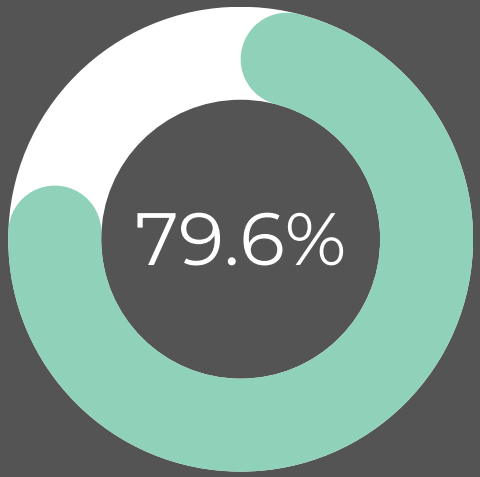
Reporte: Data Quality Assessment

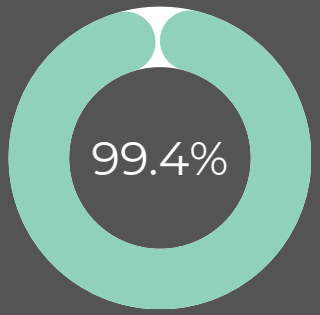
Este reporte de calidad de datos se hizo evaluando cuidadosamente las seis dimensiones primarias según DAMA UK Working Group, usando notebooks de Python junto a librerías como Pandas y Matplotlib.

Repositorio | Análisis | Video | Cristian Gabriel Torres



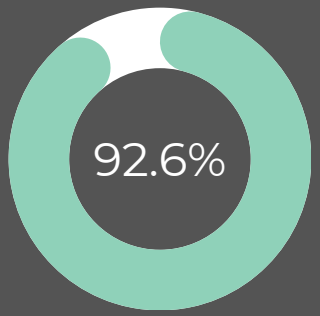
Puntuación Final

Se analizaron meticulosamente las seis dimensiones y una se le atribuyó una puntuación individual para finalmente llegar a una puntuación final del 79.63%.
Si bien la puntuación final es bastante buena, quisiera destacar que se encontraron muchos problemas de distintos tipos, lo cual hace que dependiendo del uso que se le de a la data, puedan generar varios problemas puntuales y difíciles de solucionar.



Completeness

Completeness:
Se identificaron áreas donde faltan datos.
En total 89 registros nulos , lo que representan un 99.39% de la data.

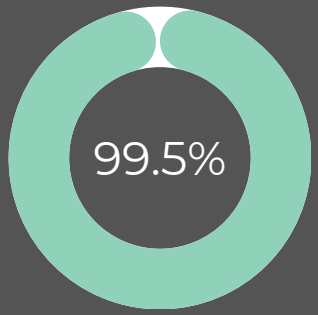


Uniqueness

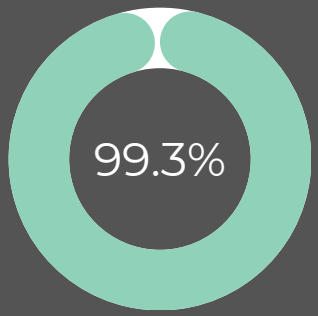
Uniqueness:
Si bien hay información redundante, como el nombre del artista, su popularidad etc; como datos duplicados se encontraron una columna, un álbum y una canción que representan un 92.59% de la data.

Timeliness

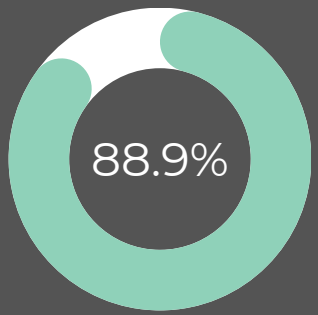
Timelines:
Esta dimensión se exploró desde varios puntos de vista, se la comparó con información actualizada realizando una nueva consulta a la API de Spotify, se evaluó el tiempo que demora Spotify en actualizar la data, e incluso se verificó la precisión que especifica Spotify mediante otra consulta a su API.



Validity



Accuracy



Consistency

Validity:

Aquí se prestó especial atención, ya que se encontraron todo tipo de datos no válidos.

Se encontraron valores fuera de rango, como la instrumentalidad, fecha de lanzamiento, duración de la pista, etc.

Y otros como Thirteen en lugar de 13, o valores como Si y No mezclados con True y False, que no solo son valores inválidos si no que generan un cambio en el tipo de dato de la variable.

Fueron pocos registros, pero muchos de distinta naturaleza.

Accuracy:

En esta dimensión es de particular importancia el conocimiento de la base de datos, por lo que se estudiaron bien todas las variables para comprender su significado.

Se encontraron inconsistencias en la etiqueta de contenido explícito, en la duración y en la popularidad de las canciones.

Consistency:

Finalmente, analizando las variables existentes y mediante otra consulta a la API de Spotify se definieron tres columnas inconsistentes, debido a errores analizados anteriormente.

Una columna que debería ser booleana y dos numéricas, se interpretan como strings debido a la inconsistencia de sus datos.