Supports :

- Vidéo : Création d'un notebook sous RStudio (https://www.youtube.com/watch?v=u6pqsK8_vO4)
- Tutoriel : Manipulation des données avec R (http://tutoriels-data-science.blogspot.com/p/tutoriels-en-francais.html#293116318583292725)
- D'autres sites très intéressants :
    - Quick-R : http://www.statmethods.net/index.html
    - Aide-mémoire R : http://www.duclert.org/
    - Graphiques sous R : https://www.cyclismo.org/tutorial/R/plotting.html

Remarque introductive

ATTENTION ! Pour chaque exercice, vous devez créer un notebook (**cf. la vidéo**). Pour chaque question, nous devons avoir : son numéro, le code R correspondant à la solution, la sortie de la commande.

Exercice 1

Le tutoriel de référence est celui indiqué en préambule ci-dessus (pour rappel : http://tutoriels-data-science.blogspot.com/p/tutoriels-en-francais.html#293116318583292725).

Fichier « Produits.xlsx »

- Si ce n'est pas déjà fait, installez le package « xlsx » (install.packages ou voir *https://www.youtube.com/watch?v=u1r5XTqrCTQ*)
- Chargez le package en mémoire (library) (tutoriel, page 8)
- Modifiez le répertoire de travail (setwd, tutoriel, page 8)
- Importez le fichier dans un data.frame en utilisant le package (read.xlsx ; tutoriel, page 8)
- Afficher un résumé des données [cf. *print(summary(...))*]. Notez bien les noms de variables !!!
- Extraire le **Nom**, **Catégorie**, **Origine** et **Prix** des produits, pour (voir tutoriel – **page 13**)
    1. catégorie = boissons
    2. catégorie = boissons et prix >100
    3. catégorie = boissons et origine=CEE et prix > 100
    4. catégorie = boissons ou catégorie = condiments
    5. (catégorie = boissons et origine = CEE) OU (catégorie = condiment)
    6. (catégorie = viande ET origine = CEE) OU (catégorie = condiment ET origine = extérieur)
    7. prix > 70 et prix <=100

8. Lister les aliments dont le prix est compris entre 100 FF et 200 FF, et qui sont des « viandes »
9. Lister les 15 produits les moins chers (cf. piste possible **page 16**)
10. Calculer la moyenne de prix des boissons distribuées à Lyon (tapply peut être ? **page 14**)
11. Quel est le nombre de produits : (catégorie = boissons et prix <100) OU (ville = lyon et stock > 20) (nrow peut être, après avoir filtré le tableau de données).

Exercice 2

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

« Classe » joue un rôle particulier, la variable indique les personnes qui ont un revenu annuel supérieur (more) ou inférieur (less) à un seuil quelconque.

Des indications sur les commandes à utiliser sont données. Après, si vous avez des solutions qui semblent plus appropriées, vous pouvez les utiliser également.

1. Si ce n'est pas déjà fait, installer le package « xlsx ». Le charger par la suite (library).
2. Charger le fichier « **census.xlsx** » sous R (read.xlsx)

Que constatez-vous au chargement des données ? Interrompez l'opération. Pour remédier à ce problème vous pouvez :

   a. Essayer le package « readxl » pour voir s'il a plus de succès.
   b. Ou bien, ouvrir le fichier « census.xlsx » dans Excel et l'exporter le au format texte avec séparateur tabulation (census.txt). Puis charger le fichier « census.txt » sous R (read.table, attention aux options)

3. Combien y a-t-il de variables dans le fichier ? Combien y a-t-il d'observations ? (ncol, nrow)
4. Afficher le résumé des données (summary)
5. Essayer de répondre aux différentes questions suivantes : quelle est la *proportion* des hommes (sex = male) ? celle des « classe = more » ? (table ou filtrage + calcul)
6. Construire le diagramme à bandes pour les variables « marital_status » et « relationship » (http://www.statmethods.net/graphs/bar.html).
7. Pour les mêmes variables, construire les diagrammes à secteurs (http://www.statmethods.net/graphs/pie.html).
8. 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 ? (table permet de croiser deux variables, on a une matrice – type matrix - qu'on peut indicer de différentes manières, on peut aussi effectuer des calculs récapitulatifs)

9. Calculer le KHI-2 du tableau croisé entre « classe » et « sex » (chisq.test, pas de correction de continuité). Puis en déduire le v de Cramer (ex. http://www.r-bloggers.com/example-8-39-calculating-cramers-v/ ; l'objet généré par chisq.test possède la propriété statistic que l'on peut exploiter).
10. Confronter le résultat obtenu avec ce que fourni le package « lsr » (http://www.inside-r.org/packages/cran/lsr/docs/cramersV). Vos résultats concordent-ils ?
11. 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 ? (*il faut appliquer un which.max pour chaque modalité ligne (et colonne pour la 2ème partie de la question)* ; *avant de se lancer dans une boucle, voir du côté de apply* : « Tableau et matrices sous R », http://tutoriels-data-science.blogspot.com/p/tutoriels-en-francais.html#1340916549748592669, *pages 12 et 13*).
12. Penchons-nous maintenant sur la variable « age ». Calculer sa moyenne et son écart-type (mean, sd).
13. Centrer et réduire « age » (scale). Recalculer la moyenne et l'écart-type sur les données transformées. Que constate-t-on ?
14. Calculer la médiane et les quartiles d'ordre 1 et 3 des variables (median, quantile).
15. Construire le graphique BOXPLOT (boîte de Tukey) pour la variable « âge » (boxplot). Que remarque-t-on ?
16. Produire l'histogramme de la variable âge (http://www.statmethods.net/graphs/density.html)
17. Calculer la corrélation entre « age » et « hours per week » (cor). Peut-on dire que ces deux variables sont liées ? Réaliser le graphique nuage de points entre ces deux variables pour affiner votre réponse (plot). Que conclure ?
18. Construire le boxplot de « âge » selon « relationship ». Il y a des choses à remarquer dans ce graphique ? (http://www.statmethods.net/graphs/boxplot.html)
19. Calculer la moyenne de l'âge pour chaque valeur de « relationship » (tapply). Le calcul confirme l'impression laissée par le graphique précédent ?
20. Quelle est le nombre de personnes travaillant pour le gouvernement (workclass contenant le terme ''gov'') ? (grep + table)
21. « 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 » ?

22. Quelle est la proportion de personnes qui ont au moins le niveau « Bachelors » ?