In-class Exercise 14

Started: Nov 23 at 10:32am

Quiz Instructions

Question 1 — 1.66 pts

Please download **NASM_ICE14.rds** file. You will use this dataset in Questions 1 and 2. **This is different from the previously North American Stock Market dataset. Therefore, you should use this one.**

Load the dataset with readRDS and called the data frame as `ice14`.

Similar to our usual North American Stock Market dataset, gvkey and fyear combination is used to uniquely identify each firm's observation in a given financial year in `ice14`. However, you are told that there are exactly **two** observations in `ice14` that share the same `gvkey` and `fyear`.

Find those two observations and report which `gvkey` and `fyear` are duplicated in the dataset (Hint: you can use the same approach used in the lecture).

`gvkey:` 112982

`fyear:` 2013

Note: Enter numbers only; do NOT include any quotation marks, commas, or any other special characters.

Question 2 — 1.66 pts

You are asked to identify the duplicates in `ice14` by not using the `duplicated()` function. Complete the formula to output the gvkey and fyear combination that is duplicated.

```
identifiers <- ice14 %>%
  group_by (gvkey, fyear) %>%
  summarise (marker = n ())
```

```
identifiers %>% filter (marker == 2)
```

Note: You can use function names, arithmetic or logical operators to fill the blanks.

Question 3

1.66 pts

Suppose you have conducted a small survey on college students. The survey is about schooling and employment and each participant is assigned to a unique id. You created `base_survey` data frame as follows:

```
base_survey <- data_frame(id = c(1,2,3,4,5,6,7,8,9,10),
                          school = c("UBC", "UBC", "SFU", "UBC", "BCIT", NA,
"Langara", "Douglas", NA, "Emily Carr"),
                          part_time = c(TRUE, FALSE, FALSE, TRUE, TRUE, FALSE,
TRUE, TRUE, FALSE, FALSE))
```

You then conducted a follow-up survey 2 months later because you realized that you had forgotten to ask students about their hourly wages (for those who did work part time). Many of the participants of the base survey participated in the follow-up survey. Some of those who participated in the follow-up survey chose not to disclose their wages.

Run the following code to create the `follow_up_survey` data frame with the responses.

```
follow_up_survey <- data_frame(id = c(1,2,3,4,5,7,8,10),
                               hourly_pay = c(21, NA, NA, 17, NA, 19, 14, NA))
```

By using the unique id of a participant, you combine responses from both survey. You now want to combine the data you collected in these two surveys. Note that you want to retain all rows from base_survey.

Then to merge the two datasets you would run:

```
merge1 <- left_join (base_survey, follow_up_survey)
```

You now want to only keep observations where the respondent have participated in both surveys.

Then ,to merge the two datasets you would run:

```
merge2 <- inner_join (base_survey, follow_up_survey)
```

Question 4

1.68 pts

Assume that you are using `data1` and `data2` which have a common identifier called `ID`. Please select the correct statement(s) regarding `inner_join()` and `left_join()`. There is at least one correct statement.

- [x] the number of columns in the merged data frame produce by `left_join(data1, data2)` is always equal to the number of columns in the merged data frame produce by `inner_join(data1, data2)`.
- [ ] the number of columns in the merged data frame produce by `left_join(data1, data2)` is smaller than or equal to the number of columns in the merged data frame produce by `inner_join(data1, data2)`.
- [ ] the number of columns in the merged data frame produce by `left_join(data1, data2)` is greater than or equal to the number of columns in the merged data frame produce by `inner_join(data1, data2)`.
- [ ] the number of rows in the merged data frame produce by `left_join(data1, data2)` is always equal to the number of rows in the merged data frame produce by inner `inner_join(data1, data2)`.
- [ ] the number of rows in the merged data frame produce by `left_join(data1, data2)` is smaller than or equal to the number of rows in the merged data frame produce by `inner_join(data1, data2)`.
- [x] the number of rows in the merged data frame produce by `left_join(data1, data2)` is greater than or equal to the number of rows in the merged data frame produce by `inner_join(data1, data2)`.

Question 5

1.68 pts

You can change the order of the data frames that you pass onto `inner_join()` and `left_join()`.

For example, if you have two data frames; `data1` and `data2`, it is possible to do `inner_join(data1, data2)` or `inner_join(data2,data1)`.

While swapping the order of the data frames is possible, the merged datasets may or may not have the same number of rows.

Assume you are merging `data1` and `data2` by a common variable called `ID`. Please select the correct statement(s) below. There is at least one correct option. Note that you are only asked to comment on **only** the *number of rows* produced.

- [x] the number of observations in the merged dataset when `inner_join(data1, data2)` is used is always the same as that when `inner_join(data2, data1)` is used
- [ ] `left_join()` always produces the same number of rows in the merged data frame regardless of the order of the two data frames.
- [ ] `inner_join()` can produce the same number of rows in the merged data frame when you change the order of the two data frames, but not always.
- [x] `left_join()` can produce the same number of rows in the merged data frame when you change the order of the two data frames, but not always.
- [ ] `inner_join()` can never produce the same number of rows in the merged data frame when you change the order of the two data frames.
- [ ] `left_join()` can never produce the same number of rows in the merged data frame when you change the order of the two data frames.

Question 6

1.66 pts

Suppose you have two data frames called `data3` and `data4`. You would like to merge them by inner joining them with two matching variables. In `data3`, they are named `id` and `time`, while in `data4`, they are named `ID` and `TIME`, respectively. You would then use the following code to merge the two datasets:

```
data5 <- inner_join(data3, data4, [by] = c
( ["id"] = "ID" , ["time"] = "TIME" ))
```

No new data to save. Last checked at 11:10am

Submit Quiz