ANÁLISIS Y VISUALIZACIÓN DE TENDENCIAS MUSICALES EN SPOTIFY

REALIZADO POR:
MEDINA SANCHEZ SUGEY
HERNANDEZ MUÑOZ KAROL

INTRODUCCION A LA CIENCIA DE DATOS

DOCENTE:
JAIME ALEJANDRO ROMERO SIERRA

INTRODUCCION

Evaluar como los algoritmos de recomendación en las listas de reproducción y las interacciones de los usuarios en Spotify afectan la popularidad mediante lo explicito, la duración o listas de reproducción de las canciones, esto para desarrollar estrategias comerciales que maximicen el crecimiento de los artistas y optimicen el retorno de inversión en el marketing musical.

Esto permitirá analizar no solo las reproducciones, sino también las implicaciones económicas y las oportunidades para monetizar esa popularidad.

JUSTIFICACION

1. Mejorar la experiencia del usuario
2. Optimización de algoritmos de recomendación
3. Monetización y estrategias de negocio
4. Desarrollo de nuevos proyectos o características
5. Estudio de tendencias globales y locales
6. Investigación de impactos sociales y culturales

En resumen, resolver o estudiar problemas en entorno a un dataset de Spotify tiene localizaciones directas en las experencias del usuario, la optimización de servicios, el crecimiento de la plataforma y la compresión de fenómenos culturales y comerciales.Ademas, puede ofrecer a las empresas

Fuentes de datos:

	Artista	Url_spotify	Pista	Album	Url	Danzabilidad	Energia	Tonalidad	Volumen	Habla	...	Sensacion_Emocional	Pulsos	Duracion_ms	Url_youtube	Titulo	Canal	Vistas	Comentarios	Descripcion	Reproducciones
0	Gorillaz	https://open.spotify.com/artist/3AA28KZvwAUcZu...	Feel Good Inc.	Demon Days	spotify:track:0d28khcov6AiegSCpG5TuT	0.818	0.705	6.0	-6.679	0.1770	...	0.772	138.559	222640.0	https://www.youtube.com/watch?v=HyHNuVaZJ-k	Gorillaz - Feel Good Inc. (Official Video)	Gorillaz	693555221.0	169907.0	Official HD Video for Gorillaz' fantastic trac...	1.040235e+09
1	Gorillaz	https://open.spotify.com/artist/3AA28KZvwAUcZu...	Rhinestone Eyes	Plastic Beach	spotify:track:1foMv2HQwfQ2vntFf9HFeG	0.676	0.703	8.0	-5.815	0.0302	...	0.852	92.761	200173.0	https://www.youtube.com/watch?v=yYDmaexVHic	Gorillaz - Rhinestone Eyes [Storyboard Film] (...	Gorillaz	72011645.0	31003.0	The official video for Gorillaz - Rhinestone E...	3.100837e+08
2	Gorillaz	https://open.spotify.com/artist/3AA28KZvwAUcZu...	New Gold (feat. Tame Impala and Bootie Brown)	New Gold (feat. Tame Impala and Bootie Brown)	spotify:track:64dLd6rVqDLtkXFYrEUHIU	0.695	0.923	1.0	-3.930	0.0522	...	0.551	108.014	215150.0	https://www.youtube.com/watch?v=qJa-VFwPpYA	Gorillaz - New Gold ft. Tame Impala & Bootie B...	Gorillaz	8435055.0	7399.0	Gorillaz - New Gold ft. Tame Impala & Bootie B...	6.306347e+07
3	Gorillaz	https://open.spotify.com/artist/3AA28KZvwAUcZu...	On Melancholy Hill	Plastic Beach	spotify:track:0q6LuUqGLUiCPP1cbdwFs3	0.689	0.739	2.0	-5.810	0.0260	...	0.578	120.423	233867.0	https://www.youtube.com/watch?v=04mfKJWDSzI	Gorillaz - On Melancholy Hill (Official Video)	Gorillaz	211754952.0	55229.0	Follow Gorillaz online:\nhttp://gorillaz.com \...	4.346636e+08
4	Gorillaz	https://open.spotify.com/artist/3AA28KZvwAUcZu...	Clint Eastwood	Gorillaz	spotify:track:7yMiX7n9SBvadzox8T5jzT	0.663	0.694	10.0	-8.627	0.1710	...	0.525	167.953	340920.0	https://www.youtube.com/watch?v=1V_xRb0x9aw	Gorillaz - Clint Eastwood (Official Video)	Gorillaz	618480958.0	155930.0	The official music video for Gorillaz - Clint ...	6.172597e+08

5 rows × 23 columns

Esta es nuestra base de datos que fue sacada de HitgGut, con esas variables.

METODOLOGIA

- **Proceso de limpieza de datos**:

```
#Cantidad de valores nulos por columna
df.isnull().sum()
```
✓ 0.0s

```
Artista                0
Url_spotify            0
Pista                  0
Album                  0
Url                    0
Danzabilidad           2
Energia                2
Tonalidad              2
Volumen                2
Habla                  2
Acustica               2
Instrumentalidad       2
Vivacidad              2
Sensacion_Emocional    2
Pulsos                 2
Duracion_ms            2
Url_youtube          470
Titulo               470
Canal                470
Vistas               470
Comentarios          569
Descripcion          876
Reproducciones       576
dtype: int64
```

```
#valores que no se puedan convertir a numéricos serán reemplazados por NaN
df['Canal'] = pd.to_numeric(df['Canal'], errors='coerce')
df['Vistas'] = pd.to_numeric(df['Vistas'], errors='coerce')
df['Comentarios'] = pd.to_numeric(df['Comentarios'], errors='coerce')
df['Reproducciones'] = pd.to_numeric(df['Reproducciones'], errors='coerce')
df['Tonalidad'] = pd.to_numeric(df['Tonalidad'], errors='coerce')
df['Danzabilidad'] = pd.to_numeric(df['Danzabilidad'], errors='coerce')
df['Instrumentalidad'] = pd.to_numeric(df['Instrumentalidad'], errors='coerce')
```
✓ 0.0s

```
#Reemplaza todos los valores NaN en las columnas por el promedio
df['Canal'].fillna(df['Canal'].mean(), inplace=True)
df['Vistas'].fillna(df['Vistas'].mean(), inplace=True)
df['Comentarios'].fillna(df['Comentarios'].mean(), inplace=True)
df['Reproducciones'].fillna(df['Reproducciones'].mean(), inplace=True)
df['Tonalidad'].fillna(df['Tonalidad'].mean(), inplace=True)
df['Danzabilidad'].fillna(df['Danzabilidad'].mean(), inplace=True)
df['Instrumentalidad'].fillna(df['Instrumentalidad'].mean(), inplace=True)
df.head(5)
```
✓ 0.0s

```
#elimina todas las filas de df que contengan al menos un valor NaN
df2=df.dropna()
df2.head()
```
✓ 0.0s

```
#Cantidad de valores nulos por columna de df2
df2.isnull().sum()
```
✓ 0.0s

```
Artista                0
Url_spotify            0
Pista                  0
Album                  0
Url                    0
Danzabilidad           0
Energia                0
Tonalidad              0
Volumen                0
Habla                  0
Acustica               0
Instrumentalidad       0
Vivacidad              0
Sensacion_Emocional    0
Pulsos                 0
Duracion_ms            0
Url_youtube            0
Titulo                 0
Canal                  0
Vistas                 0
Comentarios            0
Descripcion            0
Reproducciones         0
dtype: int64
```

Hicimos un proceso de limpieza de datos, con el proceso de promedio por sustitución de nan por promedio en unas columnas en especifico y luego elinamos algunas filas que tenían valores nulos (nan) y finalmente comprobamos que ya no existían valores nulos, cabe recalcacr que dedsde el inicio no hubo valores duploicados.

Resumen Estadístico

Nuestro resumen oficial se puede visualizar mediante los graficos que realizamos como; histogramas, boxplots , gtraficos de pastel y graficos debarras.

Uno de los principales como:

1.Las reproducciones

Ocparemos el promedio, mediana, máximo, minuimo y desviación estándar.

Su distribución seria ver como se distribuyen las reproducciones en diferentes rangos.

2.Descripción

El numero de registros con descripción, longitud promedio de las descripciones, análisis de palabras clave.

3.Comentarios

El promedio de los comentarios por pista, máximo y mínimo, analisi de sentimiento (positivo/negativo)

4.Me gusta

Promedio, máximo, minimo y desviación estándar de la cantidad de “me gusta” por pista.

5.Cana

Frecuencia de los canales mas populares, distribución de la pista entre canales.

6.Titulo

Numero de títulos únicos, longitud promedio de los títulos

7.Url_Youtube

E l numero de pistas con url_youtube, porcentaje de canciones con y sin URL

8.Duración ms

Promedio, mediana, máximo, mínimo y desviación estándar de la duración de las pistas en milisegundos

La conversión seria de convertir de milisegundos a minutos y segundos para una mejor compresión.

9.Sensación emocional

Analisis de las emociones predominantes (alegría, tristeza,ira,etc)

Poercentaje de canciones en categoría emocional.

10.Tonalidad

Distribución de las tonalidades (mayor, menor) de las pistas

Visualización y Distribución de Variables Individuales

Variables Numéricas:

Pera object: Son columnas con datos no numéricos (textos, URLS, descripciones).

int64: Son columnas con números enteros (por ejemplo, tipo de álbum o licencias).

float64: Son columnas con números decimales (por ejemplo, puntajes musicales o métricas de interacción como vistas y "me gusta").

Columna	Tipo de dato	Descripción
Artista	`object`	Cadena de texto (probablemente el nombre del artista).
Url_spotify	`object`	Cadena de texto (URL a la página de Spotify).
Pista	`object`	Cadena de texto (nombre o identificador de la pista).
Album	`object`	Cadena de texto (nombre del álbum al que pertenece la pista).
Tipo_de_album	`int64`	Número entero (probablemente indica el tipo de álbum, por ejemplo, álbum de estudio o sencillo).
Url	`object`	Cadena de texto (URL de la pista o del álbum).
Danzabilidad	`float64`	Número decimal (probablemente una puntuación de 0 a 1 que indica cuán bailable es la pista).
Energia	`float64`	Número decimal (probablemente una puntuación de 0 a 1 que mide la energía de la pista).
Tonalidad	`float64`	Número decimal (indica la tonalidad de la pista, quizás en una escala de 0 a 1 o una medida numérica).
Volumen	`float64`	Número decimal (probablemente el volumen percibido de la pista).
Habla	`float64`	Número decimal (indica el nivel de habla en la pista, puede ser un índice de cuán hablada es la canción).
Acustica	`float64`	Número decimal (puede indicar cuán acústica es la pista).
Instrumentalidad	`float64`	Número ↓ mal (indica cuán instrumental es la pista).
Vivacidad	`float64`	Número decimal (probablemente mide cuán vibrante o animada es la pista).
Sensacion_Emocional	`float64`	Número decimal (una puntuación emocional de la canción, indicando cuán emocional es la pista).
Pulsos	`float64`	Número decimal (probablemente indica el ritmo o tempo de la pista).
Duracion_ms	`float64`	Número decimal (duración de la pista en milisegundos).
Url_youtube	`object`	Cadena de texto (URL de la pista en YouTube).
Titulo	`object`	Cadena de texto (título de la canción).
Canal	`float64`	Número decimal (probablemente relacionado con el canal de distribución o el número de visitas).
Vistas	`float64`	Número decimal (número de vistas que ha recibido la canción o video).
Descripcion	`object`	Cadena de texto (descripción de la pista, álbum o video).
Licencias	`int64`	Número entero (probablemente el número de licencias o tipo de licencias asociadas).

Variables Categóricas:

Estas variables contienen texto y se utilizan para representar categorías como los nombres de artistas, canciones o álbumes. No tienen un valor numérico o una escala continua asociada, por lo que se consideran categóricas de tipo nominal.

*Artista (object): Es una variable categórica, ya que contiene el nombre del artista, que es una categoría que no tiene un orden específico.

Url_spotify (object): Aunque es una cadena de texto, no es estrictamente una variable categórica en el sentido de clasificación o agrupación. Es más bien una URL, por lo que no la consideramos como categórica.

*Pista (object): Es el nombre de la pista, que también es una variable categórica nominal, ya que es un identificador o nombre de la canción.

*Album (object): Es una variable categórica nominal, ya que representa el nombre del álbum al que pertenece la pista, y los nombres de los álbumes no tienen un orden intrínseco.

*Url (object): Similar a Url_spotify, no es categórica en el sentido de clasificación.

*Titulo (object): El nombre de la canción. Esta es una variable categórica nominal.

*Url_youtube (object): Similar a las URLs anteriores, no es una variable categórica en el sentido clásico.

Resumen de las variables categóricas:

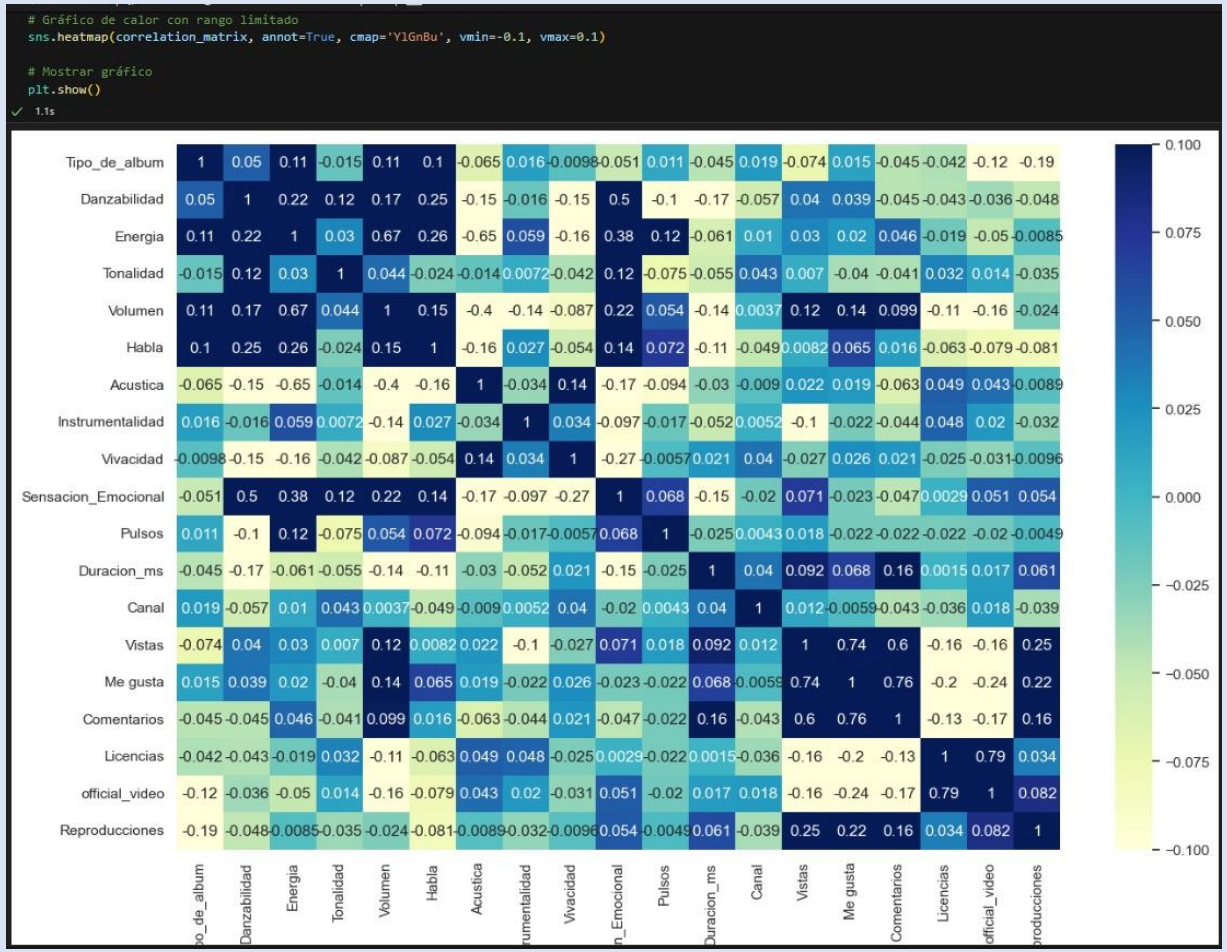
Artista: Nominal (categoría sin orden).

Pista: Nominal (nombre de la canción, sin orden).

Album: Nominal (nombre del álbum, sin orden).

Titulo: Nominal (nombre de la pista, sin orden).

Correlación entre Variables



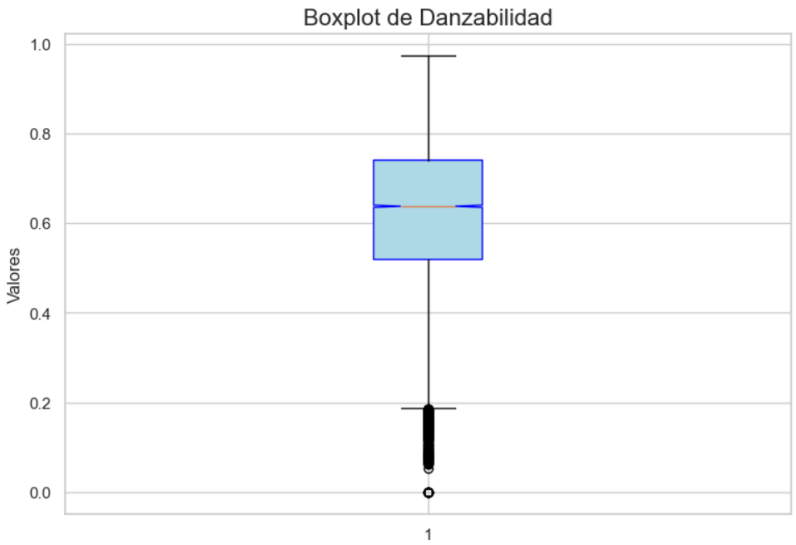
Como podemos observar en nuestro mapa de calor no hay correlación alguna entre las variables, sin embargo tenemos algunas que destacan entre lo inusual, como lo son:

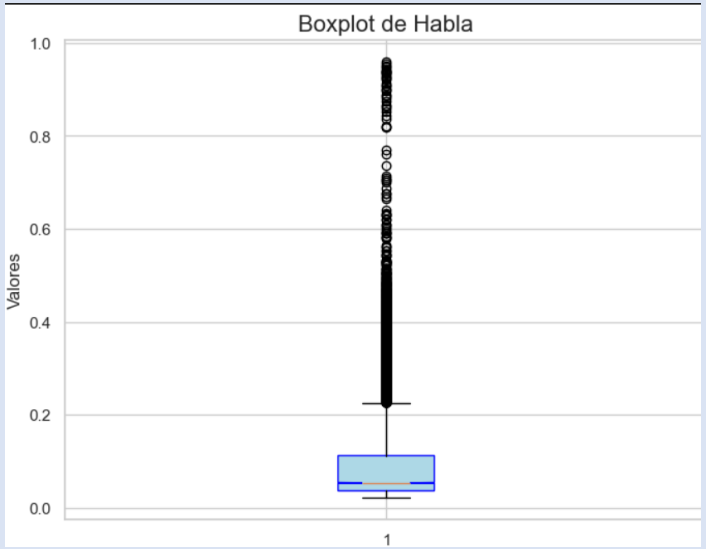
Parejas de Variables:

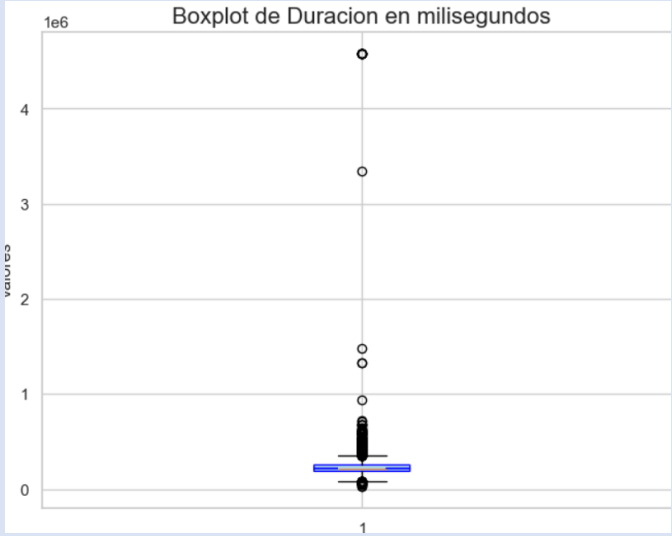
| Me gusta con comentarios con una correlación de 0.77
| Vistas con me gusta con una correlación de 0.73
| Licencias con video oficial con una correlación de 0.77
| Comentarios con vista con una correlación de 0.66
| Volumen con energía con una correlación de 0.65

Análisis de Valores Atípicos (Outliers)

- Para generar valores atípicos, se generaron boxplots para cada variable numérica. A través de esta visualización, se detectaron valores atípicos en las columnas: danzabilidad, tonalidad, habla, sensación emocional, vistas, duración en ms, comentarios y reproducciones.
- Este método consistió en calcular la diferencia entre los cuartiles y definir en rangos validos los valores.
- Los valores fuera de este rango fueron considerados atípicos y manejados de acuerdo con los objetivos del analisis., lo que permitió garantizar la calidad de los datos para nuestro modelo predictivo y reducir todas las posibles distorsiones.





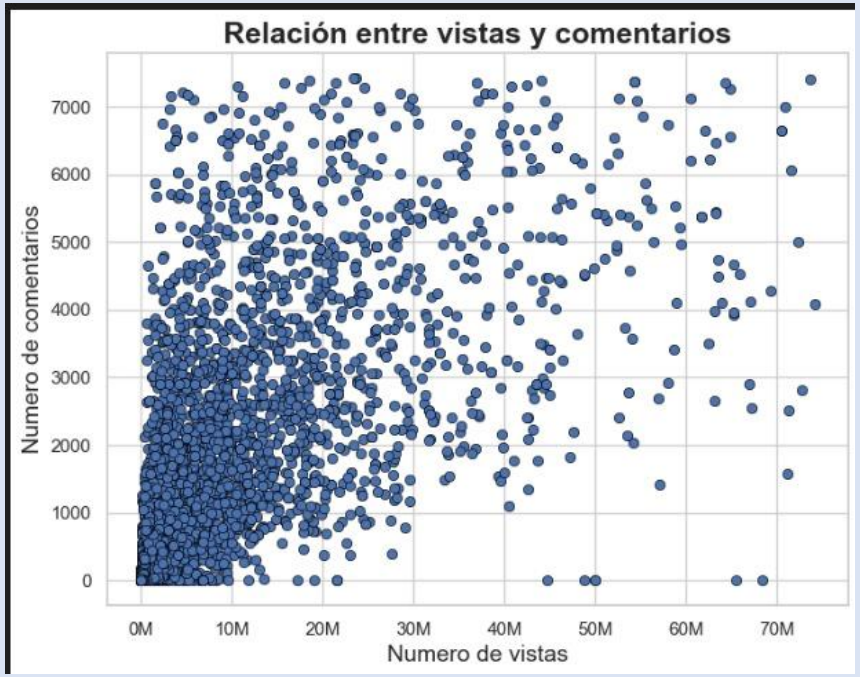


HALLAZGOS IMPORTANTES

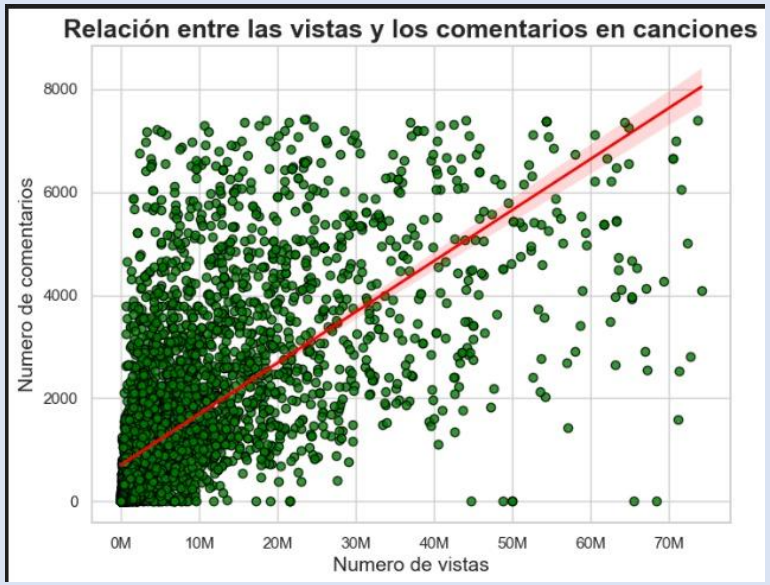
- Lo que mas nos llamo la atención fue la relación entre los me gusta con los comentarios y las vistas con las reproducciones.
- Dado a las reproducciones que pueden llegar a tener las canciones en esta plataforma, pueden tener una correlación alta o baja, como por ejemplo dependiendo la cantidad de likes que tenga una canción en Spotify las reproduciones pueden entrar a la par o en una correlación de promedio.

ANALISIS DE DATOS:

- Como en nuestra grafica de mapa de calor no obtuvimos correlación alguna decidimos hacer un analisis de datos exploratorio como, hacer correlaciones con nuestra variable mas compleja ante nosotros la cual era: “REPRODUCCIONES”
- *La primera que decimos hacer fue con “relación entre vistas y comentarios”



- La segunda fue “relación entre las vistas y los comentarios”



La tercera “relación entre producciones y vistas”