# STAT 491: Final Exam
# Name:

For motivation, we will be thinking about a dataset containing reviews of businesses by Yelp customers. The goal will be to predict the average star rating using other variables in the dataset.

Many of these variables are self explanatory, but this dataset contains:

```
- *stars:* average user rating
- *review_count:* number of reviews
- *categories:* character string denoting type of business
- *business_age:* how long a business has been open in years
- *is_chain:* binary variable denoting whether the business is a chain
```

## 1. Generalized Linear Models (17 points)

### a. (3 points)

Define a generalized linear model.

### b. (5 points)

Assume you decide to fit a multiple regression model using the GLM framework. Write out the necessary model components. Note you do not need to consider a prior at this point.

### c. (3 points)

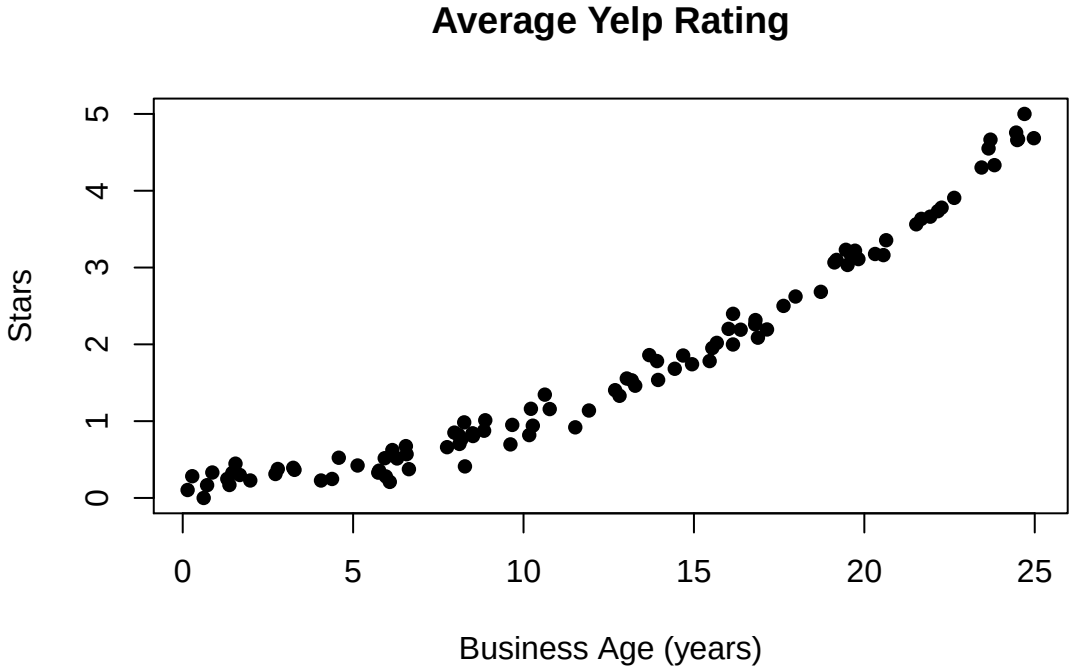
Clearly state the question that you are attempting to answer with the model in 1b. This should be free of statistical lingo and could be tailored to your entrepreneur sister who is considering opening a new business.

**d. (3 points)**

What are the implications of using a t-distribution vs. a normal distribution for a sampling model in a regression setting?

**e. (3 points)**

Assume you observe the following relationship between business age and average stars.




Describe how would you include business age in your model and include a rough approximation of what this would look like by adding a model fit line or curve to the figure.

## 2. Priors (9 points)

### a. (5 points)

Continuing the model specified in 1b, please state and defend prior distributions for all parameters in your model. Note you should include parameters in the distributions.

### b. (4 points)

Assume it is likely that there is a different relationship between business age and average star rating across the business categories. Briefly discuss two approaches to handle this issue.

## 3. Posterior (8 points)

### a. (4 points)

Assume you decide to start with a simple linear model using a normal sampling model and just business age as a predictor variable. Sketch out the model statement in JAGS for this statement, this should include the sampling model and the priors. You can use pseudocode and comments to explain your answer.

**b. (4 points)**

Assume you looked at a set of models, but decided on the simple model $Y_{stars} = \beta_0 + \beta_1 x_1$, where $x_1$ is the age of the business. If the 95 % HDI intervals are $\beta_0 = (1.1, 1.9)$ and $\beta_1 = (.20, .30)$, explain the results so that your entrepreneurial sister can understand.

## 4. Binomial Extension (6 points)

Suppose you have another dataset that contains the individual reviews for each user. We decide to use a binomial sampling model with the form: $Pr[Y_{stars} = k] = \binom{5}{k}\theta^k(1-\theta)^{1-k}$, where $Y_{stars}$ is the response and must be an integer value in {0,1,2,3,4,5}, $k$ is the number of stars, $\theta$ is the probability of getting a star, and $(1-\theta)$ is the probability of not getting star.

**a. (3 points)**

Identify and justify a conjugate prior for $\theta$. If you cannot do this, select and justify another prior distribution.

**b. (3 points)**

Assume the posterior predictive distribution and data are shown side-by-side below. What does this tell you about the sampling model / prior combination. Be specific.