STA035B Midterm Practice

Spencer Frei

This provides an idea of what types of questions and material you will have on your midterm exam.

Problem 1

Consider the following code.

```
scores <- tribble(
  ~name, ~wage, ~hours,
  "Mary", 25, 40,
  "Jose", 35, 38,
  "Ali", 37, NA,
  "Pat", NA, 42
)
cleaned_scores <- scores %>%
  mutate(
    wage = replace_na(wage, 0),
    hours = replace_na(hours, 0)
)
```

For each of the following, explain if the code is a valid command. If it is, describe what the output of the command is (if it is a tibble, draw the tibble). If it not, explain why.

a.

```
scores %>%
  mutate(a = wage * hours)
```

b.

```
scores %>%
  mutate(b = product(wage, hours, na.rm=TRUE))
```

c.

```
cleaned_scores %>%
  mutate(c = wage * hours)
```

Problem 2

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

origin	temp	dewp	humid	wind_dir	wind_speed	wind_gust	pressure	visib	time_hour
EWR	39.02	26.06	59.37	270	10.35702	NA	1012.0	10	2013-01-01 01:00:00
EWR	39.02	26.96	61.63	250	8.05546	NA	1012.3	10	2013-01-01 02:00:00
EWR	39.02	28.04	64.43	240	11.50780	NA	1012.5	10	2013-01-01 03:00:00
EWR	39.92	28.04	62.21	250	12.65858	NA	1012.2	10	2013-01-01 04:00:00
EWR	39.02	28.04	64.43	260	12.65858	NA	1011.9	10	2013-01-01 05:00:00
EWR	37.94	28.04	67.21	240	11.50780	NA	1012.4	10	2013-01-01 06:00:00

For each of the following, explain if the code is a valid command. If it is, describe what the output of the command is. If it not, explain why.

a.

```
weather[,"EWR"]
```

b.

```
weather %>%
  group_by("origin") %>%
  summarize(a = mean(wind_gust))
```

c.

```
str_replace(weather$origin, '[ER]', 'X')
```

Problem 3

Consider the following vector of strings.

```
strings <- c("William; Grade: A", "Jenny; Grade: B-", "Alex; Grade: B+")
```

Suppose we want to use regex to return a vector containing strings that indicate only the student's grades (i.e., A, B-, B+ for the vector `strings`). Which of the following options correctly does this task?

(A) `str_replace(strings, "^[A-Za-z]+: (.*)", "\\1")`

(B) `str_replace(strings, ".*: .*", "\\2")`

(C) `str_replace(strings, "(.*): (.*)", "\\2")`

(D) `str_replace(strings, "^[A-Za-z]*: (.*)", "\\2")`

Problem 4

Suppose we have a tibble `car_data` which has the following variables:

- "city", a string
- "day_number", an integer describing which day of the year it was (1-365),
- "year", an integer,
- "accidents", an integer, the number of accidents observed that day,
- "rained", a boolean, indicating whether there was rain in that city on that day.

Every observation in `car_data` corresponds to an observation for a city on a given day, and assume there is no missing data.

Suppose we want to compute the average number of accidents per day when it is raining vs. when it is not raining, for every city in the tibble: namely, we want to be able to look at what the average number of accidents per day is under different raining conditions in every city. Which of the following options correctly does this task?

(A)
```
car_data %>%
    group_by(city, rained, day_number) %>%
    summarize(avg_accidents = mean(accidents))
```

(B)
```
car_data %>%
    group_by(city, rained) %>%
    summarize(avg_accidents = mean(accidents))
```

(C)
```
car_data %>%
    summarize(avg_accidents = mean(accidents))
```

(D)
```
car_data %>%
    summarize(avg_accidents = mean(accidents)) %>%
    group_by(city, rained)
```