Data Wrangling Applied to Capstone Project

There are some steps needed to clean the dataset:

1. Creating Dataframe
   1. There exists three datasets: cases, controls, and Dates/Clinics
   2. The cases and controls will need to be merged vertically.
   3. This will then need to be merged horizontally to the Dates/Clinics by the primary key(HMCID)
2. Remove any duplicates
3. Assess any missingness.
   1. Although the unit of measurement is blobs per record. It will be of interest to see if there are any time trends or confounding between the two given hospitals.
4. Extract year from dates
   1. To assess possible trend over a 3 year period.
5. Assess text for any text oddities
   1. After importing into python, certain characters such as bc(referring to because) as well as /t were found. Will need to eliminate these.
6. Remove medication information.
   1. It was determined by our project working group that medicine and biophysical measurements were not to be analyzed in this project. This information is grouped together in each blob and will need to be excluded.
7. Merge words into a phrase so they do not get counted as a stop word.
   1. The project working group came up with a list of phrases they believe are important.
   2. In order to prevent these from being eliminated after comparing to a stop-word list, these phrases will need to be protected by concatenating the words together that make up the phrase.
8. Add to stop-word list.
   1. Additional assessment of the blobs show extra words that can be eliminated.
   2. These words need to be added to the stop-word list.
   3. E.g.: IDCC , a, is, to, etc….