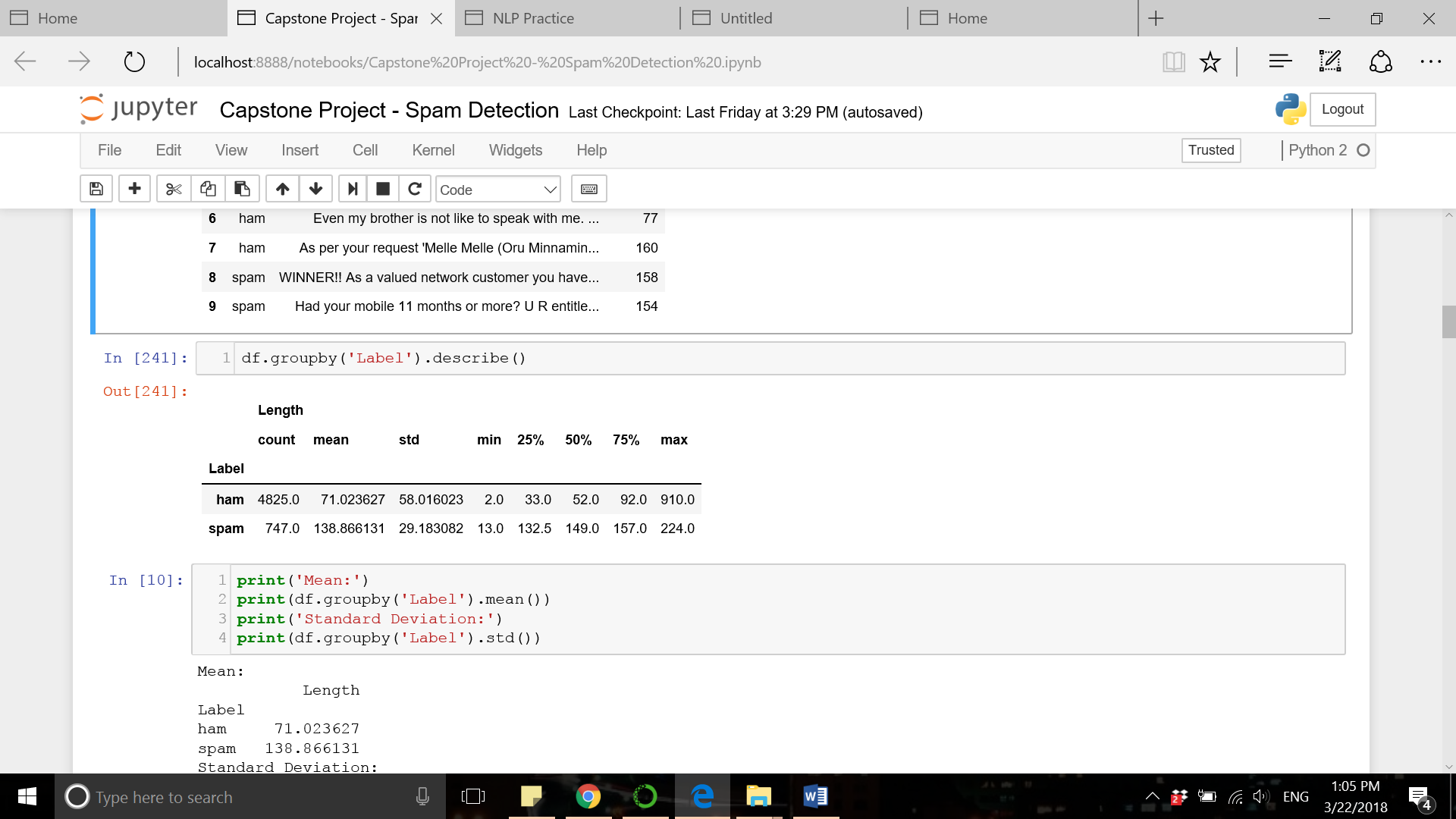


Capstone Project – Inferential Statistics

At this point, you have obtained the data set for your Capstone project, cleaned and wrangled it into a form that's ready for analysis. It's now time to apply the inferential statistics techniques you have learned to explore the data. For example, are there variables that are particularly significant in terms of explaining the answer to your project question? Are there strong correlations between pairs of independent variables, or between an independent and a dependent variable?

**Submission:** Write a short report (1-2 pages) on the inferential statistics steps you performed and your findings. Check this report into your github and submit a link to it. Eventually, this report can be incorporated into your Milestone report.

The question associated with this capstone project is which features in our data set could be used to detect spam messages from ham messages. The statistics about the length of each message clearly show that spam messages are longer than ham messages:



In addition to that, the word frequency data can provide further insights into the spam/ham messages and the features useful for detection purposes. The two charts below show the most frequent words in spam/ham messages and the total number of their occurrence at each group. It is evident that specific words are frequently found in spam messages which are used in ham messages less often. The most frequent word in spam messages is “free” which occurred 224 times while this word occurs only 60 times in ham messages (on average 1 out of 80 ham messages includes the word “free”, while 1 out of 3 spam messages includes this word). Therefore, vectorized representation of text messages created based on the vocabulary obtained from the existing corpus can provide the features for spam detection.

