Lab 4

Anya Conti

February 22, 2017

1. First I used readLines() to get the information from the website, and this was stored inside Forbes. To find the number of lines, I used length() on Forbes. To get the number of characters, I used nchar() on Forbes which returned a vector that stored how many characters were in each line of Forbes, and then this was summed to obtain the total number of characters. These are shown below.

```
Forbes <- readLines("http://people.math.umass.edu/~wei/Teaching/STAT597_Spring17/rich.html")
length(Forbes)
```

```
## [1] 1991
```

```
sum(nchar(Forbes, type="char"))
```

```
## [1] 80375
```

2. I used grep() on the Forbes data to obtain the line numbers which had the name "Bill Gates" and stored that in GatesLine (only 1 line). This was then repeated with "Warren Buffet" and stored in BuffetLine. Then I used those numbers as indices to obtain those particular lines of Forbes which had their names, as well as the lines right after them which is where their net worth was stored.

```
GatesLine <- grep("Bill Gates", Forbes)
BuffetLine <- grep("Warren Buffet", Forbes)

Forbes[(GatesLine):(GatesLine+1)]
```

```
## [1] "\t\t\t<h3>Bill Gates</h3></a></td>"
## [2] "\t\t<td class=\"worth\">$72 B</td>"
```

```
Forbes[(BuffetLine):(BuffetLine+1)]
```

```
## [1] "\t\t\t<h3>Warren Buffett</h3></a></td>"
## [2] "\t\t<td class=\"worth\">$58,5 B</td>"
```

3. First I used grep() on Forbes to obtain the line numbers that had with the particular formatting tags that went along with the only names (

), and used these as indices to obtain all the lines that had names in them, and stored them in NameLines. I similarly used that for the net worth, but added 1 to the indices obtained using that method, since the net worth is the line right below the name.

To get rid of the characters before name, I then used strsplit on NameLines, breaking each entry into 2 at the tag right before the name (), then I used unlist on this to turn it into a vector, and then used sequence to provide only even indices, and so only every other entry (the second part of the original line which is the part with the name in it) of this split vector was stored in FixedNames. I then did something similar to get rid of the characters on the end of the line. I used strsplit on NameLines, breaking each entry into 2 at the tag right after the name (), then I used unlist on this to turn it into a vector, and then used sequence to provide only even indices, and so only every other entry (the first part of the original line which is the part with the name in it) of this split vector was stored in FixedNames. After both of these, FixedNames only contained the Names on their own. This process was then used on worth, but this time a dollar sign "$" was used to split the stuff at the start from the data, and a "B" for the parts at the end.

The worth data however still had commas instead of periods. I then created a for loop which went through every entry of FixedWorth, and if a comma existed in that entry (grepl() returns TRUE if a comma at the i'th index of FixedWorth has a comma) then the if statement would carry out. If there is a comma, then first

the entry (the ith entry of FixedWorth) is split around the comma using strsplit(). This is then unlisted so the values themselves can be accessed, and the first value is obtained (the billions place, or the integer if in billions), made numeric with as.numeric(), and stored in int. The second value is obtained (the hundred millions place, or the decimal if in billions) with as.numeric(). Then the int plus .1 times the dec is stored in full, this is the full value of the billions. Then This new number is stored in the ith entry of FixedWorth again. Once the for loop is over, FixedWorth is made numeric. Then a new dataframe is created with two columns: One called Name which has the vector of FixedNames, and the other called NetWorth_Billions which has the vector of FixedWorth.

```
NameLines <- Forbes[grep("</h3></a></td>", Forbes)]
WorthLines <- Forbes[(grep("</h3></a></td>", Forbes)+1)]

FixedNames <- unlist(strsplit(NameLines,split="<h3>"))[seq(from=2,to=200,by=2)]
FixedNames <- unlist(strsplit(FixedNames,split="</h3>"))[seq(from=1,to=199,by=2)]


FixedWorth <- unlist(strsplit(WorthLines,split="\\$"))[seq(from=2,to=200,by=2)]
FixedWorth <- unlist(strsplit(FixedWorth,split=" B"))[seq(from=1,to=199,by=2)]

for(i in 1:length(FixedWorth)){
  if(grepl(",",FixedWorth[i])){
    int <- as.numeric(unlist(strsplit(FixedWorth[i], split=","))[1])
    dec <- as.numeric(unlist(strsplit(FixedWorth[i], split=","))[2])
    full <- int + (.1*dec)
    FixedWorth[i] <- full
  }
}

FixedWorth <- as.numeric(FixedWorth)

PeopleAndWorth <- data.frame("Name" = FixedNames, "NetWorth_Billions" = FixedWorth)

PeopleAndWorth
```

```
##                          Name NetWorth_Billions
## 1                   Bill Gates              72.0
## 2                 Warren Buffett            58.5
## 3                  Larry Ellison            41.0
## 4                   Charles Koch            36.0
## 5                     David Koch            36.0
## 6       Christy Walton & family             35.4
## 7                     Jim Walton            33.8
## 8                   Alice Walton            33.5
## 9                S. Robson Walton           33.3
## 10             Michael Bloomberg            31.0
## 11               Sheldon Adelson            28.5
## 12                    Jeff Bezos            27.2
## 13                    Larry Page            24.9
## 14                   Sergey Brin            24.4
## 15             Forrest Mars, Jr.            20.5
## 16               Jacqueline Mars            20.5
## 17                     John Mars            20.5
## 18                    Carl Icahn            20.3
## 19                  George Soros            20.0
## 20               Mark Zuckerberg            19.0
```

```
## 21                Steve Ballmer            18.0
## 22                Len Blavatnik            17.8
## 23              Abigail Johnson            17.2
## 24                  Phil Knight            16.3
## 25                Michael Dell             15.9
## 26                  Paul Allen             15.8
## 27                Donald Bren              14.0
## 28              Ronald Perelman            14.0
## 29            Anne Cox Chambers            13.5
## 30      Rupert Murdoch & family            13.4
## 31                   Ray Dalio             12.9
## 32                Charles Ergen            12.5
## 33                  Harold Hamm            12.4
## 34                James Simons             12.0
## 35  Laurene Powell Jobs & family           11.7
## 36                 John Paulson            11.4
## 37         Jack Taylor & family            11.4
## 38              Philip Anschutz            10.3
## 39               Richard Kinder            10.2
## 40                George Kaiser            10.0
## 41               Harold Simmons            10.0
## 42                  Andrew Beal             9.8
## 43                  Steve Cohen             9.4
## 44         Edward Johnson, III.             9.3
## 45          Patrick Soon-Shiong             9.0
## 46         Samuel Newhouse, Jr.             8.9
## 47        Charles Butt & family             8.5
## 48               Pierre Omidyar             8.5
## 49     Elaine Marshall & family             8.3
## 50           Hank & Doug Meijer             8.3
## 51                 Eric Schmidt             8.3
## 52              Donald Newhouse             8.2
## 53                 David Tepper             7.9
## 54                 Ralph Lauren             7.7
## 55           Stephen Schwarzman             7.7
## 56               Leonard Lauder             7.6
## 57             John Menard, Jr.             7.5
## 58              James Goodnight             7.2
## 59                    Eli Broad             6.9
## 60                Richard DeVos             6.8
## 61                  Jim Kennedy             6.7
## 62                  John Malone             6.7
## 63                    Elon Musk             6.7
## 64           Blair Parry-Okeden             6.7
## 65                David Duffield             6.4
## 66  Herbert Kohler, Jr. & family            6.4
## 67              Thomas Peterffy             6.4
## 68              S. Truett Cathy             6.0
## 69                 David Geffen             6.0
## 70                 Micky Arison             5.9
## 71              Sumner Redstone             5.8
## 72            Dennis Washington             5.8
## 73                Leslie Wexner             5.7
## 74                  Ray Lee Hunt            5.6
```

```
## 75              Charles Johnson             5.6
## 76       Richard LeFrak & family            5.6
## 77                Dannine Avara             5.5
## 78                 Scott Duncan             5.5
## 79                Milane Frantz             5.5
## 80            Jeffery Hildebrand            5.5
## 81           Rupert Johnson, Jr.            5.5
## 82                  Ira Rennert             5.5
## 83               Randa Williams             5.5
## 84              Stanley Kroenke             5.3
## 85                   Leon Black             5.2
## 86                   Gayle Cook             5.2
## 87              Dustin Moskovitz            5.2
## 88              Patrick McGovern            5.1
## 89                Charles Schwab            5.1
## 90       Jin Sook & Do Won Chang            5.0
## 91    Thomas Frist, Jr. & family            5.0
## 92                  David Green             5.0
## 93                Robert Rowling            4.9
## 94                  Stephen Ross            4.8
## 95                  Bruce Kovner            4.7
## 96                  Henry Kravis            4.7
## 97            Ann Walton Kroenke            4.7
## 98                  Gordon Moore            4.6
## 99                   Daniel Ziff            4.6
## 100                    Dirk Ziff            4.6
```