In-Class Exercise 4

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 1.25 pts

For ALL Questions in this exercise, please use the North American Stock Market 1994-2018 dataset.

Please load the data into a data frame called *companies* by reading the appropriate .rds file. See lecture slides to see how to do this.

Make sure to run the code library(tidyverse) so you load tidyverse.

For each fiscal year (fyear), you want to calculate the *median* value of assets (at) for all companies recorded in that fyear. Make the assumption that each company is listed only once per year.

Which fyear had the *highest* median at?

○ 2001

○ 2009

○ 2011

○ 2018

○ None of the above.

Question 2 1.25 pts

Suppose you want to create a new dataset. To create this new dataset, start with the usual **companies** dataset. Then, for each firm, drop *all* observations (of that firm) if the firm **has never reached $100 million in total assets (i.e., at) at any time in the dataset.** Note that assets is listed in millions, so $100 million would be 100.

To elaborate, let's consider the following hypothetical case:

Suppose there are 6 observations for company ABC. If *at least one* of these 6 observations has at >= 100, you must NOT drop ANY of the observations for firm ABC.

How many observations (i.e., rows) are in the new dataset after dropping observations indicated above?

Hint: Each company is uniquely identified by its *gvkey*, and each row is uniquely identified by the combination of *gvkey* and *fyear*.

○ 9,003

○ 144,655

○ 163,777

○ 177,391

○ None of the above.

Question 3

1.25 pts

Remember that *gvkey* is the unique identifier given to each firm in the dataset, *fyear* is the fiscal year, *loc* is a 3-letter abbreviation for the country of headquarters of that firm, and *sale* is total sales for that firm in that fiscal year.

Assume that *gvkey* and *fyear* together uniquely identify all observations in *companies*.

Now, you run the following code in RStudio:

```
new_dataset <-companies %>%
  filter(!is.na(fyear), !is.na(loc), !is.na(sale)) %>%
  group_by(fyear, loc) %>%
  summarise(max_sale = max(sale))
```

Which of the options below ***best*** describes the contents of *new_dataset*?

○ The maximum sales of each firm in the companies dataset.

○ For every country in the companies dataset, a listing of the largest firms headquartered in those countries, and the corresponding sales of those firms.

○ For every fiscal year in the companies dataset, a listing of the largest firms in each of those years, and the corresponding sales of those firms.

○ For every country-year combination in companies, the maximum value of sales for firms headquartered in each country in each year.

○ The maximum of sales for each fiscal year in the dataset.

Question 4

1.25 pts

Suppose you are asked to calculate the **total** sales (sale) of **all** firms in each country in each fiscal year (fyear). Call this new variable total_country_sales. Remember we use *loc* to denote the country of headquarters for each firm.

We want a dataset, called *q4*, that is based on the dataset *companies* and we want all of the original 41 variables (i.e., columns) in *companies* to be retained in *q4.* (Recall we also want *q4* to have the new variable total_country_sales, so that *q4* has 42 variables in the end.)

We also want to drop any observations (i.e., rows) that have missing values for **any** of the following variables: sales, headquarters, or fiscal year.

What would be the correct code to do this?

○
```
q4 <- companies %>%
  filter(!is.na(fyear), !is.na(loc), !is.na(sale)) %>%
  group_by(fyear,loc) %>%
  mutate(total_country_sales = sum(sale, na.rm = TRUE))
```

○
```
q4 <- companies %>%
  filter(is.na(fyear), is.na(loc), is.na(sale)) %>%
  group_by(fyear,loc) %>%
  mutate(total_country_sales = sum(sale, na.rm = TRUE))
```

○
```
q4 <- companies %>%
  select(sale, fyear, gvkey) %>%
  filter(!is.na(fyear), !is.na(loc), !is.na(sale)) %>%
  group_by(fyear,loc) %>%
  mutate(total_country_sales = sum(sale, na.rm = TRUE))
```

○ q4 <- companies %>%

```
filter(!is.na(fyear), !is.na(loc), !is.na(sale)) %>%
group_by(fyear,loc) %>%
summarise(total_country_sales = sum(sale, na.rm = TRUE))
```

○ None of the above.

Question 5

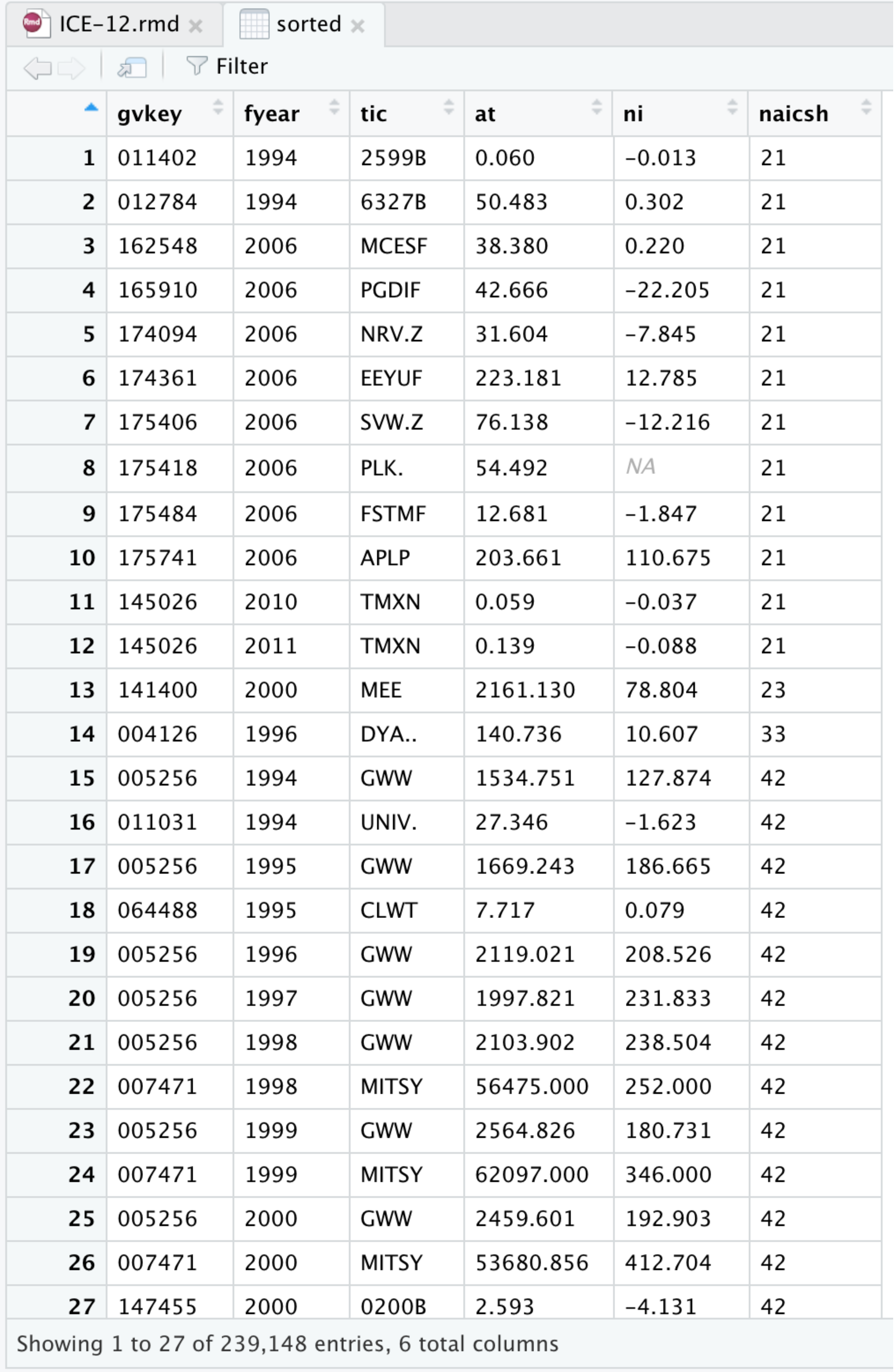
1.25 pts

Assume you have opened the usual dataset, the North American Stock Market from 1994-2018, and successfully loaded it as companies.

You then run the following:

```
sorted <- companies %>%
  filter(!is.na(fyear), !is.na(naicsh)) %>%
  #Command(s) you need to fill in here
  select(gvkey, fyear, tic, at, ni, naicsh)
View(sorted)
```

Now refer to the figure below to answer this question:



	gvkey	fyear	tic	at	ni	naicsh
1	011402	1994	2599B	0.060	-0.013	21
2	012784	1994	6327B	50.483	0.302	21
3	162548	2006	MCESF	38.380	0.220	21
4	165910	2006	PGDIF	42.666	-22.205	21
5	174094	2006	NRV.Z	31.604	-7.845	21
6	174361	2006	EEYUF	223.181	12.785	21
7	175406	2006	SVW.Z	76.138	-12.216	21
8	175418	2006	PLK.	54.492	NA	21
9	175484	2006	FSTMF	12.681	-1.847	21
10	175741	2006	APLP	203.661	110.675	21
11	145026	2010	TMXN	0.059	-0.037	21
12	145026	2011	TMXN	0.139	-0.088	21
13	141400	2000	MEE	2161.130	78.804	23
14	004126	1996	DYA..	140.736	10.607	33
15	005256	1994	GWW	1534.751	127.874	42
16	011031	1994	UNIV.	27.346	-1.623	42
17	005256	1995	GWW	1669.243	186.665	42
18	064488	1995	CLWT	7.717	0.079	42
19	005256	1996	GWW	2119.021	208.526	42
20	005256	1997	GWW	1997.821	231.833	42
21	005256	1998	GWW	2103.902	238.504	42
22	007471	1998	MITSY	56475.000	252.000	42
23	005256	1999	GWW	2564.826	180.731	42
24	007471	1999	MITSY	62097.000	346.000	42
25	005256	2000	GWW	2459.601	192.903	42
26	007471	2000	MITSY	53680.856	412.704	42
27	147455	2000	0200B	2.593	-4.131	42

Showing 1 to 27 of 239,148 entries, 6 total columns

The command(s) that could be used instead of the commented part of the code above to sort observations in the dataset, so that the above result is produced is/are:

○ arrange(gvkey, fyear, naicsh) %>%

○ arrange(naicsh, fyear, gvkey) %>%

○ arrange(gvkey) %>%
arrange(desc(fyear), naicsh) %>%

○ arrange(gvkey) %>%
arrange(desc(fyear)) %>%
arrange(naicsh) %>%

○ arrange(naicsh) %>%
arrange(desc(fyear), gvkey) %>%

○ arrange(naicsh) %>%
arrange(fyear) %>%
arrange(gvkey) %>%

○ None of the above.

Question 6

1.25 pts

Suppose you have a dataset, called mystery, with only two variables: a and b. Both variables are numerical variables (stored as doubles in R). Other than the fact you know that no values are missing, suppose you don't know any of the values for either variable.

Now, someone tells you that if you run the following:

mystery %>% arrange(a, b)

mystery %>% arrange(a, desc(b))

the mystery dataset would end up in the identical order either way.

Which of the following statements, if TRUE, would be SUFFICIENT ENOUGH BY ITSELF to produce this interesting result? In other words, as long as the statement is true, you would always get the above result as long as the rest of the question is also satisfied. Evaluate each of the statements below independently--as you consider each answer, do not depend on whether the other statements are true or false.

There is AT LEAST ONE correct option, but you MUST SELECT ALL correct options.

□ There are no duplicates for a in the entire dataset

□ There are no duplicates for b in the entire dataset

□ All values of a are identical to each other

□ All values of b are identical to each other

□ None of the above (if you select this option, do NOT select any other options)

Question 7

1.25 pts

Now, start with the full **companies** dataset again.

Suppose you want to drop any *observation* (i.e., any row) that has less than $100 million in total assets (at < 100) **or,** if it has less than sales of $100 million (sale <100). You also want to drop any observations with a missing value for any of the following: employment (emp), sales (sale), or assets (at).

After performing this screening procedure, you want to know what is the average employment *per firm* that is headquartered in the United States (loc == "USA") that was listed at any time over the years (fyear) 2016 to 2018 inclusive.

Note that in the 2016-2018 period, some American firms will be listed in just one year only. Some of them will be in there in for all three years. And some will be in there in between. This is fine. **As long as a firm is listed in that time period at all,** we want to include it in the calculation.

Remember that the unique firm identifier is *gvkey.*

First, before we get to the average employment, how many eligible firms would be included in this calculation at all?

Question 8

1.25 pts

Continue the previous question with all of the relevant screening procedures on the companies dataset.

Now that you have figured out how many firms that would be included in this average, what is the average employment *per firm* that was headquartered in the United States (loc == "USA") and listed anytime during the years *(fyear)* 2016 to 2018 inclusive? Remember that the unique firm identifier is *gvkey.*

Employment is in thousands, so list your answer in thousands, just as it would appear in RStudio. You may round your answer to one decimal place.

Not saved Submit Quiz