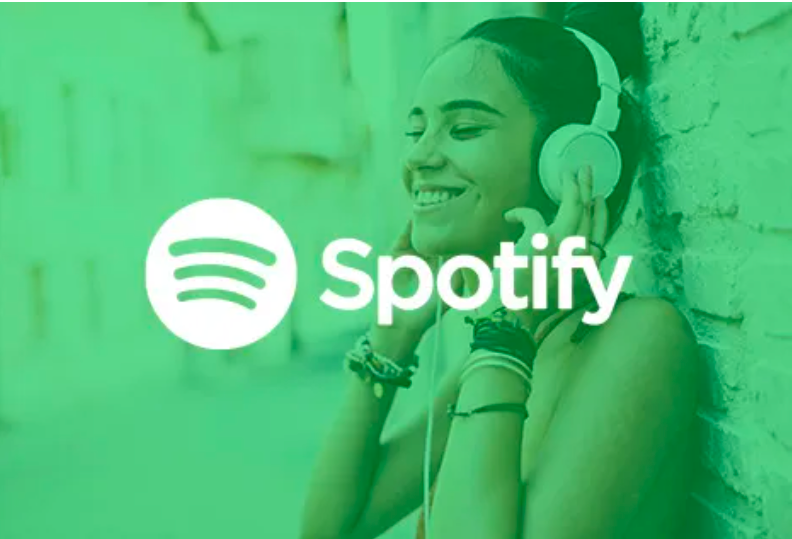
December 1, 2022



**Final Project - SPOTIFY**

**Assignment M10.7**

aJANYA SHARMA

BIA-610 APPLIED ANALYTICS

Stevens Institute of Technology – Summer 2021

Table of Contents

[1 Introduction 3](#_Toc102914244)

[1.1 The problem statement 3](#_Toc102914245)

[1.2 problem statement ScOPE 3](#_Toc102914246)

[1.3 Approach 3](#_Toc102914247)

[1.4 use case for the analysis 3](#_Toc102914248)

[2 Data preparation 4](#_Toc102914249)

[2.1 Source of the data 4](#_Toc102914250)

[2.2 data processing & cleaning 7](#_Toc102914251)

[2.3 final representation of the data 8](#_Toc102914252)

[2.4 Data summary 9](#_Toc102914253)

[3 exploratory data analysis 10](#_Toc102914254)

[3.1 the exploration 10](#_Toc102914255)

[3.2 The findings 11](#_Toc102914256)

[3.2.1 Artists and the albums released 11](#_Toc102914257)

[3.2.2 Correlation between artist, track name and their most popular track 12](#_Toc102914258)

[3.2.3 Playlist genre and their popularity over the decades 13](#_Toc102914259)

[3.2.4 Albums released per decade 14](#_Toc102914260)

[3.2.5 Average popularity of tracks released per decade 15](#_Toc102914261)

[3.2.6 Top popular tracks ever by Artist name and track name 16](#_Toc102914262)

[3.2.7 Correlation of Popularity with music characteristics 16](#_Toc102914263)

[3.2.8 Correlation between intruementalness, valence and genre 18](#_Toc102914264)

[3.2.9 Correlation between energy, danceability and genre 18](#_Toc102914265)

[3.3 Graphical representations 19](#_Toc102914266)

[3.3.1 Analysis of artists, albums, tracks, and their popularity 20](#_Toc102914267)

[3.3.2 Correlation between genre, track, playlist, artist, and popularity 21](#_Toc102914268)

[3.3.3 ﻿Correlation between Popularity vs genre vs valence and intruementalness 22](#_Toc102914269)

[3.3.4 Correlation between ﻿Popularity vs genre vs energy and danceability 23](#_Toc102914270)

[3.4 Tabular representations 23](#_Toc102914271)

[3.5 insight and analysis 23](#_Toc102914272)

[4 Summary 25](#_Toc102914273)

[4.1 Summarizing the problem statement 25](#_Toc102914274)

[4.2 adressal of the problem statement 25](#_Toc102914275)

[4.3 summary of final insight & analysis and implications 25](#_Toc102914276)

[4.4 Limitations & recommendationS 26](#_Toc102914277)

[4.4.1 Limitations 26](#_Toc102914278)

[4.4.2 Recommendations 26](#_Toc102914279)

Table of Figures

[Figure 1: General Overview of the Dataset 7](#_Toc102594991)

[Figure 2: Data Cleaning Removing Duplicates 8](#_Toc102594992)

[Figure 3: Processed Overview of the Dataset 9](#_Toc102594993)

[Figure 4: Example of Categorical Values in the Dataset 10](#_Toc102594994)

[Figure 5:Example of Numerical Values in the Dataset 10](#_Toc102594995)

[Figure 6: Artists and the number of albums released 12](#_Toc102594996)

[Figure 7: Findings for Correlation between track character and Popularity 13](#_Toc102594997)

[Figure 8: Playlist genre and their popularity over the decades 14](#_Toc102594998)

[Figure 9: Albums released per decade 15](#_Toc102594999)

Figure 10: Average popularity of tracks released per decade………………………………………………………16

Figure 11:[Top tracks ever and by whom……………………………………………………………………………………..16](#_Toc102595000)

[Figure 12: Correlation of popularity with musics characterstics 17](#_Toc102595001)

[Figure 13:Correlation between instruementalness,valence and genre 18](#_Toc102595002)

[Figure 14: Correlation between energy,danceability and genre………………………………………… .19](#_Toc102595003) [Figure 15: Analysis of artists, albums, tracks, and their popularity 21](#_Toc102594997)

[Figure 16: Correlation of popularity with music characteristics 22](#_Toc102594998)

[Figure 17: ﻿Correlation between Popularity vs genre vs valence and intruementalness………………..23](#_Toc102594999)

Figure 18:Correlation between Popularity vs genre vs energy and danceability………………………….24

# Introduction

## The problem statement

There are tons of artists coming up with new music every other day and they create music of different genre such as edm, rock, pop, r&b, rap, and Latin. With so much music being made for the last 7 decades, the objective is to identify what type of music was made, which tracks became a hit and which characters tics of music like danceability, liveliness impacted a track being popular.

## problem statement ScOPE

The data from Spotify provides us with the perfect resource to identify which track was a hit among music lovers and why it became popular. This will also help us identify the correlations between different music characteristics such as danceability, energy, loudness, liveliness, and valence to name a few which are preferred by the listeners. We further dive deep to identify the trends in the taste of music genres over last 7 decades and the albums released by the different artists. This analysis can be extended to further study the impact of artists’ collaboration and its impact on a track being popular. This project aims at creating Dashboards that will provide insights on the music industry which may help music companies and artists to identify their future music creations and collaborations.

## Approach

We are doing exploratory data analysis on Spotify data to discover patterns, to spot anomalies with the help of some statistics and graphical representations performed to visualize trends. This is done in four stages starting with Data preparation which involved data cleaning, Loading the data set to Tableau, analyzing data, and creating visualizations and identify information from the data. We try to explore each of these through excel and Tableau Dashboards in the following sections of the report.

## use case for the analysis

This Spotify data can be leveraged by music companies and artists to understand user preferences for different genres of music. Based on the user behavior on the platform, Music companies can identify how the music listing trends have been changing over decades which will allow them to collaborate with artists and come with music which will resonate more with the users. They will also be able to gain other insights on the characteristics of music like livability, danceability, energy and loudness contribute to a track being popular and their correlation with each other and across different genres. These insights will provide music companies and artists with great information about customer response to music and in turn help them monetize their music more effectively in future.

# Data preparation

## Source of the data

This data is [sourced](tihttps://github.com/rfordatascience/tidytuesday/blob/master/data/2020/2020-01-21/readme.md) from TidyTuesday repository on GitHub. To be precise, it is extracted from the [Spotifyr Package](https://www.rcharlie.com/spotifyr/) authored by Charlie Thompson, Josiah Parry, Donal Phipps & Tom Wolf. The data explained

The data involves various parameters which are classified into character and double type. Most of the variables which are non-metric such as track\_name, track\_artist is classified as characters and metric variables are of double type. The following is the data dictionary for the dataset.

| **Variable** | **Class** | **Description** |
| --- | --- | --- |
| track\_id | character | Song unique ID |
| track\_name | character | Song Name |
| track\_artist | character | Song Artist |
| track\_popularity | double | Song Popularity (0-100) where higher is better |
| track\_album\_id | character | Album unique ID |
| track\_album\_name | character | Song album name |
| track\_album\_release\_date | character | Date when album released |
| playlist\_name | character | Name of playlist |
| playlist\_id | character | Playlist ID |
| playlist\_genre | character | Playlist genre |
| playlist\_subgenre | character | Playlist subgenre |
| danceability | double | 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. |
| energy | double | Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. |
| key | double | The estimated overall key of the track. Integers map to pitches using standard Pitch Class notation. |
| loudness | double | The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks. Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typical range between -60 and 0 db. |
| mode | double | 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 | double | Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks. |
| acousticness | double | A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic. |
| instrumentalness | double | Predicts whether a track contains no vocals. “Ooh” and “aah” sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly “vocal”. The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0. |
| liveness | double | Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live. |
| valence | double | 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). |
| tempo | double | The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration. |
| duration\_ms | double | Duration of song in milliseconds |

Below is screenshot of the Spotify data in the spreadsheet:

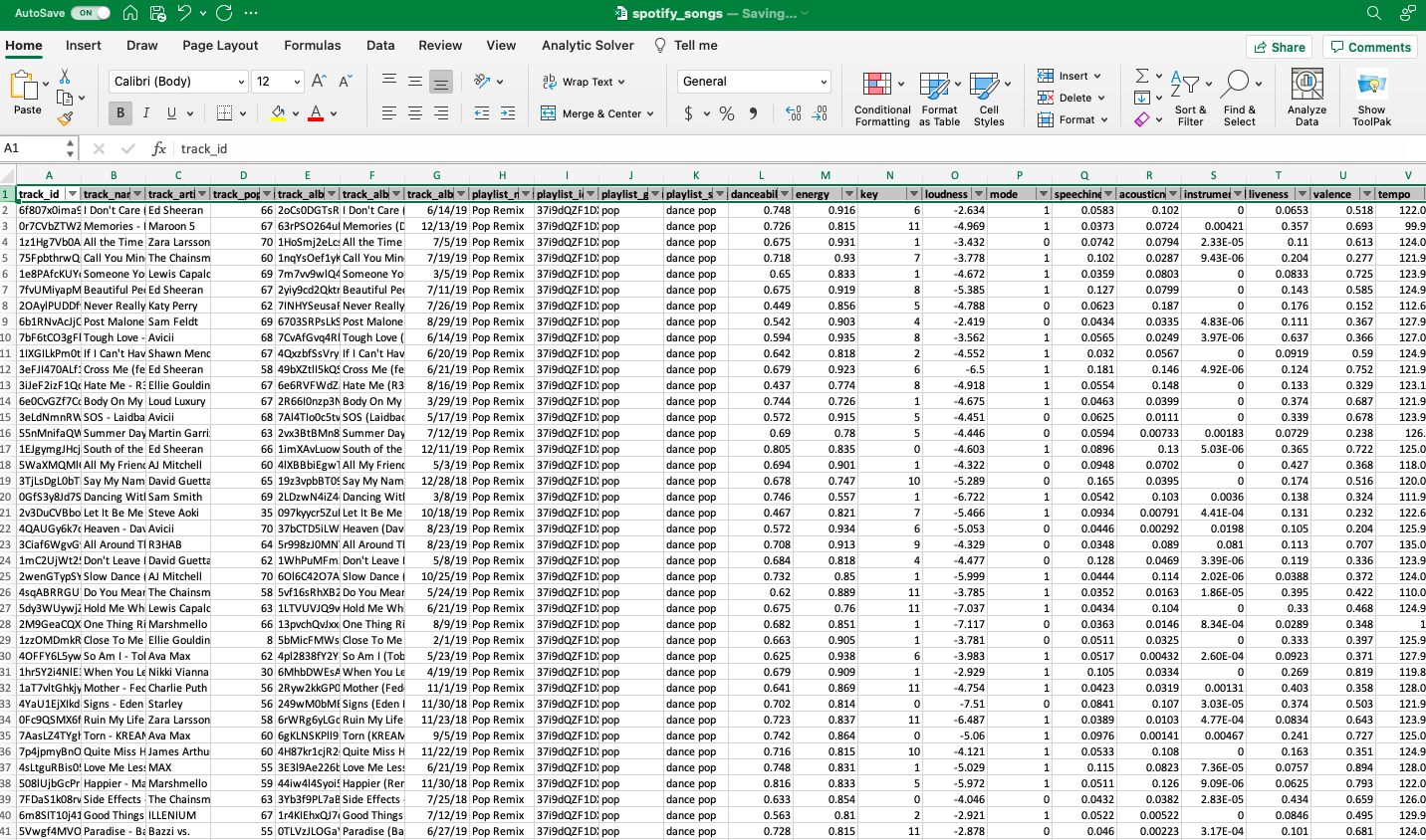


Figure : General Overview of the Dataset

## data processing & cleaning

Cleaning and converting raw data before processing and analysis is known as data preparation. It's a crucial stage before processing that often include reformatting data, making data changes, and integrating data sets to improve data quality.

For the Spotify data set, we have tried and looked for various inconsistencies in our data set and removed them so that we can do analysis on good quality data. We have performed the following operations on the data in excel:

1. Duplicates: No duplicates were found in the data set. This operation was performed using excel inbuilt functionality of “removing” duplicates from data tab.
2. Missing values or null values: The data was very consistent as no missing value or null values were encountered in the dataset.

The process was executed in excel using formulas, functional columns, and cleansing tables. A snapshot of an example is attached below,

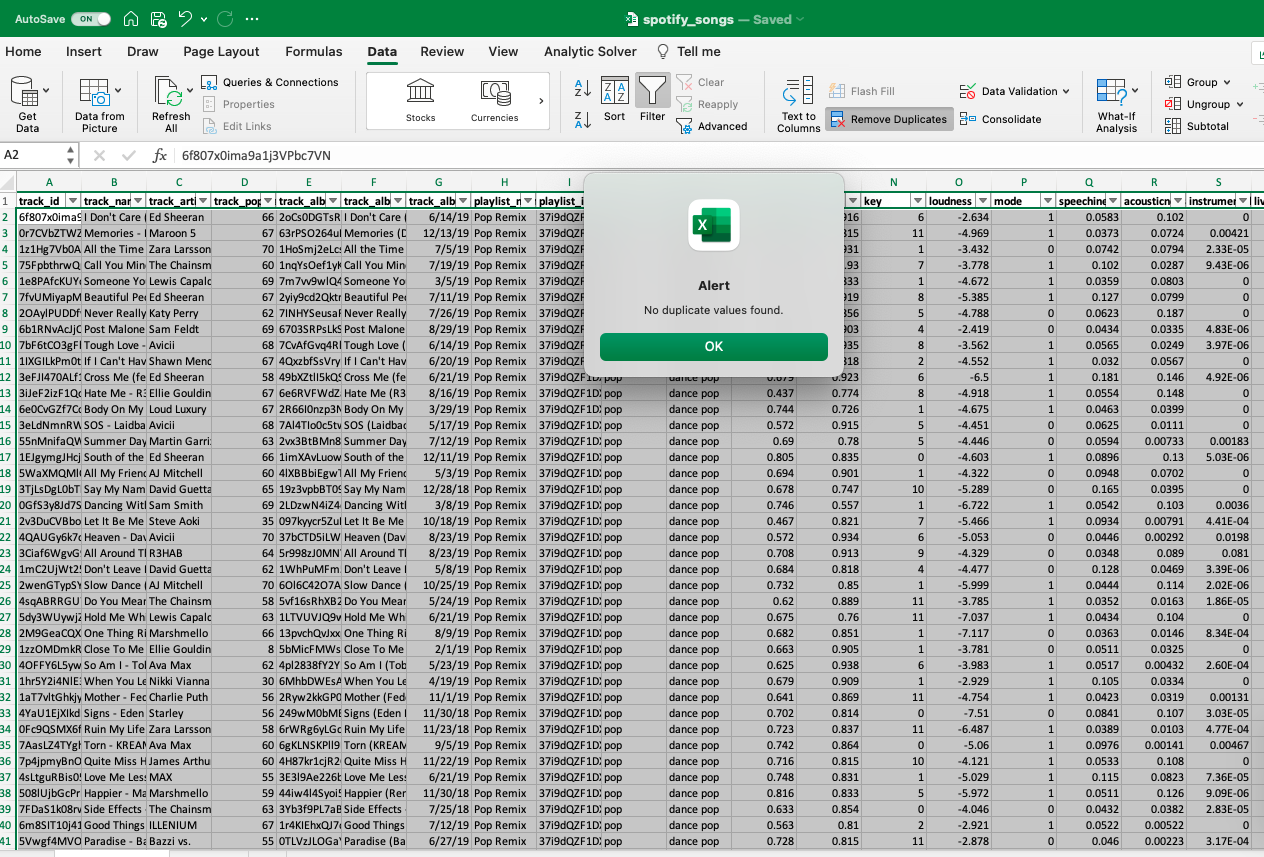


Figure : Data Cleaning Removing Duplicates

The only major problem faced during this stage was there inconsistency in data of few dates column track\_album\_release\_datewhich was taken care by using exclude Null functionality provided by tableau while doing visualizations which will be discussed in the following sections.

## final representation of the data

The following is the final representation of the data post cleansing and pre-processing and is loaded in tableau ready for analysis.

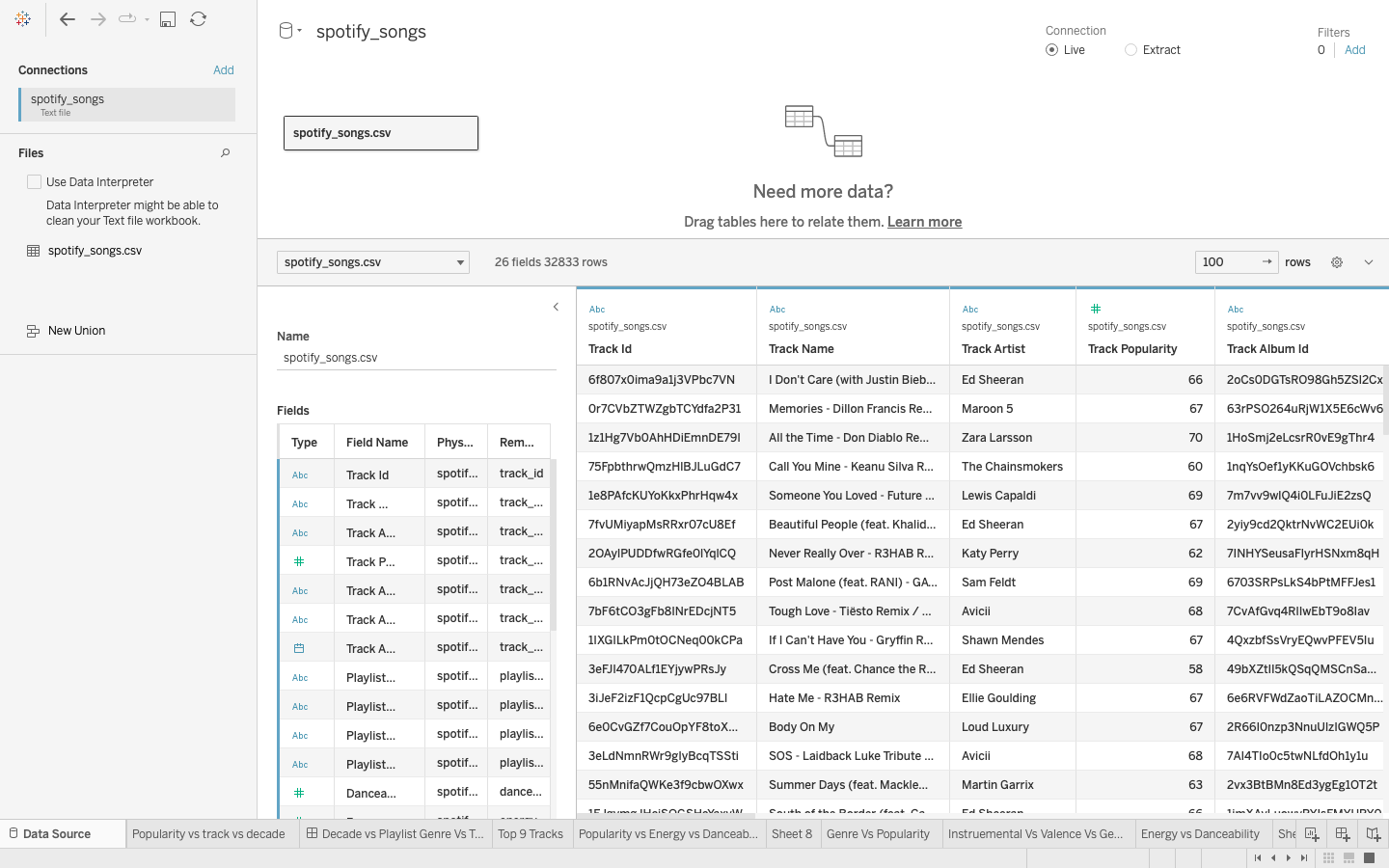


Figure : Processed Overview of the Dataset

## Data summary

This data comes from the TidyTuesday repository on GitHub. It was produced by Charlie Thompson, Josiah Parry, Donal Phipps, and Tom Wolf and is based on the Spotifyr Package. They also made it easier to acquire your own data or generic metadata about music using the Spotify API. Character and Double classes are used to represent a variety of variables, including metric and non-metric variables, dates, and other measures. The data dictionary for the dataset is listed below. This gives us a general idea of the database's purpose and nature. This helps us orient the study and make any necessary tweaks or changes based on our specific analytical goals.

# exploratory data analysis

## the exploration

Data exploration is an initial step in data analysis where we will try to get an overview about the data set and try to identify the properties of data and look for some patterns in the data.

When I did exploratory analysis, it was evident that the Spotify data set had categorical (non-metric), double integer (Metric) and date field which made up the entire data set.

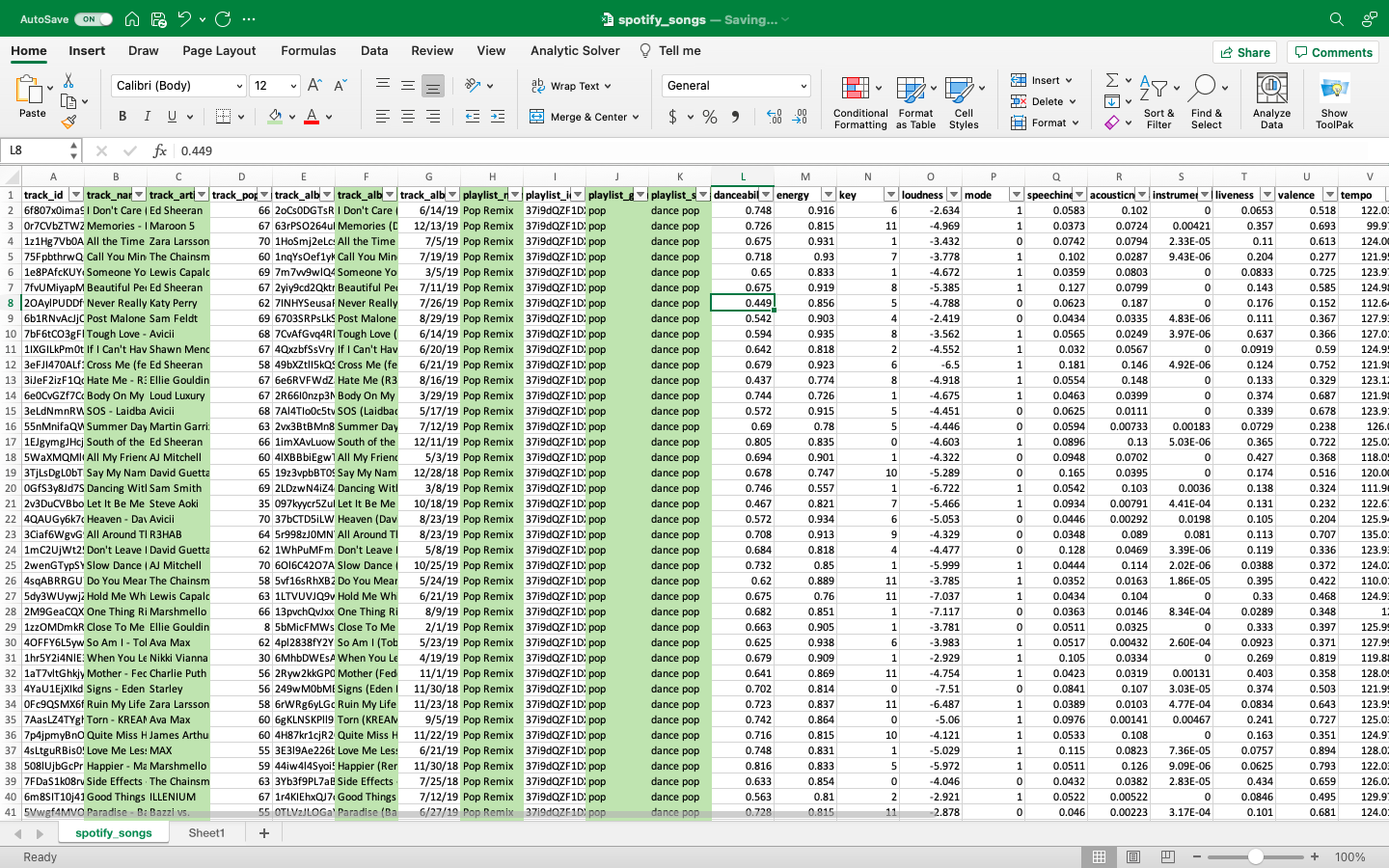


Figure : Example of Categorical Values in the Dataset

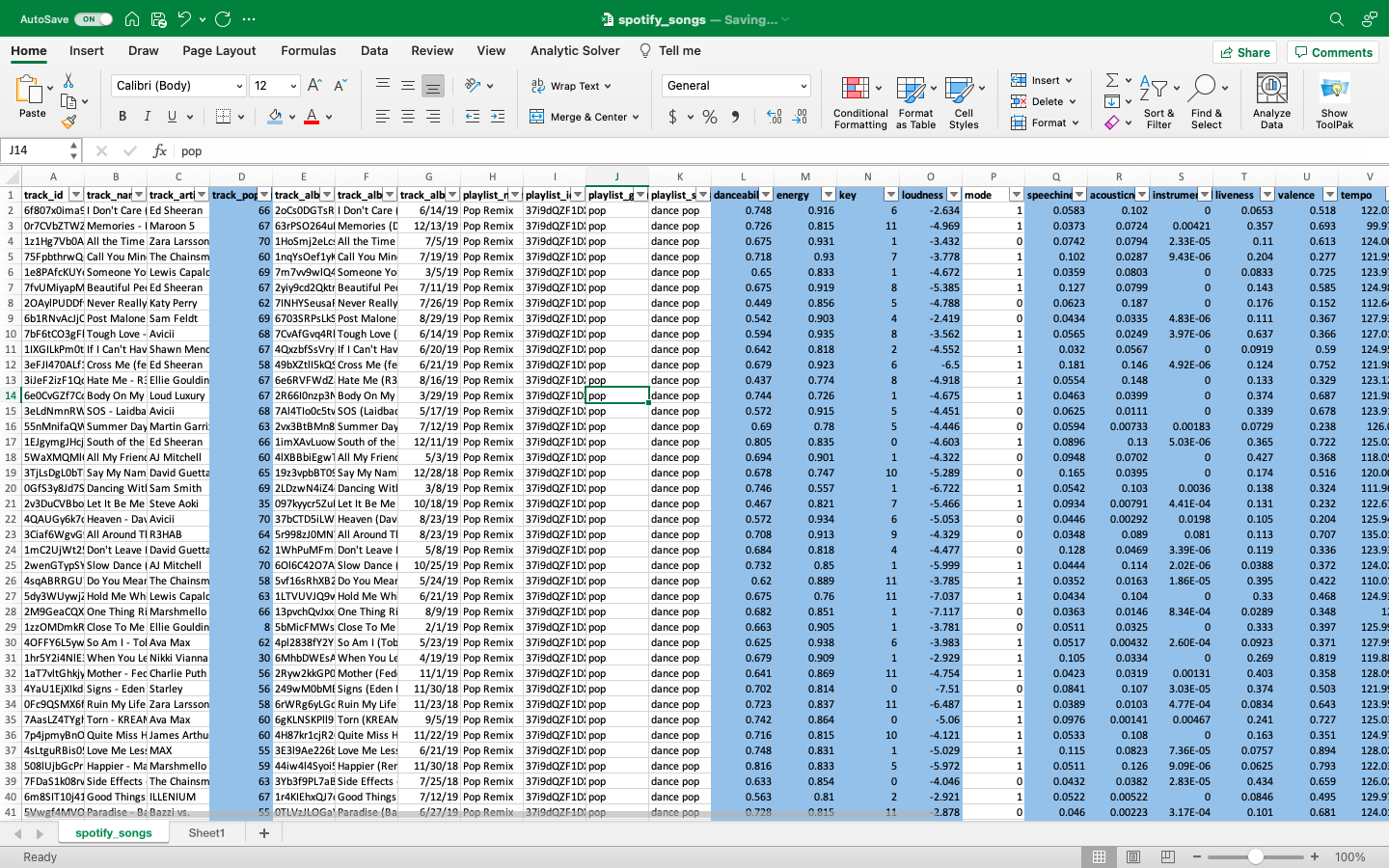


Figure :Example of Numerical Values in the Dataset

When you look at the data dictionary, you'll notice that some values range from 0 to 1, while others range from 1 to 10. Normalizing some of these values would have been ideal, but it didn't seem necessary in this case.

Text

Description automatically generated

## The findings

Once the data has been loaded, we have used tableau to do visualization and try to identify the patterns in the data set. I have analyzed and visualized data and arrived at the conclusion that **time series and correlations** will be very helpful in our data analysis. The following findings have been accumulated and listed below in this section.

### Artists and the albums released

The bar graph shows the number of distinct albums released by the artists shown in decreasing order. We can see that Maroon 5 has the highest 25 albums released among all the artists followed by The Weekend who has 14 releases, almost half of the album release count. Also, a filter has been added in the visualization to see if we want to check for some artist.

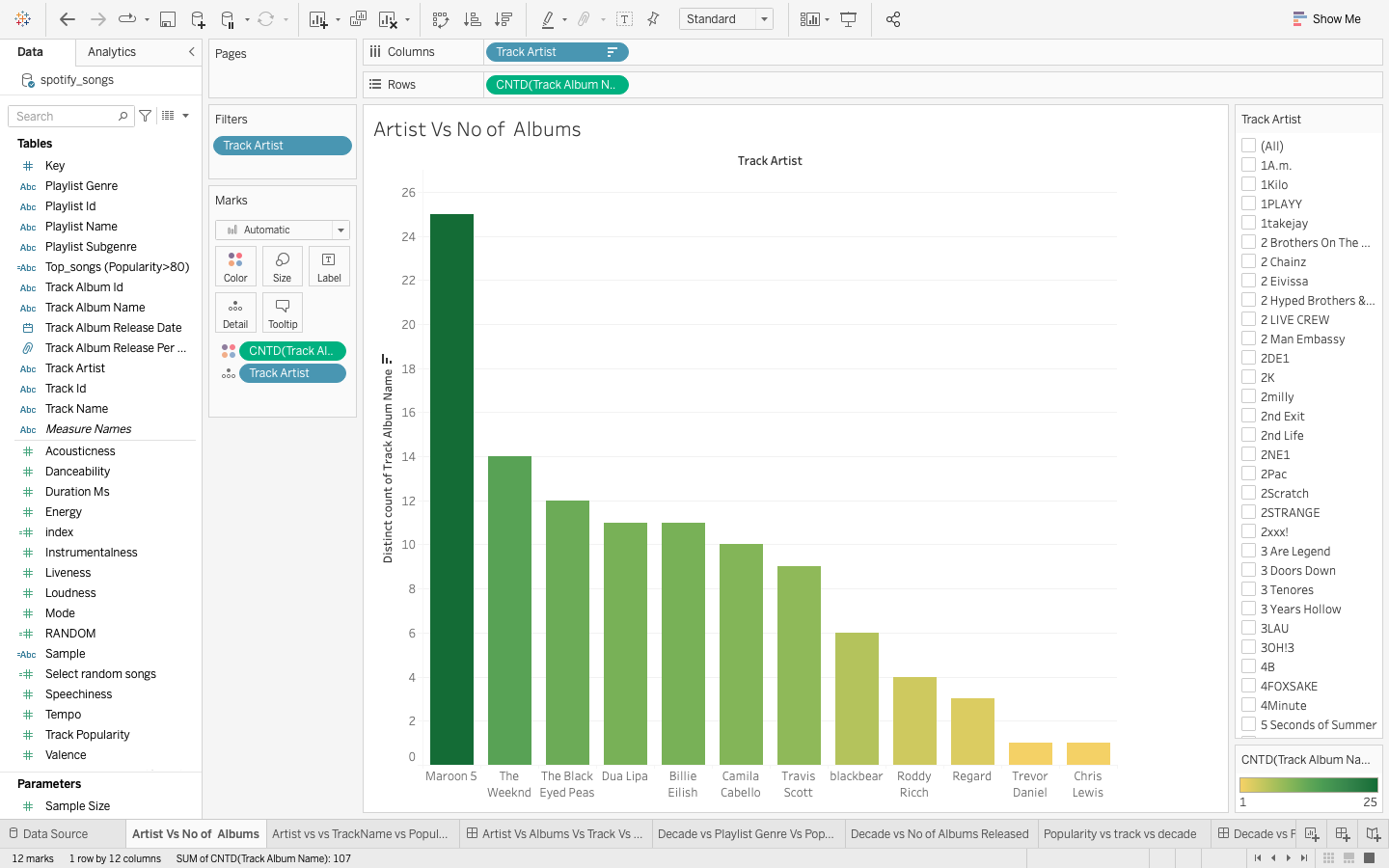


Figure : Findings for artists and number of albums released

This decrease has been consistently for all other artists so it highly unlikely that the artists prefer releasing lot of albums instead of making less but popular tracks.

### Correlation between artist, track name and their most popular track

This visualization helps us to find what is the most popular track for an artist based on the filter of artist applied on the data. The chart also gives information about the for example: Falling is the most popular track for Trevor Daniel with a popularity of 97.00

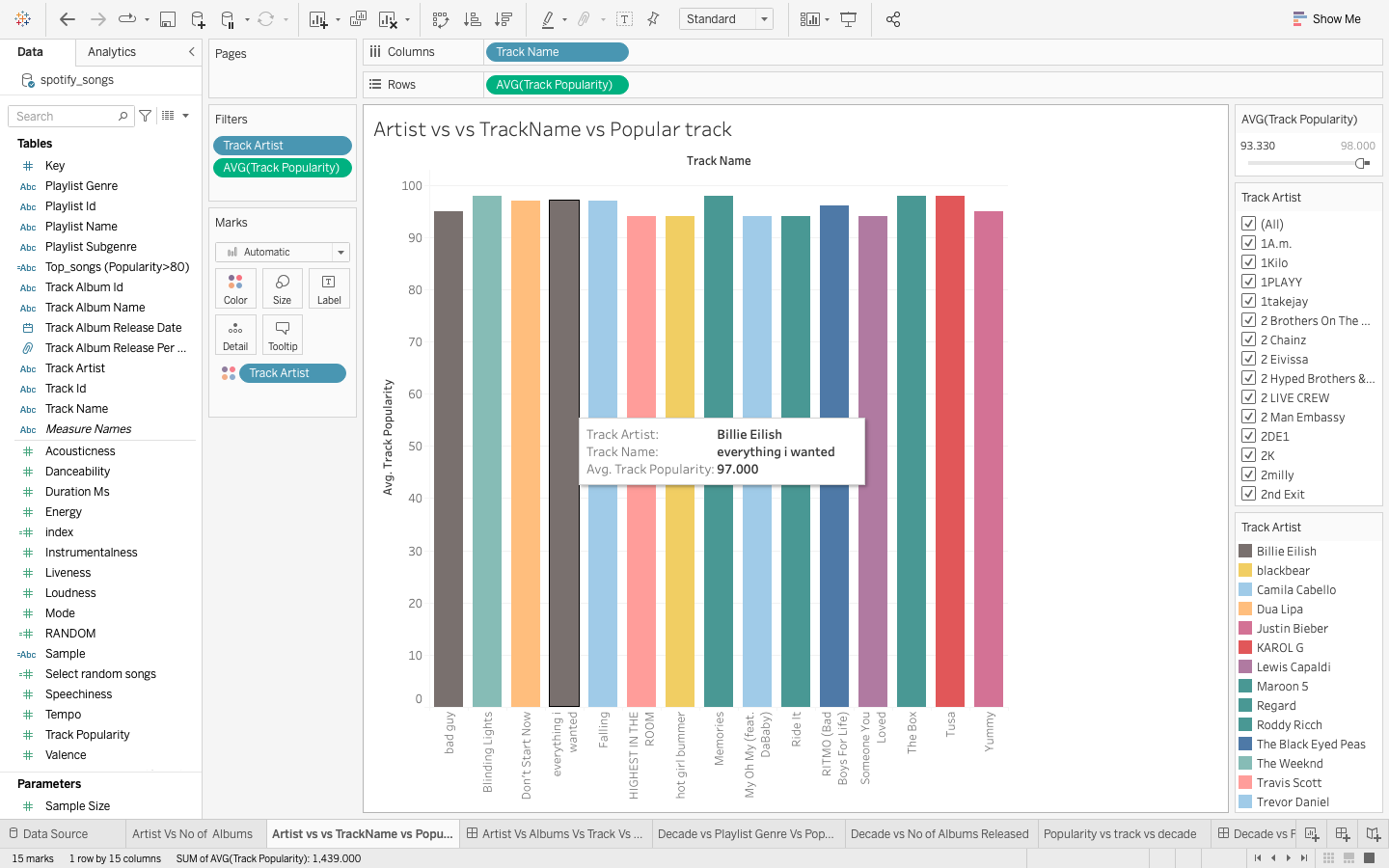


Figure : Findings for Artists and their most popular tracks

Here, we get a good understanding of the hits given by different artists.

### Playlist genre and their popularity over the decades

This observation falls in a time series and our aim here is to find how the genre preferences of the user have changed across the decades. Chart, treemap chart

Description automatically generated

Figure : Playlist genre and their popularity over the decades

This visualization can help the artist or the music companies to see the change in genre trends over the decades and by how much. We can see that how the popularity of genres has changed across the decades. It will be a much-informed choice for a company or an artist to decide which genre-based music they should as per the decade’s popularity trend.

### Albums released per decade

Here in this visualization, we are again using time series approach to mapping the distinct number albums released each decade from 1950 to 2020.

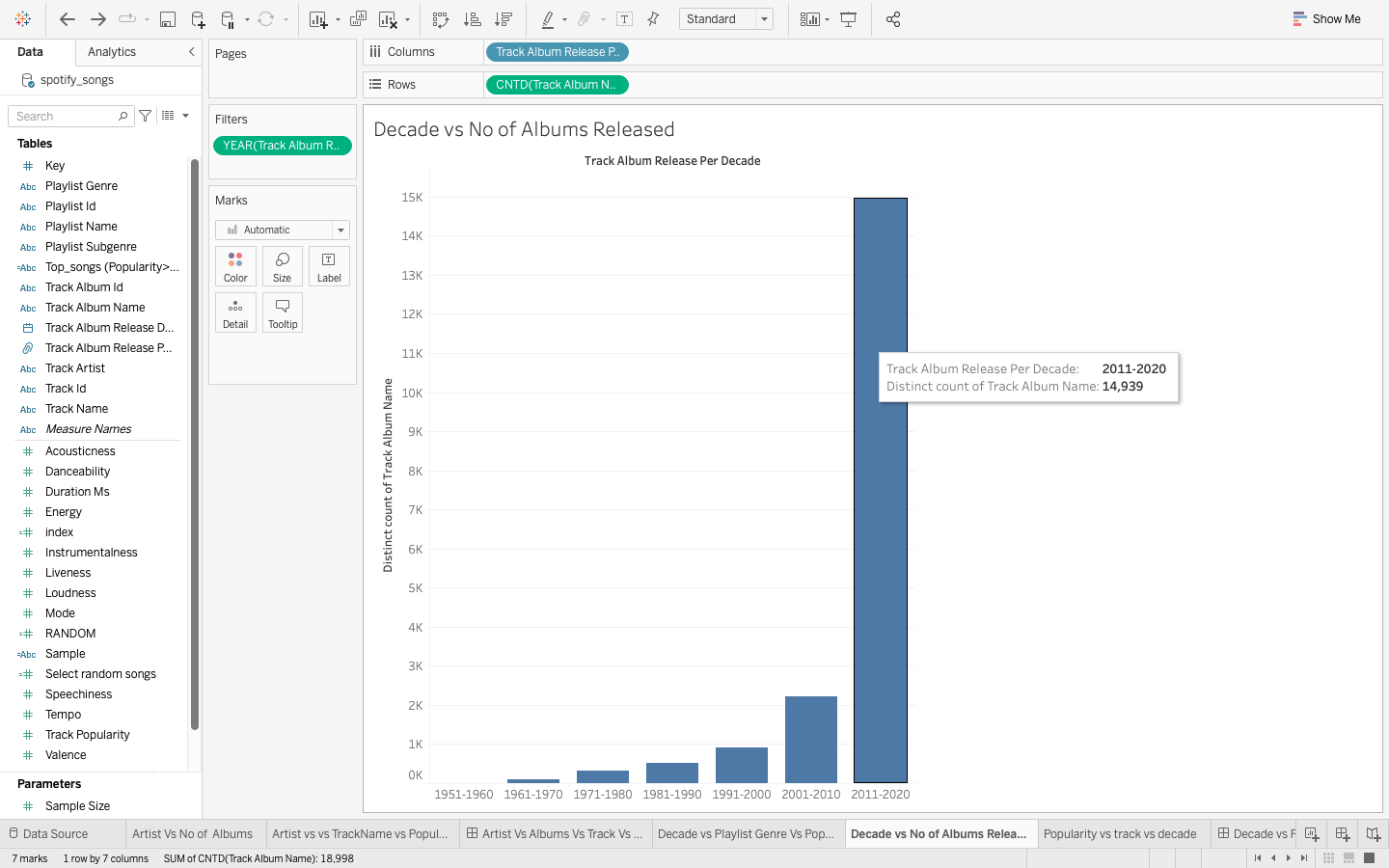


Figure : Albums released per decade

An interesting finding here is the rapid increase of around 400 % in the number of albums released from 2001-2010 to 2011-2020.The business opportunities looked bright for both the artist and music companies in this decade. On the contrary, we can see the number of increases in albums released in not exponential and is in fact a consistent proportional change.

### Average popularity of tracks released per decade

Here in this visualization, we are again using time series approach to identify the patterns in the popularity changes across different decades.

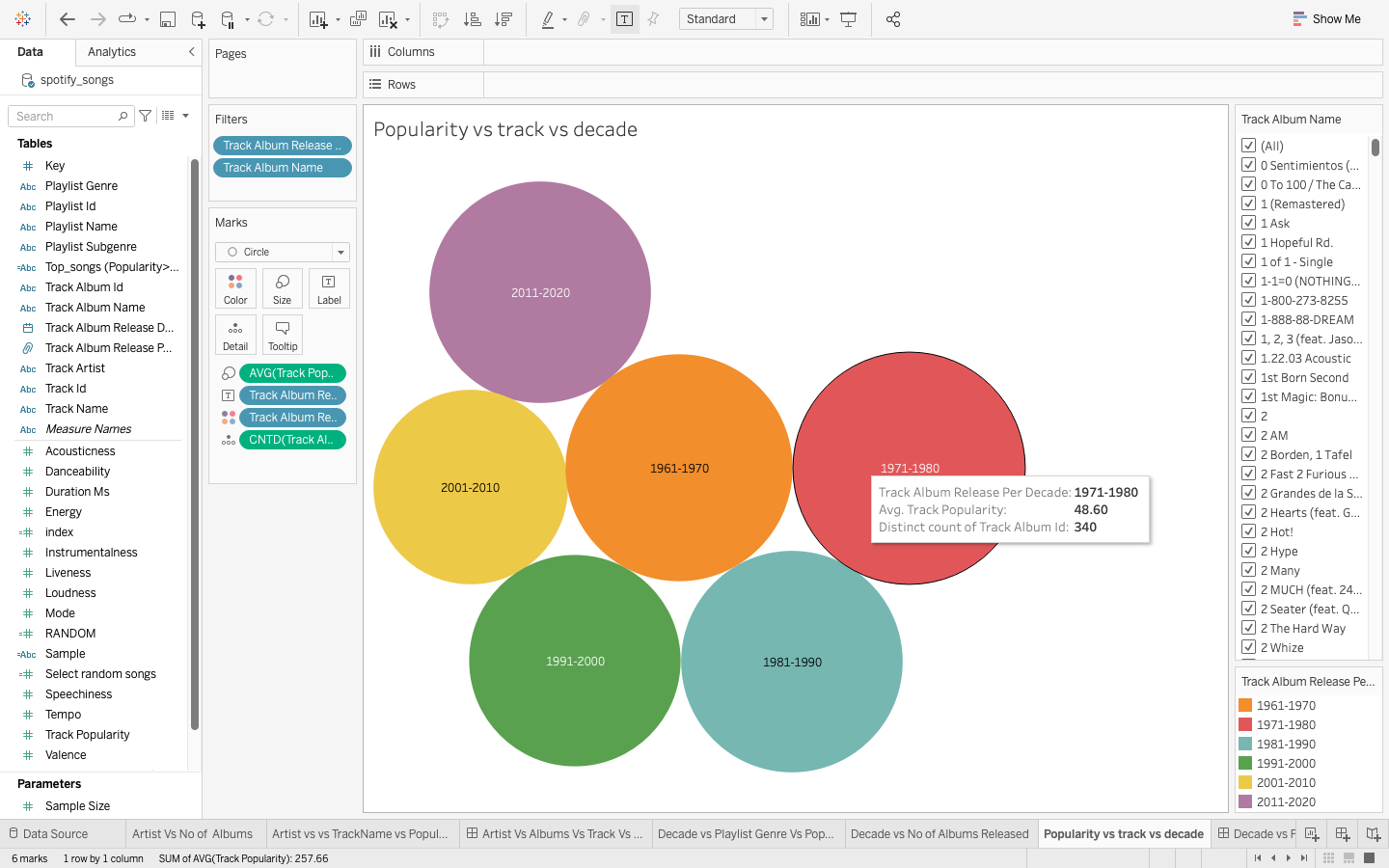


Figure 10: Average popularity of tracks released per decade

On an average, in decade 1971-1980 the tracks released the highest average popularity of 48.60 when rated on a scale of 0-100.This was closed followed by decade 1961-1970 at 46.39.

I have considered decade 1951-1960 an outlier here because of very few tracks released that decade s per our data set and that will result in a bias observation.

### Top popular tracks ever by Artist name and track name

Here in this visualization, I am trying to see what the top tracks of all times and those tracks were given by which artists.

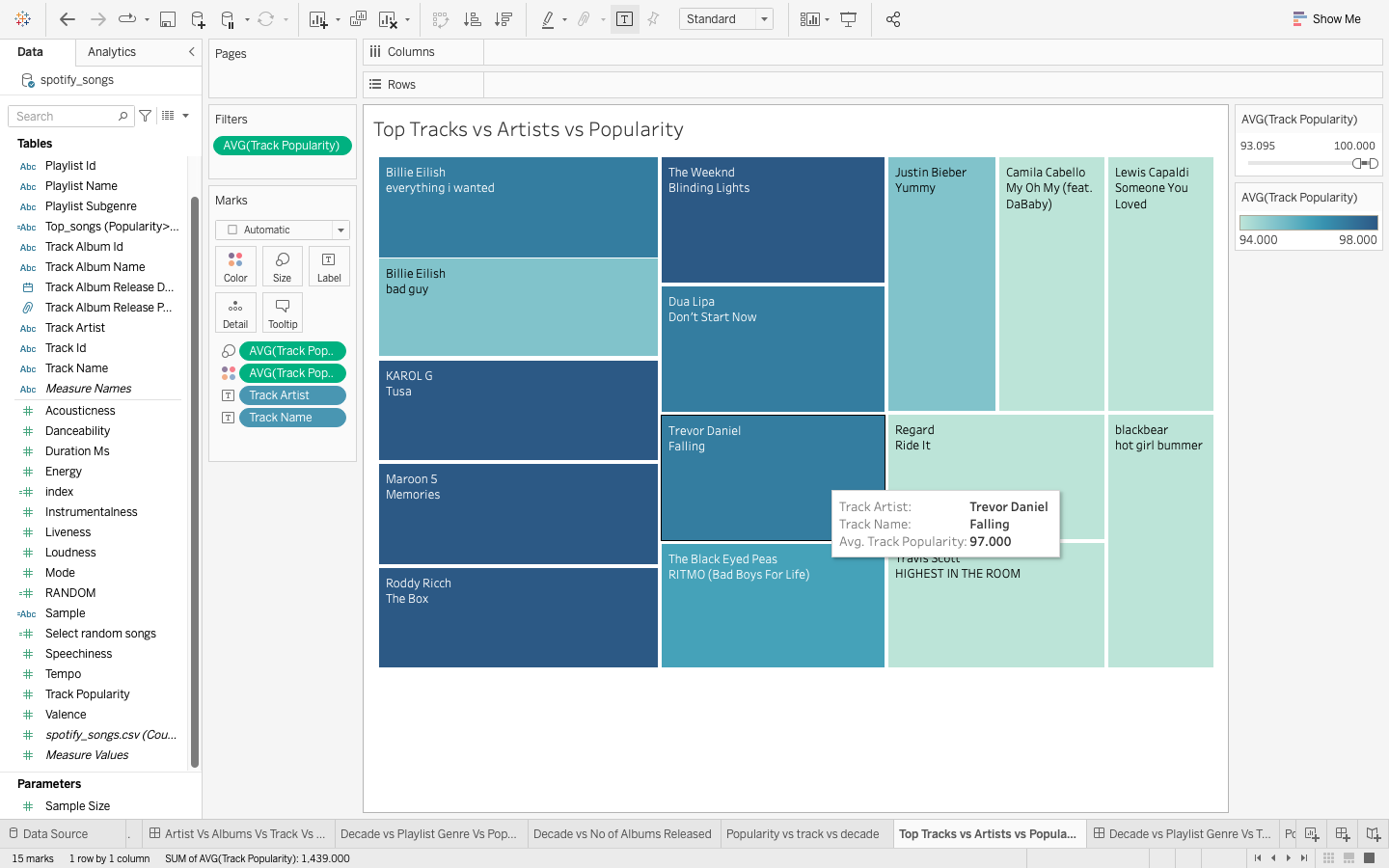


Figure 11: Top tracks ever and by whom

Top tracks have been grouped here with popularity greater that 94 on scale of 0-100.

### Correlation of Popularity with music characteristics

Here in this visualization, I am trying to find a correlation of track being popular with music characteristics like danceability, energy, liveliness, valence, and few others.

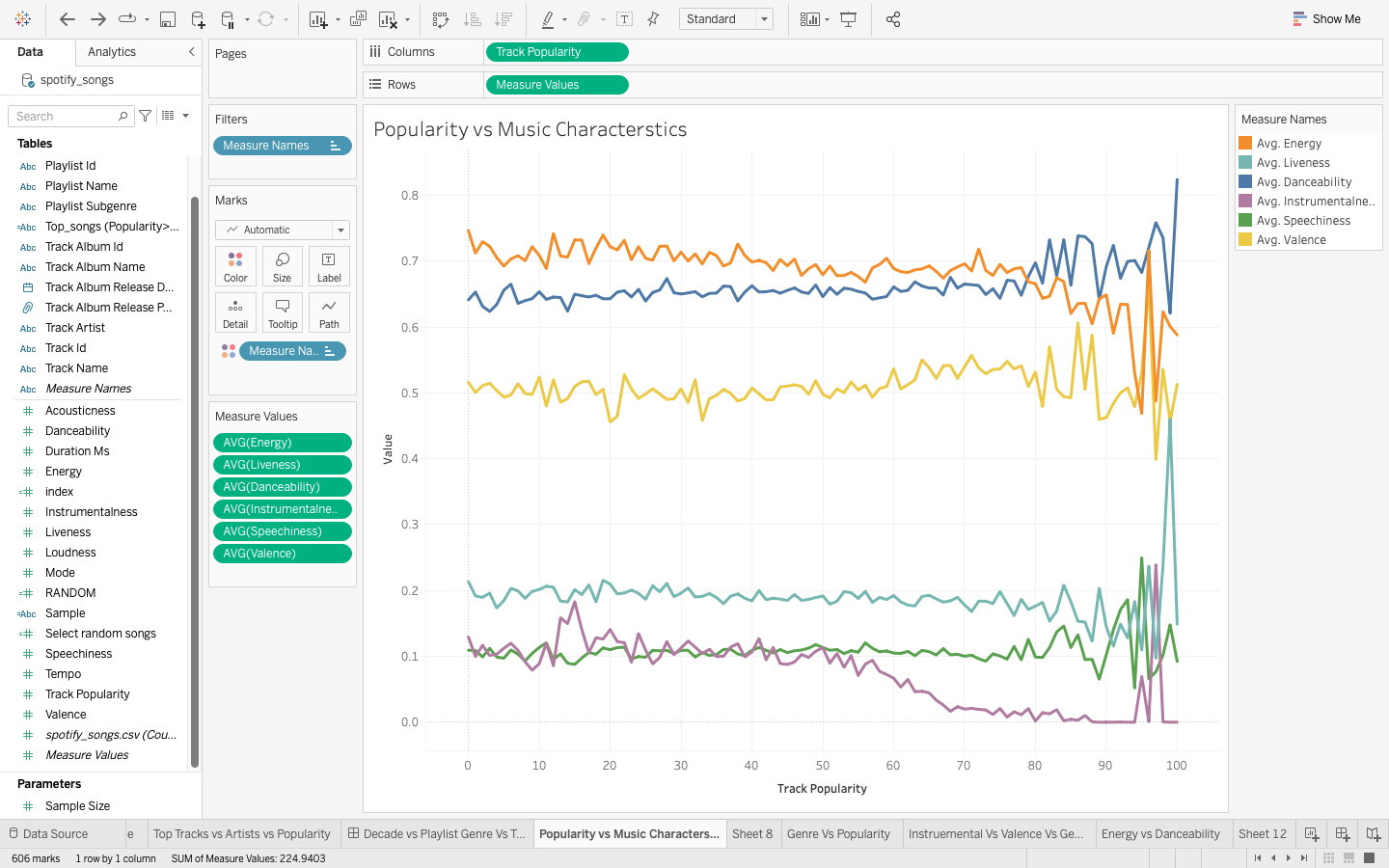


Figure 12: Correlation of popularity with music characteristics

An interesting finding here is that tracks with higher popularity rating of 60 usually has almost negligible instruementalness and higher values of liveness, danceability and energy. We can see that danceability is close to almost one for the highest popular songs so this is a very important in song being popular. The song’s danceability is directly proportional to popularity.

### Correlation between intruementalness, valence and genre

Here in this visualization, we are using correlation to analyze the relationship between instruementalness, valence and genre.

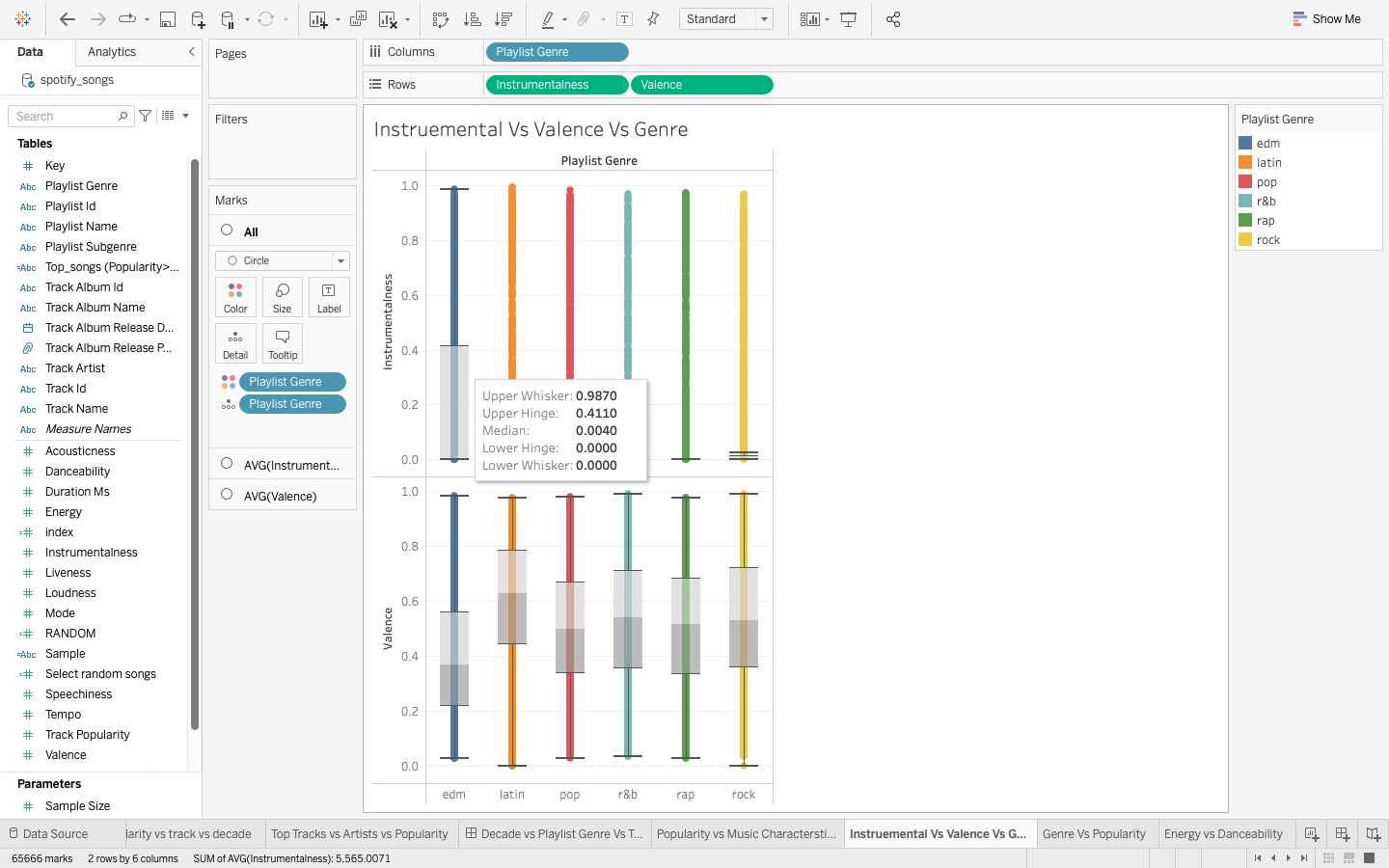


Figure 13: Correlation between instruementalness, valence and genre

Valence is the measure of positiveness of a track. Higher the valence, higher is the positiveness conveyed by the track. Intruementalness predict whether a track has vocals or not. Closer the value of instruementalness to 1, lower are the number of vocals in the track. Rap has the lowest upper whisker which means that out of all the genres, rap has the most vocals. For valence, rock has the highest whisker at 0.9910 hence higher music positiveness.

### Correlation between energy, danceability and genre

Here in this visualization, we are trying to analyze how danceability and energy is spread across different genres.

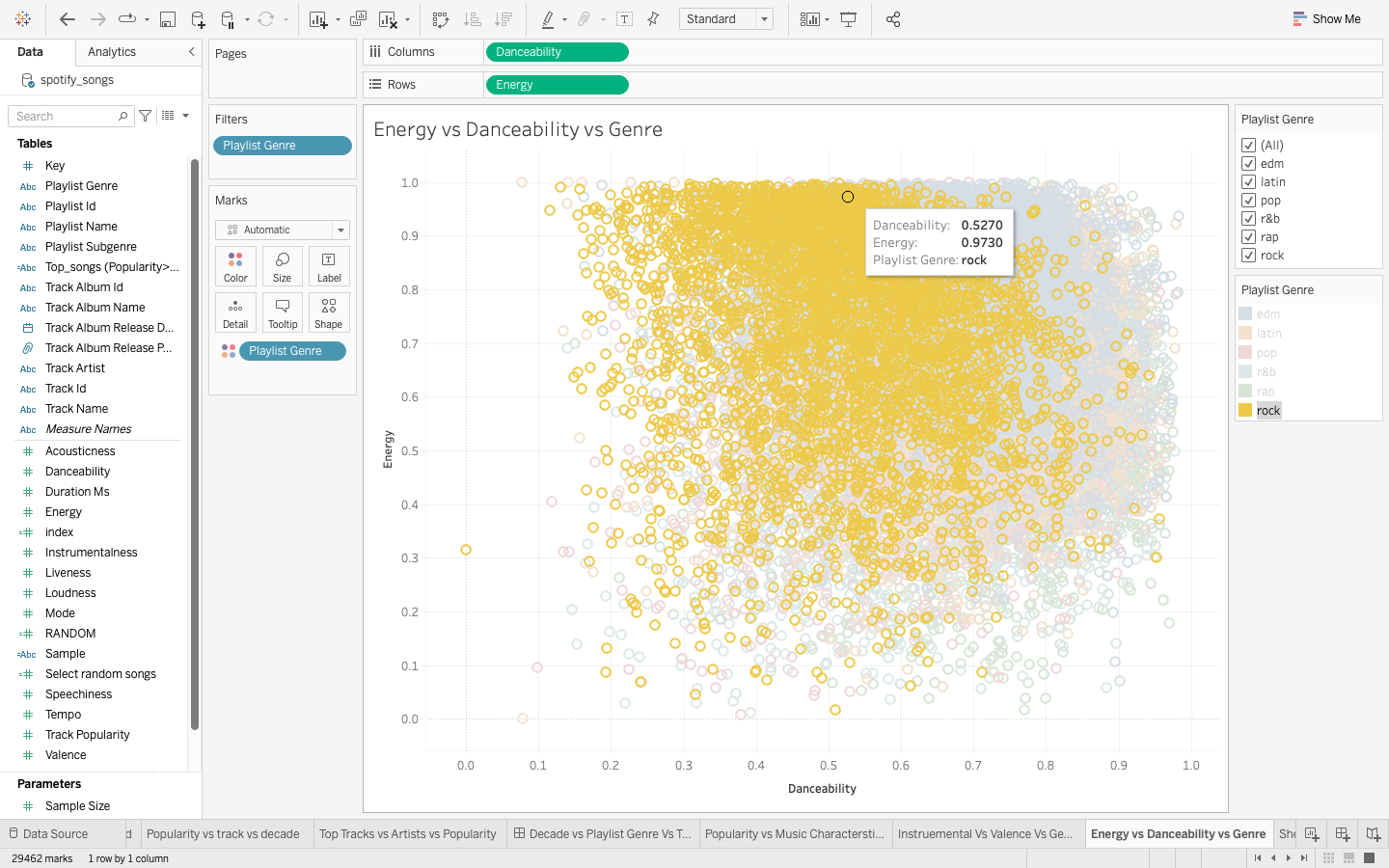


Figure 14: Correlation between energy, danceability and genre

An interesting finding here is that rock and pop music has generally high energy and lower danceability for most of the data spread whereas the correlation between energy and danceability is positive for edm and Latin. Rap mostly has high energy and danceability.

## Graphical representations

The final graphical representations for each of the four sections mentioned in the findings may be found here. We have created 4 dashboards to link the insights analyzed in the visualization done in the above section.

### Analysis of artists, albums, tracks, and their popularity

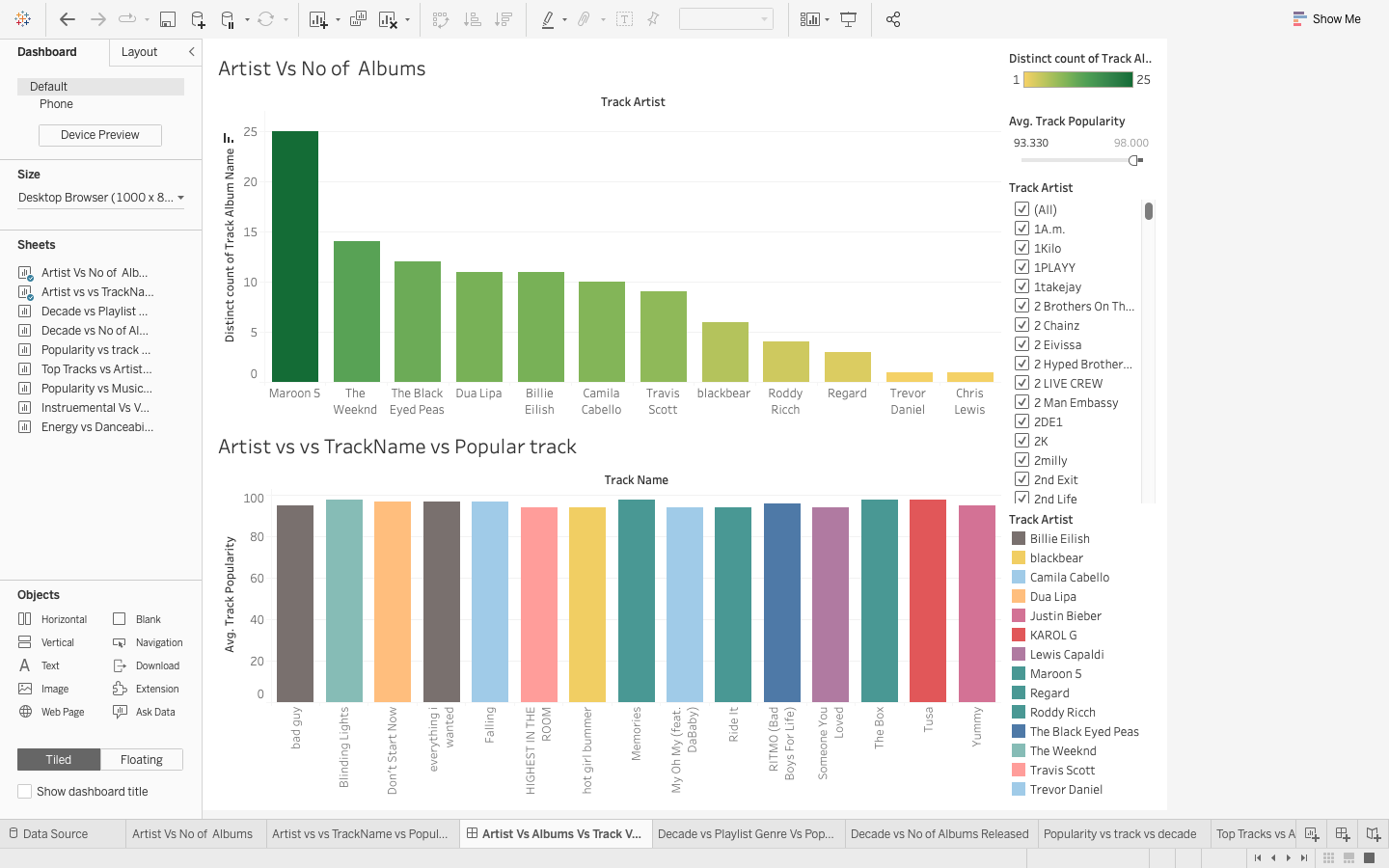


Figure 15: Analysis of artists, albums, tracks, and their popularity

This dashboard is intended to be interactive. This Dashboards helps one to choose an artist, the number of albums he has created and what is the popularity of track of an artist.

### Correlation between genre, track, playlist, artist, and popularity

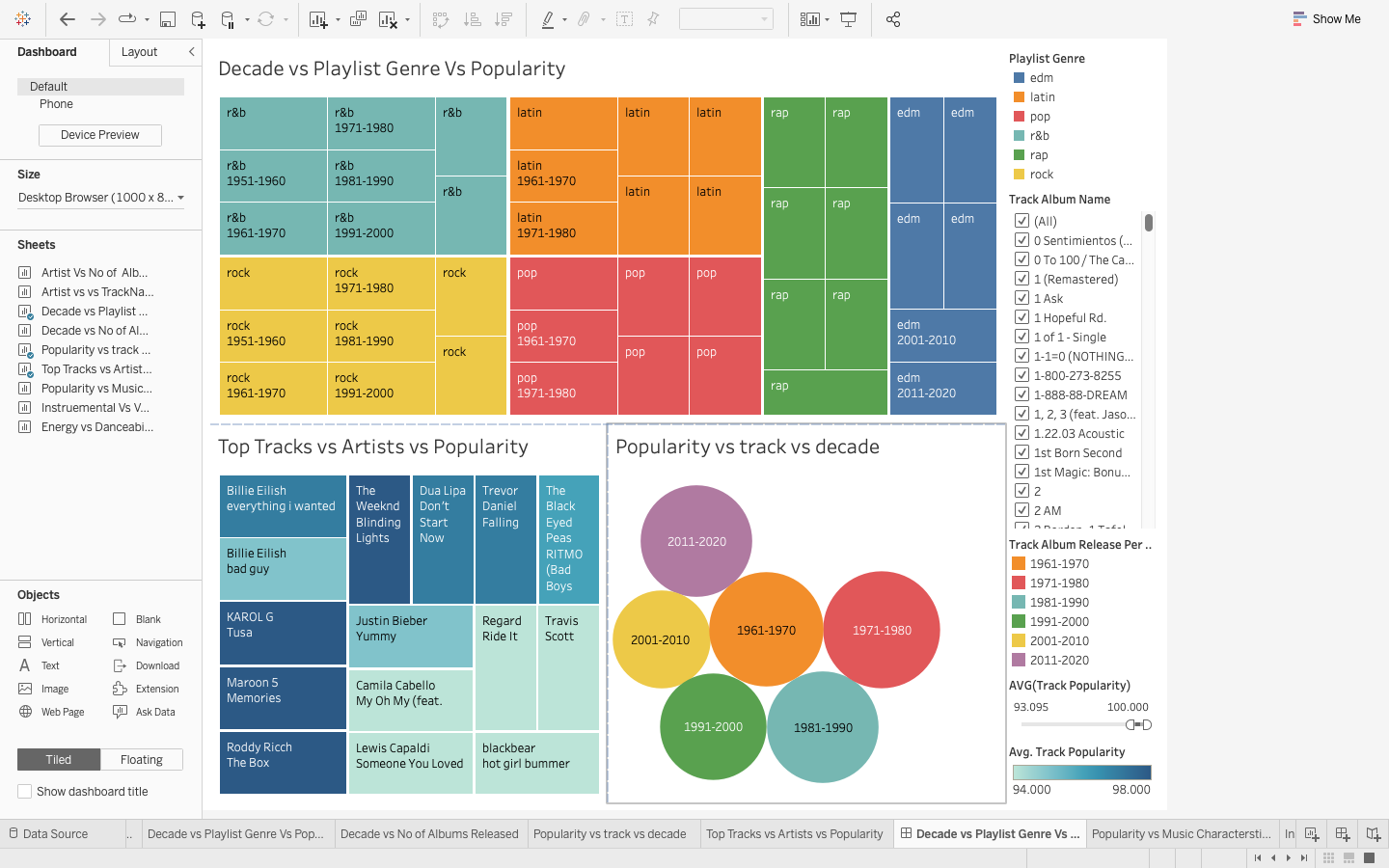


Figure16: Correlation between genre, track, playlist, and popularity

This dashboard is again designed for interaction. This one is intended to find any correlation between different genres such as pop, Latin and edm, their popularity across different decades and the popular tracks produced by artists in the respective genres. This will help music companies to evaluate with which artist they should collaborate while targeting their preferred genre.

### ﻿Correlation between Popularity vs genre vs valence and intruementalness

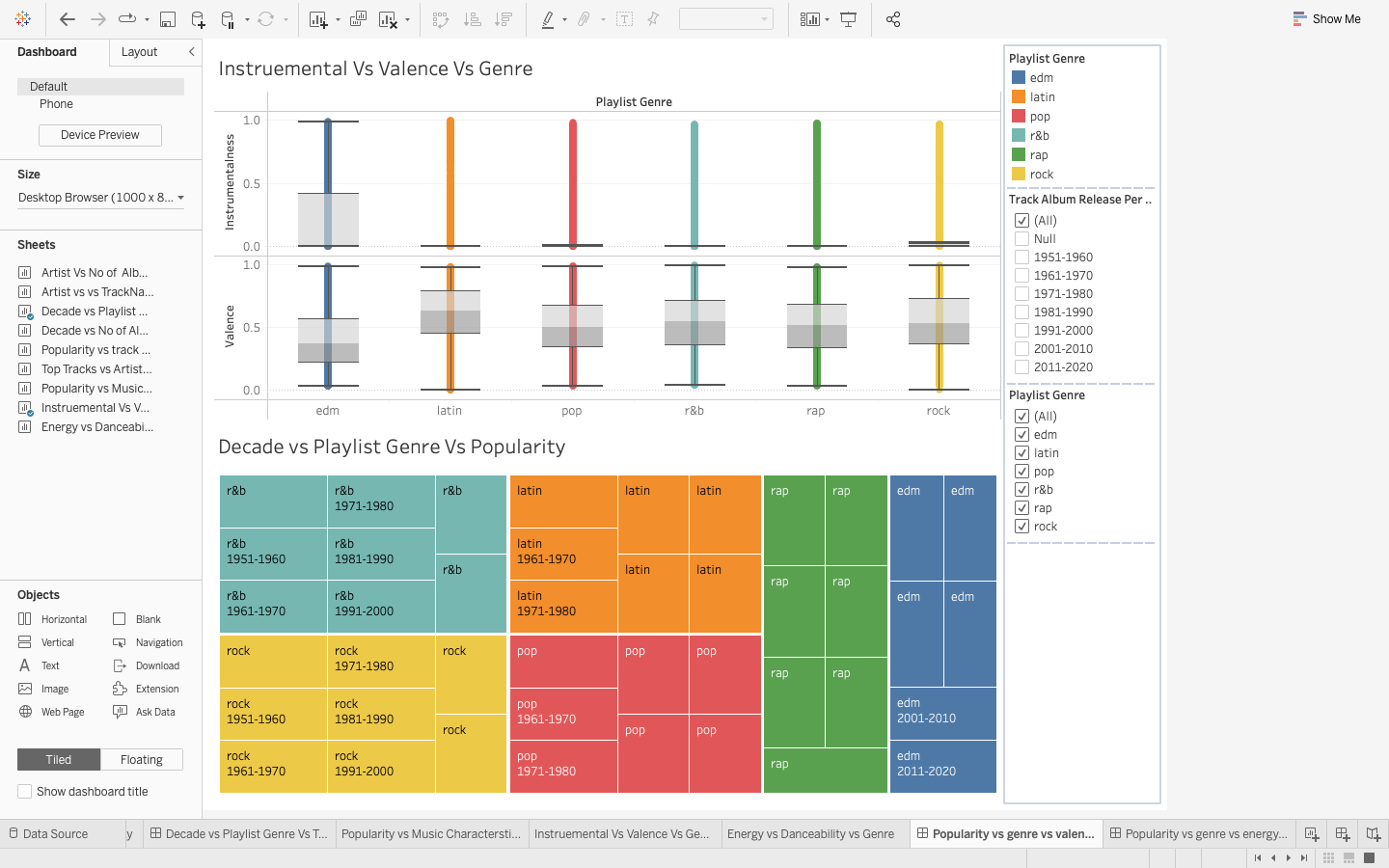


Figure 17: Correlation between popularity vs genre vs valence and intruementalness

This dashboard has also been designed to be interactive. This dashboard gives insights about correlation of music characteristics like valence and instruementalness across different genres and how it is related to their popularity. This gives us an idea about what parameters a musician must aim for in a particular genre if he wants to have a higher probability of the track being successful.

### Correlation between ﻿Popularity vs genre vs energy and danceability

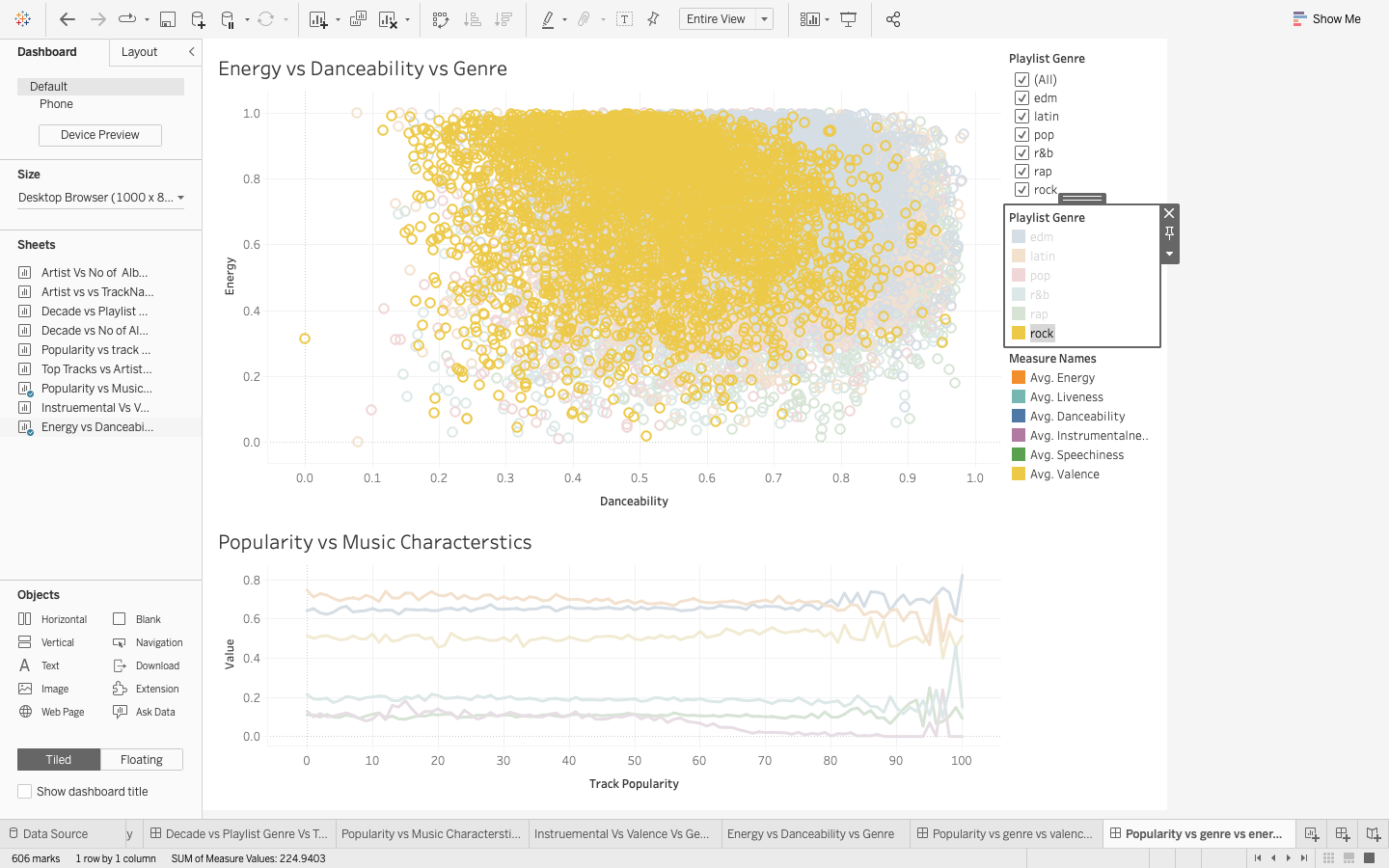


Figure 18: ﻿Correlation between Popularity vs genre vs energy and danceability

This dashboard has also been designed to be interactive. This dashboard gives insights about correlation of music characteristics like energy and danceability across different genres and how it is related to their popularity. This gives us an idea about what parameters a musician must aim for in a particular genre if he wants to have a higher probability of the track being successful.

## Tabular representations

This analysis does not make use of Tabular representations.

## insight and analysis

This Spotify data can be leveraged by music companies and artists to understand users' preferences for different genres of music. Based on user behavior on the platform, music companies can identify how playlist trends have changed over the decades, allowing them to collaborate with artists and provide music. Music will resonate more with users. They will also be able to gain other insights into musical characteristics such as playability, danceability, energy and volume that contribute to the popularity of the piece and how it correlates with each other and on different genres. This information will provide music companies and artists with valuable insights into their customers' reactions to music and thereby help them more effectively monetize their music in the future.

For the analysis, we have leveraged the **Time series data** that is recorded over consistent intervals of time and the **Cross-sectional data** consisting of several music characteristics recorded at the same time.

﻿ The bar graph in Figure 6 shows the number of individual albums released by featured artists in descending order. Maroon 5 can be seen with the 25 most released albums of all artists, followed by The Weekend with 14 releases, almost half the number of album releases.

This image shown in Figure 7 helps us find which track is most popular with an artist based on the artist filter applied to the data. The chart also provides information about an example: Falling is Trevor Daniel's most popular track with a popularity rating of 97.00.

Our goal here is to determine how users' genre preferences have changed over the decades. This is done using Dashboard where we observe correlation between genres, artists and track popularity which can help artists or music companies see how genre trends have changed over the decades and to what extent. We can see that the popularity of genres has changed over the decades. It would be a wise choice for a company or an artist to decide which genre-based music they should use according to the decade's popular trends. Using the time series approach to map the separate release albums released in each decade from 1950 to 2020. An interesting finding here is that the number of albums released from 20012010 to 20112020 has increased rapidly by about 400%. Business opportunities look promising for artists and music companies this decade. On the contrary, we can see that the number of albums released is not exponential and is in fact a proportional change. over other different decades. On average, during the decade 1971-1980, the releases with the highest popularity averaged .60 when rated on a scale of 0-100. This ends, followed by the decade 1961-1970 at 6.39. by our data set and this will lead to observation bias.

For dashboard three, there are some interesting insights about the correlation between a song's popularity and musical characteristics such as choreography, energy, liveliness, melody, and some other characteristics. difference. An interesting finding here is that tracks with popularity ratings above 60 often have near-insignificant instrumental properties and higher liveliness, choreography, and energy values. We can see that the choreography is almost the same for most popular songs, so it is a very important factor in the popularity of a song. The song's choreography is commensurate with its popularity.

For dashboard four, the correlation helps us find the relationship between instrumentalness, values, and genre. The higher the valence, the higher the positivity of the trail. The instrumentalness predicts whether a track has vocals or not. An interesting finding here is that rock and pop music are generally more energetic and less danceable while the correlation between energy and danceability is positive. On top of that, the rap part has a lot of energy.

All these insights provide useful information about our use case and arrive at good probable choice of collaboration of an artist with other artists or music companies and genre they can create for their next musical project to be a success.

# Summary

## Summarizing the problem statement

It's a daunting task for the music companies and the artists to go to the market and to identify other artists and subjects with which they should work to be more successful. Their success can be defined as a measure for track being popular and reaching more diversified audiences, resulting in a larger fan base, and creating higher profits.

## adressal of the problem statement

Musicians all over the world are always looking to collaborate while making music and it’s a very common practice among music companies to create music which has multiple artists. This is a common strategy used by musicians and record labels to grow their audiences and, therefore**,** their revenue. Spotify data is agreat resource for determining the righttype of collaborator based on a variety of qualities and characteristics**.** This project aims to create dashboards that provide insightinto the music industry and its general indicators**,** allowing for more strategic collaboration between artist and music companies and analyze music trends to reach their most preferred audiences.

## summary of final insight & analysis and implications

***In short, the complete analysis focuses on analyzing historical trends in the music industry across various genres and studying different music characteristics and their relationship with popularity. The better popularity will help artists and music companies make better music in the future.***

To reach a bigger audience and promote its unique styles and cultures, the music industry frequently engages in artistic and creative partnerships with one another. This is a typical strategy employed by musicians and record labels to grow their following and thus their revenue. We want to use this as a valuable tool for discovering the right kind of partners based on a range of characteristics and characteristics. This project aims to produce dashboards that give a broad picture of the music industry and its main parameters, allowing artists and music companies to collaborate more strategically.

## Limitations & recommendationS

### Limitations

The dataset and the subsequent analysis have the following limitations:

1. The dataset doesn’t have user’s demographic data like age and taste of music.
2. The dataset also doesn’t have details like how many times a track was liked or played and how it impacts the popularity of a song.
3. The dataset can have information like money earned by an artist based on the popularity so monetizing factor can be included in the analysis.

### Recommendations

The following are the recommendations for the dataset and the subsequent analysis:

1. Do some analysis about how duration of a song and the number of times it is played impacts a popularity.
2. Also, we can try to find a correlation between how may time a track is played vs how many times it was completely played and how it is related to the popularity of the song.
3. This dataset needs to be combined with official track performance data to understand the full music industry scenario under consideration. In addition, radio performance and other song mentions should also be included in the analysis.

#### *References:*

1. Spotify Songs Data Source:

<https://github.com/rfordatascience/tidytuesday/blob/master/data/2020/2020-01-21/readme.md>

1. Exploratory Data Analysis reference:

<https://pasaentuciudad.com.mx/visualizing-spotify-songs-with-python-an-exploratory-data-analysis/>

1. Box and whisker charts reference:

<https://www.tableau.com/data-insights/reference-library/visual-analytics/charts/box-whisker>