

# 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 `%>%`:

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

# 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

each column a variable

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

each row an observation

# 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.

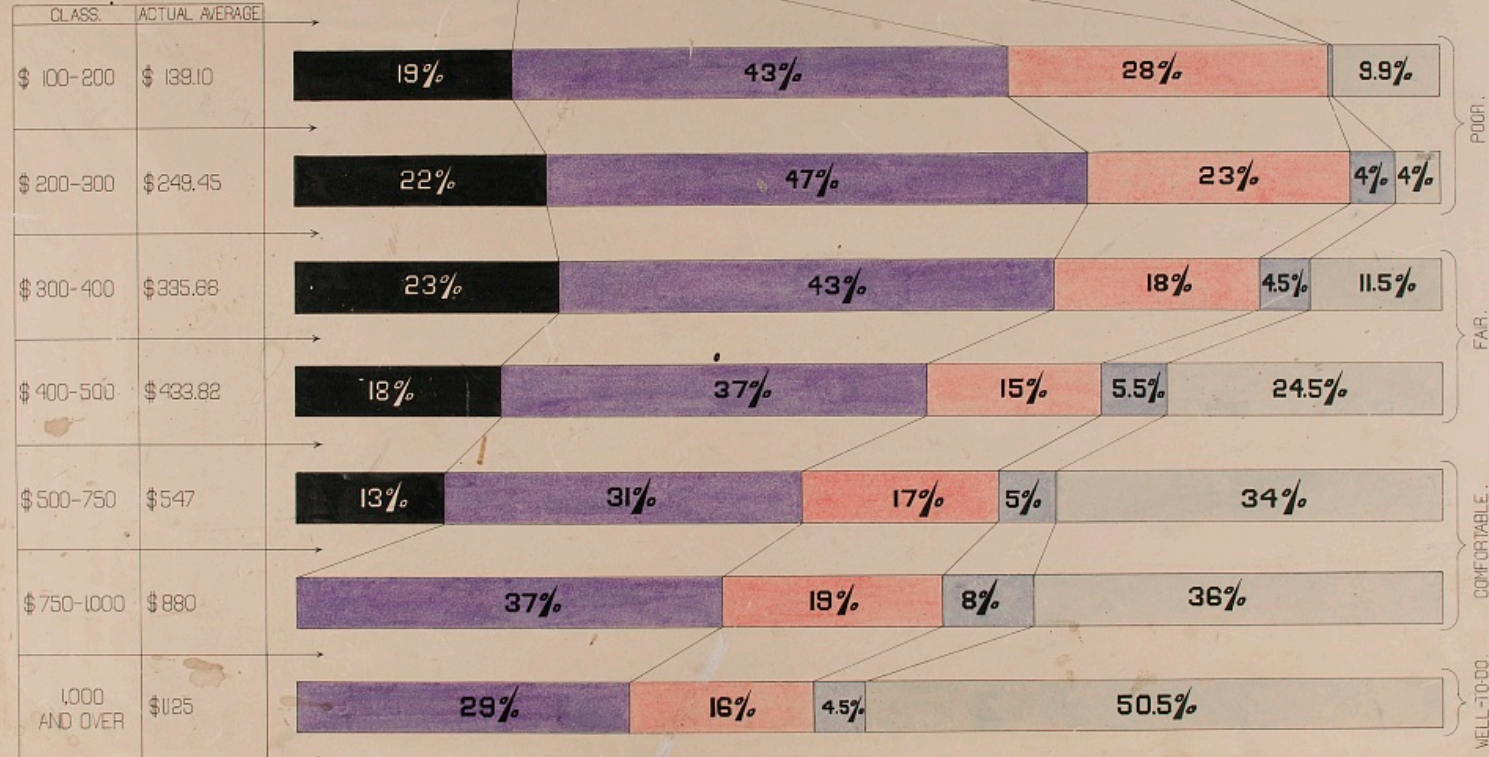
**Tidy**  
data  
is data in  
**long**  
format

# Beautiful visualizations

# INCOME AND EXPENDITURE OF 150 NEGRO FAMILIES IN ATLANTA, GA., U.S.A.



ANNUAL EXPENDITURE FOR				
RENT.	FOOD.	CLOTHES.	DIRECT TAXES.	OTHER EXPENSES AND SAVINGS.
	<p>DETAILED LIST OF WELL-TO-DO NEGRO FAMILY FROM BULLETIN U.S. DEPARTMENT OF AGRICULTURE NO. 71.</p> <p>FOOD AND other provisions of food measured during year and in dollar.</p> <p>WELL-TO-DO NEGRO FAMILY</p> <p>1. Food and other provisions of food measured during year and in dollar.</p> <p>2. Clothing and other expenses measured during year and in dollar.</p> <p>3. Direct taxes measured during year and in dollar.</p> <p>4. Other expenses and savings measured during year and in dollar.</p> <p>5. Total annual income measured during year and in dollar.</p>		<p>THE STATE TAX RATE IS:</p> <p>1880 - \$ 3.50 PER \$1000</p> <p>1885 - \$ 3.50 .. ..</p> <p>1890 - \$ 3.86 .. ..</p> <p>1895 - \$ 4.56 .. ..</p> <p>1899 - \$ 5.36 .. ..</p> <p>STATE AND COUNTY TAXES RAISE THIS TO \$21 PER \$1,000 IN ATLANTA.</p>	<p>THE HIGHER LIFE.</p> <p>RELIGION.</p> <p>ART.</p> <p>EDUCATION.</p> <p>SICKNESS.</p> <p>SAVINGS.</p> <p>AMUSEMENTS.</p> <p>BOOKS AND PAPERS.</p> <p>TRAVEL.</p>



FOR FURTHER STATISTICS RAISE THIS FRAME.



# What makes a great visualization?

Truthful

Functional

Beautiful

Insightful

Enlightening

Alberto Cairo, *The Truthful Art*

# What makes a great visualization?

**Good aesthetics**

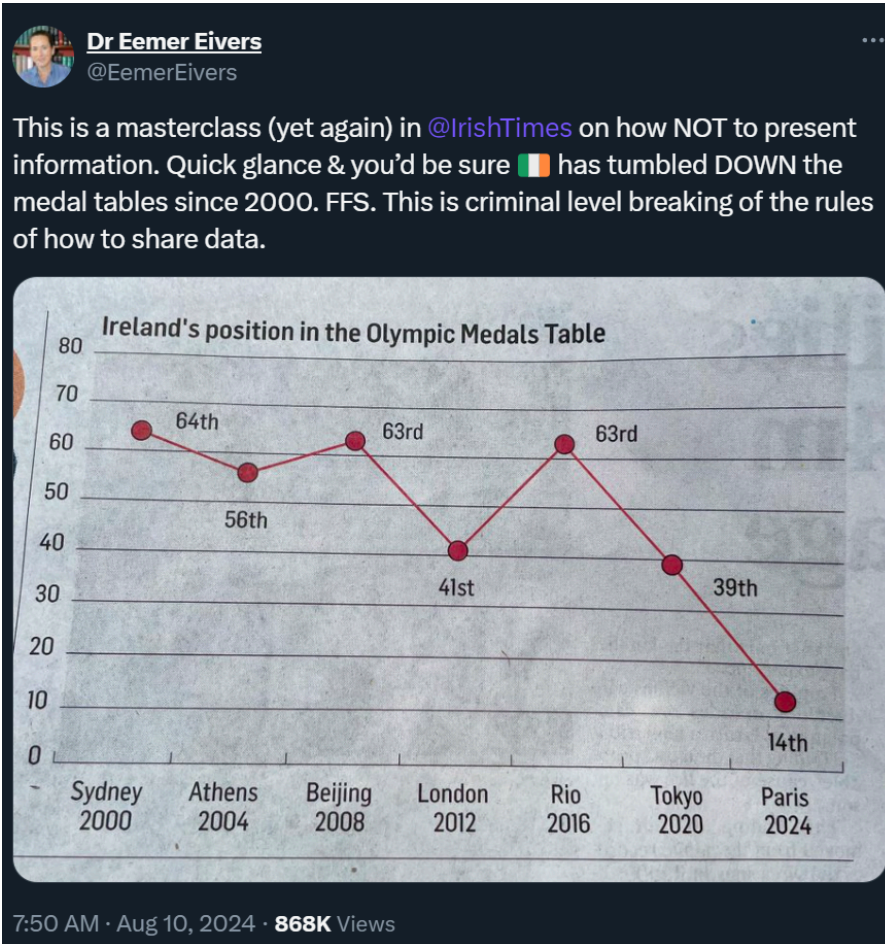
**No substantive issues**

**No perceptual issues**

**Honesty + good judgment**

**Kieran Healy, *Data Visualization: A Practical Introduction***

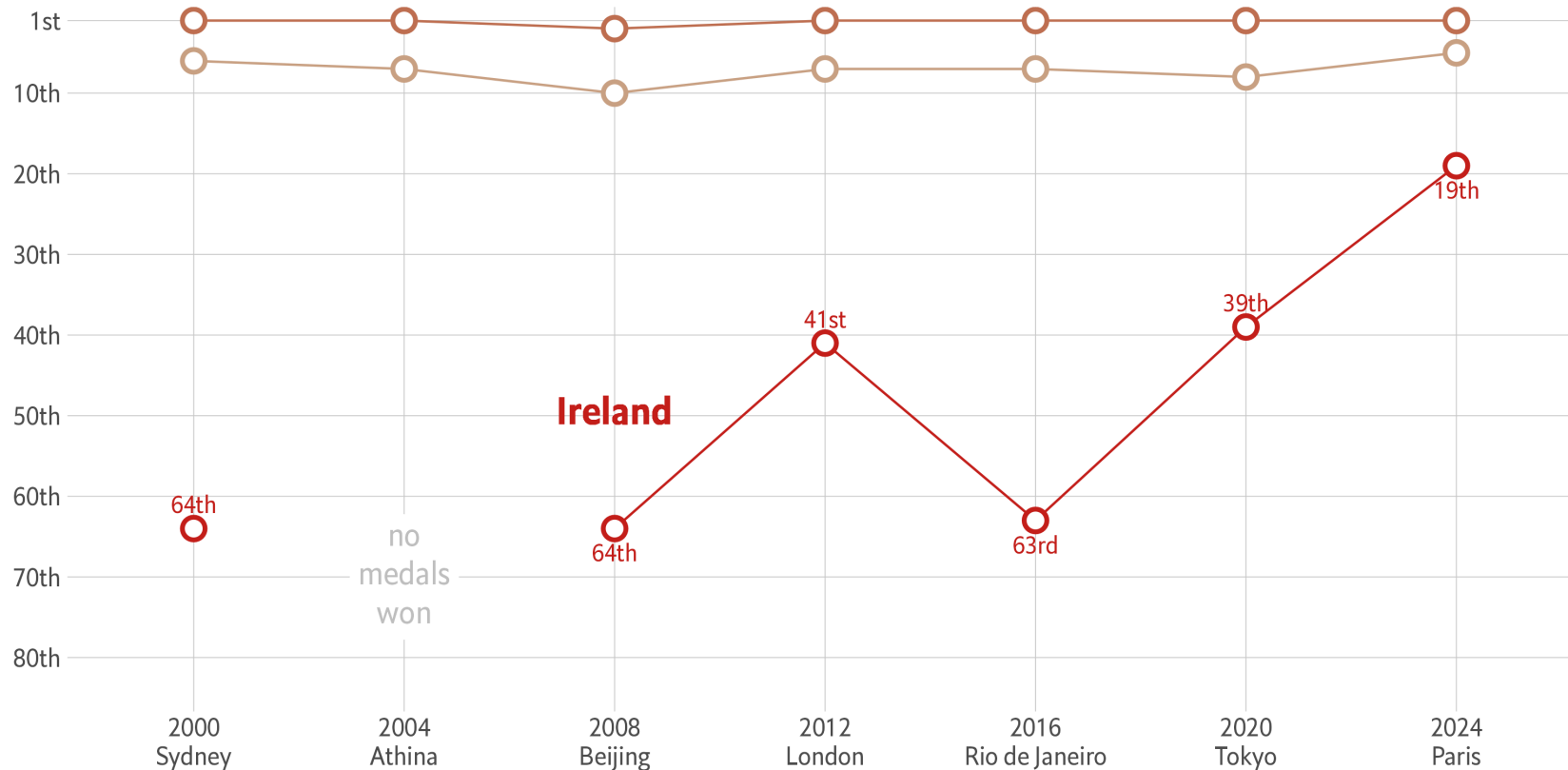
# You see bad plots everywhere: What's wrong?



# 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>  
  ) +  
  <COORDINATE_FUNCTION> +  
  <FACET_FUNCTION> +  
  <SCALE_FUNCTION> +  
  <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](#).

# Time to code



via GIPHY

# Our goal

## Trump Vote Share vs. College Education

Did Trump win the State? • No • Yes

