Project 7 Solutions

Kevin Choe

Collaborators: N/A

TA help: Katie Brinkers guided with problems 1-7, explained new concepts

Online resources used: Stat 190 Example Book (All problems)

Question 1

```
books <- read.csv("/class/datamine/data/goodreads/csv/goodreads_books.csv")
authors <- read.csv("/class/datamine/data/goodreads/csv/goodreads_book_authors.csv")
#Shows number of rows and columns
dim(books)
```

```
[1] 1000000      26
```

```
dim(authors)
```

```
[1] 829529      5
```

Question 2

```
#Breaks books into 4 Categories
book_size <- cut(books$num_pages, breaks=c(0,250,500,1000,Inf), labels = c("small","medium","large","hug
table(book_size)
```

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

Question 3

```
#Calculates the mean of each Category
tapply(books$publication_year, book_size, mean, na.rm=T)
```

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

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

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

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

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

The results did actually surprise me because there was not much of a difference between each category.

Question 4

```
#Created list of 4 dataframes
booksubset <- books[,c("publication_year","average_rating","text_reviews_count")]
books_by_size <- split(booksubset,book_size)
#Apply column means to each of new dataframe
lapply(books_by_size, colMeans, na.rm=T)
```

```
$small
  publication_year     average_rating text_reviews_count
       2007.62348            3.81663           19.16754

$medium
  publication_year     average_rating text_reviews_count
      2008.410163           3.863392          52.577575

$large
  publication_year     average_rating text_reviews_count
      2006.426014           3.994815          57.662953

$huge
  publication_year     average_rating text_reviews_count
      2000.011787           4.203271          51.415847
```

Question 5

```
#Creates an equivalent data frame of my own, by using the subset function
res <- subset(books, subset=language_code %in% c("en-US", "en-CA", "en-GB", "eng", "en", "en-IN") & publ

en_books <- books[books$language_code %in% c("en-US", "en-CA", "en-GB", "eng", "en", "en-IN") & books$pu
dim(en_books)
```

```
[1] 325499      8
```

```
dim(res)
```

```
[1] 243269      8
```

The differnece is that the subset function removes the NA rows while the other one does not.

Question 6

```
#Combines res and authors in a way which appends all information from author when there is a match in r
mymergedDF <- merge(res, authors, by="author_id")
dim(mymergedDF)
```

```
[1] 243269     12
```

Question 7

```
#Prints authors and looks into author's highest rated book after declaring an author
abigail <- mymergedDF[mymergedDF$name == "Abigail Thomas",]
abigail[which.max(abigail$ratings_count.x),]$title
```

```
[1] "A Three Dog Life"
```

I agree the book to be the highest rated book from the author because I personally have read that book and enjoyed it very much. I also looked up reviews on Google as well.

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.