Vincent GUIBLAIN

Etape 1 : Compréhension du problème
Décrivez les variables disponibles.
Formulez clairement le problème métier.
Identifiez la variable cible et les variables explicatives.
Quelle est la problématique centrale pour la ferme ?

- surface_ha : Surface cultivée en hectares
- type_sol : Type de sol (argileux, sableux, limoneux)
- engrais_kg/ha : Quantité d'engrais utilisée en kg/ha
- precipitations_mm : Précipitations moyennes mensuelles en mm
- temperature_C : Température moyenne mensuelle en °C
- rendement_t/ha : Rendement obtenu en tonnes par hectare

Les variables disponibles sont celles citées ci-dessus.Elles sont toutes numériques sauf la variable type_sol qui prend les valeurs: “argileux”,”sableux” et “limoneux”.
On cherche ici à diminuer les coûts et augmenter les rendements. Ce qui revient à augmenter le rendement (rendement_t/ha) en optimisant la quantité d’engrais (engrais_kg/ha) en fonction du lieu (type_sol) et des conditions météo (precipitations_mm et temperature_C). La variable surface_ha est utile mais peu nécessaire redondante car les variables d’engrais et de rendement sont exprimées par hectare.

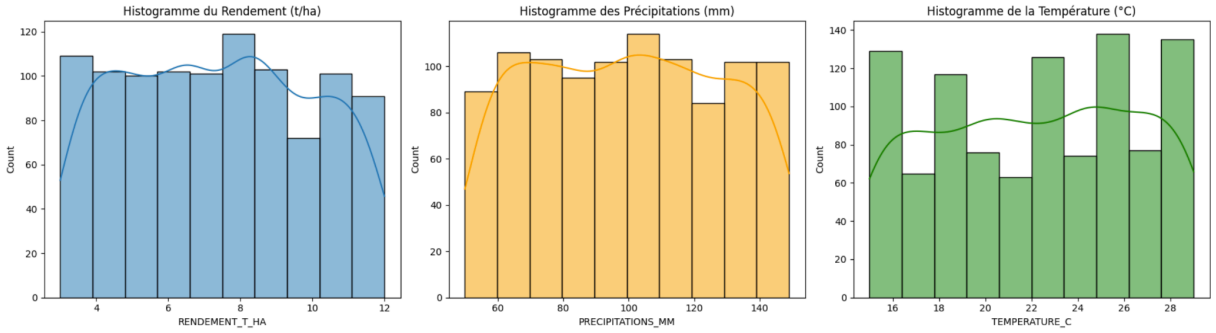
Etape 2 : Analyse statistique descriptive
2.1 Mesures de tendance centrale
Calculez la moyenne, médiane, et mode du rendement.

```
Moyenne Rendement: 7.378418687218944
Mediane Rendement: 7.349138167259971
Mode Rendement: 3.000276469608442
```

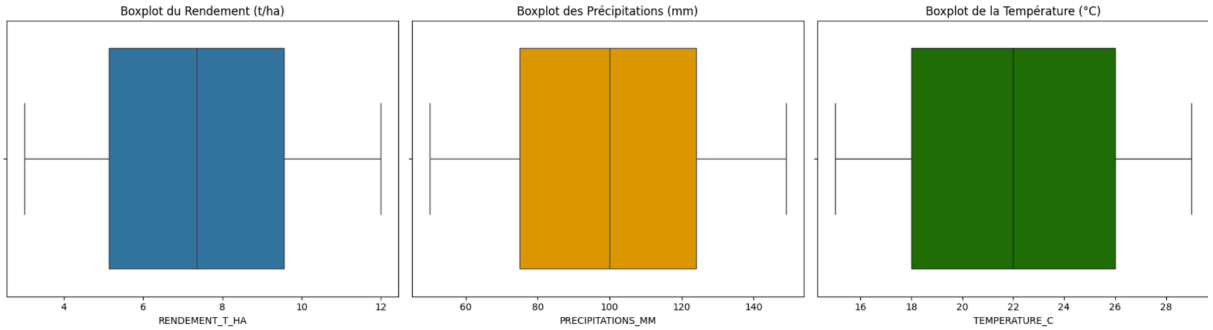
2.2 Mesures de dispersion
Calculez l'écart-type, variance, et étendue du rendement.

```
Écart-type: 2.5699909853267067, Variance: 6.604853664660536, Étendue: 8.99574285964550
```

2.3 Visualisation des données
Créez des histogrammes pour le rendement, les précipitations, et la température.



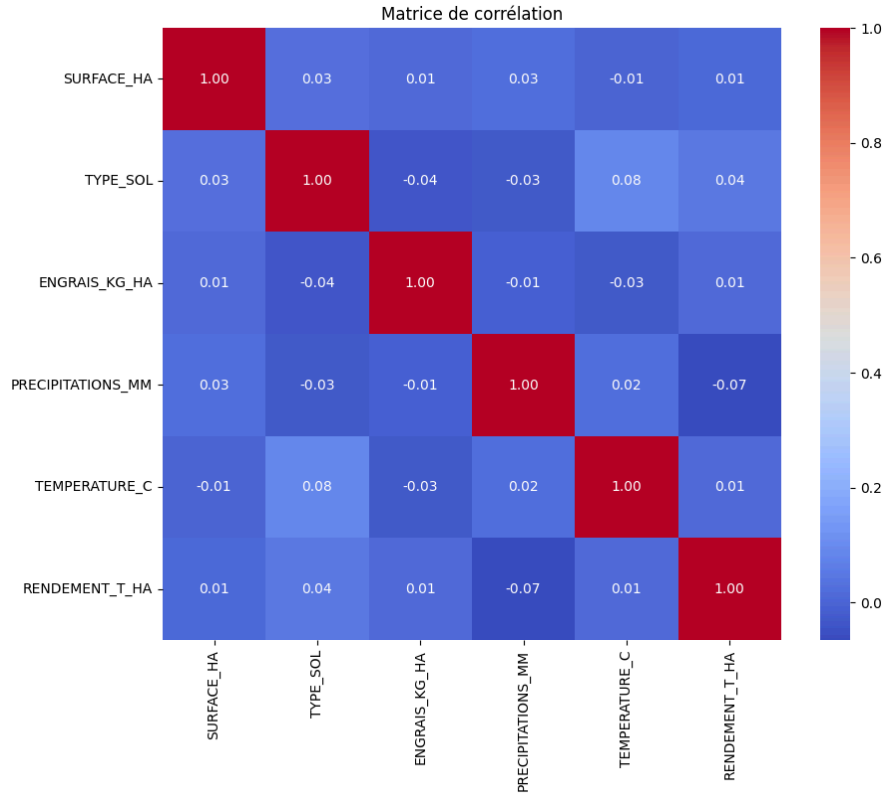
Affichez des boxplots pour identifier d'éventuels outliers.



2.4 Corrélations

Calculez la matrice de corrélation entre les variables numériques.

Affichez une heatmap pour visualiser les corrélations.



Quelles variables semblent avoir le plus d'impact sur le rendement ?

La matrice de corrélation nous donne des résultats extrêmement faibles (proche de 0) et donc peu exploitables mais pour la bonne exécution de ce TD nous allons tenter d'en extraire des informations judicieuses.

On remarque que la corrélation la plus forte est de -0.07 (sans compter les diagonales et les corrélations avec la variable type_sol) et elle concerne les précipitations et le rendement. Si ces deux variables ont une relation linéaire on peut en conclure que plus il y a de pluie et moins les rendements sont élevés. On peut également supposer que la pluie impacte les rendements dans une certaine mesure mais qu'à partir d'un moment, si les quantités d'eau déversées sont trop importantes, les plants ont plus de chances de se noyer ce qui diminue les rendements.

Etape 3 : Analyse de la variance (ANOVA)
3.1 Hypothèses
H0 : Le type de sol n'influence pas le rendement.
H1 : Le type de sol influence le rendement.
3.2 Test ANOVA
Réalisez une ANOVA sur le type de sol.
Interprétez la p-value obtenue.
Le type de sol a-t-il une influence significative sur le rendement ?

OLS Regression Results

Dep. Variable:	RENDEMENT_T_HA	R-squared:	0.007
Model:	OLS	Adj. R-squared:	0.002
Method:	Least Squares	F-statistic:	1.303
Date:	Mon, 31 Mar 2025	Prob (F-statistic):	0.260
Time:	21:34:09	Log-Likelihood:	-2359.1
No. Observations:	1000	AIC:	4730.
Df Residuals:	994	BIC:	4760.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	7.5239	0.585	12.870	0.000	6.377	8.671
SURFACE_HA	0.0094	0.031	0.302	0.763	-0.052	0.071
TYPE_SOL	0.1331	0.101	1.322	0.187	-0.065	0.331
ENGRAIS_KG_HA	0.0008	0.002	0.424	0.672	-0.003	0.004
PRECIPITATIONS_MM	-0.0058	0.003	-2.042	0.041	-0.011	-0.000
TEMPERATURE_C	0.0070	0.019	0.370	0.711	-0.030	0.044

Omnibus:	451.006	Durbin-Watson:	1.988
Prob(Omnibus):	0.000	Jarque-Bera (JB):	53.819
Skew:	0.055	Prob(JB):	2.06e-12
Kurtosis:	1.869	Cond. No.	1.21e+03

La p-value obtenue est environ égale à 0.15 et elle est supérieure à notre seuil de signification alpha(0.05) donc on ne rejette pas H0. Autrement dit, le type de sol n'a pas une influence significative sur le rendement.
Conclusion: Le type de sol n'influence pas le rendement.

	sum_sq	df	F	PR(>F)
TYPE_SOL	13.130921	1.0	1.990042	0.158648
Residual	6585.117890	998.0	NaN	NaN

Etape 4 : Modélisation
4.1 Séparation des données
Divisez les données en train (80%) et test (20%).
4.2 Création du modèle
Entraînez des modèles de votre choix vu précédemment pour prédire le rendement.
4.3 Évaluation du modèle
Calculez les métriques : MAE, RMSE, et R2 de ces modèles.
Lequel des modèles est-il performant (pourquoi d'après vous) ?

```
MAE: 2.0954865111086325, RMSE: 2.462539510233106, R²: -0.028294715676091986
```

Le MAE varie de manière linéaire car il s'agit de la valeur absolue de la différence entre la prédiction et la valeur réelle tandis que le RMSE est plus sujet aux fortes variations entre prédiction et valeur réelle. Ici le R^2 est négatif et très proche de 0. En d'autres termes, le modèle est inadapté et n'explique rien (0%).

Etape 5 : Interprétation et recommandations
Analysez l'importance des variables.
Proposez des recommandations concrètes pour augmenter le rendement (ex : ajuster l'engrais, choisir un type de sol particulier, etc.).
Identifiez les limites du modèle et proposez des pistes d'amélioration.
Quelles décisions la ferme pourrait-elle prendre pour optimiser sa production ?

La température moyenne par mois est importante mais loin d'être suffisante. En effet, afin de réunir les conditions optimales il est nécessaire de connaître constamment la température à laquelle les plants sont exposés. Pour obtenir un rendement maximal, il est important d'apporter à la plante les besoins qu'elle requiert et la température doit toujours être comprise entre certaines valeurs. Ainsi, il est intéressant d'avoir la température minimale et maximale pour chaque jour. En ce qui concerne les précipitations, les données sont trop imprécises. D'après le boxplot il semblerait que nos valeurs soient quasi-continues mais il se pourrait qu'une journée soit trop pluvieuse et qu'on n'arrive pas à le détecter avec notre modèle.
Il est également important de savoir si le maïs est traité contre les parasites ou non.

La ferme pourrait équiper ses plantes de divers capteurs afin de réunir les données nécessaires. Elle peut également faire appel à un fermier expérimenté qui lui apportera toutes sortes de bons conseils.