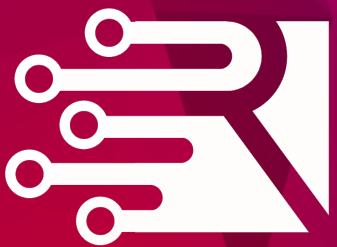
Jackson State U · U Michigan · UT San Antonio

DAIR^3

2026 Data Challenge: Vital Statistics



Juan B. Gutiérrez

Contents

Chapter 1

2022 Data Challenge - Vital Statistics

1.1 Problem Description

Your team is part of a consultancy supporting a planning commission for the State of Texas. The commission is planning budgetary requirements for various State services in 2030. The commission requests the following:

- Projections of underweight newborns by county in Texas. The CDC offers a data table of infant weight for age, available at:
  https://www.cdc.gov/growthcharts/html_charts/wtageinf.htm
  You can extract from this table the information related to weight at birth.
- Projections of newborn mortality by county in Texas. According to the CDC, a stillbirth is classified as either early, late, or term. An early stillbirth is a fetal death occurring between 20 and 27 completed weeks of pregnancy. A late stillbirth occurs between 28 and 36 completed pregnancy weeks. A term stillbirth occurs between 37 or more completed pregnancy weeks.
- Identification of the socioeconomic factors associated with these two outcomes.
- Comparison to other states.

Your team will produce an executive report accompanied by an appendix of technical material.

1.2 Data Description

The commission has approved specific data sources for this analysis. You can only use approved data available at this hyperlink.

National Center for Health Statistics (NCHS): Since 1969, all births recorded in the US are available as digital records from the NCHS, from the Centers for Disease Control and Prevention (CDC). The date and time of birth is publicly available only between 1969 and 1988; starting in 1989, only the week of birth is recorded. The NCHS keeps records of place of birth, assistance during delivery (at home, with doctors, with midwives), level of education of the parents, place of residence, weight at birth, number of weeks of gestation, number of siblings, birth order, etc. In total, over 100 variables are recorded. You have access to records form 1969 to 1988.

Surveillance, Epidemiology, and End Results Program (SEER): The U.S. Census Bureau annually releases unabridged population estimates for five-year age groups and race at the county level. The Census Bureau does not release bridged race estimates by single year of age at the county level due to concerns about the reliability of these estimates. However, these estimates are provided to the National Cancer Institute through SEER to meet programmatic needs such as the creation of age groupings that differ from the standard groupings used by the Census Bureau. Users of the single-year-of-age county-level interpolated race population estimates should carefully consider the limited reliability of these

estimates. County-level population files with 19 age groups (¡1, 1-4, ..., 80-84, 85+) and with 86 single-year age groups (¡1, 1, 2, ..., 84, 85+) are provided.

Socioeconomic Data and Applications Center (SEDAC): SEDAC, the Socioeconomic Data and Applications Center, is one of the Distributed Active Archive Centers (DAACs) in the Earth Observing System Data and Information System (EOSDIS) of the U.S. National Aeronautics and Space Administration (NASA). SEDAC contains georeferenced U.S. county-level population projections, total and by sex, race and age, based on shared socioeconomic pathways (SSPs). This data set, produced by Mathew E. Hauer, consists of county-level population projection scenarios of population in five-year intervals for all U.S. counties for the period 2020 - 2100. Obtain the data description from https://doi.org/10.1038/sdata.2019.5.

The NCHS data set covers from 1969 through 1986; it provides individual birth data. The SEER data set provides age-bracketed population estimates from 1969 to 2020. Since for this exercise we do not have birth data beyond 1986, you will have to use the SEER data to infer births in the period 1987-2020; alternatively, you could go to the NCHS source and obtain more recent data, however the NCHS data is complex and downloading additional years is not advised in the short time available to complete the Rowdy Datathon (but we will not stop you). The SEDAC data has total population estimates per county from 2020 through 2100 categorized in four ethnicities: Hispanic, white, black, and other. A viable sequence of analysis is NCHS → SEER → SEDAC.

Table 1.1: File sizes. If you have limited storage, plan your analysis in stages.

DATA SET	FILE	ZIPPED	UNZIPPED
NCHS	US1969-1986.zip	2.7 GB	22.8 GB
	natalityConfBackup_PostgreSQL.sql	2.42 GB	25 GB
	US1969.zip	32.6 MB	381 MB
SEDAC	hauer_county_NH_pop_SSPs.xlsx	N/A	15.1 MB
SEER	SEER Data Dictionary.pdf		73 KB
	tx.1969_2020.19ages.adjusted.txt.gz	5.3 MB	35 MB
	tx.1969_2020.singleages.adjusted.txt.gz	18.8 BM	19 MB
	tx.1990_2020.19ages.adjusted.txt.gz	6.5 MB	6 MB
	tx.1990_2020.singleages.adjusted.txt.gz	20.9 MB	21 MB
	us.1969_2020.19ages.adjusted.txt.gz	66.7 MB	430 MB
	us.1969_2020.singleages.adjusted.txt.gz	238.8 MB	1.6 GB
	us.1990_2020.19ages.adjusted.txt.gz	76.7 MB	520 MB
	us.1990_2020.singleages.adjusted.txt.gz	246.3 MB	1.8 GB
TOTAL		5.7 GB	52 GB

1.3 Criteria

Results will be evaluated according to requirements set by the commission:

- **Compelling presentation**: You must enable the commission to share your numerical and graphical results directly with legislators and citizens through executive summaries. This lay audience should find your summaries and implications to be understandable and convincing. Express your results in ways that can be acted on to plan e.g. funding of schools, care for the elderly, etc. The supporting documentation can be technical.
- **Analysis comprehension**: Before a single line of code is written, before a single byte of raw data is processed, you must be able to tell the story of what is the progression of steps that will be undertaken in analysis.
- **Sound technical methods**: You may cite the analyses of others, but the commission wants to see the methods that you have invented or adopted to calculate these projections (which should be accompanied by error bars, if possible). The commision must have confidence in your results in order to present those results to others.
- **Awareness of the data context**: All data have bias. Before, during and after analysis, it is essential to identify biases in the data and articulate clearly how these biases influence all steps of analysis and interpretation.
- **Reproducible results**: You must enable the commission to have your results confirmed by an independent team. That is, enable the independent team to replicate your results by describing your data and methods in detail.

Table 1.2: Evaluation Criteria

CRITERION	% WEIGHT
1. Compelling presentation - Informative	10%
2. Compelling presentation - Understandable	10%
3. Analysis comprehension	20%
4. Sound technical methods	20%
5. Awareness of the data context	20%
6. Reproducible results	20%

Chapter 2

Analysis Activities

2.1 A single year of vital statistics

A problem most people can relate to is demography. Millions of people are born in the US every year. Recording birth events is necessary for legal matters such as obtaining a driver's license or other forms of government-issued identification. However, recording, keeping, and using this information has challenges that exemplify many aspects of data analysis, as this exercise will demonstrate.

Since 1969, all births recorded in the US are available as digital records from the National Center for Health Statistics (NCHS) from the Centers for Disease Control and Prevention (CDC). Only between 1969 and 1988 the date and time of birth is publicly available; starting on 1989, only the week of birth is recorded. Why is the date and time of birth no longer recorded? Record your explanation as answer #1.

The NCHS keeps records of place of birth, assistance during delivery (at home, with doctors, with midwives), level of education of the parents, place of residence, weight at birth, number of weeks of gestation, number of siblings, birth order, etc. In total, over 100 variables are recorded.

I have made available two files for you: a compressed file with birth records from 1969, and a data dictionary. The uncompressed data file is about 380 MB in size. The data is contained in a "flat file". This means that every line of text in this file is a continuous chain of characters. We must extract information from this type of files with a "dictionary" that tells us the beginning and ending columns of a given variable. Why was this format used? Record your explanation as answer #2.

Please answer the following questions:

1. How many live births occurred in Texas in 1969 from mothers residing in Texas?
    - Bonus question: How would you visualize births from each state with respect to every other state?
2. Show graphically how the level of education of the mother is related to the birth order (1st born, second child, third, etc.)
    - Bonus question: How would you visualize each variable with respect to every other variable?

It is possible that you might not know how to answer some of these questions on first contact with this problem. As a senior undergraduate or beginning graduate student, you are expected to figure things out and solve new problems to which you have not been previously exposed... which involves reading, and asking questions to your peers and instructors. A common approach to extract information from flat files is by importing it into Excel. The "*Text Import Wizard*" would guide you through the process of identifying variables by column. However, this results in a problem. Describe it. Record your explanation as answer #3.

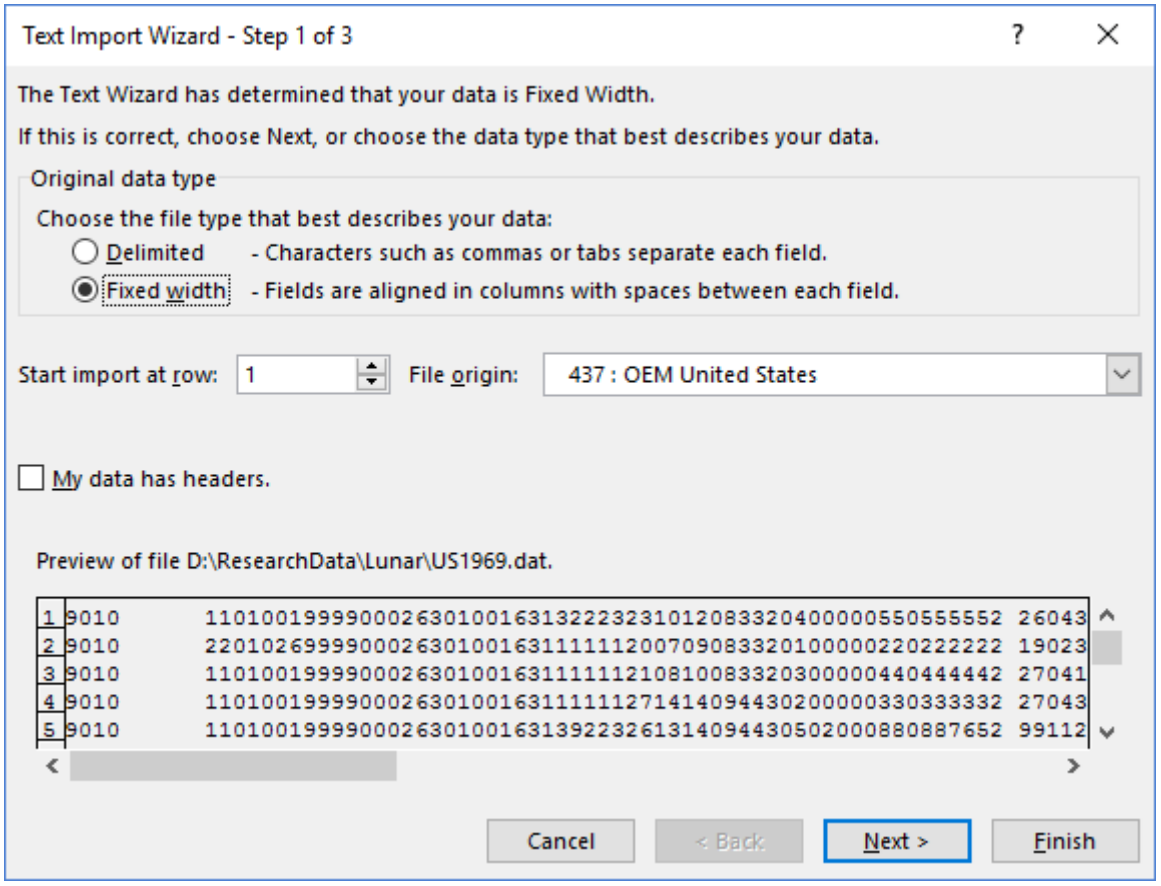


Figure 2.1-1: Excel's Import Wizard

A simple program that can help you explore this file is

```
f = open("US1969.dat", "r", encoding="cp1252")
counter = 0;
for x in f:
    if x[25] == "7" and x[26] ==  "4": # other conditions?
        counter = counter + 1
print(counter)
```

The instruction *encoding="cp1252"* is necessary in non-Windows systems due the presence of single-byte character encoding of the Latin alphabet, used by default in the legacy components of Microsoft Windows for English and many European languages including Spanish, French, and German. If you remove this instruction, the following error might show up in non-Windows operating systems: "'utf-8' codec can't decode byte..."

Chapter 3

Analysis Activities

3.1 Unit 1: Responsible Conduct of Research and Ethics

3.1.1 Learning Objectives

- Analyze the sociotechnical system of biomedical data science
- Differentiate between traditional bioethical, sociotechnical, and other ethical approaches
- Identify ethical issues in secondary use of vital statistics data

3.1.2 Activities

3.1.2.1 1.1 Ethical Considerations in Vital Statistics Analysis

Consider the NCHS vital statistics data you will be analyzing:

1. Why is the date and time of birth no longer recorded after 1988, only the week of birth? Discuss privacy implications. Record your explanation as answer #4.

2. Identify at least three ethical concerns when projecting underweight newborns and infant mortality by county. Consider:
   - Stigmatization of specific counties or populations
   - Secondary use of data originally collected for other purposes
   - Potential biases in historical data collection
3. Develop a stakeholder engagement strategy for presenting your findings to affected communities
4. Write a brief anticipatory governance framework for how your projections might be used by policymakers

3.2 Unit 2: Data Management, Representation, and Sharing

3.2.1 Learning Objectives

- Understand data management principles and metadata requirements
- Choose appropriate data representations for analysis tasks
- Apply FAIR principles to data sharing

3.2.2 Activities

3.2.2.1 2.1 Data Management - Understanding the NCHS Dataset

1. Document the metadata for the NCHS vital statistics:
    - List 5 critical metadata categories needed to reproduce your findings
    - Create a data dictionary for the key variables you will use
    - Explain why the data was stored in "flat file" format. What are the advantages and disadvantages? Record your explanation as answer #5.

3.2.2.2 2.2 Data Representation Choices

1. The NCHS data is provided as a flat file with over 100 variables. Design three different representations:
    - A relational database schema (tables and relationships)
    - A document-oriented (NoSQL) structure
    - A graph database representation
2. For each representation, identify which analyses become easier or harder
3. Implement your relational design using the PostgreSQL import process described

3.2.2.3 2.3 From CSV to PostgreSQL Implementation

[Keep existing PostgreSQL section from original document]

3.2.2.4 2.4 Data Sharing Plan

1. Create a 2-page Data Management and Sharing Plan following NIH requirements for your analysis results
2. Apply FAIR principles to your processed datasets:
    - Findable: Assign persistent identifiers
    - Accessible: Define access protocols
    - Interoperable: Use standard vocabularies
    - Reusable: Document provenance and license

3.3 Unit 3: Rigorous Statistical Design

3.3.1 Learning Objectives

- Develop appropriate study designs for research aims
- Create analytic plans with power analyses
- Identify sources of bias and interpret findings appropriately

3.3.2 Activities

3.3.2.1 3.1 Study Design for Projection Analysis

Research Aim: Project underweight newborns and infant mortality by Texas county for 2030

1. Propose a study design that addresses:
    - Historical trend analysis using NCHS data (1969-1986)
    - Integration with SEER population data (1987-2020)
    - Projection using SEDAC data (2020-2030)
2. Discuss strengths and weaknesses of your approach
3. Identify potential confounders and how to address them

3.3.2.2 3.2 Analytic Plan and Power Analysis

1. Develop a detailed analytic plan including:
    - Time series regression methods for projection
    - Handling of missing data and data gaps (1987-present birth records)
    - Uncertainty quantification for projections
2. Conduct power analysis for detecting county-level differences
3. Calculate sample sizes needed for reliable projections

3.3.2.3 3.3 Bias Assessment and Causal Interpretation

1. Identify specific biases in the observational data:
    - Selection bias in birth recording
    - Information bias in weight measurements
    - Confounding by socioeconomic factors
2. Produce mock results with confidence intervals
3. Write an interpretation addressing limitations

3.4 Unit 4: Design and Reporting of Predictive Models

3.4.1 Learning Objectives

- Apply dimension reduction techniques
- Build and evaluate classification models
- Follow TRIPOD guidelines for reporting

3.4.2 Activities

3.4.2.1 4.1 Data Preparation and Feature Engineering

1. Implement Principal Component Analysis (PCA): [Keep existing PCA implementation from Linear Discriminants section]
2. Apply feature selection for predicting underweight births:
    - Use correlation analysis
    - Apply LASSO regularization
    - Implement recursive feature elimination
3. Document your choices following TRIPOD guidelines

3.4.2.2 4.2 Predictive Modeling

1. Build models to predict:
    - Birth weight categories (underweight, normal, overweight)
    - Risk of infant mortality
2. Compare multiple approaches:
    - Logistic regression
    - Random forests
    - Gradient boosting
3. Evaluate using appropriate metrics (AUC, calibration plots)
4. Address data leakage concerns

3.4.2.3 4.3 TRIPOD-Compliant Reporting

1. Complete the TRIPOD checklist for your predictive model
2. Create a model card documenting:
    - Model purpose and limitations
    - Training data characteristics
    - Performance metrics by subgroups
    - Potential biases and fairness considerations

3.5 Unit 5: Reproducible Workflows

3.5.1 Learning Objectives

- Create reproducible analysis pipelines
- Implement version control and containerization
- Follow best practices for scientific computing

3.5.2 Activities

3.5.2.1 5.2 Code Notebooks for Exploratory Analysis

1. Create Jupyter notebooks for:
    - Initial data exploration (EDA)
    - Single year analysis (1969 Texas births)
    - Time series visualization
2. Use Quarto to create reproducible reports
3. Include markdown documentation throughout

3.5.2.2 5.3 Automation and Containerization

1. Create a Makefile with targets:
    - download-data: Fetch NCHS, SEER, SEDAC files
    - process-data: Clean and merge datasets
    - run-analysis: Execute all analyses
    - generate-report: Create final report
2. Build Docker container with all dependencies
3. Test full pipeline reproducibility

3.6 Unit 6: Meta-analysis Integration

3.6.1 Learning Objectives

- Synthesize findings across multiple data sources
- Account for heterogeneity and dependencies
- Conduct systematic comparisons

3.6.2 Activities

3.6.2.1 6.1 Systematic Comparison with Other States

1. Conduct meta-analysis of infant mortality across states:
    - Calculate effect sizes for each state
    - Test for heterogeneity (I^2 statistic)
    - Create forest plots
2. Account for correlation between neighboring states
3. Identify states with similar patterns to Texas

3.6.2.2 6.2 Integrating Multiple Evidence Sources

1. Combine projections from:
    - Your time series model
    - SEDAC population projections
    - Published literature on infant mortality trends
2. Weight evidence by quality and relevance
3. Produce ensemble projections with uncertainty

3.7 Unit 7: Transformer-based AI Applications

3.7.1 Learning Objectives

- Apply LLMs to assist in data analysis
- Create consensus-based analytical approaches
- Develop AI-augmented analysis pipelines

3.7.2 Activities

3.7.2.1 7.1 LLM-Assisted Code Generation

1. Use LLMs to generate code for:
    - Data cleaning and validation scripts
    - Statistical analysis functions
    - Visualization templates
2. Validate generated code against test cases
3. Document prompts that produced best results

3.7.2.2 7.2 Consensus Analysis Framework

1. Query multiple LLMs about:
    - Interpretation of unusual patterns in data
    - Selection of appropriate statistical methods
    - Identification of potential confounders
2. Aggregate responses using consensus methods
3. Compare LLM suggestions with traditional approaches

3.7.2.3 7.3 Automated Report Generation

1. Create pipeline using LLMs to:
    - Summarize findings for executive summary
    - Generate plain-language explanations of methods
    - Produce policy recommendations
2. Review and validate AI-generated content
3. Ensure ethical use and proper attribution

3.8 Integration Exercise: Complete Analysis Pipeline

3.8.1 Comprehensive Project Requirements

Integrate all units to produce the final deliverable for the planning commission:

1. **Executive Report** (3-5 pages):
    - County-level projections for 2030
    - Key socioeconomic factors identified

- Comparison with other states
- Policy recommendations

2. **Technical Appendix** (unlimited):
   - Complete methodology
   - Reproducible code and workflows
   - Validation results
   - Uncertainty quantification
   - Ethical considerations
   - Data management plan
3. **Reproducibility Package**:
   - GitHub repository with version control
   - Docker container
   - Automated pipeline (Makefile)
   - Complete documentation