Analyse de Données
TP3 – Langage R

Ouvrir RStudio. Dans la partie Console :

1. Spécifier le répertoire de travail
2. Charger le fichier de données « percVec.csv »
3. Faire les transformations nécessaires pour convertir les attributs de 3 à 7 à des facteurs (variables qualitatives), et pour convertir les attributs 8 à 39 à des valeurs numériques (variables quantitatives).
4. Afficher les valeurs minimale et maximale des variables *MontMeuble* et *MontBoutique* représentant les pourcentages d'achat pour chaque rayon (Meuble et Boutique) par rapport au total d'achat. Que peut-on conclure ?

On veut étudier si l'on peut prédire la valeur de *MontMeuble* si l'on connait la valeur de *MontBoutique* pour un client donné.

5. Tracer le nuage de point représentant *MontMeuble* en fonction de *MontBoutique*. Peut-on dire qu'il existe une relation linéaire entre les deux variables ?
6. Supprimer du data set les lignes où *MontMeuble* et/ou *MontBoutique* dépassent 100%.
7. Retracer le nuage de points. Que peut-on conclure ?
8. Calculer le coefficient de corrélation de Pearson entre *MontMeuble* et *MontBoutique*. Que peut-on déduire ?
9. Construire un modèle de régression linéaire prédisant *MontMeuble* à partir de *MontBoutique*. Afficher le résumé de la modélisation.
10. Lire les résultats de la modélisation :
    a. Quelle est l'équation établie ?
    b. En vérifiant la nullité des coefficients estimés par la p-value fournie (test de significativité), peut-on garder toujours la même équation ?
    c. Quelle est la qualité du modèle ? Qu'est-ce qu'elle représente ?
    d. Que signifie la statistique F ? d'après la p-value fournie, que peut-on dire ?
11. Valider le modèle établi :
    a. Vérifier la normalité des valeurs prédites (tracer QQ-plot et calculer la valeur de Kolmogorov-Smirnov)
    b. Vérifier la normalité des résidus
    c. Vérifier l'hétéroscédasticité des résidus
    d. Vérifier la non-autocorrélation des résidus (calculer le coefficient de Durbin-Watson)
12. Repérer les points influents en utilisant la distance de Cook.
13. Supprimer les points détectés et refaire la modélisation. Est-ce que la qualité du modèle a été améliorée ?
14. Construire un modèle de régression linéaire prédisant *NbPassage* à partir de *NbTotal* et *NbRetours*. Quelle est la qualité du modèle ?

15. Refaire la modélisation en rajoutant une autre variable explicative *HorsWknd*. Est-ce que la qualité est améliorée ?
16. Remplacer la variable *HorsWknd* par la *DiversiteProduits*. Quelle est la qualité maintenant ?
17. Rajouter en plus la *Tenure* (standardisée). Quelle est l'équation à établir si l'on vérifie les tests de significativité ?