Exercices du Chapitre 19 : Tableaux et analyse de données avec Pandas

Pour les exercices suivants, il sera nécessaire d'importer les fonctions suivantes :

```
from math import log, sqrt
from sympy import Li
from pandas import Series, DataFrame, read_excel, read_csv
```

Exercice 1. (Hypothèse de Riemann [1]).

- Constuire un tableau de données de 1000 lignes ($0 \leq x < 1000$) contenant les trois colonnes : $\pi(x)$, le nombre de nombres premiers inférieurs ou égal à x ; $x/\log x$, l'approximation de $\pi(x)$ faite par Gauss ; $\mathrm{Li}(x) = \int_2^x \frac{dt}{\log t}$, l'approximation de $\pi(x)$ faite par Dirichlet.
- Ajouter une colonne dans le tableau qui donne la valeur absolue de la différence entre $\pi(x)$ et $\mathrm{Li}(x)$.
- Trouver (approximativement) une constante C telle que $|\pi(x) - \mathrm{Li}(x)|$ vaut à peu près $C \cdot (\sqrt{x}\log x)$.

Remarque : Helge von Koch a montré en 1901 [2] que l'Hypothèse de Riemann est vraie si et seulement si

$$\pi(x) = \mathrm{Li}(x) + O\left(\sqrt{x}\log x\right).$$

Remarque : Les fonctions `read_csv` et `read_excel` peuvent prendre des URL en entrée ce qui évite d'avoir à télécharger les fichiers localement. Par exemple :

```
read_csv("http://www.slabbe.org/Enseignements/MATH2010/capital21c_tableauSI1.csv")
```

Exercice 2 (Les arbres de Namur)**.** Créer un tableau de données de Pandas à partir du fichier `arbresremarquables.xls` au sujet des *Arbres remarquables Namur* disponible sur le site `http://data.gov.be`. Afficher les premières et les dernières lignes du tableau. Quelle est la circonférence moyenne des arbres remarquables de Namur ? Quelle essence d'arbre est la plus représentée ? Est-ce que l'arbre le plus grand de Namur est situé dans un domaine privé ou public ? De quelle essence s'agit-il ?

Exercice 3 (Le capital au 21^e siècle, Tomas Pikkety)**.** Créer un tableau de données de Pandas à partir du fichier `capital21c_tableauSI1.csv` disponible sur la page du cours. Utiliser ce tableau pour recréer le graphique *La part du décile supérieur dans le revenu total (y compris plus-values) aux Etats-Unis, 1910-2010* du chapitre d'introduction du livre de Piketty [3]. En observant le graphique, compléter les trous de la citation suivante :

> *Lecture : la part du décile supérieur dans le revenu national américain est passée de ..... % dans les années 1910-1920 à moins de .....% dans les années 1950 (il s'agit de la baisse mesurée par Kuznets) ; puis elle est remontée de moins de .....% dans les années 1970 à .....% dans les années 2000-2010.*

1. Mazur, Barry, et William Stein. Prime Numbers and the Riemann Hypothesis. Cambridge University Press, 2015. `http://wstein.org/rh/`
2. Von Koch, Helge (1901). "Sur la distribution des nombres premiers". Acta Mathematica 24 (1) : 159–182. `https://dx.doi.org/10.1007/BF02403071`
3. Thomas Piketty, Le capital au 21e siècle, Editions du Seuil, Septembre 2013, `http://piketty.pse.ens.fr/capital21c`

Exercice 4 (La lotterie)**.** En novembre 2015, le site *BuzzFeedNews* a fait des simulations pour savoir quelles sont les chances de perdre de l'argent à la loterie. On peut consulter leur jupyter notebook ici :

`https://github.com/BuzzFeedNews/2015-11-lottery-simulations`

Faire la lecture du texte et du code de ce notebook Jupyter. Quelles fonctionalités de la librairie `pandas` ont-ils utilisées ? Quelles sont les chances de perdre de l'argent en achetant des billets de loterie selon les simulations effectuées par BuzzFeedNews ?

Exercice 5 (BuzzFeedNews)**.** Faîtes la lecture d'un autre sujet de votre choix parmi la liste de la page :

`https://github.com/BuzzFeedNews/everything`

Dans le code associé à l'article, reconnaissez-vous des fonctions ou librairies Python que vous connaissez ? Trouver 3 choses que vous connaissez et que vous avez appris dans ce cours. Trouver 3 choses que vous ne connaissez pas et que vous n'avez pas appris dans ce cours.

Exercice 6 (Nombre de naissances par jour)**.** Créer un tableau de données de Pandas à partir du fichier sur le *Nombre de naissances par jour* pour la période du 1er janvier 2008 au 31 décembre 2014 disponible sur le site `http://data.gov.be`. En moyenne, combien d'enfants naissent chaque jour en Belgique ? Entre 2008 et 2014, quelle année y a-t-il eu le plus de naissance le jour du 10 mai ? Y a-t-il une grande différence entre le minimum et le maximum pour le jour du 10 mai ? Plus difficile : quel mois a vu naître le plus d'enfants pendant ces sept années ?

Exercice 7 (`data.gov.be`)**.** Choisir un fichier de données de votre choix parmi les 4800+ jeux de données disponibles sur `http://data.gov.be`. Tenter de l'ouvrir dans pandas (il y a parfois des erreurs d'importation avec le format excel, les données sont rarement parfaites dans la vraie vie. Dans ce cas, une option est de sauvegarder la feuille excel sous le format csv à partir d'Excel ou Libre Office). Réfléchissez à une question de votre choix au sujet de ces données. Répondez à votre question.