ANALYSE STATISTIQUE & LANGAGE R

Statistiques descriptives

Présentation

Gaëtan DION

- **Euro Information – Groupe Crédit Mutuel – CIC**
- Elaboration et développement d'un SIG (système d'information géographique) groupe
- Développeur et Lead Technique (architecture) .NET / C# / SQL Server
- Développement de projets BI / Décisionnel
- Gestion de projets - analyses de données : Big Data (Spark, NoSQL…), couplé aux statistiques / Machine Learning

Planning

- 09/11/2020 : 6h
- 10/11/2020 : 2h
- 30/11/2020 : 2h
  - *1e évaluation théorie / pratique*
- 25/01/2020 : 2h
- 29/01/2020 : 2h
  - *2e évaluation théorie / pratique*

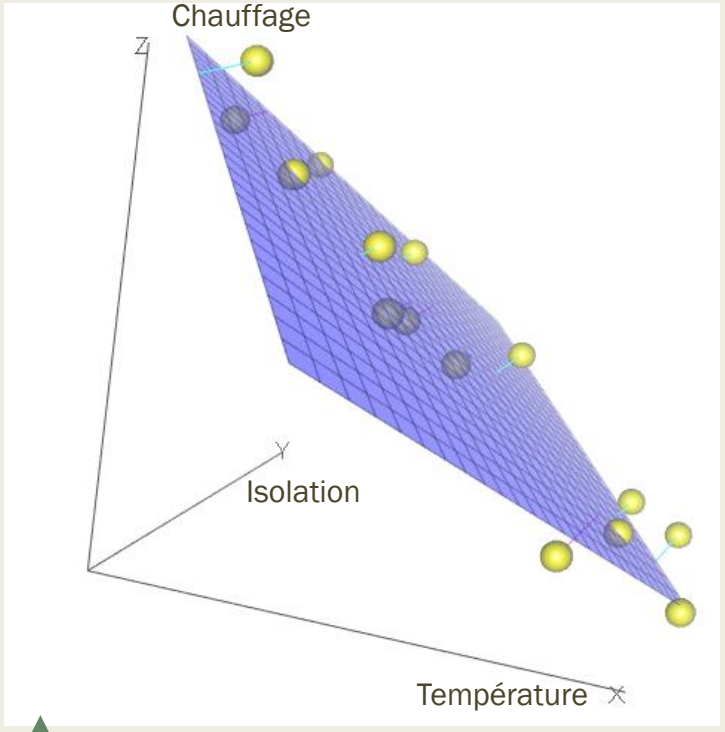
- **Séance 1** : statistiques descriptives
- **Séance 2** : tests d'hypothèses et régression linéaire
- **Séance 3** : modélisation (régression logistique), évaluation du modèle
  - *Analyses factorielles (PCA)*

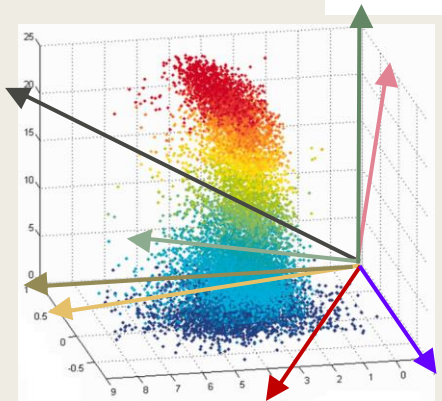
Déroulement
Séance 1

- Qu’est-ce que le Machine Learning?
- A quoi ça sert?
- Comment le situer parmi tous les types d’intelligence artificielle?
- Distinguer classification / régression, apprentissage supervisé / non supervisé
- La démarche d’un data scientist
- Les statistiques descriptives
  - *Les différents types de variables*
  - *Analyser un jeu de données*

Qu’est ce que le Machine Learning?

- Traitements informatiques basés sur des méthodes mathématiques, qui apprend sur un jeu de données (distributions, hypothèses…), afin de déterminer des règles automatiquement, en minimisant l’erreur
- Sortir une prédiction, classification d’individus (labellisé ou non labellisé)
- Processus non déterministe, chaque résultat possède une probabilité d’appartenance à cette classification
- 2 phases :
  - *Entrainement du modèle (modélisation)*
  - *Test / évaluation du modèle*

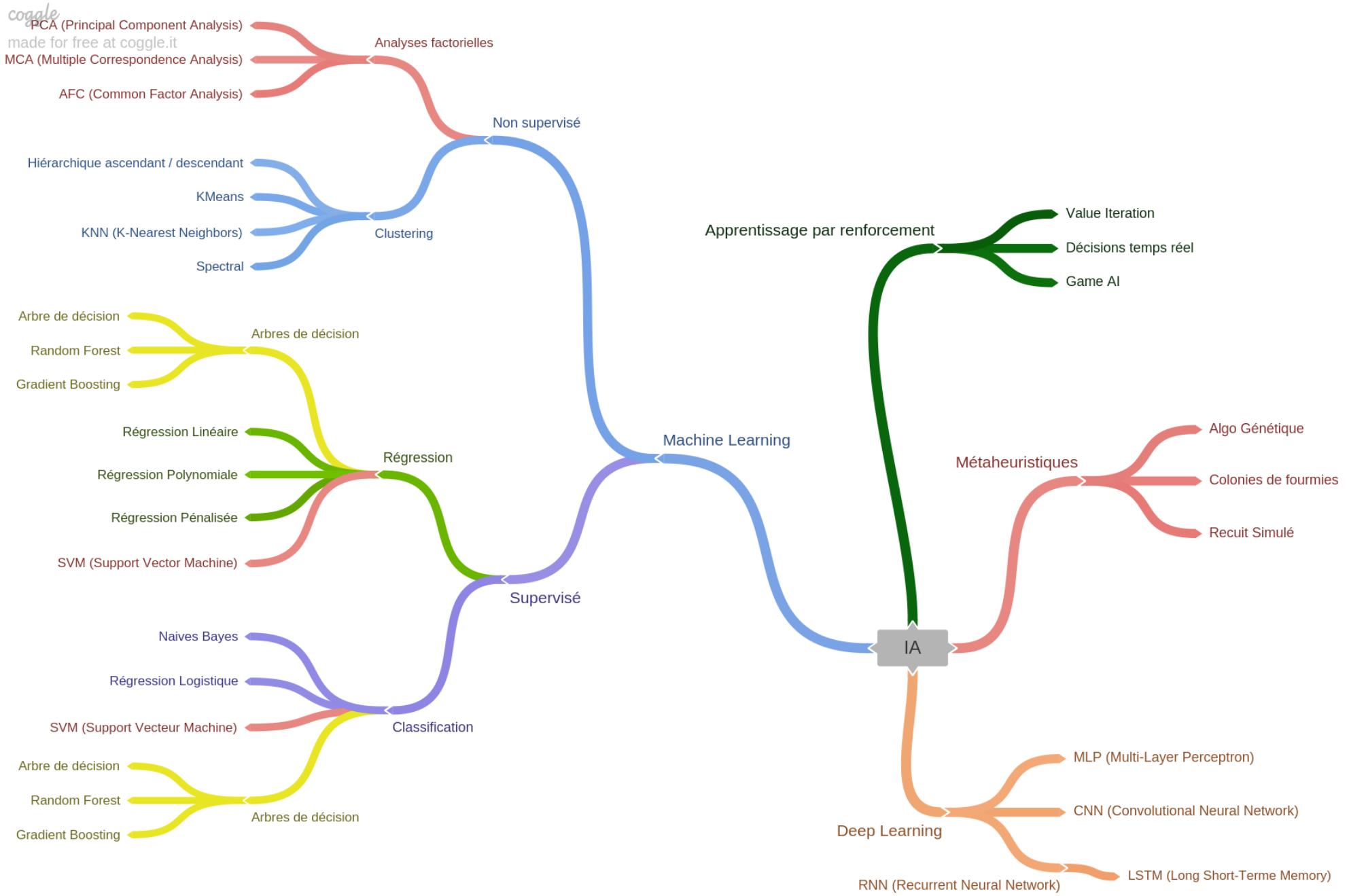




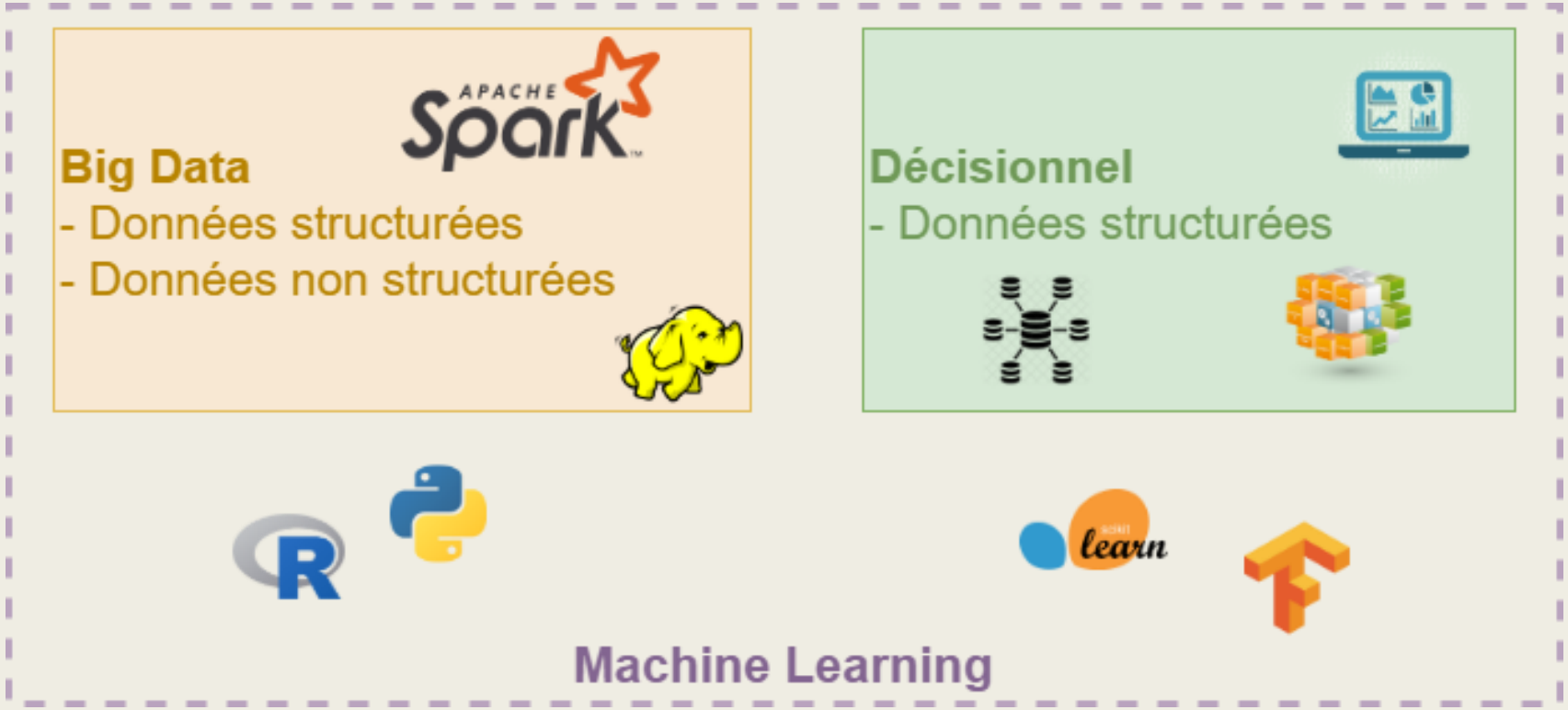
A quoi ça sert?

- Ces modèles permettent d'aborder des problèmes à espace multidimensionnel ou non linéaire, réalisation de multiples croisements, variations…
- Analyse de gros volumes de données, centaines de variables
- **Concrètement?**
  - *Détection de fraude (données labellisées)*
  - *Ciblage Marketing (appétence)*
  - *Score d'attrition (CHURN)*
  - *Bourse, Trading…*
  - *Analyse de sentiments, d'images, reconnaissance vocale (SVI), chatbot…*
  - *Actuariat, assurance…*

Une vision d'ensemble



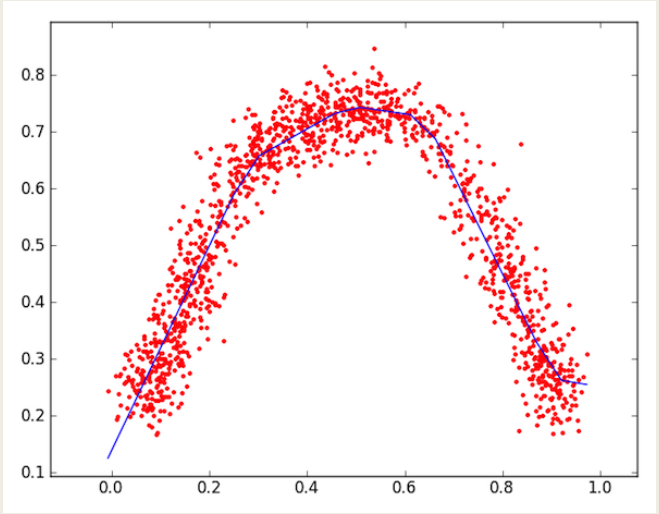
Le Big Data, l'IA, la BI



Classification / régression

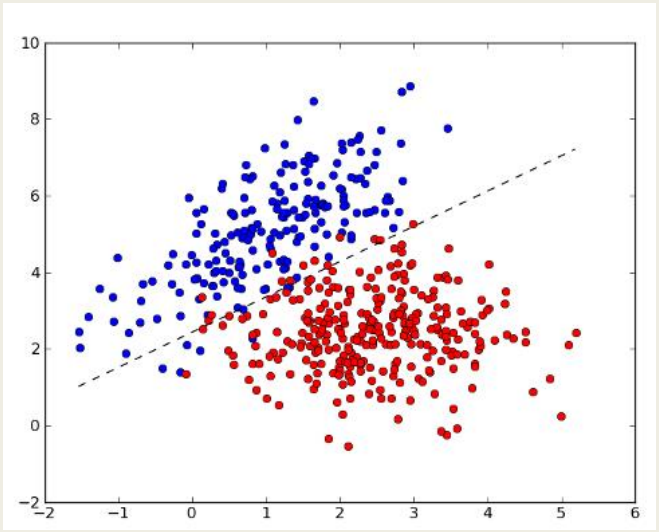
2 types de problèmes

- **Régression** : $Y \in R$, une infinité de valeurs continues réelles => taux de chômage, PIB, températures…



- **Classification** : Y = {0 ; 1}, un nombre fini de valeur.

- Une classification binaire : fraude 1/0, appétence 1/0, churn 1/0
- Multi Class : {0 ; 9} => mnist les nombres de 0 à 9
- Multi Label : Une observation peut appartenir à plusieurs classes (pas d'exclusivité) => 1 article pour plusieurs sujets

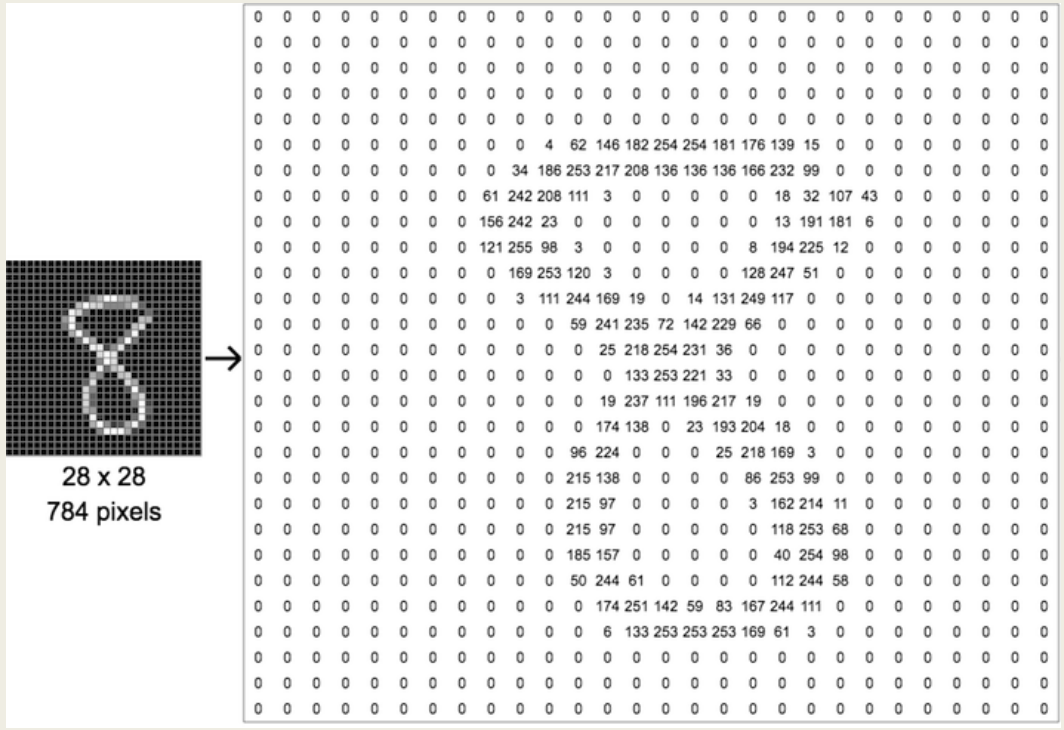


Supervisé / non supervisé

2 méthodes d'apprentissage

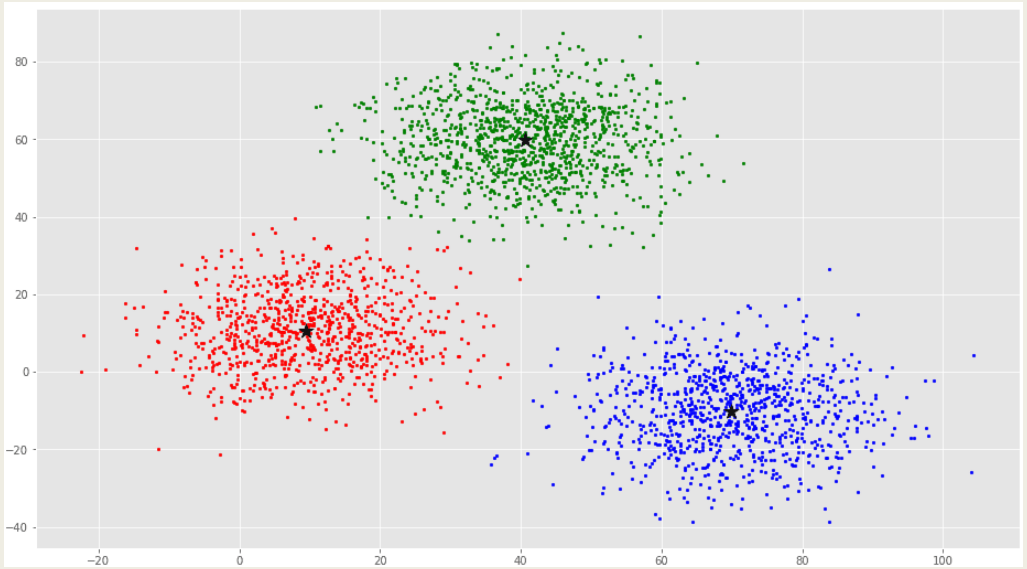
- **Supervisé** : ces algo extraient la connaissance d'un jeu de données de type entrée / sortie. On connait la variable à expliquer, et on se sert des variables explicatives pour expliquer la distribution des données.

781	pixel782	pixel783	indexedLabel
0.0	0.0	0.0	5.0
0.0	0.0	0.0	5.0
0.0	0.0	0.0	5.0
0.0	0.0	0.0	5.0
0.0	0.0	0.0	6.0
0.0	0.0	0.0	5.0
0.0	0.0	0.0	6.0
0.0	0.0	0.0	5.0
0.0	0.0	0.0	2.0
0.0	0.0	0.0	2.0
0.0	0.0	0.0	5.0



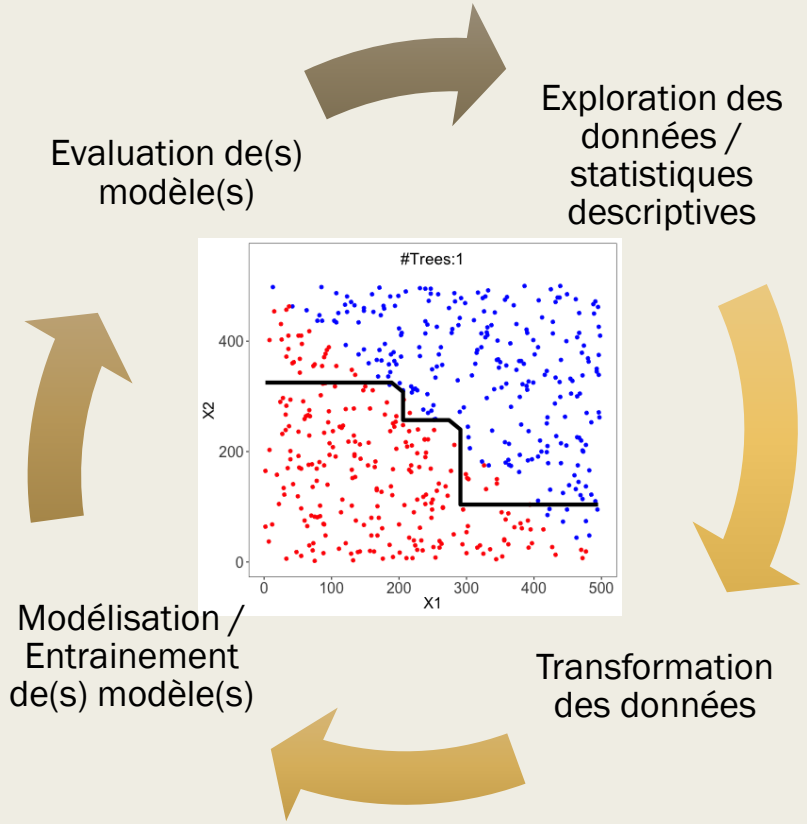
Non supervisé

- Ici pas d'entrée / sortie, toutes les données sont équivalentes, on cherche à organiser les données en groupe. Les groupes comportent des données similaires, et les données différentes se trouvent dans des groupes distincts.



La démarche d'un data scientist

1. Compréhension / exploration du jeu de données
   a. *Statistiques descriptives*
      a. Valeurs uniques / manquantes par variable
      b. Univariées (moyenne, médiane, quartiles, variance, écart type, distribution…)
      c. Bivariées (nuage de points, boites à moustaches…)
      d. Multivariées (matrice de corrélations…)
   b. *Nettoyage des données (remplacement de valeurs manquantes, uniformisation…)*
   c. *Création de nouvelles variables (agrégats…)*
   d. *Création d'un ou plusieurs modèle statistique*
   e. *Evaluation d'un ou plusieurs modèle*

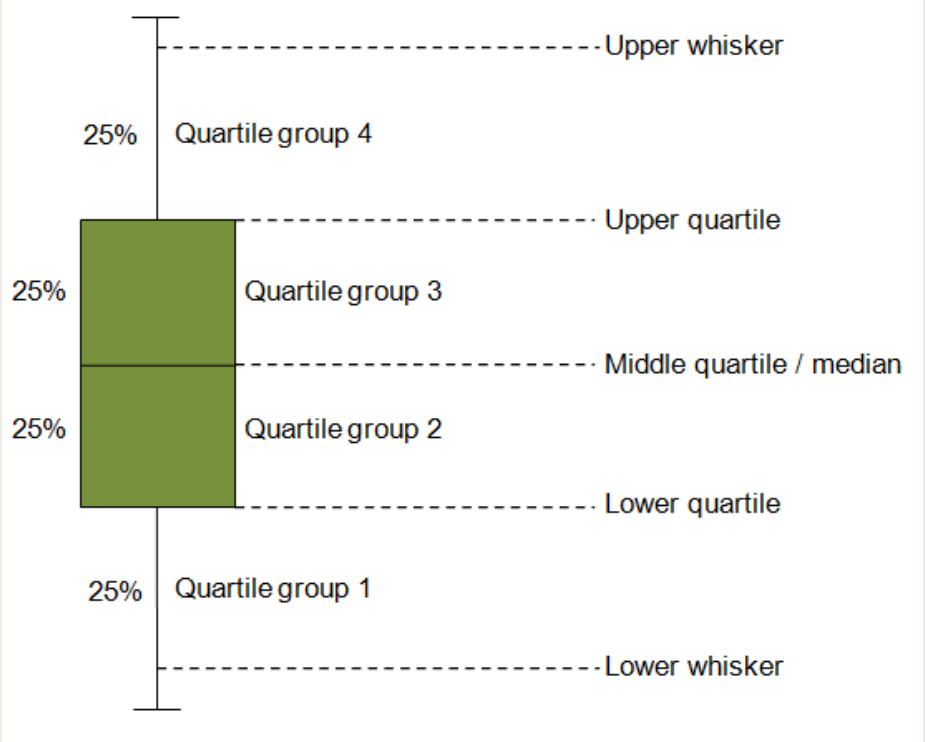


Les types de données

- Quantitative / qualitative
  - *Quantitative : données numériques*
  - *Qualitative : données alphanumériques*
- Quantitative discrète / quantitative continue
  - *Variable qui peut prendre un nombre fini de valeurs*
    - Mr/Mme, secteur d'activité, Marié/célibataire…
  - *Variable qui peut prendre un grand nombre de valeurs dans un intervalle réel donné*
    - âge, poids, taille, revenu…
- Nominale / Ordinale
  - *L'ordre n'a pas d'importance (ville, profession, …)*
  - *Variables dont le classement à une importance : (S, M, L, XL) ou (âge, poids, taille, revenus)*

Quelques fondamentaux

- Variance ?
  - *Mesure la dispersion autour de la moyenne*
  - *Evolue entre [0 ; +∞], toujours positif*
  - *2, 4, 6 => $[(2 - 4)^2 + (4 - 4)^2 + (4 - 6)^2] / 3$*
- Ecart type ?
  - $\sqrt{variance}$
  - *Coefficient de variation : (σ / μ)*
    - si >= (μ/2) : forte dispersion
- Box plot / boite à moustaches ?



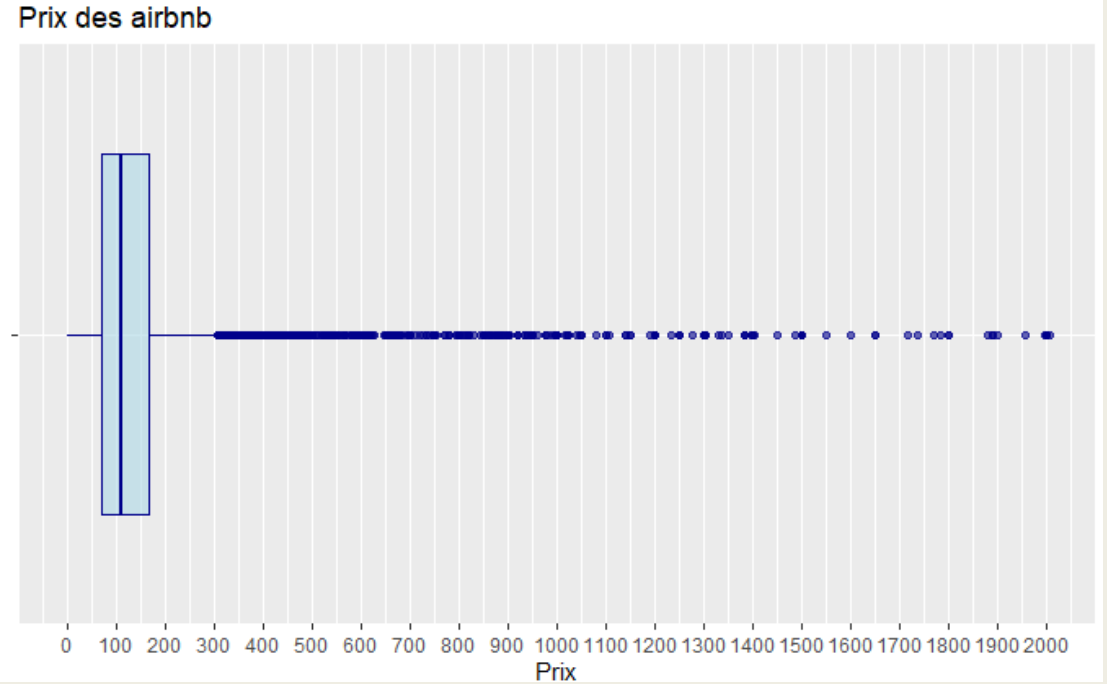
Prise en main de R Studio

- Se connecter à https://rstudio.cloud/
- Suivre le notebook : Les types de données R
  - *Comprendre et manipuler les types de données*
- Suivre le notebook : Mettre en forme des données
  - *Lecture d'un fichier*
  - *Manipulation / transformation de données*
  - *Ecriture d'un fichier résultat*

Exercice box plot

- Récupérer des données Airbnb de Melbourne sur le site Kaggle : https://www.kaggle.com/tylerx/melbourne-airbnb-open-data#listings_summary_dec18.csv
- Data" > "Data Sources" > Sélectionnez "listings_summary_dec18.csv" > en-dessous cliquez sur l'icone download

- **Analyse univariée**
  - *Afficher moyenne, médiane, variance, écart type, coef. de variation des prix*
  - *Afficher une boite à moustache / box plot sur la variable prix*
  - *Fonction **geom_boxplot**()*
  - *Afficher un graphique lisible (échelle, pas…)*

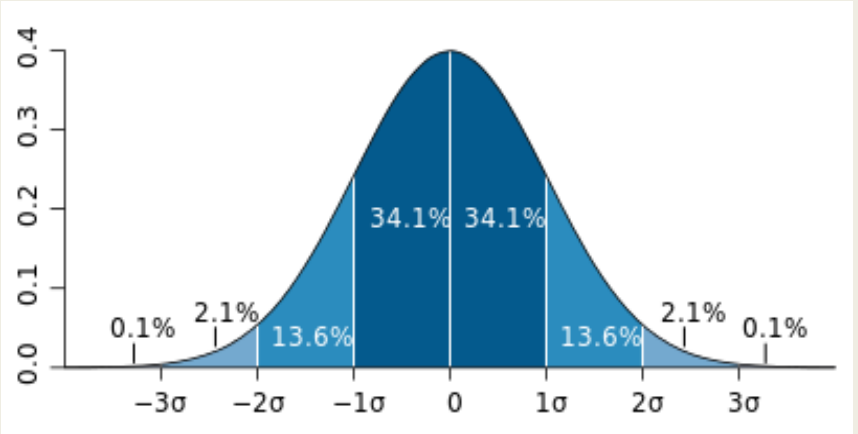


Distribution, loi normale ?

Permet d'appréhender la distribution de la série

Dans une distribution normale, la tendance centrale : la moyenne, la médiane et le mode ont des valeurs identiques.

68% de la population est concentrée entre +/- 1 écart type



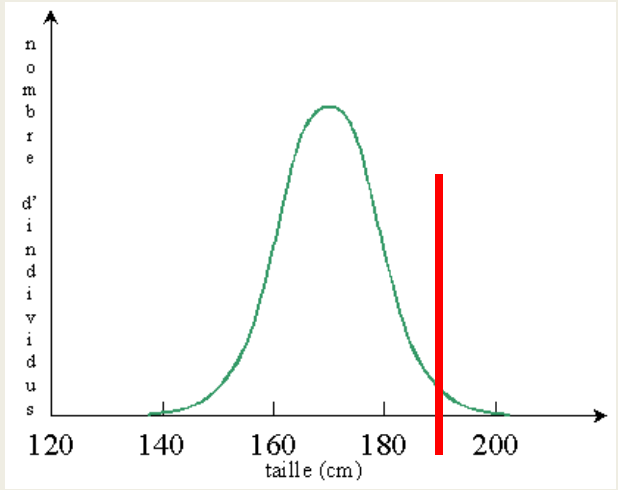
X suit une loi normale N(170, 10)

Loi normale centrée réduite : (x - μ) / σ => N(0, 1)

P(x>190)=2,2% (2 écart-type)

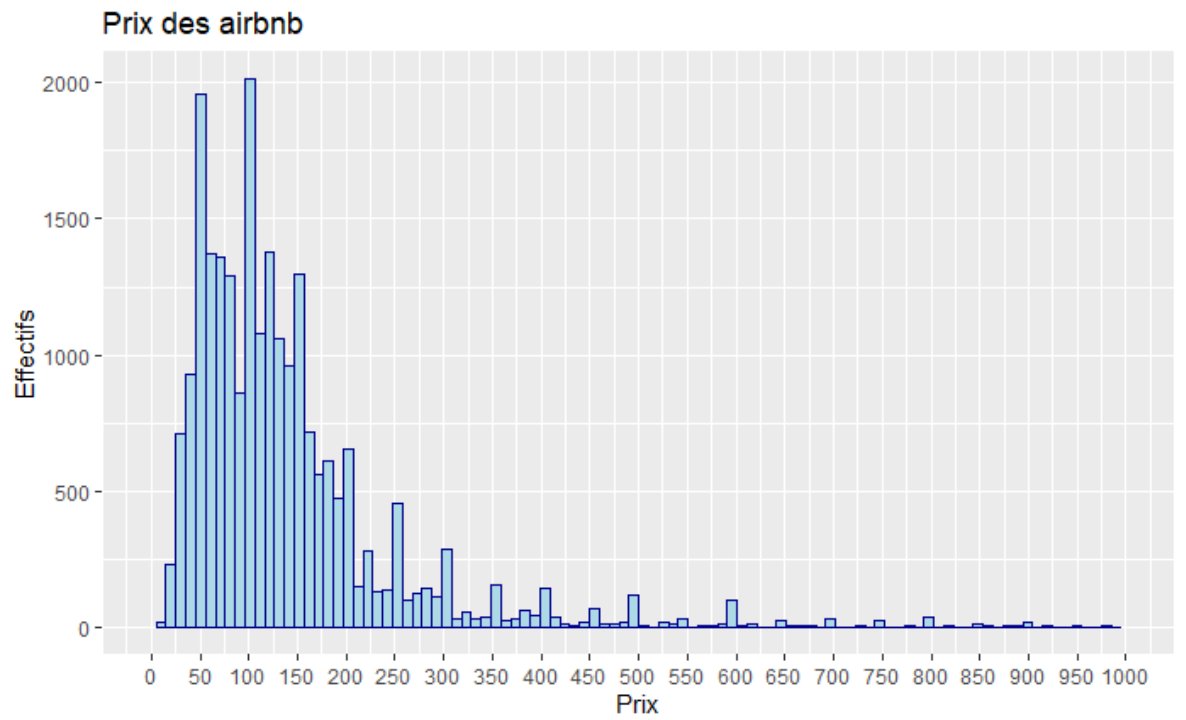
La loi normale centrée réduite permet de se référer à une table de probabilité

Elle permet d'uniformiser le poids des variables pour un modèle



Exercice histogramme

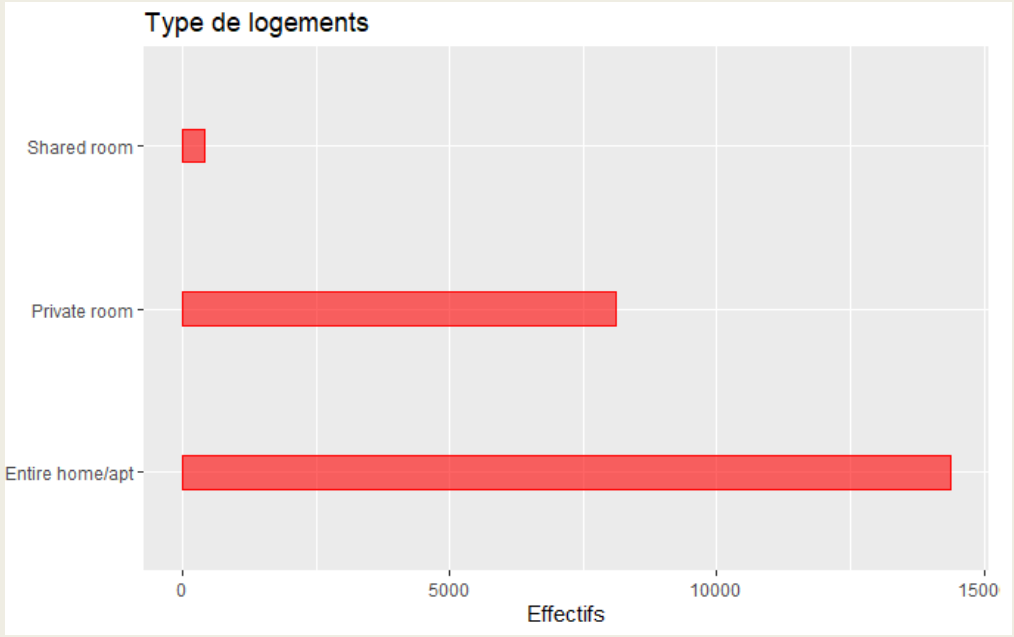
- **Analyse univariée**
  - *Afficher une distribution du prix des airbnb via un histogramme*
  - *Fonction : geom_histogram()*



Exercice diagramme en bâtons

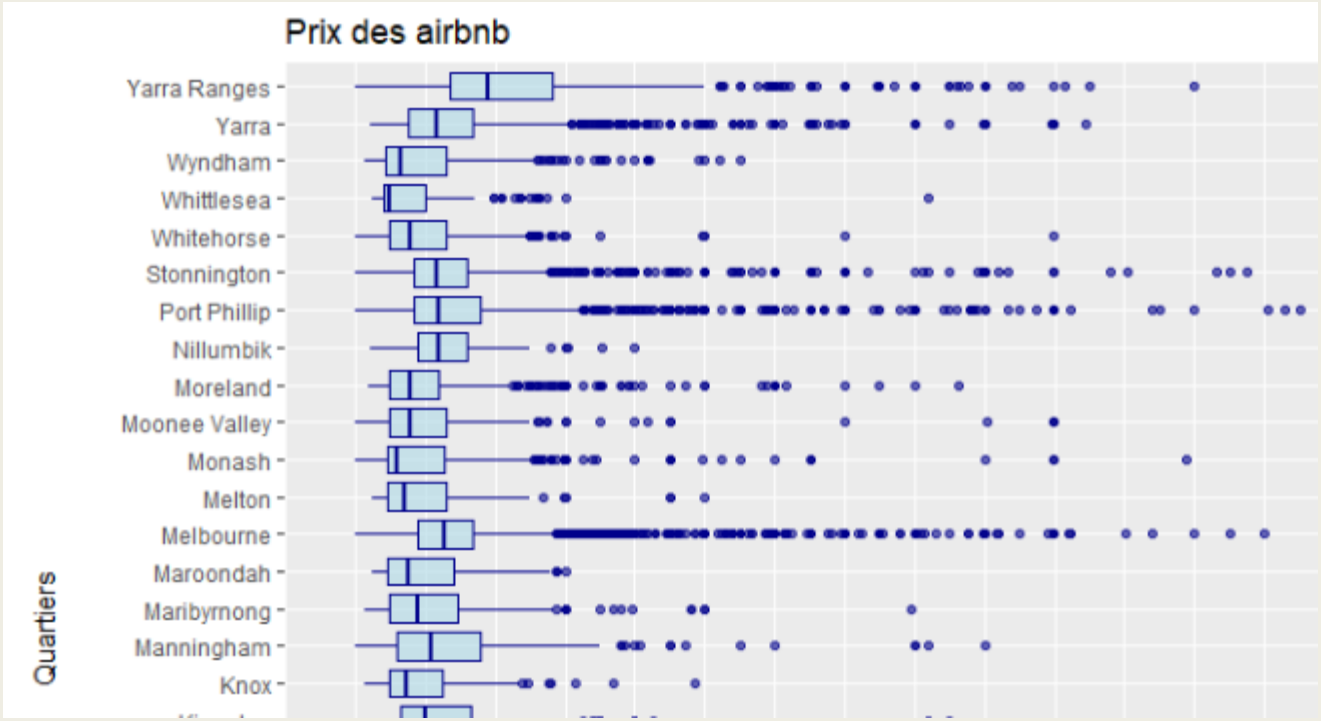
- **Analyse univariée**
  - *Diagramme de fréquence sur la variable qualitative « room_type » et « neighbourhood »*
  - *Fonction geom_bar()*
  - *Afficher le graphique à l'horizontal (90°)*

- Qu'en déduisez-vous?



Exercice bivariée

- **Analyse bivariée**
  - *Prix des logements en fonction des quartiers (variable neighbourhood)*

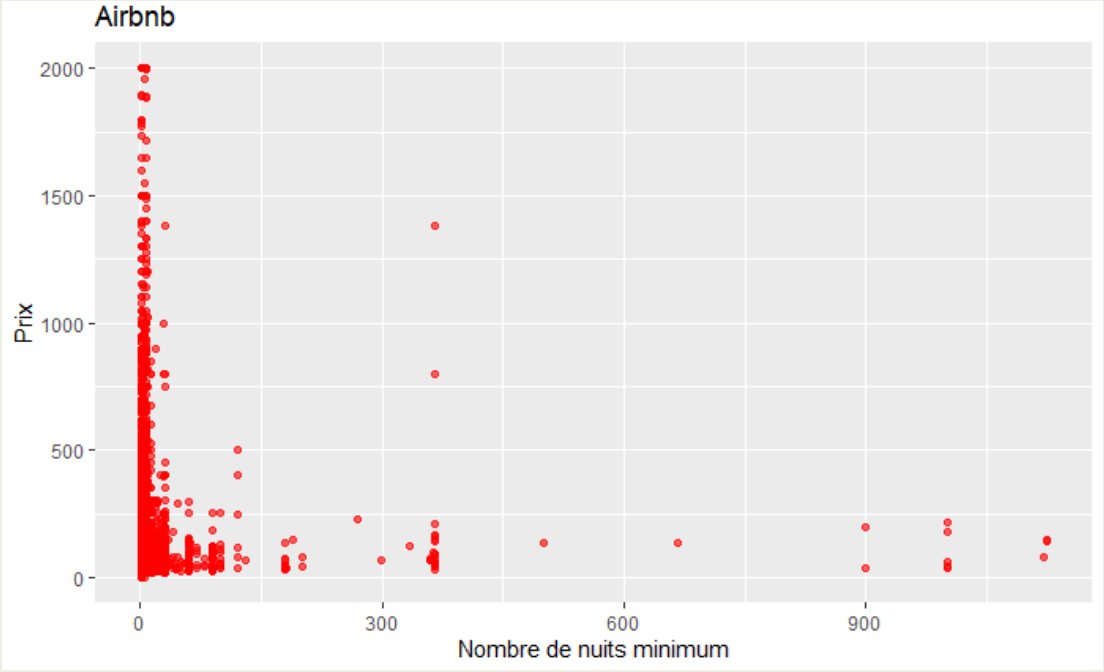


Exercice nuage de points

- **Analyse bivariée**
  - *Prix des logements (ordonnée) en fonction :*
    - du type de logement (room_type)
    - du quartier (neighbourhood)
    - du nombre de nuits minimum (minimum_nights)

Abscisse

  - *Que peut-on en déduire?*



Résumé

Utiliser la « distribution » tidyverse :

```
install.packages("tidyverse", repos='http://cran.r-project.org')
library(tidyverse)
```

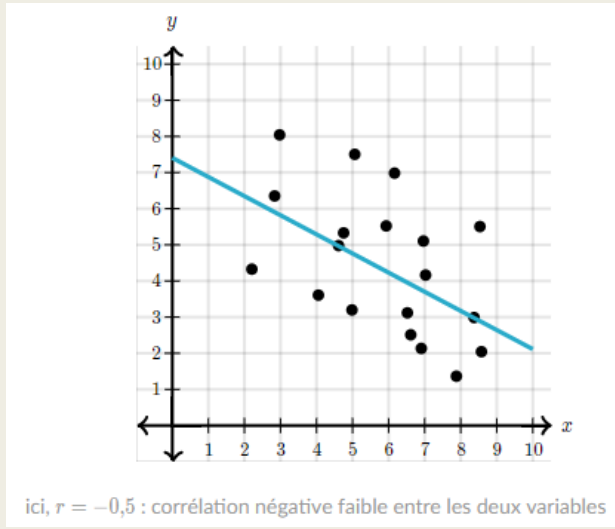
La librairie graphique est **ggplot2**

Fonctions R	Univariées / biavariées / Multivariées	Types de graphiques
Histogramme	Univarié	geom_histogram()
Courbe gaussian ou non	Univarié	geom_density()
Boite à moustache	Univarié et bivarié	geom_boxplot()
Diagramme en bâtons	Univarié	geom_bar()
Nuage de points	Bivarié	geom_point()

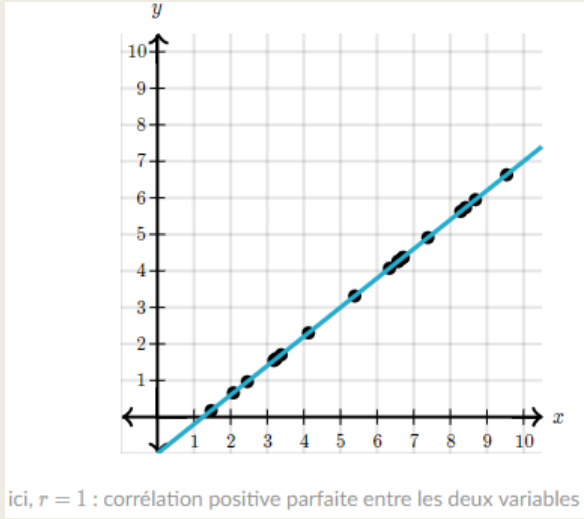
Les corrélations

- Etudier l'intensité de la liaison qui peut exister entre deux ou plusieurs variables
  - *Test de dépendance ou d'indépendance*
- Le coefficient de corrélation est compris entre [-1 ; 1]
  - *0 : absence de relation*
  - *1 : relation linéaire positive forte*
  - *-1 : relation linéaire négative forte*

$$r = \frac{\sum(xi - \bar{x})(yi - \bar{y})}{SxSy}$$



ici, $r = -0{,}5$: corrélation négative faible entre les deux variables



ici, $r = 1$: corrélation positive parfaite entre les deux variables

Exercice matrice de corrélation

- Analyse multivariée
  - *Matrice de corrélation*
  - *Données quantitatives*
- Fonction corrplot() / package corrplot
  - *Sélectionner seulement les variables numériques (plusieurs solutions)*
  - *Déterminer le coefficient de corrélation pour chaque couple de variables avec cor()*
  - *Afficher la matrice de corrélation à partir de ce résultat*