Méthodologie

* Approche méthodologique
  + Comment analyser les concerts en Suisse ?
    - Analyse quantitative pas encore réalisée
  + Quelles données utiliser ?
    - Données des pratiques spectatorielles difficile d’accès (cf. SOTA : OFSP)
    - Données des événements disponibles et exploitables
* Récupération des données (introduction + récupération + traitement)
  + Songkick
  + Spotify
  + Wikidata
* Méthodes d’analyse
  + Calcul mobilité
  + Calcul fréquence
  + Visualisation de données
    - Quels outils de visualisation ?
      * Distribution géographique
      * Distribution des genres
      * Caractéristique des lieux
      * Similarité des lieux (cluster K-means)
      * Mobilité des artistes
      * Activité des artistes
    - Public visé ?
      * Deux options de visualisation : simple et avancé
      * Simple : public non familier avec termes, paramètres. Facilité d’utilisation, didactique
      * Avancé : paramétrable, complexe, plus d’utilisation des données
    - Difficultés ?
    - => chapitre résultat pour discuter du résultat de la visualisation de données en fonction de ces critères
* Evaluer et justifier les choix méthodologiques
  + Fiabilité des données récupérées (check Romandie, Spotify matches similarity-sample)
  + Pertinence des données (limitation des données)

Cette question permet d’aborder plusieurs axes à l’aide de la visualisation de données : à quoi faut-il être attentif ? Quels sont les difficultés ou à l’inverse les facilités rencontrées ? Pour quels usages cette visualisation de données est-elle conçue ? Comment l’interface doit-elle être pensée en fonction de quel public ? Quelles sont les limites des données utilisées ?

Dans le prolongement de cette question principale, il est possible d’aborder également des questions secondaires relatives aux artistes (mobilité ? récurrence ? répartition géographique ?) et aux salles de concerts (diversité de l’offre ? concurrence entre les salles ? densité de population ?).

Songkick

* Introduction
  + Créé en 2007
  + Recensement des concerts partout dans le monde
  + 6 millions événements, 15 millions d’utilisateurs
    - songkick.com
  + Fonctionnement : crawl de plusieurs sources + partenariats avec sociétés de vente de billets
    - CEO Ian Hogarth. "What we've done is crawl and cull this information from thousands of sites all across the web, and stitched it together enough to make a giant database of live musical history."
    - To get the ball rolling on this partially crowdsourced resource, Songkick staff added more than a million shows to the database, with new show information coming from agreements with 29 ticketing agencies
    - vanbuskirk\_songkick\_2009
  + Business model : commission billets vendus sur site
    - Now for the money question \x84 as in, how does Songkick intend to make any? Hogarth told us Songkick receives fees from ticketing agencies when someone buys a ticket through the site
    - vanbuskirk\_songkick\_2009
  + Check Romandie 2019 : distinction entre concerts et soirées, très fiable pour les headlines, bonne fiabilité pour premières parties
* Récupération
  + Méthode : web scraping plutôt qu’API parce que plus flexible
  + 1 : identifier salles se trouvant en Suisse
  + 2 : récupérer toutes les salles en Suisse
  + 3 : récupérer tous les concerts dans ces salles
  + 4 : récupérer tous les artistes à ces concerts
  + Chaque élément identifié par son id songkick
  + Résultat : 3 jeux de données : salles, concerts, artistes
  + Salles : nom, code postal, rue, localité, pays, coordonnées
  + Concerts : nom, type, salle, artistes, date (date début uniquement)
    - = élément central qui fait le lien entre salles et artistes
  + Artistes : nom
  + Stats brutes toutes périodes : concerts = 82'011, artistes = 32'231, salles = 7'657
* Traitement
  + Doublon de salles
    - Crawling de songkick automatisé donc pas parfait
    - Doublon identifiable par nom, adresse, coordonnées ou manuellement
    - Modification manuelle des doublons : réunion des doublons sous un seul élément unique choisi en fonction des données les plus complètes (nom, adresse, coordonnées)
  + Salles hors de la Suisse : coordonnées erronées pour des salles véritablement en Suisse
  + Stats doublons:
    - Adresse: 102
    - Coordonnées: 267
    - Nom: 60
    - Manuellement: 21
    - Total: 450
    - Changements effectués : 433 (certains doublons à double, par ex. nom et coordonnées)
    - Hors de Suisse: 6
    - Salles totales: 7'224
  + Un artiste inconnu sans ID : 32'230 artistes

Spotify

* Introduction
  + Service de streaming de musique, 2008
  + 70 millions de morceaux
    - spotify.com
  + Succès grâce à son système de recommandations
    - mood categorization represents a much more natural way of thinking about music; one that is more representative of how people speak about music
    - prey\_knowing\_2018, 46
* Récupération
  + Méthode : API
  + 1 : chercher artistes Spotify correspondant aux artistes Songkick
    - Utiliser fonction de recherche de l’API
    - Trouver similarité entre terme cherché et chaque terme trouvé
    - Garder l’artiste dont le nom a la plus haute similarité avec le nom cherché
    - Si plusieurs artistes, garder celui avec la plus haute popularité
    - Métadonnées : Spotify ID, nom, popularité
  + 2 : pour chaque artiste Spotify, obtenir genres musicaux, nombre d’auditeurs et de followers
  + 3 : pour chaque artiste, obtenir les morceaux les plus populaires (max. 10)
  + 4 : pour chaque morceau, obtenir les métadonnées audio
    - https://developer.spotify.com/documentation/web-api/reference/#/operations/get-audio-features
    - danceability: Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable.
      * 0 <= x <= 1
    - energy : Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy.
      * 0 <= x <= 1
    - key: The key the track is in. Integers map to pitches using standard Pitch Class notation. E.g. 0 = C, 1 = C♯/D♭, 2 = D, and so on. If no key was detected, the value is -1.
    - loudness: The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks. Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typically range between -60 and 0 db.
      * -60 <= x <= 0
    - mode: Mode indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived. Major is represented by 1 and minor is 0.
      * x = 0 || x = 1
    - speechiness: Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks.
      * 0 <= x <= 1
    - acousticness: A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic.
      * 0 <= x <= 1
    - instrumentalness: Predicts whether a track contains no vocals. "Ooh" and "aah" sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly "vocal". The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0.
      * 0 <= x <= 1
    - liveness: Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live.
      * 0 <= x <= 1
    - valence: A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).
      * 0 <= x <= 1
    - tempo: The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration.
    - duration\_ms: The duration of the track in milliseconds.
    - time\_signature: An estimated time signature. The time signature (meter) is a notational convention to specify how many beats are in each bar (or measure). The time signature ranges from 3 to 7 indicating time signatures of "3/4", to "7/4".
  + Stats: 32'230 artistes Songkick, 28'571 artistes Spotify (3'659 sans match), 3744 genres musicaux
* Traitement
  + Doublon
    - Différents artistes Songkick avec le même Spotify ID
    - DB de Songkick référence plusieurs fois le même artiste
      * Ex : groupe lausannois Larytta : LARYTTA et Larytta
    - Garder le doublon dont le nom a la plus haute similarité
    - Stats : 913 artistes redondants (17 triples, 431 doublons), 448 uniques, 28'106 artistes uniques.
  + Vérification
    - Low similarity : similarity < 0.7
    - Nbr low similarity checked: 343
    - Nbr error similarity: 228
    - Ratio error in low similarity artists: 0.66
    - Nbr sample checked: 300
    - Nbr error sample: 5
    - Ratio error in sample: 0.02

Wikidata

* Introduction
  + Base de connaissance libre et collaborative
  + Base centrale pour les données structurées qui alimentent les autres projets de Wikimedia (chaque article de Wikipédia a un équivalent sur wikidata)
  + Chaque élément a : nom, description propriétés, alias
    - Exemple : élément « disco »
      * https://www.wikidata.org/wiki/Q58339
      * Description : genre of dance music and a subculture that emerged in the 1970s from the United States' urban nightlife scene
      * Instance de « musical form » et « music genre »
      * Sous-classe de « dance music »
      * Partie de « African-American music »
  + Nécessaire pour grouper les genres donnés par Spotify
    - it was deemed important to also look directly at the data contained within the largest online, global knowledge-base
    - gagen\_hybrids\_2019, 88
* Récupération
  + Obtenir tous les éléments sur Wikidata instances de « music genre »
    - Genre parent
    - Alias everynoise
    - Alias discogs genre
    - Alias discogs style
  + SPARQL query to fetch data from wikidata graph  
    SELECT ?genre ?genreLabel ?parent\_genre ?parent\_genreLabel ?everynoise\_id ?discogs\_genre\_id ?discogs\_style\_id WHERE {  
     SERVICE wikibase:label { bd:serviceParam wikibase:language ",en". }  
     ?genre wdt:P31 wd:Q188451.  
     OPTIONAL { ?genre wdt:P279 ?parent\_genre. }  
     OPTIONAL { ?genre wdt:P9881 ?everynoise\_id. }  
     OPTIONAL { ?genre wdt:P9218 ?discogs\_genre\_id. }  
     OPTIONAL { ?genre wdt:P9219 ?discogs\_style\_id. }  
    }
* Traitement
  + Trouver le genre parent le plus haut pour chaque élément
    - A genre (level = 1) is a subclass of a top genre (level = n, n > 1), such that parent class of top genre is not an instance of music genre
    - Popular music est trop vague, prendre le genre inférieur (rock music, jazz, etc.)
    - Un genre peut avoir plusieurs genres parents
      * Garder les top genres pour chaque genre
    - Des genres parents différents peuvent avoir le même top genre
      * Garder les top genres qu’une seule fois pour chaque genre
    - Faux top genre : des genres avec plusieurs top genres dont lui-même
      * Rock and roll subclass of rock music
      * Rock and roll subclass of popular music
      * Rock music subclass of popular music
      * 309 faux top genres
    - Des top genres n’apparaissent qu’une seule fois (top genre de lui-même)
      * 1323 unique top genres
    - 4579 genres, 80 top genres
  + Correspondance avec Spotify
    - Correspondance avec nom de Wikidata, nom everynoise, nom discogs
      * 3011 correspondances, 733 sans correspondance
    - Top genre selon les genres similaires
      * Si un genre Spotify n’a pas de correspondance avec Wikidata, prendre la liste de tous les artistes associés à ce genre Spotify. Prendre le top genre de ces artistes le plus fréquent pour l’associer au genre Spotify jusque-là sans correspondance.
      * 626 correspondances, 107 sans correspondance
    - Correspondance manuelle
      * 107 correspondances, 0 sans correspondance
  + Correction de top genre pour Spotify
    - Unique top genre
      * Exemple : argentine punk ne doit pas être un top genre
      * 128 unique top genres manuellement changés en 15 top genres
    - Changement manuel de top genre
      * 48 top genres manuellement changés en 25 top genres

Méthodes d’analyse

* Analyse spatiale et temporelle avec le calcul de la mobilité et d’activité
* Calcul mobilité
  + Mobilité d’un artiste utile pour voir quels sont les artistes qui se déplacent le plus en Suisse
  + Artistes avec un intérêt régional ou national
  + Coefficient de mobilité : nombre de lieux divisés par le nombre de concerts
    - Proche de 0 : très peu mobile
    - 1 : jamais deux concerts au même endroit
  + Coefficient de parcours : distance moyenne entre chaque lieu et leur centroïde divisée par 175
    - 175 = distance moyenne maximale entre des lieux et leur centroïde en Suisse puisque largeur maximale de la Suisse est de 350 km environ.
    - Centroïde d’un nombre fini de points (formule math)
    - 0 : aucune distance parcourue
    - 1 : très mobile et beaucoup de distance
* Calcul activité
  + Activité d’un artiste utile pour voir si sa présence en Suisse s’étend dans le temps ou si elle est concentrée à un moment précis
  + Fenêtre d’activité en Suisse : dates du premier et dernier concert
    - Entre 2010 et 2019
  + Fréquence de concerts : nombre de concerts divisé par nombre de jours de la fenêtre d’activité
    - 0 : aucun concert donné dans la fenêtre d’activité
    - 1 : concert donné tous les jours de la fenêtre d’activité
  + Tournée : période durant laquelle au moins deux concerts ont lieu sans avoir plus de 90 jours entre chaque concert
    - 90 jours : au-delà d’une pause de 3 mois pas considéré comme une tournée (par exemple pause du 15 juin au 15 septembre sans festival estival).
  + Intensité d’une tournée : nombre moyen de concerts dans une tournée divisé par nombre moyen de jours dans une tournée
    - 0 : tournée tranquille
    - > 0.03 : 1 concert par mois
    - > 0.14 : 1 concert par semaine
    - 1 : tournée intense (concert tous les jours)
* Visualisation de données

Justification des choix méthodologiques

* Fiabilité des données récupérées
  + Vérification manuelle
    - Similarité < 0.7 = 338 artistes
    - Similarité >= 0.7 = 27'768 artistes
    - Spotify matches similarity-sample)
    - Nbr low similarity checked: 338
    - Nbr error similarity: 226
    - Ratio error in low similarity artists: 0.67
    - Nbr sample checked: 300
    - Nbr error sample: 5
    - Ratio error in sample: 0.02
* Fiabilité audio features
* Pertinence des données (limitation des données)