Analyse de données avec Python
Spotify

Vous connaissez tous spotify, un des plus importants site d'écoute de musique. Je vous propose dans ce TD/TP d'analyser et de visualiser des éléments du TOP50 de la musique (au niveau mondial) des années 2010 à 2019.

Notre première approche avant de visualiser quoi que ce soit, sera bien évidemment de nous familiariser avec ces données.

Charger et se familiariser avec les données

La première étape est de charger les données, attention elles n'ont pas été sauvegardées en UTF8 (malheureusement) mais au format ISO-8859-1. Vous devez passer ce codage en option lors du chargement du fichier CSV.

- Examinez les premières lignes du fichier
- Donnez le nombre de lignes et de colonnes du fichier
- En utilisant la méthode isnull() de pandas (https://www.w3resource.com/pandas/isnull.php). Analysez le fichier afin de rechercher les valeurs manquantes, vous pouvez également utiliser la méthode count()
- En utilisant la méthode nunique() (https://www.geeksforgeeks.org/python-pandas-series-nunique/), analyser les données afin d'en extraire le nombre de chansons, le nombre d'artistes ainsi que le nombre de chansons. Vous pourrez ainsi constater que plusieurs artistes ont plusieurs chansons différentes dans ce top

J'ai trouvé sauf erreur de ma part qu'il y avait 184 artistes, 584 chansons et 50 genres différents

Visualisation de plusieurs informations

Artistes les plus écoutés

Je vous propose comme première visualisation de visualiser les 10 artistes les plus écoutés (en nombre de titres) dans ce top 50. La méthode value_counts() (https://www.w3resource.com/pandas/series/series-value_counts.php) qui est une méthode extrêmement utilisée permet de compter le nombre d'occurrences d'un même objet.

Affichez ces 10 artistes et en utilisant seaborn utiliser un « barplot » pour les afficher.

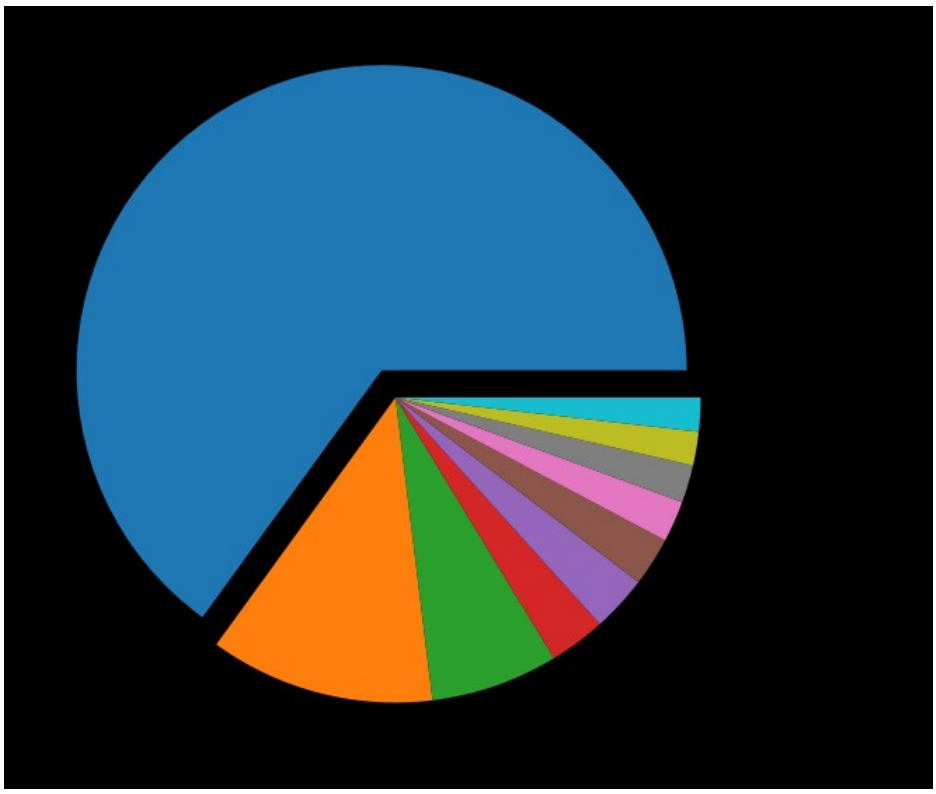
Quel(le) est le/la chanteur(se) la plus écouté(e) ces 10 dernières années, avec combien de titres ?

Pouvez-vous donner les titres les plus écoutés (indice : il y en a 17)

Genres les plus populaires

Je vous propose dans le même ordre d'idée de visualiser les 10 genres les plus populaires. Utiliser un « plot » et une graphique sous forme de camemberts afin de visualiser les titres les plus populaires. Seaborn ne propose pas de graphismes de type camembert mais matplotlib sait le faire (http://www.python-simple.com/python-matplotlib/pie.php)

Vous devriez obtenir quelque chose qui ressemble à la figure ci-dessous :



Les noms des genres sont volontairement absents mais si vous aimez la musique « dance », vous n'êtes pas très très original !

Durée des chansons

D'après les spécialistes de la musique, afin de survivre dans un monde de plus en plus concurrentiel et des plus en plus pressé, les musiques doivent être de plus en plus courtes sinon les auditeurs ne prennent plus le temps de les écouter dans leur entièreté ! Il serait intéressant de vérifier, si au cours de cette décade, la durée de la musique a effectivement diminué en moyenne.

Extraire la durée des chansons sur la décennie
Vérifier on non ce fait avec un simple « lineplot ». Qu'en déduisez-vous ?

Il serait intéressant de vérifier cette durée en la comparant aux durées moyennes des artistes du top10, en profiter pour extraire la chanson la plus courte et la plus longue de chacun de ces artistes.

On peut transformer un index d'une Series ou DataFrame avec la commande Python tolist()

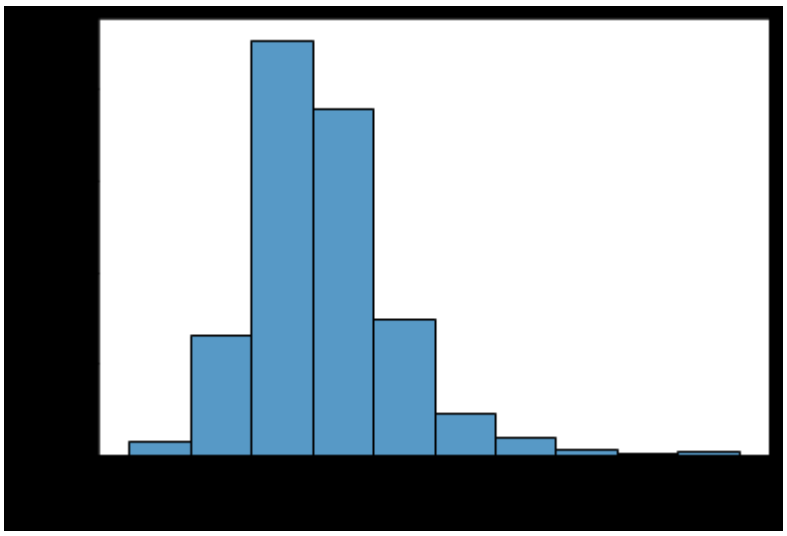
```
Katy Perry',
 'Justin Bieber',
 'Maroon 5',
 'Rihanna',
 'Lady Gaga',
 'Bruno Mars',
 'Shawn Mendes',
 'Pitbull',
 'The Chainsmokers',
 'Ed Sheeran']
```

BPM ses chansons

Il est intéressant également de faire une analyse complémentaire en analysant le BPM moyen des morceaux ces dernières années. Utilisez lineplot pour visualiser l'évolution des BPM des années 2010 à 2019. Qu'en déduisez-vous ?

Autre forme de visualisation

Une visualisation complémentaire de la durée (ou des BPM) de celle que l'on a faite est la visualisation sous un histogramme afin de représenter tous les morceaux et leur durée correspondante (voir le tutoriel de seaborn à cette adresse https://seaborn.pydata.org/tutorial/distributions.html)



Vous devriez obtenir un histogramme semblable à celui ci-dessus. Que pouvez-vous en déduire ?

Dépendance des variables

Nous abordons ici les limites du cours car nous ne faisons que de l'analyse de données mais le cours d'explicabilité ou d'utilisation de ML pour l'analyse de données (analyse par composantes principales, régression des variables ou encore apprentissage automatique) sont des notions que vous verrez en dernière année du cycle ingénieur, mais il est possible assez facilement d'explorer quelques corrélations basiques entre les variables.

Si on examine les données fournies par Spotify, on s'aperçoit qu'il y a de nombreuses colonnes dont il pourrait être intéressant d'analyser les corrélations.

top.columns
→ Index(['Unnamed: 0', 'title', 'artist', 'top genre', 'year', 'bpm', 'nrgy',
`'dnce', 'dB', 'live', 'val', 'dur', 'acous', 'spch', 'pop'],`
`dtype='object')`

Je ne suis pas capable d'expliquer toutes les colonnes mais la colonne dance correspond à la « dansabilité » du morceau. Nous pouvons étudier la corrélation entre la dansabilité du morceau et le BPM.*

Une droite de régression se dessine avec seaborn avec la commande regplot et en lui passant simplement les valeurs en x et en y

Étudiez la corrélation entre les BPM et la dansabilité des morceaux ?

Essayez de trouver les deux variables les plus corrélées