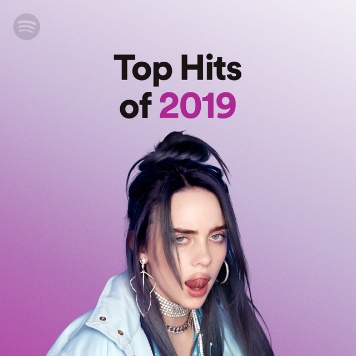
**Desafío Clase 37. PRIMERA ENTREGA DEL PROYECTO FINAL**

**Top Hits Spotify from 2000-2019**







Contenido

[Objetivos Generales: 3](#_Toc109745718)

[Objetivos Específicos: 3](#_Toc109745719)

[Presentación de la empresa, organización o problema específico. 4](#_Toc109745720)

[Top Hits Spotify from 2000-2019 4](#_Toc109745721)

[Preguntas y objetivos de la investigación. 5](#_Toc109745722)

[Conformación del equipo de trabajo. 5](#_Toc109745723)

[Indicación de la fuente del dataset y los criterios de selección (Data Acquisition) 5](#_Toc109745724)

[Descripción de las variables 5](#_Toc109745725)

[Generación del primer Data Wrangling y EDA, apuntado a sus datos (insights) univariado, bivariado y multivariado. 7](#_Toc109745726)

[Análisis de componentes principales. 7](#_Toc109745727)

[Contar la historia de sus datos 7](#_Toc109745728)

[Filtros aplicados a los datos. Distribución. Dataset final para analizar. 7](#_Toc109745729)

# Objetivos Generales:

1. Entender el problema de negocio e identificar los elementos a ser considerados para el planteamiento de un Modelo de Data Science.
2. Describir los datos de negocio y las relaciones entre datos mediante el Análisis Exploratorio de Datos.
3. Construir una presentación ejecutiva para la alta gerencia mostrando los resultados obtenidos.

# Objetivos Específicos:

1. Desarrollar las instancias de Data Acquisition y Data Wrangling en el trabajo final.
2. Lograr una articulación en equipo y una división de tareas adecuadas a los objetivos.
3. Realizar Filtrado
4. Describir qué significa cada variable, cómo se comporta.
5. Especificar las distribuciones y relaciones (géneros, sexo, edad, IVA, tipo de empresa).

# Presentación de la empresa, organización o problema específico.

### [Top Hits Spotify from 2000-2019](https://www.kaggle.com/code/thomaslazarus/visualizing-spotify-top-hits/data)

Hace 20 años era sencillo identificar quienes eran los reyes y reinas de la música pop en el mundo: bastaba con mirar quién encabezaba las listas de popularidad de singles y discos vendidos. Sin embargo la industria discográfica ha pasado a tener un sistema múltiple de medición, que se asemeja un poco al sistema utilizado por el Boxeo, en el que hay múltiples campeones para muchas categorías diferentes.

Los fans más adultos recordarán las listas de Billboard como el indicador más importante de quienes eran los artistas más exitosos en el mundo, pero esto ya no es el caso, pues ahora existen multitud de listas, todas con campeones diferentes.

Por ejemplo, están las listas de Spotify (la diaria y la semanal), y están las listas de Apple Music. YouTube también tiene sus listas, al igual que Shazam, Deezer y también iTunes. Cada servicio monitoriza un comportamiento distinto. Las listas oficiales toman una combinación de ventas, reproducciones en streaming y descargas, pero con un límite en cuanto a la cantidad de veces que un solo artista puede aparecer en la lista. Shazam mide cuántas veces una canción es buscada a través de la app, las demás tabulan las reproducciones de sus propias plataformas. Debido a que cada una mide algo diferente, los resultados también tienden a ser muy variados.

Spotify es un servicio de música, podcasts y videos digitales que le brinda acceso a millones de canciones y otro contenido de creadores de todo el mundo. Spotify está disponible en una gran variedad de dispositivos, como ordenadores, teléfonos, tabletas, altavoces, televisores o coches, y puedes pasar fácilmente de uno a otro.

El problema está en que, mientras más listas existan, menos relevancia tienen y esto dificulta identificar y medir el éxito de un artista frente a otro. Si un artista es Nº1 en YouTube pero Nº20 en la lista semanal de Spotify, ¿significa que es más o menos popular que alguien que ocupe el Nº8 en ambas listas? ¿Qué pasa con artistas que no aparecen en listas de popularidad pero que hacen llenos totales de sus conciertos en recintos más grandes que las estrellas del streaming? ¿Qué tan populares son?

Este conjunto de datos contiene estadísticas de audio de las 2000 mejores pistas en Spotify de 2000 a 2019. Los datos contienen alrededor de 18 columnas, cada una de las cuales describe la pista y sus cualidades.

En el presente proyecto analizaremos el dataset para predecir la popularidad de las canciones, con base en la relación de sus variables, utilizando las bibliotecas de Python:

1. Numpy link: [https://numpy.org](https://numpy.org/)
2. Pandas link: [https://pandas.pydata.org](https://pandas.pydata.org/)
3. Matplotlib link: [https://matplotlib.org](https://matplotlib.org/)
4. Plotly link: [https://plotly.com](https://plotly.com/)

# Preguntas y objetivos de la investigación.

**El objetivo principal sería comprender qué variables infieren sobre la popularidad de las canciones.**

Resultaría interesante también ver qué características hacen que ciertas canciones o géneros sean o no populares, y ver cómo difieren los hábitos musicales a través de los años, es decir, como saber las características de las canciones populares con base en el análisis de diversas variables, tales como: bailabilidad, si son cantadas, etc y con esto saber el impacto que tienen sobre los oyentes estos artistas y estos géneros.

También resulta importante de este dataset la evolución de la música a través de los años con la aparición de nuevos subgéneros musicales pos sobre otros que tenían mayor popularidad al inicio de este siglo.

# Conformación del equipo de trabajo.

1. Sebastian Vadra
2. Santiago Pradini
3. Tania Garnica

# Indicación de la fuente del dataset y los criterios de selección (Data Acquisition)

Fuente: <https://www.kaggle.com/datasets/paradisejoy/top-hits-spotify-from-20002019>

Elegimos el dataset del Top Hit de Spotify, ya que deseamos analizar las variables, y con ello descubrir la relación que hay con la popularidad, y con esto predecir qué características deberían tener las canciones para tener alta popularidad, es decir, predecir si las canciones tendrán éxito o no.

Consideramos que este dataset tiene características que pueden ayudarnos en nuestro aprendizaje de Data Science, al tener variables valiosas para este análisis, que pueden ser bien aprovechadas.

### Descripción de las variables

• **artist:** Nombre del Artista.

• **song:** Nombre de la Pista.

• **duration\_ms:** Duración de la pista en milisegundos.

• **explicit:** La letra o el contenido de una canción o un video musical contienen uno o más de los criterios que podrían considerarse ofensivos o inadecuados para los niños.

• **year:** Año de lanzamiento de la pista.

• **popularity:** Cuanto mayor sea el valor, más popular será la canción.

• **danceability:** La bailabilidad describe qué tan adecuada es una pista para bailar en función de una combinación de elementos musicales que incluyen tempo, estabilidad del ritmo, fuerza del ritmo y regularidad general. Un valor de 0,0 es menos bailable y 1,0 es más bailable.

• **energy:** La energía es una medida de 0,0 a 1,0 y representa una medida perceptible de intensidad y actividad.

• **key:** La clave en la que se encuentra la pista. Los números enteros se asignan a tonos utilizando la notación estándar de clase de tono. P.ej. 0 = C, 1 = C♯/D♭, 2 = D, y así sucesivamente. Si no se detectó ninguna clave, el valor es -1.

• **loudness:** El volumen 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. El volumen es la cualidad de un sonido que es el principal correlato psicológico de la fuerza física (amplitud). Los valores suelen oscilar entre -60 y 0 db.

• **mode:** Mode indica la modalidad (mayor o menor) de una pista, el tipo de escala de la que se deriva su contenido melódico. Mayor está representado por 1 y menor es 0.

• **speechiness:** Speechiness detecta la presencia de palabras habladas en una pista. Cuanto más parecida a la voz sea la grabación (por ejemplo, programa de entrevistas, audiolibro, poesía), más cerca de 1,0 será el valor del atributo. Los valores superiores a 0,66 describen pistas que probablemente estén formadas en su totalidad por palabras habladas. Los valores entre 0,33 y 0,66 describen pistas que pueden contener tanto música como voz, ya sea en secciones o en capas, incluidos casos como la música rap. Los valores por debajo de 0,33 probablemente representen música y otras pistas que no sean de voz.

• **acousticness:** Una medida de confianza de 0,0 a 1,0 de si la pista es acústica. 1.0 representa una alta confianza en que la pista es acústica.

• **instrumentalness:** Predice si una pista no contiene voces. Los sonidos "Ooh" y "aah" se tratan como instrumentales en este contexto. Las pistas de rap o de palabras habladas son claramente "vocales". Cuanto más cerca esté el valor de instrumentalidad de 1,0, mayor será la probabilidad de que la pista no contenga contenido vocal. Los valores superiores a 0,5 pretenden representar pistas instrumentales, pero la confianza es mayor a medida que el valor se acerca a 1,0.

• **liveness:** Detecta la presencia de una audiencia en la grabación. Los valores de vivacidad más altos representan una mayor probabilidad de que la pista se interprete en vivo. Un valor superior a 0,8 proporciona una gran probabilidad de que la pista esté activa.

• **valence:** Una medida de 0.0 a 1.0 que describe la positividad musical transmitida por una pista. Las pistas con una valencia alta suenan más positivas (p. ej., felices, alegres, eufóricas), mientras que las pistas con una valencia baja suenan más negativas (p. ej., tristes, deprimidas, enfadadas).

• **tempo:** El tempo general estimado de una pista en pulsaciones por minuto (BPM). En terminología musical, el tempo es la velocidad o ritmo de una pieza dada y se deriva directamente de la duración promedio del tiempo.

• **genre:** Género de la pista.

# Generación del primer Data Wrangling y EDA, apuntado a sus datos (insights) univariado, bivariado y multivariado.

# Análisis de componentes principales.

# Contar la historia de sus datos

# Filtros aplicados a los datos. Distribución. Dataset final para analizar.