## Description statistique

### Variables numériques

Comme indiqué en introduction, le jeu de données contient **14 variables numériques,** ainsi que **2 variables catégorielles** que nous choisissons de traiter comme des variables numériques pour cette étude.

Table 1 présente des statistiques descriptives pour l’ensemble de ces **16 variables.**

Table 1. Statistiques descriptives pour les 16 variables numériques.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **MinTemp** | **MaxTemp** | **Rainfall** | **Evaporation** | **Sunshine** |
| count | 143975.000000 | 144199.000000 | 142199.000000 | 82670.000000 | 75625.000000 |
| mean | 12.194034 | 23.221348 | 2.360918 | 5.468232 | 7.611178 |
| std | 6.398495 | 7.119049 | 8.478060 | 4.193704 | 3.785483 |
| min | -8.500000 | -4.800000 | 0.000000 | 0.000000 | 0.000000 |
| 25% | 7.600000 | 17.900000 | 0.000000 | 2.600000 | 4.800000 |
| 50% | 12.000000 | 22.600000 | 0.000000 | 4.800000 | 8.400000 |
| 75% | 16.900000 | 28.200000 | 0.800000 | 7.400000 | 10.600000 |
| max | 33.900000 | 48.100000 | 371.000000 | 145.000000 | 14.500000 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **WindGustSpeed** | **WindSpeed9am** | **WindSpeed3pm** | **Humidity9am** | **Humidity3pm** |
| count | 135197.000000 | 143693.000000 | 142398.000000 | 142806.000000 | 140953.000000 |
| mean | 40.035230 | 14.043426 | 18.662657 | 68.880831 | 51.539116 |
| std | 13.607062 | 8.915375 | 8.809800 | 19.029164 | 20.795902 |
| min | 6.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 25% | 31.000000 | 7.000000 | 13.000000 | 57.000000 | 37.000000 |
| 50% | 39.000000 | 13.000000 | 19.000000 | 70.000000 | 52.000000 |
| 75% | 48.000000 | 19.000000 | 24.000000 | 83.000000 | 66.000000 |
| max | 135.000000 | 130.000000 | 87.000000 | 100.000000 | 100.000000 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Pressure9am** | **Pressure3pm** | **Cloud9am** | **Cloud3pm** | **Temp9am** | **Temp3pm** |
| count | 130395.00000 | 130432.000000 | 89572.000000 | 86102.000000 | 143693.000000 | 141851.00000 |
| mean | 1017.64994 | 1015.255889 | 4.447461 | 4.509930 | 16.990631 | 21.68339 |
| std | 7.10653 | 7.037414 | 2.887159 | 2.720357 | 6.488753 | 6.93665 |
| min | 980.50000 | 977.100000 | 0.000000 | 0.000000 | -7.200000 | -5.40000 |
| 25% | 1012.90000 | 1010.400000 | 1.000000 | 2.000000 | 12.300000 | 16.60000 |
| 50% | 1017.60000 | 1015.200000 | 5.000000 | 5.000000 | 16.700000 | 21.10000 |
| 75% | 1022.40000 | 1020.000000 | 7.000000 | 7.000000 | 21.600000 | 26.40000 |
| max | 1041.00000 | 1039.600000 | 9.000000 | 9.000000 | 40.200000 | 46.70000 |

À première vue, tout paraît cohérent. Par exemple, les valeurs minimales et maximales semblent plausibles.

Toutefois, en regardant de plus près, nous décelons **quelques irrégularités.** Par exemple, pour la variable Rainfall, 75% des valeurs se trouvent en dessous de 0.8, alors que le maximum est de 371. Son écart-type (8.48) est également plusieurs fois plus grand que sa moyenne (2.36). Nous pouvons constater un phénomène similaire pour la variable Evaporation, dont 75% des valeurs se trouvent en dessous de 7.4, alors que le maximum est de 145.

Ces observations semblent **difficilement imputables aux valeurs manquantes.** À titre d’exemple, la variable Sunshine, dont le taux de NaN est de 48.01%, présente une distribution tout à fait cohérente.

Afin de mieux appréhender ces tendances, il est utile de les **visualiser** après suppression des valeurs manquantes et mise à l’échelle par standardisation :

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

Description générée automatiquement

Figure 1. Les distributions statistiques des 16 variables numériques.

**Figure 1** met en évidence la présence de **plusieurs distributions** aux **schémas inhabituels,** caractérisés par de **nombreuses valeurs extrêmes,** notamment dans le cas des variables suivantes :

* Rainfall
* Evaporation
* WindGustSpeed
* WindSpeed9am
* WindSpeed3pm
* Humidity9am
* Pressure9am
* Pressure3pm

Regardons une de ces variables de plus près.

Figure 2 présente la distribution de Evaporation (en valeurs absolues, et non plus standardisées).

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

Description générée automatiquement

Figure 2. La distribution statistique de Evaporation.

Ce graphique rejoint les premières observations effectuées à partir des statistiques descriptives, à savoir, qu’une grande majorité des valeurs (75%) se trouvent en dessous d’un seuil très faible (7.4).

Les variables sous étude étant des **paramètres météorologiques,** nous pourrions supposer qu’elles soient **fonction de la date et du lieu** d’enregistrement des données ; une décomposition sur ces axes pourrait donc se révéler instructive.

Figure 3 présente la distribution de Evaporation selon les stations météorologiques. Elle montre que **les valeurs extrêmes**—presque toutes supérieures—**varient effectivement en fonction des stations.** Plus important, elle permet aussi de constater que **certaines stations ne disposent d’aucune entrée** pour Evaporation dans le jeu de données.

Ces informations sont importantes à prendre en compte dans la gestion des valeurs manquantes, les choix d’algorithmes et de modèles d’apprentissage automatique et l’interprétation des résultats.

Concernant les valeurs manquantes en particulier, **deux stratégies** se dessinent :

1. L’imputation d’un critère statistique de position (moyenne / médiane / mode), ou
2. La suppression.

Nous tenterons d’explorer ces deux pistes en parallèle dans la suite du projet.

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

Description générée automatiquement

Figure 3. La distribution de Evaporation en fonction des stations météorologiques.   
Les noms des stations ne disposant d’aucune donnée pour cette grandeur sont surlignés.   
Afin d’améliorer la lisibilité du graphique, l’abscisse a été arrêtée à 80,   
excluant 4 valeurs extrêmes supérieures.

### Variables catégorielles

Les variables catégorielles sont les suivantes :

* Date
* Location
* WindGustDir
* WindDir9am
* WindDir3pm
* RainToday
* RainTomorrow

Il n’y a rien d’inattendu dans la distribution des 16 modalités (correspondant aux 16 points cardinaux) des 3 variables liées au vent (WindGustDir, WindDir9am, WindDir3pm).

En revanche, les 2 variables liées à la **pluie** (RainToday, RainTomorrow) présentent un **fort déséquilibre,** comme illustré dans Figure 4**.**

|  |  |
| --- | --- |
| Une image contenant texte, capture d’écran, nombre, diagramme  Description générée automatiquement | Une image contenant texte, capture d’écran, nombre, diagramme  Description générée automatiquement |

Figure 4. Distributions de RainToday et de RainTomorrow.

Comme RainToday est la variable cible principale de l’étude, il sera nécessaire de mobiliser des méthodes conçues pour le traitement des échantillons contenant une forte disparité entre les classes à prédire (par exemple : rééchantillonnage, techniques de classification avancée).

En outre, ces deux variables catégorielles sont intimement liées à la variable numérique Rainfall (dont la distribution reste à clarifier) : les 3 sont toujours présentes ou absentes ensemble.

## Corrélations

### Variables numériques

Figure 5 présente une carte de corrélation des variables numériques avec RainToday et RainTomorrow établie en première approche, avec :

* La binarisation de ces deux variables catégorielles sous les noms RainTodayNum et RainTomorrowNum ;
* La simple suppression des valeurs manquantes ; et
* L’utilisation du coefficient de corrélation de Pearson comme métrique.

La ligne et la colonne qui correspondent à RainTomorrowNum, laquelle représente la **variable cible** RainTomorrow, ne montre pas de corrélation particulièrement importante : la plus forte est celle avec Sunshine (–0.45), suggérant la nécessité d’apporter un soin particulier à la stratégie de gestion des valeurs manquantes de cette dernière.

En considérant uniquement les variables **explicatives,** nous pouvons formuler les observations suivantes :

1. Les variables liées à la température (MinTemp, MaxTemp, Temp9am, Temp3pm) semblent être relativement corrélées entre elles ().
2. L’ensoleillement (Sunshine) et la couverture nuageuse (Cloud9am, Cloud3pm) semblent être négativement corrélés ().
3. L’ensoleillement (Sunshine) et le taux d’humidité à 15 h (Humidity3pm) semblent également être négativement corrélés ().

Ces observations pourraient informer la stratégie de simplification du jeu de données à travers la gestion des valeurs manquantes (en justifiant la suppression de certaines variables si elles peuvent être « représentées » par d’autres variables avec lesquelles elles sont fortement corrélées), la réduction de dimensions et d’autres méthodes.

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

Description générée automatiquement

Figure 5. Carte de corrélation des variables numériques avec RainToday et de RainTomorrow.

### Variables catégorielles

**Table 2** présente les résultats du test de d’indépendance visant à déterminer si un lien statistique existe entre chacune des variables catégorielles explicatives et la variable cible RainTomorrow, elle aussi catégorielle.

Les hypothèses formulées sont les suivantes :

Le seuil de la valeur-p fixé pour le rejet de l’hypothèse nulle est de 0.05.

Comme les valeurs-p sont toutes en-dessous de ce seuil, nous rejetons l’hypothèse et concluons à l’hypothèse  : les variables explicatives ne sont pas indépendantes de RainTomorrow.

**Table 2.** Résultats du test de d’indépendance entre   
les variables catégorielles explicatives et RainTomorrow.

|  |  |  |
| --- | --- | --- |
| **Variable** | **Statistique du test** | **Valeur-p** |
| Date | 16735 | 0.0 |
| Location | 3563 | 0.0 |
| WindGustDir | 1517 | 0.0 |
| WindDir9am | 2178 | 0.0 |
| WindDir3pm | 1283 | 0.0 |
| RainToday | 13799 | 0.0 |