**Quiz 8, Version 1**

**Figure 1. On the left, there are two dataframes called “songs” and “artists”. On the right is a dataframe called “combined”, which is the result of merging the two dataframes on the left. Use these dataframes to answer Questions 1-3.**



Q1. If we wanted to combine the two dataframes on the left in Figure 1 to create the dataframe on the right, which R code would we use?

1. left\_join(songs, artists, by=“name”)
2. arrange(songs, by=“name”)
3. bind\_rows(songs, artists)
4. rbind(songs, artists, by=“name”)

Q2. If we wanted to sort the whole songs dataframe by the name column so that the names were in alphabetical order, which R code would we use?

1. songs %>%

 arrange(name)

1. sort(songs$name)
2. arrange(songs)
3. group\_by(name) %>%

 songs

Q3. If you ran the code is.na(combined$plays), the result would be FALSE FALSE FALSE TRUE.

1. TRUE
2. FALSE

Q4. Which of these statements is **false**?

1. In R, data wrangling is often completed using the functions in the ggplot2 and readr packages.
2. Data wrangling generally takes longer than most statistical analyses; once datasets are cleaned and have the correct structure, most analyses are quick.
3. Reshaping dataframes and creating new columns are examples of data wrangling.
4. Dealing with missing values is part of data wrangling.

Q5. Which of these would **NOT** be included in the metadata for a dataset?

1. The list of references from the journal article that the dataset was first published in.
2. A description of each column or variable in the dataset that explains any abbreviations, the units of measurement, etc.
3. Information regarding how the data were collected.
4. Information regarding who collected the data, who owns the data, and what the data can be used for.

Q6. Which of these is a reason why you should **NOT** make a research dataset publicly available in an online repository?

1. Your data include the geographic coordinates where endangered species were found during wildlife surveys.
2. Publishing the data will maximize research transparency and accountability by allowing other researchers to check the statistical analyses for errors.
3. Publishing the data will reduce the costs of future research by helping researchers to avoid collecting the same data again.
4. Publishing the data will advance science by letting others use the data in new, innovative ways.

**Quiz 8, Version 2**

Q1. Which of these is a reason why you should **NOT** make a research dataset publicly available in an online repository?

1. Your data include the names and addresses of people who were surveyed in the research.
2. Publishing the data will allow other researchers to check and validate the data and analyses.
3. Publishing data will reduce the costs of research by helping researchers to avoid collecting data that someone else has already collected.
4. Tax dollars paid for the research, so the data should be available for anyone to use.

Q2. Which of these would **NOT** be included in the metadata for a dataset?

1. The statistical tests used to analyze the dataset
2. The research questions and hypotheses addressed by the project that the data were collected for
3. Known problems that might limit how the data can and should be used; for instance, if the study was based on a biased sample of a population
4. Explanations for why some values are missing in the dataset

Q3. Which of these statements is **false**?

1. Running regression analyses is an example of data wrangling.
2. Data wrangling must happen before most plotting and statistical analyses can be conducted.
3. Data wrangling generally takes longer than most statistical analyses; once datasets are cleaned and have the correct structure, most analyses are quick.
4. In R, data wrangling is often completed using the functions in the dplyr and tidyverse packages.

Q4. These two lines of code do the same thing; they both select a column called breed from a database called frisbeedogs.

select(data=frisbeedogs, breed)

frisbeedogs %>% select(breed)

1. TRUE
2. FALSE

**Figure 1. Below are two datasets called Cases\_Wide and Cases\_Long. Use them to answer Questions 5 and 6.**



Q5. Which R code would transform Cases\_Wide into Cases\_Long from Figure 1?

1. gather(Cases\_Wide, “Year”, “n”, 2:4)
2. gather(Cases\_Long, “Year”, “n”, 2:4)
3. spread(Cases\_Wide, Year, n)
4. spread(Cases\_Long, Year, n)

Q6. In Cases\_Long from Figure 1, which of these would sort the dataframe by Year and show only the Country and Year columns?

1. Cases\_Long %>%

 arrange(Year) %>%

 select(Year, Country)

1. Cases\_Long %>%

 sort(Year) %>%

 print(Year, Country)

1. Cases\_Long %>%

 group\_by(Year) %>%

 mutate(Year, Country)

1. Cases\_Long %>%

 group\_by(Year) %>%

summarise(Year, Country)