UPEC
Connaissance - Action

UNIVERSITÉ
PARIS-EST CRÉTEIL
VAL DE MARNE

TP2: TD ATDN2

Analyse de Données et Méthodes d'Ensemble

GAKOU YOUSSOUF

27/03/2025

1.Introduction

Dans ce TP j'ai étudié un jeu de données sur l'élevage de poulets l'objectif était d'utiliser des outils statistiques pour comprendre les données détecter si il ya anomalies et ensuite réduire leur dimension pour mieux les visualiser et ensuite utiliser des algorithmes d'apprentissage pour faire des prédictions sur la survie et la prise de poids des poulets je vais présenter les résultats obtenus.

Description de la base de données utilisée

J'ai travaillé avec une base de donnees nommée « donnees_elevage_poulet.csv » Dans ce base de donnees il ya des informations provenant d'un élevage de poulets elle contient 200 observations. Chaque ligne représente un poulet spécifique avec différentes mesures réalisées pendant son élevage.
Voici les variables étudiées :

```
df = pd.read_csv('donnees_elevage_poulet.csv')
df.head()
```

	Poids_poulet_g	Nourriture_consommee_g_jour	Temperature_enclos_C	Humidite_%	Age_poulet_jours	Gain_poids_jour_g	Taux_survie_%	Cout_elevage_FCFA
0	3974	52	27.6	79.3	24	12.0	81.1	2682
1	1660	152	31.7	62.5	42	12.2	89.1	6626
2	2094	186	30.1	64.8	29	18.8	90.4	8424
3	1930	111	29.2	87.0	63	13.8	92.9	1933
4	1895	100	26.1	78.2	21	5.5	93.0	4598

Partie 1 : Analyse exploratoire des données

Exercice 1 : Statistiques descriptives

Question 1 *: Calculez la moyenne, médiane, écart-type, variance et les quartiles pour les variables poids, nourriture et température.*

On peut voire que le poids moyen des poulets est d'environ 2510 grammes et que la moitié des poulets pèse moins de 2482 grammes médiane le poids lui varie beaucoup car l'écart type est élevé il est de 898 grammes.

```
stats.loc['variance'] = df[['Poids_poulet_g', 'Nourriture_consommee_g_jour', 'Temperature_enclos_C']].var()
stats
```

	Poids_poulet_g	Nourriture_consommee_g_jour	Temperature_enclos_C
count	200.000000	200.000000	200.000000
mean	2509.580000	129.745000	28.389000
std	898.436875	44.006166	2.065724
min	821.000000	51.000000	25.000000
25%	1810.750000	95.750000	26.600000
50%	2481.500000	135.500000	28.500000
75%	3356.500000	165.250000	30.300000
max	3974.000000	199.000000	31.900000
variance	807188.817688	1936.542688	4.267215

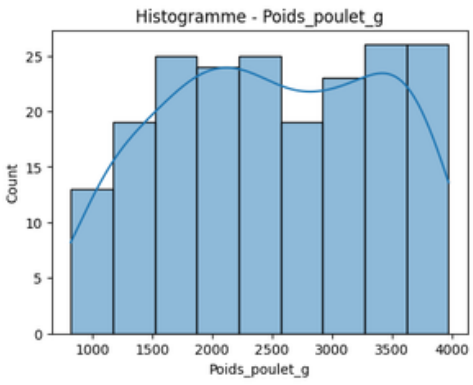
Pour la nourriture consommée en moyenne chaque poulet consomme environ 130 grammes par jour avec une grande variation entre les poulets l'ecart type et d'environ 44 grammes

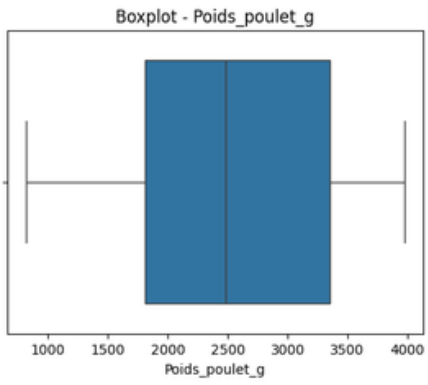
et pour la température de l'enclos la moyenne est d'environ 28,4degres avec une faible variation lécart type et d'environ 2degres.

Ces valeurs montre que le poids et la quantité de nourriture varient beaucoup d'un poulet à l'autre alors que la température est plus stable.

Exercice 1 : Statistiques descriptives

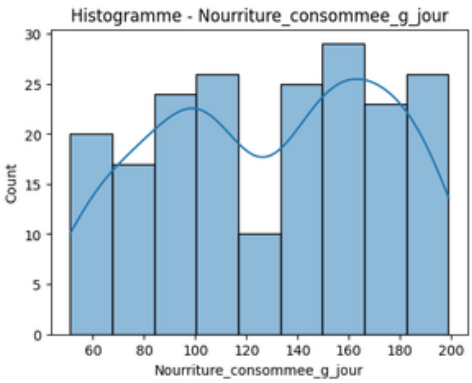
Question 2 *:Tracez des histogrammes et des boxplots pour visualiser la répartition des données. Que pouvez-vous déduire de ces graphiques ? Les données semblent-elles homogènes ou dispersées ?*

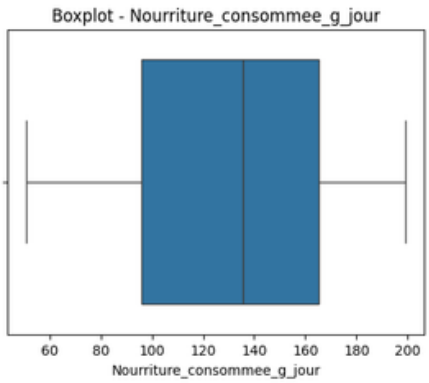




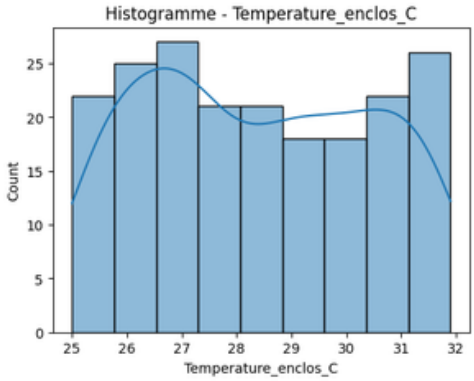
Pour le poids on peut voire que les données sont dispersées certain poulets sont légers environ 800 grammes d'autres sont très lourds jusqua presque 4000 grammes.

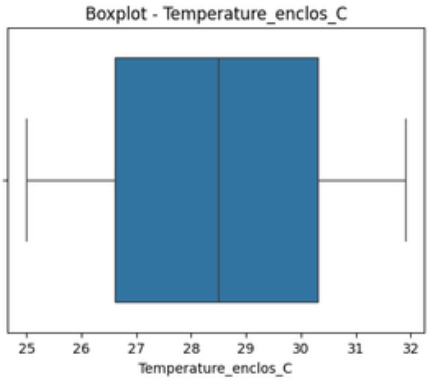
Partie 1 : Analyse exploratoire des données





Pour la consommation de nourriture on vois aussi une forte dispersion certains poulets mangent très peu, d'autres beaucoup.





Et pour la température est plus homogène la plupart des valeurs se regroupent autour de 28 degres

le poids et la nourriture sont très variables alors que la température est plus régulière

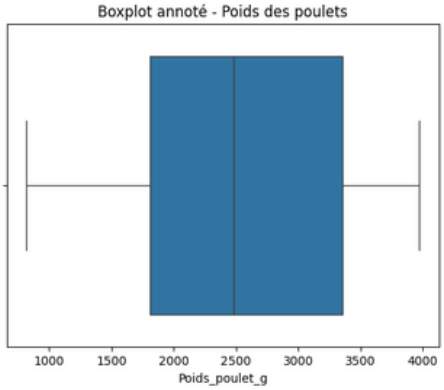
Exercice 2 : Détection des outliers

Question 3 *:Détectez les outliers (valeurs extrêmes) avec les méthodes IQR et Z-score. Comparez les résultats.*

Question 4 *:*

Visualisez ces outliers sur un boxplot annoté. Les outliers détectés sont-ils réalistes ou issus d'erreurs de mesure ? Faut-il les exclure ou les garder ? Justifiez votre choix.

J'ai testé les deux méthodes IQR et Z-score pour détecter des valeurs anormalement élevées ou basses, mais aucune méthode n'a détecté de valeurs extrêmes dans les données par contre lorsque je modifie les paramètre des méthodes j'obtient des résultats mais cela ne respecte pas les convention je donc décide de garder les paramètre par défaut



Comme il ya aucune valeur extrême sans que je ne modifie les paramètre de depart il ny avait rien a annoter sur les boxplots ducoup les données semblent réalistes et bien mesurées donc je décide de les garder toutes.

Partie 1 : Analyse exploratoire des données

Exercice 3 : Tests paramétriques

Question 5 *:Testez la normalité des variables (poids, nourriture, température) avec le test de Shapiro-Wilk. Expliquez ce que vous observez.*

le test de Shapiro avec un seuile α=0.05 renvoie :

- Poids_poulet_g : p-value = 9.09e-06
- Nourriture_consommee_g_jour : p-value =6.23e-07
- Temperature_enclos_C : p-value = 4.40e-07

Tous les p value sont inférieurs a 0.05 donc on rejette l'hypothese. Ces variables ne suivent donc pas une distribution normale.

Exercice 3 : Tests paramétriques

Question 6*:Comparez les moyennes de deux groupes avec le test t de Student, puis utilisez une ANOVA pour comparer les moyennes de plusieurs groupes. Interprétez les résultats.*

```
group1 = df[df['Age_poulet_jours'] <= df['Age_poulet_jours'].median()]['Temperature_enclos_C']
group2 = df[df['Age_poulet_jours'] > df['Age_poulet_jours'].median()]['Temperature_enclos_C']
ttest_result = ttest_ind(group1, group2)

anova_result = f_oneway(
    df[df['Age_poulet_jours'] <= 20]['Temperature_enclos_C'],
    df[(df['Age_poulet_jours'] > 20) & (df['Age_poulet_jours'] <= 40)]['Temperature_enclos_C'],
    df[df['Age_poulet_jours'] > 40]['Temperature_enclos_C']
)

ttest_result, anova_result
```

```
(TtestResult(statistic=np.float64(0.2527089007576175), pvalue=np.float64(0.80075532778353), df=np.float64(198.0)),
 F_onewayResult(statistic=np.float64(0.4596773459717096), pvalue=np.float64(0.632162962501347)))
```

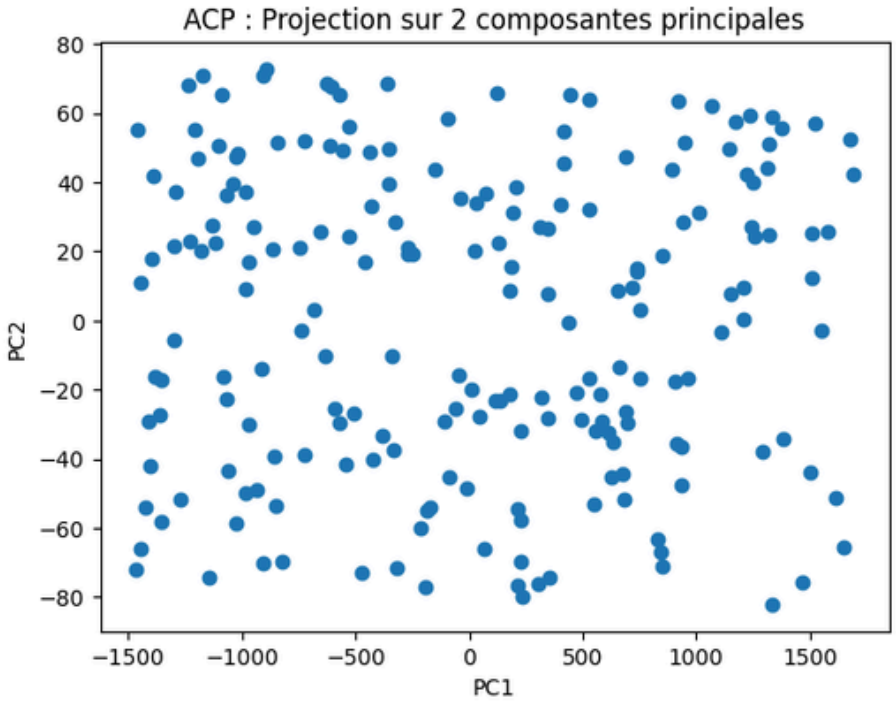
L'anova sur les trois groupes dâge distincts a une p-valeur de 0,63 ca veut dire que il nya pas de grande différence la température ne varie donc pas de manière significative en fonction de l'âge.

Partie 2 : Réduction de dimensionnalité

Exercice 4 : ACP classique

Question 7:*Implémentez une ACP sans scikit-learn (numpy). Calculez la matrice de covariance, valeurs propres et vecteurs propres.*

Question 8:*Projetez les données sur les deux premières composantes principales et visualisez le résultat. Combien de composantes gardez-vous ?*

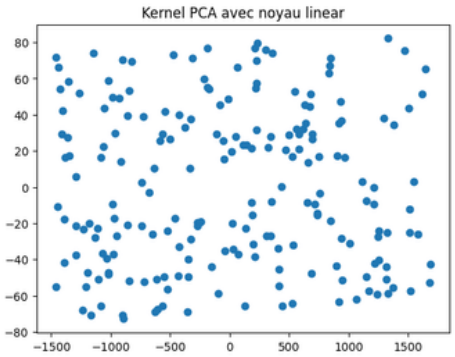


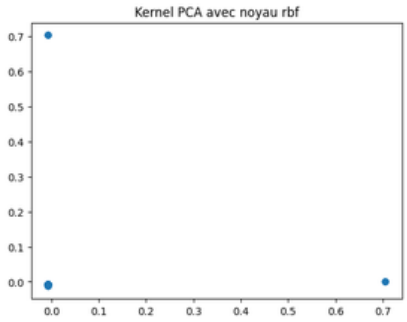
Après avoir projeté les données sur PC1 et PC2 comme le montre le nuage de points on observe la repartition dans un plan 2D ici on garde 2 composantes parce que cela suffit à représenter l'essentiel de la variance et à visualiser le jeu de données

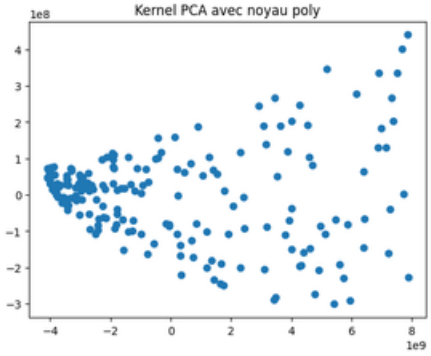
Exercice 5 : ACP ACP à Noyau

Question 9:*Appliquez KernelPCA (avec scikit-learn) sur les données et testez différents noyaux (linéaire, RBF, polynomiQuestion)*

Question 10: *Comparez les résultats avec l'ACP classique. Dans quels cas l'ACP à noyau donne-t-elle de meilleurs résultats ?*







Avec l'ACP classique on se base sur des combinaisons linéaires des variables, donc si les données sont à peu près alignées sur un plan, ça suffit par contre si les donnees ont des formes plus complexes l'acp classique ne les voit pas. L'ACP a noyau va transformer l'espace pour mieux séparer ces formes. Du coup c'est utile quand on pense que les variables sont reliées de façon non linéaire et qu'on veut capturer des structures plus compliques