# ETC1010: Data Modelling and Computing

# Lecture 6: Missing data, descriptive statistics

Di Cook (dicook@monash.edu, @visnut)

Week 6



Processing math: 100%

# Overview

- naniar
- data set overviews
- which summary to use

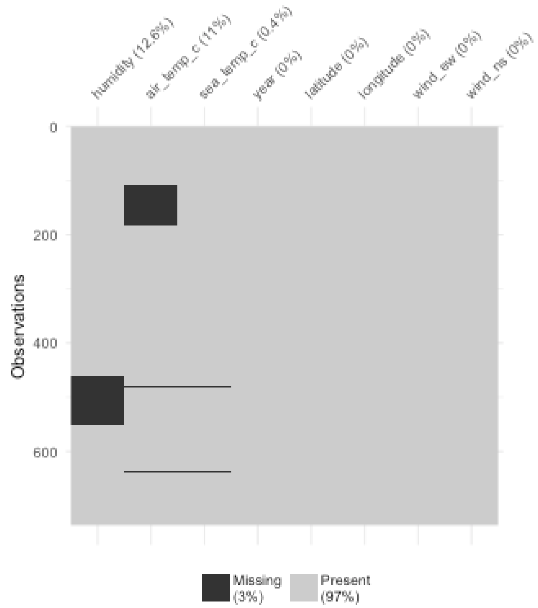
# Exploring missings

West Pacific Tropical Atmosphere Ocean Data, 1993 & 1997, for improved detection, understanding and prediction of El Nino and La Nina, collected from http://www.pmel.noaa.gov/tao/index.shtml

```
Observations: 736
Variables: 8
$ year       <fctr> 1997, 1997, 1997, 1997, 1997, 1997, 1997, 1997, 19...
$ latitude   <fctr> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
$ longitude  <fctr> -110, -110, -110, -110, -110, -110, -110, -110, -1...
$ sea_temp_c <dbl> 27.59, 27.55, 27.57, 27.62, 27.65, 27.83, 28.01, 28...
$ air_temp_c <dbl> 27.15, 27.02, 27.00, 26.93, 26.84, 26.94, 27.04, 27...
$ humidity   <dbl> 79.6, 75.8, 76.5, 76.2, 76.4, 76.7, 76.5, 78.3, 78....
$ wind_ew    <dbl> -6.4, -5.3, -5.1, -4.9, -3.5, -4.4, -2.0, -3.7, -4....
$ wind_ns    <dbl> 5.4, 5.3, 4.5, 2.5, 4.1, 1.6, 3.5, 4.5, 5.0, 3.5, 2...
```

# Missingness map

Heatmap display showing where missing values are in the data table.






Processing math: 100%

# Numerical summaries

Proportion of observations missing:

```
[1] 0.03006114
```

Proportion of variables missing:

```
[1] 0.375
```

How many observations have $k$ missings?

```
[[1]]
# A tibble: 4 x 3
  n_missing_in_case n_cases    percent
              <int>   <int>      <dbl>
1                 0     565 76.7663043
2                 1     167 22.6902174
3                 2       2  0.2717391
4                 3       2  0.2717391
```

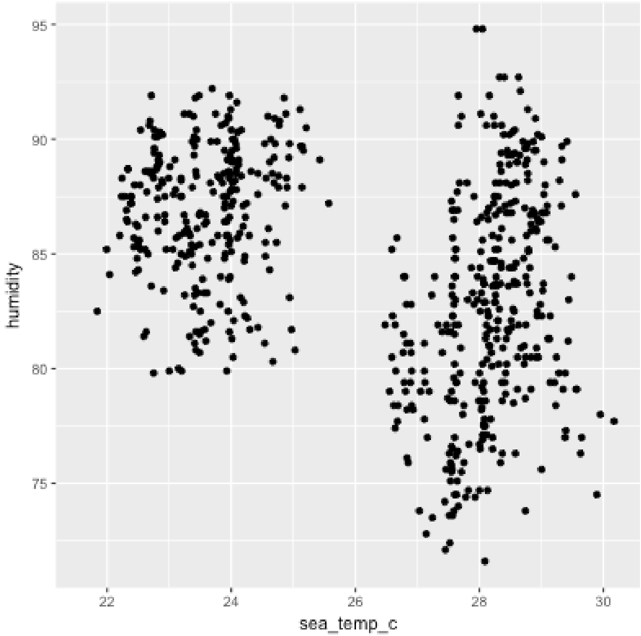
# By group

```
[[1]]
# A tibble: 6 x 4
    year n_missing_in_case n_cases    percent
  <fctr>             <int>   <int>      <dbl>
1   1997                 0     291 79.0760870
2   1997                 1      77 20.9239130
3   1993                 0     274 74.4565217
4   1993                 1      90 24.4565217
5   1993                 2       2  0.5434783
6   1993                 3       2  0.5434783
```

# Missings shouldn't be ignored

but most software will simply drop them!

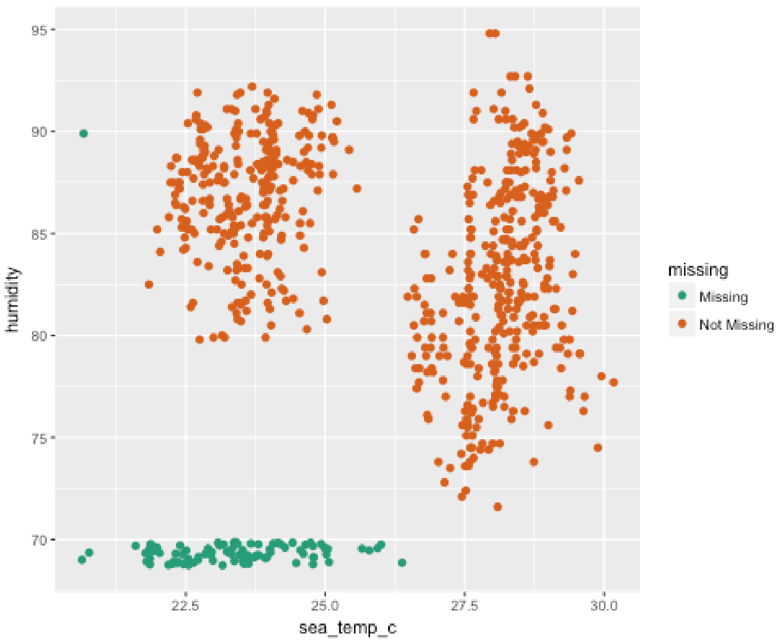
```
Warning: Removed 94 rows containing missing values (geom_point).
```



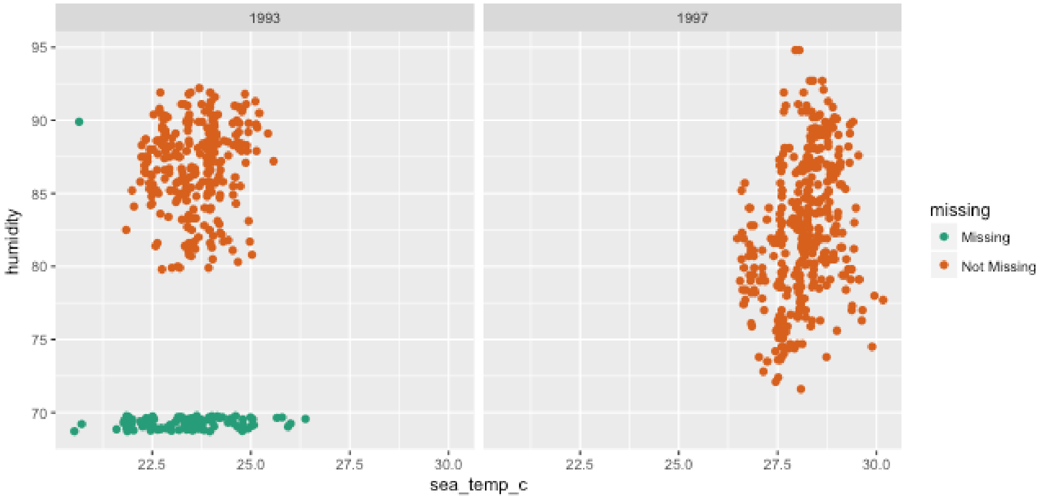



Processing math: 100%

# Keep them in the plot



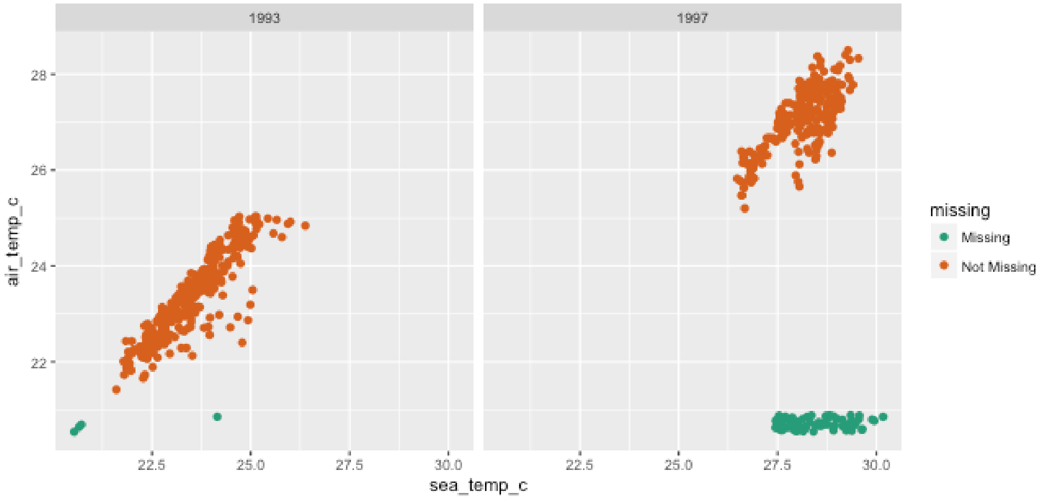
# by year






Processing math: 100%

# Understanding missing dependencies

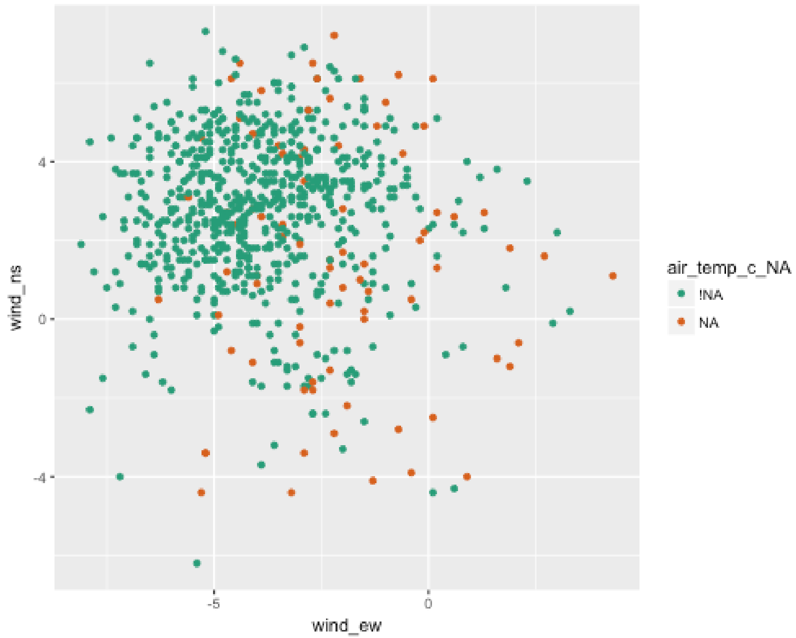



Year needs to be accounted for in finding good substu=itute values.



Processing math: 100%

# Relationship with other variables






Processing math: 100%

# Handling missings

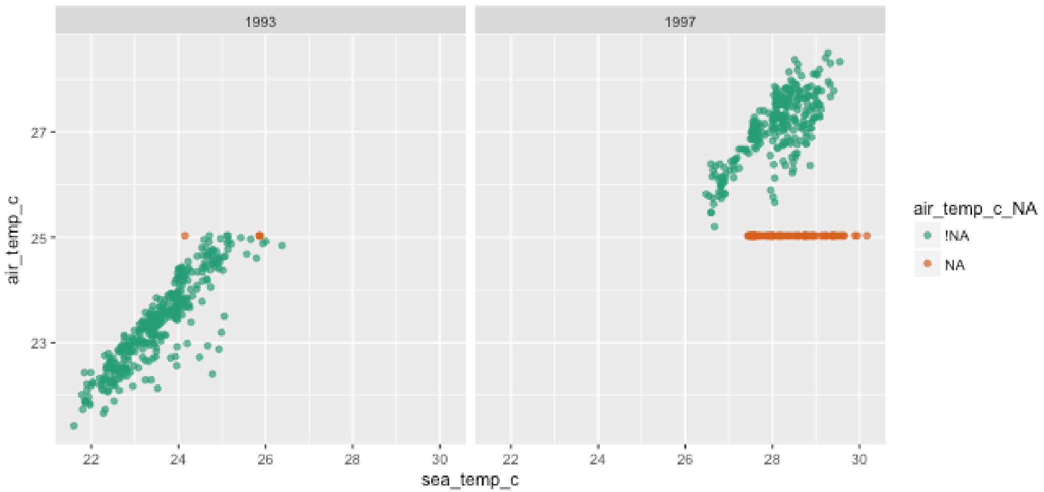
- An small fraction of cases have several missings, drop the cases
- A variable or two, out of many, have a lot of missings, drop the variables
- If missings are small in number, but located in many cases and variables, you need to impute these values, to do most analyses
- Designing the imputation should take into account dependencies that you have seen between missingness and existing variables.
- For the ocean buoys data this means imputation needs to be done separately by year

# Common ways to impute values

- Simple parametric: use the mean or median of the complete cases for each variable
- Simple non-parametric: find the $k$ nearest neighbours with a complete value and average these
- Multiple imputation: Use a statistical distribution, e.g. normal model and simulate a value (or set of values, hot deck imputation) for the missings

# Examples - using the mean

and ignoring year.

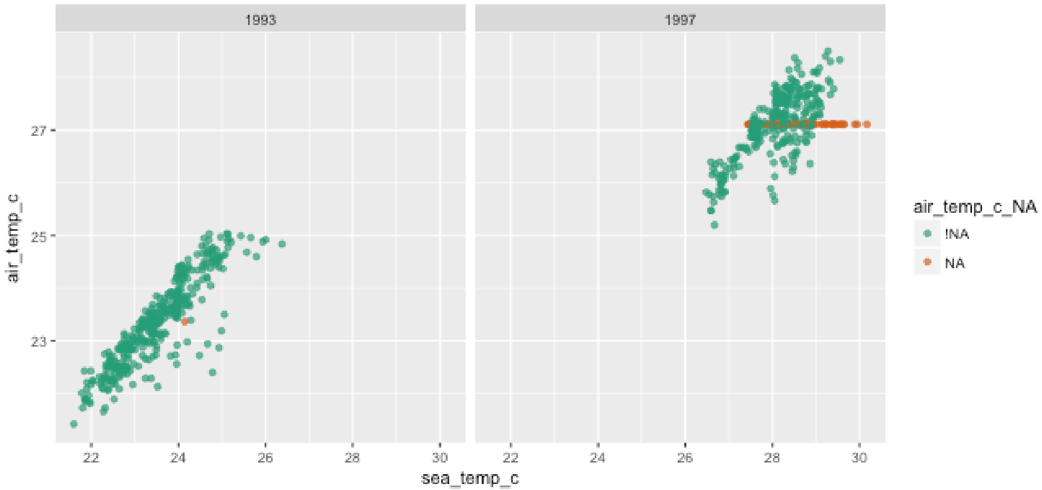



POOR MATCH!



Processing math: 100%

# by year

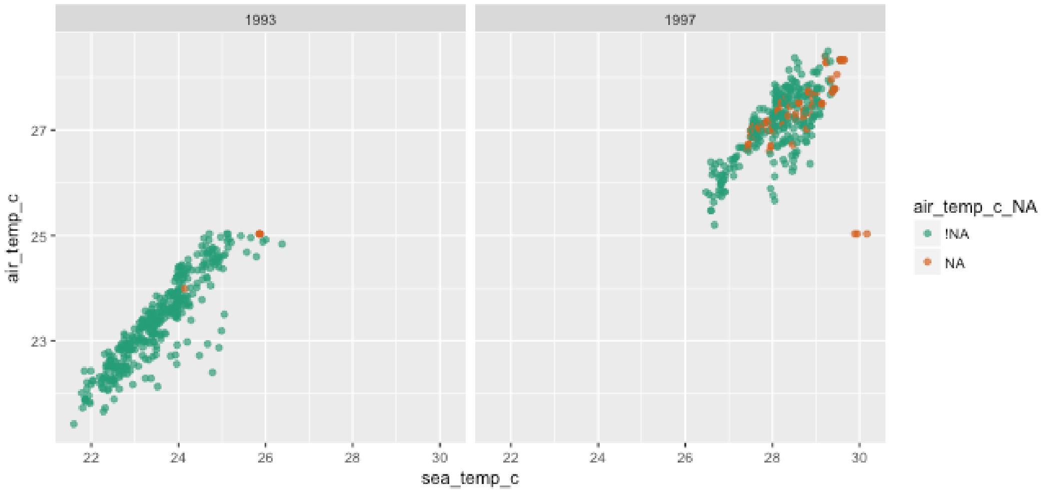



Better, but still a bit weird!



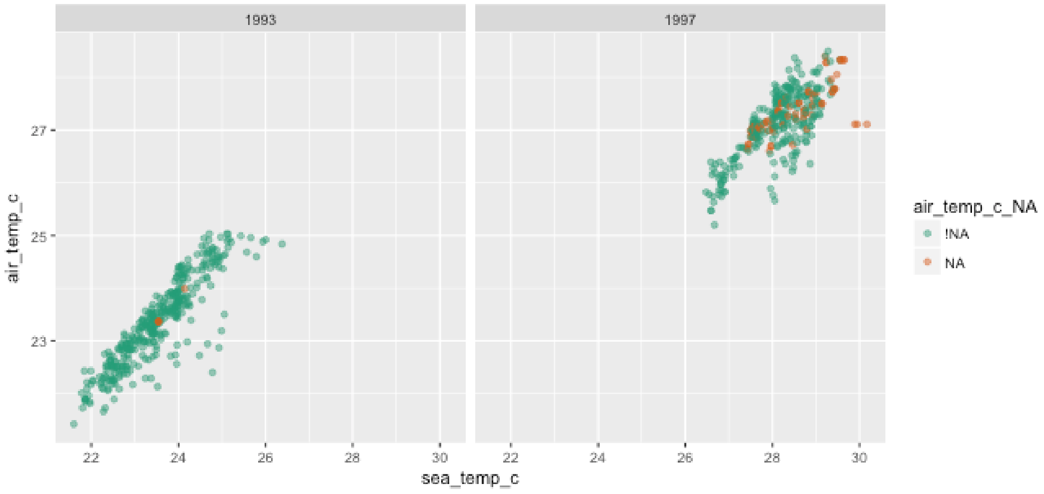
Processing math: 100%

# Nearest neighbors imputation




A LITTLE BETTER!

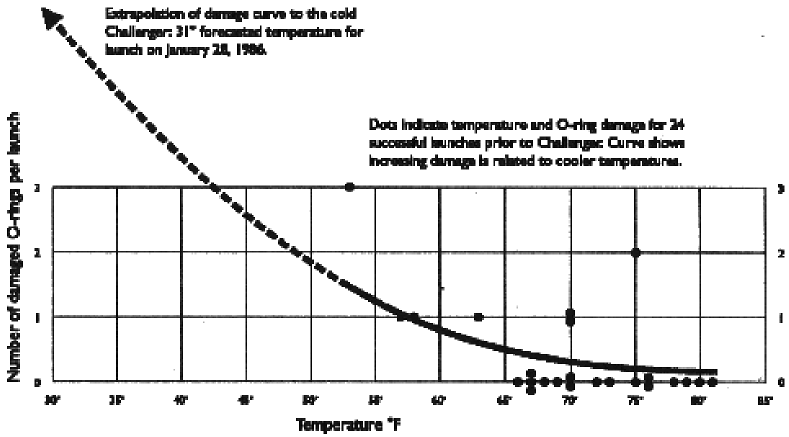
# by year




MUCH BETTER!

# Famous example of ignoring missings

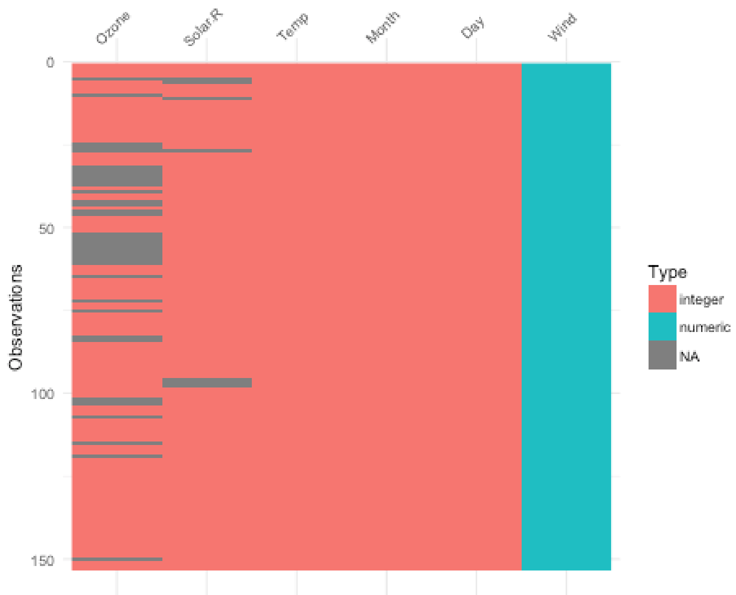
- Subsequent investigation determined that the cause was failure of the O-ring seals used to isolate the fuel supply from burning gases.
- NASA staff ignored observations where no O-rings failed.




http://www.asktog.com/books/challengerExerpt.html

# Data summary

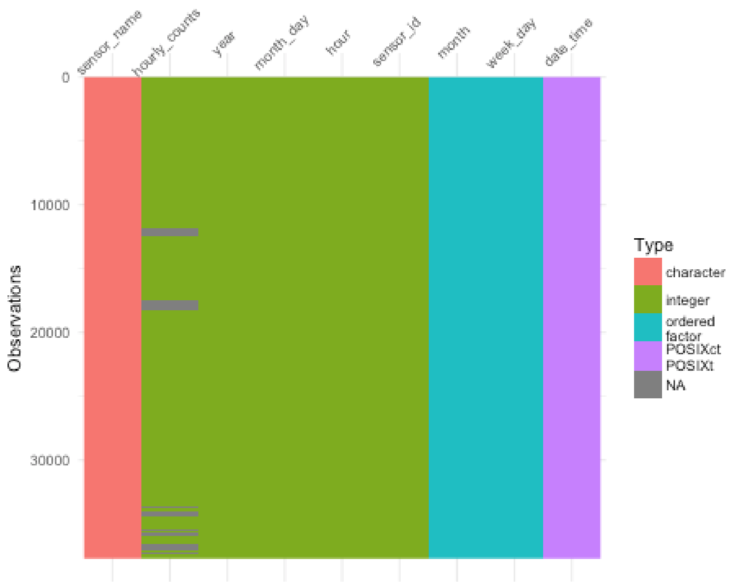
In general, there is a nice way to get a quick overview of your data






Processing math: 100%

```
Observations: 37,700
Variables: 9
$ hourly_counts <int> 883, 597, 294, 183, 118, 68, 47, 52, 120, 333, 7...
$ date_time     <dttm> 2016-01-01 00:00:00, 2016-01-01 01:00:00, 2016-...
$ year          <int> 2016, 2016, 2016, 2016, 2016, 2016, 2016, 2016, ...
$ month         <ord> January, January, January, January, January, Jan...
$ month_day     <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
$ week_day      <ord> Friday, Friday, Friday, Friday, Friday, Friday, ...
$ hour          <int> 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14...
$ sensor_id     <int> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, ...
$ sensor_name   <chr> "Bourke Street Mall (South)", "Bourke Street Mal...
```


sensor_name
hourly_counts
year
month_day
hour
sensor_id
month
week_day
date_time
0
10000
20000
30000
Observations
Type
character
integer
ordered factor
POSIXct POSIXt
NA

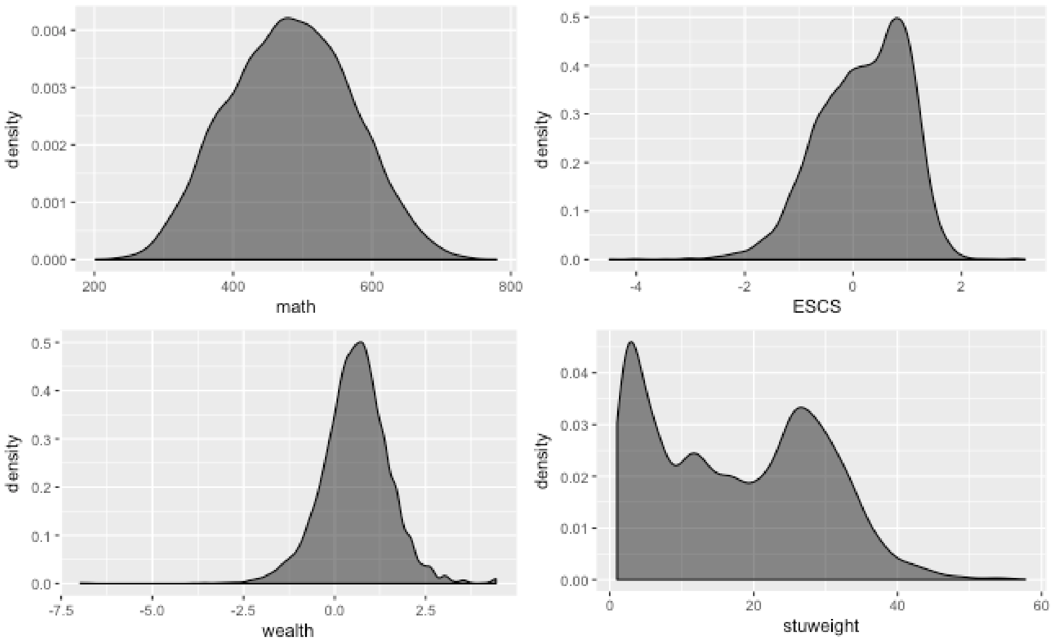
# Data descriptions

We have seen a lot of descriptive statistics thus far. Here is a summary of good practice:

- Depending on the variable type some summaries are appropriate and others are not
- For quantitative variables, you need to examine the distribution to determine to use *mean/sd* or *median/IQR* statistical summaries
- For categorical variables, summarise using counts and proportions
- For two quantitative variables, if the distributions are both symmetric and unimodal, correlation is a good numerical statistic
- Two categorical variables are typically summarised using a contingency table, which has counts, and several different proportion calculations
- Mix of a categorical and a quantitative variable, numerical summary by category!

# Which statistics?

In the PISA data, looking at some of the demographics index variables, would you use *mean/sd* or *median/IQR* to summarise these quantitative variables?

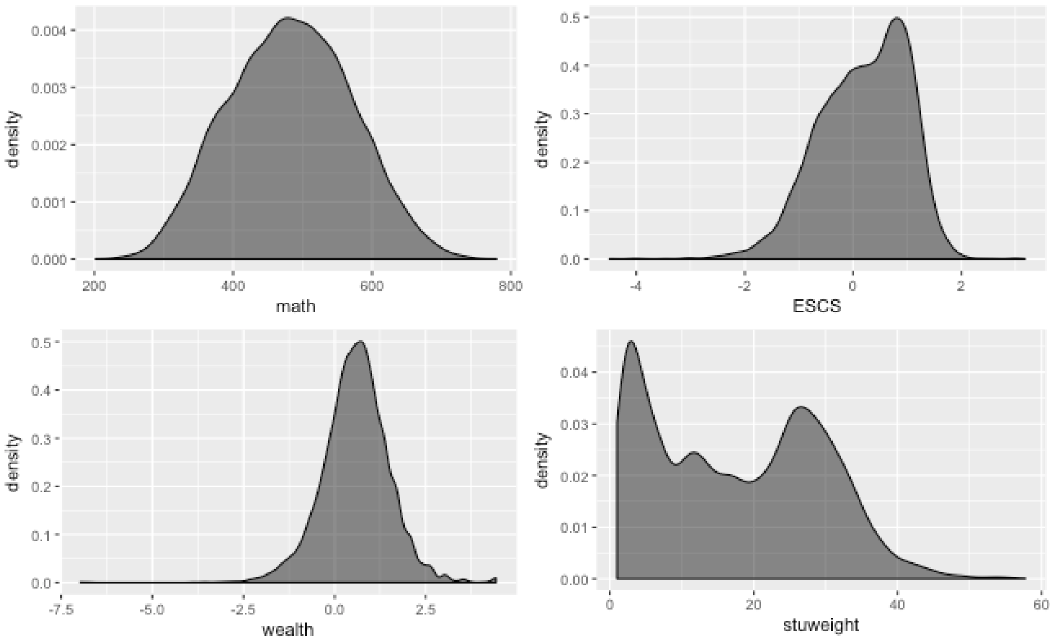





Processing math: 100%

# Which statistics?

In the PISA data, looking at some of the demographics index variables, would you use *mean/sd* or *median/IQR* to summarise these quantitative variables?




*mean/sd*, *median/IQR*, *mean/sd* and point out the long tail of low values,



Processing math: 100%

# Categorical variable

```
# A tibble: 2 x 3
  gender     n       p
  <fctr> <int>   <dbl>
1 female  7163 0.49298
2   male  7367 0.50702
```



Processing math: 100%

# How many digits should you use?

- Recommendation (Chatfield, 1991 The Practice of Statistics): Two-three variable digits
- Gender proportions: 0.49298 round to 0.49, 0.50702 round to 0.51
- or 0.49298 round to 0.493, 0.50702 round to 0.507

# Contingency tables

- Two categorical variables, count the unique combinations
- Add the marginal counts
- Add proportions by dividing by (1) overall count, (2) row marginal count, (3) column marginal count

e.g. Gender by TVs in the household

```
           1     2     3     4   Sum
female    92  1153  3044  2547  6836
male     114  1106  3025  2689  6934
Sum      206  2259  6069  5236 13770
```

# Proportions

```
           1     2     3     4   Sum
female    92  1153  3044  2547  6836
male     114  1106  3025  2689  6934
Sum      206  2259  6069  5236 13770
```

Overall:

```
             1       2       3       4     Sum
female 0.00633 0.07935 0.20950 0.17529 0.47047
male   0.00785 0.07612 0.20819 0.18507 0.47722
Sum    0.01418 0.15547 0.41769 0.36036 0.94769
```

By row:

```
            1     2     3     4   Sum
female     92  1153  3044  2547  6836
male      114  1106  3025  2689  6934
Sum       206  2259  6069  5236 13770
```

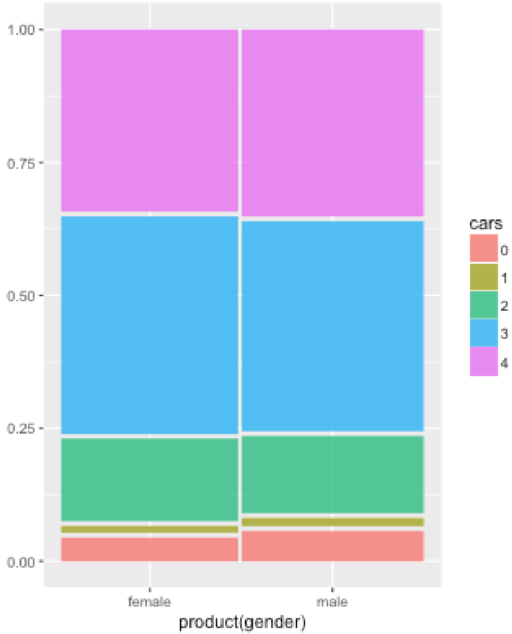
```
             1       2       3       4     Sum
female 0.00673 0.08433 0.22264 0.18629 0.50000
male   0.00822 0.07975 0.21813 0.19390 0.50000
Sum    0.00748 0.08203 0.22037 0.19012 0.50000
```

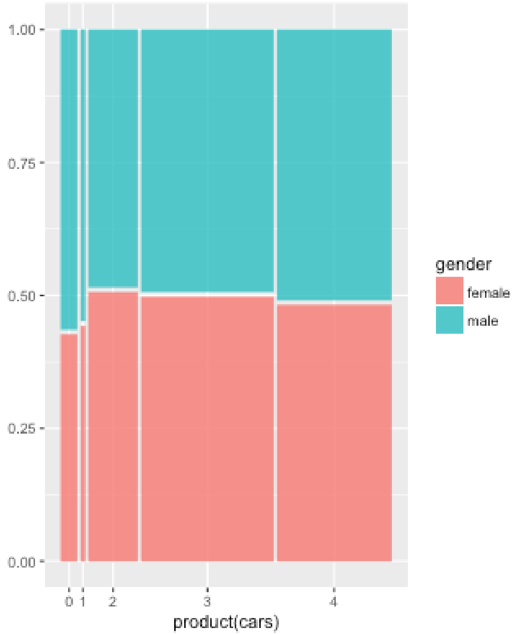
By column:

```
            1     2     3     4   Sum
female     92  1153  3044  2547  6836
male      114  1106  3025  2689  6934
Sum       206  2259  6069  5236 13770
```

```
            1     2     3     4   Sum
female 0.223 0.255 0.251 0.243 0.248
male   0.277 0.245 0.249 0.257 0.252
Sum    0.500 0.500 0.500 0.500 0.500
```

# Contingency tables and mosaic plots






Processing math: 100%

# Share and share alike





Processing math: 100%