Analyse des données Licence Pro 2025-2026

TD n°2- L'analyse univariée

Florian Bayer

Objectif du TD 2

Les objectifs de ce TD sont de mettre en application les acquis du cours 2 sur l'analyse d'une série de données

- avec des graphiques
- les valeures centrales
- les paramètres de dispersion

Vous apprendrez à utiliser un outil d'analyse de données : Orange

Orange : présentation

Orange est un logiciel open source dédié à l'analyse de données, à l'exploration visuelle et à l'apprentissage automatique.

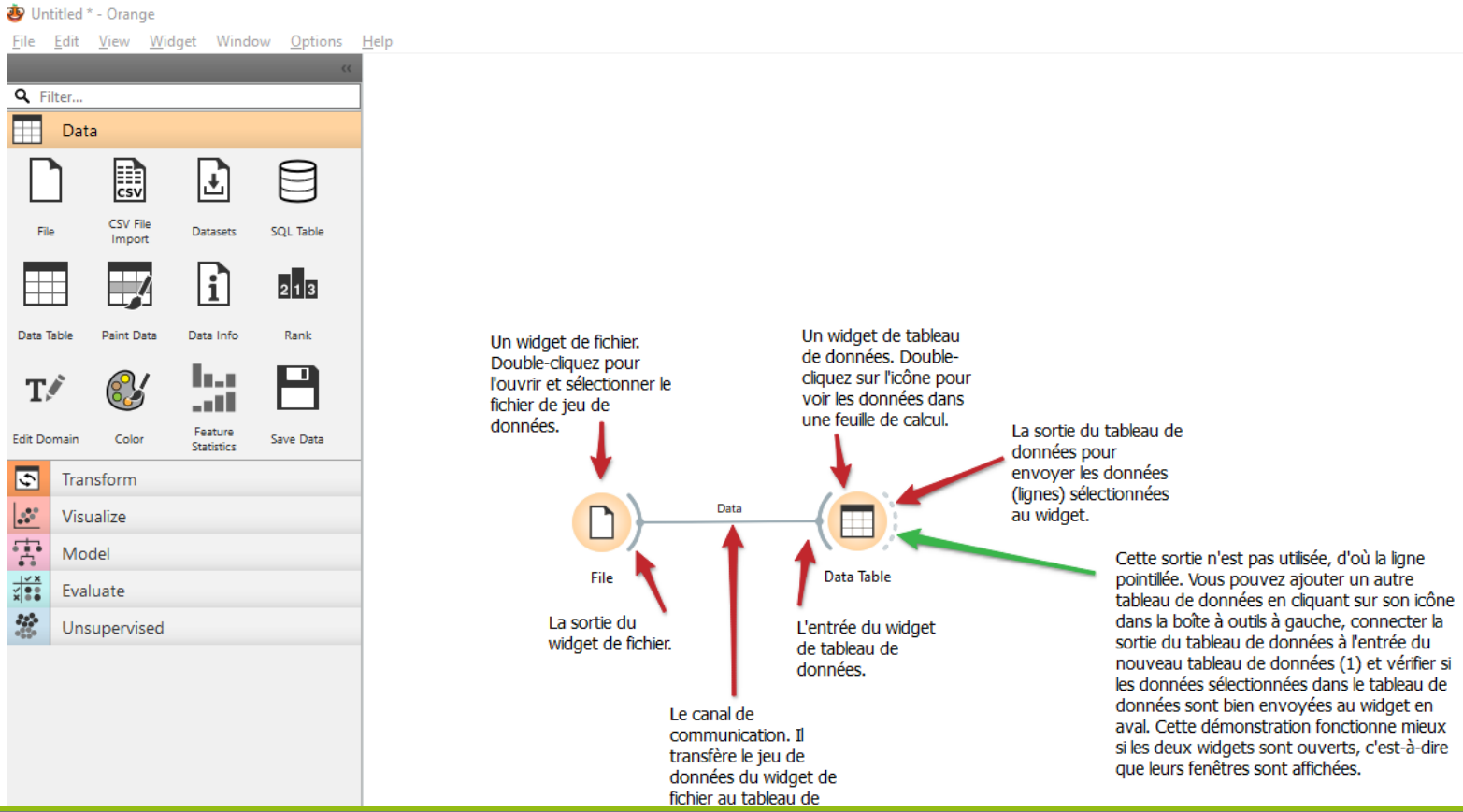
Basé sur des packages Python pour les analyses, son interface graphique intuitive permet de construire des flux de travail ou workflow, sans avoir à écrire de code.

Contrairement à Excel et comme lorsque l'on utilise du code, l'avantage d'Orange est de pouvoir relancer chaque étape du calcul pour le vérifier ou le modifier.

Orange : présentation

La vue principale d'Orange ressemble à une toile vide où vous pouvez commencer à ajouter des widgets pour créer un workflow.

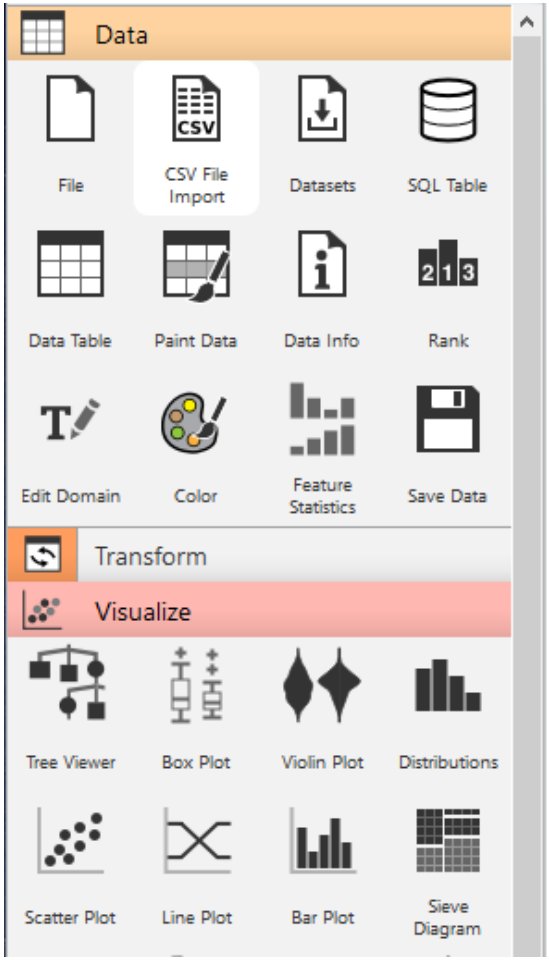
A gauche la zone de widgets, à droite l'espace de travail.



Orange : les widgets

Les widgets sont des modules préconfigurés qui permettent d'importer, de traiter, d'analyser et de visualiser des données. À gauche de l'écran se trouve une barre d'outils avec les catégories de widgets comme :

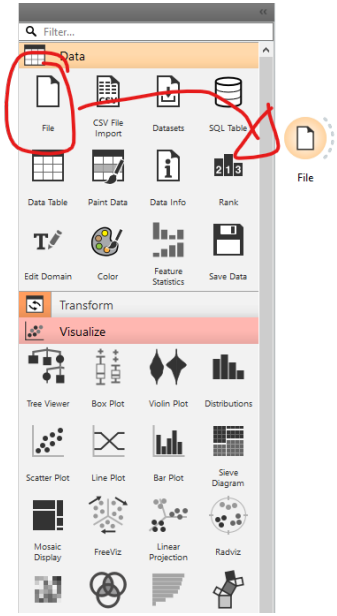
- Data : Chargement, transformation, filtrage des données
- Visualize : Graphiques et visualisations
- Model : Apprentissage automatique (classification, régression)
- Evaluate : Validation de modèles

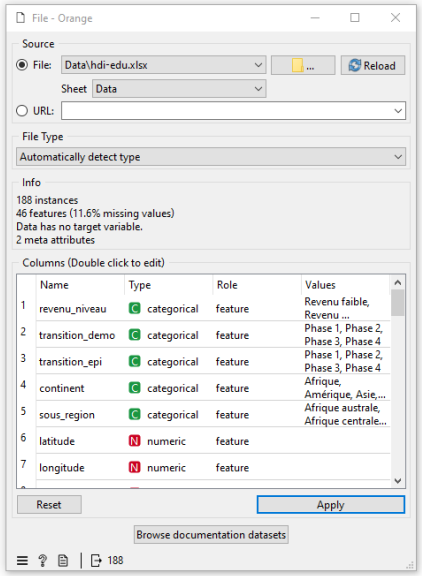


Orange : chargement des données

Le chargement des données se fait à travers le widget "File" dans la catégorie "Data". Il vous permet d'importer des fichiers sous différents formats, y compris CSV ou Excel

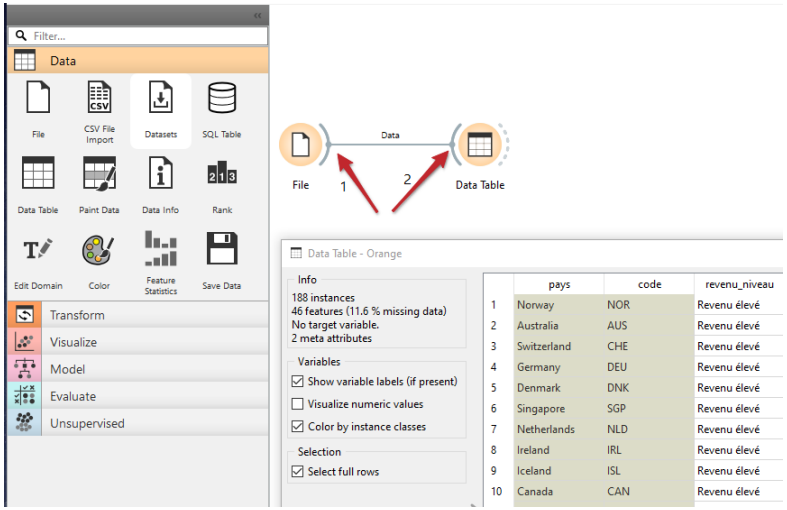
- Ajoutez le widget "File" : Glissez-déposez le widget *File* depuis la barre d'outils dans l'espace de travail.
- Sélectionnez le fichier : Cliquez sur le widget "File" puis ouvrez hdi-edu.xlsx. Vérifiez que vous avez bien la feuille de calcul Data
- Orange détermine seul chaque type de données, mais vous pouvez les modifier manuellement via la colonne Type
- Fermez le widget *File*





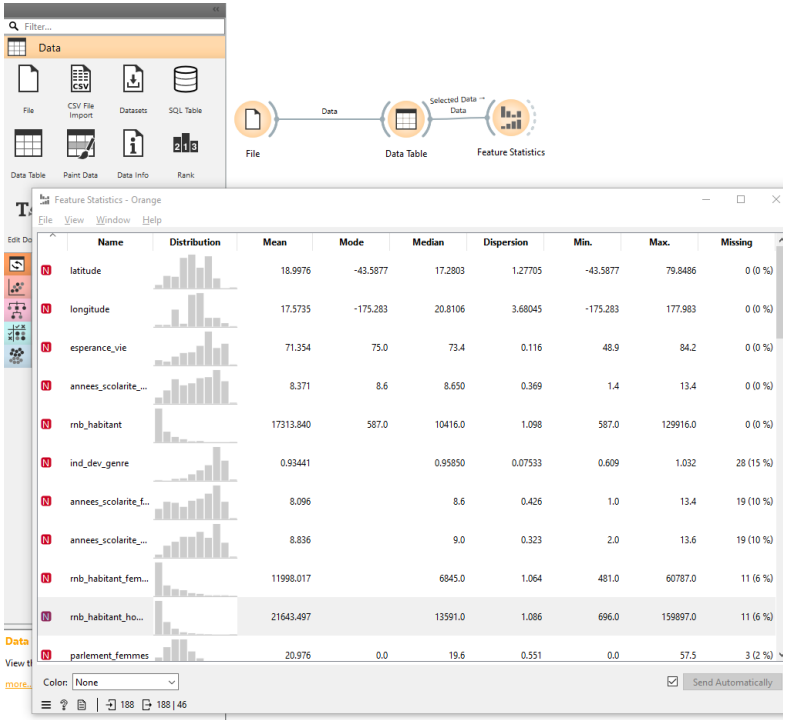
Orange : afficher les données

- Dans la catégorie Data, sélectionnez et faites glisser le widget Data Table
- A l'aide de la souris, connectez la sortie du widget *File* à l'entrée du widget *Data Table*
- Vous pouvez voir maintenant le contenu des données via le widget *Data Table*
- Comme ils sont liés, tous changements dans le widget *File* entraînera une modification dans le widget *Data Table*
- Vous pouvez sélectionner des lignes dans *Data Table*, mais cela aura aussi un impact sur les futurs calculs. Ils ne se feront que sur les lignes sélectionnées



Orange : Statistiques univariée

- Ajoutez le widget *Feature Statistics*, toujours dans la catégorie Data.
- Connectez le à la sortie de *Data Table*
- Un histogramme, des valeurs centrales et des paramètres de dispersion sont disponibles
- Notez que Dispersion correspond au coefficient de variation pour les données quantitatives
- Attention, pensez à vérifier que vous n'avez pas de ligne sélectionnée dans Data Table, sinon les calculs de *Feature Statistics* seront uniquement fait sur votre sélection
- Notez que vous pouvez appliquer une couleur aux histogrammes pour les distinguer à l'aide d'une variable qualitative. Par exemple par continent.

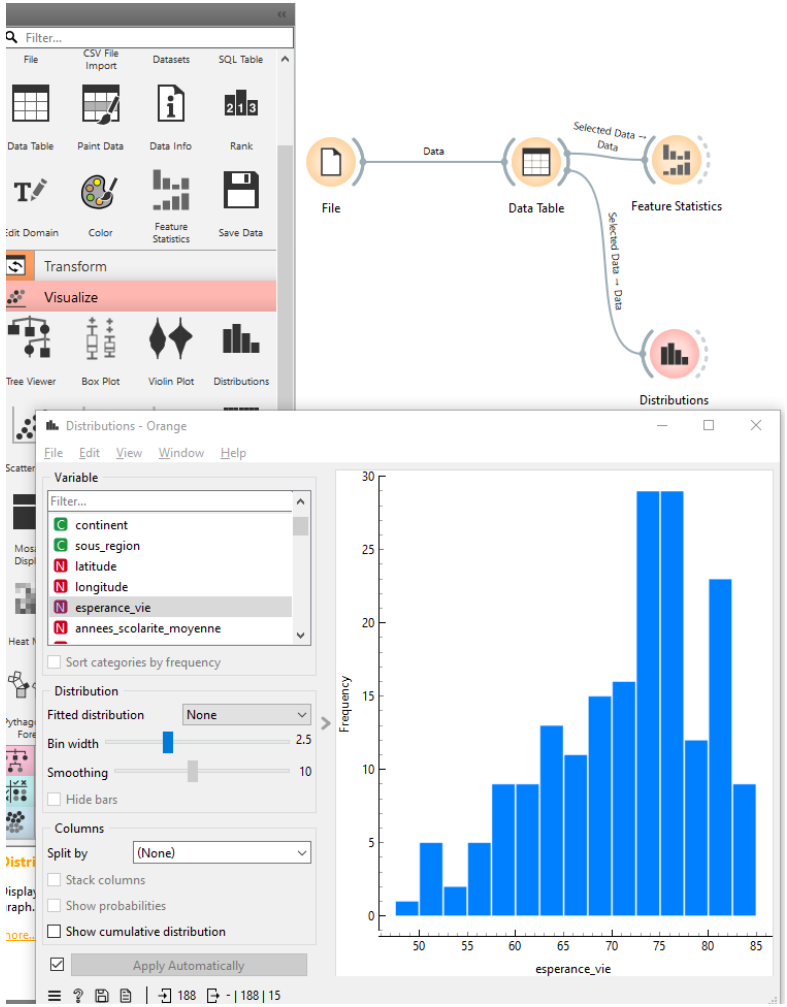


Faites l'analyse univariée de la variable esperance_vie. Que pouvez-vous en conclure ?

Orange : Histogramme

- A partir de la catégorie Visualize, ajoutez le widget *Distributions*.
- Connectez le à la sortie de *Data Table*
- Les données qualitatives (Category) sont représentées par un diagramme en bâton
- Les données quantitatives par un histogramme
- Pour ce dernier, vous pouvez modifier le nombre de *bins*

Comme précédemment, vous pouvez appliquer une catégorie pour découper l'histogramme selon les modalités de cette dernière. Pour la variable esperance_vie, faites un split par transition_epi. Que pouvez-vous en conclure ? N'oubliez pas de consulter les métadonnées (onglet Meta du fichier Excel) pour plus de détails



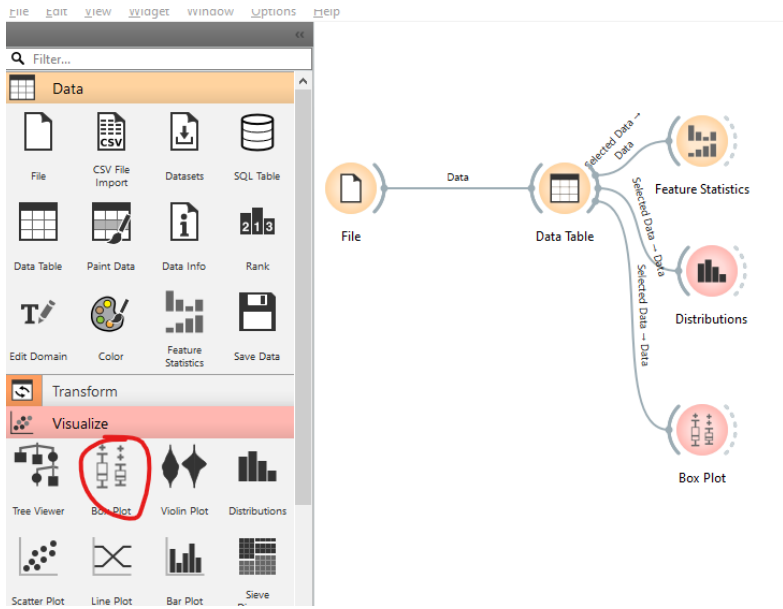
Orange : Box Plot

- Ajoutez maintenant un Box Plot.
- Connectez le à la sortie de *Data Table*

Analysez la distribution de taux_fertilite, puis faites un sous-groupe avec transition_demo.

Que pouvez-vous en conclure ?

Pensez à sauvegarder votre projet Orange, nous le réutiliserons pour le TD4



Orange : Analyse comparative

Consignes :

1. Formez des binômes
2. Analysez la variable `esperance_vie` avec Feature Statistics et Box Plot

Étape 1 - Vue d'ensemble :

- Analysez les moyennes et écart-types pour chaque continent (split)
- Utilisez le Box Plot avec split par Continent pour comparer l'Afrique aux autres continents
- Que révèle ces analyses ?

Étape 2 - Analyse par pays africains :

- Maintenant et l'aide d'un nouveau flux, faites l'analyse pour les pays africains
- Identifiez les pays avec les `esperance_vie` extrêmes (outliers + min/max)
- Choisissez 2 pays opposés pour votre analyse.

Comprendre les outliers africains

À partir de votre analyse des pays africains, vous devez expliquer pourquoi certains pays ont des espérances de vie si différentes.

Étape 3 - Identifier les cas extrêmes :

- Pays avec la plus haute espérance de vie : _
- Pays avec la plus basse espérance de vie : _
- Outliers détectés (si il y en a) : _

Étape 4 - Formuler des hypothèses : Avant de regarder les données, quelles pourraient être les raisons de ces différences ?

- 1 _
- 2 _
- 3 _

Étape 3 - Tester vos hypothèses :

Explorez les autres variables (économiques, sociales, sanitaires) pour vos pays extrêmes. Utilisez l'onglet Meta pour comprendre les indicateurs.

Restitution collective

Tour de table : Chaque binôme présente en 1 minute :

- Leurs pays extrêmes choisis et leurs hypothèses principales
- Une découverte surprenante dans les données

Questions de débriefing :

- Quelles variables explicatives reviennent le plus souvent ?
- Y a-t-il des pays difficiles à expliquer avec vos hypothèses ?
- Vos hypothèses initiales étaient-elles confirmées ou infirmées ?

Synthèse pédagogique :

- Dispersion : L'écart-type révèle des sous-groupes de pays avec des profils similaires
- Corrélation ≠ causalité : Vous avez identifié des associations, mais attention aux explications simplistes
- Contextualisation : Les chiffres racontent des histoires humaines et géographiques
- Démarche scientifique : Hypothèse → Test → Interprétation critique
- Préparation bivariée : Vous avez intuitivement cherché des corrélations, nous allons maintenant les mesurer !

Dans ce TD, vous avez appris à

Compétences techniques :

- Charger et manipuler des données dans Orange
- Calculer et interpréter les mesures de tendance centrale (moyenne, médiane)
- Calculer et interpréter les mesures de dispersion (écart-type, coefficient de variation)
- Créer et analyser différents types de graphiques (histogrammes, box plots)
- Comparer des distributions entre groupes (continents)

Compétences analytiques :

- Formuler des hypothèses avant l'analyse des données
- Identifier et expliquer les valeurs aberrantes (outliers)
- Comprendre qu'une moyenne cache des disparités (importance de l'écart-type)
- Contextualiser les résultats statistiques avec des connaissances géographiques
- Passer de l'observation à l'explication causale

Cette démarche d'investigation vous a préparés à l'analyse bivariée : vous avez commencé à chercher des liens entre variables (espérance de vie ↔ revenus, éducation, santé...). Dans les prochains cours, nous formaliserons ces intuitions avec les techniques de corrélation et la cartographie statistique.