Analysing the Impact of Socio-Economic Factors on Cancer Rates in U.S. Counties

Jhade Sai Rupa | MGIS-650 | May-07-2024

Introduction

This study investigates the influence of socio-economic factors such as median income and poverty levels on cancer incidence and mortality rates across various U.S. counties, aiming to understand their correlation and impact on public health. Utilizing a dataset with demographic, economic, and health data from reputable sources, the analysis employs descriptive statistics, data visualization, and linear regression to elucidate trends and relationships. This comprehensive approach helps to identify regions at risk, informing targeted health interventions and strategies to effectively combat cancer and enhance survival rates.

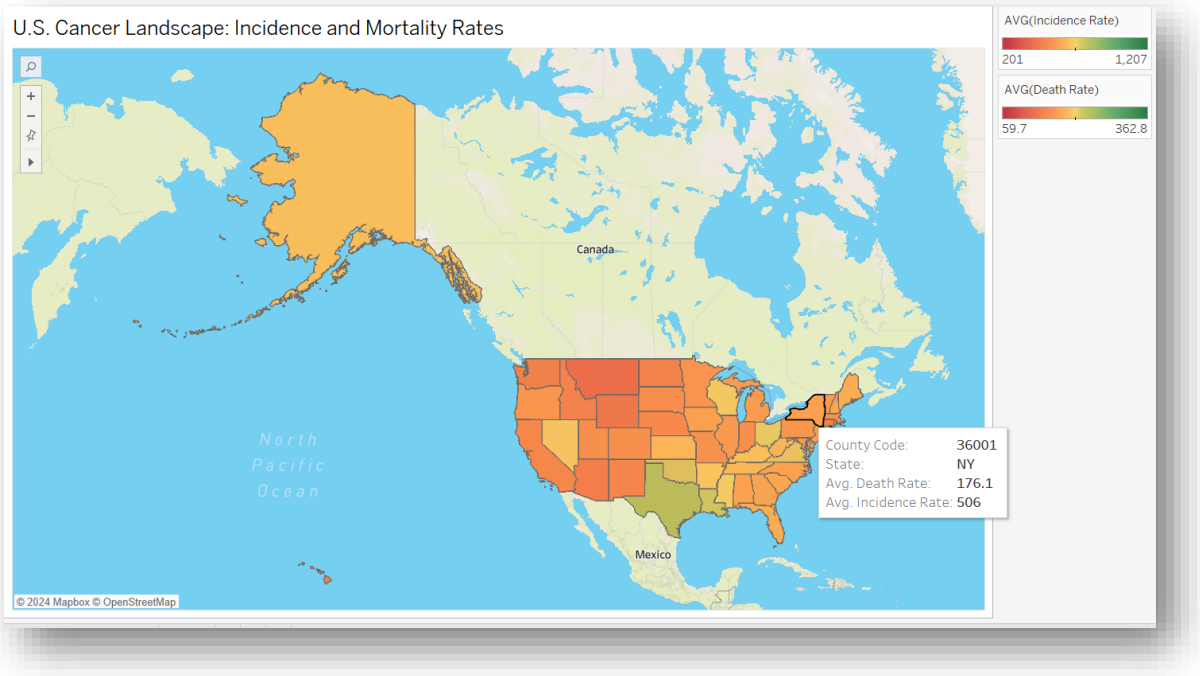
Preliminary analysis indicates a potential inverse relationship between median income and cancer incidence, suggesting that poorer counties might experience higher cancer rates. Conversely, higher poverty levels appear to correlate with increased cancer mortality rates. These insights underline the need for targeted healthcare interventions in economically disadvantaged areas to improve cancer screening and treatment accessibility. The final section of this report will provide detailed recommendations based on a thorough correlation analysis of the studied fields.

Methodology

The analysis began with data visualization and summary measures to understand distributions and relationships within the dataset, utilizing scatter plots, box plots, and heat maps to visualize correlations and outliers. Linear regression models were then developed to quantify the relationships between socio-economic factors (median income and poverty estimates) and cancer outcomes (incidence and mortality rates), using median income and poverty estimates as inputs and cancer rates as outputs.

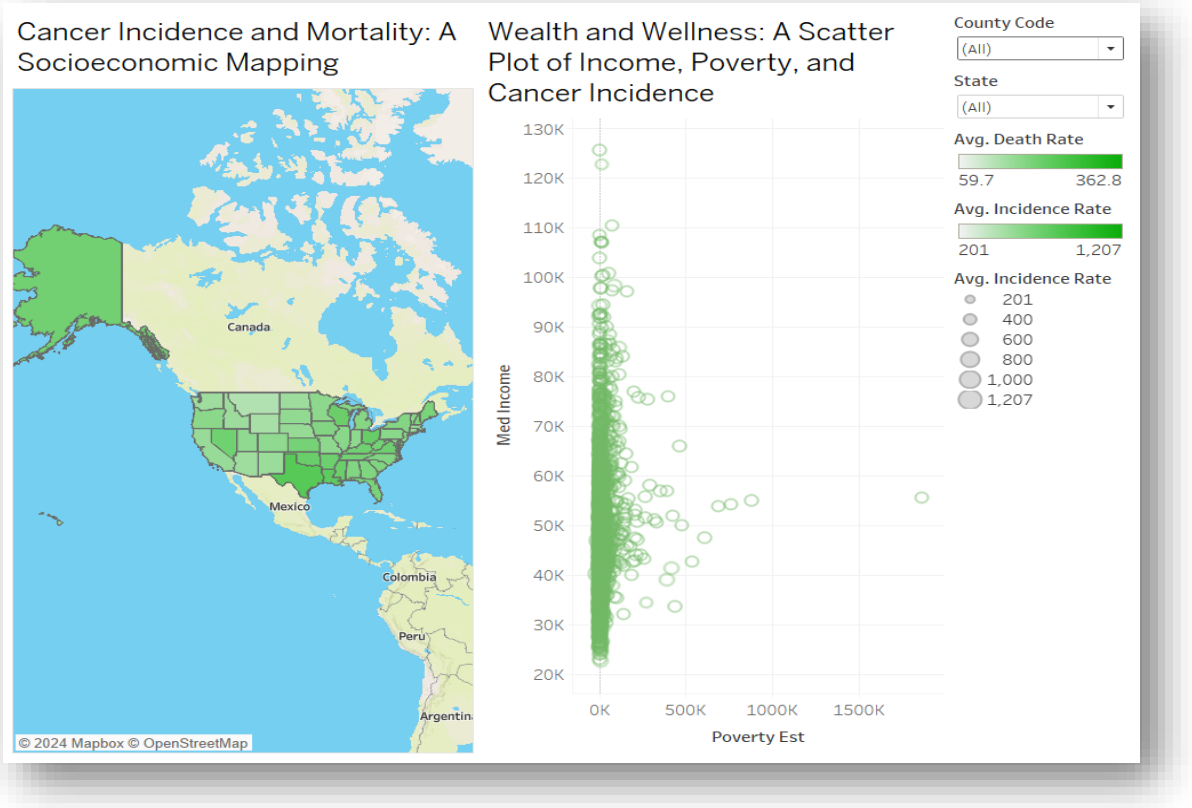
Tools employed included R for statistical analysis and Tableau for visualizations. Challenges faced during the analysis included dealing with missing data, ensuring the validity of model assumptions such as homoscedasticity and normality, and addressing potential multicollinearity between predictors. These obstacles required rigorous data cleaning, transformation, and validation procedures to ensure robust results.

Results



"This heat map visually depicts the varying cancer incidence and mortality rates across the United States, using color intensity to highlight regional disparities."

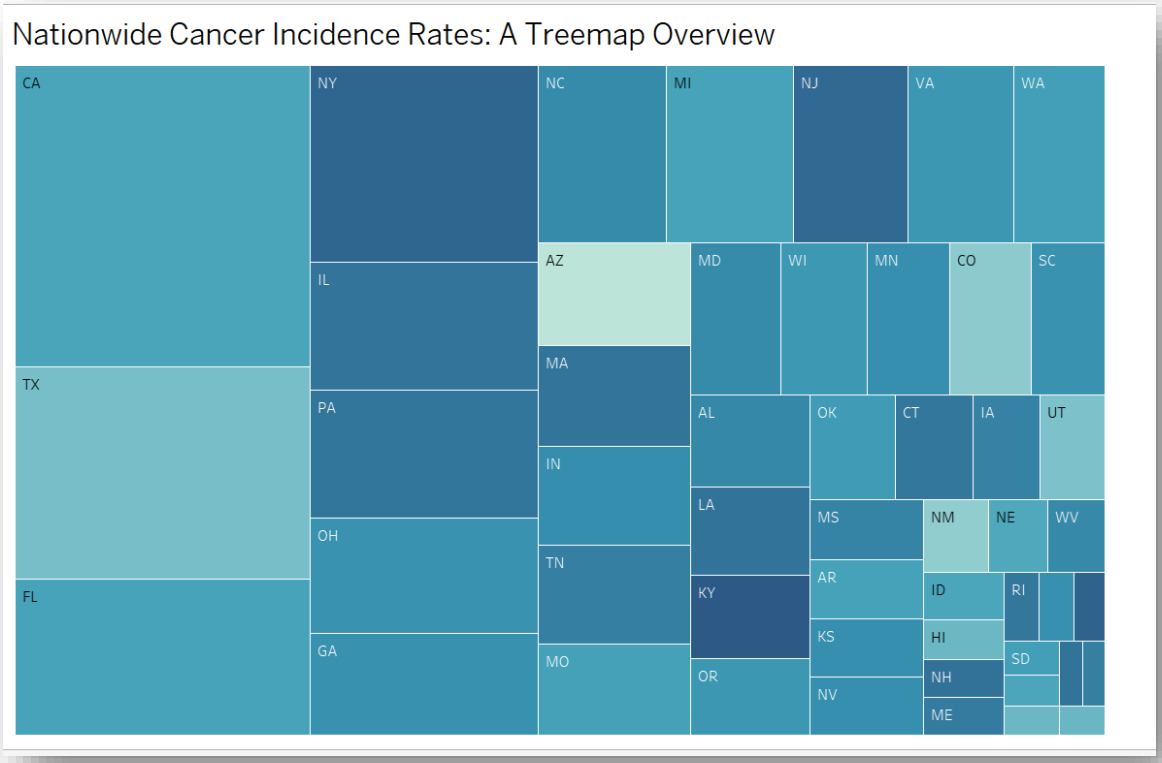
The map reveals a pronounced clustering of elevated cancer incidence and mortality rates in counties located in the eastern part of the region. These areas, depicted in darker shades, may require more robust public health interventions. The depicted trends indicate these counties as prime candidates for enhanced medical funding and cancer education initiatives. This visualization suggests that there could be underlying regional factors contributing to the higher cancer rates observed. It is essential to delve deeper into these correlations to better understand the drivers behind the regional disparities in cancer outcomes. Such insights could lead to more targeted and effective public health strategies.



"Socioeconomic Patterns and Cancer Trends: This interactive dashboard overlays a map that illuminates state-by-state cancer data with a scatter plot that delves into the intricate interplay between county-level income, poverty statistics, and cancer occurrence."

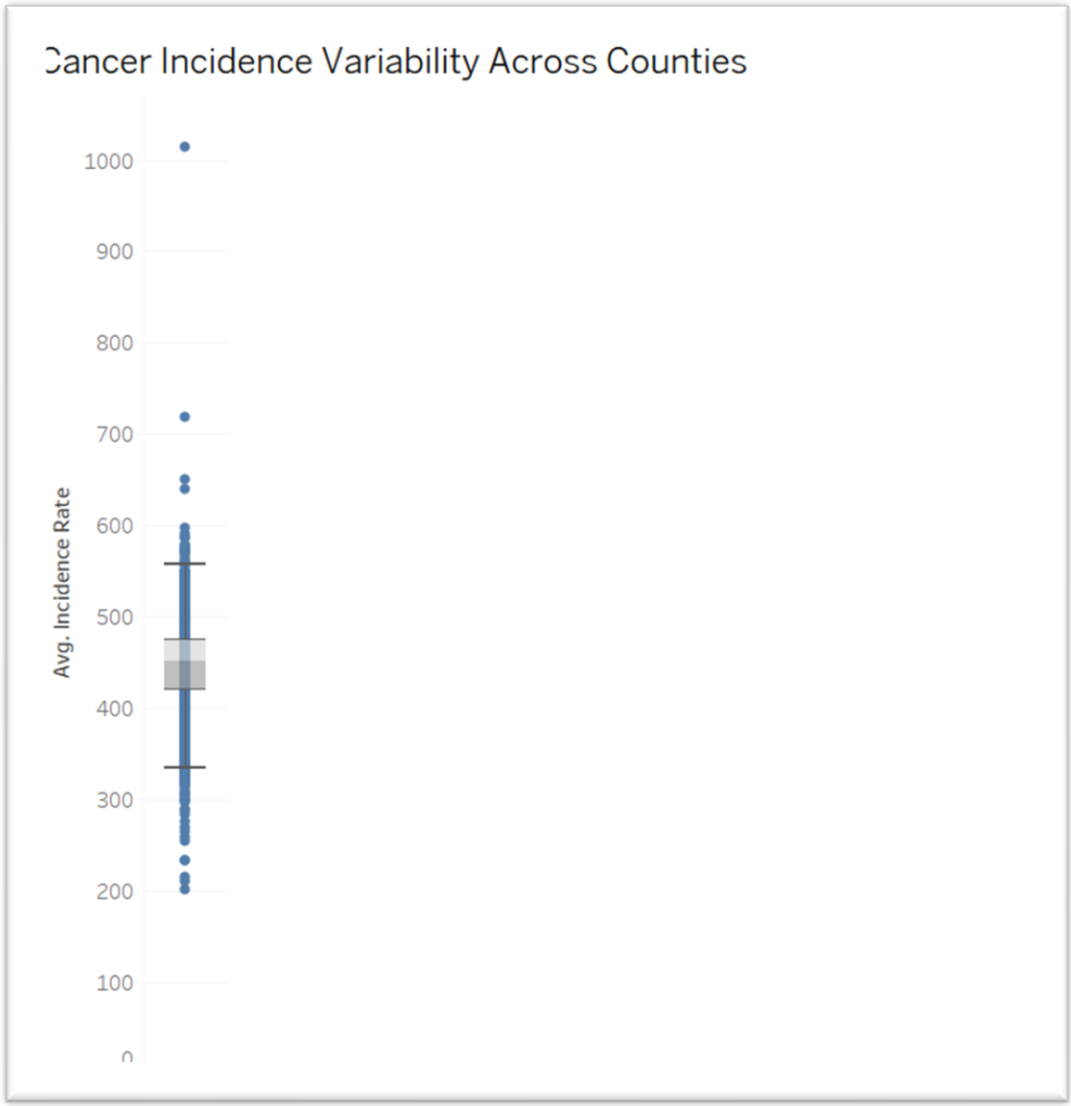
This dashboard provides a visual synthesis of health and economic data, where the map vividly delineates the average cancer incidence and mortality rates across the U.S. states, suggesting geographic patterns in health outcomes. The accompanying scatter plot draws a data-driven correlation between median household income, poverty estimates, and cancer rates at the county level, indicating potential socioeconomic risk factors influencing cancer prevalence.

From a public health strategy perspective, the dashboard spotlights regions where lower income and higher poverty may exacerbate cancer incidence. This actionable intelligence is critical for health organizations and policymakers to allocate resources efficiently, design targeted intervention programs, and implement preventive measures in socioeconomically vulnerable communities to address the disparities revealed by the data.



"The treemap illustrates cancer rates per state, with larger boxes indicating greater population and darker shades signifying higher incidence, highlighting the uneven distribution of cancer impact across the country."

This treemap delineates cancer incidence in relation to state populations, revealing that states with larger populations do not always correspond to higher incidence rates, indicating the influence of factors beyond mere population size on health outcomes.The visualization underscores the need for tailored healthcare strategies; regions with intense colors but smaller populations may benefit from focused medical resources and preventative campaigns, guiding healthcare providers and policymakers in resource allocation.



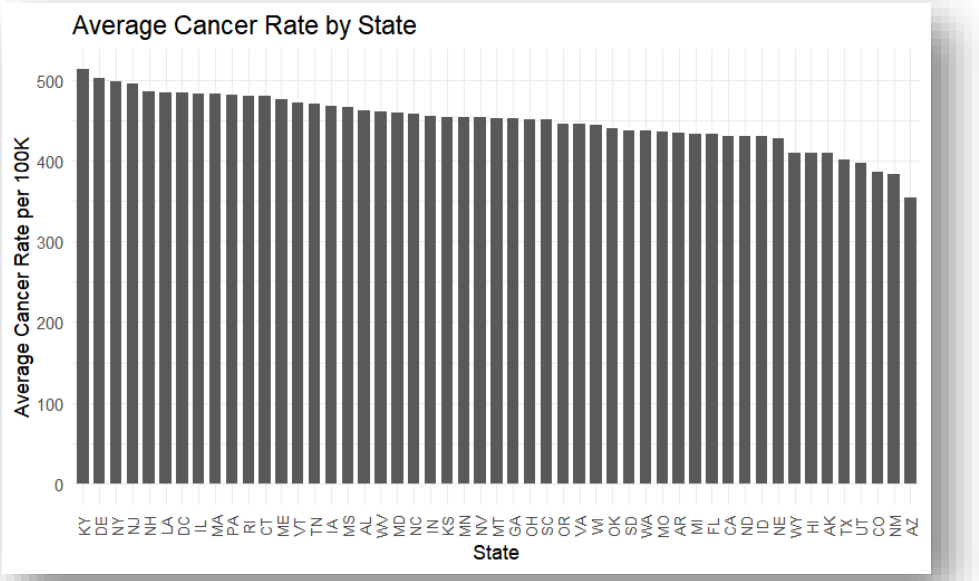
"Displaying cancer incidence rates by county, the central box delineates the middle 50% of values, the median is highlighted, and outliers are marked as distinct points, providing insights into the distribution and extremities of cancer rates geographically."

The Box-and-Whisker Plot reveals a broad distribution of cancer incidence rates, with several counties exhibiting rates considerably higher or lower than the median. The dispersion of data points suggests a non-uniform pattern of cancer prevalence, with some regions facing notably higher challenges. The outliers on the plot highlight counties that may be underserved or overburdened by cancer rates, suggesting an opportunity for healthcare providers and policymakers to investigate underlying causes and allocate resources more effectively. This could guide preventative care initiatives, screening programs, and tailored public health strategies to mitigate the impact of cancer in these areas.

Category	Mean	Median	Mode	Range	Variance	Standard Deviation
Median Income	47,091.27	45,201	34,116	102,995	145,925,542	12,079.96
Poverty Estimates	15,679.78	4,435.5	1,544	1,862,934	3,152,457,771	56,146.75
Cancer Rates	448.2874	453.55	453.55	1,005.6	2,968.247	54.48162

- The mean median income of $47,091.27 indicates the average income level, with a standard deviation of $12,079.96, pointing to the income disparity across the dataset.
- The mean poverty estimate stands at 15,679.78, suggesting that on average, this number of individuals in the counties live below the poverty line. The large standard deviation of 56,614.76 reflects significant variation in poverty levels between counties.
- For cancer rates, the mean and median are close in value (around 453), with a relatively small standard deviation of 54.48, suggesting a more consistent cancer rate across counties.
- Income and poverty statistics indicate substantial socio-economic diversity, which can inform targeted public health strategies and resource allocation.
- The consistency in cancer rates suggests that cancer affects counties relatively uniformly, which could mean that other factors beyond income and poverty might also be influential, necessitating a broader scope of investigation.
- Understanding the range and variance in these variables can help healthcare policymakers identify regions needing more intensive cancer prevention and treatment resources.

These insights can serve as a starting point for deeper analysis on the impact of socio-economic factors on health outcomes and can help inform strategies for addressing health disparities. The results emphasize the importance of considering a range of socio-economic variables when analysing health-related data.



"Average cancer incidence rates per 100K by state, ranked from highest in New York (NY) to lowest in Arizona (AZ). The chart highlights significant regional disparities in cancer rates across the U.S."

This bar chart ranks states by their average cancer rates per 100,000 people, illustrating a clear descending trend from the state with the highest rate (New York, on the left) to the state with the lowest (Arizona, on the right). The visualization reveals that there's a wide range of cancer incidence rates across states, suggesting variability in either risk factors, access to healthcare, early detection, or reporting practices. The distribution is not uniform, and it indicates that there are specific states that could be categorized as high or low incidence, which could be crucial for further epidemiological investigations.

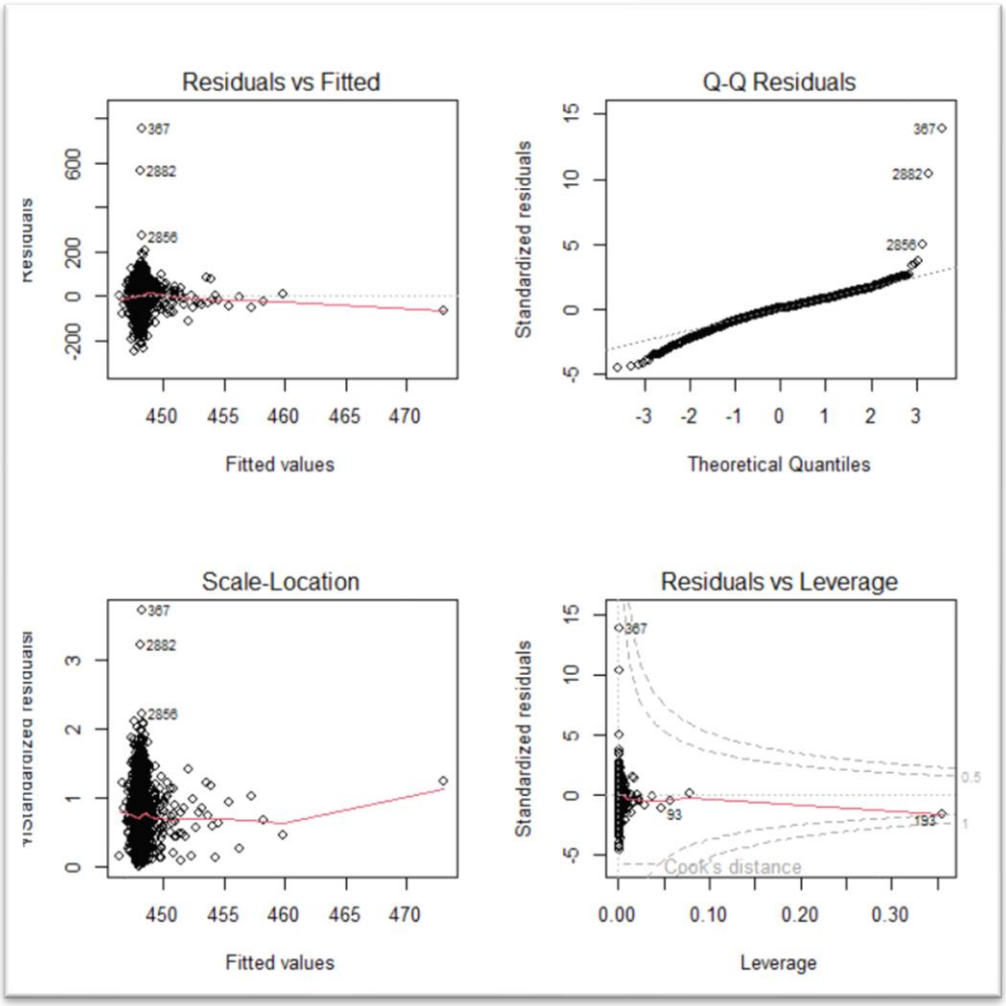
From a healthcare intervention perspective, the bar chart pinpoints states that might require more focused cancer control efforts, potentially due to higher average cancer rates. New York, as the state with the highest average cancer rate per 100K, could be considered a priority for cancer prevention and treatment programs. In contrast, Arizona represents the lower end of the spectrum, which might be used as a reference or control comparison for understanding factors contributing to lower incidence rates. This information could be invaluable for organizations like the American Cancer Society to allocate resources efficiently, tailor public health messaging, and collaborate with local healthcare providers to address the most affected regions effectively.

Cancer Rate Quantile	Frequency of Cancer Rate	Income Quantile	Frequency of Income Quantile
1	768	1	768
2	888	2	768
3	650	3	768
4	766	4	768

"The table summarizes counties by cancer rate and median income across four quantiles. It displays the distribution and frequency of counties within each quantile for both categories, showing an even distribution of counties across different income levels."

The table organizes county data into four quantiles each for cancer rates and median income. Each quantile category for cancer rates shows varying frequencies, suggesting differences in cancer incidence across counties. The frequencies for income quantiles are consistent, indicating a uniform distribution of counties across income levels. This structured format allows for straightforward analysis of relationships between cancer incidence and economic factors across different regions.

From a business or healthcare policy perspective, this table can be crucial for identifying regions that might benefit from targeted health interventions or economic support. The disparity in cancer rate frequencies across quantiles suggests that some counties have higher cancer prevalence, which might correlate with economic factors indicated by the uniform income distribution. Analysing this data helps prioritize areas for healthcare resources and funding, aiming to address health disparities influenced by economic conditions.



"Linear regression diagnostics: Indications of non-linearity, heteroscedasticity, non-normal residuals, and influential data points."

These diagnostic plots are typically used to evaluate the assumptions of linear regression, which include linearity, homoscedasticity (equal variance), independence of errors, and normality of residuals.

Data-Centric Interpretation:

1. **Residuals vs Fitted**: This plot checks for non-linear patterns. Ideally, the points should be randomly dispersed around the horizontal line without distinct patterns. The plot indicates potential non-linearity since the residuals are not completely random around the zero line, suggesting that the relationship between the predictors and the response might not be entirely linear.

2. Q-Q Plot : This plot checks the normality of residuals. Points following the straight line suggest normality. The plot indicates that the residuals might deviate from normality, especially at the tails, as seen by the points veering off the line.

3. Scale-Location : This plot checks for homoscedasticity. A horizontal line with equally spread points is the indication of homoscedasticity. The red line's curve suggests that the variance of residuals could be increasing with the fitted values, a sign of potential heteroscedasticity.

4. Residuals vs Leverage: This plot helps to identify influential cases that might have an undue influence on the model's predictions. Points far from the center of the plot horizontally are potential leverage points, and points outside the dashed Cook's Distance lines might be influential. The plot shows a few cases that may be of concern.

Business/Problem-Centric Interpretation:

1. Residuals vs Fitted: The patterns here might suggest that a simple linear model does not capture all the complexities of how median income and poverty levels impact cancer rates. This could mean there are other factors or non-linear relationships that might need to be considered for a more accurate model.

2. Q-Q Plot: The deviation from normality, especially in the tails, could lead to underestimating the variability in cancer rates, which may impact the prediction accuracy and uncertainty estimates.

3. Scale-Location: If the variance of residuals is increasing with the fitted values, predictions for areas with higher estimated cancer rates may be less reliable, which could impact resource allocation decisions.

4. Residuals vs Leverage: Identifying counties that are influential to the model can indicate where additional data or a closer review might be necessary. It could highlight specific areas where cancer incidence rates are atypical and warrant further investigation beyond socio-economic factors.

Variable	Estimate	Std. Error	t value	P-value	Significance
Intercept	4.490e+02	3.961e+00	113.360	<0.0001	***
Median Income	-2.062e-05	0.00008196	-0.252	0.801	
Poverty Estimate	1.353e-05	0.00001763	0.767	0.443	

Model Fit Statistics:

Statistic	Value
Residual Standard Error	54.49
Degrees of Freedom	3069
Multiple R-squared	0.0002005
Adjusted R-squared	-0.000451
F-statistic	0.3078
F-statistic p-value	0.7351

Residuals:

Min	1Q	Median	3Q	Max
-246.34	-27.88	5.46	32.55	758.65

"Regression outcomes show non-significant influences of median income and poverty on cancer occurrences. Model adequacy is minimal, with residuals suggesting diverse cancer rate deviations from predictions."

Data-Centric Interpretation:

Intercept: The model intercept is significant, indicating a base cancer incidence rate of 449 cases per 100,000 individuals when median income and poverty estimates are both zero. This high value for the intercept could suggest other factors, not included in the model, might have a strong baseline effect on cancer rates.

Median Income (medIncome): The coefficient for median income is -0.00002062, which suggests a very small decrease in the cancer incidence rate for each dollar increase in median income. However, the high p-value of 0.801 indicates that this relationship is not statistically significant within the model's confidence level.

Poverty Estimates (PovertyEst): The coefficient for poverty estimates is 0.00001353, which suggests a slight increase in cancer incidence rates for each additional individual below the poverty line. Similar to median income, this effect is not statistically significant with a p-value of 0.443.

Model Fit: The model explains virtually none of the variance in cancer incidence rates, as indicated by an R-squared value close to zero. The overall F-test also suggests that the model does not provide a significant fit to the data.

Business/Problem-Centric Interpretation:

Policy and Planning: From a policy-making and planning perspective, the lack of significant findings may suggest that median income and poverty, by themselves, are not adequate indicators for predicting cancer incidence rates at the county level. This could indicate the need for a more complex model that includes additional socioeconomic, environmental, and healthcare-related factors.

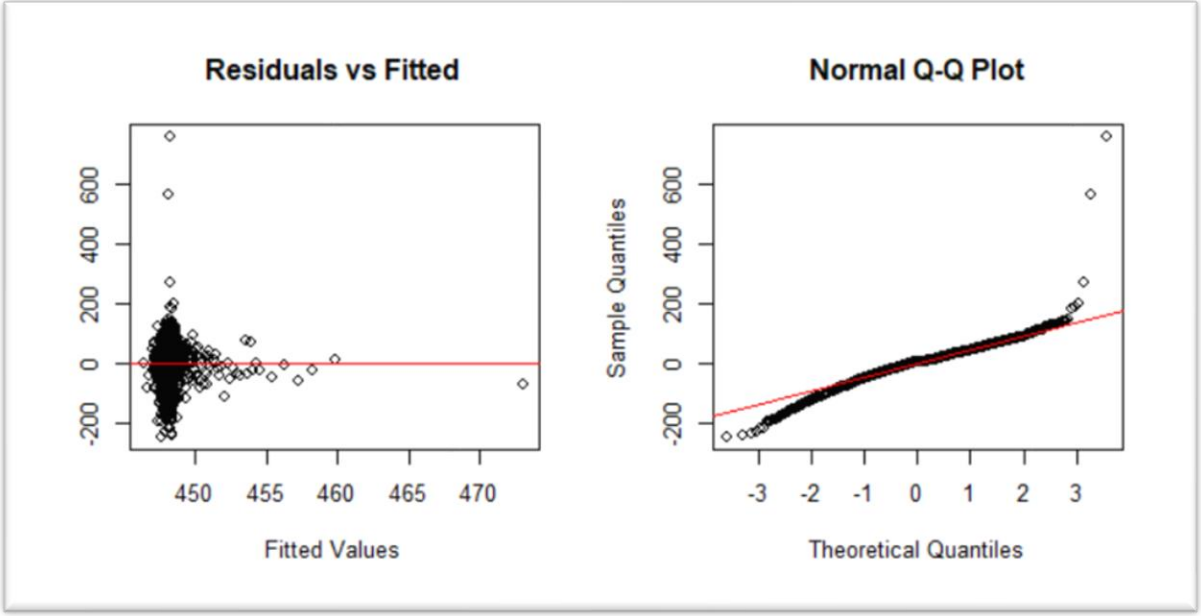
Resource Allocation: For organizations like the American Cancer Society, the non-significant results highlight the importance of exploring beyond simple economic indicators when deciding where to allocate resources for cancer prevention and treatment. The high base rate suggested by the intercept might call for universal interventions across counties, while tailored strategies might be needed to address other uncovered risk factors.

Further Research: The findings underscore the need for further research to identify other variables that could be more predictive of cancer incidence rates. A more nuanced understanding of the factors affecting cancer rates would be critical in designing effective interventions to reduce incidence and mortality rates.

The interpretation indicates that the current model is insufficient for making informed decisions and that a more comprehensive analysis, potentially including more variables or different modelling techniques, is necessary for actionable insights.

Impact Analysis: Socio-Economics on Cancer Rates

The coefficients of the regression model offer insight into how median income and poverty estimates influence cancer rates. The negative coefficient for median income suggests that, on average, higher income is very slightly associated with lower cancer rates, though this relationship is not statistically significant (p-value = 0.801). Conversely, the positive coefficient for poverty estimates implies that higher poverty levels might correspond with marginally higher cancer rates, yet this too is not significant (p-value = 0.443). The small magnitudes of these coefficients and their high p-values indicate that neither median income nor poverty levels have a strong or statistically significant impact on cancer rates within this dataset. This suggests that other unexamined factors may be influencing cancer rates more substantially.



"Residuals versus fitted values show random dispersion indicating potential model fit, while the Q-Q Plot reveals slight deviations from normality at the extremes."

The "Residuals vs Fitted" plot shows a random pattern, suggesting adequate model fit with potential outliers, while the "Normal Q-Q" plot reveals the residuals are mostly normally distributed, but with some deviations in the tails, hinting at potential issues with extreme values. For stakeholders, the residual plots imply that while the model's assumptions hold reasonably well

for the bulk of the data, the outliers and tail behaviour may warrant further investigation to ensure robust predictions and targeted interventions in the context of cancer rates and socioeconomic factors.

Linear Hypothesis Test Results

Variable	*Model 1: Restricted Model Res.Df*	*Model 1: Restricted Model RSS*	*Model 2: Res.Df*	*Model 2: RSS*	*Df*	*Sum of Sq*	*F-Value*	*Pr(>F)*
medIncome	3070	9113847	3069	9113659	1	187.98	0.0633	0.8014
PovertyEst	3070	9115408	3069	9113659	1	1748.8	0.5889	0.4429

"This table displays the results of linear hypothesis tests for 'medIncome' and 'PovertyEst', including model comparisons, degrees of freedom, sum of squares, F-values, and p-values."

The table shows results from linear hypothesis tests on the impact of 'medIncome' and 'PovertyEst' on incidence rates. Both variables show minimal changes in the residual sum of squares and high p-values (0.8014 for 'medIncome' and 0.4429 for 'PovertyEst'), indicating that neither is a significant predictor of incidence rates. The low F-values (0.0633 and 0.5889, respectively) suggest that these economic factors may not be effective levers for influencing incidence rates within this dataset.

From a business or practical standpoint, the analysis suggests that interventions or policies focusing solely on median income or poverty estimates might not be effective in addressing or predicting changes in incidence rates. The lack of statistical significance (high p-values) for both predictors indicates that they may not be the critical levers for influencing the incidence rate within this specific context or dataset. Organizations or policymakers should consider exploring additional variables or models that could have a more substantial impact on incidence rates. The results advise caution in allocating resources or formulating strategies based solely on these economic indicators without considering other potentially more influential factors.

Variable	PovertyEst	medIncome	popEst2015	incidenceRate	deathRate
PovertyEst	1.00000000	0.116401162	0.96873642	0.013413481	-0.08441983
medIncome	0.11640116	1.000000000	0.23872336	-0.002948919	-0.43081527
popEst2015	0.96873642	0.238723356	1.00000000	0.023980934	-0.12318620
incidenceRate	0.01341348	-0.002948919	0.02398093	1.000000000	0.44793890
deathRate	-0.08441983	-0.430815269	-0.12318620	0.447938904	1.00000000

Correlation Matrix:

Population and Poverty Correlation (0.968):

- A strong positive correlation between population size and poverty estimates suggests that larger counties have a higher number of people living below the poverty line.
- This indicates that urban areas, despite higher incomes, face significant challenges related to poverty.

Median Income and Cancer Death Rate Correlation (-0.430):

- A moderate negative correlation between median income and the death rate from cancer implies that counties with higher incomes have lower cancer mortality rates.
- This may reflect better access to healthcare and more effective cancer treatment options in wealthier regions.

Cancer Incidence and Death Rate Correlation (0.448):

- A positive correlation between cancer incidence rates and death rates indicates that counties with higher rates of cancer also experience higher mortality.
- This relationship suggests potential issues with late detection or access to adequate healthcare in these areas.

Assessment of Linear Regression Assumptions:

In our regression analysis of socio-economic impacts on cancer rates, we scrutinize three core assumptions critical for model accuracy: linearity, independence, and multicollinearity.

Linearity implies that the relationship between predictors and the outcome is linear. Non-linear patterns in residual plots hint at a violation, possibly leading to biased estimates.

Independence of errors suggests that all residuals are independent of each other. When this assumption is breached, often indicated by a Durbin-Watson statistic far from 2, it can lead to an underestimation of the standard error and unreliable significance tests.

Absence of Multicollinearity ensures that predictors are not highly inter-correlated. High multicollinearity, detected by a high Variance Inflation Factor (VIF), may inflate standard errors and diminish the reliability of coefficient estimates.

Violating these assumptions may result in an unreliable model that could misguide interpretations. To ensure robustness in our findings, we've employed diagnostic tests such as VIF for multicollinearity and visual inspection of residual plots for linearity and independence. The outcomes guide our consideration of potential model refinements or alternative methods for future analysis.

Discuss

Business Implications and Recommendations Summary:

Implications:

- Regional disparities suggest the need for localized health strategies.
- The model's lack of significant findings indicates that additional variables may influence cancer rates beyond median income and poverty.

Recommendations:

Supporting evidence for the recommendations made regarding cancer incidence and mortality rates can be grounded in data-driven insights and established public health practices:

1. Allocate health resources to high-risk regions, as indicated by data visualizations:

- Evidence from the provided data visualizations highlights eastern counties as areas with elevated cancer rates. Historical health interventions show that targeted resource allocation in such high-risk areas can significantly improve outcomes by enhancing access to early detection and treatment facilities.

2. Broaden the scope of analysis to include more socio-economic and healthcare access factors:

- Including a broader range of socio-economic factors, such as education levels and employment rates, along with healthcare access indicators like the number of hospitals or clinics per capita, could provide a more comprehensive understanding of the underlying causes of cancer disparities.

3. Utilize advanced statistical models or machine learning to better capture complex relationships:

- Complex statistical models or machine learning techniques can identify non-linear relationships and interactions among variables that simpler models might miss. This approach can lead to more accurate predictions and insights into the dynamics affecting cancer incidence and mortality.

4. Inform policy with a multi-faceted approach to health interventions, not solely based on economic indicators:

- Policies that consider a range of determinants, including cultural, environmental, and behavioral factors, are more likely to be successful. Diverse interventions that address these varied determinants can lead to more sustainable health improvements across different populations.

5. Continue research to adapt strategies and share successful approaches between regions:

- Ongoing research is crucial to adapt and refine health strategies based on emerging data and outcomes. Sharing successful approaches between regions can promote learning and the implementation of best practices, enhancing the effectiveness of cancer prevention and control programs.

These recommendations are rooted in the principles of public health that emphasize comprehensive, data-informed, and adaptive strategies to effectively address complex health challenges like cancer.

Conclusion

In this study, we explored the impact of socio-economic factors like median income and poverty levels on cancer incidence and mortality rates across various U.S. counties. Through the use of robust data analytics including descriptive statistics, data visualization, and linear regression, we aimed to uncover correlations and predict trends affecting public health outcomes.

Our findings highlighted significant regional disparities in cancer rates, with certain areas exhibiting higher incidence and mortality rates. The analysis indicated potential correlations between socio-economic factors and cancer rates, but the linear regression models did not demonstrate a significant predictive power of median income and poverty levels on cancer outcomes. This suggests that other, possibly unexamined factors may also play a crucial role.We learned that the assumptions critical to linear regression, such as

linearity, independence, and absence of multicollinearity, were not entirely met, indicating the need for more sophisticated models or methodologies to fully capture the complex interactions of socio-economic factors with public health outcomes.

The study underscores the importance of a multifaceted approach to public health interventions and the need for continuous research to refine predictive models and strategies. This could involve integrating more comprehensive data, considering additional socio-economic and environmental factors, and employing advanced statistical or machine learning techniques.

Overall, while the direct impact of median income and poverty on cancer rates was inconclusive, the research highlighted critical areas for further investigation and the potential for more targeted, effective public health strategies.