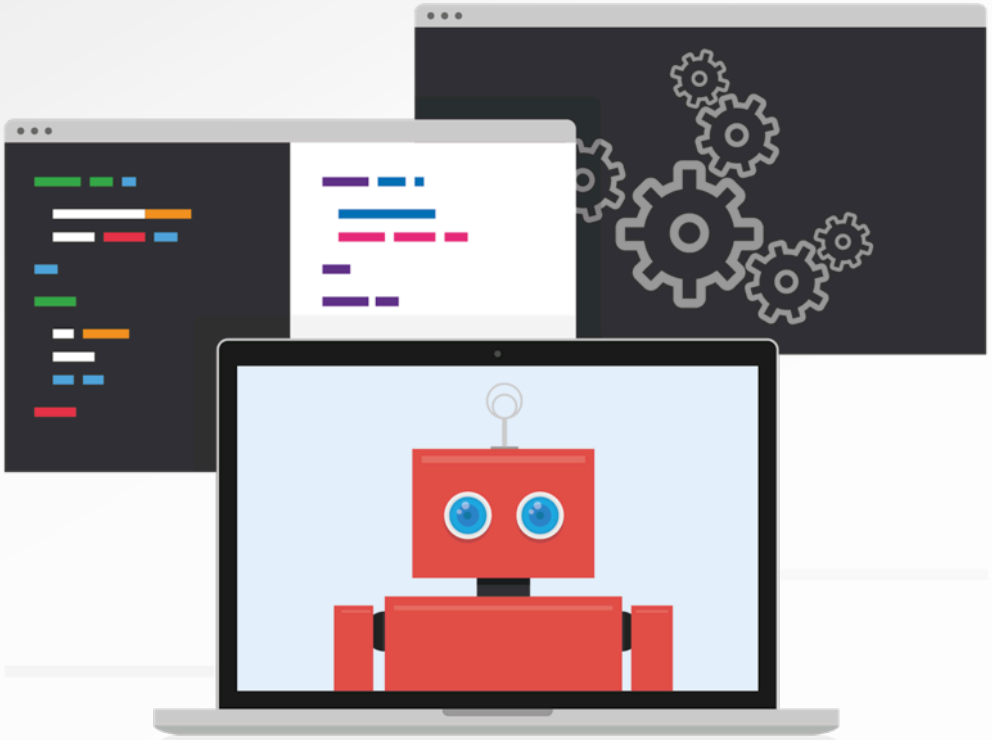
Cambridge Coding Academy

Exploratory data analysis & interactive figures with Plotly



cambridgecoding.com

@cambridgecoding

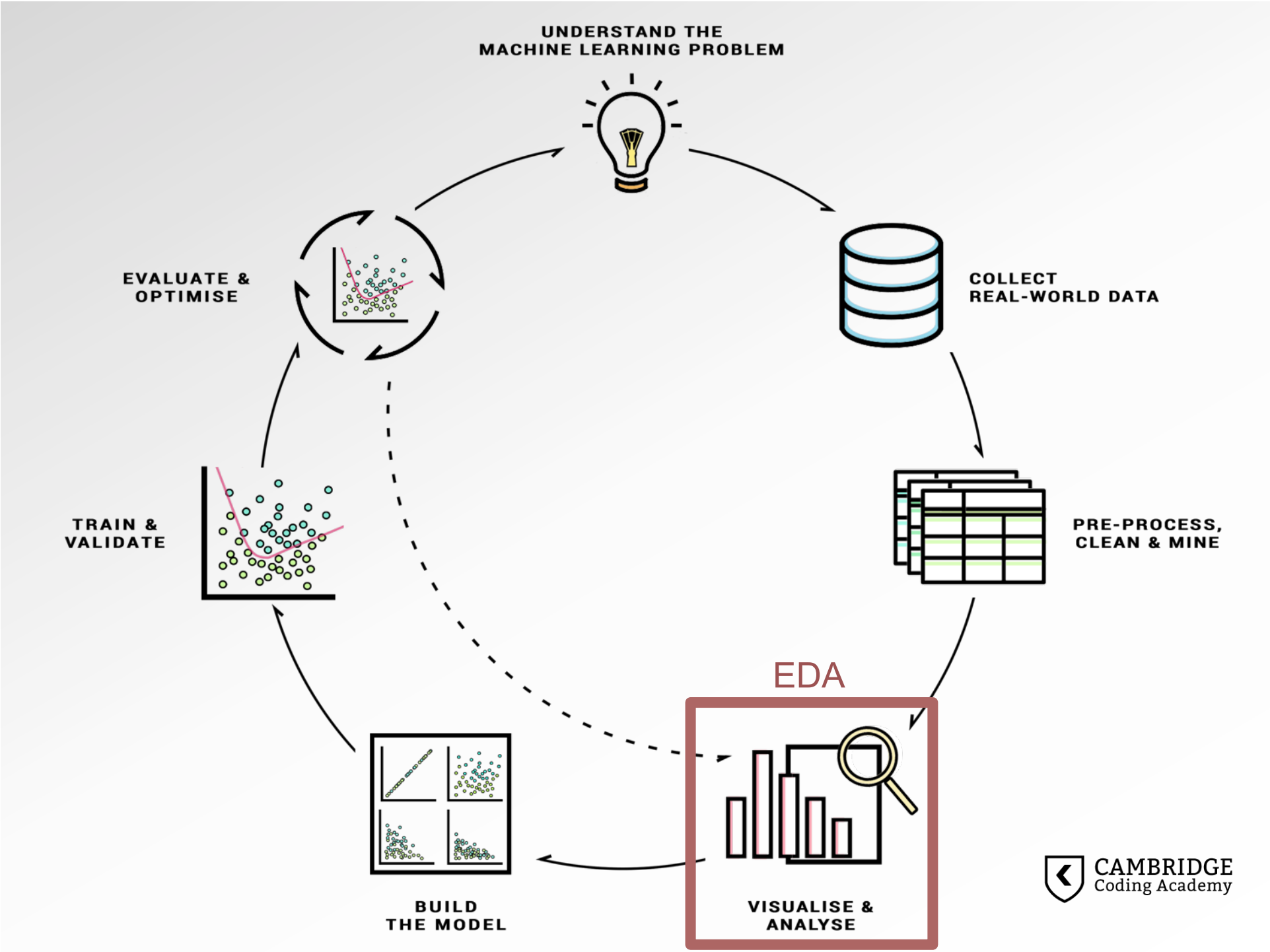
facebook.com/cambridgecoding

CAMBRIDGE
Coding Academy

Agenda

- EDA - When to use it and why?
- The dataset and our tool belt
- Basics in plotting and data summarizing
- Hands-on coding session

CAMBRIDGE
Coding Academy


UNDERSTAND THE
MACHINE LEARNING PROBLEM
COLLECT
REAL-WORLD DATA
PRE-PROCESS,
CLEAN & MINE
EDA
VISUALISE &
ANALYSE
BUILD
THE MODEL
TRAIN &
VALIDATE
EVALUATE &
OPTIMISE
CAMBRIDGE
Coding Academy

Online Platform

http://online.cambridgecoding.com/

Interested in a two-day data science bootcamp using Python?

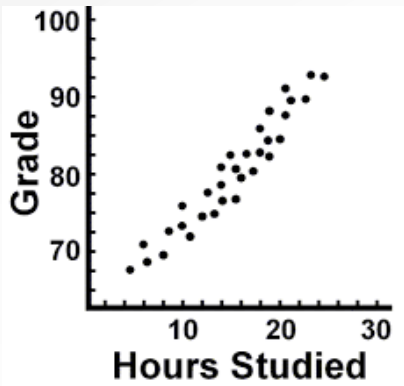
Check out our upcoming bootcamp on 02-03 July in London

http://cambridgecoding.com/datascience-bootcamp

CAMBRIDGE
Coding Academy

What is EDA?

Friend	Number of hours of studying per week	Grade Point Average (out of 5.0)
Allie	14	3.91
Samantha	42	4.98
Hayley	10	3.22
Jessica	32	4.81
Megan	5	2.0
Rachel	10	2.82
Briley	25	3.79
Lauren	18	3.48



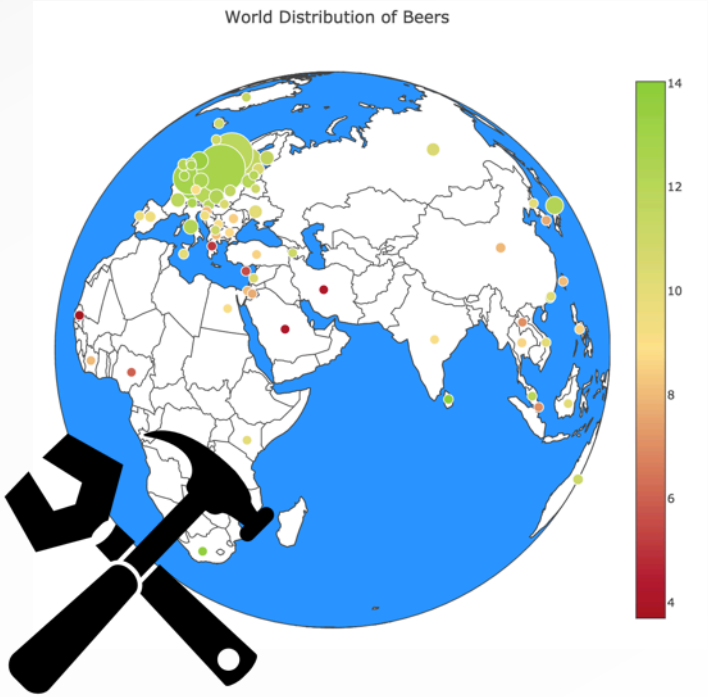
- Mapping data to visual object to make it relatable and understandable
- Goes hand in hand with statistics and machine learning
- Communicate your decision to somebody efficiently

CAMBRIDGE Coding Academy

The Dataset

What country in the world produces the most and the best beer?

http://www.cambridgebeerfestival.com/



CAMBRIDGE Coding Academy

Our EDA setup

Python

Tool belt

plotting

data processing

Coding environment

CAMBRIDGE Coding Academy

What is Plotly

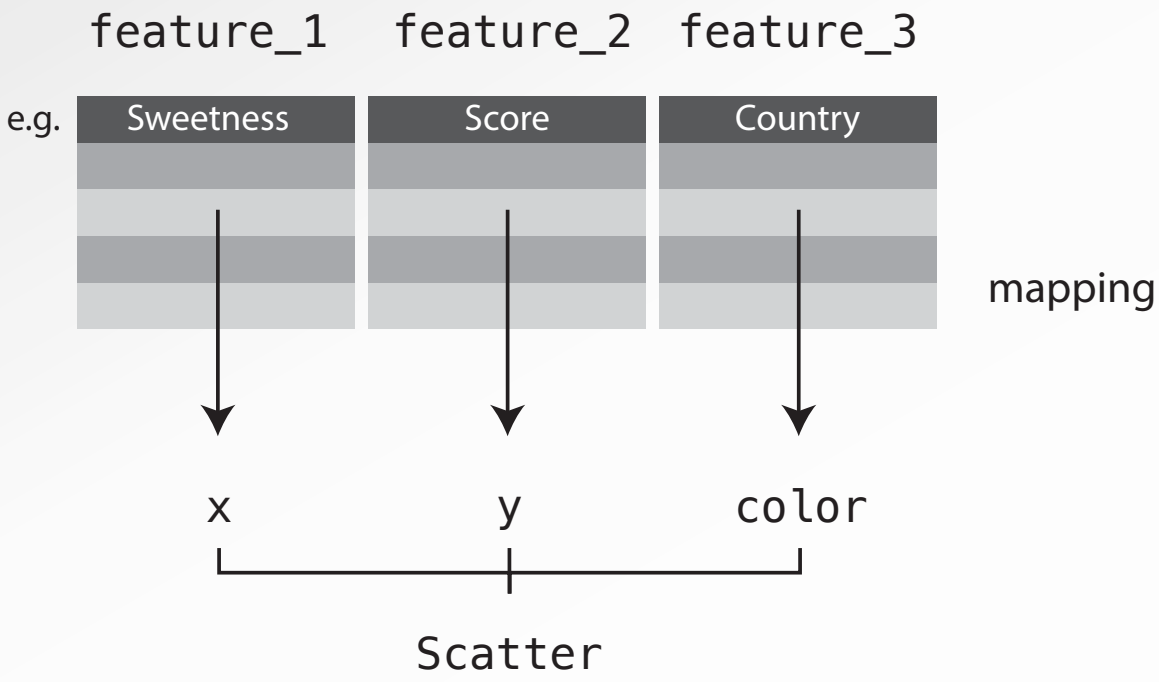
https://plot.ly/

- Produce interactive graphics
- Share with plots with people (if registered)
- Integrates with other frameworks (R, Python…)
- Easy to pick up
- Many different plots types available

CAMBRIDGE
Coding Academy

Plotting basics

1. collect data



3. map to aesthetics

2. pick a geometry

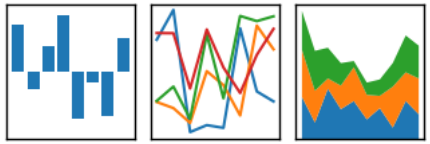
```
trace = go.Scatter(x = feature_1 , y = feature_2, color = feature_3)
```

CAMBRIDGE Coding Academy

What is Pandas

pandas

$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$



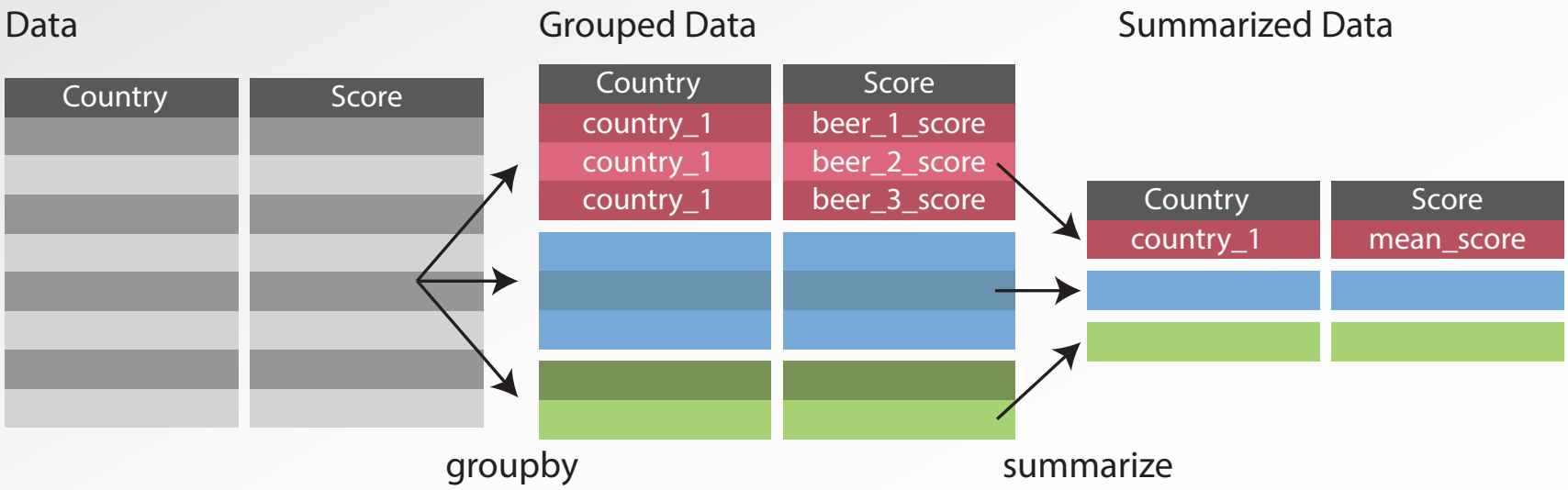
http://pandas.pydata.org/

- Convenient data analysis in python
- Easy to read in data
- Reshaping of and preprocessing data
- Filtering data
- Calculate summary statistics

CAMBRIDGE
Coding Academy

Basics of data processing

- Break process up into steps

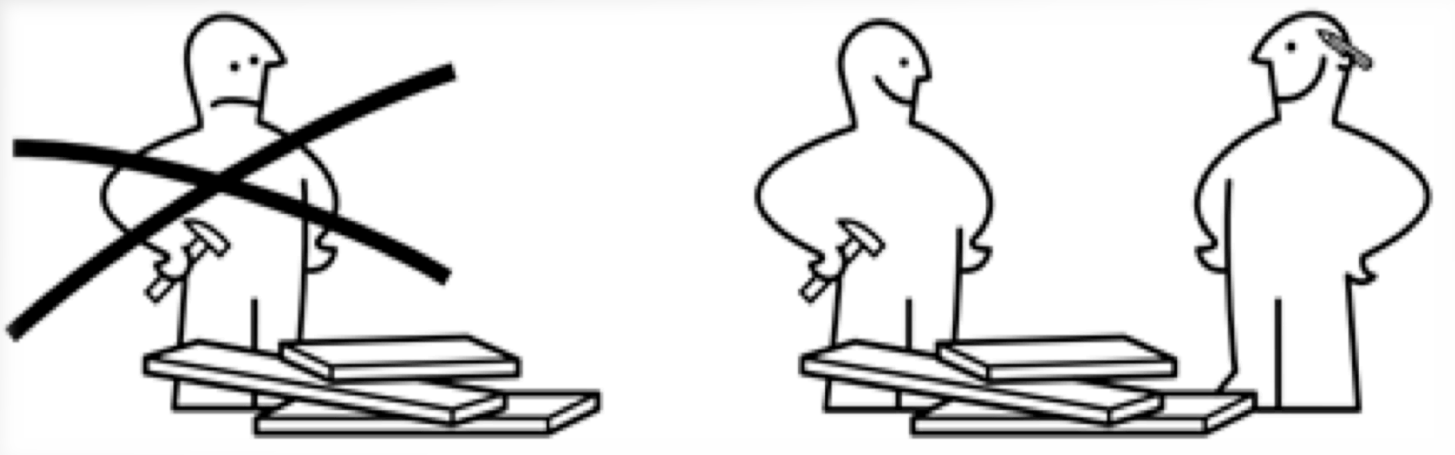



```
dataset['Score'].groupby(dataset['Country']).mean()
```

CAMBRIDGE Coding Academy

Hands-on session

Pair Programming



- Reduces risk, bug fixing and overall working time
- Shared knowledge, discussion and QA

CAMBRIDGE
Coding Academy

Change code and see what happens.

- In the interest of time the code is all there for you
  - run it
  - change it
  - break it
  - learn from it
- Notebook is standalone - go at your own pace!
- Take the notebook home and run your own data

CAMBRIDGE
Coding Academy

Using Jupyter

- Download the workshop from: https://github.com/cnjr2/eda_tech_talk
- Start the notebook:

```
Charless-MacBook-Pro:~ charles$ jupyter notebook eda_tech_talk.ipynb
```

- Double-click a cell to edit its contents (its border will become green)
- Execute the code in a cell by pressing 'ctrl+ enter' or press button.
- You can add cells with plus + button
- Put short chunks of code into each cell to go step-by-step
- Inactivate a line of code preceding it with a #; the re-run the chunk to see what changes!

CAMBRIDGE Coding Academy