STA130 - Week 1 Problem Set (Fall 2025)

Nathalie Moon

Problem Set Objectives

The weekly problem sets are designed to do the following:

- Provide structured practice with immediate feedback, allowing students to identify and correct misconceptions early
- Build practical data analysis skills through hands-on experience with R, reinforcing statistical concepts through active application
- Develop reproducible workflow habits, namely through the use of Quarto to combine code, output, and text.

Instructions

How do I check my work

You will access the Problem Set and do your work on JupyterHub.

When you click `Render` to create a pdf of your solutions, you may get a popup that says "Popup Blocked - We attempted to open an external browser window…". If you get this message, click `Cancel` and then you will be able to open your .pdf file by selecting it in the bottom right window (in the same folder as your .qmd file).

Once you are done with a question and want to check if it is correct, you will need to download your .qmd and .pdf files and upload them to MarkUs to run the tests.

Usually when you do an assignment, you don't find out whether your answers are correct until *after* the deadline, when you get your grade back. However, using MarkUs, you can submit your work before the deadline and run tests to check your solutions!

Note: Some parts of some questions may not be covered by tests in MarkUs, but you're still responsible for reviewing the posted solutions and make sure you understand them.

To download your files from JupyterHub, go to the bottom right window and do the following:

- Select the files you want to download (likely .qmd and .pdf)
- Click on More ⇒ Export

To upload your work to MarkUs to run the tests:

- Go to MarkUs: https://markus.teach.cs.toronto.edu/markus/courses
- Open the current assignment and upload your file(s) (note that by default, the files you download from JupyterHub will likely be in your `Downloads` folder)
- Run the tests

What to do if a test fails on MarkUs

- Take a deep breath! Your work won't really be graded until the deadline, so start early to make sure you have lots of time to resolve issues before the deadline.
- Read the message to get hints about what the problem is. For example "variable X not present" means that you may have a typo in your variable name.
- Search on Piazza to see if other classmates have encountered a similar error (and if not, consider posting a screenshot of the error message)
- Come to TA or instructor office hours with your issue

How do I submit my Problem Set?

You will submit your solutions (.qmd and .pdf) on MarkUs at the link above. You can submit as many times as you like but only your latest submission will be counted.

Question 1

In this question, you'll use R as a calculator to get familiar with the kinds of operations it can do.

(a) What is 420 times 85? Save the answer in the variable below called Q1a.

```
# Replace NULL with your answer below
Q1a <- NULL
Q1a
```

```
NULL
```

(b) What is 2565 divided by 3? Save the answer in the variable below called Q1b.

```
# replace NULL with your answer below
Q1b <- NULL
Q1b
```

```
NULL
```

(c) What is the sum of all positive integers from 8 to 15? Save the answ

```
# Replace NULL with your answer
Q1c <- NULL
Q1c
```

```
NULL
```

(d) What is 0.25 to the fourth power? Save the answer in the variable below called Q1d.

```
# Replace NULL with your answer
Q1d <- NULL
Q1d
```

```
NULL
```

(e) Calculate the absolute value of -17.8using the `abs()` function. Save the answer is the variable called Q1e.

```
# Replace NULL with your answer
Q1e <- NULL
Q1e
```

```
NULL
```

(f) Calculate the average of the numbers 5, 12, 18, 23, and 32 using the `mean()` function.

```
# Here is a vector containing these numbers, constructed using the c() function
Q1f_vector <- c(5, 12, 18, 23, 32)

# Replace NULL with your answer
Q1f <- NULL
Q1f
```

```
NULL
```

Question 2

In this question, you'll experiment with the `length()` function. Run the code chunk below (using the green arrow in the top right corner of the gray block) and answer the questions below.

```
# Below, the hobbies object is a vector with the list of a student's hobbies
hobbies <- c("reading", "gaming", "cooking", "hiking")
length(hobbies)
```

```
[1] 4
```

(a) Write one sentence describing what the `length()` function does?

Your answer: The `length()` function returns the number of values in a vector.

(b) Suppose the student picks up a new hobby, photography. Create a new vector with all the hobbies from before, as well as this new hobby as the final element of the vector. Save the answer in the variable below called `Q2` (replace NULL with your answer).

```
Q2 <- NULL
Q2
```

```
NULL
```

Question 3

For this question we will work with data about the TV show Avatar: The Last Airbender.

(a) The name of the data set is `avatar.csv`. Load the data using `read_csv()` and save it under the name "avatar".

```
# Tip: don't forget to put quote marks around the name of the dataset
# inside the function
```

(b) We have learned two functions this week that let us quickly get an idea of our data. Apply both of them to the avatar data.

```
# Your code here
```

(c) Based on your answer to b) answer the following:

- How many observations does the `avatar` data frame include? Save your answer in the R object below called `Q3_num_observations` (replace NULL with your answer).
- How many variables are measured for each observation? Save your answer in the R object below called `Q3_num_variables` (replace NULL with your answer).
- What is the name of the third variable in the `avatar` tibble? Save your answer in the R object below called `Q3_fifth_variable_name` (replace NULL with your answer). Hint: when your answer is a word, make sure to put it in quotation marks.
- What is the value of the `director` variable for the first observation in the `avatar` tibble? Save your answer in the R object below called `Q3_first_value_of_director` (replace NULL with your answer). Hint: when your answer is a word, make sure to put it in quotation marks.

```
Q3_num_observations <- NULL
Q3_num_variables <- NULL
Q3_fifth_variable_name <- NULL
Q3_first_value_of_director <- NULL

# Hint: You can use the nrow() and ncol() functions to get the number of rows
```

```
# and number of columns from the avatar tibble, and make sure you see where
# these numbers appear in the glimpse() as well
# If your answer is one or more words, make sure to use quotation marks.
```

Question 4 (you'll use your answers to this question in this week's Friday tutorial)

In this question, you will identify different types of variables using data you might collect about recent blockbuster movies:

- Movie title
- Release year
- IMDb rating
- Runtime (in minutes)
- Genre (animation, comedy, drama, sci-fi, action, etc)
- Rating (G, PG, PG-13, R, etc)

For each variable below, identify whether it is quantitative (numerical) or categorical. If quantitative, specify whether it's discrete or continuous. If categorical, specify whether it's nominal, ordinal, or binary. For each variable, rate how confident you are in your answer (note: there may be more than one correct answer here, and you'll discuss these further in tutorial).

Tip: To fill in the table below, you may find it easier to use the visual editor (click on "Visual" / "Source" at the top-left corner of this window to switch between the two)

Variable	Type of Variable	Your confidence level (very confident / somewhat confident / not at all confident)
Movie title		
Release Year		
IMDb Rating		
Runtime (minutes)		
Genre		
Rating		

Question 5 (you'll use your answers to this question in this week's Friday tutorial)

In this question, you will consider another example of survivorship bias in business success stories.

A popular business magazine publishes an article titled "The Secrets of Successful College Dropouts" featuring entrepreneurs like Bill Gates and Mark Zuckerberg who dropped out of college and became billionaires. The article concludes that "dropping out of college can be a path to extraordinary business success" and interviews 20 college dropouts who are now running successful companies.

(a) Is the sample of college dropouts featured in this article representative of all people who drop out of university? What group of people is missing from these data?

Write your answer here:

(b) If someone uses these data (and this article) to conclude that dropping out of university leads to business success, what could be wrong with this conclusion?

Write your answer here:

(c) This is another example of survivorship bias, similar to the airplane armor problem discussed in class. In the airplanes example, we only saw data from planes that returned safely. In this example, what data are we not seeing?

Write your answer here: