Northeastern University

Course: DA5020

Assignment: Module 2 - Data transformation with dplyr

Total Points: 100

Date Due: Posted on Blackboard

Learning Objectives

In this assignment, you will learn how to:

- Manipulate data objects
- Filter, select, summarise, order data in a data object
- Derive domain knowledge on two datasets

Tasks

Please write this week's assignment in R or in R markdown. Questions that are not code should be in comments in an R file or as a text chunk in R markdown. Install the "gapminder" package in R and load the gapminder dataset to answer questions 6 - 10. Load the attached file surveys.csv file into R using the code below:

```
surveys <- read.csv("surveys.csv", header = T, sep = ",")
```

This loads the surveys.csv data into the surveys data frame. The fields in the data table are the following:

- Record_id : a unique number for each row in the table
- Month : Month when the observation was made
- Day: Calendar day the observation was made
- Year: Year the observation was made
- Plot_id: the area the measurement was taken
- Species_Id: species id, please see https://github.com/weecology/portal-teachingdb/blob/master/species.csv for more information
- Sex: sex of the observation Male or Female
- Hindfoot: length of the hindfoot
- Weight: weight of the animal

Answer questions 1-5 using this data frame.

1. Write R code to extract the survey observations for the first three months of 1990 using the filter() function. (5 points)

2. Sort the 1990 winter surveys data by descending order of record ID, then by ascending order of weight. (10 points)

3. Write code that returns the record_id, sex and weight of all surveyed individuals of Reithrodontomys montanus (RO), (10 points)

4. Write code that returns the average weight and hindfoot length of Dipodomys merriami (DM) individuals observed in each month (irrespective of the year). Make sure to exclude NA values. (10 points)

5. Write code that determines the number of individuals by species observed in the winter of 1990. (15 points)

The following questions are questions on the gapminder data. Please review the description of the data using RStudio. The gapminder data is not a data frame, you need to use the as.data.frame() function to convert it to one. The fields on the data frame are the following:

- Country: the country the statistics are collected for
- Contiient: the continent where the country resides
- Year: the year when the statistics were collected
- LifeExp: the life expectancy for a person living in that country in that particular year
- pop : the population for the country in that particular year
- gdpperCap: the GDP per capita (person) GDP is gross domestic product, the total economic output of a country, i.e., the amount of money a country makes. GDP per capita is the total output divided by the number of people in the population. This measure provides an average output of each person, i.e., the average amount of money each person makes.

6. Create a dataframe named gapminder_df and mutate it to contain a column that contains the gross domestic product for each row in the data frame. (5 points)

7. Calculate the Mean GDP for Cambodia for the years within the dataset. (15 points)

8. Find the year with the maximum life expectancy for countries in Asia and arrange them in descending order by year, The result should contain the country's name, the year and the life expectancy. (15 points)

9. Count the number of observations per continent. (5 points)

10. Compute the average and median life expectancy and GDP per capita by continent for the years 1952 and 2007. Should we be optimistic given the results? (10 points)

Deliverables

You need to submit either an R or an .Rmd extension file. Be sure to state all the assumptions and give explanations as comments in the .Rmd file wherever needed to help us assess your submission. **Please name the submission file LAST_FirstInitial_2.Rmd** for example for John Smith's 2nd assignment, the file should be named **Smith_J_2.Rmd**.