21DS636- Statistical Modelling

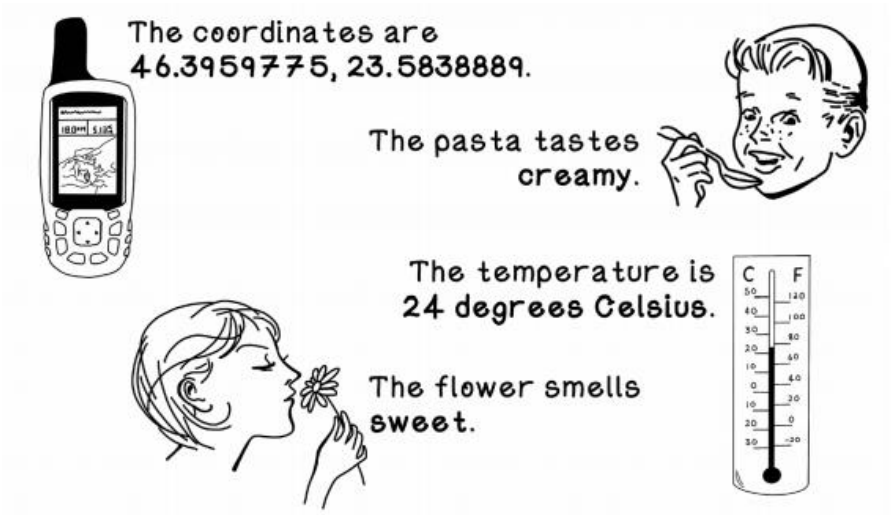
AMRITA VISHWA VIDYAPEETHAM
DEEMED TO BE UNIVERSITY

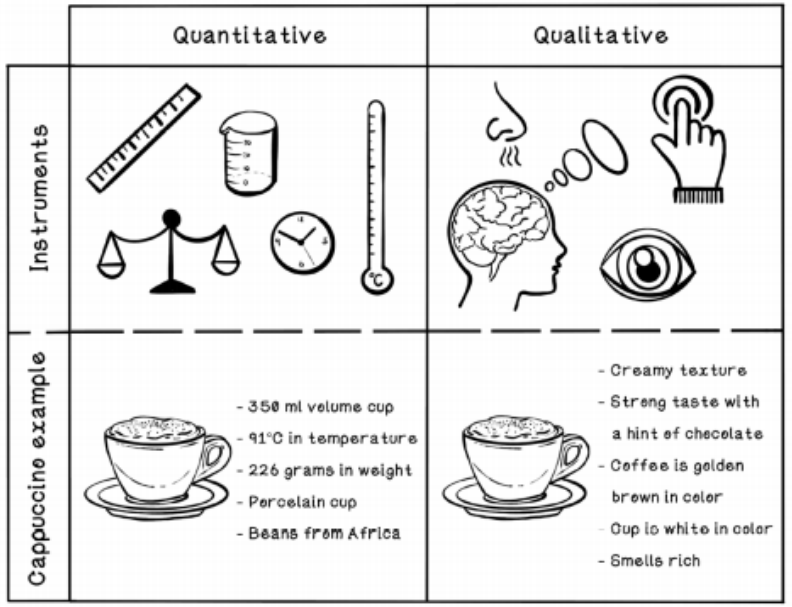
Exploratory Data Analysis

Dr. Deepa Gupta, Professor
Human Language Technology Lab
Dept. of computer Science and Engineering
Amrita School of Engineering, Bangalore

Data is Core to AI/ML/DM/DS Algorithms

- When we look at living organisms and machines, we see that the core element for operation is data.
  - Visuals that we see are data;
  - sounds that we hear are data;
  - measurements of the things around us are data.
- We consume data, process it all, and make decisions based on it;
- So, a fundamental understanding of the concept's surrounding data is important for understanding **AI/ML/DM/DS Algorithms**





What is Data?

- Data is different types of information usually formatted in a particular manner.

size of house (square feet)	# of bedrooms	price (1000$)
523	1	115
645	1	150
708	2	210
1034	3	280
2290	4	355
2545	4	440

image	label
	cat
	not cat
	cat
	not cat

How do you think about data? --Think of a spreadsheet

Data As you Know It

	A	B	C	D
1		Column 1	Column 2	Column 3
2	Row 1	2.2	2.3	1
3	Row 2	2.3	2.6	0
4	Row 3	2.1	2	1
5				

Column: All the data in one column will have the same scale and have meaning relative to each other.

Row: A row describes a single entity or observation and the columns describe properties about that entity or observation. The more rows you have, the more examples from the problem domain that you have.

Cell: A cell is a single value in a row and column. It may be a real value (1.5), an integer (2) or a category (red).

NOTE: we can call this type of data: tabular data. This form of data is easy to work with in AI/ML/DS.

Statistical Learning Preceptive

The statistical perspective frames data in the context of a **hypothetical function (f)** that the machine learning algorithm is trying to learn. That is, given some input variables (input), what is the predicted output variable (output).

Output = f(Input)

Those columns that are the inputs are referred to as input variables. you would like to predict for new input data in the future is called the output variable. It is also called the response variable.

Output Variable = f(Input Variables)
Output Variable = f(Input Vector)
Dependent Variable = f(Independent Variables)

	A	B	C
1	X1	X2	Y
2	2.2	2.3	1
3	2.3	2.6	0
4	2.1	2	1
5			

The standard shorthand used in the statistical perspective is to refer to the input variables as capital X and the output variables as capital Y . →Y = f(X)

There is a lot of overlap in the **computer science terminology** for data with the statistical perspective.

Output Attribute = Program(Input Attributes)

Computer Science perspective

	A	B	C	D
1		Attribute 1	Attribute 2	Output Attribute
2	Instance 1	2.2	2.3	1
3	Instance 2	2.3	2.6	0
4	Instance 3	2.1	2	1
5				

Output = Program(Input Features)

Acquiring Data

- **Manual labeling**


cat


not cat


cat

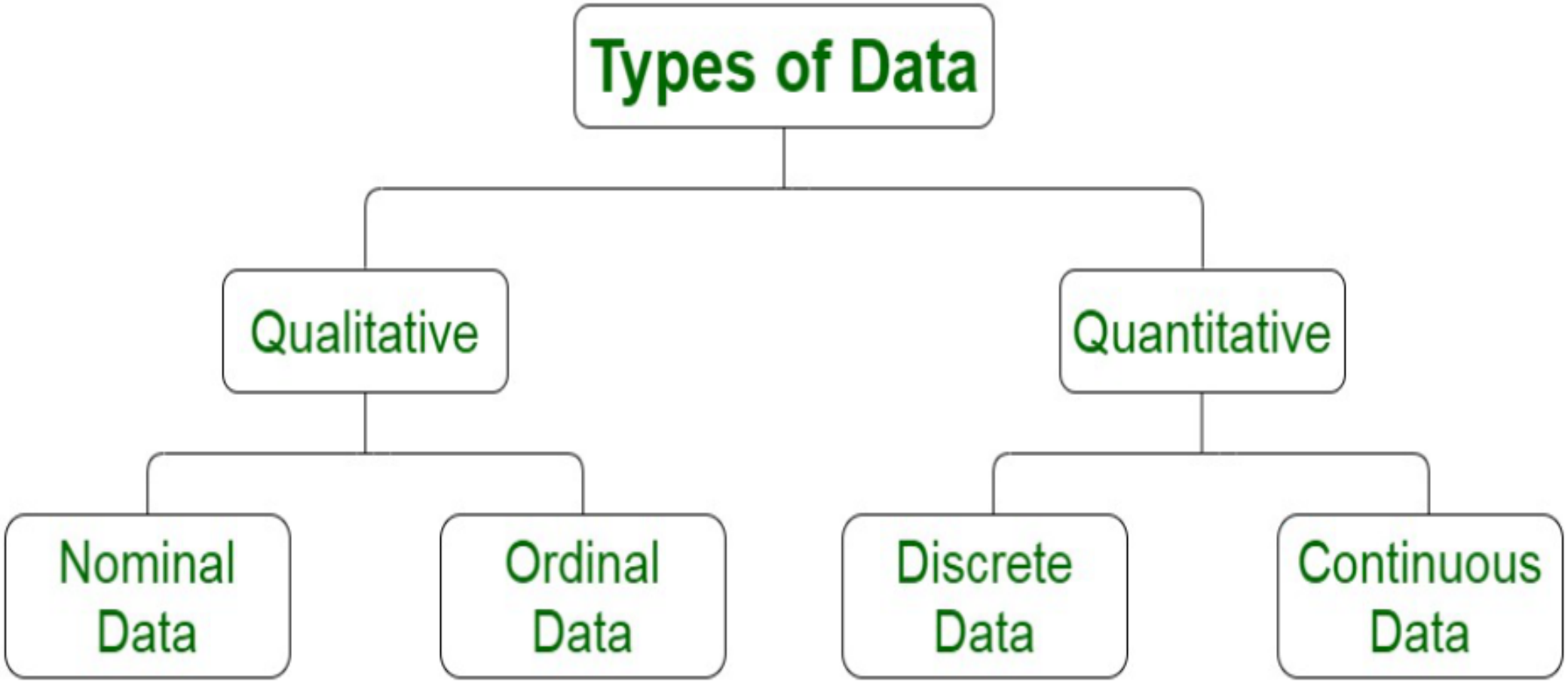

not cat

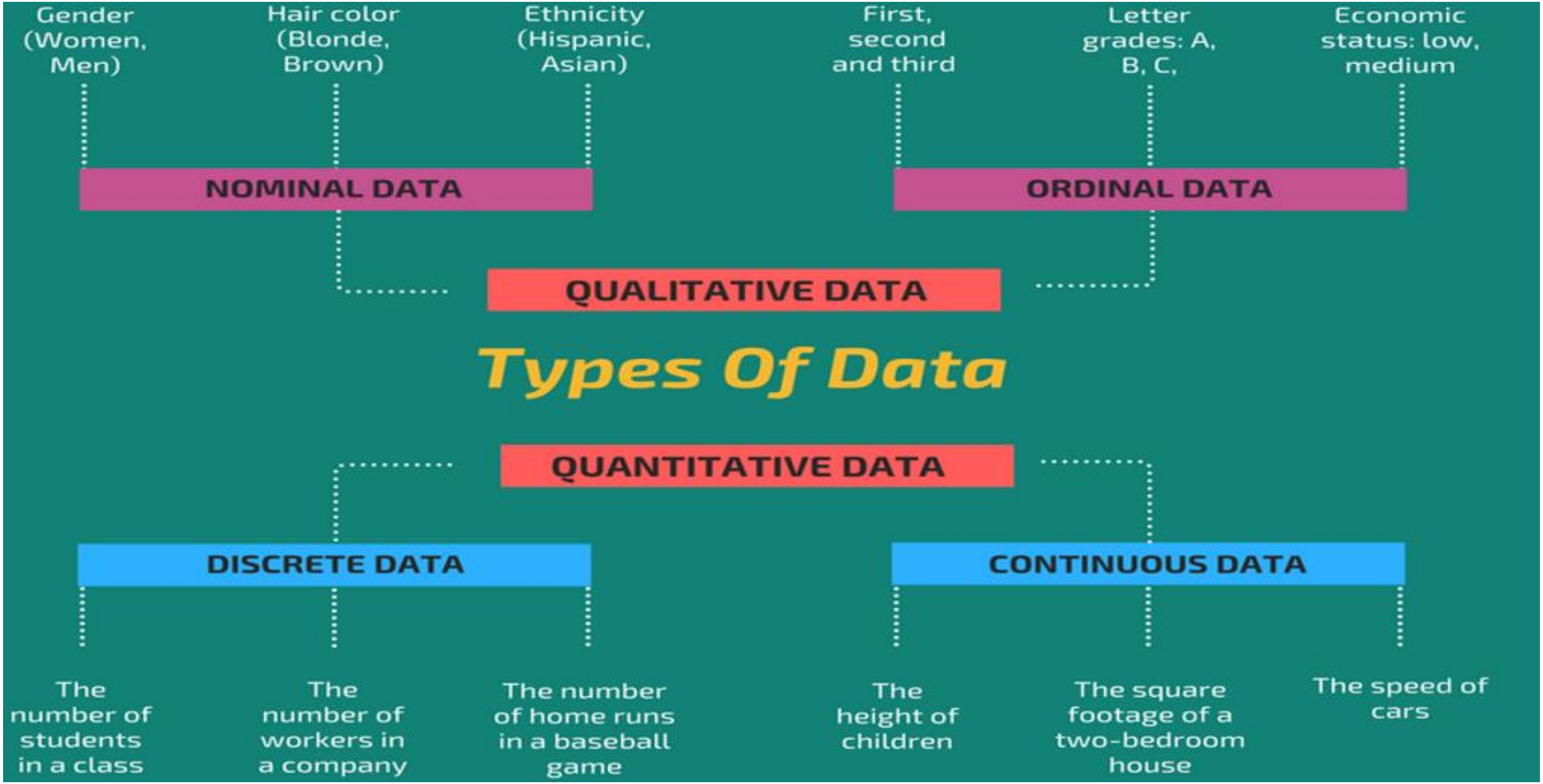
- **From observing behaviors**

user ID	time	price ($)	purchased
4783	Jan 21 08:15.20	7.95	yes
3893	March 3 11:30.15	10.00	yes
8384	June 11 14:15.05	9.50	no
0931	Aug 2 20:30.55	12.90	yes

machine	temperature (◦C)	pressure (psi)	machine fault
17987	60	7.65	N
34672	100	25.50	N
08542	140	75.50	Y
98536	165	125.00	Y

- **Download from websites/partnerships**


Types of Data
Qualitative
Quantitative
Nominal Data
Ordinal Data
Discrete Data
Continuous Data


Gender (Women, Men)
Hair color (Blonde, Brown)
Ethnicity (Hispanic, Asian)
First, second and third
Letter grades: A, B, C,
Economic status: low, medium
NOMINAL DATA
ORDINAL DATA
QUALITATIVE DATA
Types Of Data
QUANTITATIVE DATA
DISCRETE DATA
CONTINUOUS DATA
The number of students in a class
The number of workers in a company
The number of home runs in a baseball game
The height of children
The square footage of a two-bedroom house
The speed of cars

Key Terms for Data Types

Continuous

Data that can take on any value in an interval.

Synonyms

interval, float, numeric

Discrete

Data that can take on only integer values, such as counts.

Synonyms

integer, count

Categorical

Data that can take on only a specific set of values representing a set of possible categories.

Synonyms

enums, enumerated, factors, nominal, polychotomous

Binary

A special case of categorical data with just two categories of values (0/1, true/false).

Synonyms

dichotomous, logical, indicator, boolean

Ordinal

Categorical data that has an explicit ordering.

Synonyms

ordered factor

Why do we bother with a taxonomy of data types?

- It turns out that for the purposes of data analysis and predictive modeling, the data type is important to help determine the type of visual display, data analysis, or statistical model.
- In fact, data science software, such as R and Python, uses these data types to improve computational performance.
- More important, the data type for a variable determines how software will handle computations for that variable.
- Knowing that data is categorical can act as a signal telling software how statistical procedures, such as producing a chart or fitting a model, should behave.
- Storage and indexing can be optimized (as in a relational database).
- The possible values a given categorical variable can take are enforced in the software (like an enum).

Key Take Away

- Data is typically classified in software by type.
- Data types include continuous, discrete, categorical (which includes binary), and ordinal.
- Data typing in software acts as a signal to the software on how to process the data

How the actual data look Like…

Data Set

Input Attributes

Target Attribute

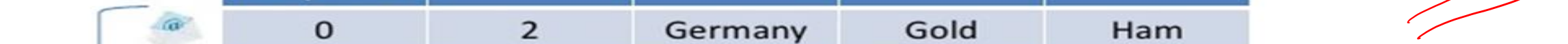
Number of new Recipients	Email Length (K)	Country (IP)	Customer Type	Email Type
0	2	Germany	Gold	Ham
1	4	Germany	Silver	Ham
5	2	Nigeria	Bronze	Spam
2	4	Russia	Bronze	Spam
3	4	Germany	Bronze	Ham
0	1	USA	Silver	Ham
4	2	USA	Silver	Spam

Instances

Numeric

Nominal

Ordinal



More Examples

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

More Examples

Play golf dataset

Independent variables				Dep. var
OUTLOOK	TEMPERATURE	HUMIDITY	WINDY	PLAY
sunny	85	85	FALSE	Don't Play
sunny	80	90	TRUE	Don't Play
overcast	83	78	FALSE	Play
rain	70	96	FALSE	Play
rain	68	80	FALSE	Play
rain	65	70	TRUE	Don't Play
overcast	64	65	TRUE	Play
sunny	72	95	FALSE	Don't Play
sunny	69	70	FALSE	Play
rain	75	80	FALSE	Play
sunny	75	70	TRUE	Play
overcast	72	90	TRUE	Play
overcast	81	75	FALSE	Play
rain	71	80	TRUE	Don't Play

More Examples

An example: data (loan application)

Approved or not

ID	Age	Has_Job	Own_House	Credit_Rating	Class
1	young	false	false	fair	No
2	young	false	false	good	No
3	young	true	false	good	Yes
4	young	true	true	fair	Yes
5	young	false	false	fair	No
6	middle	false	false	fair	No
7	middle	false	false	good	No
8	middle	true	true	good	Yes
9	middle	false	true	excellent	Yes
10	middle	false	true	excellent	Yes
11	old	false	true	excellent	Yes
12	old	false	true	good	Yes
13	old	true	false	good	Yes
14	old	true	false	excellent	Yes
15	old	false	false	fair	No

More Examples

No.	Experience	Test Score	Salary
1	4	78	24
2	7	100	43
3	1	86	23.7
4	5	82	34.3
5	8	86	35.8
6	10	84	38
7	0	75	22.2
8	1	80	23.1
9	6	83	30
10	6	91	33

No	Experience	Test Score	Salary
11	9	88	38
12	2	73	26.6
13	10	75	36.2
14	5	81	31.6
15	6	74	29
16	8	87	34
17	4	79	30.1
18	6	94	33.9
19	3	70	28.2
20	3	89	30

Table 2: Percentage of different crimes to overall crime on women in India

Year	Rape	Kidnapping and Abduction	Dowry Deaths	Cruelty By husband and relatives	Molestation	Eve teasing	Importation of girls	Sati Prevention Act	Immoral Traffic	Indecent Representation of women	Dowry Prohibition Act
2000	11.67	9.87	4.86	32.27	23.30	7.80	0.05	0.00	6.73	0.47	2.03
2001	11.18	10.18	4.76	34.19	23.73	6.78	0.08	0.00	6.12	0.73	2.24
2002	11.09	9.82	4.62	33.34	22.98	6.88	0.05	0.00	7.61	1.70	1.91
2003	11.27	9.46	4.42	36.06	23.43	8.77	0.03	0.00	3.92	0.74	1.91
2004	11.81	10.09	4.55	37.66	22.40	6.48	0.06	0.00	3.72	0.89	2.33
2005	11.80	10.13	4.36	37.49	21.97	6.42	0.10	0.00	3.80	1.88	2.06
2006	11.74	10.57	4.62	38.31	22.22	6.05	0.04	0.00	2.76	0.95	2.73
2007	11.19	11.02	4.37	40.97	20.90	5.91	0.03	0.00	1.93	0.65	3.03
2008	10.96	11.71	4.17	41.53	20.63	6.24	0.03	0.00	1.36	0.52	2.84
2009	10.50	12.63	4.11	43.94	18.99	5.40	0.02	0.00	1.21	0.41	2.77
2010	10.38	13.95	3.93	44.03	19.01	4.66	0.02	0.00	1.17	0.42	2.43
2011	10.59	15.55	3.77	43.36	18.79	3.75	0.03	0.00	1.06	0.20	2.89
2012	10.20	15.66	3.37	43.61	18.57	3.76	0.02	0.00	1.05	0.06	3.70
2013	10.89	16.76	2.61	38.40	22.85	4.07	0.01	0.00	0.83	0.12	3.46
2014	10.87	16.96	2.50	36.36	24.34	2.88	0.00	0.00	0.61	0.01	2.97
2015	10.58	18.11	2.33	34.64	25.18	2.65	0.00	0	0.74	0.01	3.02
Average	11.05	12.65	3.96	38.51	21.83	5.53	0.04	0.00	2.79	0.61	2.65

Note: Percentage of each crime is calculated by dividing the incident of each crime to the total number of incidence of a particular year.

Source: The above data is compiled into a table by the author ,by collecting the data from NCRB,

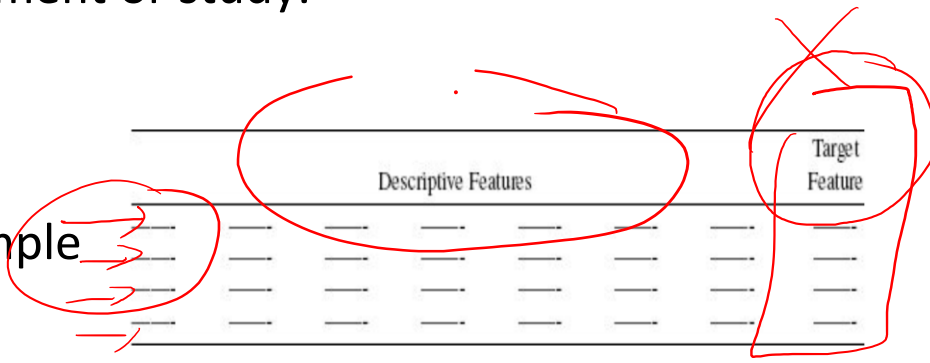
More Examples

Earthquakes by Magnitude

Obs	Latitude	Longitude	Depth	Magnitude	dNearestStation	RootMeanSquareTime	Type
1	36.8408	-97.875	4.2510	2.5	.	0.4600	earthquake
2	36.0210	-97.097	4.0420	2.5	.	0.6300	earthquake
3	36.7038	-98.041	12.7600	2.5	0.20300	0.2500	earthquake
4	36.7554	-97.556	6.2220	2.5	.	0.7600	earthquake
5	41.9035	-119.662	1.7742	2.5	0.47700	0.2545	earthquake
6	36.3510	-84.995	20.7000	2.5	0.34136	0.2000	earthquake
7	35.7968	-97.444	5.4500	2.5	.	0.3500	earthquake
8	36.7484	-97.649	3.3550	2.5	.	0.5500	earthquake
9	41.8692	-119.615	5.0000	2.5	0.65577	0.6500	earthquake
10	41.8652	-119.673	0.2000	2.5	0.46712	0.4500	earthquake

Rectangular Data/ Analytics Base Table (ABT)

- The typical frame of reference for an analysis in data science is a rectangular data object, like a spreadsheet or database table.
- **Data frame:** Rectangular data (like a spreadsheet) is the basic data structure for statistical and machine learning models.
- **Feature:** column in the table is commonly referred to as a feature.
  - **Synonyms:** attribute, input, predictor, variable
- **Outcome:** Many data science projects involve predicting an outcome—often a yes/no out- come. The features are some- times used to predict the outcome in an experiment or study.
  - **Synonyms:** dependent variable, response, target, output
- **Records:** A row in the table is commonly referred to as a record.
  - **Synonyms:** case, example, instance, observation, pattern, sample



Data Exploration

- Before attempting to build predictive models based on an ABT/Rectangular Data, it is important that we undertake some exploratory analysis, or data exploration, of the data contained in the ABT/ Rectangular Data

- Data exploration is a key part of both the **Data Understanding** and, **Data Preparation.**

Data Exploration-Goals

- **Fully understand the characteristics of the data in the ABT**
  - types of values a feature can take
  - the ranges into which the values in a feature fall
  - how the values in a dataset for a feature are **distributed** across the range.

- **Data quality issues**
  - missing values
  - instance has an extremely high value for a feature
  - invalid data
  - perfectly valid data that may cause difficulty to some machine learning techniques.

- The most important tool used during data exploration is the **data quality report**.

Data Quality Report

- The data quality report is the most important tool of the data exploration process.

- A data quality report includes tabular reports (one for continuous features and one for categorical features) that describe the characteristics of each feature in an ABT using standard statistical measures of central tendency (mean, mode, and median) and variation (standard deviation and percentiles).

- The tabular reports are accompanied by data visualizations (bar plots, histograms, and box plots) that illustrate the distribution of the values in each feature in an ABT.

Data Quality Report - Continuous Features

- The table in a data quality report that describes continuous features should include:
  - ✓ minimum,
  - ✓ 1st quartile,
  - ✓ mean,
  - ✓ median,
  - ✓ 3rd quartile,
  - ✓ maximum,
  - ✓ standard deviation
  - ✓ the percentage of instances in the ABT that are missing a value for each feature
  - ✓ the cardinality of each feature, (cardinality measures the number of distinct values present in the ABT for a feature).

(a) Continuous Features

Feature	Count	% Miss.	Card.	Min.	1st Qrt.	Mean	Median	3rd Qrt.	Max.	Std. Dev.
—	—	—	—	—	—	—	—	—	—	—
—	—	—	—	—	—	—	—	—	—	—
—	—	—	—	—	—	—	—	—	—	—

Data Quality Report - Categorical Features

- The table in the data quality report that describes categorical features should include:

  - ✓ A row for each feature in the ABT that contains the two most frequent levels for the feature **(the mode and 2nd mode)**

  - ✓ The frequency with which these appear (both as **raw frequencies** and as a **proportion of the total number of instances** in the dataset).

  - ✓ missing a value for the feature

  - ✓ The cardinality of the feature.

(b) Categorical Features

Feature	Count	% Miss.	Card.	Mode	Mode Freq.	Mode %	2nd Mode	2nd Mode Freq.	2nd Mode %
—	—	—	—	—	—	—	—	—	—
—	—	—	—	—	—	—	—	—	—
—	—	—	—	—	—	—	—	—	—

Data Quality Report - continuous features & categorical features

- ✓The data quality report should also include a histogram for each continuous feature in an ABT.
- ✓ For continuous features with cardinality less than 10, we use bar plots instead of histograms as this usually produces more informative data visualization.
- ✓For each categorical feature in an ABT, a bar plot should be included in the data quality report

Descriptive Statistics

Central Tendency

- The arithmetic mean is the best-known measure of central tendency.
- The **arithmetic mean** (or **sample mean** or just **mean**) of a set of n values for a feature a, a_1, a_2 . . . a_n, is denoted by the symbol $\bar{a}$, and is calculated as

$$\bar{a} = \frac{1}{n}\sum_{i=1}^{n} a_i$$

- Mean is commonly used as part of the data exploration process as a good estimate of the central tendencies of features in an ABT

Example

ID	1	2	3	4	5	6	7	8
Height	150	163	145	140	157	151	140	149

Figure: The members of a school basketball squad. The dashed grey line shows the arithmetic mean of the players' heights.

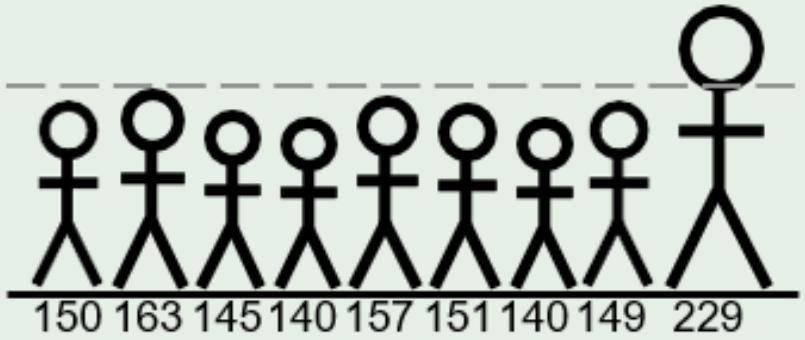
$$\overline{\text{HEIGHT}} = \frac{1}{8} \times (150 + 163 + 145 + 140 + 157 + 151 + 140 + 149)$$
$$= 149.375$$

1. The **arithmetic mean** is one measure of the **central tendency** of a **sample** (for our purposes a sample is just a set of values for a feature in an ABT).

1. Any measure of **central tendency** is, however, just an approximation.

Example

- Suppose our basketball squad manage to sign a *ringer* measuring in at 229cm



- The arithmetic mean for the full group is 158.235cm and no longer represents the central tendency of the group.
- An unusually large or small value like this is referred to as an **outlier** - the arithmetic mean is very sensitive to outliers.

Contd..

- There are other statistics that we can use to measure central tendency that are not as sensitive to outliers.

- The **median** of a set of values can be calculated by ordering the values from lowest to highest and selecting the middle value.

  - If there is an even number of values in the sample, then the median is obtained by calculating the arithmetic mean of the middle two values.

Example

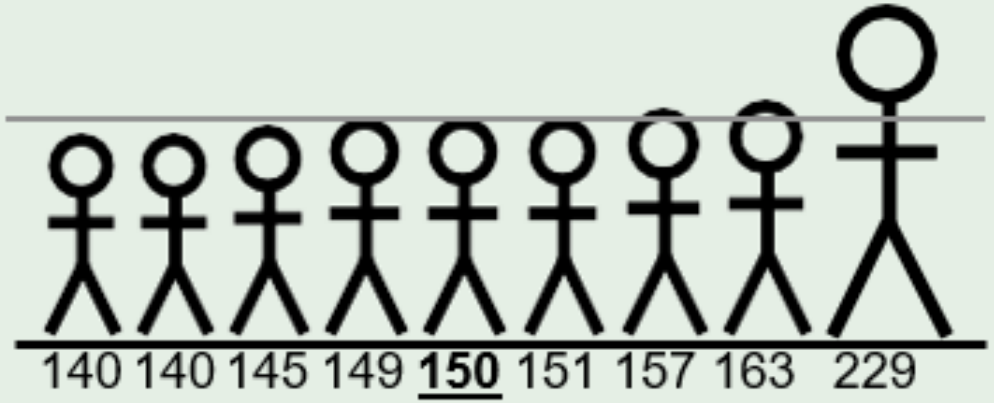
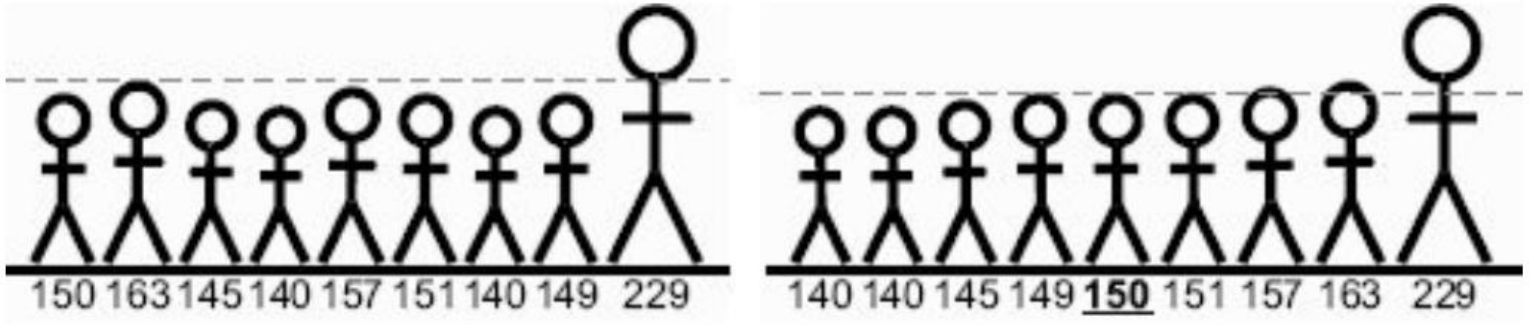


Figure: The members of the school basketball squad ordered by height, the dashed grey line shows the **median**.

ID	4	7	3	8	1	6	5	2	9
Height	140	140	145	149	**150**	151	157	163	229

- ✓ Median is not as sensitive to outliers as mean is.
- ✓ A **large difference** between the mean and median of a feature is an indication that there may be **outliers among the feature values**



- We also measure the **variation** in our data

- In essence, most of statistics, and in turn analytics, is about describing and understanding variation.

- **Range** –most easily calculated measure of variation

$$range = max(a) - min(a)$$

- **Variance** - measures the average difference between each value in a sample and the mean of that sample

$$var(a) = \frac{\sum_{i=1}^{n} (a_i - \overline{a})^2}{n-1}$$

Example

What is the variance of the heights of the two basketball squads?

ID	1	2	3	4	5	6	7	8
Height	150	163	145	140	157	151	140	149

ID	1	2	3	4	5	6	7	8
Height	192	102	145	165	126	154	122	188

Example

What is the range of the heights of the two basketball squads?

$$\text{range} = 163 - 140 = 23$$

$$\text{range} = 192 - 102 = 90$$

Variance

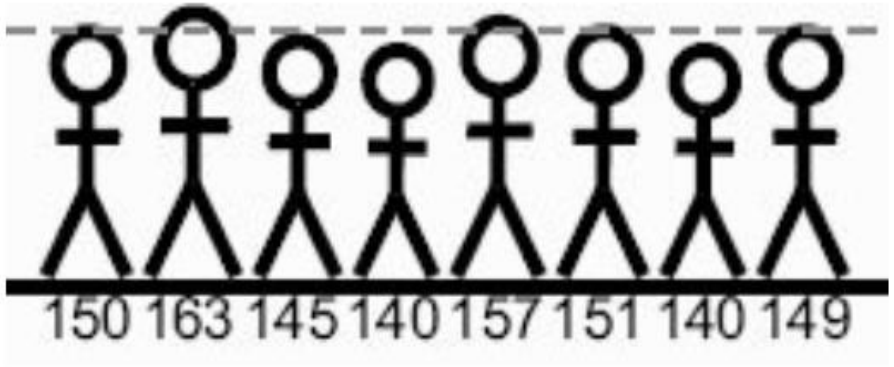


Fig. 1

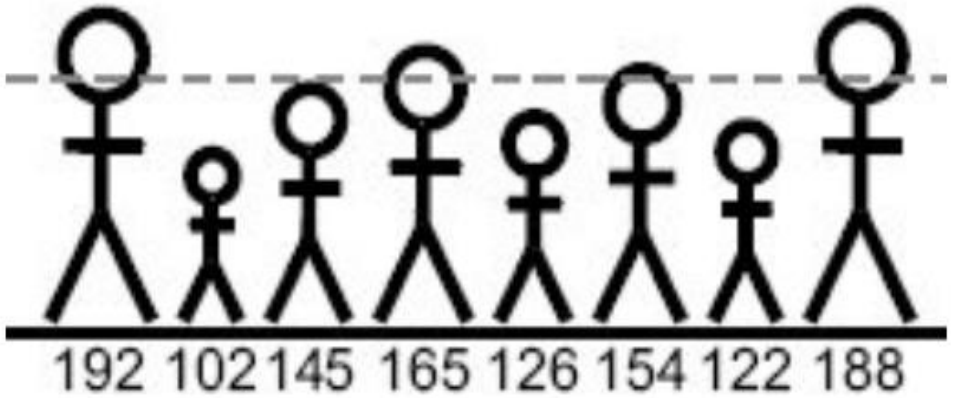
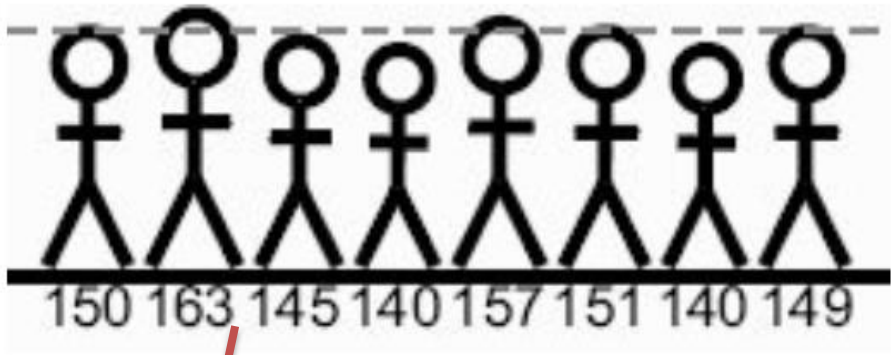
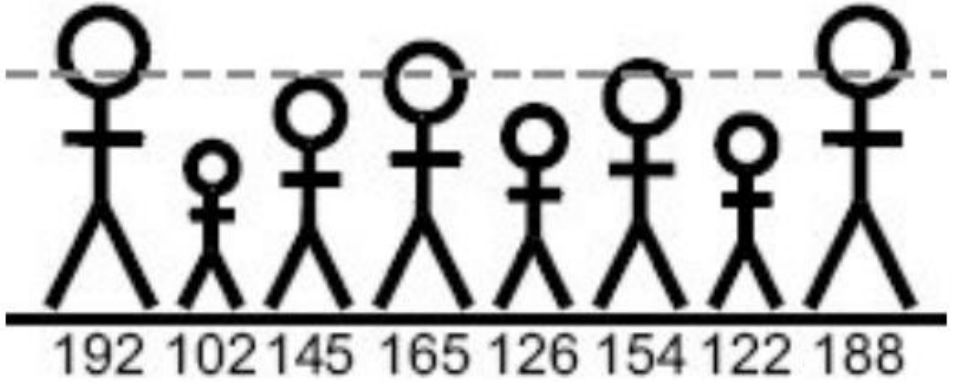


Fig. 2

- ✓ **The data represented by both Fig 1 and Fig 2 has same Mean (149.39).**
- ✓ **But Fig 2 has more variation in the data.**





$$var(\text{HEIGHT}) = \frac{(150 - 149.39)^2 + (163 - 149.39)^2 + \ldots + (149 - 149.39)^2}{8 - 1}$$

$$= 65.282$$

$$var(\text{HEIGHT}) = \frac{(192 - 149.39)^2 + (102 - 149.39)^2 + \ldots + (188 - 149.39)^2}{8 - 1}$$

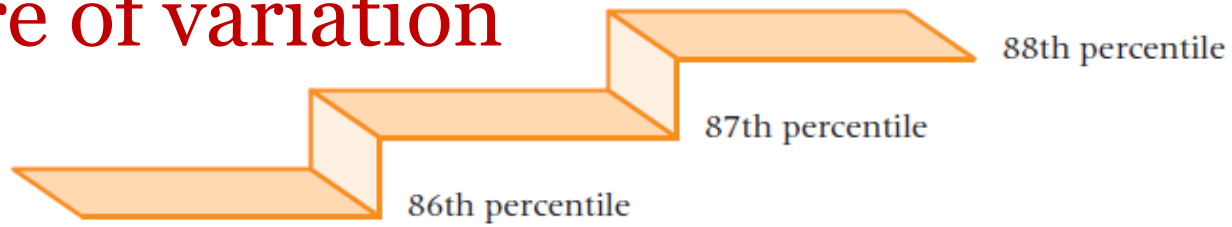
$$= 1{,}020.348$$

Standard Deviation:

$$sd(a) = \sqrt{var(a)}$$

- ✓ The standard deviation of the heights of the players on the first basketball team is 8.080 and for the second team is 31.943.
- ✓ These measures are in the same units as the heights.
- ✓ We can say that, on average, players on the first team vary by 8cm from the average of 149.39cm, while on the second team, they vary by approximately 32cm

Percentiles –another measure of variation



- ✓ Percentiles are measures of central tendency that divide a group of data into 100 parts.
- ✓ There are 99 percentiles because it takes 99 dividers to separate a group of data into 100 parts.
- ✓ The nth percentile is the value such that at least n percent of the data are below that value and at most (100 - n) percent are above that value.
- ✓ Specifically, the 87th percentile is a value such that at least 87% of the data are below the value and no more than 13% are above the value.
- ✓ Percentiles are “stair-step” values, as shown in Figure , because the 87th percentile and the 88th percentile have no percentile between.

- Percentiles are widely used in reporting test results.
- Almost all college or university students have taken the SAT, GRE, or GMAT examination.
- In most cases, the results for these examinations are reported in percentile form and also as raw scores.

- ✓ **Interquartile range (IQR)**: another measure of variation

- ✓ The inter-quartile range is calculated as the difference between the 25th percentile (Q1) and the 75th percentile(Q2). About 50% of the data is between *Q1* and *Q3*.

- ✓ These percentiles are also known as the lower quartile (or 1st quartile) and the upper quartile (or 3rd quartile)

Example

What is the variance of the heights of the two basketball squads?

ID	1	2	3	4	5	6	7	8
Height	150	163	145	140	157	151	140	149

ID	1	2	3	4	5	6	7	8
Height	192	102	145	165	126	154	122	188

- ✓ The inter-quartile range is 151 −140 = 11, while for the second team, it is 165 − 122 = 43.

Key Ideas

- ✓ The basic metric for location is the mean, but it can be sensitive to extreme values (outlier).
- ✓ Other metrics (median, trimmed mean) are more robust.
- ✓ The variance and standard deviation are the most widespread and routinely reported statistics of variability.
  - ✓ Both are sensitive to outliers.
- ✓ More robust metrics include mean, median, absolute deviations from the mean and percentiles.

- ✓ The mode is a measure of the central tendency of a categorical feature and is simply the most frequent level.
- ✓ We often also calculate a second mode which is just the second most common level of a feature.
- ✓ For the data in Table the mode is $19.00 because the offer price that recurred the most times (four) was $19.00.
- ✓ Organizing the data into an ordered array (an ordering of the numbers from smallest to largest) helps to locate the mode.

7.00 11.00 14.25 15.00 15.00 15.50 19.00 19.00 19.00 19.00 21.00 22.00 23.00 24.00 25.00 27.00 27.00 28.00 34.22 43.25

Offer Prices for the 20 Largest U.S. Initial Public Offerings in a Recent Year

$14.25	$19.00	$11.00	$28.00
24.00	23.00	43.25	19.00
27.00	25.00	15.00	7.00
34.22	15.50	15.00	22.00
19.00	19.00	27.00	21.00

Mode

- In the case of a tie for the most frequently occurring value, two modes are listed.
- Then the data are said to be **bimodal**.
- If a set of data is not exactly bimodal but contains two values that are more dominant than others, some researchers take the liberty of referring to the data set as bimodal even without an exact tie for the mode.
- Data sets with more than two modes are referred to as **multimodal**.

Example

- 4,6,4,1,8,7,7,2,5,7
- 1,2,4,4,5,6,7,7,7,8 - mode – 7

- 4,3,7,1,8,9,6
- 1,3,4,6,7,8,9 – no value repeating – NO MODE

- 2,7,5,2,8,9,7,3
- 2,2,3,5,7,7,8,9 – 2 n 7 repeated equally – Modes – 2 n 7 – Bi modal

- For categorical features we are interested primarily in frequency counts and proportions.
  - The frequency count of each level of a categorical feature is calculated by counting the number of times that level appears in the sample.
  - The proportion for each level is calculated by dividing the frequency count for that level by the total sample size.
  - Frequencies and proportions are typically presented in a frequency table.

Table: A dataset showing the positions and weekly training expenses of a school basketball squad.

ID	Position	Training Expenses	ID	Position	Training Expenses
1	center	56.75	11	center	550.00
2	guard	1,800.11	12	center	223.89
3	guard	1,341.03	13	center	103.23
4	forward	749.50	14	forward	758.22
5	guard	1,150.00	15	forward	430.79
6	forward	928.30	16	forward	675.11
7	center	250.90	17	guard	1,657.20
8	guard	806.15	18	guard	1,405.18
9	guard	1,209.02	19	guard	760.51
10	forward	405.72	20	forward	985.41

Table: A frequency table for the POSITION feature from the professional basketball squad dataset in Table.

Level	Count	Proportion
guard	8	40%
forward	7	35%
center	5	25%

Data Visualization

- ✓ When performing data exploration **data visualization** can help enormously.

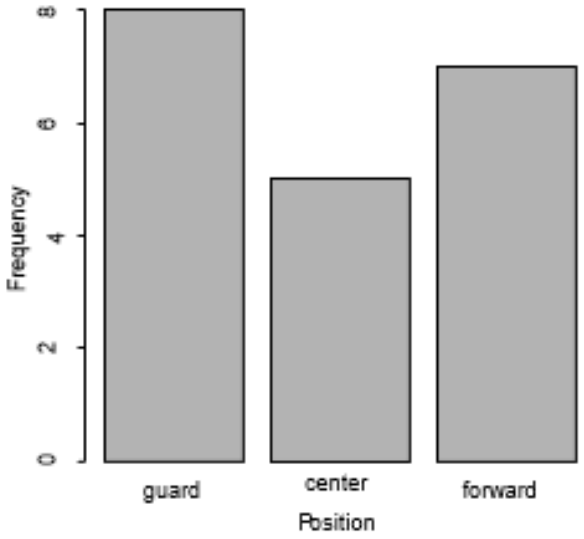
- ✓ We will describe three important data visualization techniques that can be used to visualize the values in a single feature:
  - ✓ the **bar plot**
  - ✓ the **histogram**
  - ✓ the **box plot**

Table: A dataset showing the positions and weekly training expenses of a school basketball squad.

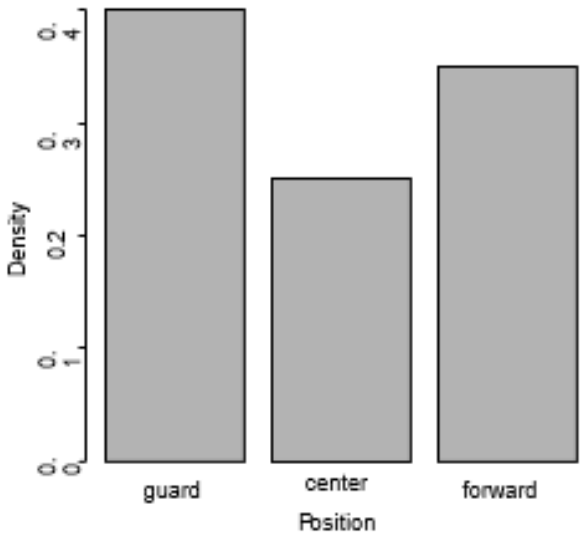
ID	Position	Training Expenses	ID	Position	Training Expenses
1	center	56.75	11	center	550.00
2	guard	1,800.11	12	center	223.89
3	guard	1,341.03	13	center	103.23
4	forward	749.50	14	forward	758.22
5	guard	1,150.00	15	forward	430.79
6	forward	928.30	16	forward	675.11
7	center	250.90	17	guard	1,657.20
8	guard	806.15	18	guard	1,405.18
9	guard	1,209.02	19	guard	760.51
10	forward	405.72	20	forward	985.41

Bar plots are great for categorical features

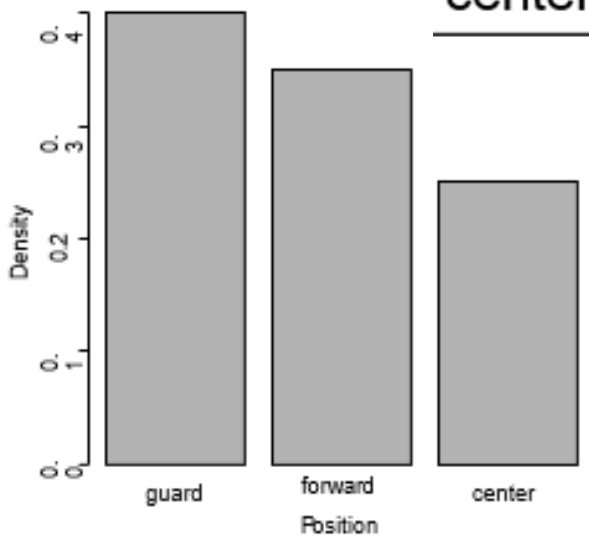
Level	Count	Proportion
guard	8	40%
forward	7	35%
center	5	25%



(a) Frequency



(b) Proportion

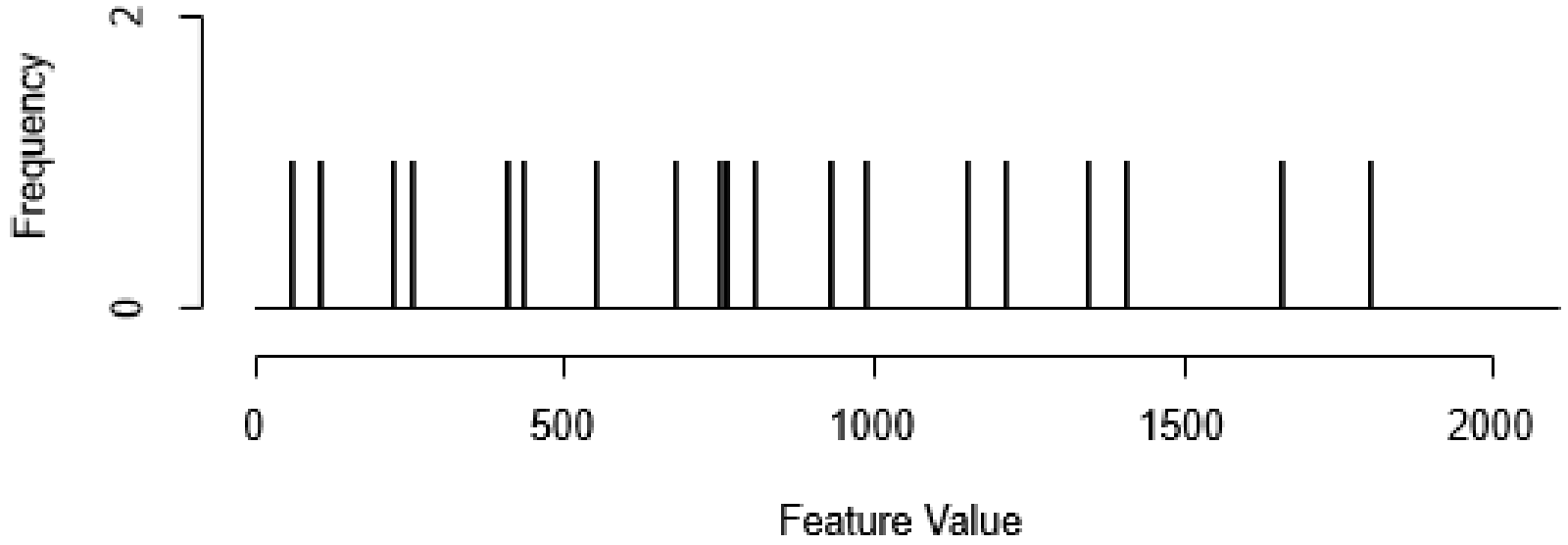


(c) Ordered

Bar charts: The frequency or proportion for each category plotted as bars.

Pie charts: The frequency or proportion for each category plotted as wedges in a pie.

Bar plots don’t work for continuous features



2/(20x200)

By dividing the range of a variable into intervals, or bins, we can generate histograms

(a) 200 unit intervals

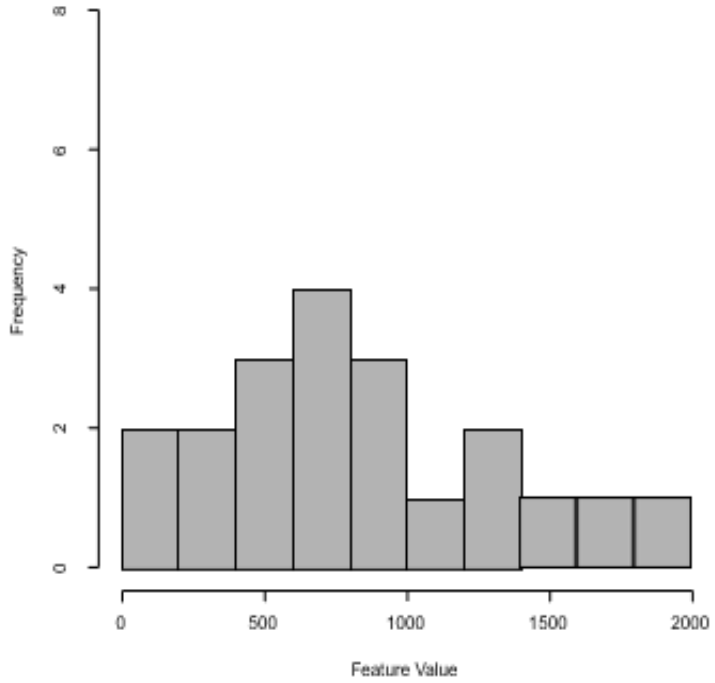
Interval	Count	Density	Prob
[0, 200)	2	0.0005	0.1
[200, 400)	2	0.0005	0.1
[400, 600)	3	0.00075	0.15
[600, 800)	4	0.001	0.2
[800, 1000)	3	0.00075	0.15
[1000, 1200)	1	0.00025	0.05
[1200, 1400)	2	0.0005	0.1
[1400, 1600)	1	0.00025	0.05
[1600, 1800)	1	0.00025	0.05
[1800, 2000)	1	0.00025	0.02

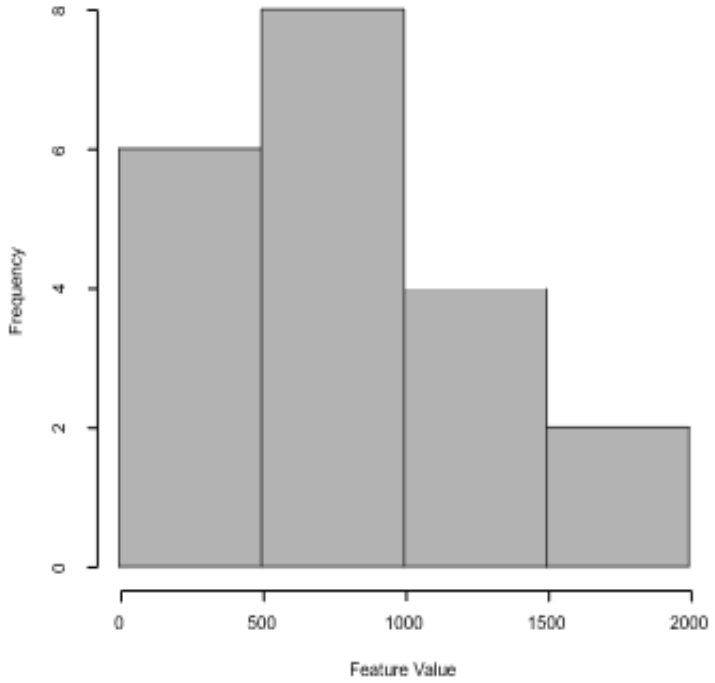
(b) 500 unit intervals

Interval	Count	Density	Prob
[0, 500)	6	0.0006	0.3
[500, 1000)	8	0.0008	0.4
[1000, 1500)	4	0.0004	0.2
[1500, 2000)	2	0.0002	0.1

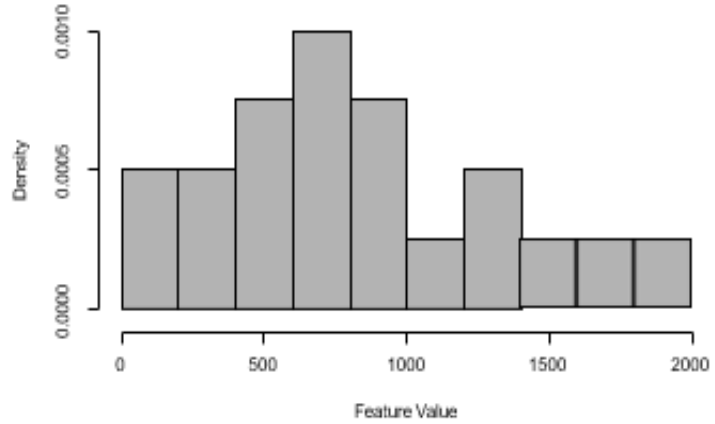
Frequency and density histograms for the continuous Training Expenses feature from Table

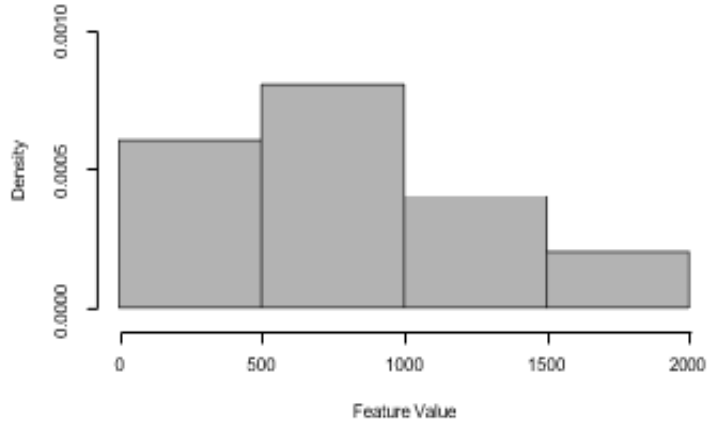
A **frequency histogram plots** frequency counts on the y-axis and variable values on the x-axis; it gives a sense of the distribution of the data at a glance





A **density plot** is a smoothed version of a histogram; it requires a function to estimate a plot based on the data (multiple estimates are possible, of course)





Box plots/ whisker plots

are another useful way of visualising continuous variables

- ✓ **A box and whisker plot—also called a box plot—displays the five-number summary of a set of data.**
- ✓ The five-number summary is the minimum, first quartile, median, third quartile, and maximum.

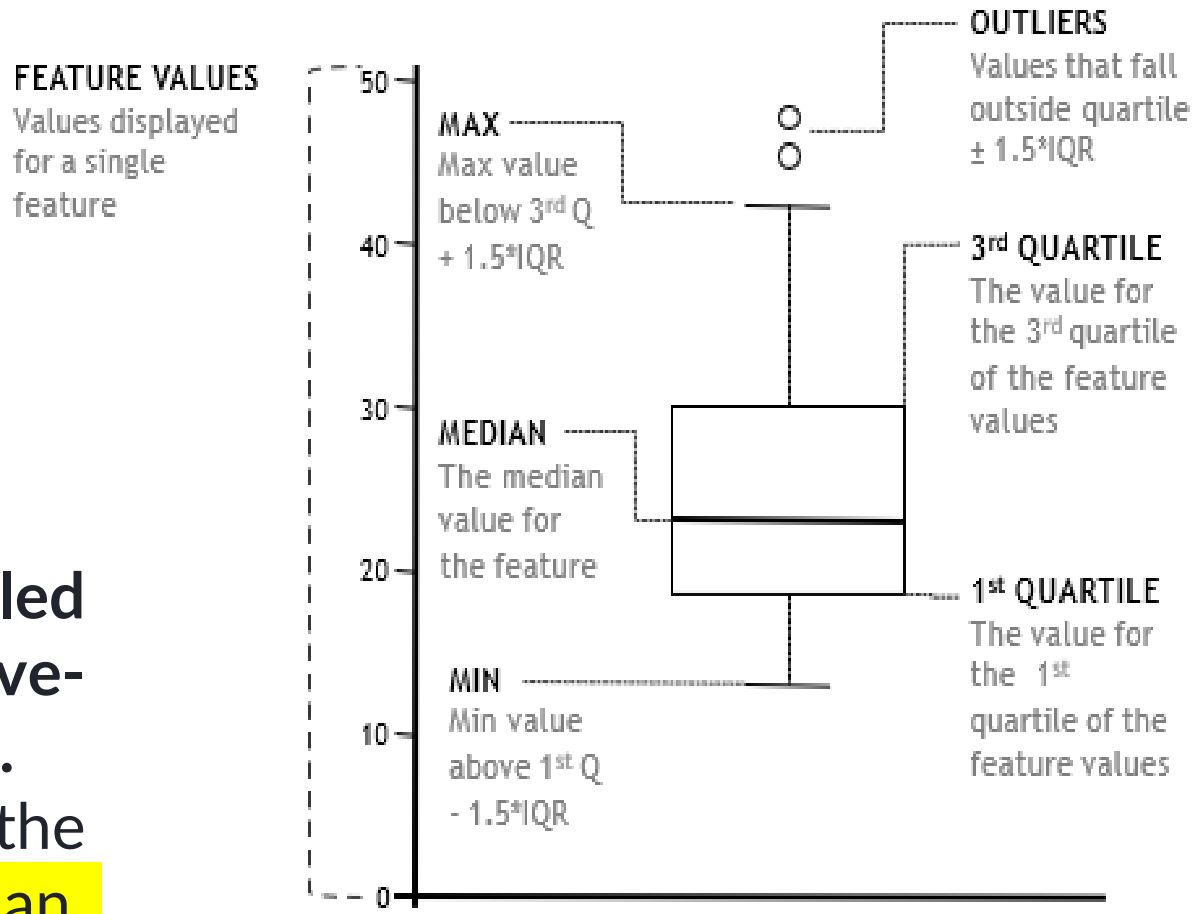


Figure: The structure of a box plot.

Key points-Box plot

- ✓ A boxplot—with the top and bottom of the box at the 75th and 25th percentiles, respectively—also gives a quick sense of the distribution of the data; it is often used in side-by-side displays to compare distributions

- ✓ The whiskers that emerge from the top and bottom of the main rectangle in a box plot are designed to show the **range of the data**

- ✓ Values that fall outside the whiskers are referred to as **outliers and are shown as small circles**

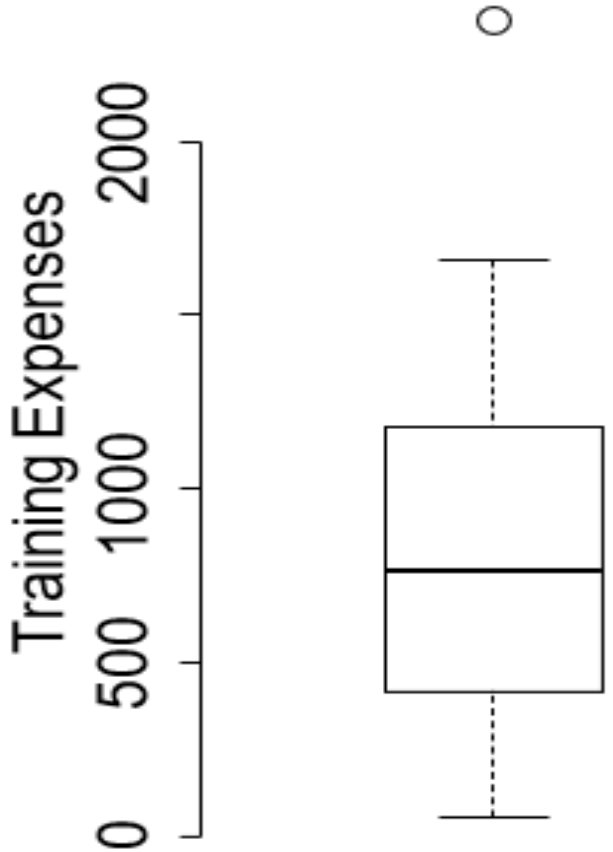
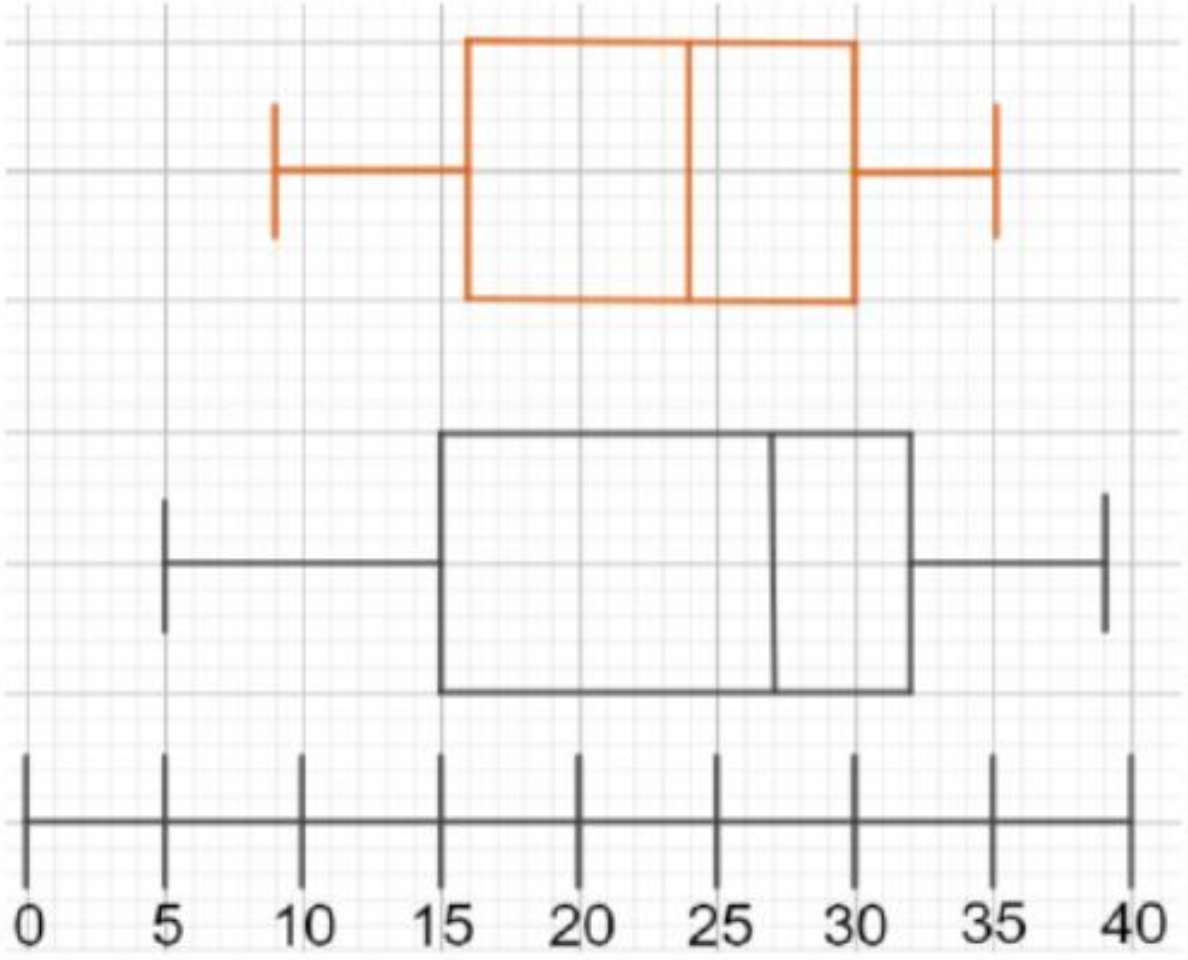


Figure: A box plot for the TRAINING EXPENSES feature from the basketball squad dataset in Table

Question: The following box plots show how many hours of TV is watched by a year 11 class (orange) and a year 9 class (grey) in a given month. Compare the box plots

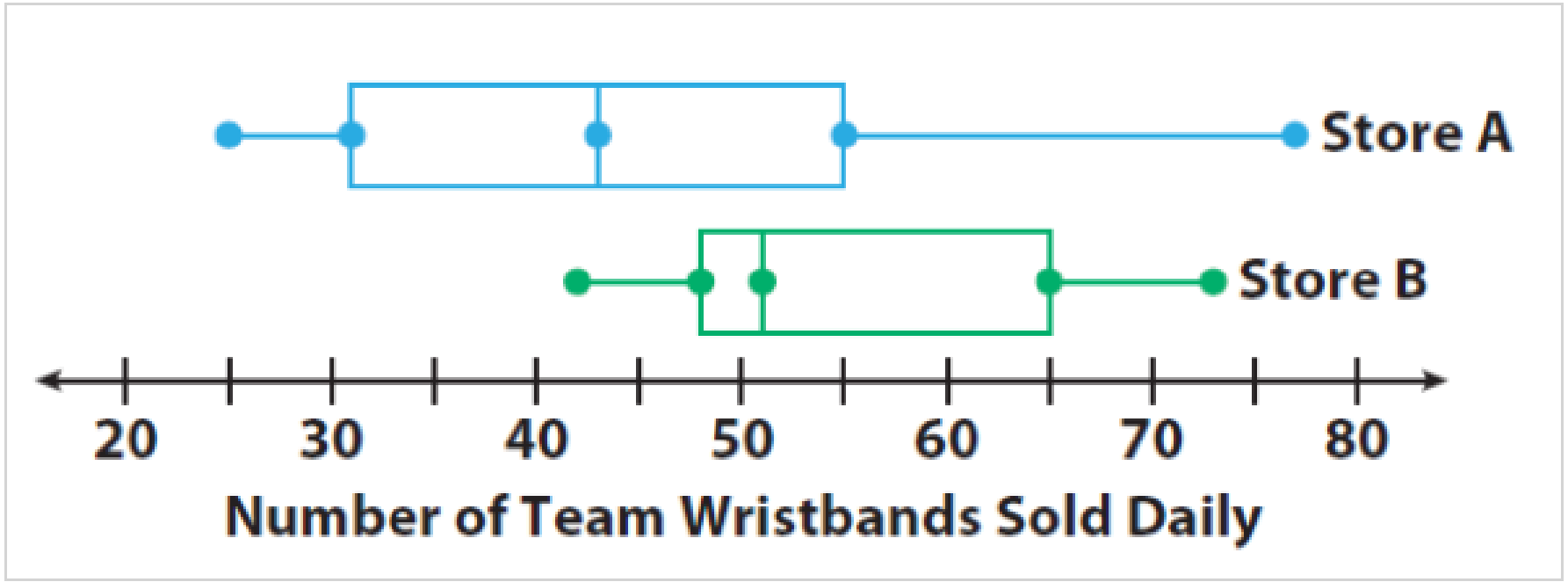


1. When comparing box plots you want to look at the **median** and **interquartile range** as your first two comparisons.

2. The **median** time is greater for the year 9 class.

3. The year 9 class also have a larger **interquartile range**.

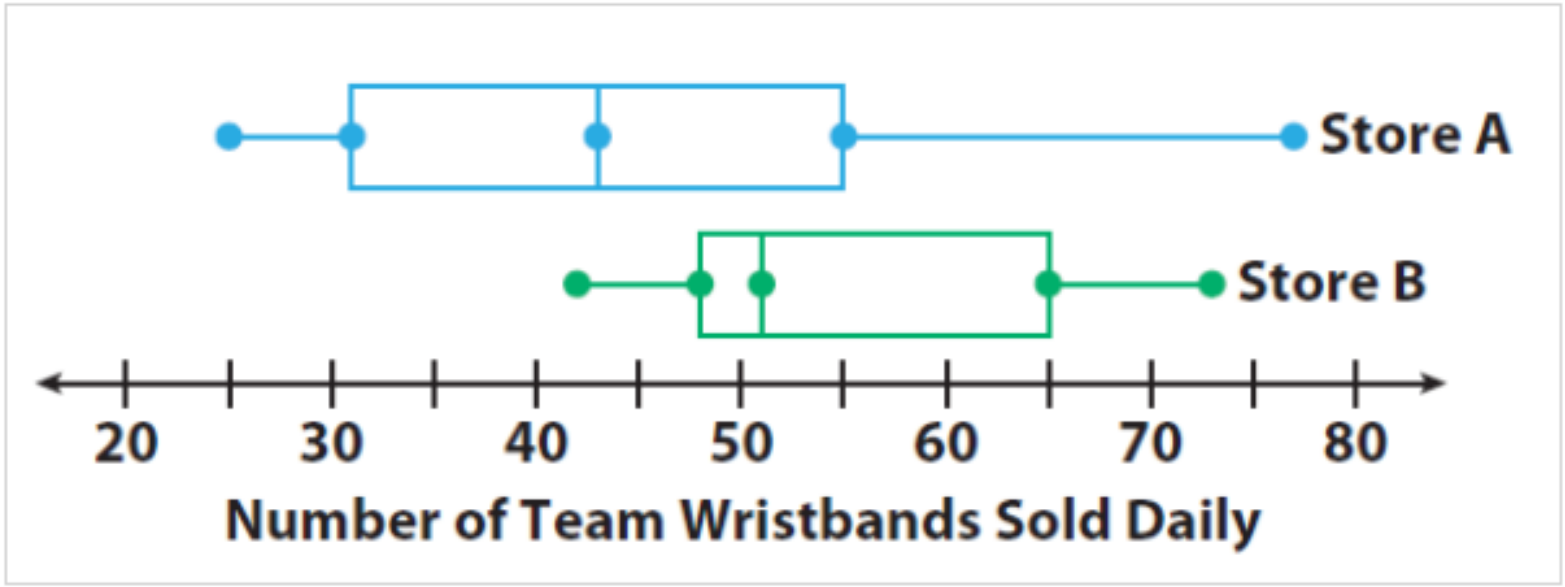
Question: The box plots show the distribution of the number of team wristbands sold daily by two different stores over the same time period.



1. Compare the shapes of the box plots.

Answer: Store A's box and right whisker are longer than Store B's.

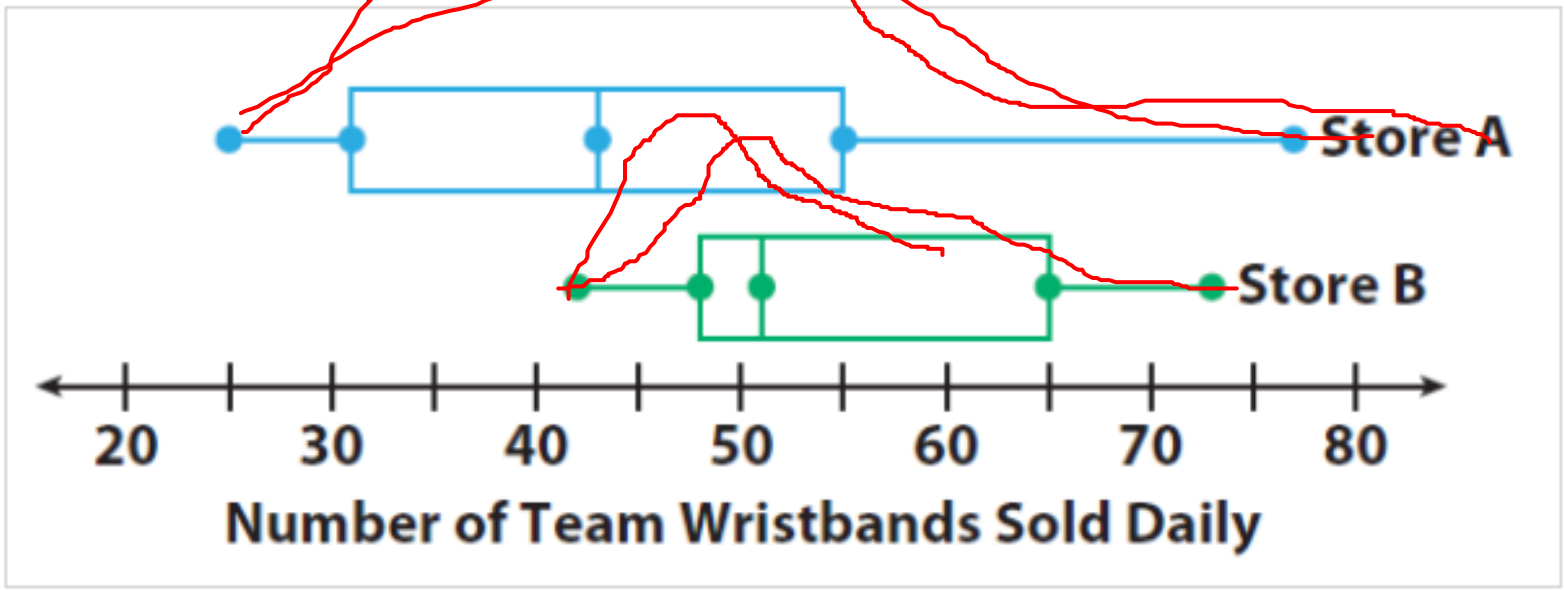
Question: The box plots show the distribution of the number of team wristbands sold daily by two different stores over the same time period.



1. Compare the centers of the box plots.

Answer: Store A's median is about 43, and Store B's is about 51. Store A's median is close to Store B's minimum value, so about 50% of Store A's daily sales were less than sales on Store B's worst day.

Question: The box plots show the distribution of the number of team wristbands sold daily by two different stores over the same time period.



1. Compare the spreads of the box plots.-

Answer: Store A has a greater spread. Its range and inter quartile range are both greater. Four of Store B's key values are greater than Store A's corresponding value. Store B had a greater number of sales overall.

Identifying Data Quality Issues

- A data quality issue is loosely defined as anything unusual about the data in an ABT/Rectangular Data frame.

- The most common data quality issues are:
  - ✓ missing values
  - ✓ irregular cardinality
  - ✓ outliers

- The data quality issues we identify from a data quality report will be of two types:
  - ✓ Data quality issues due to invalid data.
  - ✓ Data quality issues due to valid data.

Data Quality issues

Valid data –need not take any coercive action unless required by ML models

Invalid data- immediate actions to correct them

Missing Data

- Data is not always available
  - ✓ e.g., many tuples have no recorded value for several attributes

- Missing data may be due to
  - ✓ equipment malfunction
  - ✓ inconsistent with other recorded data and thus deleted
  - ✓ data not entered due to misunderstanding
  - ✓ certain data may not be considered important at the time of entry
  - ✓ not register history or changes of the data

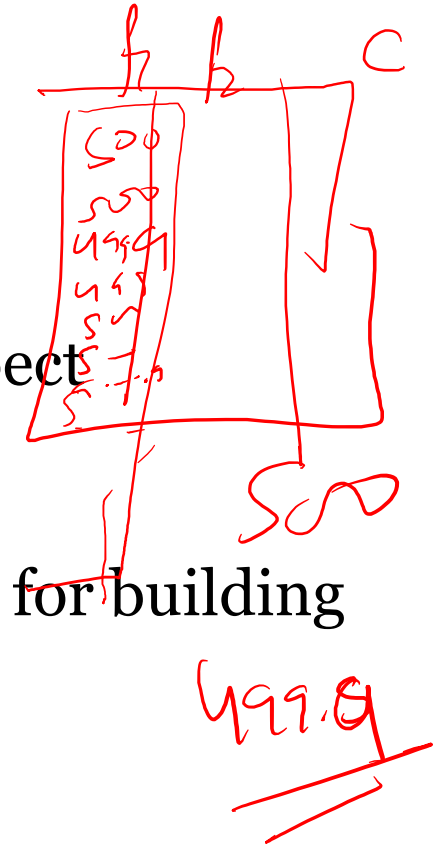
Missing Values

- Highlight the percentage of missing values for each feature
- If proportion of missing values in a feature is >60%, good idea to remove

Missing Values Handling

- **Imputation** replaces missing feature values with a plausible estimated value based on the feature values that are present.
- The most common approach to imputation is to replace missing values for a feature with a measure of the central tendency of that feature.
- We would be reluctant to use imputation on features missing in excess of 30% of their values and would strongly recommend against the use of imputation on features missing in excess of 50% of their values.

Irregular cardinality

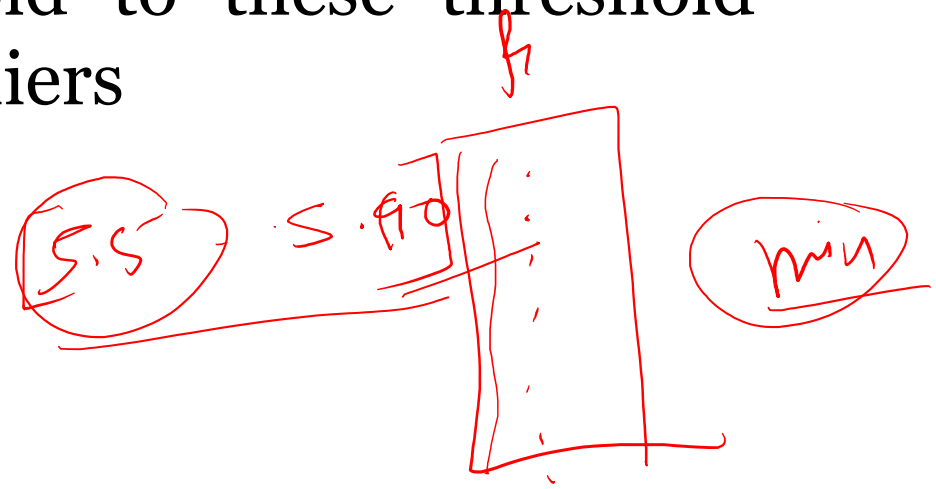


- Cardinality-number of distinct values present for a feature
- Issue- when the cardinality for a feature does not match what we expect
- features with a cardinality of 1.
  - ✓ same value for every instance and contains no information useful for building predictive models
- Categorical features incorrectly labeled as continuous.
  - ✓ the number of children a person has 1 for female and 0 for male
- Categorical feature has a much higher cardinality
  - ✓ categorical feature storing gender with a cardinality of 6
  - ✓ categorical feature simply has a very high number of levels- above 50

Handle Outliers

- The easiest way to handle outliers is to use a clamp transformation that clamps all values above an upper threshold and below a lower threshold to these threshold values, thus removing the offending outliers

$$a_i = \begin{cases} lower & \text{if } a_i < lower \\ upper & \text{if } a_i > upper \\ a_i & otherwise \end{cases}$$



- where a_i is a specific value of feature a, and lower and upper are the lower and upper threshold

Handle Outliers

- The upper and lower thresholds can be set manually based on domain knowledge or can be calculated from data.

- calculate clamp thresholds-
  - lower threshold to the 1st quartile value minus 1.5 times the inter-quartile range i.e. Q1-1.5xIQR
  - upper threshold to the 3rd quartile plus 1.5 times the inter-quartile range. i.e. Q3+1.5IQR
  - setting the upper and lower thresholds to the mean value of a feature plus or minus 2 times the standard deviation

- **Strong Outlier**
  - Outliers with extreme deviation from the rest of the dataset.

- **Weak Outlier**

Determining Outliers:. **Data point < Q1-1.5xIQR**
Data point> Q3+1.5xIQR

Data point is considered as outliers

Strong Outliers : **Data point < Q1-3.0xIQR**
Data point> Q3+3.0xIQR

Data point is considered as Strong outliers

Weak Outliers : Besides strong outliers, there is another category for outliers. If a data value is an outlier, but not a strong outlier, then we say that the value is a weak outlier.

Example 1

First, suppose that we have the data set {1, 2, 2, 3, 3, 4, 5, 5, 9}. The number 9 certainly looks like it could be an outlier. It is much greater than any other value from the rest of the set.

To objectively determine if 9 is an outlier, we use the above methods. The first quartile is 2 and the third quartile is 5, which means that the interquartile range is 3.

We multiply the interquartile range by 1.5, obtaining 4.5, and then add this number to the third quartile. The result, 9.5, is greater than any of our data values. Therefore, there are no outliers.

Example 2

Now we look at the same data set as before, with the exception that the largest value is 10 rather than 9: {1, 2, 2, 3, 3, 4, 5, 5, 10}.

$Q_1 = 2$ $Q_3 = 5$

$Q_2 = 3$

10 is a outlier ?

? strong or weak

10 ≯ 14

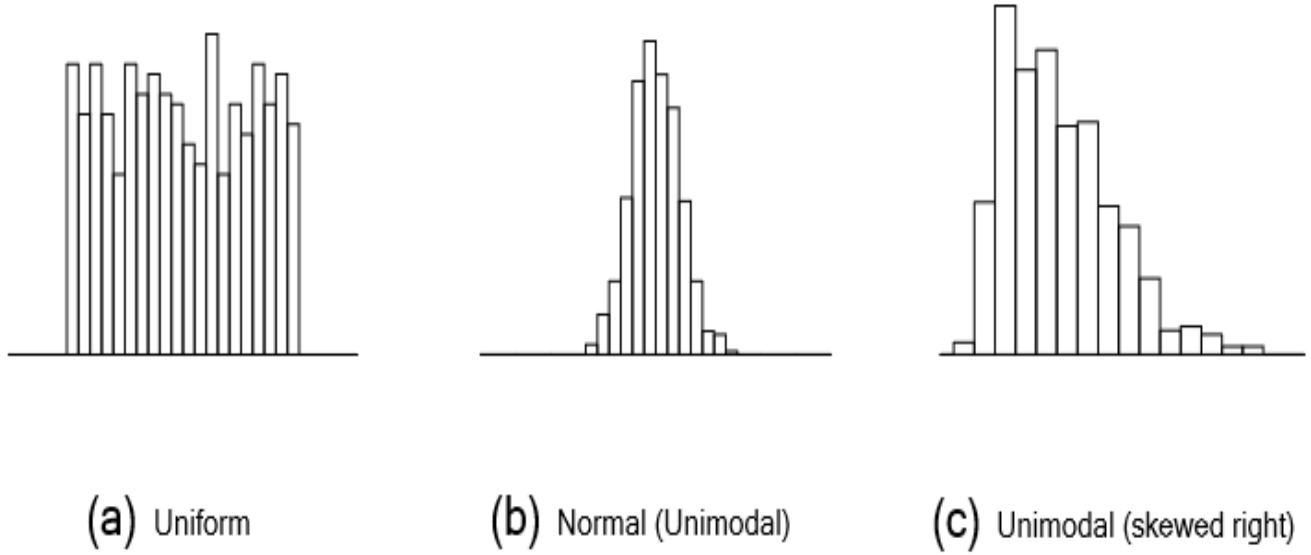
$\Rightarrow Q_3 + 3\,IQR = 5 + 3 \times 3$

$= 5 + 9 = 14$

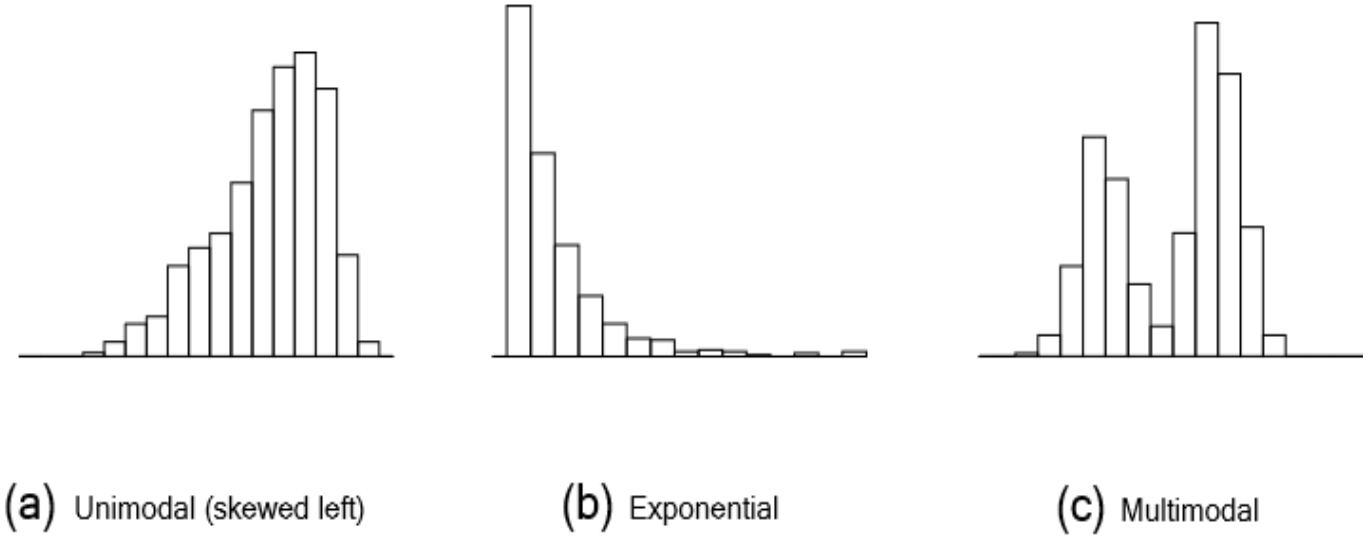
The first quartile, third quartile, and interquartile range are identical to example 1. When we add 1.5 x IQR = 4.5 to the third quartile, the sum is 9.5. Since 10 is greater than 9.5 it is considered an outlier.

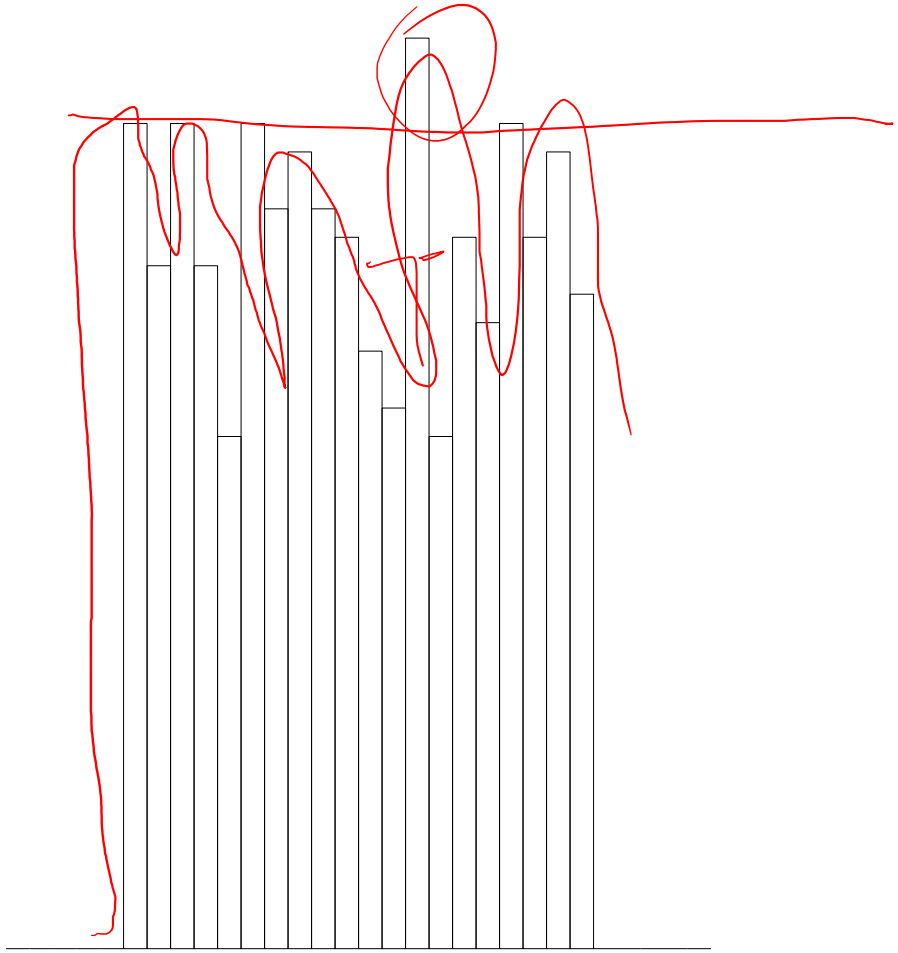
Is 10 a strong or weak outlier? For this, we need to look at 3 x IQR = 9.
When we add 9 to the third quartile, we end up with a sum of 14. Since 10 is not greater than 14, it is not a strong outlier. Thus, we conclude that 10 is a weak outlier.

When we generate histograms of features there are a number of common, well understood shapes that we should look out for.

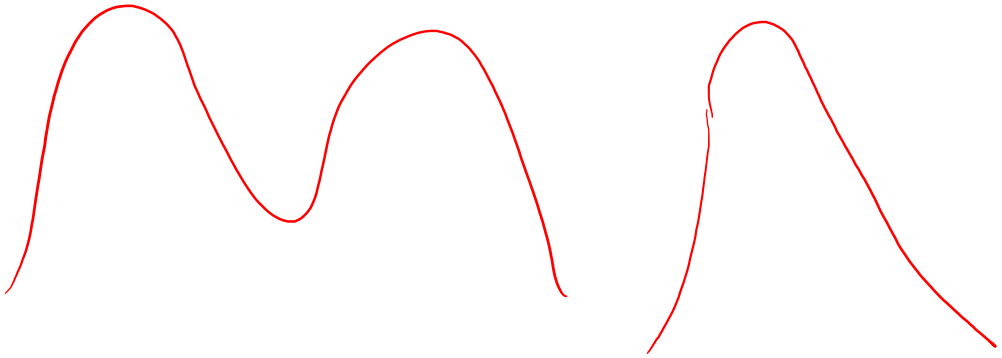


Histograms for different sets of data each of which exhibit well-known, common characteristics.

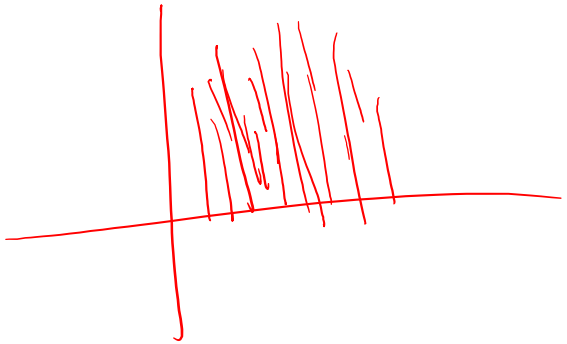


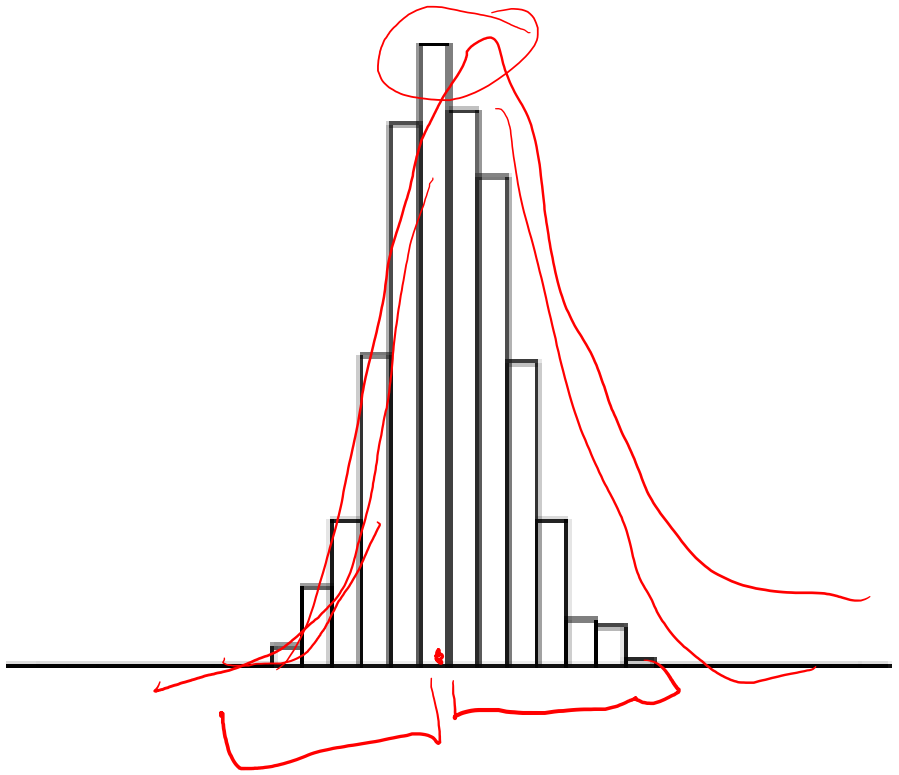


Uniform



A uniform distribution indicates that a feature is equally likely to take a value in any of the ranges present.



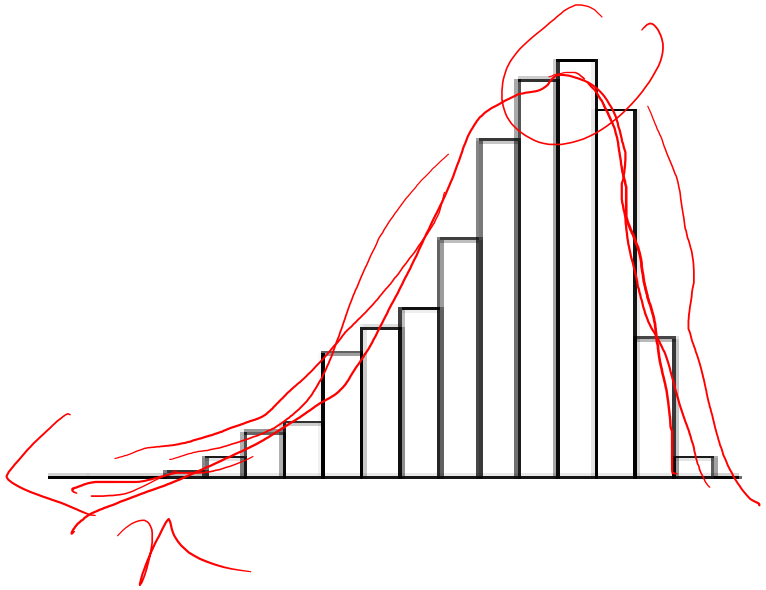


Normal (Unimodal)

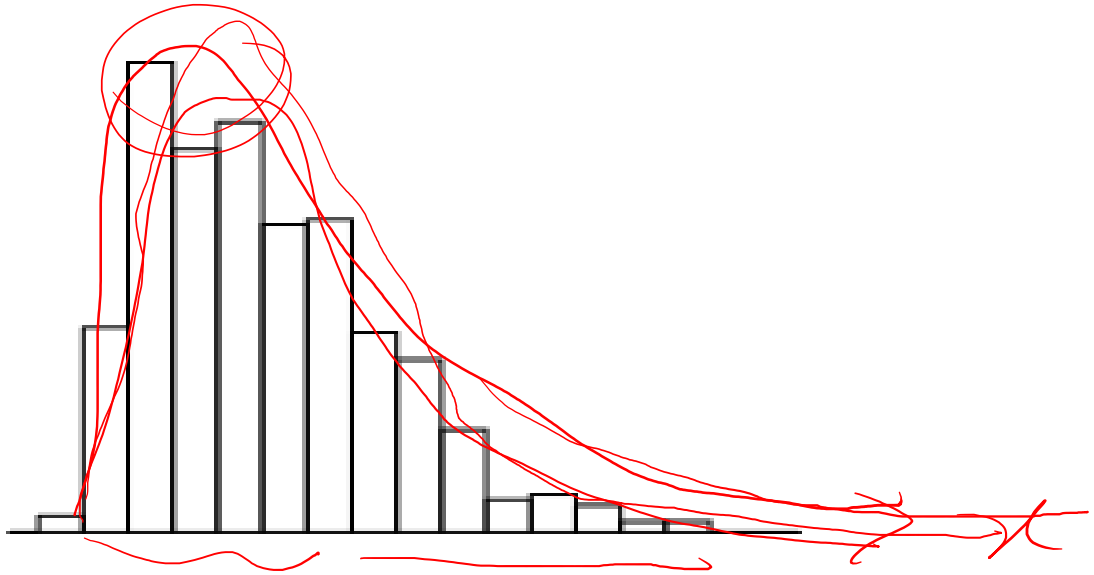
Features following a normal distribution are characterized by a strong tendency towards a central value and symmetrical variation to either side of this.

mean = median = mode

- A fundamental task in many statistical analyses is to characterize the location and variability of a data set.
- A further characterization of the data includes skewness and kurtosis.



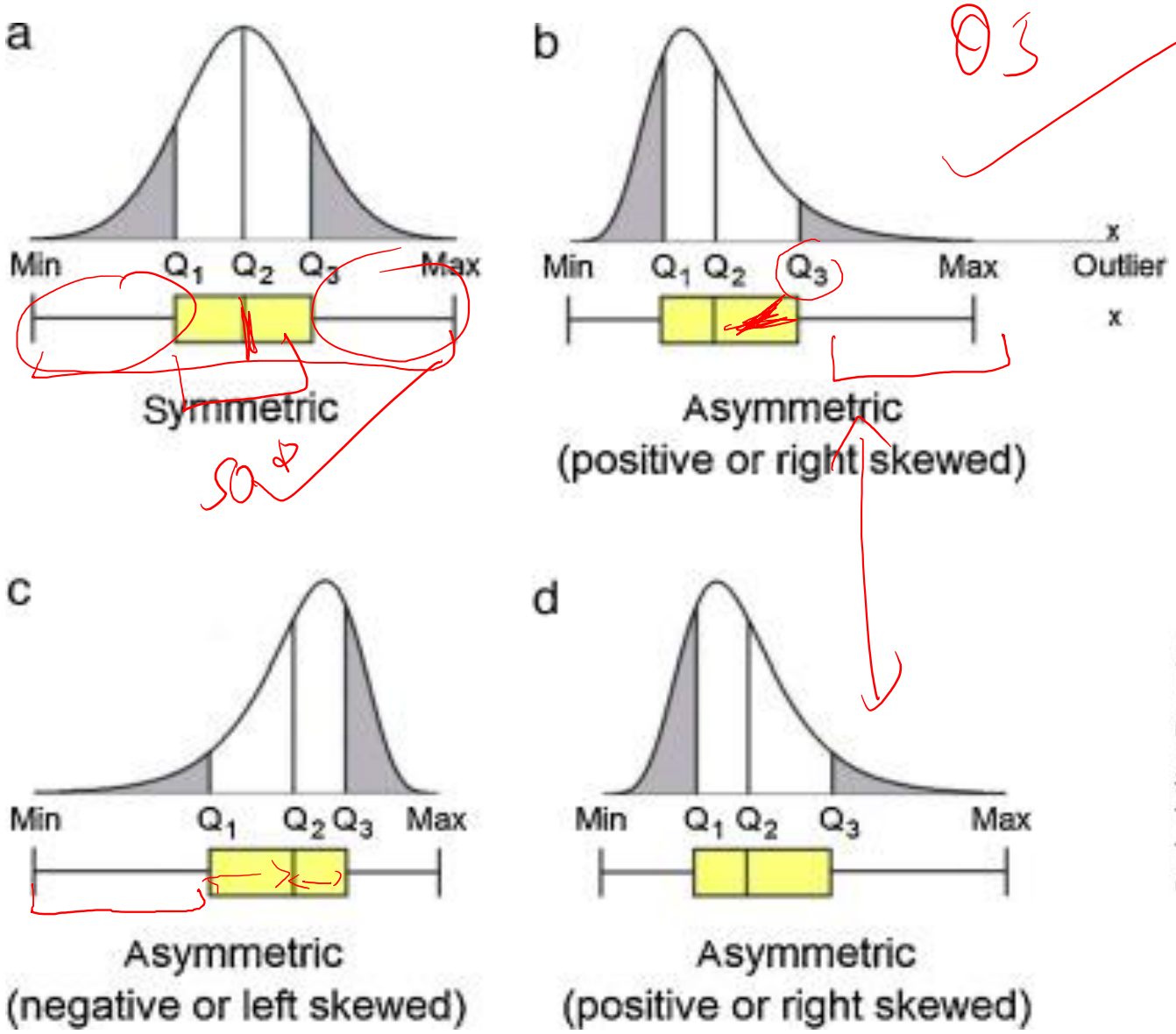
Unimodal (skewed left)

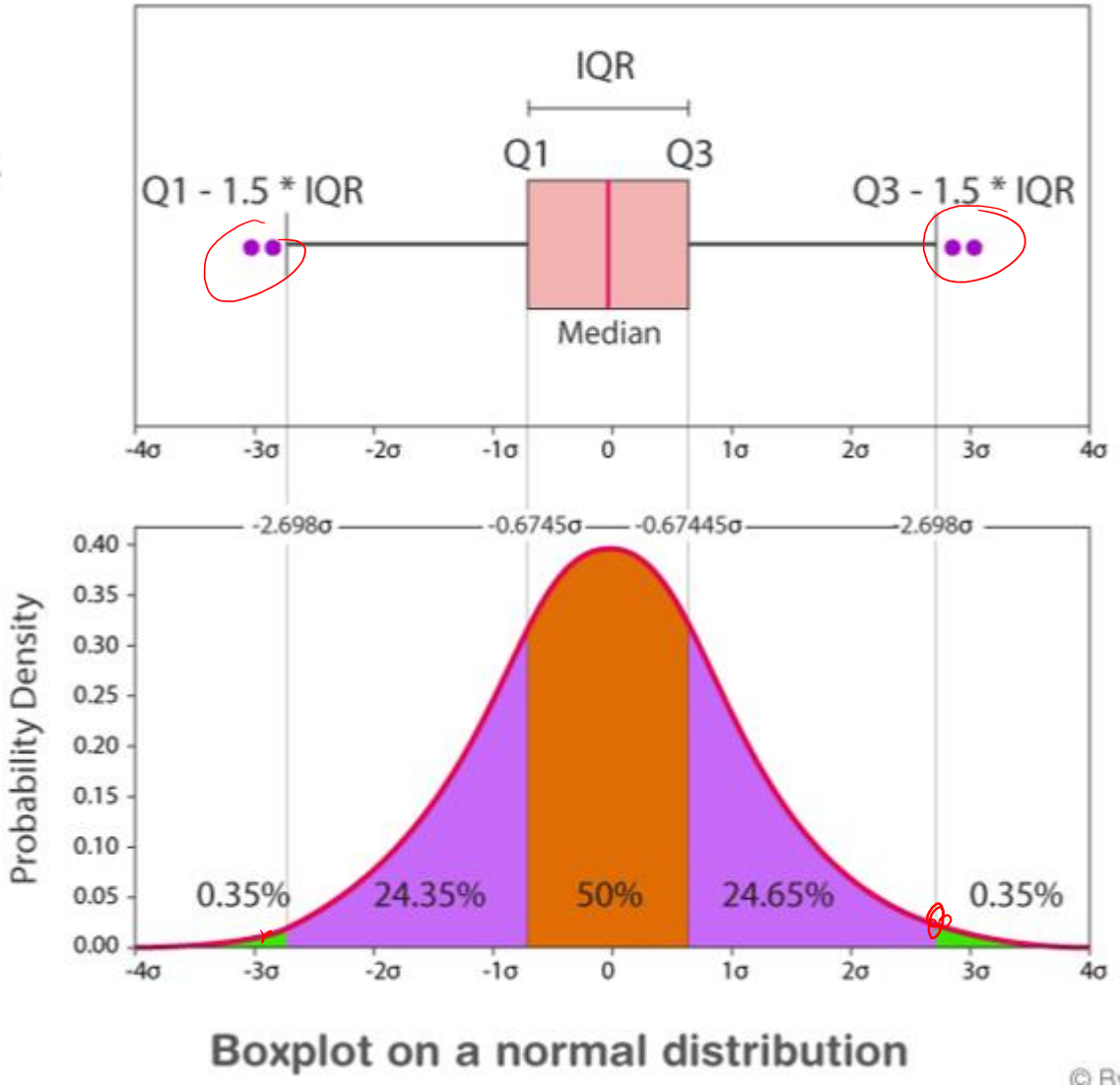


Unimodal (skewed right)

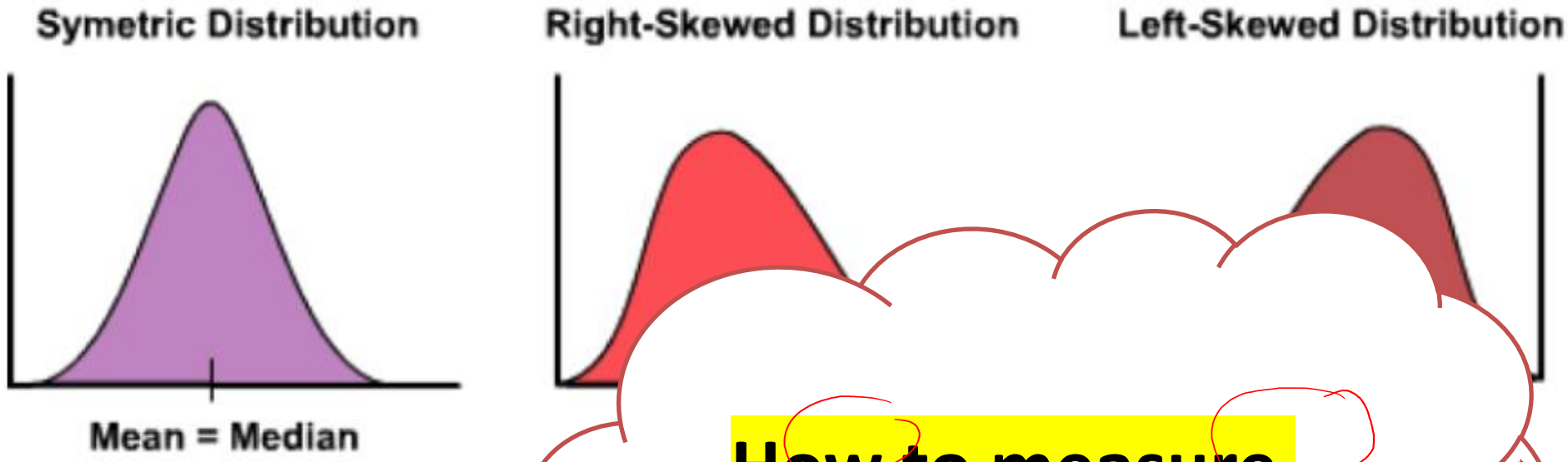
- Data is skewed when its distribution curve is asymmetrical (as compared to a normal distribution curve that is perfectly symmetrical) and *skewness* is the measure of the asymmetry.

- The skewness for a normal distribution is 0.





Boxplot on a normal distribution



How to measure skewness??

- There are 2 different types of skew

- **Effects of skewed data:** Degrade models) to describe typical cases as it has to d skewed data will predict better on data points with lower va es.

- Skewed data also does not work well with many st cal n . However, ***tree based models are not affected***

- To ensure that the machine learning model capabilities is not affected, skewed data has to be transformed to approximate to a normal distribution.

Pearson's Coefficient of Skewness

- ✓ Karl Pearson developed two methods to measure skewness in a sample.
- ✓ Pearson's Coefficient of Skewness #1 uses the mode.
  - ✓ The formula is:

$$Sk_1 = \frac{\bar{X} - Mo}{s}$$

where $\bar{X}$ = the mean, Mo = the mode and
s = the standard deviation for the sample.

- ✓ Pearson's Coefficient of Skewness #2 uses the median.
  - ✓ The formula is:

$$Sk_2 = \frac{3(\bar{X} - Md)}{s}$$

Where $\bar{X}$ = the mean, Md = the mode and
s = the standard deviation for the sample.

Example problem: Use Pearson's Coefficient #1 and #2 to find the skewness for data with the following characteristics:

- Mean = 70.5.
- Median = 80.
- Mode = 85.
- Standard deviation = 19.33.

$$Sk_1 = \frac{\bar{X} - Mo}{s}$$

$$Sk_2 = \frac{3(\bar{X} - Md)}{s}$$

Pearson's Coefficient of Skewness #1 (Mode):

Step 1: Subtract the mode from the mean: 70.5 – 85 = -14.5.

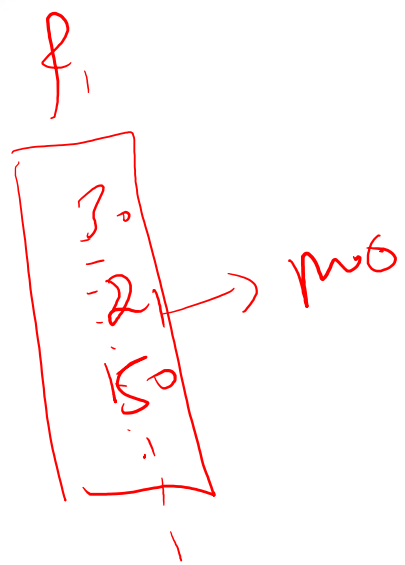
Step 2: Divide by the standard deviation: -14.5 / 19.33 = -0.75.

Pearson's Coefficient of Skewness #2 (Median):

Step 1: Subtract the median from the mean: 70.5 – 80 = -9.5.

Step 2: Multiply Step 1 by 3: -9.5(3) = -28.5

Step 2: Divide by the standard deviation: -28.5 / 19.33 = -1.47.



Interpretation

In general:

- ✓ The direction of skewness is given by the sign.
- ✓ The coefficient compares the sample distribution with a normal distribution.
- ✓ The larger the value, the larger the distribution differs from a normal distribution.
  - ✓ A value of zero means no skewness at all.
  - ✓ A large negative value means the distribution is negatively skewed.
  - ✓ A large positive value means the distribution is positively skewed.

Caution

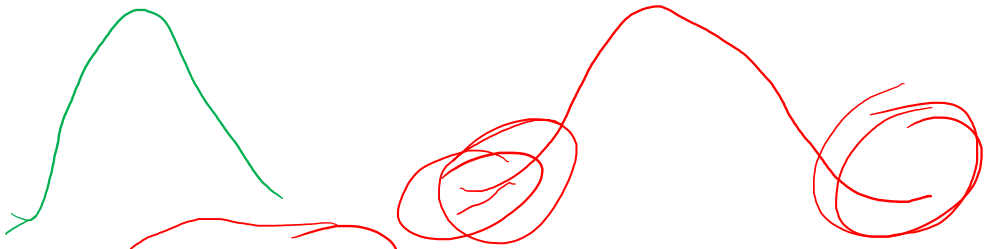
- Pearson's first coefficient of skewness uses the mode. Therefore, if the mode is made up of too few pieces of data it won't be a stable measure of central tendency.

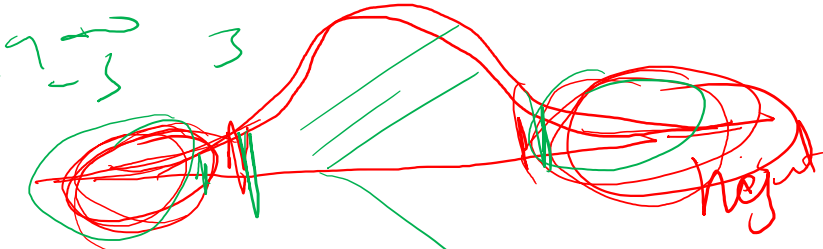
- For example, the mode in both these sets of data is 9:
  1 2 3 4 5 6 7 8 9 9.
  1 2 3 4 5 6 7 8 9 9 9 9 9 9 9 9 9 9 9 10 12 12 13.

- In the first set of data, the mode only appears twice. This isn't a good measure of central tendency so you would be cautioned *not* to use Pearson's coefficient of skewness.

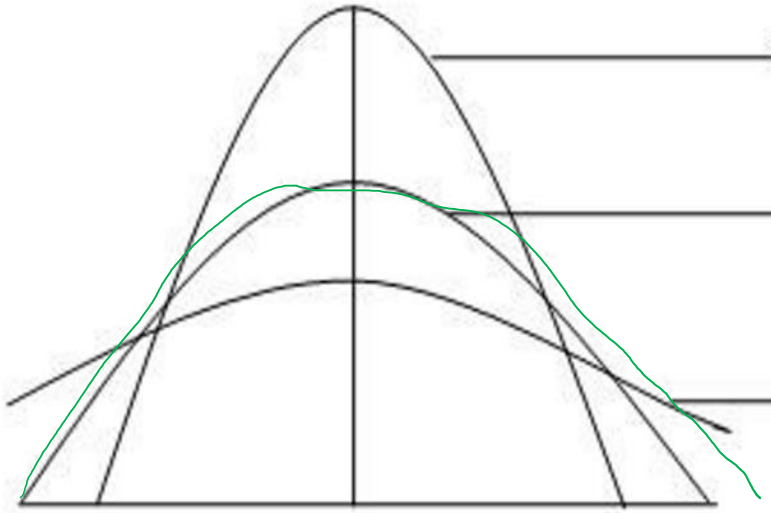
- The second set of data has a more stable set (the mode appears 12 times). Therefore, *Pearson's coefficient of skewness will likely give you a reasonable result.*

Kurtosis





- Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution. That is, data sets with high kurtosis tend to have heavy tails, or outliers. Data sets with low kurtosis tend to have light tails, or lack of outliers.
- Kurtosis is a measure of the tailedness of a distribution. Tailedness is how often outliers occur. Excess kurtosis is the tailedness of a distribution relative to a normal distribution.
  - Distributions with medium kurtosis (medium tails) are mesokurtic.
  - Distributions with low kurtosis (thin tails) are platykurtic.
  - Distributions with high kurtosis (fat tails) are leptokurtic.
- The histogram is an effective graphical technique for showing both the skewness and kurtosis of data set.



Leptokurtic, Kurtosis > 3, Excess Kurtosis (+ve)

Mesokurtic, Kurtosis = 3, Excess Kurtosis (0)

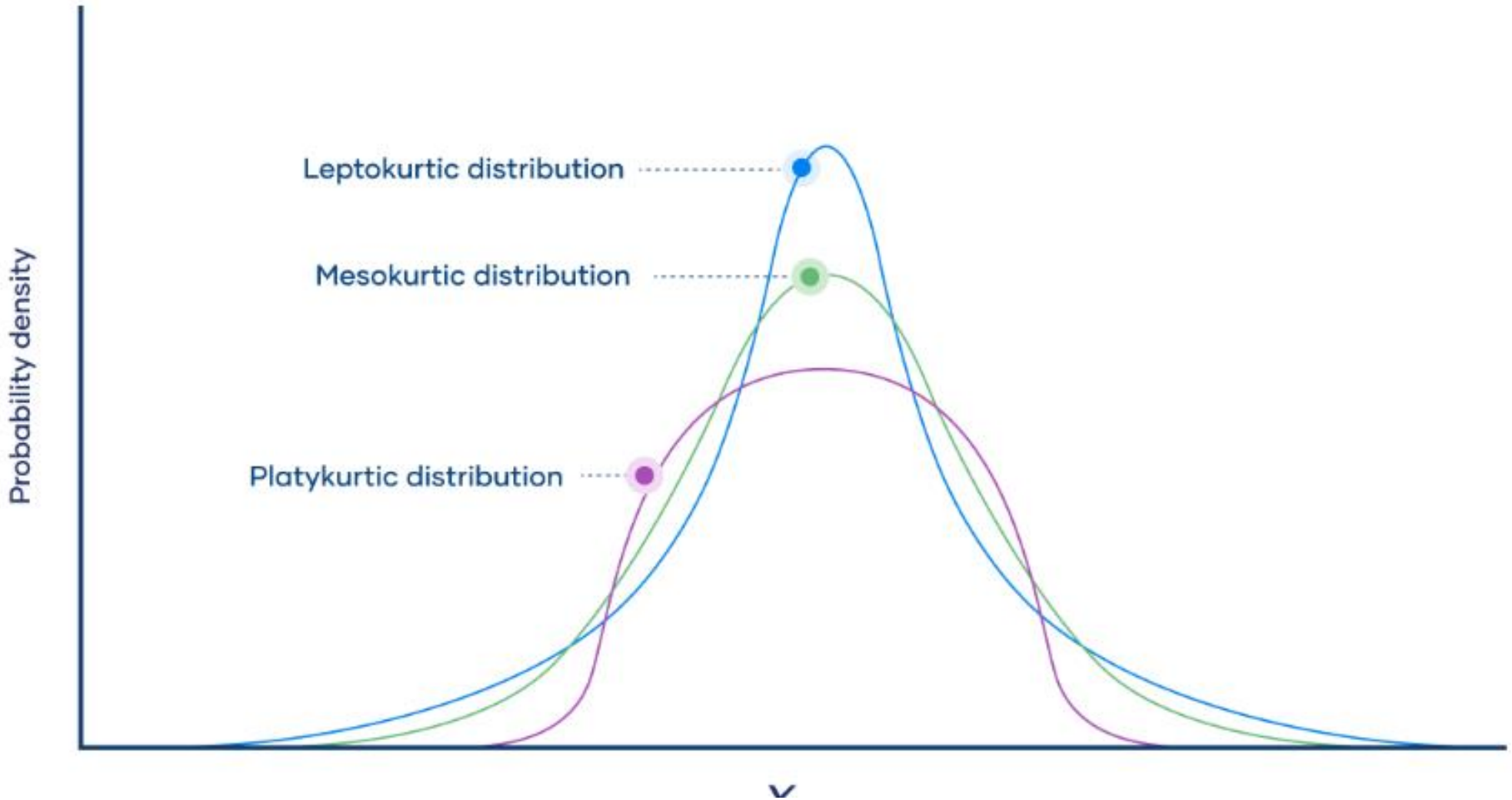
Platykurtic, Kurtosis < 3, Excess Kurtosis (-ve)

- A normal distribution has a kurtosis of 3 and is called mesokurtic.

	Category		
	Mesokurtic	**Platykurtic**	**Leptokurtic**
Tailedness	Medium-tailed	Thin-tailed	Fat-tailed
Outlier frequency	Medium	Low	High
Kurtosis	Moderate (3)	Low (< 3)	High (> 3)
Excess kurtosis	0	Negative	Positive

- **excess kurtosis=** kurtosis − 3.
- The greater the value more the peakedness.

Kurtosis


Leptokurtic distribution
Mesokurtic distribution
Platykurtic distribution
Probability density

Kurtosis

The Kurtosis of a given set of ungrouped data values can be calculated using the formula,

Ungrouped Data Kurtosis Calculation

$$\text{Kurtosis} = \frac{\sum(x_i - \bar{x})^4}{n\sigma^4}$$

where, $\bar{x}$ denotes the mean and σ denotes the standard deviation.

Grouped Data Kurtosis Calculation

$$\text{Kurtosis} = \frac{\sum_{i=1}^{n} f_i(x_i - \bar{x})^4}{n\sigma^4}$$

Frequency count=f_i

Example 1: Kurtosis for Ungrouped Data

Consider the following set of ungrouped data values,

$$23, 34, 38, 47, 59, 63, 84.$$

We first calculate the values of the mean and the standard deviation.

$$\bar{x} = \frac{\sum x_i}{n} = \frac{23 + 34 + 38 + 47 + 59 + 63 + 84}{7} = \frac{348}{7} = 49.7143.$$

x_i	$(x_i - \bar{x})$ $=(x_i - 49.7143)$	$(x_i - \bar{x})^2$ $=(x_i - 49.7143)^2$	$(x_i - \bar{x})^3$ $=(x_i - 49.7143)^3$	$(x_i - \bar{x})^4$ $=(x_i - 49.7143)^4$
23	-26.714	713.653	-19065	509301
34	-15.714	246.939	-3880.5	60978.8
38	-11.714	137.225	-1607.5	18830.6
47	-2.7143	7.3673	-19.997	54.2778
59	9.2857	86.2245	800.656	7434.66
63	13.2857	176.51	2345.06	31155.9
84	34.2857	1175.51	40303.2	1381824
TOTAL = 348	TOTAL = 0	TOTAL = 2543.43	TOTAL = 18876.2	TOTAL = 2009579

$$\text{Standard Deviation } \sigma = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n}} = \sqrt{\frac{2543.3}{7}} = 19.0617$$

We now calculate the Kurtosis using the formula,

$$\text{Kurtosis} = \frac{\sum(x_i - \bar{x})^4}{n\sigma^4}$$

$$\text{Kurtosis} = \frac{2009579}{7 \times 19.0617^4} = 2.1745$$

- Excess Kurtosis= kurtosis-3=2.1745-3=-0.8255
- Platykurtic
- Tailedness is Thin-tailed and outlier frequency is low

Example 2: Grouped Data Kurtosis Calculation

Consider the following set of data values given in the form of a grouped frequency distribution table.

Class Intervals	Frequency
0-5	2
5-10	3
10-15	1
15-20	4
20-25	5
25-30	9
30-35	6
35-40	12
40-45	8
45-50	7

We calculate the values required to calculate the mean and the standard deviation,

Classes	Class Mark (x_i)	Frequency (f_i)	$f_i x_i$	$x_i - \bar{x}$	$f_i(x_i - \bar{x})^2$	$f_i(x_i - \bar{x})^4$
0 – 5	2.5	2	5	−28.86	1665.76	1387376
5 – 10	7.5	3	22.5	−23.86	1707.85	972249
10 – 15	12.5	1	12.5	−18.86	355.686	126513
15 – 20	17.5	4	70	−13.86	768.36	147594
20 – 25	22.5	5	112.5	−8.8596	392.467	30806.1
25 – 30	27.5	9	247.5	−3.8596	134.072	1997.26
30 – 35	32.5	6	195	1.1404	7.8024	10.1462
35 – 40	37.5	12	450	6.1404	452.447	17059
40 – 45	42.5	8	340	11.1404	992.859	123221
45 – 50	47.5	7	332.5	16.1404	1823.58	475062
		n=57	$\sum f_i x_i = 1787.5$		$\sum f_i(x_i - \bar{x})^2 = 8300.877$	$\sum f_i(x_i - \bar{x})^4 = 3281887$

$$\text{Mean } \bar{x} = \frac{\sum f_i x_i}{n} = \frac{1787.5}{57} = 31.3596$$

$$\text{Standard Deviation } \sigma = \sqrt{\frac{\sum f_i (x_i - \bar{x})^2}{n}} = \sqrt{\frac{8300.8772}{57}} = 12.0677$$

The formula for calculating kurtosis for a set of grouped data values is as follows,

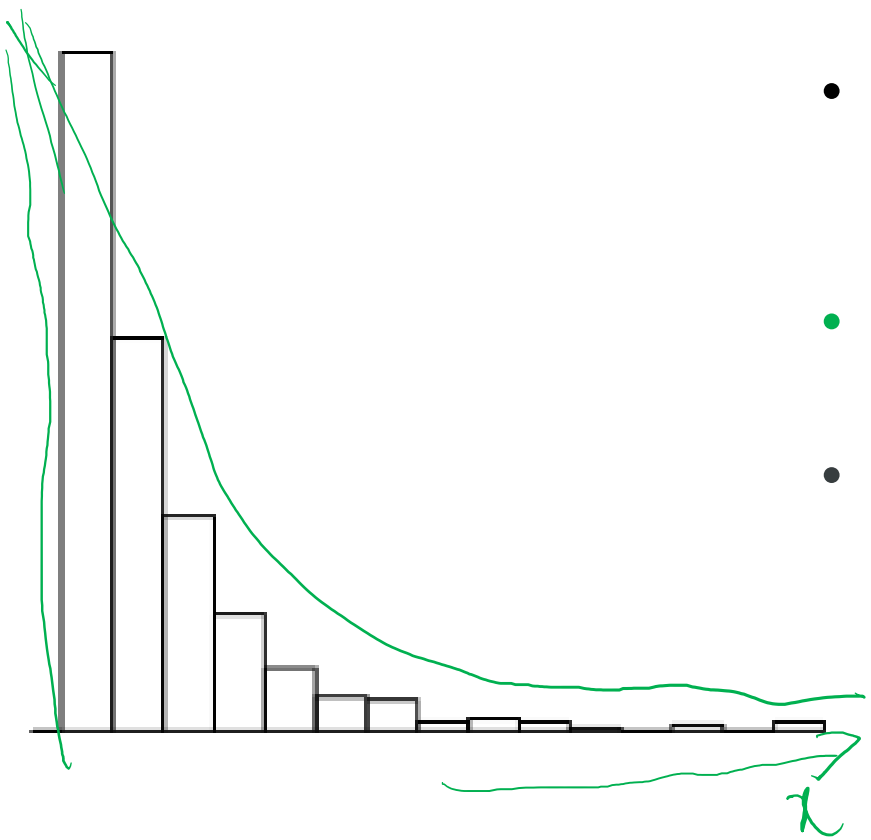
$$\text{Kurtosis } = \frac{\sum_{i=1}^{n} f_i (x_i - \bar{x})^4}{n\sigma^4} = \frac{3281887.0786}{57 \times 12.0677^4} = 2.7149$$

- Excess Kurtosis= kurtosis-3=2.7149-3=-0.2851
- Platykurtic
- **Tailedness is Thin-tailed and outlier frequency is low**

Note

Although a population's probability distribution can have a kurtosis of exactly 3, real data is almost always at least slightly platykurtic or leptokurtic.

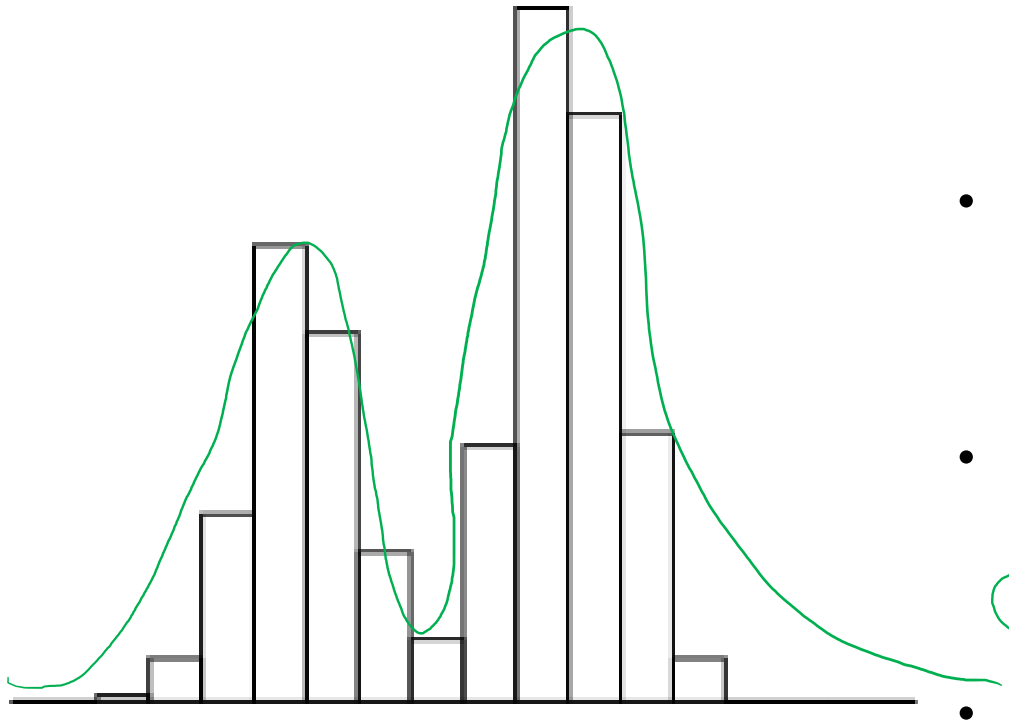
If a sample has a kurtosis of approximately 3, you can assume it's drawn from a mesokurtic population (i.e. normal population.



Exponential

In a feature following an **exponential distribution** the likelihood of occurrence of a small number of low values is very high, but sharply diminishes as values increase.

- **Values for an exponential random variable occur in the following way**.
- There are fewer large values and more small values.
- For example, the amount of money customers spend in one trip to the supermarket follows an exponential distribution. There are more people who spend small amounts of money and fewer people who spend large amounts of money.
- The value of the change that you have in your pocket or purse approximately follows an exponential distribution



Multimodal

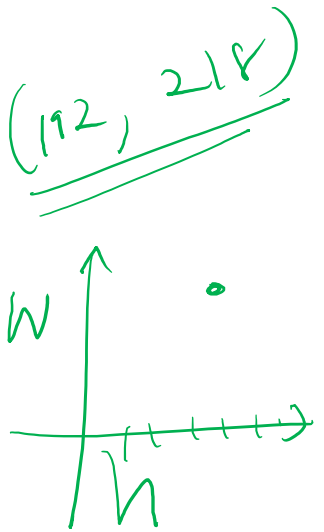
A feature characterized by a multimodal distribution has two or more very commonly occurring ranges of values that are clearly separated.

- Figure shows a bi-modal distribution with two clear peaks—we can think of this as two normal distributions pushed together.
- Multimodal distributions tend to occur when a feature contains a measurement made across a number of distinct groups.
- For example, if we were to measure the heights of a randomly selected group of Irish men and women, we would expect a bi-modal distribution with a peak at around 1.635m for women and 1.775m for men.

Advanced Data Exploration

The details of a professional basketball teams

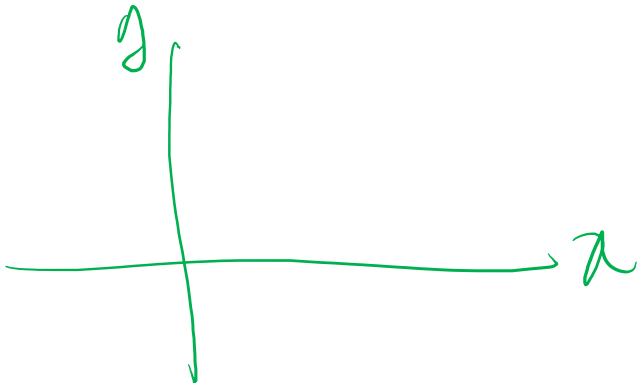
ID	Position	Height	Weight	Career Stage	Age	Sponsorship Earnings	Shoe Sponsor
1	forward	192	218	veteran	29	561	yes
2	center	218	251	mid-career	35	60	no
3	forward	197	221	rookie	22	1,312	no
4	forward	192	219	rookie	22	1,359	no
5	forward	198	223	veteran	29	362	yes
6	guard	166	188	rookie	21	1,536	yes
7	forward	195	221	veteran	25	694	no
8	guard	182	199	rookie	21	1,678	yes
9	guard	189	199	mid-career	27	385	yes
10	forward	205	232	rookie	24	1,416	no
11	center	206	246	mid-career	29	314	no
12	guard	185	207	rookie	23	1,497	yes
13	guard	172	183	rookie	24	1,383	yes
14	guard	169	183	rookie	24	1,034	yes
15	guard	185	197	mid-career	29	178	yes
16	forward	215	232	mid-career	30	434	no
17	guard	158	184	veteran	29	162	yes
18	guard	190	207	mid-career	27	648	yes
19	center	195	235	mid-career	28	481	no
20	guard	192	200	mid-career	32	427	yes
21	forward	202	220	mid-career	31	542	no
22	forward	184	213	mid-career	32	12	no
23	forward	190	215	rookie	22	1,179	no
24	guard	178	193	rookie	21	1,078	no
25	guard	185	200	mid-career	31	213	yes
26	forward	191	218	rookie	19	1,855	no
27	center	196	235	veteran	32	47	no
28	forward	198	221	rookie	22	1,409	no
29	center	207	247	veteran	27	1,065	no
30	center	201	244	mid-career	25	1,111	yes

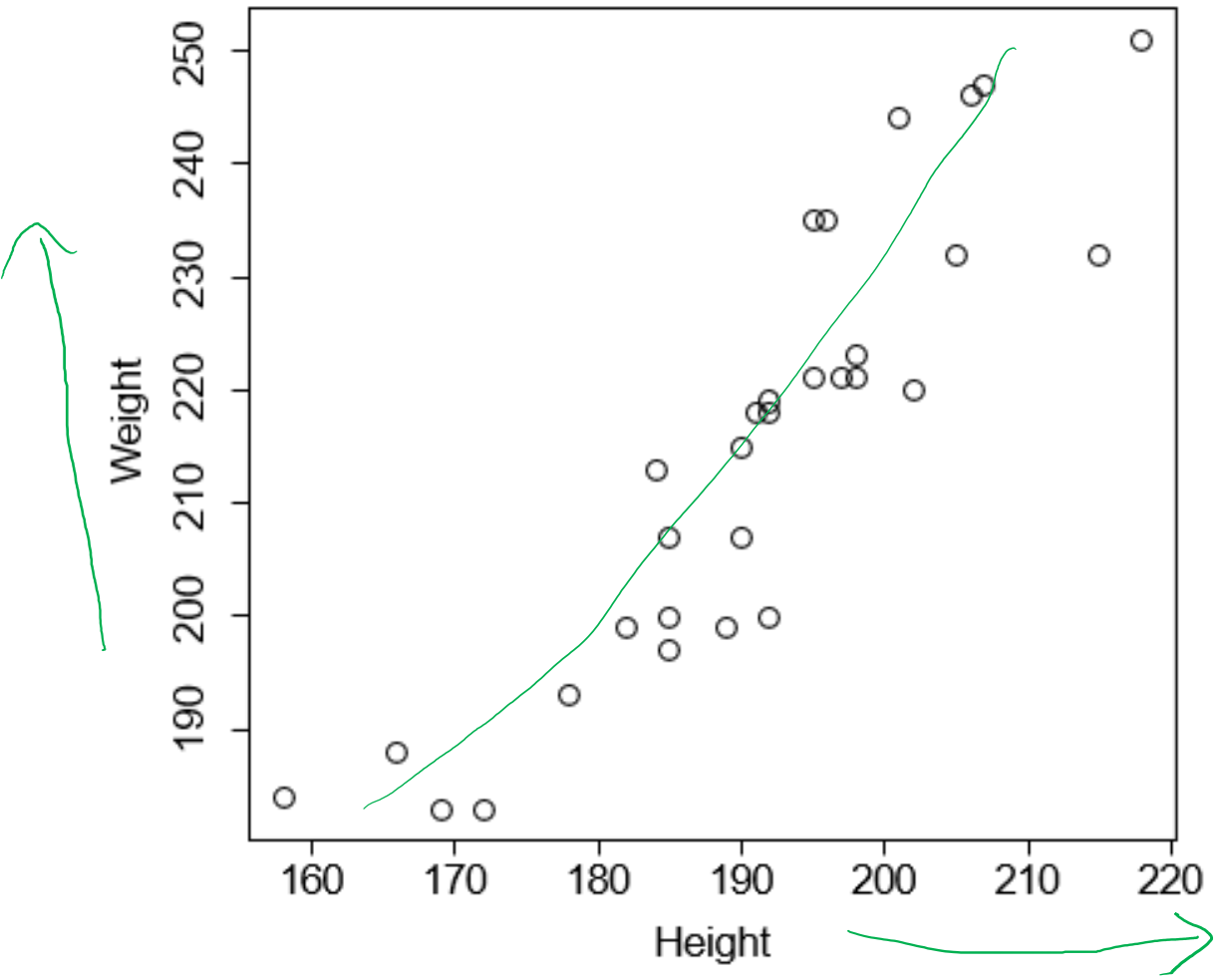


scatter plot

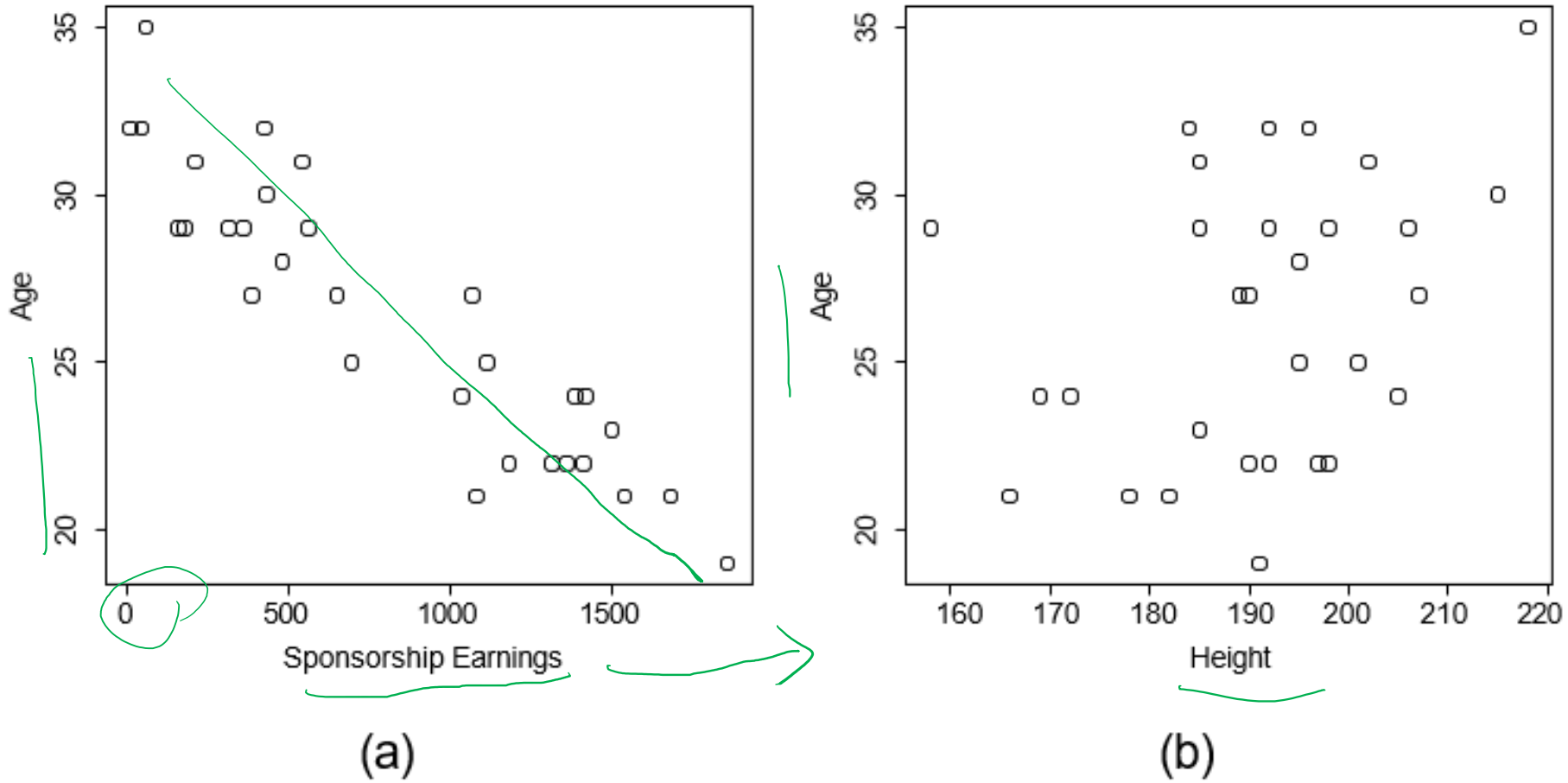
- A **scatter plot** is based on two axes: the horizontal axis represents one feature and the vertical axis represents a second.

- Each instance in a dataset is represented by a point on the plot determined by the values for that instance of the two features involved.

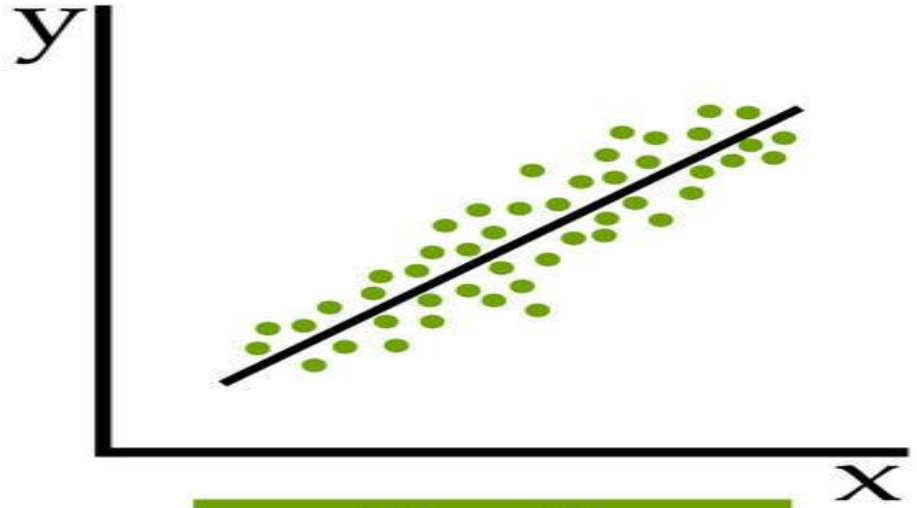




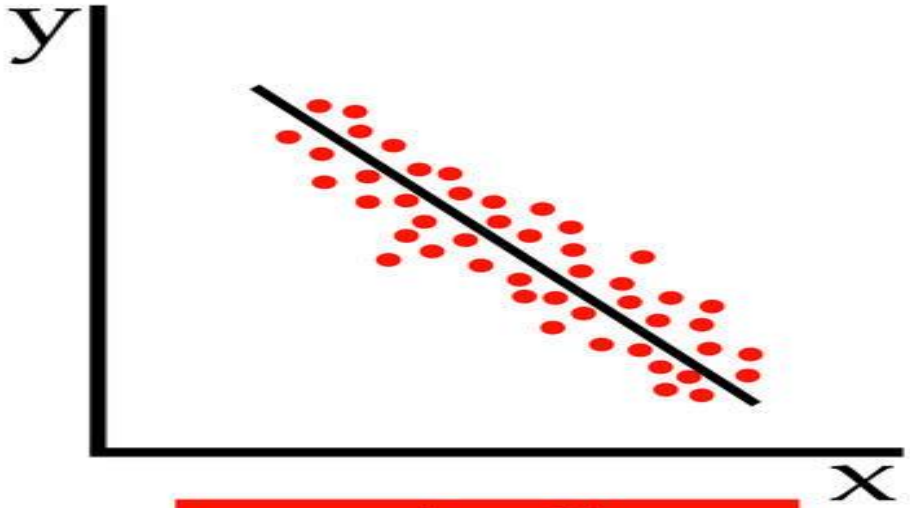
An example scatter plot showing the relationship between the HEIGHT and WEIGHT features from the professional basketball squad dataset



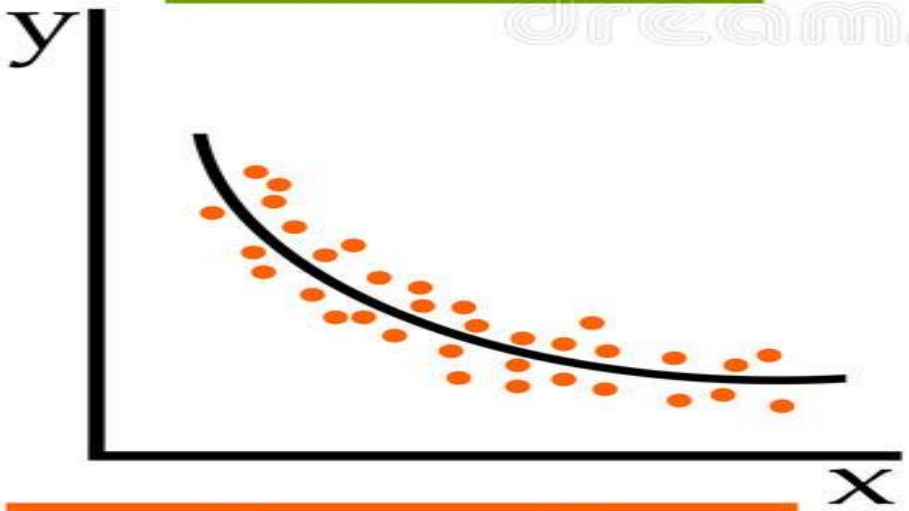
- Example scatter plots showing:
  - (a) the strong negative covariance between the SPONSORSHIP EARNINGS and AGE features and
  - (b) the HEIGHT and AGE features from the dataset



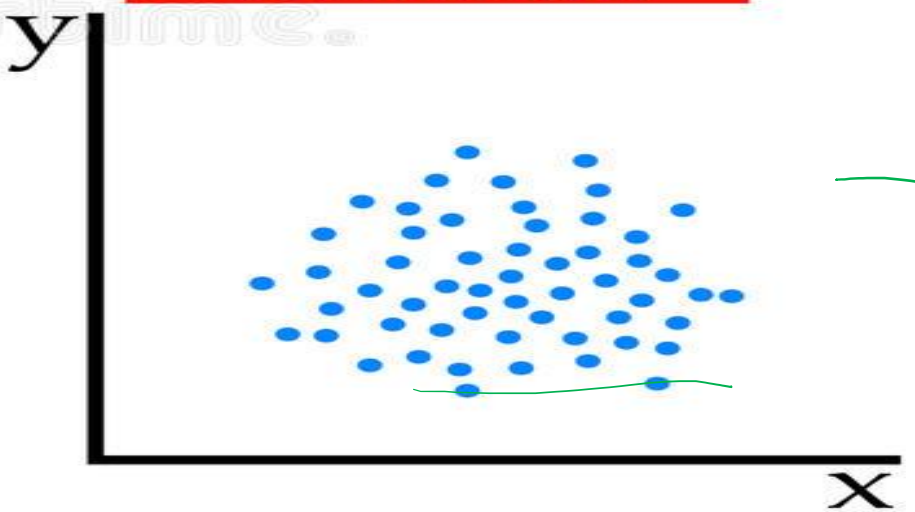
positive linear correlation



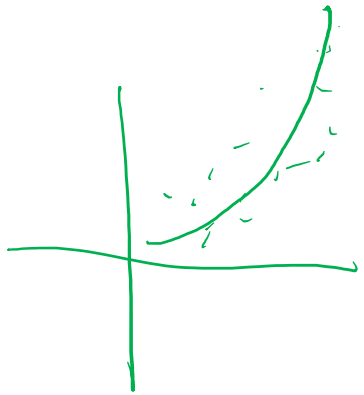
negative linear correlation



negative non-linear correlation



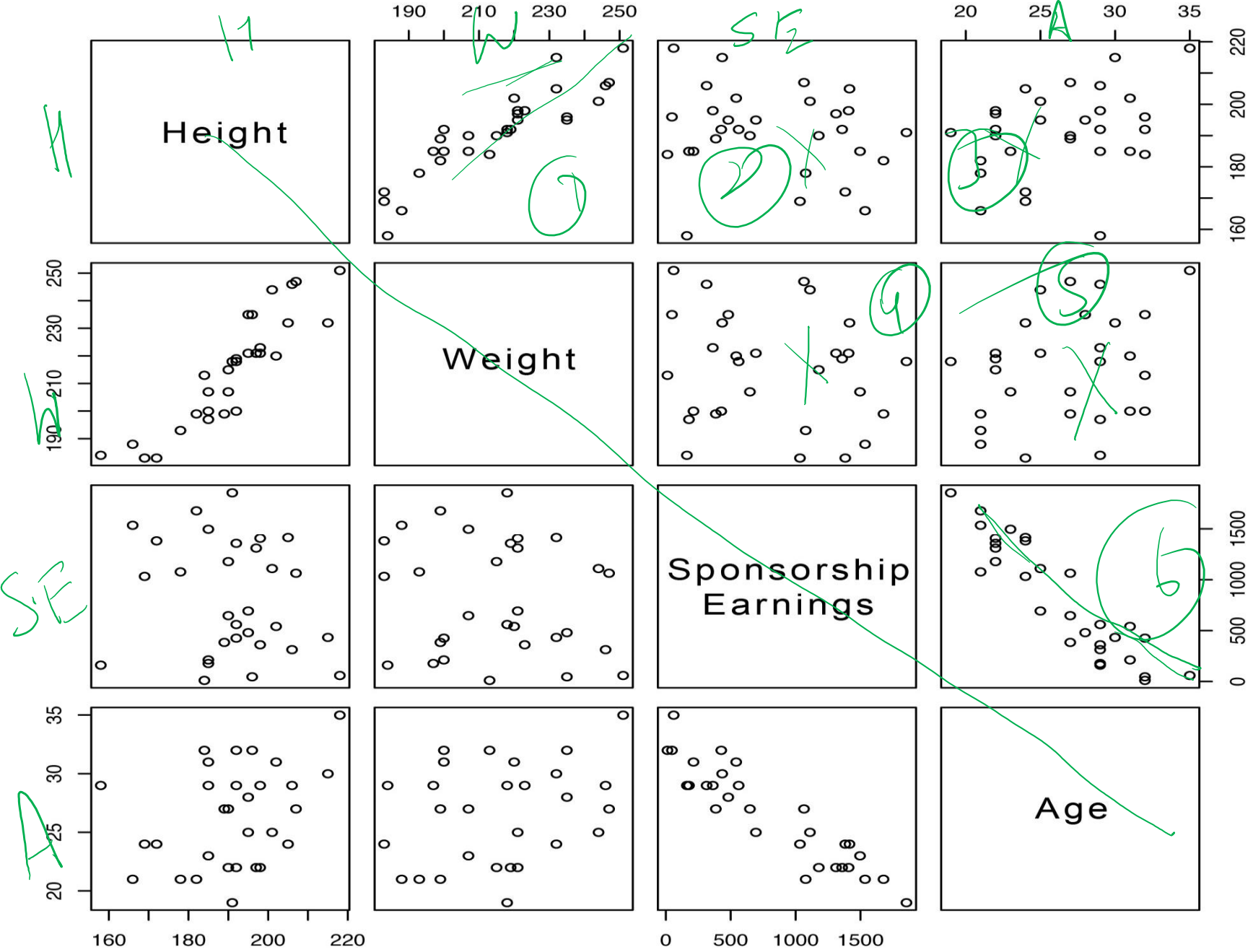
no correlation



Scatter Plot Matrix (SPLOM)

- A **scatter plot matrix (SPLOM)** shows scatter plots for a whole collection of features arranged into a matrix.

- This is useful for exploring the relationships between groups of features - for example all of the continuous features in an ABT.



A scatter plot matrix showing scatter plots of the continuous features from the professional basketball squad dataset.

The simplest way to visualize the relationship between two categorical variables is to use a collection of **small multiple** bar plots

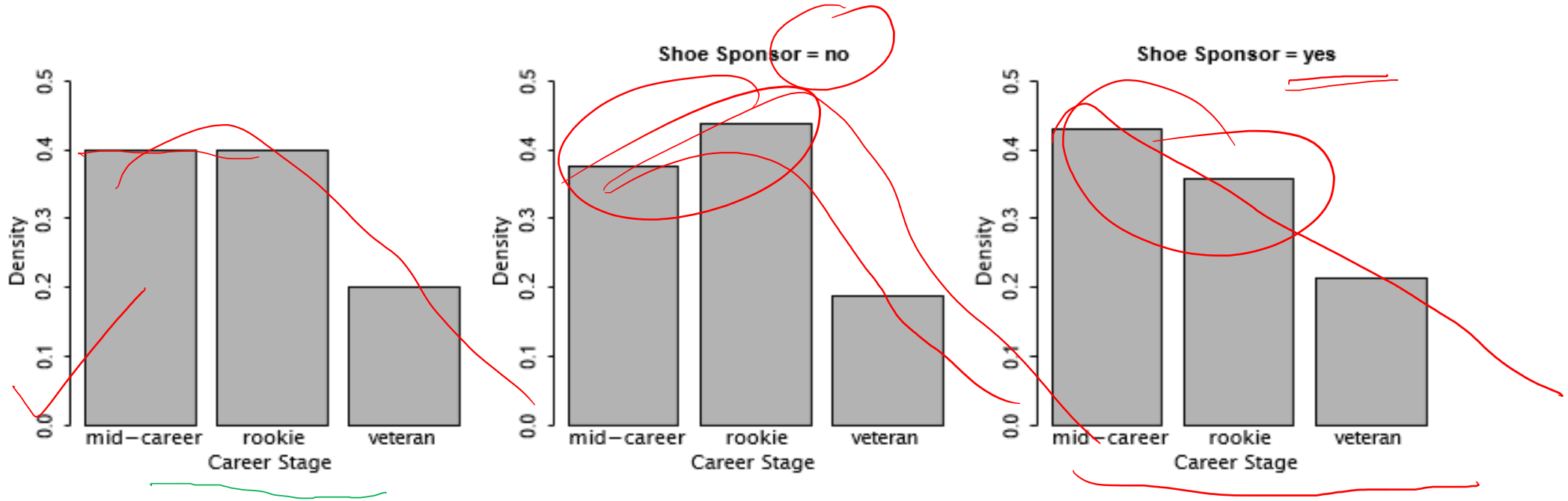


Figure: Using small multiple bar plot visualizations to illustrate the relationship between the CAREER STAGE and SHOE SPONSOR features.

- ✓ The bar plot on the left shows the distribution of the different levels of the CAREER STAGE feature across the entire dataset. The two plots on the right show the distributions for those players with and without a shoe sponsor.
- ✓ Since all three plots show very similar distributions, we can conclude that no real relationship exists between these two features and that players of any career stage are equally likely to have a shoe sponsor or not.

Figure: Using small multiple bar plot visualizations to illustrate the relationship between the POSITION and SHOE SPONSOR features

- ✓ In this case, the three plots are very different, so we can conclude that there is a relationship between these two features. It seems that players who play in the guard position are much more likely to have a shoe sponsor than forwards or centers.

Visualizing a Categorical Feature and a Continuous Feature

- ✓ To visualize the relationship between a continuous feature and a categorical feature a **small multiples** approach that draws a histogram of the values of the continuous feature for each level of the categorical feature is useful.

- ✓ Each histogram includes only those instances in the dataset that have the associated level of the categorical feature. Similar to using small multiples for categorical features, if the features are unrelated (or independent) then the histograms for each level should be very similar.

- ✓ If the features are related, however, then the shapes and/or the central tendencies of the histograms will be different.

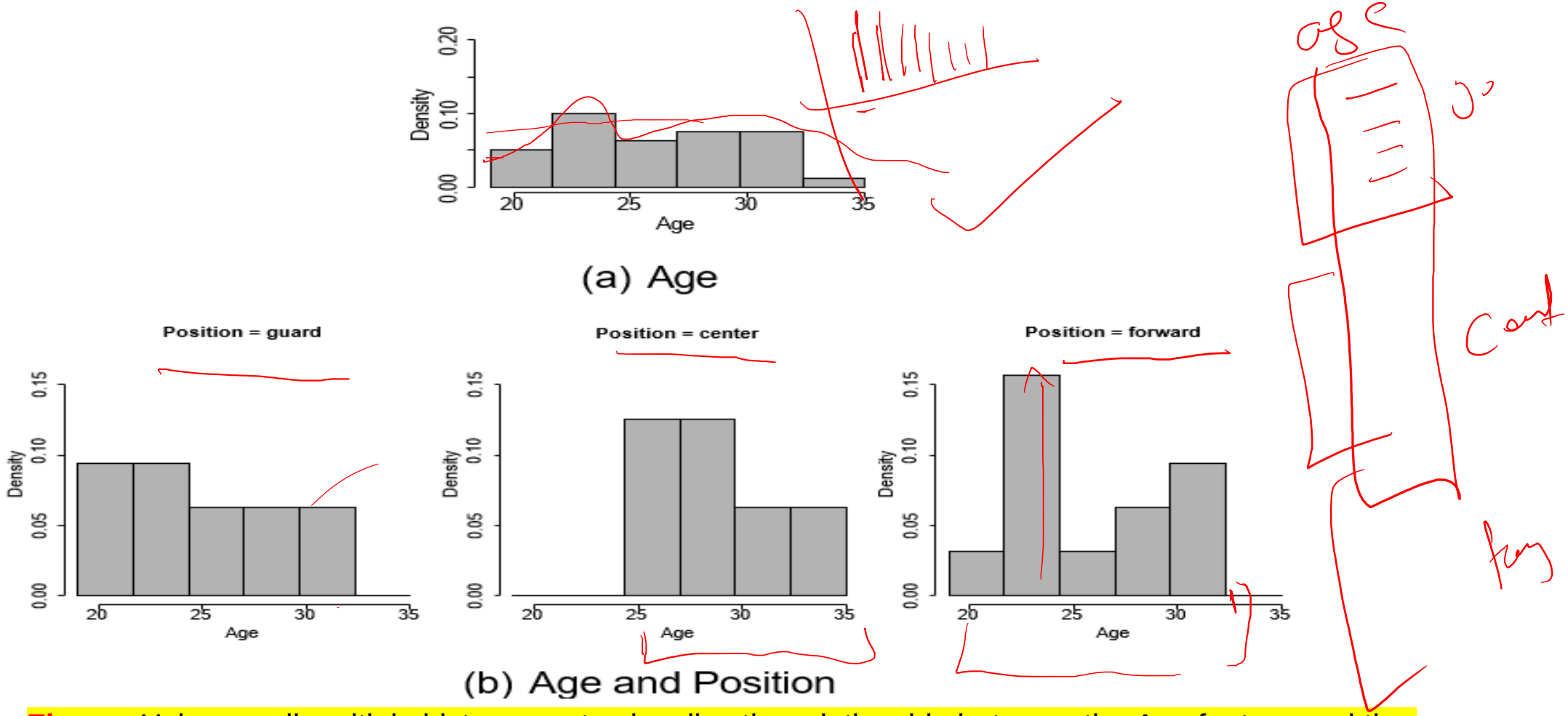
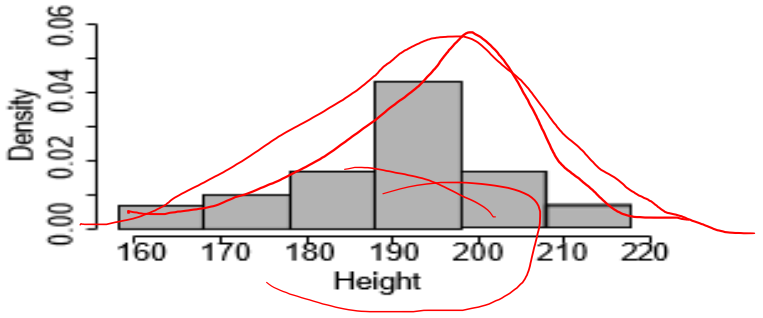
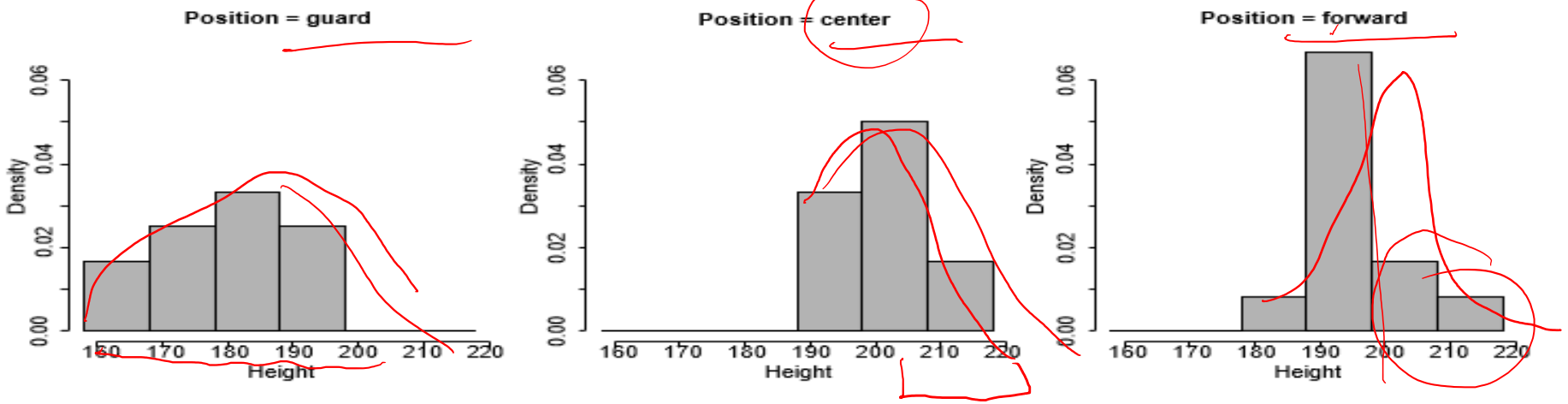


Figure: Using small multiple histograms to visualize the relationship between the AGE feature and the POSITION FEATURE.

- ✓ It shows small multiple histograms for values of AGE broken down by the different levels of the POSITION feature. These histograms show a slight tendency for centers to be a little older than guards and forwards, but the relationship does not appear very strong as each of the smaller histograms are similar to the overall uniform distribution of the AGE feature



(a) Height



(b) Height and Position

Figure: Using small multiple histograms to visualize the relationship between the HEIGHT feature and the POSITION feature.

- ✓ HEIGHT follows a normal distribution centered around a mean of approximately 194. The three smaller histograms depart from this distribution and suggest that centers tend to be taller than forwards, who in turn tend to be taller than guards.

- ✓ A second approach to visualizing the relationship between a categorical feature and a continuous feature is to use a collection of box plots.

- ✓ For each level of the categorical feature a box plot of the corresponding values of the continuous feature is drawn.

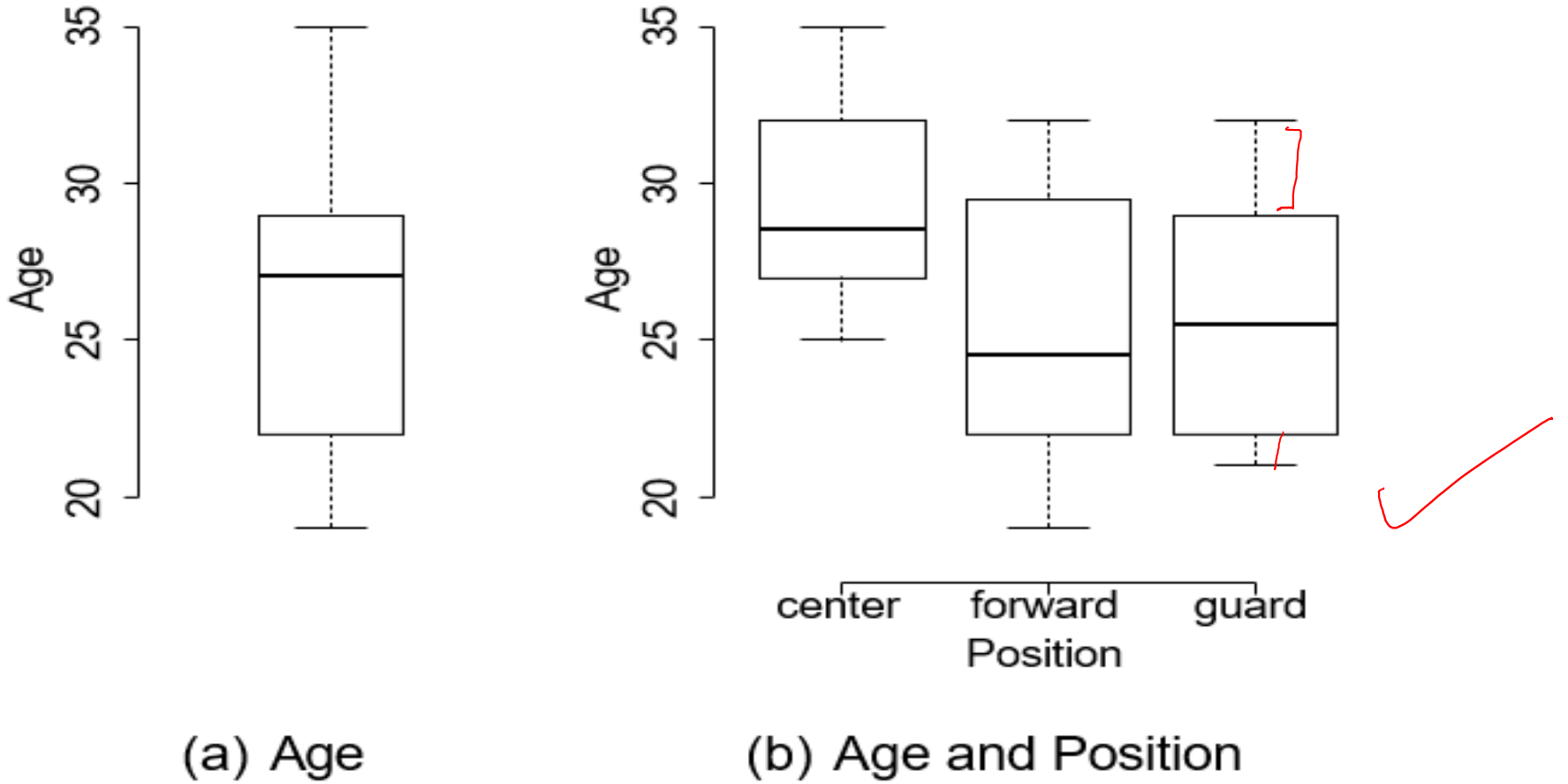


Figure: Using box plots to visualize the relationship between the AGE and the POSITION feature.

- ✓ This visualization shows a slight indication that centers tend to be older than forwards and guards, but the three box plots overlap significantly, suggesting that this relationship is not very strong.

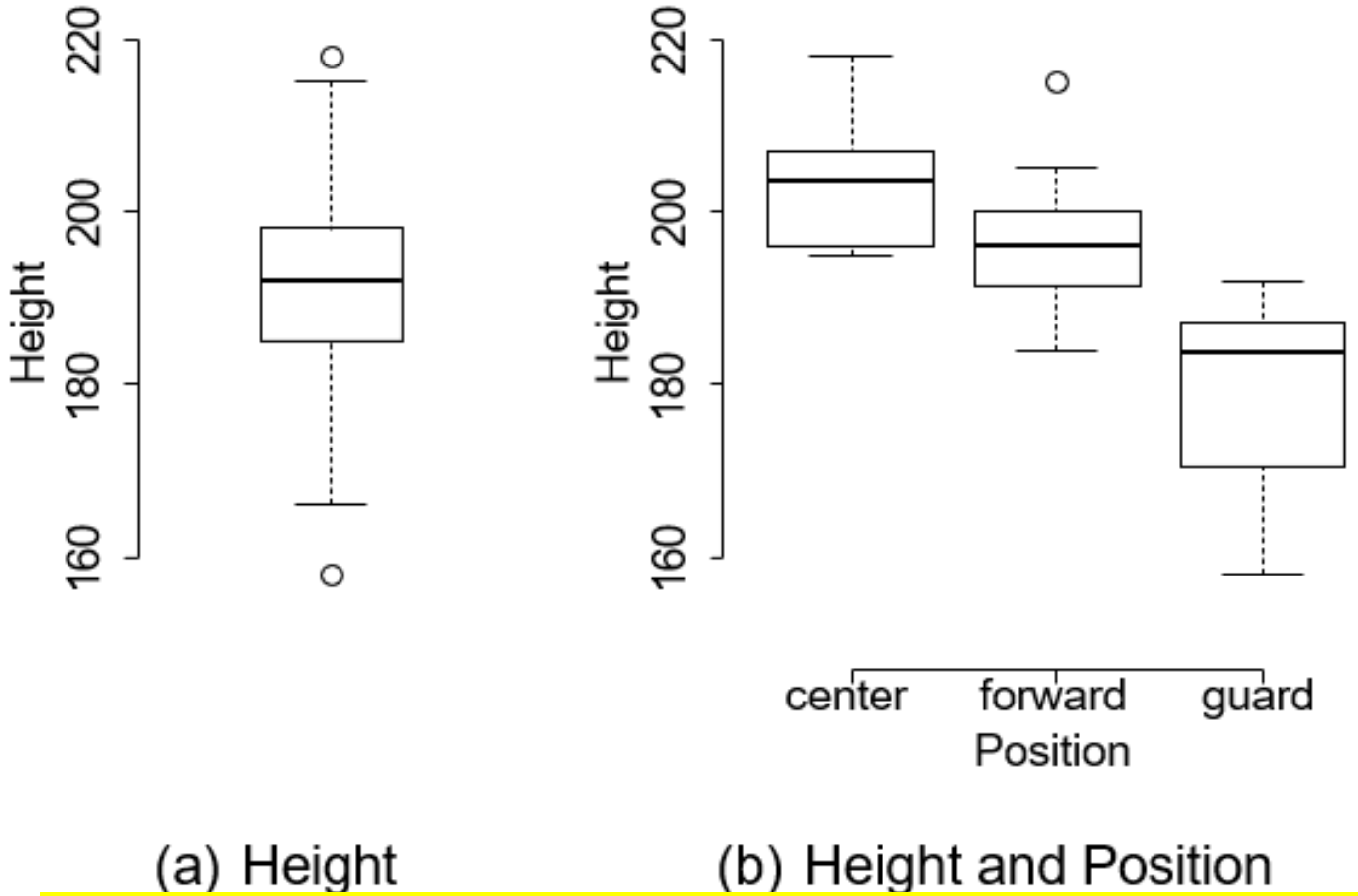


Figure: Using box plots to visualize the relationship between the HEIGHT feature and the POSITION feature.

✓ The average height of centers is above that of forwards, which in turn is above that of guards. Although the whiskers show that there is some overlap between the three groups, they do appear to be well separated.

covariance and correlation.

- As well as visually inspecting scatter plots, we can calculate formal measures of the relationship between two continuous features using **covariance** and **correlation**.
- For two features, a and b, in a dataset of n instances, the **sample covariance** between a and b is

$$cov(a,b) = \frac{1}{n-1}\sum_{i=1}^{n}\left((a_i - \bar{a}) \times (b_i - \bar{b})\right) \quad (1)$$

- where a_i and b_i are values of features a and b for the i^{th} instance in a dataset, and $\bar{a}$ and $\bar{b}$ are the sample means of features a and b.
- Covariance values fall into the **range $[-\infty, \infty]$** where negative values indicate a negative relationship, positive values indicate a positive relationship, and values near zero indicate that there is little or no relationship between the features

Calculating covariance between the HEIGHT feature and the WEIGHT and AGE features from the basketball players dataset.

ID	HEIGHT (h)	$h-\bar{h}$	WEIGHT (w)	$w-\bar{w}$	$(h-\bar{h})\times(w-\bar{w})$	AGE (a)	$a-\bar{a}$	$(h-\bar{h})\times(a-\bar{a})$
1	192	0.9	218	3.0	2.7	29	2.6	2.3
2	218	26.9	251	36.0	967.5	35	8.6	231.3
3	197	5.9	221	6.0	35.2	22	-4.4	-26.0
4	192	0.9	219	4.0	3.6	22	-4.4	-4.0
5	198	6.9	223	8.0	55.0	29	2.6	17.9
				...				
26	191	-0.1	218	3.0	-0.3	19	-7.4	0.7
27	196	4.9	235	20.0	97.8	32	5.6	27.4
28	198	6.9	221	6.0	41.2	22	-4.4	-30.4
29	207	15.9	247	32.0	508.3	27	0.6	9.5
30	201	9.9	244	29.0	286.8	25	-1.4	-13.9
Mean	191.1		215.0			26.4		
Std Dev	13.6		19.8			4.2		
Sum					7,009.9			570.8

Calculating covariance between the HEIGHT feature and the WEIGHT and AGE features from the basketball players dataset.

$$cov(\text{HEIGHT}, \text{WEIGHT}) = \frac{7{,}009.9}{29} = 241.72$$

$$cov(\text{HEIGHT}, \text{AGE}) = \frac{570.8}{29} = 19.7$$

$n = 30$

$n - 1 = 29$

- ✓ **Correlation (Pearson correlation coefficient)** is a normalized form of covariance that ranges between −1 and +1.
- ✓ The correlation between two features, *a* and *b*, can be calculated as

$$corr(a,b) = \frac{cov(a,b)}{sd(a) \times sd(b)} \quad (2)$$

strength and direction of the linear relationship between your two variables,

- ✓ where $cov(a, b)$ is the covariance between features *a* and *b* and $sd(a)$ and $sd(b)$ are the standard deviations of *a* and *b* respectively.

- ✓ Correlation values fall into the range [−1, 1], where values close to −1 indicate a very strong negative correlation (or covariance), values close to 1 indicate a very strong positive correlation, and values around 0 indicate no correlation.
- ✓ Features that have no correlation are said to be independent (not always true)

Calculating correlation between the HEIGHT feature and the WEIGHT and AGE features from the basketball players dataset.

$$corr\,(Height, Weight) = \frac{241.72}{13.6 \times 19.8} = 0.898$$

$$corr\,(Height, Age) = \frac{19.7}{13.6 \times 4.2} = 0.345$$

- ✓ In the majority of ABTs there are multiple continuous features between which we would like to explore relationships.
- ✓ Two tools that can be useful for this are the covariance matrix and the correlation matrix.

The covariance matrix, usually denoted as $\sum$, between set of continuous features, $\{a, b, \ldots, z\}$, is given as

$$\sum_{\{a,b,\ldots,z\}} = \begin{bmatrix} var(a) & cov(a,b) & \cdots & cov(a,z) \\ cov(b,a) & var(b) & \cdots & cov(b,z) \\ \vdots & \vdots & \ddots & \vdots \\ cov(z,a) & cov(z,b) & \cdots & var(z) \end{bmatrix}$$

Similarly, the correlation matrix is just a normalized version of the covariance matrix and shows the correlation between each pair of features:

$$\text{correlation matrix}_{\{a,b,\ldots,z\}} = \begin{bmatrix} corr(a,a) & corr(a,b) & \cdots & corr(a,z) \\ corr(b,a) & corr(b,b) & \cdots & corr(b,z) \\ \vdots & \vdots & \ddots & \vdots \\ corr(z,a) & corr(z,b) & \cdots & corr(z,z) \end{bmatrix}$$

Calculating covariances matrix for the HEIGHT feature and the WEIGHT and AGE features from the basketball players dataset.

$$\sum_{<Height, Weight, Age>} = \begin{bmatrix} 185.128 & 241.72 & 19.7 \\ 241.72 & 392.102 & 24.469 \\ 19.7 & 24.469 & 17.697 \end{bmatrix}$$

Calculating correlation matrix for the HEIGHT feature and the WEIGHT and AGE features from the basketball players dataset.

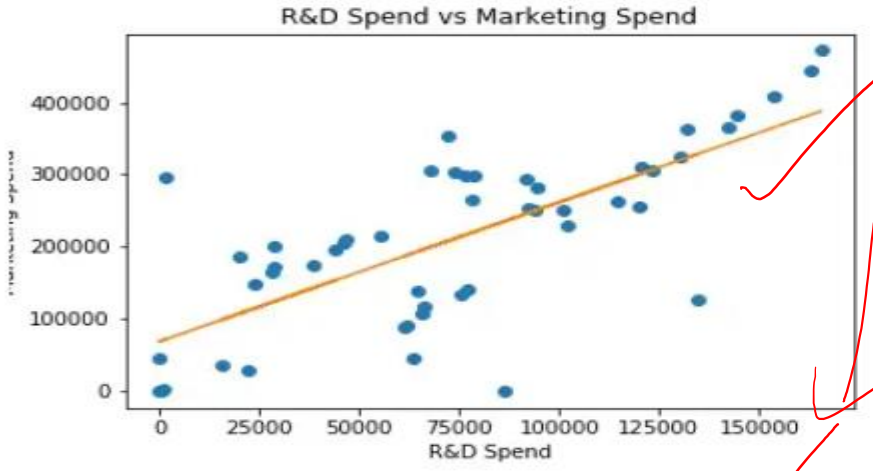
$$\underset{<Height, Weight, Age>}{\text{correlation matrix}} = \begin{bmatrix} 1.0 & 0.898 & 0.345 \\ 0.898 & 1.0 & 0.294 \\ 0.345 & 0.294 & 1.0 \end{bmatrix}$$

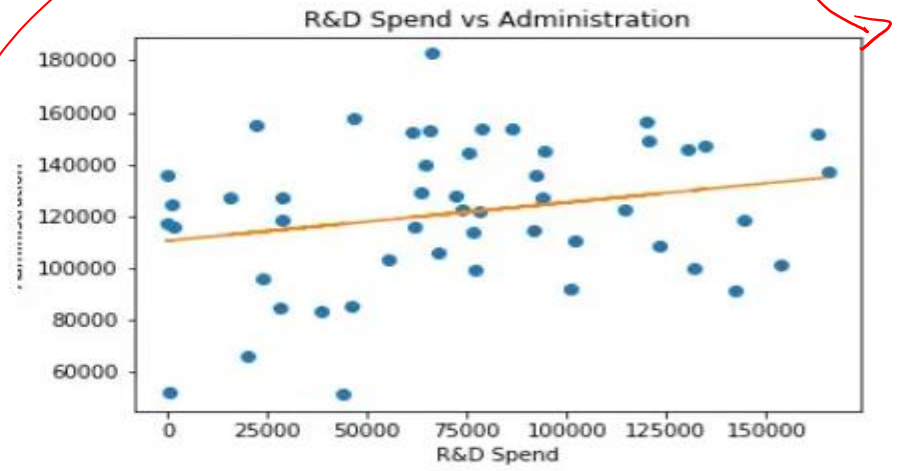
$A^T = A$

3 x 3

- ✓ The **scatter plot matrix** (SPLOM) is really a visualization of the correlation matrix.
- ✓ This can be made more obvious by including the correlation coefficients in SPLOMs in the cells above the diagonal.

Note: A high correlation between dependent and independent variables is desired whereas the high correlation between 2 independent variables is undesired.

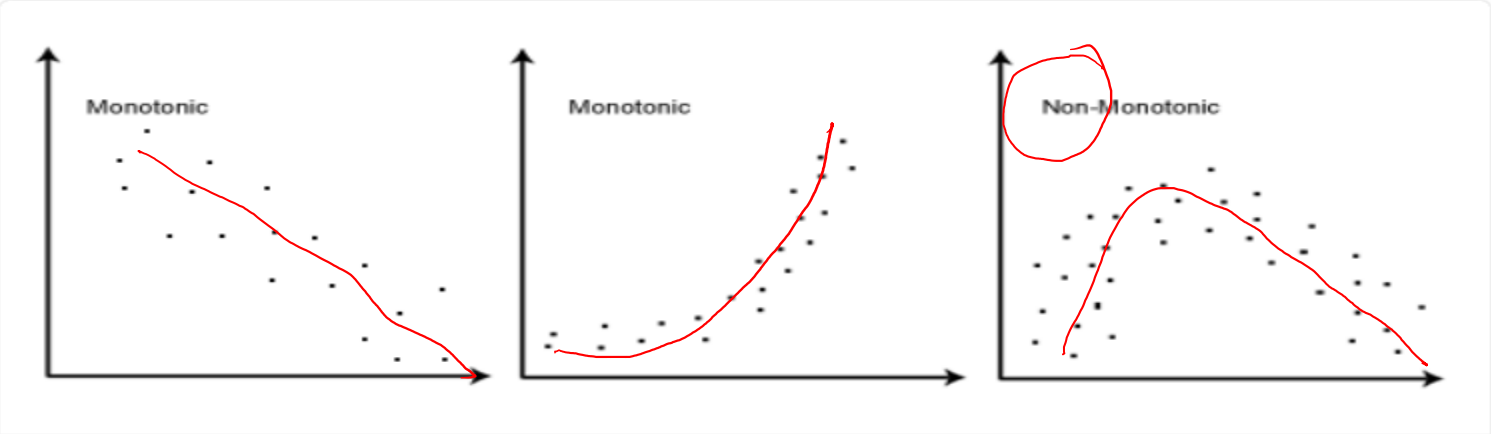




- ✓ The above 2 graphs show the correlation between independent variables. We can see a higher correlation in the first graph whereas very low correlation in the second.
- ✓ This means we can exclude any one of the 2 features in the first graph since the correlation between 2 independent variables causes redundancy.
- ✓ But which one to remove? The answer is straightforward. The variable with a higher correlation with the target variable stays and the other is removed.

- ✓ Spearman's correlation coefficient, (ρ, also signified by r_s) measures the strength and direction of association between two ranked variables.
- ✓ Spearman's correlation determines the strength and direction of the **monotonic relationship** between your two variables
- ✓ What is a monotonic relationship?

A monotonic relationship is a relationship that does one of the following: (1) as the value of one variable increases, so does the value of the other variable; or (2) as the value of one variable increases, the other variable value decreases.



- ✓ Why is a monotonic relationship important to Spearman's correlation?

Monotonicity is "less restrictive" than that of a linear relationship. For example, the middle image above shows a relationship that is monotonic, but not linear.

How to rank data?

	Marks									
English	56	75	45	71	61	64	58	80	76	61
Maths	66	70	40	60	65	56	59	77	67	63

English (mark)	Maths (mark)	Rank (English)	Rank (maths)
56	66	9	4
75	70	3	2
45	40	10	10
71	60	4	7
61	65	6.5	5
64	56	5	9
58	59	8	8
80	77	1	1
76	67	2	3
61	63	6.5	6

descending order

$\frac{6+7}{2} = \frac{13}{2}$

$= 6.5$

What is the definition of Spearman's rank-order correlation?

There are two methods to calculate Spearman's correlation depending on whether: (1) your data does not have tied ranks or (2) your data has tied ranks. The formula for when there are no tied ranks is:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

where d_i = difference in paired ranks and n = number of cases. The formula to use when there are tied ranks is:

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

where i = paired score.

Spearman's Rank-Order Correlation

	Marks									
English	56	75	45	71	62	64	58	80	76	61
Maths	66	70	40	60	65	56	59	77	67	63

English (mark)	Maths (mark)	Rank (English)	Rank (maths)	d	d^2
56	66	9	4	5	25
75	70	3	2	1	1
45	40	10	10	0	0
71	60	4	7	3	9
62	65	6	5	1	1
64	56	5	9	4	16
58	59	8	8	0	0
80	77	1	1	0	0
76	67	2	3	1	1
61	63	7	6	1	1

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

$\sum d_i^2 = 25 + 1 + 0 + 9 + 1 + 16 + 0 + 0 + 1 = 54$

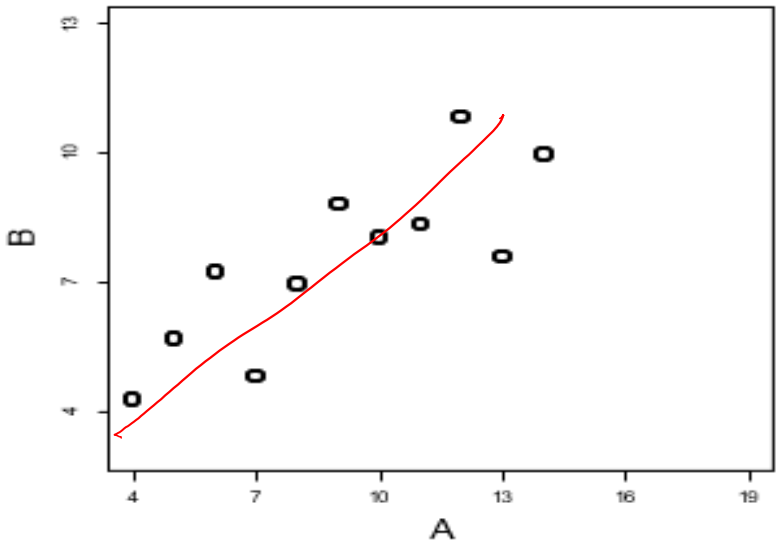
$\rho = 1 - \frac{6 \times 54}{10(10^2-1)}$

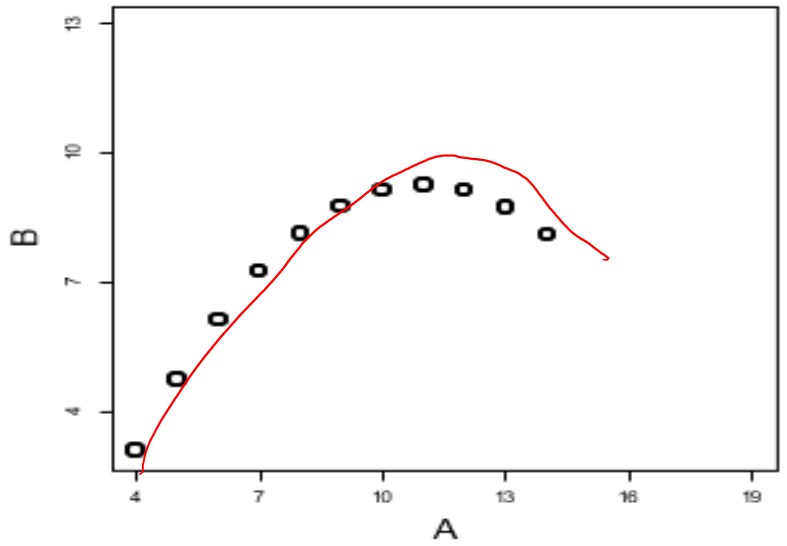
$= 1 - \frac{6 \times 54}{990}$

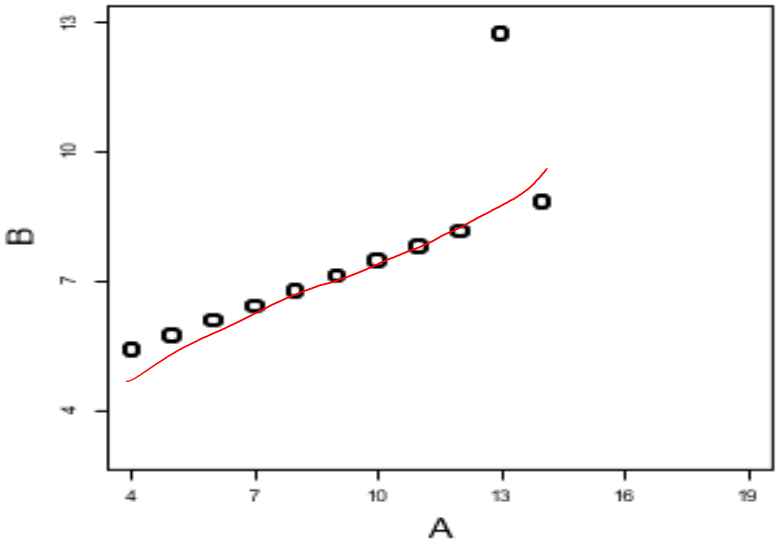
- ✓ Correlation is a good measure of the relationship between two continuous features, but it is not by any means perfect.
- ✓ Some of the limitations of measuring correlation are illustrated very clearly in the famous example of Anscombe's quartet by Francis Anscombe.
- ✓ It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties.

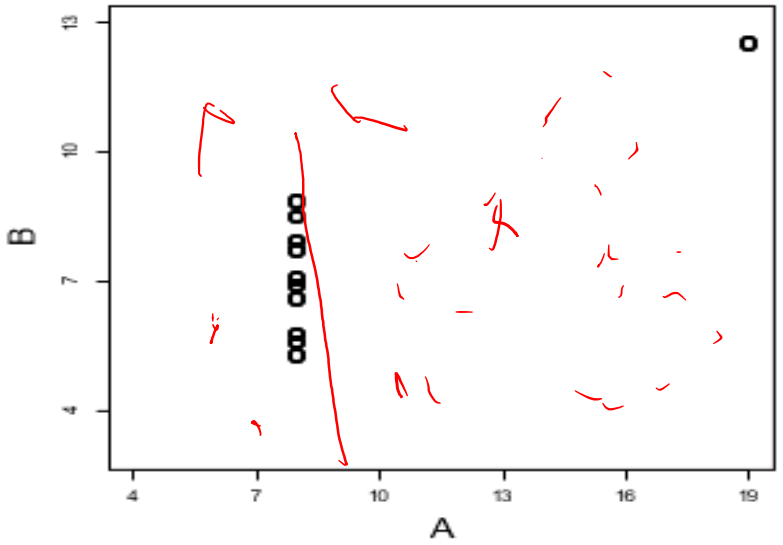
Anscombe's Data									
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89

Anscombe's Data											
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
				Summary Statistics							
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	









Perhaps the most important thing to remember in relation to correlation is that correlation does not necessarily imply causation.

The observed correlation could be due to the effects of a hidden third variable, or just entirely down to chance.

Takeaway

We have described the four datasets that were intentionally created to describe the importance of data visualisation and how any regression algorithm can be fooled by the same.

Hence, all the important features in the dataset must be visualised before implementing any machine learning algorithm on them which will help to make a good fit model.

Correlation Does Not Imply Causation

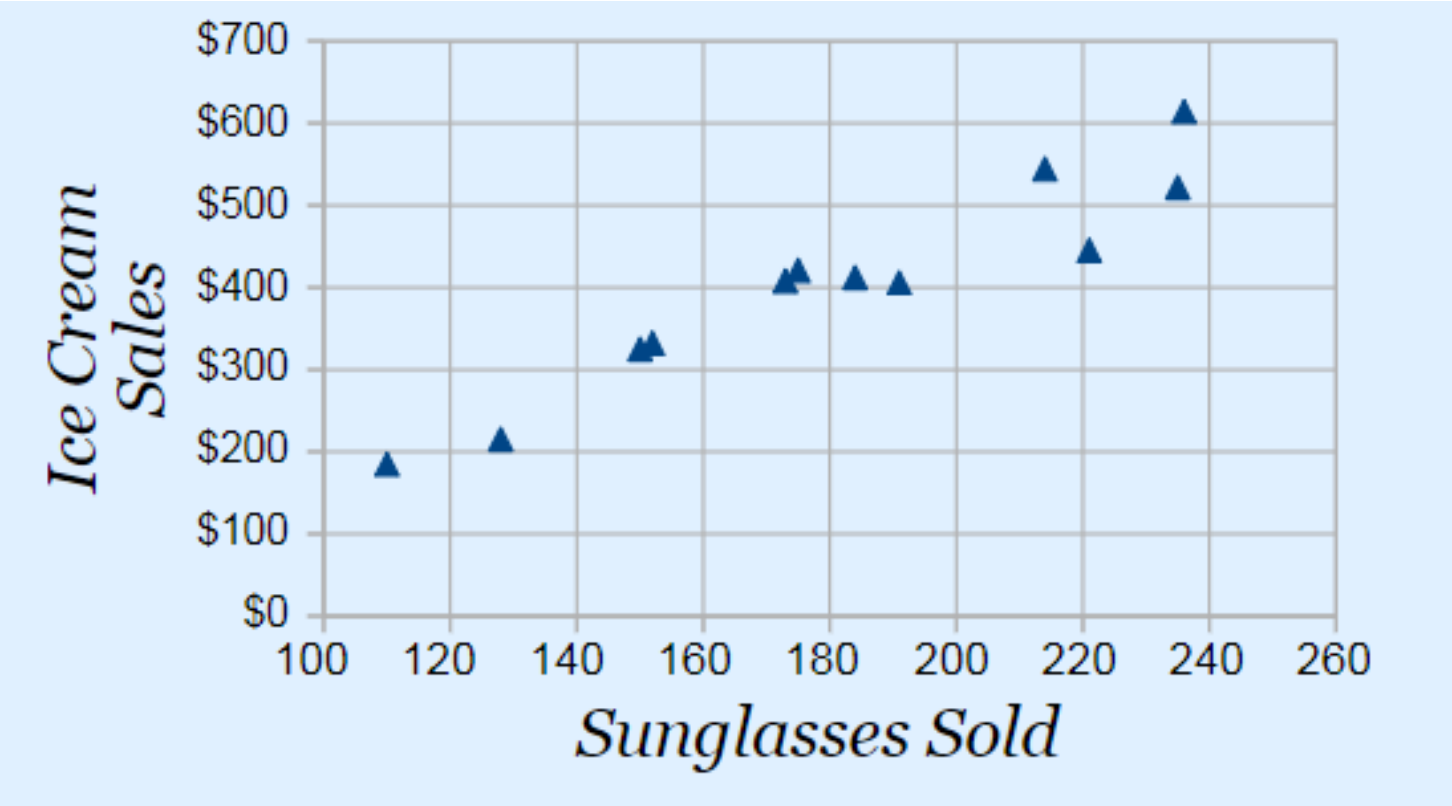
- ✓ When 2 unrelated things tied together, so these can be either bound by causality or correlation.

- ✓ The phrase "**correlation does not imply causation**" is often used in statistics to point out that correlation between two variables does not necessarily mean that one variable causes the other to occur.

- ✓ **Correlation** is a statistical technique which tells us how strongly the pair of variables are linearly related and change together. It does not tell us why and how behind the relationship, but it just says the relationship exists.

- ✓ **Example:** When a person is exercising then the amount of calories burning goes up every minute. Former is causing latter to happen

Example: Correlation between Ice cream sales and sunglasses sold.



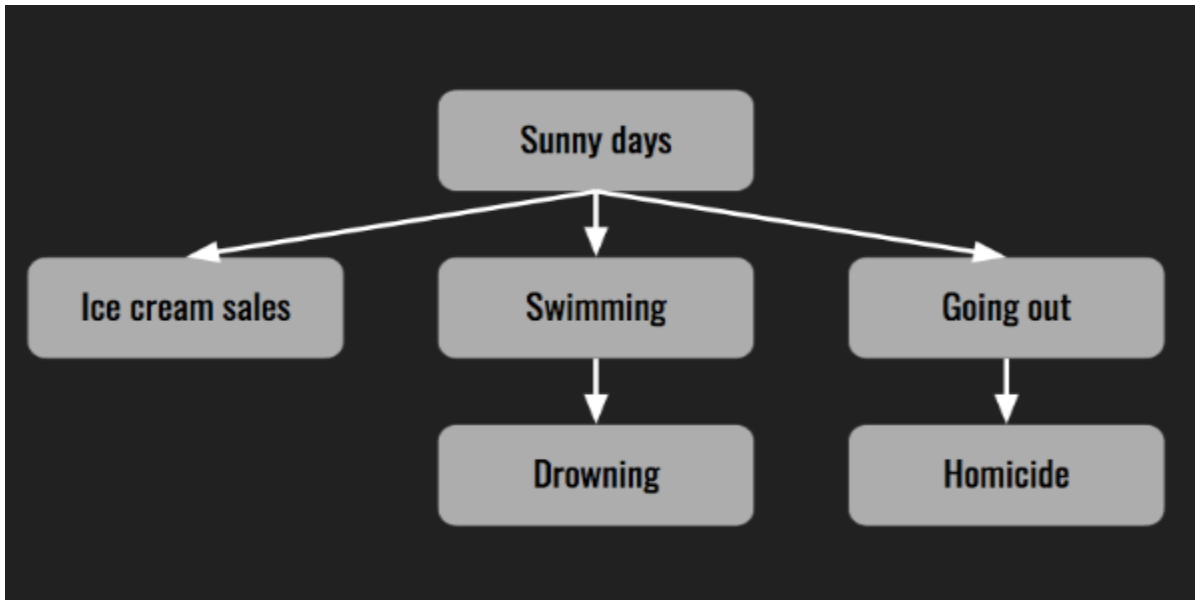
Causation takes a step further than correlation. It says any change in the value of one variable will **cause** a change in the value of another variable, which means one variable makes other to happen. It is also referred as cause and effect.

Ice cream sales is correlated with homicides in New York (Study)

- ✓ As the sales of ice cream rise and fall, so do the number of homicides. Does the consumption of ice cream causing the death of the people?

- ✓ No. Two things are correlated doesn't mean one causes other.

- ✓ Correlation does not mean causality or in our example, ice cream is not causing the death of people.



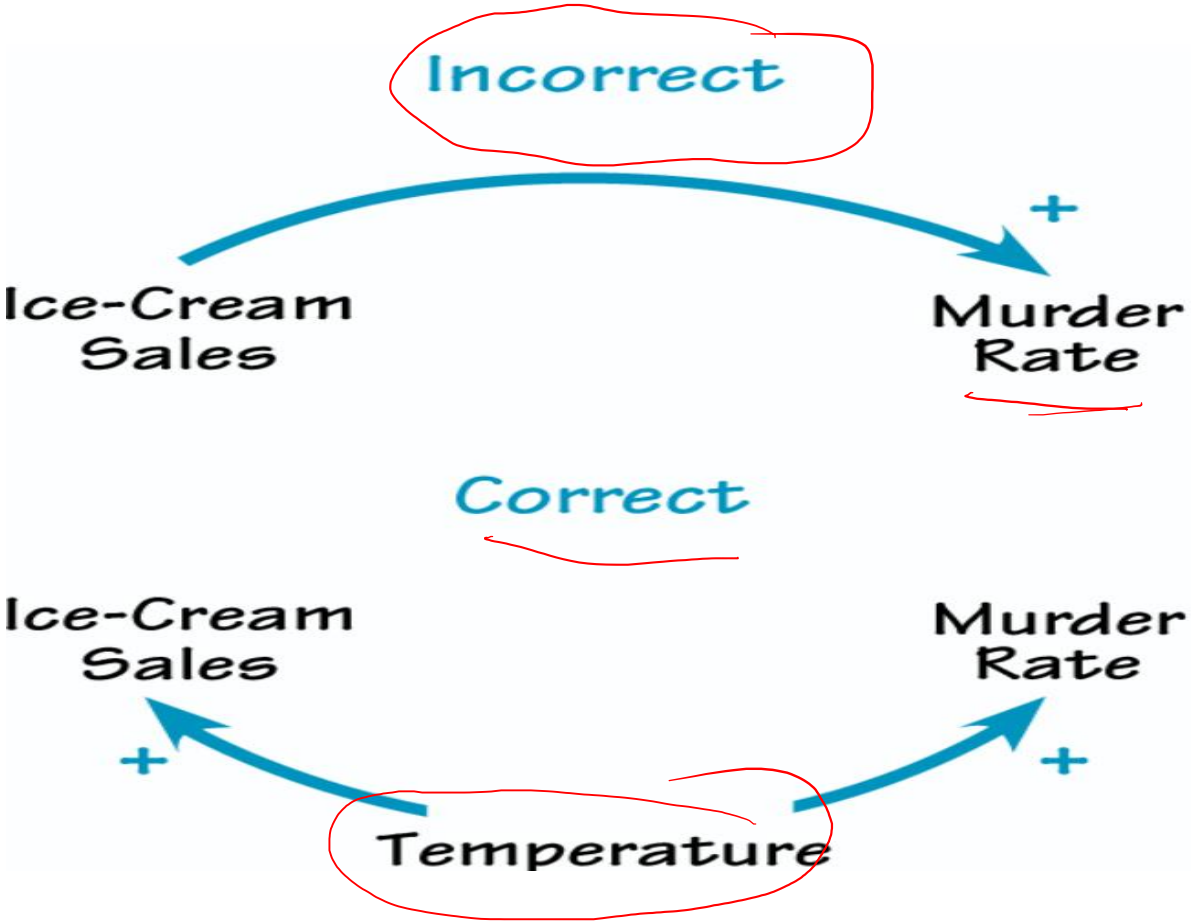
Relationship of sunny days with ice-cream sales and homicide

Consider Underlying Factors Before Conclusion

- ✓ In some cases, there are some hidden factors which are related on some level.
- ✓ Like in our example of ice cream sales and homicide rates , **weather** is the hidden factor which is causing both the things.

Don't conclude too fast!

Just after finding correlation, don't draw the conclusion too quickly. Take time to find other underlying factors as correlation is just the first step. Find the hidden factors, verify if they are correct and then conclude.



Data Preparation

- ✓ Some data preparation techniques change the way data is represented just to make it more compatible with certain machine learning algorithms.
  - ❑ Normalization
  - ❑ Binning
  - ❑ Sampling

What is Normalization

- ✓ Normalization techniques can be used to change a continuous feature to fall within a specified range while maintaining the relative differences between the values for the feature.

- ✓ For machine learning, every dataset does not require normalization. It is required only when features have different ranges.

Normalization Approaches

- ✓ min-max normalization
- ✓ z-score normalization
- ✓ normalization by decimal scaling

Min-Max Normalization

We use **range normalization/min-max normalization** to convert a feature value into the range [*low*, *high*] as follows:

$$a_i' = \frac{a_i - min(a)}{max(a) - min(a)} \times (high - low) + low$$

Ex. Let income range \$12,000 to \$98,000 normalized to [0.0, 1.0]. Then \$73,600 is mapped to

$$\frac{73{,}600 - 12{,}000}{98{,}000 - 12{,}000}(1.0 - 0) + 0 = 0.716$$

Min-Max Normalization: Example

[2,3] [-1,1]

Apply Min-max Normalization Technique on the following data point. The normalization range is [0,1]

a	marks
a1	8
a2	10
a3	15
a4	20

a	marks	Min-max normalization
a1	8	0
a2	10	0.16
a3	15	0.58
a4	20	1

$$a_i' = \frac{a_i - min(a)}{max(a) - min(a)} \times (high - low) + low$$

min(a) = 8

max(a) = 20

$a_1' = 0$

$a_4' = 1$

$a_2' = \frac{a_2 - 8}{20 - 8}$

$= \frac{10-8}{20-8} = \frac{2}{12} = \frac{1}{6}$

z-score Normalization

- ✓ Another way to normalize data is to **standardize** it into **standard scores**.
- ✓ A standard score measures how many standard deviations a feature value is from the mean for that feature. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.
- ✓ We calculate a standard score as follows:

$$a_i' = \frac{a_i - \bar{a}}{sd(a)}$$

z-score Normalization : Example

Apply **standard scores** Technique on the following data point.

Mean =13.25, sd=5.377

$$a_i' = \frac{a_i - \bar{a}}{sd(a)}$$

a	marks
a1	8
a2	10
a3	15
a4	20

s.no.	marks	Min-max normalization
a1	8	-0.9768
a2	10	
a3	15	
a4	20	

The Big Question – Normalize or Standardize?

- ✓ Normalization is good to use when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbors and Neural Networks.

- ✓ Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization.

Note: You can always start by fitting your model to raw, normalized and standardized data and compare the performance for best results.

Normalization by decimal scaling

$$v' = \frac{v}{10^j}$$

Where j is the smallest integer such that Max(|v'|) < 1

Decimal scaling. Suppose that the recorded values of A range from -986 to 917. The maximum absolute value of A is 986. To normalize by decimal scaling, we therefore divide each value by 1000 (i.e., $j = 3$) so that -986 normalizes to -0.986 and 917 normalizes to 0.917. ■

Why Should we Use Feature Scaling?

- ✓ Some machine learning algorithms are sensitive to feature scaling while others are virtually invariant to it.

- ✓ **Gradient Descent Based Algorithms**
  - ❑ Machine learning algorithms like linear regression, logistic regression, neural network, etc. that use gradient descent as an optimization technique require data to be scaled.

  - ❑ Having features on a similar scale can help the gradient descent converge more quickly towards the minima.

- ✓ **Distance-Based Algorithms**
  - ❑ Distance algorithms like KNN, K-means, and SVM are most affected by the range of features. This is because behind the scenes they are using distances between data points to determine their similarity.

Links

1. https://www.analyticsvidhya.com/blog/2021/10/handling-missing-value/

2. https://www.thoughtco.com/what-is-an-outlier-3126227

Thank you