Université de Technologie de Compiegne

SY09

Data Mining

Premier Rendu

Oumaima Talouka Zineb SLAM

25 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.

0.1 Statistique descriptive

0.1.1 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. Nous avons omis les etudiants qui n'avaient pas de résultats (*NA*) ou qui ont '*ABS*' en résultat de l'UV parce que ce qu'on a pense que ce sont des etudiants qui s'étaient de-inscris ou n'ont pas suivi l'UV et donc ne constituent pas d'information pour notre dataset. Nous gardons néanmoins les etudiants qui ont NA en médian ou en final mais qui ont obtenu un résultat final.

```
dataset = dataset[!is.na(dataset$resultat), ]
dataset = dataset[dataset$resultat!='ABS', ]
```

Nous gardons donc $N = 284$ individus pour notre étude

Description des variables

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

Nous allons a présent définir chaque variable en précisant son intervalle ou les valeurs qu'elle peut prendre.

- **Nom :** chaine de caractère identifiant chaque étudiant.
- **Spécialité :** la branche de l'étudiant : *GB*, *GM*, *GSM*, *GP* et *GI*.
- **Niveau** : Semestre d'étude de chaque étudiant de 1 a 6.
- **Statut** : Soit l'étudiant est de l'*UTC* ou en s*emestre d'échange*
- **dernier.diplome.obtenu :** *BAC, DUT, CPGE, ETRANGER SUPERIEUR, LICENCE, OTHER, NA* NA sont les etudiants qui n'ont pas informe leur diplôme
- **Note Médian :** note de l'examen Médian : Un ensemble de réel de 0 a 20.
- **Correcteur Médian :** ID du correcteur du médian de 1 a 8 de la forme *Corr1* à *Corr8*
- **Note.final :** Note de l'examen final. Un ensemble de réel de 0 a 20.
- **Note.totale :** Note totale obtenue a partir de la note du médian et la note du final
- **Correcteur.final** : identifiant du correcteur du final. Ce sont les mêmes 8 correcteurs que le médian mais qui n'ont pas forcement corrige les même copies que le médian/
- **Résultat** : Résultat obtenu en SY02 : *A, B, C, D, E et FX et FX. ABS*. ABS est pour indiquer que l'étudiant était absent pour l'examen en question.

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 résultat 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.

Liens entre les variables

Après avoir affiche la matrice de graphes pour chaque variable nous avons pu observe que les variables : note.median, note.final, et note.totale sont linéaire ce qui confirme note hypothèse précédente Le graphe matriciel ci-dessous présente aussi les coefficients de corrélation entre chacune de ces variables.

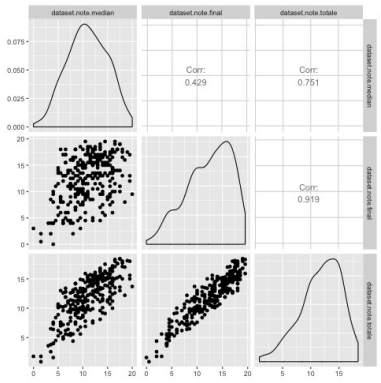


FIGURE 1 – Plot general

Performance et homogénéité

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

On remarque les diagrammes sont petit ce qui laisse penser que la variance des résultats est relativement petite et donc aussi que le niveau des etudiants en statistiques l'est aussi. En comparant les diagrammes du final et du médian on remarque que les notes ont augmente. Enfin si on analyse le dernier diagramme on voit que les deux boites du part et d'autre de la médiane ont la même taille, on peut donc mettre l'hypothèse que les résultats finaux suivent une *loi normale*.

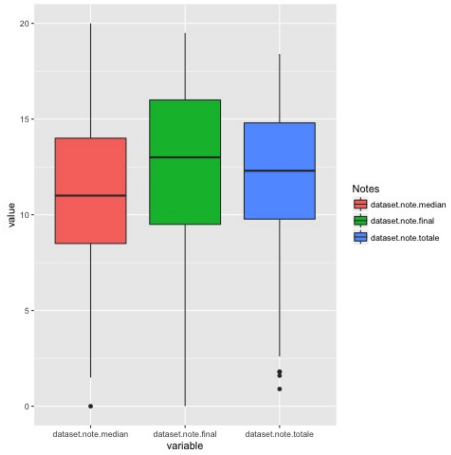


FIGURE 2 – Boxplot des Notes de Median, Final et Totale

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 – Notes

Il est tout a fait logique que le digramme de boites des notes finales se situe entre les 2 puisque que la note finale est une moyenne des deux notes du médian et du final.

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

Comme les etudiants de chaque branche n'ont pas les *même effectifs* on a choisi de représenter les données sous forme de diagrammes de moustache pour mieux pouvoir les comparer.

D'après la figure 3 on remarque que les etudiants venant durant les premiers semestres sont ceux qui réussissent le mieux leur examens de SY02 , suivi par ceux en GX02. On remarque que les notes des etudiants en GX04 et GX05 ont une grande variance. Les étudiants en GX04 et en GX05 ont . Il faut aussi remarquer que les étudiants en 4em et 5 em sont peu comme c'est des semestres de départ en stage.

En ce qui concerne l'influence de la spécialité sur les notes on remarque qu'il y'a une grande variance dans les notes chez les etudiants en TC et ISS et GSM, contrairement au GI et GP qui ont plutôt un niveau homogène et qui semble bien réussir l'UV. De plus les TC sont aussi ceux qui réussissent le mieux l'UV.

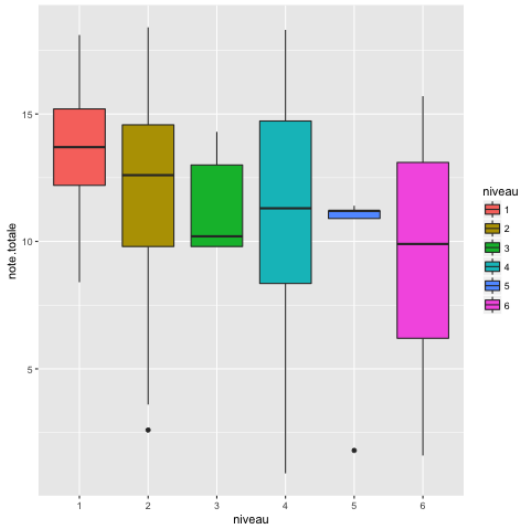


FIGURE 3 – Diagramme en boite de lien entre le niveau et le resultat

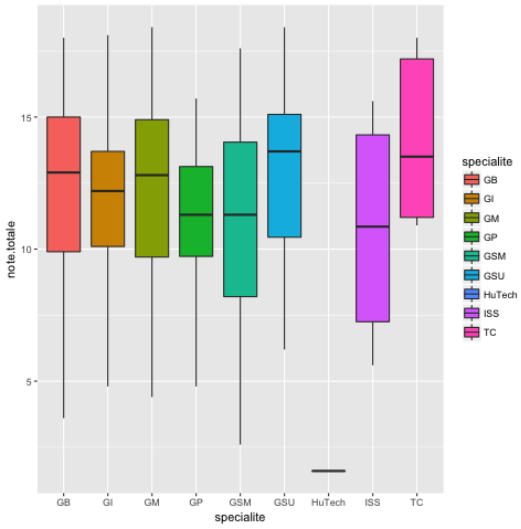


FIGURE 4 – Diagramme en boite de lien entre la branche et le resultat

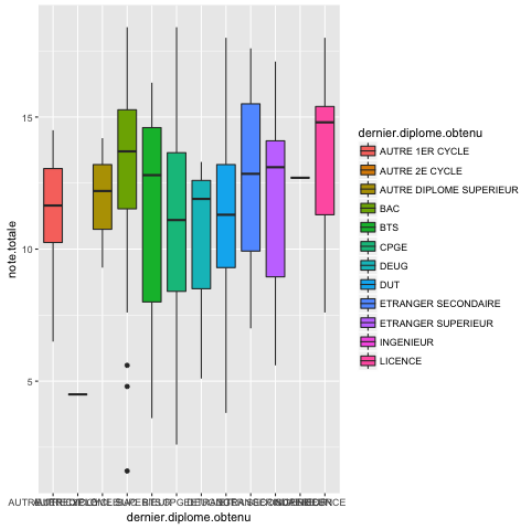
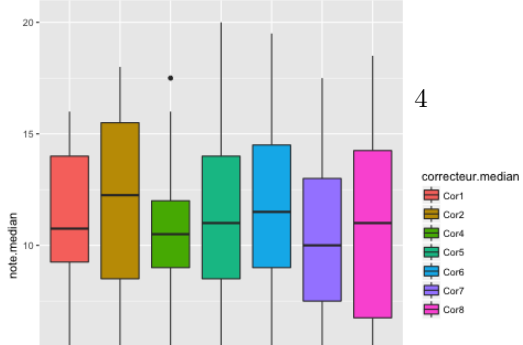


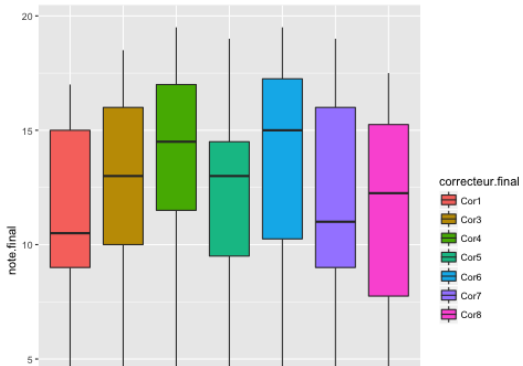
FIGURE 5 – Diagramme en boite de lien entre la formation et le resultat

On remarque qu'il y'a une grande variance chez presque tous les diplômes obtenus. On peut remarquer cependant que les élèves du autre premier cycle et second cycle ainsi qu'autres diplômes réussissent en général bien en SY02 et ont un petite variance. Seulement on ne peut généraliser le cas ici vu que leur effectif est faible par rapport aux autres élèves des autres branches **TO CHECK !** De plus pour les élèves provenant de License ou de Deug on remarque que le niveau est hétérogène vu que les boites du dessous sont plus longues et donc **? ? ?**.

Influence du correcteur sur la note

Les deux diagrammes ci-dessous montrent la dispersion des notes en final et en médian chez chaque correcteur.





Conclusion

Cette première analyse data nous a permis d'étudier quels sont les facteurs qui influent sur la réussite d'un étudiant dans l'UV SY02 comme le dernier diplôme obtenu ou le niveau et ceux qui n' influence pas comme le correcteur. Néanmoins ces conclusions sont propre a la population du P17 et donc biaisées ; pour pouvoir généraliser il faut analyser les notes de SY02 sur plusieurs semestre avec des populations différente

0.1.2 Crabs

0.1.3 Pima

0.2 Analyse Composantes Principales