Homework Set 5 - EMR

Purpose: In class we learned to ask the question "Where's the Data Dictionary?" In this homework you will be working with a dataset from the medical world. Fortunately for you it has a nice data dictionary with primary/foreign key mapping and everything. As you work the exercise you will see how there is information you need and how you can acquire that from the metadata. Put another way, when you pickup someone else's dataset it is really helpful if it has good documentation. In this homework you will experience that first hand and take that with you when you have to make your own data and documentation.

Instructions: Put your solutions for all problems in a google colaboratory notebook. Create a level-1 header for each problem and then present the solution in subsequent cells. The header must contain the problem number followed by "confidence x" where x is an integer from 1 to 10 with 10 representing high confidence. For full credit you must show your work. The submission is made via a link on UVACanvas.

Domain Extras: We will be working with the data from the Synthea project. Their systems allows you to generate synthetic medical data. For this homework we will use the data in their csv format (which is just all of the tables from a relational database stored individually). We will see this data again when we work with it in its true form, FHIR, and we will be using is to populate a MongoDB.

Data

The data is available from the synthea downloads site `https://synthea.mitre.org/downloads`. Use the file linked behind "1K Sample Synthetic Patient Records, CSV ". It is a zip file that is 56 MB.

N.B. For problems 5.1 through 5.5 we will be using pandas. In problem 5.6 we will use SQL.

5.1 Data Preparation (5ψ)

Confirm that there are no duplicate patients in the dataset.

5.2 Data Analysis (10ψ)

Print out the patient name and out-of-pocket cost for medication for each patient with the acute bronchitis condition. Also print the total number of patients and total out-of-pocket costs.

5.3 Data Analysis (10ψ)

Print out the patient name and total duration of all care plans for each patient with the acute bronchitis condition. Also print the total number of patients and total duration of all care plans.

5.4 Data Visualization (10ψ)

5.4.1

Make a chart showing the histogram of total amount paid on each claim.

5.4.2

Make a second chart for total amount paid up to $80,000

5.5 Press Release (10ψ)

Explore the data dictionary and create your own query (have a little fun, be an investigator). Then create a press release explaining your finding. Make sure it includes:

- Headline
- Statement of why this is important
- Detailed description of the finding
- Visualization explaining the finding

5.6 Apply in new context (15ψ)

5.6.1

Repeat problem 5.2 but by populating a RBD with the necessary tables and using a SQL query.

5.6.2 Short answer

1. Describe the difference between working in pandas and SQL.
2. Which do you think is better and why? (for this problem)
3. What would it take for that answer to change? (how would the problem have to change)

5.7 Review: Definitions (20ψ)

Provide definitions, or describe the distinction, for the following words:

1. Metadata
2. Data Dictionary
3. Established Data
4. FAIR vs. CARE
5. Context
6. Provenance vs. DMP
7. License
8. Internal/Embedded vs External

5.8 Review: Multiple Choice (20ψ)

Instructions: Select the best answer for each question. Do not show work.

5.8.1 Which of the following best defines metadata?

(A) Data that describes other data

(B) The largest or most important values in a dataset.

(C) A backup copy of a dataset stored in a different location.

(D) The statistical summary (e.g., mean, median) of a numeric variable.

5.8.2 Structural metadata primarily describes:

(A) Who created the data and when.

(B) The quality or accuracy of the data.

(C) How the data is organized

(D) Who is allowed to access the data.

5.8.3 In data management, a schema is best described as:

(A) A graph or chart that visualizes the data.

(B) A list of recommended analyses for the dataset.

(C) The raw data values in the first few rows.

(D) The definition of the data's structure

5.8.4 A data dictionary is used to:

(A) Document what each field means

(B) Store the actual data values in a compressed form.

(C) Define security rules for who can edit the data.

(D) Automatically correct errors in the data.

5.8.5 Provenance in the context of metadata refers to:

(A) The cost of storing and processing the data.

(B) The origin and history of the data

(C) The number of rows and columns in a table.

(D) The visual layout of a report or dashboard.

5.8.6 Which of the following is typically considered administrative metadata?

(A) The main topic or subject of a document.

(B) The definitions of technical terms in the data.

(C) Creation date, file type, and access permissions.

(D) A summary statistic such as the average of a column.

5.8.7 Metadata is said to be "embedded" when it:

(A) Is stored in a separate catalog or database from the data.

(B) Is derived by software when the file is opened.

(C) Is written in a programming language rather than plain text.

(D) Is stored within the same file as the data

5.8.8 Why is metadata important for data reuse and sharing?

(A) It provides context so others can understand, interpret, and trust the data.

(B) It reduces the total size of the dataset.

(C) It guarantees that the data has no errors.

(D) It replaces the need for a data dictionary or schema.

5.8.9 In a CSV of patient claims, which of the following is metadata rather than data?

(A) A single value such as $150.00 in the "amount paid" column.

(B) A patient ID such as "abc-123."

(C) The date of a specific claim.

(D) The column header "amount_paid" and its meaning (e.g., "total paid by insurer in USD").

5.8.10 This homework set is:

(A) Too short and Too easy

(B) Too short and Too hard

(C) Too long and Too easy

(D) Too long and Too hard