Guide nettoyage de données

Bien préparer ses datas

Les couturiers inclusifs

Procédure de data analyse N°1

Sommaire :

**Étape 1 : Collecte des Données**

1. Définir les Objectifs de Collecte

2. Identifier les Sources de Données

3. Évaluer la Qualité des Données

4. Mettre en Place un Plan de Collecte en lien avec les normes RGPD

5. Importer les Données dans Excel

**Étape 2 : Exploration des Données**

1. Acquérir une Compréhension Globale des Données

2. Identifier les Anomalies et les Valeurs Manquantes

3. Visualiser les Données

4. Résumer les Statistiques descriptives

**Étape 3 : Nettoyage des Données**

1. Vérification des Problèmes avec Excel

2. Allègement du Poids du Fichier

3. Traitement des Données aux Mauvais Formats

4. Traitement des Valeurs Manquantes

5. Détection des Valeurs Aberrantes

6. Élimination des Doublons

7. Transposition des Données

**Étape 4 – Examiner, adapter, répéter**

Introduction :

Cette procédure vise à vous guider à travers les étapes de collecte, d'exploration et de nettoyage des données en utilisant Excel, adaptée aux débutants. Ces étapes sont essentielles pour garantir la qualité et la fiabilité des données utilisées dans vos analyses.

****Définition du nettoyage des données :

Le nettoyage des données est un processus qui vise à identifier et corriger les données altérées, inexactes ou non pertinentes. Cette étape fondamentale du traitement des données améliore la cohérence, fiabilité et valeur des données.

**Étape 1 : Collecte des Données**

1. Définir les Objectifs de Collecte :

• Avant de commencer, identifiez clairement ce que vous souhaitez accomplir avec vos données. Par exemple, si vous souhaitez analyser les ventes d'une entreprise, *déterminez les informations nécessaires* telles que les dates de vente, les produits vendus et les montants.

2. Identifier les Sources de Données :

• Recensez toutes les sources potentielles de données disponibles, comme des fichiers Excel, des bases de données internes, ou des données disponibles en ligne.

3. Évaluer la Qualité des Données :

• Avant de les importer dans Excel, examinez les données pour identifier toute anomalie ou incohérence qui pourrait affecter leur qualité.

• Il s’agit d’examiner les données brutes : Ouvrez le fichier de données dans un éditeur de texte ou dans un logiciel spécialisé pour examiner les données brutes dans leur format d'origine. Cela vous permettra de voir les données telles qu'elles sont enregistrées.

4. Mettre en Place un Plan de Collecte :

• Créez un plan détaillé indiquant comment vous allez collecter les données, quels fichiers vous allez utiliser, et quand vous allez les collecter.

La collecte doit respecter Le Règlement Général européen sur la Protection des Données (RGPD) qui vise à renforcer la protection des données personnelles des individus.

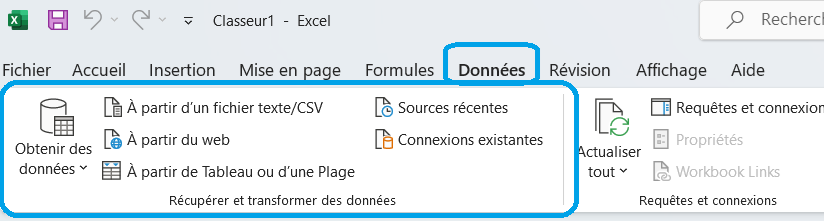
Voici quelques-unes des règles importantes à suivre :

1. **Consentement explicite :** Vous devez obtenir le consentement explicite de la personne pour collecter et traiter ses données personnelles à des fins spécifiques. Le consentement doit être donné de manière libre, spécifique, éclairée et univoque.
2. **Finalité limitée :** Vous ne devez collecter les données personnelles que pour des finalités spécifiques, légitimes et clairement définies. Vous ne pouvez pas utiliser les données à d'autres fins sans obtenir un consentement supplémentaire.
3. **Minimisation des données :** Vous ne devez collecter que les données personnelles strictement nécessaires pour atteindre les finalités spécifiées. Vous devez limiter la quantité de données collectées au minimum nécessaire.
4. **Exactitude des données :** Vous devez prendre des mesures pour garantir que les données personnelles que vous collectez sont exactes et tenues à jour. Vous devez également corriger ou effacer les données inexactes ou incomplètes dès que possible.
5. **Durée de conservation limitée :** Vous ne devez pas conserver les données personnelles plus longtemps que nécessaire pour atteindre les finalités pour lesquelles elles ont été collectées. Vous devez établir des politiques de conservation des données et les respecter.
6. **Intégrité et confidentialité :** Vous devez prendre des mesures techniques et organisationnelles appropriées pour protéger les données personnelles contre la perte, la destruction, la divulgation non autorisée ou l'accès non autorisé.
7. **Droits des personnes concernées :** Vous devez informer les personnes concernées de leurs droits en vertu du RGPD, y compris leur droit d'accès, de rectification, d'effacement et d'opposition au traitement de leurs données personnelles.
8. **Transferts internationaux :** Si vous transférez des données personnelles en dehors de l'Union européenne, vous devez vous assurer que des mesures appropriées sont en place pour garantir un niveau de protection adéquat des données.

Il est important de noter que ces règles s'appliquent à toutes les étapes du processus de traitement des données, y compris la collecte, le stockage, le traitement et la destruction des données personnelles. Il est également crucial de documenter et de tenir des registres de toutes les activités de traitement des données pour démontrer la conformité au RGPD.

5. Importer les Données dans Excel :

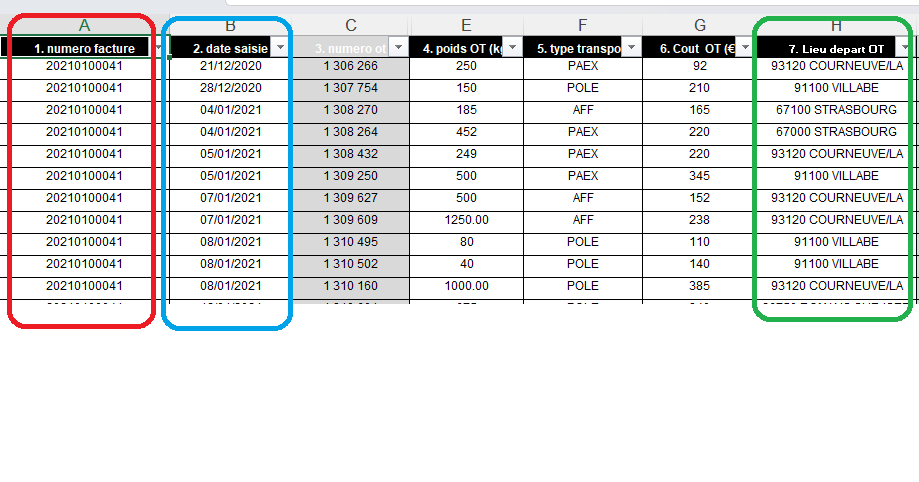
• Utilisez l'option "obtenir des données" dans l’onglet Données d’Excel pour charger vos données depuis les différentes sources identifiées.



**Étape 2 : Exploration des Données :**

1. Acquérir une Compréhension Globale des Données :

• Une fois les données importées, examinez-les pour comprendre leur contenu. Parcourez les différentes colonnes pour identifier les types de données qu'elles contiennent.



Encadré rouge : données de nombre

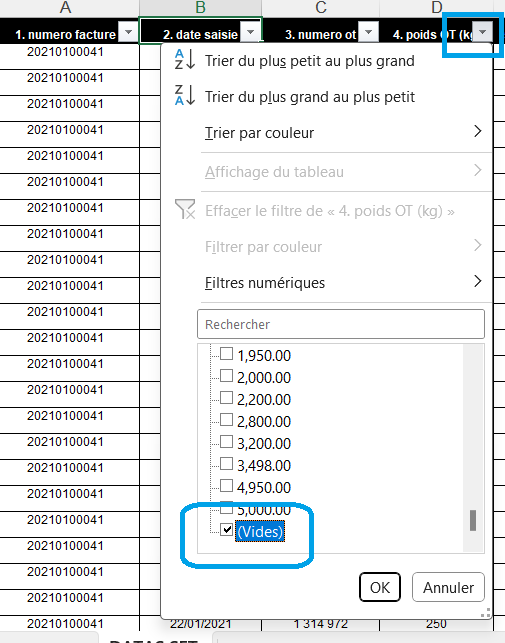
Encadré bleu : données de date

Encadré vert : données de texte

2. Identifier les Anomalies et les Valeurs Manquantes :

• Utilisez les outils de filtrage et de tri dans Excel pour repérer les valeurs manquantes ou les anomalies dans vos données.

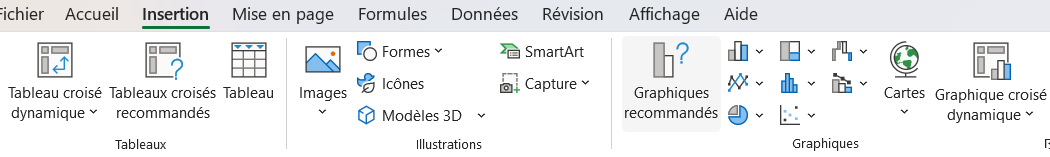
Cliquez sur le bouton grisâtre avec la flèche dans les en-têtes des colonnes, les options de tris et filtres s’affichent, descendez jusqu’à la dernière valeur, si un champs « vides » apparait c’est qu’il y a des cellules où rien n’est renseigné :



3. Visualiser les Données :

• Créez des graphiques simples, comme des histogrammes ou des graphiques à barres, pour visualiser la répartition des données dans chaque colonne.

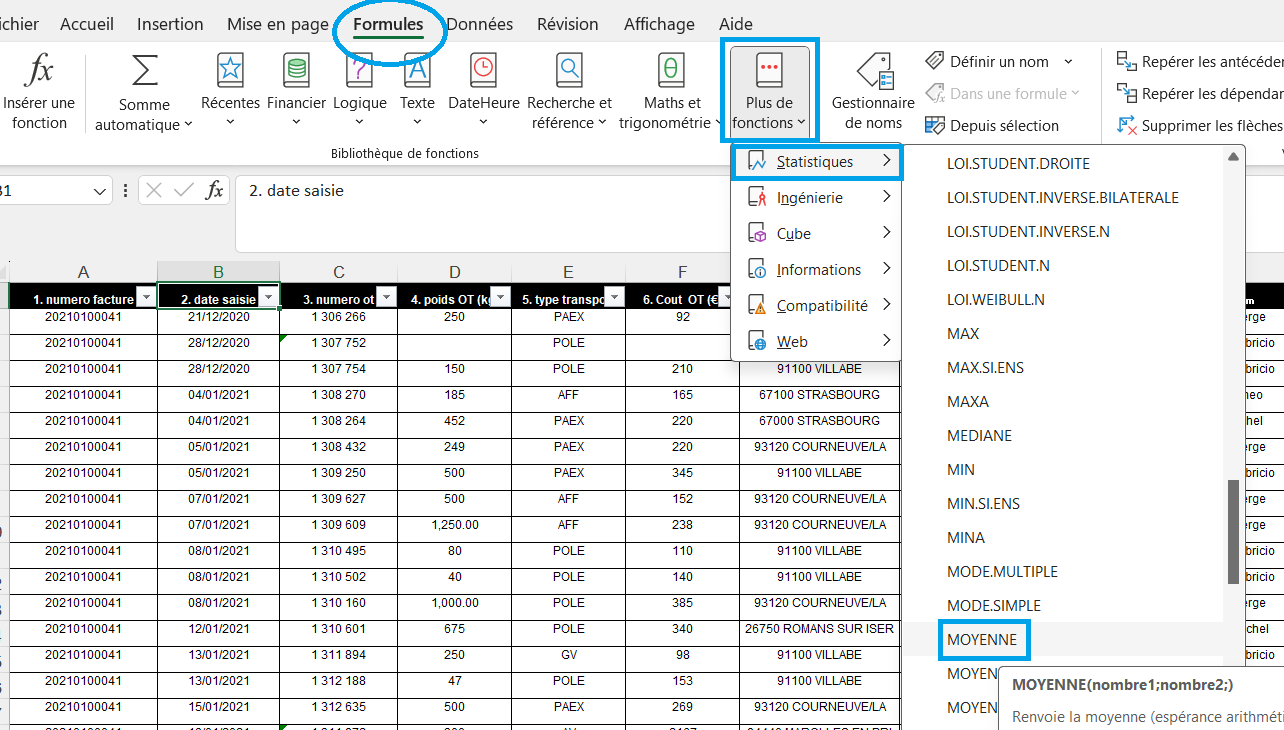
Cliquez sur l’onglet insertion, puis choisir le type de graphique :



4. Résumer les Statistiques descriptives :

• Utilisez les fonctions intégrées d'Excel telles que "MOYENNE", "MEDIANE", et "ECART.TYPE" pour calculer les statistiques descriptives

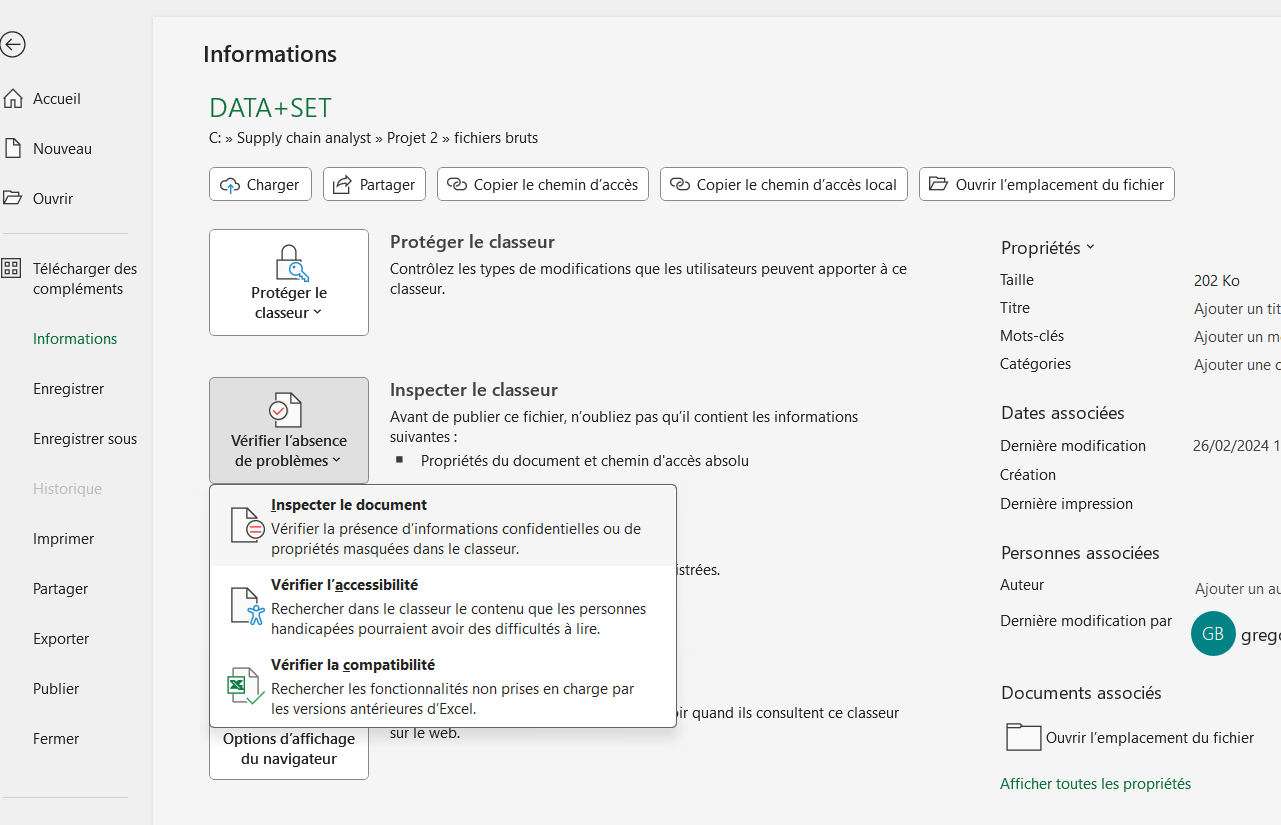
Cliquez sur l’onglet « formule », puis « plus de fonctions », « statistiques » et choisissez la formule, exemple « MOYENNE » :



**Étape 3 : Nettoyage des Données :**

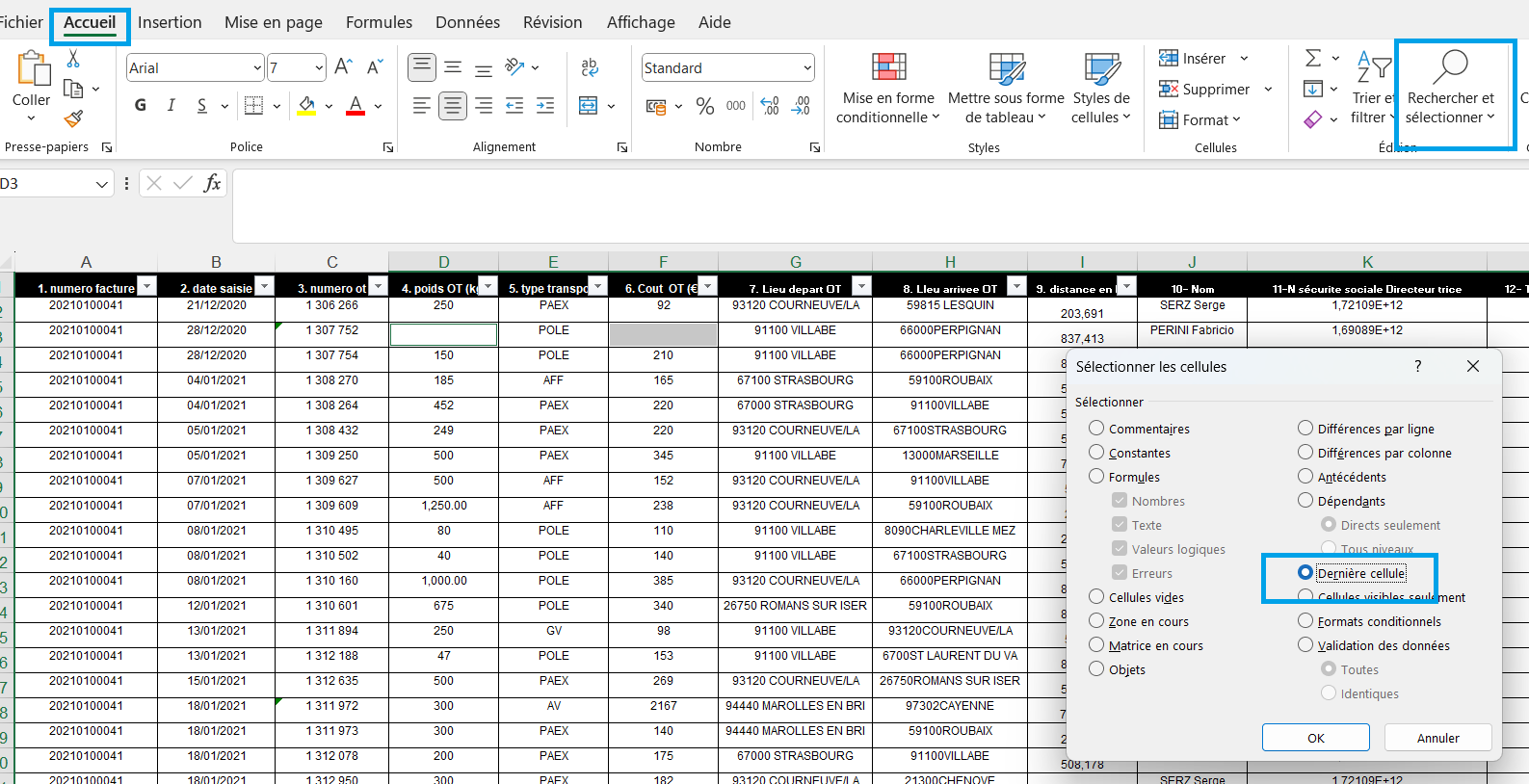
1. Vérification des Problèmes avec Excel :

• Cliquez sur l'onglet "Fichier", puis sélectionnez le menu "Informations" « Vérifiez l'absence de problèmes » , cliquez ensuite sur « inspecter le document » puis « inspecter »



2. Allègement du Poids du Fichier :

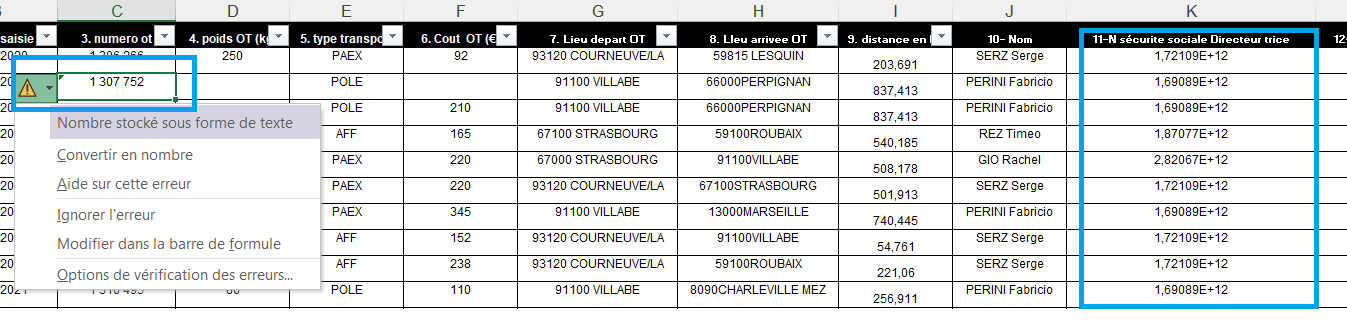
• Onglet « accueil » Cliquez sur la flèche sous l'icône "Rechercher et sélectionner", puis sélectionnez la commande "Sélectionner des cellules". Choisissez "Dernière cellule" pour alléger le poids du fichier en réduisant la portée de votre sélection :



Une fois la dernière cellule trouvée, si la cellulle est trop éloignée dans le fichier, on peut supprimer cette cellule.

3. Traitement des Données aux Mauvais Formats :

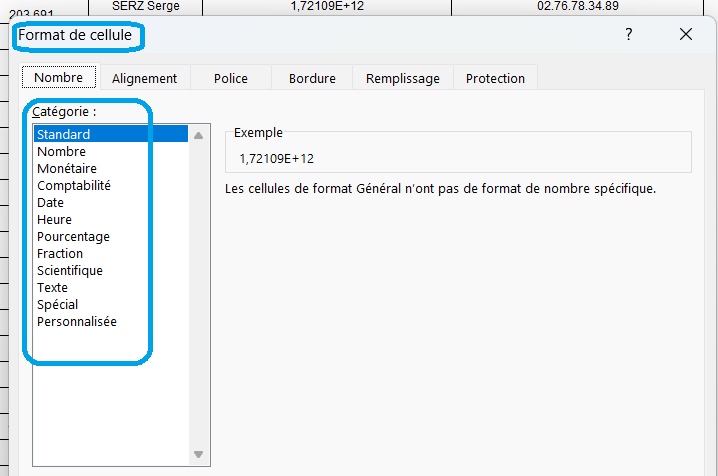
• Identifiez et corrigez les données qui ne sont pas dans le bon format, telles que les dates mal formatées ou les nombres en format texte. Assurez-vous également que les entrées sont placées dans les emplacements adéquats et corrigez les fautes de frappe.



Exemples :

colonne 11 : le format est en « standard » et la donnée n’apparait pas, il faut que le format soit « nombre ».

clique droit sur la cellule, choisir « format cellule » puis choisir le format :



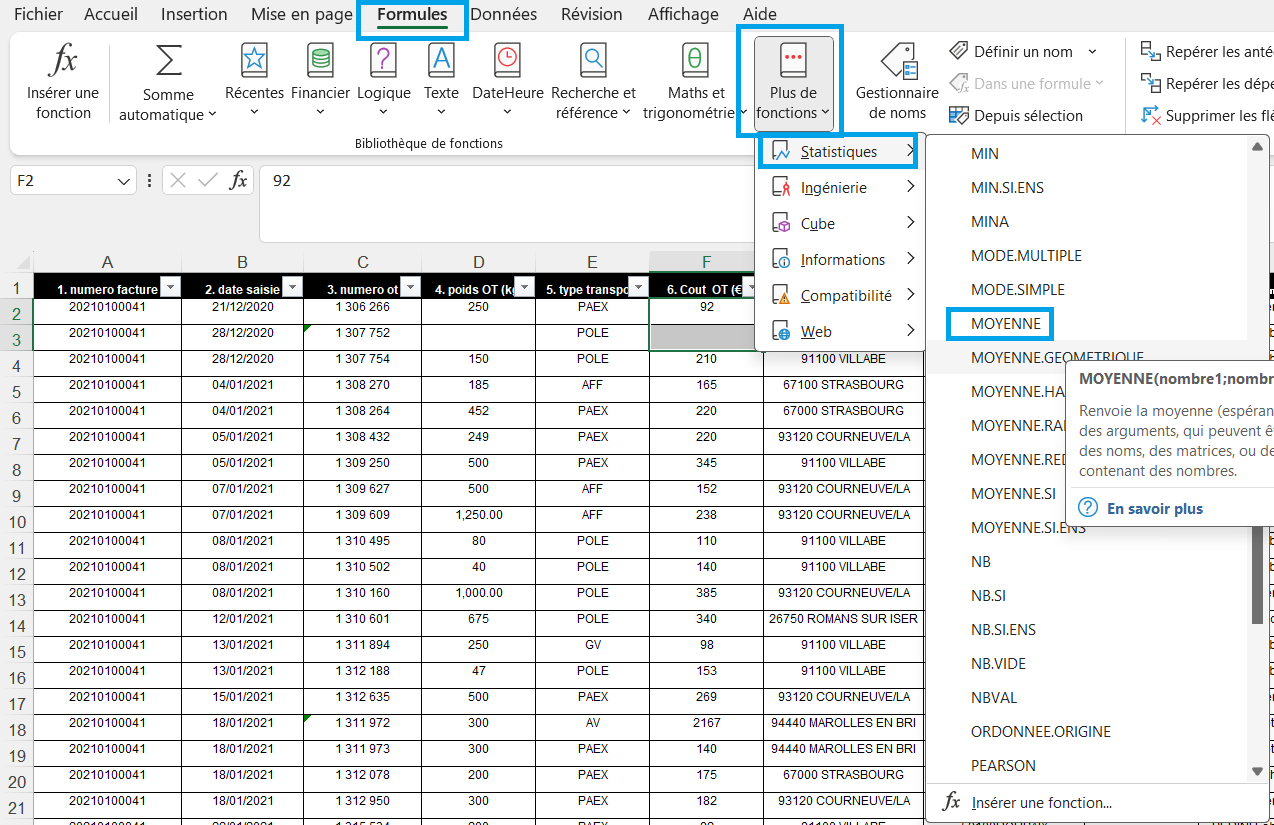
colonne 3 : un logo avec un triangle apparait ce qui indique qu’il y a une erreur de format, en cliquant sur ce dernier des options apparaissent : choisir « convertir en nombre »car la donnée est un nombre.

4. Traitement des Valeurs Manquantes :

• Identifiez la plage des valeurs manquantes et formulez des stratégies pour les traiter en fonction de leur importance. Vous pouvez choisir de supprimer les champs inutiles, de remplir le contenu manquant en utilisant des indicateurs de position (moyenne, médiane, mode, etc.), ou en effectuant des estimations plus complexes.

Si l’on choisit de donnée un valeur moyenne sur une cellule vide d’une plage de données on peut utiliser la formule moyenne

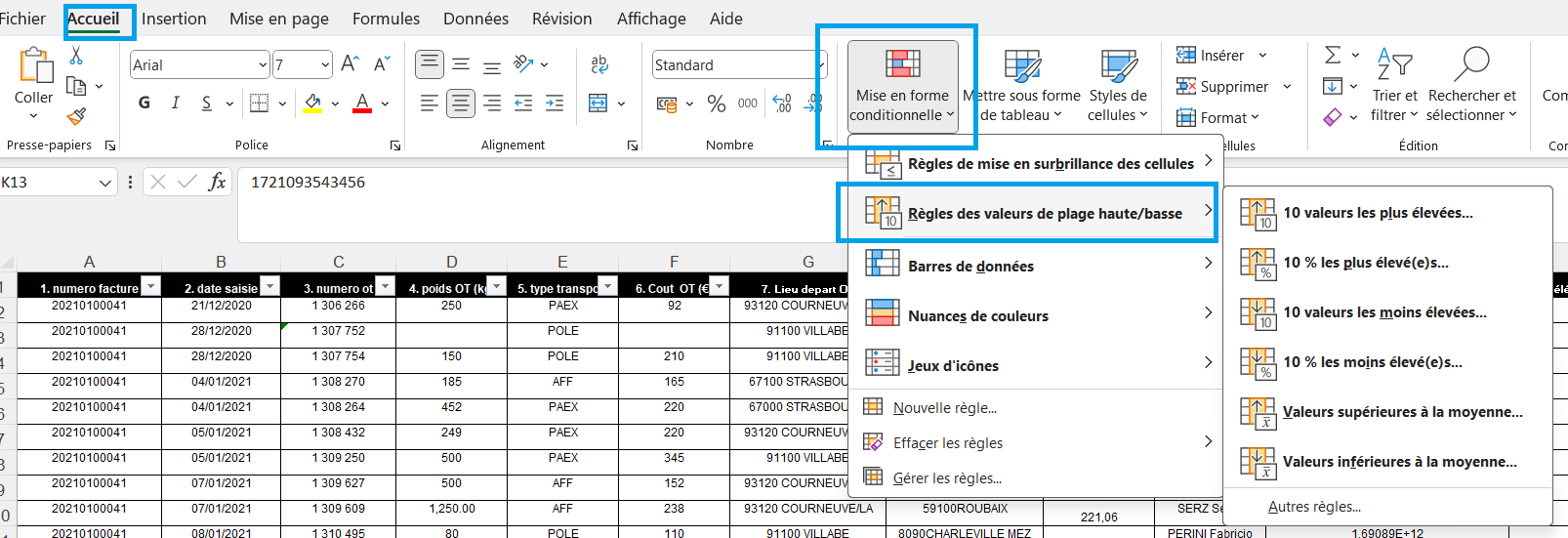
Cliquez sur l’onglet « formules » puis le menu « plus de fonctions », champs « statistiques » et ensuite la fonction « moyenne » :



5. Détection des Valeurs Aberrantes :

• Utilisez des méthodes d'analyses statistiques pour identifier les valeurs d'erreurs ou les valeurs aberrantes possibles, telles que la mise en forme conditionnelle des valeurs :

Cliquez sur l’onglet « accueil », menu « mise en forme conditionnelle », choisir « règles des valeurs de plages haute/basse » :



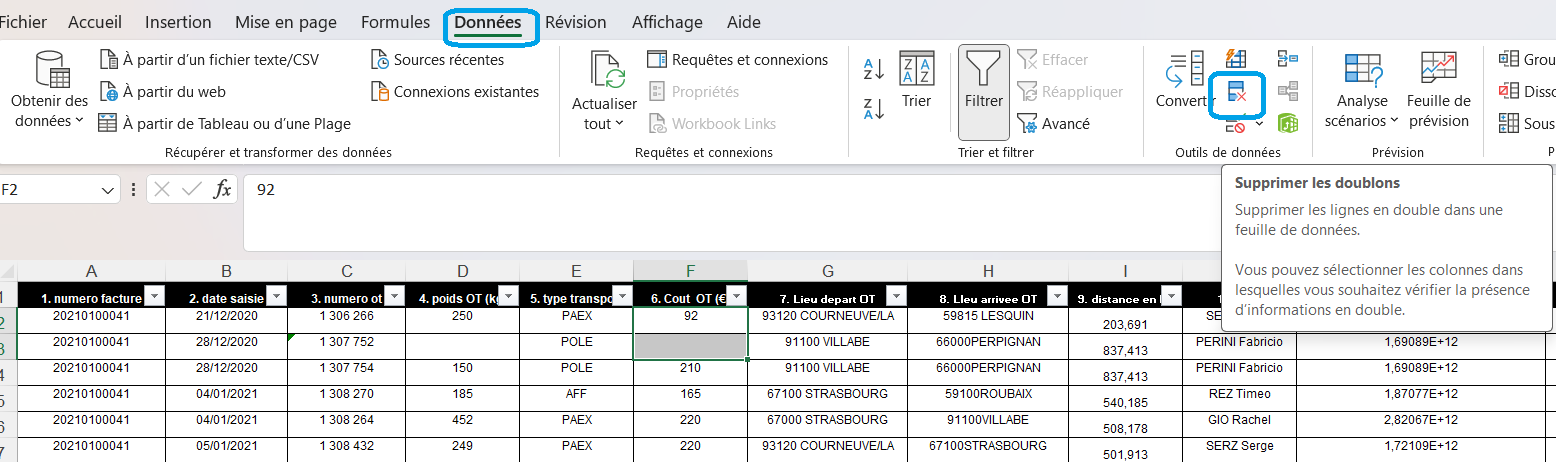
ou l'analyse des écarts ou encore l'utilisation de visualisations comme la Boxplot.

6. Élimination des Doublons :

• Les enregistrements ayant la même valeur d’attribut dans la base de données sont considérés comme des enregistrements en double, ou redondants, et sont fusionnés en un seul enregistrement. Ce phénomène se produit lorsqu’on combine des ensembles de données provenant de plusieurs sources, ou lorsque l’on reçoit des données du client.

• Utilisez l'outil de suppression de doublons dans Excel pour éliminer les lignes qui sont des doublons exacts les unes des autres, en vous assurant de ne conserver qu'une seule occurrence de chaque entrée.

Cliquez sur l’onglet « données » puis dans la section « outils de données » cliquez sur le bouton représentant 2 cellules bleues et une cellule blanche ainsi qu’une croix rouge :

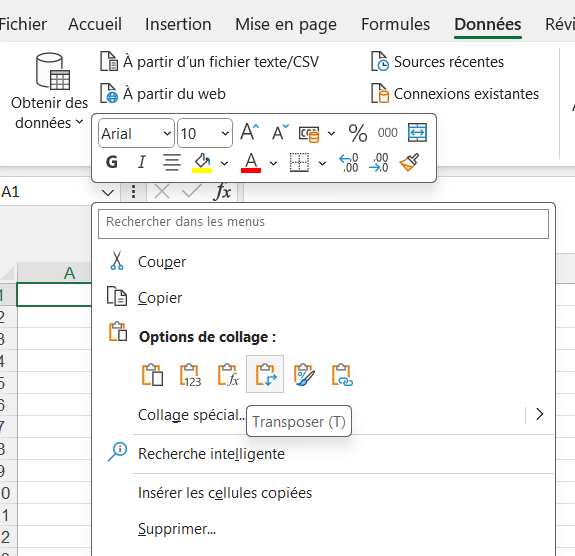


7. Transposition des Données :

• Si nécessaire, transposez les données pour réduire le nombre de colonnes et les présenter de manière plus compacte. Cela peut être utile lorsque vous avez une liste de données présentée horizontalement et que vous souhaitez la convertir en une présentation verticale.

Cela revient à faire un copier/collage spécial de données

Lorsque l’on colle les données que l’on veut transposées dans une autre configuration, il faut faire un clic droit à l’endroit où l’on veut coller les données et choisir en option de collage « transposer » :



**Étape 4 – Examiner, adapter, répéter**

En fonction de la fréquence choisie, vous devez réserver du temps chaque jour, semaine ou mois pour réévaluer votre processus de nettoyage des données. Quels sont les points positifs ? Certains aspects peuvent-ils être améliorés ? Pouvez-vous identifier des problèmes évidents ou récurrents ?