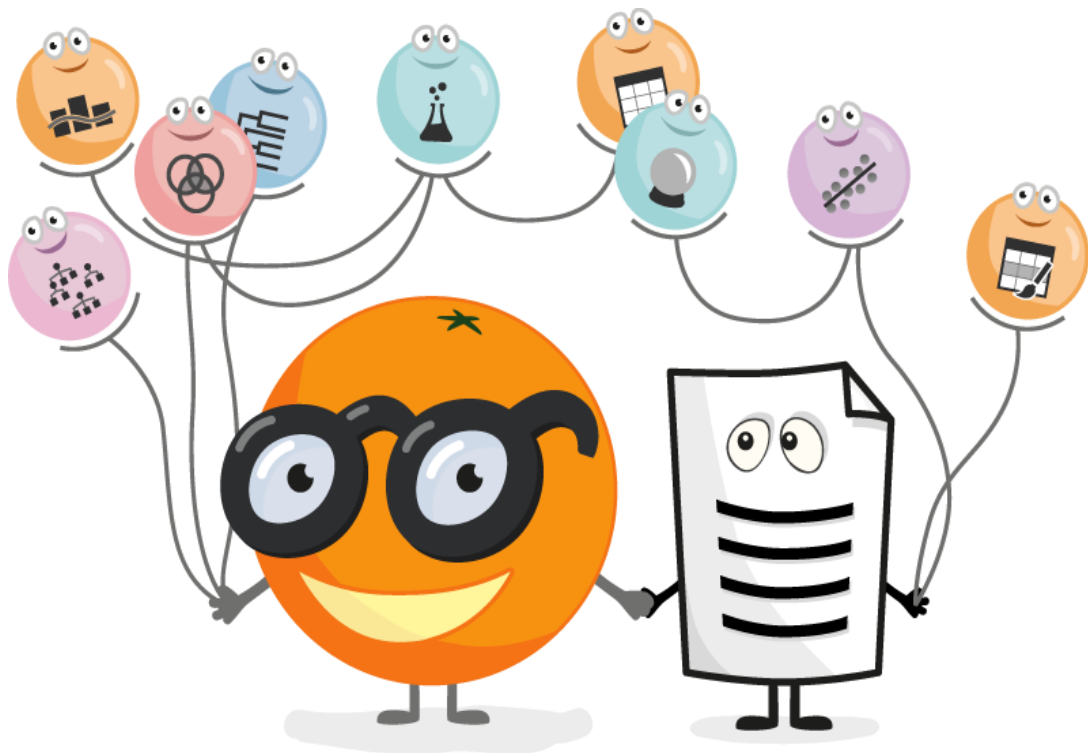


# Rapport du projet :

## Pré-traitement de données

avec Orange

[ DATA MINING ]



## ✓ L'objectif :

Appliquer les étapes et les procédures de **pré-traitement** de données sur des *ensembles de données*, à l'aide de logiciel **Orange Data Mining**. Pour l'objectif de transformer les données brutes dans un seul format utile et efficace.

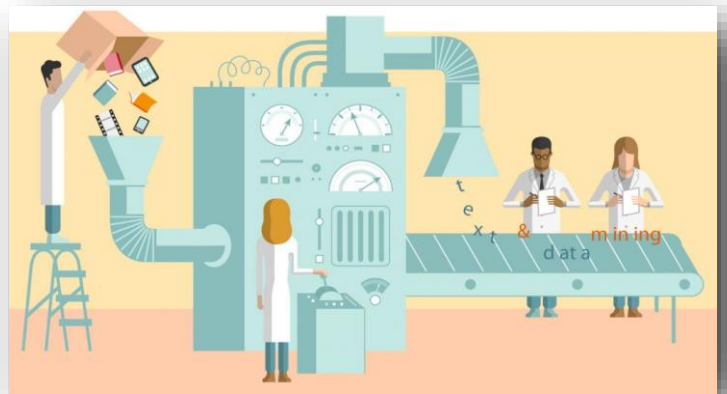
## ✓ Introduction :

Le data mining, dans sa forme et compréhension actuelle, à la fois comme champ scientifique et industriel, est apparu au début des années 90. Cette émergence n'est pas le fruit du hasard mais le résultat de la combinaison de nombreux facteurs à la fois technologiques, économiques et même sociopolitiques.

On peut voir le data mining comme une nécessité imposée par le besoin des entreprises de valoriser les données qu'elles accumulent dans leurs bases. En effet, le développement des capacités de stockage et les vitesses de transmission des réseaux ont conduit les utilisateurs à accumuler de plus en plus de données.

Certains experts estiment que le volume des données double tous les ans. Que doit-on faire avec des données coûteuses à collecter et à conserver ?

Une confusion subsiste encore entre data mining, que nous appelons en français « fouille de données », et knowledge discovery in data bases (KDD), que nous appelons en français « extraction des connaissances à partir des données » (ECD). Le data mining est l'un des maillons de la chaîne de traitement pour la découverte des connaissances à partir des données.

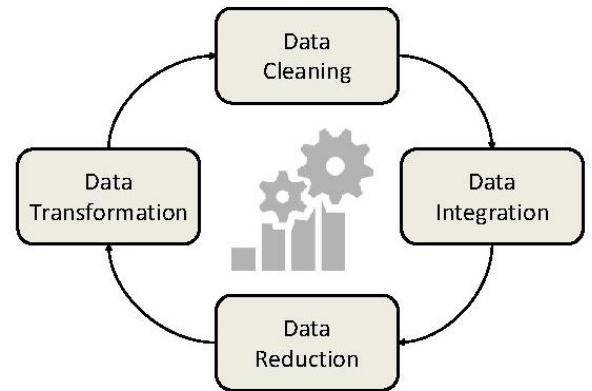


L'ECD, par le biais du data mining, est alors vue comme une ingénierie pour extraire des connaissances à partir des données.

## ✓ Le pré-traitement de données (data preprocessing) :

Le prétraitement des données est le processus de transformation des données en un format compréhensible. C'est aussi une étape importante dans l'exploration de données, car nous ne pouvons pas travailler avec des données brutes.

Le prétraitement des données est une étape très importante de préparation de données, parce que la qualité des données doit être vérifiée avant d'appliquer des algorithmes d'apprentissage automatique ou d'exploration de données.

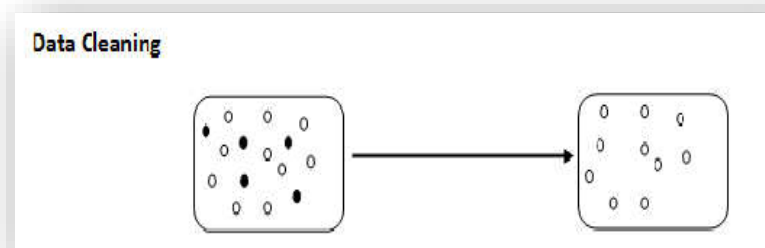


## ✓ Les étapes principales du pré-traitement de données :

Après la partie de collection de données, la première étape du traitement des données. Les données proviennent de toutes les sources disponibles, y compris les *data lakes* et les *data warehouses*, suit la préparation des données ou ce que n'appelée « pré-traitement ».

Voici les étapes principales de cette partie :

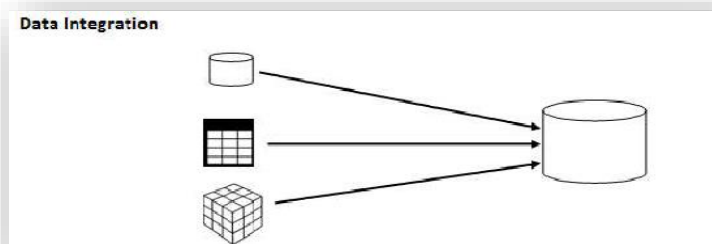
1. **Nettoyage de données** : le processus pour supprimer les données incorrectes, les données incomplètes et les données inexacts des ensembles de données, et il remplace également les valeurs manquantes.



Il existe certaines techniques de nettoyage des données :

- Des valeurs standard comme « Not Available » ou « NA » peuvent être utilisées pour remplacer les valeurs manquantes.
- Les valeurs manquantes peuvent également être remplies manuellement (mais il n'est pas recommandé)
- La valeur moyenne de l'attribut peut être utilisée pour remplacer la valeur manquante.

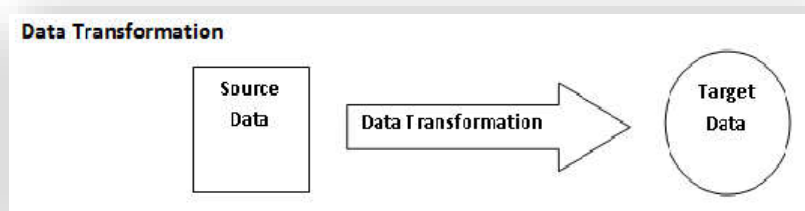
2. **Intégration de données** : le processus de combinaison de plusieurs sources en un seul ensemble de données.



Il y a certains problèmes à considérer lors de l'intégration des données.

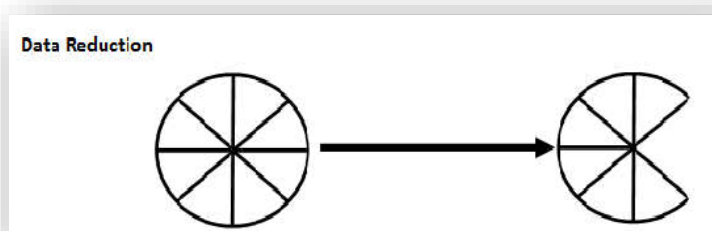
- Intègre des métadonnées (un ensemble de données qui décrivent d'autres données) provenant de différentes sources.
- Problème d'identification de l'entité.
- Détection et résolution des concepts de valeur des données, par exemple la date dans une source représenté comme JJ/MM/AAAA, et dans une autre source nous avons MM/JJ/AAAA.

3. **Transformation de données** : la modification apportée au format ou à la structure des données est appelée transformation des données. Cette étape peut être simple ou complexe en fonction des exigences.



4. **Réduction de données** : Ce processus permet de réduire le volume des données, ce qui facilite l'analyse tout en produisant le même résultat ou presque. Cette réduction permet également de réduire l'espace de stockage.

Il y a quelques-unes des techniques de réduction des données sont réduction de dimensionalité, réduction de numération, compression des données

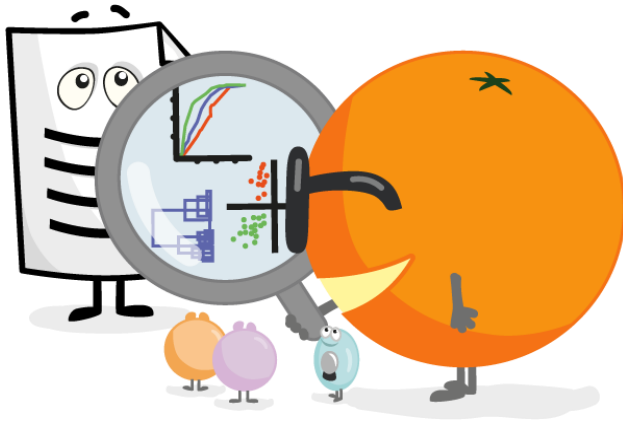


## ✓ Le Logiciel Orange :

Orange est un logiciel libre d'exploration de données (data mining).

Il propose des fonctionnalités de modélisation à travers une interface visuelle, une grande variété de modalités de visualisation et des affichages variés dynamiques.

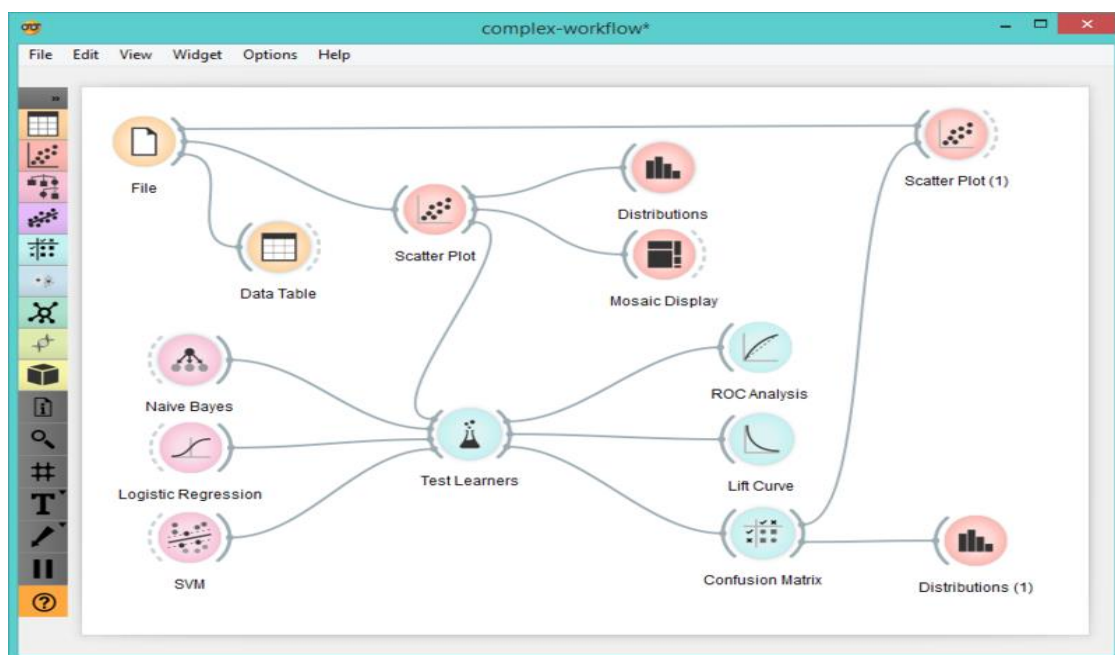
Développé en Python, il existe des versions Windows, Mac et Linux.



Effectuez une analyse de données simple avec une visualisation de données intelligente. Explorez les distributions statistiques, les diagrammes de boîtes et les diagrammes de dispersion, ou plongez plus profondément avec les arbres de décision, les regroupements hiérarchiques, les cartes thermiques, les MDS et les projections linéaires. Même vos données multidimensionnelles peuvent devenir sensibles en 2D,

en particulier avec le classement des attributs et les sélections intelligentes.

Exploration interactive des données pour une analyse qualitative rapide avec des visualisations nettes. L'interface utilisateur graphique vous permet de vous concentrer sur l'analyse exploratoire des données au lieu de coder.



[Orange Data Mining - Data Mining](#)

## ✓ Les widgets utilisent dans ce projet :

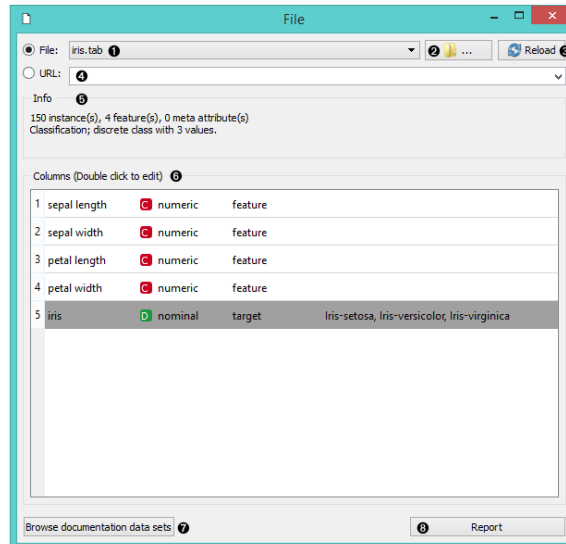
### File :



File

L'icône Fichier lit le fichier de données d'entrée (table de données avec instances de données) et envoie le jeu de données à son canal de sortie. L'historique des derniers fichiers ouverts est conservé dans le widget.

Le widget lit les données à partir d'Excel (.xlsx), de simples tabulations (.txt), de fichiers séparés par des virgules (.csv) ou d'URL, etc.

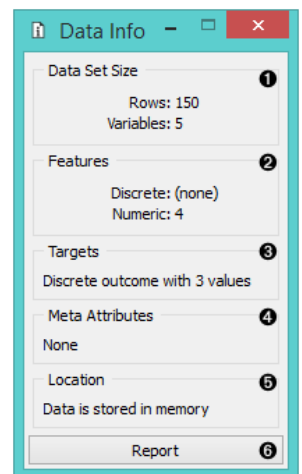


### Data Info :



Data Info

Un simple widget qui présente des informations sur la taille de l'ensemble de données, les caractéristiques, les cibles, les méta-attributs et l'emplacement.

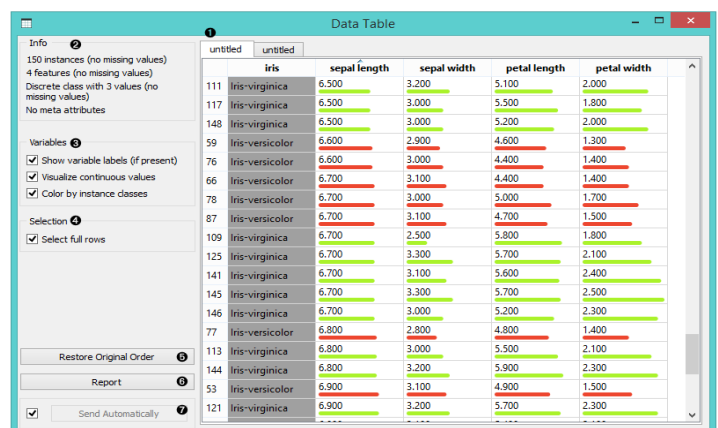


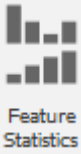
### Data Table :



Data Table

Le widget Data Table reçoit un ou plusieurs ensembles de données dans son entrée et les présente comme une feuille de calcul. Les instances de données peuvent être triées par valeurs d'attribut. Le widget prend également en charge la sélection manuelle des instances de données.





## Feature Statistics :

Cette icône fournit un moyen rapide d'inspecter et de trouver des fonctionnalités intéressantes dans un ensemble de données donné et pour visualiser les entités qui contiennent des données manquantes.



## Select Rows :

Ce widget sélectionne un sous-ensemble à partir d'un ensemble de données d'entrée, en fonction des conditions définies par l'utilisateur. Les instances qui correspondent à la règle de sélection sont placées dans le canal de sortie *Matching Data*.

Les termes de condition sont définis en sélectionnant un attribut, un opérateur dans une liste d'opérateurs et si nécessaire en définissant la valeur à utiliser dans le terme de condition. Les opérateurs sont différents pour les attributs discrets, continus et chaîne.



## Select Columns :

Utilisé pour composer manuellement votre domaine de données. L'utilisateur peut décider quels attributs seront utilisés et comment. Orange distingue les attributs ordinaires, les attributs de classe (optionnels) et les attributs méta.

Les attributs orange ont un type et sont soit discrets, continus ou une chaîne de caractères.

Le type d'attribut est marqué par un symbole apparaissant avant le nom de l'attribut (D, C, S, respectivement).



Save Data

### Save Data :

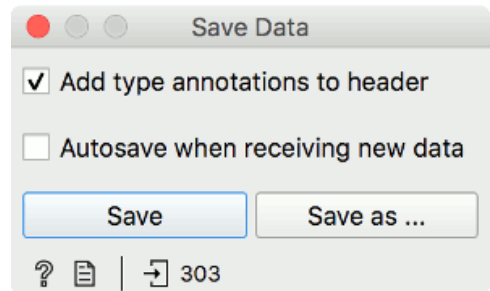
Le widget Enregistrer les données considère un ensemble de données fourni dans le canal d'entrée et l'enregistre dans un fichier de données avec un nom spécifié. Il peut enregistrer les données comme:

- ✓ un fichier délimité par des tabulations (.tab)
- ✓ fichier séparé par des virgules (.csv)
- ✓ pickle (.pkl), utilisé pour stocker le prétraitement des objets Corpus
- ✓ Feuilles de calcul Excel (.xlsx)
- ✓ spectra ASCII (.dat)
- ✓ carte hyperspectrale ASCII (.xyz)
- ✓ formats compressés (.tab.gz, .csv.gz, .pkl.gz)

Le widget n'enregistre pas les données chaque fois qu'il reçoit un nouveau signal dans l'entrée, car

cela écraserait constamment (et, surtout, par inadvertance) le fichier. Au lieu de cela, les données sont enregistrées uniquement après qu'un nouveau nom de fichier est défini ou que l'utilisateur appuie sur le bouton Enregistrer.

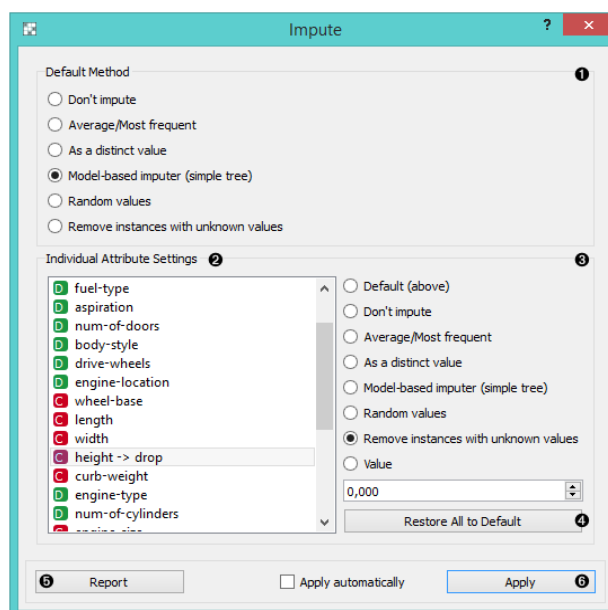
Si le fichier est enregistré dans le même répertoire que le workflow ou dans la soustraction de ce répertoire, le widget se souvient du chemin relatif. Sinon, il stockera un chemin absolu mais désactivera l'enregistrement automatique pour des raisons de sécurité.



Impute

### Impute :

Certains algorithmes et visualisations d'Orange ne peuvent pas gérer les valeurs inconnues dans les données. Ce widget fait ce que les statisticiens appellent l'imputation : il substitue les valeurs manquantes par des valeurs calculées à partir des données ou définies par l'utilisateur.







Concatenate

### Concatenate :

Le widget concaténé plusieurs ensembles d'instances (ensembles de données). La fusion est « verticale », en ce sens que deux ensembles de 10 et 5 instances donnent un nouvel ensemble de 15 instances.



Edit Domain

### Edit Domain :

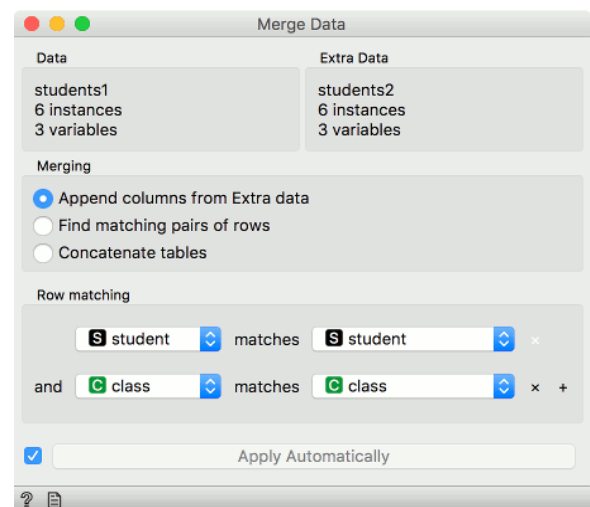
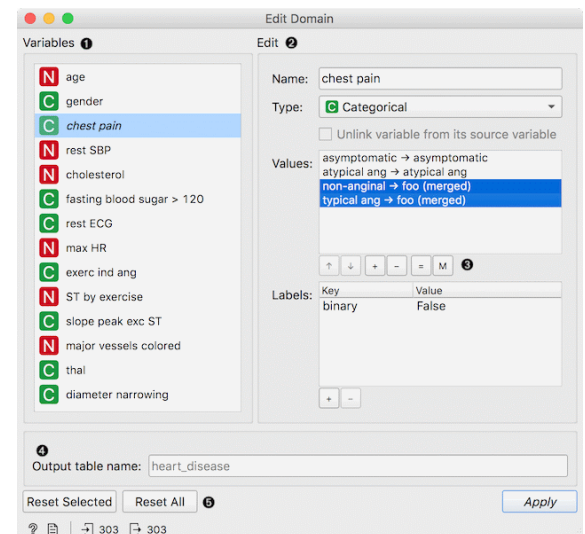
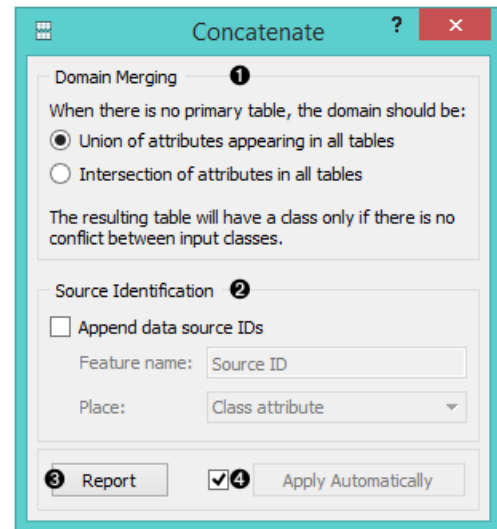
Ce widget peut être utilisé pour modifier le domaine d'un jeu de données - renommer des fonctionnalités, renommer ou fusionner des valeurs de fonctionnalités catégorielles, ajouter une valeur catégorielle et attribuer des étiquettes.



Merge Data

### Merge Data :

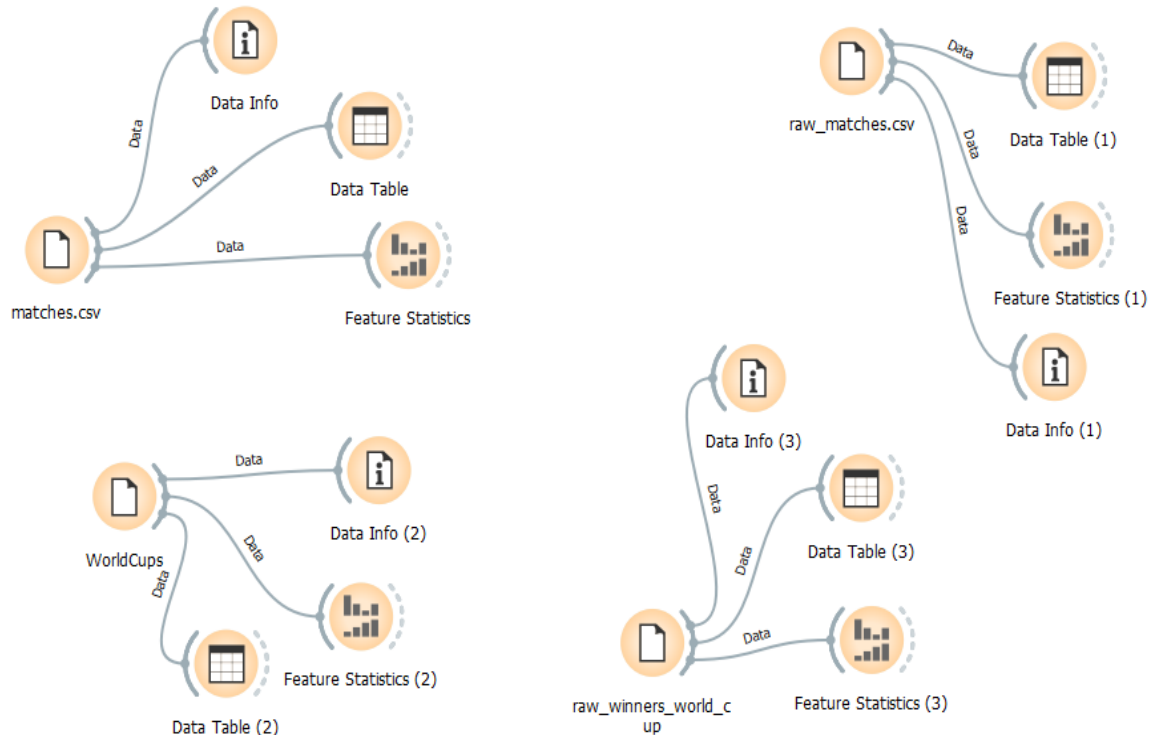
Le widget Fusionner des données est utilisé pour fusionner horizontalement deux ensembles de données, en fonction des valeurs des attributs sélectionnés (colonnes). Dans l'entrée, deux ensembles de données sont nécessaires, des données et des données supplémentaires. Les lignes des deux ensembles de données sont appariées par les valeurs de paires d'attributs, choisies par l'utilisateur. Le widget produit une sortie. Il correspond aux instances des données d'entrée auxquelles les attributs (colonnes) des données d'entrée supplémentaires sont ajoutés. Si la paire d'attributs sélectionnée ne contient pas de valeurs uniques (en d'autres termes, les attributs ont des valeurs dupliquées), le widget donnera un avertissement. Au lieu de cela, on peut faire correspondre par plus d'un attribut. Cliquez sur l'icône plus pour ajouter l'attribut à fusionner. Le résultat final doit être une combinaison unique pour chaque ligne individuelle.



## ✓ La visualisation et la compréhension des données :

Cette étape est très importante pour bien comprendre notre donnée avec ses différents features, dans les différents fichiers et pour extraire les relations entre ils avec quelques remarques sur les problématiques de prétraitement s'ils existent, et essayer de les modifier.

Pour réaliser cette tâche nous avons utilisé le workflow suivant :



## ✓ Quelques problèmes de prétraitement à résoudre :

- 64 (8%) des données dans les entités **attendance**, **time** et **phase** du fichier **matches** sont des données manquantes (Missing data).

|   | Name       | Distribution | Mean     | Median        | Dispersion | Min.     | Max.     | Missing |
|---|------------|--------------|----------|---------------|------------|----------|----------|---------|
| N | attendance |              | 43632.74 | 39873.50      | 0.56       | 2000     | 173850   | 64 (8%) |
| T | time       |              | 17:07:05 | 17:00:00      | ~9 hours   | 11:30:00 | 21:00:00 | 64 (8%) |
| C | phase      |              |          | Group matches | 1.25       |          |          | 64 (8%) |

### [L'étape de nettoyage de données]

- Nous avons 4 fichiers (sources différentes de données).
- Nous avons quelques données répéter avec nom d'entités différents, par exemple nous avons **home – away** dans le fichier **matches** ont les mêmes significations de **team1 – team2** du fichier **raw\_matches**.

- Il y a des entités de même nom dans différents fichiers par exemple **attendance**, dans le fichier **matches** représenté le nombre des participants pour chaque matche et dans le fichier **WorldCups** la somme des participants pour tous les matches d'un mondiale.
- Chaque fichier couvre un intervalle de temps différent :
  - ✓ Le fichier **matches** à les données des mondiales de 1930 jusqu'à 2010.
  - ✓ Le fichier **raw\_matches** à les données des mondiales de 1950 jusqu'à 2014.
  - ✓ Le fichier **Worldcups** à les données des mondiales de 1930 jusqu'à 2018.
  - ✓ Le fichier **raw\_winners\_world\_cup** à les données des mondiales de 1950 jusqu'à 2010.
- Dans le fichier **matches** la date est représenté avec des manières différentes.
- Les données de fichier **raw\_winners\_world\_cup** sont déjà existent dans le fichier **WorldCup**.

[L'étape d'intégration de données]

### ✓ Nettoyage de données :

#### • Probleme :

Les données manquantes des entités **attendance**, **time** et **phase** du fichier **matches** sont les données de 2010 World Cup South Africa.

|    | year | home        | away           | phase | attendance | time |
|----|------|-------------|----------------|-------|------------|------|
| 64 | 2010 | Netherlands | Spain          | ?     | ?          | ?    |
| 63 | 2010 | Uruguay     | Germany        | ?     | ?          | ?    |
| 62 | 2010 | Germany     | Spain          | ?     | ?          | ?    |
| 61 | 2010 | Uruguay     | Netherlands    | ?     | ?          | ?    |
| 60 | 2010 | Paraguay    | Spain          | ?     | ?          | ?    |
| 59 | 2010 | Argentina   | Germany        | ?     | ?          | ?    |
| 58 | 2010 | Uruguay     | Ghana          | ?     | ?          | ?    |
| 57 | 2010 | Netherlands | Brazil         | ?     | ?          | ?    |
| 56 | 2010 | Spain       | Portugal       | ?     | ?          | ?    |
| 55 | 2010 | Paraguay    | Japan          | ?     | ?          | ?    |
| 54 | 2010 | Brazil      | Chile          | ?     | ?          | ?    |
| 53 | 2010 | Netherlands | Slovakia       | ?     | ?          | ?    |
| 52 | 2010 | Argentina   | Mexico         | ?     | ?          | ?    |
| 51 | 2010 | Germany     | England        | ?     | ?          | ?    |
| 50 | 2010 | USA         | Ghana          | ?     | ?          | ?    |
| 49 | 2010 | Uruguay     | Korea Republic | ?     | ?          | ?    |
| 48 | 2010 | Switzerland | Honduras       | ?     | ?          | ?    |
| 47 | 2010 | Chile       | Spain          | ?     | ?          | ?    |
| 46 | 2010 | Spain       | Honduras       | ?     | ?          | ?    |
| 45 | 2010 | Chile       | Switzerland    | ?     | ?          | ?    |
| 44 | 2010 | Spain       | Switzerland    | ?     | ?          | ?    |
| 43 | 2010 | Honduras    | Chile          | ?     | ?          | ?    |
| 42 | 2010 | Korea DPR   | Côte d'Ivoire  | ?     | ?          | ?    |
| 41 | 2010 | Portugal    | Brazil         | ?     | ?          | ?    |

L'entité **attendance** représenté le nombre des participants pour chaque matche, **time** représenté l'heure de départ de matche et l'entité **phase** classier chaque matche dans des catégories bien ordonnées, par exemple Group matches, Final et Quarter-Final, etc.

- **Solution :**

2010 World Cup South Africa, 2006 World Cup Germany et 2010 World Cup Korea / Japan, ont le même nombre des matches, 64 matches.

- ✓ Les 48 premières sont de phase Group matches, 8 suivants sont de phase Round of 16, Quarter-finals pour les 4 matches suivants, Semi-finals pour les 2 suivants, puis un match de phase Third place et finalement un seule matche de Final.

Alors on peut remplir ces différentes catégories de l'entité **phase** pour les données manquantes de 2010 World Cup South Africa on utilise l'entité **new\_match\_number** qui contient le nombre de matche.

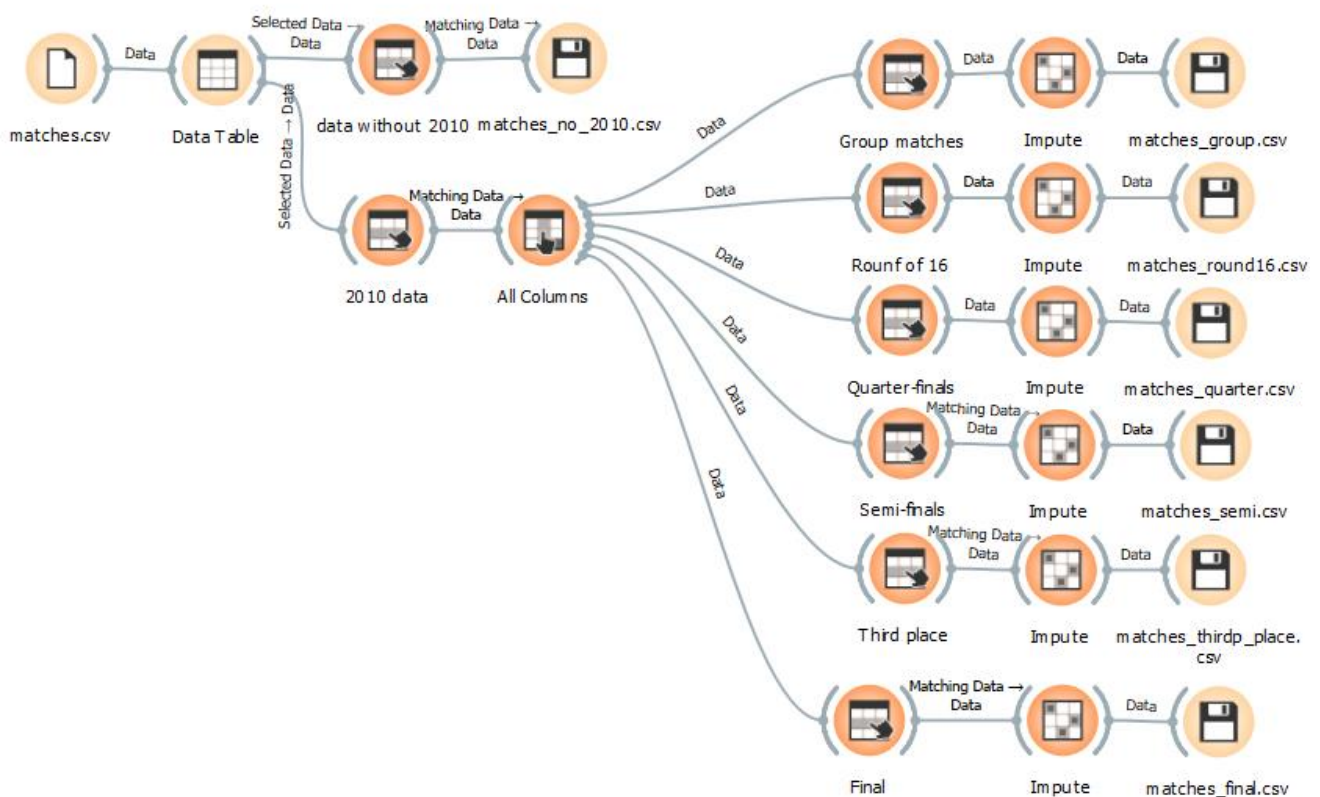
D'après le fichier **WorldCups** le nombre total des participants de 2010 World Cup South Africa est 3178856, donc on peut divise ce nombre sur le nombre des matches (64 matches) pour extraire une moyenne pour remplir les cases de données manquantes.

- ✓  $Moyenne = 3178856 / 64 = 49\,669,625$

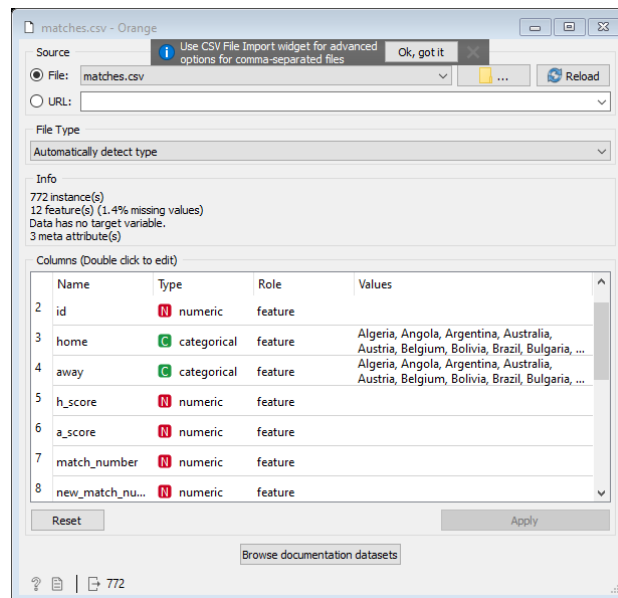
Pour times on va considérons un temp par default.

- ✓ On va remplir les cases vides avec 00 :00 :00.

❖ Pour appliquer ces modifications on va utiliser le *workflow* suivant sur :

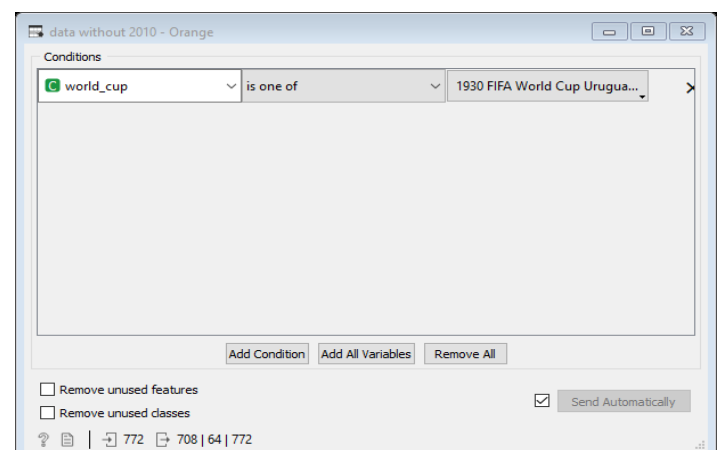
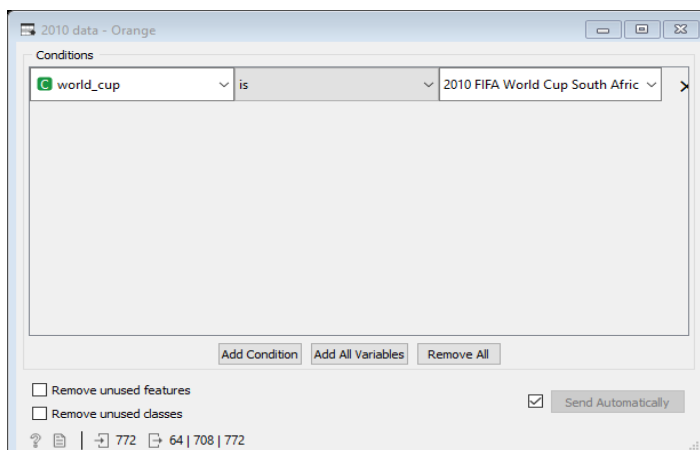


## 1. Importer le fichier matches.csv avec l'icône File.

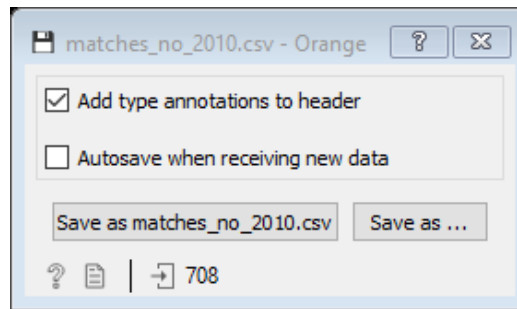


|    | stadium           | phase | world_cup         | id        | home           |
|----|-------------------|-------|-------------------|-----------|----------------|
| 1  | Johannesburg S... | ?     | 2010 FIFA Worl... | 300061454 | South Africa   |
| 2  | Cape Town Cap...  | ?     | 2010 FIFA Worl... | 300061453 | Uruguay        |
| 3  | Tshwane/Pretor... | ?     | 2010 FIFA Worl... | 300061452 | South Africa   |
| 4  | Polokwane Pet...  | ?     | 2010 FIFA Worl... | 300061451 | France         |
| 5  | Rustenburg Ro...  | ?     | 2010 FIFA Worl... | 300061450 | Mexico         |
| 6  | Mangaung/Blo...   | ?     | 2010 FIFA Worl... | 300061449 | France         |
| 7  | Johannesburg E... | ?     | 2010 FIFA Worl... | 300061460 | Argentina      |
| 8  | Nelson Mandel...  | ?     | 2010 FIFA Worl... | 300061459 | Korea Republic |
| 9  | Mangaung/Blo...   | ?     | 2010 FIFA Worl... | 300061457 | Greece         |
| 10 | Johannesburg S... | ?     | 2010 FIFA Worl... | 300061458 | Argentina      |
| 11 | Durban Durban...  | ?     | 2010 FIFA Worl... | 300111115 | Nigeria        |
| 12 | Polokwane Pet...  | ?     | 2010 FIFA Worl... | 300061455 | Greece         |
| 13 | Rustenburg Ro...  | ?     | 2010 FIFA Worl... | 300061466 | England        |
| 14 | Polokwane Pet...  | ?     | 2010 FIFA Worl... | 300061465 | Algeria        |
| 15 | Johannesburg E... | ?     | 2010 FIFA Worl... | 300061463 | Slovenia       |
| 16 | Cape Town Cap...  | ?     | 2010 FIFA Worl... | 300061464 | England        |
| 17 | Nelson Mandel...  | ?     | 2010 FIFA Worl... | 300061462 | Slovenia       |
| 18 | Tshwane/Pretor... | ?     | 2010 FIFA Worl... | 300061461 | USA            |
| 19 | Durban Durban...  | ?     | 2010 FIFA Worl... | 300111116 | Germany        |
| 20 | Tshwane/Pretor... | ?     | 2010 FIFA Worl... | 300061471 | Serbia         |

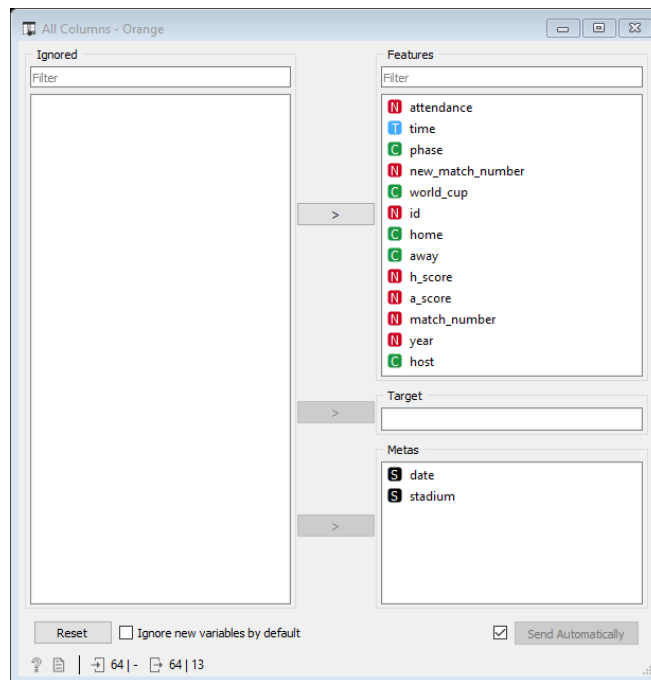
## 2. Diviser les données en deux parties, une partie contient les données de 2010 World Cup South Africa et une deuxième contient les autres données complètes, on utilise **Select Rows**.



3. Enregistrer les données complètes d'autres mondiales dans un fichier **matches\_no\_2010.csv** à l'aide de l'icône **Save Data**.

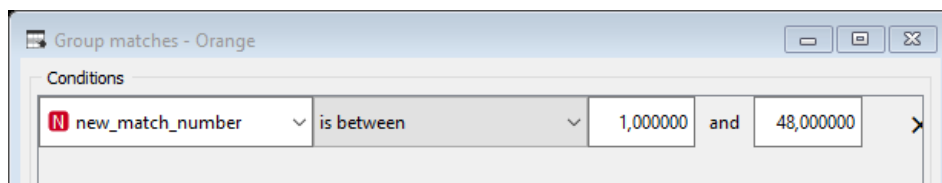


4. Sélectionner toutes les entités des données de 2010 World Cup South Africa à l'aide de **Select Columns**.

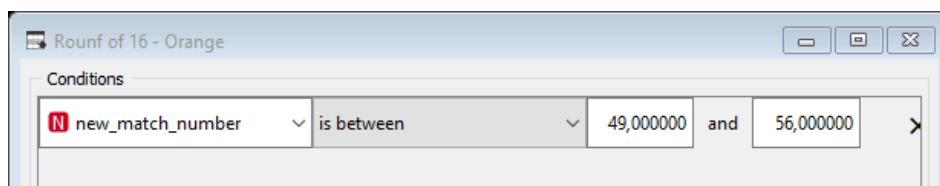


5. Diviser les des données de 2010 World Cup South Africa en 6 parties de phase différente, on utilise l'entité **new\_match\_number** et l'icône **Select Rows**.

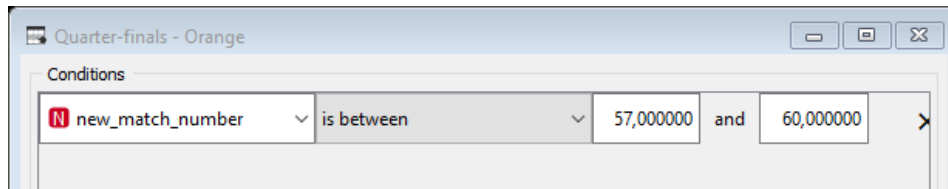
- ✓ Les lignes de 1 jusqu'à 48 sont de phase Group matches.



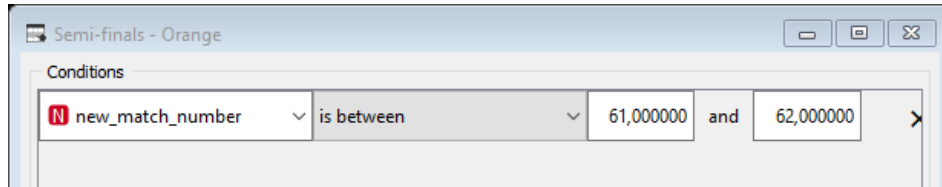
- ✓ Les lignes de 49 jusqu'à 56 sont de phase Round of 16.



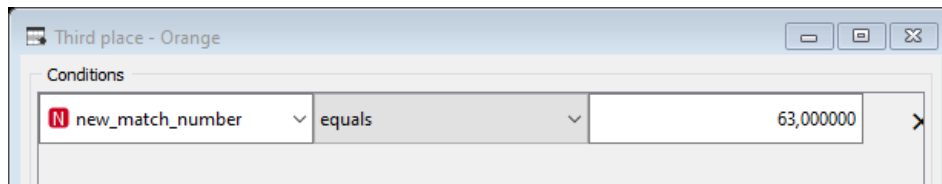
- ✓ Les lignes de 57 jusqu'à 60 sont de phase Quarter-Finals.



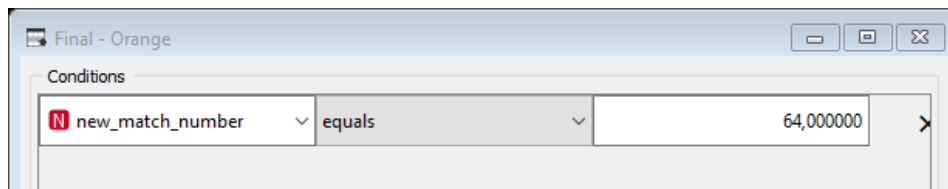
- ✓ Les lignes de 61 jusqu'à 62 sont de phase Semi-Finals.



- ✓ La ligne 63 est de phase Third place.

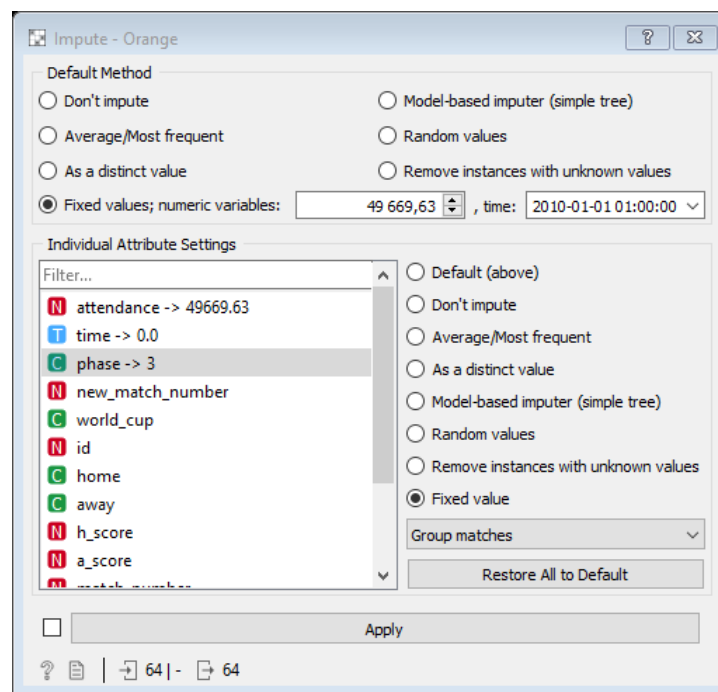


- ✓ La dernière ligne (ligne 64) est de phase Final.



- Appliquer les changements sur les entités **attendance**, **time** et **phase** pour chaque partie des 6 parties de données, avec l'icône **Impute**.

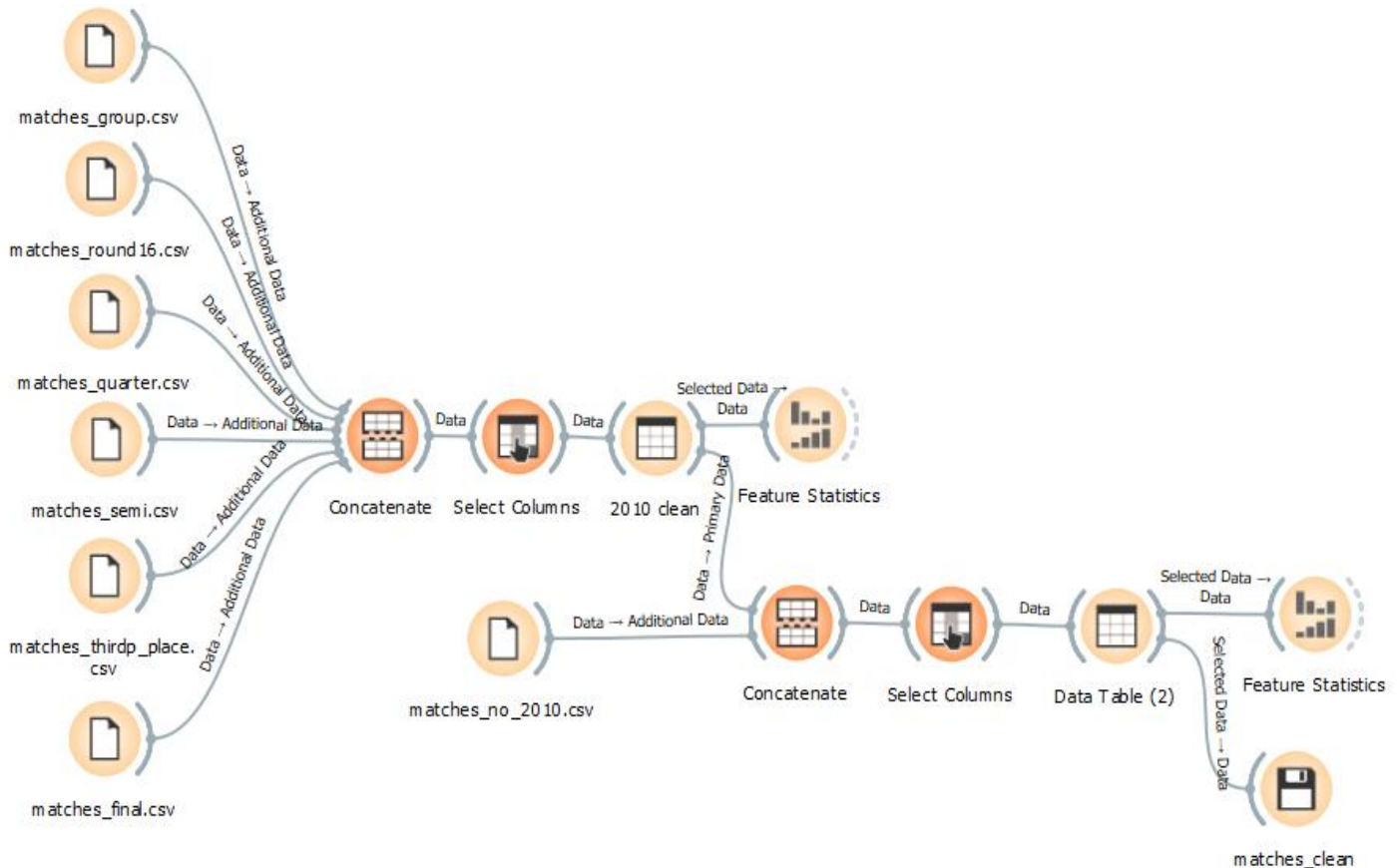
Exemple de Group matches :





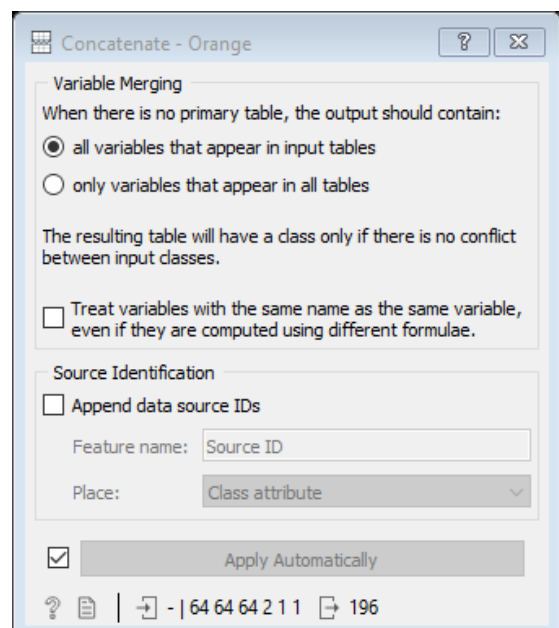
7. Enregistrer ces données sur des fichiers à l'aide de **Save Data** avec cet ordre :

- ✓ **matches\_group.csv**
- ✓ **matches\_round16.csv**
- ✓ **matches\_quarter.csv**
- ✓ **matches\_semi.csv**
- ✓ **matches\_thirdp\_place.csv**
- ✓ **matches\_final.csv**



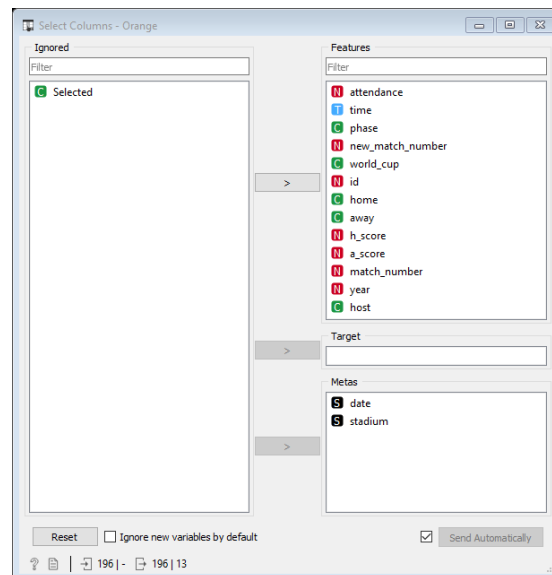
8. Importer ces 6 fichiers dernières avec **File**.

9. Rassembler ces fichiers pour voir un seule data de 2010 World Cup South Africa dans toutes les données sont complètes, en utilise **l'icône Concatenate**.

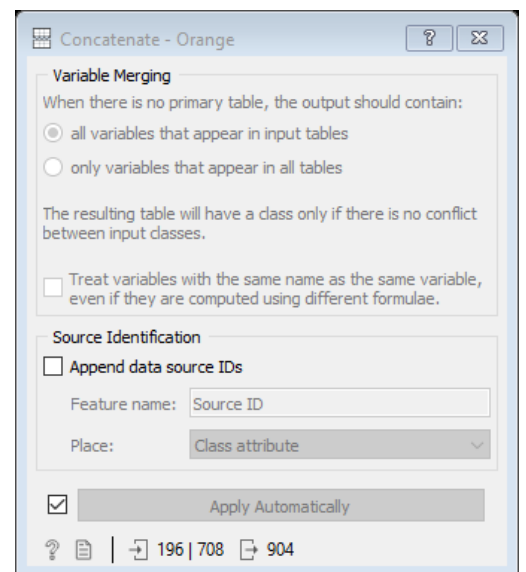




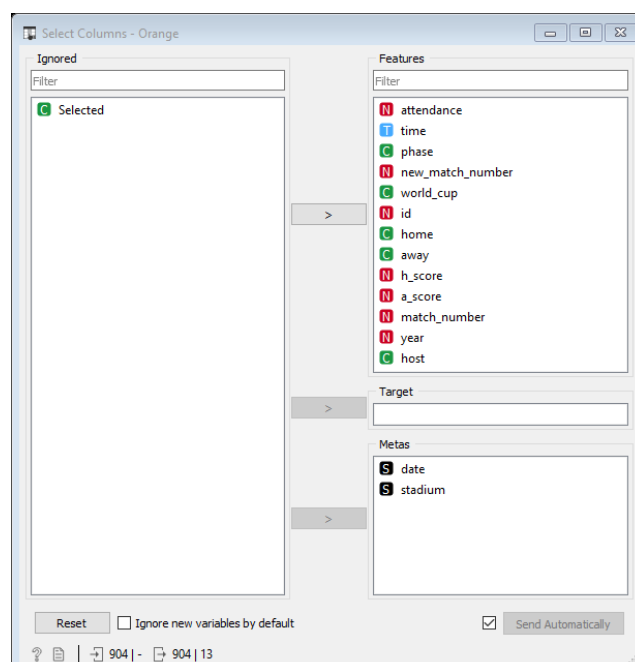
10. Supprimer l'entité **Selected** ajoutée par l'icône **Concatenate** à l'aide de l'icône **Select Columns**.



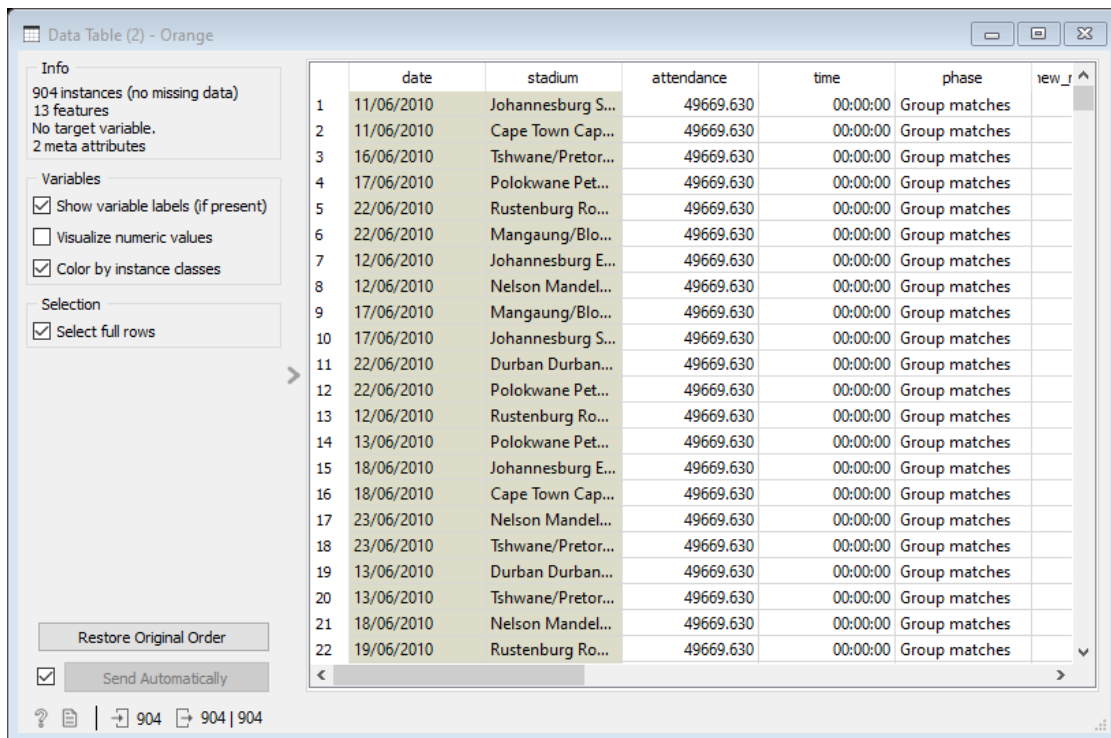
11. Rassembler les données de 2010 World Cup South Africa avec les données du fichier **matches\_no\_2010.csv** avec **Concatenate**.



12. Supprimer l'entité **Selected** ajoutée par l'icône **Concatenate** à l'aide de l'icône **Select Columns**.

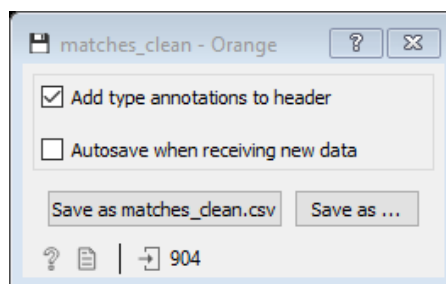


13. Vérifier qu'il n'y a pas des données manquantes avec **Data Table** et **Feature Statistics**.



|    | date       | stadium           | attendance | time     | phase         | new_id |
|----|------------|-------------------|------------|----------|---------------|--------|
| 1  | 11/06/2010 | Johannesburg S... | 49669.630  | 00:00:00 | Group matches |        |
| 2  | 11/06/2010 | Cape Town Cap...  | 49669.630  | 00:00:00 | Group matches |        |
| 3  | 16/06/2010 | Tshwane/Pretor... | 49669.630  | 00:00:00 | Group matches |        |
| 4  | 17/06/2010 | Polokwane Pet...  | 49669.630  | 00:00:00 | Group matches |        |
| 5  | 22/06/2010 | Rustenburg Ro...  | 49669.630  | 00:00:00 | Group matches |        |
| 6  | 22/06/2010 | Mangaung/Blo...   | 49669.630  | 00:00:00 | Group matches |        |
| 7  | 12/06/2010 | Johannesburg E... | 49669.630  | 00:00:00 | Group matches |        |
| 8  | 12/06/2010 | Nelson Mandel...  | 49669.630  | 00:00:00 | Group matches |        |
| 9  | 17/06/2010 | Mangaung/Blo...   | 49669.630  | 00:00:00 | Group matches |        |
| 10 | 17/06/2010 | Johannesburg S... | 49669.630  | 00:00:00 | Group matches |        |
| 11 | 22/06/2010 | Durban Durban...  | 49669.630  | 00:00:00 | Group matches |        |
| 12 | 22/06/2010 | Polokwane Pet...  | 49669.630  | 00:00:00 | Group matches |        |
| 13 | 12/06/2010 | Rustenburg Ro...  | 49669.630  | 00:00:00 | Group matches |        |
| 14 | 13/06/2010 | Polokwane Pet...  | 49669.630  | 00:00:00 | Group matches |        |
| 15 | 18/06/2010 | Johannesburg E... | 49669.630  | 00:00:00 | Group matches |        |
| 16 | 18/06/2010 | Cape Town Cap...  | 49669.630  | 00:00:00 | Group matches |        |
| 17 | 23/06/2010 | Nelson Mandel...  | 49669.630  | 00:00:00 | Group matches |        |
| 18 | 23/06/2010 | Tshwane/Pretor... | 49669.630  | 00:00:00 | Group matches |        |
| 19 | 13/06/2010 | Durban Durban...  | 49669.630  | 00:00:00 | Group matches |        |
| 20 | 13/06/2010 | Tshwane/Pretor... | 49669.630  | 00:00:00 | Group matches |        |
| 21 | 18/06/2010 | Nelson Mandel...  | 49669.630  | 00:00:00 | Group matches |        |
| 22 | 19/06/2010 | Rustenburg Ro...  | 49669.630  | 00:00:00 | Group matches |        |

14. Enregistrer les données complètes de fichier matches nettoyé dans un fichier **matches\_clean.csv** à l'aide de l'icône **Save Data**.



**Remarque :** L'utilisation de widget **Data Table** et **Feature Statistics** est importante et favorable pour vérifier si vos modifications sont bien appliquées ou non.

## ✓ Intégration de données :

### • Problème :

Nous avons 4 fichiers différents contiennent des informations pour des intervalles de temps différents.

- ✓ Le fichier **matches** contient les données des matches de la compétition World Cup de 1930 jusqu'à 2010, par exemple la date, stadium, phase, les résultats, etc.
- ✓ Le fichier **WorldCups** a les données de chaque mondiale en générale c'est-à-dire les informations de la compétition dans sa totalité, les quatre premières positions par exemple. Pour les années 1930 jusqu'à 2018.
- ✓ Le fichier **raw\_matches** contient les données des résultats des matches de la compétition World Cup de 1950 jusqu'à 2014.

- ✓ Le fichier **raw\_winners\_world\_cup** contient les données de position de 4 premières des mondiales de 1950 jusqu'à 2010.

Les données sont représentées avec des manières différentes par exemple dans le fichier matches nous avons une représentation différent de la date et aussi pour la phase (Third place – Match for Third place).

Quelques données répéter dans des fichiers différents.

Il y a des entités différentes de même nom dans des fichiers différents.

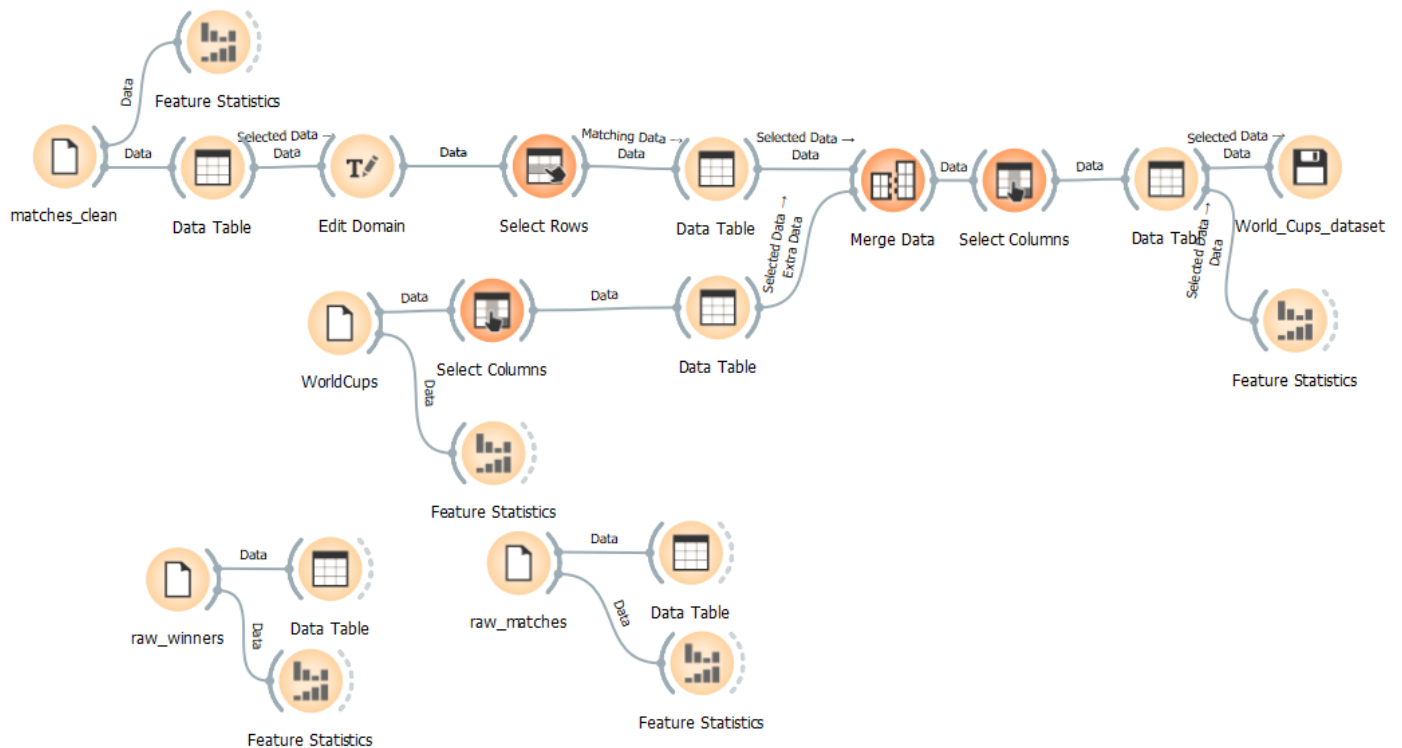
- **Solution :**

Pour les intervalles de temps différent on va travailler avec un seul intervalle commun 1950 – 2010.

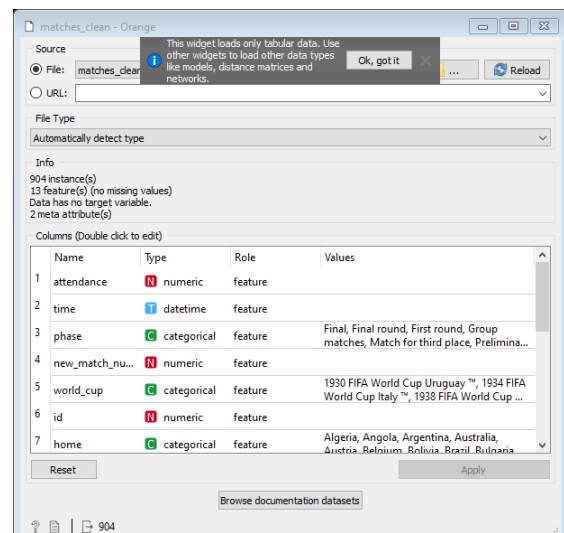
Les données de fichier **raw\_matches** et **raw\_winners\_world\_cup** sont déjà existents dans **matches** et **WorldCups**, donc on va travailler juste avec ces deux dernières.

Pour corriger les problèmes d'harmonisation des entités date et phase, on va utiliser **l'icône Edit Domain**.

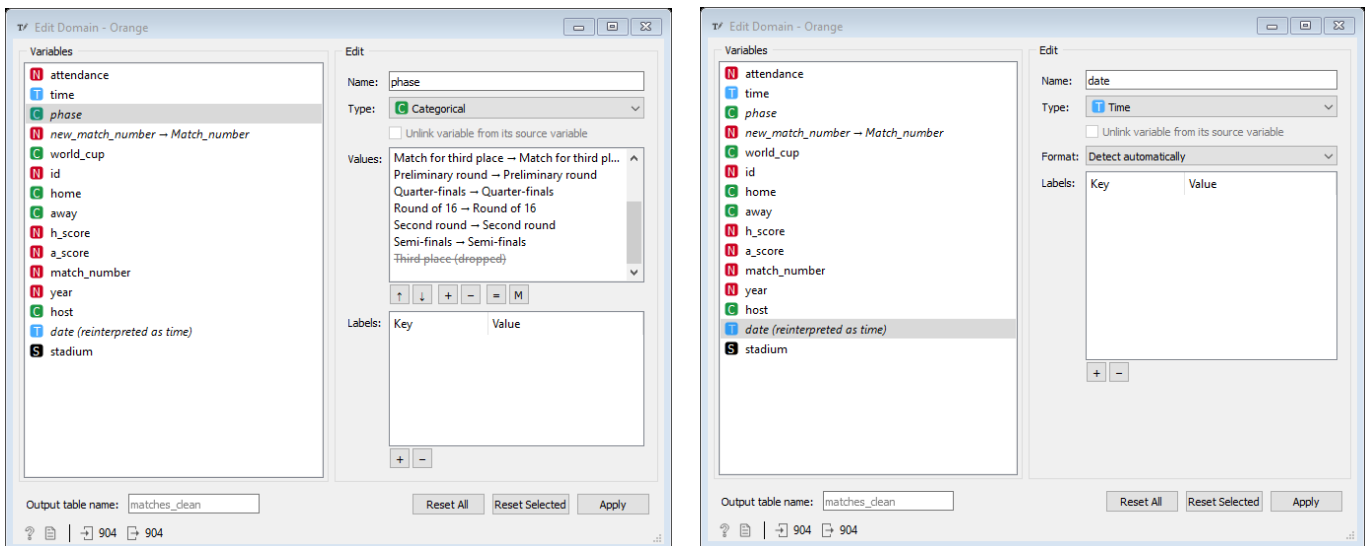
Nous avons faire ces corrections avec ce workflow :



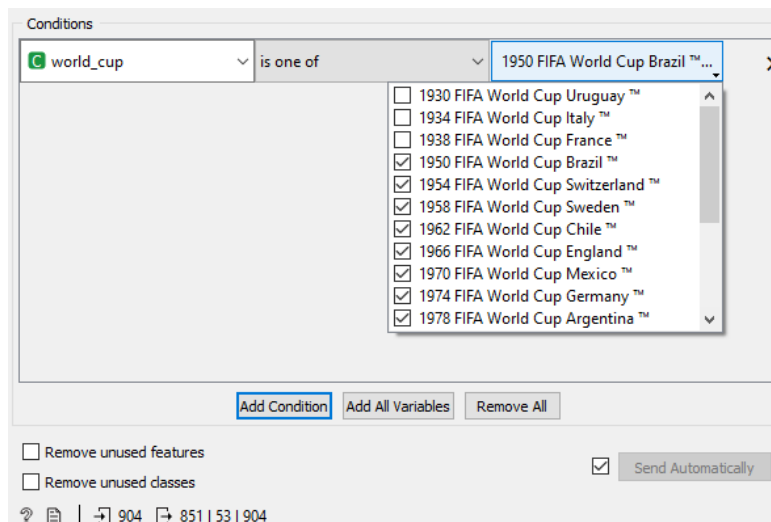
## 1. Importer le fichier **matches\_clean** avec **File**.



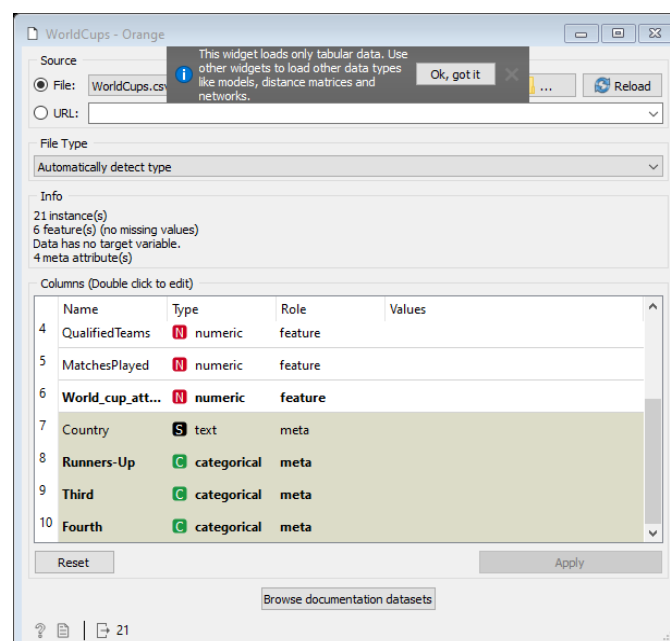
## 2. Corriger l'harmonisation de la **date** et **phase** avec **Edit Domain**.



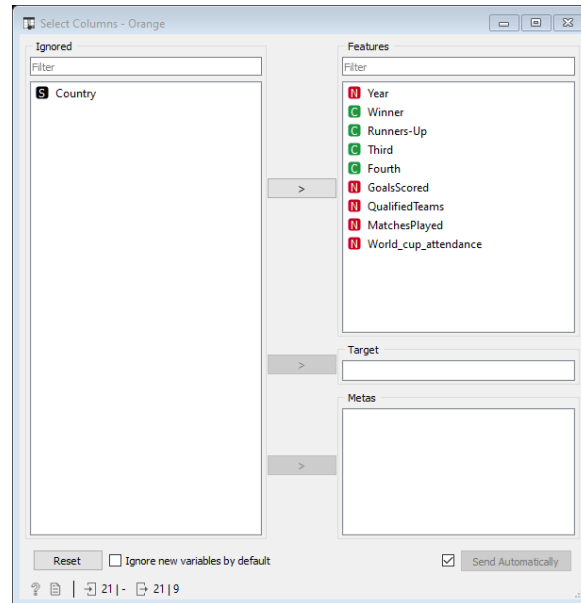
## 3. Sélectionner les mondiales entre les années 1950 et 2010 avec **Select Rows**.



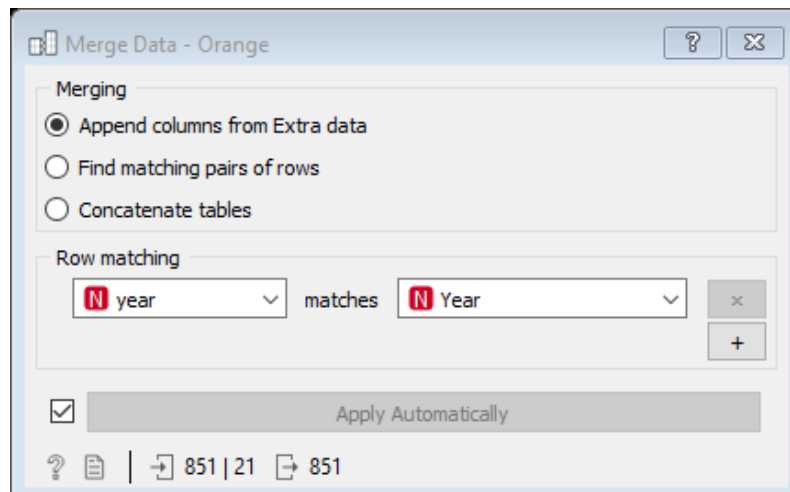
## 4. Importer le fichier **WorldCups** à l'aide de **Picône File**.



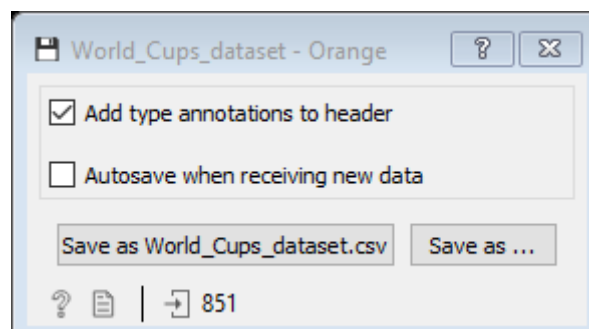
5. Supprimer l'entité **country** qui y'a les mêmes données existent dans **host** de fichier **matches** avec l'icône **Select Columns**.



6. Rassembler ces données avec **Merge Data**.

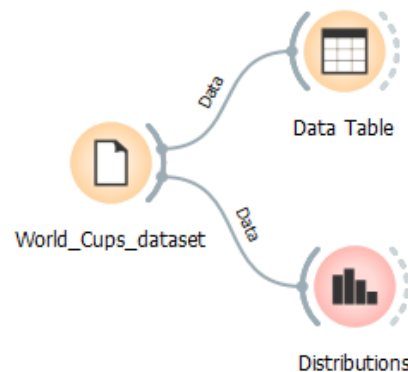


7. Enregistrer les données finales à l'aide de **Save Data** avec le nom **World\_CupsI\_dataset**.



## ✓ Conclusion :

L'ensemble des données à l'état final superficiel c'est le même de départ parce que contient les mêmes informations et connaissances mais à la profondeur des choses ce n'est pas les mêmes, cette dernière compose d'une seule source d'informations structures et complètes.

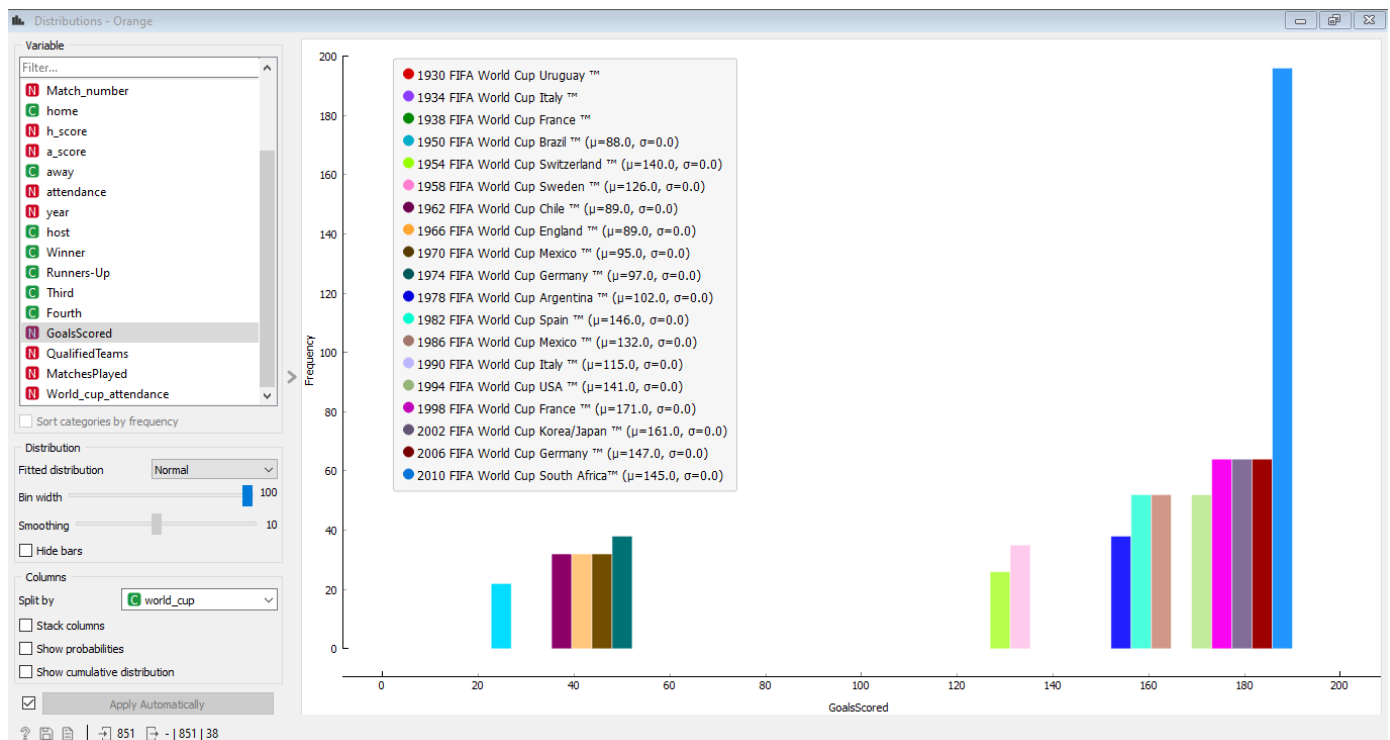


- Dataset final :

The screenshot shows the 'Data Table - Orange' window. On the left, the 'Info' panel displays: 851 instances (no missing data), 20 features, Target with 19 values, and 1 meta attribute. The 'Variables' panel has 'Show variable labels (if present)' checked, 'Visualize numeric values' unchecked, and 'Color by instance classes' checked. The 'Selection' panel has 'Select full rows' checked. The main table displays 22 rows of data with columns: Id\_cup, stadium, date, time, id, phase, Match\_number, and home\_team. The bottom status bar shows 851 instances and 851 features.

|    | Id_cup     | stadium           | date                | time     | id          | phase       | Match_number | home_team |
|----|------------|-------------------|---------------------|----------|-------------|-------------|--------------|-----------|
| 1  | FA Worl... | Johannesburg S... | 2010-11-06 00:00:00 | 00:00:00 | 300061454.0 | Round of 16 | 1            | South A   |
| 2  | FA Worl... | Cape Town Cap...  | 2010-11-06 00:00:00 | 00:00:00 | 300061453.0 | Round of 16 | 2            | Urugua    |
| 3  | FA Worl... | Tshwane/Pretor... | 2010-06-16 00:00:00 | 00:00:00 | 300061452.0 | Round of 16 | 16           | South A   |
| 4  | FA Worl... | Polokwane Pet...  | 2010-06-17 00:00:00 | 00:00:00 | 300061451.0 | Round of 16 | 20           | France    |
| 5  | FA Worl... | Rustenburg Ro...  | 2010-06-22 00:00:00 | 00:00:00 | 300061450.0 | Round of 16 | 36           | Mexico    |
| 6  | FA Worl... | Mangaung/Blo...   | 2010-06-22 00:00:00 | 00:00:00 | 300061449.0 | Round of 16 | 35           | France    |
| 7  | FA Worl... | Johannesburg E... | 2010-12-06 00:00:00 | 00:00:00 | 300061460.0 | Round of 16 | 5            | Argent    |
| 8  | FA Worl... | Nelson Mandel...  | 2010-12-06 00:00:00 | 00:00:00 | 300061459.0 | Round of 16 | 3            | Korea R   |
| 9  | FA Worl... | Mangaung/Blo...   | 2010-06-17 00:00:00 | 00:00:00 | 300061457.0 | Round of 16 | 19           | Greece    |
| 10 | FA Worl... | Johannesburg S... | 2010-06-17 00:00:00 | 00:00:00 | 300061458.0 | Round of 16 | 18           | Argent    |
| 11 | FA Worl... | Durban Durban...  | 2010-06-22 00:00:00 | 00:00:00 | 300111115.0 | Round of 16 | 34           | Nigeria   |
| 12 | FA Worl... | Polokwane Pet...  | 2010-06-22 00:00:00 | 00:00:00 | 300061455.0 | Round of 16 | 33           | Greece    |
| 13 | FA Worl... | Rustenburg Ro...  | 2010-12-06 00:00:00 | 00:00:00 | 300061466.0 | Round of 16 | 4            | Englan    |
| 14 | FA Worl... | Polokwane Pet...  | 2010-06-13 00:00:00 | 00:00:00 | 300061465.0 | Round of 16 | 6            | Algeria   |
| 15 | FA Worl... | Johannesburg E... | 2010-06-18 00:00:00 | 00:00:00 | 300061463.0 | Round of 16 | 22           | Sloveni   |
| 16 | FA Worl... | Cape Town Cap...  | 2010-06-18 00:00:00 | 00:00:00 | 300061464.0 | Round of 16 | 23           | Englan    |
| 17 | FA Worl... | Nelson Mandel...  | 2010-06-23 00:00:00 | 00:00:00 | 300061462.0 | Round of 16 | 40           | Sloveni   |
| 18 | FA Worl... | Tshwane/Pretor... | 2010-06-23 00:00:00 | 00:00:00 | 300061461.0 | Round of 16 | 37           | USA       |
| 19 | FA Worl... | Durban Durban...  | 2010-06-13 00:00:00 | 00:00:00 | 300111116.0 | Round of 16 | 7            | German    |
| 20 | FA Worl... | Tshwane/Pretor... | 2010-06-13 00:00:00 | 00:00:00 | 300061471.0 | Round of 16 | 8            | Serbia    |
| 21 | FA Worl... | Nelson Mandel...  | 2010-06-18 00:00:00 | 00:00:00 | 300061470.0 | Round of 16 | 21           | German    |
| 22 | FA Worl... | Rustenburg Ro...  | 2010-06-19 00:00:00 | 00:00:00 | 300061469.0 | Round of 16 | 26           | Ghana     |

- GoalsScored en fonction de world\_cup (les buts marquent pour chaque compétition mondiale) :



- World\_cup\_attendance en fonction de world\_cup (le nombre des participants pour chaque compétition mondiale) :

