TD SAS - Manipulation des données (2)

Exercice 1 :

1. Sachant que Federer est plus vieux de Nadal de 5 ans et que le cumul de leur âge vaut 71 (qui est par ailleurs l'année de naissance de Pete Sampras), trouvez l'âge de ces deux joueurs de tennis émérites. La résolution d'un système matriciel (relativement simple…) est évidemment attendu car il est interdit de demander la réponse à Google !

Indications : pensez à poser le problème sous forme de deux équations à deux inconnues qu'on pourra écrire matriciellement AX=b. On aura donc X=A-1b (sous SAS, utiliser la fonction inv()

2. Créez un vecteur V contenant le nombre de victoires entre 2001 et 2018 de Roger Federer (1 124) et de Rafael Nadal (985). Dans un second vecteur appelé ratio, stockez le nombre moyen de victoire par an de chacun de ces joueurs. En concaténant (operateur ||) le vecteur des âges, des victoires et du ratio, stockez dans une table "recap" les précédentes informations (*utilisez create, voir page 8 du cours*). N'oubliez pas de préciser le nom du joueur pour la ligne correspondante, et à quoi correspondent les colonnes.
3. Nicolas, né en 1996, a gagné quant à lui 200 matchs sur les 10 dernières années. Créez un dataset Nico contenant les informations du dataset précédent (*page 7 du cours*), et à l'aide d'une proc append, insérez cette ligne dans le dataset Recap.

Exercice 2 :

Nous disposons d'un fichier **gains.xlsx** contenant, pour un numéro de joueurs de tennis donnée, le nom du tournois, l'année, le montant de la prime et le sponsor.

1. Créez la bibliothèque "TD2", importez le fichier « **gains.xlsx** » et sauvegardez le dans cette bibliothèque.
2. Calculez les fréquences pour la variable « lieutournoi » (proc freq).
3. Filtrez les valeurs où le lieu de tournoi est Roland Garros et sponsor est Peugeot.
4. Calculez la moyenne de la variable « prime » par l'année (proc sort et proc means).
5. Créez une nouvelle variable « country » selon le critère de « lieutournoi ». Les valeurs de la variable « country » sont France, UK, et US (if else ou encore select when). Utilisez proc freq pour calculer les fréquences de la variable «country ».

Exercice 3 : Manipulation SQL et macro-programme

- **rencontres.xlsx** contient pour chaque rencontre, le numéro du joueur gagnant, celui du perdant ainsi que le nom et l'année du tournois.
- **joueurs.xlsx** contient le nom, le prénom, la date de naissance et la nationalité des joueurs ayant participé à des tournois de tennis dans les années 1990.

Importer les fichiers **gains.xlsx** et **rencontres.xlsx** dans la bibliothèque TD2 créé à l'exercice précédent.

A l'aide de procédure SQL :

1. Mettre à jours la table joueurs afin de renommer l'ensemble des nationalités des joueurs avec les deux première lettre ,en majuscule, du pays, à l'exception des Etat-Unis pour lequel le sigle US doit apparaître. (*Utilisation de **proc sql** - cours p19, et de **substr** -* http://support.sas.com/documentation/cdl/en/imlug/64248/HTML/default/viewer.htm#imlug_langref_sect292.htm)

2. Créer une vue indiquant le nombre de joueur par nationalité.

3. Afficher le nombre de rencontre gagnées pour chaque nationalité.

4. Quel joueur a remporté le plus de match ? Afficher son nom, prénom, nationalité et le nombre de rencontres qu'il a remporté.

5. Ecrire un macro-programme nommées "gagnant" prenant en paramètre une année et un tournoi et affichant les joueur ayant gagné au moins un match lors de ce tournoi. (Utilisation de proc sql au sein de la macro, voir cours p24 à p27)

   Résultat attendus :
   %gagnant('*Roland-Garros*',*1990*);

nom	prenom
NAVRATILOVA	Martina
WILANDER	Mats
McENROE	John
McENROE	John

Exercice 4 : Fusions/Jointures avec MERGE et exportation

1. Reprenez les datasets **gains.xlsx** et **joueurs.xlsx (ou GAINS et JOUEURS)**.
   Effectuez une fusion de ces deux datasets en une seule table par la variable commune **num_joueur**. Appelez-la **TENNIS**.
   Remarque : La variable commune ne porte pas le même nom dans les deux datasets. De plus, n'oubliez pas de trier vos jeux de données sur cette variable.
   (Cours Manipulation des données (2) pages 12 à 14).
2. Affichez-la pour vous assurer que la fusion s'est bien déroulée.
   On remarque que des joueurs n'ont pas gagné de compétitions mais sont présents dans la table fusionnée. (Joueurs présents dans le dataset **joueurs** mais pas dans celui de **gains**).
3. Dans la question 2, vous avez réalisé une *FULL OUTER JOIN.* Rajoutez une condition sur l'étape DATA *(IF Statement)* pour ne retourner que les joueurs (et leurs gains) ayant gagné des compétitions. Autrement dit, vous devez mettre en place une *INNER JOIN*.

4. Réalisez maintenant la fusion (*INNER JOIN*) des deux datasets avec PROC SQL. Appelez la table créée **TENNIS_SQL**. Vous devriez obtenir le même résultat que dans la question 3.
   Remarque : La syntaxe de la requête suit cette forme :

```
PROC SQL NOPRINT;
CREATE TABLE NOM_TABLE AS
      SELECT  *  FROM  TABLE1  T1,  TABLE2  T2  WHERE  T1.variable  =
T2.variable
      ORDER BY variable;
QUIT;
PROC PRINT DATA=NOM_TABLE;
```

5. Exportez la table **TENNIS_SQL** en un fichier que vous nommerez **tennis.csv**.
   (Cours Manipulation des données (2) page 28).
6. Ouvrez le fichier **tennis.csv** nouvellement créé et vérifiez que l'export s'est bien effectué.