Exercise 1

Simple Statistics

1. Let the following data set be given (sample size 60):
   4, 3, 2, 5, 4, 6, 3, 7, 4, 1, 4, 0, 6, 4, 3, 5, 2, 3, 5, 1, 4, 4, 9, 5, 4, 3, 3, 5, 2, 4,
   3, 6, 5, 2, 6, 2, 4, 5, 5, 1, 5, 4, 4, 2, 7, 1, 3, 3, 4, 7, 3, 4, 4, 6, 6, 3, 3, 2, 6, 1.

   Calculate
   a. the mean
   b. the mean recursively
   c. the standard deviation over the sample

Simple Statistics on a Stream

2. Consider the Yelp dataset Challenge and the city assigned to your group. For the user reviews generate a feature vector comprising of the review_id, user who posted the review, the business for which the review was posted, the star rating of the review, the text of the review, the date of posting, the number of words in the text, the number of positive words and the number of negative words. We consider the reviews as a stream with the timepoint being the 1st of each month.
   Now answer the following questions
   a. For timepoint t_i, compute the number of reviews n_i seen so far.
   b. For timepoint t_i, how many positive/negative reviews have been seen till now? (HINT: We use a simple way to identify if the review is positive or negative. If the rating of the review is 2 stars or below it is a negative review)
   c. Compute the mean and standard deviation of the number of words, number of positive words and number of negative words seen in a review until t_i.
   d. For timepoint t_i, how many users have written a review so far?

   For the list of positive and negative words we will use a fixed list. The list of words can be downloaded from https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html under the section Opinion Lexicon (or sentiment Lexicon)

Sufficient Statistics

3. Discuss the disadvantages of the methods used to answer the questions in (2)

4. What is the importance of sufficient statistics in streams?

Sampling

5. During the lectures we discussed reservoir sampling. Another sampling technique could be where each instance of the stream has a 50% chance of being considered. Use this alternate sampling technique to count the number of users who have written reviews and compare your answer with (2).

Visualization

6. An important part of data mining is to be able to visualize the data. Use appropriate visualization techniques to represent the data in 2a, 2b, and 2d.