Tutoriels de référence :

http://tutoriels-data-mining.blogspot.fr/2017/02/python-manipulations-des-donnees-avec.html

Questions :

On souhaite traiter le fichier « **Census.xlsx** ».

« classe » joue un rôle particulier, la variable indique le niveau de revenu c.-à-d. les personnes qui ont un revenu annuel supérieur (more) ou inférieur (less) à un seuil quelconque.

1. Importez la libraire **Pandas** (import)
2. Chargez le fichier « **Census.xlsx** » (read_excel)
3. Combien y a-t-il de variables dans le fichier ? Combien y a-t-il d'observations ? (info)
4. Afficher le résumé des données (describe)
5. Essayer de répondre aux différentes questions suivantes : quelle est la *proportion* des hommes (sex = male) ? celle des « classe = more » ? (value_counts)
6. Construire le diagramme à bandes pour les variables « marital_status » et « relationship » (http://pandas.pydata.org/pandas-docs/version/0.18.1/visualization.html ; bar).
7. Pour les mêmes variables, construire les diagrammes à secteurs (http://pandas.pydata.org/pandas-docs/version/0.18.1/visualization.html ; pie).
8. Croiser les variables « classe » et « sex ». Quelle est la proportion des « more » parmi les hommes ? Parmi les femmes ? Est-ce que ce résultat nous permet de conclure que le niveau de revenu est différent selon que l'on est un homme ou une femme ? (crosstab)
9. Croiser maintenant « relationship » et « marital status ». Pour chaque valeur de « relationship », quelle est la modalité de « marital status » qui lui est le plus associé ? (crosstab + idxmax). Que peut-on en conclure ?
10. Nous souhaitons quantifier l'intensité de la liaison entre ces deux variables. Calculez le KHI-2 (https://docs.scipy.org/doc/scipy-0.16.1/reference/generated/scipy.stats.chi2_contingency.html) du test d'indépendance. En déduire le V de Cramer (https://lemakistatheux.wordpress.com/2013/05/31/le-v-de-cramer/). Comment pourrait-on qualifier la relation entre ces deux variables ?
11. Penchons-nous maintenant sur la variable « age ». Calculer sa moyenne et son écart-type (mean, std) (http://pandas.pydata.org/pandas-docs/stable/generated/pandas.Series.html)
12. Centrer et réduire « âge » c.-à-d. on lui retranche la moyenne et on divise le tout par l'écart-type. Recalculer la moyenne et l'écart-type de la variable transformée. Que constate-t-on ?
13. Calculer la médiane et les quartiles d'ordre 1 et 3 de l'âge (quantile).

14. Construire le graphique BOXPLOT (boîte de Tukey) pour la variable « âge » (boxplot). Que remarque-t-on ?
15. Produire l'histogramme de la variable âge (hist).
16. Calculer la corrélation entre « age » et « hours per week » (corr). Peut-on dire que ces deux variables sont liées ? Réaliser le graphique nuage de points entre ces deux variables pour affiner votre réponse (scatter). Que conclure ?
17. Construire le boxplot de « âge » selon « relationship » (boxplot). Il y a des choses à remarquer dans ce graphique ?
18. Calculer la moyenne de l'âge pour chaque valeur de « relationship » (pivot_table). Le calcul confirme l'impression laissée par le graphique précédent ?
19. On s'intéresse à l'influence du niveau d'instruction (education) sur le revenu (classe). Créez une variable qui permet d'identifier les personnes ayant un des niveaux suivants : "Bachelors", "Masters", "Prof-school", "Doctorate". Combien d'observations répondent à ce critère ? (isin)
20. Quelle est la proportion de classe = more parmi ces individus, quelle est cette même proportion chez les autres (qui n'ont pas ce niveau d'études). Est-ce que le niveau d'instruction a un impact sur le revenu ? (crosstab)