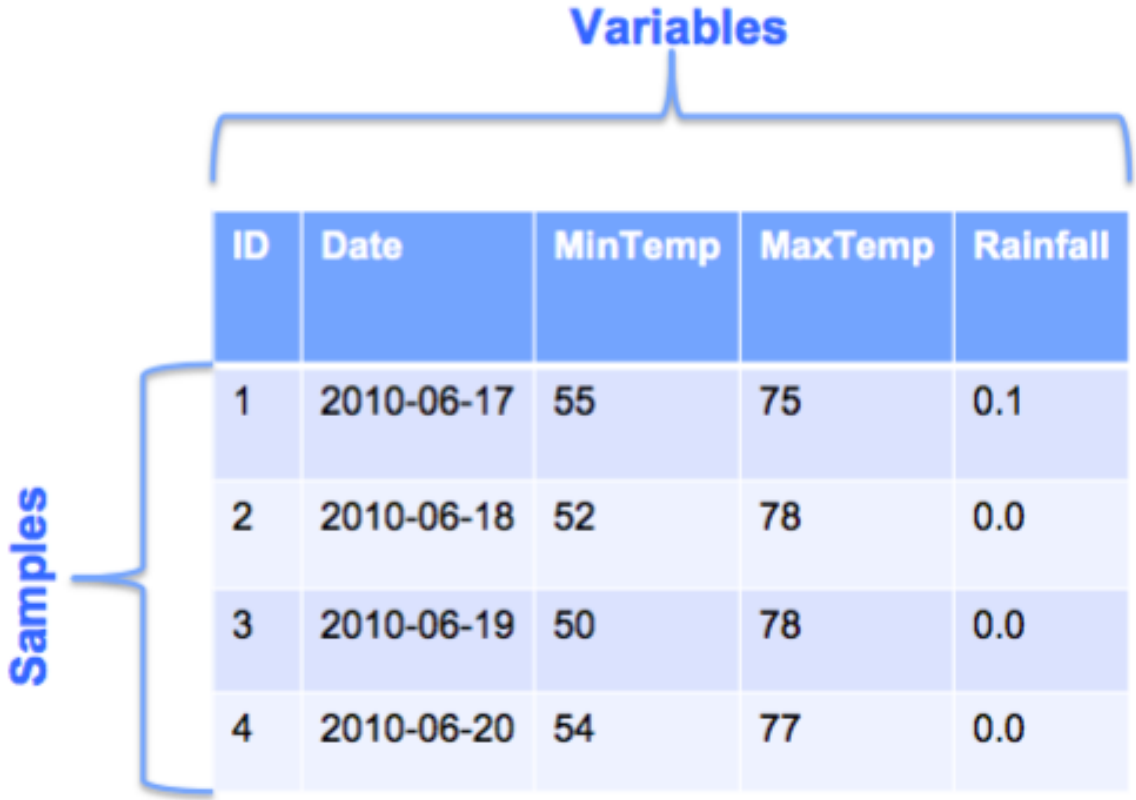
Python for Data Science

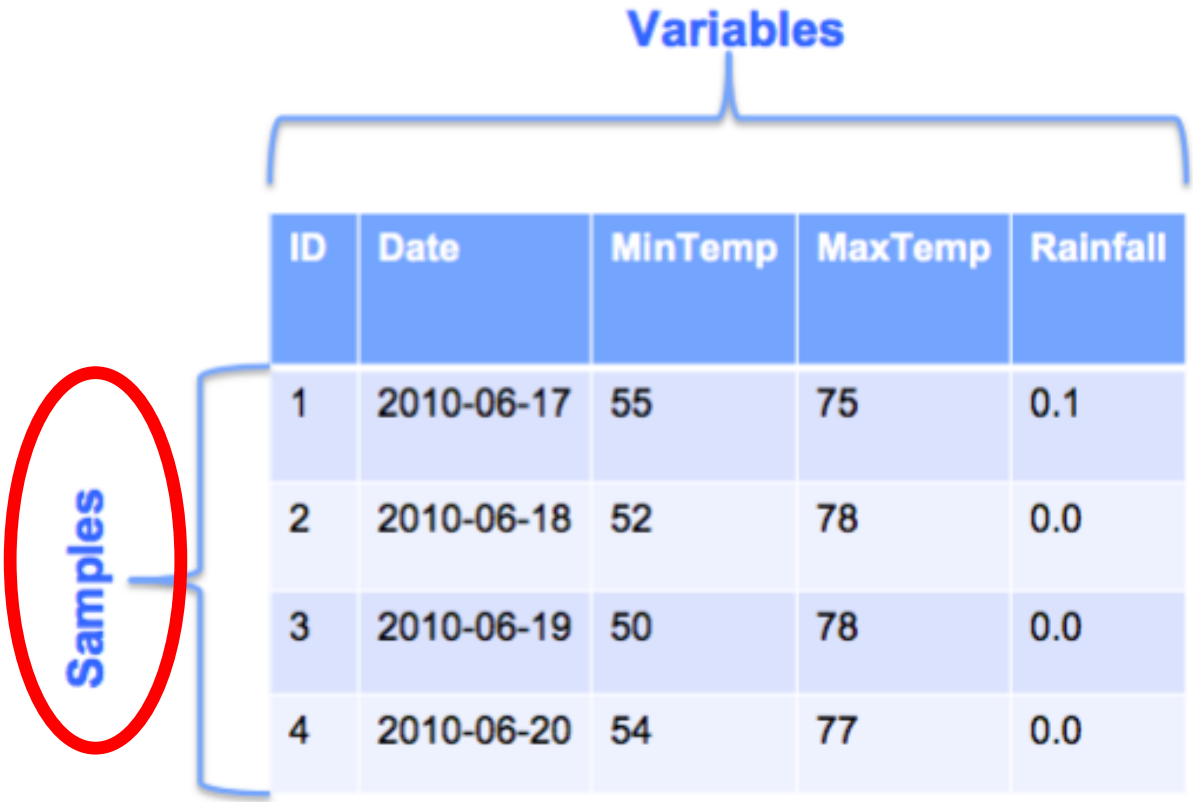
Exploratory Data Analysis

Dr. Ilkay Altintas

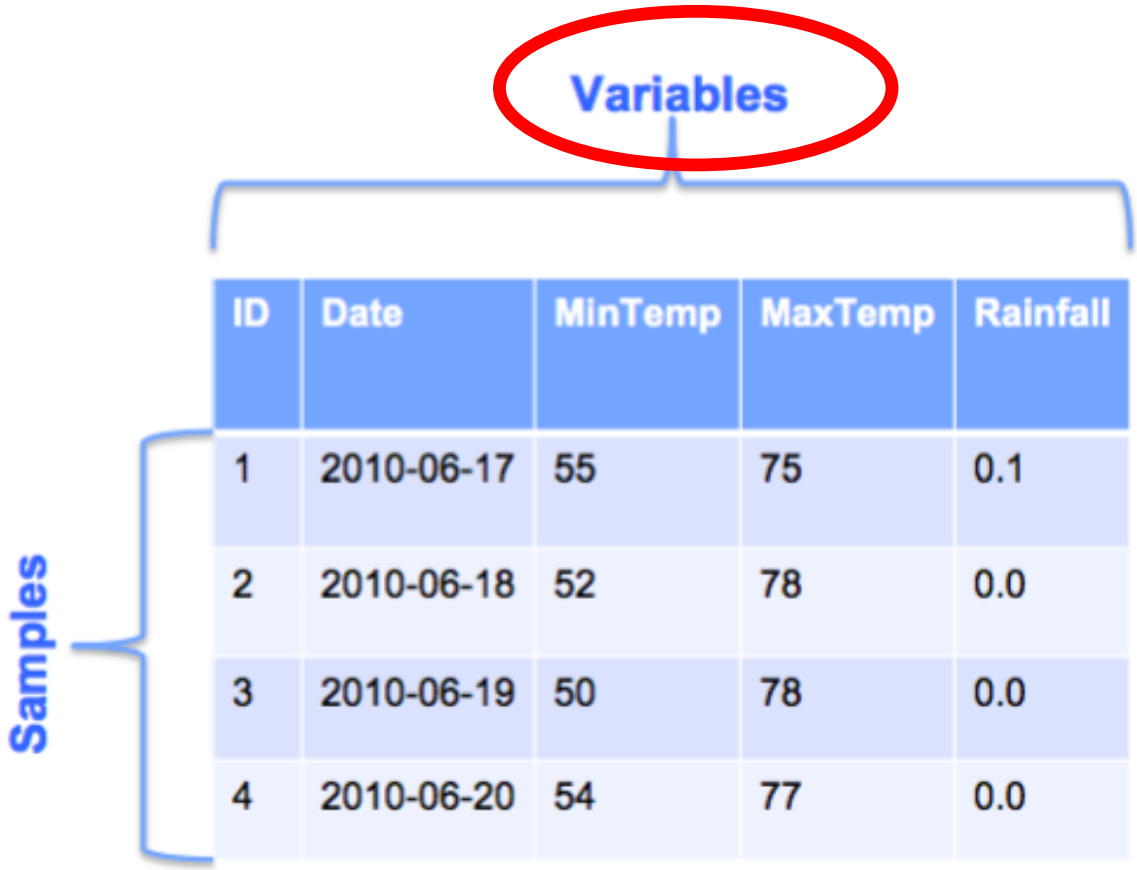
Terms to Describe Data



Terms to Describe Data



Terms to Describe Data



ID	Date	MinTemp	MaxTemp	Rainfall
1	2010-06-17	55	75	0.1
2	2010-06-18	52	78	0.0
3	2010-06-19	50	78	0.0
4	2010-06-20	54	77	0.0

Other Names for 'Sample'

sample

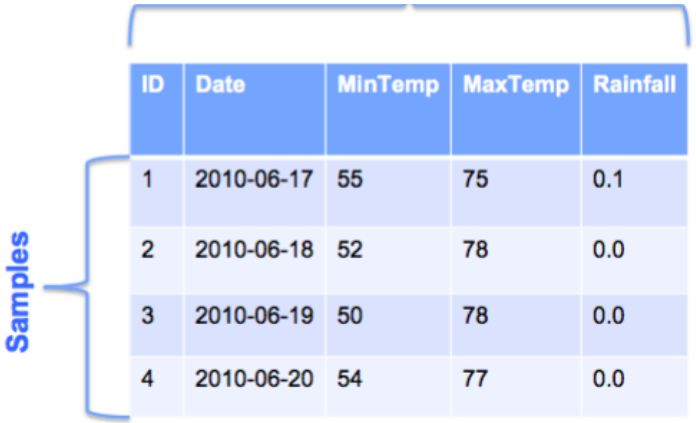
row

instance

observation

record

example



Samples

ID	Date	MinTemp	MaxTemp	Rainfall
1	2010-06-17	55	75	0.1
2	2010-06-18	52	78	0.0
3	2010-06-19	50	78	0.0
4	2010-06-20	54	77	0.0

Other Names for 'Variable'

variable

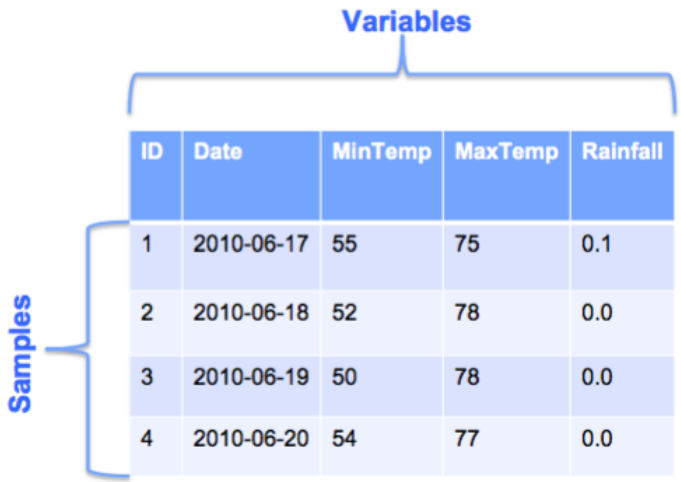
feature

column

dimension

attribute

field



ID	Date	MinTemp	MaxTemp	Rainfall
1	2010-06-17	55	75	0.1
2	2010-06-18	52	78	0.0
3	2010-06-19	50	78	0.0
4	2010-06-20	54	77	0.0

Data Types

- Most common

Numeric **Categorical**

- Others

String **Date** **…**

Numeric Variables

- Values are numbers
- Also called ‘quantitative’

1

$7x10^5$

-0.4902

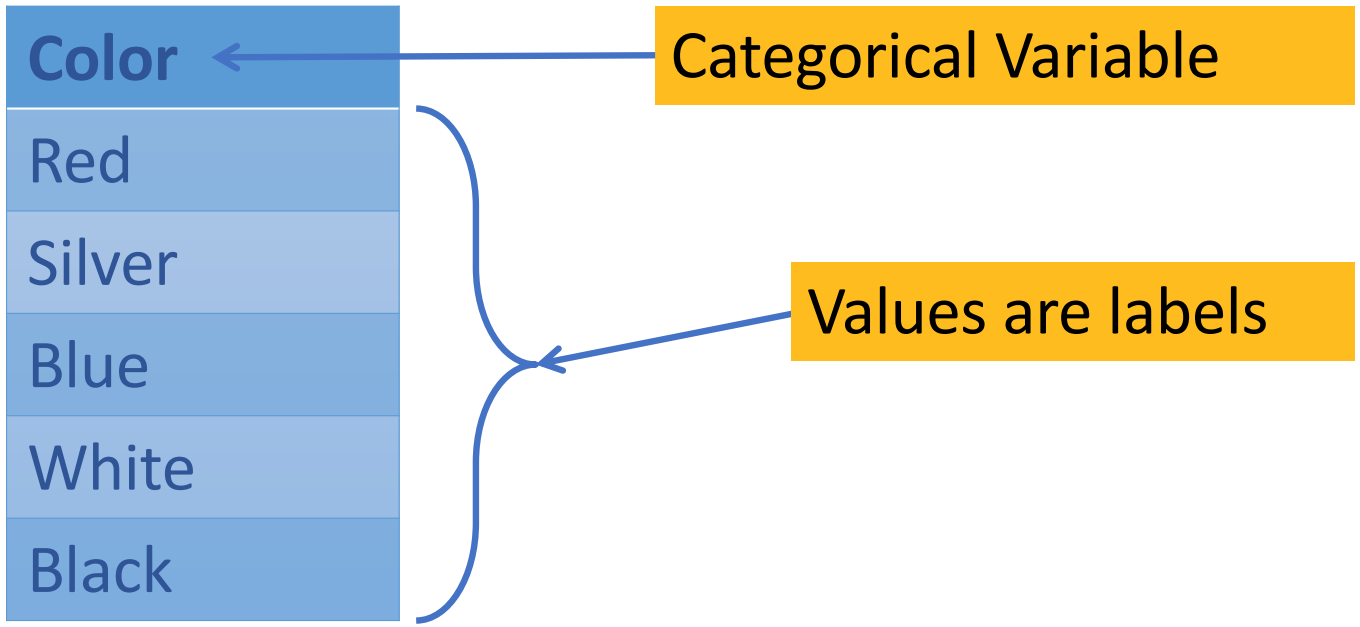
163.92

Examples of Numeric Variables

- Height
- Score on an exam
- Number of transactions per hour
- Change in stock price

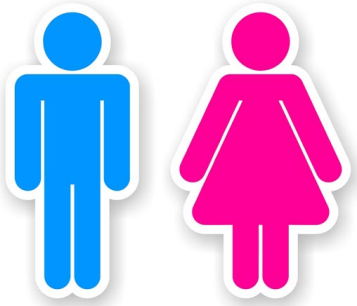
Categorical Variables

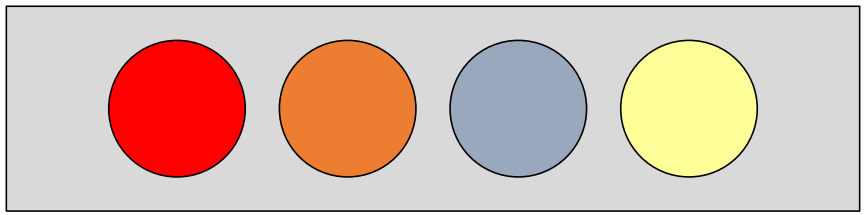
- Values are labels, names, or categories
- Also called ‘qualitative’ or ‘nominal’

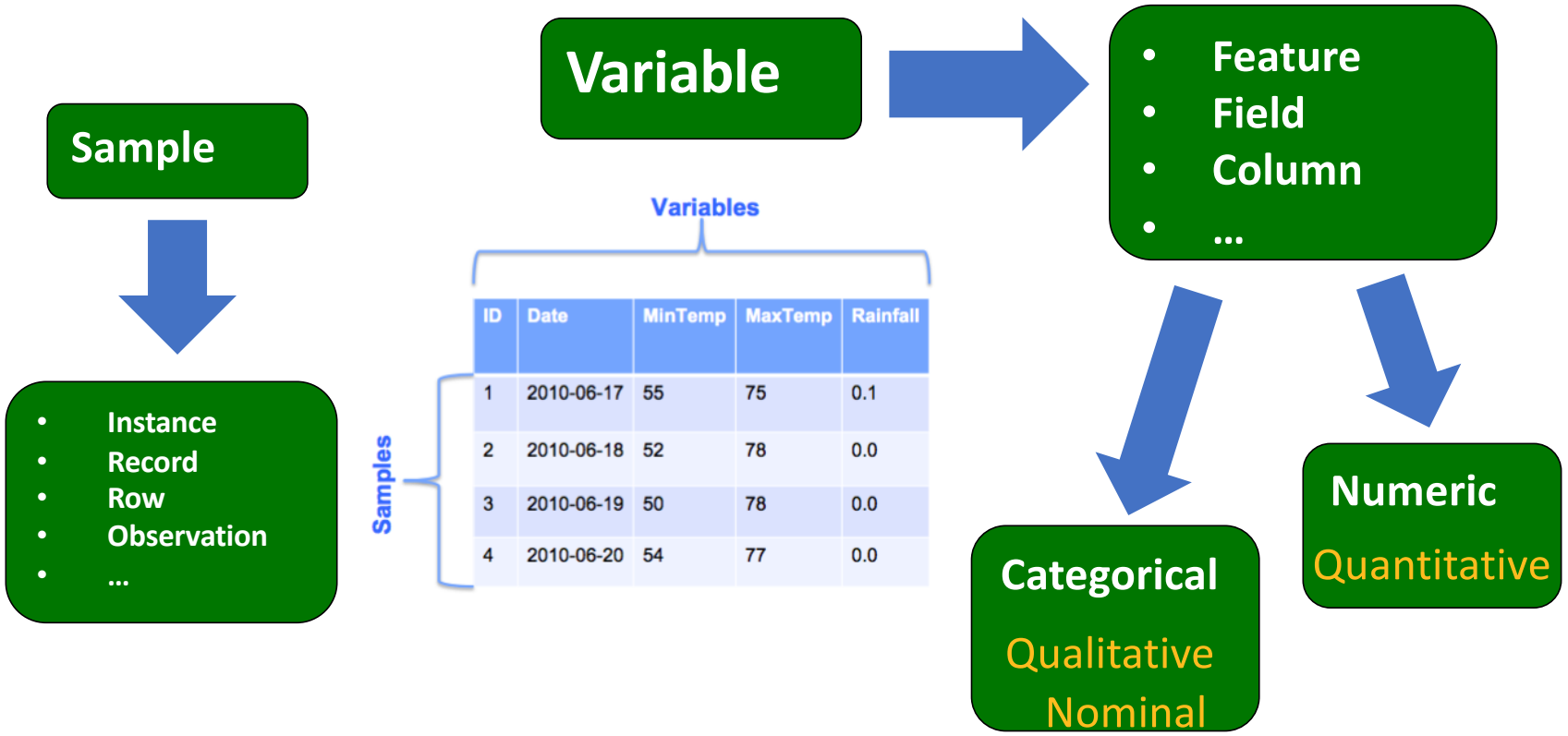


Examples of Categorical Variables

- Gender
- Marital status
- Type of customer
- Product categories
- Color of an item

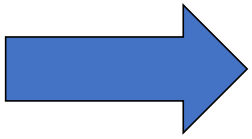




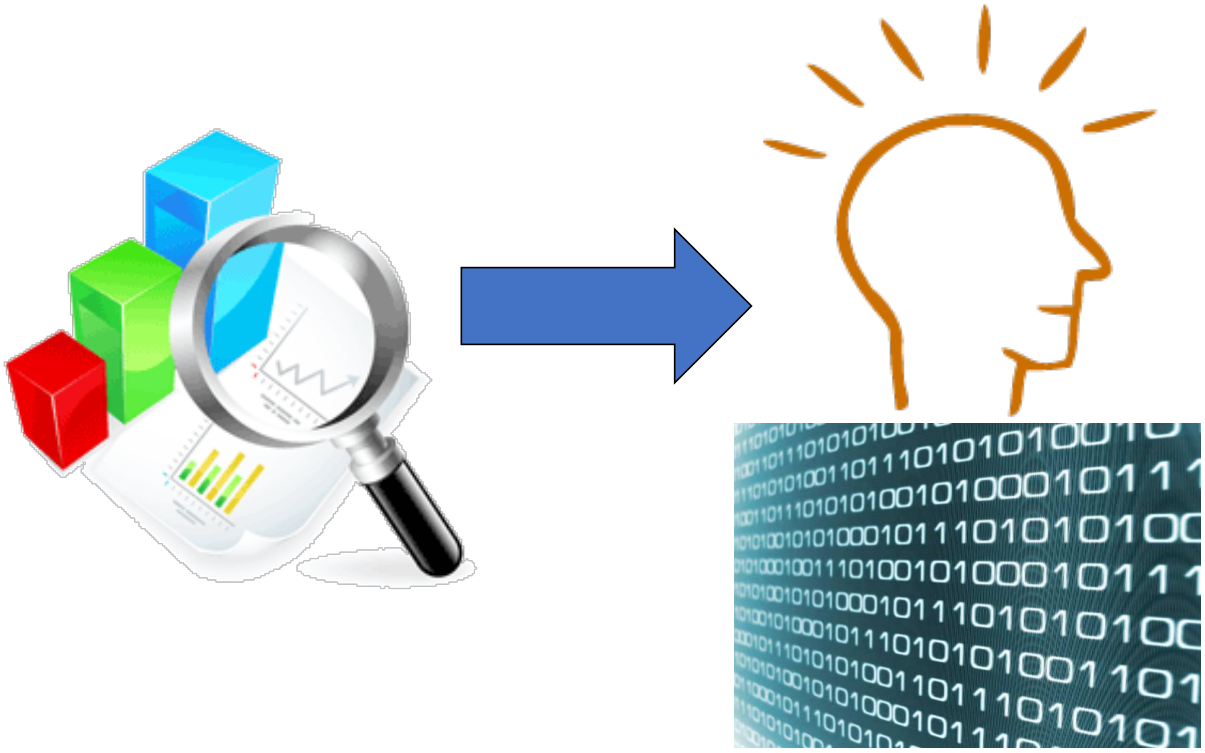

Sample
Instance
Record
Row
Observation
...
Variable
Variables
Samples
ID Date MinTemp MaxTemp Rainfall
1 2010-06-17 55 75 0.1
2 2010-06-18 52 78 0.0
3 2010-06-19 50 78 0.0
4 2010-06-20 54 77 0.0
Feature
Field
Column
...
Categorical
Qualitative
Nominal
Numeric
Quantitative

Why Explore Data?

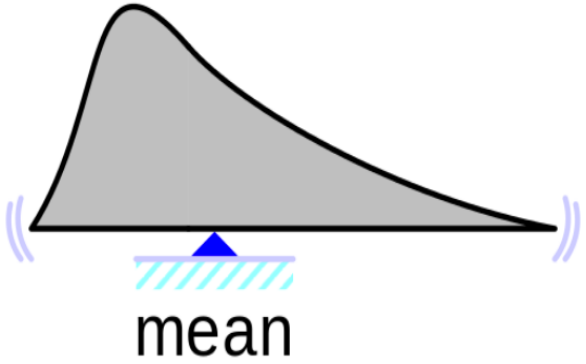
Goal: To understand your data

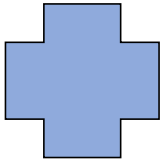


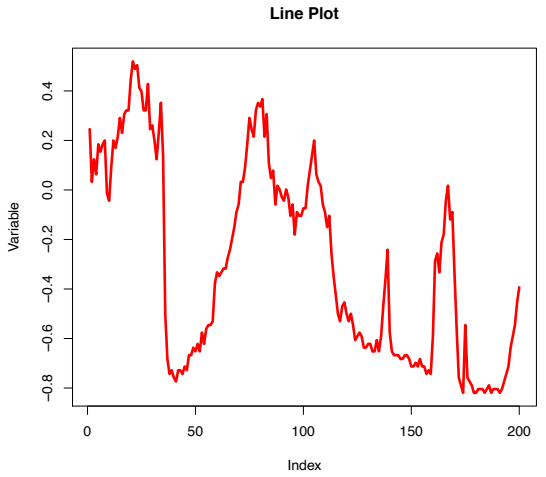
Exploratory Data Analysis (EDA)



Ways to Explore Data





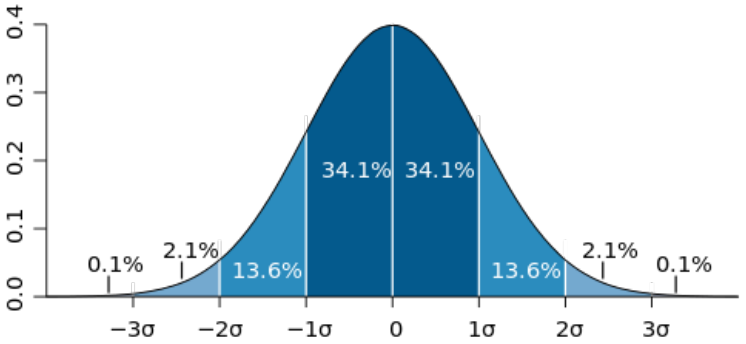


Summary Statistics

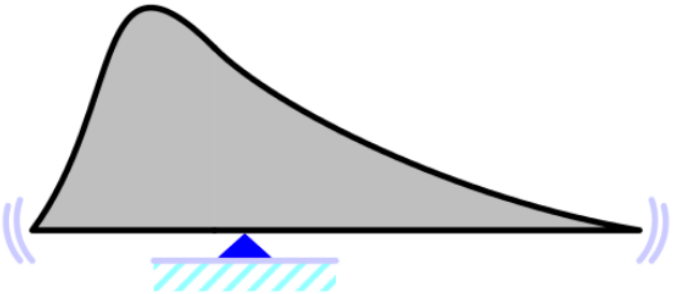
Visualization

Summary Statistics

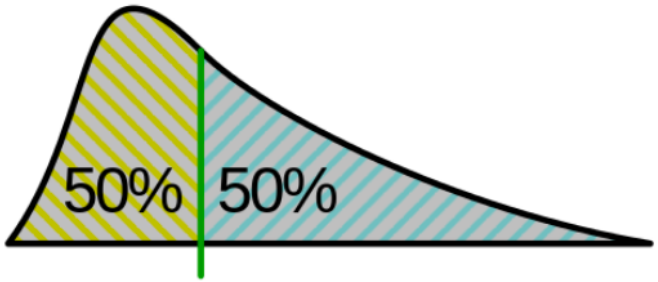
- Information that summarizes dataset



standard deviation



mean

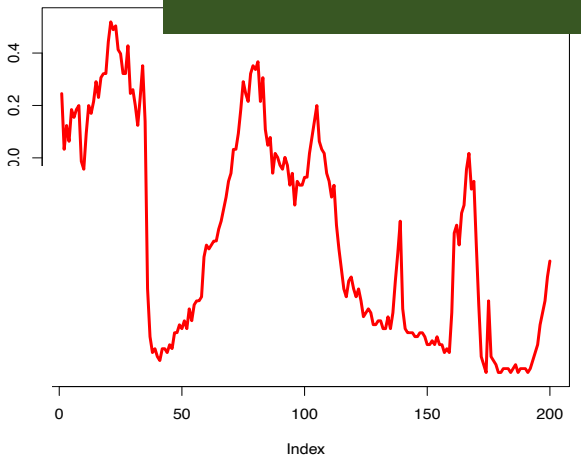


median

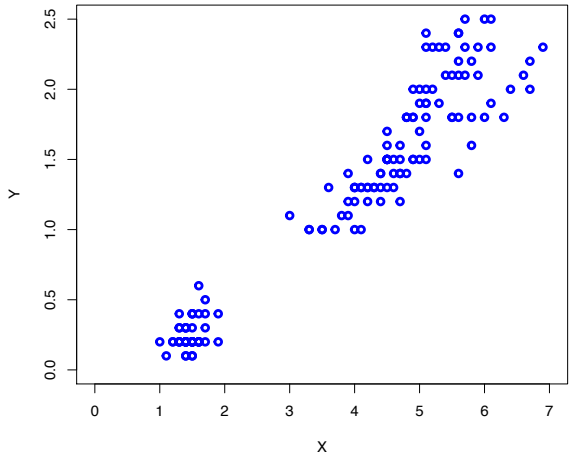
Data Visualization

- Look at data graphically

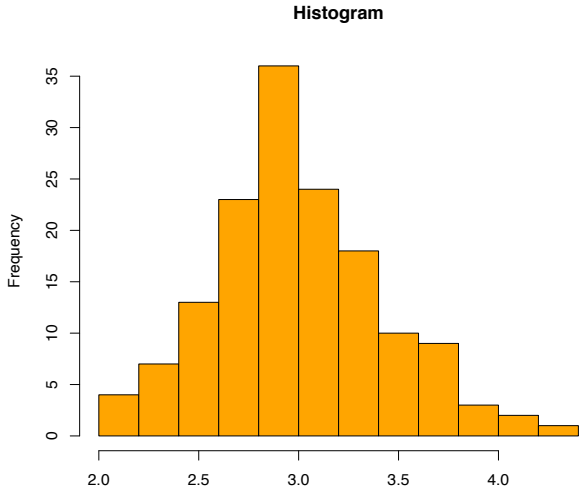
Line Plot



Scatter Plot

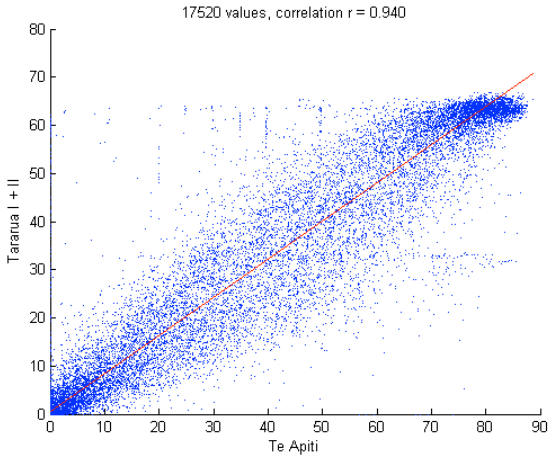


Histogram

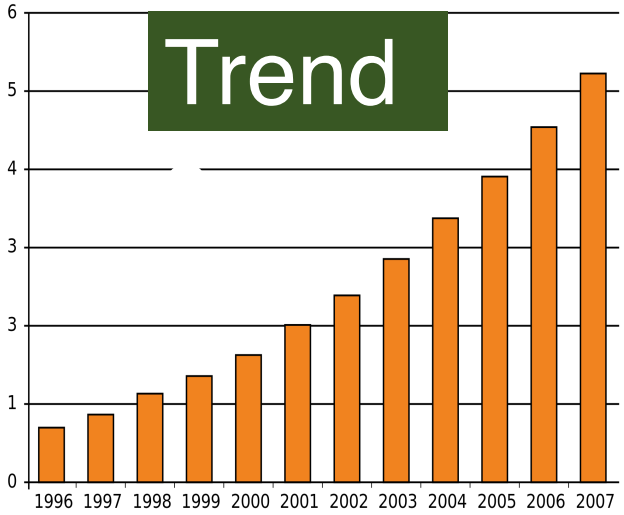


Some Things to Look For

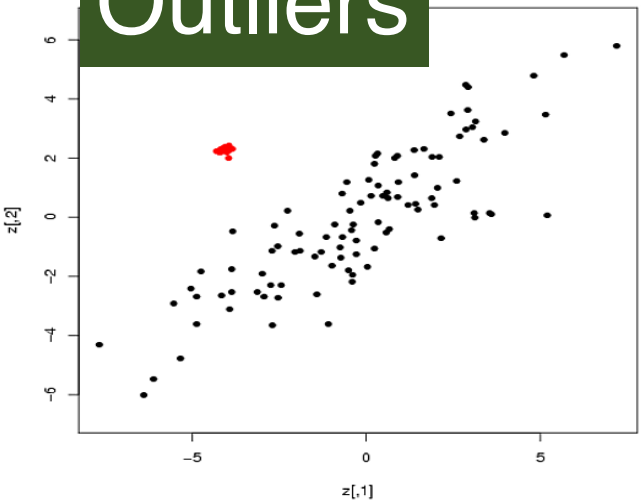
Correlations



Trend

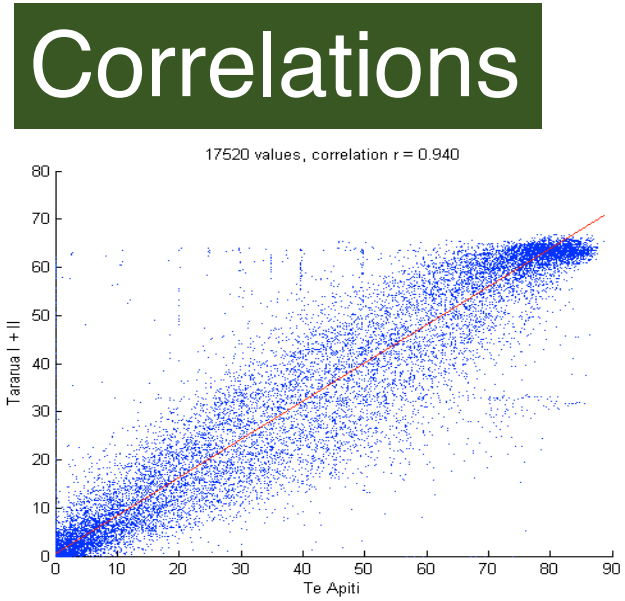


Outliers



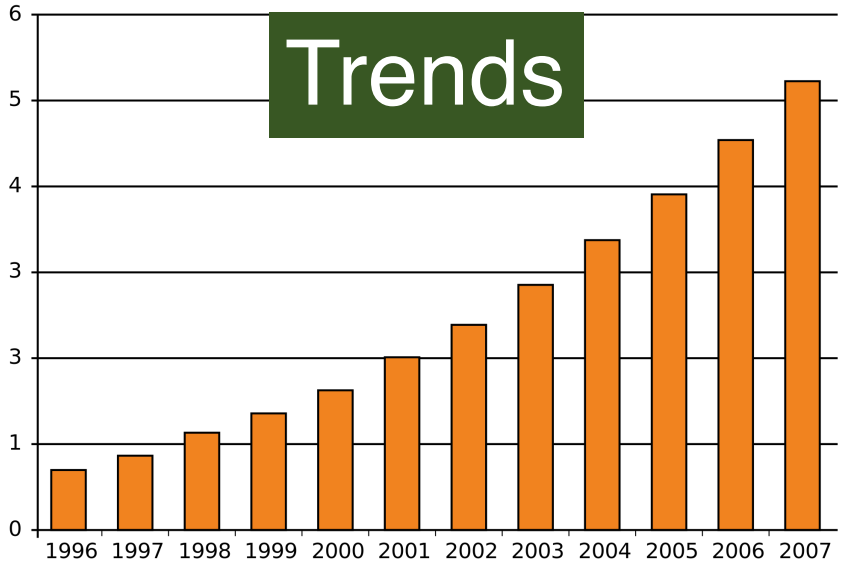
Correlations

- Provide information about relationship between variables



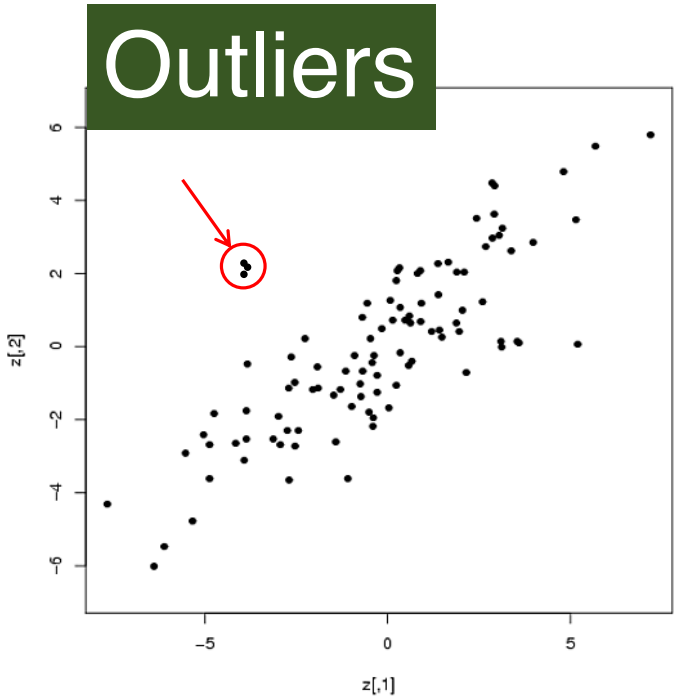
Trends

- Indicate general characteristics of data

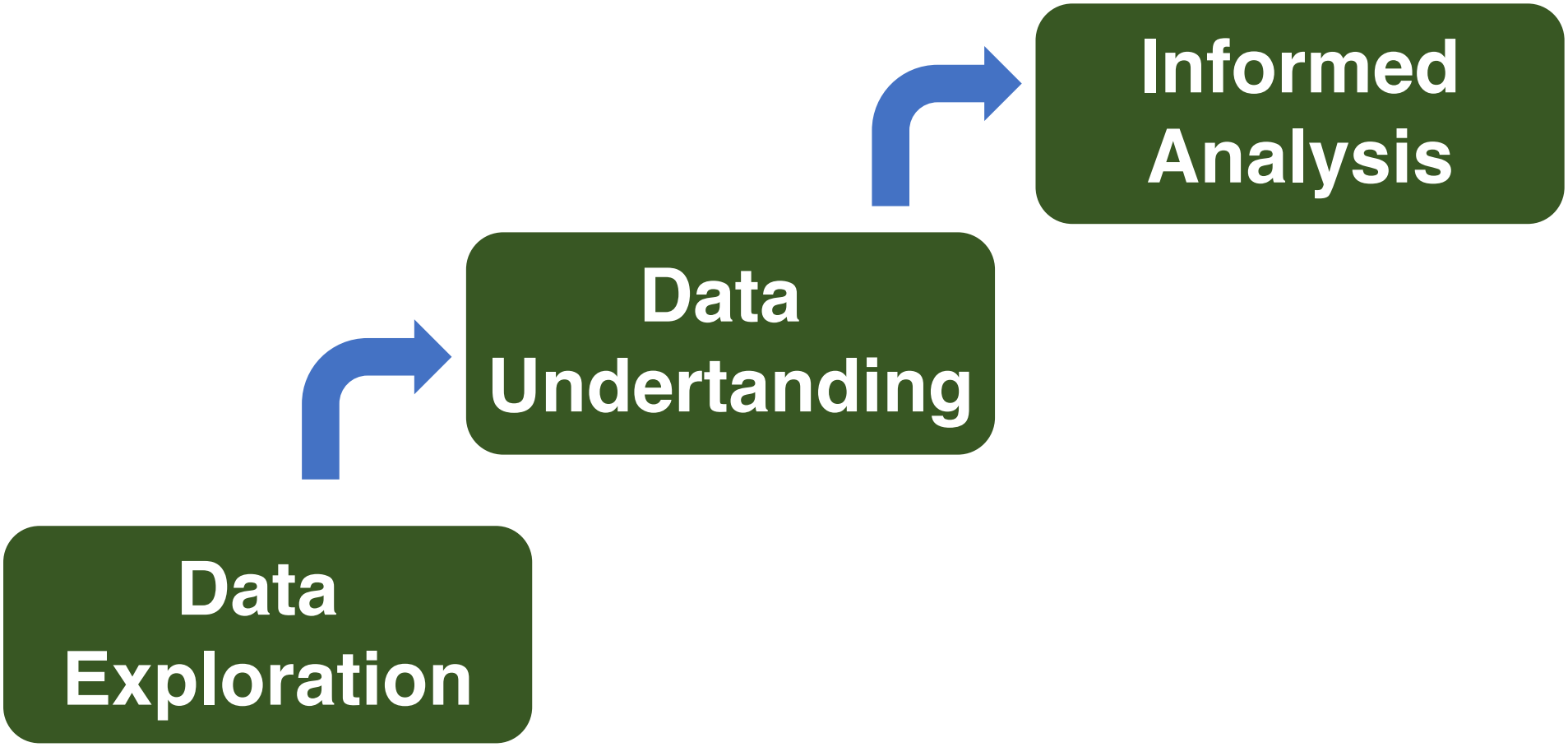


Outliers

- Indicate potential problems with data



Data Exploration



Exploring Data through Summary Statistics

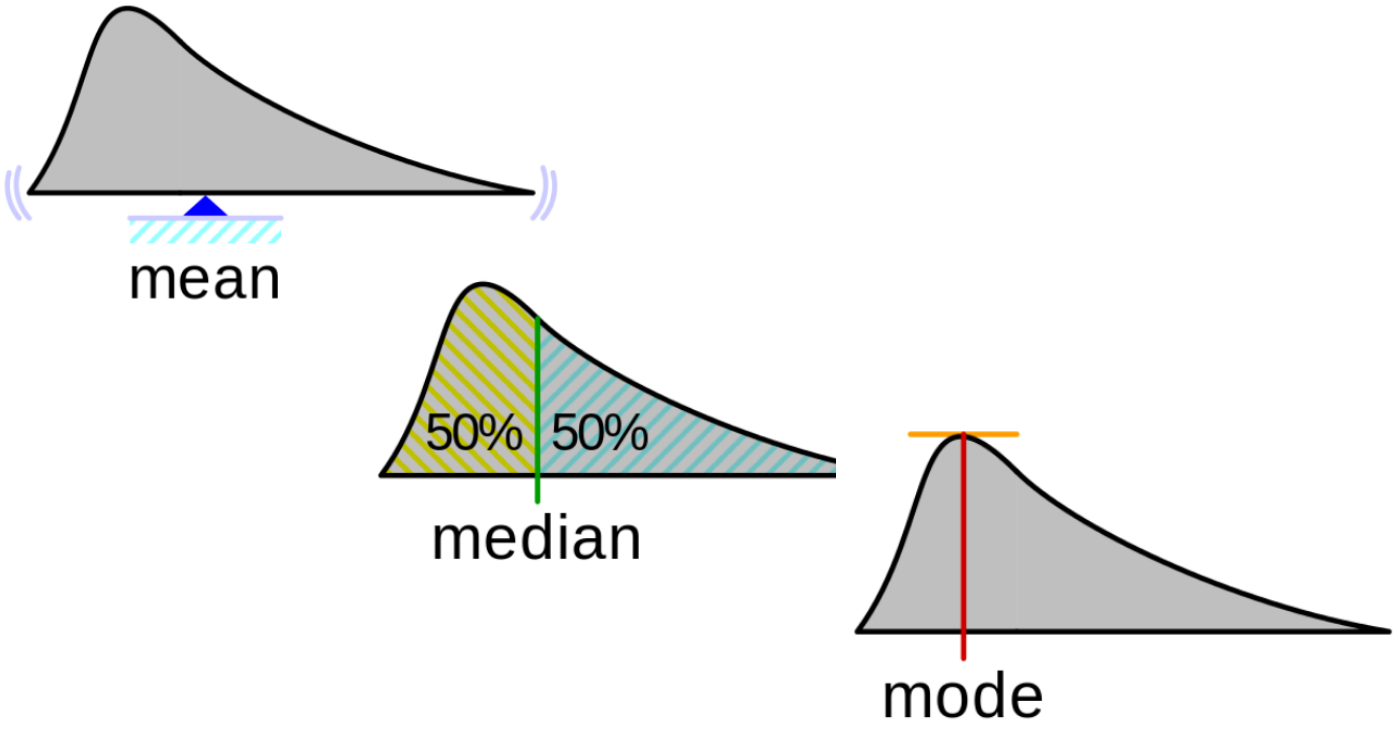
- Define what a summary statistic is
- List common summary statistics
- Explain how summary statistics are useful in exploring data

What are summary statistics?

- Quantities that summarize and describe a set of data values
- Measures of
  - Location: mean, median
  - Spread: standard deviation
  - Shape: skewness

Measures of Location

Describe central or typical value of dataset



Measures of Location - Example

Age
35
42
78
22
56
50
42
78
21
87

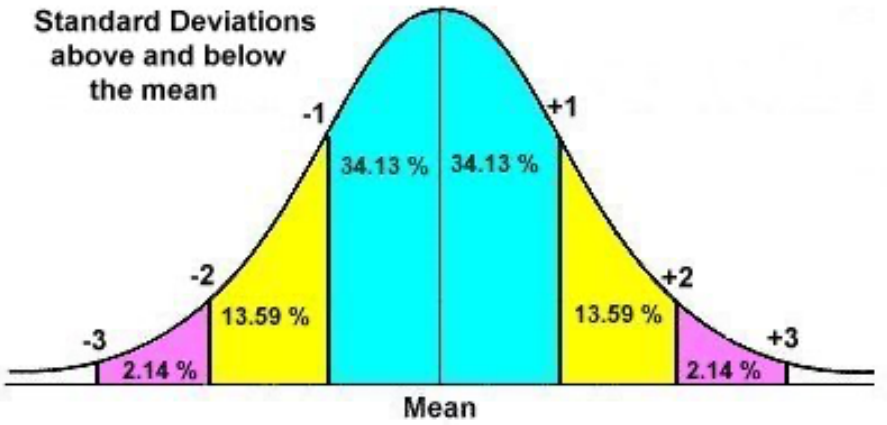
Age (sorted)
21
22
35
42
42
50
56
78
78
87

Mean = 51.1

Median = (42+50)/2 = 46

Mode = 42 & 78

Measures of Spread



Describe how dispersed or varied data is

minimum

maximum

standard deviation

variation

range

Measures of Spread – Example

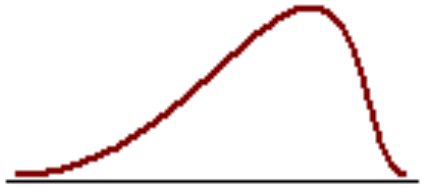
Age	Age (sorted)
35	21
42	22
78	35
22	42
56	42
50	50
42	56
78	78
21	78
87	87

Range = 87 – 21 = 66

Variance = 548.767

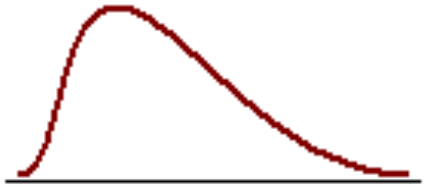
Standard deviation = 23.426

Measures of Shape



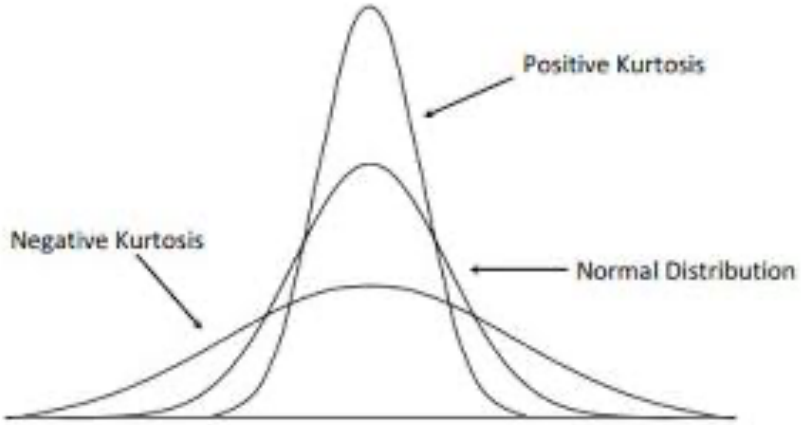
Negatively skewed distribution
or Skewed to the left
Skewness <0

Normal distribution
Symmetrical
Skewness = 0



Positively skewed distribution
or Skewed to the right
Skewness > 0

skewness



kurtosis

Measures of Shape – Example

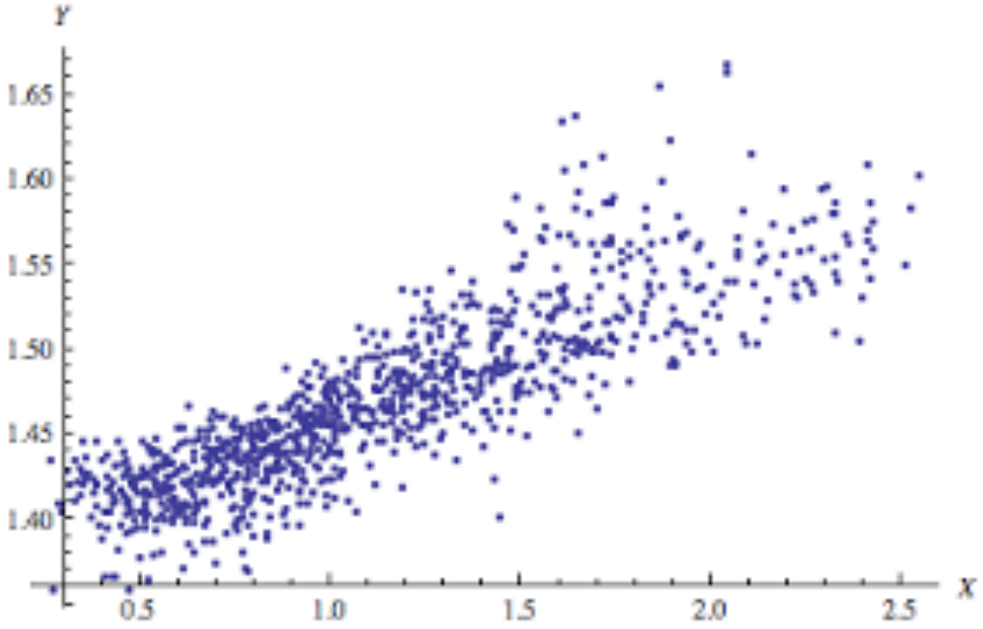
Age
35
42
78
22
56
50
42
78
21
87

Skewness = 0.2995

Kurtosis = -1.2028

Measures of Dependence

Describe relationship between variables



correlation

Measures of Dependence – Example

Height	Weight
180	68
153	70
204	84
133	44
208	81
142	53
122	40
168	50
175	64
200	72

Correlation = 0.8906

Statistics on Categorical Variables

Describe number of categories and frequency of each category

Color/Pet	White	Brown	Black	Orange	Total
Dog	34	44	32	0	110
Cat	25	2	43	0	70
Fish	1	0	5	33	39
Total	60	46	80	33	219

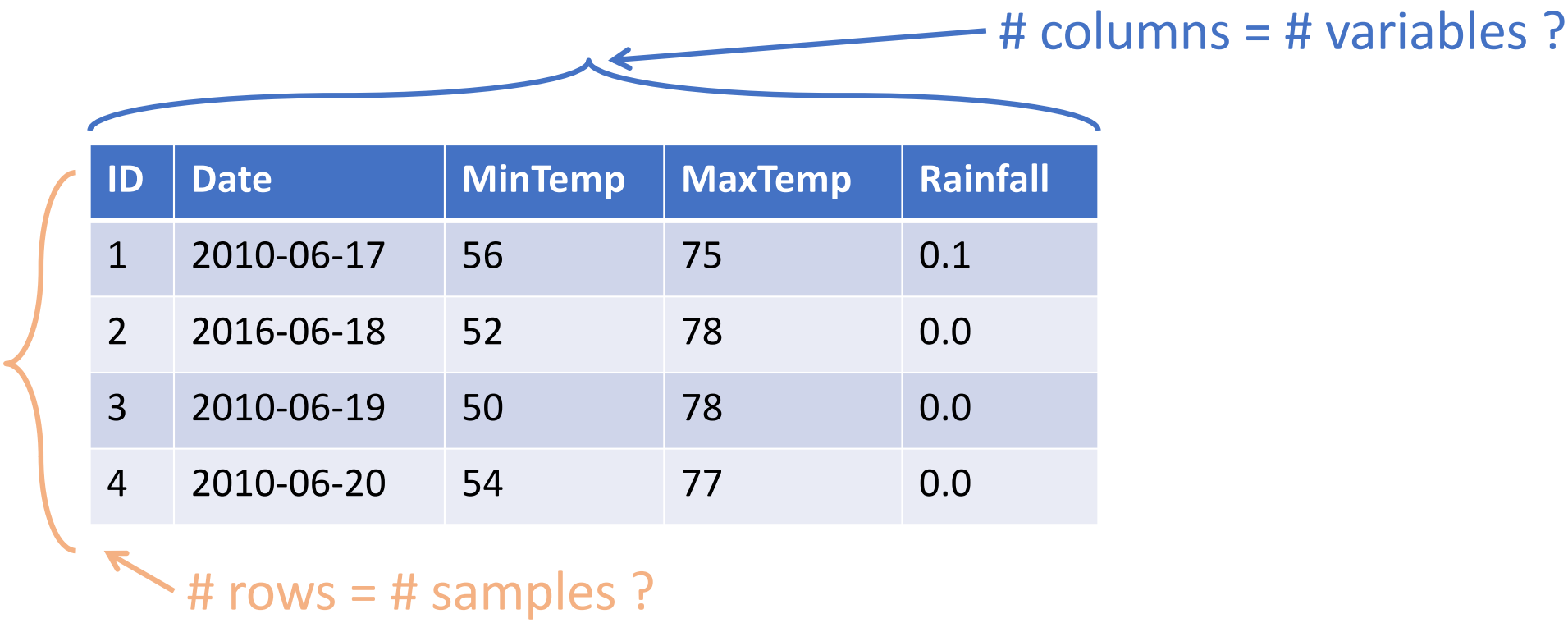
contingency table

Contingency Table - Example

Color/Pet	White	Brown	Black	Orange	Total
Dog	34	44	32	0	110
Cat	25	2	43	0	70
Fish	1	0	5	33	39
Total	60	46	80	33	219

Check Dimensions

- Check number of rows and columns



ID	Date	MinTemp	MaxTemp	Rainfall
1	2010-06-17	56	75	0.1
2	2016-06-18	52	78	0.0
3	2010-06-19	50	78	0.0
4	2010-06-20	54	77	0.0

Check Values

- Check values in some samples

Should temperature values in F or C?

Is this correct?

ID	Date	MinTemp	MaxTemp	Rainfall
1	2010-06-17	56	24	0.1
2	2016-06-18	52	26	3,678.9
3	2010-06-19	50	26	0.0
4	2010-06-20	54	25	0.0

Is this date or timestamp?

Check Missing Values

ID	Date	MinTemp	MaxTemp	Rainfall
1	2010-06-17	56	75	--
2	2016-06-18	52	78	--
3	2010-06-19	--	78	0.1
4	2010-06-20	54	77	--

Does feature have mostly missing values?

How many samples have missing values?

Summary Statistics

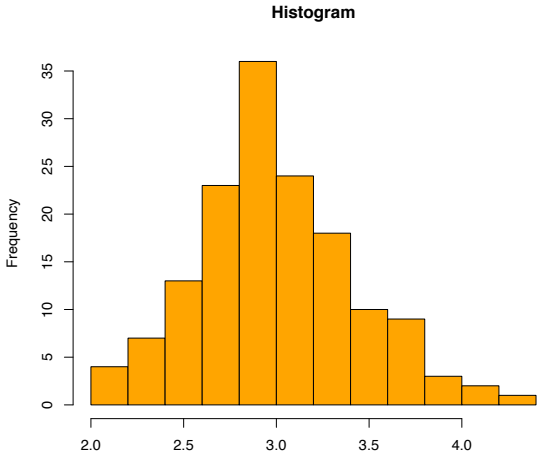
- Measures of
  - Location, spread, shape, dependence
- Contingency table
  - For categorical variables
- Data validation
  - Dimensions, missing values

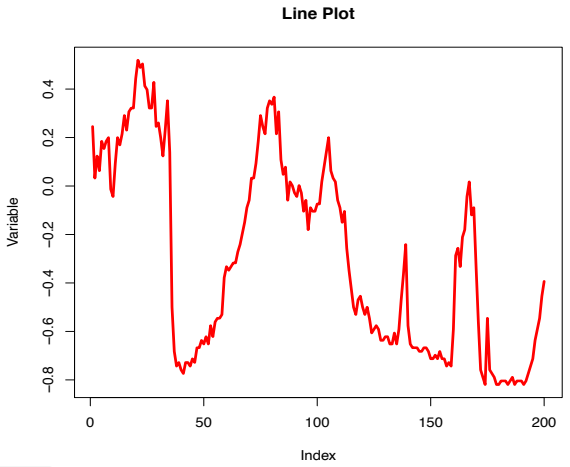
Exploring Data through Plots

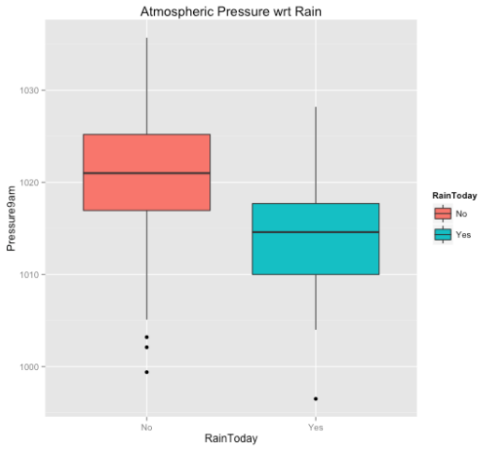
- Discuss how plots can be useful in exploring data
- Describe how you would use a scatter plot
- Summarize what a boxplot shows

Python for Data Science

Visualizing Data





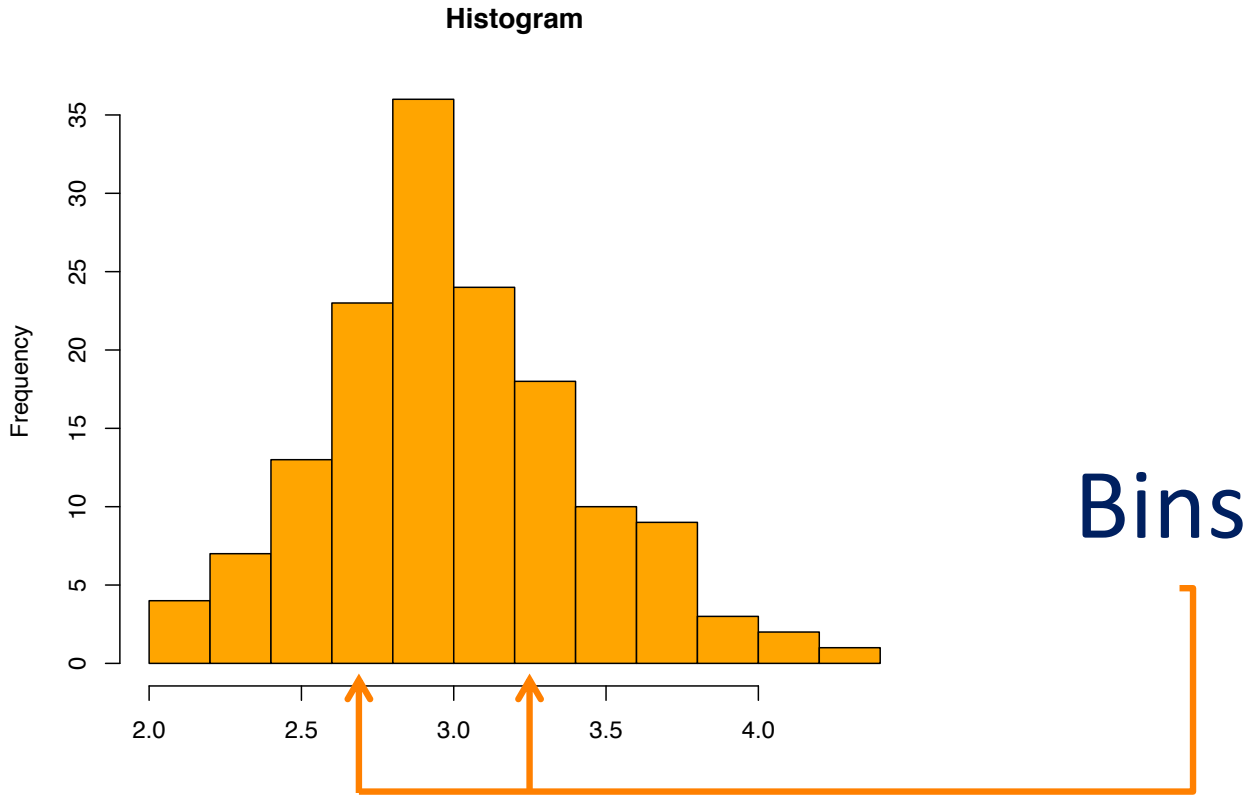


Types of Plots

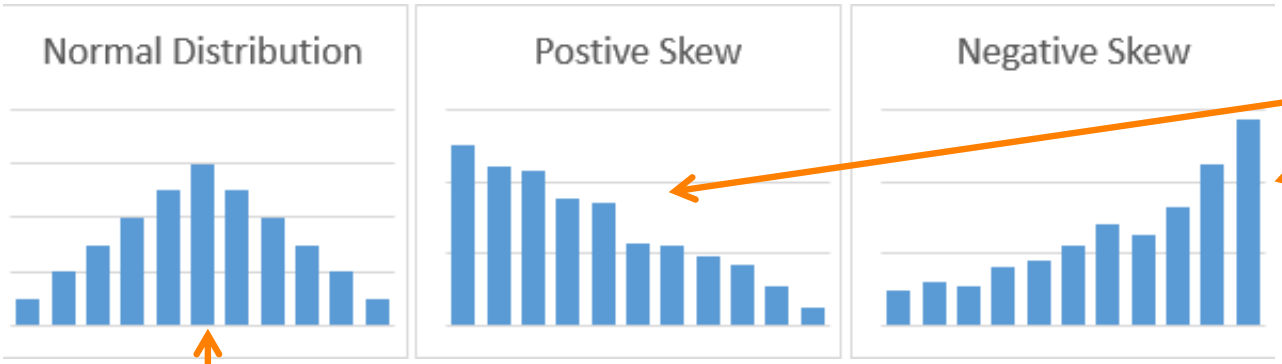
- Histogram
- Line plot
- Scatter plot
- Bar plot
- Box plot
- others

Histogram

- Shows distribution of numeric variable



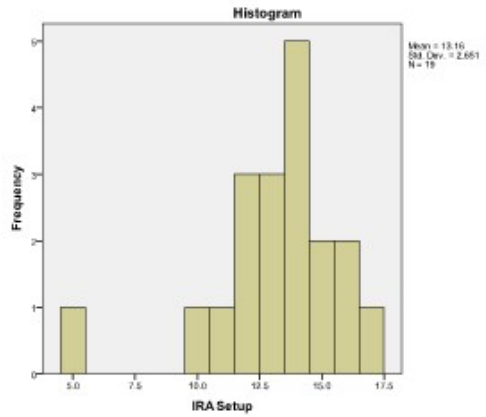
What a Histogram Shows



Skewness

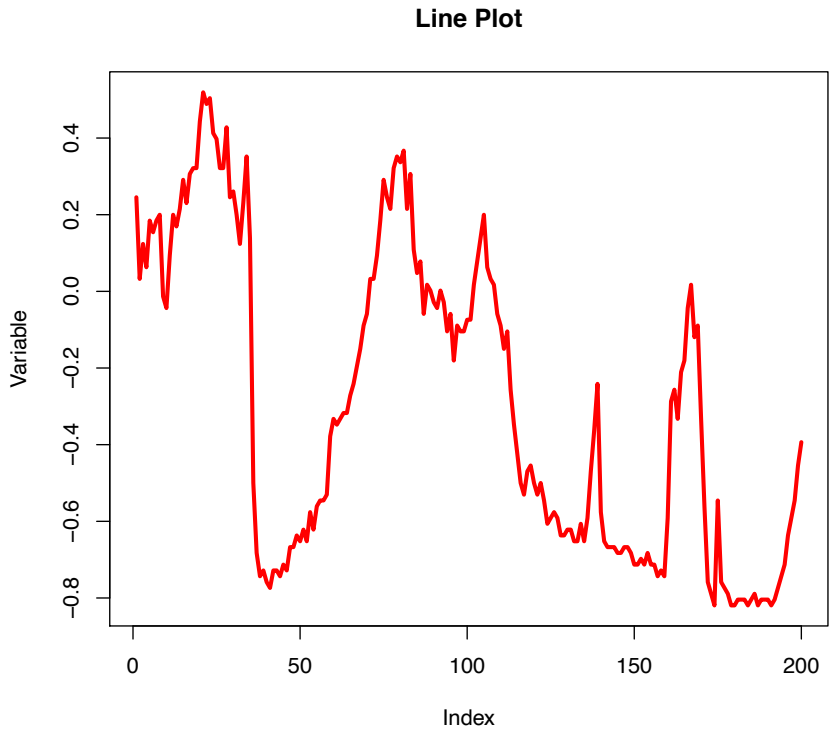
Central Tendency

Outlier



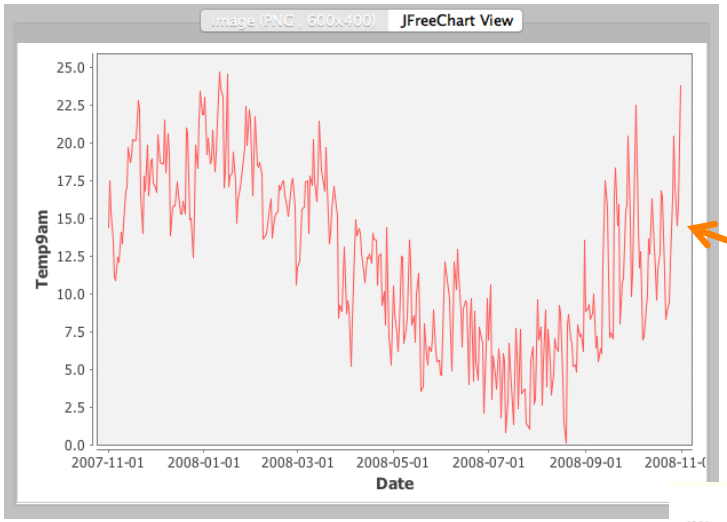
Line Plot

- Shows change in data over time



Python for Data Science

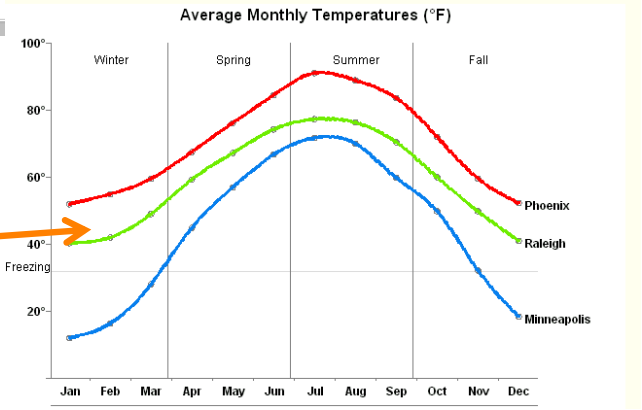
What a Line Plot Shows



Trend

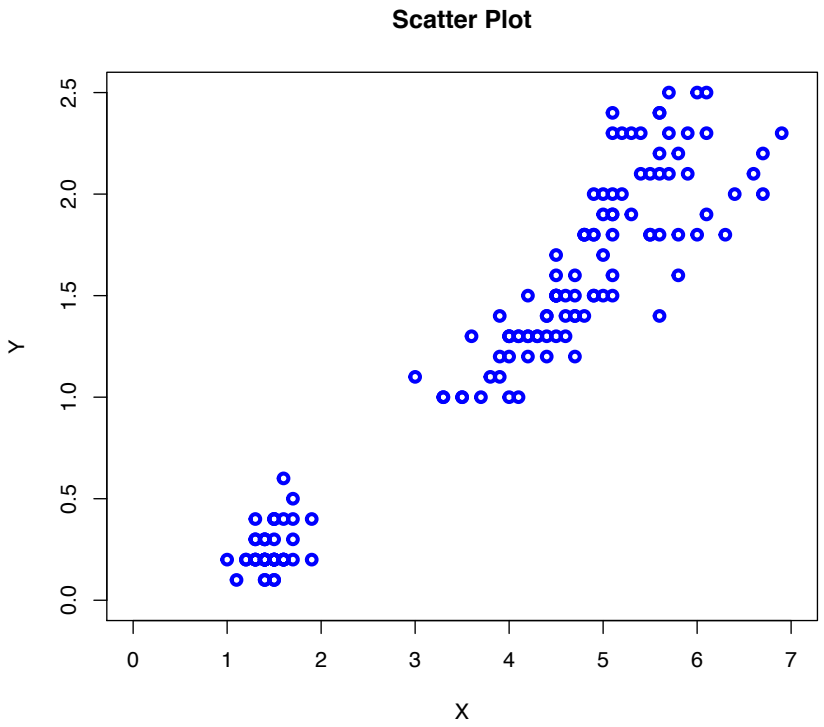
Cyclical pattern

Compare variables



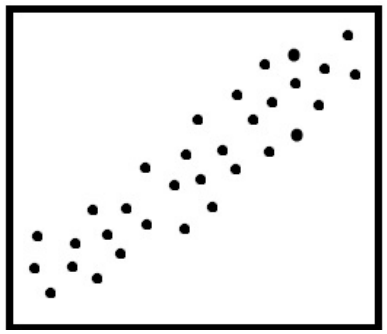
Scatter Plot

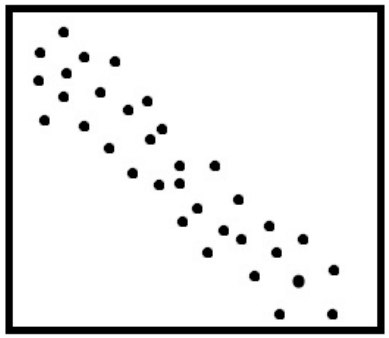
- Shows relationship between two variables



What a Scatter Plot Shows

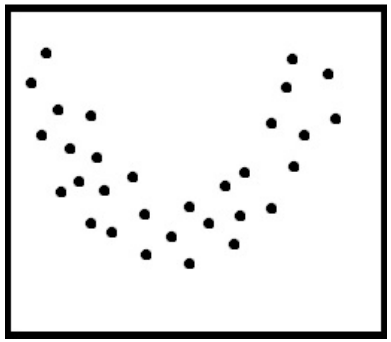
Positive Correlation

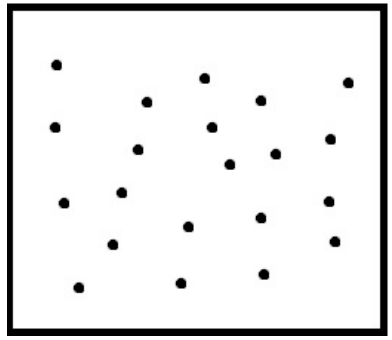




Negative Correlation

Non-Linear Correlation

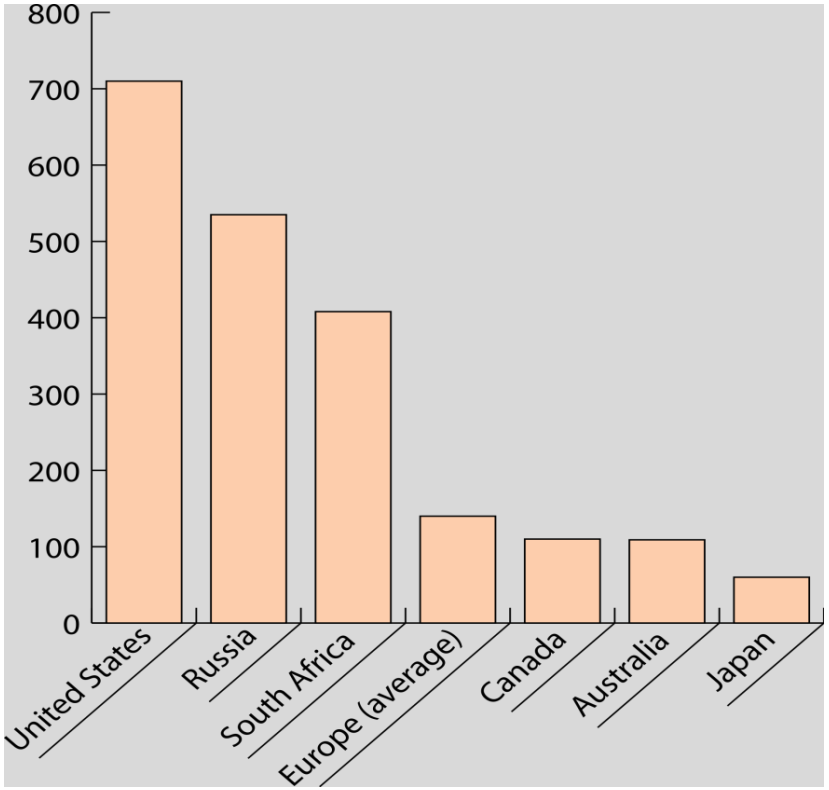




No Correlation

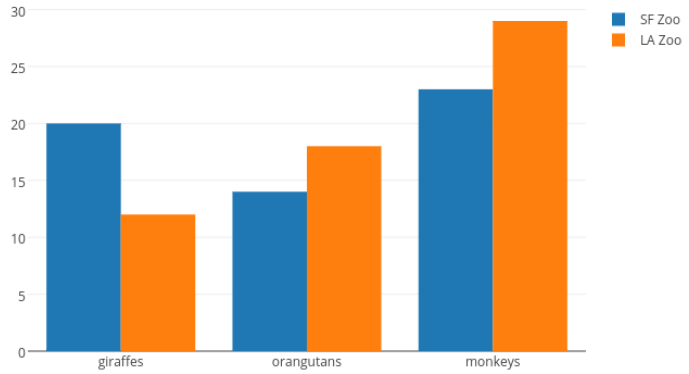
Bar Plot

- Shows distribution of categorical variable

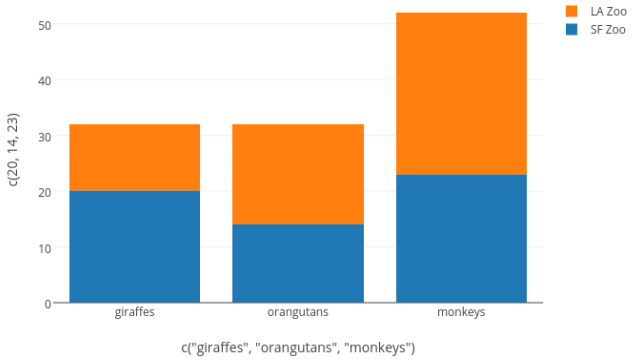


What a Bar Plot Shows

Grouped Bar Chart

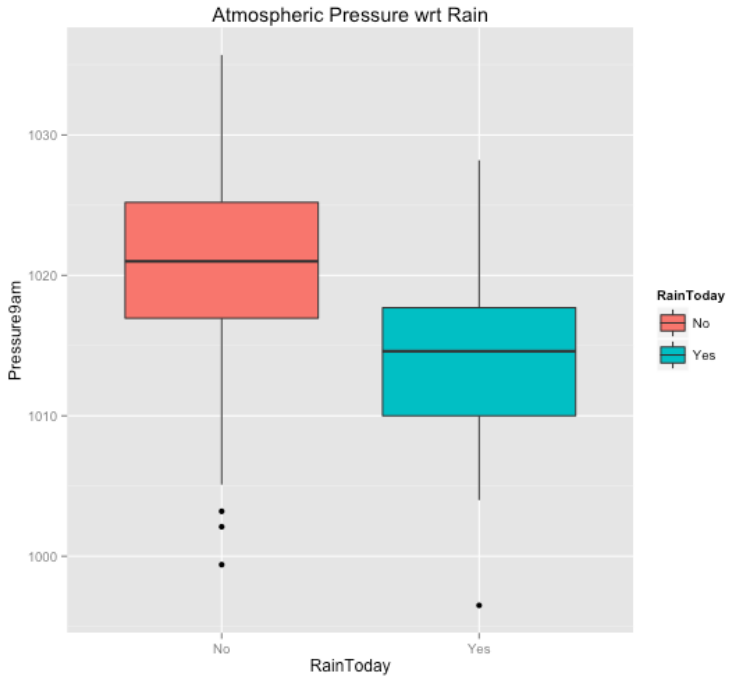


Stacked Bar Chart

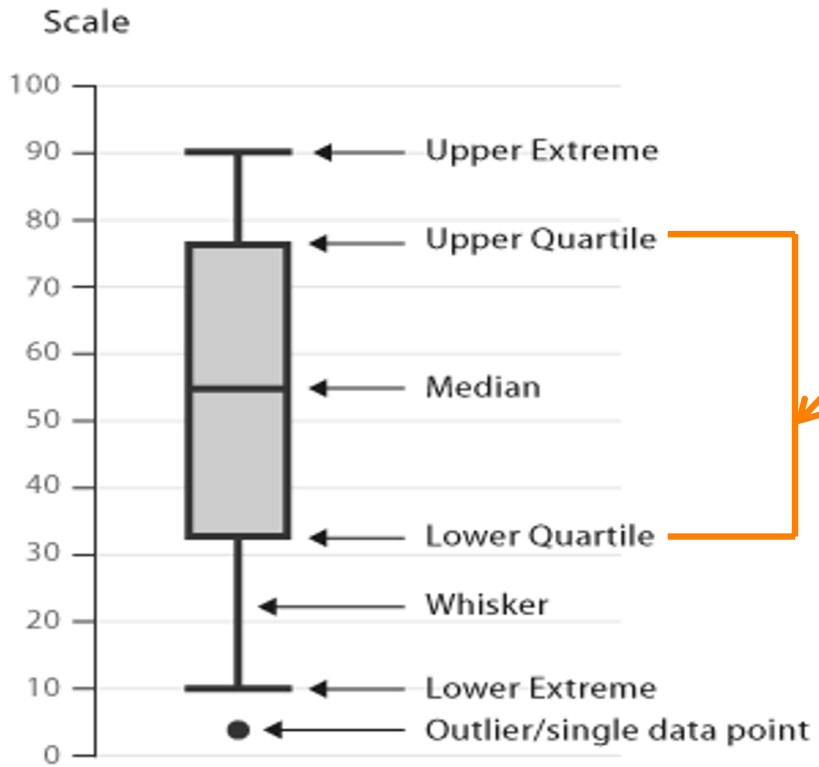


Box Plot

- Compares distributions of variables

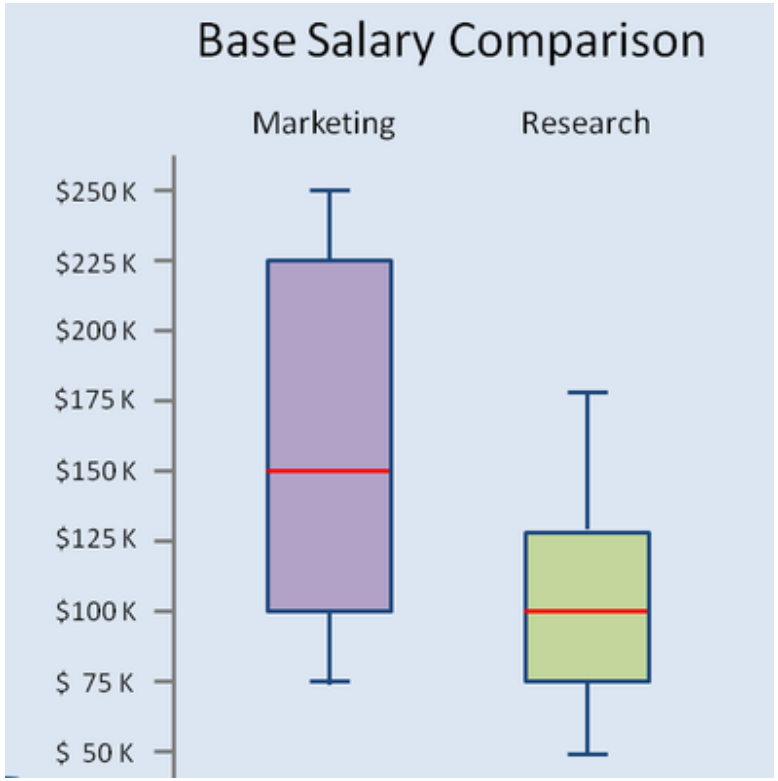


Components of a Box Plot

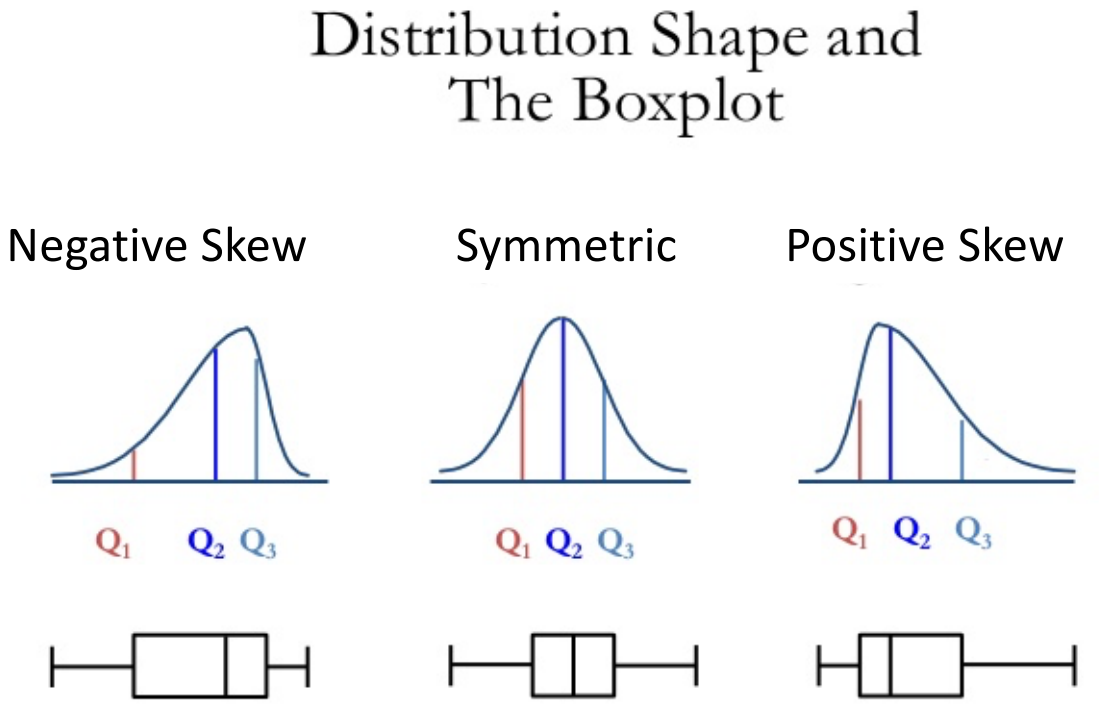


The middle 50% of data are in this region

What a Box Plot Shows



What a Box Plot Shows



Data Visualization

- Provides intuitive way to look at data
- Should be used with summary statistics for data exploration
- Are also useful for communicating results