CPSC 447, Foundations 1

Out: Thu Sep 14, 2023. Due: Wed Sep 20, 2023, 6pm.
Submit to Gradescope, as a file PDF format following the answer template

Administrative overhead penalty of up to 10% will be imposed if additional processing work is required to handle your assignment, such as: did not follow answer template, unreadable file loaded into Gradescope, pages and questions not matched in Gradescope, etc.
Your assignment will not be marked until mandatory question 1 is completed on Piazza.

Learning goal

This assignment has two learning goals: to get practice with analyzing dataset and attribute types, and to grapple with the complexities and ambiguities of real-world datasets.There can be multiple legitimate answers: in situations where you think there's ambiguity, make a judgment call and defend it by providing your rationale.

1 Self-introduction (0%, Mandatory)

Introduce yourself on Piazza, in thread https://piazza.com/class/lm6wxwo7yn91ly/post/7

2 Data Abstraction: Attributes (5%)

What type of attribute (categorical, ordinal, quantitative) are the following?

1. 100 meter race times
2. College major
3. Amazon rating for a product
4. Product name
5. Weight of your favorite cat
6. Personality of your favorite cat
7. Rank of your favorite cat in the local feline dominance hierarchy
8. Length of your favorite cat's longest whisker
9. Size of your favorite sweatpants
10. Color of your favorite sweatpants

Rubric: .5% each
Exercise Credit: adapted from Lex & Meyer, Utah

3 Data Abstraction: Central Park Squirrel Census (30%)

2018 Squirrel Census Data

https://data.cityofnewyork.us/Environment/2018-Central-Park-Squirrel-Census-Squirrel-Data/vfnx-vebw
For each field, write down the attribute type (categorical, ordinal, quantitative). Your answer should match the order of fields/columns within the dataset. Determine its cardinality (number of levels) for categorical or ordinal, or its range for quantitative.

The answer template already has field names in each row; do not change the order. To make your solution more concise, we have "collapsed" together the fields that have identical attribute type *and* categorical values (e.g. all the binary attributes like *eating*, *foraging*); they are in the answer template together already. Also, you do not have to separately characterize the fully redundant field of combined lat/long (since you will have specified x/latitude and y/longitude already).

The cardinality/range should be based on the actual values within the specific dataset, not the theoretical possible values. For each field, if there is a discrepancy between the dataset itself and the data dictionary, determine the cardinality/range from the dataset itself and explicitly describe the discrepancy; otherwise, just say the field is an exact match (between the data dictionary and the actual dataset). Possible sources of discrepancies include potential null or empty values, spaces, and so on. Consider whether such values are semantically important and should be counted, or are unimportant and should be ignored, and include your rationale within the discrepancy discussion.

Rubric: 18 fields. For each .5% type (9%), .5% cardinality/range (9%), .66% discrepancies (12%)

4 Data & Task Abstraction: Foreign Aid (65%)

Foreign aid dataset

For this dataset, here are the field descriptions in domain language:

DATA USER'S GUIDE

FIELD DESCRIPTIONS

Records in the AidData main table and research releases contain over 100 fields, but many of these are tailored for specialized project reports or research questions. For a complete list of fields and descriptions, please see the AidData User's Guide. Most users will find the most useful information in the following fields:

Donor name: Name of the donor country or multilateral organization. For more detailed information, use Implementing Agency and/or Financing Agency.

Recipient name: Name of the recipient country or region. In some cases, Private Recipient, Beneficiary or Borrower may contain relevant information.

Year: Commitment year. Other date fields may also contain useful values, but Year is always populated.

Commitment Amount: Amount the donor has agreed to provide for the duration of the project, often disbursed over the following years. Note that there are actually several commitment amount fields:

Nominal/Current: As reported by the donor, in the reported currency. Current (USD): As reported by the donor, converted to nominal USD at the average exchange rate in effect in the commitment year. Constant (USD): The reported amount converted to USD and adjusted for inflation and exchange rate changes. Constant amounts are all presented in USD2009 (i.e. at 2009 prices and exchange rates). See the AidData User's Guide for conversion and deflation methods.

Title, Short Description, Long Description: These fields contain descriptive information as provided by the donor. Long descriptions range from only a few sentences to several paragraphs in length.

Purpose Code: AidData has developed a granular system of sector coding, which expands the OECD's purpose code scheme. However, coding is still underway. AidData researchers have coded projects from non-CRS sources and work is underway to add these codes to CRS-sourced data as well, but new codes have not yet been released for CRS projects. Therefore, CRS purpose codes for CRS-sourced records and AidData activity codes for non-CRS records should be used complementarily. See the AidData User's Guide for a full description of AidData's codes and how to use them.

4.1: Overall

- What is the dataset type(s)?
- How many fields/attributes does it have?
- How many items are there?

4.2: Analyze for each of the 7 fields
(AidData ID, Donor, Recipient, Year, Commitment Amount, Purpose Code, Purpose Name / Title Description)

- Determine its cardinality/range. For categorical or ordinal attributes, write down the number of unique levels. For quantitative attributes, specify range from min to max. If what you see in the actual data conflicts with the description of the data, use the actual data values to answer the question, and note this divergence explicitly as part of your answer the next question.

- What anomalies or discrepancies does this characterization reveal that appear to be dataset quality problems?

4.3: Write a question you would like to answer with this data set, from the point of view of an aid worker reporting to the government of a country providing aid. The question should benefit from a

visual chart to understand the answer. The question should be answerable using only the provided dataset.
a) Question
b) Which fields/attributes do you need to use to answer the question?

4.4: Write a question you would like to answer with this data set, from the point of view of a policymaker within the government of a country providing aid. The question should benefit from a visual chart to understand the answer. The question should require additional transformations of the data that can be derived from what is provided in the dataset to be answered.
a) Question
b) Which fields/attributes from the original dataset do you need to use to answer the question?
c) What additional data fields do you need that are not provided in the dataset itself? What kind of transformations or derivations are needed to generate those additional fields?

4.5: Write a question from the point of view of a journalist reporting to the citizens of a country receiving aid. The question should benefit from a visual chart to understand the answer. The question should require additional data beyond what can be derived from the provided dataset.
a) Question
b) Which fields/attributes from the original dataset do you need to use to answer the question?
c) What additional datasets (or dataset fields) do you need that are not provided and cannot be derived from the provided dataset?
d) Do you think it is plausible to get this data? How would you find or collect it?

Rubric:
4.1: 1% each, 3%
4.2: 7 fields, .5% each type, 1% each cardinality/range (10.5%). 5.5% anomalies. 16%
4.3: a) question, 5%. b) attributes, 5%. 10%
4.4: a) question, 5%. b) attributes, 5%. c) derive, 6%. 16%
4.5: a) question, 5%. b) attributes, 5%. c) additional 6%, d) find, 4%. 20%

Exercise Credit: adapted from Bertini, NYU Tandon