Tidy Data and Visualization

Lesson 2

API 209: Advanced Quantative Methods
TF: Rony Rodriguez-Ramirez
Summer 2024

Recap and Tidy Data

Wrangling your data {Recap}

- You are ***highly encouraged*** to read through Hadley Wickham's chapter. It's clear and concise.
- Also check out this great "cheatsheet" here.
- The package is organized around a set of **verbs**, i.e. *actions* to be taken.
- All *verbs* work as follows:

$$\text{verb}(\underbrace{\text{data.frame}}_{\text{1st argument}}, \underbrace{\text{what to do}}_{\text{2nd argument}})$$

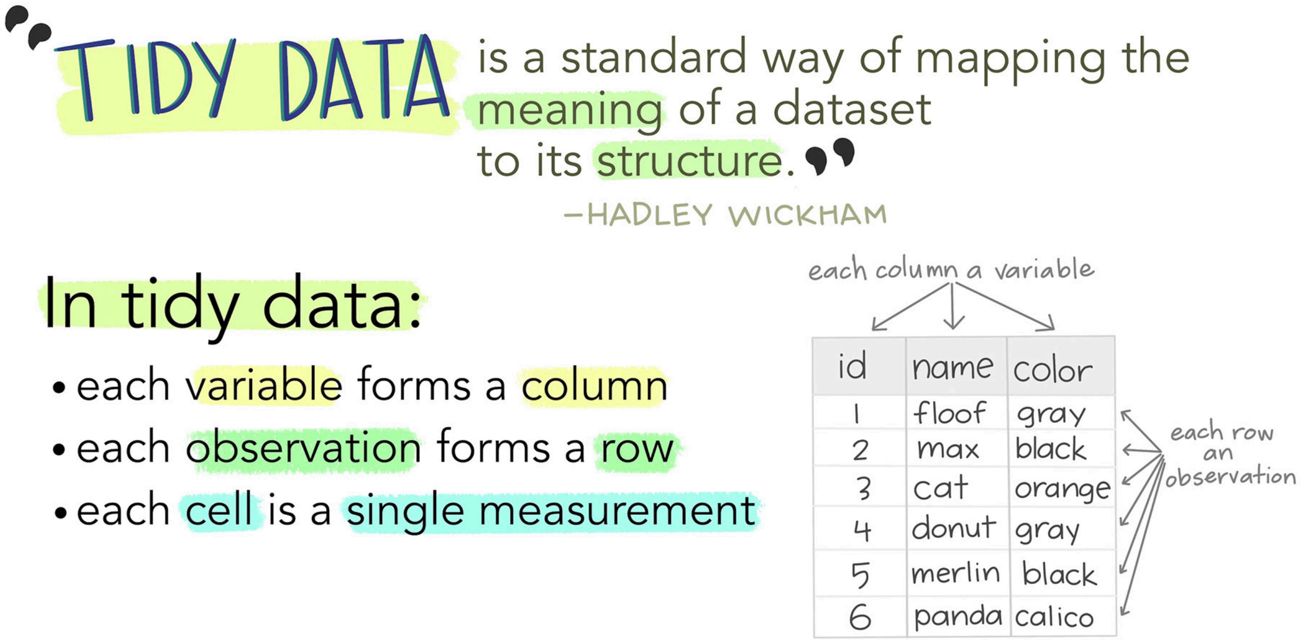
- Alternatively you can (should) use the `pipe` operator `%>%`:

$$\underbrace{\text{data.frame}}_{\text{1st argument}} \quad \underbrace{\%{>}\%}_{\text{"pipe" operator}} \quad \text{verb}(\underbrace{\text{what to do}}_{\text{2nd argument}})$$

Tidy data

- In most cases, your datasets won't be `tidy`.

> **Tidy data**: A dataset is said to be tidy if it satisfies the following conditions:



"TIDY DATA is a standard way of mapping the meaning of a dataset to its structure."

—HADLEY WICKHAM

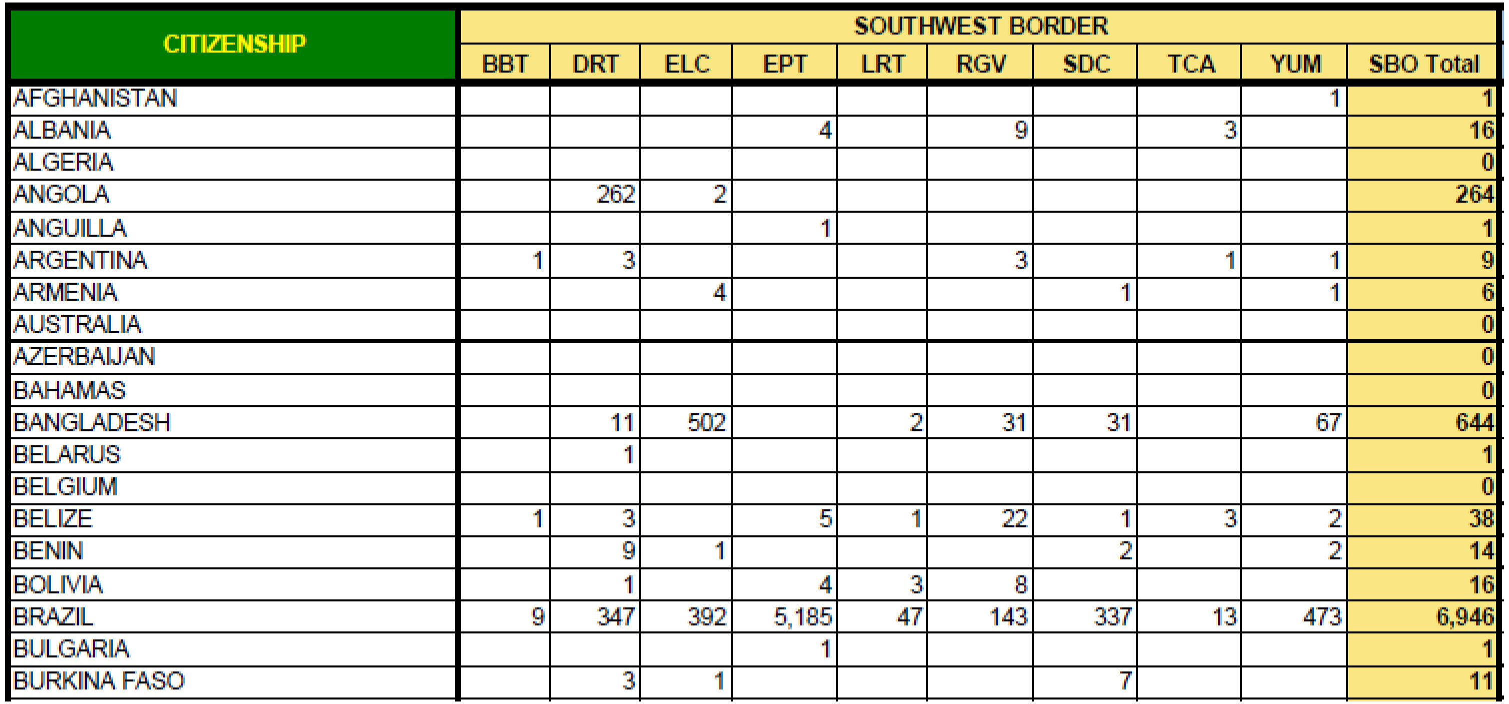
In tidy data:

- each variable forms a column
- each observation forms a row
- each cell is a single measurement

id	name	color
1	floof	gray
2	max	black
3	cat	orange
4	donut	gray
5	merlin	black
6	panda	calico

Wickham, H. (2014). Tidy Data. Journal of Statistical Software 59 (10). DOI: 10.18637/jss.v059.i10

Untidy data is pretty common

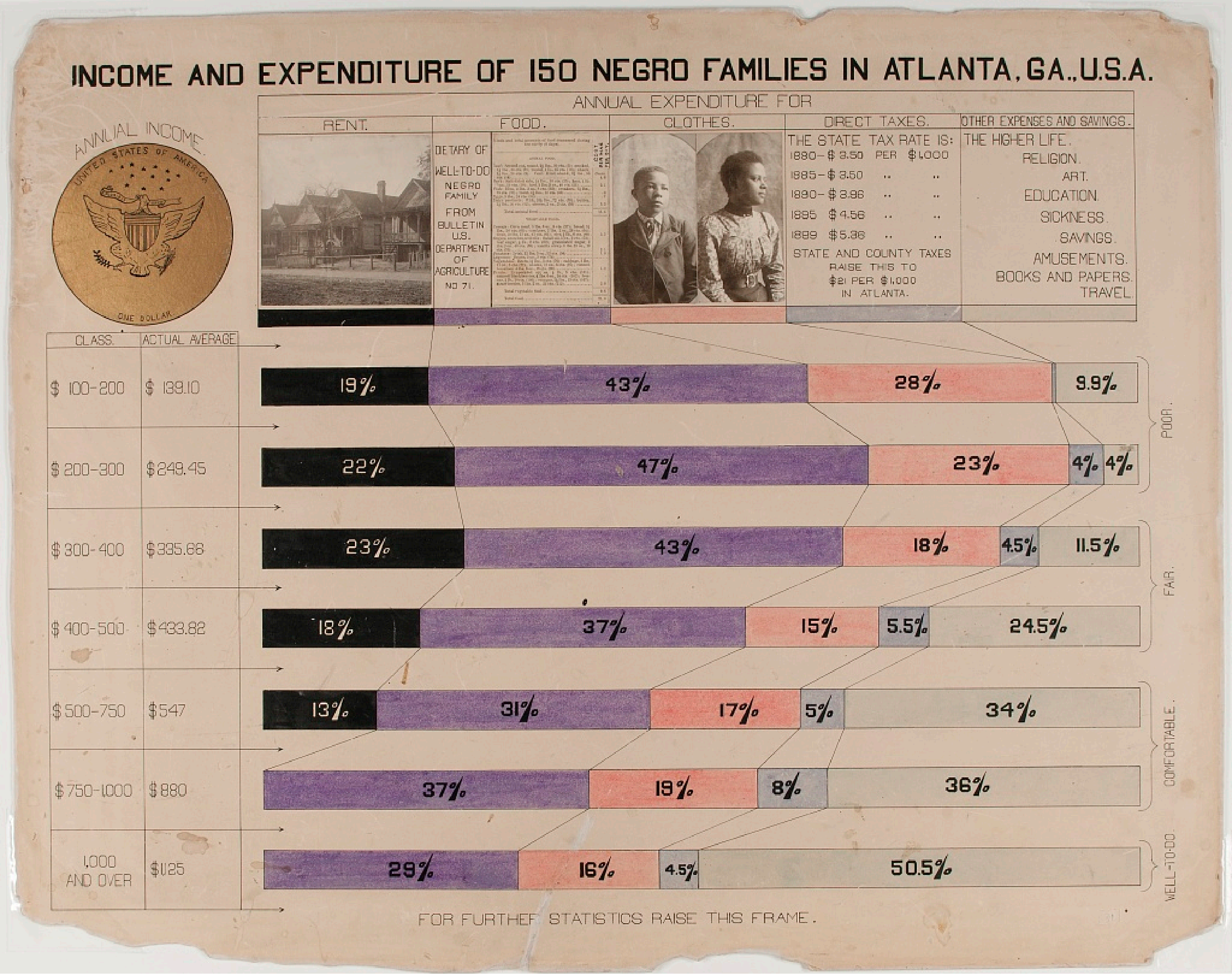


CITIZENSHIP	SOUTHWEST BORDER									
	BBT	DRT	ELC	EPT	LRT	RGV	SDC	TCA	YUM	SBO Total
AFGHANISTAN									1	1
ALBANIA				4		9		3		16
ALGERIA										0
ANGOLA		262	2							264
ANGUILLA				1						1
ARGENTINA	1	3				3		1	1	9
ARMENIA			4				1		1	6
AUSTRALIA										0
AZERBAIJAN										0
BAHAMAS										0
BANGLADESH		11	502		2	31	31		67	644
BELARUS		1								1
BELGIUM										0
BELIZE	1	3		5	1	22	1	3	2	38
BENIN		9	1				2		2	14
BOLIVIA		1		4	3	8				16
BRAZIL	9	347	392	5,185	47	143	337	13	473	6,946
BULGARIA				1						1
BURKINA FASO		3	1				7			11

However, storing data in wide form is easier to display in a printed table.

Tidy
data
is data in
long
format

Beautiful visualizations


INCOME AND EXPENDITURE OF 150 NEGRO FAMILIES IN ATLANTA, GA., U.S.A.
ANNUAL INCOME.
UNITED STATES OF AMERICA
ONE DOLLAR
ANNUAL EXPENDITURE FOR
RENT.
FOOD.
CLOTHES.
DIRECT TAXES.
OTHER EXPENSES AND SAVINGS.
DIETARY OF WELL-TO-DO NEGRO FAMILY FROM BULLETIN U.S. DEPARTMENT OF AGRICULTURE NO 71.
THE STATE TAX RATE IS:
1880 – $ 3.50 PER $1,000
1885 – $ 3.50 " "
1890 – $ 3.96 " "
1895 $ 4.56 " "
1899 $ 5.36 " "
STATE AND COUNTY TAXES RAISE THIS TO $21 PER $1,000 IN ATLANTA.
THE HIGHER LIFE.
RELIGION.
ART.
EDUCATION.
SICKNESS.
SAVINGS.
AMUSEMENTS.
BOOKS AND PAPERS.
TRAVEL.
CLASS.
ACTUAL AVERAGE.
$ 100-200
$ 139.10
19%
43%
28%
9.9%
$ 200-300
$ 249.45
22%
47%
23%
4%
4%
$ 300-400
$ 335.66
23%
43%
18%
4.5%
11.5%
$ 400-500
$ 433.82
18%
37%
15%
5.5%
24.5%
$ 500-750
$ 547
13%
31%
17%
5%
34%
$ 750-1000
$ 880
37%
19%
8%
36%
1,000 AND OVER
$ 1125
29%
16%
4.5%
50.5%
POOR.
FAIR.
COMFORTABLE.
WELL-TO-DO.
FOR FURTHER STATISTICS RAISE THIS FRAME.

What makes a great visualization?

Truthful

Functional

Beautiful

Insightful

Enlightening

Alberto Cairo, *The Truthful Art*

How do we express visuals in words?

- **Data** to be visualized
- **Geometric objects** that appear on the plot
- **Aesthetic mappings** from data to visual component
- **Statistics** transform data on the way to visualization
- **Coordinates** organize location of geometric objects
- **Scales** define the range of values for aesthetics
- **Facets** group into subplots

What makes a great visualization?

Good aesthetics

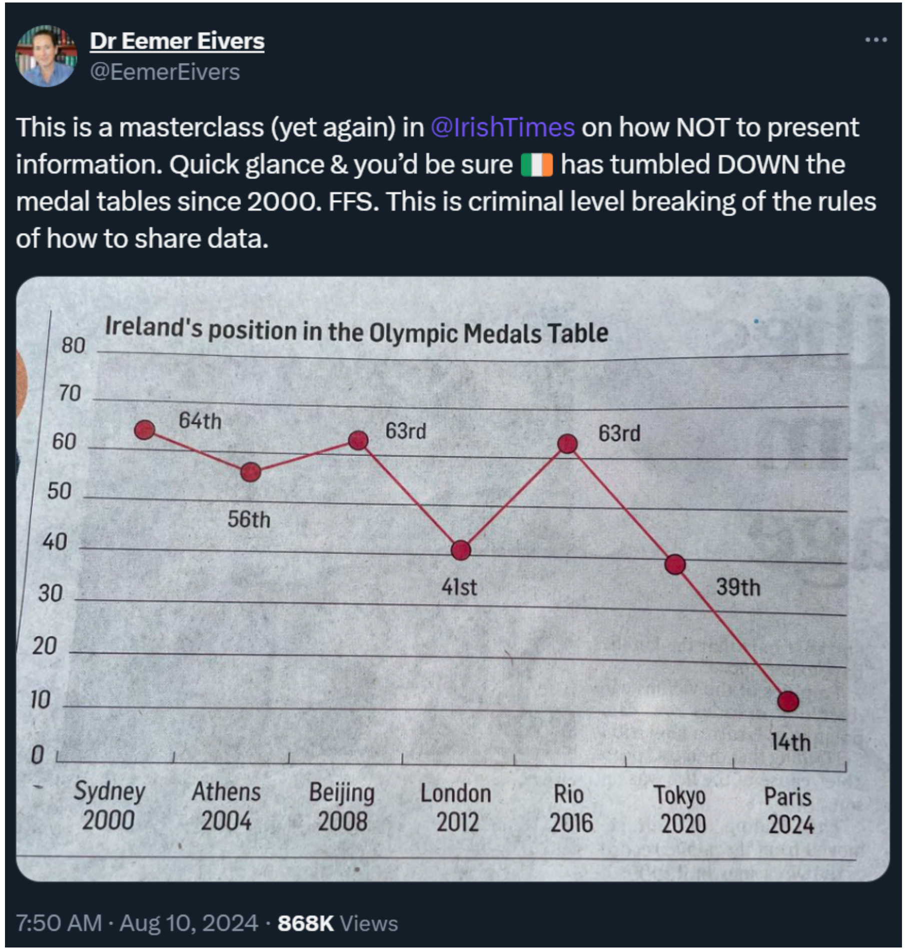
No substantive issues

No perceptual issues

Honesty + good judgment

Kieran Healy, *Data Visualization: A Practical Introduction*

You see bad plots everywhere: What's wrong?

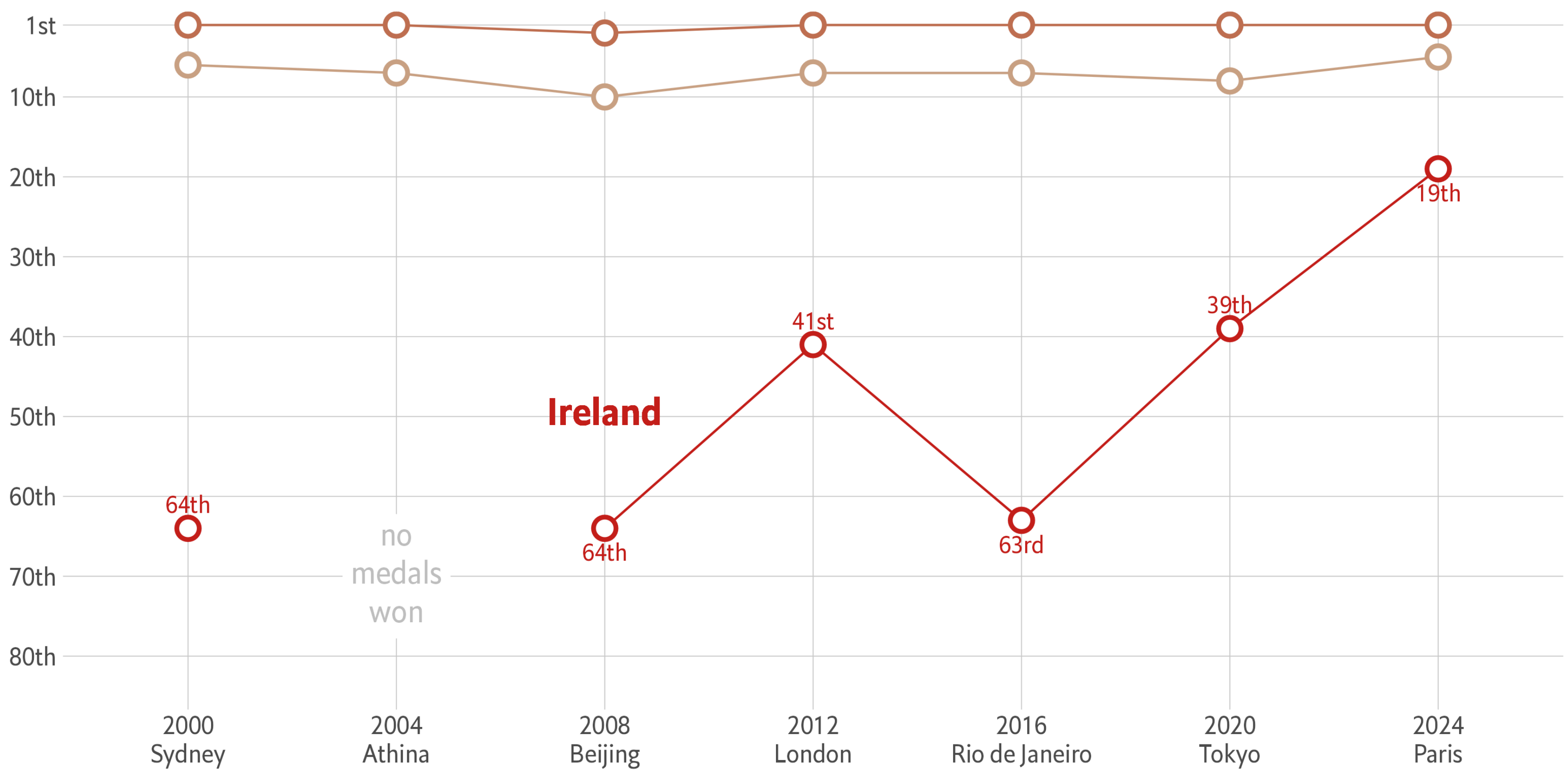


Dr Eemer Eivers
@EemerEivers

This is a masterclass (yet again) in @IrishTimes on how NOT to present information. Quick glance & you'd be sure 🇮🇪 has tumbled DOWN the medal tables since 2000. FFS. This is criminal level breaking of the rules of how to share data.

7:50 AM · Aug 10, 2024 · **868K** Views

Is this right?



Ireland's position in the Olympics Medals Table

Compared to the position of the United States and France

Source: Wikipedia via Kaggle · Created with GGPlot · **Original Chart**: @lisacmuth · **This Chart**: @rrmaximiliano

Entering ggplot

ggplot

For this session, you'll use the ggplot2 package from the tidyverse meta-package.

- So, you can just load the `tidyverse` package when using ggplot.

1. Consistency with the **Grammar of Graphics**
    - This book is the foundation of several data viz applications:

    `ggplot2, polaris-tableau, vega-lite`
2. Flexibility
3. Layering and theme customization
4. Community

It is a powerful and easy to use tool (once you understand its logic) that produces complex and multifaceted plots.

ggplot2: basic structure (template)

The basic ggplot structure is:

```
ggplot(data = DATA) +
  GEOM_FUNCTION(mapping = aes(AESTHETIC MAPPINGS))
```

Mapping data to aesthetics

Think about colors, sizes, x and y references

We are going to learn how we connect our data to the components of a ggplot.

I usually code like this:

```
DATA |>
  ggplot(aes(AESTHETIC MAPPINGS)) +
  GEOM_FUNCTION()
```

Mapping

Mappings do not directly specify the particular, e.g., colors, shapes, or line styles that will appear on the plot. Rather, they establish which variables in the data will be represented by which visible elements on the plot.

ggplot2: full structure

```
ggplot(data = <DATA>) +
  <GEOM_FUNCTION>(
     mapping = aes(<MAPPINGS>),
     stat = <STAT>,
     position = <POSITION>
  ) +z
  <COORDINATE_FUNCTION> +
  <FACET_FUNCTION> +
  <SCALE_FUNTION> +
  <THEME_FUNCTION>
```

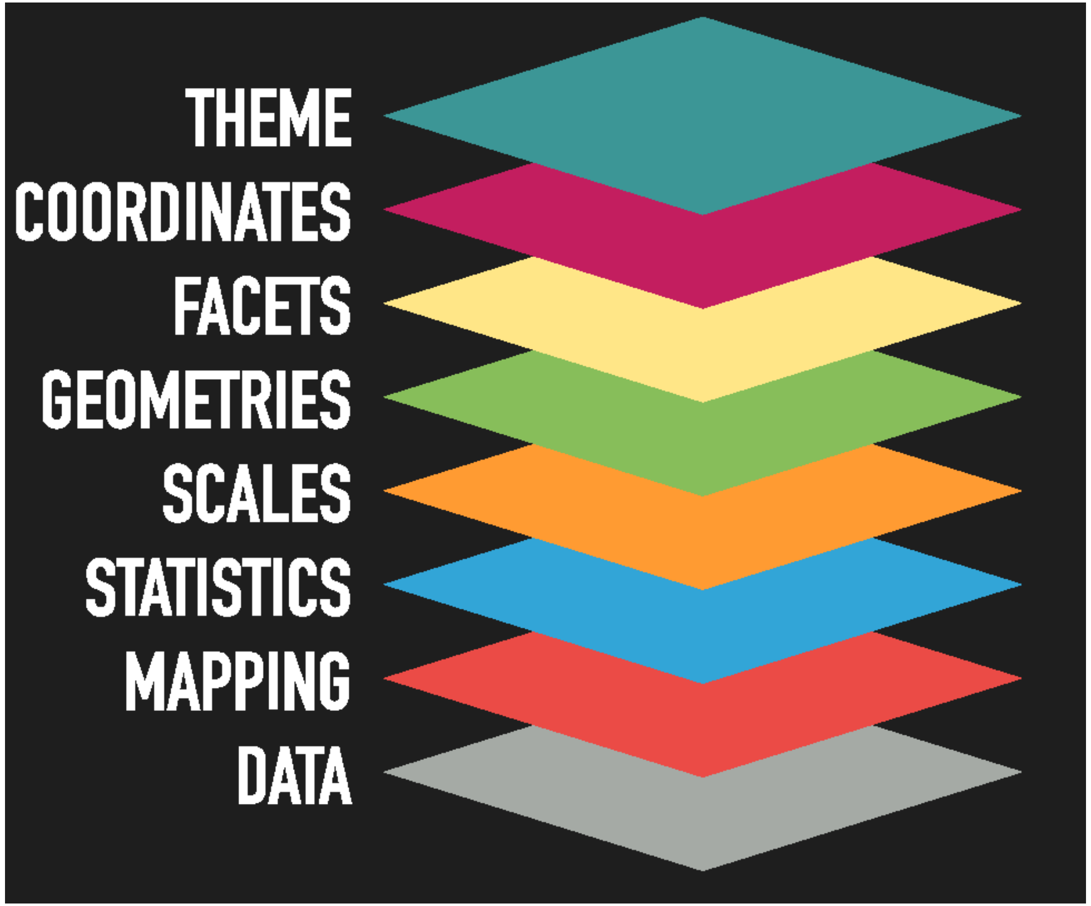
1. `Data`: The data that you want to visualize
2. `Layers`: geom_ and stat_ → The geometric shapes and statistical summaries representing the data
3. `Aesthetics`: aes() → Aesthetic mappings of the geometric and statistical objects
4. `Scales`: scale_ → Maps between the data and the aesthetic dimensions
5. `Coordinate system`: coord_ → Maps data into the plane of the data rectangle
6. `Facets`: facet_ → The arrangement of the data into a grid of plots
7. `Visual themes`: theme() and theme_ → The overall visual defaults of a plot

ggplot2: decomposition

There are multiple ways to structure plots with ggplot

For this presentation, I will stick to Thomas Lin Pedersen's decomposition who is one of most prominent developers of the ggplot and gganimate package.

These components can be seen as layers, this is why we use the + sign in our ggplot syntax.



Exploratory Analysis

The most common `geoms` are:

- `geom_bar()`, `geom_col()`: bar charts.
- `geom_boxplot()`: box and whiskers plots.
- `geom_density()`: density estimates.
- `geom_jitter()`: jittered points.
- `geom_line()`: line plots.
- `geom_point()`: scatter plots.

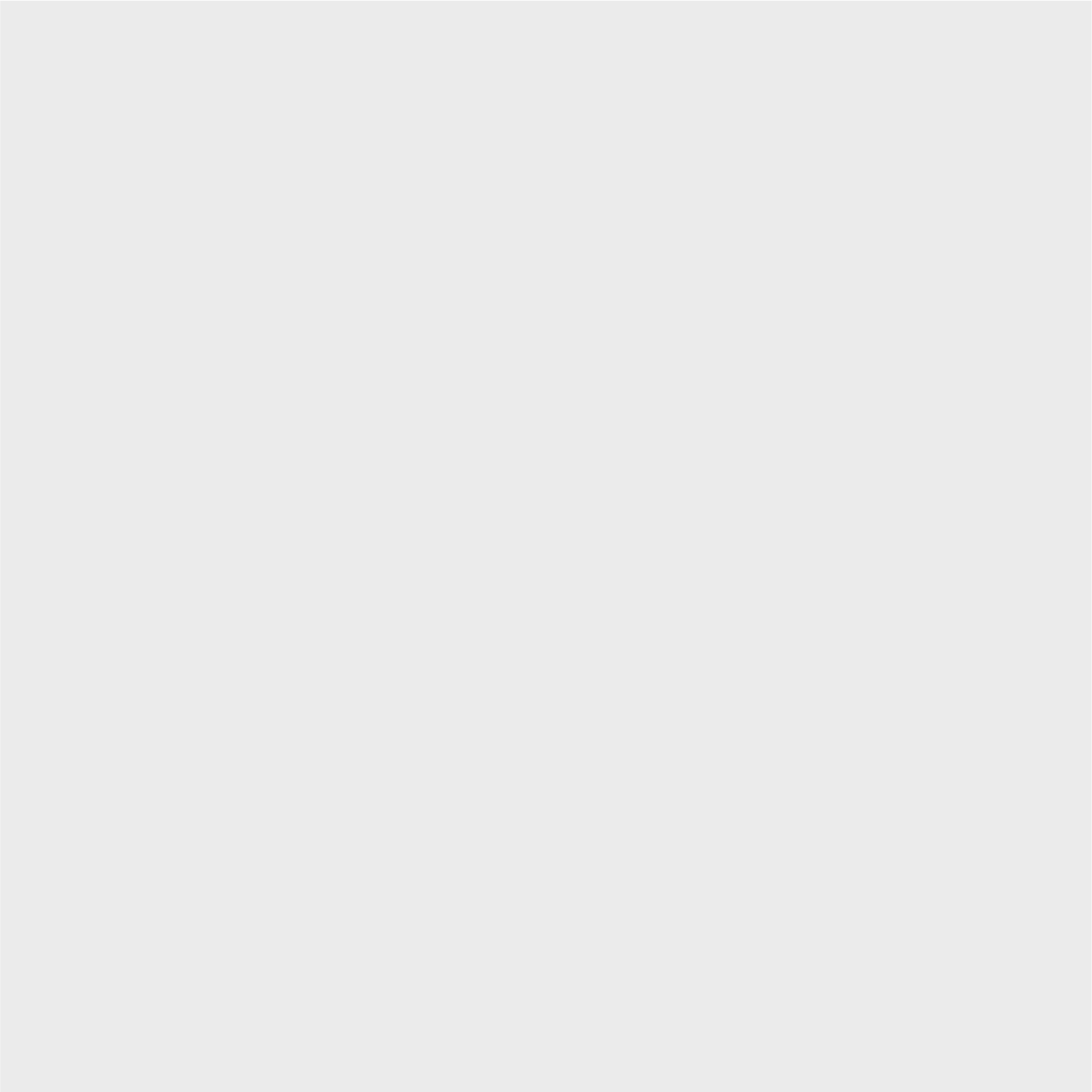
> If you want to know more about layers, you can refer to this.

Step by step from Garrick Aden-Buie's gentle guide

Using the `gapminder` package, let's start with `lifeExp` vs `gdpPercap`

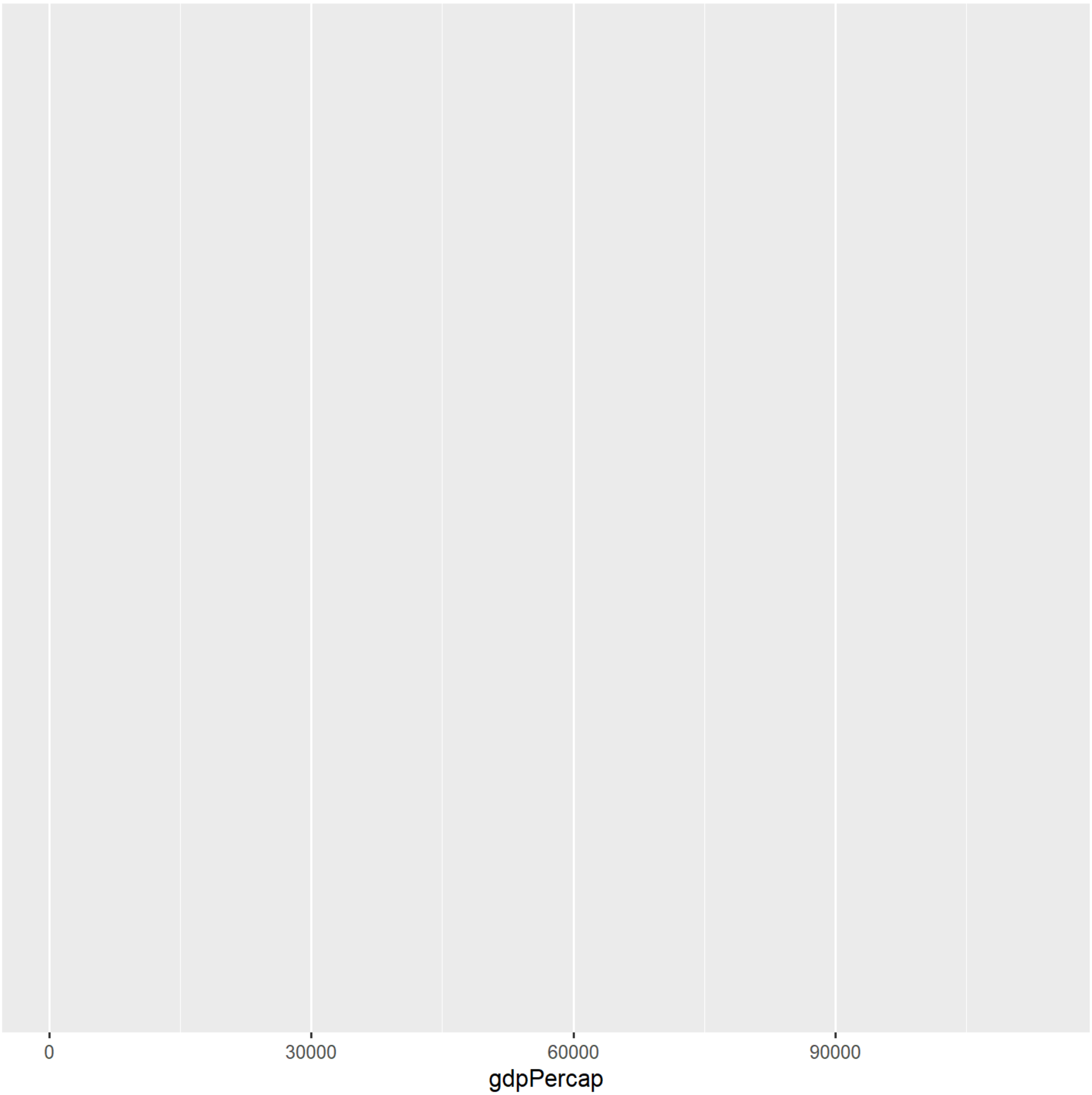
```
Rows: 1,704
Columns: 6
$ country   <fct> "Afghanistan", "Afghanistan", "Afghanistan", "Afghanistan", …
$ continent <fct> Asia, Asia, Asia, Asia, Asia, Asia, Asia, Asia, Asia, Asia, …
$ year      <int> 1952, 1957, 1962, 1967, 1972, 1977, 1982, 1987, 1992, 1997, …
$ lifeExp   <dbl> 28.801, 30.332, 31.997, 34.020, 36.088, 38.438, 39.854, 40.8…
$ pop       <int> 8425333, 9240934, 10267083, 11537966, 13079460, 14880372, 12…
$ gdpPercap <dbl> 779.4453, 820.8530, 853.1007, 836.1971, 739.9811, 786.1134, …
```

```
ggplot(gapminder)
```



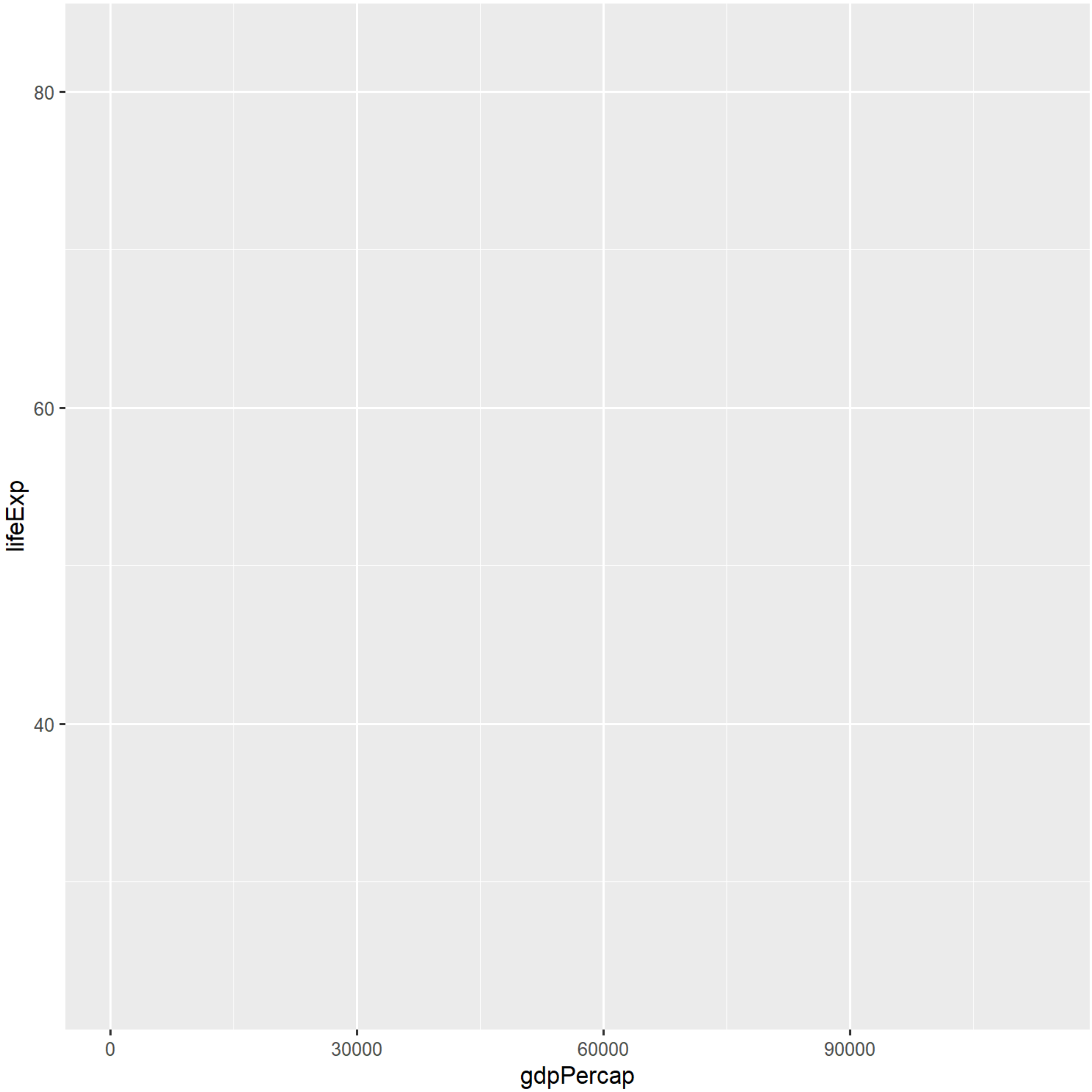
The Canvas

```
ggplot(gapminder) +
  aes(x = gdpPercap)
```



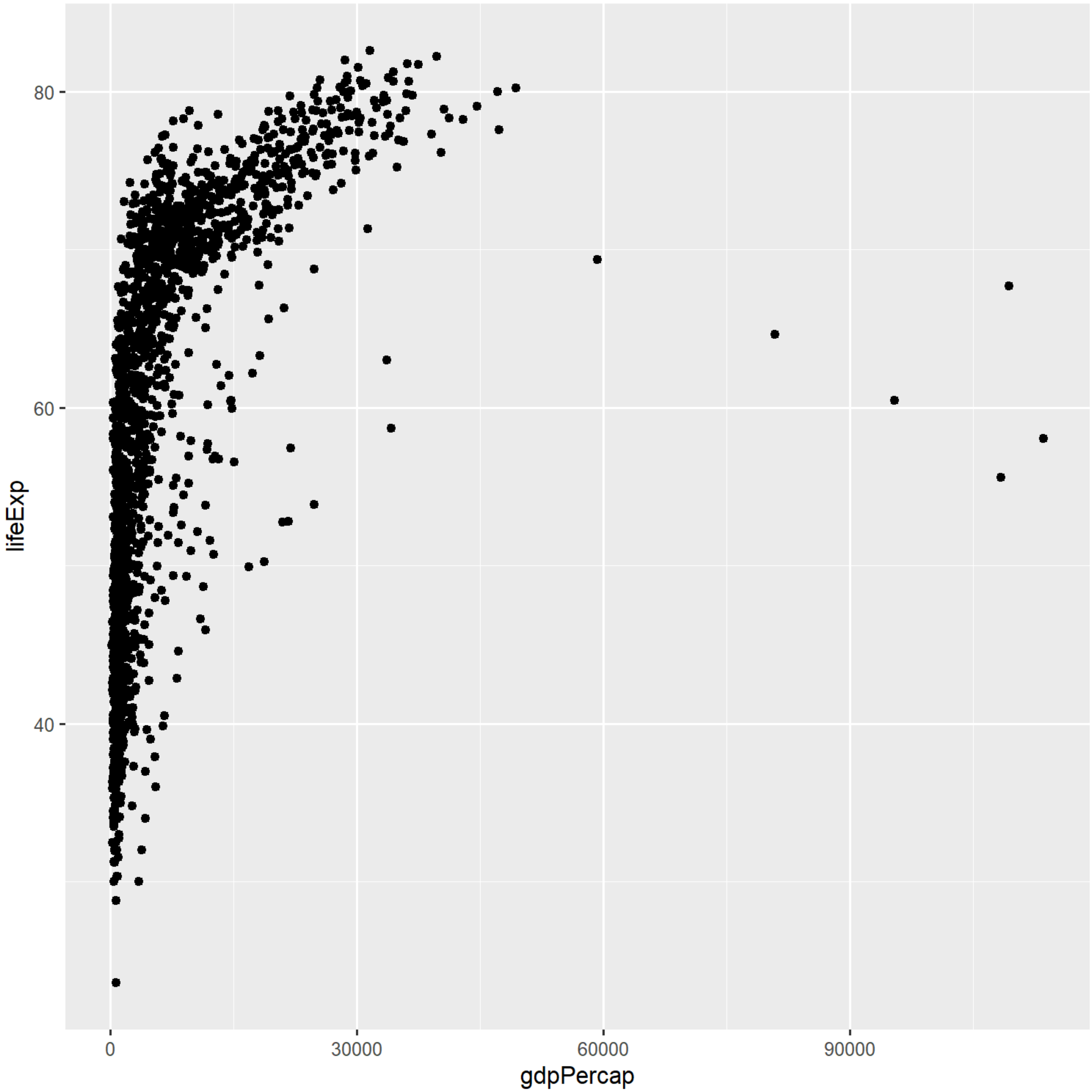
The Canvas

```
ggplot(gapminder) +
  aes(x = gdpPercap,
      y = lifeExp)
```



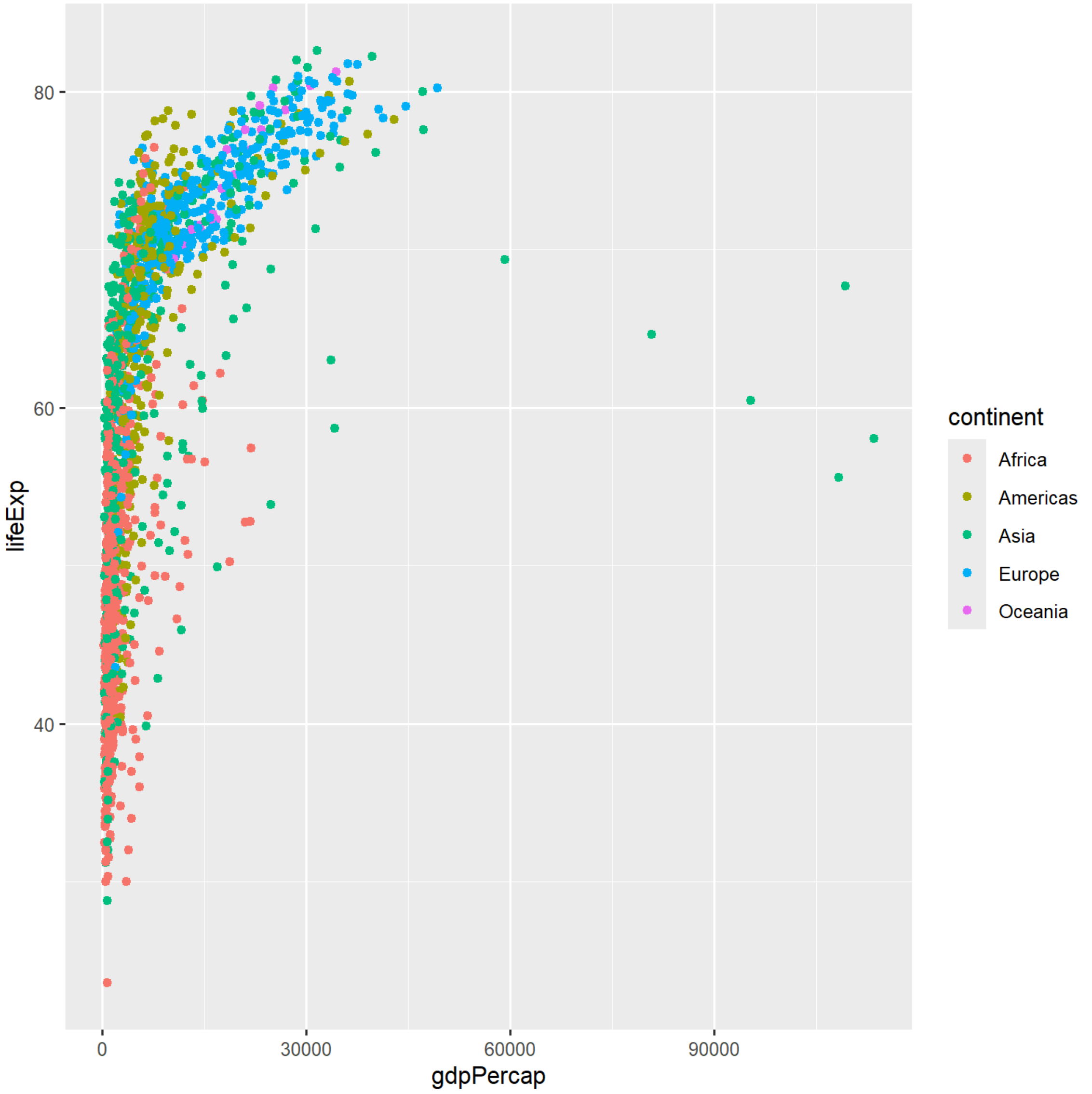
Add points...

```
ggplot(gapminder) +
  aes(x = gdpPercap,
      y = lifeExp) +
  geom_point()
```



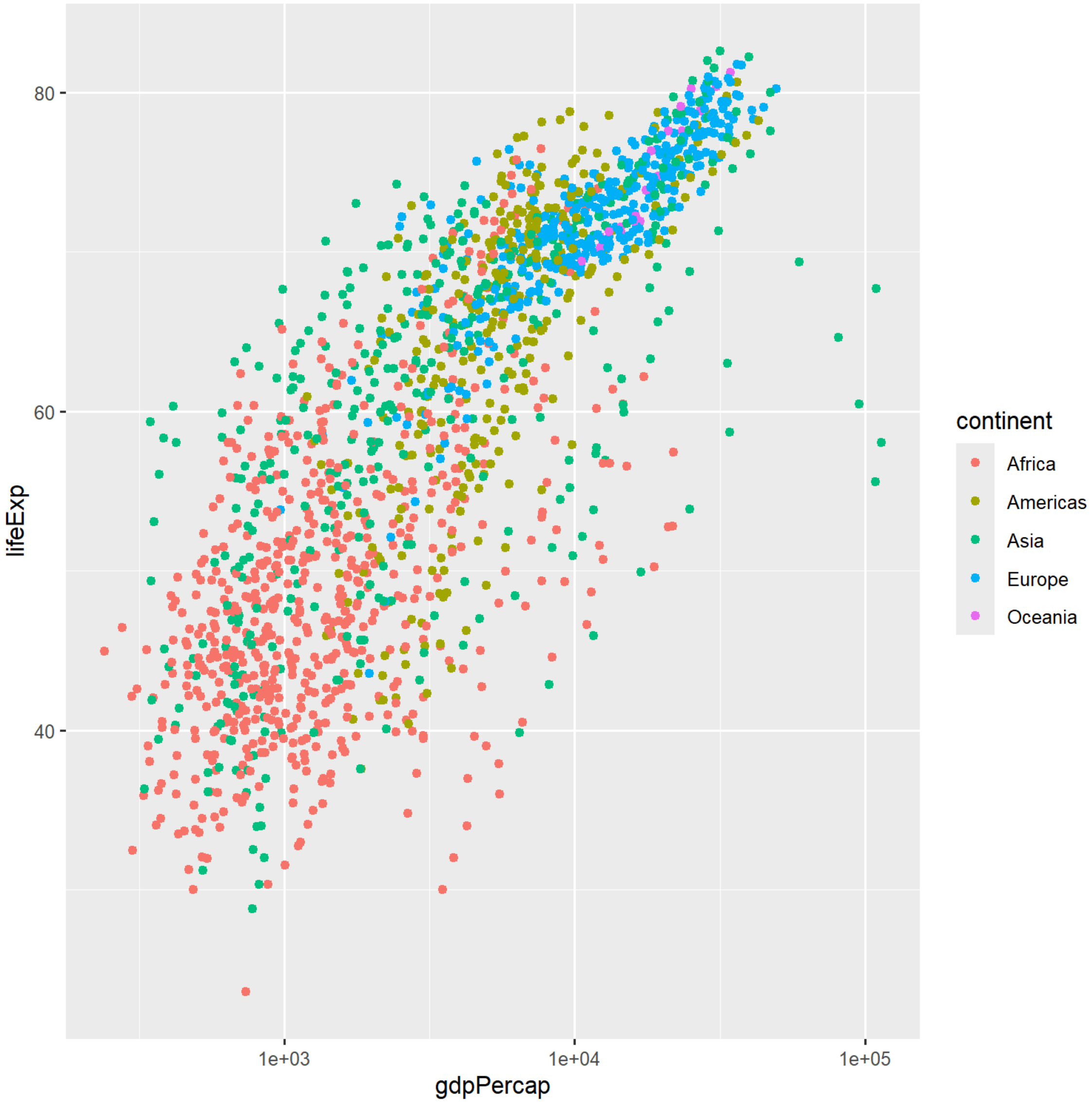
How can I tell countries apart?

```
ggplot(gapminder) +
  aes(x = gdpPercap,
      y = lifeExp,
      color = continent) +
  geom_point()
```



GDP is squished together on the left

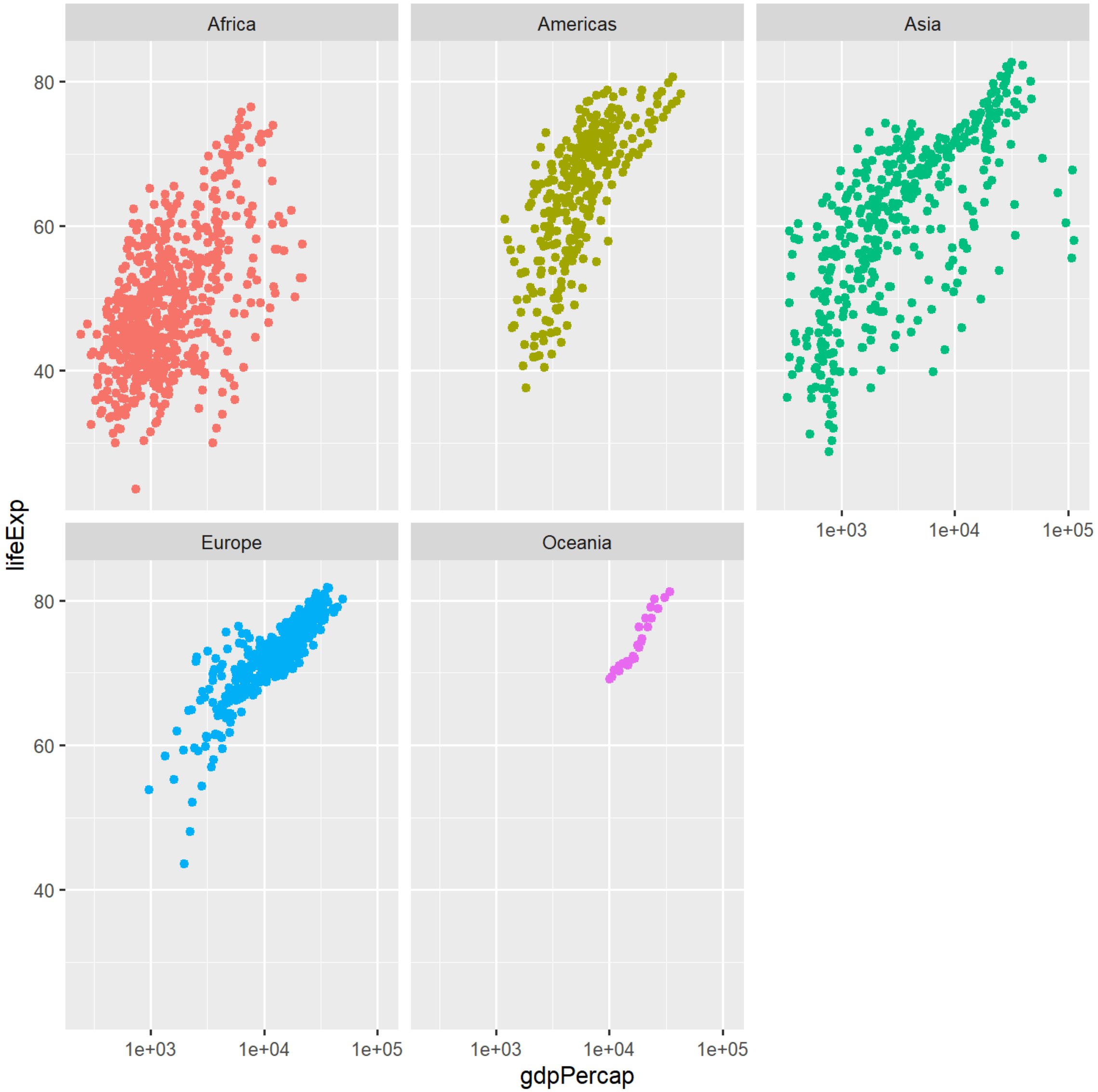
```
ggplot(gapminder) +
  aes(x = gdpPercap,
      y = lifeExp,
      color = continent) +
  geom_point() +
  scale_x_log10()
```



Still lots of overlap in the countries...

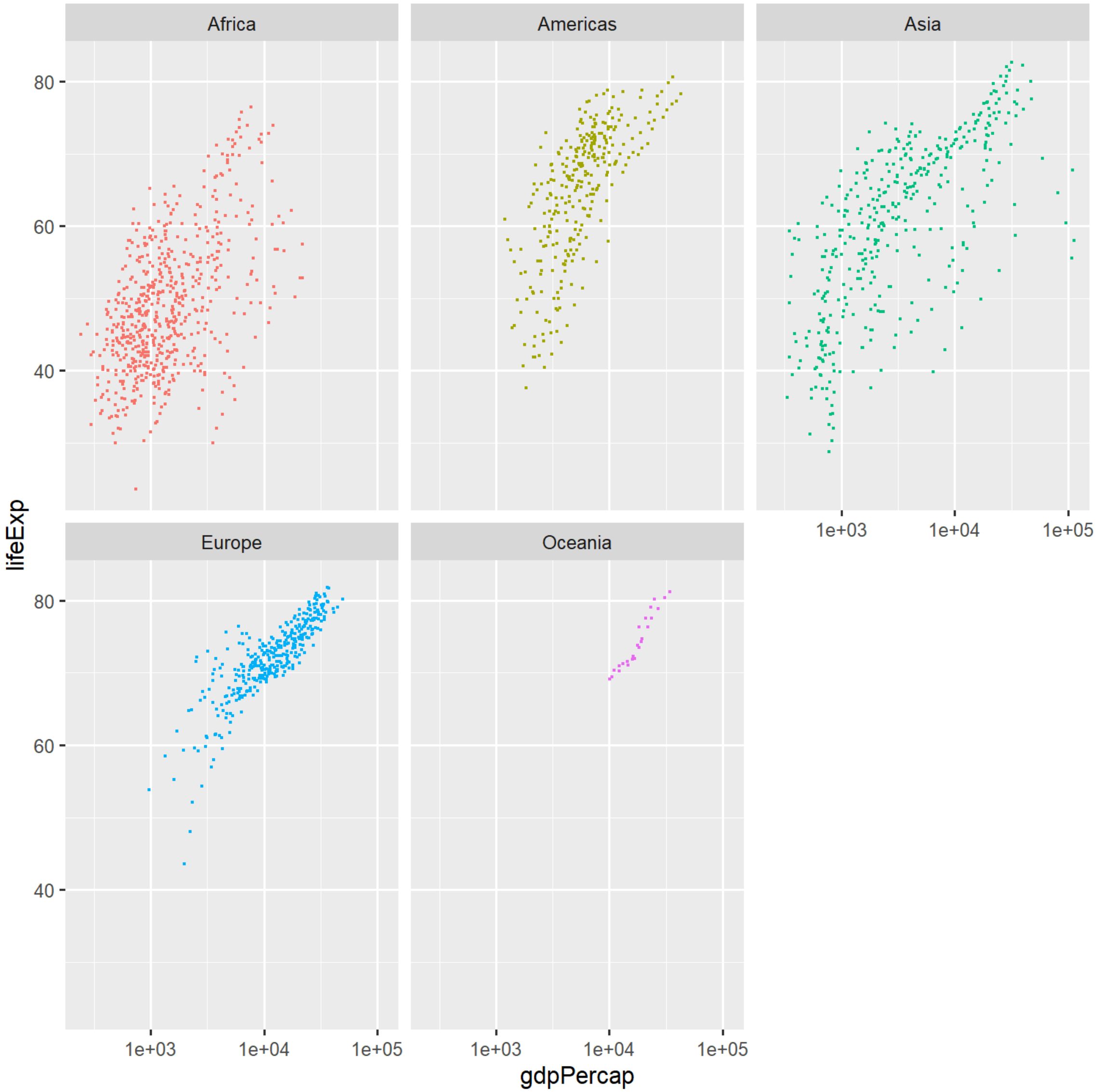
```r
ggplot(gapminder) +
  aes(x = gdpPercap,
      y = lifeExp,
      color = continent) +
  geom_point() +
  scale_x_log10() +
  facet_wrap(~ continent) +
  guides(color = FALSE)
```

No need for color legend thanks to facet titles



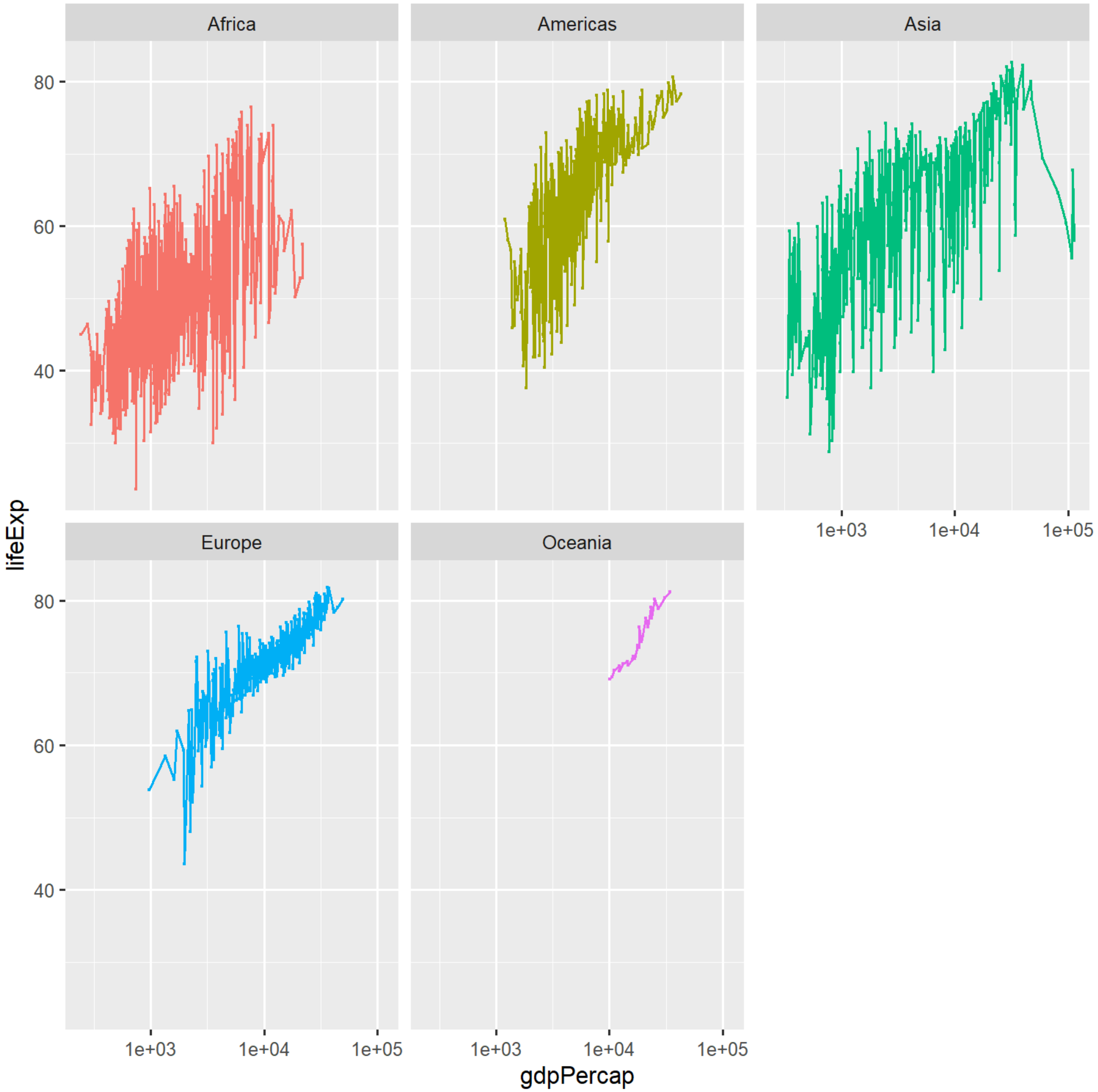
Lots of overplotting due to point size

```
ggplot(gapminder) +
  aes(x = gdpPercap,
      y = lifeExp,
      color = continent) +
  geom_point(size = 0.25) +
  scale_x_log10() +
  facet_wrap(~ continent) +
  guides(color = FALSE)
```



Is there a trend?

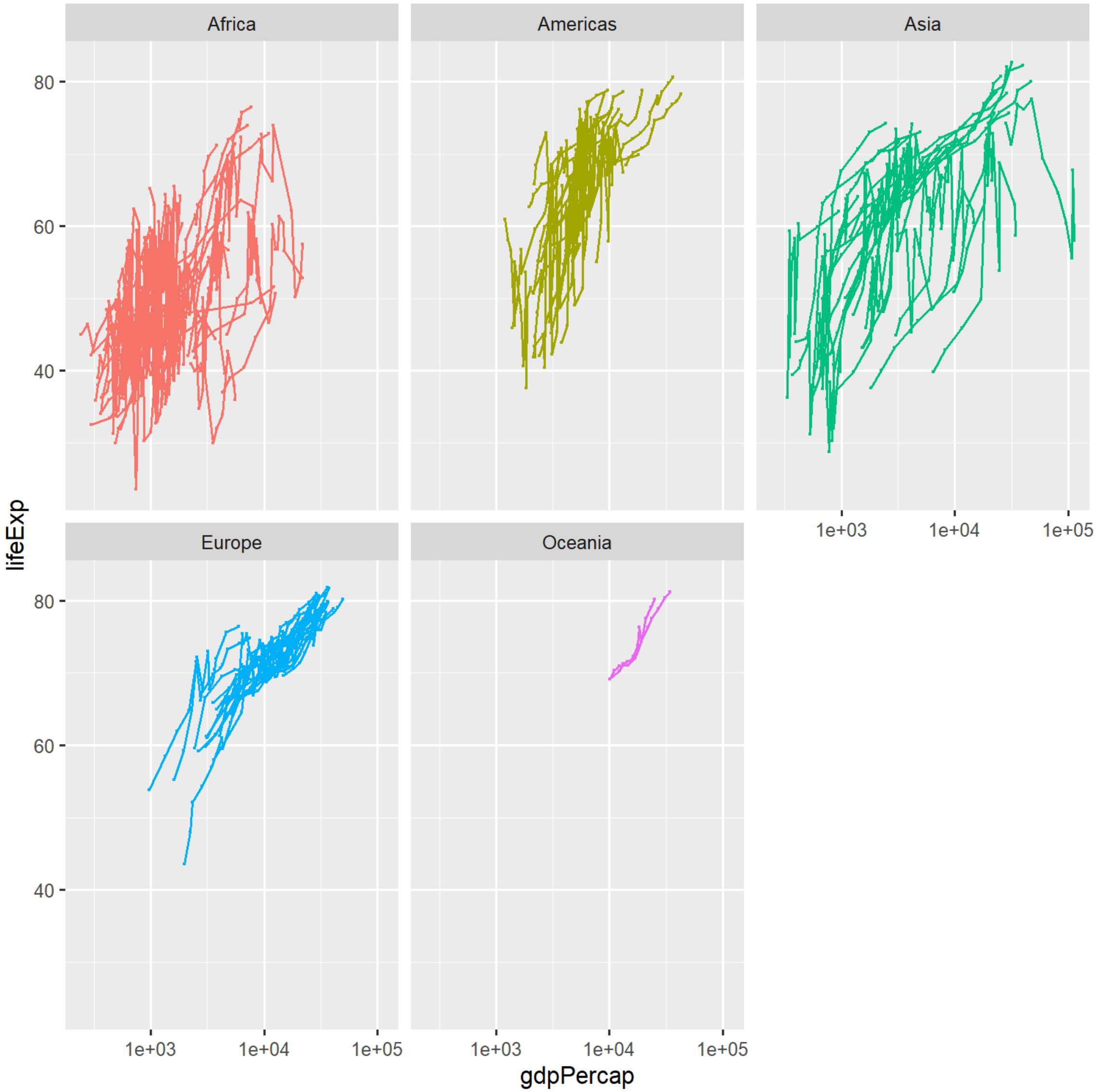
```
ggplot(gapminder) +
  aes(x = gdpPercap,
      y = lifeExp,
      color = continent) +
  geom_line() +
  geom_point(size = 0.25) +
  scale_x_log10() +
  facet_wrap(~ continent) +
  guides(color = FALSE)
```



Okay, that line just connected all of the points sequentially...

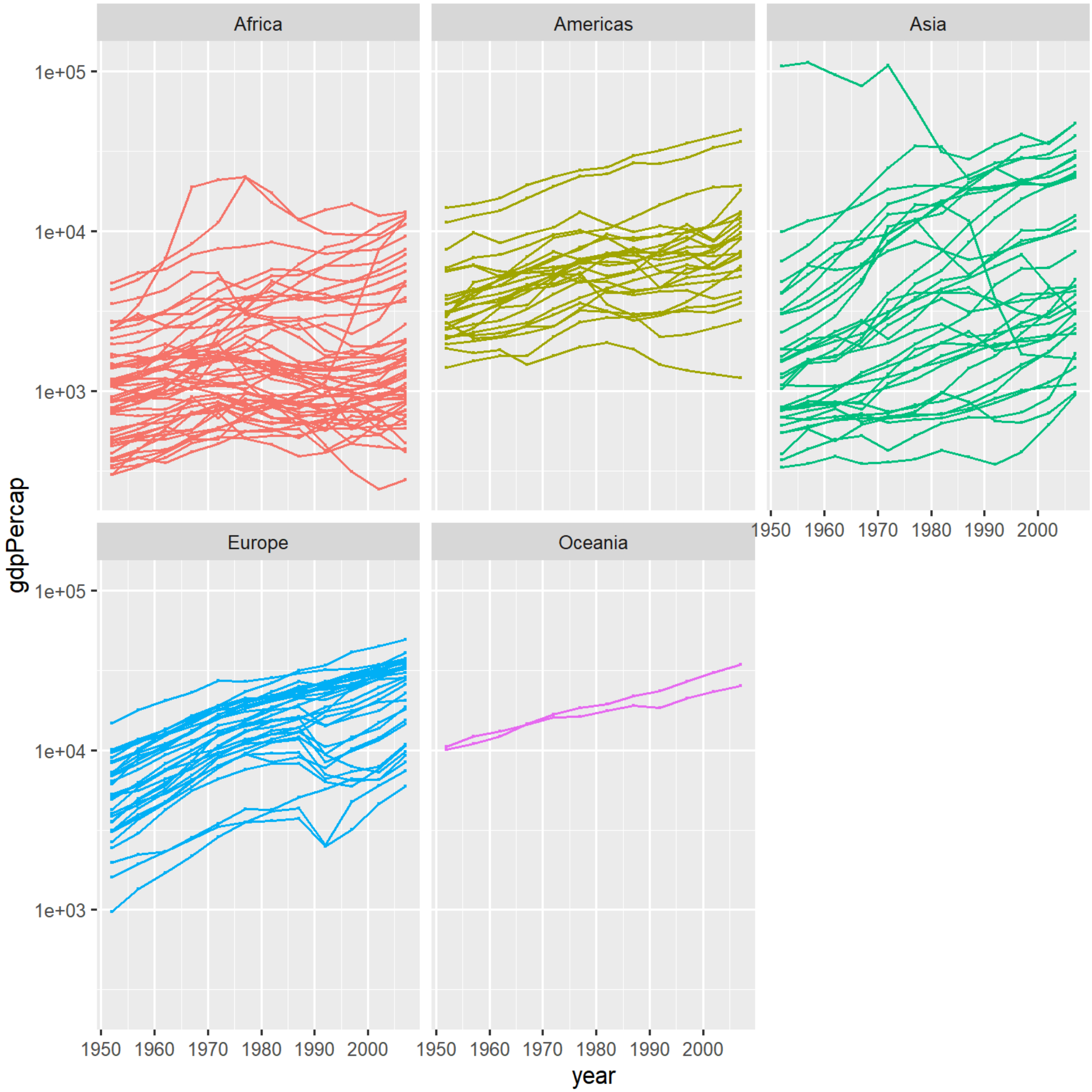
```
ggplot(gapminder) +
  aes(x = gdpPercap,
      y = lifeExp,
      color = continent) +
  geom_line(
    aes(group = country)
  ) +
  geom_point(size = 0.25) +
  scale_x_log10() +
  facet_wrap(~ continent) +
  guides(color = FALSE)
```

🤔



💡 We need time on x-axis!

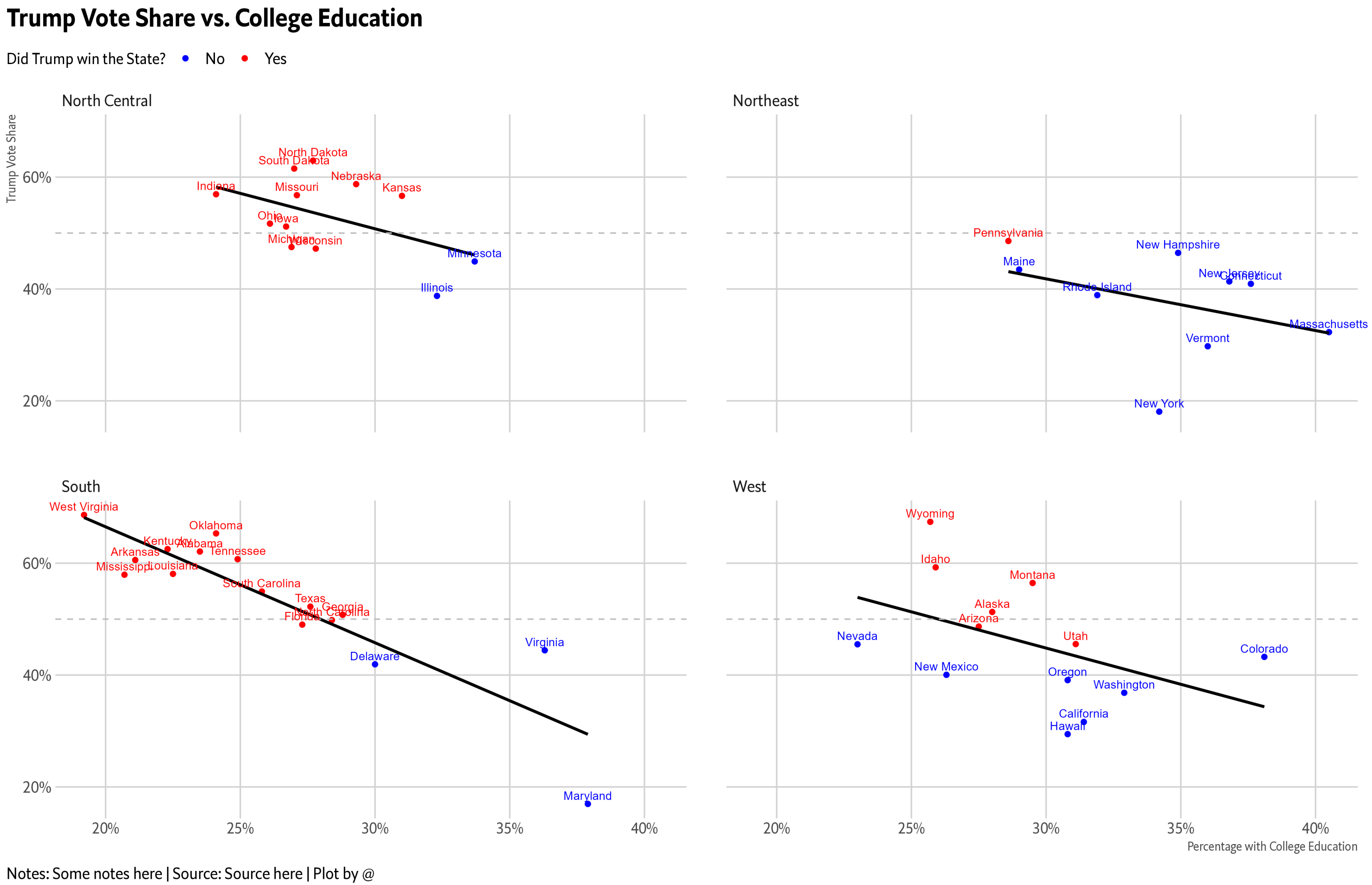
```
ggplot(gapminder) +
  aes(x = year,
      y = gdpPercap,
      color = continent) +
  geom_line(
    aes(group = country)
  ) +
  geom_point(size = 0.25) +
  scale_y_log10() +
  facet_wrap(~ continent) +
  guides(color = FALSE)
```



Time to code

via GIPHY

Our goal


Trump Vote Share vs. College Education
Did Trump win the State? ● No ● Yes
North Central
Northeast
South
West
Trump Vote Share
Percentage with College Education
20%
40%
60%
20%
25%
30%
35%
40%
North Dakota
South Dakota
Nebraska
Indiana
Missouri
Kansas
Ohio
Iowa
Michigan
Wisconsin
Minnesota
Illinois
Pennsylvania
New Hampshire
Maine
New Jersey
Connecticut
Rhode Island
Massachusetts
Vermont
New York
West Virginia
Oklahoma
Kentucky
Alabama
Arkansas
Tennessee
Mississippi
Louisiana
South Carolina
Texas
Georgia
North Carolina
Florida
Virginia
Delaware
Maryland
Wyoming
Idaho
Montana
Alaska
Arizona
Utah
Nevada
Colorado
New Mexico
Oregon
Washington
California
Hawaii
Notes: Some notes here | Source: Source here | Plot by @