TP final Régression – Petit challenge

On dispose d'un échantillon statistique de **n = 300** données de **ventes** globales journalières (en euros) d'une station-service ouverte tous les jours de la semaine. Conjointement, on fournit les températures moyennes journalières (en °C) ainsi que les jours de la semaine correspondants (lundi, …, dimanche). Ces données sont dans le fichier **'Data_app.txt'** (données d'apprentissage).

Le but est de construire le meilleur modèle prédictif des ventes en fonction des variables **'température'** et **'type de jour'**.

Le critère utilisé pour évaluer ce challenge est la moyenne des carrés des erreurs de prévision (en anglais, Root Mean Square Error):

$$\mathbf{RMSE} = \sqrt{\mathbf{MSE}} \quad \mathbf{où\ MSE} = \frac{1}{\mathbf{n.test}} \sum_{i=1}^{\mathbf{n.test}} (y_i - \hat{y}_i)^2$$

où n.test = 100 est le nombre de données 'test' (cf. fichier **'Data_test.txt'**), y_i les valeurs de vente réellement observées et $\hat{y}_i$ les prévisions associées à partir de votre modèle de Régression. On pourra s'aider au début des indications suivantes :

1. Commencer par une régression linéaire simple des ventes sur la température journalière. Obtenir le graphique montrant les données ainsi que la droite de régression. Qu'observe-t-on ? Pouvez-vous expliquer ? Tracer les résidus contre la réponse prédite. Conclusion ? Comment pourrait-on améliorer le modèle à partir de simples régressions linéaires ?

2. Envisager une régression linéaire multiple. Comparer à ce qui précède. Validez-vous le modèle ?

3. Dans le modèle RLM précédent, il peut être pertinent d'envisager une interaction entre les prédicteurs température et type de jour. Quel(s) nouveau(x) prédicteur(s) faut-il introduire ? Faire l'étude et comparer…

4. Réfléchir encore à d'autres améliorations possibles…

On déposera sur Campus en fin de séance le fichier texte correspondant à vos prévisions (fichier comportant donc une seule colonne formée de vos n.test = 100 prédictions de vente). Mettre ce fichier au format Votre**NOM_pred.txt**. Voir exemple de tel fichier (BAY_prevd.txt) correspondant à une prévision constante égale à la moyenne des ventes observées (modèle de régression linéaire vide où p = 0).