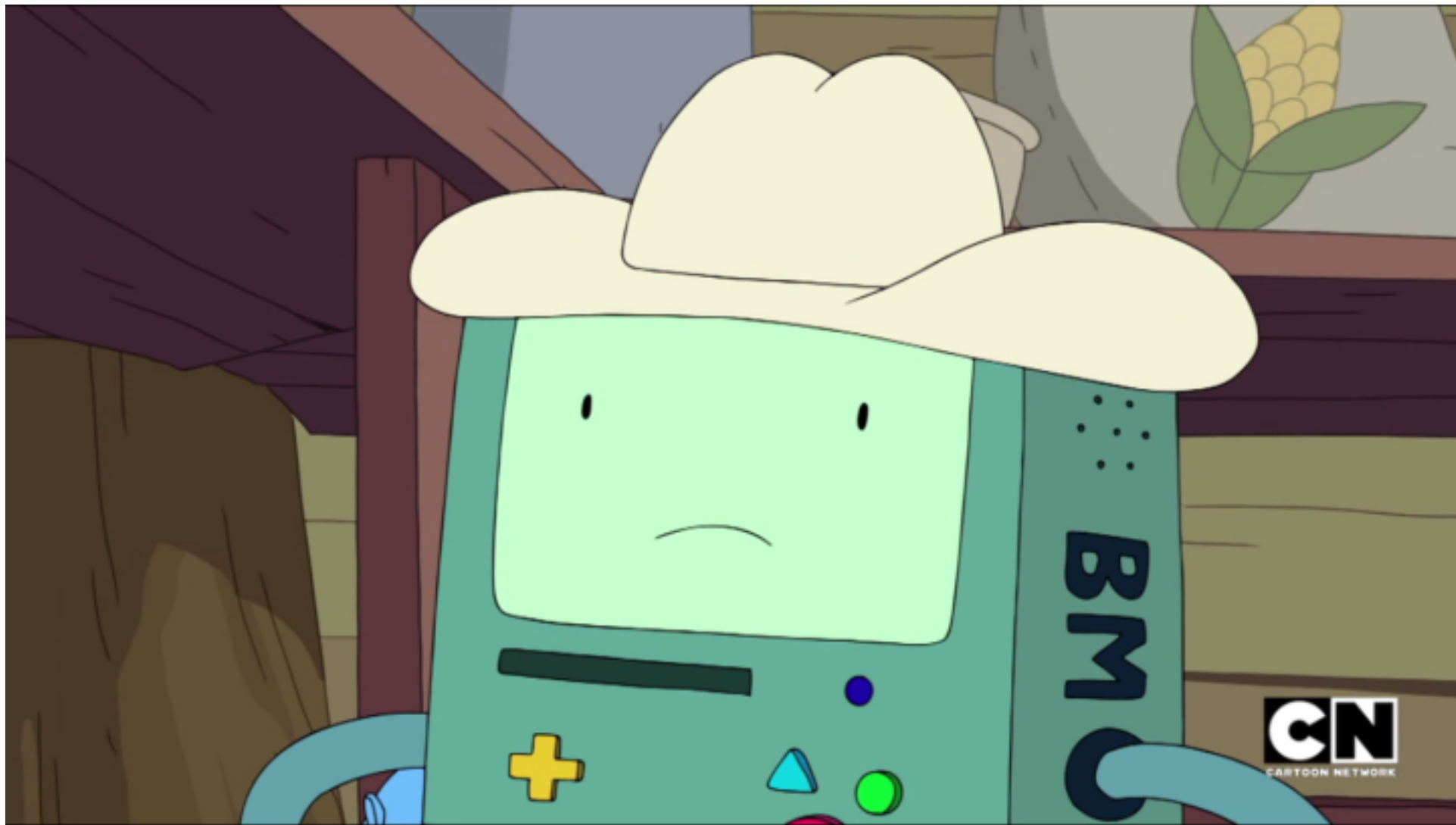


data wrangling in R



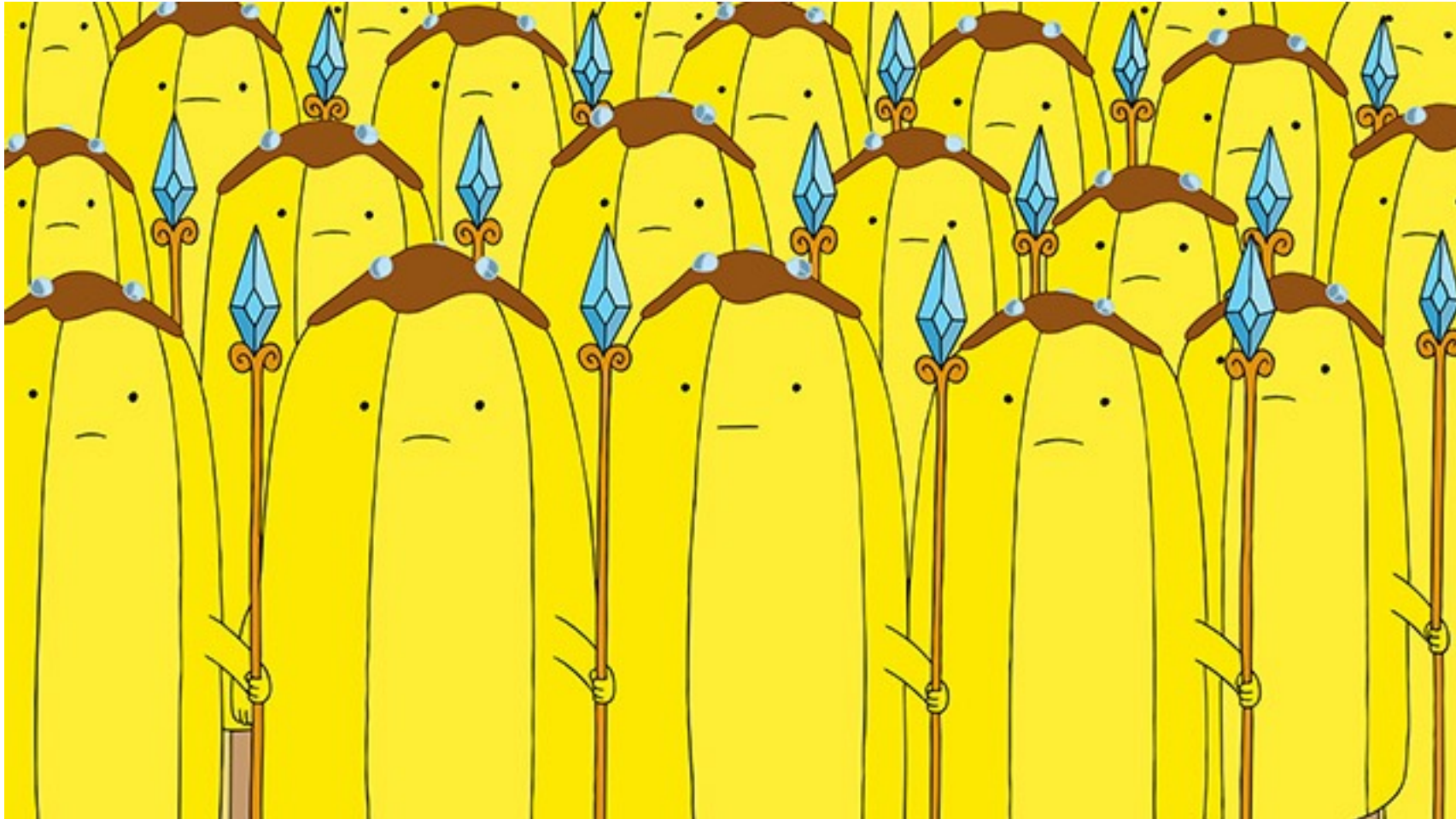
@cjlortie

more than half the battle



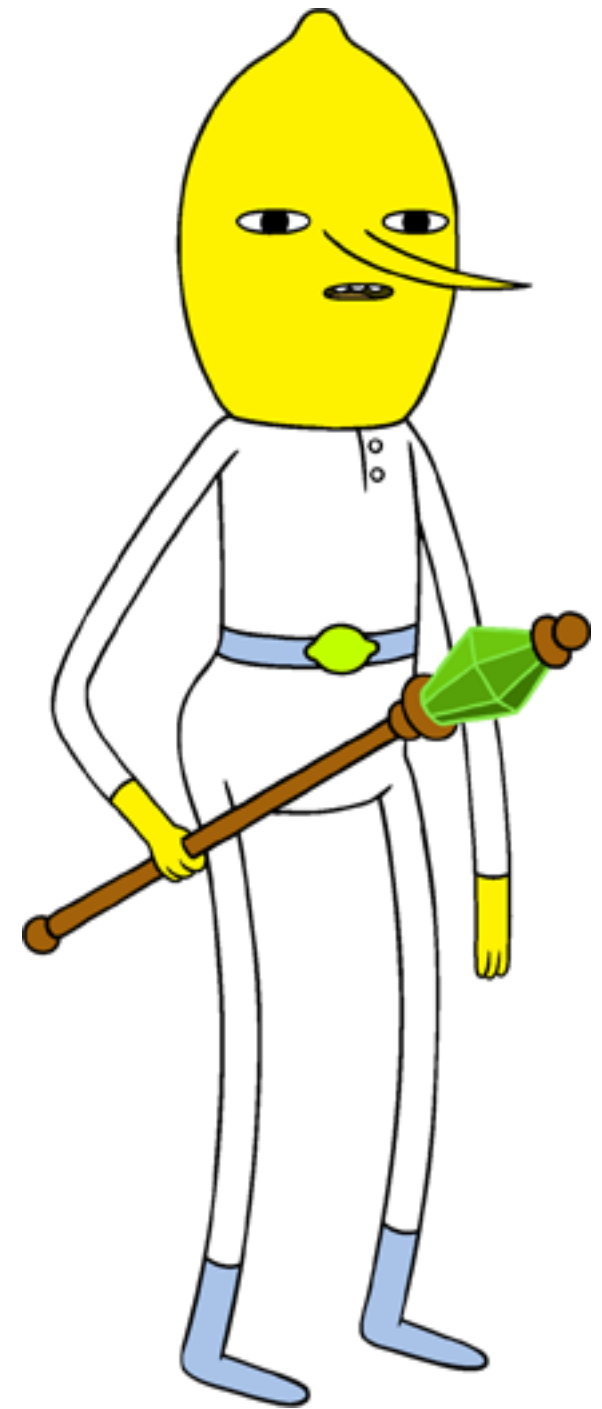
benefit of wrangling in R versus Excel

reproducibility



transparent

know thy data



values, variables, and observations

tidy your data with variables in columns



vectors of values are easier to work with in R

column headers should be variable names

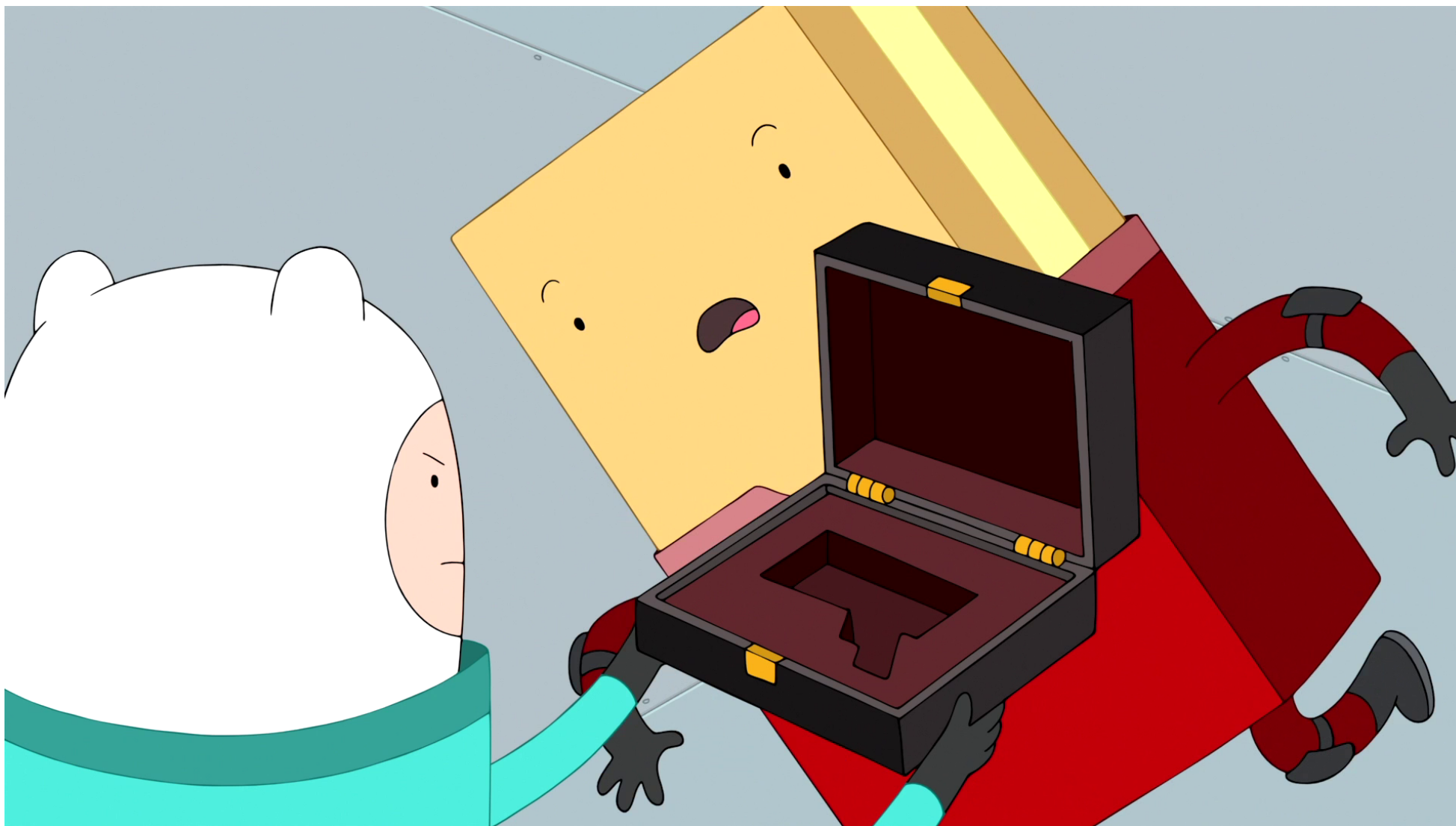
no special signs such as \$#*/ within dataframe

one variable per column and one class

one observational unit per table



missing values



happen



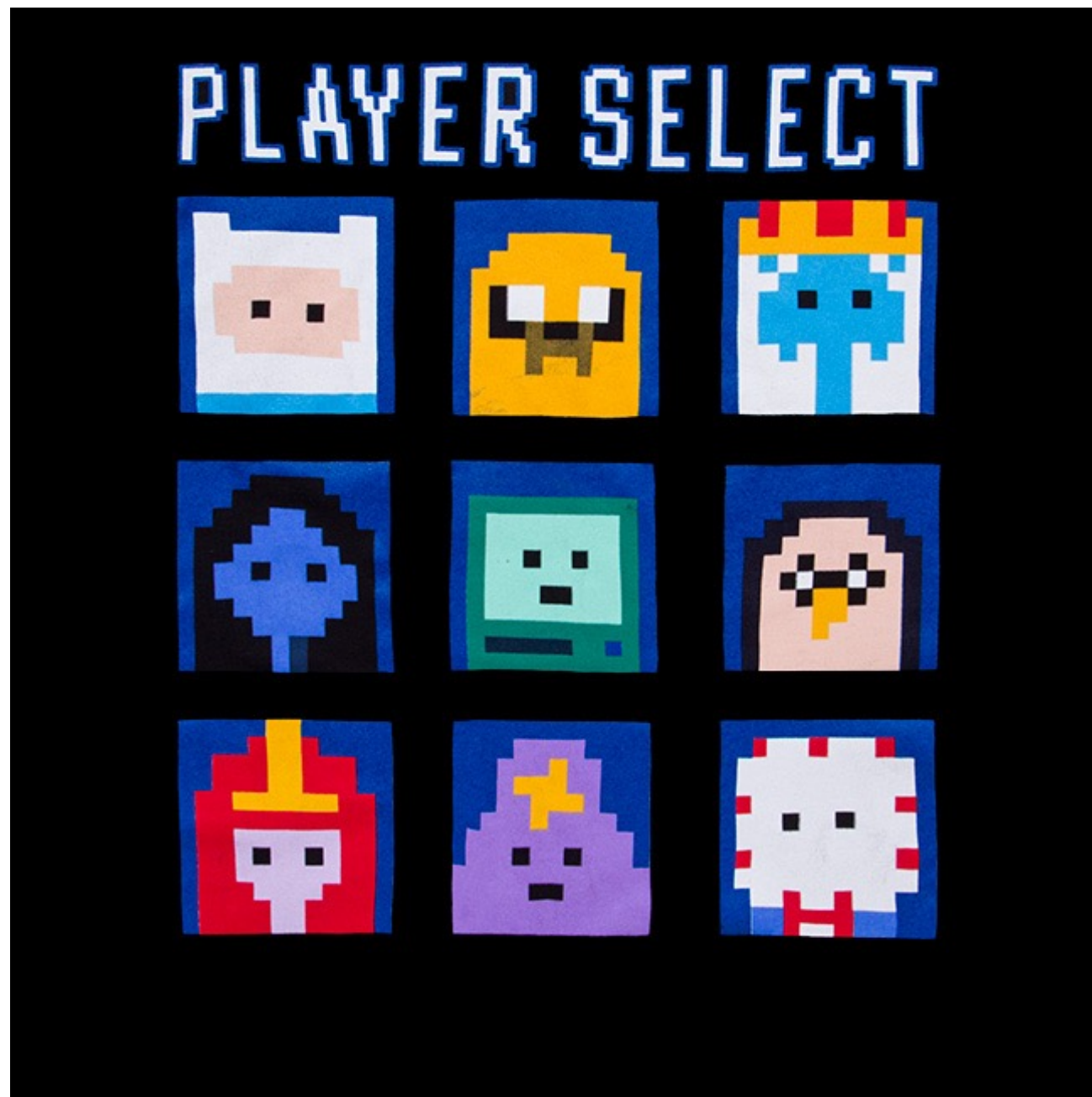
is.na()
na.omit
na.rm=TRUE

simplify & selections from dataframes

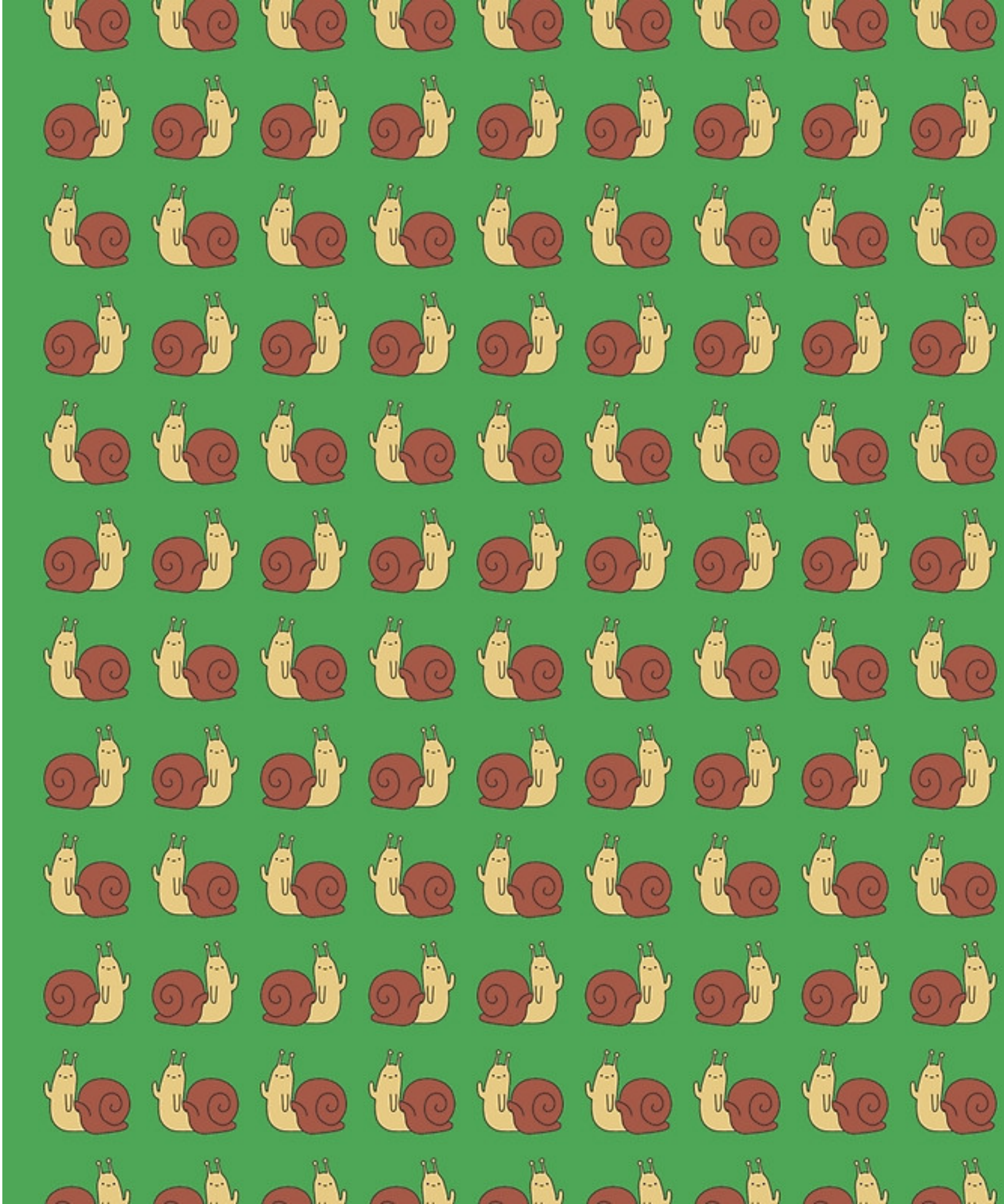
drplyr



select for columns



filter for rows



summarise & mutate

%>% to build logical wrangling pipelines



tidy data semantics from Journal of Statistical Software
reading such as 'tidy data' idea paper