Data Management

Outline

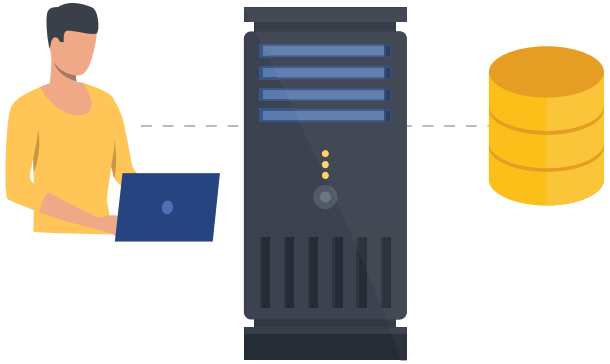
Data Collection and Acquisition

1. **Data Sources**
2. **Data Representation**
    - **Data Matrix**
    - **Types of Data**
    - **Attributes**
3. **Preparing Data**
    - **Encoding of Categorical Data**
    - **Normalization and Standardization**
    - **Data Quality**
    - **Data Cleaning**
        - **Inconsistent Datatypes**
        - **Missing data**
        - **Duplicate data**

Data Sources

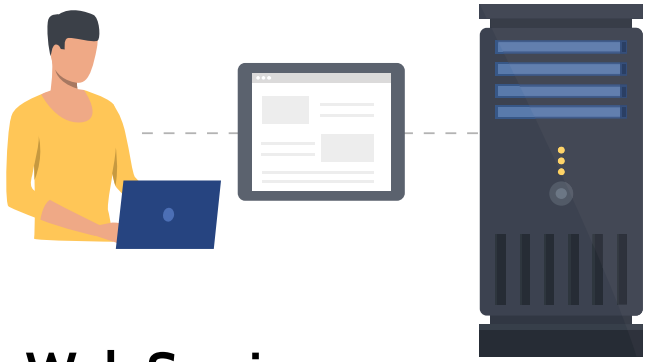
Questionnaires

- Paper-based questionnaires
- Electronic-based questionnaires
- Online questionnaires



Web Servers

Server software, or hardware dedicated to running said software, that can satisfy World Wide Web client requests.



Web Services

A service offered by an electronic device to another electronic device, communicating with each other via the World Wide Web

Data Sources

Database

An organized collection of data, generally stored and accessed electronically from a computer system

Logs

- Records of events.
- In computer, for example, a file that records either events that occur in an operating system or other software runs, or messages between different users of a communication software.

Online Repositories

- A repository is a central place in which an aggregation of data is kept and maintained in an organized way, usually in computer storage.
- An online repository is a digital library or archive which is accessible via the internet.

Data Sources

Suggested Data Sources

- **UCI Machine Learning Repository**
  https://archive.ics.uci.edu/ml/index.php
- **Kaggle**
  https://www.kaggle.com/datasets
- **Open Government Data of Thailand**
  https://data.go.th/

Data Matrix

Data Representation

Example: Cosmic Dataset

	name X_1	id X_2	align X_3	eye X_4	hair X_5	gender X_6	alive X_7	appearances X_8	first_appear X_9	publisher X_{10}
$\mathbf{x}_1$	Spider-Man (Peter Parker)	Secret	Good	Hazel Eyes	Brown Hair	Male	Living Characters	4043	Aug-62	marvel
$\mathbf{x}_2$	Captain America (Steven Rogers)	Public	Good	Blue Eyes	White Hair	Male	Living Characters	3360	Mar-41	marvel
	…	…	…	…	…	…	…	…	…	…
$\mathbf{x}_n$	Natalia Romanova (Earth-616)	Public	Good	Green Eyes	Red Hair	Female	Living Characters	1050	Apr-64	marvel

Data Matrix

Data Representation

Attributes, Variable, Features, Field,...

$$\text{Dataset} \quad D = \begin{pmatrix} X_1 & X_2 & \cdots & X_d \\ x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nd} \end{pmatrix} \begin{matrix} \\ \mathbf{x}_1 \\ \mathbf{x}_2 \\ \\ \mathbf{x}_n \end{matrix}$$

Entities, Instances, Examples, Records, Points, Feature-vectors, ...

$\mathbf{x}_i$ denotes the *i*th row which is a *d*-tuple given as

$$\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{id})$$

X_j denotes the jth column which is a n-tuple given as

$$X_j = (x_{1j}, x_{2j}, \ldots, x_{nj})$$

Data Matrix

Data Representation

Example: Cosmic Dataset

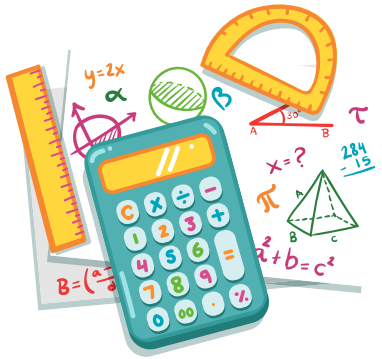
	name X_1	id X_2	align X_3	eye X_4	hair X_5	gender X_6	alive X_7	appearances X_8	first_appear X_9	publisher X_{10}
$\mathbf{x}_1$	Spider-Man (Peter Parker)	Secret	Good	Hazel Eyes	Brown Hair	Male	Living Characters	4043	Aug-62	marvel
$\mathbf{x}_2$	Captain America (Steven Rogers)	Public	Good	Blue Eyes	White Hair	Male	Living Characters	3360	Mar-41	marvel
	…	…	…	…	…	…	…	…	…	…
$\mathbf{x}_n$	Natalia Romanova (Earth-616)	Public	Good	Green Eyes	Red Hair	Female	Living Characters	1050	Apr-64	marvel

We can write an example $\mathbf{x}_2$ as

$$\mathbf{x}_2 = (Captain\ America\ (Steven\ Rogers), Public, Good, Blue\ Eyes, White\ Hair, Male, Living\ Characters, 3360, Mar-41, marvel)$$

Types of Data

Data Representation



Quantitative Data

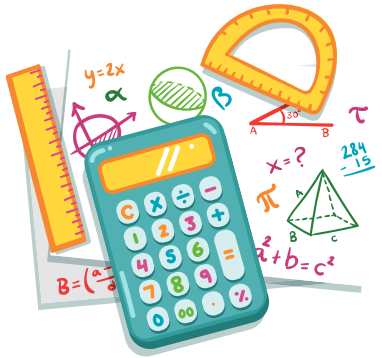
- This data can be described using **numbers.**
- **Basic mathematical procedures** are possible on the set.

Qualitative Data

- This data cannot be described using numbers and basic mathematics.
- This data is generally described using natural **categories and language**.

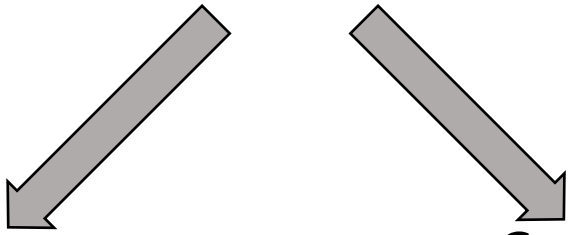
Attributes

Data Representation



Numeric Attributes - Quantitative

- One that has a real-valued or integer-valued domain.
- Such as age, height, grade, frequency, etc.



Discrete

- Take on a finite or countably infinite set
- Such as integer, grade, number of object, etc.

Continuous

- Take on any real value
- Such as height, weight, size, etc.

Categorical Attributes

- One that has a set-valued domain composed of a set of symbols.
- Such as Gender = {M,F}, Education = {High School, BS, MS, PhD}, etc.

Attributes

Data Representation

Categorical Attributes

Nominal

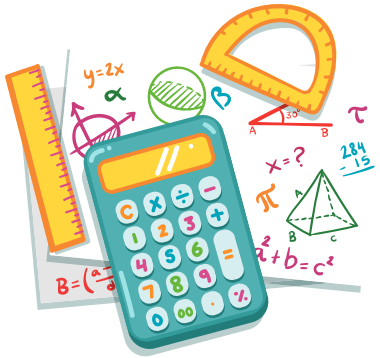
- Attribute values in the domain are unordered.
- Can only equality (=) compare.
- Such as gender, type of hair, etc.

Ordinal

- Attribute values are ordered.
- Can both equality (=) and inequality (<, >) compare.
- Such as education, feel (unhappy, OK, happy), etc.

Attributes

Data Representation



Numeric Attributes

Interval-scaled

- Can compute only differences (addition or subtraction)
- For example, temperature measured in ˚C or ˚F.
  - If it is 20 ˚C on one day and 10 ˚C on previous day
  - We **can** talk about a temperature drop of 10˚C.
  - We **cannot** say that it is twice as cold as the previous day.

Ratio-scaled

- Can compute both differences and ratio between values,
- For example, age.
  - If Jone is 20 years old and Jim is 10 years old.
  - We **can** say that Jone older than Jim with 10 years.
  - We **can** say that Jone is twice as old as Jim.

Attributes

Data Representation

Summary of data types and scale measures

Provides	Nominal	Ordinal	Interval-scaled	Ratio-scaled
The order of values is known		/	/	/
“Count,” aka “Frequency of Distribution”	/	/	/	/
Mode	/	/	/	/
Median		/	/	/
Mean			/	/
Can quantify the difference between each values			/	/
Can add or subtract values			/	/
Can multiple and divide values				/
Has “true zero”				/

https://www.mymarketresearchmethods.com/types-of-data-nominal-ordinal-interval-ratio/

Attributes

Data Representation

Cosmic Dataset

	name X_1	id X_2	align X_3	eye X_4	hair X_5	gender X_6	alive X_7	appearances X_8	first_appear X_9	publisher X_{10}
$\mathbf{x}_1$	Spider-Man (Peter Parker)	Secret	Good	Hazel Eyes	Brown Hair	Male	Living Characters	4043	Aug-62	marvel
$\mathbf{x}_2$	Captain America (Steven Rogers)	Public	Good	Blue Eyes	White Hair	Male	Living Characters	3360	Mar-41	marvel
	…	…	…	…	…	…	…	…	…	…
$\mathbf{x}_n$	Natalia Romanova (Earth-616)	Public	Good	Green Eyes	Red Hair	Female	Living Characters	1050	Apr-64	marvel

Practice: What is the type of each attribute?

Nominal, Ordinal, Interval-scaled or Ratio-scaled

Encoding of Categorical Data

Preparing Data

- Most of Machine learning algorithms can not handle categorical variables.

→We convert them to numerical values.

Nominal variable

One Hot Encoding

- Map each category to a vector that contains 1 and 0
  - 1 - presence of the feature
  - 0 - absence of the feature

Gender
Male
Female
Other

isMale	isFemale	isOther
1	0	0
0	1	0
0	0	1

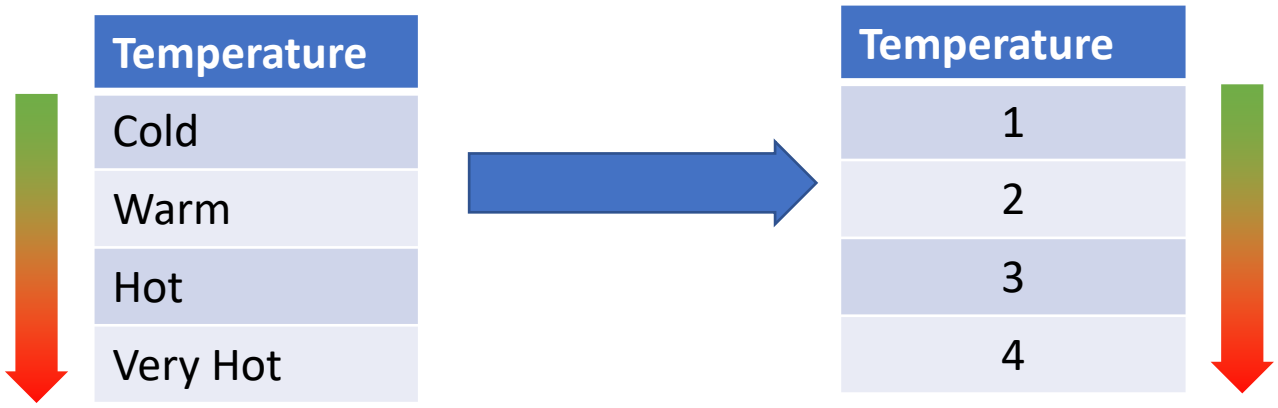
Encoding of Categorical Data

Preparing Data

Ordinal

Ordinal Encoding

- The encoding of variables retains the ordinal nature of the variable
- Each category is assigned a value from 1 through the number of possible values by considering the order of values.



Encoding of Categorical Data

Preparing Data

Practice

How can we encode the following categorical data?

Align
Bad
Neutral
Good

Hair
Black
Bronze
Brown
Gold
Gray

Normalization and Standardization

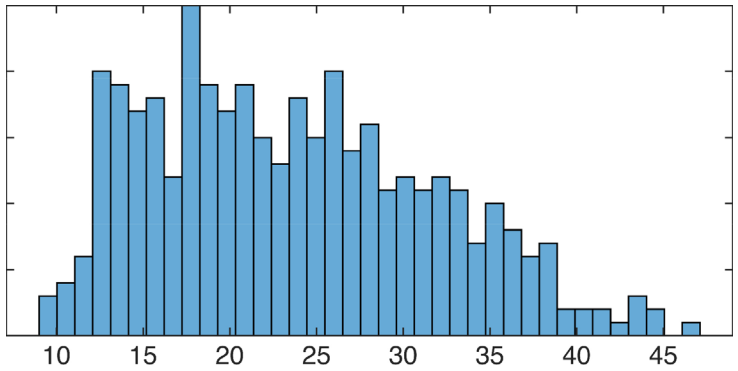
Preparing Data

Normalization

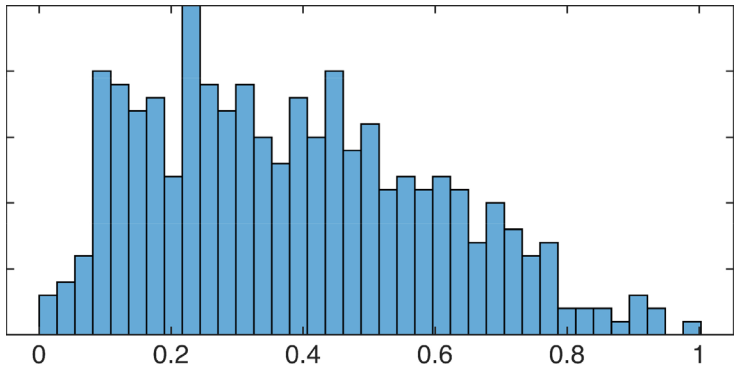
Scale a variable to have a values between 0 and 1

Min-Max Normalization:

$$x_{normalized} = \frac{x - x_{min}}{x_{max} - x_{min}}$$



Data distribution before normalized



Data distribution after normalized

Normalization and Standardization

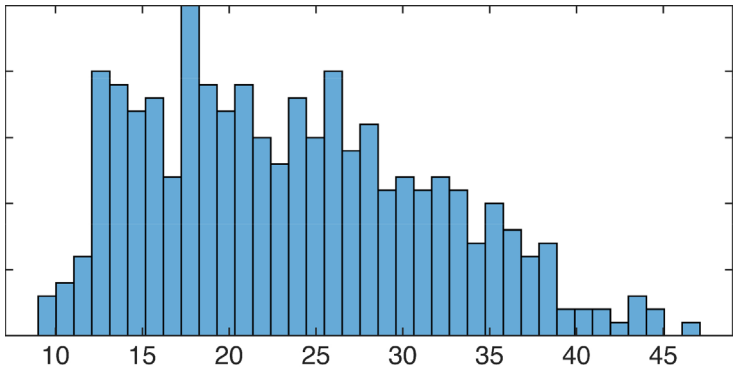
Preparing Data

Standardization

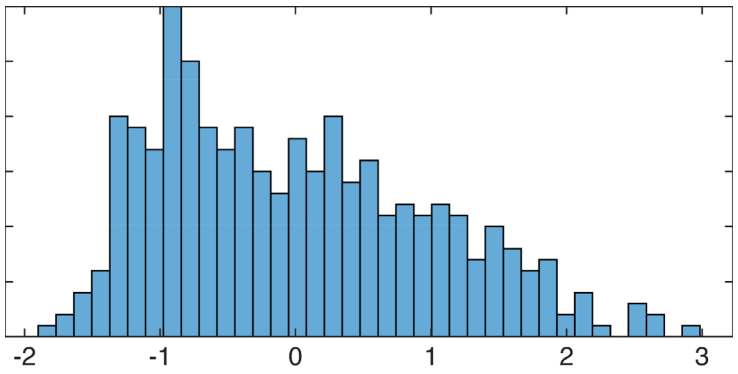
Transforms data to have a mean of zero and a standard deviation of 1.

Z-score Standardization

$$x_{standardized} = \frac{x - \bar{x}}{S.D.}$$



Data distribution before standardized



Data distribution after standardized

Data Quality

Preparing Data

Source: http://itsadeliverything.com/wordpress/images//accuracy-vs-precision.jpg

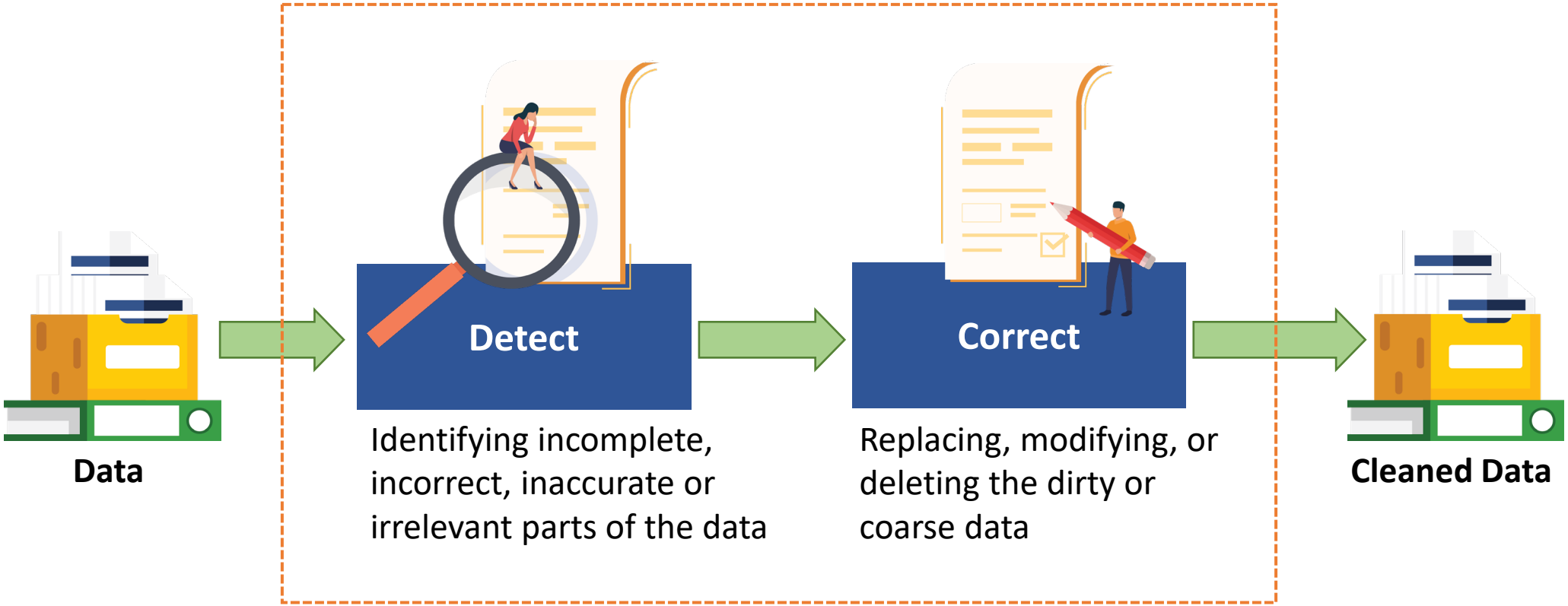
Data should be:

- **Accurate and Precise**
- **Complete** – Does not have "unknown" or "missing" values
- **Consistency** – Two data items in the data set contradict each other
- **Valid** – Conform to defined business rules or constraints
- **Uniform** – Using the same units of measure in all systems
- **Unique** – Does not contain duplicates

Data Cleaning

Preparing Data

Data Cleaning is the process of detecting and correcting/removing corrupt or inaccurate records from a record set



Inconsistent Datatypes

Preparing Data >> Data Cleaning

We expect that:

Values in a particular attribute must be of a particular datatype, e.g., Boolean, numeric (integer or real), date, etc.

Values in *align* and *alive* are inconsistent datatype

1 – Living Characters
0 – Deceased Characters

	name X_1	id X_2	align X_3	eye X_4	hair X_5	gender X_6	alive X_7	appearances X_8	first_appear X_9	publisher X_{10}
$\mathbf{x}_1$	Spider-Man (Peter Parker)	Secret	Good	Hazel Eyes	Brown Hair	Male	1	4043	Aug-62	marvel
$\mathbf{x}_2$	Captain America (Steven Rogers)	Public	Good	Blue Eyes	White Hair	Male	Living Characters	3360	Mar-41	marvel
	…	…	…	…	…	…	…	…	…	…
$\mathbf{x}_n$	Natalia Romanova (Earth-616)	Public	1	Green Eyes	Red Hair	Female	Living Characters	1050	Apr-64	marvel

1 – Good
0 – Bad

Inconsistent Datatypes

Preparing Data >> Data Cleaning

How to address the Inconsistent datatypes

- Choose an appropriate datatype
- Transform values in another datatype into the selected datatype

Values in *align* and *alive* are inconsistent datatype

1 – Living Characters
0 – Deceased Characters

	name X_1	id X_2	align X_3	eye X_4	hair X_5	gender X_6	alive X_7	appearances X_8	first_appear X_9	publisher X_{10}
$\mathbf{x}_1$	Spider-Man (Peter Parker)	Secret	Good	Hazel Eyes	Brown Hair	Male	Living Characters	4043	Aug-62	marvel
$\mathbf{x}_2$	Captain America (Steven Rogers)	Public	Good	Blue Eyes	White Hair	Male	Living Characters	3360	Mar-41	marvel
	…	…	…	…	…	…	…	…	…	…
$\mathbf{x}_n$	Natalia Romanova (Earth-616)	Public	Good	Green Eyes	Red Hair	Female	Living Characters	1050	Apr-64	marvel

1 – Good
0 – Bad

Missing Value

Preparing Data >> Data Cleaning

We expect that:

All required measures are known.

	IQ X_1	Job performance X_2
$\mathbf{x}_1$	78	NA
$\mathbf{x}_2$	84	NA
$\mathbf{x}_3$	84	NA
$\mathbf{x}_4$	85	NA
$\mathbf{x}_5$	99	7
$\mathbf{x}_6$	105	10
$\mathbf{x}_7$	105	11
$\mathbf{x}_8$	106	15
$\mathbf{x}_9$	108	10
$\mathbf{x}_{10}$	112	10
$\mathbf{x}_{11}$	113	12
$\mathbf{x}_{12}$	115	14
$\mathbf{x}_{13}$	118	16
$\mathbf{x}_{14}$	134	12

Job performances of $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$ and $\mathbf{x}_4$ are unknow. They are missing value.

Missing Value

Preparing Data >> Data Cleaning

How to deal with the missing value

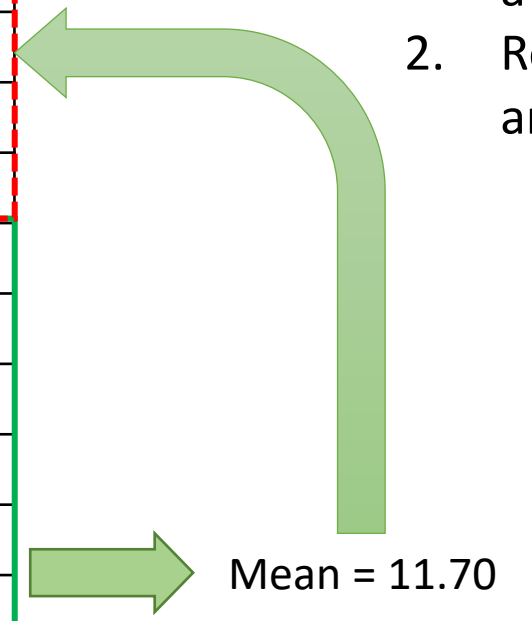
Single Imputation: Generate a single replacement value for each missing data point.

- **Arithmetic Mean Imputation**
  - replaces missing values with mean of available values
- **Regression Imputation**
  - replaces missing values with predicted scores from a regression equation
- **Hot-deck Imputation**
  - A collection of techniques that impute the missing values with scores from “similar” datapoints, such as nearest neighbor hot-deck and last observation carried forward.
- **and etc.**

Missing Value

Preparing Data >> Data Cleaning

	IQ X_1	Job performance X_2
$\mathbf{x}_1$	78	11.70
$\mathbf{x}_2$	84	11.70
$\mathbf{x}_3$	84	11.70
$\mathbf{x}_4$	85	11.70
$\mathbf{x}_5$	99	7
$\mathbf{x}_6$	105	10
$\mathbf{x}_7$	105	11
$\mathbf{x}_8$	106	15
$\mathbf{x}_9$	108	10
$\mathbf{x}_{10}$	112	10
$\mathbf{x}_{11}$	113	12
$\mathbf{x}_{12}$	115	14
$\mathbf{x}_{13}$	118	16
$\mathbf{x}_{14}$	134	12



Example of Arithmetic Mean Imputation

1. Compute the arithmetic mean of X_2 from available values
2. Replace the missing values of X_2 by the arithmetic mean

Missing Value

Preparing Data >> Data Cleaning

	IQ X_1	Job performance X_2
$\mathbf{x}_1$	78	7.529
$\mathbf{x}_2$	84	8.267
$\mathbf{x}_3$	84	8.267
$\mathbf{x}_4$	85	8.390
$\mathbf{x}_5$	99	7
$\mathbf{x}_6$	105	10
$\mathbf{x}_7$	105	11
$\mathbf{x}_8$	106	15
$\mathbf{x}_9$	108	10
$\mathbf{x}_{10}$	112	10
$\mathbf{x}_{11}$	113	12
$\mathbf{x}_{12}$	115	14
$\mathbf{x}_{13}$	118	16
$\mathbf{x}_{14}$	134	12

$$JP = 0.123(78) + (-2.065) = 7.529$$
$$JP = 0.123(84) + (-2.065) = 8.267$$
$$JP = 0.123(84) + (-2.065) = 8.267$$
$$JP = 0.123(85) + (-2.065) = 8.390$$

$$JP = \beta_1(IQ) + \beta_0 = 0.123(IQ) + (-2.065)$$

incomplete variables

complete variables

Example of Regression Imputation

1. Estimate a set of regression equations
2. Generate predicted values for the incomplete variables
3. Fill in the missing values

Duplicate Data

Preparing Data >> Data Cleaning

We expect that:

A data should appear on the dataset one time

	name X_1	id X_2	align X_3	eye X_4	hair X_5	gender X_6	alive X_7	appearances X_8	first_appear X_9	publisher X_{10}
$\mathbf{x}_1$	Spider-Man (Peter Parker)	Secret	Good	Hazel Eyes	Brown Hair	Male	Living Characters	4043	Aug-62	marvel
$\mathbf{x}_2$	Captain America (Steven Rogers)	Public	Good	Blue Eyes	White Hair	Male	Living Characters	3360	Mar-41	marvel
$\mathbf{x}_3$	Spider-Man (Peter Parker)	Secret	Good	Hazel Eyes	Black Hair	Male	Living Characters	NA	Aug-62	marvel
	…	…	…	…	…	…	…	…	…	…
$\mathbf{x}_{100}$	Natalia Romanova (Earth-616)	Public	Good	Green Eyes	Red Hair	Female	Living Characters	1050	Apr-64	marvel

We have two recodes of Spider-Man. So, the two recodes are duplicate data

Moreover, one contradicts each other

Duplicate Data

Preparing Data >> Data Cleaning

How to deal with the duplicate data

1. Select one recode that is up-to-date and accurate
2. Remove the others

	name X_1	id X_2	align X_3	eye X_4	hair X_5	gender X_6	alive X_7	appearances X_8	first_appear X_9	publisher X_{10}
$\mathbf{x}_1$	Spider-Man (Peter Parker)	Secret	Good	Hazel Eyes	Brown Hair	Male	Living Characters	4043	Aug-62	marvel
$\mathbf{x}_2$	Captain America (Steven Rogers)	Public	Good	Blue Eyes	White Hair	Male	Living Characters	3360	Mar-41	marvel
	...	...	...	...	...	...	...	...	...	...
$\mathbf{x}_{100}$	Natalia Romanova (Earth-616)	Public	Good	Green Eyes	Red Hair	Female	Living Characters	1050	Apr-64	marvel

We have two recodes of Spider-Man. So, the two recodes are duplicate data

Descriptive Statistics with Pivot Tables

Outline

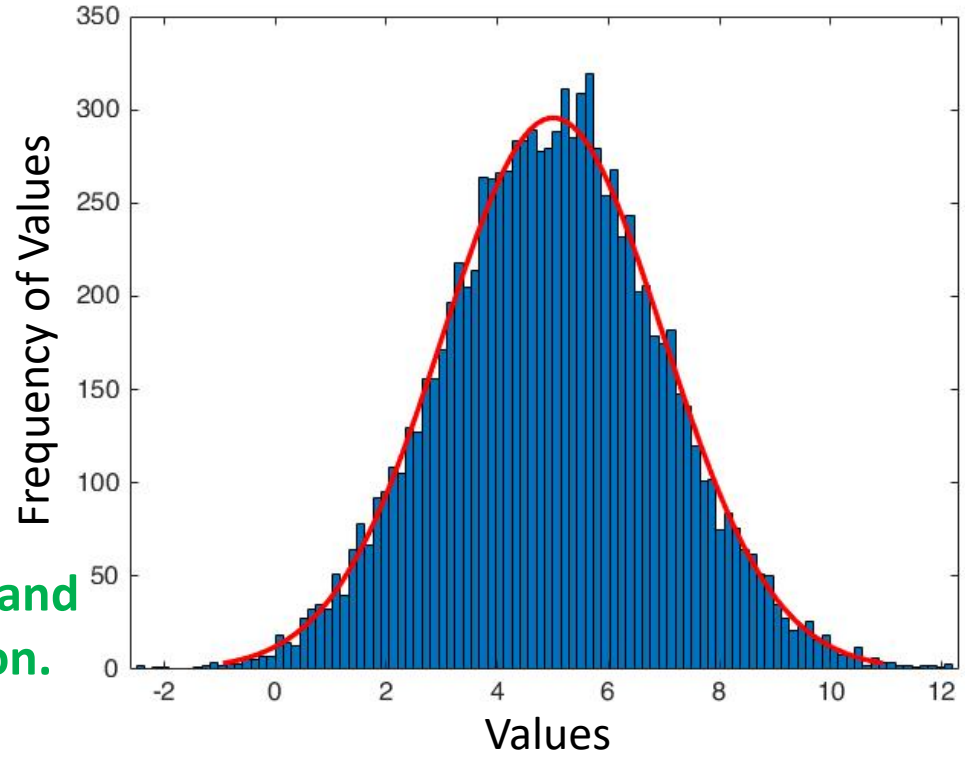
Descriptive Analysis

1. **Descriptive Statistics with Pivot Tables**
    - **Mean, Median and Mode**
    - **Variance and Standard Deviation**
    - **Skewness and Kurtosis**
    - **Covariance Matrix**
2. **Cluster Analysis**
    - **Distances**
    - **K-means Clustering**
    - **Hierarchical Clustering**
    - **Density-based Spatial Clustering**
3. **Association Analysis**
    - **Itemset Mining**
    - **Association Rules**

Mean, Median and Mode

Descriptive Statistics with Pivot Tables

	X_1	X_2	...	X_{10}
$\mathbf{x}_1$				
...				
$\mathbf{x}_n$				



We can slice a feature/variable and describe it as a data distribution.

A distribution in statistics is a function that shows:

- the possible values for a variable (x-axis)
- how often they occur (y-axis).

Mean, Median and Mode

Descriptive Statistics with Pivot Tables

Mean

- A measure of a central or typical value for a probability distribution.
- The sum of all measurements divided by the number of observations in the data set.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Example:

Job performance: 7, 10, 11, 15, 10, 10, 12, 14, 16, 12

Mean of job performance:

$$\bar{x} = \frac{7+10+11+15+10+10+12+14+16+12}{10} = \frac{117}{10} = 11.7$$

Mean, Median and Mode

Descriptive Statistics with Pivot Tables

Median

- Reflect the central tendency of the sample in such a way that it is uninfluenced by extreme values or outliners.
- The middle value that separates the higher half from the lower half of the data set.
- To compute the middle value, we need to arrange all the numbers from smallest to greatest.
- Then,

$$\tilde{x} = \begin{cases} x_{\frac{(n+1)}{2}}, & if\ n\ is\ odd, \\ \frac{\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right)}{2}, & if\ n\ is\ even, \end{cases}$$

Example:

Job performance: 7, 10, 11, 15, 10, 10, 12, 14, 16, 12

Median of job performance:

n = 10. So, n is even

$$\tilde{x} = \frac{x_5 + x_6}{2} = \frac{11 + 12}{2} = 11.5$$

11.5

7	10	10	10	11	12	12	14	15	16
				x_5	x_6				

Mean, Median and Mode

Descriptive Statistics with Pivot Tables

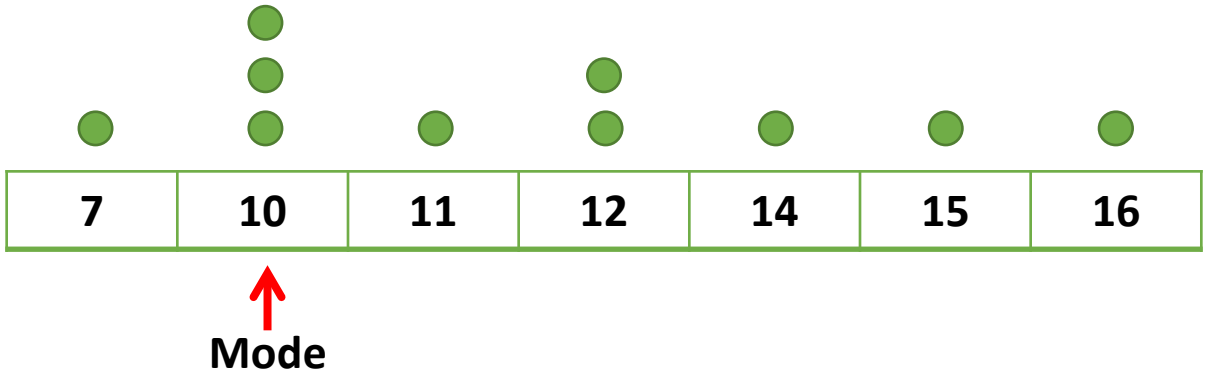
Mode

- The most frequent value in the data set.

Example:

Job performance: 7, 10, 11, 15, 10, 10, 12, 14, 16, 12

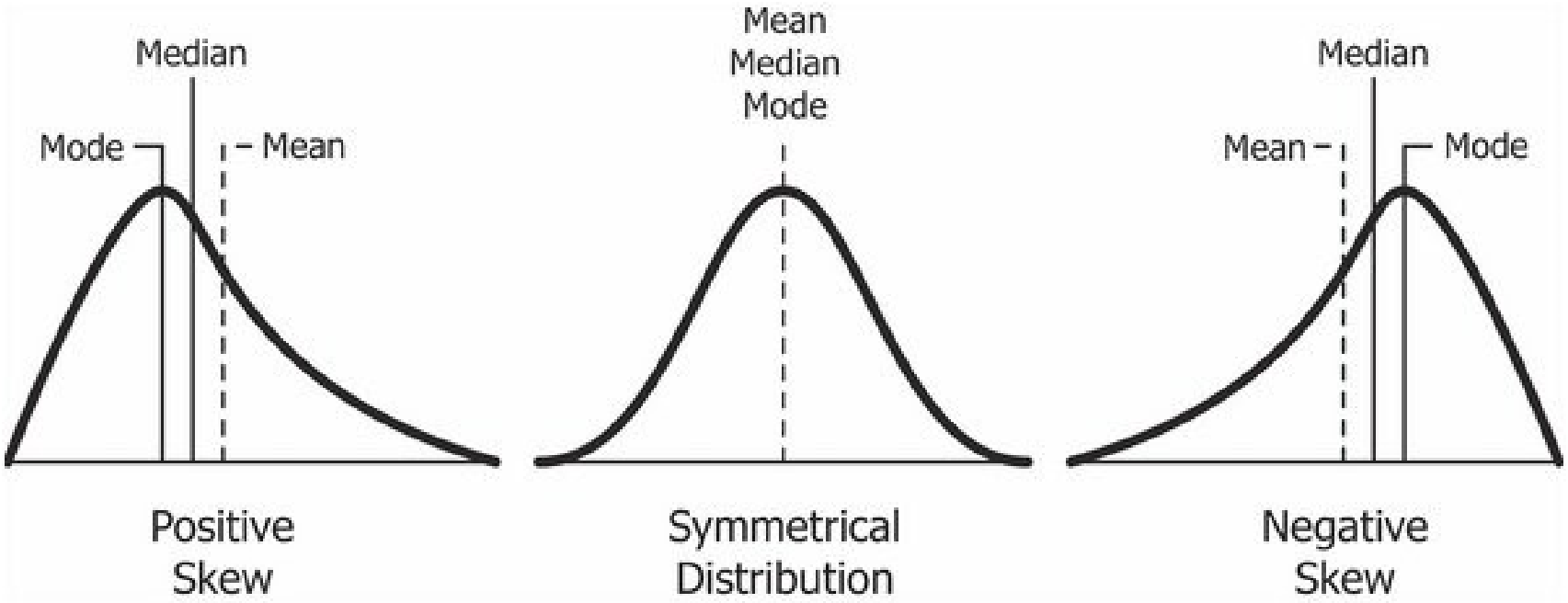
Mode of job performance:



Mean, Median and Mode

Descriptive Statistics with Pivot Tables

Geometric visualization of the mode, median and mean of an arbitrary probability density function



Source: https://codeburst.io/2-important-statistics-terms-you-need-to-know-in-data-science-skewness-and-kurtosis-388fef94eeaa

Mean, Median and Mode

Descriptive Statistics with Pivot Tables

Recall:

Provides	Categorical Attribute		Numerical Attribute	
	Nominal	Ordinal	Interval-scaled	Ratio-scaled
Mode	/	/	/	/
Median		/	/	/
Mean			/	/

Mean, Median and Mode

Descriptive Statistics with Pivot Tables

	IQ X_1	Job performance X_2
$\mathbf{x}_1$	99	7
$\mathbf{x}_2$	105	10
$\mathbf{x}_3$	105	11
$\mathbf{x}_4$	106	15
$\mathbf{x}_5$	108	10
$\mathbf{x}_6$	112	10
$\mathbf{x}_7$	113	12
$\mathbf{x}_8$	115	14
$\mathbf{x}_9$	118	16
$\mathbf{x}_{10}$	134	12
Mean		11.7
Median		11.5
Mode		10

Quiz:

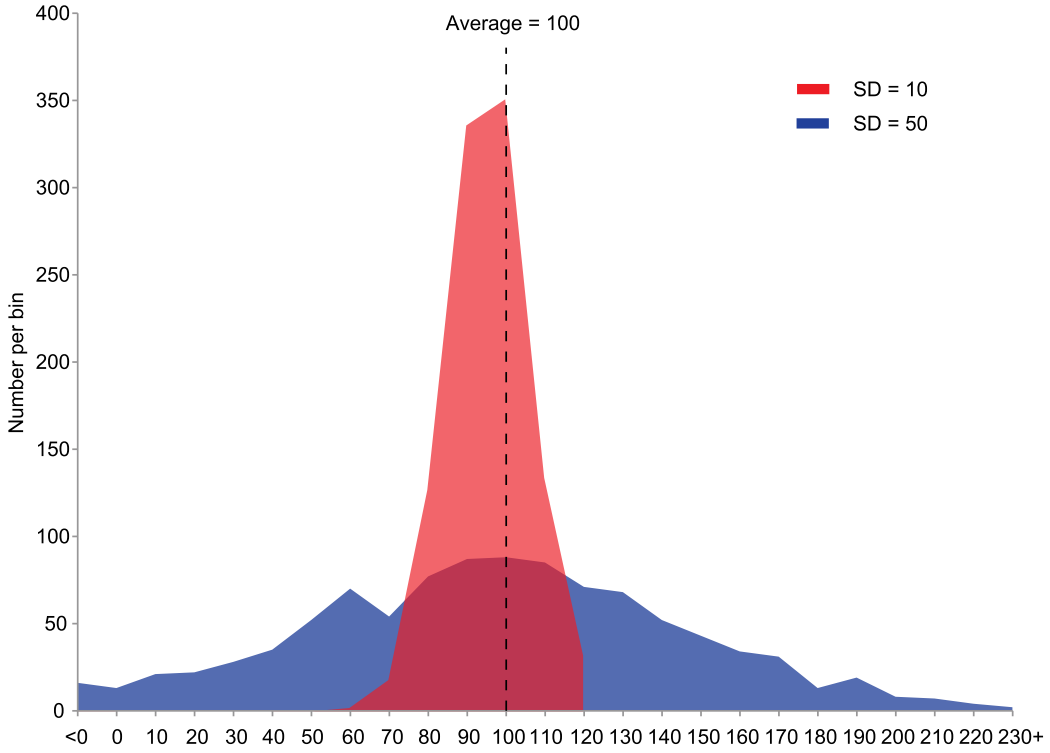
Find the mean, median and mode of IQ.

Variance and Standard Deviation

Descriptive Statistics with Pivot Tables

Standard Deviation (SD, s)

- A measure that is used to quantify the amount of variation or dispersion of a set of data values.
- A low standard deviation indicates that the data points tend to be close to the mean.
- A high standard deviation indicates that the data points are spread out over a wider range of values.



Source:
https://en.wikipedia.org/wiki/Standard_deviation#/media/File:Comparison_standard_deviations.svg

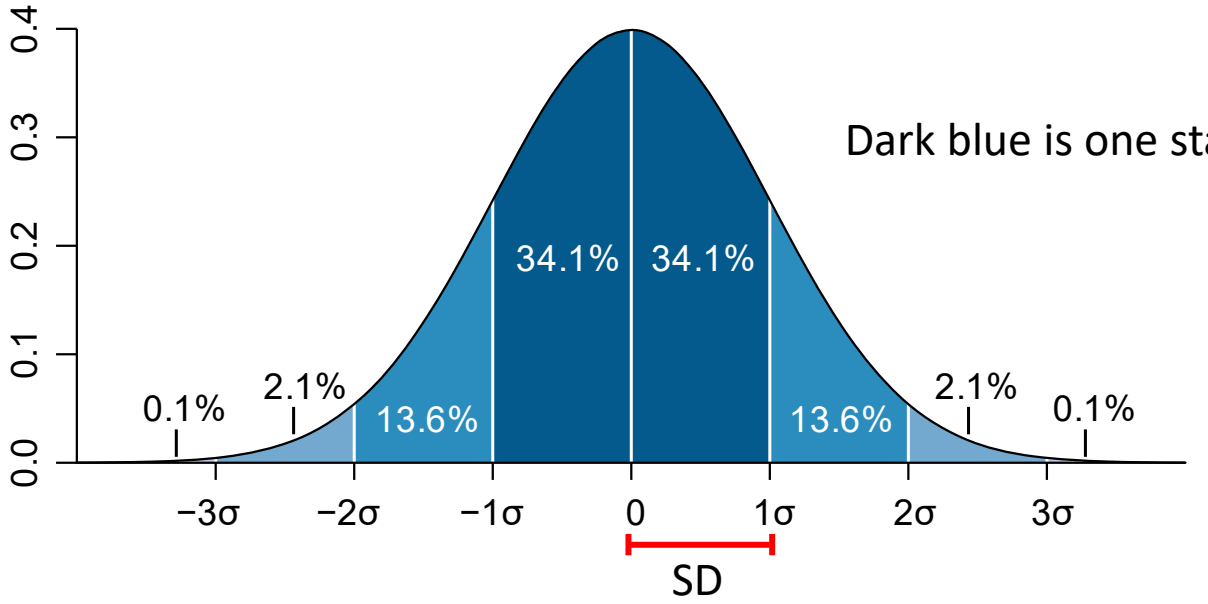
Variance and Standard Deviation

Descriptive Statistics with Pivot Tables

Standard Deviation (SD, s)

The formula for the sample standard deviation is

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$



Dark blue is one standard deviation on either side of the mean.

Source:
https://en.wikipedia.org/wiki/Standard_deviation#/media/File:Standard_deviation_diagram.svg

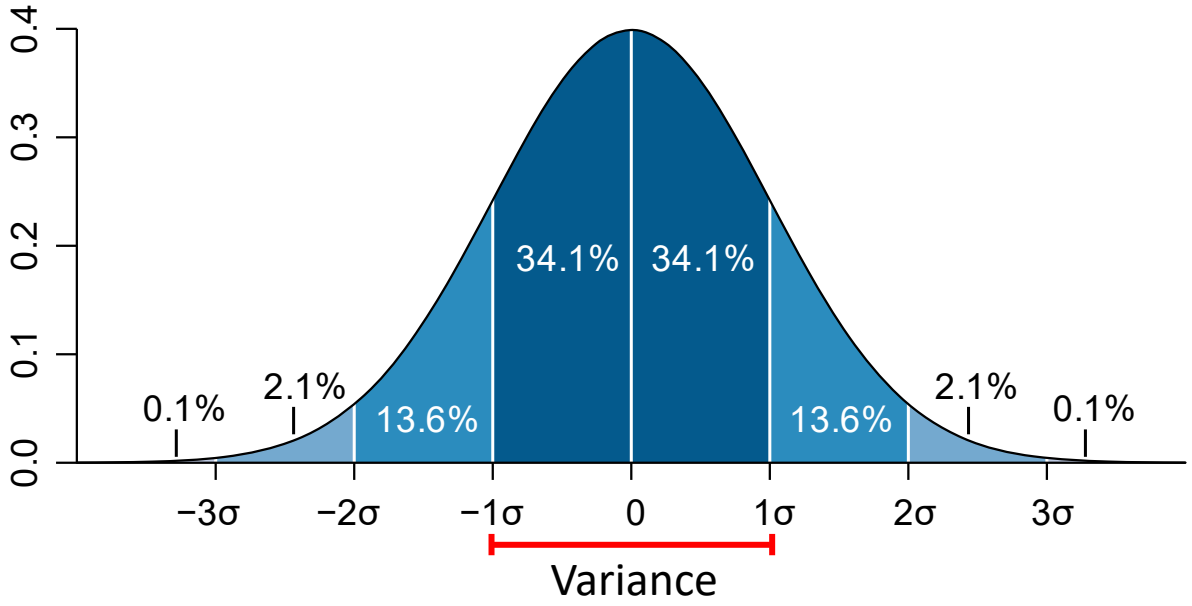
Variance and Standard Deviation

Descriptive Statistics with Pivot Tables

Variance (σ)

- How far a set of numbers are spread out from their average value.
- It is the square of the standard deviation

$$\mathrm{var}(X) = s^2 = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}^{\,2} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$



Source:
https://en.wikipedia.org/wiki/Standard_deviation#/media/File:Standard_deviation_diagram.svg

Variance and Standard Deviation

Descriptive Statistics with Pivot Tables

Example

- Job performance; X={7, 10, 11, 15, 10, 10, 12, 14, 16, 12}
- Mean of job performance $\bar{x}$: 11.7
- Standard Deviation; $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} = 2.71$
- Variance; $\mathrm{var}(X) = SD^2 = 2.71^2 = 7.34$

Job performance x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
7	-4.7	22.09
10	-1.7	2.89
11	-0.7	0.49
15	3.3	10.89
10	-1.7	2.89
10	-1.7	2.89
12	0.3	0.09
14	2.3	5.29
16	4.3	18.49
12	0.3	0.09
$\sum_{i=1}^{n}(x_i - \bar{x})^2$		66.1
$\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$		2.71

Variance and Standard Deviation

Descriptive Statistics with Pivot Tables

	IQ X_1	Job performance X_2
$\mathbf{x}_1$	99	7
$\mathbf{x}_2$	105	10
$\mathbf{x}_3$	105	11
$\mathbf{x}_4$	106	15
$\mathbf{x}_5$	108	10
$\mathbf{x}_6$	112	10
$\mathbf{x}_7$	113	12
$\mathbf{x}_8$	115	14
$\mathbf{x}_9$	118	16
$\mathbf{x}_{10}$	134	12
Mean	111.5	11.7
SD		2.71
Variance		7.34

$$\mathrm{var}(X) = s^2 = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}^{\,2} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Quiz:

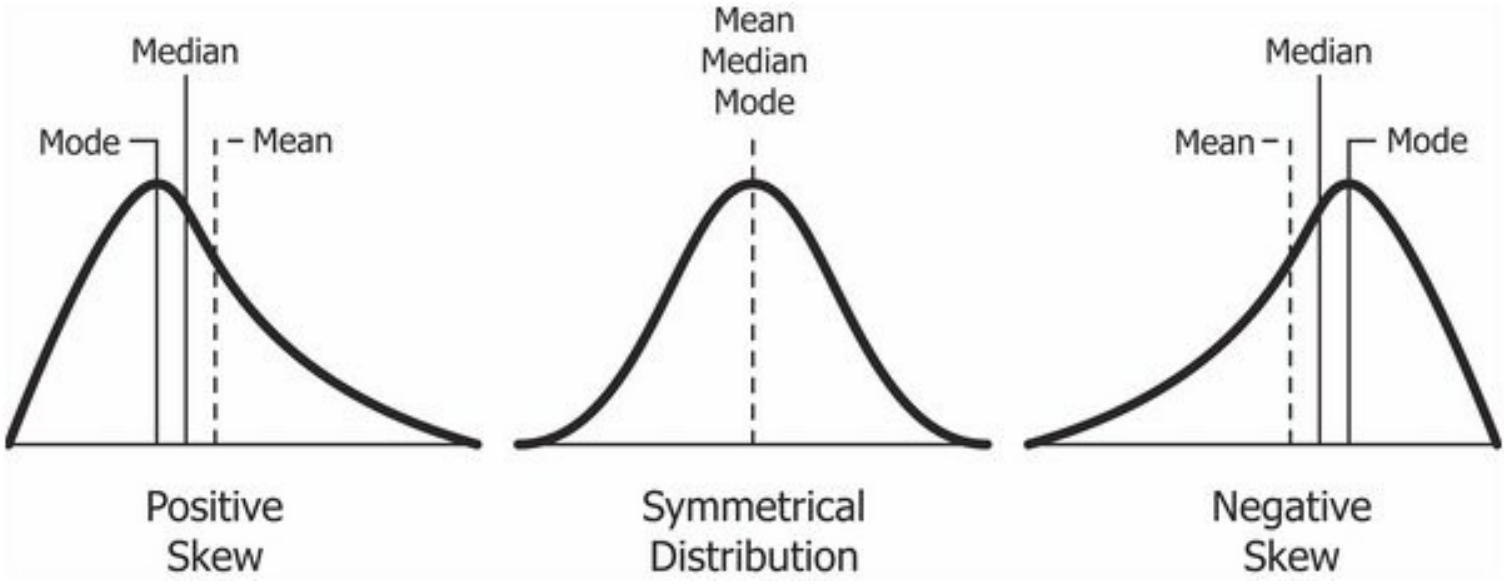
Find the SD and variance of IQ.

Skewness and Kurtosis

Descriptive Statistics with Pivot Tables

Skewness

- Skewness is usually described as a measure of a **dataset's symmetry** – or lack of symmetry.
- The normal distribution has a skewness of 0.



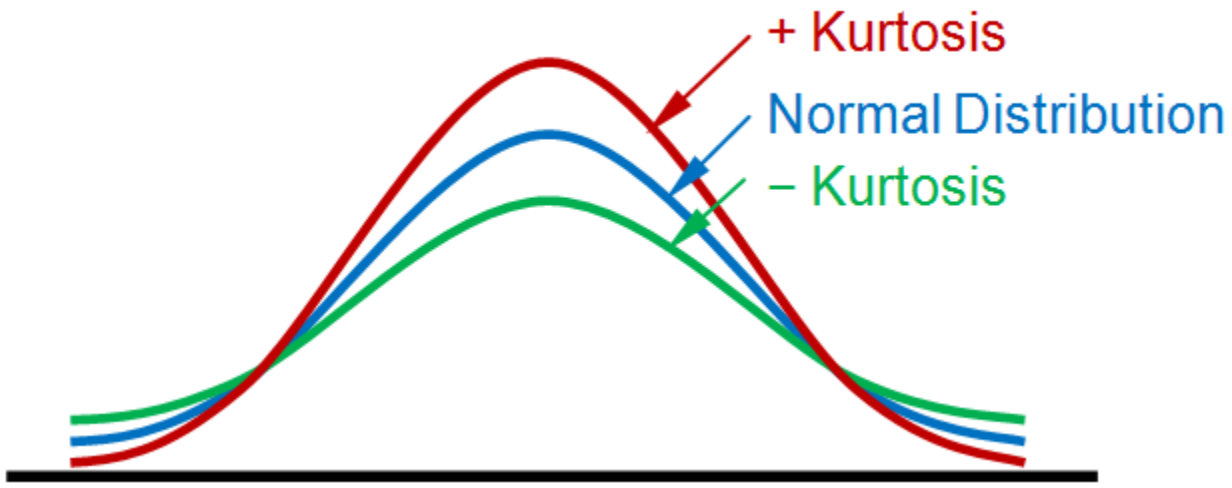
Source: https://codeburst.io/2-important-statistics-terms-you-need-to-know-in-data-science-skewness-and-kurtosis-388fef94eeaa

Skewness and Kurtosis

Descriptive Statistics with Pivot Tables

Kurtosis

- Measures the **tail-heaviness of the distribution**.
- The excess kurtosis for a standard normal distribution is 0.



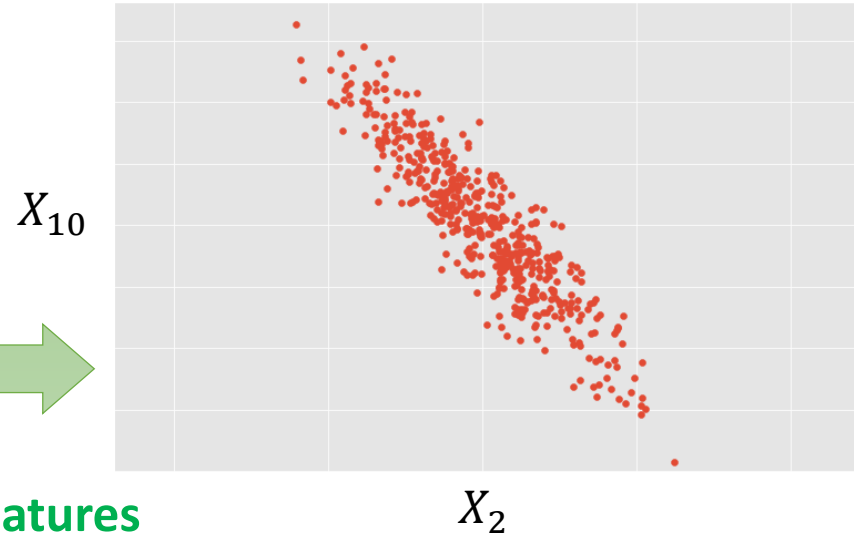
Source: https://www.statext.com/android/kurtosis.html

Covariance Matrix

Descriptive Statistics with Pivot Tables

	X_1	X_2	...	X_{10}
$\mathbf{x}_1$				
...				
$\mathbf{x}_n$				

The joint variability of two random variables can be described by **covariance**



We can slice any variables/features and display them as a scatter plot

Covariance Matrix

Descriptive Statistics with Pivot Tables

Covariance

- How much two random variables vary together.
- The covariance of random variables X and Y, denoted by $\mathrm{cov}(X, Y)$ can be computed by:

$$\mathrm{cov}(X, Y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

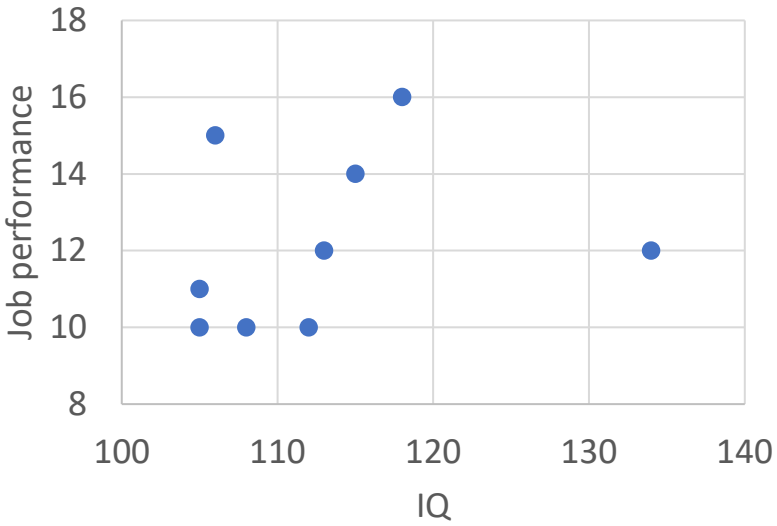
- The value of covariance lies between $-\infty$ and $+\infty$.

Covariance Matrix

Descriptive Statistics with Pivot Tables

Example

	IQ X	Job performance Y	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
$\mathbf{x}_1$	99	7	-12.5	-4.7	58.75
$\mathbf{x}_2$	105	10	-6.5	-1.7	11.05
$\mathbf{x}_3$	105	11	-6.5	-0.7	4.55
$\mathbf{x}_4$	106	15	-5.5	3.3	-18.15
$\mathbf{x}_5$	108	10	-3.5	-1.7	5.95
$\mathbf{x}_6$	112	10	0.5	-1.7	-0.85
$\mathbf{x}_7$	113	12	1.5	0.3	0.45
$\mathbf{x}_8$	115	14	3.5	2.3	8.05
$\mathbf{x}_9$	118	16	6.5	4.3	27.95
$\mathbf{x}_{10}$	134	12	22.5	0.3	6.75
Mean	111.5	11.7		SUM	104.5



$$\text{cov}(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

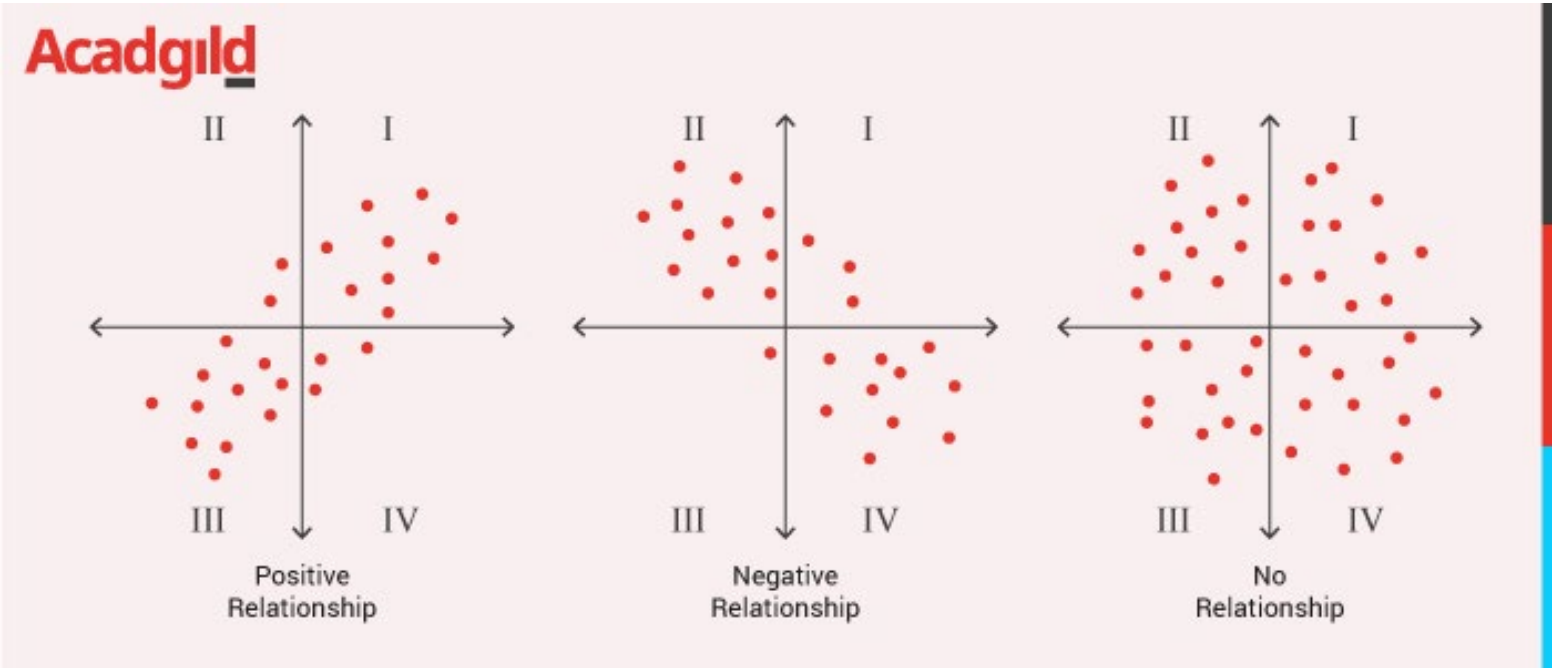
$$\text{cov}(X,Y) = \frac{104.5}{9} = 11.61$$

What dose it mean?

Covariance Matrix

Descriptive Statistics with Pivot Tables

Covariance



Source: https://acadgild.com/blog/covariance-and-correlation

A **positive covariance** means both variables tend to move upward or downward in value at the same time.

A **negative covariance** means the variables will move away from each other.

A **zero covariance** means there is no relationship.

Covariance Matrix

Descriptive Statistics with Pivot Tables

Correlation

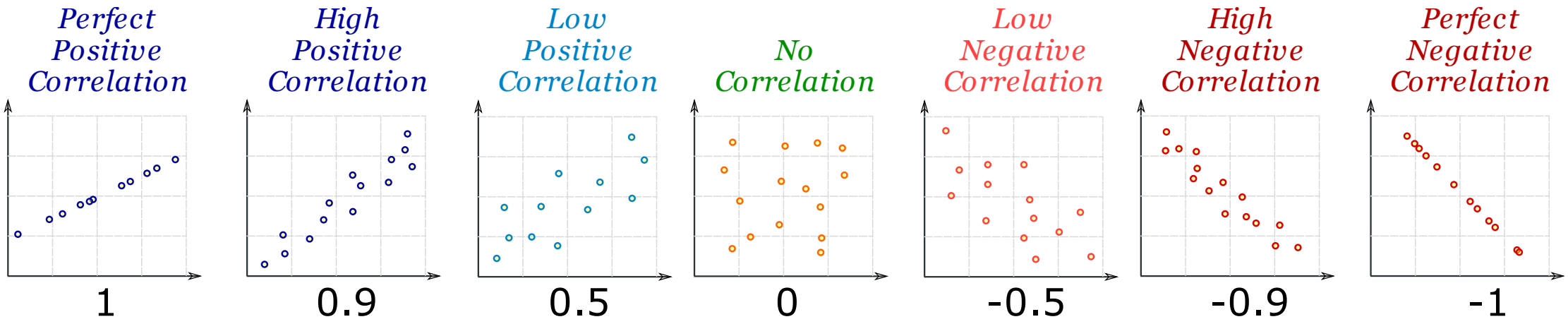
- Unit measure of change between two variables change with respect to each other.
- A normalized form of covariance.

$$\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{s_X s_Y}$$

- The value of correlation lies between -1 and $+1$.
  - If the correlation coefficient is one, it means that if one variable moves a given amount, the second moves proportionally in the same direction.
  - If correlation coefficient is zero, no relationship exists between the variables.
  - If correlation coefficient is –1, it means that one variable increases, the other variable decreases proportionally.

Covariance Matrix

Descriptive Statistics with Pivot Tables



The value of covariance lies between -1 and $+1$.

- If the correlation coefficient is one, it means that if one variable moves a given amount, the second moves proportionally in the same direction.
- If correlation coefficient is zero, no relationship exists between the variables.
- If correlation coefficient is –1, it means that one variable increases, the other variable decreases proportionally.

Covariance Matrix

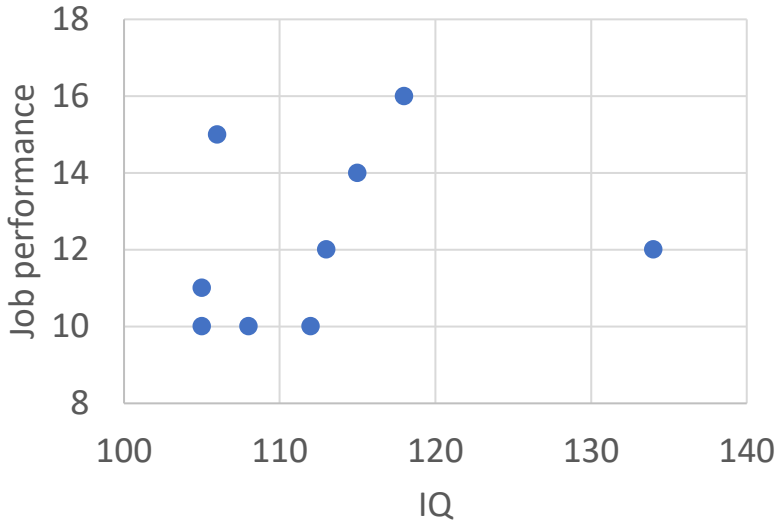
Descriptive Statistics with Pivot Tables

Example

	IQ X	Job performance Y
$\mathbf{x}_1$	99	7
$\mathbf{x}_2$	105	10
$\mathbf{x}_3$	105	11
$\mathbf{x}_4$	106	15
$\mathbf{x}_5$	108	10
$\mathbf{x}_6$	112	10
$\mathbf{x}_7$	113	12
$\mathbf{x}_8$	115	14
$\mathbf{x}_9$	118	16
$\mathbf{x}_{10}$	134	12
Mean	111.5	11.7
SD	9.70	2.71

$$\text{cov}(X, Y) = 11.61$$

$$\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{s_X s_Y} = \frac{11.61}{9.70 \times 2.71} = \frac{11.61}{26.287} = 0.44$$

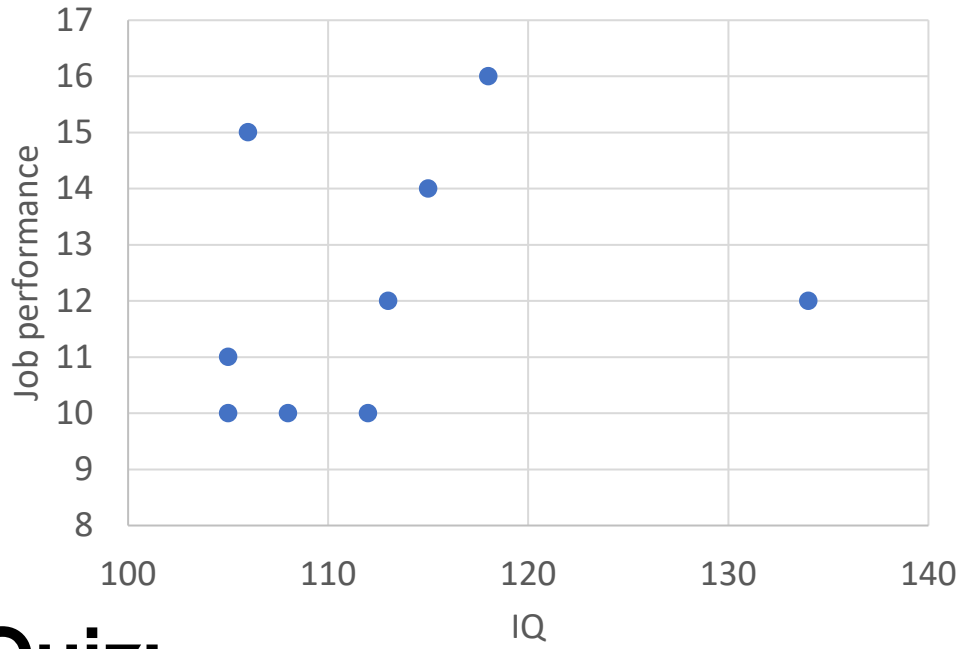


Covariance Matrix

Descriptive Statistics with Pivot Tables

Example

	IQ X	Job performance Y
$\mathbf{x}_1$	99	7
$\mathbf{x}_2$	105	10
$\mathbf{x}_3$	105	11
$\mathbf{x}_4$	106	15
$\mathbf{x}_5$	108	10
$\mathbf{x}_6$	112	10
$\mathbf{x}_7$	113	12
$\mathbf{x}_8$	115	14
$\mathbf{x}_9$	118	16
$\mathbf{x}_{10}$	134	12
Mean	111.5	11.7
SD	9.70	2.71



$$\text{cov}(X, Y) = 11.61$$

$$\text{corr}(X, Y) = 0.44$$

Quiz:

What do the covariance and correlation tell about the relation between IQ and job performance?

Covariance Matrix

Descriptive Statistics with Pivot Tables

Covariance Matrix

- A matrix whose element in the *i, j* position is the covariance between the *i*-th and *j*-th features.

	X_1	X_2	...	X_{10}
$\mathbf{x}_1$				
...				
$\mathbf{x}_n$				

Data Matrix

$$C = \begin{array}{c} \\ X_1 \\ X_2 \\ \\ X_{10} \end{array} \begin{array}{c} \begin{array}{cccc} X_1 & X_2 & & X_{10} \end{array} \\ \begin{bmatrix} \mathrm{cov}(X_1, X_1) & \mathrm{cov}(X_1, X_2) & \cdots & \mathrm{cov}(X_1, X_{10}) \\ \mathrm{cov}(X_2, X_1) & \mathrm{cov}(X_2, X_2) & \cdots & \mathrm{cov}(X_2, X_{10}) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(X_{10}, X_1) & \mathrm{cov}(X_{10}, X_2) & \cdots & \mathrm{cov}(X_{10}, X_{10}) \end{bmatrix} \end{array}$$

Covariance Matrix