Fall 2022 STAT 240 In-Person (part 1) Midterm

□

1st Letter of Family Name	Family Name	Given Name	Student ID

Instructor (Circle) **Bret Larget** **Bi Cheng Wu**

Lecture Time (Circle) 8:50 - 9:40 **9:55 - 10:45** **1:20 - 2:10**

Discussion (Circle)

TA	time 1	time 2	time 3
Ajinkya Kokandakur	M 2:25 pm	T 7:45 am	T 4:35 pm
Cameron Jones	M 1:20 pm	T 8:50 am	T 9:55 am
Hailey Louw	T 7:45 am	T 8:50 am	T 9:55 am
Dibyendu Sah	T 8:50 am	T 9:55 am	T 11:00 am
Congwei Yang	T 8:50 am	T 9:55 am	T 11:00 am
Kessys Lorranya Peralta De Oliveira	M 2:25 pm	M 3:30 pm	
Ryan Yee	M 2:25 pm	M 4:35pm	
Margaret Turner	M 3:30 pm	M 4:35 pm	

Instructions:

1. You may use one regular sheet of paper with any handwritten notes you like. You may use both sides of the paper.
2. You may not consult other resources, your phone, a computer, information online, nor your neighbor's exam.
3. Do all of your work in the space provided. Use the backs of pages if necessary, indicating clearly that you have done so (so the grader can easily find your complete answer).

Sections

- Name, Lecture, and Discussion (2 points)
- Multiple Choice (24 points, eight questions worth 3 points each).
- Short Answer (12 points, three questions worth 4 points each).
- Data Analysis (12 points, three questions worth 4 points each).

Scoring

Problems	Possible	Score
name	2	
1 - 8	24	
9 - 11	12	
12 - 15	12	
TOTAL	50	

Multiple Choice (24 points). Each problem is worth 3 points.

Circle the correct answer or answers as indicated.

Problem 1. You have a dataset with a numerical column Y and a categorical (i.e. character) column X. Suppose you wanted to visualize just the Y column to see the distribution of its values in your sample. Which of the following plots would be reasonable and useful to make? **Circle all correct responses**.

(a) `geom_line()`
(b) `geom_bar()`
(c) `geom_histogram()`
(d) `geom_density()`
(e) `geom_col()`
(f) `geom_point()`
(g) `geom_smooth()`

Problem 2. Which of the following produces the proportion of values in a column X that are positive?

(a) `sum(X > 0)`
(b) `mean(X > 0)`
(c) `count(X > 0)`
(d) `length(X > 0)`

Problem 3. Circle **all** valid names for an R object.

(a) `.bucky.badger.` (b) `bucky_badger` (c) `bucky-badger` (d) `bucky+badger`

Problem 4. Which command keeps all rows of the official Madison weather data set `official` where the maximum temperature `tmax` is **not missing** and the temperature is between 32 and 60 degrees Fahrenheit (include these two boundary values)?

(a) `official %>% filter( tmax != "NA" & between(tmax, 32, 60) )`
(b) `official %>% filter( tmax != "NA" && (tmax >= 32 | tmax <= 60) )`
(c) `official %>% filter( !is.na(tmax) & tmax > 31 & tmax < 61 )`
(d) `official %>% filter( !is.na(tmax) & (32 <= tmax <= 72) )`

Problem 5. You run `df %>% group_by(type) %>% summarise(n=n(),sum=sum(x),sd=sd(x))` and get the following output:

type	n	sum	sd
A	10	81	4.6
B	7	50	5.8
C	3	29	2.7

When you run `df %>% summarise(m=mean(x))`, the final number you get is closest to which number?

(a) 5 (b) 6 (c) 7 (d) 8 (e) 9

Problem 6. Which lubridate function converts the string “10-2022-14” into the date October 14, 2022?

(a) `dmy()` (b) `dym()` (c) `myd()` (d) `mdy()` (e) `ymd()`

Problem 7. A data set grocery_items has variables named item, type, and price. A data set named grocery_list has variables named item and n. The values in the columns item match if the same item is part of both data sets. Some items in grocery_items may not by in grocery_list and some items in grocery_list may not by in grocery_items. Which description matches the contents of df after executing the following code? No items are repeated within either data set. **Circle the correct response.**

```
df = grocery_list %>%
  left_join(grocery_items, by = "item")
```

(a) A data frame with one row for each item in both data sets and columns item and n only.
(b) A data frame with one row for each item in both data sets and columns item, n, type and price.
(c) A data frame with one row for each item in grocery_list and columns item, n, type and price.
(d) A data frame with one row for each item in either data set, columns item, n, type and price, and the value NAin columns n, type, and price in rows where this information was missing in the corresponding data set.

Problem 8

A data set obesity has no missing data, 500 rows, and columns zip, age_5_17, age_18_34, age_35-54, age_55-74, and age_75+. The values in zip are zip codes and the values in the other columns contain counts of residents of these zip codes in the corresponding age ranges. What will be the dimensions of the data set created with the following code?

```
df = obesity %>%
  pivot_longer(starts_with("age"), names_to = "age", values_to = "n")
```

Circle all correct responses.

(a) 500 rows and 5 columns
(b) 500 rows and 6 columns
(c) 2500 rows and 3 columns
(d) 2500 rows and 6 columns

Short Answer (12 points). Each problem is worth 4 points

Problems 9-11 are based on this small data set `df` which has numerical variables `a` and `b` and a categorical variable named `group`.

```
##   a  b group
## 1 1  3     X
## 2 2 -2     Z
## 3 3  0     Y
## 4 4  4     Z
## 5 5 -1     Z
## 6 6  5     X
```

Problem 9. Write the result of the following code.

```
df %>%
  mutate(c = a + b) %>%
  filter(group == "X") %>%
  arrange(desc(c))
```

Problem 10 Write the result of the following code.

```
df %>%
  group_by(group) %>%
  summarize(v = min(b)) %>%
  arrange(group)
```

Problem 11 Write the result of the following code.

```
df %>%
  filter(b > 0) %>%
  select(-b) %>%
  group_by(group) %>%
  mutate(v = sum(a))
```

Data Analysis (12 points). Three problems worth 4 points each.

Each problem asks you to interpret the output from the following data analysis of the official historical Madison weather data set. The variable `prcp` measures daily precipitation in inches.

Read the questions before attempting to read the code. Only read what is needed to answer the questions.

```
## Summaries of precipitation data
prcp_1 = official %>%
  select(date, prcp) %>%
  drop_na() %>%
  mutate(year = year(date),
         month = month(date, label = TRUE),
         day = day(date),
         wday = wday(date, label = TRUE)) %>%
  group_by(year, month) %>%
  summarize(n = n(),
            v1 = max(prcp),
            v2 = sum(prcp > 0),
            v3 = sum(prcp)) %>%
  ungroup()

prcp_2 = prcp_1 %>%
  group_by(month) %>%
  summarize(w1 = min(v1),
            w2 = mean(100*v2/n),
            w3 = mean(v3))

prcp_2
```

```
## # A tibble: 12 x 4
##    month    w1    w2    w3
##    <ord> <dbl> <dbl> <dbl>
##  1 Jan    0.04  30.6  1.40
##  2 Feb    0.02  28.5  1.34
##  3 Mar    0.08  31.8  2.10
##  4 Apr    0     36.2  2.86
##  5 May    0.19  37.8  3.56
##  6 Jun    0.1   35.9  4.09
##  7 Jul    0.13  30.1  3.79
##  8 Aug    0.21  29.3  3.56
##  9 Sep    0.06  30.7  3.44
## 10 Oct    0     28.3  2.36
## 11 Nov    0.01  29.2  2.01
## 12 Dec    0.04  30.0  1.61
```

Problems 12-14 ask you what values in the summary table represent. Example solutions include the following:

- The average daily precipitation.
- The smallest total monthly precipitation in a single year.
- The average number of days per month with positive precipitation totals.

Problem 12 What does the value `w1` in the summary table represent for each month?

Problem 13 What does the value `w2` in the summary table represent for each month?

Problem 14. What does the value `w3` in the summary table represent for each month?