Instructions

- Clear the environment
- Open a new R Script called `day3_exercise_script` where you will do the exercise and later save in the `day3` project directory.
- Add the purpose of the file and the author
- Here are the main activities for this excercise
  1) Load the `rio`, `lubridate`,`epikit`, `janitor` ,and `tidyverse` package
  2) Load the data using the `import`
  3) Create a pipe chain to clean the data
  4) Work with categorical variables
  5) Work with dates
  6) Clean the string variables
  7) Using ggplot, plot a histogram
  8) Using ggplot, plot a scatter plot
  9) Using ggplot, plot a boxplot
  10) Using ggplot, plot a barplot
  11) Using ggplot, plot a linegraph

Part 1 (*You* :-) will do this together)

1.1

- Load the `rio`, `lubridate`,`epikit`, `janitor` ,and `tidyverse` package
- Using the `hospital_df` and `location_df` as the object names, load the `line_hospital_data.csv` and the `line_hospitals_locations.xlsx` data respectively

Helper code

```
# example of laoading data
hospital_df <- import("Data/line_hospital_data.csv")
```

1.2 Create a pipe chain to clean the `hospital_df`

Create a new data called `hospital_df_clean` with the following activities

- Remove the spaces in the variable names using the `clean_names()` function
- Remove duplicates in `case_id` using the `distint()` function
- Create an age_cat variable with a split of 10yrs. Hint: use the `age_categories` function
- Create a `year_hosp` and `month_hosp` from hospital visit date data
- Create a `year_onset` and `month_onset` from date onset date data
- Use the `year_hosp` and `year_onset` to report the numbers in each year
- Create a new variable `days_to_hosp`by subtracting `hosp_date - date_onset`
- Use the `year_onset` and `month_onset` to report the numbers in each month of the year: Hint use `group_by()` and `tally()`

Helper code

```
hosptial_full %>%
  group_by(hospital) %>%
  tally()
```

1.3 Merge the datasets

- Left join the `hospital_df_clean` to the `location_df` and create `hospital_df_merged` data
- In the `hospital_df_merged` data clean the hospital variable using `recode`: Hint Correct the spelling mistakes

Helper code

```
hospital_df_merged %>%
   # re-code hospital column to have same ne
     mutate(hospital = recode(hospital,
                              # for reference: OLD = NEW
                              "Mitilary Hospital"  = "Military Hospital",
                              "Port"               = "Port Hospital",
                              "Port Hopital"       = "Port Hospital",
                              "Mitylira Hopital"   = "Military Hospital",
                              "Mitylira Hospital"  = "Military Hospital",
                              "St. Mark's Maternity Hospital (SMMH)" = "SMMH"))
```

1.4 Plot histogram and density using ggplot2

- Plot a histogram and density of the age data
- Plot a histogram and density of of the weight data
- Plot a histogram and density of of the ct_blood
- Plot the histogram and density of BMI that you calculated in exercise2
- What do you think of the distributions above?
- Repeat above and change the `bins = 5` / what changes?

```
ggplot(data = ,aes()) +
  geom_histogram()
```

1.5 Create a scatter plot

- What do you think of weight vs age?
- Plot a scatter of weight vs age
- Plot a scatter of weight vs age and color the points by gender
- Plot a scatter of weight vs age and color the points by outcome

1.6 Create a scatter plot + line graph

1. Plot a scatter of weight vs height then add a line graph
2. Plot a scatter of weight vs height then add a line graph color by gender: Adjust the size of the dots: Change the line types
3. Plot a scatter of weight vs height then add a line graph **color the points** by gender: Add the theme_bw
4. What is the difference between 2 and 3
5. What do you think of the trend?

1.7 Create boxplot using ggplot

1. Plot a box weight vs gender
2. Plot a box of height by age_group
3. Plot a box of height by age_group color by gender
4. Plot a box of ct_blood vs chills
5. Give a summary of what you observe
6. Plot a box of height by age_group color by gender add scatter. Try adding a layer of theme_bw()

1.7 Create barplot using ggplot

1. Plot a barplot of gender
2. Plot a barplot of gender and chills

3. Plot a barplot of age group and color by gender
4. Plot the count of symptoms onset per. Hint: create year variable and use that to plot

Extra to try:

Remember the `filter participants that had cough AND chills OR aches OR their ct_blood IS GREATER than 20`,

1. Do a box plot of wt_kg by gender having removes the participants with missing age.
2. Add a scatter plot using `geom_jitter`
3. Change the x and y axis labels
4. Change y limits
5. Add the `theme_bw()` in the ggplot command.

Here we compare whether there is a weight difference in participants who meet the above condition in terms of gender.