2024

Cancer, Poverty & Race

AN ANALYSIS OF CANCER DEATHRATES IN THE US

AHMAD QADAFI

Table of Contents

Introduction

Cancer remains one of the most pressing public health challenges in the United States, with significant disparities in both incidence and mortality rates across different regions and demographic groups. The primary objective of this study is to investigate the relationships between various demographic factors—such as poverty, employment, and urbanicity—and cancer death and incidence rates across U.S. states and counties. By analyzing these variables, the study aims to reveal patterns that could provide insights into the geographic and demographic determinants of cancer outcomes. Understanding these relationships is crucial for designing effective public health interventions aimed at reducing cancer-related disparities.

Literature Review

The existing literature strongly suggests that socioeconomic factors like poverty and education are key determinants of cancer outcomes. Extreme poverty levels are linked to increased cancer mortality, often due to limited access to healthcare services, delayed diagnoses, and lower treatment adherence. For instance, counties with persistent poverty have significantly higher cancer mortality rates, underscoring the long-term impact of socioeconomic deprivation on health outcomes(Comprehensive Cancer Information,USAFacts).

Urbanicity also influences cancer outcomes, with rural areas experiencing higher cancer mortality rates compared to urban areas. This disparity is partly attributed to differences in healthcare accessibility, environmental exposures, and lifestyle factors(USAFacts,Cancer Health). Educational attainment, particularly at higher levels, is associated with better cancer outcomes, likely due to improved health literacy and greater access to healthcare resources (BioMed Central).

Racial and ethnic disparities in cancer outcomes are well-documented, with minority populations often bearing a disproportionate burden. For example, Black men and women have significantly higher mortality rates for prostate and breast cancers, respectively, compared to their White counterparts. These disparities are compounded by structural inequities and systemic injustices that limit access to healthcare and other resources(Cancer Health).

This study builds on these findings by exploring these factors at a more granular level, focusing on U.S. counties. The insights gained will help inform public health strategies aimed at mitigating these disparities and improving cancer outcomes nationwide.

Study Objectives

The primary objective of this study is to investigate the complex relationships between various demographic and socioeconomic factors—such as poverty levels, employment rates, and urbanicity—and cancer incidence and mortality rates across U.S. counties. By analyzing these relationships, the study aims to identify significant patterns that could shed light on the determinants of cancer outcomes, particularly in geographically and demographically diverse regions. The insights gained from this analysis are intended to inform public health strategies and interventions, with the goal of reducing cancer disparities and improving overall health outcomes across different population groups. Additionally, the study seeks to explore how these relationships vary across different regions and demographic groups, thereby contributing to a more nuanced understanding of the public health challenges associated with cancer.

Methodology

The methodological approach of this study is grounded in rigorous data collection and cleaning processes, ensuring that the analysis accurately reflects the relationships between

demographic factors and cancer outcomes. Data were obtained from the State Cancer Profiles website, a reliable source frequently used in public health research for accessing comprehensive cancer statistics across U.S. counties(Cancer Health). This website provides critical data on cancer incidence and mortality, which are essential for understanding regional disparities in cancer outcomes.

To ensure the datasets were suitable for analysis, we undertook extensive data cleaning and preparation. This step is crucial as it enhances the reliability and validity of the analysis by eliminating inconsistencies, missing data, and errors that could skew results. Such preprocessing techniques are well-documented in epidemiological research, where clean and well-structured datasets are foundational for accurate statistical modeling and interpretation(BioMed Central,Comprehensive Cancer Information). The use of county-level data, in particular, allows for a more granular analysis, offering insights into the geographic variability in cancer outcomes that might be obscured in broader, state-level data(USAFacts).

By adhering to these methodological best practices, the study ensures that its findings are robust, reproducible, and applicable to real-world public health interventions aimed at reducing cancer disparities across the United States.

Data Wrangling

Purpose: The purpose of the data wrangling process was to prepare the datasets for subsequent statistical analysis. This process involved several steps, including data cleaning, restructuring, and merging, to create a comprehensive dataset that accurately represents the variables of interest.

Data Acquisition

The datasets death.csv and incd.csv were downloaded from the State Cancer Profiles website. The death.csv file includes data on cancer death rates by county, while the incd.csv file contains information on cancer incidence rates by county.

Data Cleaning and Restructuring

County Column Modification: The County column in both datasets was cleaned by removing trailing numbers that were not pertinent to the analysis. Additionally, state names were separated from the County column and placed into a new column labeled States. A further column, Region, was created to categorize states into the East, South, Midwest, and West regions.

Column Renaming: For death.csv, the following columns were renamed:

- **Age-Adjusted Death Rate** - deaths per 100,000 was renamed to **Death Rate**.
- **Recent 5-Year Trend in Death Rates** was renamed to **5 Year Trend in Death Rates**.
- **2023 Rural-Urban Continuum Codes** was renamed to **Urbanicity**.

For incd.csv, the following columns were renamed:

- **Age-Adjusted Incidence Rate - cases per 100,000** was renamed to **Incidence Rate**.

- **Recent 5-Year Trend in Incidence Rates** was renamed to **5 Year Trend in Incidence Rate.**

Data Type Conversion: To ensure accuracy in statistical analysis, the Death Rate, 5 Year Trend in Death Rates, Incidence Rate, and 5 Year Trend in Incidence Rate columns were converted to numeric data types. This conversion was essential for facilitating subsequent analyses and visualizations.

Handling Missing Data: Rows with missing or invalid data, such as those marked as "data not available" or containing asterisks, were removed from the primary dataset and stored in separate data frames (Missing Counties_DR for death rates and Missing Counties_IN for incidence rates). This approach ensured that the primary datasets were free from inconsistencies that could skew the results.

Creation of New Data frames

Two new data frames were created to streamline the analysis:

Cancer Death Rates: This data frame includes the columns County, Death Rate, 5 Year Trend in Death Rates, and Urbanicity.

Cancer Incidence Rates: This data frame includes the columns County, Incidence Rate, and 5 Year Trend in Incidence Rate.

Subsequently, these two data frames were merged on the County and States columns to create a unified data frame named Cancer Rates. This final data frame consolidates all relevant variables for the analysis.

```
python
Copy code
# Example code snippet illustrating the data wrangling process.
import pandas as pd

# Loading datasets
death_df = pd.read_csv('death.csv')
incd_df = pd.read_csv('incd.csv')

# Cleaning and modifying County and State columns
death_df['County'] = death_df['County'].str.replace(r'\d+$', '')
death_df['States'] = death_df['County'].apply(lambda x: x.split(',')[1])
# Similar steps for incd.csv

# Renaming columns
death_df.rename(columns={'Age-Adjusted Death Rate([rate note]) -
deaths per 100,000': 'Death Rate',
                'Recent 5-Year Trend ‡ in Death Rates': '5 Year Trend in Death
Rates',
                '2023 Rural-Urban Continuum Codes([rural urban note])':
```

[Code Implementation: Data Wrangling]

Conclusion

The data wrangling process successfully prepared the datasets for detailed analysis. By ensuring that each county is represented by a single row and that all variables are correctly formatted, we have laid the groundwork for a robust analysis of the relationships between demographic factors and cancer outcomes across U.S. counties. This comprehensive dataset is now ready for further statistical exploration and interpretation, which will be crucial in guiding

public health strategies and interventions aimed at reducing cancer incidence and mortality. Should further refinement or additional analyses be required, they can be readily incorporated into this structured framework.

Data Wrangling: Part 2 - Adding Additional Variables to the Dataset

Purpose: The aim of this stage was to enrich the existing dataset, Cancer Rates Updated.csv, by incorporating additional demographic and socioeconomic variables from multiple supplementary data sources. These variables are expected to provide a more comprehensive understanding of the factors associated with cancer rates across U.S. counties.

Data Files and Preparation

The following files were uploaded and merged with the Cancer Rates Updated.csv dataset:

Geography	Health Statistics	Economics	Education	Race	Household Characteristics
County	Death Rate	Poverty Below 150%	Bachelor's degree	White Pct	Household Characteristics
States	5 Year Trend in Death Rates	Persistent Poverty	HS Education	Hispanic Pct.	
FIPS_Code	Incidence Rate	Families Below Poverty		Asian/PI Pct	
Urbanicity	5 Year Trend in Incidence Rate	Unemployment Rate		Black Pct.	
		Persons in Poverty		AI/AN Pct	
		SES		Racial Minority Index	

Each of these files contains data relevant to specific demographic or socioeconomic variables by county. The process for merging these files with the primary dataset is described below.

Data Cleaning and Merging Process

1. **County and State Separation**:

The initial step involved separating county and state names from a single column in each file. For instance, in files where the County column also contained state information, the column was split into two separate columns: County and States.

This was accomplished using a custom function, clean_county_column, which removed extraneous characters and ensured consistency in the formatting of county and state names across all files.

2. **Handling Multi-word Names**:

The function was specifically designed to handle multi-word county and state names (e.g., "South Dakota," "Adams North") to prevent errors during the merger process. This ensured that names were accurately split without losing essential information.

3. **Merging Datasets**:

Each of the supplementary files was merged with the Cancer Rates Updated.csv file using the County and States columns as keys. This approach ensured that the merger was based on both geographic identifiers, minimizing the risk of mismatched or missing data.

The merger was conducted using a left join, meaning that all counties present in the Cancer Rates Updated.csv dataset were retained, even if no corresponding data was found in the supplementary files.

4. **Managing Duplicate Columns**:

During the merger, potential conflicts due to duplicate columns (such as multiple FIPS columns) were resolved by renaming columns or dropping unnecessary duplicates. This approach preserved the integrity of the data while preventing confusion from overlapping column names.

5. **Addressing Unmerged Entries**:

After the initial merger, it was found that some counties had not successfully merged due to inconsistencies in naming or formatting. These issues were systematically addressed by refining the county name cleaning process and ensuring that state names were consistently represented.

6. **Inclusion of FIPS Codes**:

A particular focus was placed on the inclusion of Federal Information Processing Standards (FIPS) codes. These codes are crucial for uniquely identifying counties and were included in the final merged dataset to facilitate further geographic analysis.

The FIPS.csv file was merged successfully, and the resulting dataset included a FIPS_Code column, ensuring that each county could be uniquely identified.

Results

The final merged dataset, now titled Merged_Cancer_Rates_with_FIPS_Final.csv, contains 2,852 rows and twenty-four columns. The columns include

Geography	Health Statistics	Economics	Education	Race	Household Characteristics
County	Death Rate	Poverty Below 150%	Bachelor's degree	White Pct	Household Characteristics
States	5 Year Trend in Death Rates	Persistent Poverty	HS Education	Hispanic Pct.	
FIPS_Code	Incidence Rate	Families Below Poverty		Asian/PI Pct	
Urbanicity	5 Year Trend in Incidence Rate	Unemployment Rate		Black Pct.	
		Persons in Poverty		AI/AN Pct	
		SES		Racial Minority Index	

This dataset represents a comprehensive compilation of cancer rates alongside a broad range of demographic and socioeconomic indicators, making it a valuable resource for in-depth analysis.

Conclusion

The merging process was successfully completed, resulting in a dataset that integrates cancer statistics with various demographic and socioeconomic factors. The careful handling of county and state names, as well as the inclusion of FIPS codes, ensures that the dataset is both comprehensive and ready for advanced geographic and statistical analysis.

This enriched dataset can now be used to explore correlations, identify patterns, and perform regression analyses to understand better the determinants of cancer incidence and mortality across the United States. Further analysis could involve visualizing the data, conducting multivariate regression, or applying machine learning models to predict cancer outcomes based on the integrated variables.

Exploratory Data Analysis: Part 1

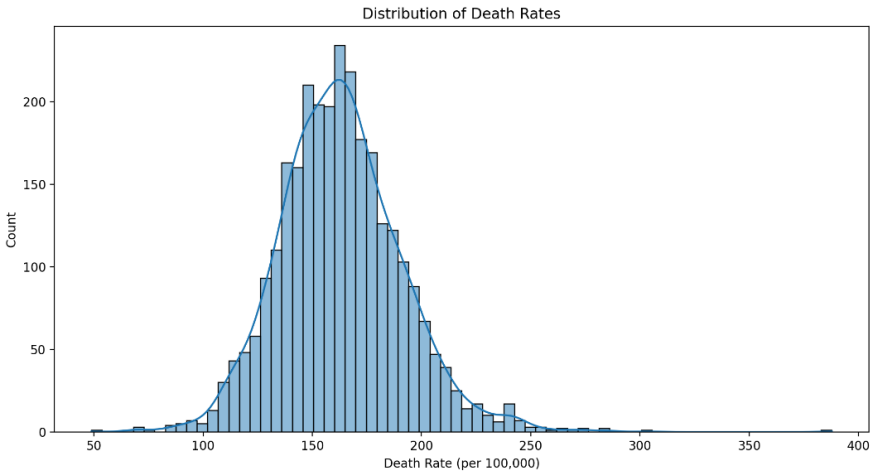
Introduction

In this section, we begin the exploratory data analysis (EDA) of the cancer rates dataset, focusing on key statistical insights and visualizations that highlight the relationships between a range of factors and cancer outcomes across U.S. counties.

Distribution of Death Rates Across Counties

The histogram illustrates the distribution of cancer death rates across all counties. The distribution is normal with a slight right skew, indicating that while most counties have death rates clustered around the mean, there are some counties with notably higher rates.

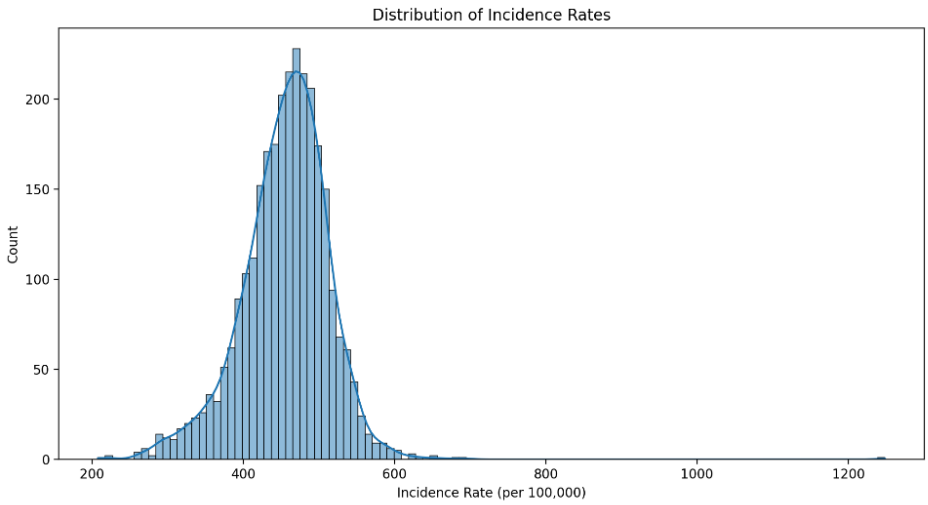
Figure 1: Distribution of Death Rates



Distribution of Incidence Rates Across Counties

Like the death rates, the incidence rates across counties exhibit a normal distribution with a right skew. This suggests that while most counties have incidence rates around the mean, there are a few counties with significantly higher incidence rates.

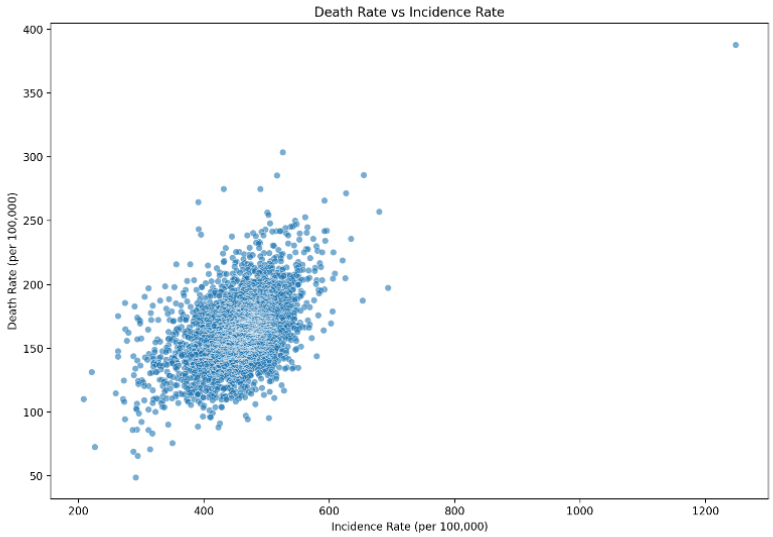
Figure 2: Distribution of Incidence Rates



Relationship Between Death Rates and Incidence Rates

I created a scatter plot to explore the relationship between cancer death rates and incidence rates across counties. The plot reveals a positive correlation, suggesting that as incidence rates increase, death rates tend to increase as well. However, there is considerable variation around this trend.

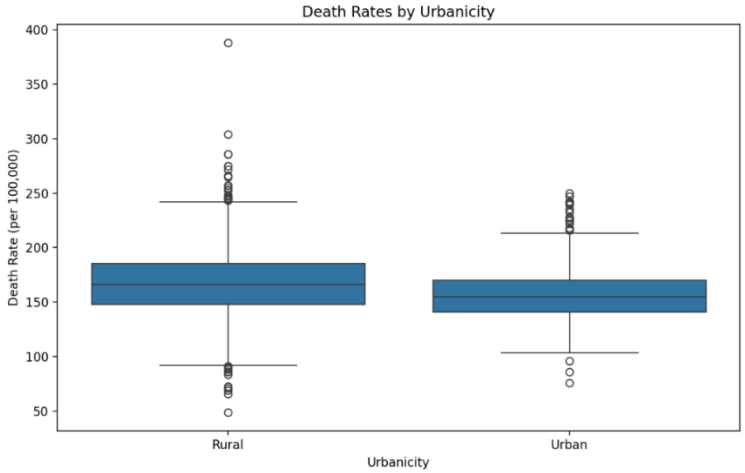
Figure 3: Death Rates vs. Incidence Rates



Comparison of Death Rates Between Rural and Urban Counties

I constructed a box plot to compare the distribution of death rates between rural and urban counties. The analysis shows that rural counties tend to have slightly higher median death rates and exhibit more variability in death rates compared to urban counties.

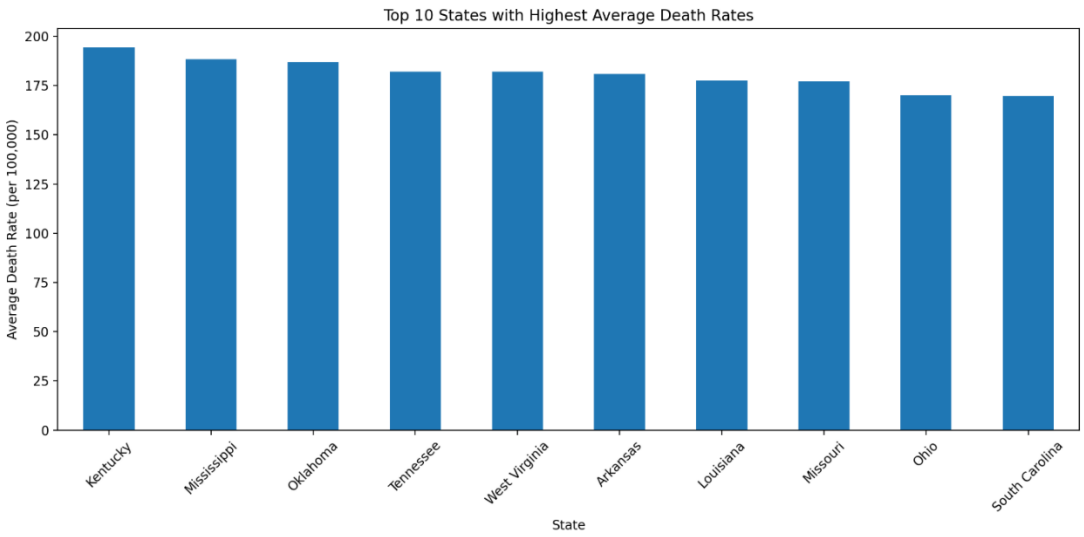
Figure 4: Death Rates by Urbanicity



States with the Highest Average Death Rates

A bar chart identifies the states with the highest average cancer death rates. The data indicates that Kentucky, Mississippi, and West Virginia have the highest average death rates among all states.

Figure 5: States with the Highest Average Death Rates



Summary Statistics

The summary statistics provide an overview of the central tendencies and variability in key cancer-related variables:

1. The average death rate across all counties is approximately 163 per 100,000 population, with a standard deviation of 28.1.
2. The average incidence rate is about 456 per 100,000 population, with a standard deviation of 58.1.

3. Both the 5-year trends for death rates and incidence rates show slight negative averages, indicating a small overall decrease in these rates over.

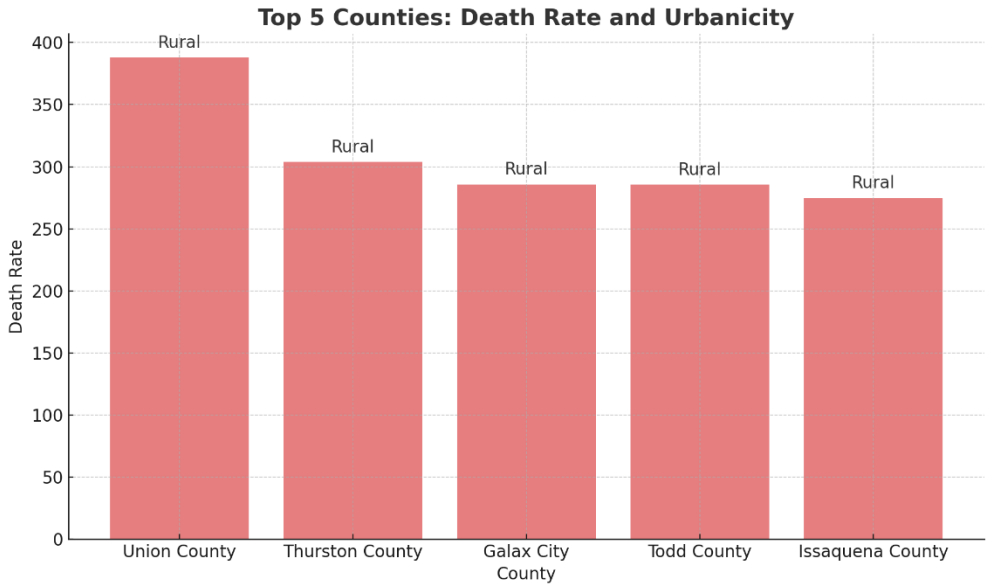
Figure 6: Summary Statistics

Calc	Death Rate	5 Year Trend in Death Rates	Incidence Rate	5 Year Trend in Incidence Rate
count	2852	2852	2852	2852
mean	163.06	-1.04	455.65	-0.28
std	28.13	1.86	58.1	2.22
min	48.9	-16.5	208	-31
25%	144.9	-1.4	424.08	-0.8
50%	161.6	-1	460.4	-0.2
75%	179.5	-0.6	492.13	0.4
max	387.8	24.5	1248.4	31.9

Key Observations

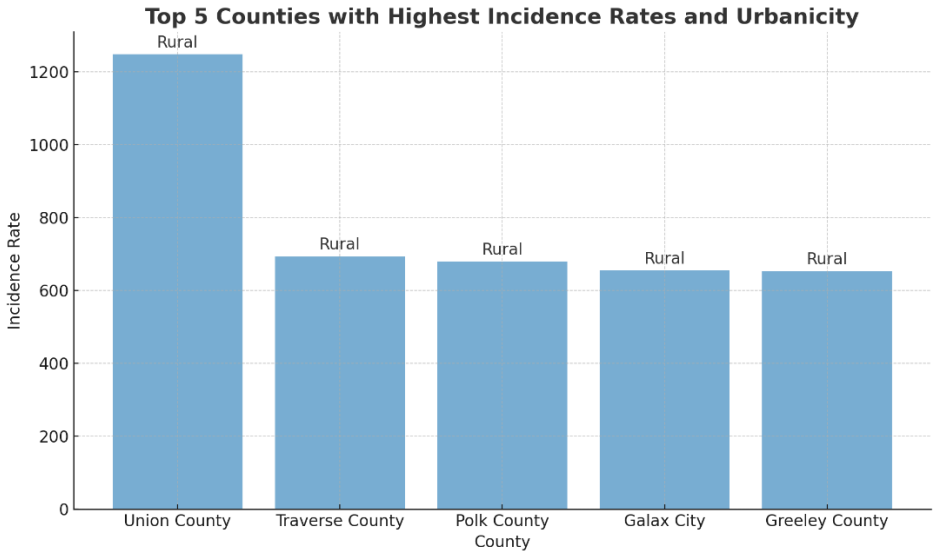
The bar hart highlights the top five counties with the highest cancer death rates with Union County, Florida leading with a death rate of 387.8 per 100,000 population.

Figure 7: Top 5 Counties with Highest Death Rates



Similarly, the chart below shows the top five counties with the highest cancer incidence rates with Union County, Florida also having the highest incidence rate of 1248.4 per 100,000 population.

Figure 8: Top 5 Counties with Highest Incidence Rates



Conclusion

These initial insights from the exploratory data analysis provide a solid foundation for understanding the landscape of cancer rates across the United States. The distributions, correlations, and key statistics highlighted in this section will guide further analyses and help identify potential areas for intervention and public health focus. The visualizations, when included, will further enhance the understanding of these relationships and trends.

Exploratory Data Analysis: Part II

The purpose of this section is to delve deeper into the relationships between various demographic factors and cancer rates across U.S. counties. The analysis will focus on examining correlations, generating visualizations, and providing summary statistics to uncover potential patterns and insights.

Dataset

The dataset comprises 2,864 entries, each representing a county or equivalent administrative division across the United States. The dataset includes twenty-four columns, with variables of interest categorized into demographic, educational, employment, household, and urbanicity indicators, as well as cancer rates.

Variables of Interest

Demographic Variables:

Racial Demographics	Educational Variables	Employment Variable	Household Variables	Urbanicity Variable	Cancer Rates
White Pct	Bachelor's Degree	Unemployment Rate	Persons in Poverty	Urbanicity	Death Rate
Black Pct	HS Education		Families Below Poverty		Incidence Rate
Hispanic Pct					
Asian/PI Pct					
AI/AN Pct					

Urbanicity

The 'Urbanicity' variable is a categorical variable with two unique values: 'Rural' and 'Urban.' This variable categorizes counties based on their rural or urban status, which is

important for understanding geographic disparities in cancer outcomes. The categorical nature of this variable required special handling in the analysis, particularly when examining correlations with numeric variables.

Data Cleaning and Preparation

Before proceeding with the exploratory analysis, the dataset underwent a cleaning process to address issues related to non-numeric values in columns that were expected to contain numeric data. Specifically:

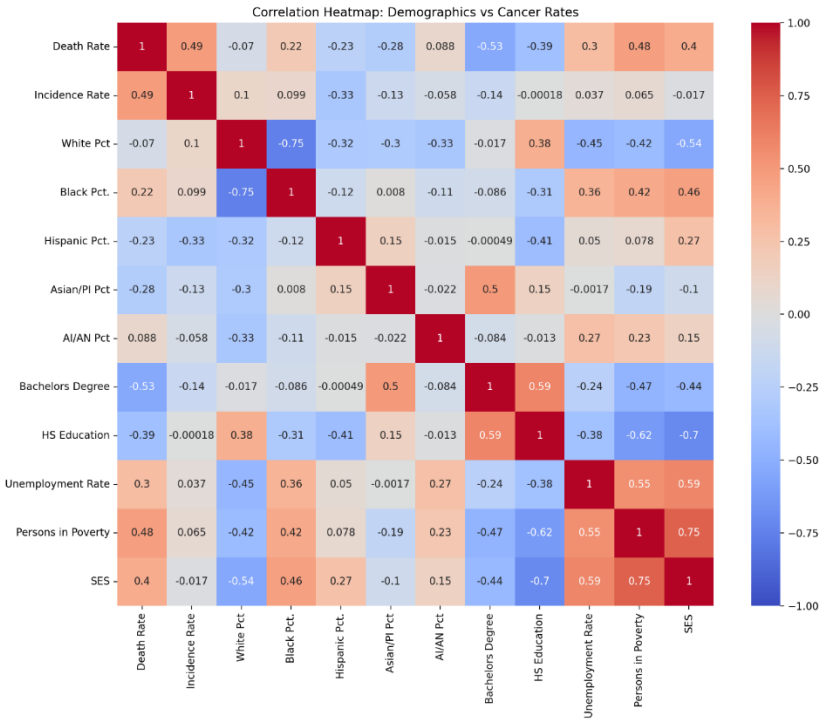
- The 'SES' (Socioeconomic Status) column contained non-numeric values labeled as "data not available." These values were replaced with NaN (Not a Number) to ensure consistency in the dataset.
- All relevant columns were then converted to numeric data types, using coercion to handle any remaining non-numeric values. This process ensured that the dataset was ready for further statistical analysis.

Data Analysis

Correlation Analysis and Heatmap

A heatmap was generated to visualize the correlations between various demographic factors and cancer rates. The darker the color, the stronger the correlation, with red indicating positive correlations and blue indicating negative correlations.

Figure 1: Correlation Between Demographic Factors and Cancer Rates



Top Correlations with Death Rate

The variables with the strongest positive correlations with death rate were *Persons in Poverty* (0.48) and *SES* (0.40). The variable with the strongest negative correlation was *Bachelor's degree* (-0.53).

Top Correlations with Incidence Rate

The variables with the strongest correlations with incidence rate included *Death Rate* (0.49) and a weak negative correlation with *Hispanic Percentage* (-0.33).

Overall, the correlations between incidence rates and demographic factors were weaker than those with death rates.

Scatter Plots with Trend Lines

Education vs. Death Rate

A scatter plot was created to explore the relationship between education levels (specifically the percentage of the population with a bachelor's degree) and cancer death rates. The plot showed a clear negative correlation, indicating that higher education levels are associated with lower cancer death rates.

Figure 2: Death Rate vs Bachelor's degree

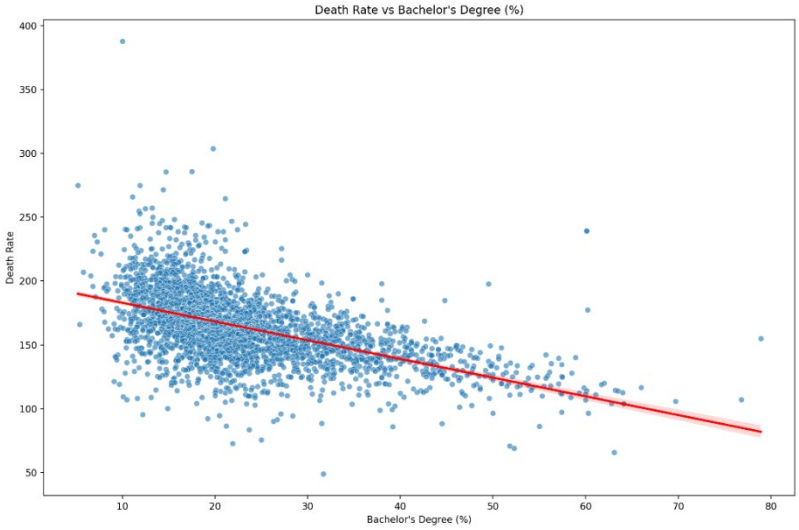
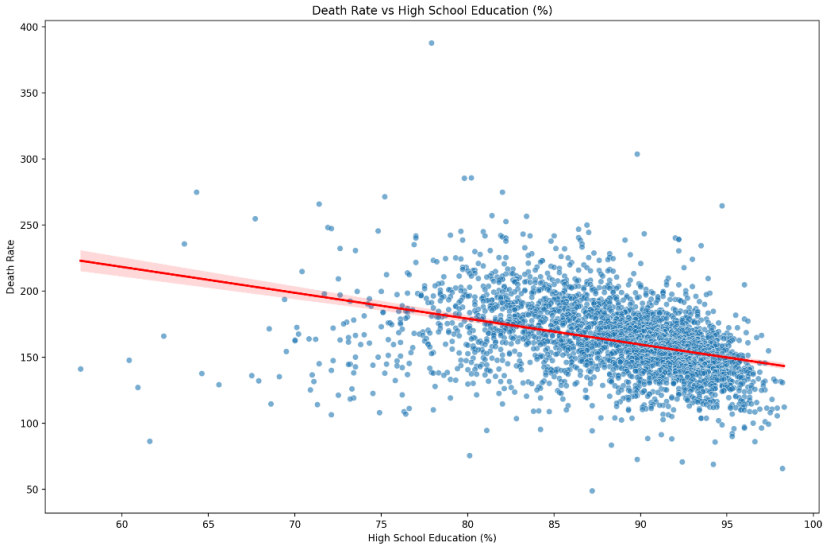


Figure 2: Death Rate vs High School Education



Race/Ethnicity vs. Death Rate

Scatter plots were created to examine the relationships between race/ethnicity percentages and cancer death rates:

The plot showed a negative correlation between Whites and Hispanics and Death Rates. However, there was a clear positive correlation between Black population and Death Rates. These plots highlight clear racial and ethnic disparities in cancer outcomes.

Figure 3: Death Rate vs White Population

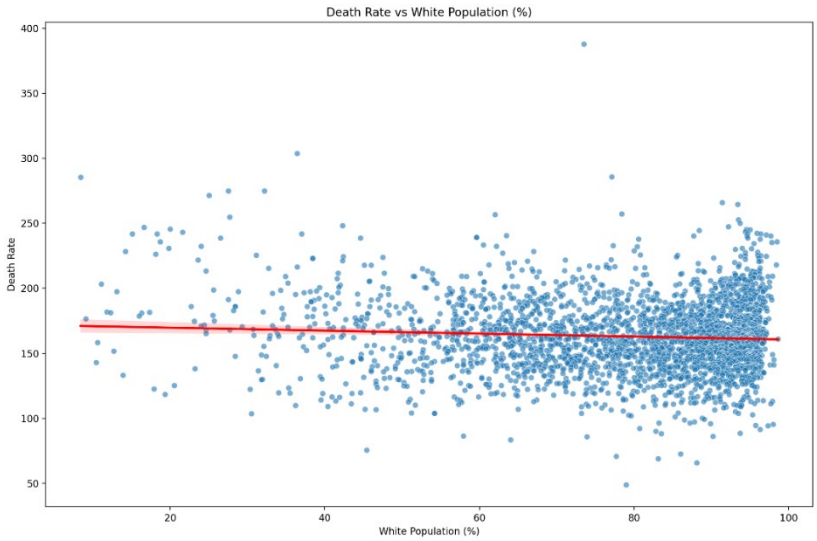


Figure 4: Death Rate vs Black Population

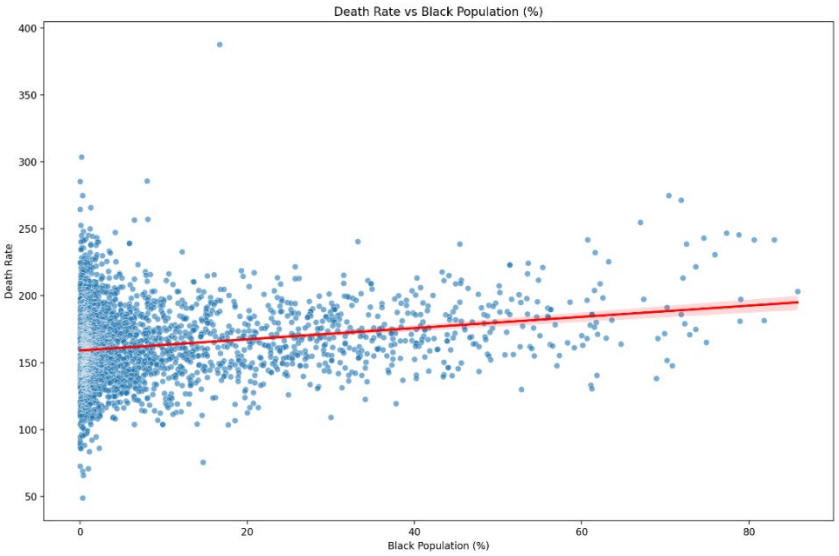
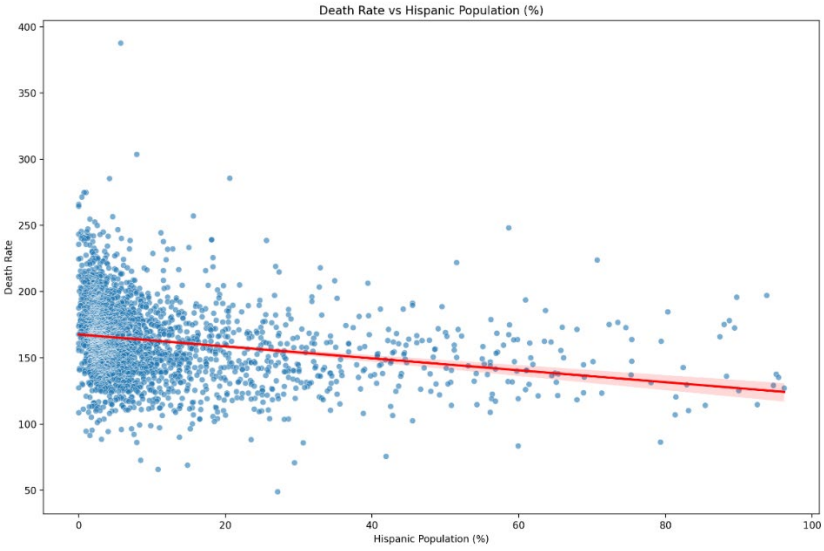


Figure 5: Death Rate vs Hispanic Population



Poverty vs. Death Rate

I created Scatter plots to assess the relationship between poverty measures and cancer death rates. There is strong correlation between Poverty, SES and Death Rates. Both individuals

and families in poverty correlate with higher death rates. SES (Measured as a social vulnerability, so higher number means more vulnerable) also has a positive correlation.

Figure 6: Death Rate vs Families Below Poverty

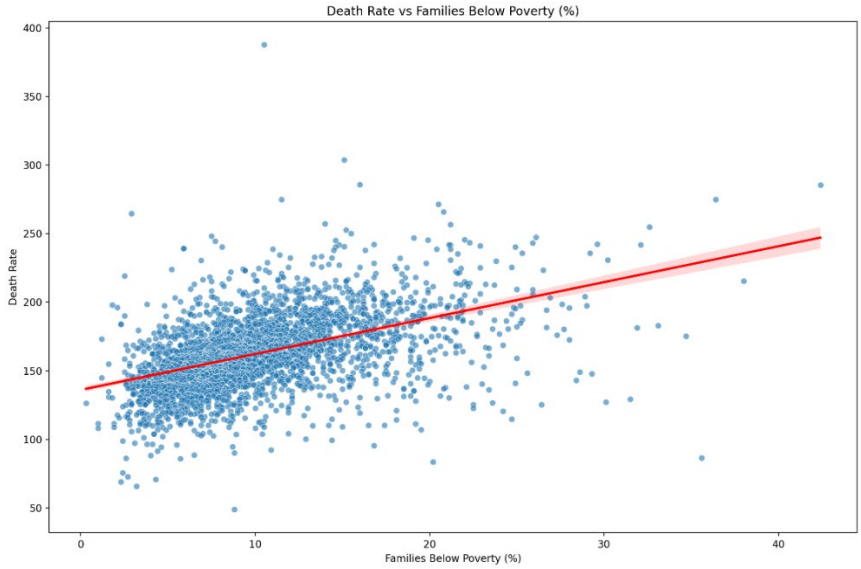


Figure 7: Death Rates vs. Below 150% Poverty Line

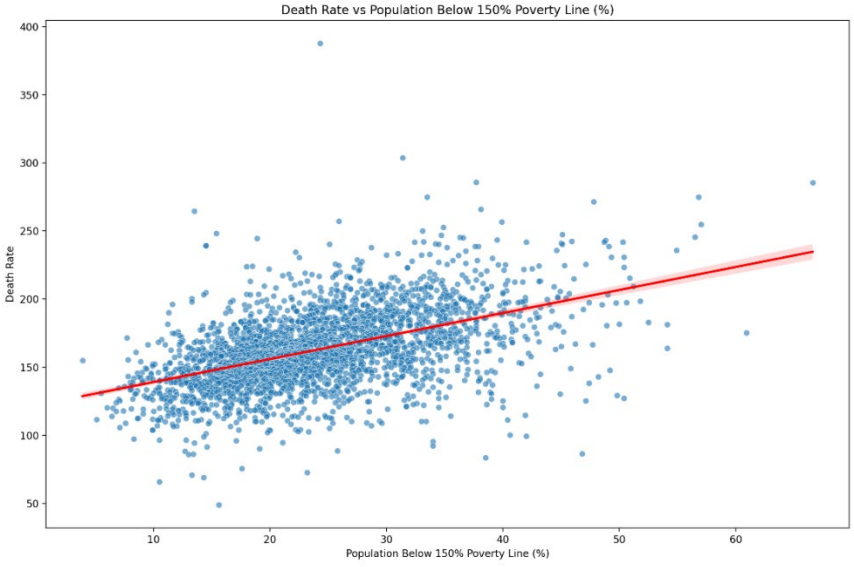
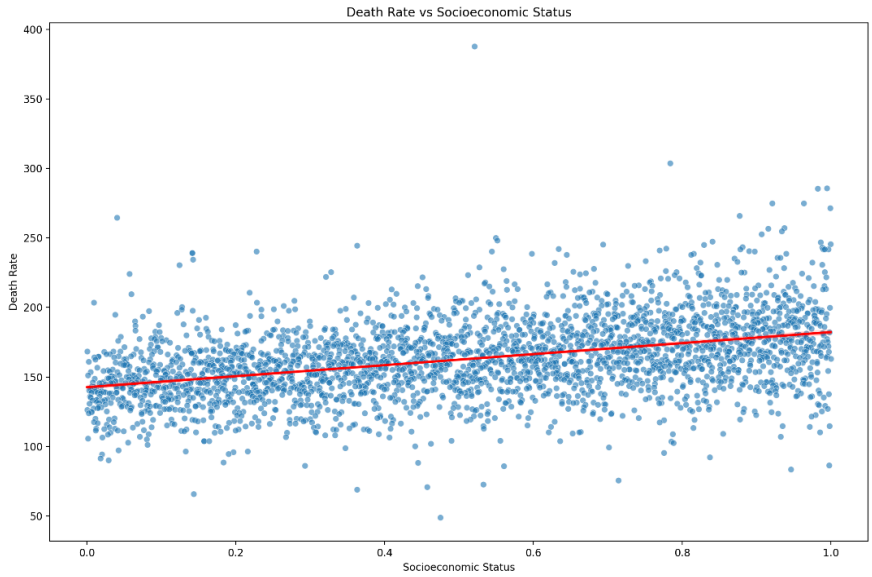


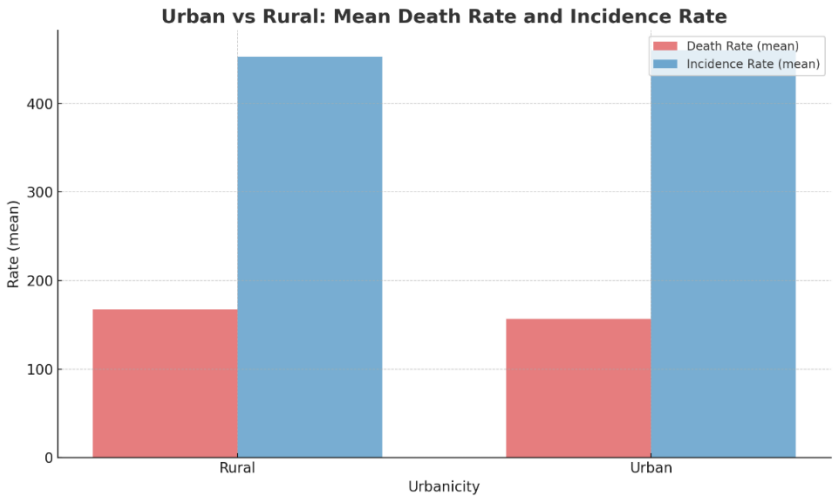
Figure 8: Death Rates vs. Socioeconomic Status



Cancer Rates by Urbanicity

A summary table was created to compare cancer rates between rural and urban counties. The analysis revealed that rural areas have a higher mean death rate (167.35 per 100,000) compared to urban areas (156.52 per 100,000). However, urban areas have a slightly higher mean incidence rate (459.89 per 100,000) compared to rural areas (452.88 per 100,000).

Figure 9: Cancer Rates by Urbanicity



Statistical Testing: Urban vs. Rural Death Rates

T-test: Death Rate by Urbanicity

A t-test was conducted to assess the statistical significance of the difference in cancer death rates between urban and rural areas. The results (t-statistic = -10.2303, p-value = 0.0000) indicate a statistically significant difference in death rates, with rural areas having higher rates than urban areas.

Key Insights

1. **Education and Poverty**:

Education levels, particularly higher education, show a strong inverse relationship with cancer death rates. In contrast, poverty measures have a strong positive correlation with cancer death rates, highlighting the impact of economic hardship on health outcomes.

2. **Racial and Ethnic Disparities**:

The analysis revealed significant racial and ethnic disparities in cancer outcomes, with areas having higher Black populations showing higher death rates, while areas with higher Hispanic populations tended to have lower death rates.

3. **Urban-Rural Divide**:

There is a notable difference in cancer death rates between urban and rural areas, with rural areas exhibiting higher death rates. However, the difference in incidence rates between these areas is less pronounced.

4. **Unexpected Correlation**:

The positive correlation between SES and death rates is counterintuitive and suggests that there may be confounding factors or data issues that need to be explored further.

Conclusion

The findings from this exploratory data analysis underscore the complex relationships between socioeconomic factors, race/ethnicity, geography, and cancer outcomes. These insights suggest that addressing cancer disparities requires a multifaceted approach that considers education, poverty reduction, and targeted interventions for specific populations and geographic areas. Further investigation is warranted, particularly regarding the unexpected positive correlation between SES and cancer death rates, to better understand the underlying factors influencing these outcomes.

Socioeconomic and Demographic Correlates of Cancer Outcomes Across U.S. States: An Exploratory Data Analysis Pt. III

Introduction

This report presents an exploratory analysis of the relationships between various socioeconomic and demographic factors and cancer outcomes across different U.S. states. Specifically, we examine how factors such as poverty, urbanicity, and racial demographics correlate with cancer death and incidence rates. The analysis aims to identify key patterns and relationships that could inform public health policy and interventions.

Data Preparation

The dataset used in this analysis initially contained 2,864 rows and twenty-five columns, covering a broad range of socioeconomic and demographic variables. After cleaning and preparing the data, we retained all rows and columns for analysis, ensuring that the dataset was free of missing or non-numeric values.

Results

Overall Correlation Analysis

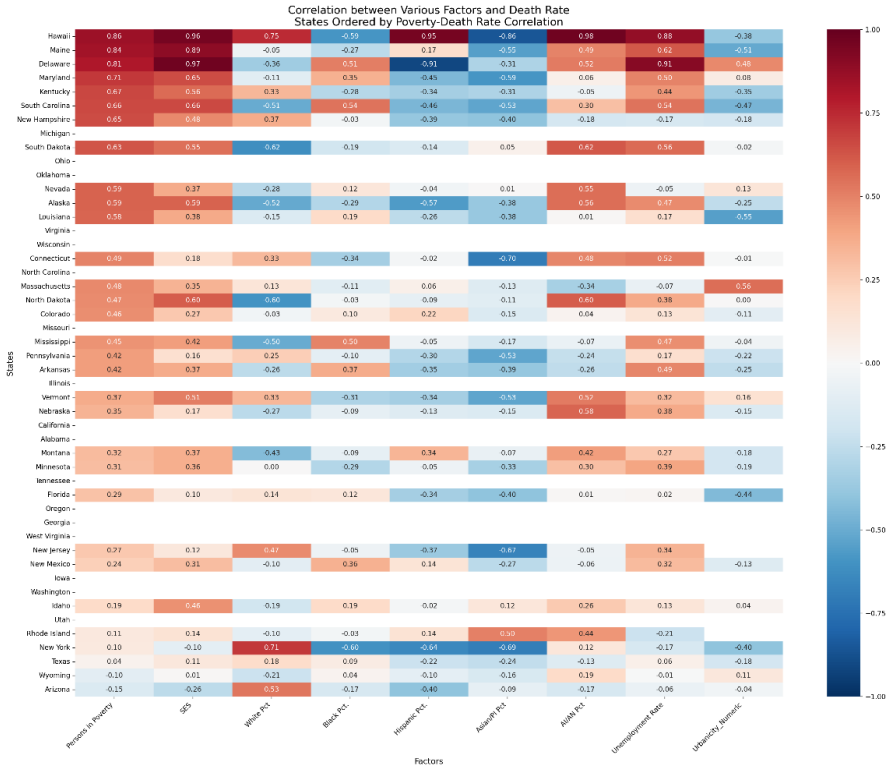
There is a moderate positive correlation (0.413) between poverty and cancer death rates across all states. The correlation between poverty and cancer incidence rates is notably weaker (0.044). Urbanicity shows a slight negative correlation (-0.166) with cancer death rates, but a

slight positive correlation (0.128) with cancer incidence rates. Racial demographics display varying correlations, with the most significant being a negative correlation (-0.316) between the percentage of Asian/Pacific Islander populations and cancer death rates.

State-Level Correlations

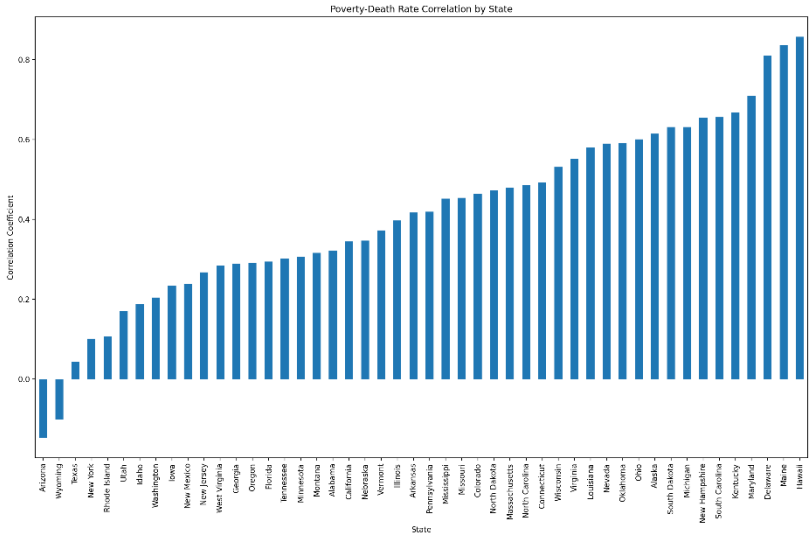
The state-level analysis further uncovers the variability in how these correlations manifest across different regions. The correlations between a range of factors and cancer death rates are visualized in the heatmap (Figure 1). States are ordered by the strength of the correlation between poverty and cancer death rates.

Figure 1. Correlation Between Various Factors and Death Rate by State



There is a moderate positive relationship (0.310) between the poverty-death rate correlation and the percentage of the Native population's death rate. A negative relationship (-0.324) is observed between the poverty-death rate correlation and the Asian population's death rate.

Figure 2. Poverty-Death Rate Correlation by State



This bar chart ranks states by their poverty-death rate correlation. Hawaii, Maine, and Delaware emerge as the states with the highest correlation, while Arizona and Wyoming show the lowest correlations. The significant variability across states underscores the complex interplay of socioeconomic and demographic factors influencing cancer mortality.

Correlations for Top and Bottom 10 States

To gain deeper insights in state level factors related to poverty and death rates, the study further examined the correlations for the top ten and bottom ten states based on their poverty-death rate correlations.

Figure 3. Correlations for Bottom 10 States

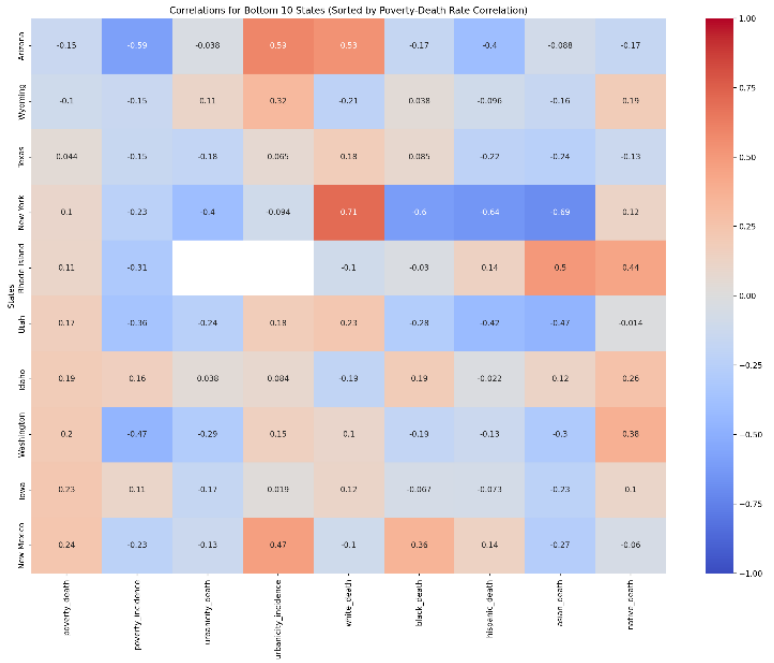
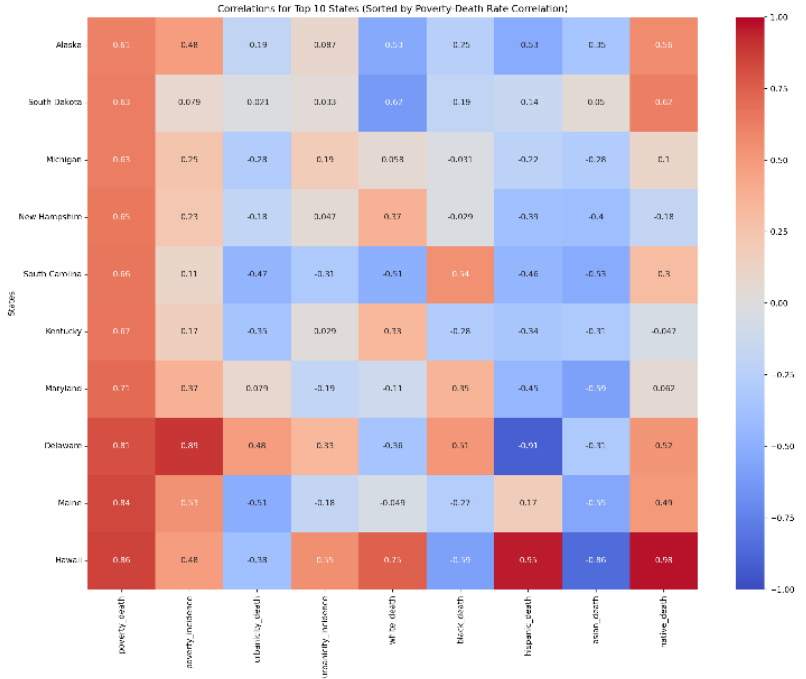


Figure 4. Correlations for Top 10 States



These figures provide a detailed view of how correlations between distinct factors and cancer death rates vary across states with the highest and lowest poverty-death rate correlations. For example, in the top ten states, strong correlations are observed between poverty and cancer death rates, while urbanicity shows mixed effects across different racial demographics.

Discussion

The state-level differences in the correlation between poverty and cancer death rates suggest that poverty alone may not fully account for variations in cancer mortality across regions. The variability in correlations points to the influence of other factors that could be crucial in understanding cancer outcomes. These additional factors include:

1. **Healthcare Access and Quality**: States with stronger poverty-death rate correlations, like Hawaii, Maine, and Delaware, may also face issues related to healthcare access or quality for low-income populations. For example, limited access to early detection and treatment facilities, or disparities in healthcare infrastructure between urban and rural areas, could amplify the effects of poverty on cancer mortality. Conversely, states with weaker correlations, such as Arizona and Wyoming, may have better access to care for low-income individuals, mitigating the impact of poverty on outcomes.

2. **Insurance Coverage and Medicaid Expansion**: The differences in state-level correlations may reflect the impact of Medicaid expansion and the availability of affordable health insurance options. States that expanded Medicaid under the Affordable Care Act (ACA) may show weaker poverty-death rate correlations, as low-income individuals gain better access to preventive care and cancer treatment services. On the

other hand, states that did not expand Medicaid might show stronger correlations due to gaps in coverage for impoverished populations.

3. **Demographic Influences**: The moderate positive correlation (0.310) between poverty-death rate and the Native population's death rate and the negative correlation (-0.324) with the Asian population's death rate indicate that racial and ethnic differences may play a significant role in cancer mortality patterns. These disparities could be cultural, social, and genetic factors, as well as differences in healthcare access and lifestyle. For instance, Native populations may face higher cancer mortality due to higher rates of poverty, limited access to healthcare, and historical inequities, while Asian populations may experience better outcomes due to healthier lifestyles or cultural emphasis on preventive care.

4. **Environmental and Occupational Factors**: Certain states may have environmental or occupational exposures that disproportionately affect low-income populations. For example, higher exposure to carcinogens in agricultural, industrial, or mining industries prevalent in certain regions could contribute to elevated cancer death rates among poorer populations. Exploring the geographical distribution of environmental risks, such as pollution levels, industrial activity, or pesticide use, could uncover important insights.

5. **Urban vs. Rural Differences**: The mixed relationships between urbanicity and cancer mortality across states highlight the need to consider how rural and urban environments differently shape cancer outcomes. Rural areas often face unique challenges, such as fewer healthcare facilities, longer travel distances for treatment, and shortages of specialized care providers. On the other hand, urban areas might have better

healthcare infrastructure but face issues such as overcrowding, environmental pollution, or barriers to care due to socioeconomic inequalities.

6. **Health Behaviors and Public Health Campaigns**: Variations in public health education, lifestyle factors (e.g., smoking rates, diet, physical activity), and the effectiveness of cancer prevention campaigns can also influence the strength of the poverty-death rate correlation. States with effective public health initiatives that target at-risk, low-income populations may see lower correlations, as these programs help mitigate the negative effects of poverty on cancer outcomes.

Further Research Directions

To gain deeper insights into the state-level differences in cancer mortality, further research should explore:

Medicaid expansion impact on cancer outcomes: A comparative analysis between states with and without Medicaid expansion could provide evidence of how insurance access influences cancer mortality.

Environmental justice: Investigating the link between environmental hazards and cancer outcomes, particularly in low-income and minority communities, may reveal significant contributing factors to state-level variations.

Cultural and social determinants of health: Examining how cultural attitudes toward healthcare, trust in medical institutions, and social support networks differ across states and their impact on cancer outcomes could shed light on disparities.

In conclusion, while poverty is a significant factor in cancer mortality, the variability in correlations across states underscores the importance of a comprehensive approach. Addressing

cancer disparities requires considering a broad array of factors, including healthcare access, insurance coverage, demographics, environmental risks, and public health interventions. Tailoring policies to the unique needs of each state and population can lead to more effective strategies for reducing cancer mortality.

The Role of Urbanicity and Race in the Relationship Between Poverty and Cancer Death Rates

Introduction

The purpose of this report is to explore the relationship between poverty and cancer death rates across different states, with a specific focus on the potential role of urbanicity as a mediating or confounding variable. The analysis seeks to determine whether urbanicity (urban or rural) influences the strength of the correlation between poverty and cancer death rates. Additionally, the report explores the possible contributions of other variables, such as race and employment, to this relationship.

Data Preparation

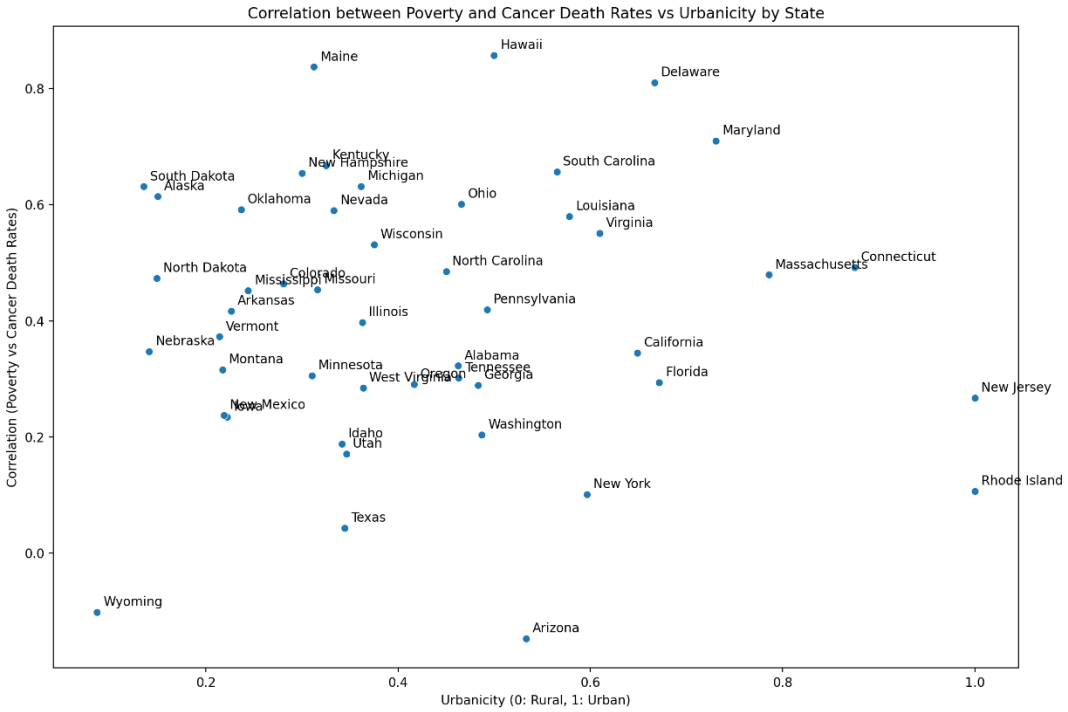
The dataset provided includes a column labeled "Urbanicity," which is a categorical variable indicating whether a state is more urban or rural. For the purposes of this analysis, the Urbanicity column was converted to a numeric scale where values closer to zero indicate a more rural environment, and values closer to one indicate a more urban environment.

Analysis

Correlation Between Poverty and Cancer Death Rates by Urbanicity

I created a scatter plot of the relationship between urbanicity and the correlation between poverty and cancer death rates. This plot compares the strength of the correlation between poverty and cancer death rates with the level of urbanicity for each state.

Figure 1: *Scatter Plot of Correlation Between Poverty and Cancer Death Rates vs. Urbanicity*



Key Observations

1. **Lack of Clear Relationship**: The scatter plot reveals that there is no clear linear relationship between urbanicity and the correlation between poverty and cancer death rates. The data points are widely scattered, indicating that urbanicity alone does not strongly determine the relationship between poverty and cancer death rates across states.

2. **Positive Correlation in Most States**: Most states exhibit a positive correlation between poverty and cancer death rates, suggesting that as poverty increases, so do cancer death rates. However, the strength of this correlation varies significantly between states.

3. **Outliers**: States like Hawaii and Maine have some of the highest correlations between poverty and cancer death rates, despite differing levels of urbanicity. Conversely, Wyoming and Arizona show negative correlations, which are atypical compared to other states.

4. **Cluster of States**: There is a noticeable cluster of states with urbanicity values between 0.2 and 0.4 that display a wide range of correlation strengths, indicating that other factors besides urbanicity may be influencing the relationship between poverty and cancer death rates.

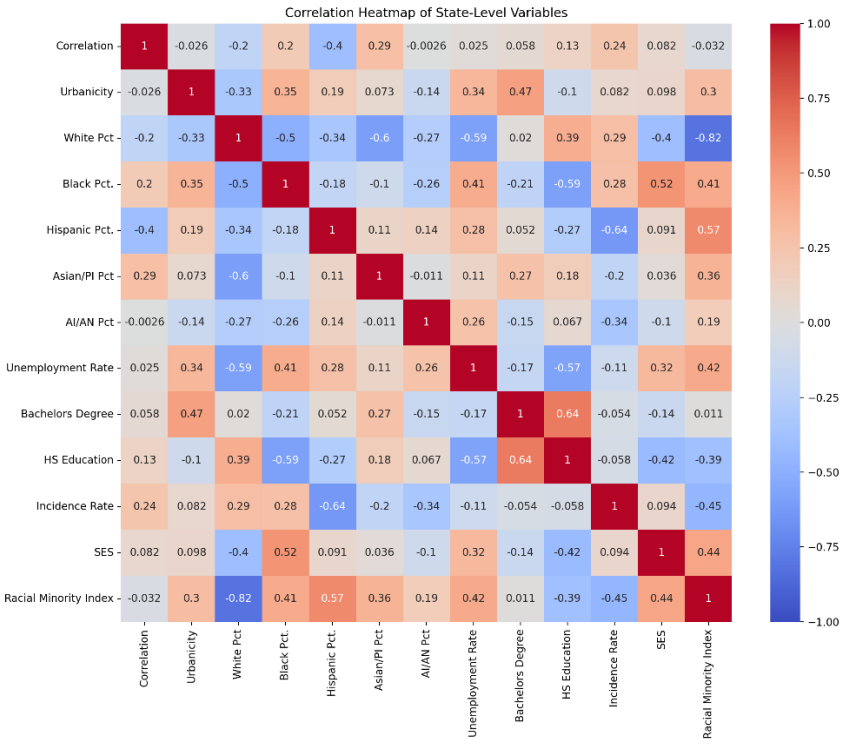
Mediating or Confounding Variables

To better understand why some states, exhibit stronger correlations between poverty and cancer death rates, we examined other variables, such as racial demographics and employment, to see if they might be mediating or confounding this relationship.

Partial Correlation Analysis

I conducted a partial correlation analysis to determine the influence of various racial demographics on the relationship between poverty and cancer death rates, while controlling for the percentage of the Hispanic population.

Figure 2: *Partial Correlation Heatmap*



Key Findings

Race and Ethnicity:

Asian/Pacific Islander (PI) Population: The analysis revealed that states with higher percentages of Asian/PI populations tend to have stronger correlations between poverty and cancer death rates, even after controlling for Hispanic populations.

Hispanic Population: States with higher percentages of Hispanic populations tend to have weaker correlations between poverty and cancer death rates. This suggests that in these states, high cancer death rates might be less strongly associated with poverty compared to states with higher Asian/PI populations.

Other Demographics:

White Population: After controlling for Hispanic population percentages, states with higher White populations showed a negative correlation with poverty-death rate correlations, suggesting that poverty is less likely to be a primary driver of cancer death rates in these states.

Racial Minority Index: When Hispanic population was controlled for, States with greater racial diversity (as measured by the Racial Minority Index) exhibited stronger correlations between poverty and cancer death rates.

Urbanicity

Urbanicity itself showed only a weak negative correlation with the poverty-death rate correlation, indicating that it may not be a strong mediator or confounder in this relationship.

Conclusions

The analysis suggests that while urbanicity does not strongly influence the relationship between poverty and cancer death rates, racial and ethnic demographics play a more significant role. States with higher Hispanic populations tend to show weaker correlations between poverty and cancer death rates, whereas states with higher Asian/PI populations tend to show stronger correlations. These findings point to the complexity of the factors influencing cancer outcomes and highlight the importance of considering multiple variables when analyzing health disparities.

Recommendations for Further Research

1. **State-Specific Analysis**: Future research could benefit from examining state-specific healthcare policies and access to better understand the variations in poverty-cancer death rate correlations.
2. **Non-Linear Relationships**: Further exploration of non-linear relationships between these variables might uncover more nuanced insights.
3. **Incorporating Environmental Factors**: Adding environmental variables to the analysis could provide additional context for understanding the disparities in cancer death rates.
4. **Policy Implications**: These findings suggest the need for tailored public health interventions that consider the specific demographic and socioeconomic context of each state.

This report provides a foundational understanding of the complex relationships between poverty, race, urbanicity, and cancer death rates, highlighting areas for further investigation and potential policy action.

The Interplay of Poverty, Education, and Race on Cancer Mortality

In this section, we explore how the combined effects of poverty, education, and race impact cancer death rates across U.S. counties. We employ visualizations to illustrate these relationships and analyze how race might mediate the influence of socioeconomic factors on cancer mortality.

Methods

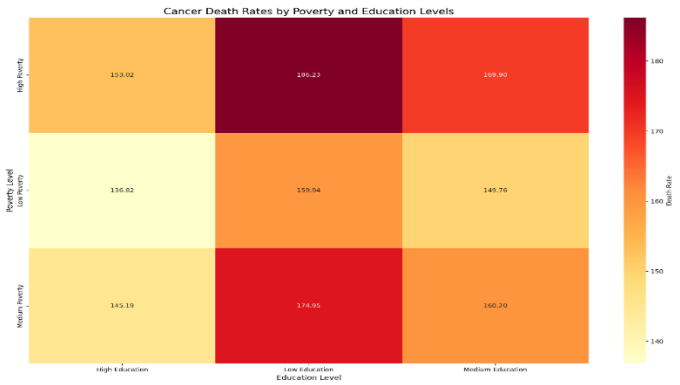
To fully capture the complex interplay between these factors, we created several heatmaps that provide a clear visual representation of the data. These heatmaps allow us to see how cancer death rates, racial composition, and socioeconomic factors intersect.

Results

Cancer Death Rates by Poverty and Education Levels

The following heatmap shows the variation in cancer death rates across different combinations of poverty and education levels. Darker colors represent higher death rates.

Figure 1: Cancer Death Rates by Poverty and Education Levels



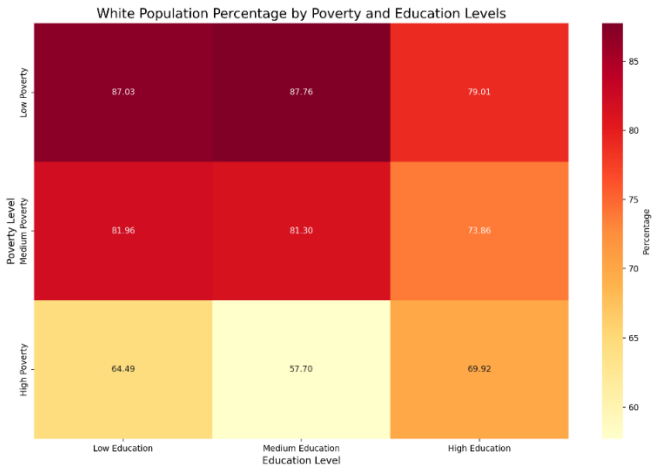
Key Observations

The highest death rate (186.48) occurs in areas with extreme poverty and low education. The lowest death rate (137.62) was found in areas with low poverty and high education. There is a clear trend of decreasing death rates as poverty decreases and education increases, highlighting the mitigating effect of education on cancer mortality.

Racial Composition by Poverty and Education Levels

To understand how race may influence the relationship between poverty, education, and cancer death rates, we examined the racial composition across different socioeconomic contexts.

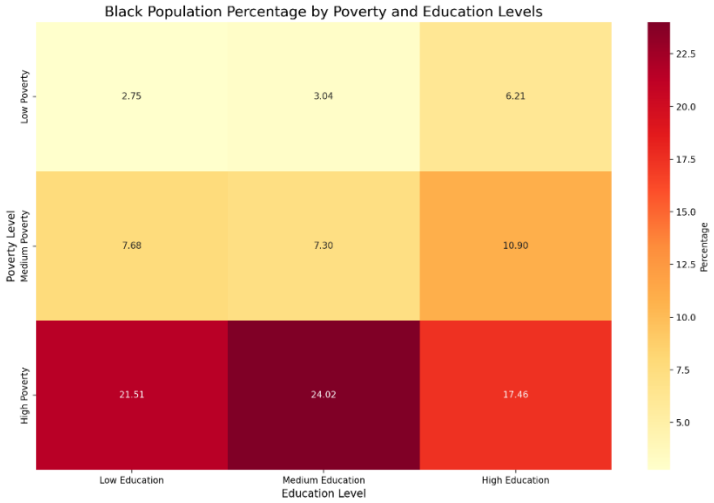
Figure 2: White Population Percentage by Poverty and Education Levels



Key Observations

A larger white population is associated with lower poverty and higher education levels, which correlate with lower death rates. The highest percentage (87.76%) is in areas with low poverty and medium education.

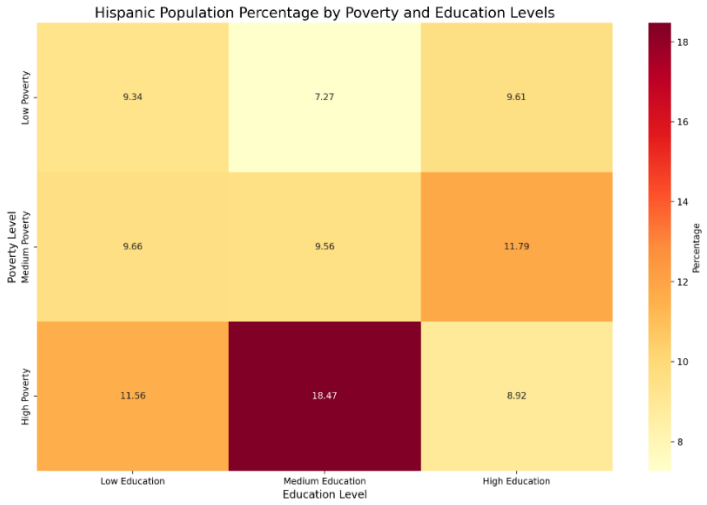
Figure 3: Black Population Percentage by Poverty and Education Levels



Key Observations

Higher percentages of Black population are concentrated in areas with higher poverty levels, which correlate with higher death rates. The highest percentage (24.02%) is in areas with extreme poverty and medium education.

Figure 4: Hispanic Population Percentage by Poverty and Education Levels



Key Observations

The Hispanic population is more evenly distributed across different poverty and education levels, but higher percentages are found in extreme poverty areas, particularly those with medium education levels.

Analysis

Based on the visualizations and data analysis, it is evident that race mediates the relationship between socioeconomic factors and cancer mortality:

Poverty and Education:

There is a strong inverse relationship between poverty/education levels and cancer death rates. Higher poverty and lower education consistently predict higher death rates.

Racial Composition:

White Population: Higher percentages of White populations are associated with lower poverty and higher education levels, which correlate with lower cancer death rates.

Black Population: Higher percentages of Black populations are associated with higher poverty levels and lower education, which correlate with higher cancer death rates.

Hispanic Population: The relationship is more complex, but higher percentages of Hispanic populations are found in extreme poverty areas, which are associated with higher cancer death rates.

Mediating Effect of Race:

Race mediates the effects of poverty and education on cancer death rates. For instance, even in extreme poverty areas, increasing education levels reduces death rates across all racial groups, but the baseline rates differ significantly by race.

Intersectionality

The data suggests an intersectional relationship between race, poverty, and education in determining cancer mortality. The effects of these factors are not independent but interact in complex ways to influence outcomes.

Potential Factors

A range of factors may influence these relationships, including access to healthcare, environmental factors, cultural differences in health behaviors, and systemic inequalities in education and healthcare systems.

Conclusion

The findings demonstrate that while poverty and education levels are strong predictors of cancer death rates, race significantly mediates these relationships. The complex interplay between race, poverty, and education suggests that efforts to reduce cancer mortality disparities must address these factors simultaneously. Targeted interventions focusing on poverty reduction, educational improvement, and addressing racial inequalities in healthcare access could be crucial in reducing cancer mortality across diverse communities.

This analysis underscores the need for multifaceted public health strategies that consider the intersectionality of race, socioeconomic status, and education to effectively address and reduce cancer disparities in the United States.

Interaction Effects on Cancer Death Rates

Introduction

This report examines the interaction effects between various socioeconomic and demographic factors on cancer death rates across different regions. Specifically, we analyze how interactions between variables such as Socioeconomic Status (SES), education, unemployment, race, and urbanicity influence the relationship between poverty and cancer death rates. By exploring these interaction effects, we aim to uncover more complex relationships that may not be evident when examining individual variables independently.

Methodology

Data Preparation

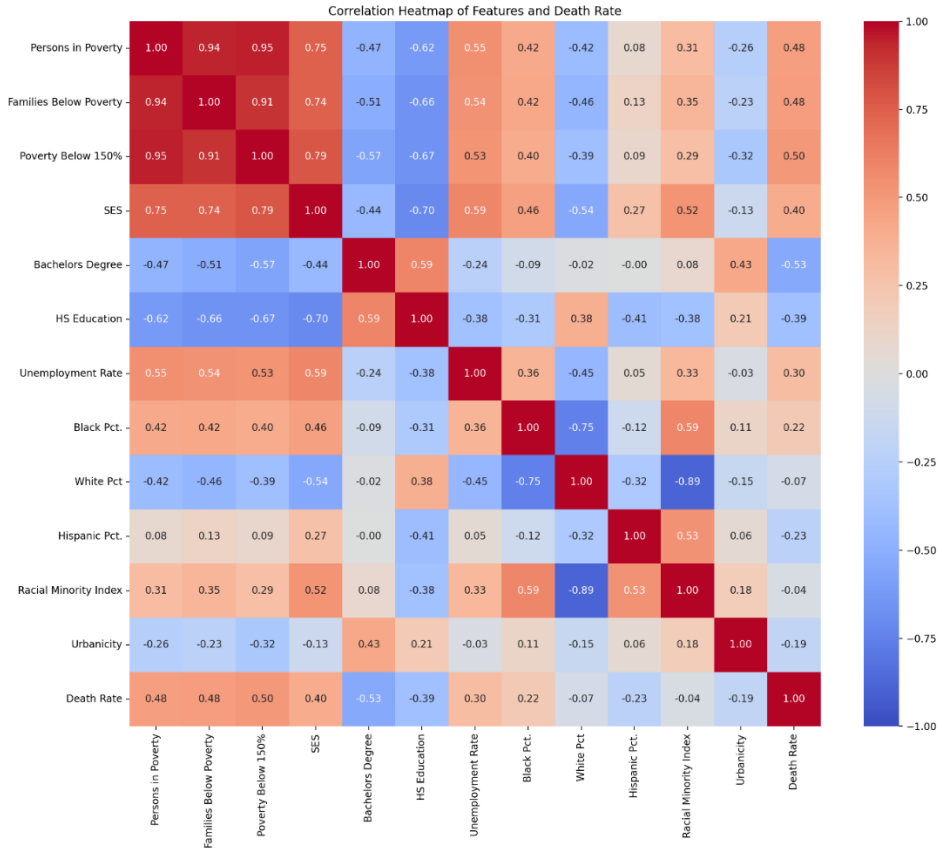
The initial dataset included the following variables: Persons in Poverty, Families Below Poverty, Poverty Below 150%, SES, Bachelor's Degree, High School Education, Unemployment Rate, Black Percentage, White Percentage, Hispanic Percentage, Racial Minority Index, and Urbanicity. Before running the regression analysis, we performed data cleaning to address any non-numeric values and standardized categorical variables.

Addressing Multicollinearity

Multicollinearity occurs when independent variables in a regression model are highly correlated, making it difficult to interpret the individual effects of each variable. To address this issue, we used the following steps:

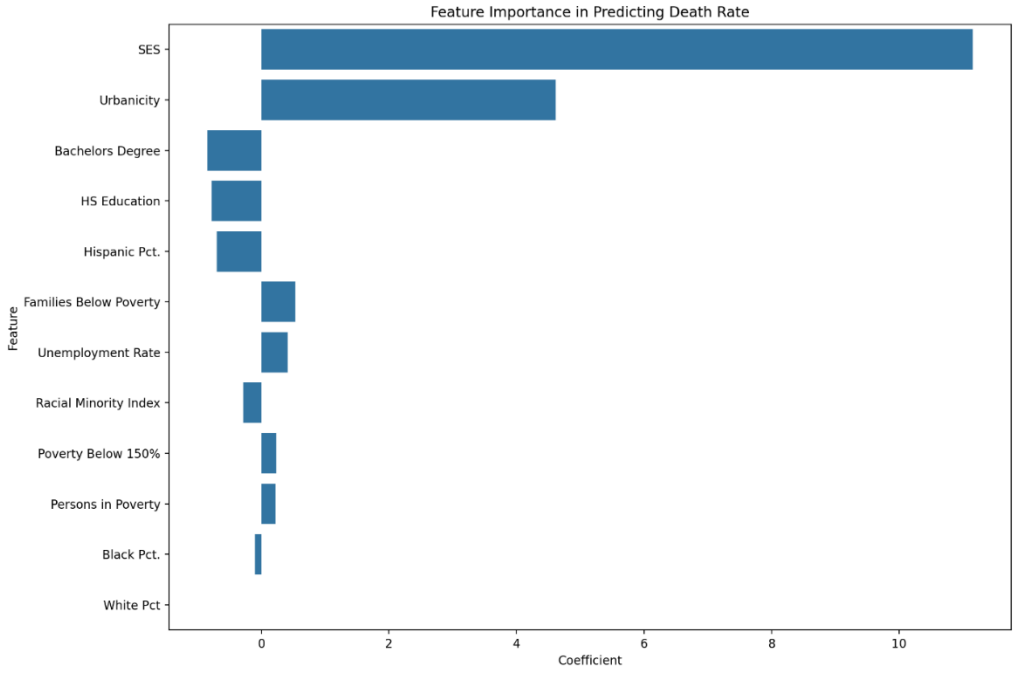
1. **Correlation Analysis:** We generated a correlation matrix to identify pairs of variables that were highly correlated (e.g., SES and Poverty Below 150%).

Figure 1: Correlation Matrix of Initial Features



2. **Variance Inflation Factor (VIF):** We calculated VIF values for all variables to quantify the severity of multicollinearity. VIF values above 5-10 indicate problematic multicollinearity.

Figure 2: Initial Variance Inflation Factors (VIF)



3. **Feature Selection:** Based on the correlation analysis and VIF values, we removed highly correlated variables to reduce multicollinearity. The final set of variables included SES, Bachelor's Degree, High School Education, Unemployment Rate, Black Percentage, Hispanic Percentage, and Urbanicity.

Figure 3: Updated Correlation Matrix of Selected Features

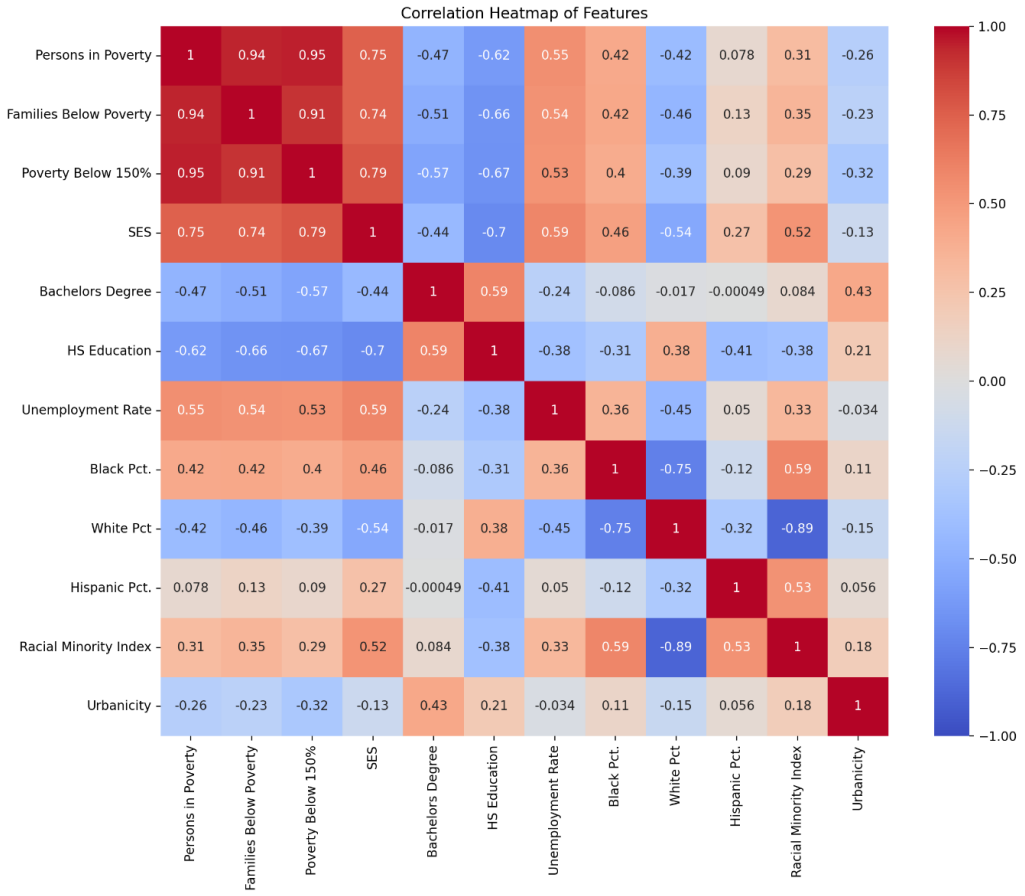
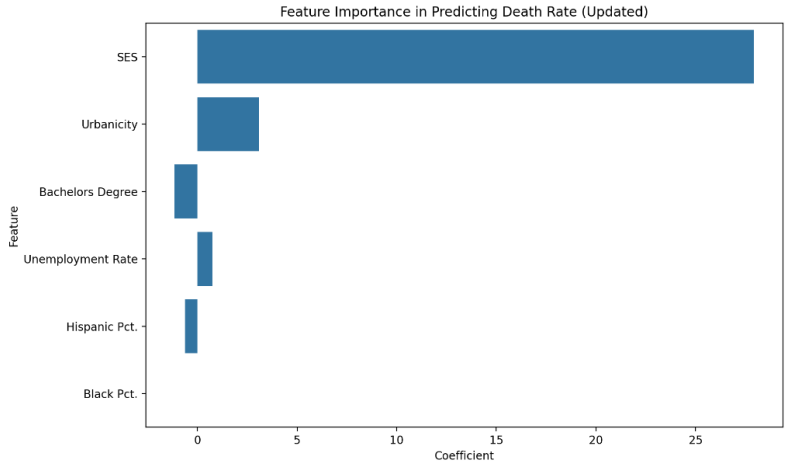


Figure 4: Updated Variance Inflation Factors (VIF)



Results

With the reduced set of features, we conducted a multiple regression analysis with cancer death rate as the dependent variable. The results are as follows:

Model Performance

Mean Squared Error: 418.4246

R-squared: 0.4340

The R-squared value of 0.4340 indicates that approximately 43.40% of the variance in cancer death rates can be explained by the selected features. Although the R-squared value decreased slightly after reducing multicollinearity, the model's interpretability has improved.

Feature Coefficients

Figure 4: Feature Coefficients in the Multiple Regression Model

Feature	Coefficient
SES	27.93
Urbanicity	3.08
Bachelor's Degree	-1.14
Unemployment Rate	0.76
Hispanic Pct.	-0.62
Black Pct.	-0.01

Interpretation of Results

Main Effects

SES: SES remains a significant predictor of cancer death rates, with a strong positive relationship. This suggests that areas with higher socioeconomic status tend to have higher death rates, which is counterintuitive and requires further investigation.

Urbanicity: Urban areas tend to have higher cancer death rates, due to factors such as environmental pollution, stress, and lifestyle differences between urban and rural areas.

Education and Race: Higher education levels (Bachelor's Degree) and higher Hispanic populations are associated with lower death rates, which aligns with the "Hispanic Paradox" observed in public health literature.

Exploring Interaction Effects

To further understand the complex relationships between these variables, we explored interaction effects between SES and other key predictors (Bachelor's Degree, Unemployment Rate, and Urbanicity). Interaction terms help us understand how the relationship between SES and cancer death rates varies across distinct levels of these variables.

Model Performance with Interaction Effects

Mean Squared Error: 405.5724

R-squared: 0.4514

Including interaction terms improved the model's explanatory power, as indicated by an increase in the R-squared value from 0.4340 to 0.4514.

Interaction Effects

Figure 5: Feature Coefficients (Including Interaction Terms)

Feature	Coefficient
SES	13.49
Bachelor's Degree	-10.26
Hispanic Pct.	-8.58
Urbanicity	4.81
SES x Urbanicity	-3.88
Unemployment Rate	3.48
SES x Unemployment	-3.35
SES x Bachelor's Degree	-2.59
Black Pct.	0.05

Figure 6: Feature Importance in Predicting Death

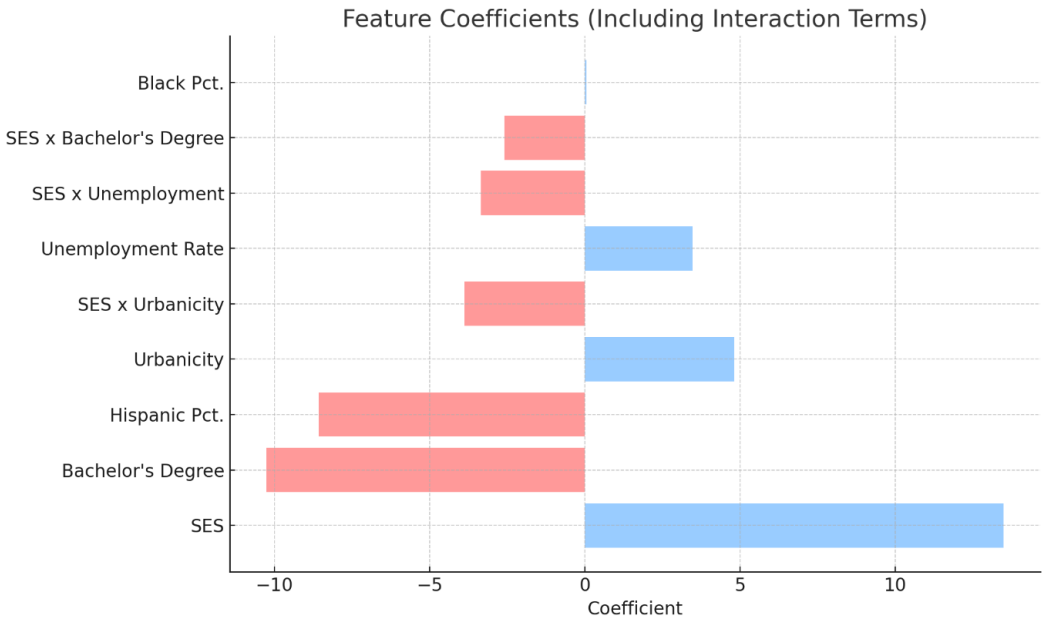
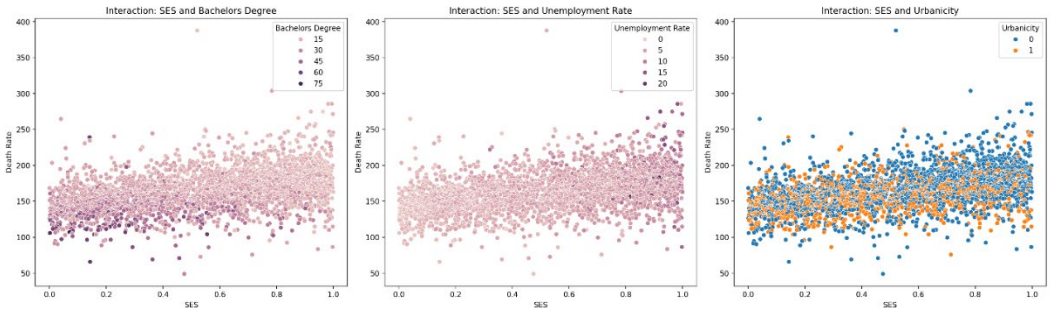


Figure7: Interaction Effects Visualization



Interpretation of Interaction Effects

1. **SES x Urbanicity:** The positive effect of SES on cancer death rates is less pronounced in urban areas. This suggests that urban areas may have resources or characteristics that mitigate some of the negative impacts associated with higher SES.

2. **SES x Unemployment:** The effect of SES on cancer death rates is reduced in areas with higher unemployment. This indicates that unemployment may have a more uniform effect across different SES levels.

3. **SES x Bachelor's Degree:** Higher education levels reduce the positive relationship between SES and cancer death rates, implying that education has a protective effect, particularly in areas with higher SES.

Implications and Next Steps

The findings from this analysis reveal that the relationship between SES and cancer death rates is not uniform and is influenced by other factors such as education, unemployment, and urbanicity. These results have several important implications:

Targeted Interventions: Public health interventions should be tailored to account for the complex interactions between SES, education, and urbanicity. For example, strategies to reduce

cancer death rates in urban areas may differ from those in rural areas due to the different dynamics at play.

Education's Protective Effect: The strong negative interaction between SES and Bachelor's Degree highlights the importance of education in improving health outcomes, suggesting that policies promoting higher education levels could be effective in reducing cancer death rates.

Further Investigation: The counterintuitive positive relationship between SES and cancer death rates, even after accounting for interactions, warrants further investigation to identify potential confounding factors or measurement issues.

Conclusion

This analysis has provided a deeper understanding of how interaction effects between socioeconomic and demographic variables influence cancer death rates. By accounting for these interactions, we have gained valuable insights that can inform more effective public health strategies. Future research should continue to explore these complex relationships and consider additional variables that may further elucidate the determinants of cancer outcomes.

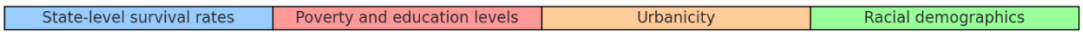
Survival Rate Analysis: Impact of Socioeconomic and Demographic Factors

Introduction

This section explores the relationship between cancer survival rates and a range of factors, including state-level differences, poverty, urbanicity, race, and education. The analysis aims to uncover patterns and correlations that may contribute to disparities in cancer outcomes across the United States.

Methodology

The dataset used for this analysis includes information on cancer survival rates across different counties in the United States, along with demographic, socioeconomic, and geographic data. The analysis focuses on:



Results

Correlation Analysis:

Correlation between Survival Rate and Poverty: r = -0.4428

The analysis reveals a moderate negative correlation between survival rates and poverty, indicating that counties with higher poverty levels tend to have lower cancer survival rates. This relationship underscores the impact of socioeconomic factors on health outcomes.

Descriptive Statistics:

Figure 1: Survival Rates

Statistic	Value
Mean Survival Rate	64%
Standard Deviation	6%
Range	32.2% to 83.2%

These statistics show considerable variation in cancer survival rates across counties, with some areas experiencing survival rates more than twice as high as others.

Figure 2: Top 5 Counties with Highest Survival Rates:

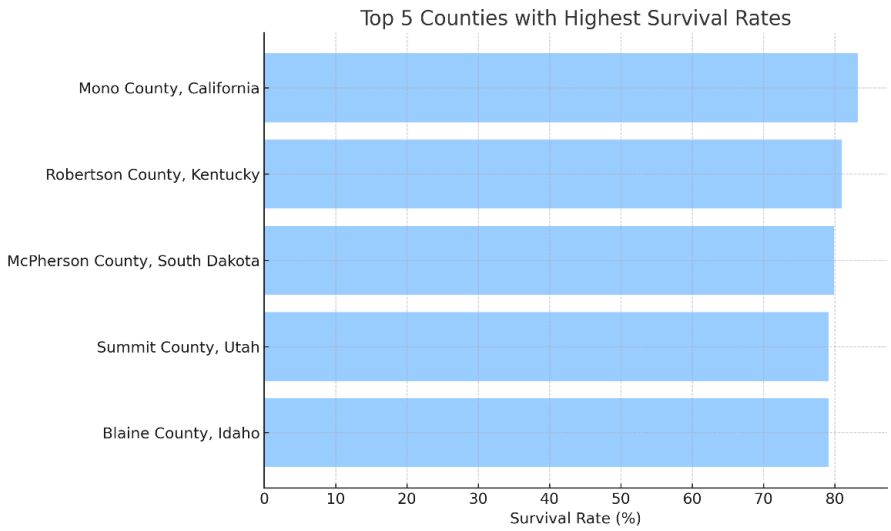
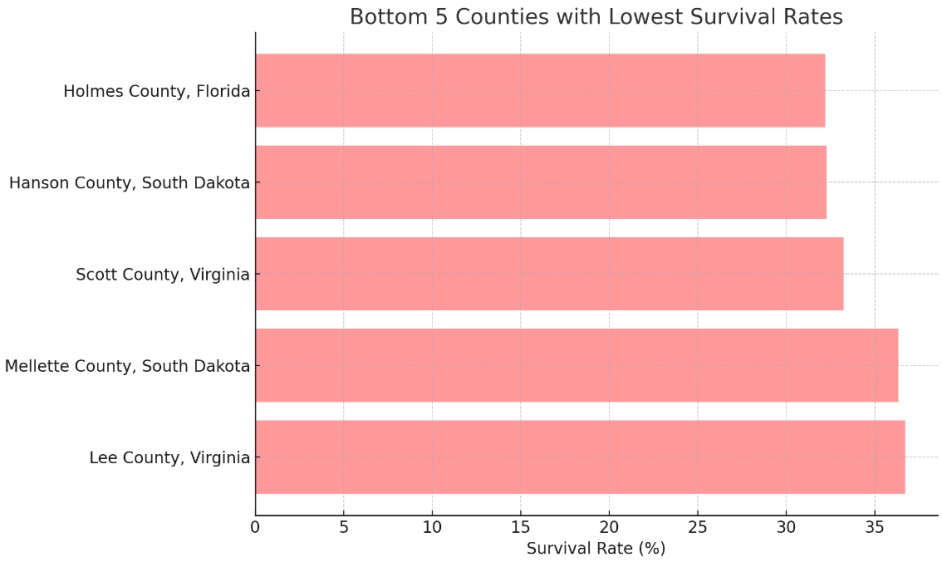


Figure 2: Bottom 5 Counties with Lowest Survival Rates:



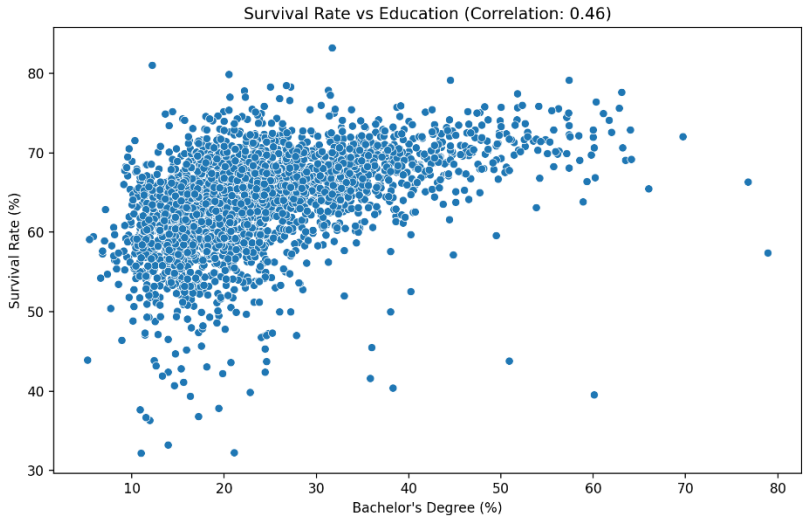
These results highlight the stark disparities in cancer survival rates between different regions, with a difference of over fifty percentage points between the highest and lowest rates.

Survival Rate and Education

Survival Rate and Education (Bachelor's Degree):

r = 0.4552: There is a moderate positive correlation between survival rates and the percentage of the population with a Bachelor's degree. This suggests that higher education levels are associated with better cancer survival outcomes.

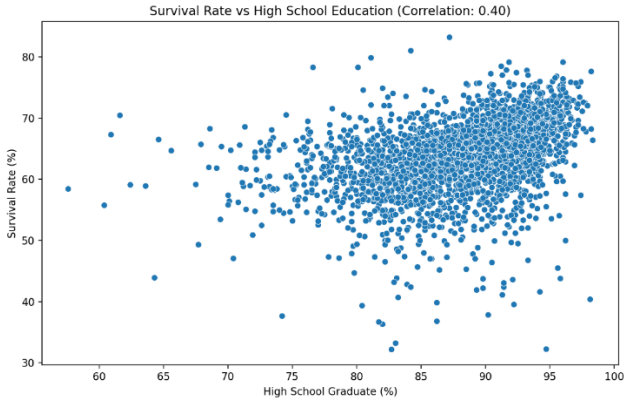
Figure 3: *Survival Rate vs. Education (Bachelor's Degree)*



Survival Rate and High School Graduation

r = 0.3974: A moderate positive correlation was observed between survival rates and the percentage of high school graduates. This relationship is slightly weaker than that observed for Bachelor's degree holders but still significant.

Figure 4: *Survival Rate vs. High School Graduation Rate*



State-Level Analysis

Figure 5: *Top 5 States by Average Survival Rate*

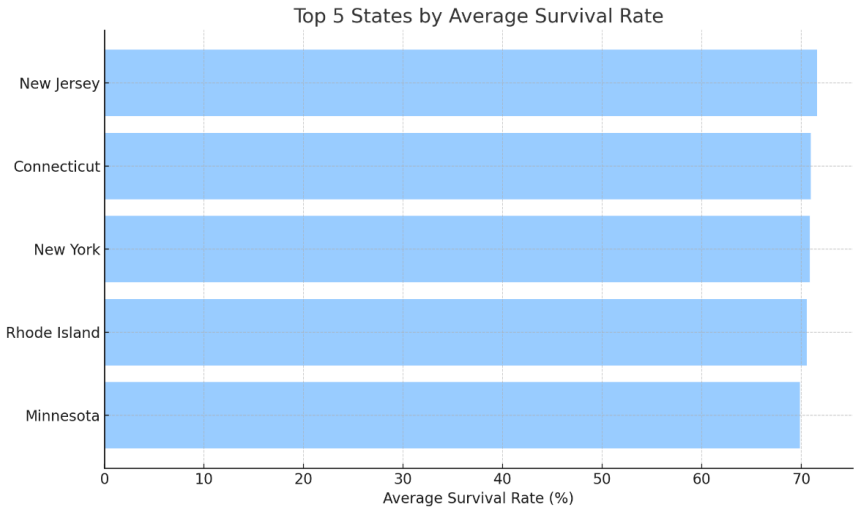
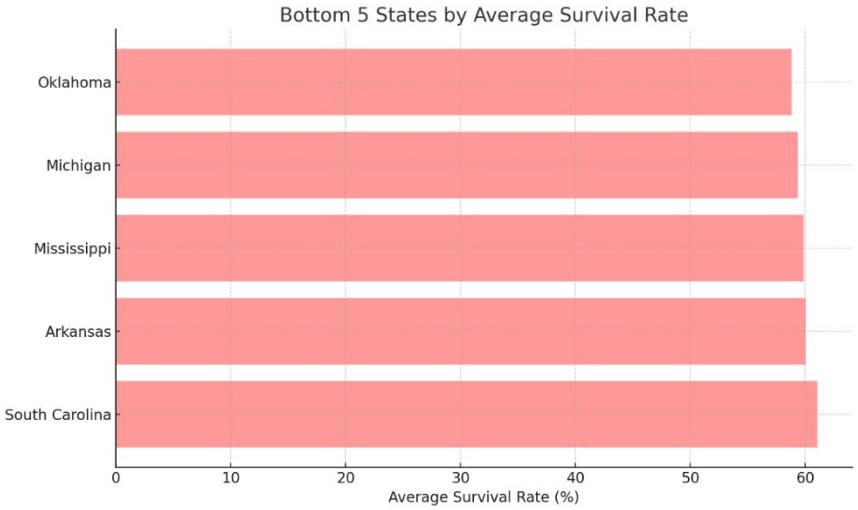


Figure 6: *Bottom 5 States by Average Survival Rate*



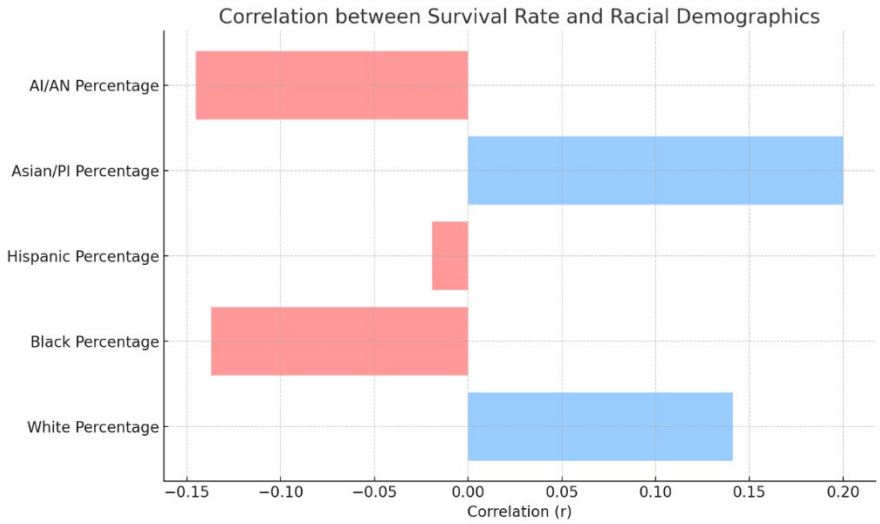
These visualizations emphasize the substantial disparities in survival rates across different states, with the top states averaging survival rates more than ten percentage points higher than the bottom states.

Survival Rate and Racial Demographics

Correlation between Survival Rate and Racial Demographics:

These correlations indicate that counties with higher percentages of White and Asian/Pacific Islander populations tend to have slightly higher survival rates, while those with higher percentages of Black, Hispanic, and American Indian/Alaska Native populations tend to have slightly lower survival rates.

Figure 5: *Survival Rate and Racial Demographics*



Discussion and Recommendations

Poverty Impact:

The moderate negative correlation between poverty and survival rates suggests that socioeconomic factors play a critical role in cancer outcomes. Areas with higher poverty levels likely face challenges related to access to healthcare, early detection, and treatment quality.

Education and Survival Rates:

The positive correlations between education levels and survival rates suggest that higher educational attainment may contribute to better health outcomes, potentially through improved health literacy, access to resources, and overall socioeconomic status.

State-Level Differences:

Significant disparities exist between states, with survival rates varying widely. States with higher average survival rates may have better healthcare infrastructure, more effective public health policies, or other advantageous factors that could be emulated in lower-performing states.

Racial Demographics:

The weak correlations between survival rates and racial demographics indicate that race alone may not be the primary determinant of survival outcomes. However, these relationships may be influenced by broader social determinants of health, including socioeconomic status, healthcare access, and environmental factors.

Recommendations:

1. **Targeted Interventions**: Focus on improving healthcare access and quality in areas with extreme poverty levels to reduce disparities in survival rates.
2. **Education Initiatives**: Promote educational programs that enhance health literacy and awareness, particularly in regions with lower survival rates.

3. **State-Level Policy**: Investigate successful policies and practices in states with high survival rates and consider implementing similar strategies in states with lower rates.

4. **Comprehensive Approaches**: Address the broader social determinants of health, including socioeconomic factors and healthcare access, to improve survival rates across diverse populations.

This analysis provides critical insights into the factors influencing cancer survival rates across the United States, highlighting the need for multifaceted, targeted approaches to reduce disparities and improve outcomes.

Multiple Regression Analysis on Survival Rates

Introduction

This analysis focuses on understanding the factors influencing cancer survival rates using a multiple regression model. The dependent variable in this study is the survival rate, and various socioeconomic, educational, racial, and geographic variables were considered as predictors. Additionally, a multicollinearity analysis was conducted to ensure the robustness of the regression model.

Methodology

A multiple regression analysis was performed with survival rate as the dependent variable. The independent variables included in the initial model were:

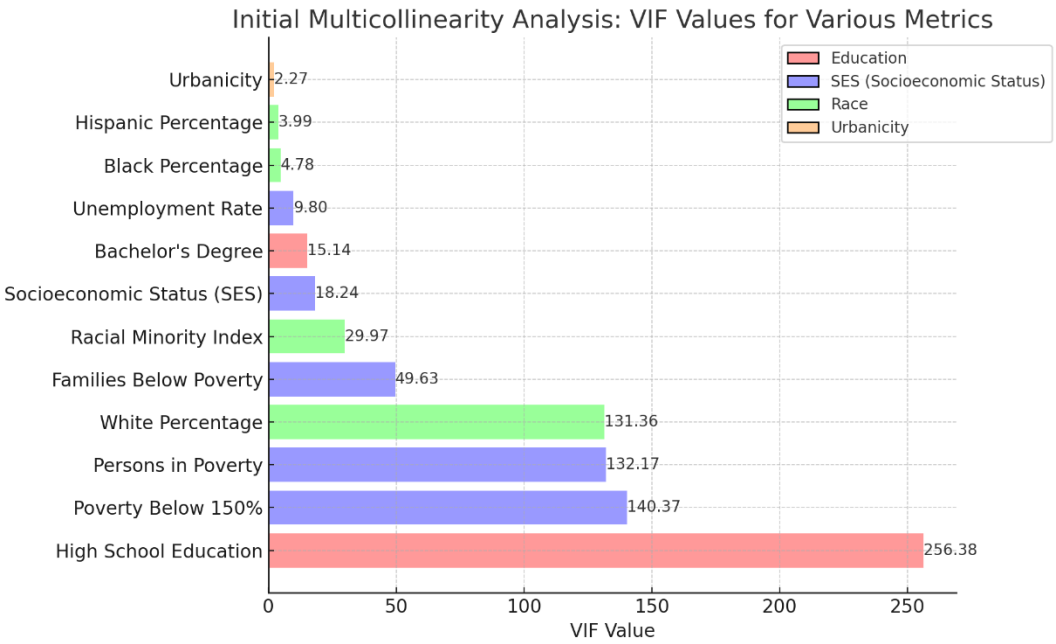
Education	Poverty	Demographics	Urbanization
High School Education	Poverty Below 150%	White Percentage	Urbanicity
Bachelor's Degree	Persons in Poverty	Black Percentage	
	Families Below Poverty	Hispanic Percentage	
	Unemployment Rate	Racial Minority Index	

A multicollinearity analysis was conducted using Variance Inflation Factors (VIF) to identify highly correlated variables, which could potentially distort the regression coefficients.

Results

1. Initial Multicollinearity Analysis

Figure 1: *Initial Variance Inflation Factors*

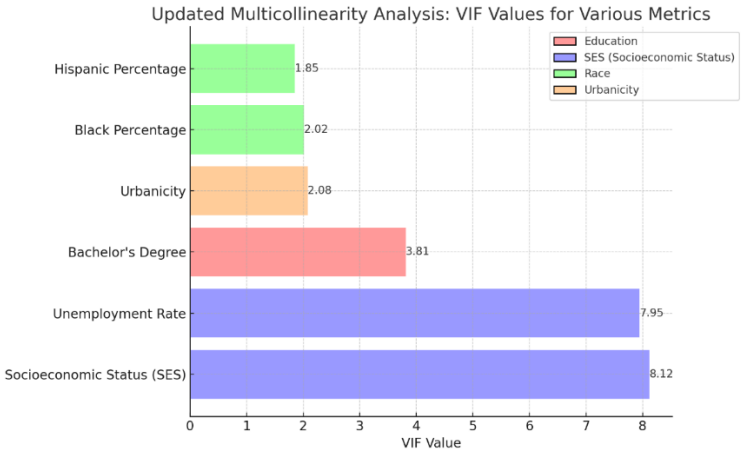


These initial VIF values indicated significant multicollinearity, particularly among education and poverty-related variables.

2. Updated Multicollinearity Analysis

After feature selection to address multicollinearity, the following variables were retained:

Figure 2: *Updated Variance Inflation Factors*



All updated VIF values are below the common threshold of ten, indicating acceptable levels of multicollinearity.

3. Model Performance

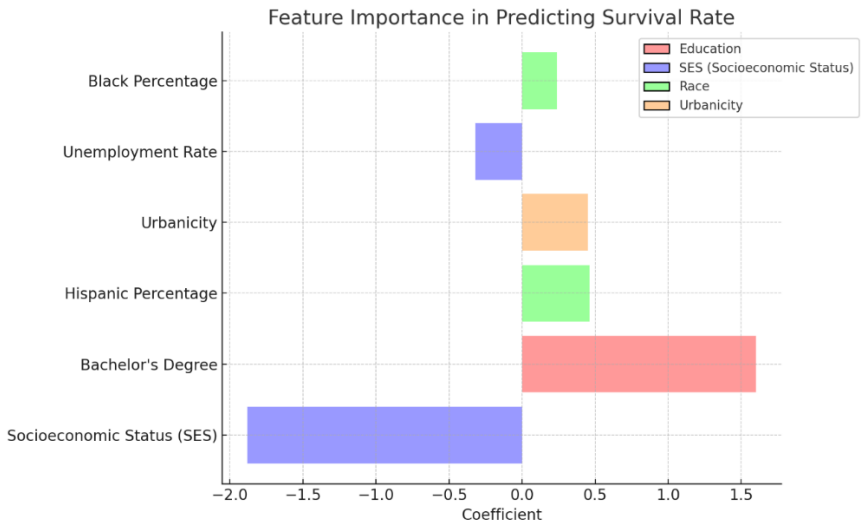
Mean Squared Error (MSE): 23.75

R-squared Score: 0.2915

The R-squared score of 0.2915 suggests that approximately 29.15% of the variance in the survival rate is explained by the selected features. This indicates that while the model has some predictive power, other key factors influencing survival rates may not have been included in the analysis.

4. Feature Coefficients

Figure 3: *Feature Coefficients for Survival Rate Prediction*



Key Findings and Interpretations

1. **SES (Socioeconomic Status)**: The negative coefficient for SES is surprising and suggests a counterintuitive relationship with survival rates. This finding warrants further investigation to understand whether this result is due to the way SES is calculated or other confounding factors.

2. **Bachelor's Degree**: A strong positive relationship with survival rates supports the well-established link between higher education levels and better health outcomes.

3. **Hispanic and Black Percentages**: The positive relationships between these racial demographics and survival rates are intriguing. These findings might relate to the "Hispanic Paradox," where Hispanic populations often experience better health

outcomes despite lower socioeconomic status. Further exploration of underlying factors is necessary.

4. **Urbanicity**: The positive coefficient for urbanicity suggests that urban areas, due to better access to healthcare facilities, tend to have higher survival rates.

5. **Unemployment Rate**: The negative relationship between unemployment rate and survival rates aligns with the understanding that employment status is a critical determinant of health.

6. **Residual Analysis**: The residual plot indicated patterns suggesting potential non-linear relationships or omitted variables not captured by the linear model.

Next Steps

1. **Investigate SES**: The unexpected negative relationship between SES and survival rates should be explored further. This may involve re-evaluating the SES calculation or identifying confounders.

2. **Explore Non-linear Relationships**: Given the patterns observed in the residual plot, it may be beneficial to consider non-linear modeling approaches or interaction terms in the regression model.

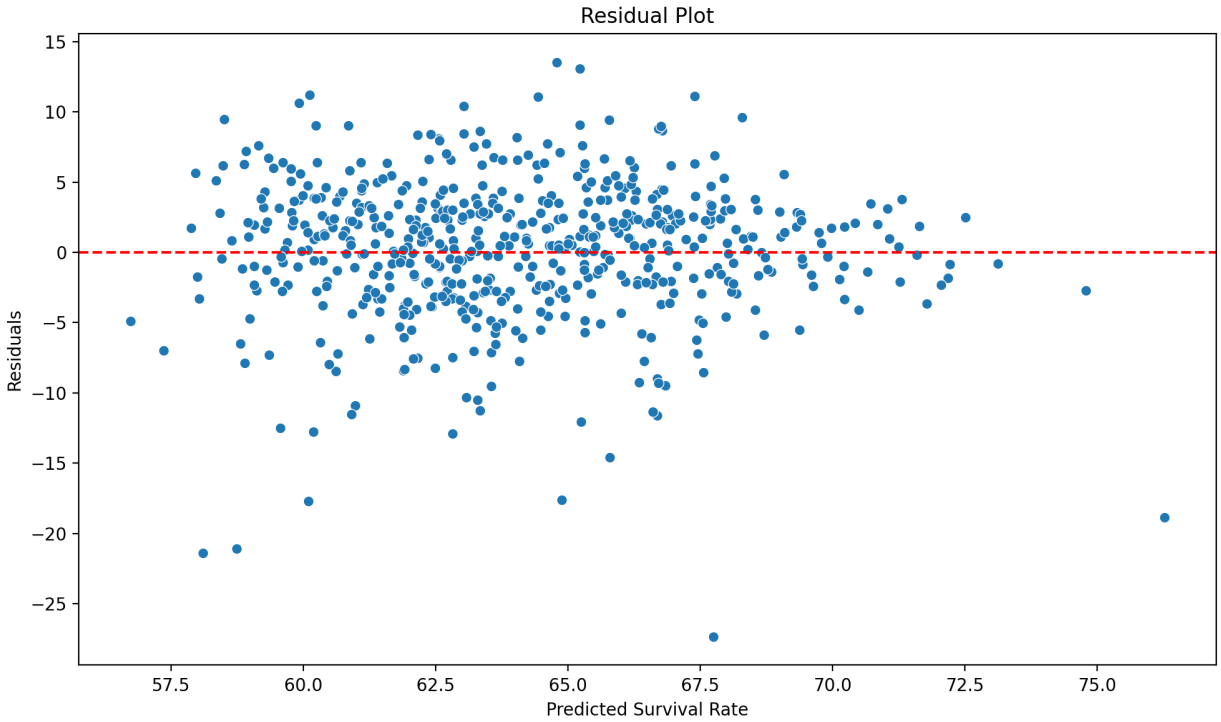
3. **Incorporate Additional Variables**: With only 29% of the variance in survival rates explained by the current model, exploring other potential predictors could improve model performance.

4. **Detailed Urbanicity Analysis**: Urbanicity's impact on survival rates could be further dissected by examining more specific urban characteristics or categorizing different urban areas.

5. **Health Policy Implications**: The positive associations with Hispanic and Black populations should be explored further to understand how these findings can inform targeted health policies and interventions.

6. **Advanced Modeling Techniques**: Consider using more sophisticated models, such as random forests or gradient boosting machines, which can automatically account for complex interactions and non-linearities in the data.

Figure 4: *Residual Plot of Survival Rate Model*



Conclusion

This multiple regression analysis provides insight into the complex factors influencing cancer survival rates. While the model highlights the significant roles of education, race, and urbanicity, the unexpected findings related to SES call for further research. Future efforts should

focus on refining the model, exploring non-linear relationships, and incorporating additional variables to enhance our understanding of survival rates and inform public health strategies.

Survival Rate and Death Rate Prediction Using Machine Learning

Introduction

This analysis employs machine learning to predict cancer survival and death rates by leveraging a comprehensive dataset containing demographic, socioeconomic, and geographic variables. The core goal is to enhance our understanding of how various factors contribute to cancer outcomes and to develop models that accurately predict these outcomes. Insights from this analysis can help guide public health policies and interventions, especially in the areas of education and poverty alleviation, where disparities in health outcomes are most pronounced.

Methodology

Machine Learning Approach:

The machine learning models employed in this analysis use Random Forest Regression. Random Forests, an ensemble learning method, were chosen due to their robustness in handling complex, non-linear relationships between features. Random Forest models create multiple decision trees during training and combine their predictions, thus reducing overfitting, improving accuracy, and handling multicollinearity between variables.

Key Variables in the Analysis: The analysis considered a variety of predictors:

- Demographic factors: racial composition (White, Black, Hispanic percentages)
- Socioeconomic indicators: Education levels (Bachelor's Degree rate, High School Education), unemployment, income, and poverty levels (Poverty Below 150%)
- Geographic information: Urbanicity(Urban vs. Rural)

Education	Economics	Demographics	Urbanization
High School Education	Poverty Below 150%	White Percentage	Urbanicity
Bachelor's Degree	Persons in Poverty	Black Percentage	
	Families Below Poverty	Hispanic Percentage	
	Socioeconomic Status (SES)	Racial Minority Index	
	Unemployment Rate		

Why Random Forest?

Random Forest models are highly effective for complex datasets that may have many interacting variables. This method also provides the benefit of ranking the importance of predictors, which is essential for understanding which factors are most influential in determining cancer survival and death rates. Additionally, Random Forest handles missing values and scaling issues well, making it suitable for this dataset's variety of variables.

Results

Random Forest Model for Death Rate:

- **Mean Squared Error (MSE):** 387.2382
- **R-squared Score:** 0.4762

The Random Forest model explained 47.62% of the variance in death rates. The moderately high R-squared value suggests that the predictors used in the model capture a substantial portion of the factors influencing death rates.

Random Forest Model for Survival Rate:

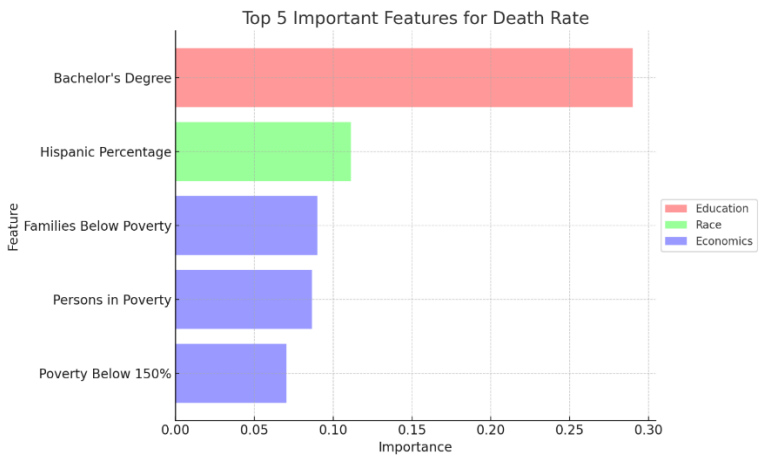
- **Mean Squared Error (MSE):** 25.0846
- **R-squared Score:** 0.2516

For survival rates, the model explained only 25.16% of the variance, indicating that survival rates might be influenced by factors not captured in the dataset or more complex non-linear interactions that the current model does not fully explain.

Feature Importance:

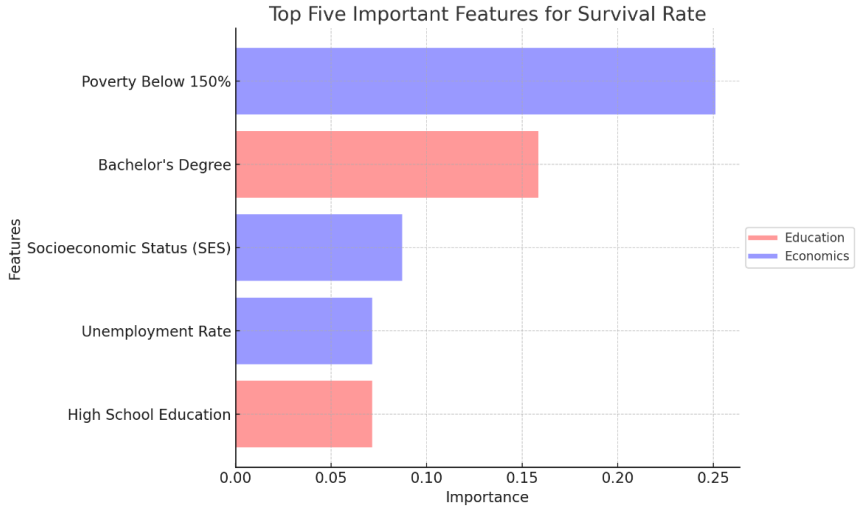
Death Rate Prediction: The most significant variable was the percentage of individuals with a Bachelor's Degree, followed by the Hispanic percentage. Education emerged as a consistent predictor across both death and survival models, signifying its central role in influencing health outcomes.

Figure 1: *Feature Importance for Death Rate Prediction*



Survival Rate Prediction: Similarly, education levels and poverty-related factors were crucial. Poverty (Poverty Below 150%) played a stronger role in predicting survival rates than death rates, suggesting economic factors significantly affect long-term cancer outcomes.

Figure 2: *Feature Importance for Survival Rate Prediction*



Discussion

1. Model Performance & Interpretations:

The Random Forest model for death rate performed notably better than the survival rate model. This could be due to the fact that death rates might be influenced by well-documented health disparities such as access to healthcare, lifestyle factors, or late-stage diagnosis, which are more readily quantifiable in the dataset. In contrast, survival rates might be influenced by complex factors not fully captured here, including treatment types, genetic predisposition, and other unrecorded individual-level data.

The lower R-squared for survival rates (25.16%) suggests that survival outcomes are more challenging to predict. Future research should explore adding clinical data (e.g., stage of diagnosis, treatment types) or patient-level data that can capture the complexities of cancer survivorship.

2. Significance of Education:

The finding that Bachelor's Degree percentage was the most influential predictor for both death and survival rates reflects the strong correlation between education and health. Individuals with higher education levels typically have better access to healthcare, more awareness of preventative measures, and higher income—all of which positively impact health outcomes.

3. Ethnic Composition and Health Outcomes:

The Hispanic percentage emerged as a significant predictor for death rates but not survival rates. This could suggest that while mortality rates among Hispanic populations are affected by socioeconomic and healthcare access disparities, survival rates may not be as significantly impacted by ethnicity alone. This raises important questions about the specific barriers facing different ethnic groups and the role of healthcare access in mitigating these disparities.

4. Economic Disparities:

The prominence of poverty-related variables (such as Poverty Below 150%) in predicting survival rates emphasizes that poverty directly affects long-term health outcomes. Economic instability can limit access to timely treatments, quality care, and long-term follow-up, leading to worse survival rates.

5. Non-linear Relationships:

The improved performance of the Random Forest models over linear models indicates the presence of non-linear relationships between predictors and outcomes. Given this, exploring more advanced non-linear models, such as **Gradient Boosting Machines** (GBM) and **Neural Networks**, may further improve predictive accuracy. GBM, in particular, has the potential to provide higher accuracy by sequentially correcting model errors, while Neural Networks could capture more intricate patterns.

6. Model Refinements:

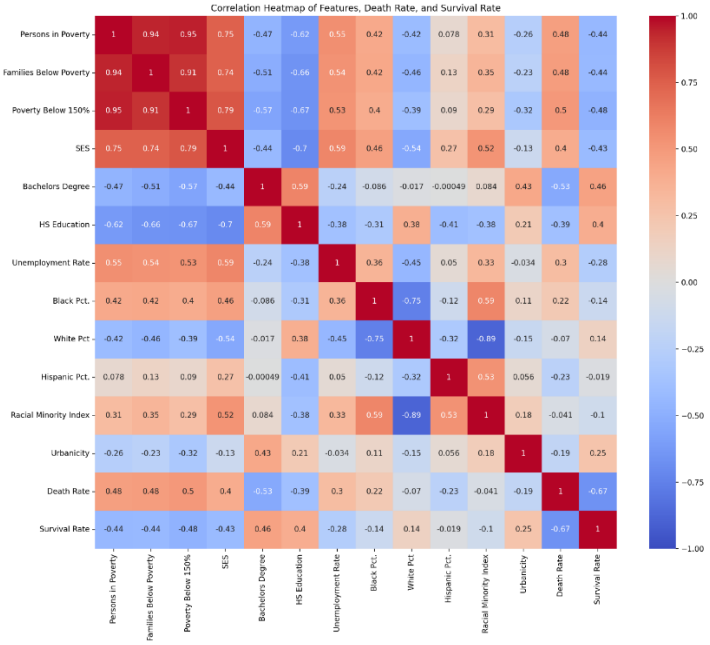
Future improvements could be achieved through:

- **Hyperparameter tuning**: Optimizing the Random Forest model parameters (number of trees, maximum depth, etc.).
- **Feature engineering**: Creating interaction terms (e.g., interaction between poverty levels and race) to better capture complex relationships.
- **Incorporating additional variables**: Integrating clinical data, individual health behaviors, and healthcare access metrics to improve model comprehensiveness.

7. Geospatial Analysis:

Incorporating geospatial data could offer new insights into how geographic disparities contribute to health outcomes. For example, regional variations in healthcare access, environmental factors, and social determinants of health could be explored using **spatial machine learning** techniques.

Figure 3: *Correlation Heatmap of Predictors*



Conclusion

This analysis successfully used machine learning to identify key predictors of cancer survival and death rates, with education and poverty emerging as the most critical factors. The Random Forest model performed well for death rate prediction, though further improvements could be made for survival rates. Policymakers should consider these findings in designing interventions

that target education and poverty to improve health outcomes, especially in underprivileged and ethnically diverse populations.

Future Directions

1. **Exploring Other Algorithms**: Testing additional models like **Gradient Boosting Machines (GBM)**, **Support Vector Machines (SVM)**, and **Neural Networks** could uncover more complex relationships.
2. **Survival Analysis**: Applying survival analysis techniques such as **Cox proportional hazards** or **Kaplan-Meier curves** could help better understand survival times and the risk factors associated with survival.
3. **Treatment Effect Modeling**: Incorporating treatment and healthcare data to predict the effectiveness of specific treatments on survival rates.
4. **Longitudinal Analysis**: Collecting and analyzing longitudinal data to better capture the changes in patient health status over time and the impact of interventions.

Suggested Machine Learning Applications:

Time-to-Event Modeling: Using survival analysis techniques (Cox models, Kaplan-Meier analysis) to predict not just survival rates but time-to-event (i.e., time until death or remission).

Healthcare Resource Optimization: Machine learning models can predict healthcare resource needs (hospital beds, staff) based on predicted death or survival rates, improving resource allocation.

Risk Stratification Models: Build models that identify high-risk individuals based on socioeconomic and demographic factors, guiding targeted interventions to improve cancer outcomes.

Works Cited

American Cancer Society. "Cancer Facts & Figures 2024." American Cancer Society, 2024. www.cancer.org.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32.

Cancer Health. "AACR Releases Cancer Disparities Progress Report 2024." 2024. www.cancerhealth.com.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). The elements of statistical learning (Vol. 1, No. 10). New York: Springer series in statistics.

National Cancer Institute. "Persistent Poverty's Impact on Cancer Death." 2024. www.cancer.gov.

USAFacts. "Which States Have the Highest Cancer Rates?" USAFacts, 2024. www.usafacts.org.

BMC Medicine. "Association between Social Determinants of Health and Survival among the US Cancer Survivors Population." 2024. bmcmedicine.biomedcentral.com.