# Introduction to Data Science

**Instructor: Daniel D. Gutierrez**

# HOMEWORK 1

NOTE: for this homework assignment, please use the `airquality` data set from base R where indicated. Please hand in an R script file with your answers – **R code and R output**.

**Question 1**

Write an R script to perform the following operations:

- Define a new matrix named `mat` with 3 row and 4 columns, filling with a sequence of values `1:12` by row
- Add a new row to the matrix containing all 9's
- Add a new column to the matrix containing all 8's

The resulting matrix should be the following:

```
     [,1] [,2] [,3] [,4] [,5]
[1,]    8    9    9    9    9
[2,]    8    1    2    3    4
[3,]    8    5    6    7    8
[4,]    8    9   10   11   12
```

**Question 2**

Write an R script to perform the following operations:

- Define a new list named `lst` with the following components: (a) a character vector containing the values "Ellen", "Catherine", and "Stephen", (b) an integer vector containing the values 90, 95, and 99, (c) a matrix containing attendance records for two class sessions where the first row represents the first session attendance for each student, and similarly for the second row. Here is what the attendance matrix should look like:

```
     [,1]  [,2]  [,3]
[1,] TRUE  TRUE FALSE
[2,] TRUE FALSE  TRUE
```

- Display the names of all students
- Display Stephen's grade

- Display Caterhine’s attendance for both sessions

**Question 3**

Write an R script to define a new character vector `gender` that is populated with `“male”` for the first 20 elements, and `“female”` for the next 30 elements. Use factors to quickly and efficiently calculate a total number of values in each gender category. [Hint: you may find the R function `rep()` useful for this question]

**Question 4**

Using the `airquality` data set, how many missing values are in the `Ozone` column of this data frame? [Hint: must use subsetting in your R code to solve]

(a) 43

(b) 37

(c) 9

(d) 78

**Question 5**

Extract the subset of rows of the data frame where `Ozone` values are above 31 and `Temp` values are above 90. What is the mean of `Solar.R` in this subset? [Hint: must use subsetting in your R code to solve]

(a) 334.0

(b) 185.9

(c) 205.0

(d) 212.8

**Question 6**

Make a copy of the `airquality` data frame so you can add a new variable `hotcold`. The new variable shall have two possible values: “hot” if the value of the `Temp` variable is greater than the median value for `Temp`, and cold otherwise. You can hand in the results of head() and tail() to show your code is working. [Hint: you may find the R functions `ifelse()` and `median()` useful for this question]

**Question 7**

Based on a traditional English children's game, write an R script that:

- Prints the numbers from 1 to 100
- For multiples of 3, print "Fizz" instead of the number
- For multiples of 5, print "Buzz" instead of the number
- For multiples of 3 and 5, print "FizzBuzz" instead of the number

Here's an example of what the print display should look like:

```
[1] 1
[1] 2
[1] "Fizz"
[1] 4
[1] "Buzz"
[1] "Fizz"
[1] 7
[1] 8
[1] "Fizz"
[1] "Buzz"
[1] 11
[1] "Fizz"
[1] 13
[1] 14
[1] "Fizz Buzz"
[1] 16
[1] 17
[1] "Fizz"
[1] 19
[1] "Buzz"
. . . and so on
```

**Question 8**

Using the following matrix definition: `mat1 <- matrix(rep(seq(4), 4), ncol = 4)` use one of R's loop functions to compute the sum of each row plus 2. [Hint: you'll need a user defined function (UDF) or an anonymous function to find the answer]

**Question 9**

Using the built-in character vector in R that contains all the U.S. state names: `state.name` write an R script to randomly select 10 names from the vector and create a new vector to store them. For reproducibility, make sure you get the same 10 names each time you run the code. Next, sort the new vector according to the state names. The resulting vector should contain 10 sorted state names. [Hint: you may find the R functions `sample(), order(),` and `set.seed()`, along with the data set `state.name` useful for this question]

**Question 10**

Create a variable in R named `xct` to store the date and time of the Apollo 11 moon landing on July 20, 1969 at 20:18 UTC. Calculate and display the number of years since the moon landing until today. [Hint: you may find the R function `Sys.time()` useful for this question]