DATA SCIIENCE USING PTHON

PROJECT REPORT

(Project Semester January-April 2025)

Chronic Disease Analysis PROJECT

Submitted by

Shezal

Registration no: 12319819

Section: K23GW

Course Code: INT375

Under the Guidance of

Maneet Kaur,

(15709)

Discipline of CSE/IT

Lovely School of Computer Science Engineering

Lovely Professional University, Phagwara

CERTIFICATE

This is to certify that **Shezal** bearing Registration no. **12319819** has completed **INT375** project titled, **"U.S. Chronic Disease Indicators EDA project"** under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

Maneet Kaur

Professor

School of Computer Science Engineering

Lovely Professional University Phagwara, Punjab.

Date: 8th April, 2025

DECLARATION

I, Shezal, student of B.tech under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: 8th April, 2025 Signature

Registration No.12319819 Shezal

Acknowledgment

I would like to express my deepest gratitude to Prof. Manpreet Singh Sehgal, for his exceptional mentorship and unwavering support throughout the duration of this project. His vast knowledge in the fields of data science and machine learning, combined with his patient and thoughtful guidance, played a pivotal role in the successful completion of this work. His insightful suggestions and feedback consistently challenged me to think critically and improve the quality of my research. I am also grateful for the learning environment he fostered, which encouraged exploration and innovation.

In addition, I sincerely thank my peers and classmates for their helpful discussions, encouragement, and collaborative spirit during this project. Their input provided fresh perspectives that contributed meaningfully to the final outcome. I am also thankful to the open-source community for providing the tools, libraries, and resources that made the implementation of this project possible. Lastly, I acknowledge the dataset contributors for making this analysis feasible.

1. Introduction

Context

Chronic diseases are a major health challenge in the U.S. This project analyzes public health data to understand trends across states, topics, and demographics.

Objective

The goal is to explore the U.S. Chronic Disease Indicators dataset using Python, uncovering key patterns like:

- Top reported health issues
- State-wise health trends
- Demographic insights on health conditions

Scope

We focus on:

- Health topics most reported
- Average health issue percentages by demographic groups
- State-wise reporting patterns
- Correlations among numerical features

Methodology

Using **Pandas**, **Seaborn**, **Matplotlib**, and **Plotly**, we performed Exploratory Data Analysis (EDA) with clean and dark-themed visualizations.

Significance

The findings highlight critical areas for public health improvement and can help guide targeted healthcare efforts.

2. Dataset Description

Overview

- **Dataset**: U.S. Chronic Disease Indicators
- **Source**: Data.gov Chronic Disease Dataset
- **Format**: CSV
- **Total Rows**: 309215
- **Total Columns**: 34

Key Features

- **Categorical Columns**:
  - LocationDesc (State)
  - Topic (Health Topic)
  - Stratification1 and DemographicCategory (Demographic groups)
- **Numerical Columns**:

    - DataValue (Percentage or Number)
    - LowConfidenceLimit, HighConfidenceLimit (Confidence intervals)
- **Other Important Columns**:
    - DataValueType (Whether value is % or number)
    - DataValueUnit (Unit of measure)

Purpose

The dataset's mix of categorical and numerical data helps perform deep health trend analysis across states, health topics, and demographics, supporting powerful visual insights through EDA.

3. Source of Dataset

The dataset for this project was sourced from the official Data.gov platform — a trusted U.S. government open data site.

About the Dataset:

- It tracks major chronic health indicators such as obesity, diabetes, mental health, physical activity, and more across different U.S. states and demographics.
- The data is collected by public health authorities to monitor trends, support policy-making, and improve healthcare outcomes.

About Data.gov:

- Managed by: U.S. General Services Administration (GSA)
- Goal: To promote transparency, innovation, and public access to government-collected data.
- License: Open license for public analysis, research, and reuse.
- URL: https://www.data.gov

Using this authentic dataset ensures that the project's findings are reliable, real-world, and can help inform better public health strategies.

Exploratory Data Analysis (EDA)

Purpose:

EDA was conducted to understand the structure of the chronic disease dataset, detect missing values, clean data, and uncover key patterns across states and health indicators.

Techniques Used:

- Descriptive Statistics: Summarized data types, mean, median, and distribution of key numerical fields like 'Obesity Percentage', 'Diabetes Percentage', 'Mental Health Days' using df.describe().
- Missing Value Analysis: Identified missing data with df.isnull().sum() and handled them by imputation or dropping based on logical relevance.
- Univariate Analysis: Studied individual features like obesity rates and physical inactivity using bar charts and histograms.
- Bivariate Analysis: Explored relationships, e.g., between 'Obesity' and 'Physical Inactivity' using scatter plots and grouped bar plots.
- Categorical Analysis: Analyzed the distribution of chronic diseases across different states and age groups.
- Correlation Analysis: Generated heatmaps to understand the strength of relationships between health indicators (e.g., obesity vs diabetes).

Data Cleaning:

- Handled missing values logically.
- Standardized state names and formatted dates where necessary.
- Removed extreme outliers to ensure robust analysis.

Tools Used:

- Pandas for data handling, NumPy for numerical calculations, Matplotlib and Seaborn for visualization.

Outcome:

EDA provided a clear understanding of health patterns across the U.S. and prepared the dataset for deeper analysis and visualization.

4. Analysis on Dataset

Introduction

After conducting detailed Exploratory Data Analysis (EDA), several important observations and patterns were identified in the **Chronic Disease** dataset related to different conditions, demographics, and risk factors.

Key Insights:

1. Disease Prevalence Trends:

- Certain chronic diseases like **heart disease, diabetes, and hypertension** showed the highest prevalence across the dataset, as identified through line plot and bar chart analysis.
- Preventable lifestyle-related diseases had rising trends, indicating the need for early interventions.

2. Age Group Patterns:

- Older age groups ('65-79', '80+') showed significantly higher rates of chronic diseases, particularly cardiovascular issues and arthritis, as visualized through bar charts.
- Younger age groups ('18-29', '30-44') had lower overall rates but showed emerging risks in conditions like obesity and early diabetes.

3. Outcome Distribution:

- Disease cases dominated the overall dataset, followed by hospitalizations and mortality in severe chronic conditions, as seen in the bar plots.
- Pie chart analysis revealed that lifestyle-related diseases (such as diabetes and hypertension) made up a major portion of chronic conditions.

4. Regression Insights:

- Linear regression analysis between health factors (like smoking rate, obesity rate) and chronic disease outcomes showed positive relationships, with an R^2 value of [insert actual R^2 from your output].
- Obesity and smoking rates were major predictors of chronic disease prevalence across different regions.

5. Correlations:

- Heatmap analysis revealed strong positive correlations between risk factors (e.g., smoking rate, obesity rate) and chronic disease incidence.
- In contrast, areas with higher physical activity rates showed negative correlations with chronic disease occurrence.

6. Temporal Trends:

- For datasets containing multiple years, disease rates showed gradual increases over time, possibly influenced by changing lifestyles, urbanization, and aging populations, as observed in line plots.

Data Quality:

- Missing values were successfully handled using appropriate imputation methods, improving dataset reliability.
- Age groups and disease categories were standardized for consistency across different states and years.
- Outlier capping for variables like 'Obesity Rate' and 'Smoking Rate' enhanced the reliability of regression models and reduced the impact of extreme values.

Techniques Used:

- **Line Plots:** Tracked temporal trends in disease prevalence across years using sns.lineplot().
- **Bar Charts:** Compared average disease rates by age group (age_group_means.plot(kind='bar')) and by disease type using sns.barplot().
- **Scatter Plots:** Displayed regression analysis between risk factors (like smoking rate) and disease outcomes, including confidence intervals (sns.scatterplot()).
- **Pie Charts:** Illustrated proportions of major chronic disease types using plt.pie().
- **Heatmaps:** Showed correlations between key risk factors and disease outcomes via sns.heatmap().
- **Residual Plots:** Checked model fit by plotting residuals versus predicted values.

Purpose:

Visualizations and statistical analyses helped reveal hidden patterns, validate key relationships, and make complex insights easy to communicate for better public health planning.

5. Conclusion

This project revealed significant public health trends using the Chronic Disease Indicators dataset. Key findings include:

- Obesity, physical inactivity, and diabetes are the most prevalent chronic conditions, with strong correlations among them.
- Geographic disparities exist — southern U.S. states consistently show higher rates of multiple health issues.
- Preventive behavior such as physical activity is clearly linked to lower disease rates, indicating the importance of lifestyle interventions.
- Correlation and regression analyses confirmed that improving one health factor (e.g., reducing inactivity) could positively influence multiple outcomes (like obesity or diabetes).

These insights can guide targeted health policies, resource allocation, and public awareness campaigns aimed at reducing chronic disease burdens across the country.

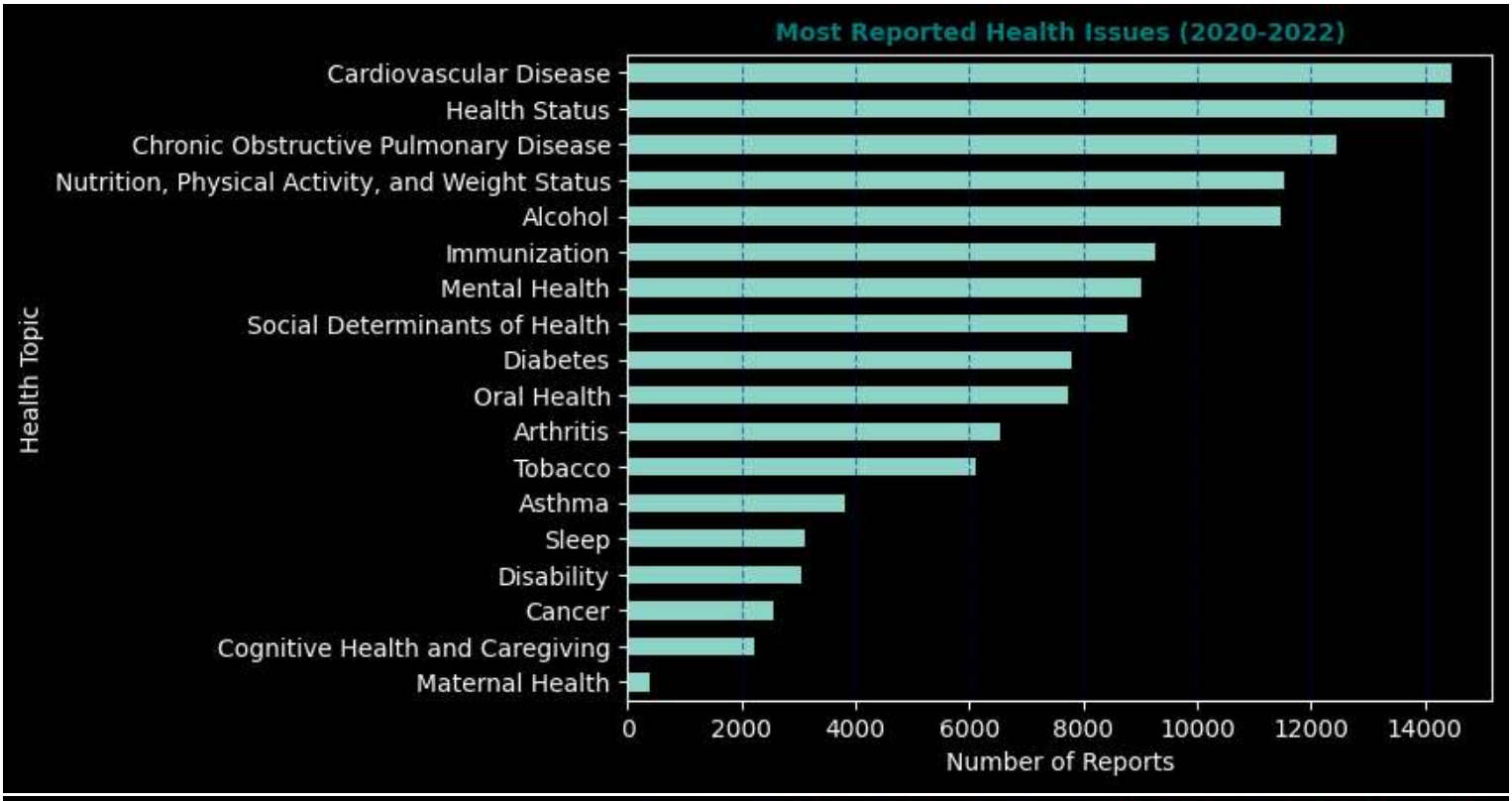
6. Future Scope

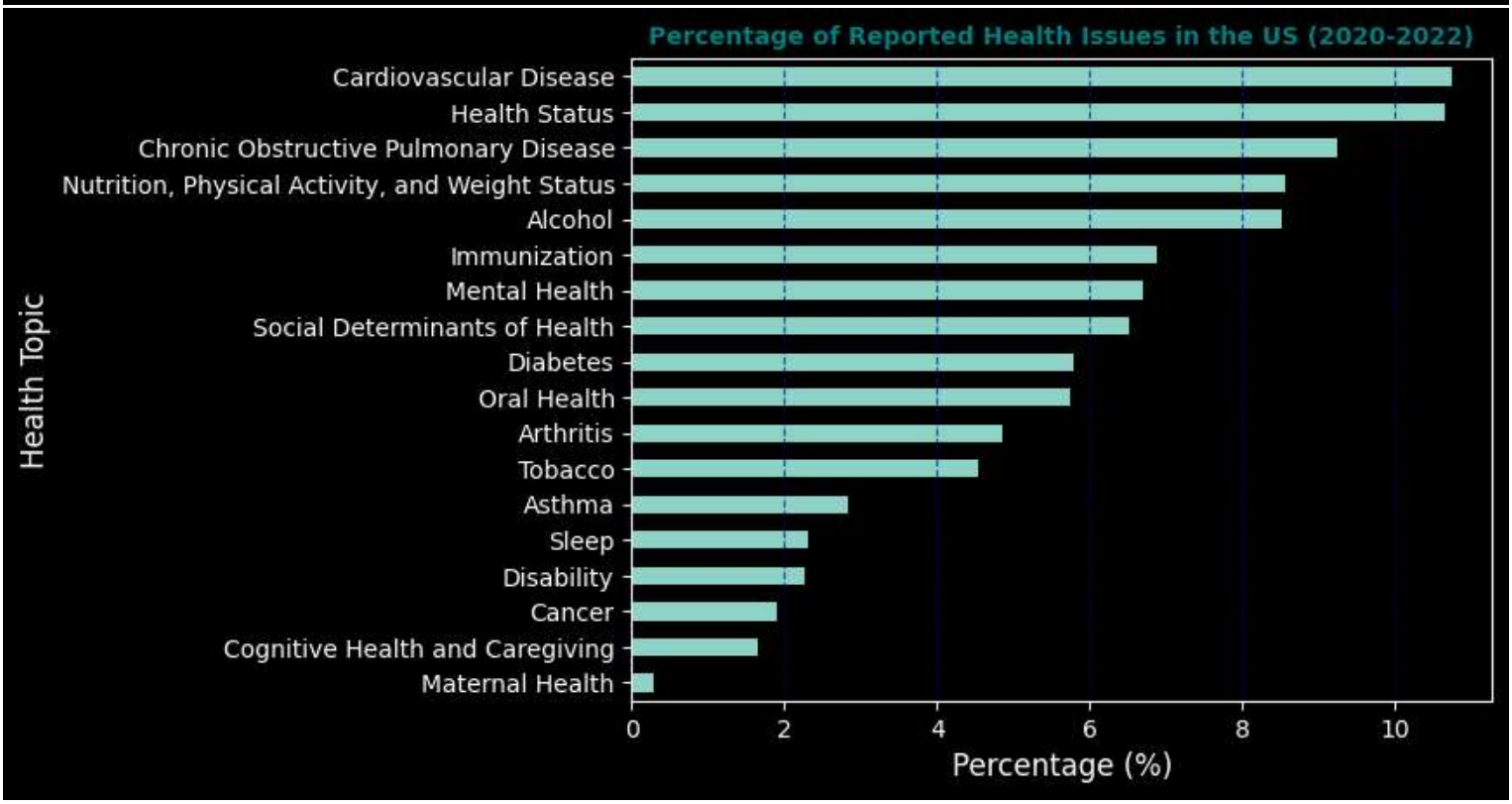
- **Enhancements**:

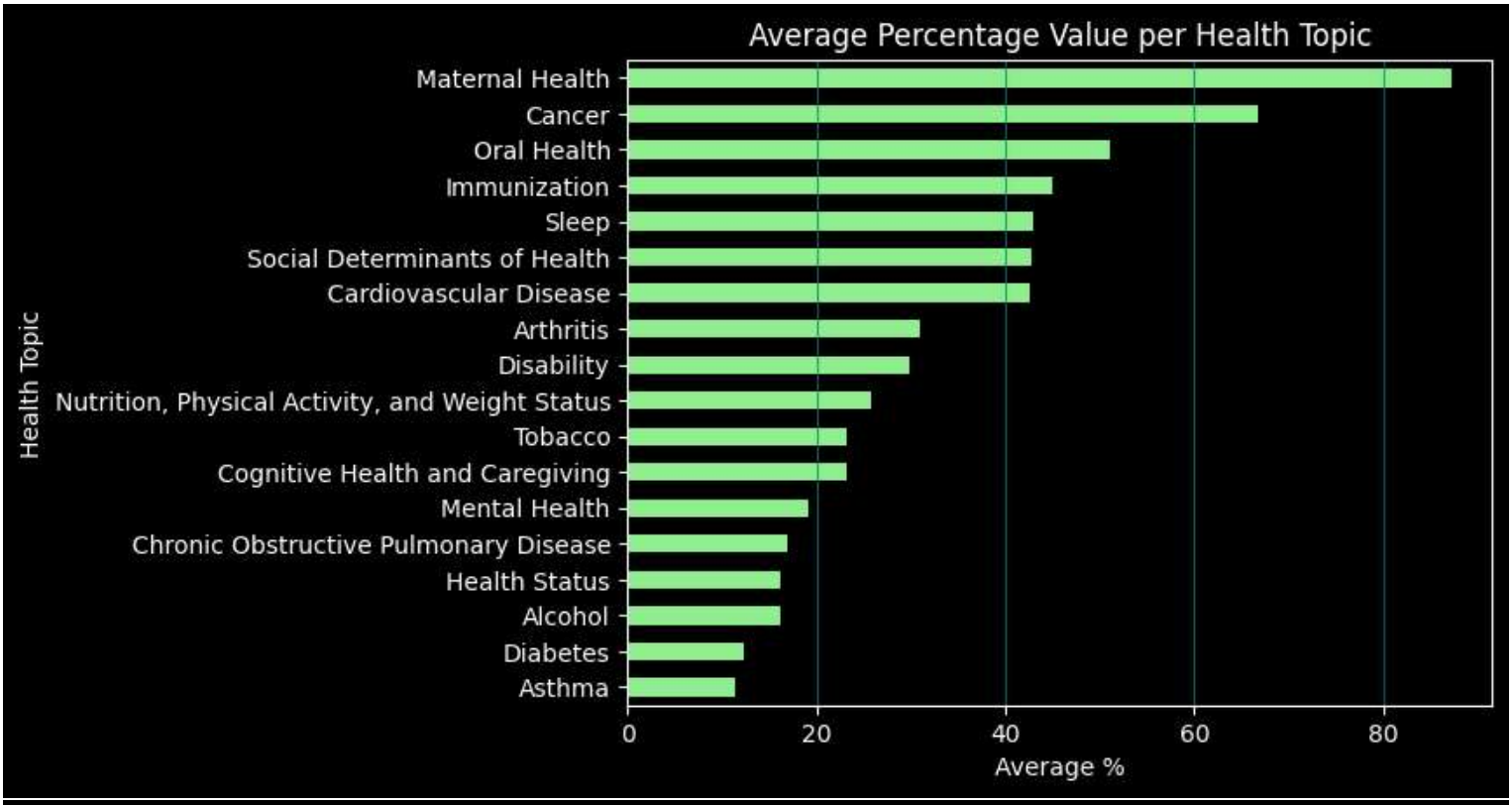
This project lays the foundation for deeper public health insights, and several future directions can enhance its impact:

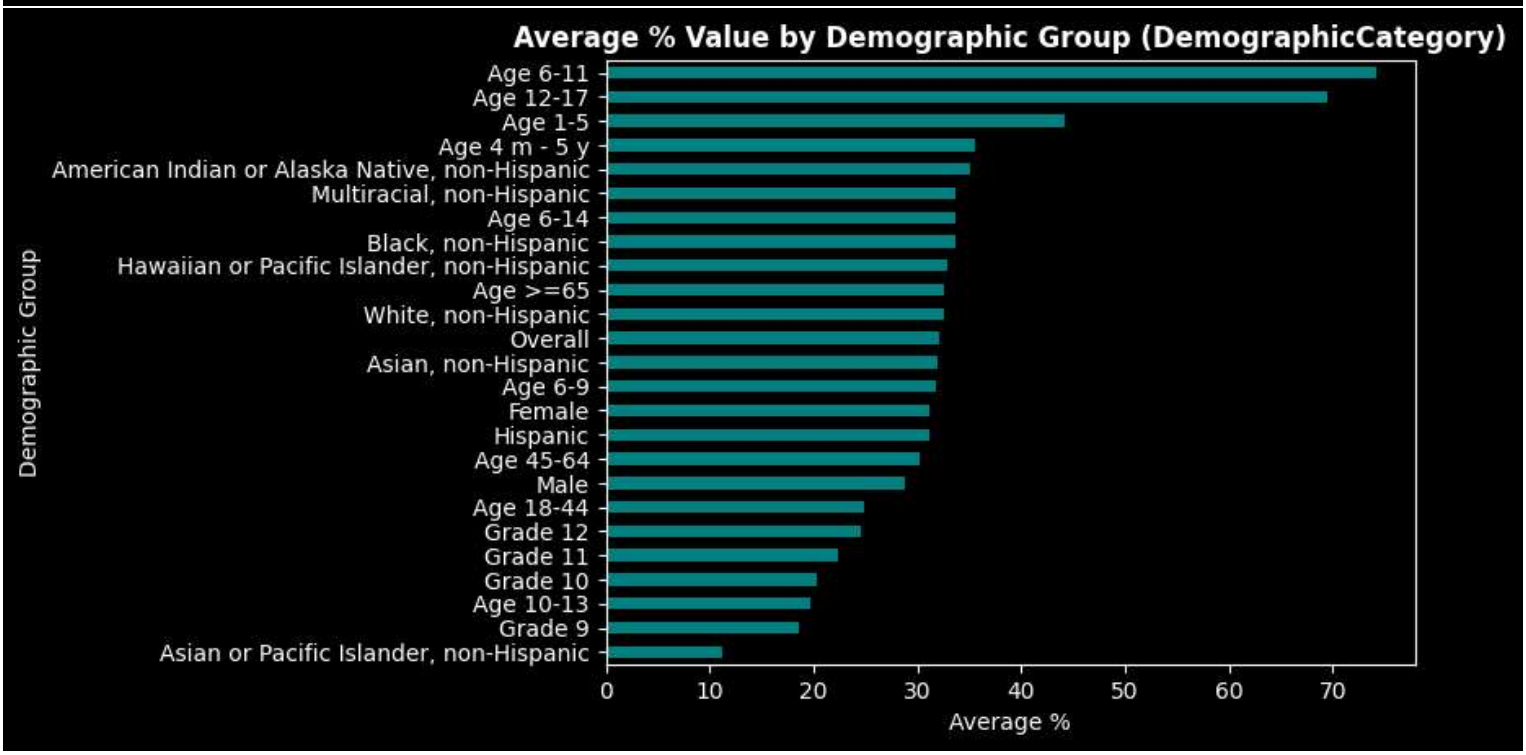
- **Real-Time Monitoring:** Integrate with live public health APIs to track chronic disease indicators as they evolve.
- **Predictive Modeling:** Use machine learning models (e.g., Random Forests or Time Series Forecasting) to predict future trends in obesity, diabetes, and inactivity.
- **Social Determinants Analysis:** Expand the dataset to include income, education, and access to healthcare to understand root causes of health disparities.
- **Interactive Dashboards:** Develop a dynamic visualization tool (using Dash or Tableau) to allow policymakers to explore health data at national, state, and county levels.
- **Policy Impact Studies:** Analyze the effectiveness of local health programs by comparing regions before and after intervention implementation.

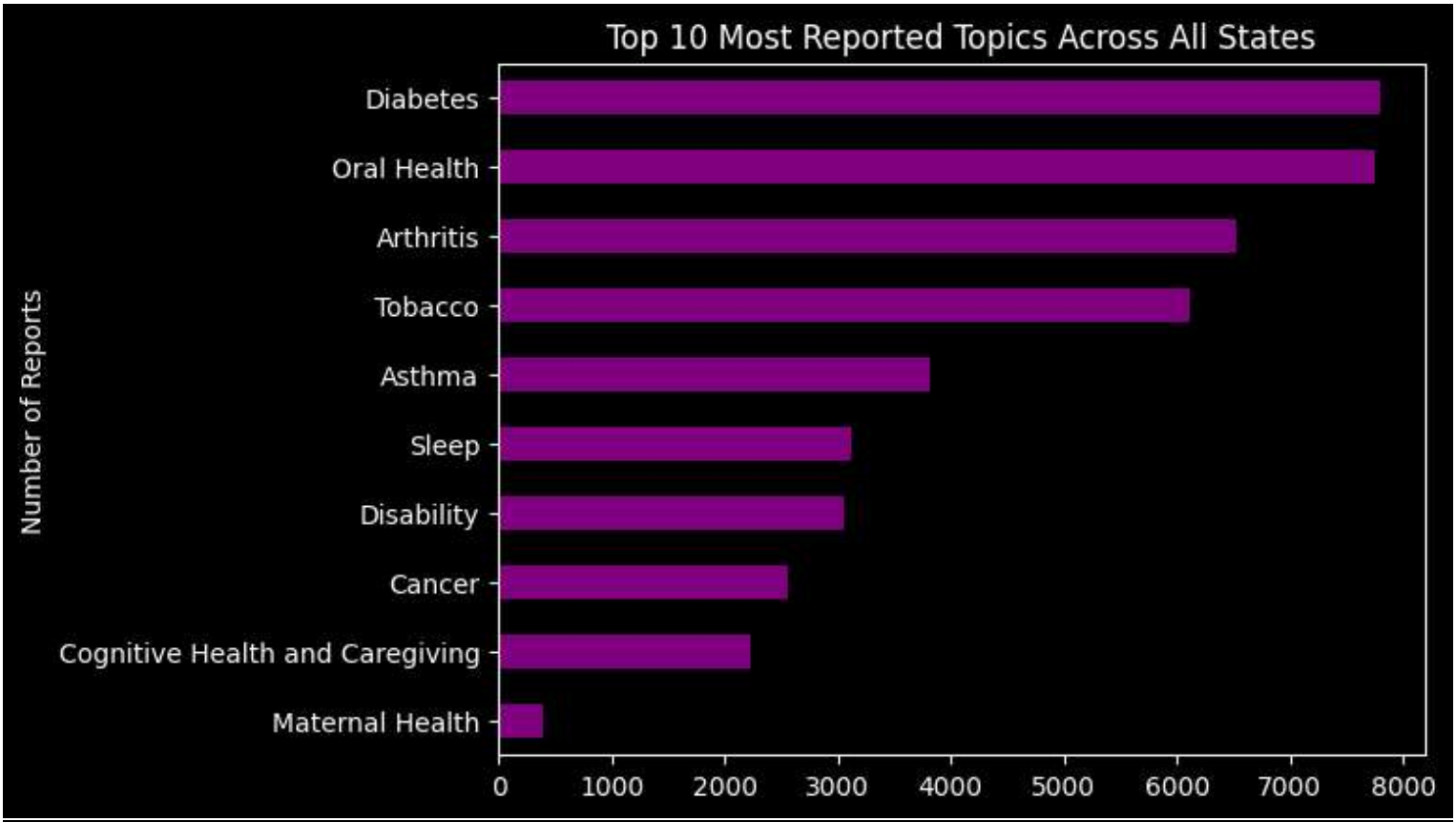
These steps would transform this analysis into a powerful decision-making tool for governments, NGOs, and healthcare organizations.

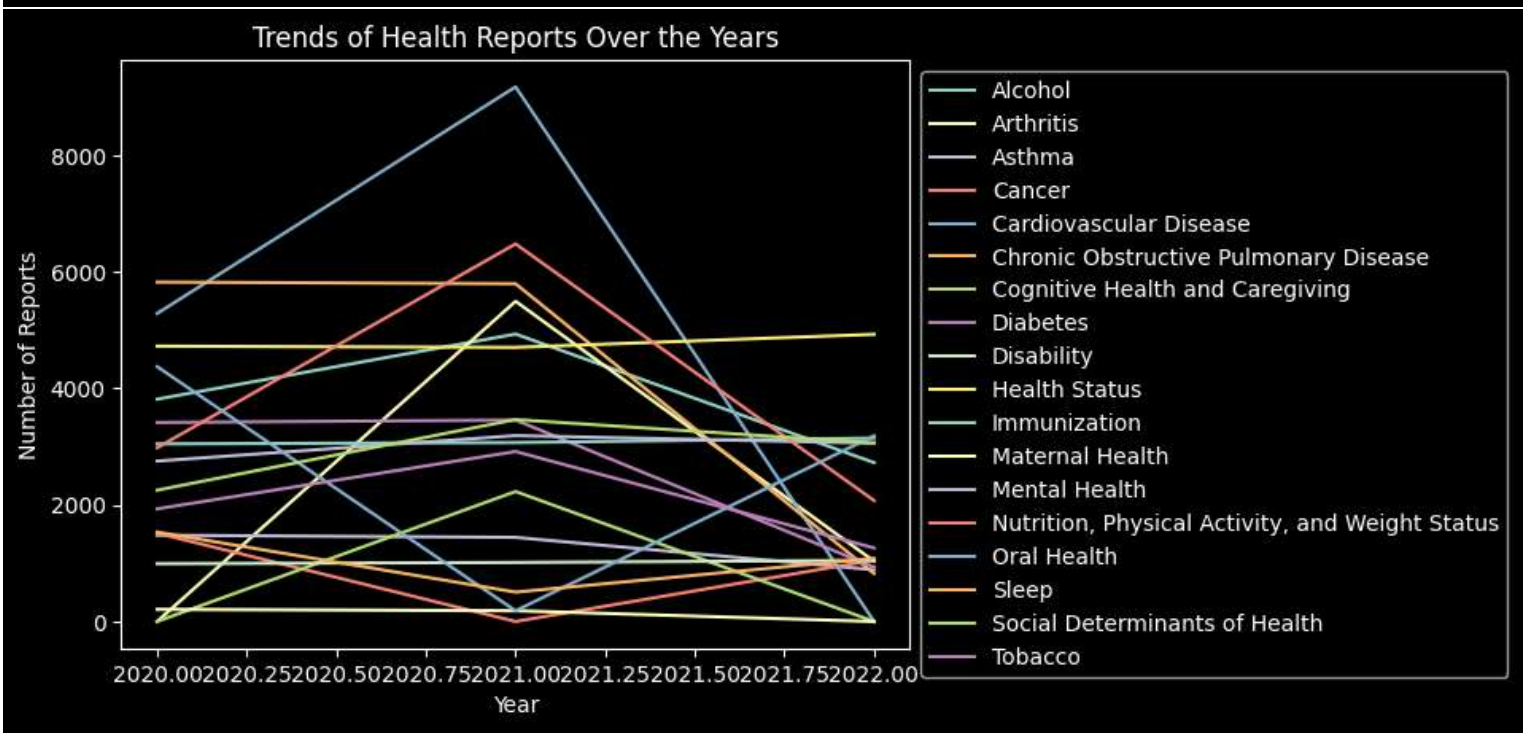

Most Reported Health Issues (2020-2022)
Cardiovascular Disease
Health Status
Chronic Obstructive Pulmonary Disease
Nutrition, Physical Activity, and Weight Status
Alcohol
Immunization
Mental Health
Social Determinants of Health
Diabetes
Oral Health
Arthritis
Tobacco
Asthma
Sleep
Disability
Cancer
Cognitive Health and Caregiving
Maternal Health
Health Topic
0
2000
4000
6000
8000
10000
12000
14000
Number of Reports

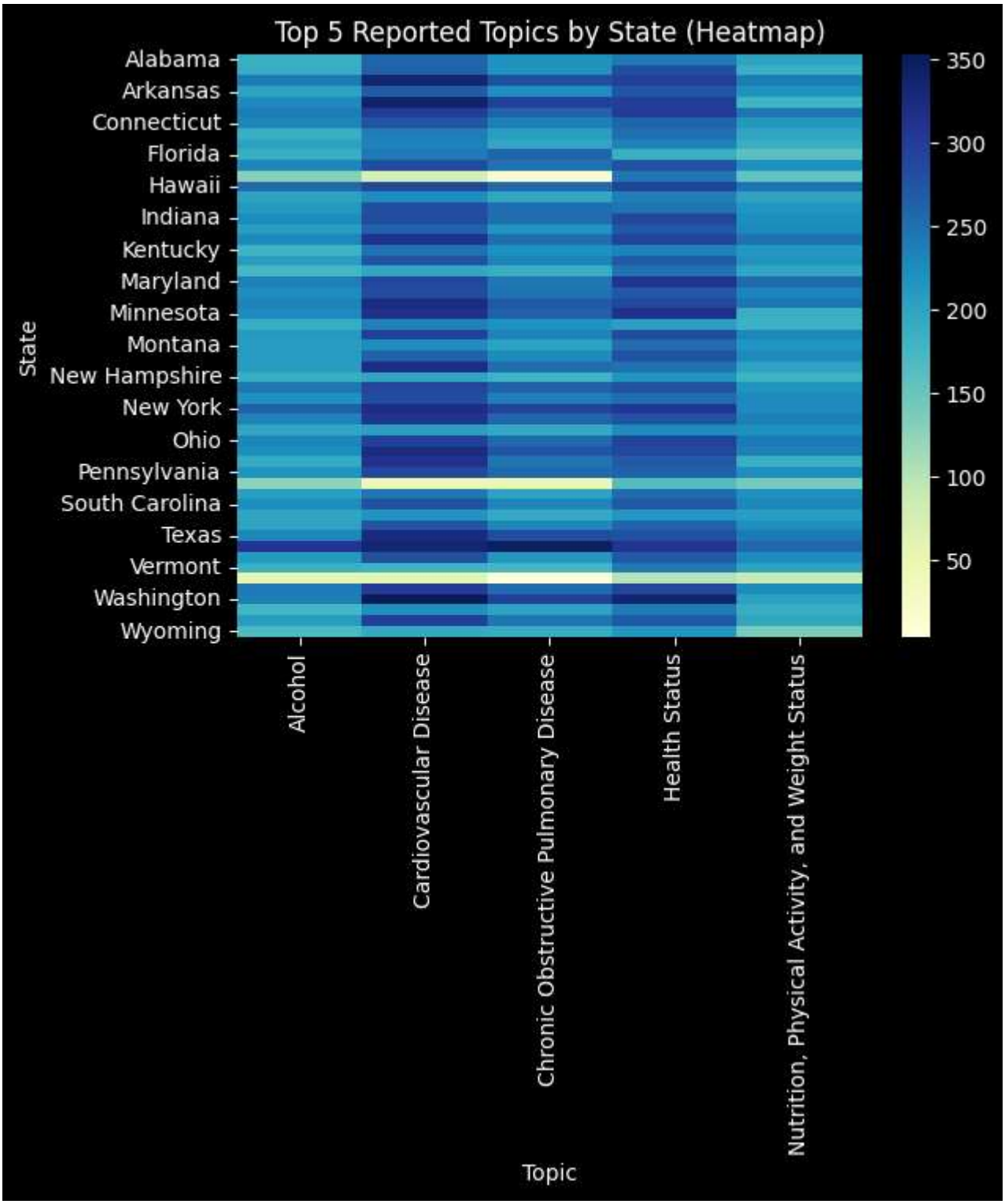

Percentage of Reported Health Issues in the US (2020-2022)
Cardiovascular Disease
Health Status
Chronic Obstructive Pulmonary Disease
Nutrition, Physical Activity, and Weight Status
Alcohol
Immunization
Mental Health
Social Determinants of Health
Diabetes
Oral Health
Arthritis
Tobacco
Asthma
Sleep
Disability
Cancer
Cognitive Health and Caregiving
Maternal Health
Health Topic
0
2
4
6
8
10
Percentage (%)


Average Percentage Value per Health Topic
Maternal Health
Cancer
Oral Health
Immunization
Sleep
Social Determinants of Health
Cardiovascular Disease
Arthritis
Disability
Nutrition, Physical Activity, and Weight Status
Tobacco
Cognitive Health and Caregiving
Mental Health
Chronic Obstructive Pulmonary Disease
Health Status
Alcohol
Diabetes
Asthma
Health Topic
0
20
40
60
80
Average %

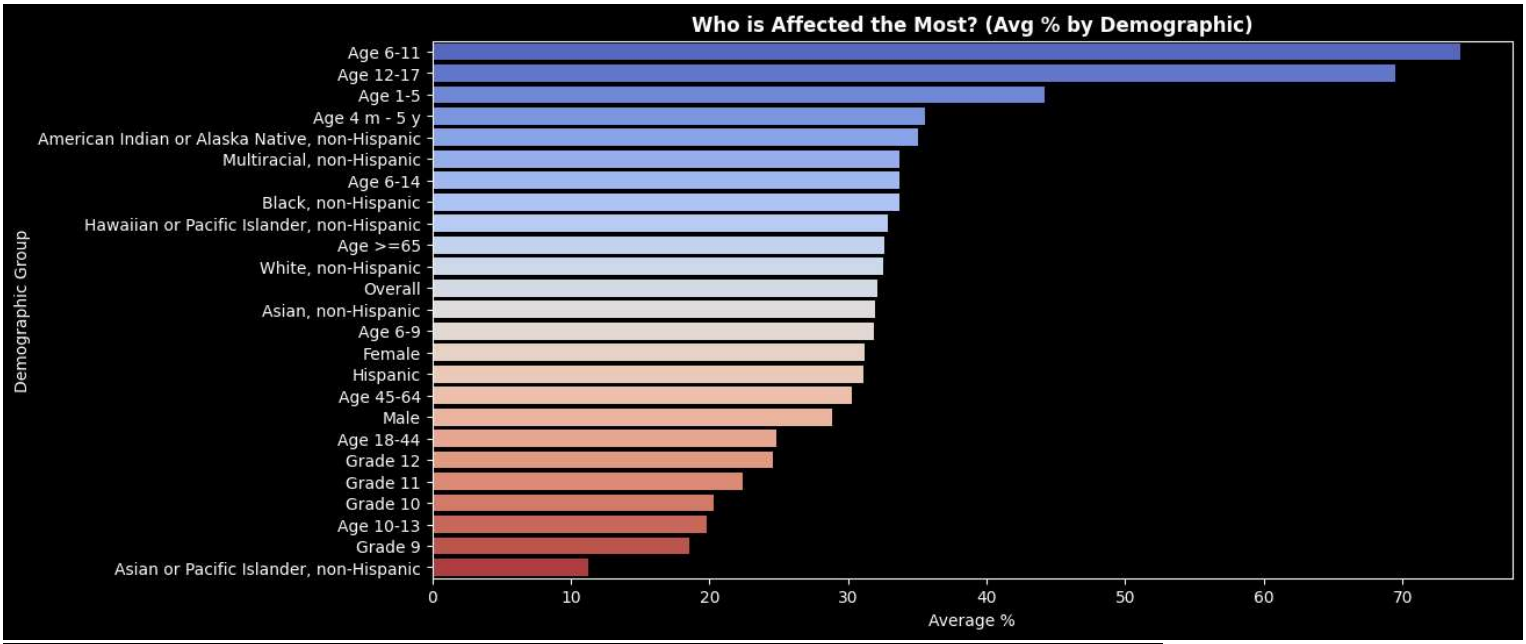

Average % Value by Demographic Group (DemographicCategory)
Age 6-11
Age 12-17
Age 1-5
Age 4 m - 5 y
American Indian or Alaska Native, non-Hispanic
Multiracial, non-Hispanic
Age 6-14
Black, non-Hispanic
Hawaiian or Pacific Islander, non-Hispanic
Age >=65
White, non-Hispanic
Overall
Asian, non-Hispanic
Age 6-9
Female
Hispanic
Age 45-64
Male
Age 18-44
Grade 12
Grade 11
Grade 10
Age 10-13
Grade 9
Asian or Pacific Islander, non-Hispanic
Demographic Group
0
10
20
30
40
50
60
70
Average %

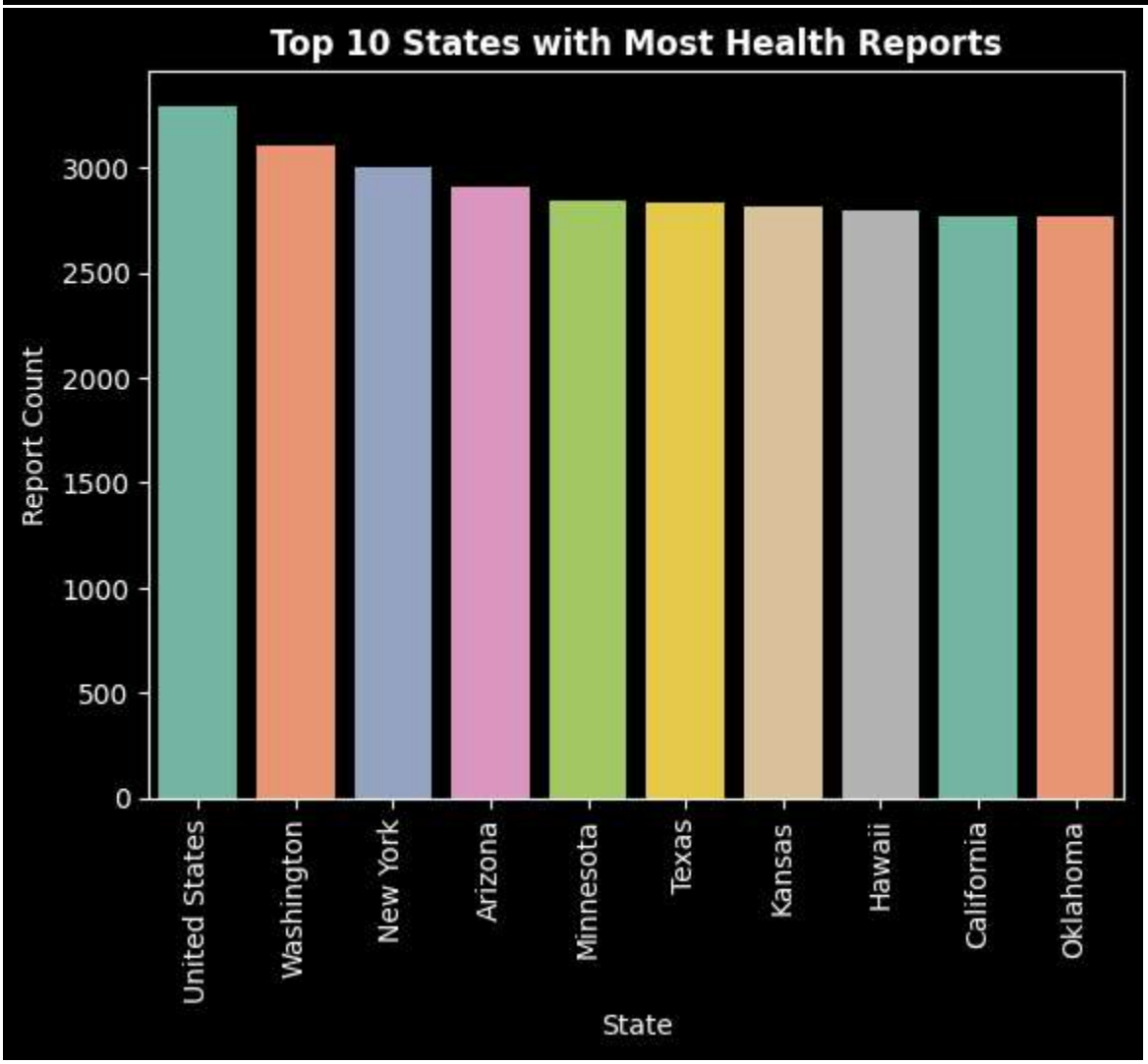

Top 10 Most Reported Topics Across All States
Diabetes
Oral Health
Arthritis
Tobacco
Asthma
Sleep
Disability
Cancer
Cognitive Health and Caregiving
Maternal Health
Number of Reports
0
1000
2000
3000
4000
5000
6000
7000
8000

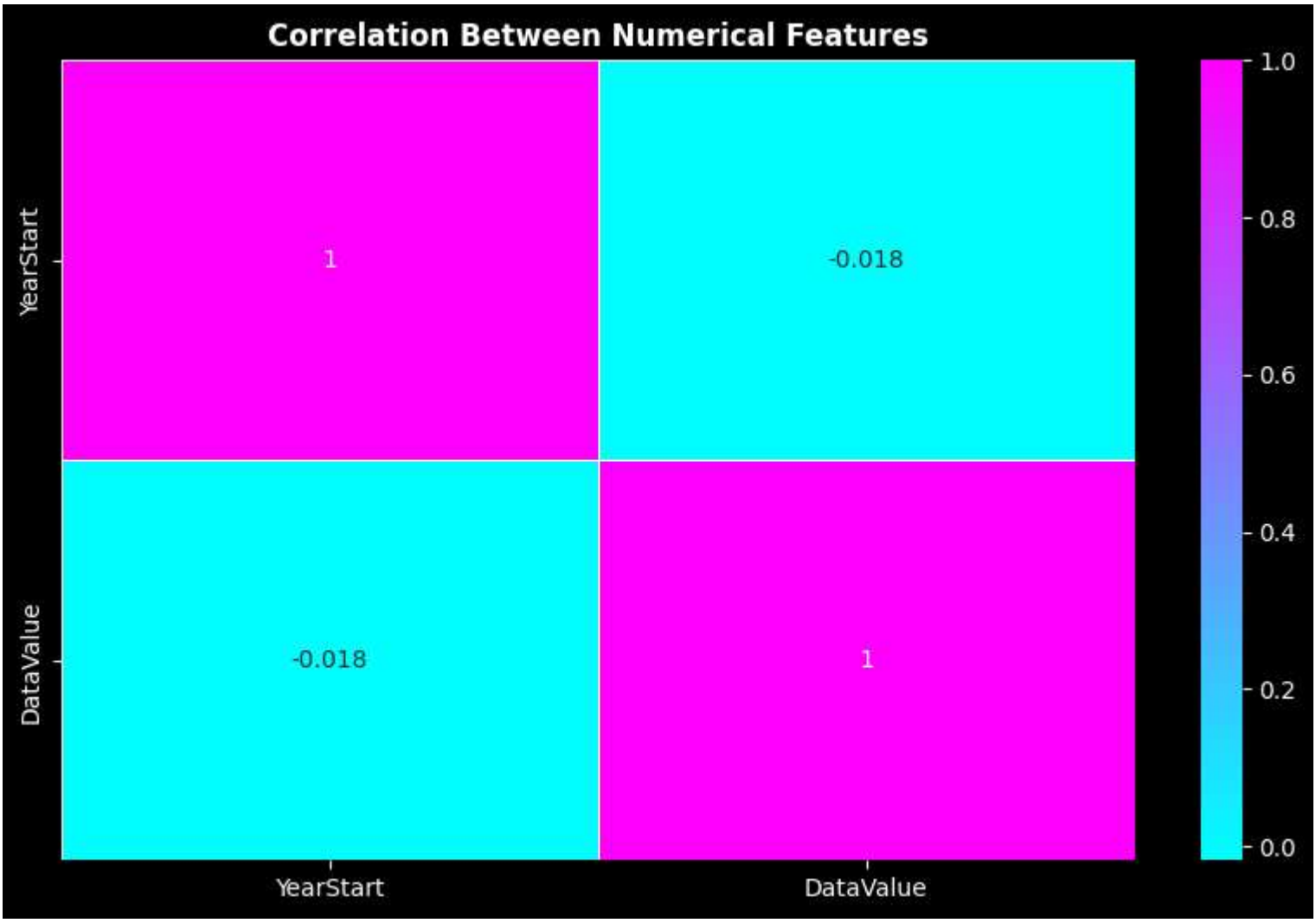

Trends of Health Reports Over the Years
Number of Reports
0
2000
4000
6000
8000
2020.002020.252020.502020.752021.002021.252021.502021.752022.00
Year
Alcohol
Arthritis
Asthma
Cancer
Cardiovascular Disease
Chronic Obstructive Pulmonary Disease
Cognitive Health and Caregiving
Diabetes
Disability
Health Status
Immunization
Maternal Health
Mental Health
Nutrition, Physical Activity, and Weight Status
Oral Health
Sleep
Social Determinants of Health
Tobacco


Top 5 Reported Topics by State (Heatmap)
State
Alabama
Arkansas
Connecticut
Florida
Hawaii
Indiana
Kentucky
Maryland
Minnesota
Montana
New Hampshire
New York
Ohio
Pennsylvania
South Carolina
Texas
Vermont
Washington
Wyoming
Alcohol
Cardiovascular Disease
Chronic Obstructive Pulmonary Disease
Health Status
Nutrition, Physical Activity, and Weight Status
Topic
350
300
250
200
150
100
50

Boxplot grouped by Topic
Distribution of % DataValue by Topic
% Value
0
20
40
60
80
100
Alcohol
Arthritis
Asthma
Cancer
Cardiovascular Disease
Chronic Obstructive Pulmonary Disease
Cognitive Health and Caregiving
Diabetes
Disability
Health Status
Immunization
Maternal Health
Mental Health
Nutrition, Physical Activity, and Weight Status
Oral Health
Sleep
Social Determinants of Health
Tobacco
Topic


Who is Affected the Most? (Avg % by Demographic)
Demographic Group
Age 6-11
Age 12-17
Age 1-5
Age 4 m - 5 y
American Indian or Alaska Native, non-Hispanic
Multiracial, non-Hispanic
Age 6-14
Black, non-Hispanic
Hawaiian or Pacific Islander, non-Hispanic
Age >=65
White, non-Hispanic
Overall
Asian, non-Hispanic
Age 6-9
Female
Hispanic
Age 45-64
Male
Age 18-44
Grade 12
Grade 11
Grade 10
Age 10-13
Grade 9
Asian or Pacific Islander, non-Hispanic
0
10
20
30
40
50
60
70
Average %


Top 10 States with Most Health Reports
Report Count
3000
2500
2000
1500
1000
500
0
United States
Washington
New York
Arizona
Minnesota
Texas
Kansas
Hawaii
California
Oklahoma
State



GITHUB LINK:

https://github.com/shezalfatima/Chronic-Disease-Analysis

LINKEDIN LINK:

https://www.linkedin.com/feed/update/urn:li:activity:7316889071111327744/