Université de Technologie de Compiegne

SY09

Data Mining

Premier Rendu

Oumaima Talouka Zineb SLAM

21 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

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

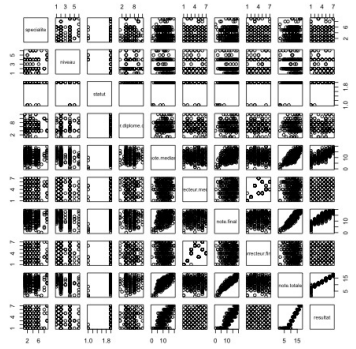

FIGURE 1.1 – Plot general

Performance et homogénéité

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

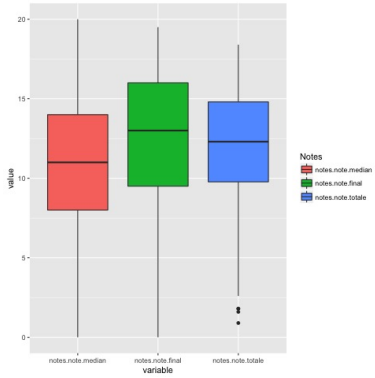

FIGURE 1.2 – Boxplot des Notes de Median, Final et Totale

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

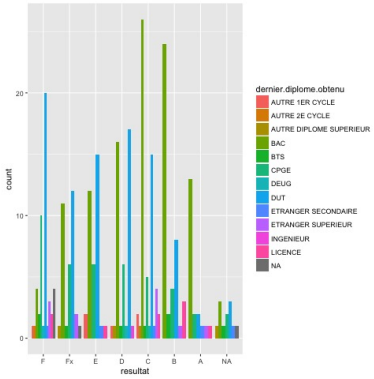

FIGURE 1.3 – Diagramme a baton de lien entre la formation et le resultat

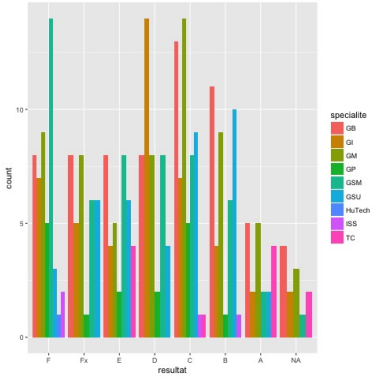

FIGURE 1.4 – Diagramme a baton de lien entre la branche et le resultat

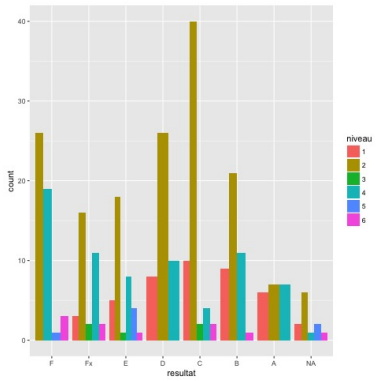

FIGURE 1.5 – Diagramme a baton de lien entre le niveau et le resultat

Influence du correcteur sur la note

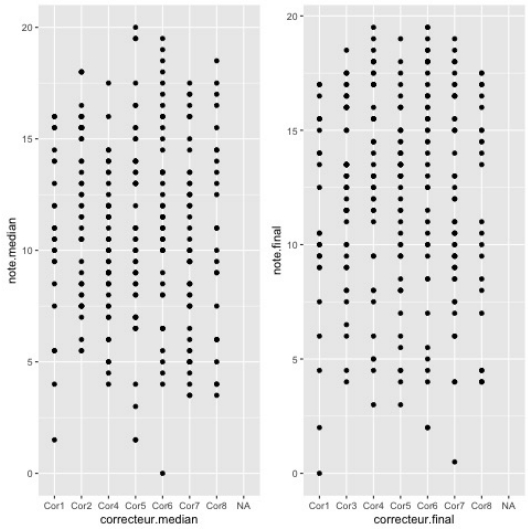

FIGURE 1.6 – Scatterplot des notes en fonction des correcteurs

1.2 Conclusion

1.3 Crabs

1.4 Pima

Chapitre 2

Analyse Composantes Principales