Midterm GR5206 October 21, 2016

This is the computer portion of the midterm. When you are done, you should turn in a knitted .html file along with the raw .Rmd file on the Canvas page. Your files should be saved as "Midterm_UNI" (Midterm underscore UNI) where 'UNI' is replaced with your UNI. The head of your .Rmd file should look like the following:

```
---
title: "Midterm Section 00#"
author: "Name (UNI)"
date: "October 21, 2016"
output: html_document
---
```

where '00#' is replaced with your section number and 'Name (UNI)' is replaced with your name and UNI. Please clearly label the questions in your responses. The code for the responses should each be contained in separate code chunks. **Note that having the correct heading in your file and saving your file with the correct name will be worth 2 points in your final midterm grade.**

This part of the exam is open notes, so feel free to use any resource form the class: labs, homeworks, notes, etc. **However, you may not communicate with your peers during this portion of the exam. If we see you on your email or messenger you will automatically receive a 0 for this part of the exam.**

Question 1: Character Data

The file `rich.html` on the Canvas page is a listing of the 100 richest people in America, according to Forbes magazine (from 2013), which I scraped from `Forbes.com`. We will use the file to practice working with character data. The file `rich_dataframe.txt` is formatted version of the data in the `.html` file.

(a) Please load `rich.html` onto your computer using `readLines()` and save it as `rich`. Load `rich_dataframe.txt` using the appropriate command so that it's a data frame in `R` and save it as `rich_df`.

(b) Your task is to extract the net worths of people listed in `rich`. The lines that contain the net worths look like the following:

```
"\t\t<td class=\"worth\">$##,# B</td>"
```

except with the '#' values are replaced by digits. For example, the first two lines which hold Bill Gates' and Warren Buffet's net worth are

```
"\t\t<td class=\"worth\">$72 B</td>"
"\t\t<td class=\"worth\">$58,5 B</td>"
```

You can find the location of these lines by running the following command:

```
worth_lines <- grep("td class=\"worth\"", rich)
```

To convince yourself, make sure the length of the above output is indeed 100.

Your first step is to create a vector called `networths` that holds the values of net worths recorded in `rich`, in the same format. The first two values of your vector should be "$72 B" and "$58,5 B" and it should be of length 100. At the end, run the following command `networths[1:5]`.

(c) The Forbes website writes net worths in the form “$75,5 B” to mean $75,500,000,000$ dollars or 75.5 billion dollars.

Write code to convert from the Forbes format in your `networths` vector to numbers (in billions), and run it to create a numeric vector of net worths, called `networths2`. (If you are unable to create a `networths` vector in part (b), use the column `rich_df$Networths` as your starting point.) The first two values of your vector should be the numbers 72 and 58.5 and the vector should be of length 100. At the end, run the following commands `networths2[1:5]` and `identical(networths2, rich_df$Worth)`.

(d) Add the vector you created in (c) as a column to the data frame `rich_df`. Call the new column `MyWorth` and run the command

```
head(rich_df)
```

If you couldn’t complete part (c), add a column `MyWorth` to the data frame `rich_df` and fill the column with NA values.

Question 2: Plotting and KNN Classification

This question uses the data frame `rich_df` you imported in the previous question.

(a) In either base `R` graphics or `ggplot2` create a scatterplot with

- `Age` on the x-axis and `Worth` on the y-axis.
- points colored according to whether or not the billionaire's industry is technology (this information is in the `Technology` variable).
- the x-axis labeled '`Age`' and the y-axis labeled '`Net Worth`'.
- the title 'labeled '`Billionaire Net Worth vs Age`'.
- a legend describing the coloring of the points.

(b) `rich_df` holds data from 2013 and so isn't the most up-to-date. Let's imagine a new billionaire has come on the scene and we know that she is 55 years old and worth 30 billion. Let's add a new point to the graph that represents her. So your tasks are the following:

- add a new point to the graph that's colored purple at location (55, 30).
- *clearly* label the new point '`New Money`'.

(c) Let's also imagine that we don't know whether the new millionaire's industry is technology or not and we'd like to make a prediction using a KNN Classification scheme. Suppose that the Euclidean distances between the new point and the points in the `rich_df` dataset are stored in a vector called `dists` (meaning the first value in `dists` in the distance between the new point and Bill Gates' point, the second value in `dists` in the distance between the new point and Warren Buffet's point, and so on).

Now suppose we have the following information:

```
> head(order(dists), 10)
 [1] 12  6  8  7 22 21 23 13 14 25
> tail(order(dists), 10)
 [1] 46 98 60  2 37  1 29 71 68 50
```

Using the above information, what would be the predicted classification if we were using a KNN classifier with $K = 7$. You don't need to write code that outputs the answer, but you should need a line or two of code to give you enough information to find the answer.