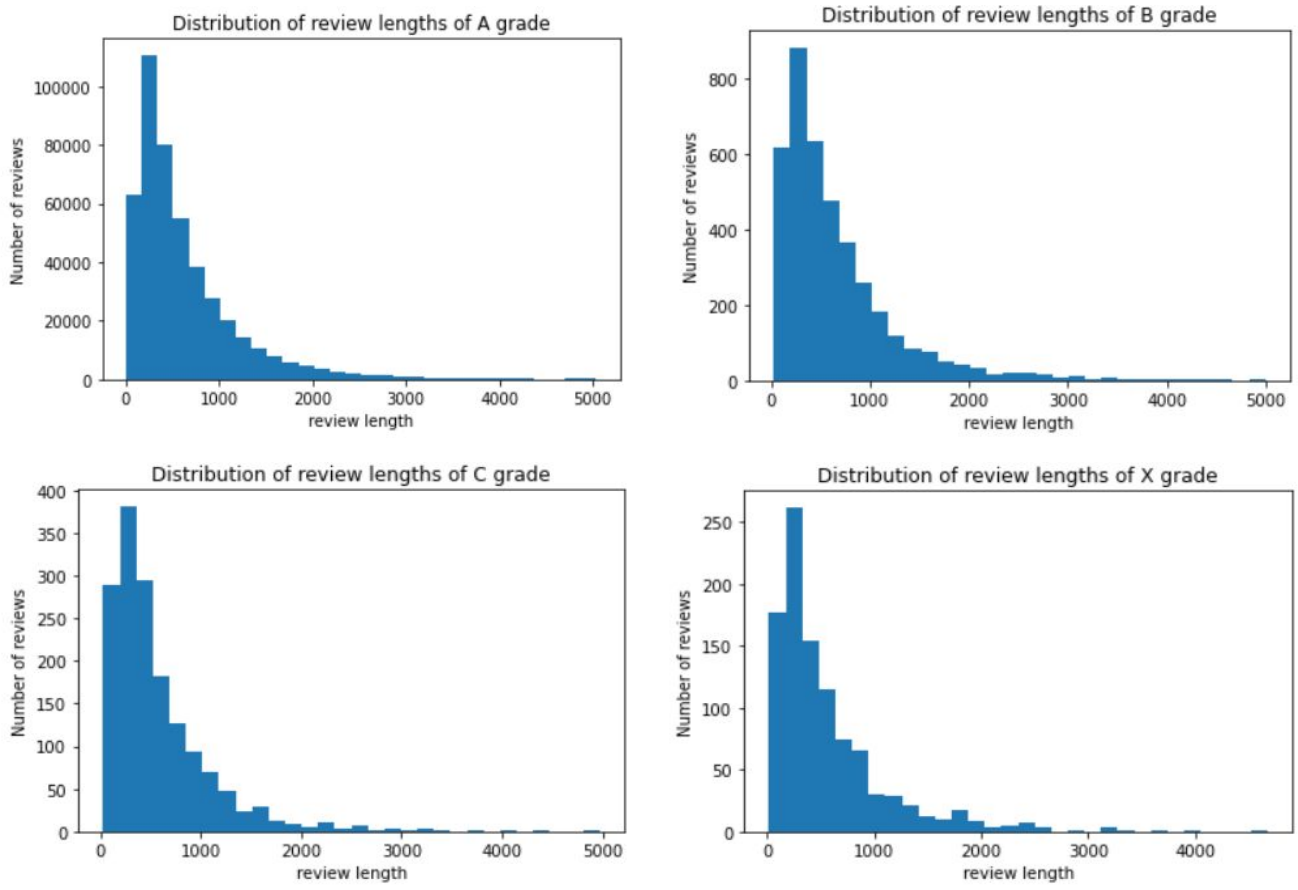
Capstone Project I - Statistical Data Analysis

What factors can determine a health inspection grade? One would think that if a patron has a negative experience, they may leave a lengthy response detailing what they were upset with or a short response because they don't want to waste time on something that left them a bad experience. Based on these two assumptions, there could be no clear cut expectations for length of reviews in determining a grade. However, what is the point of statistical data analysis if you do not use it to prove this point?

Firstly, it is important to separate the four inspection grades (A, B, C, X) to their individual dataframes. After the separation, it may be nice to view each grade's review count versus its length. Plotting each histogram reveals there is no visible difference among the grades.



In order to better understand if each dataset's review length is truly different from the other, it is necessary to perform a two sample t-test. The following parameters are used for each two sample t-test between the letter grade combinations: alpha of 0.05, null hypothesis that both grade's mean review length are the same, and alternate hypothesis that they are not equal. Running the 6 different combinations (${}_4C_2$), it is shown that the only combination where the null hypothesis cannot be rejected is when C and X are being compared. An assumption of why this could be the case is that restaurants with low C grades are very close to receiving an X grade, so the reviews for both are similarly negative.