Final Exam

1. **Format**: Submit the exam to gradescope. For this exam, please make sure that your document has correctly compiled (check all of your figures). Use the `echo = T` option to make sure your code is visible (so that you can be awarded partial credit).
2. **Advice**: Be sure to adequately justify your answers and appropriately reference any sources used. Even if you are not able to answer a question completely, do your best to provide an answer and discuss solutions that you tried. Include your code and graphics for each question and create neat output by using options like `kable()` for tables and writing results in line with R commands.
3. **Resources and Citations:** While the exam is open book and you can use any resources from class or freely available on the internet, this is strictly an individual endeavor and **you should not discuss the problems with anyone outside the course instructor including class members.** All resources, including websites, should be acknowledged.
4. **Exam Questions:** If clarification on questions is required, please email the course instructor: andrew.hoegh@montana.edu. Clarifying questions will be addressed, but troubleshooting of R code will not be provided.
5. **A note on sharing / reusing code:** There is a huge volume of code is available on the web to solve any number of problems. For this exam you are allowed to make use of any online resources (e.g., StackOverflow, ChatGPT) but you must explicitly cite where you obtained any code you directly use (or use as inspiration). Any code that is discovered and is not explicitly cited will be treated as plagiarism.

Academic Honesty Statement

Include the following statement at the beginning of your submission.

> I, ___ (your full name here) ___, hereby state that I have not communicated with or gained information in any way from my classmates or anyone other than the course instructor during this exam, and that all work is my own or appropriately cited.

In the event that you have inadvertently violated the above statement, you should not sign above and instead discuss the situation with the course instructor.

Any plagiarism (including reusing code without appropriate citations) will result in a zero for the assessment.

Question 1

This question will use data from the USDA's Economic Research Service on educational attainment https://data.ers.usda.gov/reports.aspx?ID=17829

```
some_hs <- read_csv("https://raw.githubusercontent.com/stat408/Data/refs/heads/main/some_hs.
hs_only <- read_csv('https://raw.githubusercontent.com/stat408/Data/refs/heads/main/HS.csv')
some_college <- read_csv("https://raw.githubusercontent.com/stat408/Data/refs/heads/main/some
college <- read_csv("https://raw.githubusercontent.com/stat408/Data/refs/heads/main/College.
```

Question 1a (6 points)

Use the `college` dataset and create a better version of the choropleth obtained from the Economic Research Service. In particular, use a better color scale. Don't worry if your figure is limited to the lower 48 and excludes AK and HI.

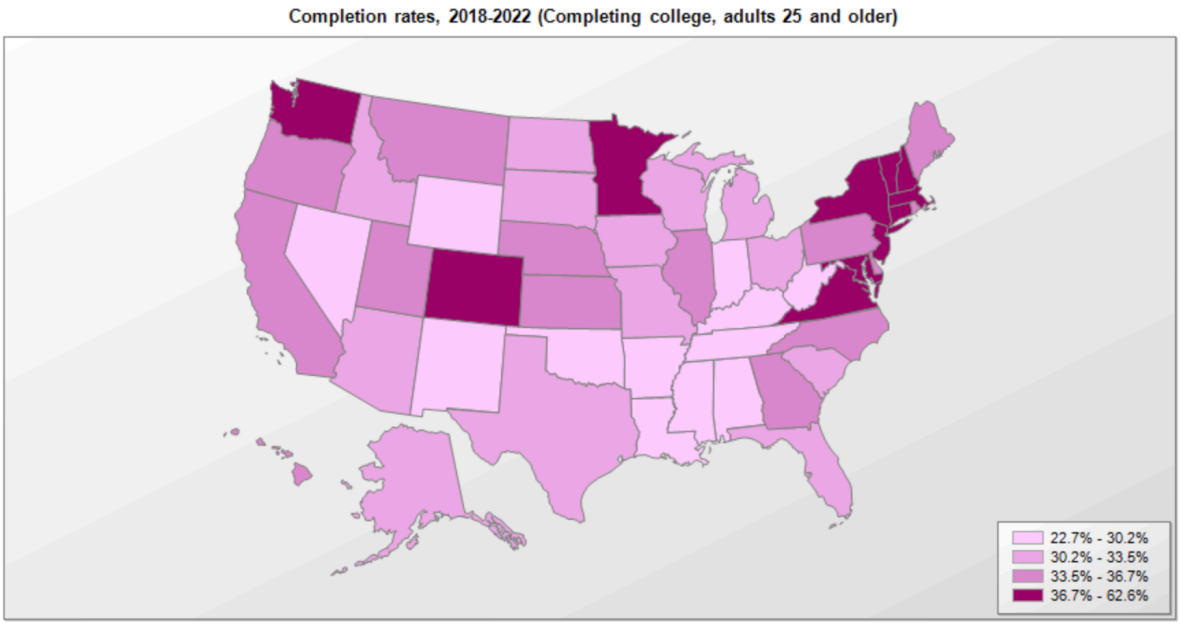

Figure 1: https://raw.githubusercontent.com/stat408/final20024/refs/heads/main/College_completion.png

Question 1b (6 points)

Consider the US Census Bureau's 9 divisions:

- Division 1: New England (Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, and Vermont)

- Division 2: Middle Atlantic (New Jersey, New York, Pennsylvania, and Maryland)
- Division 3: East North Central (Illinois, Indiana, Michigan, Ohio, and Wisconsin)
- Division 4: West North Central (Iowa, Kansas, Minnesota, Missouri, Nebraska, North Dakota, and South Dakota)
- Division 5: South Atlantic (Florida, Georgia, North Carolina, South Carolina, Virginia, Washington, D.C., Delaware, and West Virginia)
- Division 6: East South Central (Alabama, Kentucky, Mississippi, and Tennessee)
- Division 7: West South Central (Arkansas, Louisiana, Oklahoma, and Texas)
- Division 8: Mountain (Arizona, Colorado, Idaho, Montana, Nevada, New Mexico, Utah, and Wyoming)
- Division 9: Pacific (Alaska, California, Hawaii, Oregon, and Washington)

Create a table that shows the average in 1970 (within each division) of the four values:

- some HS
- HS completion
- some college
- college completion

Question 1c (6 point)

Pose a research question and create a graphic to address it. Include appropriate labels, legends, captions, and annotation.

Question 2

Consider the dice game left-center-right (LCR), which uses 6-sided dice where there are dots on 3 sides, a `R` on one side, a `L` on one side, and a `C` on the final side. For the initial turn, the beginning player rolls three dice. If a player rolls an `L` or `R`, they are required to pass one of the tokens (or dollar bills) to the player on the left or right, corresponding to `L` and `R` respectively. For a `C` the player places the token in the pot in the center. Dots are the most desirable outcome where players keep their token. All players start with three tokens (or dollar bills). If players have fewer than three tokens, they roll that many dice. Players take turns rolling until a single player has tokens remaining and they are the winner and take the center pot.

Question 2a (4 points)

Write a function called LCR with the following characteristics:

input - num_dice: the number of dice to roll

output: - the result of the *num_dice* dice as "dot", "L", "R", and/or "C"

Verify your function works by running the following unit tests

`LCR(3)`, `LCR(2)`, and `LCR(1)`.

Question 2b (2 points)

Now add appropriate error messages for the following scenarios

- `LCR(4)`: only three dice need to be rolled
- `LCR(0)`: sorry you don't have any more tokens
- `LCR('text')`: please indicate how many dice to roll

Question 2c (6 points)

Consider the situation where Champ has 3 tokens and Monte (the Griz mascot) has 1 tokens. Assume it is Champ's turn to roll. Simulate the game to the end and narrate each roll. You'll want to use `set.seed()` so the results do not change when you recompile the document. I've added Champ's first roll based on my function implementation - yours may look different.

```
set.seed(11292024)

## Starting Point
champ <- 3
monte <- 1
pot <- 0

## Champ's first roll
champ1 <- LCR(3)
pot <- pot + sum(champ1 == 'C')
monte <- monte + sum(champ1 %in% c('L','R'))
champ <- champ - sum(as.numeric(champ1 %in% c('dot', 'L', 'R')))
```

1. After Champ's first role which is `"C", "L", 'dot'` Champ has 1 token, Monte has 2 tokens, and there is 1 token in the pot.

Question 2d (4 EC points)

optional question worth extra credit points

Consider a simplified game with just one player. Assume the player starts with 3 tokens. In this case, the player will keep the tokens if an `L` or `R` are rolled and only lose tokens to the pot when rolling a `C`. Use a Monte Carlo technique to estimate how many rolls a player would go before running out of tokens. Create a figure to show the distribution of how many rolls the player would expect to go before running out of tokens.

Question 3

For this question, we will re-use the bread basket dataset from the midterm. Recall this dataset has transactions from a bakery in Edinburgh [https://www.kaggle.com/datasets/mittalvasu95/the-bread-basket](https://www.kaggle.com/datasets/mittalvasu95/the-bread-basket with the following characteristics:

- date: date of purchase
- time: time of purchase
- transaction: transaction number (there can be multiple items with each transaction)
- item: item type

```
bread <- read_csv('http://math.montana.edu/ahoegh/teaching/timeseries/data/BreadBasket.csv')
```

```
Rows: 21293 Columns: 4
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (1): Item
dbl  (1): Transaction
date (1): Date
time (1): Time

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Question 3a (3 points)

What transaction purchased the most items. How many items were in the transaction?

Question 3b (3 points)

What proportion of transactions include both `Coffee` and `Pastry`

Question 3c (3 points)

Create a table with the 5 most commonly purchased items in the afternoon (12 PM - 3:30 PM) and how many of those item types were purchased.

Question 3d (3 points)

Which hour/day had the most transactions? Note: For example, I'm not just looking for, say, 2PM but rather 2PM on a specific day.

Question 3e (6 points)

Recreate this plot.

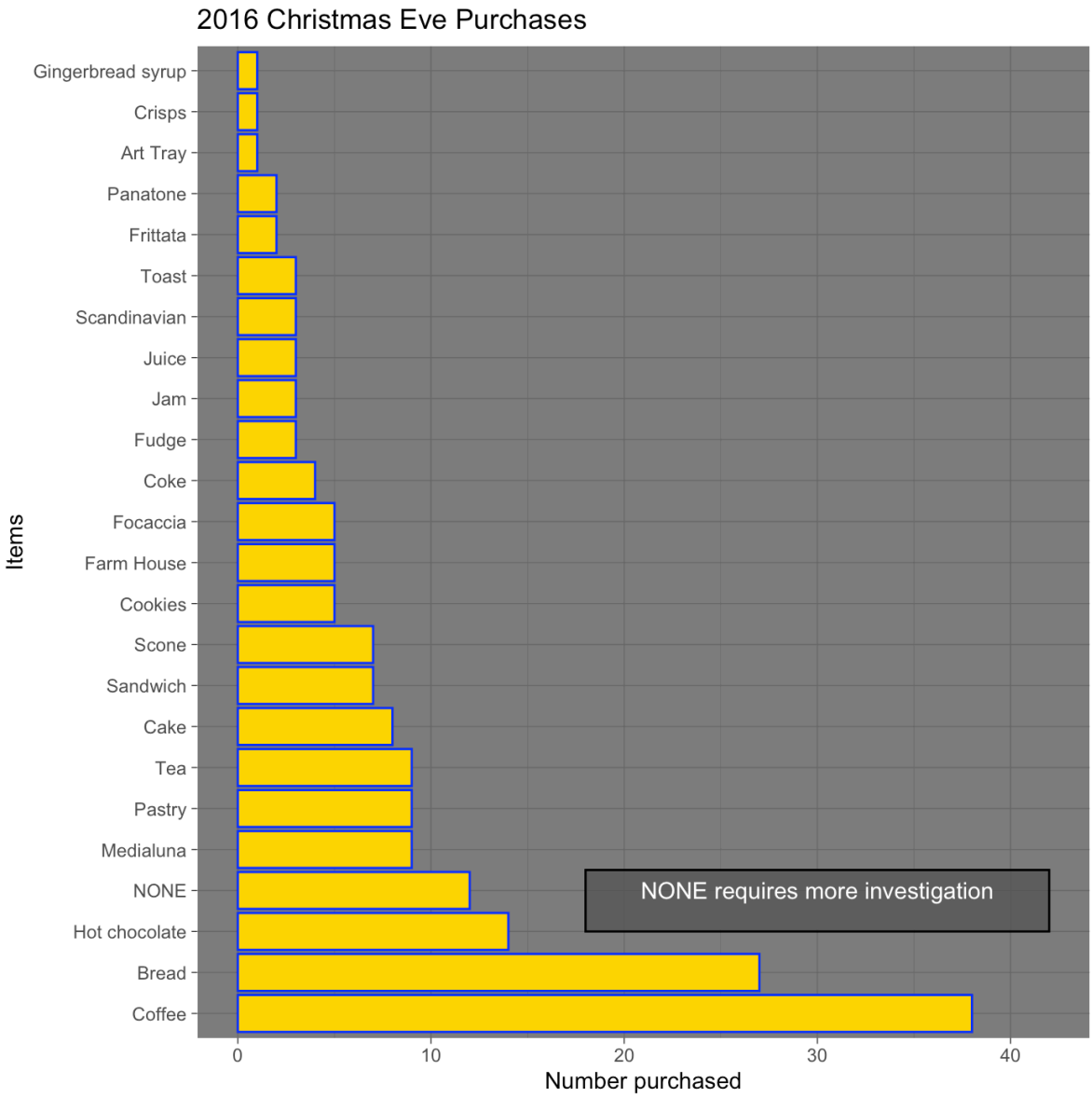

Figure 2: https://github.com/stat408/final20024/blob/main/ChristmasEve.png?raw=true