Machine Learning and Big Data: Problem Set 1 - Ian Bach

1. (5 points) Write an R function to simulate throwing a loaded 3-sided dice. More specifically, the function prints '1' with probability 0.5, '2' with probability 0.3, and '3' with probability of 0.2.

```
setwd("C:/Users/ibach/OneDrive - Terillium/Desktop/Purdue MSBA/Machine Learning/HW 1")

dice <- function() {
  # Use the sample function to draw a value from 1, 2, or 3 with specified probabilities
  result <- sample(c(1, 2, 3), size = 1, prob = c(0.5, 0.3, 0.2), replace = TRUE)

  # Print the result
  print(result)
}

# Example usage
dice()
```

```
## [1] 2
```

2. (5 points) Use the function in Question 1 to simulate rolling the 3-sided dice 10,000 times. How many times did '2' come up?

```
#Defin the funtion dice that will stimulate the rolling dice
dice <- function() {
  #use sample() to randomly select one of the number (1,2,3)
  #the probabilites of 0.5 for 1, 0.3 for 2, and 0.2 for 0.3
  result <- sample(c(1, 2, 3), size = 1, prob = c(0.5, 0.3, 0.2), replace = TRUE)
  #print the result for the roll (if needed)
  #print(result)
  #return the result for the roll
  return(result)
}
# Need to set a seed for the random number generator
set.seed(123)
#create rolling biased dice 10,000 times and store the results in 'rolls'
rolls <- replicate(10000, dice())
#count the number of times the result was a 2 in the 10,000 times rolled
count_twos <- sum(rolls == 2)
#print the count
count_twos
```

```
## [1] 2994
```

```
# Answer = 2994
```

3. (10 points) The dataset 'movies ratings.csv' contains ratings of 1,682 movies (columns) from 944 users (rows). Each row represents the ratings from one user. If a particular movie is rated by the user, a value from 1 to 5 is recorded (5 is the highest rating); if the movie is not rated, the entry is recorded as 0. (The file 'movies ids.txt' contains the names of the movies if you are interested.)

(a) Write an R script to count the number of movies the third user (third row) has rated and print the number.

```
MR <- read.csv("movies_ratings.csv")
#head(MR)

#Extract data for the 3rd users
third_user <- MR[3, ]

#count numnber of movies the 3rd users has rated. Also assume each rating is in a separate column and n
rate_movies <- sum(!is.na(third_user))

#print with the rate movies with the sentence presented
print(paste("The third user has rated", rate_movies, "movies"))
```

```
## [1] "The third user has rated 1682 movies"
```

(b) What is the fraction of rated movies (i.e., the number of non-zero entries divided by the total number of entries)?

```
#Count the number of non - zero entries
#0 represents a movie that has not been rated
NZ <- sum(MR != 0, na.rm = TRUE)

#Calculate total number of entries in the dataset
TE <- nrow(MR) * ncol(MR)

#Calculate the fraction of rated movies
FR <- NZ / TE

#Print the total within the sentence show
print(paste("The fraction of rated movies is", FR))
```

```
## [1] "The fraction of rated movies is 0.0629874644793325"
```

(c) Write an R script to compute the average rating of Toy Story (first column). [Hint: count only the non-zero entries.]

```
#Extract data from first column that is Toy Story
TS <- MR[, 1]

#Filter out the non zero ratings
NZTS <- TS[TS != 0]

#Create the average rating
AVGTS <- mean(NZTS, na.rm = TRUE)

#print with the pasted sentence
print(paste("The average rating for Toy Story is", AVGTS))
```

```
## [1] "The average rating for Toy Story is 3.88079470198675"
```

(d) For each movie, compute its average rating. Report the results (a vector of length 1,682) via a histogram.

```r
#Apply the function to each column of MR
AVG_RAT <- apply(MR, 2, function(column) {
 #Filter out the zero values from the column
  NZR <- column[column !=0]
  #Calculate the mean of the non-zero ratings ignoring any NA values shown
  mean(NZR, na.rm = TRUE)
})

#Shows the number of entries
print(length(AVG_RAT))
```

```
## [1] 1682
```

```r
#Create the histogram
hist(AVG_RAT,
     main = "AVG Movie Ratings", #Title
     xlab = "AVG Rating", # Label X axis
     ylab = "FREQ", # Label y axis
     col = "purple" # Color for the histogram
     )
```