

Cours de R et Rstudio.

Chargé de cours : Abdelghani Maddi, abdelghani.maddi@cnrs.fr

M2 Chargé(e) d'études sociologiques : conseil, enquêtes, évaluation

Dossier 1 : Manipulation des données avec dplyr et représentation avec ggplot2

Exercice 1 : Analyse des occupations selon les tranches d'âge

Contexte : Dans cet exercice, vous allez explorer comment les occupations varient selon les tranches d'âge. Vous utiliserez le jeu de données `hdv2003`, en recodant les âges en différentes tranches et en calculant la répartition des occupations pour chaque groupe d'âge.

Objectifs :

- Recoder une variable continue (l'âge) en catégories.
- Manipuler les colonnes et lignes des tableaux (filtrer, sélectionner, regrouper, etc.).
- Comptabiliser et analyser les occurrences d'une variable catégorielle (occupation) au sein des différentes tranches d'âge.
- Visualiser les résultats de manière claire et interpréter les tendances observées.

Questions :

1. **Préparation des données :**
 - À partir du jeu de données `hdv2003`, créez une nouvelle variable `age_rec` qui regroupe les individus en tranches d'âge : [18, 30), [30, 40), [40, 50), [50, 64), et [64, 97).
 - Quelles sont les principales caractéristiques des tranches d'âge définies ?
2. **Sélection des variables pertinentes :**
 - Sélectionnez les colonnes `id`, `age_rec` (tranche d'âge), et `occup` (occupation). Pourquoi est-il important d'utiliser des données sans doublons dans ce contexte ?
 - Retirez les doublons dans les données. Expliquez l'impact de cette

Cours de R et Rstudio.

Chargé de cours : Abdelghani Maddi, abdelghani.maddi@cnrs.fr

M2 Chargé(e) d'études sociologiques : conseil, enquêtes, évaluation

opération sur le jeu de données.

3. Analyse par groupe :

- Comptez le nombre d'individus dans chaque tranche d'âge et chaque occupation. Quelle est l'occupation la plus courante dans chaque tranche d'âge ?
- Quelles occupations semblent être les moins représentées ou absentes dans certaines tranches d'âge ?

4. Calcul des proportions :

- Calculez la proportion de chaque occupation dans chaque tranche d'âge par rapport au total d'individus dans cette tranche.
- Comment les proportions varient-elles selon les tranches d'âge ? Existe-t-il une tranche d'âge où une occupation domine nettement ?

5. Visualisation :

- Représentez graphiquement les proportions d'occupation pour chaque tranche d'âge à l'aide d'un diagramme en barres.
- Que pouvez-vous conclure sur la répartition des occupations selon les différentes tranches d'âge ? Identifiez des tendances claires et commentez.

Questions pour aller plus loin :

- Quels sont les facteurs, en dehors de l'âge, qui pourraient influencer les choix professionnels des individus ?

Cet exercice vous permettra d'explorer la relation entre l'âge et l'occupation, tout en développant des compétences en manipulation de données et en analyse descriptive dans R.

Cours de R et Rstudio.

Chargé de cours : Abdelghani Maddi, abdelghani.maddi@cnrs.fr

M2 Chargé(e) d'études sociologiques : conseil, enquêtes, évaluation

Exercice 2 : Lien entre le niveau d'études et le statut professionnel

Contexte : Cet exercice vous invite à analyser la relation entre le niveau d'éducation et le type de profession exercée par les individus. Vous utiliserez le jeu de données `hdv2003` pour regrouper les niveaux d'éducation en trois catégories principales et observer la répartition des qualifications professionnelles au sein de chaque groupe.

Objectifs :

- Filtrer et préparer un jeu de données en excluant les valeurs manquantes.
 - Créer des catégories regroupant les niveaux d'études.
 - Analyser la répartition des professions selon ces catégories.
 - Visualiser les résultats et interpréter les tendances.
-

Questions :

1. **Filtrage des données :**
 - Filtrez le jeu de données `hdv2003` pour conserver uniquement les observations où le niveau d'éducation (`nivetud`) et la qualification professionnelle (`qualif`) sont renseignés.
 - Combien d'observations restent après cette étape de filtrage ?
2. **Création des catégories de niveau d'études :**
 - Créez une nouvelle variable `cat_nivetud` qui regroupe les niveaux d'études en trois catégories : *Primaire*, *Secondaire* et *Supérieur*.
 - Expliquez pourquoi il peut être pertinent de regrouper les niveaux d'études de cette façon. Quels avantages présente une telle simplification ?
3. **Analyse des professions par niveau d'études :**
 - Sélectionnez les colonnes `id`, `cat_nivetud` et `qualif`, puis retirez les doublons.
 - Pour chaque niveau d'études, comptez combien d'individus se trouvent dans chaque qualification professionnelle (`qualif`). Quelle est la qualification la plus fréquente dans chaque catégorie de niveau d'études ?

Cours de R et Rstudio.

Chargé de cours : Abdelghani Maddi, abdelghani.maddi@cnrs.fr

M2 Chargé(e) d'études sociologiques : conseil, enquêtes, évaluation

4. Calcul des proportions :

- Calculez la proportion de chaque qualification professionnelle au sein de chaque groupe d'éducation, en pourcentage du total dans chaque niveau d'études.
- Les proportions de professions varient-elles significativement en fonction du niveau d'études ? Donnez quelques exemples de différences notables.

5. Visualisation :

- Utilisez un graphique à barres pour représenter la répartition des qualifications professionnelles pour chaque niveau d'éducation.
- Que pouvez-vous déduire de la répartition des professions selon les niveaux d'éducation ? Identifiez les tendances ou les anomalies dans les données.

Questions pour aller plus loin :

- Quelles pourraient être les implications de ces résultats sur le marché de l'emploi ?
- Pensez-vous que la répartition observée dans ces données est représentative de l'ensemble de la population française ? Pourquoi ou pourquoi pas ?

Cet exercice vous permettra de comprendre l'impact du niveau d'éducation sur le type de profession exercée, tout en vous familiarisant avec des techniques de recodage, de manipulation des données, et d'analyse descriptive dans R.