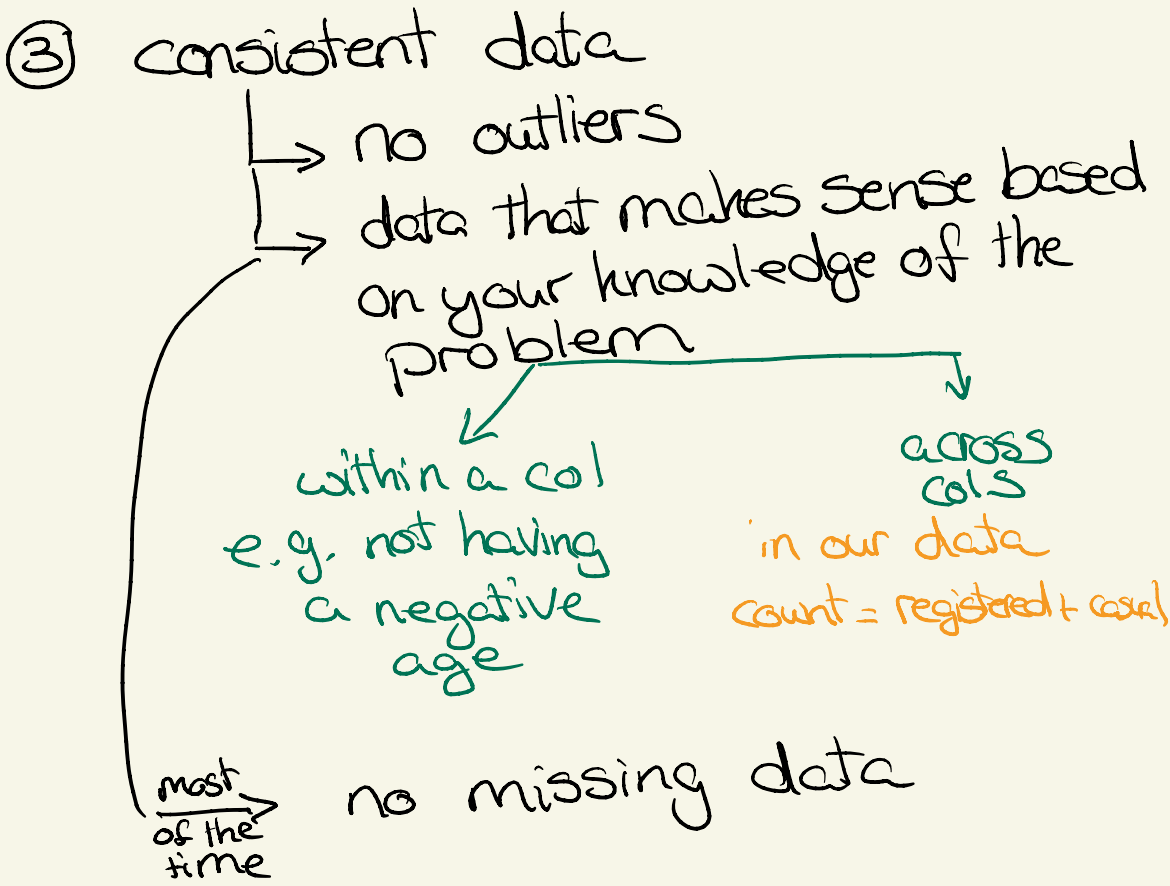
Process:

End goal is to have data that is ready for analysis (clean so that we don't have "junk in junk out")

1. Tidy data; R term but the concept extends to all software
   - ↳ put the data in a format that makes your analysis easier (ISA 245 ⇒ data model)
   - R →
     - every row is an observation
     - every variable is in one col
     - every cell a single value

2. Technically correct data
   - ↳ data understood correctly by the software
     - reasonable col name
     - correct classes for columns
     - categorical variable correct labels


3 consistent data
no outliers
data that makes sense based on your knowledge of the problem
within a col
e.g. not having a negative age
across cols
in our data
count = registered + casual
most of the time
no missing data