

TP 1 Préparer son fichier de données et découverte d'Orange

Objectifs du TP

Les objectifs de ce TP sont les suivants :

- Apprendre à préparer un fichier de données avec des expressions régulières
- Prise en main du logiciel Orange
 - Pour faire une jointure (File et Merge Data)
 - Pour sélectionner les colonnes à analyser (Select Columns)
 - Pour faire quelques visualisations simples (Box Plot et Distribution)

Contexte

Nous allons exploiter une base de données existante : la base de données publique du médicament (BDPM). Elle ne peut pas être chargée telle quelle dans le logiciel Orange : une étape de formatage des fichiers sera d'abord nécessaire.

Jointure des fichiers avec Orange

Les données sont présentes dans 2 fichiers différents : vous utiliserez ensuite Orange pour combiner les deux fichiers à l'aide d'une jointure, puis conserver uniquement les colonnes qui nous intéressent.

Nous souhaitons travailler sur un fichier contenant les informations suivantes :

En provenance du fichier BDPM :

- Id
- Présentation (ex : bain de bouche, bâton pour application,...)
- Laboratoire (ex : 3M ESPE AG (Allemagne), ABBVIE, ACCORD HEALTHCARE France,...)
- Commercialisation (ex : Commercialisée, Non commercialisée)
- Nom (ex : DAFALGAN)

En provenance du fichier CIP :

- Date commercialisation
- Taux remboursement (15%, 35%, 65%, 100%,...)
- Prix
- Remboursement

Visualisations

Enfin, une fois le fichier prêt vous pourrez effectuer quelques visualisations simples, pour répondre à l'aide d'un graphique aux questions suivantes :

- Combien y-a-t-il de médicaments commercialisés et non commercialisés ?
- Combien y-a-t-il de médicaments par taux de remboursement ? (0%, 15%, 35%, 65% et 100%)
- Quels sont les laboratoires qui proposent les médicaments remboursés les plus chers ?

Les instructions pour parvenir au résultat sont données dans la suite du TP.

Matériel

Pour ce TP, vous aurez besoin des logiciels suivants :

- Un éditeur de texte qui gère les expressions régulières (par exemple : Notepad++, Atom,...)
- Le logiciel Orange (<https://orange.biolab.si/>) (TP testé avec la version 3.20). Nous l'utiliserons également pour les TP suivants

Ce TP utilise les données issues de la BDPM (base de données publique du médicament) qui sont disponibles sur Moodle (fichiers CIS_bdpm.tsv et CIS_CIP_bdpm.tsv)

1. Préparation des fichiers

Nous allons préparer les fichiers. Je vous conseille l'utilisation de Notepad++. Pendant cette étape, vous pouvez en profiter pour télécharger et installer Orange sur votre ordinateur personnel (il est déjà installé en salle TP).

- Ouvrir le fichier CIP et remplacer « 100 % » par « 100% », « 65 % » par « 65% », « 30 % » par « 30% » et « 15 % » par « 15% »
- S'inspirer du lien suivant (<https://nliataud.developpez.com/tutoriels/web/notepadplusplus-guide-pratique/expressions-regulieres/>) pour transformer les nombres à virgule par des points

Exemple : 7,11 8,13 -> 7.11 8.13

2,850,07 2,851,09 -> 2850,07 2851,09

Astuce : vous pouvez lancer les expressions régulières plusieurs fois pour gérer les nombres > 999

- Utiliser une expression régulière pour trouver toutes les dates sous la forme jj/mm/aaaa et les transformer sous la forme jj-mm-aaaa

2. Utilisation d'Orange

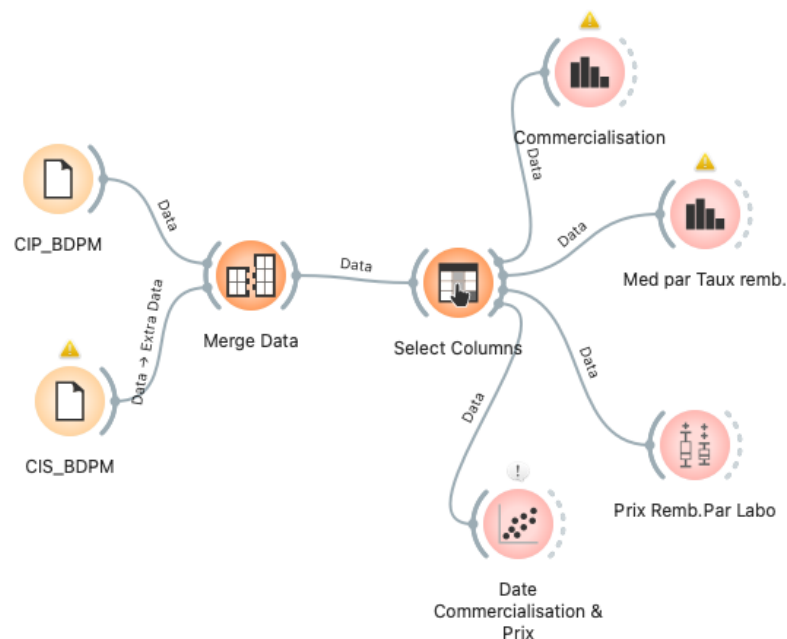
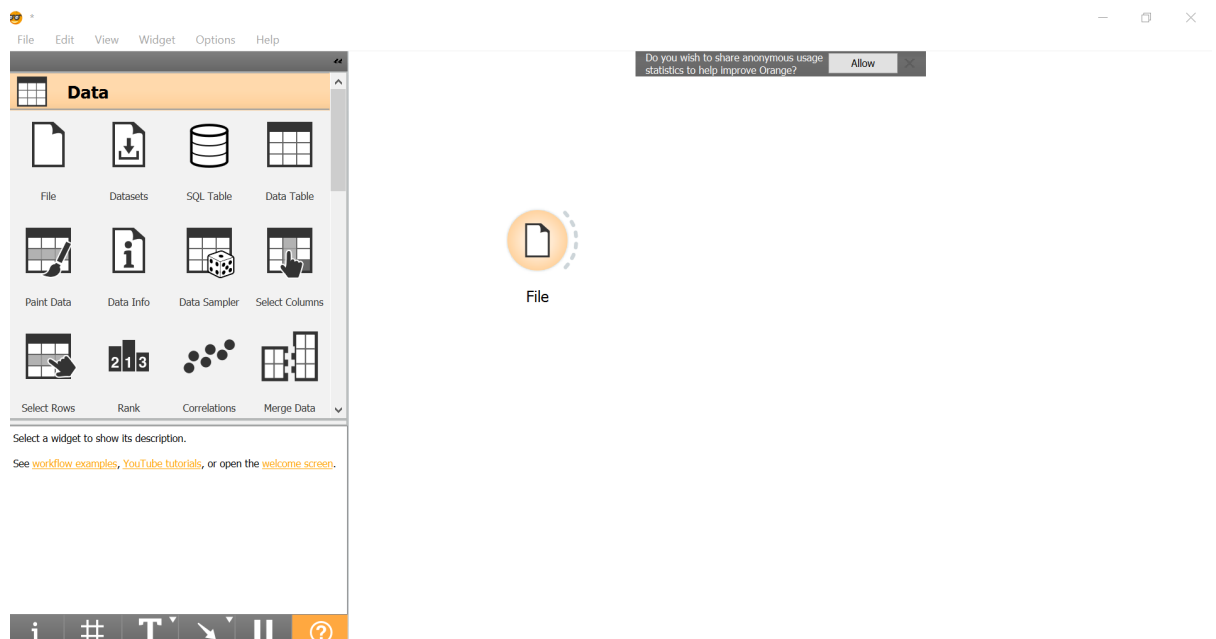


Figure 1 - Ce que vous devriez obtenir à la fin du TP

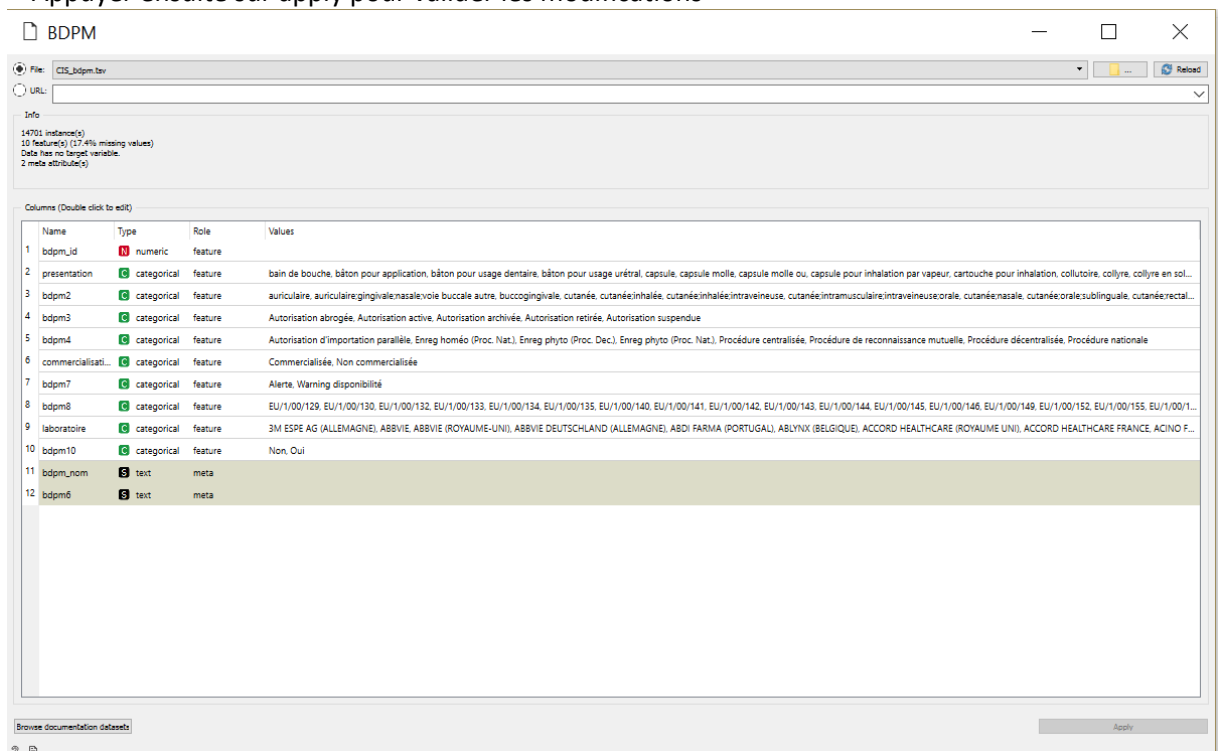
Orange dispose de composants sur la gauche, que vous pouvez glisser dans la zone de travail à droite. Chaque composant peut être renommé avec F2. Il est possible de les relier en maintenant un clic de souris sur un quart de cercle entourant un composant et en l'amenant à un autre composant.

a) Chargement du fichier BDPM

Charger le fichier CIS_bdpm à l'aide du composant file disponible dans le menu Data > File.
Un composant file apparaît dans la zone de travail (illustration 1), cliquer dessus.



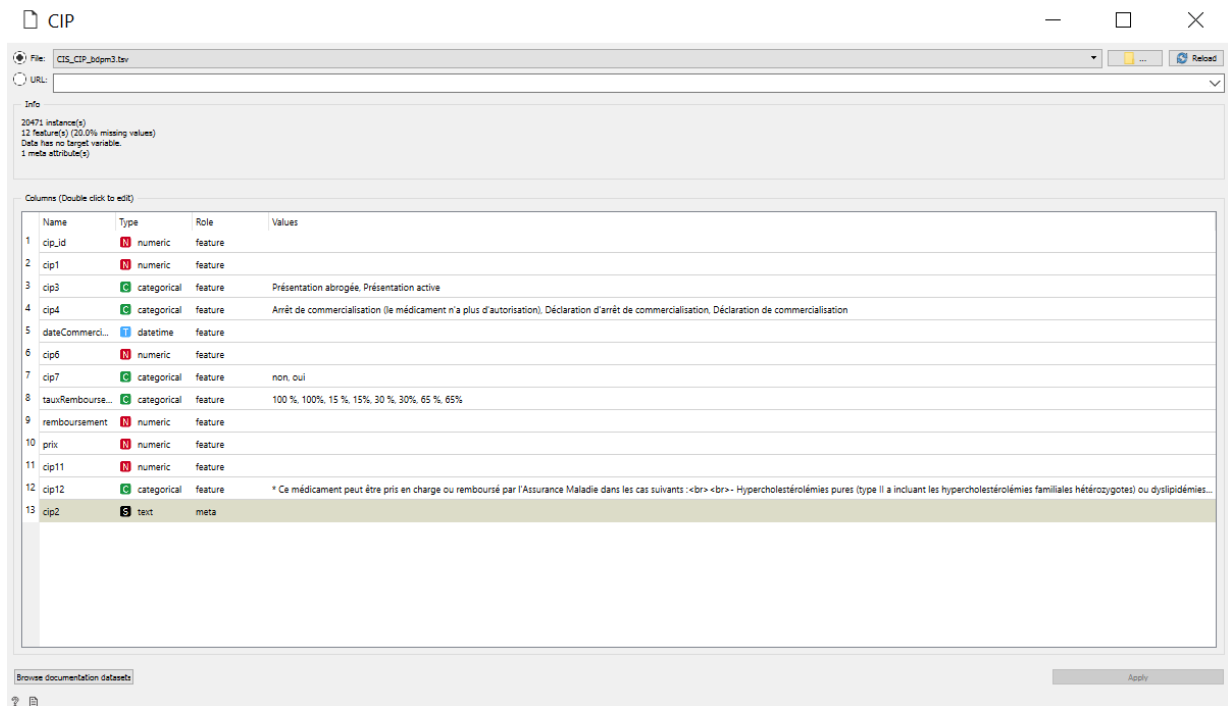
Choisir le fichier CIS_bpdm et le charger, puis renommer les colonnes qui nous intéressent.
Appuyer ensuite sur apply pour valider les modifications



b) Chargement du fichier CIP

Suivre la même démarche pour le fichier CIP. La colonne « id » doit avoir un autre nom que celle utilisée dans l'autre fichier (ici par exemple cip_id).

S'assurer que les colonnes prix et remboursement ont bien le type numeric, et que la colonne dateCommercialisation a bien le type datetime. Sinon, cela signifie qu'il faut retravailler votre fichier .tsv



File: CIS_CIP_bdpm3.tsv

Info

20471 instance(s)
12 feature(s) (20.0% missing values)
Date has no target variable
1 meta attribute(s)

Columns (Double click to edit)

	Name	Type	Role	Values
1	cip_id	numeric	feature	
2	cip1	numeric	feature	
3	cip3	categorical	feature	Présentation abrogée, Présentation active
4	cip4	categorical	feature	Arrêt de commercialisation (le médicament n'a plus d'autorisation), Déclaration d'arrêt de commercialisation, Déclaration de commercialisation
5	dateCommercial...	datetime	feature	
6	cip6	numeric	feature	
7	cip7	categorical	feature	non, oui
8	tauxRembourse...	categorical	feature	100 %, 100%, 15 %, 15%, 30 %, 30%, 65 %, 65%
9	remboursement	numeric	feature	
10	prix	numeric	feature	
11	cip11	numeric	feature	
12	cip12	categorical	feature	* Ce médicament peut être pris en charge ou remboursé par l'Assurance Maladie dans les cas suivants : - Hypercholestérolémies pures (type II a incluant les hypercholestérolémies familiales hétérozygotes) ou dyslipidémies...
13	cip2	text	meta	

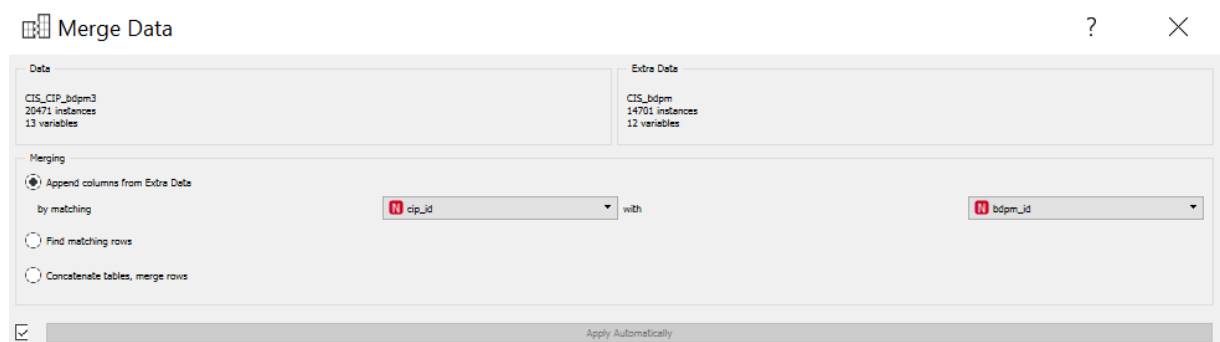
Browse documentation datasets

Apply

c) Fusion des deux fichiers

Nous allons maintenant utiliser le composant Merge Data pour joindre les deux fichiers.

- 1) Relier le fichier CIP au composant Merge Data
- 2) Relier le fichier BDPM au composant Merge Data
- 3) Cliquer sur le composant Merge Data, choisir « append columns from Extra Data » et choisir vos colonnes id (ici « cip_id » et « bdpm_id »)



Merge Data

Data: CIS_CIP_bdpm3, 20471 instances, 13 variables

Extra Data: CIS_bdpm, 14701 instances, 12 variables

Merging

☒ Append columns from Extra Data

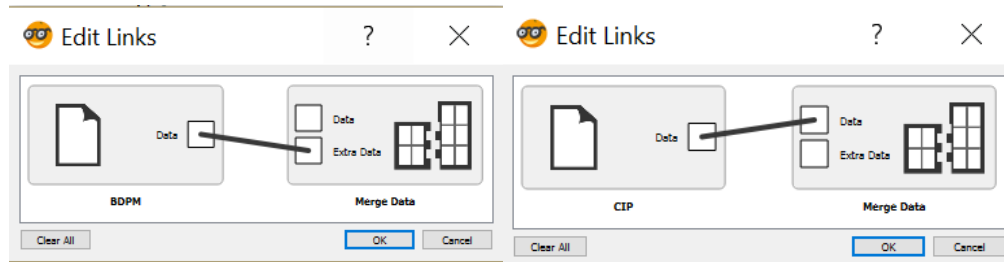
by matching: with

☐ Find matching rows

☐ Concatenate tables, merge rows

☒ Apply Automatically

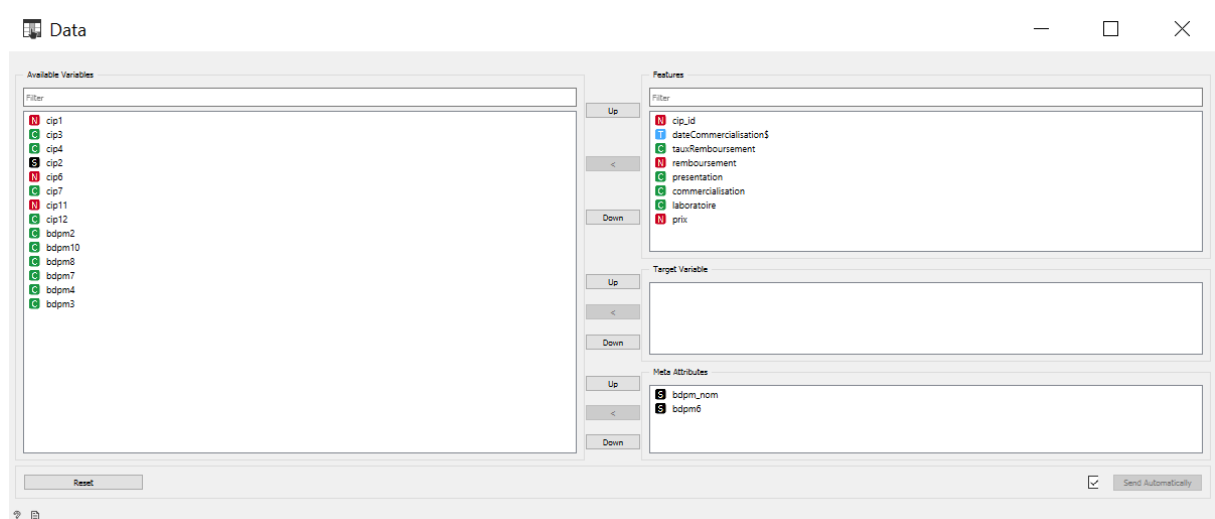
Note : si l'étape 3 ne fonctionne pas, cliquez sur les liens pour les configurer et obtenir le résultat suivant :



Note2 : à tout moment, vous pouvez visualiser le résultat avec le composant « Data Table », en le reliant à une sortie de composant puis en cliquant sur le composant « Data Table »

d) Suppression des colonnes inutiles

Pour ne pas trop polluer le reste du TP, nous allons nous débarrasser des colonnes inutiles. Ajouter un composant « Select Column ». Ce composant permet de filtrer les colonnes à conserver. Celles-ci se situent sur la droite. Basculer toutes les colonnes demandées au début du TP sur la droite, les autres sur la gauche.



e) Visualisations

Utiliser les composants « Distributions », « Box Plot », « Scatter Plot » pour répondre aux questions suivantes :

- Combien y-a-t-il de médicaments commercialisés et non commercialisés ?
- Combien y-a-t-il de médicaments par taux de remboursement ? (0%, 15%, 35%, 65% et 100%)
- Quels sont les laboratoires qui proposent les médicaments remboursés les plus chers ?
- A partir de quand apparaissent les médicaments qui coûtent plus de 9000 euros ? Quel est le premier à apparaître ?

Ces composants se situent dans le menu « visualize »