STAT 135
3. EDA

Spring 2022

Lecturer: Dr Rebecca Barter (*she/her*)
Office hours: Tu 9:30-10:30, Th 1:00-2:00
Office: Evans 339

Email: rebeccabarter@berkeley.edu
Twitter: @rlbarter
GitHub: rlbarter

Exploratory Data Analysis (EDA)

and

Explanatory Data Analysis

Only by looking at your data can you really understand it

PHOTO: DAN SAELINGER

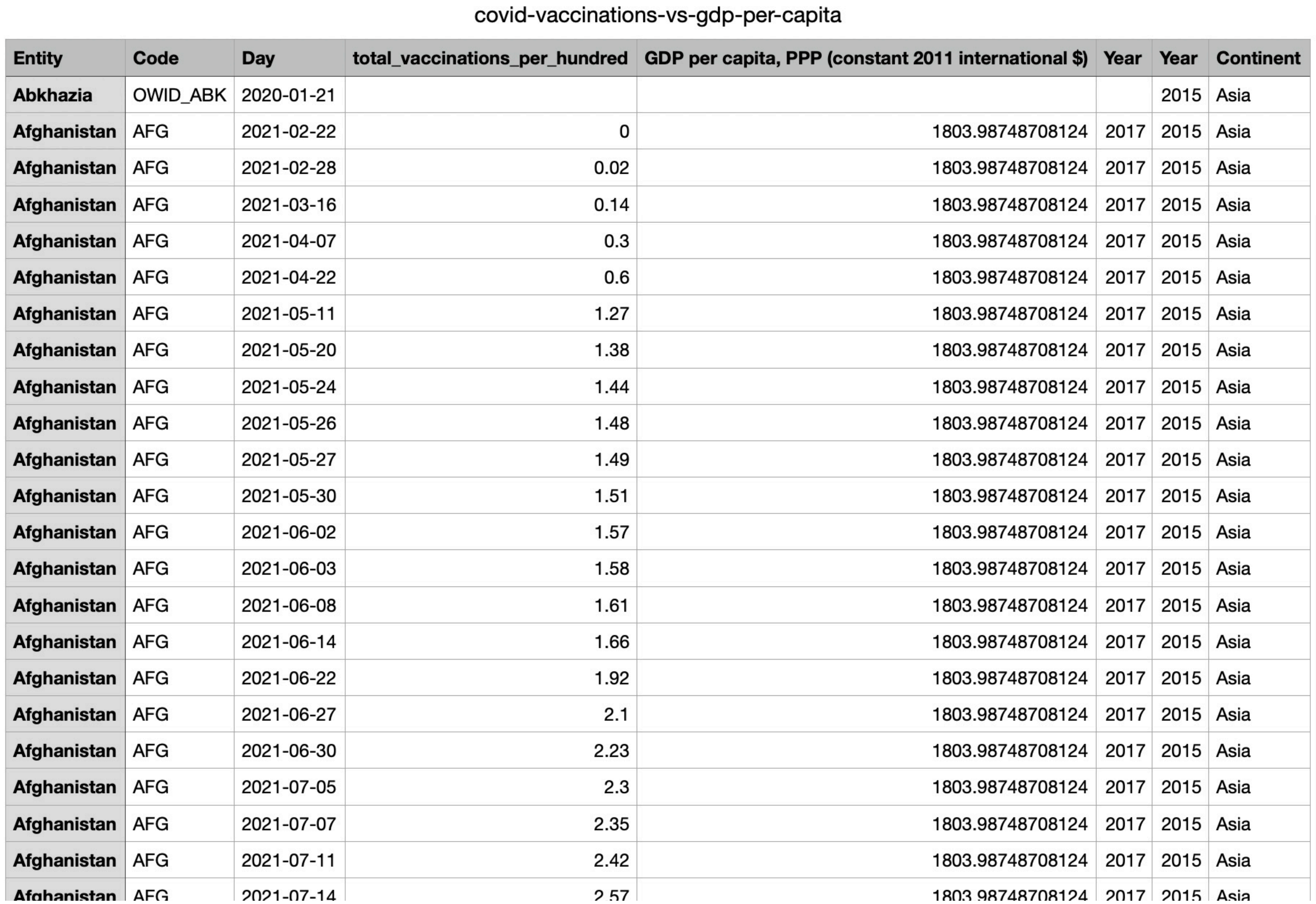
COVID-19 vaccine data example

Data collected from:

https://ourworldindata.org/grapher/covid-vaccination-doses-per-capita

Daily data for every country on:

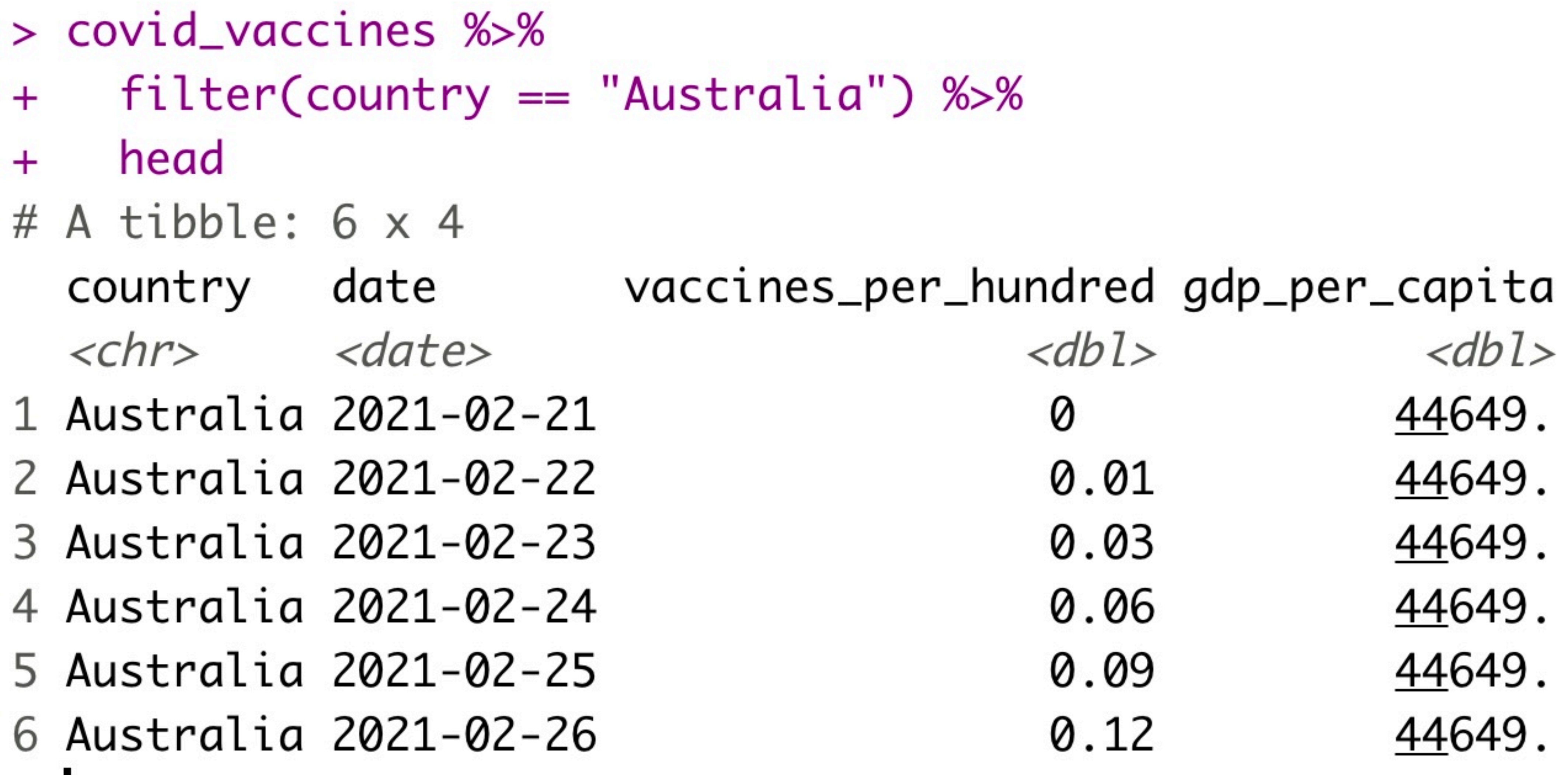
- COVID Vaccine rates (per hundred)
- GDP per capita



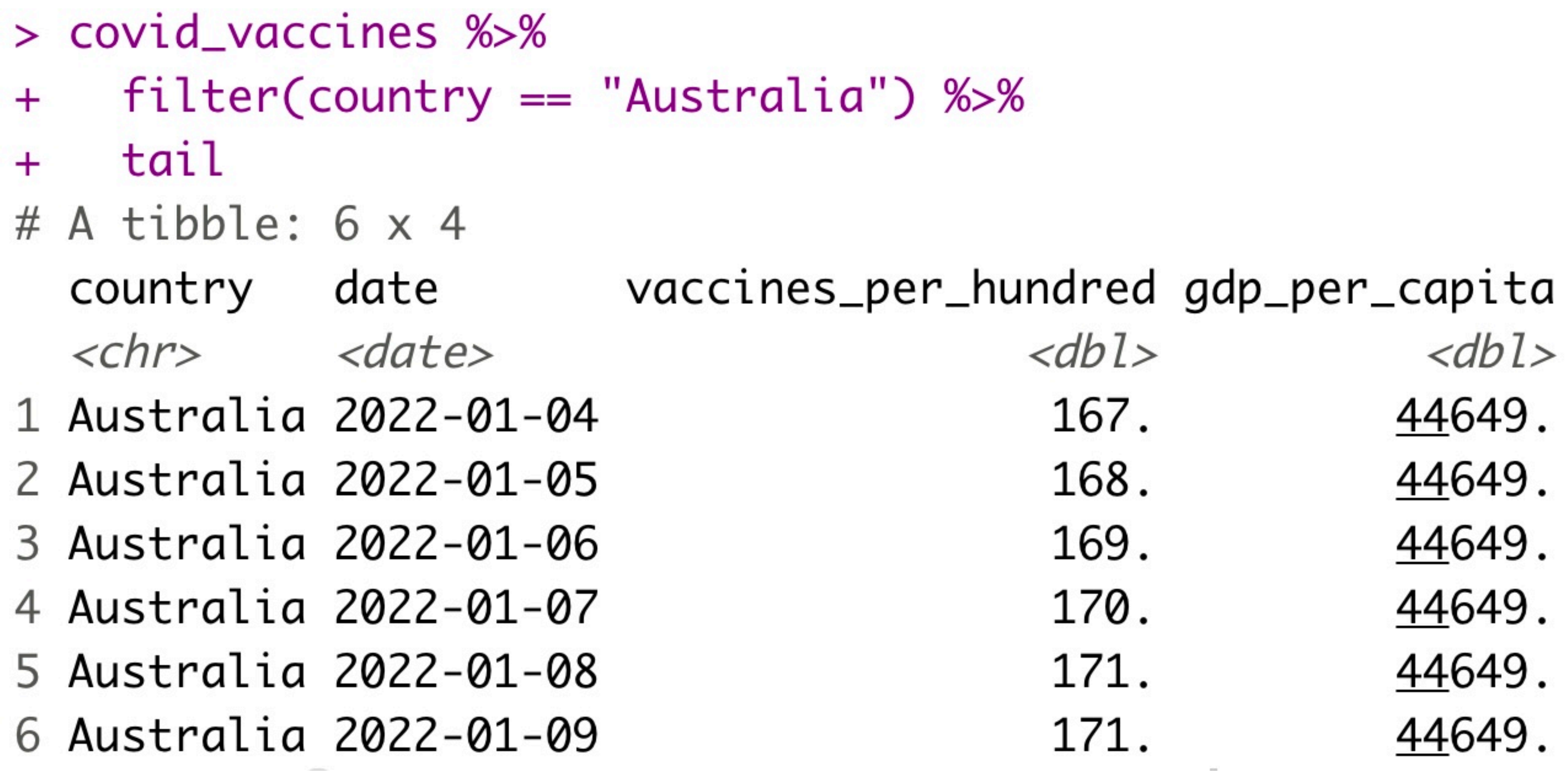
covid-vaccinations-vs-gdp-per-capita

Entity	Code	Day	total_vaccinations_per_hundred	GDP per capita, PPP (constant 2011 international $)	Year	Year	Continent
Abkhazia	OWID_ABK	2020-01-21				2015	Asia
Afghanistan	AFG	2021-02-22	0	1803.98748708124	2017	2015	Asia
Afghanistan	AFG	2021-02-28	0.02	1803.98748708124	2017	2015	Asia
Afghanistan	AFG	2021-03-16	0.14	1803.98748708124	2017	2015	Asia
Afghanistan	AFG	2021-04-07	0.3	1803.98748708124	2017	2015	Asia
Afghanistan	AFG	2021-04-22	0.6	1803.98748708124	2017	2015	Asia
Afghanistan	AFG	2021-05-11	1.27	1803.98748708124	2017	2015	Asia
Afghanistan	AFG	2021-05-20	1.38	1803.98748708124	2017	2015	Asia
Afghanistan	AFG	2021-05-24	1.44	1803.98748708124	2017	2015	Asia
Afghanistan	AFG	2021-05-26	1.48	1803.98748708124	2017	2015	Asia
Afghanistan	AFG	2021-05-27	1.49	1803.98748708124	2017	2015	Asia
Afghanistan	AFG	2021-05-30	1.51	1803.98748708124	2017	2015	Asia
Afghanistan	AFG	2021-06-02	1.57	1803.98748708124	2017	2015	Asia
Afghanistan	AFG	2021-06-03	1.58	1803.98748708124	2017	2015	Asia
Afghanistan	AFG	2021-06-08	1.61	1803.98748708124	2017	2015	Asia
Afghanistan	AFG	2021-06-14	1.66	1803.98748708124	2017	2015	Asia
Afghanistan	AFG	2021-06-22	1.92	1803.98748708124	2017	2015	Asia
Afghanistan	AFG	2021-06-27	2.1	1803.98748708124	2017	2015	Asia
Afghanistan	AFG	2021-06-30	2.23	1803.98748708124	2017	2015	Asia
Afghanistan	AFG	2021-07-05	2.3	1803.98748708124	2017	2015	Asia
Afghanistan	AFG	2021-07-07	2.35	1803.98748708124	2017	2015	Asia
Afghanistan	AFG	2021-07-11	2.42	1803.98748708124	2017	2015	Asia
Afghanistan	AFG	2021-07-14	2.57	1803.98748708124	2017	2015	Asia

COVID-19 vaccine data example



```
> covid_vaccines %>%
+   filter(country == "Australia") %>%
+   head
# A tibble: 6 x 4
  country   date       vaccines_per_hundred gdp_per_capita
  <chr>     <date>                    <dbl>          <dbl>
1 Australia 2021-02-21                 0             44649.
2 Australia 2021-02-22                 0.01          44649.
3 Australia 2021-02-23                 0.03          44649.
4 Australia 2021-02-24                 0.06          44649.
5 Australia 2021-02-25                 0.09          44649.
6 Australia 2021-02-26                 0.12          44649.
```

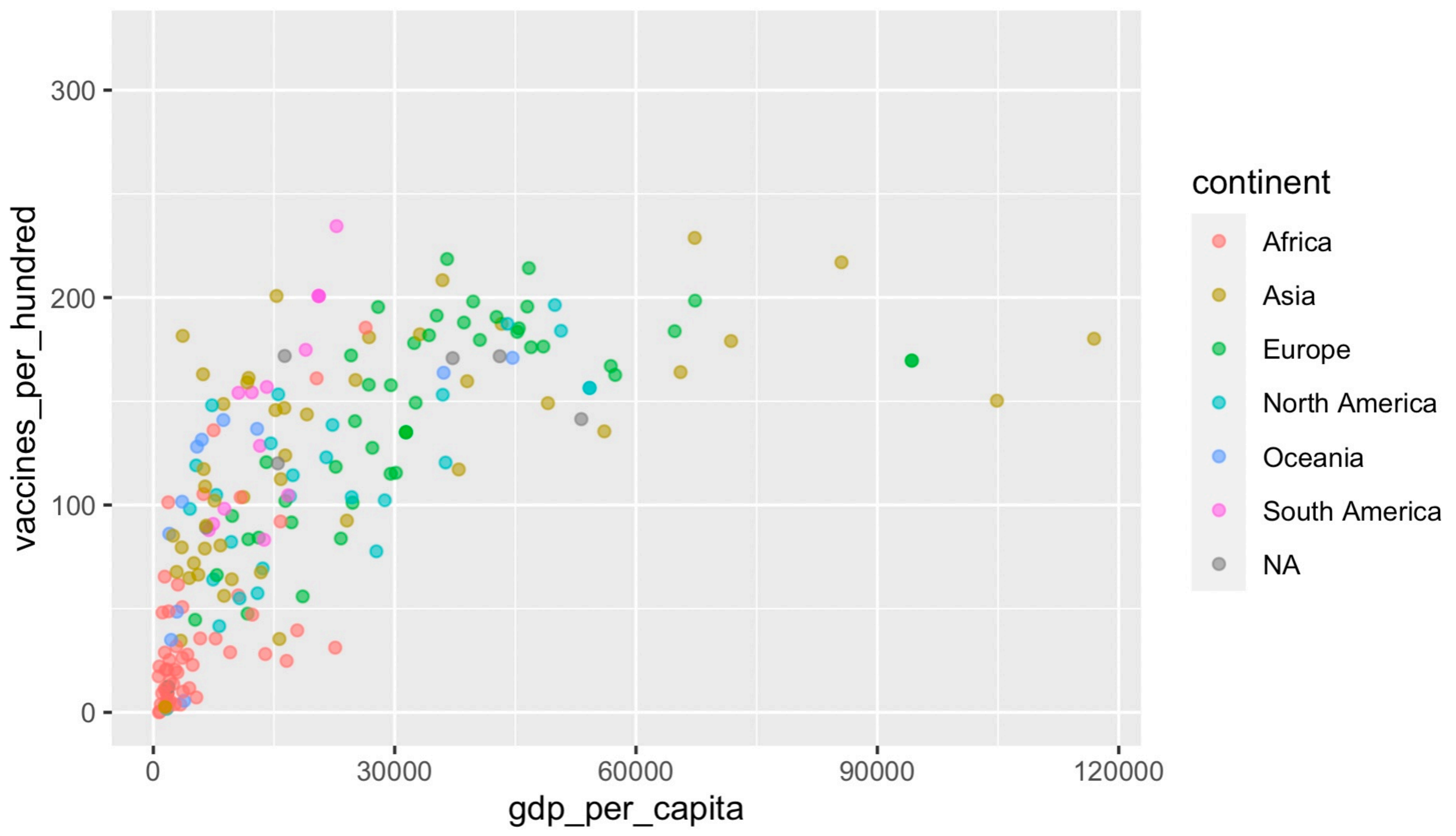


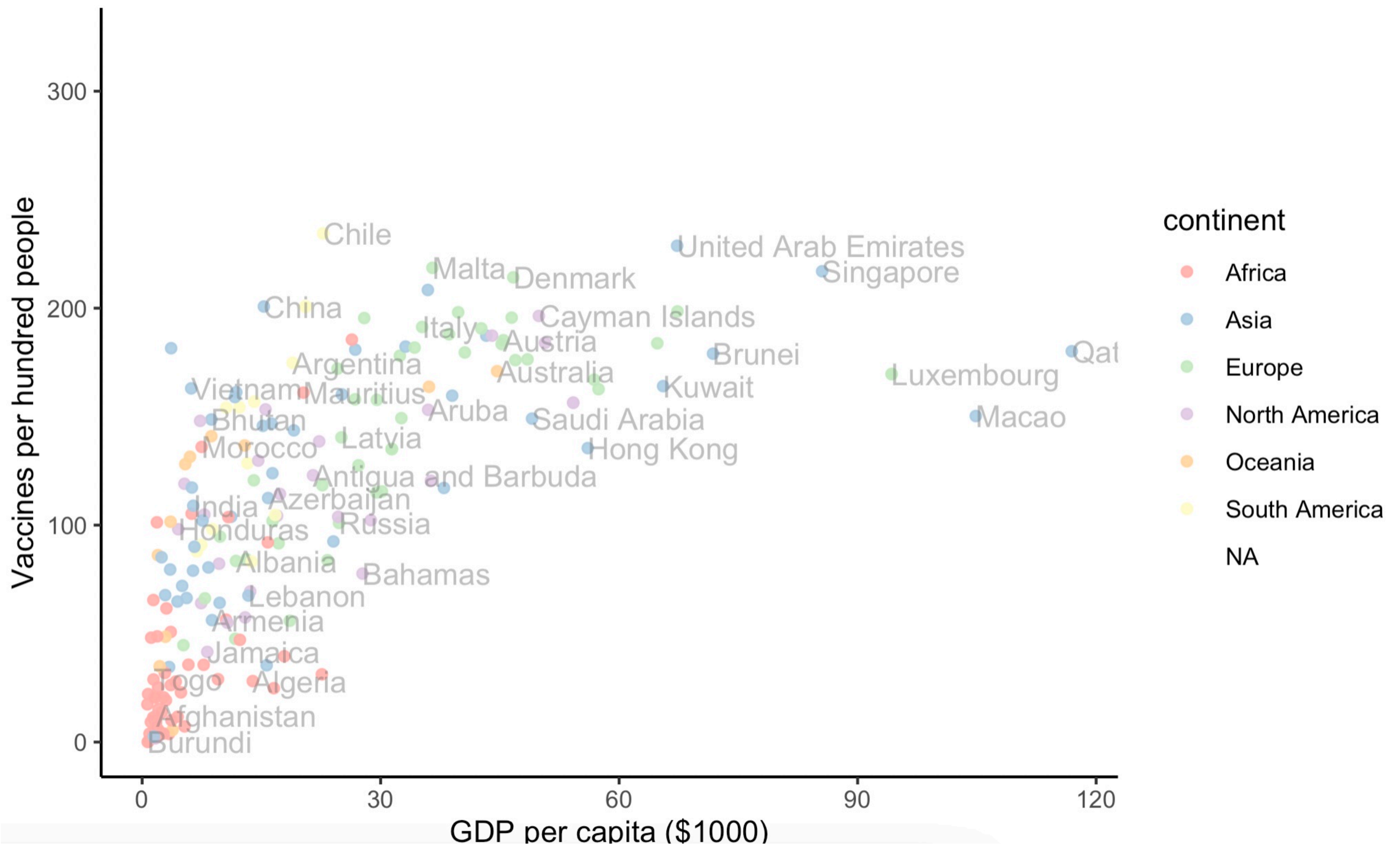
```
> covid_vaccines %>%
+   filter(country == "Australia") %>%
+   tail
# A tibble: 6 x 4
  country   date       vaccines_per_hundred gdp_per_capita
  <chr>     <date>                    <dbl>          <dbl>
1 Australia 2022-01-04                 167.          44649.
2 Australia 2022-01-05                 168.          44649.
3 Australia 2022-01-06                 169.          44649.
4 Australia 2022-01-07                 170.          44649.
5 Australia 2022-01-08                 171.          44649.
6 Australia 2022-01-09                 171.          44649.
```

What questions do you have about what the numbers in this dataset mean?

1. Are the vaccines_per_hundred column the daily number of vaccines given, or are they the cumulative?
2. Why are the vaccines per hundred *greater than 100?*

COVID-19 vaccine data example





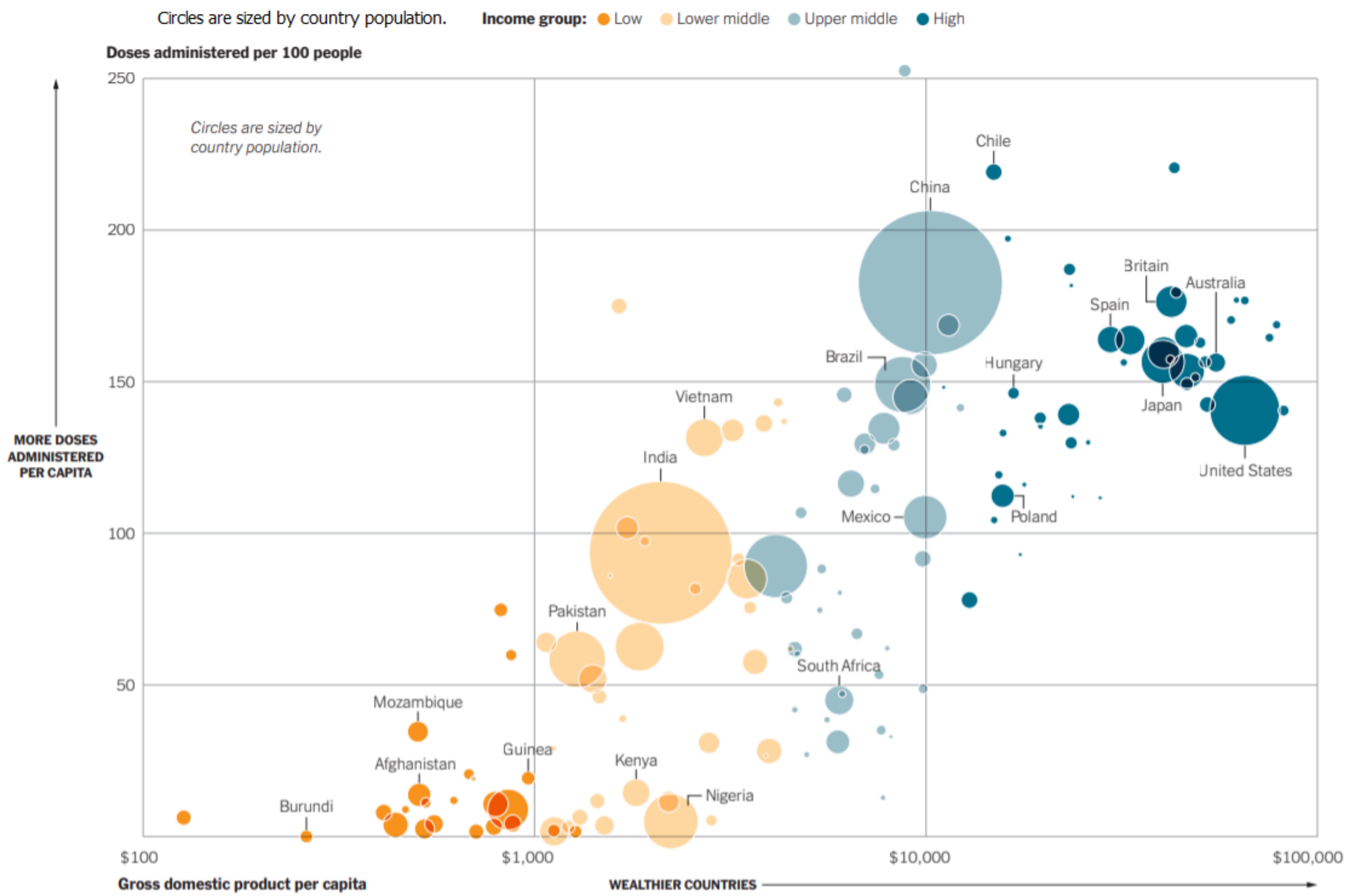
NYTimes version: data viz goals

Questions:

1. What do you take away from this figure?

2. What questions do you have about this figure?

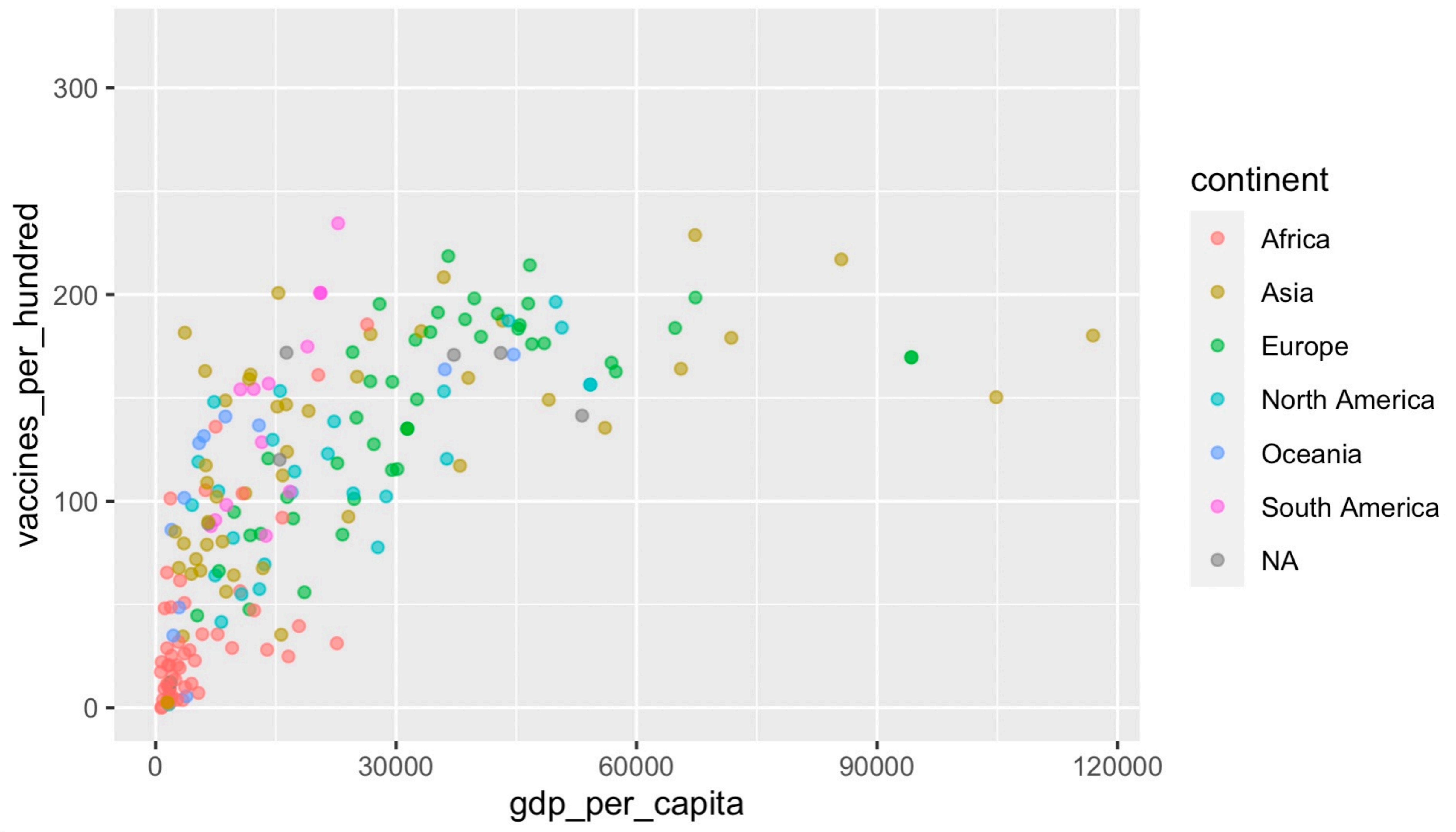
https://www.nytimes.com/2022/01/06/learning/whats-going-on-in-this-graph-jan-12-2022.html

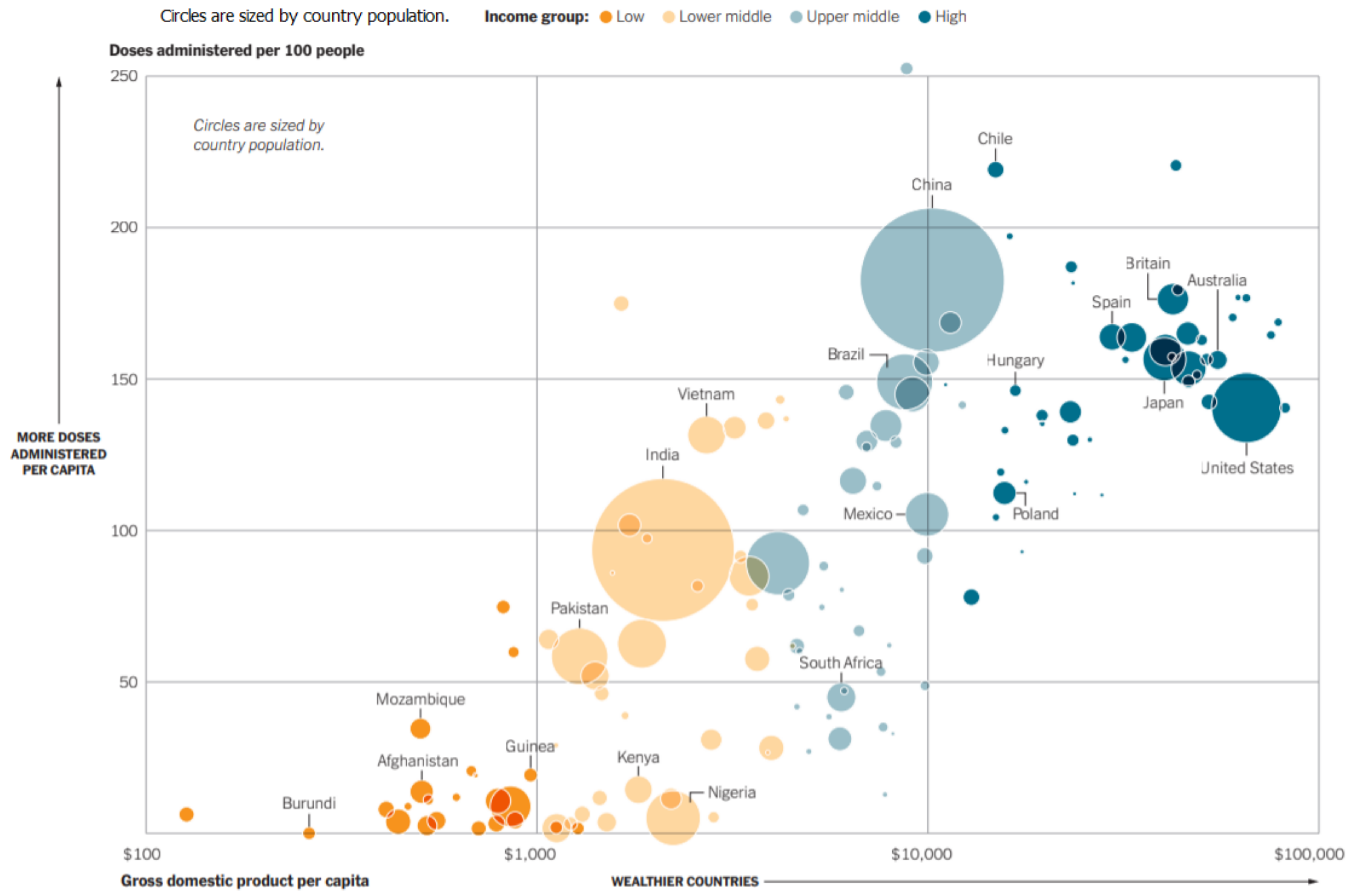


Sources: Vaccination data from local governments via Our World in Data; income classifications and gross domestic product data from the World Bank. | Note: Data is as of Dec. 8.

Glossary: G.D.P. per capita is the Gross Domestic Product, or wealth of a country divided by its population size.

Exploratory vs Explanatory





Sources: Vaccination data from local governments via Our World in Data; income classifications and gross domestic product data from the World Bank. | Note: Data is as of Dec. 8.

Glossary: G.D.P. per capita is the Gross Domestic Product, or wealth of a country divided by its population size.

Goal: for **you** to understand what the data looks like

Products: Lots of quick, messy plots

Goal: for **others** to understand what the data looks like

Products: A few highly polished figures that tell a clear story

Tips for producing impactful explanatory figures

1. Use color sparingly. Try to never use more than ~5 colors in a single figure

2. Avoid using red and green in the same plot (be mindful of color blindnes

3. Use size and color to guide the audiences attention

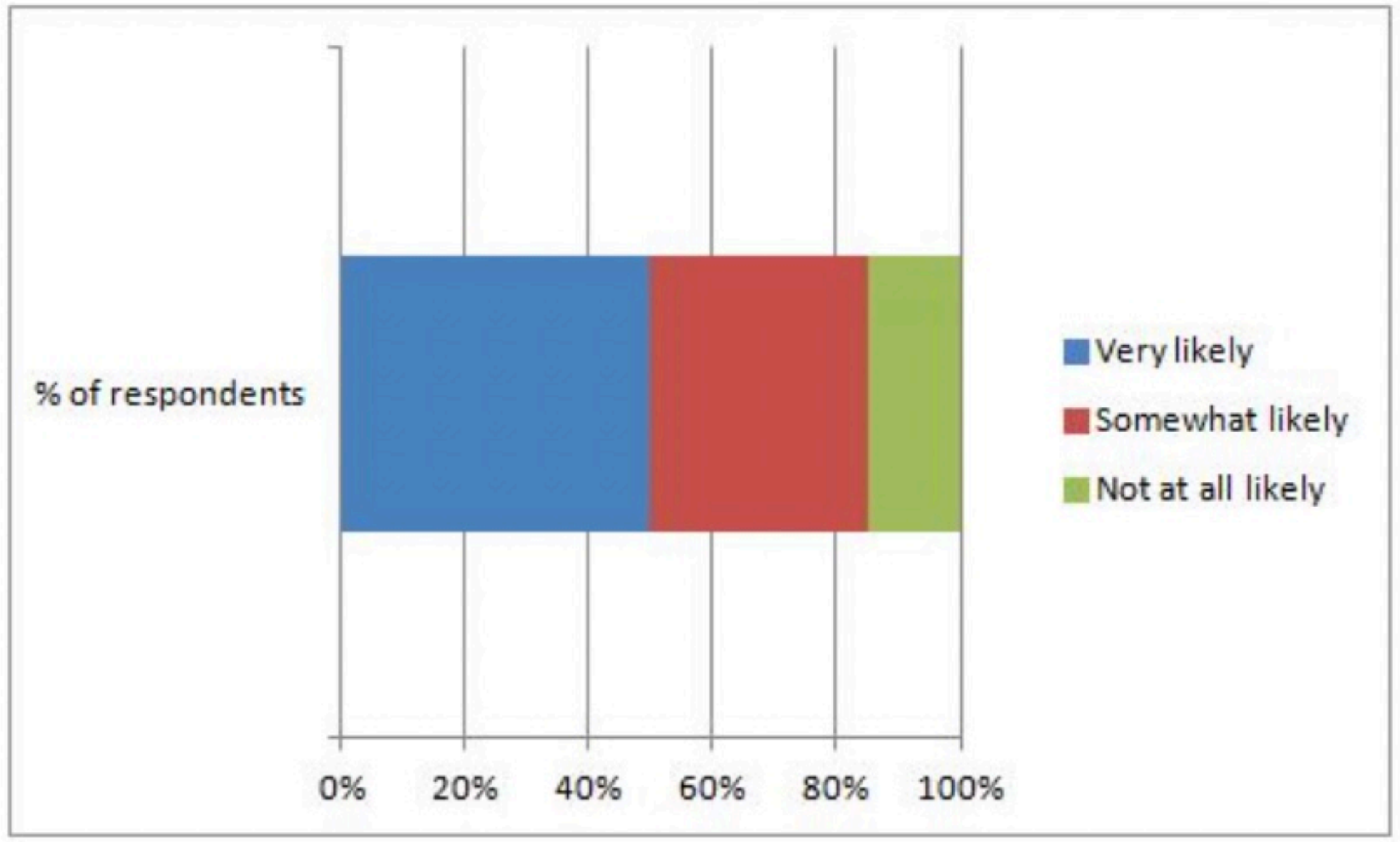
4. Use transparency to reduce over-plotting

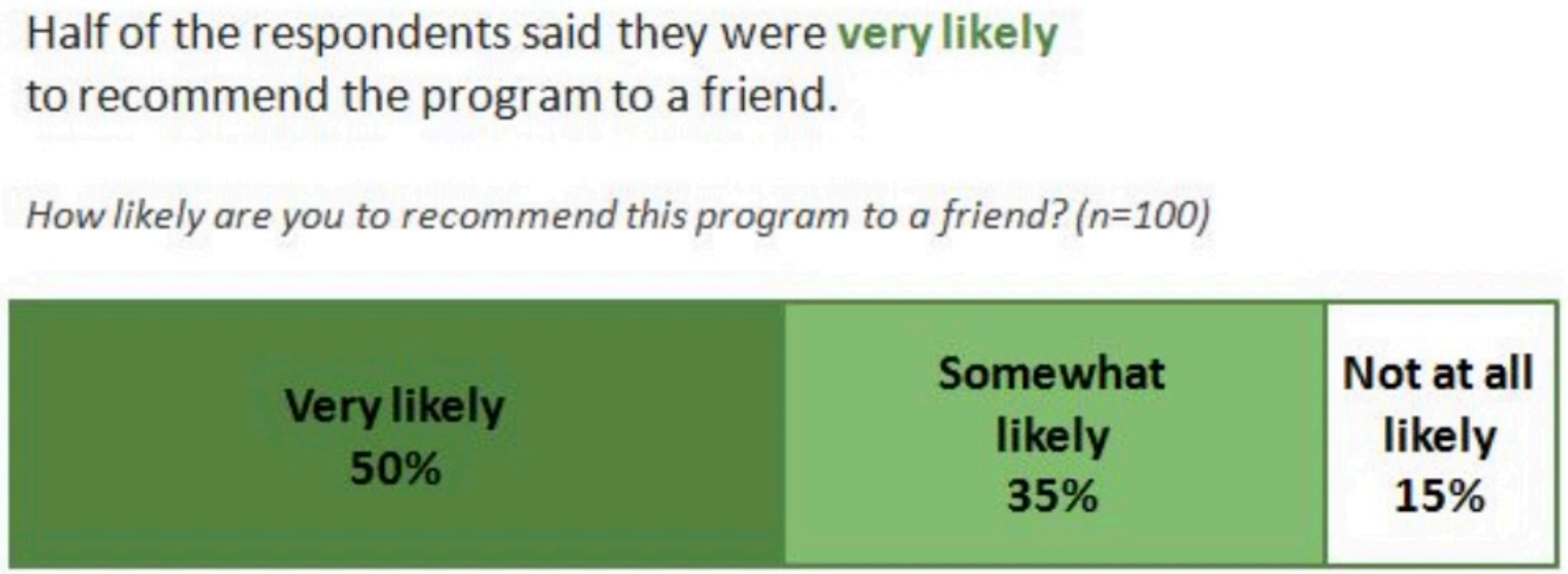
5. Use text annotation (but not excessively)

Color suggestions

Use sequential color schemes for sequential data

Direct annotation



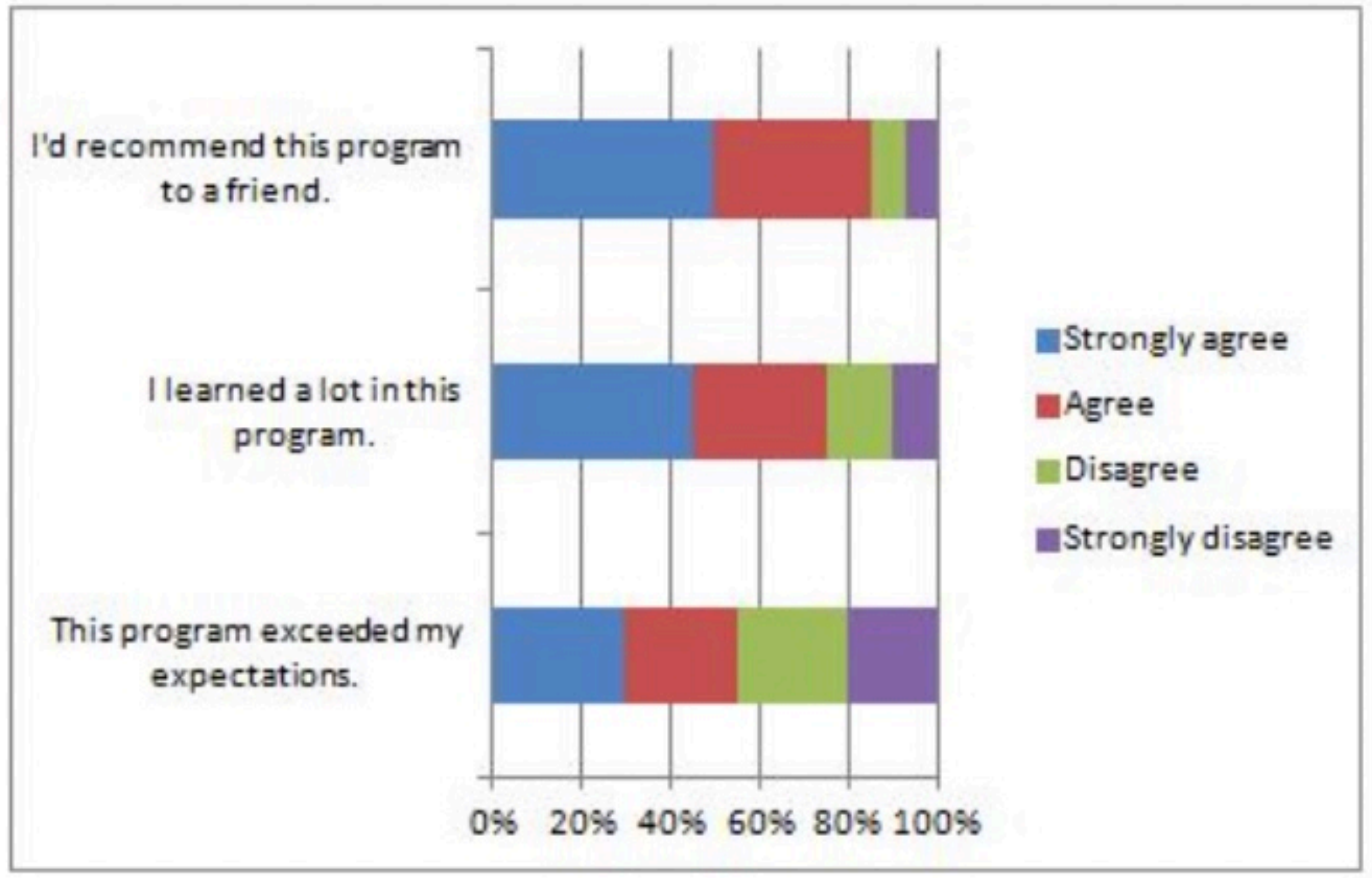

Half of the respondents said they were **very likely** to recommend the program to a friend.

How likely are you to recommend this program to a friend? (n=100)

Color suggestions

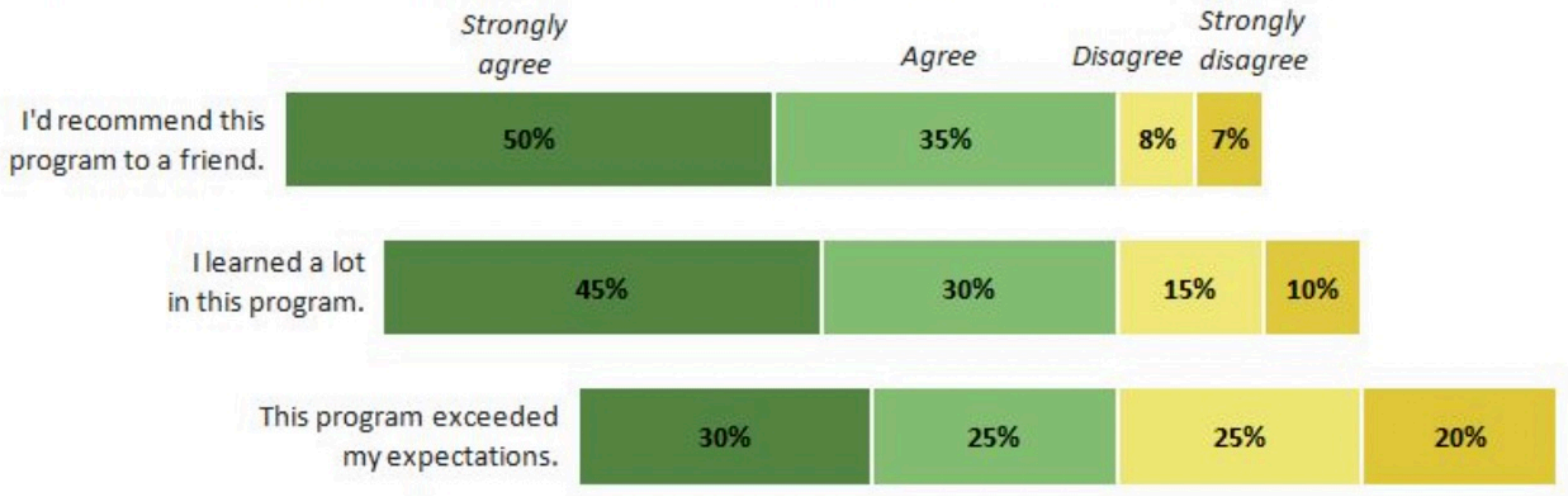
Use diverging color schemes for diverging data

Direct annotation

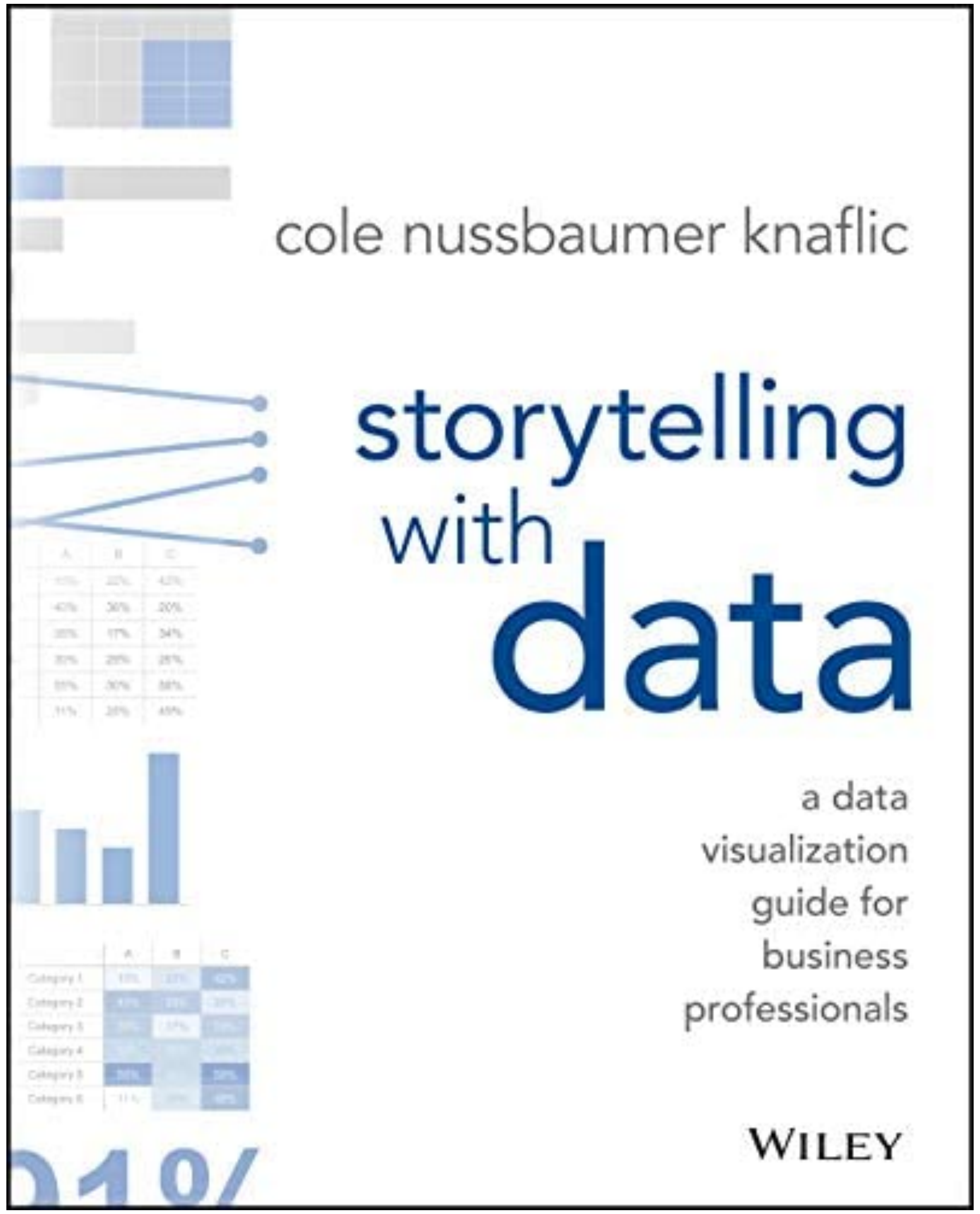


A majority of participants **would recommend this program to a friend** (85% strongly agreed or agreed). Three-quarters (75%) strongly agreed or agreed that they **learned a lot in this program,** but only half (55%) felt the program **exceeded their expectations.**

Please indicate how much you agree or disagree with the following statements. (n=100)



Tips for producing impactful explanatory figures



cole nussbaumer knaflic

storytelling with data

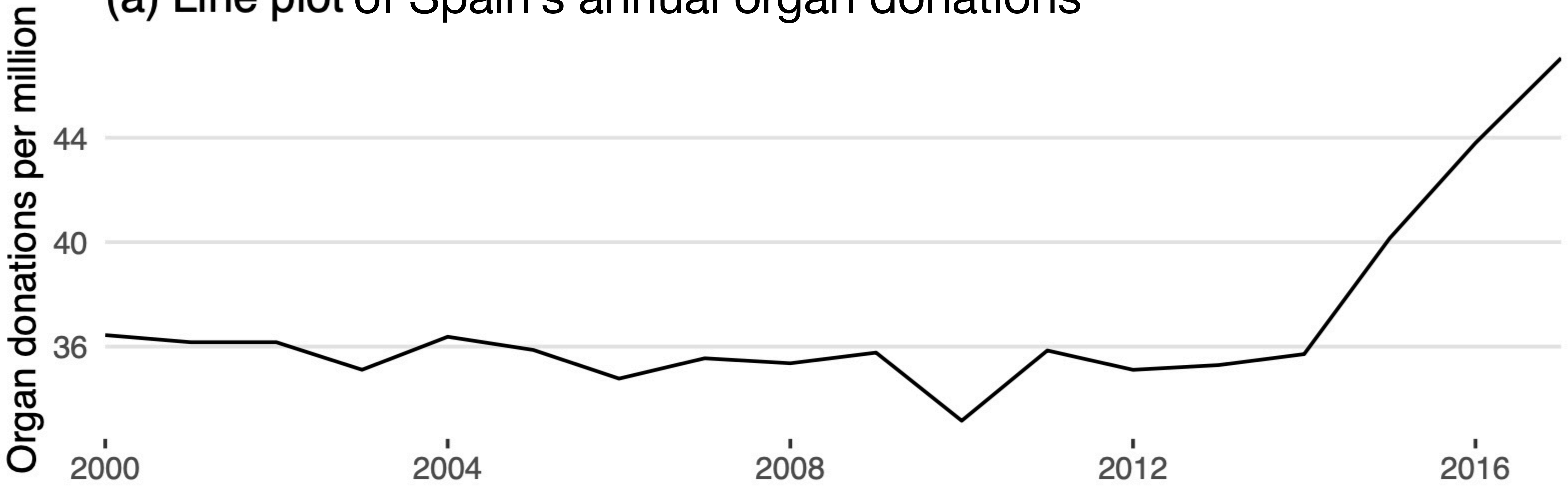
a data visualization guide for business professionals

WILEY

Presenting data honestly

What is your takeaway message from this plot?

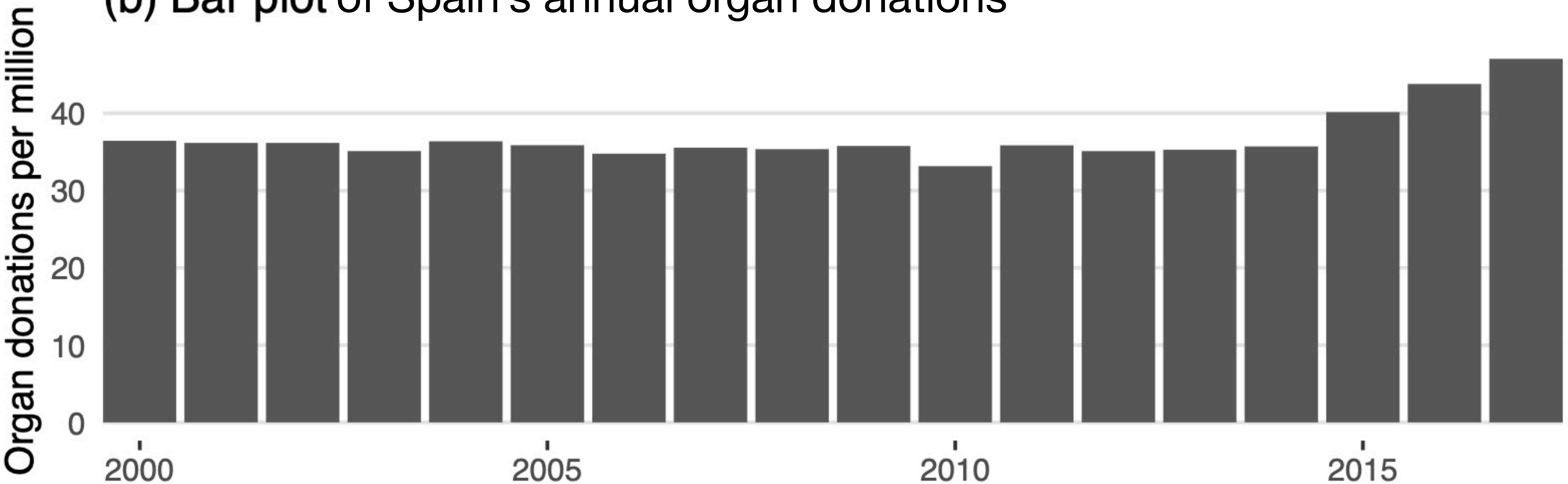
(a) Line plot of Spain’s annual organ donations



Presenting data honestly

What is your takeaway message from this plot?

(b) Bar plot of Spain's annual organ donations



Presenting data honestly

How Charts Lie

Getting Smarter about Visual Information

Alberto Cairo