Nous travaillons avec **Visual Studio Code** (https://www.youtube.com/watch?v=OKVmfEqa1jQ), **déjà installé dans notre salle informatique**. Créez un Notebook (.ipynb) pour chaque exercice.

Supports de référence – PANDAS / PYTHON :

- **Vidéo** : https://www.youtube.com/watch?v=5jggoPitXjw
- **Tutoriel** : « Manipulation de données avec Pandas », http://tutoriels-data-science.blogspot.com/p/tutoriels-en-francais.html#5944076168611428 37

Exercice 1

Fichier « **produits.xlsx** »

- Modifiez le répertoire par défaut (os.chdir ; **Vidéo, 04:28**)
- Importez la librairie **Pandas** et chargez le fichier dans un Data.frame (read_excel) (Vidéo, 04 :28 ; voir aussi http://pandas.pydata.org/pandas-docs/stable/generated/pandas.read_excel.html).
- Affichez les dimensions de la matrice des données (shape ; **Vidéo, 07:31**).
- Affichez les informations concernant l'ensemble de données importé (info ; **Vidéo, 07:31**). Combien y a-t-il d'observations et de variables dans le fichier ? Notez bien les noms des variables.
- Affichez les 10 premières lignes des données (head ; **Vidéo, 08:15**).
- Affichez la liste des variables (columns ; **Vidéo, 08:51**).
- Extraire le **Nom**, **Catégorie**, **Origine** et **Prix** des produits, pour (voir **Tutoriel** – « **Restrictions avec les conditions »,** par exemple instruction n°39)
  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 et 200, et qui sont des « viandes »
  9. Lister les 15 produits les moins chers (cf. piste possible **instruction n°22**)
  10. Calculer la moyenne de prix des boissons distribuées à Lyon (pivot_table)
  11. Quels sont les 5 produits les moins chers vendus à Lyon ?

Exercice 2

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. Chargez le fichier « **census.xlsx** » (read_excel)
2. Combien y a-t-il de variables dans le fichier ? Combien y a-t-il d'observations ? (info)
3. Affichez les10 premières lignes des observations (head)
4. Afficher le résumé des données (describe), des variables qualitatives d'une part, des quantitatives d'autre part.
5. Essayer de répondre aux différentes questions suivantes : quelle est la *proportion* des hommes (sex = male) ? (0.668442) celle des « classe = more » ? (value_counts) (0.2392818)
6. Construire le diagramme à bandes pour les variables « marital_status » et « relationship » (https://pandas.pydata.org/pandas-docs/version/0.25/user_guide/visualization.html ; bar).
7. Pour les mêmes variables, construire les diagrammes à secteurs (pie).
8. Croiser les variables « classe » et « sex ». On cherche à savoir si le niveau de revenu est différent selon que l'on est un homme ou une femme ? Quelle est la proportion des « more » parmi les hommes ? (0.303767) Parmi les femmes ? (0.109251) (crosstab). Conclusion ?
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). (Husband → Married-civ-spouse, Not-in-family → Never-married, etc.). Le résultat vous paraît logique ?
10. Nous souhaitons quantifier l'intensité de la liaison entre ces deux variables. Calculez le KHI-2 (https://docs.scipy.org/doc/scipy/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/) (0.4881598..). Comment pourrait-on qualifier la relation entre ces deux variables ? Est-ce cohérent avec ce que nous avons pu déduire dans la question précédente ?
11. Penchons-nous maintenant sur la variable « age ». Calculer sa moyenne et son écart-type (mean, std) (38.64,13.71) (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) (0.07155). 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) (43.90, 38.42, 33.42, ...). Le calcul confirme l'impression laissée par le graphique précédent ?
19. Calculer le carré du rapport de corrélation entre « âge » (Y) et « relationship » (X)

$$\eta^2_{Y/X} = \frac{SCE}{SCT} = \frac{\sum_{k=1}^{K} n_k(\bar{y}_k - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Où n est l'effectif total, n_k les effectifs conditionnellement aux modalités de X, y_i l'observation pour l'individu n°i, $\bar{y}$ la moyenne globale, $\bar{y}_k$ les moyennes conditionnelles (SCT = 9181037.53, SCE = 2060875.29, η^2 = 0.22447) (j'ai transféré les valeurs dans des vecteurs Numpy pour réaliser facilement mes calculs... la propriété .values permet de récupérer les valeurs d'un data.frame Pandas ou d'une Series sous forme de matrice ou de vecteur ; https://tutoriels-data-science.blogspot.com/p/tutoriels-en-francais.html#3846468465678573627).
20. On s'intéresse à l'influence du niveau d'instruction (education) sur le revenu (classe). Créez une variable qui permet d'identifier les personnes instruites, ayant un des niveaux suivants : "Bachelors", "Masters", "Prof-school", "Doctorate". Combien d'observations répondent à ce critère ? (isin) (12110)
21. Quelle est la proportion de classe = more parmi ces individus (0.480595), quelle est cette même proportion chez les autres (qui n'ont pas ce niveau d'études) (0.159724). Est-ce que le niveau d'instruction a un impact sur le revenu ? (crosstab)
22. Quel est l'âge moyen parmi ces personnes instruites ? (40.87) Chez les autres ? (37.90)
23. Ces écarts sont-ils significatifs à 5 % ? Effectuez un test de comparaison de moyennes (hypothèse de variances égales dans les populations) pour le vérifier (t de Student = 20.714, pvalue ≈ 0.0) (https://tutoriels-data-science.blogspot.com/p/tutoriels-en-francais.html#7873383569118807 37, cf ttest_ind)