Practical DataScience
(Data Manipulation)
Assignment-1

Problem1

The data file rainfall.dat(available at algorithmica github repository) records hourly rainfall at a certain location in Canada, every day from 1960 to 1980. Answer the following questions:

1. Load the data set into `R` and make it a dataframe called `rain.df`. What command did you use?
2. How many rows and columns does `rain.df` have? How do you know? (If there are not 5070 rows and 27 columns, you did something wrong in the first part of the problem.)
3. What command would you use to get the names of the columns of `rain.df`? What are those names?
4. What command would you use to get the value at row 2, column 4? What is the value?
5. What command would you use to display the whole second row? What is the content of that row?
6. What does `names(rain.df) <- c('year', 'month', 'day', 0:23)` do?
7. Create a new column called `daily` which is the sum of the 24 hourly columns
8. Make a histogram of the daily rainfall amounts.

Problem 2

Use the read.table function to import data from the file “ISIT.txt” (available at algorithmica github repository) in a variable called DeepSea and answer the following questions:

- Use the *names, str, dim, head* functions to check the correctness of *DeepSea*.
- Use the *unique* function to check all possible stations in DeepSea.
- Extract the subset of *DeepSea* data from station 1 in a variable called *DeepSea.sta1*.
- Compute the following: how many observations were made at station 1? What are the minimum, mean, and maximum sampled depth at station 1?
- Compare the data at stations 1,2,3,4,5, which two stations have fewest observations?
- Create a new data frame called *DeepSea.clean* by omitting the two stations with fewest observations from *DeepSea* data. Show how many observations were left.
- Extract the subset of *DeepSea* data from month Aug, Sep, Oct of all years in a variable called *DeepSea.fall*. Show the number of observations.
- Extract the subset of *DeepSea* data that were measured at depths greater than 2000 meters (from all years and months) in a variable called *DeepSea.dep2000*. Show the number of observations.
- Extract the subset of *DeepSea* data that were measured at depths greater than 2000 meters in Aug, Sep, Oct of year 2001 in a variable called *DeepSea.dep2000.fall2001*.

Problem 3

Use the DeepSea1.txt and DeepSea2.txt files from algorithmica github repository to answer the following questions:

Practical DataScience
(Data Manipulation)
Assignment-1

1. Use the read.table function to import data from file DeepSea1.txt in a variable called DeepSea12
2. Use the read.table function to import data from file DeepSea2.txt in a variable called DeepSea2
3. Check the first a few examples of *DeepSea1* and *DeepSea2* and find the common identification column in the two datasets.
4. Merge *DeepSea1* and *DeepSea2* by the common identification column into a data frame called *DeepSea*. Compare the number of observations in *DeepSea, DeepSea1, and DeepSea2.*
5. Merge *DeepSea1* and *DeepSea2* into a data frame called *DeepSea.full* so that there is no observation missing from DeepSea2.

6. Show only the columns *Year, Month, Station, SampleDepth* in *DeepSea.*
7. Show the above four columns according to increasing *SampleDepth*.
8. Show the above four columns according to increasing *Station* and then decreasing *SampleDepth*.

9. Add two new factor columns (*fYear* and *fMonth*) to *DeepSea*, to convert the year and month data to factors.
10. Show the levels of the above two new factor columns.
11. Add a new factor column (fMonthName) to DeepSea, so that months are coded using names (e.g. April, August, March, October) instead of using numbers (e.g. 4, 8, 3, 10).
12. Use the *write.table* function to export only the columns *ID, Year, Month, Station, SampleDepth, fYear, fMonth, fMonthName* of *DeepSea* to a file called DeepSea.txt. Make sure you use '\t' as separator, include quotation marks for factors and there is no extra column for row names.

Problem 4

Use the Temperature.txt file available from algorithmica github repository to answer the following questions. The file contains temperature observations made at 30 locations along the Dutch coastline from 1990 to 2005.

Practical DataScience
(Data Manipulation)
Assignment-1

1. Use the read.table function to import data from file Temperature.txt in a variable called Temp.
2. Check the correctness of *Temp* data.

3. Calculate the mean temperature per month.
   Calculate the standard deviation of temperature per month.
4. Calculate the mean temperature per month and station.
   Calculate the standard deviation of temperature per month and station.
5. Calculate the mean temperature per month in year 1990.
   Calculate the standard deviation of temperature per month in year 1990.

6. Calculate the mean of *Salinity, Temperature, CHLFa* for all observations.
   Calculate the standard deviation of *Salinity, Temperature, CHLFa* for all observations.
7. Calculate the mean of *Salinity, Temperature, CHLFa* for observations in station 'DANT'.
   Calculate the standard deviation of *Salinity, Temperature, CHLFa* for all observations in station 'DANT'.

8. Show the min, 1st quartile, median, mean, 3rd quartile, max of *Salinity, Temperature, CHLFa* for all observations.

9. How many observations were made per station?
   How many observations were made per year?
   How many observations were made at each station per year?