EDA Report

1.1 Modeling Question and Data Overview

Our research question is: What factors most affect cancer mortality rates, and what, if any, are the interactions between these factors. Using our dataset, consisting of both yearly values and multi-year averages from cancer.gov, the American Community Survey, and clinicaltrials.gov, the focus of our project is causal inference on factors such as age, gender, race/ethnicity, socioeconomic status, geographic location, and more. To establish causal relationships, we will use linear regression modeling and techniques to control for potential confounding factors and establish the causal effects on the dependent variable of Mean *per capita* (100,000) cancer mortalities. This will allow us to reach causal conclusions and identify potential interventions that can reduce cancer deaths. Cancer is a leading cause of mortality in the United States, and understanding the factors that affect cancer mortality can help the federal government as well as state governments effectively allocate capital and resources in the correct areas. Additionally, inference on variables relating to socioeconomic status can help officials in the healthcare industry, such as insurance companies, better understand the relationship between healthcare affordability and cancer mortality.

Our data set consists of 3047 observations and 34 variables, where a unit of observation is a county in the United States. As mentioned previously, a variable can either be a value collected in a certain year, or an average over a period of years. For example, our dependent variable, “TARGET_deathRate” represents the Mean *per capita* (100,000) cancer mortalities between the years 2010 and 2016. In contrast, the variable “MedianAge” represents the median age of residents in a certain county coming from the 2013 Census Estimates. Since the aim of our project is causal inference, not prediction accuracy, as well as the difficulty of consistently measuring variables such as “MedianAge,” we don’t believe there is any significant temporal aspect to our results. We don’t believe that our observations are perfectly independent, as state-wide trends can certainly affect certain variables compared to a different state, however we will control for this by considering state and even region averages in our model. While there are gaps in certain variables due to not being collected in that county, for most variables this is insignificant and randomly placed (geographically) enough to consider the data set generalizable to the entire United States. Additionally, the use of multiple years in our data set we believe can

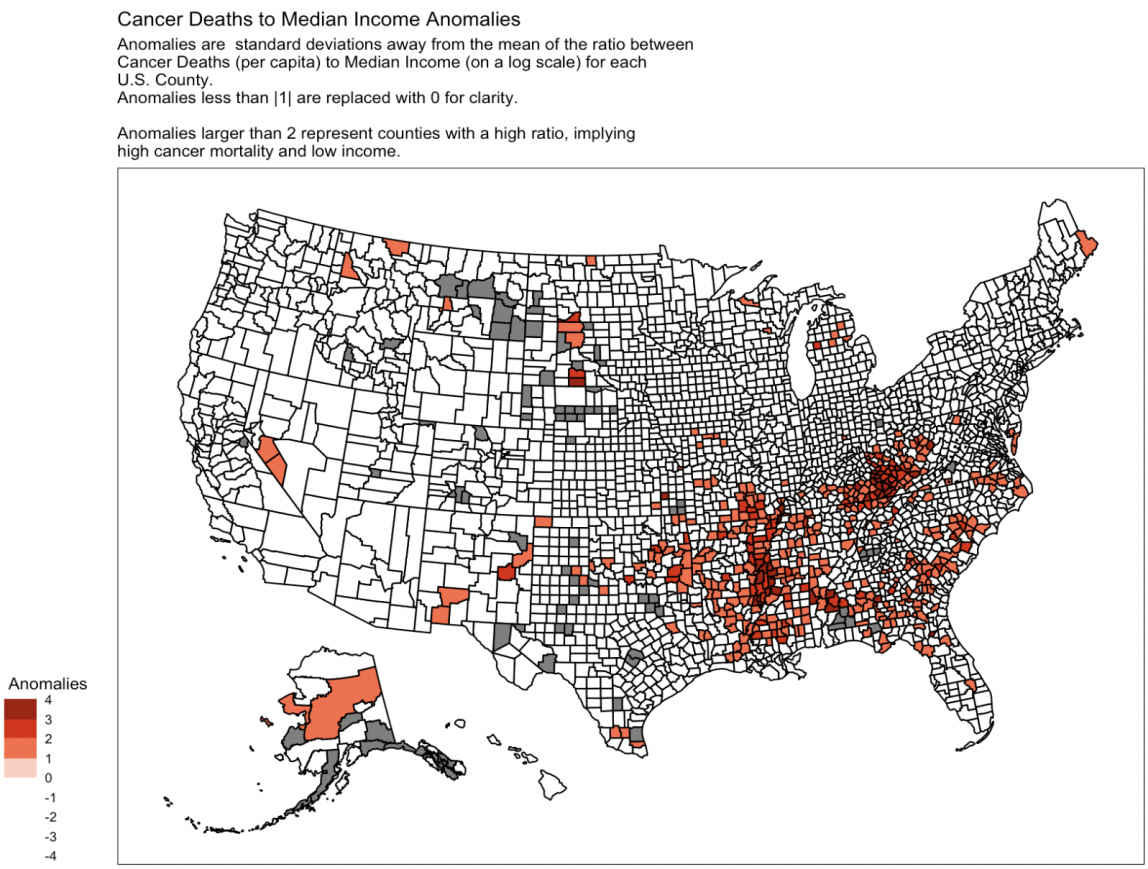
make our results more robust and generalizable for the decade of 2010-2020, as well as possibly beyond that. As mentioned previously, we anticipate understanding possible trends within states or regions to be a challenge for our analysis, as well as understanding the inevitable collinearity between certain variables, such as education level and income. Additional data not recorded in our dataset that we believe would improve our modeling is data on county occupation distributions, as it's known that certain environmental factors, such as radiation, impact the incidence and thus death rate of cancer.

Data citation:

N.Rippner. Cancer Trials, 2017. Retrieved from http://data.world/exercises/linear-regression-exercise-1/workspace/file?filename=cancer_reg.csv.

1.2 EDA

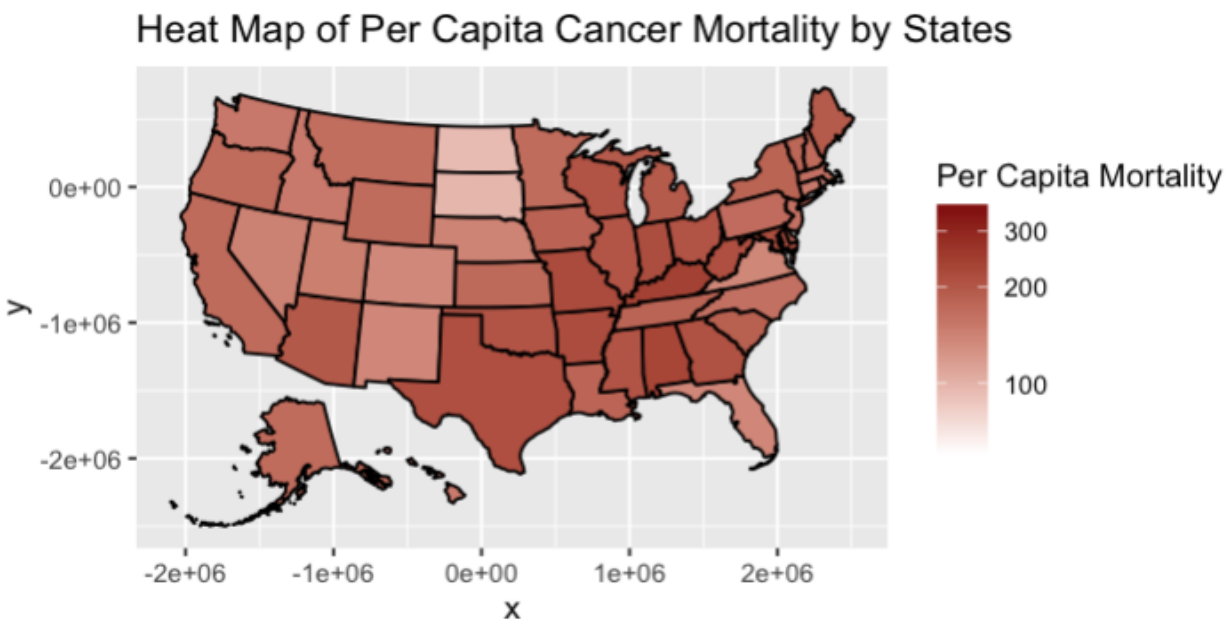
1.2.1



For our first visualization, we were interested in analyzing the relationship between income and cancer mortality, as well as identifying if there were any “problem areas” geographically that may indicate a larger state or region-wide trend.Firstly, we explored the ratio between our dependent variable, average cancer mortality per capita, and the median income for each county. To help symmetrize median income and thus our ratio, we first applied a log transformation, then calculated our ratios. Then, to identify any anomalies, that is, counties with an unusually high or small ratio, we standardized our ratio values, creating a new variable called “Anomalies,” representing the standard deviations away from the mean ratio. Finally, to focus on significant differences to the mean ratio, we changed any anomalies less than |1| to 0. We also wanted to only focus on positive anomalies, which represent a county with a high cancer mortality compared to its median income, and thus when graphing, only used color for positive anomalies larger than 1. Grey represents missing data, which mentioned previously, is not significant to consider when generalizing our results.

As seen in the graph, nearly all of the positive anomalies are in the South-East region of the United States, indicating possibly a region-wide trend. When modeling, this implies that we should consider analyzing if a dummy variable that tells us whether or not a county is in the South-East region significantly affects average cancer mortality. If this dummy variable is found to be significant, it could be used to justify additional federal spending in this certain area of the country. Additionally, if this dummy variable is found to be significant, it also supports the relationship between median income and average cancer mortality, as it’s important to note in this visualization we aren’t modeling simply cancer mortality, but the ratio between cancer mortality and median income.

1.2.2

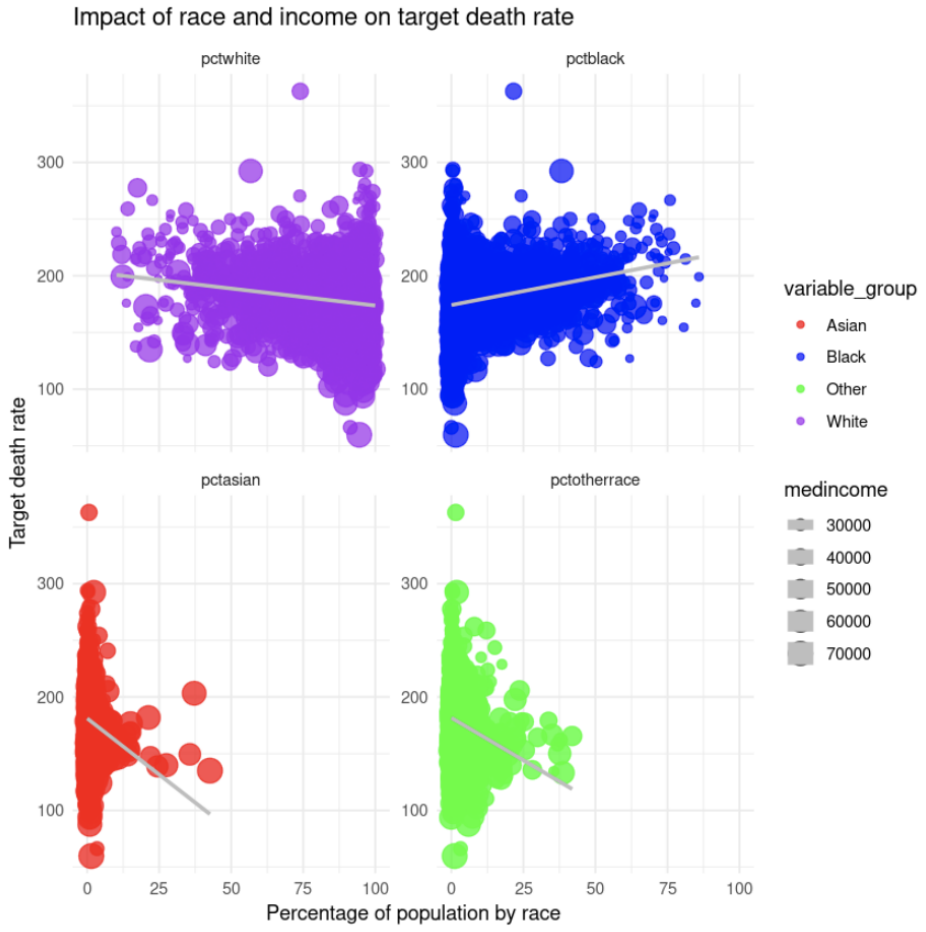


While our first visual shows the potential connections between the median income and the cancer mortality rates, this heatmap directly shows the distribution of per capita average numbers of annual deaths from Cancer in the US by state. The darker the color is, the more annual deaths occurred in that state. To create this heat map, we utilized the 'geometry' column of the dataset and used string manipulation to extract the state name for each tuple. We then used the usmap package in R to extract the map information and merged the map dataframe with our cancer dataframe on state name. We utilized ggplot to plot the final heatmap. In addition, the above data preprocessing will be carried over into the model.

The heatmap above is related to our research question in that it reveals the connection between geographical locations and cancer death rates as well as diagnostic cases (the plot for which is omitted here for clarity). It looks like the geographical location indeed has a direct impact on the per capita cancer death rates, in that the states closer to the west and east coast generally have higher death cases from cancer than states in the middle. In addition, different from our first visual which shows low mortality-to-income ratios in west-coast states (which might indicate low per capita cancer mortality rates at the first glance), this heatmap indicates that the per capita cancer mortality rates in west-coast states are actually pretty high, which suggests an interesting phenomenon that west-coast states may have a higher-than-usual average income base that reduces the mortality-to-income ratios while the actual mortality rates in those

states are definitely worth to be examined. Therefore, it may be worth it to put geographical locations into the final regression model, which is again concerned with finding the most significant factors that affect mortality rates of cancer. For future actions, we may consider splitting the states into two categories (near coast and far from coast) to minimize the number of dummy variables for the state information.

1.2.3



While the heatmap from 1.2.2 informs us that the region is an important factor in the average cancer mortality rate, we begin to think about whether the effect of geopolitical factors on cancer might be reflected through racial demographic effects on cancer mortality.

By considering distribution percentages of different races in each county respectively from the dataset, as well as the median income, a ggplot helps visualize the possible impact of race on annual average cancer mortality rate. Each point in the plot represents a county, each color of the point represents a specific race of that county. With the size of the dots representing the range of median income for that county. The grey line as trends for each racial of percentages of race vs cancer mortality is indicating they are not random and racial demographic should be classified as features. Including income helps us understand if it will be a confounding factor of races' percentage and the cancer mortality. Among each race, higher up on the vertical line, there appears to be a greater distribution of lower income (smaller points), while such distribution characteristics vary across racial density. This implies that our lower income groups appear to face greater cancer mortality, and that this effect of income on mortality varies by race.

The information obtained by this plot suggests that we should further consider the possibility of multicollinearity later in building our model, the effects of racial percentages and income on target mortality may vary by region or subpopulation, and differences will be taken into account when constructing our model and interpreting our results. The relationship between racial percentages, income, and the death rate are complex and non- linear, and we may have to better implement it through linearization and appropriate methods under specific consideration.