Table des matières

[1- Introduction 3](#_Toc146809684)

[2- Etapes de l’analyse des données 4](#_Toc146809685)

[2-1. Collecte des données 4](#_Toc146809686)

[2-2. Nettoyage des données 4](#_Toc146809687)

[2-3. Exploration des données 4](#_Toc146809688)

[2-4. Préparation des données 4](#_Toc146809689)

[2-5. Choix des méthodes d’analyse 4](#_Toc146809690)

[2-6. Modélisation des données 4](#_Toc146809691)

[2-7. Evaluation des modèles 5](#_Toc146809692)

[2-8. Interprétation – communication des résultats 5](#_Toc146809693)

[2-9. Répéter et itérer 5](#_Toc146809694)

[2-10. Mise en œuvre des recommandations 5](#_Toc146809695)

[3- Les différents types de données statistiques 6](#_Toc146809696)

[3-1. Exemple : 6](#_Toc146809697)

[3-2. Les types de variables 7](#_Toc146809698)

[3-2.1 Les variables quantitatives 7](#_Toc146809699)

[3-3. Les données qualitatives ou catégorielles 7](#_Toc146809700)

[4- Analyse des données – Variables quantitatives 8](#_Toc146809701)

[4-1. Présentation des données 8](#_Toc146809702)

[4-1.1 Variable quantitative discrète : Eval 8](#_Toc146809703)

[4-1.2 Variable quantitative continue : %FemaleWorkers 9](#_Toc146809704)

[4-2. Mesures de tendance centrale 11](#_Toc146809705)

[4-2.1 Nombre d’observations. 11](#_Toc146809706)

[4-2.2 Moyenne arithmétique 11](#_Toc146809707)

[4-2.3 Médiane 12](#_Toc146809708)

[4-2.4 Quartiles – écart interquartile 14](#_Toc146809709)

[4-2.5 Le mode. 14](#_Toc146809710)

[4-3. Mesures de dispersion 15](#_Toc146809711)

[4-3.1 L’étendue – valeur minimale – valeur maximale. 15](#_Toc146809712)

[4-3.2 La variance – écart type – coefficient de variation. 15](#_Toc146809713)

[4-4. Graphiques 17](#_Toc146809714)

[4-4.1 Variable quantitative continue - Histogramme 17](#_Toc146809715)

[4-4.2 Variable quantitative continue - Boite à moustaches (Box Plot) 18](#_Toc146809716)

[4-4.3 Variables quantitatives discrètes - Graphique en barres 19](#_Toc146809717)

[4-4.1 Variables quantitatives discrètes - Camembert 19](#_Toc146809718)

[5- Analyse des données – variables qualitatives ( categorielles) 20](#_Toc146809719)

[5-1. Présentation des données 20](#_Toc146809720)

[5-1.1 Variable qualitative : Product Type. 20](#_Toc146809721)

[5-1.2 Variable qualitative ordinale : Quality Rating 20](#_Toc146809722)

[5-2. Indicateurs de tendance centrale 21](#_Toc146809723)

[5-2.1 LE MODE 21](#_Toc146809724)

[5-3. Graphiques 21](#_Toc146809725)

[5-3.1 Graphique en camembert (Pie Plot) 21](#_Toc146809726)

[5-3.2 Graphique en barres 21](#_Toc146809727)

[5-4. Association de variables 22](#_Toc146809728)

[5-4.1 Corrélation linéaire – nuage de points 22](#_Toc146809729)

[5-4.2 Tableau croisé dynamique 23](#_Toc146809730)

[6- Nettoyage des données 24](#_Toc146809731)

[6-1. Données aberrantes 24](#_Toc146809732)

[6-1.1 Postal Code 24](#_Toc146809733)

[6-1.2 LineWorkers 24](#_Toc146809734)

[6-1.3 %FemaleWorkers 24](#_Toc146809735)

[6-1.4 Eval 25](#_Toc146809736)

[6-2. Données manquantes 25](#_Toc146809737)

Chapitre 2 : Le monde passionnant de l’analyse des données.

# Introduction

Les données jouent un rôle essentiel dans le monde connecté. Avec l’avènement du numérique, la quantité de données disponibles a explosé offrant de nouvelles opportunités d’analyse.

Que ce soit pour mieux comprendre une situation ou pour gérer l’incertitude dans les systèmes d’IA, tous les algorithmes sont basés sur les statistiques.

Les statistiques sont fondamentales en tant qu’outil d’aide à la compréhension des données dans divers domaines. En extrayant des informations significatives à partir d’ensemble de données complexes, l’analyse statistique des données massives permet de dégager des tendances, de découvrir des relations cachées, de prédire des tendances futures. Elles fournissent un moyen objectif d’analyser des phénomènes variés, qu’il s’agisse de performances sportives, de tendances économiques ou même de résultats scientifiques. Grâce à des techniques telles que la moyenne, l’écart-type, la corrélation et la régression, les statistiques permettent de transformer des données brutes en connaissances exploitables.

Dans le domaine de l’IA, les statistiques agissent comme un pont entre les données et l’intelligence. Elles permettent de créer des systèmes intelligents capables de prendre des décisions complexes en se basant sur des données historiques et en temps réel.

Il est important de noter que, pour que la machine puisse apprendre, il ne suffit pas de choisir une collection d’observations, de choisir un algorithme d’apprentissage et de cliquer sur un bouton pour obtenir une bonne prédiction. Il faut soi-même comprendre ce jeu d’observations pour en sélectionner les informations pertinentes et importantes. Pour cela, les statistiques seront d’une grande utilité. C’est pourquoi ce chapitre est dédié à l’analyse de données et aux statistiques.

# Etapes de l’analyse des données

## Collecte des données

Consiste à rassembler toutes les données pertinentes. Ces données peuvent être collectées de différentes manières :

* en provenance de bases de données : fichier des usines Nike;
* en provenance des réseaux sociaux : il faut utiliser une API appropriée
* en provenance de capteurs : Kaggle <https://www.kaggle.com/datasets/uciml/human-activity-recognition-with-smartphones>
* en provenance d’enquêtes : <https://forms.office.com/Pages/DesignPageV2.aspx?origin=NeoPortalPage&subpage=design&id=SAQmcanxKUe2MEiC5NxY9WtA0Qojja1MhXTRmko2BEtUOUFEUzFEUVVQR000SFZBNUZUSFJQU0ozVy4u>
* en provenance d’appareils photo, de caméras : Kaggle <https://www.kaggle.com/competitions/petfinder-pawpularity-score/discussion/287448>
* en provenance d’un site internet.

## Nettoyage des données

Les données brutes sont souvent incomplètes, incorrectes ou incohérentes. Le nettoyage des données implique la gestion des valeurs aberrantes, la correction des erreurs et la gestion des données manquantes (cfr ci-dessous).

## Exploration des données

A ce stade, les données sont explorées pour comprendre leur structure. Des statistiques descriptives, des graphiques et des visualisations sont utilisés pour révéler des tendances et des corrélations (cfr ci-dessous).

## Préparation des données

Etape qui consiste à préparer les données pour l’analyse proprement dite. Il peut s’agir de normaliser ou formater les données, de réduire la dimensionnalité ou de créer de nouvelles variables plus pertinentes.

## Choix des méthodes d’analyse

En fonction de la nature des données et des objectifs de l’analyse, différentes méthodes peuvent être choisies : le monitoring via un tableau de bord, la régression, la classification, le clustering, l’analyse des séries temporelles.

## Modélisation des données

Construction de modèles statistiques ou d’apprentissage automatique pour extraire les informations utiles à partir des données. Exemples : modèles de Machine Learning, ajustement des courbes ou autres techniques d’analytiques. Il est à noter que vous, en tant qu’informaticiens, n’aurez pas à élaborer ces modèles mais bien d’en comprendre les bases afin de pouvoir gérer leur mise en œuvre au niveau informatique.

## Evaluation des modèles

Etape cruciale : mesurer les performances du modèle et utiliser les métriques appropriées.

## Interprétation – communication des résultats

Les résultats sont interprétés pour tirer des conclusions qu’il faudra expliquer souvent à un public non technique. Communiquer de manière claire et efficace aux parties prenantes fera partie de votre métier.

## Répéter et itérer

Le processus est souvent itératif. Si de nouvelles données deviennent disponibles notamment, le processus peut être répété pour affiner les modèles et les conclusions.

## Mise en œuvre des recommandations

Les résultats de l’analyse peuvent être à la base de prises de décisions et de la mise en œuvre d’actions concrètes.

L’analyse des données est un processus continu et évolutif qui demande des compétences en statistiques, en programmation, en visualisation des données et en compréhension du domaine d’application. Elle joue un rôle central dans de nombreuses disciplines.

Les chapitres suivants présenteront les outils statistiques de base pour réaliser une analyse de données. L’idée n’étant pas de donner un cours de statistiques mais bien d’aborder de manière intuitive l’analyse de données.

# Les différents types de données statistiques

Dans cette section, nous allons explorer les différents types de données statistiques ainsi que leur signification en analyse des données (data analysis)

## Exemple :

Le fichier Nike est un fichier de données qui répertorie différentes usines de fabrication associées à Nike, Inc. Ce fichier contient plusieurs variables qui décrivent chaque usine notamment leur emplacement, leur type de production et d'autres détails pertinents. Les données peuvent être utilisées pour la gestion de la chaîne d'approvisionnement, la surveillance des conditions de travail, l'évaluation de la diversité de la main-d'œuvre, et d'autres analyses liées aux opérations de fabrication de Nike, Inc.

Toute étude statistique se réalise sur une **population** (ici les usines de fabrication associées à Nike), Inc, composée d'**individus** : dans cet exemple, un individu = une usine. Ces individus peuvent être des personnes ou des choses. Chaque individu dispose de **caractères** correspondant aux variables statistiques comme Factory Name.

**Attention, certaines colonnes ont été ajoutées ou modifiées à des fins pédagogiques.**Voici une description de chaque variable :

* Factory Name (Nom de l'usine) : Le nom de l'usine de fabrication.
* Factory Type (Type d'usine) : Le type d'usine de fabrication, par exemple, "FINISHED GOODS" (biens finis) ou "Equipment" (équipement).
* Product Type (Type de produit) : Le type de produit fabriqué dans cette usine, par exemple, 1 - Apparel (vêtements) // 2 – Equipment (équipement) // 3 – Footwear (chaussures).
* Postal Code (Code postal) : Le code postal de l'emplacement de l'usine.
* Country / Region (Pays / Région) : Le pays ou la région où se trouve l'usine.
* Region (Région) : La région géographique à laquelle appartient cette usine, par exemple, "AMERICAS" (Amériques), "SE ASIA" (Asie du Sud-Est) ou "EMEA" (Europe, Moyen-Orient et Afrique).
* Total Workers (Nombre total de travailleurs) : Le nombre total de travailleurs employés dans cette usine.
* Line Workers (Travailleurs en ligne) : Le nombre de travailleurs occupés dans la chaîne de production.
* % Female Workers (% de travailleurs féminins) : Le pourcentage de travailleurs de sexe féminin parmi le personnel de l'usine.
* % Migrant Workers (% de travailleurs migrants) : Le pourcentage de travailleurs migrants parmi le personnel de l'usine.
* Eval (Évaluation) : Une évaluation ou une note attribuée à cette usine. Valeurs possibles de 0 à 10.
* Quality Rating (note de qualité) : qualité des produits dans l’usine (1 mauvais – 2 moyen – 3 bon).

## Les types de variables

Il existe deux types de variables : les variables quantitatives et les données qualitatives.

### Les variables quantitatives

Les variables quantitatives sont des nombres sur lesquels il est possible d’effectuer des opérations mathématiques. Ces données peuvent être classées en deux groupes distincts :

* **les données quantitatives discrètes** : elles peuvent prendre un nombre fini de valeurs possibles, celles-ci peuvent être énumérées (ex : des notes entières, nombre de pattes d’un animal …) ;
* **les données quantitatives continues** ayant une infinité de valeurs possibles (ex : temps de parcours – vitesse – taille – poids) ;  
  Une variable discrète dont le nombre de valeurs différentes est élevé est assimilée à une variable quantitative continue.

Exemple Nike - variables quantitatives discrètes et continues

**Candidats** : Postal Code, Total workers, Line workers, %Female workers, %Migrant workers, Eval (nombres entiers) et Product Type Type.

* Postal Code : ne va pas être considérée comme une variable quantitative. En effet, même si ce sont des nombres, quel sens donner à une moyenne de codes postaux par exemple ?
* TotalWorkers et LineWorkers sont des « nombre de » => bons candidats mais il y a une grande quantité de valeurs possibles. Elles seront donc traitées comme des données quantitatives continues.
* %FemaleWorkers et %MigrantWorkers sont des pourcentages ici exprimés en nombres entiers. D’une part, la grande variété de valeurs possibles et d’autres part, le fait que le pourcentage pourrait être exprimé en valeurs réelles (avec des décimales) mèneront aussi à considérer ces deux variables comme quantitatives continues.
* Eval : seulement 11 valeurs possibles et des indicateurs de type moyenne ont un sens => variables quantitatives discrètes.
* Product Type : indicateur de type moyenne : pas de sens.

## Les données qualitatives ou catégorielles

Les données qualitatives ont pour but de décrire des qualités propres aux données : est-ce un homme, une femme? Est-ce petit ou grand? Ces données ne sont pas des nombres et sont également réparties en deux groupes distincts :

* **les données qualitatives ne pouvant être ordonnées** : homme ou femme, chien ou chat. ..
* **les données qualitatives ordinales** pouvant être ordonnées (petit, moyen, grand).

Exemple Nike - variables qualitatives

Factory Name, Factory Type, Product Type, Postal Code, Country/Region, Region, Quality Rating sont toutes des variables qualitatives mais observons plus précisément :

* Product Type Type : les valeurs 1, 2, 3 marquent les différents types de produits mais le fait de mettre des chiffres à la place de caractères induit une relation d’ordre qui en réalité n’existe pas entre les différents produits. Je vais donc dans le fichier remplacer ces chiffres par des chaines de caractères.
* Quality Rating : inversement, nous avons des chaines de caractères « mauvais – moyen – bon » qui, dans leur signification, indiquent une relation d’ordre. On pourrait donc les changer en 1,2,3 => qualitative ordinale.

Connaître les différents types de données, permet d'obtenir une première analyse des données et de définir les traitements que nous serons amenés à effectuer.

# Analyse des données – Variables quantitatives

## Présentation des données

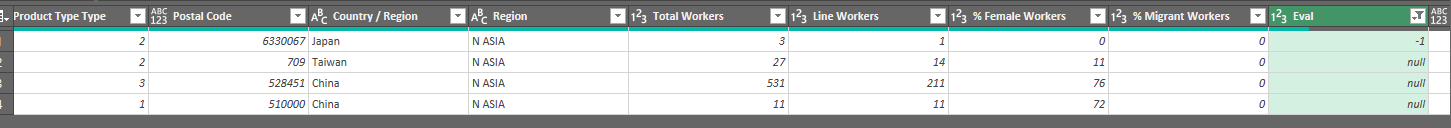
Pour cette première étape, l’idée est de passer de la donnée à l’information … quitte à perdre parfois un peu en précision par rapport aux données de départ.

### Variable quantitative discrète : Eval

Les différentes valeurs rencontrées : Une image contenant texte, capture d’écran, logiciel, Icône d’ordinateur

Description générée automatiquement

On y voit null et -1.

Null et -1 correspondent à 4 enregistrements *(cfr vidéo : Mise en forme tableau)*:   


Que faire ?

On verra plus tard : pour l’instant, nous n’allons pas en tenir compte.

Présentation

Une image contenant texte, capture d’écran, nombre, Police

Description générée automatiquement *(cfr vidéo présentation)*

Avantage : ce tableau est plus simple à lire.-

### Variable quantitative continue : %FemaleWorkers

En appliquant le même procédé que ci-dessus, la présentation n’apporte pas d’intérêt (506 valeurs différentes), c’est pourquoi nous les agencerons en classes quitte à perdre de l’information :   
Une image contenant texte, capture d’écran, nombre, logiciel

Description générée automatiquement ….. Une image contenant texte, capture d’écran, nombre, Police

Description générée automatiquement

Présentation

|  |  |  |
| --- | --- | --- |
| Limite Inférieure | Limite supérieure | Fréquence=ri |
|  |
| [0; | 10[ | 30 |  |
| [10; | 20[ | 29 |  |
| [20 | 30[ | 24 |  |
| [30; | 40[ | 42 |  |
| [40; | 50[ | 55 |  |
| [50; | 60[ | 78 |  |
| [60; | 70[ | 76 |  |
| [70; | 80[ | 100 |  |
| [80; | 90[ | 127 |  |
| [90; | 100] | 78 |  |

Nous allons dans un premier temps procéder à la déduction d'une tendance centrale, c'est-à-dire essayer de déterminer une valeur autour de laquelle se concentre l'ensemble des notes, puis nous étudierons sa dispersion et enfin nous rechercherons l'existence de données aberrantes, c’est-à-dire non représentatives de l'ensemble des notes.

Ce type d'analyse devra être effectué sur toute série d'observations que vous aurez à utiliser dans le cadre de projets de Machine Learning. En effet, réaliser les mesures de tendance et de dispersion de vos données vous permettra de comprendre leur signification, leur rôle et leur utilité dans la prédiction à effectuer, identifier les données aberrantes vous permettra d'exclure éventuellement des données non représentatives.

## Mesures de tendance centrale

Les mesures de tendance centrale servent à synthétiser la série statistique étudiée au moyen d'un petit nombre d’indicateurs. En d'autres termes, nous allons essayer de trouver un certain nombre de valeurs autour desquelles se regroupe l'ensemble des notes obtenues aux différents examens.

### Nombre d’observations.

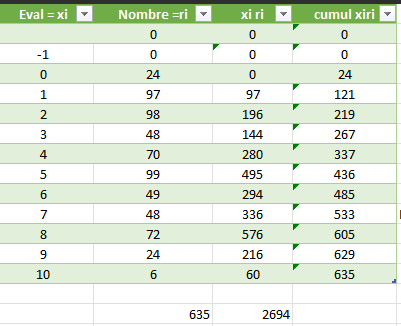
La première chose à connaître lorsque nous étudions un ensemble d'observations c'est leur nombre. En effet, en connaissant le nombre d'observations attendues, il est possible de détecter plus rapidement si certaines données sont manquantes et si nous disposons d’assez de données pour permettre un apprentissage.   
NB : pour les prochains calculs de la variable « Eval », nous ne tiendrons pas compte des données aberrantes pour l’instant.

### Moyenne arithmétique

La moyenne est un indicateur qui permet de caractériser une série statistique. Elle permet de connaître la valeur des éléments de notre série statistique si celle-ci était répartie uniformément. Dans notre cas, la moyenne permet de répondre à la question suivante :    
"Si nous devions donner la même valeur à chacune de notes d’évaluation laquelle serait-elle?".

Dans notre exemple,

Figure 1

 *(cfr vidéo moyenne arithmétique)*

Interprétation : si la même performance était réalisée dans chaque usine, la note obtenue dans chacune d’elle, aurait été de 4,24 (=2694/635).

Exemple continu : %FemaleWorkers -> *cfr vidéo fonctionMoyenne*

### Médiane

La médiane représente la valeur au centre d’un ensemble de données lorsque celles-ci sont triées par ordre croissant. La médiane peut offrir une mesure de tendance centrale plus robuste que la moyenne qui peut être influencée par les valeurs extrêmes.

Pour la calculer, il suffit de ranger les valeurs par ordre croissant et de choisir la valeur qui se trouve au milieu de la distribution. Si le nombre d’observations est impair, cette valeur est précisément la médiane. En revanche, si le nombre d’observations est pair, la médiane est la moyenne des deux valeurs du milieu.

La médiane est particulièrement utile lorsque les données contiennent des valeurs aberrantes ou sont asymétriques car elle n’est pas autant affectée par ces situations que la moyenne.

Exemple Eval

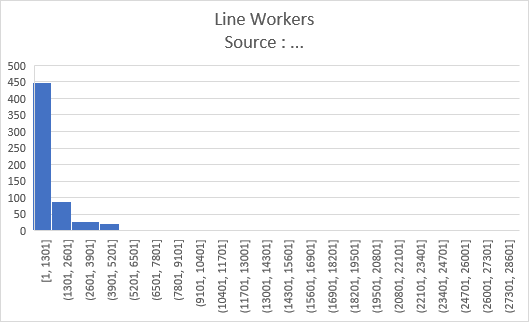
639 valeurs donc celle du milieu est la 320ème. Ainsi 319 valeurs avant et 319 valeurs après. En regardant dans le tableau figure 1 , la médiane est donc de 4. Le logiciel peut le faire via la fonction MEDIANE et donne 4.

Exemples LineWorkers et %FemaleWorkers

Pour les variables continues, le principe de la médiane est le même, à savoir subdiviser l’ensemble des données en deux parties égales. Ici, le logiciel donne comme valeur médiane 551. Ainsi, 50% des valeurs sont inférieures à 551 et, bien sûr, 50% sont au-delà de cette valeur.

La médiane est particulièrement utile lorsque les données contiennent des valeurs aberrantes ou sont asymétriques car elle n’est pas autant affectée par ces situations que la moyenne.

Voici deux schémas pour clarifier :



Médiane = 551 et moyenne = 1665. La médiane est plus représentative des données car les valeurs extrêmes, même si elles sont peu nombreuses, influencent la moyenne.

Une image contenant texte, capture d’écran, Tracé, diagramme

Description générée automatiquement

Moyenne = 62.7% et médiane = 63% donc moyenne et médiane sont assez proches l’une de l’autre, la distribution des valeurs est répartie « plus uniformément ».

### Quartiles – écart interquartile

Dans le même ordre d’idées, on peut faire la subdivision par quarts :

Quartile 1 : 25% des valeurs observées sont inférieures à la valeur du quartile 1.  
Quartile 2 = médiane.  
Quartile 3 : 75% des valeurs observées sont inférieures à la valeur du quartile 3.

L’écart interquartile est la différence entre le quartile 3 et le quartile 1 :

### Le mode.

Le mode est un autre indicateur de tendance centrale en statistique, quel que soit le type de variable y compris les qualitatives. Il représente la valeur qui apparait le plus fréquemment dans un ensemble de données. En d’autres termes, c’est la valeur qui a la fréquence la plus élevée.

Le mode peut être un indicateur utile pour identifier les valeurs les plus prédominantes dans un ensemble de données et pour déterminer les tendances dominantes au sein de la distribution.

Une distribution peut avoir un mode unique ou plusieurs modes si plusieurs valeurs ont la même fréquence maximale. Contrairement à la moyenne, le mode n’est pas affecté par les valeurs extrêmes.

Exemple Eval

Mode = 5. Il apparaît 99 fois.

Exemple %FemaleWorkers

|  |  |  |
| --- | --- | --- |
| Limite Inférieure | Limite supérieure | Fréquence=ri |
|  |
| [0; | 10[ | 30 |  |
| [10; | 20[ | 29 |  |
| [20 | 30[ | 24 |  |
| [30; | 40[ | 42 |  |
| [40; | 50[ | 55 |  |
| [50; | 60[ | 78 |  |
| [60; | 70[ | 76 |  |
| [70; | 80[ | 100 |  |
| [80; | 90[ | 127 |  |
| [90; | 100] | 78 |  |

Sur base de ce tableau, nous pouvons voir que la classe modale est [80;90[ avec 127 répétitions.

## Mesures de dispersion

La dispersion des données dans une série statistique se réfère à la manière dont les valeurs sont réparties autour d’une mesure de tendance centrale. Une mesure de dispersion permet de quantifier l’étendue des valeurs et à quel point les données sont dispersées ou regroupées.

### L’étendue – valeur minimale – valeur maximale.

Connaître la valeur minimale et la valeur maximale d'un ensemble d'observations est toujours intéressant pour établir son étendue.

Exemple Eval

Valeur min = 0 – valeur max = 10 donc l’étendue = 10

Exemple LineWorkers

Valeur min = 1 – valeur max = 27305 donc l’étendue = 27304. On constate, dans ce cas, une grande étendue des données.

### La variance – écart type – coefficient de variation.

La variance mesure la moyenne des carrés des écarts entre chaque valeur et la moyenne. L’écart-type est la racine carrée de la variance. Plus la variance ou l’écart-type est élevé, plus les données sont dispersées. Il est à noter que, pour l’interprétation, nous nous baserons plus sur la valeur de l’écart-type qui est exprimé dans la même unité que les données d’observation plutôt que sur celle de la variance qui est exprimée dans cette même unité au carré.

En général, les logiciels calculent l’écart-type de cette manière : pour des variables quantitatives discrètes et pour des quantitatives continues.

Dans Excel, utiliser la fonction ECART. STANDARD

Exemple Eval

2.63

Exemple Line Workers

3156.56

Il est à noter que ces valeurs ne sont pas interprétables en elles-mêmes. Si l’ordre de grandeur des valeurs est de 1000, un écart-type de 2.63 est très petit ; si l’ordre de grandeur est de 2, cet écart-type est très grand. C’est la raison pour laquelle on le compare souvent à la moyenne ce qui donne le **coefficient de variation** . On considère en général que si ce coefficient est inférieur à 30%, la dispersion des valeurs autour de la moyenne est faible.

## Graphiques

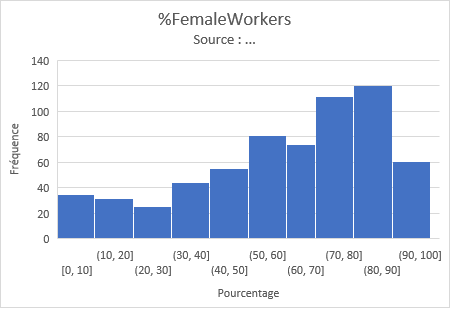
**WARNING** : un graphique doit toujours être accompagné

* de la source des données + date de recensement ;
* du titre de l’axe des abscisses et de l’unité éventuelle ;
* du titre de l’axe des ordonnées et de l’unité éventuelle ;
* du titre du graphique.

Parmi les graphiques les plus couramment utilisés :

### Variable quantitative continue - Histogramme

Graphique qui représente la distribution des données quantitatives continues en regroupant les valeurs par intervalles sur l’axe horizontal et en affichant la fréquence ou le pourcentage sur l’axe vertical. Il permet de visualiser la forme de la distribution des données.



Interprétation :

### Variable quantitative continue - Boite à moustaches (Box Plot)

Figure 2: Vidéo Box Plot - Titres

Interprétation :

Une image contenant texte, diagramme, ligne, capture d’écran

Description générée automatiquement

Une image contenant texte, capture d’écran, ligne, diagramme

Description générée automatiquement

### Variables quantitatives discrètes - Graphique en barres

Adapté pour les variables quantitatives discrètes.

### Variables quantitatives discrètes - Camembert

# Analyse des données – variables qualitatives ( categorielles)

## Présentation des données

### Variable qualitative : Product Type.

Après quelques transformations similaires à la figure 1 :

Une image contenant texte, capture d’écran, Police, nombre

Description générée automatiquement

### Variable qualitative ordinale : Quality Rating

Avec ou sans nombres

|  |  |  |
| --- | --- | --- |
| Quality rating | | |
| Valeur | fréquence | pourcentage |
| 1 | 128 | 20 |
| 2 | 113 | 18 |
| 3 | 398 | 62 |

Pourquoi les pourcentages ne correspondent-ils pas à ceux de la figure 2?

|  |  |  |
| --- | --- | --- |
| Quality rating | | |
| Mention | fréquence | pourcentage |
| Mauvais | 128 | 20 |
| Moyen | 113 | 18 |
| Bon | 398 | 62 |
|  |  |  |
|  |  |  |

## Indicateurs de tendance centrale

Dans le cas des variables qualitatives, même pour les variables qualitatives ordinales, les indicateurs numériques de tendance centrale ou de dispersion n’ont pas de sens SAUF le mode.

### LE MODE

Idem variables quantitatives.

Dans le cas « Quality Rating », le mode est « Bon » correspondant à 398 valeurs.

Pourquoi la moyenne n’a-t-elle pas de sens dans le cas du « Quality Rating » écrit sous la forme de chiffres ?

## Graphiques

### Graphique en camembert (Pie Plot)

Figure 3 : Vidéo Graphique en camembert

### Graphique en barres

De nouveau, attention. Par défaut, mauvaise interprétation des chiffres en prenant la version chiffrée.

Figure 4 : Vidéo Graphique en barres

ATTENTION : pour les variables qualitatives ordinales, veillez à garder l’ordre naturel des valeurs sur les graphiques afin que celui-ci soit plus facilement lisible.

## Association de variables

### Corrélation linéaire – nuage de points

Y a-t-il une corrélation linéaire entre le pourcentage de femmes employées dans l’usine et la note d’évaluation de l’usine ?

Aucune corrélation ci-dessus.

Sur le schéma suivant, on peut voir une répartition du nuage de points qui indique une tendance linéaire ou non linéaire

Une image contenant ligne, diagramme, Tracé

Description générée automatiquement

Corrélation non linéaire

Corrélation linéaire

### Tableau croisé dynamique

Quelle est la moyenne des évaluations par région + graphique ?

|  |  |
| --- | --- |
| **Étiquettes de lignes** | **Moyenne de Eval** |
| AMERICAS | 3,90 |
| EMEA | 4,00 |
| N ASIA | 4,28 |
| S ASIA | 4,27 |
| SE ASIA | 4,37 |
| **Total général** | **4,24** |

Figure 5 : Vidéo TCD

Une image contenant texte, capture d’écran, Police, nombre

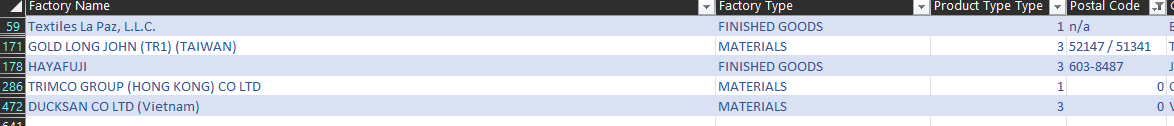
Description générée automatiquement

# Nettoyage des données

## Données aberrantes

### Postal Code

Il y a un code postal à 0, des codes avec des tirets, avec des lettres, un n/a :

Il faudrait investiguer.

### LineWorkers

On pourrait vérifier que le nombre de LineWorkers est partout inférieur au nombre TotalWorkers.  
Pas de valeur bizarre.

### %FemaleWorkers

Il y a un seul pourcentage avec des décimales et, par ailleurs, parfois des pourcentages très élevés notamment un 100%.

Une image contenant texte, nombre, Police, ligne

Description générée automatiquementOn peut remarquer que les pourcentages élevés de femmes se situent dans des pays comme la Chine, l’Indonésie, Taiwan, Sri Lanka, Vietnam… Ce ne sont donc sans doute pas des valeurs aberrantes.

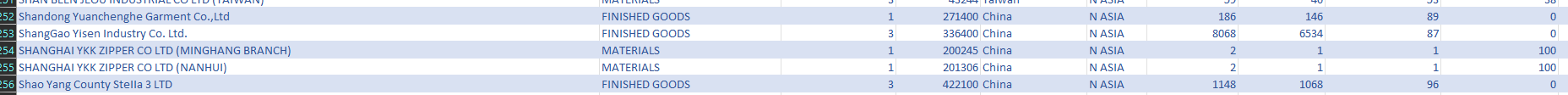
On pourrait aussi vérifier l’inverse càd « peut-on dire que toutes les usines de ces pays n’emploient que des femmes ? »

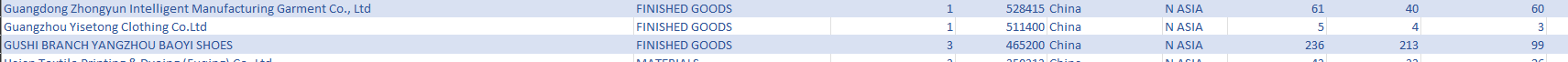
Il semble que non. Voici une partie du résultat :

Une image contenant texte, capture d’écran, nombre, ligne

Description générée automatiquement

On peut remarquer au passage les lignes suivantes qui ne semblent pas cohérentes au niveau des pourcentages :





Etc…

Dans tous ces cas, il faut investiguer. Exemple d’investigation, prenons la variable Eval.

### Eval

Cette colonne contient un -1 et du vide.

Que faire pour le -1 ? supprimer les lignes ou trouver une valeur « logique » de remplacement.

* Supprimer vu le petit pourcentage à supprimer : 1/639. Attention qu’en supprimant plusieurs lignes, on pourrait supprimer une classe ;
* Regarder la valeur de Eval dans des lignes qui sont similaires à celle que l’on veut corriger :



Japon, peu de travailleurs, bonne qualité

Une image contenant texte, nombre, ligne, Parallèle

Description générée automatiquementOn remarque que pour les usines du Japon avec peu de travailleurs et bonne qualité, en général, la note est aux alentours de 7.

* Remplacer par la moyenne de la colonne :

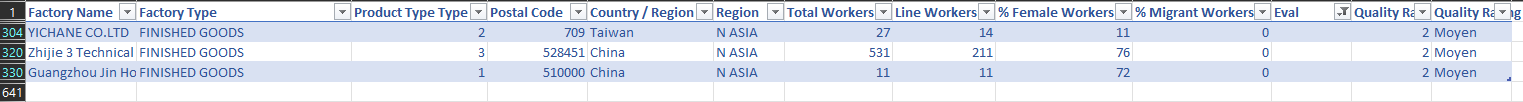
Moyenne : 4.24

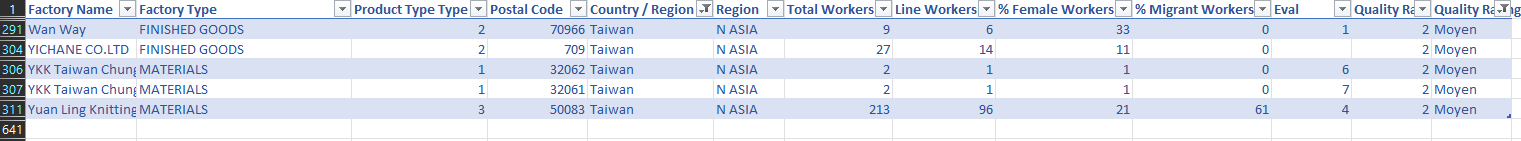
* Remplacer par la médiane de la colonne :

Médiane : 4

## Données manquantes

Même principe.

En isolant Taiwan, Moyen :

Une seule valeur pour type produit =2 => remplacement par 1 ? En isolant China, Moyen : on obtient une bonne trentaine de lignes. Remplacer par la moyenne ou la médiane et voir si ça a du sens.