1 Topics Covered:

This problem set covers the following topics:

1. Basic statistics (measures of central tendencies and dispersion)
2. Creating bar plots and histograms
3. Creating box plots
4. Doing the previous 3 in R

2 Tasks:

2.1 Tasks without R (do by hand)

For questions $1 - 4$, do the following by hand:

1. The mean of four numbers is 46.5. If three of the numbers are 50, 44, and 40.5, what is the value of the fourth number?
2. A count of size of families of a bird species was performed and the following observations were collected:

   `5, 7, 4, 9, 5, 4, 4, 3, 3, 3, 7, 4, 5, 9, 8, 2, 4, 5`

   (a) Calculate the mean and the median of the data. What observations can you make based on these mean and median figures?
   (b) Create a histogram of this data.
   (c) Create a box plot of this data.
3. Following is the number of nuts a squirrel collected on each day in a week (no particular order of days):

   `10, 39, 71, 39, 76, 38, 25`

   (a) How many nuts did she collect on average that week?
   (b) A random person from Hamburg came to you and made this claim: the squirrel collected less than 55 nuts 75% of the time. On what grounds would you challenge or defend that claim?
4. Some students at Leuphana reported the distance (k.m.) they had to travel in order to arrive at the university to take their tutorial for a statistics course. Here is what they reported:

   `2.0, 3.0, 3.0, 2.0, 3.0, 0.0, 2.0, 2.0, 2.0, 3.0, 3.0,`
   `3.5, 3.0, 3.0, 55.0, 2.0, 85.0, 4.0, 40.0, 3.0`

As you can easily see, there are some extreme cases where students have travelled very long (or very short) distances in order to arrive at the campus. We call these values outliers. Based on this:

(a) Identify the outliers from this data. *(Hint: inter-quartile range)*

(b) Would you choose mean as the appropriate measure of central tendency with this data set? Defend your choice.

2.2 Tasks in R

You have to perform the tasks on this section using R.

If you get lost at any point, use `help()` function to learn more about the task. Alternatively, you can also use the help tab from lower right hand corner of R-Studio and enter your search term there. If that does not help, you can search for the term online. Some sites that you can trust with programming questions are:

- StackExchange (`www.stackexchange.com`), StackOverflow (`www.stackoverflow.com`)
- RPubs (`www.rpubs.com`), ETH Zürich (`https://stat.ethz.ch/`)
- `www.r-bloggers.com` and instructions from YouTube.

5. Below is the data from Question 4 above:

```
2.0, 3.0, 3.0, 2.0, 3.0, 0.0, 2.0, 2.0, 2.0, 3.0, 3.0,
3.5, 3.0, 3.0, 55.0, 2.0, 85.0, 4.0, 40.0, 3.0
```

Using this data, do the following in R:

(a) Create a variable and save this data into that variable as a vector. hint: in R, use: data <- c(5,6,...)

(b) Find the summary statistics of the data using `summary()` function.

(c) Calculate the variance and standard deviation (s.d.) of the data.

(d) Create a bar plot, a histogram, and a box plot. Is the data normally distributed? Why do the bar plot and the histogram look different?

6. R provides a good assortment of datasets for us to investigate. The datasets that have been built into R are called "built-in datasets".

(a) Load a built-in dataset called `swiss` and store it in a variable named `swiss_df`.

(b) Use the help function from R to find out what this package is about.

(c) Create a variable named `swiss_agriculture` and store a column named `Agriculture` from `swiss_df`. *Hint: use the $ notation to select columns from the data frame.*

(d) What is the number that divides the data from `swiss_agriculture` into 25% and 75%? *(lower quartile)*

(e) Plot the following from `swiss_agriculture`: a scatter plot, a box plot, a histogram, a bar plot.