IUT de Montpellier
SAE 2.04 – Exploitation de base de données

Description et contexte

Contexte

On s'intéresse à une application d'écoute de musique en ligne, du même type que YouTube Music, Deezer, Spotify ou AllForMusic. L'application propose plusieurs fonctionnalités à ses utilisateurs. Elle collecte aussi des données sur leurs habitudes et préférences. Ces données sont régulièrement analysées afin d'adapter la stratégie de l'entreprise aux besoins des utilisateurs.

Dans cette SAÉ, on se propose de construire une base de données à partir d'un extrait des informations enregistrées en 2023. On effectuera ensuite une analyse statistique qu'on mettra à disposition des décideurs.

Organisation

Le travail sera segmenté en trois sections : construction de la base de données, analyse statistique et ~~gestion~~ (*annulé par manque de temps)*, qui seront mises à jour en temps et heure. Pour des raisons pratiques, le travail à réaliser dans chacune des sections sera déconnecté des autres. Ainsi, les données brutes utilisées pour la mise en place de la base de données pourront être différentes de celles analysées en statistique ou en gestion. Les trois séries de tâches pourront donc être réalisées en parallèle. On se reportera aux instructions détaillées dans chacune des sections.

Le travail sera réalisé en groupes de 3 ou 4 étudiants.

Nature des données

On dispose d'informations sur

- **Les pistes** : identifiant dans le système, album de provenance, artistes interprètes. Par ailleurs, on utilise des caractéristiques du fichiers audio provenant d'un traitement de la plateforme Spotify, sur la durée, le rythme, la tonalité, le mode, la "dansabilité", etc ...
- **Les albums** : identifiant, nom de l'album, artistes interprètes, nombre de pistes, copyright, année de sortie, marché concerné, etc ...
- **Les artistes** : identifiant, nom de l'artiste, genre, etc ...
- **Les utilisateurs** : écoutes, likes, playlists, amis, évaluations, etc ...

Les données à utiliser/analyser/exploiter seront précisées dans les différentes sections.

Livrables

Chaque section précisera les livrables attendus et les dates limites de rendu. Un seul livrable par groupe de SAÉ est attendu.

Livrable 1 – Conception et création de la base de données à partir d'un jeu de données

Date limite du rendu : lundi 20 mai (12h00)

Instructions

L'objectif de ce premier livrable est de concevoir et créer la base de données à partir des fichiers CSV qui vous sont fournis dans la section Bases de données/Fichiers CSV. Un document décrivant les données contenues dans ces fichiers est également présent dans cette section.

Les fichiers CSV qui vous sont fournis contiennent des données qui sont souvent redondantes et ne peuvent donc pas être importées telles quels dans des tables d'une base de données relationnelle.

Avant de créer les tables et insérer les données contenues dans les fichiers CSV dans ces tables, il faut réfléchir à la structure de la base de données afin que celle-ci ne contienne pas de redondance.

Dans un premier temps, vous devez analyser ces fichiers CSV et décrire la structure de la base de données à l'aide d'un modèle entité-association.

Dans un deuxième temps, vous créerez les tables sur Oracle et importerez les données contenues dans les fichiers CSV dans vos tables. Pour réaliser ces tâches, vous pouvez utiliser un client graphique tel que DBeaver qui permet d'importer facilement des fichiers CSV.

Pour insérer les données dans les tables, vous devrez procéder de la façon suivante :

- **Importer** les fichiers CSV qui vous sont fournis dans des tables temporaires sur Oracle
- **Créer** vos tables sur Oracle sans oublier toutes les contraintes d'intégrité (clé primaire, contrainte d'intégrité référentielle et éventuellement contraintes de domaine (CHECK))

- **Réaliser** des requêtes SQL pour insérer les données des tables temporaires dans vos tables

Exemple de requête d'insertion de données depuis une autre table:

```
INSERT INTO maTable (col1, col2, col3)
SELECT col1, col2, col3 FROM table_temporaire
```

Ce livrable devra être déposé au format pdf, comprenant :

- **Le modèle entité/association,**
- **Le Schéma Relationnel de la base de données,**
- **Le nom du schéma sur Oracle (c'est à dire le compte étudiant) dans lequel vous avez créé vos tables. On doit retrouver dans vos tables toutes les données contenues dans les fichiers CSV,**
- **Les requêtes SQL de création des tables et d'insertion des données dans les tables.**

Livrable 2 – Création de vues

Date de rendu : lundi 10 juin (12h)

Instructions

L'objectif de ce deuxième livrable est de créer des vues mais aussi d'améliorer le travail rendu lors du premier livrable.

Améliorer si nécessaire le modèle entité/association rendu lors du premier livrable. Puis modifier le schéma relationnel et les données comprises dans les tables en fonction des améliorations apportées.

- **Réaliser les vues contenant les informations suivantes :**
    1. Pour chaque morceau (ou piste) le nombre d'interprètes, le nombre d'écoutes, le nombre d'écoutes en cours, l'évaluation moyenne, le nombre de likes, le nombre de playlists qui contiennent le morceau, le nombre de playlists où le morceau est en première position, le nombre de partages.
    2. Pour chaque album, le nombre de morceaux, le morceau le moins écouté et le morceau le plus écouté.
    3. Le morceau le plus écouté pour chacun des signes astrologiques des utilisateurs.
    4. Les 3 artistes les plus écoutés par genre de musique (attention il peut y avoir des exæquo).
    5. Pour chaque utilisateur, le pseudo de l'utilisateur et le nom du morceau le plus écouté parmi ceux qui se trouvent dans ses playlists. Pour ce morceau, on veut indiquer le chemin dans la playlist (par exemple dossier1/dossier11/PlaylistNemard/La quête).
    6. Pour chaque utilisateur premium, les morceaux partagés avec les utilisateurs amis.
    7. Le moment de la journée où il y a le plus d'écoutes (matin 6h - 12h, après-midi 12h - 18h, soir 18h - 24h, nuit 0h - 6h).

8. Pour chaque utilisateur, la note moyenne des évaluations de chacun des morceaux évalués.
9. Pour chaque morceau, la médiane parmi les notes moyennes attribuées par chaque utilisateur (sur le morceau).
10. Le morceau qui a le plus grand écart type sur ses notes.

Pour chaque question, vous pourrez réaliser autant de vues que cela est nécessaire.

Pour ce livrable vous devrez déposer le script SQL contenant le code de création des vues avec toutes les explications nécessaires. Si vous êtes amenés à modifier le modèle entité-association du précédent livrable vous déposerez également votre nouveau modèle entité-association.

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.

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.

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.

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								

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.

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									