TP 3 Data Cleaning et Accesseurs

M1 - IA / GL
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.
- La quelle 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)

La quelle des trois colonnes est la moins significatif ?

Question 4

- 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 5

- Nous considérons un joueur comme expulsé d'un matche s'il obtient deux cartons dans le même match. Quel 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 ?

Indice :

Groupby game_id.

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 le plus fréquemment classée parmi les trois premiers du championnat (*away_club_position*, et *home_club_position*) ?

Question 2

Trier les matches de *games.csv* selon la date. Vous devez convertir la colonne en date avant le trie.

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* .