Introduction à la Science des données

Travail pratique 03 – Outils pour le calcul scientifique II

Professeurs: Andres Perez-Uribe & Carlos Peña
Assistants: Axel Fahy, Shabnam Ataee, Xavier Brochet
Emails : prenom.nom@heig-vd.ch

Objectifs:

- Se familiariser avec la bibliothèque Pandas de gestion et traitement des dataframes.
- Mise en pratique de l'analyse et la caractérisation simple des attributs d'une base de données.

1. Analyse des données socio-économiques

La fondation gapminder fondé par Hans Rosling et famille fournit une base de données sur les pays, des outils d'analyse et des études d'analyse socio-économique très intéressantes. Nous allons utiliser une petite base de données contenant la population, l'espérance de vie et le PIB par habitant pour différents pays du monde de 1952 à 2007, pour faire quelques analyses.

1) Commencez par importer le module *gapminder* avec pip. Une fois installé ce module, vous aurez accès à un dataframe appelé gapminder. Utilisez les méthodes head(), describe() et info() pour vous familiariser.
2) Utilisez la commande pandas.DataFrame.hist(gapminder) pour vous faire une idée plus précise des valeurs dans la base de données.
   Q1. Que pouvez-vous conclure à partir de ces histogrammes ?
3) Trouvez combien d'observations il y a par pays (c.a.d., combien de données par année par pays) et vérifiez s'il y a des données manquantes.
4) Listez les valeurs uniques présentes dans les colonnes 'continent', 'country', et 'year'. p.ex. utilisez la méthode unique().
5) Calculez la moyenne de l'espérance de vie de tous les pays en 1952 et en 2007. Générez un bar chart permettant la comparaison de ces moyennes.
6) Calculez la moyenne de l'espérance de vie des pays par continent en 1952 et en 2007. Générez un bar chart permettant la comparaison de ces moyennes.
   Q2. Quel continent a eu la plus grande progression ?
7) Générez un plot montrant l'évolution de 1952 à 2007 de la moyenne de la population par continent.
   Q3. Que pouvez-vous conclure à partir de ces plots ?

8) Générez deux sub-plots contenant un scatter plot avec l'espérance de vie des pays sur l'axe *y* et le PIB per capita sur l'axe *x*, pour l'année 1952 et séparément pour l'année 2007. Indiquez le continent par une couleur.

   Q4. Que pouvez-vous conclure à partir de ces scatter plots ?

9) La fonction scatter de matplotlib permet non seulement d'indiquer une couleur mais aussi la taille de chaque « point ». Générez des scatter plots de l'espérance de vie vs. PIB par habitant (années 1952 et 2007) et visualisez la population associée à chaque pays à l'aide d'un cercle de taille proportionnel à la population. Indiquez le continent par une couleur.

   Aide : `plt.scatter(x, y, s=taille, facecolors='none', edgecolors='r')`

10) Générez un scatter plot montrant l'augmentation de la population entre 1952 et 2007 (axe *y*) par rapport au PIB per capita. Indiquez le continent par une couleur.

    Q5. Que pouvez-vous observer dans cette figure et que pouvez-vous conclure ?

2. Base de données d'animaux

1) Créez un notebook pour lire la base de données et générez un dataframe.
2) Prenez la masse corporelle (en grammes) des animaux à leur âge adulte (colonne « 5-1_AdultBodyMass_g »).
3) Observez la distribution des valeurs, calculez la valeur minimale, la valeur maximale, la moyenne et la médiane. Vérifiez s'il y a des valeurs manquantes, des valeurs aberrantes, etc.
4) Générez un histogramme des masses corporelles.

   Q6. Que pouvez-vous conclure à partir de cet histogramme ?

5) Générez un histogramme des masses corporelles, mais pour les animaux qui n'atteignent pas les 50Kg à leur âge adulte.

   Q7. Que pouvez-vous conclure à partir de cet histogramme ?

3. Loi de Benford

La loi de Benford, initialement appelée loi des nombres anormaux par Benford en 1938, fait référence à une fréquence de distribution statistique observée empiriquement sur de nombreuses sources de données dans la vraie vie, ainsi qu'en mathématiques. Dans une série de données numériques, on pourrait s'attendre à voir les chiffres de 1 à 9 apparaître à peu près aussi fréquemment comme premier chiffre significatif, soit avec une fréquence de 1/9 = 11,1 % pour chacun. Or, contrairement à cette intuition, la série suit très souvent approximativement la loi de Benford : pour près du tiers des données, le 1er chiffre significatif le plus fréquent est le 1. Viennent ensuite le chiffre 2, puis le 3, etc., et la probabilité d'avoir un 9 comme premier chiffre significatif n'est que de 4,6 % (voir la table de fréquences ci-dessous).

C'est une loi observée aussi bien dans les données des sciences humaines et sociales, que dans des tables de valeurs numériques comme celles qu'on rencontre en physique (p.ex. la distance entre les étoiles), en géographie (p.ex. la longueur des fleuves et des rivières), en biologie, en économie, etc. Cette loi mathématique est utilisée pour vérifier la véracité des chiffres publiées : p.ex. dans les déclarations d'impôts, dans les publications du nombre de cas d'infections dans un pays lors d'une pandémie, etc.

chiffre	0	1	2	3	4	5	6	7	8	9
1er	-- -	30.1%	17.6%	12.5%	9.7%	7.9%	6.7%	5.8%	5.1%	4.6%

Table des fréquences d'apparition du premier chiffre significatif selon la loi de Benford

Dans le cadre de cet exercice, on vise à vérifier la loi de Benford avec la base de données des espèces d'animaux utilisée dans le point 1.

1) Prenez la masse corporelle (en grammes) des animaux à leur âge adulte (colonne « 5-1_AdultBodyMass_g ») après filtrage des données manquantes ou aberrantes.
2) Extrayez le chiffre le plus significatif de la masse corporelle des animaux.
3) Calculez la fréquence d'apparition de chaque digit pour l'ensemble d'animaux.
4) Comparer vos résultats avec la loi de Benford. Présentez une bar chart présentant cette comparaison.
   Q8. Que pouvez-vous observer ? Commentez vos résultats.
5) Essayez une autre variable pour vérifier si elle répond à la loi de Benford.
   Q9. Commentez vos résultats.

Rapport

Préparer un fichier compressé nommé nom1_nom2_ISD_TP3.zip intégrant vos deux notebooks (p.ex. gapminder.ipynb et animaux.ipynb) et le télécharger sur Cyberlearn avant la date limite indiquée sur le site du cours. Les notebooks doivent intégrer tous les points demandés et vos réponses aux questions numérotées Qi en utilisant des cellules Markdown.