STAT 331/531 – Midterm Exam
February 15, 2024

This is a three-part exam.

1. General Questions (20 points)
   - This section will be completed in class, on paper, without any resources.
   - This section must be completed first.
   - You will submit the attached question/answer sheet.
2. Short Answer (30 points)
   - This section will be completed in class, on your computer, with any non-human resources.
   - You will submit both your .html and .qmd files.
3. Open-Ended Analysis (42 points)
   - This section will be started in class (finished outside of class), on your computer, with any non-human resources.
   - You will submit both your .html and .qmd files.

Please note:

- You have 1 hour and 50 minutes to complete the in-class portion of the exam.
  - Late uploads will automatically be deducted 2 points per minute!
- The take-home portion is due 24-hours after the end of the in-class exam.
- The problems on this exam do not necessarily need to be completed in order – if you cannot accomplish problem 1, you may still be able to accomplish problem 2.
- Don't spend too long on one question. Point allocations have been given to each problem to help you manage your time.
- If any questions arise during the exam, please do not hesitate to ask!

Resource Policies

- For parts 2 and 3, you may use any online resources, including anything posted on Canvas, in the text, or in your past assignments.
- You may use ChatGPT as a *resource* but not as a *creator*. Any code you include should have been seen in class.
- You may NOT contact anyone, inside or outside this class, during the course of the exam. This includes email, chat/messenger services, and posting on online forums and message boards.
- You may NOT discuss the exam with any other students until after the exams have been returned to all students.

Name: ______________________________

Section (circle one): 9am 12pm

General Questions

__________ /20 pts.

1. (1 pt.) How often do you need to load an R package?

   A. Only once.

   B. Never, as long as you are connected to the internet.

   C. Every time you restart R.

   D. Every time you open a new R file.

2. (1 pt.) You want to select certain rows of a data frame. Which is the most appropriate function?

   A. `arrange` B. `select` C. `filter` D. `mutate`

3. (1 pt.) You need a one-dimensional object containing elements that are all of the same data type. Which is most appropriate?

   A. vector B. list C. data frame D. tibble E. matrix

4. (1 pt.) Fill in the blank.

   The package `ggplot2` implements the ________________________ of Graphics.

5. (1 pt.) You want to select certain columns of a data frame. Which is the most appropriate function?

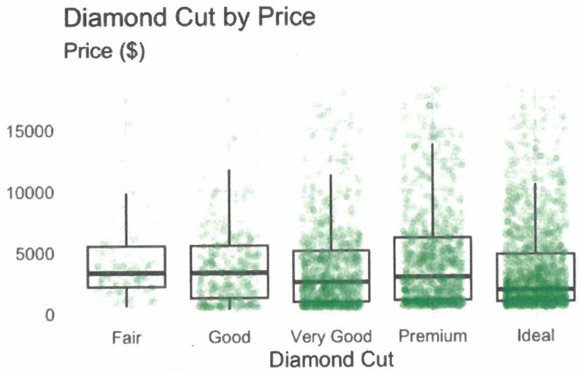
   A. `arrange` B. `select` C. `filter` D. `mutate`

6. (1 pt.) What is the data type of the variable `x`?

```
> x
 [1] SI2  SI1  VS1  VS2  SI2  VVS2 VVS1 SI1  VS2  VS1
Levels: I1, SI2, SI1, VS2, VS1, VVS2, VVS1, IF
```

   A. numeric B. character C. logical D. string E. factor

7. (4 pts.) Fill in the blanks in the following snippet of code to create this plot.



```
diamonds |>

    _______________ (aes(x = cut, y = price)) +

       _________________ (outlier.shape = NA) +

       _________________ (alpha = 0.1,

                     __________ = "forestgreen")
```

8. (4 pts.) Consider the `cereal` dataset that we have used in class. What `dplyr` verbs are needed to create the following table? Write one verb per blank.

shelf <int>	avg <dbl>	sd <dbl>	upper <dbl>	lower <dbl>
1	75.50	38.55	152.61	-1.61
2	60.75	49.07	158.90	-37.40
3	133.57	78.72	291.02	-23.88

```
cereal |>

    _____________ (potass >= 0) |>

    _____________ (shelf) |>

    _____________ (avg = mean(potass),

                   sd = sd(potass)) |>

    _____________ (upper = avg + 2*sd,

                   lower = avg - 2*sd)
```

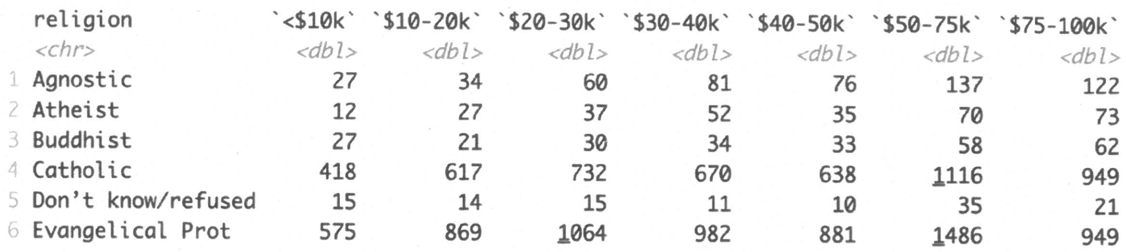
9. (1 pt.) In order to control the order of the bars in the bar chart created by the code below, what data type should `var` be?

```
data |>
    ggplot(aes(x = var) +
    geom_bar()
```

A. character B. data frame C. integer D. factor E. vector

10. (4 pts.) Fill in the blanks in the code to complete the following data transformation.

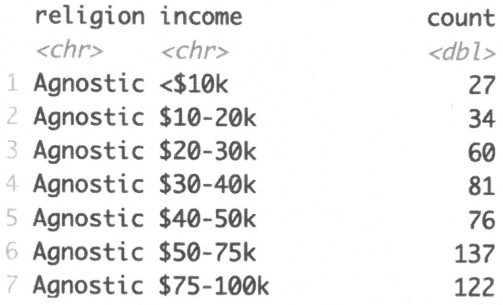
```
head(relig_income)
```



```
  religion                `<$10k` `$10-20k` `$20-30k` `$30-40k` `$40-50k` `$50-75k` `$75-100k`
  <chr>                     <dbl>     <dbl>     <dbl>     <dbl>     <dbl>     <dbl>      <dbl>
1 Agnostic                     27        34        60        81        76       137        122
2 Atheist                      12        27        37        52        35        70         73
3 Buddhist                     27        21        30        34        33        58         62
4 Catholic                    418       617       732       670       638      1116        949
5 Don’t know/refused           15        14        15        11        10        35         21
6 Evangelical Prot            575       869      1064       982       881      1486        949
```

```
new_data <- ______________ |>
    _______________ (cols = !religion,
                     names_to = "_________________",
                     values_to = "________________")

head(new_data)
```



```
  religion income   count
  <chr>    <chr>    <dbl>
1 Agnostic <$10k       27
2 Agnostic $10-20k     34
3 Agnostic $20-30k     60
4 Agnostic $30-40k     81
5 Agnostic $40-50k     76
6 Agnostic $50-75k    137
7 Agnostic $75-100k   122
```

11. (1 pt.) Which line of code would return `TRUE TRUE TRUE`? In other words, which of the regular expressions match kryptonite, pillow, and plump?

```
x <- c("kryptonite", "pillow", "plump")
```

A. `str_detect(x, pattern = "^p")`

B. `str_detect(x, pattern = "[^p]")`

C. `str_detect(x, pattern = "[^p]$")`

D. `str_detect(x, pattern = "p{2}")`