In-class Exercise 12 Results for Simran Mander

Score for this attempt: **6.4** out of 10
Submitted Nov 20 at 9:19pm
This attempt took 1,798 minutes.

Question 1

1.8 / 1.8 pts

For Questions 1-5, please use the North American Stock Market 1994-2013 dataset.

Please load the data by reading the RDS file: First, I would also suggest you to set the working directory to the folder where R related files reside. To do so, Session > Setting Working Directory > Choose Directory .... Then, make sure that you reference the RDS file with respect to your working directory. That is, if RDS file is in your working directory, use:

```
companies <- readRDS("North_American_Stock_Market_1994-2013.rds")
```

If the RDS file is saved in a folder, let's say a folder called data, which is under your working directory, use:

```
companies
<- readRDS("data/North_American_Stock_Market_1994-2013.rds")
```

You would like to create a new data frame with only the following columns: company name (ie., conm), employment (i.e., emp) and fiscal year (i.e., fyear). You would like to remove observations for which employment (i.e., emp) is `NA`. You would like your new data frame to contain observations in the 2012 fiscal year (i.e., `fyear==2012`). Hence, you run:

```
df1 <- companies %>% select (conm, emp, fyear)
%>% filter ( !is.na (emp), fyear
== 2012)
```

Now you wanted to find the difference between the maximum and the minimum values of the number of employees (i.e., emp) recorded in the

dataset in 2012. You found that the difference is 2200 .

Answer 1:

Correct! select

Answer 2:

Correct! filter

Answer 3:

Correct! !is.na

Answer 4:

Correct! ==

Answer 5:

Correct! 2200

```
df1 <- companies %>% select(conm, emp, fyear) %>%
filter(!is.na(emp), fyear == 2012)

df1 %>% summarise(difference = max(emp) - min(emp))
```

Question 2

1.8 / 1.8 pts

Use companies data frame. For each financial year (`fyear`), you wanted to calculate the median `at` value for all companies recorded in that fyear in the dataset.

Which fyear had the highest median `at`?

2011

The median `at` for that `fyear` is 332.2395 . (provide 4 decimal places)

Answer 1:

Correct!

2011

Answer 2:

Correct!

332.2395

```
df2 <- companies  %>%
select(fyear,at) %>%
group_by(fyear) %>%
summarise(med=median(at, na.rm=TRUE))

df2 %>% filter(med==max(med, na.rm=TRUE))
```

Question 3

0 / 1.8 pts

Use companies data frame. Remember each observation records information about a particular company in a given financial year.

We wanted to create a new dataset. To create this new dataset, for each company, drop all observations (of that firm) if the firm has never reached $50 million in total assets (i.e., `at`) at any time in the dataset. That is, for a given firm, if `at < 50` for all of its observations, you should remove those observations.

Let's consider these two hypothetical cases:

- Suppose there are 6 observations for company ABC. If at least one of these 6 observations has `at >= 50`, you must NOT drop ANY of the observations for firm ABC.
- However, suppose there are 10 observations for company XYZ. If all of these 10 observations has `at < 50`, you MUST drop ALL observations for company XYZ.

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

137438

Note: Each company has a unique `gvkey`. You can use it as your group_by variable if you need to group your dataset by company.

Answer 1:

You Answered: 137438

Correct Answer: 163777

```
df3<- companies %>%
  select(gvkey,at) %>%
  group_by(gvkey) %>%
  mutate(max_at=max(at, na.rm= TRUE)) %>%
  filter(max_at >= 50)
View(df3) # to confirm the # of rows 163777
```

Question 4

0 / 1.8 pts

Use companies data frame. Drop every observation which has less than $100 million in total assets (i.e. `at<100`) **or** less than $100 million in sales (i.e, `sale<100`). Hint: The sentence indicates which observations to eliminate. But, you need to think about which observations that you will retain because you will use `filter()`.

After performing the above screening process, we are interested in knowing how many unique firms appear **only once** in the screened dataset. To do so, we want to create a variable called `numberoftimes` to report how many times each gvkey appears in the screened data. Remember that firms are uniquely identified by `gvkey`.

- For example, imagine there is an observation with `gvkey` equal to `"0000001"` and `fyear` equal to `2014`. If `gvkey` `0000001` appears exactly only once in the screened dataset, then the value of `1` for the variable `numberoftimes` should be generated for this `gvkey`.
- Another example: imagine there is an observation with the `gvkey` equal to `"0000002"` and fyear equal to `2013`. Apparently, there is also another observation with the `gvkey` equal to `"0000002"` and fyear equal to `2014`. Because `gvkey` `00000002` appears twice in the screened dataset, a value of `2` for the variable `numberoftimes` should be generated for this `gvkey`.

How many firms appear only once in the screened dataset?

1803

Answer 1:

You Answered: 1803

Correct Answer: 1062

```
df4<-companies %>%
  select(gvkey,conm,sale,at) %>%
  filter(sale >= 100, at >= 100) %>%
  group_by(gvkey) %>%
  summarise(numberoftimes=n()) %>%
  group_by(numberoftimes) %>%
  summarise(numberoffirms= n()) %>%
  filter(numberoftimes==1)
```

Question 5

1.8 / 1.8 pts

Use companies data frame. You are interested in finding out how profitable these companies are. You know from another Sauder course that one metric you can look at is **EBITDA margin** which measures a company's operating profitability, and it is calculated as the ratio of EBITDA to Revenue. In this dataset there just so happens to be the columns `ebitda` and `sale` (which is the Revenue that you need for the EBITDA margin calculation). Another metric is the **Return on Total Assets (ROTA)**, calculated by the ratio of EBIT to total assets (ebit and `at` in our dataset).

You decided to calculate these two ratios for those observations in which sales are greater than 100 **billion** (i.e. `sale >100000`). First remove the other observations. Then, create the **EBITDA margin** and **ROTA** columns. Make sure to screen out observations with `NA` values for relevant columns (`ebitda, sale, ebit, at`).

Company with the ticker (i.e., `tic`) value of 3FNMA achieved the highest EBITDA margin in the fyear of 2013 .

Company with the ticker (i.e., `tic`) value of AAPL achieved the highest ROTA in the fyear of 2012 .

Answer 1:

Correct!

3FNMA

Answer 2:

Correct!

2013

Answer 3:

Correct!

AAPL

Answer 4:

Correct!

2012

```
subset6 <- companies %>%
  filter(!is.na(ebitda), !is.na(ebit), !is.na(at), !is.na(sale), sale > 100000) %>%
  mutate(em = ebitda/sale) %>%
  mutate(ROTA = ebit/at)

# highest em
subset6$tic[which(subset6$em==max(subset6$em, na.rm=TRUE))]
subset6$fyear[which(subset6$em==max(subset6$em, na.rm=TRUE))]
# alternatively
subset6 %>% filter(em==max(em,na.rm=TRUE)) %>% select(tic, fyear)

# highest ROTA
subset6$tic[which(subset6$ROTA==max(subset6$ROTA, na.rm=TRUE))]
subset6$fyear[which(subset6$ROTA==max(subset6$ROTA, na.rm=TRUE))]
#Alternatively
subset6 %>% filter(ROTA==max(ROTA, na.rm=TRUE)) %>% select(tic, fyear)
```

Question 6

1 / 1 pts

Assume we have a data frame (**not** the North American Stock Market dataset we've been

using) named `dat0` with three variables and 10 observations. The first variable is the string variable `city`. The second variable is the string variable `province`, while the last variable is the numerical variable `population`. There are no missing values anywhere in the dataset.

There are three observations where `province` is equal to `BC`, three observations where `province` is equal to `AB`, two observations where `province` is equal to `ON`, and two observations where `province` is equal to `MB`.

You run the following code:

```
dat1 <- dat0 %>% group_by(province) %>%
summarise(howmany = n())

View(dat1)
```

Evaluate each of the following statements **independently**. For each statement, determine whether it is **TRUE**, **FALSE**, or **UNCERTAIN**.

R will output a data frame with ten observations. FALSE

R will output a data frame with three columns. FALSE

R will create a new variable called `howmany`, and this variable will be a numerical variable. TRUE

The data frame has two observations where `howmany` is equal to 3. TRUE

The data frame has two different values for the variable `howmany`. TRUE

Answer 1:

Correct!

FALSE

Answer 2:

Correct!

FALSE

Answer 3:

Correct!

TRUE

Answer 4:

Correct!

TRUE

Answer 5:

Correct!

TRUE

Quiz Score: **6.4** out of 10