Topic1: Data and Graphical Summaries

Outline

Example: Australian Road Fatalities Jan-April 2016

The number of road fatalities in Australia continues to rise, given the ever increasing volume of vehicles on the road, despite preventative measures as compulsory seat belts and school zones. Last year in Australia, 1,209 died on our roads.

Data from the Australian Bureau of Statistics (ABS) from the first four months of 2016, gives the following variables:
Crash ID, State, Date, Day, Month, Year, Dayweek, Time, Hour, Min, Crash Type, Bus Involvement, Rigid Truck Invovement, Articulated Truck Involvement, Speed Limit, Road User, Gender, Age. See DataDictionary

What questions do you have?

Identifying Variables

```
## data <- read.csv("2016Fatalities.csv",header=T)
data[1,]  #Extracts the 1st row
```

```
##     Crash.ID State     Date Day   Month Year Dayweek  Time Hour Min
## 1 2.2016e+12   VIC 1-Jan-16   1 January 2016  Friday 20:30   20  30
##       Crash.Type BusInvolvement RigidTruck..Involvement
## 1 Single vehicle             No                      No
##   Articulated.Truck..Involvement. SpeedLimit         RoadUser Gender Age
## 1                              No         80 Motorcycle rider   Male  25
```

```
names(data) #Lists all the variables
colnames(data)  #Lists all the variables
head(data)  #List the 1st 5 rows of data
class(data)  #Shows the way R has stored the data
```

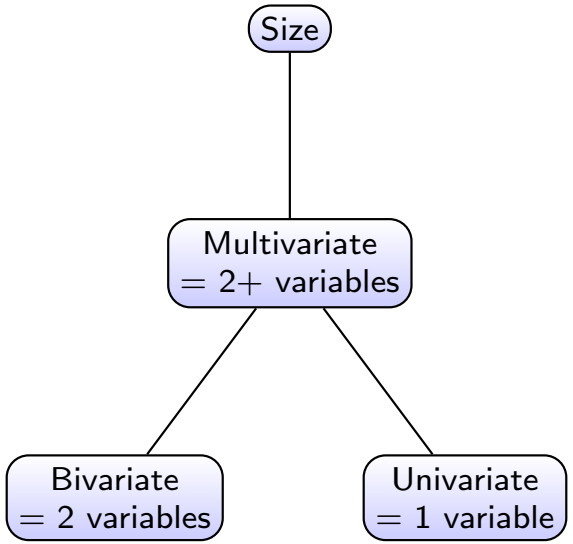
```
dim(data)
```

```
## [1] 442  18
```

The 1st step in EDA is to identify the variables, in terms of form and type.

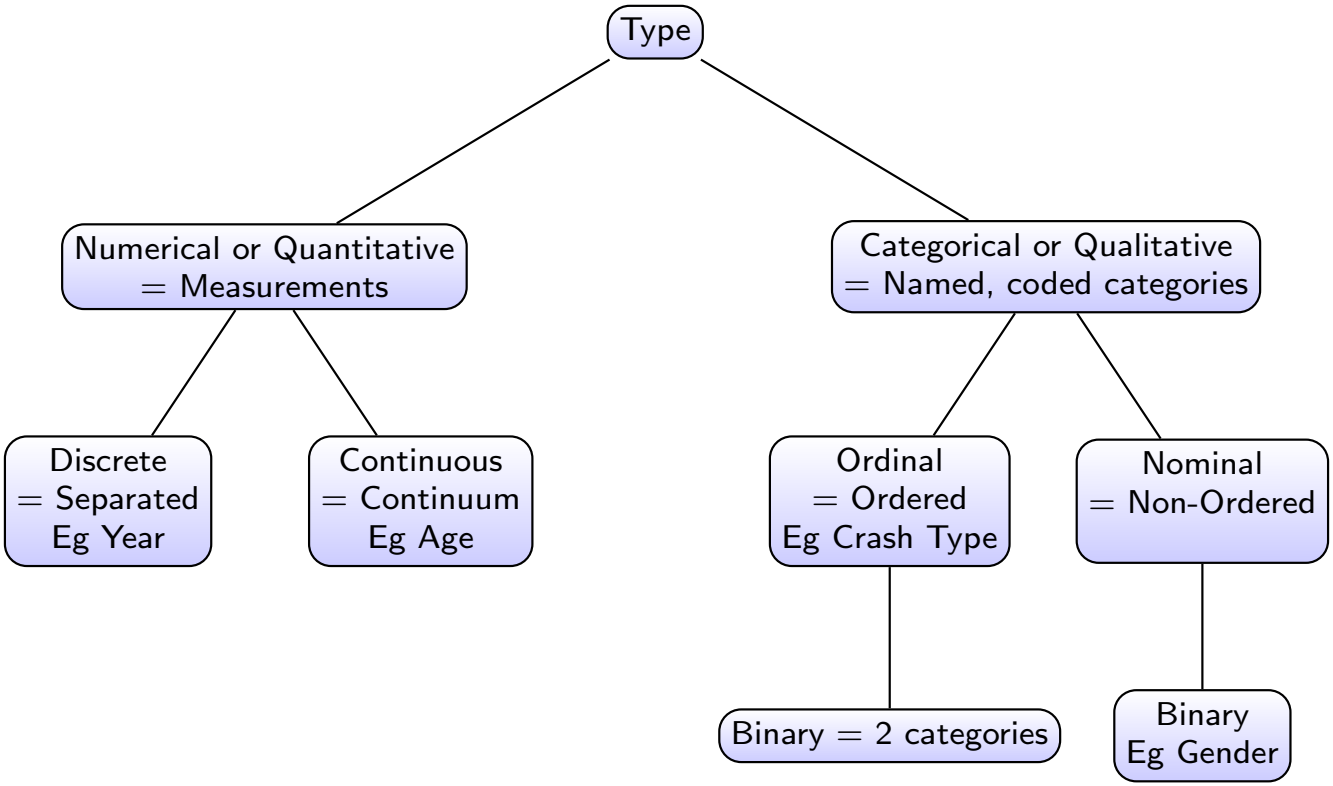
(i) Size of Variables

How many bits of information or 'variables' have been recorded?
In 'big data' we commonly have 'large p, small n' meaning that we have stacks of variables (eg gene data) relative to the data size.



(ii) Type of Variables

What is the nature of the variables – i.e. what process or situation 'produced' the data?

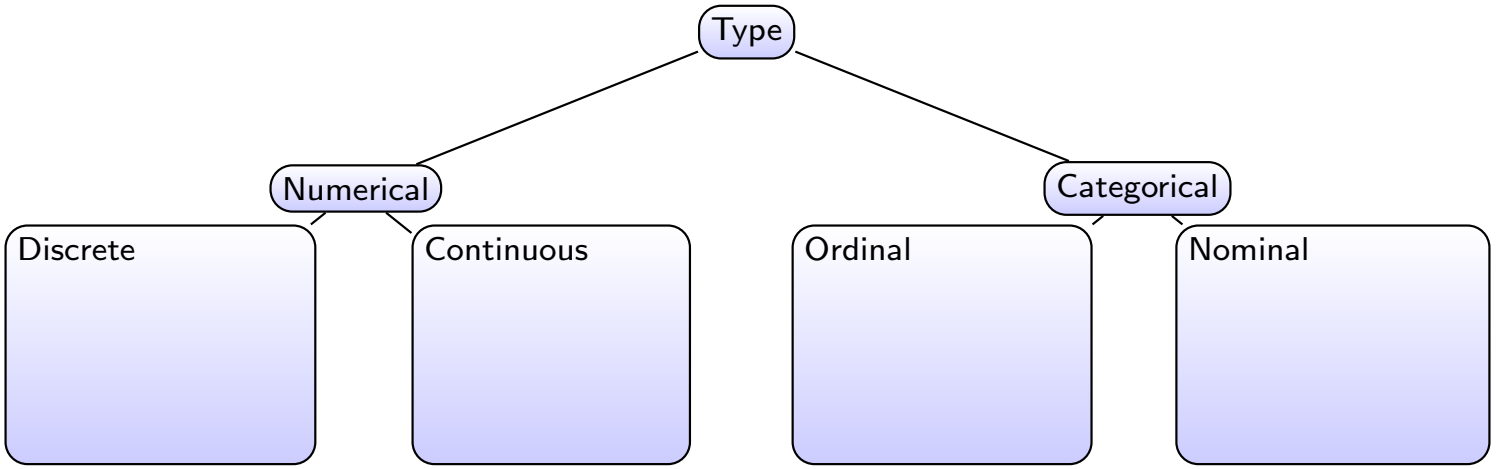


Note:

- In practise, continuous data is often reported as discrete data (by rounding), but the underlying quantity represented is still continuous (eg Age and Time).
- A helpful diagnostic for determining continuous data is to ask: "Could this data have been recorded to higher accuracy, given a more precise 'instrument'?"
- Quantitative data can be simplified to qualitative data. For example, in a survey, a respondent may feel more comfortable giving a general answer to a question about their personal income.

Have a try

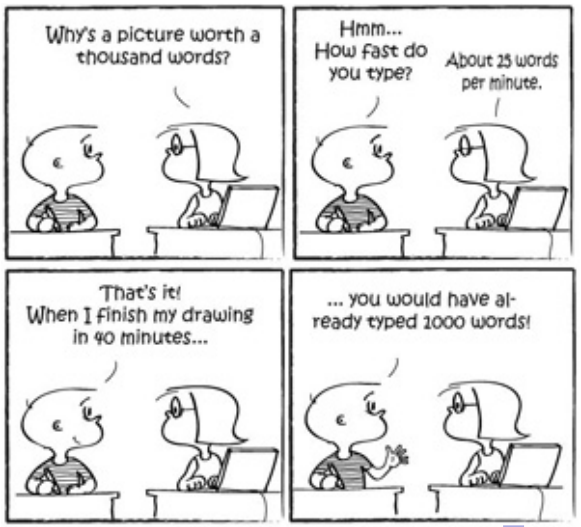
Identify all the variables for Australian Road Fatalities.



Graphical Summaries

Once we identify the variables, we can summarise the data, both graphically and numerically, in order to identify and highlight the main features of interest. A careful choice of graphical and numerical summaries can give a quick, transparent, perceptive snapshot of the data.

We often start with graphical summaries because 'A picture is worth a thousand words.' (Similar idea: Arthur Brisbane, Syracuse Advertising Men's Club, 1911)



How to choose an appropriate graphical summary?

The critical question is: 'How can I visually represent this data?' or 'What plot will best highlight features of the data?'. This knocks out pie charts and 3D charts!

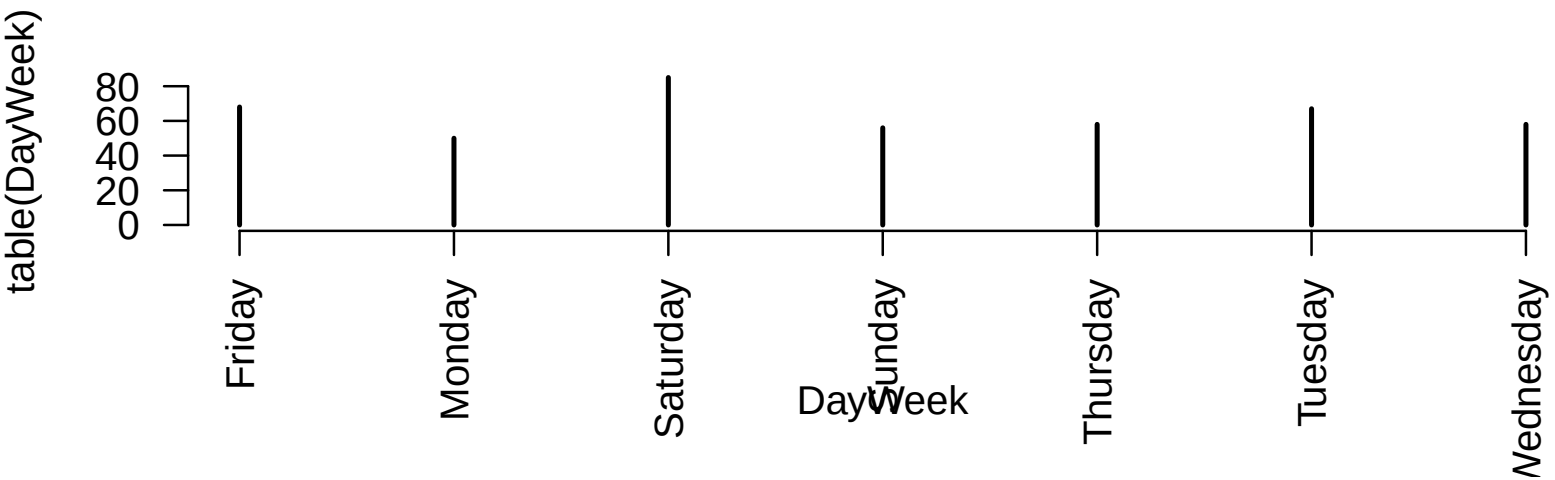
To some extent we use trial and error. We try some standard forms and see what is revealed about the data. One graphical summary can suggest another, and often a combination will highlight different features of the data

In practise we use computer packages like R to construct summaries. However, it is important to understand how to construct graphical summaries 'by hand', so that you understand how to interpret computer output and for your final exam. Some computer packages vary slightly in construction. For example, in the calculation of the quartiles or the length of the whiskers in the boxplot.

Summary0: Barplot (Categorical data)

Q: What was the most common day of road fatality?

```
DayWeek <- data$Dayweek
table(DayWeek)


## DayWeek
##    Friday    Monday  Saturday    Sunday  Thursday   Tuesday Wednesday
##        68        50        85        56        58        67        58


plot(table(DayWeek),las=2)
```



Summary1: Frequency table and ordinate diagram (discrete data)

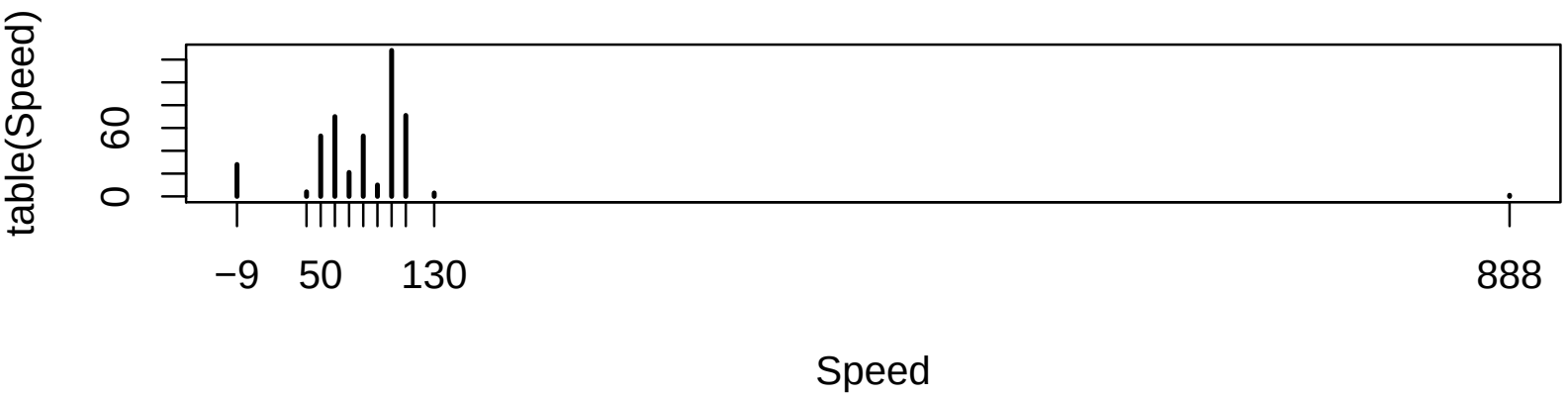
Q: What was the most common speed at which a road fatality occurred?

The frequency table is a very simple way to summarise a set of discrete data and when plotted gives an ordinate diagram.

Speed	-9	40	50	...	130	888	Total
Frequency	28	4				1	442

What is strange? Why?

```r
Speed <- data$SpeedLimit  #Extracts SpeedLimit
table(Speed)

## Speed
##  -9  40  50  60  70  80  90 100 110 130 888
##  28   4  53  70  21  53  10 128  71   3   1

plot(table(Speed))
```



Summary2: Frequency table and histograms (continuous data)

Q: What were the most common ages at which a road fatality occurred?

The frequency table can also be used to summarise a set of continuous data, by collecting it into intervals (or 'bins'). What is lost?

- For equal bin lengths, we can simply sort the data into the bins, and then plot the frequency against each bin. This is called a 'regular' histogram.
- For unequal bin lengths, we need to sort the data into the bins, then work out the relative frequency (=frequency/sample size) and the height (=relative frequency/interval length). Plotting the height against each bin is called a 'probability' histogram.

(i) Using equal bins: Regular Histogram

Bin	Frequency
[-10,0)	?
[0,10)	11
[10,20)	
...	
[90,100)	9

```
Age <- data$Age
min(Age)

## [1] -9

max(Age)

## [1] 96
```

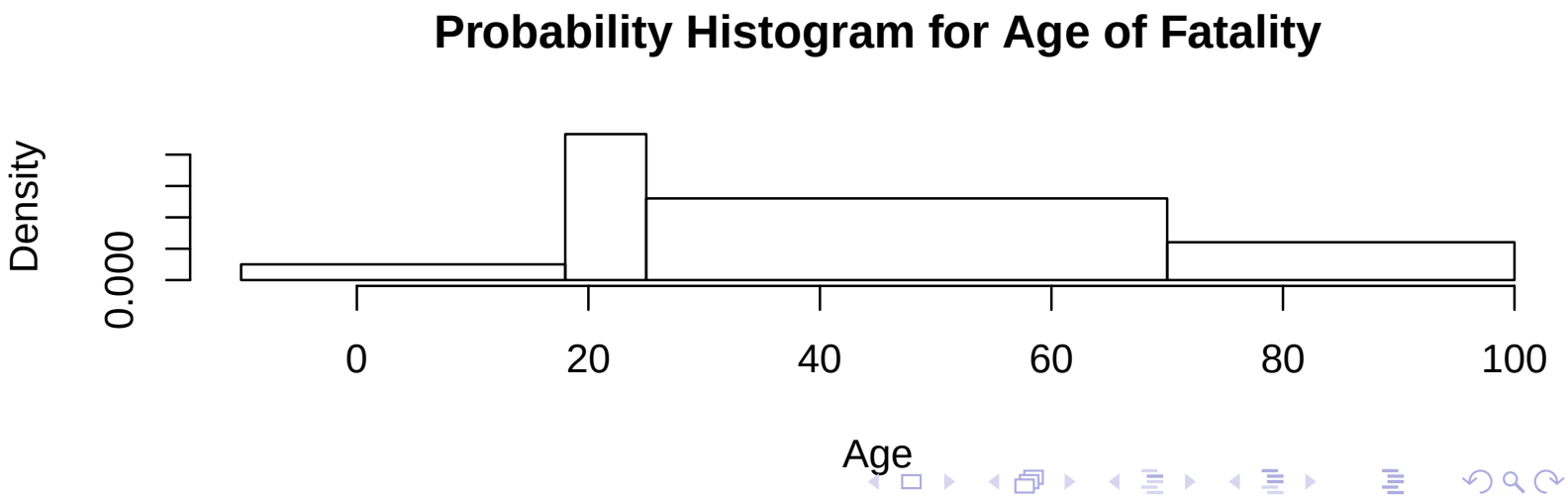
```r
Age <- data$Age
hist(Age,xlab="Age",
     main="Regular Histogram for Age of Fatality")
```

(ii) Using unequal bins: Probability Histogram

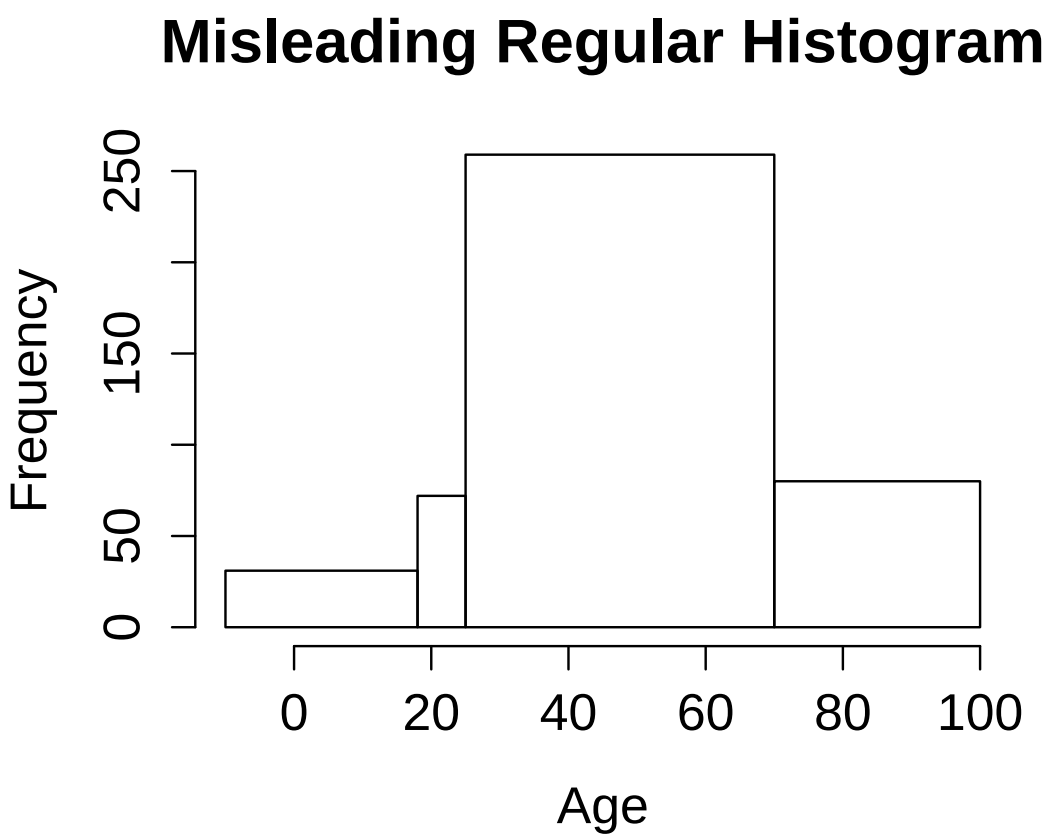
Bin	Frequency	Relative Frequency	Height
[-10,18)	31	$31/442 = 0.07$	0.0025
[18,25)	72	$72/442 = 0.16$	0.0232
[25,70)	259	$259/442 = 0.59$	0.0130
[70,100)	80	$80/442 = 0.18$	0.0060
Total	442	1	

where:
Relative Frequency = Frequency/442
Height = Relative Frequency/Bin length
Eg For bin [-10,18): height = 0.07/28 =3.6.

```r
breaks=c(-10,18,25,70,100)
table(cut(Age,breaks,right=F))

##
## [-10,18)  [18,25)  [25,70) [70,100)
##       31       72      259       80

hist(Age,br=breaks,freq=F,right=F,
     xlab="Age",
     main="Probability Histogram for Age of Fatality")
```



Note how the 'regular' histogram is misleading for unequal bin lengths, as it suggests that [25,70) is the most likely bin.



Summary3: Stem and leaf plot

Q: What were the 3 highest ages at which a road fatality occurred?

A stem and leaf plot is basically a histogram turned on its side. It is useful for moderately sized data sets. It provides both a sense of the shape and an ordering of the data, while retaining all the raw numerical data (up to a certain decimal place).

The value to the left of the | is called the 'stem' and the values on the right are called 'leaves'. The leaves should be ordered, although sorting will not affect the shape of the plot.

```r
stem(Age)
```

```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##   -0 | 99
##    0 | 011124557
##    1 | 00345556777777777778888888888888999999999
##    2 | 000000000001111112222222222333333333344444444445555555555566666666+12
##    3 | 000000000011111222223333333344455566666666677777788888899
##    4 | 00000111111222222233333333344444444445555555566666666666777777788888
##    5 | 000111112333333444555567777788888889999
##    6 | 00000001111111222333344455556777888999999999
##    7 | 000111223334444445555566667777888899
##    8 | 00011111222222222333334446778888899 9
##    9 | 001222336
```

Note that R defaults to what it considers to be a sensible layout of the data. Here R chooses a 'single' stem plot: with each stem having the leaves 0,1,2,...9. So the reading 2 | 3 is age 23. If we consider the data is over-condensed (too stretched out) or under-condensed (too bunched up), we can adjust the format by experimenting with `scale=`.

```r
stem(Age,scale=0.25)
```

```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##   -0 | 99
##    0 | 011124557003455567777777777778888888888888999999999
##    2 | 00000000000111111222222222223333333333344444444444555555555556666666666+69
##    4 | 000001111112222222333333333344444444444555555555666666666677777778888888+36
##    6 | 00000001111111222333344455556777888999999990001112233344444455555666
##    8 | 000111112222222223333344467788888999001222336
```

This is called a double leaf plot, as the stem '0' now has the leaves 0,1,2,3,4 5,6,7,8,9 (representing 00-09) and then a second set of leaves 0,1,2,3,4 5,6,7,8,9 (representing 10-19). Note you need to read carefully, as 8—0 can represent both 80 or 90.

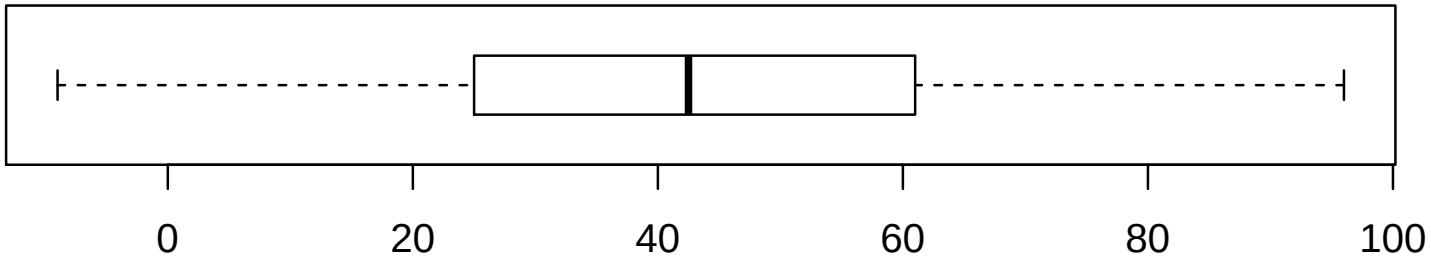
A double stem plot would have one stem '0' with leaves 0,1,2,3,4 (representing 00-04) and then a second stem 'O' with leaves 5,6,7,8,9 (representing 05-09).

Summary4: Boxplot

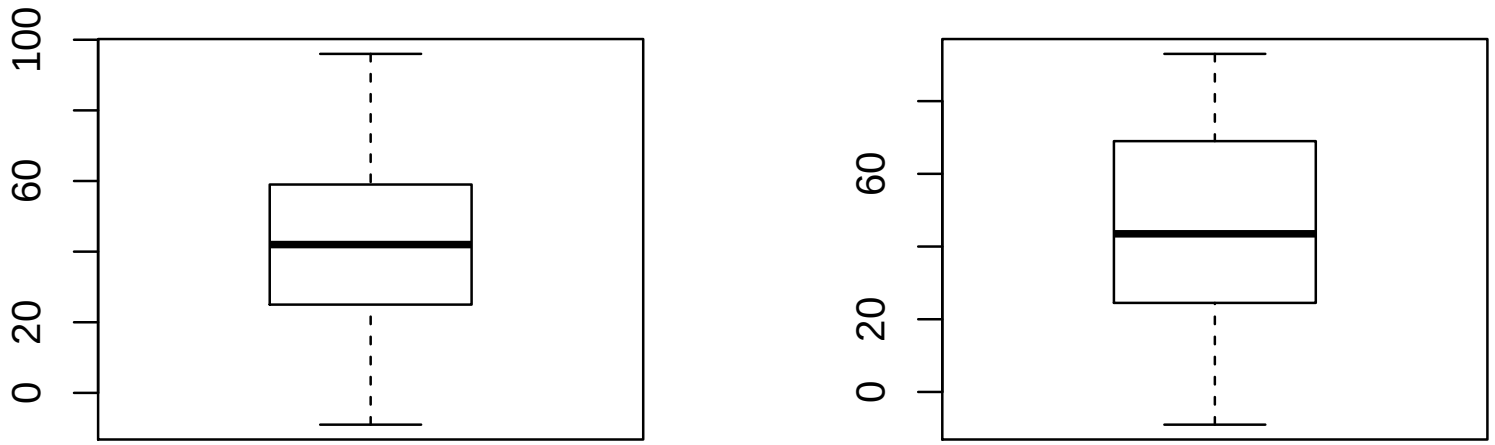
Q: Were there any unusual ages at which a road fatality occurred? Is there any difference between the ages of male and female fatalities?

Boxplots are useful for comparing data sets and identifying outliers.

```
boxplot(Age,horizontal=T)
```

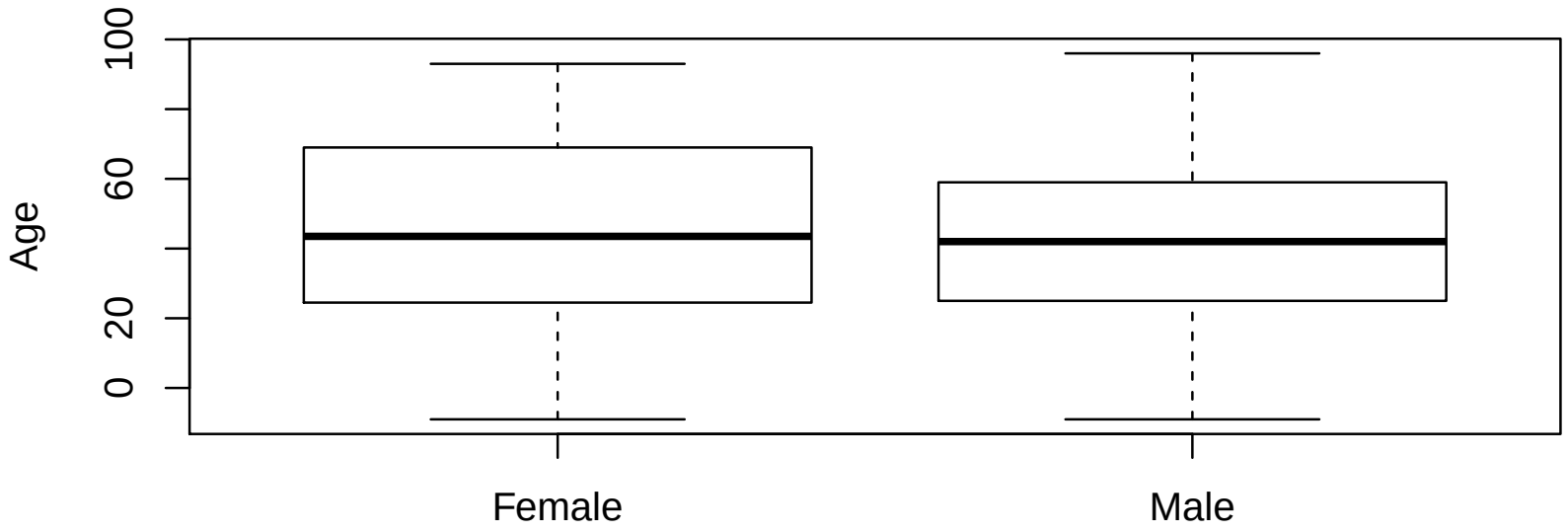


```r
AgeM <- data$Age[ data$Gender == "Male"]
AgeF <- data$Age[ data$Gender == "Female"]
par(mfrow = c(1, 2))  #Puts 2 boxplots in a row
boxplot(AgeM)
boxplot(AgeF)
```



A neat trick for producing the same boxplots:

```
boxplot(Age~data$Gender, ylab="Age")
```

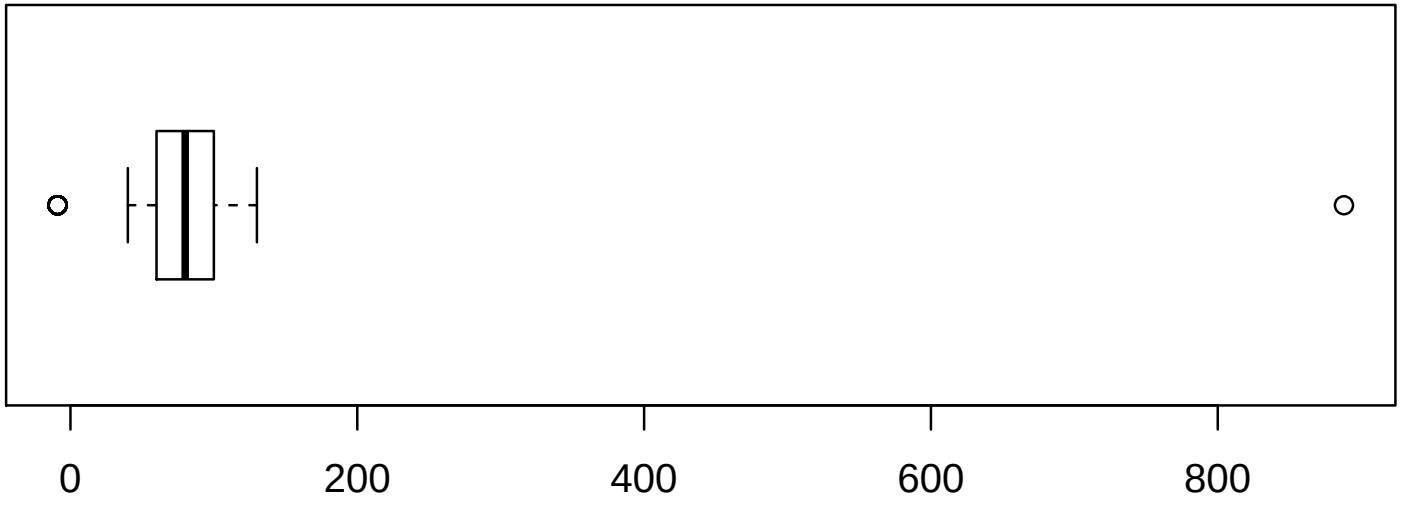


The boxplots show that the ages of road fatalities for men and women is similar. However, what do we learn from this simple commmand?

```
length(AgeM)

## [1] 326

length(AgeF)

## [1] 116
```

Q: What were there any unusual speeds at which fatalities occurred?

```
boxplot(data$SpeedLimit, horizontal = T)
```



A box plot is a visual representation of the 5 number summary (min, Q_1= 1st quartile, Q_2= median, Q_3= 3rd quartile, max), where:

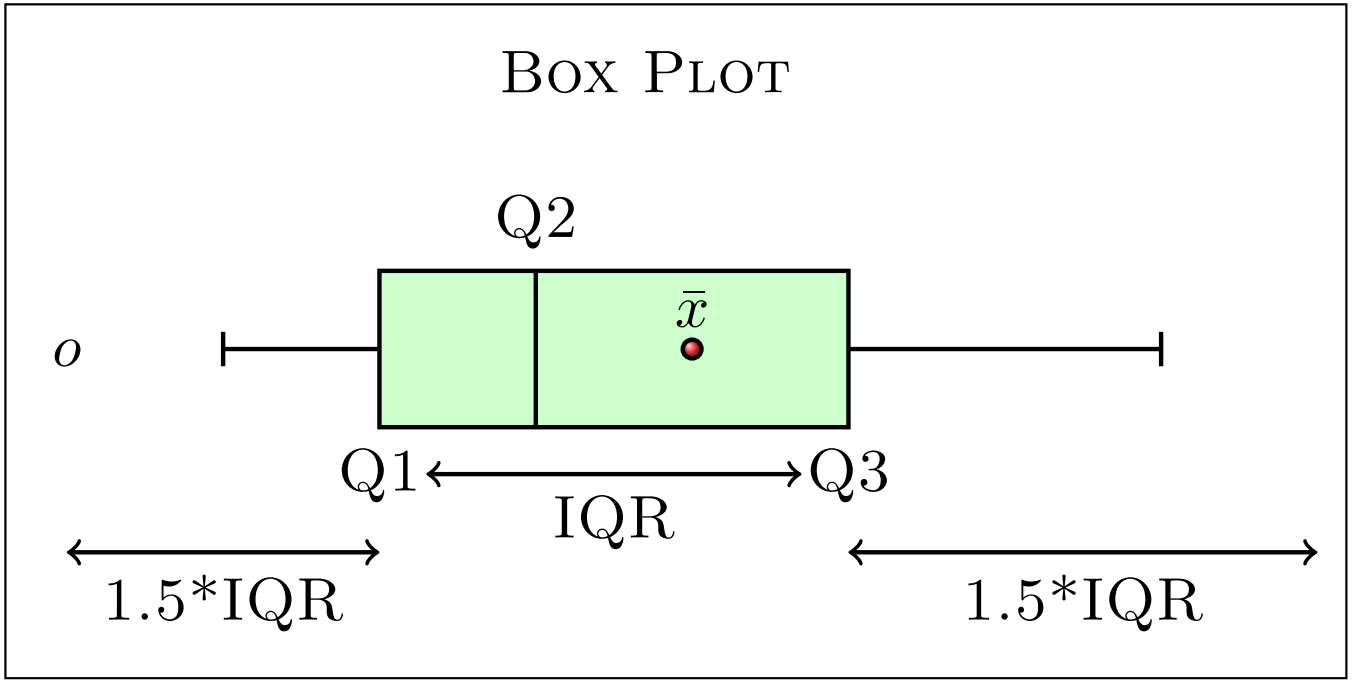
- Min = smallest data point(s)
- Max = largest data point(s)
- Q_2 = middle data point (find the average of the 2 middle points for even sized dataset.)
- Q_1 and Q_3 are the 'medians' of the half data sets: we divide the data into 2 sets at the median (including the median for an odd sized data set), and then find the median of each half set of data.

See more fuller definitions: Quartiles

There are different conventions for boxplots. We will use the convention that the whiskers extend to the minimum and maximum observations within the thresholds [LT,UT], where

- Lower Threshold $LT = Q_1 - 1.5IQR$;
- Upper Threshold $UT = Q_3 + 1.5IQR$;
- Interquartile range is $IQR = Q_3 - Q_1$.

An outlier is any observation lying outside of [LT,UT].



Note: Here we have indicated the mean $\bar{x}$ in red for comparision with the median Q_2, but normally that is not shown on the boxplot.

Steps for Constructing a Boxplot by Hand

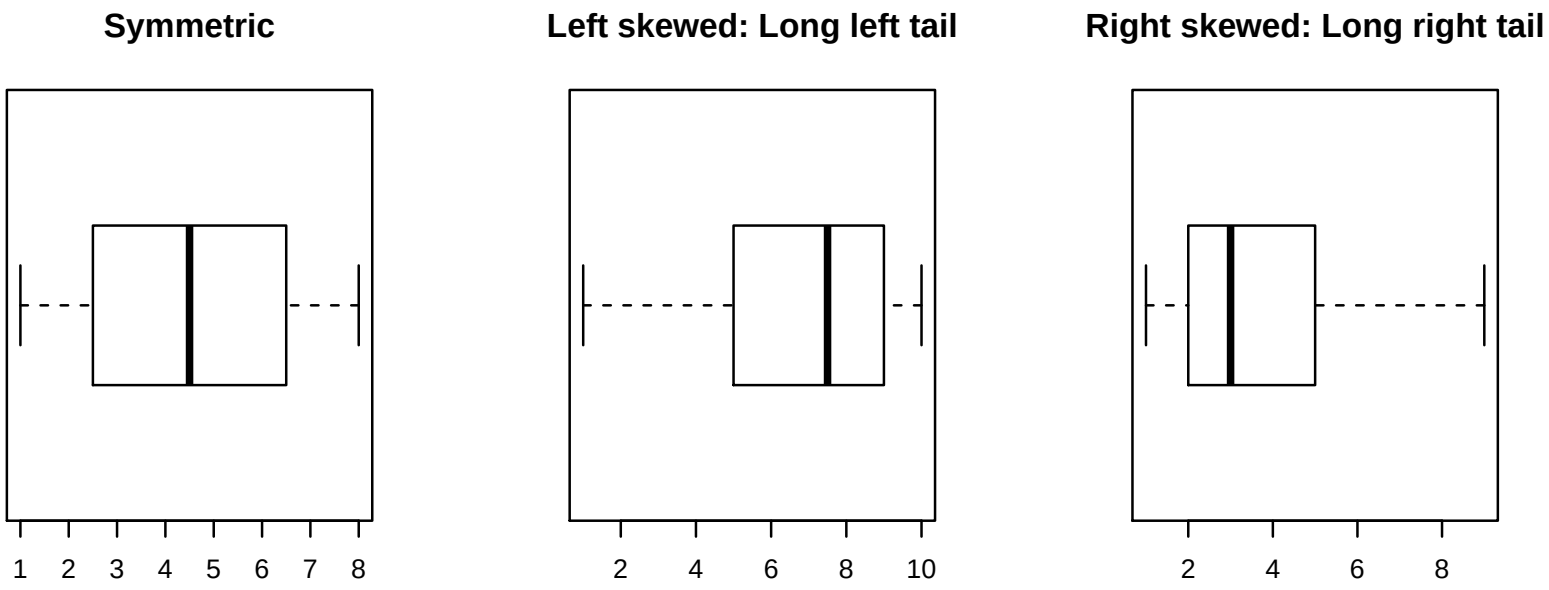
1. Calculate the quartiles Q_1, Q_2 and Q_3 and the interquartile range IQR.
2. Draw a box from Q_1 to Q_3, with a line within the box for the median$= Q_2$.
3. Calculate the upper and lower thresholds.
4. Draw a whisker from the box to the nearest points within the thresholds.
5. Any points outside the thresholds are outliers, designated by circles.

Describing the Shape of Data

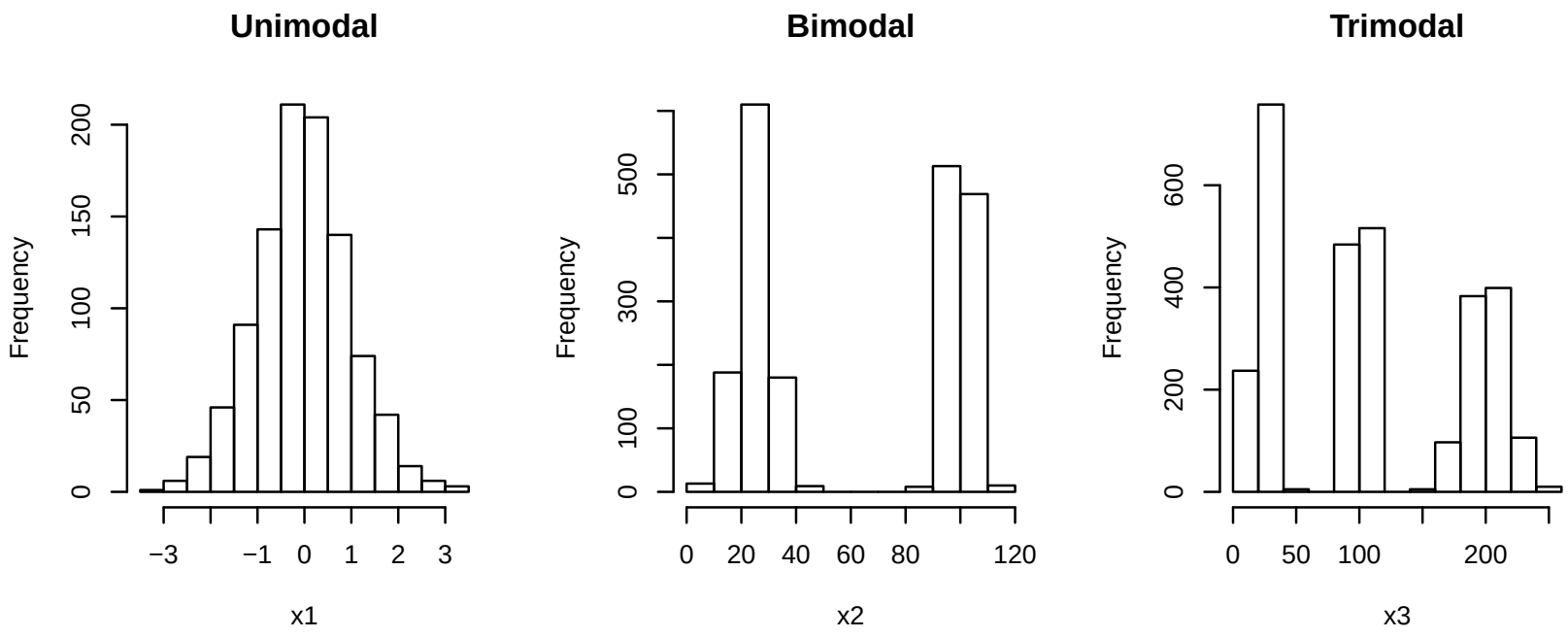
When we look at graphical summaries, we want to describe the form of the data and any 'idiosyncrasies'.

3 key questions:

1. Is it symmetric or skewed?

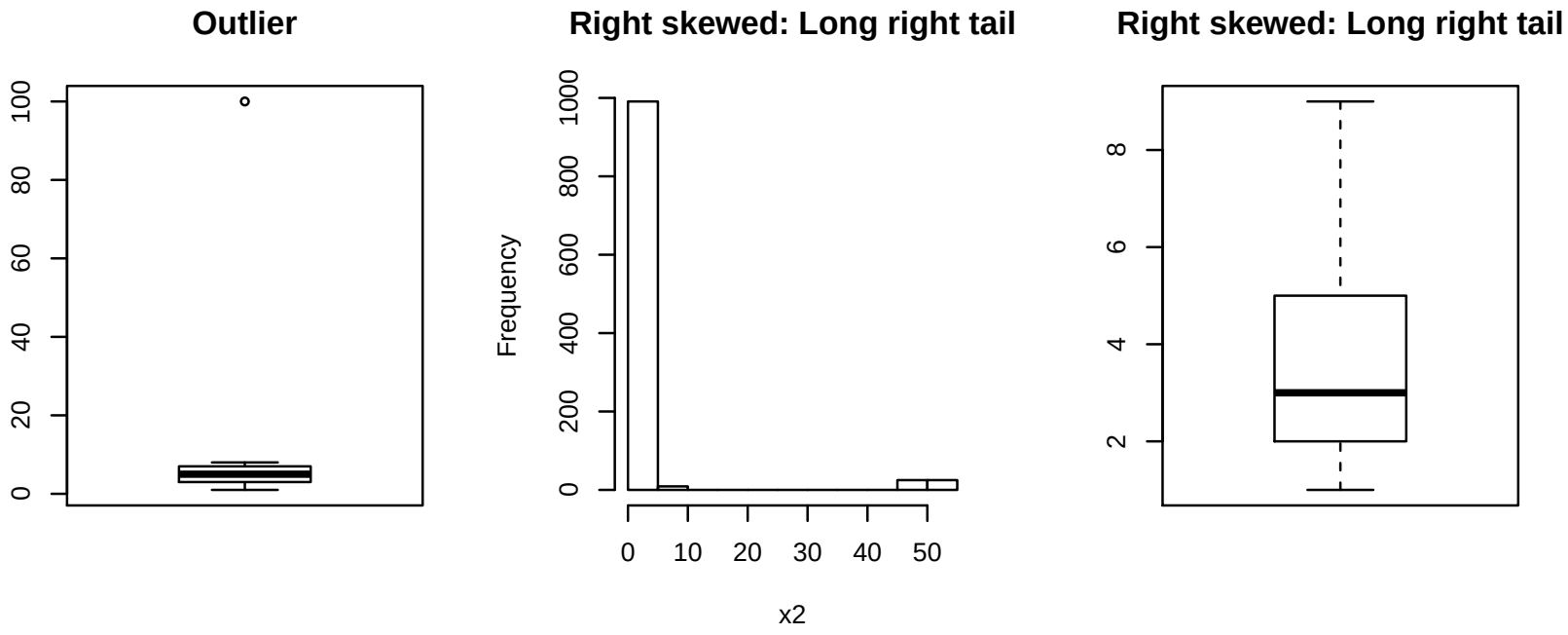


2. Is it unimodal, bimodal, trimodal, multimodal or other?



Note: Bimodality can be an indication of interesting behaviour to explore. However, it can also arise from 2 populations mistakingly put together.

3. Are there any unusual features? (eg outliers or gaps)



Note: An outlier needs to be investigated carefully, as it can be an indication of an interesting data point or possibly a mistake in the recording of data.

Dirty Data

Notice that throughout Topic1 we have deliberately showcased raw data with the missing values coded as '-9'. This is called 'dirty data' and is how real data exists.

Dealing with 'dirty data' is outside the scope of MATH1005, as it can require some sophistication. Ideally, we would replace all the missing values by a blank. However, one possible strategy is to account for the missing values in your histogram: create the bins (-10,0), [0,18] ..., so there is effectively 1 nonsense bin (although it still affects the calculation of frequencies of the other bins.)