~~~markdown
---
title: 'DTSC 650: Data Analytics In R'
subtitle: 'CodeGrade Final Project Part 2'
output: html_notebook
editor_options:
  chunk_output_type: inline
---

## Student Info

```
Name: Abigail Rhea
Term: Fall 2023
Date: 10/15/2023
```

------------------------------------------------------------------------

## General Instructions

------------------------------------------------------------------------

### Name of File

Name your assignment file **`BRFSS_Part2`**. This is a quarto "markdown" file, which has
the file has the extension '.qmd'.

------------------------------------------------------------------------

### Instructions

For the final section, you will choose four variables to explore in ways we have not in
Q1-Q9. You will choose one of those four variables as the response variable and the other
three as predictors. With those variables, complete the following. Be sure to read through
all of the instructions for Q10-Q14 before choosing your 4 variables. Feel free to create
multiple variables for Q12, Q13, and Q14, e.g. Q12a, Q12b, etc. Please make it clear,
though, the distinction between each question/problem with comments and spacing. If you
use Q12a, b, etc., be sure to print the results by doing Q12a or print(Q12a). Your answers
must be clearly identifiable. Take time to tidy your code once you are finished. The
easier it is for us to understand, the more partial credit you could receive.

------------------------------------------------------------------------

### Allowable packages

Allowable packages are `tidyverse`, `caret`, `Hmisc`, `lsr`, `olsrr`, `psych`, `lm.beta`.

-   If the allowable packages are not installed on your local computer, you'll need to do
a one-time installation *from the Console Window in RStudio* for each package like this:\
    **`install.packages('<package name>')`**\
    *Do not attempt to install packages in code that you submit to CodeGrade.*

-   Note: installing the entire tidyverse with `install.packages('tidyverse')` from the
Console Window will save you from having to install any of the tidyverse's individual
packages in the future.

-   In your code, load the package's library like this: **`library(<library name>)`**

------------------------------------------------------------------------

### Do / Do not

-   Do use tidyverse functions (dplyer verbs) for all of the questions where possible.
~~~

- Do use The Pipe.

- Do use plenty of comments throughout your code so that the grader can follow your line of thinking.

- Do not rearrange dataframe outputs unless specified by the question instructions.

- Do not create multiple copies of the BRFSS dataset in your script. Creating too many copies of the dataset can cause CodeGrade to crash with exit code -9. If you see that error on your Practice Submission, please check for this.

Data Set

- These data come from [Kaggle](https://www.kaggle.com/cdc/behavioral-risk-factor-surveillance-system).

- To answer these questions you will need to use the codebook on Brightspace, called `codebook15_llcp.pdf`. Please note that not all of the variables listed in the codebook are included in the .csv file to be downloaded from Brightspace.

- Download the `BRFSS2015_650.csv` file from Brightspace and place it in the same folder/directory as your script file. Then in RStudio, set your Working Directory to your Source File location: in the menus choose Session \| Set Working Directory \| To Source File Location. You most likely will see some warnings after it loads due to the fact that `read_csv()` will try to guess the column type but because there are so many rows it won't read enough of them to accurately make a guess.

- You must use the `read_csv()` function when loading the .csv file. Do not use read.csv().

- Do not rename the .csv file that you download from Brightspace.

- Do not edit the .csv file.

Pipe Notation

You may use the `tidyverse` pipe **`%>%`** or the new base R pipe **`|>`**. See [here](https://www.tidyverse.org/blog/2023/04/base-vs-magrittr-pipe/) for a comparison.

You are expected to use pipe notation in all of the CodeGrade assignments. Although there are alternate ways to filter, subset, and summarize data sets, using the pipe creates more readable code and is an important skill to develop.

Rounding requirement

Round all float/dbl values to two decimal places.

Dataframe vs. Tibble

Typically, in CodeGrade assignments, we expect output to be dataframes, not tibbles, unless otherwise noted.

Questions and Coded Solutions

~~~
### Preliminaries

```{r}
### It's always a good idea when working in RStudio to start with a clean environment.
### Clear objects from the environment memory that may be leftover from previous
###   versions of your code or other assignments by running the following line:
rm(list = ls())

### Load the libraries you need

library(tidyverse)
library(dplyr)
library(ggplot2)
library(olsrr)
library(tidyr)


### Load the Data
### Load the data file
brf <- read_csv("BRFSS2015_650.csv")

```

------------------------------------------------------------------------

## Questions

------------------------------------------------------------------------

### Q10

Address the values of each of the variables (use the codebook for this and include your
answer as comments). For instance, is "none" equal to a value other than 0? Are there
extra decimals implied? Are there other values that should be excluded based on the
responses they represent? Which variable will you seek to predict from the other
variables? Explain in your comments. Update the dataset you will use for the remainder of
the problems to account for these values.

```{r}
### Do not edit the following line. It is used by CodeGrade.
# CG Q10 #

### TYPE YOUR CODE BELOW ###

#Four variables:
#Response variable - GENHLTH (excellent, very good, good, fair, poor. I removed 7 for
Don't Know/Not sure, 9 for Refused, and BLANK for not asked or missing)
#Predictor variables -
#1. WEIGHT2 (weight without shoes). I only included weight in pounds by filtering in only
values greater than or equal to 50 and less than or equal to 0999. I excluded all other
values.
#2.PHYSHLTH (physical health not good on how many days during the past 30 days) I removed
77 for Don't Know/Not sure, 99 for refused, BLANK for not asked or missing, and changed
the value 88 to zero to repsent none.
#3.MENTHLTH (how many days within the past 30 days was mental health not good, i.e.
stress, depression and problems with emotions). I removed 77 for Don't Know/Not sure, 99
for Refused, and changed the value 88 to zero to represent none.
#My goal: To predict general health based on physical health, mental health, and weight in
pounds(lbs).


brf$MENTHLTH <- replace(brf$MENTHLTH, brf$MENTHLTH == 88, 0)
brf$PHYSHLTH <- replace(brf$PHYSHLTH, brf$PHYSHLTH == 88, 0)
health_prediction <- brf |>
~~~

~~~
  select(GENHLTH, PHYSHLTH, MENTHLTH, WEIGHT2, HEIGHT3) |>
  filter(GENHLTH != 7 & GENHLTH != 9 & GENHLTH != "BLANK",
         MENTHLTH !=77 & MENTHLTH != 99,
         PHYSHLTH != 77 & PHYSHLTH != 99 & PHYSHLTH != "BLANK",
         WEIGHT2 >= 50 & WEIGHT2 <= 0999) |>
  as.data.frame()
Q10 <- health_prediction





```
~~~

Q11

Remove any outliers for each applicable variable. Make sure you are updating the dataset from Q10 and using this dataset for the remainder of the problems. Briefly explain why you chose the method you used. Make sure to comment it out.

~~~
```{r}
### Do not edit the following line. It is used by CodeGrade.
# CG Q11 #

### TYPE YOUR CODE BELOW ###

weight_upper <- quantile(Q10$WEIGHT2, 0.9985, na.rm = TRUE)
weight_lower <- quantile(Q10$WEIGHT2, 0.0015, na.rm = TRUE)
weight_out <- which(Q10$WEIGHT2 > weight_upper | Q10$WEIGHT2 < weight_lower)
#WEIGHT2 has over 988 outliers present.

general_upper <- quantile(Q10$GENHLTH, 0.9985, na.rm = TRUE)
general_lower <- quantile(Q10$GENHLTH, 0.0015, na.rm = TRUE)
general_out <- which(Q10$GENHLTH > general_upper | Q10$GENHLTH < general_lower)
#No outliers in GENHLTH

physical_upper <- quantile(Q10$PHYSHLTH, 0.9985, na.rm = TRUE)
physical_lower <- quantile(Q10$PHYSHLTH, 0.0015, na.rm = TRUE)
physical_out <- which(Q10$PHYSHLTH > physical_upper | Q10$PHYSHLTH < physical_lower)
#No outliers in PHYSHLTH

mental_upper <- quantile(Q10$MENTHLTH, 0.9985, na.rm = TRUE)
mental_lower <- quantile(Q10$MENTHLTH, 0.0015, na.rm = TRUE)
mental_out <- which(Q10$MENTHLTH > mental_upper | Q10$MENTHLTH < mental_lower)
#No outliers in MENTHLTH

#After running outlier identification, I see there are no outliers in this dataset. I
chose this method because it is straightforward in detecting outliers above the 99.85%
quartile and below the 15% lower quartile.

#Percent remaining:
(nrow(Q10) - length(weight_out))/nrow(Q10)*100
#99.7 percent of values remaining after removing outliers from the WEIGHT2 column.

health_values <- Q10[-weight_out,] |>
  as.data.frame()
#updated data frame excluding outliers found in WEIGHT2

```
~~~

Q12

Complete exploratory analyses (for each variable) doing appropriate visualizations with ggplot2. Provide a discussion of your observations with comments

~~~
```{r}
### Do not edit the following line. It is used by CodeGrade.
# CG Q12 #

### TYPE YOUR CODE BELOW ###

#Histogram showing the count of people who stated their general health status.
plot1 <- ggplot(data = health_values) +
  geom_histogram(mapping = aes(x = GENHLTH), col='Blue', fill='Blue')

#Most people stated their general health status is 'Very Good' and 'Good'. Less people
stated their general health was fair or poor.

#Comparitive histogram showing the frequency of all four variables by using key-value
pairs.
plot2 <- health_values |>
  gather(key=Type, value=Value) |>
  ggplot(aes(x=Value,fill=Type)) +
  geom_histogram(position="dodge")

#Here is a bar graph showing the distribution of weight in pounds. The majority of
responses fall between 100 and 300 pounds.
plot3 <- ggplot(health_values) +
  geom_bar(mapping=aes(x = WEIGHT2), colour = 'pink')


value_factor1 = cut(health_values$WEIGHT2, 5)  #Changed the weight variable from numerical
to factor in order to show overlapping lines.
#This shows the number of days that a person experienced poor physical health, while
showing their weight. We can see that majority of people, with varying weights, experience
close to zero days of poor physical health in 1 month.
plot4 <- ggplot(data=health_values, mapping = aes(x=PHYSHLTH, colour = value_factor1)) +
  geom_freqpoly(binwidth = 0.1)

#Here is a box plot showing the distribution of mental health. Even though there weren't
any outliers in the mental health column post filtering, we can see that the median seems
to fall closer to the zero value. There are some values up to 30, but they are less often.
plot5 <- ggplot(data=health_values) +
  geom_boxplot(mapping=aes(x=MENTHLTH))


#Running a subset test to see which variable combinations give the best fit:
explore_regression <- lm(GENHLTH ~ PHYSHLTH + MENTHLTH + WEIGHT2, data = health_values)
health_subset <- ols_step_all_possible(explore_regression)

#The relationship between all three independent variables and the response variable gives
the R2 and adjusted R2 values. The Mallow's CP value for index 7 in the subset is also
closest to the number of predictors in the model above. I will use this information later
on when choosing the regression model.

```
~~~

Q13

Run basic descriptive statistics. Be sure to address each variable. What do these statistics reveal? Be sure to comment it out.

~~~
```{r}
### Do not edit the following line. It is used by CodeGrade.
# CG Q13 #

### TYPE YOUR CODE BELOW ###


summary(health_values)
~~~

~~~
#The summary function applies to health_values shows the min, median, mean, and max values
of each variable, as well as the 1st and 3rd quartile ranges.

standard <- lapply(health_values[, 1:4], sd)
#The function above shows the standard deviation of each variable in the data set,
representing the dispersion of the values within. The weight column shows the highest
standard deviation and general health shows the lowest.

variance <- lapply(health_values[, 1:4], var)
#The variance for each variable represents the squared difference between the observed
value and expected value. As with the standard deviation, the variance for the weight
variable is the highest, while general health is the lowest.

correlation <- round(cor(health_values), digits = 2)
#The correlation test for each variable reveals that the physical health predictor
variable(PHYSHLTH) has a positive correlation to general health status and is moderately
strong at 0.52. Mental Health also has a positive, but weak correlation to general health
status at 0.29. Weight in pounds has a weaker, but still positive correlation to general
health status at 0.17.

test1 <- cor.test(health_values$GENHLTH, health_values$PHYSHLTH)
#From the cor test between general health and physical health, I see that the p-value is
less than 0.05. This shows that there is significance in the hypothesis and we reject the
null hypothesis.

test2 <- cor.test(health_values$GENHLTH, health_values$MENTHLTH)
#The cor test on general health and mental health also show a p-value less than 0.05. This
relationship is significant and we can reject the null hypothesis.

test3 <- cor.test(health_values$GENHLTH, health_values$WEIGHT2)
#The cor test on general health and weight in pounds also shows a low p-value. This
relationship is significant and we can reject the null hypothesis.


```

### Q14

Finally, run at least 2 different, appropriate regressions predicting the variable you
indicated in Q10. These regressions should use different predictor(s). Identify the best
model and provide a discussion of your observations. Be sure to comment it out.

```{r}
### Do not edit the following line. It is used by CodeGrade.
# CG Q14 #

### TYPE YOUR CODE BELOW ###

#Single linear regression:
regression1 <- lm(GENHLTH ~ PHYSHLTH, data = health_values)
summary(regression1)
#Looking at the summary of linear regression between the predictor: physical health and
the response variable: general health status, I notice first that the residuals were
slightly lower than expected towards the left and very slightly higher than predicted on
the right. However, these values aren't so far off to assume the model didn't fit the data
well.
#The intercept of our response variable is ~2.27, while the slope of the predictor is
0.06. This says that as the mean of the value in general health status increases as the
value given in physical health increases.
#The standard error for the independent and dependent variables are both very low. This
shows me that there is little fluctuation in the results of this regression.
#The t-values for both the dependent and independent variables are very high and far away
from zero. This shows a relationship absolutely does exist between general health status
and physical health.
#As stated earlier in this project, the p-value is very low and indicates there is a
~~~

~~~
relationship between the two variables.
#The R2 value is at 0.2666, showing that the regression does not explain the variance in
general health status. I'd like to note that this regression only contains 2 variables,
which may contribute to the low R2 value.

#Multiple Regression:
regression2 <- lm(GENHLTH ~ MENTHLTH + WEIGHT2, data = health_values)
summary(regression2)
#Looking at the summary of multiple regression comparing mental health and weight to
general health status, I can see that the residuals are relatively symmetrical.
#This tells me that the model and the data fit well.
#The intercept of the response variable (general health status) is ~1.75 while the slopes
of mental health and weight in pounds are 4.1 and 3.7, respectively. This tells me that
the dependent variable increases in value as the two independent variables increase.
#The standard error for mental health is 2.13 and 3.56 for weight. This shows us the how
precise the models predictions are and how far away the values are from the line of best
fit. The highest calculated standard error is on the intercept, or dependent variable,
being general health status.
#The t-values are much higher than zero, which tells us that there is a relationship
between the two independent variables and the dependent variables. However, the t-values
in regression1 show a stronger relationship between general health status and physical
health.
#The p-value is well below the 0.05, stating that the relationship between the two
independent variables and the dependent variable is significant.
#The R2 value for this multiple regression is very low at 0.1108. This tells me that the
regression equation does not explain the variance in general health status. This
regression also has a low number of variables.

regression3 <- lm(GENHLTH ~ PHYSHLTH + MENTHLTH + WEIGHT2, data = health_values)
summary(regression3)

#Overall, regression3 provided the best model by showing the most significant relationship
out of the three independent variables I chose. The relationship between general health
status and physical health is the strongest, even though all independent variables have a
positive relationship with general health status. The R2 values are higher in regression3,
which is indicative of better predictor variables. Given that the t-value for physical
health is the highest, that further proves that physical health has the most impact on
general health status. The residual standard error is lower than the first two models,
which shows me that the response variable is less likely to deviate from the line of best
fit. The standard error of coefficients is also lower in regression3 than the previous
regression models. This means that the results of running regression3 are less likely to
change.



```
~~~

Before submitting to Code Grade:

1) Clear objects from your environment. Click the broom in the Environment pane in the top right. This will erase any variables (like Q1, Q2) that you've stored.

2) Rerun all your code. You can click the "Run" option above in this script pane (top right of notebook), then select "Run all". You should have all the variables stored again in the environment pane, and you should see no red error messages in the console below.

3) **Important**: You only have ONE ATTEMPT to submit Part 2 to the ACTUAL submission link! Ensure you are ready and confident in your work before submitting.
