Tidy Data and Joins Activity

Learning Objectives

- Practice identifying tidy and untidy elements of data tables.
- Recognize primary and foreign keys
- Exercise thinking about how to join two data frames

About the data

These exercises will use bird survey data collected from the central Arizona-Phoenix metropolitan area by Arizona State University researchers.

The data tables in this exercise have been adapted from this dataset (https://portal.edirepository.org/nis/mapbrowse?packageid=knb-lter-cap.256.10) for teaching purposes. They were created for this specific exercise and should not be used for any official analysis or data visualization. Please refer to the raw data from the data portal to complete any data analysis or visuals.

Set Up

1. Get together in pairs or small groups
2. Obtain materials from instructors including:
    a. Activity handout
    b. Blank paper(s)

Question 1: Assess Tidyness

Assess the "tidyness" of each of the following data tables.

1. Does each table follow the three tidy data principles? If not, which ones aren't met?
2. How would you wrangle the data to make it tidy? Describe the steps you would take to tidy the data.

Survey dataframe

survey_id	site_id	survey_date	time_start	time_end	observer	wind_speed	air_temp
23	AT-N	2000-12-29	9:10:00	9:25:00	J. Lemmer	0	50
87	AT-S	2000-12-28	9:05:00	9:20:00	B. Rambo	0	45
370	AT-E	2001-03-09	8:45:00	9:00:00	J. Lemmer	0	54
760	AT-C	2001-06-27	8:50:00	9:05:00	D. Stuart	10	85
938	AT-W	2001-09-16	6:25:00	6:40:00	B. Rambo	0	75
total_parks	**total_sites**	**total_surveys**	**total_time**	**total_observers**	**avg_air_temp**		
1	5	5	15 minutes	3	61.8		

Taxonomy dataframe

species_id	common_name	asu_itis
EUST	European Starling	211002
MODO	Mourning Dove	162560

Bird dataframe

survey_id	site_id	bird_count_EUST	distance_EUST	direction_EUST	bird_count_MODO	distance_MODO	direction_MODO
760	AT-C	*NA*	*NA*	*NA*	1	0-5	E
936	AT-E	1	FT	N	*NA*	*NA*	*NA*
370	AT-E	4	20-40	SE	*NA*	*NA*	*NA*
23	AT-N	5	20-40	E	*NA*	*NA*	*NA*
87	AT-S	4	FT	W	2	FT	W
938	AT-W	1	>40	SW	*NA*	*NA*	*NA*

Site dataframe

site_id	park_code	park_district	park_name
AT-C	AT	NE	Altadena
AT-E	AT	NE	Altadena
AT-N	AT	NE	Altadena
AT-S	AT	NE	Altadena
AT-use	AT	NE	Altadena
AT-W	AT	NE	Altadena

Question 2: Identify keys

Identify the primary and foreign keys of each table. Use the tidy data frames below. For some tables, it might be necessary to create a compound key.

Survey table

survey_id	site_id	survey_date	observer	air_temp
87	AT-S	2000-12-28	B.R.	45
370	AT-E	2001-03-09	J.L.	54
936	AT-E	2001-09-16	B.R.	75

Taxonomy table

species_id	common_name	asu_itis
EUST	European Starling	211002
MODO	Mourning Dove	162560

Bird table

survey_id	site_id	species_id	bird_count	direction
936	AT-E	EUST	1	N
370	AT-E	EUST	4	SE
87	AT-S	EUST	4	W
936	AT-E	MODO	1	S
370	AT-E	MODO	3	NE
87	AT-S	MODO	2	W

Site table

site_id	park_code
AT-E	AT
AT-N	AT
AT-S	AT

Parks table

park_code	park_district	park_name
AT	NE	Altadena

Table	Primary Key	Foreign Keys
surveys		
taxonomy		
bird_records		
sites		
parks		

Question 3: Join the data tables

Describe the steps on how you would join these data frames to have one data frame containing all the information. We want the final data frame to look like this:

survey_id	site_id	species_id	bird_count	direction	common_name	asu_itis	survey_date	observer	air_temp	park_code	park_district	park_name
936	AT-E	EUST	1	N	European Starling	211002	2001-09-16	B.R.	75	AT	NE	Altadena
370	AT-E	EUST	4	SE	European Starling	211002	2001-03-09	J.L.	54	AT	NE	Altadena
87	AT-S	EUST	4	W	European Starling	211002	2000-12-28	B.R.	45	AT	NE	Altadena
936	AT-E	MODO	1	S	Mourning Dove	162560	2001-09-16	B.R.	75	AT	NE	Altadena
370	AT-E	MODO	3	NE	Mourning Dove	162560	2001-03-09	J.L.	54	AT	NE	Altadena
87	AT-S	MODO	2	W	Mourning Dove	162560	2000-12-28	B.R.	45	AT	NE	Altadena

- Which tables would you join first and why?

- What type of join would you use and why?

- By what variable would you join each table and why?