IST 652

SCRIPTING FOR DATA ANALYSIS

FINAL PROJECT REPORT

CHRONIC DISEASE DATA ANALYSIS

"Analyzing Chronic Disease Indicators in the United States: Trends, Disparities, and Insights"

Presented By

Tharuni Tekula
Yaswanth Lalpetvari
Sindy Siddharth Reddy Kolli
Kushwanth Sai Chandu Meesala

Under the guidance of
Adjunct Prof. Hernando Hoyos

Table of Contents

Table of Figures

Project Description:

This project aims to to examine trends in chronic diseases throughout the United States using the U.S. Chronic Disease Indicators dataset. Finding regional and demographic differences in the prevalence of chronic diseases including diabetes, cardiovascular disease, and asthma is the main goal of the investigation.

Finding significant insights that can inform public health policies is the aim to guarantee that funds are distributed where they are most needed and that initiatives are designed to have a significant impact. This study aims to draw attention to health disparities to identify ways to build a more fair and healthy future.

Data Sources:

CDC's Centers for Disease Control and Prevention portal

https://data.cdc.gov/Chronic-Disease-Indicators/U-S-Chronic-Disease-Indicators/hksd-2xuw/about_data

Data type – Structured, CSV file

Dataset Description:

- Offers thorough data on long-term problems like diabetes, asthma, heart disease, and alcohol-related disorders.
- Contains state-level data, gender, age, and death rate stratifications, among other measures.

Key Statistics:

Rows: 309,216
Columns: 34

Modifications from Initial Proposal:

Our effort initially aimed to investigate the relationship between air quality and chronic diseases by combining information from the EPA Air Quality Index (AQI) and the CDC Chronic Disease Indicators datasets. Analyzing the potential relationships between variables like AQI, ozone concentration, and PM2.5 levels and outcomes including disease prevalence, hospitalization rates, and mortality was our aim. We found that the correlation between these datasets was extremely weak. This revealed that a linear relationship between air quality and chronic diseases was not evident within the available data. We concluded that other elements,

such the state of the economy, access to healthcare, and other contextual impacts, would need to be considered in order to get insightful findings. This led us to concentrate exclusively on the Chronic Disease Indicators dataset. This dataset, which compiles information from reliable sources such as the US Cancer Statistics, Behavioral Risk Factor Surveillance System, United States Renal Data System, and American Community Survey (ACS), provided a wide range of features for insightful analysis.

Preprocessing Steps:

1. Data Cleaning:

Dropped Columns:

- Dropped columns, such as "StratificationCategory2" and "Response," that have no values at all.

Handling Missing Values:

- Numeric columns were filled with their median values.
- Median was used instead of the mean to avoid distortion caused by outliers in mortality rates.
- In categorical columns such as "Geolocation," "Unknown" was used for filling data.
- Guaranteed categorical completeness while avoiding errors.

2. Data Aggregation:

- To concentrate on state-specific trends, national-level entries were eliminated.
- For thorough statistical summaries, data was grouped by year, state, and disease type.

Methods of Analysis:

Questions to Answer:

- What are the trends in asthma and diabetes mortality rates over time?
- How do cardiovascular disease prevalence rates vary by gender and state?
- What are the geospatial patterns in alcohol-related mortality rates?

Analytical Techniques:

We assessed changes in the mortality rates for diabetes and asthma from year to year using trend analysis. To investigate variations in the prevalence of cardiovascular disease by state and gender, a comparative analysis was carried out. Because of Cartopy's sophisticated mapping features, alcohol-related mortality was mapped using geospatial visualization.

Python Program Overview

The project code was created to handle the dataset in a methodical manner, carry out preprocessing, analyze the data, and provide both static and interactive visualizations. To guarantee readability and maintainability, it was divided into discrete phases. An outline of the main elements of the code is provided below:

Data Loading:

The program started by using the pandas package to load the U.S. Chronic Disease Indicators dataset. This made it possible to manage the dataset's enormous size which included over 309,216 rows and 34 columns efficiently. To comprehend the structure and find missing values, the first exploration involved looking at the first few rows and examining the column data types.

Data Preprocessing:

To keep outliers from distorting the results, median values were used to fill in missing values in numeric fields. To make sure no data points were left out because of missing information, the placeholder value "Unknown" was entered into categorical columns like "Geolocation." To simplify the dataset, completely empty columns like "Response" and "StratificationCategory2" were found and eliminated. To focus on certain patterns and insights, data was aggregated by year, state, and disease type. By doing these actions, the dataset was guaranteed to be clean and prepared for analysis.

Analysis Modules:

1. **Trend Analysis:**

   From 2019 to 2021, the program assessed patterns in the mortality rates from diabetes and asthma. The algorithm discovered important changes, including a discernible peak in 2020, by aggregating the data by year and computing the average mortality rates. Matplotlib-created line plots that showed patterns over time were used to visualize this.

2. **Comparative Analysis:**

   The prevalence of cardiovascular disease by gender was compared by state. To determine average prevalence rates, the data was stratified by gender and aggregated by state. Male and female prevalence were compared using bar plots, which showed notable differences that necessitated focused health interventions.

3. **Geospatial Analysis:**

   Using the cartopy library, the program mapped alcohol-related mortality rates across the United States. By extracting latitude and longitude coordinates from the ‘Geolocation’ column, the program plotted mortality rates geographically. Cartopy was

chosen for its advanced projection capabilities, enabling accurate and visually appealing maps that highlighted regional disparities.

Visualization Tools:

Both static and interactive visualization methods were used in the code. Trends and comparisons were shown using static graphs made with matplotlib. Dash was used to create a dynamic dashboard with interactive visuals that let users investigate "Binge Drinking Prevalence Among Adults." The interface included a drop-down menu for choosing states, and line charts that updated dynamically to show patterns over time. Stakeholders were able to successfully understand the data and derive meaningful insights because to this interaction.

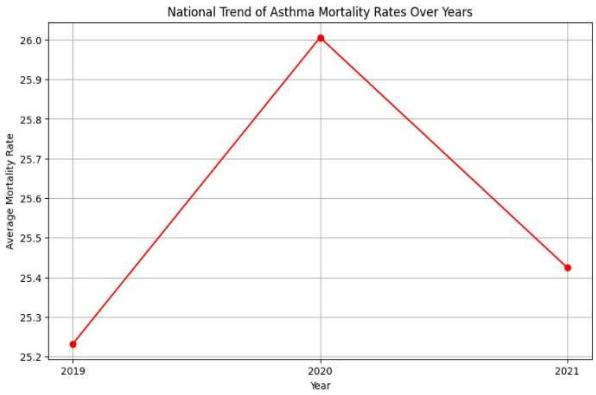
Advanced Libraries:
The program leveraged advanced Python libraries to enhance functionality:

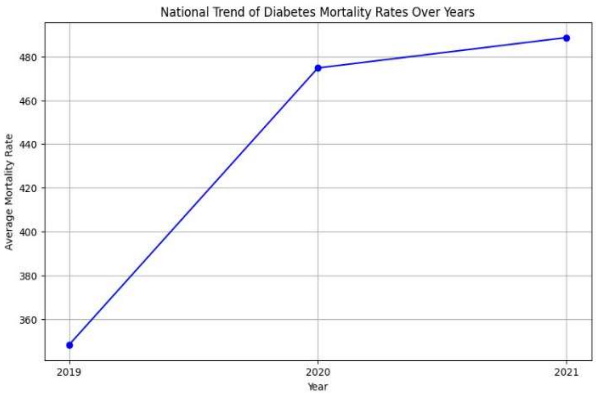
- **Pandas:** Used extensively for data loading, cleaning, aggregation, and manipulation.
- **Matplotlib:** Utilized for creating static visualizations, such as line and bar plots, to represent trends and comparisons.
- **Cartopy:** Chosen for geospatial mapping due to its flexibility and ability to handle geographical projections, surpassing simpler libraries like geopandas.
- **Dash:** Implemented for building the interactive dashboard, enabling real-time exploration of data trends.
- **Ipython.display**: Allows us to render rich HTML content, such as styled Data frames tables and enhancing the presentation and interactivity outputs.

Output Documentation:

Trends in Asthma & Diabetes Mortality (2019-2021):

The average asthma mortality rate in 2019 was 25.2, which serves as a baseline. A 2020 peak of 26.0 was followed by a slight fall to 25.3 in 2021, highlighting the COVID-19 pandemic and other potential external factors. The mortality rates from diabetes also showed similar patterns, rising from 360 in 2019 to 480 in 2020 and then marginally to 485 in 2021. A state-by-state examination showed that death rates were stable in Delaware and Alaska but high in California and New York.





Summary tables of Asthma and Diabetes Mortality Rates (2019-2021):

Asthma Mortality Rates Statewide Summary

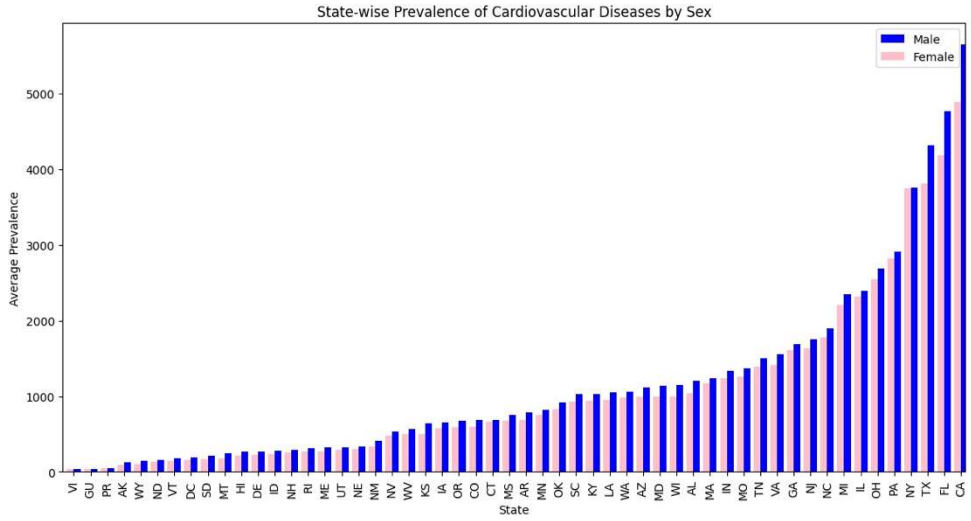
Year	2019	2020	2021
State			
Alabama	19.952778	19.680556	21.188889
Alaska	27.000000	27.000000	27.000000
Arizona	22.580556	24.619444	23.069444
Arkansas	24.194444	22.458333	24.347222
California	48.894444	56.597222	46.102778
Colorado	21.677778	21.661111	23.650000
Connecticut	21.527778	23.163889	25.772222
Delaware	27.000000	27.000000	27.000000
District of Columbia	27.000000	27.000000	27.000000
Florida	31.322222	34.638889	31.897222
Georgia	24.369444	24.686111	23.297222
Hawaii	25.533333	25.688889	25.502778
Idaho	27.000000	23.891667	27.000000
Illinois	30.388889	30.875000	25.363889
Indiana	21.233333	24.011111	23.491667
Iowa	21.897222	21.072222	23.058333
Kansas	24.047222	24.197222	24.416667

Diabetes Mortality Rates Statewide Summary

Year	2019	2020	2021
State			
Alabama	213.712500	298.495833	335.113889
Alaska	46.118056	67.680556	72.633333
Arizona	345.281944	504.169444	555.129167
Arkansas	214.475000	284.988889	299.466667
California	1897.420833	2611.287500	2759.597222
Colorado	290.169444	371.419444	409.144444
Connecticut	150.711111	201.538889	158.043056
Delaware	69.131944	108.122222	116.444444
District of Columbia	64.731944	106.694444	93.737500
Florida	1002.233333	1492.601389	1455.502778
Georgia	449.506944	600.926389	673.090278
Hawaii	99.243056	110.487500	116.731944
Idaho	131.879167	161.863889	185.043056
Illinois	501.800000	829.106944	713.816667
Indiana	412.330556	550.708333	558.127778
Iowa	233.588889	307.451389	294.265278
Kansas	173.954167	231.384722	216.658333
Kentucky	439.444444	516.100000	581.277778
Louisiana	258.809722	381.977778	381.552778

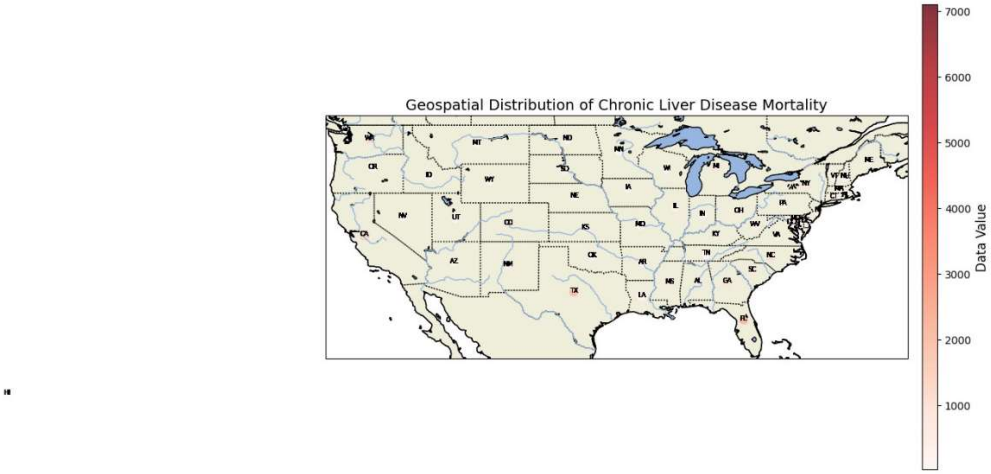
Cardiovascular Disease Prevalence:

- Higher prevalence in males, emphasizing the need for gender-specific strategies.
- States like California, Florida, and Texas require focused interventions due to high rates.



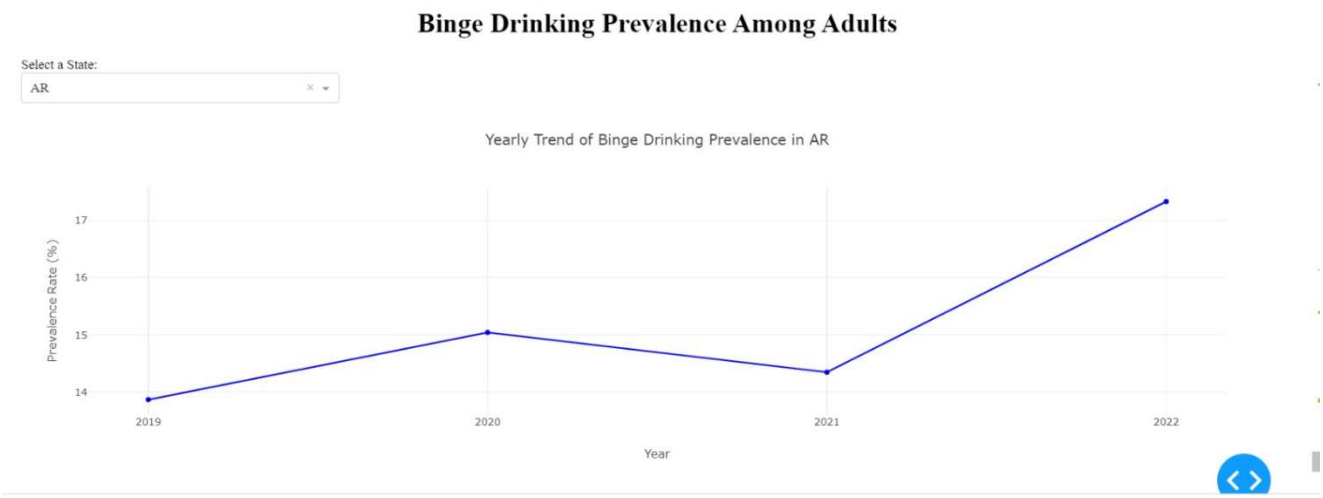
Alcohol-Related Mortality:

- High Mortality Areas: California, Texas, Florida.
- Regional Disparities: Southern and western states have higher rates.



Dash Implementation:

The creation of an interactive dashboard with Dash to examine the prevalence of binge drinking in adults was a key component of the project. The dashboard included a drop-down menu for choosing states and line charts that dynamically displayed trends. Users were able to examine trends by state and analyze data in real time using this interactive application. Users might choose a state, see the prevalence statistics for that state year-by-year, and learn about trends and inequalities, for instance. The project's objective of giving stakeholders the ability to interactively study data was perfectly embodied by the Dash application.



Conclusions:

Trends Analysis:

The mortality rates from diabetes and asthma vary significantly, peaking in 2020 most likely because of the COVID-19 pandemic.

Regional Insights:

States with continuously higher mortality rates, such as Florida and California, need for focused initiatives.

Gender Disparities:

Men are more likely than women to have cardiovascular problems, which emphasizes the necessity for specialized health interventions.

Geospatial Patterns:

Higher alcohol-related mortality rates in southern and western states point to variations in healthcare access.

Recommendations:

The report recommends focusing on gender-specific health policies and implementing targeted public health initiatives in high-mortality areas. Reaching out to high-risk individuals and closely monitoring changes in chronic conditions can help prevent future health disasters.

Team Contributions:

NAME	CONTRIBUTION
Kushwanth Sai Chandu Meesala	Data Collection, Data cleaning, Data Processing, and Geospatial visualization.
Tharuni Tekula	Cardiovascular disease analysis and gender-based comparisons, Code cleaning.
Sindy Siddarth Reddy Kolli	Dashboard creation and Interactive visualizations, Code comments.
Yaswanth Lalpet Vari	Asthma & Diabetes mortality rates trend analysis, PPT making.