STAT 505: Final Exam
Name:

Short Answer Questions (16 points)

For questions in this section, keep your answers concise. You are welcome to use a combination of prose, math, and pseudocode, but your responses should be well thought out and defended. If the class has 100% completion of course evaluations, you may choose any **4** questions to answer and cross out the question you don't want graded. If you answer all 5, the first 4 questions will be graded.

1. (4 points)

Describe your philosophy for choosing which parameters to include in a model. Highlight pros and cons of that method.

2. (4 points)

Describe statistical significance, in your own words. Then discuss whether or not you plan to use this term in your professional practice

3. (4 points)

Convince a collaborator, say a data scientist modeling vehicle prices, that using Bayesian analysis is a defensible approach.

4. (4 points)

Convince a collaborator, say a data scientist modeling vehicle prices, that using frequentist (non-Bayesian) analysis is a defensible approach.

5. (4 points)

Suppose that rather than grading final exams, your STAT505 instructor proposed to use a logistic regression model to predict the binary outcome of whether students have passed the course. Specifically, the instructor proposed to use following (fake) previous data to fit the following model.

```
## # A tibble: 10 x 3
##    pass     hw exam1
##    <chr> <dbl> <dbl>
##  1 Yes    97.1  96.2
##  2 Yes    80.8  89.7
##  3 Yes    89.0  96.2
##  4 No     95.2  94.2
##  5 Yes    74.5  87.1
##  6 Yes    92.5  96.8
##  7 No     89.3  86.0
##  8 No     75.0  89.8
##  9 Yes    75.3  90.6
## 10 Yes    74.7  90.6
```

```
glm(pass ~ hw + exam1, data = scores)
```

Beyond the obvious issues with not grading final exams, critique your instructor's use of `R` code to complete this process.

Code Interpretation (20 points)

For this question, we will reconsider the spotify dataset from the midterm to evaluate whether `acousticness` or `instrumentalness` best predict whether a song is defined as `Classical` music.

- Acoustic music generally refers to only using musical instruments without electric amplification. `acousticness` is defined as "Acousticness: A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic."
- `instrumentalness` is defined as "Instrumentalness: Predicts whether a track contains no vocals. "Ooh" and "aah" sounds are treated as instrumental in this context. The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0."
- `is_classical` is a binary variable where 1 indicates a song is from the classical music genre

1. (4 points)

Add a title and captions to the following figure.

```
fig1 <- spotify_v2 %>% ggplot(aes(x = is_classical, y = acousticness)) +
   geom_jitter(alpha = .01) + geom_violin() + theme_bw()

fig2 <- spotify_v2 %>% ggplot(aes(x = is_classical, y = instrumentalness)) +
  geom_jitter(alpha = .01) + geom_violin() + theme_bw()
```

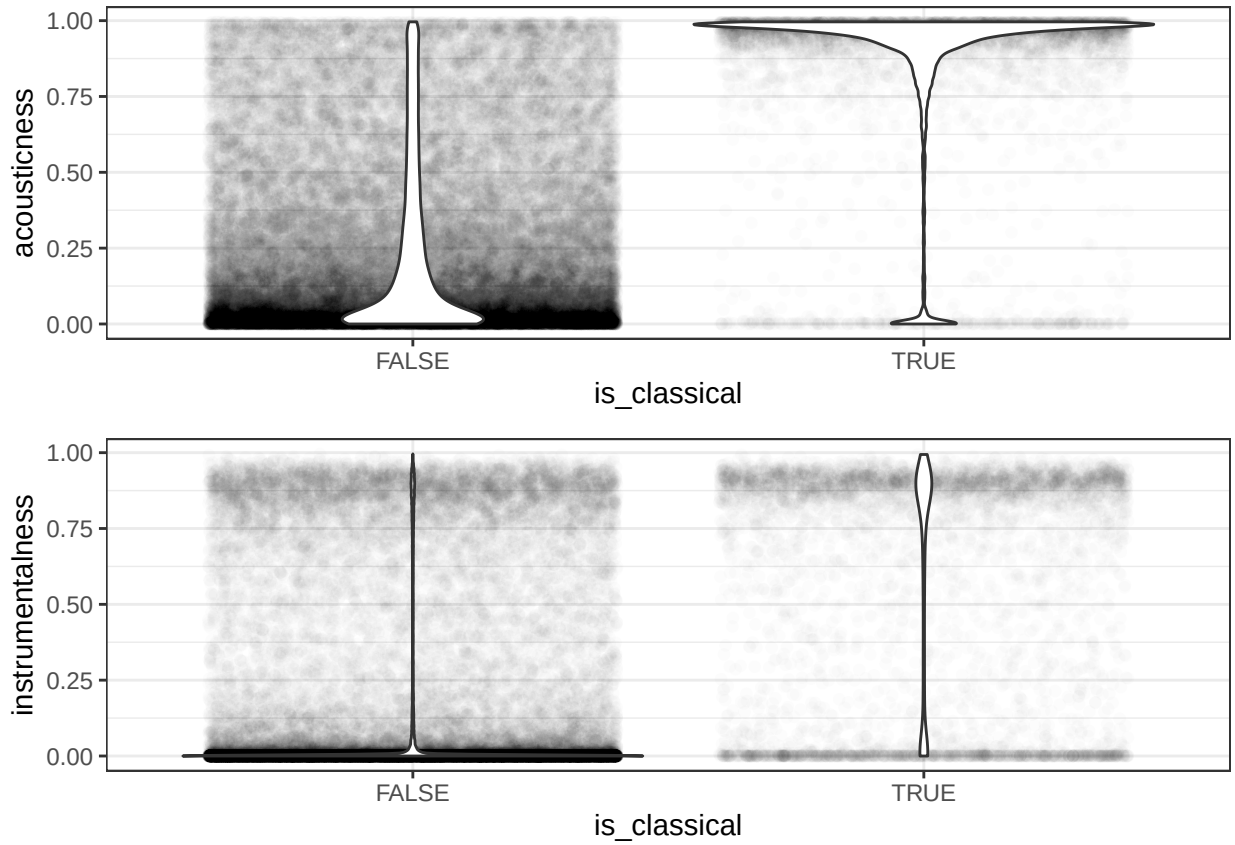

Figure 1: This figure

2. (4 points)

Write out the full statistical model implied with the following `R` code.

```
glm(is_classical ~ acousticness, data = spotify_v2,
    family = binomial(link = 'logit'))
```

3. (4 points)

Using the following model output, discuss whether higher levels of `acousticness` make it more or less likely that a song is classified as classical music

```
logistic_model <- glm(is_classical ~ acousticness, data = spotify_v2,
    family = binomial(link = 'logit'))
logistic_model %>% display()
```

```
## glm(formula = is_classical ~ acousticness, family = binomial(link = "logit"),
##     data = spotify_v2)
##              coef.est coef.se
## (Intercept)  -6.22     0.07
## acousticness  6.64     0.09
## ---
##   n = 50000, k = 2
##   residual deviance = 17903.9, null deviance = 32508.3 (difference = 14604.4)
```

4. (4 points)

Recall that the `plogis()` function performs the inverse logit or logistic function. Interpret the following model output and describe what is being calculated in each setting.

```
as.numeric(plogis(logistic_model$coefficients['(Intercept)']) +
             0 * logistic_model$coefficients['acousticness'])
```

```
## [1] 0.001987204
```

```
as.numeric(plogis(logistic_model$coefficients['(Intercept)'] +
            1 * logistic_model$coefficients['acousticness']))
```

```
## [1] 0.6039403
```

5. (4 points)

Interpret the following model output and explain the response in the context of the model selection framework.

```
AIC(glm(is_classical ~ acousticness, data = spotify_v2,
    family = binomial(link = 'logit')))
```

```
## [1] 17907.94
```

```
AIC(glm(is_classical ~ instrumentalness, data = spotify_v2,
    family = binomial(link = 'logit')))
```

```
## [1] 25351.75
```