In-Class Exercise 5

ⓘ This is a preview of the published version of the quiz

Started: Jul 21 at 3:50p.m.

Quiz Instructions

As long as you are online, your answers will be automatically saved while you complete the exercise. While you have unlimited attempts to complete this in-class-exercise before the due date, only the grade of the *last* submitted attempt will be recorded.

Question 1 — 1.11 pts

For the next few questions in this exercise, please use the North American Stock Market 1994-2018 dataset. Also, load the tidyverse library.

Start with the full North American Stock market dataset. Keep observations with fiscal year (fyear) equal to 2016 or 2017. Also, keep observations as long as they have at least $10 million dollars in assets and sales. Note that assets and sales are in millions of dollars, and are denoted by at and sale respectively.

Finally, keep the variables gvkey, fyear, conm, at, and sale.

Save this dataset as q1.

What is the unique identifier(s) in this dataset? Remember the unique identifier is the variable (or **minimum** combination of variables) that allows you to uniquely identify each observation. No two observations share the same value(s) of the unique identifier (otherwise, it is not unique =)

- ◯ gvkey
- ◯ gvkey and fyear
- ◯ gvkey, fyear, and conm
- ◯ gvkey, fyear, conm, at, and sale
- ◯ None of the above.

Question 2

1.11 pts

Now, start with the full North American Stock market dataset again. Keep observations in fiscal years 2017 or 2018, as well as observations with $100 million or more in assets and sales.

Keep the variables gvkey, fyear, tic, and ni.

Save this dataset as q2.

Merge q1 and q2 with an inner_join() by the unique identifier(s) of both datasets.

How many observations are in the merged dataset?

Question 3

1.11 pts

Start with the full North American Stock market dataset again. Keep observations with $50 million or more in assets and sales, as well as observations with nonmissing values of employment and net income (emp and ni respectively in the dataset).

We are going to calculate ROA for each observation, which is return on assets (ROA). This is net income (ni) divided by assets.

Finally, condense this dataset (*note: use a summarise() here*) so that you have only two variables (columns). The first is gvkey, and the second is ROA_avg, which is average ROA for that gvkey across all of its observations in the remaining dataset. Save this dataset as q3. What is the unique identifier(s) of q3?

- ◯ gvkey
- ◯ gvkey and fyear
- ◯ gvkey and ROA

- ◯ gvkey, fyear, and ROA
- ◯ None of the above.

Question 4

1.11 pts

Continue the previous question. How many observations are there in q3?

Question 5

1.11 pts

Now, start with the full North American stock market dataset again. Keep observations with nonmissing values of employment (emp). Calcalulate a per-firm average of employment, over the 1994-2018 time period, called emp_avg. (*Note: like q3, use the summarise() function to achieve this*).

Now, condense the dataset so there are only two variables: gvkey and emp_avg, and save this as q5. Merge q3 and q5 together using an inner_join().

How many observations are there in the merged dataset?

Question 6

1.11 pts

Note: The next four questions do not use the companies dataset anymore.

Suppose you have a data frame called students1 that has 7,000 observations and three variables stored as characters: UBC_ID (a unique eight-digit identifier), last_name, and first_name. Note that no two observations in students1 have the same UBC_ID.

Now, you also have a data frame called students2 that has 10,000 observations and two variables. The first variable is a character variable called UBC_ID, which has the same meaning as UBC_ID in students1. There is also the variable called gpa, a numerical variable representing a student's grade point average. Note that no two observations in students2 have the same UBC_ID.

You also learn there are exactly 3,000 observations with the same UBC_ID that appear in both students1 and students2. Assume that there is NO missing data anywhere in either dataset.

Suppose you run the following command in RStudio:

merged1 <- left_join(students1, students2)

Which one of the following options below is true regarding the number of observations in merged1?

- ◯ There are less than 7,000 observations in merged1.
- ◯ There are more than 10,000 observations in merged1.
- ◯ There are exactly 7,000 observations in merged1.
- ◯ There are exactly 10,000 observations in merged1.
- ◯ Without more information, we can't know how many observations are in merged1.

Question 7

1.11 pts

Now, suppose you run the following:

merged2 <- left_join(students2, students1)

Which one of the following options below is true regarding the number of observations in merged2?

- ◯ There are less than 7,000 observations in merged2.

- ◯ There are more than 10,000 observations in merged2.
- ◯ There are exactly 7,000 observations in merged2.
- ◯ There are exactly 10,000 observations in merged2.
- ◯ Without more information, we can't know how many observations are in merged2.

Question 8

1.11 pts

Now, suppose you run the following:

```
merged3 <- inner_join(students1, students2)
```

Which one of the following options below is true regarding the number of observations in merged3?

- ◯ There are less than 7,000 observations in merged3.
- ◯ There are more than 10,000 observations in merged3.
- ◯ There are exactly 7,000 observations in merged3.
- ◯ There are exactly 10,000 observations in merged3.
- ◯ Without more information, we can't know how many observations are in merged3.

Question 9

1.12 pts

True or False: All merges using an inner_join() can use at most one key variable (that is common to both datasets) to successfully merge the datasets. If there is more than one key variable to merge the datasets, then you cannot use an inner_join().

- ◯ True
- ◯ False

Not saved Submit Quiz