Hitarth Bhatt

Tirth Modi

Heet Dipeshe

ANALYSIS TRENDS OF HEART DIEASE MORTALITY IN ADULTS

Exploratory Data Analysis

Group 13

PROBLEM STATEMENT

Heart disease remains a significant cause of death worldwide with great variations in mortality rates at national and state levels. This study identifies the need to understand such variation and key trends in order to target public health interventions for improving healthcare outcomes. This project intends to carry out exploratory data analysis (EDA) on heart disease mortality rates by focusing on the following key patterns:

National Average Mortality Rate: This gives the national average mortality rate due to heart disease, which is taken as a benchmark for comparison with state-level data.

- State-Level Heart Disease Mortality: Examining variations in mortality rates across different states to identify regions with higher or lower mortality, potentially influenced by local factors such as healthcare access, lifestyle, and demographics.

- Largest Increase and Decrease in State-Level Mortality: Identifying states with the most significant increase or decrease in heart disease mortality rates over a specified period. This analysis can help pinpoint regions that have successfully improved heart disease outcomes or where the situation has worsened.

- Comparison by Gender: Analyzing the differences in heart disease mortality rates between genders, exploring how men and women may experience varying levels of risk and mortality from heart disease.

DATASET DESCRIPTION

This dataset contains detailed information regarding heart disease mortality rates of adults aged 35 and older in the United States. It covers three time periods: 2015-2017, 2017-2019, and 2019-2021. It contains mortality data for all U.S. states and counties, which allows for the analysis of both national and regional levels. The key stratifications include gender (Male, Female, Overall) and location (state-wise), which enables insights regarding how heart disease mortality rates differ across populations and regions. The dataset includes age-adjusted mortality rates per 100,000 population, which ensures that the comparison is fair between areas with different age distributions. In addition, it provides geospatial coordinates for mapping and identification of regional patterns of heart disease mortality, which supports targeted public health initiatives.

DATA INFORMATION

The first few lines of analysis using functions such as nunique(), describe(), info(), and head(10) give good insight into the structure and statistics of the dataset. The nunique() function helps in establishing the diversity of categorical variables, such as states, gender, and race. The describe() function gives key statistics, such as average mortality rates, to understand the data distribution. The function info() provides an overview of data types, number of non-nulls, and memory usage, which is quite useful in evaluating the quality of the data. Finally, head(10) previews a few rows of the start; hence one could easily spot missing or inconsistent data. These functions are extremely important for data exploration and guiding further analysis.

DATA CLEANING

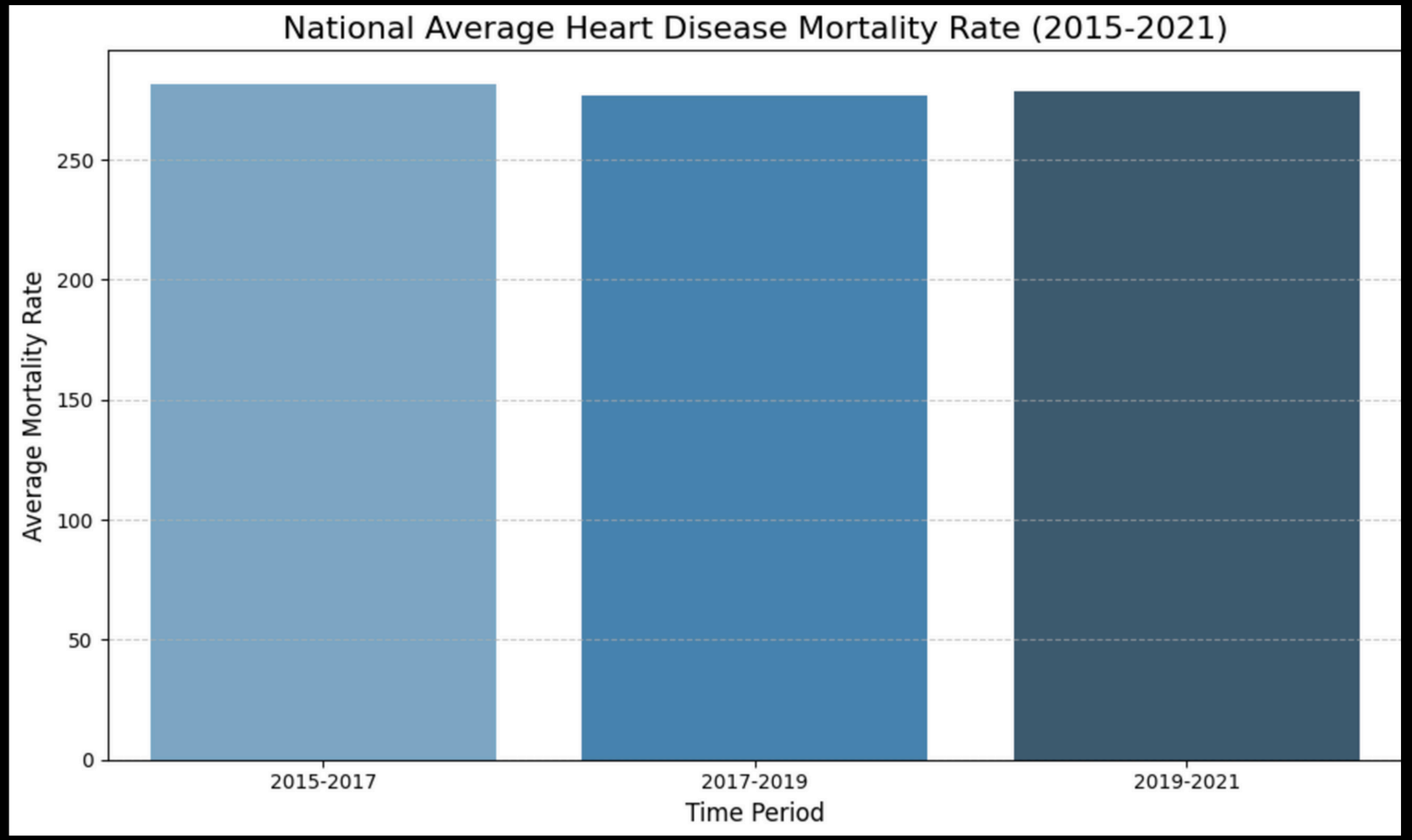
Data Cleaning Overview

- Importance: Data cleaning ensures reliable, accurate insights for analysis.
- Key Tasks:
  - Identifying and fixing errors.
  - Removing duplicates and handling missing values.
  - Standardizing formats and handling outliers.
- Process:
  - Missing Data: Found missing values in key columns (e.g., mortality rate, location, geographic info). Handled by:
    - Removing rows with missing values in critical columns (mortality rate, location).
    - Imputing missing demographic data as "Not Specified."
  - Data Standardization: Renamed columns for clarity and consistency across time periods (e.g., locationabbr → state abbr, data_value → mortality rate).
- Outcome: Clean data prepared for accurate and meaningful analysis.

Data Cleaning Implementation

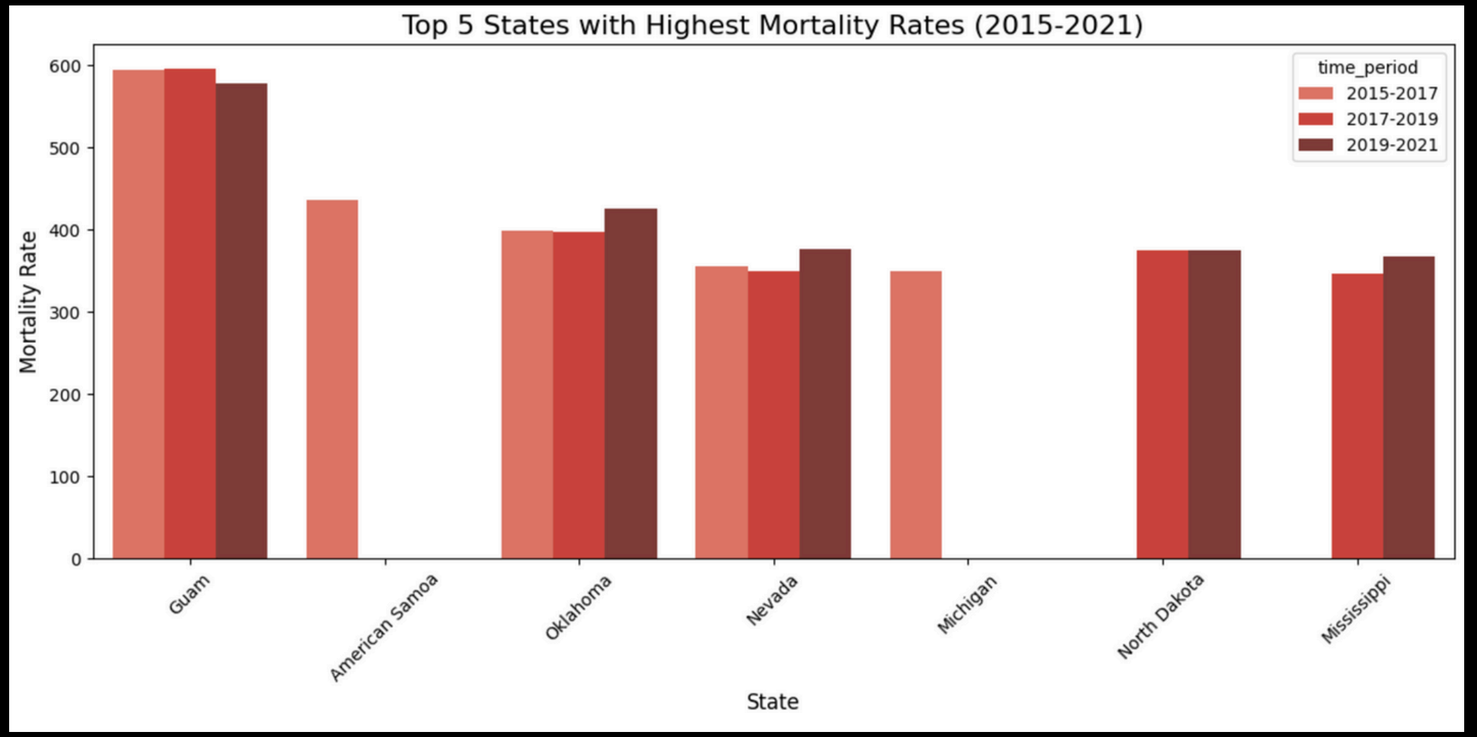
- Handling Missing Values:
  - Analyzed missing data using .isnull().sum().
  - Dropped rows with missing critical columns (e.g., mortality rate, longitude, latitude).
  - Filled missing values in demographic fields with "Not Specified."
- Data Transformation:
  - Converted the mortality rate column to numeric format, resolving non-numeric entries.
- Final Outcome:
  - Reduced data noise by focusing on essential columns.
  - Ensured data consistency across all datasets (2015-2017, 2017-2019, 2019-2021).
  - Prepared data for reliable analysis and interpretation.

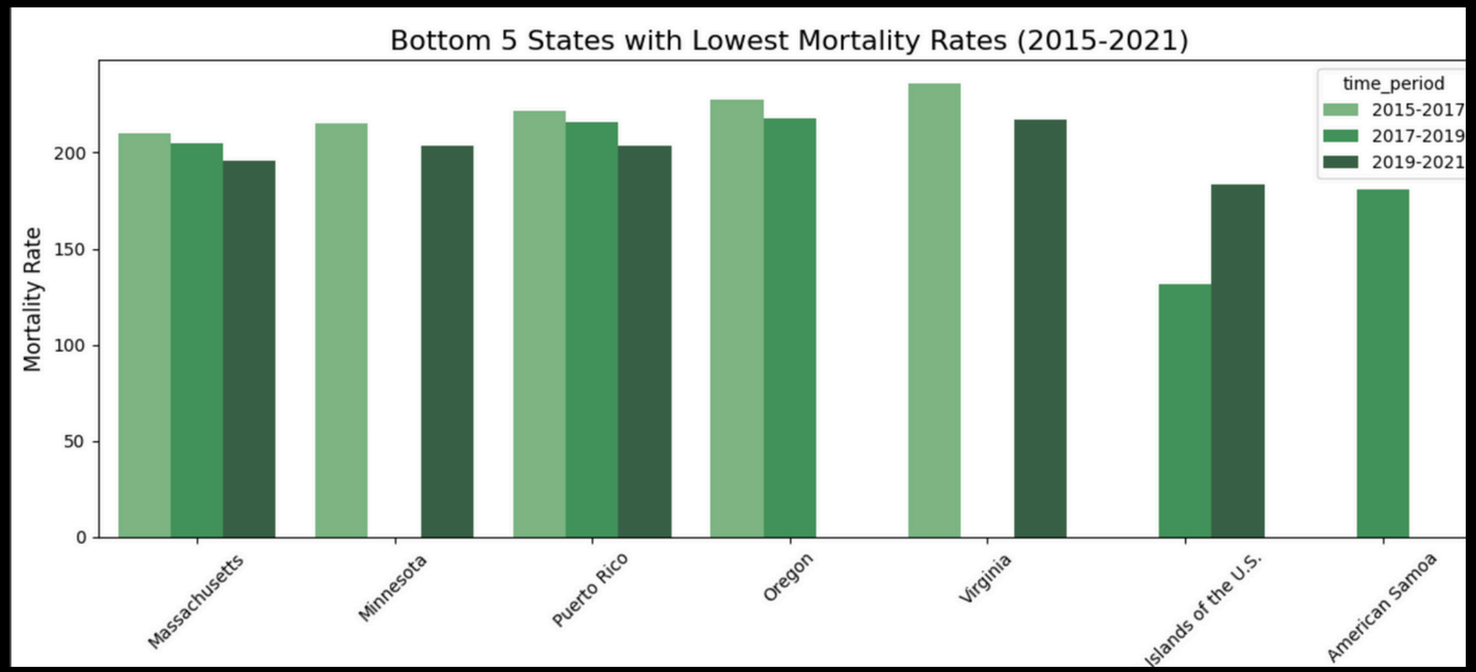
TREND ANALYSIS :-



Shows the national average heart disease mortality rate from 2015 to 2021, segmented into three time periods.

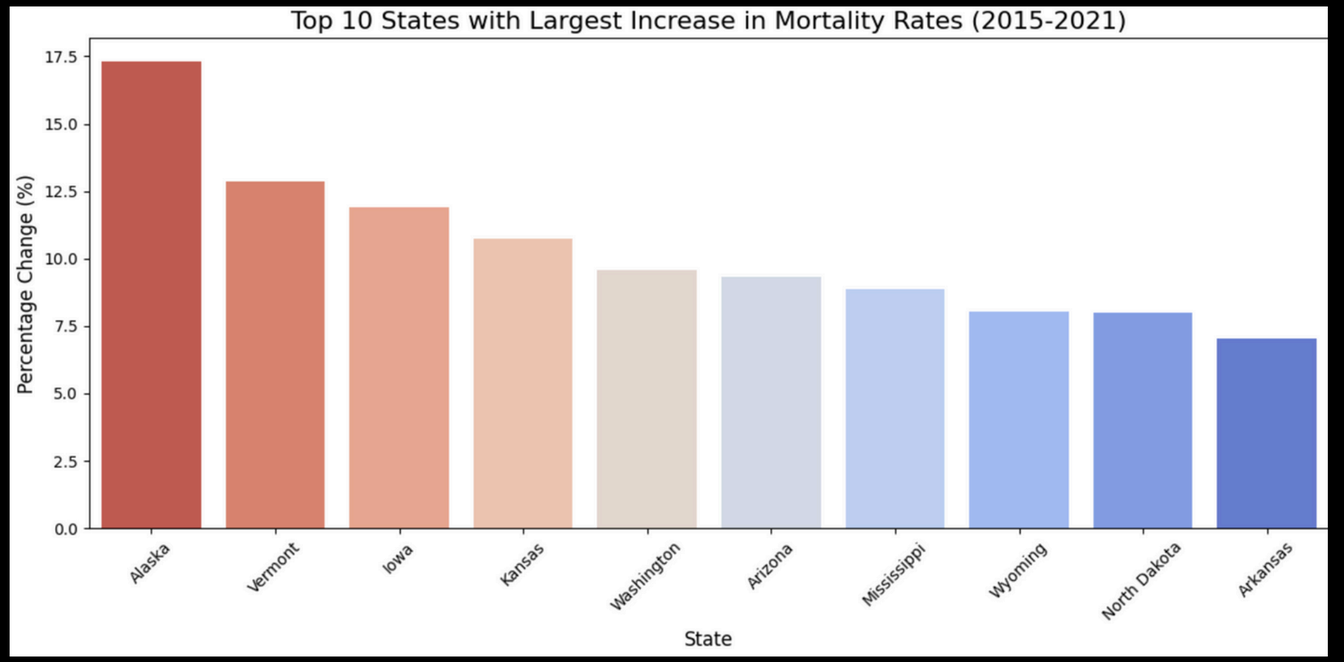
Highlights the top five states with the highest mortality rates from 2015 to 2021, segmented by three time periods.

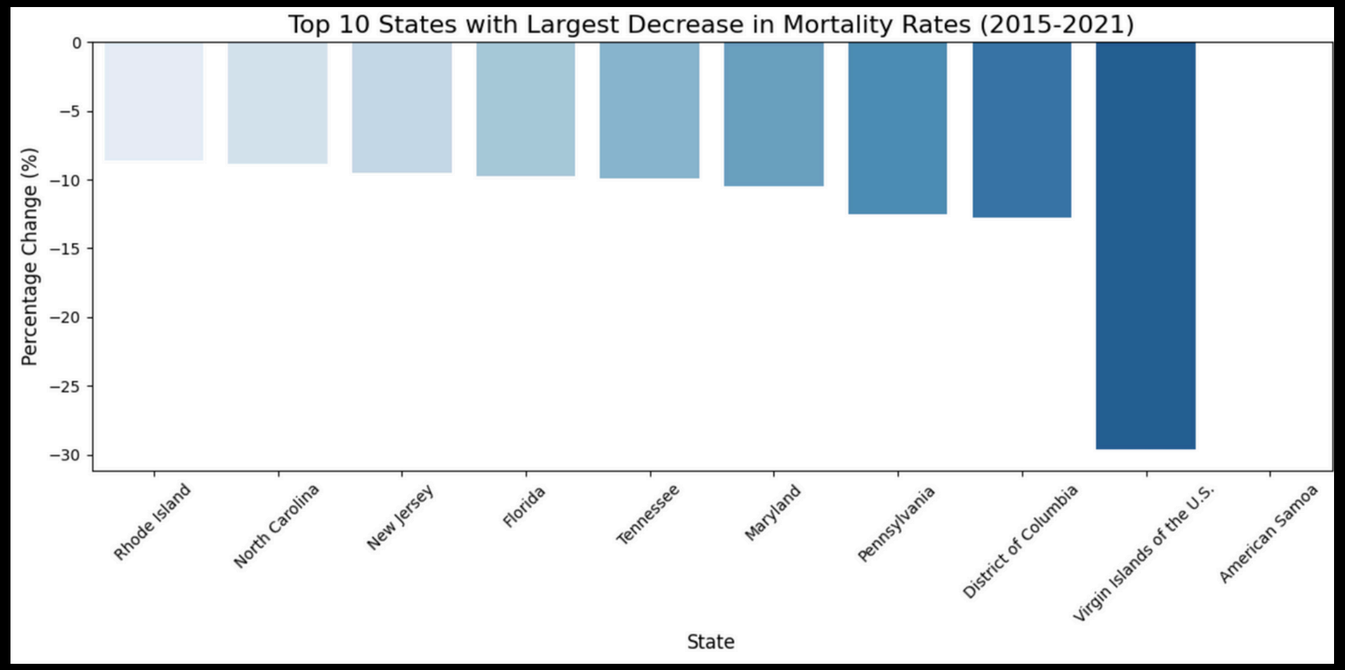




Shows the bottom five states with the lowest mortality rates from 2015 to 2021, segmented by three time periods.

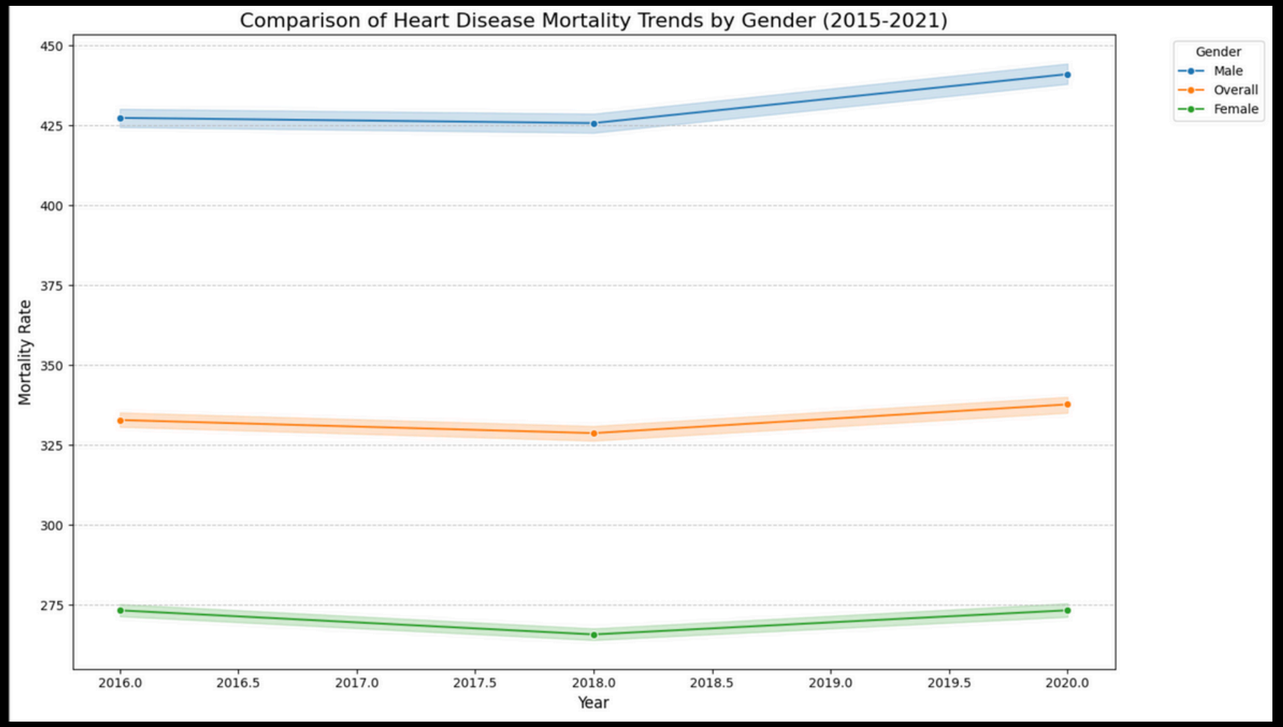
Bar chart highlighting the top 10 states with the largest increase in heart disease mortality rates from 2015 to 2021, with Alaska showing the highest rise.

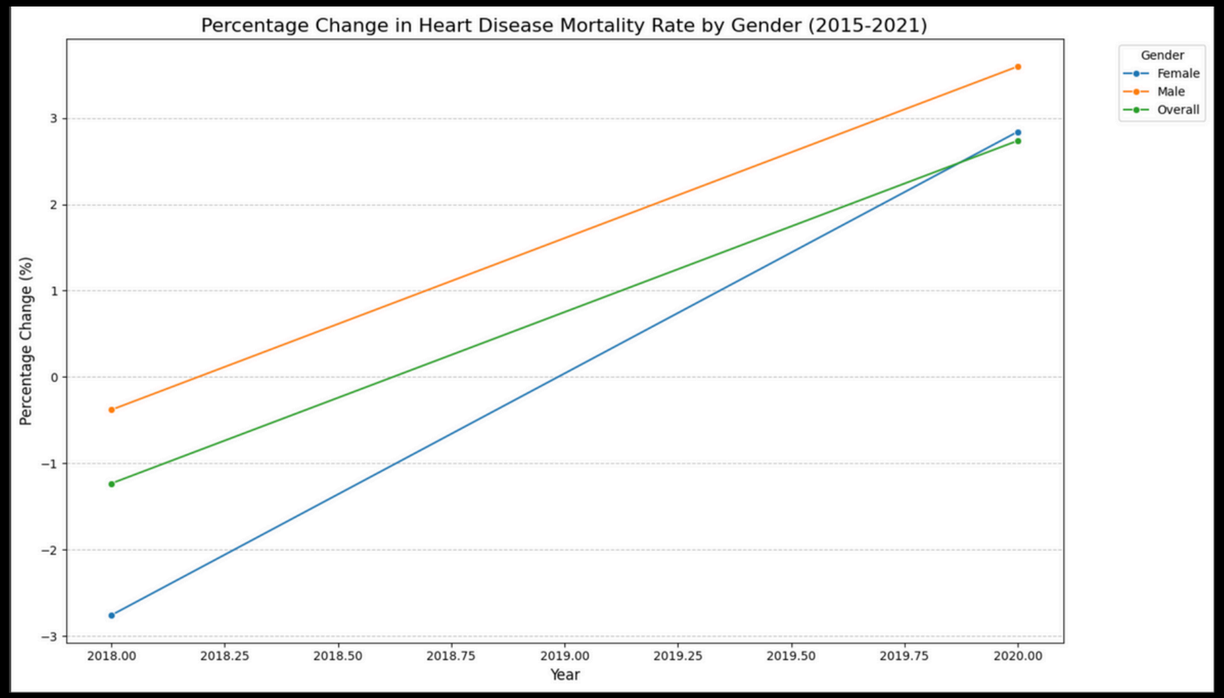




Bar chart showing the top 10 states with the largest decrease in heart disease mortality rates from 2015 to 2021, led by the Virgin Islands and American Samoa.

Line chart comparing heart disease mortality trends by gender (male, female, and overall) from 2015 to 2021.





Line chart depicting percentage change in heart disease mortality rates by gender (male, female, and overall) from 2018 to 2021.

MODEL PREDICTION

Objective:

- **Predict mortality rates across states and demographics using a Random Forest Regressor.**
- **Generate insights to guide health interventions and resource allocation.**

Key Approach:

- **Model trained on historical data (2019–2021).**
- **Evaluation metrics:**
  - **RMSE: 129.43**
  - **MAE: 104.24**
  - **R^2: 0.8236**

Outcome:

- **Reliable predictions and insights into mortality rate disparities.**

Key Findings:

- **High Mortality States: Require targeted health interventions.**
- **Low Mortality States: Serve as benchmarks for effective policies.**

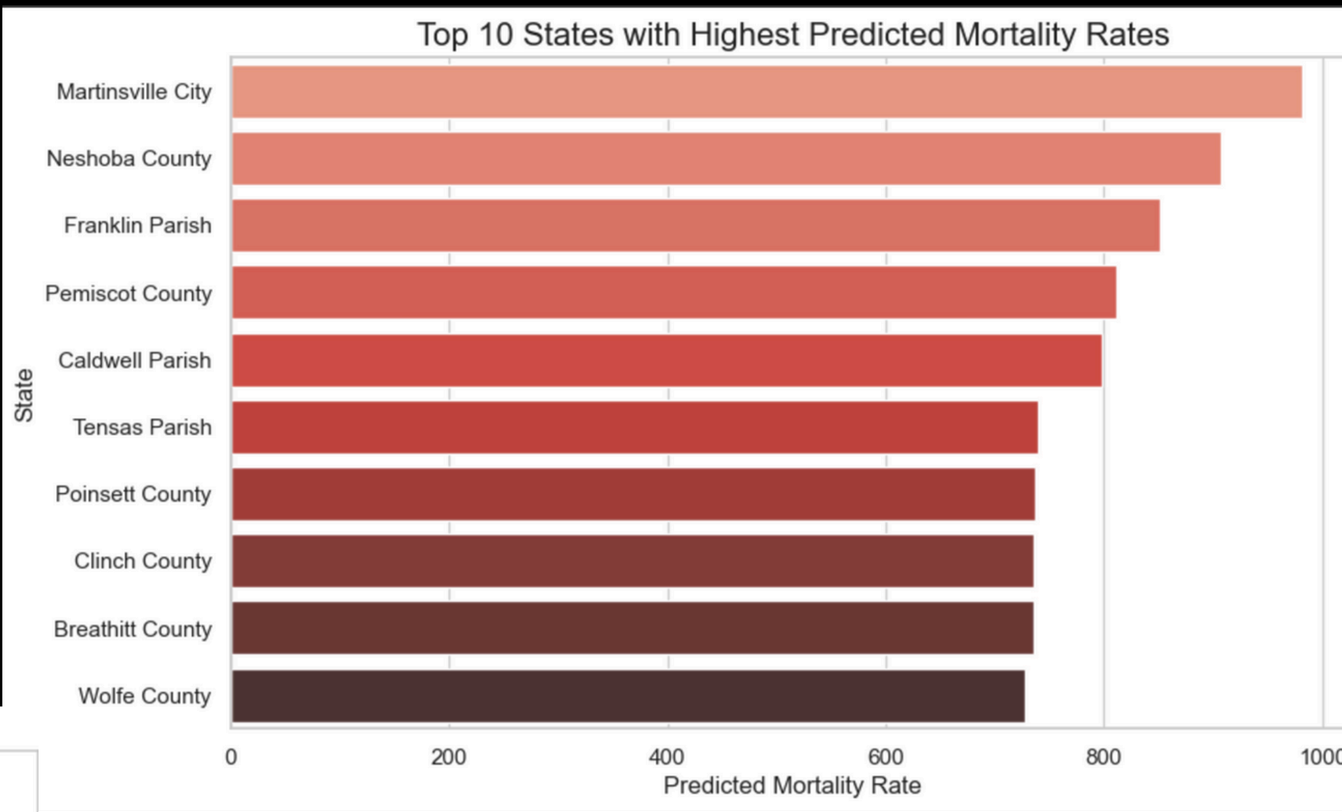
Model Validation:

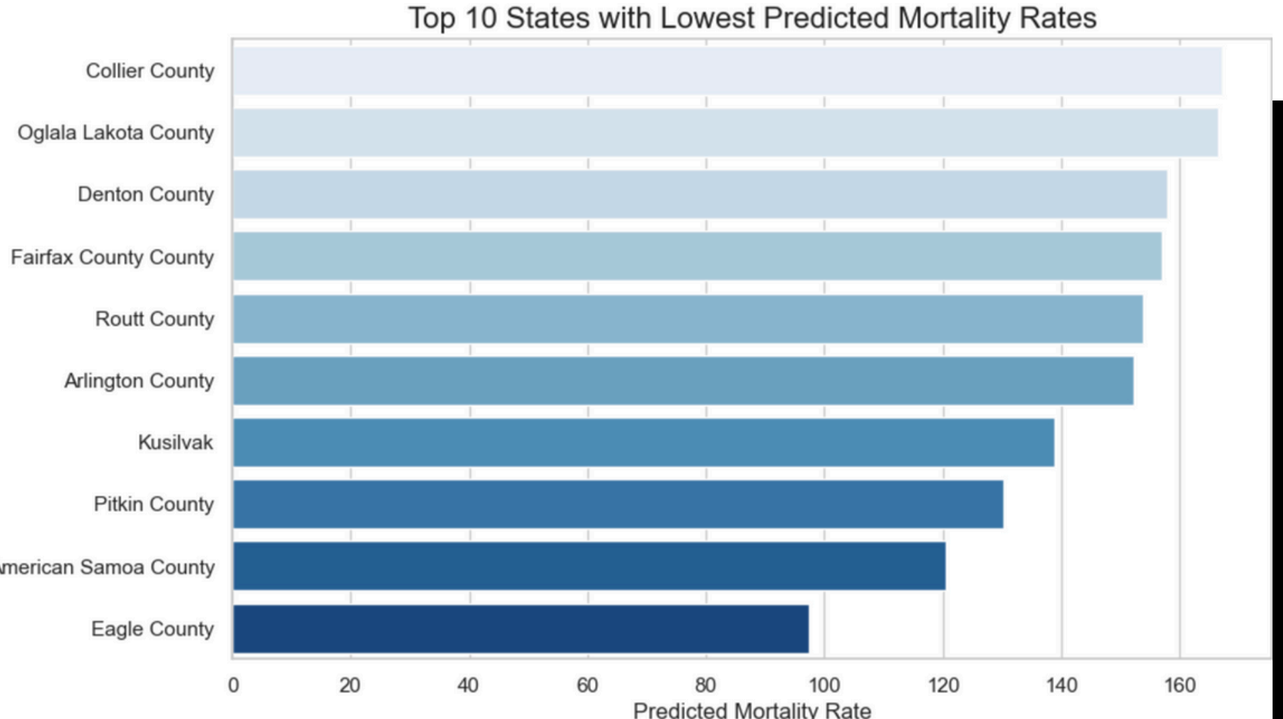
- **Cross-validation confirms stability across datasets.**

Conclusion:

The model demonstrates strong predictive power, helping identify priority regions for health improvement initiatives.

This bar plot highlights the states with the most severe predicted mortality rates based on the Random Forest model, using a red gradient to emphasize intensity.





This bar plot showcases the states with the lowest predicted mortality rates, visualized using a blue gradient to reflect better health outcomes.

CONCLUSION AND FUTURE PREDICTIONS

Objective Achieved:

Predicted heart disease mortality rates for 2022–2026 using historical data (2015–2021). The model identified key trends and provided actionable insights for policymakers.

Key Findings:

1. **Data Trends:**
   - **Mortality rates varied across states, counties, and demographics.**
   - **Combined geographic, demographic, and time features revealed significant patterns.**
2. **Prediction Accuracy:**
   - **RMSE: Good predictive performance with reliable forecasts for future trends.**
   - **Demographic and geographic attributes were the strongest predictors.**
3. **Future Predictions:**
   - **Highlighted disparities in mortality rates, pinpointing regions requiring interventions.**

Targeted Interventions:

- **Focus on high-risk regions with prevention programs, education, and healthcare investments.**

Ongoing Monitoring:

- **Regular updates with new data to refine predictions and track trends.**

Future Enhancements:

- **Explore additional predictors like socio-economic factors and healthcare access.**
- **Use findings to design equitable public health policies.**

Group Members

Hitarth Bhatt

Tirth Modi

Heet Dipeshe

THANK YOU