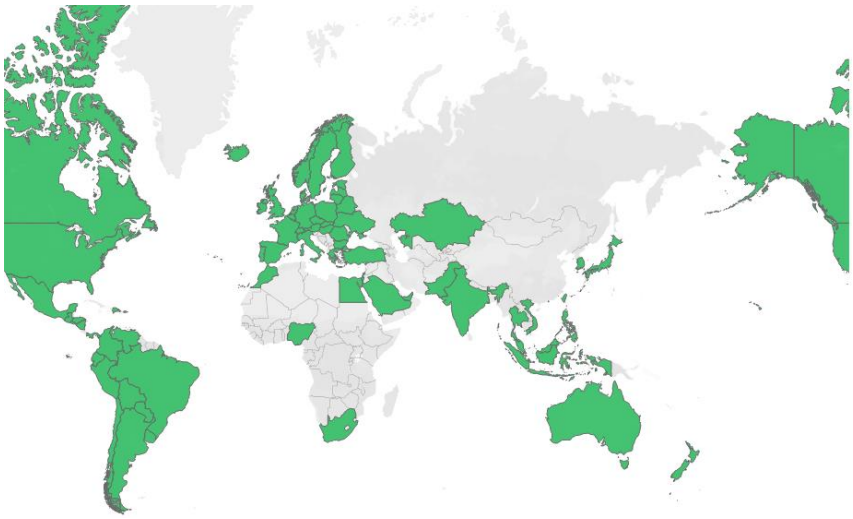
Project Report

Regional Trends and Analysis of Spotify 2024 Top Songs

1. Project Overview

This project delves into the multifaceted world of music popularity, combining regional trend analysis with predictive modeling, leveraging the Top Spotify Songs in 73 Countries. Using data sourced from Kaggle, the analysis will investigate how musical preferences vary across different geographical locations, focusing on North America, Latin America, Europe, Asia, Africa, and Oceania. It will also include a Logistic Regression Model built using Pandas and skitlearn that will help predict a song's potential popularity based on different audio features (tempo, key, mode, danceability, energy, etc.).



Countries for Analysis.

Key research questions include:

1) What are the top songs and artists globally? Do they share any identifiable features?

2) Are there any discernible musical characteristics (tempo, key, acousticness) that are more prevalent in certain regions?

3) Do any of the audio features show a high correlation with song popularity?

The analysis will involve data cleaning, exploratory data analysis using visualizations, and statistical analysis to identify regional trends. The expected outcomes include identifying distinct regional musical preferences, discovering potential correlations between popularity and musical characteristics, and gaining insights into the global music landscape and its regional variations.

Data tools:

R and Tableau for data cleaning, EDA, and visualizations. Python for the Logistic Regression Model.

2. Data

2.1 Data description

The dataset used in this analysis was sourced from Kaggle: Top Spotify Songs in 73 Countries (Daily Updated) | Kaggle. The data frame has 25 columns and 1, 468, 130 rows. The column names are:

spotify_id	weekly_movement	duration_ms	key	instrumentalness
name	country	album_name	loudness	liveness
artists	snapshot_date	album_release_date	mode	valence
daily_rank	popularity	danceability	speechiness	tempo
daily_movement	is_explicit	energy	acousticness	time_signature

There is information on 72 countries and the Global Top 50. Each country has rows for the daily rank (top 50) for each day from October 10, 2023, to November 27, 2024. To consult the variable overview, refer to appendix 6.1 at the end of the document.

2.2 Data structuring and cleaning

From the variable overview, we get that the NA values in the country column represent the playlist for "Global Top 50", the first step was changing these NAs to "GO" (to keep using the same two-letter format the rest of the columns have).

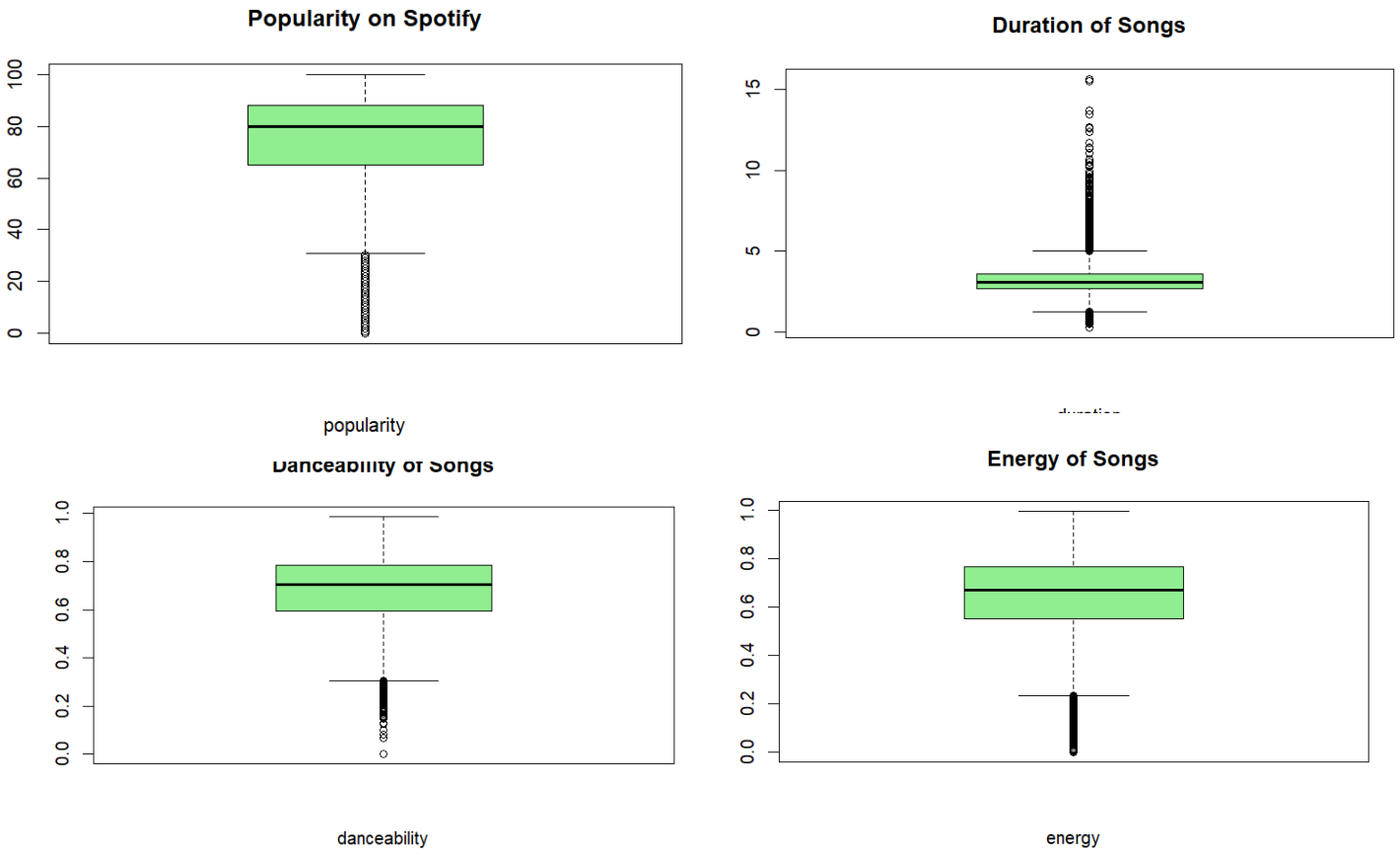
Next was to remove the NAs present in name, artists, album_name, and album_release_date. The column with the most NAs was album_name with 670 values, the rest of the NAs from the other columns are contained in album_name, so the total number of NAs in the data frame is 670, which represents less than 0.05% of the values and they were removed.

Next, we checked for duplicates and found none.

We looked for outliers in some of the numeric columns to get a feel of the data. The columns were: popularity, duration_ms, danceability, and energy.

- Popularity
    - The boxplot shows that there are several outliers at the end of the plot
    - Using the Interquartile Range we got that 50% of the values of popularity go from 65 to 88
    - Using the min and max of the boxplot we got that there are 13,895 outliers in popularity or 0.95%
    - All the outliers are on the low end of the plot

- Duration_ms
    - 50% of the values are located on a very small section of the plot, they go from 2.68 to 3.61
    - There are 36,939 outliers: 1,620 on the low end of the plot and 35,319 on the high end of the plot
    - 2.52% of the entries are outliers in duration_ms
- Danceability
    - 50% of the outliers go from 0.594 to 0.787
- Energy
    - 50% of the outliers go from 0.553 to 0.766



Boxplots of popularity, duration_ms, danceability, and energy.

These numeric columns show the presence of outliers, the decision was to not remove them given the type of analysis at hand. We need to find the trends and that includes these song features even if they are lower or higher than most of the values.

The column duration_ms is on milliseconds, so it was changed to minutes (duration_min) for a more precise understanding.

By the end of the cleaning and structuring processes, the data frame ended up with 25 columns and 1,467,460 rows.

2.3 Data exploration

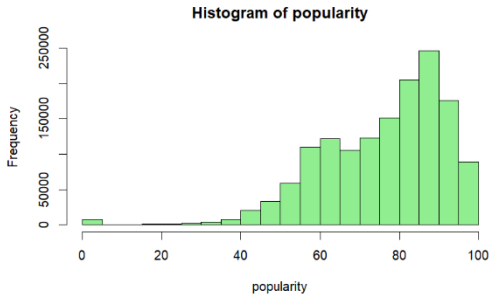
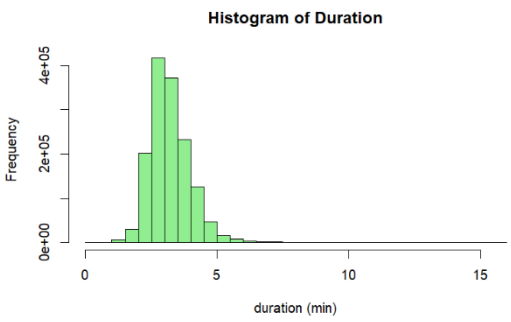
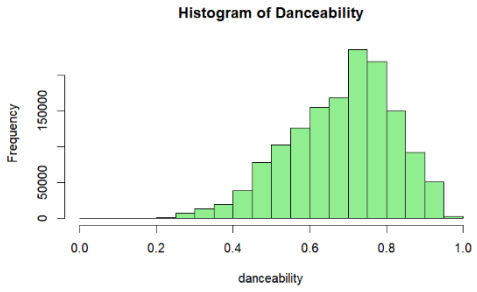
2.3.1 Summary Statistics

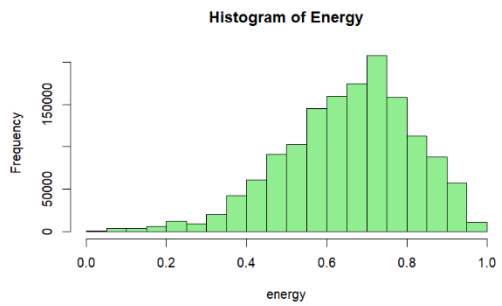
spotify_id	name	country	snapshot_date	popularity
Length: 1467460	Length: 1467460	Length: 1467460	Min: 2023-10-18	Min: 0.00
Class: character	Class: character	Class: character	1st Q:2024-01-26	1st Q: 65.00
Mode: character	Mode: character	Mode: character	Median: 2024-05-09	Median: 80.00
			Mean: 2024-05-07	Mean: 76.38
			3rd Q: 2024-08-18	3rd Q: 88.00
			Max: 2024-11-27	Max:100.00

To consult the rest of the variable's summary, refer to appendix 6.2 at the end of the document.

From the summary and the outlier's analysis for danceability and energy, we noticed the values for both are similar, so they should be investigated.

2.3.2 Distributions

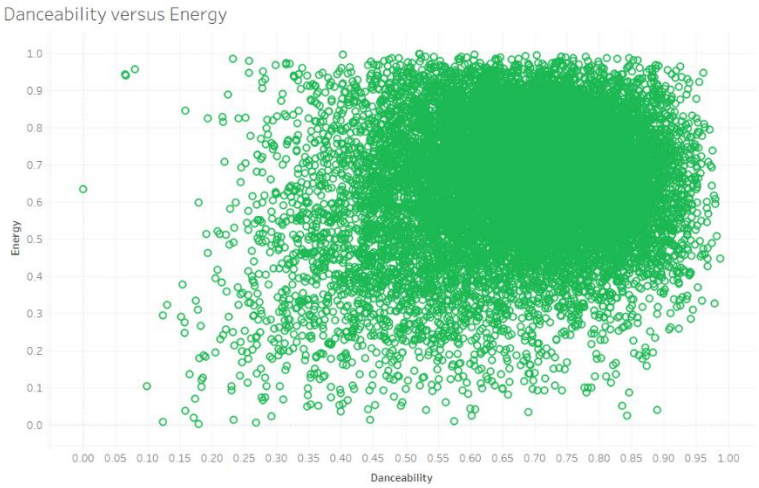
Variable	Histogram	Insights
Popularity		• The histogram is left-skewed • Most of the popularity is on higher values (which makes sense because these are the top songs) • Most of the values are spread between ~40 and 100 • The mean is 76.37
Duration_min		▪ The histogram is right-skewed ▪ Most of the songs' durations are between 0 and 5 minutes ▪ The values are not spread throughout the plot, which means most of them are similar ▪ The mean is 3.21
Danceability		▪ The histogram shows a normal distribution ▪ The values are spread from ~0.4 to ~0.9 ▪ The mean is 0.68

Energy		▪ Just like danceability, the histogram shows a normal distribution ▪ The values are spread from ~0.4 to ~0.9 ▪ The mean is 0.65

The distributions for danceability and energy show again a resemblance, so they should be compared against each other to find if there is a correlation.

2.3.3 Correlations

The visualizations for this section were made in Tableau. Given that one of the main questions is to see if song characteristics are highly correlated with popularity, most of the correlations were made to compare popularity against these audio features. But first, danceability versus energy:



Danceability versus energy Scatterplot.

- The chart shows a positive correlation between these two variables, it seems that if a song has a high danceability then it might also have a higher energy
- We can see there is a very dense cluster in the middle to upper-right corner. This suggests most songs have a middle to high level on both variables
- There are outliers present in the chart, both from low danceability and high energy and vice versa.
- The positive correlation could suggest that songs perceived as more danceable are often also perceived as more energetic
- The clustered distribution might reflect the fact that a large portion of popular music is designed to be both danceable and energetic. This could be due to listener preferences or the influence of trends in the music industry.

Next, we have the correlations between popularity versus danceability, energy, duration_min, speechiness, tempo, instrumentalness, and acousticness.

For popularity versus danceability:

- We can see a dense cluster in the upper middle of the chart, this would suggest the songs have a high danceability level. Given that all these songs are from the top 50, we can say danceability shows as a factor for popularity
- This isn't to say that if a song has a high danceability level it would also be popular, the cluster shows that the popularity is not centered in the high levels
- Now, the plot shows a line of songs that have zero popularity and are in on a variety of danceability levels, so there are exceptions

For popularity versus energy:

- This chart is more spread than the popularity versus danceability one
- There is also a dense cluster on the upper middle of the plot, but with more points on the lower side
- Energy shows a factor for popularity, but not as much as danceability

For popularity versus duration:

- There is a dense cluster on the lower half of the chart
- Most of the songs are less than 5 minutes and are spread when it comes to popularity, so we cannot say duration is directly correlated with popularity

For popularity versus speechiness:

- There is a dense cluster on the lower side of the chart, most of the values on the low end of speechiness
- The densest part seems to go from ~30 to ~80 in popularity and zero to ~0.3 in speechiness, so there might be some correlation there

For popularity versus tempo:

- There is a dense cluster in the middle of the chart
- The densest part seems to go from ~30 to ~70 in popularity and ~80 to ~160 in tempo, so there might be some correlation there

For popularity versus instrumentalness

- There is a dense cluster in the lower part of the chart, though there are several points spread all over the plot
- The densest part seems to go from ~20 to ~80 in popularity and zero to ~0.1 in instrumentalness

For popularity versus acousticness:

- There is a dense cluster of points, but they are too spread to infer a correlation between the variables

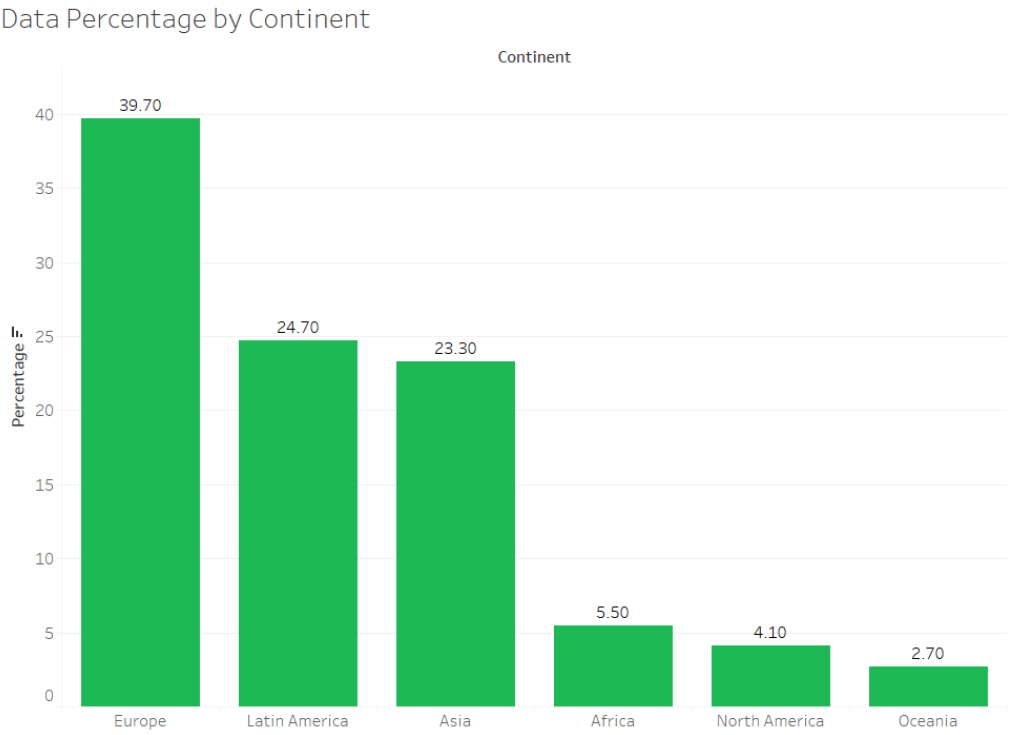
The scatterplots that were not shown can be found on Tableau Public: https://public.tableau.com/views/spotify_charts_17353242293170/SpotifyCorrelations?:language=en-US&:sid=&:redirect=auth&:display_count=n&:origin=viz_share_link

3. Methodology

3.1 Regional Trends

To find the regional trends, the data frame was divided into sections, each containing the data for North America, Latin America, Europe, Asia, Africa, and Oceania, respectively. This allowed us to compare top songs, top artists and song feature statistics.

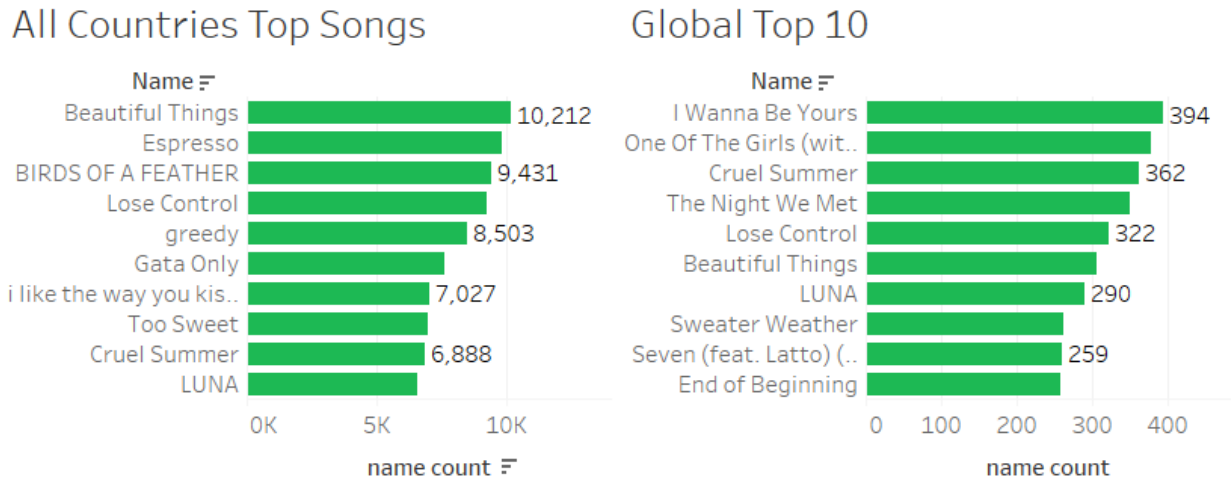
It is important to clarify that the following plots have “all countries” and “global” sections. They are not the same, given that the “all countries” charts include the data of all 72 countries, while the “global” section only includes the daily global top 50.



Continent Data Percentage.

Now, to have some perspective on the amount of data each continent has on the data set we can see on the plot that Europe is the continent with the greatest number of entries, followed by Latin America and Asia. Africa, North America, and Oceania are the ones with the least amount of data.

3.1.1 Top songs



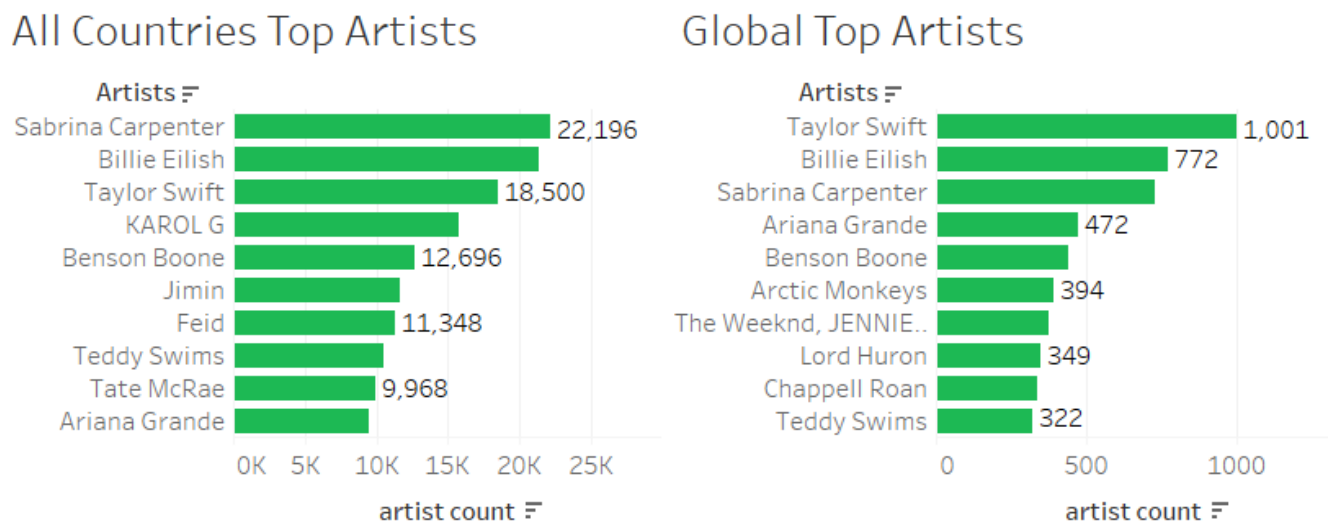
All countries and Global Top 10 songs.

To see the rest of the top songs by continent, please refer to appendix 6.3.

We compared all countries top 10 songs to see how many were on each continent's top 10:

	Global	North America	Latin America	Europe	Asia	Africa	Oceania
Beautiful Things	■	■		■			■
Espresso				■	■		
BIRDS OF A FEATHER				■	■		
Lose Control	■	■		■			■
greedy				■	■		
Gata Only							
i like the way you kiss				■			
Too Sweet		■		■			
Cruel Summer	■	■			■		■
LUNA	■		■				

3.1. 2 Top artists



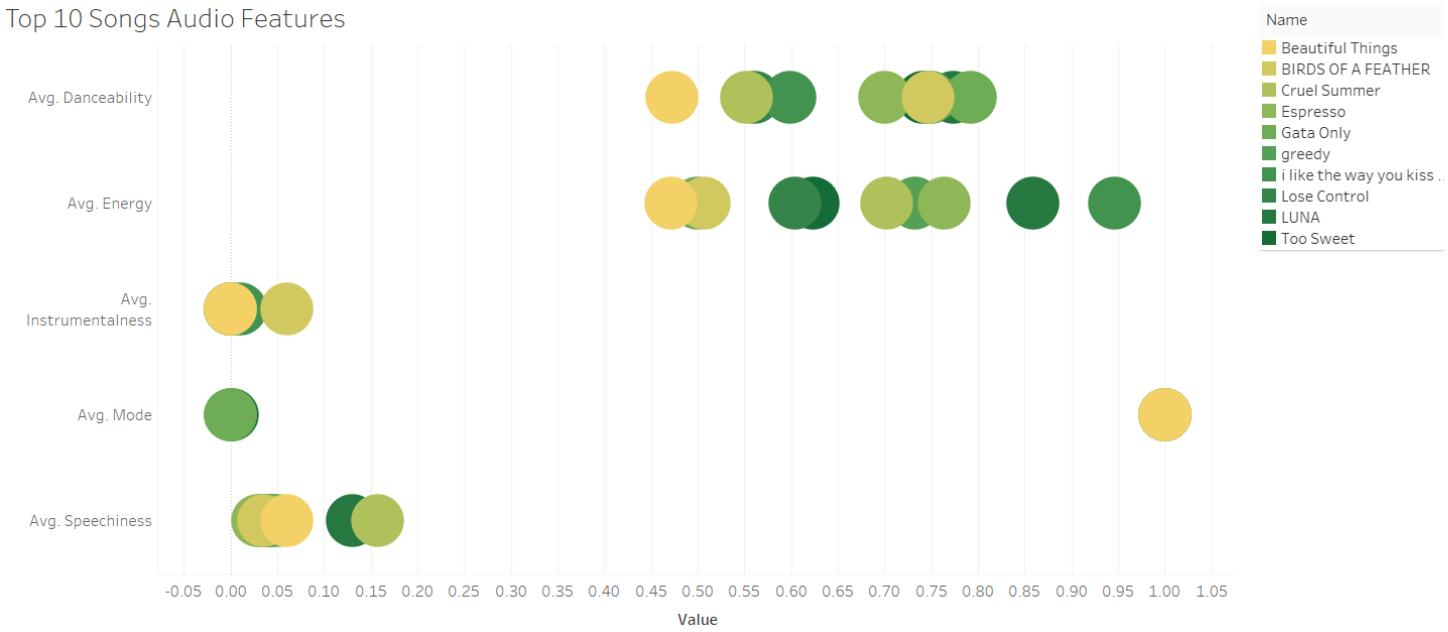
All countries and Global Top 10 artists.

To see the rest of the top artists by continent, please refer to appendix 6.4.

We compared the all country's top 10 artists to see how many were on each continent's top 10:

	Global	North America	Latin America	Europe	Asia	Africa	Oceania
Sabrina Carpenter	■	■		■	■		■
Billie Eilish	■	■		■	■		■
Taylor Swift	■	■		■	■		■
KAROL G			■				
Benson Boone	■	■		■			■
Jimin					■		
Feid			■				
Teddy Swims	■			■			■
Tate McRae				■			■
Ariana Grande	■						

3.1.3 Top 10 Songs Audio Features



Top 10 Songs Audio Features.

3.1.3 Song Feature Means by Continent

	Europe	Latin America	Asia	Africa	North America	Oceania
Popularity	73.79	81.14	75.36	64.06	86.04	87.48
Is_explicit	0.37	0.46	0.14	0.25	0.46	0.33
Duration_min	3.04	3.13	3.42	3.69	3.24	3.32
Danceability	0.68	0.74	0.64	0.71	0.67	0.64
Energy	0.66	0.69	0.61	0.65	0.64	0.62
Instrumentalness	0.02	0.006	0.02	0.03	0.009	0.01
Mode	0.52	0.47	0.62	0.39	0.62	0.72
Key	5.49	5.86	5.38	5.32	5.62	5.2
Speechiness	0.101	0.11	0.07	0.11	0.075	0.07

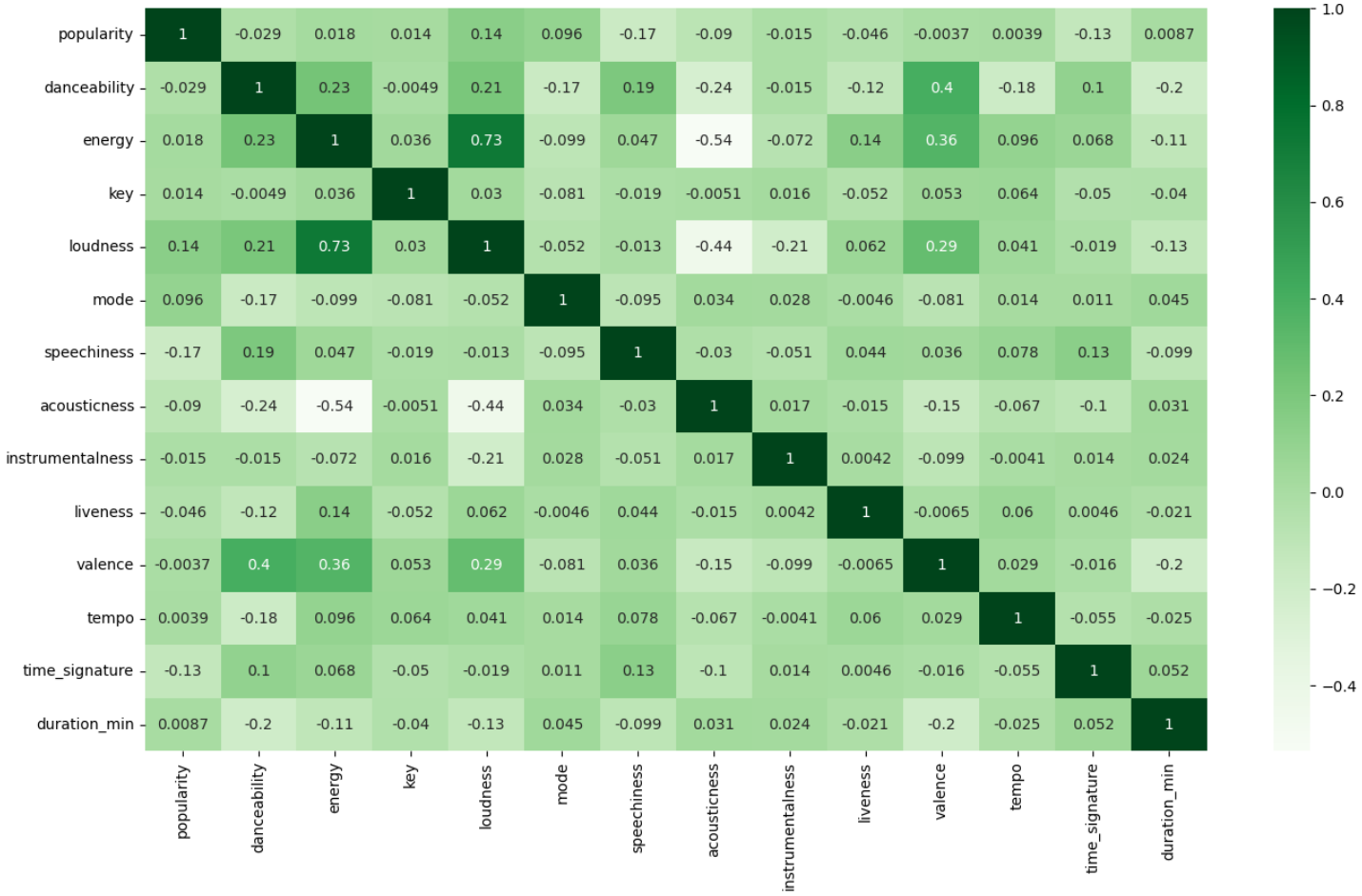
3.2 Logistic Regression Model

By selecting: danceability, energy, key, loudness, mode, speechiness, acousticness, instrumentalness, liveness, valence, tempo, time signature, and duration of the data set songs, we built a binomial logistic regression model that classifies a song into “would be popular” or “would not be popular”.

We found outliers in the selected columns and removed them, given that logistic models are sensitive to them. We plotted a heatmap of the columns to find correlations between them.

From the heatmap we noticed that the highest correlations are from energy and loudness, valence and danceability, and valence and energy. Popularity has the highest correlation with loudness, at 0.14.

To be able to classify the songs, a new column called “is_popular” was made (to have a logical column that would show if a song is popular or not). This was made by selecting the songs with a popularity equal or higher than 70 and got them marked as popular.



Heatmap of Spotify Song Features

4. Results

- We found there was a positive correlation between danceability and energy.
- Popularity seems to be correlated with danceability and energy, with exceptions.
- The top song was Beautiful Things by Benson Boone.
- The top artist was Sabrina Carpenter.
- The Top 10 Songs danceability and energy averages go from 0.45 to 0.8.
- The Top 10 Songs speechiness and instrumentalness go from 0 to 0.15.
- Seven out the 10 top Songs have a mode of one.

4.1 Regional Trends

- Seven out of the top 10 Europe songs were also in the all countries top 10.
- Two of the top 10 songs are in Spanish, the rest are in English.
- The top 3 artists (Sabrina Carpenter, Billie Eilish, and Taylor Swift) are also in the Global Top 10, North America, Europe, and Asia top 10 artists.
- Six out of the top 10 Europe artists were also in the all countries top 10.
- Three artists on the top 10 are non-English speakers (KAROL G, Jimin, and Feid).

- Latin America and North America have the highest explicit song count.
- Asia has the least explicit song count.
- Africa has the highest average song duration and instrumentalness.
- Latin America has the highest danceability, energy, and key song average.
- Oceania has the highest mode average.
- Latin America and Africa have the highest speechiness.

4.2 Model Results

While considering all the other variables constant:

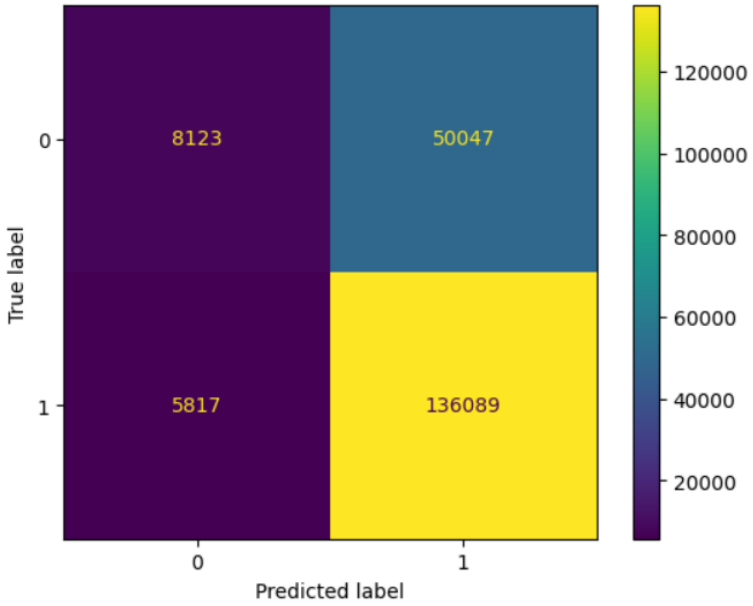
- With a one-unit increase in danceability, the odds of a song being popular decrease by 60%
- With a one-unit increase in energy, the odds of a song being popular decrease by 97%
- A one-unit increase in key means the odds of a song being popular decrease by 0.704%
- A one-unit increase in loudness means the odds of a song being popular increase by 38%
- A one-unit increase in mode means the odds of a song being popular increase by 48%
- With a one-unit increase in speechinees, the odds of a song being popular decrease by 93%
- With a one-unit increase in acousticness, the odds of a song being popular decrease by 71%
- With a one-unit increase in instrumentalness, the odds of a song being popular increase by 0.189%
- A one-unit increase in liveness means the odds of a song being popular decrease by 0.5%
- A one-unit increase in valence means the odds of a song being popular increase by 63%
- A one-unit increase in tempo means the odds of a song being popular decrease 0.683%
- With a one-unit increase in time signature, the odds of a song being popular increase by 464%
- With a one-unit increase in duration, the odds of a song being popular increase by 13%

Confusion Matrix:

Out of the 200, 076 observations

- 136,089 (68%) were correctly predicted as true
- 50,047 (25%) were incorrectly predicted as true

- 8,123 (4%) were correctly predicted as false
- 5817 (3%) were incorrectly predicted as false

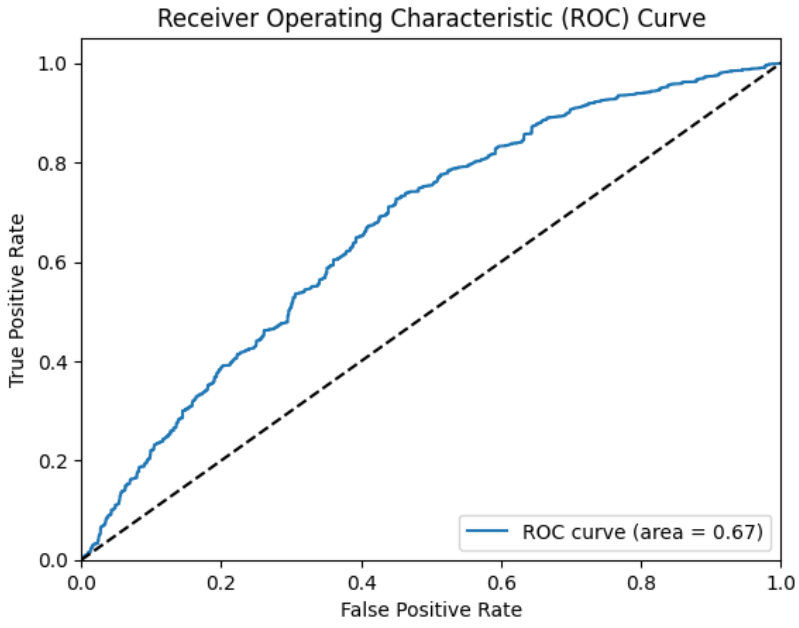


Logistic Model Confusion Matrix.

Model performance:

	precision	recall	f1-score	support
Predicted would not be popular	0.60	0.14	0.23	58170
Predicted would be popular	0.73	0.96	0.83	141906
accuracy			0.72	200076
macro avg	0.67	0.55	0.53	200076
weighted avg	0.69	0.72	0.66	200076

ROC curve:



Model ROC curve.

5. Discussion

5.1 Conclusions

The audio features for the top 10 songs show higher levels on energy and danceability, which means they are happier and upbeat songs. The speechiness and instrumentalness of the songs follow the average of the entire dataset. Most of the songs have a mode of 1, which means seven out of the 10 top songs have a major key.

The regional trends show that Europe has the most influence when it comes to top songs and artists. Latin America's presence is strong too, given that there are two songs in Spanish in the Top 10 and two non-English speakers in the top 10 artists, there is also one artist from Asia in the top 10.

North and Latin America are tied with the highest explicit song count, and Asia has the lowest one, which lets us know the preference for explicit language and how it is more accepted in some regions than others.

Latin America has the highest danceability and energy means, which lets us know in this area the happiest and most upbeat songs are well-liked. Africa has the highest instrumentalness, in other words, this region likes songs with vocals best. Lastly, Latin America and Africa have the highest speechiness, both these regions like songs with more words in them.

Based on model results, the variables that show higher odds of improving popularity are: loudness, mode, valence, time signature, and duration. The beta coefficients show that danceability and energy have a negative impact on popularity, meaning that if there is a one-unit increase in either one of them, the odds of popularity decreasing are 60% and 97%, respectively.

The model worked best at predicting the songs that would be popular. For precision, 73% of the observations predicted as "would be popular" were correct, and 58% of the observations predicted as "would not be popular" were correct. Recall had 96% where the instances predicted as "would be popular" were correct, and 14% of the instances predicted as "would not be popular" were correct. For F1-score: 83% of the observations predicted as "would be popular" were correct, and 23% of the observations predicted as "would not be popular" were correct.

The overall accuracy shows that the model correctly predicted 72% of the instances.

The Confusion Matrix shows that 68% of the test observations were correctly predicted as true, 25% were incorrectly predicted as true, 4% were correctly predicted as false, and 3% were incorrectly predicted as false.

The ROC curve shows that the AUC is 0.67 which indicates a fair model performance. The curve shows the trade-off between TPR and FPR at different classification thresholds. As the threshold decreases, more cases are classified as positive, leading to an increase in both TPR and FPR. Overall, performance could be improved.

5.2 Recommendations

Recommended song features to keep in mind to increase the popularity of a song:

- Loudness: The overall loudness of the song in decibels. The consideration should be to have a high loudness level.
- Mode: Indicates whether the song is a major or minor key. The database shows most songs (53.6%) have a major key.

- Valence: A measure of the musical positiveness conveyed by the song. A high-valence song (happy, cheerful, euphoric) would increase the chances of it becoming popular.
- Time signature: The estimated overall time signature of the song. A time signature of 4 or 5 would do best to increase the chances of the song being popular.
- Duration: the duration of the song. The model indicates that a one-unit increase in duration increases the odds of popularity by 13%, given that this percentage is not too high it should only work as a side note to take into account when considering song duration.

5.3 Next Steps

- The model shows a decent accuracy, and it works best at predicting songs that would be popular. So overall the performance could be improved.
- Danceability and energy results seem a little off given that the top songs' averages for these two variables are high. A second run would be best to confirm these two features really can have such a negative impact on song popularity.

6. Appendices

6.1 Variable overviews

Column	Description	Type
spotify_id	The unique identifier for the song in the Spotify database.	Chr
name	The title of the song	Chr
artists	The name(s) of the artist(s) associated with the song.	Chr
daily_rank	The daily rank of the song in the top 50 list.	Dbl
daily_movement	The change in rankings compared to the previous day.	Dbl
weekly_movement	The change in rankings compared to the previous week.	Dbl
country	The ISO code of the country of the Top 50 Playlist. If Null, then the playlist if 'Global Top 50'.	Chr
snapshot_date	The date on which the data was collected from the Spotify API.	Date
popularity	A measure of the song's current popularity on Spotify.	Dbl
is_explicit	Indicates whether the song contains explicit lyrics.	Lgl
duration_ms	The duration of the song in milliseconds.	
album_name	The title of the album the song belongs to.	Chr
album_release_date	The release date of the album the song belongs to.	Date
danceability	A measure of how suitable the song is for dancing based on various musical elements.	Dbl

energy	A measure of the intensity and activity level of the song.	Dbl
key	The key of the song.	Dbl
loudness	The overall loudness of the song in decibels.	Dbl
mode	Indicates whether the song is in a major or minor key.	Dbl
speechiness	A measure of the presence of spoken words in the song.	Dbl
acousticness	A measure of the acoustic quality of the song.	Dbl
instrumentalness	A measure of the likelihood that the song does not contain vocals.	Dbl
liveness	A measure of the presence of a live audience in the recording.	Dbl
valence	A measure of the musical positiveness conveyed by the song.	Dbl
tempo	The tempo of the song in beats per minute.	Dbl
Time_signature	The estimated overall time signature of the song.	Dbl

6.2 Summary Statistics Complete

spotify_id	name	artists	daily_rank	daily_movement
Length: 1467460	Length: 1467460	Length: 1467460	Min: 1.00	Min: -49.0000
Class: character	Class: character	Class: character	1st Q: 13.00	1st Qu: -1.0000
Mode: character	Mode: character	Mode: character	Median: 25.00	Median: 0.0000
			Mean: 25.49	Mean: 0.8088
			3rd Q: 38.00	3rd Qu: 2
			Max: 50.00	Max: 49

weekly_movement	country	snapshot_date	popularity	is_explicit
Min: -49.000	Length: 1467460	Min: 2023-10-18	Min: 0.00	Mode: logical
1st Qu: -3.000	Class: character	1st Q:2024-01-26	1st Q: 65.00	FALSE: 980005
Median: 0.000	Mode: character	Median: 2024-05-09	Median: 80.00	TRUE: 487455
Mean: 2.435		Mean: 2024-05-07	Mean: 76.38	
3rd Q: 5.000		3rd Q: 2024-08-18	3rd Q: 88.00	
Max: 49.000		Max: 2024-11-27	Max:100.00	

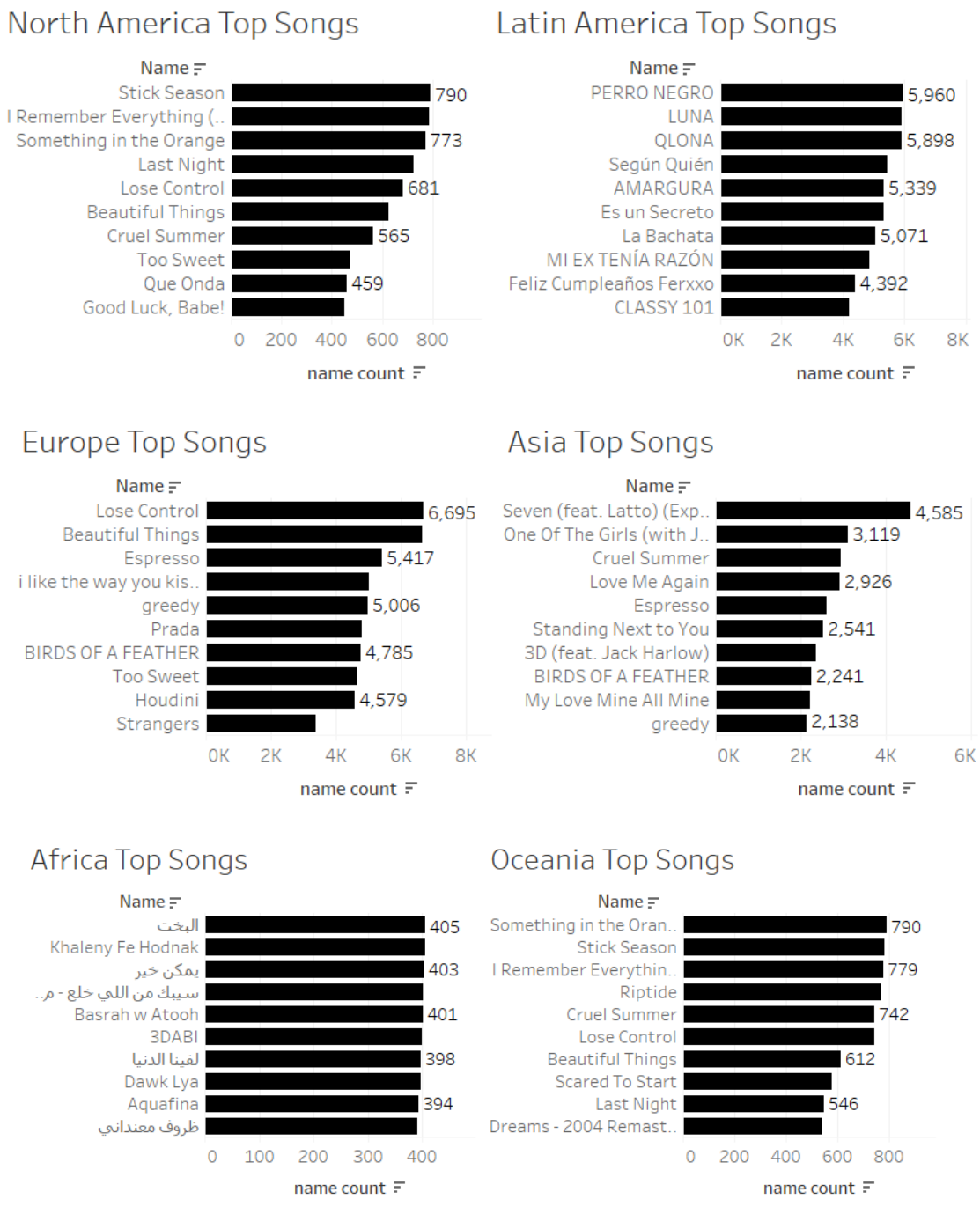
album_name	album_release_date	danceability	energy	key
Length: 1467460	Min: 1900-01-01	Min: 0.0000	Min: 0.0000201	Min: 0
Class: character	1st Q: 2023-04-27	1st Q: 0.5940	1st Q: 0.5530000	1st Q: 2
Mode: character	Median: 2023-11-09	Median: 0.7040	Median: 0.6720000	Median: 6
	Mean: 2022-04-08	Mean: 0.6848	Mean: 0.6536954	Mean: 5.54
	3rd Q: 2024-04-12	3rd Q: 0.7870	3rd Qu: 0.7660000	3rd Q: 9.00
	Max: 2024-11-25	Max: 0.9880	Max: 0.9980000	Max: 11.00

loudness	mode	speechiness	acousticness	instrumentalness
Min: -37.334	Min: 0.0000	Min: 0.00000	Min: 0.0000034	Min. :0.0000000
1st Q: -7.777	1st Q: 0.0000	1st Q: 0.03930	1st Q: 0.0661000	1st Q: 0.000000
Median: -5.967	Median: 1.0000	Median: 0.05820	Median: 0.1830000	Median:0.0000013
Mean: -6.419	Mean: 0.5361	Mean: 0.09479	Mean: 0.2704466	Mean: 0.0174593
3rd Q: -4.682	3rd Q: 1.0000	3rd Q: 0.11000	3rd Q: 0.4340000	3rd Q: 0.0000762

Max: 3.233	Max: 1.0000	Max: 0.93700	Max: 0.9960000	Max: 0.9770000

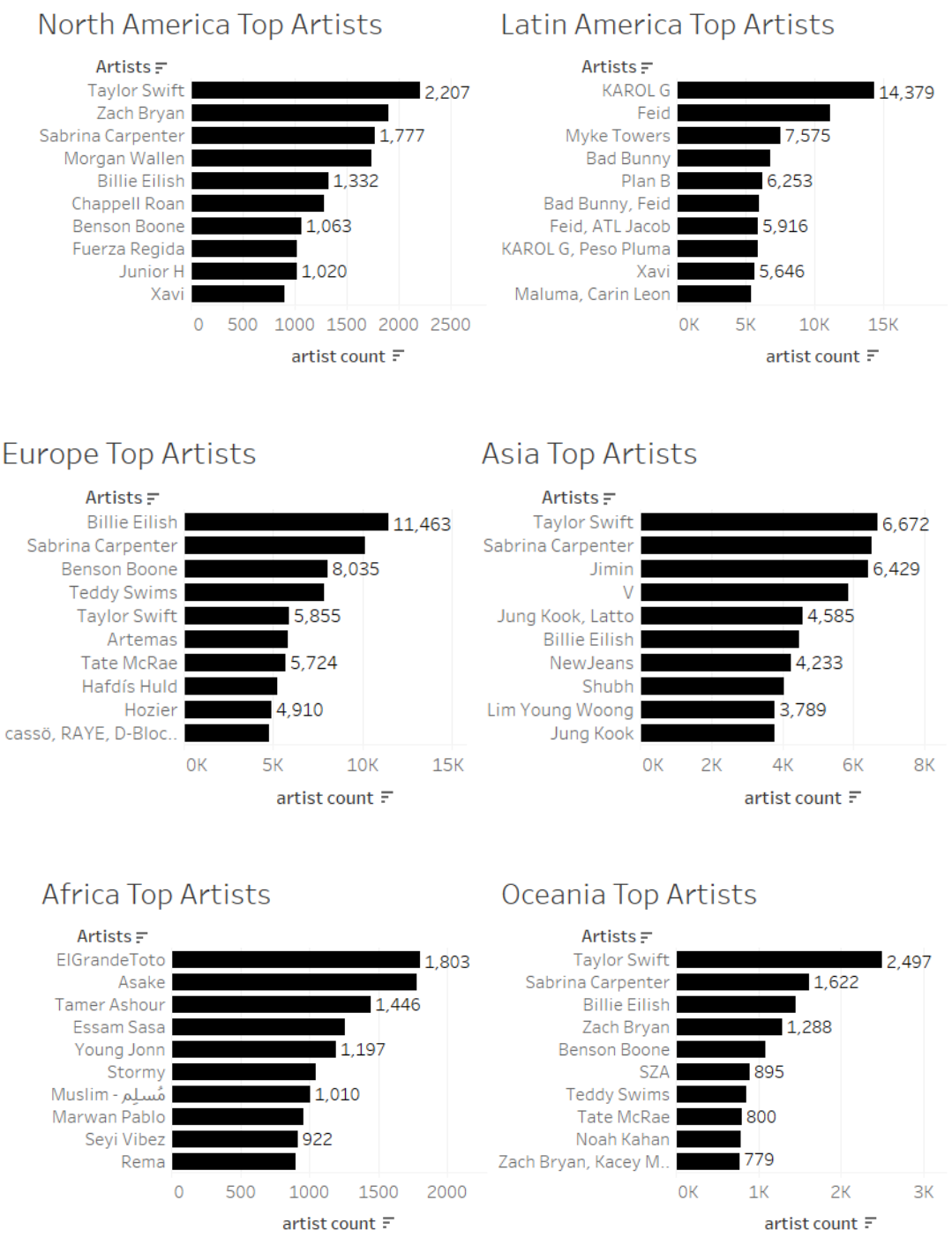
liveness	valence	tempo	time_signature	duration_min
Min: 0.0139	Min: 0.0000	Min: 0.0	Min: 0.000	Min: 0.272
1st Qu: 0.0958	1st Q: 0.3720	1st Q: 100.0	1st Q: 4.000	1st Q: 2.677
Median: 0.1200	Median: 0.5530	Median: 120.0	Median: 4.000	Median: 3.081
Mean: 0.1700	Mean: 0.5517	Mean: 122.2	Mean: 3.906	Mean: 3.206
3rd Qu: 0.2050	3rd Q: 0.7360	3rd Q: 140.1	3rd Qu: 4.000	3rd Q: 3.613
Max: 0.9780	Max: 0.9920	Max: 235.9	Max: 5.000	Max: 15.661

6.3 Top Songs by continent



Top Songs for North America, Latin America, Europe, Asia, Africa and Oceania

6.4 Top Artists by continent



Top Artists for North America, Latin America, Europe, Asia, Africa and Oceania.