Assignment 3. ST662 2023
Catherine Hurley
Due Wednesday November 22, 6pm

- Do all questions. Hand in questions 1-3 only.
- This time, no skeleton Rmarkdown file is provided. Make your own. You should knit the file to html and upload the html file to Moodle. If the uploaded file is not html, 5 marks (out of maximum of 20) will be deducted.
- The upload must be completed by time and date given above or the assignment will not be accepted.
- Help will be available during the tutorials on Friday November 17.

Install necessary packages, not needed on server. Do this once only.

```
install.packages("tidyverse")
install.packages("GGally")
install.packages("lubridate")
install.packages("mosaicData")
```

Put this code block at the start of your Rmd, to load the necessary packages.

```
library(tidyverse)
library(GGally)
library(MASS)
library(mosaicData)
```

1. Type in

   `head(Pima.tr)`

   Using the data set `Pima.tr` and ggplot2 for all plots:

   (a) Make a scatterplot plot of `bp` versus `npreg`.

   (b) Using the function `cut_interval`, construct a factor version of `npreg` with n=4 levels. Call this new variable `npregf`. Add this variable to dataset `Pima.tr`.

   (c) Plot boxplots of `bp` for each `npregf` level.

   (d) Make a scatterplot of `glu` versus `age`. Use colour to show variable `type` and add smooths for the two groups.

   (e) Redo the previous plot, separating out the two types. (Colour is not now needed)

2. For the dataset Gestation in package mosaicData, make plots which explore each of the following:

   (a) the association between a mother's and father's age.

   (b) the association between mother's race and smoking status From your graph, can you say which race has the highest proportion of never smokers?

   (c) the association of between a mother's age and race

   (d) how the association of babies weight and mother's race differs with mother's smoking status. Show the two groups for smoking status never and now only.

3. The data Hep2012.csv contains data with the points athletes received from each event in the Heptathlon in the 2012 Olympics. Access the data with

   ```
   hep <- read_csv("Hep2012.csv")
   ```

   (a) Calculate the points total for the athletes and add it as an extra variable to the data frame.

(b) Obtain a vector of the names of the athletes with the best five results overall. Hint: use `order` (Base R) or dplyr tools.

(c) Use ggduo in package GGally to plot points total versus the points in the three events Long Jump, 100m Hurdles and 800m.

(d) Make a scatterplot matrix of the event scores for the heptathlon data using ggpairs. Write code to find the name of the athletes which got zero points in some events, and the name of the athlete with an unusual shotput value. (Variables that do not begin with a letter or have spaces needed to be referred to in backquotes eg ‘200m‘.)

(e) Draw a parallel coordinate plot of the heptathlon scores, with the events in order of time 100m Hurdles,High Jump, Shot Put,200m , Long Jump,Javelin, 800m . Explain why ‘scale=”globalminmax”‘ is the appropriate choice. In this plot, can you see athletes with zero points in any event? Does this make sense vis-a-vis your findings from the previous part? As all the variables are point scores for performance, there is no need to rescale. There is joint one xero point score evident in the 800m. The other athletes with zero scores in earlier events (200m and long jump) dropped out and had NAs so are missing from this plot. The documentation for ggparcoord sats the default behaviour is to exclude cases with NA.

(f) Use the code below to make a group variable which has one level for each of the top three athletes, and a fourth level for the others. Use this to redraw the PCP with four different colours. Hint: use scale_color_manual to specify colours. Also arrange the rows in increasing order of total points so the top performers are drawn last.

```
hep <- hep |> arrange(desc(Tot))
hep$winners <- paste(1:nrow(hep), hep$Athlete, sep=" ")
hep$winners[-(1:3)] <- "Others"
```

4. The dataset BathingWaterQuality2013.csv from `https://data.smartdublin.ie/dataset/tableview/05ef6a8b-71cc-489b-9e59-57b78a9176ab` gives details of Bathing Water Quality monitoring at Fingal’s beaches. The European Union sets standards for the quality of bathing water in member states. Fingal County Council monitors the water at the beaches from mid-May to September every year. Information is made available to the public on the total number of 2 different bacteria i.e. Escherichia Coli and Enterococci, present in the water. These bacteria are not visible and so water samples are taken and tested in a laboratory. The number of bacteria in a 100ml sample of seawater is recorded. To comply with guidelines, E-coli counts must be less that 100 per 100ml for 80% of samples, and Enterococci must be less than 100 per 100ml for 90% of all samples.

```
water <- read_csv("BathingWaterQuality2013.csv")
water <- mutate(water,Date= lubridate::dmy(Date))
water4 <-
  filter(water, Beach %in% c("Sutton", "Portmarnock" ,"Skerries" ,"Howth - Claremont"))
```

(a) Using the four beaches in water4, plot Enterococci versus date, color by Beach. Use geom_point and geom_line(). Use geom_hline(yintercept=100, color=’magenta’) to show the guideline value. Which beach has the highest readings? Do any of the beaches exceed the guidelines? If so which beach and when?

(b) Repeat a. using EColi.

(c) Using the full water dataset use facet_wrap to show Enterocci time plots separately for all the beaches. Show the the guideline value.

(d) Show the Enterococci and EColi time plots overlaid for all the beaches. Use colour to distinguish the two y variables.