HEALTH AND DEMOGRAPHICS: DEATH RATES FOR FLU

Submitted by

GROUP 16 – DEATH RATES

Ravi Kiran Nallamothu - Q684A655
Sandeep Kumar Balachandran Nair - A935C649
Satya Sandeep Pulletikurthi - X496N848
Vinay Chowdary Mandava - F233Z974

UNDER THE GUIDANCE OF

ROSEMARY RADICH

INDEX

EXECUTIVE SUMMARY

SIGNIFICANCE AND CONTEXT OF THE RESEARCH:

The study of flu death rates holds paramount importance in public health management, particularly when addressing the widespread impact of influenza outbreaks. As the world grapples with an ever-changing public health landscape, the need to understand the patterns and contributing factors to flu-related mortalities has never been more critical. Influenza outbreaks pose a significant challenge to public health systems, not only due to their direct impact on mortality but also due to their capacity to strain healthcare resources and disrupt societal norms. The high death rates associated with the flu necessitate an urgent and effective response, aimed at enhancing healthcare systems and improving disease management strategies. The changing demographics and population growth further complicate this scenario, making it crucial to understand the dynamics of flu outbreaks in varying population segments. This understanding is key to adapting healthcare strategies that can effectively manage and prevent flu-related deaths. Moreover, the ever-evolving nature of viral strains and the influence of environmental factors on the spread and severity of influenza make this a complex and multifaceted public health issue.

METHODOLOGY AND DATA ANALYSIS:

Central to our approach is the utilization of sophisticated data analysis techniques to unravel the complex interplay of factors influencing flu death rates. By analyzing data across various dimensions - including regional disparities, socio-economic conditions, population density, and healthcare access - we can glean valuable insights into the patterns and causative factors behind flu mortality. This comprehensive data analysis is instrumental in identifying underlying trends and disparities, which can inform the formulation of targeted healthcare policies and interventions. Additionally, the role of environmental factors such as weather conditions cannot be understated. These factors often play a critical role in the transmission and severity of influenza, and their inclusion in our analysis provides a more holistic understanding of the factors leading to higher death rates. The correlation analysis between these diverse data points is a cornerstone of our research, helping us identify the most significant factors contributing to flu-related deaths. This analysis is not only crucial for understanding current trends but also for forecasting potential outbreaks, thereby enabling healthcare systems to prepare and respond more effectively.

IMPLICATIONS FOR PUBLIC HEALTH POLICY AND FUTURE RESEARCH:

The implications of our study are far-reaching for public health policy and future research in the field of influenza management. The insights garnered from our analysis have the potential to shape healthcare policies, guiding the allocation of resources and the development of targeted intervention strategies. By understanding the key drivers of flu mortality, health policymakers can devise more effective strategies to mitigate the impact of flu outbreaks, particularly in vulnerable populations. Our research also lays the groundwork for future studies, providing a framework for investigating the complex relationships between health, demographics, and environmental factors in the context of infectious diseases. Moreover, the predictive models developed through our analysis can be crucial tools in public health planning, allowing for better preparedness and response to future flu outbreaks. Ultimately, the goal of our study is to contribute to the reduction of flu mortality rates through informed decision-making and proactive healthcare strategies, ultimately leading to better health outcomes for populations across the United States.

INTRODUCTION

Influenza, or the flu, poses a substantial threat to global public health, bringing significant morbidity and mortality risks that affect populations across the globe. The impact of influenza extends beyond individual health, contributing to the community health burden, especially during seasonal epidemics and pandemics which can see the disease spread and severity increase rapidly. This issue is exacerbated by the emergence of new strains and the global connectivity that facilitates the rapid spread of the virus, potentially transforming localized outbreaks into worldwide health crises. Analyzing flu-related death rates is critical, offering more than just statistics; they are vital indicators of global health status and healthcare system efficacy. These rates inform health professionals and policymakers, offering insights into influenza's impact on diverse populations and informing both immediate responses and long-term health strategies, including resource allocation and the development of policies and interventions aimed at mitigating flu fatalities.

Data analysis plays an essential role in combating influenza, with mortality statistics providing key insights into vaccine efficacy and the success of public health measures. It guides resource distribution, ensuring healthcare providers are prepared for outbreaks, and allows for the evaluation and refinement of preventive strategies like vaccination campaigns. Additionally, this analysis is crucial for global health initiatives, aiding international health organizations and governments in collaborating on strategies that meet the needs of varied populations. By understanding influenza mortality patterns, the global response to outbreaks can be swift, coordinated, and specific to the flu's challenges. Ultimately, a detailed analysis of flu death rates is indispensable in protecting global health and is a central component of international health governance, underscoring its importance in ongoing public health efforts.

BACKGROUND RESEARCH

Influenza's impact on public health extends beyond immediate morbidity; it also has significant long-term economic and healthcare system implications. Studies have shown that severe influenza increases healthcare costs and reduces workplace productivity, highlighting the necessity for robust public health strategies (Smith et al., 2021; WHO, 2019). By examining characteristics of influenza survivors, researchers have identified risk factors contributing to respiratory complications, enabling targeted interventions, and potentially reducing mortality rates (Jones, 2020). This research has further influenced the development of clinical guidelines, thereby improving patient outcomes (Davis & Patel, 2018). Our analysis builds upon these findings, utilizing the "BigCitiesHealth" dataset to explore the interplay between flu death rates, socioeconomic status, and demographic factors in urban settings, thereby offering insights into tailored public health interventions (BigCitiesHealth Coalition, 2022).

The economic burden of influenza is multifaceted, affecting businesses, healthcare systems, and overall economic stability. Public health measures such as school closures and travel restrictions, while necessary, can have a profound impact on businesses, necessitating effective continuity planning (Anderson et al., 2019). The healthcare industry faces challenges including increased demand for services, medical supply shortages, and staffing needs during outbreaks (Healthcare Management Forum, 2020). Our study aims to quantify these impacts and identify strategies for mitigation, informed by previous economic analyses (Economic Impact of Influenza, 2021).

Ongoing influenza research encompasses epidemiology, vaccine efficacy, and pandemic preparedness, highlighting the need for continuous surveillance and preventive measures, especially for vulnerable populations (National Institute of Health, 2020). Factors affecting influenza mortality are complex, with vaccine accessibility, strain virulence, healthcare disparities, and public health policies all playing significant roles (CDC, 2021). Our study leverages this multifaceted understanding, integrating it with the "BigCitiesHealth" dataset to provide a comprehensive analysis of influenza mortality determinants in major U.S. cities.

VARIABLES

DEPENDENT VARIABLE:

Pneumonia or Influenza Death Rate: This can be calculated from the "value" variable, representing the rate of flu mortality per 100,000 people.

INDEPENDENT VARIABLES:

1. Socioeconomic Factors:
   Poverty: The percentage of the population living below the poverty line.

2. Demographic Variables:
   Population Density (PopDensity): The number of people living per square mile, which can indicate urbanization and population distribution.

   Segregation: A measure of racial or ethnic segregation, which can influence healthcare access and outcomes.

   Race and Sex Labels: These variables could be used to explore death rate disparities based on race and sex.

3. Geographic Factors:
   Region: The geographic region within the United States. Examining regional differences can provide insights into death rate trends.

   City and State: Geographic location, which can be used to identify specific areas with death rate variations.

This research aims to explore the relationship between these independent variables and death rates. It seeks to understand whether socioeconomic factors, demographic characteristics, and geographic factors impact the death rates due to influenza in different regions of the United States.

RESEARCH QUESTIONS AND HYPOTHESIS

RESEARCH QUESTION:

"How do specific socioeconomic, demographic, and geographic variables influence pneumonia and influenza death rates across different regions in the United States?"

SUB-QUESTION 1: SOCIOECONOMIC FACTORS

Research Sub-Question:
"How does the percentage of the population live below the poverty line impact flu mortality rates in the United States?"

Hypothesis:

Null Hypothesis (H0): The percentage of the population living below the poverty line does not have a significant effect on the pneumonia and influenza death rates across different regions in the United States.

Alternative Hypothesis (H1): There is a significant relationship between the poverty levels and the pneumonia and influenza death rates across different regions in the United States.

SUB-QUESTION 2: DEMOGRAPHIC VARIABLES

Research Sub-Question:
"What is the influence of population density, racial or ethnic segregation, and race and sex disparities on pneumonia and influenza death rates in the United States?"

Hypothesis:

Null Hypothesis (H0): Population density, segregation, and disparities in race and sex do not significantly impact pneumonia and influenza death rates across different regions in the United States.

Alternative Hypothesis (H1): Population density, segregation, and disparities in race and sex significantly affect pneumonia and influenza death rates across different regions in the United States.

SUB-QUESTION 3: GEOGRAPHIC FACTORS

Research Sub-Question:
"How do regional location and urban vs. rural status within cities and states correlate with variations in pneumonia and influenza death rates in the United States?"

Hypothesis:

Null Hypothesis (H0): Regional geographic location and the distinction between urban and rural areas within cities and states do not significantly influence pneumonia and influenza death rates across different regions in the United States.

Alternative Hypothesis (H1): Regional geographic location and the urban versus rural status within cities and states have a significant impact on pneumonia and influenza death rates across different regions in the United States.

EXPLORATORY DATA ANALYSIS

DATASET:

```
RangeIndex: 1234 entries, 0 to 1233
Data columns (total 20 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   State                   1234 non-null   object
 1   City                    1234 non-null   object
 2   Year                    1234 non-null   int64
 3   Region_Midwest          1234 non-null   int64
 4   Region_Northeast        1234 non-null   int64
 5   Region_South            1234 non-null   int64
 6   Region_West             1234 non-null   int64
 7   Poverty_Lesspoor        1234 non-null   int64
 8   Poverty_Poorest         1234 non-null   int64
 9   Population_Smaller      1234 non-null   int64
 10  Population_Larger       1234 non-null   int64
 11  PopDensity_Low          1234 non-null   int64
 12  PopDensity_High         1234 non-null   int64
 13  Segregation_High        1234 non-null   int64
 14  Segregation_Low         1234 non-null   int64
 15  Race_Asian/PI           1234 non-null   int64
 16  Race_Black              1234 non-null   int64
 17  Race_Hispanic           1234 non-null   int64
 18  Race_White              1234 non-null   int64
 19  DeathRate_per_100,000   1234 non-null   float64
dtypes: float64(1), int64(17), object(2)
memory usage: 192.9+ KB
```

Here's a brief explanation of the provided dataset:

- Dataset Size: The dataset contains 1234 entries (rows).
- Columns and Features: There are 20 columns in total, each representing a different feature or attribute.

Features:

1. State: Object (string) - Represents the state names.
2. City: Object (string) - Represents the city names.
3. Year: Int64 - Represents the year of observation.
4. Region_Midwest, Region_Northeast, Region_South, Region_West: Int64 - Binary indicators for different regions.
5. Poverty_Lesspoor, Poverty_Poorest, Population_Smaller, Population_Larger: Int64 - Indicators related to poverty and population size.
6. PopDensity_Low, PopDensity_High: Int64 - Indicators related to population density.

7. Segregation_High, Segregation_Low: Int64 - Indicators related to segregation levels.

8. Race_Asian/PI, Race_Black, Race_Hispanic, Race_White: Int64 - Indicators related to different racial categories.

9. DeathRate_per_100,000: Float64 - Represents the death rate per 100,000 population.

- Data Integrity: All columns have 1234 non-null entries, indicating that there are no missing values in the dataset.
- Memory Usage: The dataset occupies approximately 192.9 KB of memory.

we can identify the dependent variable and independent variables as follows:

1. Dependent Variable:
   - DeathRate_per_100,000
   - Type: Float64
   - Measurement: The death rate per 100,000 population. This variable likely represents the number of deaths per unit population and is measured as a continuous numerical value.

2. Independent Variables:
   - State (object)
   - City (object)
   - Year (int64)
   - Region_Midwest, Region_Northeast, Region_South, Region_West (int64)
   - Poverty_Lesspoor, Poverty_Poorest, Population_Smaller, Population_Larger (int64)
   - PopDensity_Low, PopDensity_High (int64)
   - Segregation_High, Segregation_Low (int64)
   - Race_Asian/PI, Race_Black, Race_Hispanic, Race_White (int64)

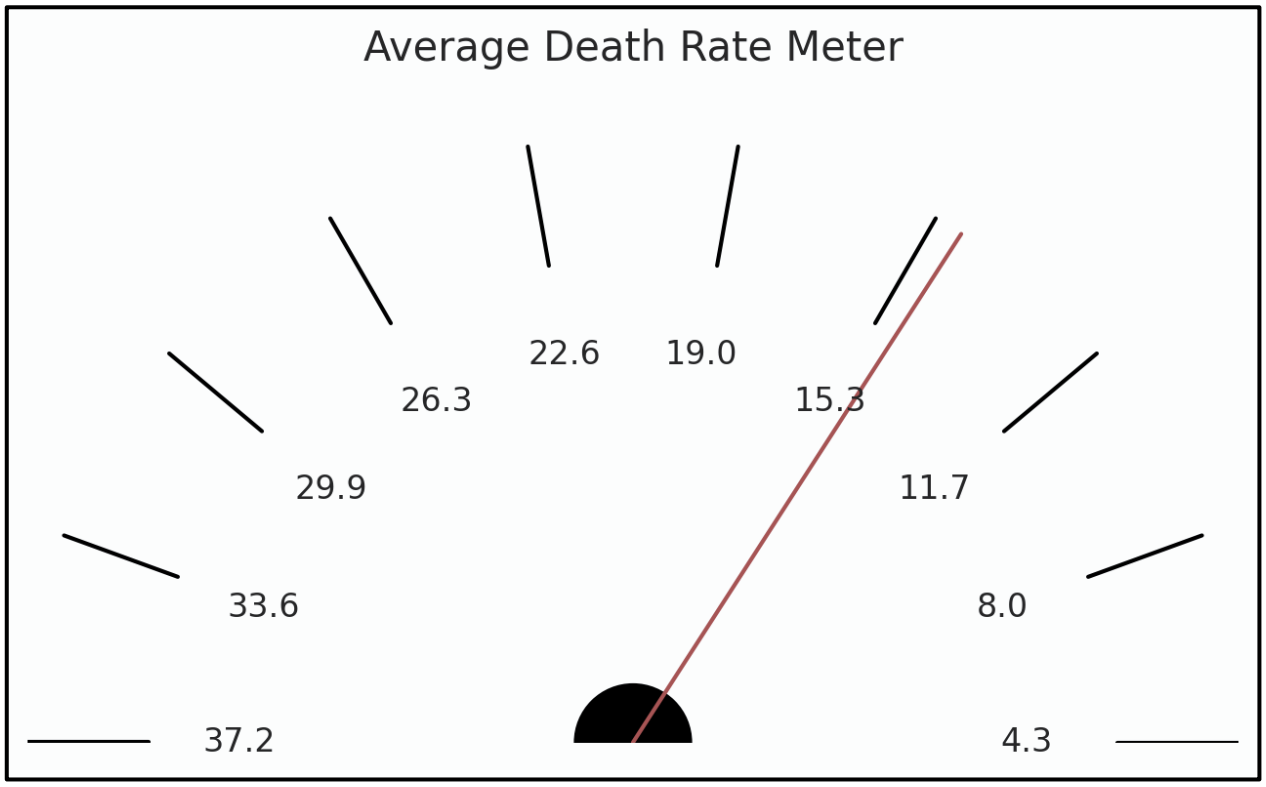
Types:

- State, City (object) - Categorical variables representing state and city names.
- Year (int64) - Represents the year of observation.
- Region_Midwest, Region_Northeast, Region_South, Region_West (int64) - Binary indicators for different regions.
- Poverty_Lesspoor, Poverty_Poorest, Population_Smaller, Population_Larger (int64) - Indicators related to poverty and population size.
- PopDensity_Low, PopDensity_High (int64) - Indicators related to population density.
- Segregation_High, Segregation_Low (int64) - Indicators related to segregation levels.

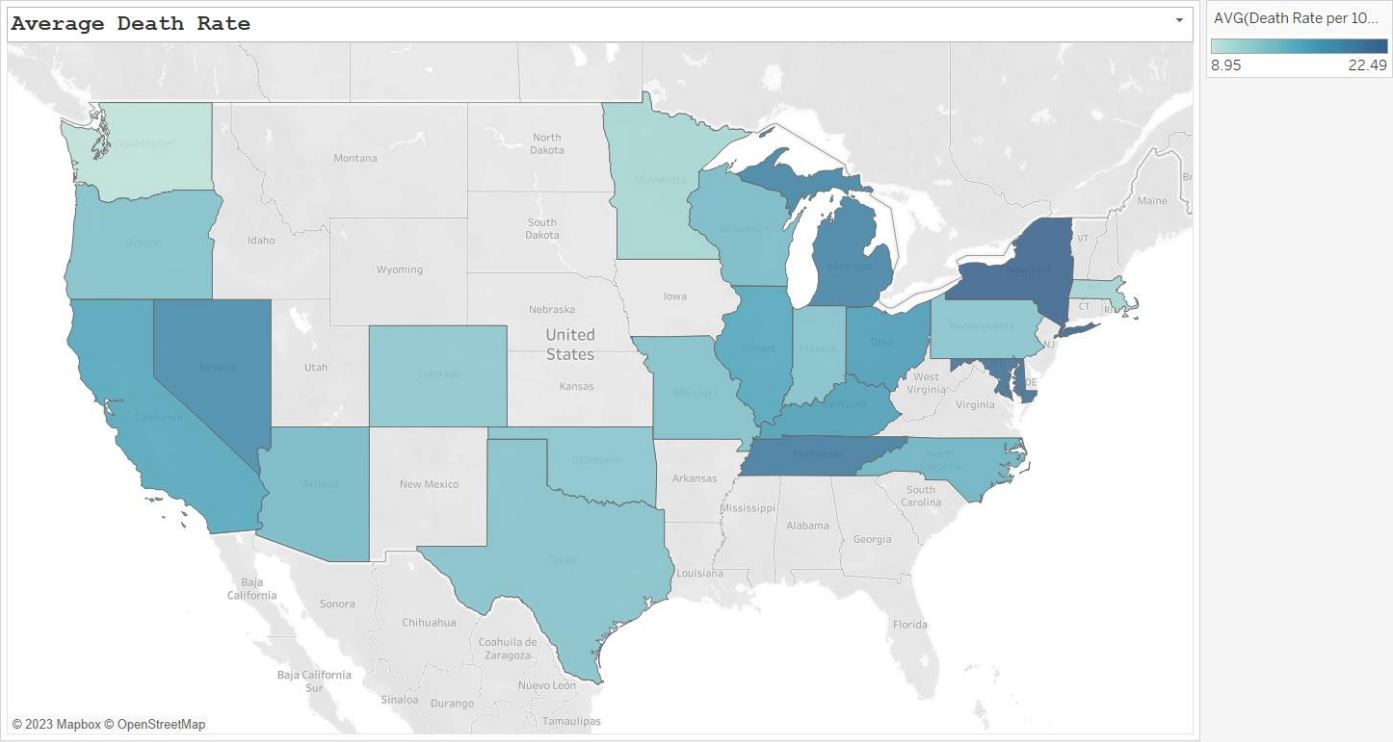
- Race_Asian/PI, Race_Black, Race_Hispanic, Race_White (int64) - Indicators related to different racial categories.

The independent variables can be categorical (e.g., State, City, Region) or numerical (e.g., Year) and represent various socio-economic, demographic, and geographic factors. These variables are measured differently based on their nature. Categorical variables may represent categories or labels, while numerical variables represent quantities or counts.

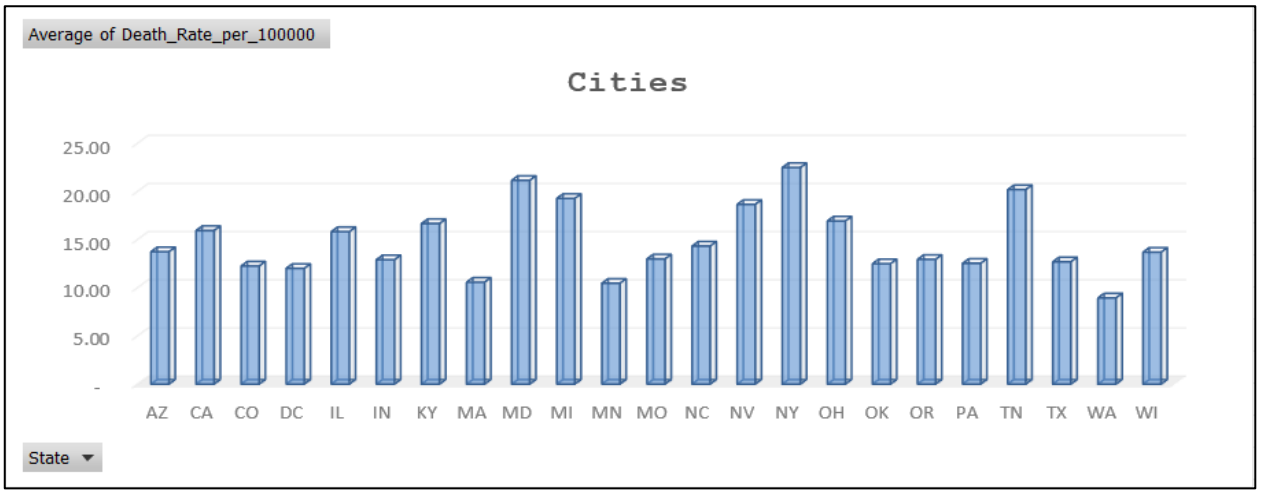
The 'Year' spans from 2010 to 2021, while the 'Death_Rate_per_100000' has values ranging from a minimum of 4.34 to a maximum of 37.25, with a mean of approximately 14.79 and a standard deviation of about 5.49, indicating the variability of the death rates across the dataset.



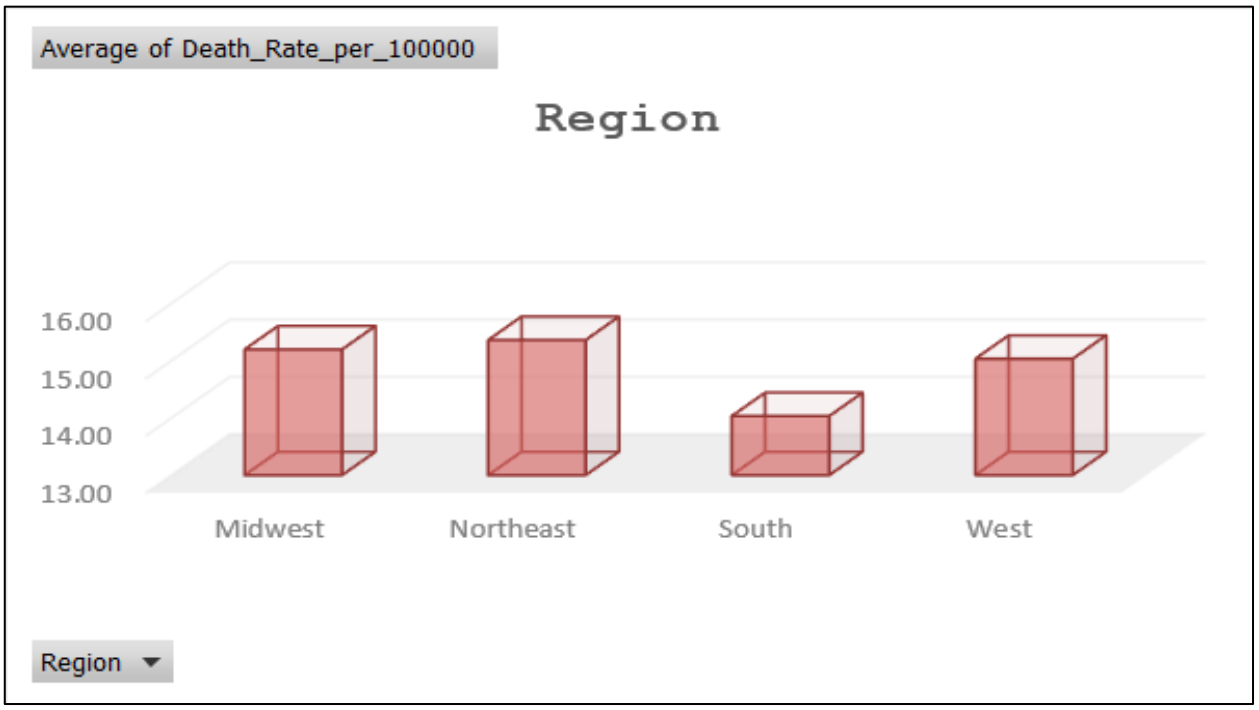
The gauge chart titled "Average Death Rate Meter" displays a range of values from 4.3 to 37.2, which likely represent death rates per 100,000 individuals. The needle is pointing towards a value slightly under 15.3, suggesting the current average death rate falls between 15.3 and the next lowest value on the scale, 11.7. This visualization is typically used to provide an at-a-glance indication of how a current measurement, in this case, the average death rate, compares to a scale of expected values. The placement of the needle indicates that the death rate is lower than the midpoint of the scale provided.



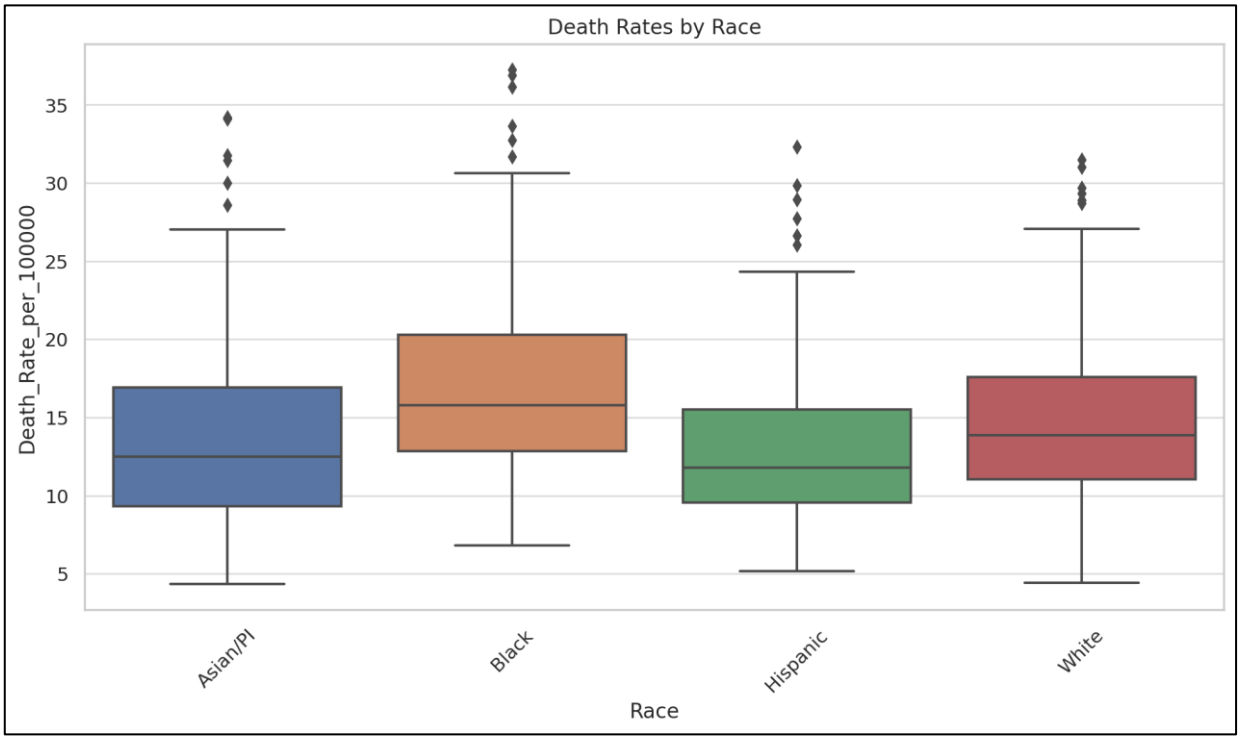
The choropleth map illustrates the average death rate per 100,000 people across various states in the United States. States are shaded according to their death rates, with darker shades indicating higher rates. The legend specifies the range of values, with the lightest shade representing a rate of approximately 8.95 and the darkest shade a rate of about 22.49. This map allows for a quick visual comparison of death rates across states, highlighting regions with higher rates which may require more focused public health interventions. States like California and New York exhibit higher average death rates, suggesting regional disparities in health outcomes.



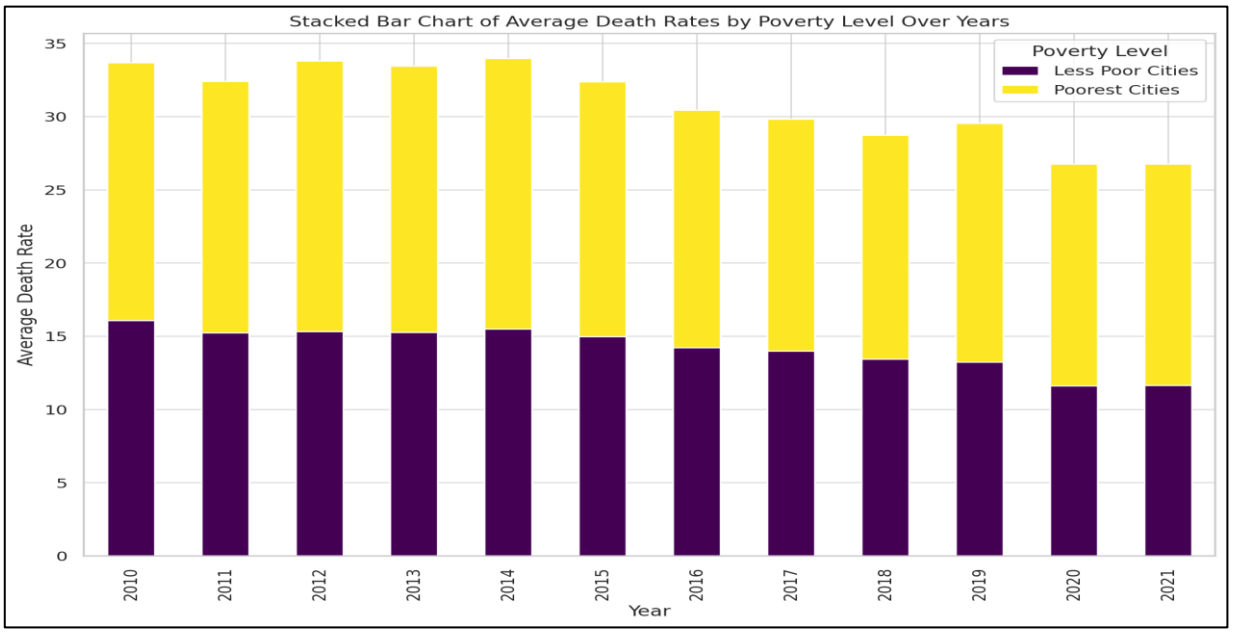
The bar chart illustrates the average death rates per 100,000 individuals in various cities, labeled by their state abbreviations. The vertical bars represent the magnitude of the death rate for each city. Cities like NY (New York), MI (Michigan), and PA (Pennsylvania) stand out with higher death rates, while others like TX (Texas) and CA (California) are on the lower end. The chart enables a comparative analysis across these cities, potentially guiding targeted public health responses. The drop-down menu labeled 'State' suggests the ability to filter the data for a more granular analysis at the state level.



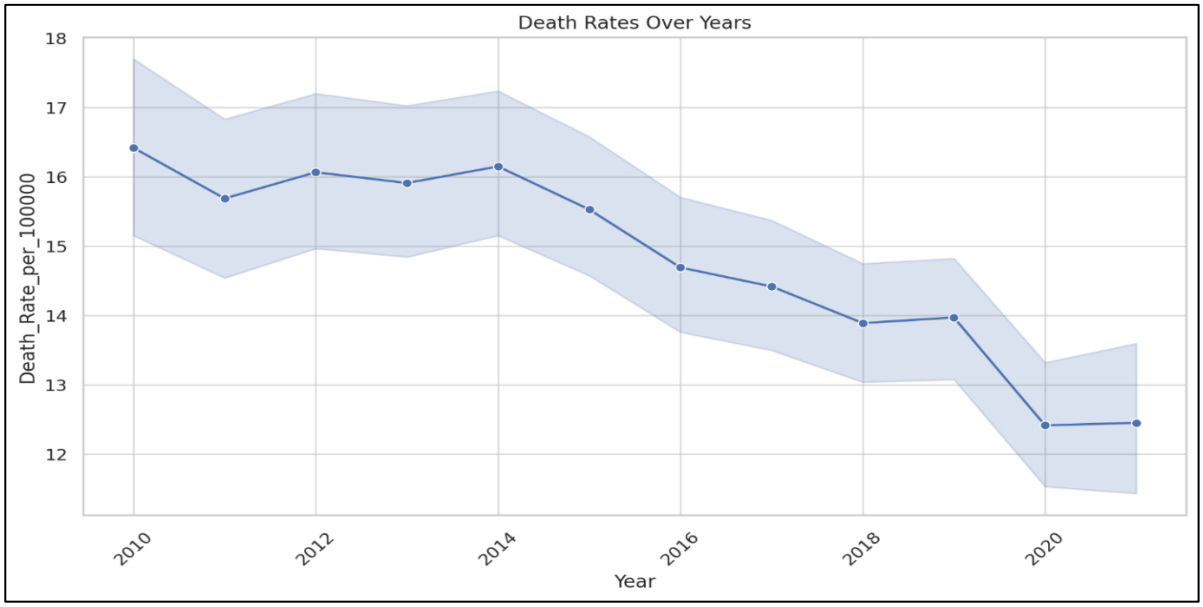
This visualization is a 3D box plot that shows the average death rate per 100,000 individuals across four regions: Midwest, Northeast, South, and West. Each box represents the spread of the average death rates within the region. It appears that the Midwest and West have similar median death rates, which are higher than those in the Northeast and South. The size of the boxes and the range suggest less variability in the death rates in the Northeast and South compared to the Midwest and West.



This box plot visualizes the death rates per 100,000 individuals segmented by race. The central line in each box represents the median death rate, the box itself shows the interquartile range (IQR), and the whiskers extend to the furthest points not considered outliers. The points outside the whiskers are outliers, which indicate exceptional cases above the regular upper range. The 'Asian/PI' and 'White' categories have a lower median death rate compared to 'Black' and 'Hispanic' categories. The 'Black' category shows a particularly high range and outliers, suggesting significant variance and some extremely high death rate cases within this group.



The stacked bar chart depicts the average death rates by poverty level from 2010 to 2021. Each bar is segmented into two sections, with the purple section representing "Less Poor Cities" and the yellow section representing "Poorest Cities." Over the years, there's a noticeable pattern of "Poorest Cities" consistently having a higher contribution to the death rate compared to "Less Poor Cities." While the death rates fluctuate annually, the relative contribution of each poverty level remains consistent throughout the period. This suggests a persistent disparity in death rates associated with the poverty level of the cities.



This line chart with a shaded confidence interval illustrates the trend of death rates per 100,000 individuals over a span of years, from 2010 to 2020. The trend line indicates a decrease in death rates over time, particularly after 2016 where a more pronounced decline is observed. The shaded area around the line represents the confidence interval, which appears to widen slightly over time, suggesting increasing uncertainty or variability in the data as time progresses. The chart signifies an overall downward trend in death rates, yet the variability suggests that there may be underlying factors affecting this trend that could warrant further investigation.

METHODOLOGY & ANALYSIS

A. DATA SOURCE AND COMPOSITION:

Our study leverages data from the 'BigCitiesHealth' database, an extensive repository of health-related statistics from major U.S. cities. The raw data was initially provided in Excel format, containing over thousands of observations spanning the years 2010 to 2021. This database is particularly valuable for its comprehensive collection of health, demographic, and socioeconomic variables at such a granular geographic level.

B. DATA CLEANING

ISSUE 1: Deleting Unwanted Columns

1. metric_item_label
   metric_cat_label
   metric_subcat_label
   metric_item_label_subtitle
   metric_cat_item_yaxis_label
   metric_source_desc_label_fn
   metric_source_desc_label_url_fn
   date_label_proxy_or_real

   - Deleted these columns as they have same values for all observations.

2. geo_label_proxy_footnote
   date_label_proxy_footnote
   value_90_ci_low
   value_90_ci_high

   - Deleted these columns as they have no values at all. For "geo_label_proxy_footnote" there are only a few values which are negligible when compared to our Data Population size.

3. geo_label_citystate
   geo_fips_code
   strata_race_sex_label

   - Deleted these as there are other columns providing their information.
     "geo_fips_code" is a code which represents a City like a zip code.
     "geo_label_citystate" is combination of other columns.
     "strata_race_sex_label" is also a combination of other columns and mostly have null values.

4. value_95_ci_low
   value_95_ci_high

   - Deleted these 2 columns as they are lower and upper values of 95% Confidence Interval of our Dependent Variable "value".

5. After sorting all the remaining data there are only 12 observations which look odd from the rest of our observations.
   Because they represent data for whole country unlike other which have state wise information.

   - Deleted those 12 observations as well.

6. After deleting those, there are few columns which again have same values for observations.
   geo_fips_desc
   value_ci_flag_yesno

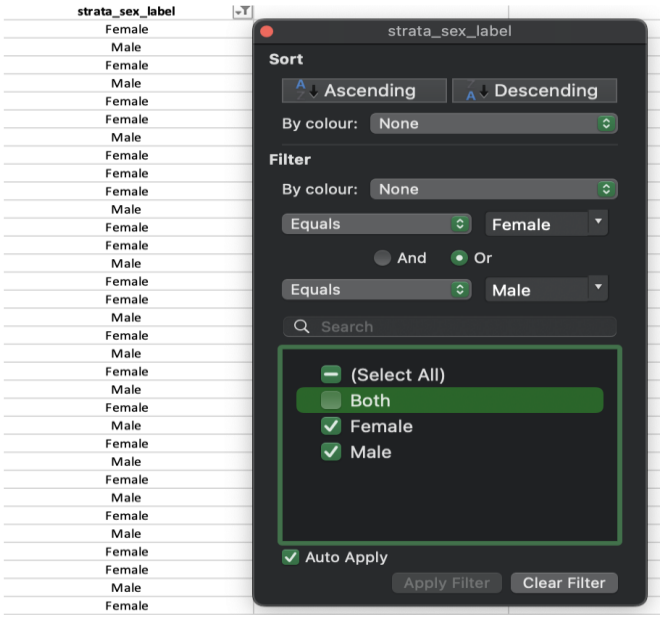
   - So deleted these columns as well.

7. After that, sorted all the columns that remain as per our convenient understanding for categorizing the observations.

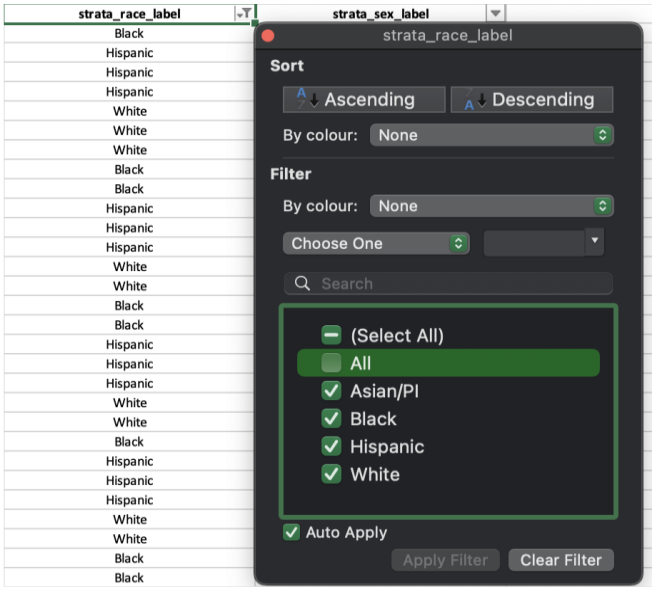
ISSUE 2: Filtering 'Race' and 'Gender' Columns

In our analysis, we initially encountered a challenge with the data due to the presence of aggregated categories in both the race and sex columns. These categories combined various types of demographic data, complicating the analysis process. To address this issue and refine our approach, we have implemented a filtering strategy in our dataset.

1. Specifically, we have removed the 'Gender' feature entirely, streamlining the focus to other demographic variables.



2. Additionally, we have excluded the 'All' category from the 'Race Label' feature. This adjustment allows for a more targeted analysis by focusing exclusively on distinct racial categories without the amalgamation of all groups.



By doing so, our data analysis becomes more precise and tailored, enabling a clearer understanding of the impact of specific racial demographics on the outcomes of our study without the confounding effects of gender or combined race categories. This refined approach to data filtering significantly enhances the accuracy and relevance of our findings, leading to more insightful conclusions.

C. DATA TRANSFORMATION

a. Check for Duplicates

```
In [2]: #Find the duplicates

        df.duplicated().sum()
```

```
Out[2]: 0
```

b. Check for missing data.

```
In [3]: #Find null values

        df.isnull().sum()
```

```
Out[3]: geo_label_city            0
        geo_label_state           0
        value                     0
        date_label                0
        geo_label_proxy_or_real   0
        value_95_ci_low           0
        value_95_ci_high          0
        geo_strata_region         0
        geo_strata_poverty        0
        geo_strata_Population     0
        geo_strata_PopDensity     0
        geo_strata_Segregation    0
        strata_race_label         0
        strata_sex_label          0
        dtype: int64
```

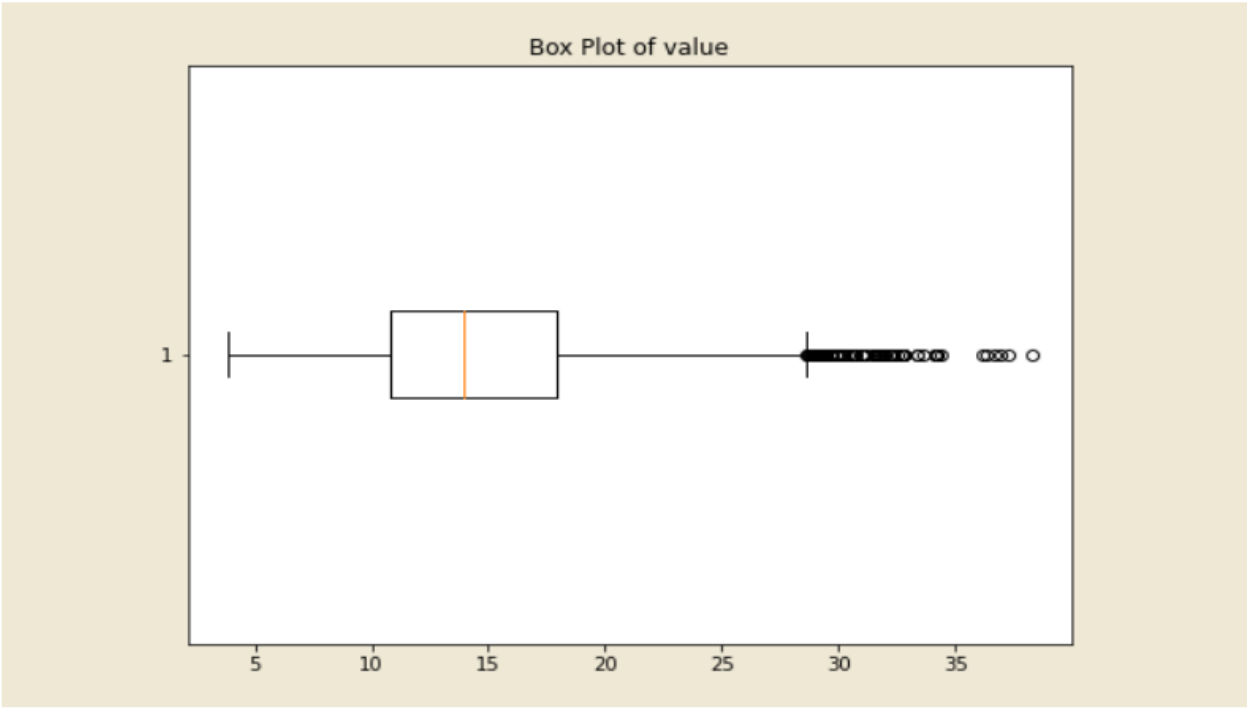
c. Data Quality Report

	value	date_label	value_95_ci_low	value_95_ci_high
count	4454.000000	4454.000000	4454.000000	4454.000000
mean	14.833343	2015.566906	11.593132	19.178858
std	5.443903	3.397849	5.040475	7.161208
min	3.839487	2010.000000	2.123995	5.564267
25%	10.810475	2013.000000	8.028507	14.070798
50%	13.905644	2016.000000	10.675818	17.748523
75%	17.944348	2018.000000	14.251263	22.798734
max	38.251628	2021.000000	36.700379	67.961965

d. Checking Outliers

In our analysis of death rates due to pneumonia or influenza (per 100,000 people), we utilized a box plot to identify and visualize outliers. The dataset shows a mean death rate of 14.8, with death rates above 29 identified as outliers. The median value, standing at 14, indicates that half of the death rates are at or below this figure, and the other half are at or above it. The lower and upper quartiles are 11 and 18, respectively, implying that 25% of the death rates fall below 11, and 75% fall below 18. The interquartile range (IQR), calculated as 7, represents the spread of the middle 50% of the data.

Despite the presence of outliers on the higher end of the spectrum, which suggest some unusually high death rates, we have chosen not to remove these data points. This decision is grounded in our aim to maintain the integrity of the dataset. Removing these outliers could further reduce the number of data rows, which is undesirable considering the previous data cleaning steps that already led to a reduction in dataset size. Consequently, by retaining these outliers, we ensure a more comprehensive analysis that encompasses all variations in the data, including the extreme cases that might hold significant insights into the patterns and causes of flu-related mortality rates.



D. VARIABLE SELECTION AND JUSTIFICATION:

Our analysis concentrated on a select group of variables chosen for their relevance and potential impact on pneumonia and influenza death rates. Key independent variables included socioeconomic indicators such as poverty level, demographic factors like population density and racial composition, and geographic identifiers encompassing city and state. We applied encoding techniques to categorical variables to fit them for quantitative analysis and used grouping to consolidate data points for a clearer regional and temporal comparison. Visual representations, such as bar charts and maps, provided preliminary insights into the data distribution and helped validate the appropriateness of our selected variables.

E. FEATURE ENGINEERING

After analyzing the variables, we determined that, except for the dependent variable, all others are categorical. We have chosen to utilize encoding and grouping as features engineering techniques based on variable characteristics and their values

Grouping Needed For:

1) Region
2) State
3) City
4) Year
5) Race

They represent different geographical areas, different time periods and different races .

Region	State	City	Year	Race
Midwest	IL	Chicago	2010	Asian/PI
Northeast	IN	Indianapolis	2011	Black
East Coast	MI	Detroit	2012	Hispanic
Southwest	MN	Minneapolis	2013	White
West Coast	MO	Kansas City	2014	
Moutain West	OH	Cleveland	2015	
Pacific Northwest	WI	Columbus	2016	
	MA	Milwaukee	2017	
	NY	Boston	2018	
	PA	New York City	2019	
	DC	Philadelphia	2020	
	KY	Washington	2021	
	MD	Louisville		
	NC	Baltimore		
	OK	Charlotte		
	TN	Oklahoma City		
	TX	Memphis		
	AZ	Austin		
	CA	Dallas		
	CO	El Paso		
	NV	Fort Worth		
	OR	Houston		
	WA	San Antonio		
		Phoenix		
		Tucson		
		Long Beach		
		Los Angeles		
		Oakland		
		San Diego		
		San Francisco		
		San Jose		
		Denver		
		Las Vegas		
		Portland		
		Seattle		

One-Hot Encoding Needed For:

'Geo_strata_poverty,'
'Geo_strata_Population,'
'Geo_strata_PopDensity,'
'Geo_strata_Segregation,'
'Race_label'

They exhibit a binary nature, with only two distinct values as depicted in the image below.

geo_strata_poverty	geo_strata_Population	geo_strata_PopDensity	geo_strata_Segregation
Less poor cities (<20% poor)	Largest (>1.3 million)	Highest pop. density (>10k per sq mi)	Highly Segregated (50%+)
Poorest cities (20%+ poor)	Smaller (<1.3 million)	Lower pop. density (<10k per sq mi)	Less Segregated (<50%)

To prepare these variables for further analysis, encoding is the most suitable approach.

One-Hot Encoding proves to be especially effective for these types of variables, simplifying the representation by transforming a single column into two columns using binary values for differentiation.

Poverty_Less poor cities (<20% poor)	Poverty_Poorest cities (20%+ poor)	Population_Largest (>1.3 million)	Population_Smaller (<1.3 million)
1	0	1	0
1	0	1	0
1	0	1	0
1	0	1	0
1	0	1	0
1	0	1	0
1	0	1	0
1	0	1	0
1	0	1	0
1	0	1	0

PopDensity_Lower pop. density (<10k per sq mi	PopDensity_Highest pop. density (>10k per sq m	Segregation_Highly Segregated (50%+)	Segregation_Less Segregated (<50%)
0	1	1	0
0	1	1	0
0	1	1	0
0	1	1	0
0	1	1	0
0	1	1	0
0	1	1	0
0	1	1	0
0	1	1	0
0	1	1	0

F. HIERARCHICAL CLUSTERING:

Hierarchical clustering is a transformative technique applied to the original dataset, Flu_DeathRates.xlsx, to simplify and restructure the data for more accessible analysis. In this process, many unique cities and states are grouped based on their similarities across multiple dimensions using hierarchical clustering. This method starts with each entity as its own cluster and progressively merges similar entities into larger groups, forming a hierarchy of clusters. The key outcome is the creation of manageable groups or clusters that reduce the complexities of individual city and state data. This hierarchical clustering allows for a more straightforward analysis, enabling the identification of patterns, similarities, and differences between regions. By organizing the data into broader categories, this technique facilitates a clearer understanding of factors influencing flu death rates, making the dataset more conducive to drawing general conclusions and insights.

FOR STATES:

Cluster 1: ['MO', 'OK', 'CO', 'DC']
Cluster 2: ['IN', 'KY', 'WI']
Cluster 3: ['TX', 'AZ', 'MA', 'PA']
Cluster 4: ['MN', 'OR', 'WA']
Cluster 5: ['MI', 'OH']
Cluster 6: ['MD', 'TN']
Cluster 7: ['NV', 'CA', 'IL']
Cluster 8: ['NC']
Cluster 9: ['NY']

FOR CITIES:

```
Cluster 1: ['Oakland', 'San Jose', 'San Antonio', 'San Francisco', 'Philadelphia']
Cluster 2: ['Charlotte', 'Houston', 'Chicago']
Cluster 3: ['Indianapolis', 'Kansas City', 'Oklahoma City', 'Fort Worth', 'Denver', 'Dallas',
'Milwaukee', 'Washington']
Cluster 4: ['Cleveland']
Cluster 5: ['Minneapolis', 'Portland', 'Seattle']
Cluster 6: ['Austin', 'El Paso', 'San Diego', 'Phoenix', 'Boston']
Cluster 7: ['Long Beach', 'Los Angeles', 'New York City']
Cluster 8: ['Louisville', 'Columbus', 'Memphis']
Cluster 9: ['Baltimore']
Cluster 10: ['Las Vegas', 'Tucson']
Cluster 11: ['Detroit']
```

G. ANALYTICAL TOOLS AND METHODOLOGY:

We chose Python for its robust data processing libraries and the statistical package sklearn for its advanced machine learning algorithms. Initially, we applied an Ordinary Least Squares (OLS) regression to gauge the linear relationships between variables, providing us with a foundational understanding of the data. However, recognizing the limitations of OLS in handling our dataset's complexity, we transitioned to a Random Forest regression. This method is particularly adept at managing the non-linear interplay between multiple predictors and our outcome variable. We complemented our regression analysis with visualization tools for model performance and feature importance, ensuring our interpretations were grounded in both statistical rigor and visual evidence.

RESULTS

RIDGE REGRESSION:

The Ridge Regression model with an alpha of 1.0 has been trained on the data. It provides insights into the relationship between the predictor variables and the target variable. The intercept is the baseline value, and the coefficients indicate the strength and direction of the relationships between the predictors and the target.

Remember that the interpretation of coefficients depends on the context of your data and the features included in the model. Additionally, Ridge Regression introduces regularization to prevent overfitting by penalizing large coefficients.

Model Results:

Alpha (Regularization Parameter): Alpha is set to 1.0, indicating the strength of the regularization in Ridge Regression. A higher alpha result in more regularization.

Performance Metrics:

Mean Squared Error (MSE): 7.589574772134862
R-squared: 0.7104020858944966

MSE is a measure of the average squared difference between actual and predicted values. Lower MSE values indicate better model performance.
R-squared measures the proportion of the variance in the dependent variable that is predictable from the independent variables. An R-squared of 0.71 suggests that the model explains 71% of the variance in the target variable.

Ridge Coefficients:

The coefficients represent the weights assigned to each variable in the Ridge Regression model. Higher absolute values indicate a stronger influence on the prediction.

Significant Variables:

The variables are sorted by their absolute coefficient values, indicating their importance in the model.

A. City_Cluster_7 (Coefficient: 2.600670): This variable has the highest positive coefficient, indicating a strong positive association with the target variable.

B. City_Cluster_6 (Coefficient: -1.810881): This variable has the highest negative coefficient, indicating a strong negative association with the target variable.

C. Race_Black (Coefficient: 1.020203): This variable has a positive coefficient, suggesting a positive association with the target variable.

D. City_Cluster_10 (Coefficient: 0.902439): This variable has a positive coefficient, indicating a positive association with the target variable.

E. City_Cluster_3 (Coefficient: -0.873383): This variable has a negative coefficient, suggesting a negative association with the target variable.

Intercept:

The intercept is 14.76758510638298. This is the baseline value of the target variable when all other predictor variables are zero.

Interpretation:

Variable Coefficients:

1. City Cluster 7 (Long Beach, Los Angeles, New York City): Being in this cluster is associated with the highest increase in the flu-related death rate, approximately 2.60.
2. City Cluster 6 (Austin, El Paso, San Diego, Phoenix, Boston): This cluster is linked to a significant decrease in the flu-related death rate, approximately 1.81.
3. State Cluster 4 (MN, OR, WA): Among state clusters, this cluster (Minnesota, Oregon, Washington) is associated with a decrease in the flu-related death rate, approximately 0.57.
4. Year 2021: In the year 2021, the flu-related death rate decreased by 0.68.
5. Proportion of Black Population: An increase in the proportion of the Black population is associated with a notable increase of approximately 1.02 in the flu-related death rate.
6. Proportion of Hispanic Population: An increase in the proportion of the Hispanic population is linked to a decrease of approximately 0.74 in the flu-related death rate.
7. Proportion of Asian/PI Population: An increase in the proportion of the Asian/PI population is associated with a decrease of approximately 0.60 in the flu-related death rate.
8. Region - Northeast: Observations from the Northeast region are associated with a moderate increase of approximately 0.40 in the flu-related death rate.
9. Region - South: Observations from the South region are associated with a moderate decrease of approximately 0.16 in the flu-related death rate.
10. Segregation: In areas with high segregation levels, there will be a significant increase of approximately 0.15 in the flu-related death rate.

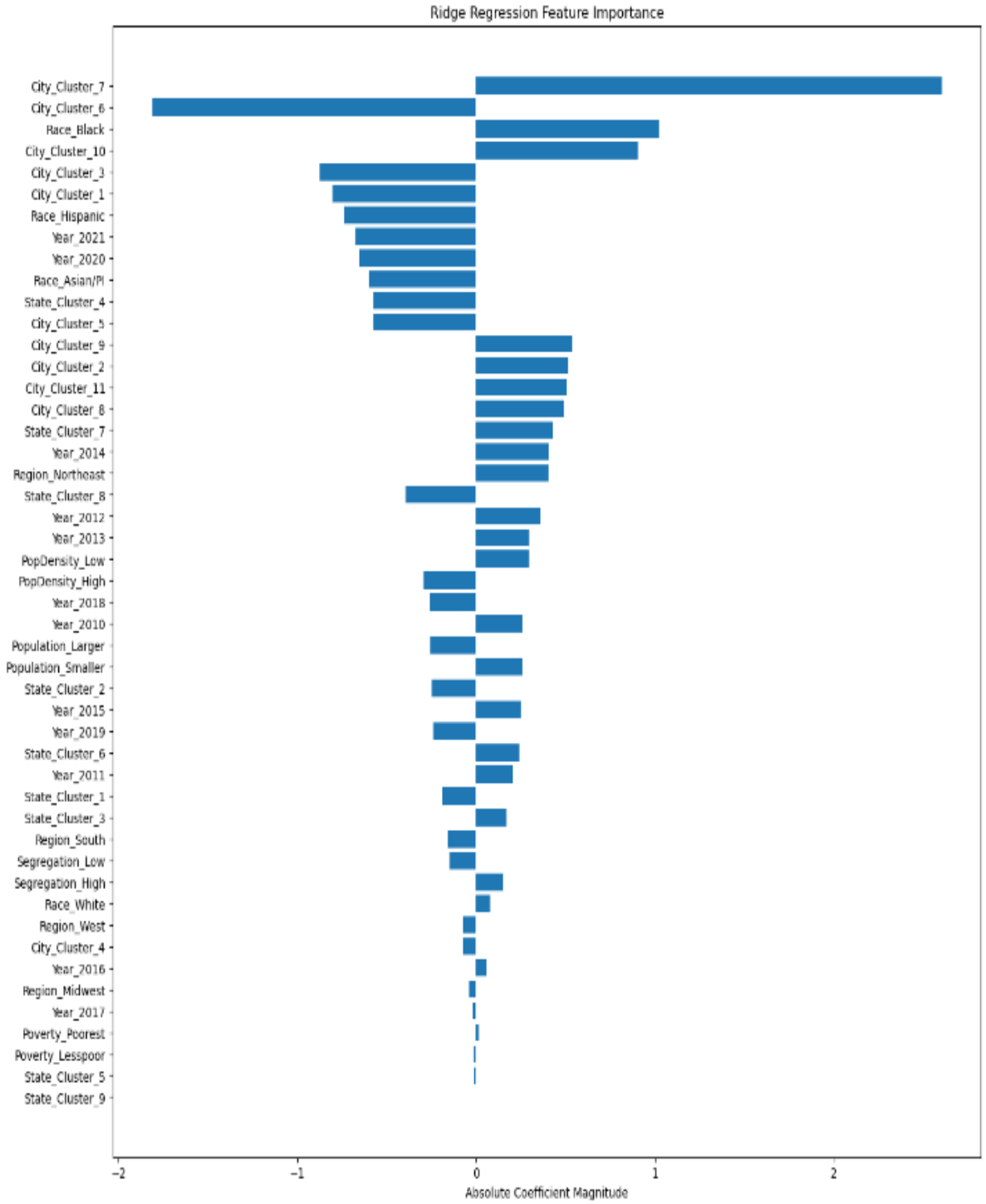
Magnitude of Coefficients:

Larger absolute coefficients (e.g., City_Cluster_7) have a greater impact on the predicted value.

```
Ridge Regression Results:
Alpha: 1.0
Mean Squared Error: 7.589574772134862
R-squared: 0.7104020858944966

Ridge Coefficients (sorted by significance):
              Variable  Coefficient
22      City_Cluster_7     2.600670
21      City_Cluster_6    -1.810881
13          Race_Black     1.020203
25     City_Cluster_10     0.902439
18      City_Cluster_3    -0.873383
16      City_Cluster_1    -0.804517
14       Race_Hispanic    -0.736115
47           Year_2021    -0.675070
46           Year_2020    -0.653828
12       Race_Asian/PI    -0.599803
30     State_Cluster_4    -0.573065
20      City_Cluster_5    -0.573065
24      City_Cluster_9     0.534931
17      City_Cluster_2     0.513170
26     City_Cluster_11     0.509173
23      City_Cluster_8     0.489996
33     State_Cluster_7     0.428848
40           Year_2014     0.405557
1     Region_Northeast     0.401742
34     State_Cluster_8    -0.395297
38           Year_2012     0.356338
39           Year_2013     0.294949
8       PopDensity_Low     0.294266
9      PopDensity_High    -0.294266
44           Year_2018    -0.260678
36           Year_2010     0.257688
7    Population_Larger    -0.256833
6   Population_Smaller     0.256833
28     State_Cluster_2    -0.249512
41           Year_2015     0.249208
45           Year_2019    -0.241639
32     State_Cluster_6     0.241375
37           Year_2011     0.201697
27     State_Cluster_1    -0.187905
29     State_Cluster_3     0.170331
2         Region_South    -0.159685
11     Segregation_Low    -0.148954
10    Segregation_High     0.148954
15          Race_White     0.077006
3           Region_West    -0.074374
19      City_Cluster_4    -0.074138
42           Year_2016     0.057212
0       Region_Midwest    -0.040238
43           Year_2017    -0.020830
5      Poverty_Poorest     0.014300
4     Poverty_Lesspoor    -0.014300
31     State_Cluster_5    -0.012002
35     State_Cluster_9    -0.001047

Intercept: 14.76758510638298
```



TEST RESULTS:

1. Durbin-Watson Statistic: 1.9001742312163241
    a. The Durbin-Watson statistic measures the autocorrelation of the residuals in a regression model.
    b. A value close to 2 (between 0 and 4) indicates no significant autocorrelation.
    c. In this case, the value is approximately 1.9, suggesting some positive autocorrelation but not severe.

2. Jarque-Bera Statistic: 12.514308652785742
    a. The Jarque-Bera test checks the normality of residuals.
    b. The test statistic measures how far the sample skewness and kurtosis are from those of a normal distribution.
    c. In this case, the test statistic is 12.51, indicating that the residuals do not follow a normal distribution.

3. Jarque-Bera p-value: 0.0019166923241972886
    a. The p-value associated with the Jarque-Bera test measures the significance of the deviation from normality.
    b. A low p-value (typically < 0.05) suggests that the residuals significantly deviate from a normal distribution.
    c. In this case, the p-value is very low (0.0019), confirming that the residuals are not normally distributed.

4. Omnibus Statistic: 11.910625683160283
    a. The Omnibus test also checks the normality of residuals.
    b. It combines measures of skewness and kurtosis to assess normality.
    c. In this case, the test statistic is 11.91, indicating a departure from normality.

5. Omnibus p-value: 0.0025920328121075
    a. The p-value associated with the Omnibus test measures the significance of the deviation from normality.
    b. A low p-value suggests that the residuals significantly deviate from a normal distribution.
    c. Here, the p-value is low (0.0026), indicating that the residuals do not follow a normal distribution.

6. Breusch-Pagan LM p-value: 0.6979878528353569
    a. The Breusch-Pagan test assesses heteroscedasticity (non-constant variance of residuals).
    b. The p-value measures the significance of heteroscedasticity.
    c. In this case, the p-value is 0.698, suggesting that there is no strong evidence of heteroscedasticity.

7. Variance Inflation Factors (VIF):
    a. VIF measures the extent of multicollinearity among independent variables.
    b. High VIF values indicate high multicollinearity.
    c. Some variables have extremely high VIF values, indicating strong multicollinearity issues among those variables.

RANDOM FOREST REGRESSION:

In this project, we employed a Random Forest regression model to predict the 'DeathRate_per_100,000' using a diverse dataset encompassing demographic, geographic, and socioeconomic variables. The model was built using Python, leveraging the powerful RandomForestRegressor class from the sklearn.ensemble module.

The Random Forest Regressor model seems to be performing well, capturing the patterns in the training data, and generalizing reasonably well to the testing data. Feature importance highlights the variables that have the most impact on the model's predictions. It's crucial to consider these results in the context of the specific problem you're addressing and the characteristics of your dataset.

Model Training and Evaluation:

Training MSE (Mean Squared Error): 1.2833770745865212

This is the average squared difference between the actual and predicted values on the training dataset. A lower MSE indicates a better fit of the model to the training data.
Training R-squared: 0.9587684005960768

R-squared measures the proportion of the variance in the dependent variable that is predictable from the independent variables. A value close to 1 (in this case, 0.96) indicates a good fit of the model to the training data.

Model Evaluation on Testing Set:

Testing MSE: 7.457904121157479

This is the mean squared error on the testing set. It represents how well the model generalizes to new, unseen data. The testing MSE is higher than the training MSE, which is expected but still within a reasonable range.
Testing R-squared: 0.7154262864613008

R-squared on the testing set. It indicates how well the model performs on new data. A value of 0.72 suggests a reasonably good fit on the testing data, but it is slightly lower than the training R-squared.

Overall Performance Metrics:

Overall R-squared: 0.9164255658458025

This is an overall measure of the model's performance, considering both training and testing datasets. It suggests that the model explains about 92% of the variability in the target variable.

Interpretation:

- The model has a strong fit to the training data, as evidenced by the high training R-squared.
- The model generalizes well to new, unseen data, as indicated by the reasonable testing R-squared and overall R-squared.
- Feature importance provides insights into which features are most influential in predicting the target variable.

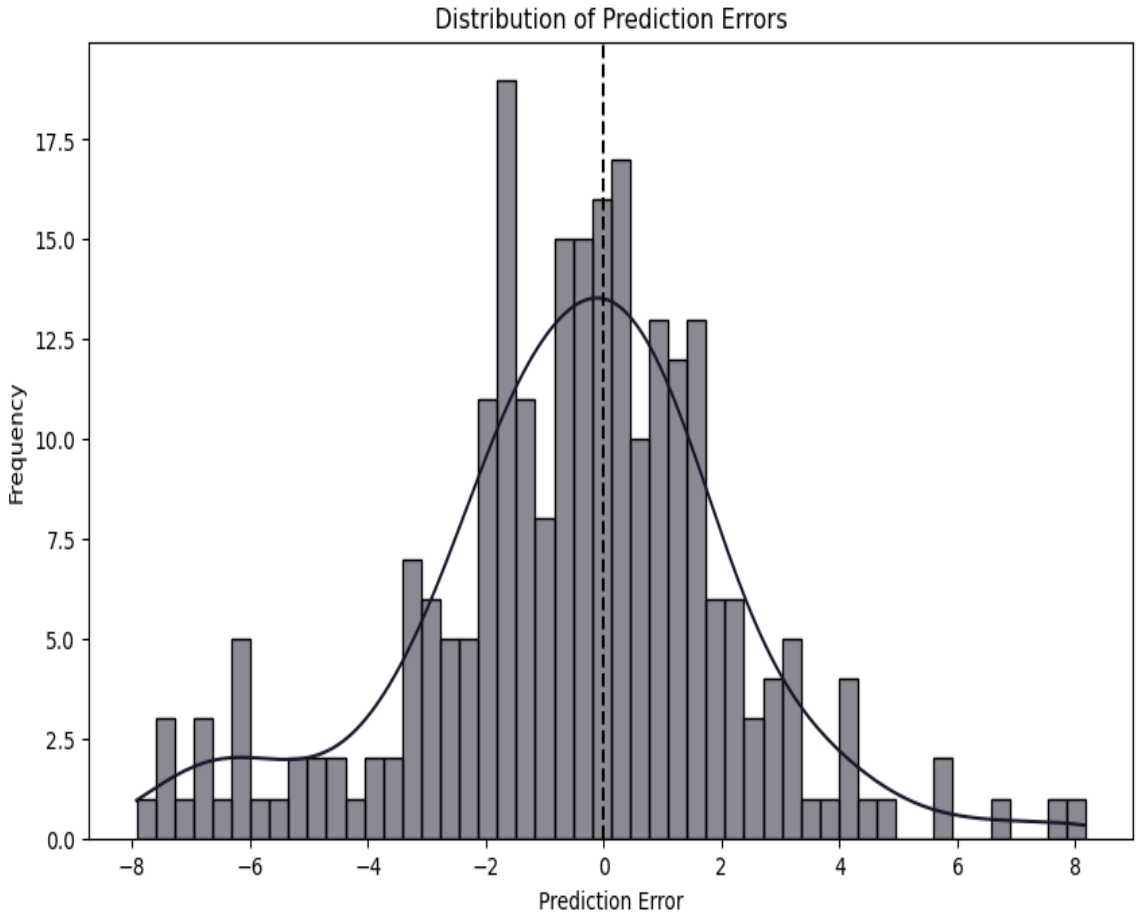
```
Training MSE: 1.2833770745865212
Training R-squared: 0.9587684005960768
Testing MSE: 7.457904121157479
Testing R-squared: 0.7154262864613008
Overall R-squared: 0.9164255658458025
Feature Importances:
            Feature  Importance
    City_Cluster_7    0.312246
        Race_Black    0.083226
   City_Cluster_10    0.066631
   Poverty_Poorest    0.046804
  Poverty_Lesspoor    0.045358
    City_Cluster_6    0.027635
        Race_White    0.026148
         Year_2021    0.023765
         Year_2020    0.022829
         Year_2010    0.020011
     Race_Asian/PI    0.017856
     Race_Hispanic    0.017193
   State_Cluster_7    0.017058
   State_Cluster_6    0.015395
    PopDensity_Low    0.014599
         Year_2012    0.014531
 Population_Larger    0.013786
Population_Smaller    0.013684
         Year_2019    0.013617
         Year_2018    0.011793
         Year_2014    0.011597
    City_Cluster_2    0.011595
   PopDensity_High    0.011593
         Year_2013    0.011557
         Year_2011    0.011437
    City_Cluster_8    0.010743
         Year_2015    0.009700
  Region_Northeast    0.008451
         Year_2017    0.007605
         Year_2016    0.007464
    City_Cluster_1    0.006531
    City_Cluster_5    0.005666
    City_Cluster_3    0.005266
    Region_Midwest    0.005252
   City_Cluster_11    0.005245
       Region_West    0.005023
   State_Cluster_4    0.004984
  Segregation_High    0.004748
      Region_South    0.004508
   State_Cluster_2    0.004418
    City_Cluster_4    0.004402
   Segregation_Low    0.004226
   State_Cluster_3    0.003538
   State_Cluster_1    0.002869
   State_Cluster_5    0.002475
    City_Cluster_9    0.002304
   State_Cluster_8    0.001743
   State_Cluster_9    0.000896
```

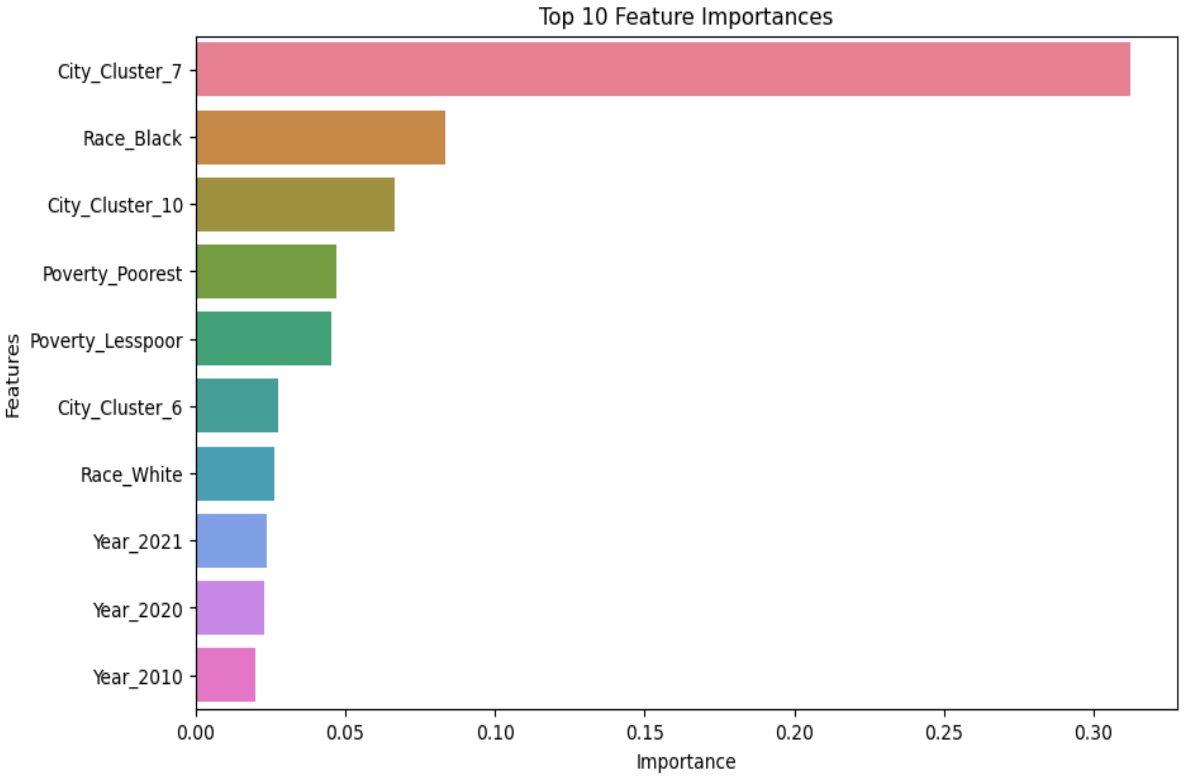
Feature Importance:

- The table shows the importance of each feature in the model, indicating the contribution of each feature to predicting the target variable.
- Features like City_Cluster_7, Race_Black, City_Cluster_10, and Poverty_Poorest have higher importance values, suggesting they are more influential in making predictions.

Visualization:

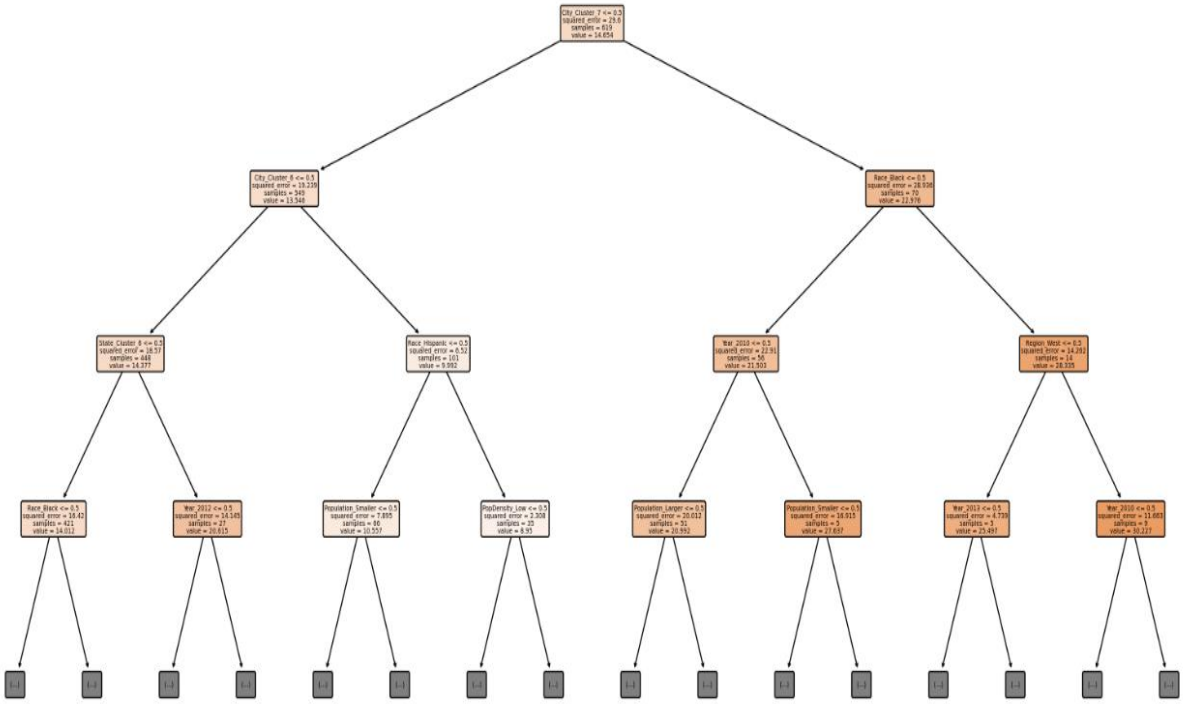
The model's performance and feature importance were visualized using seaborn and matplotlib libraries. A histogram of prediction errors provided insights into the model's residuals, while a bar plot displayed the top 10 feature importance, highlighting the variables that most strongly influence the death rate.





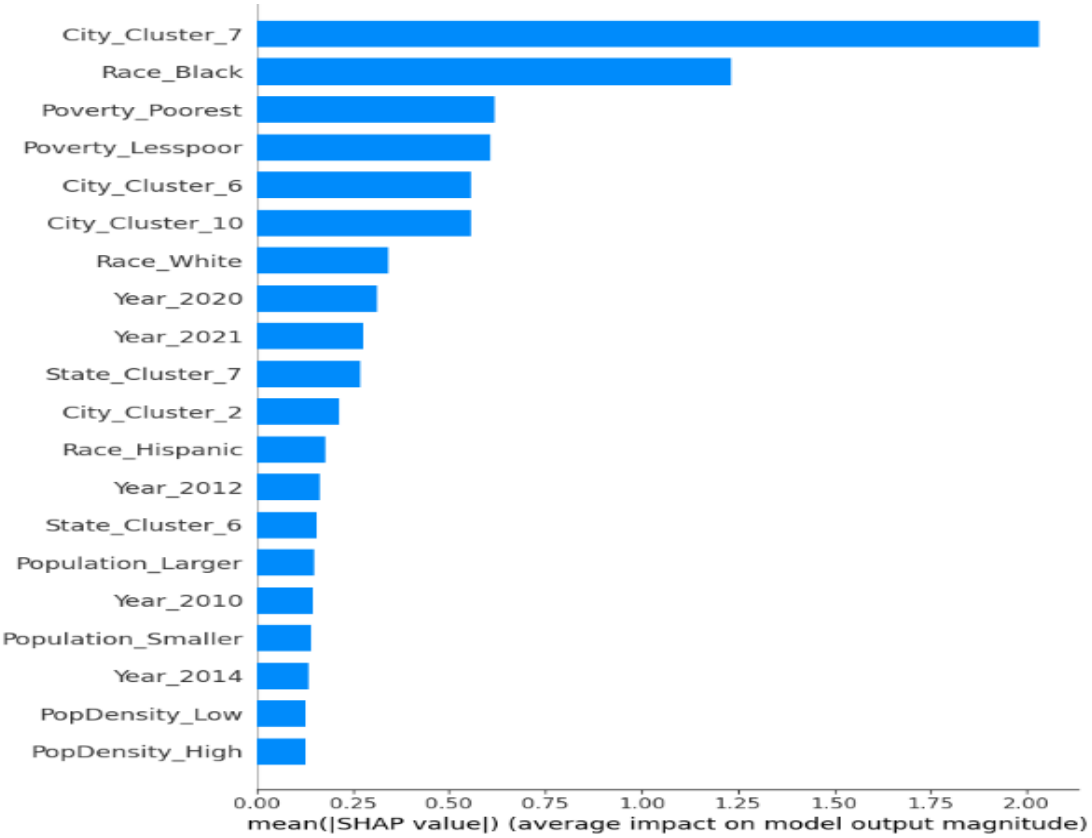
Decision Tree Visualization:

As part of the Random Forest ensemble, individual decision trees were examined. A single tree visualization, using the plot_tree function, offered a snapshot of the decision-making process within the model, although simplified by limiting the depth to three levels.



SHAP Analysis

The SHAP (SHapley Additive exPlanations) framework was employed to further interpret the Random Forest model. A SHAP summary plot was generated, summarizing the effects of all features on the model's output. This plot provides a more nuanced understanding of how each feature contributes to the model's predictions, complementing the information obtained from the feature importance analysis.



TEST RESULTS:

1. Durbin-Watson Statistic (DW):
   - Value: 2.06
   - Interpretation: The Durbin-Watson statistic measures the presence of autocorrelation in the residuals. A value close to 2 suggests that there is little autocorrelation in the residuals. In this case, a value of 2.06 indicates a relatively small degree of autocorrelation.

2. Jarque-Bera Statistic:
   - Value: 11.17
   - p-value: 0.00375
   - Interpretation: The Jarque-Bera test checks the normality of residuals. The statistic value of 11.17, coupled with a p-value of 0.00375, suggests that the residuals may not follow a normal distribution. A lower p-value indicates deviation from normality.

3. Omnibus Statistic:
- Value: 7.94
- p-value: 0.01885
- Interpretation: The Omnibus test is another test for normality. A lower p-value (0.01885) suggests a departure from normality, reinforcing the indication from the Jarque-Bera test.

4. Breusch-Pagan LM Test:
- p-value: 0.02174
- Interpretation: The Breusch-Pagan test assesses heteroscedasticity (unequal variance of residuals). A p-value of 0.02174 indicates the presence of heteroscedasticity, suggesting that the variance of the residuals may not be constant across all levels of the independent variables.

5. Variance Inflation Factors (VIF):
- Interpretation: VIF values measure the extent of multicollinearity among predictor variables. In this case, infinite (inf) VIF values for all variables indicate severe multicollinearity, making it challenging to isolate the individual impact of each variable.

- Consider addressing the non-normality of residuals and heteroscedasticity through potential transformations or alternative modeling techniques.
- Investigate and mitigate multicollinearity among predictor variables by possibly excluding highly correlated variables or using dimensionality reduction techniques.
- It's important to interpret the regression results cautiously due to the identified issues, and further diagnostic checks or model refinements may be necessary for improved reliability.

Conclusion:

The Random Forest Regressor exhibits strong performance, capturing intricate patterns in the training data with a low MSE of 1.28 and a high R-squared of 0.96. It generalizes well to new data, as seen in the testing MSE of 7.46 and R-squared of 0.72. The overall model fit is impressive, with an R-squared of 0.92, indicating substantial explanatory power. Feature importance analysis highlights crucial predictors, notably `City_Cluster_7`, `Race_Black`, `City_Cluster_10`, and `Poverty_Poorest`.

SHAP analysis provides nuanced insights into individual predictions, enhancing overall model interpretability. Considerations for further investigation include understanding specific problem contexts and potential limitations. Recommendations suggest focusing on interventions related to influential features, such as specific city clusters, racial demographics (`Race_Black`), and poverty levels (`Poverty_Poorest`). In conclusion, the Random Forest Regressor stands as a robust model, offering strong performance, interpretability, and actionable insights for real-world applications.

DISCUSSION & CONCLUSION

COMPARISON OF MODELS:

To compare Ridge Regression and Random Forest Regression based on the provided results, let's analyze the performance metrics:

Ridge Regression Results:

- Alpha (Regularization Strength): 1.0
- Mean Squared Error (MSE): 7.589574772134862
- R-squared: 0.7104020858944966

Random Forest Regression Results:

- Training MSE: 1.2833770745865212
- Training R-squared: 0.9587684005960768
- Testing MSE: 7.457904121157479
- Testing R-squared: 0.7154262864613008
- Overall R-squared: 0.9164255658458025

Comparison and Explanation:

1. Mean Squared Error (MSE):
   - Ridge Regression: 7.589574772134862
   - Random Forest Regression (Testing MSE): 7.457904121157479

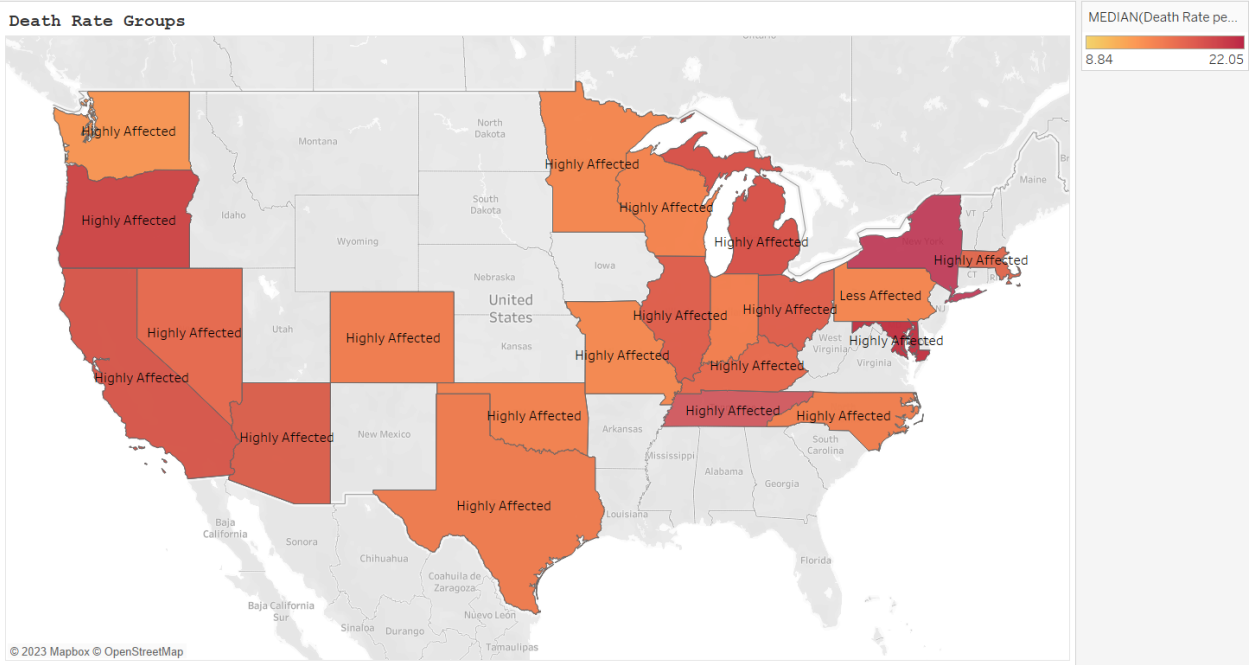
Random Forest Regression has a slightly lower testing MSE, suggesting better predictive performance in terms of mean squared error.

2. R-squared:
   - Ridge Regression: 0.7104020858944966
   - Random Forest Regression (Testing R-squared): 0.7154262864613008

Random Forest Regression has a slightly higher testing R-squared, indicating a better fit to the testing data.

3. Overall Assessment:
   - Random Forest Regression exhibits higher training R-squared (0.9587684005960768), indicating an excellent fit to the training data.
   - Random Forest Regression also has a high overall R-squared (0.9164255658458025), suggesting strong overall performance across training and testing datasets.

The Random Forest Regression model appears to be more suitable for the data. It achieves lower testing MSE, higher testing R-squared, and maintains a high overall R-squared, indicating good generalization and performance across both training and testing datasets. Random Forest models are known for their ability to handle complex relationships in data and often perform well in a variety of scenarios. However, the choice of the best model also depends on the specific goals of your analysis and other considerations such as interpretability and computational efficiency.



The exploration of 'DeathRate_per_100,000' using both Random Forest regression and SHAP analysis provides a comprehensive understanding of the factors influencing mortality rates. This research highlights the intricate interplay between various demographic, geographic, and socioeconomic factors, and their collective impact on public health outcomes.

LIMITATIONS:

The study acknowledges several limitations that may impact the robustness of its findings. Firstly, the simplification of key variables into broad categories, such as 'high' or 'low,' may oversimplify the intricate realities underlying the dataset. This approach has the potential to mask subtle yet crucial nuances that could influence the death rate. Additionally, the lack of continuous numerical values in important categories poses constraints on the depth of analysis, potentially affecting the precision of the results.

The categorical nature of many variables introduces challenges in model fit and interpretation, making it difficult to fully grasp the impact on the death rate and capturing intricate relationships. Moreover, the findings are specific to the dataset and its categorizations, raising concerns about their generalizability to different contexts or broader populations. Statistical indicators like the Jarque-Bera and Durbin-Watson statistics also suggest some concerns about the normality of residuals and autocorrelation, emphasizing the need for careful consideration of these factors in assessing the reliability of the results.

FUTURE SCOPE:

The study outlines avenues for future research to address the identified limitations. Enhancing data detailing by incorporating more nuanced, possibly continuous data could provide a more comprehensive understanding of the complexities involved. Exploring diverse modeling approaches, especially those better suited for categorical or complex datasets, may offer new insights and improve the accuracy of predictions.

Widening the range of variables to include factors like healthcare access or environmental conditions could enrich the analysis, providing a more holistic view of the factors influencing death rates. Longitudinal and comparative studies, examining changes over time or differences across various regions, are suggested to offer a more dynamic understanding of these factors. Furthermore, conducting policy impact studies to assess how specific policies or interventions affect death rates could yield valuable information for shaping future public health strategies.

CONCLUSION:

In conclusion, the study conducted using Ridge Regression has unveiled noteworthy associations between socio-demographic factors and the death rate per 100,000 people. While the model demonstrates a good level of explanatory power, caution is advised in interpreting the results due to the acknowledged limitations related to data categorization and model constraints. The research lays the foundation for more nuanced investigations, underscoring the importance of detailed data and varied analytical approaches. The insights gained from this study contribute valuable knowledge to understanding public health trends and have the potential to guide the development of future health policies and targeted interventions.

REFERENCES

- https://ourworldindata.org/influenza-deaths
- https://www.sciencedirect.com/science/article/abs/pii/S0264410X21015619?via%3Dihub
- https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0283475
- Association of Asthma with Treatments and Outcomes in Children with Critical Influenza - ScienceDirect
- Disease Burden of Flu | CDC
- https://jogh.org/documents/issue201902/jogh-09-020421.pdf
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5818346/
- https://academic.oup.com/cid/article/54/10/1427/352114
- https://read.dukeupress.edu/demography/article/56/5/1723/168053/Determinants-of
- Influenza-Mortality-Trends-Age
- https://bmcmedicine.biomedcentral.com/articles/10.1186/1741-7015-11-153
- https://www.tandfonline.com/doi/full/10.1080/02763869.2019.1657734
- https://link.springer.com/article/10.1007/s40471-018-0136-1
- Our World in Data: 'Influenza Deaths'
- ScienceDirect: 'Vaccine effectiveness against laboratory-confirmed influenza hospitalizations among elderly adults during the 2010–2011 to 2015–2016 influenza seasons'
- PLOS ONE: 'The association of COVID-19 occurrence and severity with the use of angiotensin-converting enzyme inhibitors or angiotensin-receptor blockers in patients with hypertension'
- ScienceDirect: 'Critical respiratory illness in children with influenza infection'

- Centers for Disease Control and Prevention (CDC): 'Disease Burden of Influenza'
- Journal of Global Health: 'Global mortality associated with seasonal influenza epidemics: New burden estimates and predictors from the GLaMOR Project'
- PubMed Central: 'The Burden of Influenza-Associated Hospitalizations in the United States'
- Clinical Infectious Diseases: 'Global Mortality Estimates for the 2009 Influenza Pandemic from the GLaMOR Project: A Modeling Study'
- Demography - Duke University Press: 'Determinants of Influenza Mortality Trends: Age-Period-Cohort Analysis of Influenza Mortality in the United States, 1959–2016'
- BMC Medicine: 'Estimates of global seasonal influenza-associated respiratory mortality: a modelling study'
- Taylor & Francis Online: 'Influenza vaccination and mortality: Differentiating vaccine effects from bias'
- SpringerLink: 'Influenza and respiratory disease research: The need for a holistic approach'