STAT 408: Midterm Exam
Name:

Please turn in the exam to D2L and include the R Markdown code, a Word or PDF file with output. You are welcome to turn in code and output for each question separately. Please verify that all of the code has compiled and the graphics look like you think they should on your Word or PDF file, you are welcome to upload image files directly if they look distorted in the Word or PDF file.

While the exam is open book, meaning you are free to use any resources from class, this is strictly an individual endeavor and **you should not discuss the problems with anyone outside the course instructor including group mates or class members.** The instructor will answer questions related to the data, expectations, and understanding of the exam, but will not fix broken code.

The first two questions are designed to walk you through the start of the data analysis cycle using a dataset with customer transactions at a bakery.

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

```
## # A tibble: 21,293 x 4
##    Date       Time     Transaction Item
##    <date>     <time>         <dbl> <chr>
##  1 2016-10-30 09:58:11           1 Bread
##  2 2016-10-30 10:05:34           2 Scandinavian
##  3 2016-10-30 10:05:34           2 Scandinavian
##  4 2016-10-30 10:07:57           3 Hot chocolate
##  5 2016-10-30 10:07:57           3 Jam
##  6 2016-10-30 10:07:57           3 Cookies
##  7 2016-10-30 10:08:41           4 Muffin
##  8 2016-10-30 10:13:03           5 Coffee
##  9 2016-10-30 10:13:03           5 Pastry
## 10 2016-10-30 10:13:03           5 Bread
## # ... with 21,283 more rows
```

The dataset belongs to "The Bread Basket" a bakery located in Edinburgh. The dataset has 21293 entires, over 6000 transactions and 4 columns:

- Date: Categorical variable that tells us the date of the transactions (YYYY-MM-DD format).
- Time: Categorical variable that tells us the time of the transactions (HH:MM:SS format).
- Transaction: Quantitative variable that allows us to differentiate the transactions. The rows that share the same value in this field belong to the same transaction, that's why the data set has less transactions than observations.
- Item: Categorical variable with the products.

1. (18 points - Data manipulation)

This question will focus on data manipulation using the Bread Basket dataset, which can be accessed using the link above. For full credit, include your code and create neat output by using options like `kable()` for tables and writing results in line with r commands.

a. (2 points)

Print a table that contains the 5 days with the most items purchased and the number of items purchased on each day.

b. (3 points)

Print a table that contains the 5 days with the most transactions and the number of transactions on each day.

c. (3 points)

Print a table that contains the number of items purchased during each of the 24 hours of the day. Note the table should have 24 rows.

d. (2 points)

What proportion of the items in the dataset are Coffee.

e. (3 points)

What proportion of the transactions in the dataset include Coffee.

f. (2 points)

Print the complete transactions that include Crepes.

g. (3 points)

What are the 3 most common items purchased in the same transaction as a Sandwich?

2. (14 points - Data Visualization)

Continuing with the bakery dataset, we are know going to focus on data visualization.

a. (10 points)

Create a set of **two** graphics to illustrate components of the dataset. These graphs should be compelling and stand-alone with complete titles, labels, and axes. At least one of these figures should have annotation and at least one of the figures should be made with `ggplot2`.

b. (4 points)

Write a set of captions for each figure. The captions should be 2 - 3 sentences and fully describe the figures.

3. (15 points - Shake-a-day)

Shake a day is a common bar game that involes rolling dice. Here is a link to official rules in the State of Montana

For this problem we are going to write functions to automate the shake a day procedure. Assume the following rules:

- The pot starts at $100
- It costs $.50 to play.
- You win the pot when all 5 dice match (or in other words a Yahtzee)

a. Shake function (6 points)

Write a function called `Shake` that takes the following inputs: - pot: numeric value that has the starting value, which must be greater than $100

The function should return the following: - a list that contains two elements - winner: a logical value for whether the pot is won - pot: an updated pot value.

Demonstrate the function works using the following test cases:

```
Shake(99)
Shake('$101')
Shake(100.50)
```

b. (6 points)

Simulate 1000 games of Shake-a-Day and create a figure to show the total winnings for each of the 1000 games.

c. (3 points)

Assume that the house takes 10% of the pot (in other words if the pot is 1000, you would win 900 and the house would win 100). Compute your expected return for pots of the following size: $100, $500, $1000, and $5000.

Note the expected return in this case, with a 50 cent wager, can be calculated as:

$$0.5 - P_{win} * .9 * pot$$

where P_{win} is the probability of winning (rolling 5 dice of the same number) and *pot* is the current size of the pot.