Actividad 2 – Reporte

Este código se desarrolló usando las librerías CSV y Pandas, y 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 ¿Qué es? ¿De dónde se obtuvo? ¿Qué representa?

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.

Cantidad de datos que tienes, las variables que contiene cada vector de datos y el tipo de variables.

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.

Para las dos variables que escogiste:
Los rangos de las variables que escogiste
Basándote en la media, mediana y desviación estándar de cada variable,
¿Qué conclusiones o asunciones puedes obtener de los datos?
¿Parecen muy dispersos?
¿Alcanzas a vislumbrar algún patrón? ¿Parecen relacionados?

```
Estadisticas de la columna 'duration_ms'
hay 193 diferentes duraciones
la maxima fue 10.920216666666667
la minima fue 1.2867166666666667
la media fue 3.556815555555555
la mediana fue 3.4
la desviacion estandar fue de 1.2025398810702026

Estadisticas de la columna 'tempo'
hay 194 diferentes tempos
la maxima fue 180.036
la minima fue 60.171
la media fue 121.08617435897445
la mediana fue 124.896
la desviacion estandar fue de 28.084828828756923
```

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.