UNIVERSIDAD CATÓLICA BOLIVIANA “SAN PABLO”

UNIDAD ACADÉMICA REGIONAL COCHABAMBA

Departamento de Ciencias Exactas e Ingeniería

Carrera de Ingeniería de Sistemas



Proyecto Minería de Datos y Visualización de Resultados

*Proyecto Final de la Materia de Sistemas de Información II*

**Julia Valentina Gutiérrez Munzón**

**Héctor Marcelo Hidalgo Quena**

**Daniela Michel Vela Oblitas**

Cochabamba – Bolivia

Junio de 2019

**ÍNDICE GENERAL**

[1. PROBLEMA DE INVESTIGACIÓN 3](#_Toc11846102)

[2. FORMULACIÓN DEL PROBLEMA 3](#_Toc11846103)

[3. JUSTIFICACIÓN DE LA INVESTIGACIÓN 3](#_Toc11846104)

[4. MARCO REFERENCIAL 3](#_Toc11846105)

[4.1. Origen/fuente de los datos 3](#_Toc11846106)

[4.2. Contextualización de los datos 4](#_Toc11846107)

[4.3. Descripción de las variables más importantes 4](#_Toc11846108)

[5. MARCO TEÓRICO 5](#_Toc11846109)

[5.1. Metodología aplicada a la minería de datos 5](#_Toc11846110)

[5.2. Algoritmo de minería de datos seleccionados y empleados 6](#_Toc11846111)

[6. MARCO APLICATIVO 6](#_Toc11846112)

[6.1. Preparación y limpieza de datos 6](#_Toc11846113)

[6.2. Fases de la metodología seleccionada 6](#_Toc11846114)

[6.3. Resultados de algoritmos de minería de datos aplicado 7](#_Toc11846115)

[6.4. Beneficios y costos 7](#_Toc11846116)

[*6.5.* *Dashboard* 8](#_Toc11846117)

[6.6. Resultados 14](#_Toc11846118)

[6.7. Conclusión 14](#_Toc11846119)

[7. BIBLIOGRAFÍA 14](#_Toc11846120)

[8. ANEXO 15](#_Toc11846121)

[Anexo 1 15](#_Toc11846122)

[Anexo 2 15](#_Toc11846123)

[Anexo 3 19](#_Toc11846124)

# PROBLEMA DE INVESTIGACIÓN

La música se encuentra presente en nuestro diario vivir, cuando uno se encuentra en la calle, en el transporte público, en el trabajo o en el centro de estudios. Ésta es útil para relajarse, trabajar o simplemente distraerse un poco. Las aplicaciones para reproducir música juegan un papel muy importante en esto, ya que influyen en la música que escuchamos. Hay una variedad de aplicaciones y plataformas de música y videos, pero en esta ocasión se tomará una de las más conocidas y populares de hoy en día: Spotify.

Spotify es una plataforma que ofrece servicios de música, videos y podcasts, la aplicación nos permite ver canciones en base a ciertos criterios lo cual nos facilita la búsqueda. Es por eso que se considera necesario ver la relación que tienen estos criterios. Esta relación es relevante tanto para los artistas que publican sus canciones por este medio y para los usuarios para que sepan que música escucharan.

# FORMULACIÓN DEL PROBLEMA

¿Cómo se clasifican y seleccionan las canciones de Spotify para dar recomendaciones a los usuarios?

# JUSTIFICACIÓN DE LA INVESTIGACIÓN

Hoy en día, Spotify es una de las plataformas *streaming* más usadas para escuchar música, que almacena millones de canciones de diferentes género para todos los gustos. Es por eso que se considera importante poder realizar un análisis para la industria de la música.

# MARCO REFERENCIAL

## Origen/fuente de los datos

* <https://www.kaggle.com/edalrami/19000-spotify-songs>
* <http://millionsongdataset.com/>

## Contextualización de los datos

La base de datos contiene datos sobre canciones de Spotify y los datos relacionados con las mismas, se tiene:

* Las canciones más populares.
* Más de 18 000 registros que contienen información relevante de canciones.

## Descripción de las variables más importantes

Algunas de las variables más relevantes del *dataset* y su explicación [1] son:

* *song\_name*.- Nombre de la canción
* *song\_popularity*.- Nivel de popularidad
* *danceability*.- Describe qué tan adecuada es una pista para bailar basada en una combinación de elementos musicales que incluyen el tempo, la estabilidad del ritmo, la fuerza del ritmo y la regularidad general.
* tempo.- El tempo total estimado de una pista en tiempos por minuto (BPM). En terminología musical, el tempo es la velocidad o el ritmo de una pieza dada, y se deriva directamente de la duración media del ritmo.
* *energy*.- Representa una medida perceptiva de intensidad y actividad. Típicamente, las pistas energéticas se sienten rápidas, ruidosas y ruidosas.
* *loudness*- La sonoridad general de una pista en decibelios (dB). Los valores de sonoridad se promedian en toda la pista y son útiles para comparar la sonoridad relativa de las pistas.
* *speechiness*.- Detecta la presencia de palabras habladas en una pista. Cuanto más exclusiva sea la grabación similar a un discurso (por ejemplo, un programa de entrevistas, audiolibro, poesía), más cerca del valor de atributo será 1.0.
* *instrumentalness*.- Predice si una pista no contiene voces.
* *liveness*.- Detecta la presencia de un público en la grabación. Los valores de vida más altos representan una probabilidad mayor de que la pista se realizó en vivo.
* *audio\_valence*.- Describe la positividad musical transmitida por una pista. Las pistas con alta valencia suenan más positivas, mientras que las pistas con baja valencia suenan más negativas.

# MARCO TEÓRICO

## Metodología aplicada a la minería de datos

Se eligió la metodología CRISP-DM que incluye un modelo y una guía, estructurados en seis fases (ver anexo 1), algunas de las cuales son bidireccionales, es decir que de una fase en concreto se puede volver a una fase anterior para poder revisarlas, por lo que la sucesión de fases no tiene porqué ser ordenada desde la primera hasta la última. Las fases son:

* Comprensión del negocio: se deben entender los objetivos del proyecto y los requerimientos desde una perspectiva del negocio y luego convertir este conocimiento en una definición de un problema de explotación de información y diseñar un plan preliminar para lograr dichos objetivos.
* Comprensión de los datos: comienza con la recolección inicial de datos y procede con las acciones para familiarizarse con ellos, identificar problemas de calidad, identificar primeras pautas en los datos o detectar subconjuntos interesantes de las hipótesis de información oculta.
* Preparación de los datos: cubre todas las actividades para construir el conjunto de datos final desde los datos iniciales, las tareas de esta fase pueden ser realizadas muchas veces y sin un orden preestablecido, incluye tanto la selección de tablas, registros y atributos como transformación y limpieza de datos para herramientas de modelado.
* Modelado: construcción de uno o varios modelos que aparentan tener la mayor calidad desde una perspectiva de análisis.
* Evaluación: se evalúa el modelo y revisión de los pasos ejecutados para la construcción del modelo con la finalidad de asegurar el logro de los objetivos.
* Implantación: podría ser la fase más simple, como la generación de un reporte, o más compleja, con implementaciones por toda una empresa.

La elección de la metodología CRISP-DM fue debido a que esta metodología no es rígida, por el contrario es bastante flexible, además es un proceso bastante sencillo de comprender. [2]

## Algoritmo de minería de datos seleccionados y empleados

Se escogió algoritmos de regresión, que predicen una o más variables numéricas continuas, como pérdidas o ganancias, basándose en otros atributos del conjunto de datos. Se empleó este tipo de algoritmo ya que nos permite ver las dependencias entre las variables. [3]

# MARCO APLICATIVO

## Preparación y limpieza de datos

Se realiza una limpieza de nuestros datos, eliminando los valores que pueden causar problemas como los valores en blanco.

## Fases de la metodología seleccionada

* + 1. **Comprensión del negocio**

Primero que nada se analizó el objetivo planteado previamente para definir el problema y diseñar un plan para saber que decisiones tomar.

* + 1. **Comprensión de los datos**

En base al *dataset* obtenido de canciones de Spotify, se analizó los diferentes datos para ver de que forma se podría trabajar con ellos. Se vio que podría existir una relación entre algunas variables como entre *popularity* y *danceability*, al igual que entre *energy* y *danceability*.

* + 1. **Preparación de los datos**

Se volvió a realizar la limpieza de los datos y se verificó que los datos estén en el formato establecido para poder trabajar de manera correcta con los mismos.

* + 1. **Modelado**

Se usó un modelo de regresión lineal para ver si existe una relación entre la variable.

* + 1. **Evaluación**

Se analizó el modelo para ver si cumple con los objetivos esperados.

* + 1. **Implantación**

## Resultados de algoritmos de minería de datos aplicado

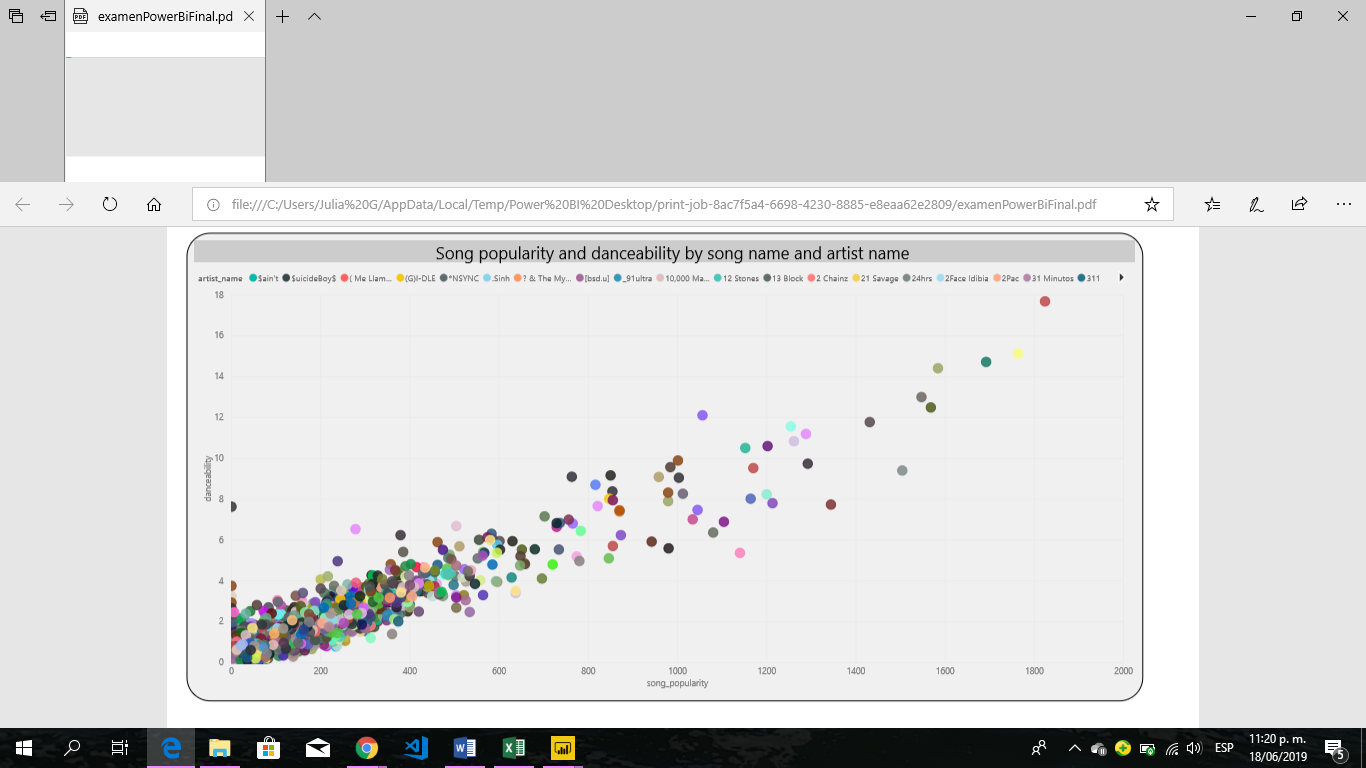
## Beneficios y costos

Armado con sólidas herramientas de procesamiento de datos y posibilidades infinitas de crecimiento, está bastante claro que Spotify puede alcanzar el dominio mundial. Con más de 100 millones de usuarios (50 millones de pago) y un catálogo de más de 30 millones de canciones, Spotify está presente en todo el globo y es una de las mayores fuentes de ingresos para los artistas.

Uno de los puntos fuertes de la compañía es su constante mejora de las posibilidades que ofrece a los artistas para hacer marketing a través del propio Spotify, es por eso la importancia de hacer un análisis de su información, para que los artistas puedan tomar decisiones en base a la música más escogida en base a los gustos de los usuarios.

## *Dashboard*

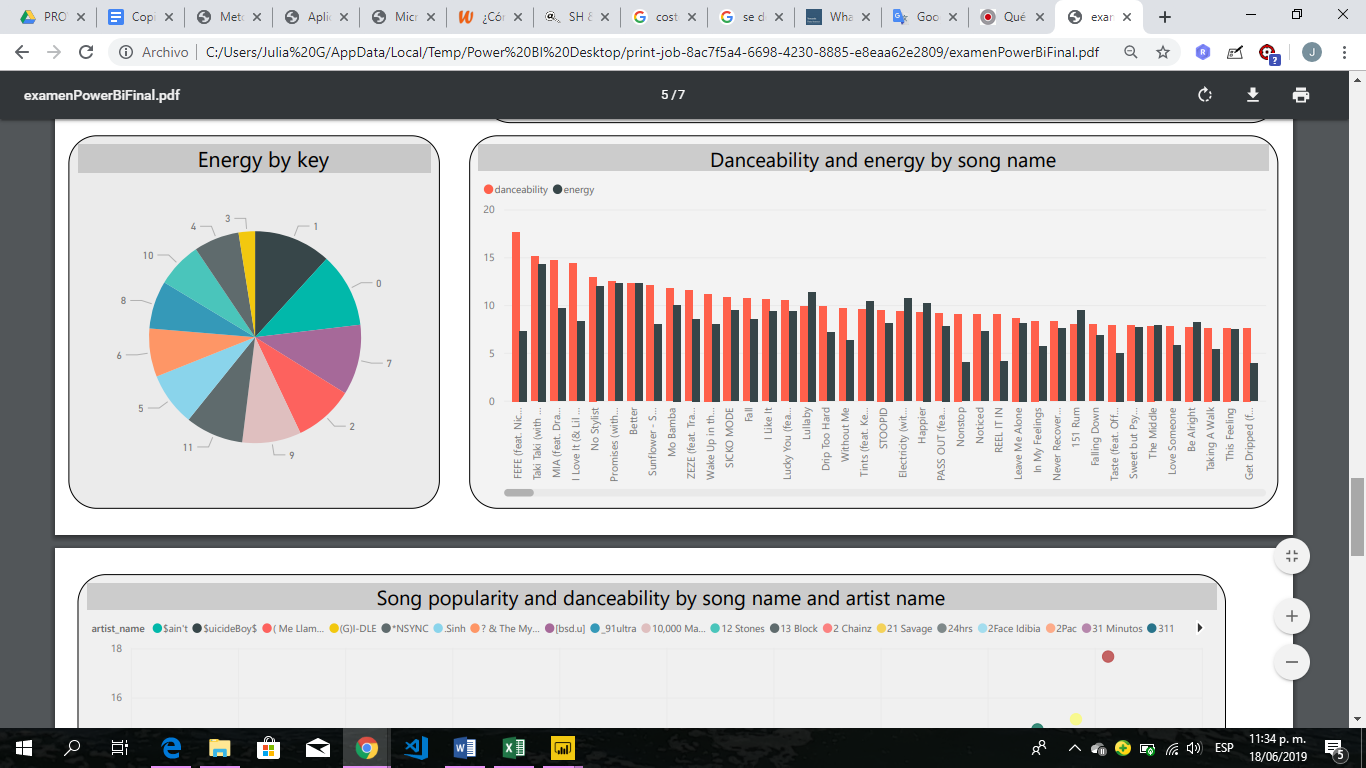
Figura 1 Popularidad y *danceability* por nombre de la canción y artista



Fuente: Elaboración propia 2019

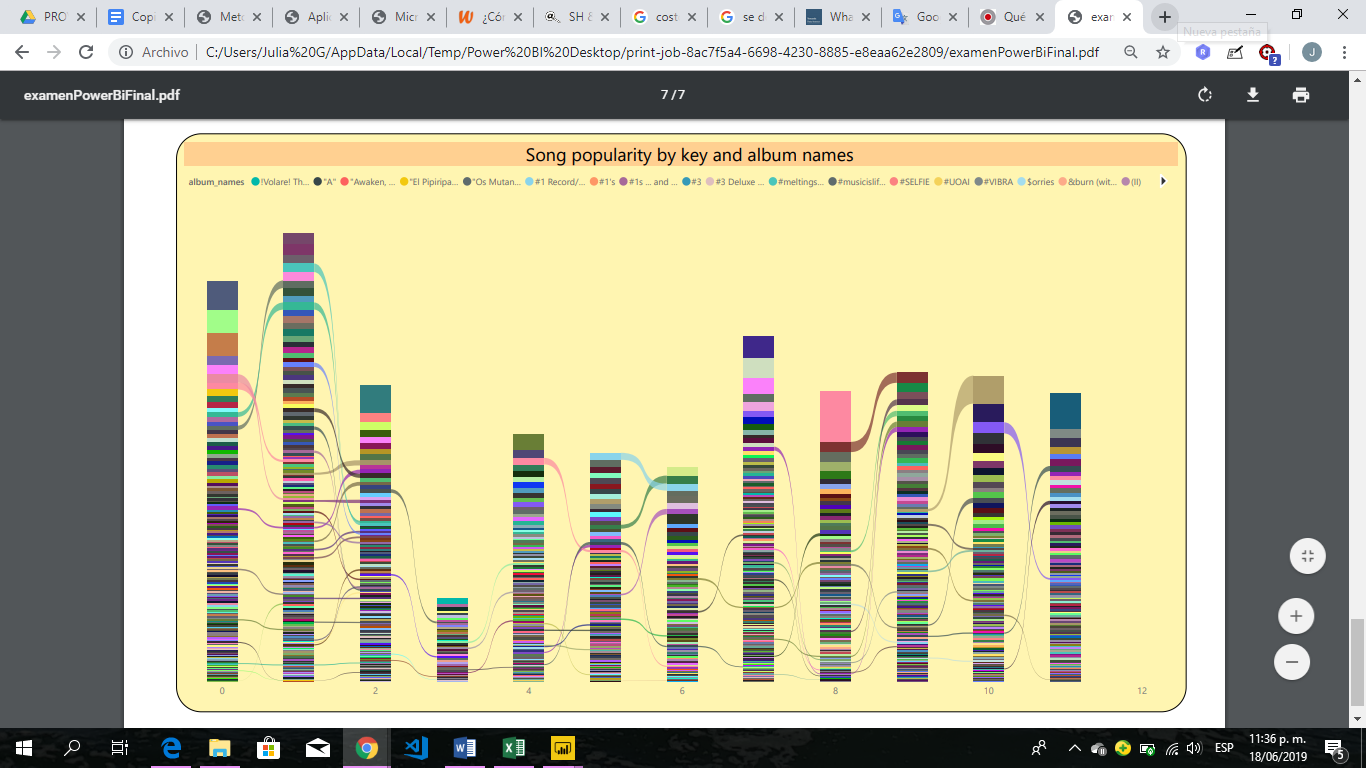
Primero se realizó una gráfica para ver la relación entre la popularidad y la *danceability* agrupando por nombre de la canción y el artista. Así se puede ver que hay una relación entre la popularidad y la *danceability* son proporcionalmente incrementales.

Figura 2 *Danceability* y energía por nombre de la canción



Fuente: Elaboración propia 2019

Al realizar esta gráfica se vio que existe una relación entre la variable de *danceability* y energía, si una canción tiene un mayor nivel de *danceability* lo más probable es que sea porque su energía es elevada también.

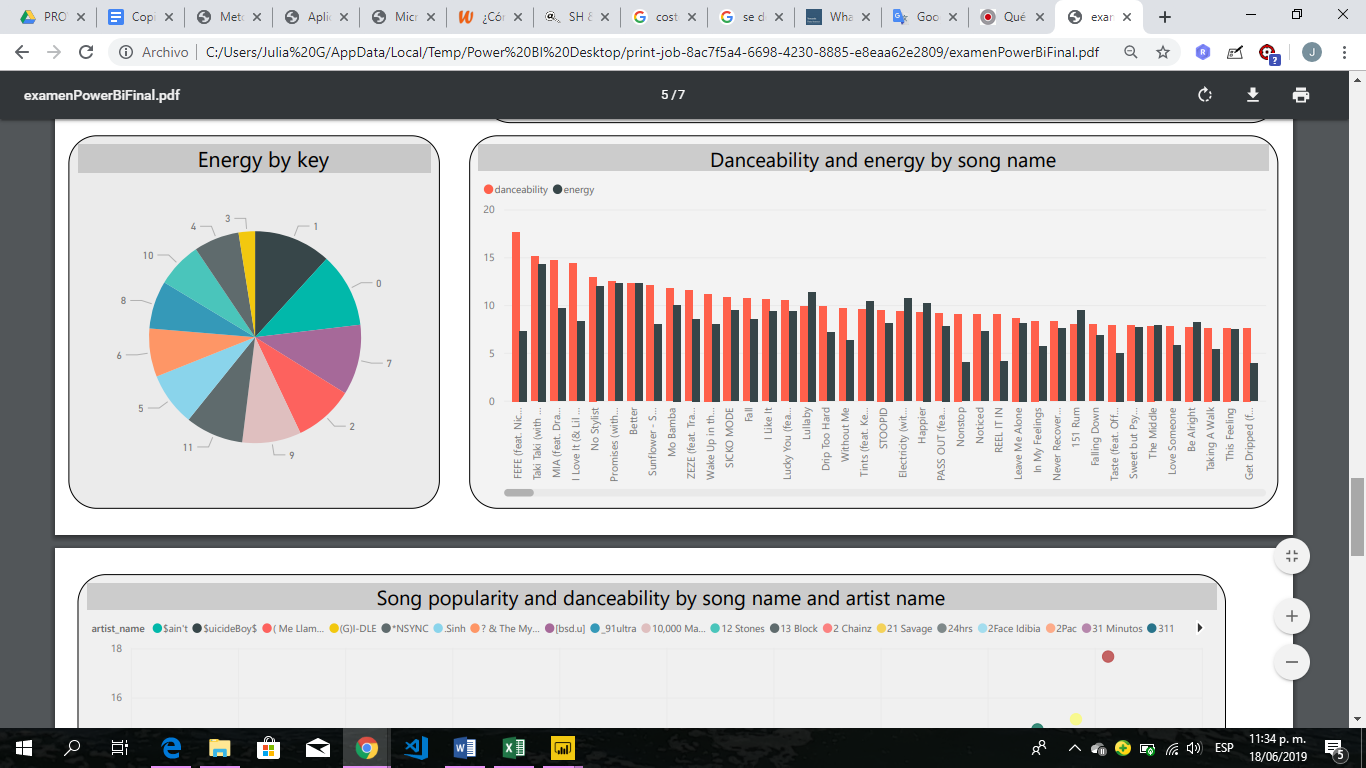
Figura 3 Popularidad por clave y album

Fuente: Elaboración propia 2019

**Figura 4 *Danceability* por clave y artista**

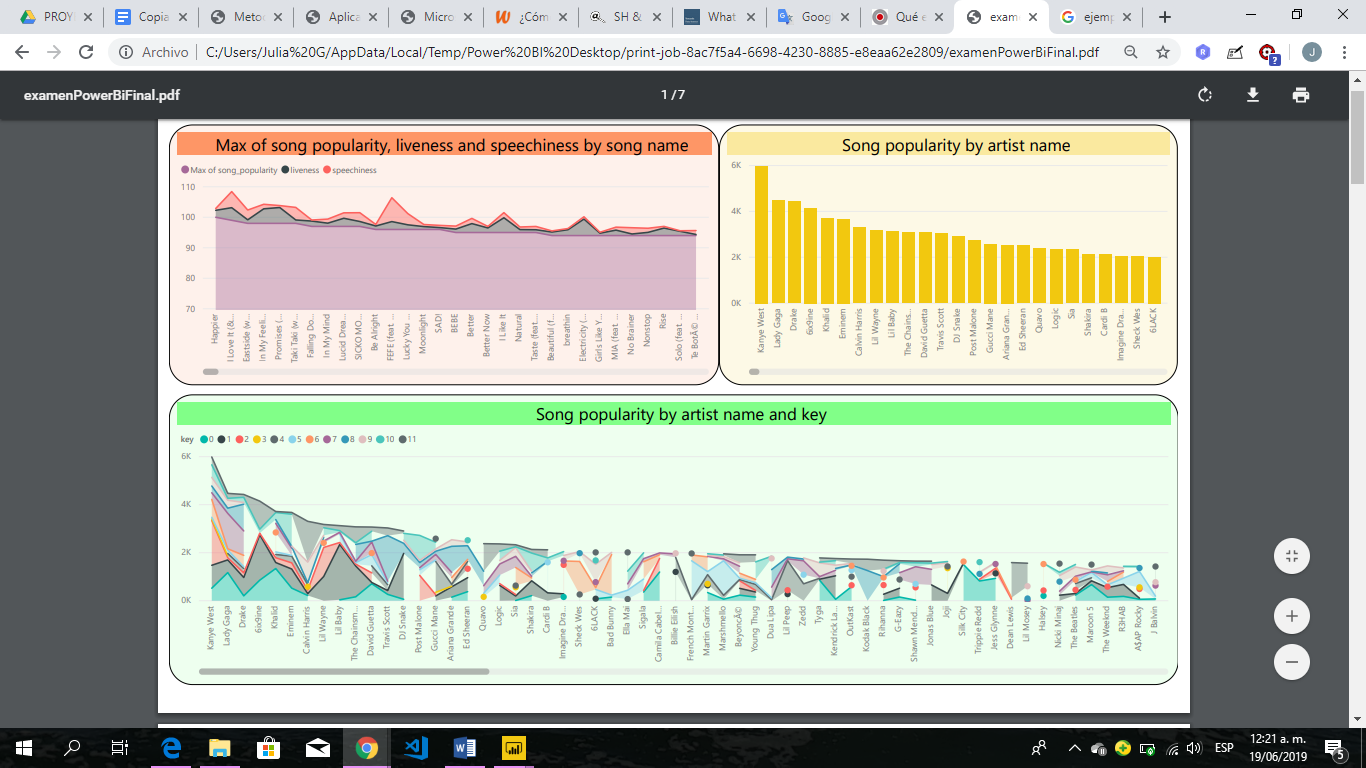
Fuente: Elaboración propia 2019

**Figura 5 Energía por clave**



Fuente: Elaboración propia 2019

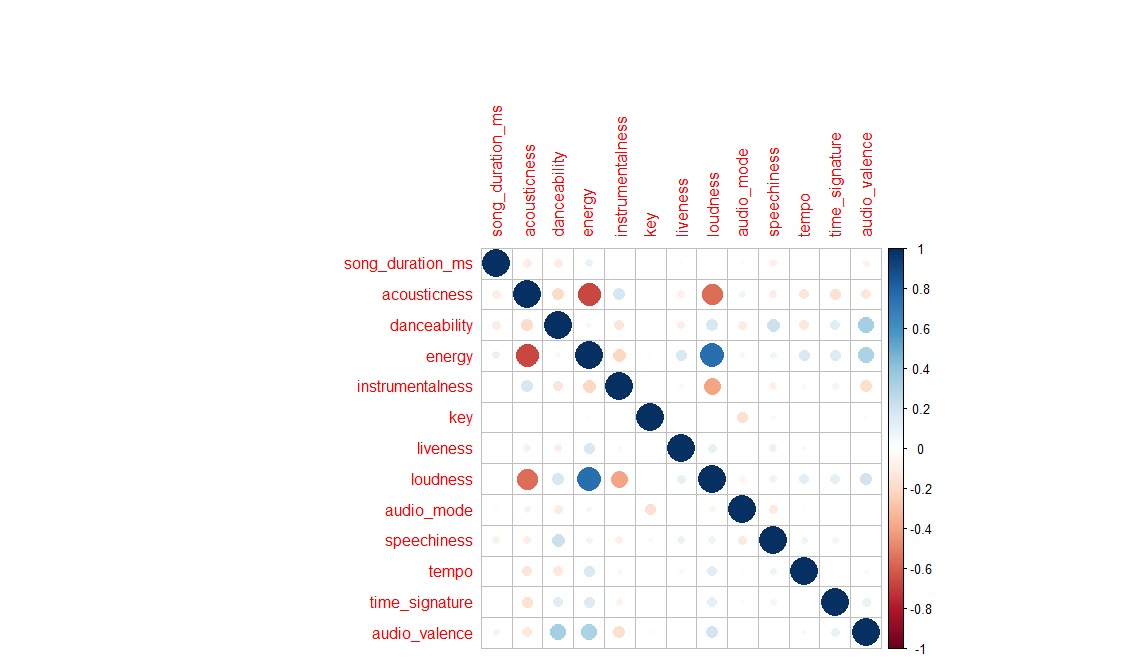
**Figura 6 Popularidad por nombre del artista y clave**



Fuente: Elaboración propia 2019

Luego se analizó el papel que juega la clave en la que esta escrita la canción en estas relaciones. Se vio que algunas claves pueden alcanzar mayor nivel de popularidad y otras no, es por eso que algunas claves tienen menos canciones.

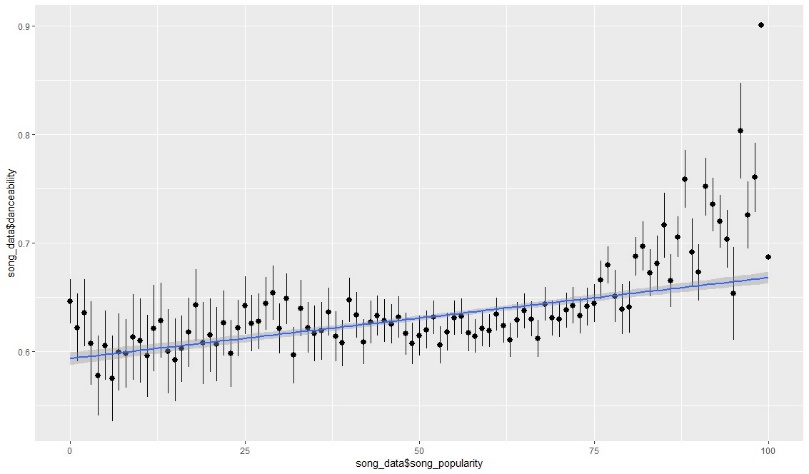
**Figura 7 Matriz de correlación**



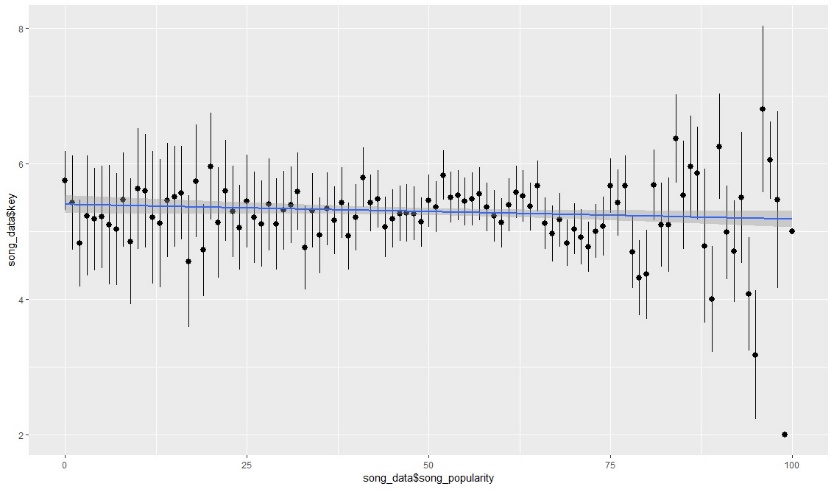
Fuente: Elaboración propia 2019

Además se realizó una matriz de correlación para ver las dependencias de las variables. En donde los puntos están pintados más fuertes se puede ver que existe una mayor relación. También se puede apreciar el comportamiento de las variables en las gráficas de barras. (ver anexo 2)

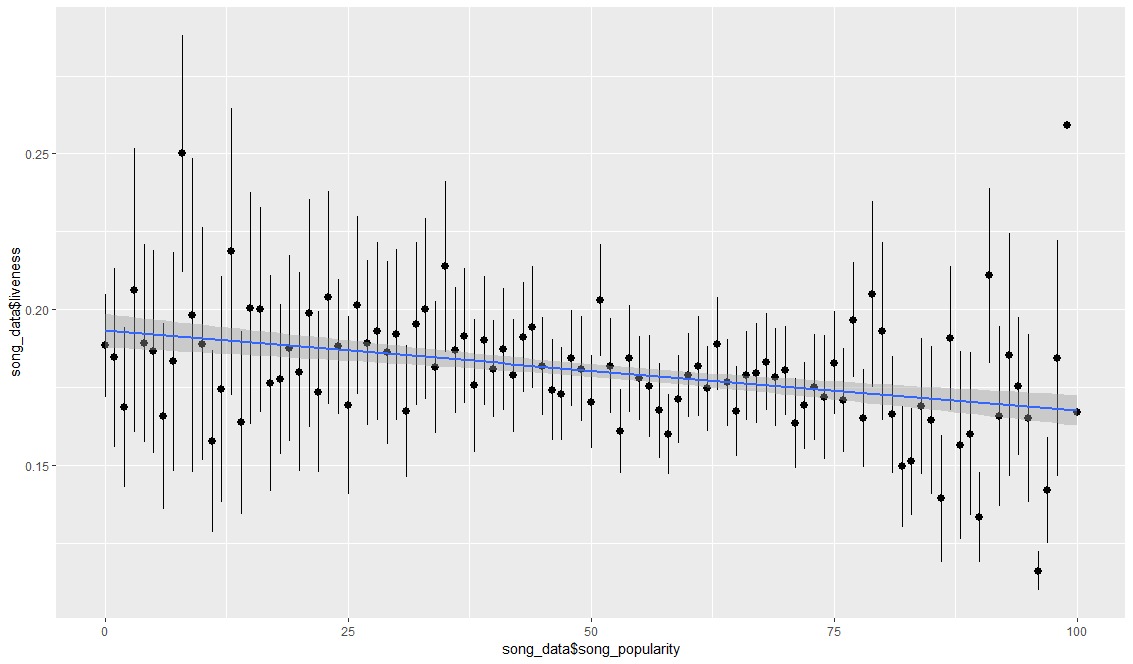
**Figura 8 Regresión lineal *danceability* por popularidad**



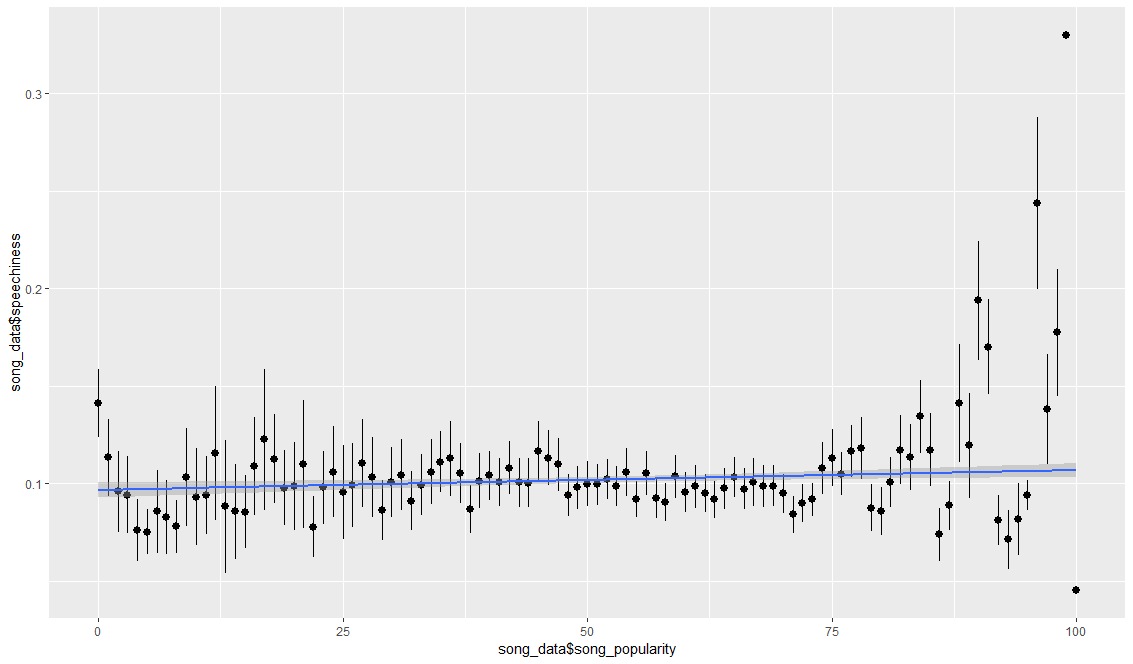
Fuente: Elaboración propia 2019

**Figura 9 Regresión lineal *danceability* por key**

Fuente: Elaboración propia 2019

**Figura 10 Regresión lineal *liveness* por popularidad** 

Fuente: Elaboración propia 2019

**Figura 11 Regresión lineal *speechiness* por key** 

Fuente: Elaboración propia 2019

Y se usó el modelo de regresión lineal para ver las relaciones.

A su vez se vio el rol que cumplen otras variables, pero no se encontró una relación directa como se esperaba. (ver anexo 3)

## Resultados

Existen relaciones entre algunas variables, podemos ver que las canciones que tienen mayor danceability tienden a ser más populares, lo cual esta relacionado íntimamente con las claves. La clave 1 es una de las más bailables y por eso tiende a ser más popular. Si se tuviera un conjunto más grandes de datos se podría lograr mejores resultados.

## Conclusión

La capacitación para entender el funcionamiento y uso del análisis de gustos musicales que usa Spotify es bastante interesante ya que esta nos muestra una forma práctica del uso de predicciones utilizando las variables que se presentan en este informe. Se pudo ver que existen relaciones entre estas, lo cual facilita la tarea ya mencionada .

# BIBLIOGRAFÍA

[1] ASHRITH (2018). “What Makes a Song Likeable?”. En: < <https://towardsdatascience.com/what-makes-a-song-likeable-dbfdb7abe404>> , (fecha de consulta 14/05/2019).

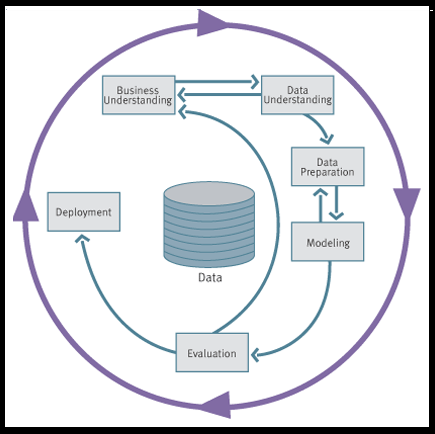
[2] MOINE, HAEDO Y GORDILLO (s.f). Estudio comparativo de metodologías para minería de datos. Argentina: s.e.

[3] “Algoritmos de minería de datos” (2018). En: < <https://docs.microsoft.com/es-es/sql/analysis-services/data-mining/data-mining-algorithms-analysis-services-data-mining?view=sql-server-2017>> , (fecha de consulta 15/05/2019).

# ANEXO

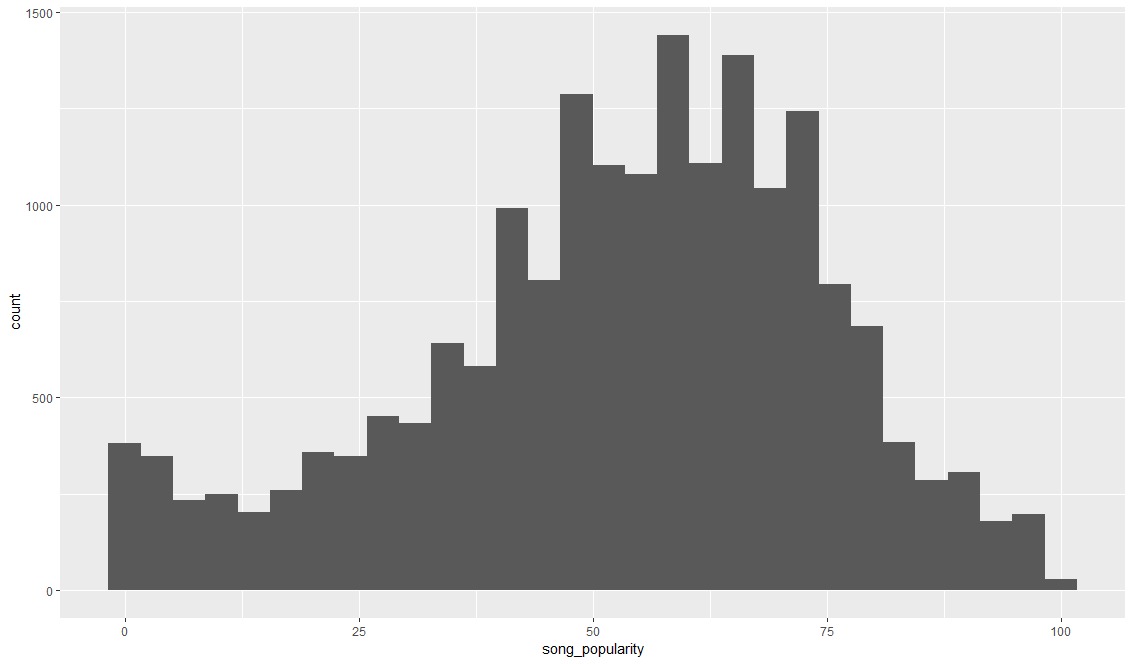
# Anexo 1

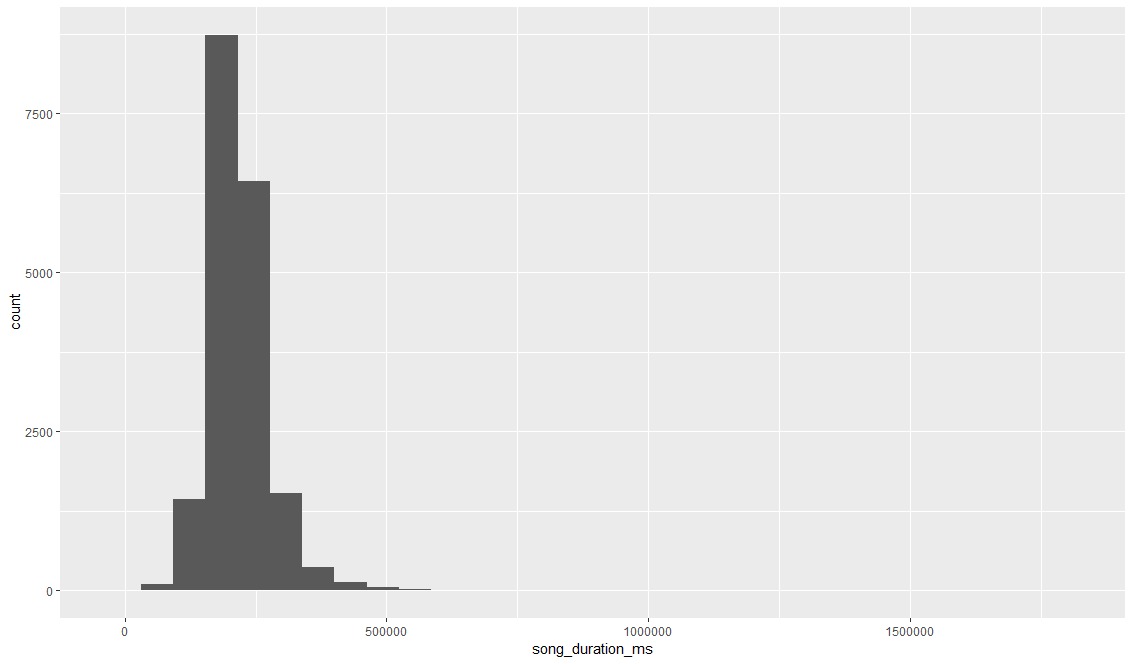
**Flujo de trabajo de la metodología CRISP-DM**

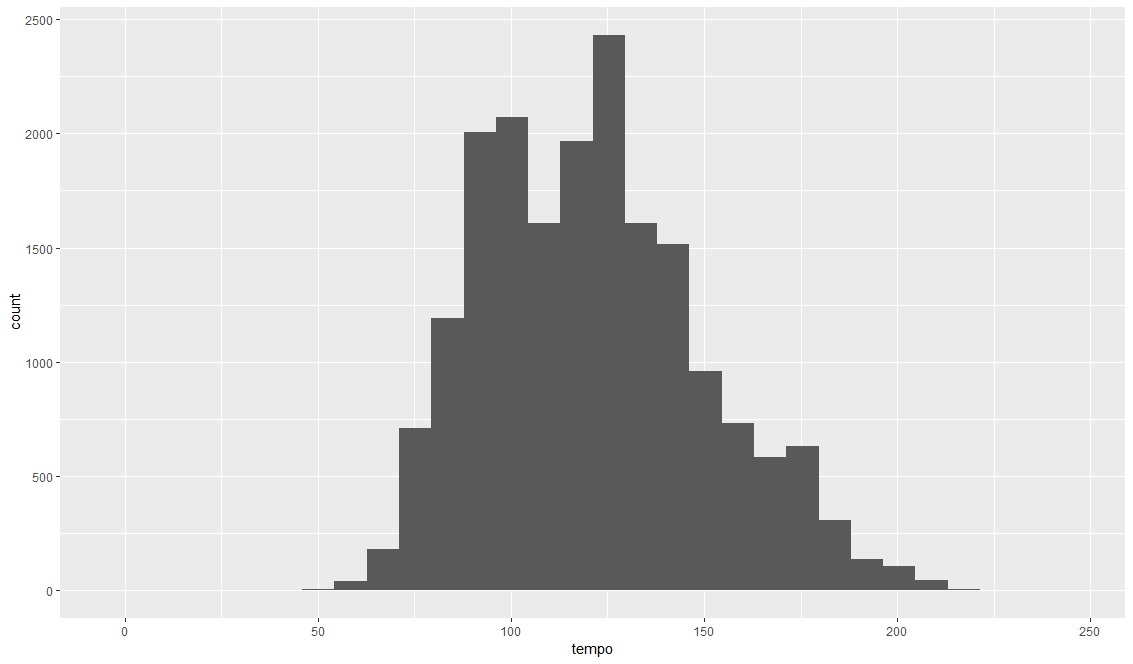
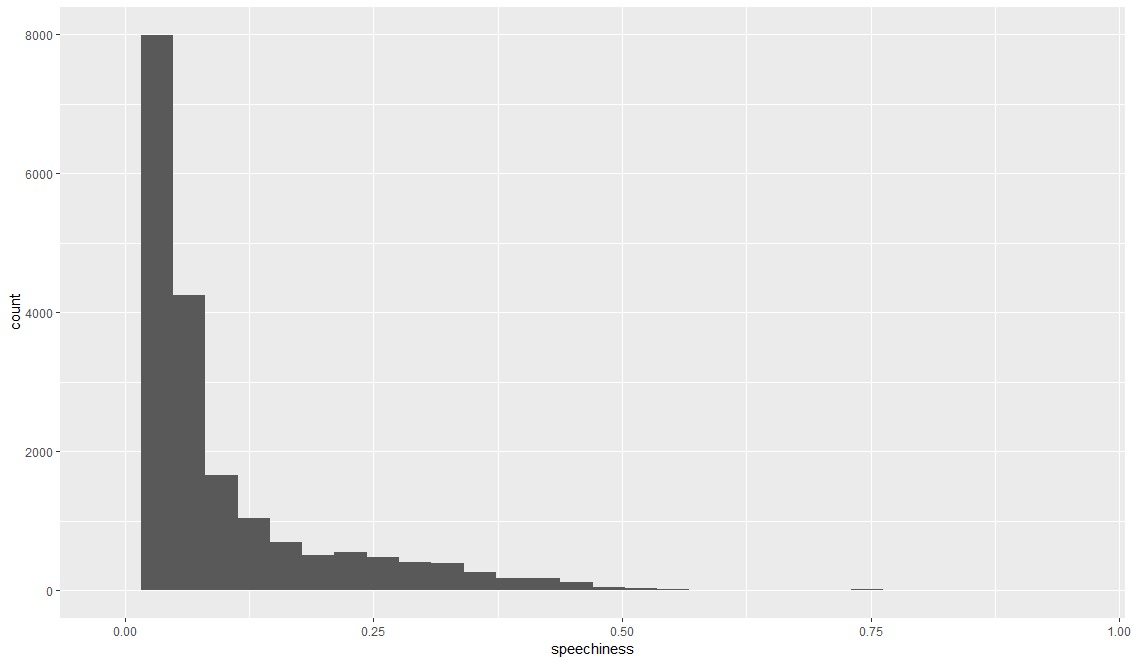
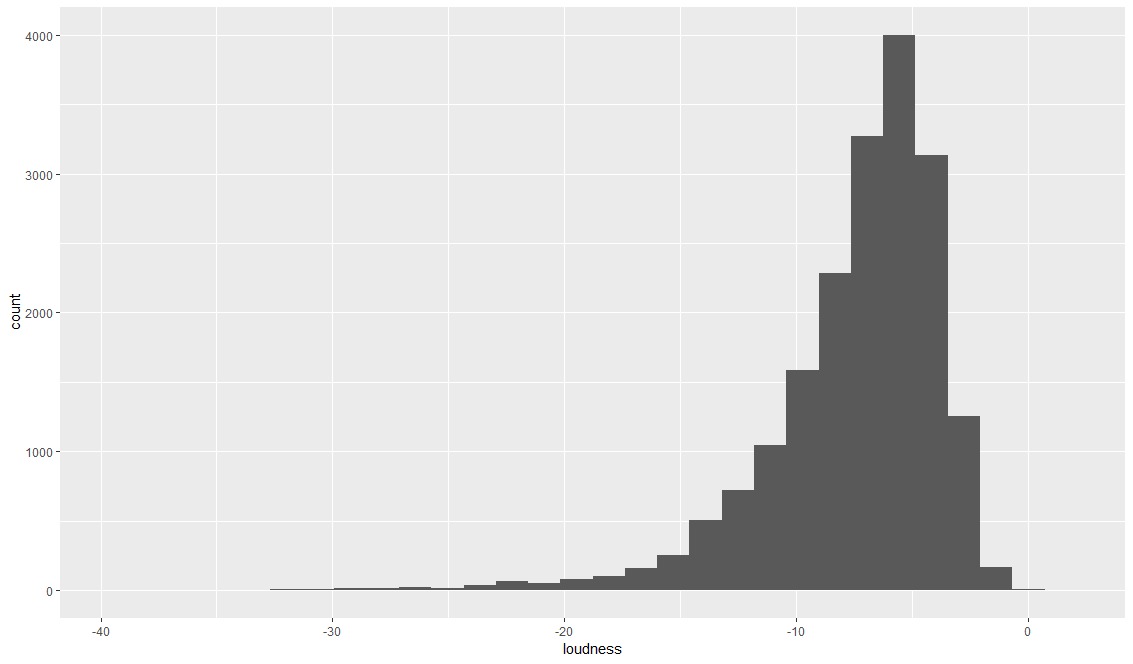
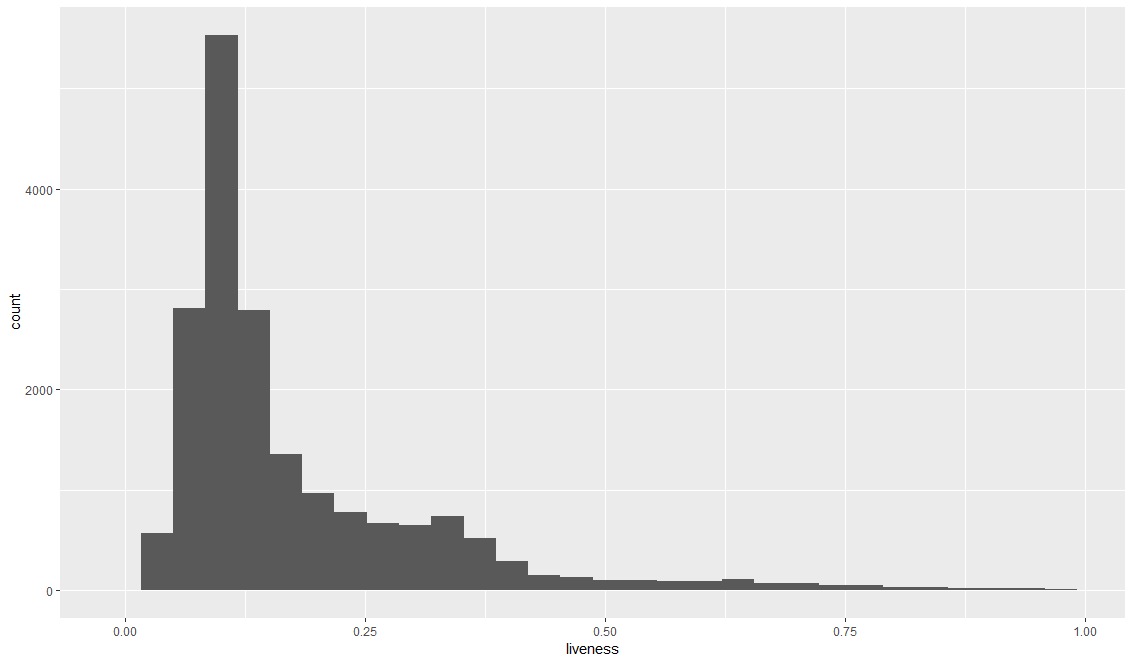
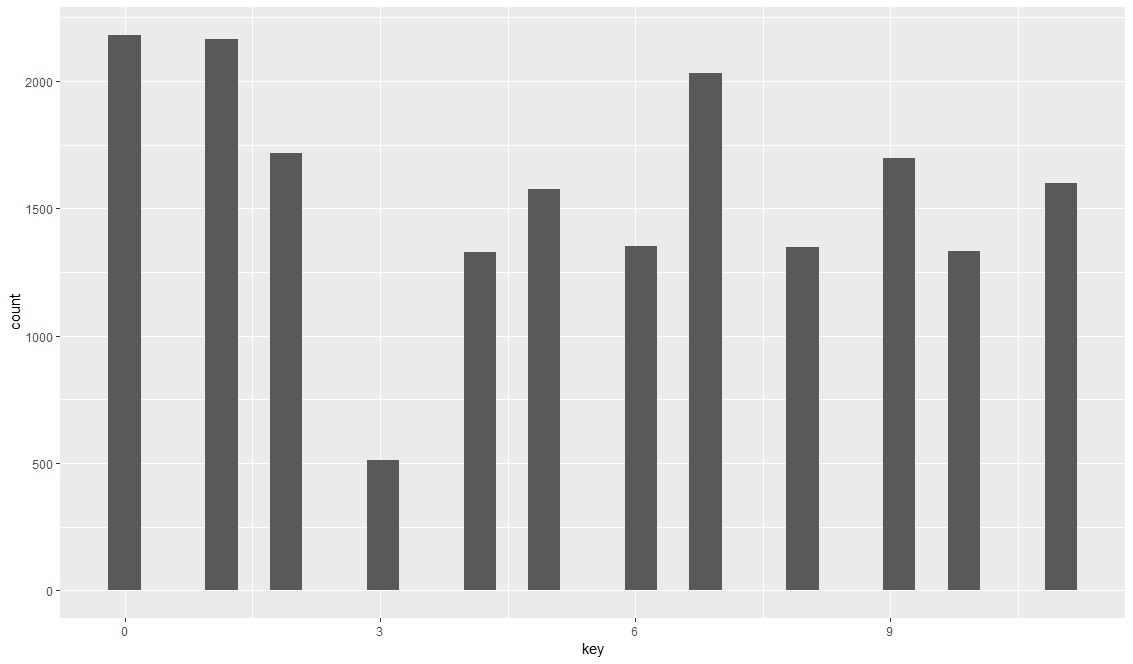
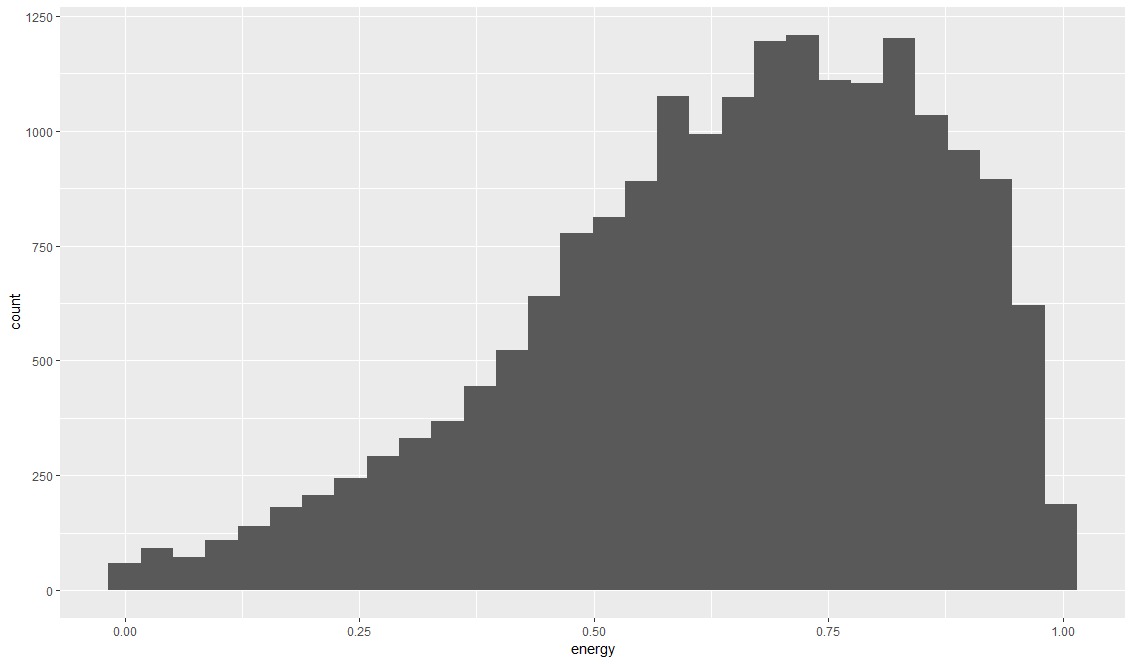
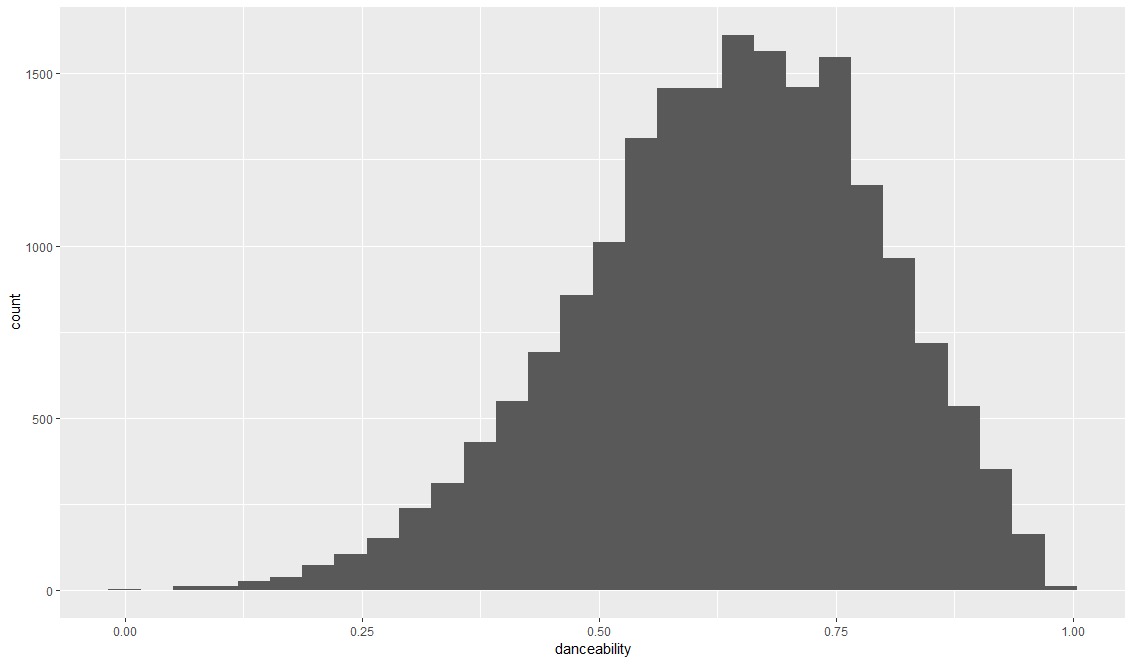
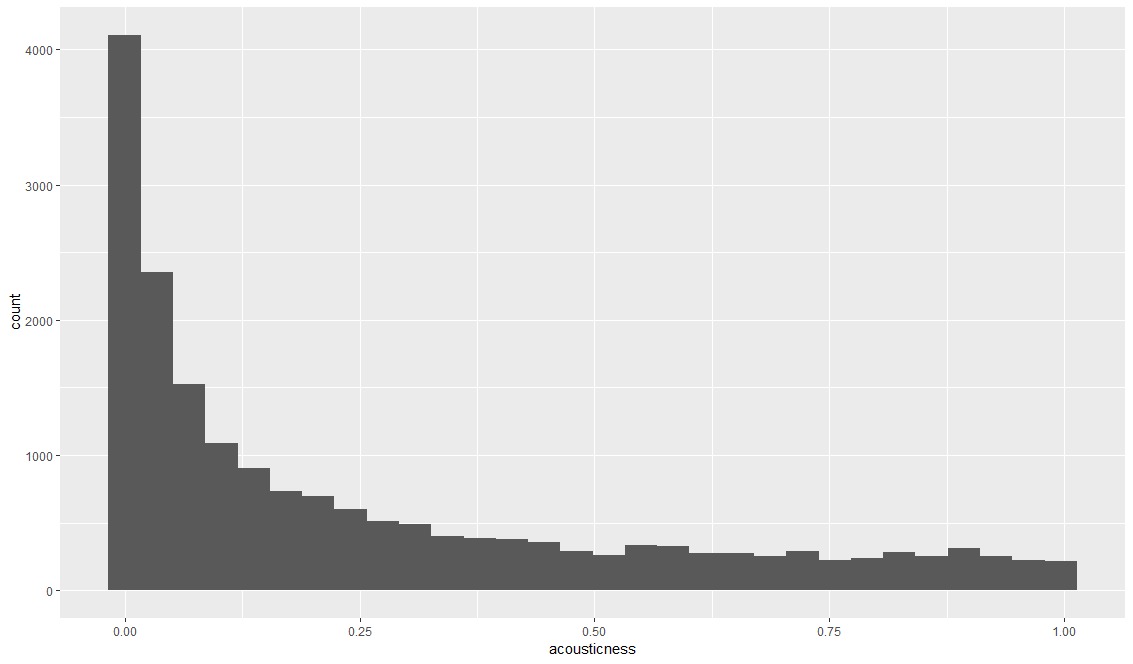
****

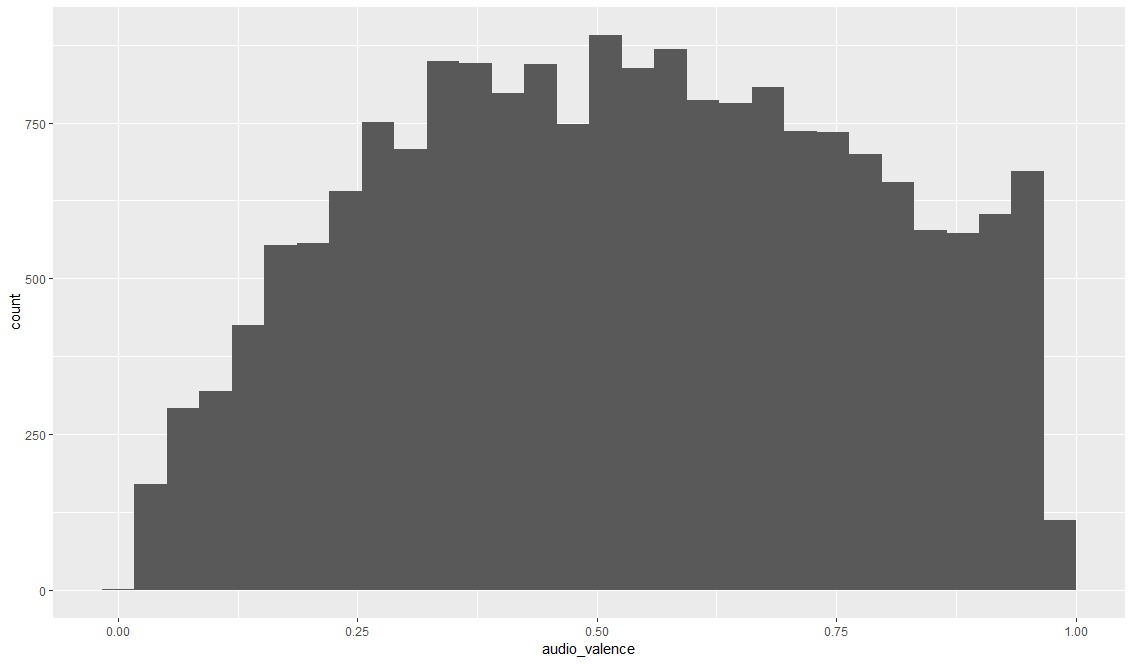
# Anexo 2

Gráficas de barras de las variables









# Anexo 3

Gráficas de relación

