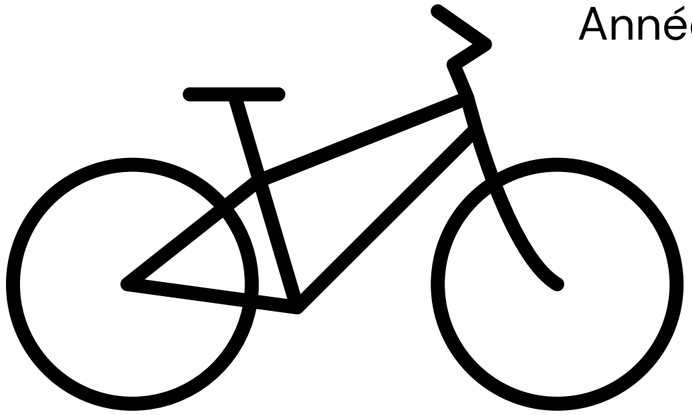
Projet Statistiques & Fouille de données

CNAM-MEDAS
Année 2024-2025



Seoul Bike Sharing Demand

Léa Dufraigne et Célia van den Eynde

Table des matières

Problématique

Le jeu de données sur la demande de vélos en libre-service à Séoul contient des informations sur la location de vélos à Séoul entre 2017 et 2018.

Il comprend des observations horaires sur 14 variables, telles que la date, l'heure, le nombre de vélos loués, les conditions météorologiques, ainsi que d'autres facteurs pouvant influencer la demande de location de vélos.

Ce jeu de données contient 8 760 lignes et 14 colonnes.

Objectif:

Construire un modèle de prédiction du nombre de vélos loués dans une heure, en fonction des variables explicatives disponibles (météo, jour, saison, etc.).

Méthodologie

Nous nous sommes inspirés de la méthode utilisée dans l'article suivant :

A rule-based model for Seoul Bike sharing demand prediction using weather data

de Sathishkumar V E et Yongyun Cho

L'un des objectifs était de comprendre leur démarche.

De nombreux modèles furent utilisés dont les KNN et le Random Forest que nous avons étudié en cours.

Variables

Date	La date à laquelle la location a eu lieu.
Hour	L'heure précise de la journée à laquelle la location a été enregistrée.
Temperature	La température extérieure mesurée en degrés Celsius.
Humidity	Le taux d'humidité dans l'air, exprimé en pourcentage.
Wind speed	La vitesse du vent, en mètres par seconde.
Seasons	La saison pendant laquelle la location a eu lieu (hiver, printemps, été, automne).
Holiday	Indique s'il s'agit d'un jour férié ou non.
Visibility	Le niveau de visibilité, mesuré par tranches de 10 mètres.
Dew point temperature	La température à laquelle la vapeur d'eau se condense (point de rosée), en degrés Celsius.
Solar Radiation	L'énergie solaire reçue au sol, exprimée en mégajoules par mètre carré.
Rainfall	La quantité de pluie tombée, en millimètres.
Snowfall	La quantité de neige accumulée, en centimètres.
Functioning Day	Indique si le service de location de vélos fonctionnait ce jour-là.

Rented Bike Count	Le nombre total de vélos loués au cours de l'heure observée (variable cible à prédire).

Préparation des données

Il n’y a aucune ligne dupliquée dans le jeu de données.

Il n’y a aucune valeur manquante ou valeur nulle dans le jeu de données.

Les colonnes Holiday et Functioning Day et Seasons ont été transformées en variables numériques.

Préparation des données

Le jeu de données a été divisé en deux :
80% base d'entrainement / 20% base test

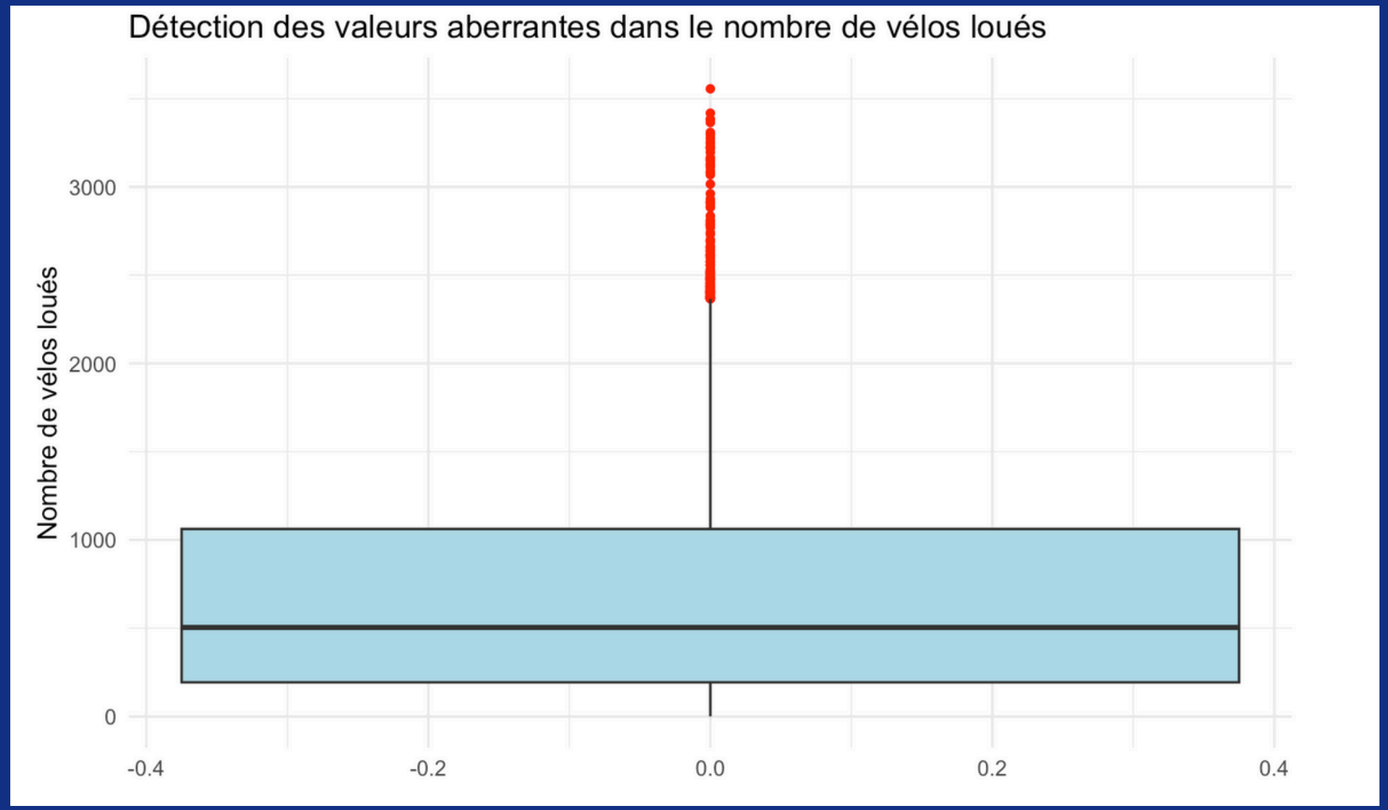
```
{r cars}
summary(data$Rented.Bike.Count)
summary(test_data$Rented.Bike.Count)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    0.0   193.0   507.0   706.6  1069.0  3556.0 
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    0.0   193.0   522.0   708.9  1088.5  3404.0 
```

Comparaison des variances des deux bases de données

Analyse descriptive

Rented Bike Count



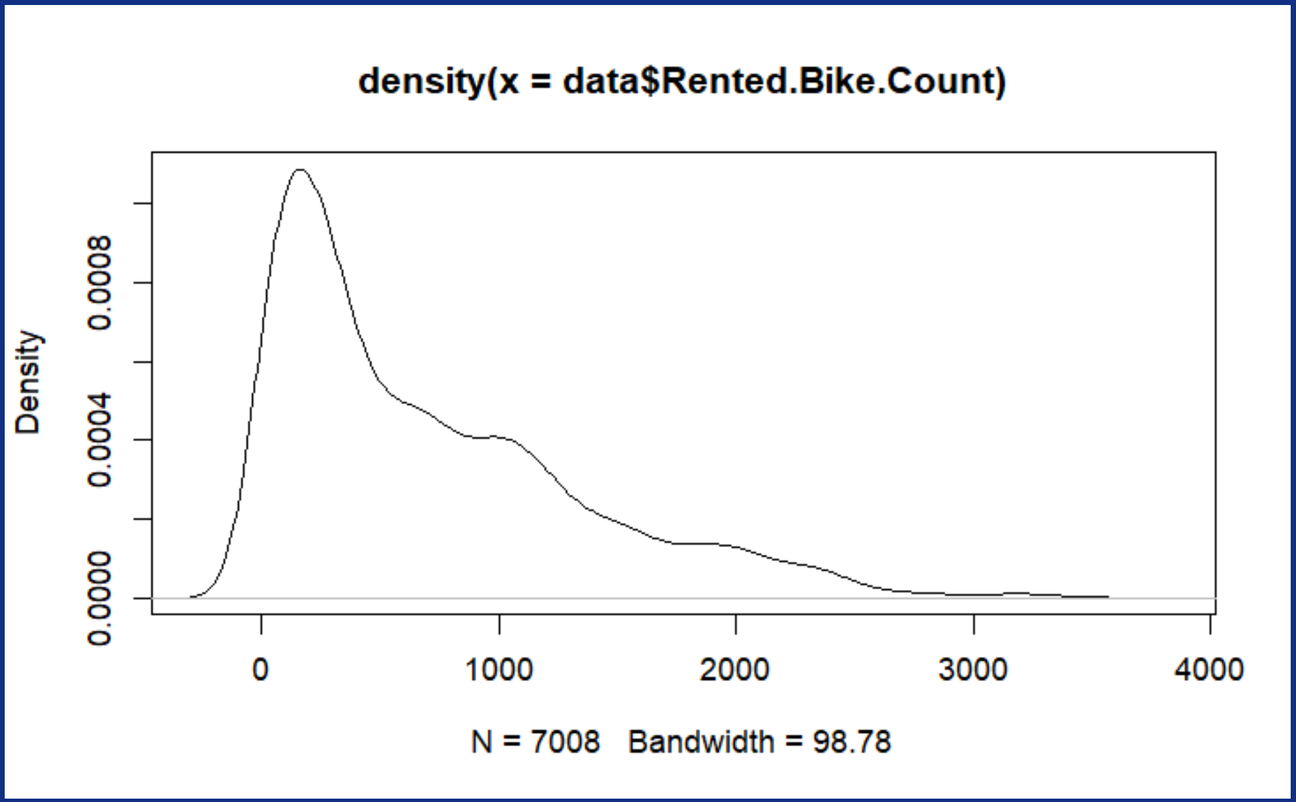
La médiane du nombre de vélos loués est d'environ 500. La moitié des heures ont vu moins de 500 vélos loués.

La boîte s'étend approximativement de 200 à 1100 vélos loués. La majorité des observations (environ 50 %) se situent dans cette plage.

Les observations situées au-dessus de 2500 locations par heure sont des outliers

Analyse de densité

Rented Bike Count



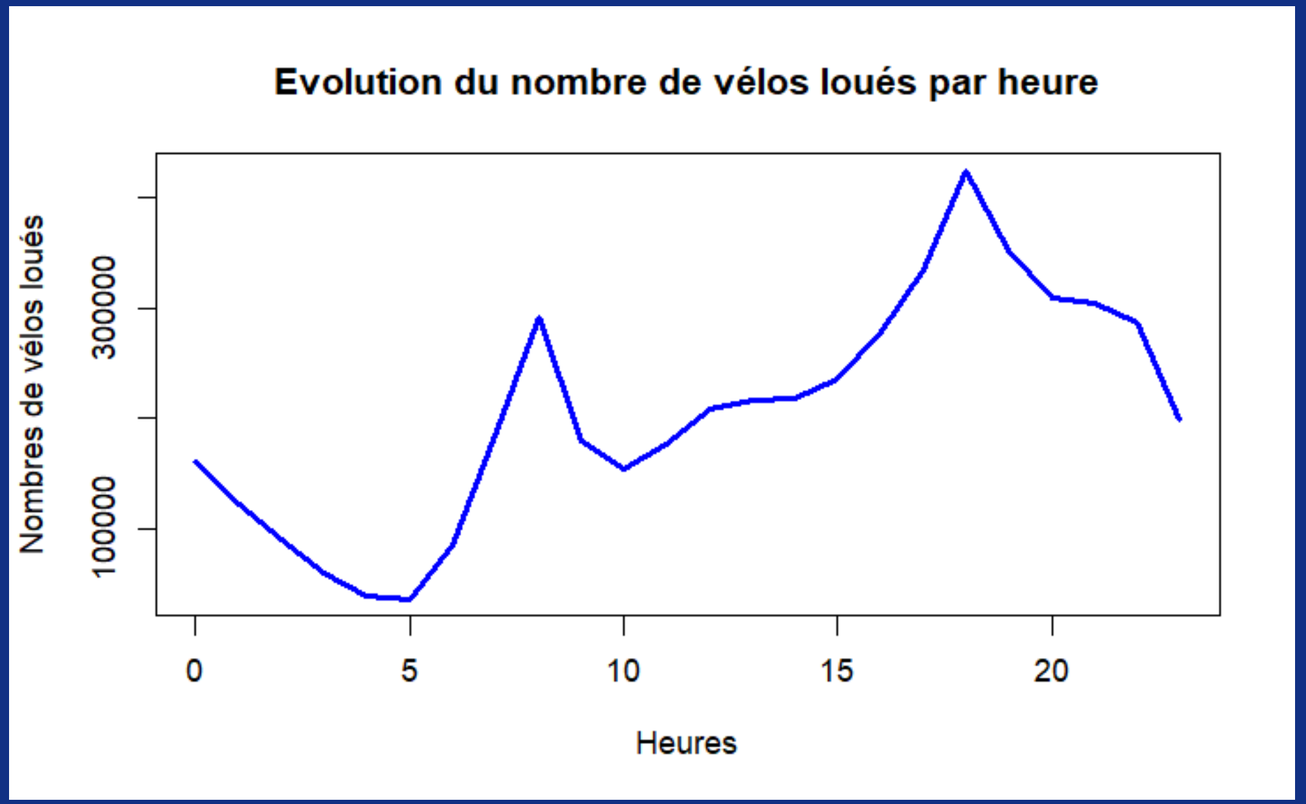
La distribution est fortement asymétrique à droite : les valeurs extrêmes existent mais sont peu fréquentes,

Le pic principal (mode) est situé vers 100 à 300 locations par heure : c'est la plage la plus fréquente dans les données.

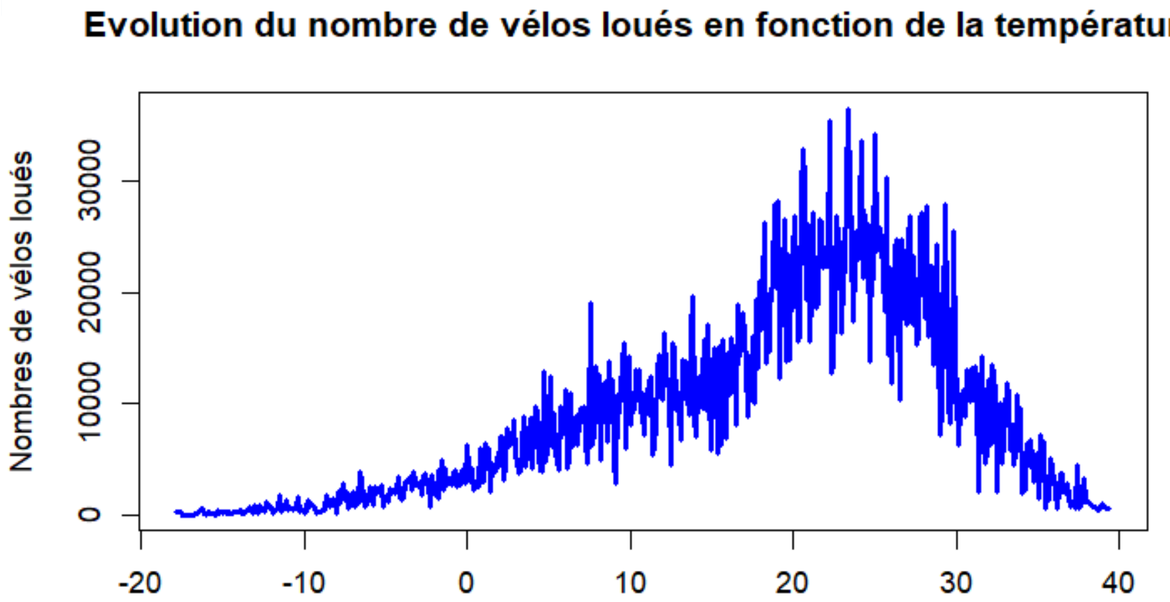
Nombres de vélos loués par heure

Répartition horaire du nombre de vélos loués
Une analyse agrégée du nombre total de vélos loués selon l'heure de la journée met en évidence deux pics d'utilisation :

- Un premier pic vers 8h, correspondant aux trajets domicile-travail du matin ;
- Un second pic plus marqué vers 18h, typique du retour à la maison en fin de journée.

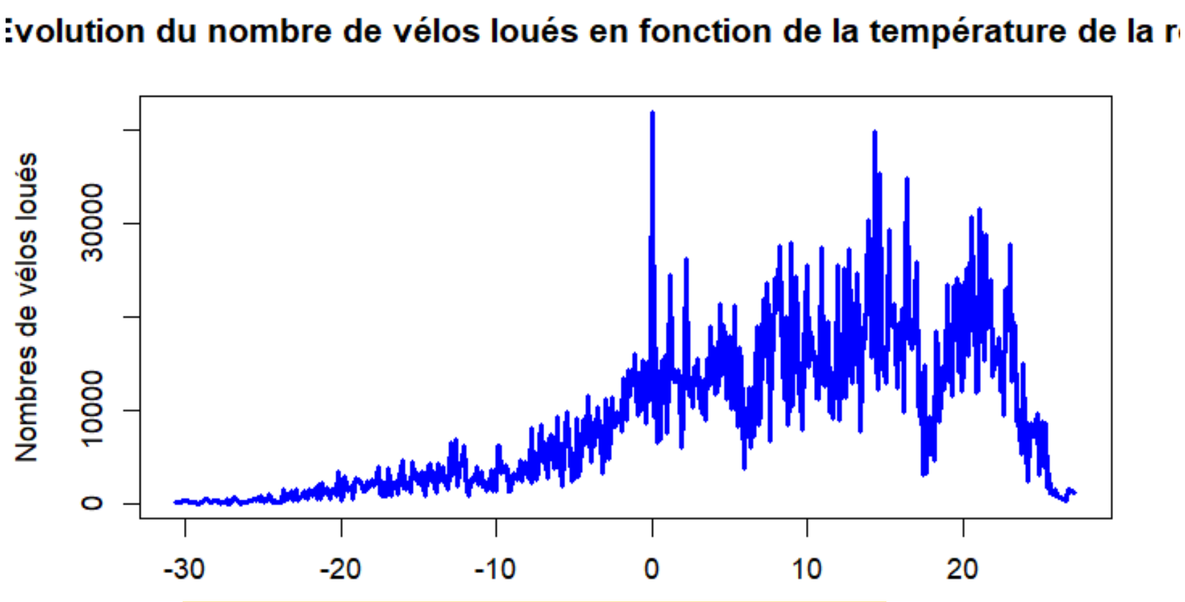


Variables sur le temps



La temperature

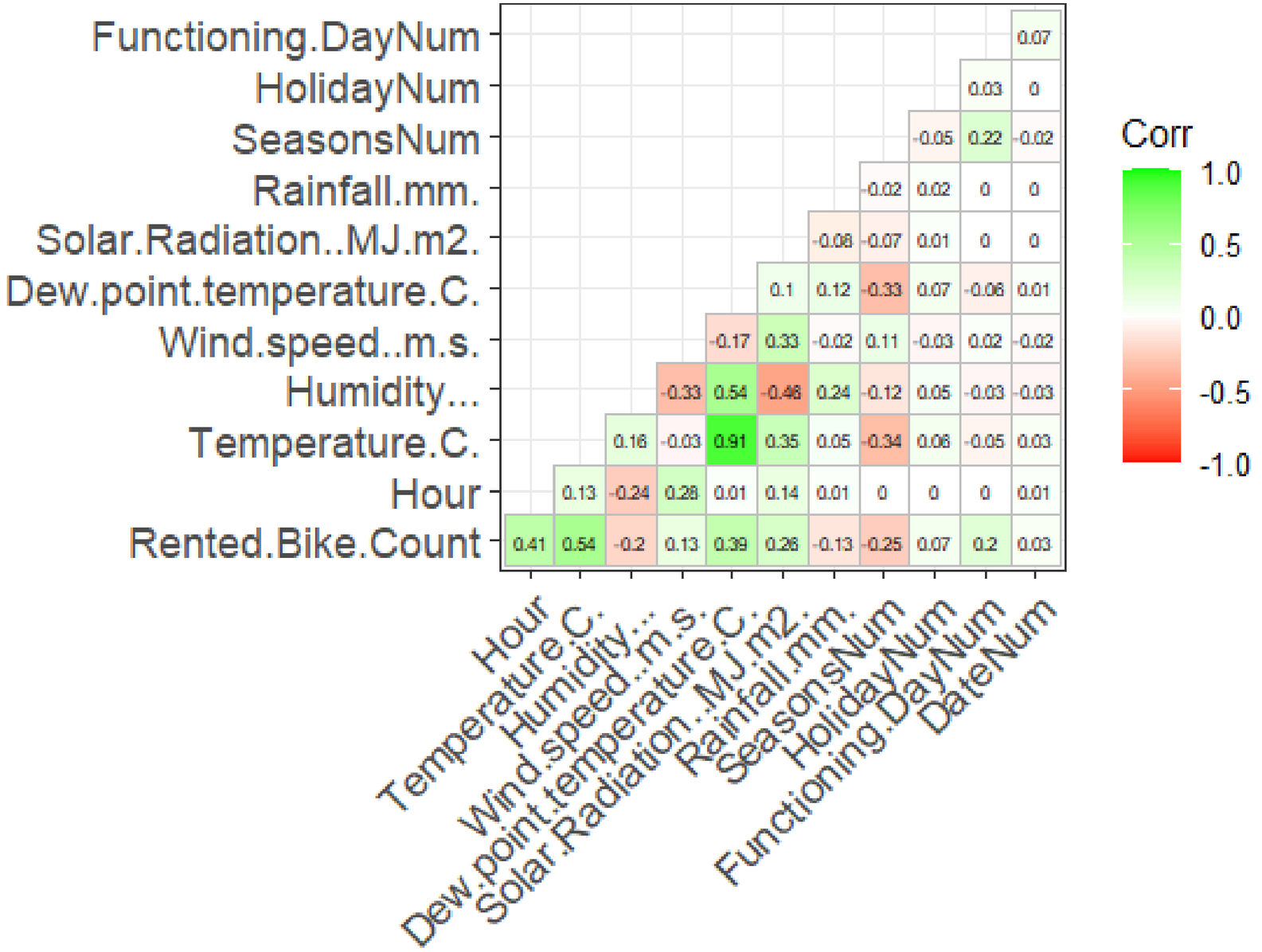
Locations de vélos est maximal autour de 22 à 25°C.



La température de la rosée

Relation croissante globale entre la la rosée et le nombre de vélos loués.

Matrice de correlation



Les variables les plus corrélées **positivement** au nombre de vélos loués sont :

- Température (0.54)
- Température du point de rosée (0.50)
- Rayonnement solaire (0.36)
- Heure (0.41)

Les variables corrélées **négativement** sont :

- Humidité (−0.36)
- Pluie (−0.25)

Régression Linéaire - choix du modèle

Comparaison des RMSE des modèles linéaires

```{r cars}
print(" model_all - Toutes les variables :")
sigma(model_all)
print("model_1 Uniquement les variables fortement corrélées avec notre variable cible :")
sigma(model_1)
print("model_2 Variables corrélées avec notre variable cible et variable les moins corrélées :")
sigma(model_2)
print("model_3 Variables peu corrélées avec notre variable cible :")
sigma(model_3)
```

```
[1] " model_all - Toutes les variables :"
[1] 432.5814
[1] "model_1 Uniquement les variables fortement corrélées avec notre variable cible :"
[1] 478.9277
[1] "model_2 Variables corrélées avec notre variable cible et variable les moins corrélées :"
[1] 475.6352
[1] "model_3 Variables peu corrélées avec notre variable cible :"
[1] 617.1721
```

```{r cars}
print("AIC")
sigma(AIC_model)
```

```
[1] "AIC"
[1] 433.2712
```

Comparaison du RMSE pour chacun des modèle de régression linéaire

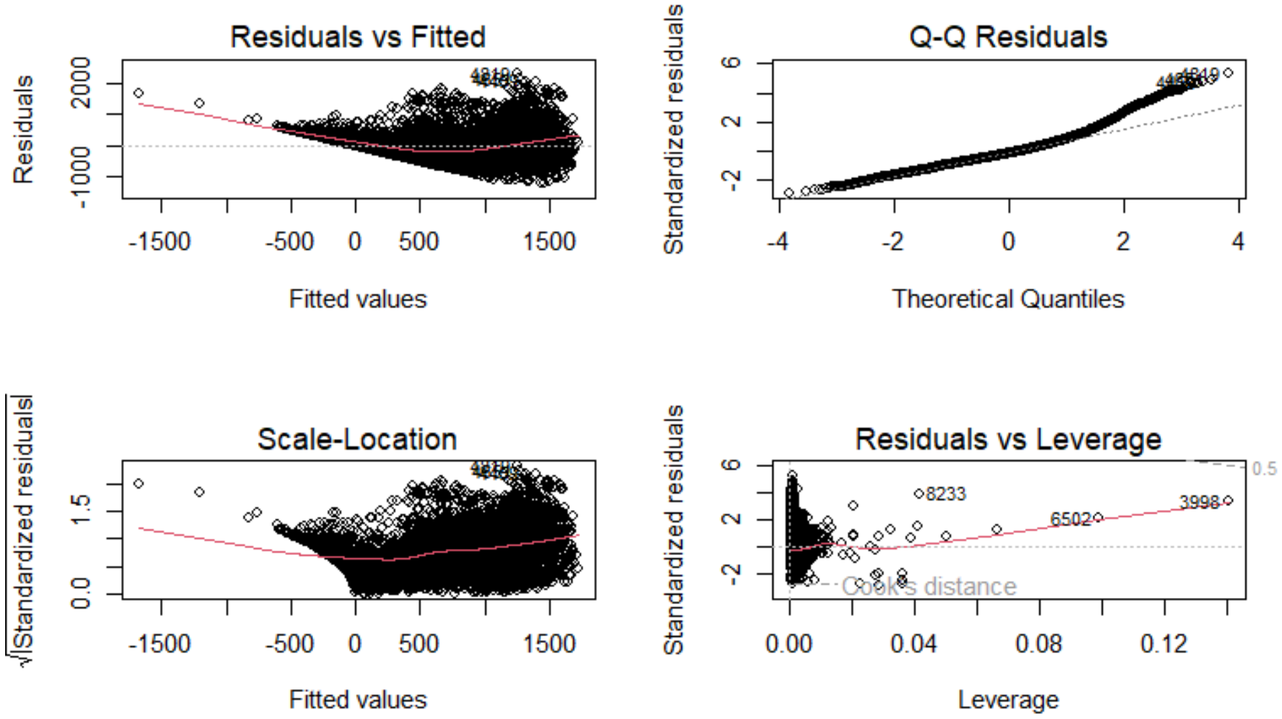
Différentes spécification du modèle ont été testées :

- toutes les variables
- avec les variable fortement corrélées avec notre variable cible
- avec les variables moins bien corrélées
- avec les variables les plus corrélées et les moins corrélées
- avec une AIC à partir du modèle prenant en compte toutes les variables

Choix du modèle : modèle avec toutes les variables

Verification de linéarité

Sur le modèle choisi



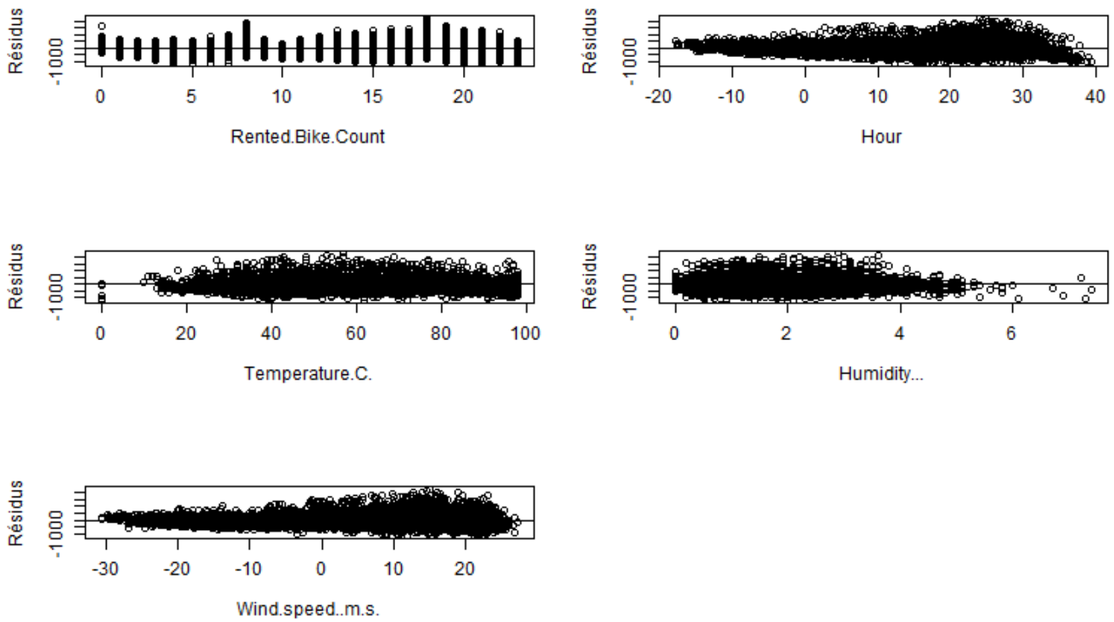
Les graphiques de diagnostic montrent que certaines hypothèses de la régression linéaire (linéarité, homoscédasticité) ne sont pas parfaitement respectées.

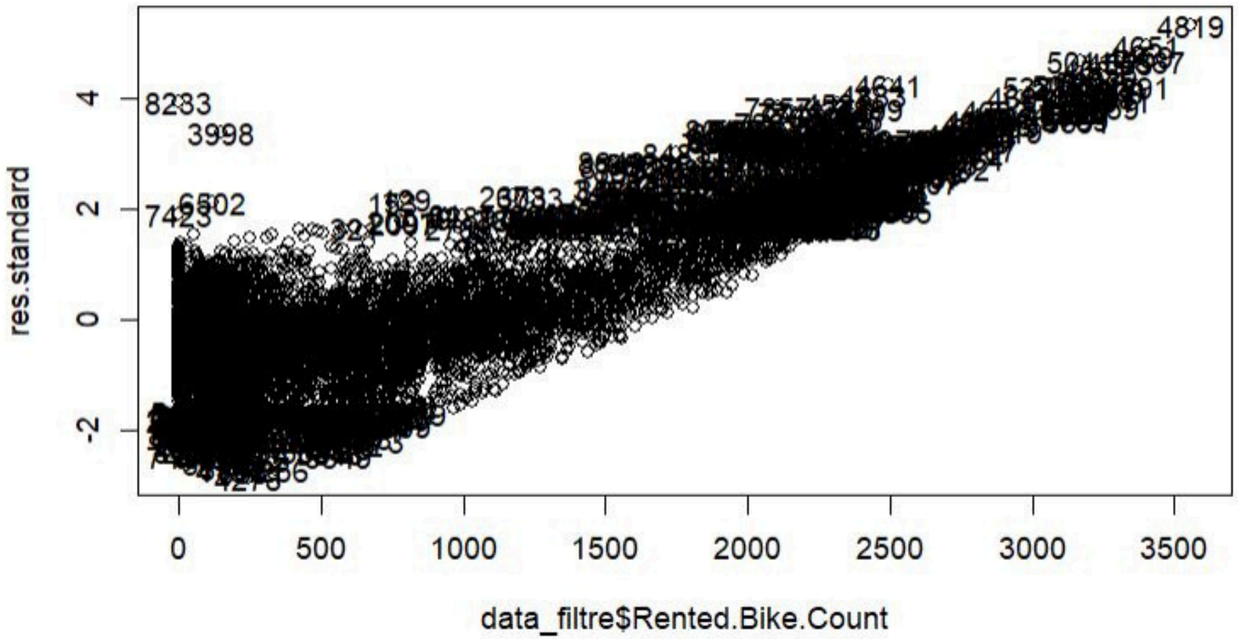
On observe des signes de non-linéarité, de variance non constante et la présence de points très influents.

Cela justifie l'intérêt de comparer le modèle linéaire à d'autres approches plus flexibles (arbres, forêts, KNN...).

Dispersion des résidus

Sur le modèle choisi





Les variables indépendantes

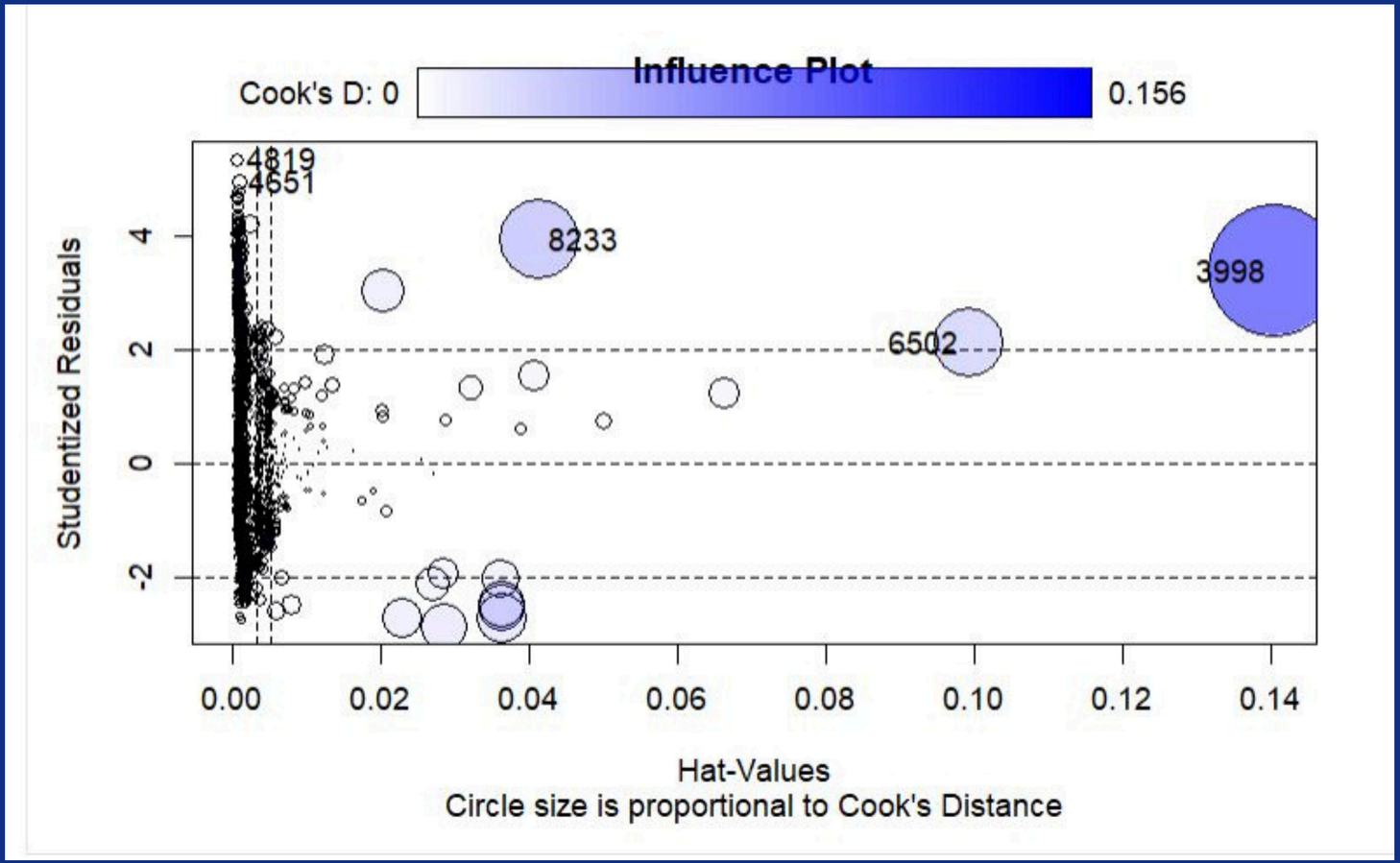
Variable cible - vélos loués

Distance de Cook

Sur le modèle choisi

Nous avons utilisé la distance de Cook pour détecter les points ayant un fort impact sur la régression.

Plusieurs observations (par exemple les lignes 3998, 8233, etc.) présentent une influence significative sur le modèle.



Conclusion du diagnostic post-regression

Grâce aux graphiques précédents, nous constatons la présence d’outliers qui influencent le modèle.
Nous émettons l’hypothèse que leur suppression améliorerai les performances du modèle.

Choix du modèle

```
{r cars}
print("modèle avec les outliers")
print("train")
sigma(model_all)
print("test")
sigma(model_test)
print("modèle sans les outliers")
print("train")
sigma(model_clean)
print("test")
sigma(model_test_clean)
```

```
[1] "modèle avec les outliers"
[1] "train"
[1] 421.6476
[1] "test"
[1] 426.4173
[1] "modèle sans les outliers"
[1] "train"
[1] 284.8649
[1] "test"
[1] 286.8765
```

En moyenne le taux d'erreur du modèle sans outliers est plus faible que celui avec les outliers.

De plus, nous remarquons que le modèle sans outliers se généralise mieux sur la base test que le modèle avec les outliers. L'écart du taux d'erreur moyen est plus faible sur la base test que sur la base entrainement.

Métrique	Valeur
R^2 (Multiple)	0.6806
R^2 ajusté	0.6779
RMSE *(erreur standard des résidus)*	284.9
F-statistic	249.8
p-value (global du modèle)	< 2.2e-16
Nombre de variables	10
Nombre d'observations	1406

Performance Regression linéaire

Le modèle explique environ 68 % de la variance du nombre de vélos loués. C'est un bon niveau de qualité de prédiction pour un phénomène complexe.

Le modèle est globalement **significatif**

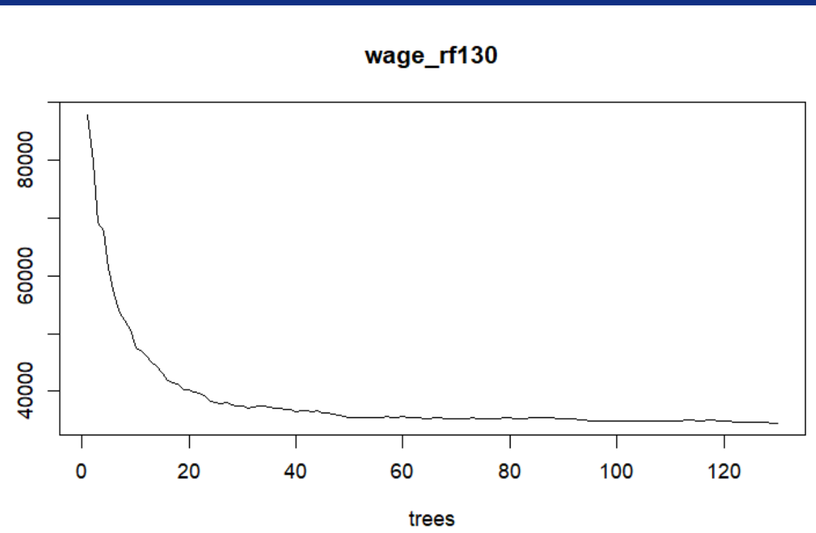
En moyenne, le modèle fait une erreur de ±284.9 vélos loués par heure

Les variables les plus significatives: conditions météorologiques (humidité, pluie, rosée), le jour (férié ou non), les saisons, les vacances et l'heure apporte une amélioration significative de l'explication de la variabilité de la variable dépendante

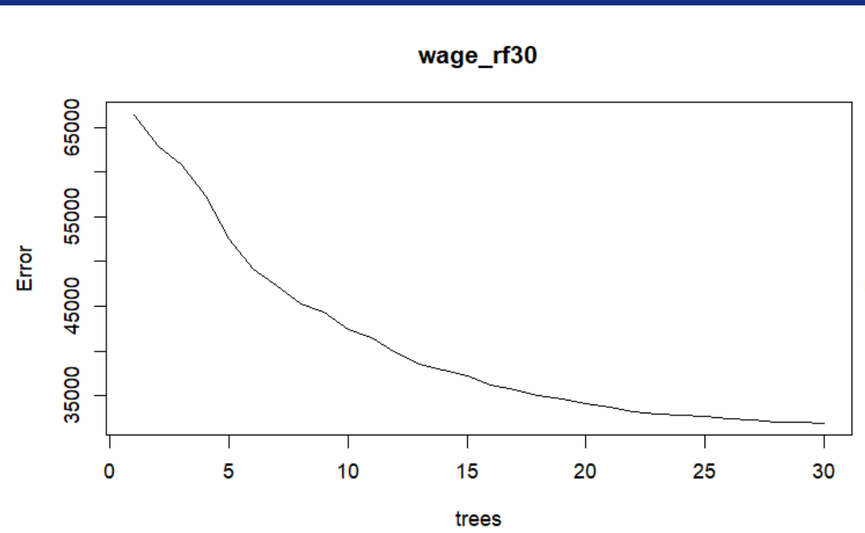
Random forest : Choix du nombre d’arbres

nodesize=5

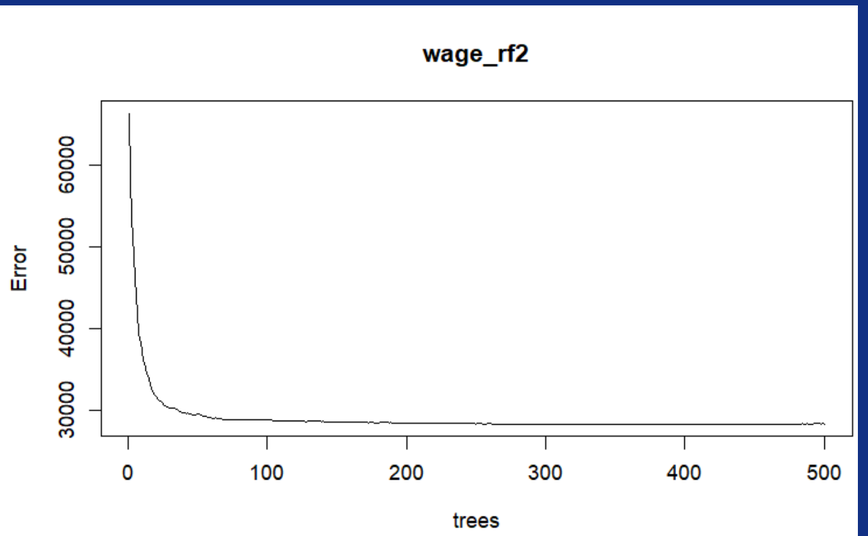
Stabilisation de l’erreur OOB selon le nombre d’arbres



130 arbres



30 arbres



500 arbres

- 500 arbres : stabilité du nombre d’erreur de prédiction rapide mais trop étalée
- 30 arbres : aucune stabilité
- 130 arbres : stabilité du nombre d’erreur de prédiction rapide et moins étalée que pour le random forest avec 500 arbres
- **Choix du modèle : 130 arbres**

Random forest : Choix du nombre d'arbres

Description: df [3 × 4]

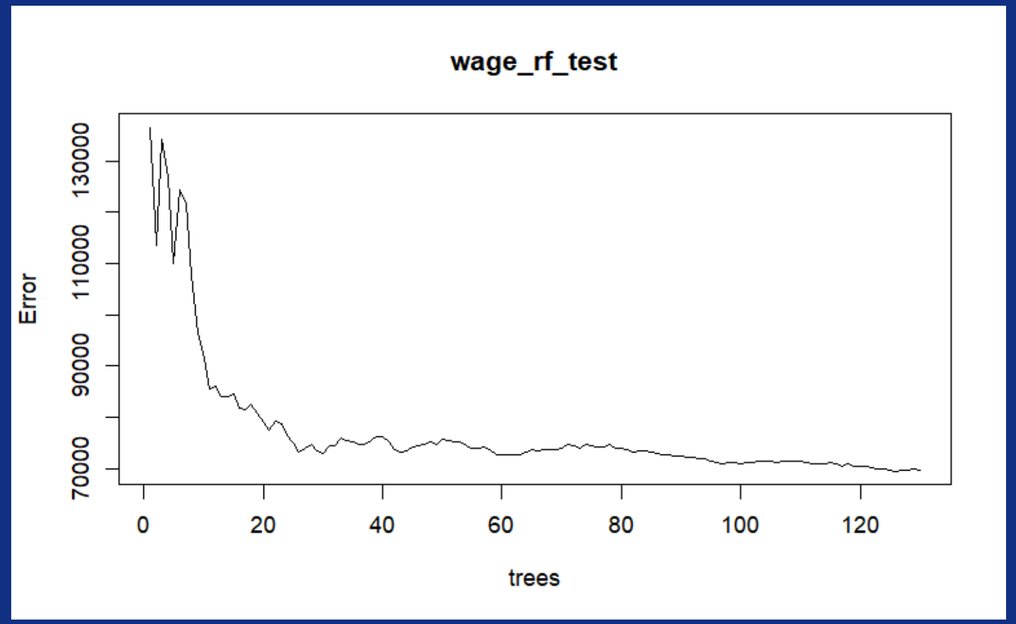
Nombre_arbre <chr>	RMSE <dbl>	R2 <dbl>	MAE <dbl>
500 arbres	189.3632	0.8575415	66.18086
30 arbres	50375.6114	0.7998675	68.47309
130 arbres	195.8381	0.8476327	65.97224

Nous n'observons pas d'écart significatif entre le RMSE, le R2 et le MAE du modèle composé de 500 arbres et celui de 130 arbres. Ainsi, les modèles composés de plus de 130 arbres sont légèrement plus performant que celui de 130 arbres.

La valeur du RMSE du random forest composé de 30 arbres s'écarte plus clairement de celle des deux autres modèles.

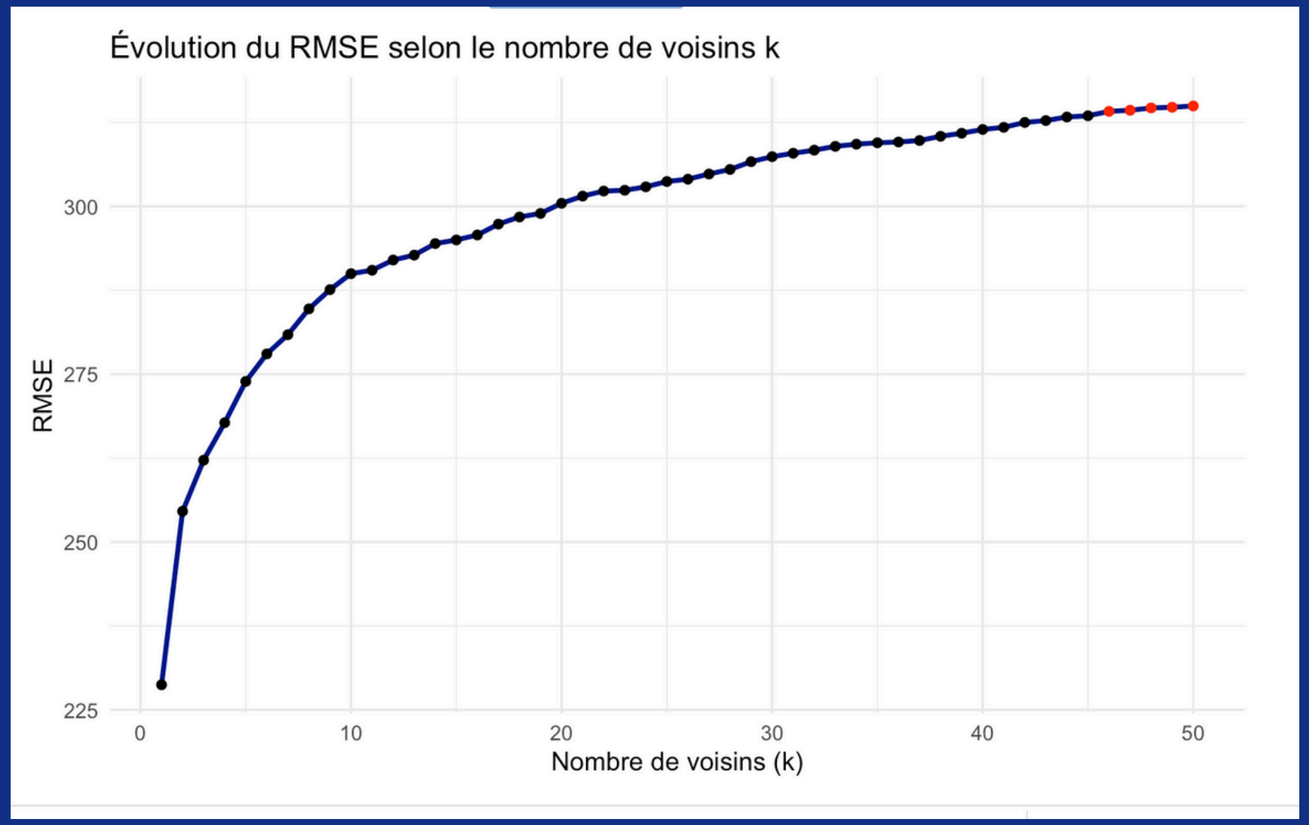
Random forest : Généralisation sur la base test

Stabilisation de l'erreur selon le nombre d'arbres



Les pics montrent le nombre d'erreur de prédiction dont la valeur est aberrante.
La courbe n'est pas lisse et semble se stabiliser difficilement.
Les erreurs sont causées en partie par la taille de la base test réduite de la base test par rapport à la base entrainement.
Le modèle se généralise mal sur la base test.
Nous pouvons y voir un surajustement.

KNN : Méthode “manuelle”

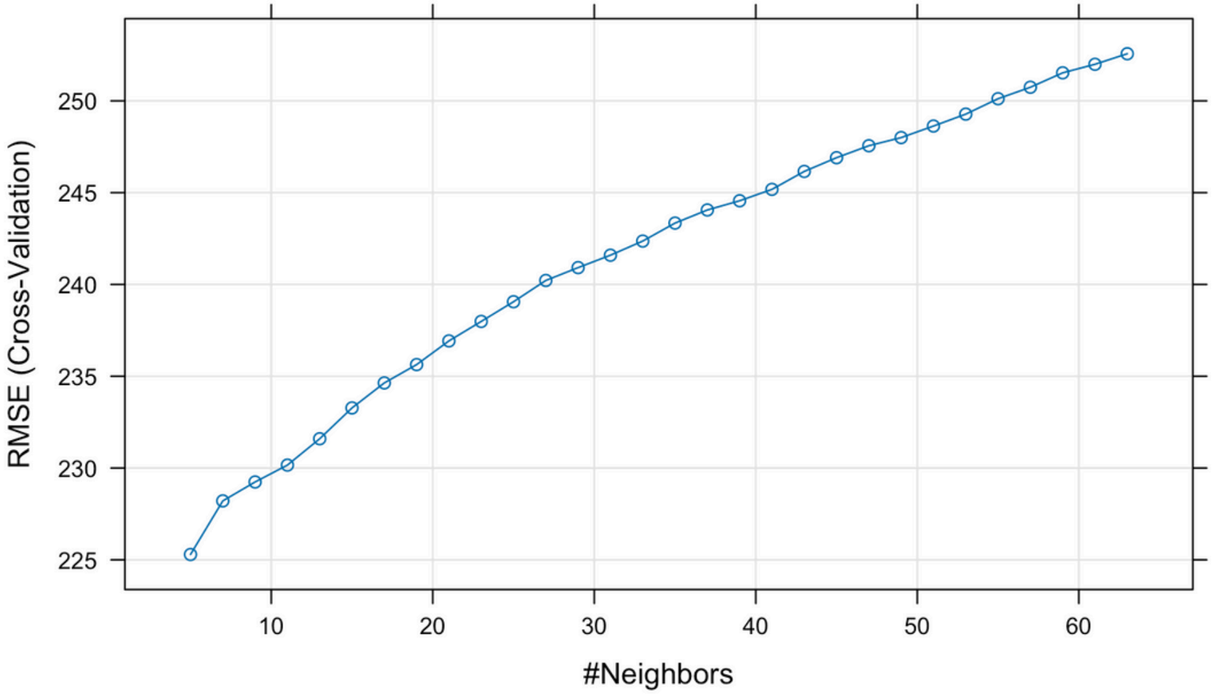


k <int>	RMSE <dbl>
1	228.74
2	254.58
3	262.19
4	267.77
5	273.92

- On teste différentes valeurs de k (1 à 50)
- Pour chaque k : modèle entraîné sur train_data, évalué sur test_data
- Le RMSE est le plus bas pour k = 1 : 228.74
- Mais risque de surapprentissage pour de petites valeurs de k
- Mais ce résultat dépend fortement du découpage train/test unique.
- Le RMSE augmente avec k → modèle devient plus stable mais moins précis

Résultat : R^2 = 0.825 ; RMSE = 228.74

KNN : Validation croisée 5-plis



k	RMSE
5	225.29
7	228.21
9	229.24
11	230.16
13	231.60
15	233.28

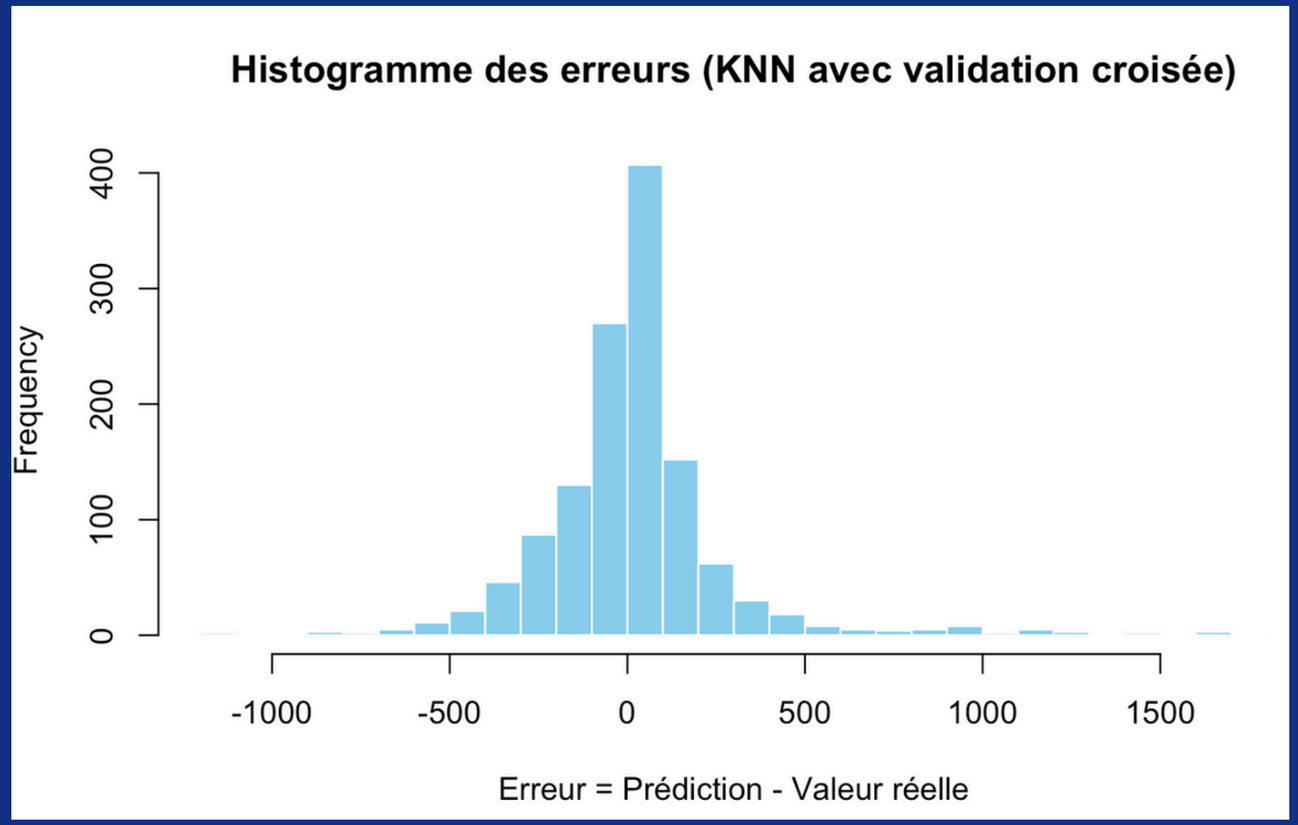
- On découpe le train_data en 5 blocs → CV 5-fold
- Chaque k est testé sur 5 itérations
- RMSE moyen minimum pour k = 5
- Plus fiable que la méthode manuelle : meilleure estimation de la performance
- Ne utilise pas le test final
- Elle garantit une meilleure généralisation.
- Le modèle le plus performant en validation croisée correspond à k = 5.

Résultat : k = 5, R^2 = 0.737, RMSE = 225.29

KNN : Évaluation finale du modele sur les données test

Méthode	k_optimal	RMSE	R2
KNN manuel (inspiré du cours)	1	228.74	0.825
KNN avec validation croisée (5-fold)	5	273.92	0.737

- Résultats calculés sur le jeu de test
- Le modèle manuel donne meilleurs scores, mais risque de surapprentissage
- On sacrifie un peu de performance brute, mais on gagne en stabilité.



- Histogramme : erreur centrée autour de 0 → pas de biais global
- Asymétrie à droite → le modèle sous-estime les grandes valeurs

Comparaison des modèles

Modele_test <chr>	Critères.optimisés <chr>	RMSE_test <dbl>	R2_test <dbl>	MAE_test <dbl>
Forêt aléatoire	nombre d'arbre = 130, nombre de noeuds =5	195.8381	0.7021110	158.9270
KNN (validation croisée) / regression	trControl = vc / 5 plis,tuneLength = 30	245.8591	0.7147423	177.5595
Regression linéaire	optimisation du nombre de variables	280.3082	0.7490000	219.8651

Modele <chr>	Critères.optimisés <chr>	RMSE <dbl>	R2 <dbl>
Forêt aléatoire	nombre d'arbre = 130, nombre de noeuds =5	180.5739	0.8777464
KNN (validation croisée) / regression	trControl = vc / 5 plis,tuneLength = 30	225.4825	0.8095398
Regression linéaire	optimisation du nombre de variables	299.1863	0.6644000

Régression linéaire: point de départ, mais limité dans les cas réels complexes.

KNN : bon compromis pour modéliser sans hypothèses, mais sensible au bruit et aux outliers, surtout pour k petit

Forêt aléatoire : modèle le plus performant dans notre cas, très robuste et efficace sur données réelles.

Importance des variables