CMPU 101

DCIC Chapter 8
Processing Tables

Cleaning Data

Task Plans

Preparing Data

Managing and Naming Data

Visualizations and Plots

8.1 Cleaning Data Tables (review)

Last time

- Loading Data Tables

  From a Google Sheet (using **load-table:**)

- Dealing with Missing Entries

  Use sanitizers: num-sanitizer, string-sanitizer, …

- Normalizing Data

  Within a column, consistent values: transform-column(…)

- Normalization, Systematically

  Prefer drop-downs over text-entry boxes on forms

  Using Programs to Detect Data Errors (that need to be fixed)

8.2 Task Plans

Strategy: Creating a Task Plan

1. Develop a concrete example (of desired output)

    a table with 4-6 rows usually

2. Identify functions useful to transform data

    functions you already know, or look up in the documentation

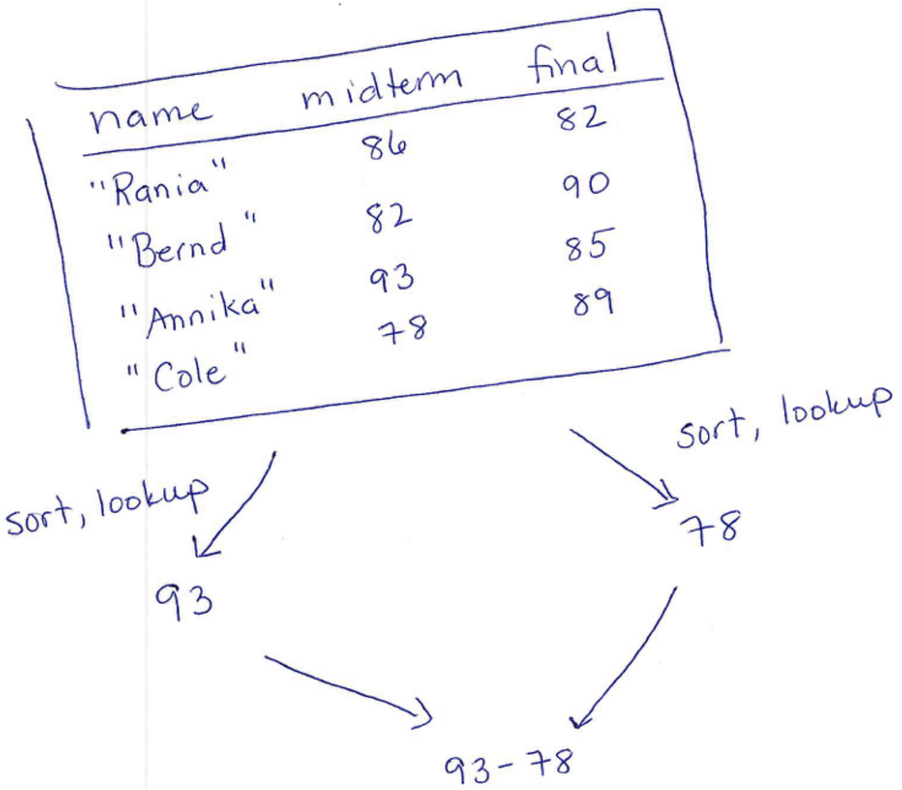
3. Develop a sequence of steps to transform data

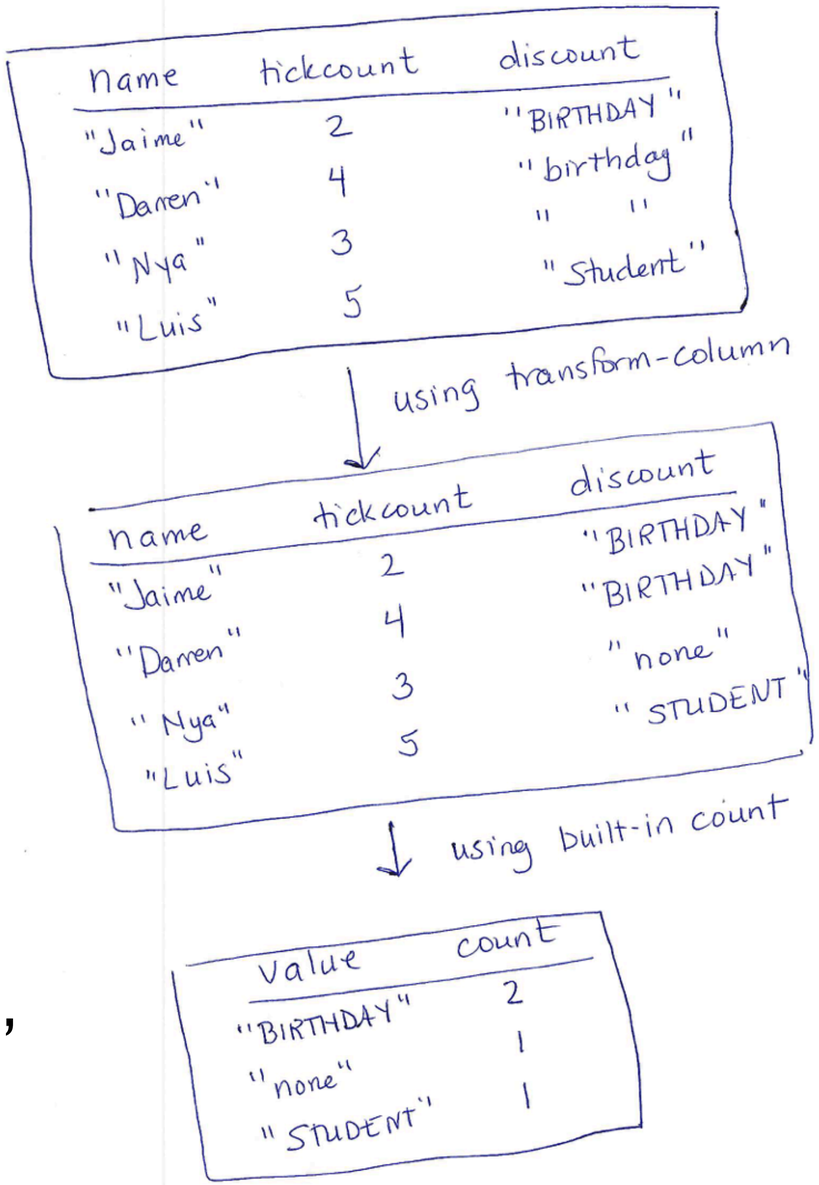
    draw as pictures, use textual descriptions, or a combination of the two

    use functions from previous step

4. Repeat step 3 to further break down steps until you can write expressions/functions to perform each step

Examples of Diagrams





If you aren't sure how to approach a problem, don't start by trying to write code. Plan until you understand the problem.

8.3 Preparing Data Tables

8.3.1 Creating Bins

Binning:

- the act of reducing one set of values into a smaller set of categories (e.g., actual tickcounts into small/medium/large)

- the bins are the categories

See the order-scale-label function in the CPO notes…

8.3.2 Splitting Columns

Use string functions to split strings:

- string-split( ) splits a string into two strings
- string-split-all( ) splits a string into two or more strings
- both functions return a *list* of strings
- we'll learn about lists soon…

```
››› string-split("Mathew Vassar", " ")
[list: "Mathew", "Vassar"]
››› string-split("Marc L. Smith", " ")
[list: "Marc", "L. Smith"]
››› string-split-all("Marc L. Smith", " ")
[list: "Marc", "L.", "Smith"]
›››
```

We can use these string-split( ) functions along with existing table functions to split a column into two or more columns—once we learn about lists…

Managing and Naming Data Tables

We have worked with several versions of the events table

- The original dataset that we tried to load
- The new sheet of the dataset with manual corrections
- The version with the discount codes normalized
- Another version that normalized the delivery mode
- The version extended with the order-scale column

Question: which tables should we name explicitly?

We have worked with several versions of the events table

- The original dataset that we tried to load
- The new sheet of the dataset with manual corrections
- The version with the discount codes normalized
- Another version that normalized the delivery mode
- The version extended with the order-scale column

Rule of thumb: we usually maintain separate names for the initially-loaded table, the cleaned table, and for significant variations for analysis purposes.

8.5 Visualizations and Plots

Ready to Analyze our Data

Now that we have cleaned and prepared the data in our table

Type of plot depends on type of data in column(s):

- quantitative: numeric values that can be ordered
- categorical: fixed set of values

Common Plots

- Scatterplots:

  shows relationships between two quantitative variables—one on each axis

- Frequency Bar charts:

  shows the frequency of each categorical value within a column

- Histograms:

  segment quantitative data into equal-size intervals, showing distribution of values across each interval

- Pie charts:

  show the proportion of cells in a column across the categorical values in a dataset

Wisdom (especially for large datasets)

Good data scientists never trust a dataset without first making sure that the values make sense.

Visualizations and plots can help data scientists identify data they might have missed that still needs to be cleaned/normalized.

8.6 Managing a Data Analysis

Steps for a Data Practitioner doing Data Analysis

1. Think about the data in each column

2. Check the data for errors

    via manual inspection, plots, and filter-with( ) expressions; normalize of correct

3. Store the normalized/cleaned data table

    either as a name in your program; or saving back out to a new file

4. Prepare the data based on the questions you want to ask

    compute new columns, bin existing columns, or combine data across tables

5. Perform your analysis, using statistical methods, visualizations, and interpretations that make sense for the questions and kinds of variables