Predicting What Effects a Song's Popularity

David Hohn

University of Washington-Tacoma

TBNALT 485: Business Intelligence

March 13th, 2024

Abstract

In the music industry there are many variables that could be attributed to a song's success. Things such as danceability, tempo, key, etc. This project aims to predict the variables that have the highest impact on the success of a song. The research uses a spotify dataset with 114,000 rows, and will implement the use of machine learning models and Tableau for visualization. By identifying key factors, this research should be able to provide insight and understanding on song popularity and what makes it resonate with an audience.

Data Dictionary

The descriptions were from the author of the dataset. There was no missing data in the set and was ready for use.

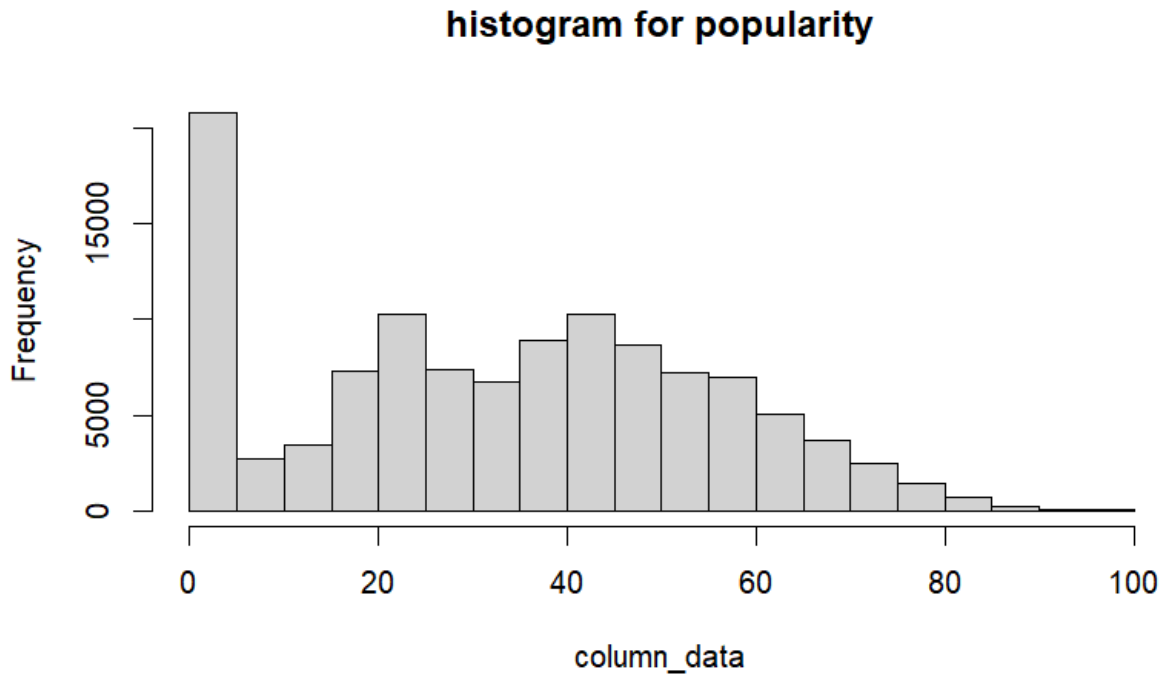
- **popularity:** The popularity of a track is a value between 0 and 100, with 100 being the most popular. The popularity is calculated by algorithm and is based, in the most part, on the total number of plays the track has had and how recent those plays are. Generally speaking, songs that are being played a lot now will have a higher popularity than songs that were played a lot in the past. Duplicate tracks (e.g. the same track from a single and an album) are rated independently. Artist and album popularity is derived mathematically from track popularity
- **duration_ms**: The track length in milliseconds
- **explicit**: Whether or not the track has explicit lyrics (true = yes it does; false = no it does not OR unknown)
- **danceability**: Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable
- **loudness**: The overall loudness of a track in decibels (dB)
- **key**: The key the track is in. Integers map to pitches using standard Pitch Class notation. E.g. 0 = C, 1 = C♯/D♭, 2 = D, 3 = D#, 4 = E, 5 = F, 6 = F#, 7 = G, 8 = G#, 9 = A, 10 = A#, 11 = B. If no key was detected, the value is -1
- **mode**: Mode indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived. Major is represented by 1 and minor is 0
- **valence**: A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry)
- **tempo**: The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration
- **track_genre**: The genre in which the track belongs
- **liveness**: Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live

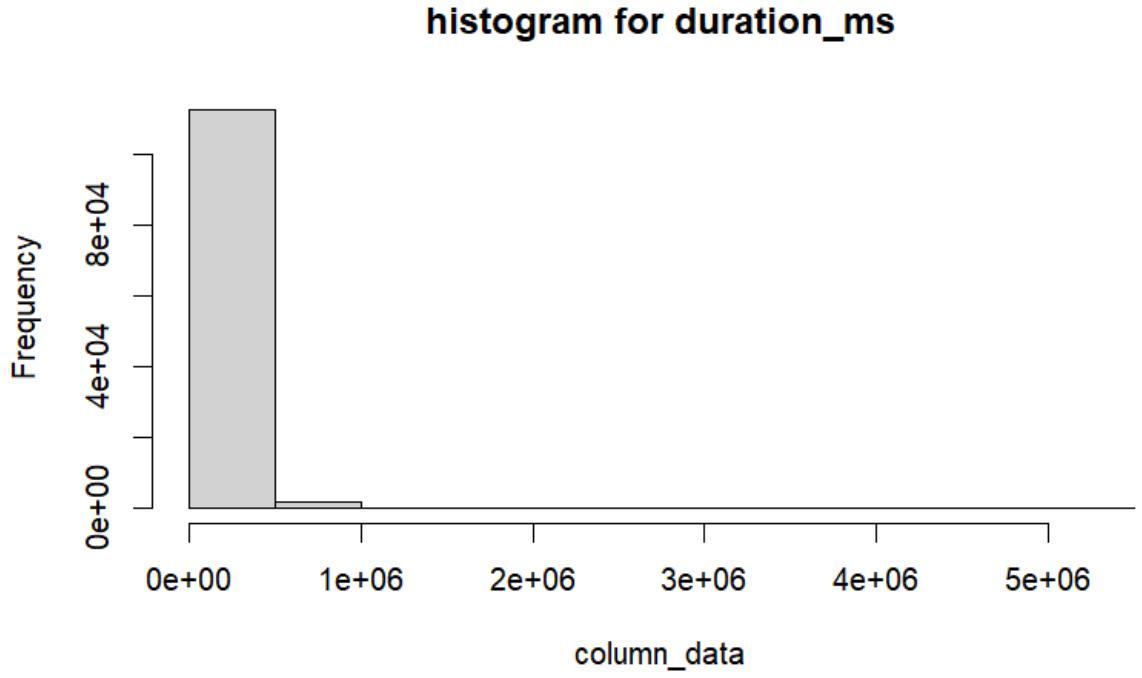
Univariate Properties

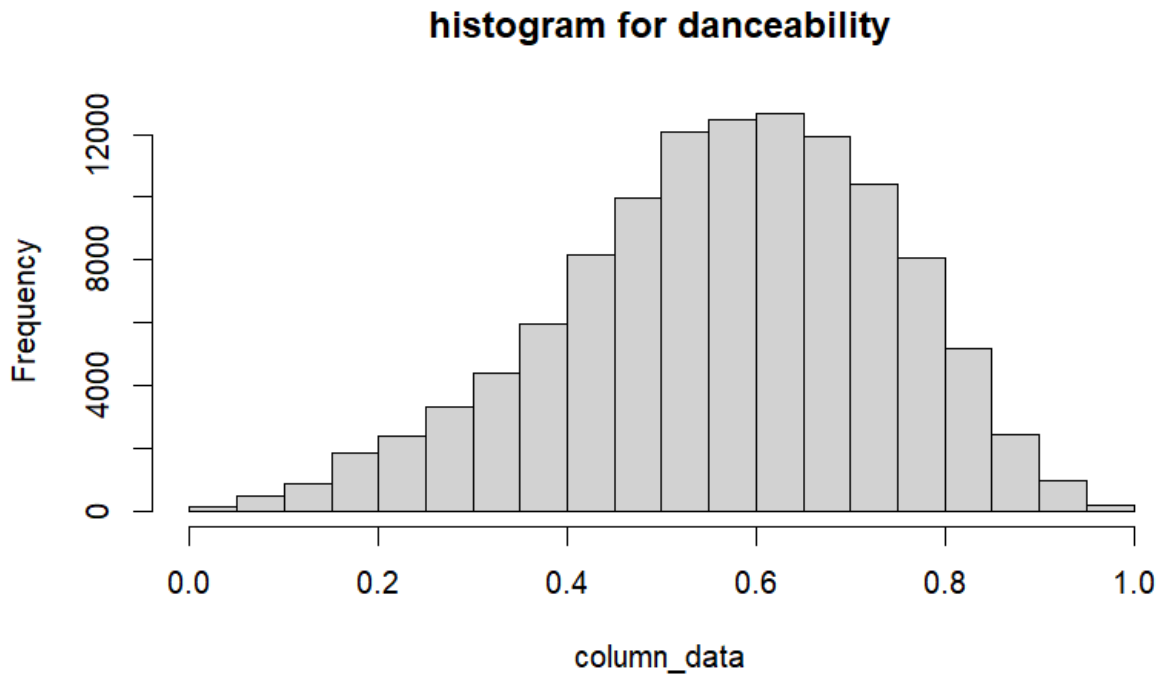
Feature	Var Type	Data Type	Count	Missing	%	Unique	Min	Q1	Med	Q3	Max	Mean	SD	Skew	Kurt
Popularity	Label	Numeric	114000	0		101	0.00	17.00	35.00	50.00	100.00	33.24	2.230508e+01	0.046401 91	2.072232 74
Danceability	Feature	Numeric	114000	0		1174	0.00	0.4560	0.5800	0.6950	0.9850	0.5668	1.735422e-01	-0.39 94914	2.81 54530
Loudness	Feature	Numeric	114000	0		19480	-49.531	-10.013	-7.004	-5.003	4.532	-8.259	5.029337e+00	-2.00 6516	8.89 5967
Tempo	Feature	Numeric	114000	0		45653	0.00	99.22	122.02	140.07	243.37	122.15	2.997820e+01	0.23 22918	2.89 13715
Track_genre	Feature	Categorical	114000	0		114							NA	NA	NA
Liveness	Feature	Numeric	114000	0		1722	0.00	0.0980	0.1320	0.2730	1.00	0.2136	1.903777e-01	2.10 5710	7.37 8024
Valence	Feature	Numeric	114000	0		1790	0.00	0.260	0.4640	0.6830	0.9950	0.4741	2.592611e-01	0.115 0765	1.97 25627
Key	Feature	Numeric	114000	0		12	0.00	2.0	5.0	8.0	11.0	5.309	3.559987e+00	-0.00 8500 249	1.72 3432 172
Explicit	Feature	Categorical	114000	0		2							2.796255e-01	NA	NA
Duration_ms	Feature	Numeric	114000	0		50697	0	174066	212906	261506	5237295	228029	1.072	11.19	357.

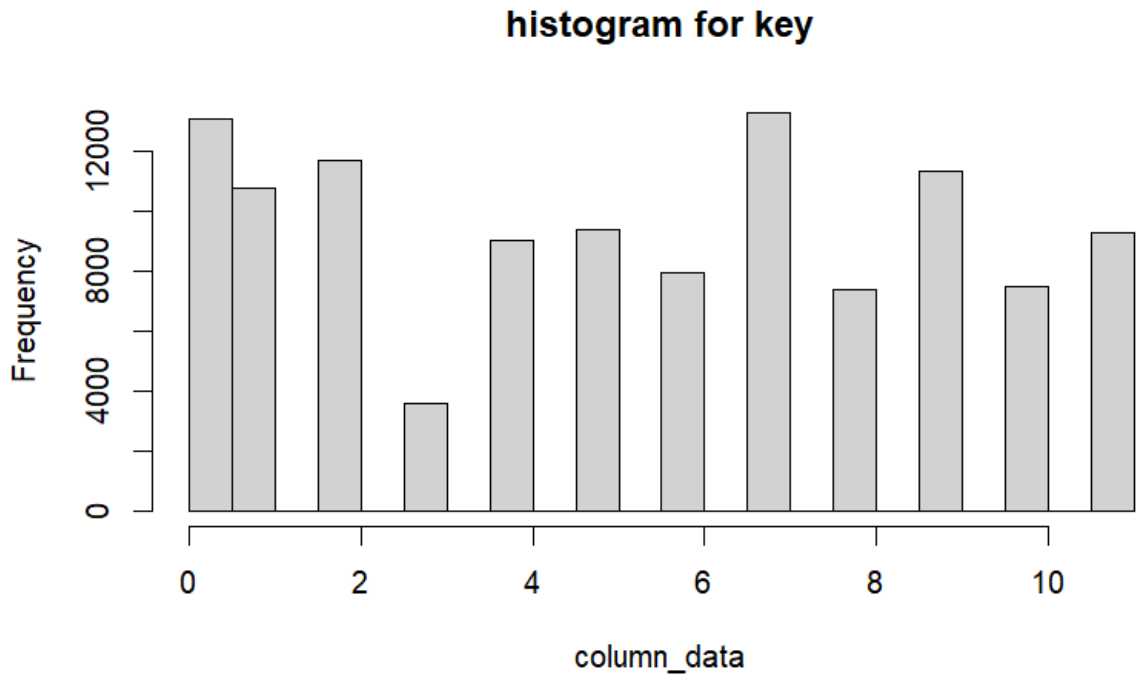
													977e+05	503	93680
Mode	Feature	Numeric	114000	0		2	0.0	0.0	1.0	1.0	1.0	0.6376	4.807092e-01	-0.572293	1.327519

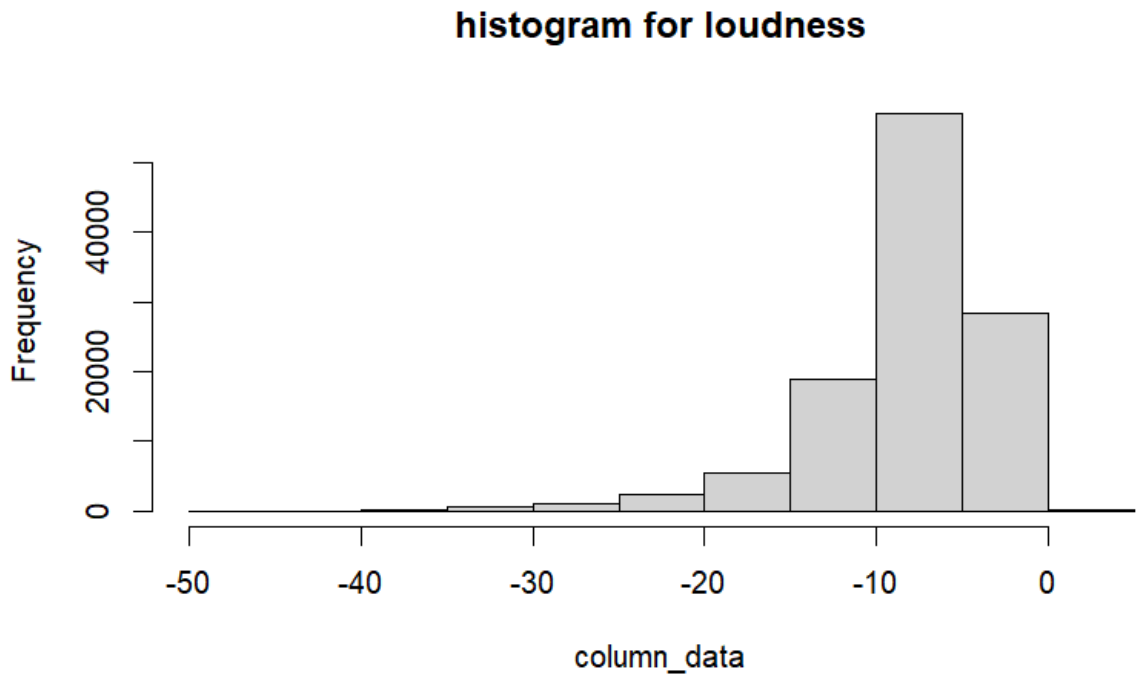
Histograms of numeric values

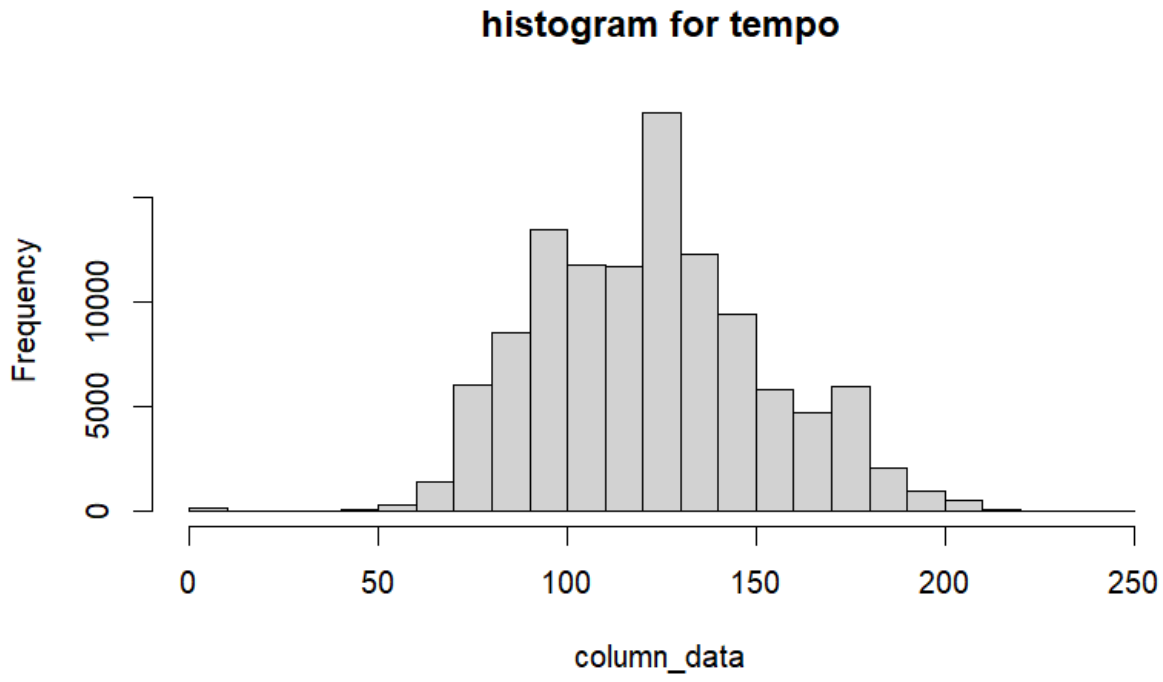


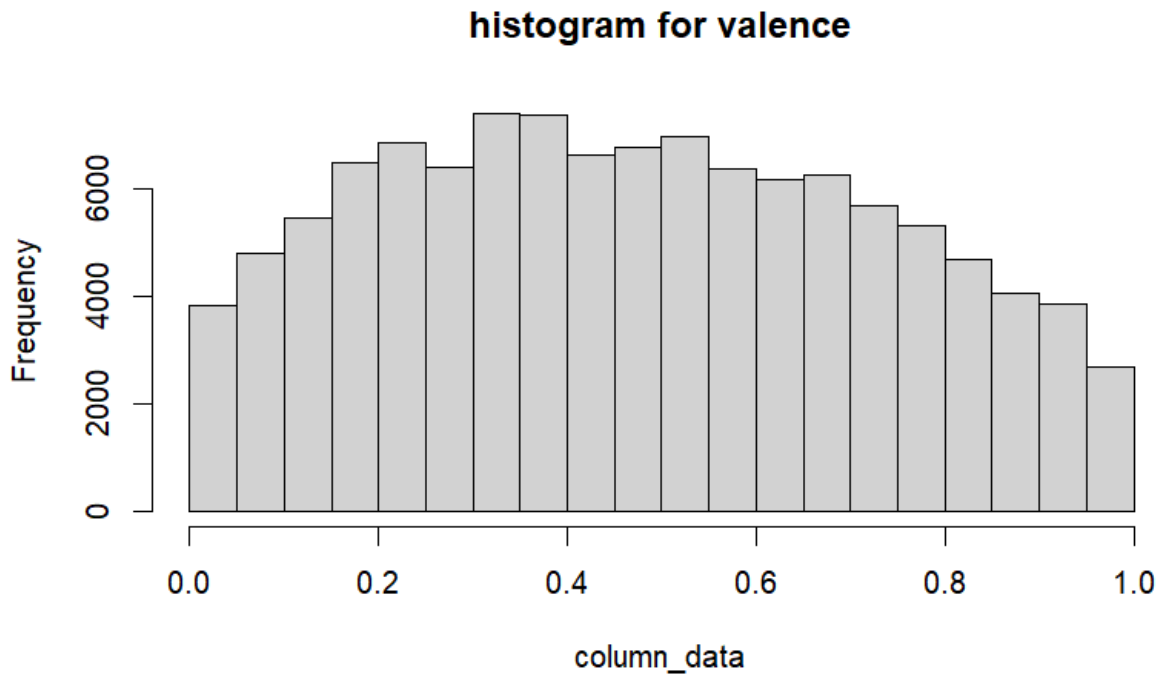


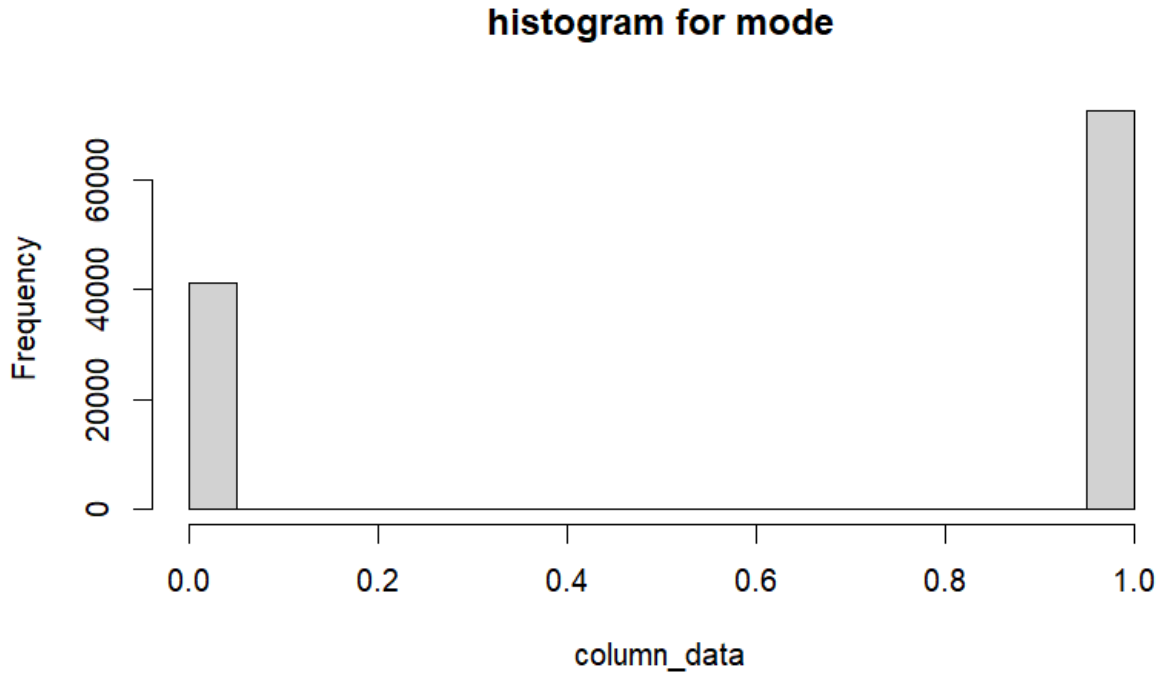

histogram for danceability
Frequency
0
4000
8000
12000
0.0
0.2
0.4
0.6
0.8
1.0
column_data

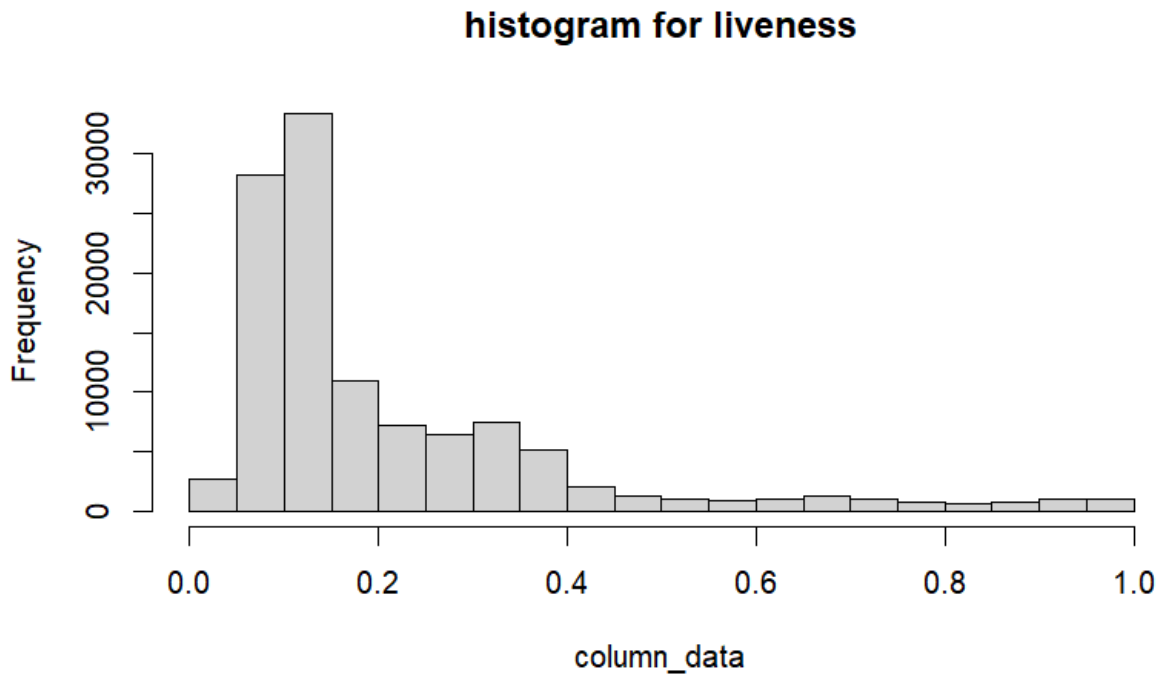

histogram for key
Frequency
0
4000
8000
12000
0
2
4
6
8
10
column_data


histogram for loudness
Frequency
0
20000
40000
-50
-40
-30
-20
-10
0
column_data


histogram for tempo
Frequency
0
5000
10000
0
50
100
150
200
250
column_data


histogram for valence
Frequency
0
2000
4000
6000
0.0
0.2
0.4
0.6
0.8
1.0
column_data


histogram for mode
Frequency
0
20000
40000
60000
0.0
0.2
0.4
0.6
0.8
1.0
column_data

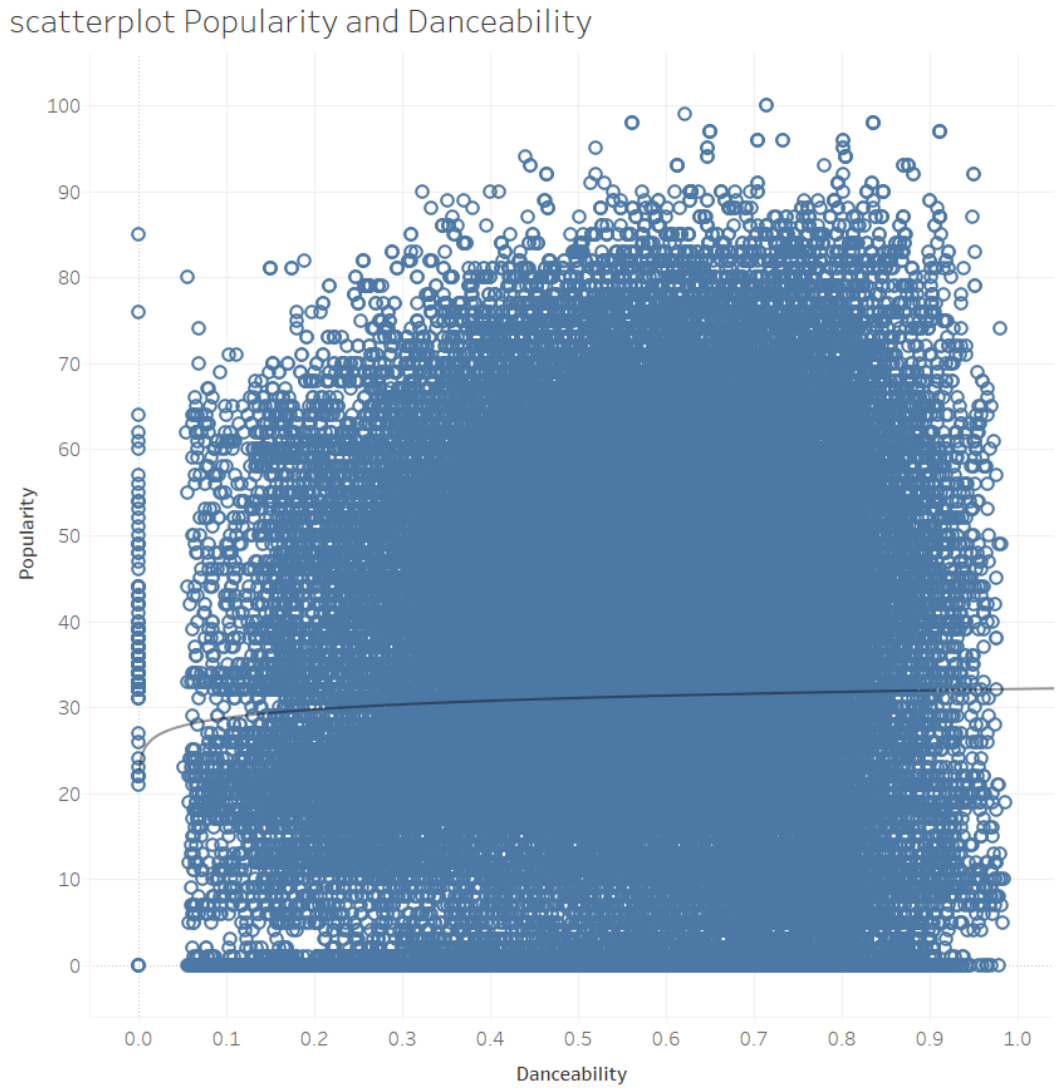

histogram for liveness
Frequency
0
10000
20000
30000
0.0
0.2
0.4
0.6
0.8
1.0
column_data

Summary Table for relationship of potential features and label “popularity”

Feature	**Analysis**	**Effect size**	**P-value**
Danceability	R squared	0.0004764	< 0.0001
Loudness	R squared	1.817e-05	0.96949
Tempo	R squared	9.926e-06	0.32442
Duration_ms	R-squared	2.878e-05	0.0931115
Liveness	R squared	6.194e-05	0.0137557
Valence	R squared	3.364e-05	0.0696928
Mode	T-stat	4.6709	3.003e-06
Key	F-stat	1.693	0.193
Track_genre	F-stat	343.5	<2e-16
Explicit	T-stat	-13.755	2.2e-16

Danceability

Hypothesis: I assume that higher danceability in a song can garner it more replays and popularity.



Danceability vs. Popularity.

Regression Equation: ln(Popularity) = 0.0477705*ln(Danceability) + 3.46925

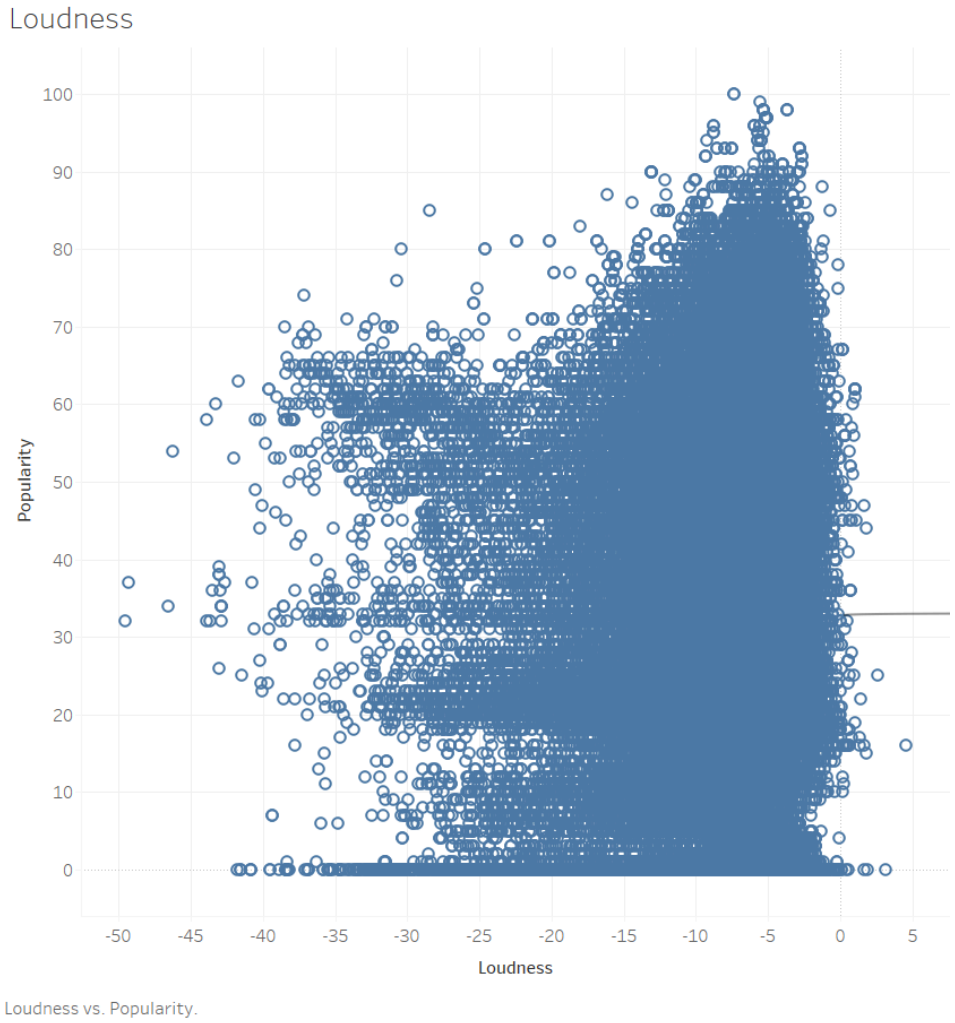
R squared: 0.0004764

P value: < 0.0001

Summary: There is an effect of danceability on popularity

Loudness

Hypothesis: I assume that loudness will not have a direct effect on popularity



Loudness vs. Popularity.

Regression Equation: ln(Popularity) = 0.00171577*ln(Loudness) + 3.49291

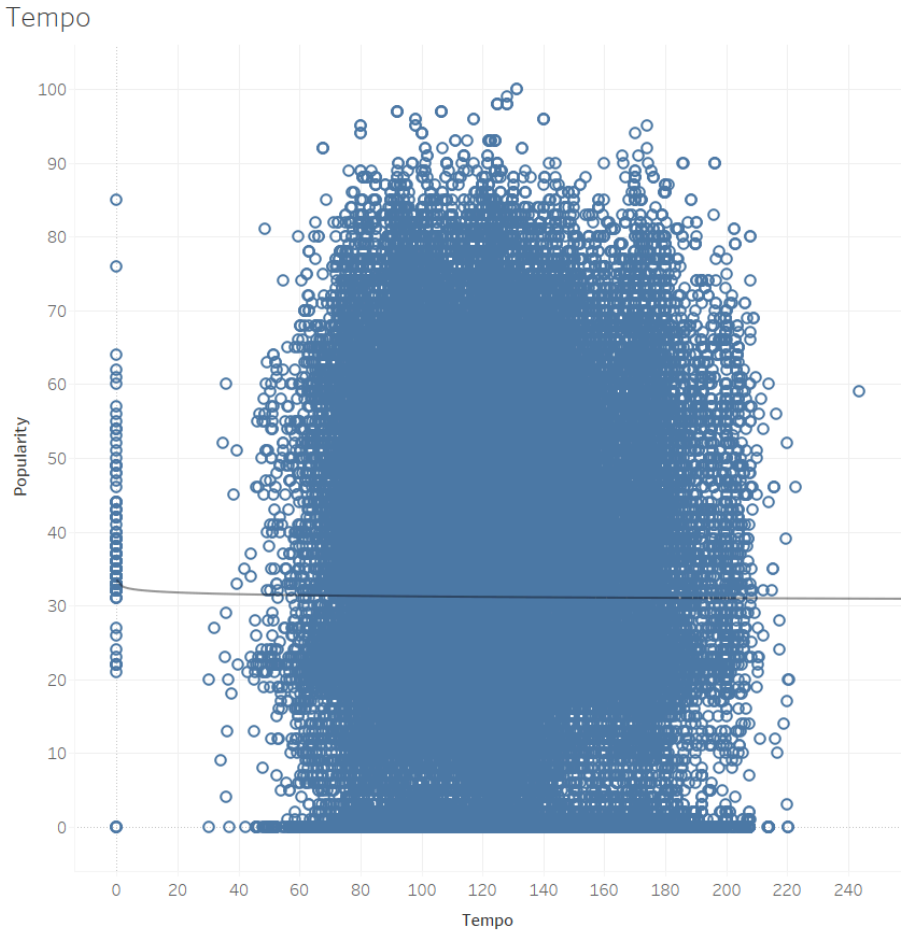
R squared: 1.817e-05

P value: 0.96949

Summary: With a p value of 0.96949, loudness does not have an effect on popularity.

Tempo

Hypothesis: I assume that the tempo of a song will not have correlation to popularity.



Tempo vs. Popularity.

Regression Equation: ln(Popularity) = -0.0105876*ln(Tempo) + 3.48986

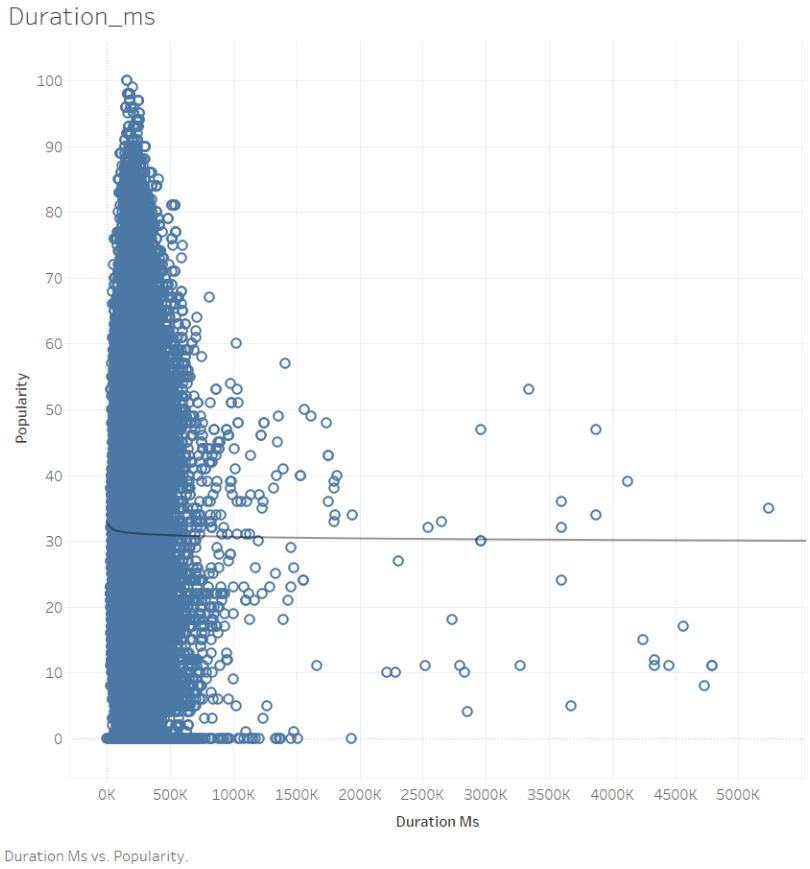
R squared: 9.926e-06

P value: 0.32442

Summary: With a p value of 0.32442, tempo does not have an effect on popularity.

Duration_ms

Hypothesis: I assume that the length of song can impact on it's popularity. If a song is too long listeners may not want to finish it multiple times.



Regression Equation:ln(Popularity) = -0.0114678*ln(Duration Ms) + 3.58025

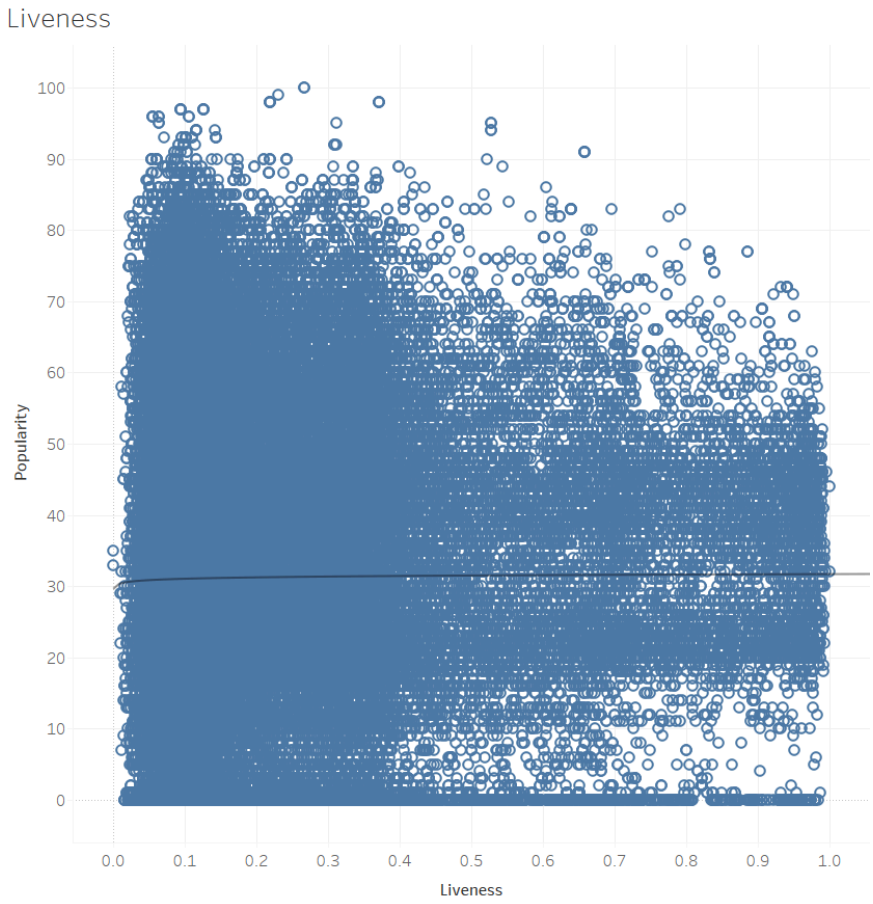
R squared: 2.878e-05

P value: 0.0931115

Summary: With a p value of 0.0931115, Duration does not have an effect on popularity.

Liveness

Hypothesis: Liveness can have an impact on the popularity of a song, as a listener may feel that they connect with the crowd that was there.



Liveness vs. Popularity.

Regression ln(Popularity) = 0.00898588*ln(Liveness) + 3.45583

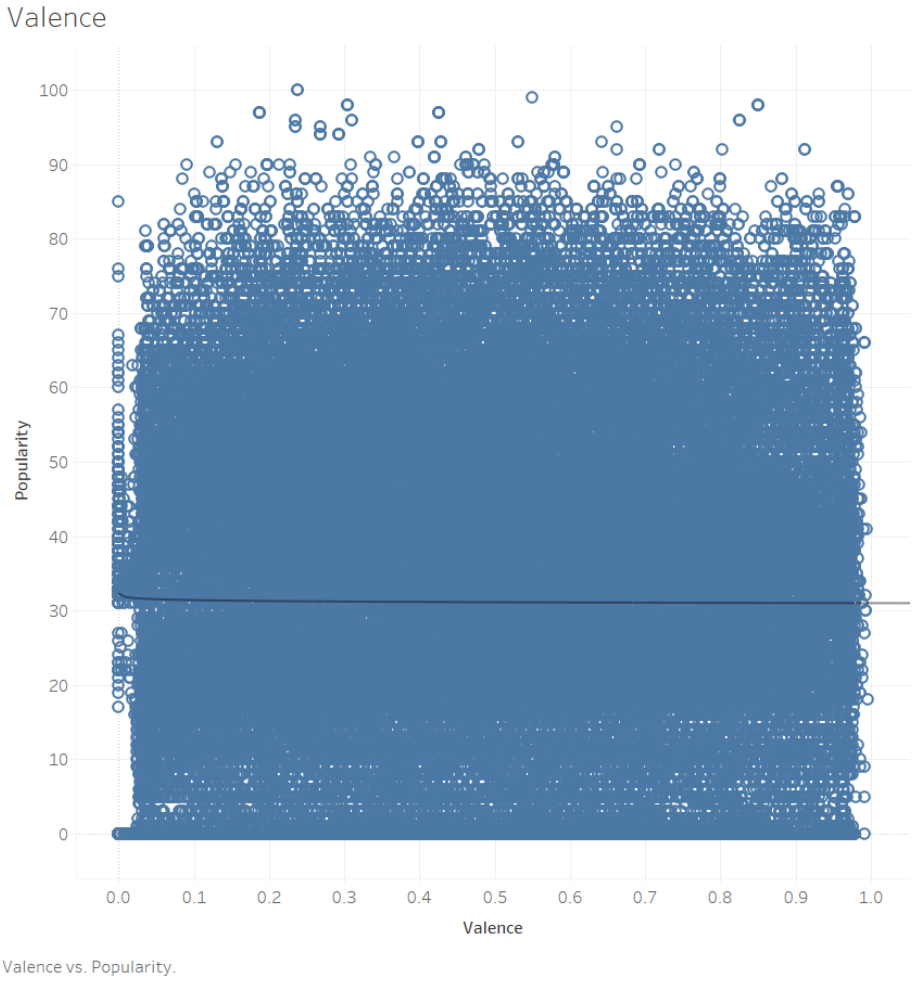
R squared: 6.194e-05

P value: 0.0137557

Summary: With a p value of 0.0137557, Liveness does have an effect on popularity.

Valence

Hypothesis: I assume that valence will not have an affect on popularity as tracks with a range of meanings have gained popularity.



Valence vs. Popularity.

Regression: ln(Popularity) = -0.00556143*ln(Valence) + 3.4337

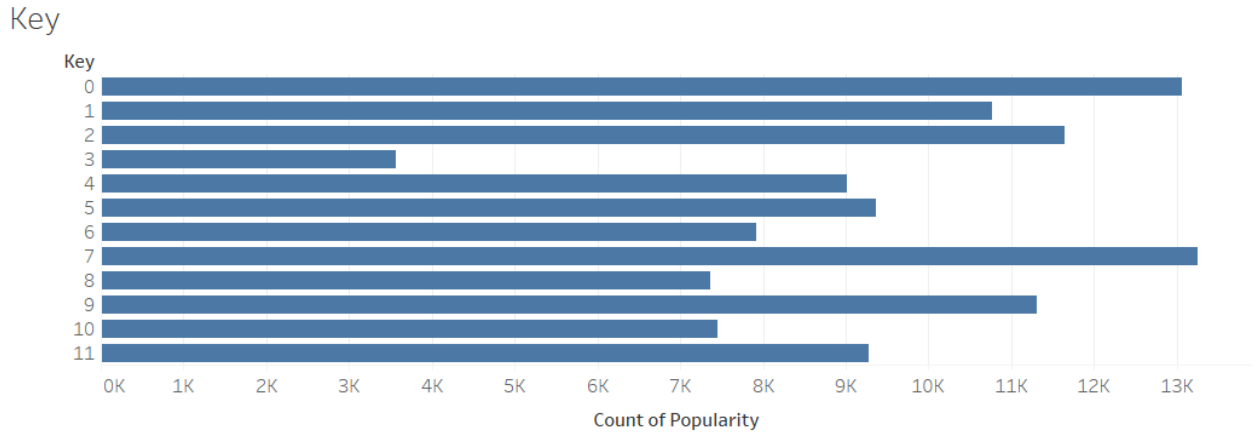
R squared: 3.364e-05

P value: 0.0696928

Summary: With a p value of 0.0696928, Valence does not have an effect on popularity.

Key

Hypothesis: I predict that the key of the song can have an effect on popularity. Some keys are easier to sing along too, and this can affect the popularity of a song.



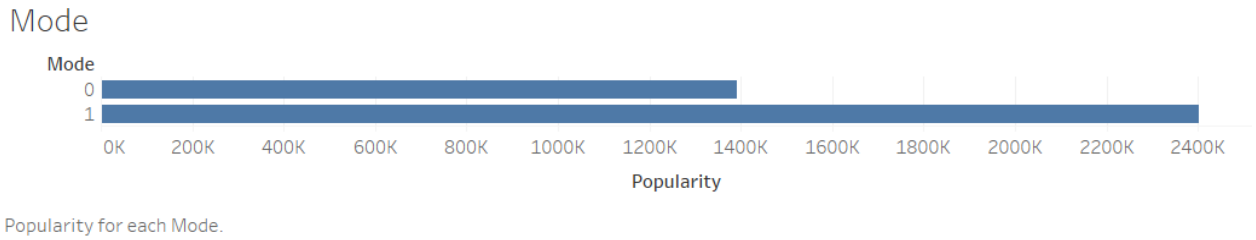
Count of Popularity for each Key.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
spotify.df$key	1	842	842.1	1.693	0.193
Residuals	113998	56715544	497.5		

Summary: With a p value of 0.193, Key does not have an effect on popularity.

Mode

Hypothesis: I predict that the mode will not have an effect on popularity as many big songs have been played in both major and minor.



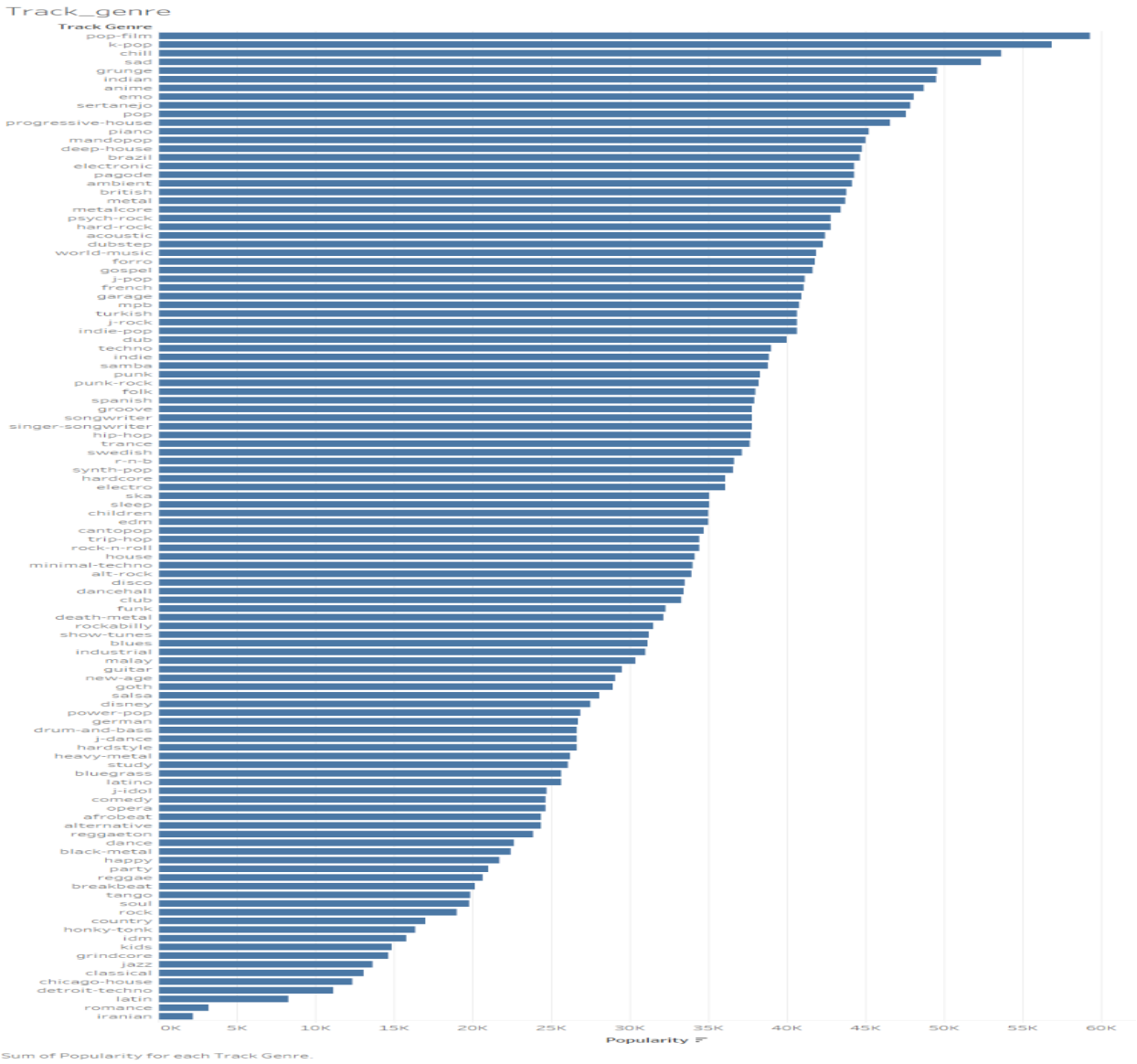
Popularity for each Mode.

```
data:  popularity by mode
t = 4.6709, df = 84062, p-value = 3.003e-06
alternative hypothesis: true difference in means between group 0 and group 1 is
not equal to 0
95 percent confidence interval:
 0.3751582 0.9176289
sample estimates:
mean in group 0 mean in group 1
       33.65064        33.00425
```

Summary: Mode does not have an effect on popularity.

Track Genre

Hypothesis: I predict that Genre will have an impact on popularity. Often the top 100 billboards have a large amount of pop-esque songs.

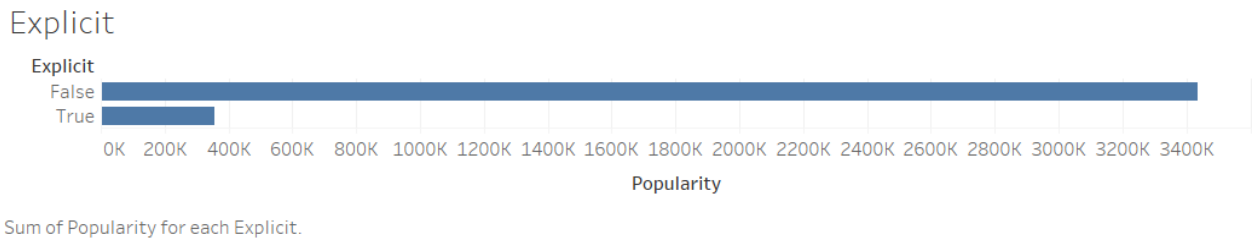



```
                          Df   Sum Sq Mean Sq F value Pr(>F)
spotify.df$track_genre    113 14415666  127572   343.5 <2e-16 ***
Residuals              113886 42300721     371
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Summary: Track_genre does indeed have an impact on popularity.

Explicit

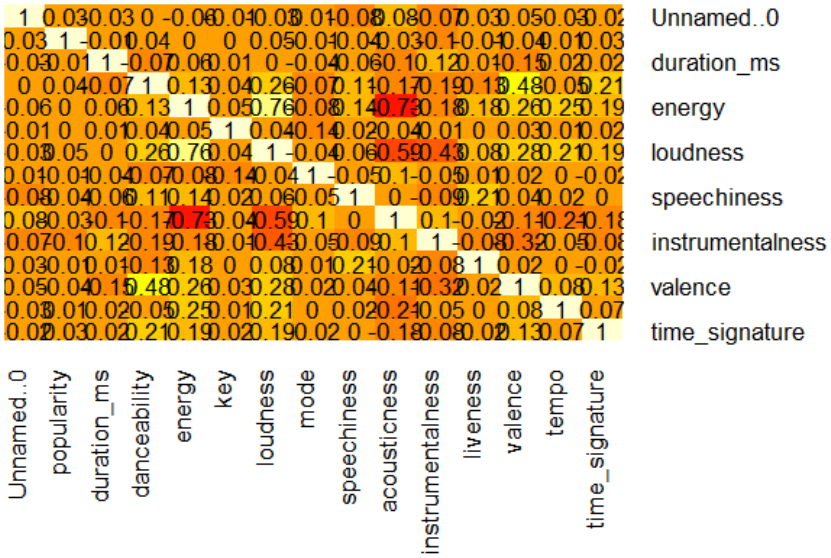
Hypothesis: I predict that the explicitness of a song can have an effect on a song's popularity as some people may avoid song's with explicitness.




```
data:  popularity by explicit
t = -13.755, df = 11301, p-value < 2.2e-16
alternative hypothesis: true difference in means between group FALSE and group
TRUE is not equal to 0
95 percent confidence interval:
 -4.017380 -3.015219
sample estimates:
mean in group FALSE  mean in group TRUE
           32.93789            36.45419
```

Summary: Explicitness does not have an effect on a song's popularity.

Correlation Matrix:



Model Testing

Feature List	Algorithm	R squared	RMSE
Track_genre, liveness, valence, tempo, key, loudness, mode, popularity, duration_ms, explicit, danceability	Boosted Decision Tree Regression	0.438507	1.043717
Track_genre, liveness, valence, loudness, mode, popularity, duration_ms, explicit, danceability	Boosted Decision Tree Regression	0.368927	1.106497
Tempo, Duration, Danceability	Boosted Decision Tree Regression	0.200468	1.245456
Key, mode, liveness, valence	Boosted Decision Tree Regression	0.183152	1.258871
Danceability, tempo	Decision forest regression	0.376572	1.099774
Danceability, tempo, valence, mode	Decision forest regression	0.42759	1.053814
Danceability, tempo, key, mode	Neural network regression	-0.00277	1.394798
Tempo, Duration, Danceability	Neural network regression	0.001734	1.391662
Key, mode, liveness, valence, loudness, explicit	Poisson Regression	0.007098	1.387918
Tempo, Duration, Danceability	Poisson Regression	0.002742	1.390959

Findings

After running a series of regression analyses to find the factors that contribute most to a song's popularity, that was measured by Root Mean Square Error and Coefficient of Determination, the results were insightful as to what factors contributed most.

- Track Genre had the highest impact on the popularity of a song. This was not a surprise as I assumed that it would have the largest impact. Billboard top 100's are typically full of pop-esque songs, and we can derive that the genre of a song will be most impactful on the popularity that it can reach.

- I found that tempo, danceability, and duration were the other largest factors in determining a song's popularity.
    - Tempo: This emerged as a crucial element in deciding a song's popularity. It can have effects on the song's danceability and energy as well.

    - Danceability: This reflects the song's to encourage movement and can highlight how an audience connects with the song.

    - Duration: Another key factor as it points to listener preference on song length. Are people willing to listen to a longer drawn out ballad or is it more in tune with the average listener's attention span?

- These insights, while interesting, are additionally nuanced, and can only help identify a few factors that play into a song's popularity. There are of course many factors that help as time of release in correlation with a world event, a song that unites people together, etc. But, these insights could still be critical for an artist in determining their market strategies.

Links

(all links were checked in an incognito browser to verify that they worked.)

Tableau:

https://public.tableau.com/app/profile/david.hohn/viz/finalforTBNALT485/Dashboard1

Azure Machine Learning Studio:

https://gallery.cortanaintelligence.com/Experiment/David-Hohn-485-Final-Project