DATA 6 Fall 2025

Lisa Yan Quiz 2

Solutions last updated: Monday, October 13, 2025

Your name: _____

Your student ID: _____

Your Berkeley email: _____

Your room location: _____

Student ID of the person to your left: _____

Student ID of the person to your right: _____

You have 50 minutes. There are 3 questions of varying credit. (34 points total)

Question:	HC	1	2	3	Total
Points:	1	12	21	0	34

For questions with **circular bubbles**, you may select only one choice.

- ○ Unselected option (Completely unfilled)
- ◌ Don't do this (it will be graded as incorrect)
- ● Only one selected option (completely filled)

For questions with **square checkboxes**, you may select one or more choices.

- ■ You can select
- ■ multiple squares
- ☑ (Don't do this)

Anything written outside the answer boxes or ~~crossed out~~ will not be graded. If you write multiple answers, your answer is ambiguous, or the bubble/checkbox is not entirely filled in, we will grade the worst interpretation. For coding questions with blanks, you may write at most one statement per blank and you may not use more blanks than provided.

As a member of the UC Berkeley community, I act with honesty, integrity, and respect for others. I will follow the rules of this exam.

Honor Code (HC): I have read and agree to the honor code above.

(1 point) Sign your name: _____

Q1 *Yelp Business Reviews* **(12 points)**

Yelp is a website where users can review businesses on a 5-star rating scale.

name	stars	review_count	category
Bellevue Nails	3	47	Beauty & Spas
TLC Dental Center	4	10	Health & Medical
Bowmans Tavern	4	278	Restaurants
Enzo Custom	4	54	Shopping
Winter The Dolphins Beach Club	2	22	Hotels & Travel

Table 1: The first 5 rows of the `businesses` table (110,678 rows total).

Each row of `businesses` represents a business. Variable descriptions:

- `name`: Business name. There are over 80,000 unique business names.
- `stars`: Rating as an *integer*, 1 (lowest) to 5 (highest).
- `review_count`: Number of reviews for this business.
- `category`: Category of business: Restaurants, Shopping, Health & Medical, Beauty & Spas, Automotive, Hotels & Travel. Each business is in exactly one category.

Q1.1 (1 point) Which of the below would be best for visualizing how many businesses are in each category?

● Bar chart ○ Scatter plot ○ None of the above

○ Histogram ○ Line plot

Q1.2 (1 point) Which of the below would be best for visualizing the distribution of reviews per business?

○ Bar chart ○ Scatter plot ○ None of the above

● Histogram ○ Line plot

Q1.3 (1 point) Which of the below would be best for visualizing the names of businesses?

○ Bar chart ○ Scatter plot ● None of the above

○ Histogram ○ Line plot

Q1.4 (1 point) The review counts of the five Automotive businesses are defined as `automotive_counts = make_array(10, 30, 5, 40, 50)`.

The p-th percentile of an array is the smallest value that is at least as large as p% of the values. What is the 65th percentile of `automotive_counts`?

○ 10 ○ 30 ○ 50

○ 25 ● 40 ○ None of the above

SID: _____

Consider the `stars` table, which counts how many businesses received each star rating:

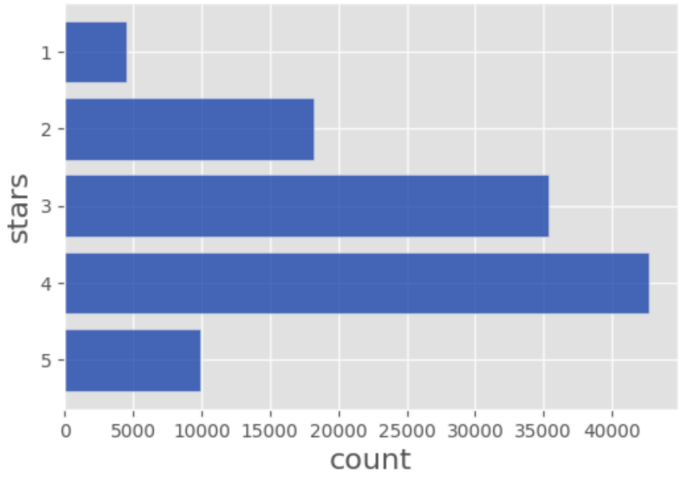
stars	count
1	4506
2	18197
3	35332
4	42689
5	9954

Table 2: The `stars` table.

Q1.5 (2 points) Which line of code correctly uses `businesses` to create `stars`? Select all that apply.

- ■ `stars = businesses.group("stars")`
- ☐ `stars = businesses.group("stars", np.sum)`
- ☐ `stars = businesses.group("stars", "count")`
- ☐ `stars = businesses.group("stars", len)`
- ○ None of the above

Q1.6 (3 points) Assume `stars` is correctly assigned above. Which line of code produces the following visualization? Select all that apply.



- ☐ `stars.barh()`
- ■ `stars.barh("stars")`
- ■ `stars.barh("stars", "count")`
- ☐ `businesses.barh("stars")`
- ☐ `businesses.hist("stars")`
- ○ None of the above

Q1.7 (3 points) Is the above visualization a histogram? Why or why not?

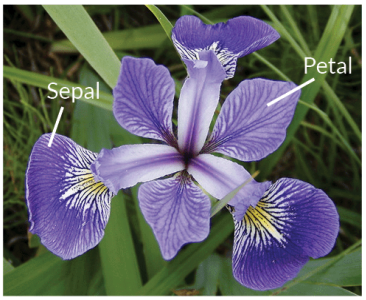
Solution: No; it is a bar chart, because stars are ordinal categorical variables. The spacing between stars on the y-axis doesn't mean anything.

SID: _____

Q2 *The Iris Dataset* **(21 points)**

The `iris` dataset measures flowers of three different species of the iris plant: *iris versicolor*, *iris setosa*, and *iris virginica*.

The petal and sepal are different parts of a flower. Each row in the `iris` table records one flower's species, its petal dimensions, and its sepal dimensions.



Iris Versicolor

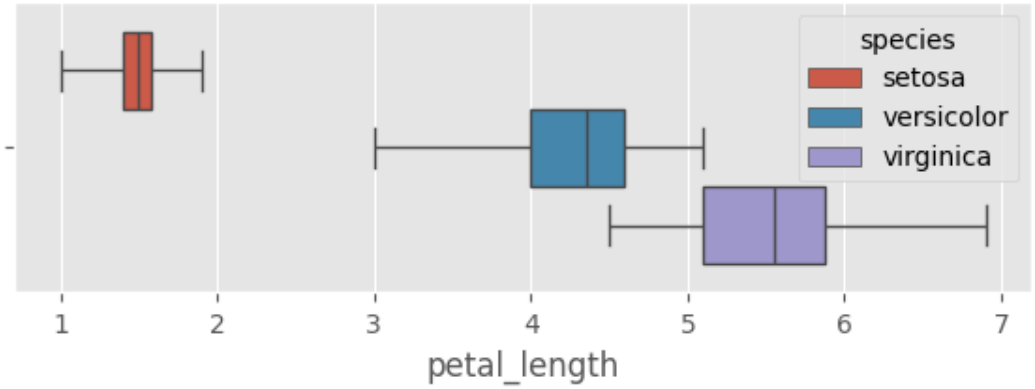
Iris Setosa

Iris Virginica

sepal_length	sepal_width	petal_length	petal_width	species
7.1	3	5.9	2.1	virginica
5.8	2.6	4	1.2	versicolor
5.8	4	1.2	0.2	setosa
4.9	2.5	4.5	1.7	virginica
6.1	2.8	4	1.3	versicolor

Table 3: The first 5 rows of the `iris` table (150 rows total). Measurements are in **centimeters (cm)**.

The `iris` table was used to generate the following boxplots of petal lengths (in cm) by species:



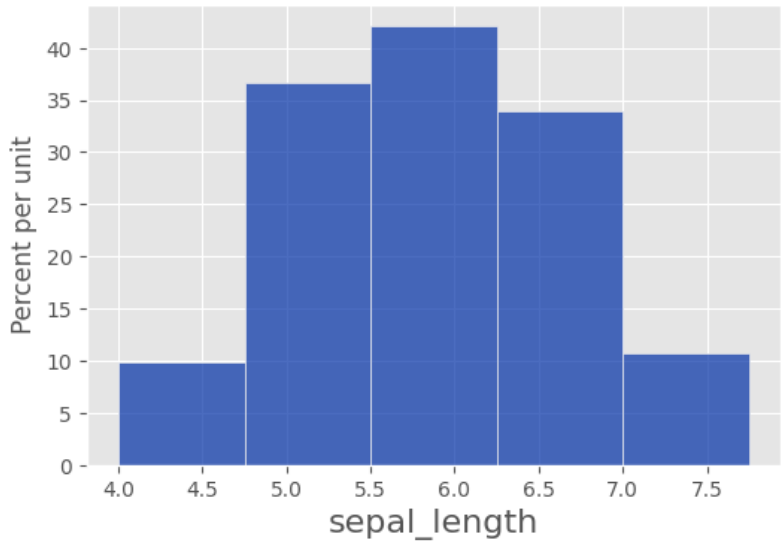
Q2.1 (4 points) Using the boxplots above, which of the following are true statements about the flowers in the `iris` table?

- ■ The bottom 50% of *versicolor* flowers have shorter petals than all *virginica* flowers.
- ☐ There are fewer *setosa* flowers than *virginica* flowers.
- ■ The distribution of *setosa* petal lengths is roughly symmetric.
- ☐ For *versicolor* flowers, the mean petal length is greater than the median petal length.
- ○ None of the above

Q2.2 - Q2.6 (5 points) Complete the code below so that supplying `my_bins` as the optional `bins` argument to the provided `hist` method produces the histogram shown. The histogram plots sepal lengths in the `iris` table with bins: [4, 4.75), [4.75, 5.5), [5.5, 6.25), [6.25, 7), and [7, 7.75].

```
my_bins = np.arange( 4 , 7.8 , .75 )
                     Q2.2   Q2.3   Q2.4

iris.hist("sepal_length", bins=my_bins)
 Q2.5              Q2.6
```



Q2.7 (1 point) Consider the y-axis in the histogram above. What is the "unit" in "Percent per unit"?

- ○ percent
- ○ sepal width
- ○ None of the above
- ○ flower
- ● centimeter

Q2.8 (1 point) Based on the histogram above, which bin contains the mean sepal length?

- ○ [4, 4.75)
- ● [5.5, 6.25)
- ○ [7, 7.75]
- ○ [4.75, 5.5)
- ○ [6.25, 7)
- ○ None of the above

Q2.9 (1 point) Based on the histogram above, which bin contains the median sepal length?

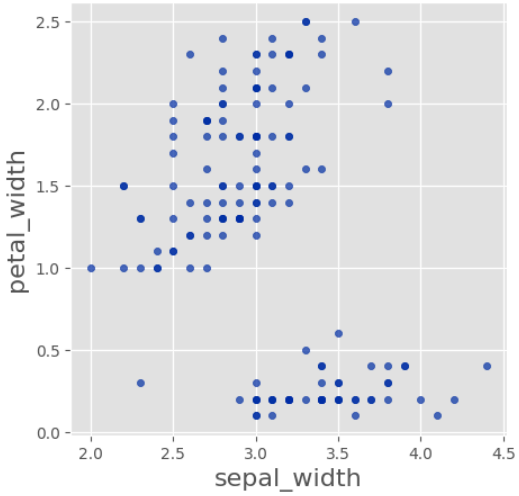
- ○ [4, 4.75)
- ● [5.5, 6.25)
- ○ [7, 7.75]
- ○ [4.75, 5.5)
- ○ [6.25, 7)
- ○ None of the above

Q2.10 - Q2.11 (2 points) Based on the histogram above, complete the code below to get all rows of the `iris` table with sepal lengths in the bin with the most flowers.

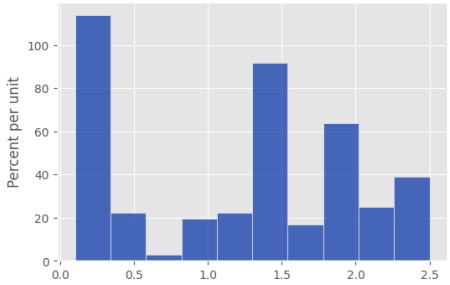
```
iris.where("sepal_length",are.between( 5.5 , 6.25 ))
                                       Q2.10   Q2.11
```

Q2.12 - Q2.14 (3 points) Complete the code below to produce the scatter plot shown, which plots petal width by sepal width for all flowers in the `iris` table.

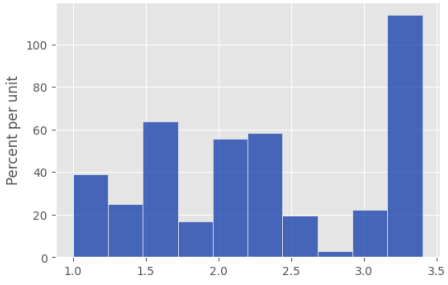
```
iris.scatter("sepal_width", "petal_width")
Q2.12              Q2.13          Q2.14
```



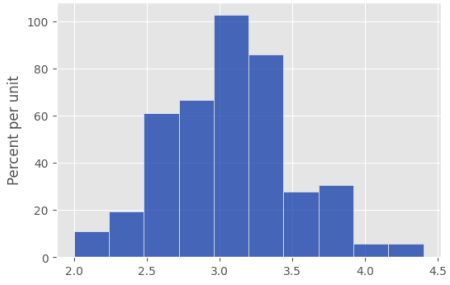
Still considering the scatter plot above, you are given three histograms. One corresponds to petal widths, one corresponds to sepal widths, and one corresponds to neither.



Histogram A



Histogram B



Histogram C

Q2.15 (2 points) If we run `iris.hist('petal_width')`, which histogram is produced?

- ● Histogram A
- ○ Histogram B
- ○ Histogram C
- ○ None of these

Q2.16 (2 points) If we run `iris.hist('sepal_width')`, which histogram is produced?

- ○ Histogram A
- ○ Histogram B
- ● Histogram C
- ○ None of these

SID: _____

Q3 *Just for fun!* **(0 points)**

Q3.1 Draw something fun, or write a message for the staff! Or leave this blank!

Everyone had cool drawings/messages! :D