S&DS 123 YData: An Introduction to Data Science
Spring 2019 Practice Midterm

INSTRUCTIONS

- You have 50 minutes to complete the exam.
- The exam is closed book, closed notes, closed computer, closed phone, and closed calculator, except one hand-written 8.5" × 11" crib sheet of your own creation and the official study guide provided with the exam.
- Mark your answers **on the exam itself**. We will *not* grade answers written on scratch paper.

First name	
Last name (Surname)	
NetID	

1. (12 points) Expressions

(a) (10 pt) An array of integers named `ca` contains the (estimated) population of California every 10 years. It has 11 items. The first item is the population of California in 1900. The last is the population in 2000.

```
array([ 1485053,  2377549,  3426861, ..., 23667902, 29760021, 33871648])
```

Write a Python expression below each of the following descriptions that computes its value. The first one is provided as an example. Do not include numbers above (e.g., 1485053) in your solutions.

- The population in 1900.

  `ca.item(0)`

- The population change from 1940 to 1960, expressed as a number of persons (not a proportion).

- Whether the population ever grew by less than 500000 in a decade represented by `ca`. (`True` or `False`)

- The *annual* (yearly) growth rate from 1920 to 1930.

- The population in 1924, assuming a fixed exponential annual growth rate from 1920 to 1930.
  You may use the name `g` for the annual growth rate from 1920 to 1930 (computed above).

- The number of items in `ca` that are at least twice as large as the population in 1960.

(b) (2 pt) You have 1000 different shirts in your huge closet, but only 2 of them are red. Each morning, you pick one uniformly at random, wear it, then put it back at night. Write a Python expression to compute the chance that, during a 30-day month, you *never* wear a red shirt.

2. (10 points) Tables

(a) (2 pt) The table named `twins` has a row for each pair of twins that contains the height of each twin in inches. The following code computes the average absolute difference in heights among the twins.

```
def diff(height1, height2):
    return abs(height1 - height2)
diffs = Table(['Absolute Differences'])
for i in np.arange(twins.num_rows):
    diffs.append([diff(twins.column(0).item(i), twins.column(1).item(i))])
np.mean(diffs.column(0))
```

Complete the expression below to compute the same result **without calling `diff` or using a `for` statement**.

```
np.mean(_____________________________________________________________________________________________)
```

(b) (8 pt) Each row of the `trip` table from lecture describes a single bicycle rental in the San Francisco area. Durations are integers representing times in seconds. The first three rows out of 338343 appear below.

Start	**End**	**Duration**
Ferry Building	SF Caltrain	765
San Antonio Shopping Center	Mountain View City Hall	1036
Post at Kearny	2nd at South Park	307

Write a Python expression below each of the following descriptions that computes its value. You *may* use up to two lines and introduce variables.

- The average duration of a rental that lasted more than 2 minutes.

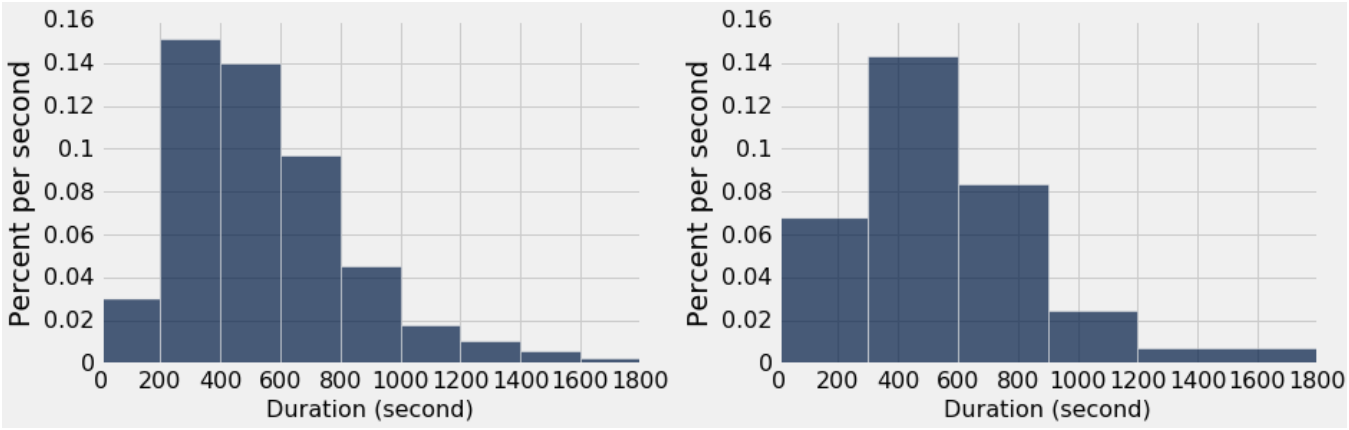
- The number of rentals that started at the `SF Caltrain` station.

- The name of the station where the most rentals ended (assume no ties).

- The number of stations for which the average duration ending at that station was at least 300 seconds.

3. (11 points) Distributions

The two histograms of bike trip durations below were both generated by `trip.hist(...)` using different bins.



(a) (8 pt) Write the proportion of trips that fall into each range of durations below. *Show your work.* If it is not possible to tell from the histograms, instead write **Not enough information**.

- Between 200 (inclusive) and 400 (exclusive) seconds

- Between 300 (inclusive) and 900 (exclusive) seconds

- Between 400 (inclusive) and 900 (exclusive) seconds

- Between 200 (inclusive) and 300 (exclusive) seconds

(b) (3 pt) A study followed 369 people with cardiovascular disease, randomly selected from hospital patients. A year later, those who owned a dog were four times more likely to be alive than those who didn't.

- Circle *True* or *False*: This study is a randomized controlled experiment.

- Circle *True* or *False*: This study shows that dog owners live longer than cat owners on average.

- Circle *True* or *False*: This study shows that for someone with cardiovascular disease, adopting a dog will probably cause them to live longer.