Herramientas Computacionales:

El Arte de la Analítica

Naomi Anciola Calderón

A01750363

Invierno 2021

Reporte

Instrucciones de Uso

Este código se desarrolló usando las librerías **CSV**, **Pandas**, **Numpy**, **Matplotlib**, **Seaborn** y **ScikitLearn**. Lo único necesario para que corra el programa es que el archivo py este en la misma carpeta que el archivo csv. Puede ser usado para otros datos csv y lo único que hay que modificar es el nombre del archivo dentro del programa

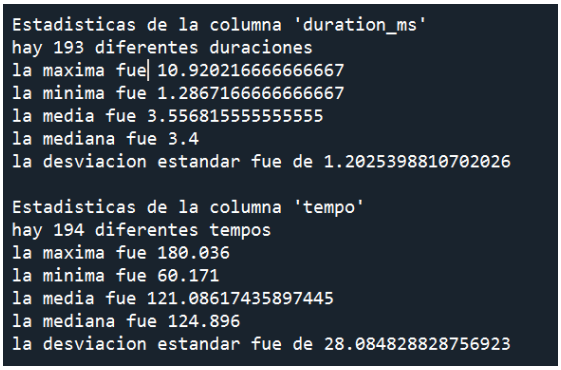
Una introducción al conjunto de datos

Este es un conjunto de datos de **Spotify**, el cual contiene información sobre diferentes canciones y sus atributos como tempo, emoción, voz, etc, además de si si le gusto a la persona o no. Esto se hizo con el propósito de construir un modelo para determinar si era probable que una canción le gustara o no a este usuario basado en sus atributos. Fue recuperado de kaggle y nos lo dio el profesor.

Datos, Variables, Tipos de Datos

Hay 14 variables: danceability, energy, key, loudness, mode, speechiness, acousticness, instrumentalness, liveness, valence, tempo, duration_ms, time_signature y liked. Todas estas son de tipo numérico, y varían entre int64 y float, es decir, que son cuantitativas. Por otra parte hay 195 registros.

Parte 1. Obtención de estadísticas descriptivas



Resultado en consola del programa

Escogí las variables que describen la **duración** de la canción y el **tempo**, ya que pensé que de estas se podrían sacar conclusiones más claras. Para las estadísticas de duration_ms además le agregue una división entre 60000 para convertir los datos a minutos, para que se volviera más intuitivo.

Resalta que para ambas la cantidad de diferentes duraciones y tempos es casi igual a la cantidad de registros, y esto es natural ya que pequeñas variaciones son contadas como únicas, y esto no nos dice mucho.

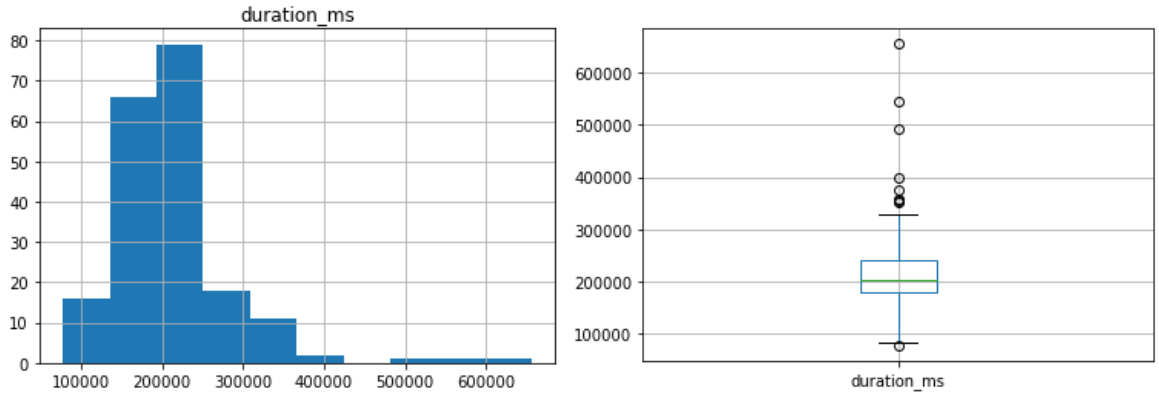
Para la duración vemos que a pesar de que entre la mínima y la máxima hay un gran salto, la media y la mediana nos dicen que la mayoría de las canciones registradas están alrededor de los 3 minutos y medio, y la desviación estándar no es tan grande, por lo que en general parece que esta persona tiene más canciones cortas en su biblioteca.

En las estadísticas del tempo también podemos sacar varias conclusiones. Vemos que hay un gran rango, yendo desde 60 bpms a 180, lo cual indicaría música de diferentes géneros musicales. Sin embargo de la media y la mediana vemos que la mayoría se encuentra alrededor de 120 bpms, con una desviación estándar de casi 30 bpms. La mayoría de la música pop según lo que se encuentra alrededor de los 100-130 bpms y el rock entre 110 y 140, así que tal vez esta persona escuche en su mayoría estos géneros.

Parte 2. Mapas de calor y boxplots

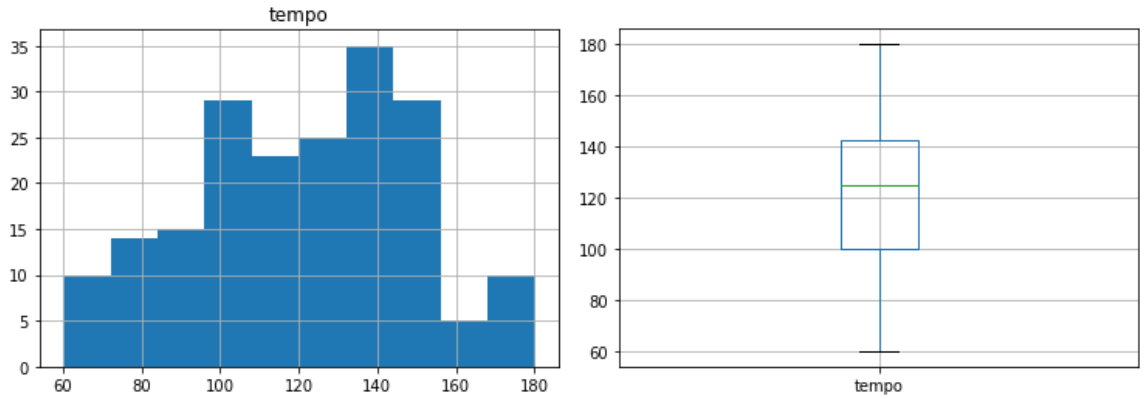
Una vez más escogí duration_ms y tempo, para continuar con el análisis del punto pasado, y para los mapas de calor me enfoque en la variable liked, porque pensé que con esta se verían cosas más interesantes.

En las instrucciones se mencionó que revisáramos los outliers para ver si tenía sentido en el contexto en el que se encontraban, y que si no tenían sentido los quitáramos. En este caso se observó que en duration_ms había varios outliers, y esto se ve tanto en el histograma como en el gráfico de cajas y bigotes. Sin embargo su significado aquí es que hay muy pocas canciones que duren mucho, y esto me imagine que podría tener peso en las correlaciones, por lo que decidí no quitarlos



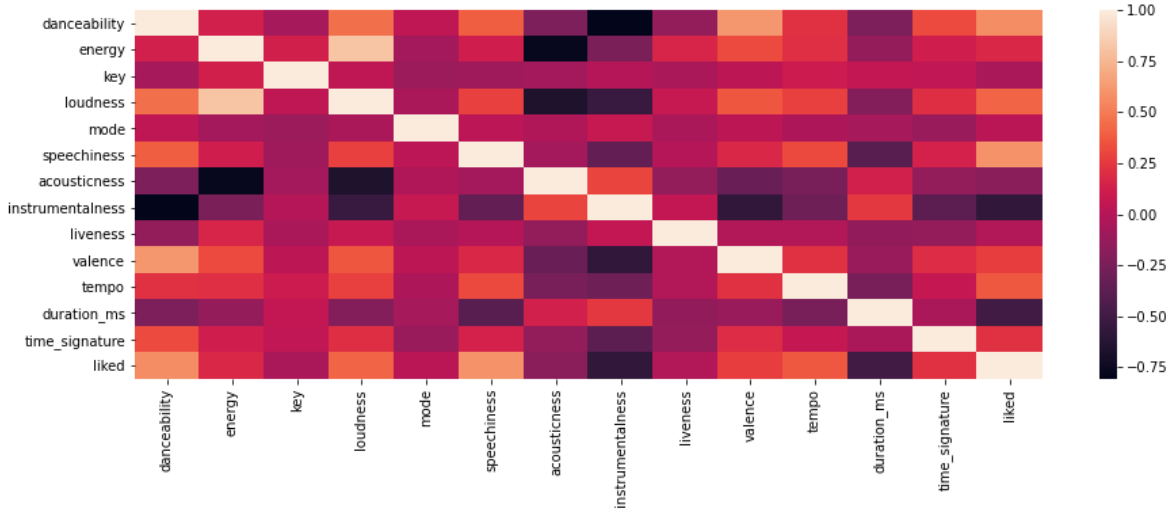
Histograma, Grafico de Cajas y bigotes de "duration_ms"

Por su parte en "tempo" no hubo outliers, por lo que no hubo necesidad de modificar los datos, ya que estos estaban distribuidos de forma más o menos uniforme.

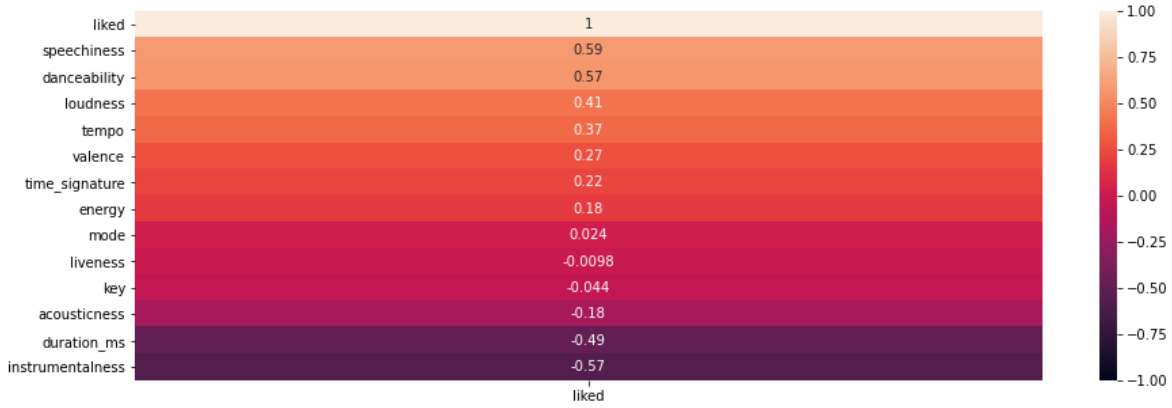


Histograma, Grafico de Cajas y bigotes de "tempo"

Finalmente el mapa de calor. En este se pueden observar varias cosas interesantes. Como sospechaba, al haber pocas canciones largas, la duración esta correlacionada negativamente con la variable “liked”. A partir del mapa de calor podemos concluir que esta persona prefiere canciones cortas, que se puedan bailar, en las que se hable mucho y que no sean instrumentales.



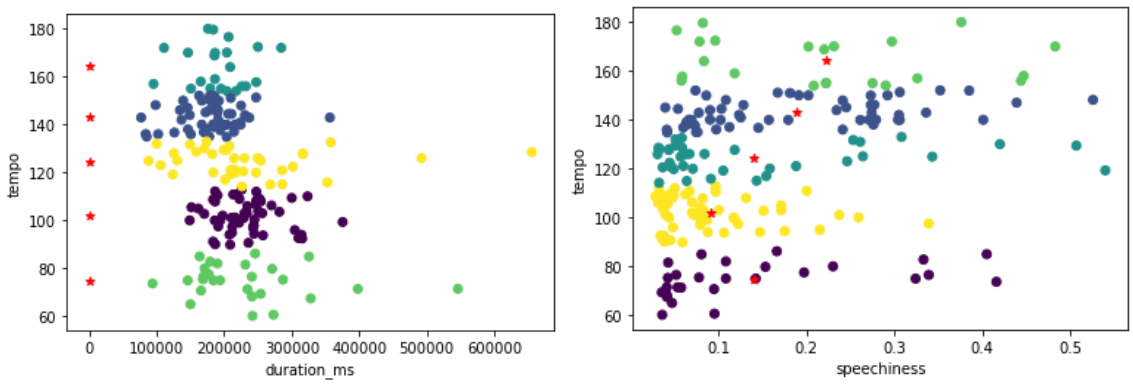
Mapa de calor



Mapa de calor para la columna “liked”

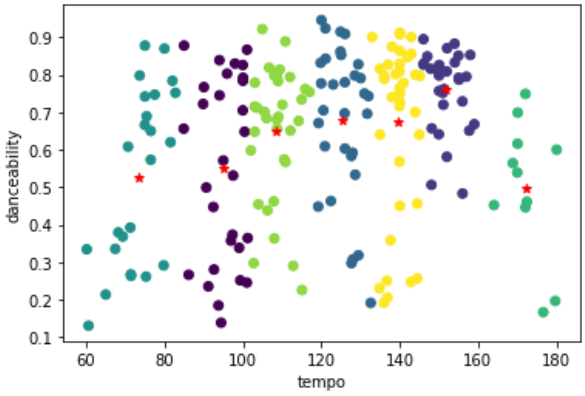
Parte 3. Patrones con K-means

Empecé usando varias combinaciones de tempo con otras variables, ya que esta había sido una de las variables en las que más me había enfocado para las otras actividades y pensé que podría haber correlaciones interesantes de esta con otros géneros. Pero me di cuenta que por alguna razón el algoritmo de K-means prefería agrupar siempre basándose solo en esa variable, ignorando la otra que le diera; lo que daba lugar a franjas en vez de clusters.



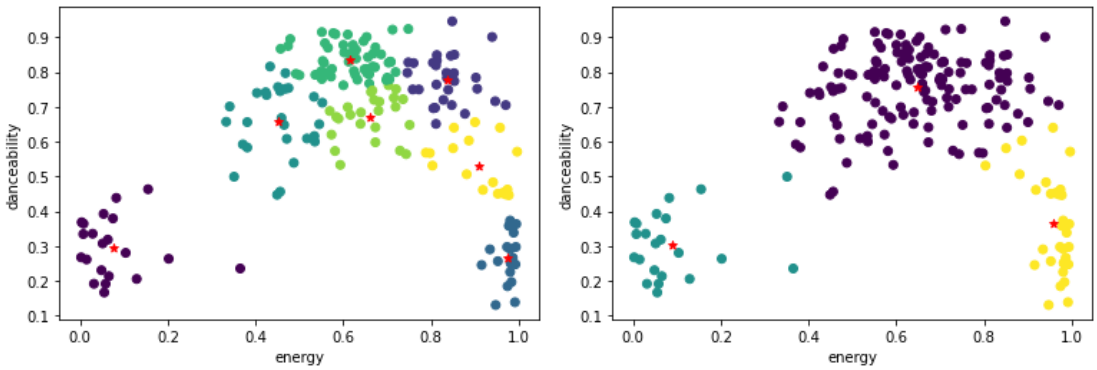
Scatter Plot con coloración a partir de K-means, tempo contra duración y tempo contra speechiness

Pensé que esto se podría deber al valor de K que estaba usando, e hice un intento más dándole un valor de 7, pensando que quizás así dividiría las franjas tomando en cuenta la otra variable. No sucedió esto, y solo conseguí franjas más delgadas.



Intente con otras variables, ahora energy y danceability, y esto resulto en una gráfica más bonita. Pensé que para lo que resultaba quizás eran demasiados centros, y como en la gráfica había 3 grupos de puntos principales, cambié K a 3.

Resultan entonces 2 gráficos, uno con 7 centros que divide el cluster más grande del segundo en otros más pequeños, y otro solo con 3. En ambos se ven 3 grupos principales, uno de canciones con poca energía y que no se pueden bailar, otro con demasiada energía y que no se pueden bailar y finalmente uno donde la energía esta justo al nivel adecuado que además se pueden bailar.



Como ya mencione, lo que varía de uno a otro es que el cluster más grande se divide en otros más pequeños, y esto creo que es informativo acerca de los diferentes géneros que existen, y como esto se ve reflejado en agrupaciones que a simple vista podrían estar cerca.

El elegir cuantos centros usar creo que debe de salir de observar los datos, intentar, ver que pasa y modificar de acuerdo en lo que observas, y creo también que no existe un número ideal, ya que depende de que es lo que quieras sacar de los datos. El usar un número más alto evidentemente ayuda a dividir en grupos más pequeños, pero corres el riesgo de perder la generalidad y con ella información que podría ser útil.

En el plot con 3 ks los 3 centros están relativamente alejados, aunque los de energía más alta están más cerca entre sí. En el de 7 ks todos los centros que aparecen dentro del cluster más grande están bastante cerca entre sí.

Si hubiera muchos outliers esto afectaría la posición de los centros, y esto a su vez haría más difícil la clasificación, o si hubiera suficientes centros tal vez uno quedaría asignado a los outliers si fueran suficientes.

Creo que más allá de decir cosas de los datos basándote en los centros es decir cosas basándote en la posición relativa entre los centros, ya que los centros en si tú decides cuantos son. Quizás sería diferente en otros datos más difíciles de clasificar o con un algoritmo que decidiera la cantidad optima de centros, pero en este caso fue que vi que había 3 grupos principales y dije, que haya 3 centros, o: quisiera que clasificara el cluster más grande, y escogí 7 centros.

Bibliografía

All Music Genres And Their Typical BPMs - 2022 UPDATED. (2021, 31 diciembre). Gemtracks Beats. Recuperado 11 de enero de 2022, de https://www.gemtracks.com/guides/view.php?title=music-genres-and-their-typical-bpms&id=823

Spotify Recommendation. (2021, 27 julio). Kaggle. Recuperado 11 de enero de 2022, de https://www.kaggle.com/bricevergnou/spotify-recommendation