Branch: master ▾

Find file | Copy path

DSC180A-Fair-Policing / assignment-1.md

colerichmond fixed typo in assignment-1.md

f4b98f7 5 days ago

2 contributors

Raw | Blame | History

99 lines (71 sloc) 3.69 KB

Assignment #1: The Data (Due Date: Jan 24th, 11:59 PM)

In this assignment, you will:

1. Write a survey of the data and the context in which it was created (report).
2. Describe and justify the data ingestion process in part 3 (report).
3. Develop code for ingesting and storing the data for later use (code).

The report portion of the assignment should be written in *markdown*, saved as a pdf, and uploaded to Gradescope.

The data ingestion code should be submitted to Gradescope as a programming assignment.

Part 1

Introduce the problem being investigated and describe the data being used to approach the problem. That is, describe the investigation into racial discrepancies in traffic stops by the SDSU investigation into the San Diego Police department (referring to the investigation in *Suspect Citizens* for context).

Address the appropriateness of the data design and collection:

- Why is the data appropriate to address the problem?
- What are the potential shortcomings of the data for addressing the problem?
- What data have been used to address this problem in the past? (Historical context).

Summarize relevant details of the data generating process, describing the population that the data represents, whether that population is relevant to the question at hand, while addressing possible questions of data reliability.

The material in this section should be informed by the listed background readings and the introduction/data explanation sections of the main paper (with particular attention made to Appendix 2 in the SDSU study).

Part 2

Describe the data ingestion process you designed. This description should:

- Specify from where the data originates, addressing legal issues pertaining to access.
- Address any data privacy concerns and how your data pipeline handles them.
- Lay out the schema and justify the decisions (what's the unit corresponding to an observation? What are the storage considerations?)

- Address the applicability of the pipeline to similar data sources you might anticipate using in your future work on the subject (what might those be?).

Part 3

In a private GitHub repository for your project, structured according to the methodology portion of the course, create a data ingestion pipeline for the result-replication project. The pipeline should:

- Ingest Traffic Stops data from the San Diego Open Data Portal into local file(s) on disk, according to best practices laid out in the methodology HW. All files should have the same schema! The ingestion pipeline should take in the year (between 2014 and 2019) as a parameter. Note that data post-2018 is structured differently according to RIPA (Racial and Identity Profiling Act).
- As a bonus, write your data ingestion code to write the data to a local sqlite database. This will be useful when working with multiple years of data (or multiple geographies).
- Store the data according to your designed schema, taking care to appropriately type the data and implement the best storage design (which columns are needed and appropriate). The data will eventually be stored in a database format both due to the size of the total collection of data, as well as for comparison to the Stanford Open Policing datasets.
- The stored data should be in a form most appropriate for assessment and cleaning (EDA). You may find it useful to compare your dataset with the cleaned SDPD data in the Stanford Open Policing Dataset.