HERTIE'S DATA SOCIETY

Exploratory Data Analysis (EDA)

Golo
21.10.2019

Today's agenda

WHAT IS EDA?
WHY DO EDA?
AN EDA CHECKLIST + PRACTICE
VISUALIZATION WITH GGPLOT + PRACTICE

What is EDA?

EXPLORATORY DATA ANALYSIS IS...

... an approach to analyzing data sets to summarize their main characteristics, often with visual methods. It is a creative process with no strict rules.

EDA IS AN ITERATIVE CYCLE. YOU...

1. Generate questions about your data.
2. Search for answers by visualising, transforming, and modelling your data.
3. Use what you learn to refine your questions and/or generate new questions.

EDA ALLOWS YOU TO...

... better understand your data, build an intuition about your data, generate hypothesis and find insights

Why do EDA?

YOU NEED TO UNDERSTAND YOUR DATA AND KNOW IF YOUR DATA MEETS YOUR EXPECTATIONS OR NOT

Let us ask Marina why EDA is important

An EDA checklist

1. WHAT QUESTION ARE YOU TRYING TO SOLVE (OR PROVE WRONG)?
2. WHAT KIND OF DATA DO YOU HAVE?
3. WHAT'S MISSING FROM THE DATA AND HOW DO YOU DEAL WITH?
4. WHERE ARE THE OUTLIERS AND WHY SHOULD PAY ATTENTION TO THEM?
5. HOW CAN YOU ADD, CHANGE OR REMOVE FEATURES TO GET MORE OUT OF YOUR DATA?

Most importantly...

WHAT TYPE OF VARIATION OCCURS WITHIN MY VARIABLES?

WHAT TYPE OF COVARIATION OCCURS BETWEEN MY VARIABLES?

STAY OPEN AND SCEPTICAL!

Let's practice

TITANIC: MACHINE LEARNING FROM DISASTER

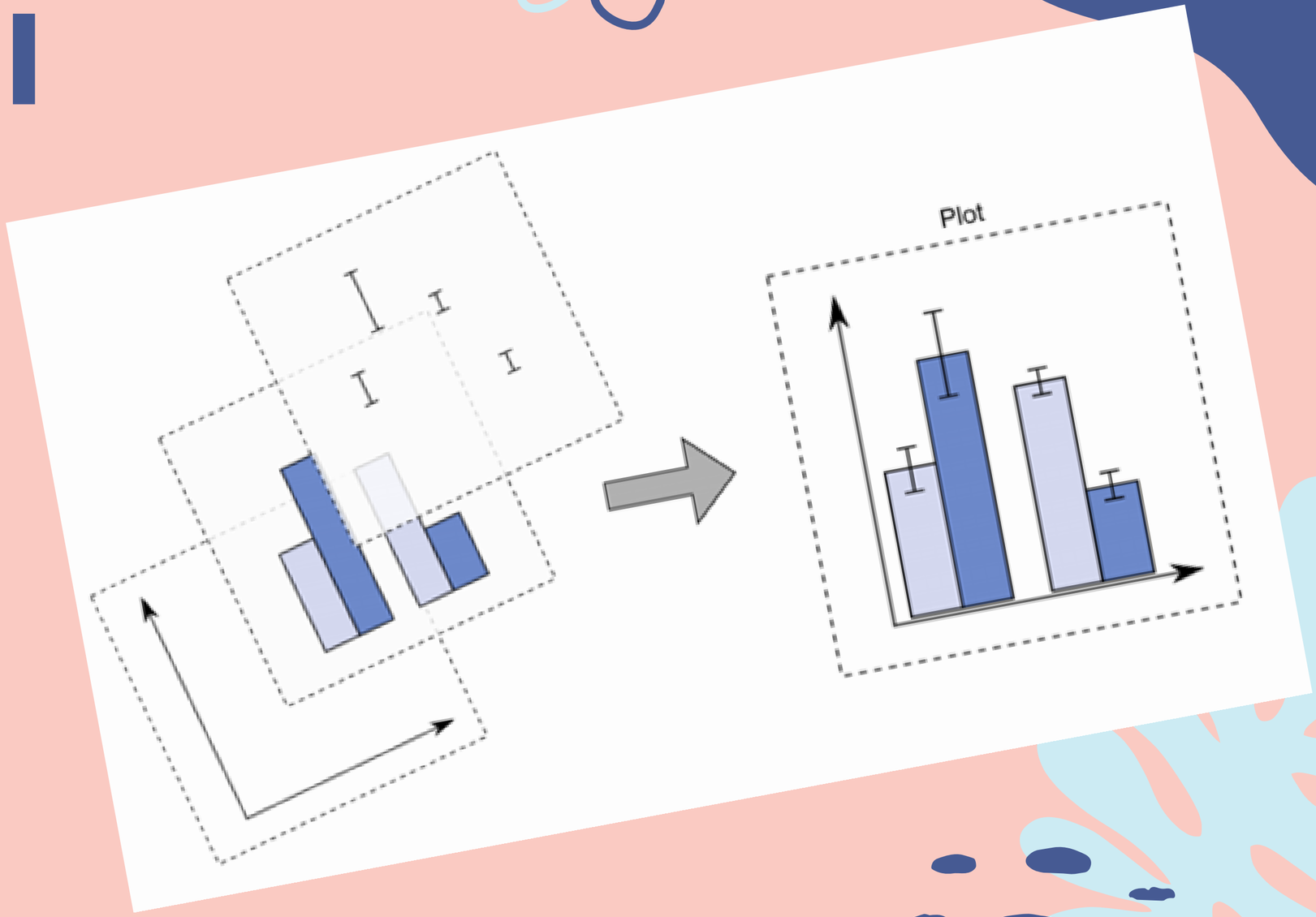
HTTPS://WWW.KAGGLE.COM/C/TITANIC

The Grammar of Graphics I

IN GGPLOT, A GRAPH IS MADE UP OF A SERIES OF LAYERS.

CHEAT SHEET:

HTTPS://RSTUDIO.COM/WP-CONTENT/UPLOADS/2015/03/GGPLOT2-CHEATSHEET.PDF

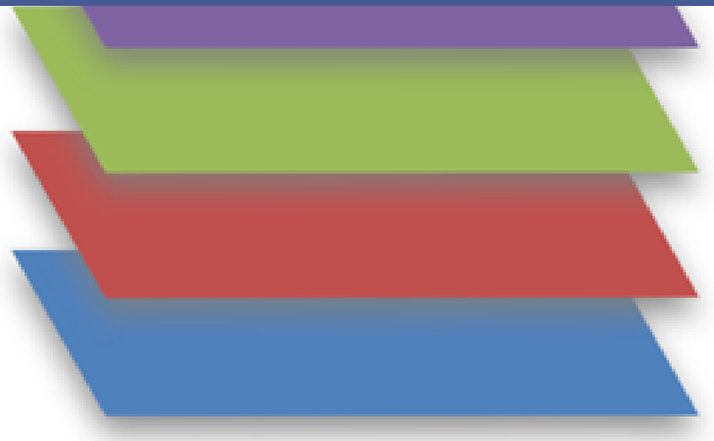


The Grammar of Graphics II

Describes all the non-data ink	Theme
Plotting space for the data	Coordinates
Statistical models & summaries	Statistics
Rows and columns of sub-plots	Facets
Shapes used to represent the data	Geometries
Scales onto which data is mapped	Aesthetics
The actual variables to be plotted	Data

The Grammar of Graphics II cont.

Shapes used to represent the data	Geometries
Scales onto which data is mapped	Aesthetics
The actual variables to be plotted	Data



AESTHETICS:

- Controls appearance (looks) and location (defines what will be on which axis)
- aes(x= var1, y= var2, colour/fill/size/etc.=var3)

GEOMETRIES:

- Scatterplot: geom_point()
- Histogram: geom_histogram()
- Barplot: geom_bar()
- Boxplot: geom_boxplot()
- Density: geom_density()
- Adding an abline: geom_smooth(model = lm)
- Or a combination of multiple.