Assignment 4: exploratory data analysis and linear models (8 marks)

To submit this assignment, upload the full document on blackboard, including the original questions, your code, and the output. Submit you assignment as a knitted `.pdf` *(prefered) or* `.html` *file.*

1. Visualization (2 marks) Import the tidyverse library. We will be using the same beaver1 dataset that we used in last week's assignment.

```
library(tidyverse)
```

   a. Using what you learned during your last assignment, turn the "activ" column into a categorical variable (factor). Then, create a plot with body temperature on the x-axis. Visualize body temperature separately by whether the beaver is active on the same graph. Hint: we learned about a new useful function that does this during our exploratory data lesson. (1 mark)

   b. Create a dataframe (summary table) that counts the number of temperature recordings that fall into bins with a width of 0.1. Group by day. (1 mark)

2. Outliers and Missing Values (2.5 marks)

   a. Identify any temperature scores that are two standard deviations away from the mean. Replace them with NA's (missing values). (1 mark)

   b. Run the following code:

```
set.seed(10)
beaver1[sample(1:nrow(beaver1), 20), "temp"] <- NA
```

Run multiple imputation (set a seed of 100) to replace the missing values. Exclude columns as predictors if necessary. Compare the means of the imputed and original values. (1.5 marks)

3. Generalized Linear Models (3 marks)

```
co2_df <- as_data_frame(as.matrix(CO2)) %>%
    mutate(conc = as.integer(conc),
           uptake = as.numeric(uptake))
```

   a. Look through the help documentation (?CO2) to understand what each variable means. Which variable(s) do you think would be the y in the GLM model? Which variable(s) would be the x? Briefly defend these choices. (1 mark)

   b. How much does `uptake` change if `conc` goes up by 10 mL/L? (*Note:* it is intentional that there is no mention of the other variables in the model.) Write out the interpretation as a simple statement of this contribution of `conc` on `uptake`, when the other variables are also in the model. (2 marks)

   c. Run the following code if you need to download our survey data.

```
download.file("https://ndownloader.figshare.com/files/2292169", "survey.csv") #if you need to re-downlo
survey <- read_csv("survey.csv")
```

Use logistic regression to see if weight significantly predicts sex. Make a concluding statement as to indicate whether the model is significant and create a plot to visualize the linear model. Run the following code to ensure sex is treated as a factor variable:

```
survey$sex <- as.factor(survey$sex)
```

Hint: you need to make sure there are only two levels to this variable: `"F"` and `"M"`. (0.5 marks)