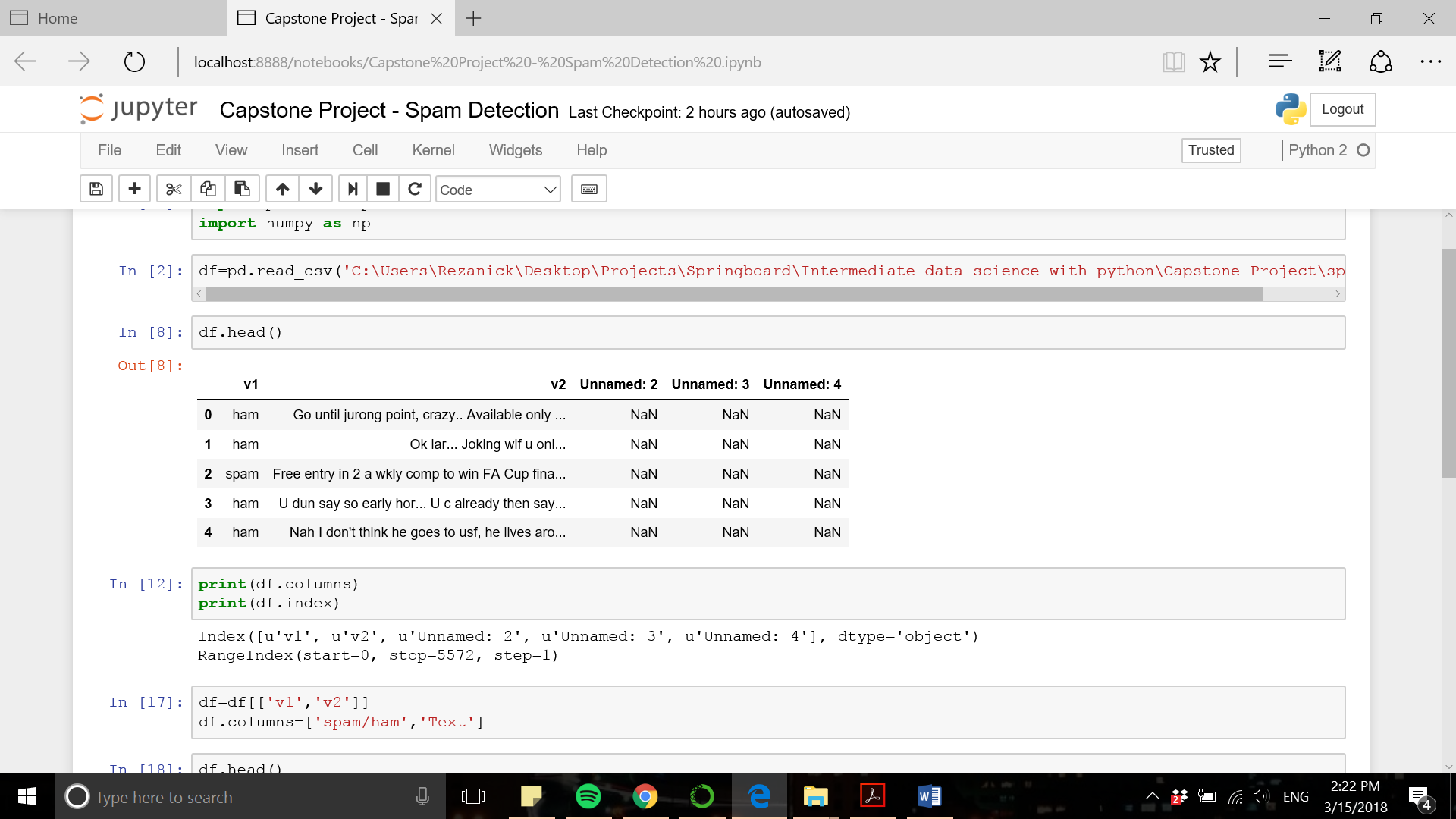


Capstone Project - Data Wrangling

Now that you have a basic ideas of the various data wrangling steps and techniques available, let's apply it to your Capstone Project! By now, you probably have a data set in mind for your project (If you don't have a data set yet, come back to this assignment once you have one). Apply some of the data wrangling techniques you have learned to this data set.

Submission: Create a short document (1-2 pages) in your github describing the data wrangling steps that you undertook to clean your capstone project data set. What kind of cleaning steps did you perform? How did you deal with missing values, if any? Were there outliers, and how did you decide to handle them? This document will eventually become part of your milestone report.

SMS spam detection dataset consisted of 5572 entries (4825 ham messages and 747 spam messages). As shown below, the column ‘v1’ represented the tag for each message labeled as ‘ham’ or ‘spam’. The column ‘v2’ consisted of text messages in string format. There are three redundant columns (‘Unnamed: 2’, ‘Unnamed: 3’, ‘Unnamed: 4’) with NaN values.



The first two columns (‘v1’ and ‘v2) are renamed into more meaningful column names (‘Label’ and ‘Text’, respectively) and the three redundant columns are dropped. The cleaned data frame is shown below and is ready for exploratory data analysis.

