

Introduction to Data Science

Data and Visualization

Joanna Bieri DATA101

Intro

Important Information

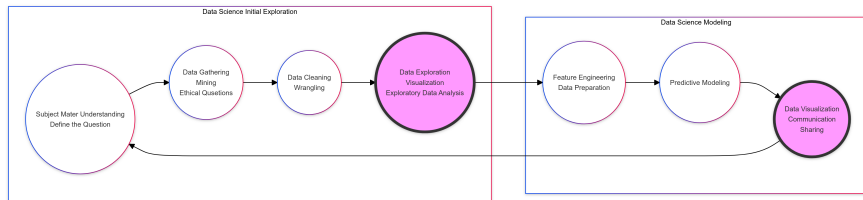
- Email: joanna_bieri@redlands.edu
- Office Hours: Duke 209 [Click Here for Joanna's Schedule](#)

Announcements

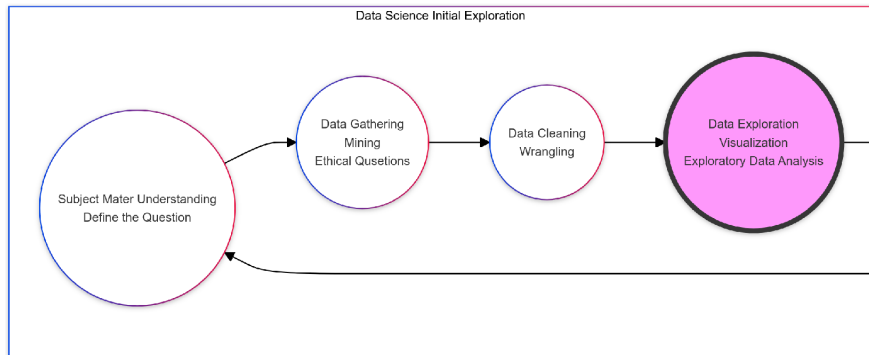
Please come to office hours to get help!

I can be **REALLY FLEXIBLE** with deadlines in the first few weeks of class, so if you run into technology issues please don't panic.

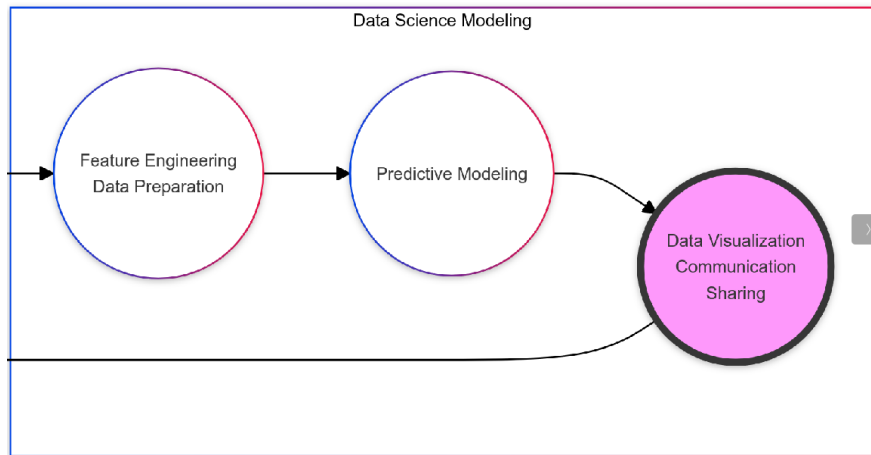
The Data Science Lifecycle



The Data Science Lifecycle



The Data Science Lifecycle



Some important things to keep in mind!

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

Data and Visualization

Data and Visualization

“The simple graph has brought more information to the data analyst’s mind than any other device.” — John Tukey

Data and Visualization:

We will do this by exploring the Star Wars Characters data set:

Star Wars Data on GitHub

Data and Visualization:

Data terminology

- Each row is an **observation**
- Each column is a **variable**

Example Observation

name: Luke Skywalker

height: 172.0

mass: 77.0

hair_color: blond

skin_color: fair

eye_color: blue

birth_year: 19.0

sex: male

gender: masculine

homeworld: Tatooine

species: Human

films: A New Hope, The Empire Strikes Back,

Return of the Jedi, Revenge of the Sith, The Force Awakens

vehicles: Snowspeeder, Imperial Speeder Bike

starships: X-wing, Imperial shuttle

Questions

```
observation_name = 'Luke Skywalker'  
DF[DF['name']==observation_name]
```

Q How would you update the cell above to search for a different character? Make a change to the code above to find information about C-3PO.

Q In the data for C-3PO one of the entries says NaN. What does this mean?

Q Look at some observations for other characters. Just choose a few different names and see what their data says.

Q What happens if you spell a name wrong or forget to capitalize?

Other Code

```
# Number of Rows - Observations  
print(len(DF))
```

```
# Number of Columns - Variables  
print(len(DF.columns))
```

```
# Total dimensions of the data  
print(DF.shape)
```

87 observations and 14 variables

Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA)

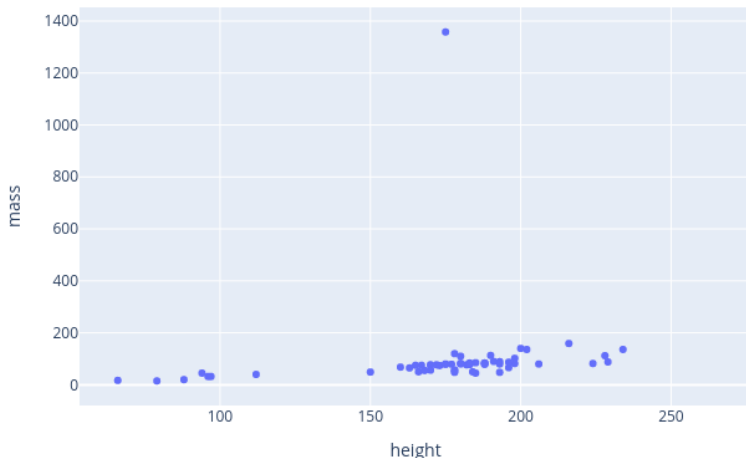
Now we will start really analyzing the data! This might consist of 1. Summary Statistics 2. Visualization (Today's Class) 3. Data Wrangling (Future Class)

Mass vs. Height

```
fig = px.scatter(DF,  
                 x='height',  
                 y='mass',  
                 title='Mass vs. Height of Starwars Characters')  
fig.write_image('images/starwars.png')  
fig.show()
```

Mass vs. Height

Mass vs. Height of Starwars Characters



Mass vs. Height

```
fig = px.scatter(DF,  
                 x='height',  
                 y='mass',  
                 title='Mass vs. Height of Starwars Characters',  
                 hover_data='name')  
fig.show()
```



Figure 1: Jabba

Birth Year vs Mass

Q See if you can figure out how to make a plot of the birth_year vs mass with the hover data being the name?

Visualization

Why Visualize?!?!

Anscombe's Quartet

x	y1	y2	y3	y4
0	10.0	8.04	9.14	7.46
1	8.0	6.95	8.14	6.77
2	13.0	7.58	8.74	12.74
3	9.0	8.81	8.77	7.11
4	11.0	8.33	9.26	7.81
5	14.0	9.96	8.10	8.84
6	6.0	7.24	6.13	6.08
7	4.0	4.26	3.10	5.39
8	12.0	10.84	9.13	8.15
9	7.0	4.82	7.26	6.42
10	5.0	5.68	4.74	5.73

Summary Statistics

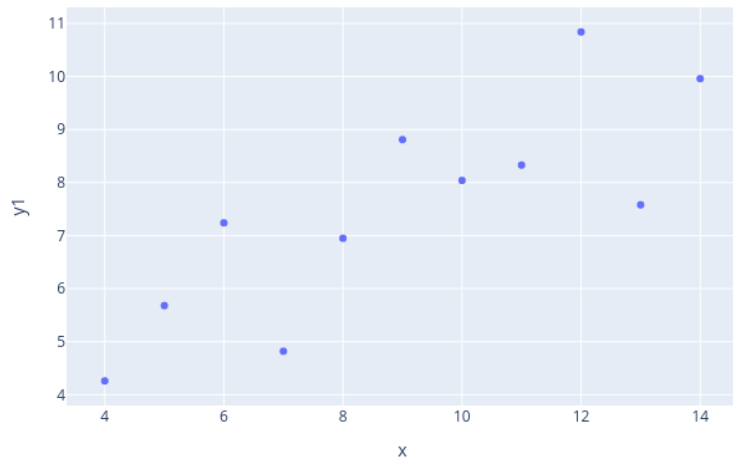
	x	y1	y2	y3	y4
count	11.000000	11.000000	11.000000	11.000000	11.000000
mean	9.000000	7.500909	7.500909	7.500000	7.500909
std	3.316625	2.031568	2.031657	2.030424	2.030579
min	4.000000	4.260000	3.100000	5.390000	5.250000
25%	6.500000	6.315000	6.695000	6.250000	6.170000
50%	9.000000	7.580000	8.140000	7.110000	7.040000
75%	11.500000	8.570000	8.950000	7.980000	8.190000
max	14.000000	10.840000	9.260000	12.740000	12.500000

Visualization

```
fig = px.scatter(DF_new,x='x',y='y4',title='x vs y4')  
fig.write_image('images/y4.png')  
fig.show()
```

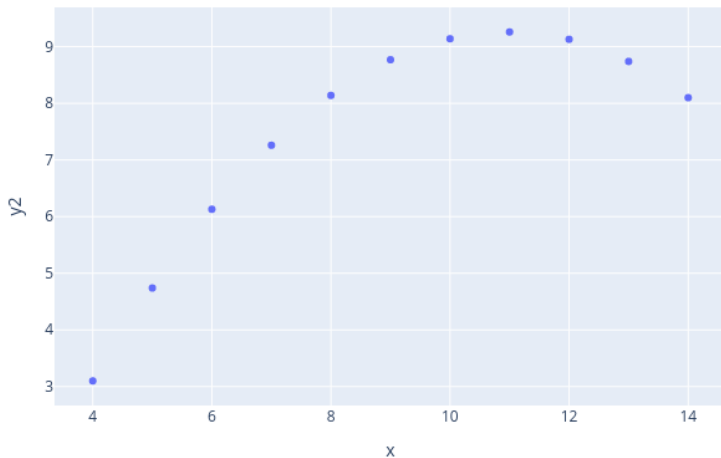
Visualization

x vs y1



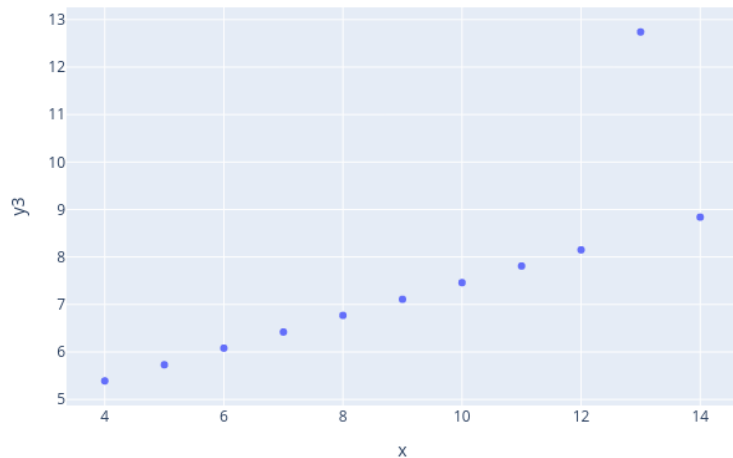
Visualization

x vs y2



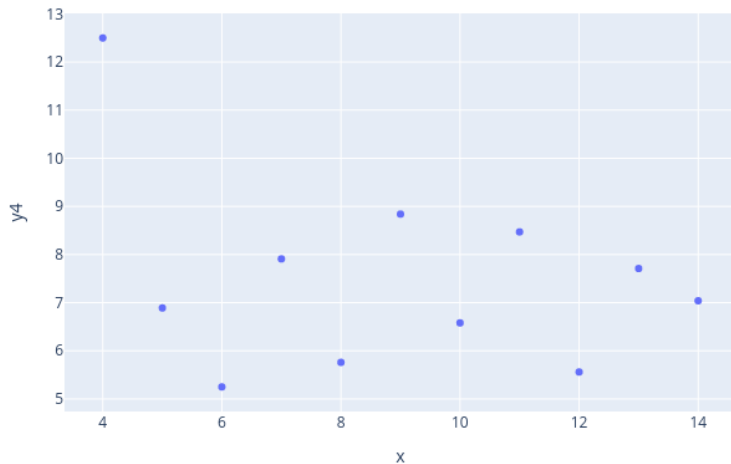
Visualization

x vs y3



Visualization

x vs y4



Data and Visualization

Questions?