FOUNDATIONS OF DATA ANALYTICS

2022 Data Companion

Juan B. Gutiérrez

Cover art: This image was created with the assistance of DALL·E 2 by OpenAI.
“Lady Justice holding uneven scales. She is in front of a computer. The scales
are tipped toward the right. One-line drawing.” Generated on July 5, 2023.

Contents

1 2022 Data Challenge - Vital Statistics 2
1.1 Problem Description 3
1.2 Data Description 3
1.3 Criteria 5

2 Analysis Activities 6
2.1 A single year of vital statistics 7
2.2 Linear Discriminants 8
2.3 How to import a CSV file into PostgreSQL 9
2.3.1 From CSV to PostgreSQL 9
2.3.2 1. Create a PostgreSQl function 10
2.3.3 2. Split the large file into small multiple files 10
2.3.4 3. Iterate through all the small files 10
2.3.5 Questions regarding data imports 10
2.4 Multiple years of vital statistics 10
2.5 General Questions 11

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:

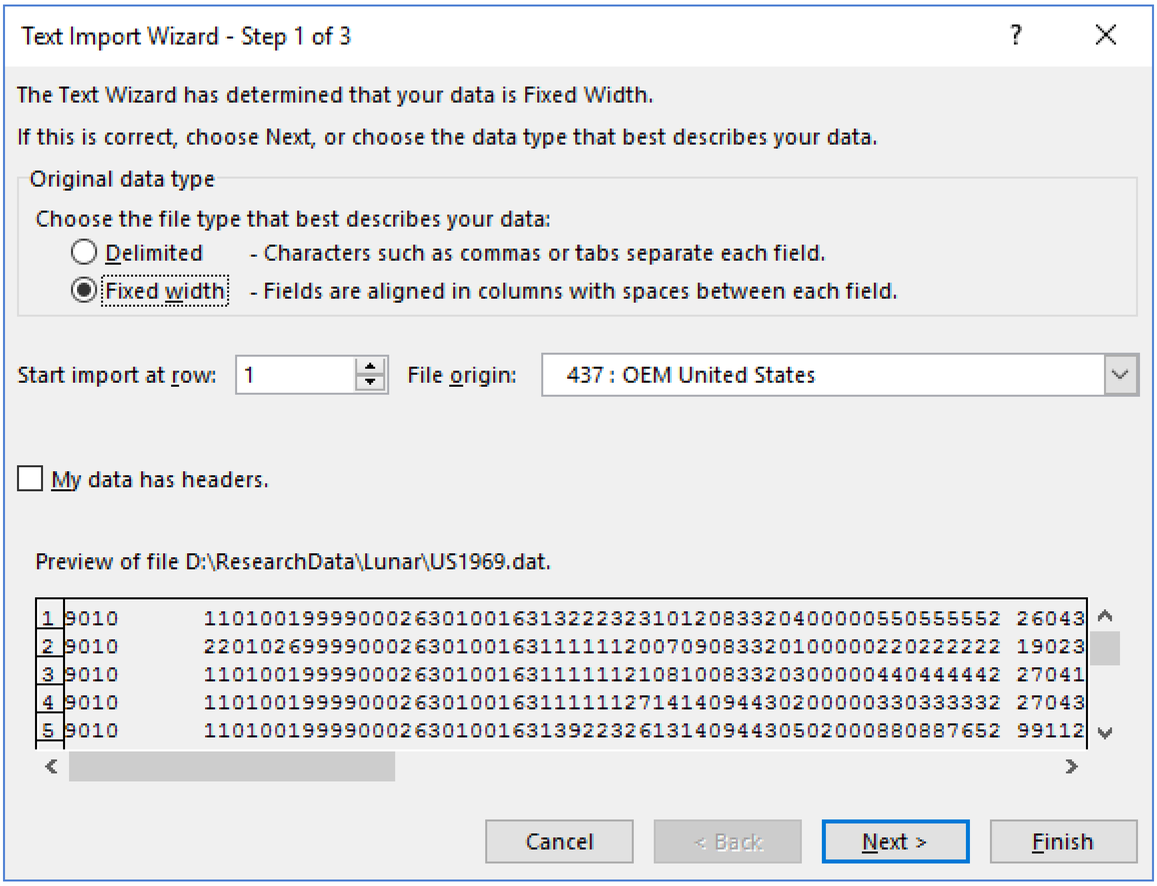
1. How many live births occurred in Texas in 1969 from mothers residing in Texas?

    - 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.



The Text Wizard has determined that your data is Fixed Width.

If this is correct, choose Next, or choose the data type that best describes your data.

Original data type

Choose the file type that best describes your data:

○ Delimited - Characters such as commas or tabs separate each field.

● Fixed width - Fields are aligned in columns with spaces between each field.

Start import at row: 1 File origin: 437 : OEM United States

□ My data has headers.

Preview of file D:\ResearchData\Lunar\US1969.dat.

```
1 9010       110100199990002630100163132223231012083320400000550555552 26043
2 9010       220102699990002630100163111111200709083320100000220222222 19023
3 9010       110100199990002630100163111111210810083320300000440444442 27041
4 9010       110100199990002630100163111111271414094430200000330333332 27043
5 9010       110100199990002630100163139223261314094430502000880887652 99112
```

Figure 2.1-1: Excel's Import Wizard

2.2 Linear Discriminants

Implement a principal component analysis according the the following procedure (use NumPy to conduct matrix operations):

1. Select weight at birth plus other numerical variables from the NCHS data set. We will call this the *feature matrix.*

2. Ensure that the mean is zero for every variable. How do you achieve this? Record your explanation as answer #4. We will call this the *normalized feature matrix.*

3. Compute the inner product of the normalized feature matrix. What is the dimension of the resulting matrix? Record your explanation as answer #5.

4. Compute eigenvalues and eigenvectors for the inner product of the normalized feature matrix. You will need to sort eigenvalues by magnitude. How do you do that in Python? Record your explanation as answer #6.

5. Project the data in the feature matrix along the eigenvectors corresponding to the two largest eigenvalues. Plot each observation in this projected space

according to quartile of weight at birth. You might need to pick different variables until a rough pattern emerges. Interpret the plot. Record your explanation as answer #7.

Note that you are not expected to simply call a PCA library.

2.3 How to import a CSV file into PostgreSQL

This is many times the most difficult step. Data is often presented in CVS format. PostgreSQL offers an easy way to import CVS files through the *COPY* instruction, provided that a table exists to hold the data. Hence, the first step is to create a table.

For instance, if you have a CVS file with three columns, the table could be something like:

```
CREATE TABLE MyDataTable (
  column1 SERIAL,
  column2 VARCHAR(50),
column3 DATE,
  PRIMARY KEY (column1)
)
```

This instruction contains the statement *CREATE TABLE* which is part of a subset of the SQL language called Data Definition Language (DDL). Note that each column has to be defined with a data type. Once the table is created, you have to import the data with the *COPY* instruction.

```
COPY MyDataTable(column1, column2, dob, column3)
FROM 'MyDataTable.csv'
DELIMITER ','
CSV HEADER;
```

2.3.1 From CSV to PostgreSQL

In this section you will use a file that already has been organized with 17 years of data. This is a comma separated value (CSV) file created from public records obtained from the Center for Disease Control, Division of Vital Statistics. To produce the file, the data dictionaries for all 17 years were analyzed and homogenized. The source code to import this data is presented here for reference. In total there are 56,545,325 records.

The file is located at Google Drive (follow this link)

The challenge is that the *COPY* instruction does not work for such a large file. Instead, we will have to: (1) create a function that resides inside of PostgreSQL, (2) split the large file into small multiple files, and (3) iterate through all the small files.

2.3.2 1. Create a PostgreSQl function

Download the file *spImportNatality.sql* from Blackboard. Execute it to generate the function inside of PostgreSQL. You can attempt this step on your own. This step will be completed in class.

2.3.3 2. Split the large file into small multiple files

Download the file *CSVSplitter.zip* from Blackboard. It is a compressed file that contains an executable file. We will use this program to split the large file into many small files, each with 100,000 rows. You can attempt this step on your own. This step will be completed in class.

2.3.4 3. Iterate through all the small files

Observe where the iteration occurs in the SQL function. Try to use all the files. Document the difficulties you will find.

2.3.5 Questions regarding data imports

1. Did the CVS splitter run out of memory? How could you tell? What alternative solution could you produce?

2. The iteration in the function works well for files with suffix '.000', '.001', etc., but it will fail with suffix '.010' and above. How can you fix this?

3. Solve the questions related to SQL from the latest homework assignment.

2.4 Multiple years of vital statistics

1. Write a Python program that computes an integral. How do you demonstrate that it works? Show it. Record your explanation as answer #8.

2. Retrieve the time series of sex ratios for every day of the year using all the data the natality data set. Pick one month. Approximate this function with a Fourier series. Decide how many basis functions you need so that you can represent the data "faithfully" (you also have to decide what '*faithfully*' means). Explain. Now repeat with a monomial basis. Record your explanation as answer #9.

3. Repeat the previous step with a discrete Fourier transform. You need to address how to compute a Fourier inverse. This is trivial, look it up. Compare the results of the Fourier transform and a Fourier series. Record your explanation as answer #10.

4. Select several variables from the Natality table. Conduct a PCA, FLD, and MD analysis. How can you use a pseudoinverse to describe the data? Why and how is the vector **x** found with a pseudoinverse different from the mean? Record your explanation as answer #11.

2.5 General Questions

- **Beginner**: Conduct Exploratory Data Analysis (EDA) on the dataset. What can you say about the number of births in Texas? What variables are correlated to the number of births? Here are some questions to get you started:
  - What can you say about the number of births in Texas?
    - How does the number of births vary over time? (e.g.: births in the year 1969 versus births in the year 1987)
    - What about different regions in Texas? Do the number of births differ?
    - What is the relation between weight at birth and gestation length?
    - How does the sex ration changes with gestation period?
  - What variables are correlated to the number of births?
    - What variables are most strongly correlated to the number of births?
    - What variables are directly correlated to the number of births?
    - What variables are inversely correlated to the number of births?
  - Answer any other questions you're wondering! *Get curious!*
- **Intermediate**: Produce a time series regression to project the number of stillbirths and underweight babies to the year 2030 .
- **Advanced**: Build a deep learning model to predict a child's weight at birth with ranges based on the CDC table of infant weight for age, as explained in the problem definition. Additionally, build a different learning model to try to estimate the risk of death; note that you will need to identify the variables that correlate with mortality.