Project 7 Solutions

(Abhimanyu Agarwal)

Collaborators: N/A

TA help:

1) Melissa : Helped me go through Question 4 and 5.

Online resources used: N/A

Question 1

```
#Loads into two distinct dataframes using read.csv()
books <- read.csv("/class/datamine/data/goodreads/csv/goodreads_books.csv")
authors <- read.csv("/class/datamine/data/goodreads/csv/goodreads_book_authors.csv")

#Dimension for books dataframe
dim(books)
```

```
[1] 1000000      26
```

```
#Dimension for authors dataframe
dim(authors)
```

```
[1] 829529      5
```

Question 2

```
#Using the cut function to break into 4 categories
book_size <- cut(books$num_pages, breaks = c(0,250,500,1000,Inf), labels = c("small", "medium", "large"
table(book_size)
```

```
book_size
 small medium  large   huge
346804 283880  41828   3559
```

Question 3

```
#calculate the mean average_rating
tapply(books$average_rating, book_size, mean, na.rm = T)
```

```
   small   medium    large     huge
3.816630 3.863392 3.994815 4.203271
```

```
#calculate the mean text_reviews_count
tapply(books$text_reviews_count, book_size, mean, na.rm = T)
```

```
   small   medium    large     huge
19.16754 52.57758 57.66295 51.41585
```

```
#calculate the mean publication_year
tapply(books$publication_year, book_size, mean, na.rm = T)
```

```
   small   medium    large     huge
2007.623 2008.410 2006.426 2000.012
```

```
#From the mean of average rating, it can be observed that the books that fall under the 'huge'
#category have the highest average rating compared to all other categories

#From the mean of text reviews count, it can be observed that the books that fall under the 'large'
#category have the highest reviews count compared to all other categories.

#The average publication year shows 'huge' books were published in the early 2000's.
#Trends probably changed towards Medium sized books around 2008's.
```

Question 4

```
#Using lapply to perform same function as we did in part - 3
books_by_size <- split(data.frame(books$average_rating, books$text_reviews_count, books$publication_yea
lapply(books_by_size, colMeans, na.rm=T)
```

```
$small
    books.average_rating books.text_reviews_count   books.publication_year
                 3.81663                 19.16754                2007.62348

$medium
    books.average_rating books.text_reviews_count   books.publication_year
                3.863392                52.577575                2008.410163

$large
    books.average_rating books.text_reviews_count   books.publication_year
                3.994815                57.662953                2006.426014

$huge
    books.average_rating books.text_reviews_count   books.publication_year
                4.203271                51.415847                2000.011787
```

```
#Class
class(books_by_size)
```

```
[1] "list"
```

###Question 5

```
en_books <- books[books$language_code %in% c("en-US", "en-CA", "en-GB", "eng", "en", "en-IN") & books$pu
res <- subset(books, subset=(language_code %in% c("en-US", "en-CA", "en-GB", "eng", "en", "en-IN")) & (p
dim(en_books)  #325499 X 8
```

```
[1] 325499      8
```

```
dim(res)       #243269 X 8
```

```
[1] 243269      8
```

```
#Answer why dimension is different
#The dimensions are different because the en_books comprises of NA values which can be checked using th
#These NA values arent included on the creation of the dataframe 'res' using the subset command. This i
```

```
#command inherently eliminates the NA values from the newly created dataframe.
```

###Question 6

```
mymergedDF <- merge(res, authors, by="author_id")
dim(mymergedDF)
```

```
[1] 243269     12
```

###Question 7

```
df <- mymergedDF[mymergedDF$name == 'Bill Bryson',]
head(df)
```

```
   author_id  book_id average_rating.x
64         7  9349220             4.05
65         7 15829463             3.79
66         7  6562877             3.89
67         7 27905084             3.71
68         7  6857029             3.79
69         7  3022170             3.85

64
65
66 In the world of contemporary travel writing, Bill Bryson, the bestselling author of A Walk in the Woo
67
68
69
                                                                                 title
64           A Walk in the Woods: Rediscovering America on the Appalachian Trail
65                                                 Shakespeare: The World as a Stage
66 I'm a Stranger Here Myself: Notes on Returning to America After 20 Years Away
67            The Road to Little Dribbling: Adventures of an American in Britian
68                                                                     Shakespeare
69                                    Bryson's Dictionary: for Writers and Editors
   ratings_count.x language_code publication_year average_rating.y
64                7           eng             2010             4.01
65                4           eng             2012             4.01
66              611           eng             2008             4.01
67               39           eng             2016             4.01
68              141           eng             2009             4.01
69               19         en-GB             2009             4.01
   text_reviews_count        name ratings_count.y
64               61884 Bill Bryson         1014813
65               61884 Bill Bryson         1014813
66               61884 Bill Bryson         1014813
67               61884 Bill Bryson         1014813
68               61884 Bill Bryson         1014813
69               61884 Bill Bryson         1014813
```

```
dim(df) #43x12
```

```
[1] 43 12
```

```
max(df$average_rating.x) #For the books
```

```
[1] 4.22
```

```r
df[which.max(df$average_rating.x),]$title
```

```
[1] "A Really Short History of Nearly Everything"
```

```r
#I agree with "A Really Short History of Nearly Everything" to be the highest rated
#book written by Bill Bryson as it demonstrated to have the highest average rating of 4.22.
#It stood up because of its rating out of all the 43 books by the author
```

Submitting deliverables: project07.RMD, project07.R and project07.pdf

Pledge

By submitting this work I hereby pledge that this is my own, personal work. I've acknowledged in the designated place at the top of this file all sources that I used to complete said work, including but not limited to: online resources, books, and electronic communications. I've noted all collaboration with fellow students and/or TA's. I did not copy or plagiarize another's work.

> As a Boilermaker pursuing academic excellence, I pledge to be honest and true in all that I do. Accountable together - We are Purdue.