

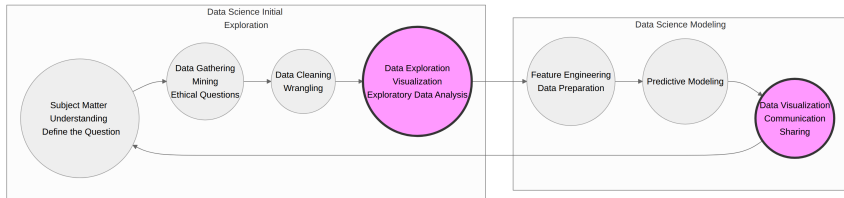
Introduction to Data Science

Doing Data Science and Web Scraping

Joanna Bieri DATA101

Mini Review - What Does It Mean to Do Data Science?

REVIEW



- It is your job to make sure your results are **reproducible**.
- It is your job to really **get to know your data**.

Mini Review - Data Science Questions

REVIEW

- 1 **Descriptive:** summarize a characteristic of a set of data
- 2 **Exploratory:** analyze to see if there are patterns, trends, or relationships between variables
- 3 **Inferential:** analyze patterns, trends, or relationships in representative data from a population
- 4 **Predictive:** make predictions for individuals or groups of individuals
- 5 **Causal:** whether changing one factor will change another factor, on average, in a population
- 6 **Mechanistic:** explore “how” as opposed to whether

Mini Review - Exploratory Data Analysis Checklist

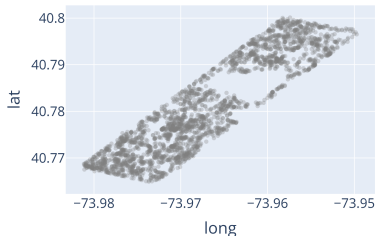
REVIEW

- Formulate your question
- Read in your data
- Check the dimensions
- Look at the data
- Validate with at least one external data source
- Make a plot
- Try the easy solution first

Mini Review - Example - Squirrels

REVIEW

```
fig = px.scatter(Df, x='long', y='lat', opacity=.3,  
                 color_discrete_sequence=['gray'])
```



3,023 squirrel sightings. Is this actually Central Park? YES - we can google the coordinates.

Mini Review - Web Scraping Ethical Issues

REVIEW

- Some data is private or protected. Just because you have access to a websites data doesn't mean you are allowed to scrape it.
- Some websites have rules against scraping. If the site has a policy about scraping data then you should follow it.
- Always cite the source of your data and make sure you are doing something new with it.
- **The Ethical Scraper:** I will request data at a reasonable rate. I will only save the data I absolutely need from your page.

Mini Review - Using pandas to Get Table Data

REVIEW

```
DF = pd.read_html(my_website)
ACES = DF[0].copy()
```

- **.isna()** creates a mask for whether or not there is a NaN in each row of the data.
- **.fillna()** will replace NaN in your data set with whatever you put inside the parenthesis.
- **.dropna()** will drop all rows that contain NaN.
- Sometimes you get errors! A 403 Forbidden error means the website is trying to stop you from scraping.

Mini Review - Using BeautifulSoup to Get HTML Code

REVIEW

```
raw_code = requests.get(website)
html_doc = raw_code.text
soup = BeautifulSoup(html_doc, 'html.parser')

result = soup.find_all('h3', class_="country-name")
```

- 1 Start an EMPTY data frame using `DF = pd.DataFrame()`
- 2 Add our **soup.find_all** results as a column of the data frame
- 3 Fix the data to strip off all of the unneeded text

Data Science Questions

ASK

A. What is the average price of the houses in our data?

B. Will this house sell for more than \$500,000?

What type of question is each one?

Data Science Questions

ANSWER

- A is **descriptive**: it summarizes a characteristic of the data we have.
- B is **predictive**: it makes a prediction for one house.
- The type of question tells you what kind of analysis you need and what you are allowed to say at the end.

Is the Data Right for the Question?

ASK

We survey University of Redlands students:

"How many people in your household went to college?"

We use the answers to describe all households
in Southern California. What is wrong?

Is the Data Right for the Question?

ANSWER

- Every household in the survey has at least one college student.
- The sample is **biased** toward households that go to college.
- Ask: was the data collected in a way that introduces bias?

Dealing with NaNs

PREDICT

df

	player	pts	award
0	Ana	10.0	NaN
1	Bea	NaN	NaN
2	Cam	7.0	NaN

`df.dropna()`

What comes out? How many rows are left?

Dealing with NaNs

ANSWER

```
df.dropna()
```

```
Empty DataFrame
```

```
Columns: [player, pts, award]
```

```
Index: []
```

Zero rows! `dropna()` removes any row with a NaN in it, and every row has a NaN in the award column. Look at where the NaNs are before you drop anything: `df.isna().sum()` gives player 0, pts 1, award 3.

Getting the Text

PREDICT

```
<h3 class="team-name">  
    Aces  
</h3>
```

```
result = soup.find_all('h3', class_="team-name")  
result[0].text
```

What comes out? What do we add to clean it up?

Getting the Text

ANSWER

```
result[0].text
```

```
'\n    Aces\n'
```

```
result[0].text.lstrip().rstrip()
```

```
'Aces'
```

.text keeps the spaces and the new lines that were inside the tag.
.lstrip() and .rstrip() take them off the left and the right.

Getting the Link

REVERSE

```
<a href="/pages/simple/">Countries of the World</a>
```

```
result = soup.find_all('a')
```

We want the words. We also want the link. What are the two commands?

Getting the Link

ANSWER

```
result[0].text          # 'Countries of the World'  
result[0].get('href')   # '/pages/simple/'
```

The words are between the tags, so `.text` gets them. The link is inside the tag, so we ask for it by name with `.get('href')`. You need both for Exercise 1 in the homework.

Web Scraping Ethics

ASK

You scrape the scores for your favorite team
every day, just for you.

Your friend scrapes the same site
and posts the scores on a website with ads.

Are the ethics the same?

Web Scraping Ethics

ANSWER

- No. Ethics are different for a personal project and for a business or a website.
- If you publish the data or make money from it, it is EXTRA important to respect the original owner of the data.
- Drive web traffic back to their site, check with them about using their data, and cite them.

Questions?

Questions?