**Final Presentation Speaker Notes**

**Team Introductions - ALL - 1 minute:**

* Hayley (sharing screen) - 15 seconds
* Jennifer - 15 seconds
* Maya - 15 seconds
* Lijoy - 15 seconds

**Introduce Topic - Hayley - 3 minutes:**

* SLIDE 1
  + Today we’ll be talking to you about the relationship between the Billboard Hot 100 chart and the Grammy Awards
* SLIDE 3
  + We’ll begin by introducing our central question - “Does inclusion on the Billboard Hot 100 chart increase the likelihood that a song will win a Grammy Award?”
  + We chose this topic because the 2021 Grammy Awards had just taken place, and I had prior experience in the music industry which led me to be curious about consumer opinions vs. awards/critical recognition. The rest of my team agreed that this would be a topic with many avenues for exploration, so we began our search for a dataset. - 30 secs
* SLIDE 4
  + We found a detailed Dataset on Kaggle.com that provided data on the Billboard Hot 100 list from 1999-2020. The Dataset also included several CSVs containing information on Spotify listens, Grammy Award Wins, and iTunes purchases. We decided to compare the Billboard Hot 100 Data with Grammy Award Wins, because both the Grammy awards and the Billboard Hot 100 chart are determined by elite committees. Therefore, an additional question for analysis is “Are these committees in agreement on which songs deserve commendation?” - 1 min
  + Once we had chosen our data and central question, the next step was the Data Exploration Phase. We used Excel to examine the data and determine what data cleaning steps would be necessary. Fortunately both the Billboard and Grammy CSVs had consistent formatting and contained a reasonable number of null values. At this point in our exploration we identified additional variables that would be useful in our analysis - Weeks on the Chart, Peak Position, and Genre. The next step was to create visualizations for our initial analysis.
* SLIDE 5
  + Our first chart examines the relationship between the total number of weeks a song was included on the Billboard chart, and the total number of grammy wins. In order to visualize this relationship, we used Excel to create a basic bar chart. We included bins to easily visualize our data. The results of this initial analysis showed that there is little correlation between weeks on the chart, and a grammy win.
* SLIDE 6
  + Our second chart examines the relationship between the peak position that a song reached on the Billboard chart, and the total number of grammy wins. Again we used Excel to create the bar chart, and created the same bins for consistency. Unlike the relationship between weeks on the chart and grammy wins, the results of this initial analysis showed that there did appear to be some correlation between the peak position on the chart, and the likelihood that a song will win a grammy award
  + Now that we had completed our initial data exploration, we were ready to move on to the Data Analysis phase of our project. I will hand it over to Lijoy for an explanation of our Database and Dashboard.

**Dashboard - Lijoy - 3 minutes:**

* Tools used - Python, Postgres, SQL, Tableau
* How was the Database created?
* We have taken Billboard and Grammy data set for our analysis
* We have used quick database diagram tool to design our database
* As a team we have decided to use Postgres as our database
* Used Python to clean Billboard and Grammy data set
  + Got single instance of each song by dropping duplicates
  + Dropped "weekly\_rank", "writing\_credits", "lyrics", because they are not relevant for this analysis
  + Totaled number of Grammy Awards for each song to get total number of Grammy Wins
* Loaded these cleaned CSVs into Postgres database using **sqlalchemy**
* Then performed a join between Billboard and Grammy tables to get one complete dataset.
* Later we decided to add genre also into our final dataset to check if genre has any impact on grammy awards, so performed these steps again in an iterative design mode to add genre
* Now our final data set is ready in Postgres - We have used this for ML and in Tableau for dashboard creation.
* Dashboard Walkthrough
  + Explain initial analysis graphs
  + Indicate MLM accuracy score & that Maya would be explaining further
* Interactive features?
  + Walk through at least 2 genres

**Machine Learning - Maya - 3 minutes:**

* Explanation of model choice,including limitations and benefits
  + For our project we chose a Logistic Regression model to best serve our needs in predicting whether or not an artist that is on the Billboard Top 100 Chart will win a Grammy because we are dealing with binary outcomes
  + Limitations
    - Cannot work with non-linear outcomes
    - Need large data set
    - Challenges faced in looking at remixed versions of songs that included a featured artist included on Billboard Top 100 versus original songs and original artist included winning a Grammy
  + Benefits:
    - Simple to implement and make predictions for binary outcomes (yes/no)
    - Simple to understand and update with new data to be used in the future
* SLIDE 2
* Feature engineering and Feature Selection
  + Dropped Variables: "weekly\_rank", "writing\_credits", "lyrics"
  + These variables are not informative for our MLM analysis.
* Description of how data was split into training and testing sets
  + We are using X to predict y.
  + y is the “GrammyAward” column, or the output
  + X, or features, is created by dropping the “artists’’ and “name” columns from the DataFrame.
  + Use the train\_test\_split module to split X and y into training and testing sets: X\_train, X\_test, y\_train, y\_test.
* Description of how they have trained the model thus far, and any additional training that will take place
  + We are comparing the actual outcome values from the test set against the model's predicted values.
  + y\_test are the outcomes (whether or not an artist that is on the Billboard Top 100 Chart will win a Grammy) from the original dataset that were set aside for testing. The model's predictions, y\_pred, were compared with these actual values, y\_test.
* Description of current accuracy score
  + 0.9411764705882353 (The accuracy score is simply the percentage of predictions that are correct.)
  + The accuracy of the MLM, 94.12%. This shows that the MLM will accurately predict whether an artist will receive a grammy based on their Billboard Top 100 Chart performance 94.12% of the time.
  + Confusion Matrix
    - Useful to include both the F1 score and confusion matrix to further assess the precision and sensitivity of the MLM.

**Results - Jennifer - 3 minutes:**

* Does the model work? Can wins be predicted with new data
  + Add in missing information with other data source
  + Grammys started 1959 Billboard started 1894 - locate all possible historical data
* Future features
  + Scraper to add weekly data and populate database each week
  + Use additional, modern data sets from spotify or pandora - does population streams have impact?
* Additional analysis
  + Does genre have an impact on win? Interact with dashboard
  + When is a song released in relation to the Grammy awards?