README

The training data (*labeled\_users.csv*) consists of ~4000 users and their Twitter user\_id, the year they are born (can be used to infer their age this year), gender and race.  
  
A few variables are missing for various reasons - these are 'nan' in the table. You will see fractional values for some attributes - this means that those users did the survey multiple times and entered different values for that item and the number is an average.  
  
For race, the coding is:  
1 - African-American  
2 - Latino/Hispanic  
3 - Asian  
4 - White  
5 - Multiracial  
NULL - didn't answer or other race (usually native american - these are all US users)