IUT de Montpellier
SAE 2.04 – Correction livrable statistique descriptive

Instructions

Le livrable pour chaque équipe devra être déposé sous la forme :

- D'un document **`pdf`** d'une ou deux pages de texte au maximum précisant les détails, les choix réalisés ainsi que les commentaires. Ces commentaires devront être précis et concis. Les tableaux et graphiques demandés seront présentés dans des annexes.
- D'un document **`ods`** contenant les données, les formules de calculs et les résultats (tableaux, graphiques) présents dans les annexes du document `pdf`. Le classeur sera organisé clairement et utilisera une feuille par question abordée. On utilisera pour cela le modèle donné par le classeur `rendu.ods`. **Tout chiffre ou graphique présenté dans le rapport devra être obtenu par une formule du fichier `ods`.**

Les documents seront nommés `Si-j.pdf` et `Si-j.ods` respectivement, où `Si-j` est le nom du groupe de SAÉ.

Données utilisées : Dans cette partie Statistique, on utilisera les données du classeur `data_stat.ods` qui contient les extractions nécessaires à l'étude. Les données de ce classeur ont déjà subi un pré-traitement de correction des erreurs et peuvent différer des données de la partie BD ou Gestion. On a extrait dans le fichier `data_stat.ods` les titres provenant d'un album dont l'un (au moins) des titres a fait partie du top 200 hebdomadaire des écoutes, entre mars et décembre 2023.

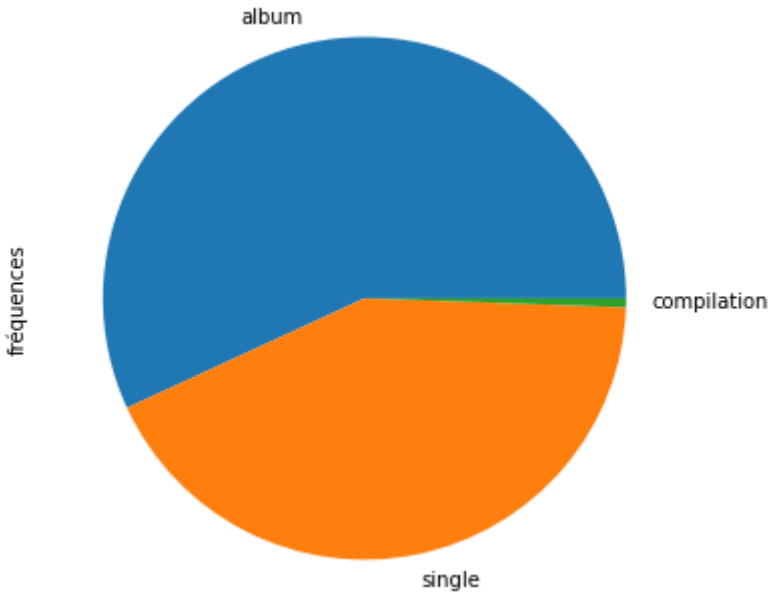
Albums

Variable `type`

Dresser le tableau des effectifs et fréquences du caractère statistique `type` et faire une représentation graphique appropriée.

	Effectifs	fréquences
type		
album	314.0	0.568841
single	235.0	0.425725
compilation	3.0	0.005435
Total	552.0	1.000000

La représentation appropriée est le diagramme en secteurs, plus lisible que le diagramme en bâtons sur cet exemple.



Variable `label`

Donner le mode du caractère `label` pour chacune des années 2021, 2022 et 2023. On précisera le mode de chaque effectif et le nombre de modalités pour les labels de l'année en remplissant le tableau suivant :

	mode	effectif du mode	nb de labels
année			
2021			
2022			
2023			

Quel commentaire sur la diversité des labels observés peut-on faire à la lecture de ce tableau ? Proposer ensuite de compléter ce tableau par une autre colonne qui permettrait de modérer ou d'affiner votre commentaire.

Le tableau demandé est :

	mode	effectif du mode	nb de labels
année			
2021	Universal Music Division Capitol Music France	4	23
2022	Columbia	6	72
2023	Play Two	14	148

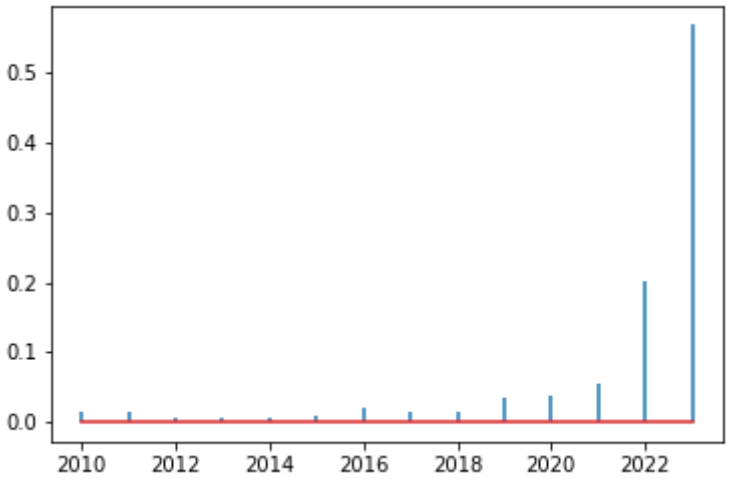
On remarque d'abord que pour les albums écoutés, le label le plus fréquent (le mode) n'est pas stable d'une année sur l'autre. On ne constate donc pas d'hégémonie flagrante d'un label particulier sur ces albums. Ensuite, il semble en première approche que le nombre de labels différents croît très rapidement avec le temps. Cependant, cette conclusion est trop rapide si on ne considère pas aussi la taille de la population d'albums de chaque année : ce nombre plus élevé de labels distincts en 2023 pourrait simplement provenir d'un plus grand nombre d'albums parus en 2023. D'où la pertinence de la question suivante.

Variable `year`

On se restreint aux albums dont l'année de sortie est au moins 2010. Donner le tableau des fréquences de la variable `year` et faire une représentation graphique adaptée. Rajouter également la colonne permettant de lire directement la part des albums récents : par exemple, on souhaite lire dans cette colonne la proportion des albums sortis après 2021 inclus. Commenter ensuite ces résultats.

	effectif	fréquence	fréquence cumulée
year			
2010	7	0.013807	1.000000
2011	7	0.013807	0.986193
2012	3	0.005917	0.972387
2013	3	0.005917	0.966469
2014	3	0.005917	0.960552
2015	5	0.009862	0.954635
2016	10	0.019724	0.944773
2017	8	0.015779	0.925049
2018	7	0.013807	0.909270
2019	17	0.033531	0.895464
2020	19	0.037475	0.861933
2021	28	0.055227	0.824458

	effectif	fréquence	fréquence cumulée
year			
2022	102	0.201183	0.769231
2023	288	0.568047	0.568047



On remarque qu'en 2023, les albums les plus écoutés, et de très loin (57 %), sont des albums de 2023. Sur le graphique adapté qui est le diagramme en bâtons, on constate que les deux dernières années sont largement plus représentées que les autres. Les utilisateurs ont donc écouté principalement des albums récents, de l'année en cours ou de l'année précédente. Cela est confirmé par le tableau des fréquences cumulées décroissantes sur lequel on peut lire que les albums de 2021 et plus représentent 82% des écoutes.

Variable `total_tracks`

On se restreint aux albums parus après 2021 inclus et qui ne sont pas du type 'single'. On s'intéresse au nombre total de pistes de l'album : total_tracks. Compléter le tableau avec les indicateurs de position et de dispersion de la variable total_tracks.

	min	max	Q_1	médiane	Q_3	moyenne	écart type	IQ
total_tracks								

Le tableau demandé est :

	min	max	Q_1	médiane	Q_3	moyenne	écart type	IQ
total_tracks	7	89	14	16	19	17,44	7,91	5

Pistes

On s'intéresse maintenant aux chansons interprétées par des artistes dont l'un au moins est classé dans le genre "`french hip hop`".

Variable `tempo`

On étudie la vitesse moyenne d'exécution de la musique, caractérisée par la variable `tempo` mesurée en bpm (beats per minute). Construire l'histogramme de la variable (considérée comme) continue `tempo` en utilisant les classes données ci-dessous.

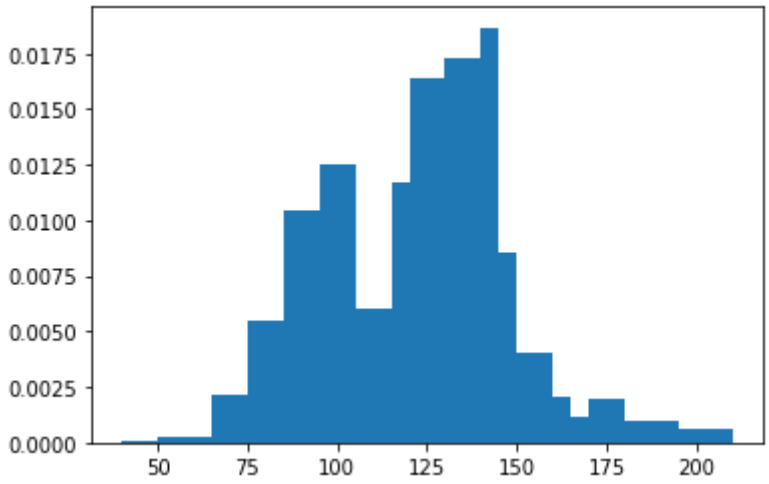
	effectif	fréquence	amplitude	hi
classe				
[40 50[				
[50 65[				
[65 75[				
[75 85[				
[85 95[				
[95 105[				
[105 115[				
[115 120[				
[120 130[				
[130 140[				
[140 145[				
[145 150[				
[150 160[				
[160 165[				
[165 170[				
[170 180[				
[180 195[				
[195 210[				

On détaillera la construction dans le fichier ods en précisant le tableau avec ses formules de calculs. Commenter ce graphique en proposant une classification des musiques en termes de vitesse d'exécution.

On construit le tableau de manière classique :

	effectif	fréquence	amplitude	hi
classe				
[40 50[	1.0	0.000373	10.0	0.000037
[50 65[	9.0	0.003361	15.0	0.000224

	effectif	fréquence	amplitude	hi
classe				
[65 75[	57.0	0.021285	10.0	0.002128
[75 85[	147.0	0.054892	10.0	0.005489
[85 95[	280.0	0.104556	10.0	0.010456
[95 105[	335.0	0.125093	10.0	0.012509
[105 115[	161.0	0.060119	10.0	0.006012
[115 120[	157.0	0.058626	5.0	0.011725
[120 130[	438.0	0.163555	10.0	0.016355
[130 140[	462.0	0.172517	10.0	0.017252
[140 145[	250.0	0.093353	5.0	0.018671
[145 150[	114.0	0.042569	5.0	0.008514
[150 160[	108.0	0.040329	10.0	0.004033
[160 165[	28.0	0.010456	5.0	0.002091
[165 170[	15.0	0.005601	5.0	0.001120
[170 180[	53.0	0.019791	10.0	0.001979
[180 195[	40.0	0.014937	15.0	0.000996
[195 210[	23.0	0.008588	15.0	0.000573
Total	2678	1		



On constate que la distribution est multimodale : les classes [95 ; 105[, [140 ; 145[et [170 ; 180[ont une valeur de la fonction histogramme plus élevée que leurs proches voisines. La classe [105 ; 115[délimite les chansons lentes et les chansons de vitesse intermédiaire. La classe [165 ; 170[délimite les chansons de vitesse intermédiaire et les chansons (très) rapides. Les chansons de vitesse intermédiaire sont les plus représentées. Les chansons rapides sont peu nombreuses.

On pourrait proposer de classer les chansons en trois groupes :

- Les chansons lentes de 40 à 115 (ou 105) bpm
- Les chansons de vitesse intermédiaire de 115 (ou 110) bpm à 165 (ou 170) bpm
- Les chansons rapides de 165 (ou 170) à 210 bpm

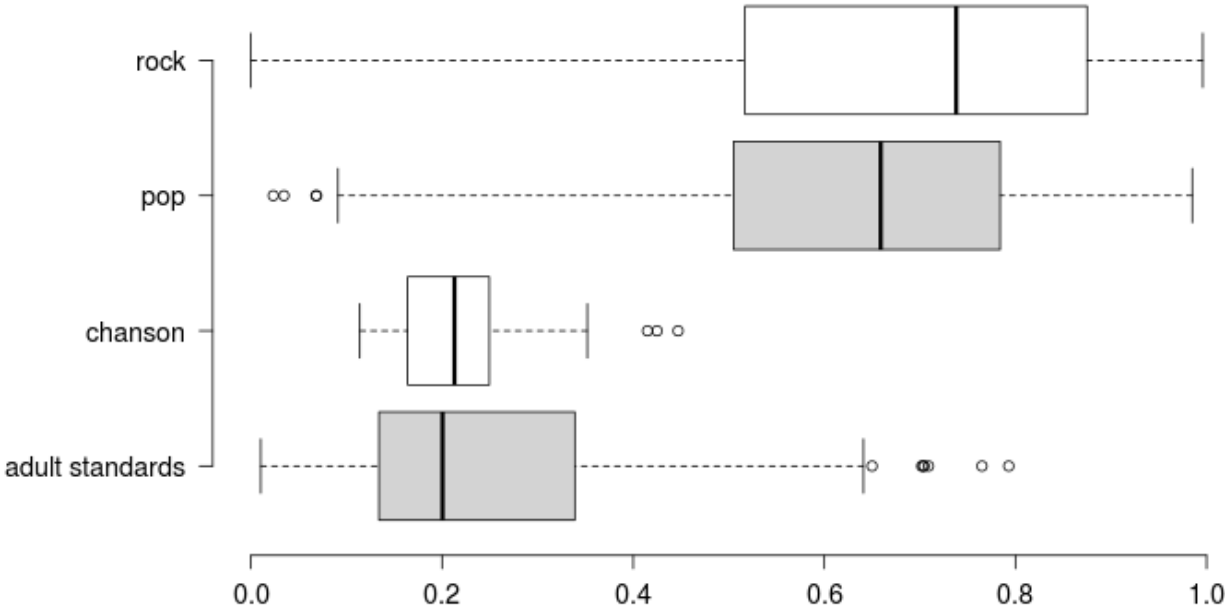
La dernière classe des chansons très rapides est discutable. On peut aussi bien considérer qu'il n'y a que deux groupes différents : les chansons lentes et les chansons rapides, ce dernier groupe autorisant des valeurs très grandes du tempo assez éloigné du tempo central de ce groupe.

Variable `energy`

On se restreint maintenant aux chansons interprétées par des artistes dont l'un au moins est classé dans le genre "pop", "rock", "chanson" ou "adult standards" et on étudie la variable `energy` qui est réel entre 0 et 1 caractérisant l'énergie du fichier audio. Afficher sur un même graphique les 4 boxplots de l'énergie selon le genre musical et commenter ce graphique. Donner ensuite les indicateurs de position et de dispersion de l'énergie pour chacun des genres musicaux et utiliser ces chiffres pour confirmer (ou pas) votre lecture du boxplot en complétant le tableau :

	min	max	moyenne	Q1	mediane	Q3	écart type	variance	IQ
name_genre									
adult standards									
chanson									
pop									
rock									

Le graphique est le suivant :



On constate que les genres musicaux "pop" et "rock" ont des valeurs de l'énergie bien plus élevées que les deux autres genres. Cependant, leur variabilité est aussi beaucoup plus importante. En particulier, le genre "rock" a les valeurs les plus élevées pour l'énergie mais aussi la plus grande variabilité. Une conséquence est qu'aucune valeur particulière ne peut être considérée comme exceptionnelle (pas de cercle sur le graphique). À l'inverse, le genre musical "chanson" est très homogène. Il serait bon de vérifier si les trois titres considérés comme particulièrement "énergiques" (cercles sur le graphique) sont à leur place dans cette catégorie.

Le tableau des indicateurs est :

	min	max	moyenne	Q1	mediane	Q3	écart type	variance	IQ
name_genre									
adult standards	0.0103	0.793	0.256443	0.134	0.200	0.339	0.164647	0.027109	0.205
chanson	0.1140	0.447	0.221098	0.163	0.213	0.252	0.078643	0.006185	0.089
pop	0.0234	0.985	0.629256	0.505	0.659	0.784	0.197548	0.039025	0.279
rock	0.0000	0.996	0.679515	0.517	0.736	0.875	0.243564	0.059324	0.358

On peut se demander aussi s'il n'y a pas d'erreur pour les pistes qui ont 0 pour énergie. Par ailleurs on retrouve bien les résultats précédents en lisant les moyennes et médianes, puis les écart-type par genre. À noter la dissymétrie en faveur des faibles valeurs de l'énergie pour le genre "rock" qui entraîne une moyenne notablement plus faible que la médiane.