TP 4 Transformation et Visualisation

TP Data Science M1 IA

Ilyas Bambrik

Table des matières

Exercice : Transformation des données

I

Lisez le Dataset *"players.csv"* présenté dans les cours et TPs précédants.

Question 1

Transformez la colonne *current_club_id* en variable catégorique et affichez la description statistique approprié.

Question 2

- Filtrez les lignes de votre Dataframe pour retrouvez seulement les joueurs avec un contrat qui expire en 2024 ou plus tard. Placez le résultat dans un nouveau Dataframe.
- Pour le résultat précédant, ajoutez une nouvelle colonne '*year_remaining*' qui prend le nombre d'années restants dans le contrat du joueur (utilisez l'année courante).

Question 3

- Créez une colonne *age* contenant l'age du joueur (nombre d'années depuis la date de naissance).
- Créez une nouvelle colonne '*age_bracket*' correspondant à l'une des tranches d'age suivantes : (12,17] , (17,20], (20,24], (24,28] ,(28,35] et (35,60].
- Quelle est la tranche d'age la plus fréquente dans le Dataset.
- Quelle est la valeur marchande (*market_value_in_eur*) maximal pour chaque tranche d'age.
- Créez une nouvelle colonne '*market_value_bracket*' qui divise d'une manière approximativement uniforme les joueurs sur six catégories selon la colonne *market_value_in_eur*. Affichez le DataFrame trié selon cette colonne.

Exercice : Data Cleaning et transformation avec MySQL (Exercice Supplémentaire +1 moyenne TP finale)

II

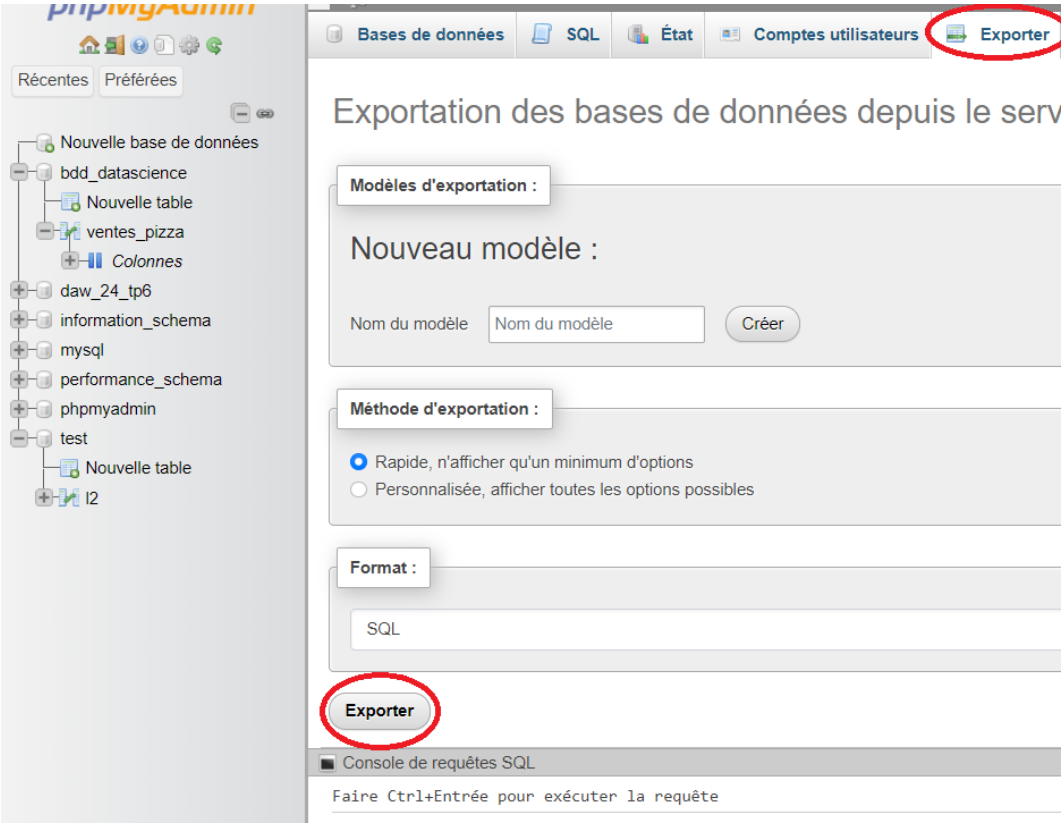
Le fichier *bdd_datascience.sql* contient un DataSet sous la forme d'une table SQL. Ce DataSet contient les ventes des pizzas dans la table *ventes_pizza*.

- Créez la base de donnée *bdd_datascience* puis exécutez le script *bdd_datascience.sql* pour créer la table *ventes_pizza*.
- Vous devez installer *MariaDB* indépendamment ou avec *Xamp*, ou bien *PostgreSQL*.
- La figure suivante montre comment exécuter le contenu du script *.sql* dans *Xamp*. Vous pouvez utiliser n importe quel outil d'exécution de requête SQL, il suffit de connecter correctement au SGBD.
- Dans Anaconda Prompt, installez *mysql-connector-python* afin d'accéder aux BDDs/tables MariaDB à partir de python par la commande suivante :

  ```
  pip install mysql-connector-python
  ```

  Si vous utilisez PostgreSQL installez et utilisez le module *psycopg2* au lieu de *mysql-connector-python* (voir le tutoriel *https://www.freecodecamp.org/news/postgresql-in-python/*):

  ```
  pip install psycopg2
  ```



Question 1

Donnez un programme python qui exécute une *seule une requête* SQL pour transformer de DataSet contenu dans *ventes_pizza* en Dataframe avec les propriétés suivantes (la requête doit retourner toutes les colonnes présentes dans la table) :

- Les valeurs nulles de la colonne *pizza_category* doivent être remplacées par *Classic*. Utilisez la fonction *COALESCE https://www.w3schools.com/sql/func_mysql_coalesce.asp* pour remplacer les valeurs nulles.
- Les valeurs nulles de la colonne *quantity* doivent être remplacées par la moyenne de cette colonne. La valeur de cette colonne doit être un *entier*.
- La colonne *pizza_size* doit contenir seulement les valeurs *'M', 'L', 'S', 'XL', 'XXL'* (toutes les valeurs doivent être des chaînes de caractères majuscules). Les valeurs non saisies ('-') doivent être changées par la valeur la plus fréquente de cette colonne. Les valeurs de cette colonne ne doivent pas avoir des espaces au début ou à la fin (utilisez la fonction SQL *TRIM*). Ne pas *hardcoder* la valeur la plus fréquente pour que votre code s'adapte aux changement dans le DataSet.

Votre programme doit seulement afficher la table transformée. N'enregistrez pas les changements dans la table pour l'instant.

Indice :

Pour récupérez les noms des colonnes du résultat après l'exécution d'une requête SQL, utilisez l'attribut *column_names* du curseur MySQL.

Question 2

Donnez *des requêtes SQL* pour tester votre solution en regardant les valeurs uniques du résultat pour les colonnes catégoriques *pizza_category*, *pizza_size*. Pour la colonne *quantity*, donnez un requête qui teste que aucune valeur nulle n'existe et que le résultat est un entier.

Question 3

En regardant la colonne *quantity*, certaines commandes sont enregistrées pour 100 pizzas. Modifiez la première requête pour que les valeurs manquantes de *quantity* sont remplacées par la *médiane* au lieu de la moyenne.

Indice :

Utilisez la fonction d'intervalle *PERCENTILE_DISC().*

Question 4

Sauvegardez le résultat de la question précédente dans un fichier *Excel .xlsx* (créez un Dataframe à partir de l' exécution de la requête de Data Cleaning).

Exercice : Visualisation des données 1

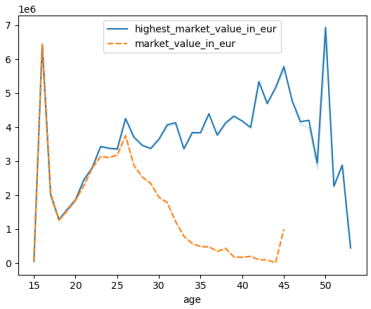
III

Question 1

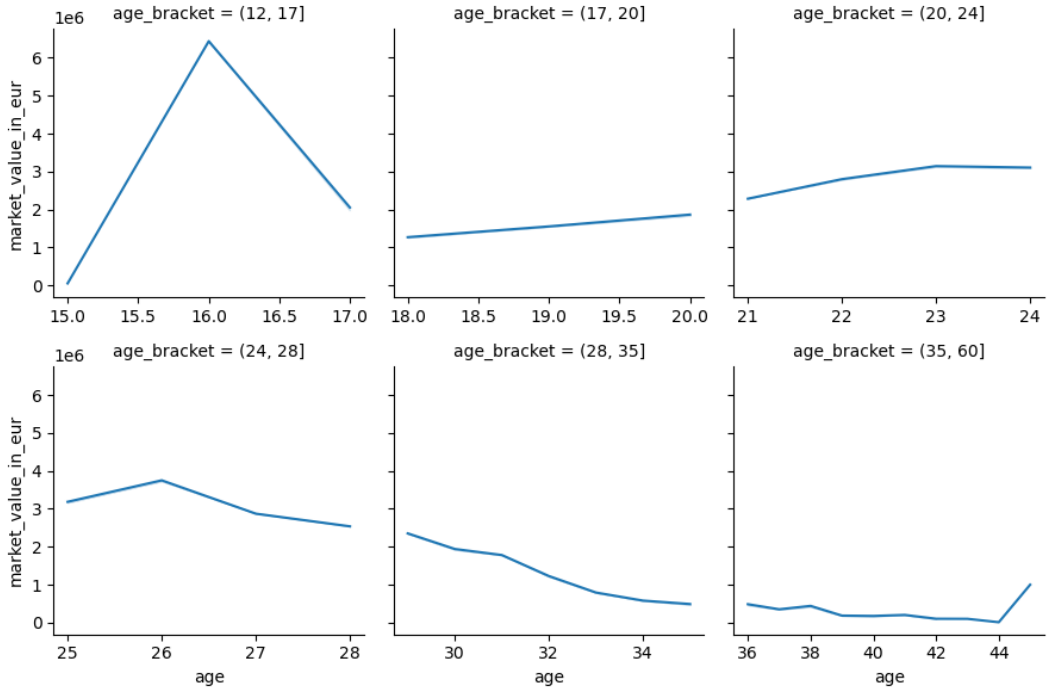
Tracez *market_value_in_eur* en fonction de *year_remaining*. Que remarquez vous (la relation entre le deux variables tracées)?

Question 2

Tracez *market_value_in_eur* et *highest_market_value_in_eur* en fonction de la colonne *age* dans la même figure.



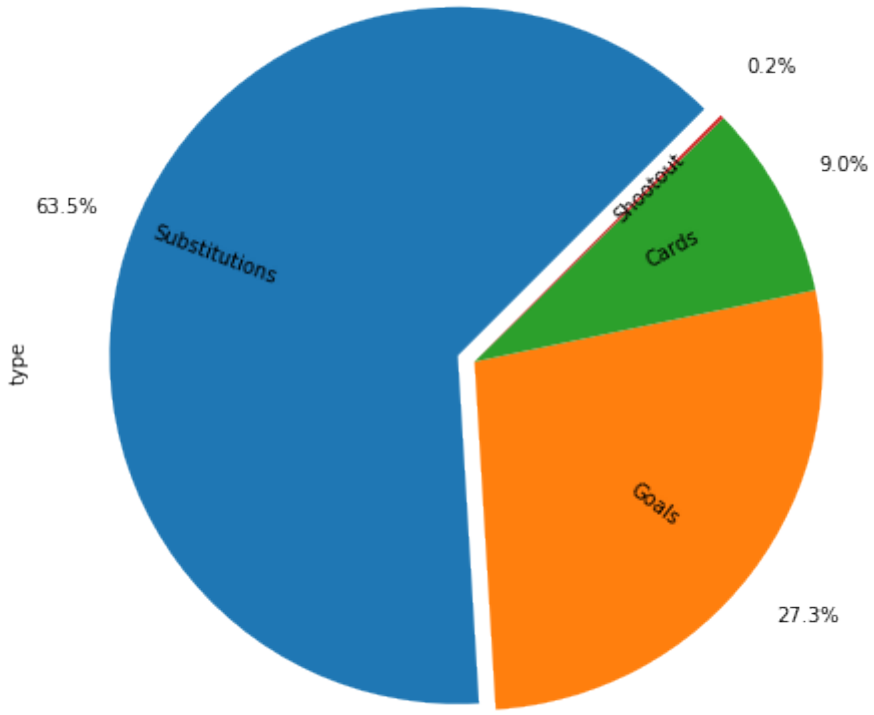
Utilisez *FacetGrid* pour tracez *market_value_in_eur* en fonction de la colonne *age* pour chaque tranche d'age définie dans *age_bracket*.



Question 3

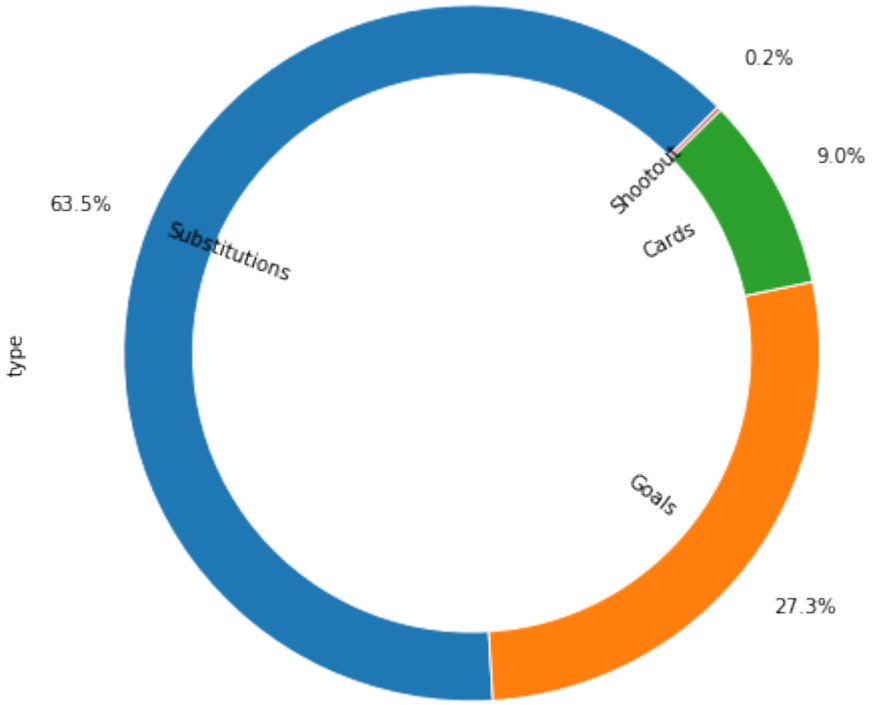
A partir du DataSet *game_events.csv*, créez un graphe Pie plot qui reflète les pourcentages des types d' événements (colonne *type*).

- La première proportion doit commencer à partir de l'angle 45°.
- Les labels doivent être inscrit à l'intérieur du disque et les pourcentages à l'extérieur.
- La plus grande proportion doit être coupé du graphique avec une distance de 0.05.



Question 4

Transformez le Pie plot précédant en Donut plot.



Exercice : Visualisation des données 2

IV

Lisez le DataSet *canada.csv*.

Question 1

- A partir du DataFrame *canada.csv*, sélectionnez seulement les colonnes *RegName*, *1980*,*1981*, *1982*, etc jusqu'à *2013*.
- A partir du résultat précédant, créez un DataFrame contenant le nombre d'émigrants pour chaque région et chaque année indexé par le nom de la région (continent).

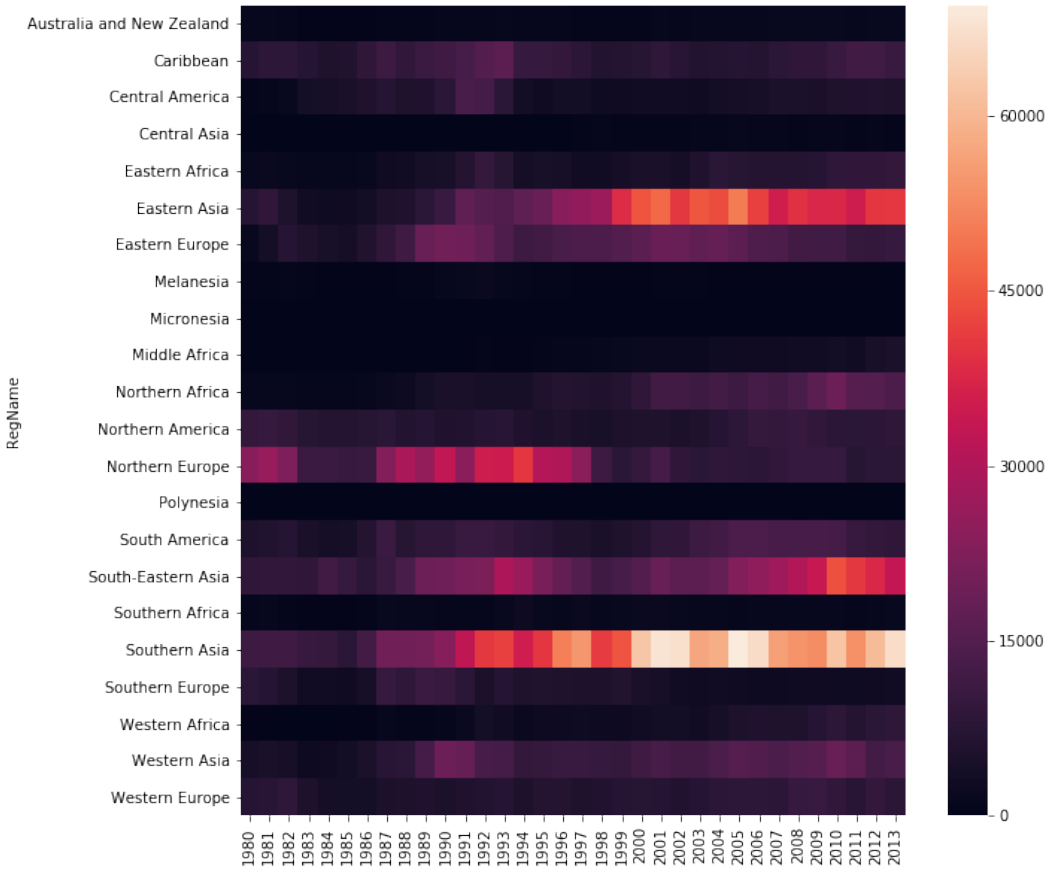
Indice :

Utilisez la fonction range pour créer la liste des années.

RegName	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	...	2004	2005	2006	2007	2008	
Australia and New Zealand	1304	1119	848	457	481	467	532	675	610	790	...	1280	1279	1193	1383	1498	
Caribbean	7045	8310	8326	6998	5553	6048	8716	10932	9229	10786	...	6630	6816	6652	7826	8862	
Central America	734	921	1612	3648	4087	4862	5909	6804	5596	5821	...	3346	3990	4140	5039	4891	
Central Asia	0	0	0	0	0	0	0	0	0	0	...	995	1134	903	936	805	
Eastern Africa	1471	1641	1426	1094	1187	1134	1454	2734	3237	4094	...	7726	7083	6750	6669	6705	
Eastern Asia	6836	8895	5481	3254	2624	2979	3416	5403	5887	7796	...	43550	50306	41763	35591	39602	3
Eastern Europe	1467	3698	6873	5459	4433	3969	6144	8861	11499	18671	...	18360	16724	14389	13886	11988	1

Question 2

Affichez un Heatmap des nombres immigrants en fonction de l'année et la région en utilisant le résultat de la question précédant. Modifiez la taille de la figure pour quelle soit de taille 10x10 pouces.



Question 3

Tracer un Area plot de nombres d'immigration de chaque région en fonction de l'année.

Indice :

Transposez le résultat de la question précédente.