Rapport analyse de donnéess
Dataset Villes

Abderahman Larbi, Mohammed Lacarne, Mohammed Dahi
Université Paris Descartes
Master 1 MLDS
17 janvier 2019

Résumé

Ce rapport à pour but d'analyser un dataset comportant quelques villes partout dans le monde afin de faire une comparaison. Pour cela, on optera pour des methodes de pretraitement, reduction dimentionnelle et d'analyses.

1 Présentation du Dataset :

On s'intéresse à 51 villes. Les données recueillies ne font pas seulement référence aux salaires mais elles constituent un ensemble plus vaste de 40 variables concernant aussi les prix et quelques indicateurs essentiellement économiques. Les villes sont réparties dans plusieurs régions du monde, les observations sont connues à 2 dates (1991 et 1994). On ne considérera ici que les données de l'édition 1994. Comme le but de l'étude est de comparer les villes selon leur niveau de salaires, seules les 12 dernières variables seront retenues comme actives pour l'analyse.

2 pétraitement :

2.1 restructuration et nettoyage :

Extraction : suppression des lignes qui represente une observation de l'année 1991, vu qu'on nous demande de faire une analyse seulement sur les observations de l'année 1994.

Variable manquante : detection d'une case manquante dans la variable "Bus" qu'on va la considerer comme supplémentaire. On a remplacé cette cas avec la moyenne de la colonne respective.

3 Analyse descriptive

3.1 Individus et variables :

Individus : il y'a 51 villes dans le dataset : AbuDhabi, Amsterdam, Athenes, Bangkok, Bogota, Bombay, Bruxelles, Budapest, BuenosAires, Caracas, Chicago, Copenhague, Dublin, Dusseldorf, Frankfurt, Geneve, Helsinki, Hongkong, Houston, Jakarta, Johannesburg, Lagos, Lisboa, London, LosAngeles, Luxembourg, Madrid, Manama, Manila, Mexico, Milan, Montreal, Nairobi, NewYork, Nicosia, Oslo, Panama, Paris, Prague, RiodeJaneiro, SaoPaulo, Seoul, Singapore, Stockholm, Sidney, Taipei, Tel-Aviv, Tokyo, Toronto, Vienna, Zurich.

Variables : on a 40 variables mais seules les dernieres 12 colonnes seront reternues comme actives pour l'analyse.

instit : salaire moyen d'un instituteur.

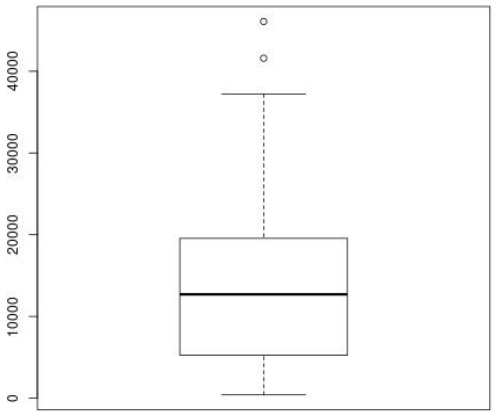
chauffeur : salaire moyen d'une chauffeur d'autobus.

meca : salaire moyen d'un mécanicien.

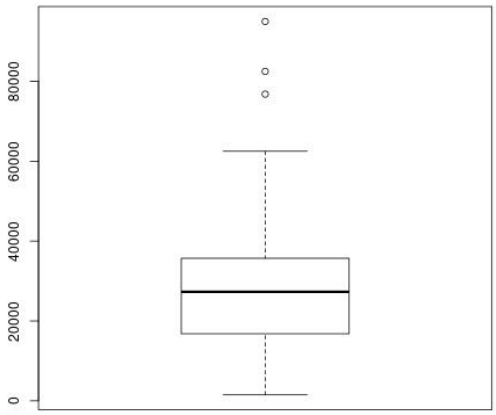
man : salaire moyen d'un manœuvre en bâtiment.

tourneur : salaire moyen d'un ouvrier tourneur.
cuisinier : salaire moyen d'un chef cuisinier.
chefserv : salaire moyen d'un chef de service (cadre).
inge : salaire moyen d'un ingénieur.
banque : salaire moyen d'un employé de banque.
secr : salaire moyen d'une secrétaire de direction.
vendeuse : salaire moyen d'une vendeuse.
ouvriere : salaire moyen d'une ouvrière du textile.

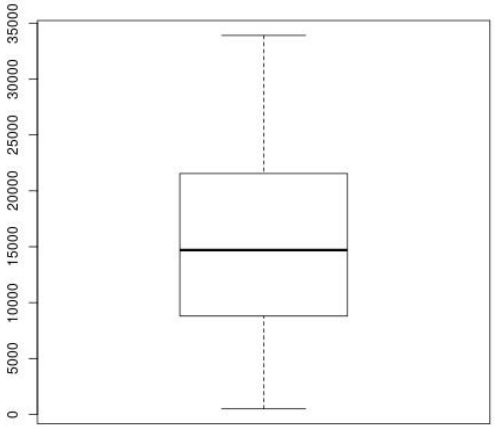
3.2 Boxplots



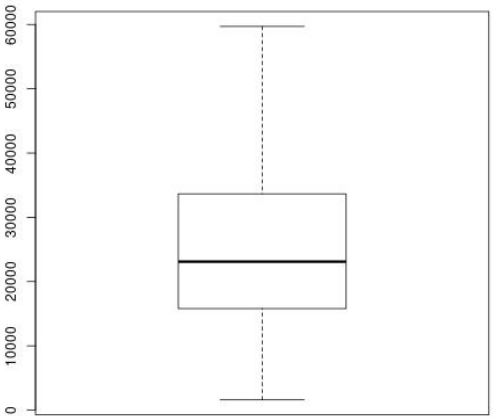
(a) Chauffeur



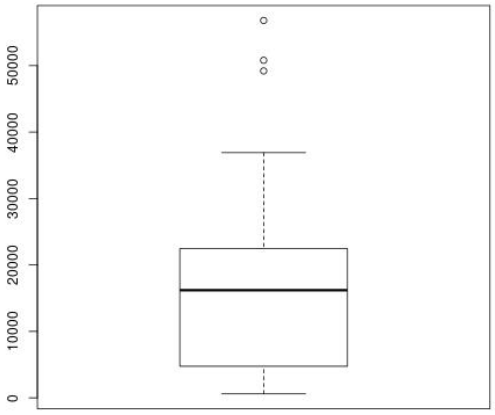
(b) Cher de service



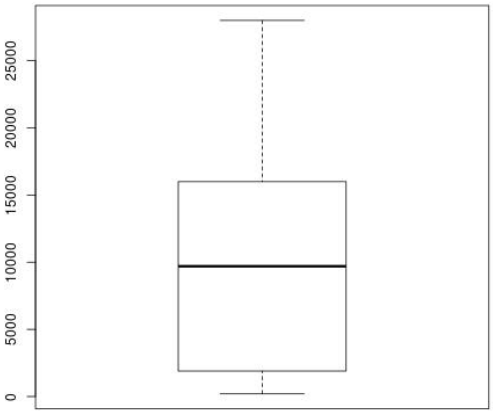
(a) Cuisinier



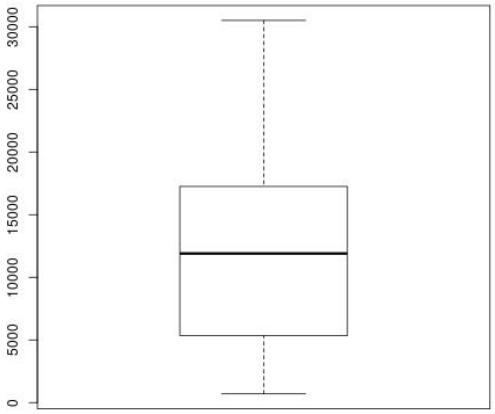
(b) Ingénieur



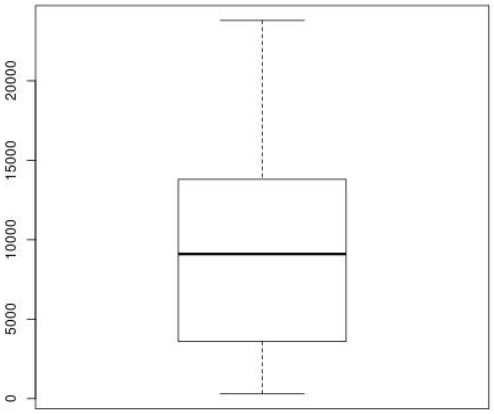
(a) Instituteur



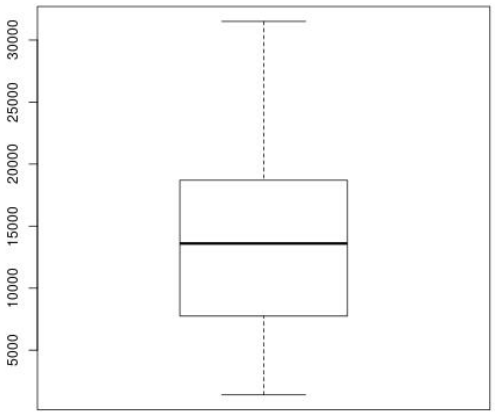
(b) Manœuvre



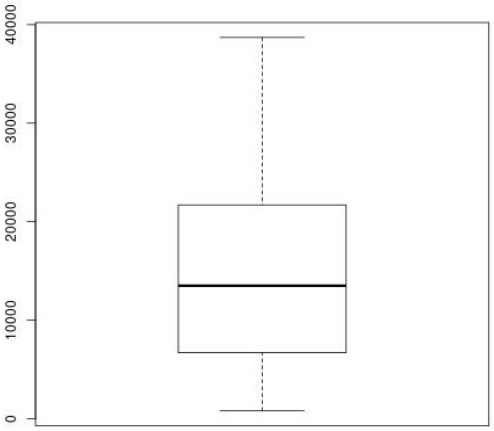
(a) Mécanicien



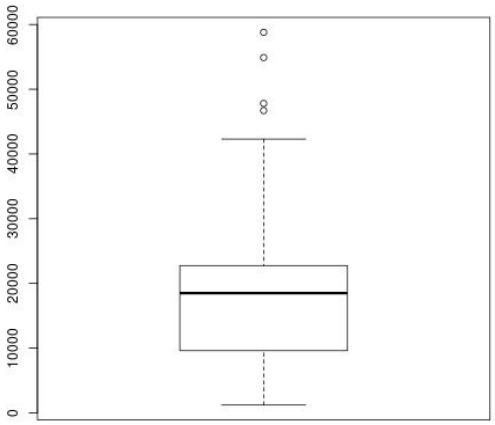
(b) Ouvrière



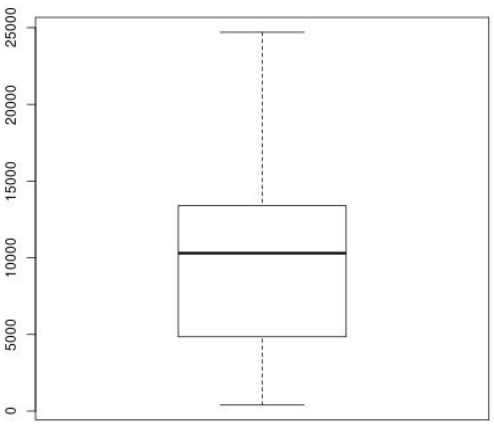
(a) Secretaire



(b) Tourneur



(a) Employé banque



(b) Vendeuse

4 Analyse :

Variables supplémentaires : on a considéré la variable "région" comme variable qualitative supplémentaire.

	Eigenvalue	percentage of variance
comp 1	10.13899999	84.4916666
comp 2	0.86118840	7.1765700
comp 3	0.32480801	2.7067334
comp 4	0.17145196	1.4287663
comp 5	0.14839937	1.2366614
comp 6	0.09728986	0.8107488

TABLE 1 – Tableau de valeurs propre des composantes

La matrice des valeurs propres : montre que les composantes principale 1 et 2 a elles seules decrivent 92% de la variabilité des données.

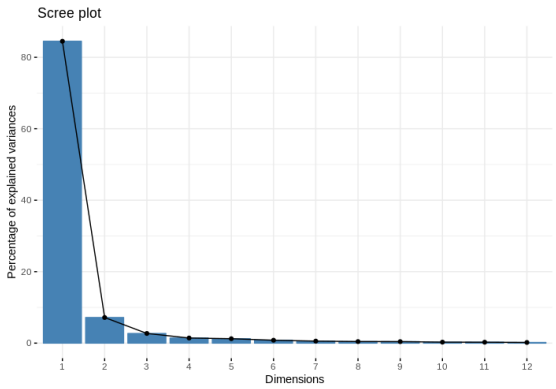
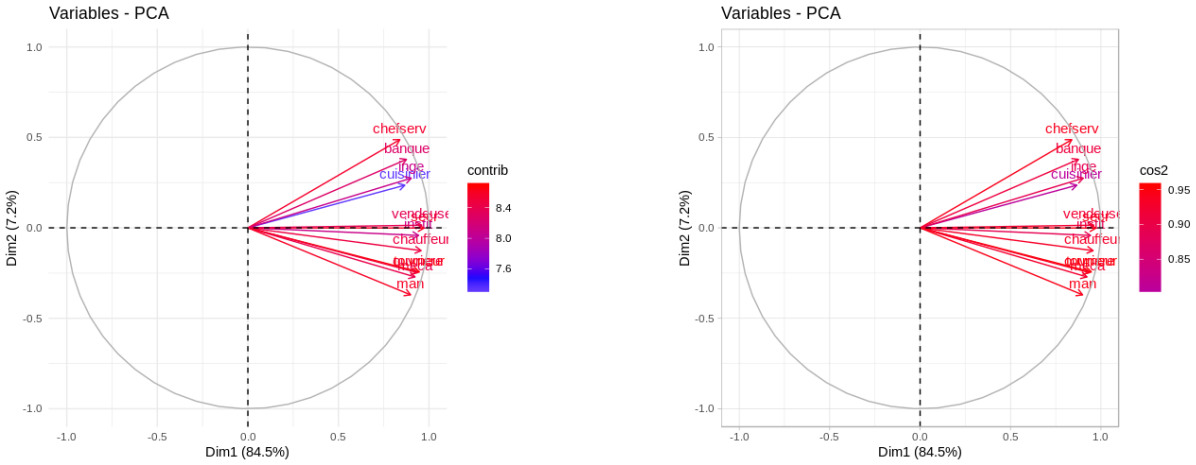


FIGURE 1 – Graphe des dimensions

Cercle de correlations : A partir des cercles de corrélation on voit que toutes les variables sont corrélées positivement avec la pca1. Le graphe montre la correlation des variables avec les axes. Les variables sont colorés en fonction de la valeur des cosinus au carré(cos2) . Le graphe (kima bghit) montre la correlation des variables avec les axes. Les variables sont colorés en fonction de la valeur contrib .



(a) Cercle de correlation (contribution)

(b) Cercle de correlation (cos2)

Contribution des variables a la pca1 : Le graphe montre la contribution des variables par rapport a la pca1, on voit que les variables qui contribuent le plus sont : secretaire, vendeuse, chauffeur , tourneur, instit , ouvrière et mécanicien contribuent le plus par rapport a pca1. La ligne en rouge signifie le seuil minimum pour lequel on décide si la variable apporte une contribution importante ou pas par rapport à l'axe 1.

Contribution des variables a la pca 2 : Les variables qui contribuent le plus a la pca2 sont : chef de service ,banquier,manoeuvre,ingénieur et mécanicien.

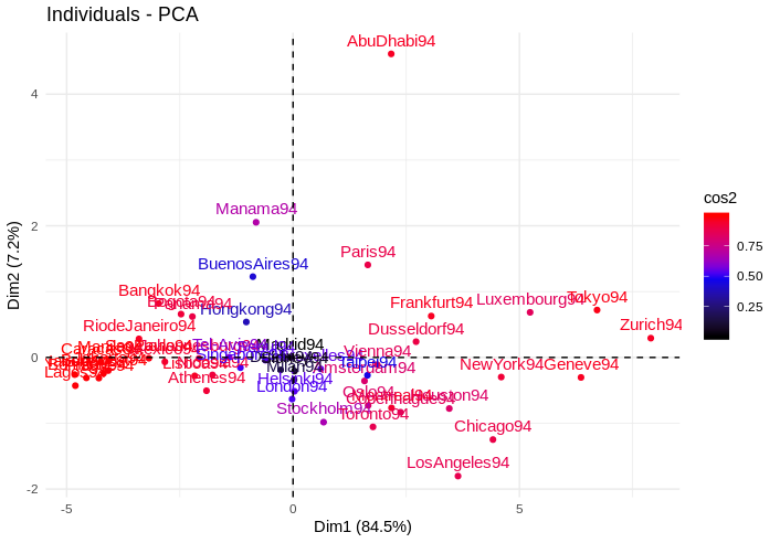


FIGURE 2 – graph des individus

Pour avoir une meilleure visualisation des individus, on decide des les colorées en fonction de leur cos2 afin de faire le lien avec les variables de notre dataset.

5 Conclusion

:

Par rapport a l'axe 1 : La pca1 represente les salaires d'emplois de classe moyenne(secretaire, vendeuse, chauffeur , tourneur ,instituteur , ouvriere). Ce qui signifie que dans le graphe des individus ,les villes qui se situent a gauche de l'axe 1 sont corrélées positivement avec celui-ci, plus les villes sont situés a gauche de l'axe 1 plus on a un salaire elevé pour ces emplois. Plus les villes sont a droite de l'axe 1 et plus les salaires sont bas pour ces emplois. Les villes se situant au milieu de l'axe correspondent aux villes ou ces emplois ont un salaire moyen.

Par rapport a l'axe 2 : Les villes ayant une grande valeur par rapport a l'axe 2 correspondent aux villes ou les emplois(banquier,chef de service et ingenieur) sont tres bien rémunérés. Plus les villes prennent de petites valeurs pour l'axe 2, plus les salaires sont faible pour(banquer,chef de service et ingenieur) et les salaires des manoeuvres et mecaniciens sont moyens.

Interpretation par région : en rajoutons la variable region comme variable supplementaire qualitative, on remarque que pour les villes d'afrique, d'amerique du sud et d'asie du sud ,l'ensemble des emplois ne sont pas bien rémunérés. On remarque aussi que pour les villes du proche orient, les emplois(chef de service, banquier et ingenieur) sont tres bien rémunérés tandis que les emplois mecanicien et manoeuvre ont une faible remuneration, tandis que pour le reste on a une remuneration moyenne.

Les villes d'europe du sud , asie de l'est(a l'exception de tokyo) , europe du nord correspodent aux villes ou l'on a un salaire moyen pour pratiquement tous les emplois.

Les villes d'europe centrale et d'amerique du nord correspondent aux villes ou les emplois de classe moyenne sont tres bien rémunérés.

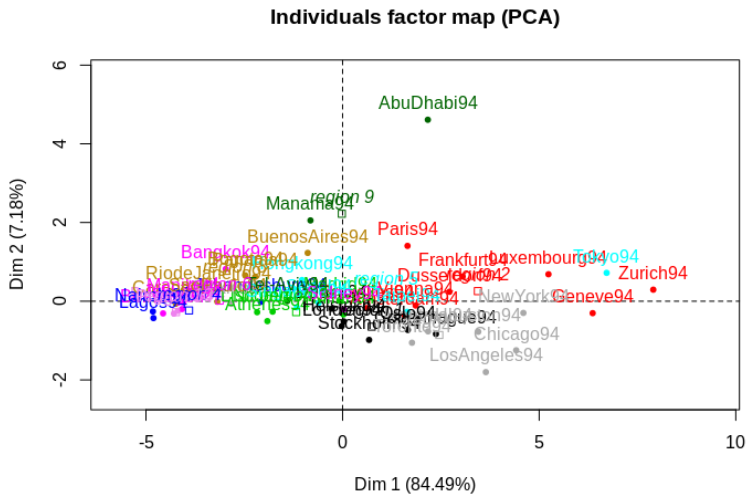


FIGURE 3 – graphe individus par régions