STA035B Midterm 1, Winter 2024

Name _____

Student ID _____

problem	points
1a	
1b	
2a	
2b	
2c	
3	
4a	
4b	
total	

Problem 1

Consider the following code.

```
scores <- tribble(
  ~name, ~midterm1, ~midterm2,
  "Mary", 80, 90,
  "Jose", NA, 100,
  "Ali", 75, 95,
)
cleaned_scores <- scores %>%
  mutate(
    midterm1 = replace_na(midterm1, 100),
    midterm2 = replace_na(midterm2, 100)
)
```

For each of the following, draw the tibble which results from the following code. With words, describe how many rows and columns there are in the resulting tibble, and say whether or not there are missing values and, if there are any, where they appear in the tibble.

(a) 5 points:

```
scores %>%
  mutate(a = pmin(midterm1, midterm2))
```

```
# A tibble: 3 x 4
  name  midterm1 midterm2     a
  <chr>    <dbl>    <dbl> <dbl>
1 Mary        80       90    80
2 Jose        NA      100    NA
3 Ali         75       95    75
```

The tibble has 3 rows and 4 columns. The row for Jose has missing values for `midterm1` and `a`, and these are the only missing values.

(b) 5 points:

```
cleaned_scores %>%
  mutate(c = pmin(midterm1, midterm2))
```

```
# A tibble: 3 x 4
  name  midterm1 midterm2     c
  <chr>    <dbl>    <dbl> <dbl>
1 Mary        80       90    80
2 Jose       100      100   100
3 Ali         75       95    75
```

The tibble has 3 rows and 4 columns. There are no missing values.

Problem 2

Suppose we have a tibble `flights` whose first few rows look like this:

time_hour	carrier	flight	tailnum	origin	dest	air_time
2013-01-01 05:00:00	UA	1545	N14228	EWR	IAH	227
2013-01-01 05:00:00	UA	1714	N24211	LGA	IAH	227
2013-01-01 05:00:00	AA	1141	N619AA	JFK	MIA	160
2013-01-01 05:00:00	DL	725	N804JB	JFK	BQN	183
2013-01-01 06:00:00	DL	461	N668DN	LGA	ATL	116
2013-01-01 05:00:00	UA	1696	N39463	EWR	ORD	150

Describe the outputs of the following lines of code.

(a) 5 points:

```
flights %>%
  group_by(origin) %>%
  summarize(n = n())
```

```
# A tibble: 3 x 2
  origin      n
  <chr>   <int>
1 EWR    120835
2 JFK    111279
3 LGA    104662
```

This is valid code. It computes the number of flights per origin: it returns a tibble with two columns, one with `origin` and another with `n`.

(b) 5 points:

```
str_remove(flights$dest, '^[AEIOU]')
```

This is valid code. It returns the vector of strings with values given by the `dest` variable in `flights`, except that if `dest` begins with a capital vowel, then it removes the first letter of `dest`.

Problem 3 (3 points)

Consider the following tibbles:

```
df1 <- tribble(
  ~product, ~q1, ~q2,
  "A", 150, 200,
  "B", 120, 180
)

df2 <- tribble(
  ~product, ~quarter, ~sales,
  "A", "q1", 150,
  "A", "q2", 200,
  "B", "q1", 120,
  "B", "q2", 180
)
```

Which of the following code correctly transforms df2 into df1?

(A) `df2 %>% pivot_wider(id_cols = c(product, quarter), names_from = quarter, values_from = sales)`

(B) `df2 %>% pivot_wider(id_cols = product, names_from = quarter, values_from = sales)`

(C) `df2 %>% pivot_wider(id_cols = c(product, quarter), names_from = sales, values_from = quarter)`

(D) `df2 %>% pivot_wider(id_cols = quarter, names_from = product, values_from = sales)`

(B) is the correct answer

Problem 4 (5 points)

Consider the following vector of strings.

```
strings <- c("William;Order 1", "Jenny;order 2", "Alex;order 25")
```

Suppose we want to use regex to return the strings vector but where we erase the name preceding the semicolon and delete the semicolon. Which of the following options correctly does this task? Explain. (If you get the answer correct, you don't need an explanation. If you get it incorrect, any explanations for why some of the options are incorrect can get you partial points.)

(A) `str_remove(strings, "^\\b+;")`

(B) `str_remove(strings, "^\\w+;")`

(C) `str_remove(strings, "$[a-z]*;")`

(D) `str_remove(strings, "$[A-Za-z]+;")`

(B) is the correct answer.

The first line is not valid code, `\\b` must appear twice to enclose something. The third and fourth options don't do anything since we are starting with "$" which indicates the **end** of a string, thus no parts of the string ar ematched.

Problem 5

Consider the two following tibbles:

```
majors <- tribble(
  ~student_id, ~major,
  123, "Math",
  234, "Statistics",
  345, "Literature",
)

grades <- tribble(
  ~student_id, ~course, ~grade,
  345, "Machiavelli", "B",
  123, "Analysis", "A",
  456, "Organic Chemistry", "C"
)
```

For each of the following, draw the tibble which results from the following code. With words, describe how many rows and columns there are in the resulting tibble, and describe any missing values.

(a) 5 points:

```
majors %>% left_join(grades)
```

```
Joining with `by = join_by(student_id)`
```

```
# A tibble: 3 x 4
  student_id major      course      grade
       <dbl> <chr>      <chr>       <chr>
1        123 Math       Analysis    A
2        234 Statistics <NA>        <NA>
3        345 Literature Machiavelli B
```

There are 3 rows, 4 columns. There are missing values for `student_id = 234` for `course` and `grade`.

(b) 5 points:

```
grades %>% left_join(majors)
```

```
Joining with `by = join_by(student_id)`
```

```
# A tibble: 3 x 4
  student_id course            grade major
       <dbl> <chr>             <chr> <chr>
1        345 Machiavelli       B     Literature
2        123 Analysis          A     Math
3        456 Organic Chemistry C     <NA>
```

There are 3 rows, 4 columns. There are missing values for `student_id = 456` for `major`.