lec01

October 10, 2024

1 Lecture 1 – Introduction

1.1 Data 6, Summer 2022

This is a Jupyter notebook. We'll write all of our code in this class in a Jupyter notebook.

Today, don't worry about how any of this works. Throughout the summer, we'll learn how each of these pieces work.

Note: If you're having trouble loading any plots or maps, try using Google Chrome.

[8]:

```
from datascience import *
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
import plotly.graph_objects as go
```

1.2 California universities

Here, we'll load in data about all public universities in California. The data comes from this Wikipedia article.

[9]:

```
# Load in the "california_universities.csv" file in the "data" folder
uni = Table.read_table('data/california_universities.csv')

# Remove irregular formatting
uni = uni.with_columns(
    'Enrollment', uni.apply(lambda s: int(s.replace(',', '')), 'Enrollment'),
    'Founded', uni.apply(lambda s: int(s.replace('*', '')), 'Founded')
)
```

Data is often stored in tables. In about a few weeks, we'll become very, very familiar with how tables work. But for now, let's just observe.

[10]:

```
# Let's see what the table looks like
uni.show(5)
```

```
<IPython.core.display.HTML object>
```

Let's start asking questions.

1.2.1 What are the largest public universities in California?

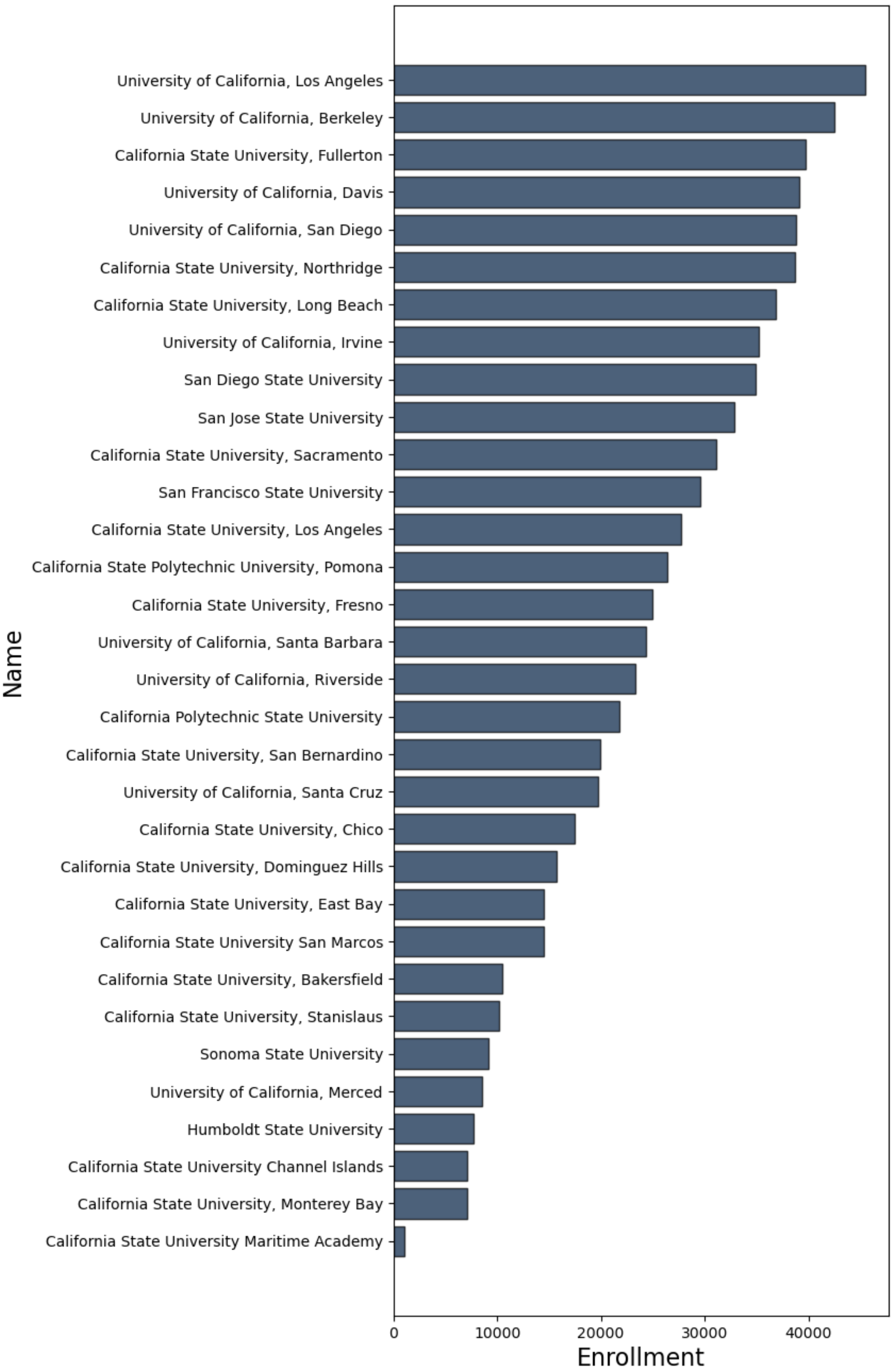
[11]:

```python
# Largest universities - table format
uni.sort("Enrollment", descending=True).show(5)
```

```
<IPython.core.display.HTML object>
```

[12]:

```python
# Can we visualize the sizes of each university?
uni.sort("Enrollment", descending=True).barh("Name", "Enrollment")
```


University of California, Los Angeles
University of California, Berkeley
California State University, Fullerton
University of California, Davis
University of California, San Diego
California State University, Northridge
California State University, Long Beach
University of California, Irvine
San Diego State University
San Jose State University
California State University, Sacramento
San Francisco State University
California State University, Los Angeles
California State Polytechnic University, Pomona
California State University, Fresno
University of California, Santa Barbara
University of California, Riverside
California Polytechnic State University
California State University, San Bernardino
University of California, Santa Cruz
California State University, Chico
California State University, Dominguez Hills
California State University, East Bay
California State University San Marcos
California State University, Bakersfield
California State University, Stanislaus
Sonoma State University
University of California, Merced
Humboldt State University
California State University Channel Islands
California State University, Monterey Bay
California State University Maritime Academy
Name
0
10000
20000
30000
40000
Enrollment

1.2.2 What's the oldest public university in California?

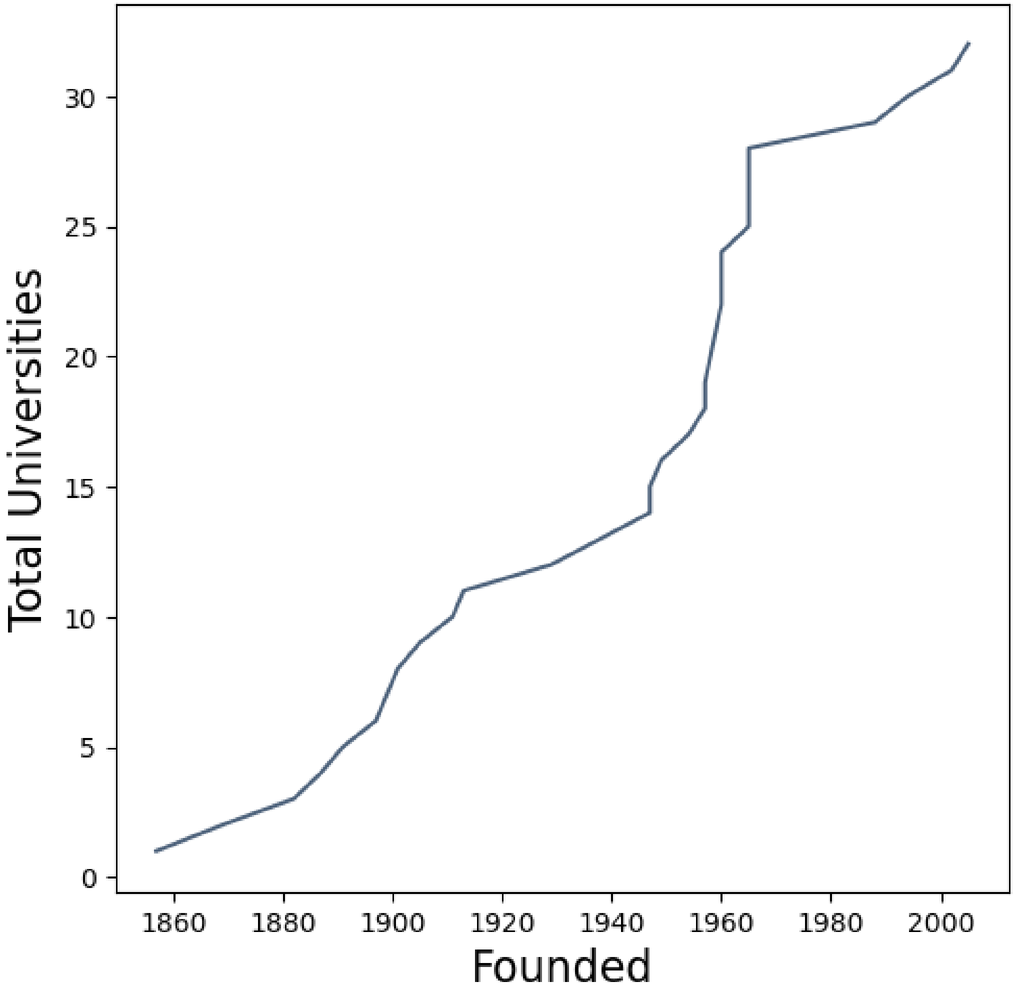
[13]:

```
# Oldest university - table format
uni.sort("Founded", descending=False).show(1)
```

```
<IPython.core.display.HTML object>
```

[14]:

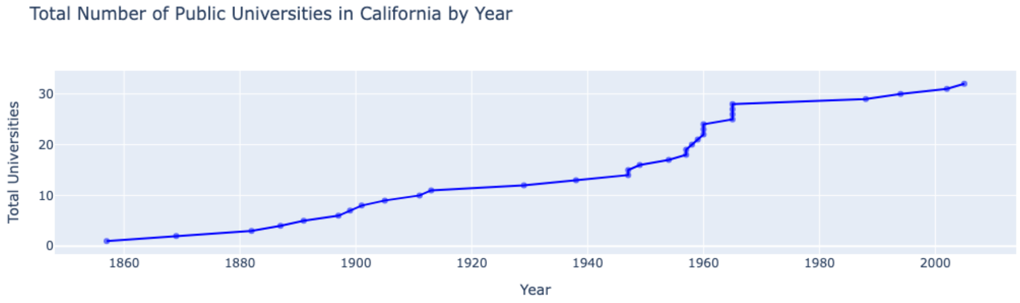
```
# How can we visualize the ages of the universities?
uni_copy = uni.sort('Founded').with_columns('Total Universities', np.arange(1,␣
 ↪uni.num_rows + 1))
uni_copy.plot('Founded', 'Total Universities')
```



Let's add some spice.

[15]:

```
# Just run me
fig = go.Figure()

fig.add_trace(
    go.Scatter(x = uni_copy.column('Founded'),
               y = uni_copy.column('Total Universities'),
               hovertext = uni_copy.column('Name'),
               mode = 'markers',
              )
)

fig.add_trace(
    go.Scatter(x = uni_copy.column('Founded'),
               y = uni_copy.column('Total Universities'),
               line = dict(color = 'blue'),
              )
)

fig.update_layout(title = 'Total Number of Public Universities in California by␣
 ↪Year',
                  xaxis_title = 'Year',
                  yaxis_title = 'Total Universities',
                  showlegend = False)

fig.show()
```



1.3 Public Universities in California (and you!)

1.3.1 Where are the public universities in California located?

First, we need some additional information:

[16]:

```
# Load in the "california_universities.csv" file in the "data" folder
uni_locations = Table.read_table('data/uni_locations.csv')
```

```python
uni_locations
```

[16]:

```
Latitude | Longitude | University
37.8719  | -122.259  | University of California, Berkeley
38.5382  | -121.762  | University of California, Davis
33.6405  | -117.844  | University of California, Irvine
34.0689  | -118.445  | University of California, Los Angeles
37.3661  | -120.422  | University of California, Merced
33.9737  | -117.328  | University of California, Riverside
32.8801  | -117.234  | University of California, San Diego
34.414   | -119.849  | University of California, Santa Barbara
36.9881  | -122.058  | University of California, Santa Cruz
38.0689  | -122.23   | California State University Maritime Academy
… (22 rows omitted)
```

Let combine some data.

[17]:

```python
# Join the `uni` and `uni_locations` tables
unis_with_location = uni.join("Name", uni_locations, "University")
unis_with_location
```

[17]:

```
Name                                             | City            | County
| Enrollment | Founded | Latitude | Longitude
California Polytechnic State University          | San Luis Obispo | San Luis
Obispo | 21812      | 1901    | 35.305   | -120.662
California State Polytechnic University, Pomona | Pomona          | Los Angeles
| 26443      | 1938    | 34.0589  | -117.819
California State University Channel Islands     | Camarillo       | Ventura
| 7095       | 2002    | 34.1621  | -119.043
California State University Maritime Academy    | Vallejo         | Solano
| 1017       | 1929    | 38.0689  | -122.23
California State University San Marcos          | San Marcos      | San Diego
| 14511      | 1988    | 33.1295  | -117.16
California State University, Bakersfield        | Bakersfield     | Kern
| 10493      | 1965    | 35.3487  | -119.103
California State University, Chico              | Chico           | Butte
| 17488      | 1887    | 39.7298  | -121.846
California State University, Dominguez Hills    | Carson          | Los Angeles
| 15741      | 1960    | 33.8662  | -118.257
California State University, East Bay           | Hayward         | Alameda
| 14525      | 1959    | 37.6571  | -122.057
California State University, Fresno             | Fresno          | Fresno
| 24995      | 1911    | 36.8134  | -119.746
… (22 rows omitted)
```

What if we want to plot these on a map?

We can use the `plotly` API (essentially a library of additional things we can do with Python)!

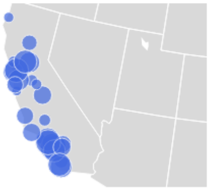
[18]:

```
# Just run me

def bubble_plot(tbl, text, size=None, lat="Latitude", lon="Longitude",␣
 ↪color=None, title=None, scale_factor=150):
    fig = go.Figure()

    if not color:
        color_arr = ['royalblue'] * tbl.num_rows
    else:
        color_arr = tbl.column(color)

    if not size:
        size_arr = [1 / scale_factor] * tbl.num_rows
    else:
        size_arr = tbl.column(size) / scale_factor

    fig = fig.add_trace(go.Scattergeo(
                            lat = tbl.column(lat),
                            lon = tbl.column(lon),
                            text = tbl.column(text),
                            marker = dict(
                                size = size_arr,
                                sizemode = 'area',
                                color = color_arr
                            )
                        ))

    fig.update_geos(fitbounds="locations")
    fig.update_layout(
        geo = dict(
                scope = 'usa',
                landcolor = 'rgb(217, 217, 217)',
            ),
        title = title
    )

    return fig
```

[19]:

```
# Call the `bubble_plot` function, passing in the proper arguments
fig = bubble_plot(unis_with_location, text="Name", size="Enrollment",␣
 ↪title="Public Universities in California")
fig.show()
```

Public Universities in California



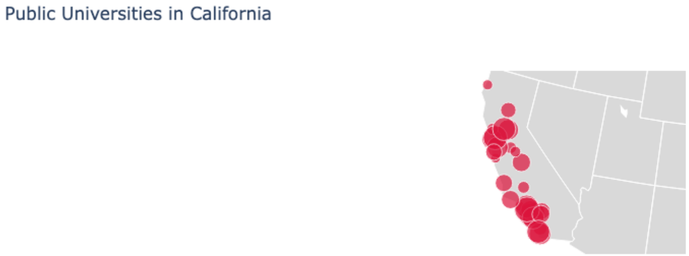
Can we add more information?

[20]:
```
# Let's add a color column
unis_with_color = unis_with_location.with_column('Color', ['crimson'] *␣
 ↪unis_with_location.num_rows)
unis_with_color
```

[20]:
```
Name                                          | City            | County
| Enrollment | Founded | Latitude | Longitude | Color
California Polytechnic State University       | San Luis Obispo | San Luis
Obispo | 21812      | 1901    | 35.305   | -120.662  | crimson
California State Polytechnic University, Pomona | Pomona          | Los Angeles
| 26443      | 1938    | 34.0589  | -117.819  | crimson
California State University Channel Islands  | Camarillo       | Ventura
| 7095       | 2002    | 34.1621  | -119.043  | crimson
California State University Maritime Academy  | Vallejo         | Solano
| 1017       | 1929    | 38.0689  | -122.23   | crimson
California State University San Marcos       | San Marcos      | San Diego
| 14511      | 1988    | 33.1295  | -117.16   | crimson
California State University, Bakersfield     | Bakersfield     | Kern
| 10493      | 1965    | 35.3487  | -119.103  | crimson
California State University, Chico           | Chico           | Butte
| 17488      | 1887    | 39.7298  | -121.846  | crimson
California State University, Dominguez Hills | Carson          | Los Angeles
| 15741      | 1960    | 33.8662  | -118.257  | crimson
California State University, East Bay        | Hayward         | Alameda
| 14525      | 1959    | 37.6571  | -122.057  | crimson
California State University, Fresno          | Fresno          | Fresno
| 24995      | 1911    | 36.8134  | -119.746  | crimson
… (22 rows omitted)
```

[21]:
```
# Use the `bubble_plot` function to map the universities, this time specifying␣
 ↪the bubble color
```

```
fig = bubble_plot(unis_with_color, text="Name", size="Enrollment",␣
 ↪color="Color", title="Public Universities in California")
fig.show()
```



Public Universities in California

It would be nice if this were color-coded based on UC vs. CSU. We can do that!

[22]:

```
#Just run me
def code_uc(name):
    if 'University of California' in name:
        return 'royalblue'
    else:
        return 'crimson'
```

[23]:

```
# Apply the `code_uc` function to the 'Name' column to color-code the␣
 ↪universities
uni_locations_separate = unis_with_color.with_column('Color', unis_with_color.
 ↪apply(code_uc, 'Name'))
uni_locations_separate
```

[23]:

```
Name                                          | City            | County
| Enrollment | Founded | Latitude | Longitude | Color
California Polytechnic State University       | San Luis Obispo | San Luis
Obispo | 21812      | 1901    | 35.305    | -120.662  | crimson
California State Polytechnic University, Pomona | Pomona        | Los Angeles
| 26443      | 1938    | 34.0589  | -117.819  | crimson
California State University Channel Islands  | Camarillo       | Ventura
| 7095       | 2002    | 34.1621  | -119.043  | crimson
California State University Maritime Academy | Vallejo         | Solano
| 1017       | 1929    | 38.0689  | -122.23   | crimson
California State University San Marcos       | San Marcos      | San Diego
| 14511      | 1988    | 33.1295  | -117.16   | crimson
California State University, Bakersfield    | Bakersfield     | Kern
| 10493      | 1965    | 35.3487  | -119.103  | crimson
California State University, Chico           | Chico           | Butte
```

```
| 17488       | 1887     | 39.7298   | -121.846   | crimson
California State University, Dominguez Hills     | Carson             | Los Angeles
| 15741       | 1960     | 33.8662   | -118.257   | crimson
California State University, East Bay            | Hayward            | Alameda
| 14525       | 1959     | 37.6571   | -122.057   | crimson
California State University, Fresno              | Fresno             | Fresno
| 24995       | 1911     | 36.8134   | -119.746   | crimson
… (22 rows omitted)
```

[24]:

```python
# Plot the color-coded universities on the map with the `bubble_plot` function
fig = bubble_plot(uni_locations_separate, text="Name", size="Enrollment", color="Color", title="UCs and CSUs")
fig.show()
```

UCs and CSUs

Violà!

1.3.2 Where are you all from?

Using the responses from the welcome survey, let's use our knowledge of Python to plot the hometowns of the students in Data 6!

[25]:

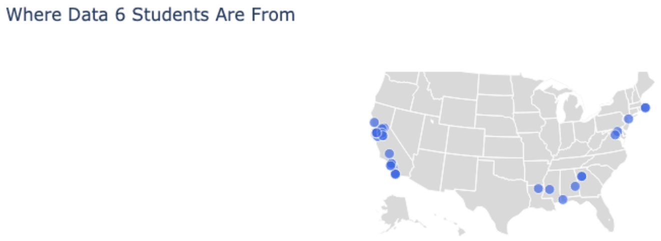
```python
# Load in the "student_hometowns.csv" file from the "data" folder
hometowns = Table.read_table("data/student_hometowns.csv")
hometowns
```

[25]:

```
City                  | State | Latitude | Longitude
Modesto               | CA    | 37.6391  | -120.997
Miami                 | FL    | 41.6688  | -70.2962
Tuskegee              | AL    | 32.4302  | -85.7077
South San Francisco   | CA    | 37.6547  | -122.408
San Diego             | CA    | 32.7153  | -117.157
San Gabriel           | CA    | 34.0961  | -118.106
Atlanta               | GA    | 33.749   | -84.388
Orange County         | CA    | 33.7175  | -117.831
Granite Bay           | CA    | 38.7632  | -121.164
```

```
Oakland              | CA     | 37.8044   | -122.271
… (23 rows omitted)
```

```
[26]: # Plot the hometowns of Data 6 students using the `bubble_plot` function
fig = bubble_plot(hometowns, text="City", title="Where Data 6 Students Are From", scale_factor=0.02)
fig.show()
```



Where Data 6 Students Are From

The end!