1. Statistique Descriptive

**Données babies**

Le but de cet exercice est d’effectuer des études sur des données des babies, qui contiennent 1236 individuels et 23 variables. Dans notre problème, nous étudions 8 variables : le poids à la naissance, la durée de gestation, le nombre de grossesses précédentes, la taille de la mère et le poids de la mère, l’âge de la mère, si la mère fume ou non et le niveau d’éducation de la mère.

**Question 01** **Relation entre du poids du Nouveau Né et la mère qui fume pendant la grossesse ou qui ne fume pas.**

**A. Résumé numérique**

Nous avons observé le résumé numérique du poids de nouveau né par rapport au tabagisme des mères.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Min | 1st.Qu. | Médian | Moyenne | 3rd.Qu | Max |
| 55.0 | 113.0 | 123.0 | 123.0 | 134.0 | 176.0 |

Résultat des valeurs du poids pour bébés nés d’une mère non fumeuse.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Min | 1st.Qu. | Médian | Moyenne | 3rd.Qu | Max |
| 58.0 | 102.0 | 115.0 | 114.1 | 126.0 | 163.0 |

Résultat des valeurs du poids pour bébés nés d’une mère fumeuse

D’après ce résumé, nous remarquons que les valeurs du poids du nouveau né d’une mère non fumeuse est généralement supérieures que celles d’une mère fumeuse.

**B. Boxplot**

Ensuite, nous désignons un Boxplot pour voir plus en détail.



Ce Boxplot confirme parfaitement notre conclusion. Nous trouvons que la majorité des bébés nés d’une mère non fumeuse ont du poids entre 110 et 130, et celles de mère fumeuse ont du poids entre 100 et 120

**Question 02 Relation entre du temps de gestation et la mère qui fume pendant la grossesse ou qui ne fume pas.**

**A. Résumé numérique**

D’abord, nous avons également fait le résumé numérique du temps de gestation par rapport au tabagisme des mères.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Min | 1st.Qu. | Médian | Moyenne | 3rd.Qu | Max |
| 148 .0 | 273.0 | 281.0 | 282.2 | 289.0 | 353.0 |

Résultat des valeurs du temps de gestation pour bébés nés d’une mère non fumeuse.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Min | 1st.Qu. | Médian | Moyenne | 3rd.Qu | Max |
| 223 .0 | 271.0 | 279.0 | 278.0 | 286.0 | 330.0 |

Résultat des valeurs du temps de gestation pour bébés nés d’une mère fumeuse.

Nous comparons la médian et la moyenne de deux groupe. Il paraît qu’il n’y a une grande différence. Donc, d’après ce résumé, nous ne pouvons pas être sûr que le tabagisme à une influence évident sur du temps de gestation.

**B. Boxplot**



Ce graphique montre également que les deux groupes n’ont pas de différences claire. Par conséquent, nous ne pouvons pas dire si le tabagisme influence du temps de gestation.

**Question 03 Relation entre d’étude et le tabagisme.**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | 0 | 1 | 2 | 3 | 4 | 5 | 7 |
| fumeuse | 4 | 102 | 176 | 33 | 102 | 65 | 1 |
| non fumeuse | 1 | 79 | 264 | 30 | 194 | 154 | 6 |

D’abord, c’est un tableau de niveau d’étude des mère fumeuse ou non fumeuse.

D’après ce tableau, nous constatons des différences claires entre différents niveaux d’étude. Parmi ces valeurs, les groupes de niveau 2,4 et 5 ont des différences plus importantes. D’ailleurs, il n’existe pas des mères qui sont dans niveau 6.



Ce graphe nous montre plus clairement que deux groupe différencient plus au niveau d’étude 2 ,4 ,5.

2. Analyse ACP

**2.2 Utilisation des outils R**

**Question 01 En utilisant ces fonctions, effectuer l’ACP du jeu de données notes étudiées en cours.**

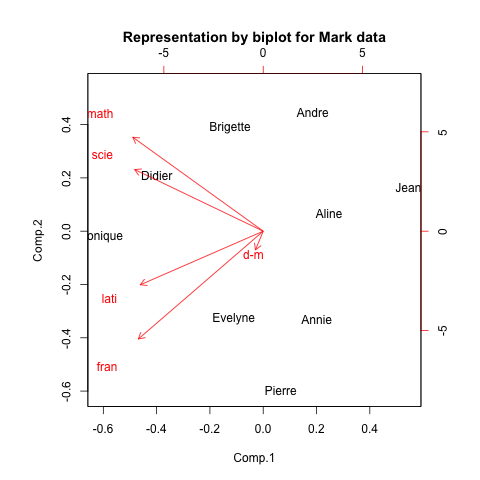
Dans cet exercice, nous utilisons la fonction R, *princomp* pour mise en œuvre l’ACP en formule :

Valeurs propres:

Vecteurs propres :

Composantes :

Et puis, nous avons désigné le graphe de la représentation par premières deux composantes principales.



D’après ce graphe, nous avons constatons que sauf la variable ‘*d-m*’, les autres sont assez bien représentés.

**Question02 La fonctionnement de plot et biplot**

D’abord, ‘*plot’* est une fonction générale pour designer des graphes. Par conséquent, il peut designer des graphes de représentation pour deux composantes

D’ailleurs, *‘biplot’* nous permet de mettre les individus et les variables dans le même graphe.

Ensuite, *‘biplot.princomp’* est une fonction plus avancé et spécialisé pour l’ACP. Sauf la représentation de deux composantes, il peut également évaluer la corrélation entre chaque variable et sa composante, qui nous permet d’évaluer la qualité de représentation.

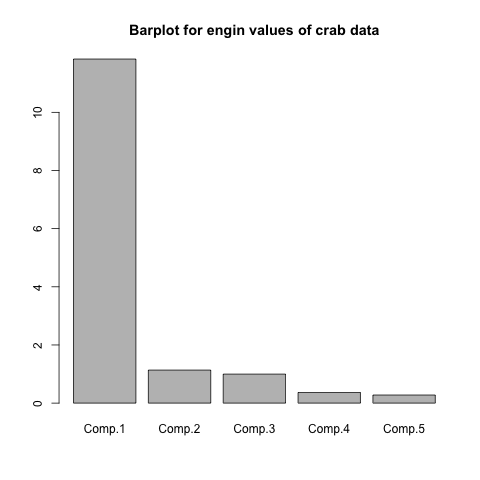
Nous avons aussi étudié les options de *‘biplot.princomp’.* L’option *‘choices’* se sert de définir la taille de vecteurs et *‘scale’* est pour obtenir une représentation standard avec des données standard (*scaled*). En finale, l’option *‘pc.plot’* se sert de désigner un plot avec des observations standardisé par la racine de N.

**2.3 Traitement des données Crabs**

**Question01 Tester tout d’abord l’ACP sur *crabsquant* sans traitement préalable.**

Nous avons d’abord effectué l’ACP sans traitement :

D’abord, nous regardons les valeurs propres (Inertie d’expliquée pour chaque axe)

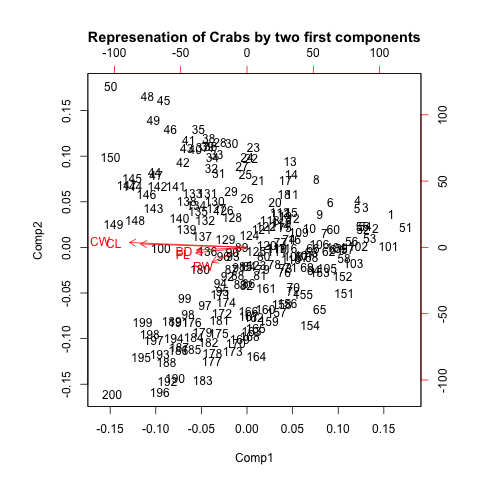


Nous remarquons que la première composante a un pourcentage très important par rapport les autre composantes.

Voici le tableau de pourcentages cumulés :

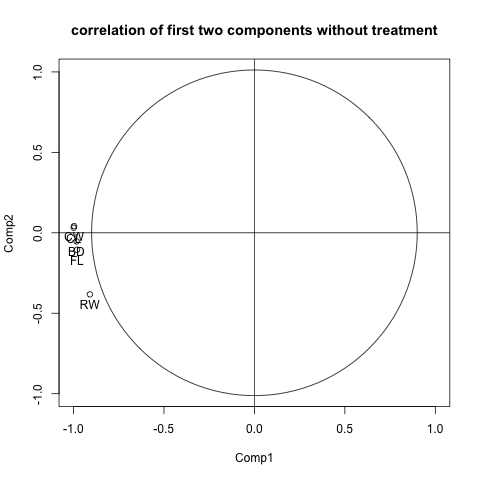
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 1 | 2 | 3 | 4 | 5 |
| 80.9% | 88.7% | 95.4% | 98.0% | 100% |

Ensuite, nous étudions la représentation par biplot.



D’après ce graphe, les angles entre les variables sont très petits. Autrement dit, toutes les variables sont fortement corrélées.

Egalement, nous regardons le graphe de corrélation :



Nous constatons que les variables sont assez bien représentées au niveau du pourcentage de l’inertie. Par contre, d’après le graphe de corrélation, les variables sont corrélés avec le premier axe et mal corrélés avec le deuxième axe.

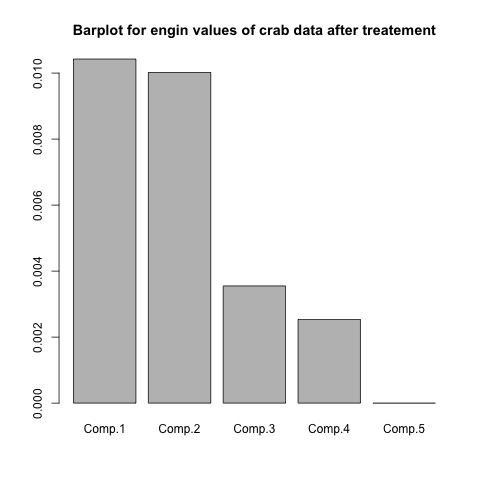
Les variables sont trop corrélées à cause de l’effet taille.

**Question02 Trouver une solution pour améliorer la qualité de votre représentation**

**Solution01**

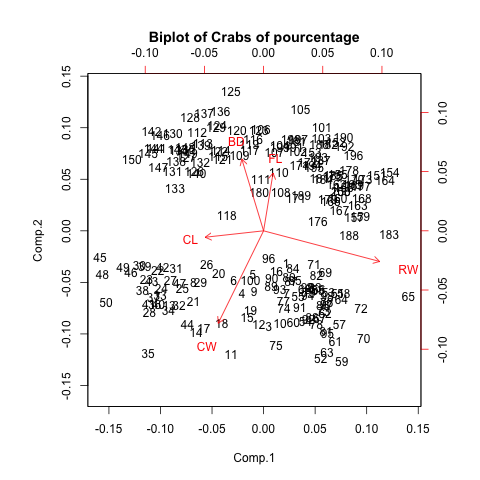
Nous avons fait la même manipulation comme l ‘exercice précédant, c’est que nous divisons chaque variable par le somme de toutes des variables et obtenons des pourcentage. Comme cela, nous éliminons l’effet de taille.

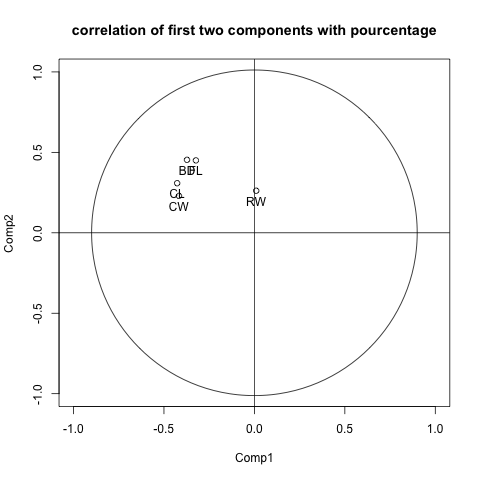
Nous avons effectué également la barplot pour des inerties expliquées



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 1 | 2 | 3 | 4 | 5 |
| 39.2% | 77.0% | 90.4% | 99.9% | 100% |

Et puis, nous avons effectué le biplot :





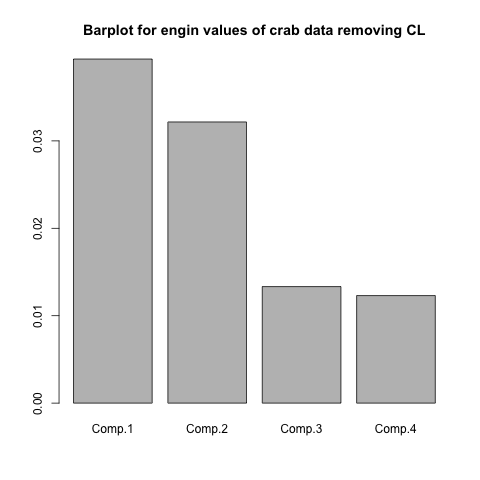
Nous trouvons que les données sont beaucoup moins corrélées maintenant. Et puis, nous trouvons que la qualité de représentation est pire que celle sans traitement au niveau de pourcentage de l’inertie. En revanche, sauf RW, les autres variables ont mieux corrélés selon le graphe de corrélation.

D’ailleurs, selon le biplot, nous pouvons constatons généralement une séparations des données. (Il y a généralement 4 groupes pour des crabs de deux espèces et deux sexes).

**Solution02**

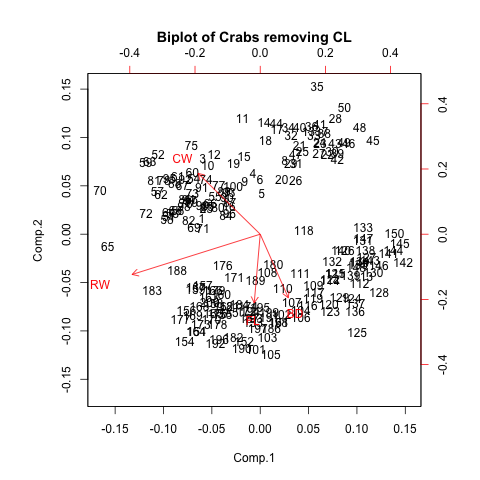
Il y a une autre méthode pour éliminer l ‘effet taille. C’est que nous divisons chaque variable par le troisième variable *‘CL’*. Et puis, nous enlevons cette variable. Comme cela, nous pouvons aussi éliminé l’effet taille.

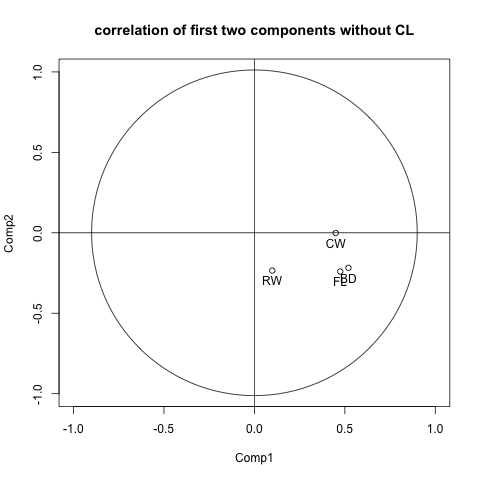
D’abord, nous avons effectué le barplot pour inerties expliquées :



|  |  |  |  |
| --- | --- | --- | --- |
| 1 | 2 | 3 | 4 |
| 40.2% | 73.5% | 87.2% | 100% |

Ensuite, nous avons désigné le biplot et graphe de corrélation:





Le résultat ressemble à première solution.

C’est que la qualité de représentation est pire que celle sans traitement au niveau du pourcentage d’inertie et elle se présente mieux au niveau du graphe de corrélation.

Et puis, nous pouvons constaté en gros une séparation des groupes d’après le graphe de biplot.

**Conclusion**