**UNIVERSITÉ NATIONALE DE HO CHI MINH VILLE**

**UNIVERSITÉ DES SCIENCES NATURELLES**

**FACULTÉ DES TECHNOLOGIES DE L'INFORMATION**



**PROJET**

**(TRAITER DES CAS CONCRETS D’ANALYSE DE DONÉES )**

**Groupe 10**

|  |  |
| --- | --- |
| **Nom et Prénom** | **Code d’étudiant** |
| Phạm Quang Huy | 20126016 |
| Nguyễn Huỳnh Mẫn | 20126041 |
| Thiều Vĩnh Trung | 20126062 |

**ANALYSE STATISTIQUE MULTIVARIÉE**

Ho Chi Minh ville – 2022

**UNIVERSITÉ NATIONALE DE HO CHI MINH VILLE**

**UNIVERSITÉ DES SCIENCES NATURELLES**

**FACULTÉ DES TECHNOLOGIES DE L'INFORMATION**

A picture containing logo

Description automatically generated

**ĐỒ ÁN**

**(TRAITER DES CAS CONCRETS D’ANALYSE DE DONNÉES )**

**|** **Instructeurs |**

Mme. Nguyễn Thị Mộng Ngọc

M. Nguyễn Văn Thìn

**ANALYSE STATISTIQUE MULTIVARIÉE**

Ho Chi Minh ville – 2022

**INTRODUIRE**

Một nghiên cứu tốt, một báo cáo khoa học có căn cứ được người đọc chấp nhận về mặt học thuật đòi hỏi phải có phương pháp tốt, áp dụng các công cụ kỹ thuật để cung cấp các thông tin xác thực. Đặc biệt trong các vấn đề kinh tế - xã hội và khi nghiên cứu số lớn chúng ta cần phải quan tâm đến các công cụ kỹ thuật như thống kê.

Thống kê học là một lĩnh vực khá rộng, do vậy trong phạm vi của môn học này mong muốn trang bị cho người học những kiến thức cơ bản trong việc phân tích số liệu thống kế nhằm mục đích có thể khai thác hiệu quả các thông tin thu thập được, để phục vụ cho công tác nghiên cứu khoa học của các khoa học về kinh tế - xã hội.

*Khái niệm phân tích số liệu thống kê*: Là sự kết hợp giữa thống kê, sự tư duy và hiểu biết các vấn đề kinh tế.

Yêu cầu: Để có thể nắm vững kiến thức của môn học này đòi hỏi người học phải có kiến thức sâu về thống kê, về kinh tế cũng như những hiểu biết thực tế của vấn đề nghiên cứu. Ngoài ra, cần phải có kiến thức về tin học và các công cụ lượng hoá khác để kết hợp trong nghiên cứu.

Ở đồ án lần này, nhóm chúng em xin trình bày về ……. Trong đồ án này, sẽ có 3 bài toán cần giải quyết giúp chúng ta hiểu rõ hơn về quy trình xử lý bài toán cũng như có cái nhìn tổng quát về ….

**MỤC LỤC**

[**INTRODUIRE** 1](#_Toc122876384)

[**Modèle de régression linéaire multiple (sujet de votre choix)** 3](#_Toc122876385)

[**1.** **Présentation du sujet** 3](#_Toc122876386)

[**2.** **Afficher la modèle d'origine** 3](#_Toc122876387)

[**3.** **Visualisation des données** 4](#_Toc122876388)

[**4.** **Éliminer les valeurs outliers** 7](#_Toc122876389)

[**5.** **Deuxième vérification des valeurs outliers** 8](#_Toc122876390)

[**6.** **Modélisation des ensembles de données nettoyés** 8](#_Toc122876391)

[**7.** **Méthode AIC** 9](#_Toc122876392)

[**8.** **Méthode BIC** 10](#_Toc122876393)

[**9.** **Meilleure modélisation d'ensemble de données** 11](#_Toc122876394)

[**10.** **Calculez la valeur résiduelle et voyez si le nouveau modèle suit une distribution normale** 12](#_Toc122876395)

[**Deux cas études** 1](#_Toc122876396)

[**I.** **Choix du modèle Exercice:** Taux d’accidents 1](#_Toc122876397)

[**1.** **Afficher la modèle d'origine** 1](#_Toc122876398)

[**2.** **Remplacer la colonne de données en position médiane** 2](#_Toc122876399)

[**3.** **Remplacez les données de la colonne par Q3** 3](#_Toc122876400)

[**4.** **La nouveau modèle basé sur les données filtrées du T3 (modele\_2)** 4](#_Toc122876401)

[**5.** **Calculez les résidus et vérifiez si le nouveau modèle suit une distribution normale** 6](#_Toc122876402)

[**II.** **ANOVA à deux facteurs Exercice:** Agents toxiques 6](#_Toc122876403)

[**1.** **Shapiro-Wilk normalité test** 7](#_Toc122876404)

[**2.** **Test Fisher hypothesis avec a = 5%:** 8](#_Toc122876405)

[**3.** **TurkeyHSD Test** 10](#_Toc122876406)

[**Les références** 1](#_Toc122876407)

[**Consulter le guide de traitement des données et la théorie** 1](#_Toc122876408)

[**Les sources de données** 1](#_Toc122876409)

# **Modèle de régression linéaire multiple (sujet de votre choix)**

1. **Présentation du sujet**

Construire une modele qui prédit le poids des poissons vendus sur le marché.

Avec l'ensemble de données fourni comme suit :

* Weight: poids (gram) → weight of fish in Gram g
* Length1: longueur verticale (cm) → vertical length in cm
* Length2: longueur diagonale (cm) → diagonal length in cm
* Length3: longueur de croix (cm) → cross length in cm
* Height: taille (cm) → height in cm
* Width: largeur diagonale (cm) → diagonal width in cm

1. **Afficher la modèle d'origine**

Text

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

Nous choisissons le modèle en fonction des données d'entrée.

- Après avoir utilisé la commande summary(), nous commentons le modèle d'origine, il semble que le modèle n'ait aucun sens.

Nous pouvons voir que, bien que la valeur de R-carré ajusté : 0,8815 soit assez élevée, il reste des colonnes qui ne sont pas marquées d'un astérisque (\*) → cela signifie que notre modèle à 5 variables n'est pas significatif pour prédire le poids .quantité.

🡪 Il est possible qu'il y ait des défauts dans les données d'entrée d'origine. Nous devons vérifier cela.

1. **Visualisation des données**

La première chose que nous pouvons faire est de voir si nos données ont des valeurs outliers.

Nous devons boxplot() sur les 5 variables de nos données pour avoir un aperçu de notre jeu de données.



Chart, box and whisker chart

Description automatically generated

Selon le graphique ci-dessus, nous pouvons voir que nos colonnes ont la présence de valeurs outliers. Exactement comme nous l'espérions.

Besoin d'une vue plus claire des colonnes des données. Le but est de localiser les valeurs outliers dans les données et de les supprimer.

Utilisez boxplot() sur chaque variable des données.

Text

Description automatically generated

Chart, box and whisker chart

Description automatically generatedChart, box and whisker chart

Description automatically generated

Colonne de valeur Longueur1 Colonne de valeur Longueur2Chart, box and whisker chart

Description automatically generatedChart, box and whisker chart

Description automatically generated

Colonne de valeur Hauteur Colonne de valeur Longueur3

Chart, box and whisker chart

Description automatically generated

Colonne de valeur Largeur

1. **Éliminer les valeurs outliers**

Une fois que nous connaissons l'emplacement des valeurs outliers, nous procédons à leur suppression de nos données.

Text

Description automatically generated

Vérifiez à nouveau si ces valeurs ont été supprimées.

A picture containing text

Description automatically generated

On confirme que nous avons éliminé les valeurs outliers que nous venons de trouver.

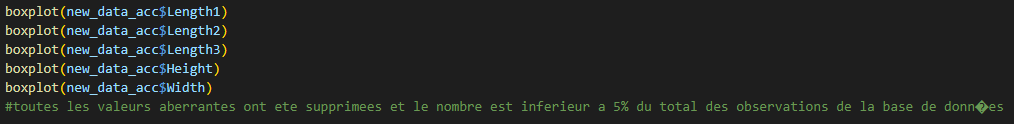
1. **Deuxième vérification des valeurs outliers**

Étant donné que le nombre de valeurs outliers supprimées est inférieur à 5 % de nos observations totales, le modèle ne sera pas affecté. Pour être plus approfondi, vérifions s'il y a des valeurs outliers dans notre ensemble de données.



Chart, box and whisker chart

Description automatically generated



Toutes les valeurs outliers ont été complètement supprimées du jeu de données.

1. **Modélisation des ensembles de données nettoyés**

Une fois les données nettoyées, nous utilisons la régression linéaire pour modéliser le nouvel ensemble de données.

A screenshot of a computer

Description automatically generated with medium confidence



On peut commenter, le modèle a été amélioré. Maintenant, la valeur R-carré ajusté : 0,9006. Cela signifie que le modèle est meilleur que l'original.

Mais je ne peux toujours pas dire que c'est un bon modèle car la colonne Length2 a une valeur de 0,19224 > 0,5. Cela peut signifier que la colonne n'a aucune signification dans notre modèle.

Pour tester cela, nous utilisons la méthode AIC pour estimer la qualité du modèle → à travers laquelle nous pouvons choisir le meilleur modèle.

1. **Méthode AIC**

Utiliser la méthode AIC pour sélectionner le meilleur modèle à partir de l'ensemble de données traitées.

Text

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

A partir de là, on choisit le meilleur modèle avec 4 variables : Longueur1, Longueur3, Hauteur, Largeur.

Cela confirme que l'hypothèse selon laquelle la colonne Length2 ne sera pas significative dans le modèle est vraie.

Après avoir exécuté step() sur l'ensemble de données traité, nous choisissons le meilleur modèle de régression linéaire avec la formule :Weight ~ Length1 + Length3 + Height + Width.

Régression linéaire → y = -431.98 + 63.64(Length1) - 44.51(Length3)

+37.76(Height) + 45.06(Width).

1. **Méthode BIC**

Utilisez la méthode BIC pour confirmer si le modèle sélectionné ci-dessus est le meilleur modèle ou non.

Parce que la probabilité de choisir le bon modèle de cette méthode est supérieure à celle de la méthode AIC. Alors peut-être qu'avec cette méthode, nous pouvons avoir un meilleur modèle plus petit que le modèle sélectionné dans la méthode AIC.

Text

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

Après avoir sélectionné le nombre d'observations de l'ensemble de données traité et exécuté step() selon la méthode BIC sur l'ensemble de données traité, nous obtenons toujours le meilleur modèle de régression linéaire avec la formule: Weight ~ Length1 + Length3 + Height + Width.

Régression linéaire → y = -431.98 + 63.64(Length1) - 44.51(Length3)

+37.76(Height) + 45.06(Width).

À partir de là, nous pouvons confirmer que la formule ci-dessus avec un nombre variable de 4 (Longueur1, Longueur3, Hauteur, Largeur) est notre meilleur modèle.

1. **Meilleure modélisation d'ensemble de données**



A screenshot of a computer

Description automatically generated with low confidence

En observant le tableau de statistiques traité et amélioré ci-dessus, nous pouvons confirmer qu'il s'agit de notre meilleur modèle pour expliquer le poids, toutes les colonnes sont significatives, toutes les valeurs sont inférieures à 0,001.

La valeur du R-carré ajusté : 0,9002, est proche de 1. Cette valeur s'est considérablement améliorée par rapport à l'ensemble de données d'origine fourni.

1. **Calculez la valeur résiduelle et voyez si le nouveau modèle suit une distribution normale**



Text

Description automatically generated

Utilisez shapeiro.test() pour voir si les résidus du nouveau modèle suivent une distribution normale.

Nous obtenons le p-valu = 2.298e-06 < 0.5 → Cela signifie que le modèle ne suit pas une distribution normale.

1. **Conclusion**

Au départ, on nous donne une donnée. Après avoir modélisé les données, nous avons remarqué que le modèle des données d'origine n'était pas bon.

Faites le pas pour voir si l'ensemble de données a des valeurs outliers, afin de les supprimer, afin d'amener les observations à un certain modèle.

Modélisez à nouveau l'ensemble de données pour la deuxième fois. Le modèle s'est considérablement amélioré, mais il reste encore de la place pour d'autres améliorations. Nous remarquons que l'une des cinq colonnes du jeu de données (Length2) peut ne pas être significative pour ce modèle. Passez à l'étape suivante du test.

Adoptez la méthode AIC et BIC. Nous pouvons confirmer que la prédiction ci-dessus est correcte. Continuez à supprimer cette colonne non significative (Length2) du modèle et nous avons le meilleur modèle.

Modélisez le meilleur ensemble de données. Nous considérons ce modèle comme notre meilleur modèle. Et il peut prédire pour notre colonne Poids. Passez à l'étape suivante.

Vérifiez si le modèle que nous avons construit suit une distribution normale. Nous constatons que ce modèle ne suit pas une distribution normale. Malheureusement, à ce stade, nous ne pouvons pas améliorer davantage le modèle.

Nous concluons qu'il s'agit du meilleur modèle de prédiction du poids des poissons vendus sur le marché que nous puissions construire.

# **Deux cas études**

1. **Choix du modèle Exercice:** Taux d’accidents
2. **Afficher la modèle d'origine**

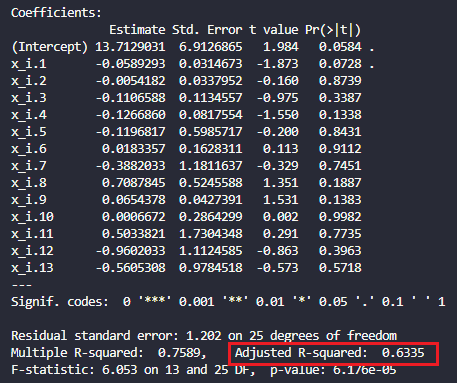
Text

Description automatically generated

Nous choisissons le modèle en fonction des données d'entrée.

- Après utiliser la commande summary(), on retrouve le modèle d'origine, les variables semblent n'avoir aucune signification.

🡪 Il peut y avoir des valeurs outliers dans les données d'entrée d'origine.



Nous pouvons voir que la valeur du R au carré ajusté : 0,6335 est encore faible.

Utilisation de boxplot() sur 13 variables (de x\_i.1 à x\_i.13).

Text

Description automatically generated

Chart, box and whisker chart

Description automatically generated

→ Il y a 17 lignes contenant des valeurs outliers sur un total de 39 observations dans les lignes 1, 2, 3, 4, 5, 6, 9, 11, 19, 21, 25, 27, 29, 31, 32, 34, 37.

→ Le nombre de lignes contenant des valeurs aberrantes représente 43,59 % du total des observations.

→ Il n'est pas possible de supprimer toutes les lignes avec des valeurs outliers, car elles représentent plus de 5 % du total des observations.

→ **Par conséquent**, nous pouvons remplacer les valeurs outliers par la moyenne ou la médiane de la colonne respective pour réduire les valeurs outliers.

*Le graphique Boxplot affiche 5 paramètres :*

* **Median**: La médiane de l'ensemble de données, c'est-à-dire la valeur de l'élément du milieu.
* **First quartile (Q1)**: Médiane entre la médiane et le plus petit élément de l'ensemble de données. Aussi connu sous le nom de 25e centile.
* **Third quartile (Q3)**: Médiane entre la médiane et le plus grand élément de l'ensemble de données. Également connu sous le nom de 75e centile.
* **Minimum**: Le plus petit élément ne fait pas exception.
* **Maximum**: Le plus grand élément ne fait pas exception.

1. **Remplacer la colonne de données en position médiane**

- Faites une copie des données d'origine (au cas où les modifications directes des données d'origine ne fonctionneraient pas → Difficile de revenir en arrière).



- Ensuite, nous remplacerons les colonnes outliers par la médiane de la colonne correspondante.

Text

Description automatically generated

- En utilisant boxplot() pour observer, nous voyons.

Chart, box and whisker chart

Description automatically generated

→ Le nombre de valeurs outliers après remplacement par la médiane est encore beaucoup (aux lignes 3, 5, 6, 9, 10, 12, 16, 20, 26) par rapport au nombre total d'observations.

→ Donc, remplacer par la médiane n'est toujours pas vraiment bon.

→ Remplacer par Q3.

1. **Remplacez les données de la colonne par Q3**

- Faire une copie des anciennes données.



- Remplacer les colonnes avec des valeurs outliers par Q3.

Text, chat or text message

Description automatically generated

- Utilisez boxplot(), pour afficher les données sur les nouvelles données.



Nous pouvons voir qu'en remplaçant les colonnes de valeurs outliers par la valeur de Q3, les données n'ont pas de valeurs outliers.

→ On voit que, lorsque nous le remplaçons par Q3, le nombre de valeurs outliers est bien inférieur.

Chart, box and whisker chart

Description automatically generated

- Après traitement des valeurs outliers, nous choisirons un nouveau modèle basé sur les nouvelles données.

1. **Utilisez AIC pour choisir le meilleur modèle (modele\_2)**

- Après traitement des valeu xrs outliers, nous choisirons un nouveau modèle basé sur les nouvelles données.

Text

Description automatically generated

A picture containing text, plaque

Description automatically generated

- Après avoir traité les valeurs outliers, nous choisirons un nouveau modèle basé sur de nouvelles données en utilisant la méthode AIC.

Text

Description automatically generated

Text

Description automatically generated with low confidence

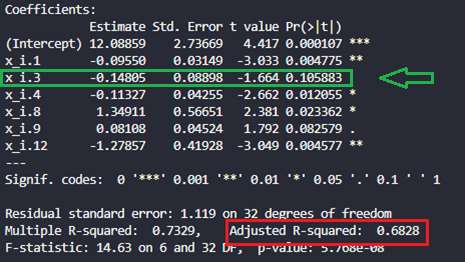
🡺 Modele Choisi: **y\_i = x\_i.1 + x\_i.3 + x\_i.4 + x\_i.8 + x\_i.9 + x\_i.12**

🡺 Équation linéaire: y\_i = 12.08859 - 0.09550(x\_i.1) - 0.14805(x\_i.3) - 0.11327(x\_i.4) + 1.34911(x\_i.8) + 0.08108(x\_i.9) - 1.27857(x\_i.12).

- Nous choisissons le meilleur modèle :

A picture containing text

Description automatically generated

****

**🡪** Avec R-carré ajusté : 0,6828 de ce nouveau modèle est meilleur que le modèle original (0,6828 > 0,6335).

**🡪** Cependant, après sélection du modèle par AIC, nous prédisons que la variable x\_i.3 n'est pas significative dans le modèle.

1. **Utilisez BIC pour choisir le meilleur modèle (modele\_3)**

Text

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

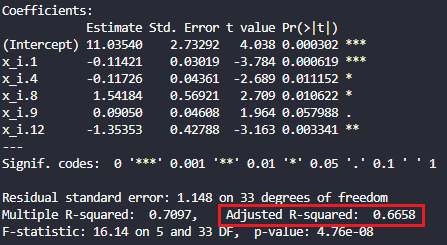
🡺 Modele Choisi: **y\_i = x\_i.1 + x\_i.4 + x\_i.8 + x\_i.9 + x\_i.12**

🡺 Équation linéaire: y\_i = 11.0354 - 0.1142(**x\_i.1**) - 0.1173(**x\_i.4** ) + 1.5418(**x\_i.8**) + 0.0905(**x\_i.9**) - 1.3535(**x\_i.12**)

- Nous choisissons le meilleur modèle :

A picture containing graphical user interface

Description automatically generated

****

Avec R-carré ajusté : 0, 6658 de ce nouveau modèle est meilleur que le modèle original (0,6658 > 0,6335).

1. **Calculer les résidus et vérifiez si le nouveau modèle suit une distribution normale**

Text

Description automatically generated

Text

Description automatically generated

- Nous calculons les résidus puis utilisons shapero.test() pour vérifier si le modèle suit une distribution normale.

- On obtient p-value : 0.6972 → Suit une distribution normale.

1. **ANOVA à deux facteurs Exercice:** Agents toxiques

Modèle linéaire général : Durée en fonction de Poison; Traitement .

Facteur type niveaux Valeurs.

Poison fixe 3 I; II; III.

Traitement fixe 4 A; B; C; D.

Visualisation de données.

Text

Description automatically generated

Chart, line chart

Description automatically generated

L'axe Y montre les temps de survie de quatre traitements A, B, C et D.

L'axe X montre les classes de poison de quatre traitements A, B, C et D.

On peut voir ça:

* Le traitement B a le plus de temps, puis le traitement D, puis le traitement C et enfin le traitement A.
* Le temps de survie du poison P3 est le plus court, suivi du poison P2 et enfin du poison P1.

### **Shapiro-Wilk normalité test**

Graphical user interface, text

Description automatically generated

Nous pouvons voir le p-valu = 0.6507 > 0,05.

Par conséquent, nous n'avons pas suffisamment de raisons de rejeter l'hypothèse H0 ou que la variable aov\_residuals obéit à la distribution normale.

Text

Description automatically generated

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Source de variation** | **SC** | **ddl** | **MC** | **Fos** | **P-value** |
| Poison | 1,0208 | 2 | 0,5104 | 34,32 | 0,000 |
| Traitement | 0,7448 | 3 | 0,2483 | 16,69 | 0,000 |
| Interaction Poison\*Traitement | 0,1801 | 6 | 0,0300 | 2,02 | 0.08 |
| Intérieur | 0,5354 | 36 | 0,0149 |  |  |
| Total | 2,4811 | 47 |  |  |  |

### **Test Fisher hypothesis avec a = 5%:**

1. **Type de Poison**

# H0: il N'a pas effect "type de Poison ".

# H1: il y a effect "type de Poison ".

P-valu ~= 0.000 < a = 0.05 🡪 rejeter H0 avec le seuil a=0.05

* il y a effect "type de Poison " dans la modele.

1. **Type de Traitement**

# H0: il n'a pas effect "type de Traitement "

# H1: il y a effect "type de Traitement "

P-valu ~= 0.000 < a = 0.05 🡪 rejeter H0 avec le seuil a=0.05

* il y a effect "type de Traitement " dans la modele.

1. **Type de Poison \* Traitement**

# H0: il N'a pas effect "type de Poison \* Traitement "

# H1: il y a effect "type de Poison \* Traitement "

P-valu = 0.08 < a = 0.05 🡪 pas assez de base pour rejeter H0 avec le seuil a=0.05

* il y n’a pas effect "type de Poison \* Traitement " dans la modele.
* il n'y a pas d'interaction entre les 2 Poison et Traitement.

Text

Description automatically generated

#H0: La variance entre chaque groupe est égale.

#H1: Au moins un groupe a une variance qui n'est pas égale aux autres.

P-valu = 0.2711 > a = 0.05 🡪 pas assez de base pour rejeter H0 avec le seuil a=0.05

* La variance entre chaque groupe est égale.

Text

Description automatically generated

#H0 : les variances entre les échantillons sont égales.

#H1 : au moins un échantillon a une variance différente.

Levene’s Test identique au test de Bartlett:

P-valu = 0.06283 > a = 0.05

* Nous ne pouvons pas rejeter H0 selon laquelle la variance est égale dans tous les échantillons au seuil de signification de 0,05.
* Les variances entre les échantillons sont égales.

### **TurkeyHSD Test**

Text

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Intervalles de confiance simultanés de Tukey = 95,0 %

On a p-adj P1-P2 = 0.04 < a = 0.05 🡪 Il existe une différence entre les poisons P1 et P2 et est statistiquement significative.

Semblable à p-adj P1-P3 = 0.0000 < a = 0.05 🡪 Il existe une différence entre les poisons P1 et P3 et est statistiquement significative.

Semblable à p-adj P2-P3 = 0.0000 < a = 0.05 🡪 Il existe une différence entre les poisons P2 et P3 et est statistiquement significative.

Text

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Intervalles de confiance simultanés de Tukey = 95,0 %.

On a p-adj B-A = 0.0000 < a = 0.05 🡪 Il y a une différence entre le traitement B et le traitement A et elle est statistiquement significative.

Semblable à p-adj D-A = 0.003 < a = 0.05 🡪 Il y a une différence entre le traitement D et le traitement A et elle est statistiquement significative.

Semblable à p-adj C-B = 0.000 < a = 0.05 🡪 Il y a une différence entre le traitement C et le traitement B et elle est statistiquement significative.

On a p-adj C-A = 0.46 > a = 0.05 🡪 Il n'y a pas de différence statistiquement significative entre le traitement C et le traitement A.

Semblable à p-adj D-B = 0.08 > a = 0.05 🡪 Il n'y a pas de différence statistiquement significative entre le traitement D et le traitement B.

Semblable à p-adj D-C = 0.12 > a = 0.05 🡪 Il n'y a pas de différence statistiquement significative entre le traitement D et le traitement C.

# **Les références**

## **Consulter le guide de traitement des données et la théorie**

* <https://dzchilds.github.io/stats-for-bio/two-way-anova-in-r.html?fbclid=IwAR0VPgVpS4ZCF1IwKDS55SXAQM4VgWnNSVxxWldgwrO-cPNPH_7n4MCKY0w>

## **Les sources de données**

* <https://www.kaggle.com/datasets/aungpyaeap/fish-market?resource=download>
* <https://www.kaggle.com/> (Vous pouvez rechercher de nombreux ensembles de données sur ce site. C'est le site que de nombreux analystes de données utilisent pour pratiquer)