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

References

- Lecture Notes Ch. 2-4

Measurement types

Level 1

- categorical
- ordered categorical (ordinal)
- discrete numerical
- continuous numerical
- text, datetime

Level 2

- factors/levels in R
- the importance of units

References

- Lecture Notes Ch. 5
- HW 1

Data Manipulation Operations

Level 1

- single table operations (selecting attributes, filtering entitites)
- more single table operations (sorting, creating new variables, summarization, grouping entities *group by*)
- `dplyr` operation pipelines
- the multiple types of joins

References

- Lecture Notes, Ch. 6,7,13
- HW 1, 2

Basic plotting

Level 1

- The data/mapping/geometry definition of data visualizations

Level 2

- Frequent used plots: scatterplot, bar graph, histogram, boxplot

References

- Lecture Notes, Ch. 8

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

References

- Lecture Notes, Ch. 10

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

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

Rerefernces

- Lecture Notes, Ch. 11

SQL and Database Systems

Level 1

- the difference between declarative and procedural representation of data operations
- the Select-From-Where SQL query
- Joins in SQL

Level 2

- SQL as a data definition language
- Views
- Database query optimization principles
- JSON

References

- Lecture Notes, Ch. 12, 14, Ch. 15
- HW 2

Data scraping

Level 1

- The hierarchical structure of HTML documents
- Basic CSS selector syntax: type, class, id, attribute

References

- Lecture Notes 16.2

Data cleaning

Level 1

- Common problems in data tidying
- The gather and spread data tidying operations (data values as headers)
- Normalizing data tables (More than one entity in a table)
- 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

References

- Lecture Notes 17, 18.1

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). You can bring 1 double sided 8.5x11in sheet of notes to the exam.