Grupo R5
Calle 127b Bis 49 - 38, Bogotá - Colombia

PRESENTACIÓN PRUEBA DATA QUALITY ENGINEER JUNIOR

FECHA DEL INFORME	NOMBRE DEL PROYECTO	PREPARADO POR
09/01/2023	Prueba	Juan Manuel García

PASOS PARA LA SOLUCIÓN DE LA PRUEBA

1- Se procede a leer toda la prueba para estructurar la mejor solución.
2- Se descarga la información necesaria para empezar a realizar el análisis y estado actual de la información.
3- Se validan paquetes y software necesario para no presentar problemas a la hora de ejecutar los scripts en Python y las diferentes librerías.
4- Se realizan pruebas iniciales para validar resultados y se evidencia que existe un problema a la hora de transformar los datos, ya que la información contenida em json, esta segmentada de una manera diferente, a la del formato que están solicitando, se realiza indagación con diferentes herramientas web, y se descubre que se debe separar la información en cuatro grupos importantes, artistas, albums, tracks, audio_features, una vez separada la información y extraída, se evidencia que el grupo de tracks, es el csv objetivo de evaluación y se renombra a data set.
5- Se procede a realizar el respectivo análisis de calidad de los datos con otro script de py y se procede a realizar un informe de los hallazgos.
6- A continuación se encuentra el desarrollo del análisis del punto dos.
7- Finalmente, se determina que en el análisis del conjunto de datos, se identificaron varios errores como datos vacíos, datos duplicados. La estadística descriptiva reveló una media de aproximadamente 236 segundos. Se detectó al menos un registro con duración de pista negativa, indicando posibles errores en la recopilación de datos. Además, 'explicit' y 'album_total_tracks' presentan tipos de datos no numéricos presentando datos no válidos. Estos hallazgos apuntan a la necesidad de una limpieza y normalización del conjunto de datos para mejorar la calidad y confiabilidad de la información.

ANALISIS DE LOS DATOS OBJETIVO

Estadisticas descriptivas	disc_numbe	duration_m	track_numbe	audio_features_valen	audio_features_temp	audio_features_time_signatu
count	539	539	539	539	538	538
mean	1.03154	236003.725	11.280148	0.39841	122.362639	3.986989
std	0.174934	55019.871	7.965621	0.199409	30.485522	0.197323
min	1	-223093	1	0.0374	68.097	3
25%	1	209486.5	5	0.23	96.6845	4
50%	1	233626	10	0.386	119.0005	4
75%	1	259045.5	15	0.535	143.939	4
max	2	613026	46	0.943	208.918	5

Valores Nulos	
disc_number	0
duration_ms	0
explicit	0
track_number	0
track_popularity	0
track_id	8
track_name	7
audio_features_danceability	2
audio_features_energy	2
audio_features_key	1
audio_features_loudness	2
audio_features_mode	0
audio_features_speechiness	1
audio_features_acousticness	1
audio_features_instrumentalness	0
audio_features_liveness	1
audio_features_valence	0
audio_features_tempo	1
audio_features_id	0
audio_features_time_signature	1
album_id	0
album_name	62
album_release_date	0
album_total_tracks	0

Valores Duplicados	18

Valores Nulos:
Algunas columnas tienen valores nulos, especialmente 'track_name' y 'album_name'.

Valores Duplicados:

Se detectaron 18 registros duplicados en el conjunto de datos.
Estadísticas Descriptivas:

Tipos de Datos:

La mayoría son numéricos

Discos con Errores:

Un registro tiene un valor negativo para 'duration_ms'.