Data representation

Lecture 2

IFT6758, Fall 2020; Reading: IDS - Chapters 8, 9, 10

Extracting information from data

Extracting information from data

A picture is worth a thousand words: can we see the data in pictorial format?

Extracting information from data

A picture is worth a thousand words: can we see the data in pictorial format?

Effect of informative data representation: e.g., News organizations increasingly embracing *data journalism* and including effective *infographics* as part of their reporting.

Extracting information from data

A picture is worth a thousand words: can we see the data in pictorial format?

Effect of informative data representation: e.g., News organizations increasingly embracing *data journalism* and including effective *infographics* as part of their reporting.

Data visualization is the strongest tool of what we call *exploratory data analysis* (EDA). **John W. Tukey**, considered the father of EDA, once said,

Extracting information from data

A picture is worth a thousand words: can we see the data in pictorial format?

Effect of informative data representation: e.g., News organizations increasingly embracing *data journalism* and including effective *infographics* as part of their reporting.

Data visualization is the strongest tool of what we call *exploratory data analysis* (EDA). **John W. Tukey**, considered the father of EDA, once said,

"The greatest value of a picture is when it forces us to notice what we never expected to see."

PART-1: Visualizing data distributions

Types of variables

Types of variables

- Types of variables: **Categorical** (**ordinal or not**) and **numerical** (**discrete or continuous**)

Types of variables

- Types of variables: **Categorical** (**ordinal or not**) and **numerical** (**discrete or continuous**)

- Example:
  - **Categorical**: Sex (Male, Female), Regions (North, South, East, West). Ordinal: when there is a sense of order, Spiciness (Mild, Medium, Hot)..
  - **Numerical**: Height (continuous), Price (continuous), Population sizes (discrete)..

Types of variables

- Types of variables: **Categorical** (**ordinal or not**) and **numerical** (**discrete or continuous**)

- Example:
  - **Categorical**: Sex (Male, Female), Regions (North, South, East, West). Ordinal: when there is a sense of order, Spiciness (Mild, Medium, Hot)..
  - **Numerical**: Height (continuous), Price (continuous), Population sizes (discrete)..
  - Discrete numeric data can be considered ordinal.
    - **Conventionally**, ordinal for variables belonging to a small number of different groups, with each group having many members: e.g.: the number of packs of cigarettes a person smokes a day, rounded to the closest pack
    - Discrete numerical for many groups with few cases in each group: the actual number of cigarettes in each pack

Distribution function

Distribution function

- With **categorical data**, the distribution describes **the proportion of each unique category**

Distribution function

- With **categorical data**, the distribution describes **the proportion of each unique category**
- **Two-category frequency table** is sufficient for comprehension:

  Example: Female 0.227, Male 0.773

Distribution function

- With **categorical data**, the distribution describes **the proportion of each unique category**
- **Two-category frequency table** is sufficient for comprehension:

  Example: Female 0.227, Male 0.773

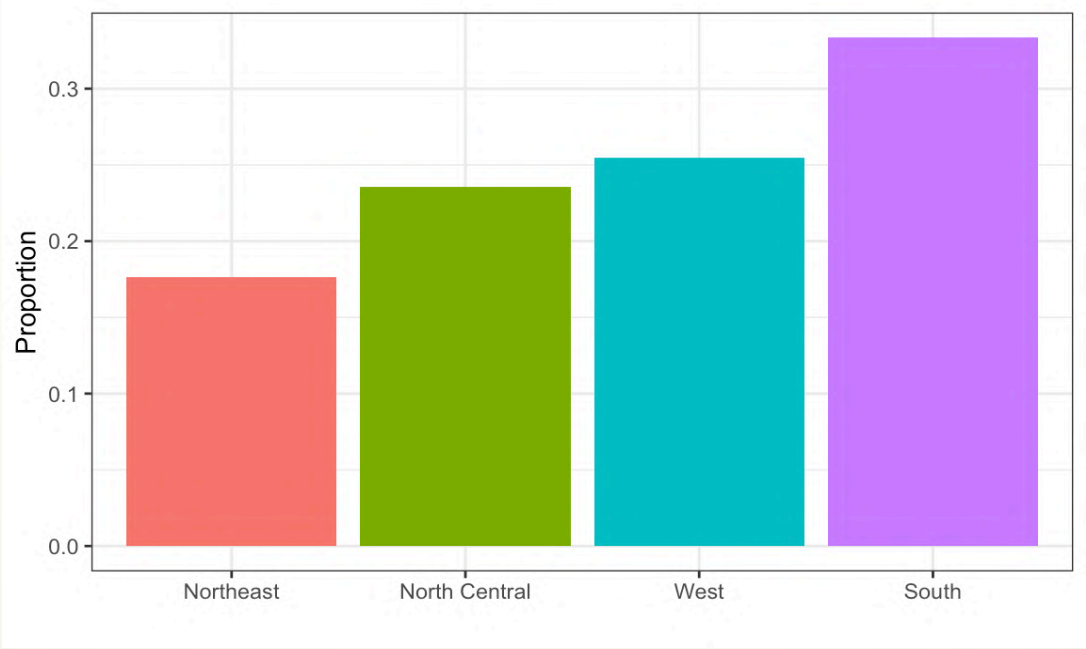
- **Visualize** for more than two categories: example – **bar plot for US population**

Distribution function

- With **categorical data**, the distribution describes **the proportion of each unique category**
- **Two-category frequency table** is sufficient for comprehension:

  Example: Female 0.227, Male 0.773

- **Visualize** for more than two categories: example – **bar plot for US population**



Cumulative distribution functions

Cumulative distribution functions

- Numerical variable: **cumulative distribution** is an effective summary

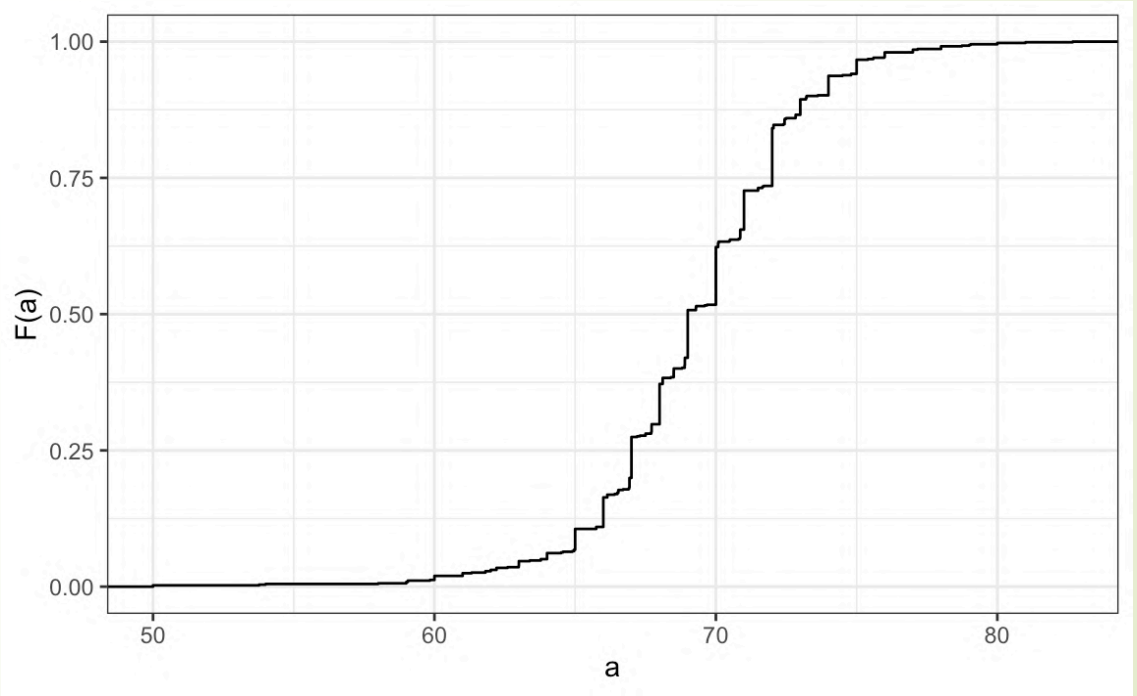
Cumulative distribution functions

- Numerical variable: **cumulative distribution** is an effective summary
- In statistics, the following notation is used: $\boldsymbol{F(a) = \Pr(x \leq a)}$

Cumulative distribution functions

- Numerical variable: **cumulative distribution** is an effective summary
- In statistics, the following notation is used: $\boldsymbol{F(a) = \Pr(x \leq a)}$
- $\boldsymbol{F(66) = 0.164, F(72) = 0.841}$

Male height data



Cumulative distribution functions

- Numerical variable: **cumulative distribution** is an effective summary
- In statistics, the following notation is used: $\boldsymbol{F(a) = \Pr(x \leq a)}$
- $\boldsymbol{F(66) = 0.164, F(72) = 0.841}$
- **Does not answer:**
  - At what value is the distribution centered?
  - Is the distribution symmetric?
  - What ranges contain 95% of the values?

Histograms

Histograms

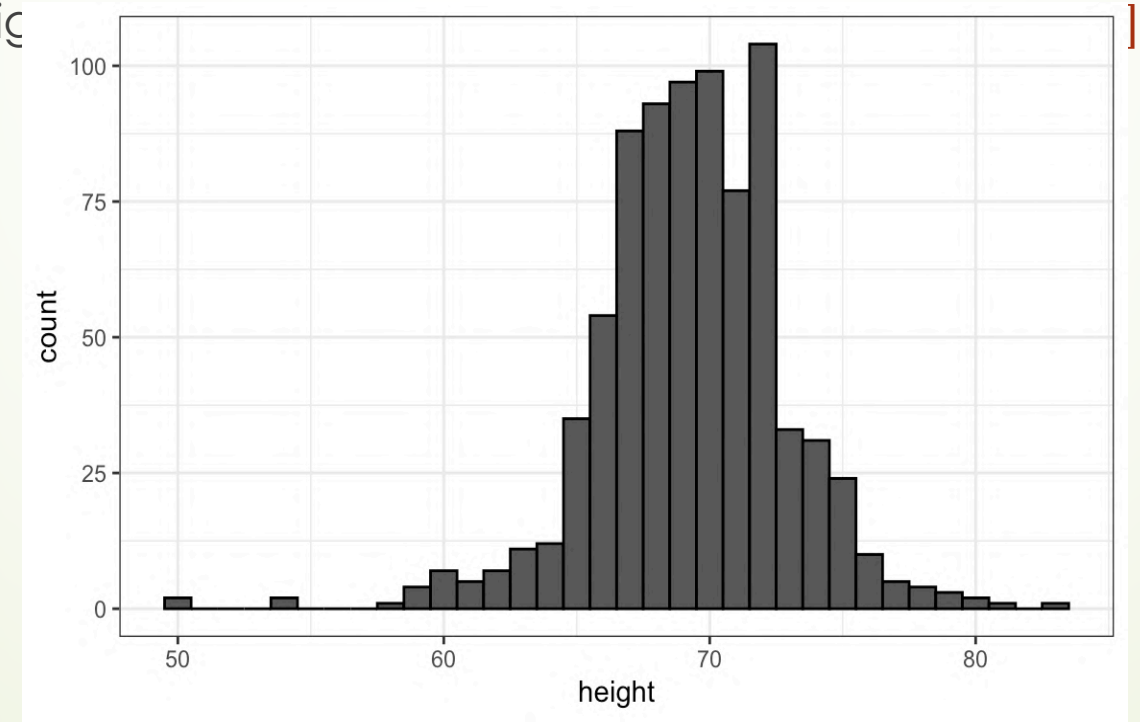
- Divide the span of our data into non-overlapping **bins** of the same size

Histograms

- Divide the span of our data into non-overlapping **bins** of the same size
- For each bin, we count the number of values that fall in that interval

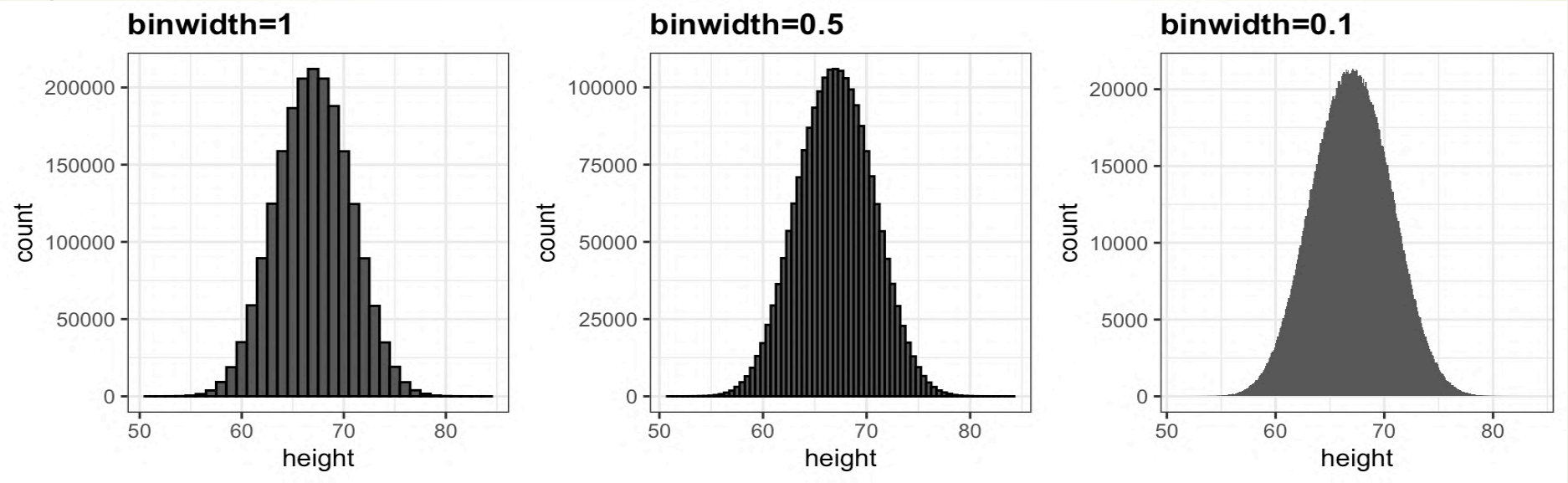
Histograms

- Divide the span of o*u9.5, 50.5 , 50.5, 51.5 , ..., (82.5, 83.5]*
- o non-overlapping **bins** of the same size
- For each bin, we count the number of values that fall in that interval
- Heig]

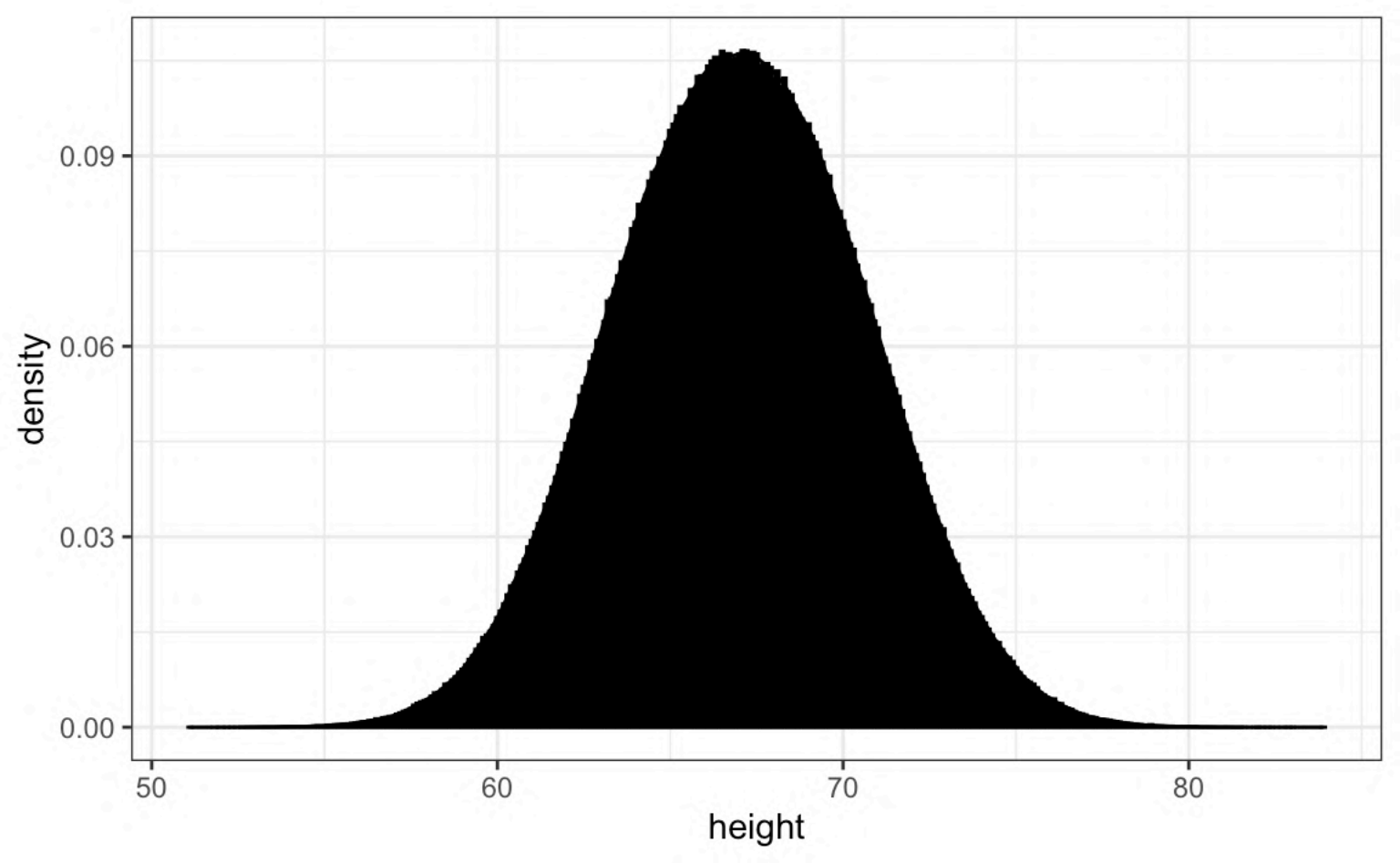


From histogram to smooth density

From histogram to smooth density

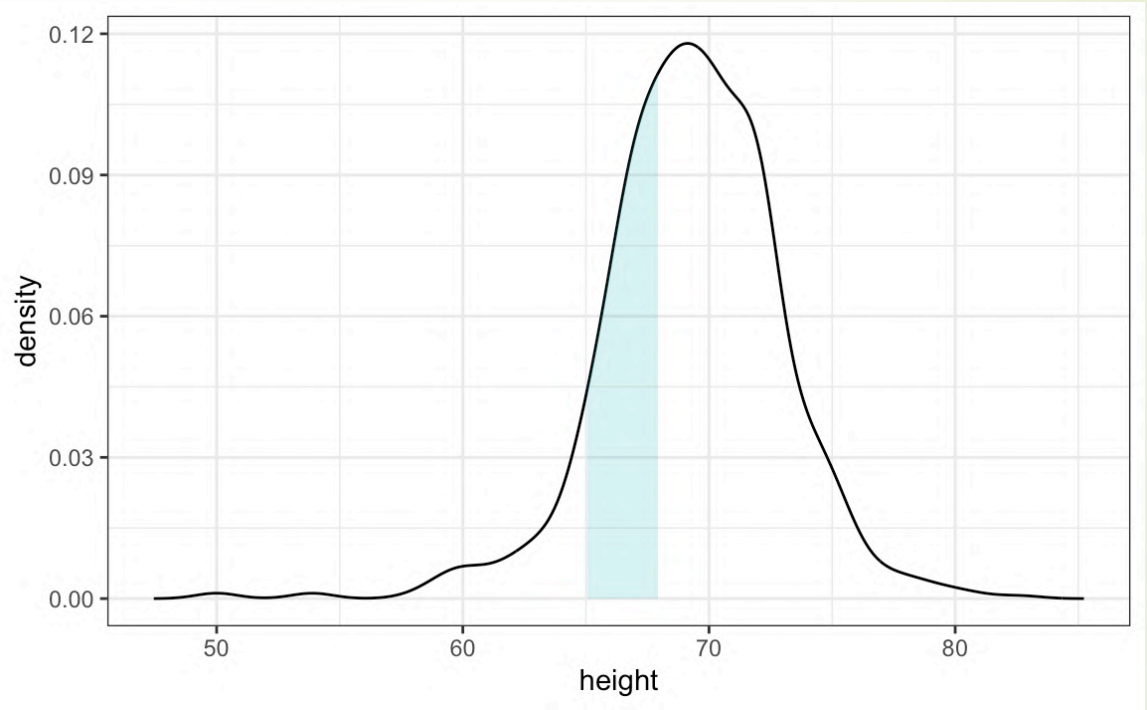


From histogram to smooth density

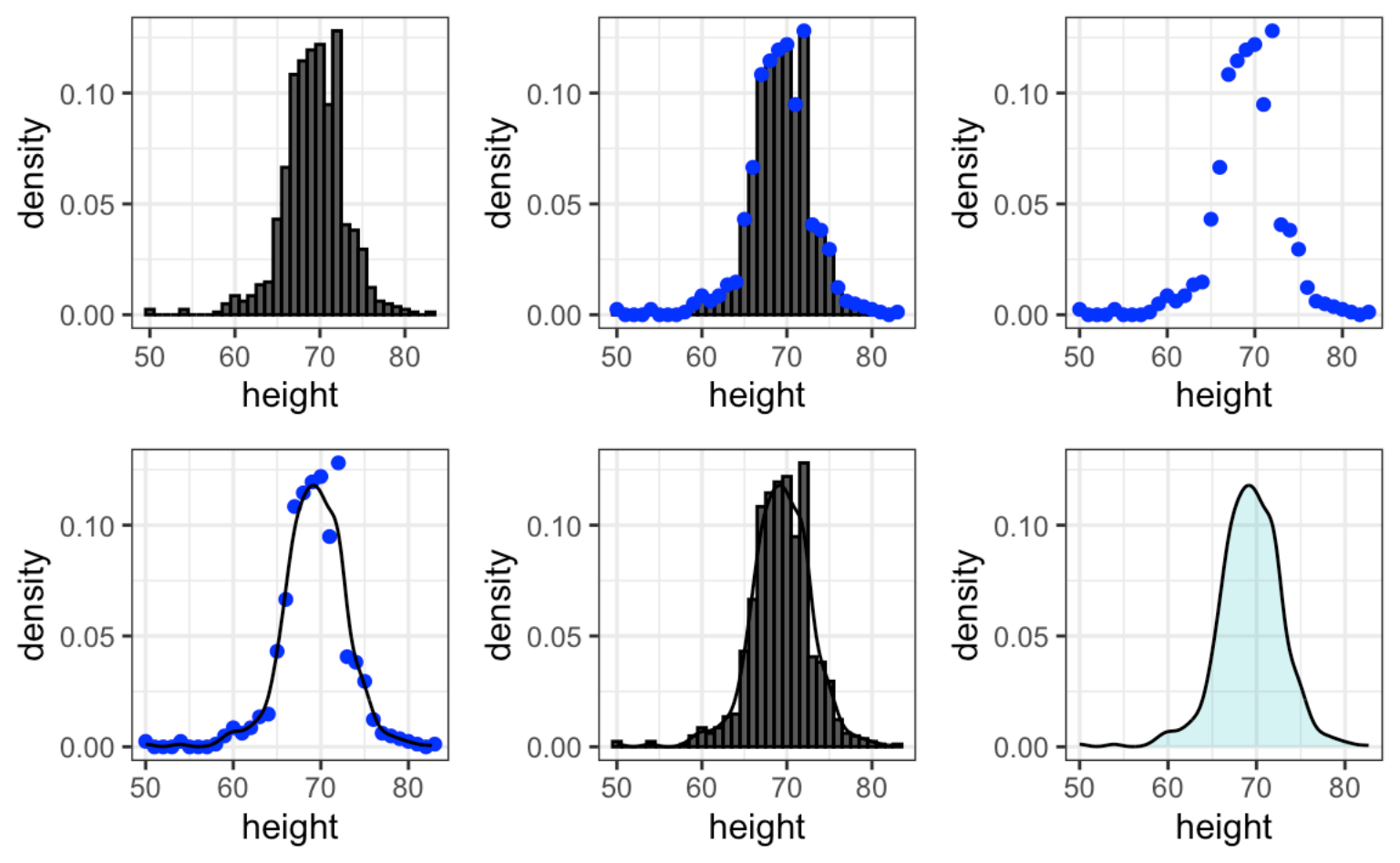


What is density?

- Proportion of values between 65 and 68
- **The proportion of this area** is about **0.3**, meaning that about **30%** of male heights are **between 65 and 68 inches.**

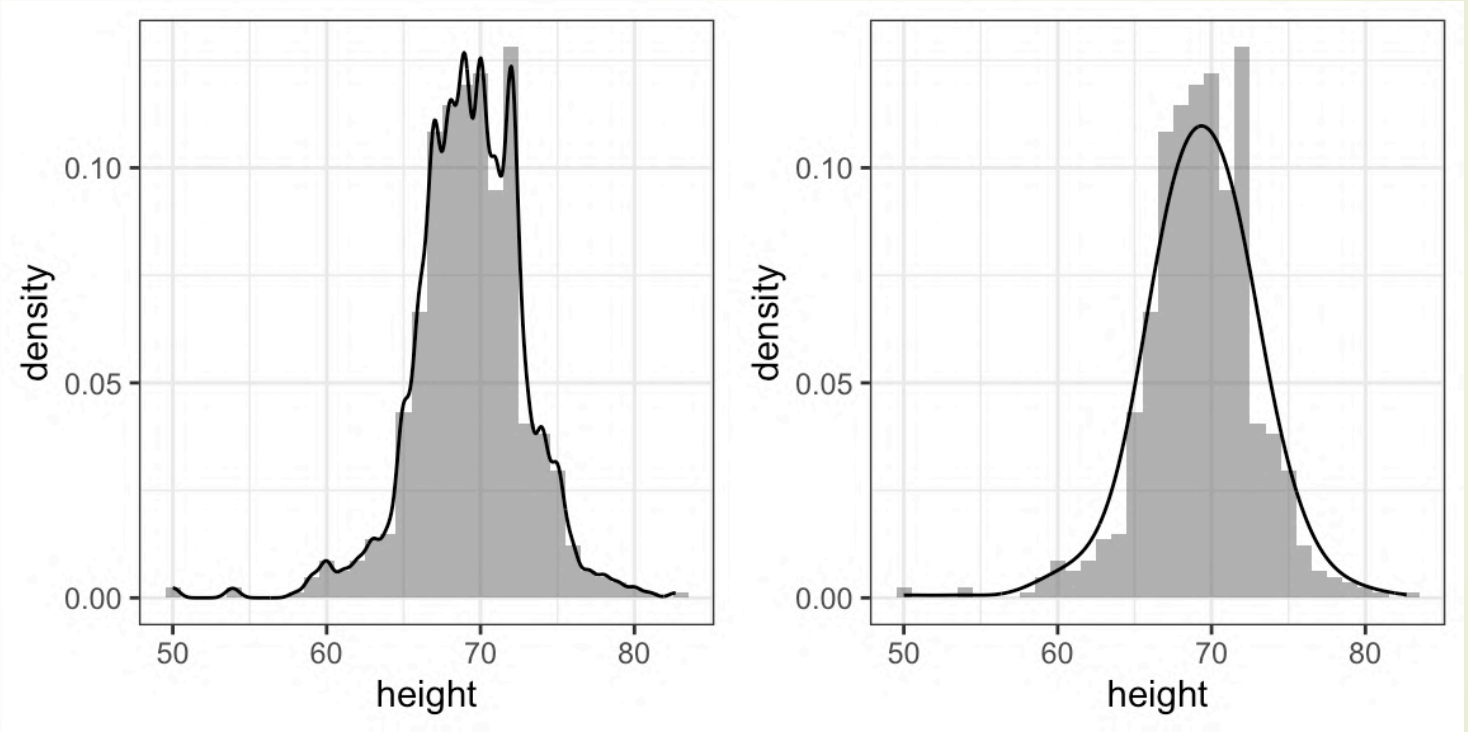


The transition to smoother density



Smoothness is relative

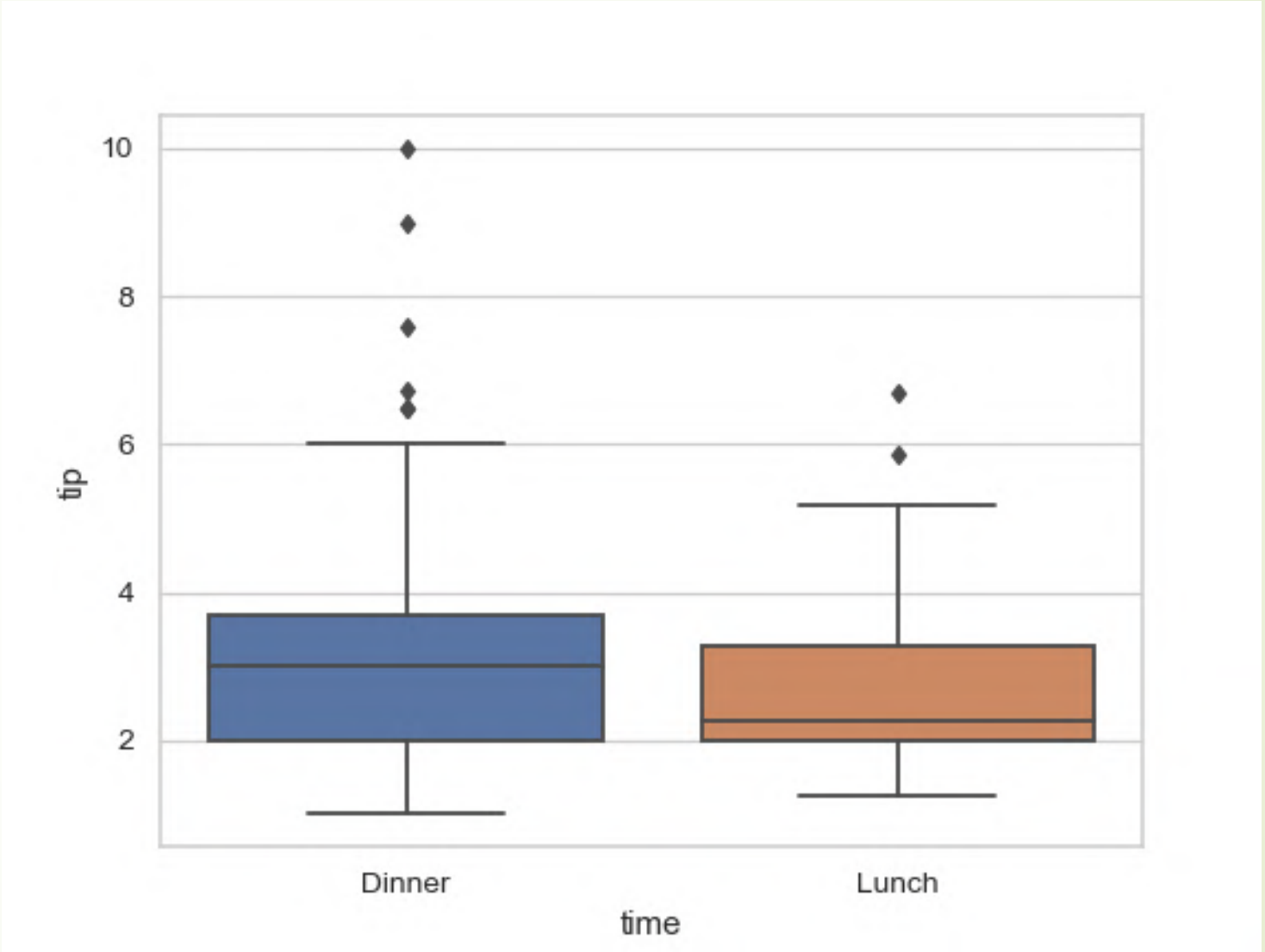
- Kernel Density Estimator
- Scipy, Scikit-Learn
- Smoothness varies with **Bandwidth**



Boxplots

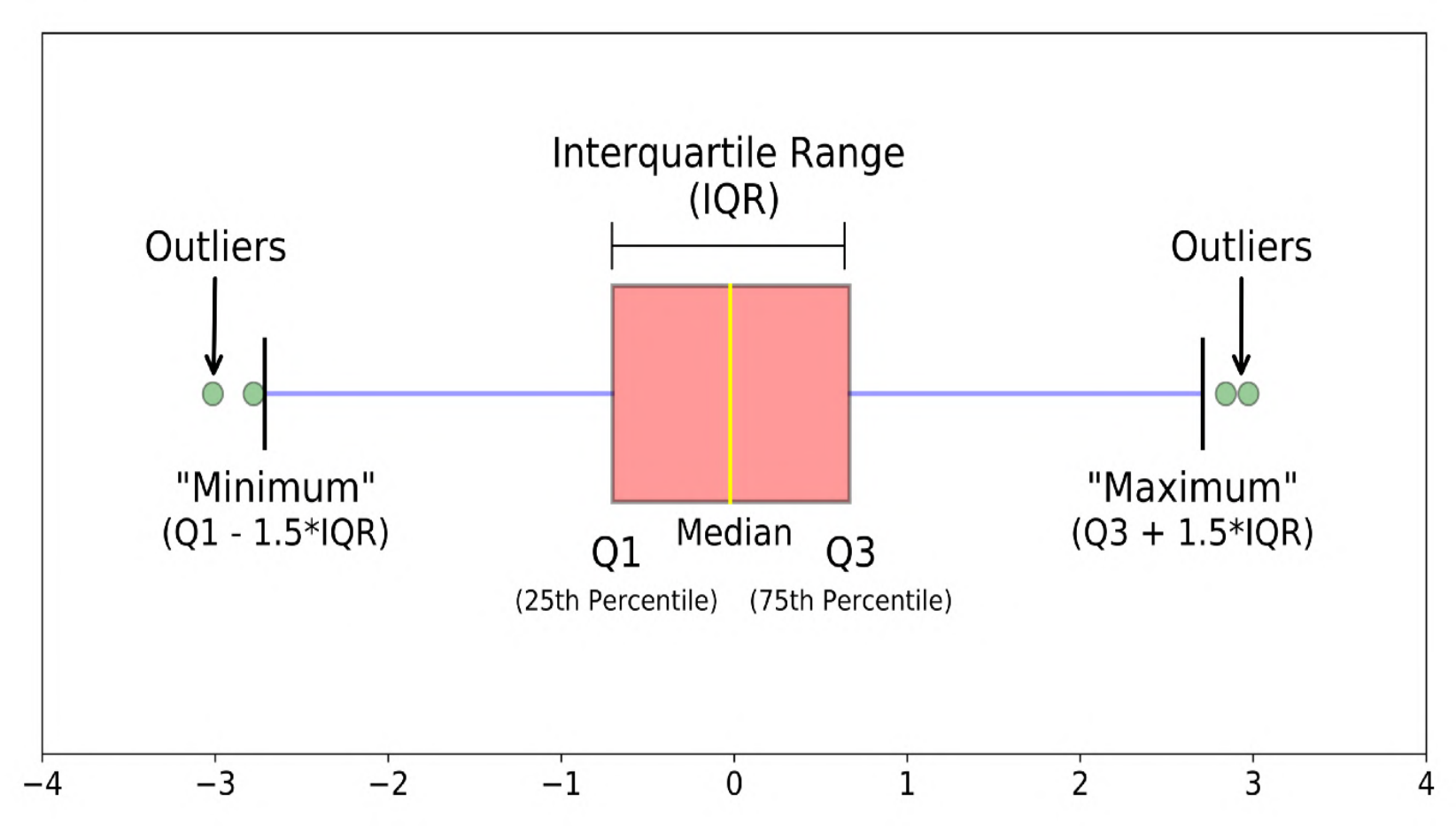
Boxplots

- Matplotlib, Seaborn



Boxplots

- Matplotlib, Seaborn
- Helps detecting outliers



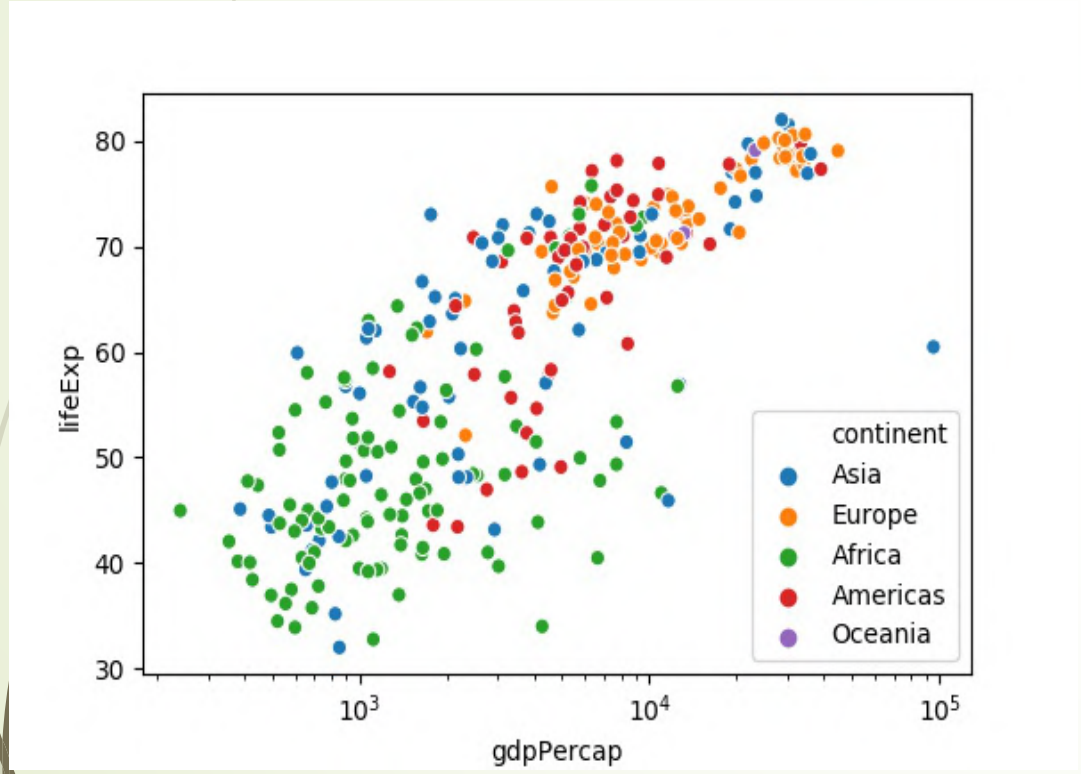
Many more to go…

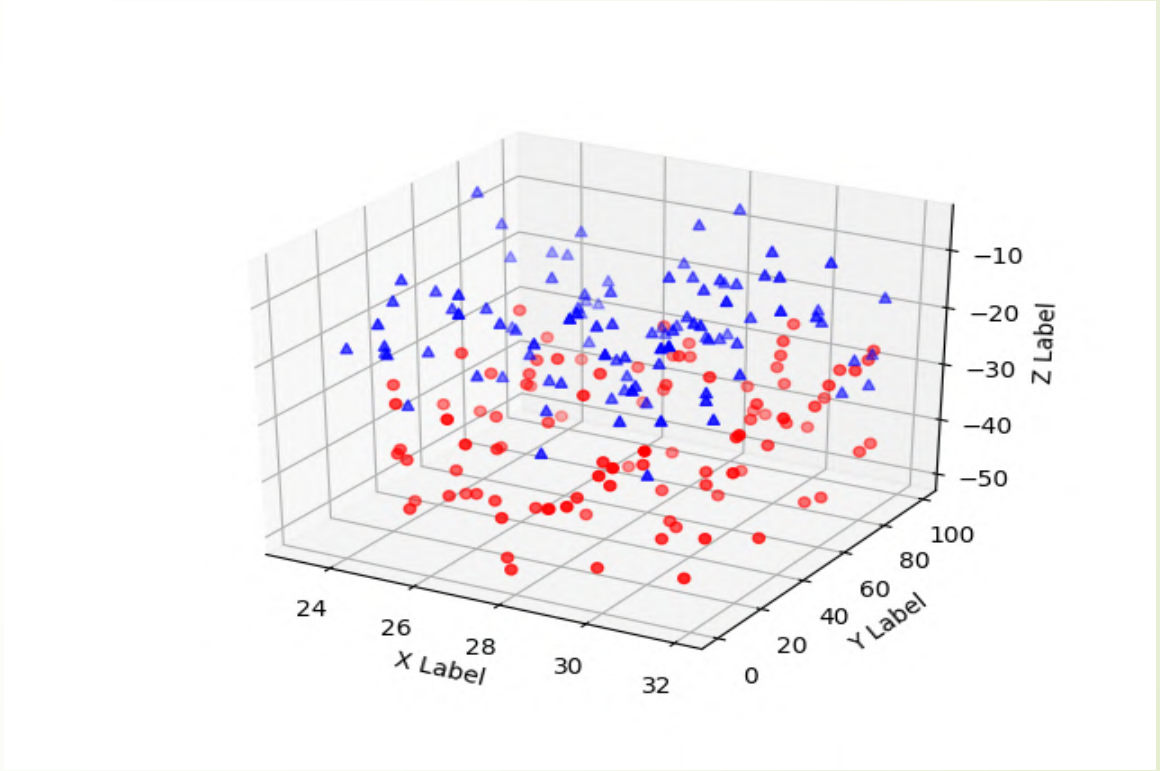
Many more to go…

- **E.g. Scatterplot**

Many more to go…

- **E.g. Scatterplot**
  - **Matplotlib, Seaborn**





Many more to go…

- **E.g. Scatterplot**
  - **Matplotlib, Seaborn**

PART-2: Data visualization in practice

An example: Gapminder data

An example: Gapminder data

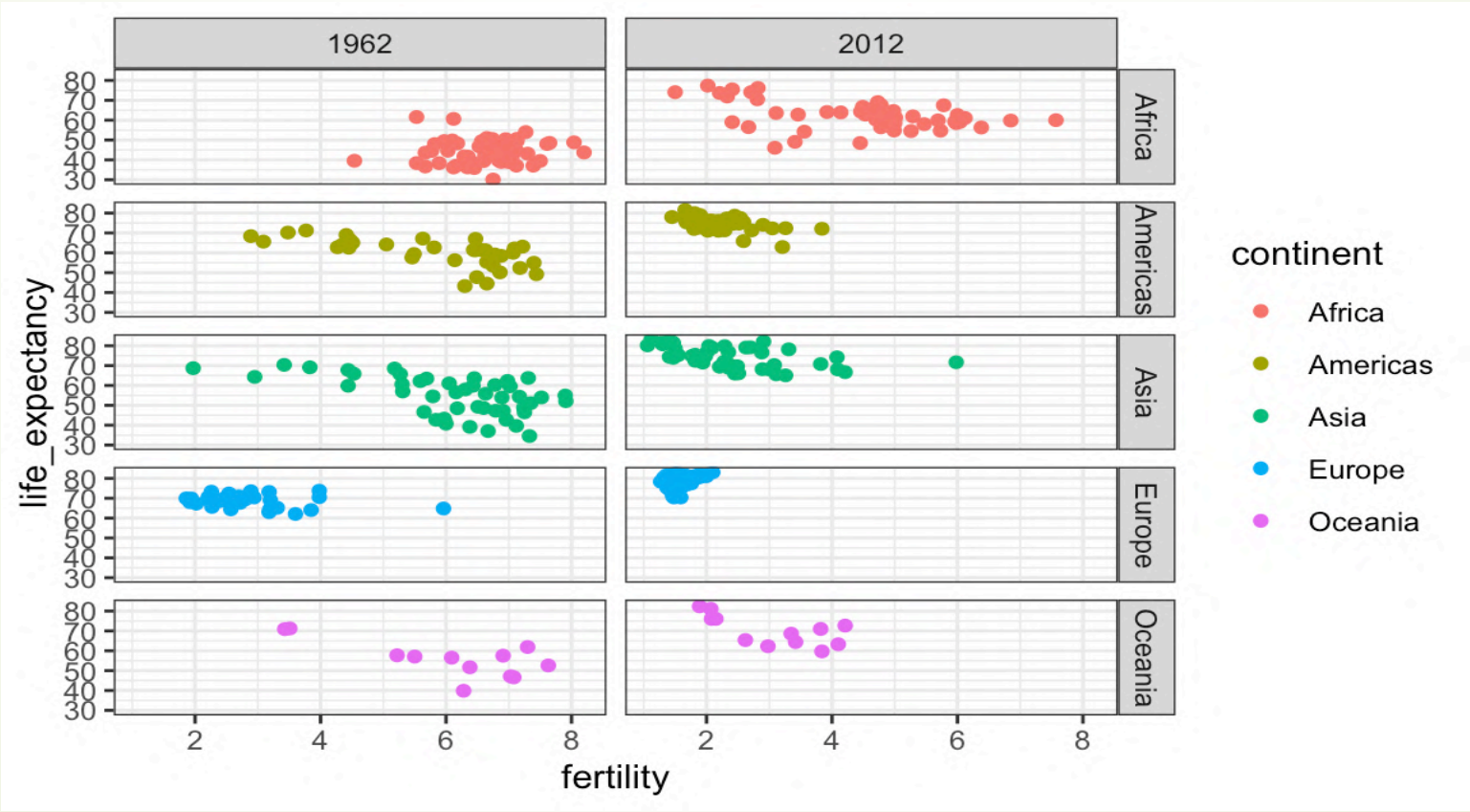
- Independent Swedish foundation with huge **global datasets** based on reliable statistics; develops data visualization tools; Gapminder homepage

An example: Gapminder data

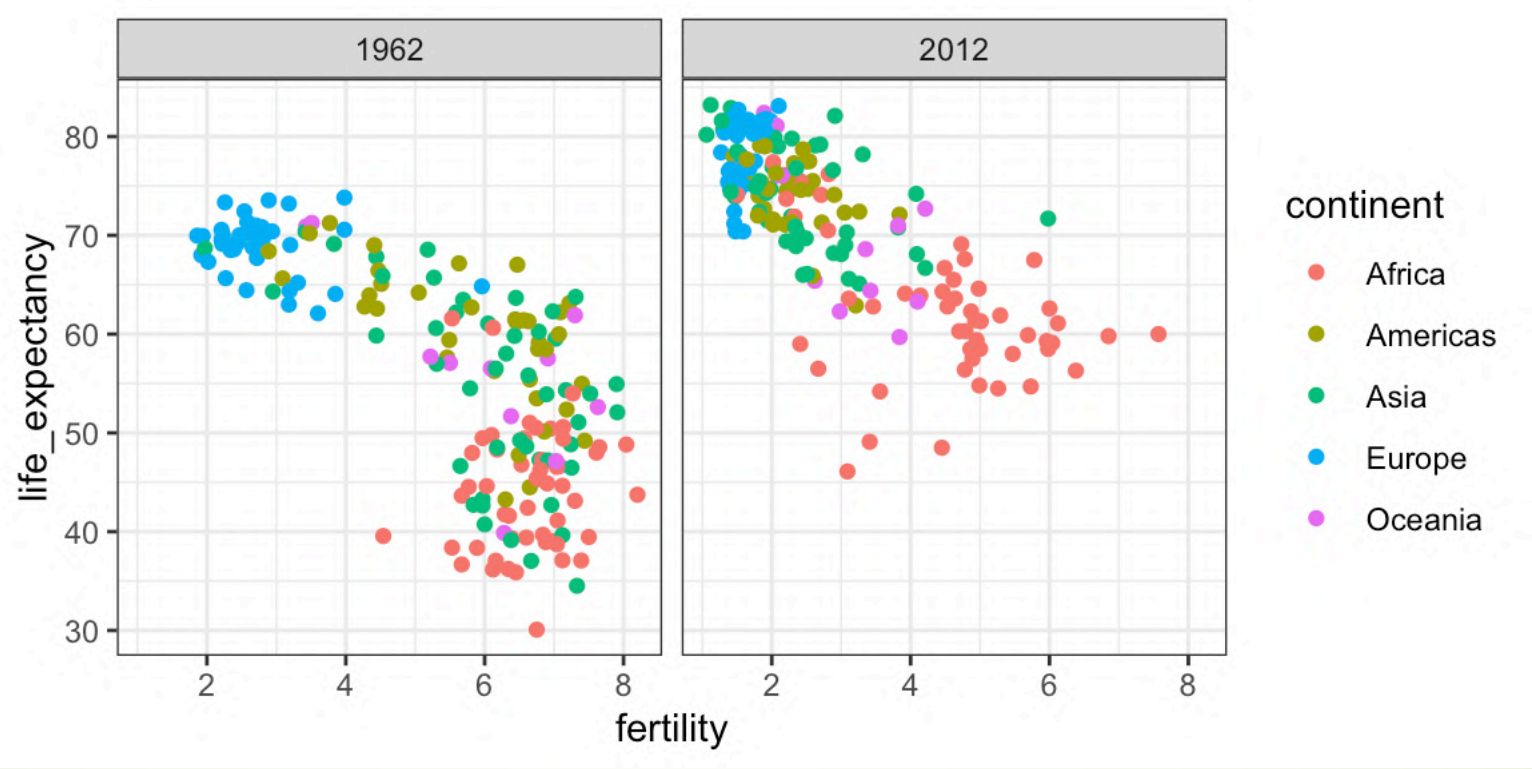
An example: Gapminder data

An example: Gapminder data

An example: Gapminder data



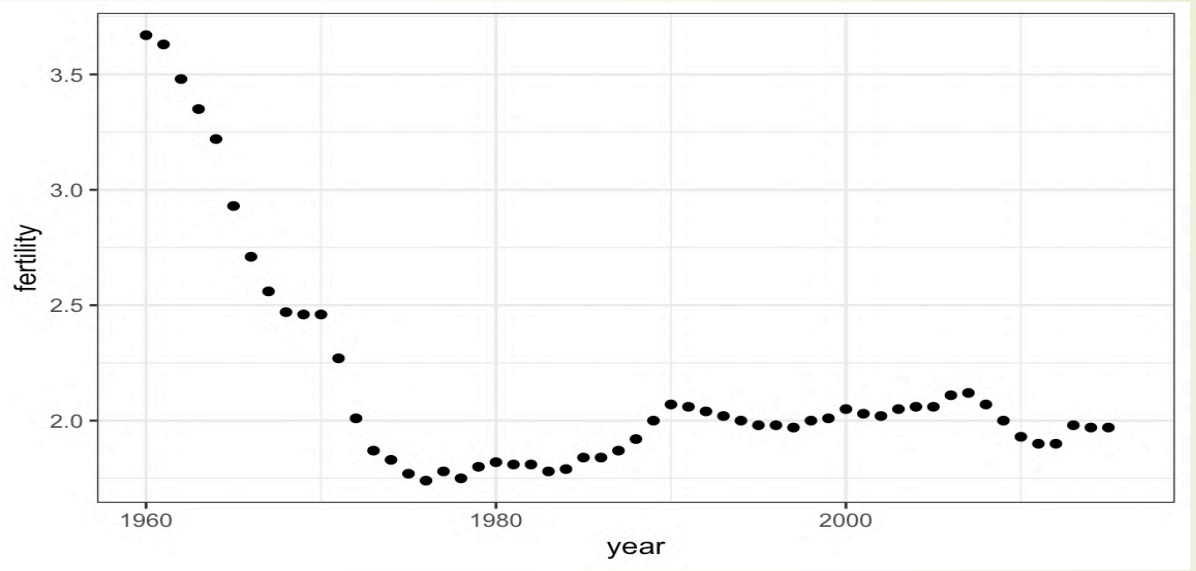
An example: Gapminder data

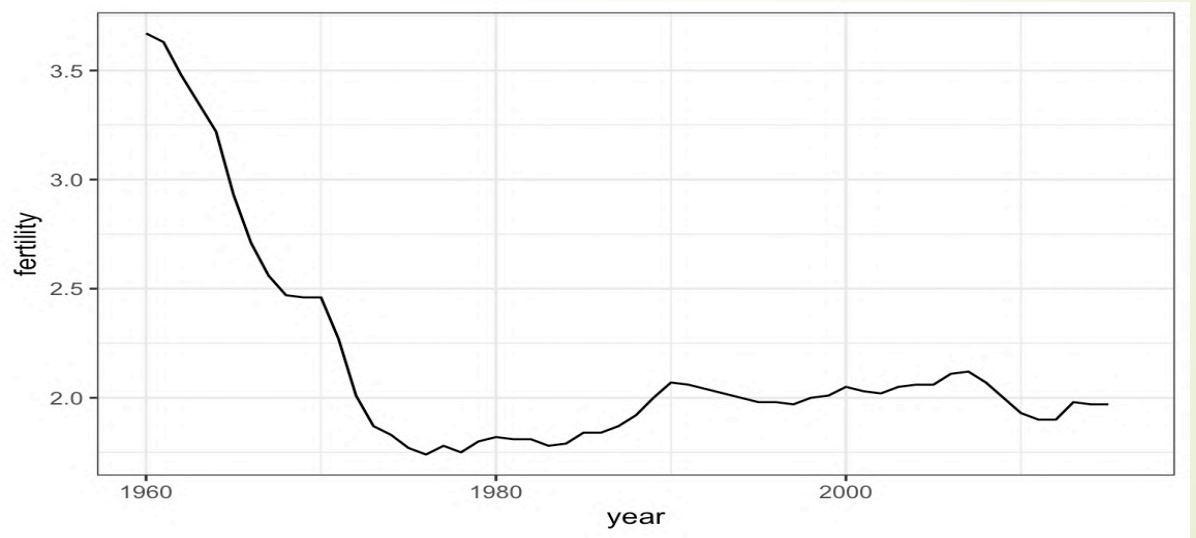


Time series plots

Time series plots

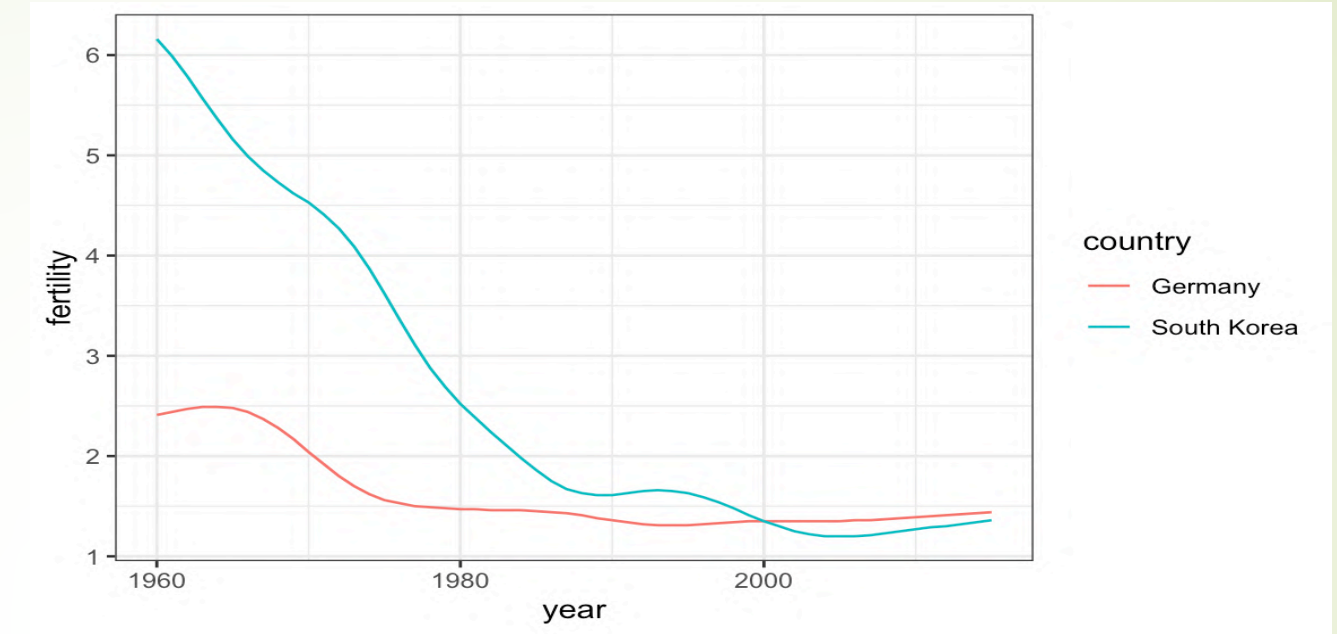
- Use different line styles

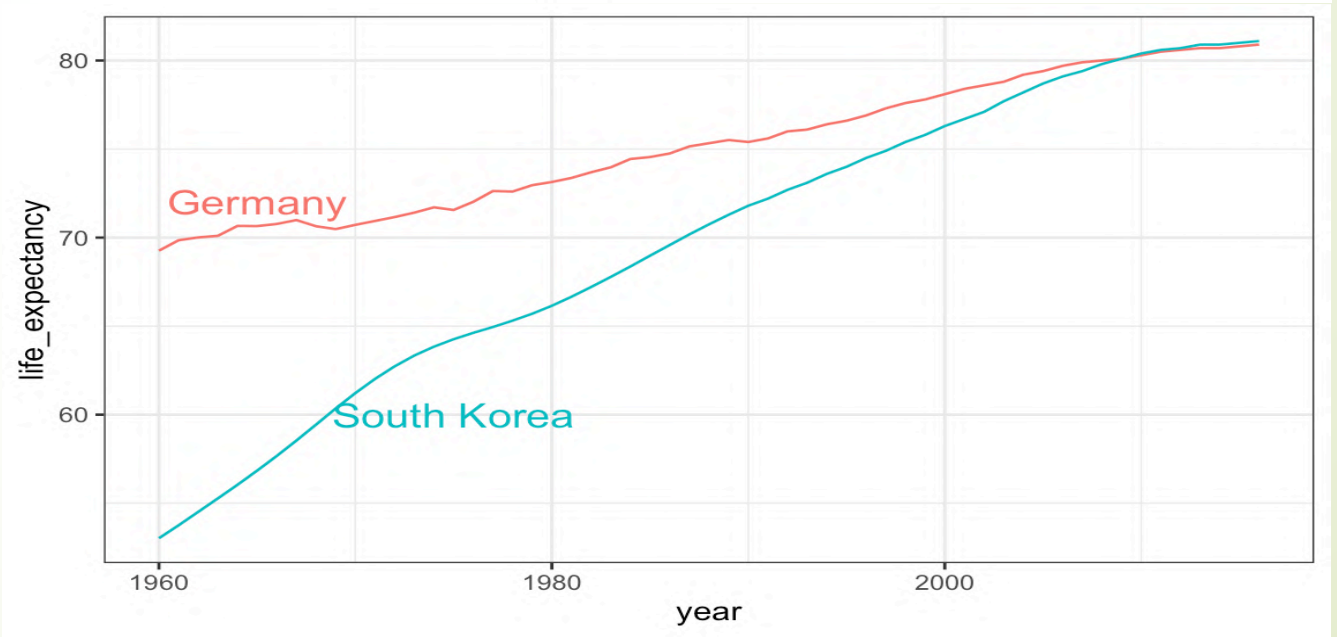




Time series plots

- Use different line styles
- Use colors and legends





PART-3: Data transformation

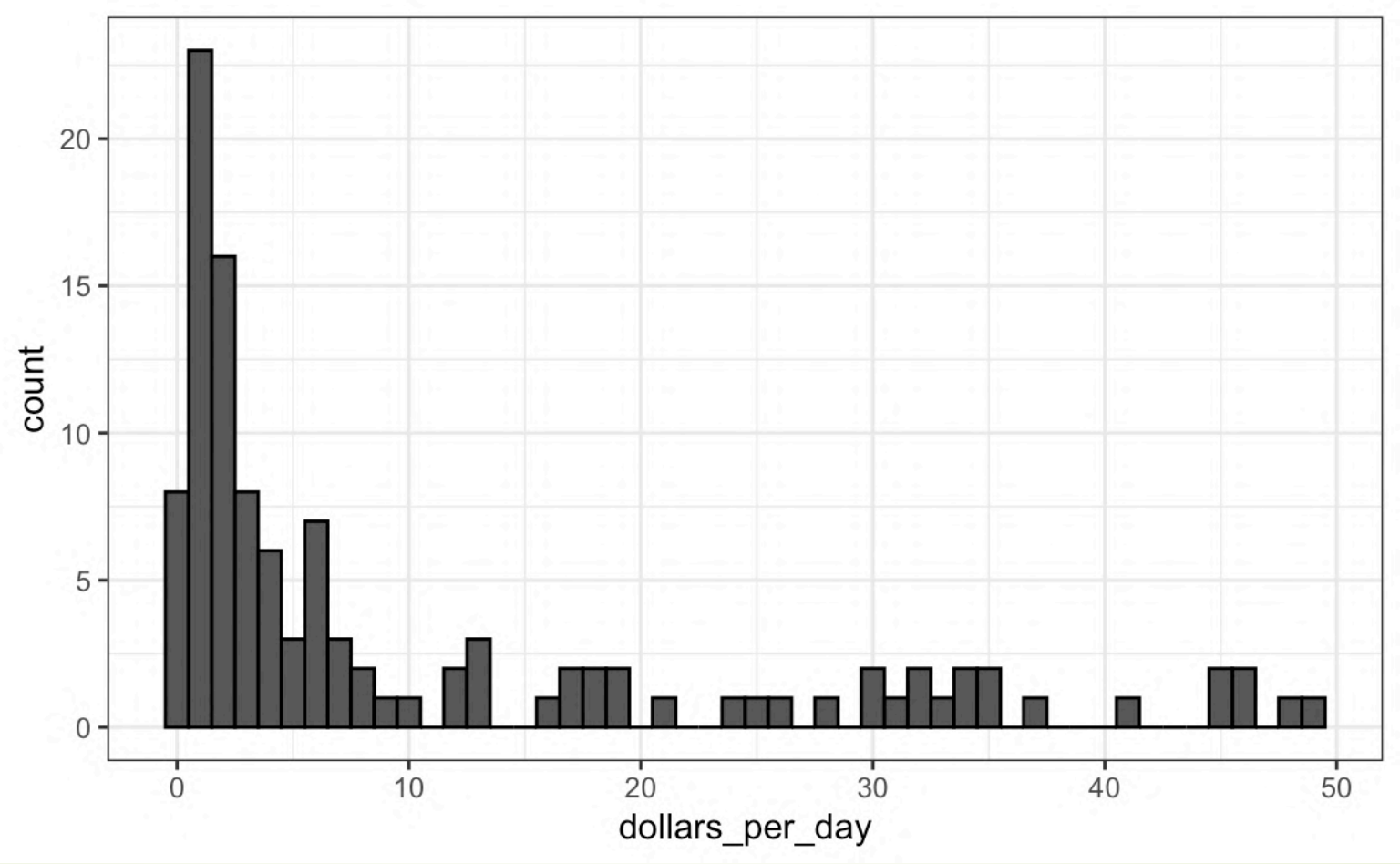
1. Log-transform

1. Log-transform

- Example: Histogram of US per-day income in 1970

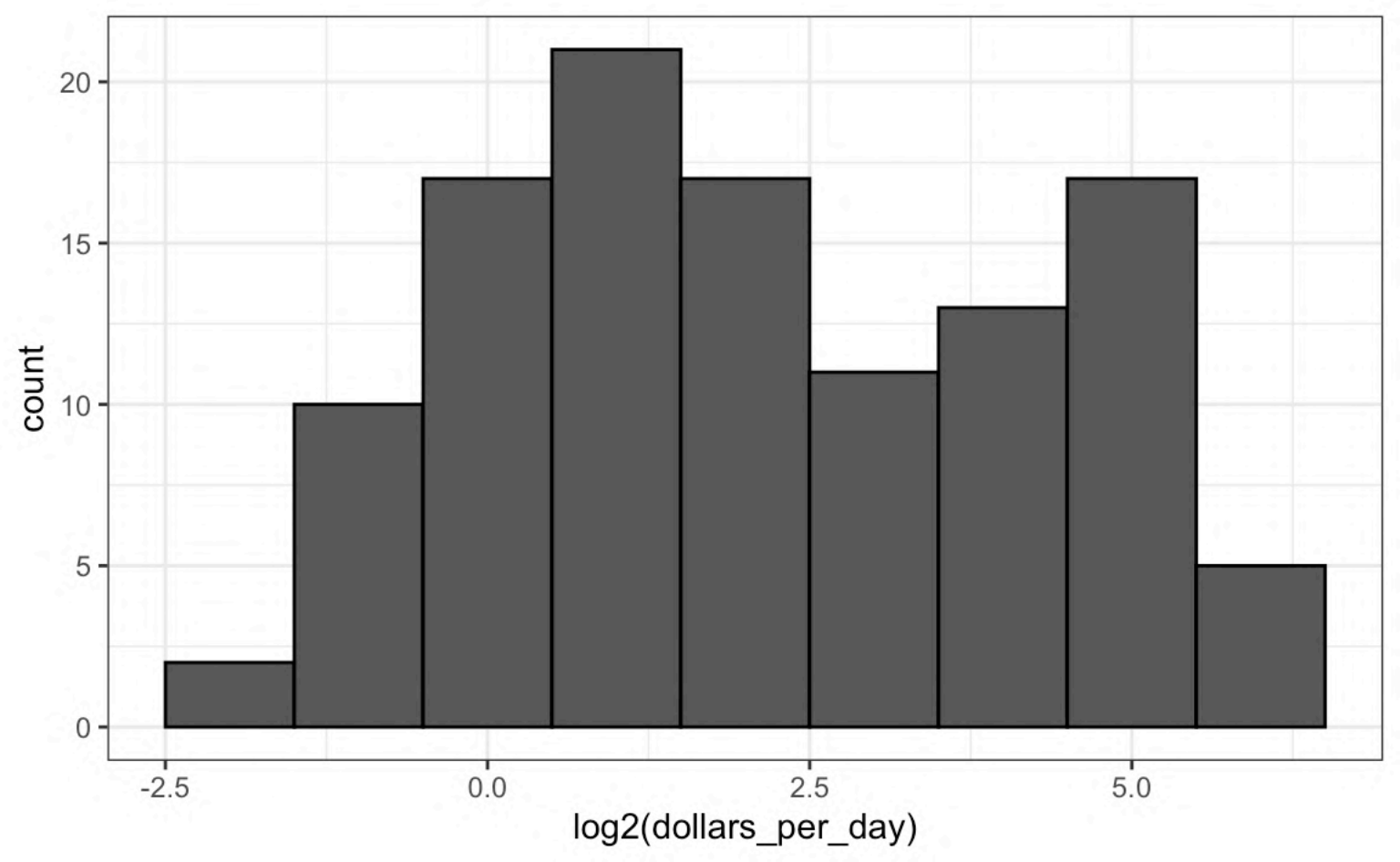
1. Log-transform

- Example: Histogram of US per-day income in 1970



1. Log-transform

- Example: Histogram of US per-day income in 1970



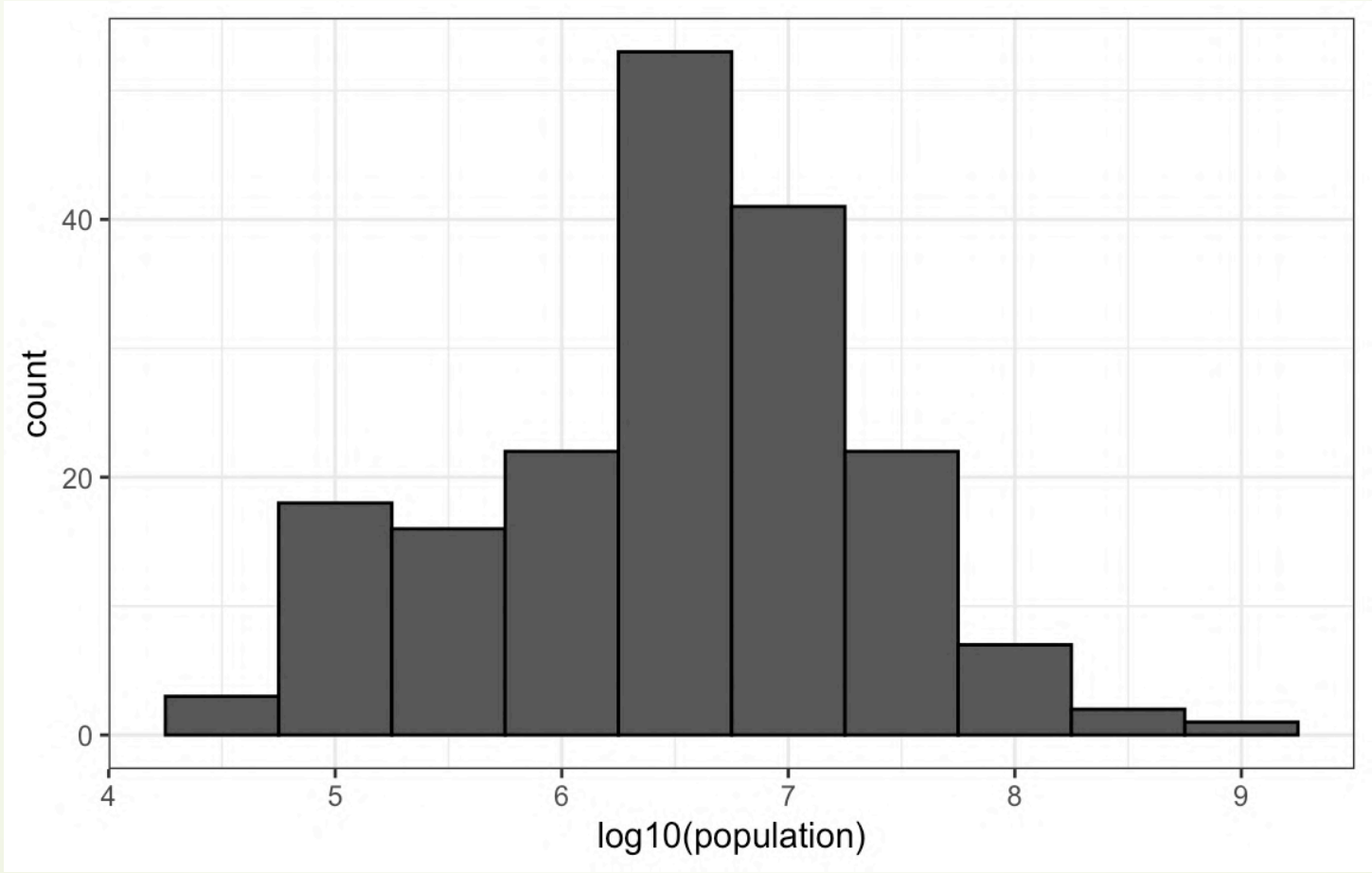
Log-transform

Log-transform

- A different base might be better: Population

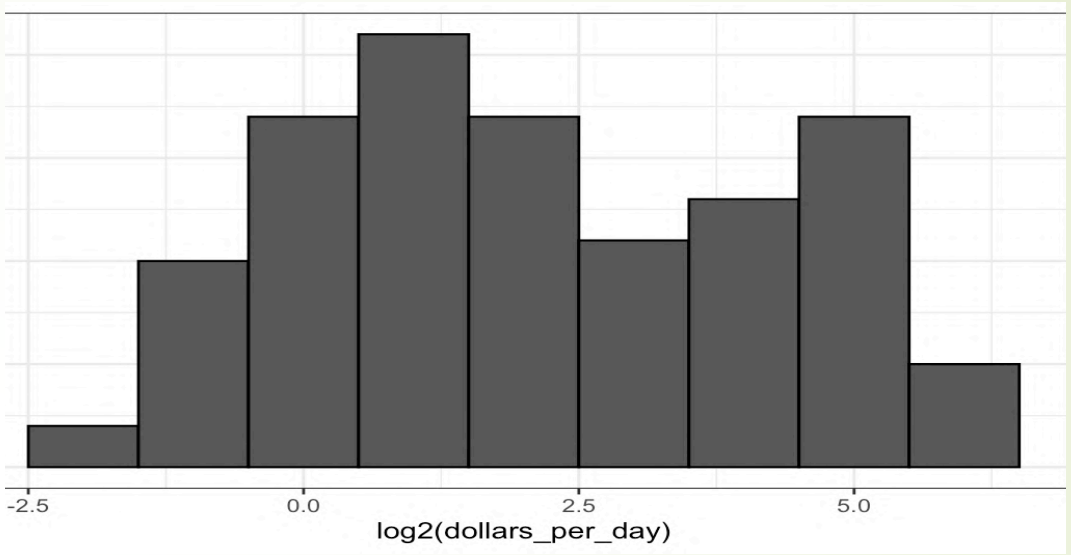
Log-transform

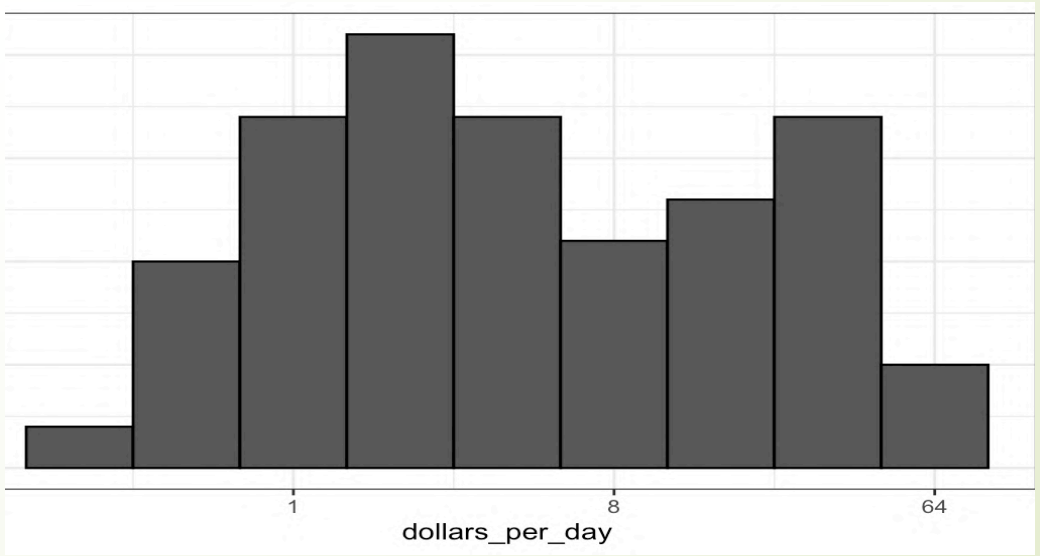
- A different base might be better: Population



Transform value or scale?

- Transforming scale sometimes provides better interpretation





2. Tidying data

2. Tidying data

- Tidying: **Structuring datasets to facilitate analysis**

2. Tidying data

- Tidying: **Structuring datasets to facilitate analysis**

- Rules of thumb
  - Functional relationships are easiest to see through columns
  - Group comparisons are easiest to see through rows
  - An observation is the smallest unit you'd like to draw conclusions or make predictions about

2. Tidying data

- Tidying: **Structuring datasets to facilitate analysis**

- Rules of thumb
  - Functional relationships are easiest to see through columns
  - Group comparisons are easiest to see through rows
  - An observation is the smallest unit you'd like to draw conclusions or make predictions about

- *Codd's 3rd normal form*:
  - Each variable forms a column.
  - Each observation forms a row.
  - Each type of observational unit forms a table.

2. Tidying data

- Tidying: **Structuring datasets to facilitate analysis**

- Rules of thumb
  - Functional relationships are easiest to see through columns
  - Group comparisons are easiest to see through rows
  - An observation is the smallest unit you'd like to draw conclusions or make predictions about

- *Codd's 3rd normal form*:
  - Each variable forms a column.
  - Each observation forms a row.
  - Each type of observational unit forms a table.

- A simple example

2. Tidying data

- Tidying: **Structuring datasets to facilitate analysis**

- Rules of thumb
  - Functional relationships are easiest to see through columns
  - Group comparisons are easiest to see through rows
  - An observation is the smallest unit you'd like to draw conclusions or make predictions about

- *Codd's 3rd normal form*:
  - Each variable forms a column.
  - Each observation forms a row.
  - Each type of observational unit forms a table.

- A simple example
- Case study: billboard data

2. Tidying data

- Tidying: **Structuring datasets to facilitate analysis**

- Rules of thumb
  - Functional relationships are easiest to see through columns
  - Group comparisons are easiest to see through rows
  - An observation is the smallest unit you'd like to draw conclusions or make predictions about

- *Codd's 3rd normal form*:
  - Each variable forms a column.
  - Each observation forms a row.
  - Each type of observational unit forms a table.

- A simple example
- Case study: billboard data

2. Tidying data

- Tidying: **Structuring datasets to facilitate analysis**

- Rules of thumb
  - Functional relationships are easiest to see through columns
  - Group comparisons are easiest to see through rows
  - An observation is the smallest unit you'd like to draw conclusions or make predictions about

- *Codd's 3rd normal form*:
  - Each variable forms a column.
  - Each observation forms a row.
  - Each type of observational unit forms a table.

- A simple example
- Case study: billboard data

2. Tidying data

- Tidying: **Structuring datasets to facilitate analysis**

- Rules of thumb
  - Functional relationships are easiest to see through columns
  - Group comparisons are easiest to see through rows
  - An observation is the smallest unit you'd like to draw conclusions or make predictions about

- *Codd's 3rd normal form*:
  - Each variable forms a column.
  - Each observation forms a row.
  - Each type of observational unit forms a table.

- A simple example
- Case study: billboard data

2. Tidying data

- Tidying: **Structuring datasets to facilitate analysis**

- Rules of thumb
  - Functional relationships are easiest to see through columns
  - Group comparisons are easiest to see through rows
  - An observation is the smallest unit you'd like to draw conclusions or make predictions about

- *Codd's 3rd normal form*:
  - Each variable forms a column.
  - Each observation forms a row.
  - Each type of observational unit forms a table.

- A simple example
- Case study: billboard data

2. Tidying data

- Tidying: **Structuring datasets to facilitate analysis**

- Rules of thumb
  - Functional relationships are easiest to see through columns
  - Group comparisons are easiest to see through rows
  - An observation is the smallest unit you'd like to draw conclusions or make predictions about

- *Codd's 3rd normal form*:
  - Each variable forms a column.
  - Each observation forms a row.
  - Each type of observational unit forms a table.

- A simple example
- Case study: billboard data

2. Tidying data

- Tidying: **Structuring datasets to facilitate analysis**

- Rules of thumb
  - Functional relationships are easiest to see through columns
  - Group comparisons are easiest to see through rows
  - An observation is the smallest unit you'd like to draw conclusions or make predictions about

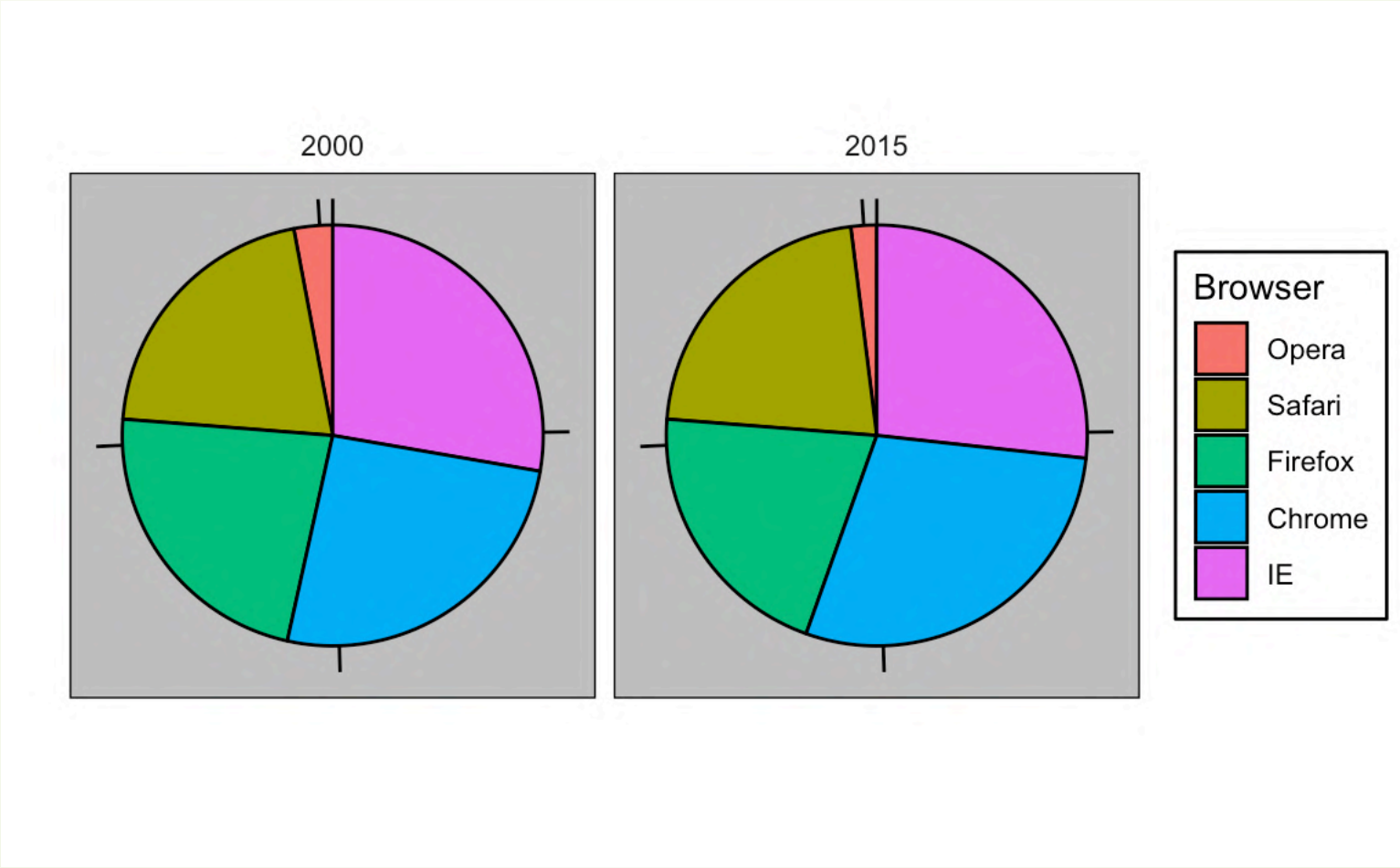
- *Codd's 3rd normal form*:
  - Each variable forms a column.
  - Each observation forms a row.
  - Each type of observational unit forms a table.

- A simple example
- Case study: billboard data

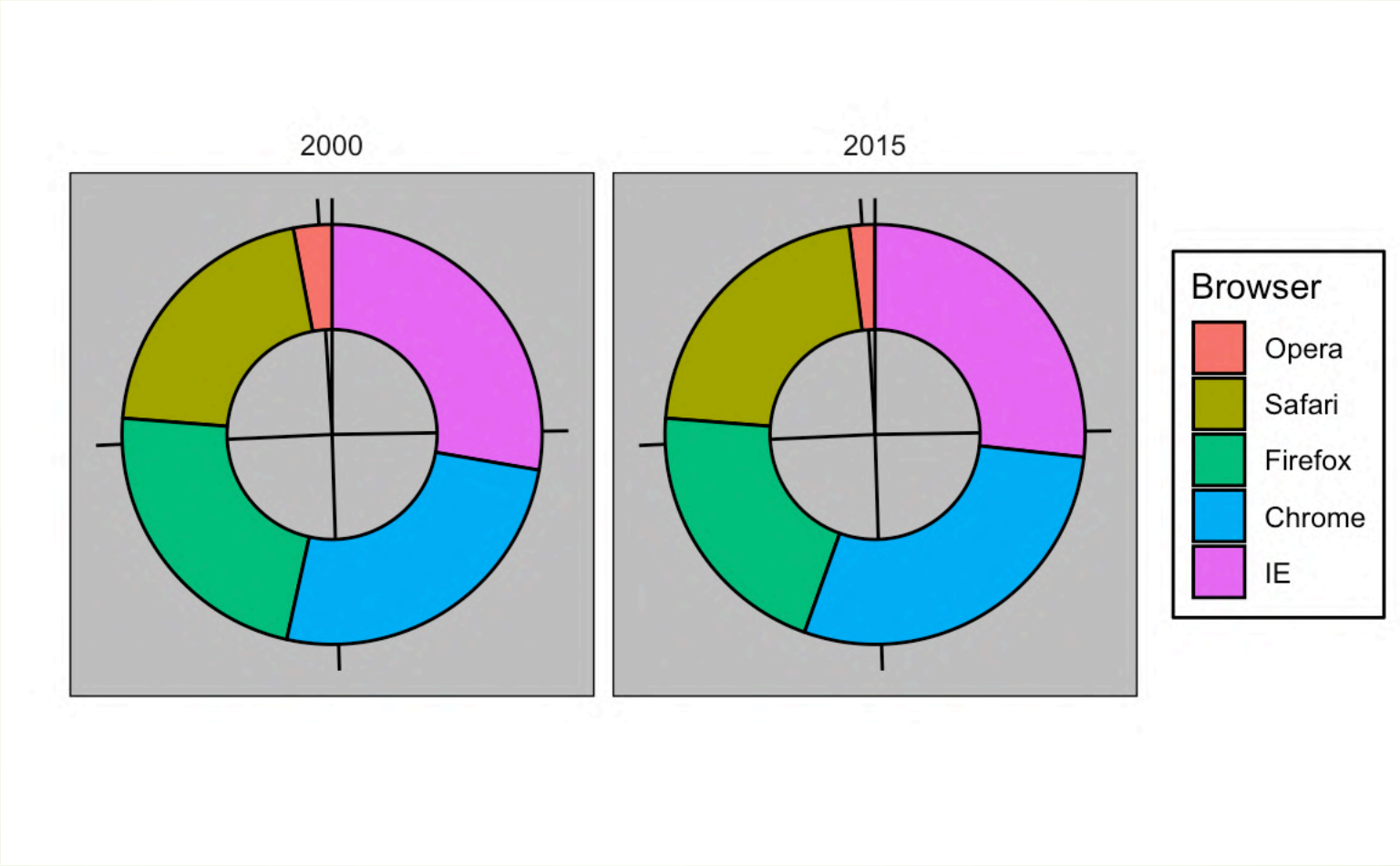
PART-4: Principles of Data Visualization

1. Use visual cue

1. Use visual cue



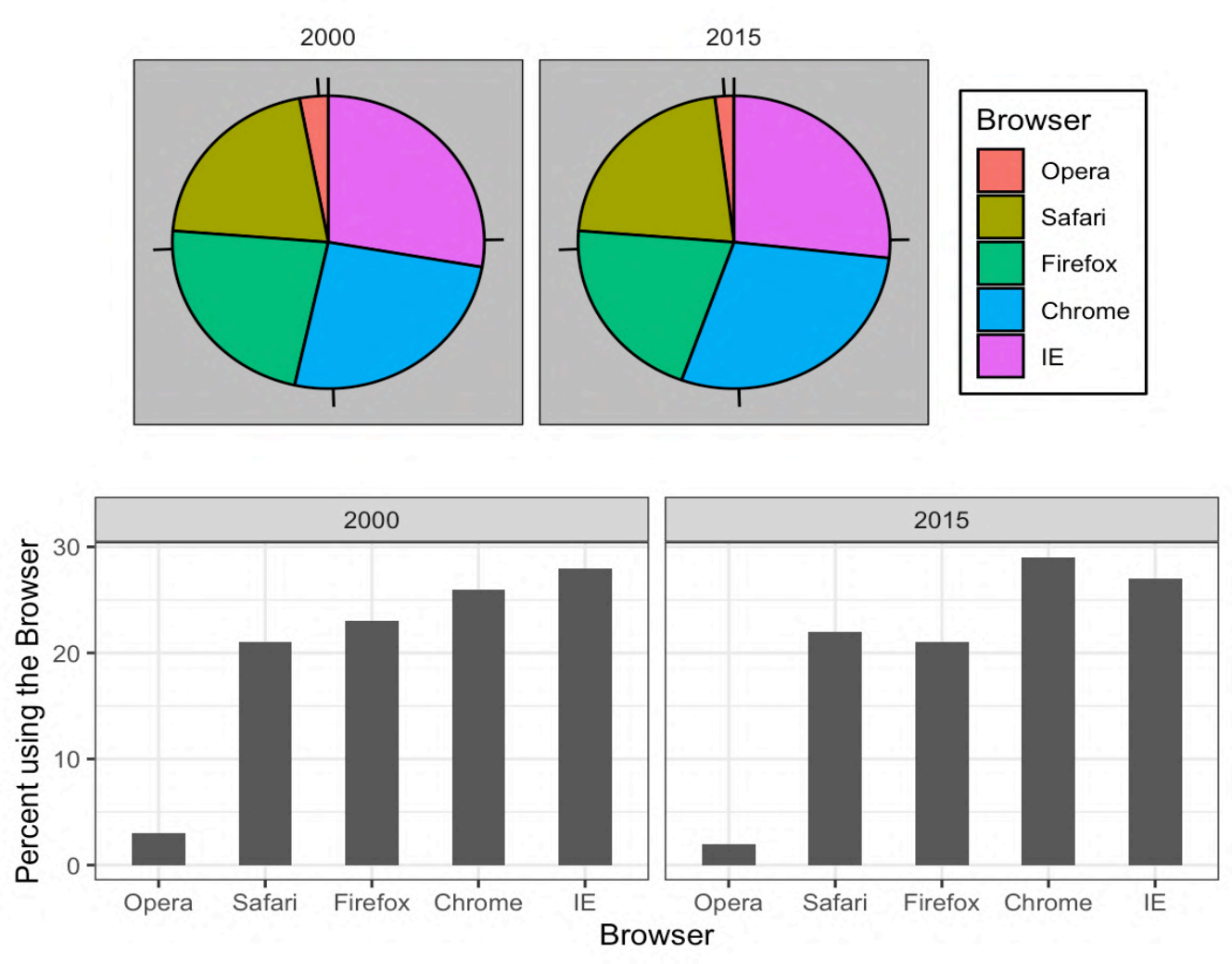
1. Use visual cue



1. Use visual cue

Browser	2000	2015
Opera	3	2
Safari	21	22
Firefox	23	21
Chrome	26	29
IE	28	27

1. Use visual cue



2. Know when to include zero

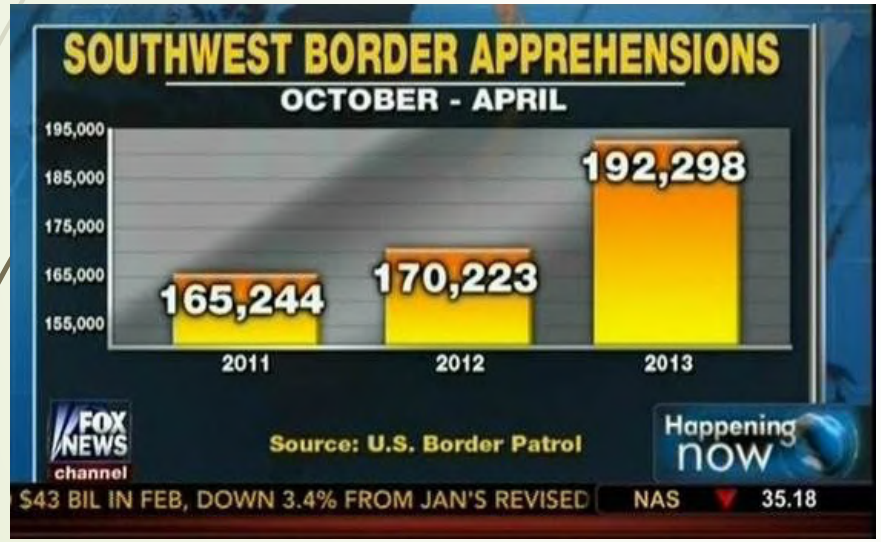
2. Know when to include zero

Important to include when **comparison is made based on length**

2. Know when to include zero

Important to include when **comparison is made based on length**

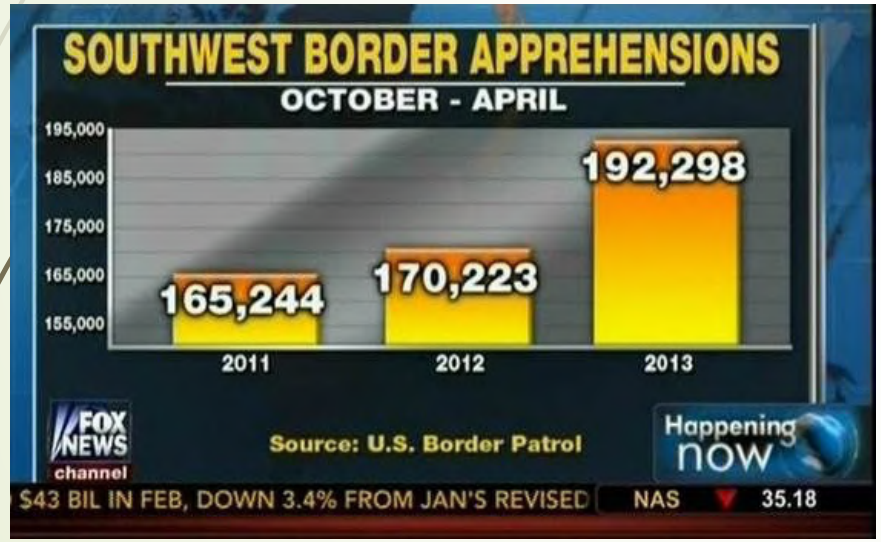
(Source: Fox News, via Media Matters)

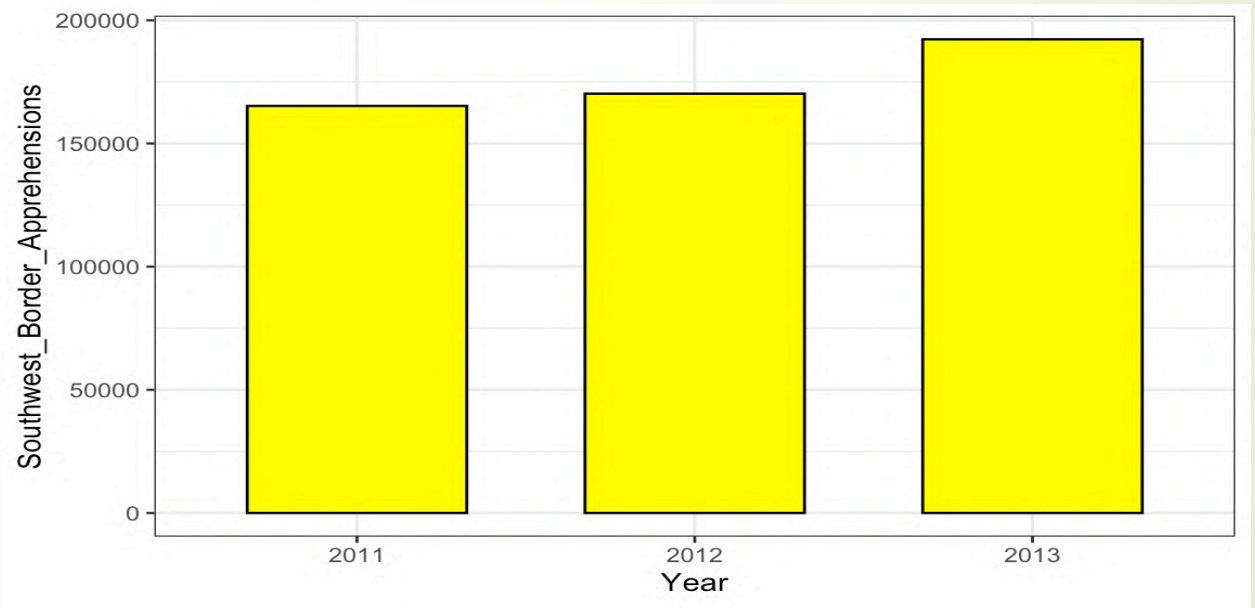


2. Know when to include zero

Important to include when **comparison is made based on length**

(Source: Fox News, via Media Matters)





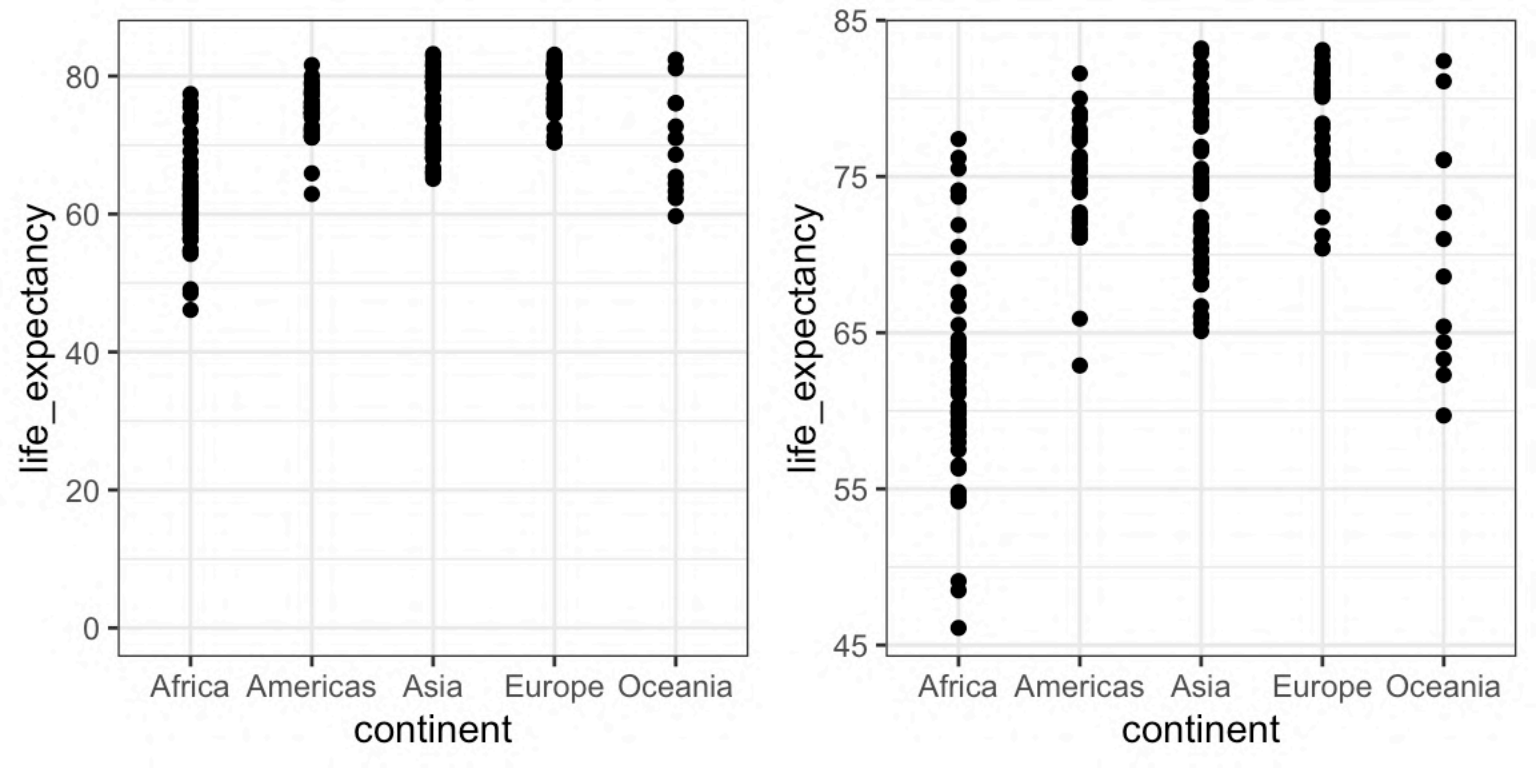
2. Know when to include zero

2. Know when to include zero

- **Not important** when **position is used for comparison**

2. Know when to include zero

- **Not important** when **position is used for comparison**



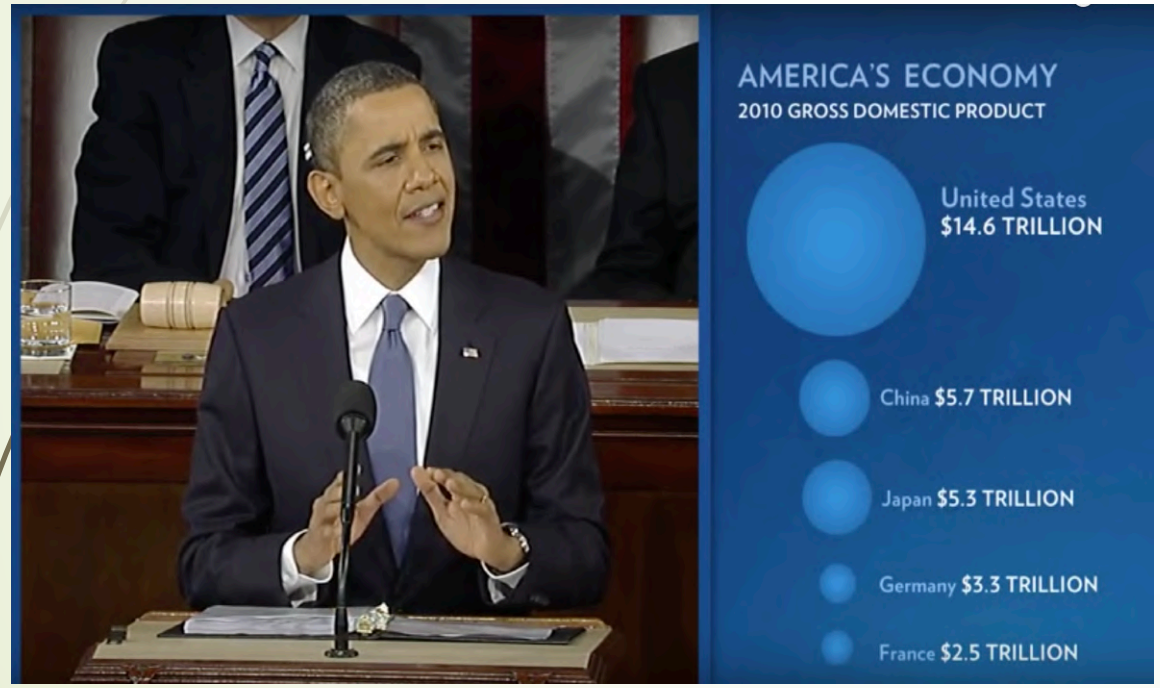
3. Do not distort quantities

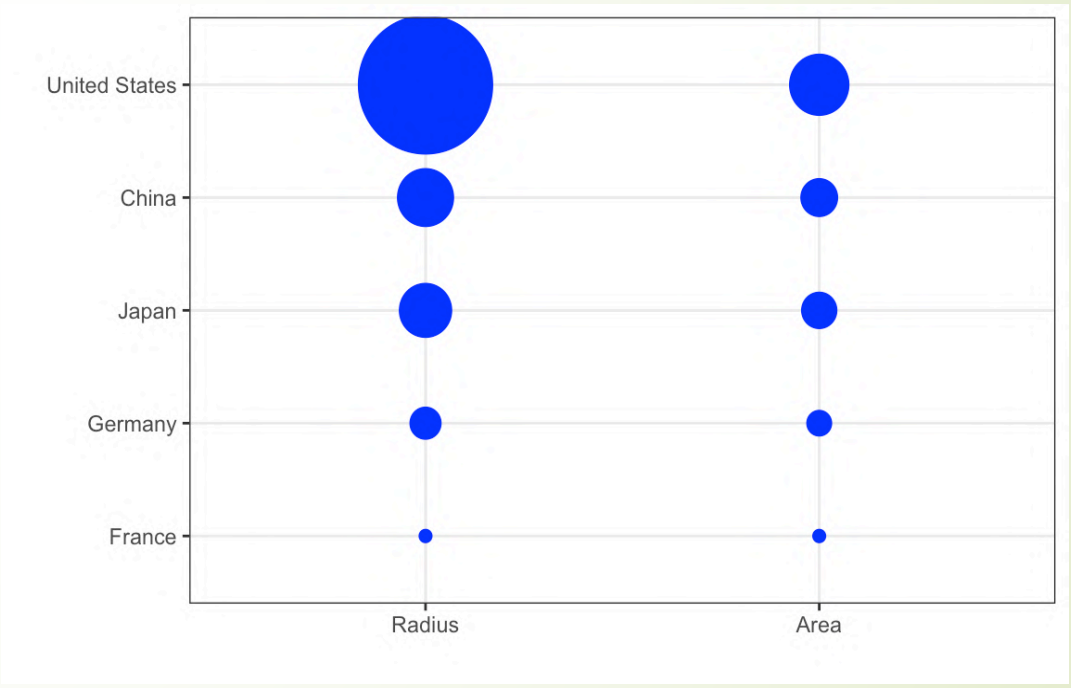
3. Do not distort quantities

(Source: The 2011 State of the Union Address)

3. Do not distort quantities

(Source: The 2011 State of the Union Address)





3. Do not distort quantities

(Source: The 2011 State of the Union Address)