Preparation.

In this workshop, we will need to install and use the following packages:

- tidyverse
- emmeans
- car
- mice

Run the following codes to install these packages (Skip this if you have already installed the packages)

```
install.packages("tidyverse")
install.packages("emmeans")
install.packages("car")
install.packages("mice")
```

Run the following codes to load the libraries.

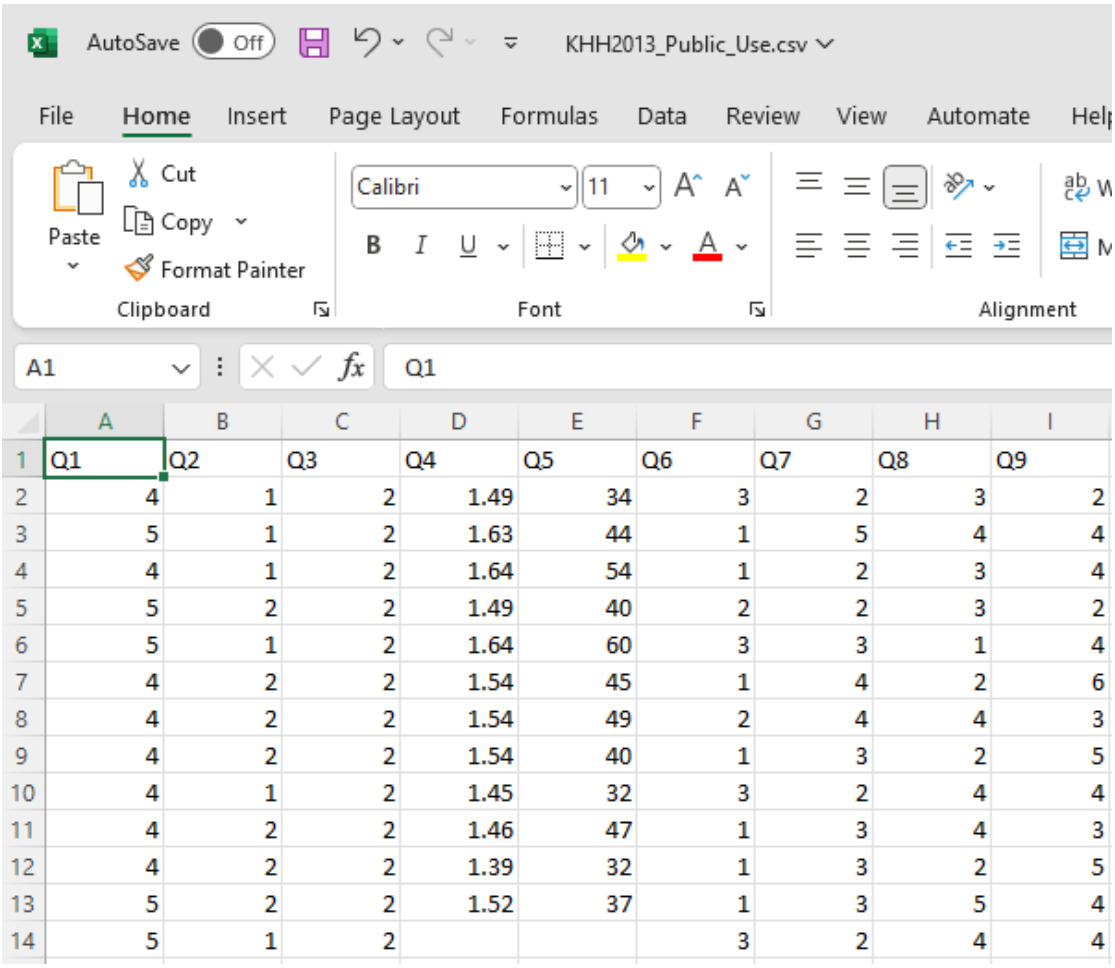
```
library(tidyverse)
library(emmeans)
library(car)
library(mice)
```

Exercise 1.

The Global school-based student health survey (GSHS) is a collaborative surveillance project designed to help countries measure and assess the behavioural risk factors and protective factors in 10 key areas among young people aged 13 to 17 years. The GSHS is a relatively low-cost school-based survey which uses a self-administered questionnaire to obtain data on young people's health behaviour and protective factors related to the leading causes of morbidity and mortality among children and adults worldwide.
https://www.who.int/teams/noncommunicable-diseases/surveillance/systems-tools/global-school-based-student-health-survey

In this exercise, we will use the data collected from Cambodia in 2013. This dataset is publicly available from the WHO. The SPSS data file has already been included in the workshop material.
SPSS data file: **KHH2013_Public_Use.sav**
Data codebook: **KHH2013_public_use_codebook.pdf**

If you have SPSS, try saving the data file into CSV and loading it up in R. If you don't have SPSS, a CSV file is included in the workshop material
CSV data file: **KHH2013_Public_Use.csv**



	A	B	C	D	E	F	G	H	I
1	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9
2	4	1	2	1.49	34	3	2	3	2
3	5	1	2	1.63	44	1	5	4	4
4	4	1	2	1.64	54	1	2	3	4
5	5	2	2	1.49	40	2	2	3	2
6	5	1	2	1.64	60	3	3	1	4
7	4	2	2	1.54	45	1	4	2	6
8	4	2	2	1.54	49	2	4	4	3
9	4	2	2	1.54	40	1	3	2	5
10	4	1	2	1.45	32	3	2	4	4
11	4	2	2	1.46	47	1	3	4	3
12	4	2	2	1.39	32	1	3	2	5
13	5	2	2	1.52	37	1	3	5	4
14	5	1	2			3	2	4	4

Task:

1. Set up your working directory. This will be the folder where you save your files and scripts. Hints: use **setwd()**
2. Load the CSV data file into R and store the data into a data frame called **cambodia_data**. Hint: use **read.csv()**
3. Find out the number of observations and number of variables in the dataset. Hint: use **summary().**
4. The GSHS data is a huge dataset. Let's focus on the following variables. Details about each variable are documented in the data code book.
    a. Q1: How old are you?
    b. Q2: What is your sex?
    c. Q6: During the past 30 days, how often did you go hungry because there was not enough food in your home?
    d. Q20: During the past 30 days, on how many days were your bullied?
    e. Q22: During the past 12 months, how often have you felt lonely?
    f. Q23: During the past 12 months, how often have you been so worried about something that you could not sleep at night?
    g. Q24: During the past 12 months, did you ever seriously consider attempting suicide?
    h. Q26: During the past 12 months, how many times did you actually attempt suicide?
    i. Q27: How many close friends do you have?
    j. Q29: During the past 30 days, on how many days did you smoke cigarettes?
    k. Q35: During the past 30 days, on how many days did you have at least one drink containing alcohol?
    l. Q49: During the past 7 days, on how many days were you physically active for a total of at least 60 minutes per day?
    m. Q50: During the past 7 days, on how many days did you walk or ride a bicycle to or from school?

    Now create a subset of the data and store it into a data frame called **cambodia_small.** Hint: use **select()**
5. Find out the number of missing data in each variable. Hint: use **summary()**

Exercise 2.

Task:

1. Find out the individuals who have missing data in age (Q1) or sex (Q2). Among these individuals, what is the mean score of their measure of loneliness (Q22)? Hint: use **filter()** to select cases with missing data, store them into a new data frame object called **cambodia_no_age_sex**. Use **summary()** on this new data frame.
2. **For the rest of the workshop**, we will only focus on individuals who are between 12 and 17 years old. Create a new data frame object, **cambodia_1217**, with only these individuals. Hint: use **filter()**. Please note that age was coded from 1 to 8, with 1 = 11 years old or younger and 8 = 18 years old or older (see data code book).
    a. What is the new sample size?
    b. How many females and males in the data? Hint: use table()
    c. What is the proportion of participants who have considered attempting suicide? Hint: use **table()** and **prop.table()**
    d. Within each sex, what is the proportion of participants who have considered attempting suicide? Hint: use **xtabs()**
    e. What is the mean and standard deviation of loneliness score (Q22) by sex? Hint: use **groupby()** and **summarize()**
3. Create a new age variable that its score reflects the actual age of the participants (i.e., 12 means 12 years old). Hint: In the original coding, 2 = 12 years old, 3 = 13 years old, etc. We can create a new variable by adding 10 to the existing variable. Name this new variable as "**age**".

Exercise 3.

Task

1. Test if sex is associated with suicide attempt (Q26) in a linear regression.
    a. Prior to running this analysis, based on the value in Q2, create a sex variable such at 0 is female and 1 is male, and convert this into a factor variable. Hint: use **mutate()** and then **factor()**. Use **table()** to check if this variable is properly created.
2. Test if being bullied was associated with suicide attempt in a linear regression.
    a. Prior to running this analysis, create a bullied variable such that 0 represents 0 days being bullied, 1 represents 1 or 2 days, 2 represents 3 days or more. Convert this into a factor variable.
    b. Conduct pairwise comparisons between the three levels.
        i. What do you find?
3. Test if age, sex, being bullied (as a 3-level categorical variable), alcohol use (Q35) are associated with suicide attempt (Q26) in a linear regression.
    a. Find all regression coefficients and the associated 95% confidence intervals.
    b. Perform residual analysis and model diagnostic. What is your conclusion?
4. Following up the model in (3), test if sex moderates the effect of alcohol on suicide attempt in a linear regression.
    a. What are the slopes of alcohol use for different sex?
    b. Are the differences in slope statistically significant?
    c. Produce an interaction plot.

Exercise 4.

Task

1. Create a binary suicide attempt (Q26) variable such that 0 represents no suicide attempt and 1 represents one or more attempt. Store this variable as **suicide_attempt.**
2. Convert this into a factor variable with label “No suicide attempt” and “One or more attempt”.
3. Create a binary alcohol use (Q35) variable such that 0 represents no alcohol use and 1 represents alcohol use in at least 1 day. Store this variable as **alcohol**.
4. Convert this into a factor variable with label “No alcohol use” and “At least one day”.
5. Test if age, sex, being bullied and alcohol use (using the new binary variable) are associated with suicide attempt (the new variable) in a logistic regression.
    a. Find all regression coefficients and the associated 95% confidence intervals.
    b. Convert all coefficients and the confidence intervals into odds ratio.
6. Test of sex moderates the effect of alcohol on suicide attempt.

RStudio problem

https://github.com/rstudio/rstudio/issues/13188