Midterm material

CMSC 320

This document describes material that will be fair game in the midterm exam. Each section is divided into two levels (level 1 and 2). Mastery of level 1 material is essential to do well in the midterm, level 2 is needed to do great in the midterm.

Preliminaries

Level 1

- Data Analysis Cycle: acquisition -> preparation -> modeling -> communication

Level 2

- Data Analysis Cycle: as presented in slides/Zumen & Mount

Measurement types

Level 1

- categorical
- ordered categorical (ordinal)
- discrete numerical
- continuous numerical

Level 2

- factors/levels in R
- the importance of units

Best practices

Level 1

- the importance of reproducibility
- tools to improve reproducibility
- data science ethics and responsible conduct of research

Level 2

- the importance of thinking like an experimentalist

Data Wrangling

Level 1

- `dplyr` single table operations
- the Select-From-Where SQL query
- different join semantics
- why are database systems helpful and useful?

Level 2

- Keys/Foreign Keys in the Entity-Relationship data model
- How an ER diagram is converted into a set of Relations (data tables)
- Database query optimization principles

Tidy Data and Data Models

Level 1

- Components of a Data Model
- Basics of the Entity-Relationship and Relational Data Models
- The components of an ER diagram
- The relationship between tidy data, the ER and the Relational models

Level 2

- JSON

Data cleaning

Level 1

- The gather and spread data tidying operations
- Regular expression basics
- Tools to extract and clean text data

Level 2

- The document-term model for text representation
- The *one_term_per_row* tidy text representation

Midterm Structure

The midterm will consist of three sections: ~8-10 multiple choice questions, ~5-7 short questions, and 1 or 2 longer questions. Multiple choice will test concepts and definitions along with problems similar to written exercises in class. Short questions will be similar to written problems done in homework, along with concept questions where longer written answers are required. Longer questions are for problem solving (e.g., design a data pipeline or SQL queries to carry out a specific task).