Socioeconomic and Demographic Correlates of Cancer Outcomes Across U.S. States

A Detailed Statistical and Machine Learning Analysis

Presented by: Ahmad Qadafi

Introduction

This study explores the relationships between demographic factors such as poverty, education, and race, and cancer outcomes across U.S. counties.

By analyzing these variables, the study aims to uncover patterns that may provide insights into the distribution and determinants of cancer-related outcomes.

Data Acquisition

The datasets were downloaded from the State Cancer Profiles website and included:

- Death.csv: Cancer death rates by county
- Incd.csv: Cancer incidence rates by county
- These datasets are critical for analyzing the distribution and determinants of cancer-related outcomes.

Data Cleaning and Restructuring

Key steps included:

- Modification of County columns by removing irrelevant trailing numbers.
- Separation of state names into a new 'States' column.
- Renaming critical columns for clarity.
- Conversion of relevant columns to numeric data types.
- Handling missing data by storing incomplete rows in separate dataframes.

Creation of New Dataframes

Two new dataframes were created:

- Cancer Death Rates: Includes County, Death Rate, 5 Year Trend in Death Rates, Urbanicity.
- Cancer Incidence Rates: Includes County, Incidence Rate, 5 Year Trend in Incidence Rate.
- These were merged into a unified dataframe 'Cancer Rates' for comprehensive analysis.

Data Wrangling: Part II

Additional demographic and socioeconomic variables were added to enrich the dataset.

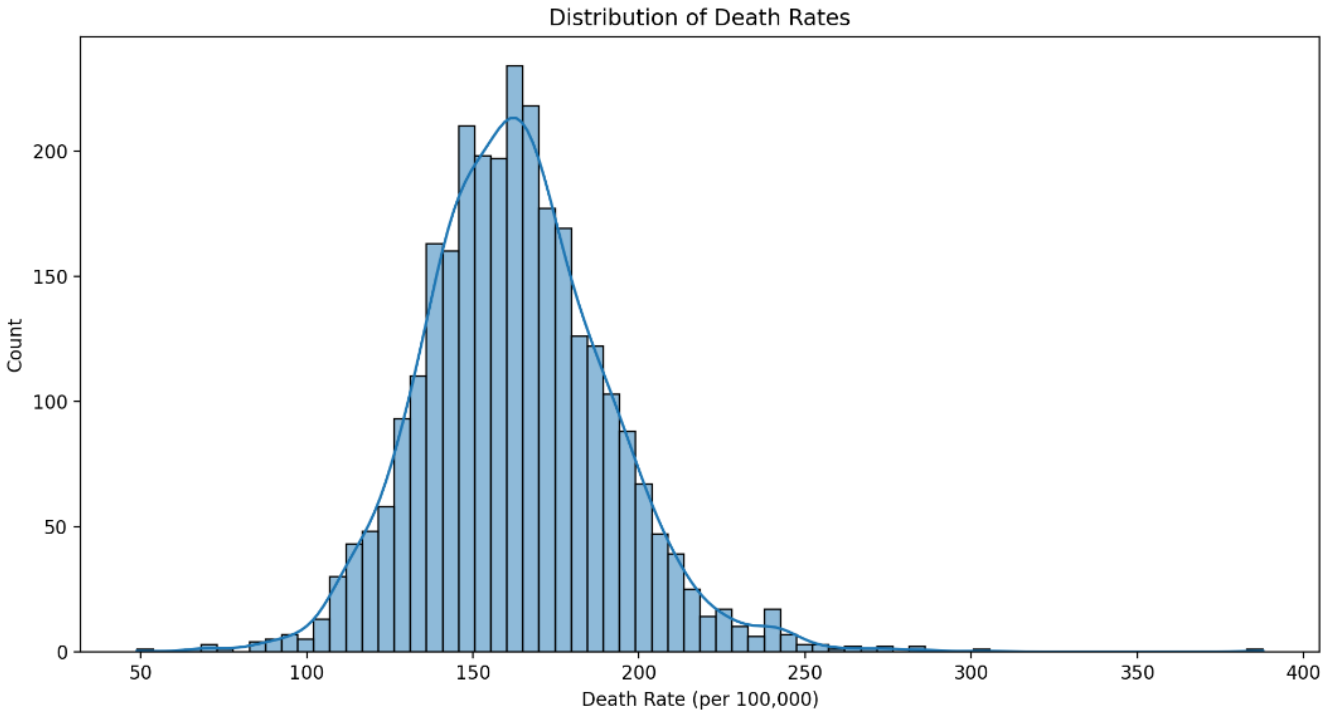
- Data files included: Poverty.csv, Education.csv, Unemployment.csv, and others.
- These were merged to create the final dataset 'Merged_Cancer_Rates_with_FIPS_Final.csv', which includes 24 columns.

https://github.com/Akqadafi/Cancer_Rates_Analysis/blob/main/CancerRates_Wrangling.ipynb

Exploratory Data Analysis: Distribution of Death Rates

- A histogram was created to analyze the distribution of cancer death rates across U.S. counties.
- The distribution appears roughly normal with a slight right skew, indicating most counties have death rates clustered around the mean.

https://github.com/Akqadafi/Cancer_Rates_Analysis/blob/main/EDA_Cancer%20Rates.ipynb

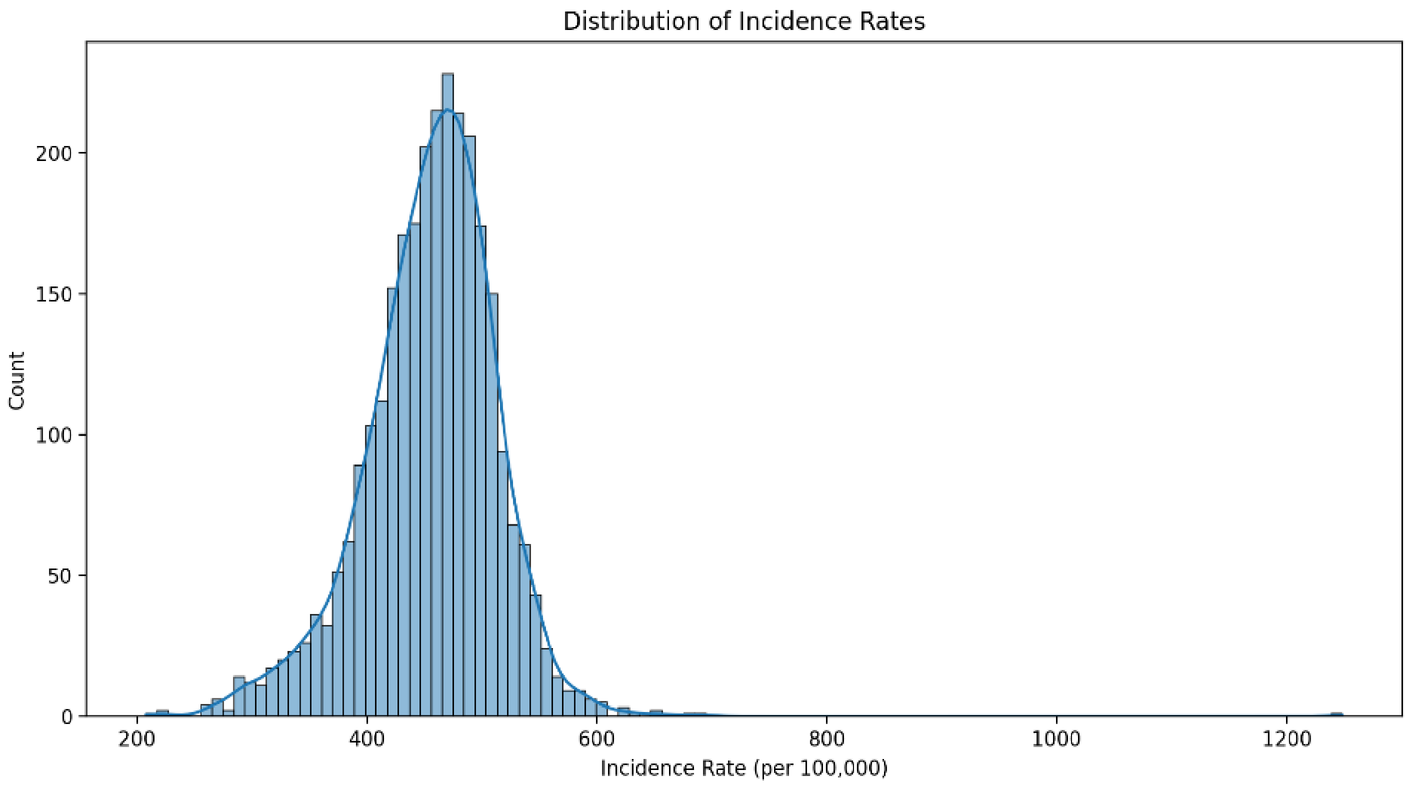


Distribution of Death Rates

Exploratory Data Analysis: Distribution of Incidence Rates

A histogram was also created for cancer incidence rates.

Like death rates, incidence rates across counties exhibit a roughly normal distribution with a right skew.



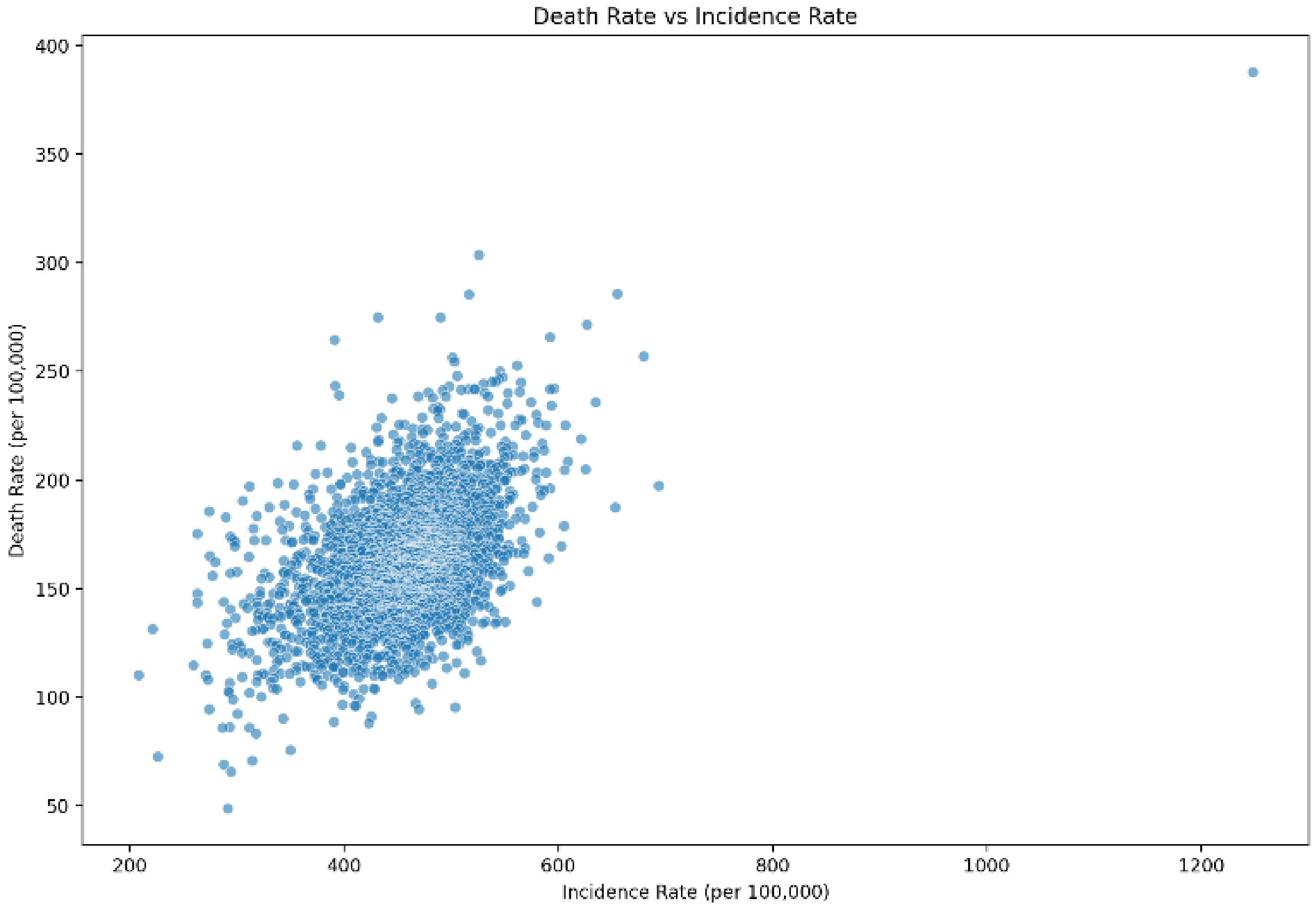
Visual: Distribution of Incidence Rates

[Placeholder for Visual: Histogram of Incidence Rates]

Exploratory Data Analysis: Death Rates vs. Incidence Rates

A scatter plot revealed a positive correlation (r = 0.49) between cancer death rates and incidence rates, indicating that higher incidence rates tend to correlate with higher death rates.

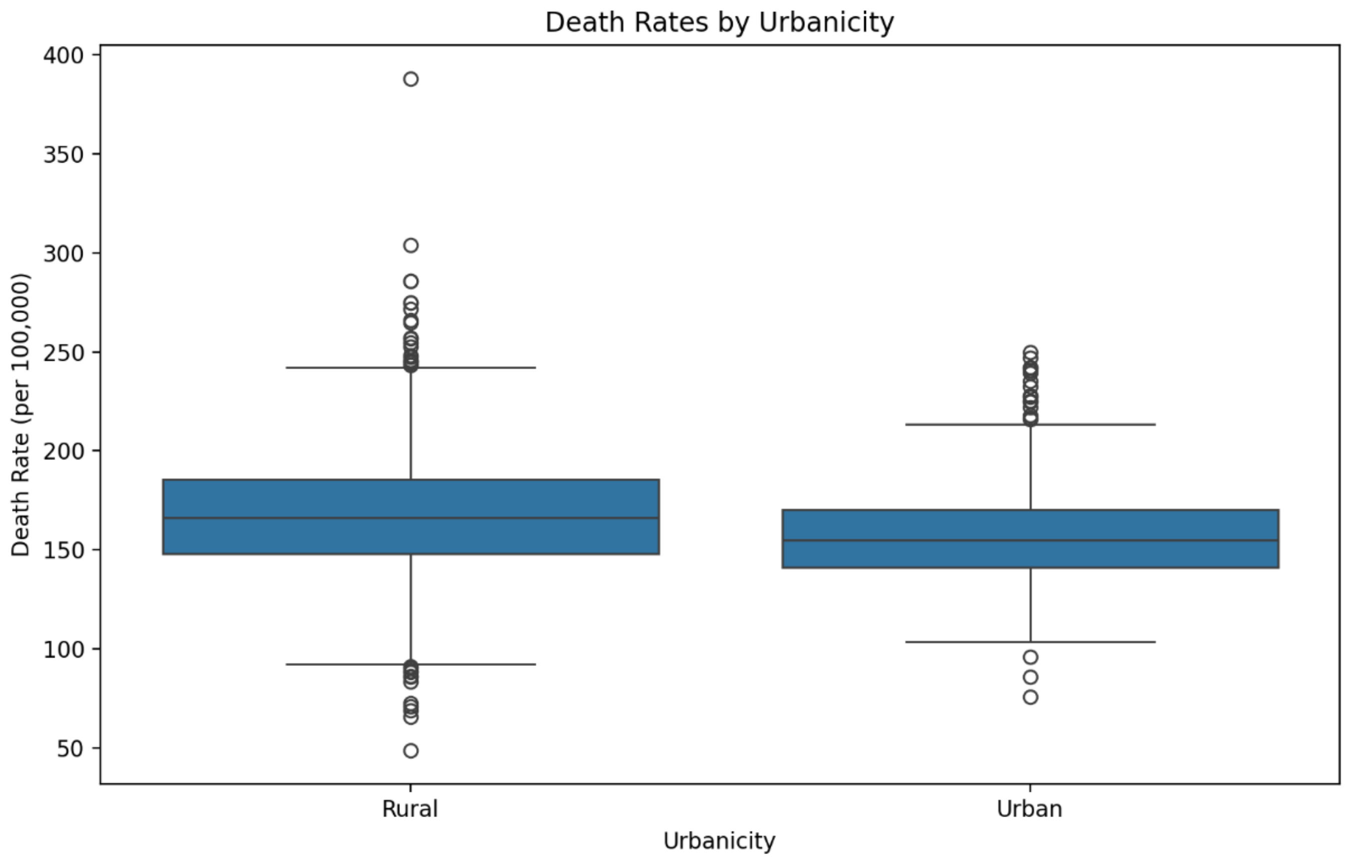
Death Rates vs. Incidence Rates



Exploratory Data Analysis: Urban vs. Rural Death Rates

An analysis comparing death rates between rural and urban counties. Showed Rural counties had higher median death rates and more variability in death rates compared to urban counties.

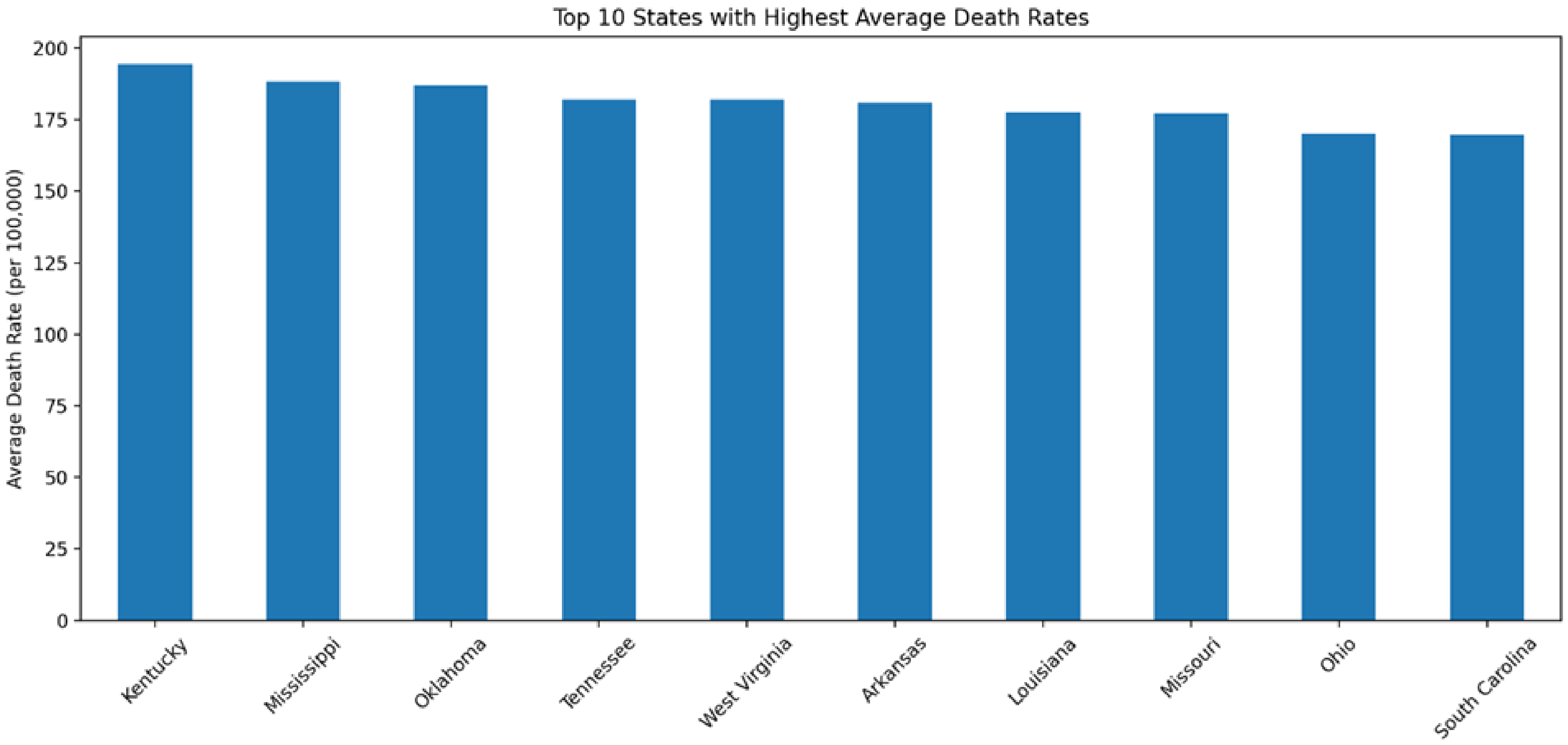
Urban vs. Rural Death Rates

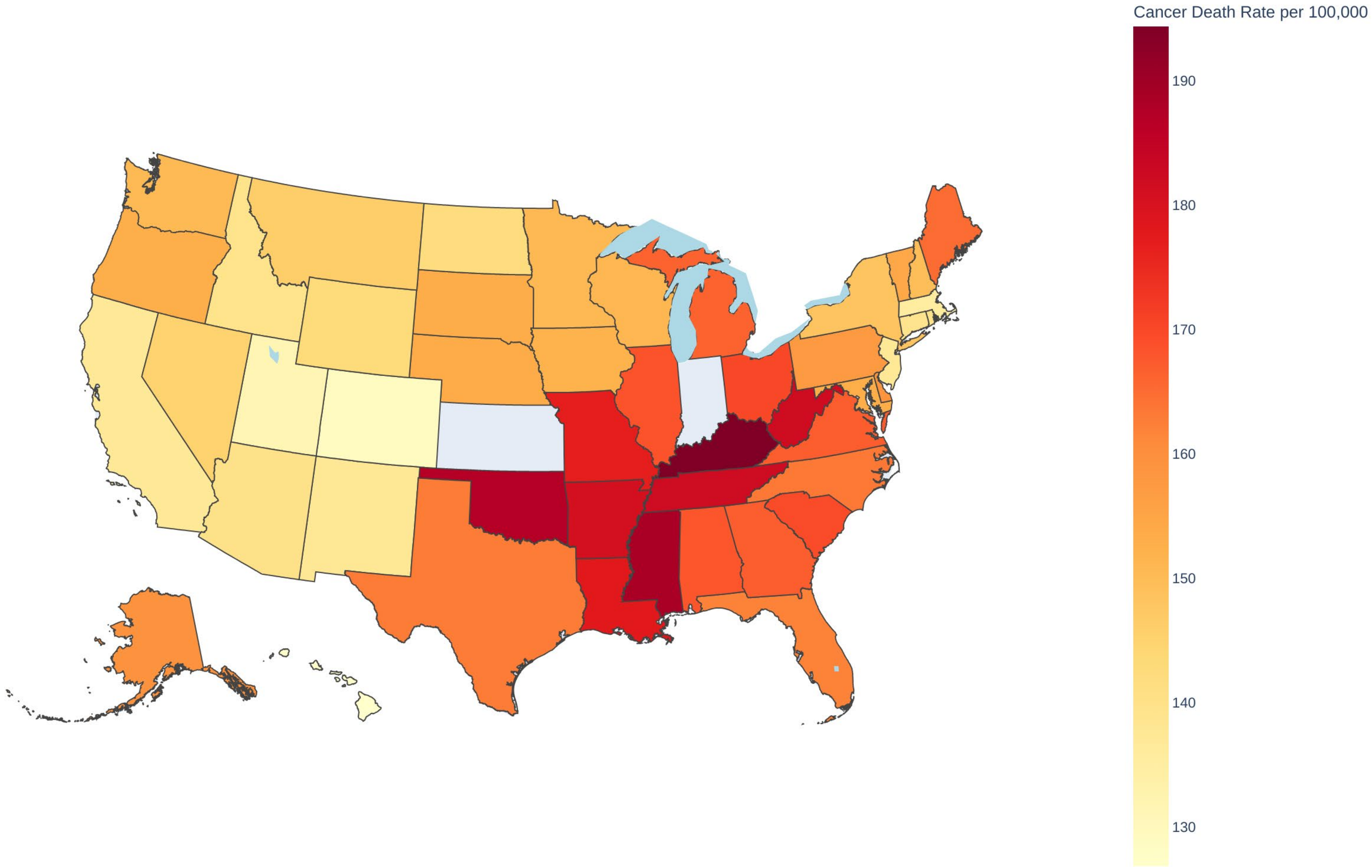


States with the Highest Average Death Rates

A bar chart was generated to identify states with the highest average cancer death rates. Kentucky, Mississippi, and West Virginia ranked highest.

Visual: States with the Highest Average Death Rates




Cancer Death Rate per 100,000
190
180
170
160
150
140
130

Summary Statistics and Key Observations

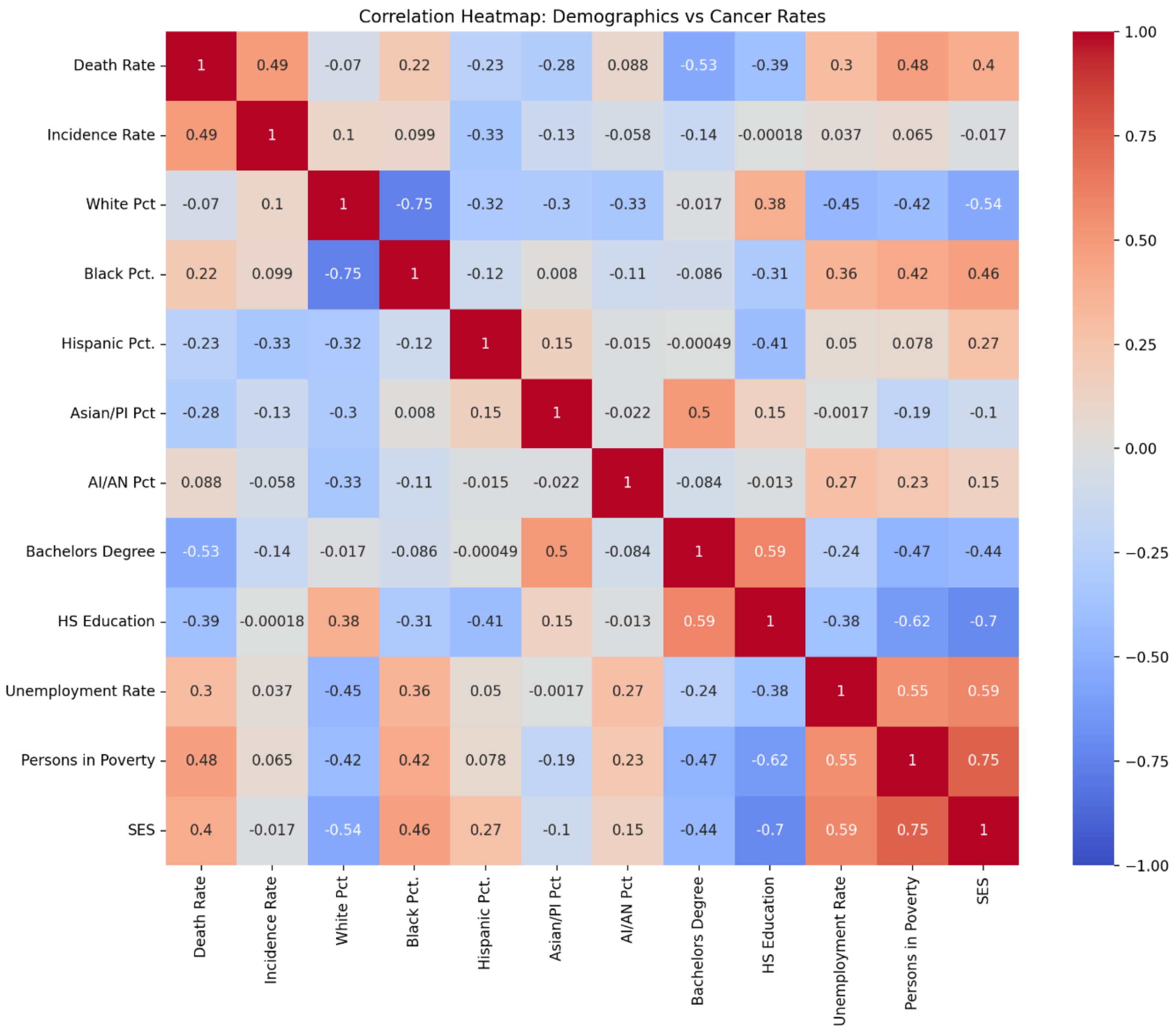
Summary statistics reveal key averages and variations in cancer-related variables, such as:

- Mean Death Rate: 163.06 per 100,000 population
- Mean Incidence Rate: 455.65 per 100,000 population
- Top counties identified by highest death and incidence rates include Union County, Florida.

Correlation Analysis

- A heatmap was generated to visualize correlations between demographic factors and cancer rates.
- Notable correlations include:
- Poverty and Death Rate: r = 0.48
- Education (Bachelor's Degree) and Death Rate: r = -0.53

Demographic Correlation Heatmap



Correlation Heatmap: Demographics vs Cancer Rates

	Death Rate	Incidence Rate	White Pct	Black Pct.	Hispanic Pct.	Asian/PI Pct	AI/AN Pct	Bachelors Degree	HS Education	Unemployment Rate	Persons in Poverty	SES
Death Rate	1	0.49	-0.07	0.22	-0.23	-0.28	0.088	-0.53	-0.39	0.3	0.48	0.4
Incidence Rate	0.49	1	0.1	0.099	-0.33	-0.13	-0.058	-0.14	-0.00018	0.037	0.065	-0.017
White Pct	-0.07	0.1	1	-0.75	-0.32	-0.3	-0.33	-0.017	0.38	-0.45	-0.42	-0.54
Black Pct.	0.22	0.099	-0.75	1	-0.12	0.008	-0.11	-0.086	-0.31	0.36	0.42	0.46
Hispanic Pct.	-0.23	-0.33	-0.32	-0.12	1	0.15	-0.015	-0.00049	-0.41	0.05	0.078	0.27
Asian/PI Pct	-0.28	-0.13	-0.3	0.008	0.15	1	-0.022	0.5	0.15	-0.0017	-0.19	-0.1
AI/AN Pct	0.088	-0.058	-0.33	-0.11	-0.015	-0.022	1	-0.084	-0.013	0.27	0.23	0.15
Bachelors Degree	-0.53	-0.14	-0.017	-0.086	-0.00049	0.5	-0.084	1	0.59	-0.24	-0.47	-0.44
HS Education	-0.39	-0.00018	0.38	-0.31	-0.41	0.15	-0.013	0.59	1	-0.38	-0.62	-0.7
Unemployment Rate	0.3	0.037	-0.45	0.36	0.05	-0.0017	0.27	-0.24	-0.38	1	0.55	0.59
Persons in Poverty	0.48	0.065	-0.42	0.42	0.078	-0.19	0.23	-0.47	-0.62	0.55	1	0.75
SES	0.4	-0.017	-0.54	0.46	0.27	-0.1	0.15	-0.44	-0.7	0.59	0.75	1

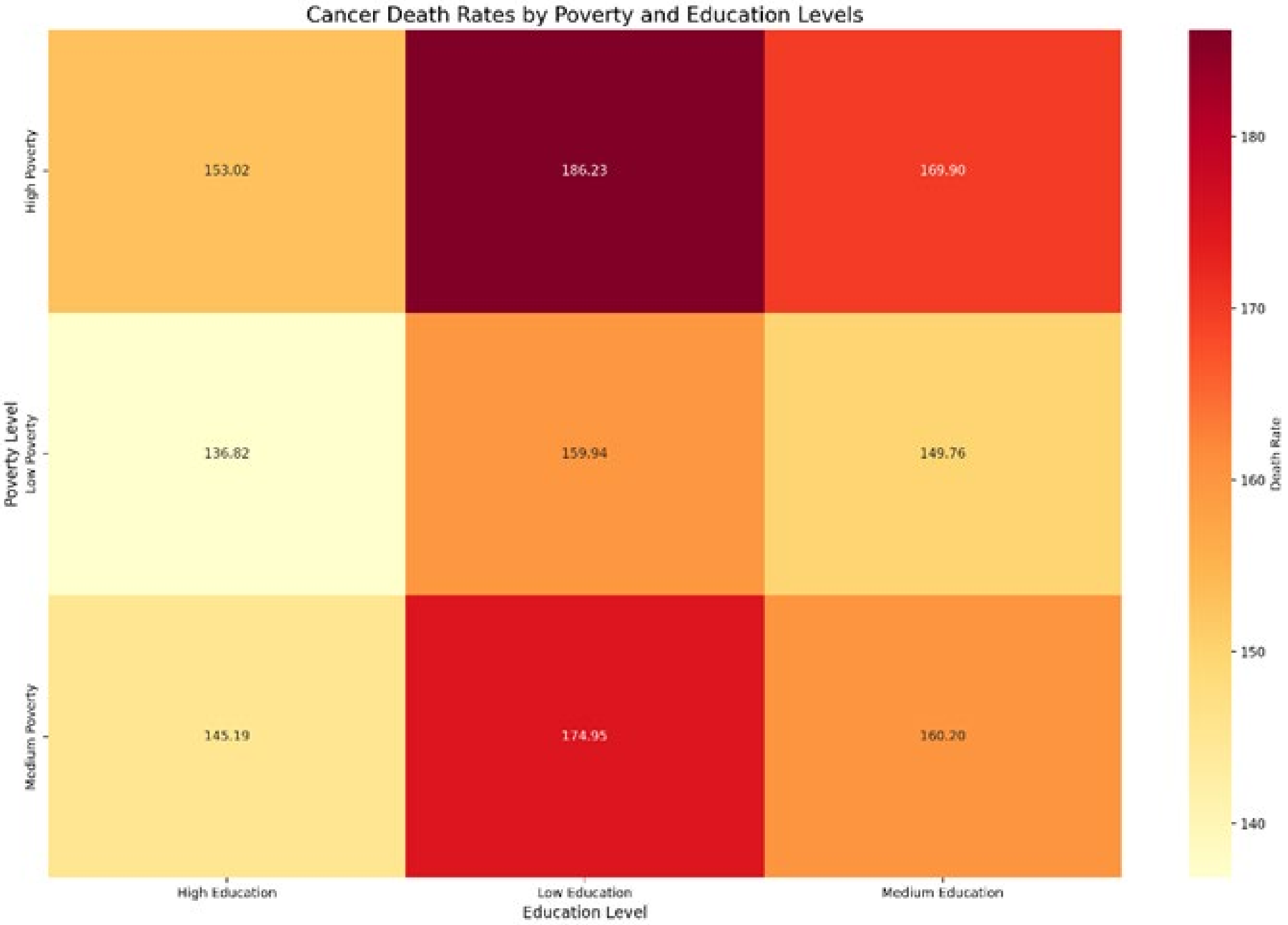
Interaction Effects: Poverty, Education, and Race

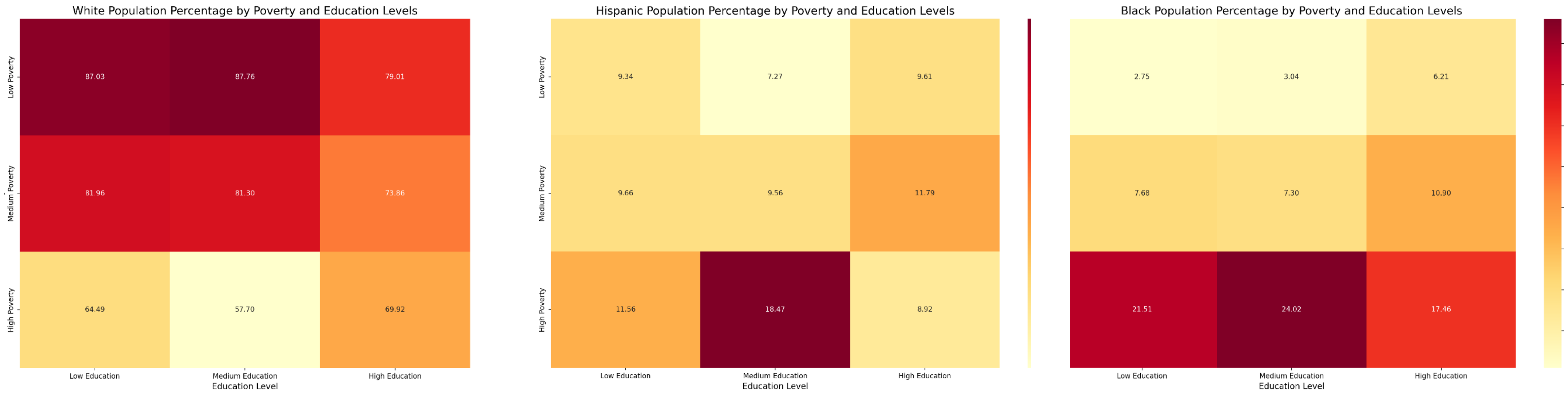
An analysis of interaction effects between poverty, education, and race revealed significant findings.

The negative effect of education on cancer death rates is stronger in counties with high poverty levels.

Racial disparities are amplified in counties with high poverty and low education levels, leading to worse cancer outcomes for minority groups.

Link: https://github.com/Akqadafi/Cancer_Rates_Analysis/blob/main/Poverty_Rate_Urbanicity_DeathRates.ipynb


Cancer Death Rates by Poverty and Education Levels
Poverty Level
High Poverty
Low Poverty
Medium Poverty
Education Level
High Education
Low Education
Medium Education
153.02
186.23
169.90
136.82
159.94
149.76
145.19
174.95
160.20
Death Rate
180
170
160
150
140



Interaction Effects of Poverty, Education, and Race

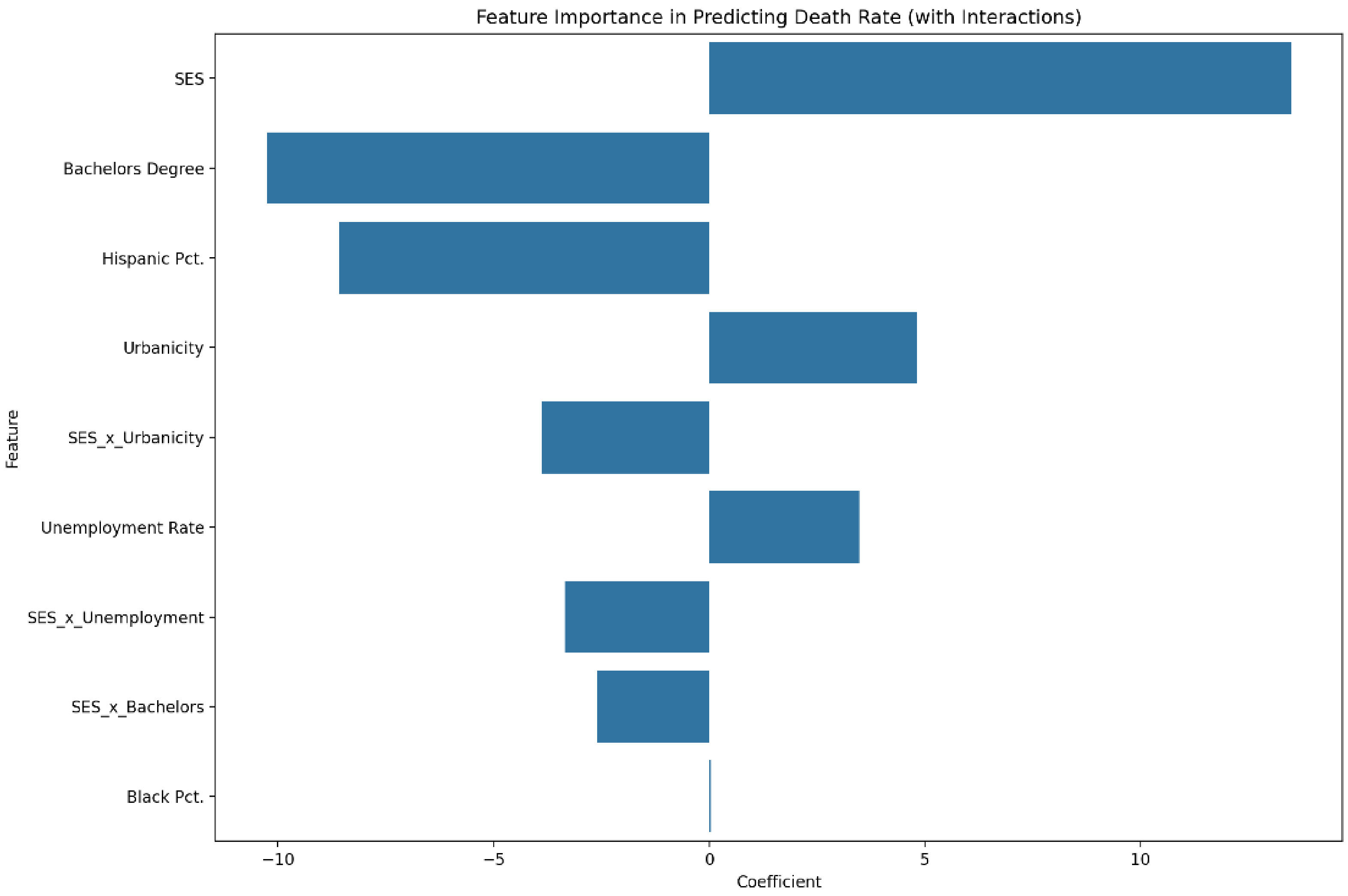
Multiple Regression Analysis

A multiple regression model was applied to assess the impact of SES, education, urbanicity, and race on cancer death rates.

- Key results include an R^2 value of 0.4340, with significant predictors including:
- SES: Positive relationship
- Education (Bachelor's Degree): Negative relationship

https://github.com/Akqadafi/Cancer_Rates_Analysis/blob/main/Multiple_Regression_Death_Rate.ipynb

Feature Importance Predicting Death Rates

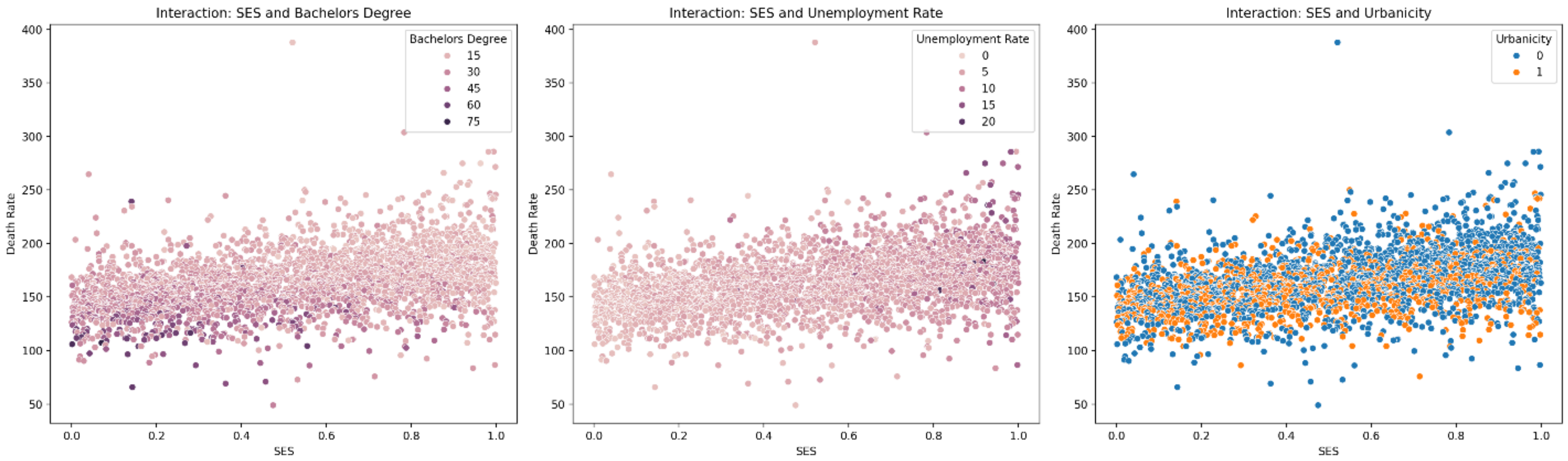


Regression Analysis - Interaction Terms

Interaction terms were included in the regression model to better understand the combined effects of:

- Poverty and Education: Strong interaction leading to higher death rates in high-poverty, low-education counties.
- Race and Poverty: Significant interaction indicating that race exacerbates the effect of poverty on cancer outcomes.

Interaction Terms in Regression

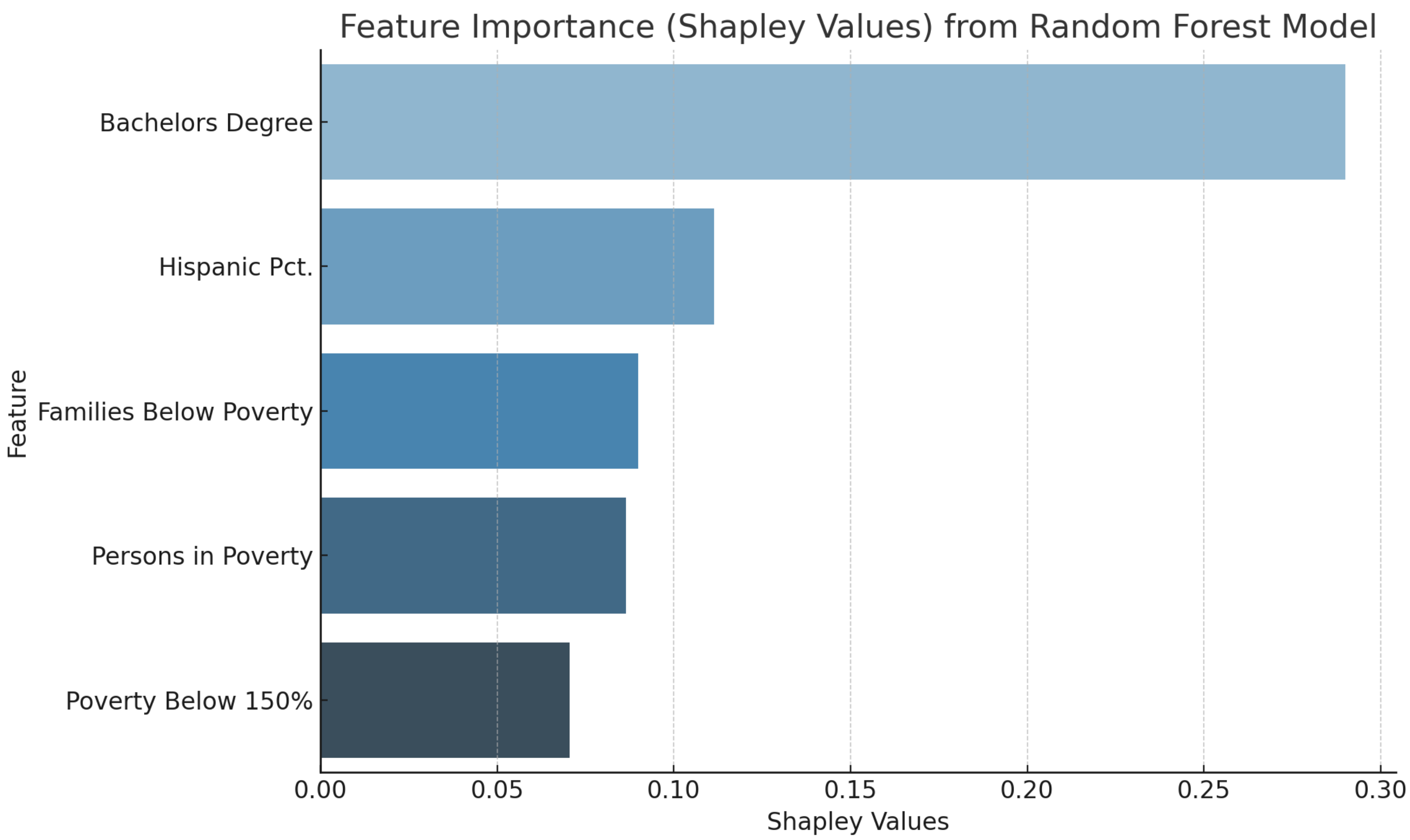


Machine Learning Analysis

A Random Forest model was applied to predict death rates yielding an R^2 of 0.4762 for death rates.

- Key predictors for Death Rates identified include:
  - Bachelor's Degree: Most important predictor

https://github.com/Akqadafi/Cancer_Rates_Analysis/blob/main/MachineLearning_DR_SR.ipynb



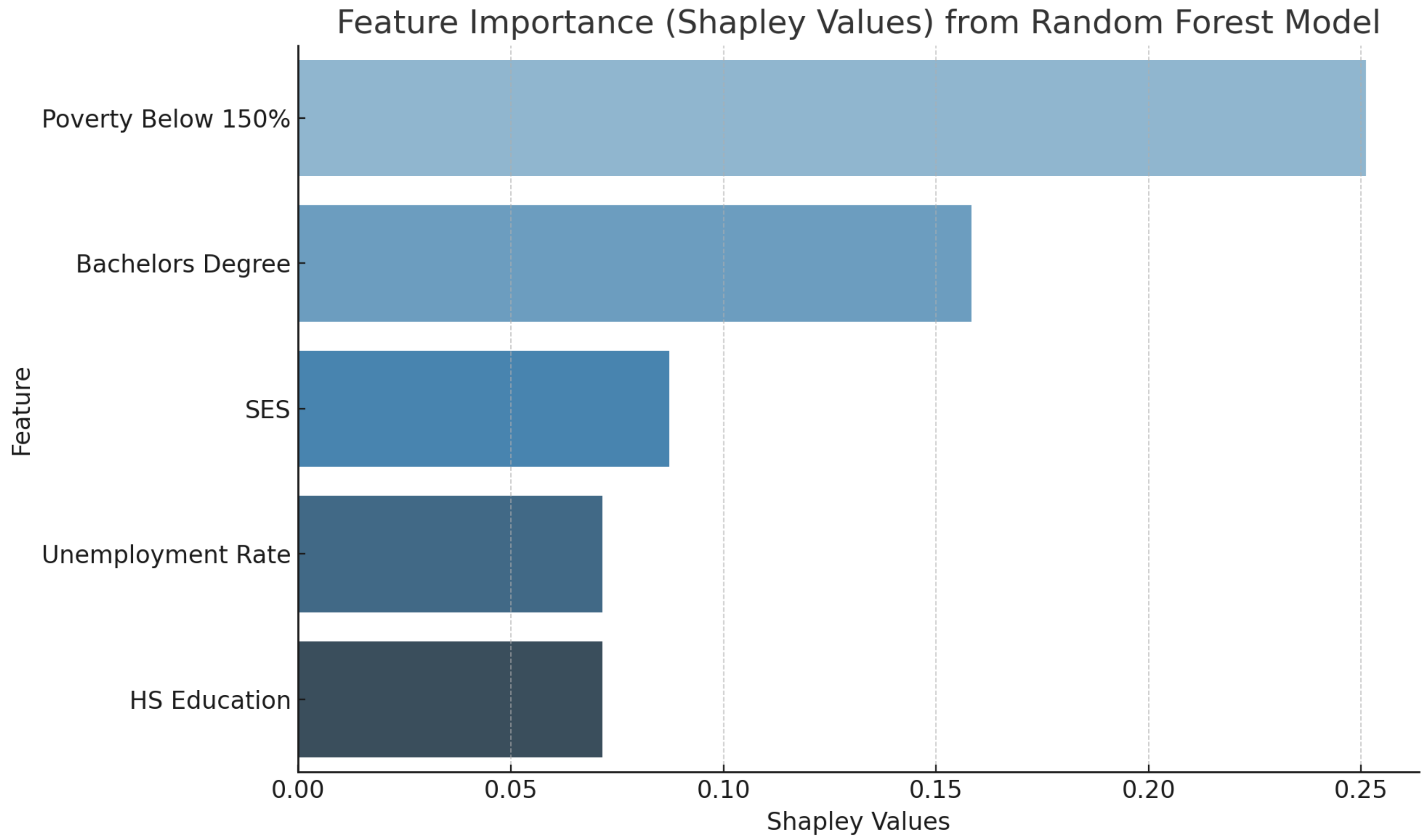
Feature Importance in Death Rates

Machine Learning Analysis – Survival Rates

Cross-validation was performed to validate the Random Forest model's predictive power:

- The model demonstrated stability across different folds with a mean R^2 value of 0.2516
- The model's performance was consistent across various subsets of the data, confirming its robustness.
- Extreme poverty proved to the most robust feature related to Survival Rate.

Feature Importance for Survival Rates



Implications and Next Steps

The analysis indicates that education, poverty, and race are critical factors in determining cancer outcomes.

Next steps include:

- Refining models to include additional variables and explore non-linear relationships.
- Developing targeted public health interventions based on these findings.

Conclusion

Summary of Key Findings:

- Education and poverty are the most significant determinants of cancer mortality and survival.
- Racial disparities are exacerbated in high-poverty, low-education areas.
- A multifaceted approach addressing these factors is essential for reducing disparities in cancer outcomes.

Call to Action

Stakeholders and policymakers must collaborate to:

- Implement targeted, evidence-based strategies to address the unique needs of different communities.
- Prioritize interventions in high-risk areas with high poverty and low education levels.
- Continue research to further understand the complex interactions between socioeconomic factors and cancer outcomes.

Thank You!

For further information, please contact:

Ahmad Qadafi
akqadafi@gmail.com
773-954-1565