Data C102 Final Project, Fall 2023

Arielle Santos, Caedi Seim, Daniel Liu, Therese Mendoza

December 10, 2023

1 Data Overview

1.1 Cancer Case Count and Socioeconomic Factors per US County

This dataset (cancer_reg.csv), which spans several US counties, includes a multitude of socioeconomic and health-related data that have been compiled from several sources, including the American Community Survey, clinicaltrials.gov, and cancer.gov. The dataset covers socioeconomic and health-related information across various US counties and was compiled from diverse sources, including the American Community Survey, clinicaltrials.gov, and cancer.gov. The American Community Survey provides census-level data on socio-demographic variables, whereas clinicaltrials.gov contributes information about specific clinical trials, and cancer.gov compiles cancer-related data from multiple sources.

The data was sourced from publicly available datasets related to health outcomes, demographics, and socioeconomic factors across US counties spanning the years 2010 to 2016. The datasets were accessed and downloaded from Kaggle, a platform that hosts various datasets and provides a collaborative environment for data science projects.

Because our data represents a census, it aims to cover the entire population within a designated geographic area. However, certain factors may lead to systematic exclusions or undercounts in a census:

1. Undercoverage: Certain groups may be underrepresented in a census due to factors such as incomplete address lists, non-response, or challenges reaching certain populations (e.g., homeless individuals, and transient populations).
2. Non-Response Bias: Individuals who choose not to participate in the census or are unable to be reached can result in non-response bias, potentially excluding certain demographic groups.
3. Hard-to-Count Populations: Some populations, such as minority groups, immigrants, or those with limited access to information, may be more challenging to count accurately, leading to systematic undercoverage.

Participants are generally aware of the collection and use of data, as the census is a large-scale, government-conducted survey that aims to enumerate and collect information from the entire population within a specific geographic area. Participation in the census is mandatory, and individuals are typically informed about the purpose of the census, the types of information being collected, and how the data will be utilized. The granularity of the data is county-level, with each row representing a unique county. This impacts the interpretations of our findings by allowing us to draw region-specific insights.

In the analysis of our dataset, several data quality considerations were examined to ensure the reliability and validity of our findings. Selection bias, a common concern in data analysis, is typically less pronounced in census data compared to many other types of data collection methods. Census data aim to enumerate the entire population within a specific geographic area, and exhaustive efforts are made to

include every individual, thereby minimizing the potential for self-selection or exclusion biases. Another critical consideration is the presence of measurement error, particularly in variables derived from self-reported data, such as income and education levels sourced from the American Community Survey. Participants may provide inaccurate or incomplete information, introducing measurement errors that could impact the reliability of socioeconomic variables. Convenience sampling, often associated with biases introduced by non-random selection methods, is not applicable in the context of our study. As we utilize census data, our approach aims to collect information from the entire population rather than a subset. This ensures a comprehensive representation of every individual or unit within the specified geographic area or population of interest. Additionally, the dataset was not modified for differential privacy. Each row represents a county, not an individual, so the data in each column is aggregated.

While the dataset provides a rich set of features relevant to our research questions on the associations between socioeconomic and demographic factors and cancer rates across different counties, there are certain additional features that, if available, could further enhance the depth of our analysis. These include:

1. Lifestyle Variables: Information on lifestyle factors such as smoking rates, physical activity levels, and dietary habits would provide valuable insights into behavioral determinants of cancer rates. Understanding the interplay between socioeconomic factors and lifestyle choices could help formulate more comprehensive public health strategies.
2. Environmental Factors: Data on environmental variables, such as pollution levels, access to green spaces, or exposure to carcinogenic substances, would contribute to a more nuanced understanding of the environmental determinants of cancer rates. This information would be particularly relevant for assessing the impact of both socioeconomic factors and environmental conditions on health outcomes.
3. Healthcare Accessibility: Metrics related to healthcare accessibility, such as the availability of healthcare facilities, the density of healthcare professionals, and the prevalence of health insurance coverage, would help evaluate the role of healthcare infrastructure

There were a few columns with missing data. These include: pctsomecol18_24, pctemployed16_overpct, privatecoveragealone. Missing values suggest that this information is not available for some counties.

- pctsomecol18_24: This column represents the percentage of people aged 18-24 who attended some college.
- pctemployed16_over: This column represents the percentage of people aged 16 and over who are employed.
- pctprivatecoveragealone: This column represents the percentage of people with private health insurance coverage who are not covered by public insurance.

We decided to drop the missing values in our data before the analysis stage.

In preparing our dataset for analysis, a series of cleaning and pre-processing steps were implemented to enhance the quality and reliability of our data. One of the initial steps in our exploratory data analysis involved deriving a percentage of cancer cases per county. This transformation was motivated by the need to standardize cancer rates across counties with varying populations. Before this, our dataset contained a column representing the average annual count of cancer cases. Converting this count to a percentage provides a more comparable metric, facilitating meaningful comparisons between counties. During the creation of visualizations for cancer rates using our newly computed percentage column, a discrepancy was identified. Specifically, numerous observations exhibited a constant value of

1962.667684 for the average annual count of cancer cases (avganncount). Recognizing the need for data integrity, we opted to address this issue by removing the rows associated with this constant value. It is essential to acknowledge that removing specific values may introduce bias if the pattern of missing values is related to other variables or if it introduces systematic bias. In this context, we carefully considered whether the constant value was indicative of missing or incorrect data. The decision to drop these rows was deemed necessary to maintain the reliability of our dataset. These cleaning and preprocessing decisions contribute to the overall quality of our data, ensuring that subsequent models and inferences are built upon a sound foundation.

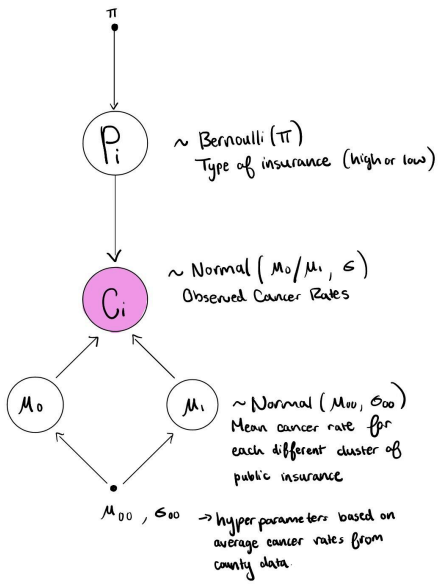
2 Research Questions

2.1.1 Question, algorithm, and modeling choices:

Question:
Can we fit a Bayesian Gaussian mixture model to the distribution of cancer rates by the percentage or type of insurance rate? How does the distribution of cancer rates change with this observed public insurance percentage?

Methods:

We first drew out a graphical model to model the variables of our question.



We have defined our variables to be as follows:

- Pi: Binary indicator Pi = 0 if county i has low insurance and Pi = 1 if county i has high insurance. For the sake of this experiment, we are saying that Pi follows a Bernoulli (0.5) distribution.
- Ci: Observed cancer rate for each county.
- μ0 and μ1: Average cancer rate from low and high insurance groups.
- Ci | Pi, μ0, μ1: Normally distributed likelihood.

Using PYMC, we aim to estimate the following pieces of information:

- We are interested in learning more about the mean cancer rates in each cluster of insurance. In other words, what are μ0 and μ1, and are they different?
- We are also interested in learning more about which cancer rates are coming from which cluster of public insurance.

Groups:
We are using our mixture model to identify if there are two different distributions of cancer coming from public insurance. Our objective is to assess whether there is a significant difference in the distribution of cancer rates between these two groups. The decision to create two clusters was influenced by the understanding that 'Uninsured adults have less access to recommended care, receive poorer quality of care, and experience worse health outcomes than insured adults do.' And, we have understood that a low level of public insurance does not necessarily correlate with a high level of private insurance. Instead, it indicates a general upward trend in the prevalence of uninsured individuals. There are potentially more levels of public insurance we could explore rather than low and high, but since we do not currently know these categorizations, we decided to explore and split the data into two categories.

Priors:
- The proportion parameter Pi in the mixture model represents the weight of each cluster. We believe low and high public insurance is equally likely as the histogram we created in our EDA of the distribution of public coverage percent showed a normal distribution, we let this be a Bernoulli random variable with 0.5 probability that reflects a neutral stance, indicating that both clusters are considered equally plausible as we do not have any information as to what constitutes high or low insurance.
- The variable μ0 should reflect the cancer mean if the cluster is coming from cluster 0, low insurance. Likewise, the variable μ1 should reflect the cancer mean if it is coming from cluster 1, high insurance. We have chosen these to be normally distributed variables with specific means based on external research on insurance rates and cancer diagnoses.
- Our last prior, cancer rate is given cluster and mean variables (Ci | Pi, μ0, μ1) is normally distributed, this reflects our belief that, within each cluster, the cancer rates follow a Gaussian distribution around the corresponding mean. We have made this assumption and are going to test its fit, as the data is continuous and the factors we anticipate influencing the cancer rate appear to be normally distributed.

2.1.2 Assumptions:

Anticipation:
We anticipate that the cancer rate comes from two different Gaussian distributions, based on a high or low insurance rate. Based on our exploratory data analysis, we found using a scatter plot that public insurance and cancer rate held a slight positive association (figure x), showing that a higher percentage of public coverage was associated with a higher proportion of cancer. Research on Rural-Urban Disparities in Cancer has also indicated that populations with low public insurance tend to have worse health outcomes and more cancer incidence per capita.

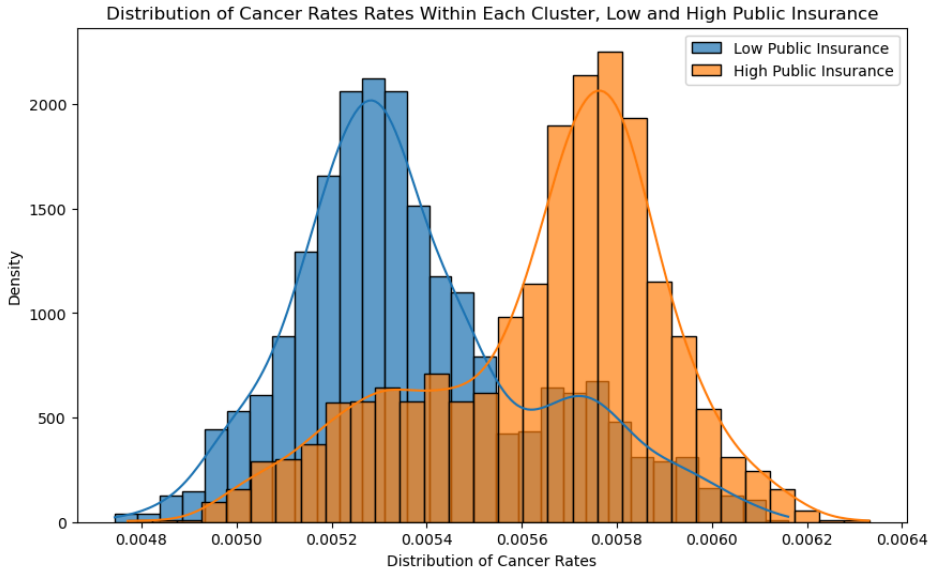
Decision on Means & Standard Deviation:

We also decided on the means and standard deviations for our μ based on the descriptive statistics of percent cancer as external research (Health Insurance Status and Clinical Cancer Screenings Among U.S. Adults) share that overall underinsured, never insured, and publicly insured women and men were less likely to receive screening. Therefore, we assumed that the lower percentile of cancer incidence rate was coming from a lower insured population, and a higher percentile of cancer incidence rate was coming from a higher insured population.

2.1.3 Implementation and statement of results:

Results Summary:
We used PyMC to sample under the posteriors using the priors and likelihood defined above..
Disclaimer: We used a random sampling of county data to create these as our notebooks ran out of memory using the full dataset of counties.

Displaying results on an overlaid histogram, we do see that there do seem to be two distinct Gaussian shapes for cancer, however, the difference between these groups is pretty small, and there seems to be some overlap.



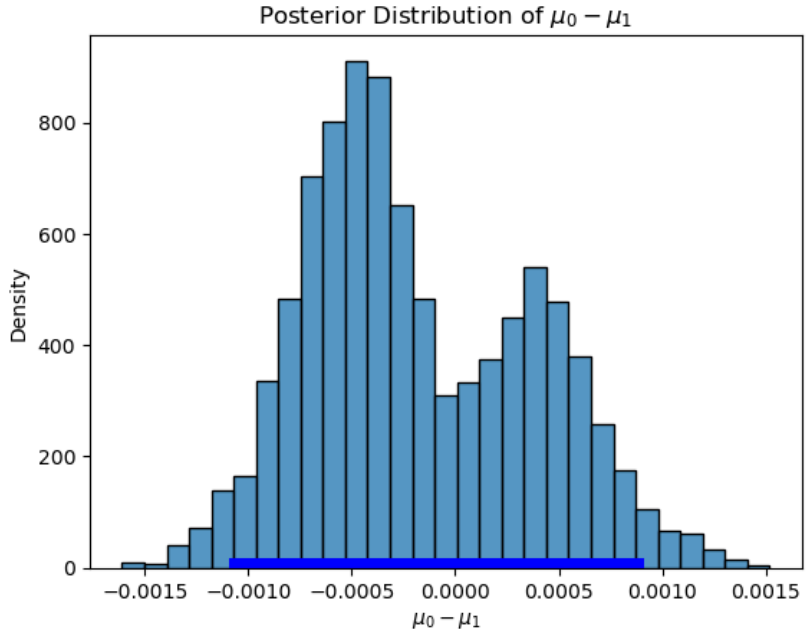
Where the median cancer rate for low public insurance is **0.00540** and high public insurance is **0.005654.**

Under the posterior, the probability that more of the cancer rates came from a cluster of low insurance was approximately **0.4585.**

Under the posterior, the probability that cluster 0, our low public insurance group, has a greater rate of cancer than that of the high insurance group is approximately **0.297.**

Quantifying Uncertainty

Using the credible interval with the posterior estimates, it appears that using the 95% credible interval, the difference in means lies between -0.0010 and 0.0010, which includes 0, meaning that there is a 95% probability that the true parameter would lie within that interval.



Where the dark blue is the credible interval.

2.1.4 Interpretation of results:

After conducting the Bayesian Gaussian Mixture Model on the distribution of cancer rates, we observed two distinct Gaussian shapes, indicating some differences in low and high public insurance clusters.that while there does seem to be two distinct peaks for cancer rates in both clusters of public insurance. However, it is important to note that these differences are very subtle, as evidenced by the overlap in the distributions and similar median values.

The posterior probabilities provide insights into the likelihood that more cancer rates originated from the low insurance cluster, which seems to be pretty even showing that the high and low clustering is balanced. It is also less likely that low public insurance groups would have a higher cancer rate than that of a high public insurance group. However, as evidenced by the credible interval the difference between the two μ0 and μ1 is not significant and any difference between the two could be due to chance as it includes 0 and is generally very small.

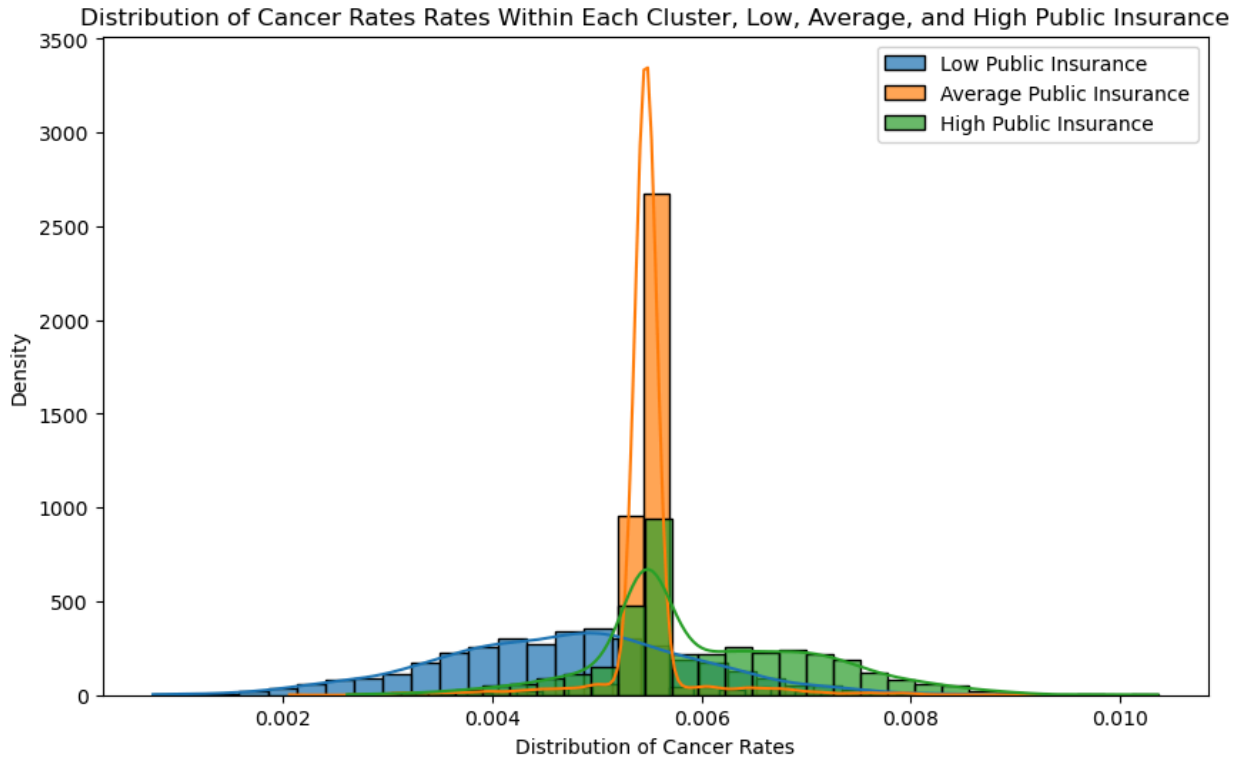
We were somewhat surprised that low public insurance was associated with a lower cancer rate and high cancer was associated with a higher cancer rate, as this does seem to be counterintuitive; however, it is important to interpret these findings with caution. The differences are very subtle, and since the cancer percentage we used was based on incidence rate, it may be that lower public insurance meant less diagnosis.

Limitations include the fact that cancer rates may be different in differing clusters due to variables that are not insurance type. For example, our dataset also includes median income level which also could have possibly had an effect on clustering cancer rates but was not explored. We also used a set standard deviation as we could not find any literature on the standard deviation for cancer means aside from our summary statistics. If possible, additional data relating to the variation of cancer rates for low-insured versus high-insured people would have likely increased the validity of our model, and we likely would have chosen stronger priors to reflect that. We also would attempt to run this on our full dataset given we have more memory.

Other Investigation

Three Clusters:

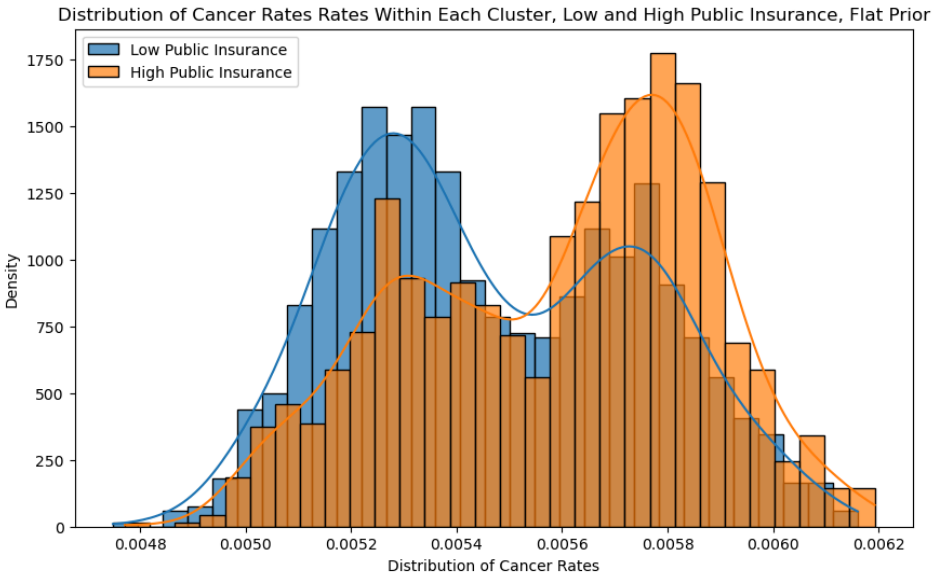
We found that using any more clusters greater than 2 resulted in a lot of overlap between the cancer rates, indicating that cancer rates per county are likely only different in at most 2 clusters.



On the other hand, although deciding on 2 clusters reduces overlap, we also subsequently categorize our data into two different binary indicators, “low” and “high”, which may risk oversimplification of the complexity of insurance coverage.

Different Prior:

We also attempted this on uninformative priors such as the flat prior for μ and sigma with initial values. We found that our normal priors did not fully converge but the flat prior did. This gave us results similar to that of the original model, although it appeared that there was more overlap between the two types of insurance.



2.2.1 Question, algorithm, and modeling choices:

Question:
What socio-economic and demographic factors are associated with variations in cancer rates across different counties?

Methods:
To answer this question, we implemented multiple hypothesis testing. We are testing for associations between cancer rate and the following variables: median income, poverty percentage, number of clinical trials per capita, educational attainment, employment status, and population density. Testing multiple hypotheses is essential to capture the complexity of these associations.

Our hypotheses all follow the following format:
Null Hypothesis (H0): There is no association between [variable] and cancer rate.
Alternative Hypothesis (H1): There is a significant association between [variable] and cancer rate.

The test we are using is Neyman-Pearson. We cannot compute the power of the test because it is a composite alternative hypothesis. Our hypotheses are as follows:

- Hypothesis 1 (Association between Cancer Rate and Median Income):
    - Test Type: Pearson's correlation coefficient.
    - Justification: Pearson's correlation is suitable for assessing the strength and direction of a linear relationship between two continuous variables. In this case, it helps determine if there is a significant negative association between cancer rate and median household income.
- Hypothesis 2 (Association with Poverty Percentage):
    - Test Type: Pearson's correlation coefficient.

    - Justification: Similar to Hypothesis 1, Pearson's correlation is appropriate for examining the relationship between two continuous variables. It will help assess whether higher poverty percentages are associated with a higher cancer rate.
- Hypothesis 3 (Association between Clinical Trials and Cancer Rate):
    - Test Type: Pearson's correlation coefficient
    - Justification: Pearson's correlation is suitable for assessing the relationship between two continuous variables. This test will help determine if clinical trial availability is associated with a significant difference in cancer rates.
- Hypothesis 4 (Impact of Education on Cancer Rate):
    - Test Type: Pearson's correlation coefficient.
    - Justification: Utilizing Pearson's correlation to assess the relationship between cancer rate and different educational attainment levels among the 18-24 age group allows for the examination of potential linear associations.
- Hypothesis 5 (Employment Status and Cancer Rate):
    - Test Type: Pearson's correlation coefficient.
    - Justification: Pearson's correlation is suitable for assessing the relationship between two continuous variables. This test will help determine if there is a significant association between employment rates and cancer rates.
- Hypothesis 6 (Impact of Population Density on Cancer Rate):
    - Test Type: Pearson's correlation coefficient.
    - Justification: Pearson's correlation is used to evaluate the linear relationship between two continuous variables. In this case, it helps assess whether there is a significant association between population density and cancer rate.

2.2.2 Assumptions:

When planning this study, we hypothesized that socioeconomic status has a significant impact on cancer rates in various counties. As a result, the hope of finding correlations between particular socioeconomic indicators and cancer rates informed many of our methodological decisions, including the variables we selected and the statistical tests we ran.

We made the assumption of non-independence of null p-values because all the variables we are testing come from the same dataset. When all the variables being tested are part of the same dataset, especially when each row represents a distinct county, it is plausible that there are interrelationships or dependencies among these variables.

In a county-level dataset, socioeconomic indicators such as median income, education levels, employment status, and population density can be interconnected. For example, counties with higher median incomes might tend to have higher educational attainment levels, and counties with higher population density might have different socioeconomic characteristics compared to less densely populated counties. To account for this assumption, we implemented the Benjamini-Yekutieli correction method as opposed to Benjamini-Hochberg, which doesn't account for dependence between the columns.

2.2.3 Implementation and statement of results:

Through this analysis, we aimed to look into the relationships that exist between different socioeconomic indicators and cancer rates across different counties. To assess the direction and strength of linear relationships between cancer rates and particular variables, we used Pearson's correlation coefficient. To account for multiple hypothesis testing, we used the Bonferroni and Benjamini-Yekutieli correction techniques.

First, we developed six hypotheses, each of which looked at the relationship between cancer rates ("pctcancer") and a particular socioeconomic variable, like poverty rate, the availability of clinical trials, median income, employment status, education levels, and population density.

In the table below, we implemented the Bonferroni correction. For the Bonferroni correction, we conducted Pearson's correlation tests for each hypothesis. The resulting p-values were then adjusted using the Bonferroni correction method to control the FWER at a significance level of 0.05. The adjusted p-values and the decision to reject the null hypothesis were recorded. Specifically targeting the FWER, ensures that the probability of making at least one false positive (Type I error) across all tests examining different factors remains at the desired level. This is to control the overall risk of falsely identifying associations.

Hypothesis object	Variable object	Correlation Coeffi...	P-Value float64	Adjusted P-Value fl...	Reject Null Hypot...
Cancer Rate and ...	medincome	-0.2556043536	1.511537974e-9	9.069227846e-9	True
Association with ...	povertypercent	-0.01771017302	0.6805079374	1	False
Association betw...	studypercap	0.03595983788	0.4029925807	1	False
Impact of Educati...	pctbachdeg18_24	-0.06410962157	0.1356996771	0.8141980625	False
Employment Stat...	pctemployed16_o...	-0.1857544111	0.00001321511229	0.000079290673..	True
Impact of Populat...	popest2015	-0.1829497779	0.000017900037..	0.0001074002223	True

Similarly, we applied the Benjamini-Yekutieli correction to control the false discovery rate (FDR). We conducted Pearson's correlation tests, adjusted the p-values using the Benjamini-Yekutieli method, and recorded the adjusted p-values along with the rejection decisions. FDR represents the expected proportion of false discoveries among the factors identified as associated with cancer rates. By using this correction, we maximize the detection of relevant factors while also tolerating a controlled proportion of false positives among the identified associations.

Hypothesis object	Variable object	Correlation Coeffi...	P-Value float64	Adjusted P-Value fl...	Reject Null Hypot...
Cancer Rate and ...	medincome	-0.2556043536	1.511537974e-9	2.221960822e-8	True
Association with ...	povertypercent	-0.01771017302	0.6805079374	1	False
Association betw...	studypercap	0.03595983788	0.4029925807	1	False
Impact of Educati...	pctbachdeg18_24	-0.06410962157	0.1356996771	0.4986963133	False
Employment Stat...	pctemployed16_o...	-0.1857544111	0.00001321511229	0.00008771018151	True
Impact of Populat...	popest2015	-0.1829497779	0.000017900037..	0.00008771018151	True

Data on the hypothesis, variable, correlation coefficient, raw p-value, adjusted p-value, and the choice to reject the null hypothesis were arranged into two DataFrames, one for each correction method. Considering the effects of multiple testing, these DataFrames, “df_bonferroni” and “df_benjamini_yekutieli”, offer a thorough summary of the relationships between cancer rates and socioeconomic variables.

There were significant associations between Hypothesis 1 (Association between Cancer Rate and Median Income), Hypothesis 5 (Employment Status and Cancer Rate), and Hypothesis 6 (Impact of Population Density on Cancer Rate). There were no significant associations between Hypothesis 2 (Association with Poverty Percentage), Hypothesis 3 (Association between Clinical Trials and Cancer Rate), and Hypothesis 4 (Impact of Education on Cancer Rate).

2.2.4 Interpretation of results:

Our first hypothesis (cancer rate and median income) revealed a negative association. This indicates that counties with higher median incomes tend to exhibit a lower cancer rate. This aligns with expectations and suggests this socioeconomic factor influences cancer rates. A similar negative association was present with employment rate, thus higher employment rates are associated with lower cancer rates. This result could be related to improved access to healthcare or other socioeconomic factors linked to employment. Our last significant result was a negative association between population density and cancer rate. Counties with higher population density have a different cancer rate compared to counties with lower population density. This might be indicative of urban-rural differences in healthcare accessibility, lifestyle, or environmental factors.

Since the null hypothesis was not rejected for our hypotheses about the relationship between cancer rates and poverty percentage, clinical trial availability, and education levels, we were not able to find a clear correlation between these variables. It's essential to consider that lack of significance doesn't necessarily mean an absence of relationships as complex interactions may contribute to cancer rates.

Limitations could include temporal limitations, as this dataset spans from 2010 to 2016. Changes in socio-economic factors, healthcare systems, or population characteristics beyond this period may not be captured. Data quality could also be considered a limitation. We had to drop some rows with inconsistent or missing data to perform these tests, and this could’ve introduced some bias. Additionally, assuming uniformity within counties could pose an issue. There could be variations within a county in terms of healthcare facilities, socioeconomic circumstances, and other elements. Lastly, external factors like policy changes, economic recessions, or major health events can influence cancer rates and socioeconomic conditions. The dataset may not fully account for these external influences.

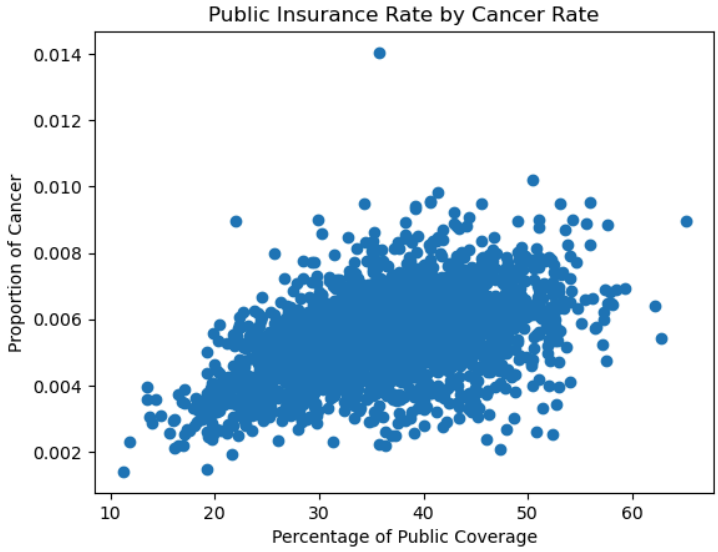
3 Exploratory Data Analysis (EDA)

We initially had to create new columns for the data we aimed to estimate using our two research questions. First, we created a percent cancer column using the data’s average annual count of cancer and

its county's population estimate. Additionally, we created an uninsured column by taking one minus the public and private insurance percentages for each column.

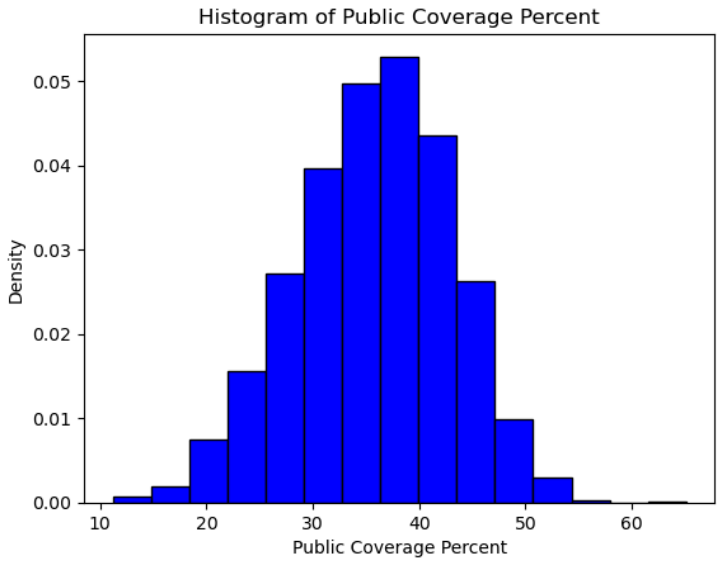
After creating our percent cancer column, and noticing some very odd high values (greater than 90%), we noted that there were some inconsistencies regarding some of the county's average annual count, where multiple rows had the same inconsistent value of 1962. After dropping these values, we did not see any glaring inconsistencies and effectively dropped our dataset to around 2841 values from 3047.

To confirm that there was an association between public insurance rates we used a scatterplot to explore this trend, which resulted in a slight positive association between public coverage and the proportion of cancer. We assumed that because these two variables are correlated, we anticipate that if we group the public insurance rate into two clusters each cluster will have its own cancer rate.

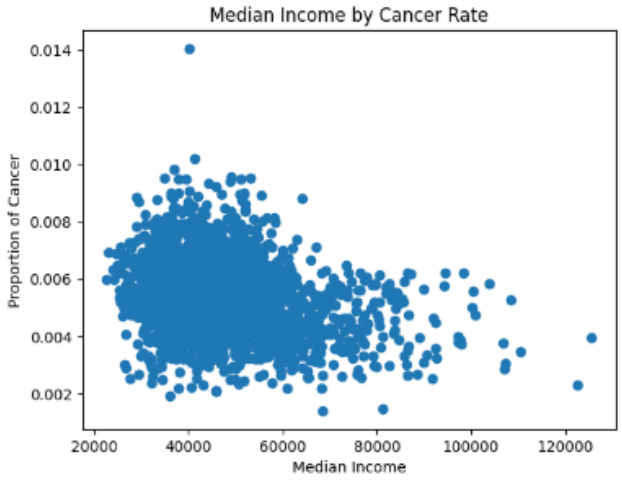


Additionally, we attempted to find associations between total insurance coverage and private insurance but these had less distinct patterns.

We checked out the histogram of public insurance coverage which appeared pretty normal leaning left skewed. We assumed that each cluster of public insurance was equally likely because it was quite normal.

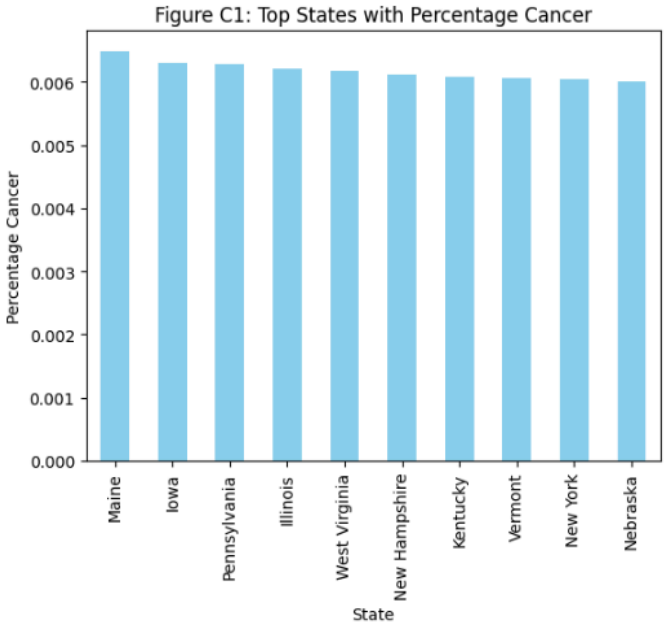


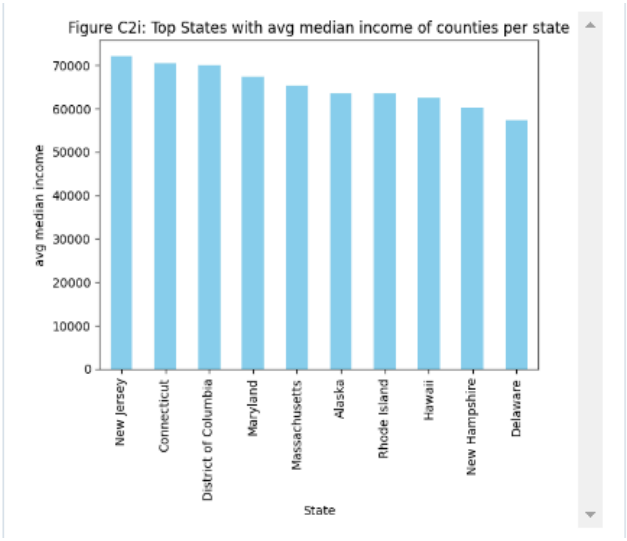
To gain important context for our next research question we created a scatter plot of cancer rates by median income to help visualize potential patterns or trends in the relationship between these variables across different counties. By plotting cancer rates on the y-axis and median income on the x-axis, we could discern whether there is a discernible correlation or any apparent clusters that suggest socioeconomic factors play a role in cancer rate variations. The scatter plot allows us to observe potential outliers, identify any nonlinear associations, and form initial hypotheses about how socioeconomic disparities might contribute to differences in cancer rates among counties.

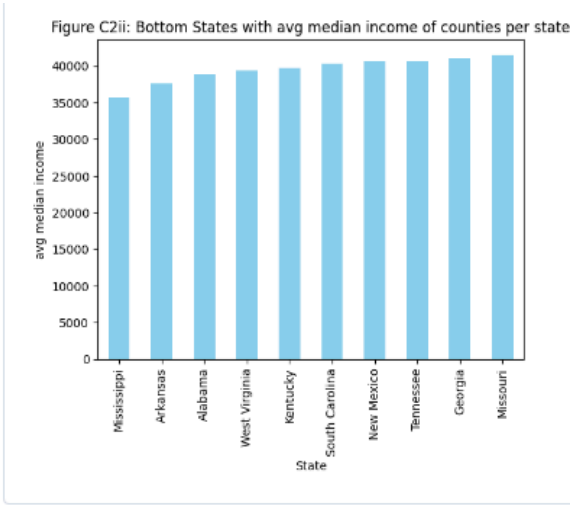


We suspected that a lower median income would correlate with a higher cancer rate. This scatterplot confirms this suspicion because there seems to be a downward trend in the cancer rate as the median income increases. This visualization helps us to answer our research question of "What socioeconomic and demographic factors are associated with variations in cancer rates across different counties?" This plot helps to suggest an answer being the income being a socioeconomic factor.

Expanding on our initial suspicion that lower median income might be correlated with a higher cancer rate, we plotted the top ten states in terms of cancer rate as well as the ten states with the highest and lowest median income to look for similarities.







The observation that two states with the lowest median income also appear among the ten states with the highest cancer rates, while no matching states were found among the highest median income and highest cancer rate states, further reinforces our initial hypothesis. This pattern aligns with the notion that lower median income is correlated with a higher cancer rate

Conclusion

After fitting a Bayesian Gaussian model into our data, we observed a difference between our 2 low and high public insurance clusters, although not a significant amount. We also observed that counties

with low public insurance clusters showed lower cancer rates, while those with high public insurance clusters exhibited higher cancer rates. To add on, of the six socioeconomic/demographic factors we observed, we distinguish median income, employment status, and the impact of population density to be some of the factors associated with variations in cancer rates across different countries. High median income correlates with lower cancer rates, increased employment rates are associated with lower cancer rates and varying population densities influence cancer rates. These findings underscore the significant socioeconomic impact on cancer rates in the U.S.

The dataset appears to offer a reasonably broad view of health outcomes and socioeconomic factors across different US counties over a six-year period (2010-2016). It's not the most current data, which could pose a problem. The results may be broadly applicable because a broad geographic scope and a variety of factors were included.

Based on our results, we believe it is crucial to advocate for improving overall public health and addressing socioeconomic disparities so that the bridges between those with low income vs. high income, opportunities given to those with employment vs. unemployment, etc. may be built and so rates of cancer may be lower and equal among all socioeconomic/demographic groups. To start, healthcare should be made accessible, if not free, for all in the U.S. In addition, it is important to address socioeconomic disparities so we may lessen the socioeconomic disparities implied from our results drawn in *section 2.4.4*. More should be done to improve educational opportunities in counties with lower education levels. It is also crucial that we work towards reducing the housing crisis/housing gap in America.

Aside from the limitations mentioned in our overview regarding the nuance of interpreting census data as well as the values we had to drop from our data, some limitations in our data that we could not account for in our analysis is mostly in regards to our first research question. We still need to do further research regarding what number is determined to mark low insurance and high insurance – or even further – what bins of insurance mark certain levels of insurance. With this knowledge, there could be a possibility of our model choice changing for our observation of cancer rates by percentage and type of insurance rate. As mentioned, the validity of our model could be higher with this knowledge. Furthermore, we had a lack of access to private data. We tried to observe private data, especially as it pertained to private health insurance, but we unfortunately did not find much and were not provided much from our dataset.

There are future studies that could build on our work. One example is studies exploring a more extensive set of variables associated with rates of cancer – ones that are more specific than the ones given in our data set (e.g. smoking rates, diet, pollution levels, access to clean water, etc.). Another is studies exploring changes in cancer rates over time, given its associated factors. Our research would be able to provide information from 2010 to 2016. To add on, studies exploring the patterns of healthcare over time, especially if they are also observing cancer rates, may benefit from our research. Lastly, studies regarding intersectionality and how they may take a role in health disparities may build on our findings having to do with socioeconomic/demographic factors.

We were able to learn that although people with cancer can come from many different backgrounds, we were able to find in our research that specific socioeconomic factors have an association with cancer rates, and those with low vs. high insurance rates are associated with slightly different rates of cancer.