Assignment 3 STATS 220 Data Technologies Due 12:00pm Thursday 2 June

All questions in this assignment relate to crime statistics in New Zealand.[1]

The data are available as 18 plain text files:
`http://stat220.stat.auckland.ac.nz/stats220/2016/notes/burglaries201407.csv`,
`http://stat220.stat.auckland.ac.nz/stats220/2016/notes/burglaries201408.csv`
to
`http://stat220.stat.auckland.ac.nz/stats220/2016/notes/burglaries201512.csv`.

There is also a ZIP file containing all 18 text files in a single download:
`http://stat220.stat.auckland.ac.nz/stats220/2016/notes/burglaries.zip`.

Each file contains data for one month and, within each file, each row of data gives several different measures for one "meshblock." A meshblock is a relatively small geographic region in New Zealand. Each meshblock belongs to an "AreaUnit", which is a larger area, roughly corresponding to a suburb, and to a "DHB" (District Health Board), which is an even larger area. The "Count" is the total number of burglaries in the meshblock for the month. The first few lines of the file for July 2014 are shown in Figure 1.

There are **9** questions in this assignment, each requiring you to write some R code to work with this data set. All questions are worth equal marks.

```
"MeshBlock","Count","AreaUnit","TerritorialAuthority","DHB"
100,0,"North Cape","Far North District","Northland"
200,0,"North Cape","Far North District","Northland"
300,0,"North Cape","Far North District","Northland"
400,0,"North Cape","Far North District","Northland"
502,0,"North Cape","Far North District","Northland"
600,0,"North Cape","Far North District","Northland"

...
```

Figure 1: The first few lines of the New Zealand crime data from July 2014.

NOTE: You should submit a file containing R code that assigns the appropriate values to the appropriately named symbols. I will mark your code by running your code and inspecting the values that have been assigned to the relevant symbols.

NOTE: You can check whether your answers are correct by running the following R code:

```
source("http://stat220.stat.auckland.ac.nz/stats220/2016/notes/ass3marking.R")
```

This will print the value `TRUE` for each answer that you have correct and `FALSE` for each answer that you have wrong. If you have an answer wrong, it may also print out some messages about how your answer differs from the correct answer.

NOTE: You should submit your answers via the submission form in the "Submissions" section of the STATS 220 web site.

[1] The data are a modified version of original data from the New Zealand Police and were obtained from a github account created by Harkanwal Singh, Data Editor at the New Zealand Herald. `https://github.com/kamal-hothi/nz_burglary_data`. These data were the basis for a week-long series of Herald articles on crime. `http://www.nzherald.co.nz/nz/news/article.cfm?c_id=1&objectid=11600999`.

Each question has an indication of when we will have covered the relevant material in the course. For example, "(Week 8)" means that you should be able to attempt the question by the end of Week 8 of this semester.

1. (Week 8)
   Use the `read.csv()` function to read the file `"burglaries201407.csv"` into R. Your code should assume that the file is in the current working directory.

   You should end up with a **data frame** called `july2014` that looks like this:

```
> head(july2014)

  MeshBlock Count   AreaUnit TerritorialAuthority       DHB
1       100     0 North Cape   Far North District Northland
2       200     0 North Cape   Far North District Northland
3       300     0 North Cape   Far North District Northland
4       400     0 North Cape   Far North District Northland
5       502     0 North Cape   Far North District Northland
6       600     0 North Cape   Far North District Northland
```

2. (Week 9)
   Use the `max()` function to calculate the largest number of burglaries in one month in the data frame `july2014` and assign the result to the symbol `july2014max`.

   You should end up with a **numeric vector** called `july2014max` that looks like this:

```
> july2014max

[1] 11
```

3. (Week 10)
   Use the `rep()` function to create a numeric vector containing the four-digit year components of the file names.

   You should end up with a **numeric vector** called `years` that looks like this:

```
> years

 [1] 2014 2014 2014 2014 2014 2014 2015 2015 2015 2015 2015 2015 2015 2015 2015
[16] 2015 2015 2015
```

4. (Week 10)
   Use the `c()`, `rep()`, and `paste()` functions to create a character vector containing the two-digit month components of the file names.

   You should end up with a **character vector** called `months` that looks like this:

```
> months

 [1] "07" "08" "09" "10" "11" "12" "01" "02" "03" "04" "05" "06" "07" "08" "09"
[16] "10" "11" "12"
```

   Hints:

   One approach to this question is to generate a sequence of numbers for the first digit ...

```
 [1] 0 0 0 1 1 1 0 0 0 0 0 0 0 0 0 1 1 1
```

   ... and a separate sequence of numbers for the second digit ...

```
 [1] 7 8 9 0 1 2 1 2 3 4 5 6 7 8 9 0 1 2
```

   ... and `paste()` those two sequences together.

   Another approach is to generate the sequence of months ...

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

   ... and then use the modulo operator (`%%`) to find the second digit ...

```
 [1] 7 8 9 0 1 2 1 2 3 4 5 6 7 8 9 0 1 2
```

   ... and integer division (`%/%`) to find the first digit ...

```
 [1] 0 0 0 1 1 1 0 0 0 0 0 0 0 0 0 1 1 1
```

5. (Week 10)
   Use the `paste()` function to combine `months` and `years` to create the 18 file names.

   You should end up with a **character vector** called `filenames` that looks like this:

```
> filenames

 [1] "burglaries201407.csv" "burglaries201408.csv" "burglaries201409.csv"
 [4] "burglaries201410.csv" "burglaries201411.csv" "burglaries201412.csv"
 [7] "burglaries201501.csv" "burglaries201502.csv" "burglaries201503.csv"
[10] "burglaries201504.csv" "burglaries201505.csv" "burglaries201506.csv"
[13] "burglaries201507.csv" "burglaries201508.csv" "burglaries201509.csv"
[16] "burglaries201510.csv" "burglaries201511.csv" "burglaries201512.csv"
```

6. (Week 11)
Write a loop that reads in all of the CSV files, calculates the maximum value from each file, and determines an overall maximum across all of the files. Assign the result to the symbol `overallMax`. Your code should assume that the file is in the current working directory.

You should end up with a **numeric vector** called `overallMax` that looks like this:

```
> overallMax

[1] 12
```

Hints:

Some intermediate steps on the way to the final answer for this quesion might include ...

... a loop that prints out all file names.

... a loop that prints out the maximum value for each of the files.

An example of a loop that updates a value each time the loop runs is shown below:

```
# Initialize overall maximum to small value
maximum <- -Inf
for (i in 1:5) {
    # Generate random number
    newValue <- rnorm(1)
    print(newValue)
    # If random number larger than current maximum, update maximum
    if (newValue > maximum) {
        maximum <- newValue
    }
}
print(maximum)
```

It may help to print out a message each time through the loop like this ...

```
The max for file burglaries201407.csv is 11; the overall max is 11
The max for file burglaries201408.csv is  8; the overall max is 11
The max for file burglaries201409.csv is  9; the overall max is 11
The max for file burglaries201410.csv is  7; the overall max is 11
The max for file burglaries201411.csv is 10; the overall max is 11
The max for file burglaries201412.csv is 12; the overall max is 12
The max for file burglaries201501.csv is  7; the overall max is 12
The max for file burglaries201502.csv is  9; the overall max is 12
The max for file burglaries201503.csv is  8; the overall max is 12
The max for file burglaries201504.csv is 10; the overall max is 12
The max for file burglaries201505.csv is 10; the overall max is 12
The max for file burglaries201506.csv is 10; the overall max is 12
The max for file burglaries201507.csv is  7; the overall max is 12
The max for file burglaries201508.csv is  9; the overall max is 12
The max for file burglaries201509.csv is  8; the overall max is 12
The max for file burglaries201510.csv is  7; the overall max is 12
The max for file burglaries201511.csv is 10; the overall max is 12
The max for file burglaries201512.csv is  8; the overall max is 12
```

7. (Week 12)
   Use subsetting to extract from the `july2014` data frame only the rows that are meshblocks from the Area Unit "Glendowie."

   You should end up with a **data frame** called `july2014glendowie` that looks like this:

```
> head(july2014glendowie)

     MeshBlock Count  AreaUnit TerritorialAuthority      DHB
5606    512000     0 Glendowie              Auckland Auckland
5607    512100     0 Glendowie              Auckland Auckland
5608    512200     0 Glendowie              Auckland Auckland
5609    512300     0 Glendowie              Auckland Auckland
5610    512500     0 Glendowie              Auckland Auckland
5611    512600     0 Glendowie              Auckland Auckland

> dim(july2014glendowie)

[1] 27  5
```

8. (Week 12)
   Write a loop that reads in all of the CSV files and calculates the maximum number of burglaries from meshblocks in Glendowie. Assign the result to the symbol `glendowieMax`.

   You should end up with a **numeric vector** called `glendowieMax` that looks like this:

```
> glendowieMax

[1] 2
```

Hints:

It may help to print out a message each time through the loop like this ...

```
The max for file burglaries201407.csv is  1; the overall max is  1
The max for file burglaries201408.csv is  1; the overall max is  1
The max for file burglaries201409.csv is  1; the overall max is  1
The max for file burglaries201410.csv is  1; the overall max is  1
The max for file burglaries201411.csv is  1; the overall max is  1
The max for file burglaries201412.csv is  1; the overall max is  1
The max for file burglaries201501.csv is  1; the overall max is  1
The max for file burglaries201502.csv is  2; the overall max is  2
The max for file burglaries201503.csv is  2; the overall max is  2
The max for file burglaries201504.csv is  2; the overall max is  2
The max for file burglaries201505.csv is  1; the overall max is  2
The max for file burglaries201506.csv is  1; the overall max is  2
The max for file burglaries201507.csv is  1; the overall max is  2
The max for file burglaries201508.csv is  2; the overall max is  2
The max for file burglaries201509.csv is  2; the overall max is  2
The max for file burglaries201510.csv is  0; the overall max is  2
The max for file burglaries201511.csv is  1; the overall max is  2
The max for file burglaries201512.csv is  1; the overall max is  2
```

9. (Week 12)
   Write a loop that reads in all of the CSV files and calculates the **total** number of burglaries from meshblocks in Glendowie. Assign the result to the symbol `glendowieTotal`.

   You should end up with a **numeric vector** called `glendowieTotal` that looks like this:

```
> glendowieTotal

[1] 71
```

Hints:

It may help to print out a message each time through the loop like this ...

```
The total for file burglaries201407.csv is  1; the overall total is  1
The total for file burglaries201408.csv is  1; the overall total is  2
The total for file burglaries201409.csv is  4; the overall total is  6
The total for file burglaries201410.csv is  1; the overall total is  7
The total for file burglaries201411.csv is  3; the overall total is 10
The total for file burglaries201412.csv is  7; the overall total is 17
The total for file burglaries201501.csv is  4; the overall total is 21
The total for file burglaries201502.csv is  9; the overall total is 30
The total for file burglaries201503.csv is  7; the overall total is 37
The total for file burglaries201504.csv is  7; the overall total is 44
The total for file burglaries201505.csv is  4; the overall total is 48
The total for file burglaries201506.csv is  2; the overall total is 50
The total for file burglaries201507.csv is  3; the overall total is 53
The total for file burglaries201508.csv is  9; the overall total is 62
The total for file burglaries201509.csv is  6; the overall total is 68
The total for file burglaries201510.csv is  0; the overall total is 68
The total for file burglaries201511.csv is  2; the overall total is 70
The total for file burglaries201512.csv is  1; the overall total is 71
```

10. **[EXTRA for EXPERTS - NO MARKS]**

Write a loop to print out the names of all Area Units in which the maximum number of burglaries occurred each month.

Your output should look like this:

```
For file burglaries201407.csv, the max (11) occurred in Porirua Central
For file burglaries201408.csv, the max ( 8) occurred in Penrose and Dannemora
For file burglaries201409.csv, the max ( 9) occurred in Russell
For file burglaries201410.csv, the max ( 7) occurred in Auckland Central East, Kingsland, and St Albans East
For file burglaries201411.csv, the max (10) occurred in Petone Central
For file burglaries201412.csv, the max (12) occurred in Highbrook
For file burglaries201501.csv, the max ( 7) occurred in Three Kings, Papatoetoe Central, Manukau Central, and Halswell West
For file burglaries201502.csv, the max ( 9) occurred in Porirua Central
For file burglaries201503.csv, the max ( 8) occurred in Tahunanui
For file burglaries201504.csv, the max (10) occurred in Fairfax
For file burglaries201505.csv, the max (10) occurred in Greenmount
For file burglaries201506.csv, the max (10) occurred in Manukau Central
For file burglaries201507.csv, the max ( 7) occurred in Leabank
For file burglaries201508.csv, the max ( 9) occurred in Manukau Central
For file burglaries201509.csv, the max ( 8) occurred in Kingsland
For file burglaries201510.csv, the max ( 7) occurred in Clendon South, Nawton, Karamu, and St Kilda East
For file burglaries201511.csv, the max (10) occurred in Te Rerenga
For file burglaries201512.csv, the max ( 8) occurred in Red Hill
```

NOTE: the how the grammar varies with the number of Area Units being reported (the use of commas and "and").