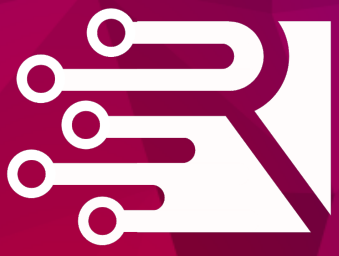
University of Texas at San Antonio

2023 Rowdy Datathon

Data Challenge



Juan B. Gutiérrez

October 28, 2023

Contents

1 2023 Data Challenge - Educational Attainment **3**
1.1 Problem Description 4
1.2 Data Description 4
1.3 Criteria 5

2 Analysis Activities **7**
2.1 Data Source 8
2.2 Tasks for all Preparation Levels 8
2.2.1 Background Information 8
2.2.2 Generative AI 8
2.2.3 Data Challenge Text File 8
2.3 Tasks for the Beginner Level 9
2.4 Tasks for the Intermediate Level 10
2.5 Tasks for the Advanced Level 10

Chapter 1

2023 Data Challenge - Educational Attainment

1.1 Problem Description

NOTE: What follows is a hypothetical scenario to provide context, not an actual consultation from a congressperson.

In this scenario, as part of a team of congressional interns working in the office of a notable US Representative, your job is to design a strategy to allocate resources to increase educational attainment.

The report assigned to you must include:

- Determination of the socioeconomic and/or environmental factors that influence educational attainment.
- A recommendation on how to allocate resources based on the need to address the most significant effector(s) identified.
- In this scenario, the congressperson you work for represents a district in Texas. Thus, you must conduct the analysis for Texas and compare it to other states.

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

1.2 Data Description

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

2017-18 Civil Rights Data Collection (CRDC)
(https://ocrdata.ed.gov/). This resource informs achievement metrics in every middle and high school in the nation.

Small Area Income and Poverty Estimates (SAIPE)
https://www.census.gov/data/datasets/2017/demo/saipe/2017-school-districts.html
The "Small Area Income and Poverty Estimates (SAIPE)" program provides annual estimates of income and poverty statistics for school districts.

2017 School Locations & Geoassignments (SLGA)
(https://nces.ed.gov/programs/edge/geographic/schoollocations). This data set contains address geocodes (estimated latitude/latitude values) and other geographic indicators to public schools, public local education agencies, private schools, and postsecondary schools. The geographic data are provided as shapefiles, and basic attribute data are available as Excel and SAS tables

2017 Home Mortgage Disclosure Act (HMDA) database
(https://www.consumerfinance.gov/data-research/hmda/historic-data/).
This resource informs the amount and rate of mortgages by census tract.

2017 School District Geographic Relationship Files (SDGR)
(https://nces.ed.gov/programs/edge/Geographic/RelationshipFiles).

National Center for Education Statistics at the Department of Education. Most school districts in the U.S. are independent local governments that have authority to determine their geographic boundaries. These boundaries may or may not be consistent with boundaries for other types of legal and statistical areas like counties, Congressional Districts, or Census tracts. As a result, school districts may have multiple spatial associations with other types of geographic areas e.g., a school district boundary may include territory in two different counties, or intersect three different Congressional Districts. The NCES EDGE school and agency location files and the NCES Common Core of Data (CCD) provides a limited set of geographic associations based on the location of the district administrative office.

2020 Address Count Listing Files of the Bureau of the Census (ACLF) (https://www.census.gov/geographies/reference-files/ 2020/geo/ 2020addcountlisting.html). The files include total housing units (including transitory units) and total group quarters counts, by 2020 census tabulation block. You have the Texas file available in the data folder. You need to download each state individually.

Opportunity Atlas (OA)
(https://www.opportunityatlas.org/) The Opportunity Atlas uses anonymous data following 20 million Americans from childhood to their mid-30s to determine the level of income related to where people grew up, for those born between 1978 and 1983... which means year 2017.

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

DATA SET	FILE	ZIPPED	UNZIPPED
CRDB	2017-18-crdc-data.zip	0.1 GB	0.9 GB
SAIPE	ussd17.xls	N/A	0.001 GB
SLGA	EDGE_GEOCODE_PUBLICLEA_1718.zip	0.001 GB	0.001 GB
HMDA	hmda_2017_nationwide_all-records_labels.zip	0.9 GB	10 GB
SDGR	GRF17.zip	0.06 GB	0.17 GB
ACLF	48_Texas_AddressBlockCountList_062022.txt	N/A	0.003 GB
TOTAL		1.1 GB	11 GB

1.3 Criteria

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

- **Skill level**: The level of inquiry must be proportional to skills. A participant who claims a lower level of expertise could be disqualified.
- **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 Data Source

You should have received a flash drive with all files. As a backup, the folder 2023 Rowdy Datathon Data Challenge contains the data.

2.2 Tasks for all Preparation Levels

2.2.1 Background Information

Read the Preface and chapters 1, 2 & 3 in the *Rowdy Datathon Supplementary Material.*

2.2.2 Generative AI

Complete the following tasks:

- Go to ChatGPT. Register for a free account using Version 3.5.
- Next to your name, on the bottom left area of the interface, select the three dots, which will show Custom Instructions.
- In the field for "*What would you like ChatGPT to know about you to provide better responses?*" enter something along the lines of "I am a student working on a data competition. My areas of expertise are..."
- In the field for "How would you like ChatGPT to respond?" enter something along the lines of "I need source code with ample documentation and verification steps." Add other criteria you might want to use.

The customization of ChatGPT allows you to define parameters that will provide better answers to your queries. You might need to refine these instructions multiple times to obtain optimal results.

2.2.3 Data Challenge Text File

We will illustrate the problem of data reproducibility with three simple questions: (1) How many school districts were there in total in 2017-18 using Excel?, (2) How many school districts were there in 2017-18 using Python? (3) How many school districts were there in 2017-18 with a different data set?

We will begin by using the CRDB database. A common approach to extract information from comma-separated or flat files is by importing it into Excel. The "*Text Import Wizard*" would guide you through the process of identifying variables by column. Import it and determine the total number of school districts. Record your explanation as answer #1.

A simple program in Python to help you explore this file is described below:

```
f = open("LEA Characteristics.csv", "r", encoding="cp1252")
counter = 0;
for x in f:
    counter = counter + 1
print(counter)
```

The previous file counts the total number of rows in the file. Does this number coincide with the number produced by Excel? Record your explanation as answer #2.

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..." Why would you need *encoding="cp1252"* for US data? Record your explanation as answer #3.

Now, turn your attention to the SLGA database. The following program will count the number of records.

```
f = open("EDGE_GEOCODE_PUBLICLEA_1718.csv", "r",
    encoding="cp1252")
counter = 0;
for x in f:
  counter = counter + 1
print(counter)
```

This number does not match the CRDB database, but both databases claim to have the list of school districts in the US in the period 2017-18. Explain why. Record your explanation as answer #4.

At this point, imagine that you are given the task of finding out the number of school districts in the nation. Which file do you trust more? Both are federal databases. Two different analysts might end up with different results using different datasets, or, worse, using the same database, as we showed with the exploration of the CRDC file. Reflect on this. Record your explanation as answer #5. .

2.3 Tasks for the Beginner Level

1. Access the file:
   EDGE_GEOCODE_PUBLICLEA_1718.xlsx
   from the SLGA data set. Export it to CSV. This file contains latitude and longitude for each school district in the US. Access *"2023 Rowdy Datathon Supplementary Material"* , section 2.1. Follow these instructions to map all school districts in the US

2. Follow the instruction in section 2.2 of *"2023 Rowdy Datathon Supplementary Material"* to color each county in the US with a color map of the number of school districts per county.

3. Plot the number of students per school district vs. the number of students in poverty for all school districts in the US. You will need the CRDC and the SAIPE data sets.

2.4 Tasks for the Intermediate Level

1. Let's focus on the CRDC dataset. Open the file "`Advanced Mathematics.csv`" Copy the first two lines of text and ask chatGPT to generate the following output:

   ```
   Generate a PostgreSQL script to import a CSV file.
   Next are the first line with columns names and a line of data:

   [first two lines of text]
   ```

2. The code by ChatGPT will not work. Follow instructions on Sections 8.4 and 8.5 of *"2023 Rowdy Datathon Supplementary Material"* to import this data. Do the same for all files in the folder SCH.

3. Plot the number of students per school district vs. the number of students in poverty for all school districts in the US. Explore how all subject matters, which you have now imported into PostgreSQL, are related to poverty. **Do data aggregation and query in PostgreSQL. Plot in your preferred programming language**.

4. Select a location you are familiar with. Go to the Opportunity Atlas. Download the data for that location. It will have the census tract as a geographic ID. Use the SDGR database to map school districts to census tracts and determine whether the income reported in the Opportunity Atlas has correlation with academic performance in high school.

2.5 Tasks for the Advanced Level

1. Plot the number of students per school district vs. the number of students in poverty for all school districts in the US. You will need the CRDC and thge SAIPE data sets. There are challenges in loading the data. Do not edit any raw files.

2. Load the HMDA file. This is a 10 GB file, and you will need to filter data from multiple perspectives. Choose between RAM (e.g. a data frame) or PostgreSQL. Your choice will heavily influence your ability to query the data.

3. Use the HMDA dataset to determine the distribution of mortgage rates, amounts and denials per school district. To do this, you will need to map zip codes and census tracts to school districts through the SDGR data set.

Use the ACLF data set to estimate incidence of denials per household per census tract.

4. Explore any possible correlation between performance in mathematics and writing against mortgage data per school district.

Acknowledgements

Acknowledgements for the Rowdy Datathon must go beyond mere formalities, for the event is a monumental effort that reflects a commitment to making data accessible and comprehensible to all.

Special thanks are due to the Student Chapter of the Association for Computing Machinery at the University of Texas at San Antonio: Roni Maddox, Poonacha Cheppudira, UTSA alumna Jenelle Millison, and many others. The event would have not been possible without the support of the School of Data Science at UTSA and the National Security Agency. To all contributors, your collective efforts have culminated in a Datathon that serves as both a platform for applied data science and as an educational crucible for emerging data analysts.

The phrase "Data is everywhere, therefore data should be for everyone" is not merely a tagline; it encapsulates the philosophy that guided us through countless meetings, debugging sessions, manual tests, and problem-solving endeavors. This project is not isolated but forms a part of our larger mission: to guide aspiring data analysts through the multi-faceted landscape of data science. The Datathon aims to bridge the gaps in a fragmented educational landscape, offering a holistic view of data analytics that is sorely needed in the field. Thank you for your time, your expertise, and most importantly, your unwavering commitment to the democratization of data.