JANUARY 2025

# Project Proposal

## Chronic Disease Prevalence Analysis and Visualization


PRESENTED TO:
Eng.Ahmed Noaman

PRESENTED BY:
Maram Zoughieb

# About Dataset

Context:
CDC's Division of Population Health provides cross-cutting set of 124 indicators that were developed by consensus and that allows states and territories and large metropolitan areas to uniformly define, collect, and report chronic disease data that are important to public health practice and available for states, territories and large metropolitan areas. In addition to providing access to state-specific indicator data, the CDI web site serves as a gateway to additional information and data resources.

Content:
A variety of health-related questions were assessed at various times and places across the US over the past 15 years. Data is provided with confidence intervals and demographic stratification.

Data information: it shows that data is imbalnced.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 403984 entries, 0 to 403983
Data columns (total 34 columns):
 #   Column                      Non-Null Count   Dtype
---  ------                      --------------   -----
 0   YearStart                   403984 non-null  int64
 1   YearEnd                     403984 non-null  int64
 2   LocationAbbr                403984 non-null  object
 3   LocationDesc                403984 non-null  object
 4   DataSource                  403984 non-null  object
 5   Topic                       403984 non-null  object
 6   Question                    403984 non-null  object
 7   Response                    79323 non-null   object
 8   DataValueUnit               374119 non-null  object
 9   DataValueType               403984 non-null  object
 10  DataValue                   297817 non-null  object
 11  DataValueAlt                273666 non-null  float64
 12  DataValueFootnoteSymbol     188019 non-null  object
 13  Datavaluefootnote           187853 non-null  object
 14  LowConfidenceLimit          246819 non-null  float64
 15  HighConfidenceLimit         246819 non-null  float64
 16  StratificationCategory1     403984 non-null  object
 17  Stratification1             403984 non-null  object
 18  StratificationCategory2     79323 non-null   object
 19  Stratification2             79323 non-null   object
 20  StratificationCategory3     79323 non-null   object
 21  Stratification3             79323 non-null   object
 22  GeoLocation                 401416 non-null  object
 23  ResponseID                  79323 non-null   object
 24  LocationID                  403984 non-null  int64
 25  TopicID                     403984 non-null  object
 26  QuestionID                  403984 non-null  object
 27  DataValueTypeID             403984 non-null  object
 28  StratificationCategoryID1   403984 non-null  object
 29  StratificationID1           403984 non-null  object
 30  StratificationCategoryID2   79324 non-null   object
 31  StratificationID2           79324 non-null   object
 32  StratificationCategoryID3   79323 non-null   object
 33  StratificationID3           79323 non-null   object
dtypes: float64(3), int64(3), object(28)
memory usage: 104.8+ MB
```

**DATA SHAPE**

```
#Show data shape and type
print("Shape of the dataset:", df.shape)
```

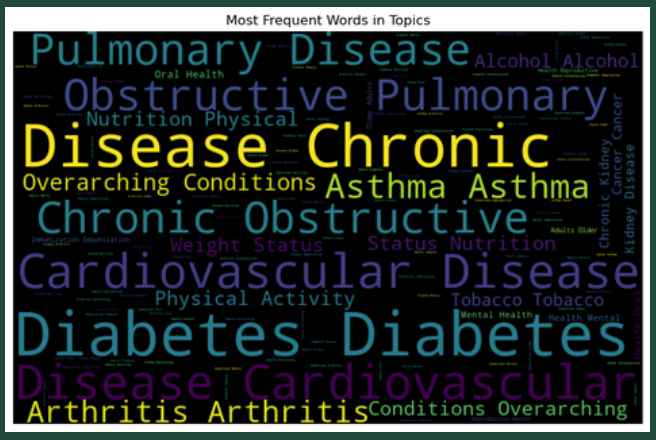
```
Shape of the dataset: (403984, 34)
```

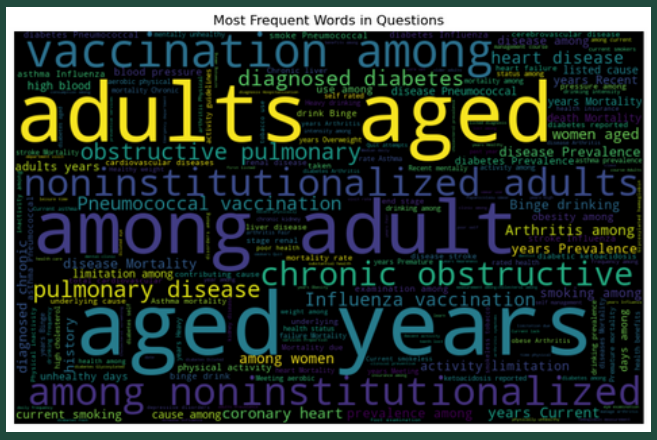
# Data Processing

1. Exploratory Data Analysis : Checking and Cleaning
2. Handle Missing Values
3. Univariate Analysis and Multivariant Analysis
4. Checking outliers for numerical data
5. Feature Engineering
6. Analysis

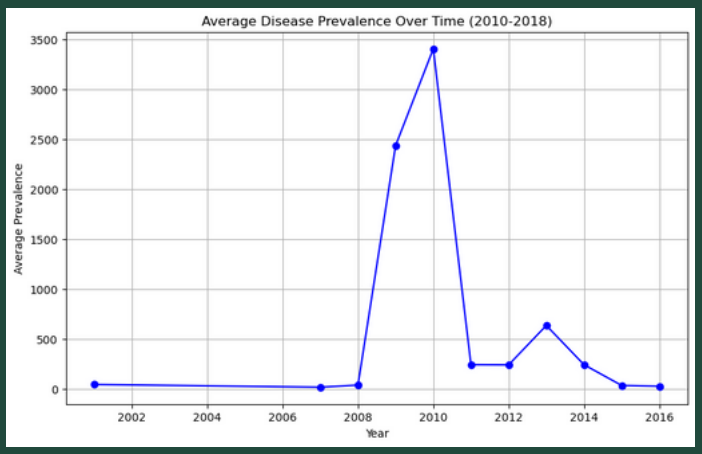
# Analysis Insights

1. checked the most frequent diseases and questions in the data

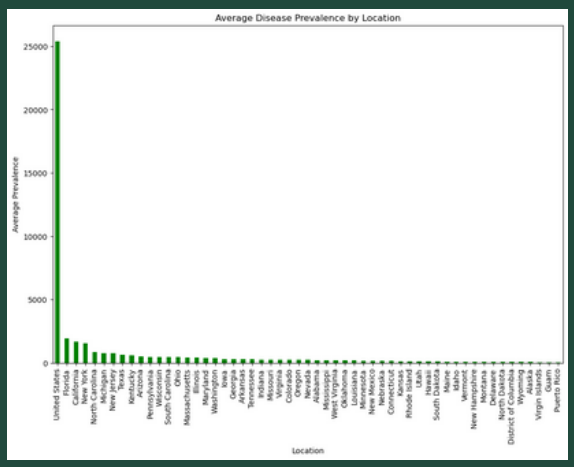


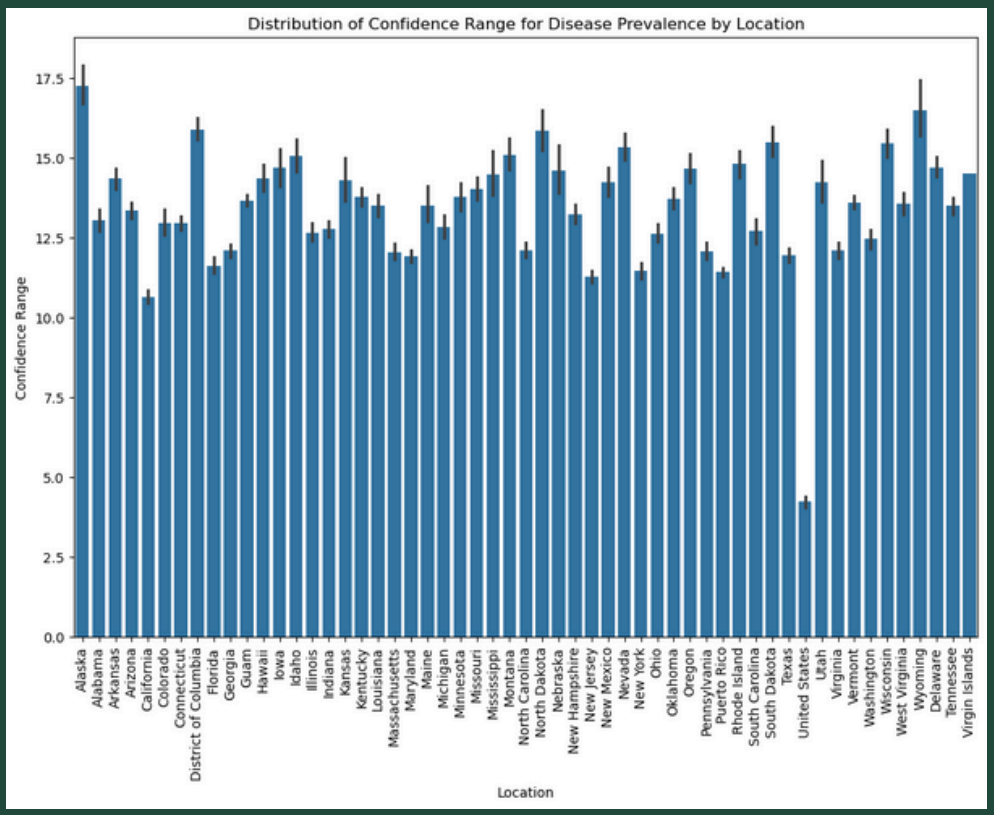

2. How have the prevalence rates of chronic diseases changed from 2010 to 2018?



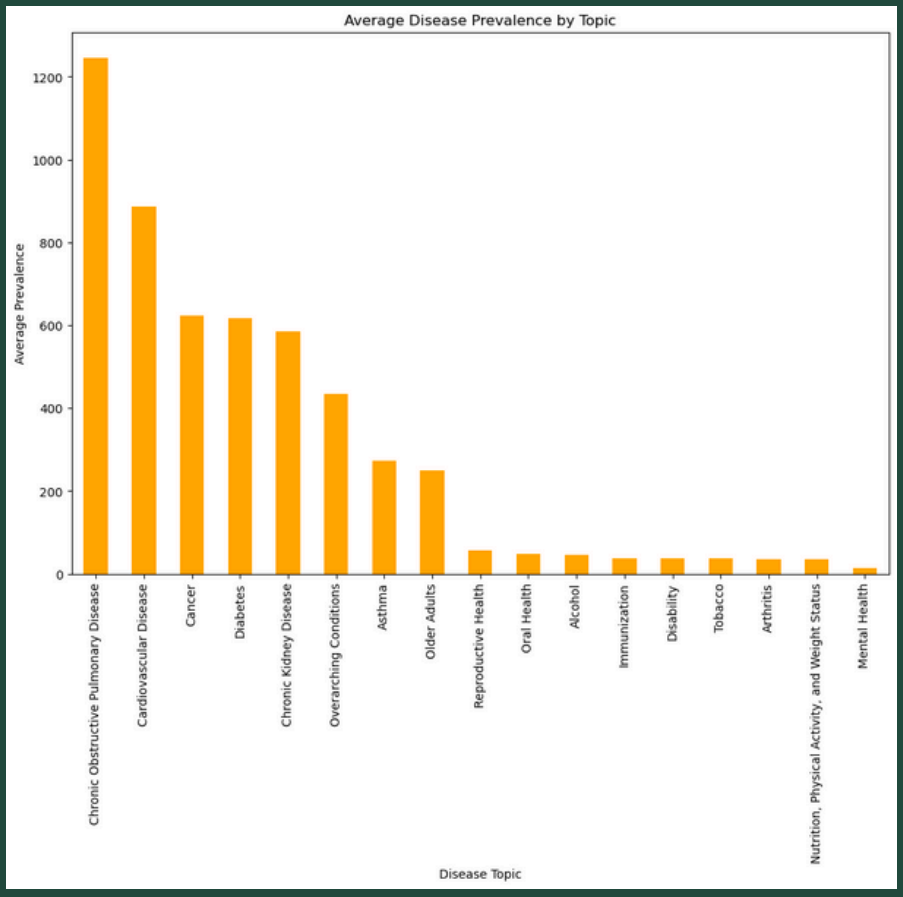
3. How do different locations (states) compare in terms of disease prevalence?



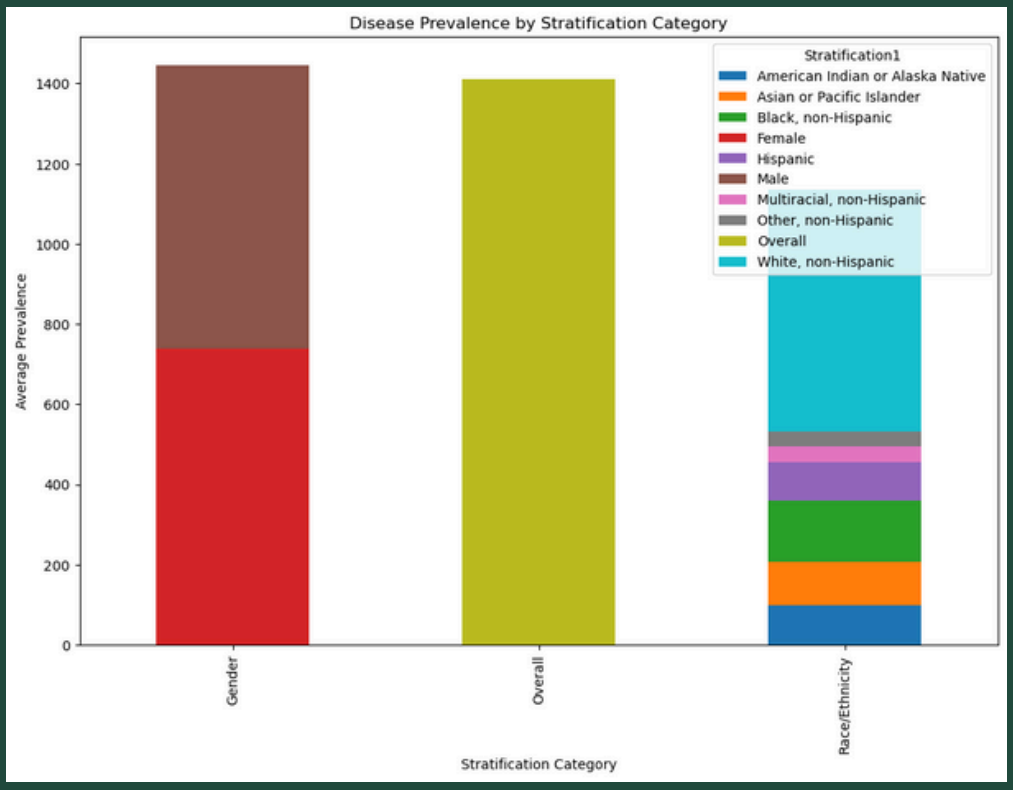

4. What is the range of confidence intervals for disease prevalence, and how does it vary across states?



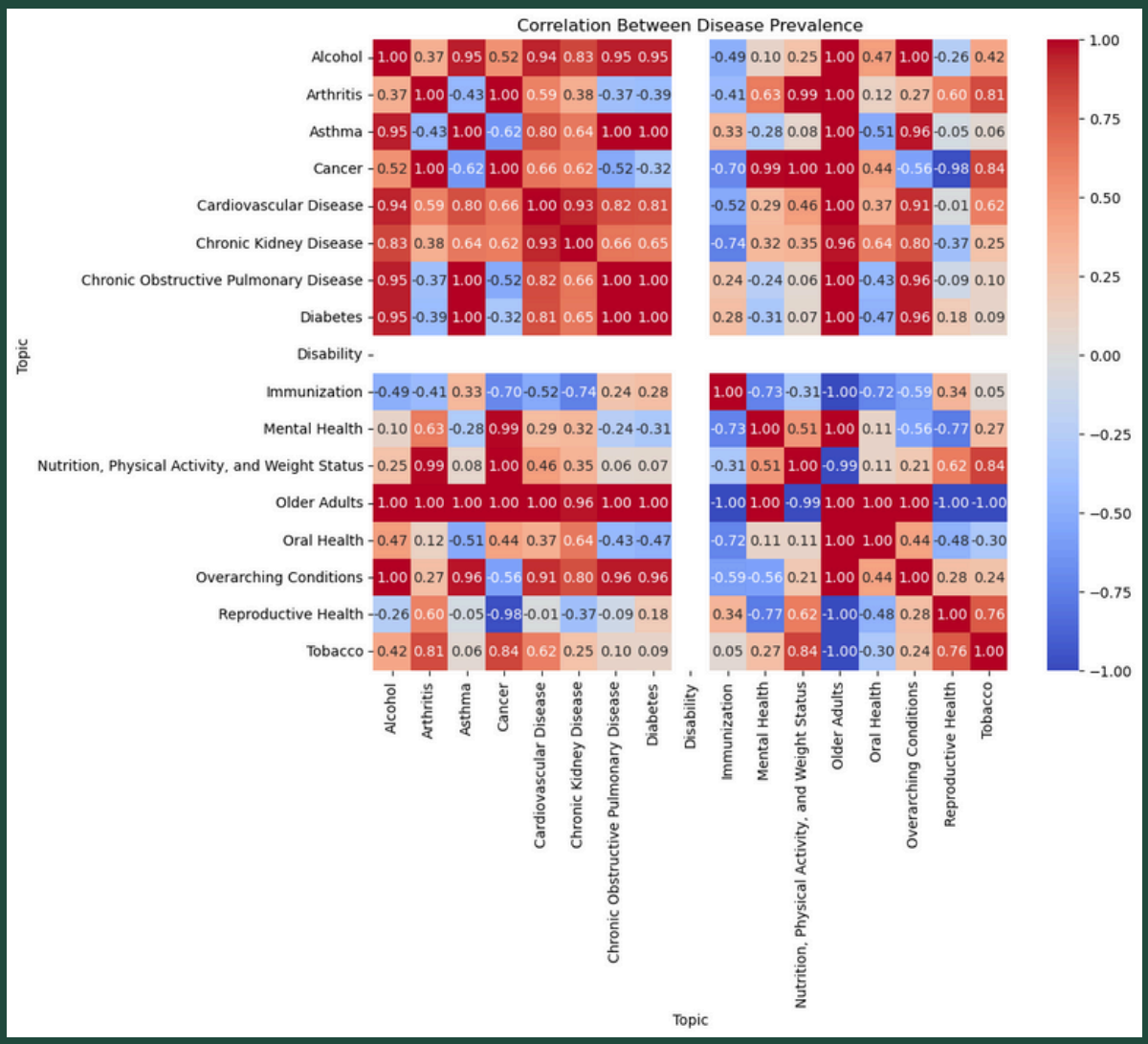
5. Which chronic disease topics (e.g., diabetes, cardiovascular disease) have the highest and lowest prevalence rates?




6.How does disease prevalence vary across different stratifications like age group, gender, or race?



7.Is there a correlation between the prevalence of different diseases (e.g., diabetes and hypertension)? Creating a correlation matrix for disease prevalence across topics



# Key Benefit

1. **Data-Driven Insights for Decision Making**

Analysis helped uncover hidden patterns in chronic disease prevalence, enabling:

- Identification of high-risk groups (ex: gender).
- Understanding disease correlations (comorbidities and risk factors).
- Spotting trends over time (rising or declining prevalence).

**2. Improved Public Health Strategies**

- Targeted interventions can be designed for vulnerable populations.
- Resource allocation can be optimized for healthcare planning.

**3. Statistical and Visualization Mastery**

- data has been cleaned and processed real-world health data.
- applied exploratory data analysis (EDA) to extract meaningful patterns.
- visualized complex trends to communicate findings effectively.

**4. Future Research and AI/ML Applications**

- **F**indings provide a foundation for predictive modeling of disease risk.
- The data can be used to train machine learning models for early diagnosis.
- Further studies can explore genetic and environmental factors influencing disease prevalence.

**Final Takeaway**

- This analysis is valuable for public health planning, medical research, and policy-making. It bridges the gap between raw data and actionable insights, empowering data-driven healthcare decisions.