MATH411 | Fall 2018 | Chapter 2: Exploratory Data Analysis

Dr. Yongtao Cao

September 10, 2018

Contents

1.1 Data **1**

1.2 Exploratory Data Analysis **5**

1.3 Examples **9**

1.1 Data

Data is anything that has been recorded.

- What is a data science project?

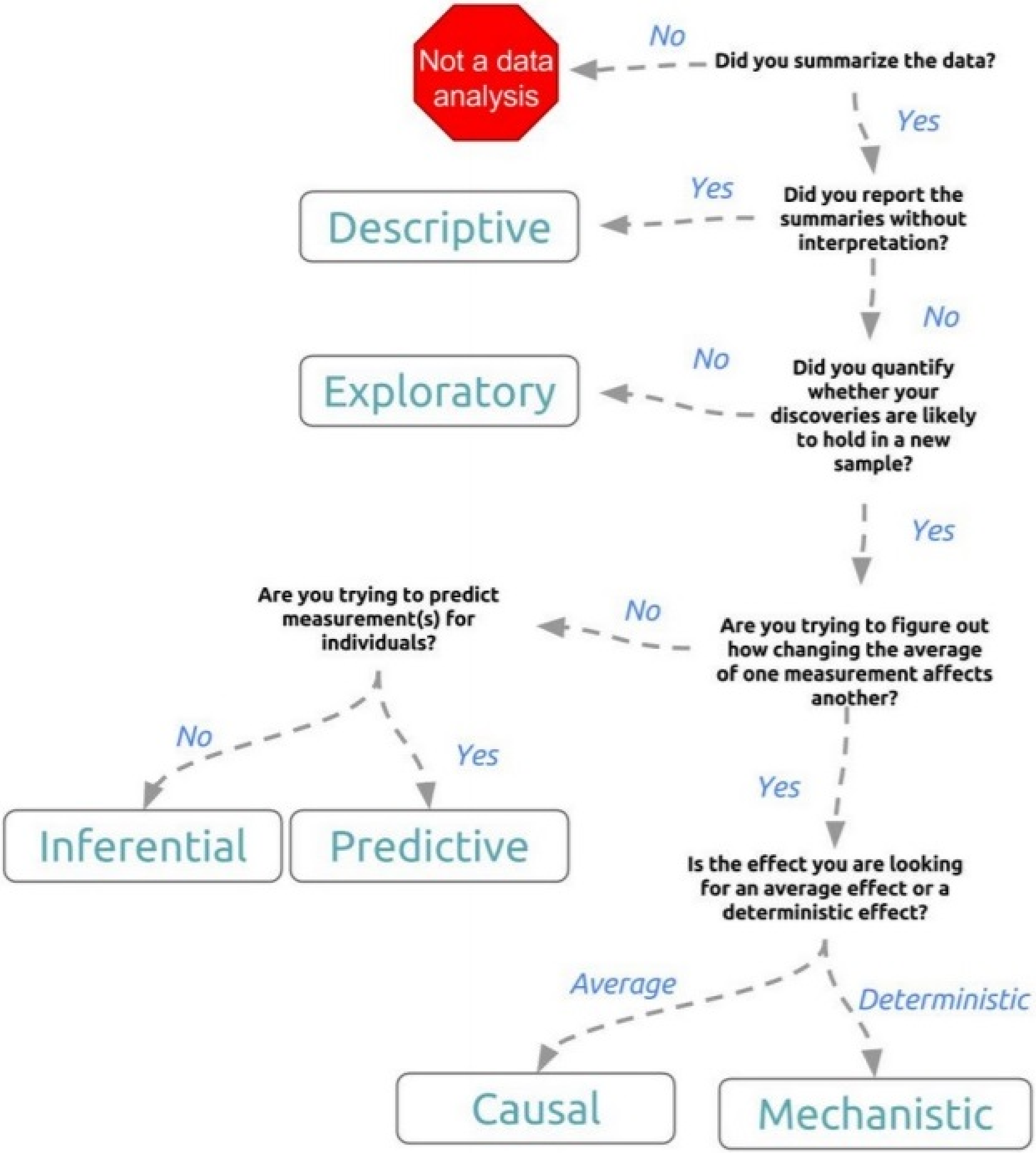

Not a data analysis
No
Did you summarize the data?
Yes
Descriptive
Yes
Did you report the summaries without interpretation?
No
Exploratory
No
Did you quantify whether your discoveries are likely to hold in a new sample?
Yes
Are you trying to predict measurement(s) for individuals?
No
Are you trying to figure out how changing the average of one measurement affects another?
No
Yes
Yes
Inferential
Predictive
Is the effect you are looking for an average effect or a deterministic effect?
Average
Deterministic
Causal
Mechanistic

Figure 1: Types of Data Analysis

- What is the typical data analysis workflow?

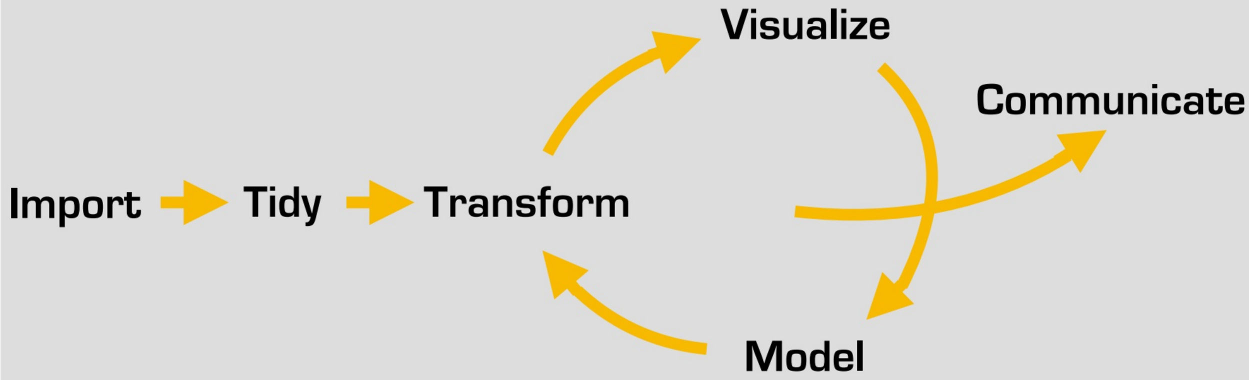
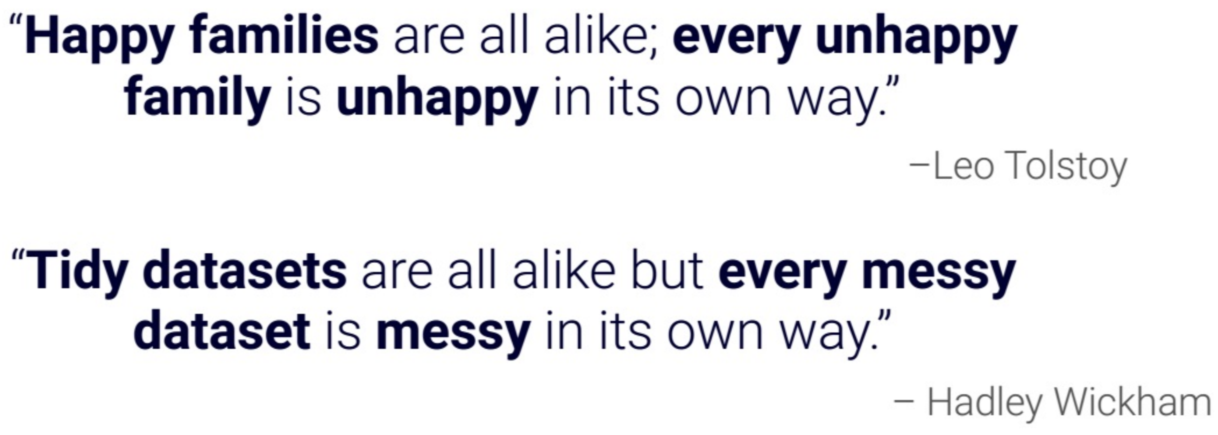


Figure 2: Data Project Workflow

1.1.1 Tidy Data

Why do we need tidy data?



“**Happy families** are all alike; **every unhappy family** is **unhappy** in its own way.”

–Leo Tolstoy

“**Tidy datasets** are all alike but **every messy dataset** is **messy** in its own way.”

– Hadley Wickham

Figure 3: Why Tidy Data

What is tidy data?

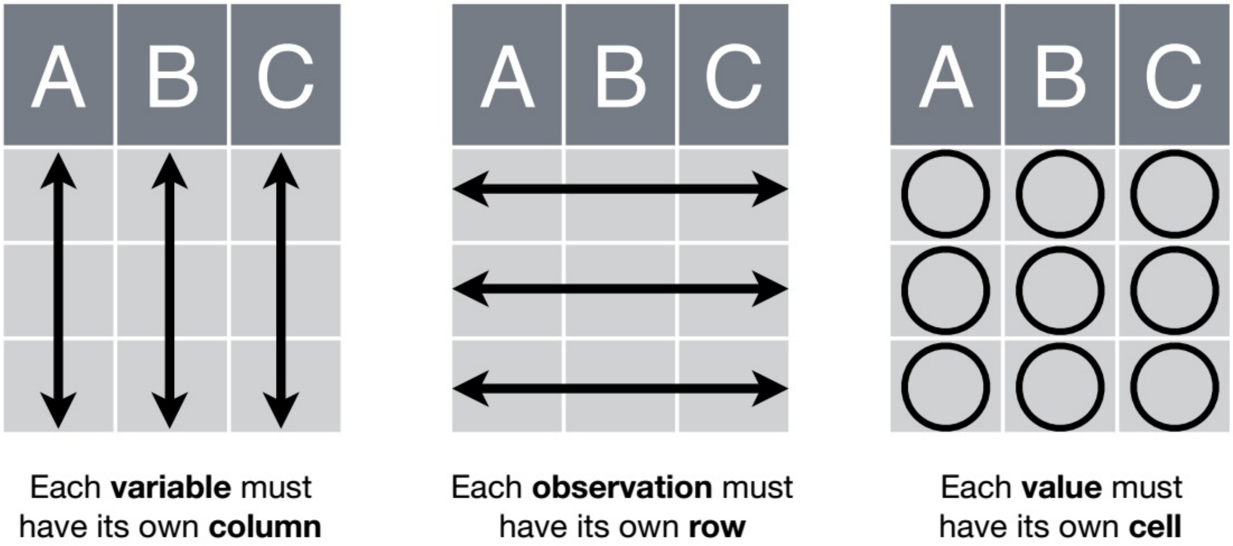

A B C
A B C
A B C
Each **variable** must have its own **column**
Each **observation** must have its own **row**
Each **value** must have its own **cell**

Figure 4: Tidy Data facilitate data modeling, graphing, aggregation with structure

Mathematically, the analyzed data can be expressed in matrix format $\mathbf{X}$.

$$\mathbf{X}_{n\times p} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

- n observations in the rows
- p variables in the columns

Then, we commonly care for two issues

- Study the resemblance between observations
- Study the relationships among variables

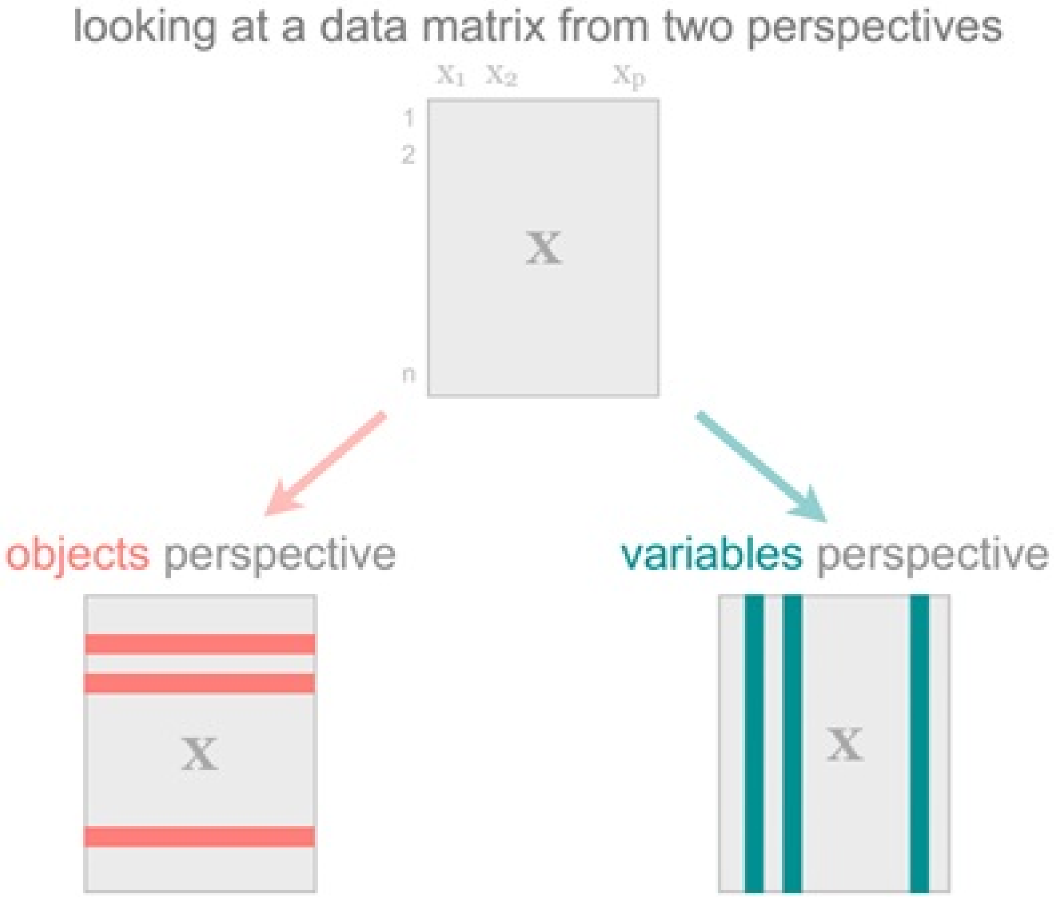


Figure 5: Data Perspectives

1.2 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a philosophy for the beginning of an analysis that describes a variety of techniques that are **quantitative** and **visual** in nature to look for patterns in data.

1.2.1 Visualization

Visualization is simply mapping data to geometric objects (points, lines, bars) and aesthetic attributes (color, shape, size).

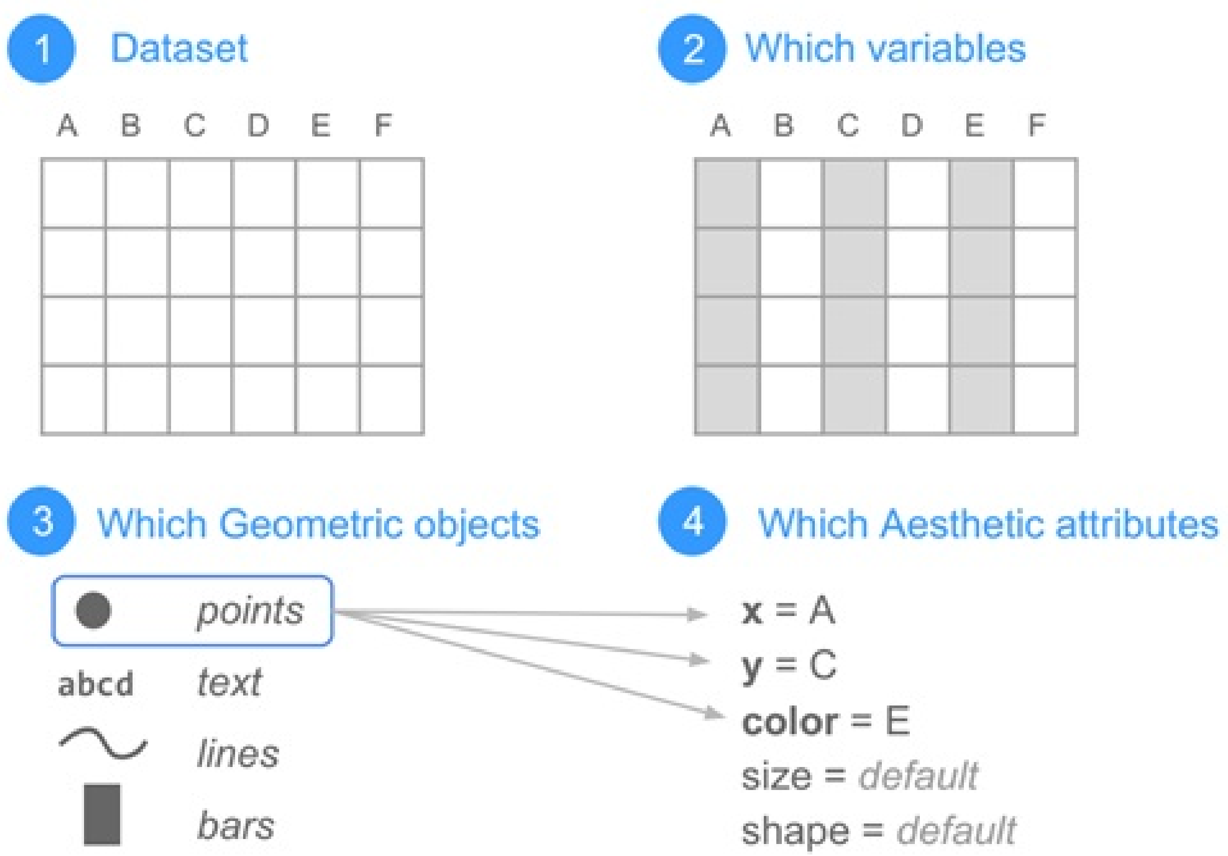


Figure 6: The Idea of Data Visualization

1.2.2 Graphing in R

There are three main **graphics systems** in R.

Figure 7: What Graph System to Choose?

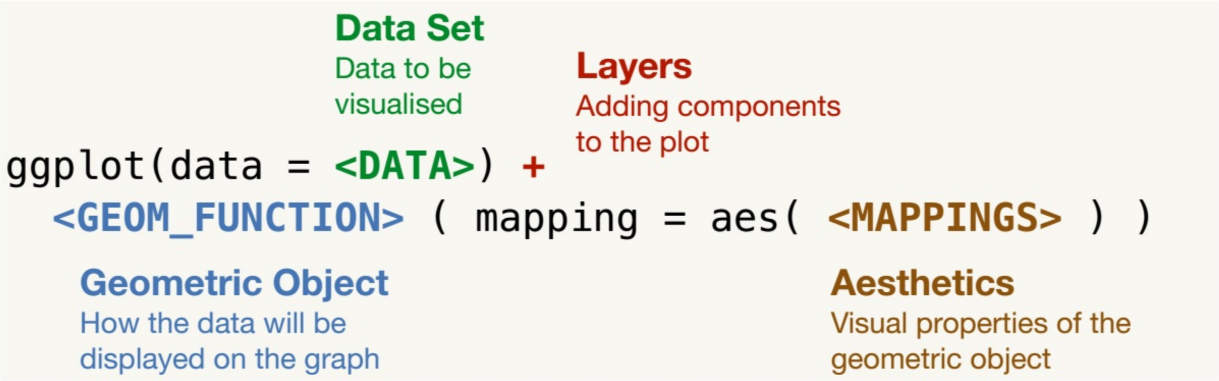
Grammar of Graphics: formal system of rules for generating graphics:

- some rules are mathematic
- some rules are aesthetic (i.e., visual)

`ggplot2` is an R package for producing statistical graphics based on the layered **Grammar of Graphics**.

1. **specification**: link data yo graphic objects
2. **Assembly**: put everything together
3. **Display**: render of a graphic

Here is a very basic `ggplot2` template:



```
ggplot(data = <DATA>) +
  <GEOM_FUNCTION> ( mapping = aes( <MAPPINGS> ) )
```

Data Set
Data to be visualised
Layers
Adding components to the plot
Geometric Object
How the data will be displayed on the graph
Aesthetics
Visual properties of the geometric object

Figure 8: Basis for Making ggplot2 Graphs

And here is the a graphing outline shows you when to use which.

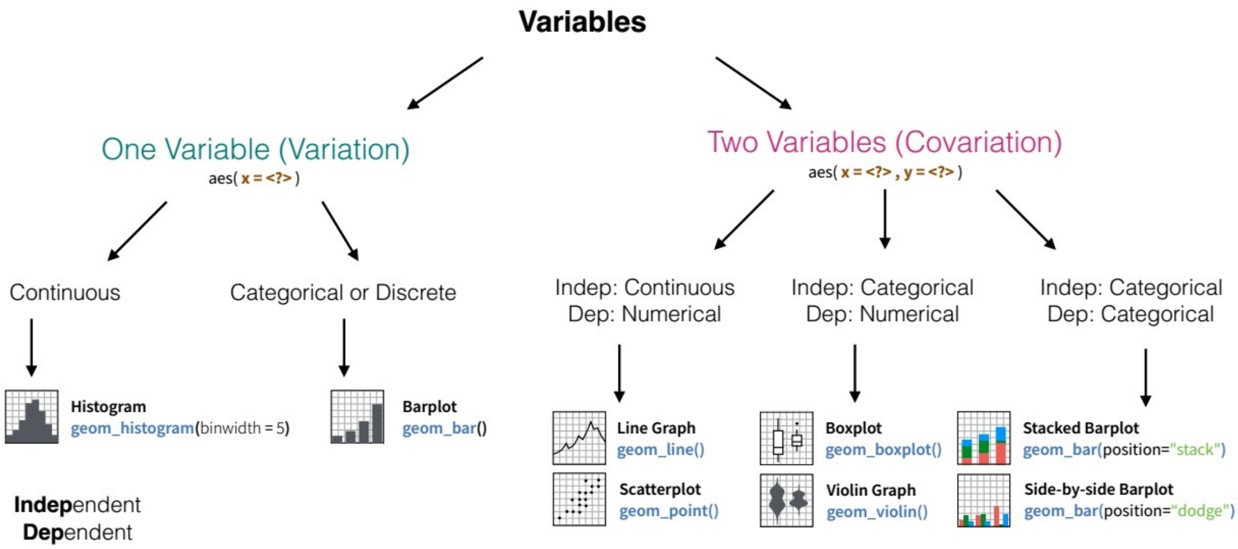


Figure 9: Graphing Outline

1.3 Examples

1.3.1 Nutrition Facts for McDonald's Menu

1.3.2 Scrape and explore `ratemyprofessors.com` data