Introduction to Data Visualization

STAT 133

Gaston Sanchez

Department of Statistics, UC–Berkeley

gastonsanchez.com
github.com/gastonstat/stat133
Course web: gastonsanchez.com/teaching/stat133

Graphics

Data Visualization

Using only numerical reduction methods in data analyses is far too limiting

Motivation

Consider some data (four pairs of variables)

```
##      x1     y1   x2    y2   x3    y3   x4    y4
## 1    10   8.04   10  9.14   10  7.46    8  6.58
## 2     8   6.95    8  8.14    8  6.77    8  5.76
## 3    13   7.58   13  8.74   13 12.74    8  7.71
## 4     9   8.81    9  8.77    9  7.11    8  8.84
## 5    11   8.33   11  9.26   11  7.81    8  8.47
## 6    14   9.96   14  8.10   14  8.84    8  7.04
## 7     6   7.24    6  6.13    6  6.08    8  5.25
## 8     4   4.26    4  3.10    4  5.39   19 12.50
## 9    12  10.84   12  9.13   12  8.15    8  5.56
## 10    7   4.82    7  7.26    7  6.42    8  7.91
## 11    5   5.68    5  4.74    5  5.73    8  6.89
```

What things would like
to calculate for each variable?

Motivation

```
##        x1              x2              x3              x4
##  Min.   : 4.0   Min.   : 4.0   Min.   : 4.0   Min.   : 8
##  1st Qu.: 6.5   1st Qu.: 6.5   1st Qu.: 6.5   1st Qu.: 8
##  Median : 9.0   Median : 9.0   Median : 9.0   Median : 8
##  Mean   : 9.0   Mean   : 9.0   Mean   : 9.0   Mean   : 9
##  3rd Qu.:11.5   3rd Qu.:11.5   3rd Qu.:11.5   3rd Qu.: 8
##  Max.   :14.0   Max.   :14.0   Max.   :14.0   Max.   :19
```

```
##        y1              y2              y3              y4
##  Min.   : 4.260   Min.   :3.100   Min.   : 5.39   Min.   : 5.250
##  1st Qu.: 6.315   1st Qu.:6.695   1st Qu.: 6.25   1st Qu.: 6.170
##  Median : 7.580   Median :8.140   Median : 7.11   Median : 7.040
##  Mean   : 7.501   Mean   :7.501   Mean   : 7.50   Mean   : 7.501
##  3rd Qu.: 8.570   3rd Qu.:8.950   3rd Qu.: 7.98   3rd Qu.: 8.190
##  Max.   :10.840   Max.   :9.260   Max.   :12.74   Max.   :12.500
```

What things would like to calculate
for each pair of variables (e.g. `x1`, `y1`)?

Motivation

```
cor(anscombe$x1, anscombe$y1)

## [1] 0.8164205

cor(anscombe$x2, anscombe$y2)

## [1] 0.8162365

cor(anscombe$x3, anscombe$y3)

## [1] 0.8162867

cor(anscombe$x4, anscombe$y4)

## [1] 0.8165214
```

Motivation

- Mean of x values $= 9.0$
- Mean of y values $= 7.5$
- least squares equation: $y = 3 + 0.5x$
- Sum of squared errors: 110
- Correlation coefficient: 0.816

Why Graphics?

Are you able to see any patterns, associations, relations?

```
##     x1     y1  x2    y2  x3     y3  x4     y4
## 1   10   8.04  10  9.14  10   7.46   8   6.58
## 2    8   6.95   8  8.14   8   6.77   8   5.76
## 3   13   7.58  13  8.74  13  12.74   8   7.71
## 4    9   8.81   9  8.77   9   7.11   8   8.84
## 5   11   8.33  11  9.26  11   7.81   8   8.47
## 6   14   9.96  14  8.10  14   8.84   8   7.04
## 7    6   7.24   6  6.13   6   6.08   8   5.25
## 8    4   4.26   4  3.10   4   5.39  19  12.50
## 9   12  10.84  12  9.13  12   8.15   8   5.56
## 10   7   4.82   7  7.26   7   6.42   8   7.91
## 11   5   5.68   5  4.74   5   5.73   8   6.89
```

Famous dataset `"anscombe"` (four data sets)

Why Graphics?

How are these two variables associated? What does these data values look like?

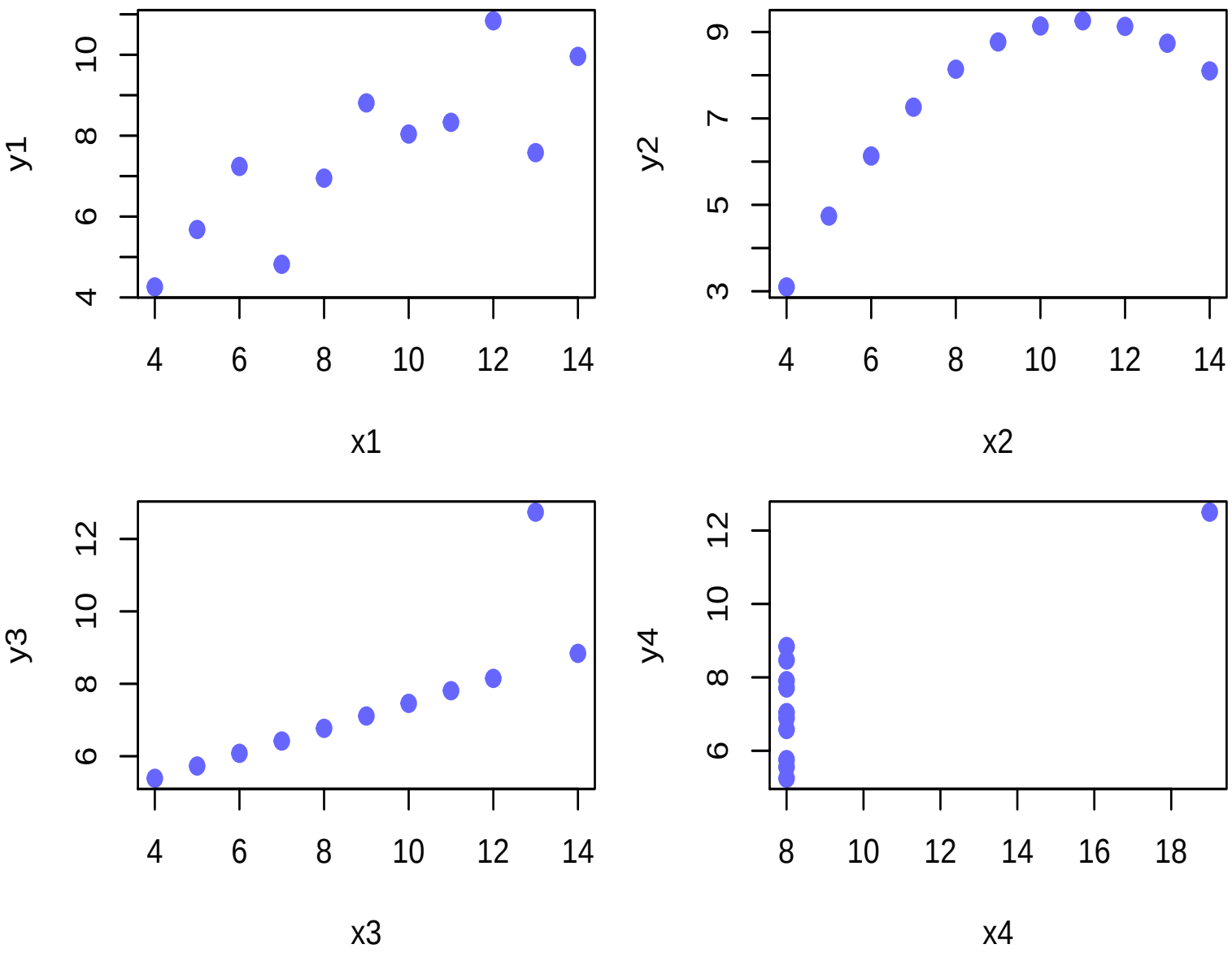
```
##     x1    y1
## 1   10  8.04
## 2    8  6.95
## 3   13  7.58
## 4    9  8.81
## 5   11  8.33
## 6   14  9.96
## 7    6  7.24
## 8    4  4.26
## 9   12 10.84
## 10   7  4.82
## 11   5  5.68
```

Our eyes are not very good at making sense when looking at (many) numbers

Our eyes are not very good at making sense when looking at (many) numbers

But they are great for looking at shapes and detecting patterns

Why Graphics



Data Visualization

Using only numerical reduction methods in data analyses is far too limiting.

Visualization provides insight that cannot be appreciated by any other approach to learning from data. (W. S. Cleveland)

Data Visualization

A key component of computing with data consists of **Data Visualization**

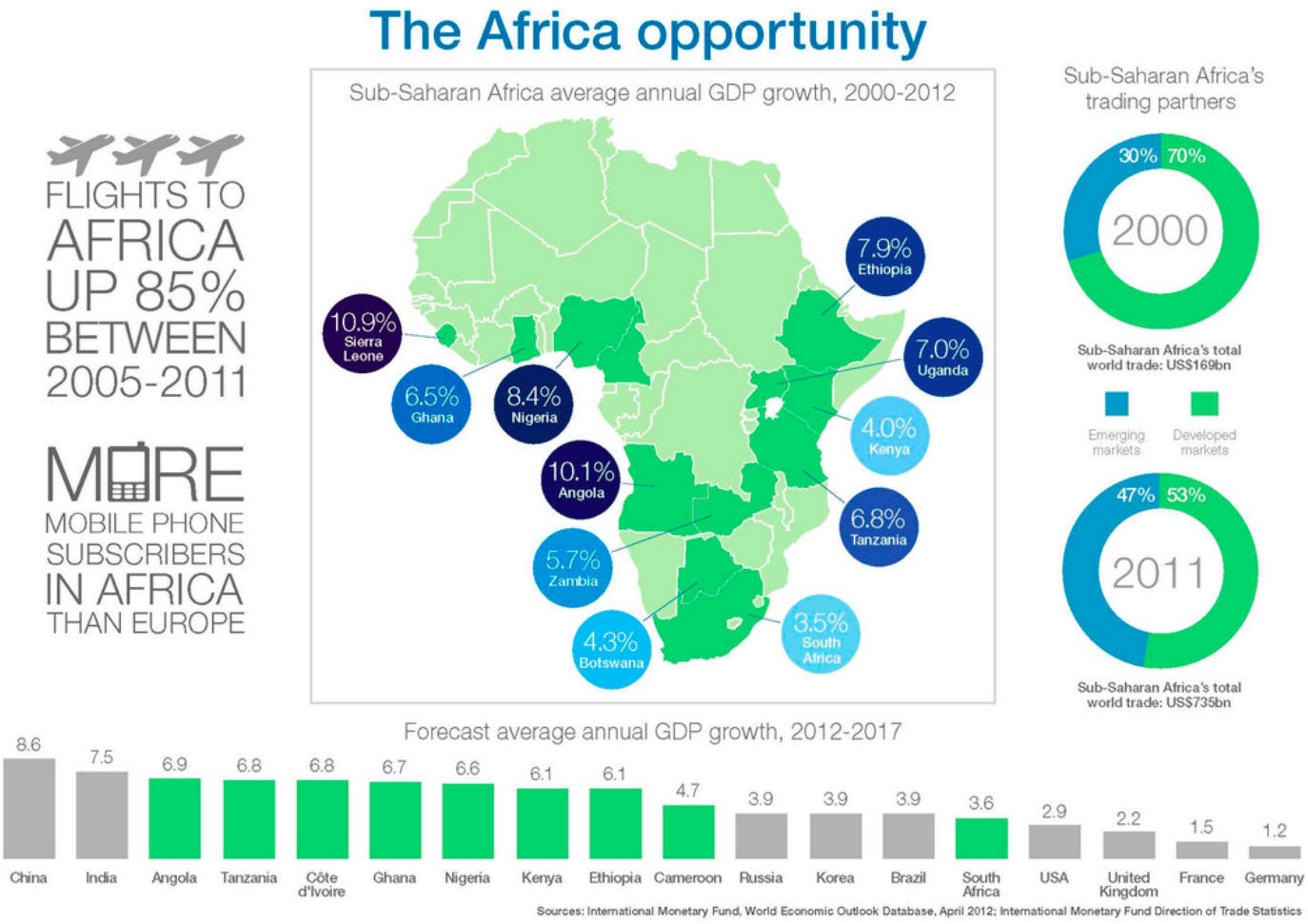
Google `"data visualization"`

Data Visualization

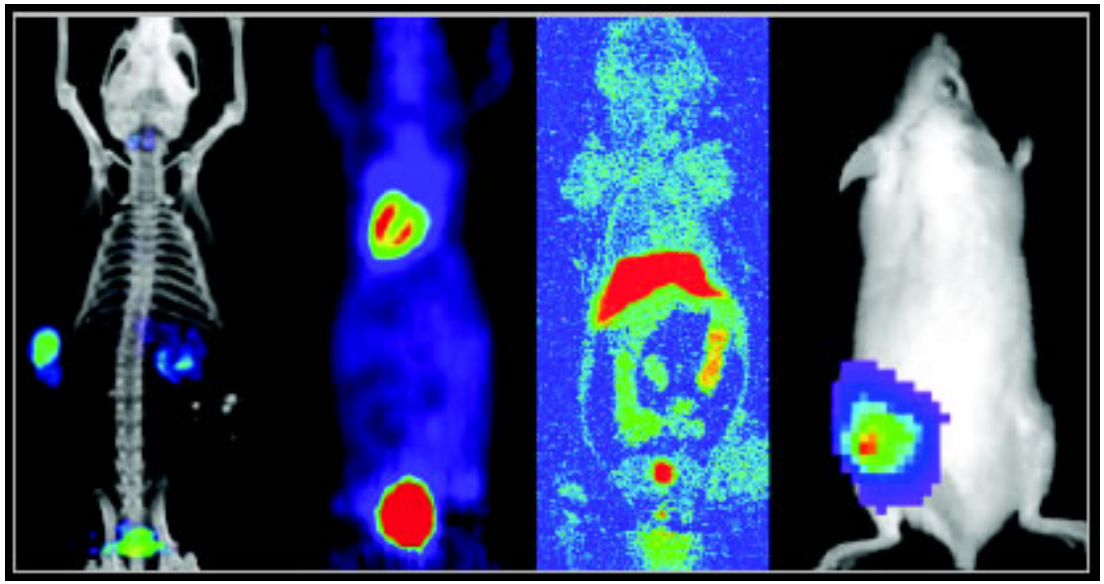
Data Visualization

- Computer Graphics?
- Infographics?
- Data Art?
- Computer Vision?

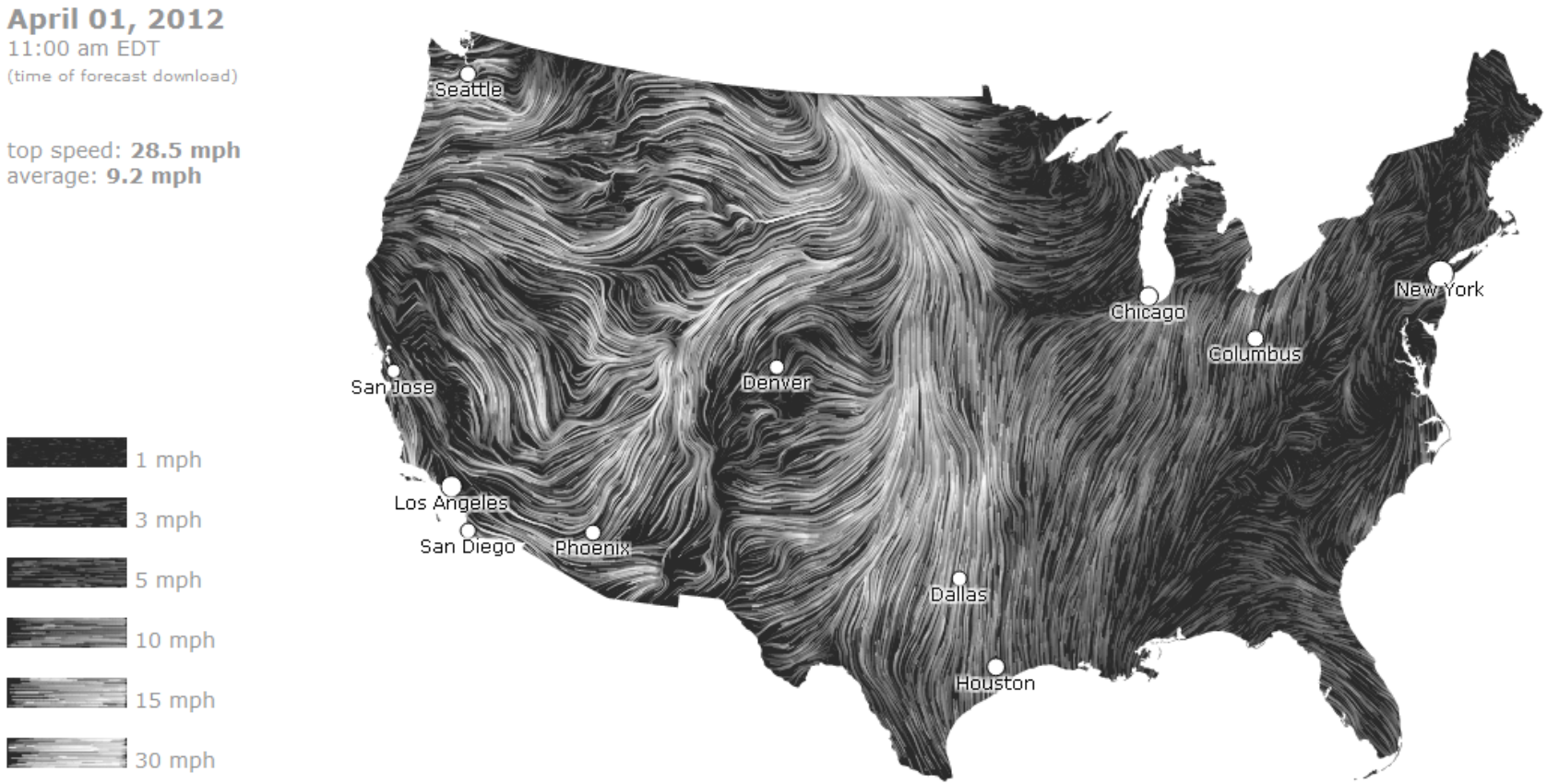
Infographic



Scientific Imaging



Data Art


April 01, 2012
11:00 am EDT
(time of forecast download)
top speed: 28.5 mph
average: 9.2 mph
1 mph
3 mph
5 mph
10 mph
15 mph
30 mph
Seattle
San Jose
Los Angeles
San Diego
Phoenix
Denver
Dallas
Houston
Chicago
Columbus
New York

Data Visualization

Things commonly said about statistical graphics

- The data should stand out
- Story telling
- Big Picture
- "The purpose of visualization is insight, not pictures" (Ben Shneiderman)

Data Visualization

We'll focus on statistical graphics and other visual displays of data in science and technology

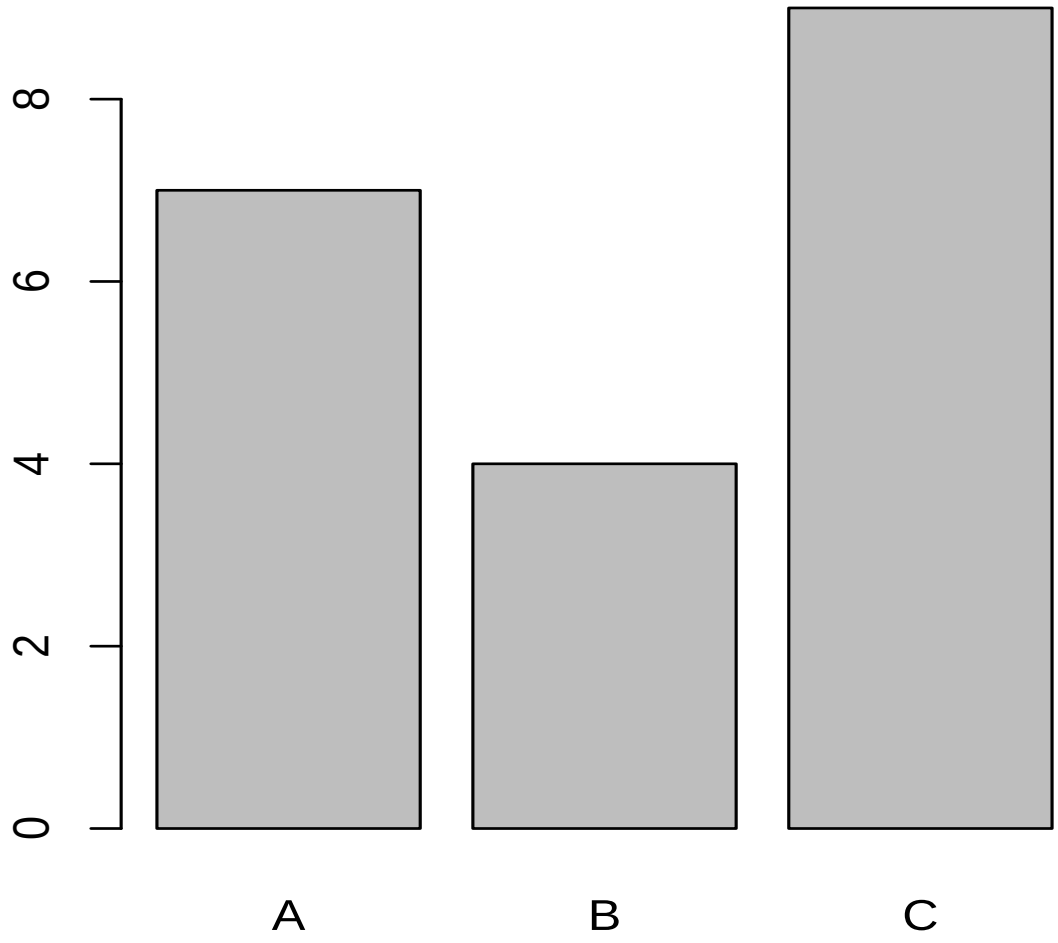
Graphics for

Exploration & Communication

Graphics for Exploration

- graphics for understanding data
- the analyst is the main (and usually only) consumer
- typically quick & dirty (not much care about visual appearance and design principles)
- lifespan of a few seconds

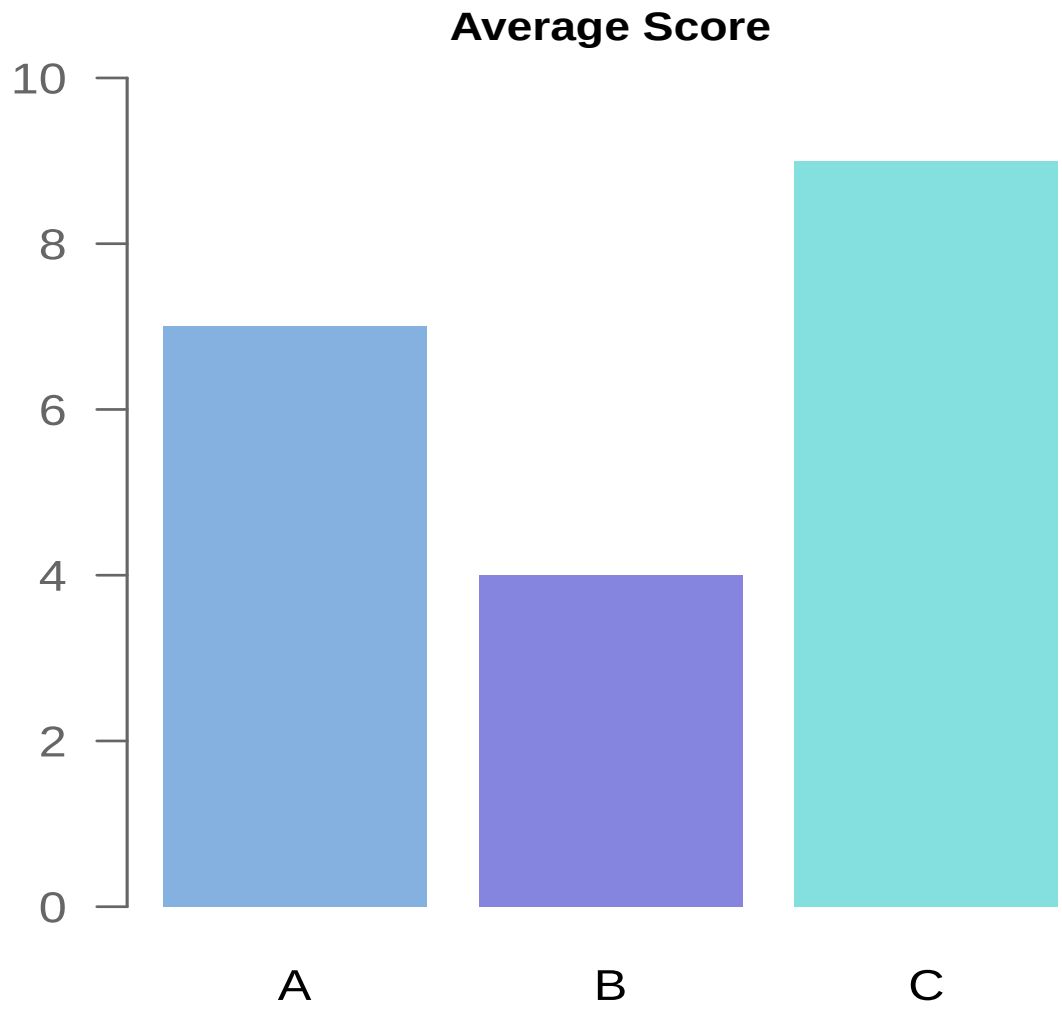
Graphics for Exploration



Graphics for Communication

- graphics for presenting data
- to be consumed by others
- must care about visual appearance and design
- require a lot of iterations in order to get the final version
- what's the message?
- who's the audience?
- on what type of media / format?

Graphics for Communication



Graphics for Communication

Use visualization to communicate ideas, influence, explain persuade

Visuals can serve as evidence or support

Visualization

- Visuals can frequently take the place of many words, tables, and numbers
- Visuals can summarize, aggregate, unite, explain
- Sometimes words are needed, however

Graphics (Part I)

In this first part of the course we'll focus on:

- graphics for exploration
- types of statistical graphics
- understanding graphics system in R
- traditional R graphics and graphics with `"ggplot2"`

Graphics (Part II)

Later in the course we'll talk about:

- graphics for communication
- design principles
- color theory and use of color
- guidelines and good practices
- `"shiny"` and interactive graphics

Considerations

Number of Variables

Type of Variables

How many variables?

Variables in datasets:

- 1 - univariate data
- 2 - bivariate data
- 3 - trivariate data
- multivariate data

What type of variables?

- Quantitative -vs- Qualitative
- Continuous -vs- Discrete

Univariate

Quantitative variable:

- Distribution
- How values are distributed
- max, min
- central values
- areas of concentration
- outliers
- patterns

Univariate

Qualitative variable:

- Counts and proportions
- Common values
- Most typical value
- Distribution of proportions

Bivariate

- Quantitative-Quantitative
- Qualitative-Quantitative
- Qualitative-Qualitative

In general we care about association (correlation, relationships)

Multivariate

- Quantitative
- Qualitative
- Mixed

In general we care about association (correlation, relationships)

What about individuals?

- Resemblance
- Similarities and disimilarities
- Typologies