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

LOAD AND VIEW DATA
Load data
Select variables of interest
See the variables in and variable type of dataframe
Rows: 359
Columns: 4
$ category <chr> "Taylor Swift", "Taylor Swift", "Taylor Swift", "Taylor Swi…
$ song_title <chr> "Tim McGraw", "Picture to Burn", "Teardrops On My Guitar", …
$ song_lyrics <chr> "['He said the way my blue eyes shined', 'Put those Georgia…
$ song_release_date <date> 2006-06-19, 2006-10-24, 2006-10-24, 2006-10-24, 2006-10-24…

Rename columns
Select year from date
Max year: 2024
Min year: 2006

Variable	Class	Description
text	character	The text variable containing 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	character	The year in which the song was released

#What are the different categories/albums in the dataset?

[1] "Taylor Swift" "Fearless"
[3] "Fearless (TV)" "Speak Now"
[5] "Speak Now (TV)" "Red"
[7] "Red (TV)" "1989"
[9] "1989 (TV)" "reputation"
[11] "Lover" "folklore"
[13] "Non-Album Songs" "evermore"
[15] "Midnights" "The Tortured Poets Department"

[17] "Other Artist Songs"

#Other Artist Songs will not be included in clean dataframe. Analysis is on Taylor Swift, however, collaborations are kept.

DATA CLEANING

#Select songs from each original album release
#Select songs that are Taylor's Version
#Select only new songs/songs that do not appear in previous release
#Keep songs with that are Taylor's Version and From Vault

#Bind the rows that contain unique songs into dataframe

EXPLORATORY DATA ANALYSIS

Load and glimpse at clean data
Rows: 266
Columns: 6
$ X <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 2…
$ category <chr> "Taylor Swift", "Taylor Swift", "Taylor Swift", "Taylor Swift", "Tay…
$ title <chr> "Tim McGraw", "Picture to Burn", "Teardrops On My Guitar", "A Place …
$ text <chr> "['He said the way my blue eyes shined', 'Put those Georgia stars to…
$ date <chr> "2006-06-19", "2006-10-24", "2006-10-24", "2006-10-24", "2006-10-24"…
$ year <int> 2006, 2006, 2006, 2006, 2006, 2006, 2006, 2006, 2006, 2006, 2006, 20…

What albums are in new dataframe
[1] "Taylor Swift" "Fearless"
[3] "Fearless (TV)" "Speak Now"
[5] "Speak Now (TV)" "Red"
[7] "Red (TV)" "1989"
[9] "1989 (TV)" "reputation"
[11] "Lover" "folklore"
[13] "evermore" "Midnights"
[15] "The Tortured Poets Department" "Non-Album Songs"

Function to keep only visible text
Convert data into vector and transform into corpus. Make words in text suitable for analysis
Keep only visible text in corpus
Convert all text to lowercase

Remove punctuation
Remove numbers from text
Remove english stop words like "is", "the", "a" etc

Convert corpus into a document term matrix

<<DocumentTermMatrix (documents: 266, terms: 5106)>>
Non-/sparse entries: 23050/1335146
Sparsity : 98%
Maximal term length: 25
Weighting : term frequency (tf)

High sparsity means many words that are not high in frequency
Reduce sparsity to 0.965
View reduced sparsity document term matrix. Sparsity at 0.89

<<DocumentTermMatrix (documents: 266, terms: 496)>>
Non-/sparse entries: 13703/118233
Sparsity : 90%
Maximal term length: 10
Weighting : term frequency (tf)

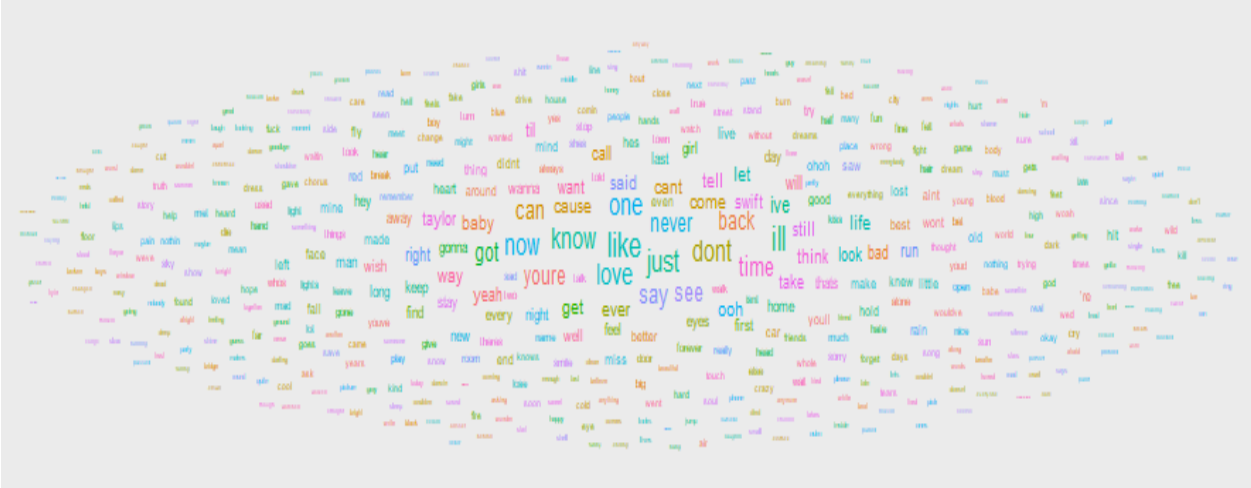
Convert document term matrix into matrix then into dataframe
Create frequency and sort/Create frequency of tokens/words in document term matrix
Make a dataframe of the sorted tokens and their frequencies with columns "word" and "frequency"
View data
Most common words are like, know, just, don't and never.
What are the common words associated with? What words are likely to come before or after frequent words?
like associated with feels and snow
$like
feels snow dream comin stars scene falling
0.49 0.44 0.40 0.35 0.34 0.33 0.31

know associated with changed and better
$know
changed better since held everything
0.58 0.55 0.48 0.36 0.33

never associated with street and walk
$never
street walk
0.35 0.33

love associated with free and came
$love
free came ohoh
0.52 0.40 0.34

Visualization
#Wordcloud of frequent words



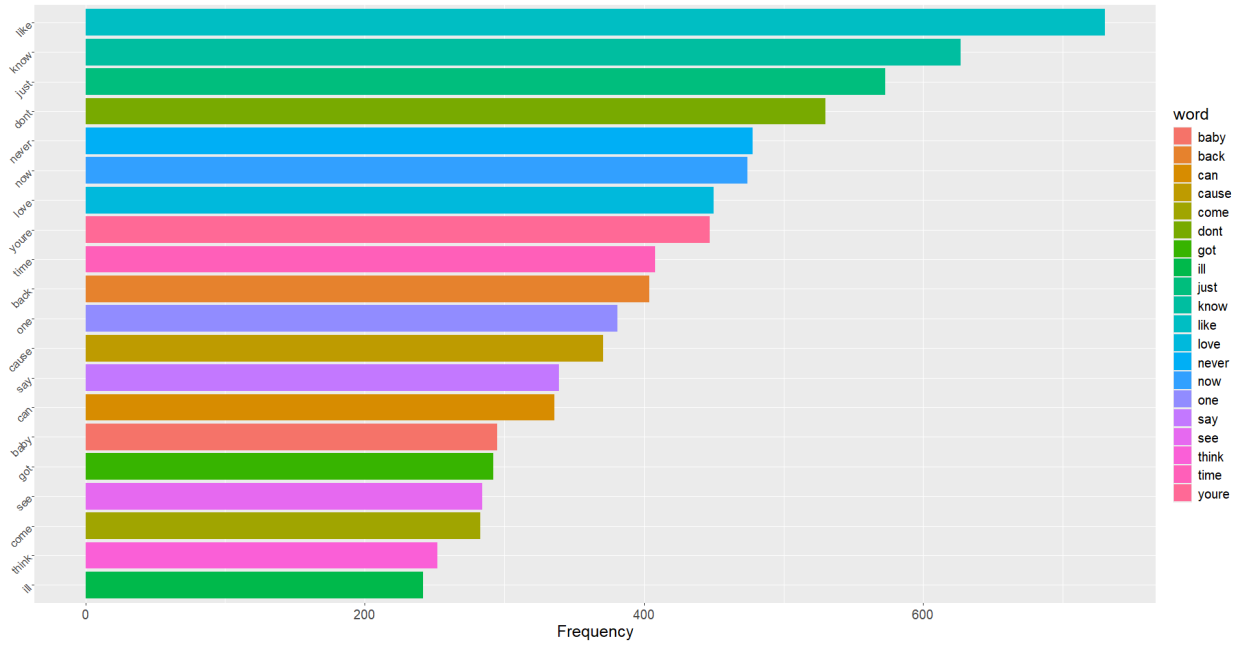
#What are the top 20 words in the lyrics?
#Select 20 rows from the frequency dataframe

	word	freq
like	like	730
know	know	627

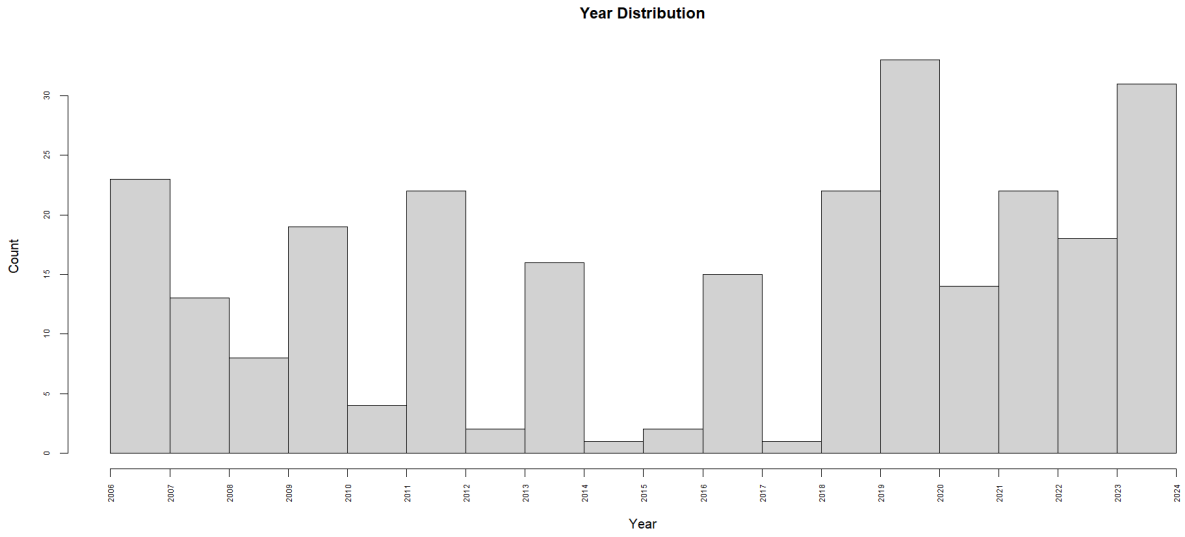
just	just	573
dont	dont	530
never	never	478
now	now	474
love	love	450
youre	youre	447
time	time	408
back	back	404
one	one	381
cause	cause	371
say	say	339
can	can	336
baby	baby	295

got	got	292
see	see	284
come	come	283
think	think	252
ill	ill	242

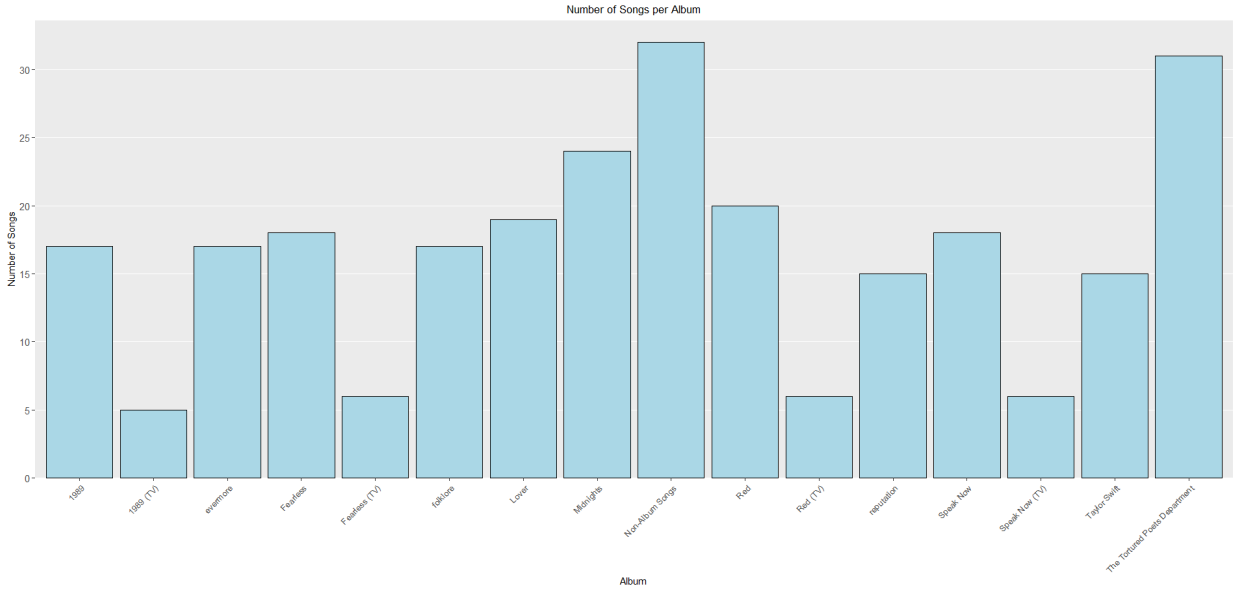
Showing 1 to 20 of 20 entries, 2 total columns



Year Distribution in dataframe



Song Distribution by Album



SENTIMENT ANALYSIS

- Only relative variables in table

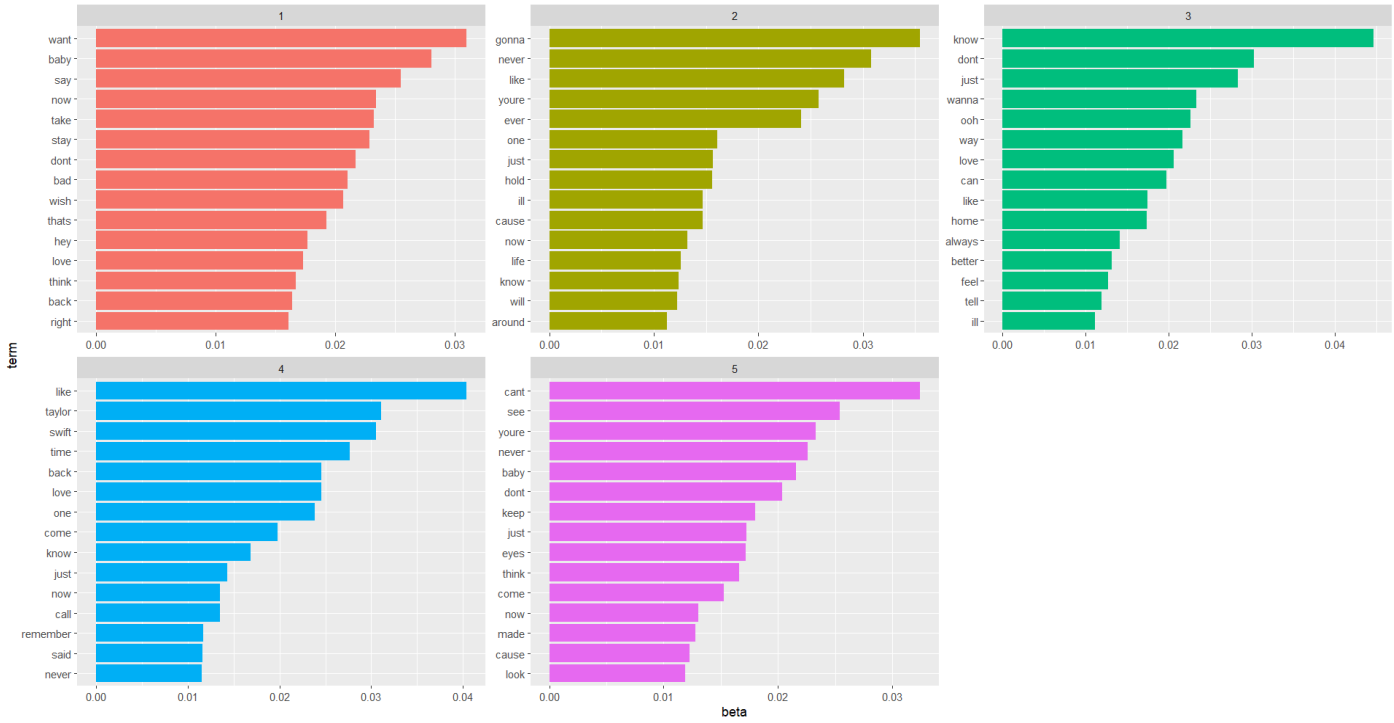
Variable	Class	Description
Text	character	The text variable containing the lyrics to corresponding Taylor Swift song
Album	character	The album each song is from
Title	character	The title of the song
Date	date	The date on which each song was released including year, month, and day
Year	character	The year in which the song was released
Sentiment HE/LM/QDAP	numeric	The sum of positive and negative sentiment for each song in each library in the Sentiment Analysis package (score between -1 and 1)
Positivity HE/LM/QDAP	numeric	The positive sentiment score for each song in each library in the Sentiment Analysis package (score between 0 and 1)
Negative HE/LM/QDAP	numeric	The positive sentiment score for each song in each library in the Sentiment Analysis package (score between -1 and 0)

#Perform sentiment analysis on text column
bind columns from sentiment dataframe to current dataframe

TOPIC MODELING

Topic modeling using topic model
Measures probability of text belonging to topic
Higher beta reflects higher probability of text belonging to specific network

Group tokens into highest probability topics
Visualize topics tokens belong



Write out grouped topics to use for next step

GEMINI API AND PYTHON

Import all libraries and dependencies
Load in dataframe from topic modeling
Connect to gemini api using key
Choose model of gemini to work with
Convert values in topics to string
Prompt gemini
Print response for prompt

Response:
Here's a breakdown of the themes and emotions associated with each topic in your dataframe:

Topic 1: Everyday Conversations
#
* **Theme:** General, casual conversations.
* **Emotions/Subjects:** Friendship, connection, simple exchanges, everyday experiences, lightheartedness. Words like "can", "like", "got", "don't", "look", "stay", "way" all point to casual interactions.

#
Topic 2: Romantic Yearning/Regret
#
* **Theme:** Longing, love, loss, reflection on past relationships.
* **Emotions/Subjects:** Hope, sadness, nostalgia, yearning for something lost. Words like "back", "baby", "come", "ever", "wish", "never", "like", "said", "right", "cause", "want" all hint at romantic themes.
#
Topic 3: Love and Loss
#
* **Theme:** Heartbreak, reflection on love, the passage of time.
* **Emotions/Subjects:** Sadness, melancholy, introspection, time passing, remembering. Words like "never", "love", "time", "still", "think", "like", "now", "didn't", "just", "know", "heart", "said", "good" are associated with feelings of loss and reflection.
#
Topic 4: Optimism and Confidence
#
* **Theme:** Positive outlook, self-assurance, determination.
* **Emotions/Subjects:** Hope, excitement, resilience, belief in oneself, making progress. Words like "know", "just", "like", "see", "gonna", "youre", "one", "keep", "cause", "never", "eyes", "better", "man", "now", "yeah" all convey a sense of confidence and forward momentum.
#
Topic 5: Pop Culture/Fan Culture
#
* **Theme:** References to a specific artist or genre (possibly Taylor Swift).
* **Emotions/Subjects:** Enthusiasm, admiration, belonging to a fandom. Words like "don't", "like", "taylor", "ooh", "swift", "wanna", "just", "ill", "say", "call", "want", "time", "know", "one", "youre" suggest a connection to a particular artist or genre.
#
Important Note: The context of the data is crucial for a more accurate interpretation. For example, the "Taylor Swift" topic might be linked to a specific song or era of her music.
#
Tips for Deeper Analysis:
#
* **Sentiment Analysis:** Use tools to analyze the emotional tone of text associated with each topic.
* **Network Analysis:** Visualize how words within each topic connect to create semantic networks.
* **Additional Context:** Look for any other available information about the data source, like genre, audience, or timeframe.

ZERO SHOT CLASSIFICATION

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 HE/LM/QDAP	numeric	The sum of positive and negative sentiment for each song in each library in the Sentiment Analysis package (score between -1 and 1)
Positivity HE/LM/QDAP	numeric	The positive sentiment score for each song in each library in the Sentiment Analysis package (score between 0 and 1)
Negative HE/LM/QDAP	numeric	The positive sentiment score for each song in each library in the Sentiment Analysis package (score between -1 and 0)
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

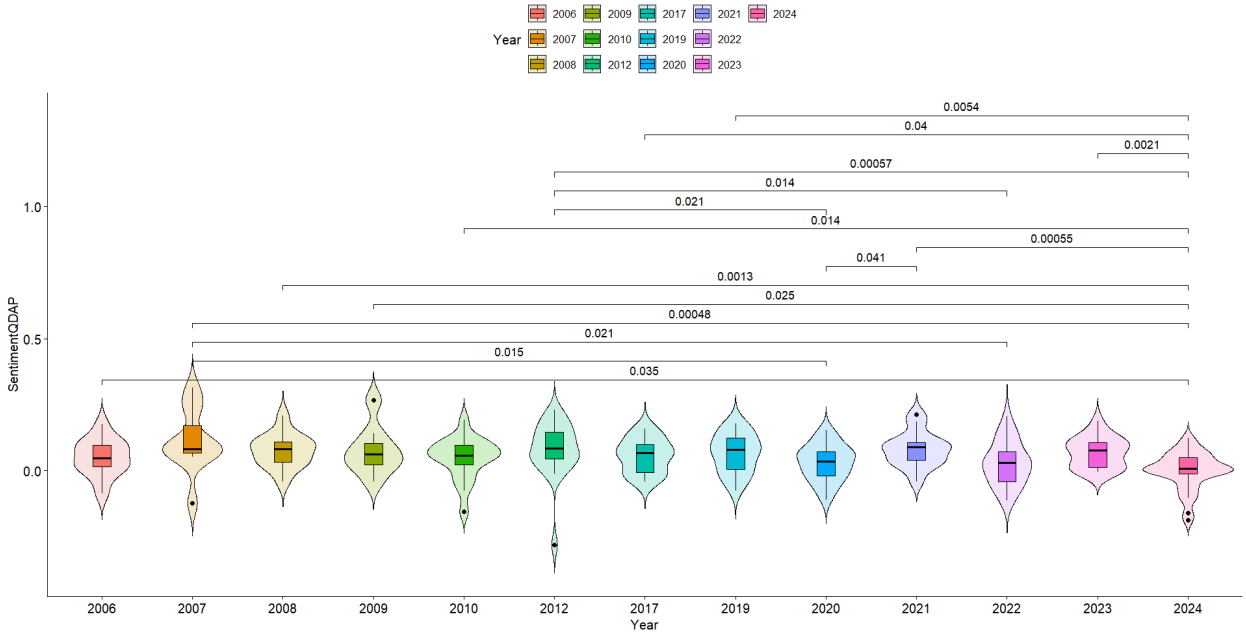
Using the themes and topics generated from gemini and topic modeling, use zero shot classification to generate scores for each theme.

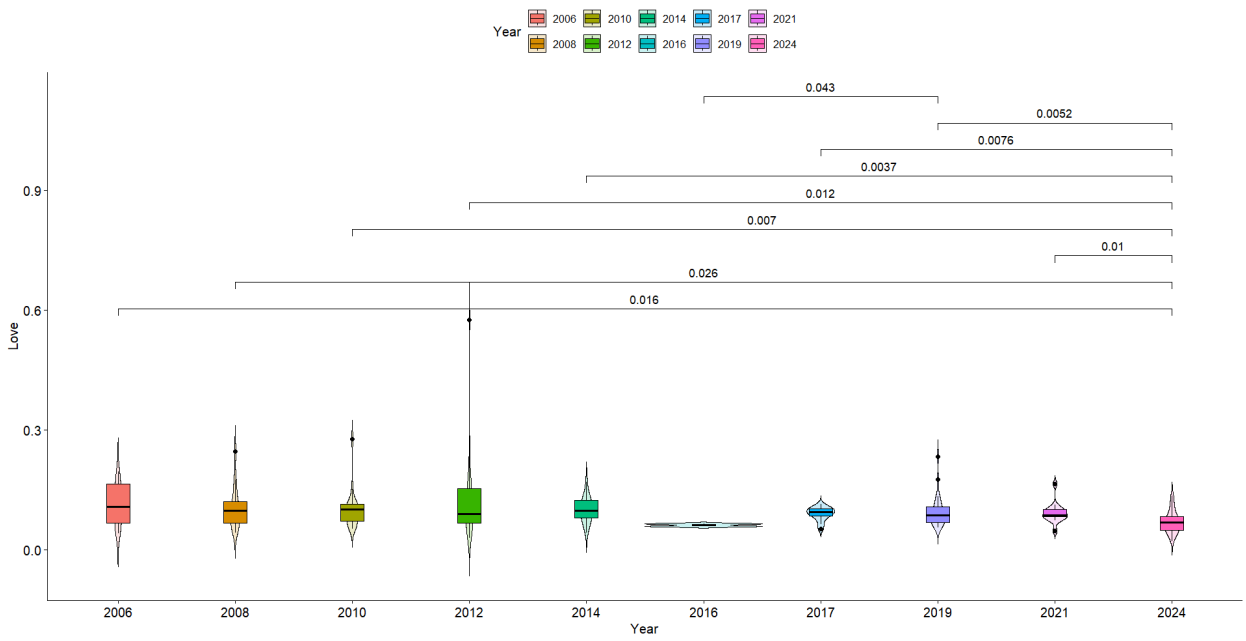
MEAN COMPARISON, SIGNIFICANCE TESTING AND VISUALIZATION

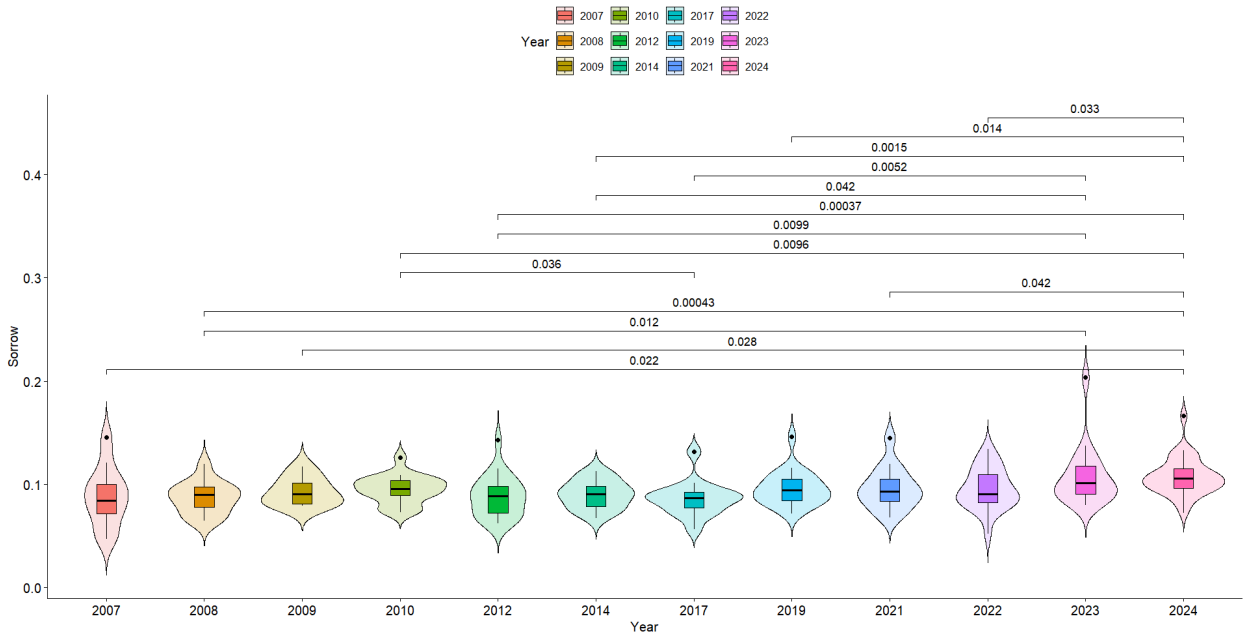
None of the results compared by year is significant for LM
Some significance in QDAP
General trend in significance: More significance in sentiments in earlier years when compared to later years
What years show significant differences when compared with each other?
Select only years where there were significant results for each sentiment and emotional theme
Clean data for significance visualization

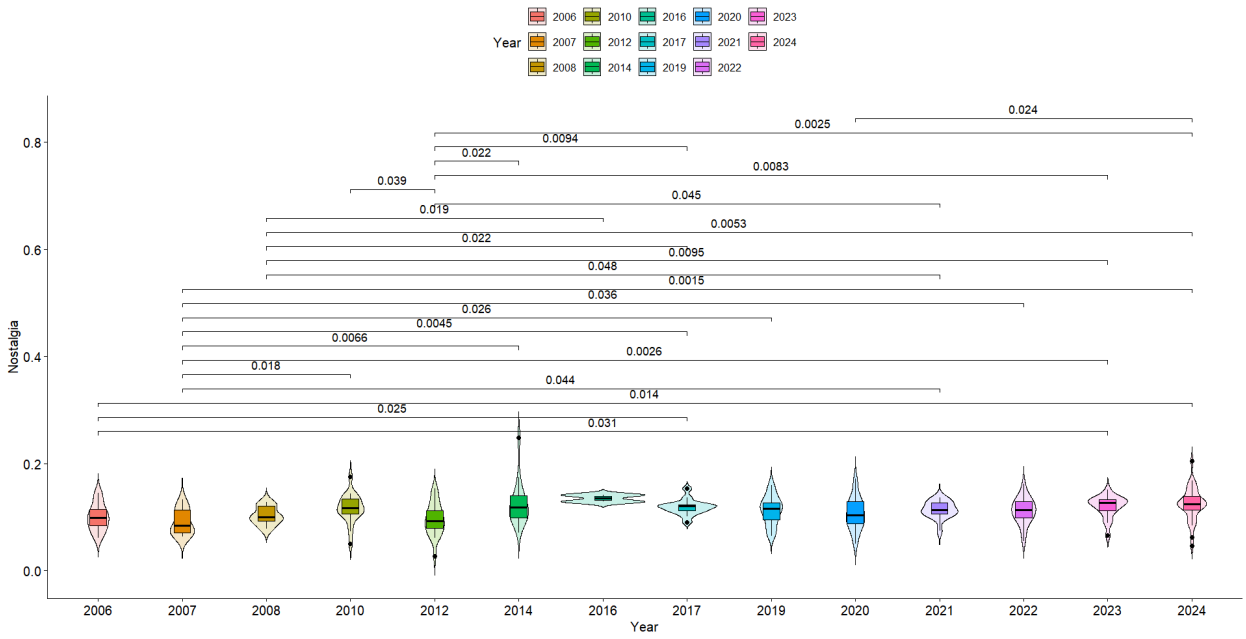
Save all variables of interest into one dataframe for visualization
Make dataframe for sentiment and topics selecting years where topic was significant
Make list of matched years in for sentiment and topic to use for visualization

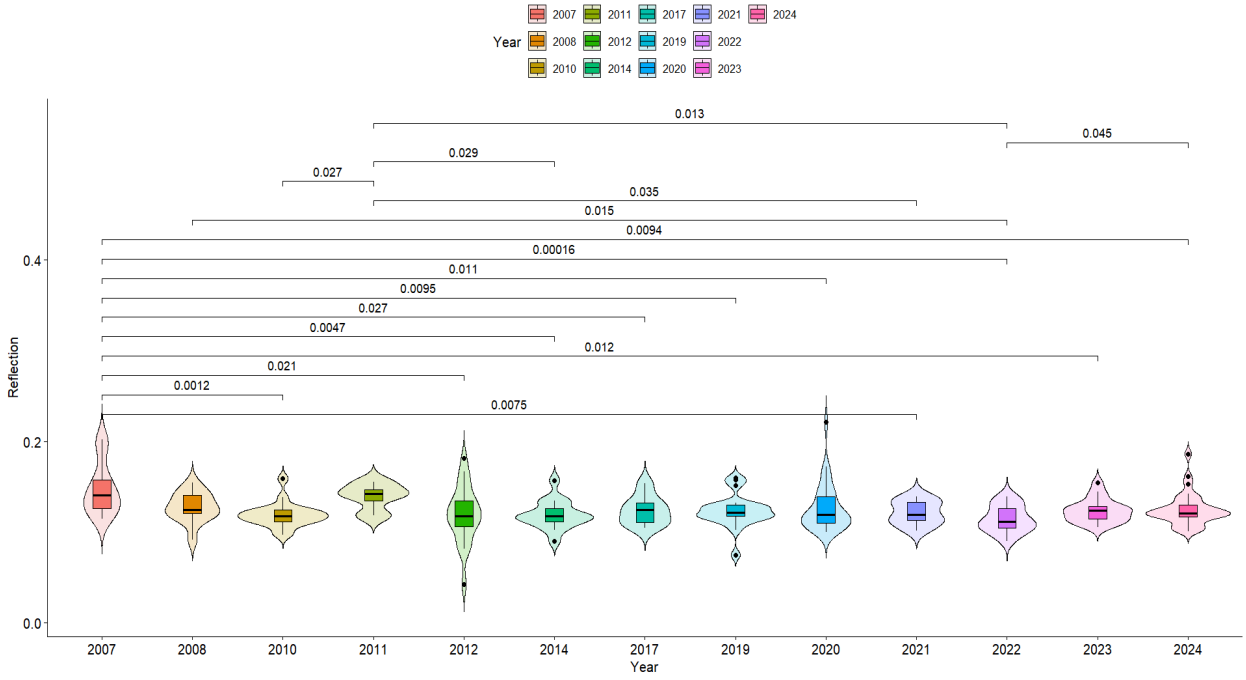
Visualize sentiment and topic significance by matched years

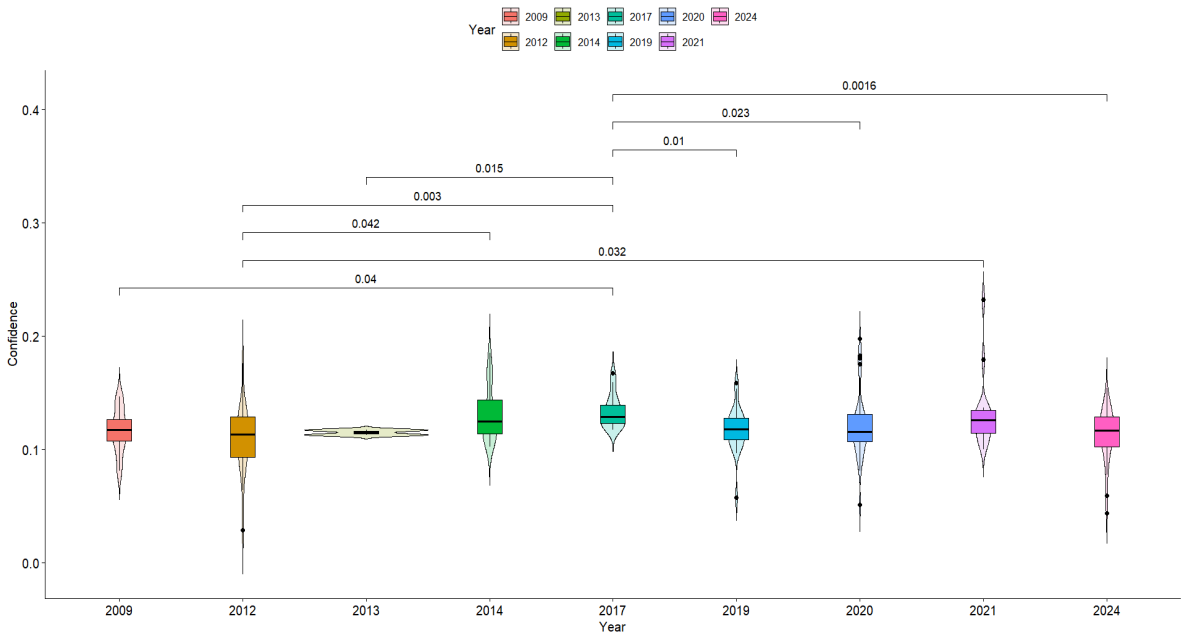


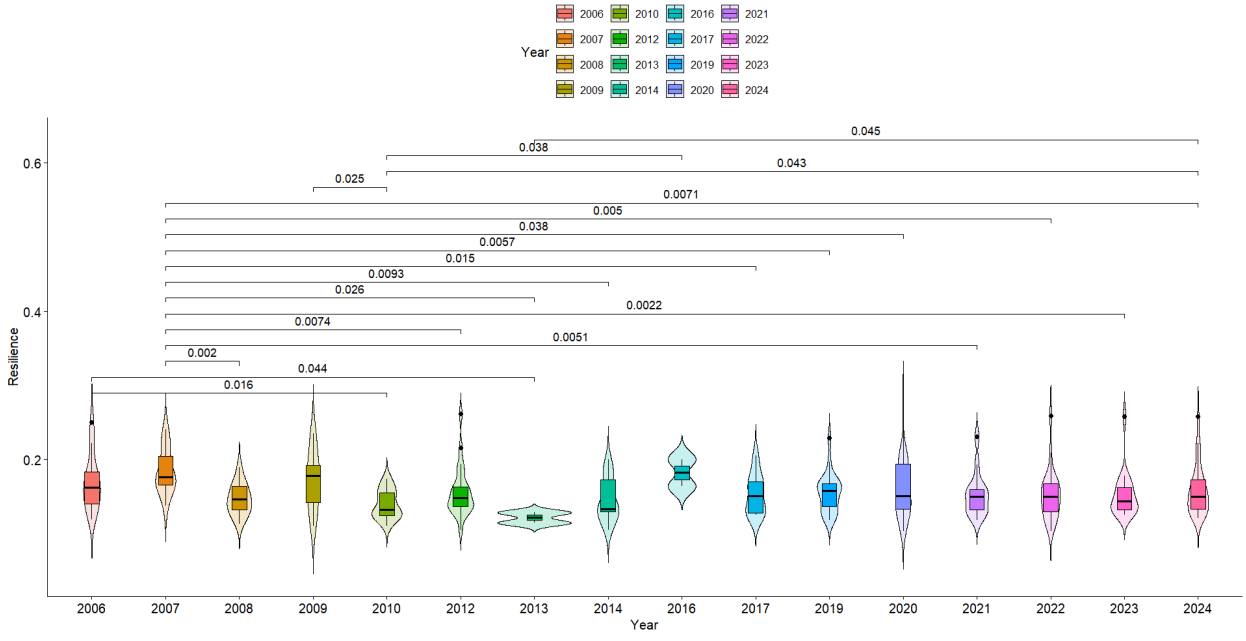

Year
2006
2008
2010
2012
2014
2016
2017
2019
2021
2024
0.043
0.0052
0.0076
0.0037
0.012
0.007
0.01
0.026
0.016
Love
0.0
0.3
0.6
0.9
2006
2008
2010
2012
2014
2016
2017
2019
2021
2024
Year

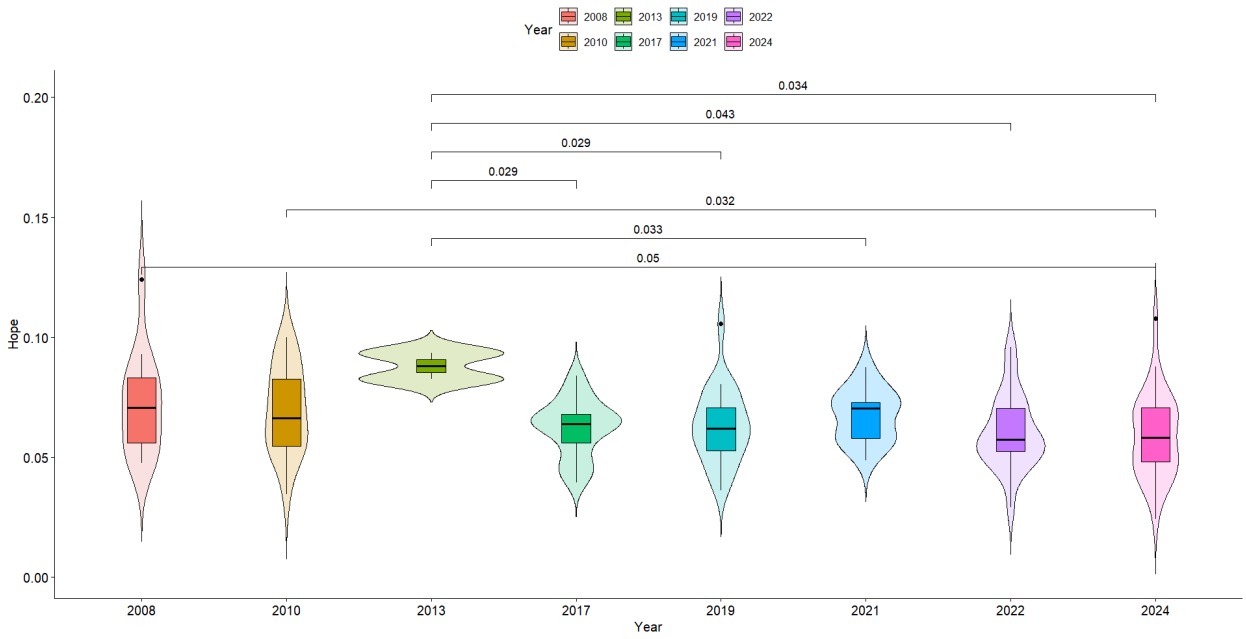

Year
2007
2008
2009
2010
2012
2014
2017
2019
2021
2022
2023
2024
0.033
0.014
0.0015
0.0052
0.042
0.00037
0.0099
0.0096
0.036
0.042
0.00043
0.012
0.028
0.022
Sorrow
0.0
0.1
0.2
0.3
0.4
2007
2008
2009
2010
2012
2014
2017
2019
2021
2022
2023
2024
Year

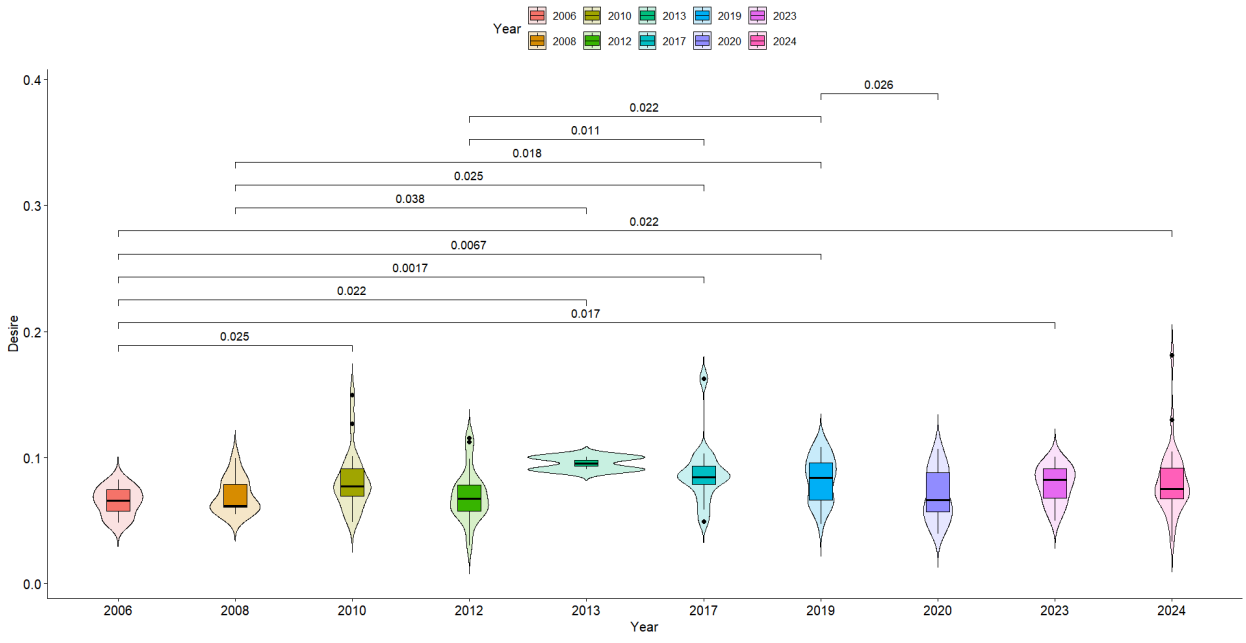

Year
2006
2007
2008
2010
2012
2014
2016
2017
2019
2020
2021
2022
2023
2024
0.024
0.0025
0.0094
0.022
0.0083
0.039
0.045
0.019
0.0053
0.022
0.0095
0.048
0.0015
0.036
0.026
0.0045
0.0066
0.0026
0.018
0.044
0.014
0.025
0.031
Nostalgia
0.0
0.2
0.4
0.6
0.8
Year

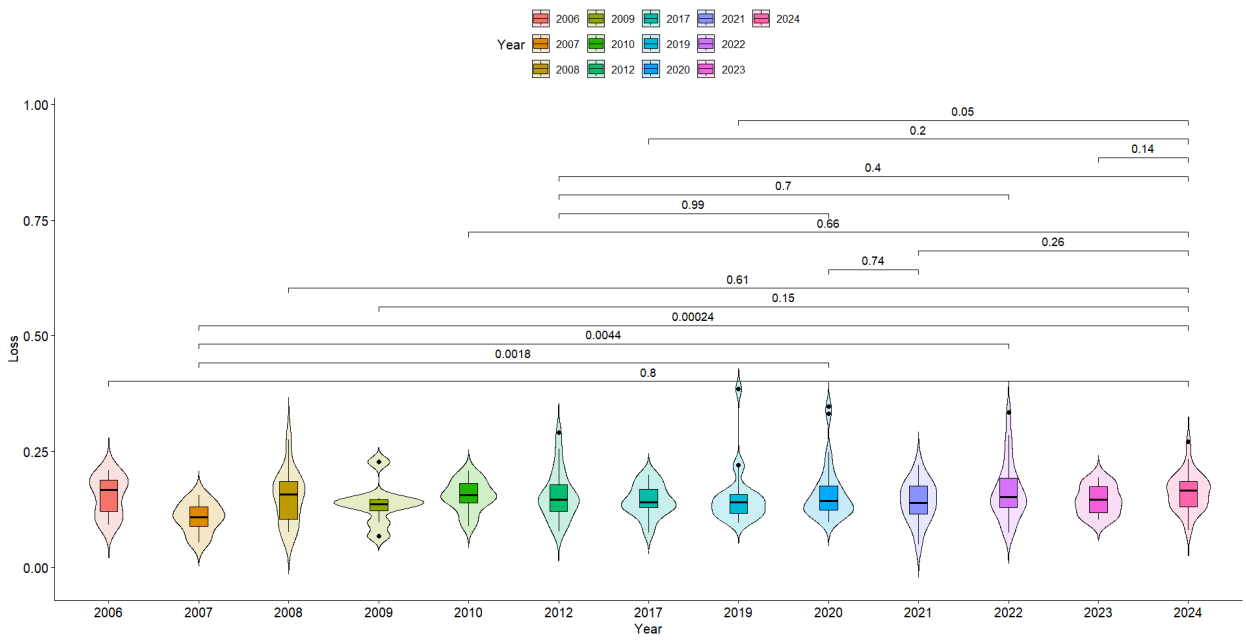

Year
2007
2008
2010
2011
2012
2014
2017
2019
2020
2021
2022
2023
2024
0.013
0.045
0.029
0.027
0.035
0.015
0.0094
0.00016
0.011
0.0095
0.027
0.0047
0.012
0.021
0.0012
0.0075
Reflection
0.0
0.2
0.4
Year


Year
2009
2012
2013
2014
2017
2019
2020
2021
2024
0.0016
0.023
0.01
0.015
0.003
0.042
0.032
0.04
Confidence
0.0
0.1
0.2
0.3
0.4
2009
2012
2013
2014
2017
2019
2020
2021
2024
Year


Year
2006
2007
2008
2009
2010
2012
2013
2014
2016
2017
2019
2020
2021
2022
2023
2024
0.045
0.038
0.043
0.025
0.0071
0.005
0.038
0.0057
0.015
0.0093
0.026
0.0022
0.0074
0.0051
0.002
0.044
0.016
Resilience
0.2
0.4
0.6
2006
2007
2008
2009
2010
2012
2013
2014
2016
2017
2019
2020
2021
2022
2023
2024
Year


Year
2008
2010
2013
2017
2019
2021
2022
2024
0.034
0.043
0.029
0.029
0.032
0.033
0.05
Hope
0.20
0.15
0.10
0.05
0.00
2008
2010
2013
2017
2019
2021
2022
2024
Year


Year
2006
2008
2010
2012
2013
2017
2019
2020
2023
2024
0.026
0.022
0.011
0.018
0.025
0.038
0.022
0.0067
0.0017
0.022
0.017
0.025
Desire
0.4
0.3
0.2
0.1
0.0
2006
2008
2010
2012
2013
2017
2019
2020
2023
2024
Year


Year
2006
2007
2008
2009
2010
2012
2017
2019
2020
2021
2022
2023
2024
Loss
1.00
0.75
0.50
0.25
0.00
0.05
0.2
0.14
0.4
0.7
0.99
0.66
0.26
0.74
0.61
0.15
0.00024
0.0044
0.0018
0.8
Year