**Introduction:** Introduce your problem(s)/objective(s). Discuss why you trying to predict this variable (motivation). Why is this useful? Describe the data source(s) you will use to build a predictive model. If you are obtaining information from websites you should site those in your **References**.

**Data Overview:** This might be more of an **Exploratory Data Analysis**; especially if the data is completely new to you. At minimum the response variable should be explored and analyzed in detail. Along with an inspection of the data for missingness and severe class imbalance of categorical data. The previous analyses should be conducted on the entire dataset. Further exploration such as exploring relationships and transformations should be conducted on either a standalone dataset used only for an EDA or some portion of the training dataset from the initial split (can put it back in to the training set when building models). Data from the final testing dataset (performance dataset) should not be used for this. This is a key step in feature engineering.

**Methods:** State and describe the model types you will be fitting. Describe any parameters that will be tuned. Describe what recipes will be used. Describe the resampling technique used. In some cases an extended discussion about recipe variations might be useful. Especially if students are using recipe variation to try and explore the predictive importance of certain variables. Explain the metric that will be used to compare and ultimately used to select a final model.

**Model Building & Selection Results:** Should reiterate the metric that will be used to compare models and determine which will be the final model. Include a table of the best performing model results. Review and analysis of tuning parameters should happen here. Should further tuning be explored? Or how should tuning be adjusted when fitting data like this in the future. This would be a good section to describe what the best parameters were for each model type. Could include a discussion comparing any systematic differences in performance between model types or recipes. If variations in recipes were used to explore predictive importance of certain variables, then it should be discussed here. The section will likely end with the selection of the final/winning model (provide your reasoning). Was it surprising or not surprising that this particular model won? Explain.

**Final Model Analysis:** This is where you fit your final/winning model to the testing data. Assess the final model’s performance with at least the metric used to determine the winning model, but it is also advisable to use other performance metrics (especially ones that might be easier to communicate/understand). Should include an exploration of predictions vs the true values (graph) or a confusion matrix (table). Remember to consider the scale of your outcome variable at this time — did you transform the target variable? If a transformation was used, then you might you should consider conducting analyses on both the original and transformed scale of the target variable. Is the model any good? It might be the best of the models you tried, but does is the effort of building a predictive model really pay off — is it better than a baseline/null model? Were there any features of the model you selected that make it the best (e.g. fits nonlinearity well)?

**Conclusion:** State any conclusions or discoveries. This is a great place for future work, new research questions, and next steps.

**References:** Any references used should be sited here. This includes but is not limited to where you got your data. There is no “formal” reference guideline but we recommend APA format. Example: Lastname, F. M. (Year, Month Date). *Title of page*. Site name. URL

Predictive Question: "Can we predict the popularity of a song based on its genre, tempo, and energy?"

Objective: To build a regression model that can accurately predict the popularity score of a song based on its genre, tempo, and energy. This model could be used by music producers to identify which genres, tempos, and energy levels are most likely to result in popular songs, and to guide their songwriting and production decisions accordingly.

Reason:

some reasons why you might choose to focus on only three variables (genre, tempo, and energy) instead of using all the available variables in your dataset:

1. Simplicity: Sometimes, using fewer variables can make it easier to build a model that is interpretable and can be easily understood by others. In some cases, using fewer variables can also make it easier to train the model and to avoid overfitting.
2. Relevance: Depending on the problem you are trying to solve, some variables may be more important or relevant than others. For example, if you are interested in predicting the popularity of a song, the genre, tempo, and energy levels of the song may be more relevant than other variables such as speechiness or instrumentalness.
3. Data quality: Sometimes, some variables in a dataset may have missing values, outliers, or other quality issues that can affect the accuracy of the model. By focusing on a smaller set of variables that are more reliable, you can potentially build a more accurate model.