Final Project

Majorz

Introduction and Data

With recent features on music apps such as Spotify Wrapped gaining massive popularity, understanding users' music taste for personalized recommendations and music trend analysis have become a critical challenge for streaming companies. To categorize and analyze the countless songs on these platforms, each are dissected into various musical elements ranging from duration and tempo to loudness and danceability. Using a real database of song tracks compiled and released by Spotify for data engineering purposes, we wanted to see whether common trends could be observed between different musical elements. Modes of songs, specifically, were of our interest since they determine the mood of the music — songs in major modes sound more bright and uplifting while those in minor modes are more calm and even sadder. We wanted to explore if musical aspects such as bounciness or tempo would be correlated to the song's mode in some way, with some of our example hypotheses being that minor songs would be slower and/or less danceable but more acoustic than major songs. Hence, we set the following: *How do different musical elements affect whether a song is in major or minor mode?*

This data was collected from the Spotify for Developers website, as the data set was published to be used as part of an open data science challenge. With no null values and well-categorized variables, our data was already cleaned and ready to be used for a complete case analysis. Minor data cleaning processes that we conducted were deleting irrelevant variables such as acoustic vectors and adding a new variable "new_mode" to express major and minor modes numerically as 1 and 0. Key variables included:

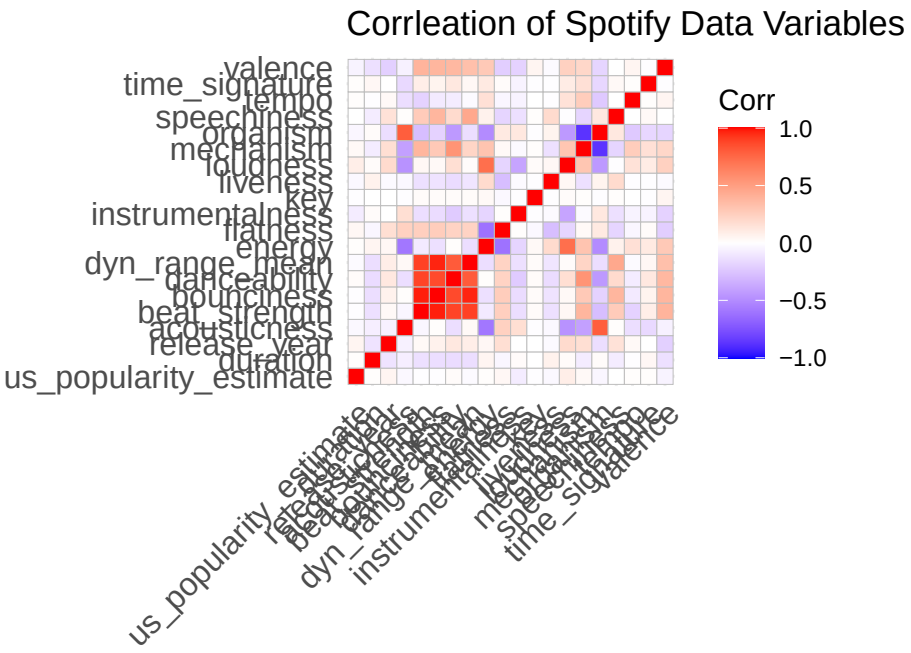
release_year: year of song released (1950-2018)	**key**: song key starting from C major (0) to B minor (11)	**mode**: song mode (major or minor)
new_mode: song mode numerized (1 = major, 0 = minor)	**tempo**: speed of song in beats per minute (bpm)	**time signature**: number of quarter notes in each measure

To get a gist of what our data was presenting, we fitted an initial logistic model using all variables as predictors.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	32.2683808337	2.3096692648	13.970996	2.343254e-44
us_popularity_estimate	-0.0112940994	0.0085641669	-1.318762	1.872486e-01
duration	-0.0008867527	0.0001370154	-6.471921	9.676459e-11
release_year	-0.0145826354	0.0010561589	-13.807236	2.305088e-43
acousticness	0.4800550240	0.1339125117	3.584841	3.372840e-04
beat_strength	2.3227248811	0.3798219868	6.115299	9.637630e-10
bounciness	-4.2116774008	0.5087117132	-8.279104	1.241051e-16
danceability	0.2508033040	0.1611182274	1.556641	1.195556e-01
dyn_range_mean	0.1188408770	0.0200061888	5.940206	2.846646e-09
energy	-0.5804580033	0.1072093530	-5.414248	6.154688e-08
flatness	0.7082199988	0.3348900418	2.114784	3.444839e-02
instrumentalness	-0.3421403175	0.0522757477	-6.544915	5.952929e-11
key	-0.0930591757	0.0026792613	-34.733146	2.490369e-264
liveness	0.3261005480	0.0588139117	5.544616	2.946002e-08
loudness	0.0223914242	0.0043965647	5.092936	3.525602e-07
mechanism	-0.8263282243	0.2122943065	-3.892371	9.926924e-05
organism	-0.3927747758	0.3168699890	-1.239546	2.151435e-01
speechiness	-1.0627013487	0.0967582511	-10.983057	4.610547e-28
tempo	0.0027562879	0.0004503870	6.119822	9.368016e-10
time_signature	-0.2081995314	0.0260102614	-8.004515	1.199383e-15
valence	0.5394631103	0.0506271725	10.655604	1.641752e-26

As demonstrated by the regression model above, there are many predictors that are statistically significant, using the significance level of $\alpha = 0.5$. However, it is critical to improve this baseline model in the following ways:

1) Confirm that there are not instances of multicollinearity (or model overfitting)
2) Ensure that the variables included are meaningfully contributing to the model
3) Optimize the model and determine if tranformations or changes are appropriate



Examining the correlation plot above, it appears there are variables that have a high positive correlation with each other. This causes great concern with multicollinearity as the model may be overfitted. For example,

- beat_strength is highly correlated with, dyn_range_mean, danceability, and bounciness
- mechanism is highly correlated with organism

Therefore, to prevent overfitting in our regression model, the following variables should be removed - beat_strength, dyn_range_mean, bounciness, and organism. However, we decided to leave in danceability and mechanism because we felt that these variables are easily understandable from a musical perspective and may be important to the model. (A revised correlation plot can be found in the appendix)

The new model:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	34.1621529993	2.2527143592	15.1648845	6.040510e-52
us_popularity_estimate	-0.0108458434	0.0085519132	-1.2682359	2.047137e-01
duration	-0.0009130731	0.0001366947	-6.6796540	2.395070e-11
release_year	-0.0151253508	0.0010349401	-14.6147115	2.263109e-48
acousticness	0.2930254311	0.0467428402	6.2688837	3.636455e-10
danceability	-0.5424777088	0.0965306501	-5.6197457	1.912387e-08
energy	-0.6337572541	0.1062485650	-5.9648547	2.448518e-09
flatness	0.1427312789	0.3270744094	0.4363878	6.625554e-01

```
instrumentalness       -0.3767551137 0.0505993590  -7.4458476  9.632399e-14
key                    -0.0928740917 0.0026767442 -34.6966634 8.846084e-264
liveness                0.3344805774 0.0585436948   5.7133493  1.107740e-08
loudness                0.0237143417 0.0043676212   5.4295784  5.648732e-08
mechanism              -0.3058568379 0.0704087463  -4.3440177  1.399003e-05
speechiness            -1.2847824470 0.0851015092 -15.0970583  1.693205e-51
tempo                   0.0017839875 0.0003600327   4.9550708  7.230395e-07
time_signature         -0.2086805576 0.0258252158  -8.0804962  6.450378e-16
valence                 0.4529713111 0.0482846337   9.3812726  6.518118e-21
```

Removing the highly related variables were essential to our analysis as some of the coefficients changed drastically, including changing signs (eg: danceability changed from a positive to negative contribution)! Additionally, the variable flatness is no longer significant in the model (at at 0.05 significance level).

In addition to removing removing the highly correlated variables, we felt it was also important to select variables that have the most impact on the model. For example, some variables may be not meaningful by nature to the outcome of interest; therefore, removal is essential. In this analysis, we decided to use a LASSO model to select variables.

```
17 x 1 sparse Matrix of class "dgCMatrix"
                                s0
(Intercept)              .
us_popularity_estimate -0.0021045272
duration               -0.0001855370
release_year           -0.0028371270
acousticness            0.0605092341
danceability           -0.1172593977
energy                 -0.1278463323
flatness                0.0209872015
instrumentalness       -0.0834989566
key                    -0.0204292745
liveness                0.0694862332
loudness                0.0046850947
mechanism              -0.0658522490
speechiness            -0.2908906556
tempo                   0.0003684788
time_signature         -0.0407137802
valence                 0.0998470360
```

LASSO kept all of the predictors, demonstrating that the predictor variables are meaningfully contributing to our outcome of interest of whether the song is on a major/minor scale. It is

important to note that LASSO does not include an intercept, as the model is centered. Since all variables are retained in the LASSO model, we decided to use the **standard logistic model** instead of the LASSO model for further analysis because LASSO adjusts coefficients (due to the intercetp) and for ease of interpretation.

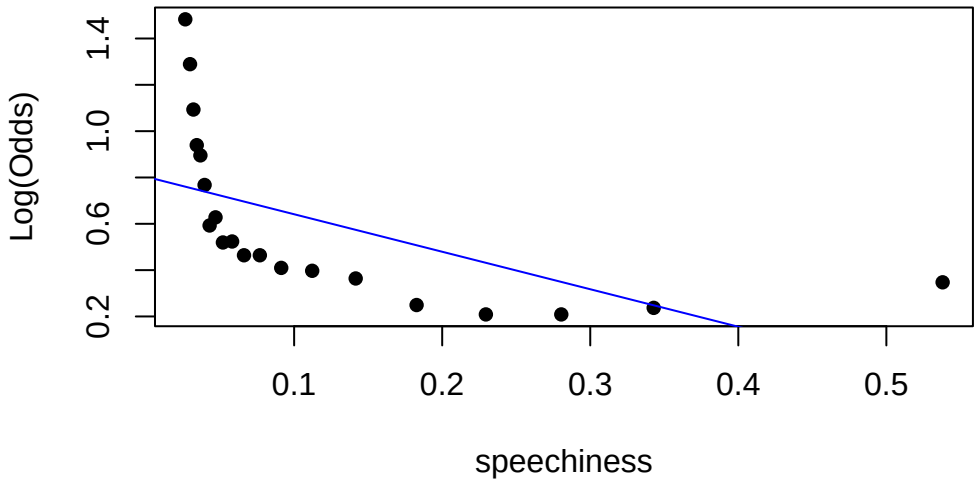
Methodology

In order to ensure our model can be interpreted in a real-world context, it is critical to check all of the assumptions for a logistic model. For logistic regression, the two most important assumptions are independence and linearity. For independence, we are checking to see if each observation in our data is independent from each other (eg: knowing about one observation does not tell us about another). On the other hand, linearity for logistic models ensures that the predictor variables generally follow a linear trend with the odds of the outcome of interest. There should not be any clear patterns or distinct trends within the data.

Independence:

The independence assumption is accepted because the observations are independent from each other. Knowing something about one song doesn't impact what we know about another song. Additionally, the data comes from 130 million users which is an extremely large amount. While the songs that a singular person listens to may be similar, we don't think that violates our independence here because there is so many different songs selected from these listeners.

Linearity:



There were fewer data points for some of the predictors because there was only so many different values and enough of them to be able to get the empirical logits. For example, with key there is only 12 unique values, but not all of them had enough values to be calculated, so we did 10 groups. I eliminated the titles to make the plots more clear and because they were repetitive. In summary, we concluded that linearity is met for time signature, tempo, mechanism, loudness, liveness, instrumentalness, key, release year and popularity because there is no major pattern in empirical logits. Linearity was not met for valence, speechiness, organism, flatness, energy, danceability, acousticness and duration because they showed patterns in empirical logits.

These are potential limitations of these variables that do not meet the linearity assumption. One issue that was tricky is attempting to transform the variables that had underlying trends. Unfortunately, we were unable to make a substantial impact, especially on the speechiness variable. However, upon further research, we discovered that the variable is impacted strongly by songs that are instrumental, as they cause a cluster of points around 0 (demonstrated on the empirical logit plot). Therefore, it is critical to understand that there may be some linearity concerns when it comes to the overall view of our model.

```
          0     1
Major 14992 30110
Minor  3157  2445
```

Using our logistic regression model as a classifier for any infection by using a threshold of 0.5 predicted probability, we are able to calculate the following values. These allow us to judge how well the outcome "classifier" does in terms of the model.

Prevalence: (30110+2445)/(30110+2445+14992+3157) = 0.642

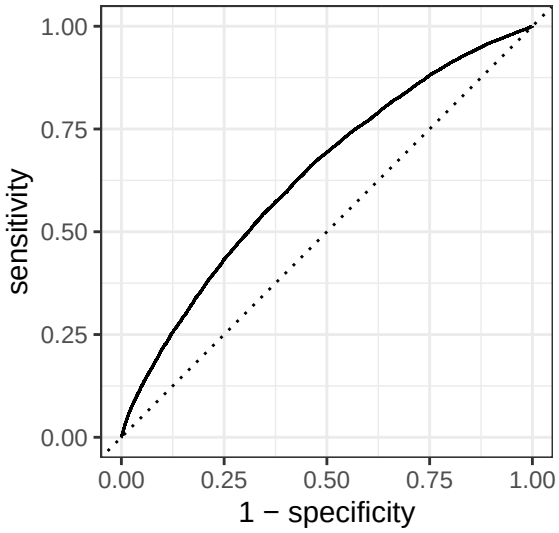
Sensitivity: 30110/(30110 + 2445) = 0.925

Specificity: 3157/(3157 + 14992) = 0.174

Positive predictive value: 30110/(30110 + 14992) = 0.667

Negative predictive value: 3157/(3157 + 2445) = 0.564

Immediately, it is clear that there is a very high sensitivity and low specificity. This means that the model (at the 0.5 threshold) may determine that a song is in a major key while in reality it doesn't. On the flip side, it also means that there are minimal songs that are considered to be in a minor key when in fact it is actually in a major key. Either way, there is a clear imbalance. Additionally, we can see that the positive and negative predictive values are not extremely high, meaning that it is fairly likely that if it is considered major/minor key, that it is truly in a major/minor key. These probabilities are not as high as expected, but they indicate that it does have downfalls.




```
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 roc_auc binary         0.637
```

The value of the area under the curve of the ROC curve is 0.64. Although it is greater than 0.5, which would imply that it would be just as effective to guess the major/minor scale, it is not as high as expected. As discussed below, there may be several reasons that this occurs in our model.

Results

One predictor that we were interested along with our outcome variable (major/minor scale) is popularity. It is a bit tough to look at some of our variables because a lot of them are scaled by tenths increases rather than whole numbers. Popularity however ranges from about 90-100 so it includes some whole number changes. When holding all other predictors constant, we for every one number increase in popularity, the odds of the song being major is predicted to be multiplied by $e^{-0.01085} = 0.989$. While we were interested in looking at popularity, the p-value is not significant. Another possible interesting predictor is danceability which is essentially looking at the song's ability to be danced to. When holding all other predictors constant, we for every one number increase in danceability, the odds of the song being major is predicted to be multiplied by $e^{-0.542} = 0.582$. This one is harder to interpret because danceability ranges from 0 to 1. This is one limitation of our interpretations because many of the scales are not with whole numbers.

Discussion

*********TALK ABOUT WHAT WE LEARNED!!!!!!********** In conclusion, this model has benefits and shortfalls. Primarily, it is clear that there are variables that do not meet the linearity assumption and create difficulties for interpretation. For example, the variable speechiness, follows a distinct pattern. However, attempts to transform these variable were not effective because of underlying issues.

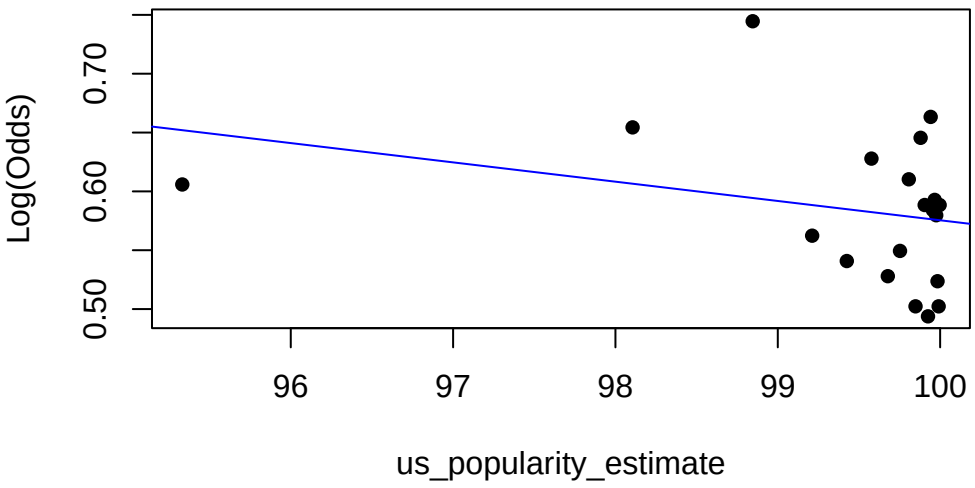
Additionally, there are challenges with some of the variables in terms of their scaling and units. For example, the variable us_popularity_estimate mostly takes on values from 97-99. Each variable is different, but they generally have unique scaling.

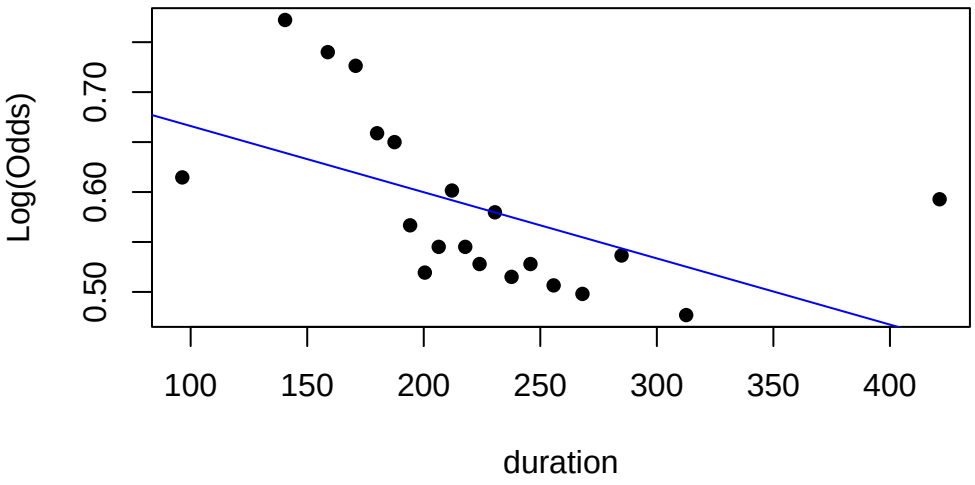
In all, our model does have downfalls, but it does have an interpretive aspect that is desirable. This allows the results to be more "reasonable" in terms of predicting if a song is in a major/minor key. The model, even though there are issues, is not extremely sophisticated or complex for a general audience. Even though the AUC value is not as strong as desirable, it is still an informative model. Generally, it provides insightful and meaningful results while simultaneously maintaining a real-world aspect.

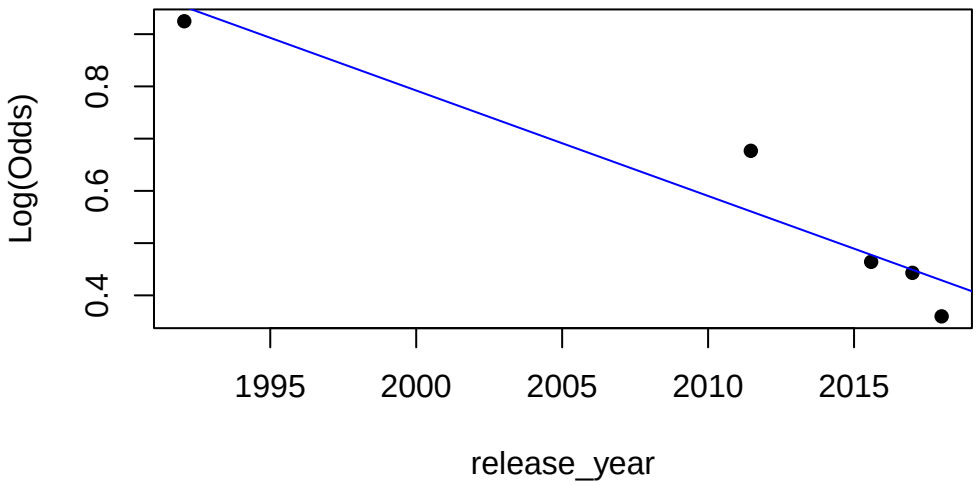
In the future, we may want to explore other data sources and outcomes to understand the media market better. For example we could compare this data from Spotify with songs played on radio stations from the 1950's to current day. Would the popularity of songs on Spotify correspond with songs frequently played on the radio? There are many questions outside of our project scope that could be answered with further research and models.

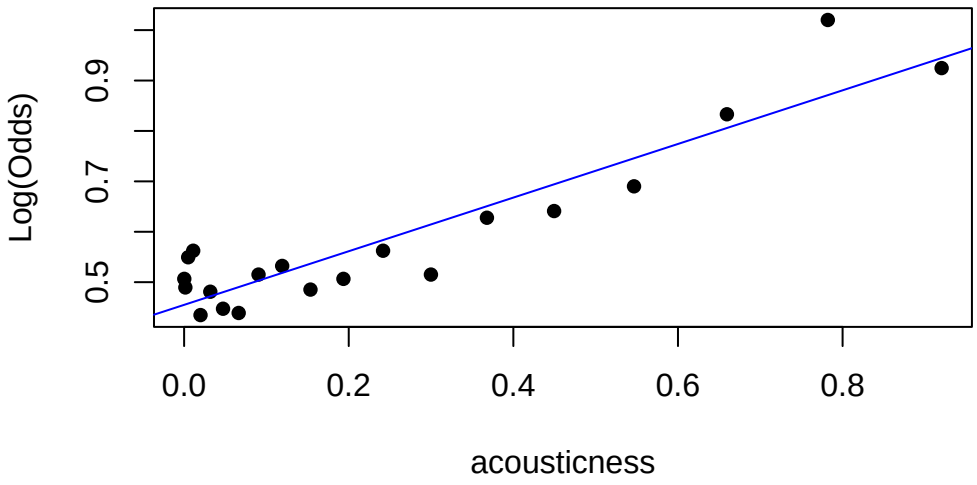
Appendix

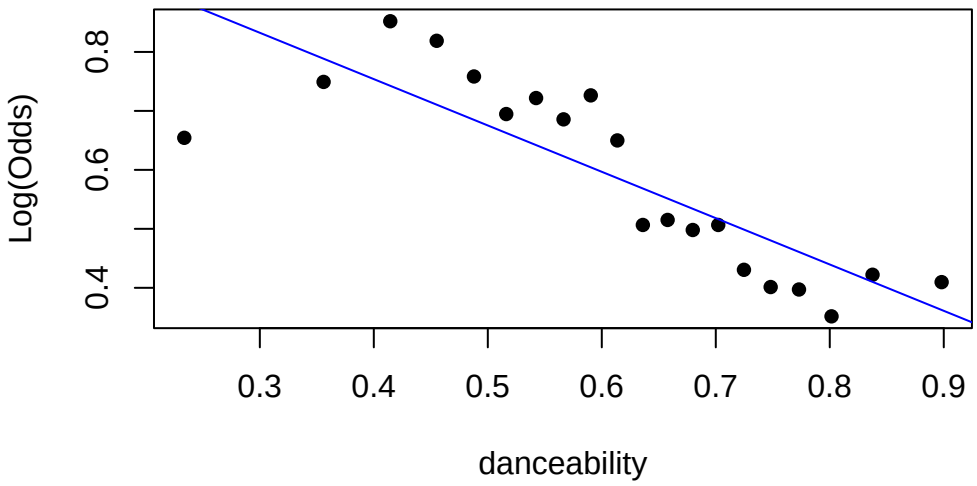
Following are the empirical logit plots as referenced in the methodology section above.

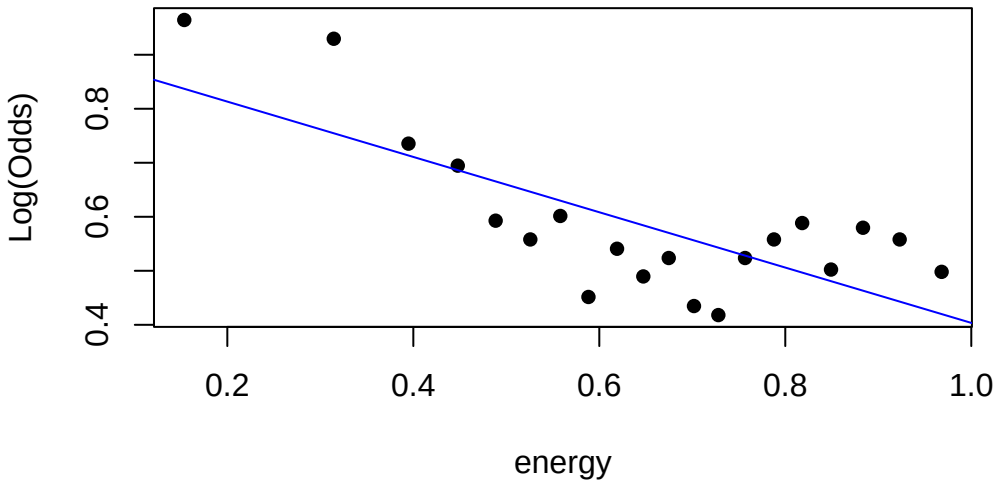

Log(Odds)
0.50
0.60
0.70
96
97
98
99
100
us_popularity_estimate

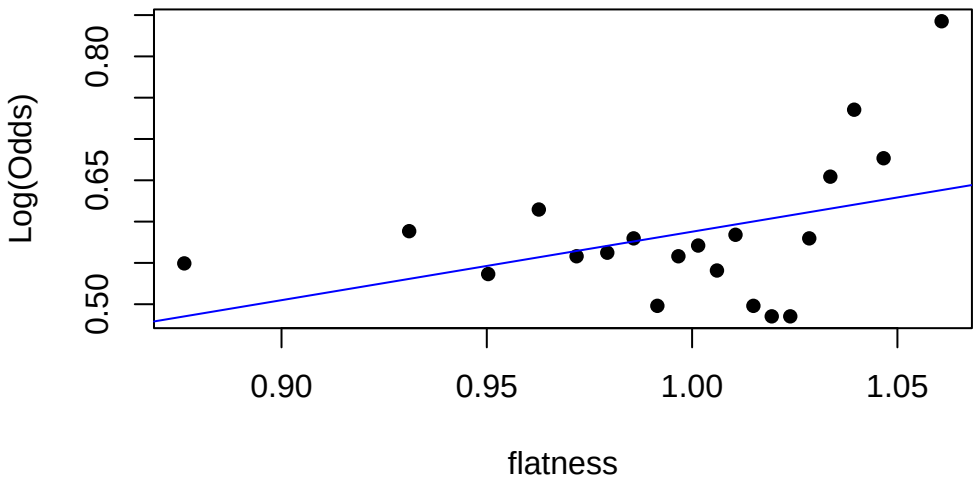

Log(Odds)
0.50
0.60
0.70
100
150
200
250
300
350
400
duration

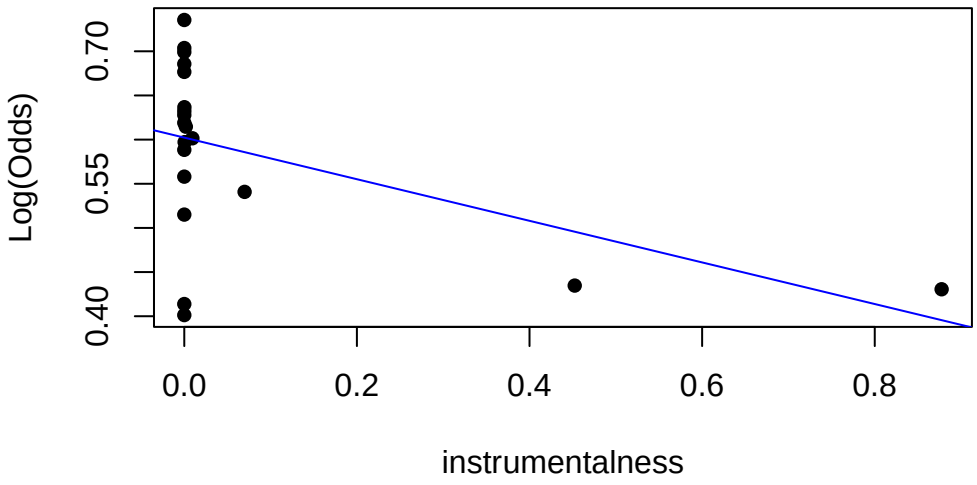

Log(Odds)
0.4
0.6
0.8
1995
2000
2005
2010
2015
release_year

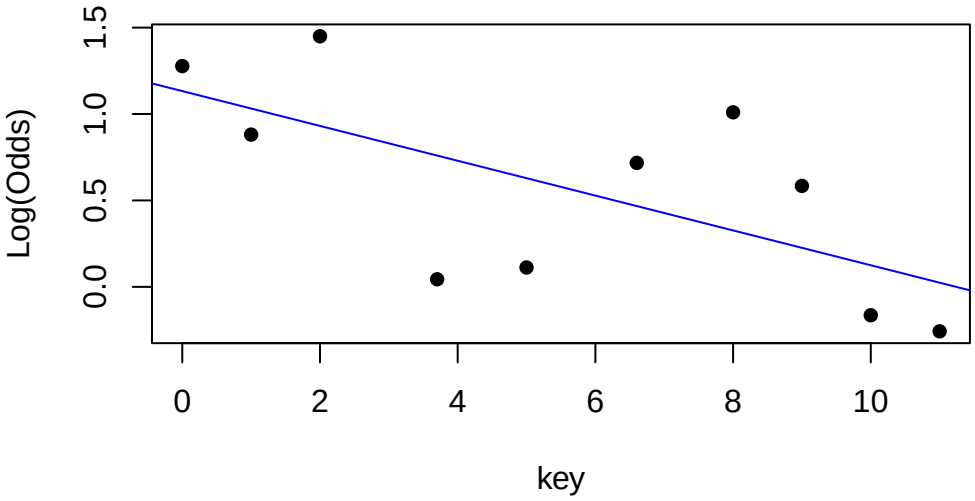

Log(Odds)
0.5
0.7
0.9
0.0
0.2
0.4
0.6
0.8
acousticness

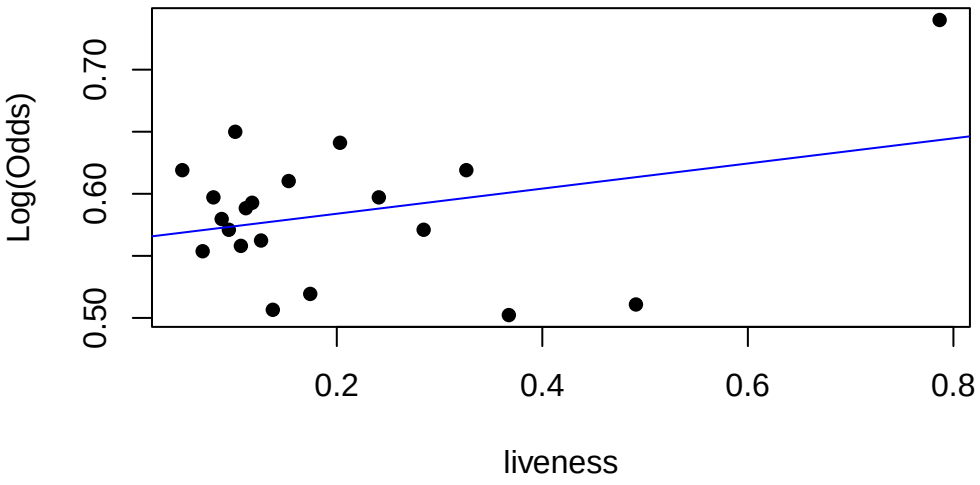

Log(Odds)
0.4
0.6
0.8
0.3
0.4
0.5
0.6
0.7
0.8
0.9
danceability

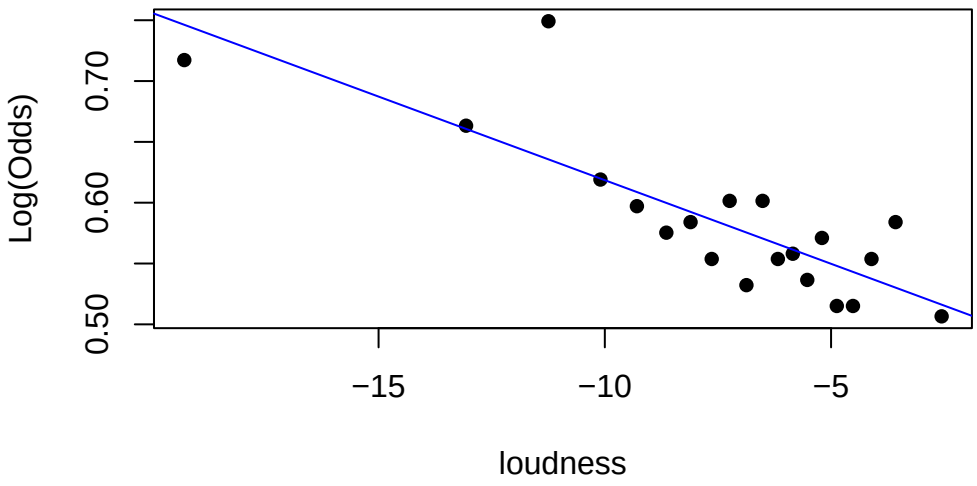

Log(Odds)
0.4
0.6
0.8
0.2
0.4
0.6
0.8
1.0
energy

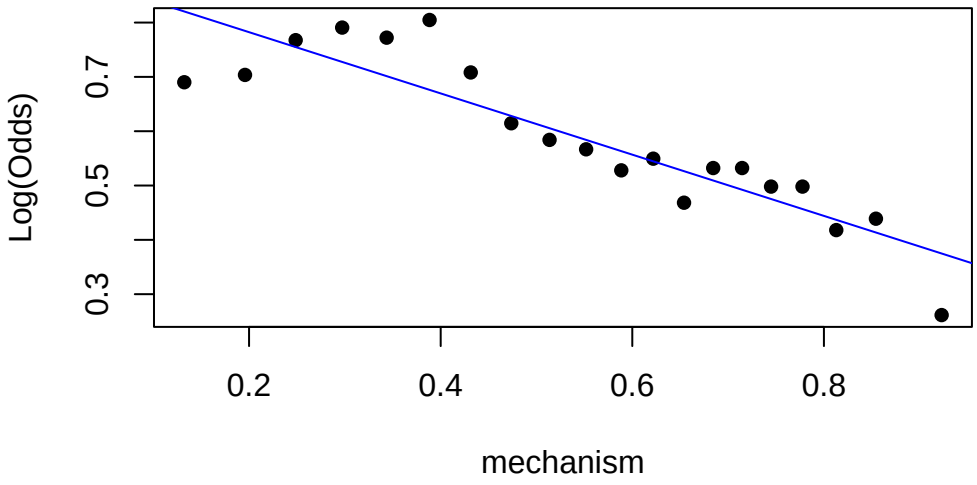

Log(Odds)
0.50
0.65
0.80
0.90
0.95
1.00
1.05
flatness

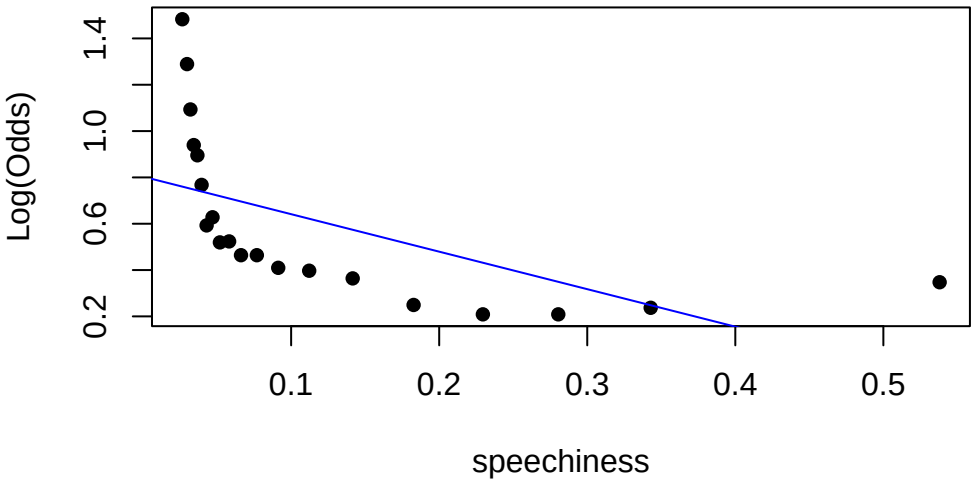

Log(Odds)
0.40
0.55
0.70
0.0
0.2
0.4
0.6
0.8
instrumentalness

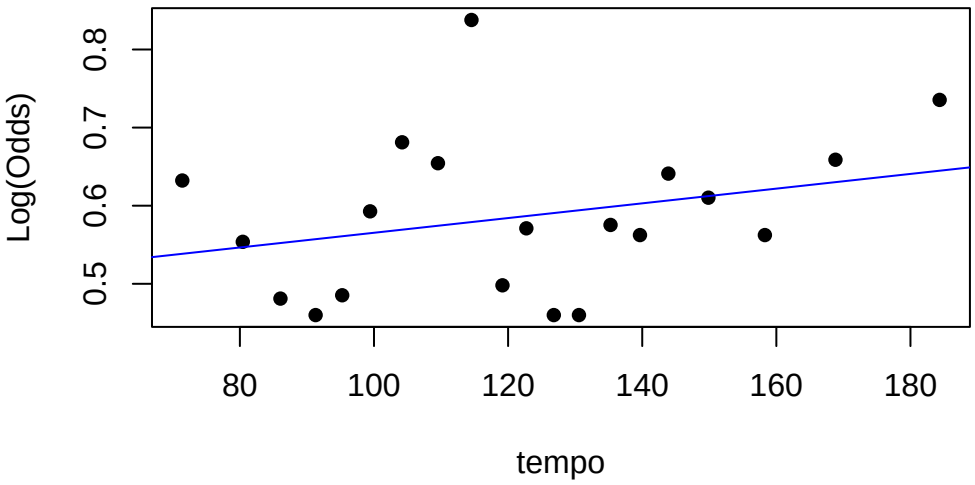

Log(Odds)
0.0
0.5
1.0
1.5
0
2
4
6
8
10
key

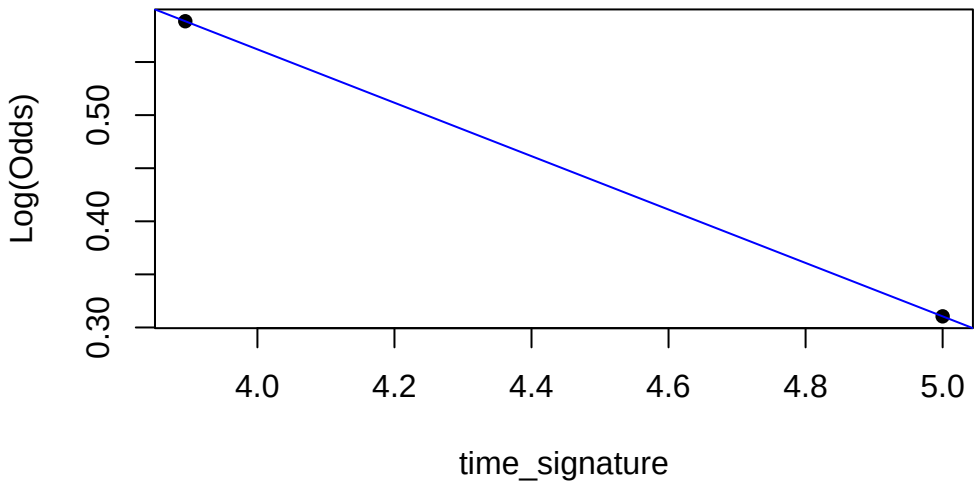

Log(Odds)
0.50
0.60
0.70
0.2
0.4
0.6
0.8
liveness

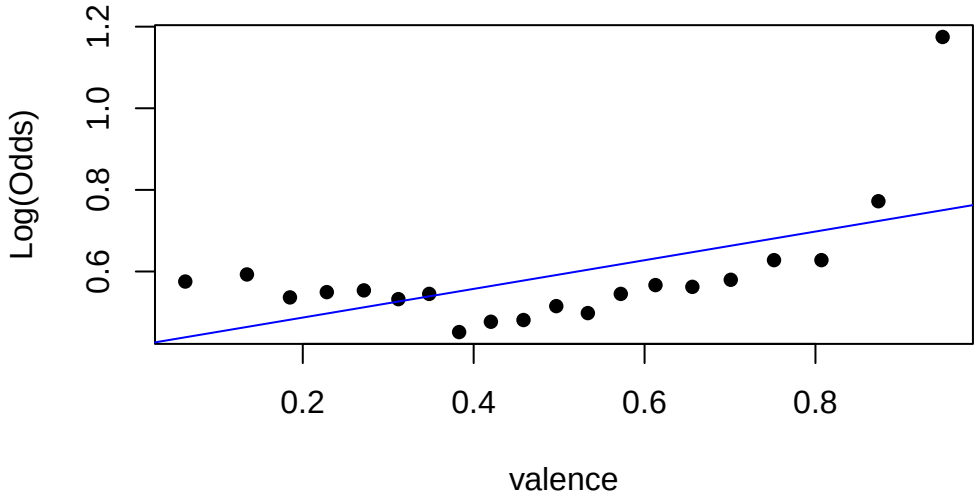

Log(Odds)
0.50
0.60
0.70
−15
−10
−5
loudness


Log(Odds)
0.3
0.5
0.7
0.2
0.4
0.6
0.8
mechanism

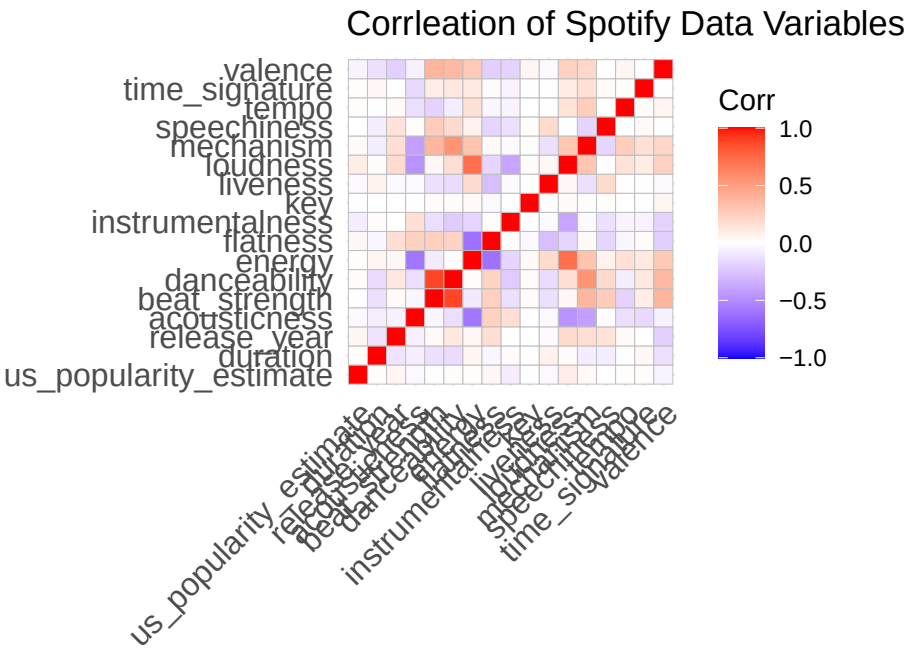

Log(Odds)
0.2
0.6
1.0
1.4
0.1
0.2
0.3
0.4
0.5
speechiness


Log(Odds)
0.5
0.6
0.7
0.8
80
100
120
140
160
180
tempo





New correlation plot after variables have been removed:



Citations

Spotify data:

https://www.aicrowd.com/challenges/spotify-sequential-skip-prediction-challenge/dataset_files (need to create an account and log in to access the dataset)

Referenced for completing the correlation matrix:

http://www.sthda.com/english/wiki/ggcorrplot-visualization-of-a-correlation-matrix-using-ggplot2#:~:text=The%20easiest%20way%20to%20visualize,ggcorr()%20in%20ggally%20package