Chapter 3

Dealing with Numerical Data

Section 3.1 Univariate EDA

In Chapter 1, we introduced two main types of variables that we will be focussing on, namely *categorical variables* and *numerical variables*. Categorical variables were discussed extensively in Chapter 2 and in this chapter, we will turn our attention to numerical variables and how they can be analysed.

Consider the following table that shows a portion of a data set relating to COVID-19 cases in Singapore.

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

An example of a numerical variable in this data set is **Age**. Can you identify another numerical variable? The analysis of data, more precisely, Exploratory Data Analysis (or

EDA) is a process of summarising or understanding the data and extracting insights or main characteristics of the data. This is a critical part of the "Analysis" step of the PPDAC problem solving cycle. In this chapter, we will discuss how numerical variables can be summarised and understood. To begin, the focus of this section will be on data exploration techniques for one variable, or *univariate exploratory data analysis*.

Example 3.1.1 In Chapter 2, the recurring data set that was used to drive the discussion on categorical variables was the patients with kidney stones data set. In this chapter, we will be using a data set closer to home.

	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

The data set (Microsoft Excel file partially shown above) that we will be looking at in this chapter corresponds to sales of Housing Development Board (HDB) resale flats within the period of January 2017 to June 2021. The entire data set contains $99,236$ rows and 11 columns. Note that each transaction is a row of the Excel file and each transaction contains information on variables (the columns) like month (of sale), flat's floor area (in square metres), resale price, etc.

The PPDAC cycle starts off with

1. **Problem.** So what is the problem that we are considering and attempting to answer? If you are a potential buyer, perhaps a question that you may be interested in investigating could be

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

2. **Plan.** Here, we need to decide what are some of the variables that are relevant and possible factors that answer the question. Suppose these variables were determined to be the 11 columns of the data set. Some of these variables are

   – "Month" - this is the month/year of the resale transaction;

- "Town" - this is the town that the resale flat belongs to;
- "Floor_area_sqm" - this is the floor size of the resale flat;
- "Resale price" - this is how much the flat was sold for.

3. **Data.** In this stage, data is collected and prepared as shown in the table above.

4. **Analysis.** We are now at this stage where the data is going to be analysed in attempting to answer the **Problem**.

Definition 3.1.2 A *distribution* is an orientation of data points, broken down by their observed number or frequency of occurrence.

Example 3.1.3 Let us look at our HDB resale flats data set. The first few rows of the data set for transactions from January to June 2021, is reproduced in the table below.

Month	Floor area sqm	Age	Resale price
1/1/2021	45	35	225000
1/1/2021	45	35	211000
1/1/2021	73	45	275888
1/1/2021	67	43	316800
1/1/2021	67	43	305000
1/1/2021	68	40	260000
1/1/2021	73	44	351000
1/1/2021	73	44	343000
1/1/2021	75	41	306000

We would like to investigate the distribution of the Age[1] variable. To do this, we would need to collate the number of flats with the same ages when the resale transaction was made and put them in a frequency table. For example the first two rows of the data indicates that the first two HDB flats in the data set had the same age of 35 years when they were sold, while the third flat was 45 years old and so on. Suppose the frequency table collated for the entire data set is as follows:

[1]The data set, which can be downloaded from https://data.gov.sg/dataset/resale-flat-prices actually does not contain the "Age" variable. The "Age" variable was created by subtracting lease_commence_date from the year the flat was sold.

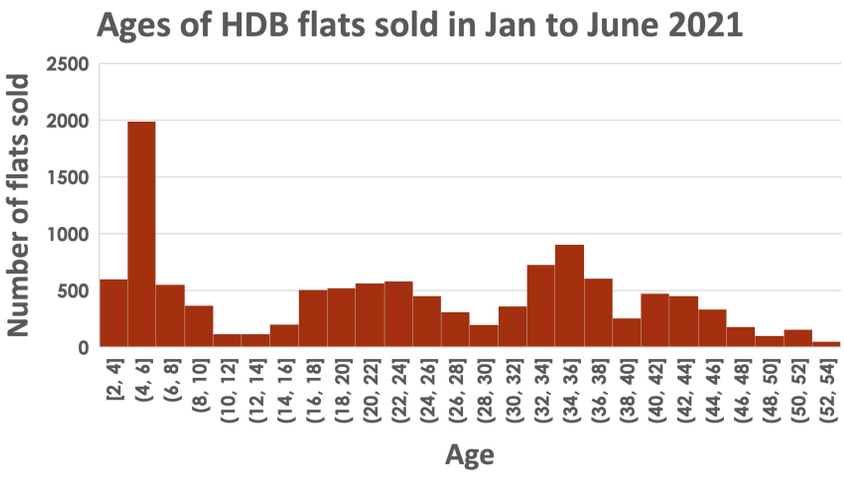
Age	Frequency
2	9
3	8
4	583
5	1105
6	884
7	295
8	255
⋮	⋮

If we simply look at the frequency values in the table, it would be hard to observe any patterns or gain insights into **how** the frequencies are distributed across the different age values. We will introduce two different graphs to present the distribution in better way.

Example 3.1.4 (Histograms for Univariate EDA) A *histogram* is a graphical representation that organises data points into ranges or bins. It is particularly useful when we have large data sets. Let us see how the histogram will look like when we use Microsoft Excel to create one based on the "Age" frequency from Example 3.1.3. To create a histogram, the variable values are "grouped" into equal size intervals called bins. For our "Age" variable, we can use bins with a width of 2 years. The number of flats in each bin are counted and tabulated.

Bins	Frequency
0-2	9
2-4	591 $(8 + 583)$
4-6	1989 $(1105 + 884)$
6-8	550 $(295 + 255)$
8-10	336 $(219 + 47)$
⋮	⋮

You may notice that for the 2-4 Bin, the frequency is obtained by adding the number of flats sold at Age 3 and Age 4 and excludes those sold at Age 2. Thus, the left-end point of the interval 2-4 is excluded. The same is observed for the rest of the bins. The histogram created by Microsoft Excel is shown below:



With the height of each bar representing the frequency for that bin range, the highest bar would represent the most frequently occurring range of values.

From the histogram above, we see that the range 4-6 years has the highest frequency as it accounts for 1989 out of the total 11644 transactions, or about 17% of the flats sold.

Remark 3.1.5

1. Notice that in the histogram above, the 0-2 bin is missing. This is actually a software limitation of the "Insert Statistics Chart → Histogram" function in Microsoft Excel, where the function automatically adds the frequency for Age 2 into the frequency for its first bin, which is the 2-4 bin.

2. You may wonder how we came to the decision to have bin widths of 2 years rather than 3 years (or any bigger number). There is no correct answer for this. Normally, we would construct several histograms with different bin widths before deciding which one is most appropriate.

Once we have obtained and visualised the distribution of a numerical variable, we would like to describe the overall pattern of the distribution as well as whether there are any deviations from the overall pattern. To describe the overall pattern of the distribution, we will focus on the

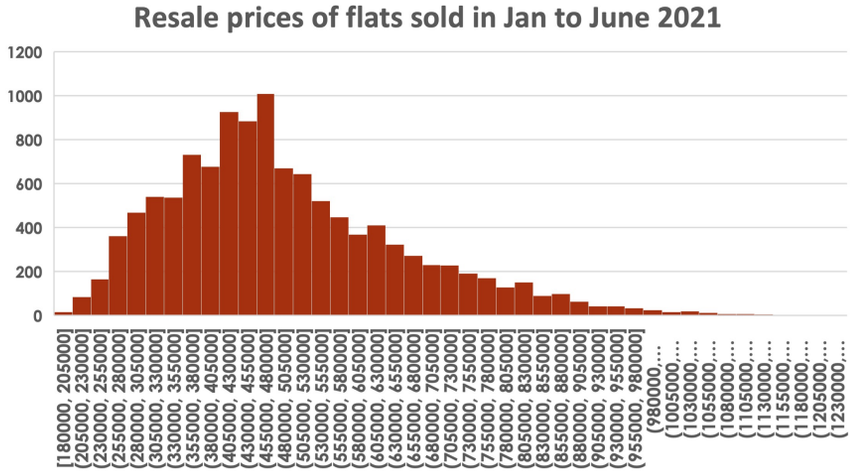
1. Shape;

2. Center; and

3. Spread of the distribution.

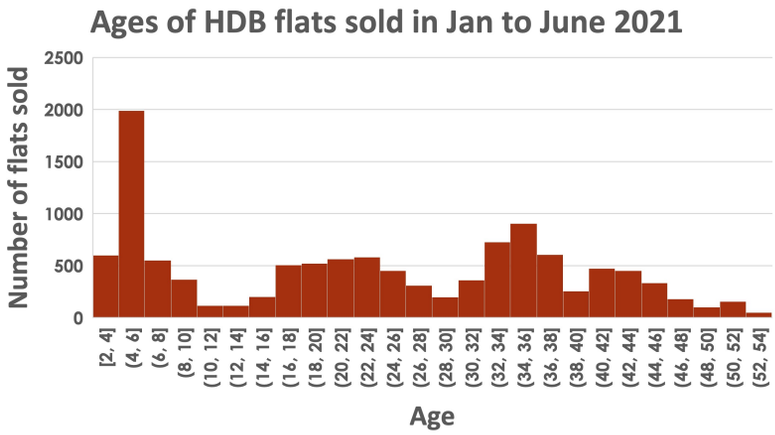
For deviations from the overall pattern, this usually refers to identifying *outliers* which will be discussed later on in this Chapter. Let us start by looking at how we can describe the shape of a distribution.

Discussion 3.1.6 (Shape - peaks and skewness). There are two important descriptors when we discuss the shape of a distribution, namely the *peaks* and the *skewness*. Let us look at another histogram plot obtained from the HDB resale data set. Rather than the age of the flat at the point of resale, we consider another numerical variable of interest, which is the "Resale Price". The following histogram was obtained when we set a bin size of 25,000.



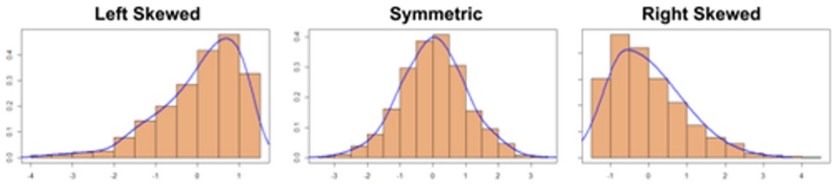
As evident from the plot, we see a peak in the interval [455000, 480000]. The distribution is *unimodal*, which means that it has one distinct peak. This tells us that the most frequent resale flat prices lies between \$455,000 and \$480,000.

Distributions are not always unimodal. Looking at the histogram we plotted earlier for the Age of the resale flats, we see that there is more than one distinct peak. In such a situation, we say that the distribution is *multimodal*. If a distribution has exactly two distinct peaks, we say it is *bimodal*.



In the histogram above, we see the highest peak in the 4-6 years range and the second highest peak occurring in the 34-36 years range. It should be noted that we say these are peaks because they occur most frequently in their immediate neighbourhoods of age ranges.

For a unimodal distribution, we can use another descriptor to describe the shape of the distribution, that is, whether the distribution is *symmetrical* or *skewed*.

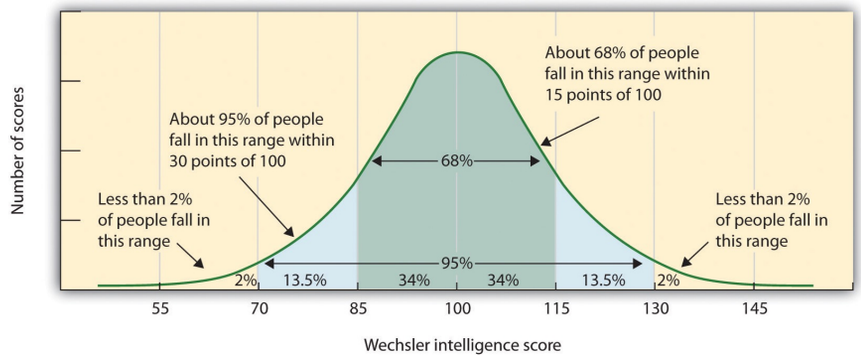


In a *symmetrical* distribution (middle picture above), the left and right halves of the distribution are approximate mirror images of each other, with the peak in the middle.

For the picture on the left, the distribution is **left skewed**, with the peak shifted to the right and a relatively long "tail" on the **left**.

The picture on the right shows a distribution that is **right skewed**. Such a distribution has the peak shifted to the left and a relatively long "tail" on the **right**. Referring back to the distribution of resale prices of HDB flats, we see that the distribution is right skewed, meaning that there are some (but few) flats sold at very high prices. These data points gave rise to the long tail to the right of the peak.

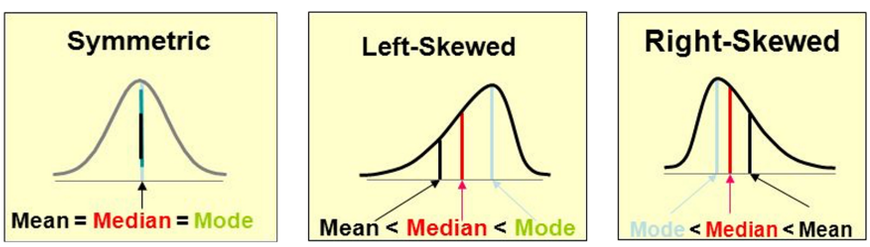
Example 3.1.7 (Symmetrical distribution - Bell curve) One of the most well-known symmetrical distributions is the *normal distribution* or what is commonly known as the bell curve. A famous example of the normal distribution is that of the IQ scores in a population, based on the Wechsler Intelligence scale.



From the figure, we see that the peak happens at 100, which means that the average IQ of a person in the population is 100. We also see that about 68% of the population has IQ scores in the range between 85 and 115, whereas about 95% of the population has IQ scores between 70 and 130.

Discussion 3.1.8 (Central tendency - mean, median and mode). Besides describing the shape of the distribution, we can also describe the characteristics of a distribution more precisely using measures of central tendency. The three most common measures of central tendency are *mean*, *median* and *mode*, which were all introduced in Chapter 1.

The three possible shapes of a distribution have different relative positions of the mean, median and mode.



1. For a symmetrical distribution, the mean, median and mode will be very close to each other near the peak of the distribution.

2. For a left skewed distribution, we usually (but not always) have

$$\text{mean} \ < \ \text{median} \ < \ \text{mode} \ .$$

   To see why this is the case, notice that the small number of extremely small values which contributes to the long tail on the left, will push down the mean/average, as compared to the median which is less affected by these extremely small values. The mode, found at the peak of the distribution is naturally the largest among the three measures of central tendency.

3. For a right skewed distribution, we have the opposite of the left skewed distribution, which is

$$\text{mode} < \text{median} < \text{mean} .$$

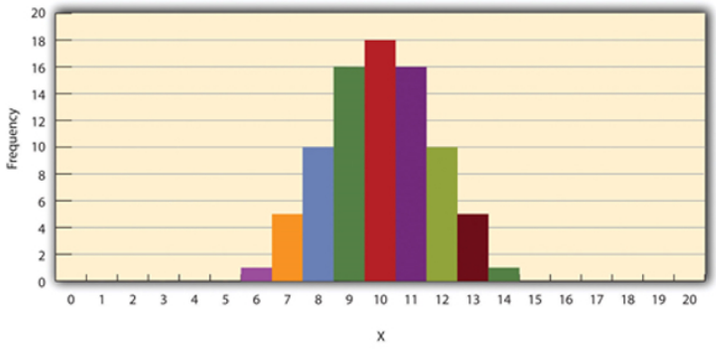
In this case, there are a small number of extremely big values which contributes to the long tail on the right. These big values will push up the mean/average as opposed to the median which is less affected by these extremely large values. The mode in such a distribution would be the smallest among the three measures of central tendency.

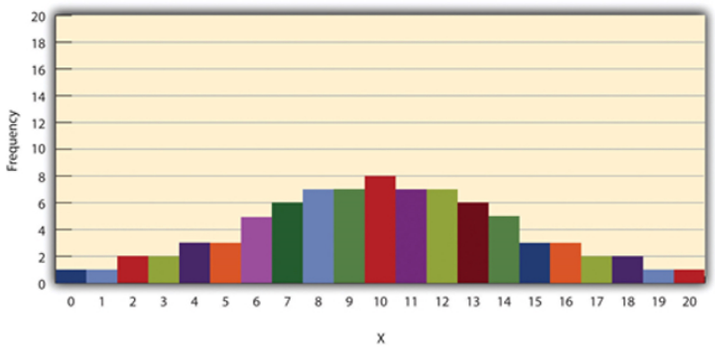
Example 3.1.9 Referring again to the resale prices distribution, we have seen the shape of the distribution and concluded that the distribution is right skewed.

The mean, median and mode of this distribution were found to be \$496,870.40, \$468,000 and \$420,000 respectively. This indeed agrees with

$$\text{mode} < \text{median} < \text{mean} .$$

Discussion 3.1.10 (Spread - standard deviation and range). Besides the shape and center of the distribution, we can also describe the *spread* of a distribution. This refers to how the data vary around the central tendency.





Take a look at the two distributions above, both of which have the same central tendencies. In fact, the mean, median and mode of both distributions are 10. However, the top distribution has a relatively lower variability compared to the distribution below. This means that the data in the top distribution are all relatively close to the center while the data in the bottom distribution are more spread out, or has more variability. We can also say that the data in the bottom distribution is spread across a much wider range.

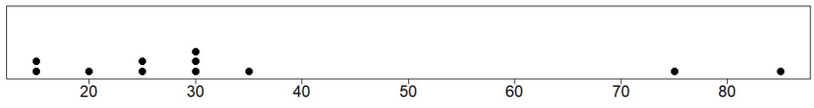
The most commonly used measure of variability is *standard deviation* which was introduced in Section 1.5. For the two distributions shown here, the top distribution has standard deviation 1.69 while the bottom distribution has standard deviation 4.30.

A simpler measure of variability is the *range* of the distribution. This is defined to be the difference between the largest and the smallest data points in the distribution. The range is simple to compute but sometimes it can be misleading. For example, if we look at the range of the HDB resale prices data, we obtain

$$\text{Range} = \text{ Highest resale price } - \text{ Lowest resale price } = \$1,250,000 - \$180,000 = \$1,070,000.$$

The range is very large and is due to the existence of a few extremely high resale prices. It is not really the case that there is great variability in resale prices as we see that most of the resale prices are actually much lower and the variability is not as big as the range indicates it to be.

Definition 3.1.11 An *outlier* is an observation that falls well above or below the overall bulk of the data.



Consider the data set with 11 data points shown above. We can consider 75 and 85 as outliers since they are way larger than the rest of the data points. At this point, we use our judgement to identify values that appear to be exceptions to the general trend in the data. Later on, we will be introducing a more precise method (boxplot) to identify outliers.

Identifying outliers can be useful when we wish to identify any strong skewness in a distribution. Sometimes the outliers are caused by erroneous data collection or data entry but this may not always be the case. It is also possible that outliers are legitimate data points that provide us interesting insights into the behaviour of the data. A general rule when we investigate a data set is that outliers should not be removed unnecessarily as they do tell us something about the behaviour of the variable and prompt us to investigate further why such extreme values can happen.

Example 3.1.12 Consider the data set below:

$$4, 4, 5, 5, 5, 5, 6, 6, 6, 7, 7, 300.$$

It is not difficult to be convinced that 300 is an outlier in the data set. The table below shows the three different central tendencies as well as the standard deviation for the entire set and also when the outlier is removed from the data set.

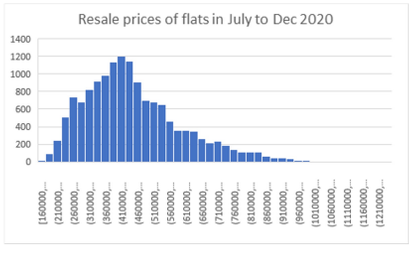
	Mean	Median	Mode	Standard deviation
Without removing 300	30	5.5	5	85.03
With 300 removed	5.45	5	5	1.04

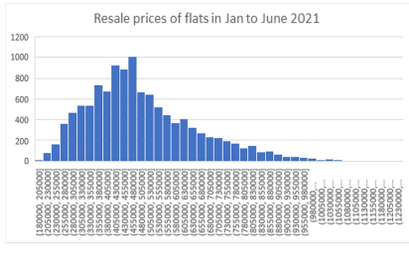
We see that between the three central tendencies, the mean seems to be the most affected by the removal of the outlier, while both the median and the mode either remained the same or only changed slightly. Without removing the outlier, the mean is pulled away in the direction of the skew (in this example, the distribution is skewed to the right). In such cases, mean may no longer be a good measure of the central tendency of the distribution. We call the median and the mode *robust statistics*.

In addition, the standard deviation also increases greatly from 1.04 to 85.03 because of the outlier. This is expected because the standard deviation measures the spread of the data points and with the outlier being far away from the other data points, the variability of the distribution is understandably high.

As mentioned above, we need to treat outliers with care. If they have minimal effect on the conclusions and if we cannot figure out why they are there, such outliers may possibly be removed. However, if they substantially affect the results, then we should not drop them without justification.

Example 3.1.13 Suppose we are interested to find out if there are significant differences in the distribution of HDB resale prices for different time periods. For example, would the distributions differ significantly if we compare the period July to December 2020 with January to June 2021? The two distributions are shown below.





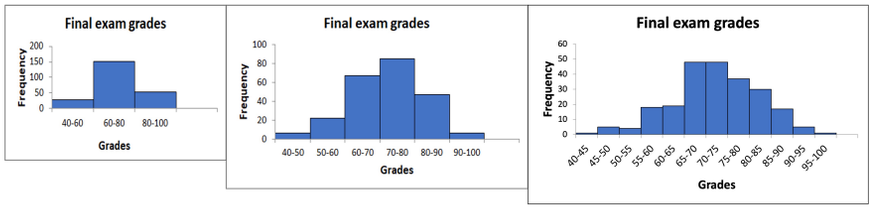
The distribution on the left corresponds to the period of resale from July to December 2020. The distribution for January to June 2021 is shown on the right. We observe that

both distributions have a similar shape which is right skewed with a single peak. Taking it one step further, we compare the central tendencies and variabilities of the data points in both periods. The values in the table can be computed using the Microsoft Excel Data Analysis Toolpak.

	Mean	Median	Mode	Range	Standard deviation
July to December 2020	\$462,827	\$435,000	\$400,000	\$1,098,000	\$155,955
January to June 2021	\$496,870	\$468,000	\$420,000	\$1,070,000	\$162,107

Observe that all measures of mean, median and mode are higher in the time period January to June 2021 compared to those in the time period July to December 2020. The range of the resale prices is lower in January to June 2021 while the standard deviation is actually higher. In conclusion, we can say that resale prices in January to June 2021 are higher, but more spread out (in terms of standard deviation) compared to the resale prices in July to December 2020.

Example 3.1.14 In Example 3.1.4, we described the setting of bin widths when creating a histogram. Deciding the bin width to use can have a big impact on how the histogram looks like and thus affect our observation and conclusion on the shape of the distribution.



The three histograms above are constructed using the same data set of 233 students' final exam scores with the only difference being the bin width settings. The histogram on the left has a bin width of 20, while the one in the middle has bin width of 10. The last histogram has bin width set at 5. What conclusions can be made on the distribution based on these histograms?

Based on the first histogram, we may make the conclusion that most students score between 60 to 80 marks, and the distribution is rather symmetric. However, with a slightly smaller bin width, the second histogram reveals that most students actually scored between 70 to 80 marks. This does not contradict the observation made earlier based on the first histogram but because of the smaller bin width, we are able to narrow the range of marks that are scored by most students. With an even smaller bin width, the third histogram suggests that most students scored between 65 and 75 marks. How do you rationalise this conclusion with the one from the second histogram?

In general, we should bear in mind the following when determining bin widths for histograms.

1. Avoid histograms with bin widths that are too large. This will result in only a few bins and information in the data will be lost when data points are grouped together into a small number of groups/bins.

2. Avoid histograms with bin widths that are too small. If we do this, there may be bins that have very few data points (or none) that does not give us a sense of the distribution.

3. Our initial choice of bin width may not be the most appropriate. Different histograms with various bin widths should be created before deciding which one is the most useful and informative.

Remark 3.1.15 We should not confuse histograms with bar graphs introduced in Chapter 2. A histogram shows the distribution of a numerical variable across a number line. So one of the axis (usually the horizontal) will display the range of values taken on by the numerical variable. On the other hand, the horizontal axis of a bar graph will show the different categories of a categorical variable.

In addition, the ordering of the bars in a histogram cannot be changed, as it progresses through the range of values, usually in an ordered manner, taken on by the numerical variable. On the other hand, the ordering of the bars in a bar graph can be switched around with little consequence. There are also usually no gaps between the bars in a histogram.

Discussion 3.1.16 (Boxplots for Univariate EDA) Besides a histogram, another way to visualise the distribution of a numerical variable is to use a *boxplot*. To construct a boxplot, we will use the five-number summary, consisting of

1. Minimum;

2. Quartile 1 (Q_1);

3. Median (Q_2);

4. Quartile 3 (Q_3);

5. Maximum.

The median and quartiles have already been introduced in Definition 1.6.1 and Definition 1.6.5. Furthermore, we have also introduced the Interquartile range

$$\text{IQR} = Q_3 - Q_1.$$

While the median can be viewed as the center of a data set, the IQR is a way to quantify the spread of a data set. We have defined an outlier in Definition 3.1.11 but did not provide an explicit way to classify a data point as an outlier. For our purpose we will adopt the following consideration to classify a data point as an outlier.

A data point is considered an *outlier* if it satisfies one of the following conditions:

- The value of the data point is **greater than** $Q_3 + 1.5 \times \text{IQR}$;
- The value of the data point is **less than** $Q_1 - 1.5 \times \text{IQR}$.

To construct a boxplot, we do the following:

1. Draw a box from Q_1 to Q_3.
2. Draw a vertical line in the box where the median (Q_2) is located.
3. Identify all the outliers by using the consideration above.
4. Extend a line from Q_1 to the smallest value that is not an outlier and another line from Q_3 to the largest value that is not an outlier. These lines are called *whiskers*.
5. Mark each of the outliers with dots or asterisks.

Example 3.1.17 Consider the following data set, with the data points already sorted in increasing order.

$$18,\ 44,\ 47,\ 55,\ 61,\ 62,\ 78,\ 79,\ 83,\ 145.$$

There are 10 data points. The median (Q_2) is the average of the fifth and sixth data points, so

$$Q_2 = \frac{1}{2}(61 + 62) = 61.5.$$

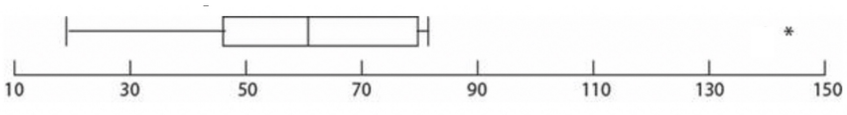
The first quartile is the median of the first five data points: $18, 44, 47, 55, 61$, so $Q_1 = 47$. The third quartile is the median of the last five data points: $62, 78, 79, 83, 145$, so $Q_3 = 79$. Following Remark 1.6.9, it should be pointed out that you may encounter slightly different ways of finding quartiles for a data set in other texts. For this course, we will adopt what is presented here.

The Interquartile Range is

$$\text{IQR} = Q_3 - Q_1 = 79 - 47 = 32.$$

To determine if we have outliers, note that $1.5 \times \text{IQR} = 48$. Since there are no data points smaller than $Q_1 - 48$, there are no small-valued outliers. On the other end, since $145 > Q_3 + 48 = 79 + 48 = 127$, we see that 145 is the only big-valued outlier.

The boxplot constructed is shown below.



Example 3.1.18 Let us return to the HDB resale flats data set. The boxplot below is based on the resale prices of flats sold in January to June 2021.

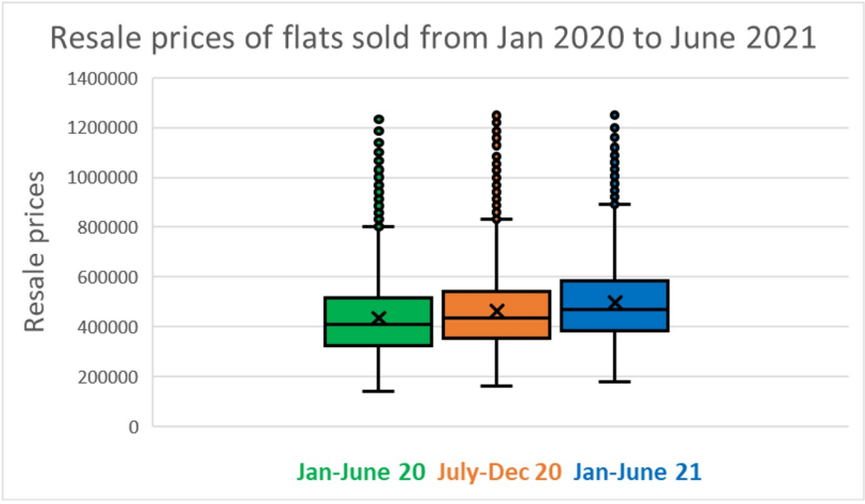
The boxplot confirms our earlier conclusion that there are outliers that correspond to very high resale prices. Note that the cross in the box, just above the median line represents the mean resale price. Recall that we have discussed the shape, center and spread of the distribution using a histogram. What can we say based on the boxplot?

1. **(Shape)** From the boxplot, we see that the variability in the upper half of the data, given by (Max − Median) is significantly larger than the variability in the lower half of the data which is equal to (Median − Min). This confirms our earlier observation that the distribution is skewed to the right and there is a relatively long tail to the upper end of the distribution due to the existence of outliers.

2. **(Center)** The center, described by the median is easily observed in the boxplot, unlike in a histogram. We can also compare the relative positions of the median and the mean from the boxplot.

3. **(Spread)** The IQR of $204,000$ gives us an idea of the spread for the middle 50% of the data set. On its own it may not be immediately informative but this would be a meaningful measure to compare across different distributions (see next example).

Example 3.1.19 The three boxplots below show the distributions of resale flat prices in three different time periods, namely January to June 2020 (call this period P1), July to December 2020 (call this period P2) and January to June 2021 (call this period P3). What can we say about the three distributions after comparing the three boxplots?



1. All three distributions are right skewed as the upper halves of the data have greater variability than the lower halves, due to (large-valued) outliers. However, upon a closer look, it is also apparent that the upper half variability in period P1 is greater than the upper half variability in P2 which in turn is greater than the upper half variability in P3.

2. The middle 50% (that is, the IQR) box of resale prices is lowest in P1, followed by P2 and then P3. Hence, the overall resale prices have increased over time. The spread (given by the height of the boxes) appears to be similar between P1 and P2 while slightly higher in P3.

3. There appears to be more outliers in P1 and P2 compared to P3.

To conclude this section, we summarise the comparison between using histograms and boxplots to represent a distribution.

1. A histogram typically gives a better sense of the shape of the distribution of a variable, compared to a boxplot. When there are great differences among the frequencies of the data points, a histogram will be able to illustrate this difference better than a boxplot.

2. If we wish to compare the distributions of different data sets, putting the different boxplots side by side is more illustrative than using histograms.

3. To identify and indicate outliers, boxplots do a better job than histograms.

4. The number of data points we have in a data set is better shown in a histogram than in a boxplot. In fact, two distributions with very different number of data points can have almost identical boxplots. On the other hand, this difference is apparent by comparing the histograms.

The bottom line is that different graphics and summary statistics have their advantages and disadvantages and they are often used together to complement each other.

Section 3.2 Bivariate EDA

In this section, we will focus on how we can investigate a relationship between two variables in a population.

Discussion 3.2.1 We start off with a relationship between two variables that is *deterministic*. This means that the value of one variable can be determined exactly if we know the value of the other variable. Perhaps the most common type of deterministic relationship is the one that involves the conversion of units of measurement from one metric to another. For example:

1. The relationship between Fahrenheit (F) and Degree Celsius (C) in the measurement of temperature. We know that F and C are related by
$$C = (F - 32) \times \frac{5}{9}.$$
This is a deterministic relationship between F and C. For example, if the temperature in the oven now is 450 degrees Fahrenheit (so $F = 450$), then the temperature in the oven now, measured in Degree Celsius is
$$C = (450 - 32) \times \frac{5}{9} = 232.22.$$

2. Meters (M) and Feet (F) are both measurements of length (or height) and they are related (approximately) by
$$F = 3.2808 \times M.$$
So, if Johnny's height is 5.9 Feet (so $F = 5.9$), then his height in meters will be
$$M = \frac{F}{3.2808} = \frac{5.9}{3.2808} \approx 1.8 \text{ meters}.$$

Discussion 3.2.2 The main focus of this section is on a relationship between two variables that is not deterministic in nature. We say such a relationship is *statistical* or non-deterministic. Recall that in a deterministic relationship, given the value of one variable, we can find a unique value of another variable. However, this is not possible for a statistical relationship, where given the value of one variable, we can describe the average value of the other variable. Such relationships between variables, called *associations* occur quite often in our daily life.

Example 3.2.3 In a Medical News Today article published in November 2020, it was reported that in a study involving more than 150,000 participants, a clear link was observed between low physical fitness and the risk of experiencing symptoms of depression, anxiety, or both.

This association between physical fitness and mental health may not be surprising but we wonder if it could be due to other factors, like a confounder. More interestingly, does having better fitness make a person mentally healthier or having better mental health make a person exercise more resulting in better physical fitness? We will not only measure the association (if one exists) between variables but also attempt to interpret any observed associations.

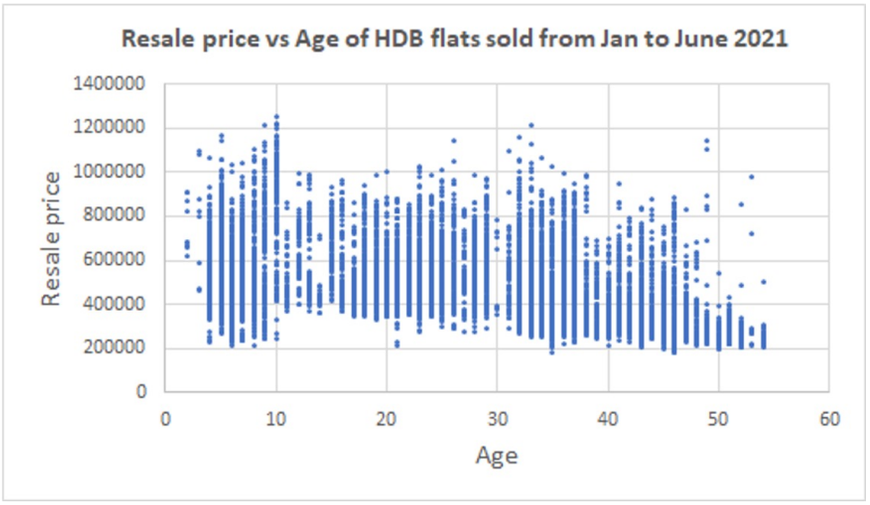
Bivariate data is data involving two variables. For example, in the HDB resale flat data set, we can study the two variables Age and Resale Price .

Month	Floor area sqm	Age	Resale price
1/1/2021	45	35	225000
1/1/2021	45	35	211000
1/1/2021	73	45	275888
1/1/2021	67	43	316800
1/1/2021	67	43	305000
1/1/2021	68	40	260000
1/1/2021	73	44	351000
1/1/2021	73	44	343000
1/1/2021	75	41	306000

In Section 3.1, we saw two ways to display univariate data, using either a histogram or a boxplot. For bivariate data, it is clear that using a table like the one above is not really useful if we wish to investigate if the two variables are associated. Instead, we will use a *scatter plot* to give us an idea of the pattern formed by the data between the two variables in question. After looking at the scatter plot, we use a quantitative measure called the *correlation coefficient* to quantify the level of linear association (if any) between the two variables. Finally, we will attempt to fit a line or a curve through the points in the data set which will enable us to make predictions on the values of the variables. This process is

known as *regression analysis*. For now, we will focus on scatter plots and defer the discussion on correlation coefficients and regression analysis to the next few sections.

Example 3.2.4 Returning to our HDB resale flats prices data set, we will focus on the bivariate data with the variables Age and Resale price. Suppose we wish to know if the age of the flat affects the resale price, with the ultimate intention to make a prediction, based on the past resale prices, of how much a 38 year old resale flat is likely going to cost. In this case, we can treat age as the independent (or explanatory) variable and resale price as the dependent (or response) variable.



Our scatter plot shown above has the age (independent) variable on the x-axis and the resale price (dependent) variable on the y-axis. Each resale transaction would be represented by an *ordered pair*

$$(x, y)$$

where x is the age of the resale flat and y is the resale price of that flat. For example, the ordered pair $(35, 225000)$ corresponds to the first resale flat listed in the table above. With a point plotted for each ordered pair, since there are $11,644$ resale transactions in the data set, there will be $11,644$ points on the scatter plot. Observe that in the scatter plot, each value of x (age of flat) corresponds to many different values of y (the resale price). This is to be expected because there are many different transactions involving flats of the same age and all these transactions are made at different resale prices.

How do we describe the relationship between two numerical variables using a scatter plot?

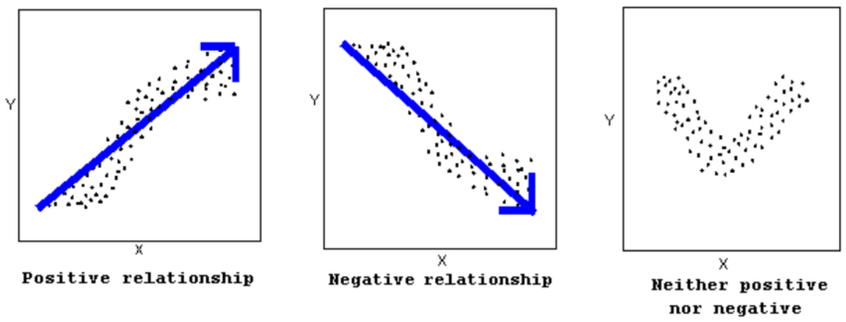
Univariate data		Bivariate data	
Overall pattern	Deviation from the pattern	Overall pattern	Deviation from the pattern
1) Shape 2) Center 3) Spread	Outliers	1) Direction 2) Form 3) Strength	Outliers

We have seen that for univariate data, we discussed the shape (symmetrical or skewed), center (median, mean and mode) and spread (interquartile range, standard deviation and range) of the distribution. For bivariate data, we will use descriptors like the direction, form and strength to describe the relationship between the two variables. For both univariate and bivariate data, data points that deviate significantly from the pattern of the main bulk of data points are called outliers.

Definition 3.2.5 The *direction* of the relationship can be either positive, negative or neither. We say that there is a positive relationship between two variables when an increase in one of the variables is associated with an increase in the other variable.

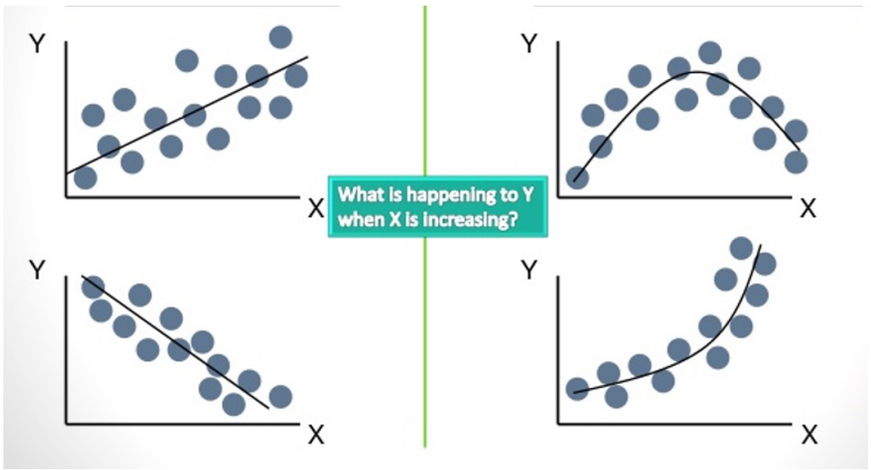
On the other hand, a negative relationship between two variables means that an increase in one variable is associated with a decrease in the other.

Not all relationships can be classified as either positive or negative and there are those that do not behave in one way or the other.



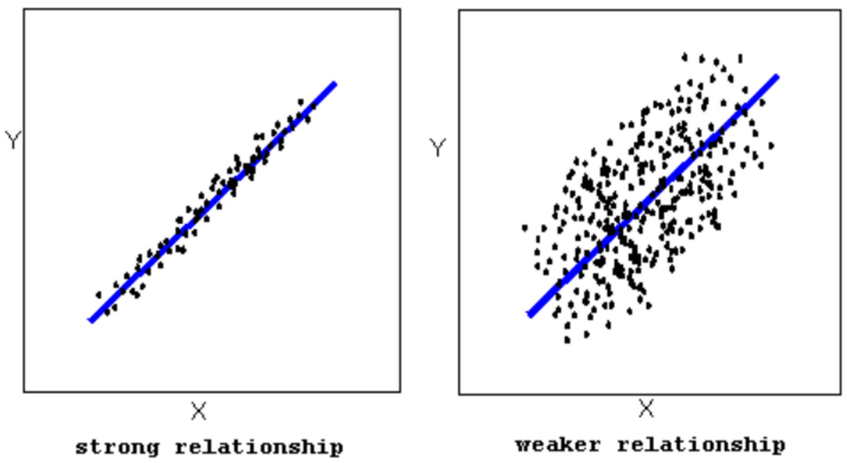
The *form* of the relationship describes the general shape of the scatter plot. In general, we can classify the form of the relationship as either linear or non-linear. The form of the relationship is linear when the data points appear to scatter about a straight line. Later in the chapter, we will use a mathematical equation to describe the straight line when the form of the relationship between two variables is linear.

When the data points appear to scatter about a smooth curve, we say that the form of the relationship is non-linear. It is beyond the scope of this course to summarise curve patterns in the data but it is useful to note that quadratic and exponential equations are examples of non-linear forms of relationship.



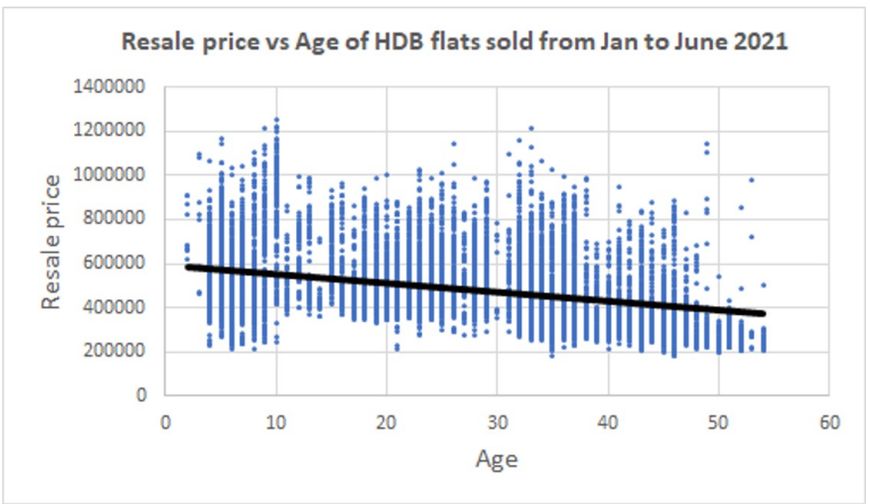
The two scatter plots on the left shows a linear form of the relationship between the two variables while the two scatter plots on the right shows non-linear forms.

The *strength* of the relationship indicates how closely the data follow the form of the relationship.



Both scatter plots above suggests that there is **positive**, **linear** relationship between the two variables. However, the scatter plot on the left shows the data points lying very close to the straight line. This indicates that the strength of the relationship is *strong*. The scatter plot on the right shows the data points scattered loosely around the straight line and thus the strength of the relationship is *weaker* than that in the scatter plot on the left.

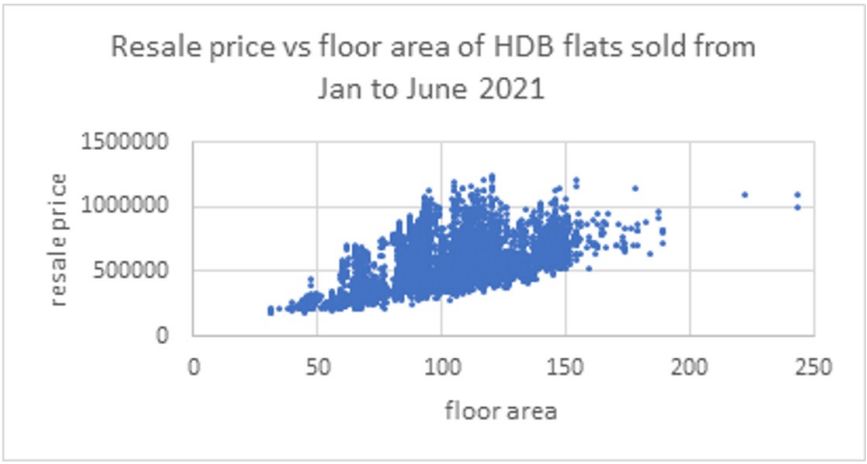
Example 3.2.6 Let us look at the scatter plot from the HDB resale flats data again. The scatter plot below is similar to the one from Example 3.2.4 except for an additional *trendline* drawn in black.



The trendline suggests that as the age of the HDB flat increases, the resale price decreases linearly on average, in the period of January to June 2021. Is this relationship strong or weak? In fact, one can argue that without the trendline, one may not even observe that there is in fact a linear relationship between age and resale price.

At this point, we cannot really tell if there is indeed a linear relationship and if there is, whether the relationship is strong or weak. Nevertheless, in the next section, we will discuss a more precise measure of the strength of a relationship.

As mentioned earlier, outliers are data points that deviate significantly from the pattern of the relationship. Consider the scatter plot shown below that plots the resale price against the floor area of the HDB resale flats. Do you observe any outliers?



Recall that for univariate data, using a boxplot, we can determine if a data point is an outlier by checking if its value is greater than $Q_3 + 1.5 \times \text{IQR}$ or smaller than $Q_1 - 1.5 \times \text{IQR}$. What about for bivariate data? We will discuss more about outliers in the next section.

Section 3.3 Correlation coefficient

In the previous section, using the HDB resale flats data set, we have observed that a flat's resale price is associated with the age of the flat. From the scatter plot, we concluded that the relationship between the age of the flat and the resale price of the flat was negative. This means that flats whose ages were higher tended to have a lower resale price. This is not surprising. However, can we say anything about whether this relationship is strong or weak? If possible, can we measure the strength of this relationship using a number?

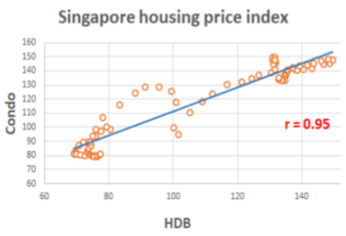
More generally, given two numerical variables, is it possible for us to measure the relationship between the two variables quantitatively?

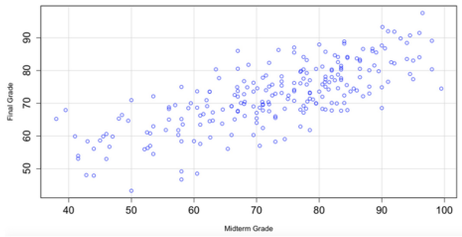
Definition 3.3.1 The *correlation coefficient* between two numerical variables is a measure of the **linear** association between them. The correlation coefficient, denoted by r, always ranges between -1 and 1. We can use this number to summarise the direction and strength of linear association between two variables.

The **sign** of r tells us about the **direction** of the linear association. If $r > 0$, then the association is *positive*, which means that when one of the variables increase, the other variable will tend to increase as well. On the other hand, if $r < 0$, then the association is *negative*, which means that when one of the variables increase, the other variable will tend

to decrease. In the event that $r = 1$ (resp. $r = -1$), we say that there is *perfect* positive association (resp. negative association). When $r = 0$, we say there is *no linear association*. Thus, while the sign of r tells us the direction of the linear association, the **magnitude** of r (that is, how close r is to 1 or -1) will tell us the strength of the linear association between two numerical variables.

Example 3.3.2 The two scatter plots below are examples of positive linear association between two variables.

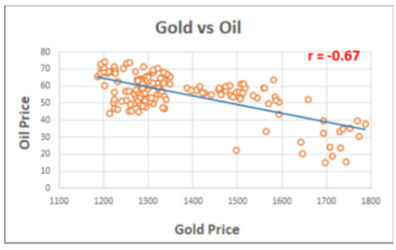


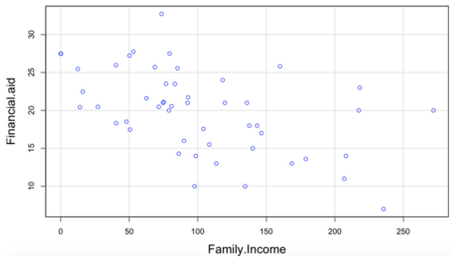


The plot on the left plots the price index of HDB flats against the price index of condominiums. We observe that there is positive linear association between the two indices, which means that as the price of HDB flats increase, it is likely that the price of condominiums would increase as well. The value of r in this case is 0.95 which indicates that the association is strong.

The plot on the right shows the midterm mark of students against the final mark. Again, we observe that there is positive linear association between the two marks and in this case, r was found to be 0.75.

The next two scatter plots are examples of negative linear association between two variables.

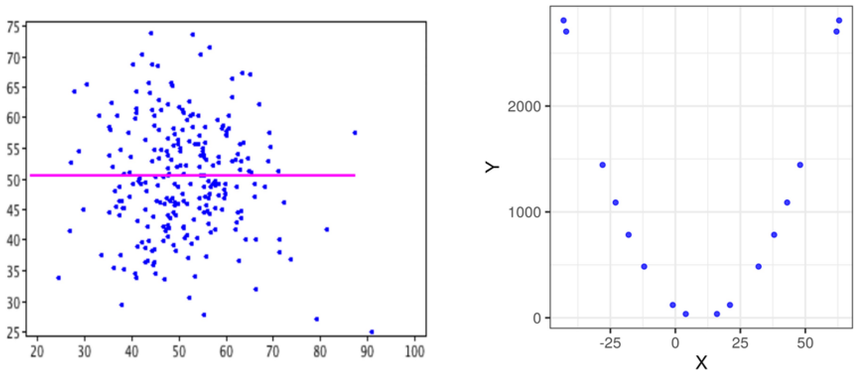




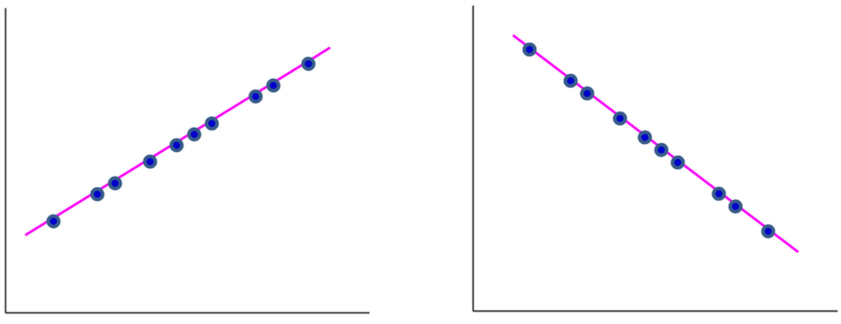
The plot on the left shows the price of oil against the price of gold. In this case, we observe that the trend is that when the price of gold increases, the price of oil tends to decrease. The value of r was found to be -0.67 and this indicates that there is negative linear association between gold and oil prices.

The plot on the right shows the amount of financial aid received by students against the students' family income. It is not surprising to find that as the family income increases, the

amount of financial aid received by students would tend to decrease. The value of r in this case is -0.49 and there is negative linear association between the two variables.

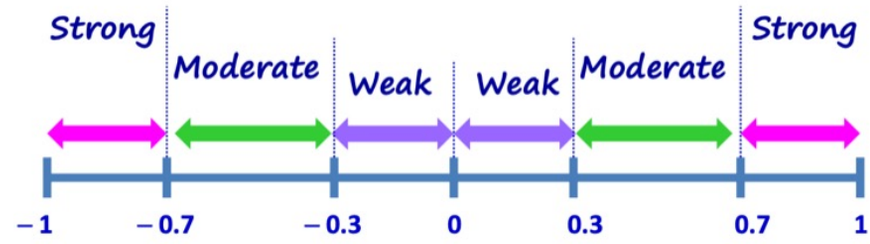


The two scatter plots above are examples where $r = 0$. This means that there is no linear association between the two variables. However, note that while $r = 0$ for the second plot, we can see that the data points fit very well onto a curve and there is a clear non-linear relationship between X and Y. More generally, no *linear association* between variables does not necessarily mean no association between variables.



The two plots above show situations where there is perfect (positive or negative) linear correlation between the two variables. In such cases, all the data points are connected by (and thus lie on) a straight line. There is however, one exception, which is when the straight line joining all the data points is actually a straight horizontal (or vertical) line. In such instances, the value of r is 0 and there is no association between the two variables. This is because when the data points are connected by a vertical or horizontal line, a change of value in one of the variables does not cause a change in the other variable.

When describing the strength of a linear relationship, we usually follow the rule of thumb as given in the diagram below.



When the magnitude of r is between 0.7 and 1, we say that the two variables have a *strong linear association*. If the magnitude is between 0.3 and 0.7, the two variables have a *moderate linear association*. If the magnitude is between 0 and 0.3, the two variables have a *weak linear association*. Do note that other sources may differentiate strong/moderate/weak linear association at other "cut-off" points that are different from 0.3 and 0.7.

In general, as the value of r becomes closer to 1 or -1, the data points will increasingly fall more closely to a straight line. Scatter plots where the data points are loosely dispersed typically mean that correlation is weak (or non-existent). We will now discuss how to compute the value of r numerically.

Example 3.3.3 We will go through the steps required to compute the correlation coefficient using an example. Consider the following table that shows a total of 10 data points of bivariate data (x, y):

x	9	4	5	10	6	3	7	2	8	1
y	41	17	28	50	39	26	30	6	4	10

1. First compute the mean and standard deviation of x and y. (Refer to Definition 1.4.1 and Definition 1.5.1 if you have forgotten how these are computed.) For this data set, we find the mean and standard deviation of x to be 5.5 and 3.03 respectively while the mean and standard deviation of y are 25.1 and 15.65 respectively.

2. Convert each value of x and y into *standard units*. To convert x (resp. y) into its standard unit, we compute
$$\frac{x-\overline{x}}{s_x} \qquad \left(\text{resp. } \frac{y-\overline{y}}{s_y}\right),$$
where s_x and s_y are the standard deviations of x and y respectively. The table below shows the values of x and y after they have been converted to standard units.

x	1.16	−0.50	−0.17	1.49	0.17	−0.83	0.50	−1.16	0.83	−1.49
y	1.02	−0.52	0.19	1.59	0.89	0.06	0.31	−1.22	−1.35	−0.96

3. Compute the product xy in their standard units for each data point. The table below has an additional row for the value xy for each data point.

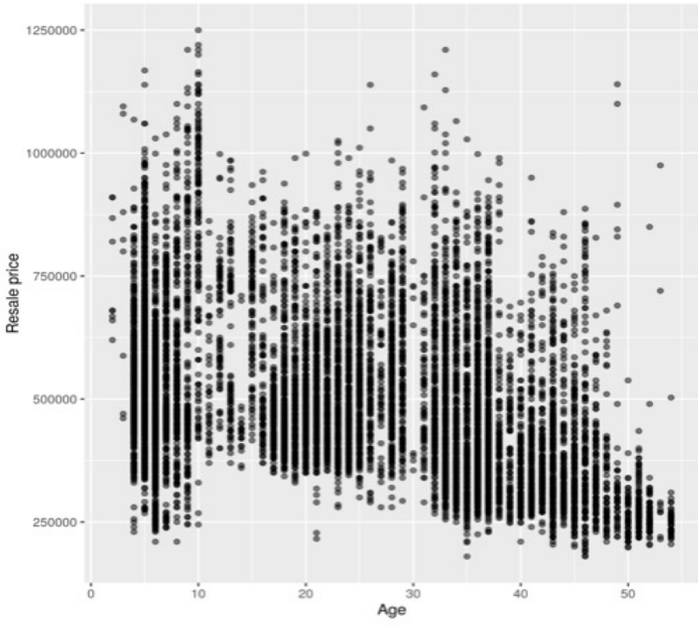
x	1.16	-0.50	-0.17	1.49	0.17	-0.83	0.50	-1.16	0.83	-1.49
y	1.02	-0.52	0.19	1.59	0.89	0.06	0.31	-1.22	-1.35	-0.96
xy	1.17	0.26	-0.03	2.36	0.15	-0.05	0.15	1.41	-1.11	1.43

4. Sum the products xy obtained in the previous step over all the data points and then divide the sum by $n-1$, where n is the number of data points. The result is the correlation coefficient r. For the data set above,

$$r = \frac{1}{9}(1.17 + 0.26 - 0.03 + 2.36 + 0.15 - 0.05 + 0.15 + 1.41 - 1.11 + 1.43) = 0.64.$$

Remark 3.3.4 For the purpose of this module, you will not be required to compute r manually, instead you should be familiar with the method of how r is computed and thereby develop some basic intuition on the properties of r.

Example 3.3.5 Let us revisit Example 3.2.6, where the scatter plot of HDB resale flat prices against the age of the flat shown below does indeed suggest that these two variables are negatively associated.



Indeed, upon computing the correlation coefficient between these two variables, we find that $r = -0.356$, confirming that there is moderate negative linear association between the age and resale price of HDB flats from the period January to June 2021.

We will now present three properties of correlation coefficients.

1. From the "Age" vs. "Price" of HDB resale flats example, we saw that $r = -0.36$ when we consider the scatter plot with Age as the x-axis and resale price as the y-axis. What would happen to r if we had done the plot with resale price as the x-axis and age as the y-axis? In other words, what happens to r when we interchange the x and y variables? If we revisit the process that describes how r is computed from a bivariate data set, you would realise that regardless of which variable is x (or y), the computation of r would not be affected in any way.

   > The correlation coefficient r is not affected by interchanging the x and y variables.

2. What would happen to the value of r if we add a constant to all the values of a variable? For example, suppose it was discovered that there was an error in the recording of all the resale prices of HDB flats and that the actual resale prices were all \$1000 higher than what was given in the data set. To correct this error, we would have to add \$1000 to all the resale prices in the data set. It turns out that such a change **does not** affect the value of r.

   > The correlation coefficient r is not affected by adding a number to all values of a variable.

   While this may not be immediately obvious, you are encouraged to verify this result by using the data set in Example 3.3.3 and adding some number to all the values of x (or y).

3. Instead of adding the same number to all the values of a variable, what would happen to the value of r if we multiply a positive number to all the values of a variable instead? For example, if the resale prices were converted to US dollars instead? This means that we have to multiply a factor of 0.73 (assuming an exchange rate of 1 Singapore dollar is to 0.73 US dollars) to all the resale prices in the data set. It turns out that such a change again **does not** affect the value of r.

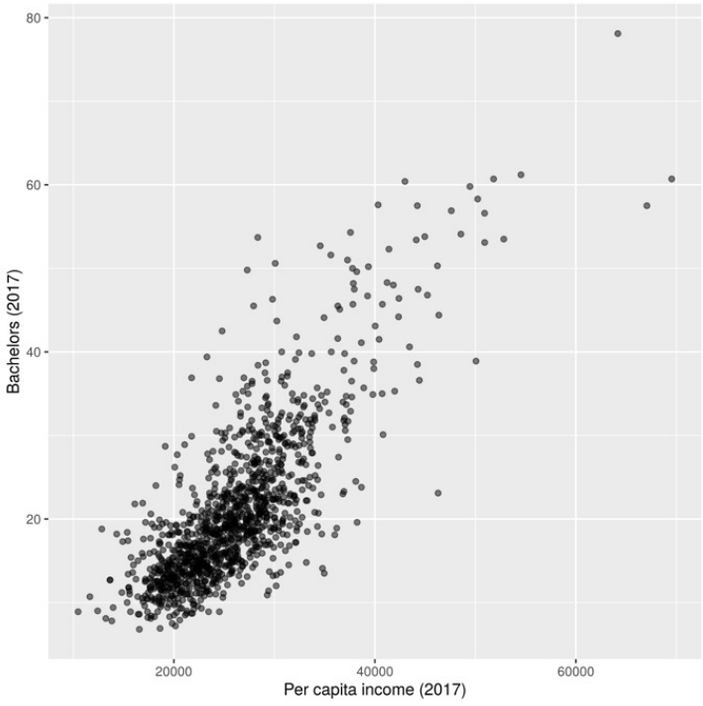
   > The correlation coefficient r is not affected by multiplying a positive number to all values of a variable.

   You are again encouraged to verify this result by adjusting the data set in Example 3.3.3 and recalculating the correlation coefficient.

While the correlation coefficient between two numerical variables is insightful, there are certain limitations.

Discussion 3.3.6

1. **Association is not causation.** To confuse association with causation is a common mistake that is made by many. Very often when there is a strong association between two variables, with a correlation coefficient of r that is close to 1 or -1, it is mistakenly concluded that any change in the explanatory variable, say x, will cause the response variable y to change. This is incorrect as what we can conclude is only a *statistical relationship* between x and y and not a *causal relationship*.



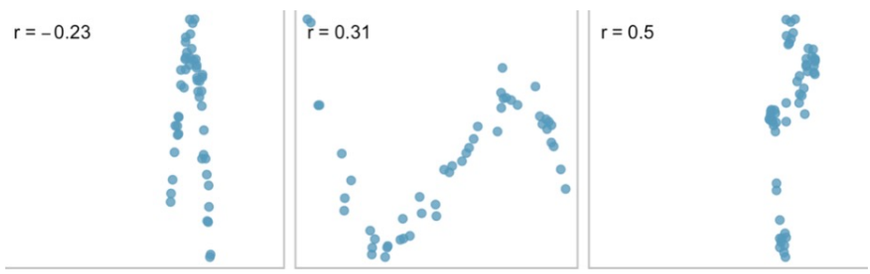
Consider the example above of a scatter plot that came from a data set containing information on the percentage of people that earned a Bachelor's Degree in 2017 across 3142 counties in the United States, as well as the per capita income of these counties in 2017.[2] Each data point in the scatter plot represents a county. The x-axis is the per capita income in the past 12 months while the y-axis is the percentage of the population

[2]Data set can be downloaded from www.openintro.org/data/?data=county_complete.

in the county that earned a Bachelor's Degree in 2017. The correlation coefficient for the two variables is 0.79, which indicates that there is strong and positive association between the two variables.

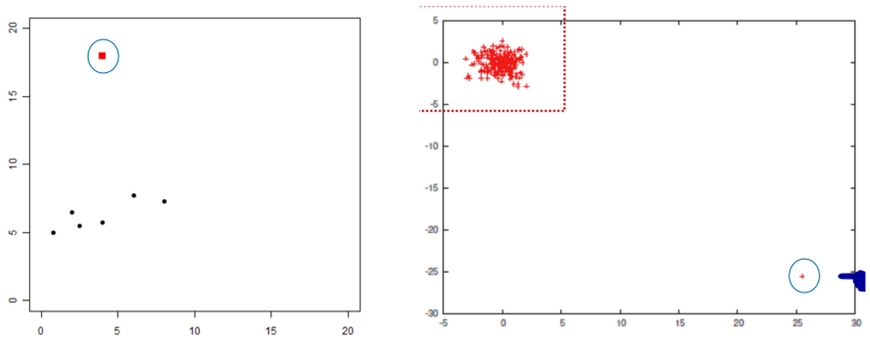
It would be tempting to conclude that the higher the per capita income of a county, the higher the percentage of the county's population would have earned a Bachelor's Degree. This is not necessarily true. The data here merely suggests association of the two variables and does not establish any causal relationship.

2. r **does not tell us anything about non-linear association.** The correlation coefficient r, as defined and described in this section, measures the degree of linear association between two numerical variables. Whatever the computed value of r is, it does not give any indication of whether the two variables could be associated in a non-linear way.



The correlation coefficients for the three scatter plots above are small but yet there is actually a strong relationship between the variables. The value of r is small because the relationship between the variables is not a linear one. It is always a good practice to look at a scatter plot of the data set and not just deduce any relationship between the variables from the computed value of r.

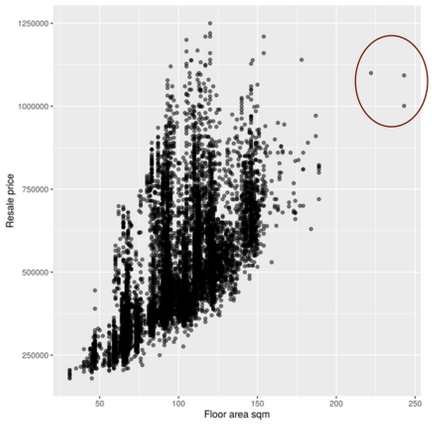
3. **Outliers can affect the correlation coefficient significantly.** Outliers are observations that lie far away from the overall bulk of the data. How do outliers affect the value of the correlation coefficient? The removal of outliers from a data set can have different effects on the correlation coefficient, depending on how the outlier is positioned in relation to the rest of the data points.



Consider the scatter plot on the left, where the outlier is circled, the correlation coefficient is 0.22 based on the data set that includes the outlier. However, when we remove the outlier, we see that there is a strong positive linear association between the remaining data points. Thus, in this case, the presence of the outlier decreases the strength of the correlation, compared to when the outlier is removed.

Consider the scatter plot on the right where again the outlier is circled. In this case, the correlation coefficient is -0.75 based on the data set that includes the outlier. When the outlier is removed, the remaining data points give a correlation coefficient of 0.01. Thus, in this case, the presence of the outlier actually increases the strength of the correlation, compared to when the outlier is removed.

Example 3.3.7 For the HDB data set that we introduced earlier, the scatter plot below shows the relationship between the resale price and the floor area of the flat. There are three outliers (circled) and these are resale flats whose floor areas are larger than 200 square meters.



Using a statistical software, it was found that the correlation coefficient was 0.626 before the outliers were removed. After the outliers are removed, the correlation coefficient becomes 0.625, which is practically the same as before.

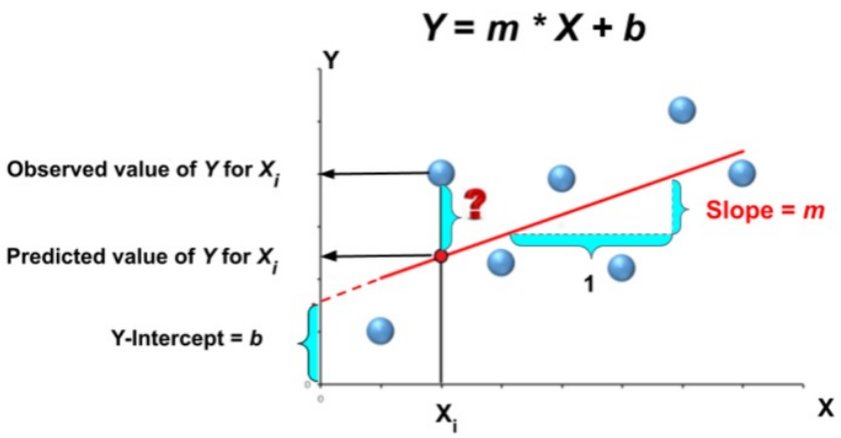
Section 3.4 Linear regression

Now that we have seen that the age of a HDB resale flat is negatively associated with the resale price, it is reasonable to wonder if we can make some predictions on the resale price of a flat given the age of the flat. For example for a flat that is 40 years old, what is our guess for its resale price?

Definition 3.4.1 If we believe that two variables X and Y are linearly associated, we may model the relationship between the two variables by fitting a straight line to the observed data. This approach is known as *linear regression*. Recall that the equation of a straight line is given by

$$Y = mX + b,$$

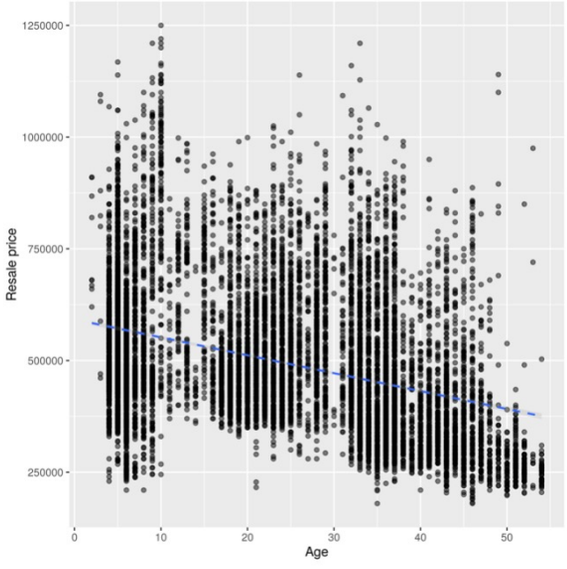
where b is the *y-intercept* and m is the *slope* or *gradient* of the line. The y-intercept is the value of Y when the value of X is 0. The slope of the line is the amount of change in Y when the value of X increases by 1.



In the figure above, the straight line in red is the regression line that is fitted to the observed data, represented by the blue dots. Consider the i-th observation (X_i, Y_i). The "?" in the figure represents the *residual* of the i-th observation, which is the observed value

of Y for X_i (that is, Y_i) minus the predicted value of Y for X_i (predicted by the straight line). This residual, denoted by e_i, is sometimes also called the *error* of the i-th observation as it measures how far the predicted value is from the observed value.

Example 3.4.2 Let us return to the question we posed at the beginning of this section. What is our prediction for the resale price of a HDB flat that is 40 years old?



With X representing the age of the resale flat and Y being the resale price, the regression line obtained from the data set is

$$Y = -4007X + 591857.$$

This means that when $X = 40$, (age of resale flat is 40),

$$Y = -4007 \times 40 + 591857 = 431577.$$

So the **predicted** resale price of a 40 year old flat is \$431,577. It is important to note that we are **not concluding** that

> A 40 year old resale flat will be sold at \$431,577.

But instead our linear regression model predicts that

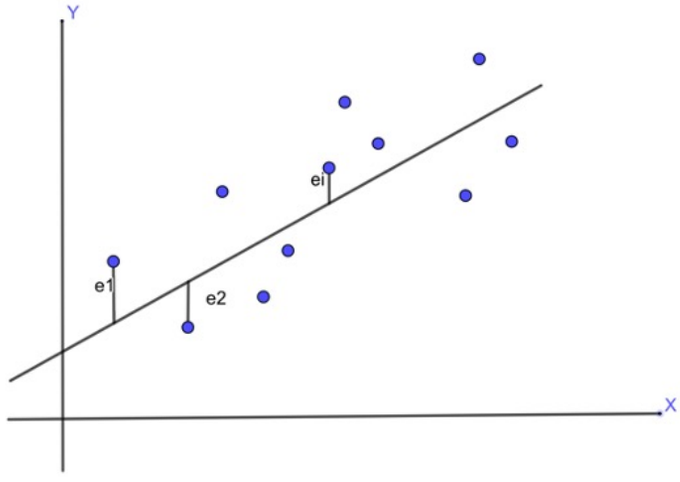
The average resale price of 40 year old HDB flats is \$431,577.

Furthermore, as the correlation between resale flat price and age of the flat is weak, the prediction obtained from the linear regression above may not be as accurate compared to the scenario where the correlation is stronger.

Now that we have seen how a regression line can be used, the question is how do we obtain such a line given bivariate data? What method and principle is used to determine the regression line? Among the many different straight lines that we can use to fit the data points, which one is the "best"?

Discussion 3.4.3 There are several ways to assess which straight line fits the observed data better. One of the most common way is the *method of least squares*. For this module, we will not go into the technicalities of this method but instead we will briefly describe the idea behind this method.

Recall that when we fit a straight line through a set of observed data points (x_i, y_i), the difference between the observed value y_i and the predicted outcome, predicted by the straight line, is known as the residual of the i-th observation. This residual, denoted by e_i is also known as the error of the i-th observation that measures how far is the observed from the predicted.



In the plot above, we see that each data point gives rise to an error term and it is reasonable to say that a line of good fit is one that keeps the error terms (considered over all data points) small. However, instead of looking at the overall error by summing up

$$e_1 + e_2 + \cdots + e_n,$$

where n is the total number of data points, the method of least squares seek to find a straight line that minimises the overall **sum of squares of errors**,

$$e_1^2 + e_2^2 + \cdots + e_n^2.$$

You may wonder why minimising $e_1^2 + e_2^2 + \cdots e_n^2$ is more appropriate than minimising $e_1 + e_2 + \cdots + e_n$. We will leave you to ponder about this question before having a discussion with your friends or instructor.

Remark 3.4.4

1. It is important to note that while we have obtained the least squares regression line that allows us to predict the average resale price for a given age of the resale flat, the same regression line cannot be used to predict the average age of resale flats for a given resale price. The reason is essentially because of the way the regression line was obtained.

   In obtaining the regression line with the independent variable (x) as age and the dependent variable (y) as the resale price, the line was fitted to minimise the square of error terms between the observed and predicted resale price.

   If the intention was to use a given resale price to predict the average age of the resale flats, then we would be looking at another regression line that minimises the square of error terms between the observed and predicted age of resale flats.

   The two regression lines are different and thus not interchangeable.

2. The correlation coefficient r between the variables X and Y is closely related to the regression line

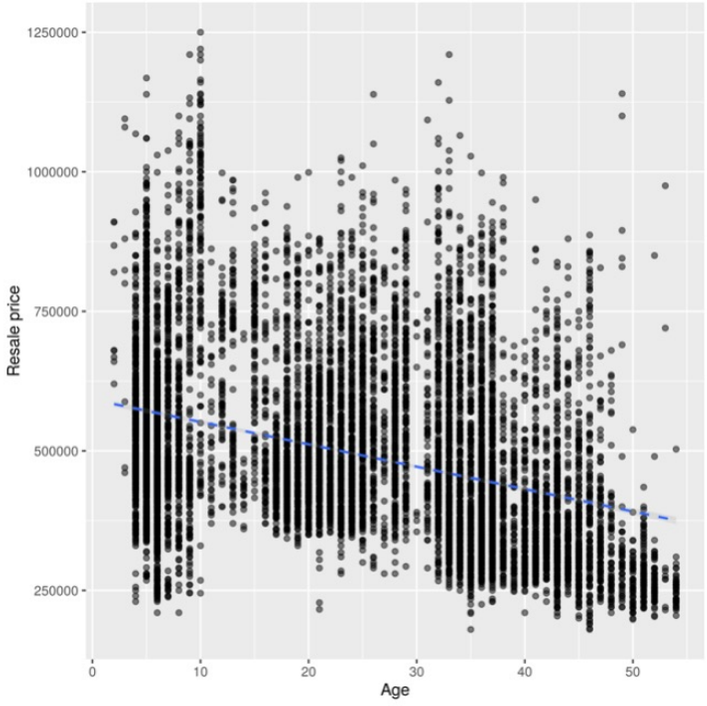
   $$Y = mX + b$$

   obtained using the method of least squares. More precisely, we have

   $$m = \frac{s_Y}{s_X} r,$$

   where s_X (resp. s_Y) is the standard deviation of X (resp. Y). With this relationship, we see that if the correlation coefficient r is positive, then the gradient of the regression line is also positive. Similarly, if the correlation coefficient is negative, then the gradient of the regression line will also be negative. However, it is important to remember that the correlation coefficient is not necessarily **equal** to the gradient of the regression line.

3. Another important point to note about the linear regression line obtained using a data set is with regards to the range of the independent variable in the data set.



Recall that we have obtained the linear regression line for the purpose of predicting the average resale price based on the age of the resale flat. From the data set, the value of the independent variable (in this case, this is the age of the resale flat) ranges from 2 to 54 years. Thus the prediction that can be arrived at using the regression line is only applicable for HDB flats whose age is between 2 and 54 years old. Outside this range, we should not use the regression line to make our prediction as the best fit regression line may change outside this range. For example, we should not use the regression line to predict the average resale price of flats that are 60 years old as our data set does not contain any information on resale flats that are more than 54 years old.

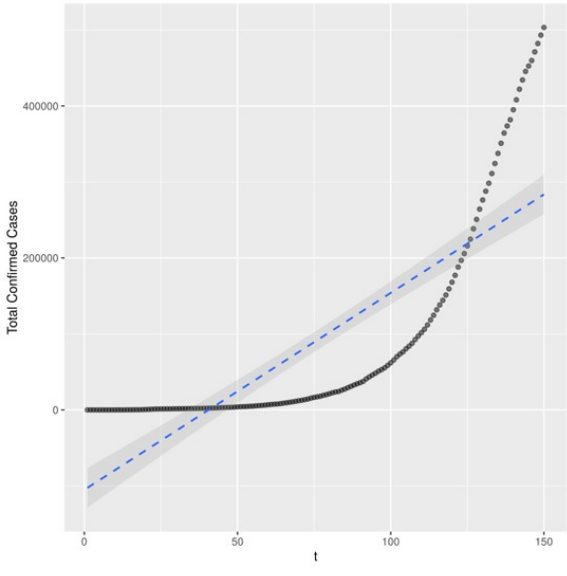
Discussion 3.4.5 To conclude this section and also the chapter, we will describe a method to study the relationship between two variables if the relationship is not linear. The following table shows part of a data set that provides the total number of confirmed COVID-19 cases in South Africa since 5 March 2020.[3].

[3]Data set can be downloaded from www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset.

t	Total confirmed cases
76	17200
77	18003
78	19137
79	20125
⋮	⋮
95	48285
96	50879
⋮	⋮

In this data set, t is the variable representing the number of days since 5 March 2020.

It can be computed using Microsoft Excel or other statistical software that the correlation coefficient between the total number of confirmed cases and t is 0.812, which indicates that there is a strong positive linear association between the two variables. Is this indeed the case? Perhaps we may make such a conclusion but as stated earlier, correlation coefficient alone does not give the entire picture. We should create a scatter plot using our bivariate data and verify if there is really a linear relationship.



Are the two variables associated linearly? It is quite clear visually that the total number of confirmed cases increases exponentially when t increases. Thus, if we let y be the variable representing the total number of confirmed cases, y and t are not linearly associated but instead the relationship between them seems to be exponential. For such a situation, can we

apply our linear regression technique to make predictions on the total number of confirmed cases? The answer is yes, but it would have to be done indirectly.

Now, if the relationship between y and t is indeed exponential in nature, we can model this relationship using the equation

$$y = cb^t,$$

where c and b are some constants that we will determine. Using the property of the logarithmic function, we see that

$$y = cb^t \quad \text{is equivalent to} \quad \ln y = \ln(cb^t) \quad \text{is equivalent to} \quad \ln y = \ln c + t \ln b.$$

Thus, instead of making a scatter plot with y plotted against t, we will make a scatter plot with $\ln y$ plotted against t. If there is indeed an exponential relationship between y and t, then we would expect to see a linear relationship between $\ln y$ and t, as indicated by the equivalent equations above. Let us go through the steps:

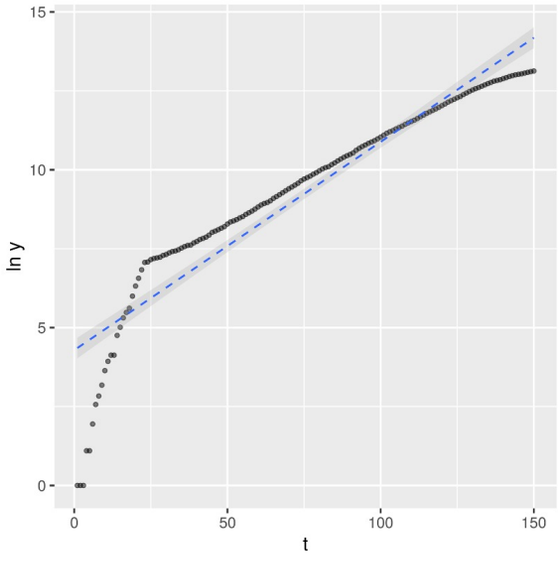
(a) Step 1: For each data point (t, y), compute $(t, \ln y)$. For our data set on COVID-19 cases in South Africa, we have the following table:

t	Total confirmed cases (y)	$\ln(y)$
76	17200	9.753
77	18003	9.798
78	19137	9.860
79	20125	9.910
⋮	⋮	⋮
95	48285	10.785
96	50879	10.837
⋮	⋮	⋮

We then plot $\ln y$ against t.

(b) Step 2: Find the linear regression line for $\ln y$ vs t. For our example, the regression line was found to be

$$\ln y = 4.287 + 0.066t.$$



This means that $\ln c = 4.287$ and $\ln b = 0.066$.

(c) Step 3: Since $\ln c = 4.287$ and $\ln b = 0.066$, we have

$$c = e^{4.287} \quad \text{and} \quad b = e^{0.066}.$$

We are now able to write down the exponential equation relating y and t:

$$y = cb^t = e^{4.287}e^{0.066t} = e^{4.287+0.066t}.$$