For the below questions, submit two separate answers: (1) MS Word file that answers all the questions, and (2) R code file that you used to answer the questions. The MS Word file and R code file should be submitted in the same format of previous weeks’ assignments.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

For this assignment, please use the dataset named “5000-movie-dataset.zip”. The data contains information on 5000 movies and user feedbacks collected from online such as Facebook. To answer the questions, please use pipe operators, “dplyr” and “ggplot” packages. Use R base or other packages only for the questions that cannot be answered by “dplyr” and “ggplot” packages.

Q1) Review the variables to understand the context of the data. Include a screenshot that shows sample size, the number of variables, variable names, data types and a few snippets of each variables etc. Provide the code and screenshot. (5 points)

Q2) Draw an appropriate plot that shows the number of movies that have been released across all years. Provide the code and screenshot. (6 points)

Q3) Draw an appropriate plot that shows the changes of average review scores across all years. Provide the code and screenshot. (10 points)

Q4) Draw an appropriate plot that shows the frequency distribution for each categories of the movie content rating (i.e., content\_rating). Provide the code and screenshot. (10 points)

Q5) From the Q4,

1. What are the most frequent 4 categories of content rating, excluding “Not Rate” and “NA” categories? Answer it. (3 points)
2. Draw a boxplot that shows the review scores (i.e., imdb\_score) for the 4 most frequent categories of content ratings (i.e., content\_rating). Provide the code and screenshot. (7 points)

Q6) Using the same data that you used for Q5, draw a density plot that shows the densities of the review scores (i.e., imdb\_score) for the 4 most frequent categories of content ratings (i.e., content\_rating). The four density plots should be overlaid upon each other with different colors so that they can be visually compared. To overlay the density plots with different colors, set the transparency value as “alpha=.3”. Provide the code and screenshot. (14 points)

Q7) Draw an appropriate plot that shows the frequency distribution of Facebook likes (i.e., “movie\_facebook\_likes”). The frequencies of Facebook likes should be log-transformed. Provide the code and screenshot. (7 points)

Q8) Draw a scatter plot that shows the relation between the frequencies of Facebook likes and the movie review scores (i.e., imdb\_score). The frequencies of Facebook likes should be plotted as frequency values squared (i.e., transformed from *X* to *X*2, where X is the frequencies of Facebook likes). To the scatter plot, overlay a smooth line (i.e., geom\_smooth()) to better represent the relation. Provide the code and screenshot. (10 points)

Q9) To a ggplot(), overlay box plots to represent the changing frequencies of Facebook likes across all years (i.e., title\_year) that movie titles have been released. As in Q8, the frequencies of Facebook likes should be represented as squared frequency values. Provide the code and screenshot. (10 points)

Q10) Draw a scatter plot to show the relation between the frequencies of Facebook likes squared and the movie review scores (i.e., imdb\_score) for the movies that have been released after 2010 in the U.S.A. To the scatter plot, overlay a smooth line (i.e., geom\_smooth()) to better represent the relation. Provide the code and screenshot. (13 points)

Q11) Report the correlation value between the frequencies of Facebook likes squared and the movie review scores (i.e., imdb\_score) for the movies that have been released after 2010 in the U.S.A. Provide the code and answer. (5 points)