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?

```
## Warning in system("timedatectl", intern = TRUE): running command 'timedatectl'
## had status 1
```

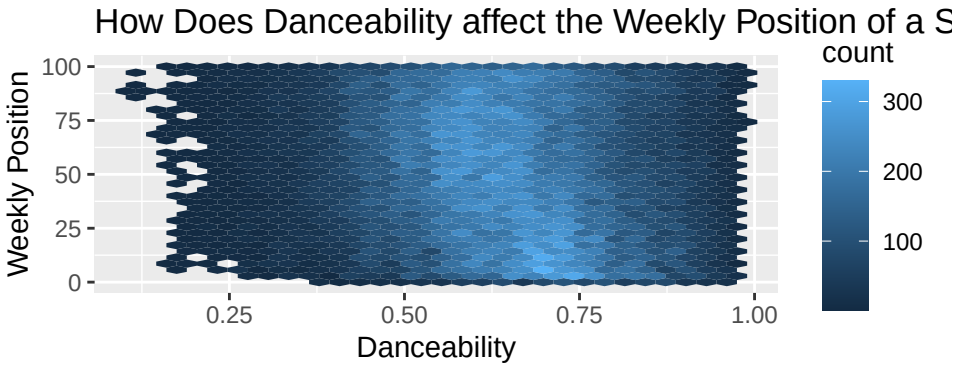
Looking into Song Rankings

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

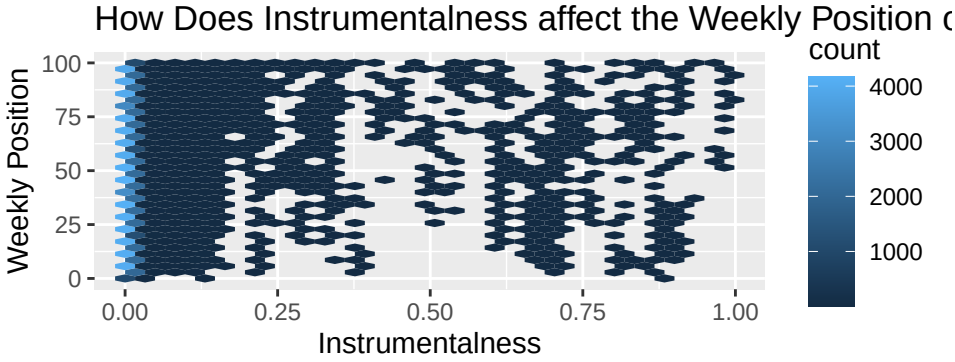
Correlation Between Song Rankings and Song Attributes

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.

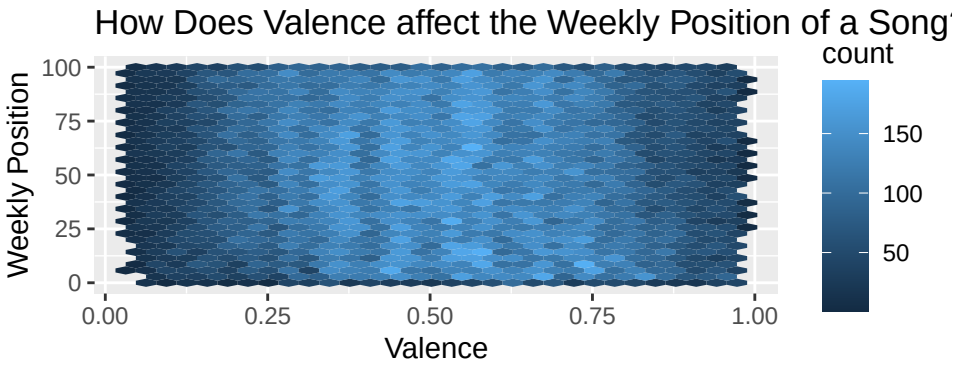
```
## Warning: Removed 8584 rows containing non-finite values (stat_binhex).
```



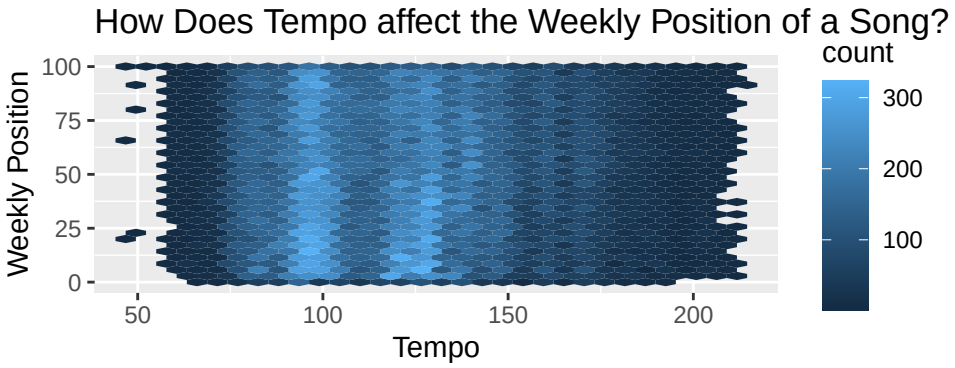

```
## Warning: Removed 8584 rows containing non-finite values (stat_binhex).
```



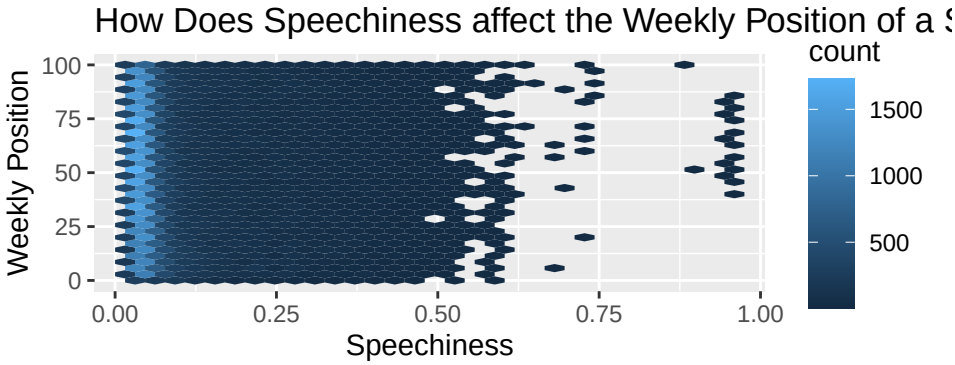

```
## Warning: Removed 8584 rows containing non-finite values (stat_binhex).
```




```
## Warning: Removed 8584 rows containing non-finite values (stat_binhex).
```




```
## Warning: Removed 8584 rows containing non-finite values (stat_binhex).
```



Insert discussion about the graph

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 value 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] 0.01778684
```

In analyzing which predictor correlated the strongest with the weekly position of the song we found that due to the

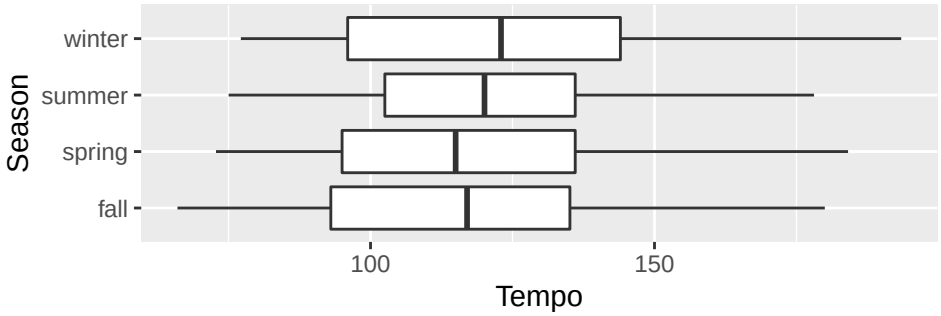
Looking into #1 Billboard Songs

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.

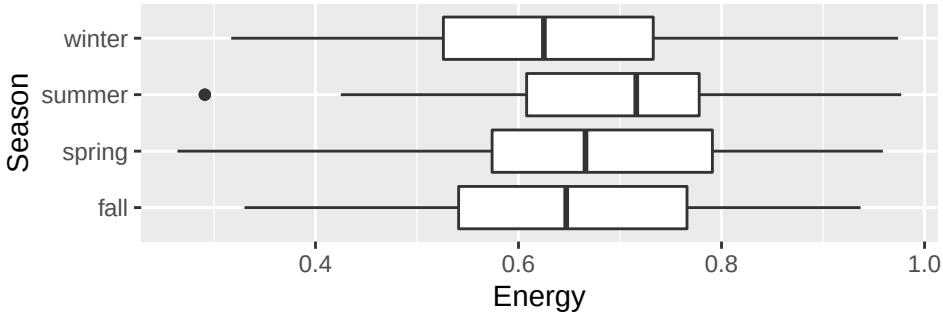
Seasons

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.

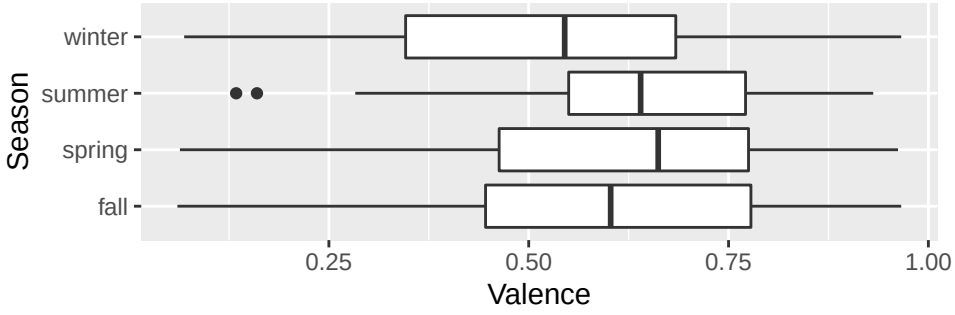
How does tempo relate to #1 songs seasonally?



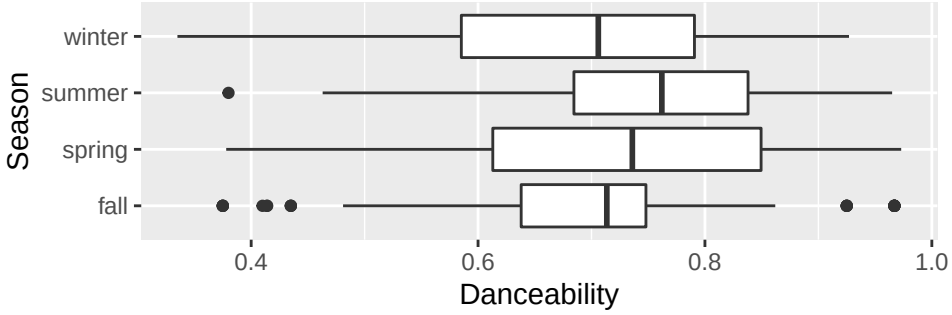
How does energy relate to #1 songs seasonally?

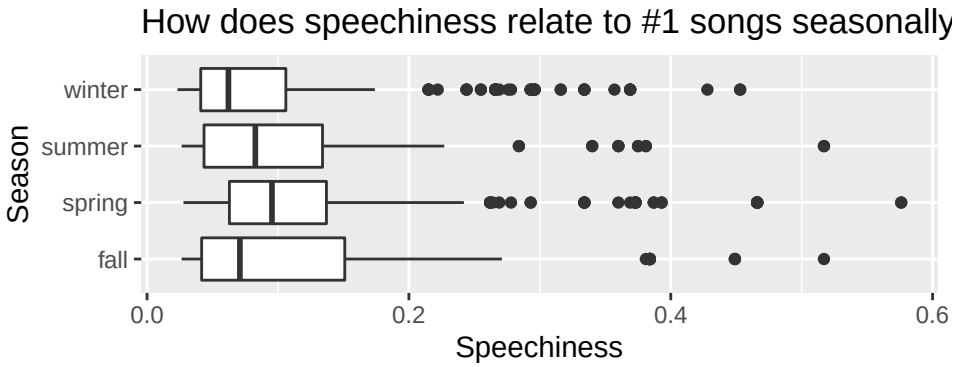


How does valence relate to #1 songs seasonally?



How does danceability relate to #1 songs seasonally?





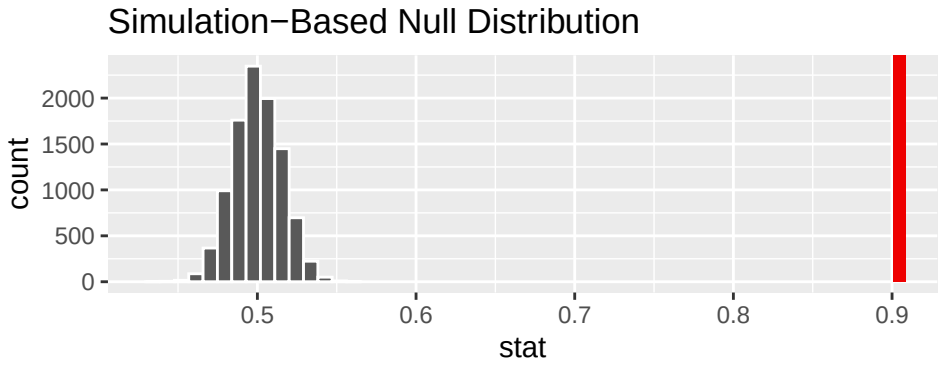
Testing Pop Genre

To examine the correlation between genre and Billboard rankings, we created a null and alternative hypothees 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 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

Need to add: This will require a summary of what you have learned about your research question along with statistical arguments supporting your conclusions. You should critique your own methods and provide suggestions for improving your analysis and future work. Issues pertaining to the reliability and validity of your data and the appropriateness of the statistical analyses should also be discussed. Also include a brief paragraph on what you would do differently if you were able to start over with the project.