21.05.2024

Dominik Czech 100529977

Case study: Bayesian regression for hit prediction based on Nirvana's discography and Spotify audio features.

Objective:

Find out what attributes make a grunge song a hit based on audio data stored by Spotify and Bayesian logistic regression models. Predict whether a song will be a hit, based on its audio features.

Data:

The data stored in the ***nirvana_discography.csv*** contains the audio features available through Spotify API for each song of Nirvana's 4 studio albums: *Bleach (1989), Nevermind (1991), Incesticide (1992)* and *In Utero (1993)*.

The dataset contains the **metadata** of a song:

- **track_id** - id of a song used to gather the audio features using Spotify API
- **song_name** - name of the song
- **album_name** - name of the album from which the song comes
- **album_release_date** - release date of the album
- **popularity** - the value of popularity of the song (between 0 and 100)
- **explicit** - whether or not song contains explicit lyrics
- **streams -** the number of times song has been played on Spotify (data for 21.05.024 16:11)

Audio features consist of the following categories:

- **Danceability**: Measures how suitable a song is to dancing on a scale from 0.0 to 1.0
- **Energy**: Measures how energetic a song is on a scale from 0.0 to 1.0. Energetic songs feel fast, loud and noisy.
- **Key:** The key the track is in, using standard Pitch Class notation (e.g., 0 = C, 1 = C♯/D♭, 2 = D, and so on). The value ranges from -1 to 11.
- **Loudness**: The overall loudness of a track in decibels (dB), averaged across the entire track. Values typically range between -60 and 0 dB.
- **Mode**: Indicates the modality of a track, with major represented by 1 and minor by 0.
- **Speechiness**: Detects the presence of spoken words in a track. Values above 0.66 likely indicate tracks made entirely of spoken words, 0.33 to 0.66 indicate tracks with both music and speech, and below 0.33 most likely represent music.
- **Acousticness**: A confidence measure from 0.0 to 1.0 of whether the track is acoustic. Higher values indicate higher confidence.
- **Instrumentalness:** Predicts whether a track contains no vocals. Values above 0.5 suggest instrumental tracks, with higher values indicating greater likelihood.
- **Liveness:** Detects the presence of an audience in the recording. Values above 0.8 indicate a strong likelihood that the track is live. Values range from 0.0 to 1.0.
- **Valence**: Measures the musical positiveness conveyed by a track. Higher values sound more positive (happy, cheerful), while lower values sound more negative (sad, angry). Values range from 0.0 to 1.0.
- **Tempo**: The speed or pace of a track, measured in beats per minute (BPM).
- **Time Signature:** The time signature of the track, indicating how many beats are in each bar. The value ranges from 3 to 7.

Methodology:

1. Data Preprocessing

Let's load our data into RStudio.

```
data <- read.csv('nirvana_discography.csv', sep =';')
attach(data)
```

```
names(data)
 [1] "track_id"             "song_name"            "album_name"
"album_release_date" "popularity"
 [6] "explicit"             "danceability"         "energy"
"key"                "loudness"
[11] "mode"                 "speechiness"          "acousticness"
"instrumentalness"   "liveness"
[16] "valence"              "tempo"
"time_signature"     "streams"
```

We need a way to describe if a song is a hit or not. In this case study we will base it on the **number of streams** (#of times a song has been played). The popularity statistic is not good for that purpose, it is based on various variables, not solely on the number of streams.

Therefore we will define a following variable:

```
bin.hit=ifelse(data$streams>150000000, 1, 0)
```

Meaning that a song is considered a hit if it has more than 150 mln streams. Here are the songs that will be considered as hits (top 11 Nirvana songs by Spotify streams in their studio discography).

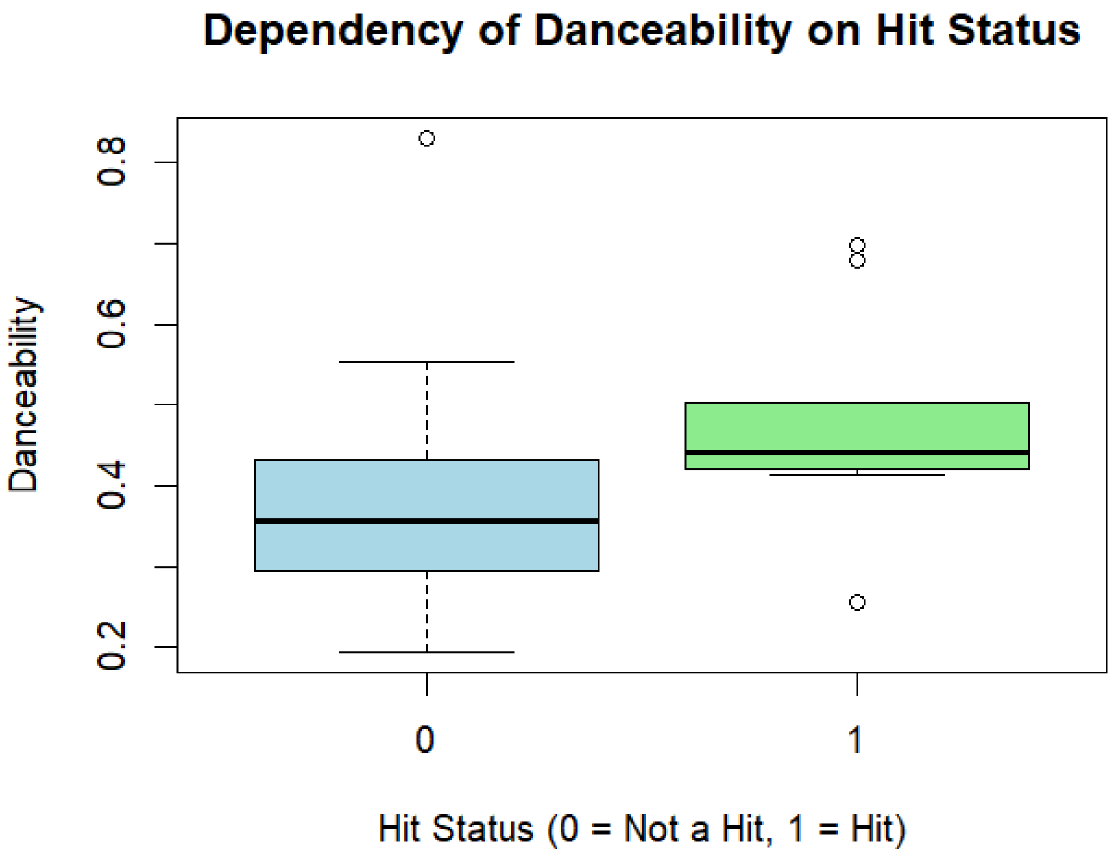
```
song_name[bin.hit==1]
 [1] "About A Girl    "Smells   Like   Teen   Spirit"   "In   Bloom"
"Come As You Are"
 [5] "Breed      "Lithium   "Something In The Way"      ″Heart-Shaped
Box"
 [9] "Rape Me    "Dumb"      All Apologies"
```

2. Data visualisation

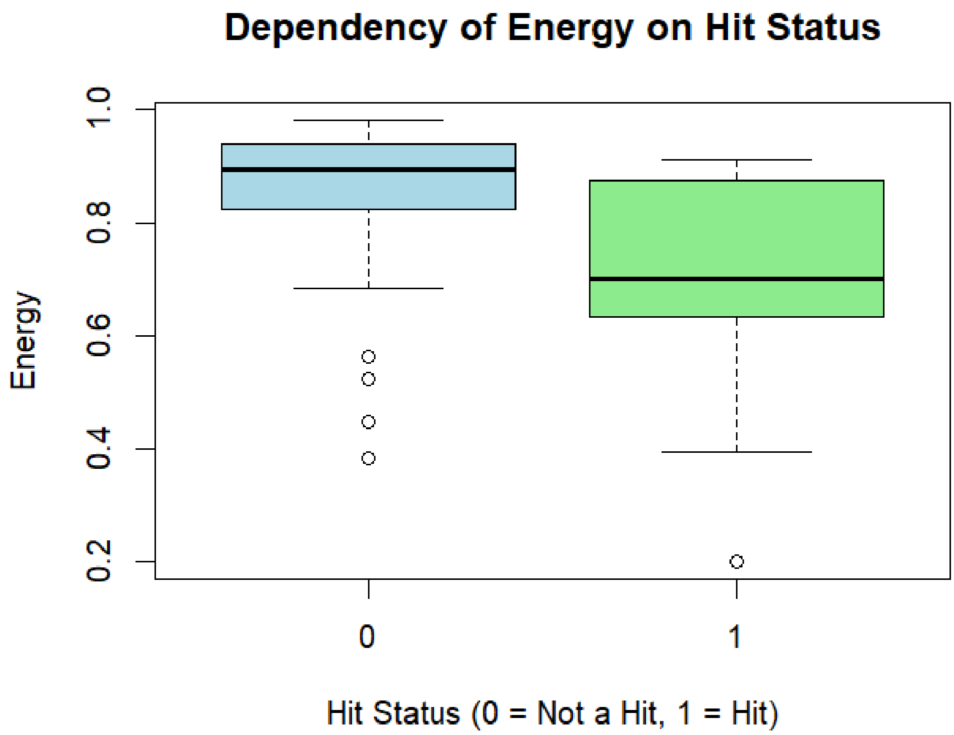
In our case study we will check if the following categories have an influence at the probability of a song being a hit:

- danceability
- energy
- loudness
- speechiness
- acousticness
- valence
- tempo

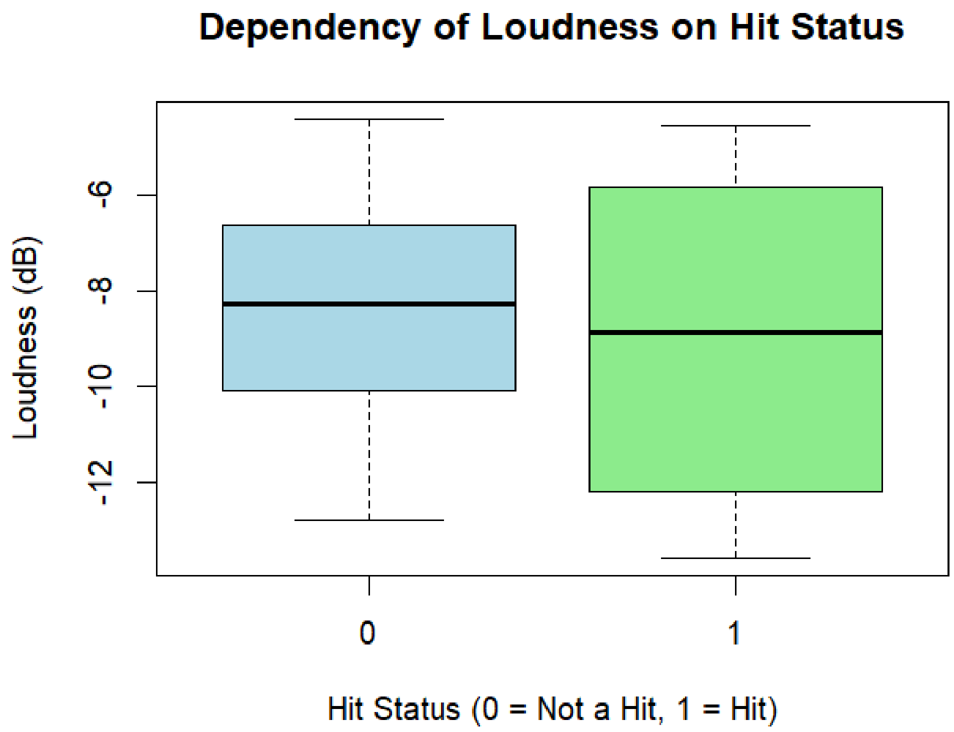
Before using the regression let's see the plots and look for dependency.



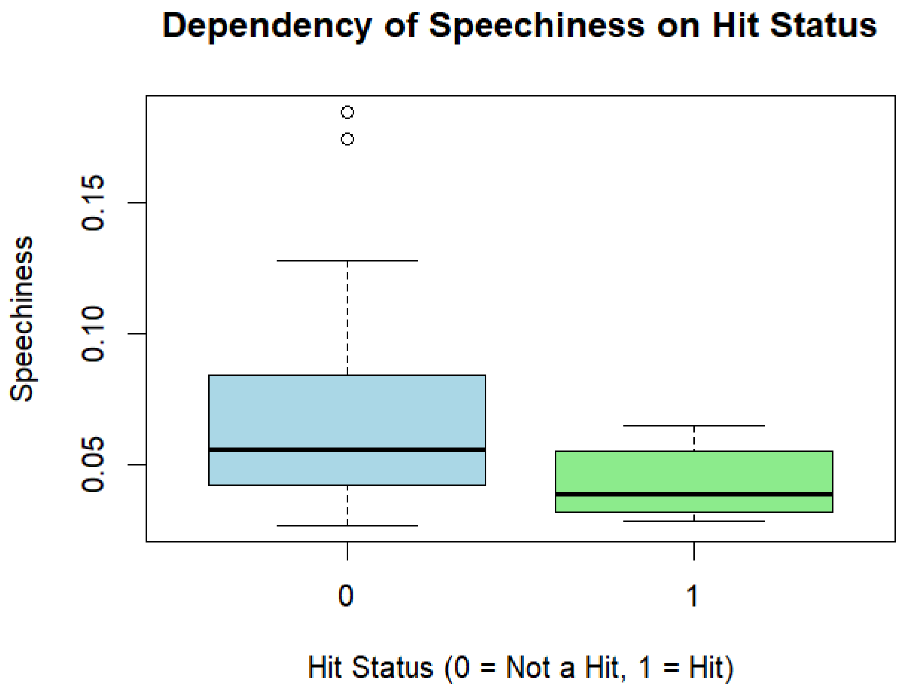
From this plot we can see that the songs that turned out to be hits tend to have a bigger danceability value. The median of hit songs' danceability was bigger than the third quartile of non hit songs.



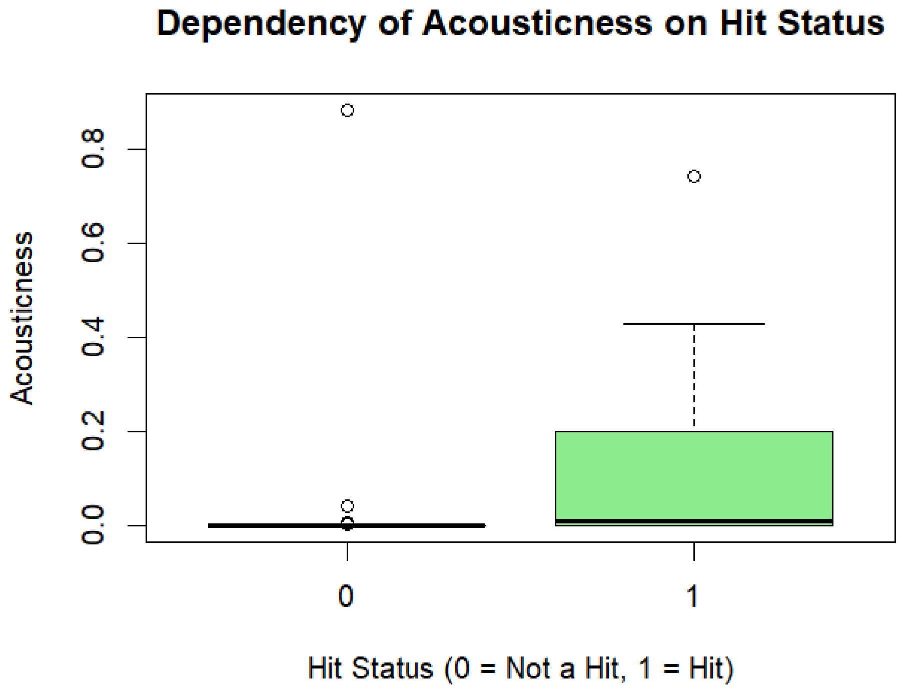
The hit songs tend to have lower energy values. Mean of hits' energy was almost at the level of minimum without outliers of non-hit songs.



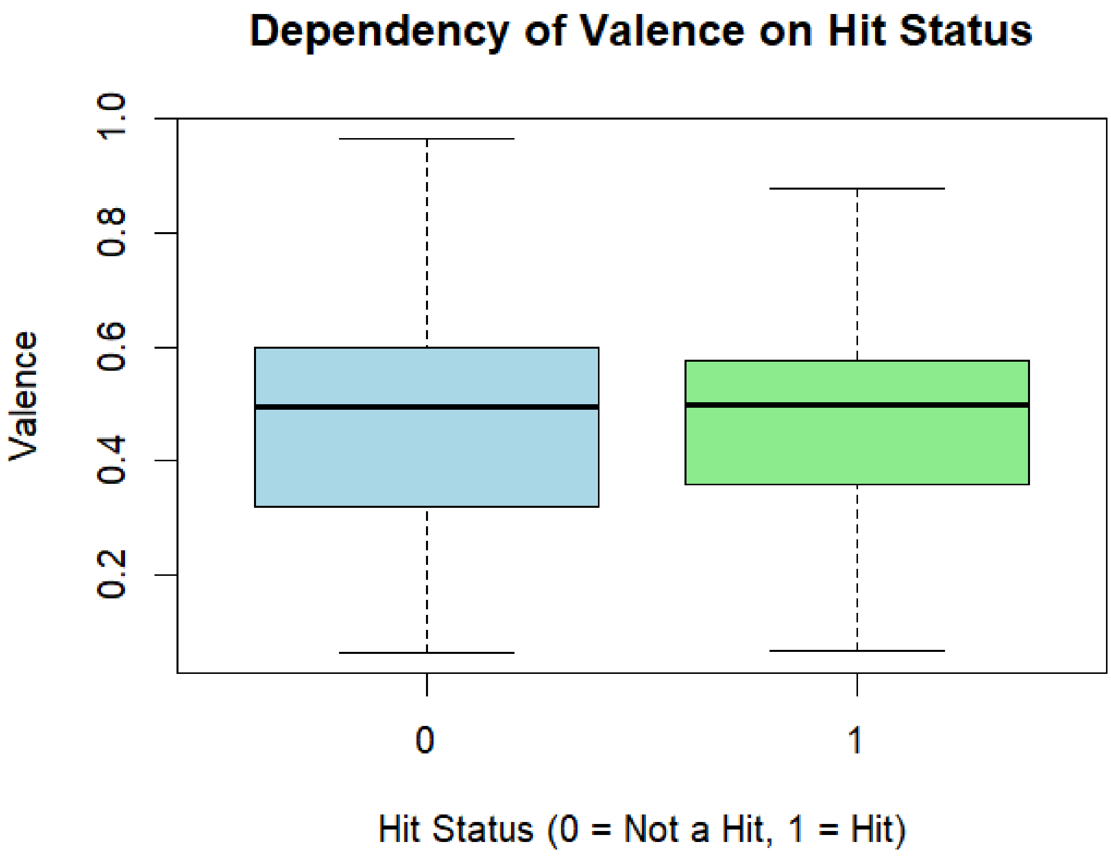
There is no clear relation between the loudness of a song and its hit status.



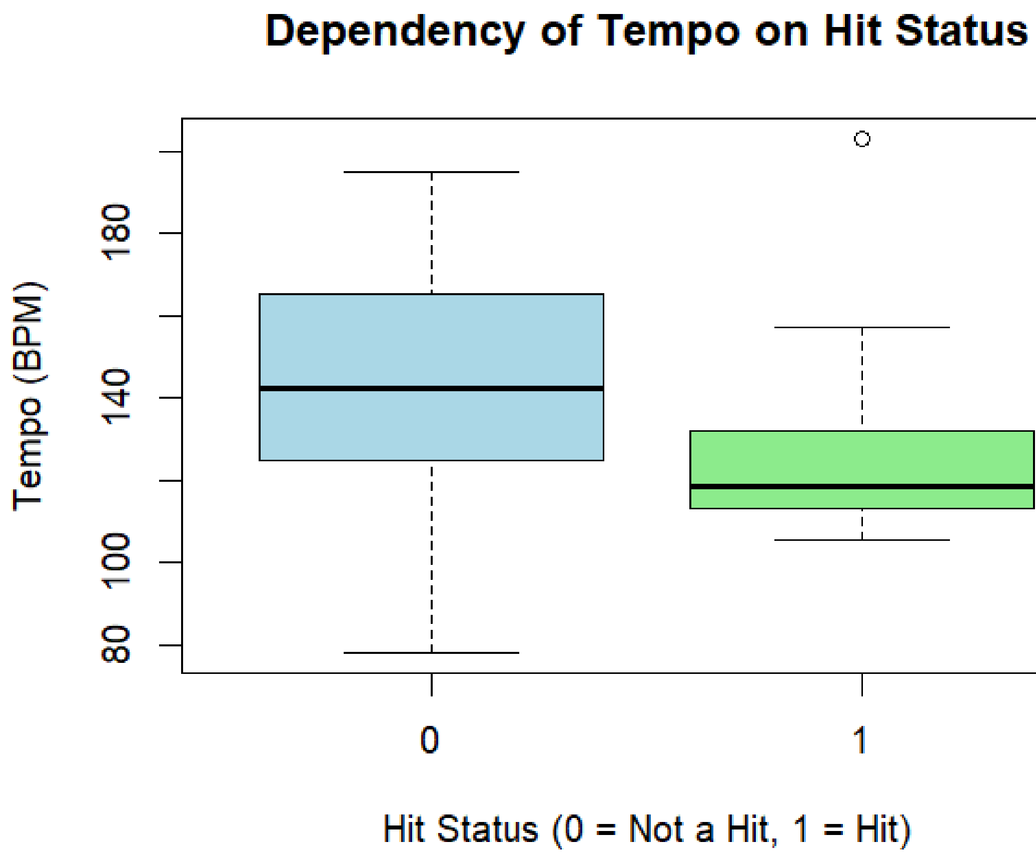
Hit songs tend to have lower speechines levels than non-hit songs.



The non hit songs usually aren't acoustic. This is an interesting case, because the most acoustic song on Nirvana's discography according to Spotify statistics (top left corner) is one of the biggest contenders to pass the hit songs threshold, being the 13th most streamed non-live songs (currently 126mln streams).



There is no clear relation between the Valence and the hit status.



Hit songs tend to be slower than non-hit songs.

3. Logistic Regression

We will try to solve this problem with logistic regression. The model we are going to use is:

$$hit[i] \mid p_i \sim Bernoulli(p_i)$$

Where hit[i] indicates if a song of index i is a hit or not and:

$$log \frac{p_i}{1-p_i} = \alpha + \sum_{j=1} (\beta_j * x_j)$$

Where x_j is the variable to which the wage β_j refers.

We also assume proper, non informative priors for the parameters:

$$\alpha \sim Normal(0,\ 1000)$$
$$\beta_j \sim Normal(0,\ 1000)$$

We will use the method MCMClogit from MCMCpack R package to obtain a sample of a posterior α and β.

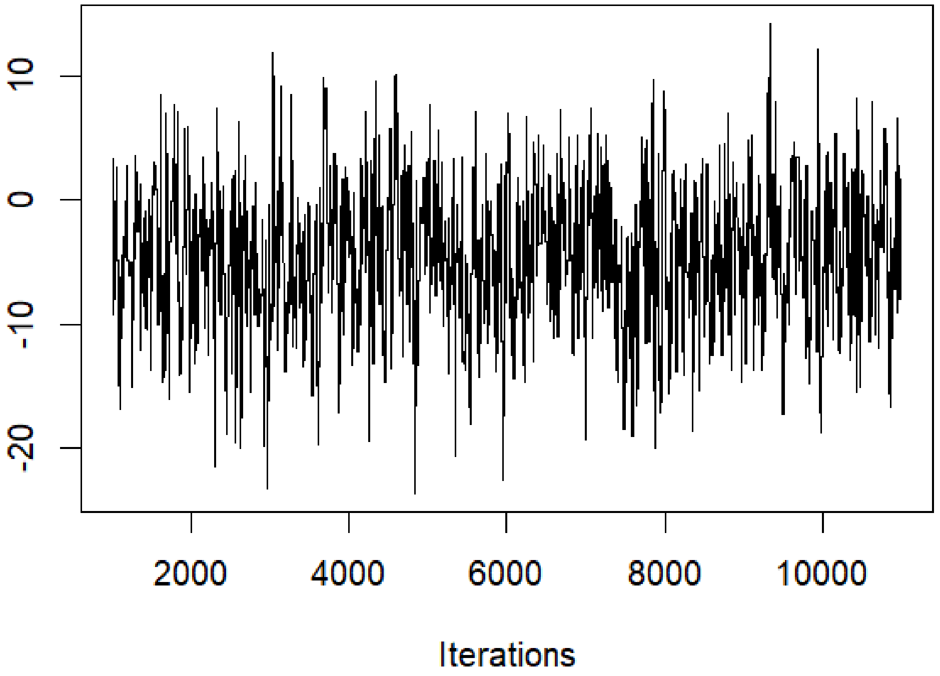
```
library(MCMCpack)

bayes.logit <- MCMClogit(bin.hit ~ danceability + energy +
speechiness + acousticness + valence + tempo, data = data,
burnin = 1000, mcmc = 10000, thin = 1)
```

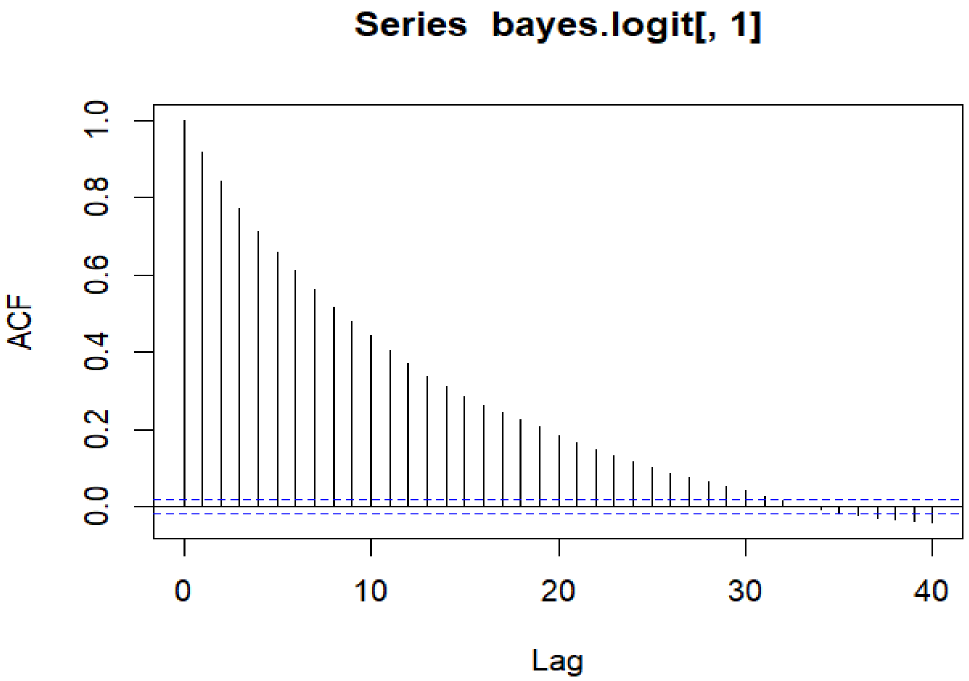
```
Iterations = 1001:11000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 10000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

                  Mean       SD  Naive SE Time-series SE
(Intercept)   -4.43590  5.20534 0.0520534       0.261112
danceability   7.30926  5.52250 0.0552250       0.263427
energy         0.93965  4.15743 0.0415743       0.193332
speechiness  -53.80689 32.28722 0.3228722       1.614852
acousticness   1.91272  3.54489 0.0354489       0.176561
valence       -3.92299  2.85323 0.0285323       0.130832
tempo          0.02704  0.02513 0.0002513       0.001194

2. Quantiles for each variable:

                  2.5%       25%       50%       75%    97.5%
(Intercept)   -14.92373  -7.70518  -4.52667  -0.76639  5.39544
danceability   -3.30535   3.58335   7.10314  10.77812 18.93527
energy         -7.17998  -1.79182   0.74798   3.57630  9.32324
speechiness  -131.92109 -70.66940 -49.69105 -31.49954  0.23419
acousticness   -4.43494  -0.46045   1.50675   4.16593  9.52318
valence        -9.75467  -5.70534  -3.93527  -1.99612  1.59146
tempo          -0.02025   0.01009   0.02549   0.04392  0.07757
```

Judging by the criteria we used in class, every predictor is insignificant, since **every interval contains a 0.** Let's check the autocorrelation and mixing of the sample distributions and see if anything has changed.

For the case of clarity in the report we will only look at the ACF and Trace plots of Intercept, but during the processing of the data every variable has been checked.

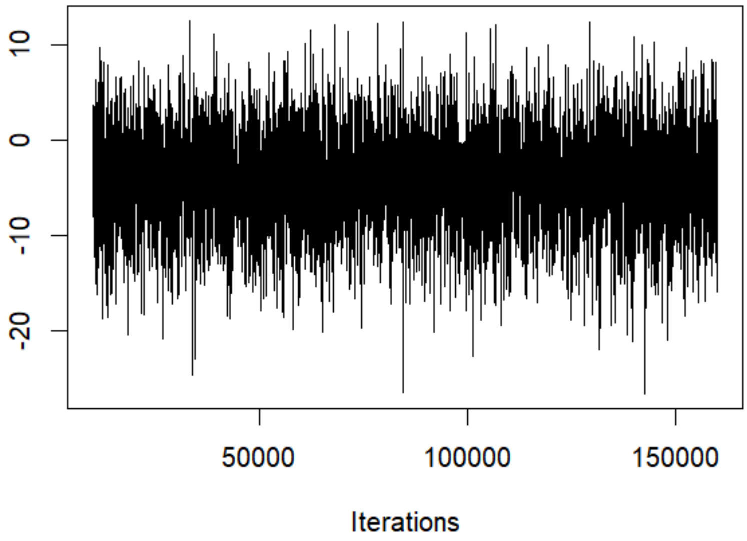
Intercept Traceplot:

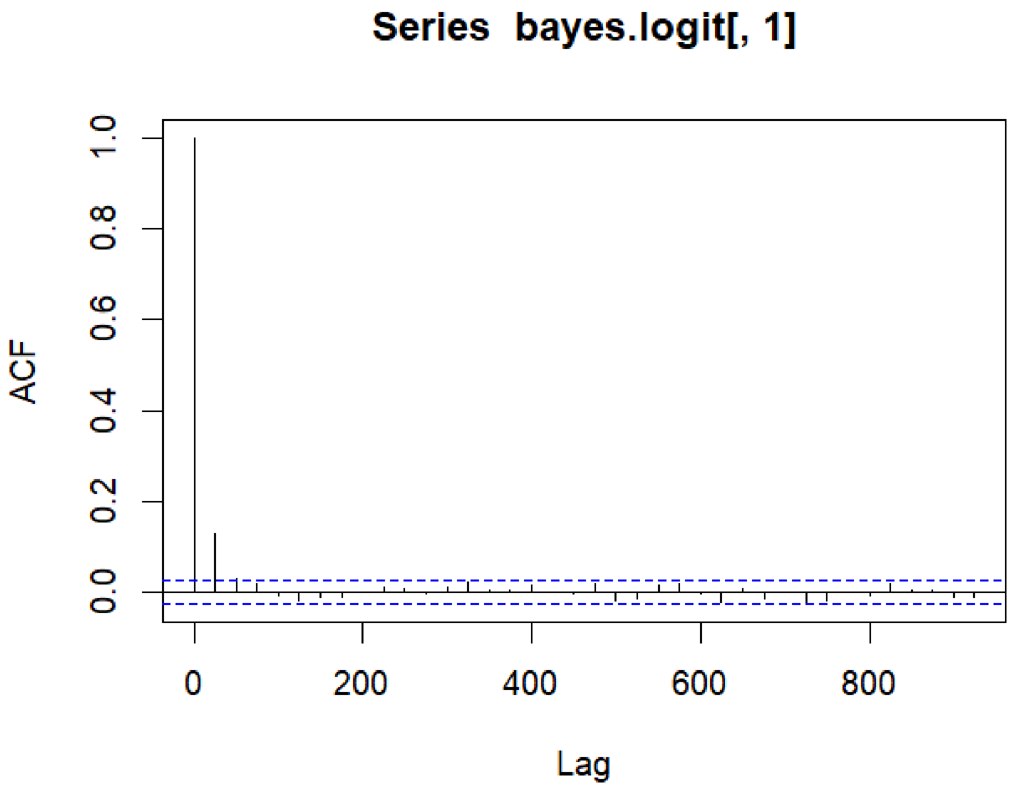




We can see that neither the mixing or the autocorrelation of the intercept look good. Let's increase the number of burnin and mcmc iterations and introduce thinning = 20.

```
bayes.logit <- MCMClogit(bin.hit ~ danceability + energy + speechiness + acousticness + valence + tempo, data = data, burnin = 10000, mcmc = 150000, thin = 25)
```





Now the autocorrelation and mixing look much better. Let’s check the summary again.

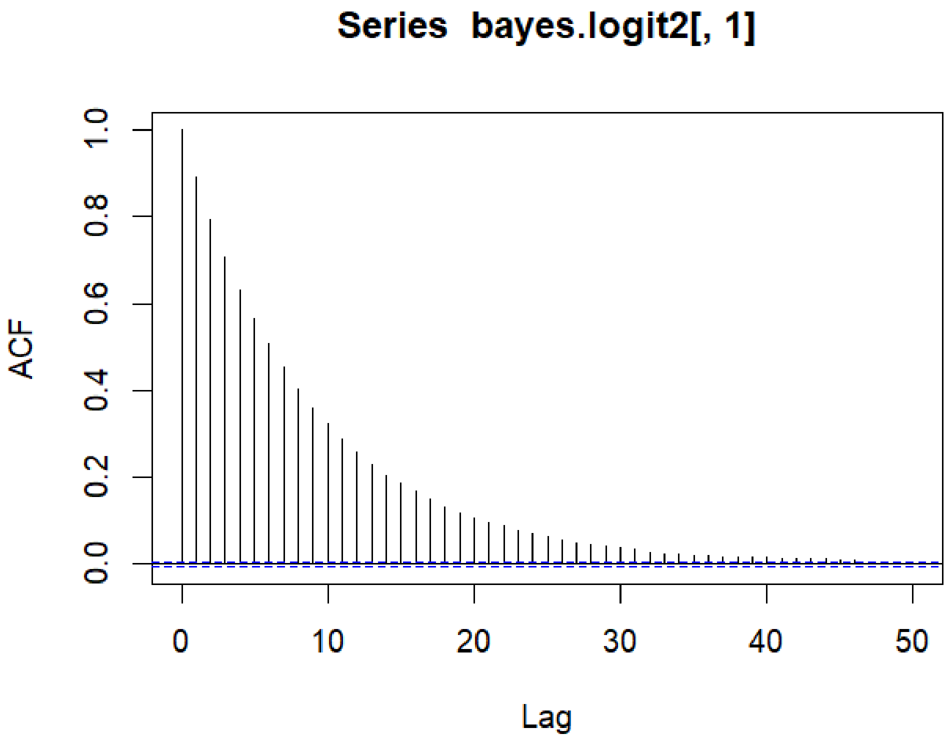
```
summary(bayes.logit)
Iterations = 10001:159976
Thinning interval = 25
Number of chains = 1
Sample size per chain = 6000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

                 Mean       SD  Naive SE Time-series SE
(Intercept)   -4.1739  5.17960 0.0668684      0.0761740
danceability   7.0234  5.51233 0.0711639      0.0811403
energy         0.9842  4.28676 0.0553419      0.0619489
speechiness  -53.7862 32.09415 0.4143337      0.4653620
acousticness   2.0227  3.67571 0.0474532      0.0557640
valence       -4.1156  2.91374 0.0376163      0.0422123
tempo          0.0263  0.02503 0.0003231      0.0003603

2. Quantiles for each variable:

                  2.5%       25%       50%       75%    97.5%
(Intercept)   -14.68004  -7.60933  -4.04473  -0.69448  5.86746
danceability   -3.40091   3.27694   6.92735  10.54643 18.39791
energy         -7.10576  -1.90313   0.87685   3.78617  9.70476
speechiness  -126.59984 -73.15402 -49.94514 -30.79713 -2.05614
acousticness   -4.85918  -0.50410   1.89136   4.43567  9.43259
valence       -10.07003  -6.03315  -4.01020  -2.14584  1.39113
tempo          -0.02049   0.00919   0.02545   0.04232  0.07833
```

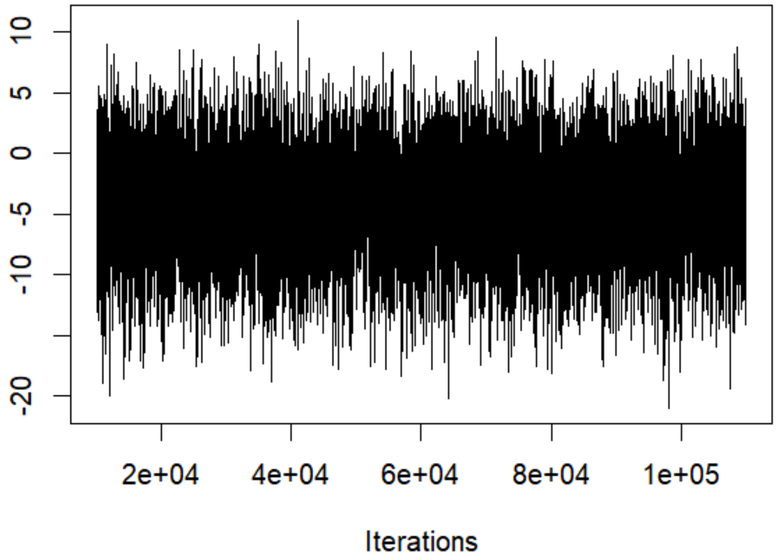
Only one predictor has proved to be significant by our previous rules. Due to the small sample size of only 54 songs, we will consider every value that has no 0 in the 50% confidence interval quantiles 25%-75% a 0. Therefore we will build a new model with the following values:

- danceability
- tempo
- speechiness
- valence

```
bayes.logit2 <- MCMClogit(bin.hit ~ danceability + tempo + speechiness +
valence, data = data, burnin = 10000, mcmc=100000)
```

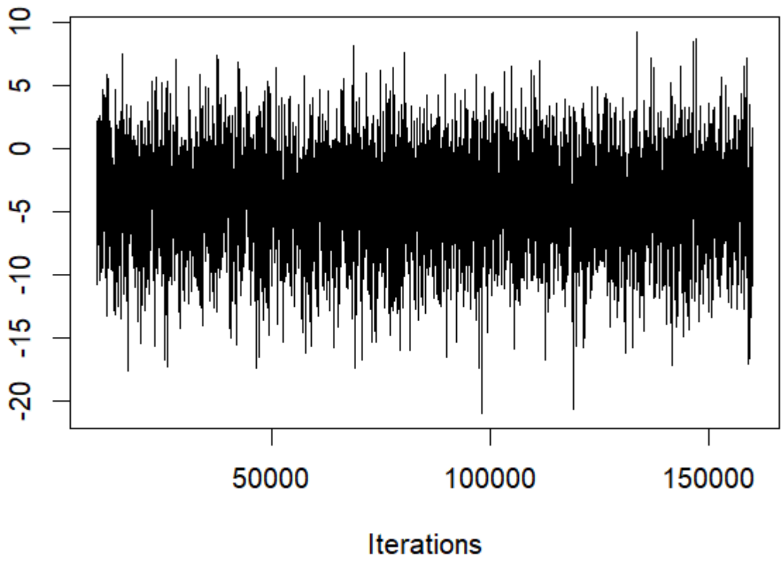
Let’s check the new models mixing and autocorrelation.

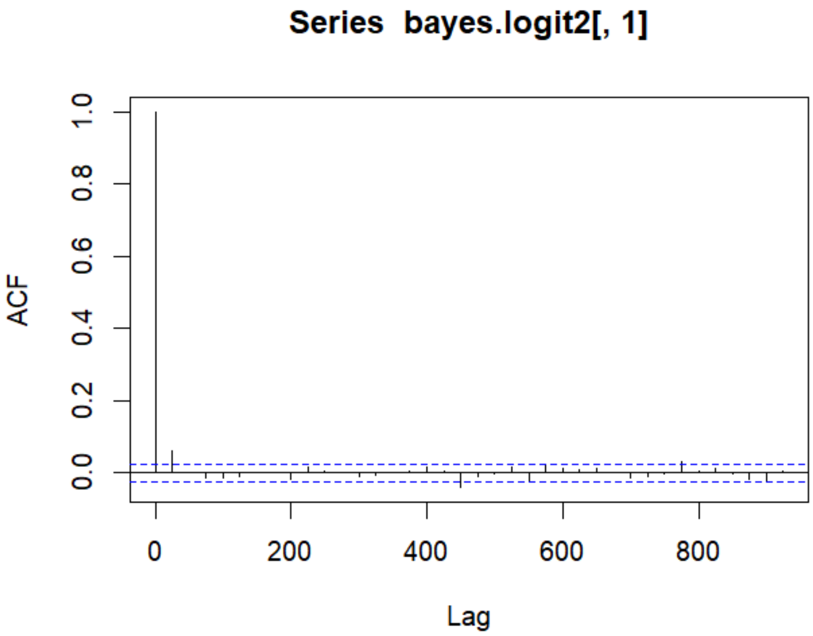




We need to get rid of the autocorrelation again, let’s implement thinning = 25.

```
bayes.logit2 <- MCMClogit(bin.hit ~ danceability + tempo + speechiness +
valence, data = data, burnin = 10000, mcmc=150000, thin=25)
```





Now, our model has been completed. Let's check out the summary of bayes.fit2 and plug in the values of Nirvana's biggest song *Smells Like Teen Spirit* to see what were the chances of it being a hit.

```
summary(bayes.logit2)
Iterations = 10001:159976
Thinning interval = 25
Number of chains = 1
Sample size per chain = 6000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

                 Mean       SD  Naive SE Time-series SE
(Intercept)  -4.10136  3.83331 0.0494878      0.0526620
danceability  7.75492  4.35891 0.0562733      0.0605026
tempo         0.02881  0.02348 0.0003031      0.0003175
speechiness -50.22236 26.74782 0.3453128      0.3723576
valence      -4.08932  2.43556 0.0314429      0.0331845

2. Quantiles for each variable:

                  2.5%       25%       50%       75%    97.5%
(Intercept)  -11.92752  -6.57627  -4.02436  -1.48541  3.13827
danceability  -0.30848   4.84013   7.54413  10.41910 16.84557
tempo         -0.01622   0.01271   0.02842   0.04432  0.07682
speechiness -111.16770 -66.50291 -47.41441 -30.92417 -6.40743
valence       -9.07020  -5.68511  -3.97678  -2.42987  0.47581
```

4. Hit chances predictions

Now, after we have obtained our linear regression model, let's predict the probability of songs becoming hits. We will take into the consideration 2 iconic grunge songs.

- ***Nirvana - Smells Like Teen Spirit*** - 1 995 705 601 streams
    - danceability: 0.502
    - tempo: 116.761
    - speechiness: 0.0564
    - valence: 0.72
- ***Soundgarden - Black Hole Sun*** - 678 729 136 streams
    - danceability: 0.35
    - tempo: 105.435
    - speechiness: 0.041
    - valence: 0.147

We're going to convert the logit(p) to p using the function inv.logit(x) from the boot library.

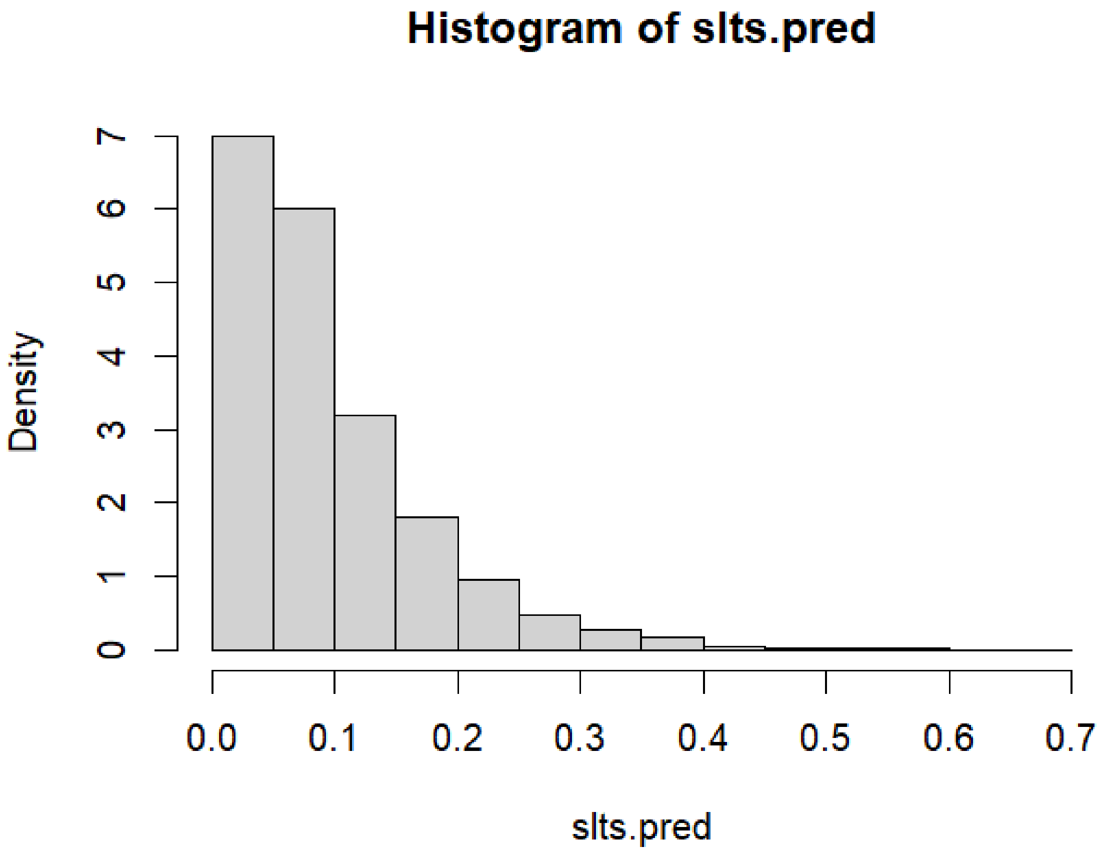
Smells Like Teen Spirit - Nirvana:

Let’s obtain the sample from predictive distribution and let’s check what was the chance of this song becoming the biggest hit in the history of grunge.

```
slts.linear=bayes.logit2[,1]+10.502*bayes.logit2[,2]
+116.761*bayes.logit2[,3]+0.0564*bayes.logit2[,4]+0.72*bayes.logit2[,5]
slts.pred = inv.logit(slts.linear)
```

Let’s check the mean probability, predictive interval and the histogram of predictive probability.

```
mean(slts.pred)
[1] 0.09435791
quantile(slts.pred, c(0.025, 0.975))
       2.5%       97.5% 
0.009010175 0.309728597
```



As we can see, the biggest song in the history of the genre, doesn’t suit the trends in the audio features of our model the best. Let’s see how it compares with the other songs.

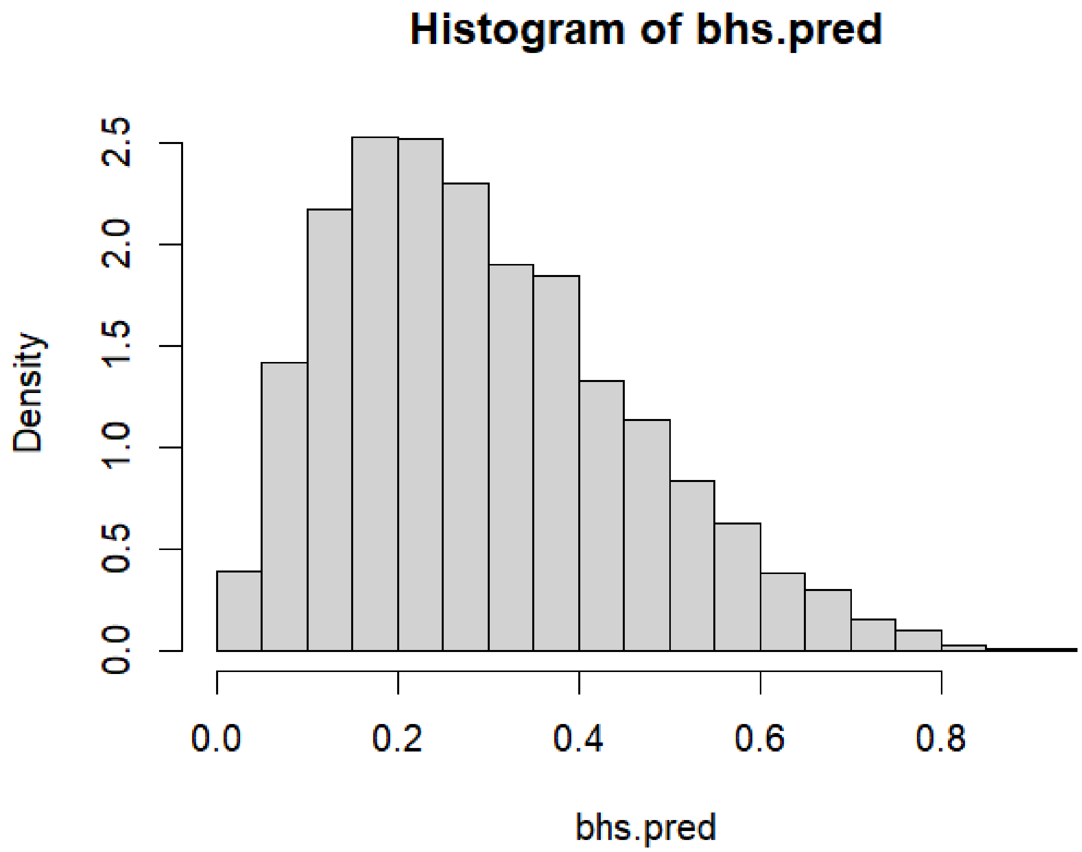
Black Hole Sun - Soundgarden:

Now let’s compare the regression based on Nirvana’s discography with other bands.

```
bhs.linear = bayes.logit2[,1] + 0.35*bayes.logit2[,2] +
105.435*bayes.logit2[,3]+0.041*bayes.logit2[,4]+0.147*bayes.logit2[,5]

bhs.pred = inv.logit(bhs.linear)
```

The mean, predictive interval and histogram:

```
> mean(bhs.pred)
[1] 0.2960849
> quantile(bhs.pred, c(0.025, 0.975))
      2.5%      97.5%
0.05633594 0.66659762
```



Ironically, Black Hole Sun matches the criteria for being a Nirvana Hit Song, better than the biggest Nirvana hit. The mean probability is 3 times bigger than the one obtained in the Smells *Like Teen Spirit* case.

5. Results and Conclusions

The Bayesian logistic regression model was used to predict the likelihood of a Nirvana song becoming a hit based on audio features such as danceability, tempo, speechiness, and valence. The model was trained on a dataset that included songs from Nirvana's four studio albums.

Key Observations:

1. **Danceability**: The analysis revealed that higher danceability values are associated with hit songs. The median danceability of hits exceeded the third quartile of non-hits, indicating a clear tendency for more rhythmic and danceable songs to gain more streams.
2. **Tempo**: Contrary to expectations, the tempo of hit songs was generally slower than that of non-hits. While high-tempo songs are often successful in other genres, this result aligns with the characteristics of Nirvana's grunge style, where slower, moodier tempos tend to prevail. The average tempo for hits was approximately 117 beats per minute.
3. **Speechiness**: There was a negative relationship between speechiness and the likelihood of a song being a hit. Lower speechiness values, indicating fewer spoken word elements, correlated with a higher probability of success. This observation is consistent with the fact that most Nirvana hits are traditional rock songs with limited speech-like components.
4. **Valence**: Hit songs were found to have lower valence scores, having a more negative or melancholic emotional tone. This result is in line with the darker, introspective mood typically associated with grunge music. The average valence of hits was significantly lower compared to non-hits, reinforcing this trend.

Comparative Analysis: Smells Like Teen Spirit vs. Black Hole Sun

As part of the evaluation, the model was used to predict the hit probability of Nirvana's most-streamed song, **"Smells Like Teen Spirit"**, and Soundgarden's **"Black Hole Sun"**.

- **Smells Like Teen Spirit**: The model predicted a hit probability of **9.4%** for this song, which is lower than expected given its status as Nirvana's most iconic track with nearly 2 billion streams. This suggests that, while the song achieved immense popularity, its audio features do not fully align with those typically associated with hits in the dataset.
- **Black Hole Sun**: In contrast, **"Black Hole Sun"** had a predicted hit probability of **29.6%**, despite having fewer streams than "Smells Like Teen Spirit." This result indicates that the model considers the audio features of "Black Hole Sun" more favorable for a hit song under the given criteria, even though its actual popularity is lower.

Conclusions

This project used Bayesian logistic regression to explore how audio features like **danceability**, **tempo**, **speechiness**, and **valence** might predict the success of a song. The model focused on Nirvana's discography and offered some interesting insights.

Key Takeaways:

- **Danceability** and **tempo** stood out as key factors. Songs that were more danceable and had slower tempos tended to perform better, which makes sense given Nirvana's moody, grunge sound.
- **Speechiness** played a negative role, meaning that songs with fewer spoken elements were more likely to become hits. This aligns with Nirvana's style of traditional rock songs with minimal spoken word content.
- **Valence**, or the emotional tone of the song, showed that darker, more melancholic tracks had better chances of being hits—again, fitting with the typical vibe of grunge music.

While these insights are useful, there are limitations to the model:

- **Audio Features Only**: The model considers audio features, which is just one piece of the puzzle. Other crucial elements, like **lyrics**, **cultural impact**, and **promotion strategies**, are not included. These non-audio factors play a massive role in a song's success, but they're harder, or even impossible to quantify and in result they weren't included in this analysis.
- **Sample Size**: The dataset consisted of only 54 songs, therefore the model's results can't be fully generalized. A larger dataset with more songs across multiple artists or genres would improve the reliability of the results and allow for a more comprehensive analysis.

Although modern bands would want to find a 'recipe' for a hit song, we need to remember that **music is subjective.** What makes a song resonate with people can't always be captured in numbers. Music is about emotion, connection, and personal experience, all of which are hard to quantify. So while this model can highlight some interesting trends, it's not a magic formula for creating a hit song. Success in music is influenced by so many **unpredictable factors**, from timing and cultural relevance to **personal taste**.

In conclusion, while the Bayesian model gives us a better understanding of how certain measurable features might affect a song's popularity, the real story of a hit is much more complex. Incorporating additional factors like **lyrics**, **cultural significance**, and **promotional strategies** in future studies would provide a fuller picture of what makes a song truly successful.