STAT/CS 94 Fall 2015 Adhikari HW03, Due: 9/16/15

NAME: SID:

Please write your answers on the printed page, in the space provided. If you print this assignment yourself, please print it *double-sided* on a single sheet of paper. (There will be a small penalty for not following this instruction; it makes the grader's job more difficult.)

Problem 1 (Tables)

In this question we'll review basic operations on tables. Refer to Section 1.5 of the course text (`data8.org/text/1_data.html#tables`). As usual, please try to work out the answers on your own, then use a Jupyter notebook (on `ds8.berkeley.edu`) to check yourself.

Suppose we are analysts at a company that makes widgets at two factories (located in Berkeley and Palo Alto), and we're investigating the productivity of our factories. The factories have reported how many widgets they produced on several days. We have created a table named `production` from this data using the following Python code:

```
production = Table()
production['Factory'] = ['Berkeley', 'Palo Alto', 'Berkeley', 'Palo Alto', 'Berkeley', 'Palo Alto']
production['Date'] = [4, 5, 6, 4, 5, 6]
production['Widgets Produced'] = [112, 65, 130, 84, 91, 105]
```

(a) Using `sort`, give a Python expression whose value is the `production` table sorted in ascending order by number of widgets produced.

(b) What is the value of the following expressions? (For any array-valued expressions, it's okay to abbreviate by writing, for example, `[True, False]` instead of `array([True, False], dtype=bool`.)

(a) `len(production.rows)`

(b) `len(production.columns)`

(c) `len(production['Date'])`

(d) `len(production['Date'] >= 5)`

(e) `production['Date'] >= 5`

(f) `len(production['Date'] >= 7)`

(g) `len(production.where(np.array([True, False, False, False, True])).rows)`

(h) `len(production.where(production['Date'] >= 5).rows)`

(i) `len(production.where(production['Factory'] == 'Berkeley').rows)`

(c) Write a short description, in English, of the value of the following Python expression:

```
production.select(['Factory','Widgets Produced']).group('Factory', collect=np.sum)
```

(d) Give a Python expression whose value is the total number of widgets produced on date 5. (There are at least 2 reasonable ways to do this using the tools we've seen so far.)

(e) Now we would like to assign bonuses to the factory bosses according to the number of widgets produced at their factories. Say that we have a second table, `bosses`, defined as follows:

```
bosses = Table()
bosses['Factory'] = ['Berkeley', 'Palo Alto']
bosses['Name'] = ['Belinda', 'Patricia']
```

Write a short description, in English, of the value of the following Python expression. (Ignore the backslash; it's used to put single-line Python expressions on multiple lines.)

```
production.select(['Factory','Widgets Produced']).group('Factory', collect=np.sum) \
  .join('Factory', bosses, 'Factory').select('Name', 'Widgets Produced sum')
```

Problem 2 Histograms

This problem is about a graphic that appears in the Visualizations chapter (Chapter 2) of the textbook. It is the pair of figures, one blue and one green, presented by Statistics Canada. In this problem we will only focus on the blue one, as the green one can be handled in a similar manner.

The blue figure attempts to show the distribution of after-tax income in Canadian families in 2005. The data are from the Canadian census of 2006.

(a) Is the graph a histogram? Why or why not?

(b) Based on the figure, you can assume that the percent of families with incomes in the range $[100,000, 125,000)$ Canadian dollars is 6%. If you were to draw a histogram of the incomes using the density scale (total area = 1), the height of the bar over the $[100,000, 125,000)$ bin would be:

(i) $6/25,000$ per thousand dollars

(ii) $6/25,000$ per dollar

(iii) $0.06/25,000$ per thousand dollars

(iv) $0.06/25,000$ per dollar

(v) $6/25$ per thousand dollars

(vi) $6/25$ per dollar

(vii) $0.06/25$ per thousand dollars

(viii) $0.06/25$ per dollar

Pick all that are correct (there may be more than one), and explain your choices below.