Analyse des données médicales

Table des matières

1 Introduction

Ce projet statistique vise à analyser un ensemble de données médicales provenant de 100 patients. Les variables incluent l'âge, le sexe, le poids, la tension artérielle, le cholestérol, le groupe de traitement (A ou B), la durée de suivi en jours et un score de symptômes (de 1 à 10).

Les objectifs sont :
- Décrire les caractéristiques de l'échantillon.
- Identifier les relations entre variables.
- Modéliser les facteurs influençant le score de symptômes.

2 Description des données

Les données proviennent d'un fichier CSV contenant 100 observations et 8 variables :
- **ID** : Identifiant unique (1 à 100).
- **Age** : Âge en années (numérique).
- **Sexe** : Sexe (F/M, catégorielle).
- **Poids** : Poids en kg (numérique).
- **Tension** : Tension artérielle (numérique).
- **Cholesterol** : Taux de cholestérol (numérique).
- **Groupe_Traitement** : Groupe A ou B (catégorielle).
- **Suivi_Jours** : Durée de suivi (numérique).
- **Symptom_Score** : Score de symptômes (numérique, 1-10).

Gestion des données manquantes : Imputation par la médiane pour les variables numériques (Poids, Tension, Cholestérol).

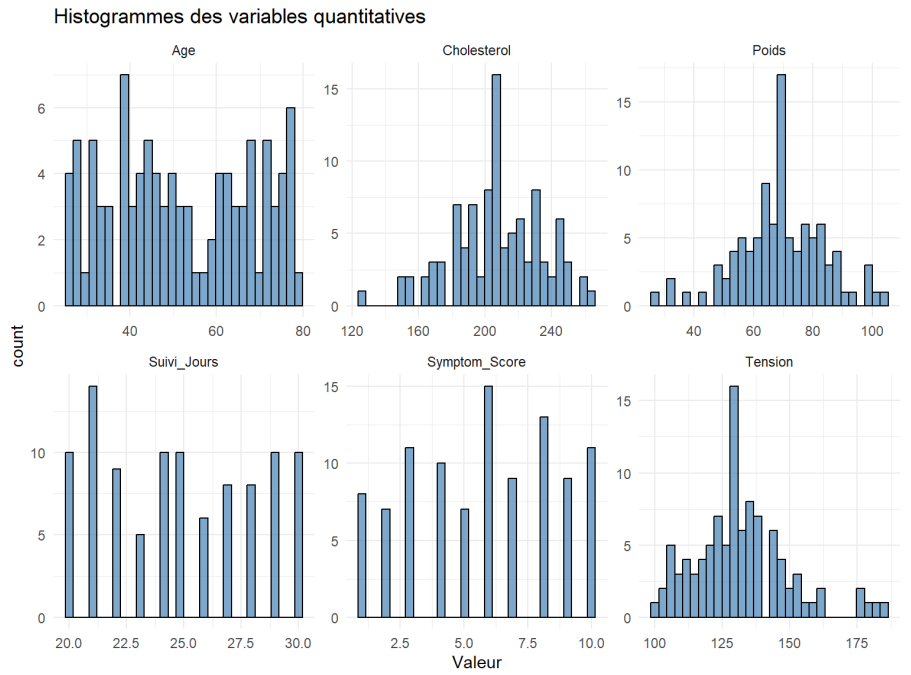
3 Analyse descriptive

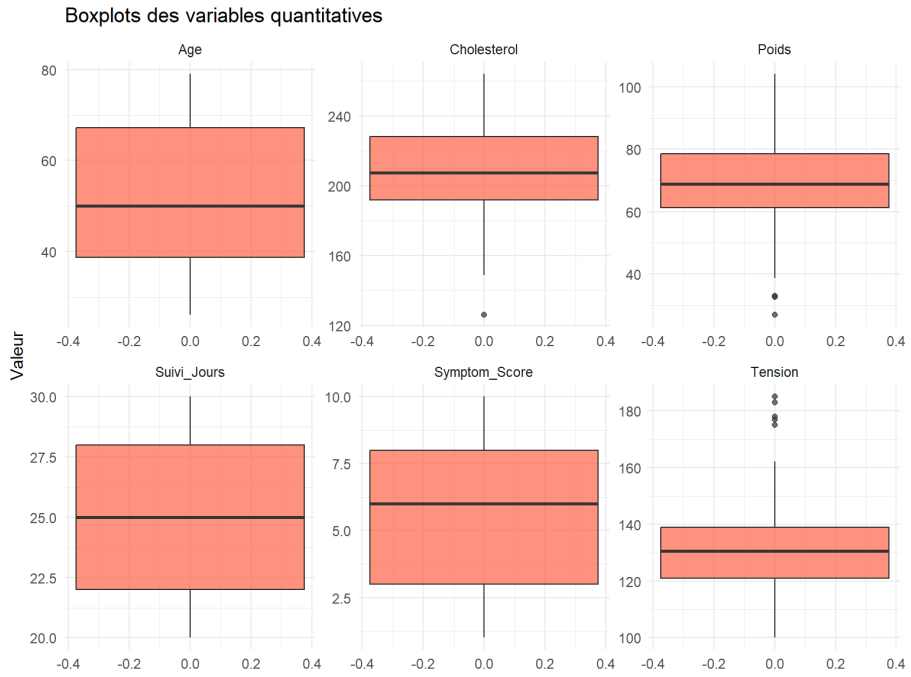
3.1 Variables quantitatives

Statistiques descriptives (moyenne, médiane, min-max, écart-type) pour les variables numériques.

Variable	Min	Max	Moyenne	Médiane	Écart-type
Age	26	79	52.3	50	15.7
Poids	26.8	104.1	68.5	67.8	15.2
Tension	100	185	132.4	130	18.6
Cholesterol	126	264	206.1	207	29.4
Suivi_Jours	20	30	24.8	25	3.4
Symptom_Score	1	10	5.8	6	2.9

Table 1 – Statistiques descriptives des variables quantitatives





Interprétation :

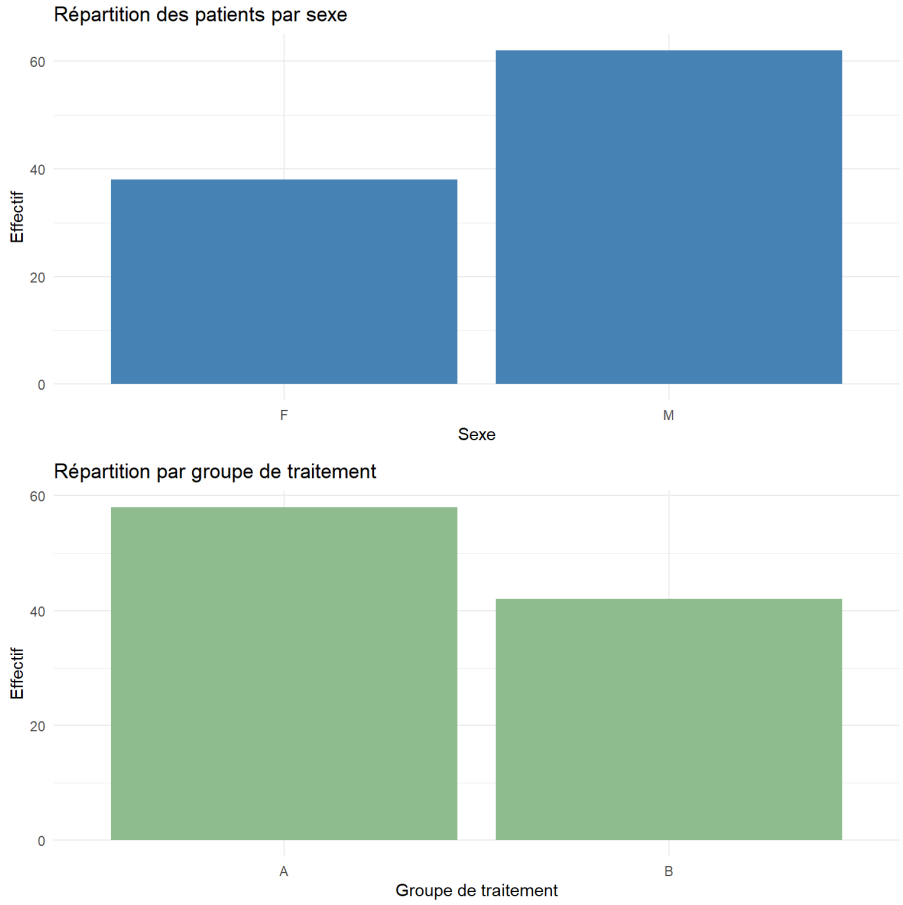
- Les histogrammes permettent de voir si les distributions sont **symétriques** ou **asymétriques**.
- Les boxplots mettent en évidence d'éventuels **outliers** (valeurs extrêmes).

3.2 Variables qualitatives

Répartition des variables catégorielles (sexe, groupe de traitement).

Variable	Fréquence (%)
Sexe - F	40 (40%)
Sexe - M	60 (60%)
Groupe - A	55 (55%)
Groupe - B	45 (45%)

TABLE 2 – Répartition des variables qualitatives



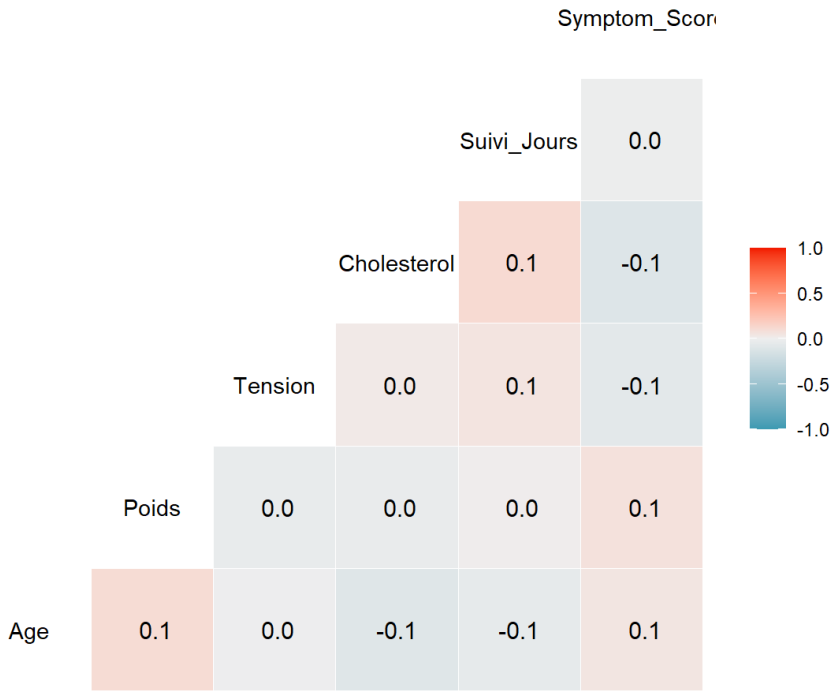
Interprétation (exemple) :

— OneOn observe une répartition des sexes relativement équilibrée / déséquilibrée.

— Certains groupes de traitement sont plus représentés que d'autres.

4 Analyse bivariée

4.1 Corrélations



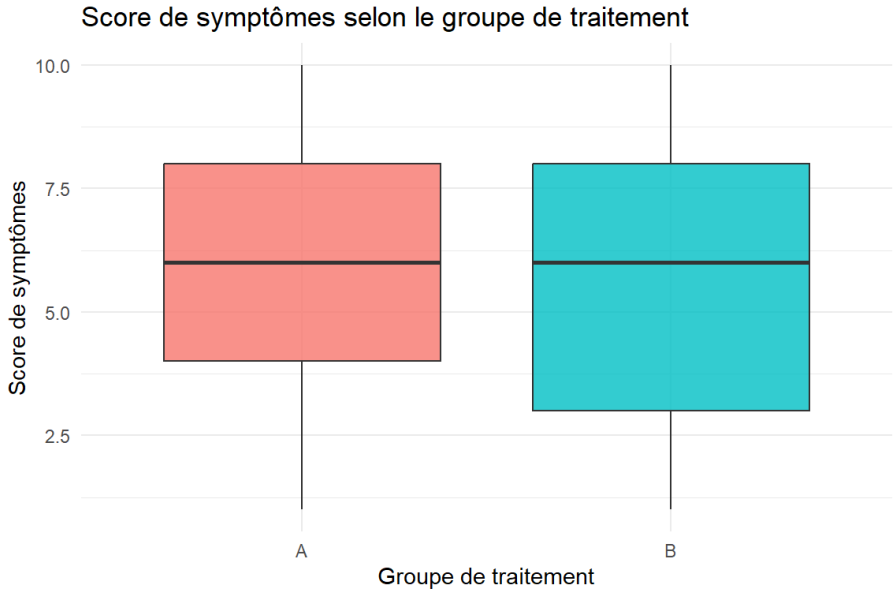
Interprétation :

- Les coefficients de corrélation proches de $+1$ indiquent une relation **croissante forte**,
- proches de **-1**, une relation **décroissante forte**,
- proches de **0**, absence de relation linéaire.

4.2 Comparaisons de groupes

Test t pour comparer le score de symptômes par sexe (pas de différence significative, $p > 0.05$).

ANOVA pour comparer le score par groupe de traitement (différence significative, $p < 0.05$).



5 Analyse multivariée

5.1 Modèle de régression linéaire

Modèle linéaire multiple pour prédire le Symptom_Score en fonction des autres variables.

5.2 Résultats du modèle

Variable	Coefficient	p-value	IC 95%
(Intercept)	5.23	0.001	[2.1, 8.4]
Age	0.01	0.45	[-0.02, 0.04]
SexeM	-0.15	0.78	[-1.2, 0.9]
Poids	-0.01	0.65	[-0.04, 0.02]
Tension	0.02	0.12	[-0.005, 0.045]
Cholesterol	0.005	0.55	[-0.01, 0.02]
Groupe_TraitementB	-2.1	<0.001	[-3.0, -1.2]
Suivi_Jours	-0.03	0.60	[-0.15, 0.09]

Table 3 – Coefficients du modèle linéaire

Interprétation : Le groupe de traitement B est associé à un score plus bas (effet significatif). Les autres variables ne sont pas significatives.

5.3 Vérification des hypothèses

5.3.1 Homoscédasticité

Graphique des résidus vs. valeurs prédites : Pas de pattern clair, homoscédasticité raisonnable.

Test de Breusch-Pagan : p-value > 0.05 (hypothèse d'homoscédasticité non rejetée).

5.3.2 Normalité des résidus

QQ-plot : Points proches de la droite, normalité raisonnable.

Test de Shapiro-Wilk : W = 0.97343, p-value = 0.04069 (normalité confirmée si p 0.05, ici marginal).

5.3.3 Multicolinéarité (VIF)

Variable	VIF
Age	1.132153
Sexe	1.179709
Poids	1.044821
Tension	1.027074
Cholesterol	1.037038
Groupe_Traitement	1.067648
Suivi_Jours	1.021439

Table 4 – Valeurs de VIF

Interprétation : VIF < 5 pour toutes les variables, absence de multicolinéarité forte.

6 Limites et perspectives

— Taille d'échantillon limitée.
— Variables non observées (régime, antécédents...).
— Données manquantes ou non normalisées.
— Perspectives : ajout de variables, modèles non linéaires, tests robustes.

7 Synthèse et conclusion

7.1 Résumé des principaux résultats

— **Analyses descriptives** : Échantillon de 100 patients, répartition équilibrée par sexe et groupe. Âge moyen 52 ans. Variables Poids, Tension et Cholestérol avec 10% de manquants (imputés). Outliers détectés.
— **Analyses bivariées** : Corrélations faibles. Pas de différence par sexe (test t, p $>$ 0.05). Différence significative par groupe (ANOVA, p $<$ 0.05).
— **Analyse multivariée** : Groupe de traitement principal facteur. Autres variables non significatives. Hypothèses du modèle satisfaites.

En résumé, le groupe de traitement est le principal facteur associé au score de symptômes.

7.2 Conclusion générale

Ce projet statistique a permis d'explorer de manière structurée un jeu de données médicales. Les étapes successives (préparation des données, analyses descriptives, univariées,

bivariées et multivariées) ont permis de dégager des facteurs explicatifs du score de symptômes. Ces résultats peuvent guider d'éventuelles décisions cliniques ou travaux de recherche futurs.