Data Terminology

After this video you will be able to..

- Describe what a feature is and how it relates to a sample
- Name some alternative terms for ‘feature’
- Summarize how a categorical feature differs from a numerical feature

Terms to Describe Data

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

Terms to Describe Data

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

Terms to Describe Data

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

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

dimension

column

attribute

field

Variables

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

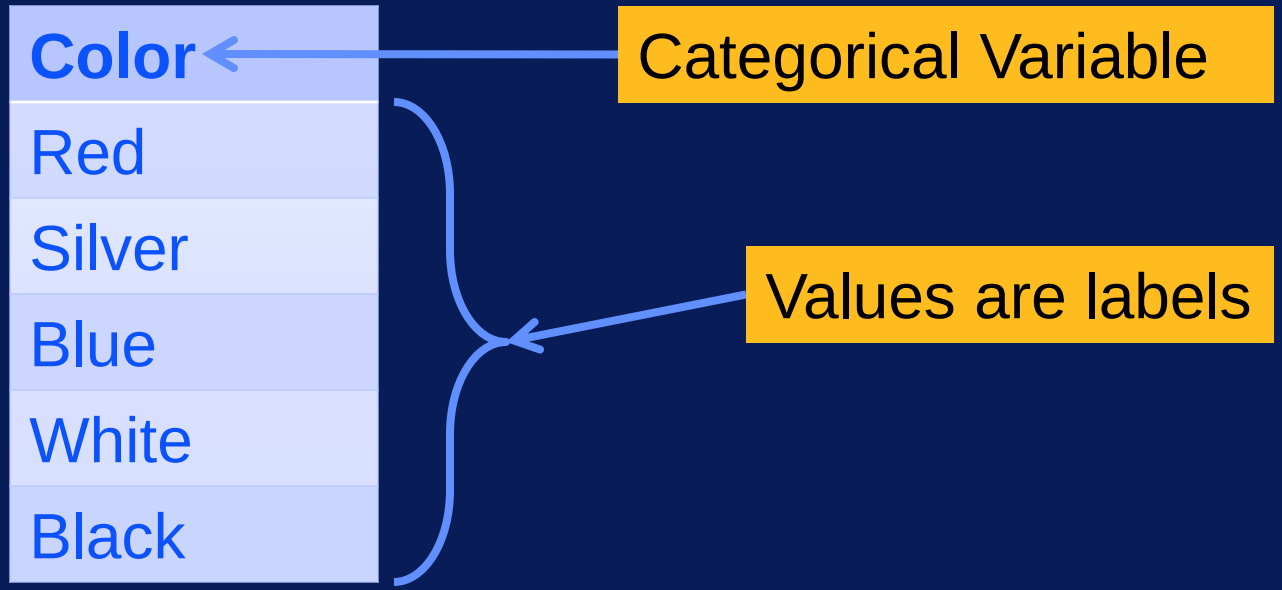
163.92

Examples of Numeric Variables

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

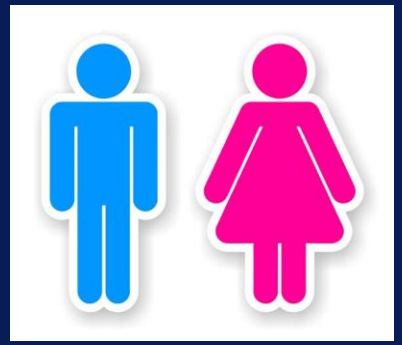
Categorical Variables

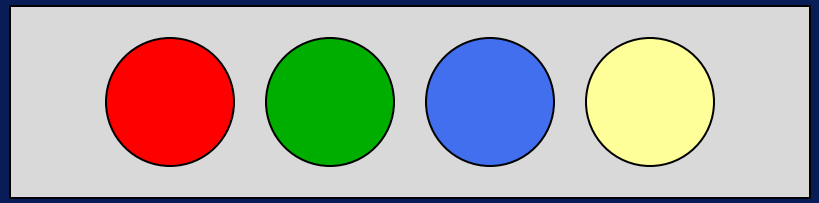
- Values are labels, names, or categories
- Also called ‘qualitative’ or ‘nominal’



Examples of Categorical Variables

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





Sample

- Instance
- Record
- Row
- Observation
- …

Variable

- Feature
- Field
- Column
- …

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

Categorical

Qualitative
Nominal

Numeric

Quantitative

Data Exploration

After this video you will be able to..

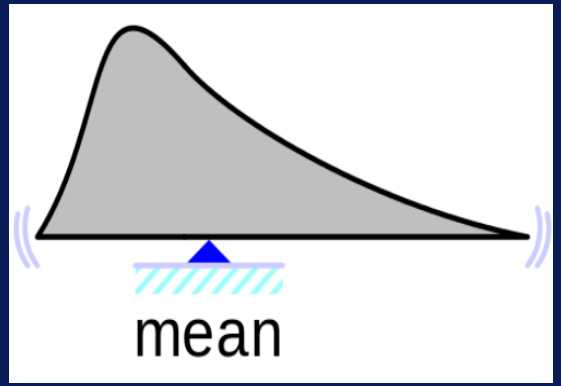
- Explain why data exploration is necessary
- Articulate the objectives for data exploration
- List the categories of techniques for exploring data

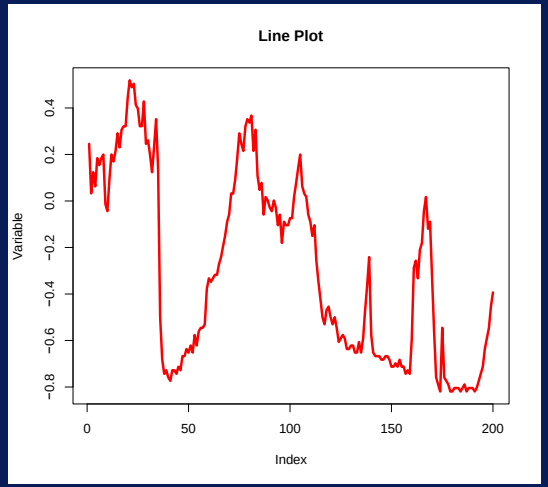
Why Explore Data?

Goal: To understand your data

Exploratory Data Analysis (EDA)

Ways to Explore Data



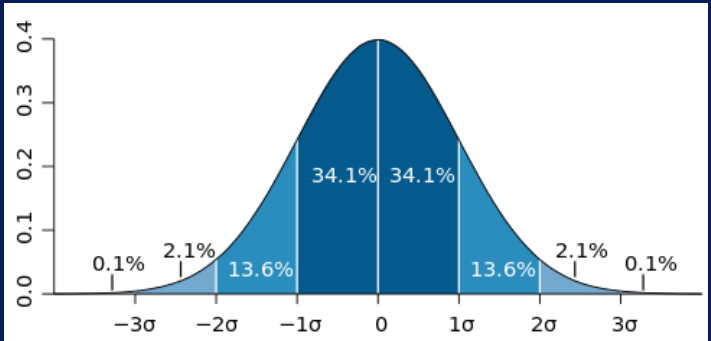


Summary Statistics

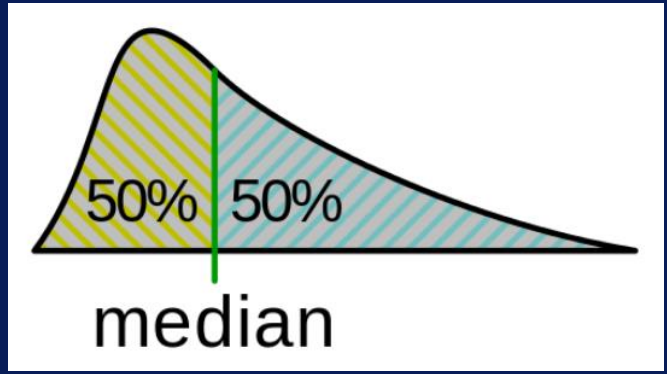
Visualization

Summary Statistics

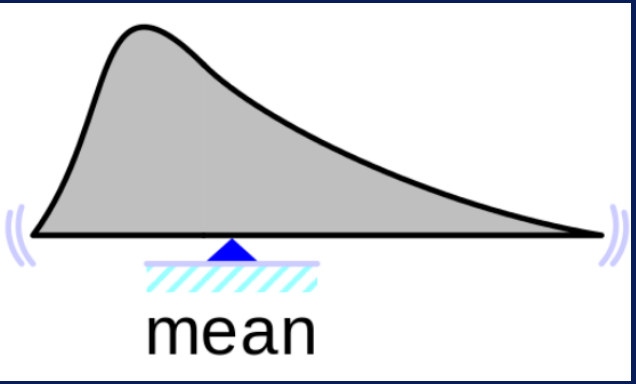
- **Information that summarizes dataset**



standard deviation



median

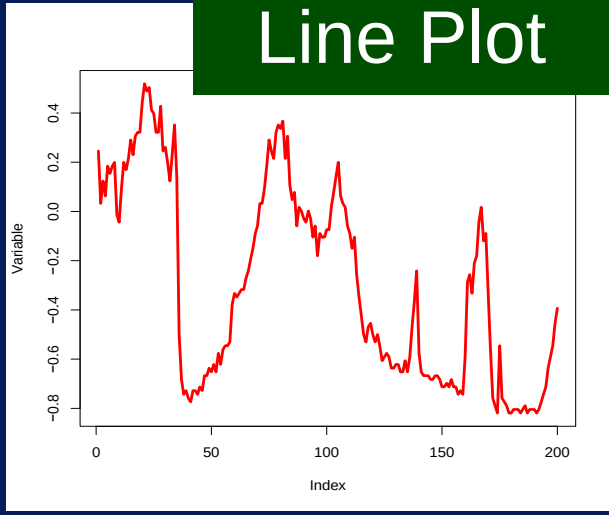


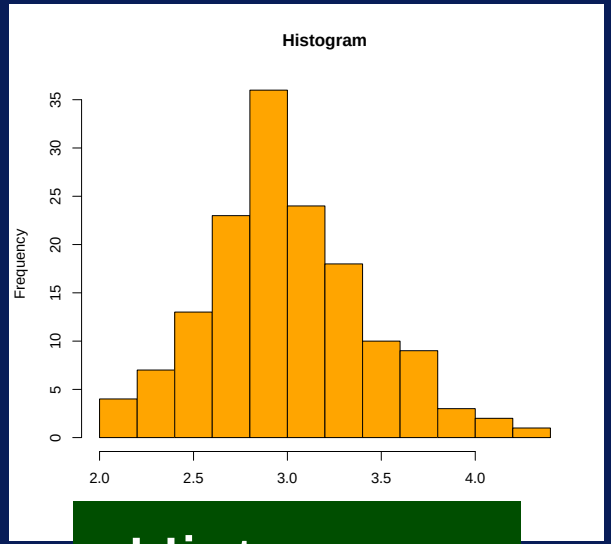
mean

Data Visualization

- **Look at data graphically**

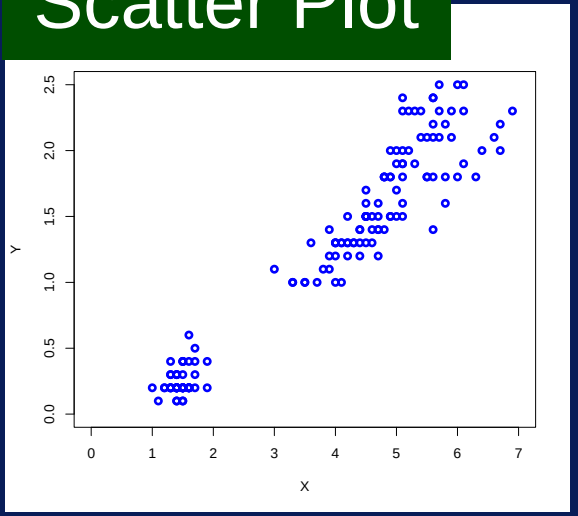
Line Plot





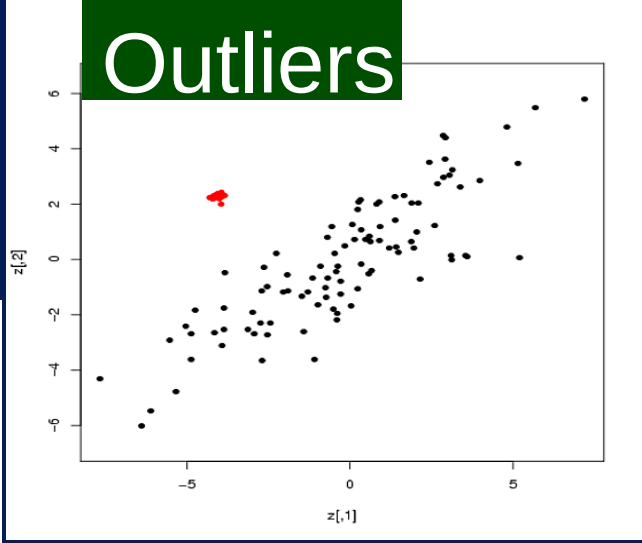
Histogram

Scatter Plot

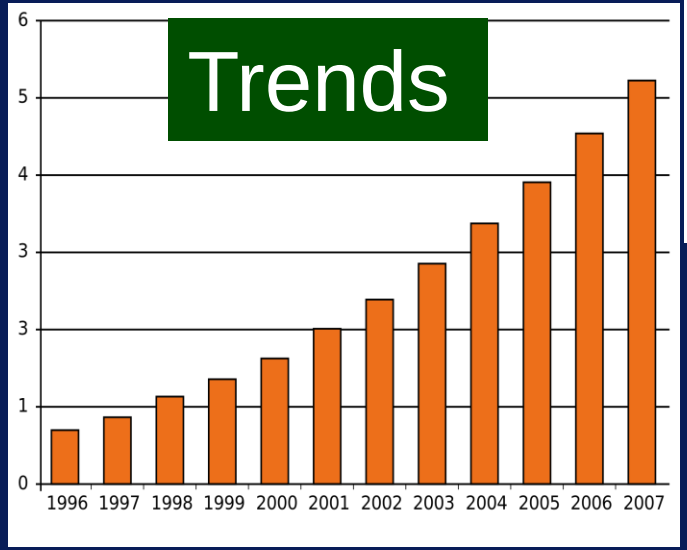


Some Things to Look For

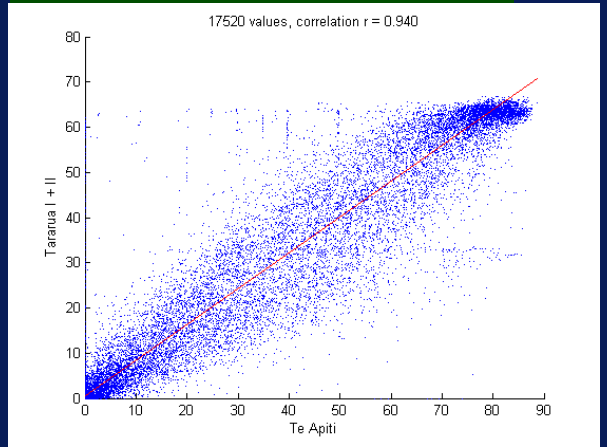
Outliers



Trends



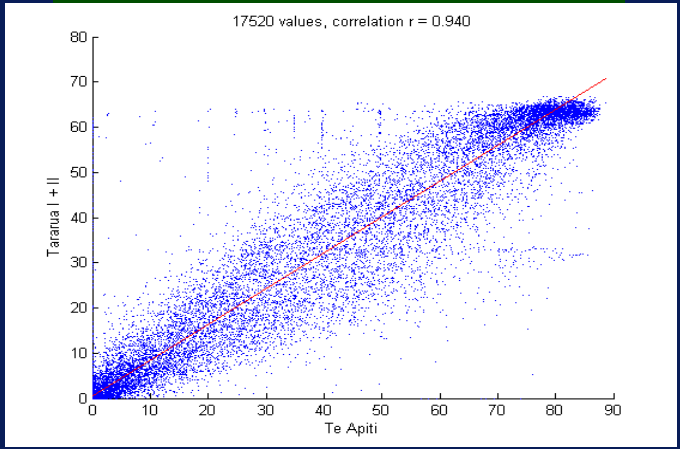
Correlations



Correlations

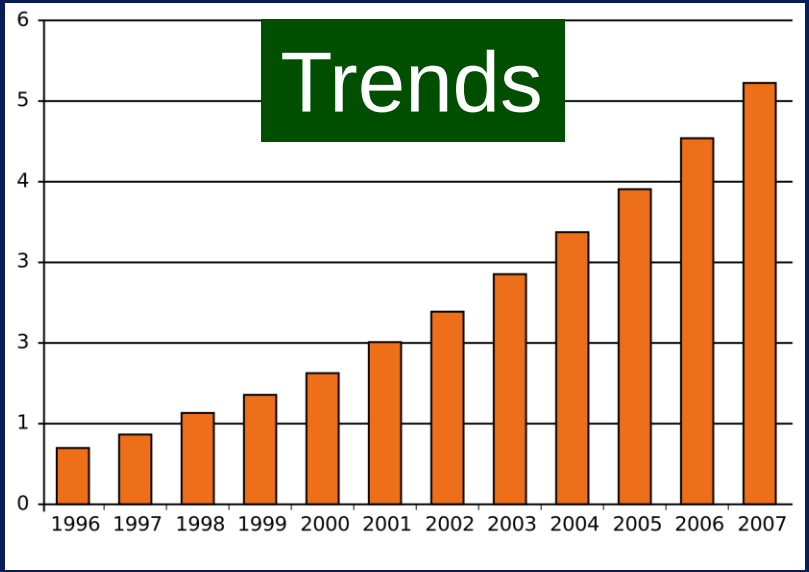
- Provide information about relationship between variables

Correlations



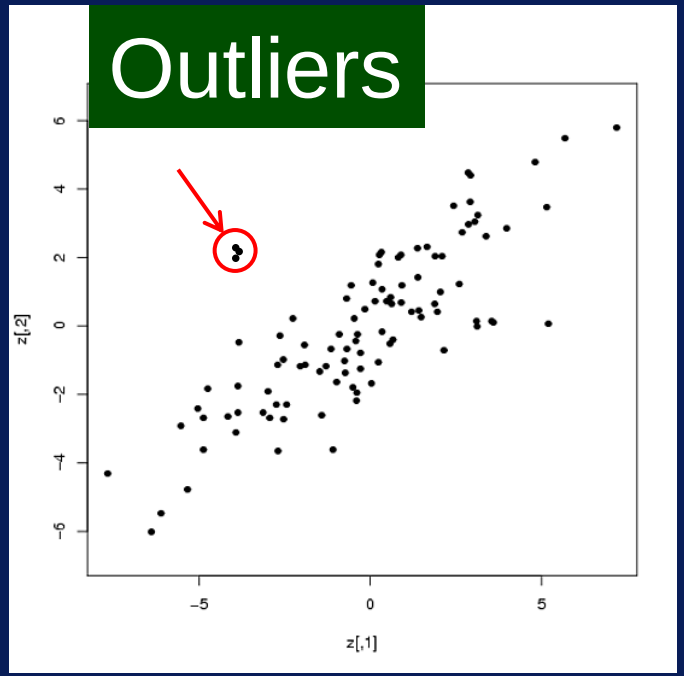
Trends

- Indicate general characteristics of data

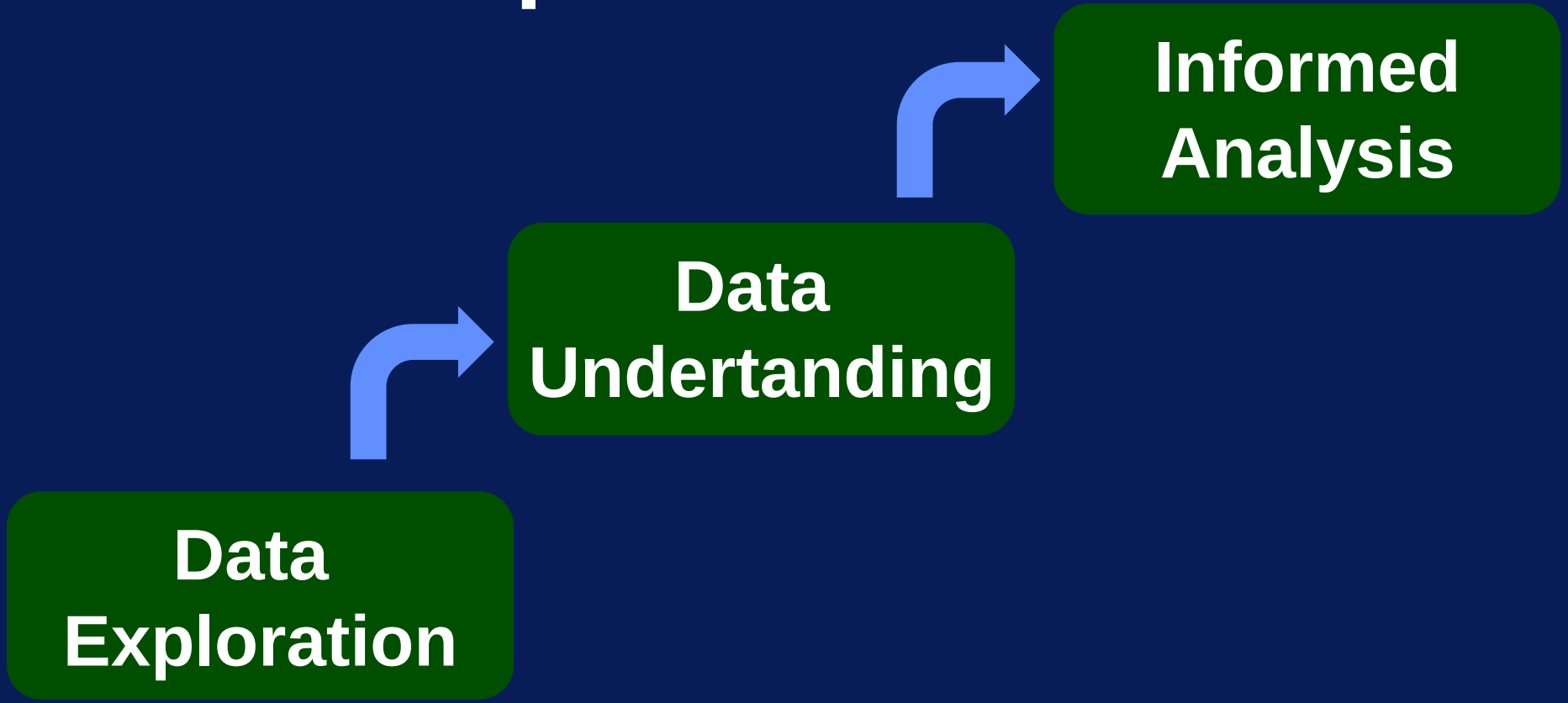


Outliers

- Indicate potential problems with data



Data Exploration



Exploring Data through Summary Statistics

After this video you will be able to..

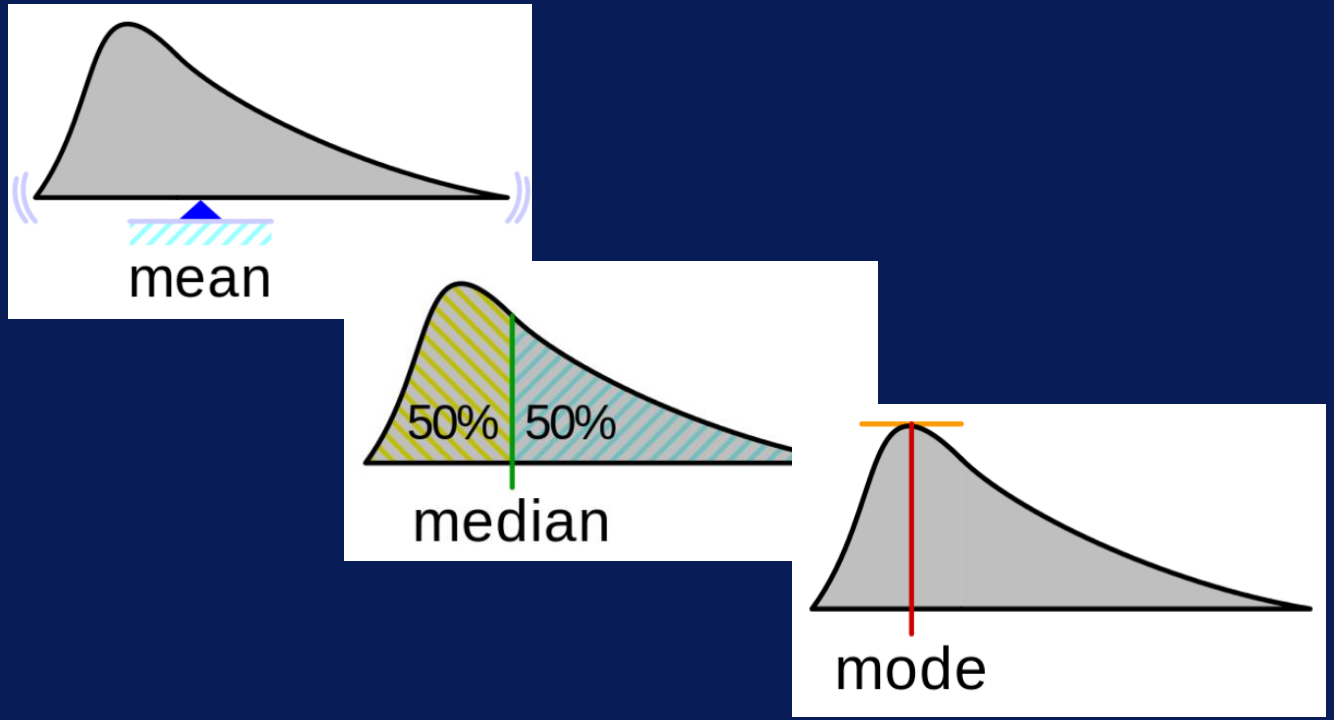
- Define what a summary statistic is
- List three common summary statistics
- Explain how summary statistics are useful in exploring data

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

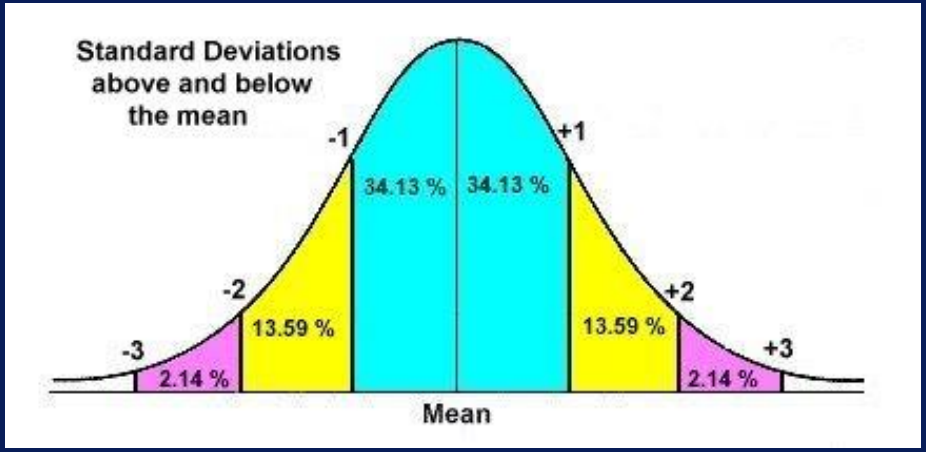
Age (sorted)
21
22
35
42
42
50
56
78
78
87

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
35
42
78
22
56
50
42
78
21
87

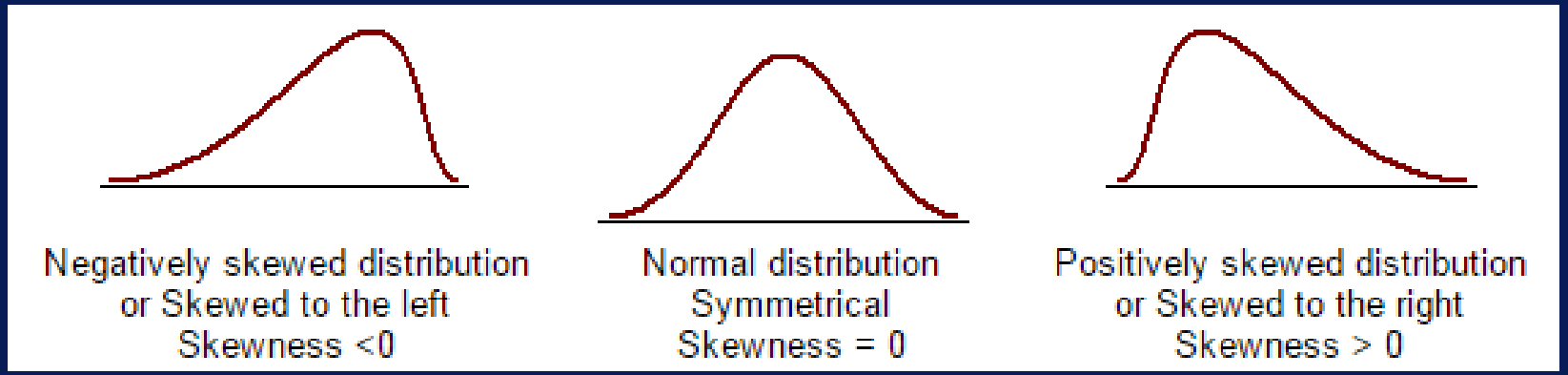
Age (sorted)
21
22
35
42
42
50
56
78
78
87

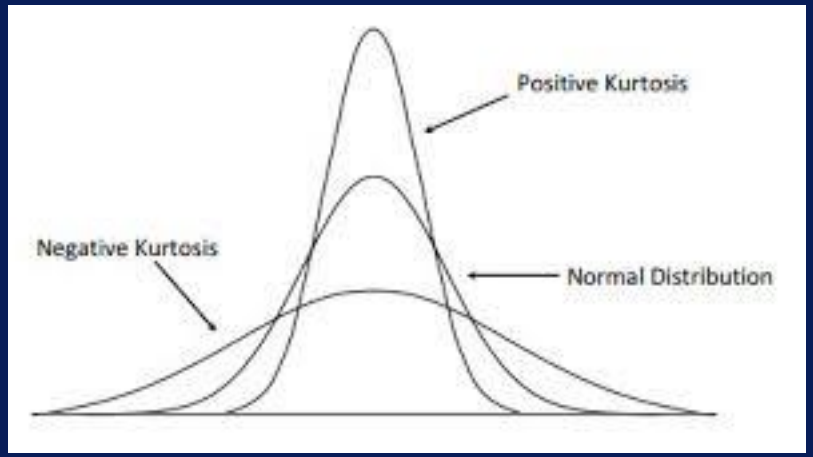
Range = 87 – 21 = 66

Variance = 548.767

Standard deviation = 23.426

Measures of Shape





skewness

kurtosis

Measures of Shape – Example

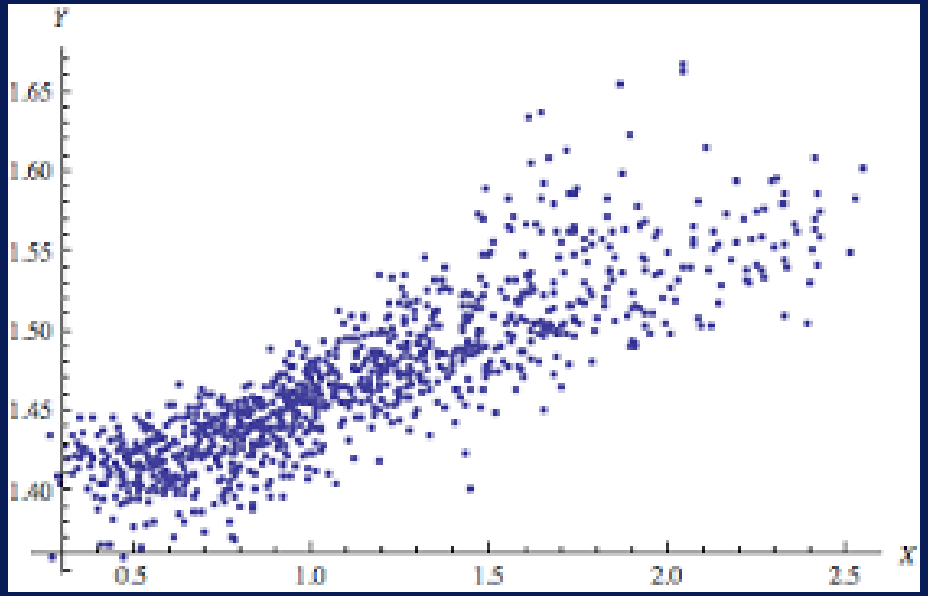
Age
35
42
78
22
56
50
42
78
21
87

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**

columns = # variables ?

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

rows = # samples ?

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

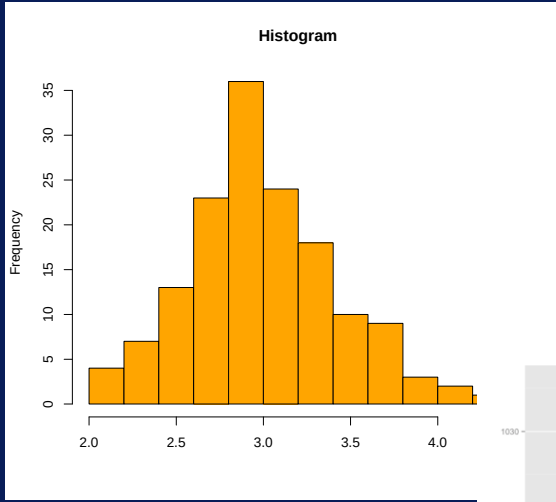
- **Measures of**
  - Location, spread, shape, dependence
- **Contingency table**
  - For categorical variables
- **Data validation**
  - Dimensions, missing values

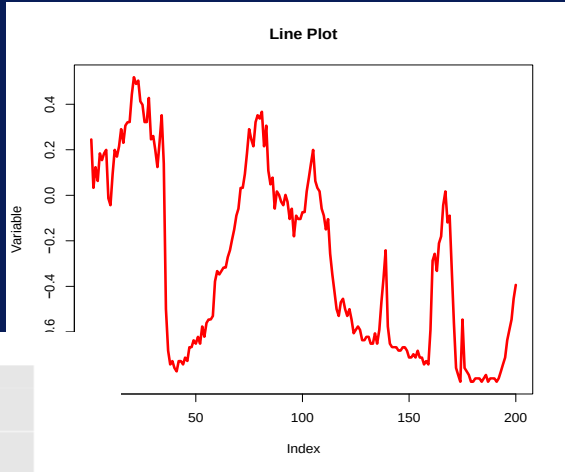
Exploring Data through Plots

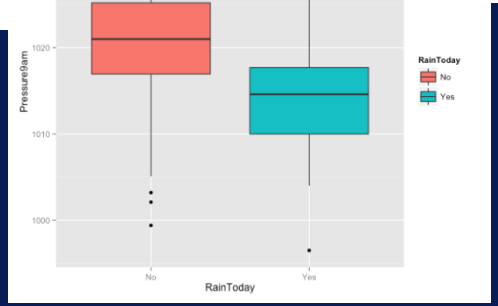
After this video you will be able to..

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

Visualizing Data



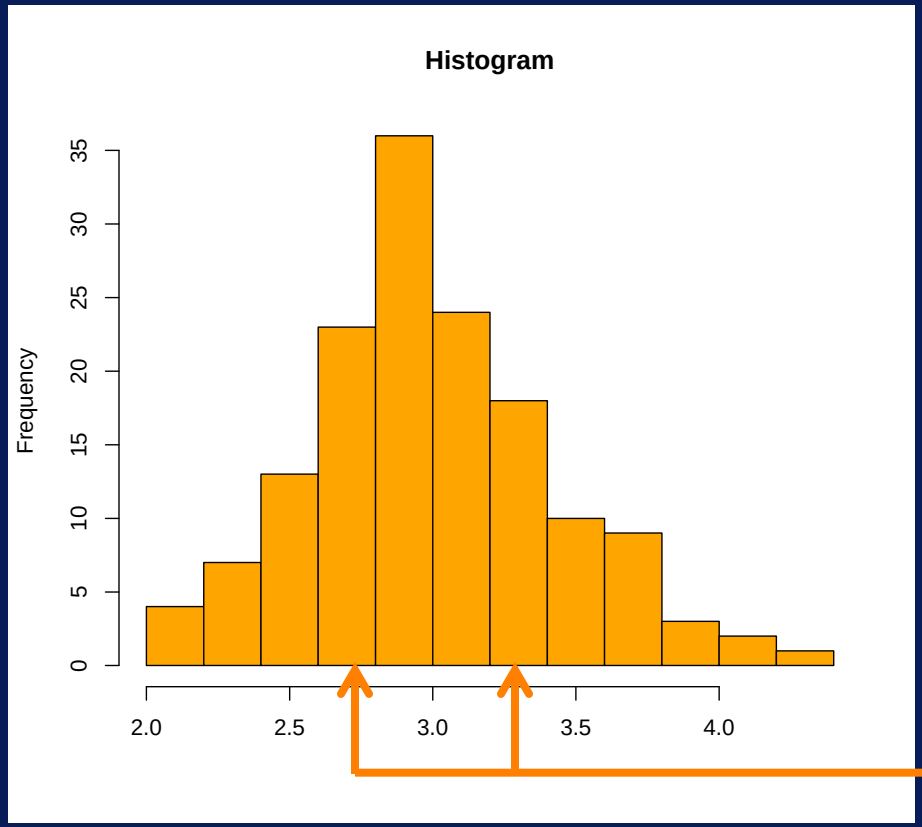




Types of Plots

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

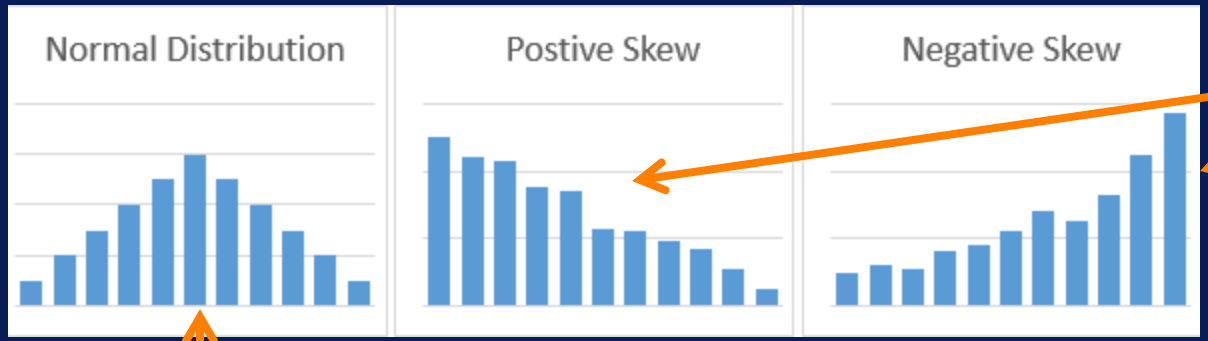
Histogram



- **Shows distribution of numeric variable**

Bins

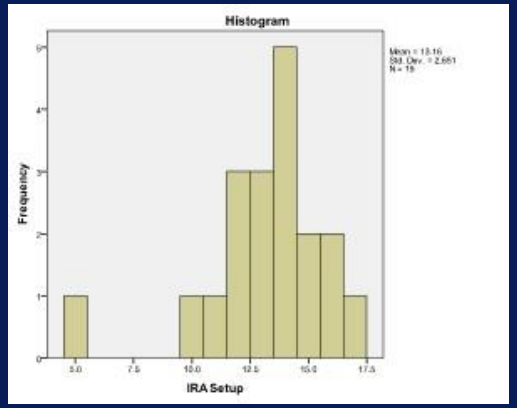
What a Histogram Shows



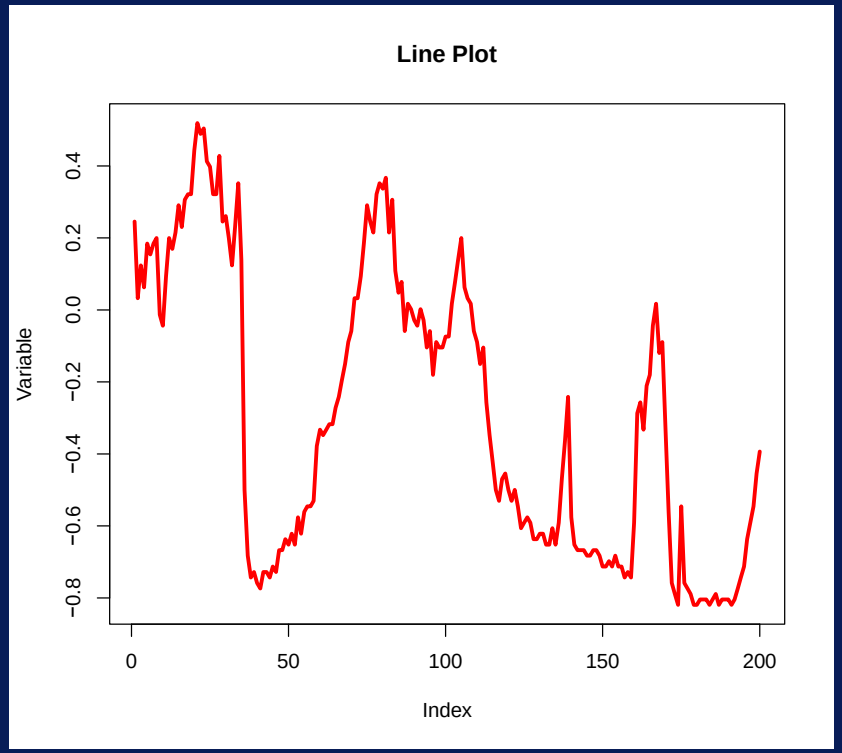
Skewness

Central Tendency

Outlier

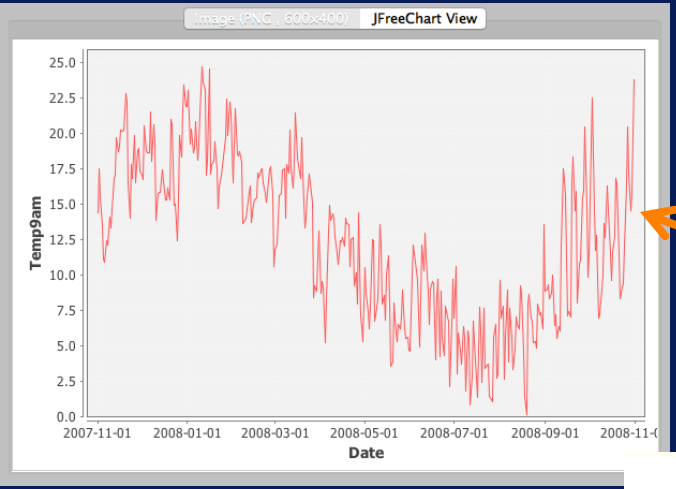


Line Plot



- **Shows change in data over time**

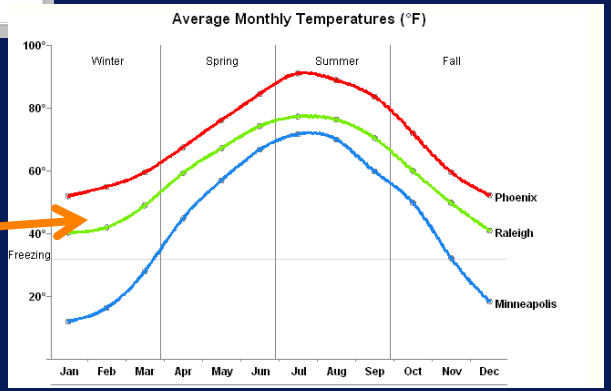
What a Line Plot Shows



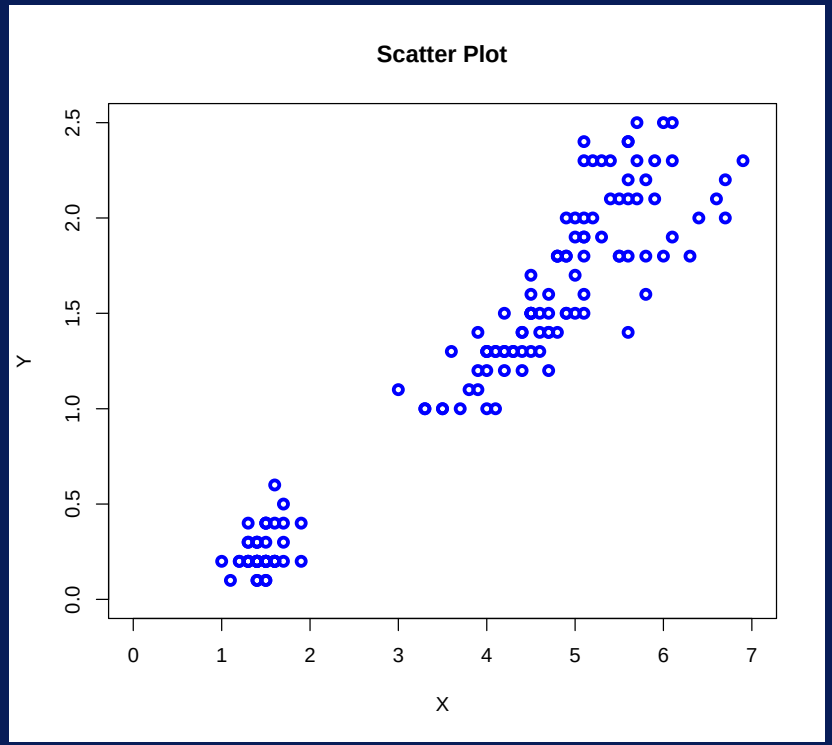
Trend

Cyclical pattern

Compare variables



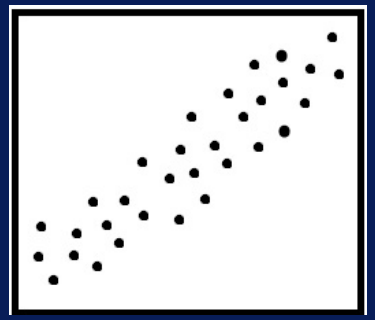
Scatter Plot

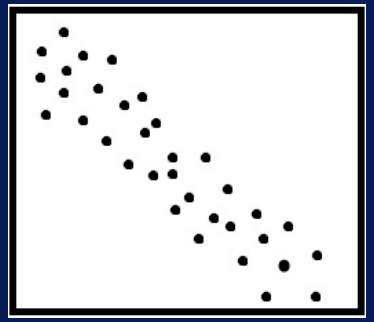


- Shows relationship between two variables

What a Scatter Plot Shows

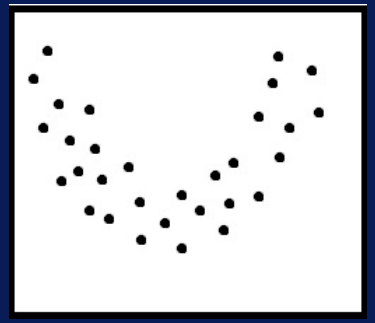
Positive Correlation

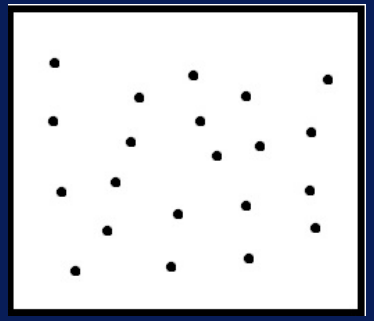




Negative Correlation

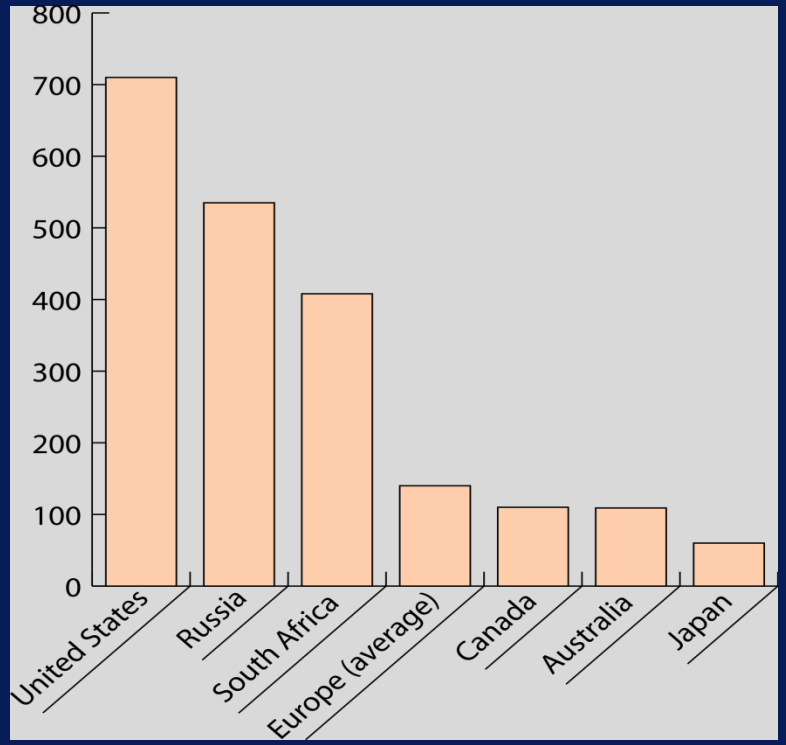
Non-Linear Correlation





No Correlation

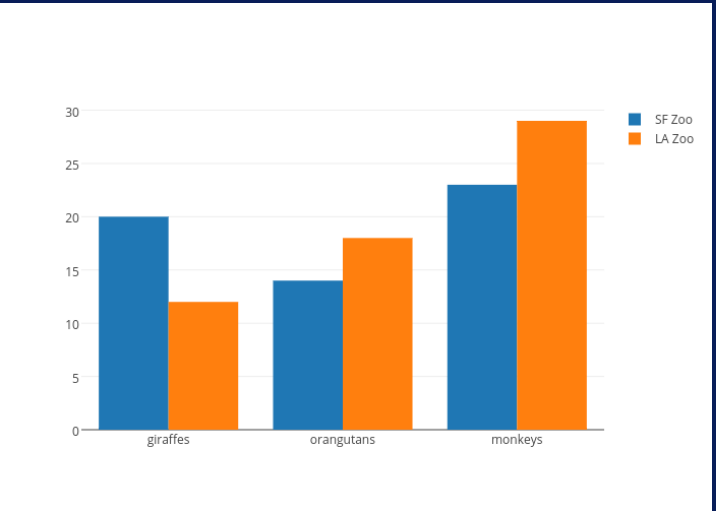
Bar Plot



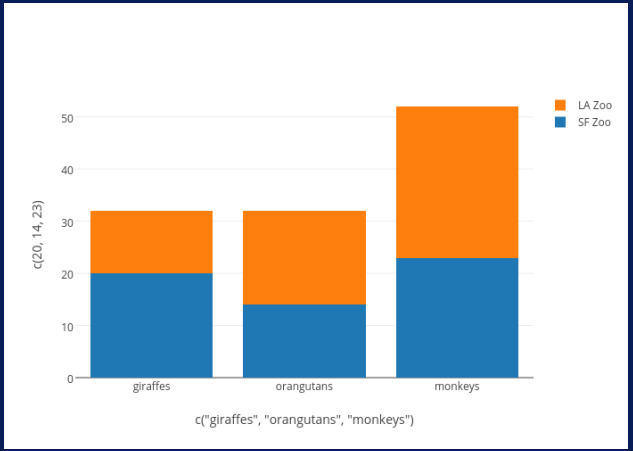
- **Shows distribution of categorical variable**

What a Bar Plot Shows

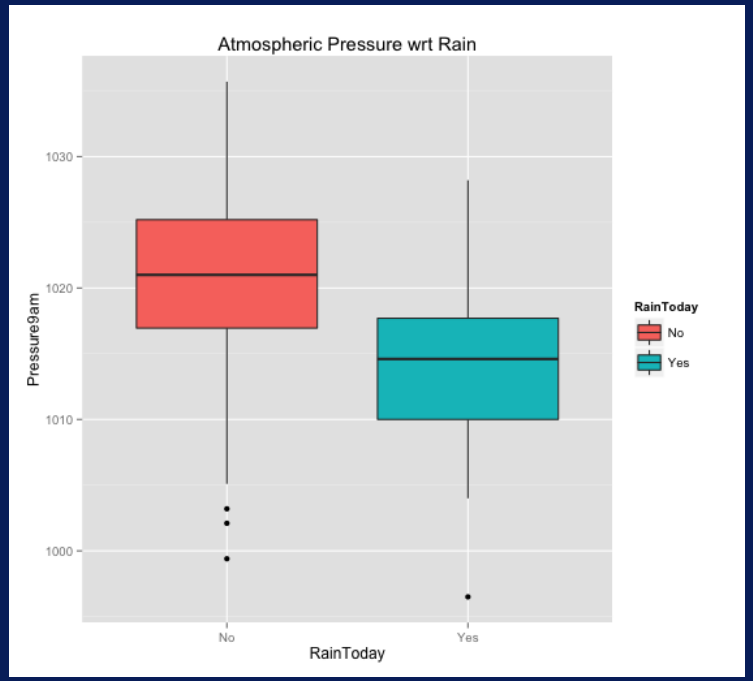
Grouped Bar Chart



Stacked Bar Chart

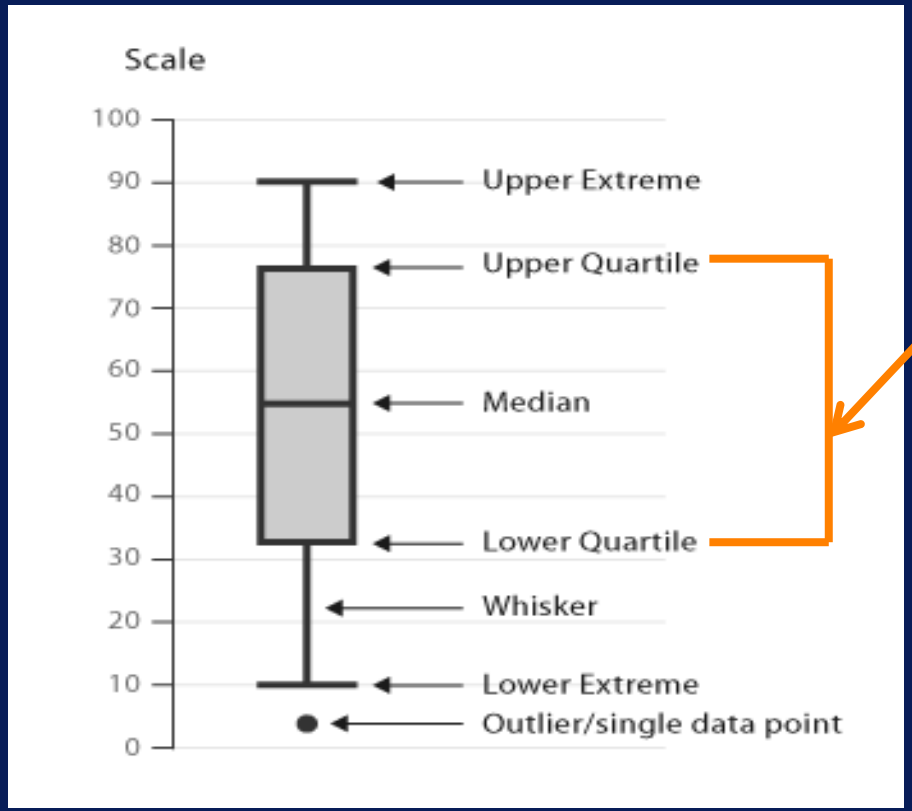


Box Plot



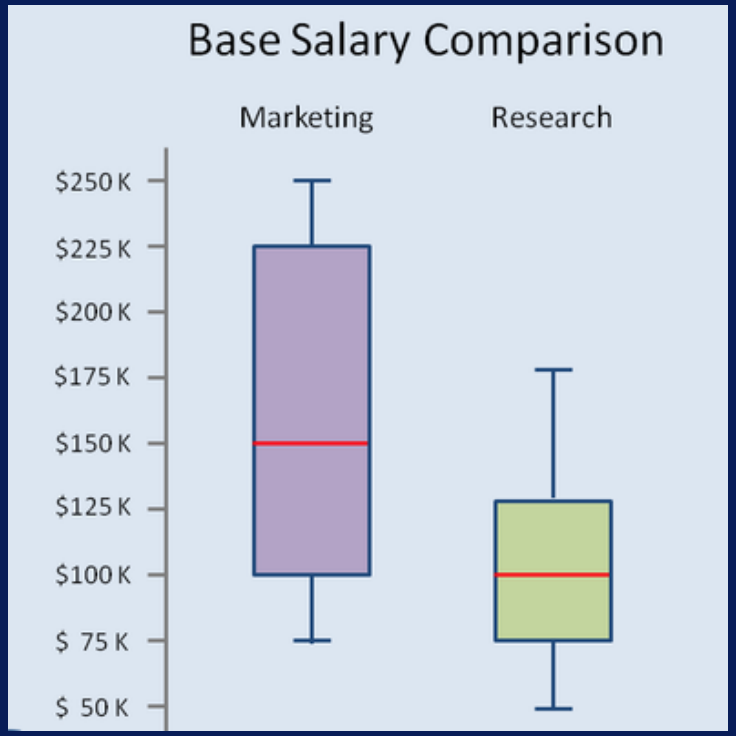
- **Compares distributions of variables**

Components of a Box Plot



The middle 50% of data are in this region

What a Box Plot Shows



What a Box Plot Shows

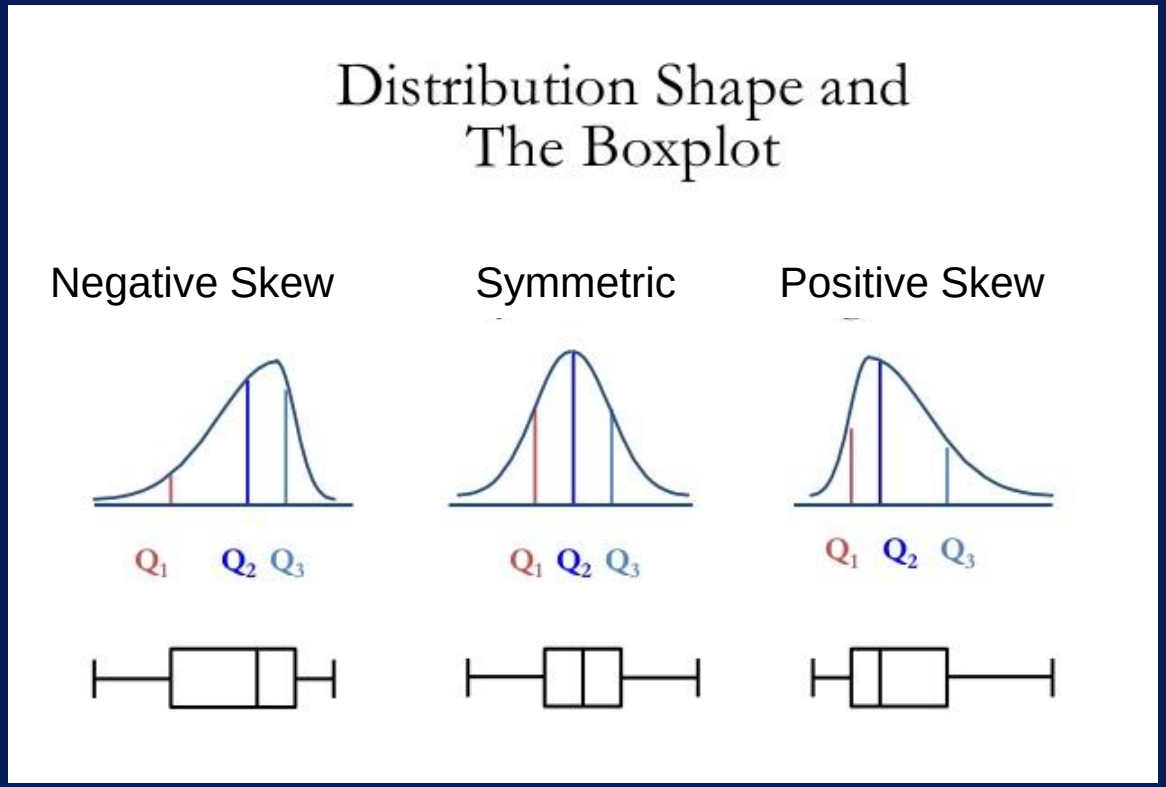


Image source: http://www.slideshare.net/mido02/chap-3gbu

Data Visualization

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

Data Preparation Overview

After this video you will be able to..

- Articulate the importance of data preparation
- Define the objectives of data preparation
- List some activities in preparing data

Preparing Data

Goal: Create data for analysis

Clean

Format

- Select features to use
- Transform data

Data Cleaning

- **Data quality issues**
  - Missing values
  - Duplicate data
  - Inconsistent data
  - Noise
  - Outliers

Addressing Data Quality Issues

- Some techniques:
  - Remove data with missing values
  - Merge duplicate records
  - Generate best estimate for invalid values

Cleaning Data

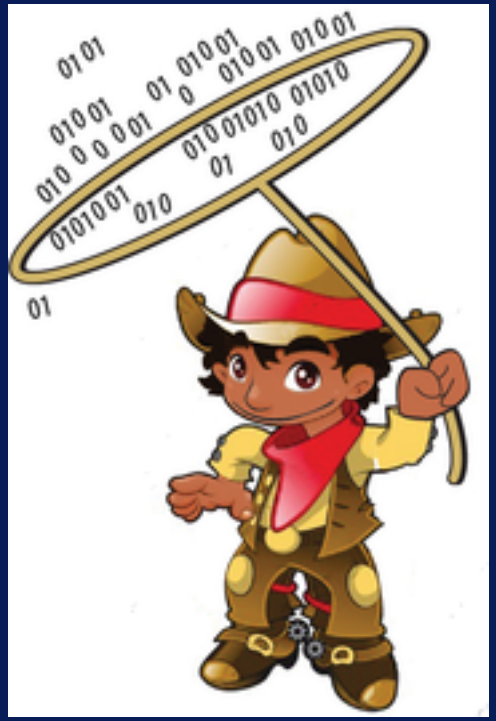
Data Cleaning

Data Cleansing

Getting Data in Shape

Data Munging

Data Wrangling

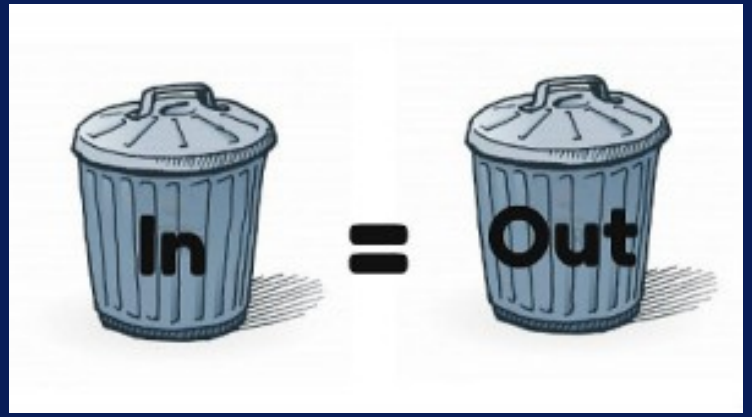


Data Preprocessing

Data Wrangling

- Feature selection
  - Combing features
  - Adding/Removing features
- Feature transformation
  - Scaling
  - Dimensionality reduction

Always Remember!



Data preparation is very important for meaningful analysis.

Garbage in = Garbage out

Data Quality

After this video you will be able to..

- Describe three data quality issues
- Name three reasons for poor data quality
- Explain why data quality issues need to be addressed

Data Quality Issues

- Real-world data is messy!

Missing Data

Name	Age	Income
Angela	34	80
Sidney	--	56
Ratan	10	--
Kiril	68	--
Zhou	45	120

Missing Values

Duplicate Data

Name	Address
Angela	430 Park Drive
Sidney	7800 West View Street
Sid	7800 West View Street
Ratan	12443 Mountain Avenue
Kiril	45 East 5th St
Kiril	1220 Mill Avenue
Zhou	4345 Apple Lane

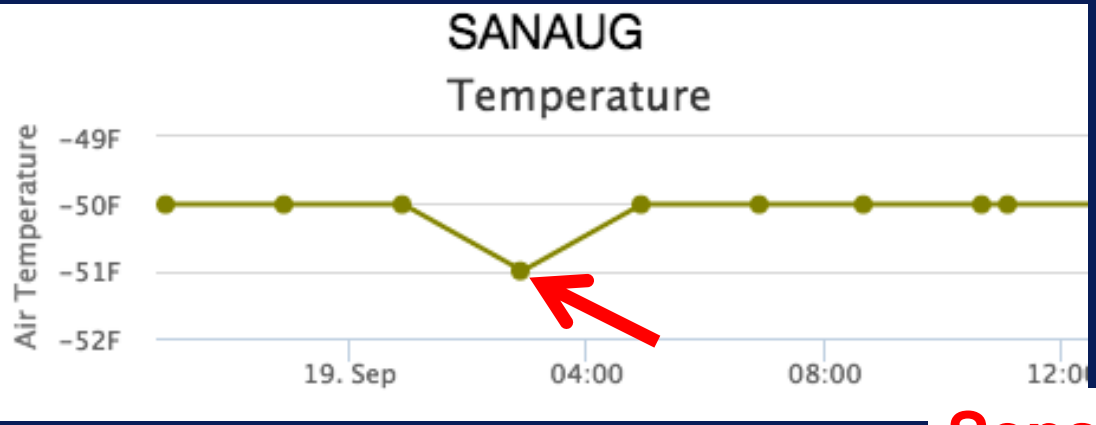
Invalid Data

Name	Zip Code
Angela	346412
Sidney	92618
Ratan	8033A
Kiril	11012
Zhou	59285

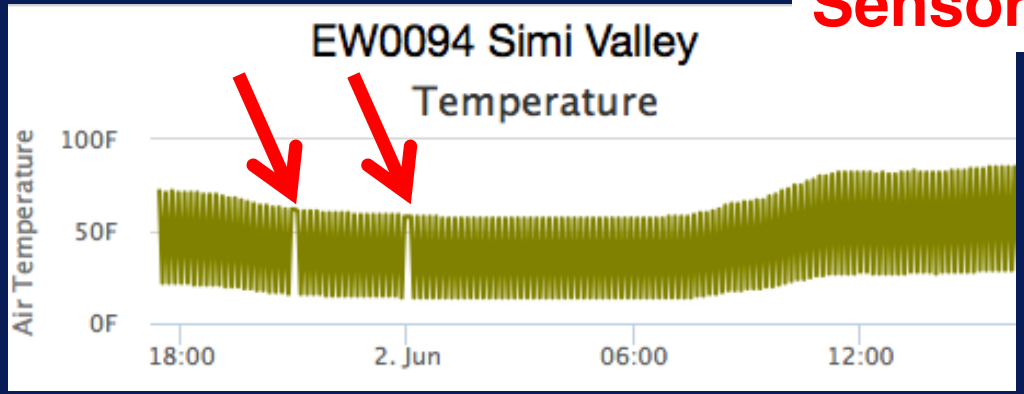
Noise

Name	Address
Angela	430 Park Drive
Sidney	780 ★❖©◆ View Street
Ratan	12443 Mountain Avenue
Kiril	1220 Mill Avenue
ZhČou	4345 Apple Lane

Outliers



Sensor Failure



Why Address Data Quality Issues?

Poor Data Quality → Poor Analysis Results

Addressing Data Quality Issues

After this video you will be able to..

- Define what ‘imputation’ means
- Illustrate three ways to handle missing values
- Describe the role of domain knowledge in addressing data quality issues

Data Quality Issues

Missing values

Duplicate data

Noise

Invalid data

Outliers

Removing Missing Data

Name	Age	Income
Angela	34	80
Sidney	--	56
Ratan	10	--
Kiril	68	--
Zhou	45	120

Name	Age	Income
Angela	34	80
~~*Sidney*~~	~~*--*~~	~~*56*~~
~~*Ratan*~~	~~*10*~~	~~*--*~~
~~*Kiril*~~	~~*68*~~	~~*--*~~
Zhou	45	120

Imputing Missing Data

Name	Age	Income
Angela	34	80
Sidney	--	56
Ratan	10	--
Kiril	68	--
Zhou	45	120

- **Replace missing values with something reasonable**

Name	Age	Income
Angela	34	80
Sidney	***50***	***56***
Ratan	***10***	***50***
Kiril	***68***	***50***
Zhou	45	120

Ways to Impute Missing Data

- **Replace missing value with**
  - Mean
  - Median
  - Most frequent
  - Sensible value based on application

Duplicate Data

- **Delete older record.**
- **Merge duplicate records**

Name	Address
Sidney	7800 West View Street
Sid	7800 West View Street
Kiril	45 East 5th St
Kiril	1220 Mill Avenue

Name	Address
Sidney	7800 West View Street
~~*Sid*~~	~~*7800 West View Street*~~
~~*Kiril*~~	~~*45 East 5th St*~~
Kiril	1220 Mill Avenue

Invalid Data

- Use external data source to get correct value
- Apply reasoning and domain knowledge to come up with reasonable value.

Name	Zip Code
Angela	346412
Ratan	8033A

Name	Zip Code
Angela	34641***2***
Ratan	8033***1***

Noise

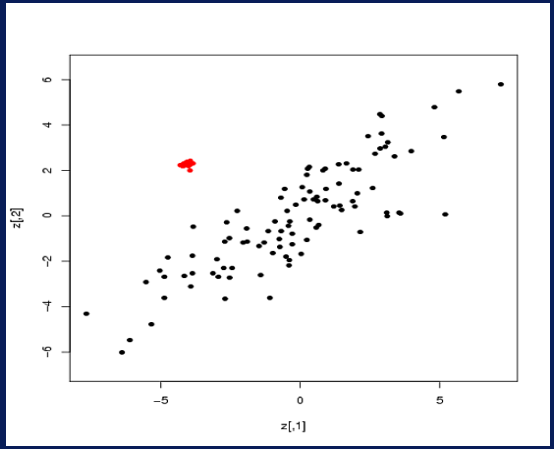
- Filter out noise component.
- May also filter out part of data, so care must be taken.

Name	Address
Sidney	7800 ★❖©◆ View Street
ZhČou	4345 Apple Lane

Name	Address
Sidney	7800 ★❖©◆ View Street
ZhČou	4345 Apple Lane

Outliers

- Remove outliers if they're not focus of analysis
- Analyze more closely if they are focus of analysis (e.g., fraud detection)



Domain Knowledge

- Required for addressing data quality issues effectively

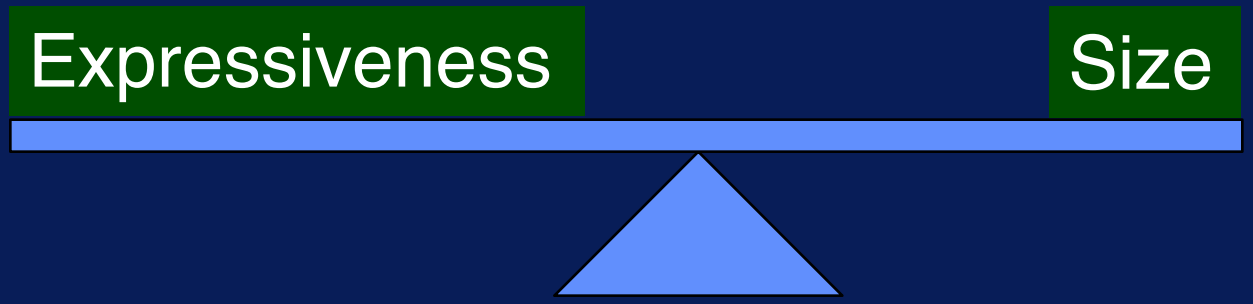
Feature Selection

After this video you will be able to..

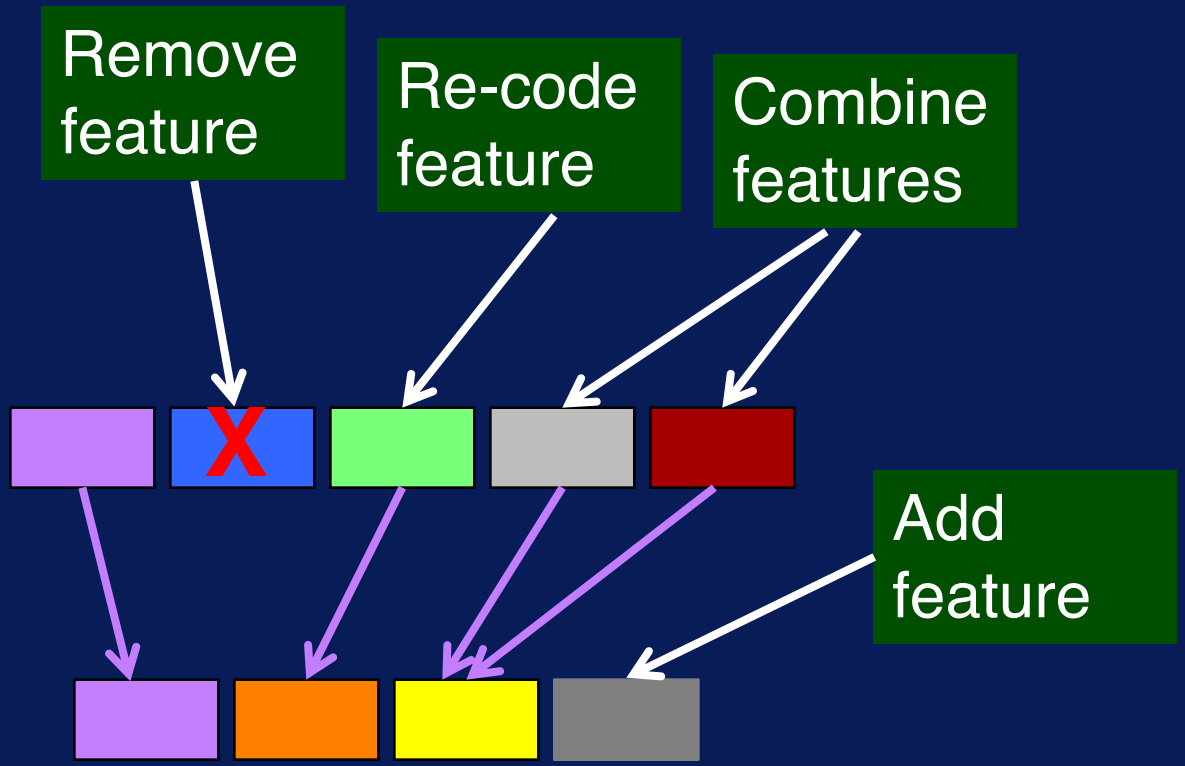
- Explain what feature selection involves
- Discuss the goal of feature selection
- List three approaches for selecting features

What is Feature Selection?

Characterize problem with smallest set of features



Feature Selection Methods



Adding Features

New features derived from existing features

Name	State
Angela	AK
Sidney	CA
Ratan	WA
Kiril	OR
Zhou	CA

Name	State	*In-State*
Angela	AK	***F***
Sidney	CA	***T***
Ratan	WA	***F***
Kiril	OR	***F***
Zhou	CA	***T***

Removing Features

- Features that are very correlated
- Features with a lot of missing values
- Irrelevant features: ID, row number, etc.

Combining Features

Name	Height	Weight
Angela	1.8	68
Sidney	1.5	70
Ratan	2.0	84
Kiril	1.3	54
Zhou	2.0	61

Name	Height	Weight	*BMI*
Angela	180	68	***21***
Sidney	153	70	***30***
Ratan	204	84	***20***
Kiril	133	44	***25***
Zhou	208	81	***19***

Recoding Features

- **Examples**
  - Discretization: re-format continuous feature as discrete
  - Customer’s age => {teenager, young adult, adult, senior}

Breaking Up Features

Address
430 Park Drive, CA, 97283
7800 W. View Street, FL, 34642
1243 Mountain Ave., CO, 80334
1220 Mill Avenue, IL, 54622
4345 Apple Lane, W

Address	State	Zip
430 Park Drive	CA	97283
7800 W. View Street	FL	34642
1243 Mountain Ave.	CO	80334
1220 Mill Avenue	IL	54622
4345 Apple Lane	WA	98421

Feature Selection Summary

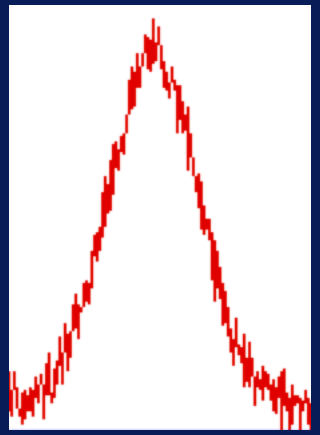
- Goal: Select smallest set of features that best captures data for application.
- Domain knowledge is important
- aka ‘feature engineering’

Feature Transformation

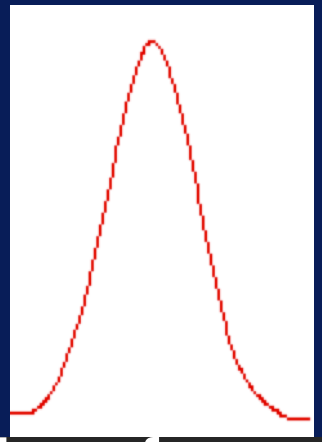
After this video you will be able to..

- Articulate the purpose of feature transformation
- List three feature transformation operations
- Discuss when scaling is important

Feature Transformation



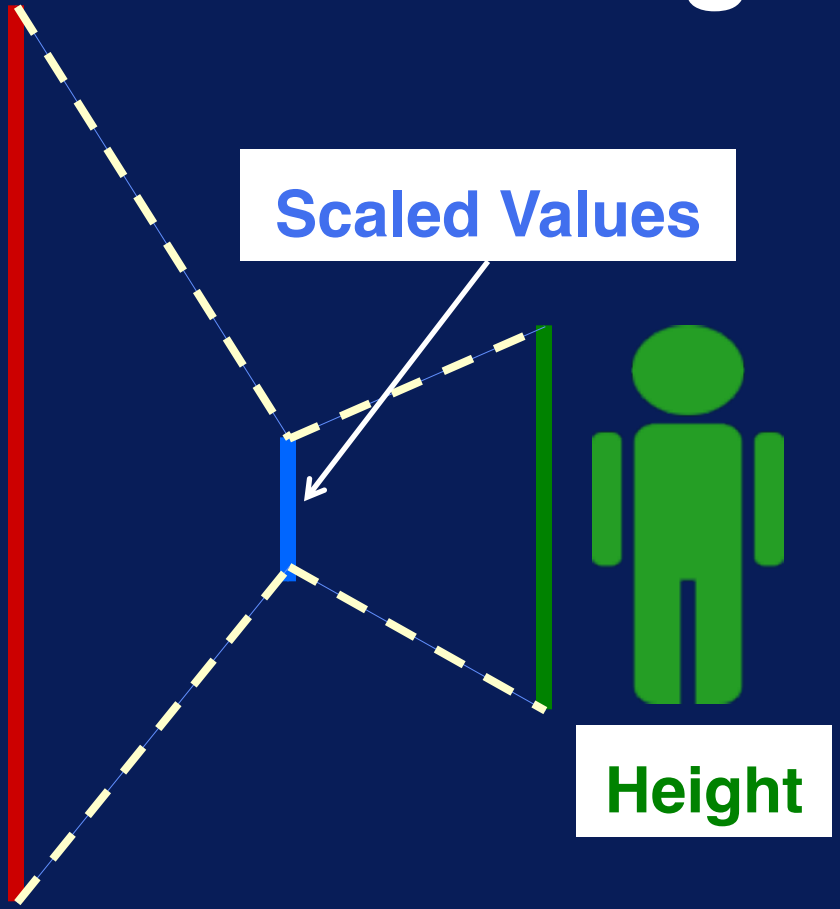
Original Data



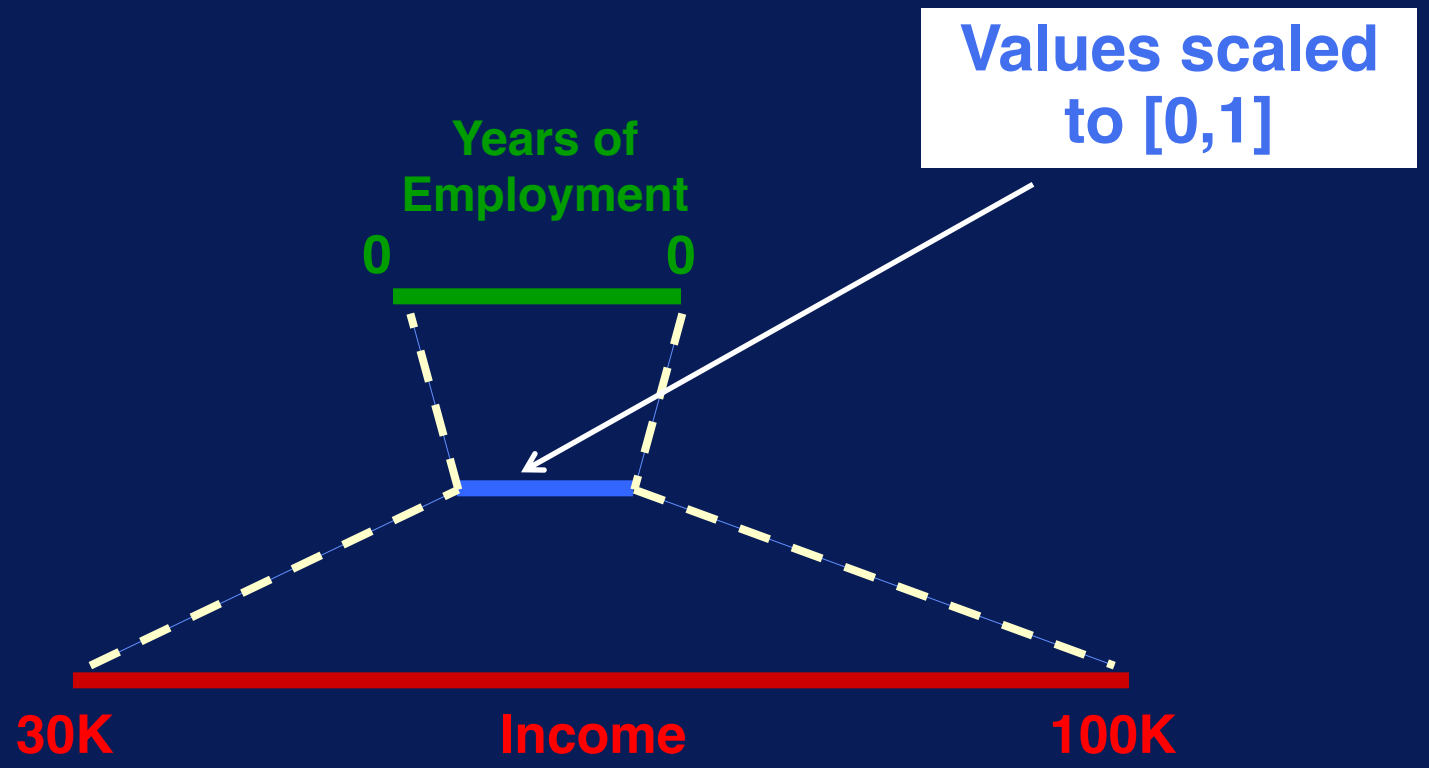
Transformed Data

Scaling

Weight

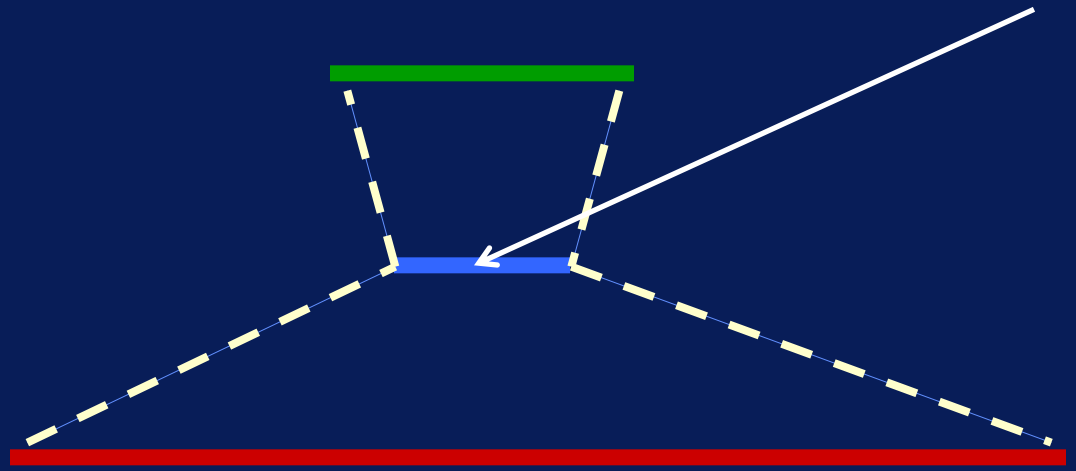


Scaling to a Range



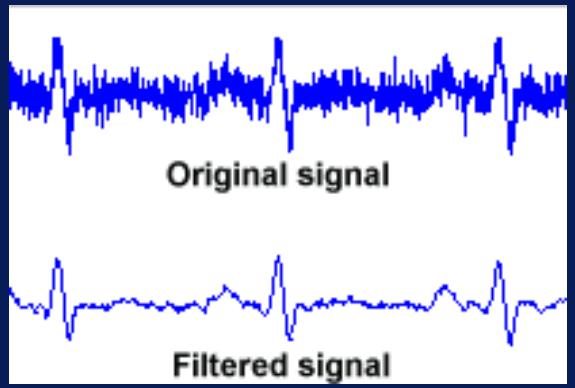
Zero-Normalization / Standardization

Mean = 0
Standard Deviation = 1

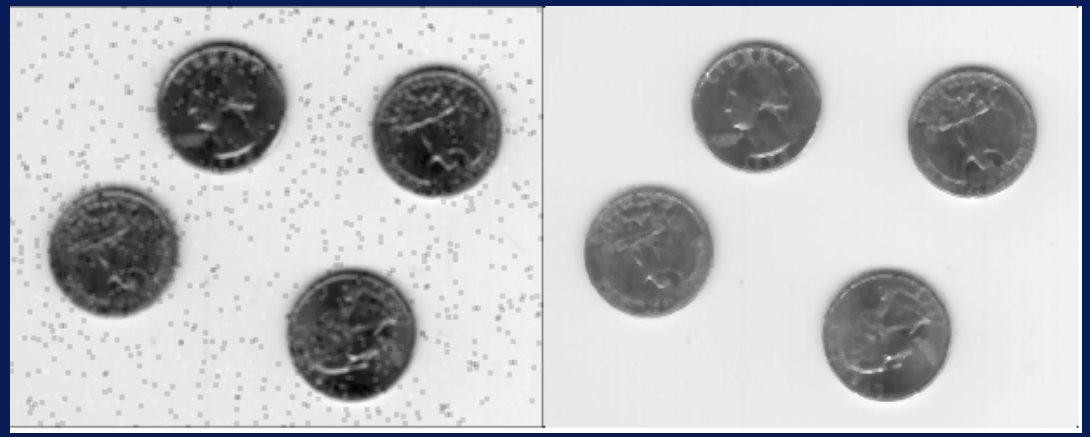


Filtering

Filter noise from audio signal

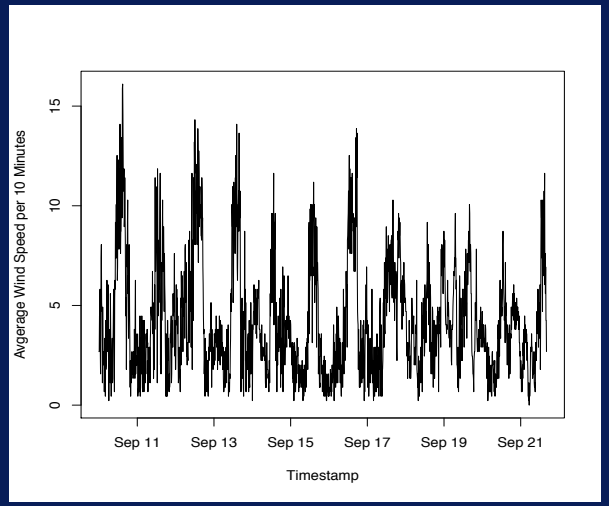


Remove grainy appearance in images

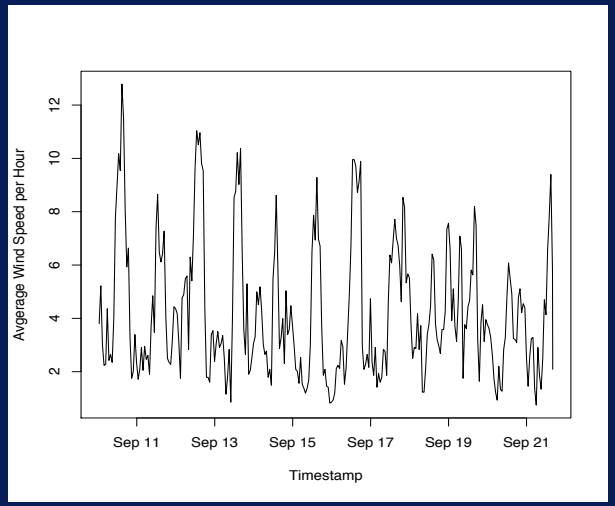


Aggregation

Avg Wind Speed
(every 10 minutes)



Avg Wind Speed
(every 60 minutes)



Feature Transformation

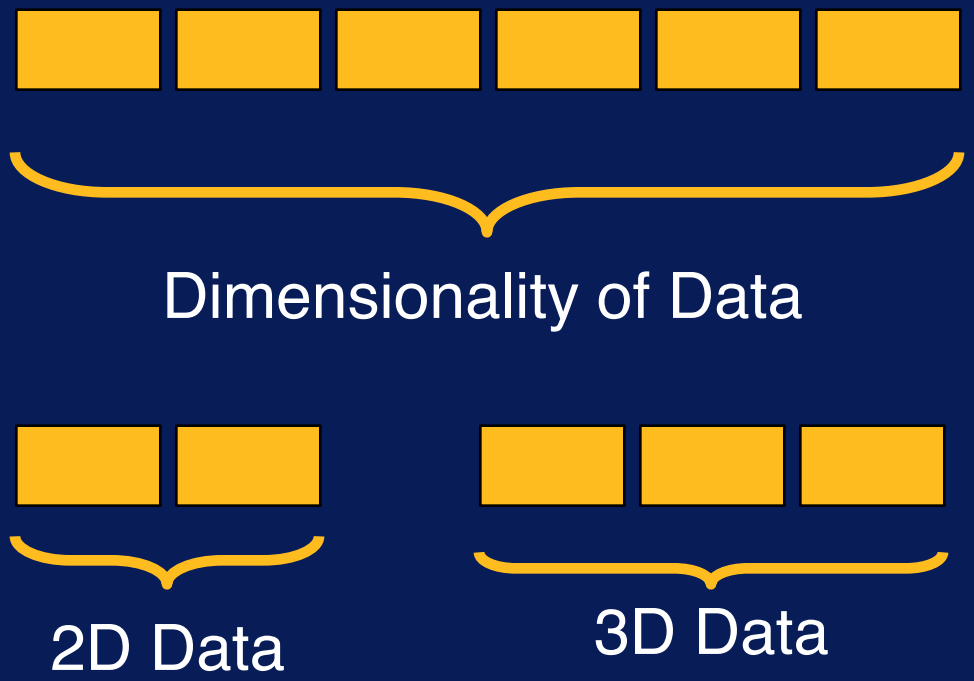
- **What:** Map feature values to new set of values
- **Why:** Have data in format suitable for analysis
- **Caveat:** Take care not to filter out important characteristics of data

Dimensionality Reduction

After this video you will be able to..

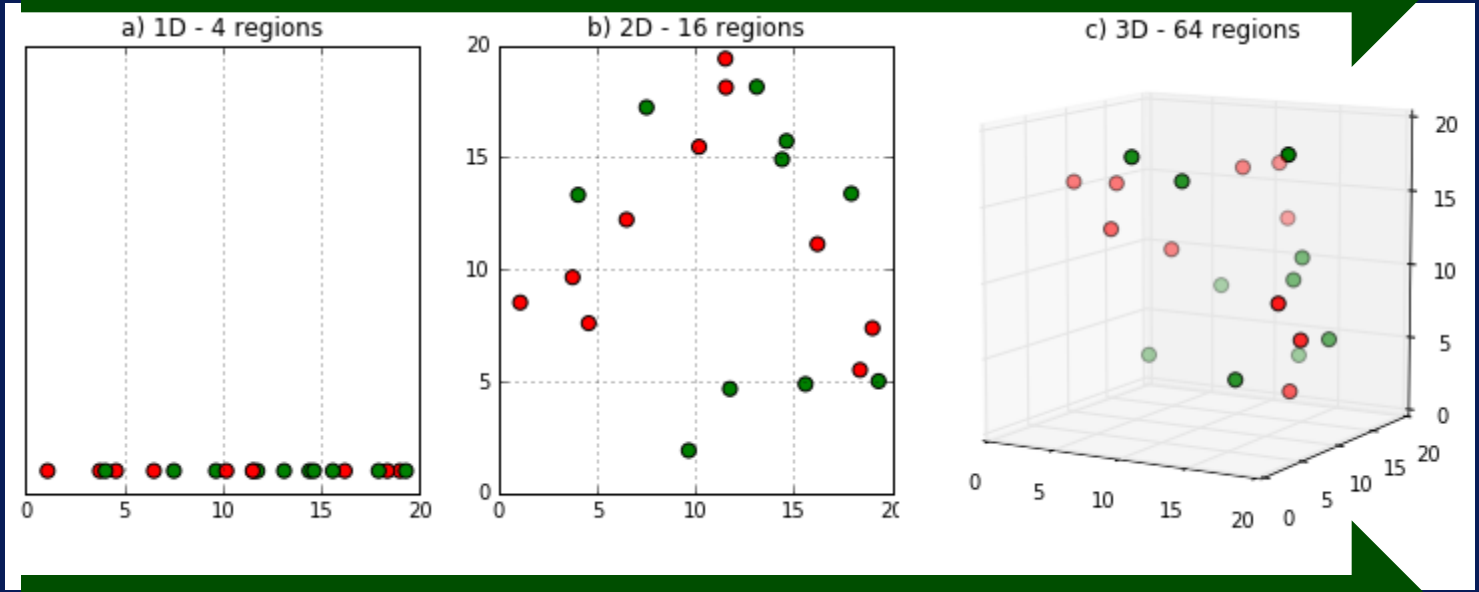
- Explain what dimensionality reduction is
- Discuss the benefits of dimensionality reduction
- Describe how PCA transforms your data

Dimensionality of Data



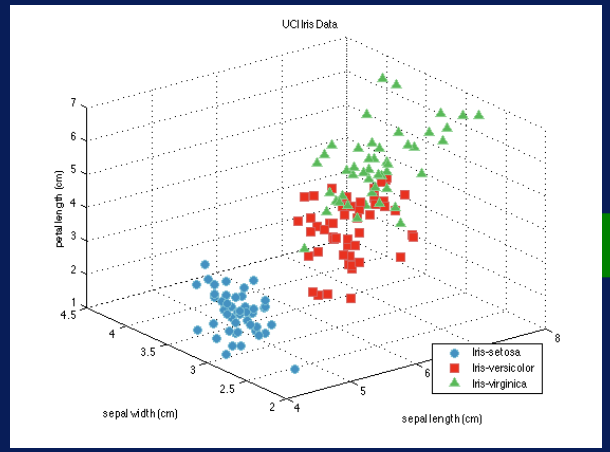
Curse of Dimensionality

Data gets increasingly sparse

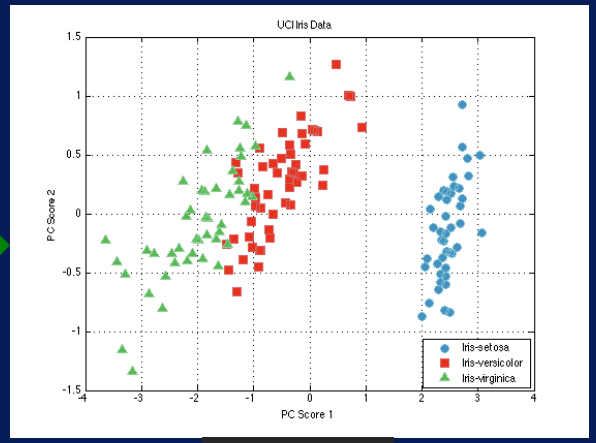


Analysis results degrade

Dimensionality Reduction

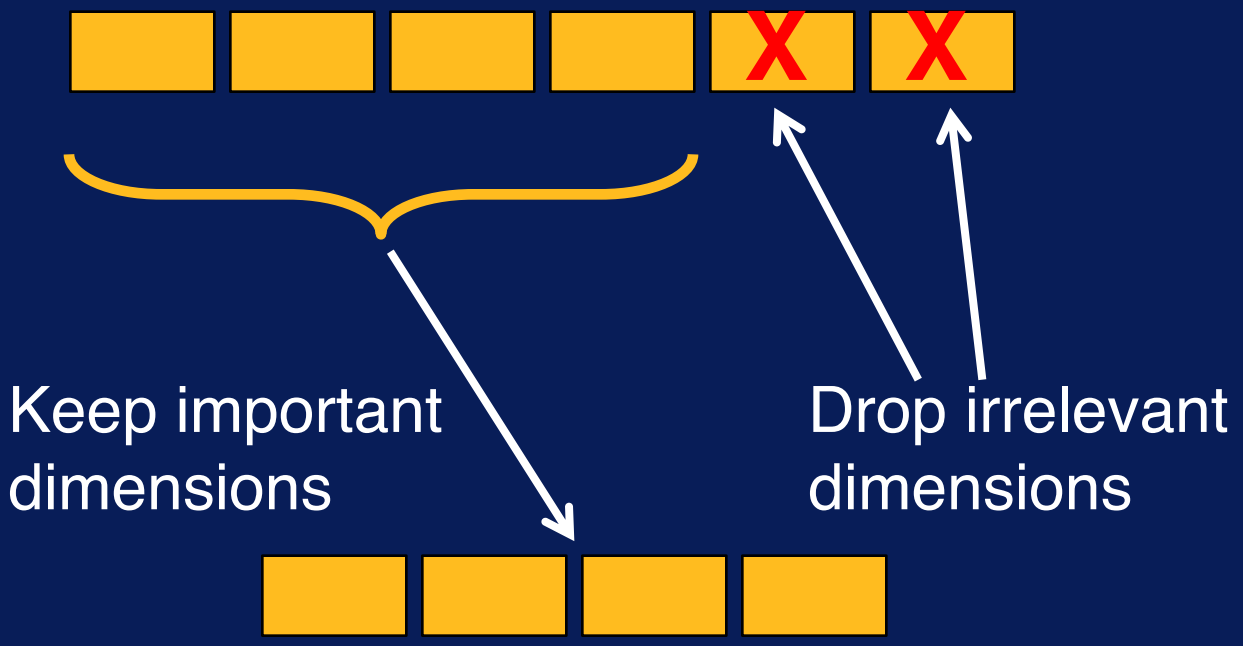


3D

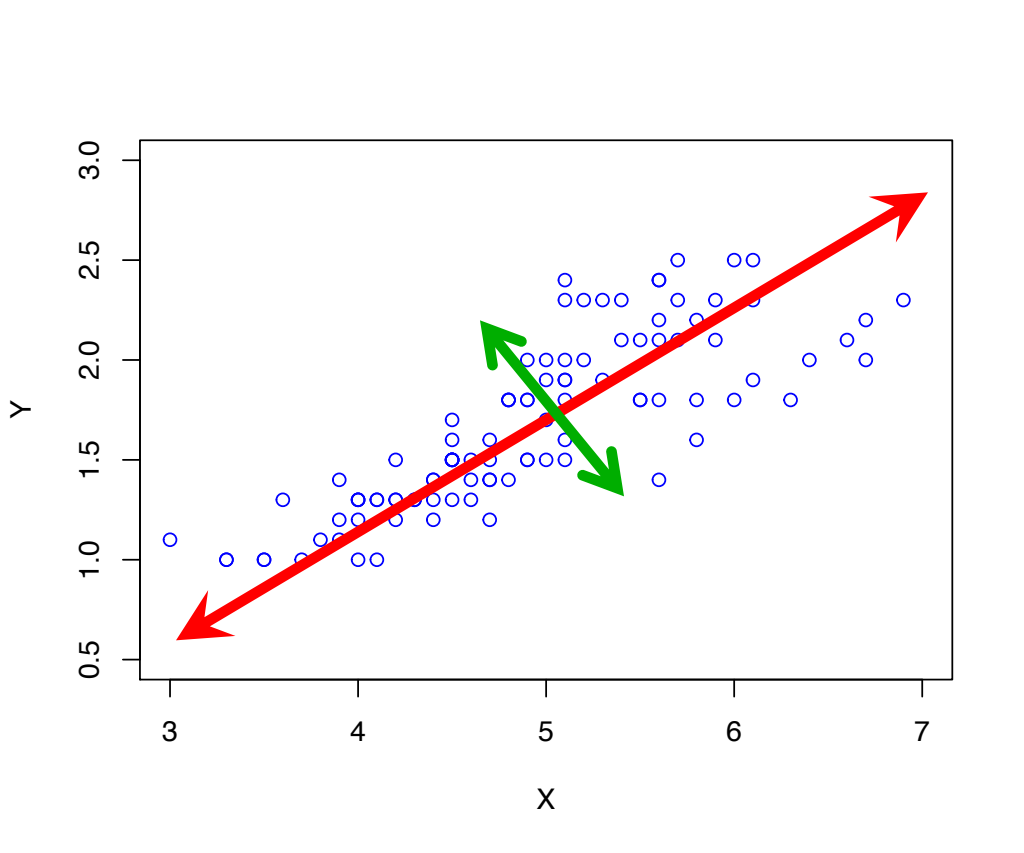


2D

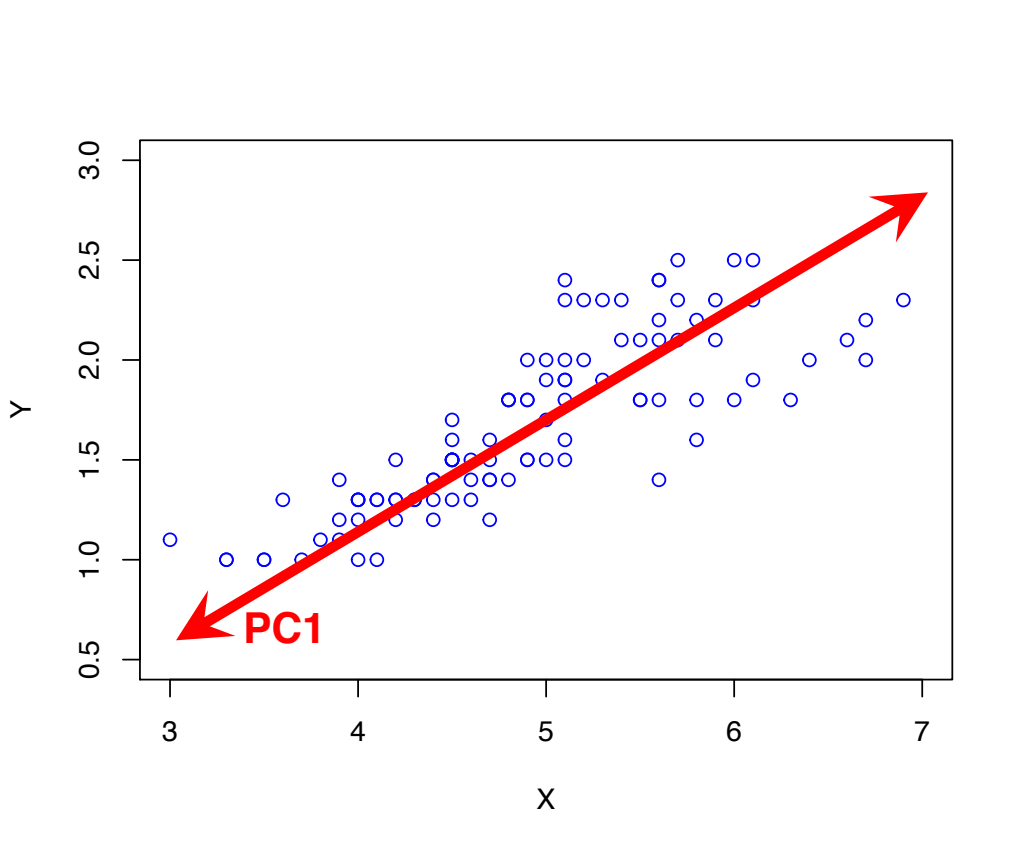
Dimensionality Reduction



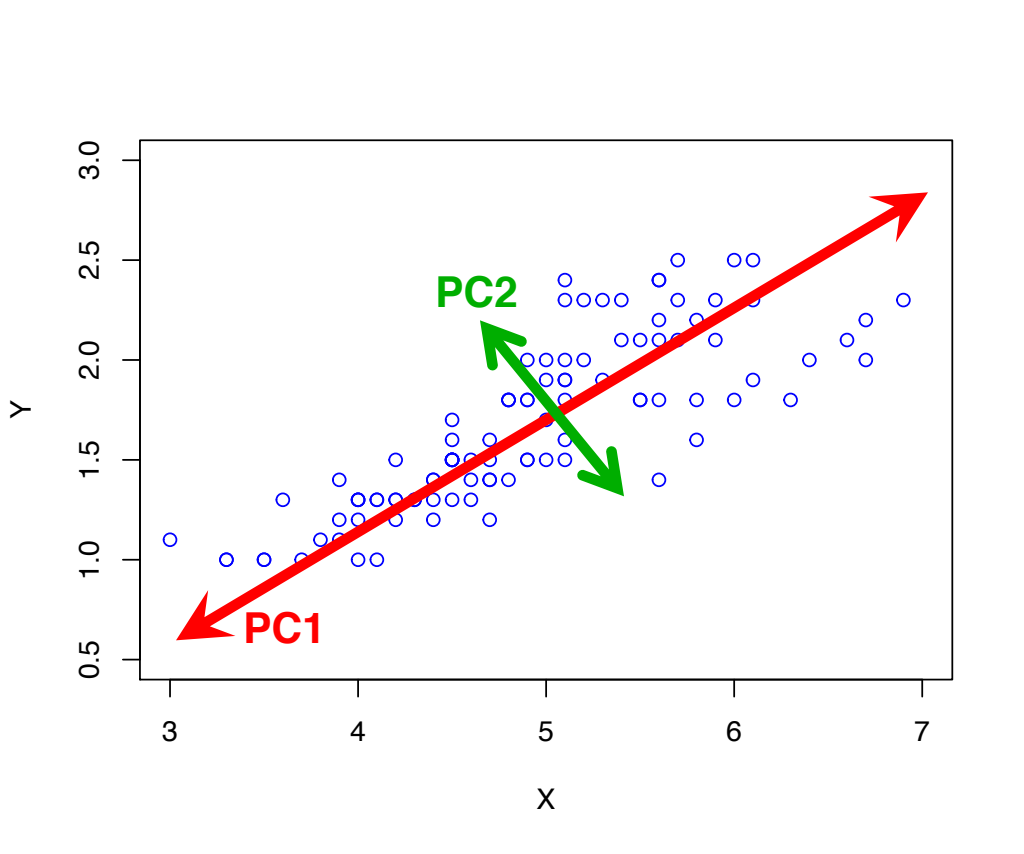
Principal Component Analysis



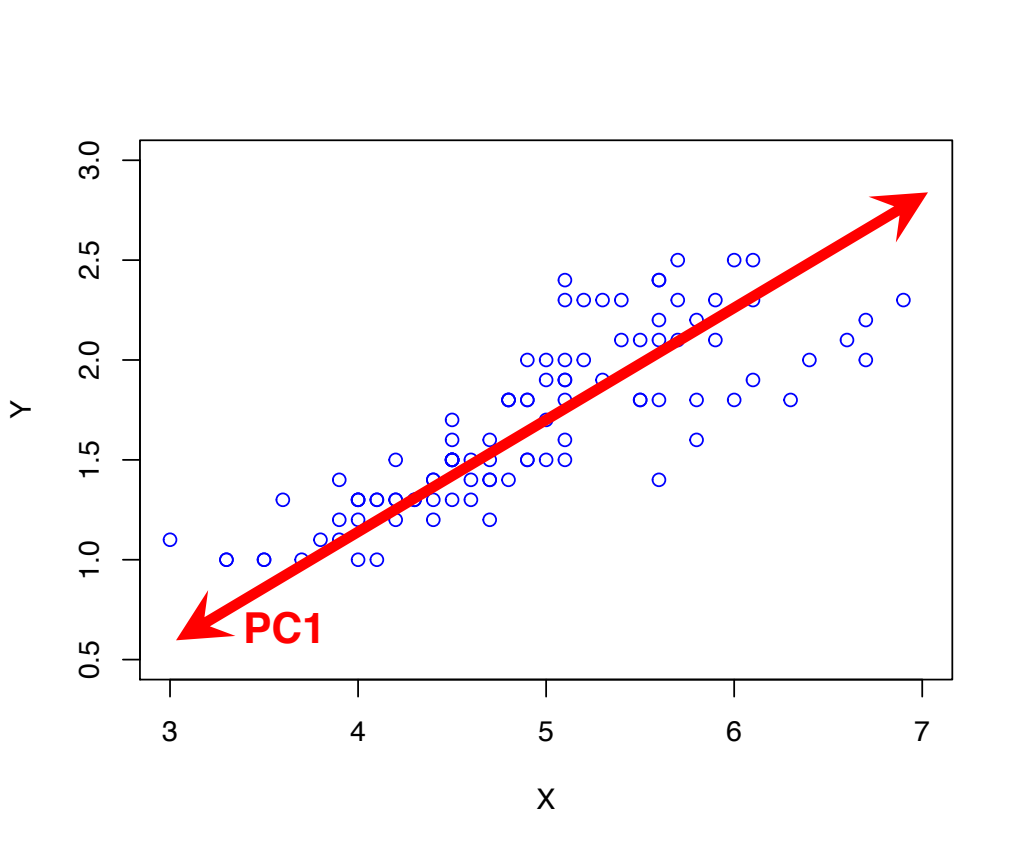
Principal Component Analysis



Principal Component Analysis



PCA for Dimensionality Reduction



PCA Main Points

- **Finds a new coordinate system such that**
  - PC1 captures greatest variance
  - PC2 captures second greatest variance, etc.
- **First few PCs capture most of variance**
  - Define lower-dimensional space for data.

PCA – To Note

- **Original dimensions**
  - Income, age, occupation, etc.
- **New dimensions**
  - PC1, PC2, PC3, etc.
- **More difficult to interpret!**