MATH 167R: Fall 2023

Final Exam

Name:______________________________

SJSU ID:______________________________

Instructions:

- Read all directions carefully and write your answers on the lines provided.
- Notes, books and all types of electronic devices must be put away. Failure to do so will result in a 0 for the exam.
- Please write darkly enough that your writing will show up on a computer scanner.
- The point total for this exam is 100 points.

Good luck!

1. Consider the following code:

```
mystery <- function(x) {
    if (x == 1) {
        return(x ^ 2)
    } else {
        return(x ^ 2 + mystery(x - 1))
    }
}
```

   (a) (4 points) What is the value of the expression `mystery(4)`?

   Answer: ____________________

   (b) (6 points) In the space below, write a new version of this function `mystery2()` that does the same task without using recursion.

2. (4 points) Which two of the four following lines of code are equivalent?

   a. `"hello" |> print()`
   b. `x <- "hello" |> print()`
   c. `print("hello")`
   d. `print |> "hello"`

   Answer: ____________________

3. (5 points) Suppose we want to write code to play the game FizzBuzz. In particular, we want to write out code that prints out the numbers 1 - 100, except that any number divisible by three (but not five) is replaced with the word "fizz", and any number divisible by five (but not three) is replaced with the word "buzz", and any number divisible by both three and five is replaced with the word "fizzbuzz." Explain the error with the following implementation:

```
for (i in 1:100) {
  if (i %% 3 == 0) {
    print("fizz")
  } else if (i %% 5 == 0) {
    print("buzz")
  } else if (i %% 3 == 0 & i %% 5 == 0) {
    print ("fizzbuzz")
  } else {
    print(i)
  }
}
```

4. (4 points) A sample of 31 trained typists was selected, and the preferred keyboard height was determined for each typist. The resulting sample average preferred height was 81cm. Assuming that the preferred height is normally distributed with standard deviation 3cm, which of the following lines of code will produce an appropriate P-value for a two-sided hypothesis test of the null hypothesis that the population mean $\mu = 80$ and the alternative hypothesis $\mu > 80$?

   a. `pnorm((81 - 80) / (3 / sqrt(31)))`
   b. `1 - pnorm((81 - 80) / (3 / sqrt(31)))`
   c. `qnorm((81 - 80) / (3 / sqrt(31)))`
   d. `1 - qnorm((81 - 80) / (3 / sqrt(31)))`

Answer: ______________________

5. This question refers to the `penguins` dataset in the `palmerpenguins` package.

   (a) (3 points) Which of the following R expressions is not equivalent to the others?

      a. `penguins$body_mass_g`
      b. `penguins[, "body_mass_g"]`
      c. `penguins[body_mass_g]`
      d. `penguins[["body_mass_g"]]`

   Answer: ______________________

   (b) (3 points) Which of the following commands will add a `body_mass_kg` column to the dataset which gives the body mass of each penguin in kilograms?

      a. `penguins |> mutate(body_mass_kg = body_mass_g / 1000)`
      b. `penguins |> append(body_mass_kg = body_mass_g / 1000)`
      c. `penguins |> mutate(body_mass_kg * 1000 = body_mass_kg)`
      d. `penguins |> append(body_mass_kg * 1000 = body_mass_kg)`

   Answer: ______________________

   (c) (3 points) Which of the following commands will create a table with three rows containing the mean body masses of the penguins for each of the three species?

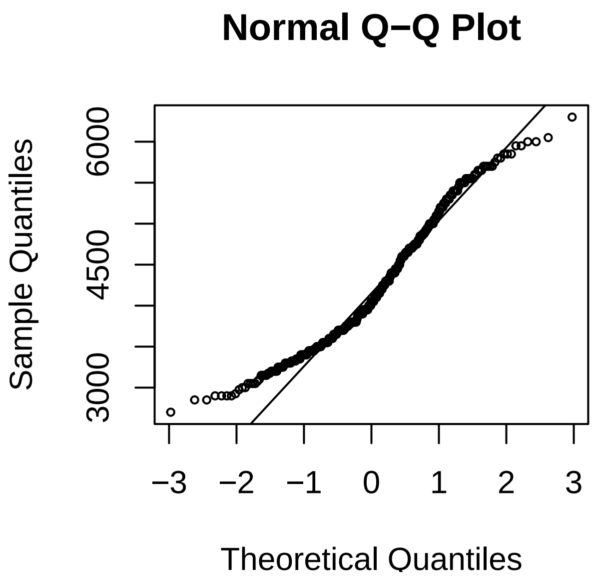
      a. `penguins |> filter(species) |> summarize(mean_mass = mean(body_mass_g))`
      b. `penguins |> summarize(mean_mass = mean(body_mass_g[species]))`
      c. `penguins |> group_by(species) |> summarize(mean_mass = mean(body_mass_g))`
      d. `penguins |> group_by(species) |> mutate(mean_mass = mean(body_mass_g))`

   Answer: ______________________

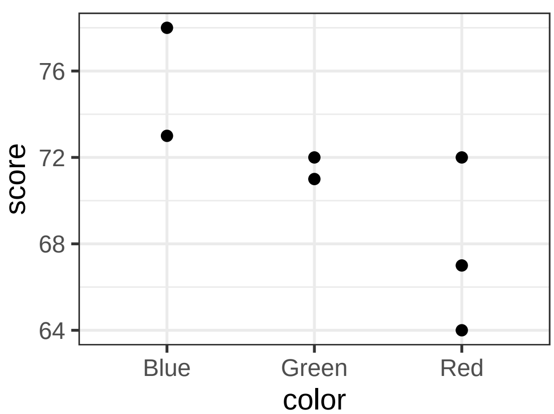
   (d) (3 points) Write a line of code that will create a new data frame called `adelie` containing only the rows of `penguins` for which the value of the `species` column is `"Adelie"`.

   Answer: ____

6. (5 points) The following quantile-quantile plot compares the quantiles of `penguins$body_mass_g` with the quantiles of a normal distribution. In the space below, describe whether `body_mass_g` appears to be normally distributed. If not, how does the shape of the distribution differ from that of a normal distribution?

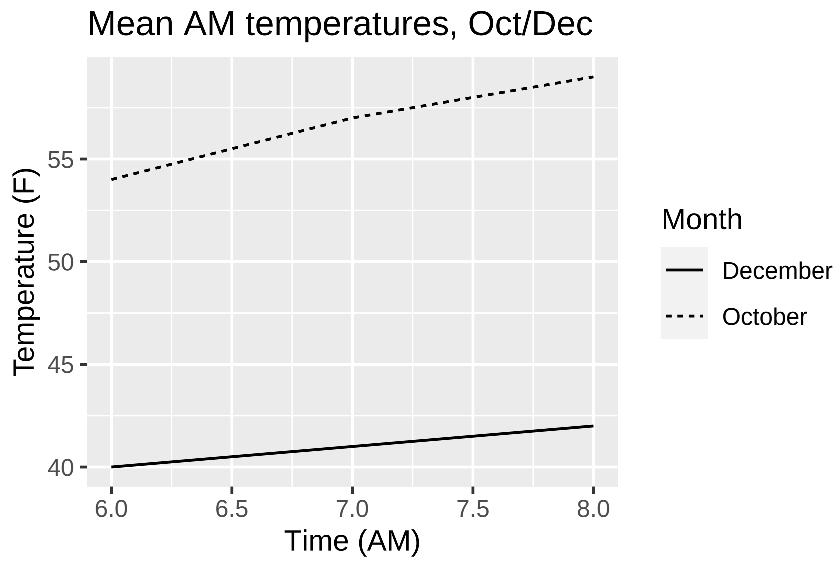


7. (5 points) Dr. Doctor printed out his final exam on three different colors of paper, recorded the scores and colors for each student, and produced the following visualization. Based on the visualization, he claims that there is a negative association between paper color and exam score. Is this an appropriate conclusion to draw? If so, explain how this can be seen in the visualization. If not, explain why not.



8. (6 points) Consider the following dataset and visualization.

```
dat <- data.frame(time = c(6, 7, 8, 6, 7, 8),
                  temp = c(54, 57, 59, 40, 41, 42),
                  mth = c(rep("October", 3), rep("December", 3)))
```



The line of code used to generate this visualization begins as follows:

```
ggplot(dat, aes(x = ____, y = ____, linetype = ____)) + ...
```

Fill in the blanks below.

Answer: `x =` ________________ , `y =` ________________ , `linetype =` ________________

9. (4 points) Which of the following is most appropriate for visualizing the relationship between a categorical variable and a numerical variable?

   a. Histogram
   b. Bar plot
   c. Pie chart
   d. Scatter plot

Answer: ____________________________

10. This question concerns a simulation that involves exponential random variables. Recall that the command `rexp(n)` generates `n` observations of exponentially distributed random variables with mean 1 and variance 1.

    (a) (5 points) The following function carries out a single simulation. Briefly summarize what this function does and what the return value represents.

    ```
    run_one_sim <- function(n) {
      X <- rexp(n)
      return(abs(mean(X) - 1) < 1.96 * sd(X) / sqrt(n))
    }
    ```

    (b) (4 points) Suppose we run the following code and obtain the following result. How does this value compare with what we would expect?

    ```
    set.seed(123)
    sim_results_20 <- replicate(10000, run_one_sim(20))
    print(mean(sim_results_20))
    ```

    ```
    ## [1] 0.9085
    ```

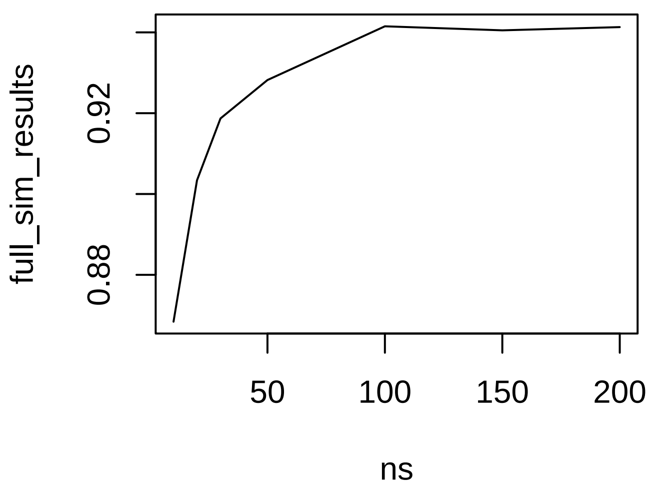
(c) (4 points) Suppose we run the following code and obtain the following result. How does this value compare with what we would expect? How does this result compare with our result for (b)?

```
set.seed(123)
sim_results_100 <- replicate(10000, run_one_sim(100))
print(mean(sim_results_100))
```

```
## [1] 0.9393
```

(d) (4 points) Suppose we run the following code and obtain the following visualization. What conclusions can we draw from these results?

```
ns <- c(10, 20, 30, 50, 100, 150, 200)
full_sim_results <-
  sapply(ns, function(x) mean(replicate(10000, run_one_sim(x))))
plot(ns, full_sim_results, type = "l")
```



11. The following functions all compute different error metrics that quantify how far a vector of predictions, `predicted`, deviate from the truth `observed`. Suppose we have n observations $x_1, \ldots, x_n$ and n predictions $\widehat{x}_1, \ldots, \widehat{x}_n$. For each of the mathematical formulas below, identify the matching function, if there is one. Write N/A if there is no matching function.

```
A <- function(observed, predicted) {
  return(mean(abs(observed - predicted)))
}
B <- function(observed, predicted) {
  return(sqrt(mean((observed - predicted) ^ 2)))
}
C <- function(observed, predicted) {
  return(mean(abs((observed - predicted) / observed)))
}
D <- function(observed, predicted) {
  return(sqrt(mean(observed - predicted)))
}
E <- function(observed, predicted) {
  return(mean(abs((observed - predicted) / predicted)))
}
```

(a) (4 points)

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \widehat{x}_i)^2}$$

Answer: ____________________

(b) (4 points)

$$\frac{1}{n}\sum_{i=1}^{n}|x_i - \widehat{x}_i|$$

Answer: ____________________

(c) (4 points)

$$\frac{1}{n}\sum_{i=1}^{n}\left|\frac{x_i - \widehat{x}_i}{x_i}\right|$$

Answer: ____________________

(d) (4 points)

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \widehat{x}_i)$$

Answer: ____________________

12. The `ChickWeight` dataset contains data on the growth of 50 baby chicks. The following code can be used to get the starting weights of all 50 chicks.

```
data("ChickWeight")
starting_wts <- ChickWeight |> dplyr::filter(Time == 0)
starting_wts$weight
```

```
##  [1] 42 40 43 42 41 41 41 42 42 41 43 41 41 41 41 41 42 39 43 41 40 41 43 42 40
## [26] 42 39 39 39 42 42 41 39 41 41 39 41 41 42 41 42 42 42 42 41 40 41 39 40 41
```

(a) (4 points) Consider the following hypothesis test code and results. In words, explain: what are the null and alternative hypotheses for this test?

```
t.test(starting_wts$weight, mu = 40)
```

```
##
##  One Sample t-test
##
## data:  starting_wts$weight
## t = 6.6197, df = 49, p-value = 2.577e-08
## alternative hypothesis: true mean is not equal to 40
## 95 percent confidence interval:
##  40.73821 41.38179
## sample estimates:
## mean of x
##     41.06
```

(b) (4 points) What conclusions can we draw based on the results of the hypothesis test?

13. (4 points) Thanks for your hard work! In one to two sentences, describe your favorite topic or concept we covered this semester.