Laboratory assignment

Component 2 - Data Analysis

Authors: Leordean Ada Alexandra, Selegean Victor
Group: 242

November 19, 2025

1 Requirements

This document presents an analysis of the features used in learning:

- correlation
- independence
- importance

Contains statistic considerations, data distributions, as well as their visualization and interpretation.

The dataset used is available on Kaggle and created by Solomon Ameh[Ame25] As of November 2025, the dataset has 4538 entries. As mentioned in the documentation of component 1, the missing values and duplicate entries make up an insignificant part of the dataset and can be dropped entirely.

2 Learning Features' Analysis

The relevant part of the dataset is made up of 14 numerical features, among which is a score from 0 to 100 representing each track's relative popularity, and one categorical feature, the track's genre. For the supervised learning task, the tracks' popularity represents the target variable. For the unsupervised learning task, the tracks' genres represent ”ideal” clusters.

For each of the 14 features, the statistical measures were computed and can be seen in table 1. Histograms were then generated in order to better visualize each range, as can be seen in figure 1. While some features such as ”energy”, ”tempo”, and ”danceability” approach something resembling a normal distribution, others such as ”valence” and ”mode” seem to map closer to a uniform one.

	Q1	Q2	Q3	Mean	Mode	Range	Variance
acousticness	0.05	0.22	0.59	0.34	0.12	1.00	0.11
energy	0.44	0.63	0.78	0.59	0.66	1.00	0.06
danceability	0.53	0.65	0.76	0.62	0.67	0.92	0.04
duration_ms	159000.00	194866.50	233478.00	206150.82	144000.00	1319885.00	6682335810.28
instrumentalness	0.00	0.00	0.20	0.20	0.00	0.99	0.12
key	2.00	5.00	8.00	5.23	1.00	11.00	12.82
liveness	0.10	0.12	0.20	0.17	0.11	0.96	0.02
loudness	-10.30	-7.19	-5.34	-9.28	-5.49	49.39	50.83
mode	0.00	1.00	1.00	0.56	1.00	1.00	0.25
speechiness	0.04	0.06	0.12	0.10	0.03	0.91	0.01
tempo	96.06	118.06	136.72	118.27	119.99	193.19	812.97
time_signature	4.00	4.00	4.00	3.94	4.00	4.00	0.17
track_popularity	41.00	56.00	72.00	54.76	68.00	89.00	393.17
valence	0.28	0.48	0.69	0.48	0.53	0.96	0.07

Table 1: Statistical Measures of the Numerical Features

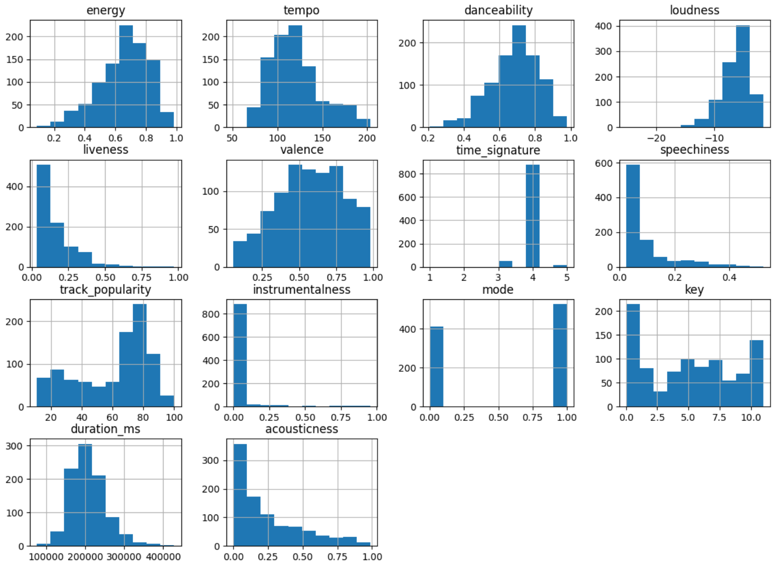


Figure 1: Histograms for the data set's numerical features

The large differences between feature ranges in table 1 suggest that, for further analysis, the features need to be normalized. This was achieved by using Scikit Learn's StandardScaler [scib]. Applying the transformation provided by StandardScaler's `fit_transform` method ensures that all features will have an expected value of 0 and a standard deviation of 1. Figure 2 shows the histograms of the features after the normalization process.

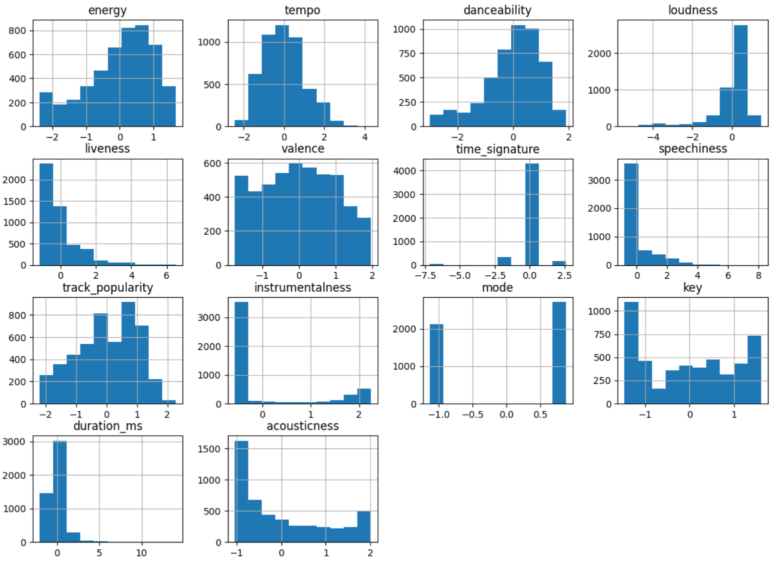
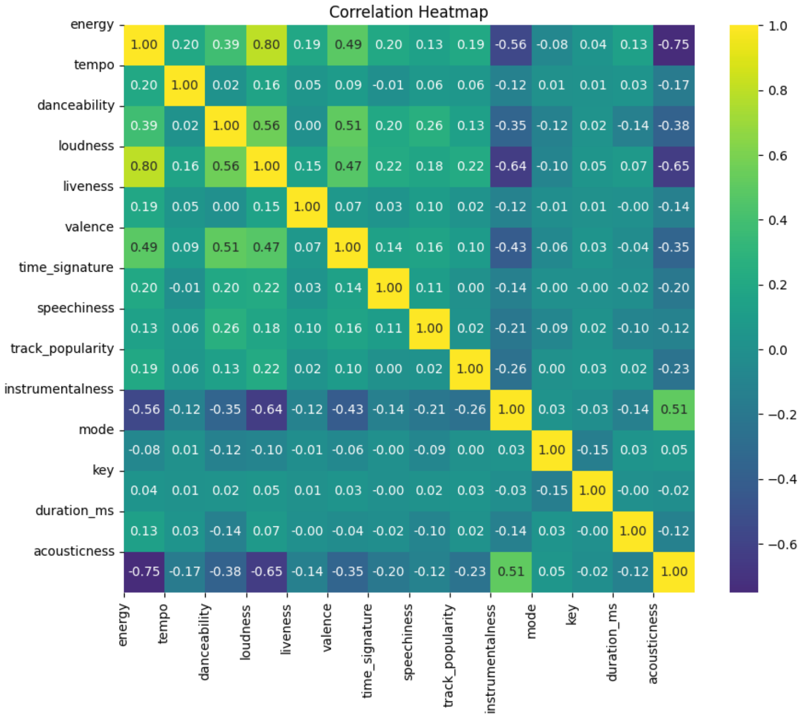


Figure 2: Normalized histograms for the data set's numerical features

The correlation between the numerical features of the dataset was analyzed using a heatmap created using Seaborn [sea], a Python data visualization library based on Matplotlib[mat]. Figure 3 shows the result of this process. The correlation coefficients range from 1.00 for each feature's correlation with itself to -0.75 for the (energy, acousticness) pair.



	energy	tempo	danceability	loudness	liveness	valence	time_signature	speechiness	track_popularity	instrumentalness	mode	key	duration_ms	acousticness
energy	1.00	0.20	0.39	0.80	0.19	0.49	0.20	0.13	0.19	-0.56	-0.08	0.04	0.13	-0.75
tempo	0.20	1.00	0.02	0.16	0.05	0.09	-0.01	0.06	0.06	-0.12	0.01	0.01	0.03	-0.17
danceability	0.39	0.02	1.00	0.56	0.00	0.51	0.20	0.26	0.13	-0.35	-0.12	0.02	-0.14	-0.38
loudness	0.80	0.16	0.56	1.00	0.15	0.47	0.22	0.18	0.22	-0.64	-0.10	0.05	0.07	-0.65
liveness	0.19	0.05	0.00	0.15	1.00	0.07	0.03	0.10	0.02	-0.12	-0.01	0.01	-0.00	-0.14
valence	0.49	0.09	0.51	0.47	0.07	1.00	0.14	0.16	0.10	-0.43	-0.06	0.03	-0.04	-0.35
time_signature	0.20	-0.01	0.20	0.22	0.03	0.14	1.00	0.11	0.00	-0.14	-0.00	-0.00	-0.02	-0.20
speechiness	0.13	0.06	0.26	0.18	0.10	0.16	0.11	1.00	0.02	-0.21	-0.09	0.02	-0.10	-0.12
track_popularity	0.19	0.06	0.13	0.22	0.02	0.10	0.00	0.02	1.00	-0.26	0.00	0.03	0.02	-0.23
instrumentalness	-0.56	-0.12	-0.35	-0.64	-0.12	-0.43	-0.14	-0.21	-0.26	1.00	0.03	-0.03	-0.14	0.51
mode	-0.08	0.01	-0.12	-0.10	-0.01	-0.06	-0.00	-0.09	0.00	0.03	1.00	-0.15	0.03	0.05
key	0.04	0.01	0.02	0.05	0.01	0.03	-0.00	0.02	0.03	-0.03	-0.15	1.00	-0.00	-0.02
duration_ms	0.13	0.03	-0.14	0.07	-0.00	-0.04	-0.02	-0.10	0.02	-0.14	0.03	-0.00	1.00	-0.12
acousticness	-0.75	-0.17	-0.38	-0.65	-0.14	-0.35	-0.20	-0.12	-0.23	0.51	0.05	-0.02	-0.12	1.00

Figure 3: Feature Correlation

From the graph, strong positive correlations can be identified between energy and loudness, danceability and loudness, and valence and loudness. Strong negative correlations seem to appear between energy and acousticness, loudness and acousticness, energy and instrumentalness, and loudness and instrumentalness. Contrary to initial expectations, instrumentalness and speechiness do not have a strong correlation, one way or another.

A Principal Component Analysis [scia] using the scikit-learn library was performed on the 14 numerical features of the dataset. These are the same features which appeared in figure 3. 14 PCs were computed, matching the dataset's dimensions. Figure 4 shows how much each PC contributes to the dataset's total variance. The first 5 of the 14 PCs contribute about 60% of the total, while the last 4 contribute less than 10%.

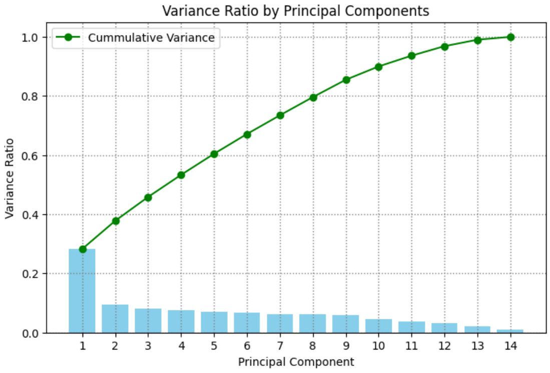


Figure 4: Relative and Cumulative Relative Variance for Principal Components

After that, a new correlation matrix was produced showing how much each feature plays into individual PCs (figure 5). By looking at the matrix' rows in relation to figure 4, it can be inferred that the most relevant distinction in the dataset is between tracks with high energy, danceability, loudness, and valence and those with high instrumentalness and acousticness. In the same way, it can be seen that the energy-loudness distinction accounts for a very small amount of variance.

Notable is the fact that the matrix as a whole shows no single feature dominates all principal components. For instance, PC2 is much more nuanced through speechiness, positive track popularity and duration in milliseconds in comparison to other principal components, which does suggest a different type of variation, more closely connected to the track's technical perception to the listener rather than its own creation characteristics. The following components show great differences between highest and lowest features, which makes it easier to distinguish more niche musical properties between them.

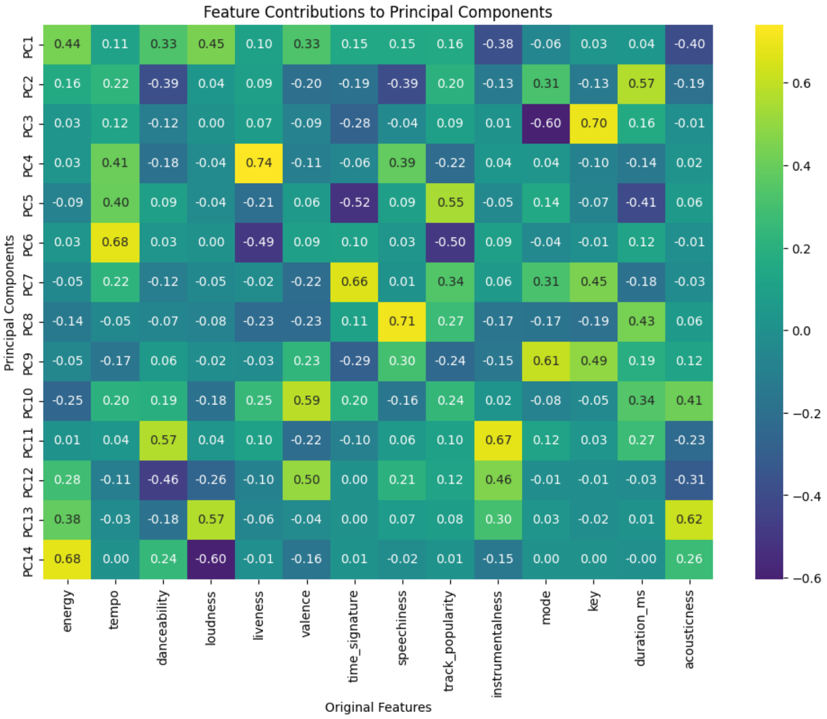


Figure 5: Feature correlation with PCs

In order to better understand the meanings in the graph above, specific graphs to aid visualize the data distribution have been created.

3 Data Distribution and Data Statistics

As seen in the previous section, the first two principal components have the greatest impact. As such, all tracks have been plotted through the viewpoint of different PCs (figures 6 and 8), with the complementary figure 7 that highlights the results from figure 4 on a small sample of tracks.

Figure 6 focuses on all tracks viewed through the first two principal components. The pattern is highly recognisable as the data forms an elongated cluster that extends from lower left regions to a much densely packed right, with upper right exceptions. Tracks on the left side (low PC1 values) correspond to low accousticness and instrumentalness, meaning low energy tracks, which those on the right side (high PC1) are characterised by loudness, energy, danceability and valence. This direction also represents the strongest axis of variation in the dataset. PC2 spreads the dataset vertically, mainly due to tempo and speechiness, which means tracks with unusual rhythm or vocal characteristics.

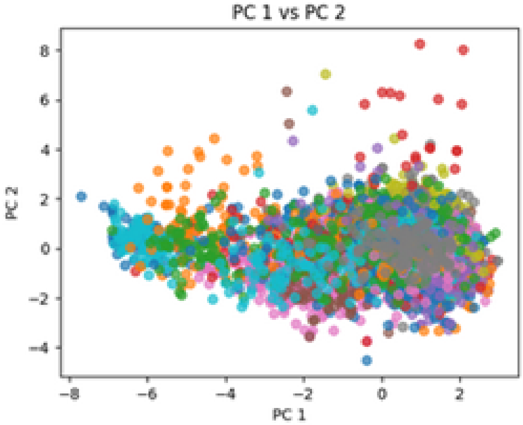


Figure 6: All data points viewed from the perspective of PC1 and PC2

While the overall shape is dense in the center towards lower right regions, the sparcity towards the edges incdicates that while most traks share common characteristics, some extreme values exist (outliers), which could impact the model performance and may require special handling while clustering.

Figure 7 focuses on tracks in the top ten most popular genres. Like this, not only is the shape still visibly similar to the previous figure, but it is also more easy to track genres based off their colour code and how they align with PC1 and PC2.

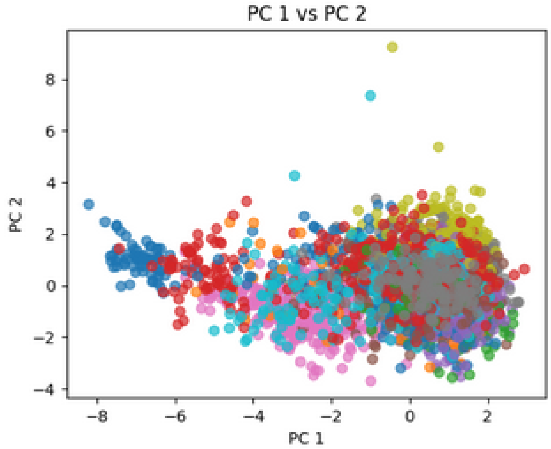


Figure 7: Data points from the most popular 10 genres viewed from the perspective of PC1 and PC2

Pop, hip-hop and punk tracks tend to appear towards the right, meaning high values of energy, danceability and loudness (positive PC1 values). In contrast, genres like ambient are concentrated towards the left, where acoustics and instrumentals are dominant.

This genre clustering also shows that some genres overlap substantially, which leads to shared musical characterstics, while more distinct and compact genres have a more individualistic profile. This shows the possibility of prediction through the features mentioned prior.

Figure 8 focuses on tracks viewed through the second and ninth principal components. Unlike figure 6, the structure is more more compact and has no directional trend. This indicates that PC9's influence in variation is rather small and does not strongly separate major musical characteristics. Additionally, the dense cluster around the center suggests that most tracks have similar values for the features interpreted through these components.

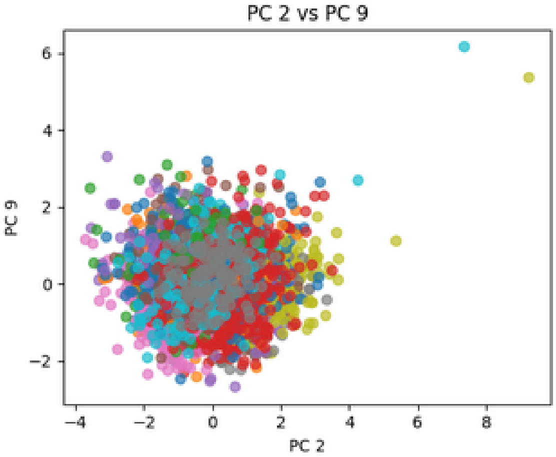


Figure 8: Data points viewed from the perspective of PC2 and PC9

The few isolated points correspond to tracks with extreme values from PC9, such as valence, speechiness or mode, which highlight rare or unconventional songs whose properties differ significantly from the mainstream patterns. This distribution shows that components beyond PC3 (as seen in figure 5) capture much more subtle or specific variations, rather than broad musical structures.

In order to better understand the plotted points above, the following section will revolve around some of the top genres of the available tracks in the dataset.

4 Data Visualization and Interpretation

Before we tackle specific genres, we first must observe the distribution of all tracks available in the dataset. Out of all 35 genres available, the following pie chart is limited to showing only the ones that quantify to more than 1% of data.

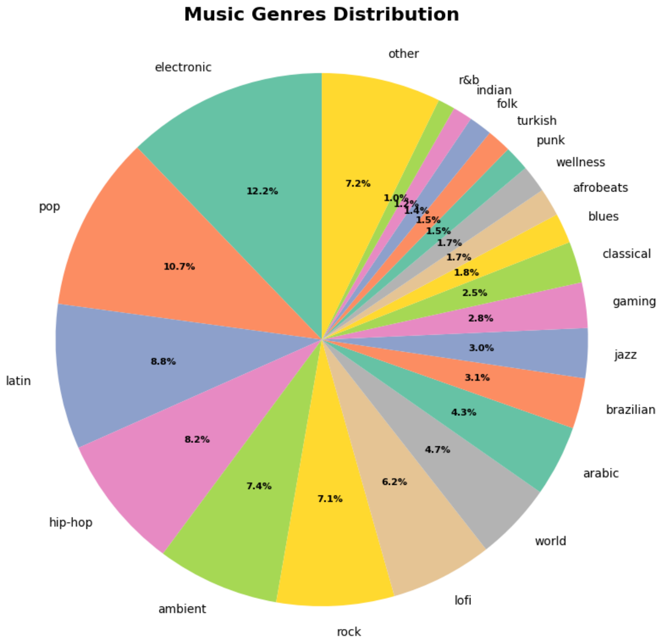


Figure 9: Genre distribution of all tracks

The first part of this section revolves around the top genres for the most popular tracks, as seen in figure 10. We note that, although electronic makes out to be a large quantity of music in the dataset, it is far below the most popular genres.

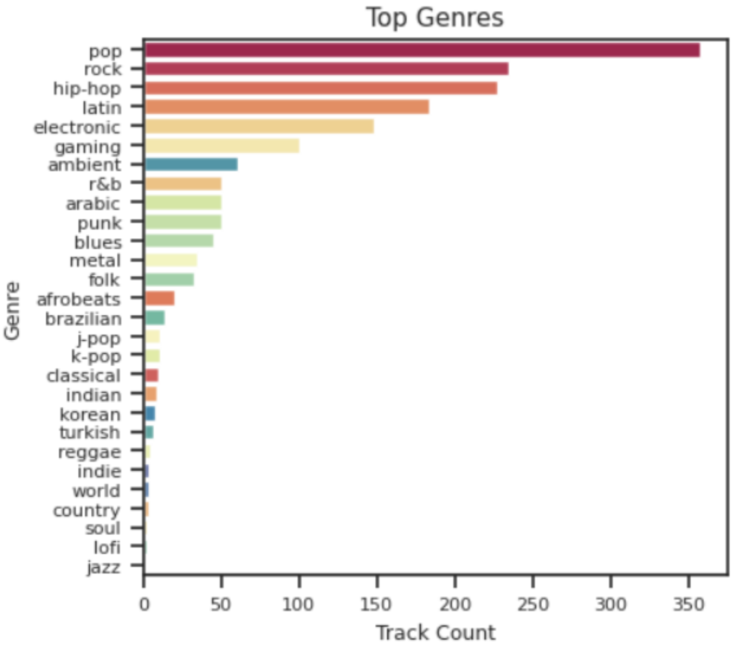


Figure 10: Top Genres for high popularity tracks

Pop, rock, hip-hop and latin take over the majority of the collection, amounting to a large portion of the tracks. These tracks share characteristics specific to the modern music such as loudness, well defined rhythm and relatively high danceability. Due to their representation taking over the majority of the high popularity tracks, it may influence some statistical measures.

Electronic, gaming, ambient and r&b cover a wide and rather drastic stylistic range, yet they make up the next top genres, from highly energetic, electronic tracks to slow, ambient pieces. Still, we notice that the genres with closer characteristics to those mentioned above are still far more numerically than the rest, which coincides with the public, general demand.

The remaining genres appear in smaller quantities and even almost infinitesimal in comparison to the rest, which reflect either niche audiences or limited dataset inputs. This distribution could lead to a class imbalance in genre prediction tasks.

In order to correlate it better to the previous section, a plot of the pop (yellow) and latin (purple) genres over PC4 and PC8 has been created. From a visual standpoint, the pattern for pop remains concentrated on the center-left side, spreading a little towards the right, while latin is majoritary fully centered, both with scattered, individualistic traits.

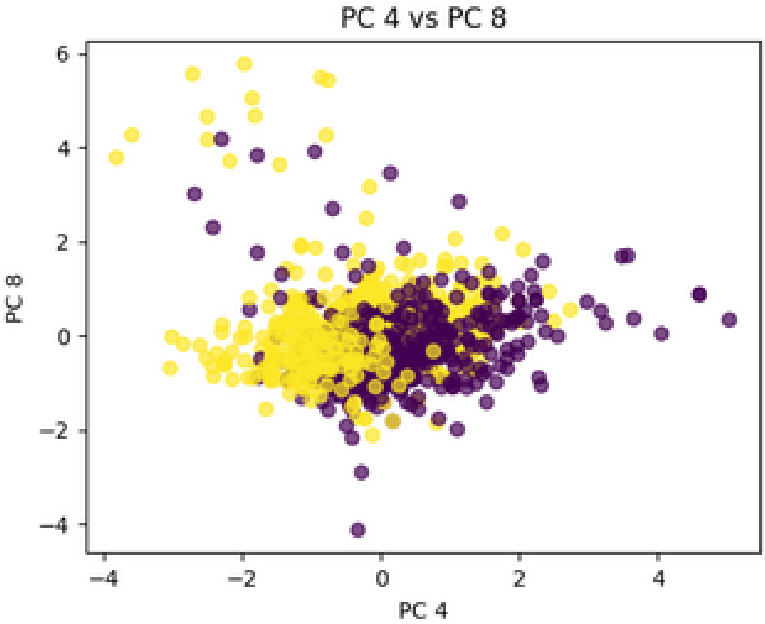


Figure 11: Difference between pop genre and latin over PC4 and PC8

For pop, we observe the fact that PC8 shows some strong characteristic for specific songs in this genre, which is mainly categorised by positive speechiness and track popularity over PC4, which correlates to a negative number. In contrast, PC4 shows a highly positive liveness and tempo over PC8's negative values, which prove to be much more influential over latin specific songs. The few latin songs on the negative side of PC8 could correspond to their valence, which represents the use of minor keys and low emotional positivity of a song. With pop being the most popular genre, it is of no surprise that it represents a high emotional positivity and major keys in more tracks, as it is preferred on a global level.

In the following part, unlike the top genres in the precedent text, the focus surrounds the top genres for low popularity tracks. Figure 12 shows that electronic music appears as the most frequent genre, whereas in the global popularity, it had only been in the fifth place. With over 400 tracks, electronic far surpasses the rest of the genres in terms of number of tracks, with ambient and lofi coming in second with around 300 tracks. It is interesting to note that these two genres are almost antithetic to the first; ambient and lofi tend to be instrumental, repetitive and even meant to set a certain mood to the listener, but these characteristics make it fitting for background music, which may contribute to the number in popularity. Latin and world music could be paired with a similar number of tracks as well, probably due to their regional constraints, where certain songs are more widely listened to within specific communities.

Most other genres amount to less than 200 tracks, which corresponds to high diversity of niche musical categories in the dataset.

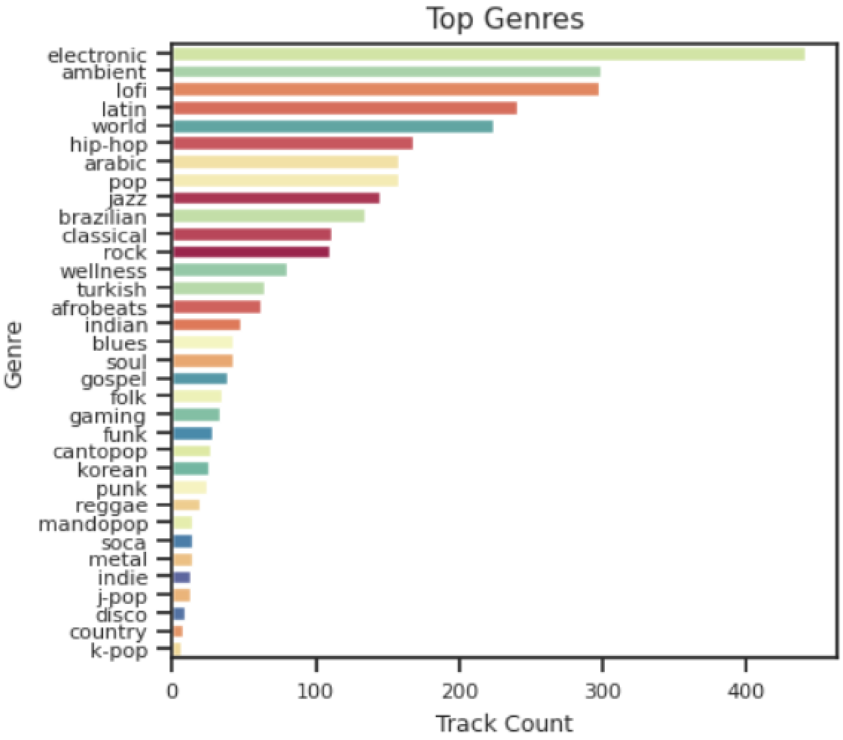


Figure 12: Top Genres for low popularity tracks

This separation between high and low popularity genres suggest that track popularity is influenced not only by audio features, but also by cultural context, listening habits and trends, besides the actual number appearing in the dataset.

References

[Ame25] Solomon Ameh. Spotify Music Dataset, 2025. Online; Accessed October 2025 at https://www.kaggle.com/datasets/solomonameh/spotify-music-dataset.

[mat] Matplotlib 3.7.10 documentation. online; Accessed November 2025 at https://matplotlib.org/stable/index.html.

[scia] scikit-learn 1.7.2 documentation. online; Accessed November 2025 at https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html.

[scib] scikit-learn 1.7.2 documentation. online; Accessed November 2025 at https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html.

[sea] Seaborn 0.13.2 documentation. online; Accessed November 2025 at https://seaborn.pydata.org/generated/seaborn.heatmap.html.