Chapitre 3:

# Analyse exploratoire des données

# Qu'est-ce que l'Analyse exploratoire des données (EDA) ?

- **Définition :** L'EDA (Exploratory Data Analysis) est une étape préliminaire essentielle qui consiste à :
  - Résumer les données,
  - Visualiser leurs caractéristiques,
  - Identifier les structures, motifs, ou anomalies.
- **Objectifs principaux**
  - Comprendre la distribution des variables.
  - Détecter les valeurs manquantes ou aberrantes.
  - Générer des hypothèses et orienter les analyses ultérieures.
- **Pourquoi est-ce important ?**
  - Évite les erreurs d'interprétation.
  - Aide à choisir les méthodes statistiques ou algorithmes adaptés.
  - Sert de base à la modélisation et à la prise de décision.

# Qu’est-ce que l’Analyse exploratoire des données (EDA) ?

- **Exemple concret**

Avant de prédire la réussite scolaire d’étudiants, on explore leurs données :

  - Distribution des notes,
  - Répartition par section / groupe de TD / groupe de TP,
  - Relation entre nombre d’heures d’étude et résultats.

# 3.1. Statistiques descriptives

# Statistiques descriptives

- Objectif :
  - Résumer et décrire les caractéristiques principales d'un jeu de données.
- Question clé :
  - *Comment donner une vision claire d'un ensemble de données sans le parcourir valeur par valeur ?*
- Deux familles de mesures :
  - **Tendance centrale** : "où se situe la majorité des valeurs ?"
  - **Dispersion** : "comment les valeurs sont-elles réparties autour de cette tendance ?"

# Mesures de tendance centrale

- **Moyenne (arithmétique)**
  - **Définition** : somme des valeurs ÷ nombre d'observations.
  - **Avantage** : facile à calculer, représentative si données homogènes.
  - **Inconvénient** : très sensible aux valeurs extrêmes (outliers).
  - Exemple :
    - Revenus de 5 personnes = {1600, 1200, 1500, 1300, 20000}
    - Moyenne = 5120 → non représentative !

# Mesures de tendance centrale

- **Médiane**
  - **Définition** : valeur centrale d'un jeu de données ordonné.
  - **Avantage** : robuste face aux valeurs extrêmes.
  - Exemple :
    - Revenus de 5 personnes = {1200, 1300, 1500, 1600, 20000}
    - Médiane = 1500 → représentative (mieux que la moyenne)

# Mesures de tendance centrale

- **Mode**
  - **Définition** : valeur la plus fréquente.
  - Utile pour les données qualitatives (ex. couleur la plus fréquente : “bleu”).
  - Exemple : {6, 3, 6, 7, 6, 3, 2, 4} → Mode = 6.

# Mesures de dispersion

- **Étendue**
  - Formule : Max – Min.
  - Indique l'amplitude totale, mais sensible aux valeurs extrêmes.
  - Exemple : {2, 5, 9, 12, 15} → Étendue = 15 – 2 = 13.

# Mesures de dispersion

- **Variance ($\sigma^2$)**
  - Mesure de la dispersion par rapport à la moyenne.
  - Plus la variance est grande, plus les données sont éparpillées.

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

# Mesures de dispersion

- **Écart-type (σ)**
  - Racine carrée de la variance.
  - Plus interprétable car exprimé dans la même unité que les données.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

  - Exemple: Données : {10, 12, 14, 16, 18}
    - Moyenne = 14
    - Variance = 8
    - Écart-type ≈ 2.83.

# Mesures de dispersion

- **Écart interquartile (IQR)**
  - Q3 – Q1 (75e percentile – 25e percentile).
  - Représente la dispersion centrale de 50 % des données.
  - Outil très utilisé pour identifier les outliers (méthode des “1.5×IQR”).

# 3.2. Visualisation des données

# Pourquoi visualiser les données ?

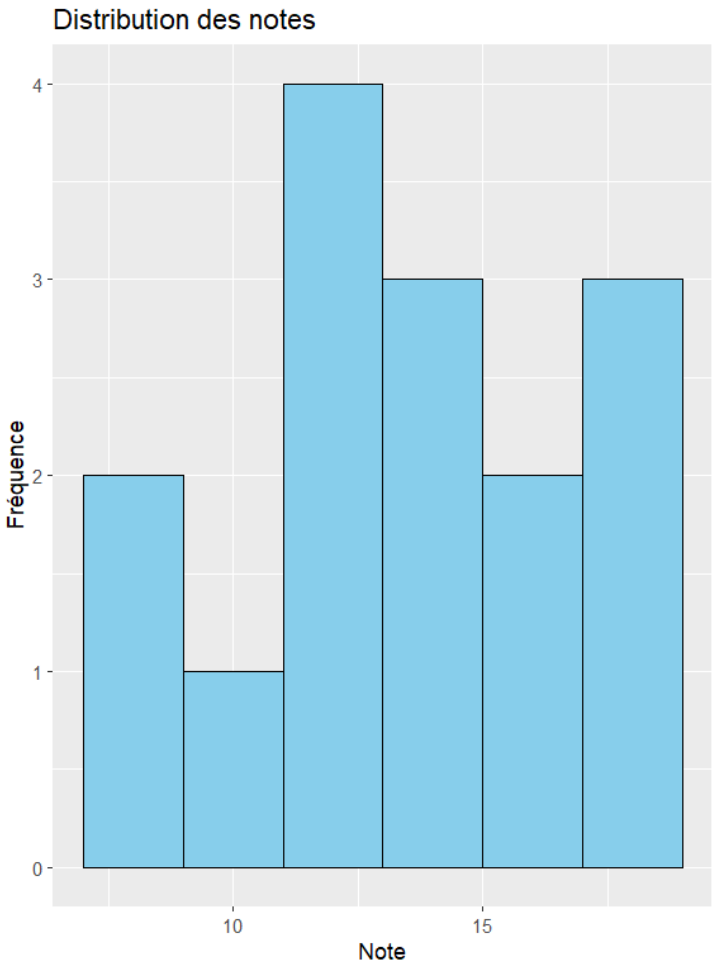
- La visualisation de données est un terme général qui décrit toute tentative visant à aider les utilisateurs à comprendre l'importance des données en les plaçant dans un contexte visuel.

- Un objet visuel peut communiquer davantage d'informations qu'un tableau dans un espace beaucoup plus restreint. Cette caractéristique des visuels les rend plus efficaces que les tableaux pour présenter les données.

- Représenter graphiquement les données pour en faciliter la compréhension.

# Histogramme

- **Définition** : représentation de la distribution d'une variable quantitative.
- **Axes** :
  - x = classes d'intervalles,
  - y = fréquences.
- **Utilité** : observer la forme de la distribution (normale, symétrique, inclinée à droite ou à gauche.
- **Exemple** : distribution des notes d'étudiants.



# Histogramme

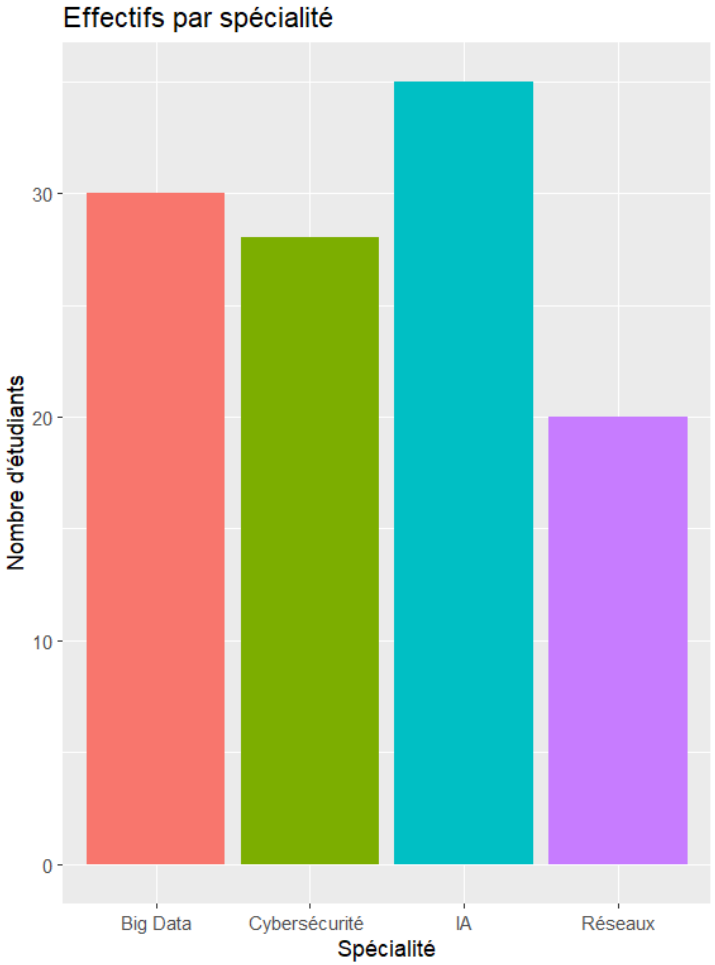
```
library(ggplot2)

notes <- data.frame(note = c(8, 9, 10, 12, 13, 14, 16, 18, 19, 12, 14, 15, 13, 17, 18))

ggplot(notes, aes(x = note)) +
  geom_histogram(binwidth = 2, fill = "skyblue", color = "black") +
  labs(title = "Distribution des notes", x = "Note", y = "Fréquence")
```

# Diagramme en barres

- **Définition** : hauteur des barres = fréquence ou valeur d'une catégorie.
- **Utilité** : comparer des variables **catégorielles** (ex. nombre d'étudiants par spécialité).
- **Attention** : largeur des barres doit être uniforme.



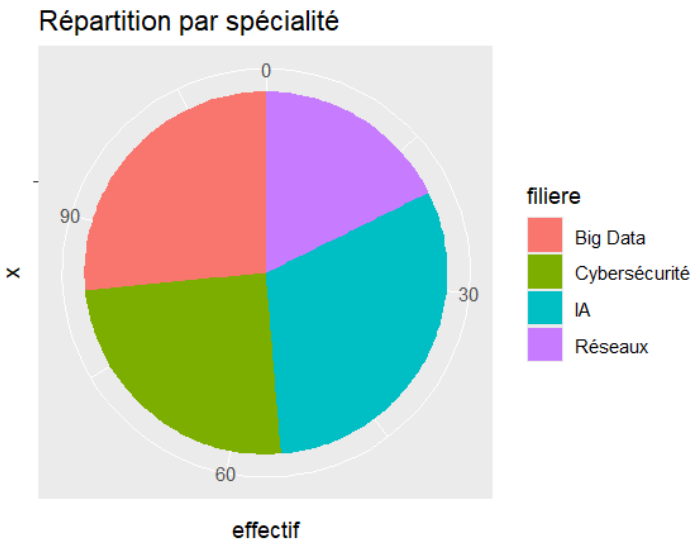
# Diagramme en barres

```
specialites <- data.frame(
  filiere = c("IA", "Big Data", "Réseaux",
"Cybersécurité "),
  effectif = c(35, 30, 20, 28)
)

ggplot(specialites, aes(x = filiere, y =
effectif, fill = filiere)) +
  geom_bar(stat = "identity") +
  labs(title = "Effectifs par spécialité", x
= "Spécialité", y = "Nombre d'étudiants") +
  theme(legend.position = "none")
```

# Diagramme circulaire (Pie chart)

- **Définition** : parts d'un cercle représentant des proportions.
- **Utilité** : montrer une répartition simple.
- **Limites** : difficile à lire avec trop de catégories → préférer un diagramme en barres.

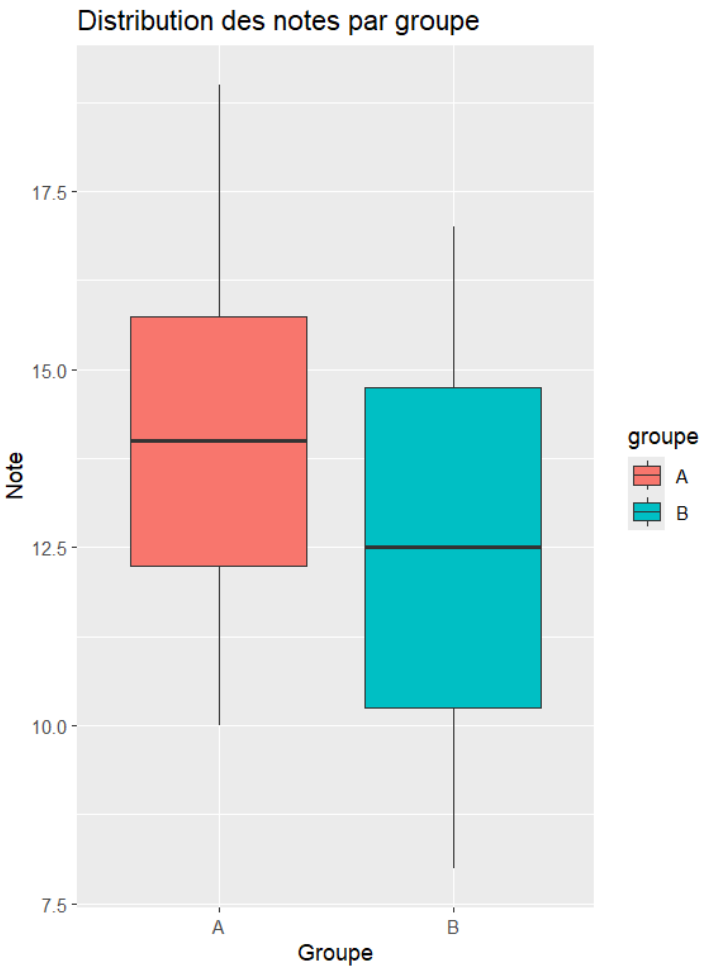


# Diagramme circulaire (Pie chart)

```
ggplot(specialites, aes(x = "", y = effectif, fill = filiere)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar("y") +
  labs(title = "Répartition par spécialité")
```

# Boîte à moustaches (Boxplot)

- **Définition** : représentation des quartiles et détection des outliers.
- **Éléments** :
  - médiane (trait dans la boîte),
  - Q1 et Q3 (boîte),
  - moustaches (valeurs normales),
  - points isolés (outliers).
- **Utilité** : comparer distributions de plusieurs groupes.



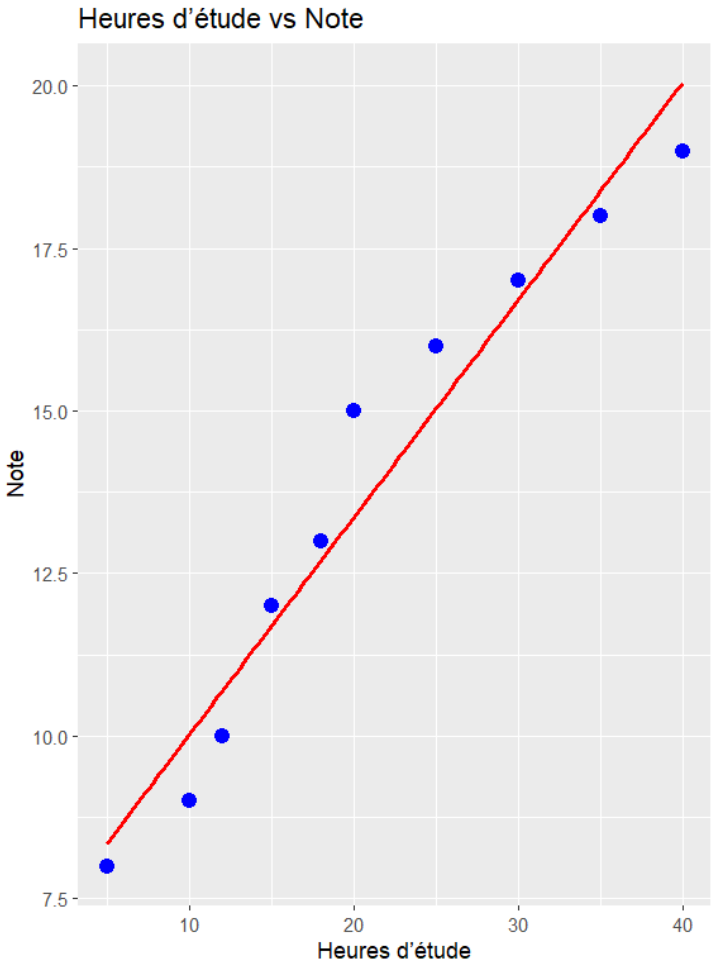
# Boîte à moustaches (Boxplot)

```
notes_groupes <- data.frame(
  groupe = rep(c("A", "B"), each = 10),
  note = c(10, 12, 14, 15, 16, 18, 19, 12, 13,
14, 8, 9, 10, 11, 13, 12, 14, 15, 16, 17)
)

ggplot(notes_groupes, aes(x = groupe, y =
note, fill = groupe)) +
  geom_boxplot() +
  labs(title = "Distribution des notes par
groupe", x = "Groupe", y = "Note")
```

# Nuage de points (Scatterplot)

- **Définition** : chaque point représente une paire (x, y).
- **Utilité** : analyser la relation entre deux variables.
- **Exemple** : nombre d'heures d'étude vs. note obtenue.
- **Attention** : sur-interprétation → corrélation ≠ causalité.



# Nuage de points (Scatterplot)

```
etudiants <- data.frame(
  heures = c(5, 10, 12, 15, 18, 20, 25, 30, 35, 40),
  note   = c(8, 9, 10, 12, 13, 15, 16, 17, 18, 19)
)

ggplot(etudiants, aes(x = heures, y = note)) +
  geom_point(color = "blue", size = 3) +
  geom_smooth(method = "lm", se = FALSE, color
= "red") +
  labs(title = "Heures d’étude vs Note", x =
"Heures d’étude", y = "Note")
```

# Graphique en ligne (Line Chart)

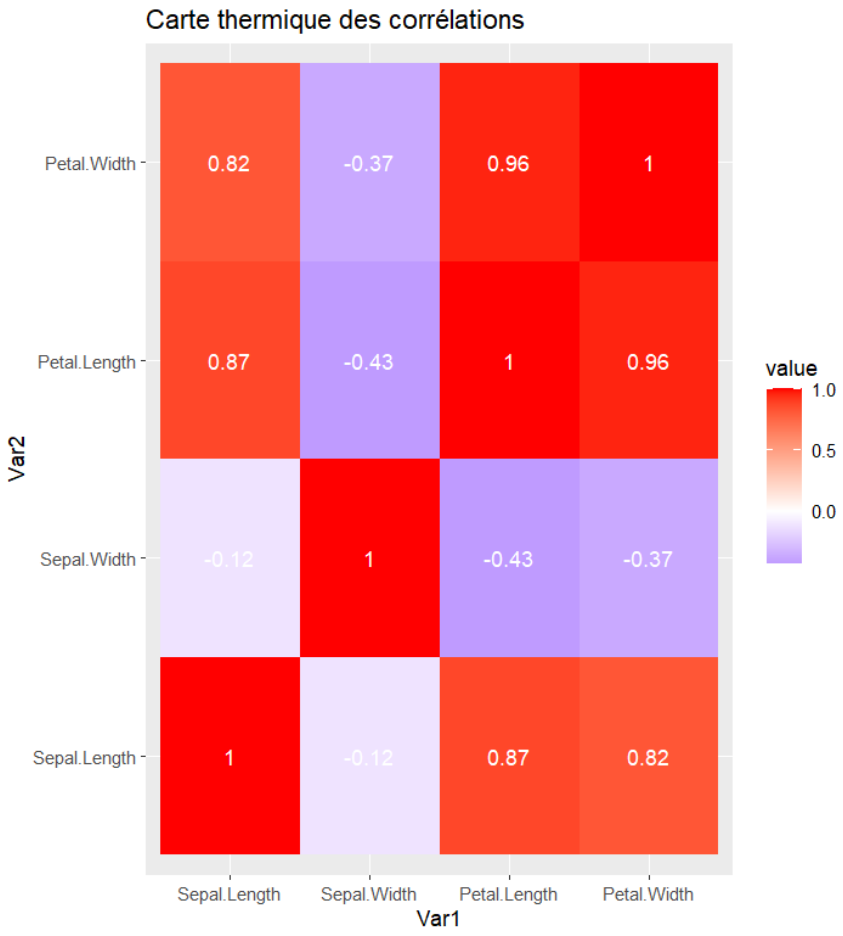
- Représentation de l'évolution d'une variable quantitative dans le temps.
- Chaque point représente une mesure à une date donnée, reliée aux autres points par une ligne.
- **Axes :**
  - **x** → temps (jours, mois, années…)
  - **y** → valeur mesurée
- **Utilité :**
  - Visualiser une **tendance** (hausse, baisse, saisonnalité)
  - Suivre une mesure dans le temps (performance, ventes…)
  - Détecter **points de rupture** ou **changements de comportement**
- **Exemple :** ventes mensuelles d'une entreprise.

# Graphique en ligne (Line Chart)

```
ggplot(economics, aes(date, unemploy)) +
  geom_line()
```

# Carte thermique (Heatmap)

- **Visualisation multivariée**
- **Carte thermique** : couleurs représentant l'intensité d'une valeur.
- **Utilité** : très utilisé pour les matrices de corrélation.
- **Exemple** : corrélation entre plusieurs variables économiques.



# Carte thermique (Heatmap)

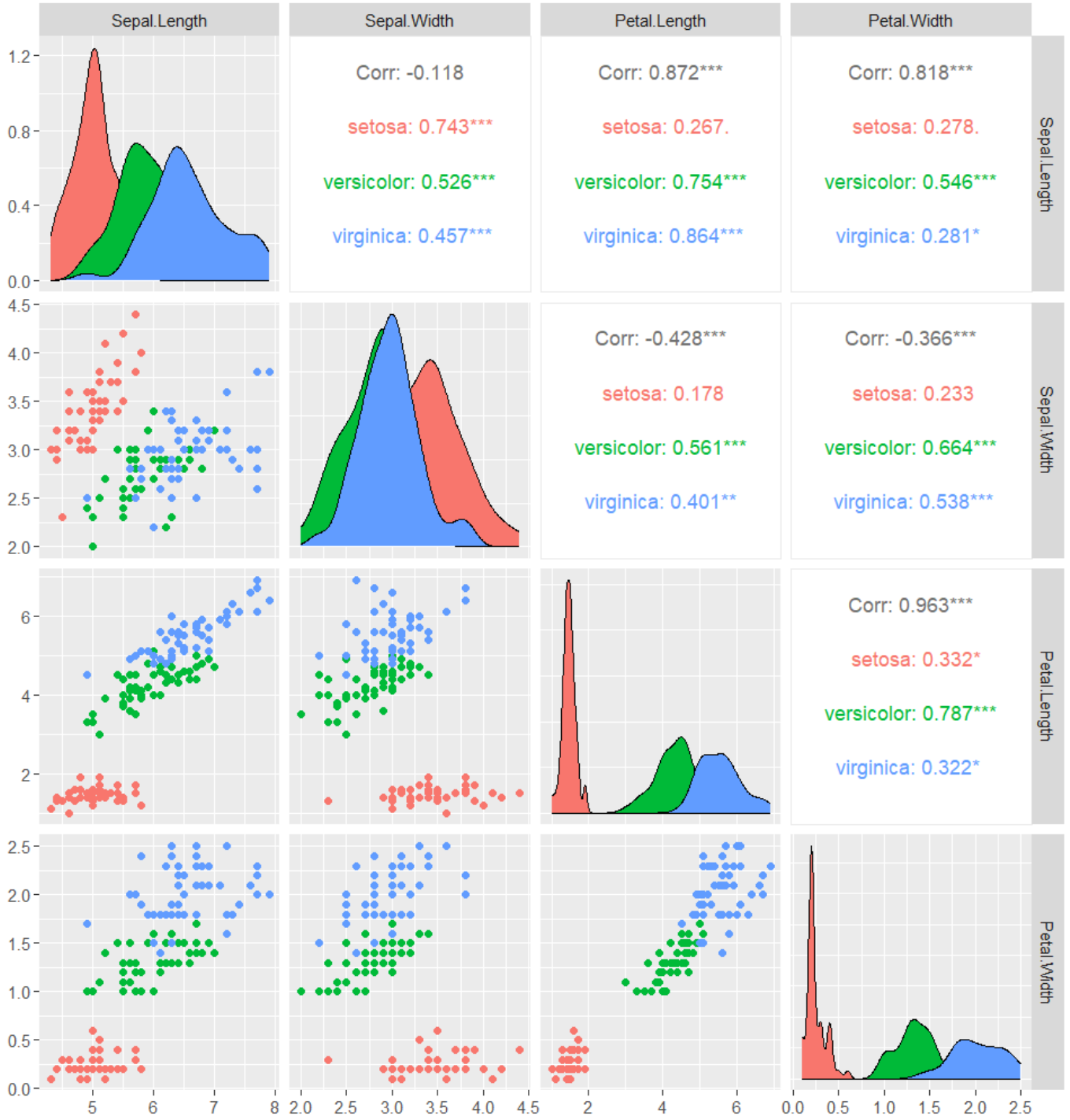
```
library(reshape2)
corr <- round(cor(iris[, 1:4]), 2)
corr_melt <- melt(corr)

ggplot(corr_melt, aes(Var1, Var2, fill = value)) +
  geom_tile() +
  geom_text(aes(label = value), color = "white") +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white", midpoint = 0) +
  labs(title = "Carte thermique des corrélations")
```

# Visualisation multivariée (Pairplot / Scatterplot Matrix)

- **Définition** : matrice de nuages de points entre toutes les variables quantitatives.
- **Utilité** : détecter corrélations, structures globales.
- **Exemple** : dataset *Iris* (relations entre longueurs/largeurs de pétales et sépales).

```
library(GGally)
ggpairs(iris[, 1:4], aes(color = iris$Species))
```


Sepal.Length
Sepal.Width
Petal.Length
Petal.Width
Corr: -0.118
setosa: 0.743***
versicolor: 0.526***
virginica: 0.457***
Corr: 0.872***
setosa: 0.267.
versicolor: 0.754***
virginica: 0.864***
Corr: 0.818***
setosa: 0.278.
versicolor: 0.546***
virginica: 0.281*
Corr: -0.428***
setosa: 0.178
versicolor: 0.561***
virginica: 0.401**
Corr: -0.366***
setosa: 0.233
versicolor: 0.664***
virginica: 0.538***
Corr: 0.963***
setosa: 0.332*
versicolor: 0.787***
virginica: 0.322*

# Principes d'une visualisation efficace

- **Clarté** : éviter la surcharge visuelle.
- **Choix du graphique adapté** (pas de pie chart pour comparer 20 catégories !).
- **Couleurs cohérentes et non trompeuses.**
- **Échelles correctes** : éviter axes tronqués qui induisent en erreur.
- Toujours mettre des titres, étiquettes, légendes.
- Éviter les représentations trompeuses.
- **Message clair** : chaque graphique doit répondre à une question précise.

# 3.3. Identification de motifs

# Identification de motifs

- Après avoir décrit et visualisé les données, l'EDA cherche à:
  - Repérer des relations entre variables
  - Détecter des comportements récurrents
  - Trouver des structures cachées
  - Mettre en évidence des anomalies
- Trois axes essentiels :
  - Corrélations & associations→ Relations entre variables numériques ou qualitatives
  - Analyse des tendances→ Identifier hausses, baisses, saisonnalité (séries temporelles)
  - Détection d'anomalies→ Valeurs atypiques qui méritent attention

**Objectif**: *Comprendre ce que racontent réellement les données pour guider les décisions ou la modélisation*.

# Corrélation

- Mesure la force et la direction de la relation entre deux variables quantitatives.
- Deux principaux coefficients :

| Coefficient | Type de relation mesurée | Intervalle | Commentaires |
|---|---|---|---|
| **Pearson** | Linéaire | [-1, 1] | Sensible aux valeurs aberrantes |
| **Spearman** | Monotone (basé sur les rangs) | [-1, 1] | Plus robuste aux outliers et aux distributions non normales |

- Critères d'interprétation généraux
  - ($|r| > 0.7$) → forte corrélation
  - ($0.3 < |r| \leq 0.7$) → corrélation modérée
  - ($|r| \leq 0.3$) → corrélation faible

# La corrélation de Pearson

- Soient deux variables quantitatives $\boldsymbol{X}$ = ($x_1$, $x_2$, …, $x_n$) et $\boldsymbol{Y}$ = ($y_1$, $y_2$, …, $y_n$) de taille $\boldsymbol{n}$.
- La corrélation de Pearson $\boldsymbol{r}$ est donnée par :

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

# Limites: Corrélation ≠ Causalité

- Une corrélation **ne prouve pas** une relation de cause à effet.
- Fausses relations possibles (facteur caché).
  - Exemple : la consommation de glace et les noyades augmentent en été, mais le vrai facteur est la température.
    - On observe une **forte corrélation** entre :

| Variable | Observation |
|---|---|
| Nombre de glaces vendues | augmente en été |
| Nombre de noyades | augmente aussi en été |

  - On pourrait croire que manger des glaces provoque des noyades.
  - Les deux variables dépendent d'un facteur commun:
    - La température / saison
      - Quand il fait chaud :
        - » Plus de gens vont à la plage ou à la piscine
        - » Plus de glaces sont vendues
  - Donc : Corrélées, mais sans relation causale directe

# Associations entre variables qualitatives

- Déterminer si **deux variables catégorielles** sont liées **statistiquement**.
- On utilise les **tableaux croisés** (de contingence)
- Tests statistiques : **Chi-2:** vérifie si deux variables qualitatives sont indépendantes.
- Exemple d'usage :
  - le type de client est-il associé à un mode de paiement ?

# Associations entre variables qualitatives: Exemple

```
set.seed(123)  # pour reproductibilité
# Nombre de clients
n <- 170
client <- data.frame(
  id = 1:n,
  type = sample(c("Particulier", "Entreprise"),
                size = n,
                replace = TRUE,
                prob = c(100, 70)),
  mode_paiement = sample(c("Espèces", "Carte"),
                         size = n,
                         replace = TRUE,
                         prob = c(0.45, 0.55)),
  montant = round(runif(n, min = 10, max = 500), 2)
)
```

# Associations entre variables qualitatives: Exemple

```
# Convertir en facteurs
client$type <- as.factor(client$type)
client$mode_paiement <- as.factor(client$mode_paiement)

# Aperçu
head(client)



# Tableaux de contingence
table(client$type, client$mode_paiement)
```

```
              Carte Espèces
  Entreprise     32      34
  Particulier    59      45
```

# Associations entre variables qualitatives: Exemple

```
# Test de Chi2
chisq.test(table(client$type, client$mode_paiement))
Pearson's Chi-squared test with Yates' continuity correction

data:  table(client$type, client$mode_paiement)
X-squared = 0.79706, df = 1, p-value = 0.372
```

L'hypothèse testée est :

- **$H_0$ (hypothèse nulle)** : les deux variables sont **indépendantes** (le mode de paiement ne dépend pas du type de client).
- **$H_1$ (hypothèse alternative)** : les deux variables sont **associées** (le mode de paiement dépend du type de client).
- Si **p-value < 0.05** → On rejette $H_0$: **association significative**

# Associations entre variables qualitatives: Exemple

```
# Test de Chi2
chisq.test(table(client$type, client$mode_paiement))
Pearson's Chi-squared test with Yates' continuity correction

data:  table(client$type, client$mode_paiement)
X-squared = 0.79706, df = 1, p-value = 0.372
```

- **X-squared = 0.79706** : c'est la statistique du test. Elle mesure l'écart entre les fréquences observées et celles attendues si les variables étaient indépendantes.
- **df = 1** : le nombre de degrés de liberté (dépend du nombre de catégories des variables).
- **p-value = 0.372** : la probabilité d'obtenir une telle statistique (ou plus extrême) si les deux variables étaient vraiment indépendantes.

# Associations entre variables qualitatives: Exemple 2

```
data <- mtcars

data$cyl <- factor(data$cyl,
     levels = c(4, 6, 8),
     labels = c("4 cylindres", "6 cylindres", "8 cylindres"))

data$am <- factor(data$am,
               levels = c(0, 1),
               labels = c("Automatique", "Manuelle"))



data
```

# Associations entre variables qualitatives: Exemple 2

```
table(data$cyl, data$am)

                Automatique Manuelle
  4 cylindres             3        8
  6 cylindres             4        3
  8 cylindres            12        2

chisq.test(data$cyl, data$am)

Pearson's Chi-squared test

data:  data$cyl and data$am
X-squared = 8.7407, df = 2, p-value = 0.01265
```

# Analyse des tendances

- **Détecter des motifs dans une série temporelle**
  - **Tendance (Trend)** : la direction générale d'une donnée sur une longue période, que ce soit une montée (augmentation) ou une descente (baisse).
  - **Saisonnalité** : sont des motifs ou comportements qui se répètent régulièrement, généralement à des intervalles précis (par exemple, chaque mois, chaque trimestre, chaque année).
  - **Cycles** : sont des variations qui suivent un rythme moins prévisible que la saisonnalité. Elles sont souvent influencées par des facteurs externes, comme les fluctuations économiques ou les crises financières sur des périodes plus longues.

# Analyse des tendances

- Exemples :
  - Ventes mensuelles d'un produit
  - Trafic web quotidien
  - Températures saisonnières
- **Méthodes simples d'exploration**
  - Courbes de tendance (régression)
  - Visualisation temporelle (line chart avec ggplot2)

# Détection d'anomalies

- **Qu'est-ce qu'une anomalie ?**
  - Une observation **très différente** des autres valeurs dans les données.
- **Méthodes basiques en EDA**

| Méthode | Utilisation |
|---|---|
| **Boîte à moustaches** | Détection visuelle via IQR |
| **Z-score** | Outliers si |
| **Scatterplots** | Outliers multivariés visibles |

# Détection d'anomalies

- Pourquoi détecter les anomalies ?
  - Éviter que des mauvaises données orientent les analyses
  - Identifier des évènements rares mais corrects et importants
- Applications :
  - Fraude bancaire
  - Capteurs défectueux
  - Données saisies erronées