

Chapter 3

Dealing with numerical data

Previously...

Case	Age	Gender	Nationality	Days to Recover	Education Level	Confirmed At	Recovered At
1	66	Male	Chinese	26	Diploma	23rd, Jan 2020	19th, Feb 2020
2	53	Female	Chinese	14	University	24th, Jan 2020	7th, Feb 2020
3	37	Male	Chinese	27	High School	24th, Jan 2020	21st, Feb 2020
4	36	Male	Chinese	17	University	25th, Jan 2020	12th, Feb 2020
5	56	Female	Chinese	21	Diploma	27th, Jan 2020	18th, Feb 2020
6	56	Male	Chinese	23	Diploma	27th, Jan 2020	20th, Feb 2020
7	35	Male	Chinese	7	High School	27th, Jan 2020	4th, Feb 2020
8	56	Female	Chinese	20	Diploma	28th, Jan 2020	18th, Feb 2020
9	56	Male	Chinese	25	University	29th, Jan 2020	23rd, Feb 2020
10	56	Male	Chinese	10	High School	29th, Jan 2020	9th, Feb 2020
11	31	Female	Chinese	11	University	29th, Jan 2020	10th, Feb 2020
12	37	Female	Chinese	13	University	29th, Jan 2020	12th, Feb 2020

Exploratory Data Analysis

Exploratory Data Analysis:

Process of summarising or understanding the data and extracting insights or main characteristics of the data.





Chapter Outline



**Univariate
EDA**



**Bivariate
EDA
(an introduction)**



**Correlation
coefficient**



**Some limitations
of correlation
coefficient**



**Linear
regression**

Housing data

	A	B	C	D	E	F	G	H	I	J	K
1	month	town	flat_type	block	street_name	storey_range	floor_area_sqm	flat_model	lease_commence_date	remaining_lease	resale_price
2	1/1/2017	ANG MO KIO	2 ROOM	406	ANG MO KIO AVE 10	10 TO 12	44	Improved	1979	61 years 04 months	232000
3	1/1/2017	ANG MO KIO	3 ROOM	108	ANG MO KIO AVE 4	01 TO 03	67	New Generation	1978	60 years 07 months	250000
4	1/1/2017	ANG MO KIO	3 ROOM	602	ANG MO KIO AVE 5	01 TO 03	67	New Generation	1980	62 years 05 months	262000
5	1/1/2017	ANG MO KIO	3 ROOM	465	ANG MO KIO AVE 10	04 TO 06	68	New Generation	1980	62 years 01 month	265000
6	1/1/2017	ANG MO KIO	3 ROOM	601	ANG MO KIO AVE 5	01 TO 03	67	New Generation	1980	62 years 05 months	265000
7	1/1/2017	ANG MO KIO	3 ROOM	150	ANG MO KIO AVE 5	01 TO 03	68	New Generation	1981	63 years	275000
8	1/1/2017	ANG MO KIO	3 ROOM	447	ANG MO KIO AVE 10	04 TO 06	68	New Generation	1979	61 years 06 months	280000
9	1/1/2017	ANG MO KIO	3 ROOM	218	ANG MO KIO AVE 1	04 TO 06	67	New Generation	1976	58 years 04 months	285000
10	1/1/2017	ANG MO KIO	3 ROOM	447	ANG MO KIO AVE 10	04 TO 06	68	New Generation	1979	61 years 06 months	285000
11	1/1/2017	ANG MO KIO	3 ROOM	571	ANG MO KIO AVE 3	01 TO 03	67	New Generation	1979	61 years 04 months	285000
12	1/1/2017	ANG MO KIO	3 ROOM	534	ANG MO KIO AVE 10	01 TO 03	68	New Generation	1980	62 years 01 month	288500
13	1/1/2017	ANG MO KIO	3 ROOM	233	ANG MO KIO AVE 3	10 TO 12	67	New Generation	1977	59 years 08 months	295000
14	1/1/2017	ANG MO KIO	3 ROOM	235	ANG MO KIO AVE 3	04 TO 06	67	New Generation	1977	59 years 08 months	295000
15	1/1/2017	ANG MO KIO	3 ROOM	219	ANG MO KIO AVE 1	07 TO 09	67	New Generation	1977	59 years 06 months	297000
16	1/1/2017	ANG MO KIO	3 ROOM	536	ANG MO KIO AVE 10	07 TO 09	68	New Generation	1980	62 years 01 month	298000
17	1/1/2017	ANG MO KIO	3 ROOM	230	ANG MO KIO AVE 3	04 TO 06	67	New Generation	1978	60 years	298000
18	1/1/2017	ANG MO KIO	3 ROOM	570	ANG MO KIO AVE 3	10 TO 12	67	New Generation	1979	61 years 04 months	3.00E+05
19	1/1/2017	ANG MO KIO	3 ROOM	624	ANG MO KIO AVE 4	04 TO 06	68	New Generation	1980	62 years 08 months	301000
20	1/1/2017	ANG MO KIO	3 ROOM	441	ANG MO KIO AVE 10	07 TO 09	67	New Generation	1979	61 years	306000



Problem → Plan → Data

What factors may affect the popularity and pricing of resale flats sold in Singapore?

Age?

Time period?

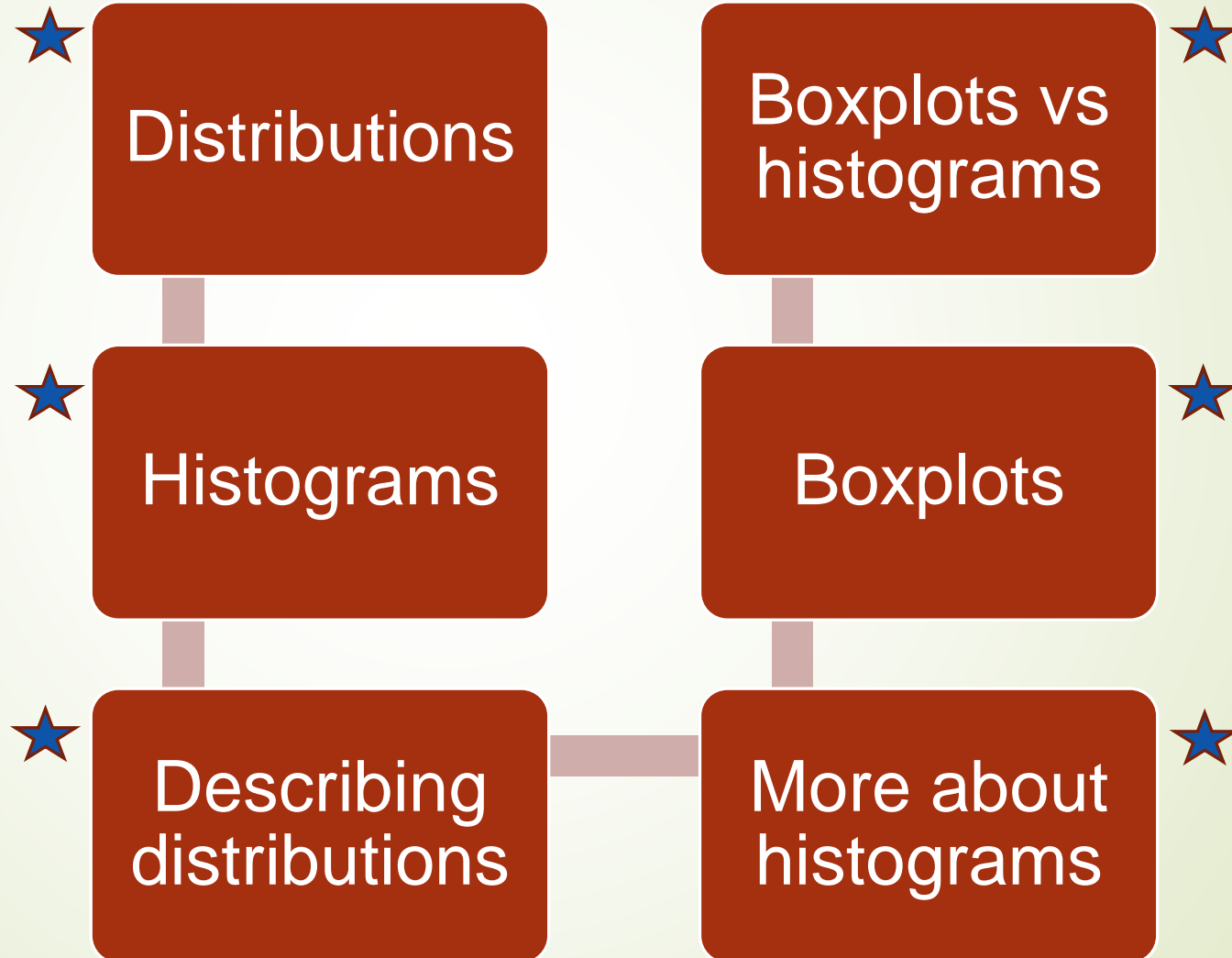
Floor area?

	A	B	C	D
1	month	floor_area_sqm	age	resale_price
2	1/1/2021	45	35	225000
3	1/1/2021	45	35	211000
4	1/1/2021	73	45	275888
5	1/1/2021	67	43	316800
6	1/1/2021	67	43	305000
7	1/1/2021	68	40	260000
8	1/1/2021	73	44	351000
9	1/1/2021	73	44	343000
10	1/1/2021	75	41	306000



Univariate Exploratory Data Analysis

Outline of unit



Distributions

	A	B	C	D
1	month	floor_area_sqm	age	resale_price
2	1/1/2021	45	35	225000
3	1/1/2021	45	35	211000
4	1/1/2021	73	45	275888
5	1/1/2021	67	43	316800
6	1/1/2021	67	43	305000
7	1/1/2021	68	40	260000
8	1/1/2021	73	44	351000
9	1/1/2021	73	44	343000
10	1/1/2021	75	41	306000




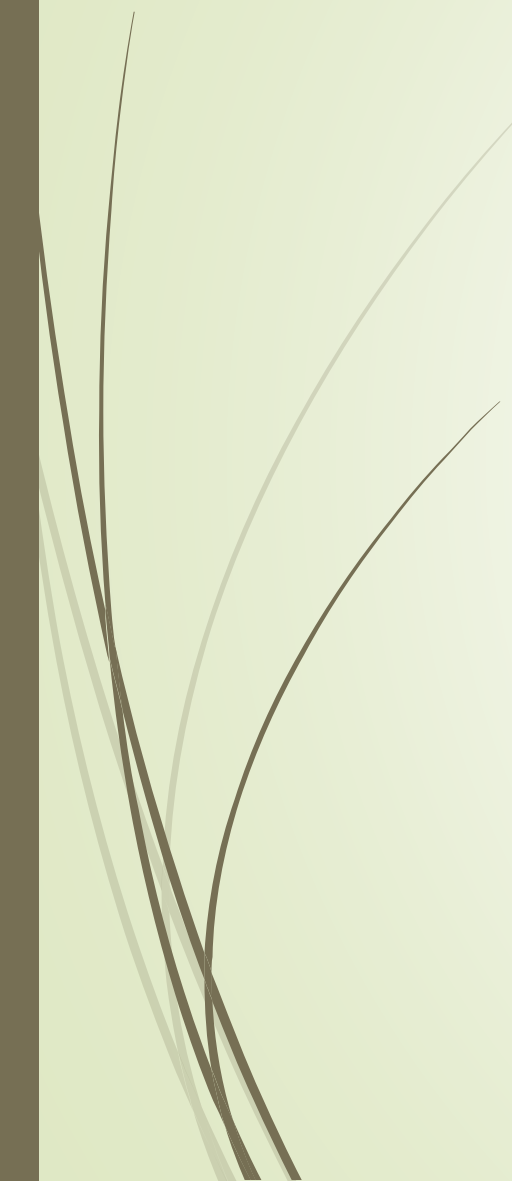
Age	Frequency
2	9
3	8
4	583
5	1105
6	884
7	295
8	255
9	219
10	147

•
•
•

Any pattern in the data?



Histograms

- 
- Graphical display of a distribution
- 
- Quick and easy to grasp
- 
- Useful for large data sets
- 

Histograms

Age	Frequency
2	9
3	8
4	583
5	1105
6	884
7	295
8	255
9	219
10	147

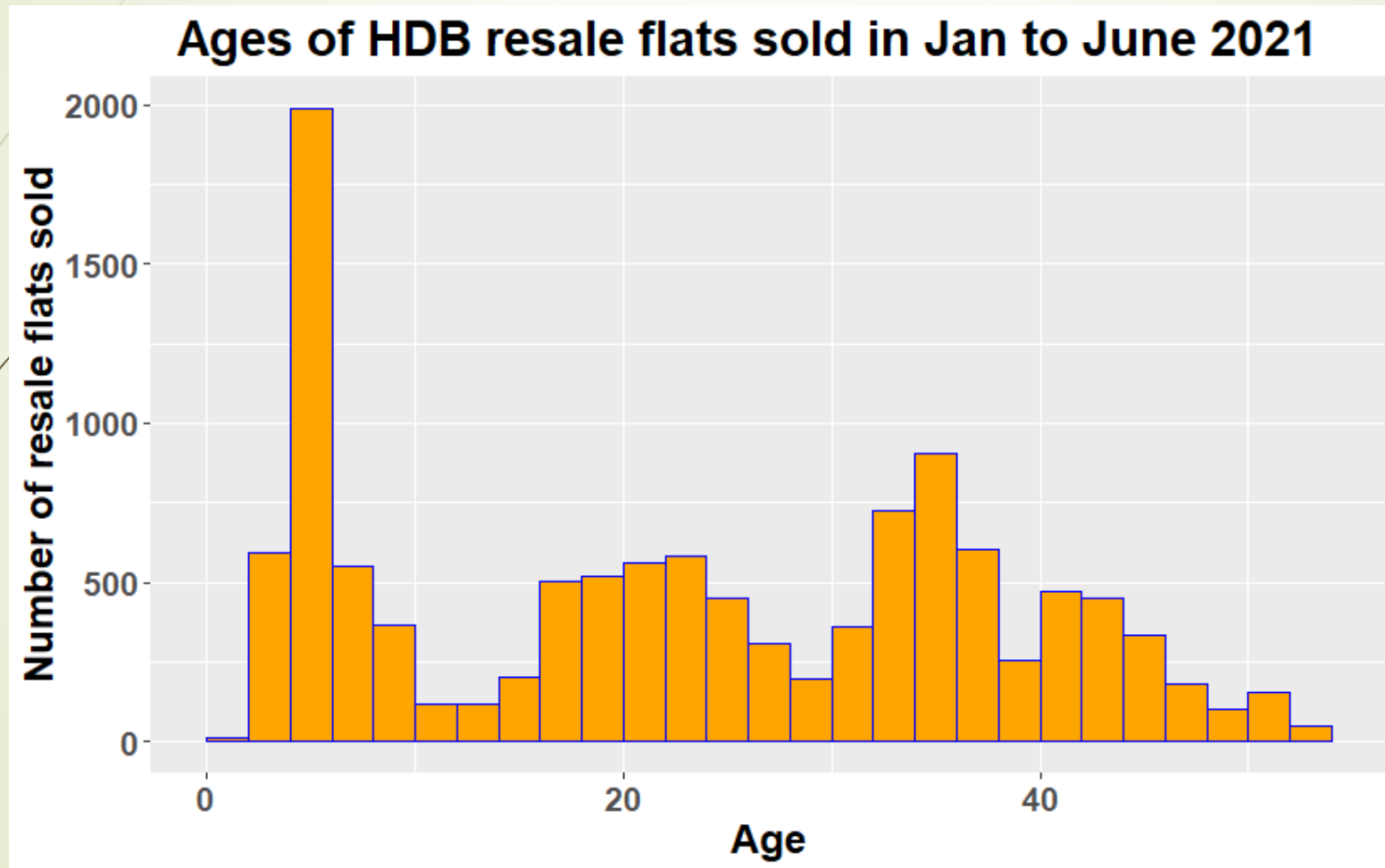
•
•
•



Bins	Frequency
0-2	9
2-4	591 (8 + 583)
4-6	1989 (1105 + 884)
6-8	550 (295 + 255)
8-10	336 (219 + 147)

•
•
•

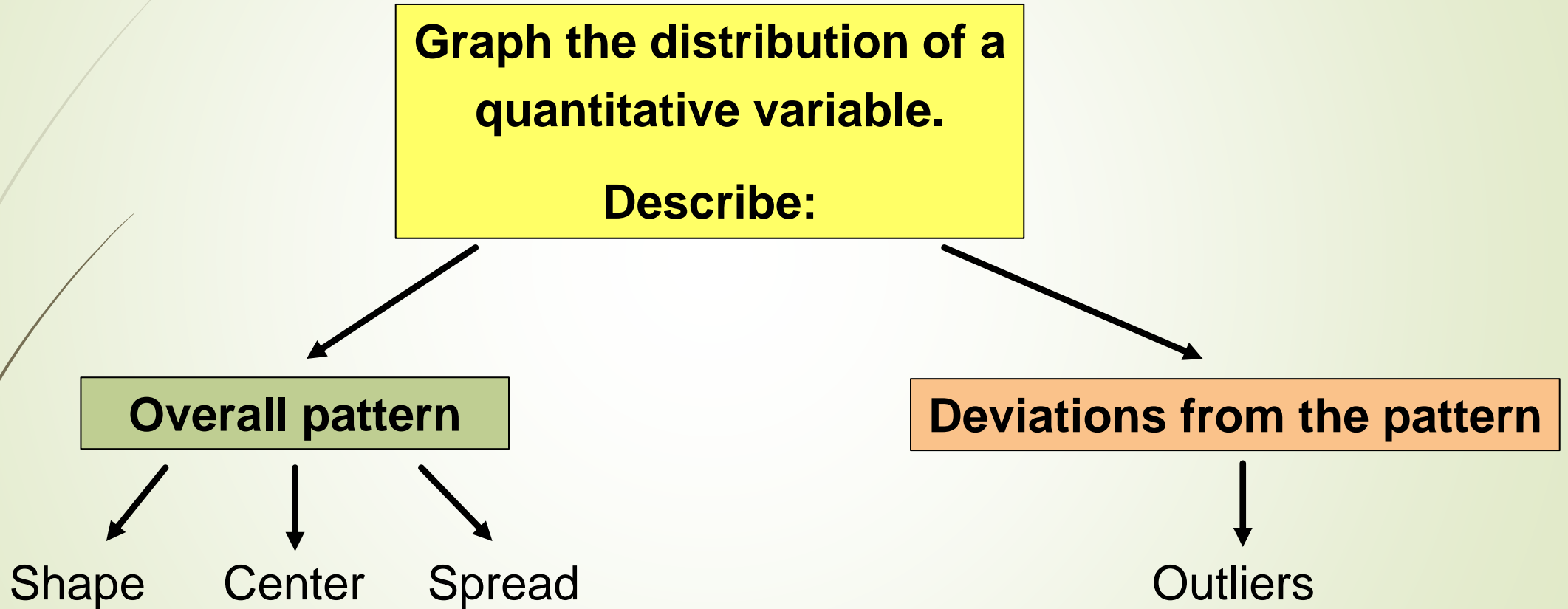
Histograms



Range 4-6 years:

- Most frequent
- 17% of flats sold

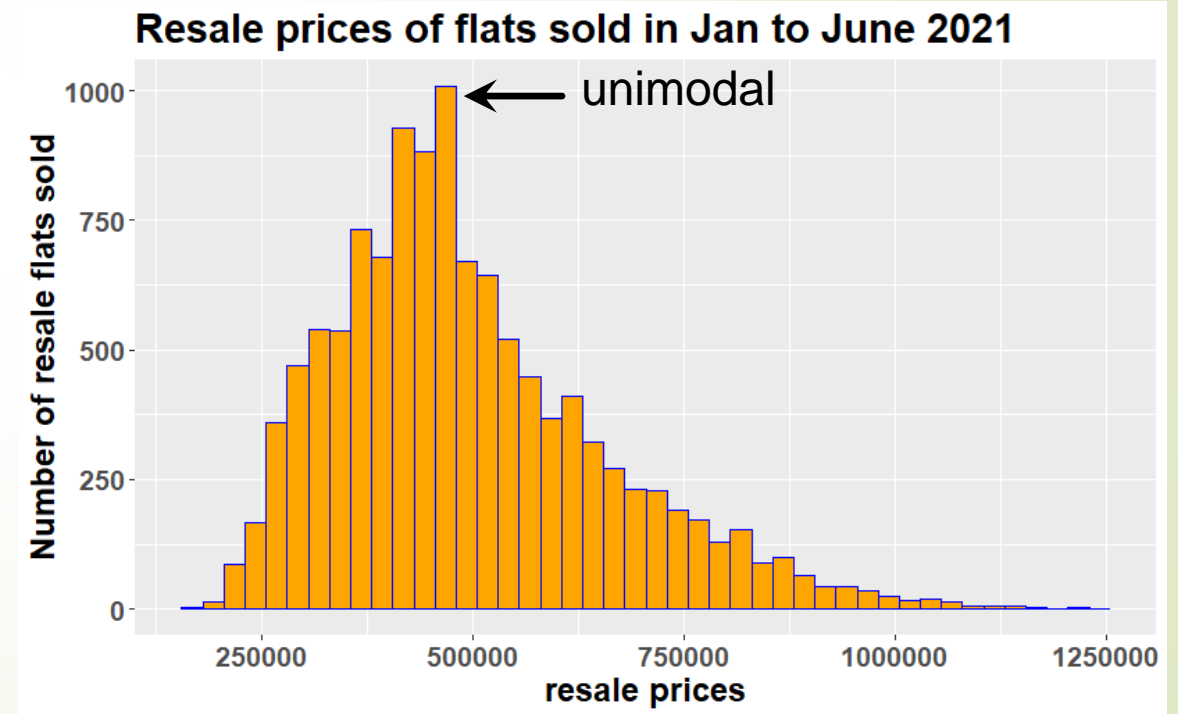
Describing distributions



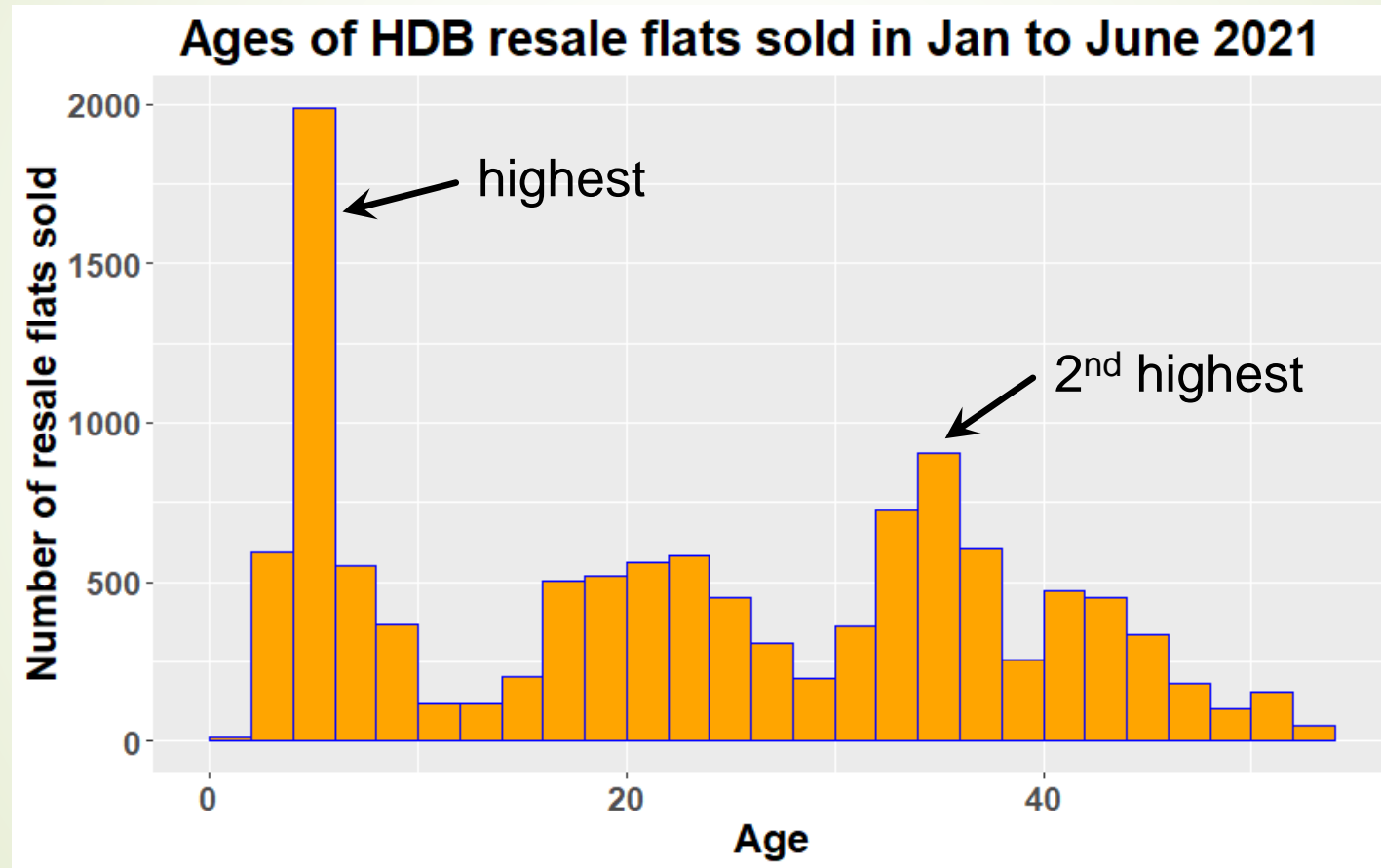
Describing distributions

Shape of a distribution: peaks and skewness

	A	B	C	D
1	month	floor_area_sqm	age	resale_price
2	1/1/2021	45	35	225000
3	1/1/2021	45	35	211000
4	1/1/2021	73	45	275888
5	1/1/2021	67	43	316800
6	1/1/2021	67	43	305000
7	1/1/2021	68	40	260000
8	1/1/2021	73	44	351000
9	1/1/2021	73	44	343000
10	1/1/2021	75	41	306000

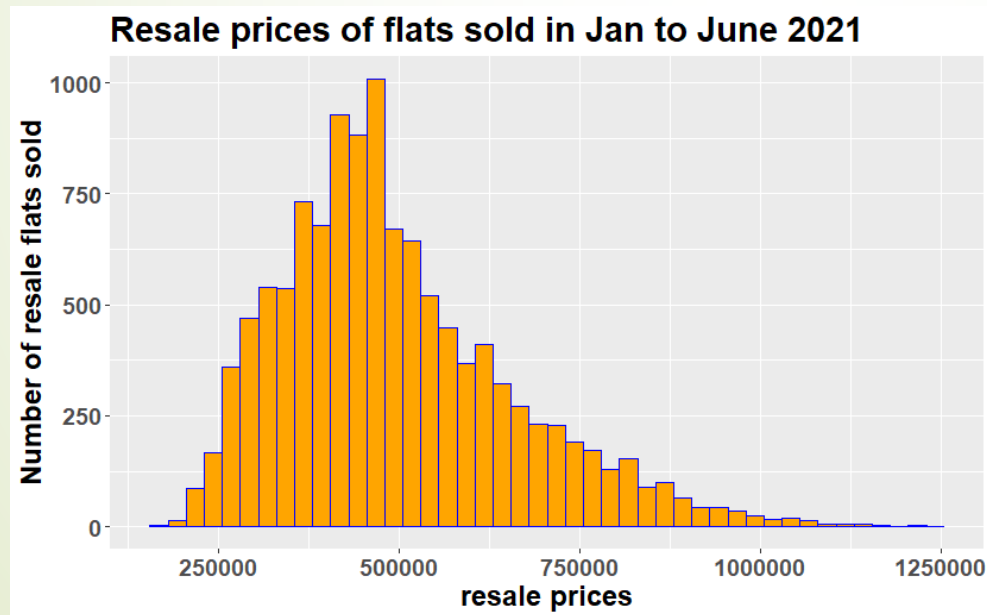
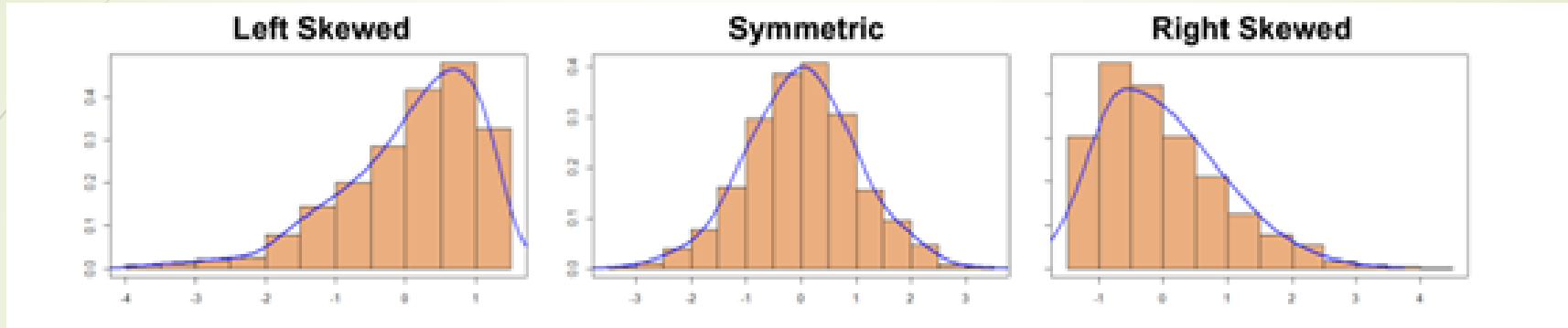


Describing distributions



Multimodal distribution!

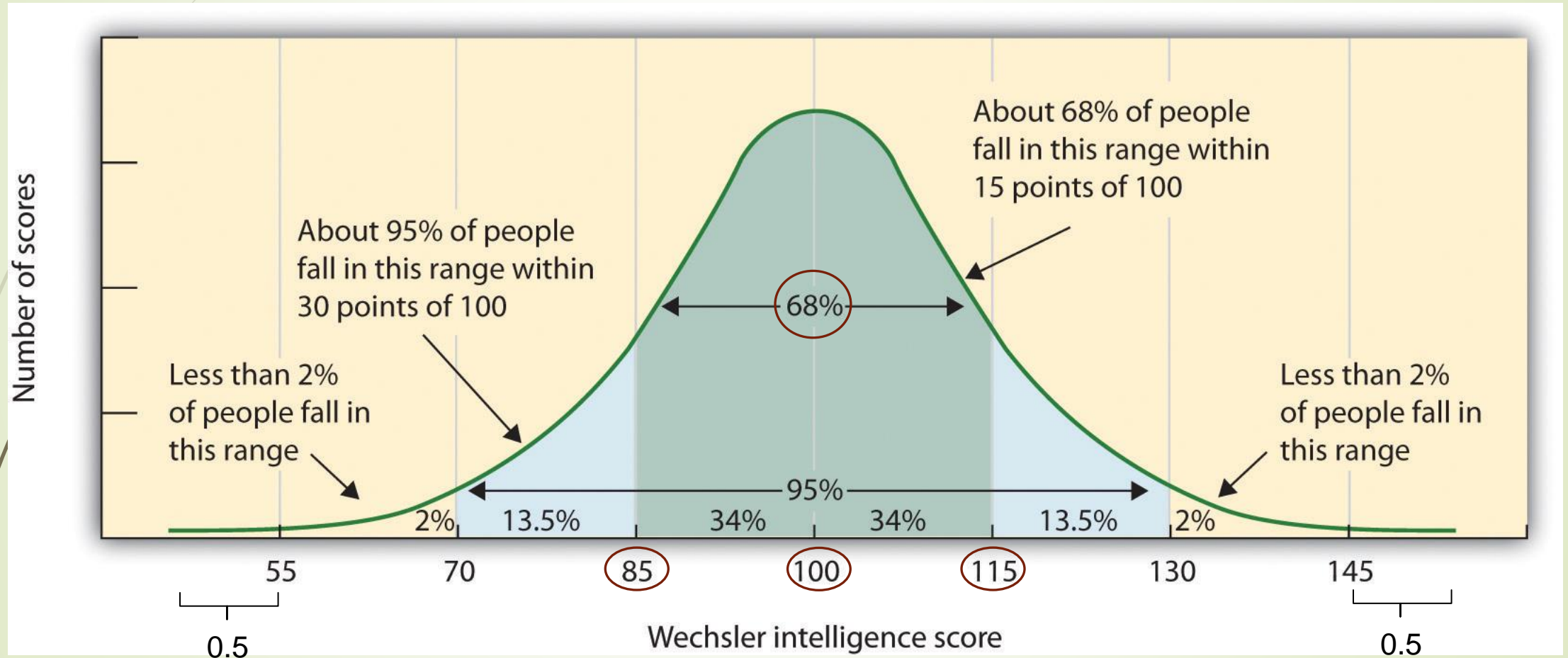
Describing distributions



Right skewed:

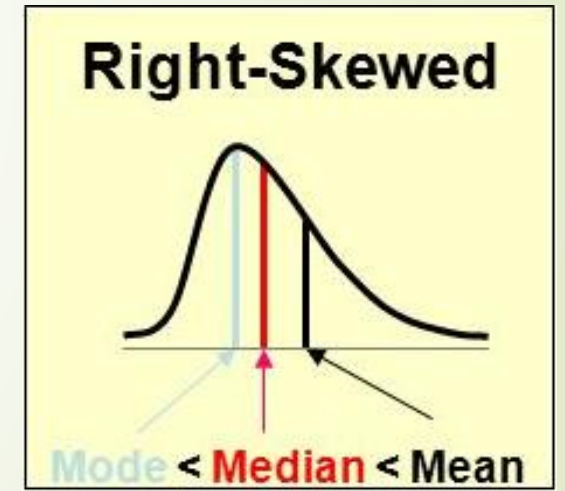
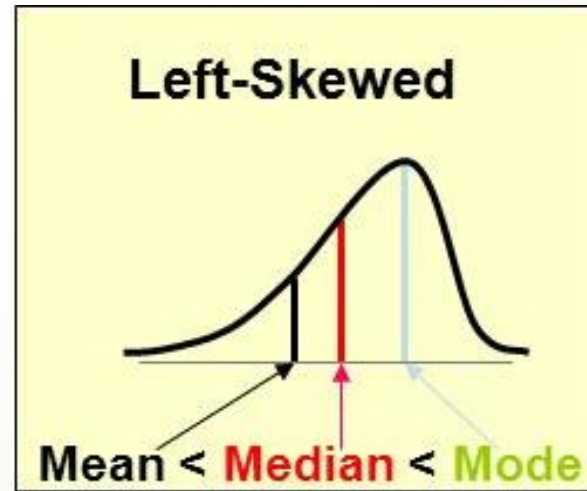
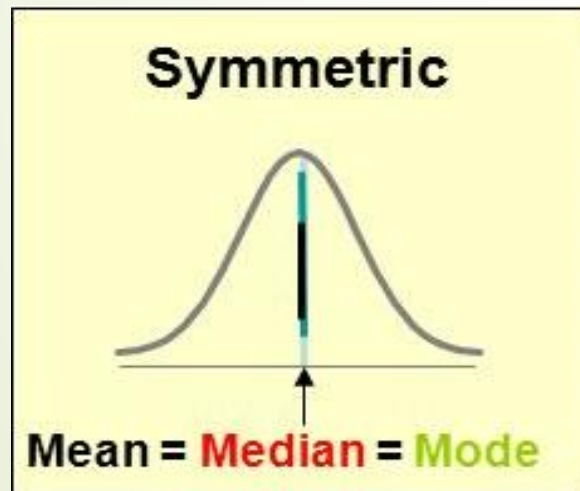
Very few flats sold at very high prices.

Describing distributions

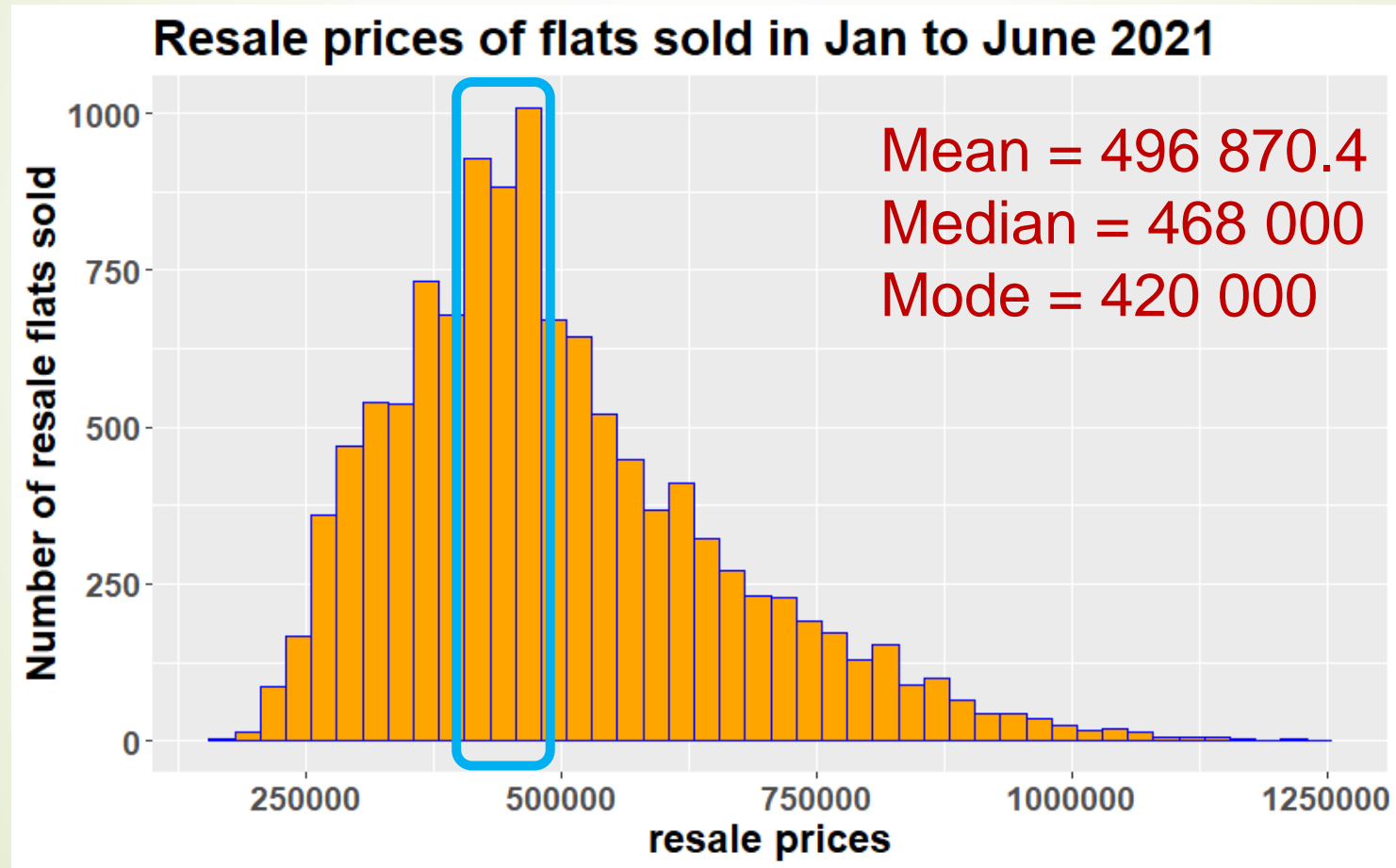


Describing distributions

Center of a distribution: mean, median and mode

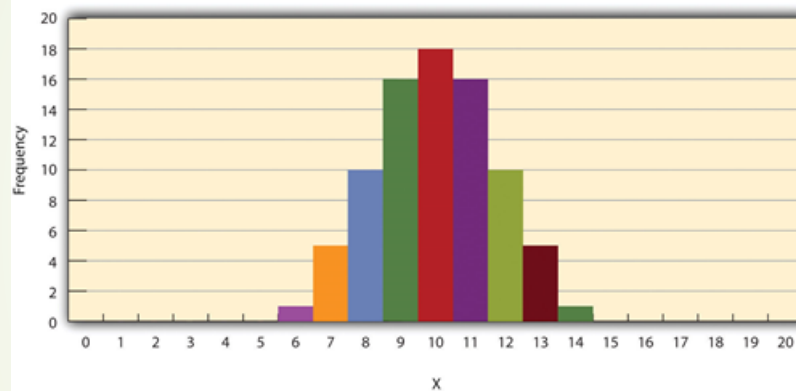


Describing distributions



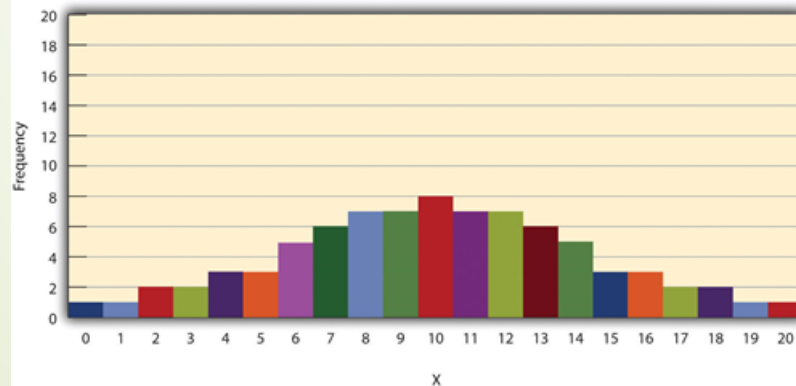
Describing distributions

Spread of a distribution: range and standard deviation



Low variability

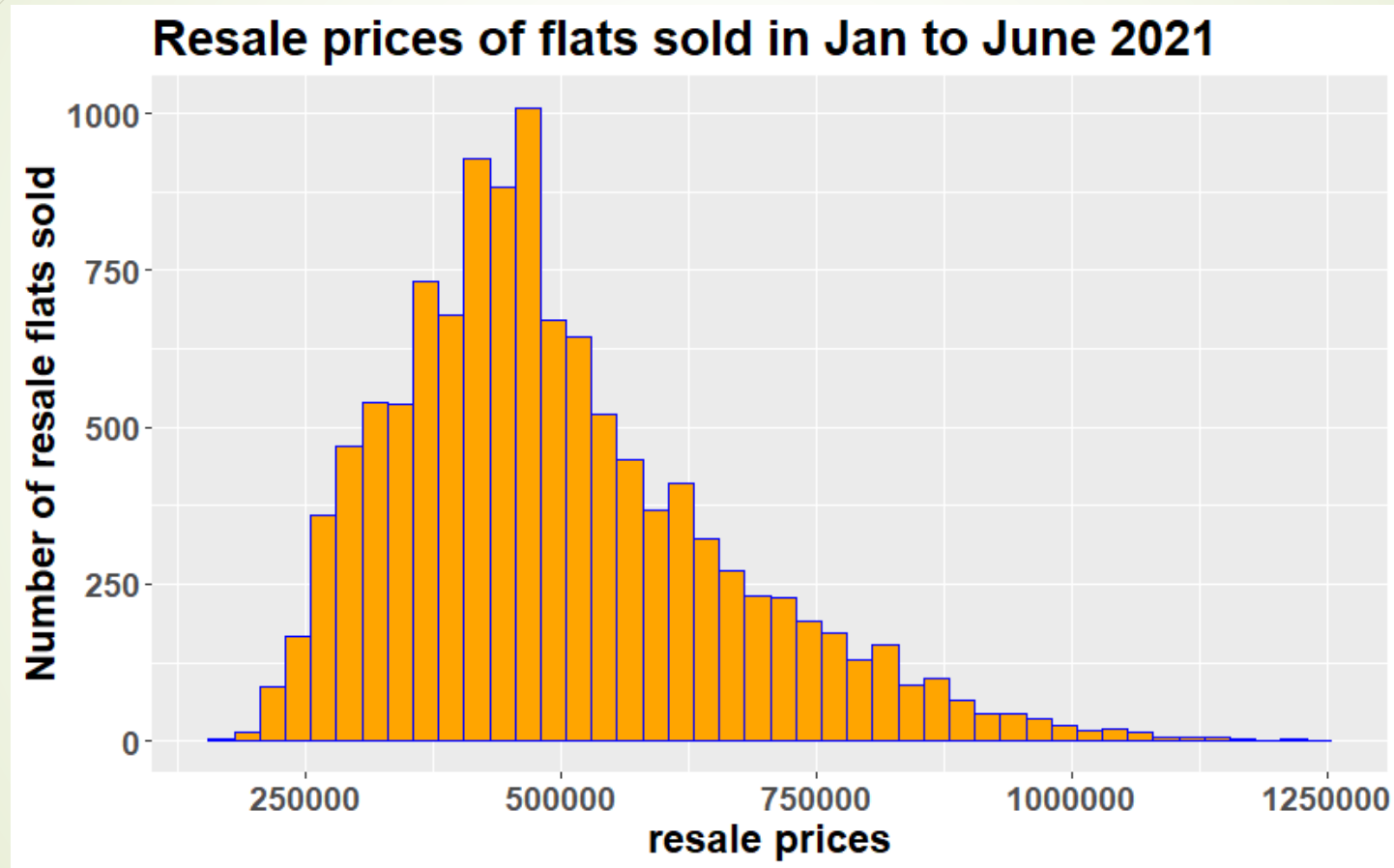
$$s = 1.69$$



High variability

$$s = 4.30$$

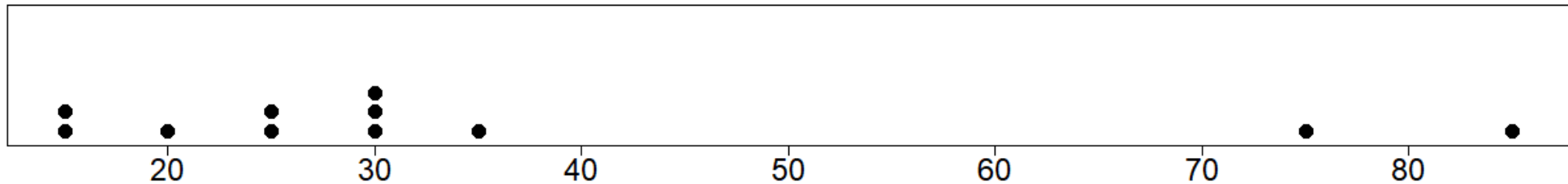
Describing distributions



1070 000

Describing distributions

Outliers: observations that fall well above or well below the overall bulk of data



Examining data for outliers can be useful in

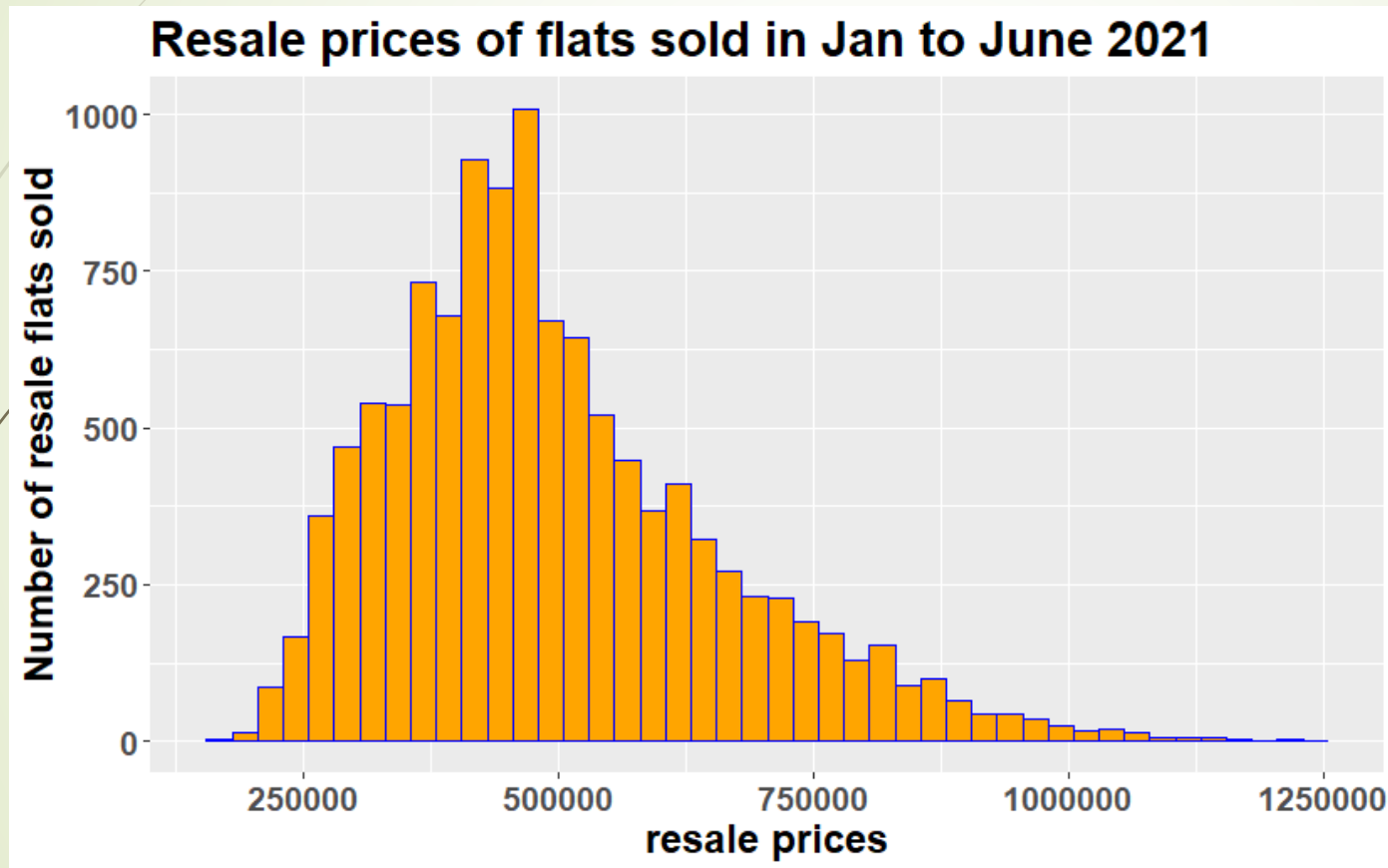
- identifying strong skew in a distribution
- identifying possible data collection or data entry errors
- providing interesting insight into the data

Describing distributions

Without Outlier	With Outlier
4, 4, 5, 5, 5, 5, 6, 6, 6, 7, 7	4, 4, 5, 5, 5, 5, 6, 6, 6, 7, 7, 300
Mean = 5.45	Mean = 30
Median = 5	Median = 5.5
Mode = 5	Mode = 5
Standard Deviation = 1.04	Standard Deviation = 85.03

Note: It may be good practice to repeat analysis with and without the outlier(s).

Describing distributions

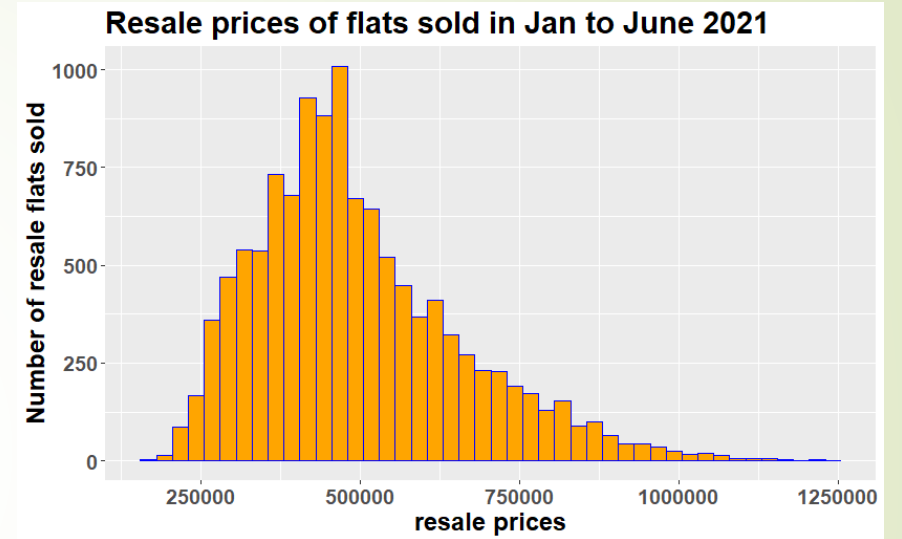
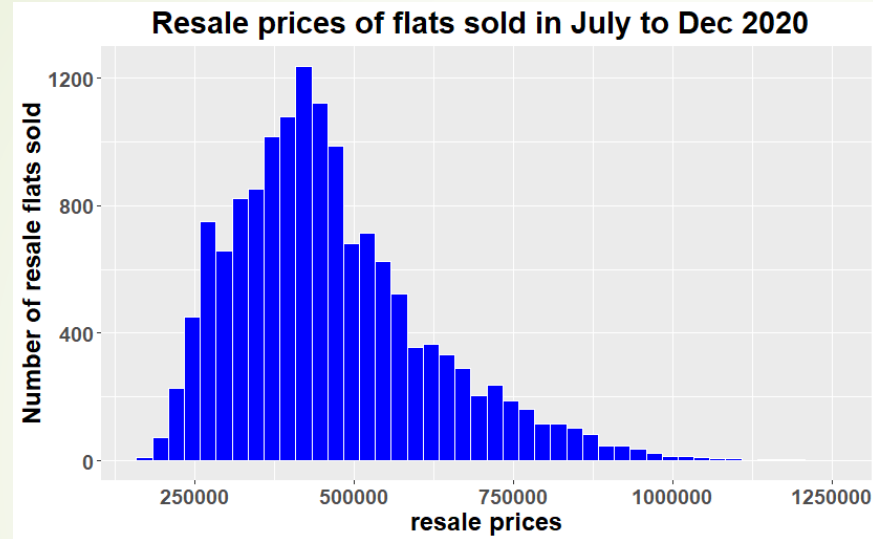


Mean = 496 870.4

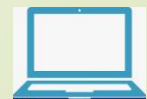
Median = 468 000

More about histograms

COMPARING HISTOGRAMS



<i>resale_price (july to Dec 2020)</i>		<i>resale_price (Jan to June 2021)</i>	
Mean	462826.8108	Mean	496870.412
Standard Error	1290.60473	Standard Error	1502.279126
Median	435000	Median	468000
Mode	400000	Mode	420000
Standard Deviation	155955.0436	Standard Deviation	162106.9861
Sample Variance	24321975624	Sample Variance	26278674930
Kurtosis	1.067523787	Kurtosis	0.82677618
Skewness	0.964942209	Skewness	0.904035777
Range	1098000	Range	1070000
Minimum	160000	Minimum	180000
Maximum	1258000	Maximum	1250000
Sum	6758197092	Sum	5785559077
Count	14602	Count	11644

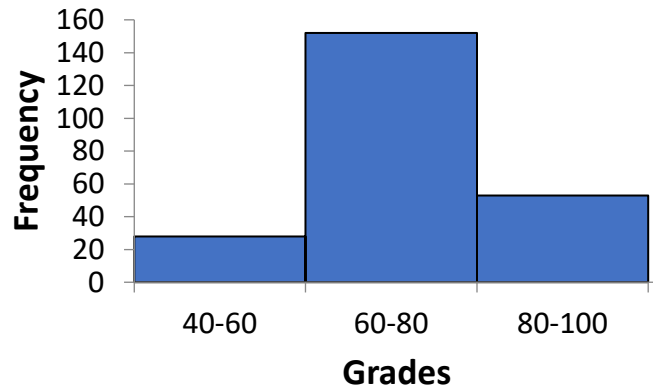


More about histograms

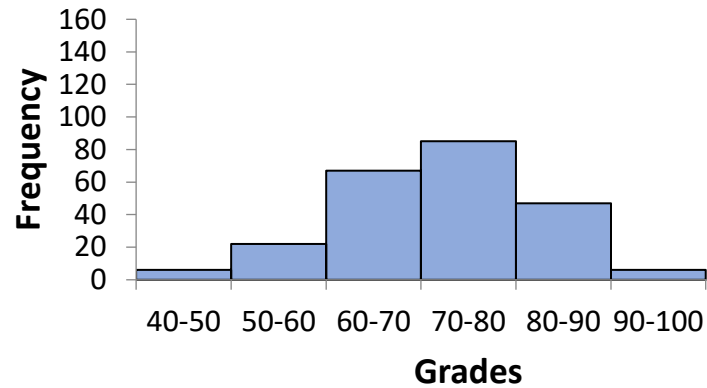
Bin size matters!



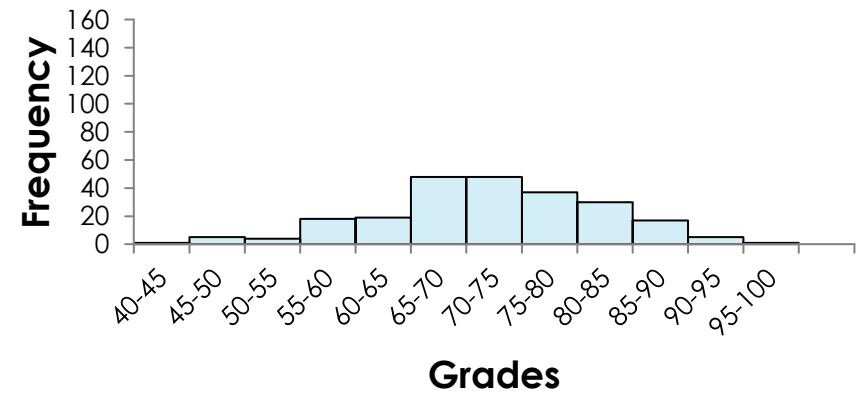
Final exam grades



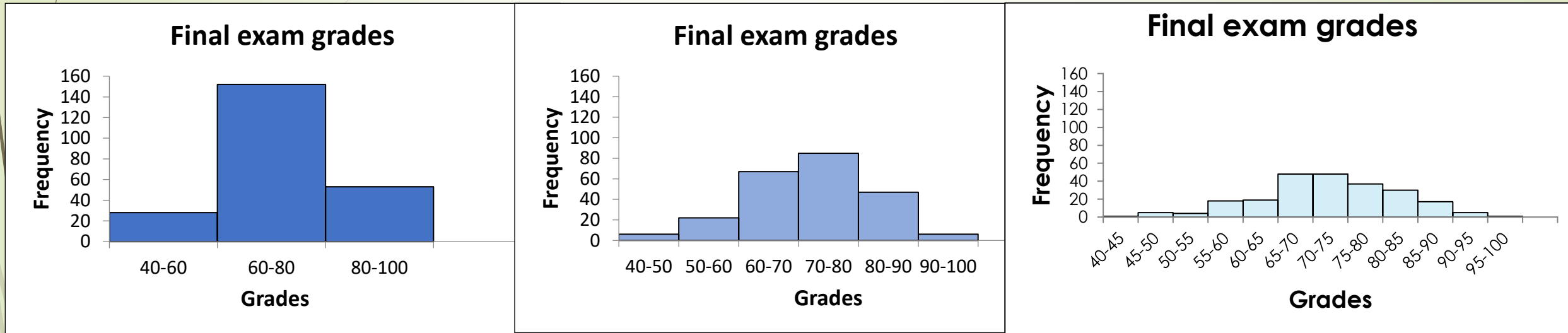
Final exam grades



Final exam grades



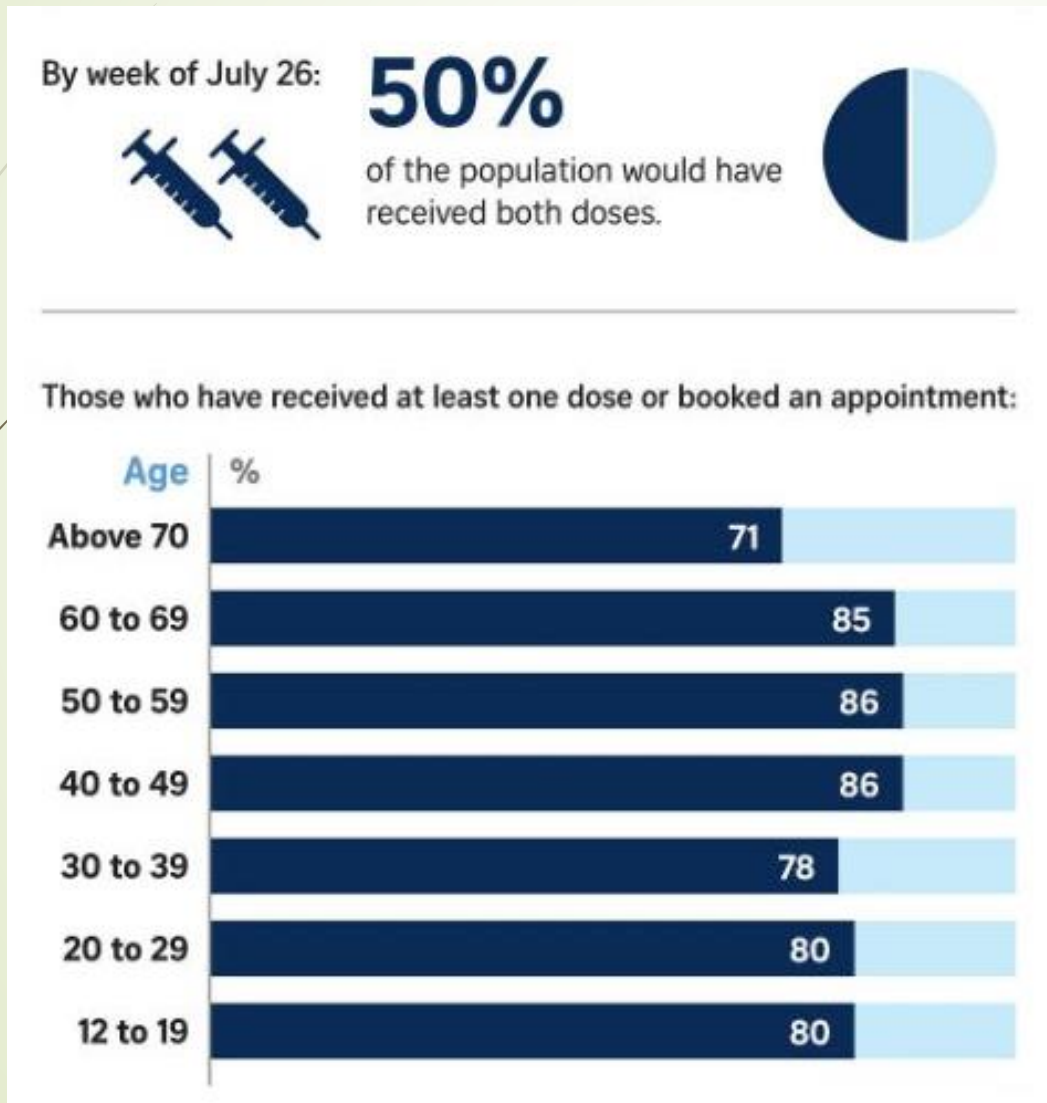
More about histograms



General advice:

- Avoid histograms with large bin widths that group data into only a few bins.
- Avoid histograms with very small bin widths that group data into too many bins.
- Construct histograms with different bin sizes to see which one is the most useful for our purpose.

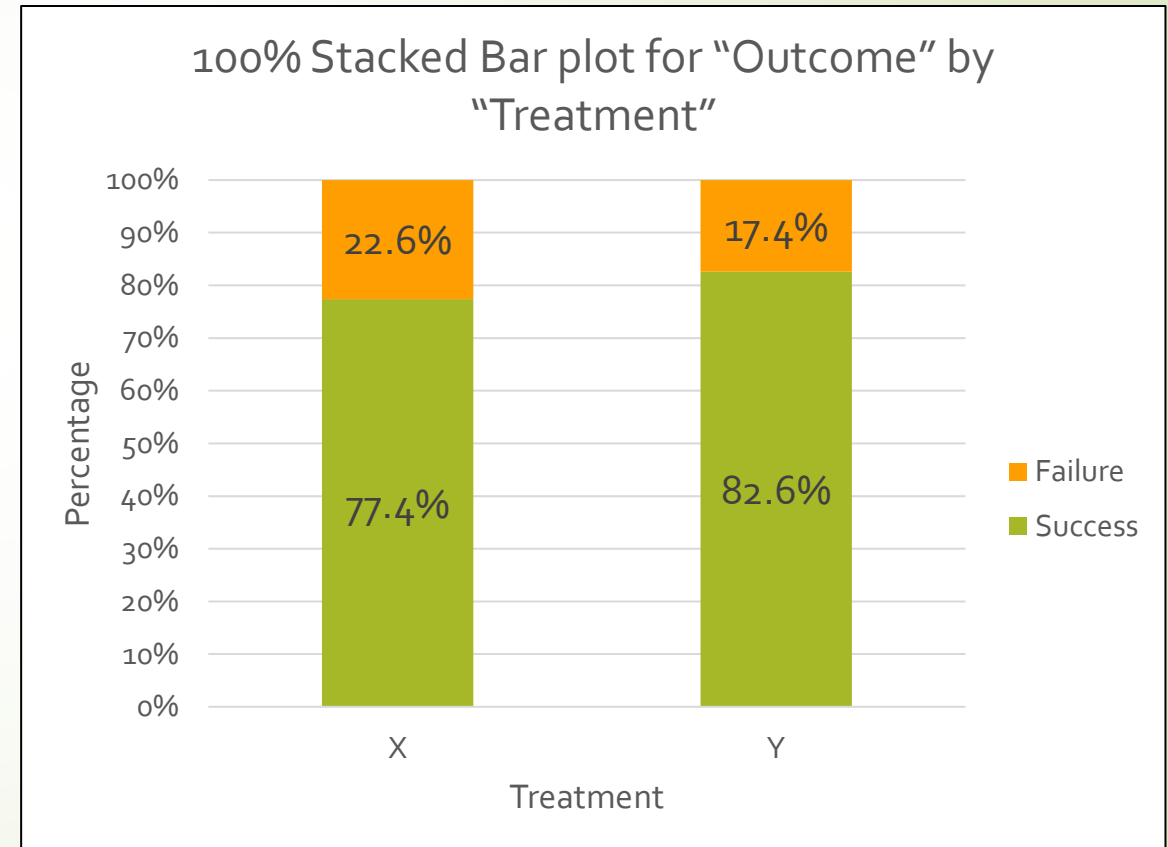
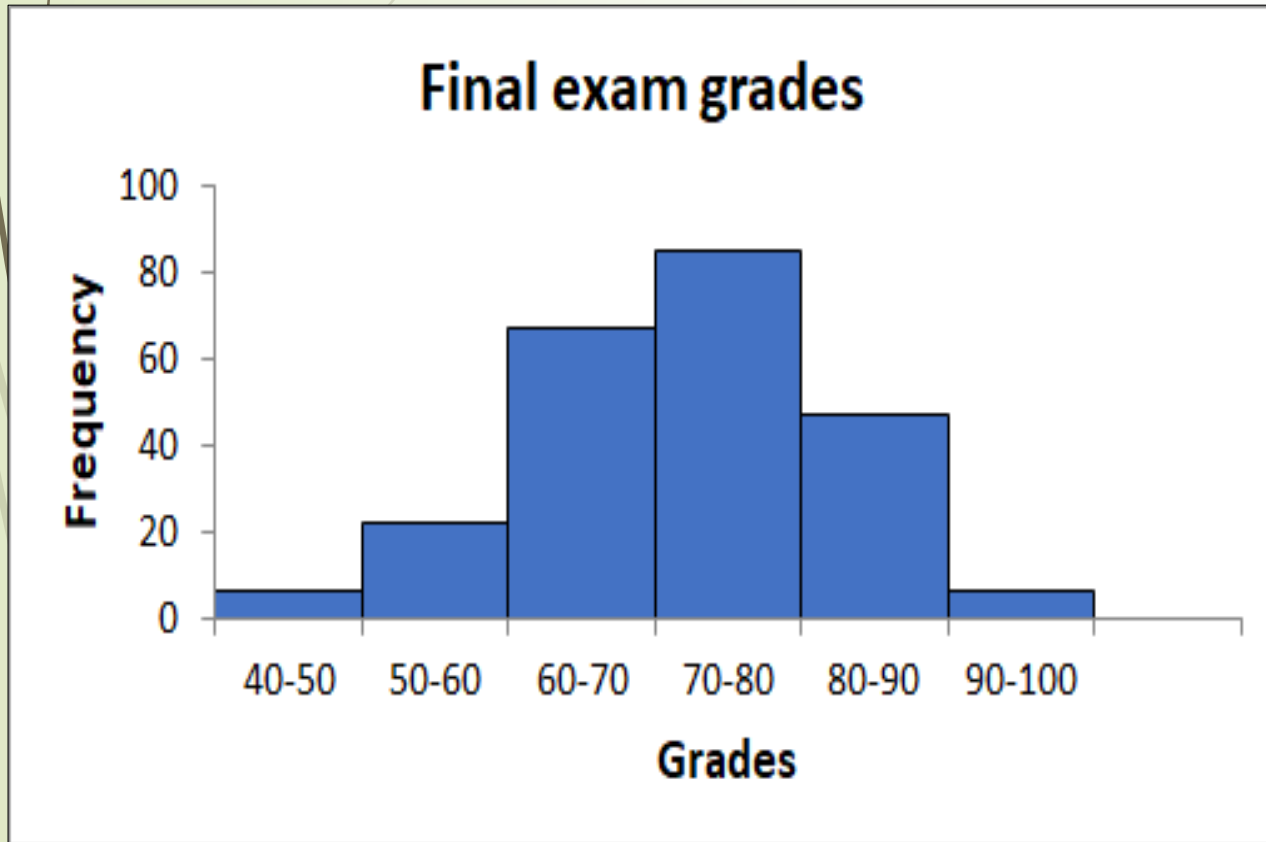
More about histograms



Histogram or bar graph?

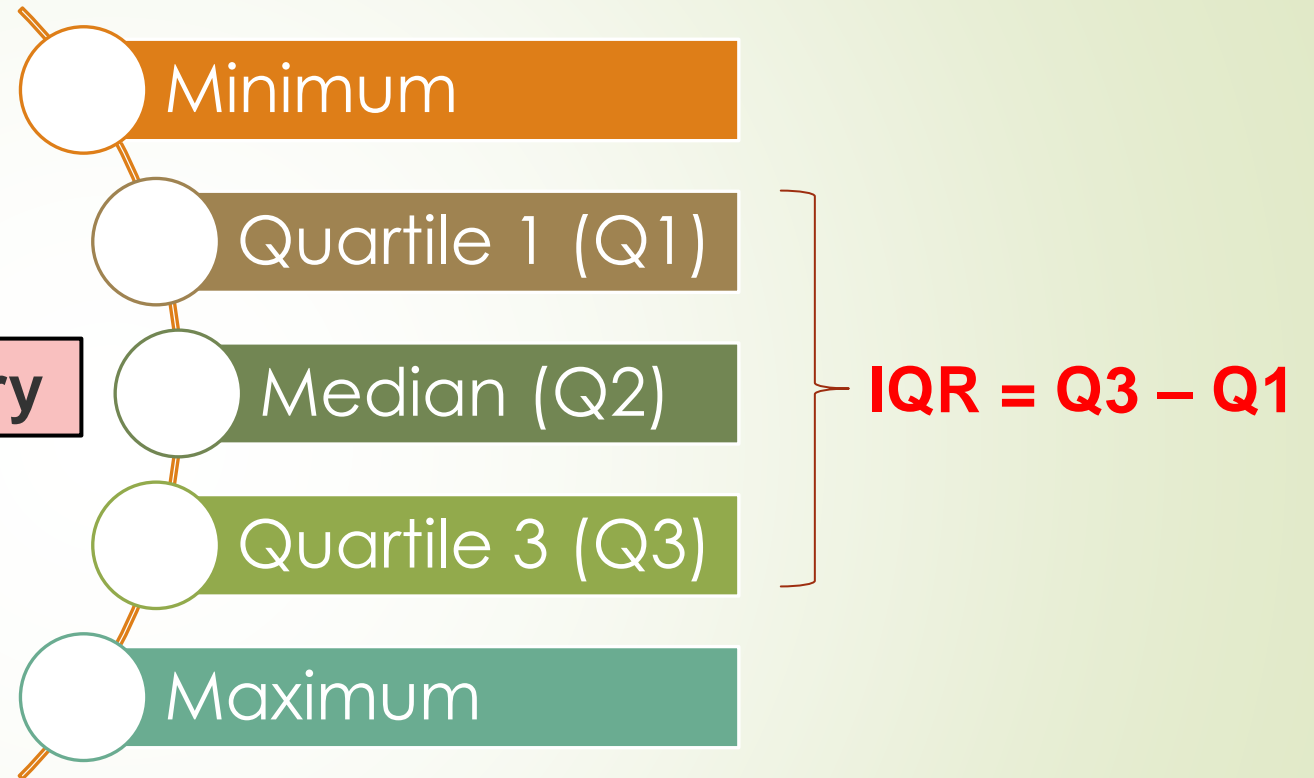
More about histograms

Histograms vs bar graphs



Boxplots

Five-number summary



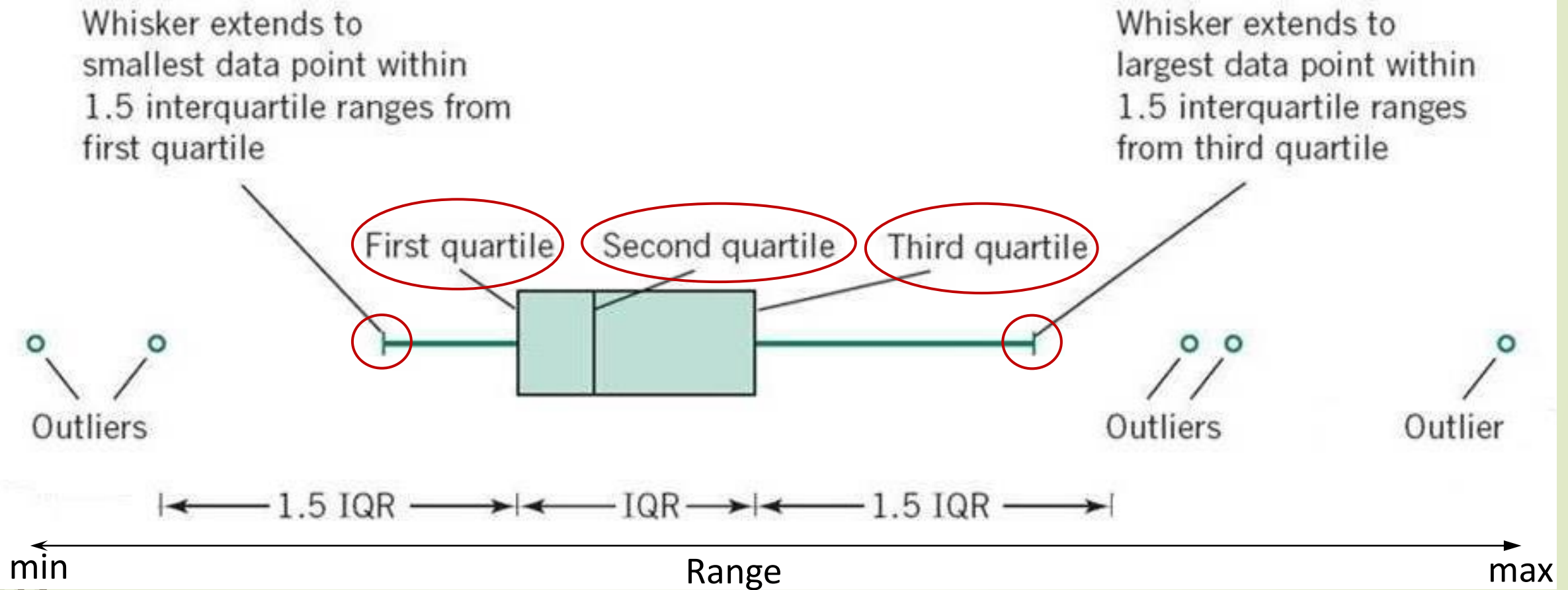
Data point is outlier if

greater than $Q3 + 1.5 * IQR$

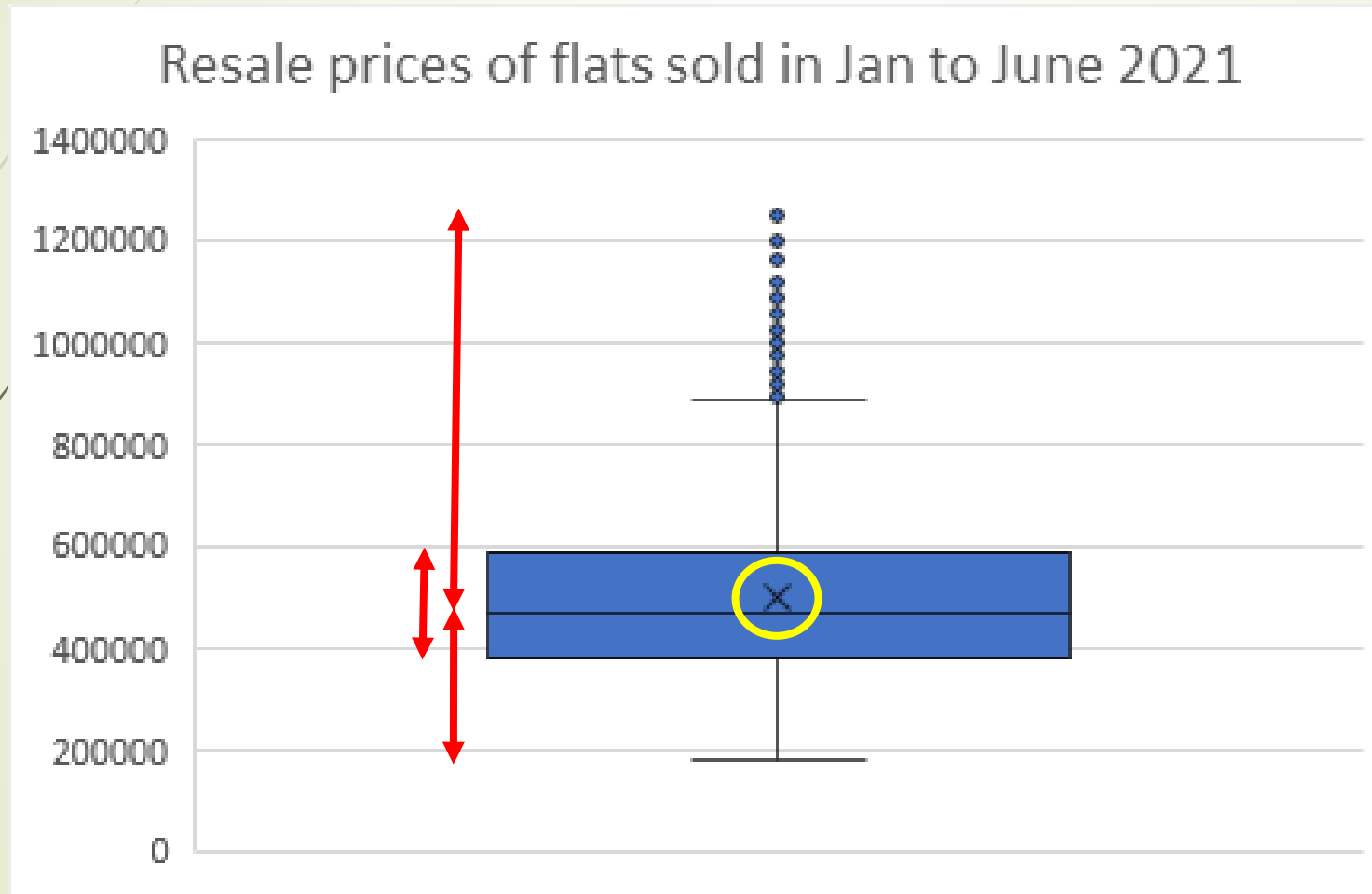
or

less than $Q1 - 1.5 * IQR$

Boxplots



Boxplots



Shape?

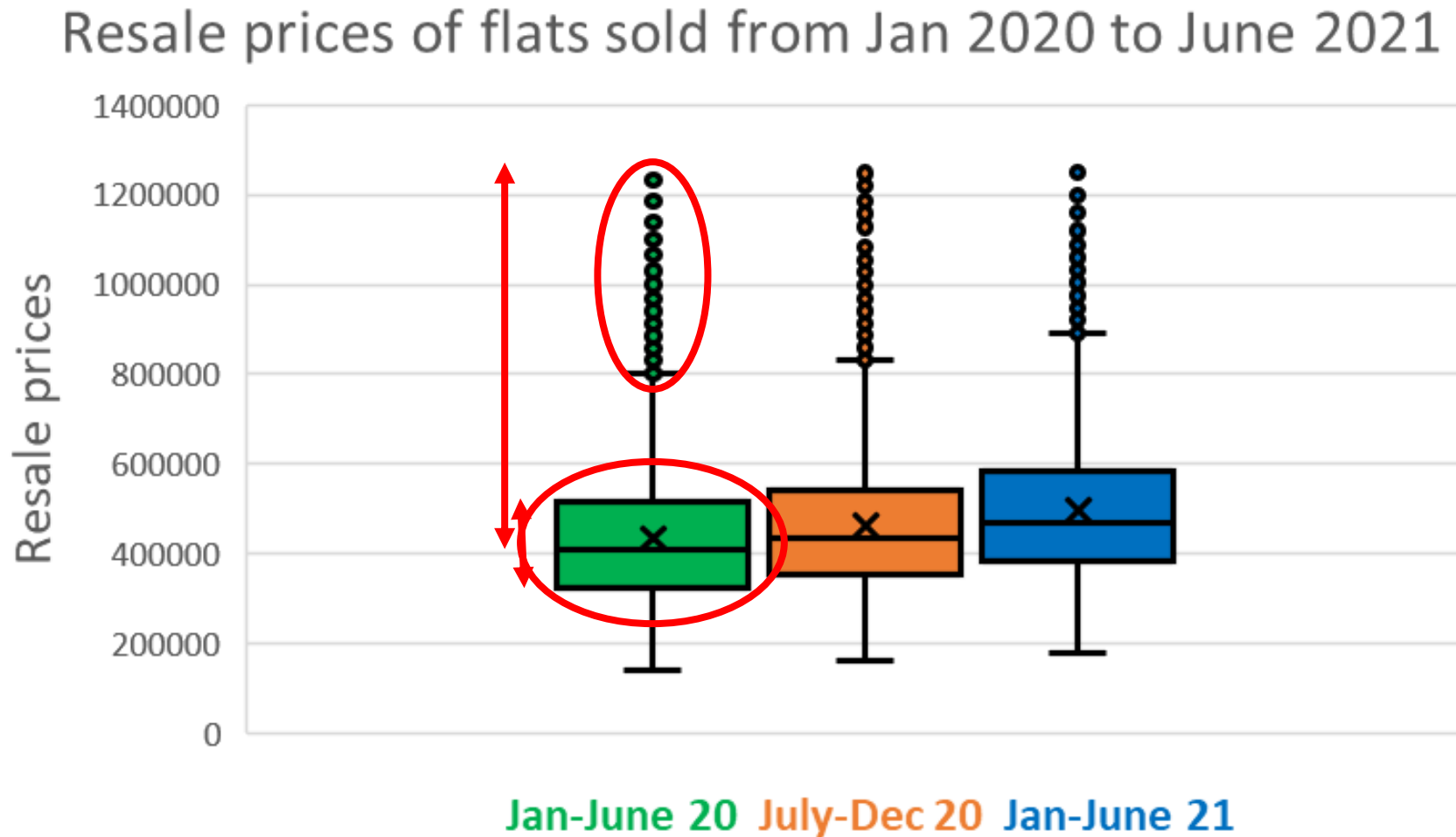
Center?

Spread?

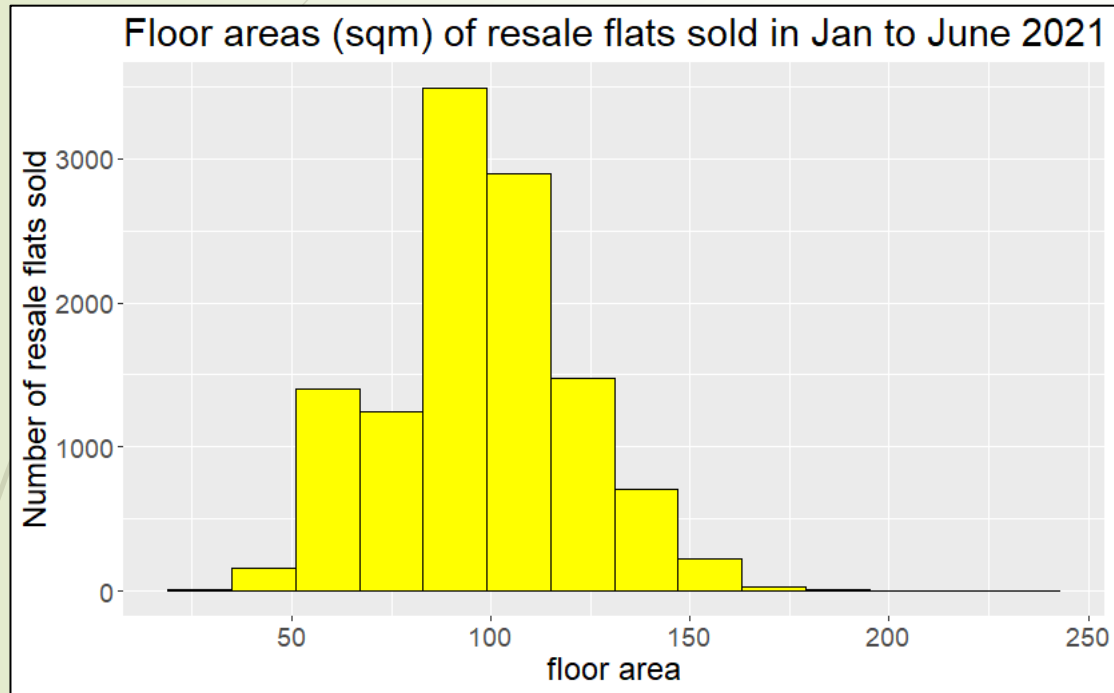
IQR = 204 000

Boxplots

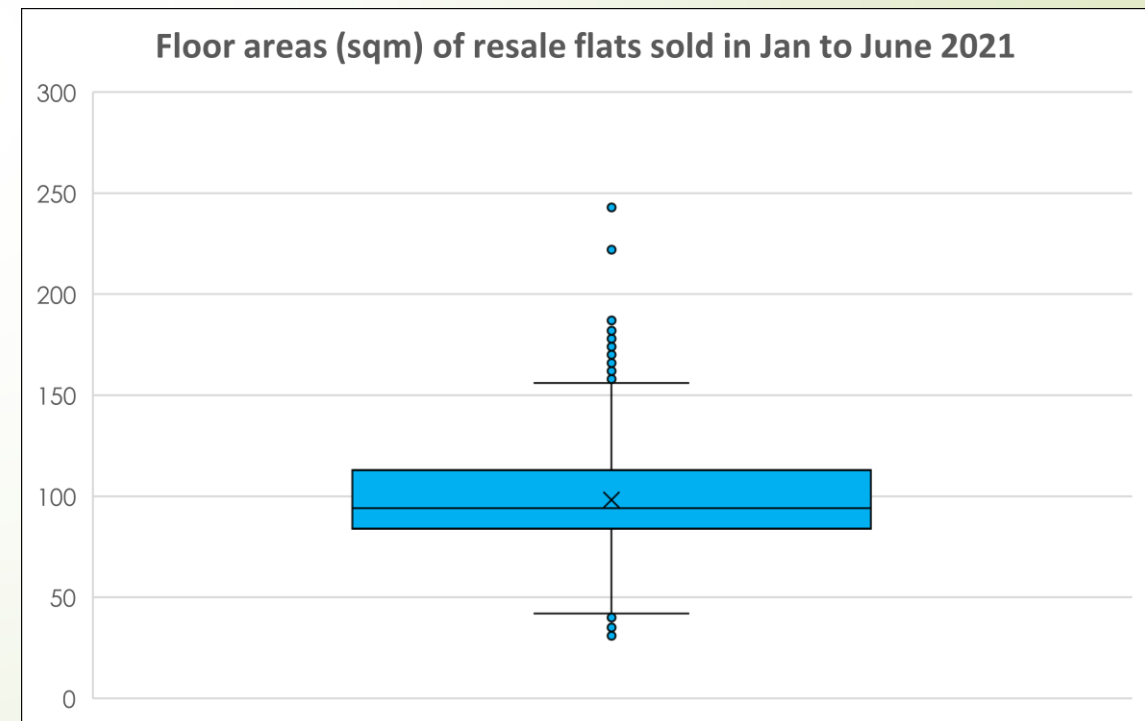
COMPARING BOXPLOTS



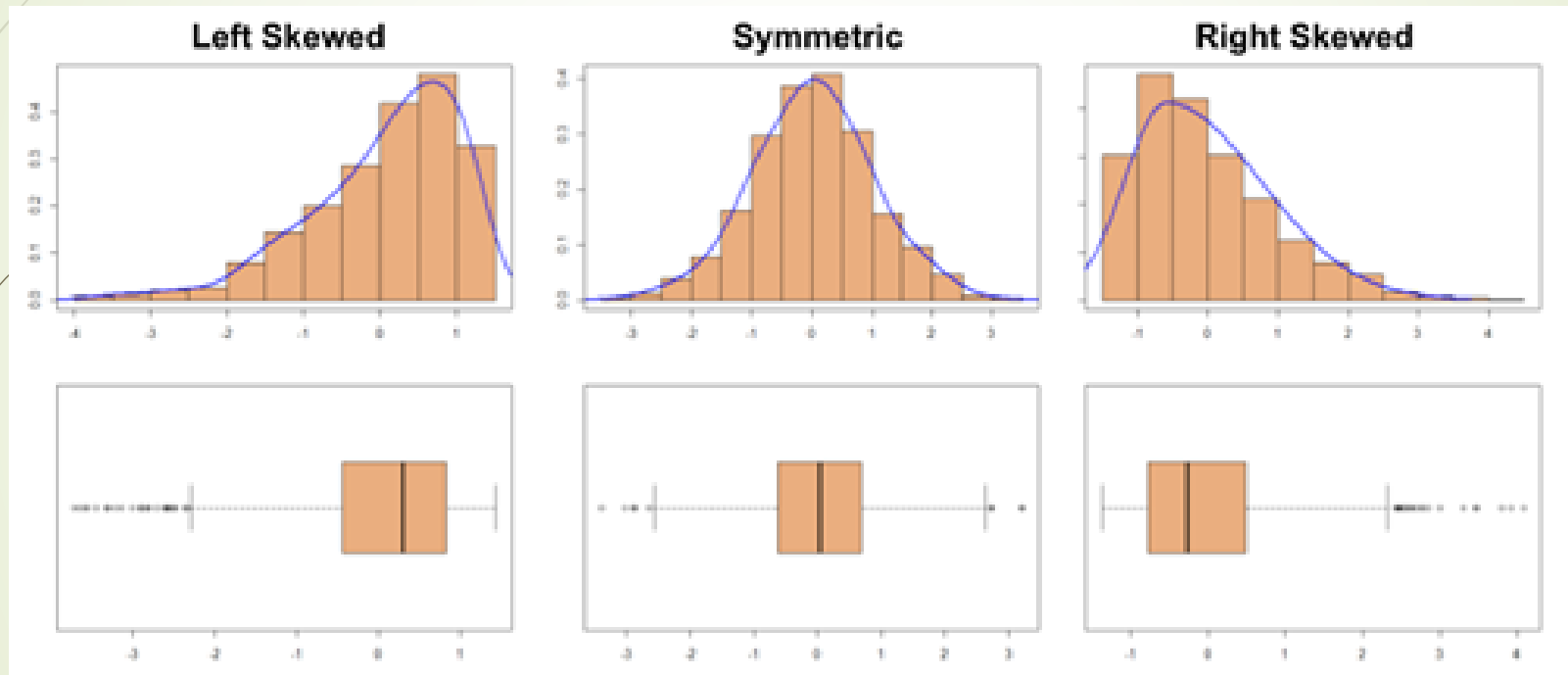
Exercise



Is IQR a robust statistic?

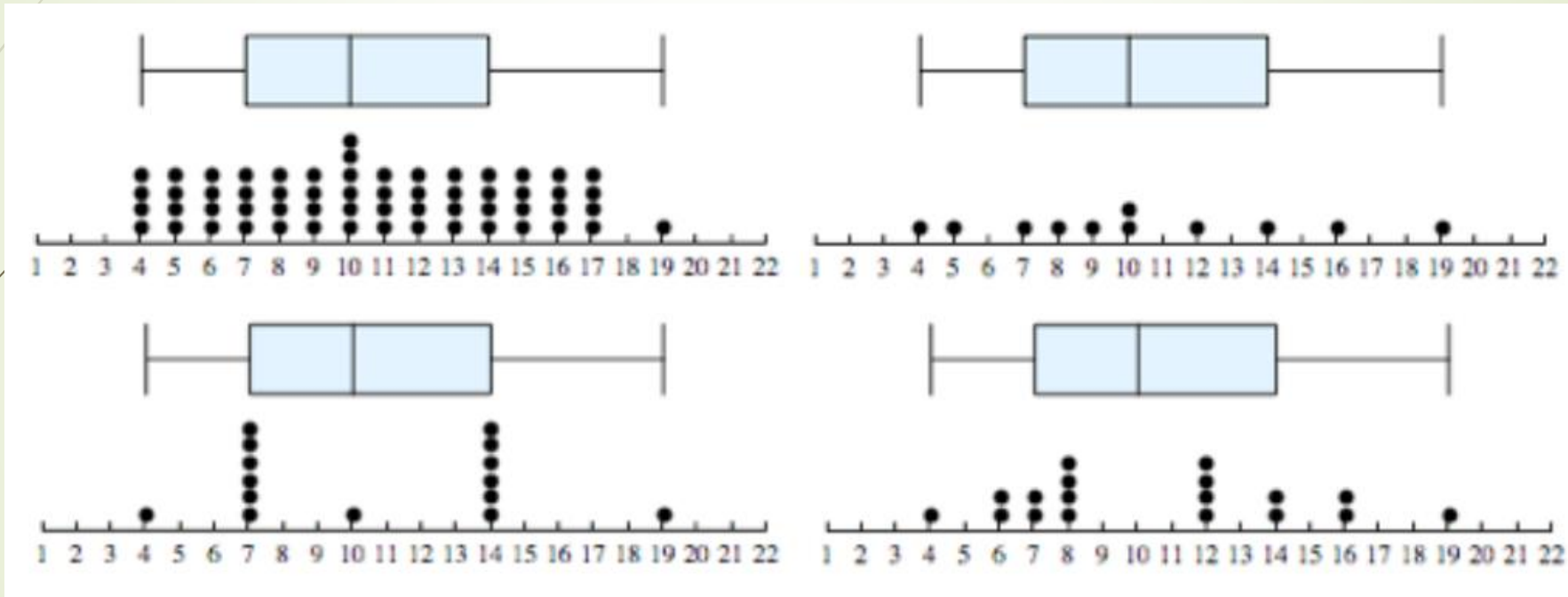


Boxplots vs histograms



Shape? Comparison of data? Outliers? Number of points?

Boxplots vs histograms



Graphics



**Summary
statistics**



Bivariate Exploratory Data Analysis (an introduction)

Previously...

Type of Research Questions	Examples
Make an estimate about the population	What is the average number of hours that students study each week?
	What proportion of all Singapore students is enrolled in a university?
Test a claim about the population	Is the average course load for a university student greater than 20 units?
	Does the majority of students qualify for student loans?
Compare two sub-populations	In university X, do female students have a higher GPA score than male students?
	Are student athletes more likely than non-athletes to do final year projects?
Investigate a relationship between two variables in the population	Is there a relationship between the average number of hours students spend each week on Facebook and their GPA?
	Does drinking coffee help students pass the math exam?

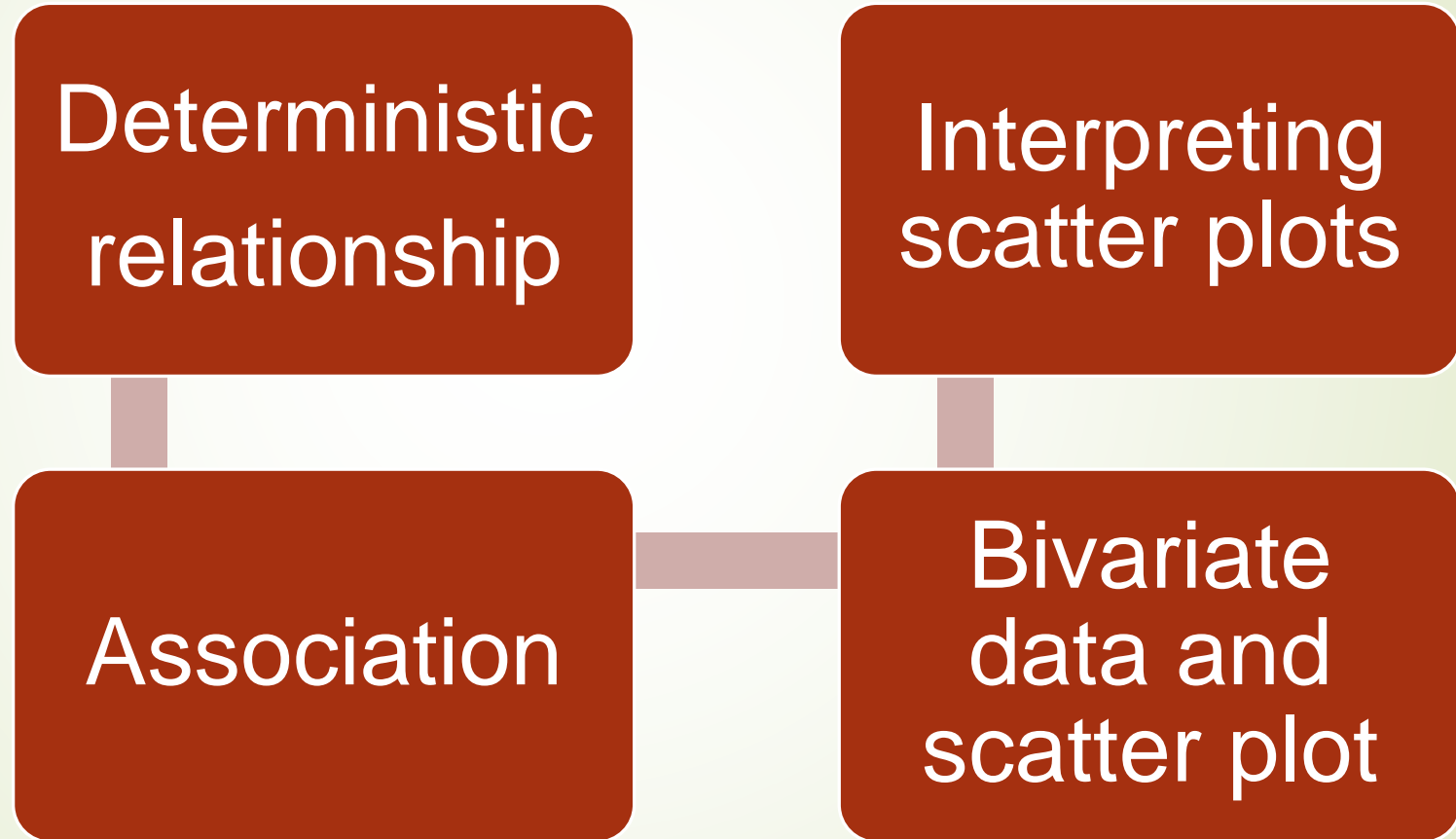
Outline of unit

Deterministic
relationship

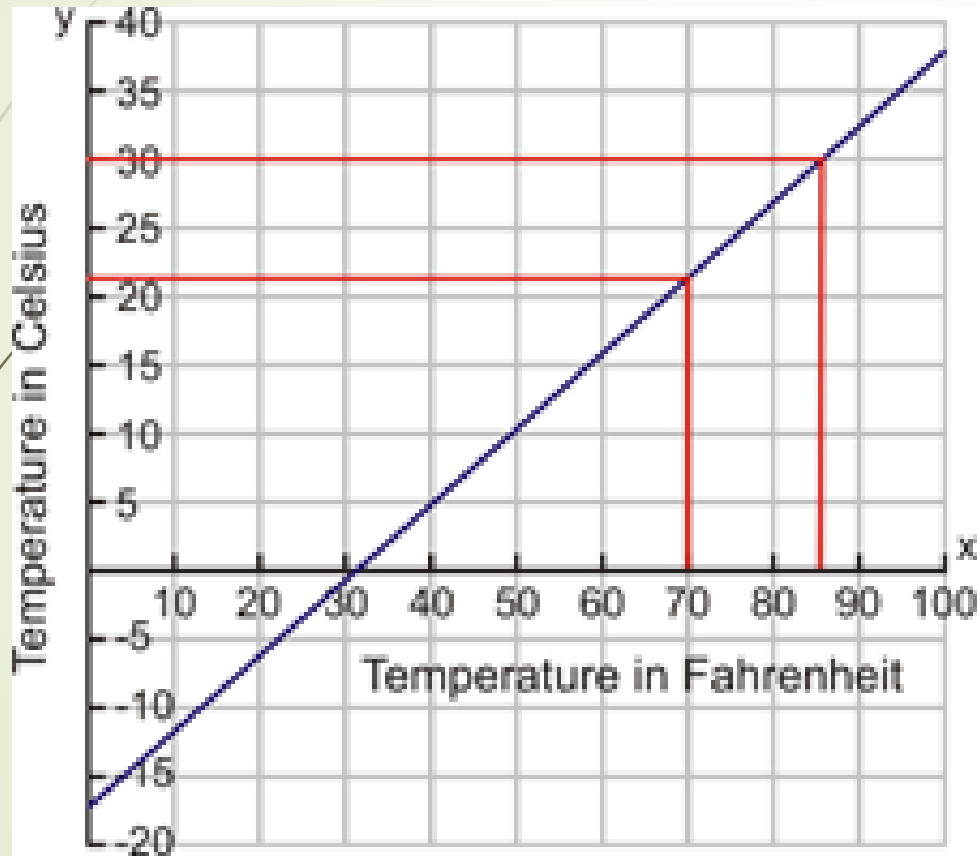
Interpreting
scatter plots

Association

Bivariate
data and
scatter plot



Deterministic relationship



$$\text{Degree Celsius} = (\text{Fahrenheit} - 32) \times (5/9)$$

Deterministic relationship

The Stuffed Chicken Breast Recipe Everyone Will Love

★★★★★ 4.9 from 18 ratings [Print](#)

We all need a fancy meal that is sure to impress, but actually only takes a few minutes of prep. Here's the stuffed chicken breast recipe everyone will love

PREP TIME
15
minutes

COOK TIME
30
minutes

SERVINGS
4
servings



TOTAL TIME: 45 MINUTES

Directions

1. Preheat the oven to 450 degrees Fahrenheit and soak 12 toothpicks in water. Place each chicken breast on a cutting board and cut a slit $\frac{3}{4}$ way through. Place the prepared chicken breasts in a high-rimmed baking sheet.

Temperature in degrees

$$\begin{aligned} &= (\text{Fahrenheit} - 32) \times (5/9) \\ &= (450 - 32) \times (5/9) \\ &= 232.22 \end{aligned}$$

Association

Statistical relationship

- Natural variability exists in measurements of two variables.
- Average value of one variable can be described given the value of the other variable.



Association

Large study finds clear association between fitness and mental health

New research from a large study demonstrates that low cardiorespiratory fitness and muscle strength have a significant association with worse mental health.



Luis Alvarez / Getty Images

DOES BETTER FITNESS
MAKE A PERSON MENTALLY
HEALTHIER?

OR

DOES BETTER MENTAL
HEALTH MAKE A PERSON
EXERCISE MORE AND BECOME
FITTER?

OR

THE ASSOCIATION IS DUE TO
OTHER REASONS?

Association

TGA says there is 'no likely association' between COVID-19 vaccine and recent deaths of two men in NSW



The TGA says current evidence does not suggest a link between the death of two men in New South Wales and the COVID-19 vaccine they received beforehand. Source: Pexels



Do not jump to conclusions easily!

Bivariate data and scatter plot

Bivariate data
=
“2 variables” data

	A	B	C	D
1	month	floor_area_sqm	age	resale_price
2	1/1/2021	45	35	225000
3	1/1/2021	45	35	211000
4	1/1/2021	73	45	275888
5	1/1/2021	67	43	316800
6	1/1/2021	67	43	305000
7	1/1/2021	68	40	260000
8	1/1/2021	73	44	351000
9	1/1/2021	73	44	343000
10	1/1/2021	75	41	306000

A picture is worth 1000 words!

Bivariate data and scatter plot

Bivariate Data Analysis

Scatter plots

- Get an idea of the pattern

Correlation coefficients

- Check for linear relationship

Regression analysis

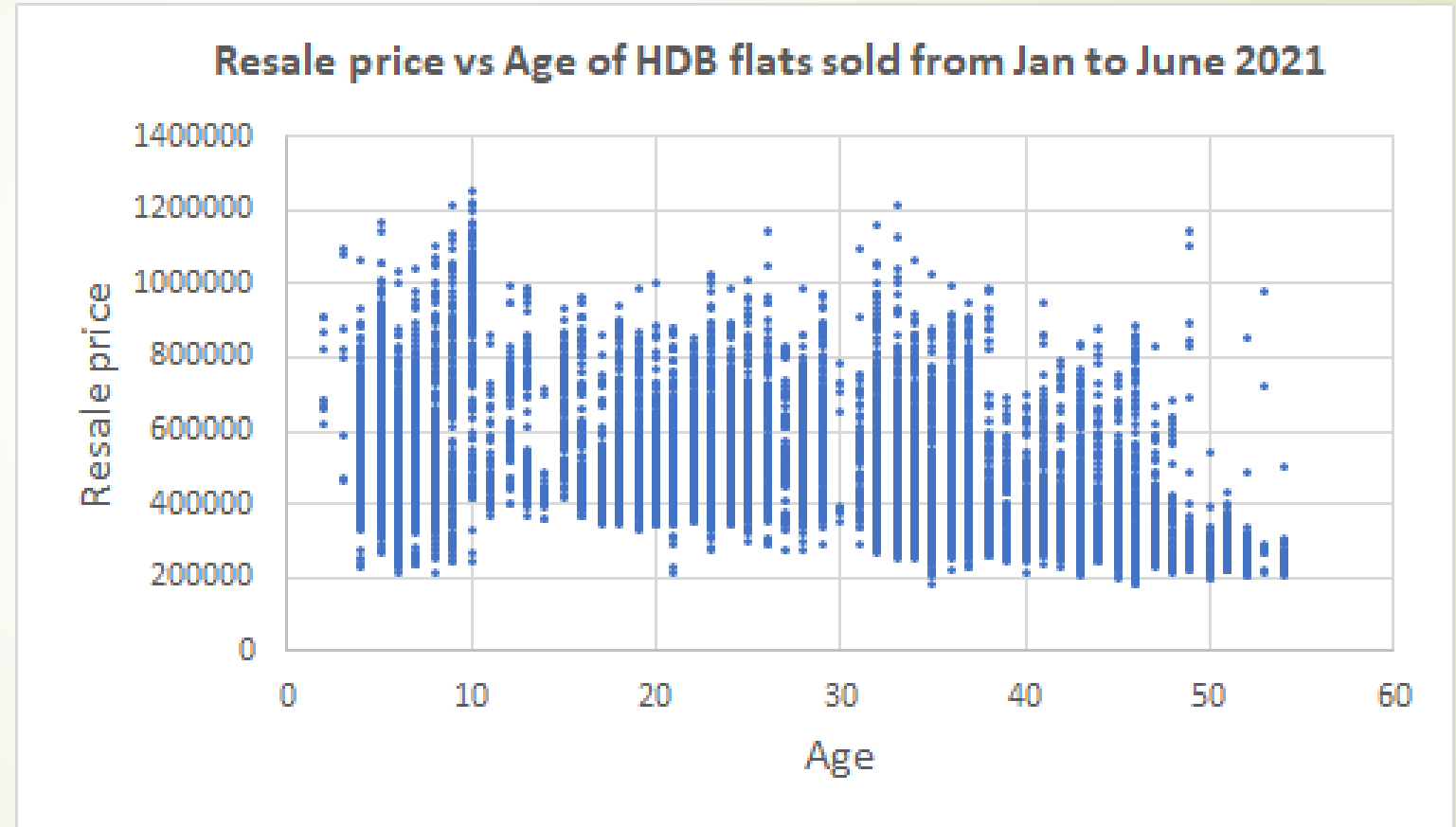
- Fit a line or curve to data

Bivariate data and scatter plot

Independent
(explanatory)

C	D
age	resale_price
35	225000
35	211000
45	275888
43	316800
43	305000
40	260000
44	351000
44	343000
41	306000

Dependent
(response)



Interpreting scatter plots

**Graph the distribution of
two quantitative variables
in a scatterplot.**

Describe:

Overall pattern

Direction

Form

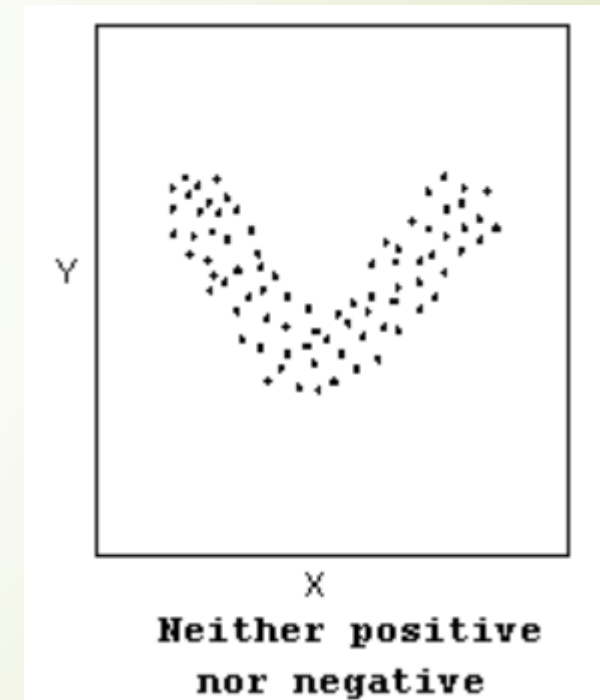
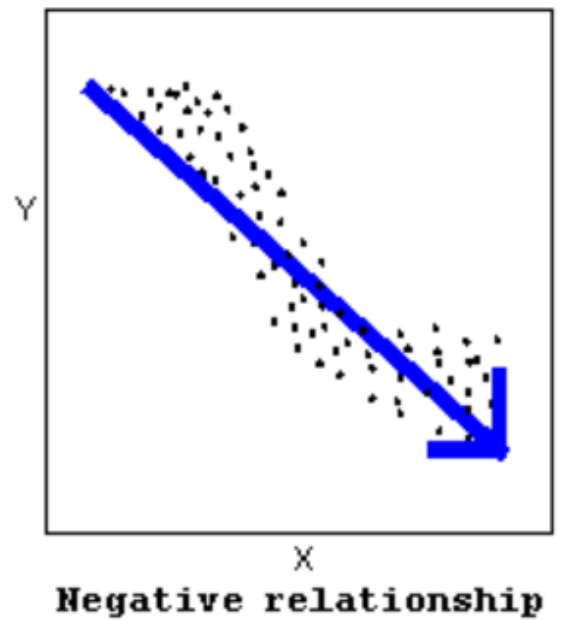
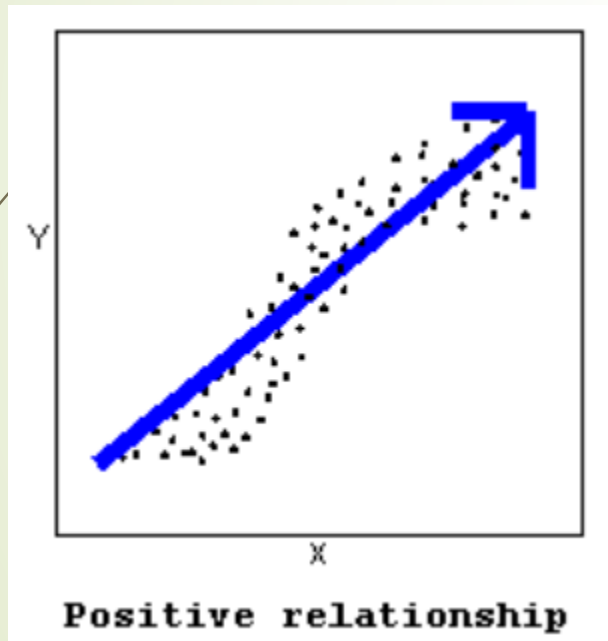
Strength

Deviations from the pattern

Outliers

Interpreting scatter plots

Direction

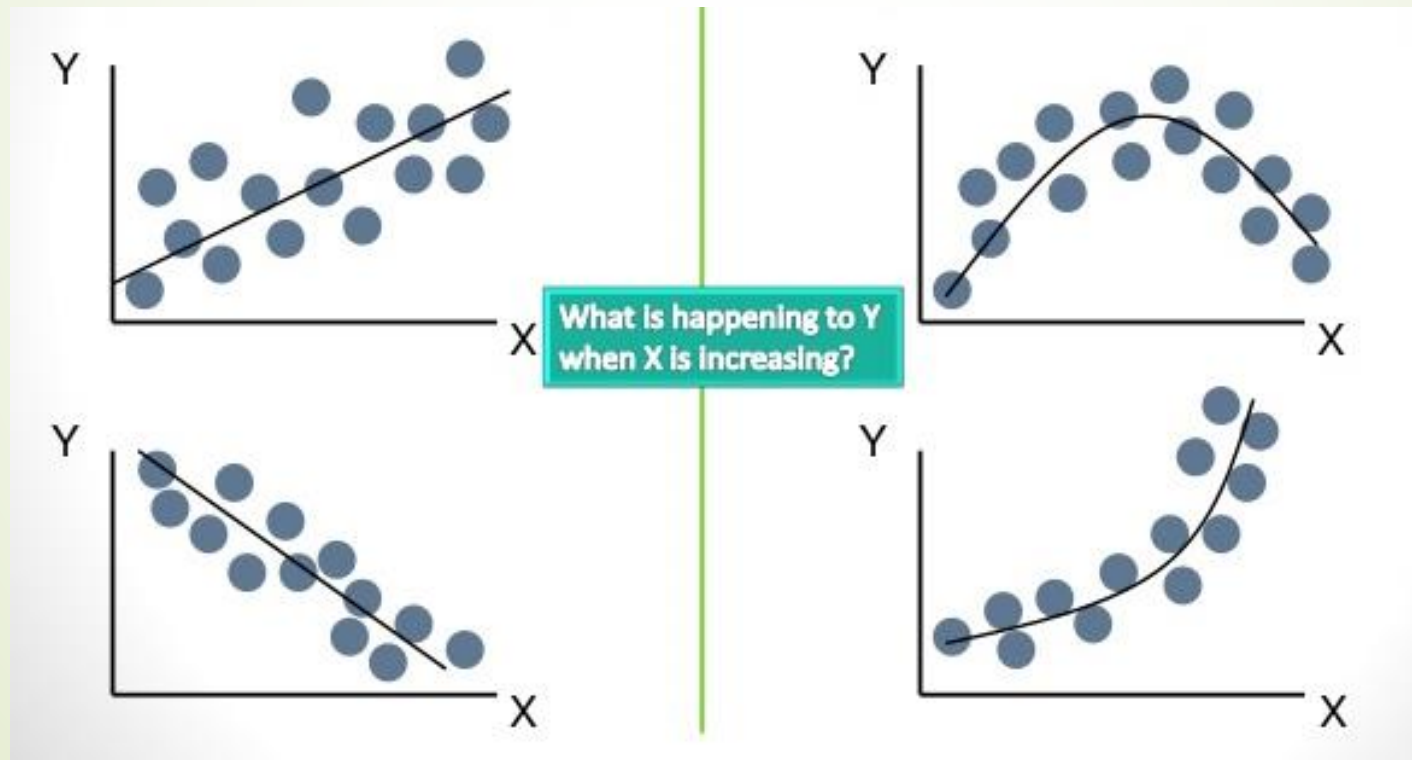


Interpreting scatter plots

Form

Linear

Non-linear

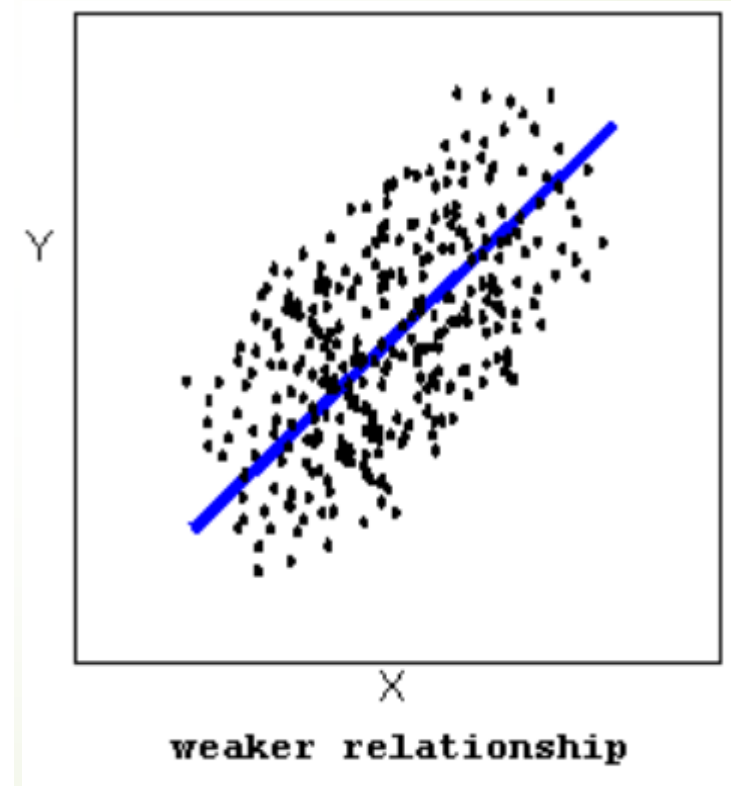
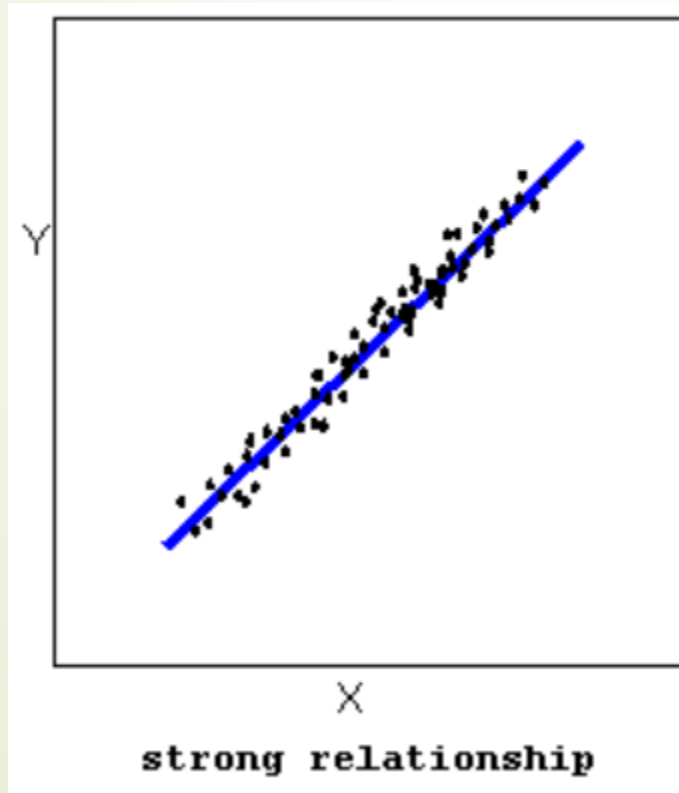


Quadratic

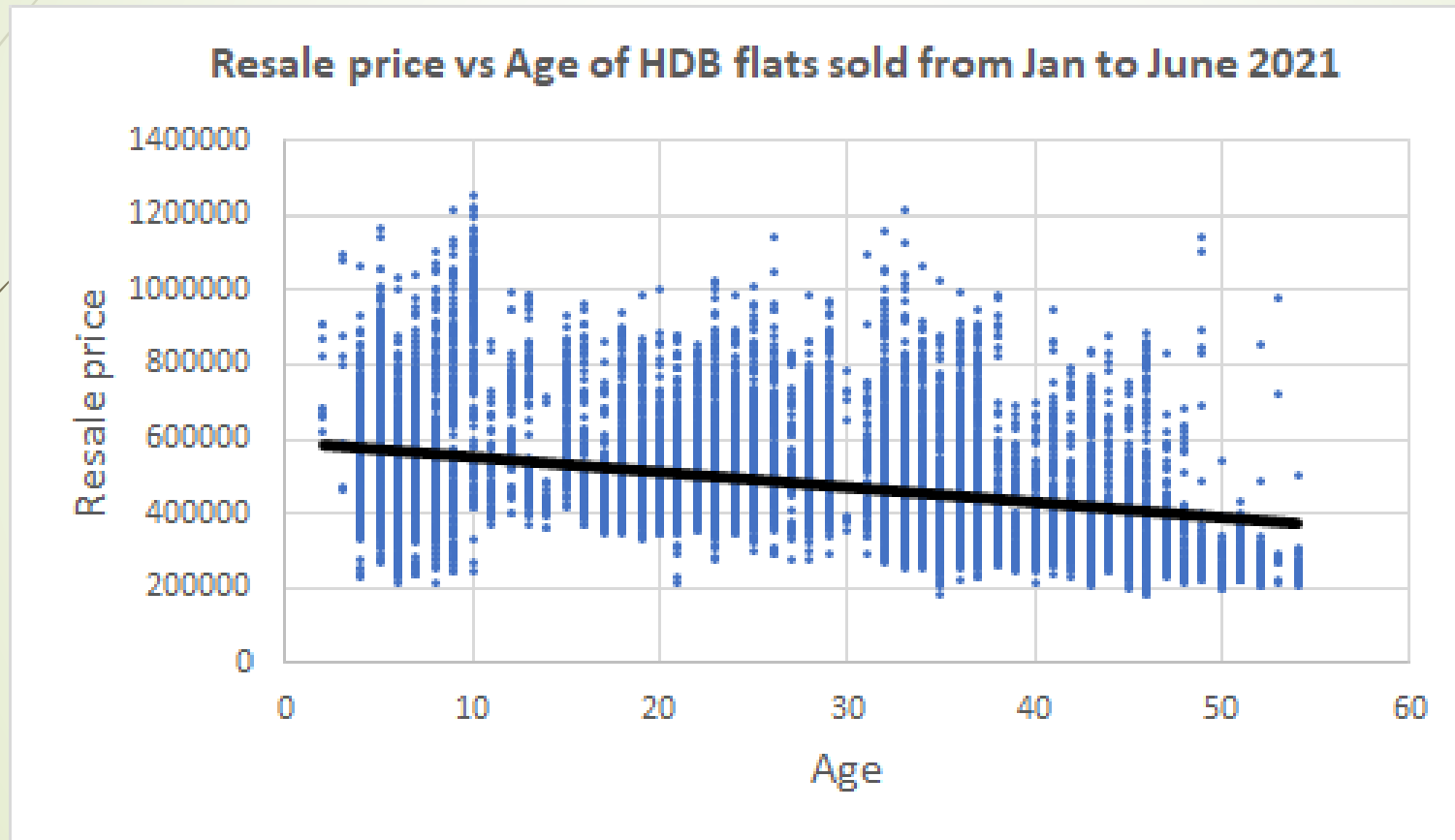
Exponential

Interpreting scatter plots

Strength



Interpreting scatter plots



Direction?

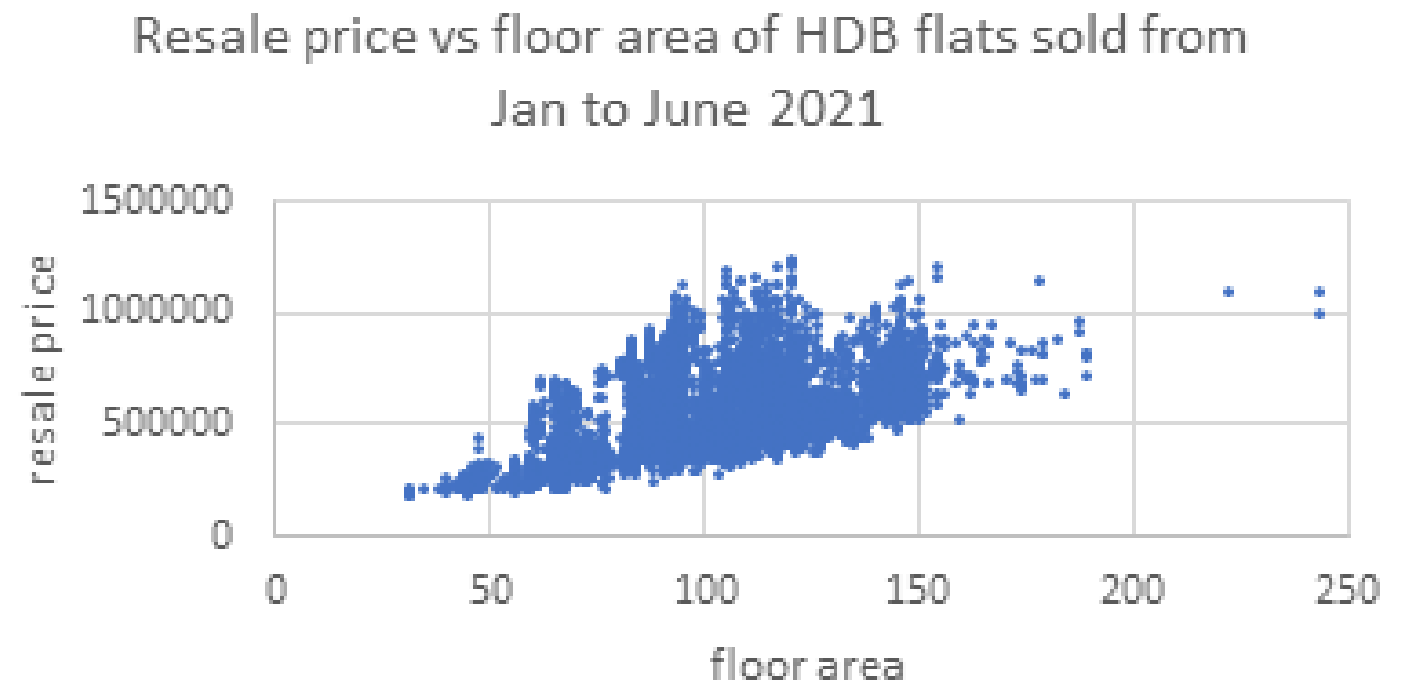
Form?

Strength?

Interpreting scatter plots

Outliers

	A	B	C	D
1	month	floor_area_sqm	age	resale_price
2	1/1/2021	45	35	225000
3	1/1/2021	45	35	211000
4	1/1/2021	73	45	275888
5	1/1/2021	67	43	316800
6	1/1/2021	67	43	305000
7	1/1/2021	68	40	260000
8	1/1/2021	73	44	351000
9	1/1/2021	73	44	343000
10	1/1/2021	75	41	306000





In summary...

- ☐ Deterministic relationship between two variables
- ☐ Association between two variables
- ☐ Bivariate data and scatter plots
- ☐ How to interpret scatter plots?

Correlation Coefficient





Learning Objectives

- By the end of this unit, you should be able to do the following:
 1. Interpret correlation coefficient
 2. Understand and apply properties of correlation coefficient

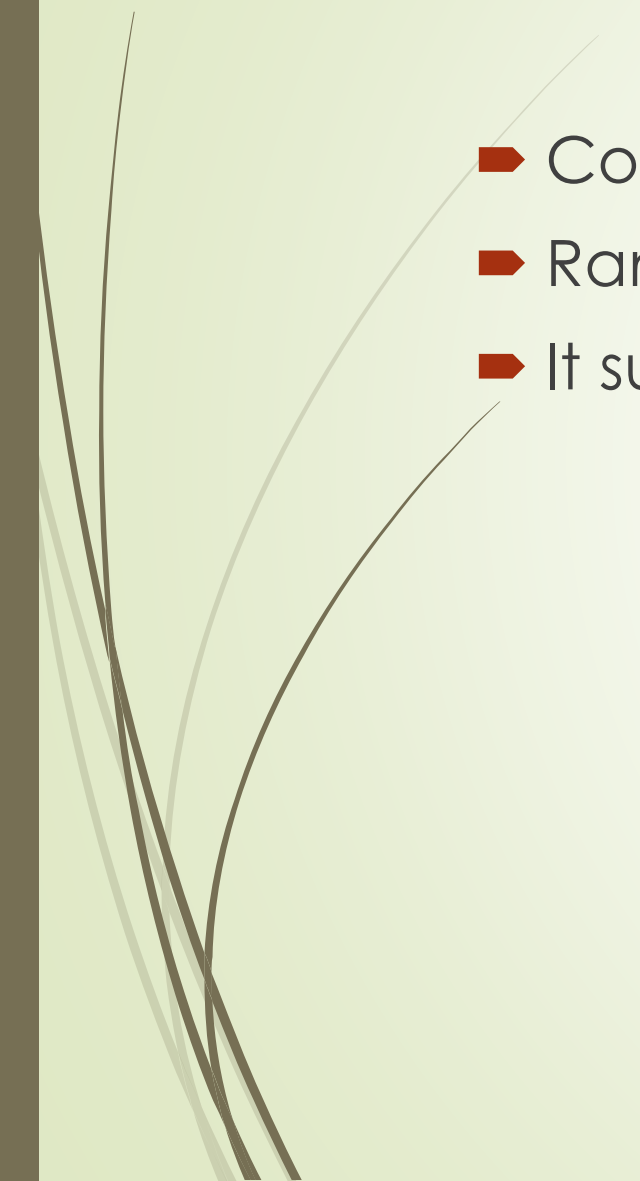
Question

- How do we measure quantitatively the relationship between the flats' resale price and age?

	A	B	C	D
1	month	floor_area_sqm	age	resale_price
2	1/1/2021	45	35	225000
3	1/1/2021	45	35	211000
4	1/1/2021	73	45	275888
5	1/1/2021	67	43	316800
6	1/1/2021	67	43	305000
7	1/1/2021	68	40	260000
8	1/1/2021	73	44	351000
9	1/1/2021	73	44	343000
10	1/1/2021	75	41	306000

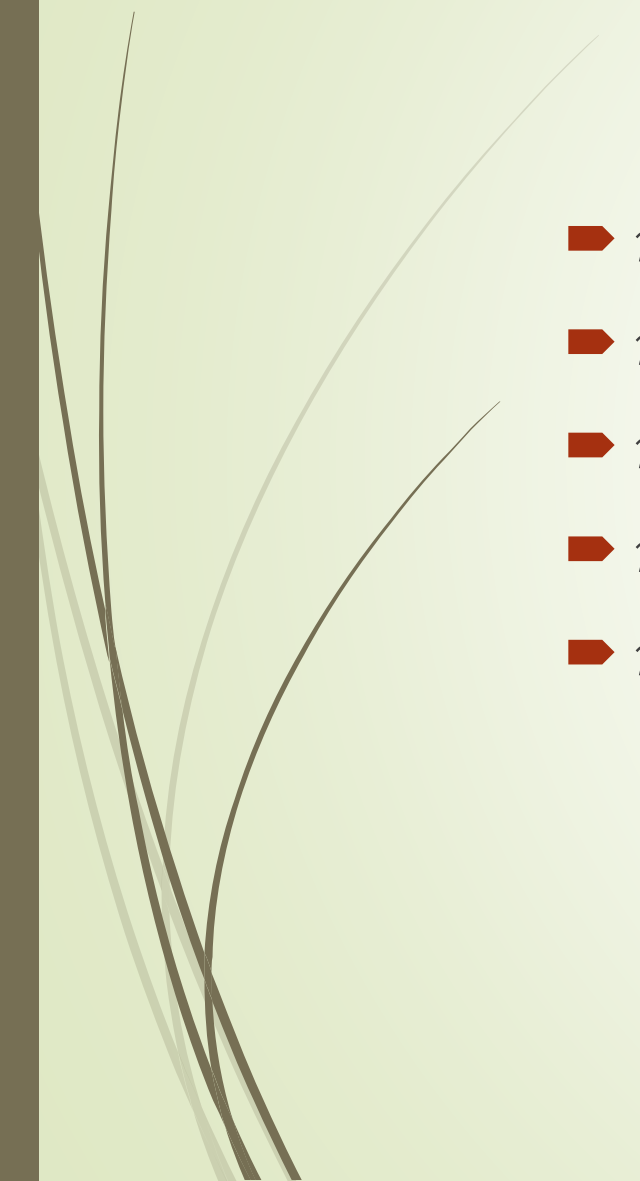


Correlation Coefficient

- Correlation coefficient is a measure of linear association
 - Range is between -1 and 1
 - It summarizes direction and strength of linear association
- 

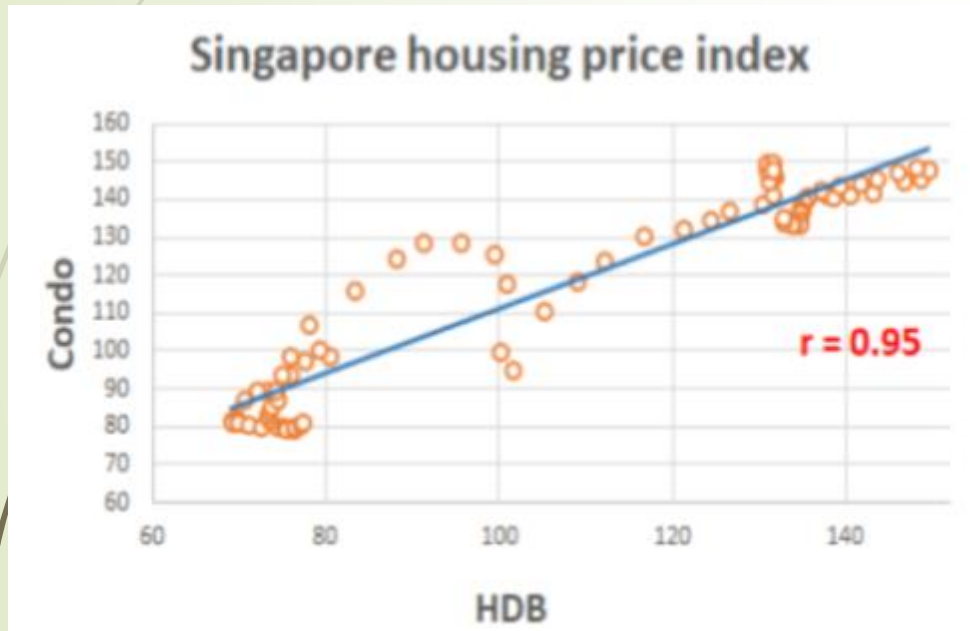


Interpreting r value

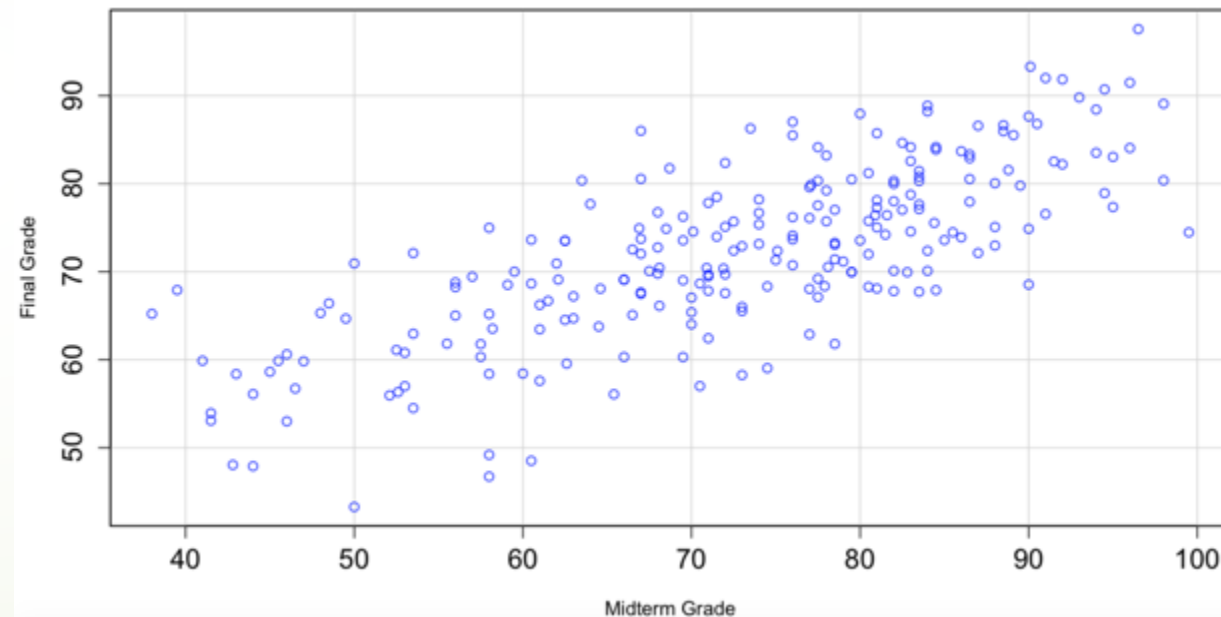
- ▶ $r > 0 \rightarrow$ positive association
 - ▶ $r < 0 \rightarrow$ negative association
 - ▶ $r = 0 \rightarrow$ no linear association
 - ▶ $r = 1 \rightarrow$ perfect positive association
 - ▶ $r = -1 \rightarrow$ perfect negative association
- 

Some Examples ($r > 0$)

- Correlation between Singapore housing price index of HDB and condo



- Correlation between students' midterm grade and final grade ($r = 0.75$).

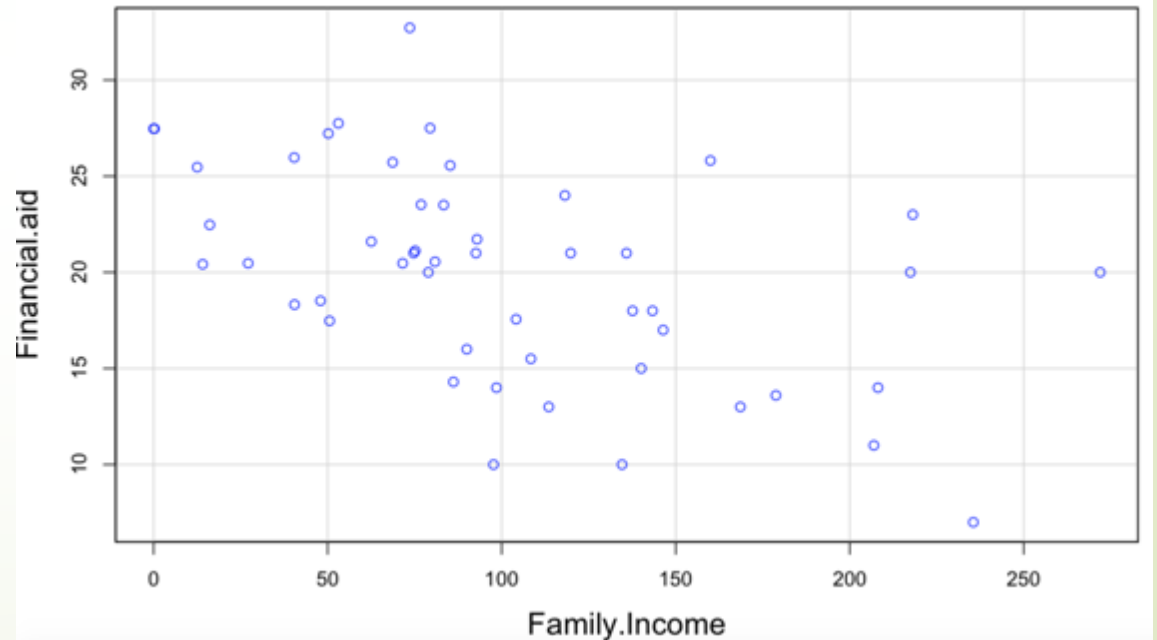


Some Examples ($r < 0$)

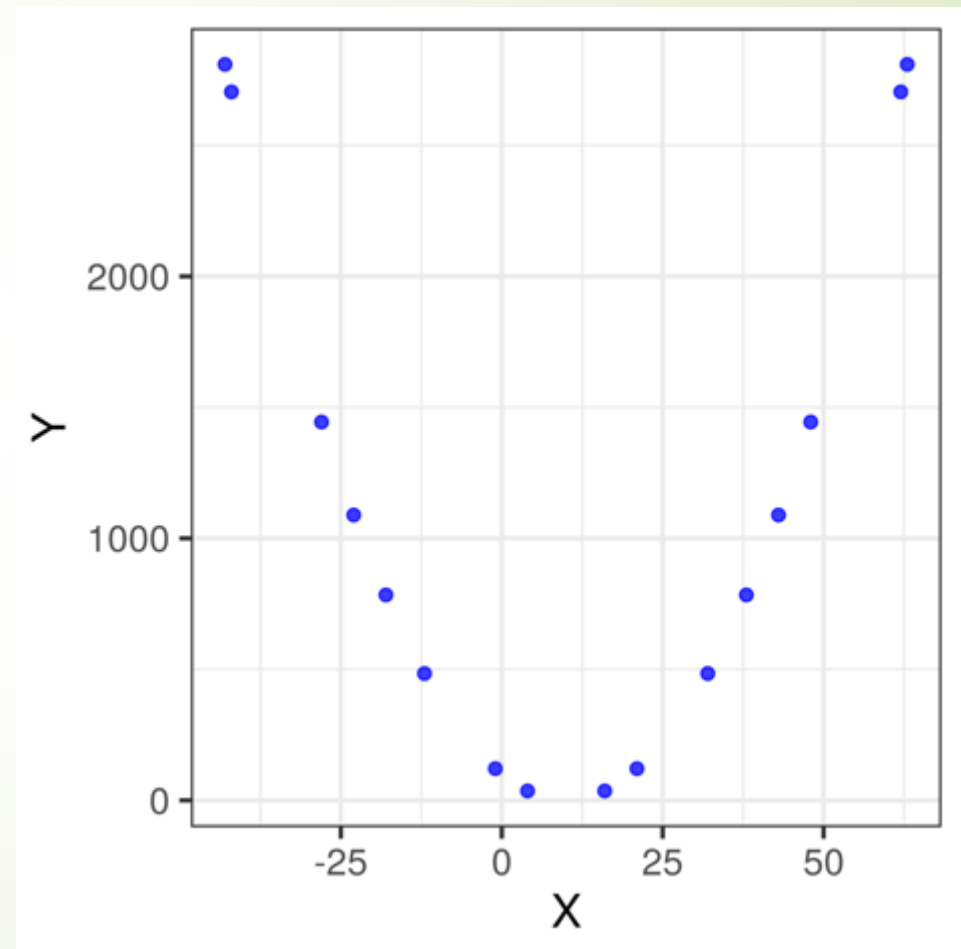
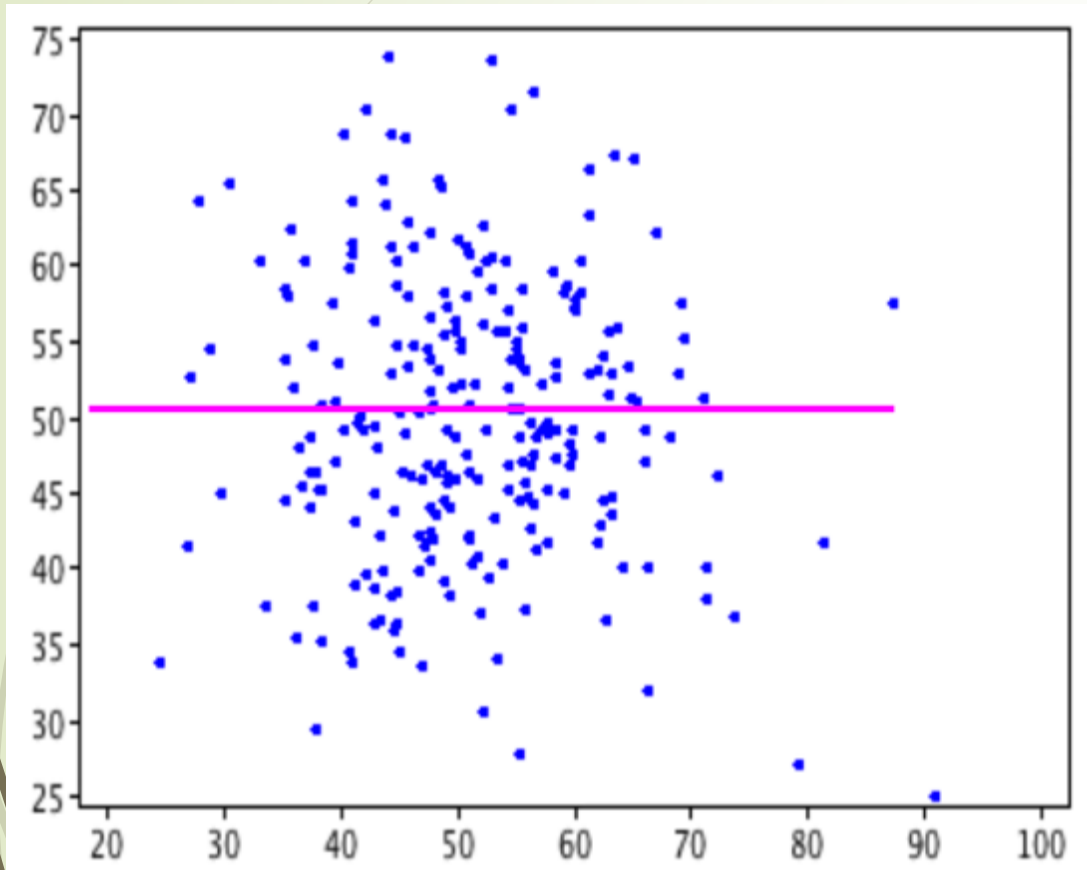
- Correlation between gold and oil price



- Correlation between financial aid and family income of students ($r = -0.49$).

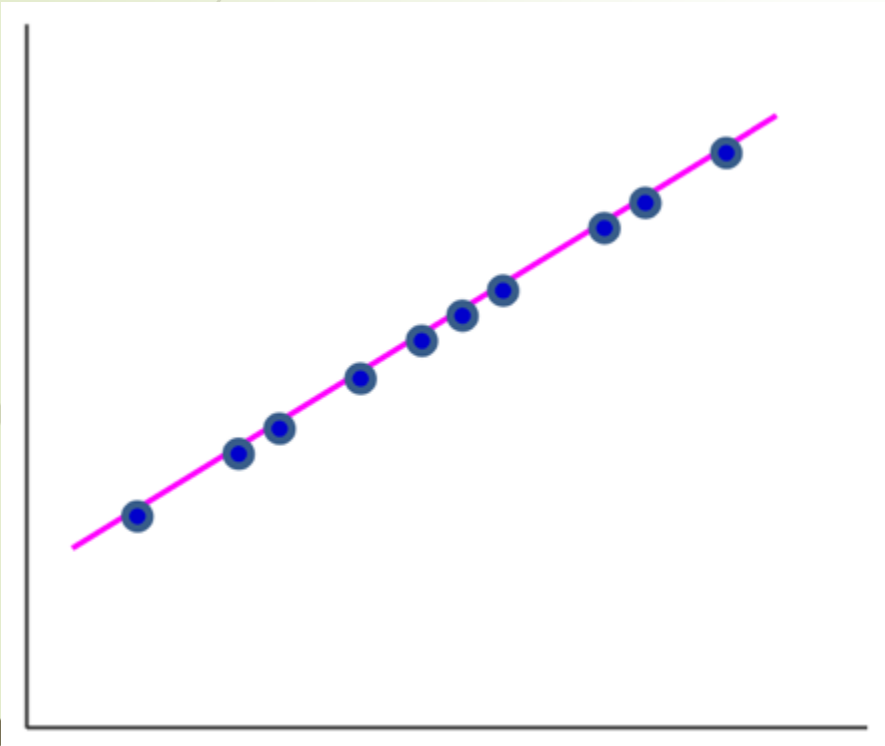


Some Examples $r = 0$

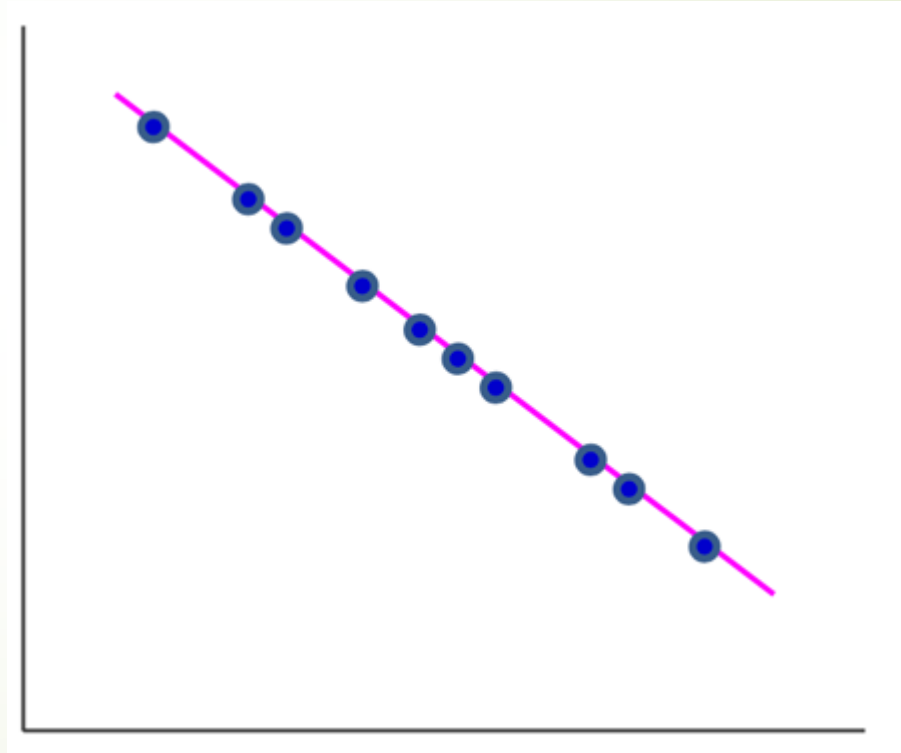


Some Examples

➤ $r = 1$

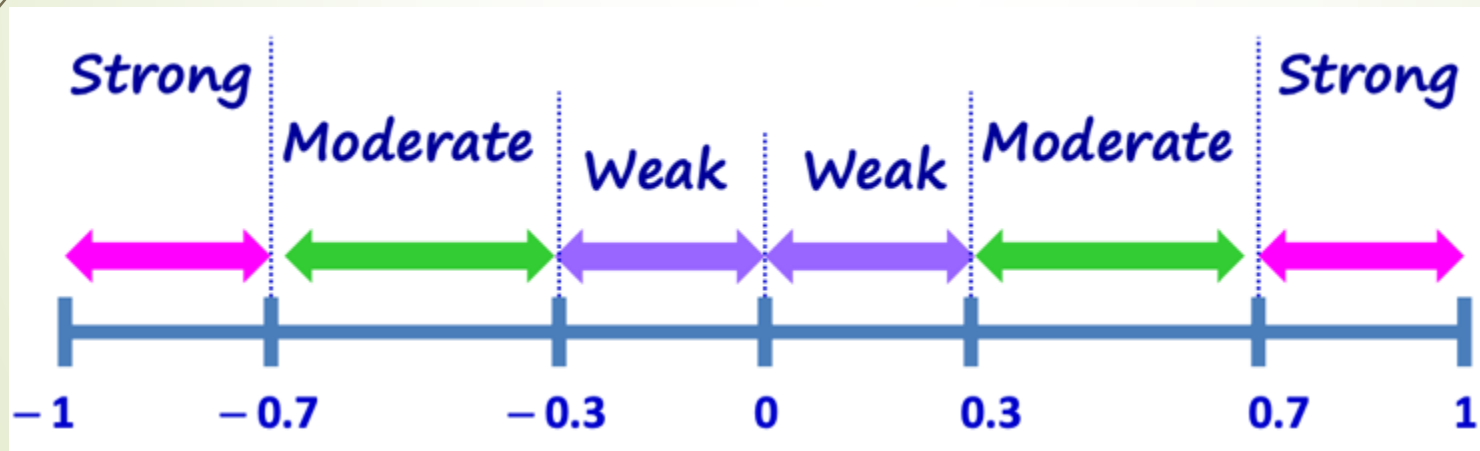


➤ $r = -1$



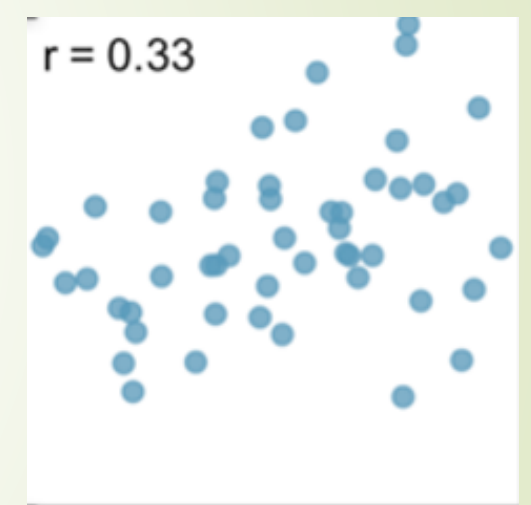
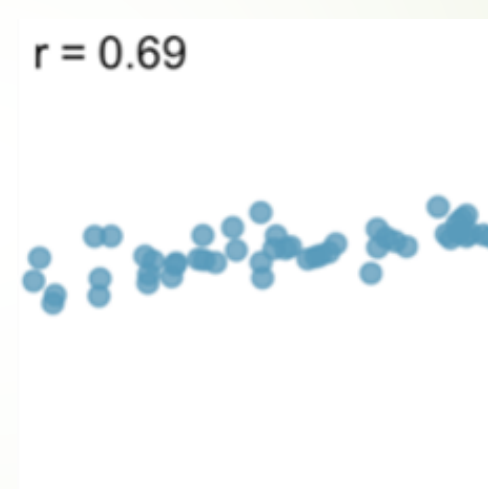
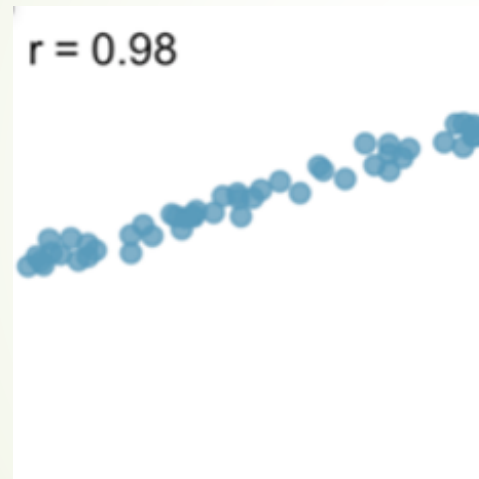
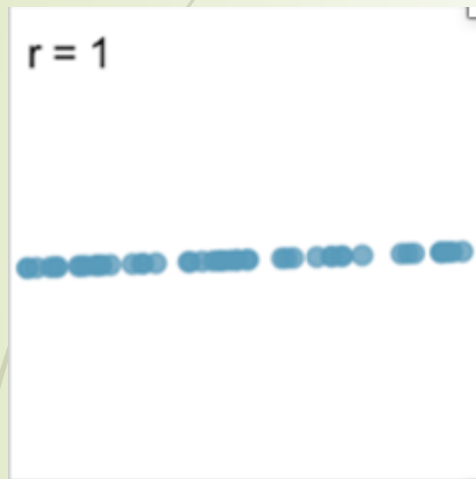
Interpreting r value

- The magnitude of r tells us about the strength of linear association
- The closer the value of r is to 1 or -1, the stronger the linear association
- The closer the value is to 0, the weaker the linear association
- Some rule of thumb



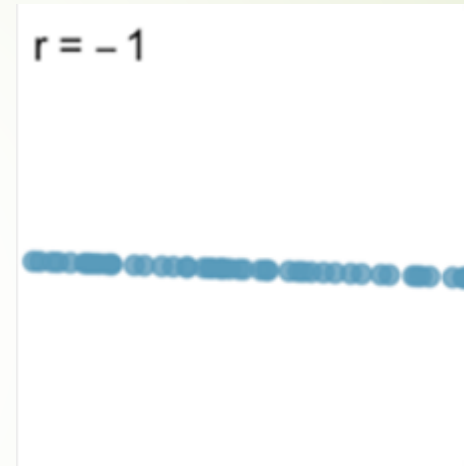
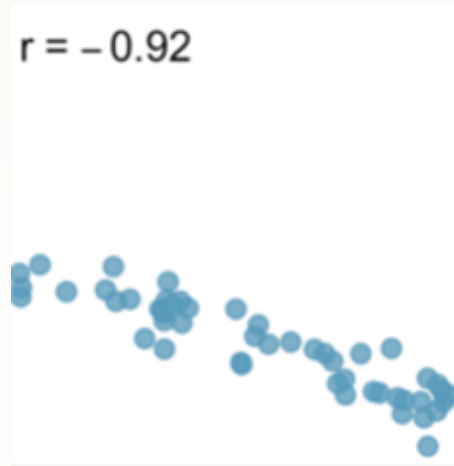
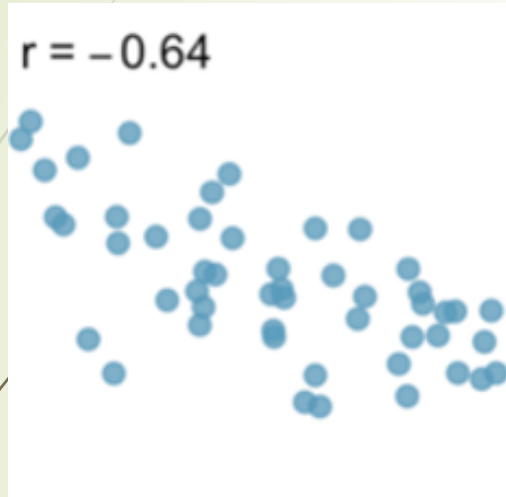
Interpreting r value

- As the value of r gets closer to 1 or -1, the data falls more closely to a straight line.



Picture credit: <https://openintro-ims.netlify.app/model-slr.html>

Interpreting r value



Picture credit: <https://openintro-ims.netlify.app/model-slr.html>

How to compute correlation coefficient?



- First, convert each data point into its standard unit
- $SU_X = \frac{X - \text{average}(X)}{s_x}$ and $SU_Y = \frac{Y - \text{average}(Y)}{s_y}$ where s_x is the standard deviation of X and s_y is the standard deviation of Y
- Second, r value is just the sum of product of X and Y in standard units divided by $n - 1$, where n is the number of data points.
- Note that you are not expected to compute the correlation coefficient by hand.

Example

$$X = 9 \rightarrow \frac{9 - 5.5}{3.03} = 1.16$$

$$Y = 41 \rightarrow \frac{41 - 25.1}{15.65} = 1.02$$

X	Y			
9	41		Average of X	5.5
4	17		Average of Y	25.1
5	28		Standard deviation of X	3.03
10	50		Standard deviation of Y	15.65
6	39			
3	26			
7	30			
2	6			
8	4			
1	10			

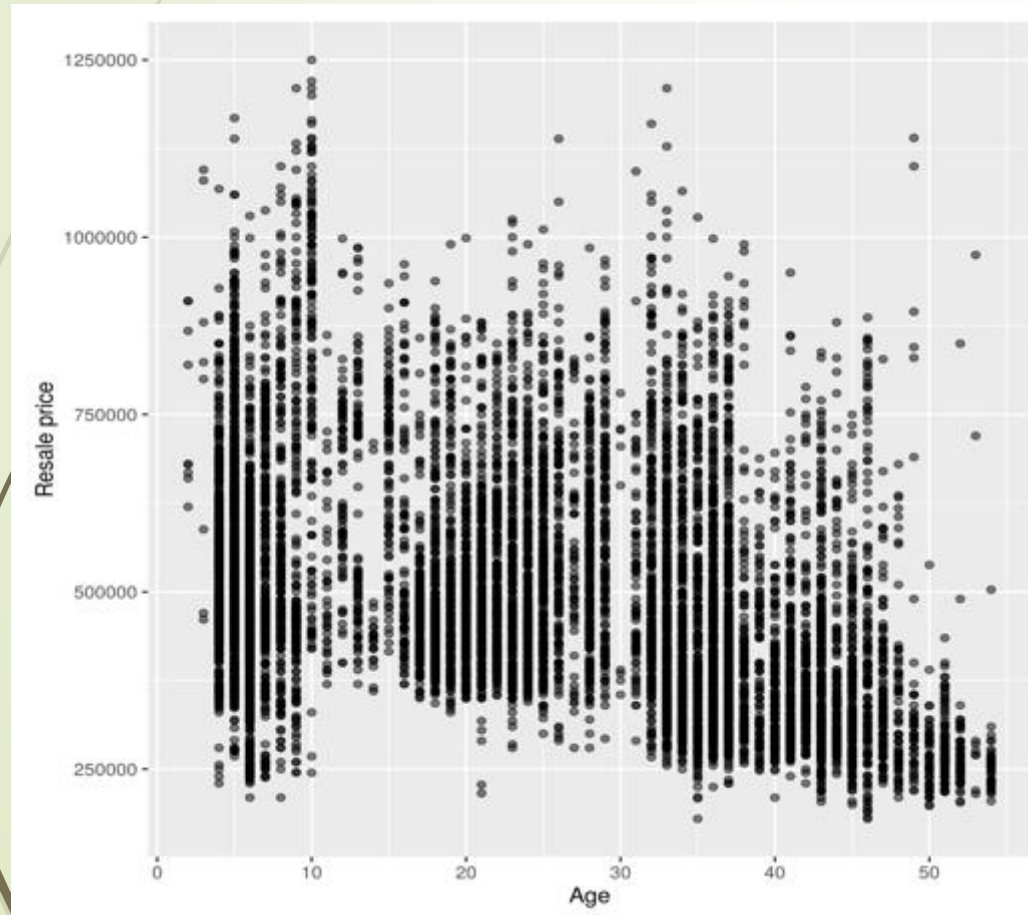



X	Y		X (standard unit)	Y (standard unit)	Product
9	41		1.16	1.02	1.17
4	17		-0.50	-0.52	0.26
5	28		-0.17	0.19	-0.03
10	50		1.49	1.59	2.36
6	39		0.17	0.89	0.15
3	26		-0.83	0.06	-0.05
7	30		0.50	0.31	0.15
2	6		-1.16	-1.22	1.41
8	4		0.83	-1.35	-1.11
1	10		-1.49	-0.96	1.43

$$r = \frac{1}{9} (1.17 + 0.26 + \dots + 1.43) = 0.64$$

Example (HDB data set)

The scatter plot of HDB resale price vs age and the correlation coefficient between resale price and age



Correlation

```
Data      : housing
Method    : Pearson
Variables  : age, resale_price
Null hyp.  : variables x and y are not correlated
Alt. hyp.  : variables x and y are correlated
```

Correlation matrix:

```
          age
resale_price -0.36
```

Questions to ponder

- The correlation coefficient between the resale price and the age of HDB flats is -0.36.
- Question: What happens to this value if we
 1. Interchange the y and the x axes?
 2. Increase the resale price by 10 000 SGD for every data point?
 3. Convert the resale prices to be in USD?

Properties of r

- r is not affected by the following operations:
 1. Interchange of two variables
 2. Adding a number to all values of a variable
 3. Multiplying a positive number to all values of a variable
- **Example:** in the HDB data set, if the resale price is converted to US dollars, the correlation coefficient remains unchanged

Example (HDB data set)

Define new variables:

Newprice=resale_price+10 000

ResalepriceUSD=resale_price in USD



Correlation matrix:

	age	resale_price	newprice
resale_price	-0.36		
newprice	-0.36	1.00	
resalepriceUSD	-0.36	1.00	1.00



Summary

➡ We have learnt

1. How to interpret r value and what it tells us about association between two variables
2. Properties of r



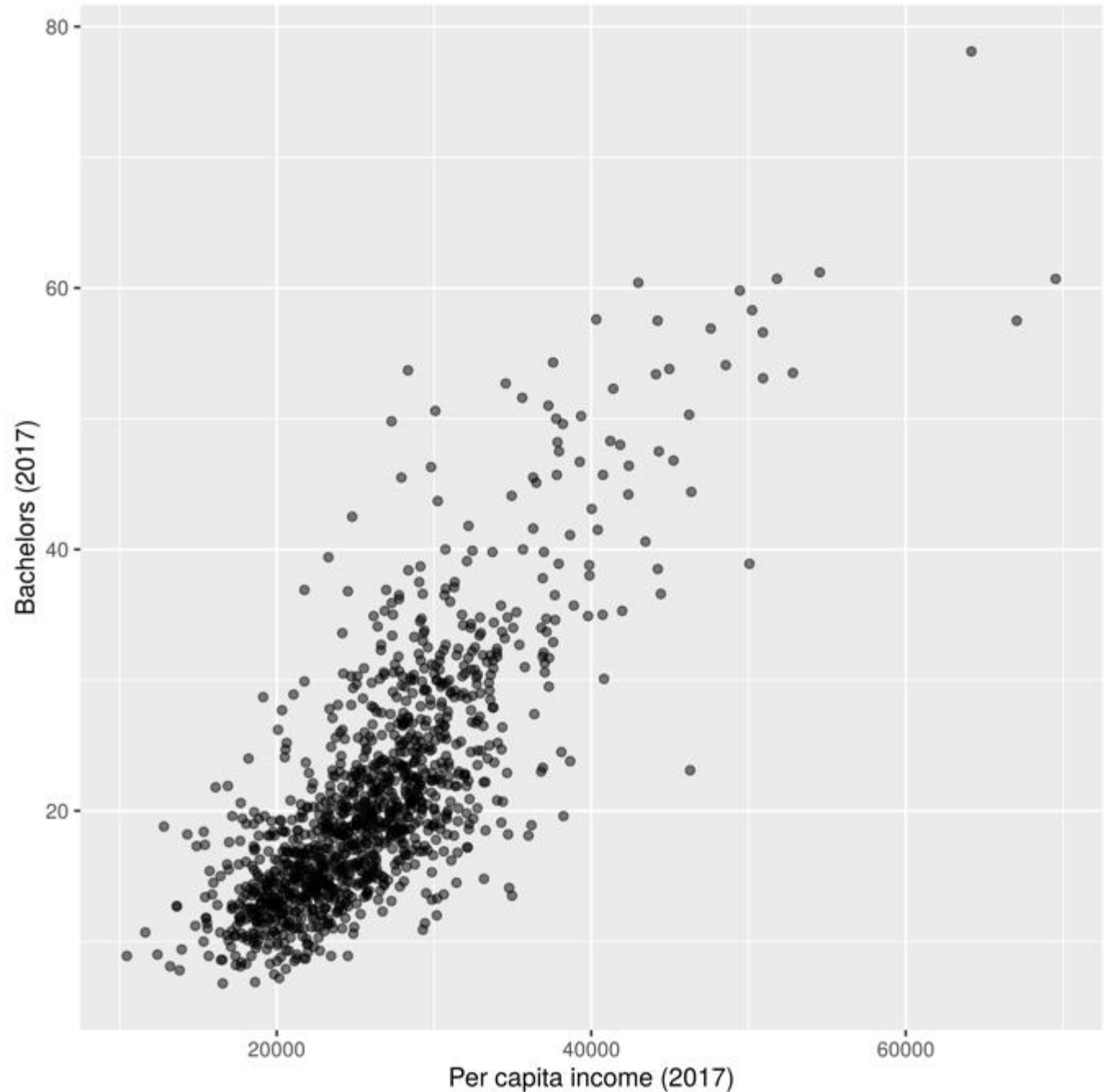
Some Limitations of Correlation Coefficient



Learning Objectives

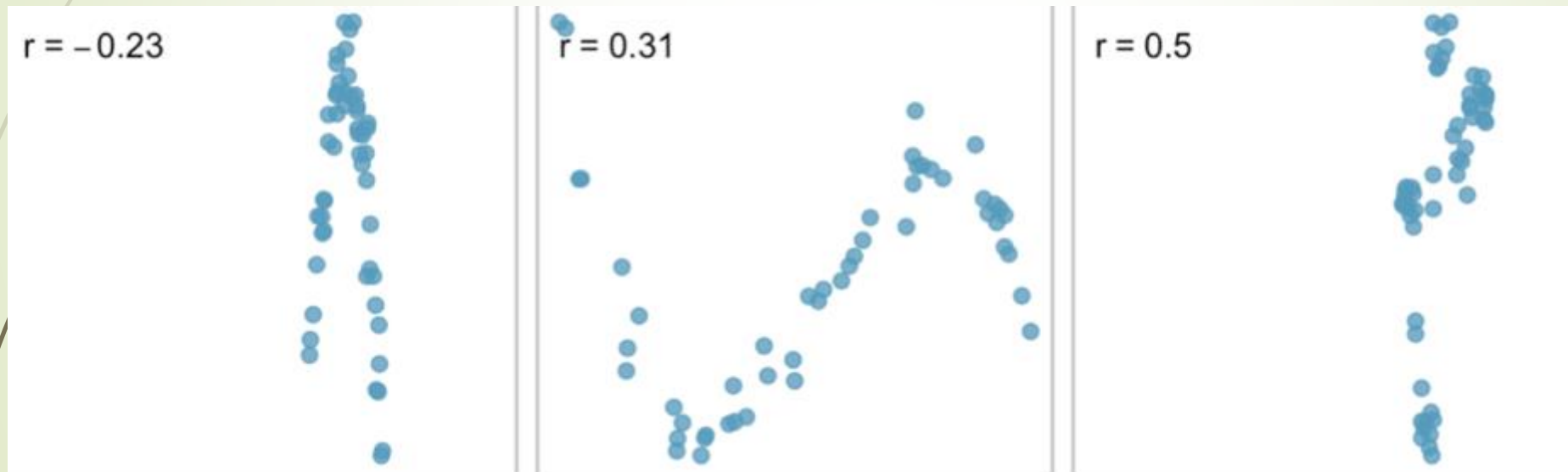
- By the end of this unit, you should be able to
 1. Identify some limitations of correlation coefficient
 2. Identify outliers and how it affects correlation coefficient

- Correlation does not imply causation.
- There might be a third variable correlated to Bachelors and per capita income.



Non-linear association

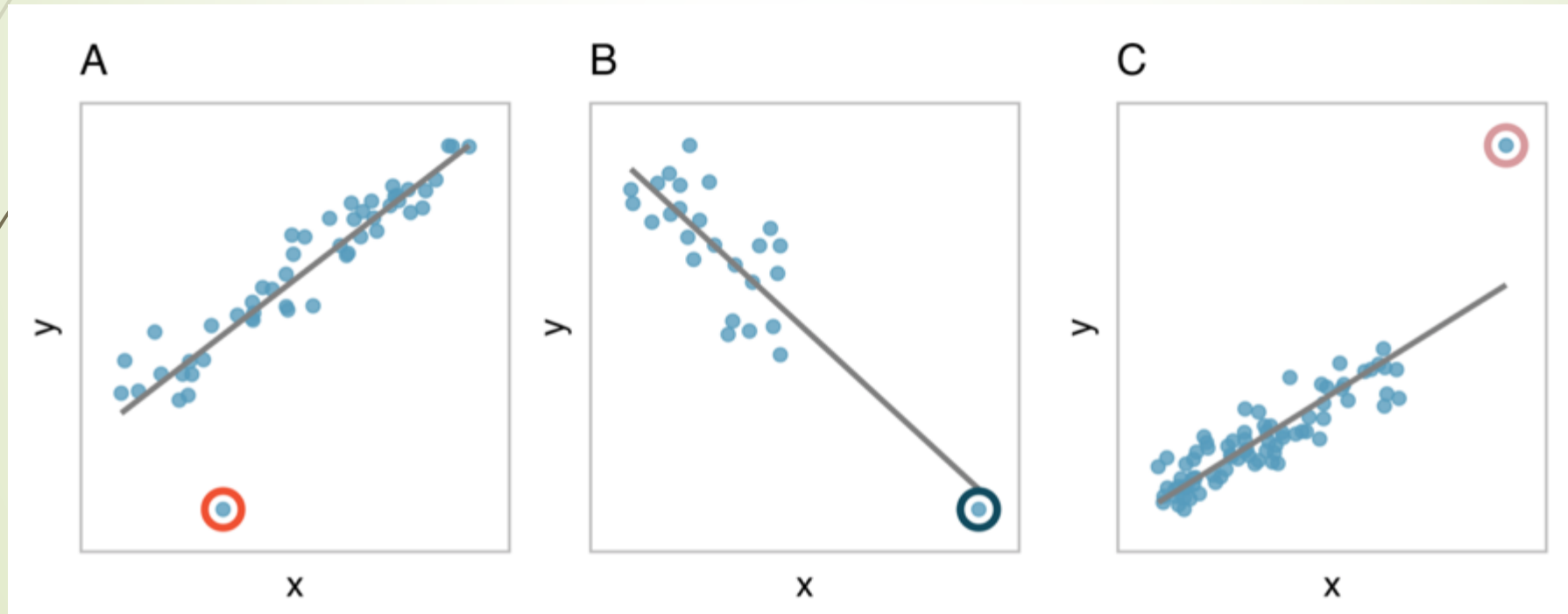
- r only measures linear association between two variables.
- Always look at the scatter plot and not only at the r value.



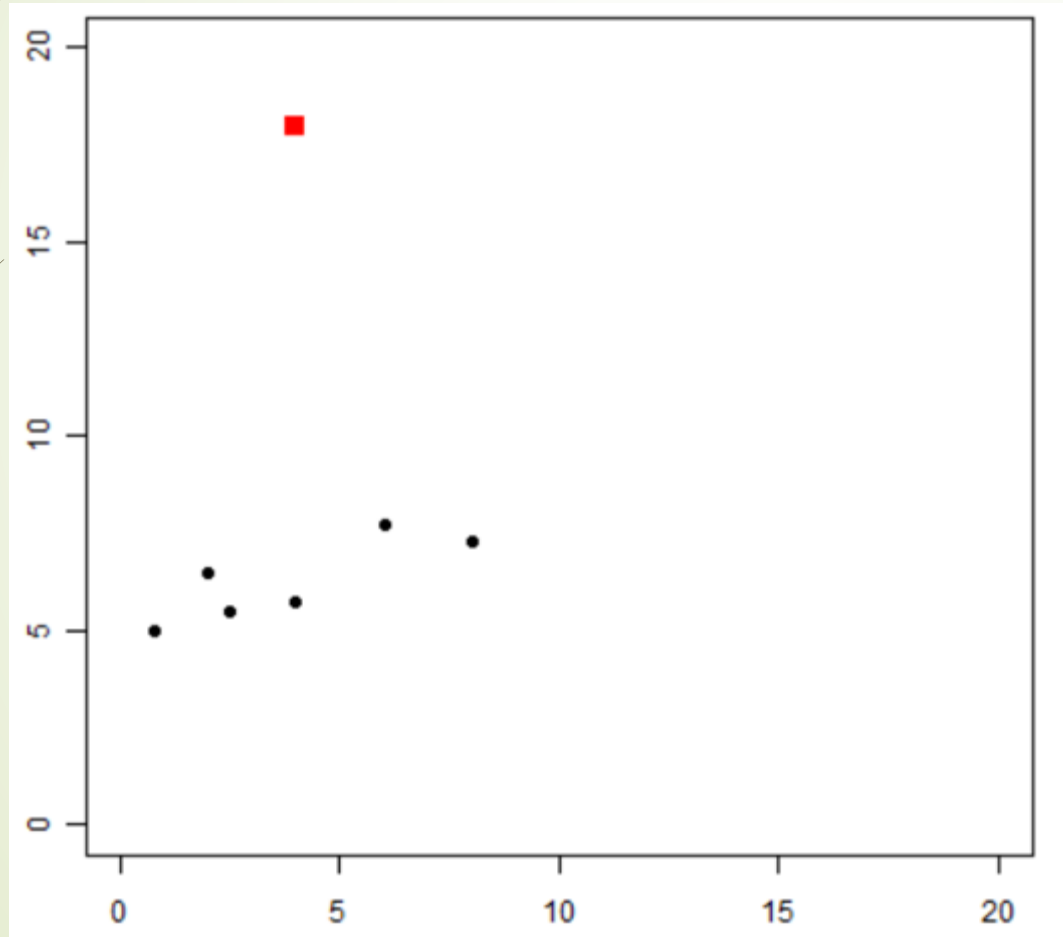
Picture credit: <https://openintro-ims.netlify.app/model-slr.html>

Outliers

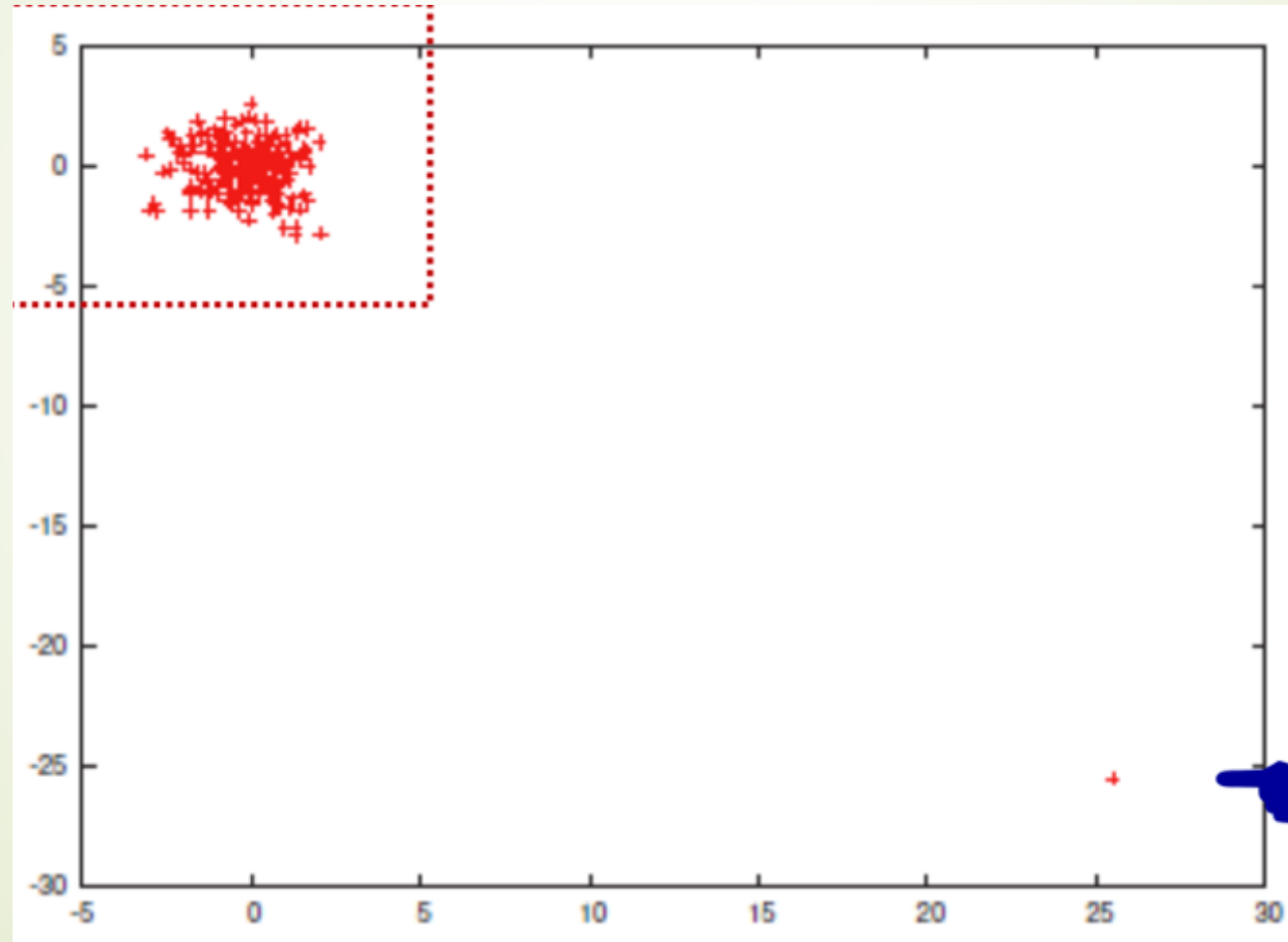
- Outliers are observations that fall far from the main bulk of points.
- How do outliers affect the correlation coefficient?



- Outliers may decrease the strength of the correlation

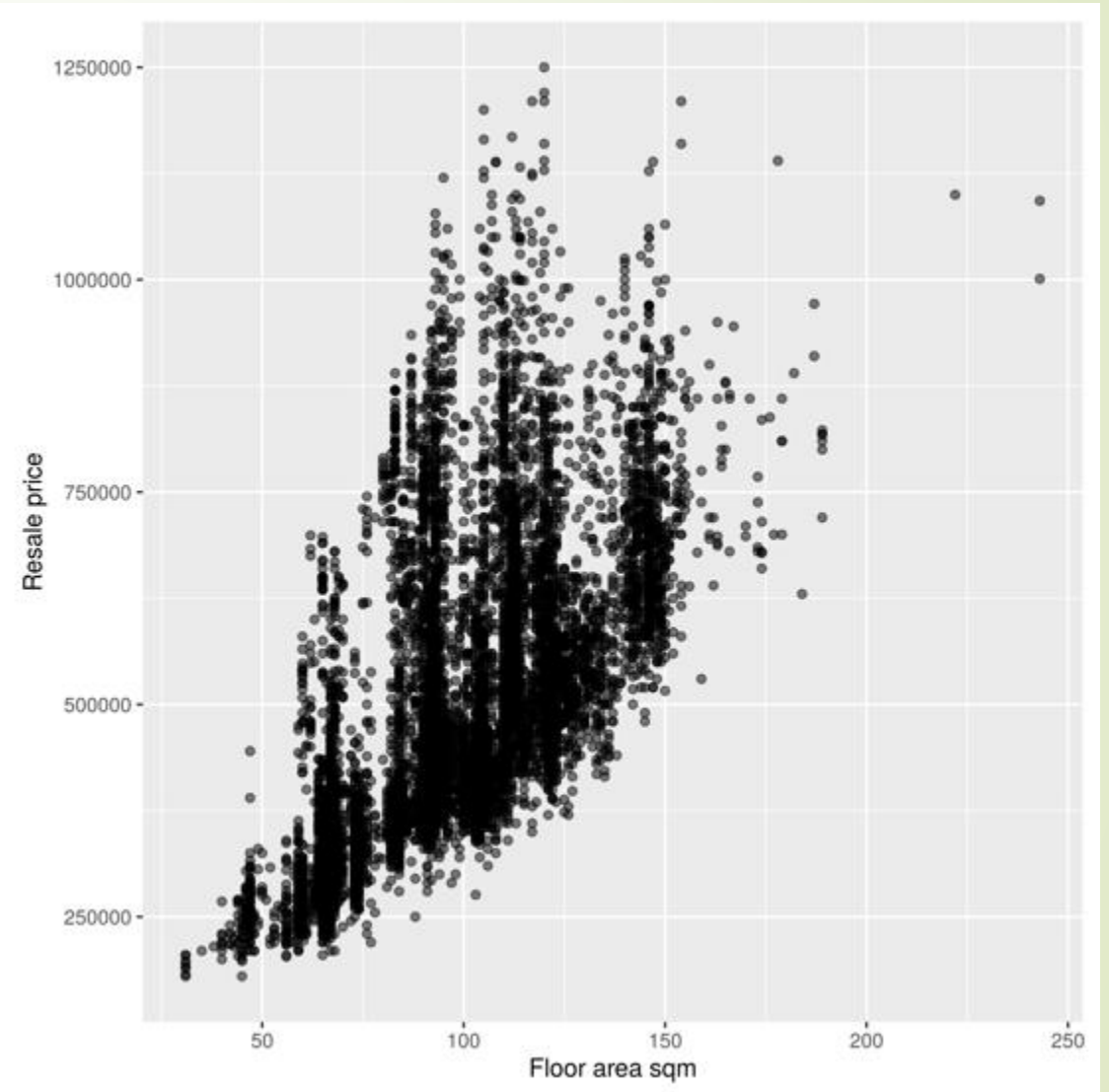


- Outliers may increase the strength of the correlation



Example (HDB data set)

What happens when we remove the outliers in the scatter plot?



Example (HDB data set)

➔ Before removing the outliers

```
Correlation matrix:  
                floor_area_sqm  
resale_price 0.63
```



➔ After removing the outliers

```
Correlation matrix:  
                floor_area_sqm  
resale_price 0.62
```



Summary

► We have learnt

1. Correlation does not imply causation
2. Correlation coefficient does not measure non-linear association
3. How outliers affect correlation coefficient

Linear Regression



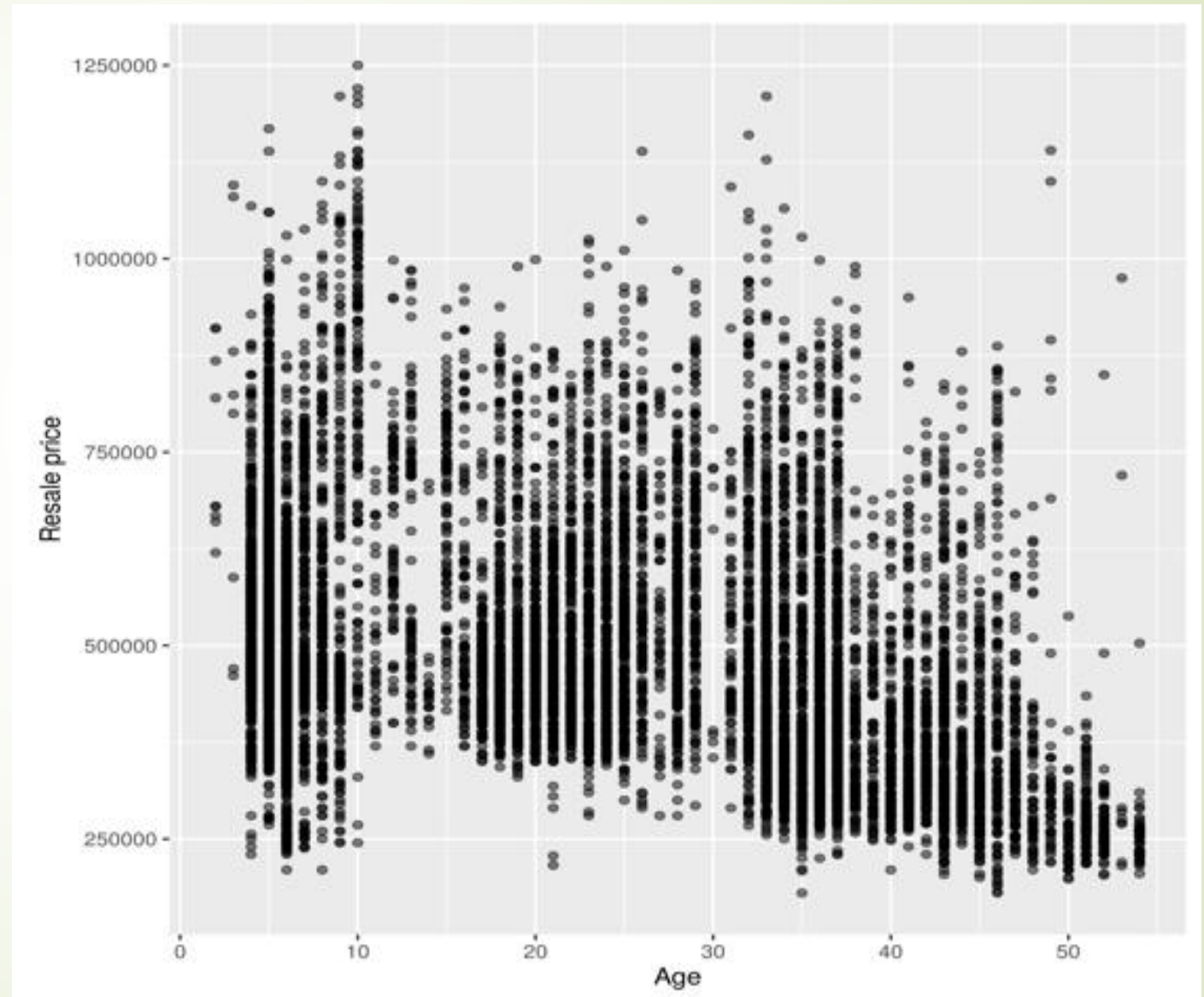


Learning Objectives

- By the end of this unit, you should be able to
 1. Use linear regression to make prediction
 2. Understand the idea behind linear regression
 3. Apply linear regression for non-linear model

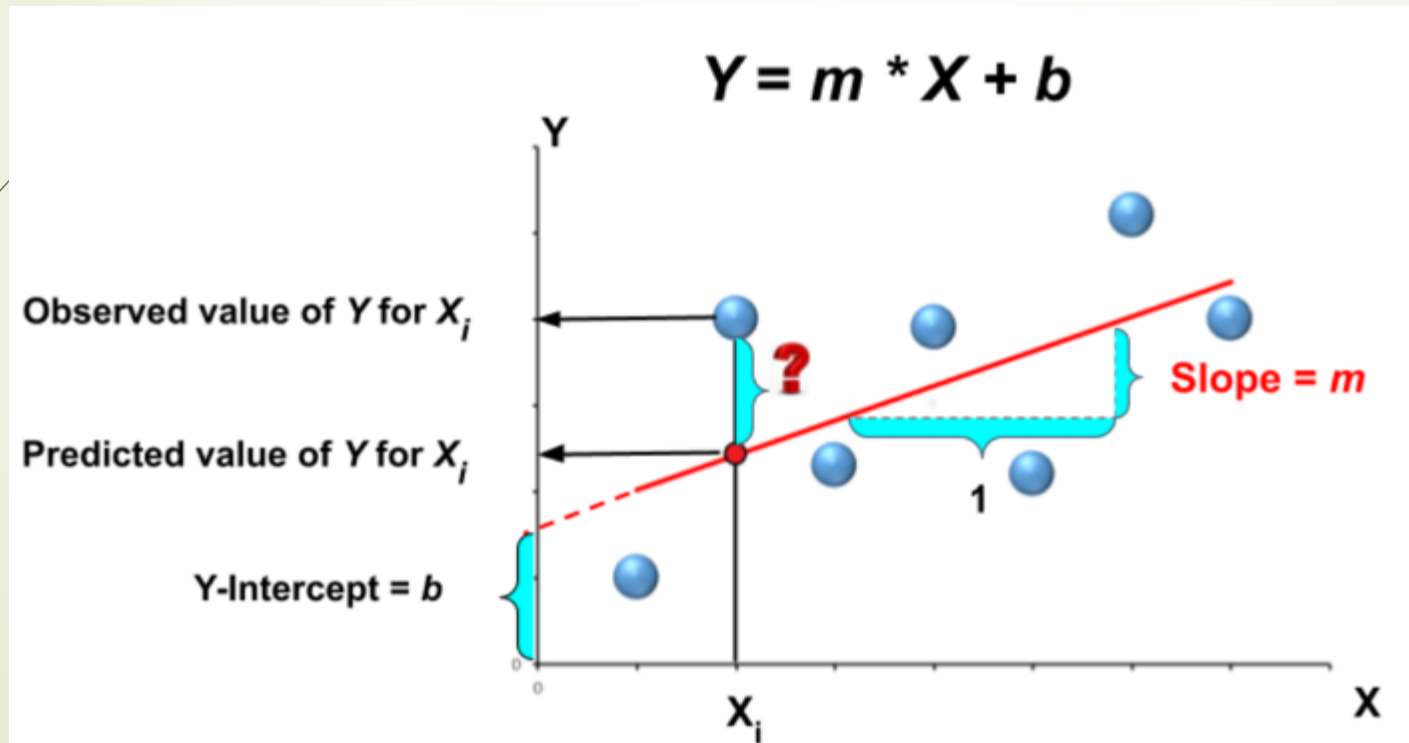
Question

- How to predict the resale price of HDB flat based on its age?



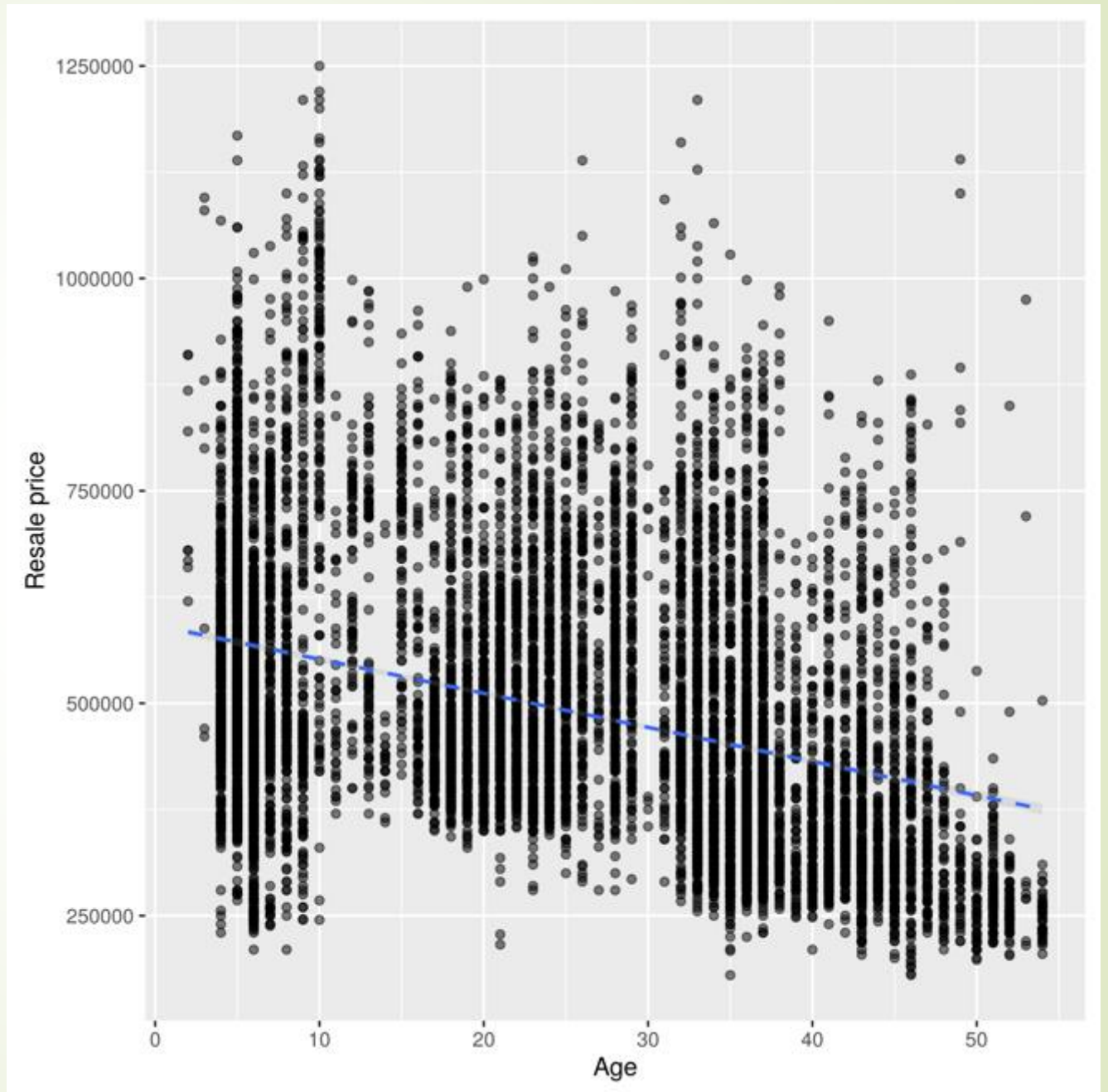
Linear regression

- We model the relationship by a straight line $Y = mX + b$.



Example (HDB data set)

- The regression line for the HDB resale price is $y = -4007x + 591857$.



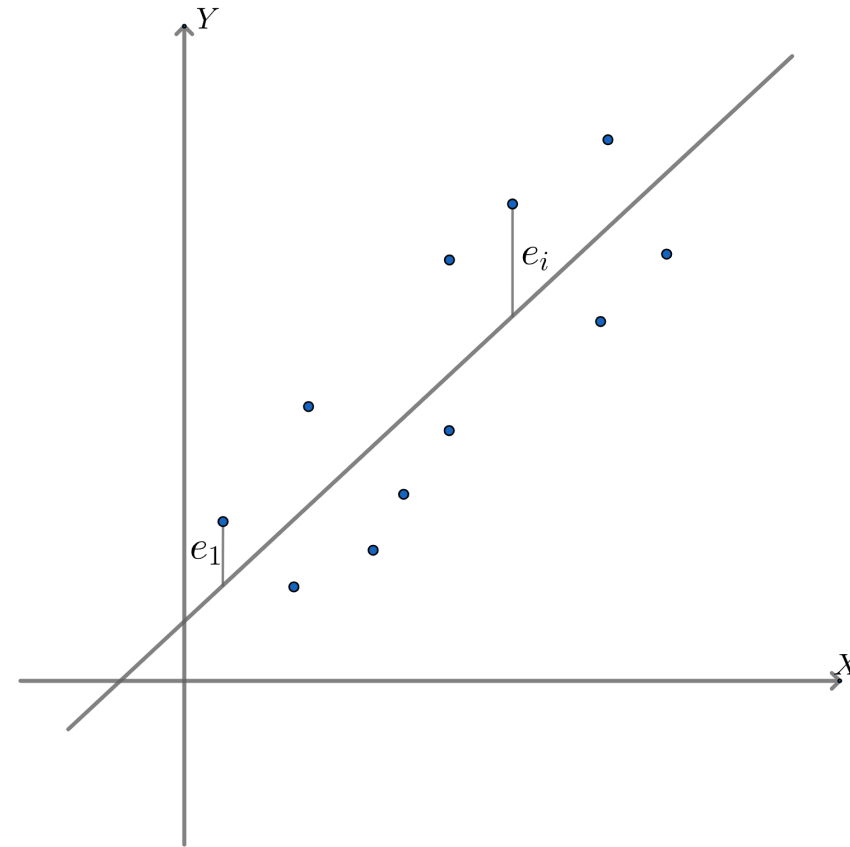


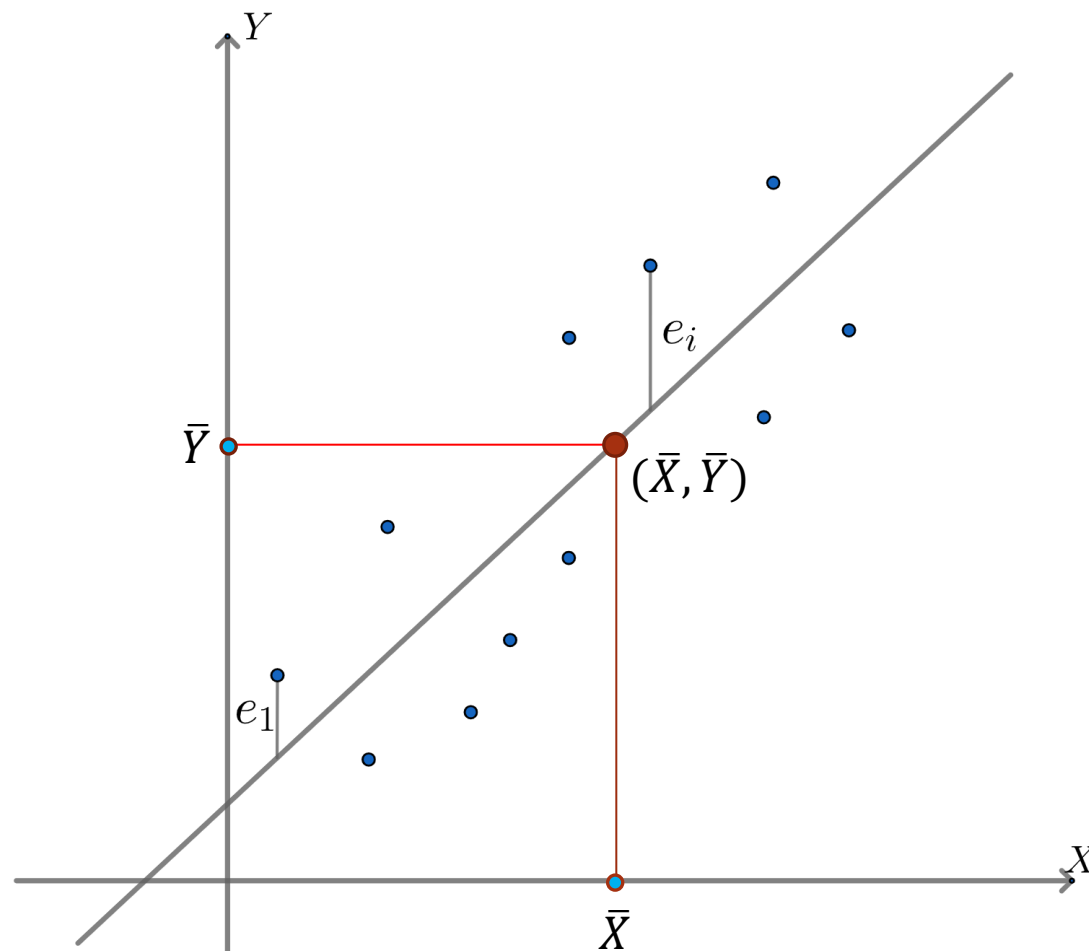
How to find regression line?

- Determine the best fit line for our data
 - We use least square method
- 

How to find regression line?

- Define the i -th residual of the observation: e_i = difference between the observed outcome and predicted outcome.
- Want to minimise $e_1^2 + \dots + e_n^2$ where n =no of data points
- https://gallery.shinyapps.io/simple_regression/





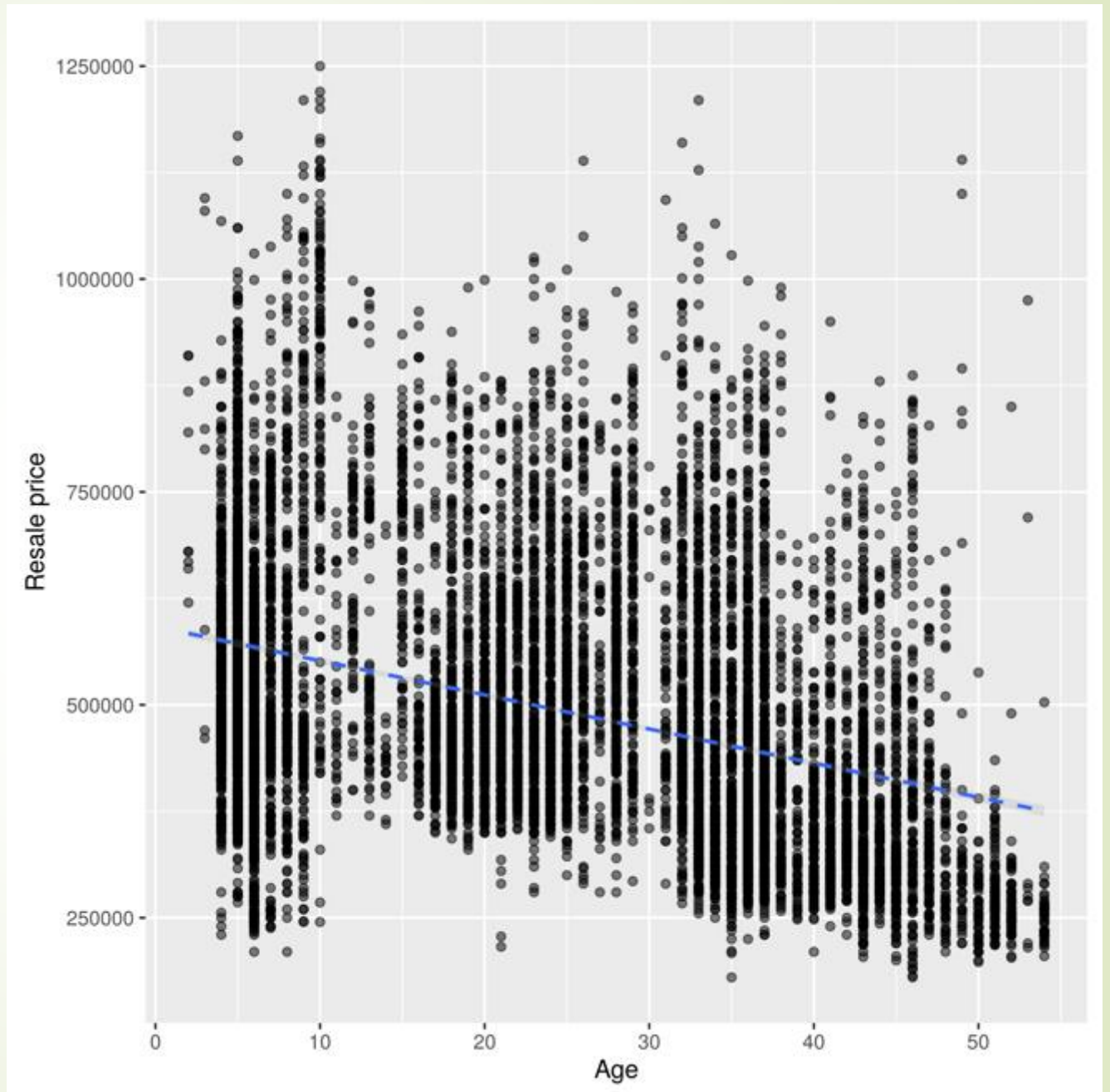
Regression line passes through (\bar{X}, \bar{Y}) .

Slope vs Correlation Coefficient

- ▶ The slope of the regression line and correlation coefficient is related by $m = \frac{s_y}{s_x} r$, where s_y is the standard deviation for y and s_x is the standard deviation for x .

Extrapolation

- Prediction of HDB resale price beyond the observed range of HDB ages is dangerous.



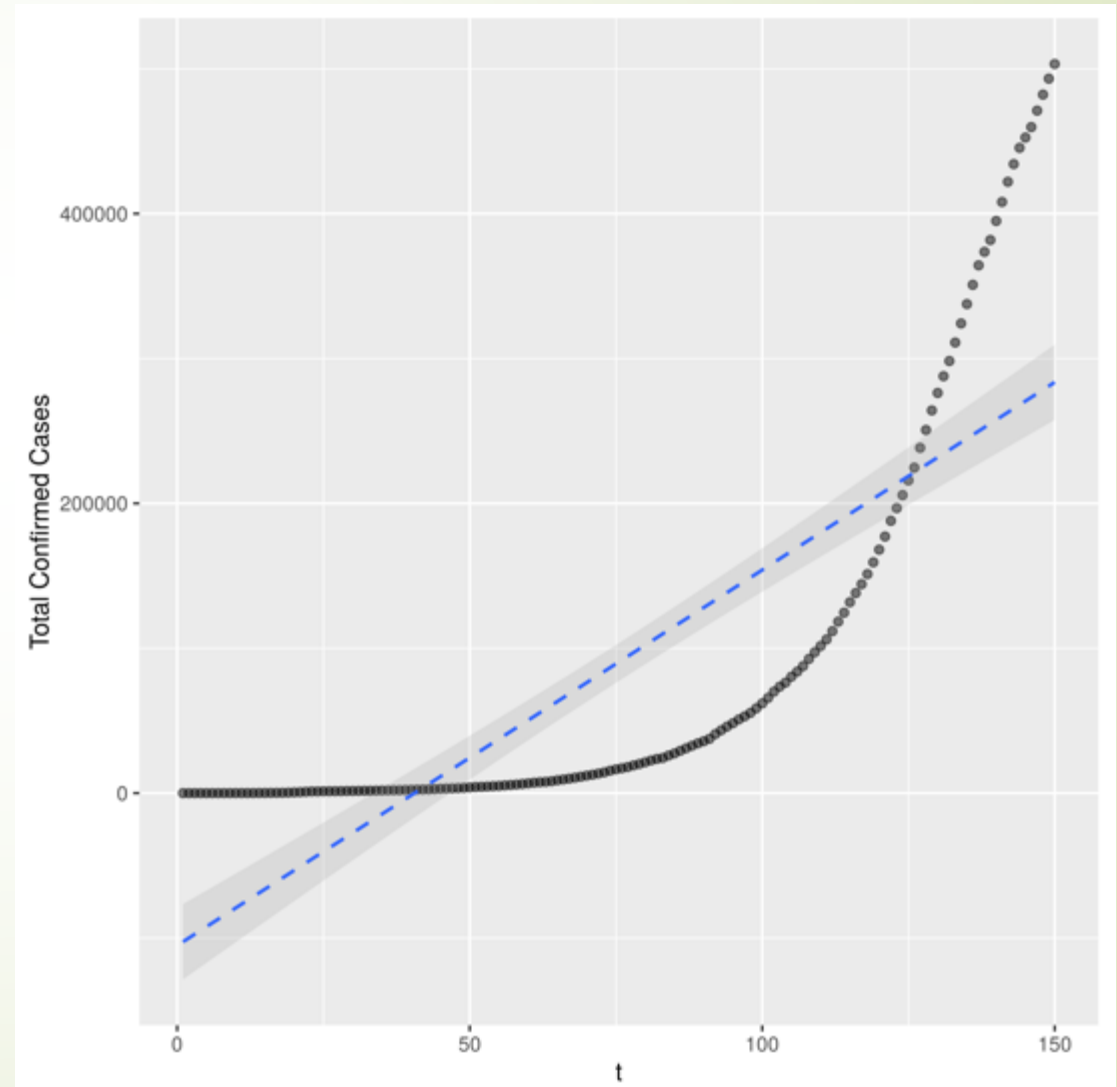
Linear regression on non-linear models

- Total number of confirmed COVID-19 cases in South Africa since 5 March 2020

t	total confirmed cases
76	17200
77	18003
78	19137
79	20125
80	21343
81	22583
82	23615
83	24264
84	25937
85	27403
86	29240
87	30967
88	32683
89	34357
90	35812
91	37525
92	40792
93	43434
94	45973
95	48285
96	50879

Total number of confirmed cases vs t

- Correlation coefficient = 0.81

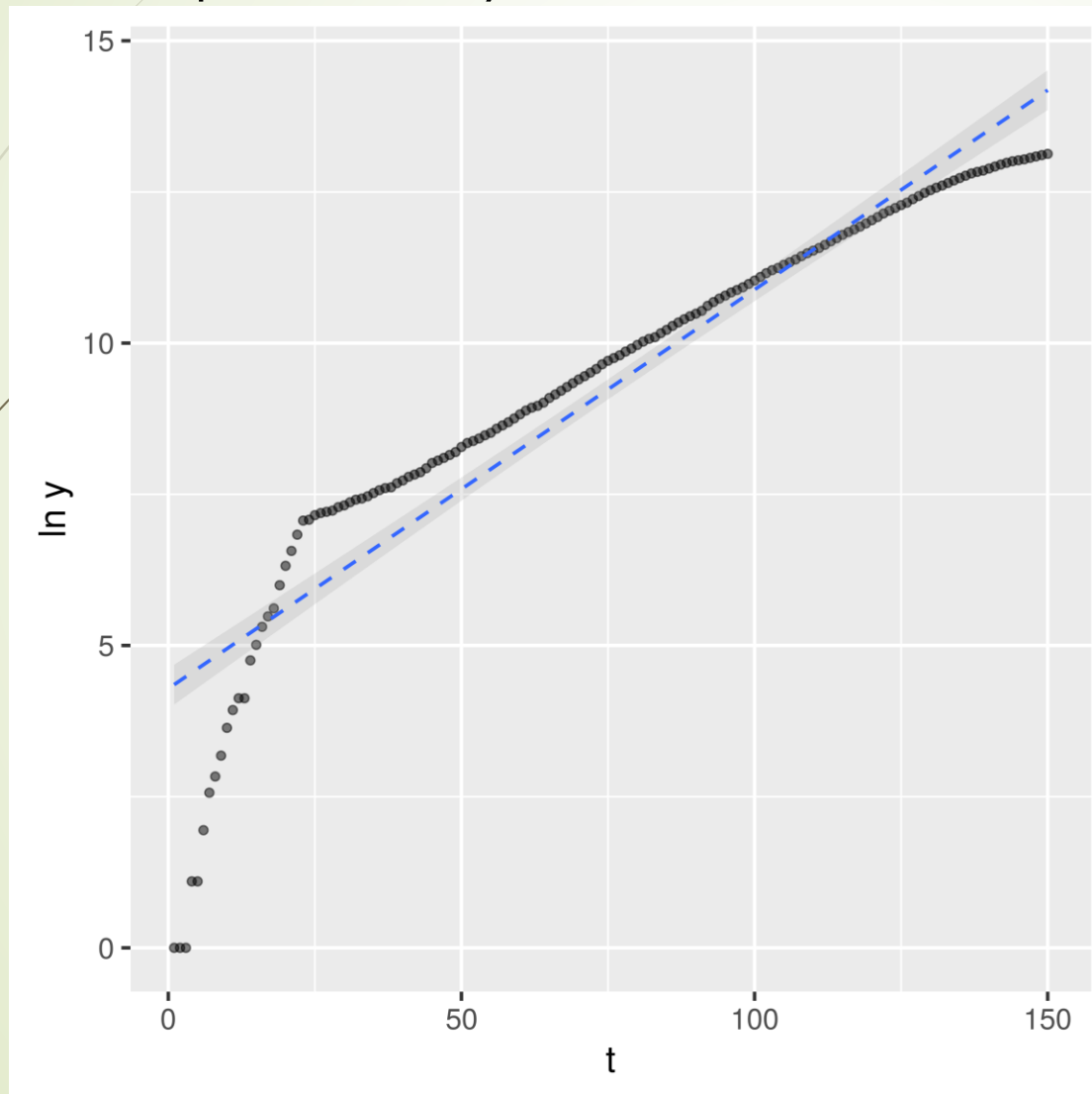




How to model?

- Model the relationship above as $y = cb^t$.
- Want to find c and b using linear regression model.
- First step: plot $\ln y$ vs t .

t	total confirmed cases	ln total
76	17200	9.7526647
77	18003	9.7982937
78	19137	9.8593789
79	20125	9.9097181
80	21343	9.9684791
81	22583	10.024953
82	23615	10.069637
83	24264	10.096749
84	25937	10.163426
85	27403	10.218408
86	29240	10.283293
87	30967	10.340677
88	32683	10.39461
89	34357	10.444561
90	35812	10.486038
91	37525	10.532763
92	40792	10.616241
93	43434	10.678998
94	45973	10.73581
95	48285	10.784876
96	50879	10.837206

- Scatter plot of $\ln y$ vs t



- 
- Apply \ln to $y = cb^t$.
 - $\ln y = \ln c + t \ln b$
 - Compare with $Y = mX + k$
 - We obtain $m = \ln b$ and $k = \ln c$
 - So, $b = e^m$ and $c = e^k$.
- 

Back to the COVID example

- find regression line for $\ln y$ vs t .
- We find $\ln y = 4.287 + 0.066t$.
- Last step: express y in terms of t .
- $y = e^{4.287} e^{0.066t}$.



In summary...

Through our attempt to answer the research question:

What factors may affect the popularity and pricing of resale flats in Singapore?

- We used univariate and bivariate EDA to explore how age, resale price and floor area can affect the number of flats sold
- Described how age, floor area and time period may affect the pricing of resale flats sold.
- Introduced correlation coefficient as a quantitative measure of linear association between two variables
- Presented a linear regression model to do predictions using data