

Cahier thématique

Pratique d’Exploration Données R

**Manipulation 1** : Description des données

**Manipulation 2** : Exploration des variables individuelles  
**Manipulation 3** : Exploration des variables multiples  
**Manipulation 4** : Visualisation des données

**Préparé par :** Nesrine Zemirli  
  
© Nesrine Zemirli 2015-2016

Ce document ne peut être utilisé dans le cadre d’une formation, publication papier, site internet ou tout support sans mon accord express. Aucune reproduction, même partielle, ne peut être faite de ce document et de l'ensemble de son contenu : textes, images, etc. sans mon autorisation express. Pour toutes informations, communiquer avec moi sur [nesrine.zemirli@bdeb.qc.a](mailto:nesrine.zemirli@bdeb.qc.a)

|  |  |  |
| --- | --- | --- |
| **Date** | **Version** | **Changement** |
| 31 Mai 2016 | 1.0 | Version initiale |
|  |  |  |
|  |  |  |

# Mise en contexte et présentation du projet

# Format : individuel

# Jeux de données : on a à notre disposition une source de données disponible sur R

## Préliminaire

* R-studio est disponible.
* Notebook sur Jupyter

# Description : Jeux de données : Iris (Fisher, 1936)



Iris est un ensemble (jeu) de données introduit en 1936 par Ronald Aylmer Fisher comme un exemple d'analyse discriminante. Cet ensemble contient 150 exemples de critères observées sur 3 espèces différentes d'iris de Gaspésie (Setosa, Versicolor, Verginica).

Chaque exemple est composé de quatre attributs :

1. Longueur des sépales = sepal length in cm

2. Largeur des sépales = sepal width in cm

3. Longueur pétales = petal length in cm

4. Largeur des pétales = petal width in cm

5. Une classe (l'espèce).

-- Iris Setosa

-- Iris Versicolour

-- Iris Virginica

## Manipulation 1 : Description des données

### 1- Vérifiez la dimension du dataframe iris

|  |  |
| --- | --- |
| dim(iris)  En sortie sur la console, on devrait avoir un résultat similaire au suivant :  [1] 150   5  2- Les noms de variables ou des colonnes names(iris)  En sortie sur la console, on devrait avoir un résultat similaire au suivant :  [1] "Sepal.Length" "Sepal.Width"  "Petal.Length" "Petal.Width"  "Species"  3- Structure  str(iris)  En sortie sur la console, on devrait avoir un résultat similaire au suivant :  'data.frame':   150 obs. of  5 variables:  $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...  $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...  $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...  $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...  $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...  4- Attribues  attributes(iris)  En sortie sur la console, on devrait avoir un résultat similaire au suivant :  $names [1] "Sepal.Length" "Sepal.Width"  "Petal.Length" "Petal.Width"  "Species"       $row.names   [1]   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20  [21]  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39  40  [41]  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58  59  60  [61]  61  62  63  64  65  66  67  68  69  70  71  72  73  74  75  76  77  78  79  80  [81]  81  82  83  84  85  86  87  88  89  90  91  92  93  94  95  96  97  98  99 100 [101] 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 [121] 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 [141] 141 142 143 144 145 146 147 148 149 150  $class [1] "data.frame" 5- Afficher les données   # les 5 premiers lignes  iris[1:5,]  En sortie sur la console, on devrait avoir un résultat similaire au suivant :    Sepal.Length Sepal.Width Petal.Length Petal.Width Species 1          5.1         3.5          1.4         0.2  setosa 2          4.9         3.0          1.4         0.2  setosa 3          4.7         3.2          1.3         0.2  setosa 4          4.6         3.1          1.5         0.2  setosa 5          5.0         3.6          1.4         0.2  setosa  # Les premières lignes  head(iris)  En sortie sur la console, on devrait avoir un résultat similaire au suivant :  Sepal.Length Sepal.Width Petal.Length Petal.Width Species  1          5.1         3.5          1.4         0.2  setosa 2          4.9         3.0          1.4         0.2  setosa 3          4.7         3.2          1.3         0.2  setosa 4          4.6         3.1          1.5         0.2  setosa 5          5.0         3.6          1.4         0.2  setosa  6 5.4 3.9 1.7 0.4 setosa  # les dernier lignes  tail(iris)  En sortie sur la console, on devrait avoir un résultat similaire au suivant :  Sepal.Length Sepal.Width Petal.Length Petal.Width Species  145 6.7 3.3 5.7 2.5 virginica  146 6.7 3.0 5.2 2.3 virginica  147 6.3 2.5 5.0 1.9 virginica  148 6.5 3.0 5.2 2.0 virginica  149 6.2 3.4 5.4 2.3 virginica  150 5.9 3.0 5.1 1.8 virginica  Nous pouvons également récupérer les valeurs d'une seule colonne.  Par exemple, les 10 premières valeurs de Sepal.Length peut être récupérée avec l'un des codes ci-dessous.  En sortie sur la console, on devrait avoir un résultat similaire au suivant :  # obtenir les 10 premières lignes la variables Sepal.Length  iris[1:10, "Sepal.Length"]  En sortie sur la console, on devrait avoir un résultat similaire au suivant :   [1] 5.1 4.9 4.7 4.6 5.0 5.4 4.6 5.0 4.4 4.9  une autre manière de faire  iris$Sepal.Length[1:10]  En sortie sur la console, on devrait avoir un résultat similaire au suivant :   [1] 5.1 4.9 4.7 4.6 5.0 5.4 4.6 5.0 4.4 4.9 Manipulation 2 : Exploration des variables individuelles 1-La distribution de toutes les variables # Explorez des variables individuelles  summary(iris)    En sortie sur la console, on devrait avoir un résultat similaire au suivant :      Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species    Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa    :50    1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50    Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50    Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199                    3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800                    Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500  Ensuite, nous vérifions la variance de Sepal.Length avec var (), et vérifier également sa distribution  avec l'histogramme et la densité en utilisant des fonctions hist () et la density (). 2- La variance # variance de la variable Sepal.Length  var(iris$Sepal.Length)  En sortie sur la console, on devrait avoir un résultat similaire au suivant :  [1] 0.6856935 3- Visualiser la distribution d’une variable # visualaiser la distribution de la variable  hist(iris$Sepal.Length)    En sortie sur la console, on devrait avoir un résultat similaire au suivant :   4- La densité  # La Densité  plot(density(iris$Sepal.Length))  En sortie sur la console, on devrait avoir un résultat similaire au suivant :  http://www.rdatamining.com/_/rsrc/1421496538197/examples/exploration/density.jpg 5- La fréquence d’une variable # Frequency  table(iris$Species)  En sortie sur la console, on devrait avoir un résultat similaire au suivant :        setosa versicolor  virginica          50         50         50  6- Visualiser la fréquence avec un graphique de Camnbert # Graphique Camembert  pie(table(iris$Species))  En sortie sur la console, on devrait avoir un résultat similaire au suivant :  http://www.rdatamining.com/_/rsrc/1421496558314/examples/exploration/pie.jpg?height=400&width=400 Manipulation 3 : Explorez des variables multiples1- Visualiser l’ensemble des variables barplot(table(iris$Species))  En sortie sur la console, on devrait avoir un résultat similaire au suivant :  2- Covariance de deux variables  # Covariance de deux variables  cov(iris$Sepal.Length, iris$Petal.Length)  En sortie sur la console, on devrait avoir un résultat similaire au suivant :  [1] 1.274315 barplot(table(iris$Species)) 3- Corrélation de deux variables # Correlation deux variables  cor(iris$Sepal.Length, iris$Petal.Length)  En sortie sur la console, on devrait avoir un résultat similaire au suivant :  [1] 0.8717538 4- Covariance de plusieurs variables # Covariance de plusieurs variables  cov(iris[,1:4])  En sortie sur la console, on devrait avoir un résultat similaire au suivant :  Sepal.Length Sepal.Width Petal.Length Petal.Width  Sepal.Length 0.6856935 -0.0424340 1.2743154 0.5162707  Sepal.Width -0.0424340 0.1899794 -0.3296564 -0.1216394  Petal.Length 1.2743154 -0.3296564 3.1162779 1.2956094  Petal.Width 0.5162707 -0.1216394 1.2956094 0.5810063 5- Corrélation de plusieurs variables # Corrélation de plusieurs variables  cor(iris[,1:4])  En sortie sur la console, on devrait avoir un résultat similaire au suivant :  Sepal.Length Sepal.Width Petal.Length Petal.Width  Sepal.Length 1.0000000 -0.1175698 0.8717538 0.8179411  Sepal.Width -0.1175698 1.0000000 -0.4284401 -0.3661259  Petal.Length 0.8717538 -0.4284401 1.0000000 0.9628654  Petal.Width 0.8179411 -0.3661259 0.9628654 1.0000000 6- Statistique aggregate() On calcule les statistiques de Sepal.Length de chaque espèce avec un aggregate ().  # Statistique sur Sepal.Length avec aggregate()  aggregate(Sepal.Length ~ Species, summary, data=iris)  En sortie sur la console, on devrait avoir un résultat similaire au suivant :  Species Sepal.Length.Min. Sepal.Length.1st Qu. Sepal.Length.Median  1 setosa 4.300 4.800 5.000  2 versicolor 4.900 5.600 5.900  3 virginica 4.900 6.225 6.500  Sepal.Length.Mean Sepal.Length.3rd Qu. Sepal.Length.Max.  1 5.006 5.200 5.800  2 5.936 6.300 7.000  3 6.588 6.900 7.900 Manipulation 4 : Visualiser les donnéesDiagramme de dispersion # visulation de diagramme de dispersion  plot(iris$Sepal.Length, iris$Sepal.Width)  En sortie sur la console, on devrait avoir un résultat similaire au suivant :  http://www.rdatamining.com/_/rsrc/1421496527116/examples/exploration/scatterplot.jpg?height=400&width=400  # avec des couleurs  with(iris, plot(Sepal.Length, Sepal.Width, col=Species, pch=as.numeric(Species)))  En sortie sur la console, on devrait avoir un résultat similaire au suivant :   Visualiser tous les diagrammes   # visualiser tous les digrammes  plot(iris)  # ou  pairs(iris)  http://www.rdatamining.com/_/rsrc/1421496504431/examples/exploration/pairplot.jpg   |  | | --- | | 3- Visualisations élaborée |   Un nuage de points 3D peut être produit avec le paquet scatterplot3d  library(scatterplot3d)  scatterplot3d(iris$Petal.Width, iris$Sepal.Length, iris$Sepal.Width)   Carte de chaleur |

Carte de chaleur présente un affichage 2D d'une matrice de données, qui peut être généré avec heatmap() R. Avec le code ci-dessous, nous calculons la similitude entre les différentes fleurs dans les données de l'iris avec dist (), puis tracer avec une carte de chaleur .

> distMatrix <- as.matrix(dist(iris[,1:4]))

> heatmap(distMatrix)

