The data are available, but a major part of the project is getting them cleaned: Perhaps it's voting data from many sources covering many counties or precincts. Your submission will focus on some basic features of the data and your plan for cleaning.

Data Overview:

Behavioral Risk Factor Surveillance System (BRFSS)

What Is BRFSS?
The Behavioral Risk Factor Surveillance System (BRFSS) is the largest continuously conducted health survey system in the world, overseen by the Centers for Disease Control and Prevention (CDC). It collects data from U.S. adults (ages 18+) about their health-related risk behaviors, chronic health conditions, and use of preventive services.

Conducted annually in all 50 states, D.C., and U.S. territories, the BRFSS plays a vital role in informing public health programs, shaping policy, and tracking national and state-level trends in health behaviors and outcomes.

Core Purpose
Since its inception, BRFSS has aimed to:

- Monitor health behaviors contributing to the leading causes of death and disability.
- Track access to health care and use of preventive services.
- Provide timely, localized health data to support public health planning and evaluation.

Significance and Use

- Public Health Agencies use BRFSS data to allocate resources and design targeted interventions.
- Researchers and Policymakers analyze trends to understand and address disparities in health outcomes.
- Local Governments tailor community health initiatives based on real-time behavioral data.

Topics include:

- Health care access
- Tobacco and alcohol use
- Physical activity

- Diet and nutrition
- Hypertension and cholesterol
- HIV/AIDS knowledge and prevention
- Cancer screening
- Immunizations
- Injury prevention

How BRFSS Has Evolved Over Time

Category	Then (e.g., 1999)	Now
Topics Covered	Smoking, alcohol use, diet, exercise, screenings	Adds mental health, insurance access, social determinants, opioids, COVID-19
Data Collection Method	Landline telephone surveys only	Landline + cell phone surveys (since 2011)
Weighting Method	Post-stratification	Raking (iterative proportional fitting) – more accurate demographics
Customization	Standard core + some optional modules	Greater flexibility for state-specific and emerging issues
Equity Focus	Limited demographic analysis	Stronger emphasis on health equity, including race, income, education, geography

Questions to Address:

- **Self-reported rates of depression/anxiety across age groups** (MENTHLTH)
- **Income level and health status relationship** (INCOME2, GENHLTH)
- **Physical activity variations by demographics** (EXERANY2, PASTAE1)
- **Chronic conditions correlation with socioeconomic factors** (DIABETE3, CVDINFR4, ASTHMA3)
- **Health insurance status and preventive care access** (HLTHPLN1)

Plan for Cleaning:

- **Data Challenges:**
    - **Skew towards more affluent populations:** The data from earlier years (e.g., 1993) might be biased due to limited phone access, especially for low-income groups.
    - **Consistency in variable definitions**: Ensure that variables across different years are comparable, especially given the evolving data collection methods (e.g., landline vs. cell phone surveys).
    - **Missing data handling**: There might be missing values in certain years, which will require imputation or exclusion.

Next Steps:

1. **Consolidate Variables**: Compile the relevant variables across all years into a single file. Make sure each year is aligned by variable name and value definition.
2. **Clean and Preprocess**: Address issues like missing values, outliers, and inconsistent variable types.
3. **Descriptive Statistics**: Generate summary statistics to understand the variation within the data and identify trends or anomalies.
4. **Visualizations**: Create visualizations that compare health outcomes, mental health, and access to care across different demographic groups.
5. **Document the Plan**: Write up the challenges you expect to face in data cleaning, as well as how the data will be useful in studying the phenomenon.