Ex 6.1: Sourcing Open Data

1. Data Source

Summary of the Data Source

The dataset utilized in this project is derived from the Centers for Disease Control and Prevention's (CDC) Behavioral Risk Factor Surveillance System (BRFSS). The BRFSS is a longstanding, large-scale telephone survey system that collects health-related data from U.S. residents. It captures information on health status, health-related risk behaviors, preventive health practices, and access to healthcare. The 2022 dataset includes indicators related to personal health behaviors, chronic conditions, preventive health screenings, and demographic variables.

While the original data can be accessed directly from the CDC's official website (https://www.cdc.gov/brfss/annual_data/annual_2022.html), for convenience and usability, this project uses a cleaned and pre-processed version available on Kaggle: Personal Key Indicators of Heart Disease (2022).

This curated dataset from Kaggle maintains the essential variables from the BRFSS 2022 data while offering a more accessible format for exploratory data analysis.

Explanation for Choosing This Dataset

This dataset is particularly well-suited for a project focused on health-related analytics for several reasons:

1. **Relevance to Public Health**: The BRFSS is a cornerstone resource for public health research in the United States. It includes a wide range of health indicators—such as chronic disease prevalence, preventive care measures, health behaviors, and demographic attributes—that support comprehensive analyses.

2. **Scope and Breadth**: The dataset covers tens of thousands of respondents across various states, providing a nationally representative sample. This broad coverage allows for in-depth examination of regional differences, demographic trends, and population-level health patterns.

3. **Richness of Variables**: Beyond core health metrics (like BMI, mental health days, and physical activity), the dataset contains information on healthcare utilization (check-ups, vaccinations), health screenings, and health conditions (e.g., heart attack, diabetes,

stroke). Such variety enables flexible and multifaceted analyses.

4. **Alignment with Personal Interests:** I have a strong personal interest in exploring healthcare data, and this dataset directly aligns with my goals to understand public health trends, healthcare disparities, and the impact of preventive medicine on health outcomes.

Limitations

1. **Self-Reported Data**: The BRFSS relies on respondents' self-reported information, which may introduce recall bias or inaccuracies due to misunderstanding survey questions.

2. **Non-Response Bias**: Certain groups may be less likely to participate in telephone-based surveys. As a result, some demographic or socio-economic segments could be underrepresented, affecting the generalizability of findings.

3. **Data Granularity**: While the dataset is large, detailed clinical measures (e.g., laboratory values, imaging results) are not included. The data provides a broad snapshot rather than in-depth clinical records.

Ethical Considerations

1. **Privacy and Confidentiality**: Although the dataset is publicly available in a de-identified format, it originally contained sensitive health information. Ensuring that no attempt is made to re-identify individuals is critical.

2. **Responsible Use of Data:** It is crucial to ensure that findings from the analysis are communicated accurately and used responsibly, as they have the potential to influence health decisions, policies, and interventions.

3. **Cultural and Social Sensitivity:** Special care must be taken when working with sensitive health data, ensuring that analysis and reporting do not perpetuate stereotypes, biases, or stigmatization of any group.

2. Data Profile

Data Profile Table

Column Name	Description	Variable Type
State	State FIPS Code	Categorical
Sex	Sex of Respondent	Categorical
GeneralHealth	General Health Rating	Categorical
PhysicalHealth	Number of Days Physical Health Not Good (Past 30 Days)	Numerical
MentalHealth	Number of Days Mental Health Not Good (Past 30 Days)	Numerical
LastCheckup	Time Since Last Routine Checkup	Categorical
PhysicalActivities	Participation in Physical Activities	Categorical
SleepHours	Average Hours of Sleep per Day	Numerical
TeethRemoved	Number of Permanent Teeth Removed	Categorical
HadHeartAttack	Ever Diagnosed with Heart Attack	Categorical
HadAngina	Ever Diagnosed with Angina	Categorical

HadStroke	Ever Diagnosed with Stroke	Categorical
HadAsthma	Ever Diagnosed with Asthma	Categorical
HadSkinCancer	Ever Diagnosed with Skin Cancer	Categorical
HadCOPD	Ever Diagnosed with COPD	Categorical
HadDepressiveDisorder	Ever Diagnosed with Depressive Disorder	Categorical
HadKidneyDisease	Ever Diagnosed with Kidney Disease	Categorical
HadArthritis	Ever Diagnosed with Arthritis	Categorical
HadDiabetes	Ever Diagnosed with Diabetes	Categorical
DeafOrHardOfHearing	Deaf or Serious Hearing Difficulty	Categorical
BlindOrVisionDifficulty	Blind or Serious Vision Difficulty	Categorical
DifficultyConcentrating	Difficulty Concentrating or Remembering	Categorical
DifficultyWalking	Difficulty Walking or Climbing Stairs	Categorical
DifficultyDressingBathing	Difficulty Dressing or Bathing	Categorical
DifficultyErrands	Difficulty Doing Errands Alone	Categorical

SmokingStatus	Smoking Status	Categorical
E-CigaretteUsage	Use of E-Cigarettes	Categorical
ChestScan	Had a CT or CAT Scan of Chest	Categorical
RaceEthnicity	Race Group	Categorical
AgeGroup	Age Group (5-Year Increments)	Categorical
HeightInMeters	Height in Meters	Numerical
WeightInKilograms	Weight in Kilograms	Numerical
BMI	Body Mass Index	Numerical
AlcoholDrinkers	Alcohol Consumption Status	Categorical
HIVTesting	Ever Tested for HIV	Categorical
FluVaxLast12	Received Flu Vaccine in the Last 12 Months	Categorical
PneumoVaxEver	Ever Received Pneumococcal Vaccine	Categorical
TetanusShot	Received Tetanus Shot in Past 10 Years	Categorical
HIVRiskBehaviorsLastYear	Engaged in High-Risk Behavior for HIV	Categorical

CovidPositive	Ever Tested Positive for COVID-19	Categorical

Summary Statistics (Numerical Fields)

Statistic	PhysicalHealth	MentalHealth	SleepHours	HeightInMeters	WeightInKilograms	BMI
count	246,013	246,013	246,013	246,013	246,013	246,013
mean	4.12	4.17	7.02	1.71	83.62	28.67
std	8.41	8.10	1.44	0.11	21.32	6.51
min	0	0	1	0.91	28.12	12.02
25%	0	0	6	1.63	68.04	24.27
50%	0	0	7	1.70	81.65	27.46
75%	3	4	8	1.78	95.25	31.89
max	30	30	24	2.41	292.57	97.65

3. Questions to Explore

1. Investigate the relationship between heart disease and variables such as age, sex, BMI, smoking status, and physical activity levels.
2. Is there a significant relationship between lifestyle choices (e.g., smoking, alcohol consumption, physical activity) and the occurrence of chronic diseases?
3. How does access to preventive healthcare services impact health outcomes?
4. Are there notable differences in health indicators across different states or regions?
5. What is the impact of sleep duration on overall health and the risk of chronic diseases?
6. How do mental health indicators correlate with physical health outcomes?
7. How prevalent are chronic conditions such as heart disease, diabetes, COPD, and arthritis in the surveyed population?