**CHAPITRE 1**

Exercice 13

Dans cette exercices nous utilisons les donnée « Poids Naissance ».

tableau 1 : Extrait des données Poids Naissance

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ID | AGE | LWT | RACE | SMOKE | PTL | HT | UI | FVT | BWT | LOW |
| 85 | 19 | 182 | 2 | 0 | 0 | 0 | 1 | 0 | 2523 | 0 |
| 86 | 33 | 155 | 3 | 0 | 0 | 0 | 0 | 3 | 2551 | 0 |
| 87 | 20 | 105 | 1 | 1 | 0 | 0 | 0 | 1 | 2557 | 0 |
| 88 | 21 | 108 | 1 | 1 | 0 | 0 | 1 | 2 | 2594 | 0 |
| 89 | 18 | 107 | 1 | 1 | 0 | 0 | 1 | 0 | 2600 | 0 |
| 91 | 21 | 124 | 3 | 0 | 0 | 0 | 0 | 0 | 2622 | 0 |

Le tableau 1 est un court extrait du jeu de données. Ici la variable LWT qui correspond au poids de la mère est exprimé en livres, nous la modifions donc pour l’avoir en kilogrammes.

tableau 2 : Extrait des données avec chagement d’unité du poids de la mère

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ID | AGE | LWT | RACE | SMOKE | PTL | HT | UI | FVT | BWT | LOW |
| 85 | 19 | 82.554 | 2 | 0 | 0 | 0 | 1 | 0 | 2523 | 0 |
| 86 | 33 | 70.307 | 3 | 0 | 0 | 0 | 0 | 3 | 2551 | 0 |
| 87 | 20 | 47.627 | 1 | 1 | 0 | 0 | 0 | 1 | 2557 | 0 |
| 88 | 21 | 48.988 | 1 | 1 | 0 | 0 | 1 | 2 | 2594 | 0 |
| 89 | 18 | 48.534 | 1 | 1 | 0 | 0 | 1 | 0 | 2600 | 0 |
| 91 | 21 | 56.245 | 3 | 0 | 0 | 0 | 0 | 0 | 2622 | 0 |

Nous obtenons le jeu de données du tableau 2.

Réalisons maintenant quelque tri à plat avec ces données.

tableau 3 : Tri à plat de l’age de la mère

|  |  |
| --- | --- |
| classes | Effectif |
| (14,20.2] | 69 |
| (20.2,26.4] | 74 |
| (26.4,32.6] | 37 |
| (32.6,38.8] | 8 |
| (38.8,45] | 1 |

Xxxx

tableau 4 : Tri à plat du poids de la mère

|  |  |
| --- | --- |
| classes | Effectif |
| (79.8,114] | 60 |
| (114,148] | 88 |
| (148,182] | 26 |
| (182,216] | 11 |
| (216,250] | 4 |

Xxxx

tableau 5 : Tri à plat de la race

|  |  |
| --- | --- |
|  | Effectif |
| Blanche | 96 |
| Noir | 26 |
| Autre | 67 |

Xxxx

tableau 6 : Tri à plat du tabagisme durant la grossesse

|  |  |
| --- | --- |
|  | Effectif |
| Non | 115 |
| Oui | 74 |

xxxx

Exercice 14

Nous allons crée un jeu de donnée personnel, pour pouvoir le manipuler.

tableau 7 : Jeu de données acteur

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Mort.a | Années.de.carrière | Nombre.de.films | Prénom | Nom | Date.du.deces |
| 93 | 66 | 211 | Michel | Galabru | 04-01-2016 |
| 53 | 25 | 58 | André | Raimbourg | 23-09-1970 |
| 72 | 48 | 98 | Jean | Gabin | 15-10-1976 |
| 68 | 37 | 140 | Louis | De Funès | 27-01-1983 |
| 68 | 31 | 74 | Lino | Ventura | 22-10-1987 |
| 53 | 32 | 81 | Jacques | Villeret | 28-01-2005 |

Le tableau 7 nous illustre le jeu de données.

tableau 8 : Prénom du jeu de données acteur

|  |
| --- |
| Prénom |
| Michel |
| André |
| Jean |
| Louis |
| Lino |
| Jacques |

Nous pouvons aussi extraire une colonne en particulier, dans le tableau 9 c’est « prénom » qui est extrait.

tableau 9: Donnée acteur trié par l’age du décès

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Age.du.décès | Années.de.carrière | Nombre.de.films | Prénom | Nom | Date.du.deces |
| 2 | 53 | 25 | 58 | André | Raimbourg | 23-09-1970 |
| 6 | 53 | 32 | 81 | Jacques | Villeret | 28-01-2005 |
| 4 | 68 | 37 | 140 | Louis | De Funès | 27-01-1983 |
| 5 | 68 | 31 | 74 | Lino | Ventura | 22-10-1987 |
| 3 | 72 | 48 | 98 | Jean | Gabin | 15-10-1976 |
| 1 | 93 | 66 | 211 | Michel | Galabru | 04-01-2016 |

Nous modifions ensuite le nom de la colonne « Mort.à » , en « Age.du.décès », comme le montre le tableau  9. Et pour finir, nous pouvons ordonner le jeu de donnée selon une conditions. Ici nous voulons ordonner par « Age.du.décès » croissant, ce qui est fait dans le tableau 9 .

Exercice 15

Dans cette exercice nous utiliserons le jeu de données « fromages ».

tableau 10 : Extrait du jeu de donnée fromage

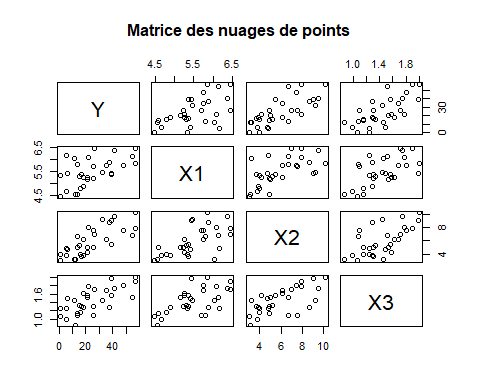
|  |  |  |  |
| --- | --- | --- | --- |
| Y | X1 | X2 | X3 |
| 12.3 | 4.543 | 3.135 | 0.86 |
| 20.9 | 5.159 | 5.043 | 1.53 |
| 39.0 | 5.366 | 5.438 | 1.57 |
| 47.9 | 5.759 | 7.496 | 1.81 |
| 5.6 | 4.663 | 3.807 | 0.99 |
| 25.9 | 5.697 | 7.601 | 1.09 |

Voici un extrait des données représenté dans le tableau 10 .

On retrouve un total de 30 individus qui corresponde à des types de fromage. Il y a 4 variables toute quantitatives. Il y a X1, X2, X3, et Y qui est le score moyen attribué par des goûteurs.

tableau 11 : Statistiques élémentaires des données fromage

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Y | X1 | X2 | X3 |
|  | Min. : 0.70 | Min. :4.477 | Min. : 2.996 | Min. :0.860 |
|  | 1st Qu.:13.55 | 1st Qu.:5.237 | 1st Qu.: 3.978 | 1st Qu.:1.250 |
|  | Median :20.95 | Median :5.425 | Median : 5.329 | Median :1.450 |
|  | Mean :24.53 | Mean :5.498 | Mean : 5.942 | Mean :1.442 |
|  | 3rd Qu.:36.70 | 3rd Qu.:5.883 | 3rd Qu.: 7.575 | 3rd Qu.:1.667 |
|  | Max. :57.20 | Max. :6.458 | Max. :10.199 | Max. :2.010 |

Le tableau 11 nous montre les statistiques élémentaire pour chacune des variables. Par exemple pour Y on trouve une valeur moyenne de 24,53, un minimum de 0.7 et un maximum de 57.20.

*Figure 1 : Matrice des nuages de points*

La figure 1 représente la Matrice de nuage de point entre chacune des variables. Ce sont les nuages de points des croisements deux à deux entre chaque variables de nos données.

tableau 12 : Extrait des données fromage filtré

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Y | X1 | X2 | X3 |
| 2 | 20.9 | 5.159 | 5.043 | 1.53 |
| 3 | 39.0 | 5.366 | 5.438 | 1.57 |
| 6 | 25.9 | 5.697 | 7.601 | 1.09 |
| 7 | 37.3 | 5.892 | 8.726 | 1.29 |
| 10 | 21.0 | 5.242 | 4.174 | 1.58 |
| 11 | 34.9 | 5.740 | 6.142 | 1.68 |

Nous allons maintenant crée avec sous jeu de données avec les contraintes suivante X1 > 5.1 et X3 < 1.77. C’est ce qui est représenté dans le tableau 12 .

tableau 13 : Statistiques élémentaires des données fromage filtré

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Y | X1 | X2 | X3 |
|  | Min. : 0.70 | Min. :5.159 | Min. :3.219 | Min. :1.080 |
|  | 1st Qu.:14.30 | 1st Qu.:5.313 | 1st Qu.:4.744 | 1st Qu.:1.295 |
|  | Median :21.00 | Median :5.455 | Median :5.438 | Median :1.460 |
|  | Mean :23.52 | Mean :5.654 | Mean :5.946 | Mean :1.435 |
|  | 3rd Qu.:33.45 | 3rd Qu.:5.968 | 3rd Qu.:6.857 | 3rd Qu.:1.575 |
|  | Max. :54.90 | Max. :6.458 | Max. :9.588 | Max. :1.740 |

Apres ce changement on trouve certaine valeur différentes des statistiques élémentaire, par exemple la moyenne de Y est maintenant de 23,52. Nous voyons les nouvelles statistiques dans le tableau .. .

Exercice 16

Les données que utiliserons sont directement implanté dan R, il s’agit des données « airquality ».

tableau 14 : Extrait des données airquality

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Ozone | Solar.R | Wind | Temp | Month | Day |
| 41 | 190 | 7.4 | 67 | 5 | 1 |
| 36 | 118 | 8.0 | 72 | 5 | 2 |
| 12 | 149 | 12.6 | 74 | 5 | 3 |
| 18 | 313 | 11.5 | 62 | 5 | 4 |
| NA | NA | 14.3 | 56 | 5 | 5 |
| 28 | NA | 14.9 | 66 | 5 | 6 |

Le tableau 14 nous montre un extrait des données.

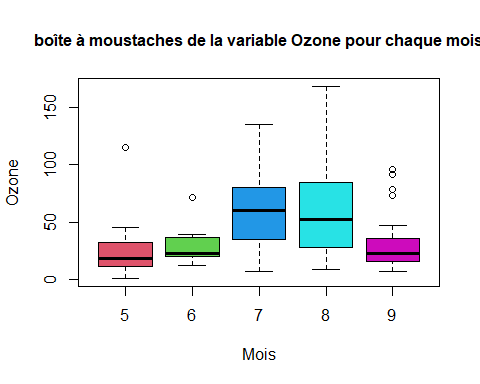
Il s’agit des relevés quotidiens des valeurs de qualité de l'air suivantes du 1er mai 1973 au 30 septembre 1973. Il y a 153 individus pour 6 variables.

* Ozone : Ozone moyen en parties par milliard de 1300 à 1500 heures à Roosevelt Island.
* R. solaire : rayonnement solaire à Langley dans la bande de fréquence 4000-7700 Angstroms de 0800 à 1200 heures à Central Park.
* Wind : Vitesse moyenne du vent en miles par heure à 0700 et 1000 heures à l'aéroport de LaGuardia.
* Temp : Température maximale quotidienne en degrés Fahrenheit à l'aéroport de La Guardia.
* Month : Le mois
* Day : Le jour

Les données ont été obtenues auprès du New York State Department of Conservation (données sur l'ozone) et du National Weather Service (données météorologiques).

tableau 15 : Statistique élémentaire des données airquality

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Ozone | Solar.R | Wind | Temp |
|  | Min. : 1.00 | Min. : 7.0 | Min. : 1.700 | Min. :56.00 |
|  | 1st Qu.: 18.00 | 1st Qu.:115.8 | 1st Qu.: 7.400 | 1st Qu.:72.00 |
|  | Median : 31.50 | Median :205.0 | Median : 9.700 | Median :79.00 |
|  | Mean : 42.13 | Mean :185.9 | Mean : 9.958 | Mean :77.88 |
|  | 3rd Qu.: 63.25 | 3rd Qu.:258.8 | 3rd Qu.:11.500 | 3rd Qu.:85.00 |
|  | Max. :168.00 | Max. :334.0 | Max. :20.700 | Max. :97.00 |
|  | NA’s :37 | NA’s :7 | NA | NA |

Le tableau 15 nous montre les statistiques élémentaires sur nos variables quantitatives, ainsi que les valeur manquante. Pour la variable ozone on remarque 37 valeurs manquante et une moyenne de 42.13.

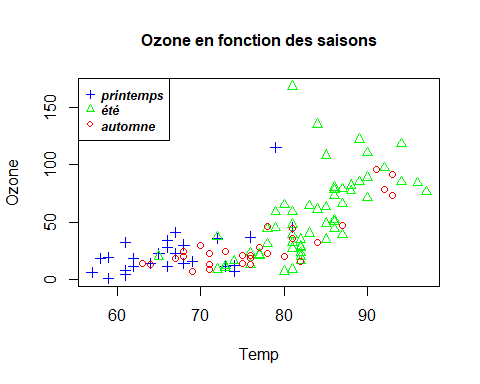
*Figure 2 : boîte à moustaches de l’Ozone pour chaque mois*

On remarque grâce à la figure 2 des diagramme à moustache avec une tendance similaire pour les mois 5, 6, et 9, avec des valeur de l’ozone peu élevé qui varie mois. Alors que pour les mois 7 et 8, la valeur de l’ozone sont plus forte et beaucoup plus répartie.

tabelau 16 : Extrait des données airquality avec la saison

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Ozone | Solar.R | Wind | Temp | Month | Day | saison |
| 41 | 190 | 7.4 | 67 | 5 | 1 | printemps |
| 36 | 118 | 8.0 | 72 | 5 | 2 | printemps |
| 12 | 149 | 12.6 | 74 | 5 | 3 | printemps |
| 18 | 313 | 11.5 | 62 | 5 | 4 | printemps |
| NA | NA | 14.3 | 56 | 5 | 5 | printemps |
| 28 | NA | 14.9 | 66 | 5 | 6 | printemps |

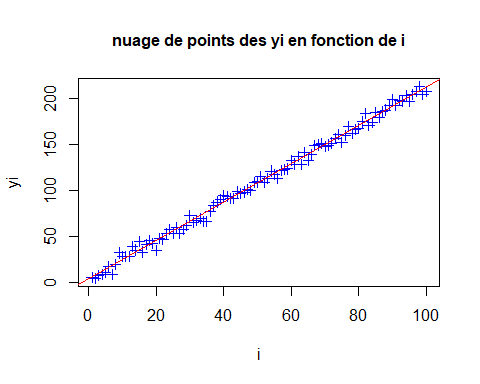
Pour notre analyse, nous rajoutons une variable saison. Le tableau 16 montre nos nouvelle données.



*Figure 3 : Ozone en fonction des saisons*

Avec la figure 3 on remarque qu’il y a une relation positive linéaire croissante entre chaque l’ozone et le temps. Cette relation est présente pour chacune des saisons. Le temps est plus élevé en été et plus faible en hiver. Et comme vu précédemment avec la figure .., la concentration d’Ozone est plus forte en été, qui correspond aux mois 7 et 8.

Exercice 17

Nous nous intéressons à la fonctions suivant yi = 1.7+2.1 i + ei , i entre 1 et 100 et les ei suivant une lois N(0,5²).

*Figure 4 : Nuage de points des yi en fonction de i*

La figure 4 nous montre le nuage de points généré avec notre fonction, avec la droite de régression. Cette droite semble être une bon ajustement de notre fonction