Designing Programs for Tables

26 September 2022
Where are we?
We can represent complex data as tables. These can be encoded directly in a program or loaded from an external source. Real data may need clean-up, which can be manual or automatic.
Automatic data clean-up includes using *sanitizers*, which ensure all data in a column is of the desired type and can provide default values for empty cells.
We can modify table data later using `transform-column`, and we can remove (apparent) bad data using `filter-with`. 
We saw how this clean-up process works last week in lab by looking at the student data from the form (some of) you filled out.
Task plans
If you aren’t sure how to approach a problem, don’t start by trying to write code!

Plan until you understand the problem.
1 Develop a concrete example of desired output
   Typically a table with 4–6 rows

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
Example: Binning
What if we want to look at the distribution of responses to the question

Would you classify your academic focus as more humanities, more STEM (science, technology, engineering, and mathematics), or in the middle?

1  2  3  4  5  6  7  8  9  10
Super humanities  ○  ○  ○  ○  ○  ○  ○  ○  ○  ○  Super STEM
We don’t particularly care about how many students rated their STEM-iness as 2 or 8 or any particular number.

Instead, we might want to *bin* the responses into a few categories.
Let’s come up with a task plan to count the number of students in these three categories.
This needs a helper function! Call it `stem-category`.

```
<table>
<thead>
<tr>
<th>stem-level</th>
<th>Stem-Category</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>'non-stem'</td>
</tr>
<tr>
<td></td>
<td>'super-stem'</td>
</tr>
</tbody>
</table>
```

Use `build-column`:

```
<table>
<thead>
<tr>
<th>value</th>
<th>count</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;non-stem&quot;</td>
<td>3</td>
</tr>
<tr>
<td>&quot;stem&quot;</td>
<td>5</td>
</tr>
<tr>
<td>&quot;super-stem&quot;</td>
<td>3</td>
</tr>
</tbody>
</table>
```

Use `pie-chart`:

Or a bar-chart:
Task plan:

1. Write *stem-category*.
2. Add stem category to table using `build-column`.
3. Summarize results using `count`.
4. Visualize the results using `pie-chart / bar-chart`. 
test-table =
  table: stem-level
    row: 10
    row: 1
    row: 5
    row: 3
  end

fun stem-category(r :: Row) -> String:
  doc: "Return a stem category (non-stem, stem, or super-stem) for a given stem-level"
  ...
  where:
    stem-category(test-table.row-n(0)) is "super-stem"
    stem-category(test-table.row-n(1)) is "non-stem"
    stem-category(test-table.row-n(2)) is "stem"
  end
fun stem-category(r :: Row) -> String:
  doc: "Return a stem category (non-stem, stem, or super-stem) for a given
  stem-level"
  s = r["stem-level"]
  if stem-level > 7:
      "super-stem"
  else if stem-level > 4:
      "stem"
  else:
      "non-stem"
  end
where:
  ...
end
data-stem-category =
    build-column(student-data-cleaned,
        "stem-category", stem-category)

counts =
    count(data-stem-category, "stem-category")

bar-chart(counts, "value", "count")
What about testing the last part of this process – building the column and counting how many times values occur?

Better make this a function.
fun count-stem-categories(t :: Table) -> Table:
    doc: "Return a table of the counts for each stem category"
    t-with-cat = build-column(t, "stem-category",
                            stem-category)
    count(t-with-cat, "stem-category")
where:
    out-table =
        table: value, count
        row: "stem", 1
        row: "non-stem", 2
        row: "super-stem", 1
end
    count-stem-categories(test-table) is out-table
end
Nested functions
(Optional) Part 3: Going further

Congratulations! You have reached the end of lab. Here is an optional exercise in case you are looking for a challenge:

Task: Write a function `percent-true` that takes a table and column name as input and returns the percent of rows that are `true` for the column specified.

```haskell
fun percent-true(t :: Table, col :: String) -> Number:
  doc: "Return the percentage of rows that are true in column 'col'"
... end
```

What’s neat about this function is it will work on any table that has a column of type `Boolean`!

Task: Use this helper function to find the percentage of survey responders who are student-athletes. Check to see if it’s the same answer you got for Exercise 2.1.

Submitting the lab

- When you’ve completed the exercises, show your code to your instructor or one of the coaches.
fun percent-true(t :: Table, col :: String) -> Number:
  doc: "Return the percentage of rows that are true in column 'col"
  ...
end
fun percent-true(t :: Table, col :: String) -> Number:
  doc: "Return the percentage of rows that are true in column 'col"
  filter-with(t, ...).length() / t.length()
end
fun true-filter(r :: Row) -> Boolean:
    doc: "Return true if 'col' is true in this row"
    r[col]
end

fun percent-true(t :: Table, col :: String) -> Number:
    doc: "Return the percentage of rows that are true in column 'col'"
    filter-with(t, true-filter).length() / t.length()
end
fun \texttt{true-filter}(r :: Row) \rightarrow \texttt{Boolean}:
  doc: "Return true if 'col' is true in this row"
  \texttt{Not} r[col] == \texttt{true!}
end

fun \texttt{percent-true}(t :: Table, col :: String) \rightarrow \texttt{Number}:
  doc: "Return the percentage of rows that are true in column 'col'"
  filter-with(t, true-filter).length() / t.length()
end
fun true-filter(r :: Row) -> Boolean:
    doc: "Return true if 'col' is true in this row"
    r[col]
end

fun percent-true(t :: Table, col :: String) -> Number:
    doc: "Return the percentage of rows that are true in column 'col'"
    filter-with(t, true-filter).length() / t.length()
end

Why doesn’t this work?
col is undefined in true-filter.

Pyret only knows the value for col when you’re inside percent-true.

This means we need to define true-filter inside percent-true!
fun percent-true(t :: Table, col :: String) -> Number:
  doc: "Return the percentage of rows that are true in column 'col"

  fun true-filter(r :: Row) -> Boolean:
    r[col]
  end

  filter-with(t, true-filter).length() / t.length()
end
As usual, we should test our function using a simple test table:

```plaintext
test-table-student-athlete =
  table:
    student-athlete
    row: true
    row: false
end

fun percent-true(t :: Table, col :: String) -> Number:
  ...
where:
  percent-true(test-table-student-athlete, "student-athlete") is 0.5
end
```
The only time you *need* to use a nested function is if that function needs data that can’t be passed in directly to the function.
Introducing λ
fun `percent-true`(t :: Table, col :: String) -> Number:
  doc: "Return the percentage of rows that are true in column 'col'"

  fun `true-filter`(r :: Row) -> Boolean:
    r[col]
  end

  filter-with(t, true-filter).length() / t.length()
end
fun percent-true(t :: Table, col :: String) -> Number:
  doc: "Return the percentage of rows that are true in column 'col"

  fun true-filter(r :: Row) -> Boolean:
    r[col]
  end

  filter-with(t, true-filter).length() / t.length()
end

This is a really simple function, which we only use in one place. Instead of defining it like normal, we can write it inline where it's used.
fun percent-true(t :: Table, col :: String) -> Number:
  doc: "Return the percentage of rows that are true in column 'col''
  filter-with(t, lam(r): r[col] end).length() / t.length()
A *lambda expression* defines an anonymous function – a function that can be passed as an argument but doesn’t have an associated name.
Lambda expressions can be convenient for giving to higher-order functions `filter-with`, `build-column`, and `transform-column`.

We’ll use them more after Exam 1!
Preview: Lists
We've seen that when you want a row of a table, you use `.row-n` and get a Row.

What about getting a column?
<table>
<thead>
<tr>
<th>timestamp</th>
<th>house</th>
<th>stem-level</th>
<th>sleep-hours</th>
<th>schoolwork-hours</th>
<th>student-athlete</th>
<th>extracurricular-hours</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;9/21/2022 21:05:52&quot;</td>
<td>&quot;OTHER&quot;</td>
<td>10</td>
<td>6</td>
<td>8</td>
<td>false</td>
<td>0</td>
</tr>
<tr>
<td>&quot;9/21/2022 21:08:21&quot;</td>
<td>&quot;Strong House (1893)&quot;</td>
<td>10</td>
<td>7</td>
<td>8</td>
<td>false</td>
<td>2</td>
</tr>
<tr>
<td>&quot;9/21/2022 21:09:01&quot;</td>
<td>&quot;Lathrop House (1901)&quot;</td>
<td>6</td>
<td>7</td>
<td>4</td>
<td>true</td>
<td>4</td>
</tr>
<tr>
<td>&quot;9/21/2022 21:09:43&quot;</td>
<td>&quot;Lathrop House (1901)&quot;</td>
<td>9</td>
<td>8</td>
<td>4</td>
<td>false</td>
<td>0</td>
</tr>
<tr>
<td>&quot;9/21/2022 21:29:32&quot;</td>
<td>&quot;Lathrop House (1901)&quot;</td>
<td>8</td>
<td>6</td>
<td>7</td>
<td>false</td>
<td>2</td>
</tr>
<tr>
<td>&quot;9/21/2022 21:33:00&quot;</td>
<td>&quot;Cushing House (1927)&quot;</td>
<td>7</td>
<td>8</td>
<td>5</td>
<td>false</td>
<td>2</td>
</tr>
<tr>
<td>&quot;9/21/2022 21:38:40&quot;</td>
<td>&quot;Josselyn House (1912)&quot;</td>
<td>8</td>
<td>7</td>
<td>4</td>
<td>false</td>
<td>1.5</td>
</tr>
<tr>
<td>&quot;9/21/2022 21:41:36&quot;</td>
<td>&quot;Jewett House (1907)&quot;</td>
<td>9</td>
<td>6</td>
<td>3</td>
<td>true</td>
<td>2</td>
</tr>
<tr>
<td>&quot;9/21/2022 21:43:53&quot;</td>
<td>&quot;Main Building (1861)&quot;</td>
<td>8</td>
<td>7</td>
<td>4</td>
<td>true</td>
<td>3</td>
</tr>
<tr>
<td>&quot;9/21/2022 22:22:48&quot;</td>
<td>&quot;Davison House (1902)&quot;</td>
<td>8</td>
<td>7</td>
<td>9</td>
<td>false</td>
<td>4</td>
</tr>
</tbody>
</table>
>>> student-data-cleaned.get-column("house")
["OTHER", "Strong House (1893)", "Lathrop House (1901)", "Lathrop House (1901)", ...]
Lists can be very convenient!
fun normalize-house(house :: String) -> String:

doc: "Return one of the nine Vassar houses or 'OTHER'"

if (house == "Main Building (1861)" or
    house == "Strong House (1893)" or
    house == "Raymond House (1897)" or
    house == "Lathrop House (1901)" or
    house == "Davison House (1902)" or
    house == "Jewett House (1907)" or
    house == "Josselyn House (1912)" or
    house == "Cushing House (1927)" or
    house == "Noyes House (1958)":

    house

else:
    "OTHER"

end

where:

    normalize-house("Main Building (1861)") is "Main Building (1861)"
    normalize-house("Offcampus") is "OTHER"
houses = [list:
    "Main Building (1861)",
    "Strong House (1893)",
    "Raymond House (1897)",
    "Lathrop House (1901)",
    "Davison House (1902)",
    "Jewett House (1907)",
    "Josselyn House (1912)",
    "Cushing House (1927)",
    "Noyes House (1958)"
]

fun normalize-house(house :: String) -> String:
    doc: "Return one of the nine Vassar houses or 'Other'"
    if member(houses, house):
        house
    else:
        "OTHER"
end

where:
    normalize-house("Main") is "Main Building (1861)"
    normalize-house("Offcampus") is "OTHER"
Just like we did when we introduced tables, we’re separating our data from our computation!
Class code:
https://tinyurl.com/101-2022-09-26
Acknowledgments

This class incorporates material from:

Kathi Fisler, Brown University
Jason Waterman, Vassar College