|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | | | |
|  |  |  | |  |
| what makes a popular song?  group-of-four | | |
| Project 1 | | December 2021 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Decorative | | | | |
|  |  |  | |  | |
|  | PRoject 1: Spotify  The aim of this project was to collect data and perform an EDA, presenting the results as a group. This project investigates what makes a popular song on the music streaming platform, Spotify.  To investigate this, the most steamed songs on the Spotify platform were chosen for analysis. | |  | |

|  |  |
| --- | --- |
|  |  |
| Introduction Spotify is an online streaming platform founded by Daniel Ek and Martin Lorentzon, in mid-2006. Spotify is currently the largest music streaming service across the globe, with almost 400 million monthly active users.  The aim of this project was to explore what makes a song popular by analysing the top 100 streamed songs on the streaming platform Spotify. In order to collate the required data, the Spotify API was utilised to return JSON metadata about the artists’ name, tracks, streams (millions), followers, and various audio features, directly from the Spotify data catalogue. | |
| Data collection The final file was created using the Spotify API to pull two dataframes:   * The first dataframe, Track information, was pulled from a Spotify playlist of the top 100 streamed songs. This dataframe contains basic track information, such as artist/album name and release dates. * This was followed by the Audio features, which pulled from the API's audio features function. This covers a wide range of information created using Spotify's own algorithms, such as danceability, acousticness, energy and valence (see Table 1, below).   Finally, the number of streams was collected from Wikipedia. | |  |

Table 1: Descriptions of Spotify Audio Features


Table 1

|  |  |  |
| --- | --- | --- |
| Findings 1: Duration, Followers and Streams |  | |
| The top 100 streamed songs on Spotify showed that the average duration was three and a half minutes long. Overall, the median for duration of the top 100 streamed songs was 3.50 minutes. The lower quartile for duration was 3.17 minutes long, whereas the upper quartile for duration was 3.90 minutes long. Therefore, the interquartile range of duration is 0.72 minutes.  Chart, scatter chart  Description automatically generated  As presented in the scatterplot above, the durations of the top 100 songs were compared against the number of streams they received. The R squared value was 0.00026483 which indicates that there was no relationship between the two variables. As displayed, the songs longer than the average sometimes performs better than songs shorter than the average in terms of the number of streams, whereas sometimes the opposite was evident.  To investigate the popularity of these songs further, the number of followers an artist has was considered as a result.  Chart, scatter chart  Description automatically generated  In order to gain a better understanding of the factors which make the top 100 streamed songs on Spotify popular, it was important to measure the number of followers an artist has against the number of streams their songs achieve. As illustrated in the scatterplot above, the number of followers an artist has against the number of streams the songs achieved also showed weak correlation. The R squared value for these variables was 0.12358. One reason could be that there were some artists in the dataset with a one-off viral song. For example, as shown in the table below, the artist of the third most streamed song has one of the lowest number of followers in the dataset.  Table  Description automatically generated  This also further indicated that there were other factors which the top 100 streamed songs have in common which show better correlation. Thus, a correlation matrix was created for the whole dataset to identify correlation coefficients between variables, which will be further explored below.  Chart  Description automatically generated Findings 2: Loudness and Energy **Is there a correlation between the loudness of track and its energy?**  **Loudness**: Overall Loudness of a track in decibels  **Energy**: Represents a perceptual measure of intensity and activity in a song  Meaning of Pearson correlation coefficient value (r). | Download Table  A Pearson’s Correlation test was chosen to measure the strength and direction of a potential linear correlation.  **Hypothesis**  **(Null Hypothesis):** There is no correlation between Loudness and Energy in the top 100 songs  H0 : r = 0  **(Alternative Hypothesis):** There is a correlation between Loudness and Energy in the top 100 songs  H1: r != 0  Significant Level = 5% (0.05) (assumed)  Critical Value(p) = 0.195  r = 0.74  R^2 = 0.54 -> x100 = 54%.  Thus, 54% of the variability in Loudness is explained by the variability in Energy. The remaining 46% is explained by the other variables that influence a track on the spotipy API.  **Conclusion:**  We can conclude from this test that our results are significant as the r value is greater than the critical value, so we would reject the null hypothesis. Concluding that there is a relationship between Loudness of a track and the Energy of a track  **Graphical Representation of the relationship between** **loudness of track and its energy:**    R squared: 0.5434088173021023 or 54%  There is a Strong positive correlation between the loudness of a track to its energy, the r value is 0.7371626803508858  **Treemap:**  A Treemap of the genre of the top 100 songs for Spotify.    We intended on making this interactive, however the time we had was limited.  To do this we would have had to create a multi-index Dataframe where we would have the first index as the main genres (Pop, Rap, Hip Hop and House and others) then have the other subgenres in the secondary index.  **Genre:**  Another visual representation of the top genres. | |

|  |  |
| --- | --- |
|  |  |

## Findings 3: Tempo, Energy and Acousticness

**Tempo:** The overall estimated tempo of a track in beats per minute (BPM).

**Energy:** Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity.

**Accousticness:** A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic.

Chart, box and whisker chart

Description automatically generatedWhen analysing tempo, the data tells us that median tempo is 115bpm. This bpm falls in the Allegro range (fast) most notably the pop genre fits this bpm rate. Looking at the data in a wider view, we see a lower quartile of 95bpm and an upper quartile of 136bpm. Showing an interquartile range of 40pbm. This shows there is a fairly wide tempo range within the most popular songs indicating tempo isn't a decisive indicator of popularity. Finally, the min (59bpm) and the max (186bpm) further iterate the previous argument as the bpm ranges from one extreme to the other.

Chart, box and whisker chart

Description automatically generatedWhen analysing energy, the data tells us that the median energy is 0.63. This energy level suggests that the majority of popular songs are slightly above average. Looking at the data in a wider view, we see a lower quartile of 0.52 and an upper quartile of 0.75. Showing an interquartile range of 0.23. This suggests that the majority of popular songs are those with a medium to high energy level. And so, songs with low energy levels tend to perform worse. However, the min (0.26) and the max (0.92) energy ratings show us that even if a song does have an extremely low energy rating, it can still perform well in popularity, and vice versa. The data shows one outlier at 0.11 bpm. This being Billie Eilish’s song, 'when the party's over'.

Chart, box and whisker chart

Description automatically generatedWhen analysing acousticness, the data tells us that the median is 0.15, an upper quartile of 0.44, and a lower quartile of 0.04. This tells us that the majority of the most popular songs tend to be less acoustic, meaning they are more lyric heavy. It is important to note that the max value of acousticness lies at 0.98. This shows that whilst the majority of data points lie in the lower acoustic ranges. It is still possible for songs to be acoustic heavy and still be very popular.

Chart, scatter chart

Description automatically generatedThe relationship between energy and tempo shows a minimal increase as illustrated by an r value of 0.28. The r squared value also shows a large distribution of data points from the regression line with a value of 0.08.

Chart, scatter chart

Description automatically generatedThe relationship between energy and acousticness is that of a strong negative one, this is validated by an r value of -0.7. Therefore, generally the less acoustic a song is, the more energetic it is in the most popular songs.

|  |  |
| --- | --- |
|  |  |
| Findings 4: Danceability & Valence How is danceability and valence represented in the 100 most streamed songs? Both of these variables are created using Spotify’s own algorithm, defined as follows:  **Danceability** - describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable.  **Valence** - A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).  In particular, I wanted to explore the relationship between the danceability and valence of a song, does a happy or more positive song mean that it more danceable?  **Summary Statistics**  A summary statistics table was created using the df.describe() function to get a feel of the shape of the data:    Overall, it appears that the danceability value of a song is higher than its valence. Looking at the mean for example, danceability is 0.67 compared to 0.50 for valence. For both variables, mean and median are quite similar (**danceability:** mean: 0.67, median: 0.69, **valence:** mean: 0.50, median: 0.48), suggesting that the distribution of data is quite symmetrical. | |
| **Visualising the Data**  **Histogram:**  A histogram can also explain the distribution of the two variables:    As mentioned above, the similar mean and median shows a symmetrical distribution for each variable. Valence peaks around 0.4, whereas danceability peaks around 0.8. Valence appears to be normally distributed, albeit with slightly more data in the right tail than the left. Danceability has a one-tailed poisson distribution, skewed towards the higher end of the scale. This shows that the songs in the top 100 tend to be more danceable.  **Box Plot:** | |
| A boxplot can describe the data even further:    The boxplots enforce the histogram’s findings that songs need to be at least slightly danceable to reach the top 100 in streams. Valence, on the other hand, is spread out across the board, suggesting that a popular song does not necessarily need to be positive.  **Correlation and Regression:**  Finally, to explain the relationship between the danceability and valence further a boxplot was created:    As expected, as the danceability level rises, as too does the valence. A correlation of 0.49 is moderately positive, but not as strong as expected. In addition, the r-square value of 0.24 suggests that the regression line is, in fact, a weak fit against the distribution of the data. There are songs that appear to go against the earlier assumptions.  **Examples of Songs that fit the model poorly**   * High Valence, Low Danceability: Shawn Mendes - Treat You Better (Danceability: 0.44, Valence: 0.75) * High Danceability, Low Valence: Kendrick Lamar - HUMBLE. (Danceability: 0.91, Valence: 0.42)   **Examples of songs that fit the model well**   * Low Valence, Low Danceability: Billie Eilish - when the party's over (Danceability: 0.37, Valence: 0.20) * High Valence, High Danceability: XXXTENTACION - Moonlight (Danceability: 0.90, Valence: 0.71) | |
|  | |
| Conclusion  The API was rather limited in terms of its practical use for data analytics. When extracting songs from a playlist you are capped at 100 songs, removing that ability to analyse large quantities of data, reducing validity. The other option was to extract random songs without a cap, however this left us to a random and small chance of using a good quality dataset. One major issue was that they didn’t have a live play count for the individual songs, meaning we had to pull this data from a less reliable source (Wikipedia), again damaging the validity of the analysis.  Whilst, Spotify allows for plenty of different analysis measures, as mentioned in this project, we were left to the Spotify algorithms interpretation of how these measures were calculated. Measures such as ‘acousticness’ are very open to interpretation and are difficult to pin a precise value to. We would have liked to delve into the artists other songs to see how their most popular songs compared to their less popular songs to see if there was any correlation.  Finally, we must understand that music itself is interpretive. Statistics alone aren’t able to justify why a song is so popular. There are external measures such as the emotional attachment people hold to an artist and similar phenomenon’s that aren’t available to pull from an API. | |