1. Intro to data visualization and graphical grammars

Lecture learning goals

1. Explain the importance of data visualization.
2. Explain what constitutes a grammar of graphics.
3. Create point and line visualizations in ggplot/Altair.
4. Transform data directly in ggplot/Altair instead of dplyr/pandas.
5. Combine geometric marks via layering.

Required activities

Before class:

- Nothing!

After class:

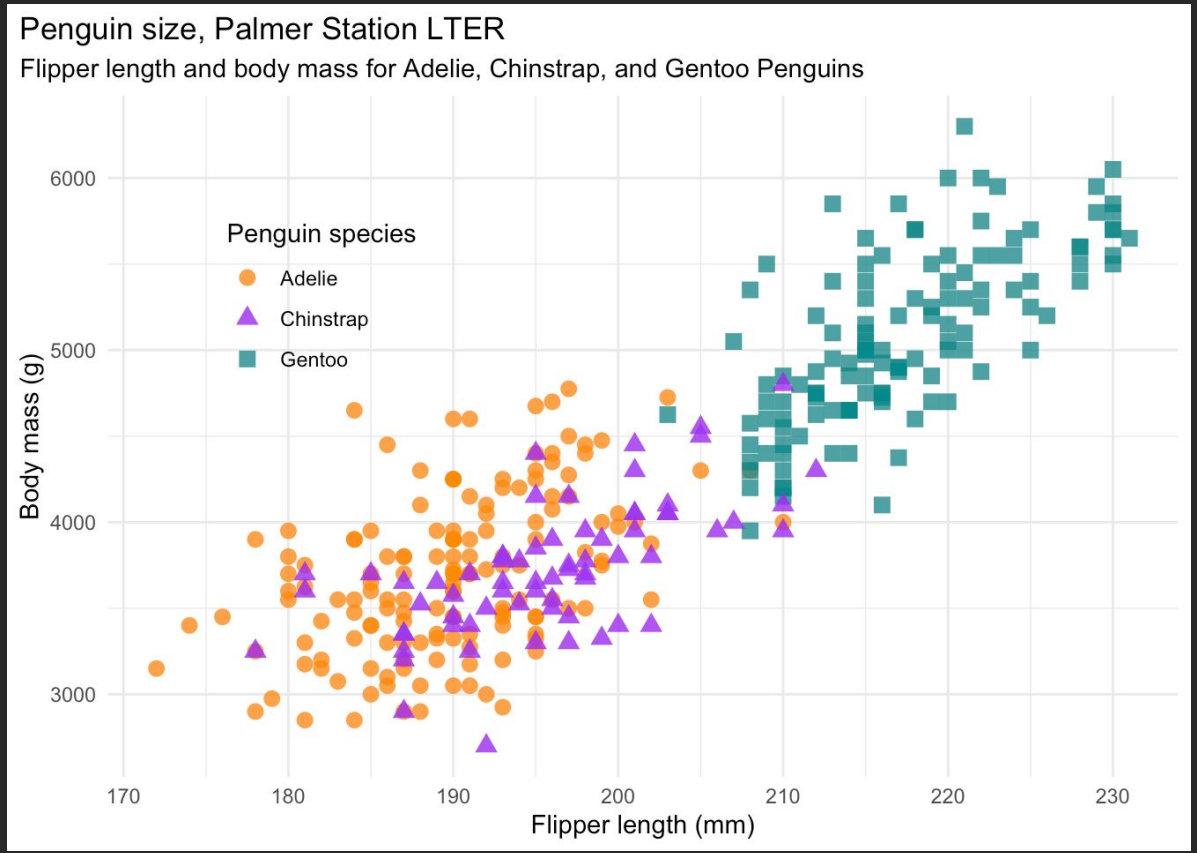
- These lecture notes.
- Data Visualization: A practical introduction by Kieran Healy, Section 1 - 1.2 (i.e. you stop when you reach 1.3) OR this video on the same topic until the “Perception” slide ~27 min.

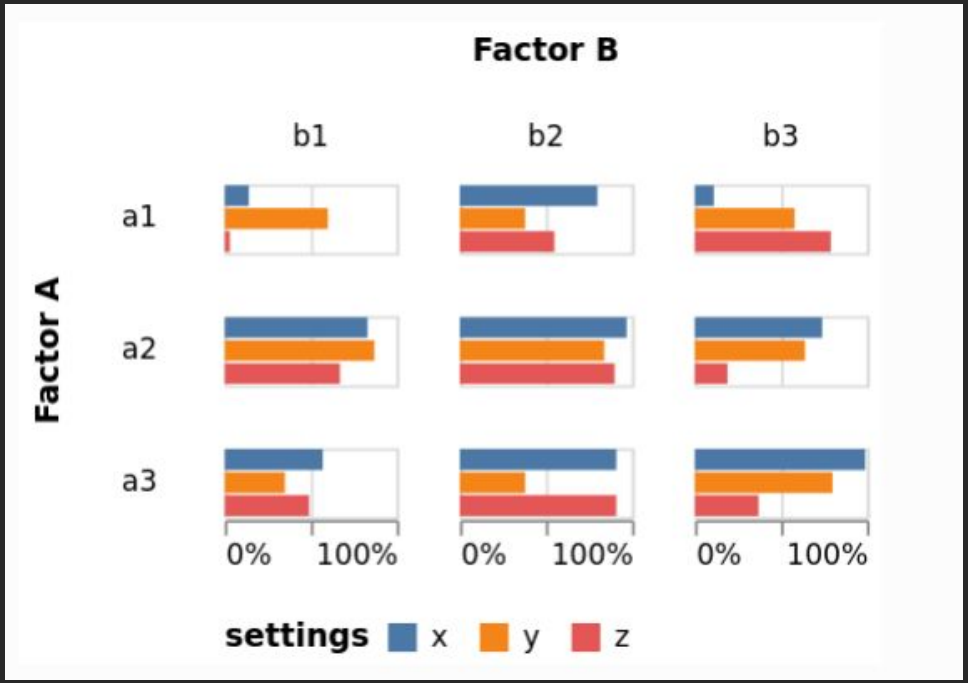
Have you done any type of data visualization before?

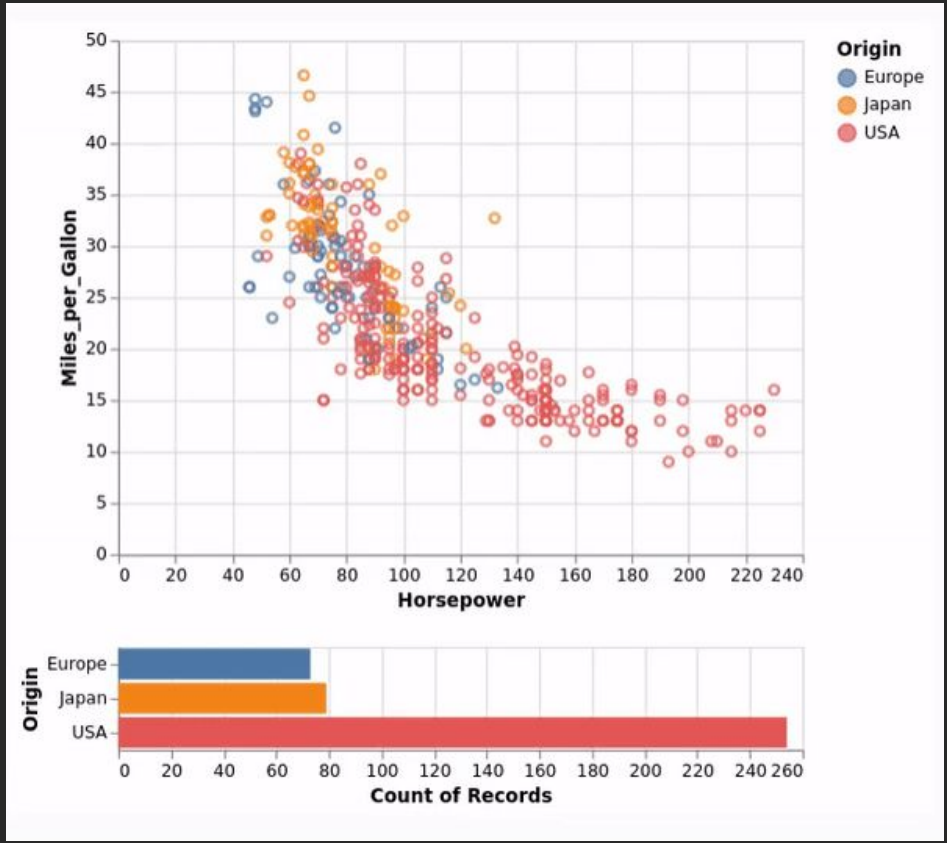
A. Yes

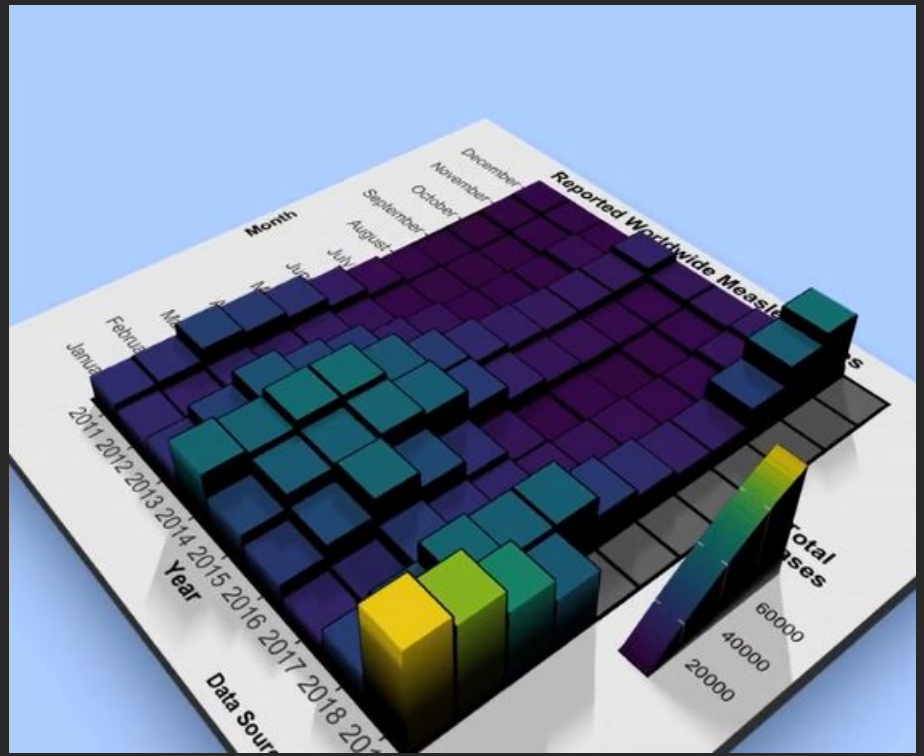
B. No

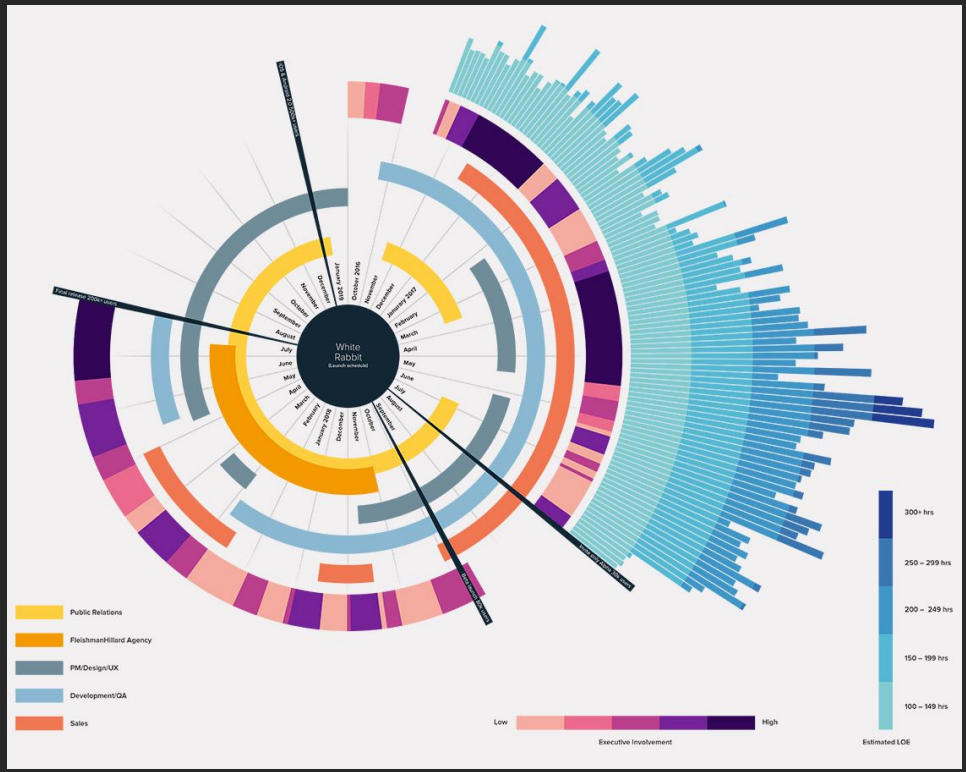
What is data viz?


Penguin size, Palmer Station LTER
Flipper length and body mass for Adelie, Chinstrap, and Gentoo Penguins
Penguin species
Adelie
Chinstrap
Gentoo
Body mass (g)
6000
5000
4000
3000
170
180
190
200
210
220
230
Flipper length (mm)


Factor B
b1
b2
b3
Factor A
a1
a2
a3
0%
100%
0%
100%
0%
100%
settings
x
y
z


Origin
Europe
Japan
USA
Miles_per_Gallon
Horsepower
Origin
Count of Records


Month
December
November
October
September
August
July
Reported Worldwide Measles
2011 2012 2013 2014 2015 2016 2017 2018
Year
Data Sour
60000
40000
20000


White Rabbit
(Launch scheduled)
Public Relations
FleishmanHillard Agency
PM/Design/UX
Development/QA
Sales
Low
High
Executive Involvement
300+ hrs
250 – 299 hrs
200 – 249 hrs
150 – 199 hrs
100 – 149 hrs
Estimated LOE

What is the purpose of visualizing data?

"Data visualization is where you start and end a data analysis”

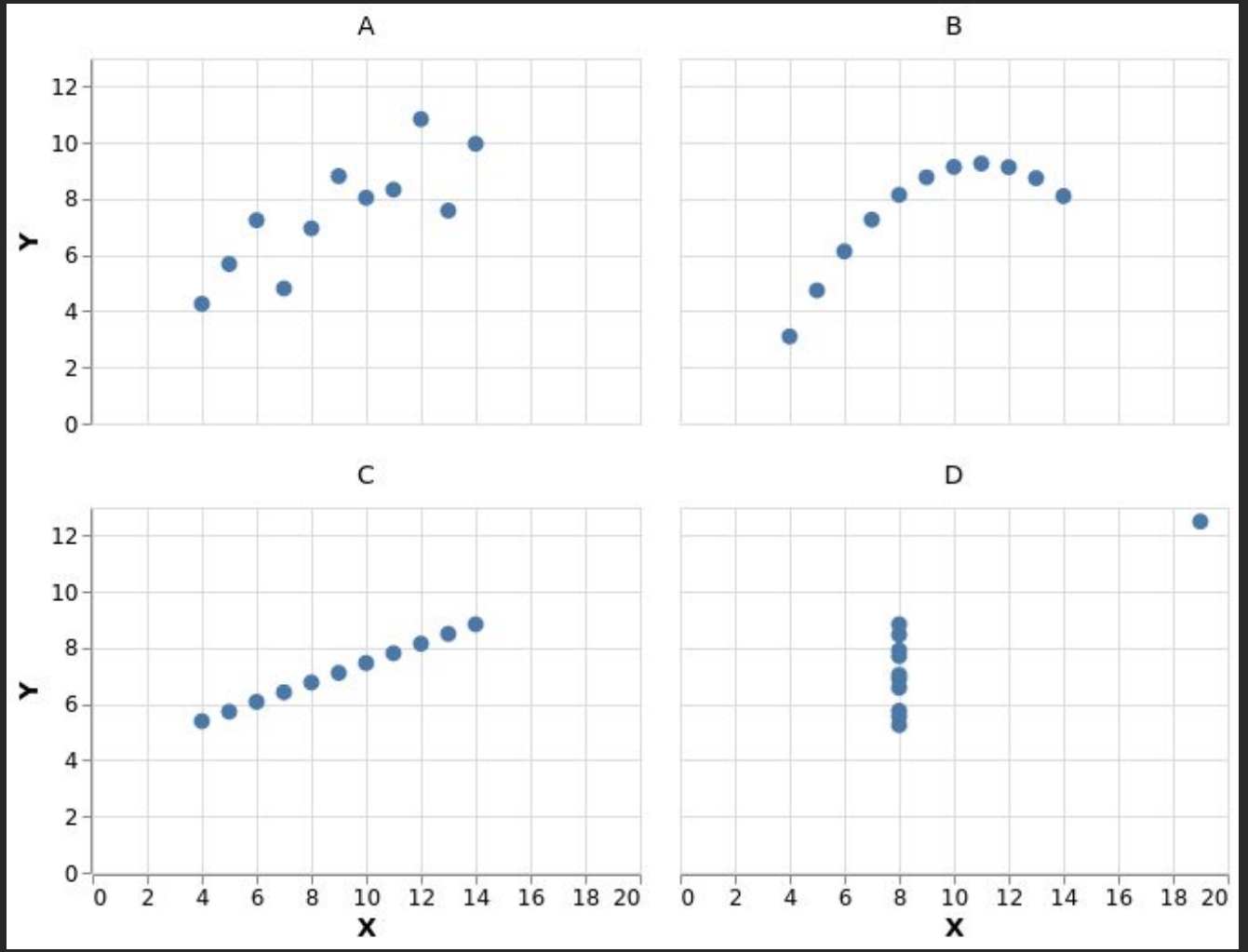
Why can’t we just show numbers?

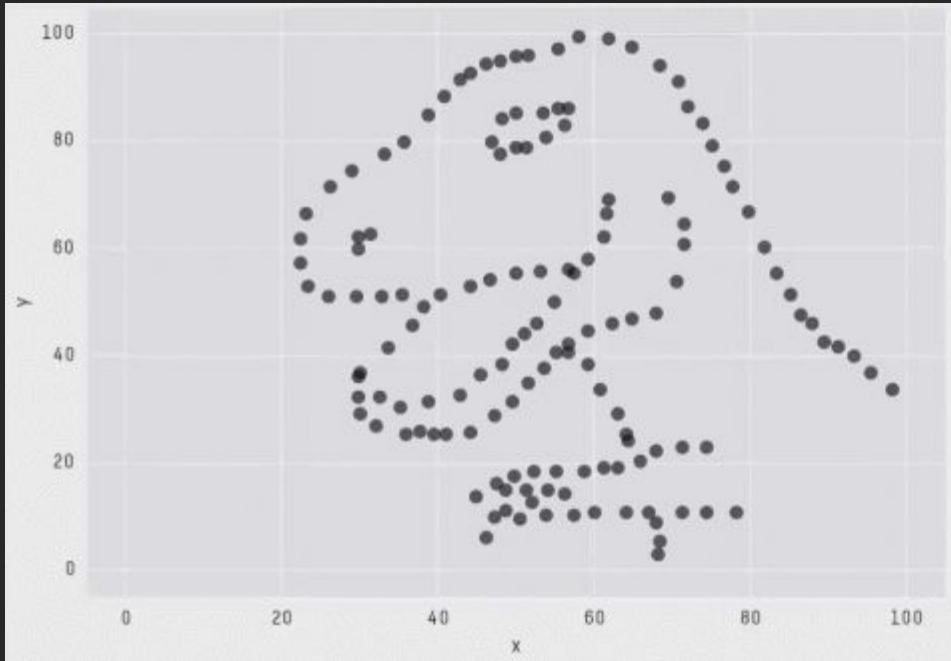
Why can’t we just show numbers?

Enter a text response in the iClicker

A		B		C		D	
X	Y	X	Y	X	Y	X	Y
10.00	8.04	10.00	9.14	10.00	7.46	8.00	6.58
8.00	6.95	8.00	8.14	8.00	6.77	8.00	5.76
13.00	7.58	13.00	8.74	13.00	8.50	8.00	7.71
9.00	8.81	9.00	8.77	9.00	7.11	8.00	8.84
11.00	8.33	11.00	9.26	11.00	7.81	8.00	8.47
14.00	9.96	14.00	8.10	14.00	8.84	8.00	7.04
6.00	7.24	6.00	6.13	6.00	6.08	8.00	5.25
4.00	4.26	4.00	3.10	4.00	5.39	19.00	12.50
12.00	10.84	12.00	9.13	12.00	8.15	8.00	5.56
7.00	4.81	7.00	7.26	7.00	6.42	8.00	7.91
5.00	5.68	5.00	4.74	5.00	5.73	8.00	6.89

	A		B		C		D	
	X	Y	X	Y	X	Y	X	Y
Mean	9.00	7.50	9.00	7.50	9.00	7.11	9.00	7.50
Std	3.32	2.03	3.32	2.03	3.32	1.15	3.32	2.03

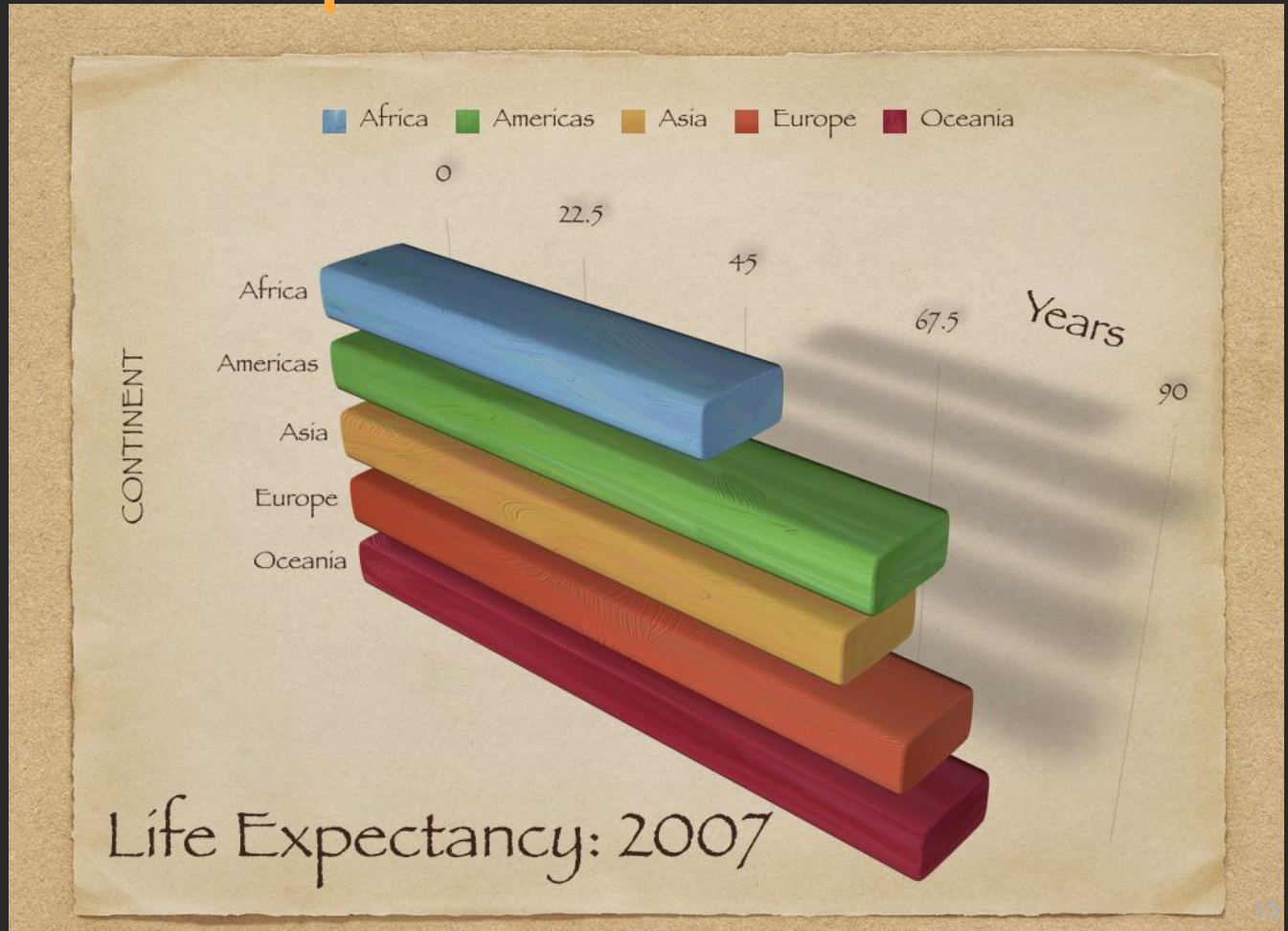

A
B
C
D
Y
X
0
2
4
6
8
10
12
14
16
18
20



X Mean: 54.2659224
Y Mean: 47.8313999
X SD : 16.7649829
Y SD : 26.9342120
Corr. : -0.0642526

What could be improved here?

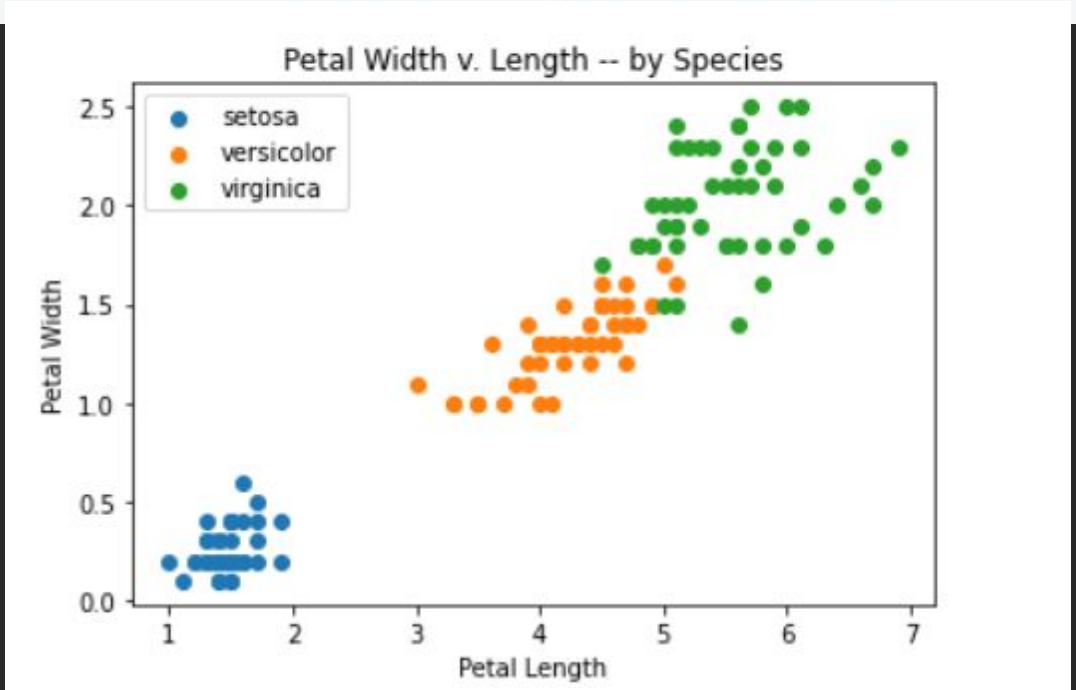
Enter a text response in the iClicker



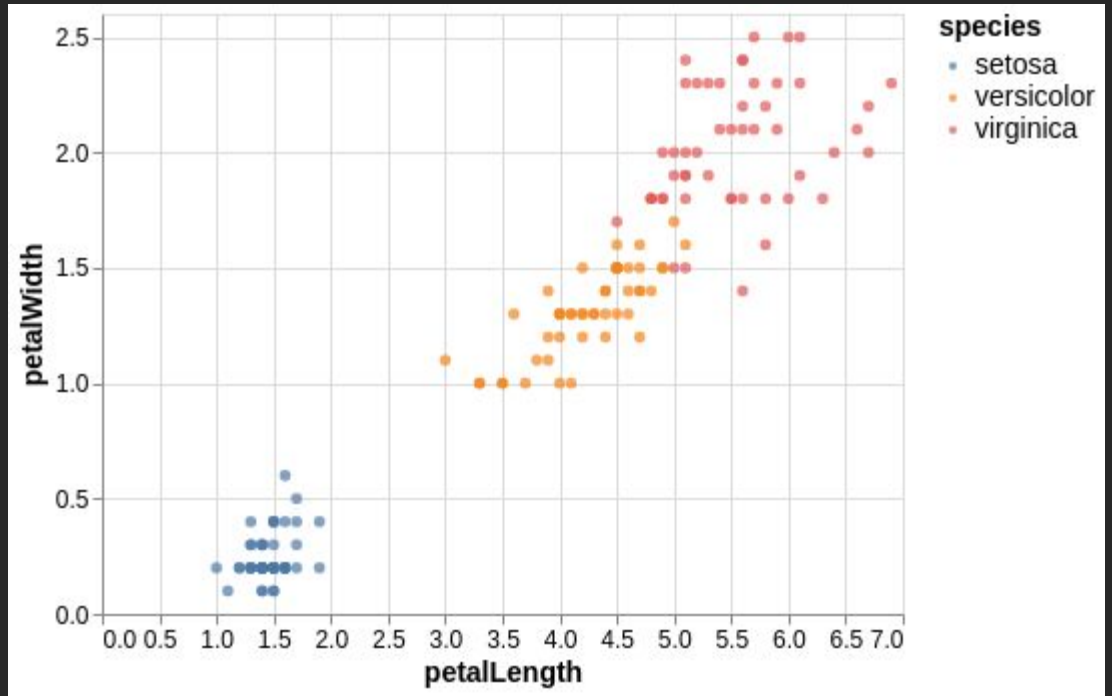
What does this code do?

```
fig, ax = plt.subplots()
colors=['#1f77b4', '#ff7f0e', '#2ca02c']

for color, species in zip(colors, df.species.unique()):
    tmp = df[df.species == species]
    ax.scatter(tmp.petalLength, tmp.petalWidth,
               label=species, color=color)
```



```
alt.Chart(df).mark_circle().encode(
    x='petalLength',
    y='petalWidth',
    color='species'
)
```



High level

Focus on the data

```
alt.Chart(df).mark_circle().encode(
    x='petalLength',
    y='petalWidth',
    color='species'
)
```

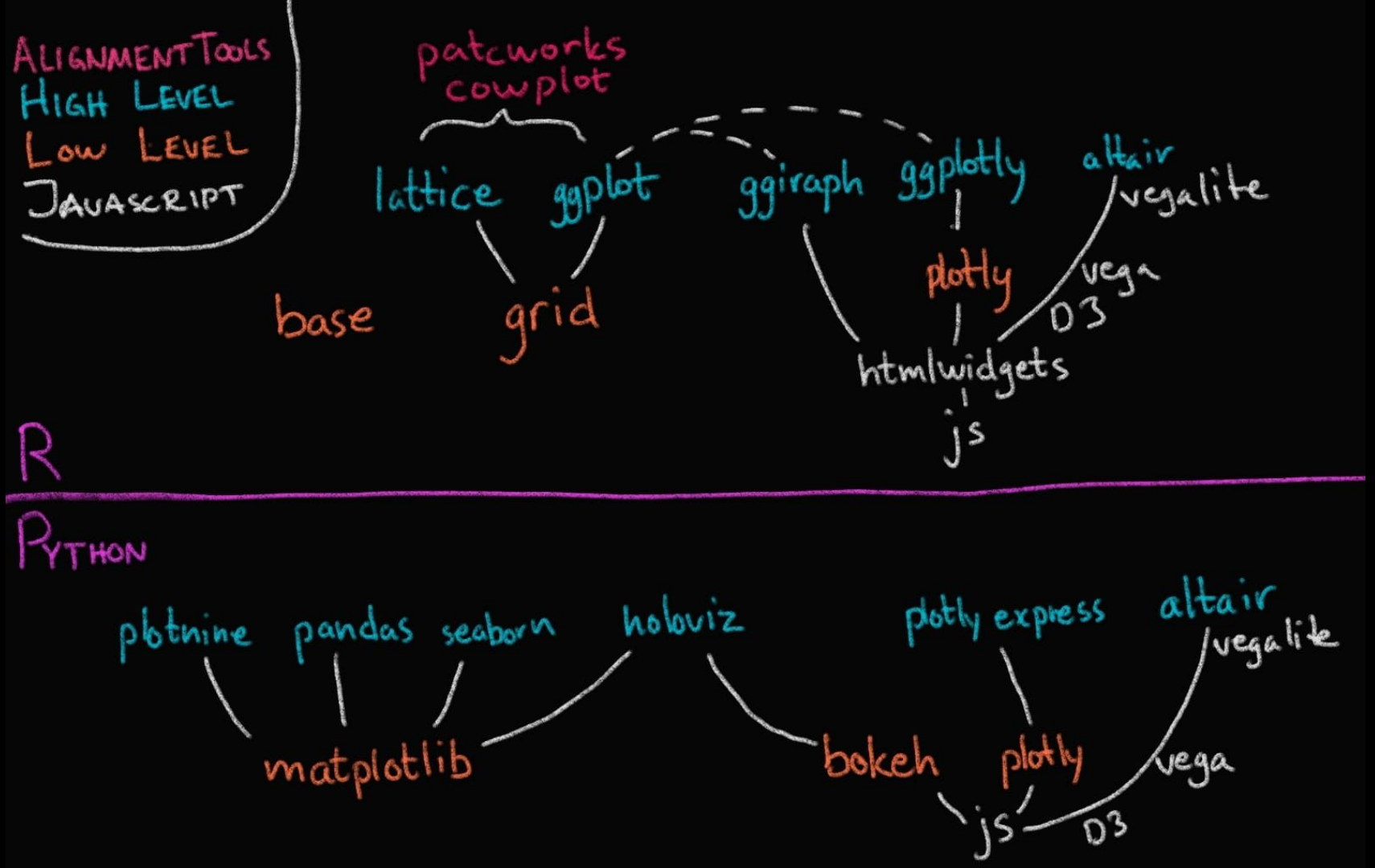
Low level

Focus on graphics details

```
fig, ax = plt.subplots()
colors=['#1f77b4', '#ff7f0e', '#2ca02c']

for color, species in zip(colors, df.species.unique()):
    tmp = df[df.species == species]
    ax.scatter(tmp.petalLength, tmp.petalWidth,
               label=species, color=color)
```

Application specific


AlignmentTools
High Level
Low Level
JavaScript
patchworks
cowplot
lattice
ggplot
ggiraph
ggplotly
altair
vegalite
plotly
vega
D3
base
grid
htmlwidgets
js
R
Python
plotnine
pandas
seaborn
holoviz
plotly express
altair
vegalite
matplotlib
bokeh
plotly
vega
js
D3

Have you used any of these visualization tools before?

A. Altair
B. ggplot
C. Something else in Py/R
D. Excel/PowerBI/Tableau etc
E. None of them

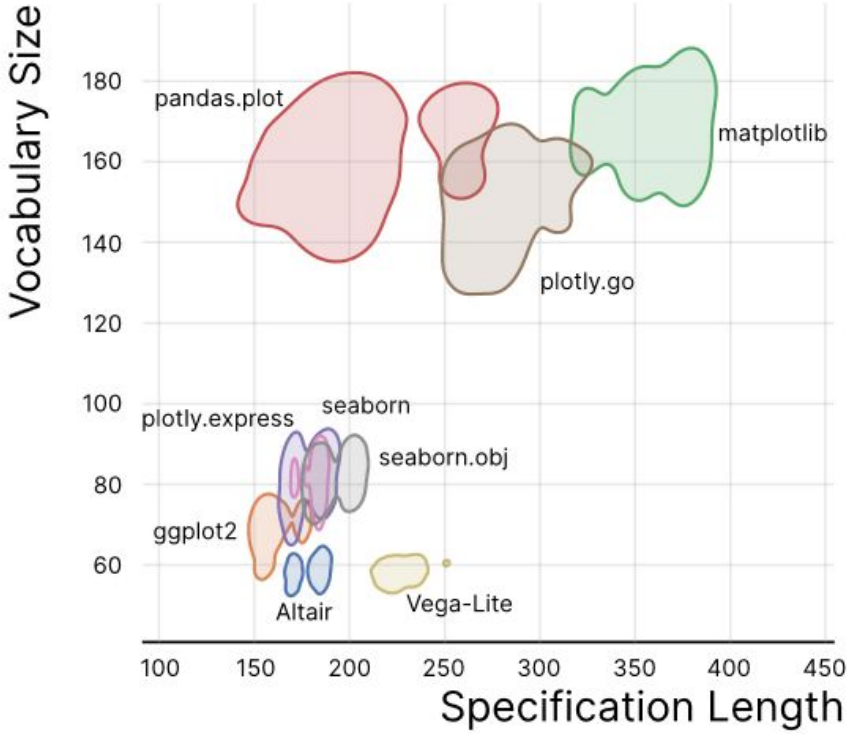
Grammar of Graphics

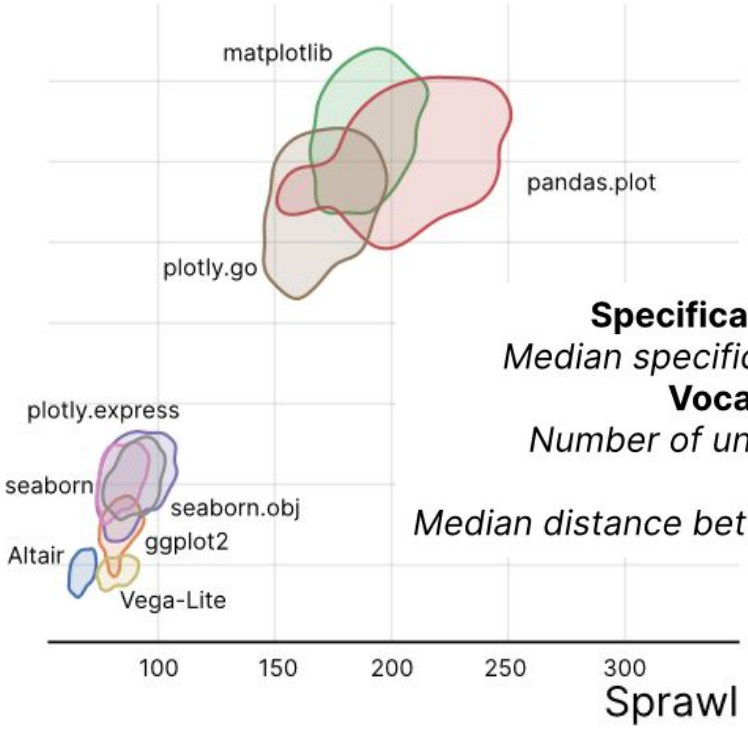
- Create a canvas/chart
- Encode visual aesthetics
- Add geometric marks

```
ggplot(data, aes(x, y)) + geom()
Chart(data).mark().encode(x, y)
```

Metrics-Based Evaluation and Comparison of Visualization Notations

Nicolas Kruchten, Andrew M. McNutt, and Michael J. McGuffin





Specification Length
Median specification bytes
Vocabulary Size
Number of unique tokens
Sprawl
Median distance between specs