Université de Technologie de Compiegne

SY09

Data Mining

Premier Rendu

Oumaima Talouka Zineb SLAM

28 mars 2017

Résumé

Dans ce rapport du TP1 de l'UV SY09 nous allons expliquer notre démarches dans l'analyse des données en expliquant les résultats obtenus. Ce TP est compose de 2 parties. La première partie a pour objectif de se familiariser avec les méthodes de traitement et de visualisation de données sur R. La deuxième partie traite de l'Analyse en composantes principales (ACP). Nous allons travailler avec 3 dataset notes de SY02, Crabs et Pima qu'on va d'abord analyser et décrire avant d'y réaliser l'ACP en seconde partie. Pour les graphes obtenus nous avons utilises la librairie ggplot2 qui offre un grand nombre de fonctionnalité Le code R sera fourni en annexe.

Table des matières

1 Statistique descriptive **2**
1.1 Notes SY02 2
1.1.1 Description des variables 2
1.1.2 Analyse descriptive des données 3
1.2 Conclusion 5
1.3 Crabs 5
1.3.1 Description des variables 6
1.3.2 Analyse descriptive des données 6
1.3.3 Analyse de la correlation 9
1.4 Pima 9
1.4.1 Description des variables 9
1.4.2 Analyse descriptive des données 10
1.4.3 Analyse de la correlation 10

2 Analyse Composantes Principales **11**

Chapitre 1

Statistique descriptive

1.1 Notes SY02

Le dataset *"sy02-p2016"* représente les notes des étudiants de l'UTC en SY02 (statistiques) durant le printemps 2016. Nos données comptent *N= 296* individus (étudiants) er *p= 11* variables. **Enlever les etudiants desinscrits**

1.1.1 Description des variables

Variables Quantitatives	note.median, note.final, note.totale
Variables Qualitatives Nominales	nom, specialite, status, dernier.diplome.obtenu, correcteur.median,
Variables Qualitatives Ordinales	niveau, resultat

Table 1.1 – Categorie de Variables

- **Nom :** chaine de caractère des identifiants de chaque étudiant
- **Spécialité :** la branche de l'etudiant : GB, GM, GSM, GP, GI
- **Niveau** : Semestre de l'etudiant de 1 a 6
- **Statut** : Soit l'étudiant est de l'UTC ou en semestre d'échange
- **dernier.diplome.obtenu :** BAC, DUT, CPGE, ETRANGER SUPERIEUR, LICENCE, OTHER, NA
- **Note Médian :** note de l'examen Médian
- **Correcteur Médian :** ID du correcteur du médian de 1 a 8.
- **Note.final :** Note de l l'examen final
- **Note.totale :** Note totale obtenue a partir de la note du médian et la note du final
- **Correcteur.final** : D du correcteur du final de 1 a 8.
- **Résultat** : Résultat obtenu de SY02 : A, B, C, D, E et F.

Pour les notes de médian et de final il y'a des notes non mentionnées (NA) par contre tous les etudiants ont un résultat final c'est pour cela on n'a pas enlevé les etudiants avec NA en médian ou/et en final.

Les variables importantes dans ce dataset sont les résultats des etudiants et comment ceux ci sont influences par d'autre variable, par exemple le niveau et la spécialité
Il est évident qu'il existe une relation linéaire entre ces 3 variables : note.median, note.final et note.totale vu que la note totale est exprimée par une relation linéaire entre la note.median et la note.final (par exemple $note.totale = 40\% * note.median + 40\% * note.final + cste$). La variable note.totale et la variable resultat sont des variables fortement corrélés En effet le résultat est une "traduction" de la note totale. On pourrait éventuellement se demander sur la relation entre les notes et le niveau, la spécialité, le diplôme ainsi que le correcteur. C'est ce qu'on va essayer d'analyser dans ce qui suit.

1.1.2 Analyse descriptive des données

Lien entre les variables

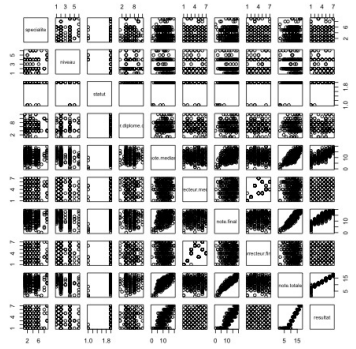


FIGURE 1.1 – Plot general

Performance et homogénéité

La figure ci-dessous représente trois diagrammes a boites des notes de médian, final et le résultat de l'UV SY02.

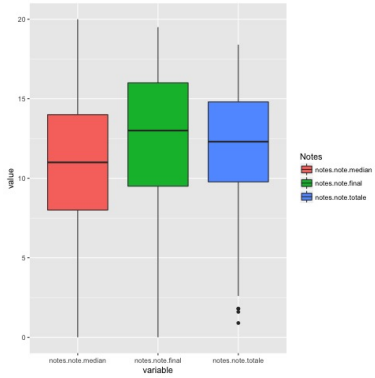


FIGURE 1.2 – Boxplot des Notes de Median, Final et Totale

On remarque que par passage du médian au final les notes ont augmente. En effet ceci est aussi témoigné par le tableau ci-dessous :

Notes	**1er Quartl**	**Medianne**	**Moyenne**	**3em Quartl**
Median	8.0	11.0	10.92	14.0
Final	9.50	13.0	12.38	16.0
Totale	9.775	12.3	11.845	14.8

TABLE 1.2 – Notes

Il est tout a fait logique que le digramme de résultats se situe entre les 2 puisque que c'est la moyenne **pondérée ? ?** des 2 notes.

Lien entre la réussite, la formation, la branche et le niveau

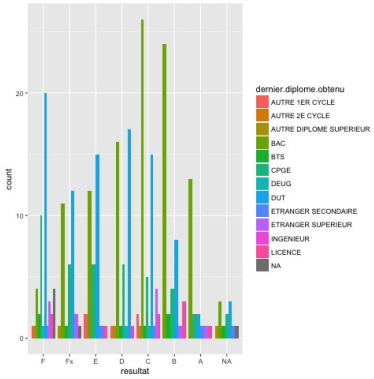


FIGURE 1.3 – Diagramme a baton de lien entre la formation et le resultat

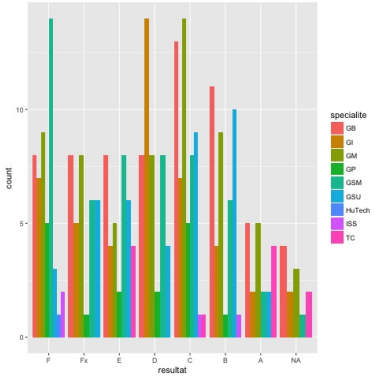


FIGURE 1.4 – Diagramme a baton de lien entre la branche et le resultat

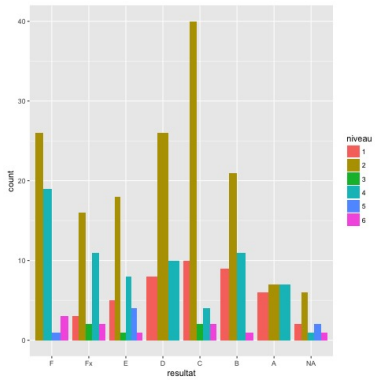


FIGURE 1.5 – Diagramme a baton de lien entre le niveau et le resultat

Influence du correcteur sur la note

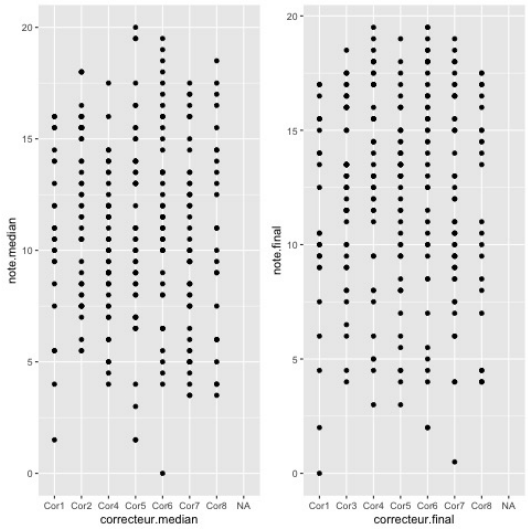


FIGURE 1.6 – Scatterplot des notes en fonction des correcteurs

1.2 Conclusion

1.3 Crabs

Le dataset *"Crabs"* représente un jeu de données de 200 crabes décrits par huit variables, trois sont qualitatives et cinq sont quantitatives.

1.3.1 Description des variables

— **Variables Qualitatives Nominales :** crabs.sp, crabs. sex, crabs.inde
— **Variables Quantitatives :** crabs.FL, crabs.RW, crabs.CL, crabs.CW, crabs.BD

Nous pouvons representer les donnees des variables quantitatives a l'aide un boxplot.

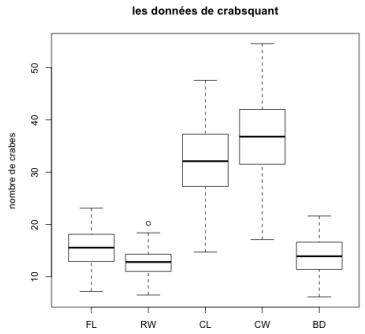


FIGURE 1.7 – Boxplot des données quantitaves

Nous remarquons d'ores et deja que deux categories de variables se distinguent, d'un cote FL, RW et BD et d'un autre, CL et CW.

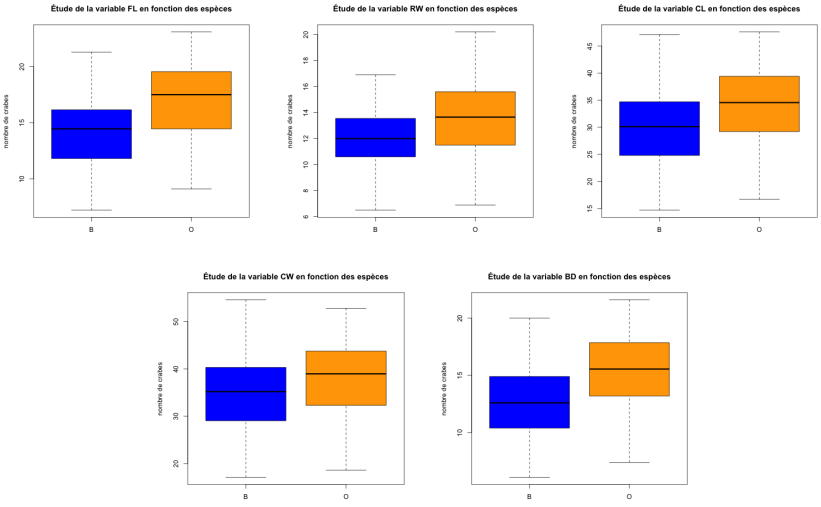
Avant de continuer l'analyse de ces donnees, nous pouvons preciser la signification de chacune de ces variables comme suit :

— **sp :** *(species)*, espece, "B" pour Bleu et "O" pour Orange
— **sex :** sexe, "F" pour Feminin et "M" pour masculin
— **index :** index de 1 a 50 pour chacune des 4 categories suivantes : {"B,M","O,M","B,F","O,F"}
— **FL :** Frontal Lobe size en mm
— **RW :** Rear Width en mm
— **CL :** Carapace Length en mm
— **CW :** Carapace Width en mm
— **BD :** Body Depth en mm

1.3.2 Analyse descriptive des données

Representation de chaque caracteristique en fonction de l'espece

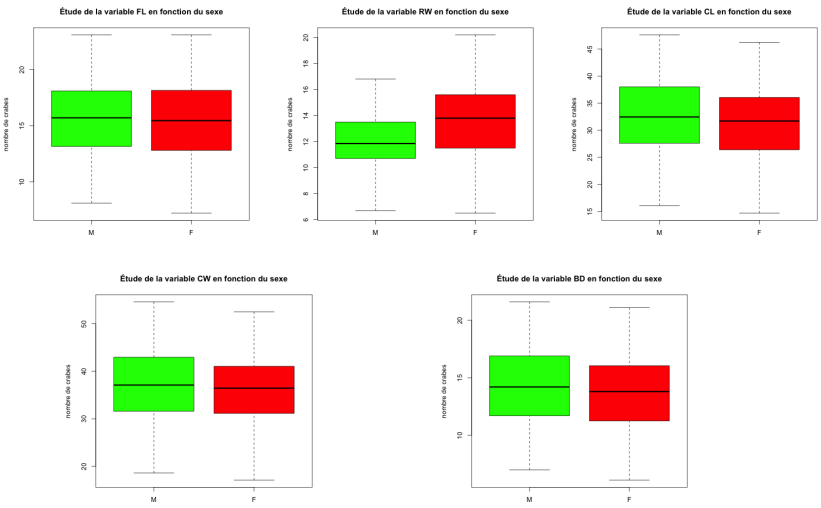
Afin de voir s'il y a une difference de caracteristiques morphologiques en fonction de l'espece d'abord, nous representons les boites a moustache de chaque variable morphologique en fonction de la variable sp comme suit :



Il y a des distributions assez differentes en fonction de l'espece. Les intervalles de confiances ne se chevauchent pas . Cependant, la dispersion des donnees reste assez homogene au vue des boxplots.

Representation de chaque caracteristique en fonction du sexe

Afin de voir s'il y a une difference de caracteristiques morphologiques en fonction du sexe, nous representons les boites a moustache de chaque variable morphologique en fonction de la variable sex comme suit :



A l'oppose des boxplot en fonction de l'espece, ceux en fonction du sexe relevent une similitude de la distribution des differentes caracteristiques pour la

plupart, ainsi que la dispersion des donnees. Nos observons que la variable RW se distingue des autres avec un largeur de l'arriere importante chez les femmes que chez les hommes.

Enfin, la nature de l'espece impacte les caracteristiques morphologiques, a la difference du sexe, qui lui n'influe que peu ces caracteristiques.

Lien entre les variables

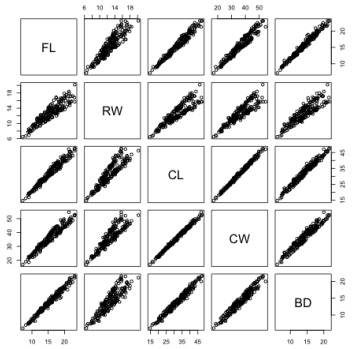


FIGURE 1.8 – Plot general

Nous pouvons representer chacune de ces variables quantitatives en fonction de l'espece puis en fonction du sexe afin de determiner la possibilite d'identifier l'une ou l'autre a partir des caracteristiques morphologiques.

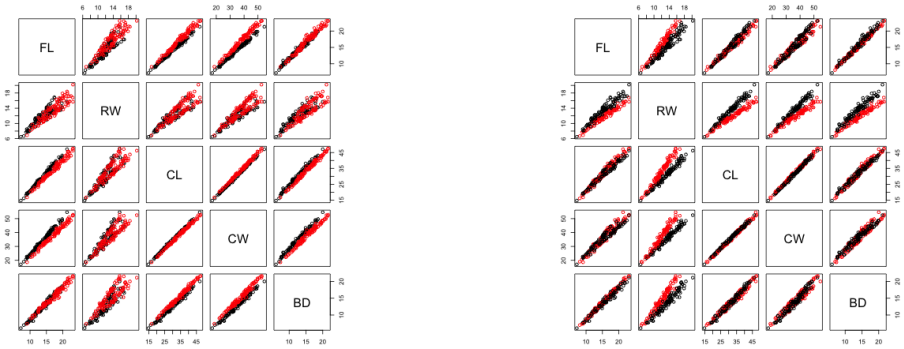


FIGURE 1.9 – Plot general en fonction de l'espece (gauche) et du sexe(droite)

Nous remarquons que ni l'espece ni le sexe ne peuvent vraiment etre identifies a partir d'une ou de plusieurs caracteristiques morphologiques. En effet, dans les cas, l'ensemble des points est represente sur une meme droite, on ne peut pas clairement distinguer l'une difference. De ce fait, il est difficile de reconnaitre une espece selon ses carateristiques morphologiques.

1.3.3 Analyse de la correlation

	FL	RW	CL	CW	BD
FL	1.000	0.907	0.979	0.965	0.988
RW	0.907	1.000	0.893	0.900	0.889
CL	0.979	0.893	1.000	0.995	0.983
CW	0.965	0.900	0.995	1.000	0.968
BD	0.988	0.889	0.983	0.968	1.000

FIGURE 1.10 – Correlation entre les variables

Il y a une forte correlation positive entre toutes les combinaisons de variables, telle que la valeur minimale observee est 0.889. Il s'agit de la taille des membres du corps d'un crabe, il semble donc logique et naturel qu'elles soient proportionnelles entre elles. Une des facons pour s'affranchir de ce phenomene est de diviser chaque valeur par la somme totale de toutes celles de l'individu.

1.4 Pima

Le dataset *"Pima"* représente un jeu de données constitue de 532 individus tous de sexe feminin décrits par huit variables dont une qualitative et sept sont quantitatives.

1.4.1 Description des variables

— **Variables Qualitatives Ordinale :** Pima.z
— **Variables Quantitatives :** Pima.npreg, Pima.glu, Pima.bp, Pima.skin, Pima.bmi, Pima.ped, Pima.age

Nous pouvons representer les donnees des variables quantitatives a l'aide un boxplot.

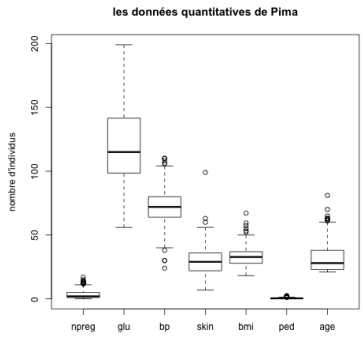
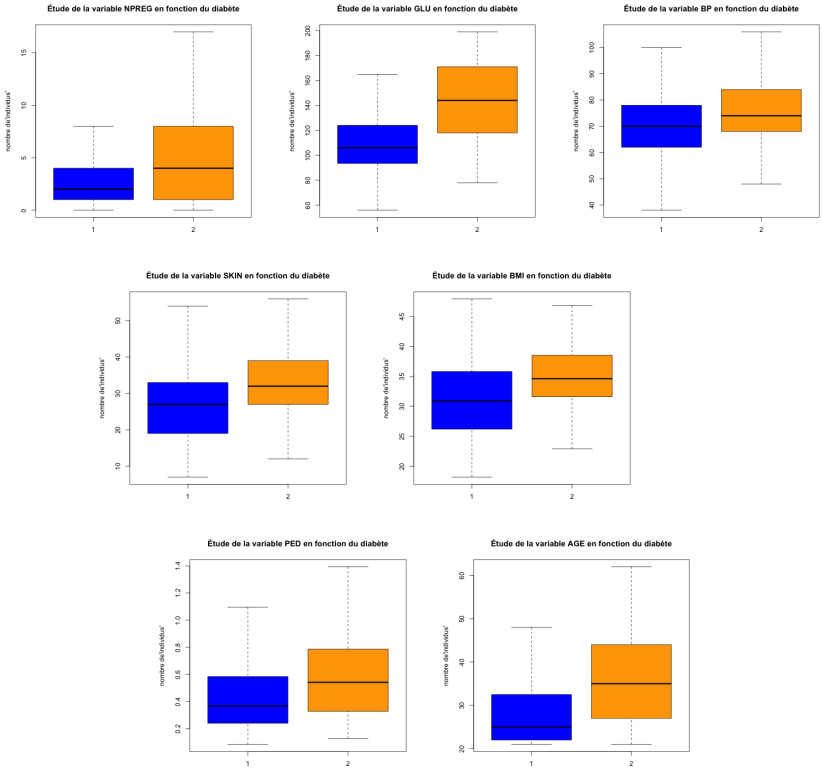


FIGURE 1.11 – Boxplot des données quantitaves

1.4.2 Analyse descriptive des données

Representation de chaque caracteristique en fonction de z (diabetique ou pas)



Lien entre les variables

Chapitre 2

Analyse Composantes Principales