YData: Introduction to Data Science

Lecture 08: Bar plots and histograms

Overview

Review and continuation of Table manipulation and data visualization

Categorical and numerical data

- **Categorical data**: frequency tables and bar charts
- **Numerical data**: Histograms

Review and continuation of Table manipulation and data visualization

COVID-19 data exploration

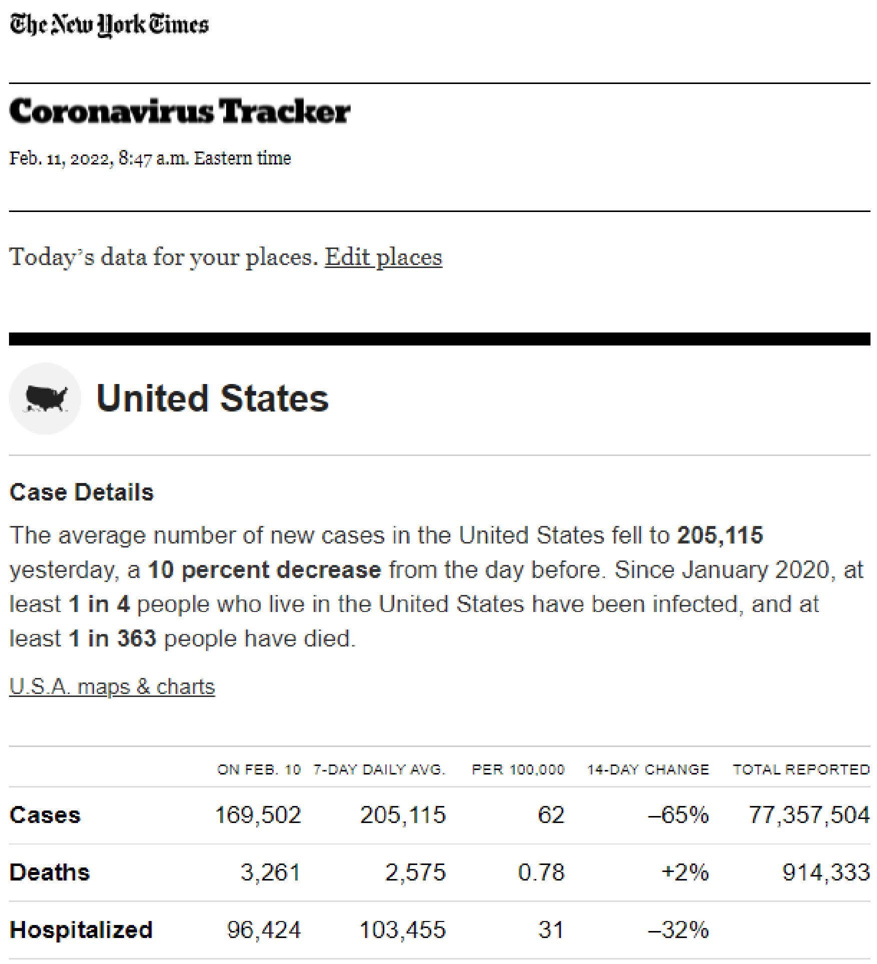
Let's examine COVID-19 data from Connecticut

Data is from the New York Times GitHub repository

Here is a site showing the number of cases at Yale:

- https://covid19.yale.edu/yale-data

Let's explore this in Jupyter!



The New York Times

Coronavirus Tracker

Feb. 11, 2022, 8:47 a.m. Eastern time

Today's data for your places. Edit places

United States

Case Details

The average number of new cases in the United States fell to **205,115** yesterday, a **10 percent decrease** from the day before. Since January 2020, at least **1 in 4** people who live in the United States have been infected, and at least **1 in 363** people have died.

U.S.A. maps & charts

	ON FEB. 10	7-DAY DAILY AVG.	PER 100,000	14-DAY CHANGE	TOTAL REPORTED
Cases	169,502	205,115	62	–65%	77,357,504
Deaths	3,261	2,575	0.78	+2%	914,333
Hospitalized	96,424	103,455	31	–32%	

Types of Data

Categorical and numeric data

Numerical: Each value is from a numerical scale

- Numerical measurements are ordered
- Differences are meaningful – can do math!

Categorical: Each value is from a distinct category

- Categories are the same or different
- May or may not have an ordering

Eats chocolate	Happiness score
Chocolate	10
No Chocolate	0
Chocolate	10
Chocolate	10

Numerical Data

Just because the values are numbers, doesn't mean the variable is numerical

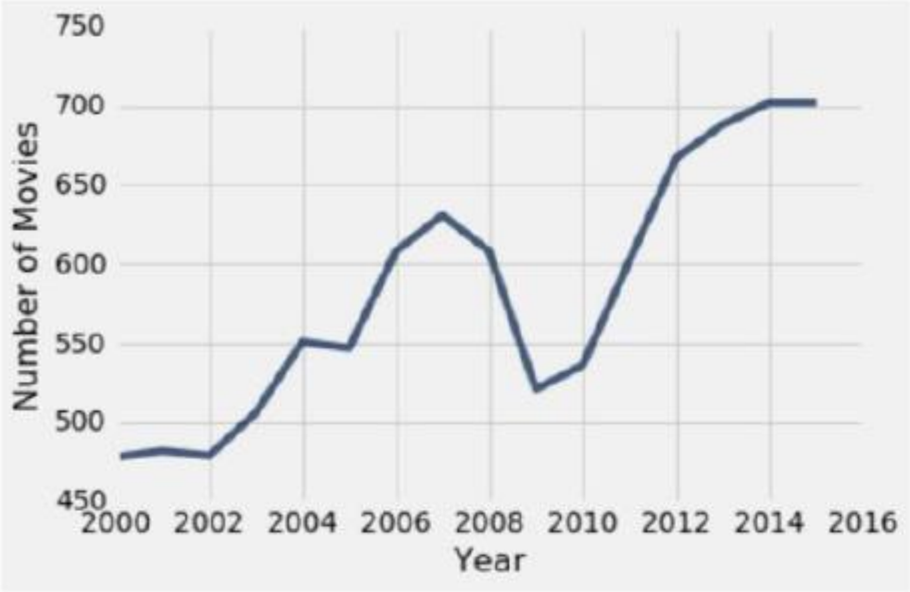
Census example had numerical SEX code (0, 1, and 2)

- It doesn't make sense to perform arithmetic on these "numbers"

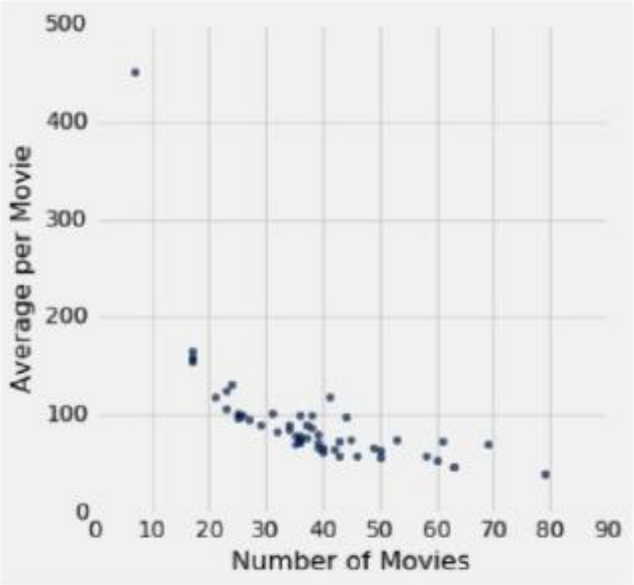
Eats chocolate	Happiness score
Chocolate	10
No Chocolate	0
Chocolate	10
Chocolate	10

Plotting two numerical variables

Line graph: `plot`



Scatter plot: `scatter`



We will discuss plotting a single numerical variable soon, but let's first discuss categorical data…

Categorical Data

Statistics and visualizing categorical data

We usually summarize categorical data by creating:

- **Frequency tables**: contain a count the number of items in each category
- **Relative frequency tables**: count the proportion of items in each category

We can use the tb.group("categorical col name") method to create frequency tables

A **bar chart** is a visual display of a frequency table

- One bar for each category
- Length of bar is the count of individuals in that category

Let's explore this in Jupyter!

Histograms

Histograms

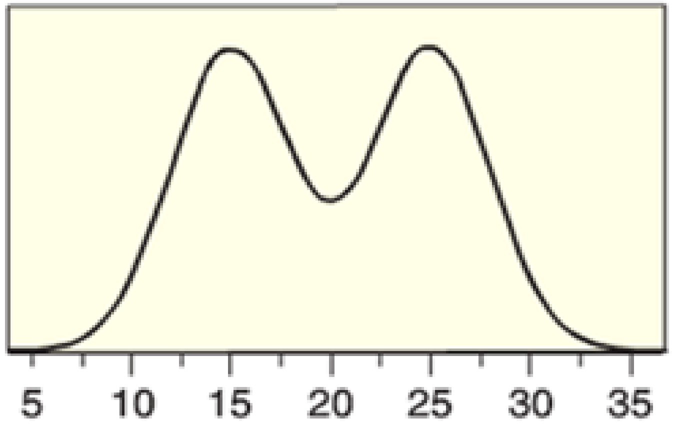
Histograms are a way to visualize numerical data that give us insight into which ranges of numerical values occur most frequently

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

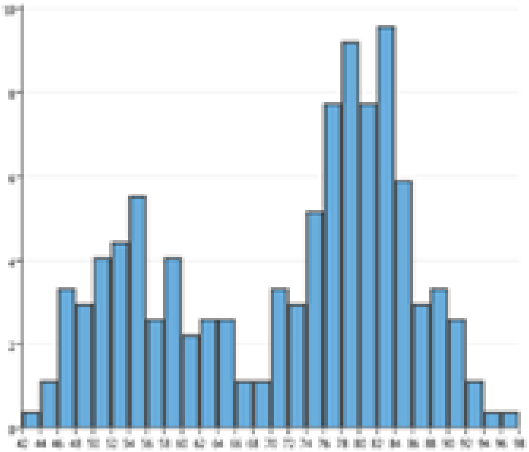
To create histograms we:

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

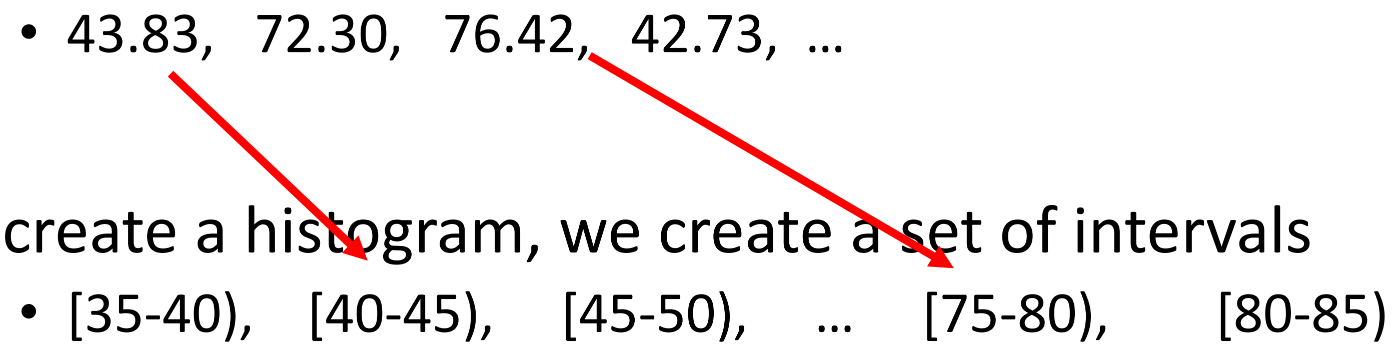


Histogram



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

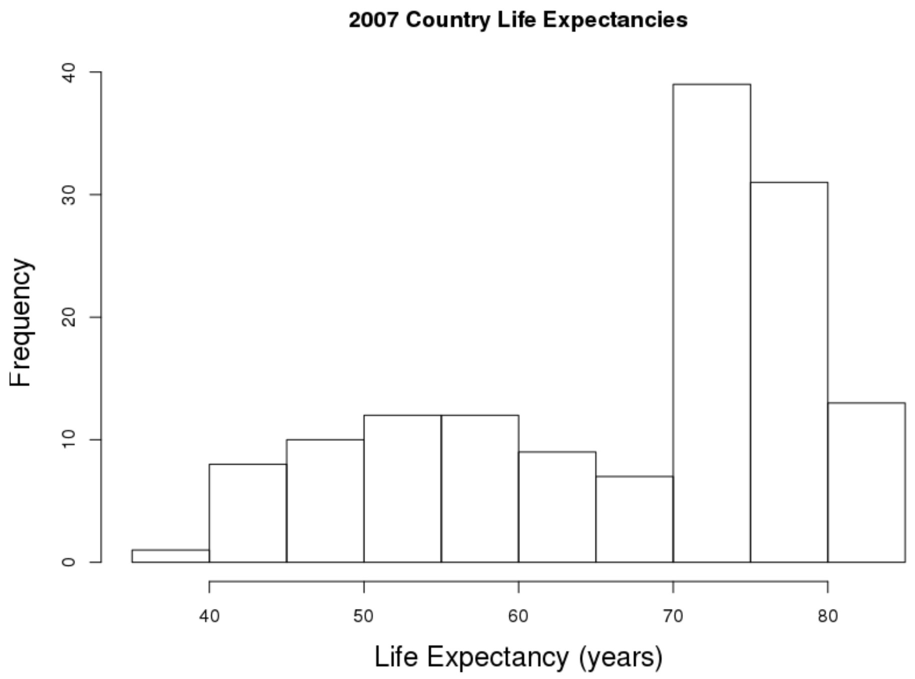
Histograms – countries life expectancy in 2007

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)
```

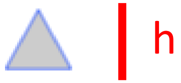


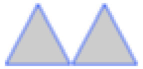
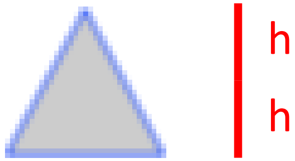
Area principal

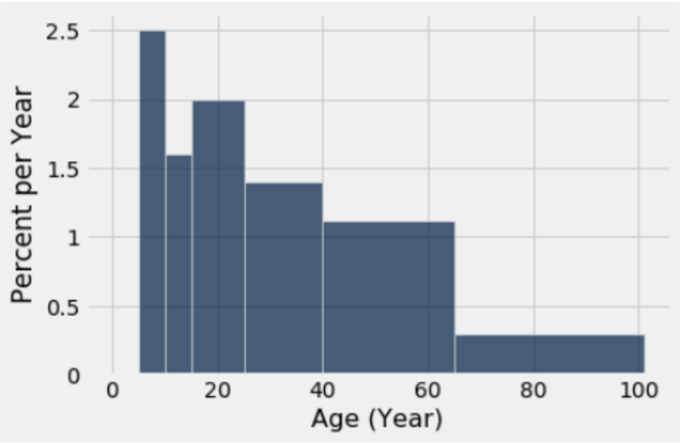
It is possible to create histograms with different sized bins

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

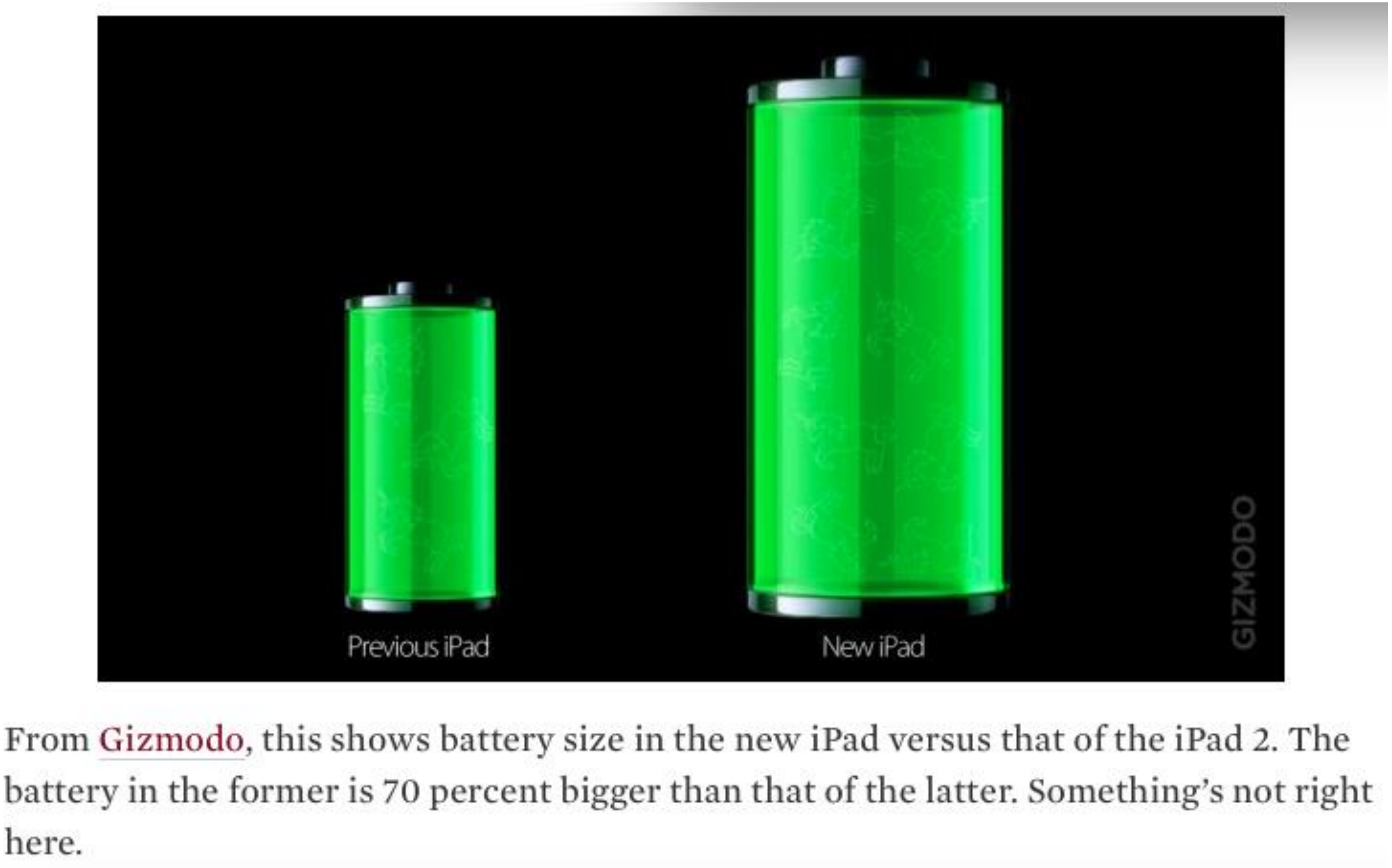
For example:

- 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.