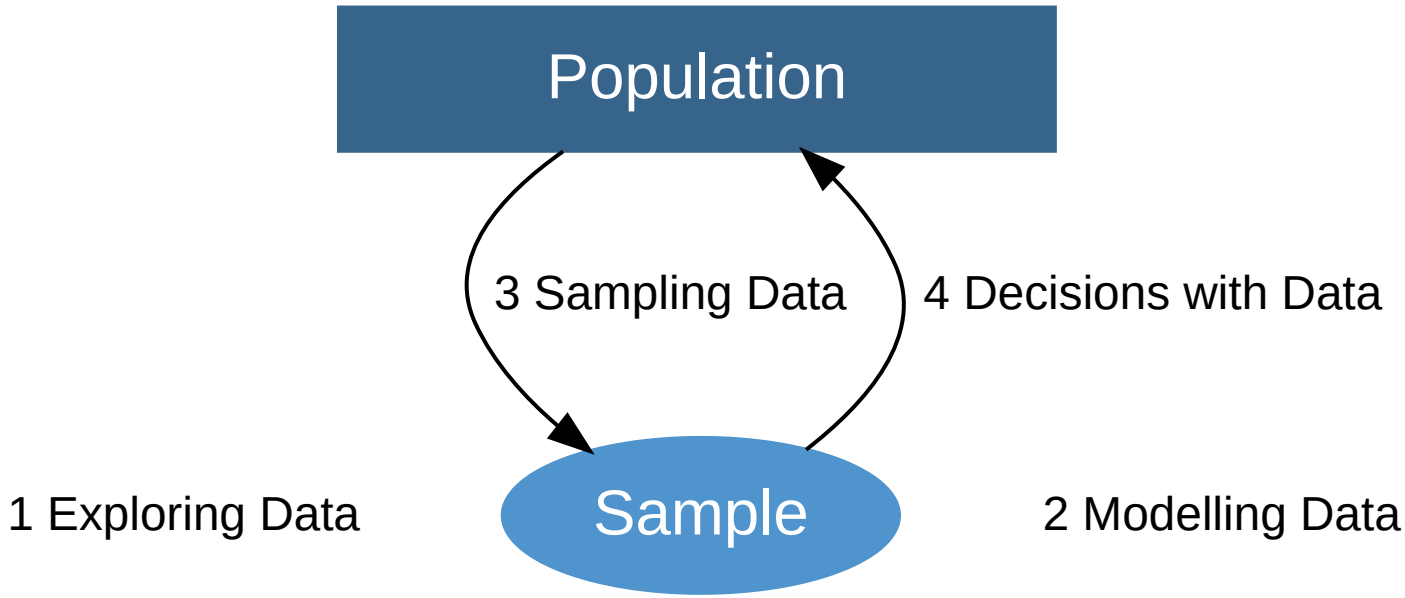
Centre

Exploring Data |Numerical Summaries

Unit Overview



Module1 Exploring Data

Design of Experiments

Where did the data come from & can we make reliable conclusions?

Data & Graphical Summaries

What type of data do we have & how can we visualise it?

Numerical Summaries

What are the main features of the data?

Centre

Data Story | How much does a property in Newtown cost?

Numerical Summaries

Mean and Median

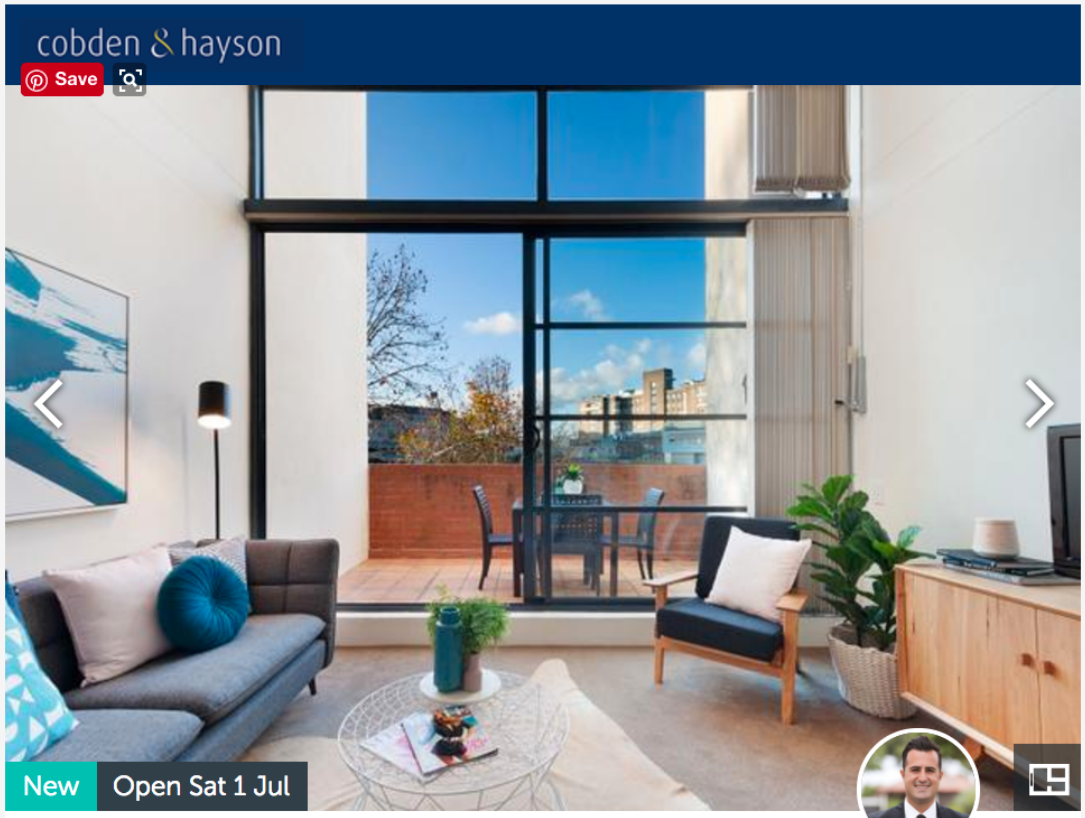
Robustness and Comparisons

More Examples

Summary

Data Story

How much does a property in Newtown cost?



Buyers Guide $600-$650k
Auction Sat 22 Jul
205w/138 Carillon Avenue, Newtown, NSW 2042
1 1 1

Jim Nikolopoulos

☆ Save Details >

BUSINESS THE ECONOMY PROPERTY MARKET

Sydney house prices fall for the first year since 2012

By Jennifer Duke
1 March 2018 — 10:49am

Cracks are showing in the Sydney property market, with prices now falling for the first time over a 12-month period since the boom began.

The harbour city recorded a 0.5 per cent drop in housing prices in the year to February, CoreLogic data released on Thursday shows. This figure includes apartments and houses.

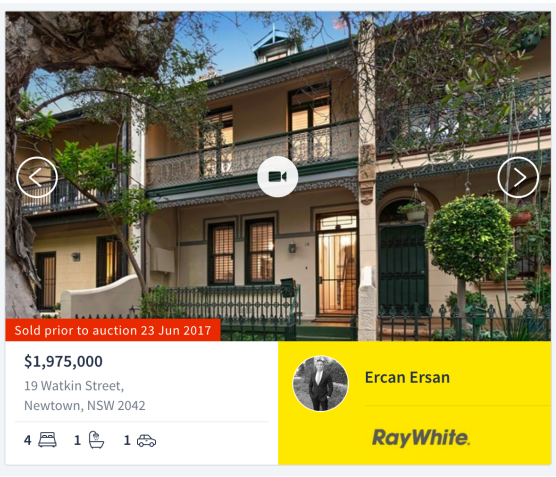
SMH

Data on Newtown Property Sales

- Data is taken from domain.com.au:
  - All properties sold in Newtown (NSW 2042) between April-June 2017.
  - The variable `Sold` has price in $1000s.

```
data <- read.csv("data/NewtownJune2017.csv",header=T)
head(data, n=2)
```

```
##                  Property  Type    Agent Bedrooms Bathrooms Carspots Sold
## 1 19 Watkin Street Newtown House RayWhite        4         1        1 1975
## 2  30 Pearl Street Newtown House RayWhite        2         1        0 1250
##      Date
## 1 23/6/17
## 2 23/6/17
```


Sold prior to auction 23 Jun 2017
$1,975,000
19 Watkin Street,
Newtown, NSW 2042
4 1 1
Ercan Ersan
RayWhite.


Save
Sold prior to auction 23 Jun 2017
$1,250,000
30 Pearl Street,
Newtown, NSW 2042
2 1 -
Ercan Ersan
RayWhite.

```
dim(data)
```

```
## [1] 56  8
```

```
str(data)
```

```
## 'data.frame':    56 obs. of  8 variables:
##  $ Property : Factor w/ 56 levels "1 Pearl Street Newtown",..: 20 26 24 23 8 55 47 33 14 54 ...
##  $ Type     : Factor w/ 5 levels "Apartment","House",..: 2 2 2 1 1 2 5 1 1 2 ...
##  $ Agent    : Factor w/ 18 levels "Belle","BresicWhitney",..: 15 15 1 15 14 15 16 6 13 10 ...
##  $ Bedrooms : int  4 2 2 1 1 5 1 1 1 3 ...
##  $ Bathrooms: int  1 1 1 1 1 1 1 1 1 2 ...
##  $ Carspots : int  1 0 0 1 1 1 0 1 1 0 ...
##  $ Sold     : int  1975 1250 1280 780 650 2100 675 740 625 1950 ...
##  $ Date     : Factor w/ 30 levels "1/4/17","1/5/17",..: 15 15 8 8 8 7 14 10 7 29 ...
```

Statistical Thinking

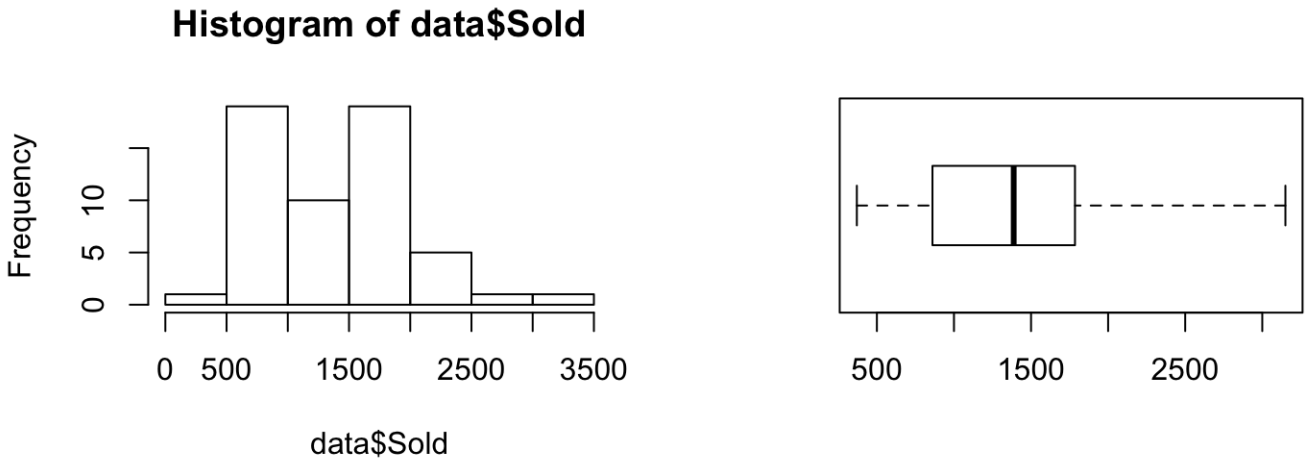
- Who might be interested in this data?
- Propose a research question and how you might investigate it.

Numerical Summaries

Recap: Graphical summaries

For the Newtown property data we could produce a histogram or boxplot.

```
par(mfrow=c(1,2))
hist(data$Sold)
boxplot(data$Sold,horizontal=T)
```



💬 What do they reveal about Newtown house prices?

Advantages of numerical summaries

- A numerical summary reduces all the data to 1 simple number (“statistic”).
  - This loses a lot of information.
  - However it allows easy communication and comparisions.
- Major features that we can summarise numerically are:
  - **Maximum**
  - **Minimum**
  - **Centre** [mean, median]
  - **Spread** [standard deviation, range, IQR]

💬 Which summaries might be useful for talking about Newtown house prices?

- It all depends how rich you are!
- Reporting the centre without the spread can be misleading!

Useful notation for data (Ext)

- In this course, we intentionally focus on statistical concepts in **words**. This is vital for collaborating with people from different fields. The mathematics is introduced in 2nd year - see the Maths Guide. However, here some simple mathematical notation is helpful.

- A dataset of size n can be represented by

$$x_1, x_2, \ldots, x_n$$

- The ranked data set (ordered from smallest to largest) is

$$x_{(1)}, x_{(2)}, \ldots, x_{(n)}$$

- The sum of the data is

$$\sum_{i=1}^{n} x_i$$

Mean and Median

Guy: You're the most average girl here

Girl: Wow you're mean

Guy: No, you are

See, math does have real world applications after all 😎

via me.me

Mean

Mean

The mean is the **average** of the data.

$$\text{Mean} = \frac{\text{Sum of data}}{\text{Size of data}}$$

or

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Note that the mean involves **all** of the data.

- The mean of all the properties sold in Newtown is:

```
mean(data$Sold)
```

```
## [1] 1407.143
```

- Focusing specifically on houses with 4 bedrooms (large), the mean is:

```
mean(data$Sold[data$Type=="House" & data$Bedrooms=="4"])
```

```
## [1] 2198.857
```

Mean as a balancing point

The mean is the unique point at which the data is **balanced**. ie. The higher readings and the lower readings all cancel each other out.

For example,

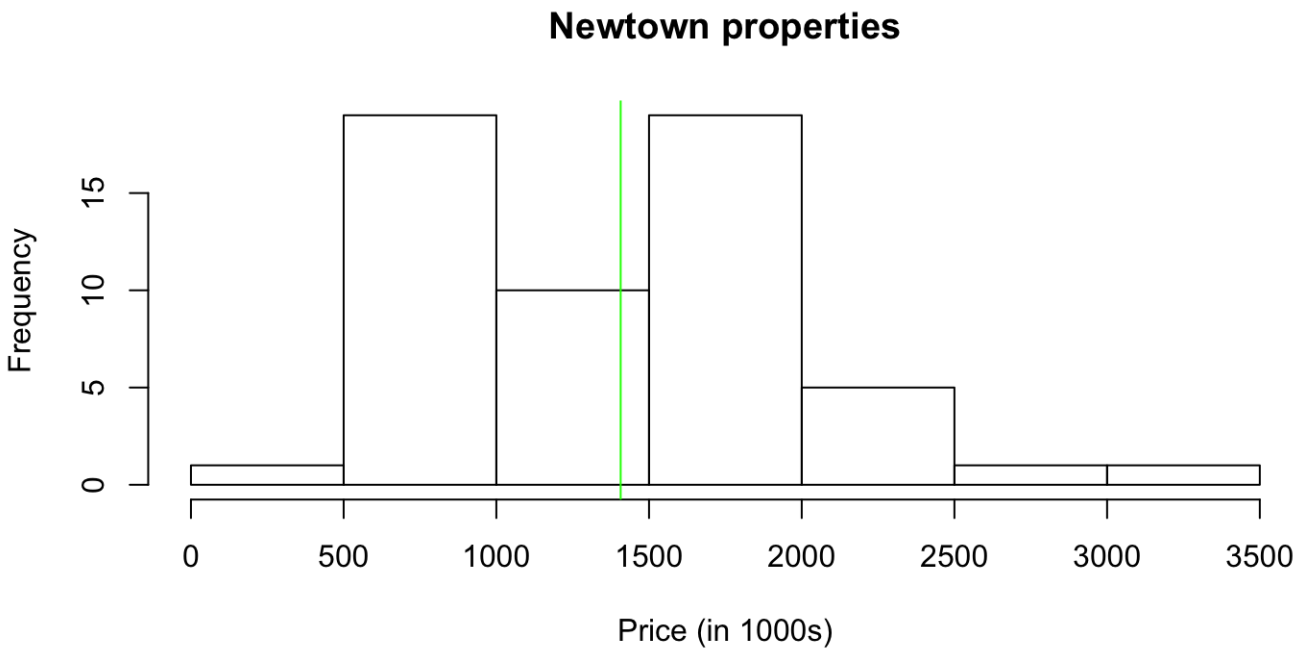
- 19 Watkin St sold for $1950 (thousands).
  - This gives a gap of (1950-1407.143).
  - This is $542.857 (thousands) **above** the mean price.
- 30 Pearl St sold for $1250 (thousands).
  - This gives a gap of (1250-1407.143).
  - This is $157.143 **below** the mean price.

💬 How could you work out what house was the best deal?

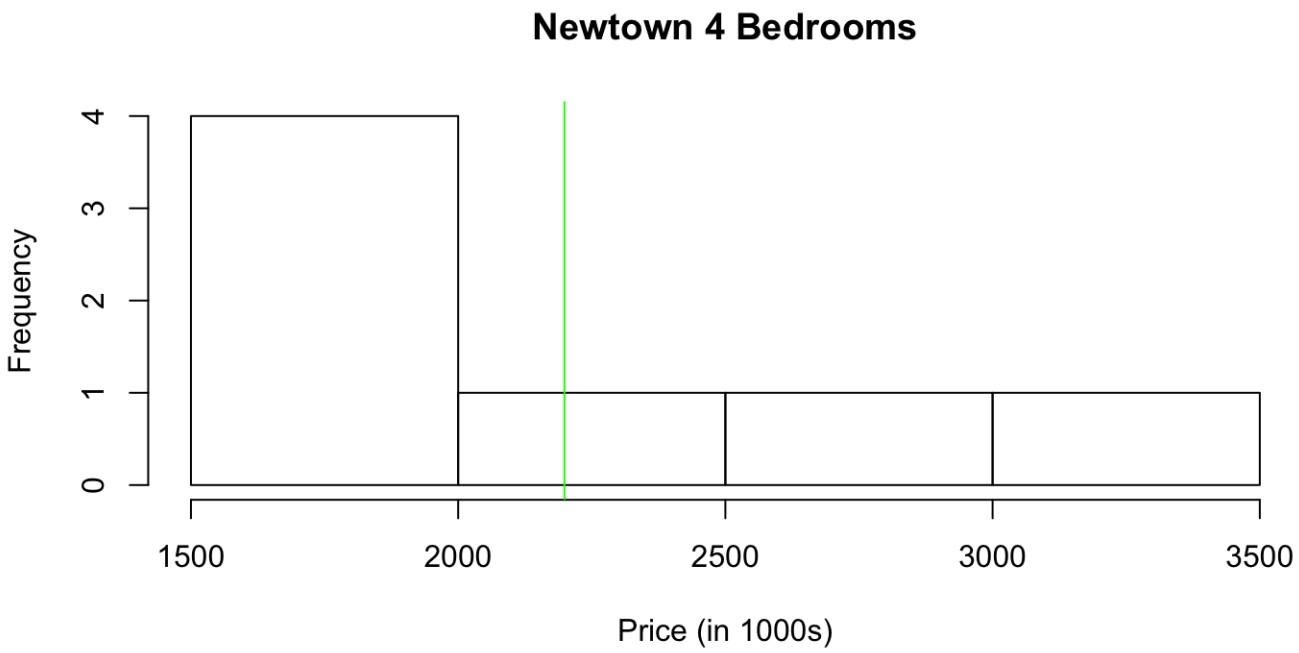
Mean on the histogram

- The mean is the **balancing point** on the histogram, where the expensive and "cheap" properties cancel each other out.

```
hist(data$Sold, main="Newtown properties", xlab="Price (in 1000s)")
abline(v=mean(data$Sold),col="green")
```



```
hist(data$Sold[data$Type=="House" & data$Bedrooms=="4"], main="Newtown 4 Bedrooms", xlab="Price (in 1000s)")
abline(v=mean(data$Sold[data$Type=="House" & data$Bedrooms=="4"]),col="green")
```



Median

Median

The median $\tilde{x}$ is the **middle data point**, when the data is ordered from smallest to largest.

- For an odd sized data set:

$$\text{Median} = \text{the unique middle point} = x_{(\frac{n+1}{2})}$$

- For an even sized data set:

$$\text{Median} = \text{average of the 2 middle points} = \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2}$$

Ordering the data

The ranked data is:

```
sort(data$Sold)
```

```
##  [1]  370  625  645  650  675  692  720  740  740  755  770  780  812  860  861
## [16]  920  935  955  955  999 1100 1240 1250 1280 1309 1315 1370 1375 1400 1460
## [31] 1553 1575 1590 1600 1600 1600 1605 1662 1701 1710 1750 1780 1790 1806 1850
## [46] 1940 1950 1975 2000 2100 2200 2235 2300 2410 2810 3150
```

```
length(data$Sold)
```

```
## [1] 56
```

As the size of the data is 56 (even), the median is found between the 28th and 29th prices, or $\frac{1375+1400}{2} = 1387.5$.

- The median of all the properties sold in Newtown is:

```
median(data$Sold)
```

```
## [1] 1387.5
```

- Focusing specifically on houses with 4 bedrooms (large), the median is:

```
median(data$Sold[data$Type=="House" & data$Bedrooms=="4"])
```

```
## [1] 1975
```

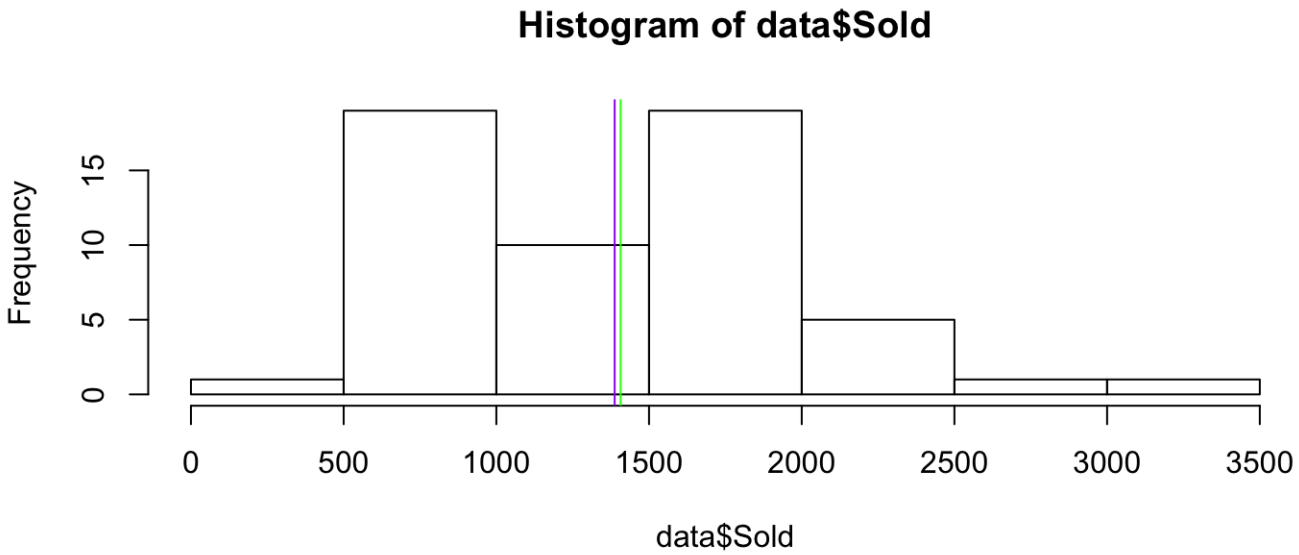
Note:

- The average of the 2 middle points is a convention. We could take the median to be anywhere in between. It is not unique.
- The median only uses 1 or 2 of the data points (after they are ranked).

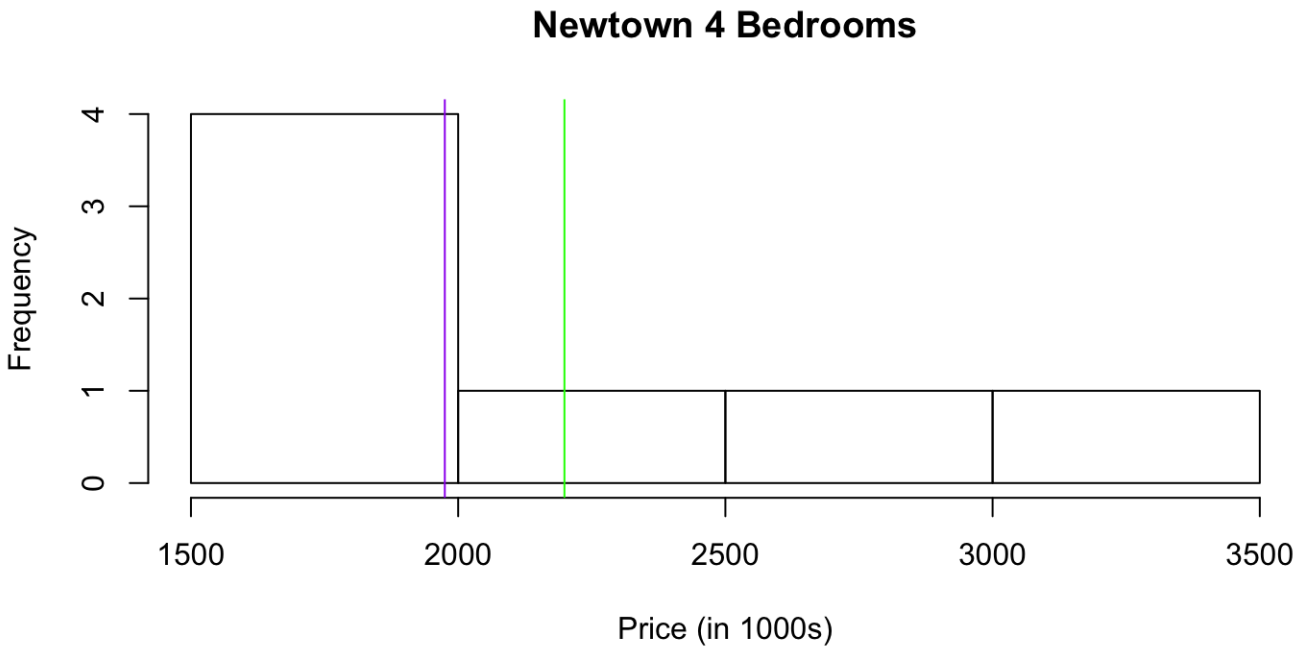
Median on the histogram

- The median is the **half way point** on the histogram - ie 50% of the houses sold are below and above $1.4 million.

```
hist(data$Sold)
abline(v=mean(data$Sold),col="green")
abline(v=median(data$Sold),col="purple")
```



```
hist(data$Sold[data$Type=="House" & data$Bedrooms=="4"], main="Newtown 4 Bedrooms", xlab="Price (in 1000s)")
abline(v=mean(data$Sold[data$Type=="House" & data$Bedrooms=="4"]),col="green")
abline(v=median(data$Sold[data$Type=="House" & data$Bedrooms=="4"]),col="purple")
```



Statistical Thinking

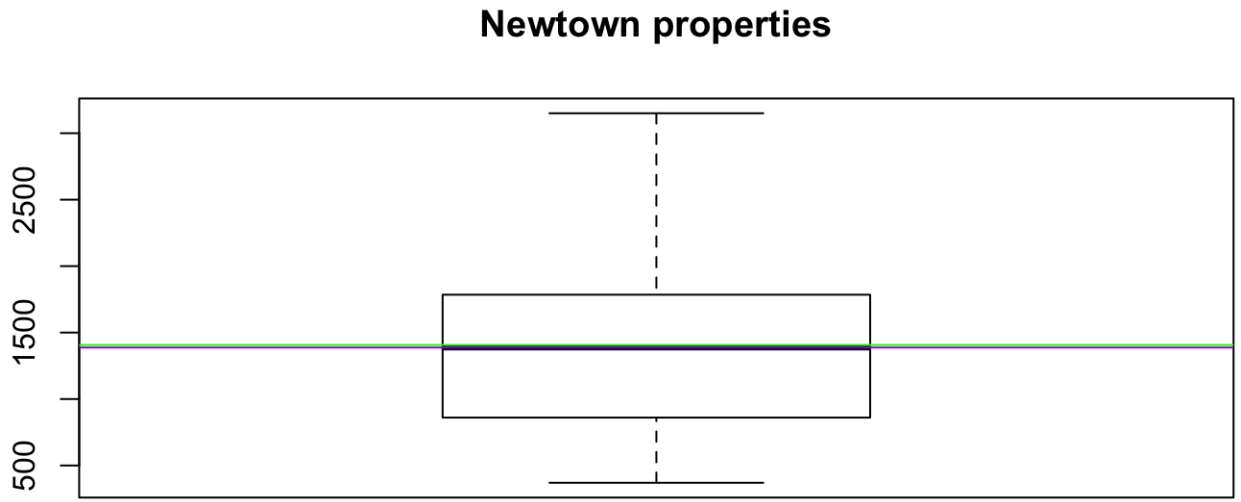
If you had to choose between reporting the mean or median for Newtown properties, which would you choose and why?

- For the full property portfolio, the mean and the median are fairly similar.
- For the 4 bedroom houses, the mean is higher than the median because it is being “pulled up” by some very expensive houses.
- For the average buyer, the median would be more useful as an indication of the sort of price needed to get into the market.
- For any agent selling houses in the area, the mean might be more useful in order to predict his average commissions!
- In practise, we can report both!

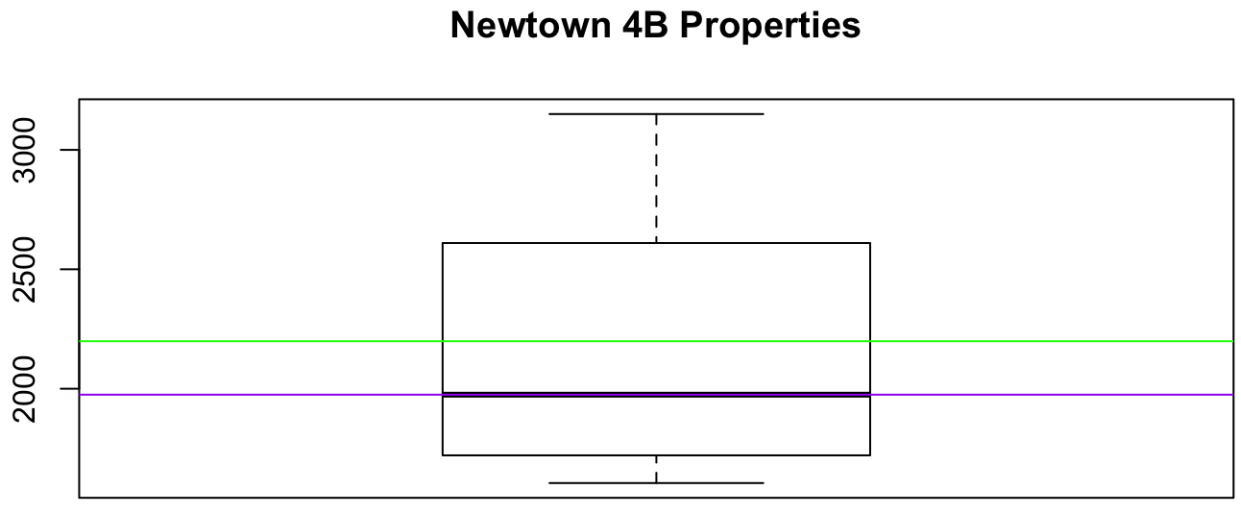
Mean and median on the boxplot

- The median is the centre line on the boxplot.

```
boxplot(data$Sold, main = "Newtown properties")
abline(h=mean(data$Sold),col="green")
abline(h=median(data$Sold),col="purple")
```



```
boxplot(data$Sold[data$Type=="House" & data$Bedrooms=="4"], main = "Newtown 4B Properties")
abline(h=mean(data$Sold[data$Type=="House" & data$Bedrooms=="4"]),col="green")
abline(h=median(data$Sold[data$Type=="House" & data$Bedrooms=="4"]),col="purple")
```



Robustness and Comparisons

Robustness

Robustness

The median is said to be **robust** and is a good summary for skewed data as it is not affected by **outliers**.

Example1

The prices of all the properties follows:

```
sort(data$Sold)
```

```
##  [1]  370  625  645  650  675  692  720  740  740  755  770  780  812  860  861
## [16]  920  935  955  955  999 1100 1240 1250 1280 1309 1315 1370 1375 1400 1460
## [31] 1553 1575 1590 1600 1600 1600 1605 1662 1701 1710 1750 1780 1790 1806 1850
## [46] 1940 1950 1975 2000 2100 2200 2235 2300 2410 2810 3150
```

```
mean(data$Sold)
```

```
## [1] 1407.143
```

```
median(data$Sold)
```

```
## [1] 1387.5
```

Suppose there was a data entry mistake, and the lowest property recorded as 370 was in fact the highest sold at 3700.

 How would the mean and median change?

- The mean would be higher, as have have replaced the smallest reading by now the maximum.
- The median would shift up, from the average of $x_{(28)}$ and $x_{(29)}$ to the average of $x_{(29)}$ and $x_{(30)}$.

```
data1 = c(sort(data$Sold)[2:56],3700)
data1
```

```
##  [1]  625  645  650  675  692  720  740  740  755  770  780  812  860  861  920
## [16]  935  955  955  999 1100 1240 1250 1280 1309 1315 1370 1375 1400 1460 1553
## [31] 1575 1590 1600 1600 1600 1605 1662 1701 1710 1750 1780 1790 1806 1850 1940
## [46] 1950 1975 2000 2100 2200 2235 2300 2410 2810 3150 3700
```

```
mean(data1)
```

```
## [1] 1466.607
```

```
median(data1)
```

```
## [1] 1430
```

Why is the the median commonly used for reporting on Sydney house prices?

FINANCIAL REVIEW

search the AFR

NEWS BUSINESS MARKETS STREET TALK REAL ESTATE OPINION TECHNOLOGY PERSONAL FINANCE LEADERSHIP LIFESTYLE ALL

Home / Real Estate

Apr 21 2016 at 12:00 AM | Updated Apr 21 2016 at 8:31 AM

Save Article | Print | License Article

Sydney median house prices below $1m, Hobart house prices rise

Major changes to negative gearing will make housing investment less attractive. **Louie Douvis**

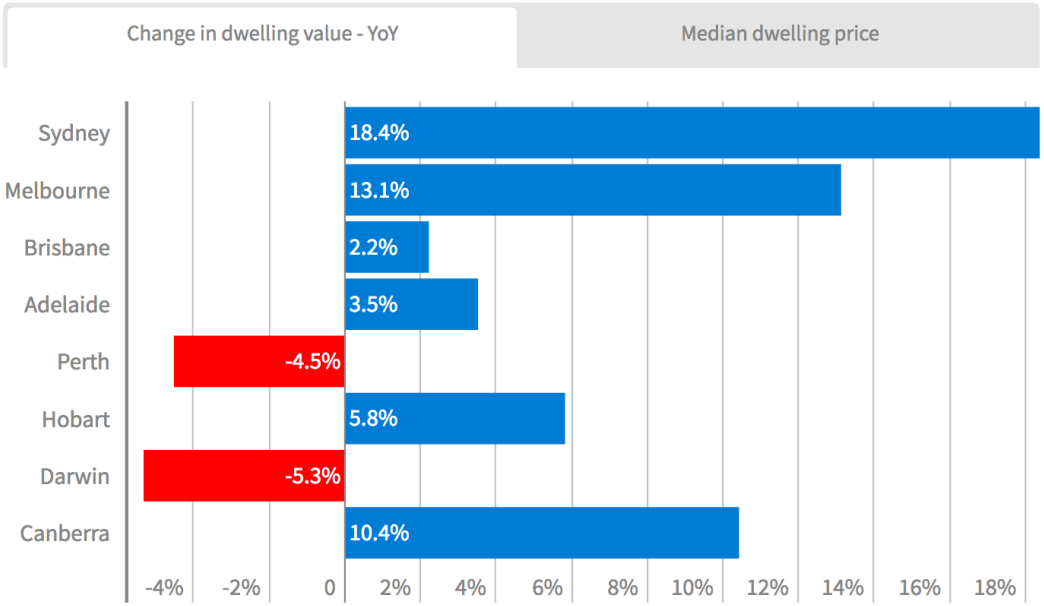
by **Su-Lin Tan** **Michael Bleby**

Sydney median house prices have dropped below $1 million for the first time in a year, offering some relief to first home and new buyers, Domain Group's latest March quarter house price report showed.

Financial Review

East coast capital city dwelling prices continue to rise

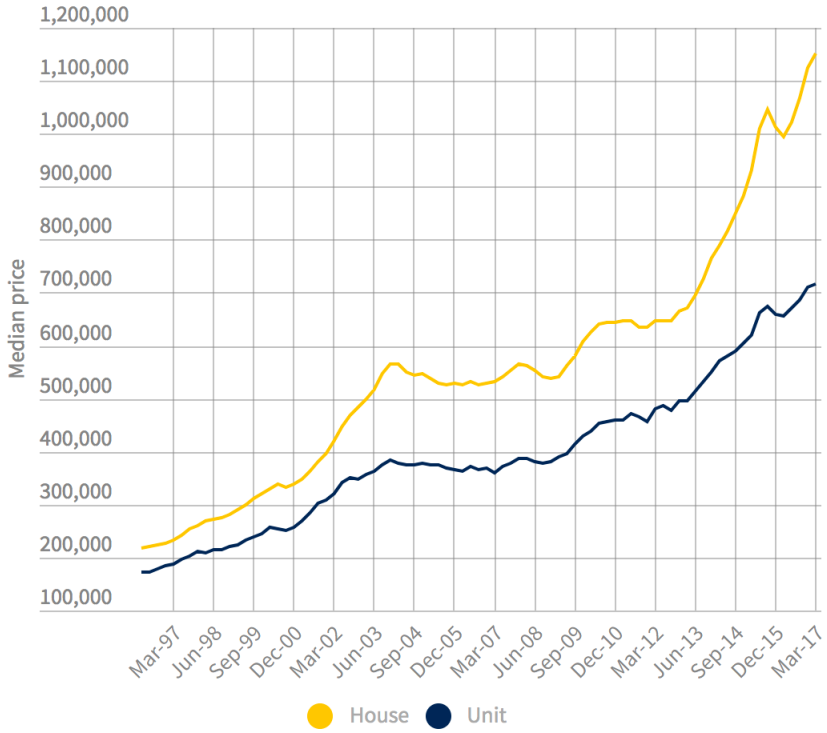
House prices in Sydney have risen by 18 per cent over the past year



Source: Core Logic Hedonic Home Value Index, February 2017. Graphic: Eryk Bagshaw

SMH

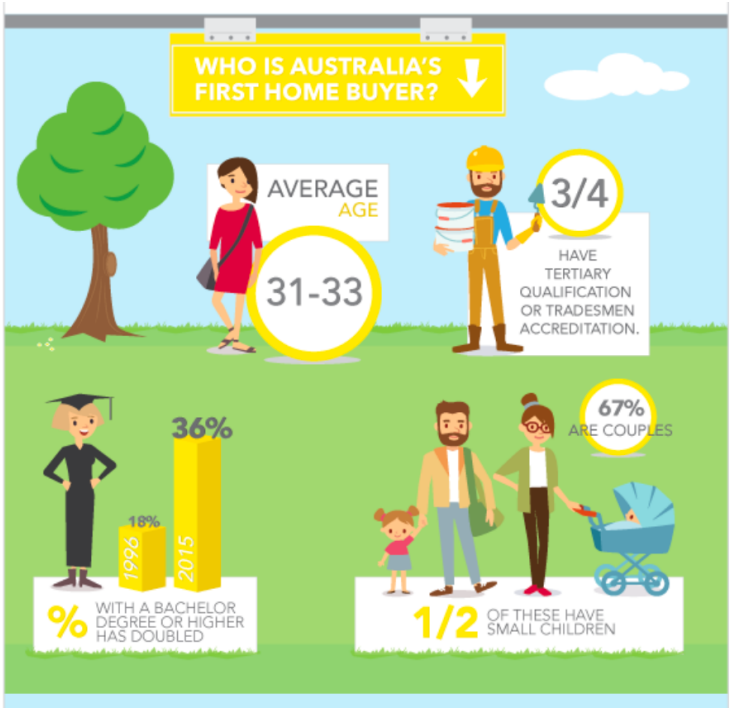
Sydney property price growth over 20 years



Source: Domain Group

Domain

Why is the age of buyer often reported in terms of the mean?

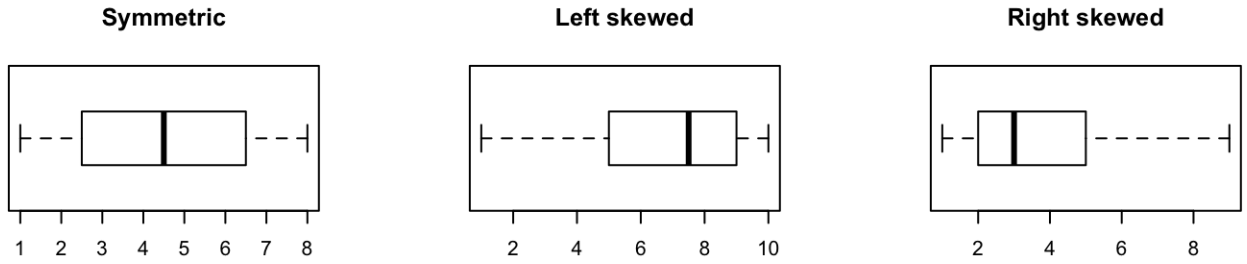


RayWhite

Comparing the mean and the median

The difference between the mean and the median can be an indication of the **shape** of the data.

- For symmetric data, we expect the mean and median to be the same: $\bar{x} = \tilde{x}$.
- For left skewed data, we expect the mean to be smaller than the median: $\bar{x} < \tilde{x}$.
- For right skewed data, we expect the mean to be larger than the median: $\bar{x} > \tilde{x}$.

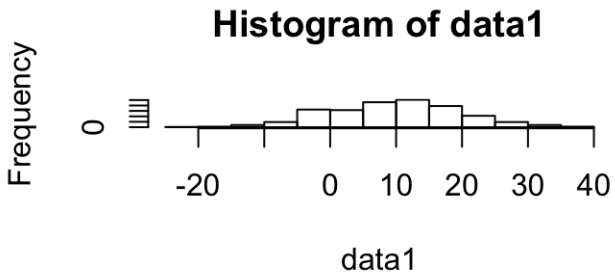


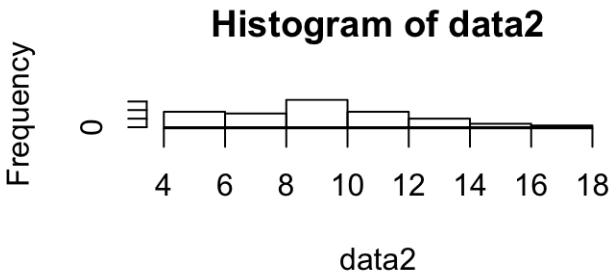
Which is optimal for describing centre?

- Both have strengths and weaknesses depending on the nature of the data.
- Sometimes neither gives a sensible sense of location, for example if the data is bimodal.
- As the median is robust, it is preferable for data which is skewed or has many outliers, like Sydney house prices.
- The mean is helpful for data which is basically symmetric, which not too many outliers, and for theoretical analysis.

Limitations of both?

- Both the mean and median allow very easy comparisons, and are easily understandable.
- However, they need to be paired with a measure of spread.
- Note in the following example, the means are the same, but the data are very different!






```
## [1] 9.700552 9.480000
```

More Examples

Example2

Suppose there was a tiny data entry mistake, and the highest property in Newtown data was in fact 3250 not 3150.

How would the mean and median change?

- The mean would be a tiny bit higher.
- The median would be unaffected.

```
data1 = c(sort(data$Sold)[1:55],3250)
data1
```

```
##  [1]  370  625  645  650  675  692  720  740  740  755  770  780  812  860  861
## [16]  920  935  955  955  999 1100 1240 1250 1280 1309 1315 1370 1375 1400 1460
## [31] 1553 1575 1590 1600 1600 1600 1605 1662 1701 1710 1750 1780 1790 1806 1850
## [46] 1940 1950 1975 2000 2100 2200 2235 2300 2410 2810 3250
```

```
mean(data1)
```

```
## [1] 1408.929
```

```
median(data1)
```

```
## [1] 1387.5
```

Example3

Recently the heritage Properts Building sold for 13 million in Newtown.

How would the mean and median change if it was added to the data?

- The mean would be a lot higher.
- The median would be a bit higher: it moves from the average of the 28th and 29th points to the 29th point.

```
data2 = c(data$Sold,13000)
sort(data2)
```

```
##  [1]   370   625   645   650   675   692   720   740   740   755   770   780
## [13]   812   860   861   920   935   955   955   999  1100  1240  1250  1280
## [25]  1309  1315  1370  1375  1400  1460  1553  1575  1590  1600  1600  1600
## [37]  1605  1662  1701  1710  1750  1780  1790  1806  1850  1940  1950  1975
## [49]  2000  2100  2200  2235  2300  2410  2810  3150 13000
```

```
mean(data2)
```

```
## [1] 1610.526
```

```
median(data2)
```

```
## [1] 1400
```

Summary of changes

Change in data	Mean	Median
Original data	1407.143	1387.5
Lowest changed to highest (x10)	1466.607	1430
Highest property changed from 3150 to 3250	1408.929	1387.5
Extra property of 13000	1610.526	1400

Summary

Both the mean and median summarise the centre of data. The median is robust making it a better choice for skewed data or where there are outliers. Both need to be paired with a measure of spread.

Key Words

numerical summary, maximum, minimum, centre, spread, mean, median, ordered data, balancing point, half way point, robust, symmetric, right-skewed, left-skewed

Further Thinking

🔗 Numerical Summaries