**DS 710**

**Homework 11**

**R assignment**

1. In this problem, you will use R to do further analysis on the Amazon reviews data.

1. In the Python assignment for homework 11, you created a .csv file with information about Amazon reviews.  **Use scan()** to read this data set into R.

* Do not read in the header row.
* It may be helpful to read in just a few rows at first, to make sure that the file is being read as you intended.

1. Convert the data into a matrix or data frame in which each row represents one review.  Read the header row into R and use it to create column names for the matrix or data frame.
2. Check whether the columns for total votes, review length, number of exclamation points, and helpful fraction are being treated as numeric vectors. If not, create new variables by converting them into numeric vectors.
3. Examine the helpful fraction vector for unrealistic values.  If you find any, set them to missing.  Also set to missing the corresponding value of the total votes vector.

(Because you computed the helpful fraction from the columns “helpful votes” and “total votes”, an unrealistic value of “helpful fraction” means that at least one of the other two values must be unrealistic.  Because we don’t know which one, the safest course is to set them both to missing.)

1. Write 1-2 sentences to document how many unrealistic values you found, what made them unrealistic, and the fact that you set those values to missing.
2. Create a new variable that describes whether more than 50% of people who voted considered it helpful.  We will call these helpful reviews.
3. Are helpful reviews longer than unhelpful ones?  Start by making appropriate graphical summaries to determine whether the review length should be transformed.  Then do a hypothesis test of whether the typical length of helpful reviews is longer the typical length of than unhelpful reviews.  State your conclusion in context.
4. In parts h-k, you will investigate whether products with more reviews tend to have more votes on their reviews. First, use tapply to find the maximum number of votes received by any of the product’s reviews.  Then count the number of reviews for each product ID (using tapply or another method you can think of).

(Hint to help you check your work:  The length of the resulting vector should be the total number of unique product IDs, which you can find using *length* and *unique*.  The sum of the number of reviews for each product ID should equal the total number of reviews.)

1. Make a scatterplot of max number of votes as a function of number of reviews.  Is there a visible trend?  If so, describe it.
2. Histograms of the review counts and number of votes indicate that both variables are right-skewed.  (You can check this for yourself.)  So, a log transformation might be helpful in investigating the relationship between them.  However, some product IDs had 0 votes, which would result in an error if we tried to take the log.  Subset the variables max.votes and number.of.reviews (or whatever you called them) to only those values corresponding to products with 1 or more votes.
3. Make a scatterplot of log(max.votes) as a function of log(number.of.reviews).   Is there a visible trend?  If so, describe it.  Does this tell us anything about the relationship between the untransformed max.votes and number.of.reviews?

Submit a report (in .docx or .pdf format) to GitHub containing your R script and output, including graphs, and written answers to the questions.