* Zoek, bijvoorbeeld via [https://datasetsearch.research.google.com,Links to an external site.](https://datasetsearch.research.google.com,/) een dataset uit waarmee je de volgende vragen kunt beantwoorden. Je hoeft de dataset NIET in te leveren, maar je inzending moet wel een link naar de bron bevatten.

* Beschrijf je dataset in minimaal een half A4: geef descriptives (zoals gemiddelde en spreiding).
* Beoordeel de dataset aan de hand van de in het hoorcollege gegeven criteria (accuracy, precision, consistency, completeness, integrity en uniqueness; intent hoeft niet, want is niet voor alle datasets gegeven.)
* Geef een onderzoeksvraag die met de betreffende data te beantwoorden is. Je hoeft GEEN daadwerkelijk antwoord te geven.

<https://www.kaggle.com/datasets/nelgiriyewithana/top-spotify-songs-2023/data>

Voor deze opdracht heb ik een dataset gekozen waar de meest gestreamde nummers op Spotify van 2023 in staan. De dataset geeft naast de standaard elementen als naam, artiesten en het tempo ook geavanceerde data als de dansbaarheid en de hoeveelheid energie in het nummer. Ik heb voor deze dataset gekozen omdat het een breder inzicht kan geven in de reden dat een nummer populair wordt. Met deze dataset kan bijvoorbeeld worden gekeken of de populariteit van een nummer te maken heeft met de dansbaarheid van een nummer, of dat het met andere aspecten te maken heeft.

De dataset bevat 24 kolommen, en per kolom wordt aangegeven wat deze inhoud en wat de relevante descriptives hiervan zijn.

**track\_name**

Dit is de naam van het nummer. In deze kolom zijn 943 unieke waarden te vinden, wat aangeeft dat er 943 nummers in de dataset zitten.

**artist(s)\_name**

De naam van de artiest(en) van het nummer. Voor deze kolom zijn er een aantal relevante descriptives, aangezien het om een categorisch variabele gaat. Ten eerste het aantal unieke waarden, wat in dit geval het aantal unieke artiesten betreft. In deze dataset zijn 645 unieke artiesten te vinden. De modus van deze kolom, ook wel de artiest die het meest voorkwam is Taylor Swift, die 34 keer voorkomt in de dataset.

**artist\_count:**

Het aantal artiesten die betrokken zijn met het nummer. In figuur 1 is de spreiding van deze variabele te zien. De grootste hoeveelheid van de nummers heeft 1-2 artiesten, met enkele uitlopers die tot de 8 artiesten gaan. De gemiddelde hoeveelheid artiesten per nummer is 1.55.

**released\_year**

Het jaar waain het nummer is uitgebracht. De modus van deze kolom, ook wel het jaar dat het meeste voorkwam in de dataset is 2022. In figuur 2 is de verdeling van de nummers per jaartal te zien, het grootste gedeelte van de nummers zijn uitgebracht in de jaren 2021, 2022 en 2023.

**released\_month**

De maand waarin het nummer is uitgebracht. De modus van deze kolom, ook wel de maand waarin het meeste is uitgebracht is de maand januari. In figuur 3 is duidelijk te zien dat er twee pieken zijn, namelijk januari en mei.

**released\_day**

De dag waarop het nummer is uitgebracht. De modus van deze kolom, ook wel de dag waarop het meeste is uitgebracht is de eerste dag van de maand . In figuur 4 is te zien wat de spreiding van de dagen is, waarbij de eerste dag van de maand de enigste piek is.

**in\_spotify\_playlists:**

Het aantal Spotify afspeellijsten waarin het nummer is opgenomen. De gemiddelde hoeveelheid afspeellijsten per nummer is 5200, maar in figuur 5 is te zien dat het grootste gedeelte tussen de 875 en 5500 afspeellijsten zit.

**in\_spotify\_charts**

De rang van het nummer in de Spotify hitlijsten. De gemiddelde positie van de nummers is 12, en in figuur 6 is te zien dat de verdeling van de nummers al begint bij 1. Het grootste gedeelte van de nummers heeft een rank tussen 1 en 16.

**streams**

Het totaal aantal streams op Spotify. De gemiddelde hoeveelheid streams van de nummers is 514,137,400. In figuur 7 is te zien dat de grootste hoeveelheid van de streams tussen 141,636,200 en de 673,869,000 ligt. De laagste hoeveelheid streams is 2762 en de hoogste hoeveelheid is 3,703,895,000

**in\_apple\_playlists:**

Het aantal Apple Music afspeellijsten waarin het nummer is opgenomen. De gemiddelde hoeveelheid afspeellijsten per nummer is 67, maar in figuur 8 is te zien dat het grootste gedeelte tussen de 13 en 88 afspeellijsten zit.

**in\_apple\_charts**

De rang van het nummer in de Apple Music hitlijsten. De gemiddelde positie van de nummers is 51, en in figuur 9 is te zien dat de verdeling van de nummers al begint bij 1. Het grootste gedeelte van de nummers heeft een rank tussen 1 en 87.

**bpm**

Het aantal ‘beats per minuut’ van het nummer. Dit geeft aan wat het tempo van het nummer is.

**key**

De toonsoort van het nummer.

**mode**

Modus van het nummer (majeur of mineur).

danceability\_%: percentage dat aangeeft hoe geschikt het nummer is om op te dansen.

valence\_%: percentage van de positiviteit van de muzikale inhoud van het nummer.

energy\_%: hoeveelheid energie in het nummer.

acousticness\_%: hoeveelheid akoestisch geluid in het nummer.

instrumentalness\_%: hoeveelheid instrumentele inhoud in het nummer.

liveness\_%: hoeveelheid live-elementen in het nummer.

speechiness\_%: hoeveelheid gesproken woorden in het nummer.

Na het landurig bestuderen van de dataset en het schrijven de descriptives is een onderzoeksonderwerp wat direct ter gedachte komt ….

Bij dit onderwerp past de volgende onderzoeksvraag: …