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:

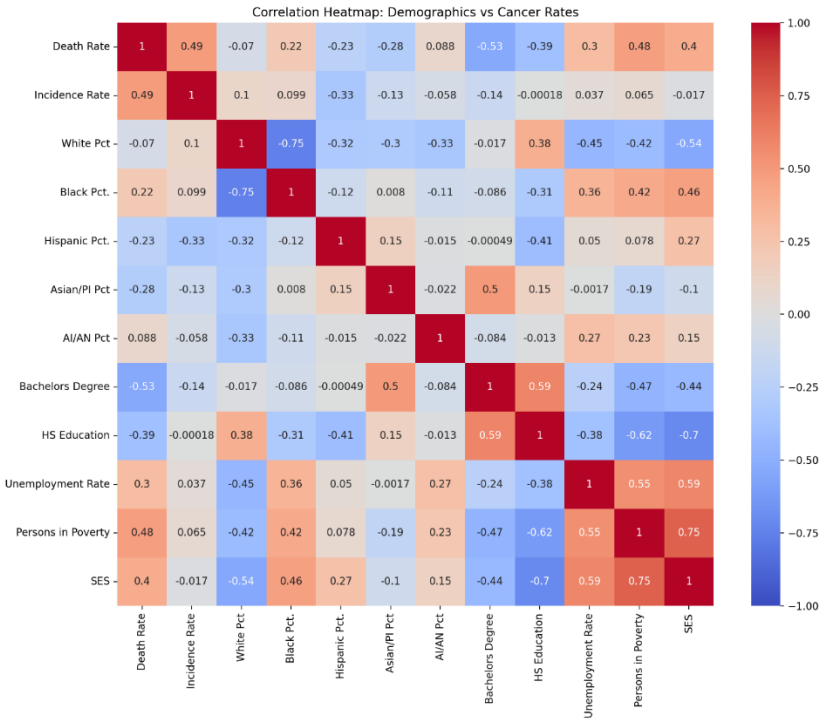
- The 'SES' (Socioeconomic Status) column contained non-numeric values labeled as "data not available." These values were replaced with NaN (Not a Number) to ensure consistency in the dataset.
- All relevant columns were then converted to numeric data types, using coercion to handle any remaining non-numeric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

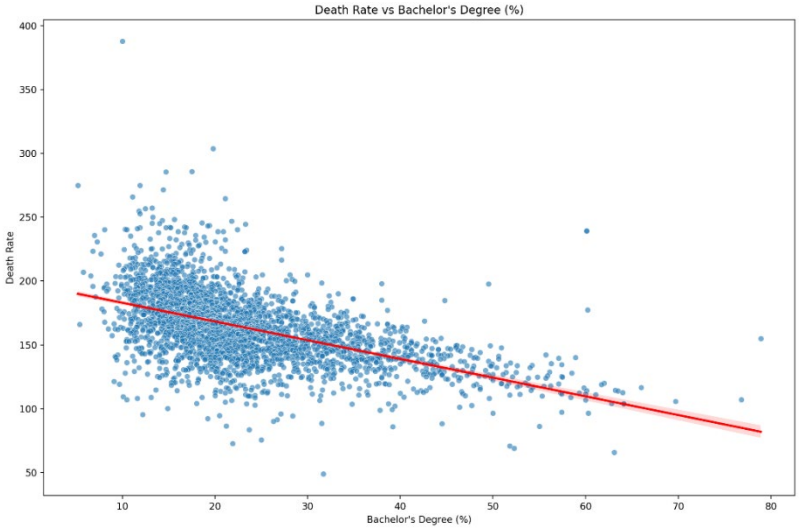
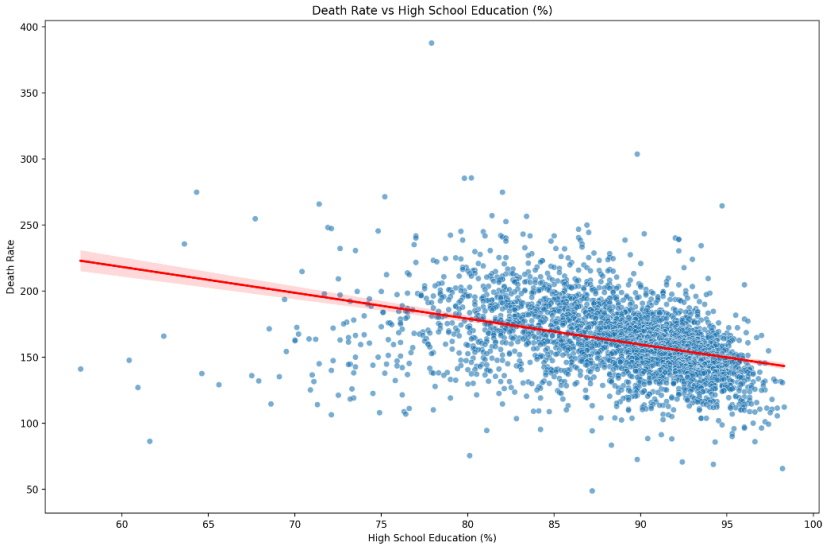


Figure 2: Death Rate vs High School Education



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

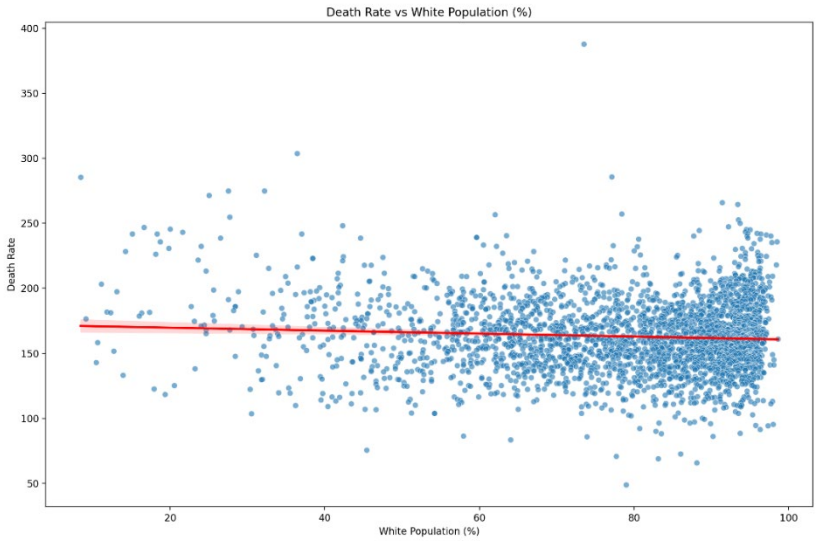


Figure 4: Death Rate vs Black Population

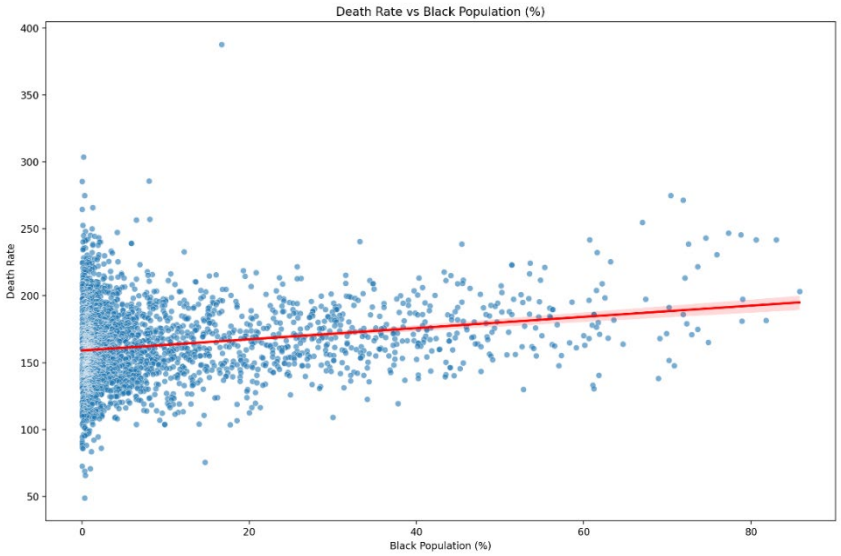
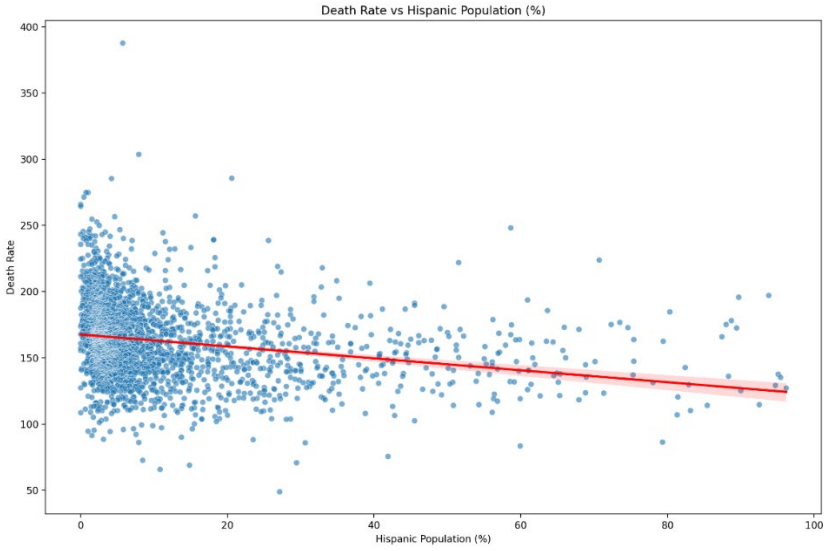


Figure 5: Death Rate vs Hispanic Population



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

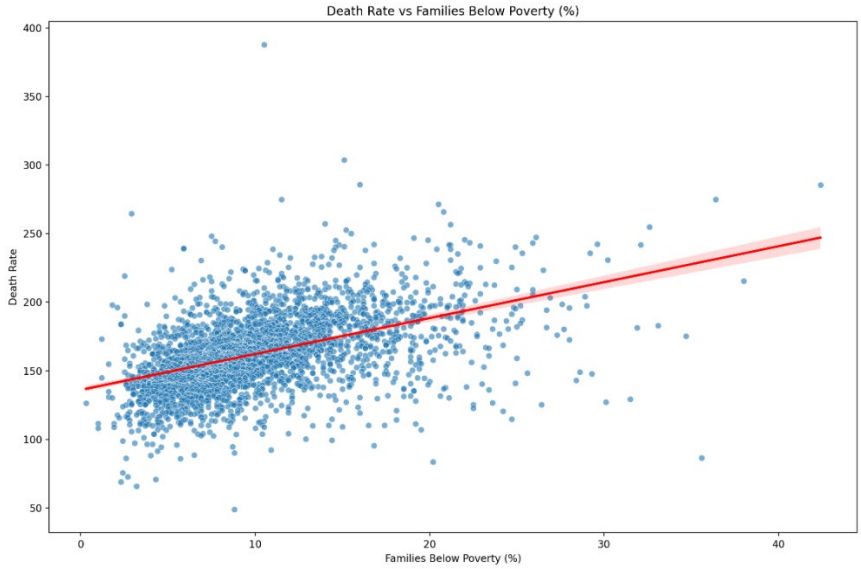


Figure 7: Death Rates vs. Below 150% Poverty Line

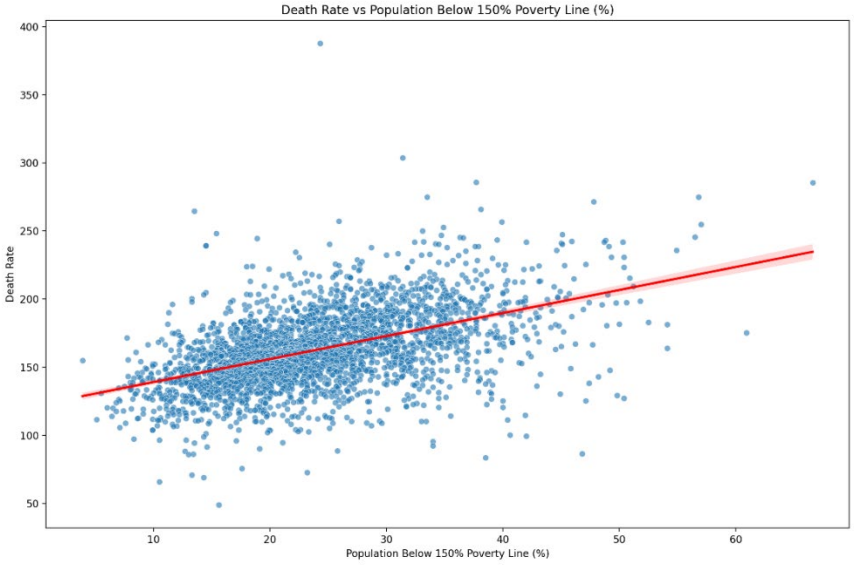
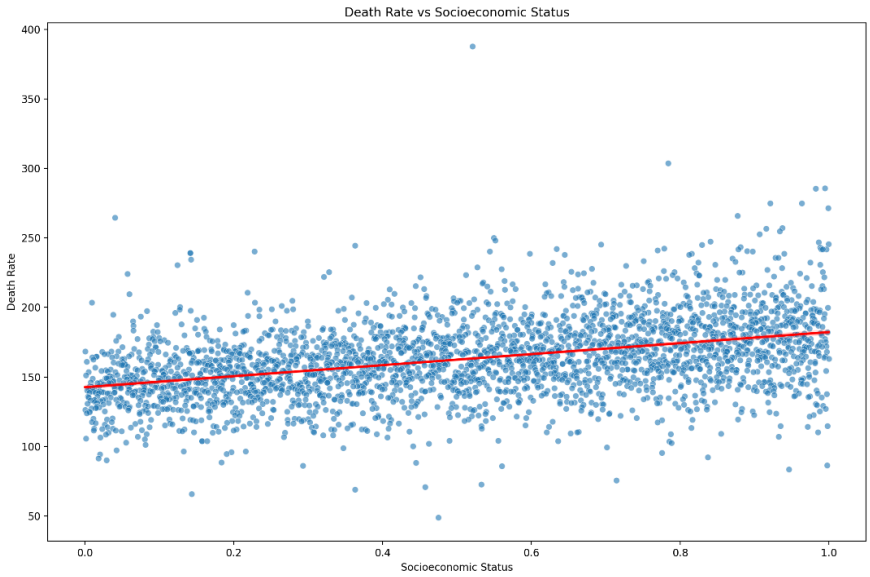


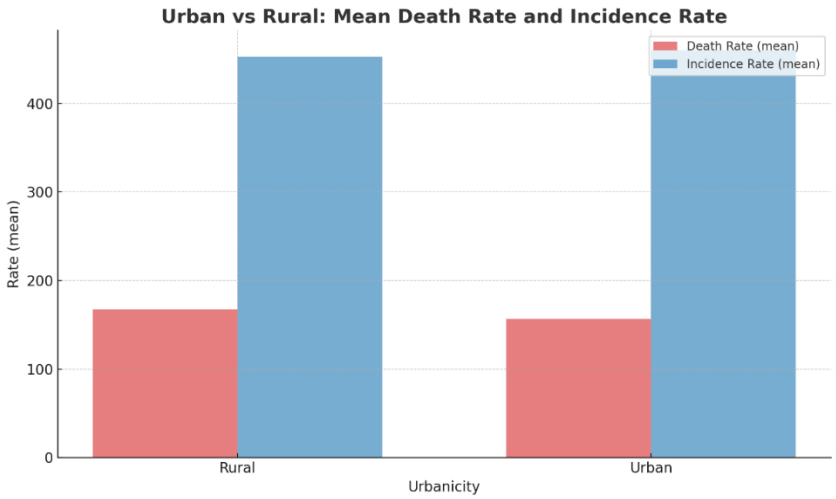
Figure 8: Death Rates vs. Socioeconomic Status



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.