Final Project

Name: Kamala Priya Lingam

Topic: Spotify’s Top 50 Songs in the Year 2021

Section: INFO 3010.002

**Abstract**

Spotify is an audio streaming and media services provider founded on 23 April 2006 by Daniel Ek and Martin Lorentzon in Sweden. It is one of the largest music streaming service providers, with over 422 million monthly active users, including 183 million paying subscribers, as of March 2022. In this project, I’ll be working on top 50 songs in Spotify for the year 2021.

**Background Information and Explanation**

For the project I’ve used the dataset that’s already create by a person from Kaggle. Since the dataset just contains 50 songs which seems pretty biased, I’ve added 10 more songs so that the question I’m working on will make more sense while comparing the factors of a song that made them reach the page.

**Data Sourcing**

The data that I used for the project is from [Kaggle](https://www.kaggle.com) and it’s created by Hanna Yukhymenko.

Here’s is the link:  [Spotify Top 50 Songs 2021](https://www.kaggle.com/datasets/equinxx/spotify-top-50-songs-in-2021)

For the details on variable values for other songs that I added, I used: [Musicstax](https://musicstax.com).

The data contains of Top 50 most listened songs in the world for year 2021 in Spotify and additional 10 songs that I added to compare.

**Data Description**

This dataset contains 14 descriptive variables about the songs which have been a major reason for their popularity.

* Popularity - The higher the value the more popular the song is
* Danceability - The higher the value, the easier it is to dance to this song
* Energy - The energy of a song - the higher the value, the more energetic song
* Key - the key the track is in. Integers map to pitches using standard Pitch Class notation.
* Loudness (dB) - The higher the value, the louder the song
* Mode - indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived. Major is represented by 1 and minor is 0
* Speechiness - The higher the value the more spoken word the song contains
* Acousticness - The higher the value the more acoustic the song is
* Instrumentalness - the number of vocals in a song. The closer the value to 1.0, the more instrumental the song is
* Liveness - The higher the value, the more likely the song is a live recording
* Valence - The higher the value, the more positive mood for the song
* Tempo - the overall estimated tempo of a track in beats per minute (BPM)
* Duration - duration of the song in ms
* Time signature - an estimated time signature. The time signature (meter) is a notational convention to specify how many beats are in each bar (or measure).

**Predictive Model**

Songs have always been a part of everyday life. And one day I wondered what makes a song popular in this digital world. Through this, I’m going to find the main reasons behind their popularity whether it’s because of the artist or is it because of the danceability, liveness, and other qualities that these songs provide. I will also check the which variable is highest for the most popular song. And by the end, I would like to know which artist had most of the popular songs in that year.

**Data Analysis**

Firstly, I want to give a summary of the data, because the audience will get to know what the data contains, which will help them get a gist of the dataset.

Graphical user interface, text

Description automatically generated

On to the main thing, which I’m working for. Now, I’ve looked at many plots having popularity as my x axis and the other variables as my y axis. I’ve discovered that some variables have an increased curve as the popularity scale increased, which means these are the main reasons which led the songs for their popularity. Where as some variables have their peak for the lower popularity values. And some have decreased slope when I came to the end of the popularity scale(with higher values).

I’ll be showing two variables which weren’t a factor in the popularity of the song.

* Graphical user interface, application

  Description automatically generatedChart

  Description automatically generated

In this line graph, we can see that the liveness for popularity at 70 and 85 is almost same, which makes it not worthy to be considered as a factor for the popularity of any song.

* Graphical user interface, application, Word

  Description automatically generated

Chart, line chart

Description automatically generated

And in this case, we can clearly see a downward slope of danceability for the highest popularity scale.

After repeating this process for all the variables, I’ve seen a increased slope for three variables, which seems to be the reason behind the popularity of the songs. The two variables that have highest values as the popularity scale increased are: Loudness, and Energy.

Line graph for Loudness:

Graphical user interface, application

Description automatically generated

Chart, line chart

Description automatically generated

Line graph for Energy:

Graphical user interface, application

Description automatically generated

Chart, line chart

Description automatically generated

In the graphs above, we can see a gradual increase in the y-variable along with increase in x-variable scale. As the popularity scale increased, both energy and loudness increased which means that these are the reason for the popularity of songs.

Now, I want to check which song is the least popular from the 60 songs we have.

Graphical user interface, text, application

Description automatically generated

I want to create histograms for two of the variables to represent the descriptive variables about the songs in the dataset.

Graphical user interface, text, application

Description automatically generated

Chart, histogram

Description automatically generated

Here we can see that most of the songs have a speechiness between 0.05 and 0.15

Graphical user interface, application, Word

Description automatically generated

Graphical user interface, application, Word

Description automatically generated

In this histogram for popularity, we can see that most of the songs have the popularity value between 85-90.

I also want to see if there’s any correlation between popularity, danceability, tempo, and key. So, I’ve created a scatterplot matrix for these variables.

Diagram

Description automatically generated with medium confidence

Now, I’m going to predict the popularity using the variables: loudness, energy, danceability, and liveness using a model.

Graphical user interface, text, application

Description automatically generated

In this above picture, we can see many statistical information for the model.

**Questions**

1. Which variables made these songs popular when compared to the least popular ones?

* Loudness and Energy

1. Who were the most popular artists for the year 2021?

* Olivia Rodrigo, Ed Sheeran, Doja Cat

Graphical user interface, text, application, Word

Description automatically generated

**Conclusion**

Through this analysis, I found that among the 14 descriptive variables that I took into consideration, there were two string variables which influenced the popularity of the songs. Even though there are many other factors that need to be taken into consideration, these are the easily accessible ones which we can work on. So, when we’re going to decide which songs are popular for the current and future years we can take these two into the consideration in predicting the most popular songs in seconds.