Homework 1- R Exercise

Instructions: You may need to google search, data.table package documentation, and other resources to answer some of questions. Please see hints for commands to look up. Also feel free to help each other on eLearning general discussion board.

1. The World Health Organization (WHO) collected data in countries across the world regarding the outbreak of swine flu cases and deaths in 2009. The data in the file SwineFlu2009.csv include counts per country by month during the epidemic. There are many variables in the raw data file with the following descriptions:

- observation_id: Unique observation id
- firstcase_date_id: ID for sorting by first case date
- firstcase_continent_id: ID (X.YY) for sorting by first case date within a continent where X represents continent X, and YY represents the YYth country with the next first case
- country: Country
- firstcasereport_date: Date of first case reported
- cum_case_[MONTH]: Number of cumulative cases reported on the first day of the month for April, May, June, July, and August (across the columns, respectively)
- cum_case_Aug09: Last reported cumulative number of cases reported to WHO as of August 9, 2009
- firstdeath_date_id: ID for sorting by first death date
- firstdeath_continent_id: ID (X.YY) for sorting by first death date within a continent where X represents continent X, and YY represents the YYth country with the next first death
- firstdeath_date: Date of first death

- cum_death_[MONTH]: Number of cumulative deaths reported on the first day of the month for May, June, July, August, September, October, November, and December (across the columns, respectively)

Your task is to read this data file into R properly

a. First, examine the raw data file SwineFlu2009.csv using Excel.

b. Read the data to memory using fread(). Examine the data in Rstudio.

c. Then, assign the proper variable name to each variable. Make sure that each variable is assigned the correct type – character or numeric. (hint: use colClasses() to examine the class of columns)

d. In R, dates can be stored as a special type of numeric data. Modify the DATA step to make sure that the dates are read in the correct R date format (not as character). (HINT: Use the correct date type format statements in as.Date(), e.g., format = “%m/%d/%Y”)

e. Calculate the date difference of the firstcasereport_date variable from the first case report date across the world, which is Apr 24, 2009

f. Subset the columns (“firstcase_date_id”, “country”) and the answer from the above question 1.e, and save it as the file “SwineFlu2009_days_from_first_incidence.csv”) using fwrite(). (HINT: the new csv file should have three columns)

(Hint 1: Read the “help (documentation)” of the fread() command in data.table package carefully.)

2. A gourmet pizza restaurant is considering adding new toppings to its menu. Each month they survey 10 customers about their preferences for three different toppings. They want data on several different toppings, so they don't always ask about the same three toppings. Customers rate each topping on a scale of 1(would never order) to 5 (would order often). The restaurant wants to compute average ratings for all toppings, so the ratings variables need to be numeric. The raw data file Pizza.csv has variables for the respondent's ID, and the ratings for five different toppings: arugula, pine nuts, roasted butternut squash, shrimp, and grilled eggplant. The first two digits in the ID correspond to the month of the survey.

a. Examine the raw data file Pizza.csv and read it into R using fread().

b. Print the data set (on the Console).

c. Examine the class of each column of data.

d. Print the summary statistics of the data using describe() in "psych" package.

e. Open the raw data file in a simple editor like WordPad and compare the data values to the output from part b) to make sure that they were read correctly into R. In a comment in your report, identify any problems with the R data set that cannot be resolved using the fread(). Explain what is causing the problem.
(Hint: You need to make sure the type of each variable is read correctly.)

f. Read the same raw data file, Pizza.csv, again. This time, make sure the issues you've identified in the previous step ls resolved.

g. Create a column that contains the average ratings for each topping. (Hint: You need to make sure "NA" entries are not included in the average. They should not be treated as zeros. See the documentation for rowMeans().)

3. The new management of a local hotel decided to update their recently acquired (and very outdated) property by installing wireless Internet service for their guests. They are also considering updating their billing system because the method used by the previous owner seems faulty. In order to conduct a billing analysis, they would like some calculations about the guests who stayed with them during the first part of February (this was the first month after the change of ownership). The raw data file Hotel.dat contains variables with information on room number, number of guests, check-in month, day, year, check-out month, day, year, use of wireless Internet service, number of days of Internet use, room type, and room rate.

a. Examine the raw data file Hotel.csv and read it into R using fread(). Is there any "problem" with this data read? Explain.

b. Assign the column names for room number and number of guests first. For other column names, you should assign them as you answer the remaining questions.

c. Create date variables for the check-in and check-out dates, and format them to display as readable dates.

(Hint

Step1: You need to combine three columns into one that looks like a date: for example 2 /7 /2014. See the documentation for paste() and paste0().

Step2: If you do step 1, you have a column that has the date. Now you need to let R know that this is actually a date. Note that R does not realize this by itself, as you have seen in question 1.

Step3: Dates are saved as numbers in R.

d. Using the data.table syntax, create a column of days of internet use. If the guest did not use the internet, assign "0". Check the class of the column you created and coerce the variable type to "numeric" as necessary. (Hint. Days of internet use is recorded only when the use of wireless internet service is YES. See the documentation for as.numeric() and as.character())

e. Using the data.table syntax, create a column of room type. (Again, use the hint from the above)

f. Using the data.table syntax, create a column of room rate. Check the class of the column you created and coerce the variable type to "numeric" as necessary. (Again, use the hint from the above)

g. Subset the cleaned variables only and create a new data.table: room number, number of guests, check-in date, check-out date, use of wireless Internet service, number of days of Internet use, room type, and room rate.

h. Create a variable that calculates the subtotal as the room rate times the number of days in the stay, plus a per person rate ($10 per day for each person beyond one guest), plus an Internet service fee ($9.95 for a one-time activation and $5.95 per day of use).

(Hint1: You can subtract dates if they have been stored as R dates. Make sure you coerce the day difference column as numeric so you don't get weird outputs.

Hint2: You may want to create intermediate variables as necessary)

i. Create a variable that calculates the grand total as the subtotal plus sales tax at 8.75%. The result should be rounded to two decimal places.

j. View the resulting data set. In a comment in your report, state the value for the grand total for room 247, checked in on Feb. 7th, 2014.