Regression Group Project:

COMPREHENSIVE ANALYSIS OF CANCER INCIDENCE AND CONTRIBUTING FACTORS ACROSS THE UNITED STATES

Group Number 4

1. *Parth Chaudhari*
2. *Sai Venkata Abhiram Karyampudi*
3. *Shriya Keshavareddygari*
4. *Naga Satya Nitish Poosarla*

Introduction.

This regression analysis project investigates trends and relationships within a dataset related to cancer mortality rates. The objective is to uncover insights and potential predictive factors associated with cancer-related deaths through a series of statistical inquiries. The analysis is structured around five distinct questions, each focusing on different dimensions of the data to provide a comprehensive understanding of the factors involved.

History.

Over the past century, data analysis has played a pivotal role in advancing cancer research and treatment in the U.S. In the early 1900s, limited diagnostic tools meant cancer was often underreported, and diseases like tuberculosis were the primary health concerns. By the 1920s-1940s, advancements like X-rays enabled the early detection of some cancers, and the establishment of the National Cancer Institute (NCI) in 1937 laid the foundation for systematic data collection and research. During this period, cancer rates began rising, driven in part by lifestyle changes such as increased smoking, which was eventually linked to lung cancer through large-scale studies.

The 1950s-1970s were characterized by significant data-driven discoveries, including the landmark 1964 U.S. Surgeon General's report confirming the link between smoking and cancer. The National Cancer Act of 1971, signed by President Nixon, increased federal funding for research, enabling more comprehensive data collection on cancer incidence, mortality, and survival rates.

During the 1980s and 1990s, data from cancer screening programs revealed the benefits of early detection for cancers like breast and cervical, but also highlighted disparities in cancer outcomes, particularly in underserved populations. These findings drove public health campaigns and refined screening protocols.

Since the 2000s, breakthroughs in genetic research, bolstered by vast amounts of genomic and clinical data, have led to the development of personalized cancer treatments. Data analysis has helped identify targeted therapies, improving survival rates for many types of cancer. Despite these advances, data continues to reveal persistent disparities in cancer rates and outcomes, particularly in regions with socioeconomic challenges, reinforcing the need for targeted interventions. Today, data remains central to shaping future cancer research, focusing on early detection, tailored treatments, and reducing health inequities.

Overall, the role of data analysis has been pivotal in shaping our understanding of cancer, from identifying risk factors to developing effective treatments. As data collection techniques and analytical methods continue to improve, they will remain critical in guiding future efforts to prevent, detect, and treat cancer more effectively.

Today, data analysis plays a critical role in understanding complex health issues like cancer, providing valuable insights into patterns, risk factors, and trends. Through advanced statistical techniques and machine learning models, researchers can identify key predictors of cancer incidence and mortality, guiding public health efforts toward more effective prevention, early detection, and treatment strategies.

Tools Used.



Minitab: All core data processing, analysis steps and data visualisation, including summary statistics, model building, and hypothesis testing, were conducted in Minitab. Minitab's capabilities in statistical modeling made it the primary tool for regression analysis, enabling detailed exploration of correlations and trends within the data.

Python: Python was employed solely for final data cleaning. Using libraries like Pandas, we ensured data quality and accuracy by performing last-minute adjustments to prepare it for analysis.

Excel: Excel's role was limited to providing the data in a compatible format, enabling easy transfer and manipulation across different tools.

Data Description

The dataset contains information related to cancer incidence and mortality rates across various U.S. counties. Each row represents a specific county or region, providing both demographic and cancer-related data. Here is a detailed description of the columns:

zipCode: The ZIP code associated with the county.

countyCode*:* A code identifying the county within each state.

studyCount: The count of study participants or cases recorded in this dataset for the specific ZIP code.

State: The U.S. state in which the county is located.

PovertyEst: Estimated number of individuals living below the poverty line in the county.

povertyPercent: The percentage of the county population living in poverty.

medIncome: The median income of the county's population.

Name: The official name of the county.

popEst2015: The estimated population of the county as of 2015.

County: The full name of the county and state, including additional identifying codes.

incidenceRate: The cancer incidence rate per 100,000 people in the county.

avgAnnCount: The average annual count of new cancer cases in the county.

recentTrend: The recent trend in cancer cases in the county (e.g., "stable," "rising," or "falling").

fiveYearTrend: The percentage change in cancer incidence over the last five years.

countyName: The name of the county along with the state for reference.

deathRate: The cancer mortality rate per 100,000 people in the county.

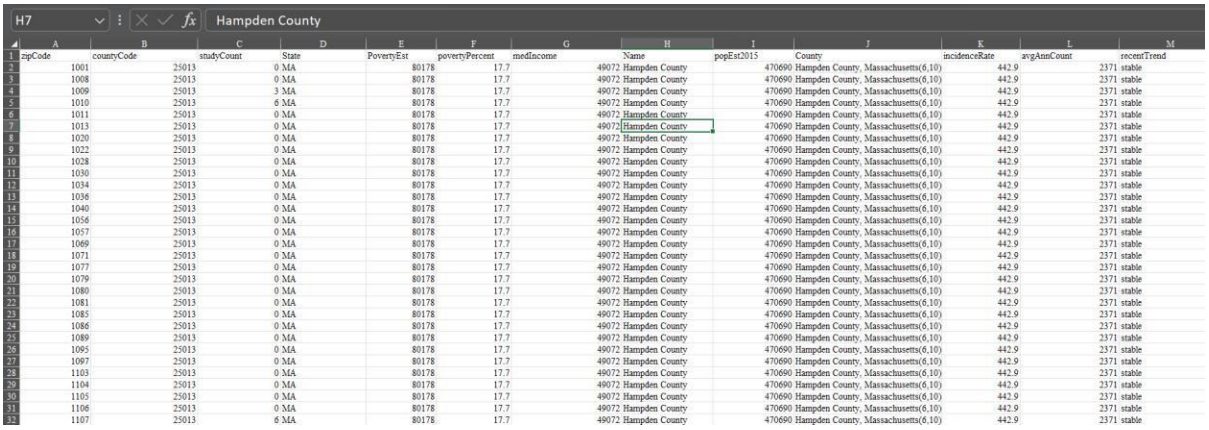
avgDeathsPerYear: The average annual count of deaths due to cancer in the county.

recTrend: The recent trend in cancer mortality rates in the county (e.g., "stable," "falling").

Purpose of the Data

This dataset provides a foundation for analyzing the relationship between socioeconomic factors (such as poverty and income) and cancer incidence and mortality rates across different regions. By examining variables like poverty levels, median income, and population trends, the dataset enables us to investigate potential socioeconomic and demographic risk factors for cancer. This information supports the development of targeted public health interventions and resource allocation to mitigate cancer risks in high-incidence regions.

Data snippet:



H7 | Hampden County

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	zipCode	countyCode	studyCount	State	PovertyEst	povertyPercent	medIncome	Name	popEst2015	County	incidenceRate	avgAnnCount	recentTrend
2	1001	25013	0	MA	80178	17.7	49072	Hampden County	470690	Hampden County, Massachusetts(6,10)	442.9	2371	stable
3	1008	25013	0	MA	80178	17.7	49072	Hampden County	470690	Hampden County, Massachusetts(6,10)	442.9	2371	stable
4	1009	25013	3	MA	80178	17.7	49072	Hampden County	470690	Hampden County, Massachusetts(6,10)	442.9	2371	stable
5	1010	25013	6	MA	80178	17.7	49072	Hampden County	470690	Hampden County, Massachusetts(6,10)	442.9	2371	stable
6	1011	25013	0	MA	80178	17.7	49072	Hampden County	470690	Hampden County, Massachusetts(6,10)	442.9	2371	stable
7	1013	25013	0	MA	80178	17.7	49072	Hampden County	470690	Hampden County, Massachusetts(6,10)	442.9	2371	stable
8	1020	25013	0	MA	80178	17.7	49072	Hampden County	470690	Hampden County, Massachusetts(6,10)	442.9	2371	stable
9	1022	25013	0	MA	80178	17.7	49072	Hampden County	470690	Hampden County, Massachusetts(6,10)	442.9	2371	stable
10	1028	25013	0	MA	80178	17.7	49072	Hampden County	470690	Hampden County, Massachusetts(6,10)	442.9	2371	stable
11	1030	25013	0	MA	80178	17.7	49072	Hampden County	470690	Hampden County, Massachusetts(6,10)	442.9	2371	stable
12	1034	25013	0	MA	80178	17.7	49072	Hampden County	470690	Hampden County, Massachusetts(6,10)	442.9	2371	stable
13	1036	25013	0	MA	80178	17.7	49072	Hampden County	470690	Hampden County, Massachusetts(6,10)	442.9	2371	stable
14	1040	25013	0	MA	80178	17.7	49072	Hampden County	470690	Hampden County, Massachusetts(6,10)	442.9	2371	stable
15	1056	25013	0	MA	80178	17.7	49072	Hampden County	470690	Hampden County, Massachusetts(6,10)	442.9	2371	stable
16	1057	25013	0	MA	80178	17.7	49072	Hampden County	470690	Hampden County, Massachusetts(6,10)	442.9	2371	stable
17	1069	25013	0	MA	80178	17.7	49072	Hampden County	470690	Hampden County, Massachusetts(6,10)	442.9	2371	stable
18	1071	25013	0	MA	80178	17.7	49072	Hampden County	470690	Hampden County, Massachusetts(6,10)	442.9	2371	stable
19	1077	25013	0	MA	80178	17.7	49072	Hampden County	470690	Hampden County, Massachusetts(6,10)	442.9	2371	stable
20	1079	25013	0	MA	80178	17.7	49072	Hampden County	470690	Hampden County, Massachusetts(6,10)	442.9	2371	stable
21	1080	25013	0	MA	80178	17.7	49072	Hampden County	470690	Hampden County, Massachusetts(6,10)	442.9	2371	stable
22	1081	25013	0	MA	80178	17.7	49072	Hampden County	470690	Hampden County, Massachusetts(6,10)	442.9	2371	stable
23	1085	25013	0	MA	80178	17.7	49072	Hampden County	470690	Hampden County, Massachusetts(6,10)	442.9	2371	stable
24	1086	25013	0	MA	80178	17.7	49072	Hampden County	470690	Hampden County, Massachusetts(6,10)	442.9	2371	stable
25	1089	25013	0	MA	80178	17.7	49072	Hampden County	470690	Hampden County, Massachusetts(6,10)	442.9	2371	stable
26	1095	25013	0	MA	80178	17.7	49072	Hampden County	470690	Hampden County, Massachusetts(6,10)	442.9	2371	stable
27	1097	25013	0	MA	80178	17.7	49072	Hampden County	470690	Hampden County, Massachusetts(6,10)	442.9	2371	stable
28	1103	25013	0	MA	80178	17.7	49072	Hampden County	470690	Hampden County, Massachusetts(6,10)	442.9	2371	stable
29	1104	25013	0	MA	80178	17.7	49072	Hampden County	470690	Hampden County, Massachusetts(6,10)	442.9	2371	stable
30	1105	25013	0	MA	80178	17.7	49072	Hampden County	470690	Hampden County, Massachusetts(6,10)	442.9	2371	stable
31	1106	25013	0	MA	80178	17.7	49072	Hampden County	470690	Hampden County, Massachusetts(6,10)	442.9	2371	stable
32	1107	25013	6	MA	80178	17.7	49072	Hampden County	470690	Hampden County, Massachusetts(6,10)	442.9	2371	stable

Area of Research

The primary focus of this research is to explore and identify the socioeconomic, demographic, and regional factors influencing cancer incidence and mortality rates across the United States. By analyzing comprehensive data from various U.S. counties, the study aims to:

1. Identify High-Risk Regions: Determine geographic areas with elevated cancer rates, enabling targeted intervention and resource allocation for cancer prevention and control.

2. Examine Socioeconomic and Demographic Factors: Assess how factors like poverty, income, population density, and healthcare access correlate with cancer incidence. This helps understand the impact of socioeconomic disparities on cancer outcomes.

3. Develop Predictive Models: Create regression models that highlight the most influential factors for both cancer incidence and mortality rates. These models aim to explain the variability in cancer rates and serve as tools for prioritizing intervention efforts.

4. Inform Public Health Strategies: Provide actionable insights to the American Cancer Society for identifying regions and demographic groups that would benefit most from cancer interventions, screening programs, and health education.

Through this research, the goal is to enable data-driven decision-making that reduces cancer disparities and improves health outcomes nationwide.

Methodology

The analysis followed a structured approach to identify factors influencing cancer incidence and mortality rates across U.S. regions:

1. Data Cleaning: Initial data cleaning was performed using Python to ensure accuracy and consistency in the dataset.

2. Descriptive Analysis: Using Minitab, descriptive statistics were calculated to highlight regions with high cancer incidence rates.

3. Hypothesis Testing: Two-sample t-tests were conducted in Minitab to test if the recent trend (stable vs. falling) and study count (high vs. low) significantly impact cancer incidence, using a 5% significance level.

4. ANOVA Analysis: An ANOVA test was applied to examine the effect of different income levels on cancer incidence by dividing income into four groups.

5. Correlation Analysis: Correlations between continuous variables (e.g., poverty, income) and cancer incidence were analyzed to identify key socioeconomic factors.

6. Regression Modeling: Regression models for cancer incidence and mortality were developed to identify significant predictors and explain variability in the data.

Mapping Cancer Hotspots Across the U.S.: A Regional Analysis

We conducted this analysis to identify which regions in the United States are most prone to cancer. This understanding can help in targeting healthcare resources, prevention programs, and support systems more effectively, especially in areas with higher cancer incidence.

To determine the regions most affected by cancer, we analyzed state-wise data on cancer incidence rates (average rate of new cancer cases per population) to see which states have higher or lower rates. By examining the average incidence rate (mean) for each state, we ranked states from those with the highest to the lowest cancer rates. This ranking provided a clear picture of regional differences in cancer prevalence.

In the Southern states, Kentucky (mean: 517.3), Alabama (459.3), and Louisiana (486.8) stand out with notably high cancer rates, suggesting a greater need for cancer-related health resources in these areas. Similarly, in the Northeastern region, New York (497.5) and Delaware (498.2) also have high incidence rates, indicating cancer is more common in these states. The Midwest also has some states with elevated rates, such as Iowa (468.8) and Indiana (453.1), putting this region among those with higher cancer cases.

By analyzing the cancer incidence rate across states, we identified key regions where cancer rates are notably higher. This allows for a more informed approach to resource allocation and program planning in areas most affected by cancer

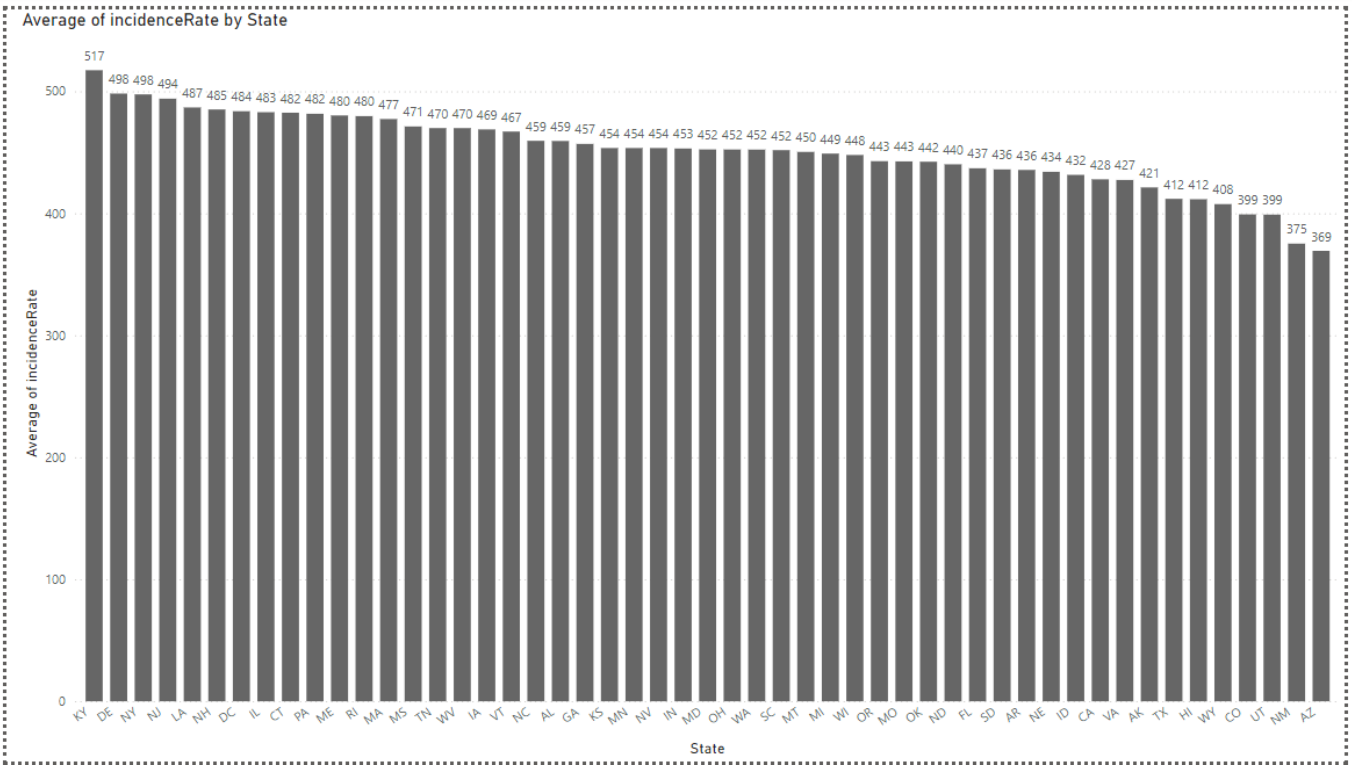


Figure 1.1

Comparing Cancer Trends: Stable vs. Falling Rates and Their Impact

To understand if changes in cancer rates over time affect how common cancer is in different areas, we divided our data into two groups: one with areas where cancer rates are decreasing ("falling") and another with areas where rates are staying the same ("stable"). We did this to see if there is a real difference in cancer cases between areas with falling rates and those with stable rates. Our goal was to find out if places with steady cancer rates might still have high numbers of cancer cases, which could indicate a need for more healthcare support or prevention efforts.

We used a method to compare the average number of new cancer cases in each group, checking whether the difference we observed was likely real or just due to chance. On average, areas with falling cancer rates had around 448.6 cases per population unit, while areas with stable rates had about 455.5 cases. When we analyzed this difference, we found a very low probability that it happened by chance, meaning the difference is statistically significant. We can be 95% confident that the true difference between these groups falls between about 5.7 to 8.1 cases.

This result suggests that even if cancer rates are stable, these areas might still have high levels of cancer cases compared to places where rates are falling. In other words, areas with stable rates could benefit from targeted healthcare resources and prevention programs to help reduce the overall number of cases.

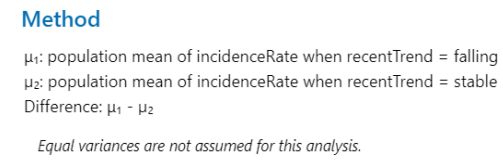


Figure 2.1.1

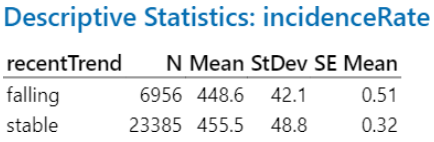


Figure 2.1.2

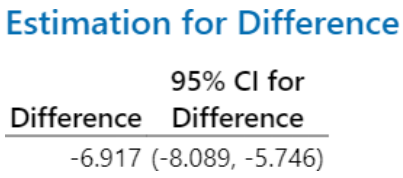


Figure 2.1.3

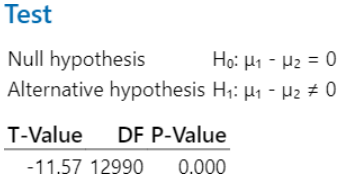


Figure 2.1.4

Impact of Study Counts on Cancer Incidence: Do More Studies Lead to Higher Rates?

To determine if the number of studies conducted in an area influences cancer rates, we divided the data into two groups: areas with a high study count and areas with a low study count. We created these groups by splitting the study count data at the average, so that areas above the average are categorized as "high" and those below are categorized as "low." This approach allowed us to compare cancer rates between areas with more studies and those with fewer studies.

The goal was to find out if areas with a higher number of studies or health initiatives experience any significant difference in cancer cases compared to areas with fewer studies. This information could help us understand whether higher study counts correlate with higher or lower cancer rates, providing insight into the potential impact of healthcare monitoring or research in these regions.

To analyze this, we conducted a two-sample test to compare the average cancer incidence rates between the "high" and "low" study count groups. In simple terms, we wanted to see if the difference in cancer rates between these two groups was real or just due to random chance.

Our findings showed that areas with a low study count have an average cancer incidence rate of around 453.2, while areas with a high study count have a slightly higher average rate of about 458.0. The probability of this difference occurring by chance was extremely low, as indicated by a p-value of 0.000, which is below our threshold of 0.05. This tells us that the difference is statistically significant, and we can be 95% confident that the true difference in cancer rates between high and low study count areas falls between about 2.86 and 6.80 cases.

In summary, areas with a higher study count show a slightly higher average cancer rate compared to those with a lower study count. This suggests that even in regions with more research and healthcare monitoring, cancer rates may remain elevated, potentially indicating a need for further preventive measures or a deeper investigation into underlying causes.

Summary

Lower End	Upper End	Recoded Value	Number of Rows
0	2.46459	low	30212
2.4646	1534.1	high	2339

Source data column studyCount
Recoded data column Recoded studyCount

Each interval includes its lower end.

Figure 2.2.1

Method

μ_1: population mean of incidenceRate when Recoded studyCount = low
μ_2: population mean of incidenceRate when Recoded studyCount = high
Difference: $\mu_1 - \mu_2$

Equal variances are assumed for this analysis.

Figure 2.2.2

Descriptive Statistics: incidenceRate

Recoded studyCount	N	Mean	StDev	SE Mean
low	30212	453.2	47.4	0.27
high	2339	458.0	39.6	0.82

Figure 2.2.3

Estimation for Difference

Difference	Pooled StDev	95% CI for Difference
-4.83	46.87	(-6.80, -2.86)

Figure 2.2.4

Test

Null hypothesis H_0: $\mu_1 - \mu_2 = 0$
Alternative hypothesis H_1: $\mu_1 - \mu_2 \neq 0$

T-Value	DF	P-Value
-4.80	32549	0.000

Figure 2.2.5

Income Levels and Cancer Incidence: Exploring the Link Between Economic Status and Cancer Rates

To understand if median income levels have an impact on cancer incidence rates, we began by categorizing our data into four income groups: very low, low, high, and very high. This process, known as "recoding," allowed us to compare cancer rates across different income brackets more effectively. We wanted to explore if income level could influence cancer rates, which would provide valuable insights into healthcare needs and potential preventive efforts across various economic backgrounds.

After creating these categories, we ran a test to see if the variation in cancer rates was different across these income groups. In simple terms, this test helps us understand whether cancer rates are more stable in some income groups and more unpredictable in others. Our results showed that areas with "very low" income had the most unpredictable cancer rates, with a higher spread or range, while other income groups had more consistent rates. The test result (p-value of 0.000) indicates that this difference is statistically significant, meaning it's unlikely to be due to random chance. In practical terms, this suggests that cancer rates are not only different across income levels but also vary in how steady or spread out they are within each group. Lower-income areas, with more unpredictable rates, may need additional healthcare support to manage this variation and address potentially fluctuating cancer cases more effectively.

Following this, we conducted a statistical comparison of average cancer incidence rates across the income groups using Welch's Test. This allowed us to determine if the observed differences in average cancer rates between income levels were significant. Our analysis found that cancer incidence rates slightly increase as median income rises, with averages ranging from 449.9 in the "very low" income group to 455.8 in the "very high" income group. The test confirmed that this difference is statistically significant (p-value = 0.000), suggesting that income does play a role in influencing cancer rates.

However, the R-squared value of 0.22% from our model indicates that income level explains only a small portion of the variation in cancer rates. In other words, while there is a statistically significant association between income and cancer incidence, income alone does not account for much of the difference. This finding highlights the complexity of cancer incidence and implies that other factors, beyond income, likely contribute more substantially to cancer rates.

In summary, our analysis shows that income levels are associated with both the average rate and variability of cancer incidence. Areas with higher median incomes have slightly higher cancer rates, but the difference is small. Additionally, the "very low" income group shows the most considerable variation in cancer rates, indicating a less consistent incidence pattern. These insights suggest that while income level has some influence on cancer rates, other factors also play a significant role, pointing to the need for a multi-faceted approach to cancer prevention and resource allocation across different economic backgrounds.

Summary

Lower End	Upper End	Recoded Value	Number of Rows
0	41800.1	very low	8138
41800.1	48544.1	low	8147
48544.1	55832.1	high	8141
55832.1	125635	very high	8125

Source data column medIncome
Recoded data column Recoded medIncome

Each interval includes its lower end.

Figure 3.1

Method

Null hypothesis	All means are equal
Alternative hypothesis	Not all means are equal
Significance level	α = 0.05

Equal variances were not assumed for the analysis.

Figure 3.2

Factor Information

Factor	Levels	Values
Recoded medIncome	4	very low, low, high, very high

Figure 3.3

Welch's Test

Source	DF Num	DF Den	F-Value	P-Value
Recoded medIncome	3	17972.9	18.41	0.000

Figure 3.4

Model Summary

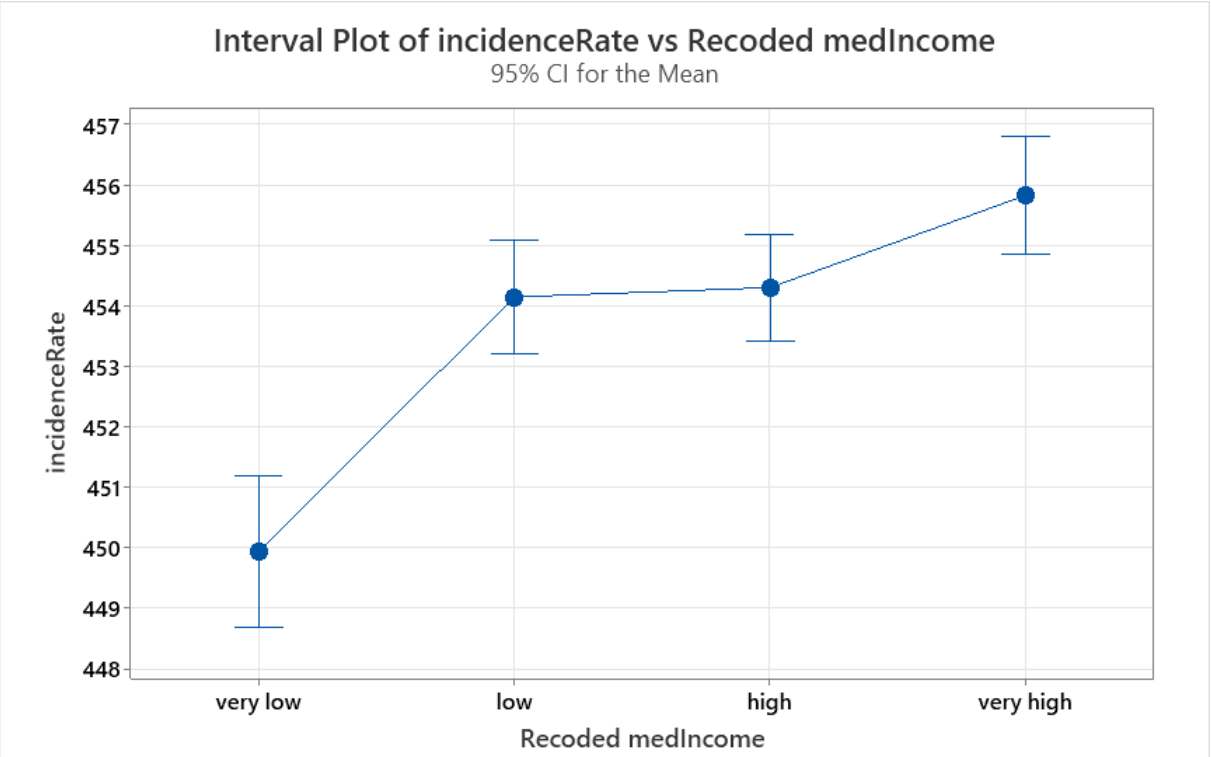
R-sq	R-sq(adj)	R-sq(pred)
0.22%	0.21%	0.19%

Figure 3.5

Means

Recoded medIncome	N	Mean	StDev	95% CI
very low	8138	449.934	57.359	(448.687, 451.180)
low	8147	454.143	43.563	(453.197, 455.089)
high	8141	454.299	40.345	(453.423, 455.176)
very high	8125	455.824	44.271	(454.862, 456.787)

Figure 3.6



Individual standard deviations are used to calculate the intervals.

Figure 3.7

Uncovering Key Factors Influencing Cancer Rates: A Closer Look at Income, Poverty, and Population

To understand what factors might be linked to cancer rates, we looked at how different things like income, poverty levels, and population size connect with cancer rates in our data. This process, called correlation analysis, gives us a sense of how closely each factor is related to cancer rates. Here's what we found:

1. Income and Cancer Rates: We found that income doesn't have much of an impact on cancer rates. In areas with different median incomes, the difference in cancer rates was minimal. So, income alone isn't a clear indicator of cancer risk.

2. Poverty and Cancer Rates: We looked at poverty in two ways: total poverty (the number of people below the poverty line) and poverty percentage (the percentage of people in poverty). Both showed very weak relationships with cancer rates, meaning that areas with higher poverty levels didn't necessarily have higher or lower cancer rates.

3. Population Size and Cancer Rates: The size of the population in an area also didn't show a strong connection with cancer rates. Whether an area was densely or sparsely populated, cancer rates didn't vary much.

4. Cancer Death Rate and Incidence Rate: One area where we saw a clear link was between cancer incidence (how often cancer occurs) and the death rate from cancer. This was the strongest connection in our data. Areas with higher cancer rates tended to also have higher death rates. This may suggest that where cancer is more common, it also tends to have worse outcomes.

5. Other Observations: We noticed that in larger populations, there were naturally more cases reported on average. This was expected, as more people usually mean more reported cases.

Overall, our analysis shows that factors like income and poverty don't have a strong influence on cancer rates. However, there is a noticeable link between how often cancer occurs and how deadly it is in those areas, which could be important for future research or planning.

Method

Correlation type	Pearson
Number of rows used	32551

Figure 4 .1

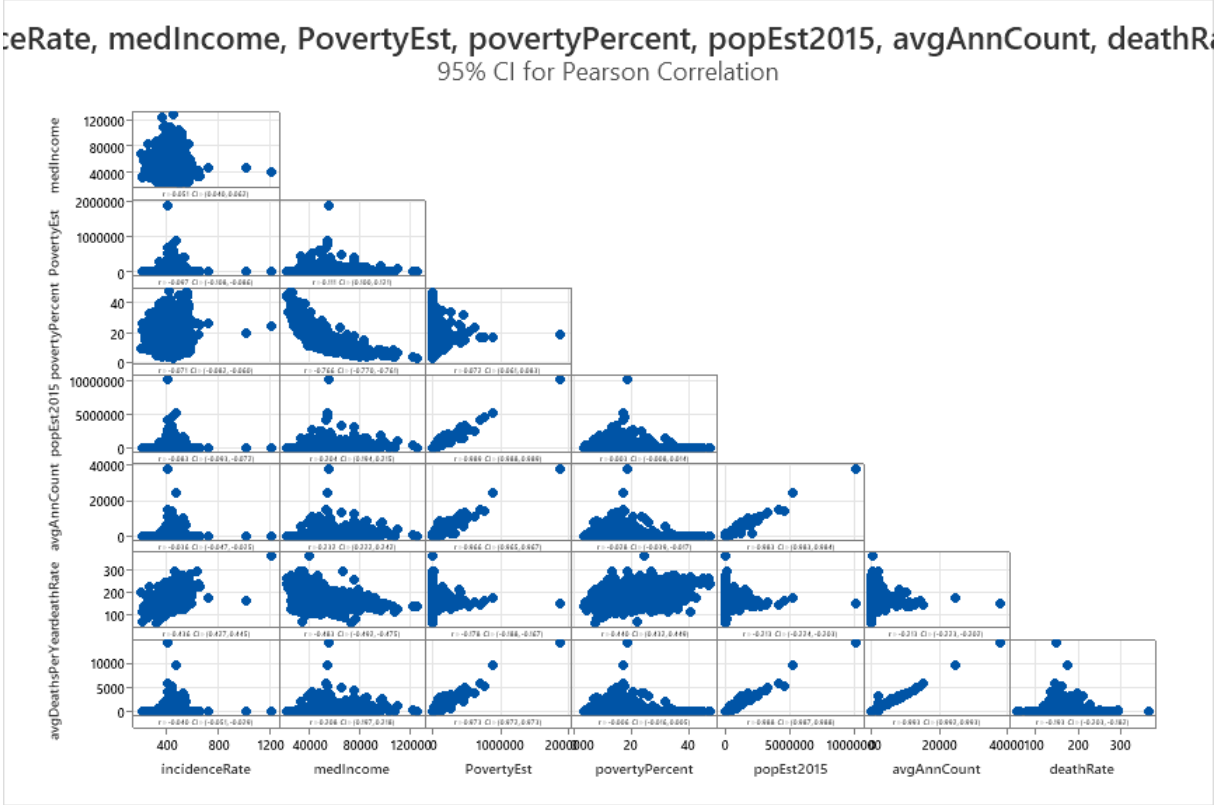


Figure 4.2

Correlations

	incidenceRate	medIncome	PovertyEst	povertyPercent	popEst2015
medIncome	0.051				
PovertyEst	-0.097	0.111			
povertyPercent	-0.071	-0.766	0.072		
popEst2015	-0.083	0.204	0.989	0.003	
avgAnnCount	-0.036	0.232	0.966	-0.028	0.983
deathRate	0.436	-0.483	-0.178	0.440	-0.213
avgDeathsPerYear	-0.040	0.208	0.973	-0.006	0.988

	avgAnnCount	deathRate
medIncome		
PovertyEst		
povertyPercent		
popEst2015		
avgAnnCount		
deathRate	-0.213	
avgDeathsPerYear	0.993	-0.193

Figure 4.3

Understanding Cancer Incidence: The Role of Death Rates, Income, and Trends

For this regression analysis, we examined how various factors might influence cancer incidence rates. Here's a straightforward breakdown:

We explored how variables like study count (the number of cases examined), poverty percentage, median income, the trend of rates over the past five years, and the overall death rate relate to the cancer incidence rate. This was done by creating a mathematical model to see how each factor contributes to or influences cancer rates.

We used the recent trend of cancer rates ("falling," "rising," or "stable") to categorize the data and understand if the incidence rates differ based on these trends.

Key Findings:

1. Death Rate Influence: Death rate (1.1701) is the most influential factor. This means that areas with higher death rates tend to have higher cancer incidence.

2. Income and Poverty: Poverty percentage (-1.0351) has a negative effect, indicating that higher poverty levels might correlate with lower reported cancer rates. Conversely, median income (0.000968) has a minor but positive influence.

3. Study Count and Trend Impact: Both the study count (0.2066) and five-year trend (1.9769) showed positive associations, though the impact is moderate.

4. Trend Differences: The trend of rates in each region ("falling," "rising," or "stable") showed distinct baseline effects on the cancer incidence rate.

The model explains approximately 31.7% of the variation in cancer incidence rates, which suggests other factors not in this model also contribute to the differences in cancer rates.

The residual plots (graphs showing the differences between actual and predicted values) suggest the model fits fairly well, with errors (or "residuals") being generally random. This consistency in residuals implies that the model's predictions are reasonable across the dataset.

In summary, this analysis helps highlight that death rates and economic conditions (poverty and income) play roles in cancer incidence, though there are likely other factors outside our model that also influence cancer rates.

Regression Equation

recentTrend	
falling	incidenceRate = 221.62 + 0.2066 studyCount - 1.0351 povertyPercent + 0.000968 medIncome + 1.1701 deathRate + 1.9769 fiveYearTrend
rising	incidenceRate = 232.14 + 0.2066 studyCount - 1.0351 povertyPercent + 0.000968 medIncome + 1.1701 deathRate + 1.9769 fiveYearTrend
stable	incidenceRate = 221.26 + 0.2066 studyCount - 1.0351 povertyPercent + 0.000968 medIncome + 1.1701 deathRate + 1.9769 fiveYearTrend

Figure 5.1.1

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	221.62	3.42	64.71	0.000	
studyCount	0.2066	0.0376	5.49	0.000	1.01
povertyPercent	-1.0351	0.0672	-15.40	0.000	2.71
medIncome	0.000968	0.000031	31.19	0.000	3.15
deathRate	1.1701	0.0108	108.12	0.000	1.29
fiveYearTrend	1.9769	0.0871	22.69	0.000	1.12
recentTrend					
rising	10.52	2.78	3.78	0.000	1.06
stable	-0.353	0.586	-0.60	0.547	1.34

Figure 5.1.2

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	7	17163776	2451968	1894.89	0.000
studyCount	1	38984	38984	30.13	0.000
povertyPercent	1	307018	307018	237.26	0.000
medIncome	1	1258606	1258606	972.66	0.000
deathRate	1	15126864	15126864	11690.09	0.000
fiveYearTrend	1	666446	666446	515.03	0.000
recentTrend	2	20916	10458	8.08	0.000
Error	28589	36993880	1294		
Lack-of-Fit	4911	36993880	7533	*	*
Pure Error	23678	0	0		
Total	28596	54157656			

Figure 5.1.3

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
35.9721	31.69%	31.68%	31.65%

Figure 5.1.4

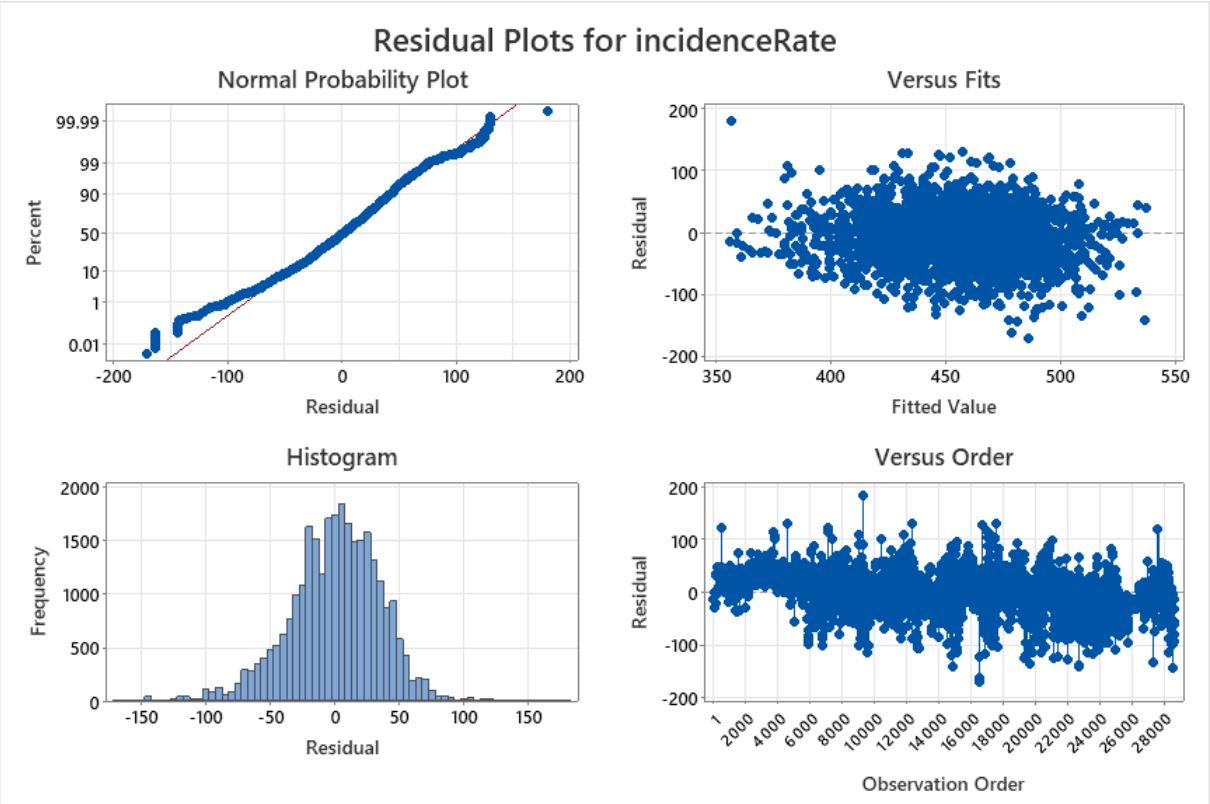


Figure 5.1.5

Understanding the Key Drivers of Cancer-Related Death Rates

In this regression analysis, we sought to understand how various factors might influence the death rate related to cancer. Specifically, we looked at the number of studies conducted (studyCount), poverty levels (povertyPercent), median income (medIncome), cancer incidence rates (incidenceRate), the trend of death rates over the past five years (fiveYearTrend), and the recent trend of cancer rates (falling, rising, or stable).

The goal was to develop a mathematical model that shows how each of these factors contributes to changes in the death rate.

Key findings:

1. **Incidence Rate Influence**: The most impactful factor was the cancer incidence rate, with a coefficient of 0.24805. This means that as the rate of new cancer cases increases, so does the death rate. This strong association suggests that areas with higher cancer occurrences tend to have worse outcomes in terms of mortality.

1. **Economic Factors**: Poverty percentage (0.5259) showed a positive association, meaning that as poverty levels increase, so does the death rate. Interestingly, median income (-0.000710) had a slight negative influence, indicating that areas with higher median incomes tend to have lower death rates, although the impact is relatively small.

2. **Study Count and Five-Year Trend**: The number of studies conducted (-0.0970) had a small but statistically significant negative impact on the death rate, suggesting that areas with more research may see slightly lower death rates. The five-year trend (-0.3350) also had a negative effect, implying that areas with improving trends in cancer death rates have better outcomes.

3. **Recent Trends**: The recent trend variable, which captures whether death rates are rising, falling, or stable, was also included. "Rising" had a positive coefficient (2.40), while "stable" had a more substantial effect (3.652), both indicating higher death rates compared to areas where death rates are falling.

The model explains approximately 44.9% of the variation in cancer death rates (R-squared value), meaning that other factors not included in the model likely contribute to the differences in death rates across areas. The residual plots showed that the errors between actual and predicted values are randomly distributed, which suggests that the model fits the data reasonably well without any major biases or patterns in the residuals.

In summary, this analysis highlights that cancer incidence rates are the most critical factor influencing death rates. Economic factors like poverty and income also play a role, though their impact is less pronounced. The model suggests that areas with higher poverty and more stable or rising trends in death rates might need more targeted interventions to address both cancer occurrences and outcomes.

Regression Equation

recentTrend

falling deathRate = 86.39 - 0.0970 studyCount + 0.5259 povertyPercent
- 0.000710 medIncome + 0.24805 incidenceRate - 0.3350 fiveYearTrend

rising deathRate = 88.79 - 0.0970 studyCount + 0.5259 povertyPercent
- 0.000710 medIncome + 0.24805 incidenceRate - 0.3350 fiveYearTrend

stable deathRate = 90.04 - 0.0970 studyCount + 0.5259 povertyPercent
- 0.000710 medIncome + 0.24805 incidenceRate - 0.3350 fiveYearTrend

Figure 5.2.1

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
16.5626	44.87%	44.86%	44.83%

Figure 5.2.2

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	86.39	1.61	53.68	0.000	
studyCount	-0.0970	0.0173	-5.60	0.000	1.01
povertyPercent	0.5259	0.0309	17.01	0.000	2.71
medIncome	-0.000710	0.000014	-51.00	0.000	2.98
incidenceRate	0.24805	0.00229	108.12	0.000	1.04
fiveYearTrend	-0.3350	0.0404	-8.29	0.000	1.14
recentTrend					
rising	2.40	1.28	1.87	0.061	1.06
stable	3.652	0.269	13.58	0.000	1.33

Figure 5.2.3

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	7	6382767	911824	3323.93	0.000
studyCount	1	8598	8598	31.34	0.000
povertyPercent	1	79389	79389	289.40	0.000
medIncome	1	713412	713412	2600.64	0.000
incidenceRate	1	3206840	3206840	11690.09	0.000
fiveYearTrend	1	18840	18840	68.68	0.000
recentTrend	2	50783	25392	92.56	0.000
Error	28589	7842568	274		
Lack-of-Fit	4911	7842568	1597	*	*
Pure Error	23678	0	0		
Total	28596	14225336			

Figure 5.2.4

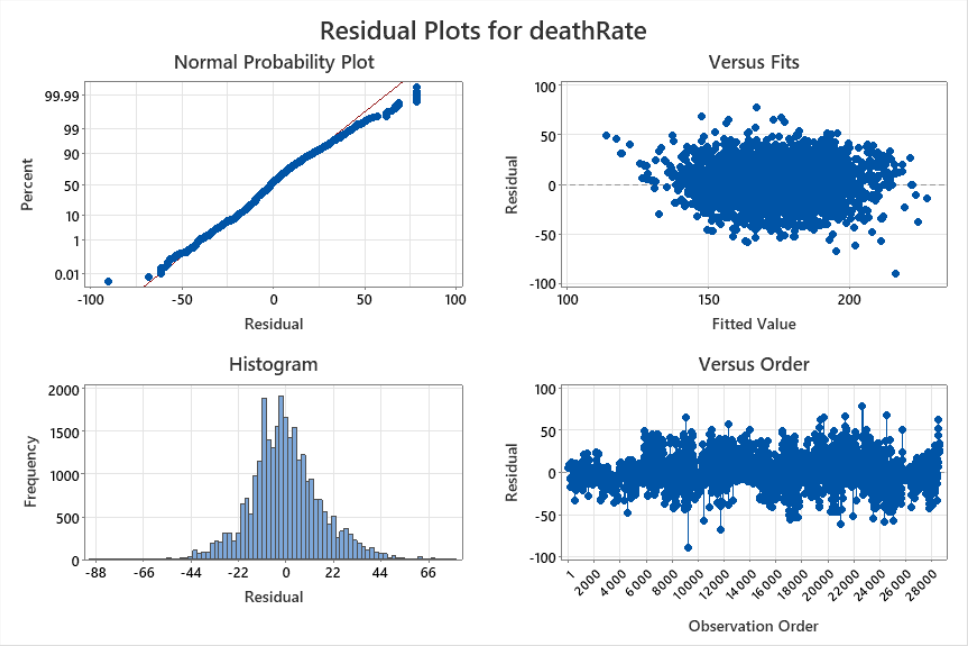


Figure 5.2.5

Conclusion

This report focused on analyzing the factors contributing to cancer incidence and death rates across different regions in the United States. Through data analysis, we explored several factors including regional cancer rates, study counts, income levels, poverty, population, and recent trends to understand their impact on cancer occurrence.

We began by identifying the regions most prone to cancer, discovering that Southern and Northeastern states such as Kentucky, Alabama, Louisiana, New York, and Delaware have higher cancer incidence rates compared to other regions. This suggests a need for greater healthcare resources and prevention programs in these areas to manage the higher prevalence of cancer cases.

When examining whether changes in cancer rates over time (falling vs. stable trends) influenced incidence, we found that even in areas where cancer rates were stable, cancer cases remained higher than in areas where rates were falling. This implies that even areas with consistent rates could benefit from ongoing healthcare support and prevention efforts.

Furthermore, we investigated the role of study count in cancer incidence. Interestingly, regions with a higher number of studies had slightly elevated cancer rates compared to areas with fewer studies. This finding suggests that while research and healthcare monitoring are critical, they alone are not enough to lower cancer rates, pointing to the complexity of cancer risk factors and the need for further exploration of underlying causes.

Income levels were also analyzed, and we found that areas with "very low" income levels exhibited the greatest variability in cancer rates. Although median income showed a minor positive association with cancer rates, the overall impact was small. However, the variability in lower-income areas highlights the need for more targeted healthcare interventions to address the unpredictable nature of cancer incidence in these regions.

Lastly, our regression analysis revealed that the death rate was the most significant predictor of cancer incidence. Other factors such as study count, poverty percentage, and median income played a role, but the model explained only about 31.7% of the variation in cancer rates, suggesting that additional factors are influencing cancer incidence beyond those we examined.

In conclusion, our comprehensive analysis highlights that cancer incidence is influenced by a variety of factors, with some regions and demographic groups more affected than others. While death rates and economic conditions play a role, the data points to the complexity of cancer incidence, suggesting that a multi-pronged approach combining research, healthcare resources, and targeted interventions is needed to address cancer effectively across the U.S.

References

American Cancer Society. (2023). *Cancer Facts & Figures 2023*. Atlanta: American Cancer Society. Retrieved from https://www.cancer.org/research/cancer-facts-statistics.html

DeSantis, C. E., Ma, J., Gaudet, M. M., Newman, L. A., Miller, K. D., Sauer, A. G., Jemal, A., & Siegel, R. L. (2019). Breast cancer statistics, 2019. *CA: A Cancer Journal for Clinicians*, 69(6), 438-451. https://doi.org/10.3322/caac.21583

National Cancer Institute. (2020). *Cancer Trends Progress Report*. Bethesda, MD: National Cancer Institute. Retrieved from https://progressreport.cancer.gov/

Siegel, R. L., Miller, K. D., & Jemal, A. (2020). Cancer statistics, 2020. *CA: A Cancer Journal for Clinicians*, 70(1), 7-30. https://doi.org/10.3322/caac.21590

Smith, R. A., Andrews, K. S., Brooks, D., Fedewa, S. A., Manassaram-Baptiste, D., Saslow, D., Brawley, O. W., & Wender, R. C. (2019). Cancer screening in the United States, 2019: A review of current American Cancer Society guidelines and current issues in cancer screening. *CA: A Cancer Journal for Clinicians*, 69(3), 184-210. https://doi.org/10.3322/caac.21557