Jefferson Bercaw
BME 580
Homework 1
1/31/2023

HW1_SP23_BME580

2023-01-25

```
## SECTION 1: Using R to calculate basic summary statistics of a dataset
# As you go through this homework, remember that reading the documentation for a given
# function can be extremely helpful. For windows, you can simply highlight a function
# in rstudio and use f1 to bring up the documentation.

# For much of this homework, we'll be using functions from an in-built R package, dplyr,
# which allows for straightforward data manipulation (dplyr documentation:
# https://dplyr.tidyverse.org/). Please note that dplyr is included in the package
# tidyverse, thus loading tidyverse instead of dplyr will also work.

# For this section, we'll use the heart disease health indicators dataset which contains
# information collected by the CDC obtained from a yearly telephone survey (The Behavioral
# Risk Factor Surveillance System). This data is a selection of responses collected in 2015
# and is available from Kaggle.

# 1. Load in the heart disease health indicators dataset from Sakai into R and print the
# first 10 rows (2 points)



# 2. How many rows and columns are present in the dataset? (2 points)
```

There are 253680 rows and 22 columns in this data set.

```
# 3. Calculate summary statistics for the heart disease dataset.
# Hint: The base function "summary" may be useful. (2 points)



# 4. Subset the original dataset into two dataframes, one containing every individual with
# heartDiseaseorAttack of 1 (yes) and another for individuals without record of heart
# disease or attack (0, 'no'). The two subsets should include all of the information from
# the original. The dplyr function filter() is a good way of doing this. You'll notice
# that filter works by cutting rows.You could also use the dplyr select() function to
# subset datasets by their columns. (4 points)



# 4a. Calculate summary statistics for both datasets you just created. Name three columns
# where the summary statistics appear noticeably different between the two datasets?
# (2 points)
```

Three of the most different categories are HighBP, Diabetes, PhysHlth

```
# 5. How many unique values of the education and income variable are there? Why would the
# education variable be better encoded as a factor rather than a numeric/double? Using a
# for loop, calculate summary statistics for each unique value of education. (7 points)
```

There are 6 unique values for education and 8 unique values for income. The education variable would be best encoded as a factor so data scientists using our data set would know that it's a categorical variable as opposed to a numerical value that could also be used to describe education, such as length of time in school.

```
# 6. How many instances are there of men (sex=0), with education greater than 3 and a
# positive record of heart disease or attack? Feel free to use any function, but filter
# or count may be of interest (5 points)
```

There are 8984 instances of this group of people.

```
# 7. When exploring datasets, you may wish to create new categories based on existing
# variables. For example, you may want to explore how the data changes between broad
# education levels. Add an additional column to the dataset that provides an education
# classification based on the education value of the subject. (12 points)
#   1-3 -> HighSchool
#   4-5 -> College
#   6 -> Graduate_or_professional



# 8. Some differences between education level may be intuitive. For example, we may
# suspect that higher educational attainment would correlate with higher incomes.
# Determine whether this is represented in the dataset by comparing the average income
# values across the education levels you just created. Do the same comparison across
# education levels for BMI. For each education level, provide the percent of individuals
# in that group who have a positive record of heart disease or attack.(10 points)
```

Average Income for Education Class 1 3.63, Average BMI for Education Class 1 29.59", For Education Class 1 the average percent of people with a positive record for heart disease or attack is 17.71
Average Income for Education Class 2 5.55, Average BMI for Education Class 2 28.96, For Education Class 2 the average percent of people with a positive record for heart disease or attack is 10.84
Average Income for Education Class 3 6.98, Average BMI for Education Class 3 27.52, For Education Class 3 the average percent of people with a positive record for heart disease or attack is 6.6

```
# 9. Sometimes it is necessary to convert chr or strings to factors for use in various
# functions or algorithms. Convert the column you created in question 7 to a factor.
# (3 points)



# 10. Write a function that selects a group of observations within the data based on an
# age range, education level, and smoking status, then returns the following: (10 points)
#        - Percent of individuals who have had a heart attack or disease
#        - Average BMI for both the positive and negative heart attack/disease groups
#        - Max and min PhysHlth of the selected group

# Be sure to include a test of your function!
```

Section 2: Understand and utilize shell script commands Working in R is sometimes limited by the amount of data that R can load and allow the user to efficiently work with. Hence, sometimes we interact with datasets using shell scripting, or bash commands. To complete this homework, you will need to access DCC and use the shell environment directly available from there. Follow the these steps to access DCC:

Connect to DukeVPN following the instructions in this website: https://oit.duke.edu/what-we-do/services/vpn Log into Duke Compute Cluster (DCC) by typing below in your terminal: ssh zq36@dcc-login.oit.duke.edu (Replace zq36 with your duke ID)

To receive full credits for each of the questions, you should copy the shell command you used to achieve the answer and also the printed answer from the shell terminal after each question. For cases where providing the entire answer (e.g. you are altering values across an entire dataset) isn't feasible, you may simply show a descriptive subset of the dataset that shows you have completed the task.

1. Upload the folder named "covid_data.csv.gz" that you can download from Sakai HW1 folder to your DCC home directory. (1 point)

2. Look at the first 5 lines and last 5 lines of the data file covid_data.csv.gz without unzipping it. Find out what "gzcat" or "zcat" does by running "man gzcat" or "gzcat –help" and use either to complete this task. (2 point)

3. How many rows and columns are there in this dataset? (4 points)

4. How many unique dates are there in this dataset? (7 points)

5. Other than 9999-99-99 (which represents that the patient survived), on which date were the most deaths recorded? (6 points)

5a. Knowing that 9999-99-99 represents that the patient survived, what percentage of patients in the dataset survived? (3 points)

6. Sometimes, it may help us to recode certain values in a file with different values that is easy to handle in R. Find and replace all of the 9999-99-99 date instances with NA then save the output to a new file. Awk and sed are both applicable here (8 points)

7. Currently, the data in this file is comma separated. Using tr, convert all commas into tabs to make this a tab delimited file. Be sure to save the output as a new file and print the first 5 lines (5 points)