DATA SOURCE
https://www.kaggle.com/datasets/madroscla/taylor-swift-released-song-discography-genius

ORIGINAL DATASET
Contained 359 observations and 13 variables. These variables included: album_title, album_url, category, album_track_number, song_title, song_url, song_artists, song_release_date, song_page_views, song_lyrics, song_writers, song_producers, and song_tags.

CLEANED DATASET
Contained 266 observations and 5 columns listed in the data dictionary below.

Variable	Class	Description
text	character	The lyrics to corresponding Taylor Swift song
category	character	The album each song is from
title	character	The title of the song
date	date	The date on which each song was released including year, month, and day
year	numeric	The year in which the song was released

FINAL DATASET

The final dataset contained 266 observations and 30 variables. The most relevant variables are described below.

Variable	Class	Description
Text	character	The text variable containing the lyrics to corresponding Taylor Swift song
Album	character	The album each song is from
Title	character	The title of the song
Date	date	The date on which each song was released including year, month, and day
Year	character	The year in which the song was released
Sentiment QDAP	numeric	The sum of positive and negative sentiment for each song in each library in the Sentiment Analysis package
Love	numeric	The numeric score of love in text
Sorrow	numeric	The numeric score of sorrow in text
Confidence	numeric	The numeric score of confidence in text
Resilience	numeric	The numeric score of resilience in text
Hope	numeric	The numeric score of hope in text
Loss	numeric	The numeric score of loss in text
Desire	numeric	The numeric score of desire in test
Reflection	numeric	The numeric score of reflection in text
Nostalgia	numeric	The numeric score of nostalgia in text

Text:
The text variable contained the lyrics and was the focal point of our analysis. This variable contained 5385 terms from 359 songs. There was no missing data. This variable was pre-processed and turned into a document text matrix. First text was turned into a corpus. All text was then converted to lowercase; all numbers, stopwords, and punctuations were removed, and only usable text was maintained. This process was achieved with the tm package in R. When data was in the form of a document term matrix, sparsity was reduced to increase the number of terms or words in the matrix that have high frequency. Using the removeSparseTerms function, sparse was set equal to 0.965, thereby decreasing sparsity to 90% and keeping only 515 terms from the original 5385 terms. This variable underwent sentiment analysis, topic modeling, and zero-shot-classification for generating emotional scores.

Figure 1: Image of the top 20 words in Taylor Swift's lyrics where words in bigger font are more frequent

Figure 2: Taylor Swift Lyrics word and frequency distribution

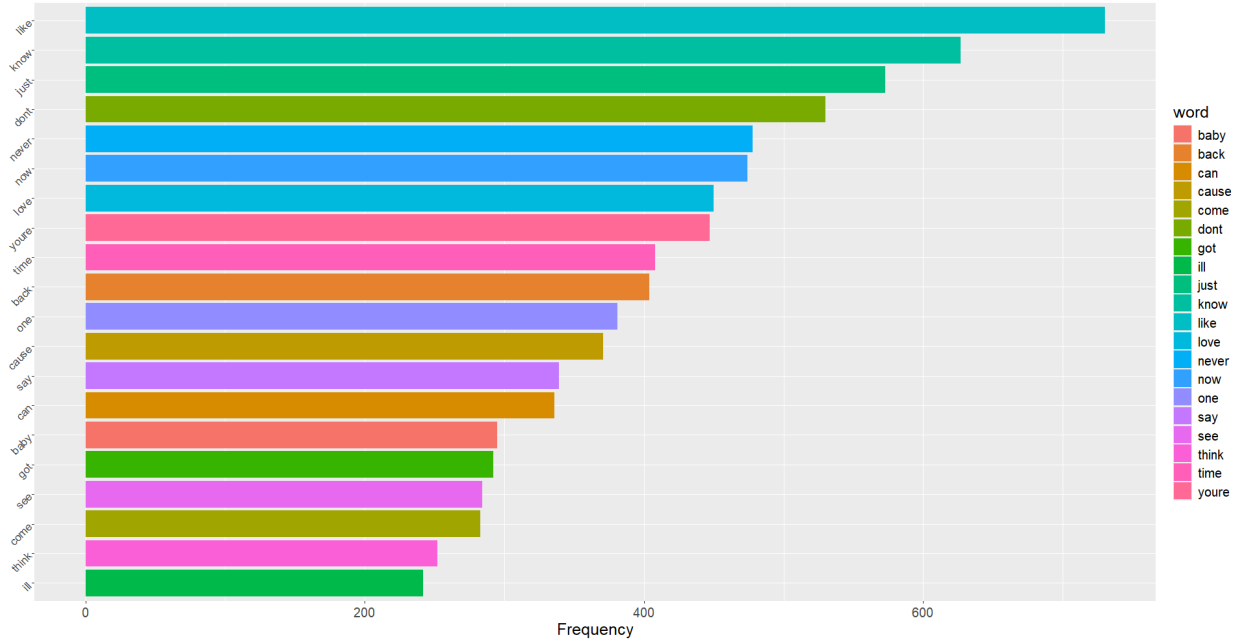
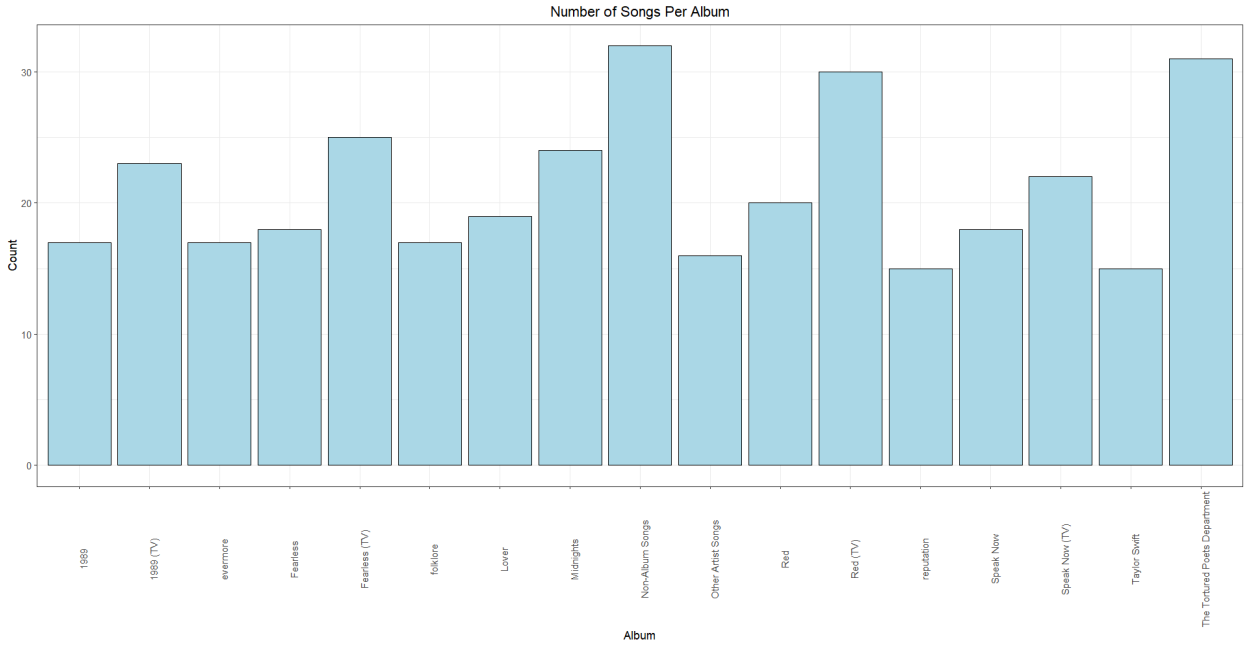


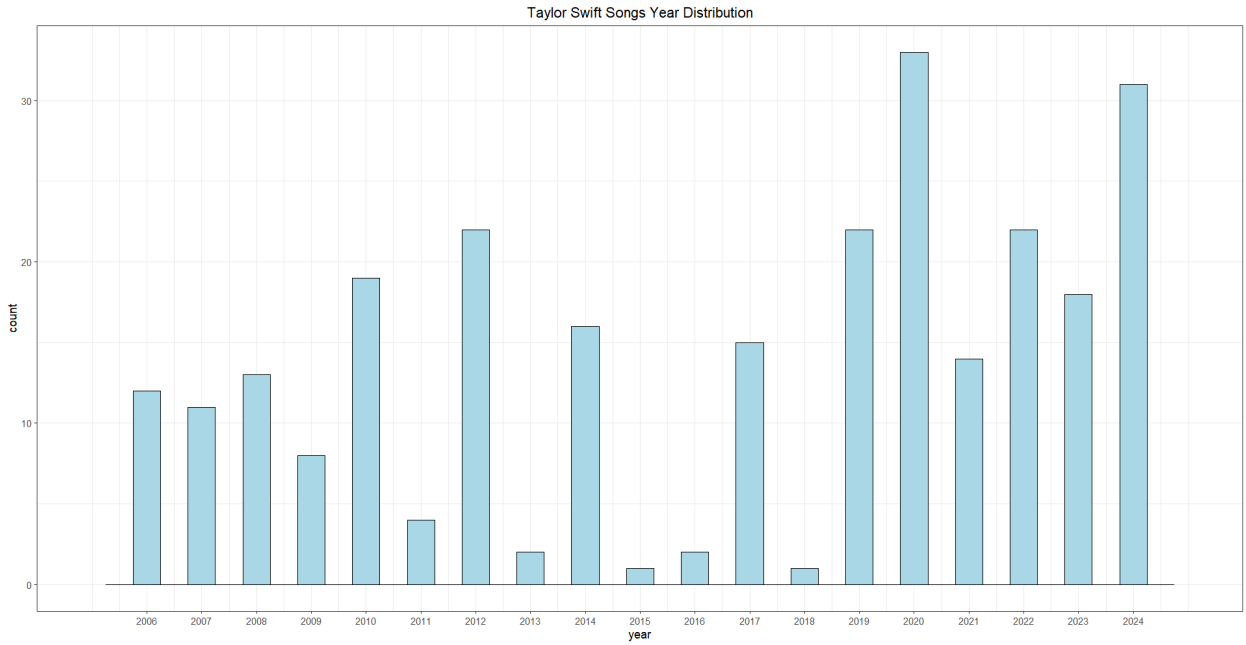
Figure 3: Topic Distribution for 75 words

Category and Title:
The category variable contains the album from which each song came. In the dataset, there were originally 379 songs from 17 albums including: Taylor Swift, Fearless, Fearless (TV), Speak Now, Speak Now (TV), Red, Red (TV), 1989, 1989 (TV), reputation, Lover, folklore, Non-Album Songs, evermore, Midnights, The Tortured Poets Department, and Other Artist Songs/covers. The clean dataset to better represent the sentiments, topics, and emotions of Taylor Swift's songs excludes the 16 other artist songs. The title variable contains the name of the songs from each album. It served best as an id for our data since it contained unique observations. The processing performed on this variable was to distinguish repeated songs
n= 266
m= 0

Figure 4: Barplot of Albums

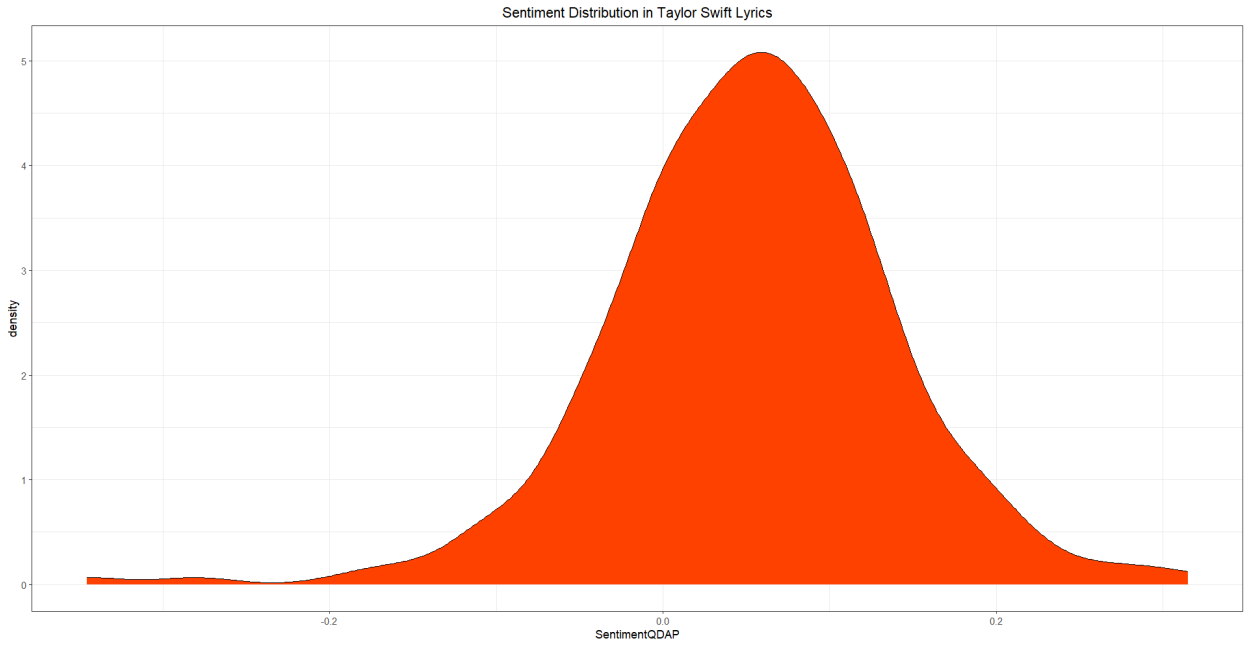


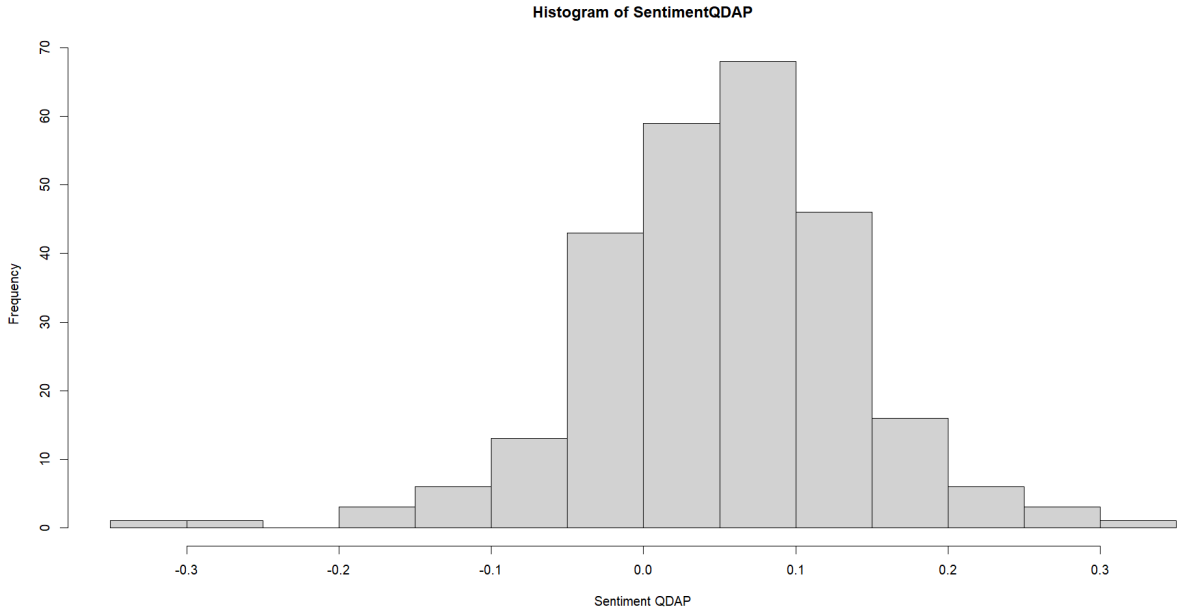
Date and Year:
The year variable was derived from the date variable, where substr() was used to select the first 4 characters in date. These 4 characters were saved in year and transformed to a numeric variable.
The year variable had a maximum year of 2024, and a minimum year of 2006, making the range for our year 2006- 2024, a span of 18 years.
n= 266
m= 0
Figure 5: Histogram of Year

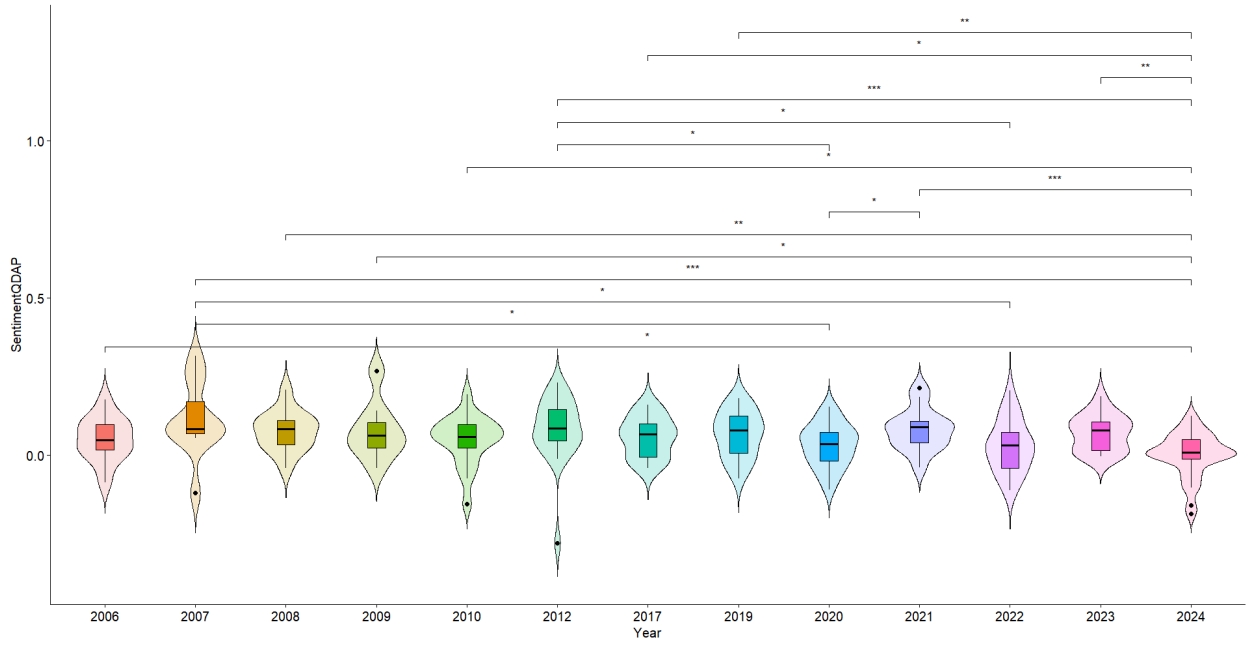


SentimentQDAP:
The Sentiment QDAP is the numerical score of sentiment calculated from the text variable using the SentimentAnalysis package. The Sentiment QDAP normally ranges from -1 to 1, indicating strong negative and positive sentiments respectively. A Sentiment QDAP of 0 is neutral. The lyrics from our dataset were not sentimentally polarized leading to most of the sentiment scores falling between -0.1 and 0.1. The range fell between -0.3 and 0.3, illustrating little variance in the sentiments present in Taylor's lyrics.

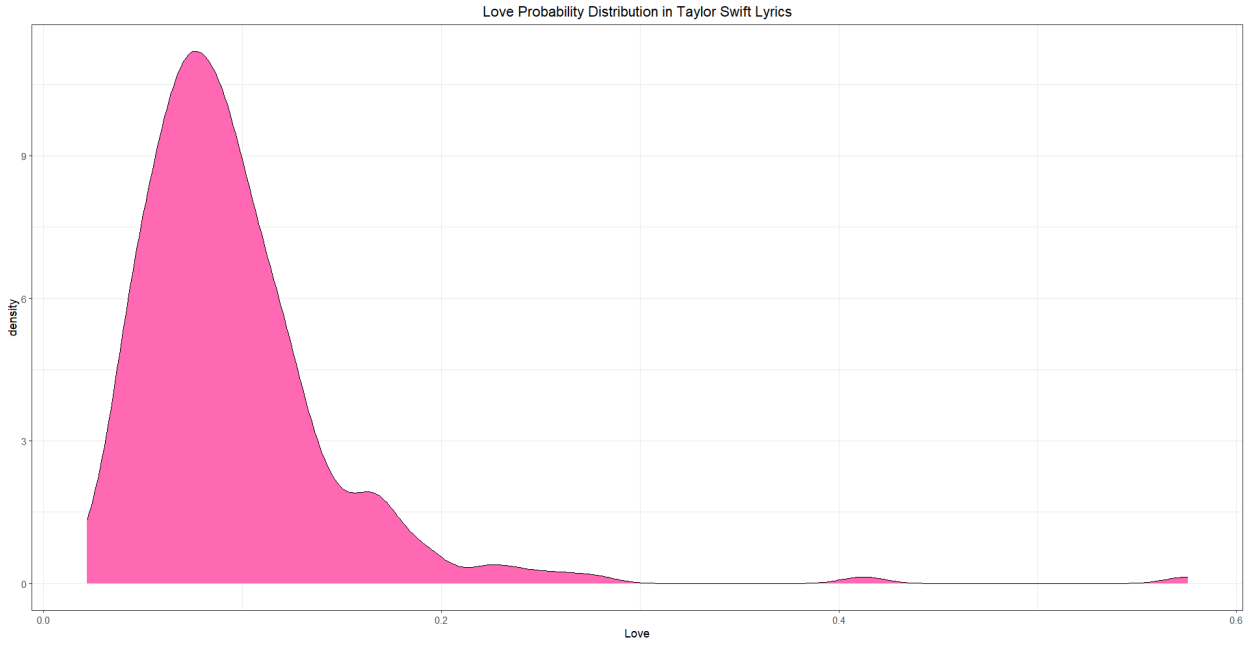
Love, Sorrow, Desire, Hope, Loss, Nostalgia, Reflection, Resilience, Confidence:
The emotional theme variables listed above were generated using the TransforEmotion package. The output for each emotion is a probability score, between 0 and 1, indicating how confident the model is that a given emotion is present in text. The range of scores for these 9 emotions was 0.15 - 0.55 indicating the model had low confidence about the emotion present in the text. This may be explained by the neutral or subtle emotions in Taylor's lyrics that could not be detected by the model, or the fact that some of these emotions are a combination of other emotions.
Images Above: Show density of the scores for each emotion
Images Below: Show the distribution of the emotional theme score for each year, determining if there is statistical significance.The following are the significance levels present in the graph: [ns](P >0.0.05), [*](P ≤ 0.05), [**](P ≤ 0.01), [***](P≤ 0.001).

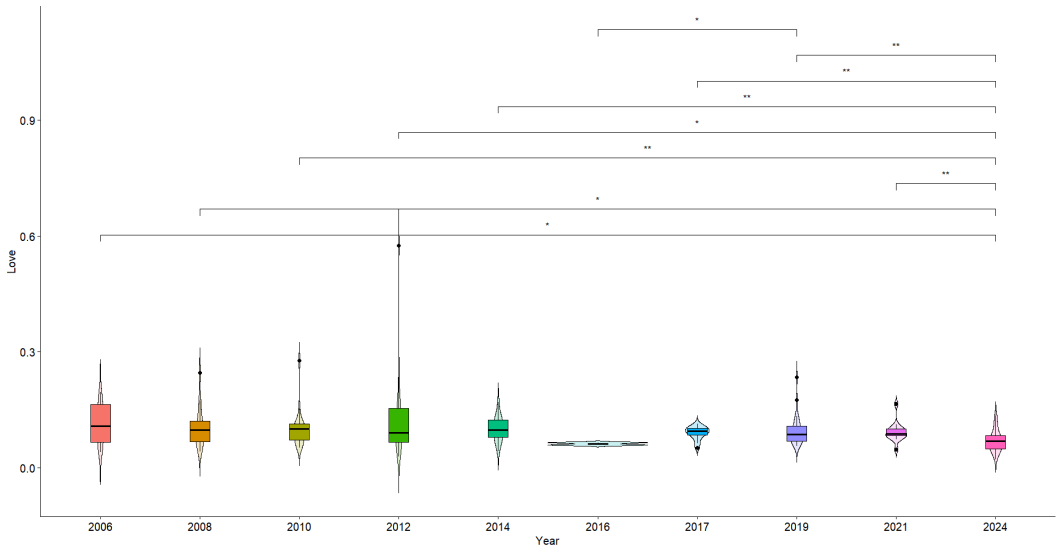



Histogram of SentimentQDAP
Frequency
0
10
20
30
40
50
60
70
-0.3
-0.2
-0.1
0.0
0.1
0.2
0.3
Sentiment QDAP

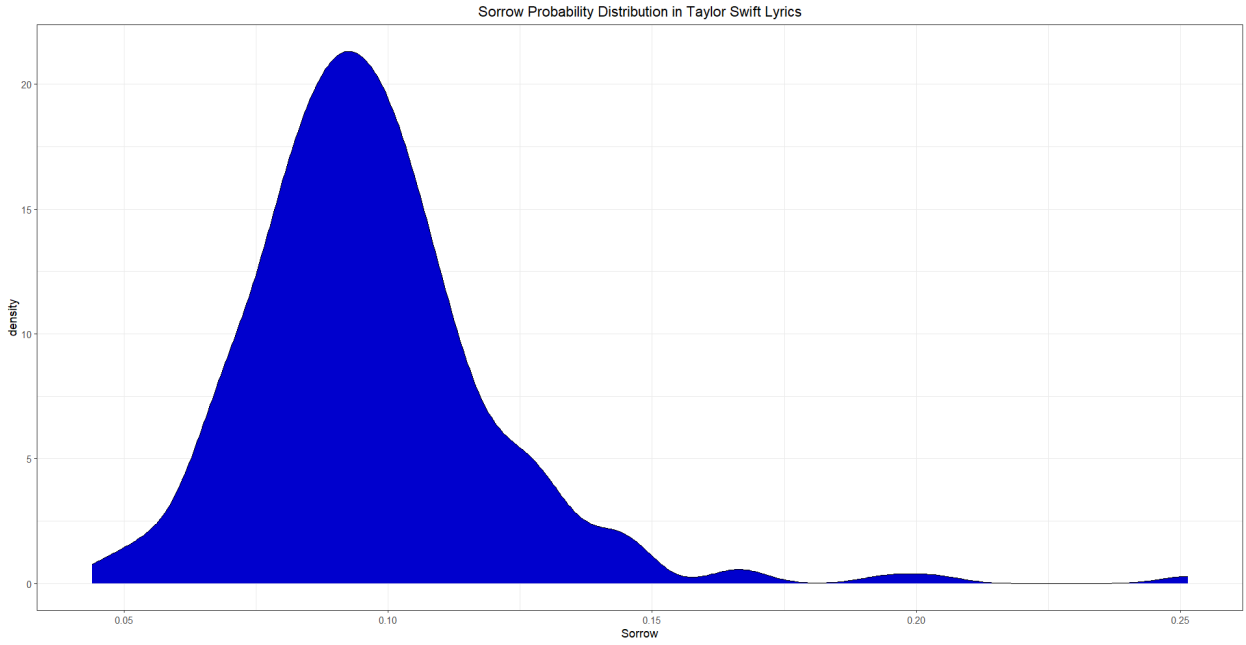

SentimentQDAP
0.0
0.5
1.0
2006
2007
2008
2009
2010
2012
2017
2019
2020
2021
2022
2023
2024
Year

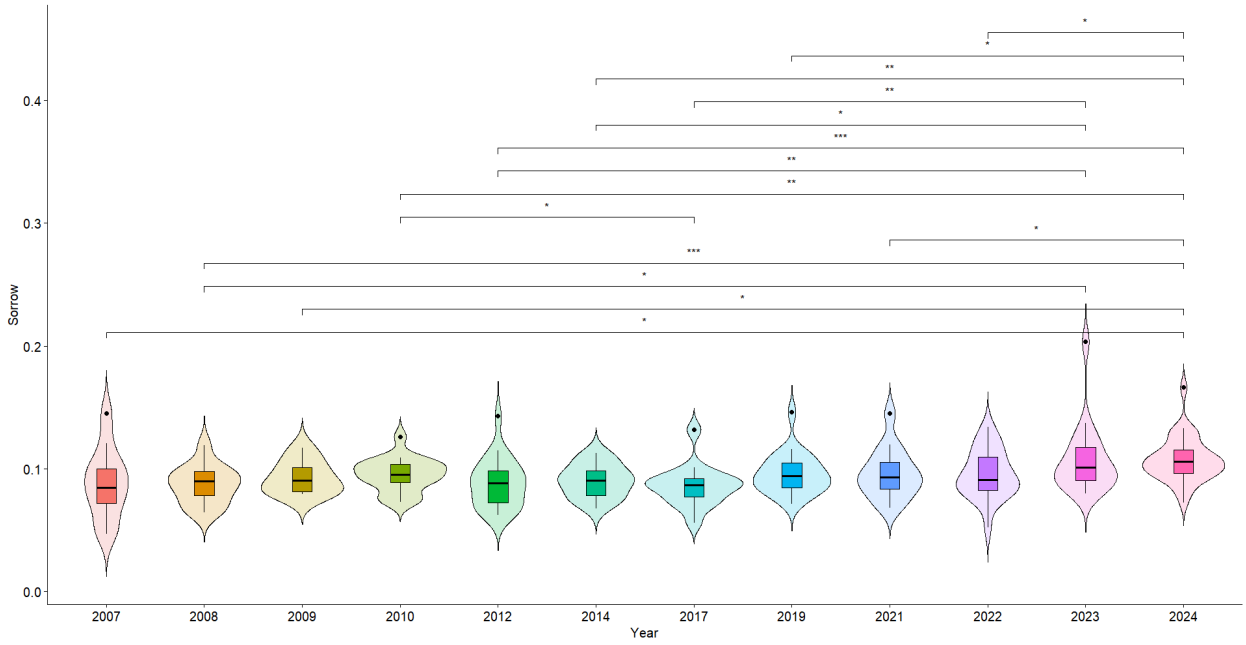
Love:


Love Probability Distribution in Taylor Swift Lyrics
density
Love
0
3
6
9
0.0
0.2
0.4
0.6

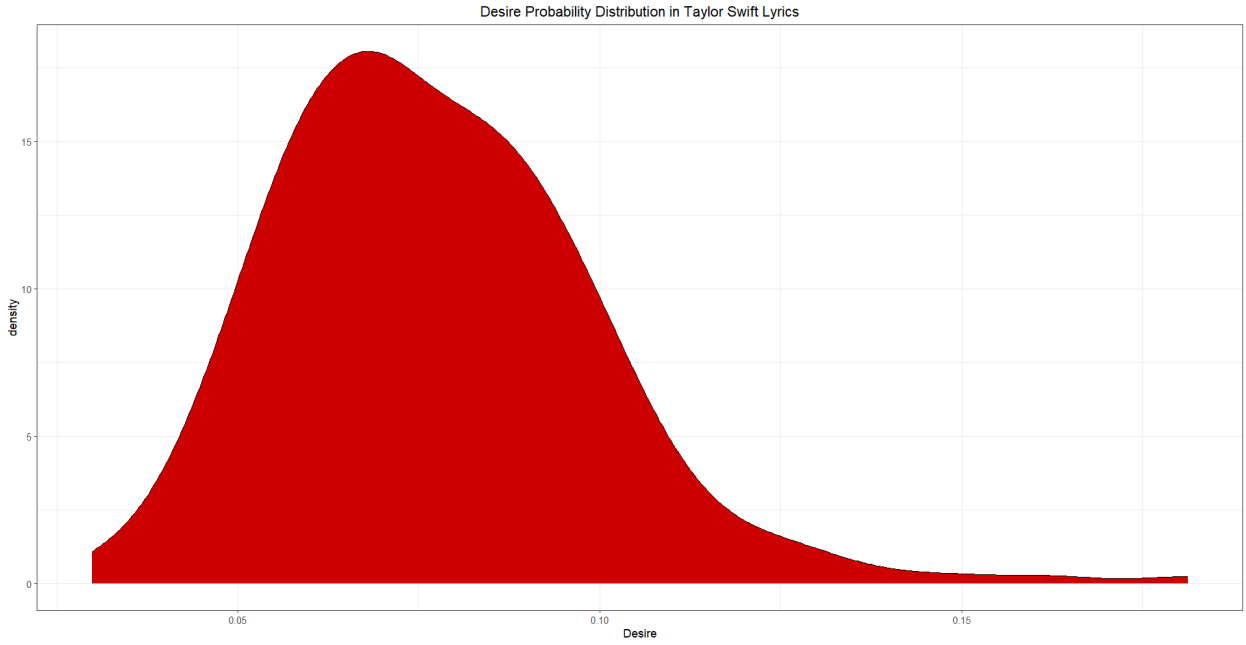

Love
Year
0.0
0.3
0.6
0.9
2006
2008
2010
2012
2014
2016
2017
2019
2021
2024
*
**
**
**
*
**
**
*
*

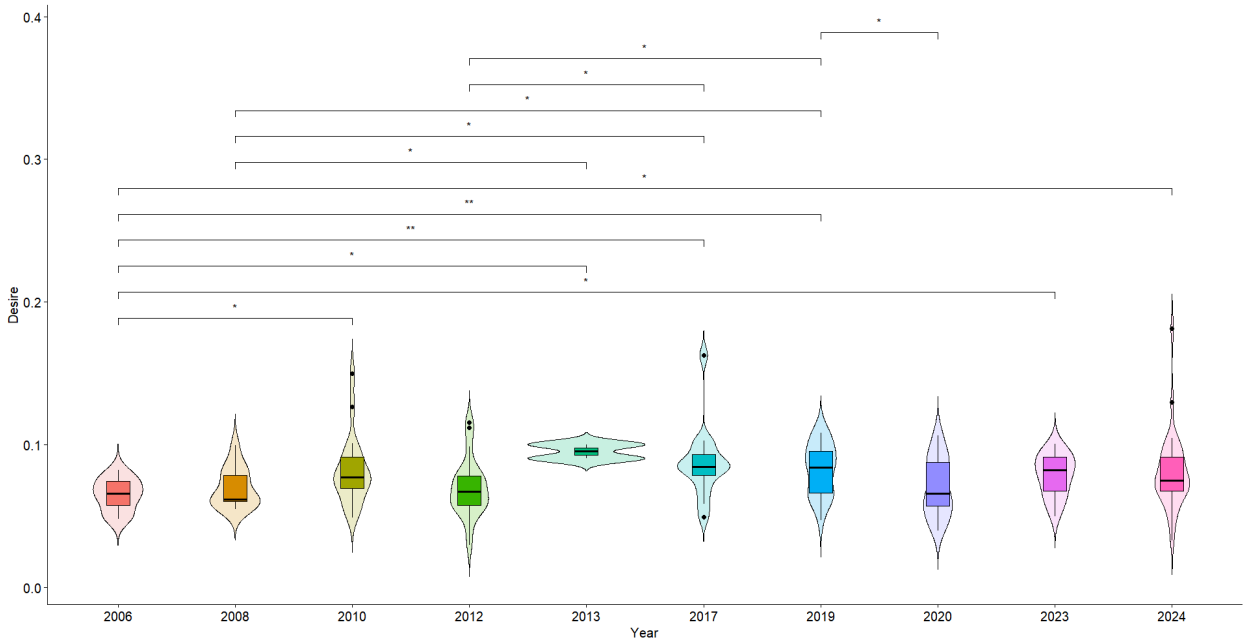
Sorrow:


Sorrow Probability Distribution in Taylor Swift Lyrics
density
Sorrow
0
5
10
15
20
0.05
0.10
0.15
0.20
0.25

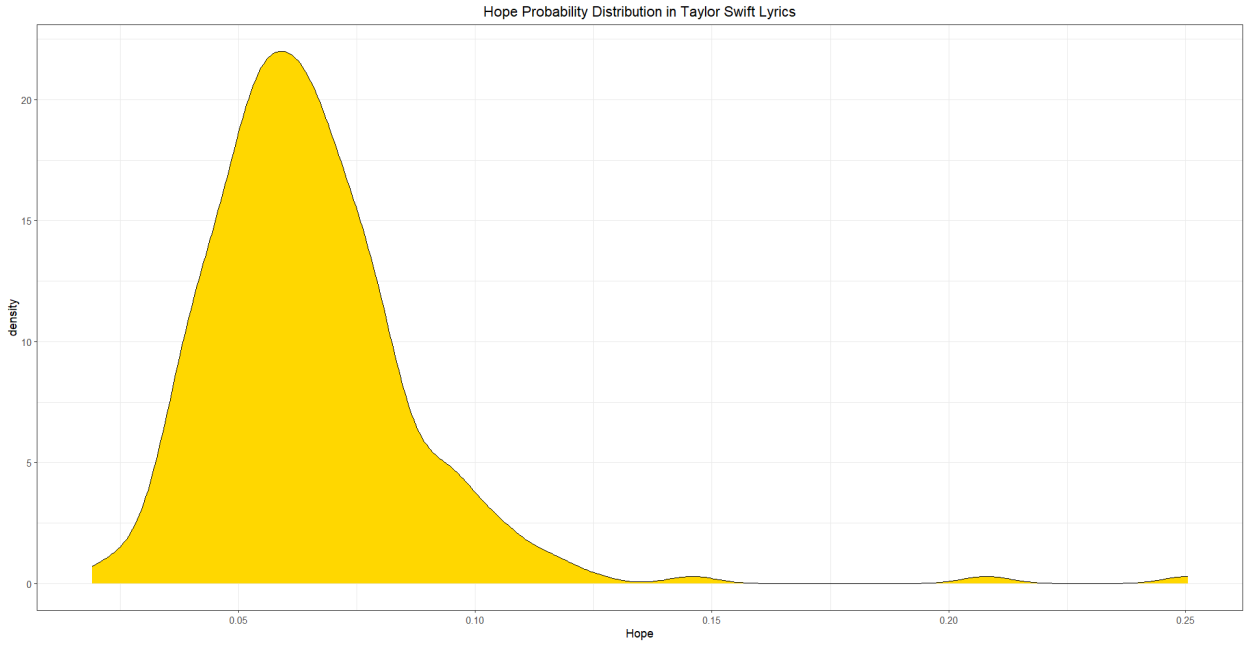

Sorrow
Year
0.0
0.1
0.2
0.3
0.4
2007
2008
2009
2010
2012
2014
2017
2019
2021
2022
2023
2024

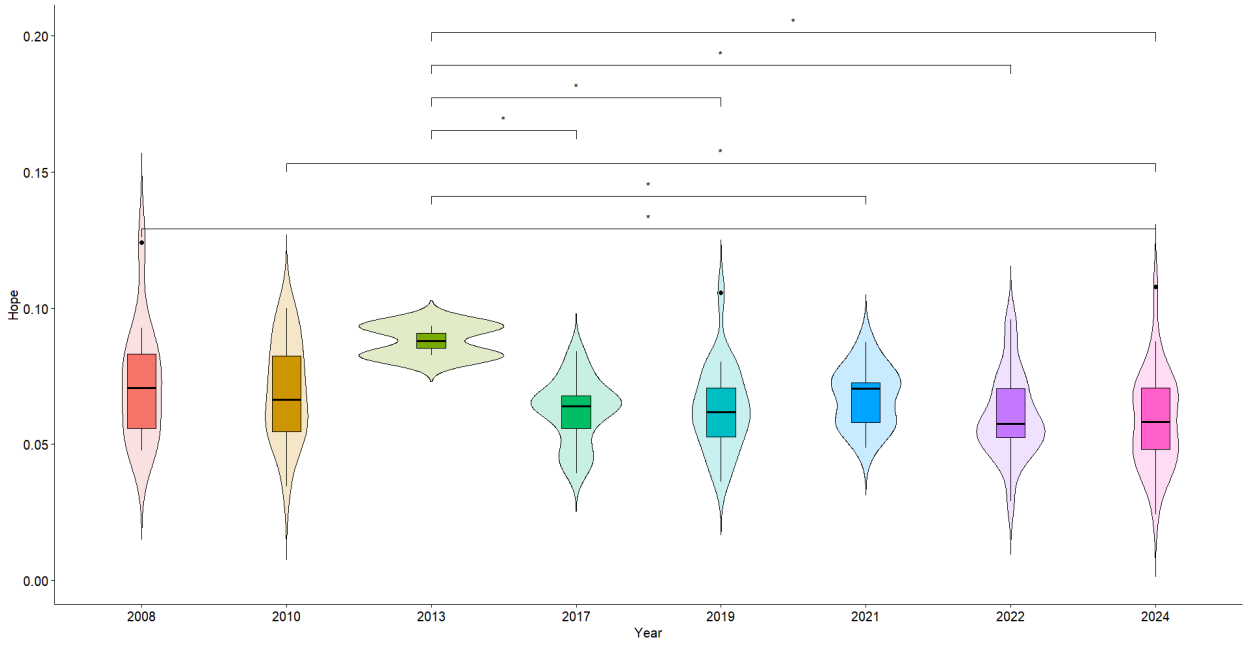
Desire:


Desire Probability Distribution in Taylor Swift Lyrics
density
Desire
0
5
10
15
0.05
0.10
0.15

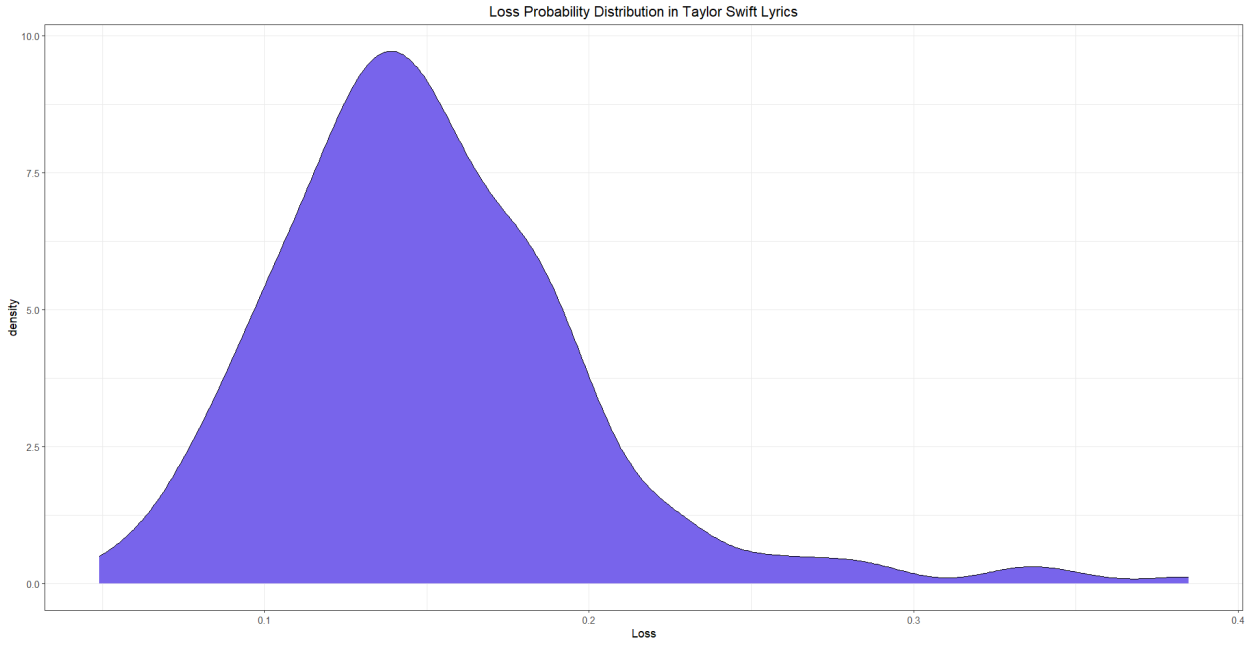

Desire
Year
0.0
0.1
0.2
0.3
0.4
2006
2008
2010
2012
2013
2017
2019
2020
2023
2024

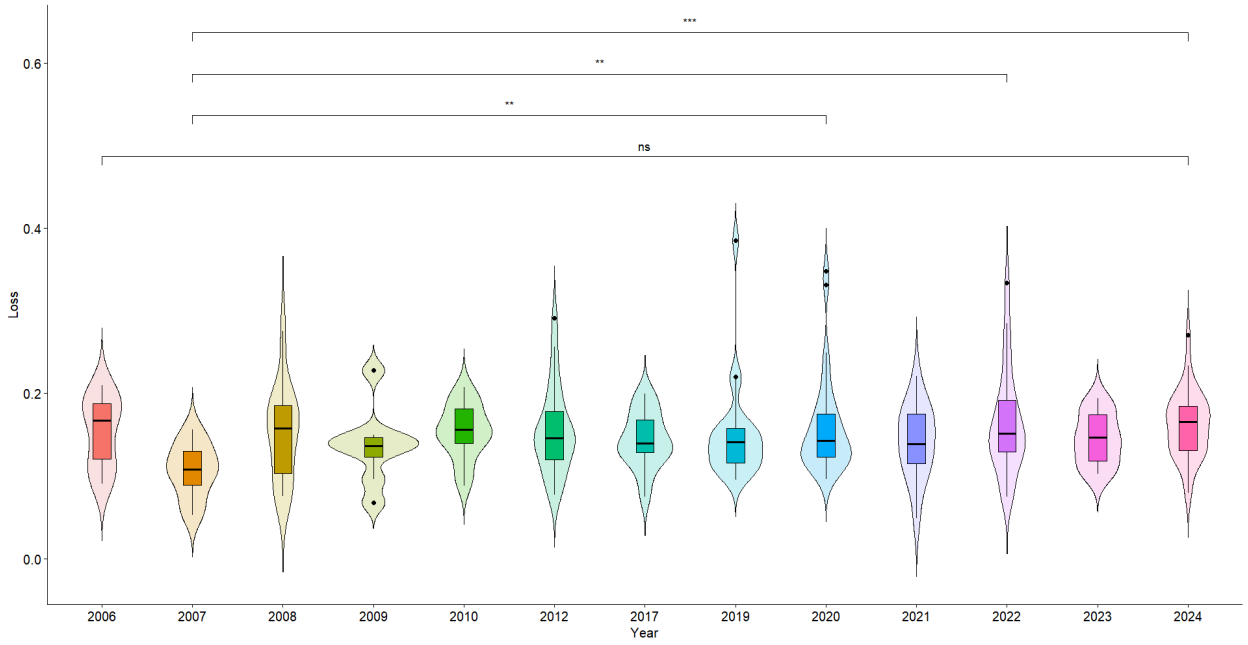
Hope:


Hope Probability Distribution in Taylor Swift Lyrics
density
Hope
0
5
10
15
20
0.05
0.10
0.15
0.20
0.25

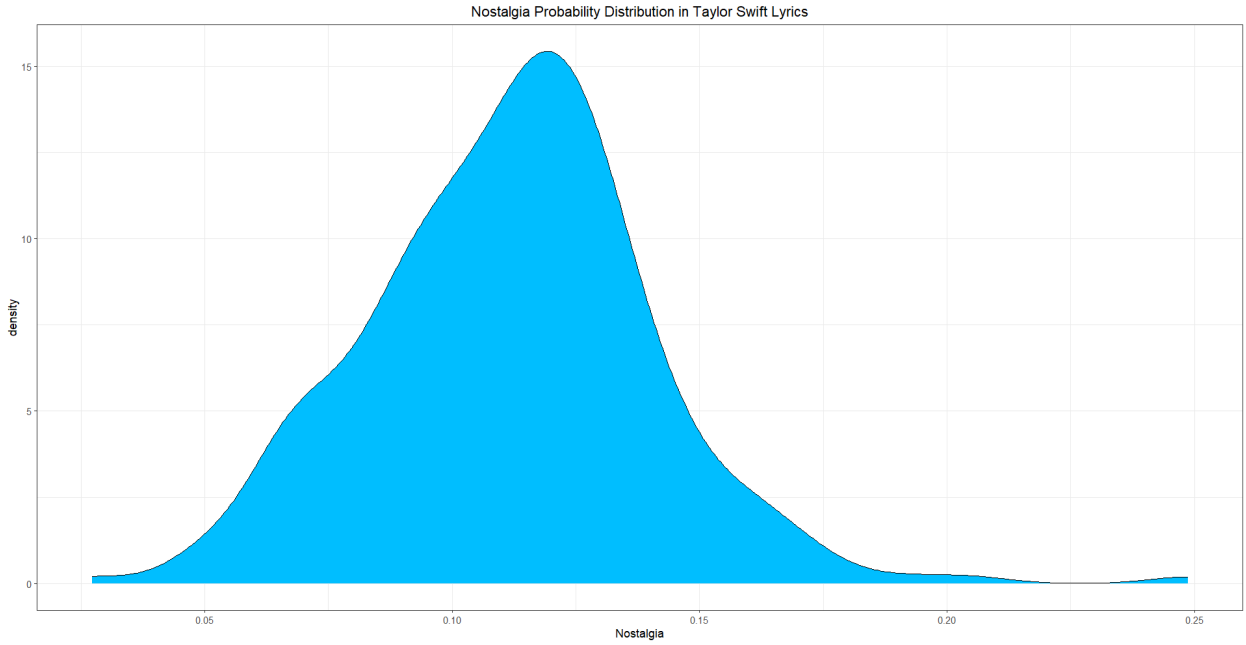

Hope
Year
0.00
0.05
0.10
0.15
0.20
2008
2010
2013
2017
2019
2021
2022
2024

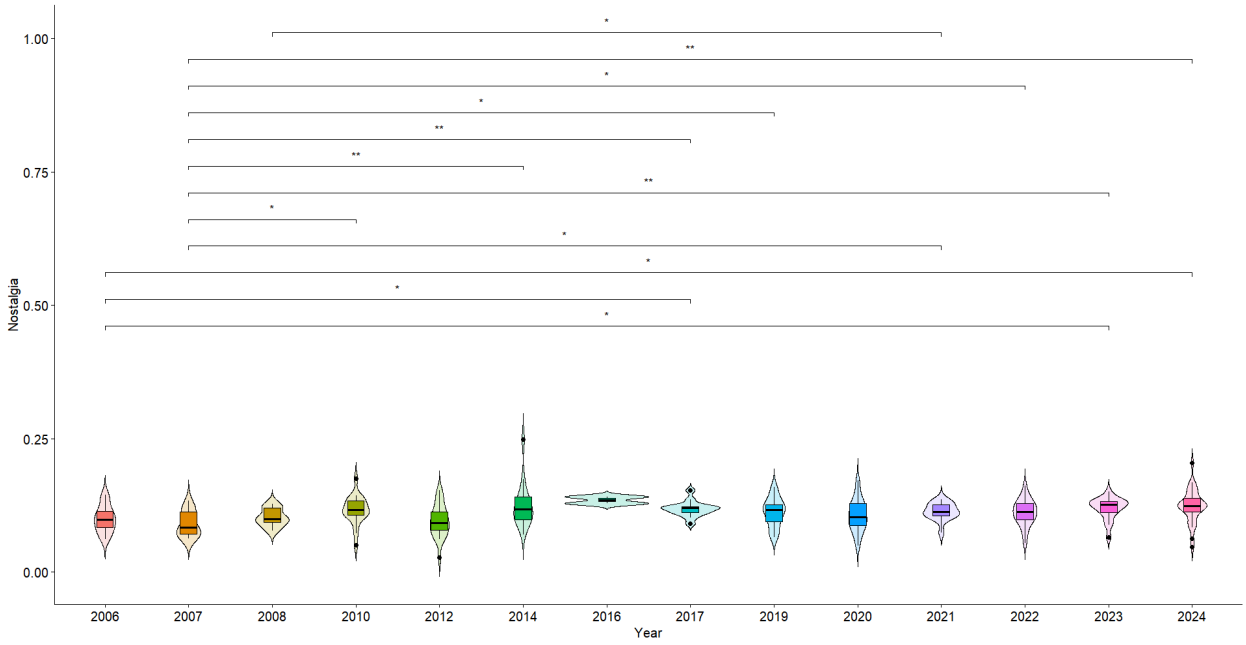
Loss:


Loss Probability Distribution in Taylor Swift Lyrics
density
Loss
10.0
7.5
5.0
2.5
0.0
0.1
0.2
0.3
0.4

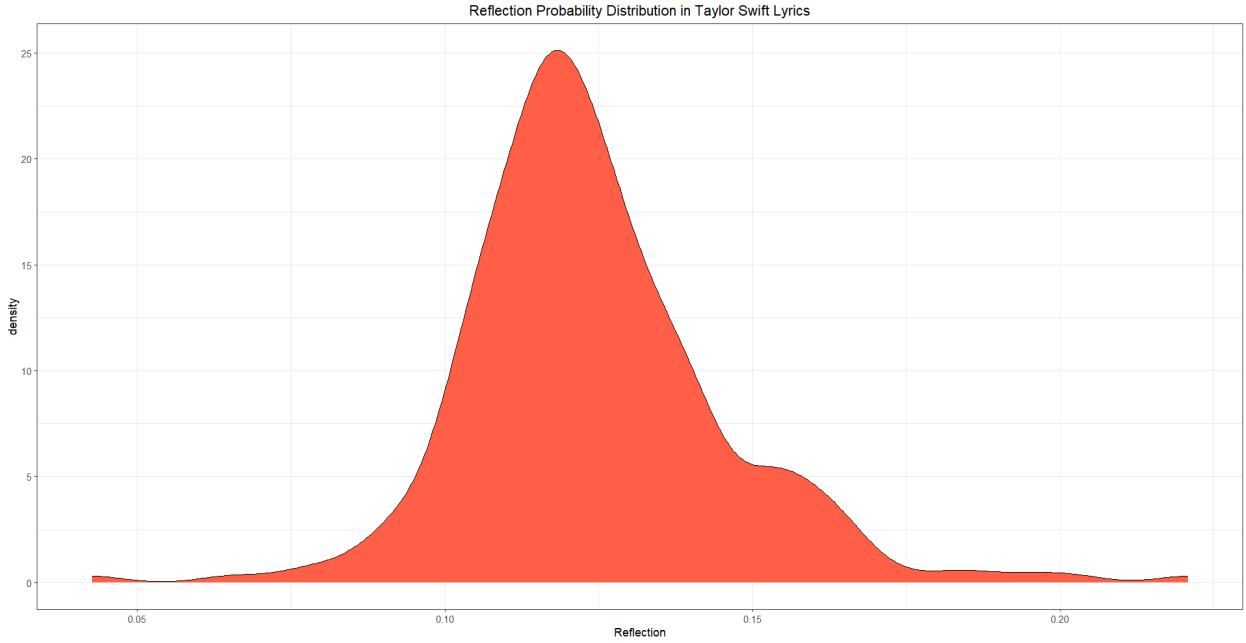

***
**
**
ns
Loss
Year
0.6
0.4
0.2
0.0
2006
2007
2008
2009
2010
2012
2017
2019
2020
2021
2022
2023
2024

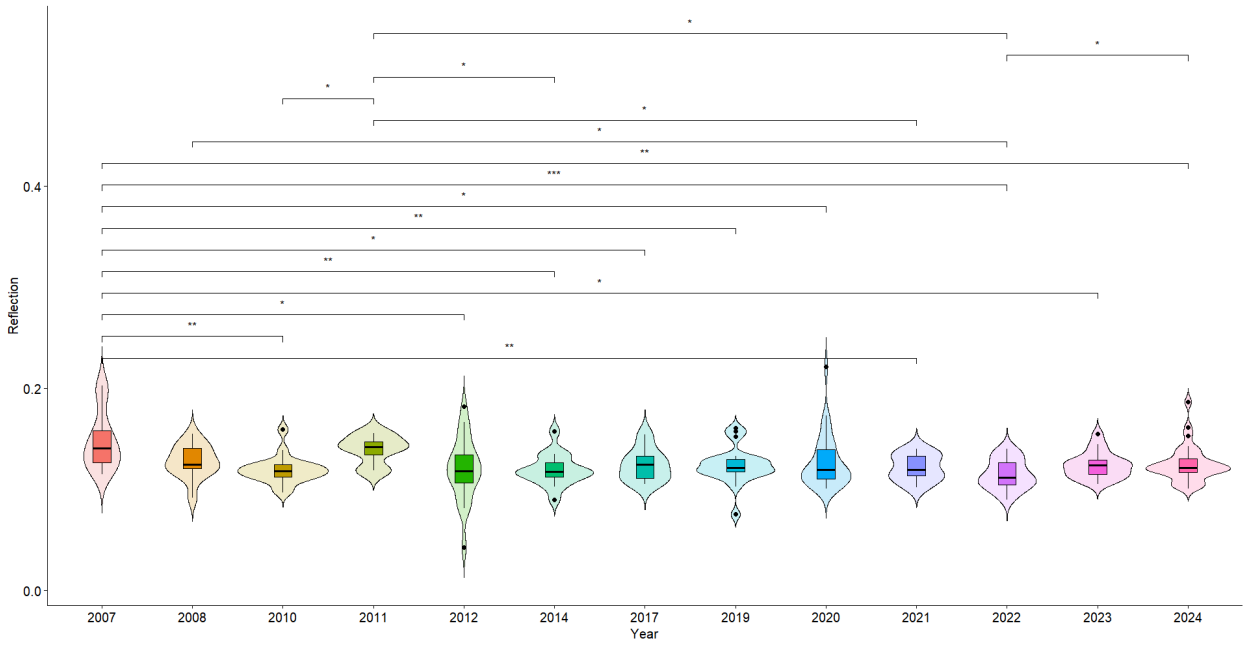
Nostalgia:


Nostalgia Probability Distribution in Taylor Swift Lyrics
density
Nostalgia
0
5
10
15
0.05
0.10
0.15
0.20
0.25

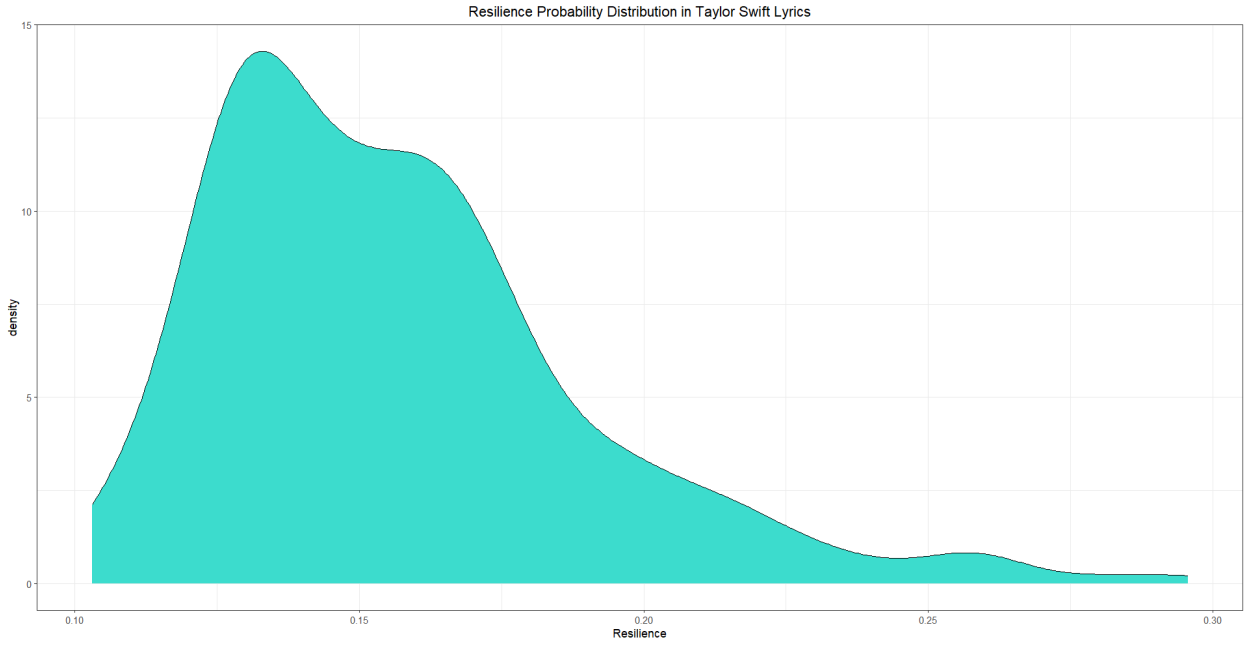

Nostalgia
Year
0.00
0.25
0.50
0.75
1.00
2006
2007
2008
2010
2012
2014
2016
2017
2019
2020
2021
2022
2023
2024

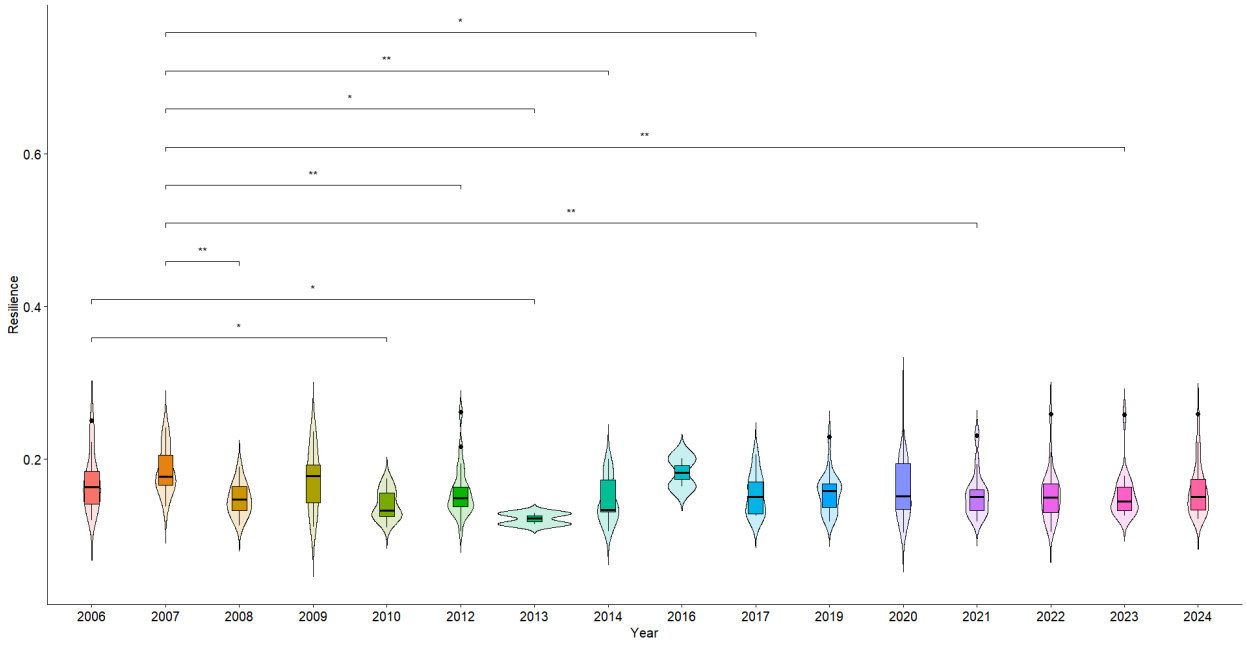
Reflection:


Reflection Probability Distribution in Taylor Swift Lyrics
density
Reflection
0
5
10
15
20
25
0.05
0.10
0.15
0.20

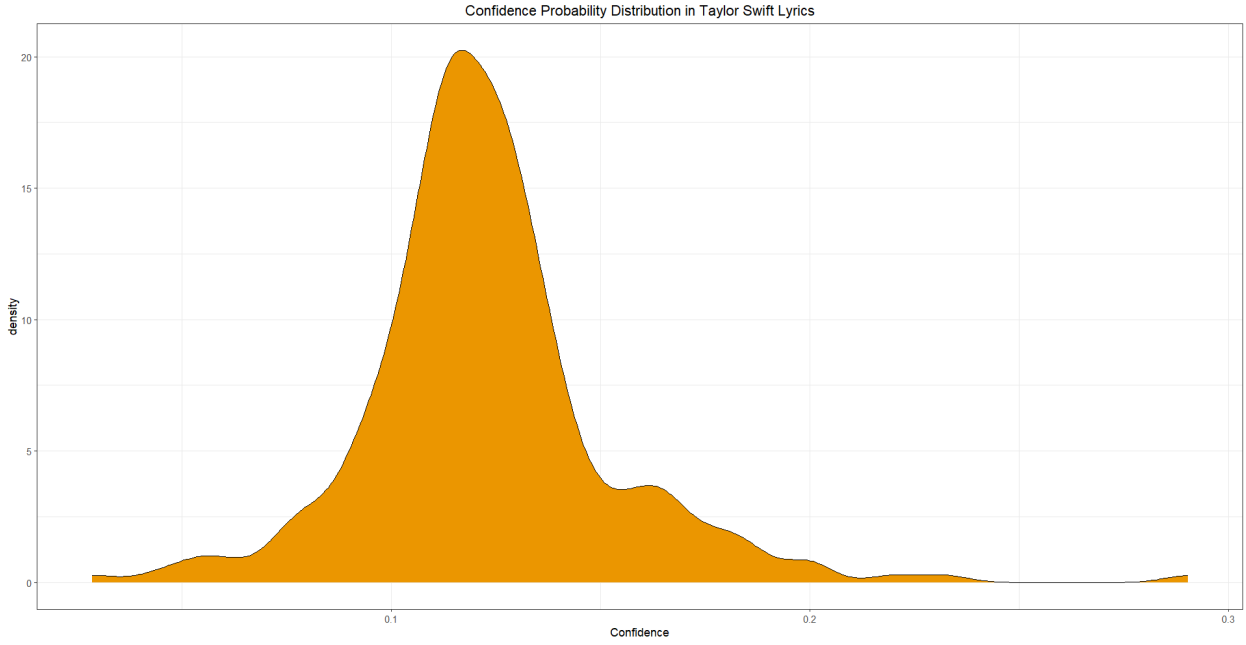

Reflection
Year
0.0
0.2
0.4
2007
2008
2010
2011
2012
2014
2017
2019
2020
2021
2022
2023
2024

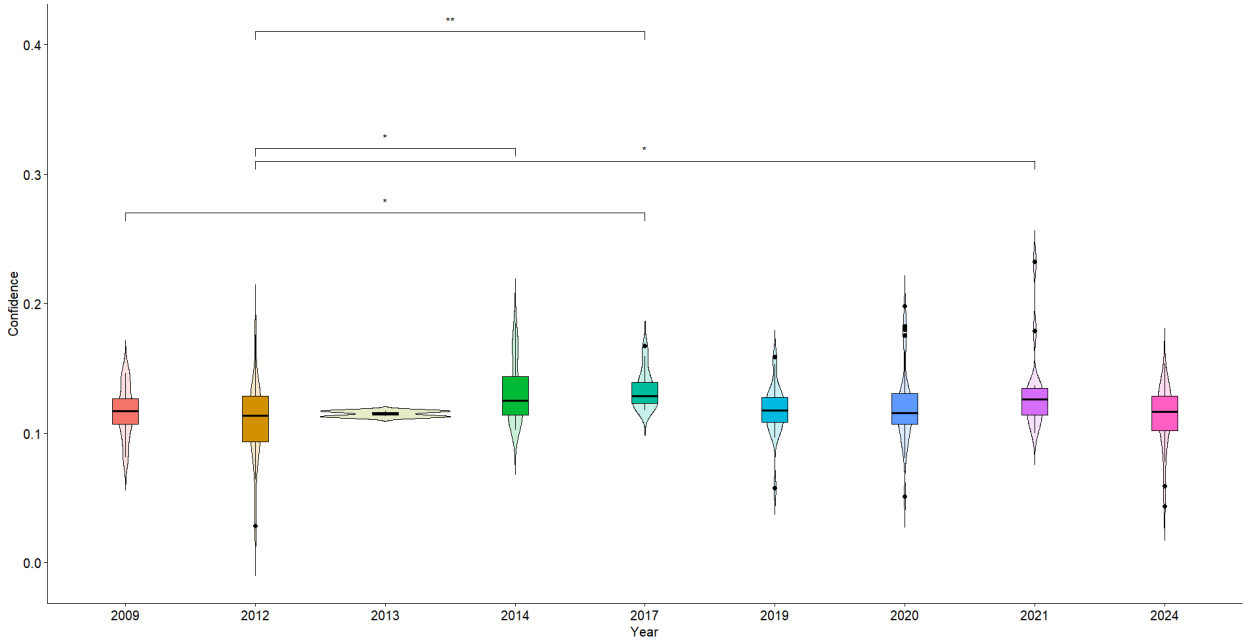
Resilience:


Resilience Probability Distribution in Taylor Swift Lyrics
density
Resilience
0
5
10
15
0.10
0.15
0.20
0.25
0.30


Resilience
Year
0.2
0.4
0.6
2006
2007
2008
2009
2010
2012
2013
2014
2016
2017
2019
2020
2021
2022
2023
2024
*
**
*
**
**
**
**
*
*

Confidence:


Confidence Probability Distribution in Taylor Swift Lyrics
density
Confidence
0
5
10
15
20
0.1
0.2
0.3


Confidence
Year
0.0
0.1
0.2
0.3
0.4
2009
2012
2013
2014
2017
2019
2020
2021
2024
**
*
*
*