


# DATA SCIENCE

## PPGla/PUCPR

Prof. Jean Paul Barddal



1

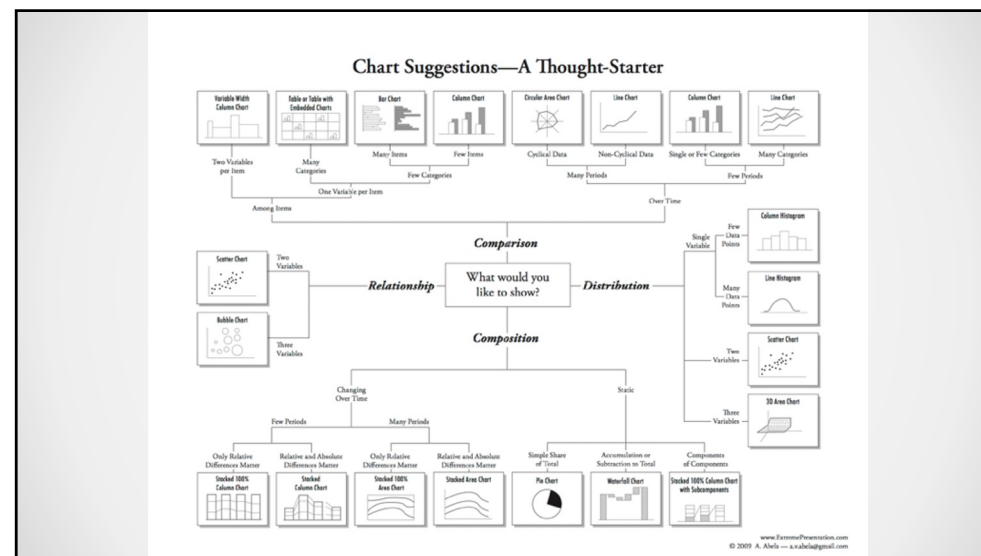
# BIVARIATE PLOTS

2

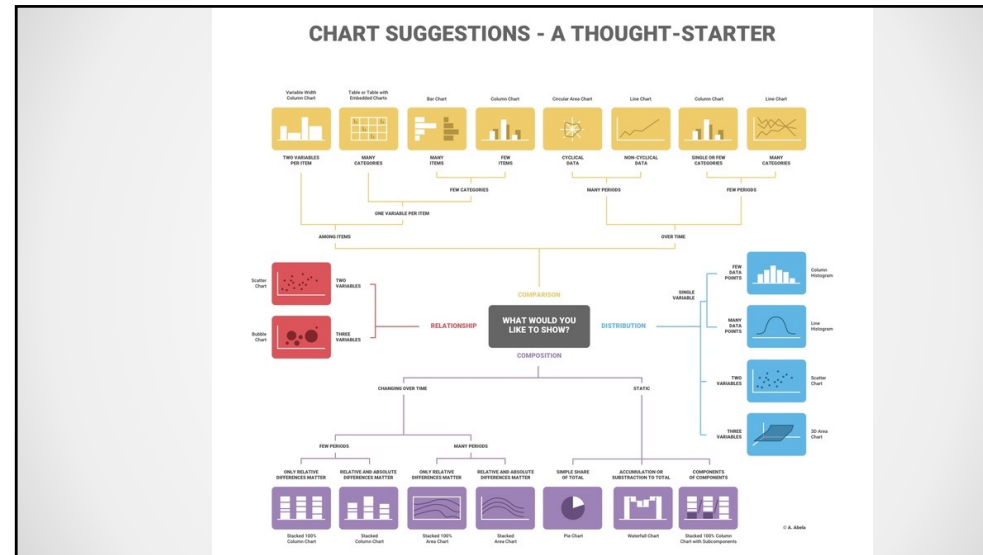
## Bivariate plots

- Main goal: verify the existence of relationships in two variables
- The definition of which type of visualization we need to build depends on the variables we are using
  - Scatterplots: two numeric variables
  - Line plots: two numeric variables
  - Box-plots and violin plots: categorical and numeric variables
  - Heatmaps: two numeric variables or one categorical and one numeric

3



4



5

## Important links

- **Datavizcatalogue**
  - <https://datavizcatalogue.com>
  - “Search by function” option
- **Data-to-viz**
  - <https://www.data-to-viz.com>
- **Python Graph Gallery**
  - <https://python-graph-gallery.com>

### What do you want to show?

Here you can find a list of charts categorised by their data visualization functions or by what you want a chart to communicate to an audience. While the allocation of each chart into specific functions isn't a perfect system, it still works as a useful guide for selecting chart based on your analysis or communication needs.

Comparisons

Proportions

Relationships

Hierarchy

Concepts

Location

Part-to-a-whole

Distribution

How things work

Processes & methods

Movement or flow

Patterns

6

### Overplotting, transparency, and jitter

- When creating multivariate plots, it is common for us to have so many data points cluttered in the same region
- This is called **overplotting**, and it prevents us from analyzing the data properly
- A few ideas to handle overplotting:

Marker size

Transparency

Density

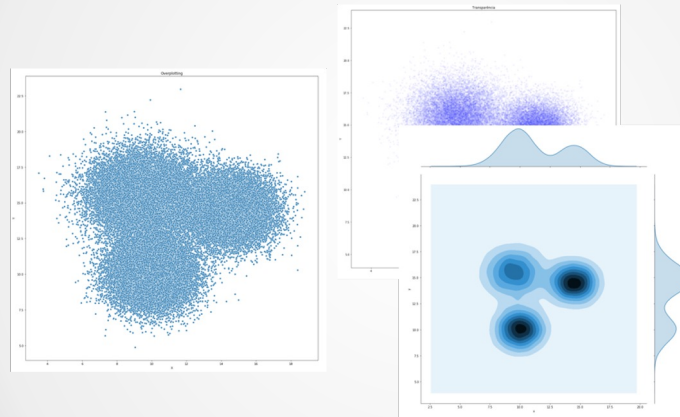
Sampling

Filtering

Clustering

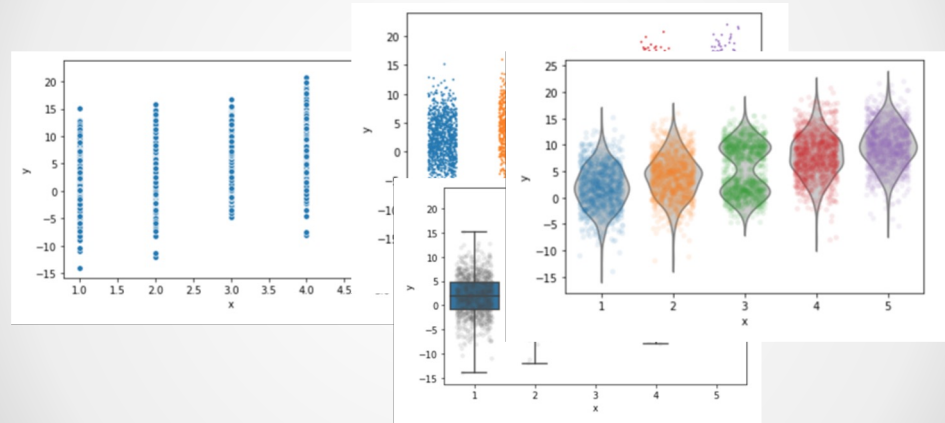
7

### Example - numeric variables



8

### Example - categorical and numeric variables



9

**DESIGN CHOICES**

10

## Overplotting

- Issue that arises when there are many data points that share the same region in the plot
- Ways to overcome overplotting:
  - *Jitter*
  - *Marker size*
  - *Transparency*
  - *Density*
  - *Sampling*
  - *Filtering*
  - *Clustering*

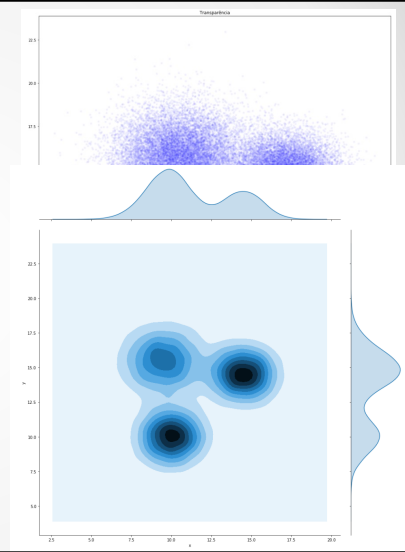
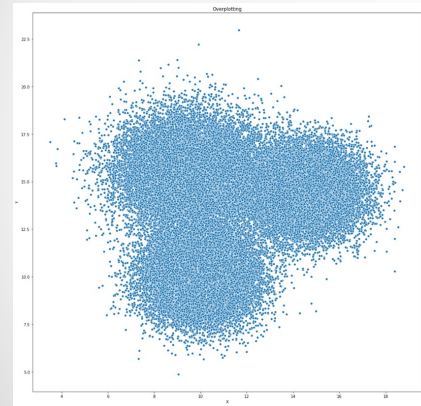
11

## Hands-on

- Let's code of the examples of bivariate plots
- And let's also overcome overplotting using the techniques mentioned earlier.

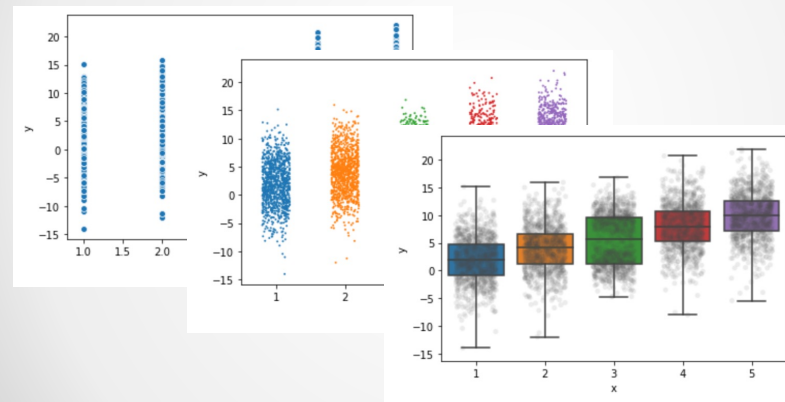
12

## Overplotting examples



13

## Numeric variable vs. categorical variable



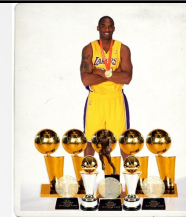
14

## ACTIVITY

36

### Back to Kobe's shots

- It is time to conduct a data analysis
- Try to work as follows:
- State an hypothesis/question about the data
- Analyze and plot the data
- Discuss the visualization, either by corroborating or invalidating the hypothesis



37