FinalProjectProposal

Chang Lu

2024-11-10

Personal Statement

For my final project, I have chosen the dataset titled "U.S. Chronic Disease Indicators (CDI) – 2023 Release." This comprehensive dataset encompasses a significant volume of information related to various chronic diseases across the United States, containing 1,185,676 rows and 34 columns. While the dataset's extensive nature presents challenges, it also offers rich opportunities for insightful analysis, which is promising for my development in data analysing skills.

Introduction to Key Variables

Given the dataset's complexity, I aim to focus on identifying and working with key variables to guide my analysis effectively. The primary variables I have selected include:

1. YearStart and YearEnd: Indicating the time frame during which the data was collected.
2. LocationAbbr, LocationDesc, and GeoLocation: Providing information on where the data was gathered, including both state abbreviations and full descriptions.
3. DataValueAlt and DataValueType: Representing the numerical values and the type of data recorded, crucial for quantitative analysis.
4. StratificationCategory1 and Stratification1: Showing how the data is segmented into various groups, which is essential for subgroup analysis and understanding disparities.

In the upcoming month, I plan to conduct detailed analyses centered around these variables. This approach will allow me to manage the dataset's breadth while yielding meaningful insights into chronic disease trends and disparities within the U.S. population.

By narrowing my focus to these core variables, I hope to produce a thorough and impactful analysis that highlights significant patterns and contributes to a deeper understanding of chronic disease data.

Question

To complete my analysis on the dataset, I have come up with several preliminary questions based on key variables mentioned above:

1. Temporal Analysis: How have trends in chronic disease indicators changed over the years from YearStart to YearEnd? Are there noticeable increases or decreases in specific conditions over time?(might change because in the most rows the startyear and endyear are the same)

2. Geographical Patterns: What are the differences in chronic disease rates across different locations (using LocationAbbr, LocationDesc, and GeoLocation)? Are certain states or regions more affected by particular chronic conditions?

3. Quantitative Insights: What is the distribution of DataValueAlt for various chronic diseases, and how does DataValueType influence the interpretation of these numerical values?

4. Group Disparities: How do the stratification categories (StratificationCategory1 and Stratification1) impact the prevalence of certain chronic diseases? Are there significant disparities among different demographic or social groups?(like hispanic or asian)

5. Correlation Analysis: Is there a relationship between certain stratification categories and changes in DataValueAlt over the observed years?

6. Comparative Analysis: Can differences be identified when comparing data between multiple states or regions for the same stratification categories?

The Data Source

I got access to the data("U.S. Chronic Disease Indicators (CDI) – 2023 Release.")(Link Text) in the website of data.gov(Link text). As the page described, this dataset is intended for public access and use.

proposed Timeline of Work

Since there is less than one month remaining, I would like to complete each part of my project at a weekly pace.

Deadline for Each Part

1. EDA: November 18
2. Data Processing: November 18
3. Modeling and Validation: November 25
4. Final submission: December 02