STATS 306 F19

Final Exam Practice Questions

1. True or false:

   - If `typeof(x)` is `"integer"`, then `is.finite(x)` equals `!is.infinite(x)`.
   - One way to get the number of rows of a data frame is by typing `length(df)`.
   - For most commands, typing `?<command>` will pull up R's built-in help page for that command.
   - Within a single column of a tibble, all of the entries have the same data type.
   - If `y` is a numerical vector and `x` is a categorical vector, the regression `lm(y ~ x)` estimates the mean of `y` within each category of `x`.
   - Each time you start R, you should use `install.packages()` to load all of the packages that you will use in your analysis.
   - When summarizing data, you should include a column showing the number of observations in each group used to compute the summary statistic(s).
   - It is impossible to model nonlinear relationships using a linear model.
   - R has different data types for scalars (single values) and vectors.
   - The difference between lists and atomic vectors is that lists can hold any type of data, whereas atomic vectors only hold a single type of data.
   - You shouldn't waste time writing code comments; your intent will be obvious to the next person who has to run your code.
   - Both `for()` loops and `map()` can be used to iterate over vectors and lists.
   - R is the greatest programming language ever invented.

2. I want to join weather data for the specific hour and departure airport corresponding to each flight in the `flights` table.

   a. About how many rows should the resulting table have?

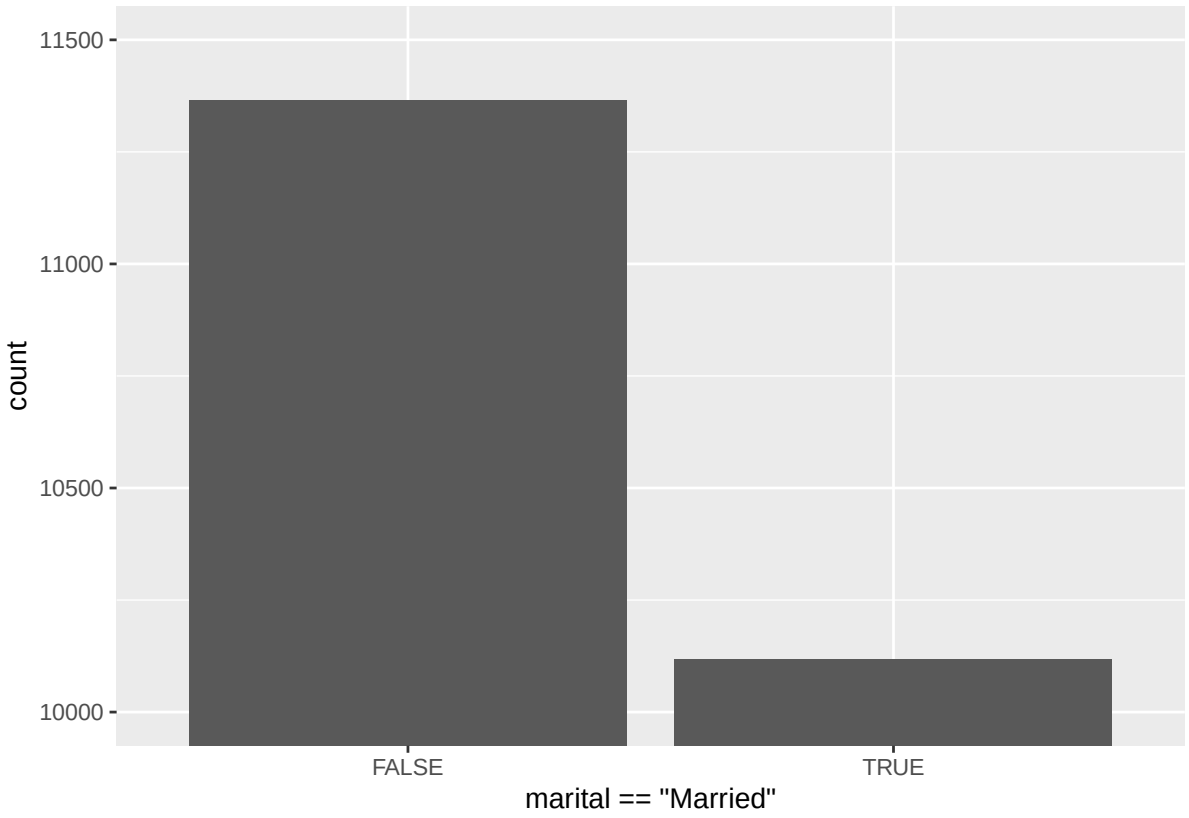
   b. Suppose I use the following command to perform this merge:

   ```
   left_join(flights, weather, by=c("origin", "hour"))
   ```

   Approximately how many rows will the resulting table have?

   c. Explain what went wrong, and provide the correct command needed to perform the merge.

3. The following plot was created from the `gss_cat` data set and compares the number of survey respondents who were married versus all other categories:



Explain why this figure is misleading, and supply code to produce a less deceiving plot.

4. The following command loads the complete text of Shakespeare's Hamlet into a vector called `hamlet`:

```
hamlet = readr::read_lines("https://git.io/vpeOg")
```

Each vector entry represents one line. Character names are in ALL CAPS: "HAMLET", "HORATIO", etc.

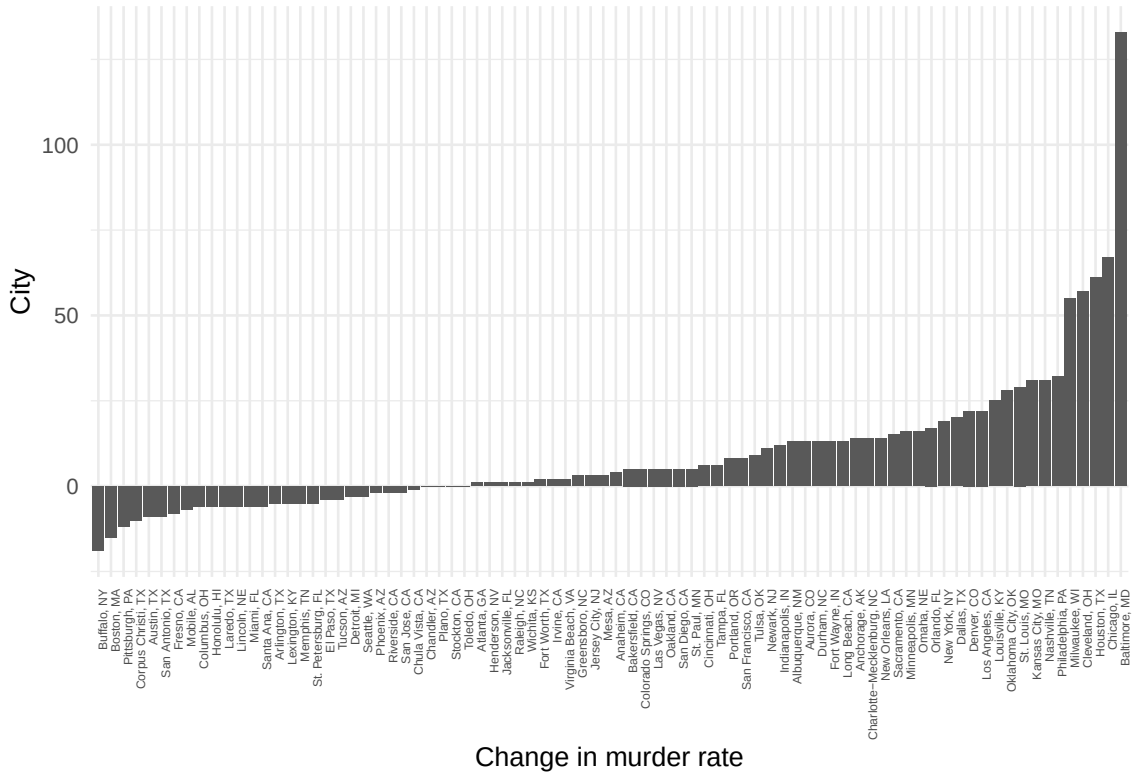
   a. What are the longest word(s) used in Hamlet? What if you include hyphenated words?
   b. What is the most common contraction (word containing an apostrophe)?
   c. Construct a table showing the most frequently used work in the play which is longer than n characters, for $n = 3, 4, \ldots, 10$. Exclude character names (HAMLET, HORATIO, etc.).
   d. A list of (almost) all the words in the English language is available at `https://git.io/JeDLe`. Ignoring contractions, how many words are there that a) occur only once in Hamlet and b) are not found in the dictionary. What are they? Which is your favorite? (Mine: offendendo).

5. The `fivethirtyeight` package contains data sets from http://fivethirtyeight.com. The data frame `fivethirtyeight::murder_2015_final` contains data on the murder rate in major US cities in 2014 and 2015.

   a. Load this data and convert it to tidy format.

   b. Convert each city and state pair to a label featuring the abbreviated state name. For example, the row for "Baltimore" and "Maryland" should become "Baltimore, MD". *Hint*: State names and abbreviations come pre-loaded into base R.

   c. Use the data to recreate the following plot:



6. Use the `optim()` function to find the unique positive root of the polynomial

$$x^4 - 16x^3 - 32x^2 - 64x - 144$$

(Hint: if r is a root of p, i.e. $p(r) = 0$, then r is a minimum of the function $p(x)^2$.)

7. A natural number greater than 1 is *prime* if it is only evenly divisible by itself and 1. Of the first thousand natural numbers, how many are prime?

8. Given a number n, suppose I do the following: start with $x = 1$; reverse the digits of x and add n to the resulting number. Repeat this process until x takes on a value that it has already taken before, and then stop. Let $r(n)$ be the number of steps needed until this process stops.

   If $n = 1$ then the sequence is

$$1 \to \underbrace{1 \leftarrow 1 + 1 = 2}_{\text{step 1}} \to \cdots \to \underbrace{10 \leftarrow 01 + 1 = 2}_{\text{step 9}}$$

   , so $r(1) = 9$. You can check that $r(2) = 81$.

   What is $r(88)$?

9. Consider the following data set:

```
n = 1000
df = tibble(x = runif(n, -1, 1), y = 4 * (x^2 - 1/2)^2 + runif(n, -1, 1) / 3)
```

   a. Compute the linear regression of y on x. Based on the regression results, is x a good predictor of y?

   b. Produce a scatter plot of x and y. Based on the plot, is x and good predictor of y?

   c. Find a better model for predicting y from x, fit it, and provide a numerical measure of how much better this model is at prediction than the model you fit in part *a)*.

10. Recall that `forcats::gss_cat` contains data from the General Social Survey.

    a. What is the median reported age for Jewish respondents?

    b. A millennial is defined to be someone who was 18 or younger in the year 2000. Are the millennials in this survey more likely to identify as atheist (`relig == 'None'`) compared to earlier generations?

    c. Consider the following two possible models relating age and hours spent watching TV:

    ```
    lm(age ~ tvhours)
    lm(age ~ poly(tvhours, 2))
    ```

    In your opinion, which model is a better fit to the data and why?

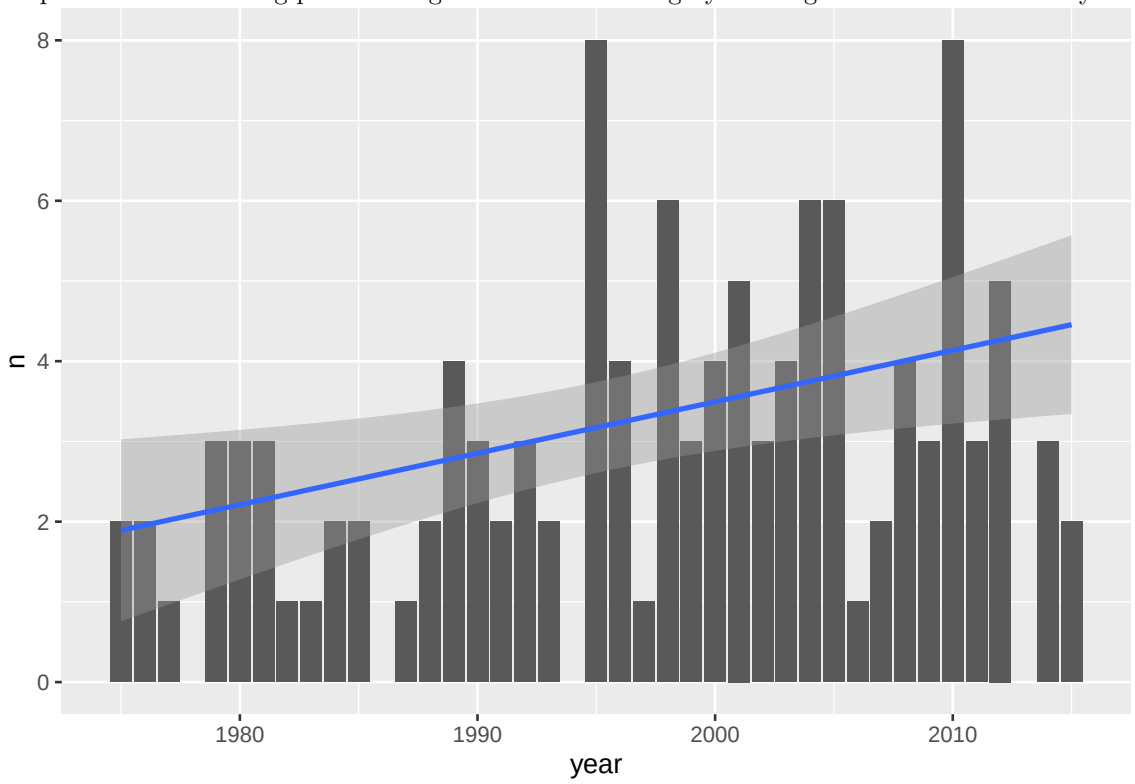
    d. Is there a statistically significant difference in the fraction of white respondents who identified as Christian, compared to non-white respondents? What is the p-value? (Define "Christian" to mean any of the following responses: "Christian", "Orthodox-christian", "Catholic", "Protestant".)

11. The following questions refer to the `dplyr::storms` data set. This data set contains tracking information on tropical storms and hurricanes in the United States over the past 40 years.

    a. Each storm is given a name which is unique for that year. Names can be re-used in later years. For example, there has been a storm named Ana in 1979, 1985, 1991, 1997, 2003, 2009 and 2015. One other storm name has been used seven times. What is it?

    b. Most of the observations in `storms` are tropical depressions or tropical storms. How many storms became category 2 or higher hurricanes at some point?

    c. Reproduce the following plot showing the number of category 2 or higher hurricanes in each year:



    d. Is the slope of the regression line in the preceding plot significantly different from zero? What does this imply?

    e. In 1985 a hurricane made landfall on Long Island, NY near JFK Airport. What was the name of that hurricane?