Team 3-1: Analyzing Billboard Songs

Aimi, Sonya, Ethan, Luca

Introduction

"The Billboard Hot 100 is the music industry standard record chart in the United States for songs, published weekly by Billboard magazine. Chart rankings are based on sales (physical and digital), radio play, and online streaming in the United States."

The data was found on TidyTuesday and is from Data.World with the original data points found on Billboard.com and Spotify. There are 2 sets of data. One is from the Billboard.com. The cases are songs from the certain week(s) in which they appeared on the Billboard 100 chart. It includes every weekly Hot 100 singles chart from Billboard.com. The other is from Spotify. The cases are songs with their audio features (such as genre, danceability) as the other variables. We combined these 2 dataset into one by joining with the common variables: song name and performer. For our final dataset, the cases are songs from the certain week(s) in which they appeared on the Billboard 100 chart along with the songs' audio features.

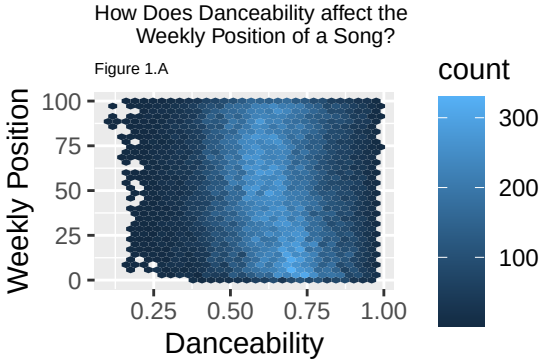
From this dataset, we're interested in examining songs from 2000 to today. This gave us 114,319 datapoints. Since there are so many song attributes in the dataset, we decided to focus on attributes that are easily identifiable through hearing by the general population: genre, danceability, energy, speechiness, valence, and tempo. For more details about our variables, please refer to our ReadMe. Our main research question is: what factors influence a song's ranking?

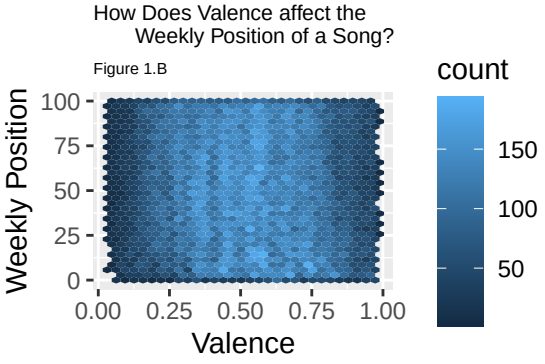
Looking into Song Rankings

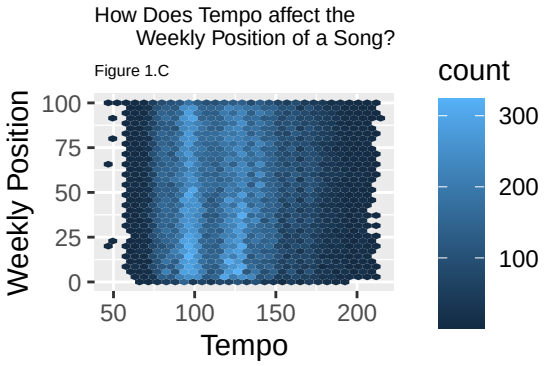
To start answering our research question, we first took a general look at how attributes influence song rankings.

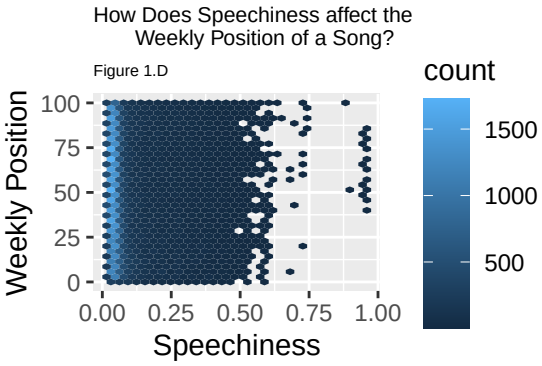
Correlation Between Song Rankings and Song Attributes

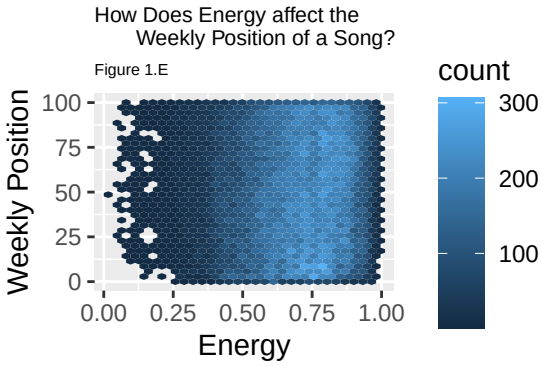
To accomplish this, we created visualizations to see the correlation between songs' weekly positions and song attributes. We hypothesize that songs with higher speechiness and tempo will be ranked higher, given the rising popularity of rap, and that other attributes would correlate less with songs' rankings.











When examining the graphs, we know there will be an even distribution of data along the y-axis, since there is one #1 song, #100 song, and everything in between from each week. If there is an association between variables, though, there would be non-vertical patterns in the data. Danceability (figure 1.A) seems like it may be slightly curved to the right as y approaches 0, which could suggest that more danceable songs tend to rank higher. Valence (figure 1.B) appears to have no pattern at all, whereas songs seem to for the most part have danceablilty in the 0.5 to 0.75 range, the valence chart has songs in all areas of the graph. Once again, this doesn't display a perceivable association. Tempo (figure 1.C) is similar, with two vertical-shaped light streaks suggesting that many songs, regardless of rank, have tempos around 100 and 120. Speechiness (figure 1.D) seems to be a variable that is centered close to 0 regardless of rank, which again suggests no perceivable association. Similarly, for energy (figure 1.E), regardless of weekly position, the range of energy for songs seem to be fall between 0.5 and 0.9. There seems to be a small cluster of high ranking songs whose energy level is at 0.75, but the visual pattern isn't that stark.

At this point, it does not seem likely that song attributes are valuable predictors of song rank/ weekly position. However, there is a lot of data displayed here, so much so that the graphs have points in a majority of the hexbins available. This could make it harder for us as humans simply looking out a plot to observe a relationships between song position and audio variables that may be hiding in the data. To investigate this, we used linear regression.

Testing Predictors

To further examine the relationship, we looked to see how well each song attribute predicted songs' weekly positions. Given the results from above, we hypothesize that danceability would be one of the highest predictors, since its visualization is one of the only visualizations to show a difference between differently ranked songs. We created linear regression models for each attribute and also a linear model with all the attributes as the explanatory variables. We then found the R-squared values for each model to see how well the explanatory variable(s) predicted songs' weekly positions.

Danceability Linear Model and its R-Squared Value:

```
## # A tibble: 2 x 5
```

```
##   term         estimate std.error statistic p.value
##   <chr>           <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)      66.9     0.404     165.        0
## 2 danceability    -25.8     0.615     -42.0       0
```

```
## [1] 0.01639905
```

Energy Linear Model and its R-Squared Value:

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    45.3      0.384     118.  0
## 2 energy          7.18     0.541      13.3 3.86e-40
```

```
## [1] 0.001661981
```

Speechiness Linear Model and its R-Squared Value:

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    51.1      0.124    412.   0
## 2 speechiness    -7.89     0.876     -9.01 2.05e-19
```

```
## [1] 0.000767592
```

Valence Linear Model and its R-Squared Value:

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    53.8      0.230     234.  0
## 2 valence        -6.64     0.403     -16.5 7.89e-61
```

```
## [1] 0.002557178
```

Tempo Linear Model and its R-Squared Value:

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)  45.2      0.379       119.  0
## 2 tempo         0.0417   0.00304      13.7 1.26e-42
```

```
## [1] 0.001769466
```

Overall Linear Model and its R-Squared Value:

```
## # A tibble: 6 x 5
##   term         estimate std.error statistic   p.value
##   <chr>           <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)   60.7      0.733       82.7  0
## 2 danceability -22.9      0.719      -31.8  4.83e-221
## 3 energy         6.66     0.610       10.9  1.10e- 27
## 4 speechiness    1.40     0.904        1.55 1.22e-  1
## 5 valence       -3.53     0.480       -7.36 1.92e- 13
## 6 tempo          0.0118   0.00313      3.76 1.68e-  4
```

```
## [1] 0.01774039
```

In analyzing which predictor correlated the strongest with the weekly position of the song we found that due to the vast amount of data that there was, no one linear model showed significant correlation in testing which attribute of a song was the strongest predictor. However, we cannot discount that the models do show the variation within the data to some extent.

Looking at the R^2 values, we can see that the danceability model and the overall model were the strongest predictors and accounted for the greatest amount of variation in the data. The danceability model presented a R^2 value of 0.0164 which signifies that about 1.64% of the variation can be accounted for through the linear model.

$$\widehat{weeklyposition} = 66.86 - 25.8(danceability)$$

This model shows that, if all else is held constant for each additional point increase in danceability causes the the weekly position of the song to decrease on average by 25.8. This shows that an increase in danceability causes a corresponding higher ranking in the song. This intuitively makes sense, as one would think that if a song seems to be more danceable then people would like it more and therefore it would be higher on the billboard.

Similarly, our adjusted R^2 value for the overall model is 0.0177. This tells us that, even after penalties for using multiple variables, the model with danceablity, energy, speechiness, valence, and tempo accounts for more of the variation in song position, 1.77%.

$$\widehat{weeklyposition} = 60.66 - 22.87(danceability) + 6.66(energy) + 1.40(speechiness) - 3.53(valence) + .0118(tempo)$$

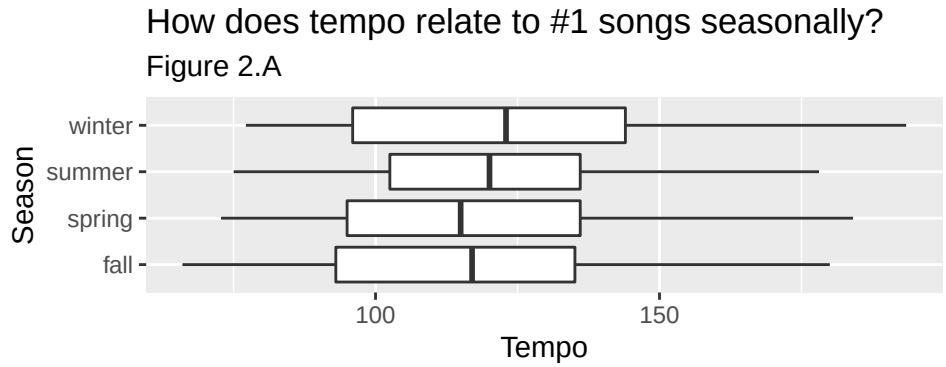
When looking at the estimates, it is important to keep in mind that negatives for slopes are better for a song's ranking. This is because a "decrease" in ranking actually brings it closer to the #1 spot! With that said, we see that increased danceability and valence seem to be better for a songs ranking, all else held constant, while increases in the other variables do not.

Looking into #1 Billboard Songs

Arguably, many artists' want to have a #1 Billboard song. So, to further investigate what factors play a role in songs' rankings, we looked at factors that influence #1 Billboard songs.

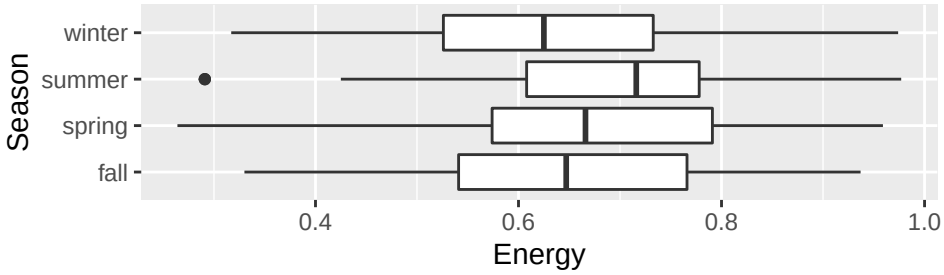
Seasons

Seasons can influence what song attributes are preferred. For example, winter has many holidays, so that we hypothesize that that happier songs with medium tempos will be #1 Billboard songs. We separated the dataset by seasons and looked at the distribution of attributes' values by creating boxplots. We defined fall as from September to November, winter as from December to February, spring as March to May, and summer as June to August.



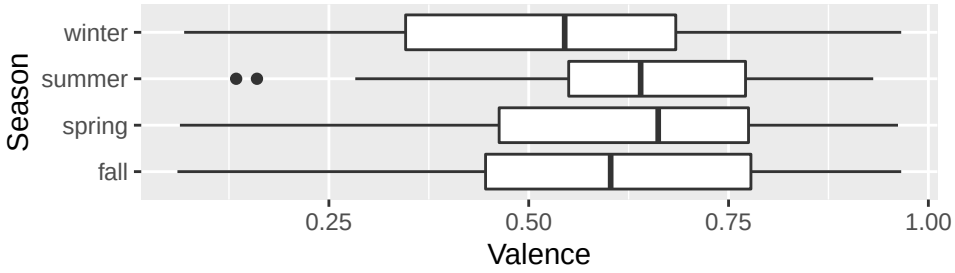
How does energy relate to #1 songs seasonally?

Figure 2.B



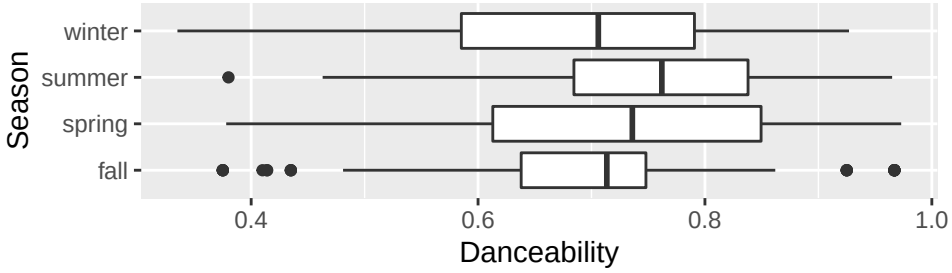
How does valence relate to #1 songs seasonally?

Figure 2.C



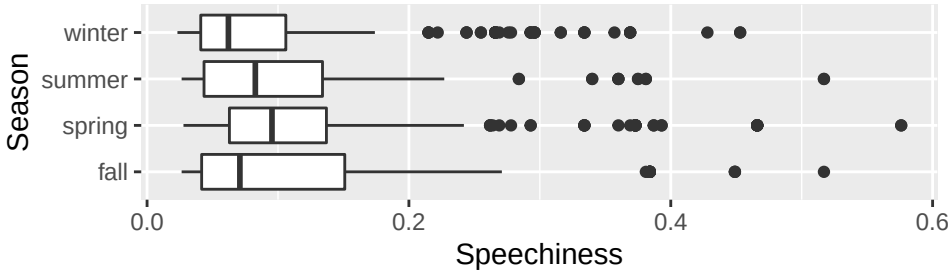
How does danceability relate to #1 songs seasonally

Figure 2.D



How does speechiness relate to #1 songs seasonally

Figure 2.E



In figure 2.A, the tempo of Billboard #1 songs are highest in the winter and lowest in the spring, although the differences between tempo across seasons is quite small. In figure 2.B, the energy of Billboard #1 songs have a clear relationship between the seasons as summer as significantly the highest energy songs compared

to winter. This makes sense because colder weather pairs better with slower songs and summery, warm weather pairs well with high-energy bops. Spring and fall have similar energy songs most likely due to the smaller difference in temperatures between the two seasons, compared to the wider temperature difference with winter and summer. In figure 2.C, winter and fall tend to have songs with lower valence compared to spring and summer. This makes sense with trends like sad-girl autumn and the winter blues. Because seasons can affect the moods of artists and consumers, people with lower moods in the cooler seasons of fall and winter, might consume more of the lower valence music in winter and fall. In figure 2.D, summer unsurprisingly has the most danceable #1 songs since summer also has the highest energy songs. Winter with lower valence, and lower energy songs, also have lowest energy songs. Spring has slightly more danceable songs than fall, but not by much. In figure 2.E, the box-plot illstrates that the most #1 songs regardless of the season tend to have low speechiness, or minimal spoken word. For some reason, #1 songs charting in the summer and spring have slighly more speechiness than winter and fall. Overall, each of the boxplots can reveal unique relationships between #1 songs and various attributes seasonally.

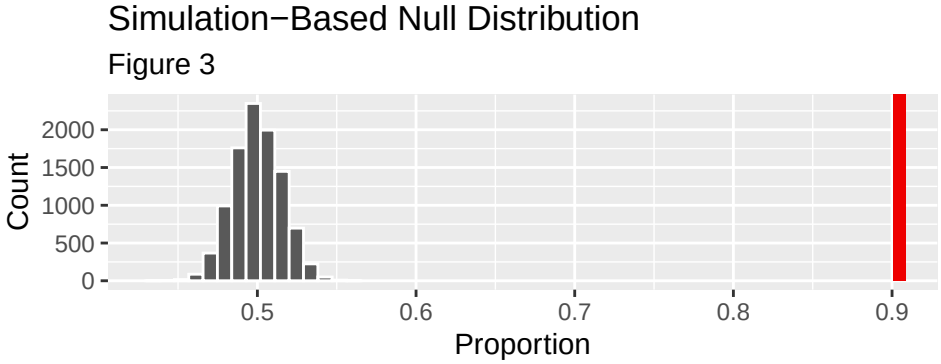
Testing Pop Genre

To examine the correlation between genre and #1 Billboard songs, we created a null and alternative hypotheses to test with simulation-based methods. Since the labels "pop genre" and "popular genre" tend to be interchangeable, we predict that the majority of songs that are #1 on the Billboard are of the pop genre.

Null Hypothesis: 50% of the #1 Billboard songs are of the pop genre. Alternative Hypothesis: Over 50% of the #1 Billboard songs are of the pop genre.

$$H_0 : p_{pop} = 0.5 \text{v.s.} H_a : p_{pop} > 0.5$$

To test our hypotheses, we first created a dataset of just the #1 Billboard songs. We then checked to see if the word "pop" appears in the Spotify genre column. With our dataset set up, we constructed a null distribution. We set the seed to make this reproducible. After constructing our null distribution, we created a visualization of it (Figure 3) and also calculated the p-value. Since our alternative hypothesis is greater than 0.5, we did not do a two-sided p-value; we only looked at it from the greater direction.




```
## [1] 0
```

Using the typical significance level of 0.05, we reject the null hypothesis. Given that the p-value is 0, there is strong evidence that the majority of #1 Billboard songs are of the pop genre.

To verify our conclusion from above, we also decided to perform a confidence interval Test. We decided to do a 95% confidence interval test since that seems to be the typical confidence interval. To perform this test, we first created a bootstrap distribution, then found the range of the middle 95% of the bootstrap distribution.

```
## # A tibble: 1 x 2
##   lower upper
##   <dbl> <dbl>
```

```
## 1 0.887 0.922
```

We are 95% confident that the proportion of #1 Billboard songs that are of the pop genre is between 0.89 and 0.92. Since this interval does not have 0.5 between it, we reject our null hypothesis, further providing evidence that the majority of #1 Billboard songs are of the pop genre.

Conclusion

To summarize, our main research question is: what factors influence a song's ranking? We visualized the relationship between weekly song positions and the audio features that we focused on. Through those, we found that danceability was the audio feature that correlated the most strongly with weekly song positions. This was confirmed with our linear models. We found that when you use all the audio features to create a linear model, it was the strongest predictor of weekly song positions. However, the R-squared values of that model and the danceability model were similar. When we looked at seasons, we realized that each season had their own "preference" for each audio feature. We confirmed our hypothesis that the majority of #1 Billboard songs are of the pop genre through hypotheses testing. In short, we found that higher ranked Billboard songs had higher danceabilities, seasons influence the audio features in #1 songs, and #1 songs are comprised mostly of pop.

Through our data analysis, we realized that the sheer amount of data made it hard for us to find patterns. For example, since there were so many data points, it made it hard to fit our linear models to our data, as evidenced by the R-squared values. We were able to create a general analysis on the factors that influence song rankings. However, we weren't able to look at specifics. For example, for genre, we only looked to see if the genre column had the word "pop" in them. Given the many genres of music and subgenres of pop, our analysis was pretty broad.

We also realized there are a multitude of other variables that our data set and our analyses didn't cover that could influence songs' weekly positions. Besides the song itself, the performers matter too. For example, if Adele and a small artist both released songs (and assuming the songs have similar numbers for their audio features), Adele's song would be expected to chart on the Billboard and rank highly. However, the small artist would not. In fact, the small artist might not even chart on the Billboard because their audience could be focused on entirely Adele's new song. Major artists could greatly impact the Billboard rankings and could influence other artists and when these smaller artists release their songs.

If we could redo this project, we would want to apply statistical techniques that would help us produce a model to predict song rankings. Our visualizations were able to show correlation between variables, our linear models gave us an idea on how the audio features predict song rankings, and we were able to conclude that #1 songs are mostly of the pop genre. However, we were not able to actually come up with a specific way/model to predict song rankings. We would love to figure out a way to simplify the data and come up with a model.