Traitement des données sous Excel. Des indications sur les fonctions à utiliser sont parfois indiquées dans le texte (en majuscule).

Des liens vers des supports devraient également vous permettre de comprendre la nature des traitements demandés.

A chaque question doit correspondre une feuille Excel, étiqueté avec le numéro de question. Expliciter les traitements effectués et les réponses aux questions dans une zone de texte.

« classe » joue un rôle particulier, la variable indique les personnes qui ont un revenu annuel supérieur (more) ou inférieur (less) à une valeur donnée.

Charger le fichier « **census.xlsx** » sous Excel.

1. Combien y a-t-il de variables dans le fichier ? Combien y a-t-il d'observations ?
2. Combien y a-t-il de variables qualitatives ? De variables quantitatives ? Scinder les données en mettant les variables qualitatives (respectivement quantitatives) dans une feuille à part.
3. Pour chaque variable qualitative, calculer les distributions de fréquences absolues et relatives (mettre tous les tableaux pour chaque variable dans une feuille unique) (cf. TABLEAUX CROISES DYNAMIQUES). Essayer de répondre aux différentes questions suivantes : quelle est la proportion des hommes (sex = male) ? celle des « classe = more » ? celle des personnes travaillant pour le gouvernement (workclass contenant le terme ''gov'').
4. Construire le diagramme à bandes pour les variables « marital_status » et « relationship » (cf. http://www.ihet.rnu.tn/download/Chapitre1.pdf). Pour les mêmes variables, construire les diagrammes à secteurs.
5. « Education » correspond en réalité à un niveau d'éducation atteint. C'est donc une variable qualitative ordinale avec les modalités suivantes { Preschool, 1st-4th, 5th-6th, 7th-8th, 9th, 10th, 11th, 12th, HS-grad, Some-college, Assoc-voc, Assoc-acdm, Bachelors, Masters, Prof-school, Doctorate }. Quelle est la proportion des personnes qui ont uniquement le niveau « Preschool » ? Quelle est la proportion de personnes qui ont au moins le niveau « Bachelors » ?
6. Croiser les variables « classe » et « sex ». Quelle est la proportion des « more » dans l'échantillon global ? 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 ?
7. Calculer le KHI-2 du tableau croisé entre « classe » et « sex » (http://eric.univ-lyon2.fr/~ricco/cours/cours/Dependance_Variables_Qualitatives.pdf, section 2.1). Puis en déduire le v de Cramer (section 2.4). Quelle est la valeur obtenue ? Quel est son domaine de variation ? Peut-on conclure dans notre cas que la liaison existe réellement ? Quelles sont les associations entre les modalités qui contribuent le plus à l'information (*voir contributions aux khi-2*) ?

8. Croiser maintenant « relationship » et « marital status ». Pour chaque valeur de « relationship », quelle est la modalité de « marital status » qui lui est le plus associé ? Et inversement ? Est-ce que la relation est symétrique ?
9. Penchons-nous maintenant sur les variables quantitatives. Calculer les moyennes et écarts-type de chaque variable (MOYENNE, ECARTYPEP). Quelle est la différence entre ECARTYPE et ECARTYPEP d'Excel ? Peut-on comparer les moyennes et écarts-type d'une variable à l'autre ?
10. Centrer et réduire chacune des variables. Recalculer la moyenne et l'écart-type sur les données transformées. Que constate-t-on ?
11. Calculer la médiane et les quartiles d'ordre 1 et 3 des variables (MEDIANE, CENTILE).
12. Construire le graphique BOXPLOT (boîte de Tukey) pour la variable « âge » (il y a un outil qui le fait directement maintenant dans Excel 2016, voir Insertion / Graphiques recommandés / Tous les graphiques / Boîte à Moustaches).
13. Découper la variable « age » en 10 intervalles de largeurs égales. Comment les bornes des classes ont-elles été déterminées ? Combien y a-t-il d'observations dans chaque classe ? (http://eric.univ-lyon2.fr/~ricco/cours/slides/discretisation.pdf, page 10). Créer explicitement une variable recodée « age_discrete ».
14. A partir des résultats de l'étape précédente, produire le graphique « histogramme de fréquences ».
15. Réitérer les deux questions précédentes pour la variable « hours per week ». Créer la variable « hours_discrete ».
16. Calculer la corrélation entre « age » et « hours per week », tout d'abord en effectuant explicitement tous les calculs intermédiaires dans Excel (http://eric.univ-lyon2.fr/~ricco/cours/cours/Analyse_de_Correlation.pdf ; section 2.3), puis en utilisant la fonction COEFFICIENT.CORRELATION. Vérifier que l'on obtient bien un résultat identique. Peut-on dire que ces deux variables sont liées ?
17. Calculer le V de Cramer entre les variables discrétisées de « age » et « hours per week » (age_discrete et hours_discrete). Peut-on faire un rapprochement avec le coefficient de corrélation calculé précédemment ?
18. L'âge est-il lié à la « classe » ? Quel outil pourrait-on utiliser pour quantifier cela ? (voir http://eric.univ-lyon2.fr/~ricco/cours/cours/Analyse_de_Correlation.pdf ; section 4.6).