Data Cleaning Notes:

* Change names
  + Changed wt to bwt to make it clear it is the baby weight we are working with
  + Changed all of the mother’s variables to have an m in front to make clear it is the mother’s values we are talking about
* Remove columns
  + Removed id column since it is useless information that will have no explanatory relationship with bwt
  + Removed plurality since every value is the same – no twins or triplets
  + Removed outcome since all values = 1 and all survived the 28 days
  + Removed sex since all the data is for male babies and the value = 1 for all
* Rename unknowns as NA
  + Gestation unknown wasn’t in descriptor readme, but replaced 999 with NA and cleaned anyway
  + Marital unknown valued at 0 wasn’t in the descriptor readme, but given the how the things are coded, another potential coding which was missed cannot be surmised so this was given as an unknown
  + Time had a not asked value on top of the unknowns which was also changed to NA since it is not useful information
  + Number has 11 values for smoked but don’t know, so had to change those to NA since it isn’t going to be useful information except that they smoke
* Keeping the values for 10 in the Mrace and Drace since it would appear this could be another race factor which is left out of the descriptor, don’t know what it is, but it could be relevant information
* Combining mixed data
  + What needs to be combined
    - Mrace – 0-5 = white for some reason
    - Med – 6-7 = trade school HS unclear but don’t know why it has been split this way – as far as we are concerned they are the same
    - Drace – 0-5 = same as mothers’
    - Ded – 6-7 = same as mothers’
  + Set race values equal to 5 since that keeps the continuity with the following race values
  + Set ed values equal to 6 for similar continuity reasons