YData: Introduction to Data Science

Lecture 09: Functions

Overview

Review and continuation of visualizing numeric data

- Binning and histograms

Writing functions

Applying functions to values in a column of a Table

If there is time: prediction example

Histograms

Histograms

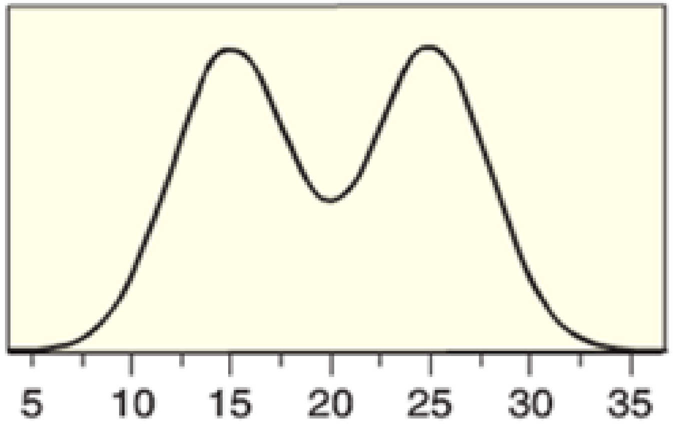
Histograms are a way to visualize which ranges of numerical values occur most frequently

- i.e., the give insight into how numerical data is ***distributed***

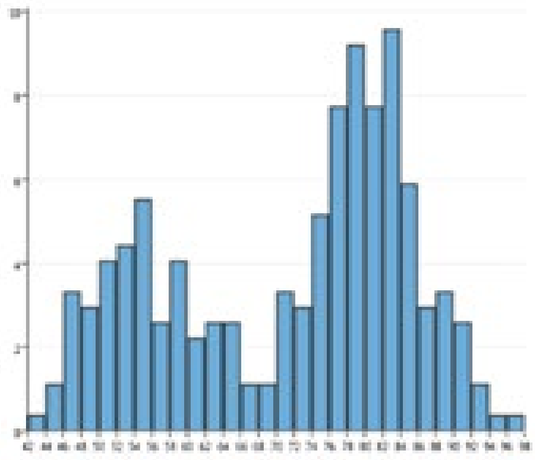
To create histograms we:

1. Create a sequence of binning intervals
2. Count the number of data points that fall into each interval
3. Plot these counts as a bar chart

Underlying distribution



Histogram



Histograms of country life expectancy in 2007

Suppose we had the average life expectancy for 142 countries in the world:

- 43.83,   72.30,   76.42,   42.73, …

To create a histogram, we create a set of intervals

- [35-40),   [40-45),   [45-50),   …   [75-80),   [80-85)

We count the number of points that fall in each interval

We create a bar chart with the counts in each bin

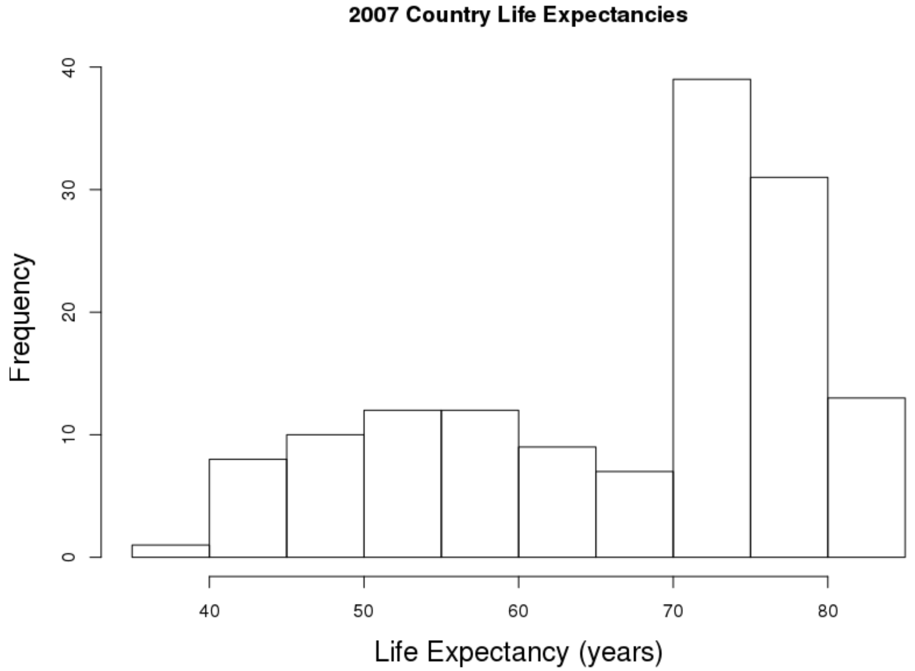
Histograms – countries life expectancy in 2007

Life Expectancy	Frequency Count
[35 – 40)	1
[40 – 45)	8
[45 – 50)	10
[50 – 55)	12
[55 – 60)	12
[60 – 65)	9
[65 – 70)	7
[70 – 75)	39
[75 – 80)	31
[80 – 85)	13

```
my_bins = np.arange(35, 86, 5)

tb.bin("numeric col", bins = my_bins)

tb.hist("numeric col"), bins = my_bins)
```

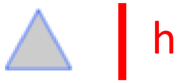


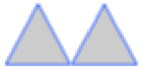
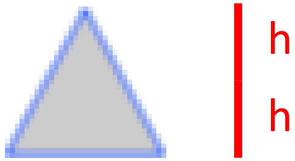
Area principal

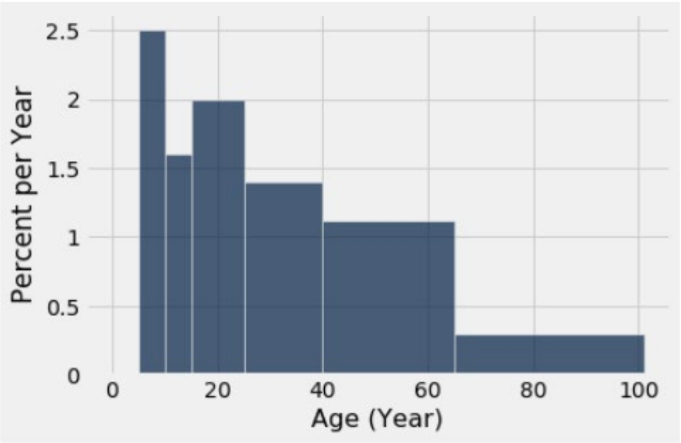
It is possible to create histograms with different sized bins

The **area** taken up by a bar in a histogram should be proportional to the **percentage** of the values represented

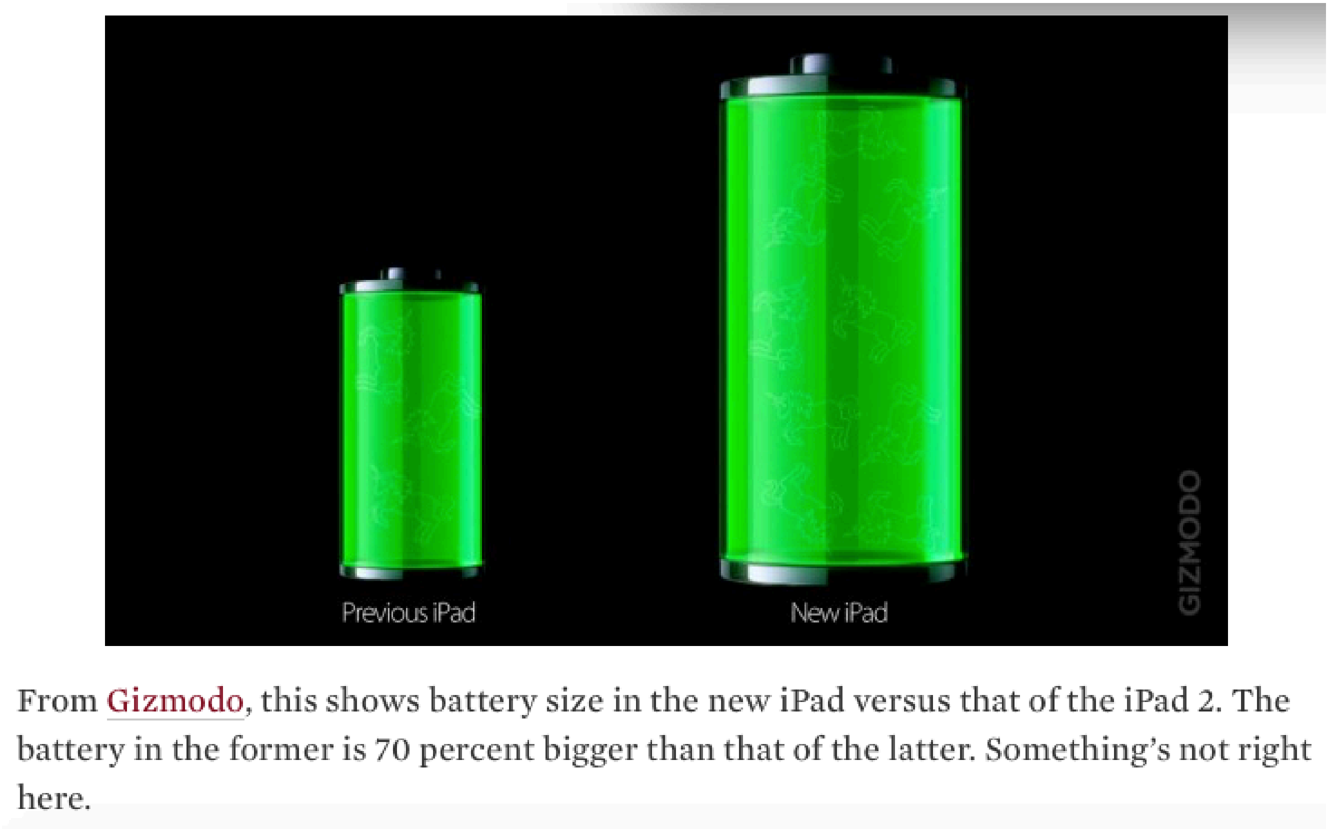
For example:

- If 20% of a population is represented by: 

- Then 40% should be represented by: 
- But not by: 



A Gizmodo article got this wrong once ☹



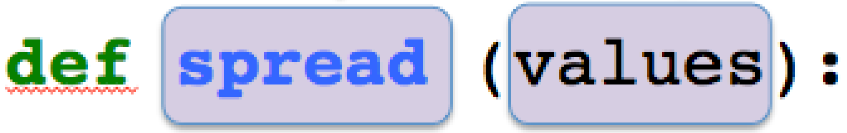
From Gizmodo, this shows battery size in the new iPad versus that of the iPad 2. The battery in the former is 70 percent bigger than that of the latter. Something's not right here.

Let's explore this in Jupyter!

Defining functions

Def statements

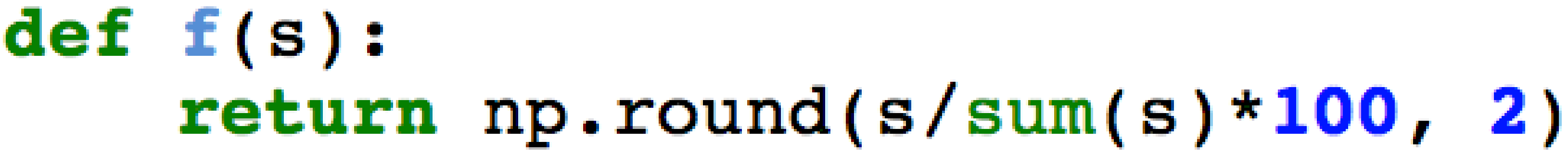
User-defined functions give names to blocks of code



```
def spread(values):
```

Let's explore this in Jupyter!

Discussion questions



```
def f(s):
    return np.round(s/sum(s)*100, 2)
```

1. What does this function do?
2. What kind of input does it take?
3. What output will it give?
4. What's a reasonable name?

Let's explore this in Jupyter!

Applying functions to columns in a Table

Apply method

The tb.apply() method creates an array by calling a function on every element in input column(s)

```
tb.apply(function_name, "numeric col")
```

- First argument: Function to apply
- Other arguments: The input column(s)

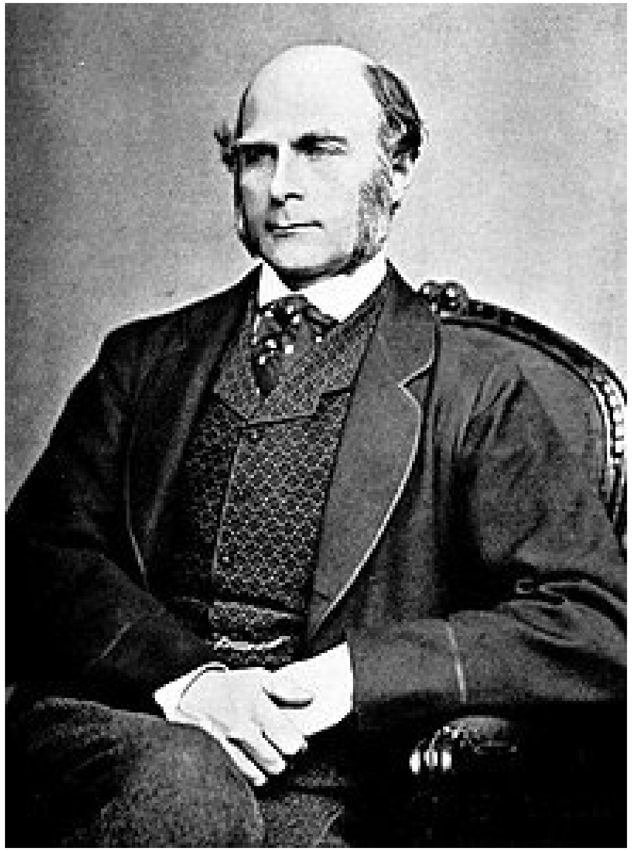
Let's explore this in Jupyter!

Example: Prediction

Francis Galton

- 1822 - 1911
  - Charles Darwin's half-cousin
- A pioneer in making predictions
- Particular (and troublesome) interest in heredity

One of his most famous results involved exploring the relationship between the heights of parents and their children



Let's explore this in Jupyter!