STAT 408: Final Exam
Due: May 3 at 8:00 AM
Name:

Note: the exam can be turned in as late as May 3 at 8:00 AM for full credit.

Please turn in the exam to D2L and include the R Markdown code, SAS code *and either* a Word or PDF file with output. You are welcome to turn in code and output for each question separately. You can also embed SAS code in the markdown document itself. 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 or troubleshoot broken code.

The exam is designed to walk you through the complete data analysis cycle using a subset of data available from the Yelp Academic Dataset: https://www.kaggle.com/yelp-dataset/yelp-dataset or https://www.yelp.com/dataset. Specifically we will use the data files corresponding to a subset of the businesses located in Las Vegas, Nevada.

There are three datasets that will be used for this exam. The datasets are also available in the class folder via SAS.

1. http://www.math.montana.edu/ahoegh/teaching/stat408/datasets/yelp_lasvegas_business.csv. This dataset contains business specific information about 500 randomly sampled businesses in Las Vegas, Nevada. Many of these variables are self explanatory, but this dataset contains:

    - *business_id:* a business specific id.
    - *name:* business name
    - *neighborhood:* neighborhood location in Las Vegas
    - *address:*
    - *state:*
    - *city:*
    - *postal_code:*
    - *latitude:*
    - *longitude:*
    - *stars:* average user rating, rounded to nearest 0.5
    - *review_count:* number of reviews
    - *categories:* string of multiple categories for the business

2. http://www.math.montana.edu/ahoegh/teaching/stat408/datasets/yelp_lasvegas_business_hours.csv. This dataset contains business specific hours for each of the seven days in a week. Look closely at the format, as the business hour listed as `8:0-17:0` corresponds to the business being open from 8AM to 5 PM.

3. http://www.math.montana.edu/ahoegh/teaching/stat408/datasets/yelp_lasvegas_reviews.csv. In this dataset, each row represents a single review of a business. Note this dataset is fairly large and might take a little time to download.

1. (28 points - Data Wrangling)

This question will focus on data manipulation using the three Yelp datasets, which can be accessed using the links above and have been made available in SAS. You are welcome to use either R or SAS for this question, but include your code for full credit.

a. (4 points)

Your first stop in Las Vegas is 'The Strip'. How many businesses in this dataset are located on "The Strip" and have star ratings equal to or greater than 4.0?

b. (4 points)

After taking a walk down 'The Strip', you are curious about which of the other neigborhoods you should visit while in Las Vegas. Use the yelp_lasvegas_business.csv dataset to compute the median star rating for each of the neighborhoods in this dataset.

c. (4 points)

You have arrived on Sunday and notice that a large share of businesses seem to be closed. Compute the proportion of businesses that are closed on Sunday.

d. (4 points)

After visiting a few 3-star establishments, you feel they are less than "average". Use the yelp_lasvegas_business.csv dataset to compute the average star review given to the businesses in Las Vegas.

e. (4 points)

Upon arriving and experiencing the desert heat of Las Vegas, you realize that you need a haircut. Identify the businesses that are in the category `Barbers` and print the name and address of these businesses. Note: the `stringr()` package can be helpful here.

f. (4 points)

After a successful night at the Casino, you are ready for a nice dinner. Identify the business with the most 5 star reviews (which happens to be a restaurant). Print the name of the restaurant and the address.

g. (4 points)

Finally at 5:59 AM on Monday morning you decide that you are once again hungry. Identify the businesses in the yelp dataset that are open as of 5:59 AM on Monday. You can exclude anything that opens at 6AM or later.

2. (18 points - Data Visualization)

Continuing with the Yelp dataset, we are know going to focus on data visualization. You are welcome to use either SAS or R for this question.

a. (12 points)

Create a set of **three** graphics to illustrate components of the dataset. These graphs should be compelling and stand-alone with complete titles, labels, and axes. You are welcome to use either SAS or R for this question.

b. (6 points)

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

3. (5 points - Clustering)

Describe how you would create clusters of businesses using these three datasets. Be specific and discuss what variables you would select and what method you would use.

4. (SAS/ SQL)

For this question you must use SAS for full credit.

a. (3 points)

Compute the average star rating across the neighborhoods in the yelp_lasvegas_business.csv

b. (6 points)

Create a subset of the data containing businesses in the "Downtown" and "The Strip" neighborhoods. Then use a t-test to compare the overall rankings across the neighborhoods. (You can do this directly from the yelp_lasvegas_business.csv file). *Make sure to explain the results of the procedure.*