Eltecon Data Science Course by Emarsys

Data Visualization

Tamás Koncz

September 25, 2019

About me - Tamás

- Spent the last 6+ years of working with data daily one way or another
- 1 year mark @ Emarsys as a Data Scientist
- CEU MSc in Business Analytics
- reach me @ t.koncz@gmail.com
- Twitter, LinkedIn

About me - Peti

- Spent 1 year in Academia
- 2.5 yrs @ Emarsys
- Economics MSc in Amsterdam
- email: peter.lukacs@emarsys.com

Section 1

QUIZ TIME

Section 2

Communication as a Data Scientist

Why should you care?

- Data Science is a very complex, technical field
- But at the end we usually want to have an impact on the business
- Business people tend not to be technical
- Our impact as a data scientist depends on the decisions (human-made or automated) that we can influence.
- Communication is the tool to transfer the right ideas, and build *trust*
- You'll most frequently communicate using charts and other visualization tools

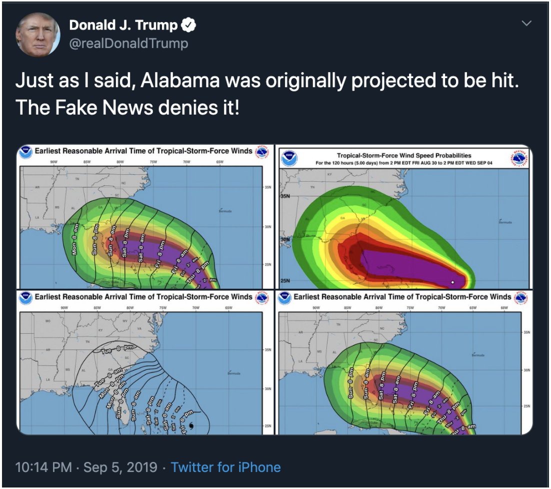
Let's see an example about Hurricanes

LIVE DEMO

OR

use `hurricane_dorian_forecast_map.pdf`...

And a tweet by Mr. Trump. . .



Donald J. Trump
@realDonaldTrump

Just as I said, Alabama was originally projected to be hit. The Fake News denies it!

Earliest Reasonable Arrival Time of Tropical-Storm-Force Winds
Tropical-Storm-Force Wind Speed Probabilities
For the 120 hours (5.00 days) from 2 PM EDT FRI AUG 30 to 2 PM EDT WED SEP 04
Earliest Reasonable Arrival Time of Tropical-Storm-Force Winds
Earliest Reasonable Arrival Time of Tropical-Storm-Force Winds

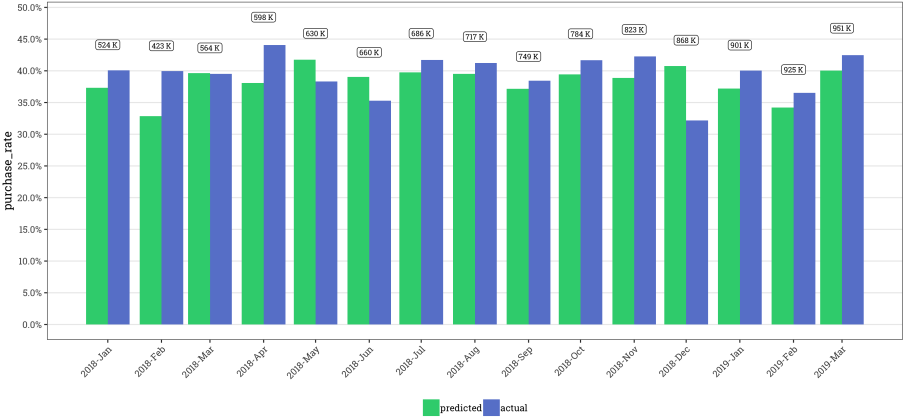
10:14 PM · Sep 5, 2019 · Twitter for iPhone

For more “fun” click here.

An example from Emarsys

Purchase Rate in AI Life Cycle Segments

Customer: brand_alley



Notes:
- Excludes contacts aquired during the month
- Includes first time and repeat buyers
- # K numbers above bars represent number of contacts

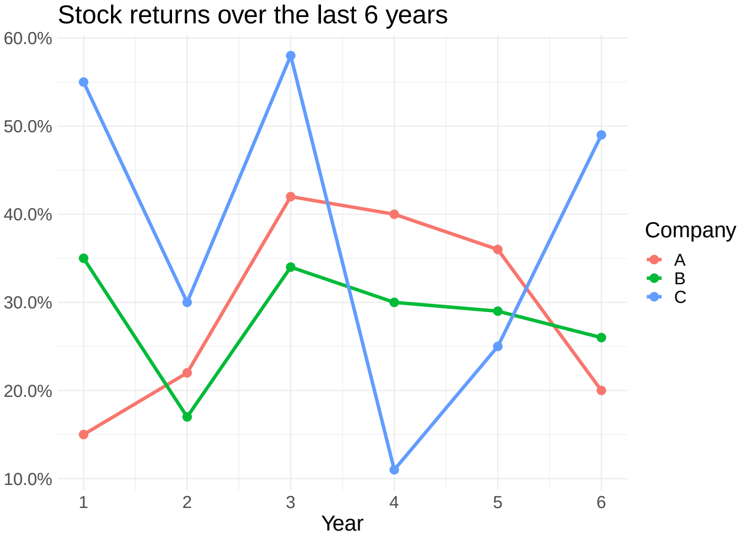
Section 3

Why does data visualization matter?

Tables vs charts

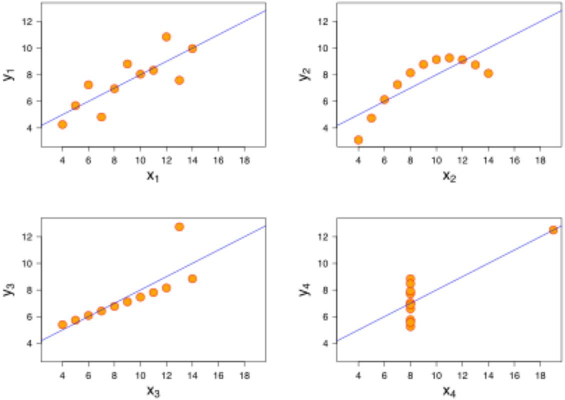
Year	A	B	C
1	0.15	0.35	0.55
2	0.22	0.17	0.30
3	0.42	0.34	0.58
4	0.40	0.30	0.11
5	0.36	0.29	0.25
6	0.20	0.26	0.49

Tables vs *charts*



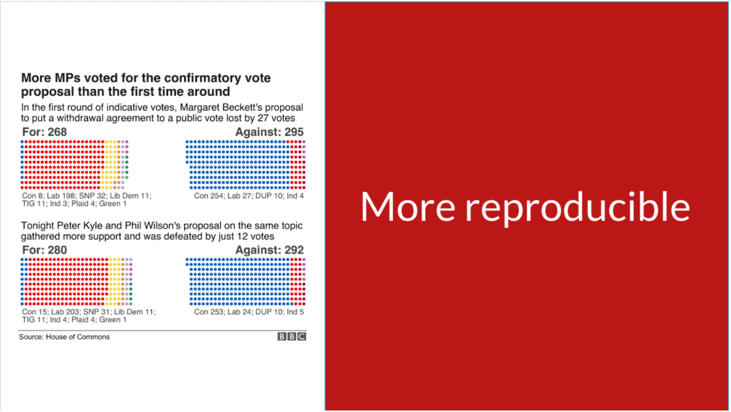
Anscombe's quartet

If it's about summarizing information, why are summary statistics insufficient? The below datasets have the same means, variances and correlations between X and Y.



Why R for data visualization?

- Reproducibility



More MPs voted for the confirmatory vote proposal than the first time around

In the first round of indicative votes, Margaret Beckett's proposal to put a withdrawal agreement to a public vote lost by 27 votes

Tonight Peter Kyle and Phil Wilson's proposal on the same topic gathered more support and was defeated by just 12 votes

Source: House of Commons

More reproducible

Source: EARL London, 2019 - How the BBC uses R for data visualisation

Why R for data visualization?

- Reproducibility
  - BBC example
  - Data wrangling is an important step we have to do
    - If it's done e.g. in Excel, the steps might not be replicable, or they just take time to do
- Fast iteration
  - Above also means that it is easy to change something,
  - Or visualize new data in "old" ways

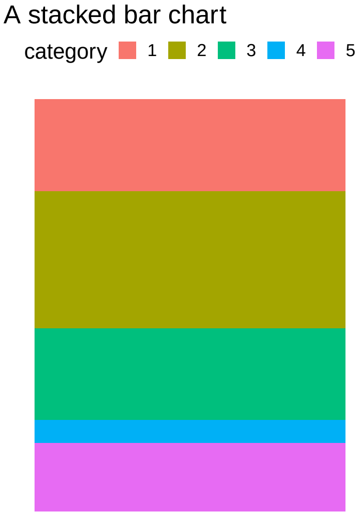
Explorative vs Descriptive Data Viz

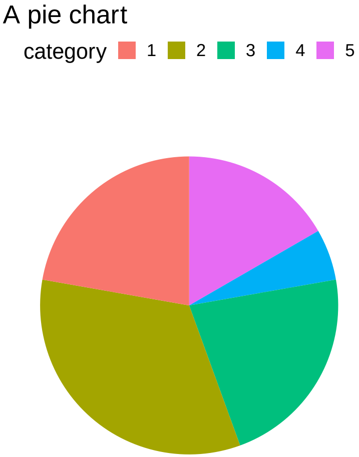
- Explorative: during research, getting to know the data
  - Interactivity! Specially when doing it for others
- Descriptive: summarizing findings, communication of results
  - Custom made
  - Know your audience: hard part. Others won't have the knowledge that you have. Right level of detail is also crucial.
  - Should show what we want it to show. Nothing more nothing less.
  - Usually 1 message / chart
  - Title, labels, etc. are all a MUST

Section 4

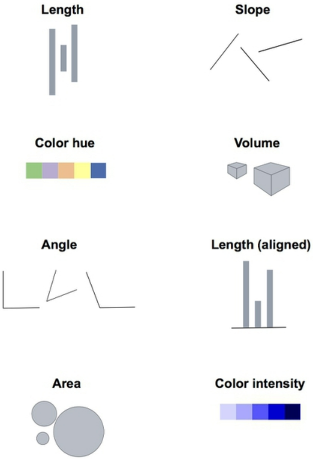
Visual Cues

Why we dislike pie charts?

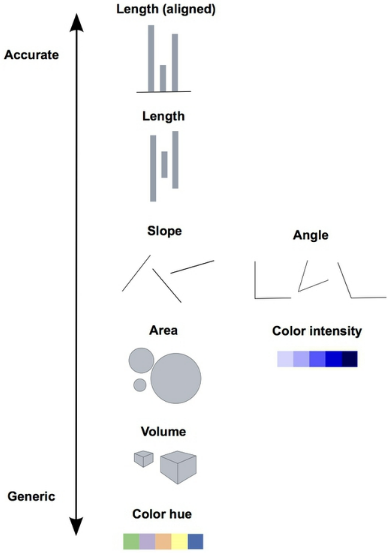




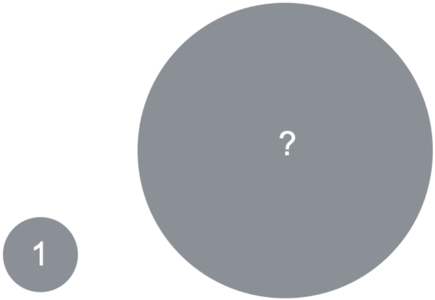
Perception of quantitative information



Perception of quantitative information

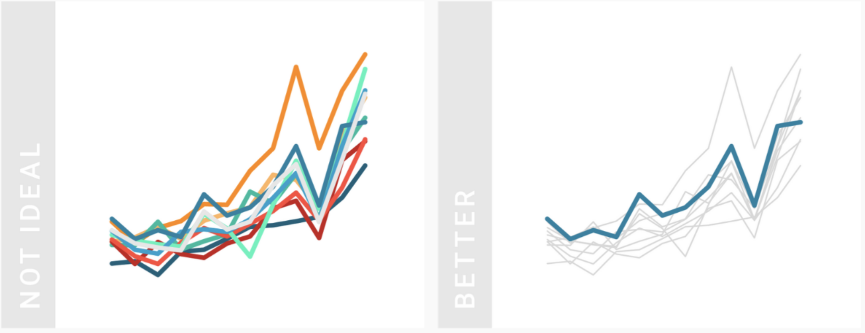


Test yourself



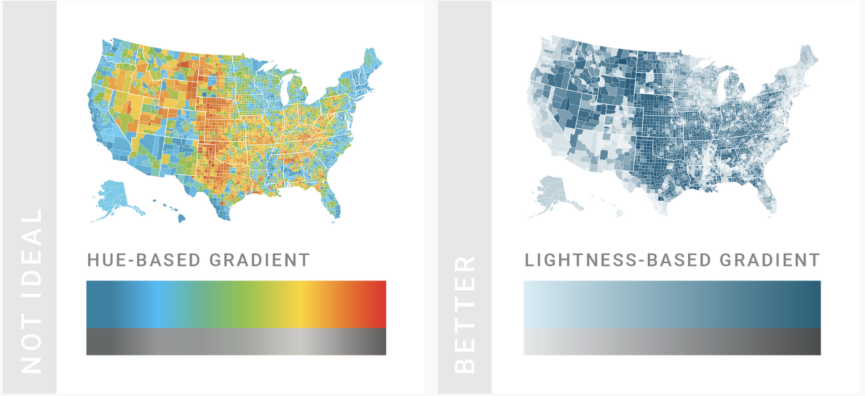
Source: Save the Pies for Dessert

About colors - highlighting



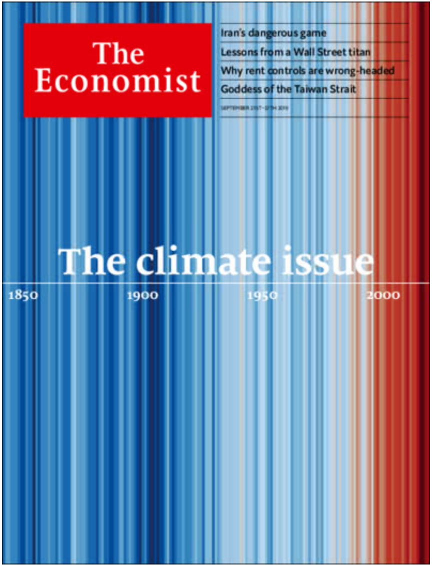
Source: What to consider when choosing colors for data visualization

About colors - hue - 1

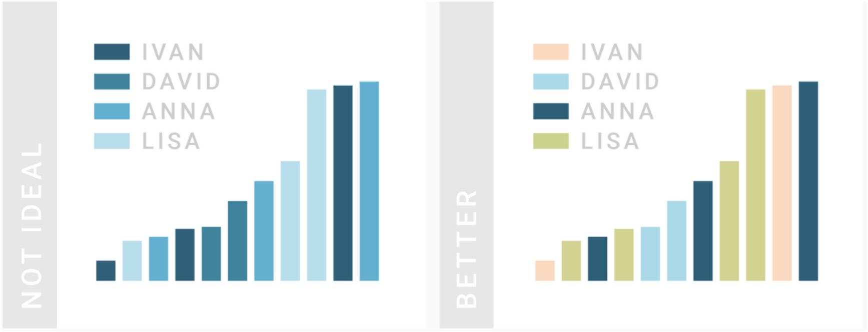


Source: What to consider when choosing colors for data visualization

About colors - hue - 1

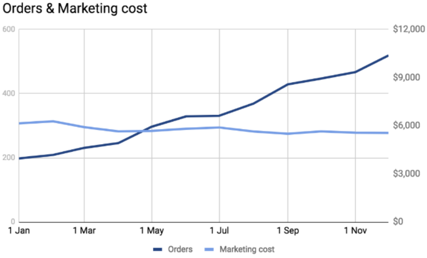

The Economist
Iran's dangerous game
Lessons from a Wall Street titan
Why rent controls are wrong-headed
Goddess of the Taiwan Strait
The climate issue
1850
1900
1950
2000

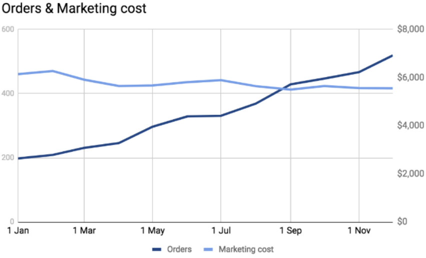
About colors - hue - 2



Source: What to consider when choosing colors for data visualization

Another pitfall - the double Y-axis trap





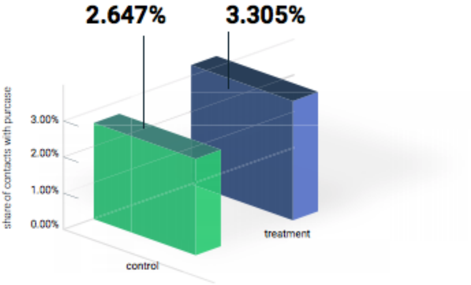
Source: Why you shouldn’t use pie charts - Tips for better data visualization

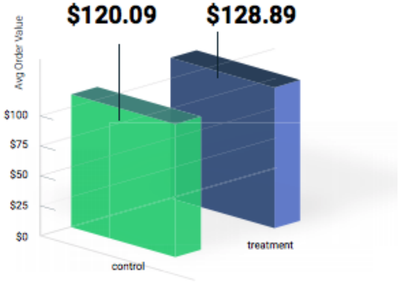
Section 5

Visualizing uncertainty

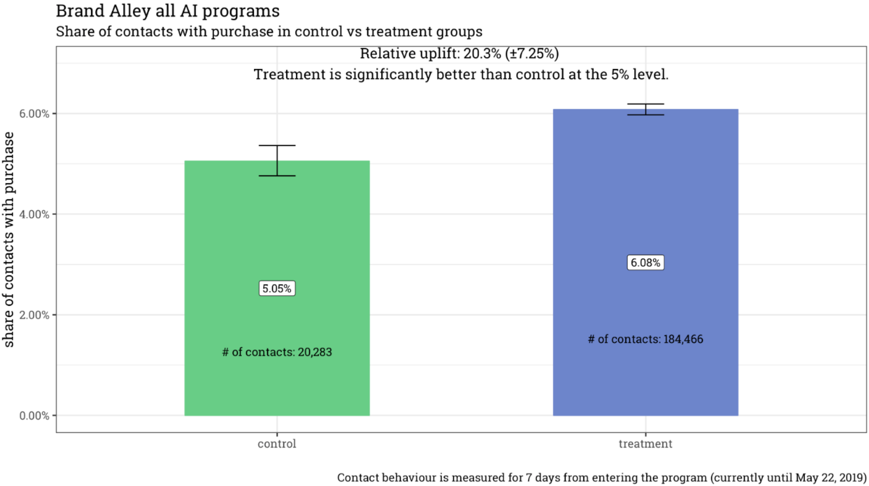
A recent example at Emarsys

List the bad (and good) things about these charts!

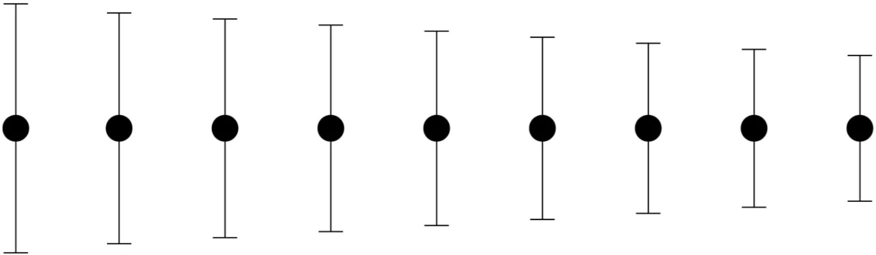




How we did it

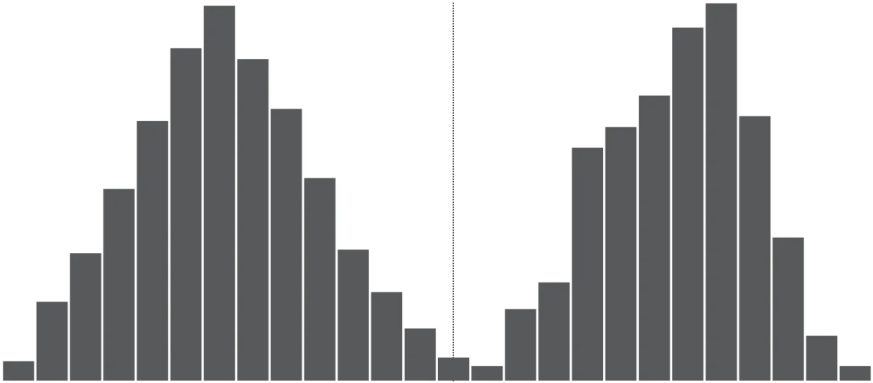

Brand Alley all AI programs
Share of contacts with purchase in control vs treatment groups
Relative uplift: 20.3% (±7.25%)
Treatment is significantly better than control at the 5% level.
share of contacts with purchase
6.00%
4.00%
2.00%
0.00%
5.05%
# of contacts: 20,283
6.08%
# of contacts: 184,466
control
treatment
Contact behaviour is measured for 7 days from entering the program (currently until May 22, 2019)

Uncertainty - Ranges



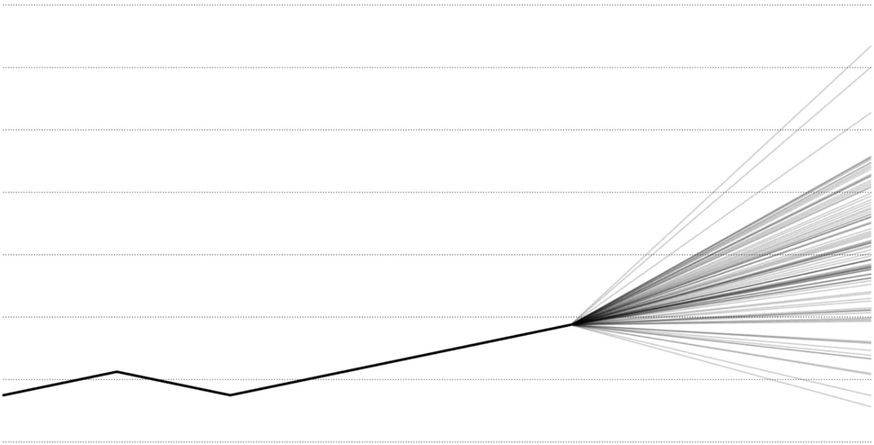
Source: Visualizing the Uncertainty in Data

Uncertainty - Distributions



Source: Visualizing the Uncertainty in Data

Uncertainty - Timeseries



Source: Visualizing the Uncertainty in Data

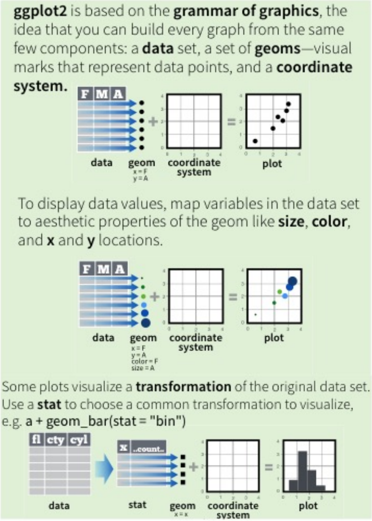
Section 6

ggplot & the grammar of graphics

Why ggplot2?

- Very mature, 10+ years in the making
- Enables fast in prototyping
- But also good enough in customization
- Great set of extensions
- Just get your data in the right format
- And then apply the “grammar of graphics”

Grammar of graphics



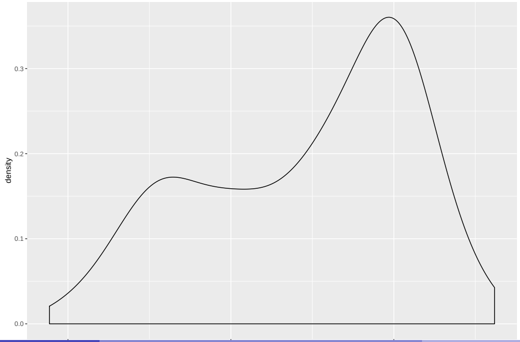
ggplot2 is based on the **grammar of graphics**, the idea that you can build every graph from the same few components: a **data** set, a set of **geoms**—visual marks that represent data points, and a **coordinate system.**

To display data values, map variables in the data set to aesthetic properties of the geom like **size**, **color**, and **x** and **y** locations.

Some plots visualize a **transformation** of the original data set. Use a **stat** to choose a common transformation to visualize, e.g. a + geom_bar(stat = "bin")

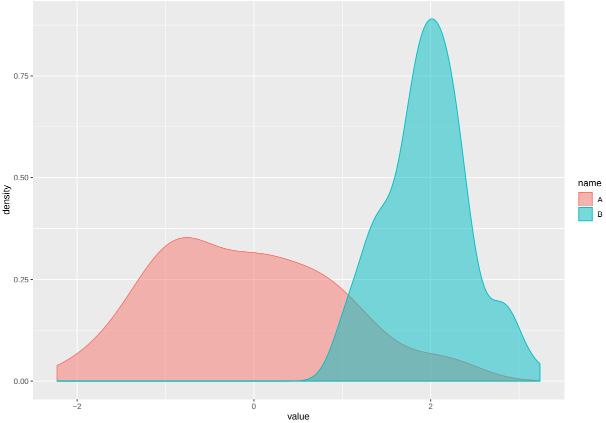
A minimal plot

```
set.seed(925)
dt <- data.table(
  name  = c(rep("A", 100),    rep("B", 100)),
  value = c(rnorm(100, 0, 1), rnorm(100, 2, 0.5))
)
ggplot(data = dt, mapping = aes(x = value)) +
  geom_density()
```



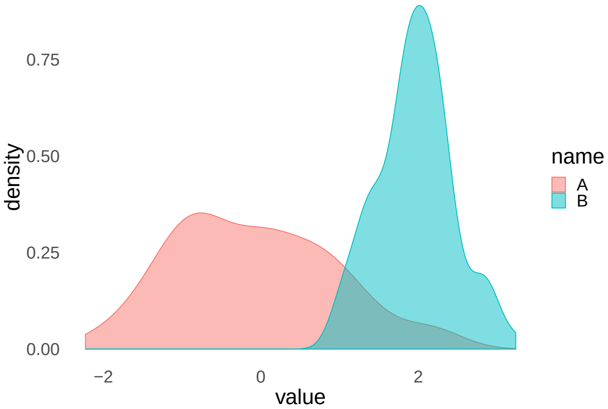
Let’s add one more aesthetic

```r
p <- ggplot(data = dt, mapping = aes(x = value)) +
  geom_density(aes(fill = name, color = name), alpha = 0.5)
p
```



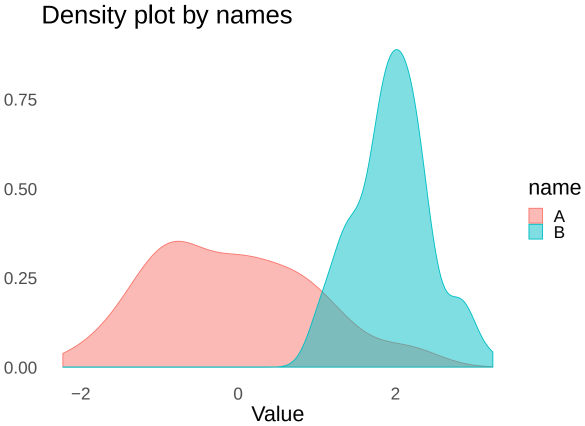
Apply some formatting

```
p <- p + theme_minimal() +
  theme(panel.grid = element_blank(), text = element_text(size = 25))
p
```



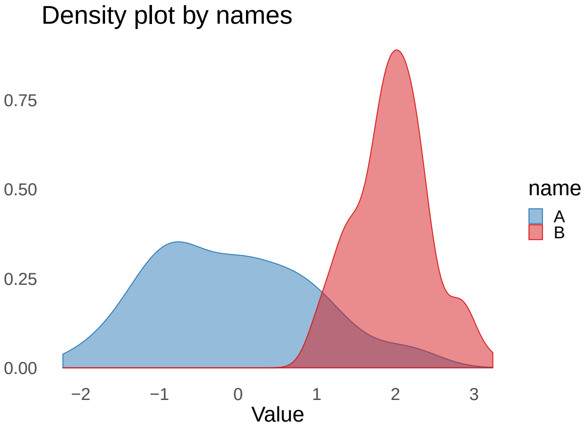
Add annotation

```
p <- p + labs(
  title = "Density plot by names", x = "Value", y = ""
)
p
```



Fix scales

```r
p + scale_x_continuous(breaks = c(-2:3)) +
  scale_color_manual(values = c("A" = "#2c7bb6", "B" = "#d7191c")) +
  scale_fill_manual(values = c("A" = "#2c7bb6", "B" = "#d7191c"))
```



Some useful resources

- RStudio ggplot2 cheatsheet
- Hadley Wickham: ggplot2: Elegant Graphics for Data Analysis
- https://www.r-graph-gallery.com/
- https://coolors.co/app
- http://colorbrewer2.org/