In-class Exercise 11 Results for Simran Mander

Score for this attempt: **7** out of 10
Submitted Nov 15 at 11:51pm
This attempt took 1,950 minutes.

Question 1

1.75 / 1.75 pts

Questions 1-5 will be based on the North American Stock Market 1994-2013 dataset.

The dataset has 232362 rows and 50 columns.

How many `NA`'s in total are there in the data set (Hint: `is.na()` would also work with data frames)?

1103773

What is the minimum value of the sale column (i.e., `sale`) ? Please round your result to the first decimal place.

-15009.3

What is the maximum value of the number of employees column (i.e., `emp`) ? Please round your result to the first decimal place.

2545.2

Answer 1:

Correct!

232362

Answer 2:

Correct!

50

Answer 3:

Correct!

1103773

Answer 4:

Correct!

-15009.3

Answer 5:

Correct!

2545.2

```
View(companies)

sum(is.na(companies))

round(min(companies$sale,na.rm=TRUE),1)

round(max(companies$emp,na.rm=TRUE),1)
```

Question 2

1.75 / 1.75 pts

Create another data frame which contains all the observations with total assets greater than $50 million (i.e., `at>50`) from `fyear 1995` (inclusive) to `2005` (inclusive). Call this data frame as `df`.

This new dataset (i.e., `df`) has 76828 rows and 50 columns.

Suppose we use the `summarise` function on the new dataset (i.e., `df`) to obtain the mean of sales (i.e `sale`) and standard deviation of the number of employees (i.e., emp) and saved the resultant data frame as `df2`.

We would obtain a dataset with 1 rows and 2 columns.

Using `View()` function, we can see that the mean of sales is 2303.616 and the standard deviation of number of employees is 37.85884 .

Note: Given that there are NAs in the dataset, it is safe to assume that we should use na.rm

=TRUE.

Answer 1:

Correct! 76828

Answer 2:

Correct! 50

Answer 3:

Correct! 1

Answer 4:

Correct! 2

Answer 5:

Correct! 2303.616

Answer 6:

Correct! 37.85884

orrect Answer 37.858840

```
df <- companies %>% filter(fyear>=1995, fyear<=2005, at>50)

#View(df) #76828 rows 50 columns

df2 <- df %>% summarise(mean_sale=mean(sale, na.rm = TRUE),
sd_emp =sd(emp, na.rm=TRUE))

df2
```

Question 3

0 / 1.75 pts

Now, create a new data frame by removing all the observations containing `NA`'s in the North America Stock Market 1994-2013 dataset. Call this data frame `companies_complete`. Hint: This can easily be done by using `na.omit` function. If your original data frame is named as `companies`, then

`companies_complete <- na.omit(companies)` would do the trick.

Note: you will use this `companies_complete` data frame in questions 4-5 as well.

Use `companies_complete` data frame. Add a new column called `x` which is equal to `1` if the company is incorporated (i.e. `fic`) in Canada (i.e., `"CAN"`), equal to `0` if the company is incorporated in the United States (i.e., `"USA"`), and equal to `-1` if the company is incorporated elsewhere (not in`"CAN"` or in `"USA"`).

For observations for which `x== 1`, the mean value of Net Income (i.e., `ni)` is 65.9567 . (round the value to its 4th decimal place. You can do this by using `round( ,4)`)

For observations for which `x== 0`, the mean value of Net Income (i.e., `ni)` is 79.3550 . (round the value to its 4th decimal place)

For observations for which `x== -1`, the mean value of Net Income (i.e., `ni)` is 533.8710 .(round the value to its 4th decimal place)

Answer 1:

You Answered: 65.9567

Correct Answer: 75.4604

Answer 2:

You Answered: 79.3550

Correct Answer: 76.8298

Answer 3:

ou Answered

533.8710

orrect Answer

603.9162

```
companies_complete <- na.omit(companies)

companies_complete$x <- -1

companies_complete$x[companies_complete$fic=="CAN"] <- 1

companies_complete$x[companies_complete$fic=="USA"] <- 0


companies_complete %>% filter(x==1) %>%
summarise(round(mean(ni),4))

companies_complete %>% filter(x==0) %>%
summarise(round(mean(ni),4))

companies_complete %>% filter(x==-1) %>%
summarise(round(mean(ni),4))
```

Question 4

1.75 / 1.75 pts

Use `companies_complete` data frame created in Q3.

Create a new data frame with all the observations where the fiscal year is 1999 (`fyear==1999`) and total liabilities (i.e., `lt`) **not** equal to `0`.

What is the highest total assets (i.e., `at`) to total liabilities (i.e., `lt`) ratio?

666

What is the `gvkey` value of the company which has the highest of this ratio (i.e., `at` to `lt` ratio) ? Note: gvkey is one of the columns.

021345

Answer 1:

Correct!

666

Answer 2:

Correct!

021345

```
df3 <- companies_complete %>% filter(fyear==1999, lt!=0) %>%
mutate(at_to_lt= at/lt)

# By using dplyr
df3 %>% summarise(max(at_to_lt))
df3 %>% filter(at_to_lt==max(at_to_lt)) %>% select(gvkey)

# by using subsetting

max(df3$ratio,na.rm=TRUE)

df3$gvkey[which(df3$ratio == max(df3$ratio,na.rm=TRUE))]
```

Question 5

1.75 / 1.75 pts

Use `companies_complete` data frame created in Q3.

In the fiscal year 2005 (i.e., `fyear==2005`), what percentage of the companies had total assets over $100 million (`at>100`)? Please round the value to 3rd decimal place.

0.574

In the fiscal year 2005 (i.e., `fyear==2005`), what percentage of the companies had sales over $100 million (`sale>100`)? Please round the value to 3rd decimal place.

0.539

Hint 1: you can use `n()` as an aggregate function to count the number of rows in a data

frame passed into `summarise()` function. For example, in the `iris` dataset you have used, if you want to find the number of flowers whose Species is virginica, you can use the following:

```
iris %>% filter(Species == "virginica") %>%
summarise(numberofvirginica = n())
```

Above `n()` counts the number of rows in the data frame (i.e., `iris %>% filter(Species == "virginica")`) passed into `summarise()`.

Hint 2: Numerator and denominator of each of the ratios can be calculated separately.

Answer 1:

Correct!

0.574

orrect Answer

57.422

Answer 2:

Correct!

0.539

orrect Answer

53.872

```
# what percentage of the companies had total assets
over $100 million
numerator1 <- companies_complete %>% filter(fyear
== 2005, at>100) %>% summarise(n())
denominator <- companies_complete %>% filter(fyear
== 2005) %>% summarise(n())
round(numerator1/denominator,3)

round(100*numerator1/denominator,3) # to give the percentage

# what percentage of the companies had total sales
over $100 million
numerator2 <- companies_complete %>% filter(fyear
== 2005, sale>100) %>% summarise(n())
round(numerator2/denominator,3)

round(100*numerator2/denominator,3) # to give the percentage
```

Question 6

0 / 1.25 pts

Suppose you have a data frame named `dat` with three columns and 1000 observations (`dat` is **not** the North American Stock Market dataset we have been using above). Three columns are

- a numerical column `year`,
- a string column `city`, and
- a numerical column `population`.

There are no missing values anywhere in the dataset.

You would like to obtain observations only from year 1995 (inclusive) to year 1999 (inclusive). Which of the following will return the desired output? (there is at least one correct option; you must select all the right options)

ou Answered

- [x] `dat %>% filter(year == c(1995,1996,1997,1998,1999))`
- [ ] `dat %>% filter(year == c(1995,1999))`
- [ ] `dat %>% filter(year == c(1995:1999))`

orrect Answer

- [ ] `dat %>% filter(year <= 1999, year >= 1995)`

Correct!

- [x] `dat %>% filter(year <= 1999 & year >= 1995)`
- [ ] `dat %>% filter(year <= 1999 | year>= 1995)`

In filter, you could either use comma to separate the conditions or combine the conditions with &. Hence, the options 4 and 5 are the right answers.

The first three options would not work. The vectors at either side of the the comparison had different length.

For your reference, the following two would have also worked. Since we have not seen %in%, you are not responsible for it.

dat %>% filter(year %in% c(1995,1996,1997,1998,1999))

dat %>% filter(year %in% c(1995:1999))

Quiz Score: **7** out of 10