CLEMS☣N
U N I V E R S I T Y

Lecture 2
Exploratory Data Analysis I
Readings: IntroStat Chapters 2-3; OpenIntro Chapter 2

STAT 8010 Statistical Methods I
May 17, 2023

Whitney Huang
Clemson University

Agenda

Exploratory Data
Analysis I

CLEMSON
U N I V E R S I T Y

Summarizing
Categorical Data

Summarizing
Numerical Data

1 **Summarizing Categorical Data**

2 **Summarizing Numerical Data**

CLEMSON
U N I V E R S I T Y

Summarizing
Categorical Data

Summarizing
Numerical Data

Last Lecture

- Stating the problem, identifying the variable(s) of interest, and gathering data

    - Types of variables and datasets

    - Observational vs. Experimental Studies

    - Methods of sampling

- Summarizing the data

- Analyzing the data

- Reporting and interpreting the results

Exploratory Data Analysis I

CLEMS☼N
U N I V E R S I T Y

Summarizing
Categorical Data

Summarizing
Numerical Data

- Stating the problem, identifying the variable(s) of interest, and gathering data

    - Types of variables and datasets

    - Observational vs. Experimental Studies

    - Sampling Techniques

- Summarizing the data

- Analyzing the data

- Reporting and interpreting the results

CLEMSON
U N I V E R S I T Y

Summarizing Categorical Variables

Example: Sport Injuries

The paper *"Profile of sport/leisure injuries treated at emergency rooms of urban hospitals."* by Pelletier et al. 1991 examined the nature and number of sport/leisure injuries treated in hospital emergency rooms in a large metropolitan city. They classified non-contact sports injuries by sport, resulting in the following data set:

Sport
Soccer
Basketball
Others
Basketball
Touch Football
Others
Touch Football
Volleyball
Baseball/softball
⋮

Question: How to summarize this data set?

Frequency Table

Exploratory Data Analysis I

CLEMSON
UNIVERSITY

Summarizing
Categorical Data

Summarizing
Numerical Data

- A frequency table for **categorical data** is a table that displays the possible categories along with the associated **frequencies** or **relative frequencies**

- The frequency for a particular category is the number of times the category appears in the data set

- The relative frequency for a particular category is the fraction or proportion of the time that the category appears in the data set.

Frequencies and Relative Frequencies

```
> table(sport)
sport
Baseball/softball        Basketball          Bicycling    Jogging/running
              11                  19                 11                 11
          Others              Soccer     Touch Football         Volleyball
              47                  24                 38                 17
> table(sport) / dim(sport)[1]
sport
Baseball/softball        Basketball          Bicycling    Jogging/running
      0.06179775          0.10674157         0.06179775         0.06179775
          Others              Soccer     Touch Football         Volleyball
      0.26404494          0.13483146         0.21348315         0.09550562
```
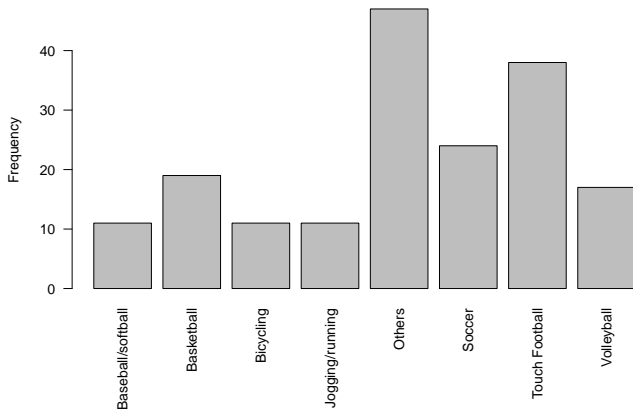
Frequencies and Relative Frequencies

Exploratory Data
Analysis I

CLEMSON
U N I V E R S I T Y

Summarizing
Categorical Data

Summarizing
Numerical Data

```
> table(sport)
sport
Baseball/softball        Basketball           Bicycling    Jogging/running
              11                  19                  11                 11
          Others              Soccer      Touch Football         Volleyball
              47                  24                  38                 17
> table(sport) / dim(sport)[1]
sport
Baseball/softball        Basketball           Bicycling    Jogging/running
      0.06179775          0.10674157          0.06179775         0.06179775
          Others              Soccer      Touch Football         Volleyball
      0.26404494          0.13483146          0.21348315         0.09550562
```
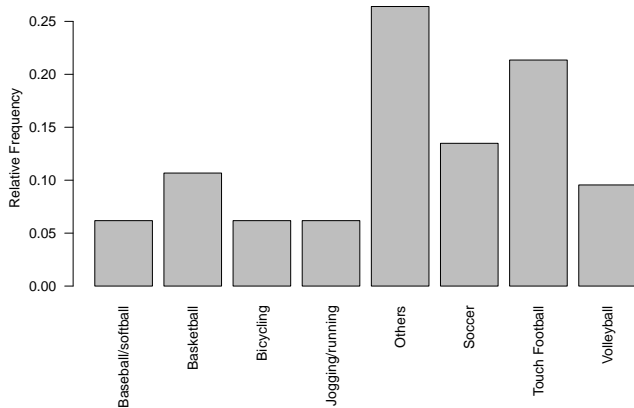
How could we visualize these information?
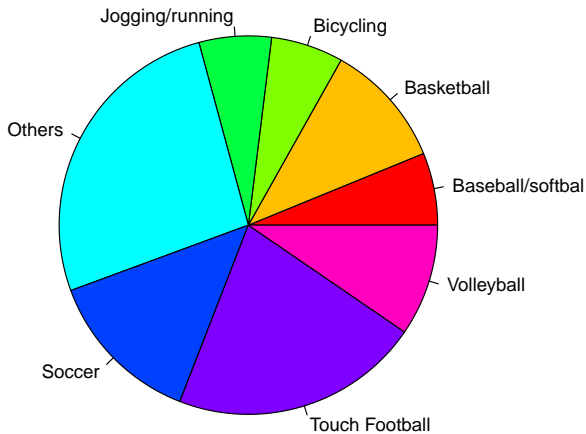⇒ Making a bar chart and/or a pie chart

Bar Charts

Exploratory Data
Analysis I

CLEMS❀N
U N I V E R S I T Y

Summarizing
Categorical Data

Summarizing
Numerical Data

A bar chart draws a bar with a height proportional to the count in the table:

Bar Charts cont'd

Pie Charts

Exploratory Data
Analysis I

CLEMS☘N
U N I V E R S I T Y

Summarizing
Categorical Data

Summarizing
Numerical Data

Exploratory Data
Analysis I

CLEMSON
U N I V E R S I T Y

Summarizing
Categorical Data

Summarizing
Numerical Data

Pie Charts cont'd

Exploratory Data Analysis I

CLEMS😼N
U N I V E R S I T Y

Summarizing
Categorical Data

Summarizing
Numerical Data
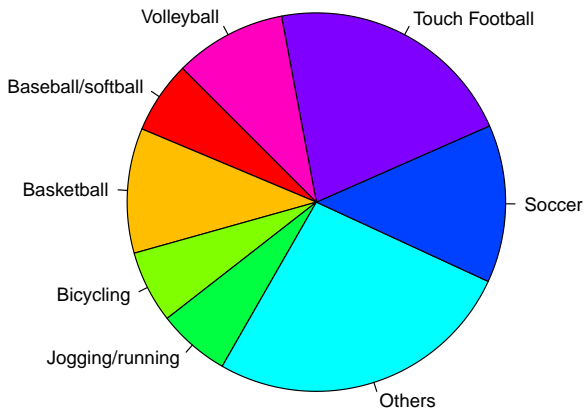
Discussion: Which one you prefer to visualize categorical variables. Why?

A Good Bar Chart

Exploratory Data
Analysis I

CLEMSON
U N I V E R S I T Y

Summarizing
Categorical Data

Summarizing
Numerical Data
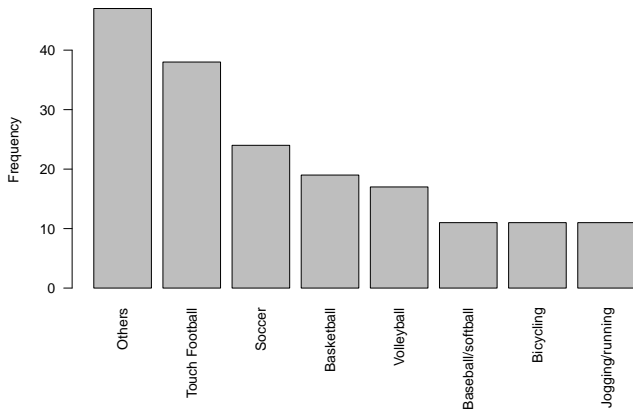
A (Potential) Misleading Bar Chart

CLEMS☼N
U N I V E R S I T Y

Same Data, Different Y-Axis
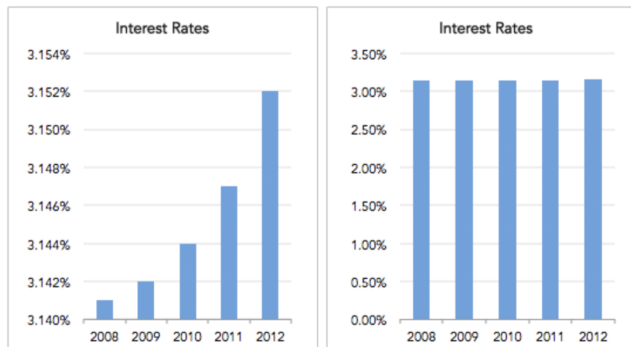
Example: O'Hare Airport Flight Data

Exploratory Data
Analysis I

CLEMS❀N
U N I V E R S I T Y

Summarizing
Categorical Data

Summarizing
Numerical Data

```
    carrier  origin
1       UA     EWR
2       AA     LGA
3       AA     LGA
4       AA     LGA
5       UA     LGA
6       UA     EWR
```
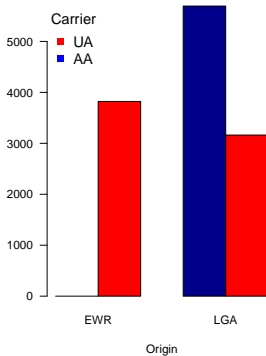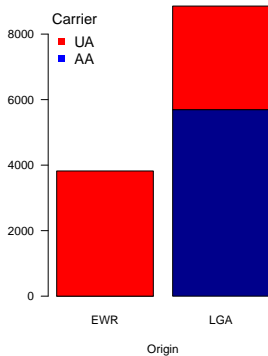
In this example, we have two categorical variables, `carrier` and `origin`, respectively. How to summarize/visualize this dataset?

ORD Flight Data Cont'd

CLEMSON
U N I V E R S I T Y

Summarizing
Categorical Data

Summarizing
Numerical Data

CLEMSON
U N I V E R S I T Y

Summarizing Numerical Variables

Example: Murder arrests (per 100,000) in US States in 1973

Data: 13.2, 10.0, 8.1, 8.8, 9.0, 7.9, 3.3, 5.9, 15.4, 17.4, 5.3, 2.6, 10.4, 7.2, 2.2, 6.0, 9.7, 15.4, 2.1, 11.3, 4.4, 12.1, 2.7, 16.1, 9.0, 6.0, 4.3, 12.2, 2.1, 7.4, 11.4, 11.1, 13.0, 0.8, 7.3, 6.6, 4.9, 6.3, 3.4, 14.4, 3.8, 13.2, 12.7, 3.2, 2.2, 8.5, 4.0, 5.7, 2.6, 6.8.
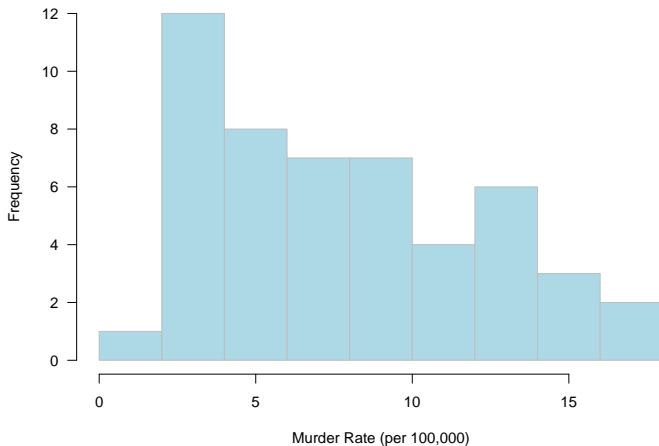
Question: How to graphically summarize this data set?

Stem-and-Leaf Plot

CLEMS☙N
U N I V E R S I T Y

Summarizing
Categorical Data

Summarizing
Numerical Data

```
The decimal point is at the |

 0 | 8
 1 |
 2 | 1122667
 3 | 2348
 4 | 0349
 5 | 379
 6 | 00368
 7 | 2349
 8 | 158
 9 | 007
10 | 04
11 | 134
12 | 127
13 | 022
14 | 4
15 | 44
16 | 1
17 | 4
```
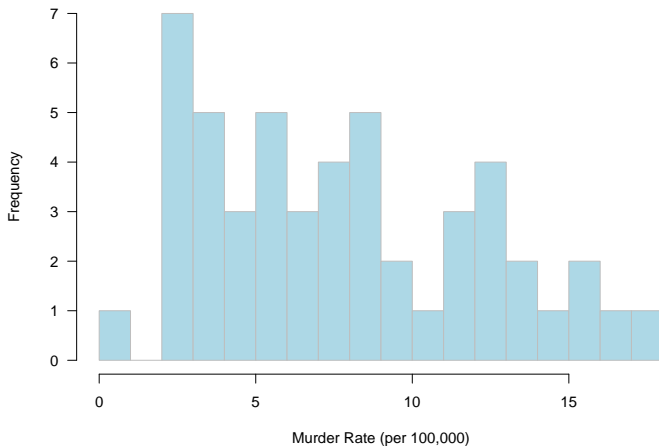
Histogram

Exploratory Data
Analysis I

CLEMS✹N
U N I V E R S I T Y

Summarizing
Categorical Data

Summarizing
Numerical Data

Histogram of US Murder Rate in 1973

Histogram

Exploratory Data
Analysis I

CLEMSON
U N I V E R S I T Y

Summarizing
Categorical Data

Summarizing
Numerical Data

Histogram of US Murder Rate in 1973

Box-and-Whisker Plot

Exploratory Data
Analysis I

CLEMS●N
U N I V E R S I T Y

Summarizing
Categorical Data

Summarizing
Numerical Data

Murder Rate (per 100,000)

Shape of Distributions

CLEMSON
U N I V E R S I T Y

Summarizing
Categorical Data

Summarizing
Numerical Data

Negative Skew   Positive Skew
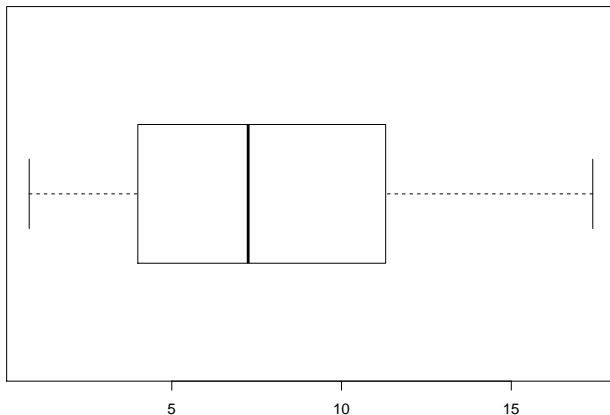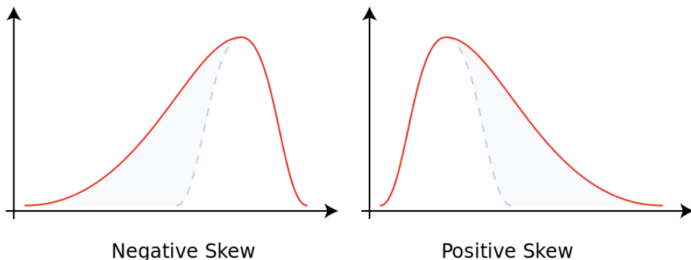
Source: Skewness - Wikipedia

In the rest of the class, we will talk about how to summarize a numerical variable in terms of its center and spread

Measures of Center

- A **measure of center** attempts to report a "typical" value for the variable

- When a measure of center is calculated with **sample data** it is a **statistic**

- When a measure of center is calculated with popular (e.g., census data) it is a **parameter**

- **Measures:** Mean, Median, Mode

Mean

Exploratory Data
Analysis I

CLEMS☙N
U N I V E R S I T Y

Summarizing
Categorical Data

Summarizing
Numerical Data

- The population mean, denoted by μ_X, is the sum of all the population values ($\{X_i, \cdots, X_N\}$) divided by the size of the population (N). That is,

$$\mu_X = \frac{\sum_{i=1}^{N} X_i}{N}$$

- The sample mean, denoted by $\bar{X}$ is the sum of all the sample values ($\{X_1, \cdots, X_n\}$) divided by the sample size (n). That is,

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Median

The median is the value separating the higher half from the lower half of a data sample

How to compute the median: Order the n observations in a data set from smallest to largest, then

$$\text{Median} = \begin{cases} \text{the single middle value,} & \text{n odd} \\ \text{the average of the middle two values,} & \text{n even} \end{cases}$$
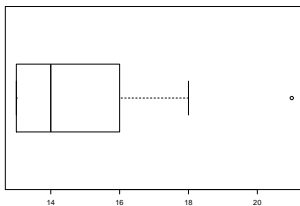
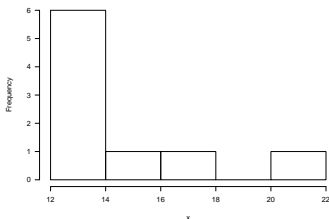# Mode

The mode is the value of the observation that appears most frequently

How to compute the mode(s): Order the observations in a data set from smallest to largest, then find the number that is repeated more often than any other

Example

Suppose we have the following list of values: 13, 18, 13, 14, 13, 16, 14, 21, 13

- Plot this "data set" and describe the shape of the distribution

Example cont'd

CLEMS😈N
U N I V E R S I T Y

Summarizing
Categorical Data

Summarizing
Numerical Data

Suppose we have the following list of values: 13, 18, 13, 14, 13, 16, 14, 21, 13

- Find the sample mean

$$\bar{X} = \sum_{i=1}^{9} \frac{13 + 18 + 13 + 14 + 13 + 16 + 14 + 21 + 13}{9} = 15$$

- Find the sample median
  1. Order the data first: 13, 13, 13, 13, 14, 14, 16, 18, 21

Example cont'd

Suppose we have the following list of values: 13, 18, 13, 14, 13, 16, 14, 21, 13

- Find the sample mean

$$\bar{X} = \sum_{i=1}^{9} \frac{13 + 18 + 13 + 14 + 13 + 16 + 14 + 21 + 13}{9} = 15$$

- Find the sample median
  1. Order the data first: 13, 13, 13, 13, 14, 14, 16, 18, 21

Example cont'd

Suppose we have the following list of values: 13, 18, 13, 14, 13, 16, 14, 21, 13

- Find the sample mean

$$\bar{X} = \sum_{i=1}^{9} \frac{13 + 18 + 13 + 14 + 13 + 16 + 14 + 21 + 13}{9} = 15$$

- Find the sample median

  1. Order the data first: 13, 13, 13, 13, 14, 14, 16, 18, 21

  2. Compute the sample size n and identify (or compute) the median value

Example cont'd

Suppose we have the following list of values: 13, 18, 13, 14, 13, 16, 14, 21, 13

- Find the sample mean

$$\bar{X} = \sum_{i=1}^{9} \frac{13 + 18 + 13 + 14 + 13 + 16 + 14 + 21 + 13}{9} = 15$$

- Find the sample median

1. Order the data first: 13, 13, 13, 13, 14, 14, 16, 18, 21

2. Compute the sample size n and identify (or compute) the median value

Example cont'd

Suppose we have the following list of values: 13, 18, 13, 14, 13, 16, 14, 21, 13

- Find the sample mean

$$\bar{X} = \sum_{i=1}^{9} \frac{13 + 18 + 13 + 14 + 13 + 16 + 14 + 21 + 13}{9} = 15$$

- Find the sample median

  1. Order the data first: 13, 13, 13, 13, 14, 14, 16, 18, 21

  2. Compute the sample size n and identify (or compute) the median value

  3. $n = 9 \Rightarrow$ the median is the 5th number, which is 14

Exploratory Data
Analysis I

CLEMSON
U N I V E R S I T Y

Summarizing
Categorical Data

Summarizing
Numerical Data

- Find the mode
  1. Order the data first: 13, 13, 13, 13, 14, 14, 16, 18, 21

Example cont'd

- Find the mode
  1. Order the data first: 13, 13, 13, 13, 14, 14, 16, 18, 21

Example cont'd

Exploratory Data
Analysis I

CLEMSON
U N I V E R S I T Y

Summarizing
Categorical Data

Summarizing
Numerical Data

- Find the mode
  1. Order the data first: 13, 13, 13, 13, 14, 14, 16, 18, 21

  2. We have 3 13 and 2 14 $\Rightarrow$ 13 is the mode

Example: Resistant (Robust) Statistics

Suppose we have the following list of values: 13, 18, 13, 14, 13, 16, 14, 210, 13

- Find the sample mean

$$\bar{X} = \sum_{i=1}^{9} \frac{13 + 18 + 13 + 14 + 13 + 16 + 14 + 210 + 13}{9} = 36$$

- Find the sample median
  1. Order the data first: 13, 13, 13, 13, 14, 14, 16, 18, 210

Example: Resistant (Robust) Statistics

Suppose we have the following list of values: 13, 18, 13, 14, 13, 16, 14, 210, 13

- Find the sample mean

$$\bar{X} = \sum_{i=1}^{9} \frac{13 + 18 + 13 + 14 + 13 + 16 + 14 + 210 + 13}{9} = 36$$

- Find the sample median
  1. Order the data first: 13, 13, 13, 13, 14, 14, 16, 18, 210

Example: Resistant (Robust) Statistics

Suppose we have the following list of values: 13, 18, 13, 14, 13, 16, 14, 210, 13

- Find the sample mean

$$\bar{X} = \sum_{i=1}^{9} \frac{13 + 18 + 13 + 14 + 13 + 16 + 14 + 210 + 13}{9} = 36$$

- Find the sample median
  1. Order the data first: 13, 13, 13, 13, 14, 14, 16, 18, 210
  2. Compute the sample size n and identify (or compute) the median value

Example: Resistant (Robust) Statistics

Suppose we have the following list of values: 13, 18, 13, 14, 13, 16, 14, 210, 13

- Find the sample mean

$$\bar{X} = \sum_{i=1}^{9} \frac{13 + 18 + 13 + 14 + 13 + 16 + 14 + 210 + 13}{9} = 36$$

- Find the sample median

  1. Order the data first: 13, 13, 13, 13, 14, 14, 16, 18, 210

  2. Compute the sample size n and identify (or compute) the median value

Exploratory Data
Analysis I

CLEMS✹N
U N I V E R S I T Y

Summarizing
Categorical Data

Summarizing
Numerical Data

Example: Resistant (Robust) Statistics

Suppose we have the following list of values: 13, 18, 13, 14, 13, 16, 14, 210, 13

- Find the sample mean

$$\bar{X} = \sum_{i=1}^{9} \frac{13 + 18 + 13 + 14 + 13 + 16 + 14 + 210 + 13}{9} = 36$$

- Find the sample median

  1. Order the data first: 13, 13, 13, 13, 14, 14, 16, 18, 210

  2. Compute the sample size n and identify (or compute) the median value

  3. $n = 9 \Rightarrow$ the median is the 5th number, which is (still) 14

Example cont'd

- Find the mode
  1. Order the data first: 13, 13, 13, 13, 14, 14, 16, 18, 210

Example cont'd

- Find the mode
  1. Order the data first: 13, 13, 13, 13, 14, 14, 16, 18, 210

Example cont'd

- Find the mode
  1. Order the data first: 13, 13, 13, 13, 14, 14, 16, 18, 210

  2. We have 3 13 and 2 14 $\Rightarrow$ 13 is (still) the mode

Example cont'd

- Find the mode
    1. Order the data first: 13, 13, 13, 13, 14, 14, 16, 18, 210
    2. We have 3 13 and 2 14 $\Rightarrow$ 13 is (still) the mode

Exploratory Data
Analysis I

Summarizing
Categorical Data

Summarizing
Numerical Data

Example cont'd

- Find the mode
  1. Order the data first: 13, 13, 13, 13, 14, 14, 16, 18, 210
  2. We have 3 13 and 2 14 $\Rightarrow$ 13 is (still) the mode

What is the take-home message?

Summary

Exploratory Data
Analysis I

CLEMSON
U N I V E R S I T Y

Summarizing
Categorical Data

Summarizing
Numerical Data

In this lecture, we learned

- Summarizing Categorical Data

- Summarizing the Central Tendency of Numerical Data

In next lecture we will learn

- How to summarize the spread of numerical data

- How to construct a boxplot

- How to visualize numerical + categorical variables and numerical + numerical variables