STAT 230 Short Report 1

Due October 18th at 10pm (no penalty late submission at 2am on 10/19)

For this short report, you will dig into the data that you started analyzing with the MUSC 232 (Golden Age of R&B) class on September 30 in class. You will need to do some background reading and cite at least two peer-reviewed sources to give context to your study (you are welcome to use additional articles/sources that are not peer-reviewed, but you must include two peer-reviewed sources).

Spotify Data

In this short report, you will continue your analysis based on the discussion of R&B music using data from Spotify, a popular music streaming service. You are encouraged to also incorporate the categorical crossover variable, in part defined by your MUSC 232 partners. This crossover variable is indicating whether or not the song was in the top 20 on both the Pop and R&B charts (whether it was able to be a hit in both categories, or "cross over" the two genres). I am not requiring a specific response variable or covariates- these choices will be based on your class conversations, your final research questions, and analysis steps (look for need for transformations/interactions, etc.).

You can read the codebook for the Spotify-defined variables here. Your unit of analysis will be individual songs. Throughout your analysis, you should keep in mind any criticism you might have about the definition of variables and you should have a healthy discussion of any necessary caveats at the conclusion of your report.

There are two Spotify datasets loaded below. One of them is the dataset with all of the songs labelled by the MUSC 232 students (412 songs). Due to time constraints, they were not able to label every song in the dataset. If you decide not to use the crossover variable in your final model (this decision must be well-justified), you may consider using the full dataset (consisting of 5,928 songs) which is also loaded below.

In your assignment, you **must** use the artists that your MUSC 232 partners were assigned (refer to your worksheet from class). You can choose additional artists if you'd like. You must also refer to the discussion of variables from the joint class with the MUSC 232 students. This should not be a blind analysis- you have had a conversation with a subject matter expert to suggest a starting point for your analysis. Your final set of variables may differ from this starting point discussed with MUSC 232 students, but this should be a well-justified analysis path. As mentioned in the joint class, if you look hard enough (varying artists, variables, etc.) you are likely to find a significant relationship just by random chance. Do not do this! Create a thorough and thought-out analysis based on your preliminary conversation with the subject matter expert and recognize that "null findings" (no significant relationships) are completely acceptable and should be interpreted fully as well.

Be careful that "key" and "mode" are categorical variables, even though they are coded as numbers. If you use these variables, you should convert them to factors using `as.factor()`.

```
#first spotify data, with crossover variable
#BE SURE TO DELETE THE LINE BREAK BEFORE RUNNING THIS
spotify_data_crossover <- read.csv("https://www.math.carleton.edu/ckelling/regression/
                                    report_cross_spotify_data.csv")

#second spotify dataset- all data, crossover labelled or not
#BE SURE TO DELETE THE LINE BREAK BEFORE RUNNING THIS
full_musc_data <- read.csv(file = "https://www.math.carleton.edu/ckelling/regression/
```

```
                              report_nocross_spotify_data.csv")


#subset to just a few artists
#spotify_data_subset <- spotify_data_crossover %>%
#  filter(artist_name %in% c(***artist names here***))
```

Overall tips

When analyzing these datasets, be sure to do the following:

- Keep in mind the goals of the analysis
- Check for interactions and the need for transformations
- To reach satisfactory depth of analysis, your final model should include at least three explanatory variables, unless fully justified that it is not necessary/appropriate. This includes quantitative and categorical variables.
- Check model assumptions often
- Be careful about independence in this analysis: what concerns do you have given the structure of the data you are using?
- Fully justify the final dataset and model using tools discussed in class
- We will learn more about multiple regression as the course progresses- you are only expected to use the tools you know of before the report is due

Report Guidelines

The main text should be **UNDER 4 PAGES**, including figures. Your report should be organized and clearly written. It should contain accurate and precise language and justifiable interpretations of statistical evidence, and should have the components listed below. The report should be a summary of your analysis, for an informed audience more interested in conclusions than in technical details. This text should include the following clearly labeled parts. *Clearly labeled* means that each section should have a bold-font section title that corresponds to the labels given below. The percentages listed below indicate how the report will be graded (**see also attached Rubric**). You should include the RMarkdown appendices (code and knitted pdf) as separately attached documents.

- **Title and Author(s)** (5%): Your title should be an informative and engaging description of your report (i.e., not "Short Report").
- **Introduction** (15%): A brief introduction to the questions addressed. Also include any relevant information about the data and how they were collected, and background information as appropriate. Avoid making broad and unsubstantiated claims, and cite evidence for claims that are not common knowledge if applicable. Give enough background in the intro so that a reader doesn't need to refer to source material, but take care not to plagiarize your source material above. Cite sources as needed. You will need to cite *at least two peer-reviewed sources/papers* to provide context and motivate your analysis!
- **Results** (50%): Both exploratory and formal statistical analyses. Include the following: (1) at least one relevant graphical exploratory display of your data (labeled appropriated, and described in text, *not involving a model*), (2) estimates and uncertainty quantification from your model, and (3) relevant interpretations in context. Before jumping into the model building process, you should start your Results section by introducing your dataset and performing some EDA to motivate your analysis. Technical language should be used correctly. This section should focus on conveying the *results* of your analysis to your reader and should not be a step-by-step description of what you did. For example, there is no need to tell the reader how many different transformations you tried before finding a linear association. Do not include screenshots of R tables- these should be formatted nicely in your report as tables.

- **Discussion** (15%): Summarize your findings, and describe any limitations of your analysis as well as questions that you would have been interested in answering but that are outside the scope of this short report. Be sure to include a healthy discussion of any necessary caveats here.
- **R Markdown Appendices** (5%): Your main text should suppress R commands and output if you write it in RMarkdown. You are welcome to write the report in Google Docs. In addition to your report, you will submit your Rmd document and your knitted code as separate documents (three files total). R code should be appropriately commented so that the technical details of your analysis are clearly communicated and the code should approximately follow the order of your report.

Finally, 10% of the grade is based on **overall presentation and group evaluation**: is the report structured according to the instructions given above? Is the text neatly presented, broken into readable paragraphs and laid out in a coherent manner? An excellent report will be written so that a reader with basic statistics training can understand your data, your analysis, and your conclusions. Do not simply write a step-by-step description of what you did in R. At the conclusion of the project, I will ask you to evaluate your group members. I expect satisfactory peer evaluations, showing that you contributed to the assignment.

In describing your conclusions, be sure to pay attention to the study design and be clear about what can be learned from these data (e.g., sampling bias, causality, etc.). You should answer a question of interest, motivated by your conversation with the subject matter expert, regardless of what the answer ends up being; in particular, a result does not need to be "statistically significant" for the analysis to be valid and interesting!

Best Practices for Partnered Work

Partners should work on this report *together*. While you are permitted to do some work separately (e.g., revising a section at a time when your schedules don't permit meeting together), I discourage you from completing substantial portions of the work individually. Each partner should contribute substantively to each component of the report, and each partner is assumed to be an equal contributor to the final product. You should be in regular contact with your partner(s) and should arrange at least two meetings to discuss your work. If you are having issues communicating with your partner(s), **please email me as soon as possible**.

My suggested timeline is as follows:

- **Week 1:** Meet as a group to explore data, solidify research questions, and develop plan for preliminary model-building. Conduct exploratory model building steps.
- **Week 2:** Early in Week 2 of the project, finalize your model and related plots. Begin writing early in Week 2 to allow for enough time for feedback and editing between group members.

Academic Honesty

Just a reminder, but submitted work is assumed to be by the author(s) unless proper attribution is made. I do not expect you to do significant outside research for this assignment (other than the two required sources for context/motivation), but remember to cite your sources that you use (using any standard citation method). I expect that the primary intellectual contribution to the project is from the authors. One example of permissible outside collaboration is receiving help formatting a graph from the stats lab. One example of impermissible outside collaboration is receiving extensive guidance on how to construct a confidence interval for the quantity you're interested in estimating. Please come to me if you need assistance for this project outside of your assigned group.

Title and Author(s) (5%)

- Has a title and lists the authors
- The title is an informative and engaging description of your report

---

Introduction (15%):

- Introduces and motivates the topic being addressed
- Includes at least two peer-reviewed sources for context for this study
- Explains the questions addressed in the report
- Gives appropriate background information (and avoids making broad and unsubstantiated claims)

Results (50%):

- Focus is on conveying the *results* of the analysis and is not a step-by-step description
- Exploratory graphical summary of data (not involving a model- just EDA)
  - Figure(s) is/are informative about the question addressed in the report
  - Figure(s) is/are labeled appropriately, and described in text
- Statistical inference
  - A complete model statement is given (give the equation!)
  - An appropriate model is chosen
  - Model assumptions are assessed and satisfactory (including independence)
  - Estimates and uncertainties of model coefficients are given
  - The model and estimates are interpreted correctly and in the context of the question being addressed
  - Technical language is used correctly

Discussion (15%)

- Summarizes the main findings and answers the question posed in the introduction
- Describes any limitations of your analysis and any necessary caveats
- Describes any follow-up questions that would be interesting to investigate

---

R Markdown Appendices (5%)

- R code used for the analysis is shown and implemented correctly
- R code is approximately ordered as the results are presented in the report

- Code is commented in a way where results are easy to understand and reproducible

---

Overall presentation (10%)

- The report is structured according to the instructions
- The text is neatly presented, broken into readable paragraphs and laid out in a coherent manner
- Peer evaluations were completed for group members and satisfactory from group members