Week 4: Tables and Histograms

Data 8 Tutoring

# 1 Tables

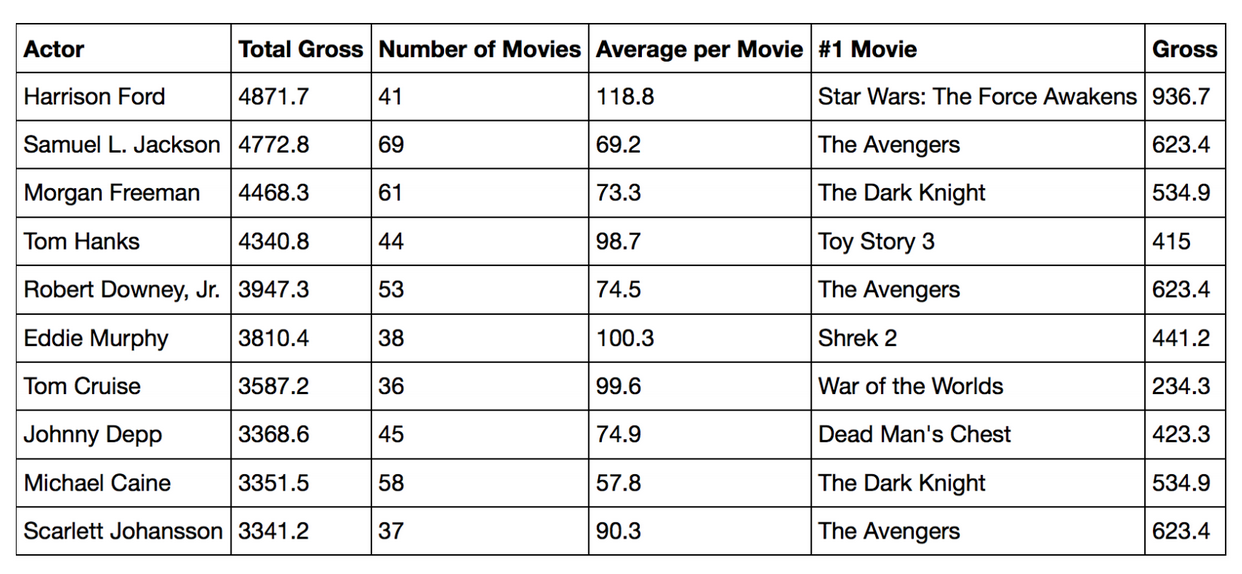
## Key Concepts

**Table Manipulation:** Table methods can be applied to clean or reshape a given table in order to conduct analysis. This is done by writing tbl.method() and passing in the necessary arguments. When using table methods, make sure to assign a variable to the result of the new Table method. *Examples*: where, select, sort, drop

## Practice Problems

Assume all imports are completed. **Pay close attention to all of the syntax, as it’s difficult to learn at first.** Feel free to ask your tutor any questions you have!

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

We will start with some simple queries on this table and move our way upwards to more advanced ones.

**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)

**1.3** What is Tom Hanks’ #1 movie? Write a line of code to find out.

actors.where("Actor", are.equal\_to("Tom Hanks")).column("#1   
Movie").item(0)

**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")

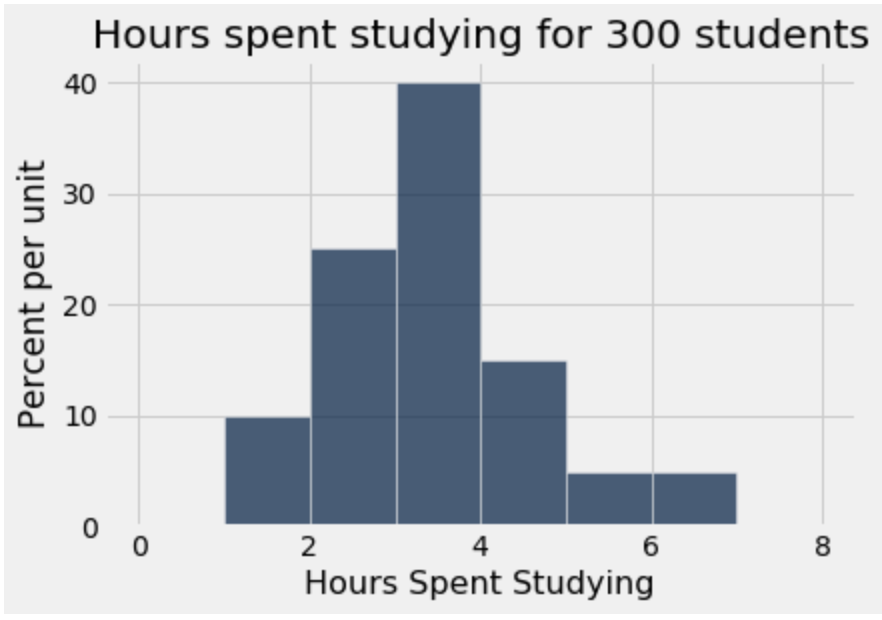
# 2 Histograms

## Key Concepts

Histograms are an important visualization for understanding the *distribution* of a single numerical variable. While they appear similar to bar charts, histograms have important differences which make them powerful visualizations for data science.

## Practice Problems

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?

(3-2)\*25 = 25 percent

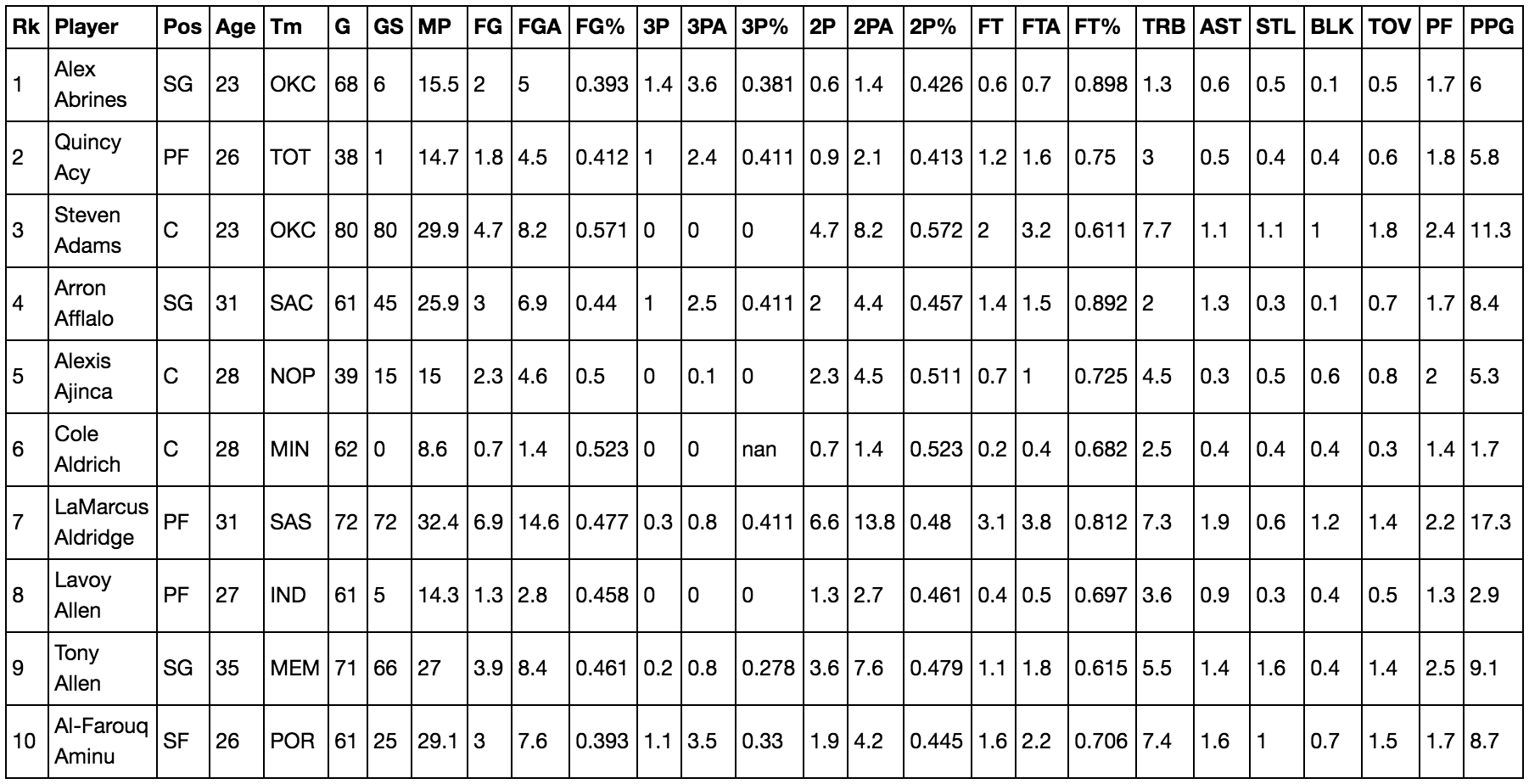
**2.2** How many students studied between three and four hours a day?

(4-3)\*40\*300/100 = 120 students

**2.3** Suppose you created a new bin for students who studied between three and five hours a day. What would be the height of the new bar?

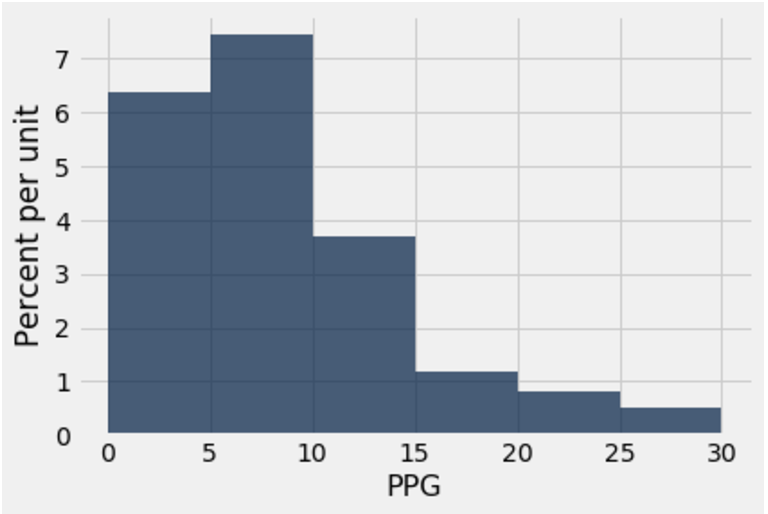
(40 + 15)/2 = 27.5

Throughout this section, we’ll be focusing on the following table: nba.csv. It describes the average statistics of NBA players for the 2016-2017 season. Pay special attention to the syntax! This can quickly become confusing, so ask your tutor if anything seems confusing.  
  
The first few rows of the nba table look like this:



If you don’t know what the acronyms stand for, don’t worry! We’ll be working with the “PPG” column, which stands for “Points per game”.

Assume all imports are correctly made.



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

1. 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?

Using only the histogram above, we cannot find the percentage of players that scored between 12 and 15 points per game. This is because the bins have regular widths of 5; we have no way to determine how many players are in the [12, 15) interval out of the [10, 15) bin using only the histogram above. If we had a bin for [12, 15) or several bins that composed it exactly, we could then find the percentage.

1. 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?

We would be able to find the percentage of players who averaged 20 or more points, but we wouldn’t be able to find the number of players who averaged 20 or more points. We would need the number of players that the histogram represents to help us answer the question.