Exploratory data analysis in big data using R

Catia Nicodemo
University of Oxford, CHSEO

PHCS, Oxford

8 February 2018

Data narrative

According to Bit.ly's Hillary Maso, data scientists generally do three fundamentally different things:

1. math
2. code
3. COMMUNICATE:
   'a goal of data visualization is to communicate abstract concepts that emerge from the word of math and metrics using the more human language of spatial representation' (in Manoochehri , 2014)

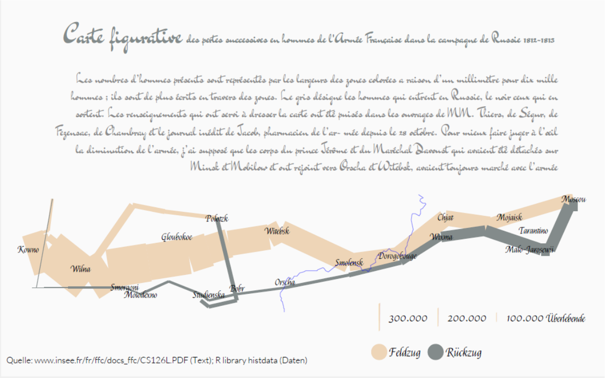
The research question

. . . (we believe) more important than the data itself

Graphs, engine for **communication**

The best graph in history!

In M. Manoochehri (2014)'s book "Data Just Right"



Carte figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812-1813

Les nombres d'hommes présents sont représentés par les largeurs des zones colorées a raison d'un millimètre pour dix mille hommes ; ils sont de plus écrits en travers des zones. Le gris désigne les hommes qui entrent en Russie, le noir ceux qui en sortent. Les renseignements qui ont servi à dresser la carte ont été puisés dans les ouvrages de MM. Thiers, de Ségur, de Fezensac, de Chambray et le journal inédit de Jacob, pharmacien de l'ar- mée depuis le 28 octobre. Pour mieux faire juger à l'œil la diminution de l'armée, j'ai supposé que les corps du prince Jérôme et du Maréchal Davoust qui avaient été détachés sur Minsk et Mobilow et ont rejoint vers Orscha et Witebsk, avaient toujours marché avec l'armée

Quelle: www.insee.fr/fr/ffc/docs_ffc/CS126L.PDF (Text); R library histdata (Daten)

Figure 1: Charles Joseph Minard's 1869 work Carte figurative des pertes successives en hommes de l'Armee Francaise dans la campagne de Rusie 1812-1813

Politics in Catalunya: a simple graph

SCC: Catalanes, La Catalunya inmune al procés.

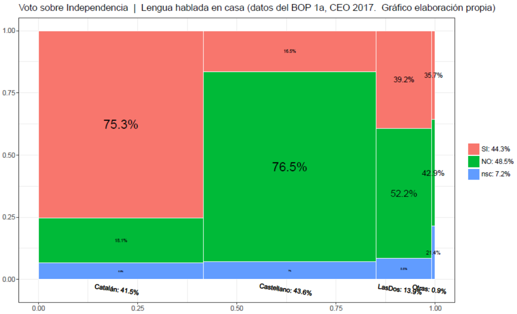

Voto sobre Independencia | Lengua hablada en casa (datos del BOP 1a, CEO 2017. Gráfico elaboración propia)

Figure 2: Fracture in Catalunya?

Elaborated from survey data of CEO (data collected in July 2017)

EDA (Tukey, 1977)

"**Exploratory data analysis is detective work**
counting detective work
graphical detective work. " (Tukey, 1977, p. 1)

EDA in big data

- Effective data visualization is an important tool for statistical analysis, it helps to convey in a direct way essential aspects of the data.
- New challenges arise in data visualisation when involving big data, the standard approaches for graphing get cluttered by an excess of data points.
- On the contrary, graphs that are uselless for small data , become highly informative when sample size is large (e.g., the boxplot)

This talk discusses changes on classical EDA graphs when involving Big Data

Data Sciences: Programming + Statistics

- Data collection, data storage, data retrieving: Task of computer Sciences
- Information retrieving: . . . Statistics!
  - All the Data, Summaries, graphs, . . . (Descriptive statistics for a population)
  - Incomplete Data, sample bias
  - Big Data, reducing dimension, clustering cases, outlier detection . . . (descriptive multivariate analysis)
  - Small Data, inferences (when) random sampling
  - Inexact Data, measurement error, indicators of latent variables: inferring true values, . . . (test scoring)
- Statistical software (we have a programming tool, open, free software, **written by statisticians!** : `R`

  The R Project for Statistical Computing

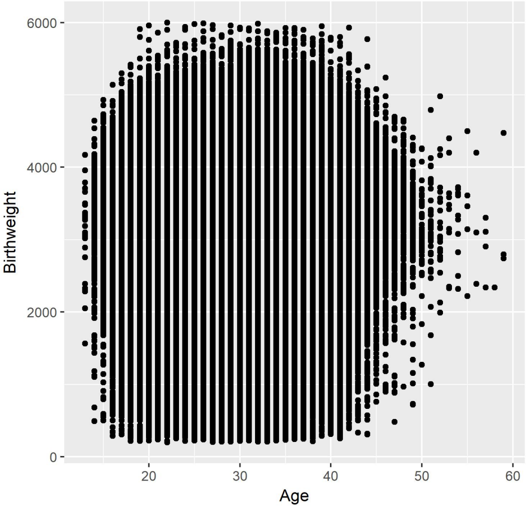
  . . . we use `ggplot`

EDA for Health Economics

Maternity Infant Health Outcomes and Unemployment Rate

- Hospital Episode Statistics Data: Maternity Data in 2011.
- We have all the baby born in the public hospital in 2011, around 500,000

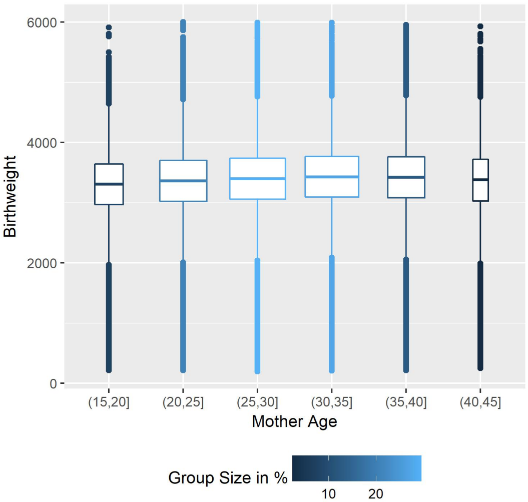
Scatterplot of Birthweight | age



Source: HES 2011

Boxplot of Birthweight | age-groups

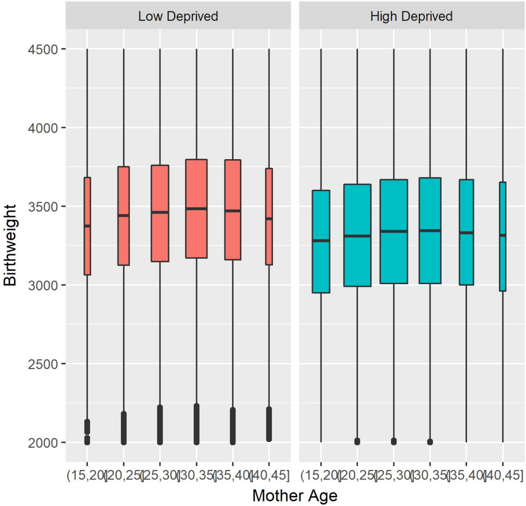
Note: width of the box proportional to group size!



Footnote: The wide is proportional to the size of age group. Source: Hospital Data

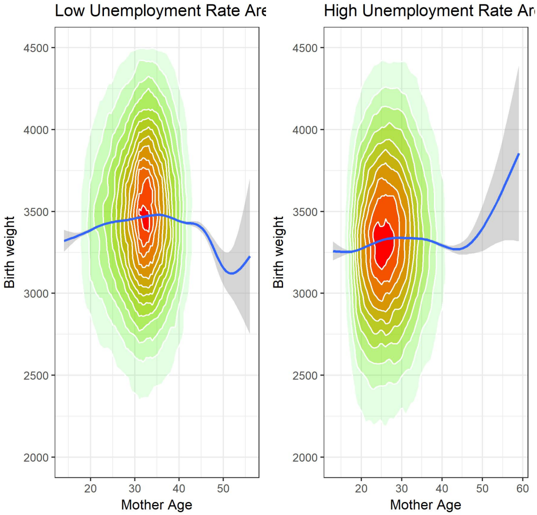
Boxplot of Birthweight | deprivation Index

Note: width of the box proportional to group size!



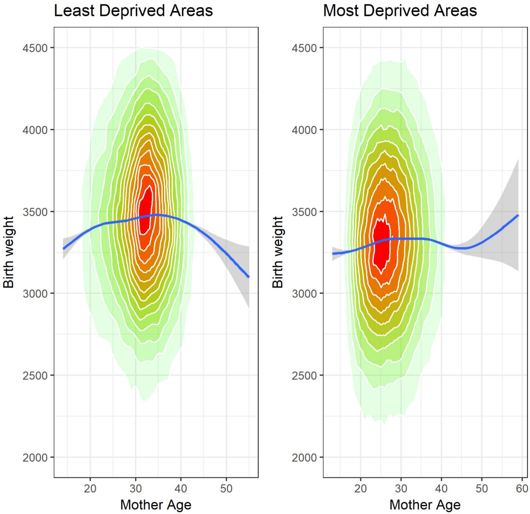
Footnote: The wide is proportional to the size of age group. Source: Hospital Data

Heat-maps of Birthweight | unemployment



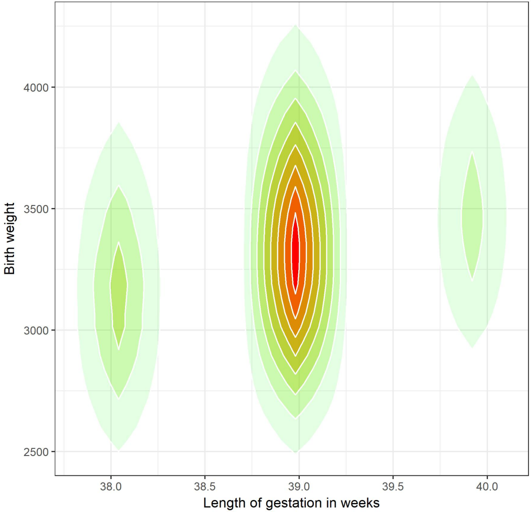
Source: Hospital Data 2011

Heat-maps of Birthweight | deprivation Index



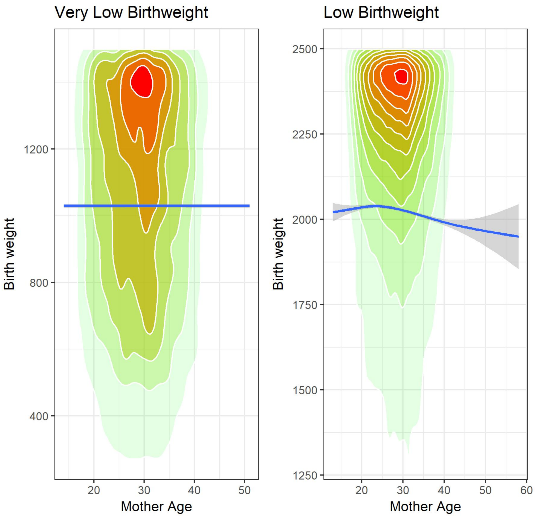
Source: Hospital Data 2011

Heat-maps of Birthweight | length of pregnancy



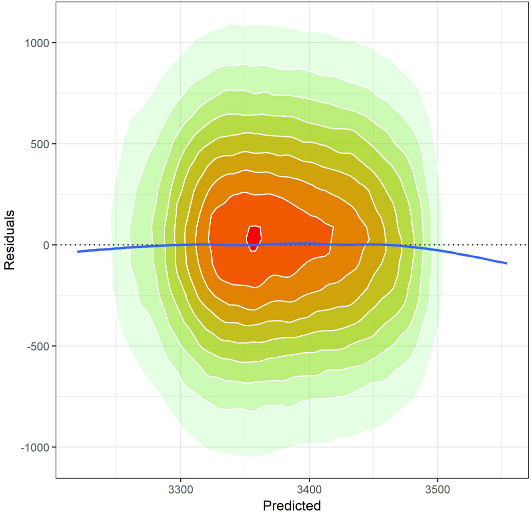
Source: Hospital Data 2011

Heat-maps of Birthweight | low Birthweight



Source: Hospital Data 2011

Heat-maps of Birthweight | residual vs fitted values



Source: MCVL 2010

Thank you!

catia.nicodemo@economics.ox.ac.uk