TP 3 Data Cleaning et Accesseurs

TP Data Science M1 IA

Ilyas Bambrik

2024

Table des matières

Exercice : Interprétation des entrées Nulles

I

Lisez les trois Datasets *game_events.csv*, *clubs.csv*, *games.csv* et *players.csv* de Transfermarket dans trois Dataframes différents.

Question 1

- Listez les colonnes avec le nombre de cellules nulles pour chacune des DataSet.
- Laquelle des colonnes de *game_events.csv* contient le plus grand nombre d'entrées vides ?

Indice :

Regardez la méthode utilisée dans le cours pour *players.csv* afin de compter le nombre d'entrées vides (page 4).

Question 2

Selon le contenu de *game_events.csv*, quelle est la meilleure façon pour éliminer les cellules vides, la suppression des colonnes ou bien la suppression des lignes ?

Question 3

La colonne *player_assist_id* (dans *game_events.csv*) indique l'identifiant du joueur qui a participer/créer un but. En outre *player_in_id* indique l'identifiant du joueur entrant en jeu à la place d'un autre joueur.

Pour les colonnes *description*, *player_in_id* et *player_assist_id*, classifier les cellules nulles de ces colonnes dans les deux catégories suivantes (*vous devez valider votre hypothèse avec des tests*) :

- Information non saisie
- Information inexistante (n'a pas de sens pour cet événement)

Question 4

- Listez les valeurs de la colonne *type*. Trouvez les erreurs de saisie dans cette colonnes à la base des fréquences de chaque valeur.
- Nettoyez cette colonne afin que que toutes les valeurs de la colonne soient *title case* (premier lettre du mot majuscule seulement).

```
""" 'SUBSTITUTIONS ', 'GOALS',
       'goals', 'cards', 'cards ', 'goals ', 'substitutions',
       'SUBSTITUTIONS', 'substitutions ', 'GOALS '"""
games.type.str.title().str.strip().unique()
```

Question 5

- Est ce qu'ils existent des joueurs dans *game_events.csv* qui ne sont pas contenus dans *players.csv* ? Démontrez votre réponse.
- Dans *game_events.csv*, quel est le nom du joueur le plus fréquemment rencontré ?
- Quel est le nom du joueur qui a obtenu le plus de cartons jaunes/rouges ?

Question 6

- Nous considérons un joueur comme expulsé d'un matche s'il obtient deux cartons dans le même match. Quels sont les joueurs qui se sont fait expulsés selon *game_events.csv* au moins une fois?
- Quel est le nom du joueur avec le maximum nombre d'expulsions ?
- Quels sont les noms des joueurs ayant marquer trois buts dans un matche au mois une fois ?
- Quelle est la compétition ayant le plus grand nombre de buts? Utilisez le DataFrame *games.csv*.

Indice :

Groupby game_id.

Question 7

- Pour la colonne *height_in_cm* de *players.csv*, utilisez la méthode describe() et trouvez les valeurs aberrantes. Remplacez la valeur aberrante par une valeur plus appropriée. Utilisez l'opérateur *.at[]* pour affecter la valeur (exemple *df.serie.at[index]=valeur*)
- Créez un DataFrame contenant pour chaque player_id, ça taille (height_in_cm) et nombre de but marqués selon le Dataset *game_events.csv*.
- Dans le résultat de la question prétendante, remplacez, les valeurs nulles e de *height_in_cm* par la médian *median()*) et les valeurs nulles de nombre de buts par 0.
- Calculez la corrélation entre *height_in_cm* et le nombre de buts marqués. Utilisez la méthode *.corr()*.

Exercice : Accesseurs

II

Question 1

Selon *games.csv* et *clubs.csv*, quel est le nom de l'équipe (*away_club_id* et *home_club_id*) qui est la plus fréquemment classée parmi les trois premiers dans son championnat (*away_club_position*, et *home_club_position*) ?

Question 2

Trier les matches de *games.csv* selon la colonne *date*. Vous devez convertir la colonne en type *datetime* avant le trie. Les valeurs qui ne sont pas correspondantes à une date valide doivent être converties en *Not A Timestamp* (*NAT*).

Question 3

Considérons que la première moitié de la saison se termine le 30 décembre. Pour chaque saison dans *games.csv* , quel est le nombre de buts marqués dans la deuxième partie de la saison ?

Question 4

Selon la colonne *description* du DataSet de *game_events.csv*, trouvez le nom du joueur qui a obtenu le plus grand nombre de carton rouge.

Indice :

Avant de procéder, regardez les valeurs uniques de la colonne *description* .

Question 5

- Dans le Dataframe, *games_event.csv*, transformez le type de la colonne type en "*category*".
- Triez le Dataframe, *games_event.csv*, de manière descendante selon le nombre de butes du match de manière descendante et puis selon le type d'événement de manière descendante. Le classement des événements doit être comme suite : *[Cards < Substitutions < Shootout < Goals]*

Indice :

- Vous devez créer un nouvelle colonne de *games_event.csv*, qui comptabilise le nombre de butes pour chaque match (*game_id*).
- Pour transformer une *Series* en *Dataframe*, vous pouvez utiliser la méthode *to_frame()*.