STAT 505: Exam I
Name:

Turn in the exam to GitHub and include the R Markdown code and a PDF file with output. 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 output file.

While the exam is open book and you can use any resources from class or freely available on the internet, 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.** All resources, including websites, should be acknowledged.

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.

Short Answer Questions (16 points)

For questions in this section, keep your answers concise. You are welcome to use a combination of prose, math, visualization, and pseudocode, but your responses should be well thought out and defended.

1. (4 points)

In your own words describe what an interaction is and why they are important to consider when fitting linear models.

2. (4 points)

If $x1$ has three categories `low`, `medium`, and `high` and $x2$ is continuous, write out the model implied by `y ~ x1 + x2 - 1`. Provide an interpretation for all parameters in the model.

3. (4 points)

Why should inferences about sampling units be characterized as differences in predictors *between* units rather than differences in predictors *within* a sampling unit?

4. (4 points)

Consider making predictions for an expected outcome (say all candy bars with chocolate, peanut butter, and the 50th percentile for sugar) and for a new observation (such as a newly developed candy bars with chocolate, peanut butter, and the 50th percentile for sugar). Describe how the point estimate and uncertainty intervals would differ (or not) between the two situations.

Simulation Question (12 points)

1. (4 points)

Consider the code below. Create a figure of x and y. What linear regression assumption does this data violate?

```
set.seed(11112021)
n <- 100
x <- seq(1,100, length.out = n)
sigma <- 1
beta <- c(1, 1)
x_star <- log(x)
y <- rnorm(n, beta[1] + beta[2] * x_star, sd = sigma)
combined <- tibble(y = y, x = x)
```

2. (4 points)

How well does a linear regression model (`y ~ x`) recover the point estimates of beta? Justify your answer and note several replications of simulation data may be useful.

3. (4 points)

How well does a linear regression model capture the uncertainty in a predictions for y conditional on

- x = 1
- x = 50
- x = 100

Modeling Questions (32 points)

For this dataset we will use a dataset pulled from Spotify. In addition to `music_genre`, `artist_name`, and `track_name`, the dataset contains the following variables:

- Danceability: Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable.
- Tempo: The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration.

```
spotify <- read_csv('spotify.csv')
```

Using this dataset we will explore a special kind of linear model referred to as an analysis of variance (ANOVA).

1. (4 points)

We will start with a 1- way ANOVA, a linear model with a single categorical variable. In this case `music_genre` to predict danceability. Write out the statistical notation (using linear models framework show in this class, hint: use β). You are welcome to use scalar or matrix notation, but clearly define all parameters in your model.

2. (4 points)

Interpret the output from the ANOVA model. Focus on describing differences in danceability across `music_genre`. If you choose to use p-values, consider the implications we've discussed in class and highlighted in homework readings.

```
lm(danceability ~ music_genre, data = spotify) %>% summary()
```

```
##
## Call:
## lm(formula = danceability ~ music_genre, data = spotify)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.49081 -0.09107  0.00820  0.09584  0.44321
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           0.299792   0.006405   46.81   <2e-16 ***
## music_genreCountry    0.282635   0.009268   30.50   <2e-16 ***
## music_genreElectronic 0.314019   0.009386   33.46   <2e-16 ***
## music_genreHip-Hop    0.424370   0.009335   45.46   <2e-16 ***
## music_genreRap        0.385353   0.009250   41.66   <2e-16 ***
## music_genreRock       0.229379   0.009310   24.64   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1348 on 2428 degrees of freedom
## Multiple R-squared:  0.524,  Adjusted R-squared:  0.523
## F-statistic: 534.5 on 5 and 2428 DF,  p-value: < 2.2e-16
```

3. (4 points)

Explain the differences in the two intervals, focus both on the actual values and the intepretation. Use language that a typical spotify user could understand.

```
anova_stan <- stan_glm(danceability ~ music_genre, data = spotify, refresh = 0)

interval1 <- posterior_linpred(anova_stan, new_data = tibble(music_genre = "Classical")) %>%
  quantile( probs = c(.025, .975))
interval2 <- posterior_predict(anova_stan, new_data = tibble(music_genre = "Classical")) %>%
  quantile( probs = c(.025, .975))
```

The first interval is 0.29, 0.73 and the second interval is 0.15, 0.91

3. (20 points)

For this question we will add a continuous variable `tempo` to assess the `danceability` as a function of `music_genre` and `tempo`. This model is referred to an an Analysis of Covariance (ANCOVA).

The `tempo` variable has been centered for you if you choose to use it, where the mean tempo value is 120

```
spotify <- spotify %>%
  mutate(tempo_centered = tempo - mean(tempo))
```

a. (4 points) Create a data visualization to explore how `music_genre` and `tempo` (or `tempo_centered`) relate to `danceability`. This figure should include an informative caption, titles, and axis labels.

```
spotify %>% ggplot() + theme_bw()
```

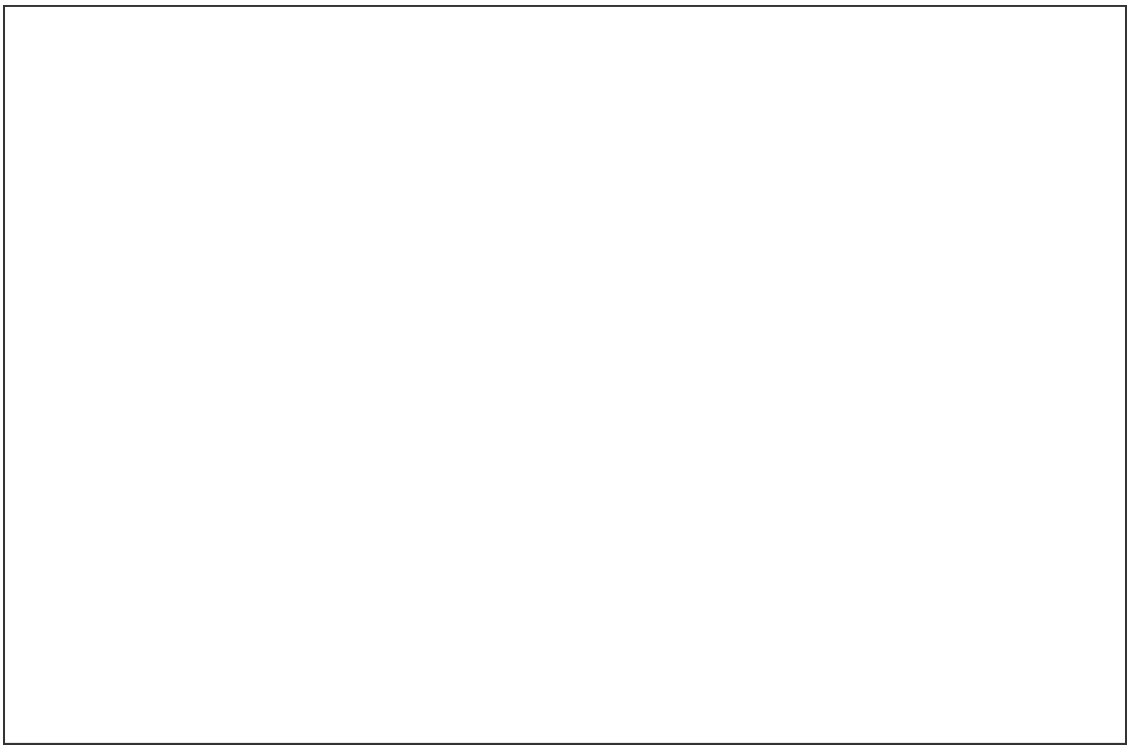

Figure 1: Caption goes here

b. (4 points) Write out and justify the model that you choose to fit that best captures the relationship for `danceability` as a function of `music_genre()` and `tempo`.

c. (4 points) Fit the model and interpret the model coefficients so that you can explain the results to your cousin who is an aspiring DJ. Data visualization may be helpful here.

d. (4 points) What assumptions are you making with fitting this model? Without a formal investigation, highlight one that are the most concerned about and describe why.

e. (4 points) Using a realistic value for `tempo`, make a recommendation for song characteristics (`tempo` and `music_genre`) that would have the highest `danceability`. Provide an interval estimate of the `danceability` of that song.