You will be working on a research paper for your final project.  This project will include identifying a topic/problem that you want to solve using data science.  While the final solution to the problem does not need to be provided via programming – you will be doing some exploratory data analysis, transformations, and summary statistics on the data via R.  You are welcome to create a model based on what you have learned in this course to solve the problem, but this is not required.  Instead, a recommendation is required for a model or method you would implement to solve the problem.  There are 3 steps to this project, with the final deliverable being a formal paper (completed in R Markdown) that outlines the problem, shows the analysis done with the data, and concludes with your recommendation for next steps.  Each step provides detailed information that you must include in each phase of the project.

* Identify a topic or a problem that you want to research.  Provide an introduction that explains the problem statement or topic you are addressing. Why would someone be interested in this?  How is it a data science problem?
  + Spotify data
  + How is my music taste different from others and the general population?
* Draft 5-10 Research questions that focus on the problem statement/topic.
  + Do I perceive certain aspects of music differently from others?
  + How do my tastes compare to my peers?
  + What statistically different deviations does my taste have against the population?
  + Can I quantify the degree of ‘indie’ my taste in music trends towards?
  + What might be some new genres I would enjoy, based on my current taste?
  + What aspects of a given song do I most enjoy?
  + What aspects do I enjoy the most in certain genres?
  + How might I weight and measure my enjoyment of a song saved in my library?
* Provide a concise explanation of how you plan to address this problem statement.
  + Getting a handle on the available data (my set vs the population) and understanding the existing measurements of song characteristics and summary statistics
  + Conducting ANOVA on samples of my taste to the population or most popular songs within a given genre
  + Predicting songs that I would like the most given a training dataset (using the variables provided by Spotify API)
* Discuss how your proposed approach will address (fully or partially) this problem.
  + The variables provided by Spotify give a cursory insight into objective qualities of music as determined by their proprietary algorithms. Leveraging these may help me better understand how my taste might evolve, and better qualify what my taste is.
* Do some digging and find at least 3 datasets that you can use to address the issue. (There is not a required number of fields or rows for these datasets)
  + Original source where the data was obtained is cited and, if possible, hyperlinked.
    - Spotify API data pulls
    - My Spotify data (Playlists, my favorites playlist, streaming history 0 & 1)
  + Source data is thoroughly explained (i.e. what was the original purpose of the data, when was it collected, how many variables did the original have, explain any peculiarities of the source data such as how missing values are recorded, or how data was imputed, etc.).
    - Request from spotify & API key
    - My data is one year snapshot of activity and playlist
* Identify the packages that are needed for your project.
  + Jsonlite
  + Spotifyr
  + dplyr
  + ggplot
* What types of plots and tables will help you to illustrate the ﬁndings to your research questions?
  + Boxplots & histograms to analyze distributions, violin & scatter plots to assess correlations,
* What do you not know how to do right now that you need to learn to answer your research questions?
  + Better usage of the jsonlite and spotifyr packages
  + Better understanding of Spotify provided variables
  + Understanding relative significance (Bayesian function?) of songs listened to but not liked (right outer join from Liked Songs to Streaming History)

You can use the following template for Step 1:

* Introduction
* Research questions
* Approach
* How your approach addresses (fully or partially) the problem.
* Data  (Minimum of 3 Datasets - but no requirement on number of fields or rows)
* Required Packages
* Plots and Table Needs
* Questions for future steps

At this point you should have framed your problem/topic, described the data, and how you plan to solve the problem.  Now you need to move on to the next step of analyzing and preparing the data. Adding on to the draft you started in Step 1:

* Data importing and cleaning steps are explained in the text and follow a logical process.  Outline your data preparation and cleansing steps.
  + Removing metadata
  + Subsetting data into concepts
  + Grouping data by genre (and playlist?)
  + Removing irrelevant nulls
  + Addressing formats and scales
* With a clean dataset, show what the final data set looks like. However, do not print off a data frame with 200+ rows; show me the data in the most condensed form possible.
  + Showcase data of favorites
* What do you not know how to do right now that you need to learn to import and cleanup your dataset?
  + Additional grouping methods; methodological grouping and sorting by model efficacy
* Discuss how you plan to uncover new information in the data that is not self-evident.
  + Attempt predictions of key values and contributing factors to assess regression assumptions (normal dist, heteroskedasticity, multicollinearity)
* What are different ways you could look at this data to answer the questions you want to answer?
  + Group by genre, song age, stream age
  + Residual vs. population
  + Popularity scale
* Do you plan to slice and dice the data in different ways, create new variables, or join separate data frames to create new summary information? Explain.
  + Above grouping and sub-setting with mirrored analysis
  + Outer join streaming to favorites playlist to determine less-liked songs; compare to liked songs
  + Generate recommendations from predictors for liked songs (training data)
* How could you summarize your data to answer key questions?
  + Assess confusion matrix and relative impact between variables to determine relative importance of song features
  + Compare genre streaming frequency to frequency within favorites playlist to assess favorite genres
  + Summarize unique genre values to identify favorite genres
  + Summarize concatenated genres to identify favorite niche/ sub-genres
* What types of plots and tables will help you to illustrate the findings to your questions? Ensure that all graph plots have axis titles, legend if necessary, scales are appropriate, appropriate geoms used, etc.).
  + Continue with density/violin plots to assess relationships of attributes to genres
* What do you not know how to do right now that you need to learn to answer your questions?
  + ???
* Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.
  + Cluster analysis would be an asset for determining natural groupings within favorites and streaming histories

Some additional questions you may want to consider asking yourself as you work through this section of the project:

1. What features could you ﬁlter on?
   1. Ranges of song features
   2. Alignment (residuals) between mean values of song features and variables
2. How could arranging your data in different ways help?
3. Can you reduce your data by selecting only certain variables?
   1. Excluding album name from songs and last modified date from playlist
   2. Only including favorites playlist
   3. Maybe including metal favorites
4. Could creating new variables add new insights?
   1. Variables storing summary statistical values and residuals to compare across sets and categories
5. Could summary statistics at different categorical levels tell you more?
   1. Aggregating by genre over simple relational subsets
   2. Summarizing preference over time series of streaming history or added date to playlist
6. How can you incorporate the pipe (%>%) operator to make your code more efﬁcient?
   1. Pipe for ddply subsets

You can use the following template for Step 2:

* How to import and clean my data
* What does the final data set look like?
* Questions for future steps.
* What information is not self-evident?
* What are different ways you could look at this data?
* How do you plan to slice and dice the data?
* How could you summarize your data to answer key questions?
* What types of plots and tables will help you to illustrate the findings to your questions?
* Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.
* Questions for future steps.