# STAT 506: Midterm Exam
# Name:

Please turn in the exam to GitHub and include the R Markdown code, a Word or MD 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 files, 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 questions marked with an * denote in-class type questions, however, you are welcome to use any resources, but please include references and/or acknowledgments of the sources you have used.

For full credit, 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.

## 1. (ANOVA: 20 points)

This question will focus on ANOVA models using a dataset with arrival delays from airplanes leaving Bozeman. The dataset is filtered to include the following destinations: Denver (DEN), Minneapolis (MSP), Chicago (ORD), and Salt Lake City (SLC). Flights that are cancelled have been removed from the dataset.

```r
planes <-  read_csv('http://math.montana.edu/ahoegh/Data/planes.csv')
```

```
## Parsed with column specification:
## cols(
##   Year = col_double(),
##   Month = col_double(),
##   DayOfWeek = col_double(),
##   UniqueCarrier = col_character(),
##   ArrDelay = col_double(),
##   DepDelay = col_double(),
##   Dest = col_character(),
##   Cancelled = col_double()
## )
```

```r
planes <- planes %>% filter(Dest %in% c('DEN','MSP','ORD', 'SLC')) %>%
    filter(Cancelled == 0) %>% mutate(Day_factor = factor(DayOfWeek))
```

### a.* (6 points)

For each of the models included below, clearly define each parameter in the model and provide an evidence and size statement for each parameter.

```
lm_dest <- lm(ArrDelay ~ Dest - 1, data = planes)
display(lm_dest)
```

```
## lm(formula = ArrDelay ~ Dest - 1, data = planes)
##         coef.est coef.se
## DestDEN  7.74     1.02
## DestMSP  0.01     1.24
## DestORD 34.46     2.02
## DestSLC -0.86     0.81
## ---
## n = 4493, k = 4
## residual sd = 36.56, R-Squared = 0.07
```

```
lm_day <- lm(ArrDelay ~ DayOfWeek, data = planes)
display(lm_day)
```

```
## lm(formula = ArrDelay ~ DayOfWeek, data = planes)
##             coef.est coef.se
## (Intercept) 3.86     1.26
## DayOfWeek   0.12     0.28
## ---
## n = 4493, k = 2
## residual sd = 37.69, R-Squared = 0.00
```

```
lm_day2 <- lm(ArrDelay ~ Day_factor, data = planes)
display(lm_day2)
```

```
## lm(formula = ArrDelay ~ Day_factor, data = planes)
##             coef.est coef.se
## (Intercept)  5.41     1.48
## Day_factor2  2.73     2.12
## Day_factor3 -5.88     2.10
## Day_factor4 -4.21     2.10
## Day_factor5 -1.53     2.08
## Day_factor6  0.29     2.14
## Day_factor7  1.08     2.06
## ---
## n = 4493, k = 7
## residual sd = 37.61, R-Squared = 0.01
```

**b. (9 points)**

Explore whether the Bozeman snowy season and destination impact the arrival delay. For season, consider the snowy season to be the following months ( 11, 12, 1, 2, 3) and non-snowymonths. Justify whether an interaction is appropriate. Define all model parameters and provide evidence and size statements. (you don't need to check model assumptions for this model)

**c. (5 points)**

For the model that you fit in the previous question, state the model assumptions. Then provide a sentence or two for each assumption detailing whether you are satisfied that the assumption is reasonable in this case.

## 2. (Predictive Modeling: 20 points)

This question will focus on predictive models using a dataset with yelp rankings for businesses in Las Vegas.

```r
library(stringr)
YelpReviews <- read_csv(
    'http://math.montana.edu/ahoegh/teaching/stat408/datasets/yelp_lasvegas_business.csv') %>%
    dplyr::select(name, neighborhood, postal_code, stars, review_count, categories) %>%
    mutate(restaurant = str_detect(categories, 'Restaurants'),
           shopping = str_detect(categories, 'Shopping'),
           health = str_detect(categories, "Health & Medical"),
           local_services = str_detect(categories, "Local Services"),
           automotive = str_detect(categories, "Automotive"),
           home_services = str_detect(categories, "Home Services"))
YelpReviews
```

```
## # A tibble: 500 x 12
##    name  neighborhood postal_code stars review_count categories restaurant
##    <chr> <chr>              <dbl> <dbl>        <dbl> <chr>      <lgl>
##  1 "\"L... Spring Vall...       89113   4.5           56 Hair Remo... FALSE
##  2 "\"B... Westside           89117   5              6 Restauran... TRUE
##  3 "\"D... Downtown           89101   4.5            6 Bail Bond... FALSE
##  4 "\"A... Centennial         89149   4             14 Laundry S... FALSE
##  5 "\"C... Downtown           89101   5              3 Professio... FALSE
##  6 "\"G... Sunrise            89115   4              4 Video Gam... FALSE
##  7 "\"T... The Strip          89109   4.5           10 Barbers;S... FALSE
##  8 "\"O... Downtown           89121   3.5            6 Auto Part... FALSE
##  9 "\"A... Westside           89102   3             15 Body Shop... FALSE
## 10 "\"N... Spring Vall...       89147   4.5           34 Local Ser... FALSE
## # ... with 490 more rows, and 5 more variables: shopping <lgl>, health <lgl>,
## #   local_services <lgl>, automotive <lgl>, home_services <lgl>
```

**a*. (4 points)**

Detail a test and training approach and describe why this is necessary in predictive modeling.

**b*. (4 points)**

Star rankings for each establishment are rounded to the nearest 1/2 star. Define a loss function and then propose a loss function for comparing predictive models with this dataset.

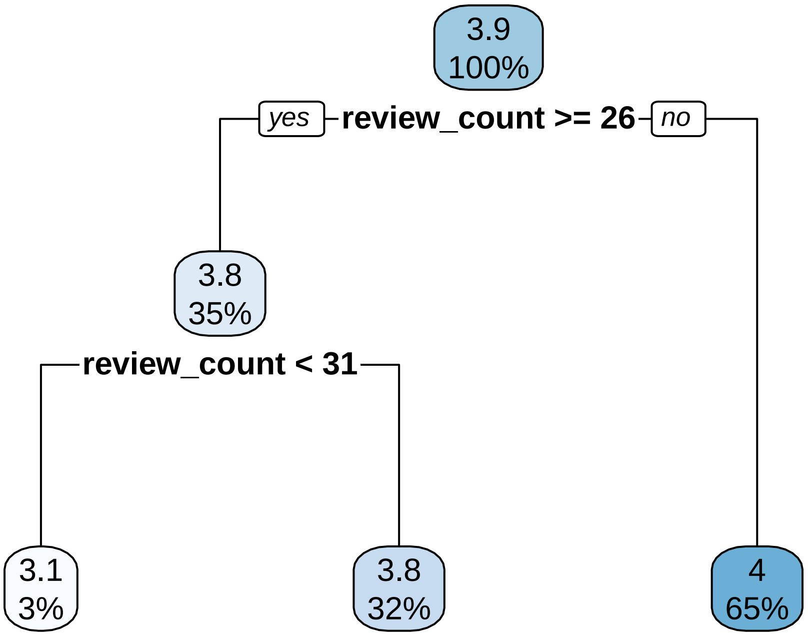
```
unique(YelpReviews$stars)
```

```
## [1] 4.5 5.0 4.0 3.5 3.0 2.5 2.0 1.5 1.0
```

**c*. (4 points)**

Interpret the following decision tree.

```
tree <- rpart(stars ~ restaurant + review_count, data = YelpReviews)
rpart.plot(tree)
```



**d. (2 points)**

Create a test and training test.

**e. (6 points)**

Using your loss function from part b, develop a predictive model that is better than the model in part c. Describe your model in a way that a Las Vegas tourist could understand.

## 3. (Regression: 24 points)

This question will focus on regression models using a dataset with rankings of wine varieties.

```
wine <- read_csv('http://www.math.montana.edu/ahoegh/teaching/stat532/data/wine_ratings.csv')
```

```
## Parsed with column specification:
## cols(
##   points = col_double(),
##   price = col_double(),
##   province = col_character(),
##   title = col_character(),
##   variety = col_character(),
##   winery = col_character()
## )
```

```r
common_wines <- wine %>% group_by(variety) %>% tally(sort = T) %>% slice(1:8)
wine <- wine %>% filter(variety %in% common_wines$variety)
```

### a. (8 points)

Create a series of data visualizations exploring the relationship between points and price, province, and variety. For each figure include an informative caption. For full credit, include informative titles, labels, and potentially annotations.

### b.* (2 points)

Based on the figures created in the previous question, do you have any reservations about fitting a regression model with normal errors?

### c. (8 points)

Regardless of the results for part b, fit a regression model for points as a function of price and variety. Include evidence and size statements for the model. Show your results graphically.

### d. (2 points)

Predict the average point rankings for a bottle of a $50 Pinot Noir. Show your results graphically.

### e. (2 points)

Predict the point rankings for a single bottle of a $50 Pinot Noir. Show your results graphically.

### f.* (2 points)

How could a hierarchical model be used for different relationship between points and priceas a function of variety or province?