Data Analysis Report on Billboard Dataset

Ajit Mallavarapu, Sean Lee

September 12 2024

Contents

Introduction 2
Funding and RA Team 3
Data Collection 3
Gathering Billboard Top 100 Songs 3
Genre Data Collection 3
Data Cleaning and Preparation 4
Adding Missing Songs, Removing Duplicate Songs 4
Adding Missing Values for Bands 4
Added Missing Genres 5
Fixing Some Errors in the "singer_songwriter" Column 5
Fixing Typos + Standardization 8
Creating a Split Dataset 8
Combining Songwriters - Concatenating & Removing Duplicates 9
Python code for the combined songwriter column for non-split data: 9
Python code used to create the combined songwriter column for the split data 10
Cleaning Trailing Commas, nan's, and Double Commas 11
Methodology for Graphs 12
Average Number of Credited Songwriters per Chart-Topping Song 12
Tableau: 13
PowerBI: 13
Average Number of Credited Songwriters per Chart Topping Song for Pop and Hip Hop vs Year 13
Tableau: 14
PowerBI: 14
Python using Plotnine: 14
% of Top 100 in which the Artist is also Credited as a Songwriter 16
Tableau: 16
PowerBI: 17

Distribution of Genres 17
Tableau: 18
PowerBI: 18
Distribution of Genres - Hip Hop & Rap 18
Tableau: 19
PowerBI: 19
Band vs. Non-Bands For Being Credited as Songwriters 20
PowerBI: 20
Anomalies and Trends 20
Anomalies 20
Trends 20
Findings 20
Warnings and Disclosures 21
Conclusion 21
Acknowledgement 21
References 21
Contact Email 22

Introduction

Data was collected for the journal article written by Dr Emma Perot, 'Music Copyright Ownership: Factors Behind the Increase in Writer Credits and Rights Clearance', published in the Berkeley Journal of Entertainment and Sports Law 2025, volume 14.

The data includes: song rank, song title, artist, songwriters, genre, whether the artiste is a band, whether the artiste is listed as a songwriter. The purpose of data collection was to examine the purported trend in increase in number of songwriters and instances of artistes being named as songwriters for year-end Billboard chart-topping songs. Data confirmed these trends and doctrinal research was conducted to analyze the key copyright cases as well as the changes in the music industry spurred on by listening tastes and technological developments that contributed to the trends. The pre-publication draft of the article can be viewed on SSRN here: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4947914.

This report details the data collection process, the modifications made to the original Billboard dataset, the methodologies employed for each graph, and discusses anomalies, trends, findings, warnings, and disclosures encountered during the analysis.

Funding and RA Team

The funding for the data collection was provided by the University of the West Indies St Augustine Campus (Trinidad and Tobago) Research and Publication Fund. The Research Assistants for this project were undergraduate law students at the UWI St Augustine: Solange Joseph (Head RA), Breana Babwah, Jessica Deonanan, Leah-Marie Lalsingh, Valini Ramkaran, Mikhila Sayers, Celine Rampaul, Catilin Alexander, and Priyanka Reemaul.

The authors of this data analysis report are: Ajit Mallavarapu BSc Informatics: Data Science with a Minor in Statistics and Sean Lee BSc Mathematics. We undertook data verification, cleaning, and created the visualizations used in the journal article. The authors were guided by Tazia Mehdi, MSc in Management Information Systems.

Data Collection

We give an overview of the data collection process done by Dr. Perot, her team, and Tazia Mehdi. We have also used their processes when adding missing songs and information into the data set when cleaning/preprocessing.

Gathering Billboard Top 100 Songs

The dataset consists of the Billboard Top 100 from 1959 onward, and year end top songs between 1946 to 1958. Prior to 1959, Billboard year-end charts contained 30 – 50 songs. The information collected for each song includes the song name, artist, genre, year, ASCAP songwriters, BMI songwriters, maximum number of songwriters, and whether the artist was a band, whether the artist was a songwriter. For the process of gathering the data, Dr Perot was assisted by a team of 9 research assistants. The names of songs, rank, and artist, were collected from the Billboard archive on Google Books [BillboardTop100 (2024)] as well as the Billboard Year-End Charts website [Billboard Year End 2006 onwards]. Of the 77 years examined, no official billboard records were located for 6 years: 1959, 1960, 1968, 1983, 1988, 1989. In these instances, alternative sources such as Spotify playlists, and the World Radio History website [World Radio History] were used.

Genre information was collected from Chosic. The ASCAP Repertory [ASCAP (2024)] and BMI repertoire [BMI (2024)] were used to find information about the ASCAP and BMI songwriters for each song.

Genre Data Collection

For more specific information about the genre, we used Chosic's Music Genre Finder [Chosic (2024)].

To analyze the genre distributions of the top 100 songs, Tazia used k-means clustering to standardize all of the genres of the songs to the following 15 genres: Pop, Hip Hop, Rap,

Country, Latin, Dance/Electronic, Funk, R&B, Rock, Jazz/Swing, Indie, Folk, Soul, Reggae, and Disco.

Standardised Genres	Pop	Hip Hop	Rap	Country	Latin	Dance/Electronic	Funk	R&B	Rock
Chosic Genres	pop	hip hop	rap	country	latin	electronic	funk	R&B	rock
	nederpop	alternative hip hop	brooklyn drill	country pop	cumbia villera	swedish electropop	latin funk	modern blues	alternative rock
	scandipop	atl hip hop	chicago drill	modern country pop	cumbia chilena	australian electropop	instrumental funk	blues	modern rock
	electrofox	underground hip hop	new york drill	classic country pop	phonk brasileiro	electropowerpop	brit funk	country blues	rock
	jpop	southern hip hop	florida drill	honky tonk	corrido	post-hardcore	new orleans funk	blues rock	modern alternative rock
	bubblegum dance	instrumental hip hop	drill	canadian country	bomba y plena	proto-techno	synth funk	jazz blues	future rock
	parody	old school hip hop	chicago drill	cowboy western	samba de roda	progressive electro house	p funk	modern blues rock	art rock
	classic girl group	minnesota hip hop	r&drill	traditional country	velha guarda	gaming edm	musica jibara	traditional blues	progressive rock
	bedroom pop	hardcore hip hop	jazz rap	country gospel	bossa nova	livetronica	g funk	rhythm and blues	country rock
	modern bollywood	conscious hip hop	freestyle	country blues	exotica	electro house	afrobeats	new orleans blues	dutch rock
	operatic pop	abstract hip hop	bronx drill	modern country rock	azonto	electro	afropop	british blues	modern southern rock
	swedish pop	seattle hip hop	philly drill	contemporary country	tejano	pixel	motown	southern soul blues	k-rock
	danish pop	east coast hip hop	rap conscient	deep contemporary country	contrabass	uk dance	g funk	louisiana blues	belgian rock
	k-pop	ohio hip hop	nyc rap	alternative country	kundiman	alternative dance	p funk	soul blues	swedish alternative rock
	swedish indie pop	detroit hip hop	underground rap	country dawn	violao	australian dance	synth funk	texas blues	southern rock
	dutch pop	bronx hip hop	chicago rap	uk americana	urbano mexicano	german dance	new orleans funk	delta blues	german rock
	classic swedish pop	latin hip hop	country rap	countrygaze	urbano latino	contemporary post-bop	jazz funk	electric blues	australian rock
	latin pop	political hip hop	pop rap	southern americana	swedish melodeath	italo dance	funktronica	memphis blues	irish rock
	classic uk pop	canadian old school hip hop	new orleans rap	classic oklahoma country	corridos tumbados	post-romantic era	neurofunk	acoustic blues	scottish rock
	australian pop	indie hip hop	rap rock	classic texas country	violao classico	speed house	liquid funk	chicago blues	christian rock

The standardized categorization of each genre into 15 *genres. Tazia used k-means clustering to determine the* 15 *clusters of standard genres.*

Data Cleaning and Preparation

Several modifications were made to the original dataset to ensure its accuracy and completeness. The following subsections detail the changes made:

Adding Missing Songs, Removing Duplicate Songs

Some songs were missing from the dataset. These were added manually based on verified sources.

We added the missing song: “I Wanna Be Around” by Tony Bennett, made in the year 1963 and was ranked 55-th place.

937	1963	54	Our Winter Love	Bill Pursell	Pop	JOHNNY COWELL	1		No	No	JOHNNY COWELL
938	1963	55	I Wanna Be Around	Tony Bennett	Jazz, Pop	Sadie Vimmerstedt, Johnny Mer	2		No	No	Johnny Mercer

The row for "I Wanna Be Around" has been added. We manually added the songwriters for the song: Sadie Vimmerstedt and Johnny Mercer.

We removed the song "Venus" by Lady Gaga, which was erroneously ranked 91-st place but was actually not in the Billboard Top 100.

This should give 100 songs for almost every year after 1960. The exception is the year 1969, where two songs were tied for 100-th place, giving 101 songs for that year’s top 100.

Adding Missing Values for Bands

There were missing values for whether the artist was a band or not. These songs were: **Having A Party, Rush Rush, Romantic, P.A.S.S.I.O.N., This House, Impulsive, Freedom 90, Jump, November Rain, Humpin Around, Hazard, Too Funky, Your Woman, Semi-Charmed Life, The Party Continues, The Way You Love Me, Classic**. We have filled those missing "Is this a band?" values in for them.

	Y		Song	Artist	Standardized	ASCAP Songwriters	I	BMI_Songwriters	Is this	si	Combined_Songwriter
835	1962	52	Having A Party	Sam Cooke	Jazz/Swing,Soul		1	COOKE SAM (BMI-	No	Yes	6711415),
3688	1991	4	Rush Rush	Paula Abdul	Pop,Rock	MORELAND PETER LORD	1		No	No	LORD (ASCAP-
3708	1991	24	Romantic	Karyn White	Jazz/Swing,Indie,		3	HARRIS JAMES	No	Yes	SAMUEL III (ASCAP-
3724	1991	40	P.A.S.S.I.O.N.	Rhythm Syndicate	Pop	CEJAS PAUL (ASCAP-	1		Yes	No	713815356),
3729	1991	45	This House	Tracie Spencer	Jazz/Swing,R&B		2	BUSH RICHARD J	No	No	(ASCAP-00074009690),
3732	1991	48	Impulsive	Wilson Phillips	Pop,Rock	CLIF MAGNESS, STEVE	2		Yes	No	KIPNER
3779	1991	95	Freedom 90	George Michael	Pop	MICHAEL GEORGE (GMR-	1		No	Yes	(GMR-125266882)
3787	1992	3	Jump	Kris Kross	Rap,Hip Hop,Pop	BERLIN IRVING (ASCAP-	13		Yes	No	2852023), DUPRI
3801	1992	17	November Rain	Guns N Roses	Rock	DUFF MCKAGAN, JEFFREY	4		Yes	Yes	JEFFREY ISBELL, SAUL
3817	1992	33	Humpin Around	Bobby Brown	Indie,R&B		5	BROWN ROBERT	No	Yes	BARRISFORD (ASCAP-
3840	1992	56	Hazard	Richard Marx	Pop,Rock	MARX RICHARD N (ASCAP-	1		No	Yes	(ASCAP-125724095),
3842	1992	58	Too Funky	George Michael	Pop	MICHAEL GEORGE (GMR-	1		No	Yes	(GMR-125266882),
4349	1997	65	Your Woman	White Town	Indie	CROSBY BING, WALLMAN	1	HOGAN DEBRA	No	No	WALLMAN IRVING,
4441	1998	57	Semi-Charmed Life	Third Eye Blind	Rock	CADOGAN KEVIN RENE (BMI-	3		Yes	Yes	(BMI-193766232),
4476	1998	92	The Party Continues	Jd	Rap,R&B,Hip	ALMEIDA EUMIRE DEODATO	13		No	No	DEODATO DE, BELL
4711	2001	27	The Way You Love Me	Faith Hill	Country	FOLLESE KEITH C (BMI-	2		No	No	343299059), DULANEY
6034	2014	50	Classic	MKTO	Pop	BOGART EVAN KIDD (BMI-	4		Yes	No	(BMI-482580437),

The rows that had missing values for "Is this a band?" We have added them in.

Added Missing Genres

Originally, 3 songs had a genre of "Other": The Next Time I Fall, I Don't Have The Heart, The Best Things In Life Are Free. We again used Chosic [Chosic (2024)] to find their specific genres. The genre "Quiet Storm" was standardized to be none of the 15 genres above. However, according to All Music [Music (2024)], Quiet Storm is a sub-genre of R&B, so we reclassified it as such. With this, we fully classified every song to contain only the 15 standardized genres above.

Year	Rank	Song	Artist	Standardized Genre	ASCAP Songwriters	Max Number of Writers	BMI_Songwriters	Is this a band?	singer_songwriter	Combined Songwriters
1987	28	The Next Time I Fall	Peter Cetera and Amy Grant	Other (Rock)	CALDWELL ROBERT J (NS-88571630), GORDON PAUL HOWARD (ASCAP-59626155)	2		No	No	CALDWELL ROBERT J (NS-88571630), GORDON PAUL HOWARD (ASCAP-59626155)
1990	36	I Don't Have The Heart	James Ingram	Other (R&B)	FRIEDMAN JUD JOSEPH (BMI-245921268), RICH ALLAN DENNIS (BMI-76316766)	2		No	No	FRIEDMAN JUD JOSEPH (BMI-245921268), RICH ALLAN DENNIS (BMI-76316766)
1992	41	The Best Things In Life Are Free	Luther Vandross and Janet Jackson	Other (Hip Hop, R&B, Soul)	BIVINS MICHAEL (ASCAP-126158585), DE VOE RONALD (ASCAP-126166881), HARRIS JAMES SAMUEL (ASCAP-125105225), LEWIS TERRY STEVEN (ASCAP-125215020)	4		No	No	BIVINS MICHAEL (ASCAP-126158585), DE VOE RONALD (ASCAP-126166881), HARRIS JAMES SAMUEL (ASCAP-125105225), LEWIS TERRY STEVEN (ASCAP-125215020)

The three songs with the genre "Other". We added their genres in the dataset. We removed the "Other" label after taking this figure.

Fixing Some Errors in the "singer_songwriter" Column

A lot of manual work was done for filling in the data value of whether the artist was credited as a songwriter or not. As a method of spotting potential errors, we have created a Python script to create two new columns that returns True when a discrepancy exists within that column. Our method is roughly as follows:

1. Create a column called 'Contains_string' which checks whether the artist's name is (as a string) is contained inside the "Combined Songwriters" column or not

2. Create a column called ’artistNotContained_IsSingerSongwriter’. This column returns True if the artist’s name was not contained in the ’Combined Songwriters’ column, but was given as a "Yes" for the ’singer_songwriter’ column. It returns False otherwise.

3. Create a column called ’artistContained_NotSingerSongwriter’. This column returns True if the artist’s name was contained in ’Combined Songwriters’ column, but was given as a "No" for the ’singer_songwriter’ column. It returns False otherwise.

The actual Python script is:

```
import pandas as pd

#Reverses name. The Artist column uses first name then last name,
#but the ASCAP/BMI uses last name then first name
def reverse_name(full_name):
    # Split the full name into individual words
    name_parts = full_name.split()

    # Rearrange the parts to create the reversed name
    reversed_name = ' '.join(reversed(name_parts))

    return reversed_name

# Load the Excel file
file_path = 'excel_file_here'  # Update this with your file path
df = pd.read_excel(file_path, sheet_name='Sheet 1')

# Check if the strings in 'column1' are contained in 'column2'
df['contains_author_string'] = df.apply(lambda row: str(row['Artist']).upper(
) in str(
                                 row['Combined Songwriters']) or  reverse_name
(str(
                                 row['Artist']).upper()) in str(
                                 row['Combined Songwriters']), axis=1)


#'artistNotContained_IsSingerSongwriter' is true if the artist's name was not
contained
#in the combined songwriter column, but the singer_songwriter column was stil
l True. This is much
#harder to debug due to an artist potentially having multiple names.
df['artistNotContained_IsSingerSongwriter'] = ((df['singer_songwriter'] == 'Y
es') &
                            df['contains_author_string'] == False)

#'artistContained_NotSingerSongwriter' is true if the artist's name was conta
ined in the
```

```
#combined songwriter column, but the singer_songwriter column was still False
. This is much
#easier to debug; the singer_songwriter must be corrected to True because the
artist was
#indeed a songwriter, unless a different songwriter happened to have the
#same name as the artist.
df['artistContained_NotSingerSongwriter'] = ((df['singer_songwriter'] == 'No'
)
& df['contains_author_string'])

# Save the cleaned data back to the same Excel files
with pd.ExcelWriter(file_path, engine='openpyxl', mode='a',
                    if_sheet_exists='replace') as writer:
    df.to_excel(writer, sheet_name='Sheet 1', index=False)
```

As mentioned in the code, the column 'artistNotContained_IsSingerSongwriter' is very hard to debug. This is because most of the rows which have a value of "True" here are due to the artist having multiple names, and so their name in the 'Artist' column is not included in the 'Combined Songwriter' column. We do not make any edits/cleaning from this column.

For the column 'artistContained_NotSingerSongwriter', we have found 10 songs which had a value of "True". The songs were: **St. George And The Dragonet, You And Me, Sometimes When We Touch, The Way You Make Me Feel, Romeo, Do You Believe In Us, I Miss You, Before You Walk Out Of My Life / Like This And Like That, Feel So Good, Cry me a River**. The only song for which our method was a false alarm was for "Romeo"; it's artist's name is Dino, but "Dino" is not actually a songwriter, rather one of the songwriter's name is "Pedinotti Sarah", so the "Dino" string was contained in "Pedinotti". So besides this song, the rest of the 9 songs do in fact have the artist as a songwriter, but mistakenly had a 'singer_songwriter' value of "No". We have fixed them to have a 'singer_songwriter' value of "Yes".

The gist of these two new columns is:

1. The column 'artistNotContained_IsSingerSongwriter' is True when the artist's name in the 'Artist' column is not contained in the 'Combined Songwriters' column, but has a 'singer_songwriter' value of "Yes". We did not clean any rows using this column; it will have to be manually done.

2. The column 'artistContained_NotSingerSongwriter' is True when the artist's name in the 'Artist' column is contained in the 'Combined Songwriters' column, but has a 'singer_songwriter' value of "No". We have used this to revise rows where this discrepancy occurs.

Fixing Typos + Standardization

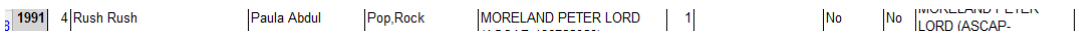
Several typos were corrected and values were standardized to ensure consistency and accuracy in the dataset. Typos such as spelling and extra characters were fixed. For the Genre section, there were a total of 15 main genres, but the dataset consisted of varying ways of displaying one (i.e. "rap" instead of "Rap"; "Hip-Hop" instead of "Hip Hop"). The Genres were manually fixed through the Find/Replace function in Excel to remove extra spaces, capitalize, and essentially standardize the names of the genres. This helped to create the Distribution of Genres graph by not identifying each different variation of one genre as separate genres.

Other typos that have been fixed included: fixed the year for the song "When Smokey Sings", which had a year value of "abc" but should have had the value "1987". Fixed the value of Max Number of Songwriters for "Scream & Shout" which had a value of "yes" but should have had the value of "10". The Song "Vaya Con Dios" has a year of "’". This is changed to 1953. The song "And Mimi" in 1947 was written as "-And Mimi". This has been changed. The song "Be My Life’s Companion"in 1952 had a missing value for ’Max Number of Songwriters’. This was changed to 2.

Creating a Split Dataset

We provide two datasets: "billboard_data_No_Missing_Songs.xlsx" and "Billboard_Splits_No_Missing_Songs.xlsx". The former dataset is closely similar to the dataset Dr. Perot has given us, but we have cleaned the data in the manner described above. We will now explain the purpose of the latter dataset.

One of the visualizations we have made includes analyzing the distributions of genres for the Top 100 songs. To prepare our data for this, we wanted each genre of a song to be counted separately. For example, the song "Rush Rush" has the genres "Pop" and "Rock". So we count "Pop" once and "Rock" once.



The row containing the song "Rush Rush".

In order to achieve this, we needed to separate each genre the song contained. Tazia created a script to split each song by their genre. Her script (slightly modified by us) was as follows:

```
import pandas as pd

# Specify the path to your CSV file
csv_file_path = "dataset_goes_here"

# Read the CSV file (replace with your actual file path)
df = pd.read_csv(csv_file_path)
```

```
# Rename the column
df.rename(columns={"Standardized Genres": "stdg"}, inplace=True)

# Split the comma-separated values into separate rows
df_split = df.assign(stdg=df['stdg'].str.split(',')).explode('stdg')

# Reset the index (optional)
df_split.reset_index(drop=True, inplace=True)

# Export the DataFrame to the CSV file
df_split.to_excel('new_dataset_name', index=False)

# Display the first few rows of the new DataFrame
print(df.head())
```

To see this in effect, continuing our example above, "Rush Rush" would be two rows where the genre has been split.

1991	4	Rush Rush	Paula Abd	MORELAND PETER LC	MORELAN	1	No	No	Pop
1991	4	Rush Rush	Paula Abd	MORELAND PETER LC	MORELAN	1	No	No	Rock

The other information about the song "Rush Rush" is simply repeated. The only information that was changed was spliting the genre into "Pop" and "Rock" respectively.

This dataset is then used for visualizations that need the distribution of genres.

Combining Songwriters - Concatenating & Removing Duplicates

The ASCAP and BMI songwriter columns were concatenated to account for overlapping and/or missing songwriters.

The idea for the code is:

1. Strip each ASCAP, BMI songwriter name using ", " as a separator
2. Concatenate them
3. Remove duplicate names

Python code for the combined songwriter column for non-split data:

```
import pandas as pd

# Load the Excel file
file_path = 'billboard_data_No_Missing_Songs.xlsx'  # Update this with your f
ile path
df = pd.read_excel(file_path, sheet_name='Sheet 1')

# Function to remove trailing commas and spaces
def remove_trailing_commas_and_spaces(songwriters):
    if pd.isna(songwriters):
```

```
        return songwriters
    return songwriters.rstrip(', ').strip()

# Function to remove duplicate names while preserving the order
def remove_duplicates(songwriters):
    if pd.isna(songwriters):
        return songwriters
    songwriters_list = songwriters.split(', ')
    unique_songwriters = []
    seen = set()
    for songwriter in songwriters_list:
        # Extract name part (remove the ID)
        name = ' '.join(songwriter.split()[:-1])
        if name not in seen:
            seen.add(name)
            unique_songwriters.append(songwriter)
    return ', '.join(unique_songwriters)

# Create 'Combined Songwriters' column by concatenating ASCAP and BMI songwri
ters
df['Combined Songwriters'] = df.apply(
    lambda row: ', '.join(
        filter(None, [str(row['ASCAP Songwriters']), str(row['BMI_Songwriters
'])])
    ),
    axis=1
)
# Apply the functions to clean up the 'Combined Songwriters' column
df['Combined Songwriters'] = df['Combined Songwriters'].apply(
                    remove_trailing_commas_and_spaces)
df['Combined Songwriters'] = df['Combined Songwriters'].apply(remove_duplicat
es)

# Save the cleaned data back to the same Excel file
with pd.ExcelWriter(file_path, engine='openpyxl', mode='a',
                    if_sheet_exists='replace') as writer:
    df.to_excel(writer, sheet_name='Sheet 1', index=False)
```

Python code used to create the combined songwriter column for the split data

```
# Import
import pandas as pd


# Load the dataset
df = pd.read_excel('dataset_name_goes_here')
```

```
# Concatenate ASCAP and BMI songwriters
df['Combined Songwriters'] = df.apply(lambda row: ','.join(
                                filter(None, [row['ASCAP Songwriters'],
                                row['BMI_Songwriters']])), axis = 1)


# Remove duplicate songwriters
df['Combined Songwriters'] = df['Combined Songwriters'].apply(lambda x: ', '.
join(
                                                                pd.unique(x.s
plit(', '))))


# Save updated dataframe to new Excel file
df.to_excel('new_dataset_name_goes_here', index=False)
```

Note that this cannot account for nicknames. For example, the Song "I Love You For Sentimental Reasons" in 1946 has William Henry “Pat” Best as a songwriter. He also went by "William Bounds". The ASCAP named him as "BOUNDS WILLIAM BEST III" while the BMI named him as "BEST WILLIAM H PAT". Because of this, our Combined Songwriter column includes both. However, regarding the graphs, this error can be considered negligible as the number of nicknames for a songwriter is very low and won’t significantly change the trends of the graphs. This has to be manually fixed.

1	ASCAP Songwriters	BMI_Songwriters
45	BOUNDS WILLIAM BEST III (ASCAP-74029192), WATSON DEEK (BMI-39595443)	BEST WILLIAM H PAT (BMI-354011994), WATSON DEEK (BMI-39595443),

Combined Songwriters
BOUNDS WILLIAM BEST III (ASCAP-74029192), WATSON DEEK (BMI-39595443), BEST WILLIAM H PAT (BMI-354011994)

We give two pictures: the one above showing the ASCAP and BMI songwriter columns, and the below showing the Combined Songwriters column. The picture above shows the ASCAP and BMI Songwriter columns. William Henry Best’s nicknames are used for ASCAP and BMI, and thus they do not match. Because of this, his name comes up twice in the combined songwriter column.

Cleaning Trailing Commas, nan’s, and Double Commas

The method used for creating Combined Songwriters can include trailing commas, nan values (which occur when the ’ASCAP Songwriters’ column is empty but the ’BMI Songwriters’ column isn’t), and double commas. We have fixed trailing commas with the steps (code included):

1. To delete nan’s, we simply use the Find and Replace option in Excel to replace all "nan," values with "".
2. To delete double commas, we simply use the Find and Replace option in Excel to replace all ’,,’ with ’,’.

3. Type Alt and F11 together
4. Press Insert and click Module
5. Paste the following Code. The idea is if the cell isn't empty, if the last character is a comma, delete it.

```
    Sub remove_trail_commas()
    Dim rng As Range
    Dim cell As Range

    ' Prompt the user to select a range
    On Error Resume Next
    Set rng = Application.InputBox("Select the range to clean", Type:=8
)
    On Error GoTo 0

    If Not rng Is Nothing Then
        For Each cell In rng
            If Right(cell.Value, 1) = "," Then
                cell.Value = Left(cell.Value, Len(cell.Value) - 1)
            End If
        Next cell
    End If
End Sub
```

6. Exit the window
7. Type ALT and F8 together
8. Click on "remove_trail_commas", select Run.
9. (Occasionally) Once you save the file, there might be a pop-up warning. If it is a message about not being able to save in a "Macro-Free workbook", you can just click save and erase features.

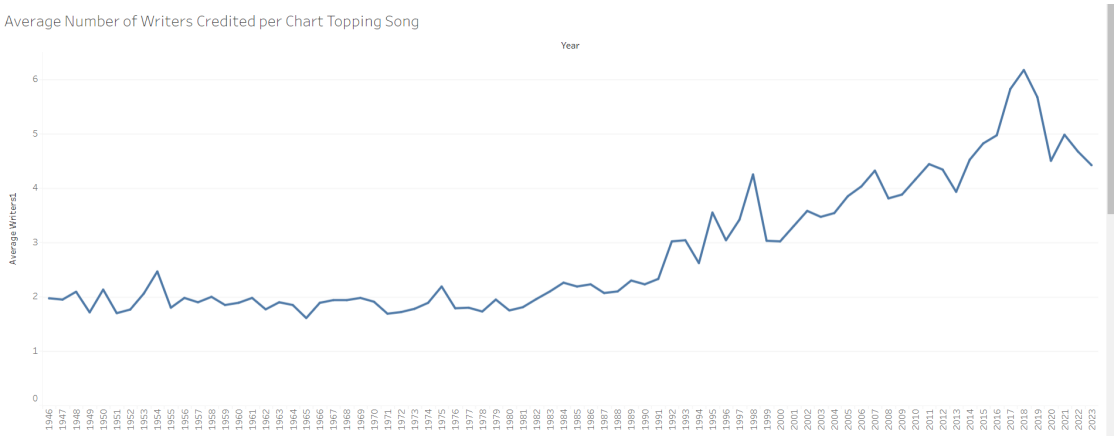
Methodology for Graphs

This section details the methodology used for each graph created during the analysis.

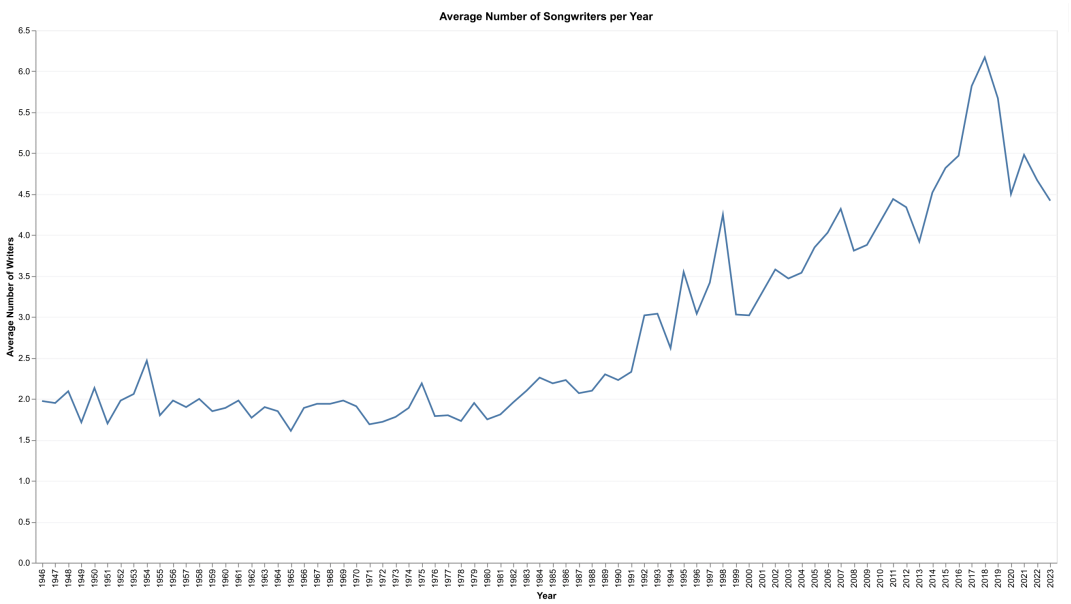
Average Number of Credited Songwriters per Chart-Topping Song

The average number of credited songwriters per chart-topping song was calculated and plotted over the years. This involved unpivoting the genre columns and calculating the average number of songwriters.

Tableau:



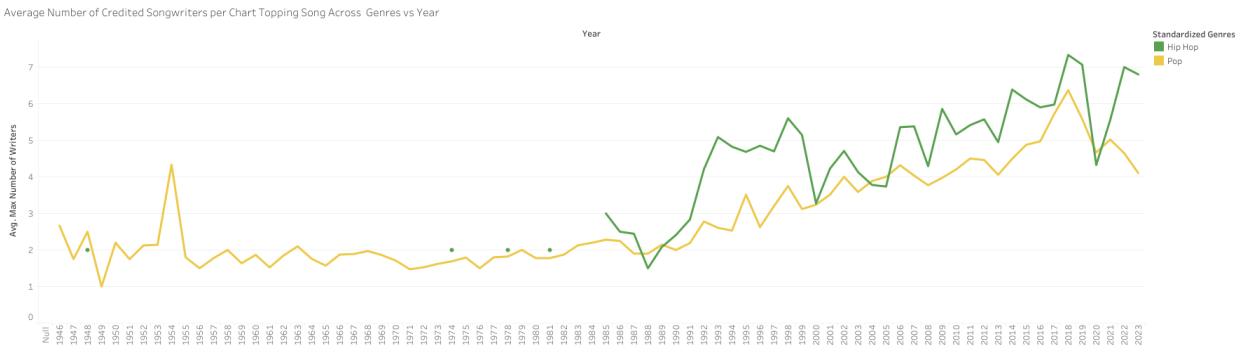
PowerBI:



Average Number of Credited Songwriters per Chart Topping Song for Pop and Hip Hop vs Year

To create a graph showing the average number of credited songwriters for Pop and Hip Hop genres, the dataset was filtered to include only these genres. The average number of songwriters was then calculated for each year.

Tableau:



PowerBI:

The "Hip Hop" genre does not have values for multiple years before 1985. PowerBI uses an interpolation for every point, which means it interpolates very large gaps between Hip Hop points. PowerBI is unable to convert these values to be discrete points while keeping the line plot. Instead, Python is able to achieve this with Plotnine, which is the Python equivalent for R's ggplot2 package.

Python using Plotnine:

The following code was used for the Python visualization:

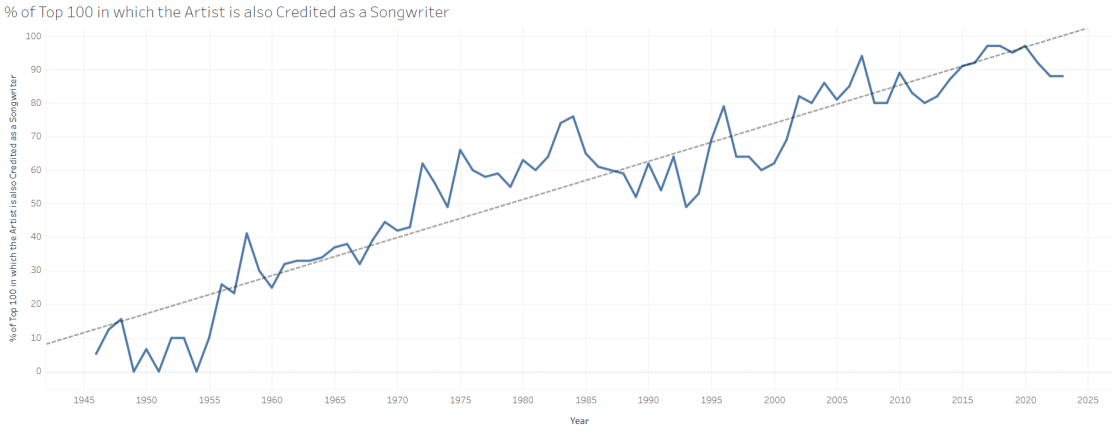
```
import pandas as pd
from plotnine import (ggplot, aes, geom_point, geom_line, labs, theme,
    element_text, element_line, element_rect, theme_set,
    theme_void, guide_legend, theme_minimal, theme_classic, theme_bw, xlim,
    scale_x_continuous, scale_color_manual)

# Specify the path to your CSV file
excel_file_path = "Billboard Data QCd Split.xlsx"

# Read the CSV file (replace with your actual file path)
df = pd.read_excel(excel_file_path)

#Filter for Pop and Hip Hop
df = df[(df['Standardized Genres'] == 'Pop') | (df['Standardized Genres'] ==
'Hip Hop')]

#Calculate Average for 'Max Number of Writers'
grouped_df = df.groupby(['Year', 'Standardized Genres'])['Max Number of Write
rs'].mean().reset_index()
# Separate data for Hip Hop and Pop. For Hip Hop, further separate before and
after 1985
df_hiphop = grouped_df[grouped_df['Standardized Genres'] == 'Hip Hop']
df_pop = grouped_df[grouped_df['Standardized Genres'] == 'Pop']
df_hiphop_before_1985 = df_hiphop[df_hiphop['Year'] <= 1985]
df_hiphop_after_1985 = df_hiphop[df_hiphop['Year'] > 1985]

# Plotting
plot = (
    ggplot() +
    #Discrete Points for Hip Hop for before 1985
    geom_point(df_hiphop_before_1985, aes(x='Year', y='Max Number of Writers'
,
                                          color = 'Standardized Genres'), size=
3) +
    #Line plot for all of Pop
    geom_line(df_pop, aes(x='Year', y='Max Number of Writers',
                          color= 'Standardized Genres'),size = .75) +
    #Line plot for Hip Hop after 1985
    geom_line(df_hiphop_after_1985, aes(x='Year', y='Max Number of Writers',
                                        color='Standardized Genres'), siz
e = .75)+
    #Set range for x axis
     xlim(1946, 2024)+ scale_x_continuous(breaks=range(1946, 2027, 10))+

     #Set colors for Hip Hop and Pop.
    scale_color_manual(values={'Hip Hop': '#744EC2', 'Pop': '#1AAB40'}) +
    #Set Titles and Labels
    labs(title='Average Number of Credited Songwriters
                per Chart Topping Song Across Pop/Hip Hop vs Year',
```

```
        x='Year',
        y='Average of Max Number of Writers') +
    #Set Theme and grid colors/ legend
    theme_bw() +
    theme(
       panel_grid_major=element_line(color = "#F4F5F7", size=0.5),
       panel_grid_minor=element_line(color = "#F4F5F7", size=0.25),
       legend_position=(0.1, 0.95)
   )
)
print(plot)
```

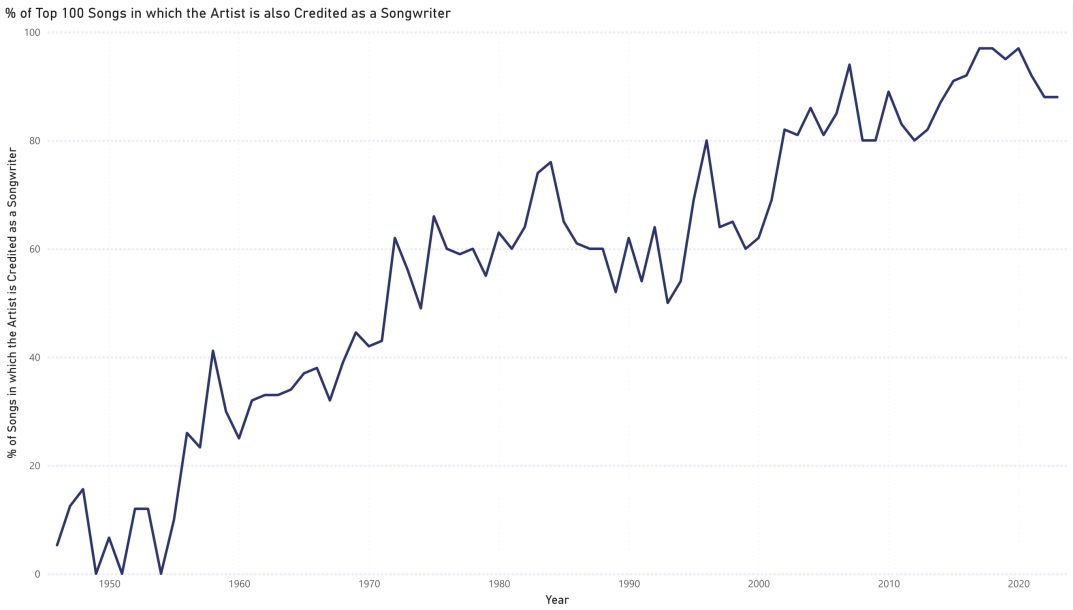
% of Top 100 in which the Artist is also Credited as a Songwriter

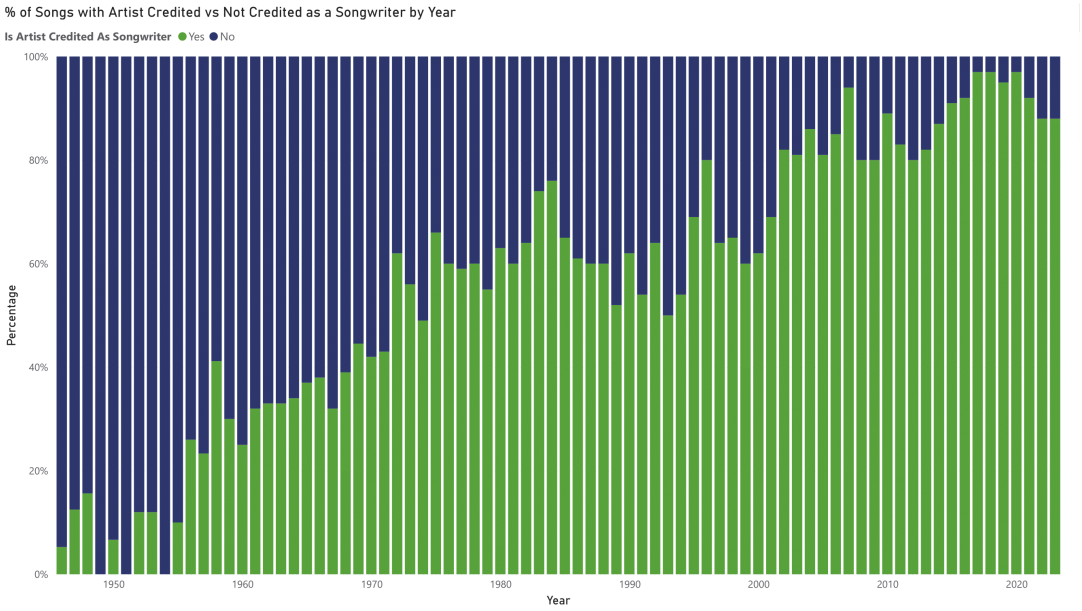
The percentage of top 100 songs in which the artist is also credited as a songwriter was calculated and plotted over the years. This involved filtering the dataset for songs where the artist is listed as a songwriter.

Tableau:



PowerBI:



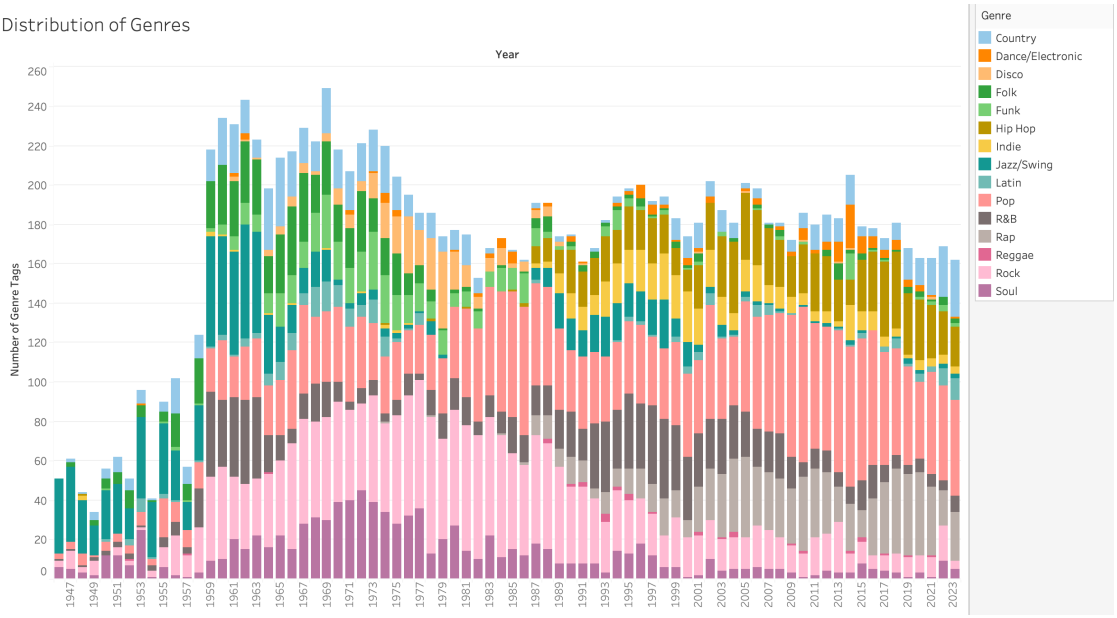


We give two interpretations. The stacked bar chart shows the rest of the proportions that also shows the percentage of artists that aren't credited as a songwriter.

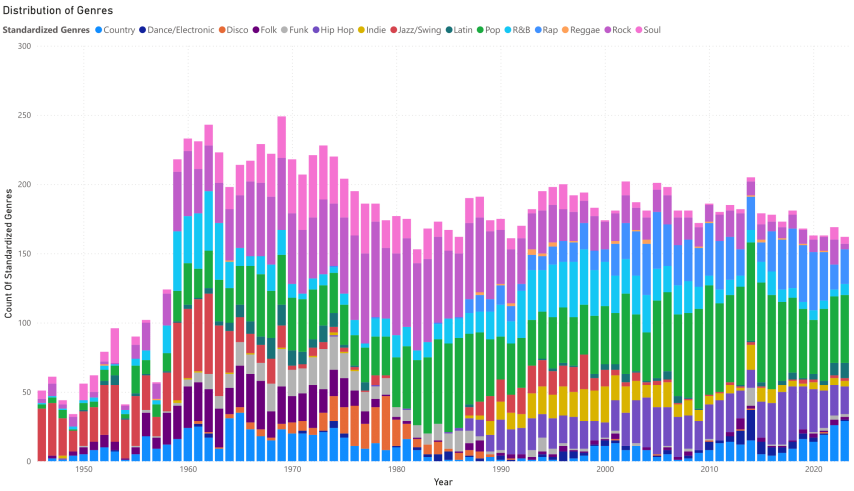
Distribution of Genres

The distribution of genres is created by plotting the count of each genre per year, using the split dataset.

Tableau:



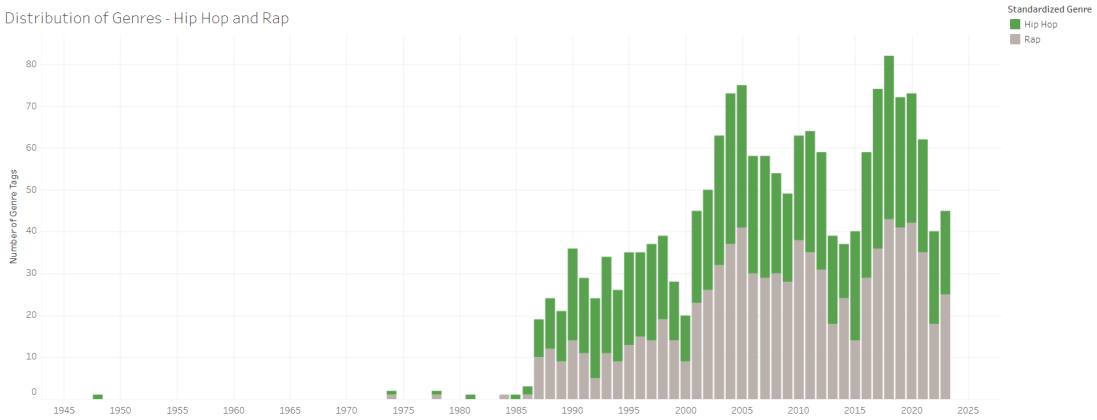
PowerBI:



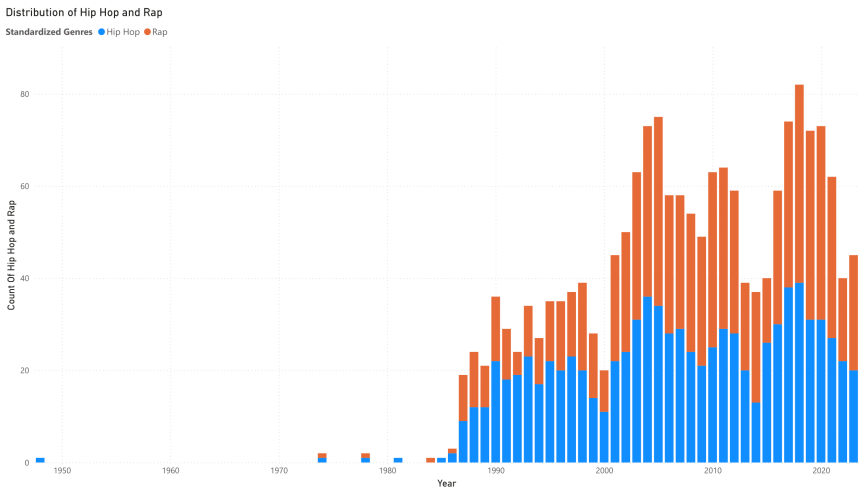
Distribution of Genres - Hip Hop & Rap

The distribution of genres is further filtered to just Hip Hop and Rap.

Tableau:

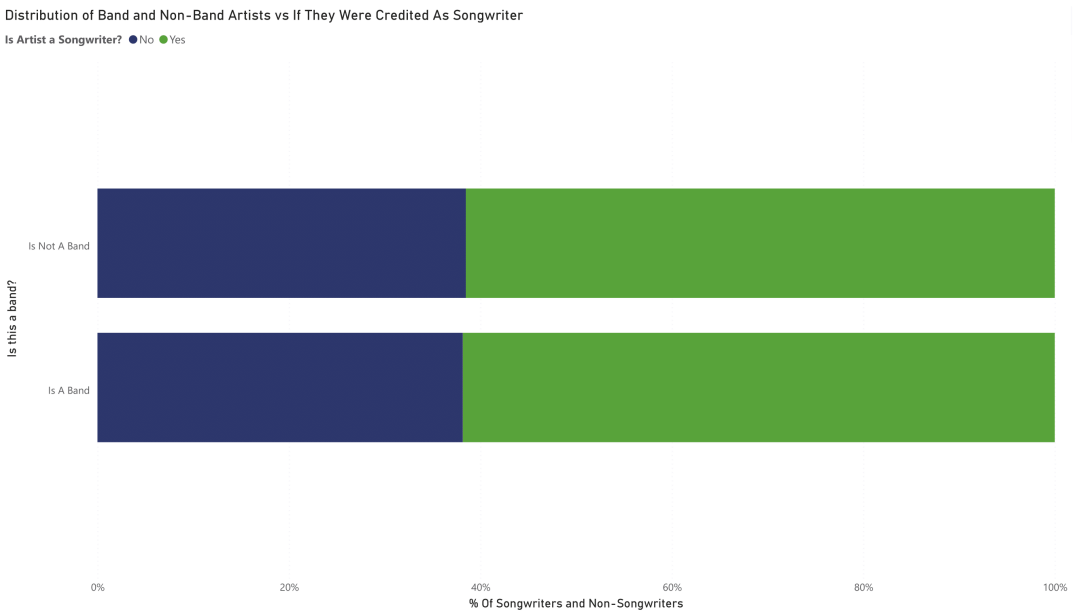


PowerBI:



Band vs. Non-Bands For Being Credited as Songwriters

PowerBI:



We see that Bands and Non-bands are credited as songwriters at very similar rates.

Anomalies and Trends

During the analysis, several anomalies and trends were observed. These are detailed below:

Anomalies

- Some songs had discrepancies in songwriter information across different years.
- Overlapping songwriter credits are apparent across ASCAP and BMI columns, as well as different nicknames

Trends

- An increase in the average number of credited songwriters over the years.
- A higher percentage of artists being credited as songwriters in recent years.

Findings

The analysis revealed several key findings:

- The music industry has seen an increase in collaborative songwriting over the years.
- Artists are more frequently involved in the songwriting process.

Warnings and Disclosures

- The dataset may contain inaccuracies due to manual entry and missing data.
- Songwriter credits may not fully represent all contributors to a song.

Conclusion

This report provides a comprehensive overview of the modifications and methodologies used in analyzing the Billboard dataset. The findings highlight significant trends in the music industry, including increased collaboration and artist involvement in songwriting.

Acknowledgement

We would like to express our gratitude to Dr. Emma Perot and Tazia Mehdi for helpful suggestions, comments, and remarks.

References

[ASCAP 2024] *ASCAP: ASCAP Repertory Search. https://www.ascap.com/repertory#/. 2024. – (accessed: 07.31.2024)*

[Billboard Archive] *Billboard Archive: Billboard https://books.google.tt/books/about/Billboard.html?id=pBQEAAAAMBAJ&redir_esc=y . – (accessed: 07.31.2024)*

[Billboard Year End 2006 onward] *Billboard: Hot 100 songs* https://www.billboard.com/charts/year-end/2006/hot-100-songs/ *(accessed: 07.31.2024)*

[BMI 2024] *BMI: Songview. https://repertoire.bmi.com/. 2024. – (accessed: 07.31.2024)*

[Chosic 2024] *Chosic: Music Genre Finder. https://www.chosic.com/music-genre-finder/. 2024. – (accessed: 07.31.2024)*

[Music 2024] *Music, All: Quiet Storm. https://www.allmusic.com/style/*

quiet-storm-ma0000004425. 2024. – (accessed: 07.31.2024)

[World Radio History] *World Radio History. https://www.worldradiohistory.com/index.htm (accessed 07.31.2024)*

Contact Email

Sean Lee: sean.lee.udub@outlook.com

Ajit Mallavarapu: aj1tmallav@gmail.com