STAT40620_FINAL_EXAM

Kamal Pradhan

17/12/2018

Question 1

(a) Loading the data

```r
load("goodreads.RData")
```

Total reviews in the dataset

```r
cat("Total reviews in the dataset",dim(goodreads)[1])
```

```
## Total reviews in the dataset 1048575
```

```r
dim(goodreads)[1]
```

```
## [1] 1048575
```

Total Unique users

```r
cat("Total different users in the dataset",length(unique(goodreads$user_id)))
```

```
## Total different users in the dataset 52927
```

```r
length(unique(goodreads$user_id))
```

```
## [1] 52927
```

(b) Book Authors and Book Review

Total book authors in the dataset

```r
cat("Total book authors in the dataset",length(unique(goodreads$authors)))
```

```
## Total book authors in the dataset 85
```

Most reviewed book

```r
freqTable<-as.data.frame(table(goodreads$title))
bookName <- freqTable$Var1[which.max(freqTable$Freq)]
#Most reviewed book
print(bookName)
```

```
## [1] The Hunger Games (The Hunger Games, #1)
## 110 Levels: 1984 ... Wuthering Heights
```

(c) Average ratings and others

Average rating per book

```
avgRatings<-aggregate(x = goodreads$rating,by = list(goodreads$title),FUN =mean)
colnames(avgRatings)<-c("BooK Title","Average Rating")
head(avgRatings)
```

```
##                                       BooK Title Average Rating
## 1                                           1984       4.046825
## 2 A Clash of Kings  (A Song of Ice and Fire, #2)       4.294218
## 3 A Game of Thrones (A Song of Ice and Fire, #1)       4.339880
## 4                           A Tale of Two Cities       3.782464
## 5                      A Thousand Splendid Suns        4.217698
## 6                                A Time to Kill        3.944559
```

Average rating per book = 5

```
cat("book with average ratings 5 = ",nrow(avgRatings[avgRatings$`Average Rating`==5,]))
```

```
## book with average ratings 5 =  0
```

Average rating per book > 4

```
cat("book with average ratings > 4 = ",nrow(avgRatings[avgRatings$`Average Rating`>4.0,]))
```

```
## book with average ratings > 4 =  44
```

reviewed > 10000 times and rating > 4.0

```
df2<-freqTable[freqTable$Freq>10000,]
df3<-avgRatings[avgRatings$`Average Rating`>4.0,]
df4<-merge(x = df2, y = df3, by.x = "Var1", by.y = "BooK Title")
cat("book with average ratings > 4  and reviewed atleast 10000 times = ",nrow(df4))
```

```
## book with average ratings > 4  and reviewed atleast 10000 times =  25
```

(d) Summarise

Average rating per book

```
class(goodreads) <- c("bookratings", "data.frame")
summary.bookratings <- function(x){

  cat("Top 10 average rated authors")

  tempdf<-aggregate(x = x$rating,by = list(x$authors),FUN =mean)
  tempdf<-tempdf[order(-tempdf$x),]
  tdf<-head(tempdf$`Group.1`,10)
```

```r
  print(tdf)

  cat("Top 10 average rated books reviewed atleast 10000 times")
  freqTable2<-as.data.frame(table(x$title))
  freqTable2<-freqTable2[freqTable2$Freq>10000,]
  avgRatings2<-aggregate(x = x$rating,by = list(x$title),FUN =mean)
  colnames(avgRatings2)<-c("BooK Title","Average Rating")
  df2<-merge(x = freqTable2, y = avgRatings2, by.x = "Var1", by.y = "BooK Title")
  df2<-df2[order(-df2$`Average Rating`),]
  ttdf<-head(df2$Var1,10)

  print(ttdf)

}

summary(goodreads)
```

```
## Top 10 average rated authors [1] J.K. Rowling               Kathryn Stockett
##  [3] Harper Lee                 George R.R. Martin
##  [5] Shel Silverstein           Markus Zusak
##  [7] William Goldman            Elie Wiesel, Marion Wiesel
##  [9] Maurice Sendak             Orson Scott Card
## 85 Levels: Aldous Huxley Alexandre Dumas, Robin Buss ... Yann Martel
## Top 10 average rated books reviewed atleast 10000 times [1] Harry Potter and the Deathly Hallows (Ha
##  [2] Harry Potter and the Half-Blood Prince (Harry Potter, #6)
##  [3] Harry Potter and the Goblet of Fire (Harry Potter, #4)
##  [4] Harry Potter and the Prisoner of Azkaban (Harry Potter, #3)
##  [5] The Help
##  [6] Harry Potter and the Order of the Phoenix (Harry Potter, #5)
##  [7] Harry Potter and the Sorcerer's Stone (Harry Potter, #1)
##  [8] A Game of Thrones (A Song of Ice and Fire, #1)
##  [9] To Kill a Mockingbird
## [10] The Book Thief
## 110 Levels: 1984 ... Wuthering Heights
```

```r
cat("Top three titles are Harry Potter and the Deathly Hallows (Harry Potter, #7),
 Harry Potter and the Half-Blood Prince (Harry Potter, #6),
 Harry Potter and the Goblet of Fire (Harry Potter, #4)")
```

```
## Top three titles are Harry Potter and the Deathly Hallows (Harry Potter, #7),
##  Harry Potter and the Half-Blood Prince (Harry Potter, #6),
##  Harry Potter and the Goblet of Fire (Harry Potter, #4)
```

```r
cat("Top three authors are J.K. Rowling, Kathryn Stockett, Harper Lee")
```

```
## Top three authors are J.K. Rowling, Kathryn Stockett, Harper Lee
```

Question 2

(a)

ii Logitstic regression

```
turtle<-read.csv("turtle.csv")

logitIt<-function(X,Y){

  model <- glm(X ~Y,family=binomial(link='logit'))
  fitted.results.cat <- ifelse(model$fitted.values > 0.5,"0","1")
  model$aic

}

logitIt(turtle$gender,turtle$length)
```

```
## [1] 53.02194
```

```
logitIt(turtle$gender,turtle$height)
```

```
## [1] 51.05099
```

```
logitIt(turtle$gender,turtle$width)
```

```
## [1] 39.74687
```

```
#length performs better
```