Data 8

Fall 2020

Tutoring Section 4

Tables and Histograms

Logistics

- Autograder/OkPy is broken
    - Homework grades are **delayed**
- If you have any comments/feedback on my teaching:
    - Complete the anonymous feedback form
        - Form: https://tinyurl.com/feedbackD8Kevin
- All resources and links can be found on my website www.kevin-miao.com (You can find the link if you go to data8.org and find me under the **staff** tab)

Today

- Weekly Check-In
- Helicopter Review
    - Tables
        - Summary of methods
    - Histograms
- Worksheet

Tables

- Creating and extending tables:
  - **Table().with_column** and **Table.read_table**
- Finding the size: **num_rows** and **num_columns**
- Referring to columns: labels, relabeling, and indices
  - **labels** and **relabeled**; column indices start at 0
- Accessing data in a column
  - **column** takes a label or index and returns an array
- Using array methods to work with data in columns
  - **item**, **sum**, **min**, **max**, and so on
- Creating new tables containing some of the original columns:
- **select, drop**

Worksheet

Link: **https://tinyurl.com/d8tutweek4**

Q1.1-1.2

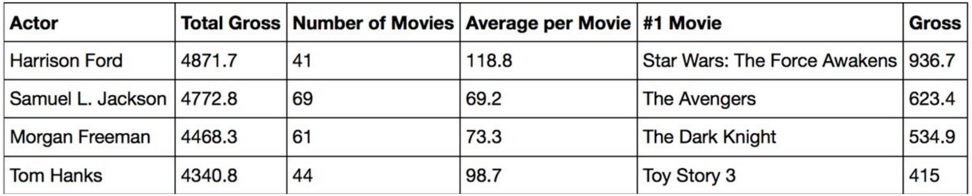
1.1 Write a line of code that returns `actors` sorted from highest to lowest number of movies.

```
actors.sort("number of movies", descending = True)
```

1.2 Now, write a line of code to find the actor who has made the most movies. Do not return a table with the actor's name; just return the name as a string.

```
actors.sort("number of movies", descending = True).column(
 "Actor").item(0)
```

For the first part, we're just going to focus on the `actors` table, which begins like this:



Actor	Total Gross	Number of Movies	Average per Movie	#1 Movie	Gross
Harrison Ford	4871.7	41	118.8	Star Wars: The Force Awakens	936.7
Samuel L. Jackson	4772.8	69	69.2	The Avengers	623.4
Morgan Freeman	4468.3	61	73.3	The Dark Knight	534.9
Tom Hanks	4340.8	44	98.7	Toy Story 3	415

Q1.3-1.4

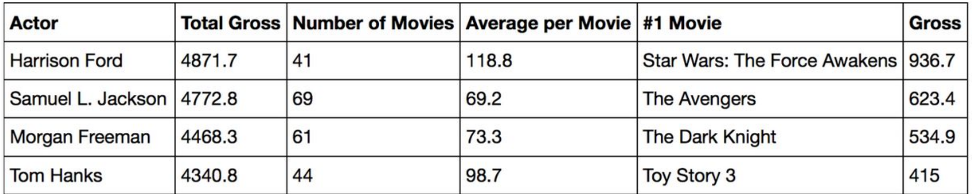
1.3 What is Tom Hanks' #1 movie? Write a line of code to return the name of the movie as a string.

actors. where ("Actor", are. equal_to ("Tom Hanks")). column ("#1 Movie").item (Ø)

1.4 Write a line of code which returns a table consisting of only the "Actor" column where the elements in the "Actor" column are the names of actors who have above 40 movies and have a total gross below 3000.

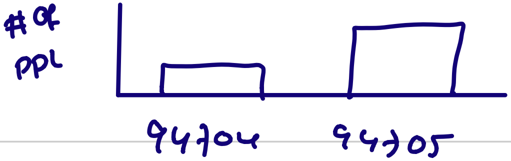
actors. where ("Number of movies", are. above (40)). where ("Total gross", are. below (3000)). select ("Actor")

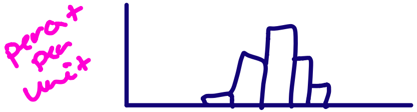
For the first part, we're just going to focus on the `actors` table, which begins like this:



Actor	Total Gross	Number of Movies	Average per Movie	#1 Movie	Gross
Harrison Ford	4871.7	41	118.8	Star Wars: The Force Awakens	936.7
Samuel L. Jackson	4772.8	69	69.2	The Avengers	623.4
Morgan Freeman	4468.3	61	73.3	The Dark Knight	534.9
Tom Hanks	4340.8	44	98.7	Toy Story 3	415

Histograms



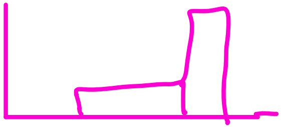


Height histogram

- **When to use a histogram vs a bar chart?**

Bar chart : Compare categorical || Histogram : Distribution of numerical Data !

- **Histograms**
  - **Areas** as **percentages**
  - **Height** as **densities**
  - The complete area under a histogram is always 1
  - Bins (can be arbitrary)
  - Formulas:

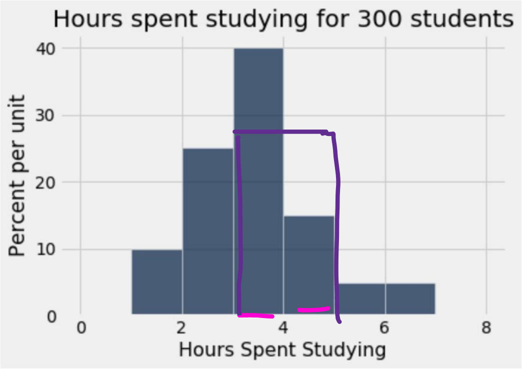


$$height = \frac{\%\ in\ a\ bin}{width\ of\ the\ bin}$$

$$area = \% \ = width\ of\ bin * height\ of\ bar$$

Q2

Suppose you are interested in the number of hours, on average, that UC Berkeley students spend studying a day. You survey 300 random UC Berkeley students, record the number of hours studying a day they reported, and plot a histogram with the data. The histogram is shown below.



2.1 What percentage of students studied between two and three hours a day?

25%

width = 1
Height = 25

2.2 How many students studied between three and four hours a day?

120 Students ||

width = 1 } → 40%) 120
Height = 40
Students = 300

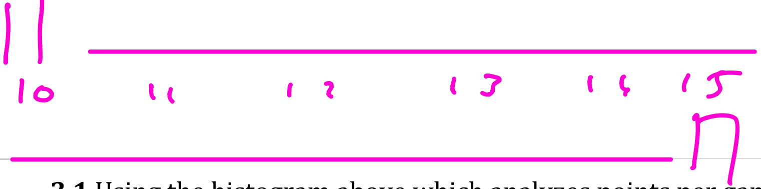
2.3 Suppose you created a new bin for students who studied between three and

5 hours a day. What would be the height of the new bar?

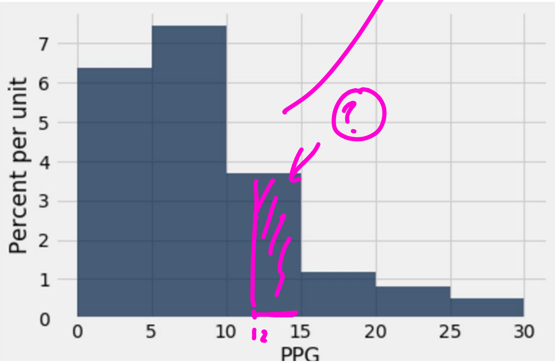
$$\frac{40*1 + 15*1}{2} = 27.5$$

Q3

?



average points per game



3.1 Using the histogram above which analyzes points per game, answer the following questions:

a. Is it possible to find the percentage of players that scored between 12 and 15 points per game? Why or why not? What piece of information could help us answer this question?

we don't know the distribution w/in the bin.

→ seperate bins

→ [10, 11] || [12, 15)

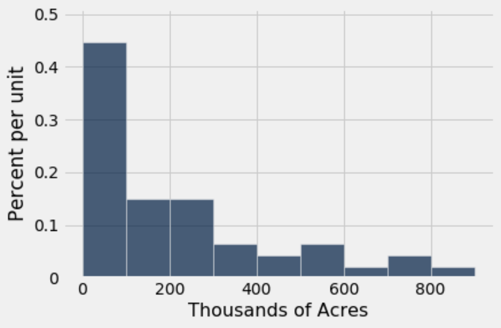
b. Can we find the total number of players who averaged 20 or more points per game? What piece of information could help us answer this question?

no, we don't have the total number of players, so we can't answer that!

Exam Prep

We gathered a data set of US national parks, and plotted below a histogram of the size of these parks (in thousands of acres). All bars are 100 wide. The area of the visible bars sum to 100%.

Calculate each quantity described below or write *Unknown* if there is not enough information above to express the quantity as a single number (not a range). It's OK to write your answer as a Python expression or an unsimplified expression. You may need to estimate the height of bars visually; if so, make your best estimate. Don't show your work.



(a) The percentage of parks that are less than 100 thousand acres in size.

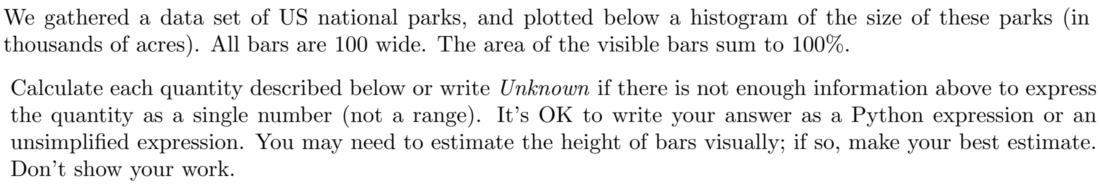
(b) The percentage of parks that are between 200 to 400 thousand acres in size (specifically, the range [200, 400), i.e., at least 200 thousand acres and less than 400 thousand acres).

(c) The percentage of parks that are less than 150 thousand acres in size.

(d) The percentage of parks that are at least 1200 thousand acres in size.

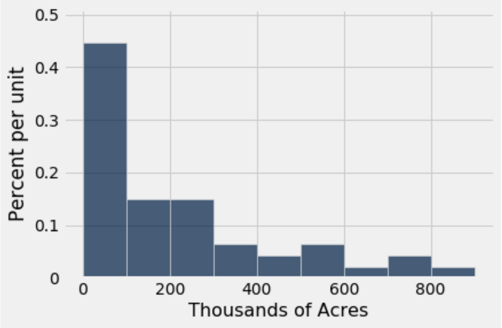
Fall 2018 – MT1

Exam Prep (Solutions)



We gathered a data set of US national parks, and plotted below a histogram of the size of these parks (in thousands of acres). All bars are 100 wide. The area of the visible bars sum to 100%.

Calculate each quantity described below or write *Unknown* if there is not enough information above to express the quantity as a single number (not a range). It's OK to write your answer as a Python expression or an unsimplified expression. You may need to estimate the height of bars visually; if so, make your best estimate. Don't show your work.



(a) The percentage of parks that are less than 100 thousand acres in size.

0.45 × 100 = 45%.

(b) The percentage of parks that are between 200 to 400 thousand acres in size (specifically, the range [200, 400), i.e., at least 200 thousand acres and less than 400 thousand acres).

(0.15 + 0.06) × 100 = 21%.

(c) The percentage of parks that are less than 150 thousand acres in size.

Unknown.

(d) The percentage of parks that are at least 1200 thousand acres in size.

0%.

Fall 2018 – MT1

End of Section

- Please complete the anonymous Feedback form so I can improve my teaching:
  - **https://tinyurl.com/feedbackD8Kevin**

- Solutions and notes will be posted next Monday!

- Stay safe in these smoky conditions.