**Spotify Tracks Data Analysis - Final Report**

This report is a breakdown of my entire data cleaning and exploratory analysis process on the Spotify dataset. The goal was to understand different aspects of the dataset, fix inconsistencies, and gain insights into song characteristics like danceability, energy, and popularity.

**Data Cleaning**

**Handling Missing Values**

When I first checked for missing values, I found that some numerical columns had gaps, which could have affected my analysis. Instead of dropping these rows (which would remove potentially useful data), I filled them with the **median**. Median is better than mean because it’s not affected by extreme values. For categorical columns, I filled missing values with the **most frequent category (mode)** to keep things consistent.

**Standardizing Categorical Data**

One thing I noticed was that category names were messy—some had extra spaces, and some were capitalized differently. To fix this, I converted all text to **lowercase and stripped extra spaces** so that Pop, pop , and POP wouldn’t be treated as different genres. Simple fix, but without it, my analysis would have been misleading.

**Dealing with Outliers**

At first, I tried the **IQR method** (Interquartile Range) to detect outliers, but that didn’t work well for some columns, especially duration\_ms and instrumentalness, because they were heavily skewed. IQR flagged too many values as outliers when, in reality, they were just naturally large numbers.

So, instead of blindly removing them, I used the **percentile-based approach** (keeping everything within the 1st to 99th percentile), which worked much better. This way, I only removed extreme cases without messing up the real data.

**Exploratory Data Analysis (EDA)**

**Univariate Analysis**

I started by plotting **histograms** to see how numerical features were distributed. danceability had a fairly even spread, but instrumentalness had a huge spike at 0, meaning most songs had no instrumental component at all. duration\_ms was another issue—it was right-skewed, meaning there were some really long tracks (probably podcasts or extended mixes).

Boxplots also helped in spotting outliers visually. The energy column had a few extreme points, but overall, most songs fell within a reasonable range.

**Bivariate Analysis**

One of the most interesting findings was the **strong correlation between energy and loudness**. This makes sense because louder songs tend to feel more energetic. A scatter plot of these two variables showed a clear upward trend.

Another thing I checked was **popularity vs. explicit songs**, but surprisingly, there was no clear relationship—both explicit and non-explicit songs had a wide range of popularity scores.

**Genre-Based Insights**

I grouped songs by genre and calculated the average danceability, energy, and popularity. As expected, **pop and electronic songs had higher danceability scores**, while rock and classical were lower. However, popularity was all over the place—some genres had both highly popular and unpopular songs, showing that popularity isn’t just about genre.

**Key Insights & What I Learned**

* **Outlier handling isn’t one-size-fits-all.** The IQR method is great for normally distributed data, but for skewed data, **percentile capping** worked much better.
* **Correlation does not imply causation.** Energy and loudness were strongly correlated, but that doesn’t mean loudness causes energy—it just means the two tend to increase together.
* **Data cleaning is way more important than I thought.** Small inconsistencies in formatting (like genre names with spaces) could have messed up my entire analysis.
* **Danceability doesn’t directly determine popularity.** While pop and electronic songs tend to have high danceability, they aren’t always the most popular. Other factors, like marketing and artist reach, probably play a bigger role.