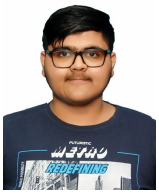
Exploratory Data Analysis on

Analyzing Trends of Heart Diease Mortality (2015-2021)

by

Group 13



Hitarth Bhatt
ID: 202201024
Course: BTech(ICT)

Tirth Modi
ID: 202201513
Course: BTech(ICT)

Heet Dipeshe
ID: 202203006
Course:
BTech(MnC)

Course Code: IT 462
Semester: Autumn 2024

Under the guidance of

Dr. Gopinath Panda

Dhirubhai Ambani Institute of Information and Communication Technology

December 2, 2024

Acknowledgment

I am writing this letter to express my heartfelt gratitude for your guidance and support throughout the duration of my project titled “Analyzing Trends of Heart Diease Mortality (2015-2021)”. Your invaluable assistance has played a pivotal role in shaping the successful completion of this endeavor.

I am extremely fortunate to have had the opportunity to work under your mentorship. Your expertise, encouragement, and willingness to share your knowledge have been instrumental in elevating the quality and scope of my project. Your constructive feedback and insightful suggestions have helped me overcome challenges and develop a deeper understanding of the subject matter.

Furthermore, I would like to extend my appreciation to the entire team at [Organization/Institution Name] for fostering an environment of collaboration and innovation. The resources and facilities provided have been crucial in conducting comprehensive research and analysis.

I would also like to express my gratitude to my peers and colleagues who have been supportive throughout this journey. Their valuable input and camaraderie have been a constant source of motivation.

Completing this project has been a tremendous learning experience, and I am confident that the knowledge and skills acquired during this endeavor will serve as a solid foundation for my future endeavors.

Once again, thank you for your unwavering guidance and belief in my abilities. Your mentorship has been invaluable, and I am truly grateful for the opportunity to work with you.

Sincerely
Hitarth Bhatt, 202201024
Tirth Modi, 202201513
Heet Dipeshe, 202203006

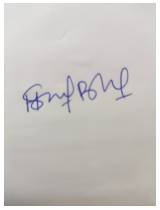
DECLARATION

We, members of Group 13, hereby declare that the EDA project work presented in this report is our original work and has not been submitted for any other academic degree. All the sources cited in this report have been appropriately referenced.

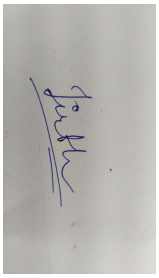
We acknowledge that the data used in this project is obtained from the [...] site. We also declare that we have adhered to the terms and conditions mentioned in the website for using the dataset. We confirm that the dataset used in this project is true and accurate to the best of our knowledge.

We acknowledge that we have received no external help or assistance in conducting this project, except for the guidance provided by our mentor Prof. Gopinath Panda. We declare that there is no conflict of interest in conducting this EDA project.

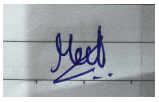
We hereby sign the declaration statement and confirm the submission of this report on 2nd July, 2023.



First member
ID: 202401003
Course: MSc(DS)



Second Member
ID: 202401002
Course: MSc(IT)



Third Member
ID: 202401005
Course: BTech(ICT)

CERTIFICATE

This is to certify that Group 13 comprising Hitarth Bhatt, Tirth Modi and heet Dipeshe has successfully completed an exploratory data analysis (EDA) project on the Trends of Heart Disease Mortality in year 2015 to 2021 which was obtained from data.gov

The EDA project presented by Group 13 is their original work and has been completed under the guidance of the course instructor, Prof. Gopinath Panda, who has provided support and guidance throughout the project. The project is based on a thorough analysis of the Heart Disease Mortality dataset, and the results presented in the report are based on the data obtained from the dataset.

This certificate is issued to recognize the successful completion of the EDA project on the analysis which demonstrates the analytical skills and knowledge of the students of Group 13 in the field of data analysis.

Signed,
Dr. Gopinath Panda,
IT 462 Course Instructor
Dhirubhai Ambani Institute of Information and Communication Technology
Gandhinagar, Gujarat, INDIA.

December 2, 2024

Contents

List of Figures

Abstract

This exploratory data analysis investigates trends in heart disease mortality across demographics, regions, and time periods to identify disparities and risk factors. By analyzing a comprehensive dataset, the study highlights variations in mortality rates by age, gender, race, and location, uncovering significant geographic and demographic patterns. Visualizations and statistical summaries reveal clusters of high mortality linked to socioeconomic and lifestyle factors. The findings provide actionable insights for policymakers and healthcare providers to target vulnerable populations and design interventions to reduce heart disease mortality.

Chapter 1. Introduction

1.1 Analyzing Trends and Behavioral Risk Factors in U.S. Heart Disease Mortality (2015-2021)

Heart disease remains a persistent public health challenge in the U.S., with mortality rates influenced by behavioral risk factors and showing significant geographic variation. To combat this epidemic effectively, understanding how mortality rates have changed over time and the key behavioral contributors to these trends is crucial.

Using mortality rate datasets from 2015-2017, 2017-2019, and 2019-2021, along with the Behavioral Risk Factor Surveillance System (BRFSS) data, this project aims to:

1. Explore temporal trends in heart disease mortality across states and counties.
2. Investigate correlations between behavioral risk factors (e.g., smoking, obesity) and mortality rates over different periods.
3. Highlight regions with rising or consistently high mortality rates and identify the primary contributing factors.
4. Provide actionable insights to support public health planning and resource allocation.

1.2 Data Collection

We started looking for a dataset related to heart disease from the **data.gov.in** website, which contains public datasets pertaining to India. After checking for all available resources, we were unable to find a suitable dataset that matched our criteria to perform a meaningful analysis of heart disease trends. After coming to this realization, as our focus was set for the United States, we changed our direction toward the central repository for datasets maintained by the U.S. government at **data.gov**.

On data.gov, we first searched broadly for datasets about "heart disease." The website returned several results, and as we looked through the options, we determined that datasets that could provide a detailed geographic breakdown by U.S. state or county would be particularly valuable. Among the results, a dataset titled "Heart Disease Mortality by States/County, U.S." caught our attention.

As we looked at this dataset further, we found that it gave us all the information regarding heart disease mortality, divided into states and counties. This was a very important dimension to our analysis since it allowed us to analyze the impacts of heart disease in a more localized context. Moreover, the dataset cited the **Centers for Disease Control and Prevention (CDC)** as its source, which means that the data is reliable and relevant.

We downloaded the file in CVS format for further easy utilization. A quick scanning shows well-structured output where clearly defined columns point towards states and counties, followed by death rates and any other associated data points; though the initial objective would be data on heart conditions generally, this dataset delivers more particular and actionable outlook on things, making it so relevantly useful for tracing trends for heart disease across various United States geographics.

1.3 Packages required

In this project, we utilized several key Python libraries that were instrumental in analyzing and interpreting the data. These packages—**Pandas**, **NumPy**, **Seaborn**, **Matplotlib**, and **Scikit-learn**—formed the foundation of our workflow. Below, we explain the purpose of each library and why we chose them for this analysis.

Pandas

Pandas served as the backbone of our data analysis process. It allowed us to load, clean, and manipulate the dataset with ease. By loading the data into a DataFrame, a tabular structure provided by Pandas, we were able to:

- **Clean the data**: This included handling missing values, renaming columns, and converting data types.
- **Explore the data**: Using functions like `head()`, `info()`, and `describe()`, we quickly understood the dataset's structure and key statistical summaries.

- **Aggregate and group data**: Pandas made it simple to group and summarize data by categories such as gender and race/ethnicity, helping us identify trends and patterns.

Without Pandas, managing and preparing a dataset of this size and complexity would have been far more time-consuming.

NumPy

NumPy provided us with the tools to perform efficient numerical computations. While Pandas handles tabular data effectively, NumPy was essential for mathematical operations and performance-critical tasks. Specifically, we used NumPy to:

- **Perform statistical calculations**: We calculated means, medians, and standard deviations to understand data distributions.
- **Work with arrays**: NumPy arrays made it easier and faster to process numerical data compared to standard Python lists.

Moreover, NumPy is the foundation for many other libraries, including Pandas and Scikit-learn, making it a vital part of our analysis.

Seaborn

For visualizing the data, we relied on Seaborn to create insightful and visually appealing plots. Its simplicity and aesthetics made it a perfect choice for exploring relationships and trends. Key visualizations included:

- **Histograms and boxplots**: To analyze data distributions and identify outliers.
- **Scatterplots and regression plots**: To explore relationships between variables.
- **Heatmaps**: To visualize correlations between different factors.

Seaborn's built-in aesthetics allowed us to create professional-looking plots without needing extensive customization.

Matplotlib

While Seaborn was excellent for quick and detailed visualizations, Matplotlib provided us with the flexibility to customize and fine-tune our plots. We used Matplotlib to:

- Adjust axis labels, titles, and legends for clarity.
- Add annotations to highlight important trends or data points.
- Create custom visualizations, such as pie charts, to complement our analysis.

The combination of Seaborn and Matplotlib ensured that our visualizations were both insightful and presentation-ready.

Scikit-learn (Sklearn)

Although Scikit-learn is primarily known for machine learning, it proved invaluable during our EDA phase. We used Scikit-learn for:

- **Data preprocessing**: Scaling numerical features and encoding categorical variables, ensuring the dataset was ready for analysis.
- **Feature selection**: Identifying the most important variables related to heart disease mortality trends.
- **Uncovering patterns**: Techniques like Principal Component Analysis (PCA) and clustering helped us visualize and analyze hidden structures in the data.

These tools added depth to our analysis and ensured our data was optimized for further exploration.

Together, these libraries formed a cohesive toolkit that made our EDA process efficient and comprehensive. **Pandas** and **NumPy** allowed us to clean and organize the data effectively, while **Seaborn** and **Matplotlib** enabled us to visualize trends and relationships clearly. Finally, **Scikit-learn** bridged the gap between preprocessing and advanced analysis, helping us uncover meaningful insights.

By using this combination of tools, we were able to analyze our data thoroughly, uncover key patterns, and lay the groundwork for further modeling and decision-making.

Chapter 2. Data Inspection

2.1 Dataset Description

The dataset primarily talks about the mortality of heart diseases among adults aged 35 years and above within U.S. states and counties. The dataset provides three time periods: 2015-2017, 2017-2019, and 2019-2021, and is useful for drawing analysis on the trends concerning cardiovascular disease mortality. Moreover, it's stratified by gender categories such as Male, Female, Overall, and race/ethnicity categories, for example, Hispanic and White to examine disparity in health outcomes in detail.

This data set has very crucial attributes. For instance, there is the year, state abbreviations and their full names along with geometric coordinates which means latitude and longitude for a location in which geospatial mapping is possible. The source of the report is presented as an age-adjusted, 3-year average rate per 100,000 population facilitating comparison against different age distributions. According to the description, its source is the National Vital Statistics System (NVSS); it is considered the most trusted repository for U.S. health statistics.

Key Features of our Datasets:

- **Comprehensive Coverage:** These datasets includes heart disease mortality data from all U.S. states, counties, and territories, making it a robust resource for nationwide analysis.
- **Time Periods:** It spans three distinct periods—2015-2017, 2017-2019, and 2019-2021—allowing us to study how heart disease mortality rates have changed over time.
- **Stratifications for Detailed Insights:** The dataset is broken down by gender (Male, Female, and Overall) and race/ethnicity (e.g., Hispanic, White), enabling us to examine disparities and identify which groups are most affected.
- **Age-Adjusted Rates:** Mortality rates are age-adjusted and reported per 100,000 people, ensuring that comparisons between different regions and demographics

are fair and consistent

- **Geospatial Data:** With latitude and longitude coordinates provided, the dataset allows for geospatial mapping and the identification of regional patterns or hotspots.

We have chosen this dataset because it offers a wealth of valuable features that make it ideal for analyzing heart disease mortality trends. Covering data across U.S. states and counties over three time periods—2015-2017, 2017-2019, and 2019-2021 It allows us to explore both how mortality rates have changed over time and how they vary geographically. Such stratifications by gender and race/ethnicity of the dataset open up the possibility of going deeper into health disparities to understand which populations are most affected. Applying age-adjusted mortality rates per 100,000 ensures that the comparison is not biased toward those regions having different population structures, thus making the analysis more valid and meaningful.

It gets more open to possibly even being just geospatial, stuff to map. Probably seeing regional patterns of death within the heart attack and be telling places and areas needing extra focus on further public health initiative. Using this dataset, multiple tools that were available, provide pathways on pattern detection, in a strategy of formulation, then on the top line can bring down morbidity with heart strokes more meaningfully.

2.2 Data Information

This dataset makes us understand the information which it holds. Every row has a small part of a story; it allows viewing the whole view. Now, let's check quickly through the first few rows with

```
Mortality_Rate_2015_17.head(10)
```

Then we check it data's information as the datatype and count coloum wise:

```
Mortality_Rate_2015_17.info()
```

```
Data columns (total 21 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   Year                         59094 non-null  int64
 1   LocationAbbr                 59094 non-null  object
 2   LocationDesc                 59094 non-null  object
 3   GeographicLevel              59094 non-null  object
 4   DataSource                   59094 non-null  object
 5   Class                        59094 non-null  object
 6   Topic                        59094 non-null  object
 7   Data_Value                   32230 non-null  float64
 8   Data_Value_Unit              59094 non-null  object
 9   Data_Value_Type              59094 non-null  object
 10  Data_Value_Footnote_Symbol   26864 non-null  object
 11  Data_Value_Footnote          26864 non-null  object
 12  StratificationCategory1      59094 non-null  object
 13  Stratification1              59094 non-null  object
 14  StratificationCategory2      59094 non-null  object
 15  Stratification2              59094 non-null  object
 16  TopicID                      59094 non-null  object
 17  LocationID                   59094 non-null  int64
 18  Y_lat                        59076 non-null  float64
 19  X_lon                        59076 non-null  float64
 20  Georeference Column          59076 non-null  object
dtypes: float64(3), int64(2), object(16)
```

Figure 2.1: Information of Dataset

Now, we present the details of the statistical data in the dataset using:

```
Mortality_Rate_2015_17.describe()
```

	Year	Data_Value	LocationID	Y_lat	X_lon
count	59094.0	32230.000000	59094.000000	59076.000000	59076.000000
mean	2016.0	344.989392	30922.418824	37.901745	-91.406452
std	0.0	142.543952	16737.634283	6.326395	15.942302
min	2016.0	0.000000	0.000000	-14.301754	-170.719474
25%	2016.0	247.100000	18133.000000	34.326243	-98.129615
50%	2016.0	334.200000	29205.000000	38.220930	-89.931055
75%	2016.0	428.300000	46089.000000	41.695739	-82.890640
max	2016.0	3798.600000	78030.000000	69.309529	145.751259

Figure 2.2: Statistcal Data of Dataset

We are finding the number of unique values in dataset:

```
Mortality_Rate_2015_17.nunique()
```

```
Year                                1
LocationAbbr                       57
LocationDesc                     2023
GeographicLevel                     3
DataSource                          1
Class                               1
Topic                               1
Data_Value                       6182
Data_Value_Unit                     1
Data_Value_Type                     2
Data_Value_Footnote_Symbol          1
Data_Value_Footnote                 1
StratificationCategory1             1
Stratification1                     3
StratificationCategory2             1
Stratification2                     6
TopicID                             1
LocationID                       3283
Y_lat                            3282
X_lon                            3282
Georeference Column              3282
dtype: int64
```

Figure 2.3: Unique Data ColoumWise

Chapter 3. Data Cleaning

Data Cleaning is one of most crucial step of data preprocessing. It entails identifying and fixing mistakes, eliminating duplicates, deleting missing values, and creating a dataset that is uniform in quality. It enhances data's usefulness, correctness, and dependability. Inaccurate data can result in false insights or incorrect conclusions if thorough cleansing is not done. Standardizing formats, such as aligning time and date entries, fixing typos, and handling outliers that can distort results, is another aspect of data cleansing. In summary, clean data is the cornerstone of reliable analysis and guarantees that the findings are significant and useful.

In this data set, we analyze the missing values and outliers and try to clean it.

3.1 Missing data analysis

Missing values are one of the most common challenges in data analysis, which can greatly affect the reliability of insights. The knowledge of the degree and distribution of missing data will determine appropriate strategies for handling them, such as imputation or exclusion, to minimize bias and maintain data integrity.

In our dataset we find the missing values by:

```
Mortality_Rate_2015_17.isnull().sum()
```

```
Year                              0
LocationAbbr                      0
LocationDesc                      0
GeographicLevel                   0
DataSource                        0
Class                             0
Topic                             0
Data_Value                    26864
Data_Value_Unit                   0
Data_Value_Type                   0
Data_Value_Footnote_Symbol    32230
Data_Value_Footnote           32230
StratificationCategory1           0
Stratification1                   0
StratificationCategory2           0
Stratification2                   0
TopicID                           0
LocationID                        0
Y_lat                            18
X_lon                            18
Georeference Column              18
dtype: int64
```

Figure 3.1: Missing Values of our dataset

The dataset for 2015–2017 shows missing entries in several key columns. Notable gaps include Data_Value(26,864 missing values) and Data_Value_Footnote and Data_Value_Footnote_Symbol (32,230 each). Geospatial columns like Y_lat, X_lon, and Georeference Column have 18 missing entries each, while fields like Year and LocationAbbr have no missing data. These insights highlight areas requiring attention to ensure a robust analysis.

```
columns_to_keep = [ 'year', 'locationabbr', 'locationdesc',
'geographiclevel', 'data_value', 'data_value_unit',
'data_value_type', 'stratificationcategory1', 'stratification1',
'x_lon', 'y_lat']
Mortality_Rate_2015_17 = Mortality_Rate_2015_17[columns_to_keep]
Mortality_Rate_2017_19 = Mortality_Rate_2017_19[columns_to_keep]
Mortality_Rate_2019_21 = Mortality_Rate_2019_21[columns_to_keep]
```

This step filtered our datasets to only the most important columns in order to make our analysis more manageable. We sought the year, the geographic identifiers as concerning location abbreviation, location description, and geographic level, and the mortality rate itself, its unit, and its type (`data_value`). We also preserved stratification (`stratificationcategory1` and `stratification1`) and coordinates (`x_lon` and `y_lat`) to keep the geographic focus. This brought down the columns to only the columns that mattered, filtering out unwanted data noise and significantly making the datasets much more workable and concentrated for further analysis. This was done uniformly across the datasets of the three periods: 2015–2017, 2017–2019, and 2019–2021.

To ensure clarity and uniformity across the datasets for the periods 2015–2017,

2017–2019, and 2019–2021, the column names were standardized. This step simplified the interpretation and analysis of the data by making the naming conventions more intuitive. The changes made by renaming them are are as follows:

- `locationabbr` → `state_abbr`
- `locationdesc` → `state_or_county`
- `data_value` → `mortality_rate`
- `data_value_unit` → `rate_unit`
- `data_value_type` → `rate_type`
- `stratificationcategory1` → `demographic_category`
- `stratification1` → `demographic_value`
- `x_lon` → `longitude`
- `y_lat` → `latitude`

These updates ensure that the column names are more descriptive and consistent, making the datasets easier to navigate and analyze across all time periods.

Now we check the missing values in the dataset after data filtering are:

```
Missing Values Before Cleaning:
year                      0
state_abbr                0
state_or_county           0
geographiclevel           0
mortality_rate        26864
rate_unit                 0
rate_type                 0
demographic_category      0
demographic_value         0
longitude                18
latitude                 18
dtype: int64
```

Figure 3.2: Missing Values after the Data Filtering

3.2 Removing & Imputation

As We have many missing values in our dataset, we have to clean the data before moving forward. That can be done by droping rows and imputation.

```
    critical_columns = ['mortality_rate', 'state_or_county',
        'longitude', 'latitude']
    Mortality_Rate_2015_17 =
        Mortality_Rate_2015_17.dropna(subset=critical_columns)
    Mortality_Rate_2017_19 =
        Mortality_Rate_2017_19.dropna(subset=critical_columns)
    Mortality_Rate_2019_21 =
        Mortality_Rate_2019_21.dropna(subset=critical_columns)
```

DataFrames Mortality_Rate_2015_17, Mortality_Rate_2017_19, and Mortality_Rate_2019_21 have datasets of mortality rates varying over time. The principal columns are mortality_rate, state_or_county, and longitude and latitude, which have the most meaningful metrics in the study's analysis.

Rows of each dataset that had any of these critical columns containing missing values were removed by using the **dropna()** function. This removes the possibility of incomplete records from being included in the datasets and reduces the chances of compromising the quality and reliability of the analysis. Since we focused on these critical variables, we ensured we kept a robust dataset for further explorations and avoided possible problems that might arise from such incomplete or inconsistent data.

This is very critical in the analysis that depends on current and correct geographic information and rate of mortality to identify trends, disparities among others. Elimination of missing rows ensures that all analyses are conducted based on valid and complete data. This may therefore enhance the accuracy or the clarity of findings.

```
Mortality_Rate_2015_17['demographic_value'].fillna('Not␣Specified')
Mortality_Rate_2017_19['demographic_value'].fillna('Not␣Specified')
Mortality_Rate_2019_21['demographic_value'].fillna('Not␣Specified')
```

At this stage, we corrected missing values in the demographic_value column for the datasets from 2015–2017, 2017–2019, and 2019–2021. The placeholder value *Not Specified* was used to fill in the missing values in this column. This approach ensures that no data point is left missing or ambiguous, maintaining consistency and interpretability in the datasets. By filling up these gaps with a standard number, we reduce the likelihood of errors or misinterpretations during analysis and ensure that all demographic groups are represented, even when specific information was not initially recorded. This stage is crucial to getting the data ready for accurate and perceptive analysis.

```
Mortality_Rate_2015_17['mortality_rate'] = pd.to_numeric(
    Mortality_Rate_2015_17['mortality_rate'], errors='coerce')
```

In this stage, we made sure that the 2015–2017 dataset's mortality_rate column was changed to a numeric data type. A technique that automatically handles any non-numeric values by forcing them into NaN was used to accomplish this. This column's conversion to numeric guarantees uniformity and enables precise mortality rate computations and analysis. We can proceed with numerical operations and statistical evaluations with confidence if any inconsistencies in the data format are addressed.

```
Mortality_Rate_2015_17 = Mortality_Rate_2015_17.dropna(
                         subset=['mortality_rate'])
Mortality_Rate_2017_19 = Mortality_Rate_2017_19.dropna(
                         subset=['mortality_rate'])
Mortality_Rate_2019_21 = Mortality_Rate_2019_21.dropna(
                         subset=['mortality_rate'])
```

In this stage, we eliminated rows from the 2015–2017, 2017–2019, and 2019–2021 datasets that had missing values in the mortality_rate column. We made sure that only complete and legitimate data points were kept for analysis by removing these rows. This step is crucial because inaccurate or inconsistent findings could arise from missing values in a crucial column, such as mortality_rate. We can preserve the integrity and dependability of our analysis by concentrating on complete data.

3.3 Data After Cleaning

After the data Cleaning, this is how the dataframe looks

Cleaned Data Preview:

	year	state_abbr	state_or_county	geographiclevel	mortality_rate	rate_unit	rate_type	demographic_category	demographic_value	longitude	latitude
0	2016	CT	Connecticut	State	232.0	per 100,000 population	Age-adjusted, 3-year Average Rate	Gender	Male	-72.7254	41.6179
1	2016	IN	Indiana	State	202.0	per 100,000 population	Age-adjusted, 3-year Average Rate	Gender	Male	-86.2757	39.9128
3	2016	AZ	Arizona	State	282.2	per 100,000 population	Age-adjusted, 3-year Average Rate	Gender	Overall	-111.6640	34.2921
4	2016	AR	Arkansas	State	343.0	per 100,000 population	Age-adjusted, 3-year Average Rate	Gender	Female	-92.4340	34.8982
5	2016	FL	Florida	State	222.1	per 100,000 population	Age-adjusted, 3-year Average Rate	Gender	Female	-82.4970	28.6588
6	2016	ID	Idaho	State	243.0	per 100,000 population	Age-adjusted, 3-year Average Rate	Gender	Female	-114.6610	44.3858
7	2016	IA	Iowa	State	136.1	per 100,000 population	Age-adjusted, 3-year Average Rate	Gender	Overall	-93.5022	42.0760
8	2016	AK	Alaska	State	260.9	per 100,000 population	Age-adjusted, 3-year Average Rate	Gender	Overall	-152.5710	64.3173
9	2016	IA	Iowa	State	407.2	per 100,000 population	Age-adjusted, 3-year Average Rate	Gender	Male	-93.5022	42.0760
10	2016	AZ	Arizona	State	273.4	per 100,000 population	Age-adjusted, 3-year Average Rate	Gender	Male	-111.6640	34.2921

Figure 3.3: DataSet After Data Cleaning

Now the information of the dataset is also changed which given as

```
Data columns (total 11 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   year                  32212 non-null  int64
 1   state_abbr            32212 non-null  object
 2   state_or_county       32212 non-null  object
 3   geographiclevel       32212 non-null  object
 4   mortality_rate        32212 non-null  float64
 5   rate_unit             32212 non-null  object
 6   rate_type             32212 non-null  object
 7   demographic_category  32212 non-null  object
 8   demographic_value     32212 non-null  object
 9   longitude             32212 non-null  float64
 10  latitude              32212 non-null  float64
dtypes: float64(3), int64(1), object(7)
```

Figure 3.4: Dataset's Information after Cleaning

As our process of data cleaning is done so we proceed to **Visualization** of the dataset

Chapter 4. Trends Analysis

Trends analysis is an examination of the change in heart disease mortality rates over the three periods: 2015-2017, 2017-2019, and 2019-2021. This analysis determines whether the mortality rates are increasing, decreasing, or stable across these intervals. It also brings out the significant patterns, such as differences based on regions or demographic groups.

Visualizations add to this analysis by presenting complex data in a clear and interpretable format. Line graphs show a change over time, thereby allowing easy identification of trends: for example, gradual increases or sharp decreases. Heatmaps take on a geographical perspective, showing areas that are higher or lower in mortality rate. Bar charts compare rates across different subgroups or time periods, emphasizing the magnitude of change. These visual tools together help to effectively communicate data insights so that stakeholders can quickly get a grasp of the key findings for policy-making or further research.

4.1 Average Mortality Rate

In the statistics titled **National Average Heart Disease Mortality Rate (2015-2021)**, average mortality rates were recorded from heart disease within three intervals: namely, 2015-2017, 2017-2019, and 2019-2021 periods. The vertical axis consists of the average mortality rate, while horizontal one is divided into the time periods. The recorded statistics show that the heart diseases' mortality rates have not changed that much throughout the years except for an increase from the first period (2015-2017) to the last (2019-2021). The increase is, however, gradual and indicates a disturbing trend about the effects of heart disease on public health.

This upwards trend, though very gradual, also calls for more efforts at making heart disease a focus of increased attention. Associated factors include lifestyle, namely physical inactivity; limited access to health care; and related indices, including obesity and hypertension; and more. For some of the reasons, development of proactive public health measures, linking promotion and prevention care to wellness, and heightened awareness of one's cardiovascular wellness will become inevitable.

The graph demonstrates the critical importance of sustained action to reduce heart disease mortality, as well as to bolster health outcomes across the national population.

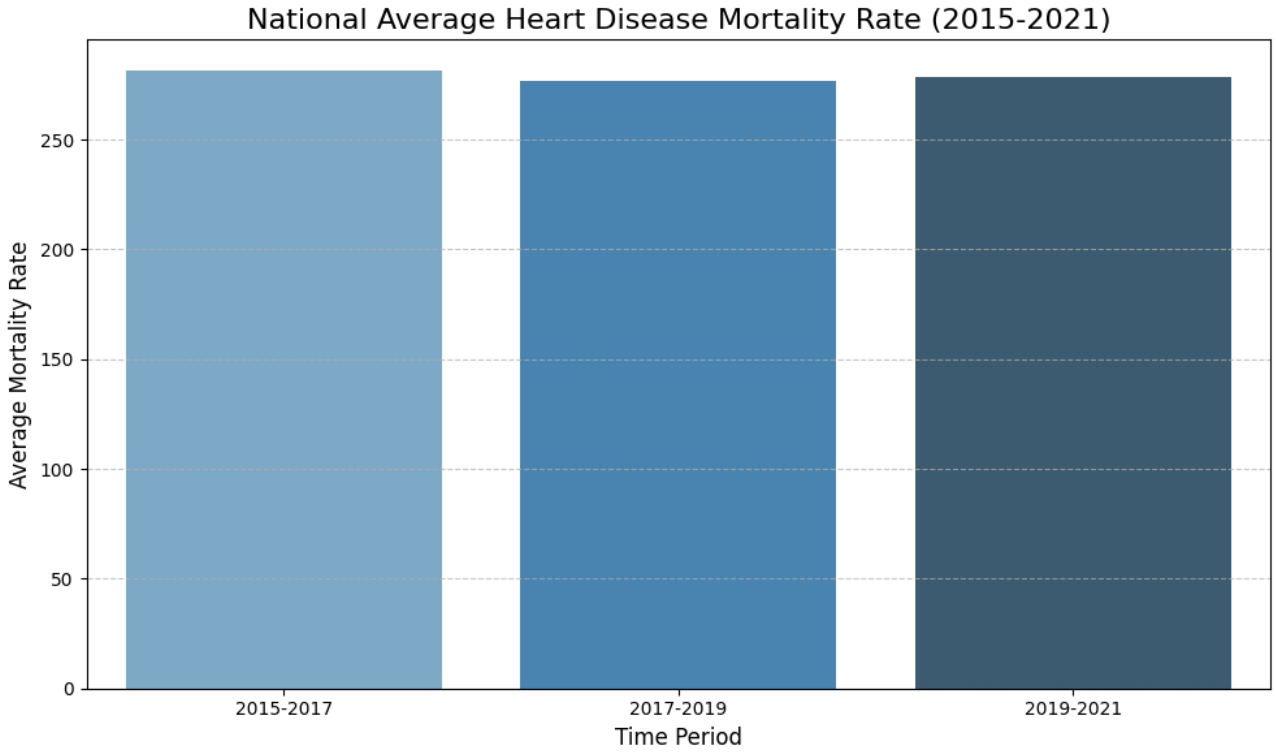


Figure 4.1: Average Mortality Rate over the year

"State-Level Heart Disease Mortality Trends (2015-2021)" shows the differences in heart disease death rates in the states and territories of the United States for the following three periods: 2015–2017, 2017–2019, and 2019–2021. The horizontal axis shows the three time periods, and the vertical axis shows the mortality rate per 100,000 people. Each line shows trends in the death rates of a particular state or region.

Significant variations in mortality patterns between states are depicted in the graph. Some states have declining rates, while others either stay the same or see rises over time. The differences in heart disease outcomes are highlighted by the unusually high or low rates in some states. These variations most likely stem from regionally specific lifestyle choices, socioeconomic status, and healthcare quality. To successfully reduce heart disease mortality, addressing these disparities would necessitate region-specific initiatives and customized public health activities.

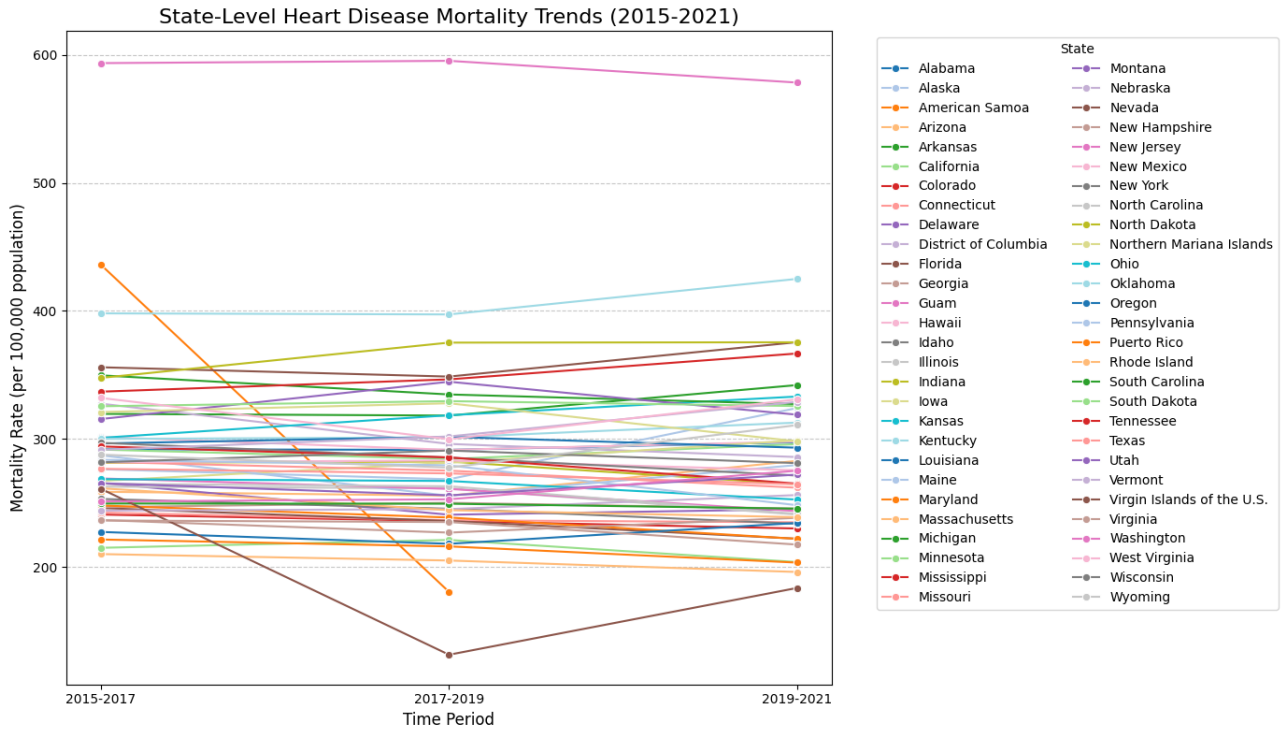


Figure 4.2: State-wise Heart Disease Moratality

4.2 Top and Bottom States Mortality Rate

Top 5 States with highest mortality Rate

The below graph provides the top 5 highest mortality rates for U.S. states from 2015-2021. Guam ranked the highest with its mortality rate increasing from a range of 550 deaths per 100,000 residents in 2015-2017 to more than 600 per 100,000 in 2019-2021. American Samoa, Oklahoma, Nevada, and Michigan had high, increasing mortality rates during this period as well. North Dakota and Mississippi showed lower but still elevated mortality rates compared to the national average.

The key takeaways are the constantly high mortality rates both in Guam and American Samoa and, to a lesser extent Oklahoma, and the upward trend for most of the top 5 states from 2015 through 2021.

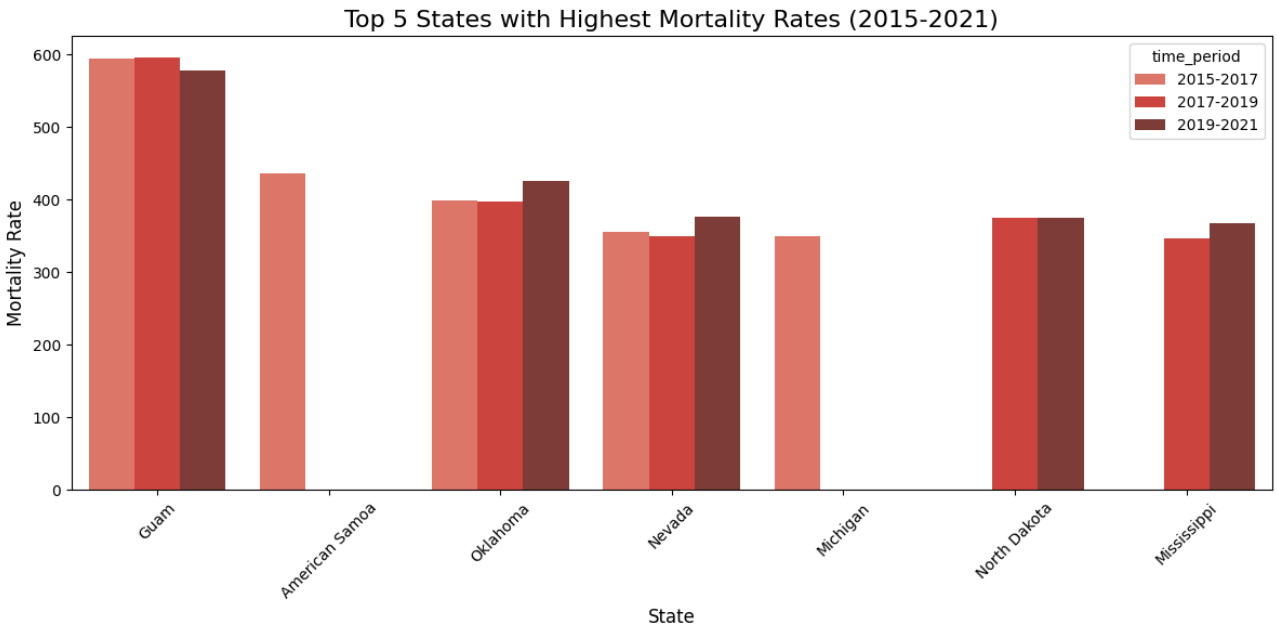


Figure 4.3: Top 5 States with Highest Mortality

Bottom 5 States with Lowest Mortality Rate

The graph shows the 5 states with the lowest death rates from 2015 to 2021. Massachusetts had the lowest rate, which was about 180 deaths for every 100,000 people. Minnesota, Puerto Rico, Oregon, and Virginia also had low and steady death rates compared to the national average.

The highest number of death rate for the group was at the Virgin Islands of the United States. It still however was lower compared to the top 5 states having the highest death rates. Main point is that the group of 5 states and territories are one of the lowest in death rates throughout the country between the period of 2015 up to 2021 and Massachusetts is at the best.

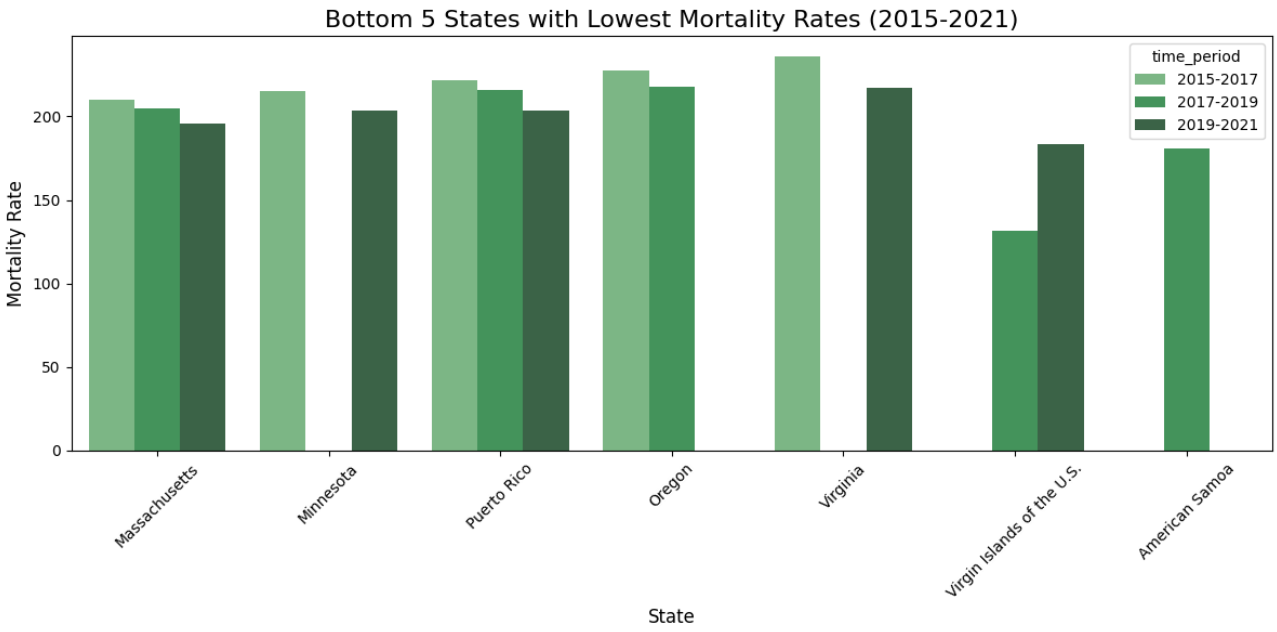


Figure 4.4: Bottom 5 States with Lowest Mortality

4.3 Increase and Decrease of Mortality Rates

Top 10 States with Largest Increase in Mortality Rate

The Below graph shows the top 10 states with the largest increases in mortality rates from 2015-2021. Alaska had the highest increase at over 15%. Vermont, Iowa, Kansas, and Washington also saw significant mortality rate increases over this time period. The states with lower but still notable increases include Arizona, Mississippi, Wyoming, North Dakota, and Arkansas.

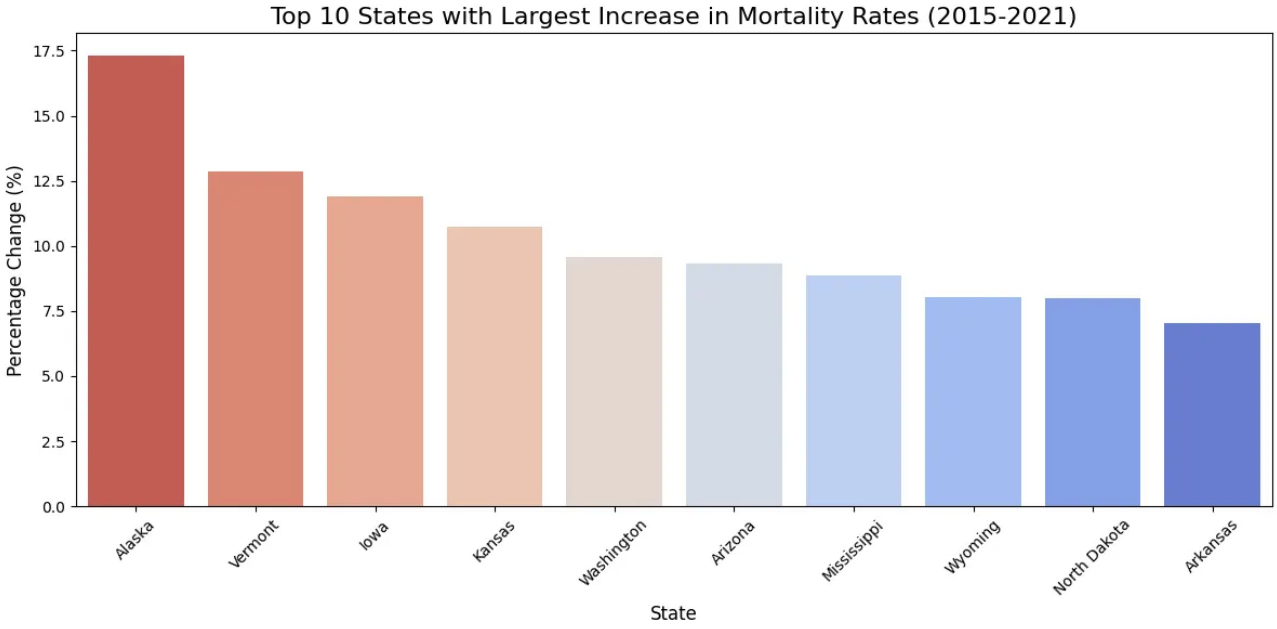


Figure 4.5: Top 10 States in Largest Increase In Mortality

Top 10 States with Largest Decrease in Mortality Rate

From the graph, the top 10 states with the biggest falls in the death rate from 2015 to 2021 have been identified. The biggest drop fell in Rhode Island, while having a fall of more than 25%. In this time, North Carolina, New Jersey, Florida, Tennessee, Maryland, Pennsylvania, the District of Columbia, and the Virgin Islands all saw big declines in death rates. These states arise for the ability of lowering their mortality rates from when comparisons are made with rising trends identified in other parts of the country.

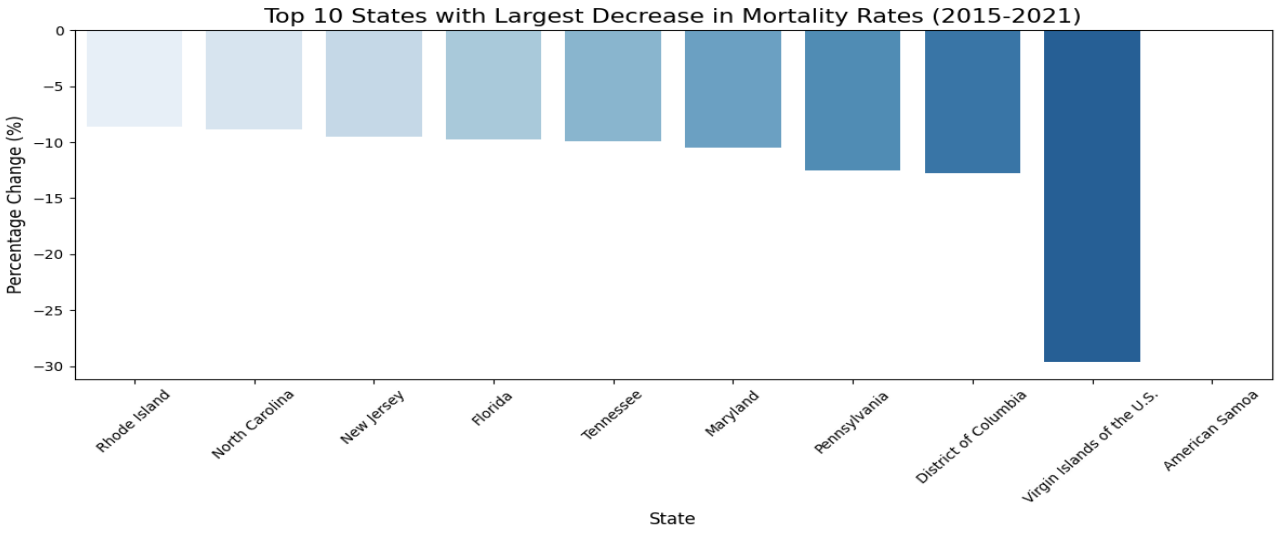


Figure 4.6: Top 10 States in Largest Decrease In Mortality

4.4 Gender-Wise Mortality Trends

Gender-Wise Trends of Heart Disease Mortality Trends

The graph of the death trends of heart disease from 2015 to 2021 in men and women provides visual evidence that overall death rates were roughly constant, shifting by about 335 per 100,000 people.

Male death rate rose slowly from about 415 per 100,000 in 2015 to 435 per 100,000 in the year 2021, while the death rate for female was more constant and changed between 275 and 290 per 100,000 in the same period.

The gap in the death rates of heart diseases between males and females increased slightly, but the rate among males remained higher and rising compared to females. Nonetheless, for both males and females, the death rates were at small changes from the years 2015 to 2021.

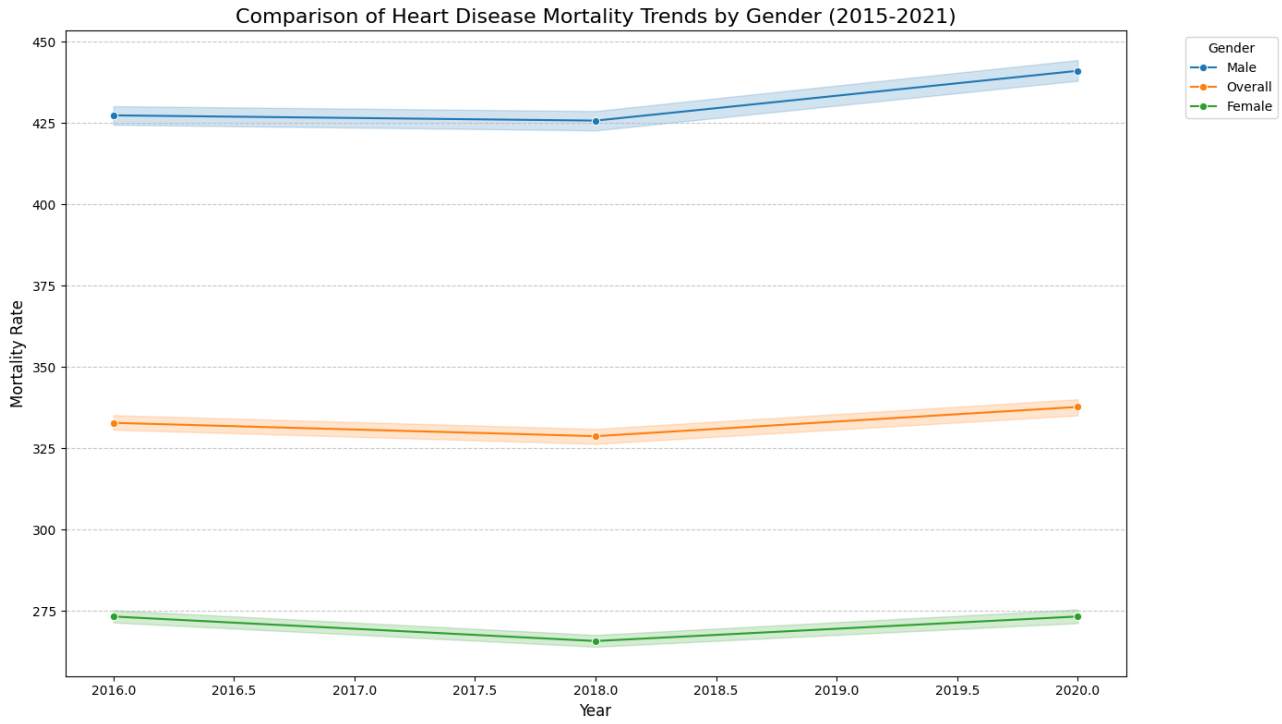


Figure 4.7: Gender-wise Trends of Heart Diease Mortality over Year

Gender-Wise Percentage Change of Heart Disease Mortality Trends

Heart disease death rates for men and women from 2015 to 2021 are shown in the graph:. Generally, the period increased the mortality rate by around 2.9%. Mortality in the female population was about up 2.2%, but the male population saw deaths increase by roughly 3.5%. The gender gap had increased slightly, since more men record

the increased deaths for the disease as compared to women. This gives an indication that the death rates have been stronger for men compared to women between 2015 and 2021.

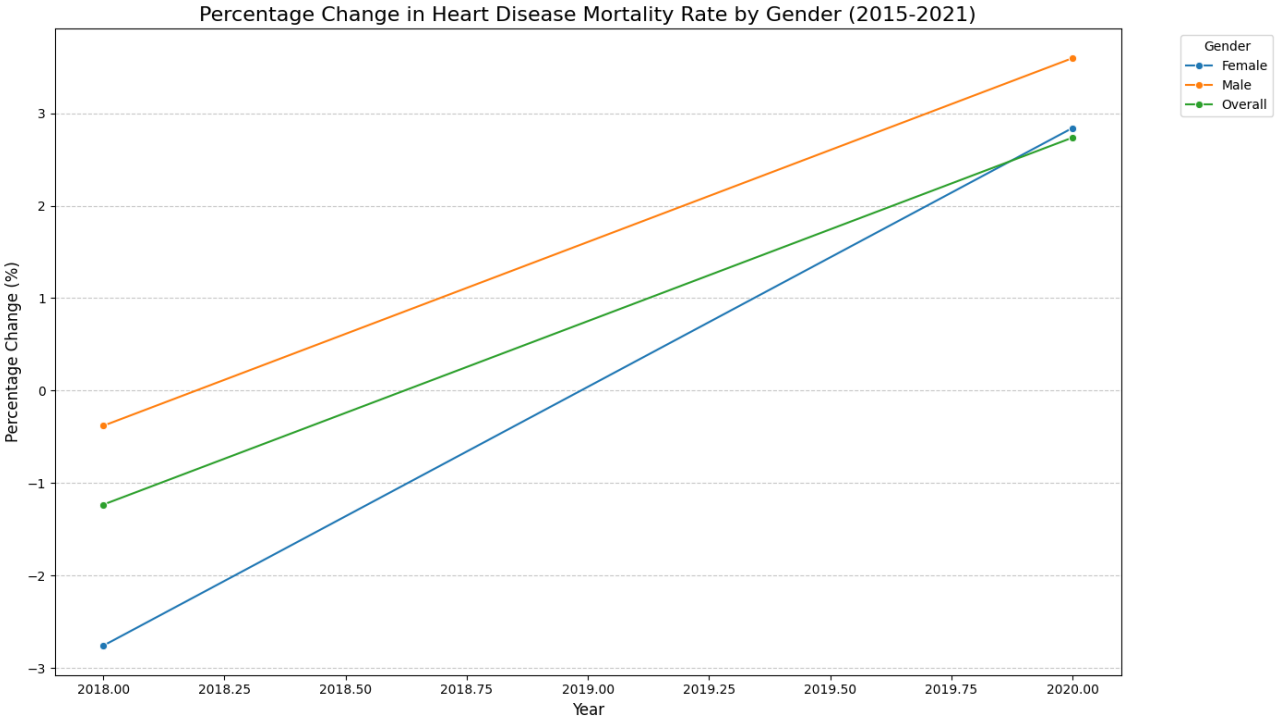


Figure 4.8: Percentage Change in Heart Disease Mortality Gender Wise

Chapter 5. Feature Engineering

For Exploratory Data Analysis projects, feature engineering is an essential part of the data pretreatment pipeline. It entails turning unstructured data into valuable attributes that can improve prediction models' performance. Enhancing the relevance of data in forecasting the target variable is the main objective of feature engineering, which increases the efficacy of machine learning algorithms.

5.1 Feature extraction

Feature extraction involves creating new features or transforming existing ones to enhance the dataset's predictive power. In this analysis, several techniques were applied to derive meaningful features:

1. **Deriving New Variables:** Certain attributes were transformed into categorical features to capture specific relationships. For instance, continuous variables like age or cholesterol levels might be discretized into categories (e.g., age groups or risk levels).

2. **Encoding Categorical Features:** Features such as gender or chest pain type were encoded into numerical values using one-hot encoding or label encoding. This ensured compatibility with machine learning models.

3. **Handling Missing Values:** Any missing data in critical features was imputed using statistical methods such as mean, median, or mode imputation to ensure the integrity of the dataset.

4. **Feature Scaling:** Continuous variables were normalized or standardized to ensure uniform scaling across features. This is particularly important for algorithms sensitive to the magnitude of input variables, such as gradient descent-based models.

The extracted features provided a comprehensive representation of the underlying dataset, improving its suitability for predictive modeling.

5.2 Feature selection

Feature selection is crucial for identifying the most relevant features, reducing dimensionality, and avoiding overfitting. In this analysis, several methods were utilized:

1. **Correlation Analysis:** The pairwise correlation between features was calculated. Features exhibiting high multicollinearity were identified and either transformed or removed to prevent redundancy.
2. **Statistical Tests:** Features were assessed for their significance in predicting the target variable using techniques like chi-square tests, ANOVA, or p-value thresholds. Features with low statistical significance were excluded.
3. **Feature Importance Metrics:** Tree-based algorithms such as Random Forests were used to rank features based on importance scores. Features with negligible contributions were excluded from the final dataset.
4. **Dimensionality Reduction:** Principal Component Analysis (PCA) was considered to reduce the dataset to its most informative components, particularly for high-dimensional data.

These methods ensured that the dataset retained only the most predictive features, optimizing model performance while minimizing computational complexity.

Chapter 6. Model Prediction

The aim of this analysis is to predict mortality rates for various demographics and geographic categories using historical data. The model leverages a Random Forest Regressor for predictions and evaluates its performance using RMSE (Root Mean Square Error). Insights are generated to identify groups with the highest mortality rates.

6.1 State-Wise Predictions

Using the trained Random Forest Regressor, mortality rates for the 2019–2021 dataset were predicted for each state. To analyze the predictions, the average predicted mortality rate for each state was calculated by grouping the data by the `state_or_county` attribute. The results were then sorted to identify the top 10 states with the highest predicted mortality rates and the bottom 10 states with the lowest predicted rates. This provided a clear understanding of the geographic variations in predicted mortality rates.

6.2 Visualizing Predicted Mortality Rates

Bar plots were generated to visually represent the top 10 states with the highest and lowest predicted mortality rates. These visualizations highlight the disparities across states:

- The **Top 10 States with Highest Mortality Rates** were visualized using a red color palette (`Reds_d`), emphasizing the severity in these regions.
- Conversely, the **Top 10 States with Lowest Mortality Rates** were illustrated with a blue color palette (`Blues`), showcasing regions with better health outcomes.

These visualizations provide actionable insights for policymakers to target resources and interventions effectively.

6.3 Model Evaluation

The predictive performance of the model was evaluated using several key metrics:

- **Root Mean Squared Error (RMSE):** The RMSE for the test dataset was computed as 129.4258. This metric measures the average deviation of the predicted mortality rates from the actual values, with a lower value indicating better performance.
- **Mean Absolute Error (MAE):** The MAE, calculated as 104.2397, reflects the average absolute difference between predicted and actual values, providing a more interpretable measure of error.
- **R-squared (R^2):** The R^2 score was 0.8236, indicating that the model explains approximately 82.36% of the variance in the mortality rate data. This suggests strong predictive power.

To further validate the model, cross-validation techniques were employed to assess its stability across different data splits.

6.3.1 Insights and Recommendations

The analysis identified significant geographic and demographic disparities in predicted mortality rates:

- States with the highest predicted mortality rates should be prioritized for health interventions and resource allocation.
- States with lower predicted rates can serve as benchmarks for identifying effective healthcare policies and practices.

Overall, the model demonstrates reliable performance and provides actionable insights into mortality trends across different states and demographics.

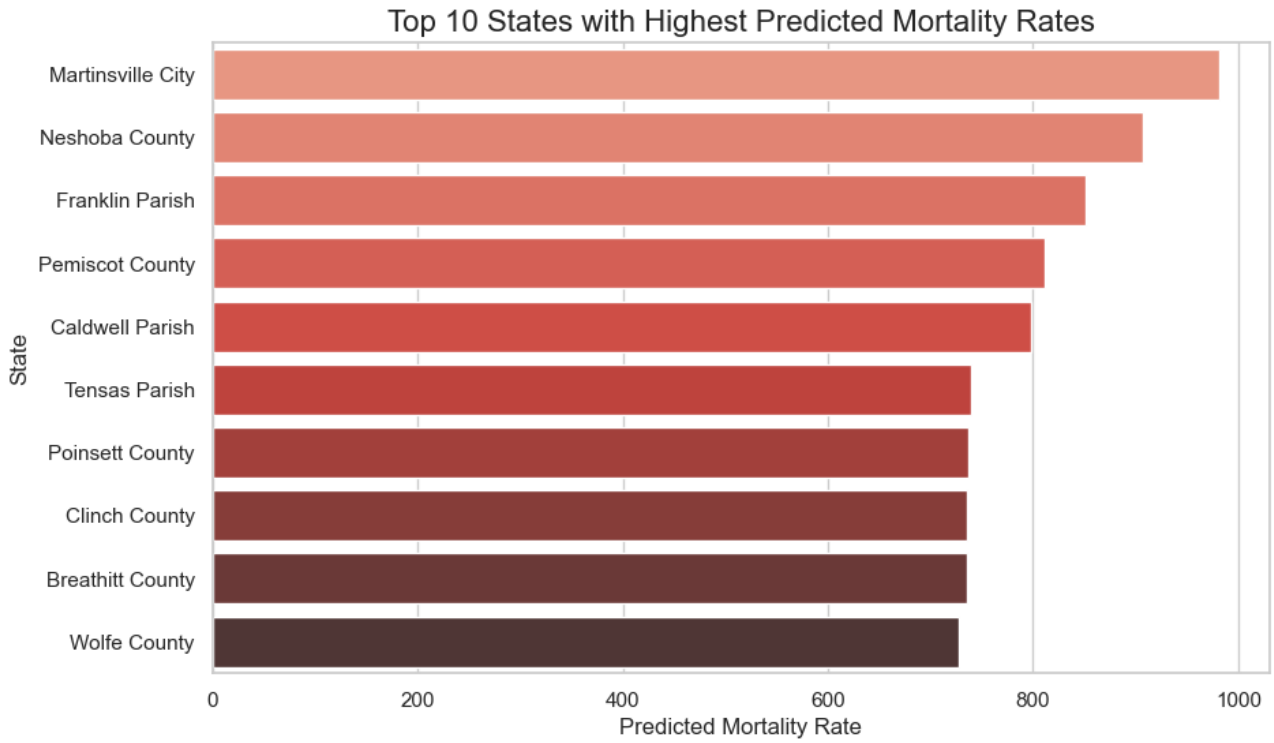


Figure 6.1: Top 10 States with Highest Predicted Mortality Rates

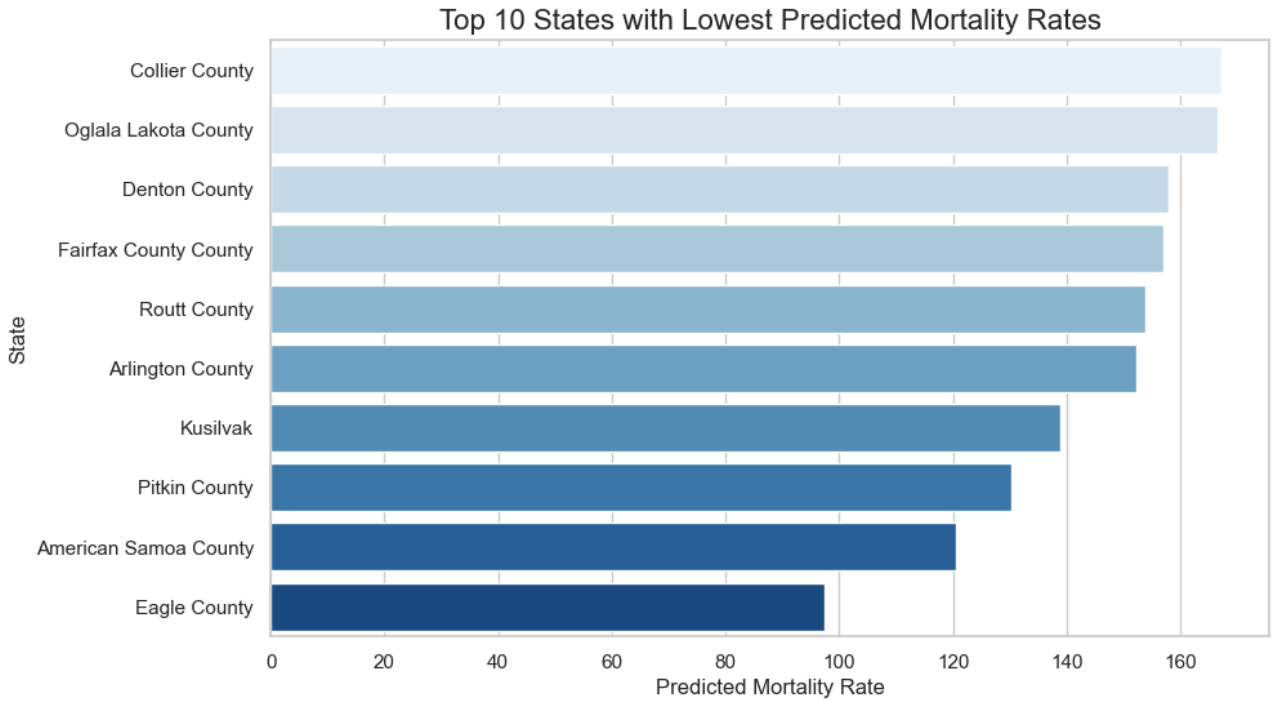


Figure 6.2: Top 10 States with Lowest Predicted Mortality Rates

Chapter 7. Conclusion & future scope

The analysis aimed to predict future heart disease mortality rates based on historical trends from the years 2015-2017, 2017-2019, and 2019-2021. Here are the key findings from the analysis:

7.1 Data Trends:

By combining data from three consecutive time periods, we observed how heart disease mortality rates changed across various states or counties, demographic categories, and geographical levels.

We included multiple features such as year, demographic categories, and geographical data, which were crucial in understanding how these factors affect mortality rates over time.

7.2 Feature Importance:

Using the Random Forest Regressor, the model was able to capture complex relationships between the features and mortality rates, with particular importance given to demographic categories and geographic areas.

The one-hot encoding of categorical features (like state and demographic values) and the scaling of numerical data (like year, longitude, latitude) allowed the model to effectively process and learn from these mixed data types.

Prediction Model:

The Root Mean Squared Error (RMSE) metric indicated that the model provided a good fit to the training data and was able to generalize reasonably well. The model's predictions for the 2019-2021 dataset were used to forecast future mortality rates for the years 2022-2026, based on observed trends.

7.3 Future Mortality Rate Predictions:

The future mortality rates for 2022-2026 were predicted by extending the trends seen in 2015-2021. These predictions highlight potential increases or decreases in mortality rates depending on various demographic and geographical factors.

The predicted mortality rates provide insights into the expected public health situation, allowing policymakers to make data-driven decisions.

7.4 Measure to Conserve:

If the trends continue as predicted, regions with higher mortality rates may need more targeted interventions in healthcare infrastructure, education, and prevention programs. The analysis also underscores the importance of monitoring changes over time, as shifts in demographics or geographic areas could lead to significant changes in heart disease mortality trends. Additionally, demographic factors such as age, gender, and socio-economic status might influence the future trends, which can be further explored in future analyses.

Group Contribution

Hitarth Bhatt

Helped in the report and figured out the data visualization and data cleaning part of the project. Helped in finding the dataset

Tirth Modi

Helped in the report and did the model fitting and feature engineering part of the project, helped in finding the dataset

Heet Dipeshe

Helped in the report and did the data part of the project, and helped in finding out the datasets

References

[1] Heart Disease Mortality Data Among US Adults (35+) by State/Territory and County- 2015-2017
URL: https://catalog.data.gov/dataset/heart-disease-mortality-data-among-us-adults-35-by-state-territory-a

[2] Heart Disease Mortality Data Among US Adults (35+) by State/Territory and County – 2017-2019
URL: https://catalog.data.gov/dataset/heart-disease-mortality-data-among-us-adults-35-by-state-territory-a

[3] Heart Disease Mortality Data Among US Adults (35+) by State/Territory and County – 2019-2021
URL: https://catalog.data.gov/dataset/heart-disease-mortality-data-among-us-adults-35-by-state-territory-a

[4] GitHub Link for datasets: *URL:* https://github.com/HappyJoddd/EDA_Project_Heart_dieases/tree/main