Lecture 1

Filtering data sets

Filtering data using dplyr filter()

- Used to sub-set rows/observations based on conditions being met as TRUE (e.g. height > 170)

gender	age	weight	height
M	18	52	165
F	67	65	171
M	40	80	183
F	23	57	154
M	26	71	173
M	34	90	167
M	55	67	169
Not disclosed	42	78	180
M	18	85	190
F	73	50	145
M	18	85	190

filter()

gender	age	weight	height
F	67	65	171
M	40	80	183
M	26	71	173
Not disclosed	42	78	180
M	18	85	190
M	18	85	190

```
install.packages("tidyverse")

library(tidyverse)
```

Filtering data using dplyr filter()

- Important arguments in the function are (i) data source and (ii) a condition whereby if TRUE the data is kept

```
filter(data source, condition, ...)
```

- For example, using the health data set, we can filter height > 170

```
filter(health, height>170)
```

- Now lets use R to filter the 'health' dataset against for the following examples:

1) Filter height > 170 and gender is M (tip "&" is the R operator for and)

2) Filter height > 170 or gender is M (tip "|" is the R operator for or)

3) Filter height > 170, gender is M and age is either 18 or 40 (tip %in% is a useful operator to deal with multiple 'or' requests within the same variable)

Filtering data using dplyr filter()

```
#Answers to example questions for filter
filter(health, height > 170)


filter(health, height > 170 & gender == "M")


filter(health, height > 170 | gender == "M")


filter(health, height > 170 & gender == "M" & age %in% c(18,40))
```

Filtering data using dplyr filter()

- Mini quiz using the iris data set

1) Filter iris for petal length less than 1.6

2) Filter iris for sepal length greater than 5.0 and sepal width greater than 3.0

3) Filter iris for petal width being either 0.2, 0.3 or 1.4, and species is virginica