Exploratory data analysis

David W. Body

2019-05-01 (updated: 2019-07-14)

Exploratory data analysis

Outline

- Iterate
  - Questions
  - Visualize, transform, model
  - Refine questions / generate new questions
- Variation and covariation
  - Continuous variables
  - Categorical variables
- More about ggplot2

Data science workflow

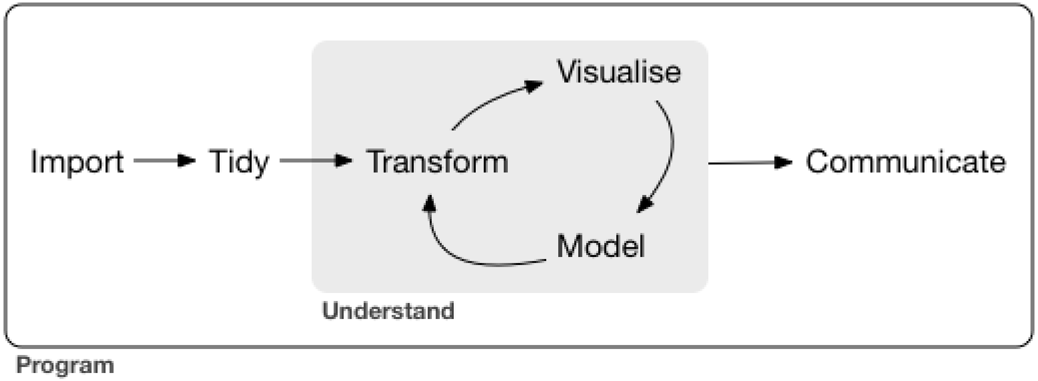


Image source: R for Data Science by Hadley Wickham & Garrett Grolemund.

Data science workflow

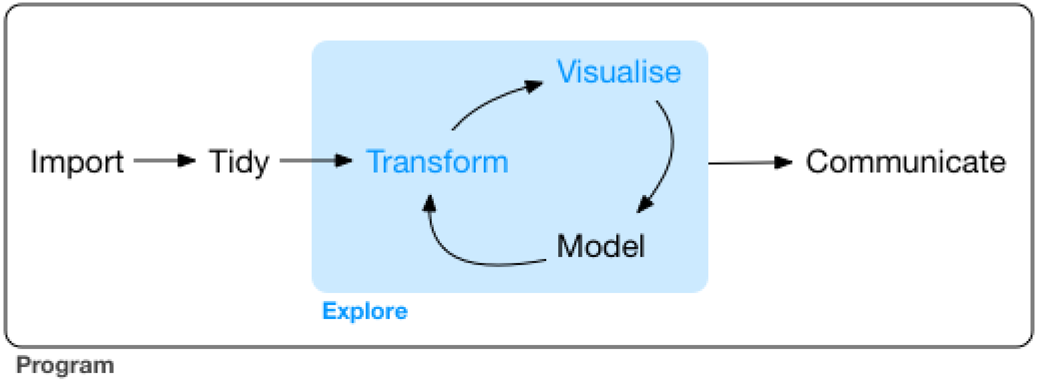
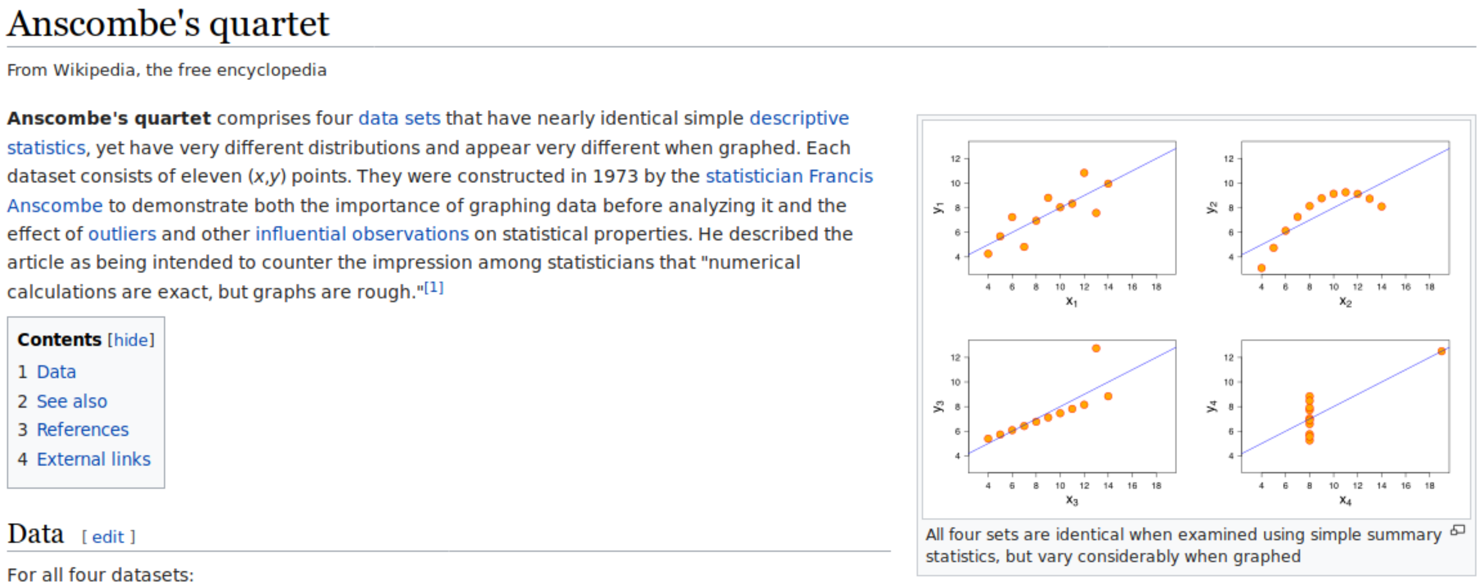


Image source: R for Data Science by Hadley Wickham & Garrett Grolemund.

EDA: Visualization is critical

Classic example: Anscombe's Quartet



Anscombe's quartet

From Wikipedia, the free encyclopedia

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (*x*,*y*) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers and other influential observations on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough."[1]

Contents [hide]

Data [edit]

For all four datasets:

All four sets are identical when examined using simple summary statistics, but vary considerably when graphed

Demo

Questions

Goal of EDA is to develop **understanding** of our data.

Standard approach is to use **questions** to guide investigation.

Specific questions will depend on the dataset, but generically

1. What type of **variation** occurs within my variables?
2. What type of **covariation** occurs between my variables?

Terminology

From chapter 7:

> - A **variable** is a quantity, quality, or property that you can measure.
>
> - A **value** is the state of a variable when you measure it. The value of a variable may change from measurement to measurement.
>
> - An **observation** is a set of measurements made under similar conditions (you usually make all of the measurements in an observation at the same time and on the same object). An observation will contain several values, each associated with a different variable. We'll sometimes refer to an observation as a data point.
>
> - **Tabular data** is a set of values, each associated with a variable and an observation. Tabular data is **tidy** if each value is placed in its own "cell", each variable in its own column, and each observation in its own row.

Variation

Variation is the tendency of the values of a variable to change from measurement to measurement.

Sources of variation include differences between subjects and measurement errors.

Types of variables

A variable is **categorical** if it can take one of a small set of values.

A variable is **continuous** if it can take any of an infinite set of ordered values.

We will use slightly different techniques to look at the variation of categorical and continous variables.

Demo

Covariation

Covariation is the tendency for the values two or more variables to vary in a related way.

Best way to find covariation is the **plot the variables**.

We'll consider the following cases:

- A categorical and continous variable
- Two categorical variables
- Two continuous variables

Demo

Your turn

Exploratory Data Analysis

your-turn/03-exploratory-data-analysis.Rmd

15:00