# Module 2 Lab

## Part 1: HTML Bio

INTRODUCE YOURSELF! Using R Markdown, create an HTML report that provides me with background information about you. Here is an old example introduction to me[1] that you can base your outline on (disregard the homework problems at the bottom). You do not need to follow this outline exactly; however, this provides the fundamental output I am looking for. I do expect you to include an image so I know who you are! Be sure to use headings appropriately and feel free to incorporate features such as italics, bold font, lists, and HTML links.[2] If you want to get saucy, show me you can write an equation and execute code chunks; however, these are not required.

### Setting up your environment

1. Create an R project for this course so that all future scripts, inputs, and outputs are organized.
2. Make sure the following RStudio preference settings are set[3]:
    - General: Set "Save workspace to .RData on exit: Never".
    - Code: In the display tab check the "Show margin" option and set "Margin Column: 88".
    - Code » Diagnostics: Make sure the "Provide R style diagnostics" is checked.
3. Create a new HTML R Markdown script.

### What to include

The basics that I want you to include in your report are:

- Which one is you? (include a picture)
- A synopsis of who you are
- Academic Background
- Professional Background
- Experience with R
- Experience with other analytic software

Once you created a Bio for yourself post it to RPubs (https://rpubs.com/about/getting-started) to obtain the URL for your report. You will submit this URL when you take the Lab Quiz. That's it! You're done with this part.

[1] https://rpubs.com/bradleyboehmke/datawrangling_week1_homework

[2] If you need help with the syntax, go to Help » Markdown Quick Reference or use the cheat sheet located at Help » Cheatsheets » R Markdown Cheat Sheet.

[3] They should be if you followed the lesson 2a; however, this will ensure these settings remain constant for this R project.

## Part 2: Importing data

1. Fill in the blanks below to import the blood_transfusion.csv file (provided via Canvas) and answer the following questions.

    - What are the dimensions of this data (number of rows and columns)?
    - What are the data types of each column?
    - Are there any missing values?
    - Check out the first 10 rows? What are the `Class` values for the first 10 observations?
    - Check out the last 10 rows? What are the `Class` values for the last 10 observations?
    - Index for the 100th row and just the `Monetary` column. What is the value?
    - Index for just the `Monetary` column. What is the mean of this vector?
    - Subset this data frame for all observations where `Monetary` is greater than the mean value. How many rows are in the resulting data frame?

```
# import data and the message print out will also tell you
# what the data types are
df <- readr::read_csv(____________)

# Are there any missing values?
sum(____(df))

# What are the dimensions of this data
___(df)

# Check out the first 10 rows
head(df, __)

# Check out the last 10 rows
____(df, 10)

# Index for the 100th row and just the `Monetary` column. What is the value?
df[___, 'Monetary']

# Index for just the `Monetary` column. What is the mean of this vector?
_____(df[['Monetary']])

# Subset this data frame for all observations where `Monetary` is greater
# than the mean value. How many rows are in the resulting data frame?
above_avg <- df[['Monetary']] ___ mean(df[['Monetary']])
df[above_avg, 'Monetary']
```

2. Fill in the blanks below to import the PDI___Police_Data_Initiative___Crime_Incidents.csv data (provided via Canvas) and answer the questions that follow. Data is taken from the City of Cincinnati Open Data Portal website [4], which you may need to read to place context in your answers.

    - What are the dimensions of this data (number of rows and columns)?
    - What do you think these columns represent?
    - Are there any missing values in this data? If so, how many missing values are in each column? Which column has the most missing values?
    - Using the `DATE_REPORTED` column, what is the `range` of dates included in this data?
    - Using `table()`, what is the most common age range for known `SUSPECT_AGE`s?

[4] https://data.cincinnati-oh.gov/safety/PDI-Police-Data-Initiative-Crime-Incidents/k59e-2pvf

- Use `table()` to get the number of incidents per zip code. Sort this table for those zip codes with the most activity to the least activity. Which zip code has the most incidents? Do you see any peculiar data quality issues with any of these zip code values?
- Using the `DAYOFWEEK` column, which day do most incidents occur on? What is the proportion of incidents that fall on this day?
- Looking at the information this data set provides, what are some insights you'd be interested in assessing? Analyze three different columns that could start to provide you with these insights. Are there missing values in these columns? What are some summary statistics you can compute for these columns? Are there any outliers or aberrant values in these columns? How do you know? Would you remove or recode them?

```
df <- readr::read_csv(_________)

# dimensions of this data
------

# Are there any missing data
------

# If so, how many missing values are in each column?
______(is.na(df))

# Using the `DATE_REPORTED` column, what is the `range` of dates included in this data?
_____(df[['DATE_REPORTED']])

# Using `table()`, what is the most common age range for known `SUSPECT_AGE`s?
_____(df[[________]])

# Use `table()` to get the number of incidents per zip code. Sort this table
# for those zip codes with the most activity to the least activity.
____(table(df['ZIP']), decreasing = _____)

# Using the `DAYOFWEEK` column, which day do most incidents occur on?  What
# is the proportion of incidents that fall on this day?
table(df[[_____]]) / sum(table(df[[_____]]))
```