Memoria

El objetivo del proyecto es realizar un análisis univariante y multivariante de una dataset extraído de Kaggle de la API de Spotify

Introducción

En este proyecto realizaremos un análisis exploratorio de datos utilizando Python, utilizando información de la biblioteca de datos Kagel sobre Spotify, que incluye información sobre canciones,artistas y características de las canciones.

Objetivo

Estudiar el comportamiento de la popularidad y el impacto del resto de variables para identificar conclusiones relevantes y significativas a través del análisis estadístico.

Hipótesis

Las hiòtesis que buscamos validar en el análisis exploratorio son las siguientes:

1. ¿Cuáles son los géneros musicales más populares?
2. ¿Las características musicales de una canción como la acústica, bailabilidad, tempo, etc, influyen significativamente en la popularidad?
3. ¿Qué característica músical impacta más en la popularidad?
4. ¿Qué relación tiene la duración de una canción con la popularidad?
5. ¿Las canciones más populares tienen alta bailabilidad?
6. ¿Las canciones acústicas son populares?
7. ¿Las canciones en vivo son populares?
8. ¿Qué factores explican la popularidad de las canciones?

Importación de librerías

La librería que vamos a usar para el análisis son las siguientes:

- import pandas as pd
- import numpy as np
- import matplotlib.pyplot as plt
- import seaborn as sns
- from scipy.stats import normaltest
- from scipy.stats import stats
- from scipy.stats import shapiro
- from scipy.stats import spearmanr
- from scipy.stats import chi2_contingency
- import warnings
- warnings.filterwarnings('ignore')

Importación de dataset

El dataset a importar se encuentra en formato csv.

Análisis de datos

Para obtener información general del dataset, usamos info(), describe(), head() y shape

De las 18 columnas, se observan:

-12 columnas:'acousticness', 'danceability', 'energy', 'instrumentalness', 'key','liveness', 'loudness', 'mode', 'speechiness', 'tempo', 'time_signature' y 'valence' corresponden a características musicales y las 6 restantes aportan otro tipo de información

-11 columnas numéricas y 7 categóricas.

Usando .isnull().sum() vemos que no hay nulos ni NaN

Iniciamos el análisis explorando el comportamiento de la variable genre, e identificamos que existen 2 valores únicos similares.Estandarizamos los géneros Children's Music y Children’s Music para dejarlo en adelante como Children Music.

Convertimos el campo duration_ms de milisegundos a minutos(en decimales) y lo llamaremos duration_min.

Análisis gráfico univariante

Para el análisis univariante, usamos un histograma para tener una idea del comportamiento de las variables en su conjunto.

Para profundizar en el comportamiento de algunas variables relevantes, hacemos uso de los ordenamientos o rankings en variables categóricas como género y artistas

Análisis bivariante

Para entender el comportamiento entre variables usamos el análisis de correlación y lo representamos a través de un mapa de calor.

Ya que una de nuestras variables relevantes en el EDA es la popularidad, podemos identificar que acousticness, energy, danceability y loudness son las variables que muestran mayor correlación. En adelante centraremos el análisis sobre estas variables y sus relación con la popularidad

También utilizamos los rankings usando 2 variables como: los 10 artistas más populares(en promedio); las 10 canciones más populares(promedio); los 10 géneros más populares(en promedio); etc

Creando un dataframe de los 5 géneros más populares para identificar patrones o comportamientos que nos permitan establecer una relación con la popularidad de sus canciones.

En base a los highlights, podemos focalizar el análisis en los géneros que mayor impacto tienen en la popularidad de las canciones.

Realizamos un análisis de correlación de géneros top vs Popularidad

Además representamos un análisis de la duración y género para identificar la presencia de outliers o valores atípicos

Por último, realizamos un análisis de normalidad para todas las variables así como de Spearmean para validar la significancia de las correlaciones y lo representamos de forma gráfica.