COMM 205

Introduction to Management Information Systems

Instructor: Adam Saunders

Lecture 23: Merging Datasets, Part II

March 29th, 2023

UBC SAUDER SCHOOL OF BUSINESS

MERGING DATASETS

- In this course, we will cover two kinds of methods to merge datasets:

```
1. inner_join()
2. left_join()
```

- Joining allows us to **match observations** from two datasets based on matching values in a particular attribute(s).
- When merging two datasets, the **variable names** of the attribute(s) in **both** datasets should be the same to use the default syntax.
- It is also recommended that the **variable type** of the key variable(s) in **both** datasets also be the same.

`left_join()` FUNCTION

- `left_join(x, y)` returns all rows from `x`, and all columns from `x` and `y`.
- Let's use the same **example1** and **example2** datasets from before. Suppose this time, you want to improve the **example1** dataset by bringing **as much information** from the **example2** as you can.
- Thus, you want to keep all the observations in the **example1** dataset **and** bring new information for the matching observations in the **example2** dataset.
- `left_join()` is suitable for this task.

`left_join()` FUNCTION

- To do this, enter into the Console:

```
merged2 <- left_join(example1,example2)
```

- R will return `Joining, by = c("gvkey", "fyear")`

Filter

	gvkey	fyear	conm	ni	at	sale
1	001004	2014	AAR CORP	10.200	*NA*	*NA*
2	001004	2015	AAR CORP	47.700	*NA*	*NA*
3	001004	2016	AAR CORP	56.500	1504.100	1767.600
4	001045	2014	AMERICAN AIRLINES GROUP INC	2882.000	*NA*	*NA*
5	001045	2015	AMERICAN AIRLINES GROUP INC	7610.000	*NA*	*NA*
6	001045	2016	AMERICAN AIRLINES GROUP INC	2676.000	51274.000	40180.000

`left_join()` FUNCTION

- For matching values of the joining attribute – in this case, the combination of `gvkey` and `fyear`, `left_join()` links the information appearing in both datasets.
- Those observations in the ***"left"*** dataset which do not have matching observations in the other dataset will be retained in the resultant merged dataset, while those observations in the ***"right"*** dataset which do not have matching observations in the other dataset will not be retained in the resultant dataset.
- This explains why assets and sales are missing for AAR corporation in 2014 and 2015. The **example2 dataset does not cover those years (it runs from fiscal year 2016 through 2018).**

NON-ONE-TO-ONE MATCHING

- When you merge two datasets whose key variable(s) uniquely identify each observation within the dataset, this kind of merge is called a **one-to-one merge**.
- However, the attribute(s) (i.e., column(s)) used as a basis of matching the observations in two data sets **do not have to be key variables for both datasets**.
- In such situations, an observation in one dataset can match with multiple observations in the other datasets through **non-one-to-one matching**.
- However, the variable or variable(s) to perform the merge should uniquely identify **at least one of the datasets.**

NON-ONE-TO-ONE MATCHING

- Suppose you have two datasets, the first is a subset of the full **North American Stock Market 1994-2018.rds** dataset.
- **example3.rds** contains `gvkey`, `fyear`, and industry code (`naicsh`) of all observations in `fyear==2016`.
- We can also create **example3** quite easily:

```
example3 <- companies %>%
   filter(fyear==2016, !is.na(naicsh)) %>%
   select(gvkey, fyear, conm, naicsh)
```

NON-ONE-TO-ONE MATCHING

- This is a picture of **example3**

	gvkey	fyear	conm	naicsh
1	001004	2016	AAR CORP	423860
2	001045	2016	AMERICAN AIRLINES GROUP INC	481111
3	001050	2016	CECO ENVIRONMENTAL CORP	333413
4	001062	2016	ASA GOLD AND PRECIOUS METALS	523999
5	001072	2016	AVX CORP	334416
6	001075	2016	PINNACLE WEST CAPITAL CORP	2211
7	001076	2016	AARON'S INC	532299
8	001078	2016	ABBOTT LABORATORIES	325412
9	001084	2016	WORLDS INC	519130
10	001094	2016	ACETO CORP	424690
11	001097	2016	ACMAT CORP -CL A	524126
12	001104	2016	ACME UNITED CORP	332215
13	001117	2016	BK TECHNOLOGIES CORP	334220
14	001119	2016	ADAMS DIVERSIFIED EQUITY FD	525990

Showing 1 to 15 of 7,159 entries, 4 total columns

NON-ONE-TO-ONE MATCHING

- **NAICS_2_6_digit_codes.rds** contains the NAICS code and industry description for more than 2,200 industries (from the 2 to the 6 digit level). This is provided for you on Canvas.

	NAICS	NAICS Description
1	11	Agriculture, Forestry, Fishing and Hunting
2	111	Crop Production
3	1111	Oilseed and Grain Farming
4	11111	Soybean Farming
5	111110	Soybean Farming
6	11112	Oilseed (except Soybean) Farming
7	111120	Oilseed (except Soybean) Farming
8	11113	Dry Pea and Bean Farming
9	111130	Dry Pea and Bean Farming
10	11114	Wheat Farming
11	111140	Wheat Farming

NON-ONE-TO-ONE MATCHING

- You want to merge the two datasets, so that each observation will have the following variables: `gvkey, fyear, naicsh`, and `NAICS_description`. This will attempt to add the industry description to every observation in the Example3 dataset.
- To do this, enter into the Console:

```
merged3 <- left_join(example3, NAICS_2_6_digit_codes, by = c("naicsh" = "NAICS"))
```

NON-ONE-TO-ONE MATCHING

- Here's what you should see as a result:

	gvkey	fyear	conm	naicsh	NAICS_Description
1	001004	2016	AAR CORP	423860	Transportation Equipment and Supplies (except Moto…
2	001045	2016	AMERICAN AIRLINES GROUP INC	481111	Scheduled Passenger Air Transportation
3	001050	2016	CECO ENVIRONMENTAL CORP	333413	Industrial and Commercial Fan and Blower and Air Pur…
4	001062	2016	ASA GOLD AND PRECIOUS METALS	523999	Miscellaneous Financial Investment Activities
5	001072	2016	AVX CORP	334416	Capacitor, Resistor, Coil, Transformer, and Other Ind…
6	001075	2016	PINNACLE WEST CAPITAL CORP	2211	Electric Power Generation, Transmission and Distribut…
7	001076	2016	AARON'S INC	532299	All Other Consumer Goods Rental
8	001078	2016	ABBOTT LABORATORIES	325412	Pharmaceutical Preparation Manufacturing
9	001084	2016	WORLDS INC	519130	Internet Publishing and Broadcasting and Web Search …
10	001094	2016	ACETO CORP	424690	Other Chemical and Allied Products Merchant Wholes…
11	001097	2016	ACMAT CORP -CL A	524126	Direct Property and Casualty Insurance Carriers
12	001104	2016	ACME UNITED CORP	332215	Metal Kitchen Cookware, Utensil, Cutlery, and Flatwar…
13	001117	2016	BK TECHNOLOGIES CORP	334220	Radio and Television Broadcasting and Wireless Com…
14	001119	2016	ADAMS DIVERSIFIED EQUITY FD	525990	Other Financial Vehicles

Showing 1 to 15 of 7,159 entries, 5 total columns