Midterm 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:** This 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.

Question 1

Recall the questions you answered on the first day of class about your self-reported enjoyment and/or ability in R. The reported values for the class can be seen in Table 1 below. Notice that either you or some of your classmates didn't report actual numbers.

Table 1: Fall 2024 STAT 408 pre-class self-reported R ability and enjoyment scores (on a scale of 1 - 5).

r_ability	r_enjoyment
3.5	3.5
3	4
2	2
4	5
3	4
3	4
3	3
2-3	3-4
3-4	2
2	2
2	2.5
3	5
3	4
3.3	4
3	4-5!
4	4
2-3ish	4
2	2
3-4	4
4	4
3	5
3	3
3	2

Calculate the average r ability of the course. For any value reported as `2-3`, `3-4` or similar use the midpoint between the two numbers.

Question 1a (5 points)

Add a column to the dataset to show the r ability as a numeric value. For full credit, you'll need to manipulate the dataset directly and create a new column that contains the numeric values. Print your table so it can be seen in the compiled document.

```
info_table <- info_table |>
  mutate(numeric_ability = case_when(
    r_ability == '2-3' ~ 2.5,
    r_ability == '3-4' ~ 3.5,
    r_ability == '2-3ish' ~ 2.5,
    TRUE ~ as.numeric(r_ability)))

info_table |> kable()
```

r_ability	r_enjoyment	numeric_ability
3.5	3.5	3.5
3	4	3.0
2	2	2.0
4	5	4.0
3	4	3.0
3	4	3.0
3	3	3.0
2-3	3-4	2.5
3-4	2	3.5
2	2	2.0
2	2.5	2.0
3	5	3.0
3	4	3.0
3.3	4	3.3
3	4-5!	3.0
4	4	4.0
2-3ish	4	2.5
2	2	2.0
3-4	4	3.5
4	4	4.0
3	5	3.0
3	3	3.0
3	2	3.0

Question 1b (5 points)

Calculate the average reported r abilty of this STAT 408 class. For full credit, report this value in a sentence using the inline r.

```r
info_table |> summarize(mean_r = mean(numeric_ability)) |> pull() |> round(2)
```

```
[1] 2.99
```

The average reported r ability of the class was 2.99

Question 2

For this question, we'll use a dataset on charging for electric vehicles, https://www.kaggle.com/datasets/valakhorasani/electric-vehicle-charging-patterns?select=ev_charging_patterns.csv

Note: don't be alarmed if the dataset seems a little off. I noticed a few problems, but thought it would still make for an interesting exam question.

```r
ev <- read_csv('https://raw.githubusercontent.com/stat408/Data/refs/heads/main/ev_charging_pa
```

Question 2a (5 points)

Note that the dataset contains a `Time of Day` variable with four outcomes: "Evening", "Morning", "Afternoon", and "Night." However, this variable seems incorrect as it doesn't correspond to the `Charging Start Time` in a coherent manner.

Create a new variable where the four times of the day correspond to `Charging Start Time` such that:

- Morning: 4AM - 11:59 AM or in military time (4 - 11:59)
- Afternoon: 12PM - 3:59PM or in military time (12 - 15:59)
- Evening: 4PM - 7:59PM or in military time (16 - 19:59)
- Night: 8PM - 12PM & 12PM - 3:59AM or in military time (20 - 24 and 0 - 3:59)

Print out the first 25 rows of the dataset that contains `Charging Start Time` and this new variable

```r
ev <- ev |>
  mutate(hour_start = hour(`Charging Start Time`)) |>
  mutate(time_of_day = case_when(
    hour_start >= 4 & hour_start < 12 ~ 'Morning',
    hour_start >= 12 & hour_start < 16 ~ 'Afternoon',
    hour_start >= 16 & hour_start < 20 ~ 'Evening',
    TRUE ~ 'Night'))
```

```
ev |>
  select(time_of_day, `Charging Start Time`) |>
  slice(1:25) |>
  kable()
```

time_of_day	Charging Start Time
Night	2024-01-01 00:00:00
Night	2024-01-01 01:00:00
Night	2024-01-01 02:00:00
Night	2024-01-01 03:00:00
Morning	2024-01-01 04:00:00
Morning	2024-01-01 05:00:00
Morning	2024-01-01 06:00:00
Morning	2024-01-01 07:00:00
Morning	2024-01-01 08:00:00
Morning	2024-01-01 09:00:00
Morning	2024-01-01 10:00:00
Morning	2024-01-01 11:00:00
Afternoon	2024-01-01 12:00:00
Afternoon	2024-01-01 13:00:00
Afternoon	2024-01-01 14:00:00
Afternoon	2024-01-01 15:00:00
Evening	2024-01-01 16:00:00
Evening	2024-01-01 17:00:00
Evening	2024-01-01 18:00:00
Evening	2024-01-01 19:00:00
Night	2024-01-01 20:00:00
Night	2024-01-01 21:00:00
Night	2024-01-01 22:00:00
Night	2024-01-01 23:00:00
Night	2024-01-02 00:00:00

Question 2b (5 points)

Create a figure to investigate whether charging times are longer at different times of day. Use your new variable from Question 2A for this question. Make sure your caption addresses whether you see any differences.

```r
ev |>
  ggplot(aes(y = `Charging Duration (hours)`, x = `Time of Day`, color = `Time of Day`)) +
  geom_boxplot(outliers = F) +
  geom_jitter() +
  theme_bw() +
  theme(legend.position = 'none')
```

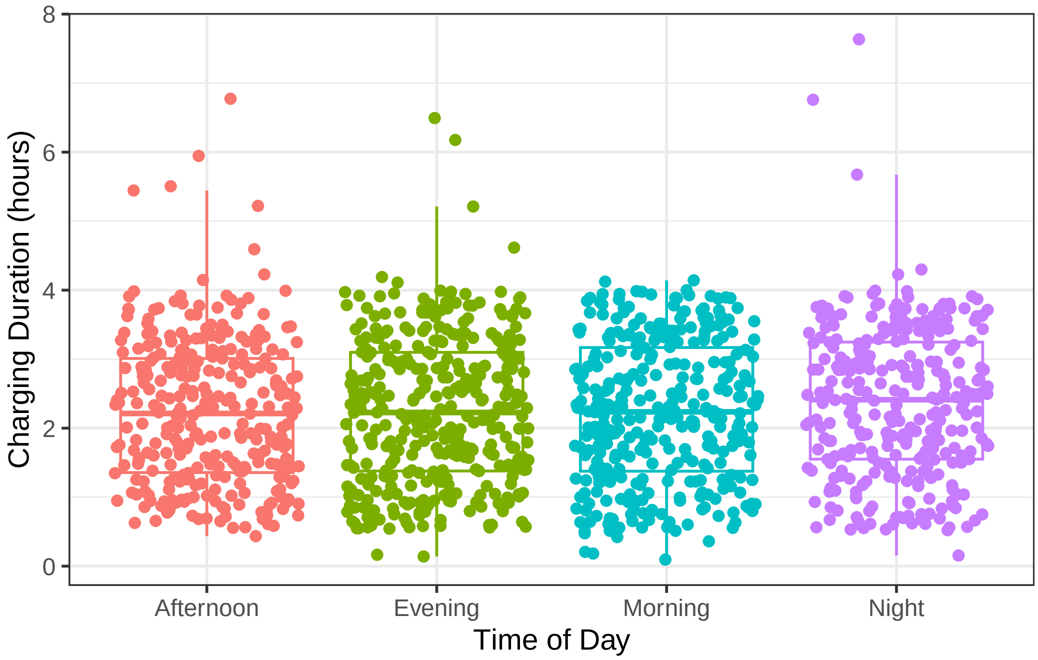

Figure 1: There don't appear to be major differences in charge time across time of day.

Question 3

For this question, we will use a cheese dataset https://www.kaggle.com/datasets/noahjanes/canadian-cheese-directory

```r
cheese <- read_csv('https://raw.githubusercontent.com/stat408/Data/refs/heads/main/cheese_da
```

Question 3a (5 points)

Print out a table the average moisture content by MilkTypeEn. Write a written summary describing the results in this table. Hint: You can, and should, filter out NAs.

```
cheese |>
  filter(!is.na(MilkTypeEn )) |>
  group_by(MilkTypeEn) |>
  summarize(mean_moisture = mean(MoisturePercent, na.rm = T)) |>
  arrange(desc(mean_moisture)) |>
  kable()
```

MilkTypeEn	mean_moisture
Buffalo Cow	72.00000
Goat	51.07290
Cow, Goat and Ewe	50.00000
Cow and Goat	49.61538
Ewe	49.25424
Ewe and Goat	48.00000
Cow	45.64631
Ewe and Cow	40.50000

The moisture in cheese made from Buffalo Cow's is substantially higher than other types.

Question 3b (10 points)

Create a single figure that explores `MoisturePercent` by `RindType`, `MilkTreatmentTypeEN`, and `FatLevel`. Hint: Again, you can, and should, filter out NAs. Make sure your caption describes the results.

```
library(viridis)
```

```
Loading required package: viridisLite
```

```
cheese |>
  filter(!is.na(MilkTreatmentTypeEn),
         !is.na(RindTypeEn),
         !is.na(FatLevel),
         !is.na(MoisturePercent)) |>
  ggplot(aes(y = MoisturePercent, x = FatLevel, color = FatLevel)) +
  geom_jitter() +
  facet_grid(RindTypeEn~MilkTreatmentTypeEn) +
  theme_bw() +
  scale_color_viridis( discrete =T) +
  geom_boxplot() +
  theme(legend.position = 'none')
```

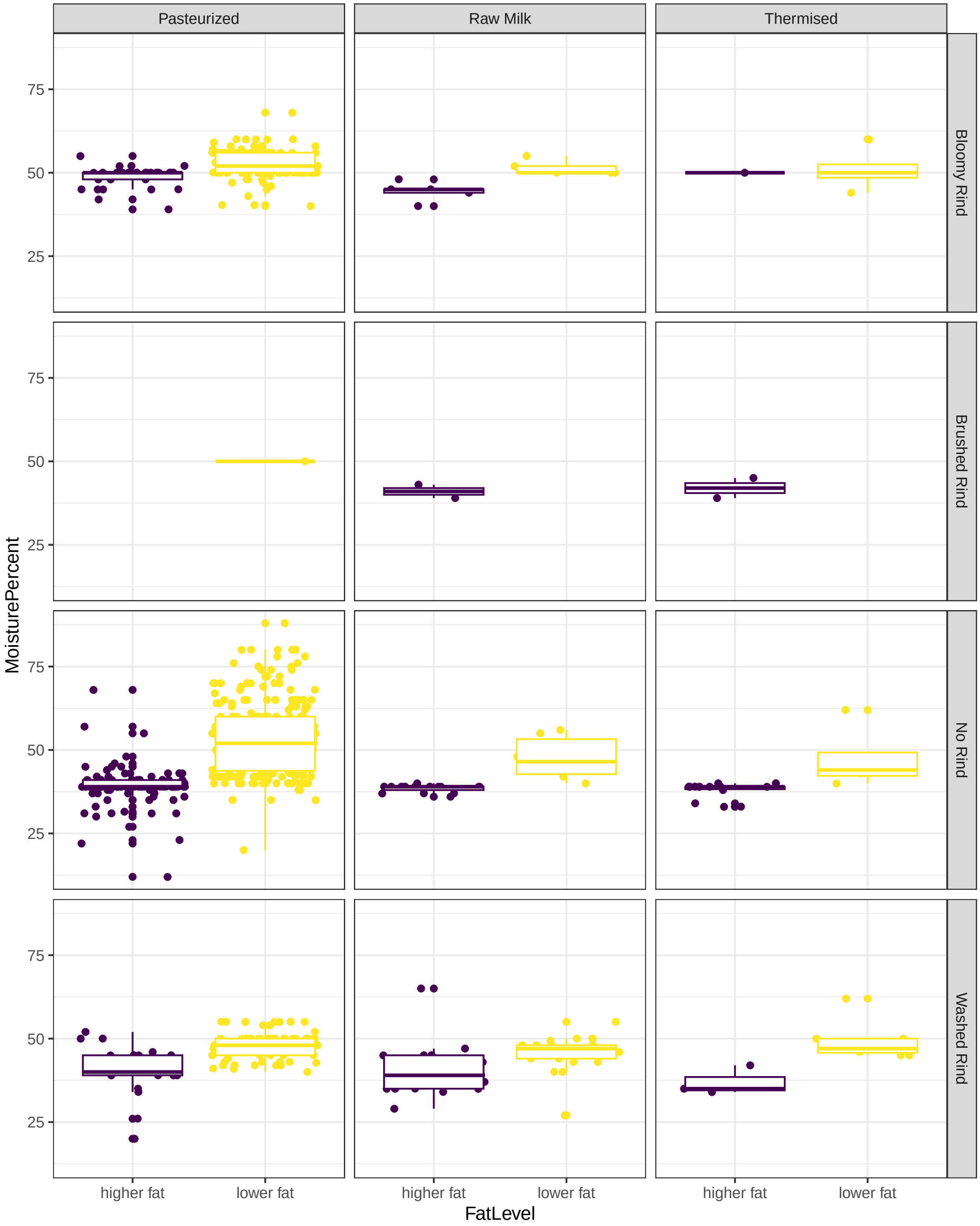

Figure 2: The lower fat cheeses tend to have more moisture, regardless of the rind type.

Question 3c (5 points)

How many of the 1042 cheeses are classified as a type of cheddar in the `CheeseName`?

```r
cheddars <- cheese |>
  filter(str_detect(CheeseName, "Cheddar")) |>
  tally() |>
  pull()
```

There are 100 different kinds of Cheddar in the dataset.

Question 4

This question uses a bread basket dataset with transactions from a bakery in Edinburgh https://www.kaggle.com/datasets/mittalvasu95/the-bread-basket.

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

```r
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 4a (2 points)

Pose a question. Which hour sees the most coffee purchases?

Question 4b (10 points)

```r
library(lubridate)
bread |>
  mutate(hour = hour(Time)) |>
  filter(Item == 'Coffee') |>
  ggplot(aes(x = hour)) +
  geom_bar(fill = 'chocolate4') +
  theme_bw() +
  ylab('Total Cups of Coffee Sold') +
  xlab('Hour of the Day') +
  labs(title = "Cups of Coffee Sold by Hour of Day at unnamed Edinburgh Coffee Shop",
       caption = paste("From ", min(bread$Date), ' to ', max(bread$Date), sep = '')) +
  annotate('text', x = 16.5, y = 950, label = "The most cups of coffee are sold during 11 c'
  annotate('segment', x= 16.5, xend = 11, y = 850, yend = 925,
           arrow = arrow())
```

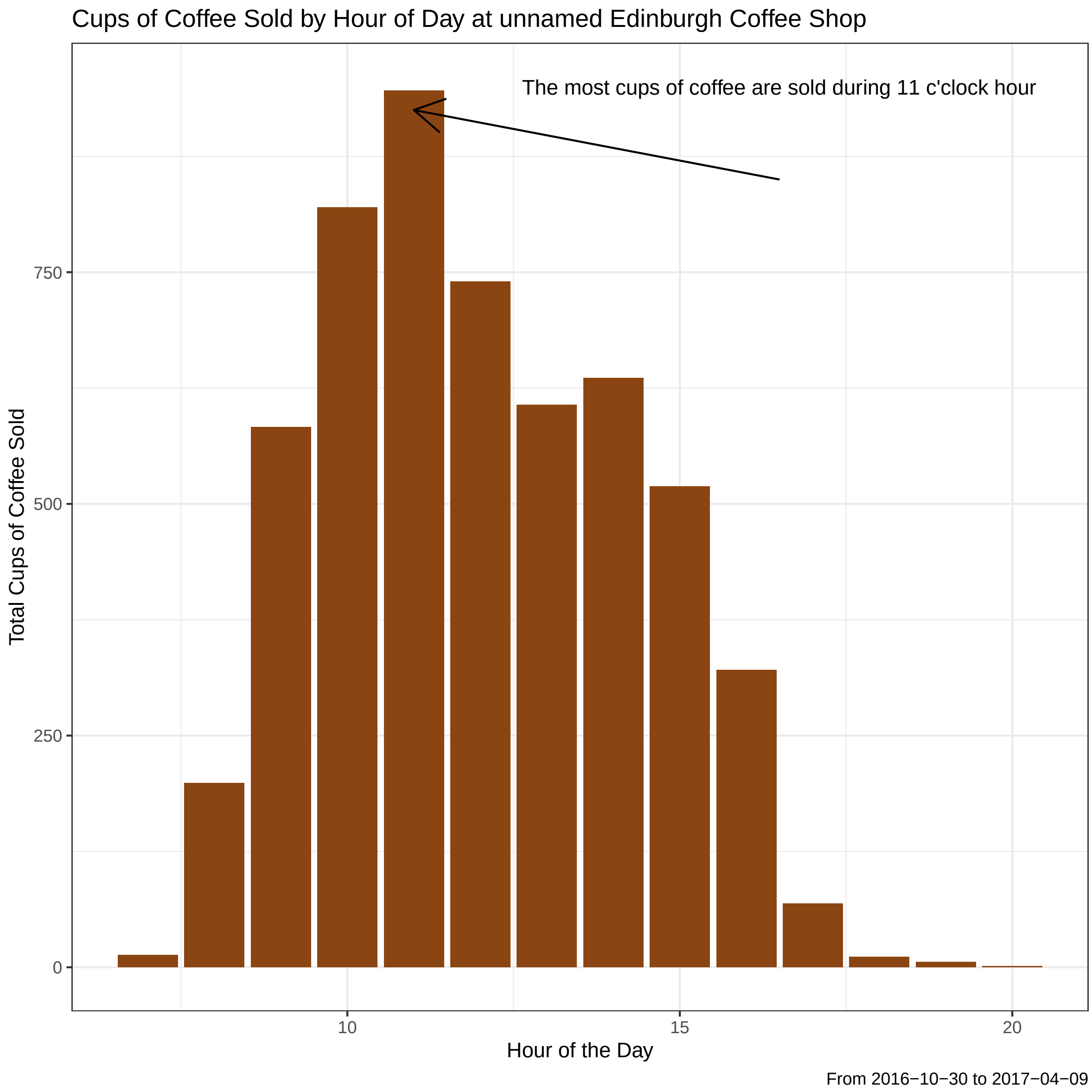

Figure 3: At this coffee shop, the eleven o'clock hour sees the most coffee sales.

Answer your question with a data visualization. For full credit, include informative labels, a caption, and at least one annotation.