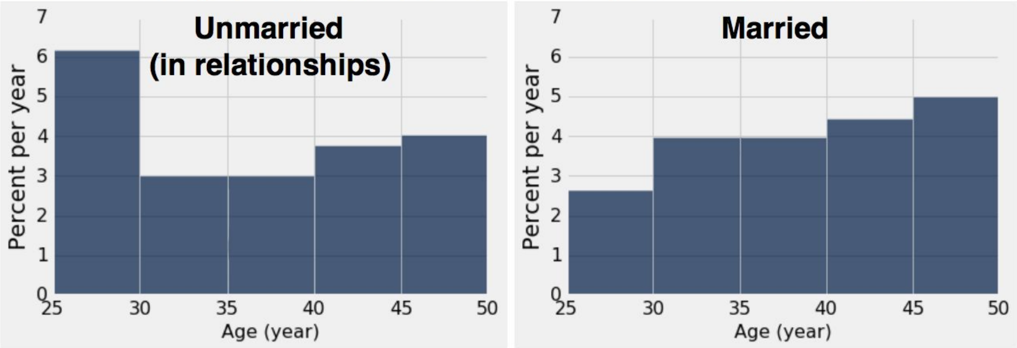
2. **(17 points) Distributions**

500 women age 25 to 49 in steady relationships were surveyed. Each woman was asked her age in years and whether she was married to her partner. There were **400 unmarried** and **100 married** women among those surveyed. The histograms below visualize the ages of these two groups of women.

```
unmarried.hist('Age',bins=[25,30,40,45,50])        married.hist('Age',bins=[25,30,40,45,50])
```



(a) **(10 pt)** For each pair of quantities, compare them using the information above and choose one of following:
(A): (I) is larger.
(B): (II) is larger.
(C): (I) and (II) are about the same.
(D): There is not enough information to compare (I) and (II).
*** **You must briefly justify your answer to receive full credit.** ***

- (I) The **number** of unmarried women age 25-29 vs (II) The **number** of unmarried women age 30-39

- (I) Among the unmarried women, the **proportion** who are of age 25-29 vs
(II) Among the married women, the **proportion** who are of age 45-49

- (I) The **number** of unmarried women age 30-39 vs (II) The **number** of married women

- (I) The **proportion** of married women age 30-34 vs (II) The **proportion** of married women age 35-39

- (I) The 20th percentile age of unmarried women vs (II) The 20th percentile age of married women

(b) **(3 pt)** What proportion of everyone surveyed were in the age range 30-39?

Note: You can ignore the `bins=[...]` argument to `hist()` in the question above. Just look at the actual histograms.

1. (16 points) Expressions

A table named `pay` contains one row for each UC Berkeley faculty member and these columns:

- **dept**: a string, the department of the faculty member.
- **name**: a string, the first name of the faculty member.
- **role**: a string, one of: Assistant Professor, Associate Professor, Professor, or Lecturer
- **salary**: an int, last year's salary paid by the university.

dept	name	role	salary
Journalism	Jeremy	Lecturer	111,528
Economics	Christina	Professor	349,727
South & Southeast Asian Studies	Penelope	Associate Professor	127,119

```
... (2056 rows omitted)
```

Fill in the blanks of the Python expressions to compute the described values. You must use *all* and *only* the lines provided. The last (or only) line of each answer should evaluate to the value described. Assume that the statements `from datascience import *` and `import numpy as np` have been executed.

(a) (2 pt) The total salary amount paid to all faculty.

```
________________(pay.____________________(_______________________________________________))
```

(b) (3 pt) The name of the third highest paid faculty member. (Assume no two faculty have the same salary.)

```
pay.__________(_________ , ______________________________).column(_________).item(_________)
```

(c) (3 pt) The number of lecturers in the department that has the most lecturers. (One has more than the rest.)

```
max(pay.__________(_________ , ____________________).__________(____________).column(’count’))
```

(d) (3 pt) The average faculty salary after all faculty members get a 10% raise each year for three years.

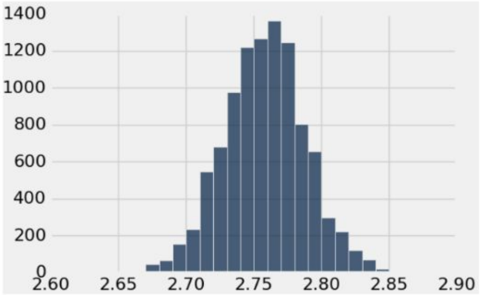
```
_____________________________________________________________________________________________
```

Note: In part (d), this means that salary is growing exponentially at a rate of 10% for three years.

3. (14 points) Sampling

This histogram shows sample means for 2,500 random samples. Each sample contains 10,000 trip distances, measured in miles, drawn at random from the distances of 1.4 million trips for New York taxis in January 2016.

```
sample_means.hist(bins=np.arange(2.65, 2.9, 0.01))
```



(a) (2 pt) What quantity is measured by the horizontal axis of this histogram?

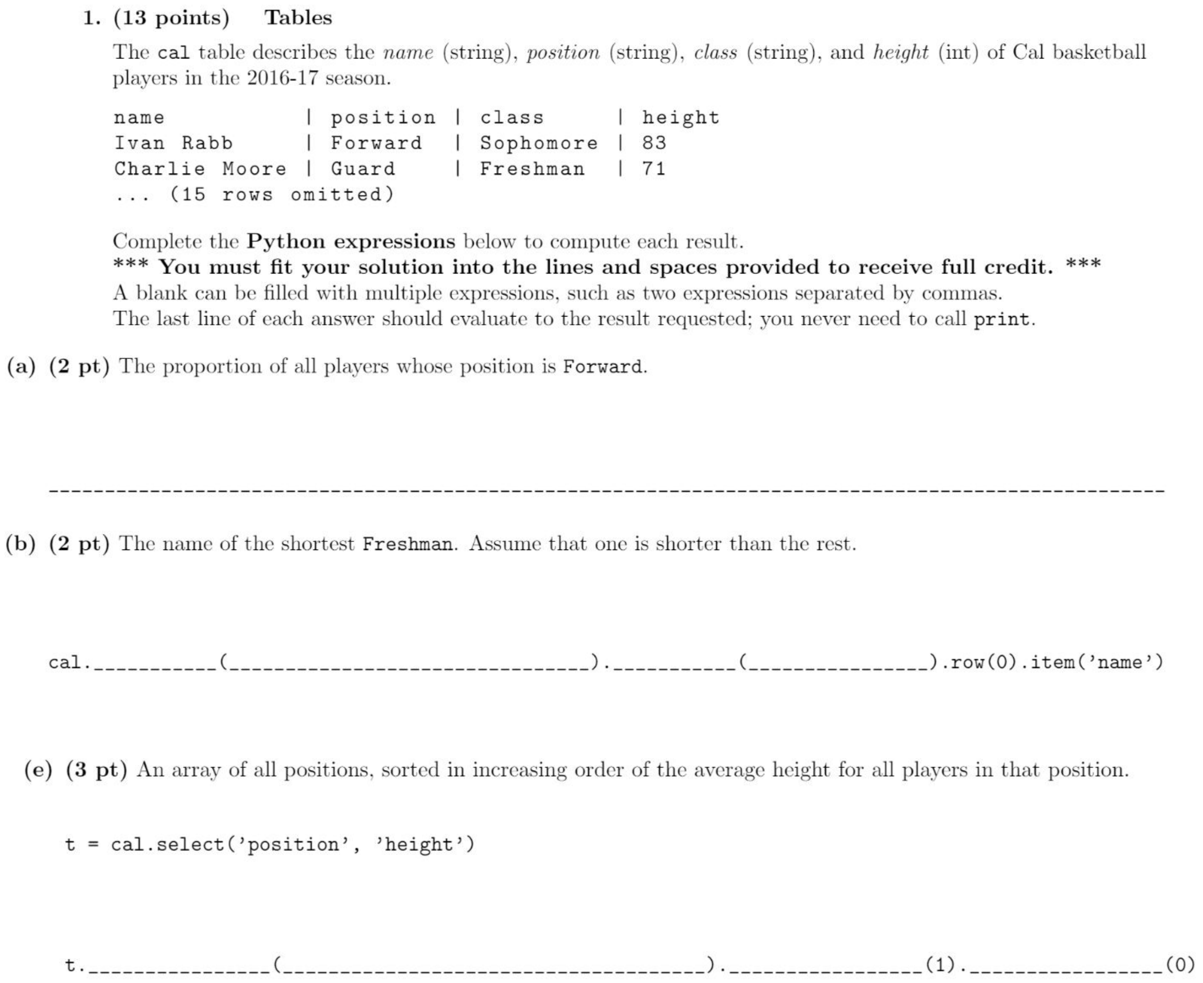
(A) Total miles for a single randomly chosen trip
(B) Total miles for a single randomly chosen sample
(C) Average miles for 10,000 randomly chosen trips
(D) Average miles for 2,500 randomly chosen samples
(E) None of the above

(b) (2 pt) What quantity is measured by the vertical axis of this histogram?

(A) Percent of trips per sample
(B) Percent of trips per mile
(C) Percent of trips per sample mean
(D) Percent of sample means per trip
(E) Percent of sample means per mile

(c) (2 pt) The percent of sample means represented by the tallest bar (with height about 1400) is closest to:

(A) 1.4 percent (B) 7 percent (C) 14 percent (D) 28 percent (E) 70 percent



1. **(13 points)** **Tables**

The `cal` table describes the *name* (string), *position* (string), *class* (string), and *height* (int) of Cal basketball players in the 2016-17 season.

```
name          | position | class     | height
Ivan Rabb     | Forward  | Sophomore | 83
Charlie Moore | Guard    | Freshman  | 71
... (15 rows omitted)
```

Complete the **Python expressions** below to compute each result.
***** You must fit your solution into the lines and spaces provided to receive full credit. *****
A blank can be filled with multiple expressions, such as two expressions separated by commas.
The last line of each answer should evaluate to the result requested; you never need to call `print`.

(a) **(2 pt)** The proportion of all players whose position is `Forward`.

```
____________________________________________________________________________________________________
```

(b) **(2 pt)** The name of the shortest `Freshman`. Assume that one is shorter than the rest.

```
cal.___________(_________________________________).___________(________________).row(0).item('name')
```

(e) **(3 pt)** An array of all positions, sorted in increasing order of the average height for all players in that position.

```
t = cal.select('position', 'height')


t.________________(_____________________________________).________________(1)._________________(0)
```

Note: In part (b), `.row(0).item('name')` is equivalent to writing `.column('name').item(0)`. (We didn't learn about `.row` because it's not very useful.) So the whole subexpression on the left should be a table.

(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. The first one is provided for you. You *may* use up to two lines and introduce variables.

- The average duration of a rental.

```
total_duration = sum(trip.column(2))
total_duration / trip.num_rows
```

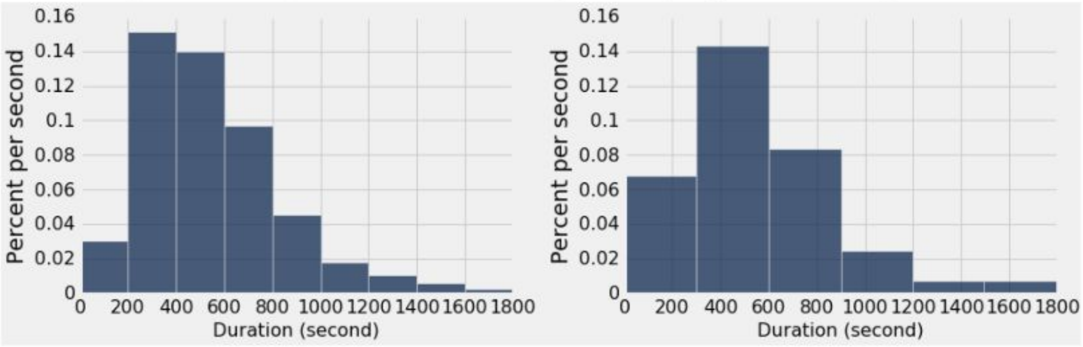
- The number of rentals that started at the `SF Caltrain` station.

- 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. (15 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

(c) (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.

1. **(16 points) Tables**

The `cafe` table (left) describes the Yelp reviews for three cafes on Euclid. Every cafe has a count for the number of 3-star, 4-star, and 5-star reviews, in that order. The `price` table (right) describes coffee prices.

```
name    | stars | count                name   | $
Nefeli  | 3     | 37                   Nefeli | 3
Nefeli  | 4     | 75                   Brewed | 3
Nefeli  | 5     | 50                   Abe    | 2
Brewed  | 3     | 56
Brewed  | 4     | 71
... (4 rows omitted)
```

Complete the **Python expressions** below to compute each result. For example, if the result prompt said, "The total number of reviews of all cafes," then you would write: `sum ( cafe.column( 2 ) )`

***** You must fit your solution into the lines and spaces provided to receive full credit. *****

The last line of each answer should evaluate to the result requested; you never need to call `print`.

(a) **(2 pt)** The total number of reviews of the cafe named `Nefeli`.

```
__________ ( cafe.____________________________________________________________________________________ )
```

(b) **(2 pt)** The total number of reviews of the cafe with the fewest reviews.

```
min ( ________________________________________________________________________________________________ )
```

(c) **(2 pt)** The average number of stars for reviews of the cafe named `Nefeli`.

```
n = cafe.where( _______________________________________________________________________________________ )

sum ( ________________________________________________ ) / sum ( ____________________________________ )
```

(e) **(2 pt)** An array containing the names of all cafes that have above-average coffee prices.

```
price.___________________________________________________________________________________________________
```

6. A clinic is trying to run a randomized controlled trial of a new appointment system. The clinic is open Monday through Saturday each week. As a way of randomization, all appointment requests that come in on Mondays, Wednesdays, and Fridays are assigned to the new system. All appointment requests that come in on Tuesdays, Thursdays, and Saturdays are assigned to the old system.

At the end of a few weeks, 212 appointment requests have been assigned to the new system and 277 to the old one.

Clinician A says, "This method is biased against the people who call on on Mondays, Wednesdays, and Fridays. If we had tossed a fair coin for each request, we wouldn't have ended up with so few requests assigned to the new system."

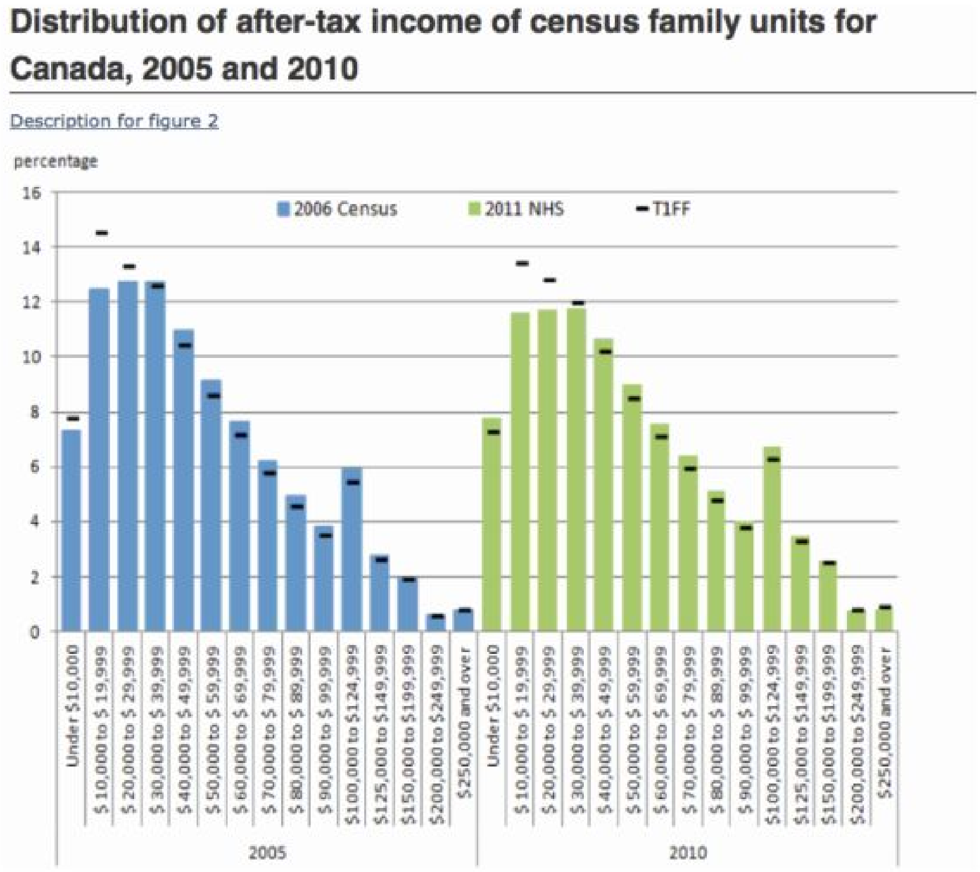
Clinician B says, "No, the results are just like tossing a fair coin."

(a) Provide null and alternative hypotheses that reflect the views of the two clinicians.

Null hypothesis:

Alternative hypothesis:

2. The figure below appears on the website of the Canadian National Household Survey. The graphs attempt to display the distribution of family income: the graph on the left shows the incomes in 2005 and the one on the right shows incomes in 2010.



Distribution of after-tax income of census family units for Canada, 2005 and 2010

Description for figure 2

In each of the two graphs, the eleventh bar from the left is unusually tall compared to the tenth bar. Explain why.