Assignment 1 – Manipulating Data

Q1. Use the data set "Airquality" in R to answer the following questions:

(1) What are the dimensions of the data?
(2) What are the column names?
(3) Remove those rows with missing data.
(4) Find the sub set of the data where the "Month" is 5.
(5) How many rows of data are there with "Wind" between 7 and 8 inclusive?
(6) *Suppose there is an air index calculated using the following formula

$$index = \frac{Solar.R \times Wind}{Temp}$$

Add a column named "Index" with the value calculated with the above formula to Airquality data set.
(7) *Find the sub set of the new Airquality data with only "Solar.R", "Index", "Month", and "Day" columns and without missing value.
(8) Find the last "Day" of each "Month".
(9) Find the maximum and minimum temperature of each month.

Q2. Construct a data frame containing all countries in the world and the corresponding continents by crawling data from http://statisticstimes.com/geography/countries-by-continents.php.

Q3. The file "MovieData.csv" lists financial data on movies released since 1980 with budgets of at least $20 million.

- Create three new variables, Ratio1, Ratio2, and Decade. Ratio1 should be US Gross divided by Budget, and Ratio2 should be Worldwide Gross divided by Budget, and Decade should list 1980s, 1990s, or 2000s, depending on the year of the release date.
- Find counts of movies by various distributors. Then create one more column, Distributor.New, which lists the distributor for distributors with at least 30 movies and lists Other for the rest.
- Show average and standard deviation of Ratio1, broken down by Distributor.New and Decade. Comment on it.
- Do the same for Ratio2

(Note: Submit your working as R Markdown file)