Assignment 4. ST661 2018
Catherine Hurley
Due on Wednesday December 5 6pm.

You should complete this assignment using Rmarkdown. Place the printed html file in in the box labelled ST661/ST663 in the ground floor of Logic house (under stairs).
Also upload the html file to Moodle.
On your own computer you will need to do (once only). This installs ggplot, dplyr and more.

```
install.packages("tidyverse")
library(tidyverse) # every time
```

The above packages are pre-installed on the Logic House computers and on our Rstudio server.

1. This question involves a dataset which gives a record of every men's singles match played in Wimbledon in 2015. Type in the following to load the data called `wim`. (File available on Moodle).

   ```
   load("h4data")
   ```

   (a) Use `mutate` (from dplyr) to add a new variable which is the difference in heights of the winner and loser. Use `ggplot` to draw a histogram of the new variable.

   (b) Construct a dataset `by_player` with one row per player, recording also the number of wins, ranking points, height and country of the player.
   Do this in the following steps:

   i. Construct a dataset `w_wim`, containing for each match the intormation `name= winner_name`, `points=winner_rank_points`, `ht= winner_ht`, `ioc= winner_ioc`, and a new variable `wins` with a value of 1. Use `select` and `mutate`.

   ii. Construct a dataset `l_wim`, containing for each match the intormation `name= loser_name`, `points=loser_rank_points`, `ht= loser_ht`, `ioc= loser_ioc`, and a new variable `wins` with a value of 0.

   iii. Use `rbind` to stick `w_wim` and `l_wim` together. Call the result `wl_wim`

   iv. Use `group_by` on `wl_wim`, to group it by `name`. and `summarise` the result of the previous step,
   with `wins=sum(wins), points=points[1], ht=ht[1], ioc=ioc[1`.

   (c) Calculate the average height for all players in the tournament. Use ggplot to plot player points versus number of wins.

   (d) Using the dataset `by_player`, write code to find the names of the tournament winner and the losing finalist. If you did not manage to correctly construct the dataset `by_player`, do this some other way.

   (e) Calculate the number of wins per country. How many matches were won by Spanish (ESP) players?

   (f) Draw a barplot showing the number of wins for the top 10 countries, preferrably in decreasing order by wins.

2. For the University ranking data from the midterm, first convert appropriate variables to numeric. Then answer these questions using dplyr tools and ggplot, in all cases except part (b).

   (a) Summarise the Location (countries) of the universities with counts. Using ggplot, draw a barplot of these location counts.

   (b) Redraw the barplot, with bars in decreasing order. Use fct_reorder.

   (c) Write a function called `top5` that counts the number of values in 1, 2, ...5 in a numeric vector, ignoring NAs. Use `%in%` .
   Check that the function gives the correct answer on vectors below

   ```
   x1 <- c(NA,10:1)
   x2 <- rnorm(8)
   x3 <- c(6,2,8,9,-1,0,4)
   ```

(d) Use `top5` to calculate for each university, how many of the categories Quality.of.Education, Publications, Influence, Citations, Broad.Impact accorded them a top 5 ranking. Add this information to your dataset. Use base R for your answer.

(e) Here is a dplyr answer to the previous question. Explain how it works. Write code which confirms the columns tc and topcount are identical. Here the base R solution is simpler, in my opinion. Can anyone supply a simpler and better dplyr-based answer?

```
d1 <- d %>% mutate_at(sel, `<=`, 5) %>%
        select(one_of(sel))%>%
        mutate(tc = rowSums(., na.rm=T))
```

(f) Construct a dataset containing the names of Universities who have at least one top 5 ranking.

(g) Add a new variable Continent to `d`, that is NorthAmerica for Universities in USA and Canada, Asia for Japan and South Korea, and Europe otherwise. Suggestion: use ifelse within mutate.

(h) Plot the Citations versus the Broad.Impact rankings,using different colours to distinguish the continents.

3. (Optional, a little advanced). On this webpage `https://www.aggdata.com/awards/oscar` you will find a complete list of Academy Award Nominees and Winners from 1927-2010. Download the data and read it in to R. Compute the number of nominations for each actress in the leading or supporting role categories, the number of wins along with the first and last year they were nominated. Use this to find the name of the actress with the biggest gap between the first and last nomination.