DATA CLEANING DEMO

04/15/2020

1. Fix typos, remove repeated rows, detect and fix (or remove) problematic values.

   -Typo essentially is caused by upper or lower case spelling

   -Specific issues to deal with:

   1). ROMS ID with different length of digits

   2). negative age, inconsistent age

   3). conflicts values in chronic pain for same person, same body region and classification, same period of therapy time

   4). inconsistent admission date and discharge date for one therapy given other conditions are the same. (holding some other criteria constant (ID, classification, body region, outcome):

   if the admission date is the same, discharge date is not, discharge date takes the last day of visit.

   if the discharge date is the same, admission date is not, admission date takes the first day of visit.

   5). remove the value of score out of range

   6). discharge date in June 2020

2. Give score either 1 or 0 based on the formula used in excel sheet

   -For pain result

   if pain relief > 2 or from less than 2 to 0, then assign 1

   -For overall outcome result

   if outcome LOWER EXTREMITY FUNC SCALE increases >=9 or increases from >71 to 80 (max), then good, assign 1, o.w., 0

   if outcome KNEE OUTCOME SURVEY, increases >= 9 or increases from >91 to 100 (max), then good, assign 1, o.w., 0

   if outcome in other 3 types , decreases >=10 or decreases from <10 to 0 (min), then good, assign 1, o.w., 0

3. Aggregate some therapy summary and evaluate by taking average.

   For same (ID, body region, classification and therapy time), if the categories of outcome are different, take the average of the pain improvement (1 or 0) and the average of overall performance (1 or 0). Since we aggregate some rows, some columns became invalid logically, so they are assigned NA.