SAE 15 : Traiter des données

Y. Haddab
yassine.haddab@umontpellier.fr

S. Druon
sebastien.druon@umontpellier.fr

Intervenants

- Sébastien DRUON
- Yassine HADDAB

Caractéristiques

- Formation encadrée : 10h, dont 8h de TP
- Projet : 20h
- Objectifs et problématiques professionnelles : Le professionnel R&T est régulièrement amené à traiter des données provenant du système d'information de l'entreprise pour ses besoins personnels ou ceux de ses collaborateurs. Ces données peuvent par exemple être liées à l'infrastructure de son réseau (état des équipements, des machines) ou aux utilisateurs. Généralement obtenues sous forme brute, elles sont ensuite traitées avec des objectifs très variés (nettoyage des données, extraction d'informations comptables, archivage, …) pour être réutilisées à d'autres fins ou être présentées dans des vues synthétiques. Ces traitements peuvent être récurrents (mensualisation de bilan, sauvegarde de données périodique, …) et gagnent à être automatisés. Le professionnel R&T doit donc développer des scripts ou des programmes pour gérer de façon efficace le traitement de ces données.
- Compétences ciblées : apprentissages critiques (RT2-Connecter)
  - AC0311 Utiliser un système informatique et ses outils
  - AC0312 Lire, exécuter, corriger et modifier un programme
  - AC0313 Traduire un algorithme, dans un langage et pour un environnement donné
  - AC0314 Connaître l'architecture et les technologies d'un site Web
  - AC0315 Choisir les mécanismes de gestion de données adaptés au développement de l'outil
  - AC0316 S'intégrer dans un environnement propice au développement et au travail collaboratif

Caractéristiques

- Ressources mobilisées et combinées :
  - R107 Fondamentaux de la programmation
  - R108 Bases des systèmes d’exploitation
  - R109 Introduction aux technologies Web
  - R110 Anglais de communication et initiation au vocabulaire technique
  - R111 Expression-Culture-Communication Professionnelles 1
  - R115 Gestion de projet

Caractéristiques

- Type de livrable ou de production :
  - Codes informatiques développés ;
  - Démonstration technique commentée ;
  - et/ou Rapport technique avec tutoriel d'installation ;
  - et/ou Soutenance orale présentant le travail réalisé.

  L'étudiant s'approprie son portfolio. Des temps sont prévus pour qu'il y synthétise sa production technique et son analyse argumentée.

- Mots-clés : Algorithmique, Programmation, Script.

Caractéristiques

- L’objectif est de valider des « compétences »
- Lien avec le Portfolio : il vous appartient d’alimenter votre portfolio tout au long du déroulement de cette SAE.

Caractéristiques

Dans le cadre de cette SAE, les objectifs visés sont :

- Collecter, traiter, présenter et publier des données.
- Réaliser un premier projet de développement informatique.
- S’initier aux différentes étapes d’un projet informatique.

Importance des données

- De nos jours, la **collecte**, le **traitement**, l'**analyse** des **données** ainsi que leur **présentation** et leur **publication** sont au cœur de l'économie numérique.

- Aujourd'hui, des millions de données circulent à travers internet. L'humanité en produit de plus en plus. La maîtrise des informations véhiculées par ces données revêt une importance stratégique et économique.

- Quelques exemples de données : données de santé des citoyens, données des réseaux sociaux, données de consommations, données de gestion des flux, données commerciales, etc.

La collecte des données

- Une première étape nécessaire : la mise en place de mécanismes de collecte des données d'intérêt.
- Selon la nature et la quantité des données à collecter, il conviendra de mettre en œuvre des mécanismes d'automatisation de la récupération des données (scripts, programmes, temporisations, etc.)

- Quelques exemples :
  - récupération de la température chaque heure dans une serre agricole,
  - récupération de l'état du trafic routier,
  - surveillance de l'état des patients dans un hôpital,
  - suivi du nombre de cas de personnes positives au Covid-19,
  - Etc.

Le traitement des données

- Une fois ces données récupérées, il convient de vérifier leur consistance et leur authenticité.
- Il conviendra également de les traiter (filtrage, élimination des données aberrantes, etc.

Extraction d'informations

- L'étape suivante consiste à extraire des informations pertinentes de ces données (caractéristiques : moyenne, moyenne glissante, écart type, tendances, évolution temporelle ou fréquentielle, élaboration de tableaux et/ou de graphes, etc.)

Présentation des données

- Mise en forme de ces données de sorte à en extraire des informations pertinentes et exploitables par un humain ou une machine.
- Passage de données brutes et en grand nombre (big data) à un ensemble d'informations de taille réduite et véhiculant des informations exploitables.

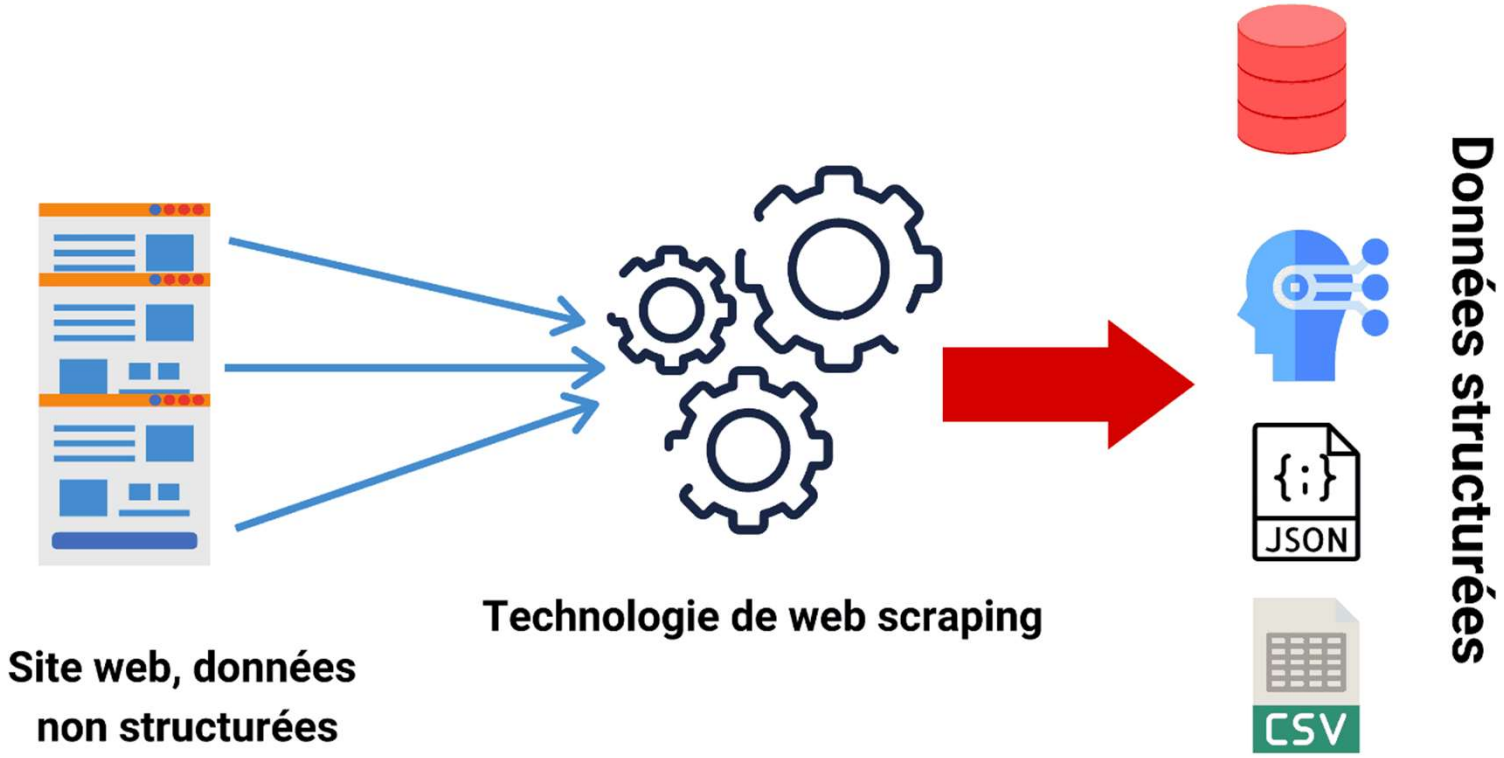
Publication des données

- Une fois les données mises en forme, dans certains cas, il conviendra de les publier (site web, dépôt logiciel, lieu de publication, etc.).

Un principe très utilisé sur internet : le web scraping

- Le « Web scaping » est une méthode automatique qui permet d'obtenir de grandes quantités de données à partir de sites web. La plupart de ces données sont des données non structurées au format HTML, XML, JSON… qui sont ensuite converties en données structurées dans un tableur ou une base de données afin de pouvoir être utilisées dans diverses applications.

Un principe très utilisé sur internet : le web scraping



Un principe très utilisé sur internet : le web scraping

- Exemples d'application
- **La surveillance des prix :** récupération des prix des concurrents en vue de l'établissement d'une stratégie commerciale.
- **Étude de marché :** analyse des tendances de consommation et choix des orientation de l'entreprise.
- **Apprentissage automatique et intelligence artificielle (IA) :** ces domaines ont besoin de très grandes quantités de données (big data).
- **Surveillance des actualités :** nécessaire pour plusieurs domaines économiques, stratégiques, politiques, etc.
- **Analyse des sentiments :** utile pour savoir comment les consommateurs accueillent un produit ou un service, etc.
- **Marketing par courriel :** collecte ciblée des e-mails, numéros de téléphones, identifiants sur les réseaux sociaux, etc.

Cas d’étude

Cas d’étude

- Données issues du site : https://data.montpellier3m.fr/

Ce sont des données mises à la disposition du public par Montpellier Méditerranée Métropole.

- Exemple de données mises à disposition :
  - Indices de qualité de l’air
  - Occupation des parkings de la ville
  - Disponibilité des places vélomag en temps réel
  - Etc.

Cas d'étude

- Nous utiliserons dans les travaux pratiques les données concernant l'occupation des parkings de la ville.
- Liste des parkings de la ville :

- Parking Antigone (FR_MTP_ANTI)
- Parking Corum (FR_MTP_CORU)
- Parking Foch (FR_MTP_FOCH)
- Parking Gare (FR_MTP_GARE)
- Parking Arc de Triomphe (FR_MTP_ARCT)
- Parking Circe (FR_MTP_CIRC)
- Parking Garcia Lorca (FR_MTP_GARC)
- Parking Mosson (FR_MTP_MOSS)
- Parking Euromédecine (FR_MTP_MEDC)
- Parking Vicarello (FR_CAS_VICA)
- Parking Gaumont OUEST (FR_MTP_GA250)
- Parking des Arceaux (FR_MTP_ARCE)
- Parking Comédie (FR_MTP_COME)
- Parking Europa (FR_MTP_EURO)
- Parking Gambetta (FR_MTP_GAMB)
- Parking du Triangle (FR_MTP_TRIA)
- Parking Pitot (FR_MTP_PITO)
- Parking Sabines (FR_MTP_SABI)
- Parking Sablassou (FR_CAS_SABL)
- Parking Saint Jean Le Sec (FR_STJ_SJLC)
- Parking Occitanie (FR_MTP_OCCI)
- Parking Gaumont EST (FR_MTP_GA109)
- Parking Charles de Gaulle (FR_CAS_CDGA)
- Parking Polygone (FR_MTP_POLY)

Cas d'étude

- Nous utiliserons dans les travaux pratiques les données concernant l'occupation des parkings de la ville.
- Pour chaque parking, un fichier XML donne les informations suivantes en temps réel (exemple pour le parking Comédie) :

```
<park>
<DateTime>2022-01-10T09:32:49</DateTime>
<Name>COME</Name>
<Status>Open</Status>
<Free>0493</Free>
<Total>0800</Total>
</park>
```

DateTime = date de la dernière mise à jour,
Name = nom du parking,
Status = statut d'ouverture,
Free = nombre de place libre,
Total = nombre total de place.

Cas d'étude

- Outils logiciels : nous utiliserons le langage Python (que vous avez étudié dans la ressource R107)

- Librairies à étudier (et à utiliser !) :
  - Librairie « **requests** » qui permet d'envoyer des requêtes HTTP
  - Librairie « **lxml** » qui permet de manipuler les données d'un fichier XML
  - Librairie « **time** » qui permet de gérer le temps

Eléments de base d’analyse des données

Analyse des données

- Les statistiques permettent d'obtenir des informations utiles à partir d'un jeu de données. Il est important de **maîtriser** les outils de base et de savoir les **programmer**.

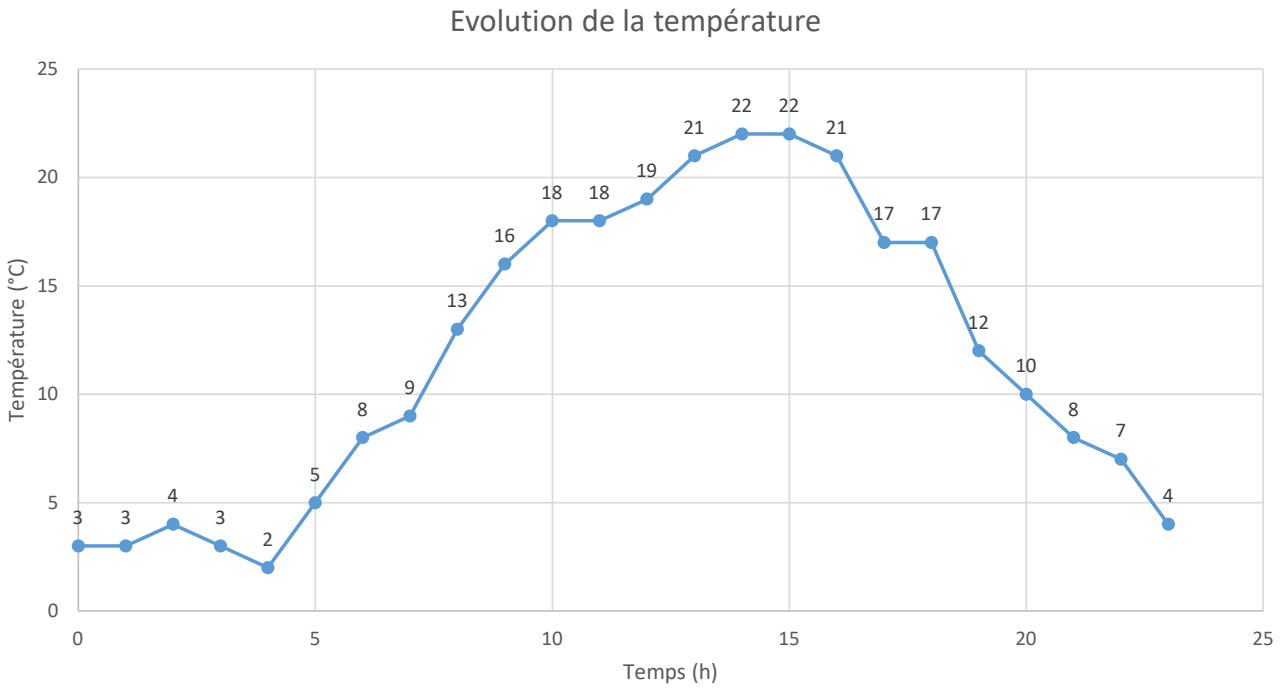
- Pour illustrer ces outils, nous considèrerons, par exemple, la température mesurée dans une salle durant 24 h (un jour entier) à raison d'une mesure par heure. On aura ainsi 24 mesures de température.

Analyse des données

- On aura ainsi 24 mesures de température. Ce signal n'est donc défini qu'à des instants t_i appartenant à un ensemble dénombrable $\{t_1, t_2, t_3, \ldots, t_n\}$.
- Dans notre exemple :
  - n=24
  - t_1 correspond à la température à 0h00
  - t_n correspond à la température à 23h00
- Nous noterons x_i les valeurs que prend la température (en °C) aux instants t_i. Cela conduit à l'ensemble de données suivant : $\{x_1, x_2, x_3, \ldots, x_n\}$.

Analyse des données

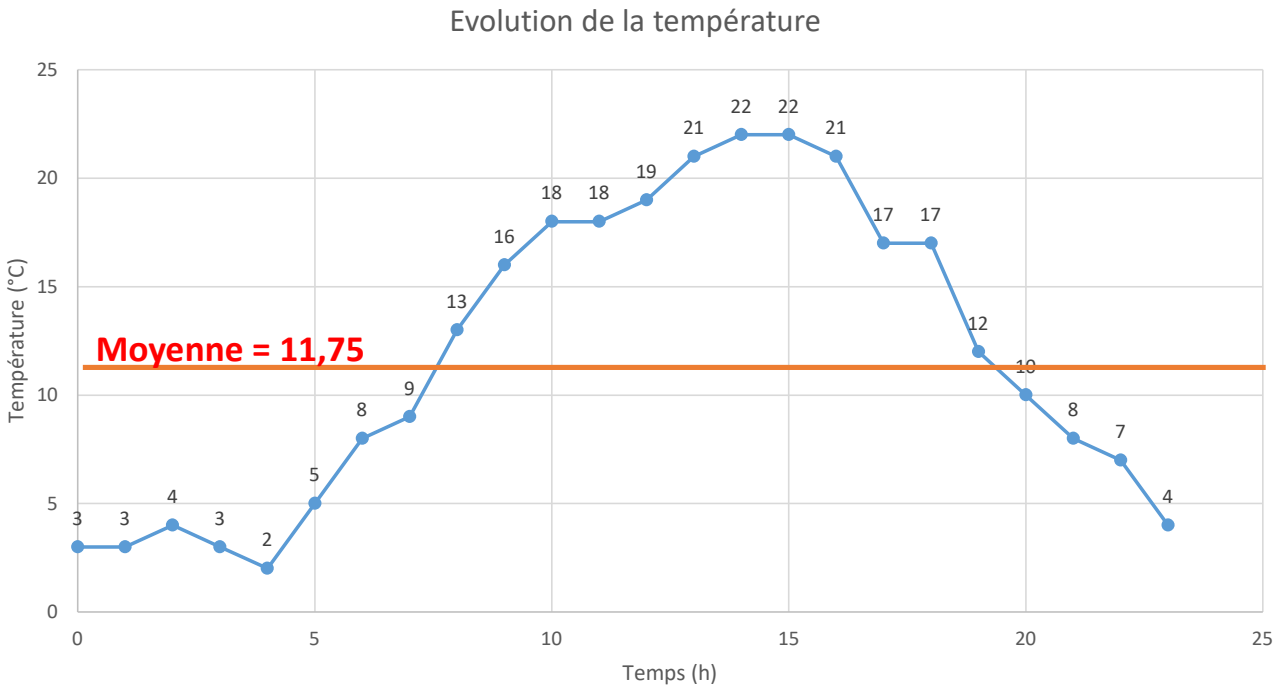
- Exemple de relevé des températures



Analyse des données

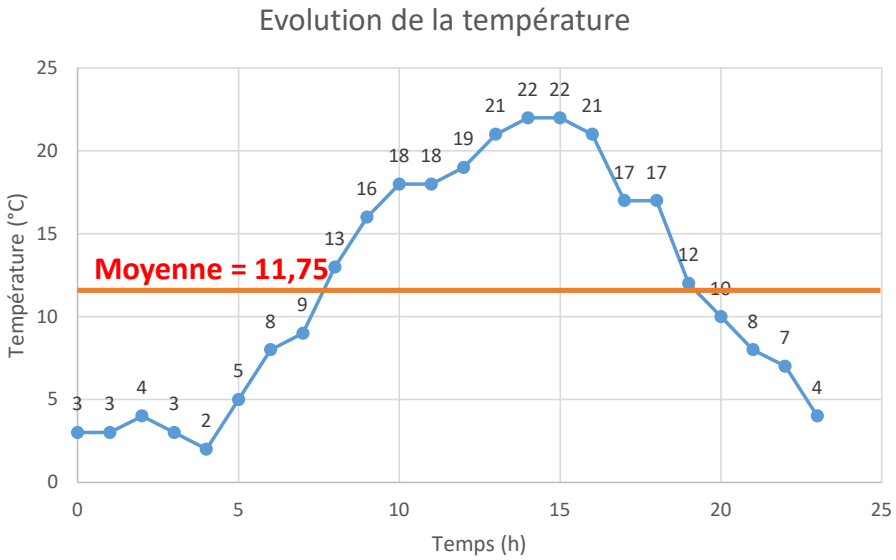
- Calcul de la moyenne

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

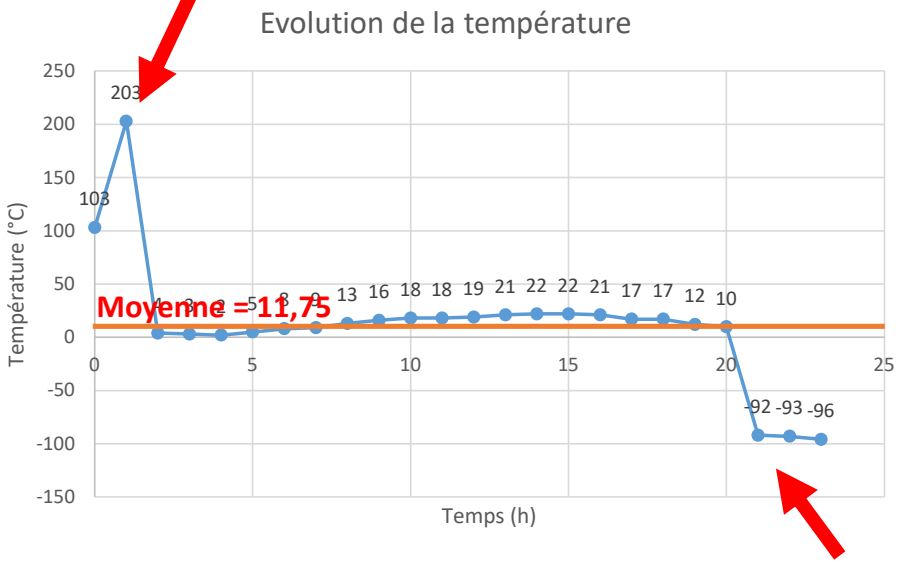


Analyse des données

- La moyenne est insuffisante pour caractériser des données. Ces deux signaux ont la même moyenne !



Un incendie ???



Tiens !!!

Pourtant, ces deux jeux de données ont la même moyenne !

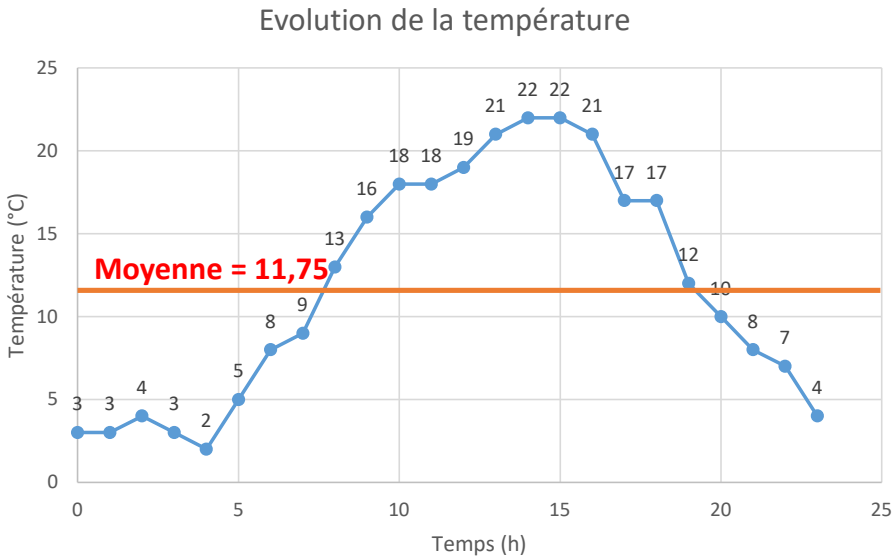
Analyse des données

- L'écart type : permet de mieux apprécier la dispersion des valeurs autour de la moyenne

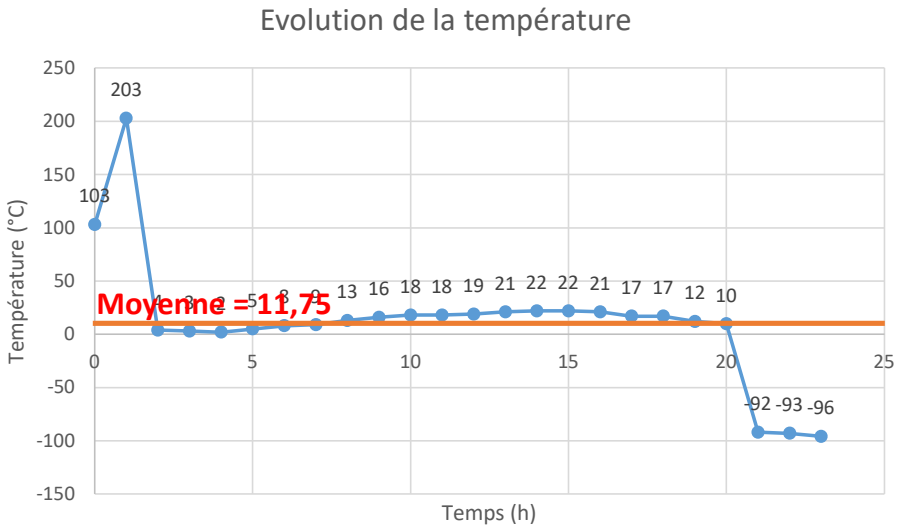
$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Analyse des données

- L'écart type



$\sigma = 7{,}04$



$\sigma = 58{,}66$

Permet de décrire la dispersion autour de la valeur moyenne.