). Epidemiology of stomach cancer. *World Journal of Gastroenterology*, *28*(12), 1187–1203. Retrieved Jul 7, 2024, from https://doi.org/10.3748/wjg.v28.i12.1187

Institute for Health Metrics and Evaluation (IHME) Global Health Data Exchange. (2024, Feb 28). *United States Stomach Cancer Mortality Rates by County, Race, and Ethnicity 2000-2019.* Retrieved Jun 30, 2024, from https://ghdx.healthdata.org/record/ihme-data/united-states-stomach-cancer-mortality-by-county-race-ethnicity-2000-2019.

Katella, K. (2024, February 7). *Stomach cancer is still a risk for many people*. Yale Medicine. Retrieved Jul 21, 2024, from https://www.yalemedicine.org/news/stomach-cancer-gastric-cancer.

Kronick R. (2016). *AHRQ's Role in Improving Quality, Safety, and Health System Performance.* Public health reports (Washington, D.C. : 1974), 131(2), 229–232. Retrieved Jul 22, 2024, from https://doi.org/10.1177/003335491613100205.

Lundberg, S. (2018). *An introduction to explainable AI with Shapley values.* SHAP Documentation. Retrieved Aug 9, 2024, from https://shap.readthedocs.io/en/latest/example_notebooks/overviews/An%20introduction%20to%20explainable%20AI%20with%20Shapley%20values.html.

Lundberg, S. (2018). *SHAP TreeExplainer.* SHAP Documentation. Retrieved Aug 9, 2024, from https://shap-lrjball.readthedocs.io/en/latest/generated/shap.TreeExplainer.html.

Peduzzi, P., Concato, J., Kemper, E., Holford, T.R., & Feinstein, A.R. (1996, Dec). A simulation study of the number of events per variable in logistic regression analysis. Journal of Clinical Epidemiology, Vol 49, Issue 12, pp. 1373-1379. Retrieved Jul 22, 2024, from https://doi.org/10.1016/S0895-4356(96)00236-3.

Rippner, N. (2018, Jul 30). *County FIPS to zip code crosswalk.* DataWorld. Retrieved Jun 30, 2024, from https://data.world/nrippner/fips-to-zip-code-crosswalk.

Sarkar, S., Dauer, M. J., & In, H. (2021). *Socioeconomic disparities in gastric cancer and identification of a single SES variable for predicting risk. Journal of Gastrointestinal Cancer, 53(1), 170–178.* Retrieved Jul 1, 2024, from https://doi.org/10.1007/s12029-020-00564-z.

Santellano, B., Agrawal, R., Duchesne, G., Sharara, M., Balas, E. A., Agrawal, G., Tsai, M.-H., Nayak-Kapoor, A., & Cortes, J. E. (2024a). *EPR24-113: The role of Social Determinants of health in gastrointestinal cancer outcomes in the United States context: A systematic review.* Journal of the National Comprehensive Cancer Network, 22(2.5). Retrieved Jul 1, 2024, from https://doi.org/10.6004/jnccn.2023.7217.

SEER. (2021). *Cancer of the stomach - cancer stat facts*. Retrieved Jul 21, 2024, from https://seer.cancer.gov/statfacts/html/stomach.html.

Scikit Learn User Guide & API. (n.d.) *StandardScaler*. Retrieved Jul 29, 2024, from https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html.

Tucker-Seeley, R., Abu-Khalaf, M., Bona, K., Shastri, S., Johnson, W., Phillips, J., Masood, A., Moushey, A., & Hinyard, L. (2024, Feb 22) *Social Determinants of Health and Cancer Care: An ASCO Policy Statement.* JCO Oncology Practice Vol 20, No 5. Retrieved Jul 1, 2024, from https://doi.org/10.1200/OP.23.00810.

United States Census Bureau. (2024, Jun 18). *American Community Survey Data.* Retrieved Jul 22, 2024, from https://www.census.gov/programs-surveys/acs/data.html.

United States Census Bureau. (2021, Oct 8). *2010 ZIP Code Tabulation Area (ZCTA) Relationship File Record Layouts.* Retrieved Jun 30, 2024, from https://www.census.gov/programs-surveys/geography/technical-documentation/records-layout/2010-zcta-record-layout.html.