Compte Rendu GLM

Kevin Wardakhan, Ibrahim Abdelatif, Erwan Ouabdesselam

2025-01-24

Sommaire

1 Introduction

Dans ce projet, nous avons cherché à comprendre les facteurs qui influencent le nombre de locations de vélos dans un système de partage urbain. En analysant un jeu de données détaillé, nous avons exploré les interactions entre plusieurs variables environnementales, temporelles et contextuelles, afin d'identifier celles qui impactent significativement la demande.

L'objectif principal de cette étude est de développer un modèle prédictif robuste basé sur des modèles linéaires généralisés (GLM). Ce modèle permettra non seulement d'expliquer les variations de la demande en fonction des conditions externes, mais aussi de fournir des prévisions fiables pour optimiser la gestion et la répartition des ressources dans ce type de système.

Ce rapport s'articule autour de plusieurs étapes clés :

1. **Préparation et exploration des données**, avec un accent sur le nettoyage et la compréhension initiale des variables.
2. **Étude des relations entre les variables**, afin d'identifier les corrélations significatives.
3. **Modélisation statistique**, qui vise à construire un modèle prédictif performant.

Interprétation des résultats, avec des recommandations pour la gestion opérationnelle des systèmes de partage de vélos.

1.1 Discussion sur la question d'intérêt

1.1.1 Pourquoi cette étude est-elle importante ?

L'analyse des systèmes de partage de vélos revêt une importance particulière pour plusieurs raisons :

1. **Impact environnemental :** Ces systèmes offrent une alternative aux modes de transport polluants, contribuant ainsi à réduire les émissions de gaz à effet de serre et à améliorer la qualité de l'air. Une compréhension fine des comportements d'utilisation permet d'encourager leur adoption, favorisant ainsi la transition vers des mobilités durables.
2. **Avantages sociaux :** Les vélos en libre-service offrent une option économique et pratique pour les déplacements urbains, en particulier pour les personnes sans accès à un véhicule personnel.
3. **Optimisation opérationnelle :** Identifier les facteurs qui influencent la demande peut permettre une gestion plus efficace, notamment en optimisant la répartition des vélos entre les stations, en planifiant leur maintenance, et en anticipant les pics de demande.

En répondant à ces enjeux, cette étude contribue à améliorer l'efficacité et l'accessibilité des systèmes de partage de vélos, tout en soutenant les politiques publiques en faveur des mobilités douces.

1.2 Description des variables prédictives et hypothèses associées

Dans cette étude, nous avons analysé plusieurs variables explicatives, regroupées en trois grandes catégories:

1.2.1 1. Variables environnementales

- **Température moyenne et ressentie :** Nous supposons que des températures agréables favorisent l'utilisation des vélos, tandis que des températures extrêmes (chaleur ou froid) pourraient la réduire.
- **Humidité :** Une humidité élevée, souvent associée à des conditions inconfortables, pourrait dissuader les utilisateurs.
- **Vitesse du vent :** Nous faisons l'hypothèse que des vents faibles ont peu d'impact, mais des vents forts pourraient réduire la demande.
- **Conditions météorologiques (pluie/neige) :** Ces conditions devraient entraîner une baisse marquée des locations.

1.2.2 2. Variables temporelles

- **Saison :** Les saisons chaudes (printemps, été) devraient afficher une demande plus élevée, contrairement aux saisons froides (automne, hiver).
- **Jour de la semaine :** Les week-ends pourraient enregistrer une demande accrue, en raison d'une utilisation davantage récréative.
- **Horaire :** Nous supposons que les heures de pointe (matin et soir) correspondent à des pics d'utilisation liés aux trajets domicile-travail.

1.2.3 3. Variables contextuelles

- **Vacances :** Pendant les vacances, la demande pourrait diminuer pour les trajets domicile-travail, mais augmenter pour les loisirs.
- **Jour ouvré ou non :** Une demande plus constante est attendue les jours ouvrés, contrairement aux jours non ouvrés où elle pourrait être plus variable.

1.3 Hypothèses globales

1. Les **conditions climatiques** (température, humidité, précipitations, vent) influencent directement la demande. Des conditions favorables encouragent les locations, tandis que des conditions défavorables les réduisent.
2. Les **facteurs temporels** (saisons, jours, horaires) déterminent des variations cycliques de la demande. Par exemple, la demande devrait être plus forte en été et pendant les heures de pointe.
3. Les **jours fériés ou périodes de vacances** pourraient présenter des schémas d'utilisation distincts, reflétant des comportements davantage orientés vers les loisirs.

2 Préparation et exploration des données

2.1 Chargement et nettoyage des données

Le jeu de données a été chargé à partir d'un fichier CSV contenant $n = 1817$ observations et 13 variables explicatives.

En analyse les données brutes nous avons décidé de retirers la première colonne, jugée inutile pour l'analyse. Les variables qualitatives (comme la saison ou le mois) ont été converties en facteurs pour permettre une analyse appropriée. Toutes les valeurs manquantes ont été identifiées pour garantir l'intégrité des données. De plus, les colonnes contenant des entiers ont été transformées en variables numériques afin d'améliorer leur compatibilité avec les outils d'analyse statistique.

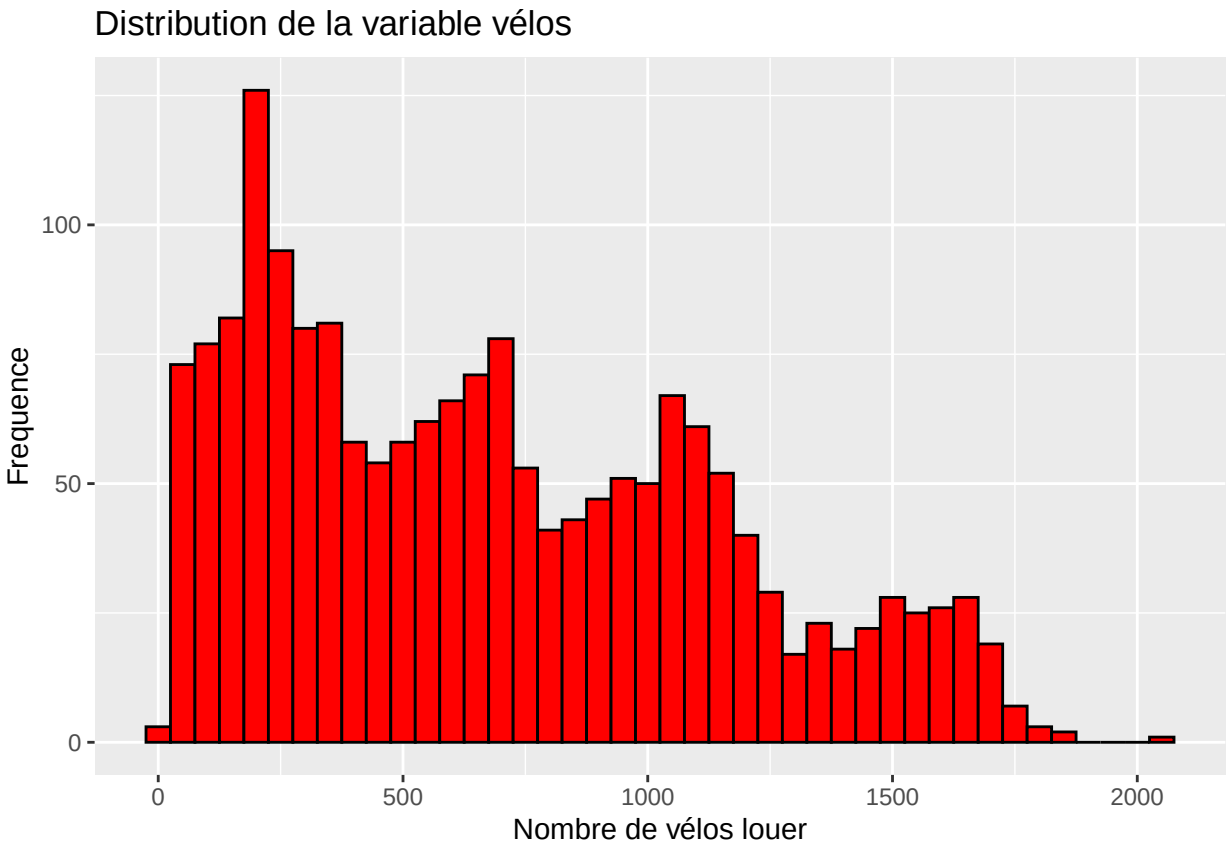
Ce prétraitement assure une structure cohérente et prête à l'emploi pour l'exploration et la modélisation des données.

2.2 Analyse descriptive :

2.2.1 Distribution des variables

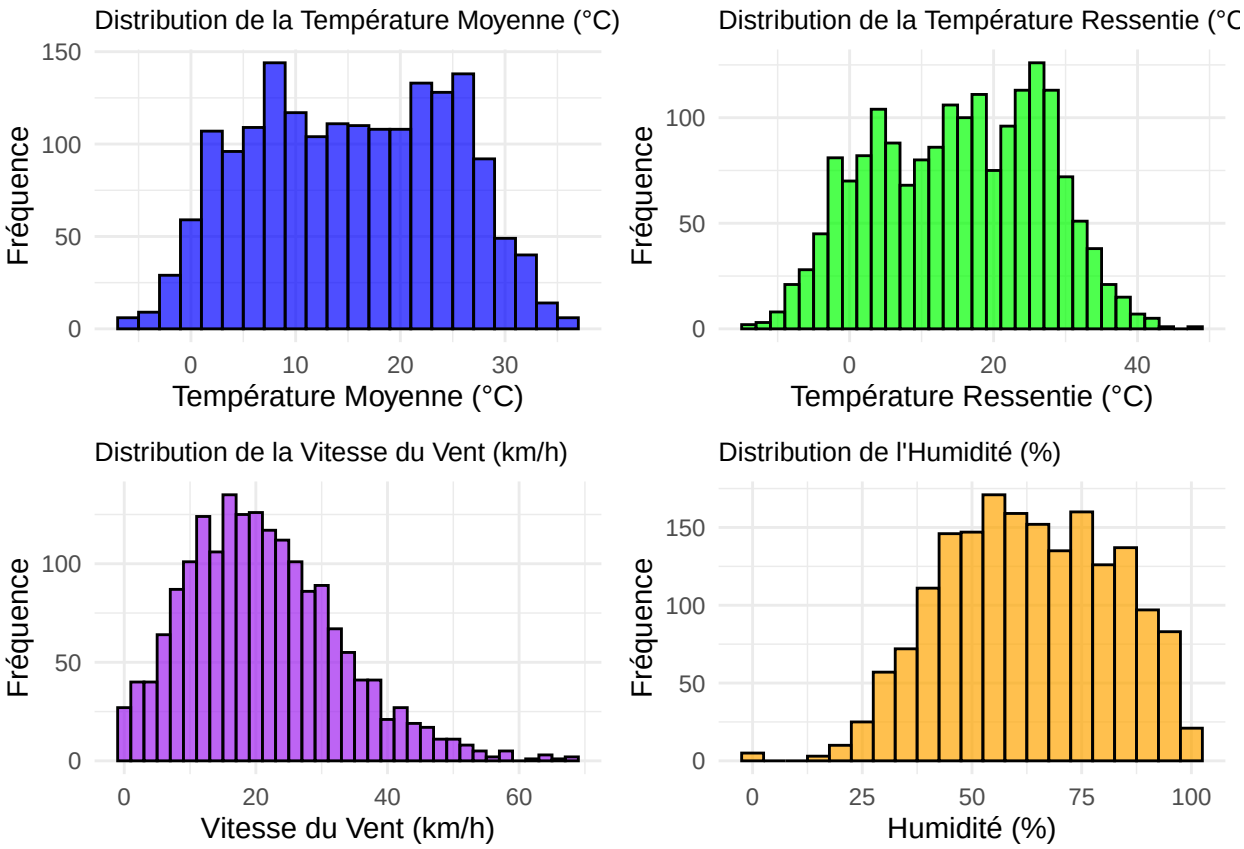
Nous avons généré des tableaux de fréquences pour analyser la répartition des variables qualitatives, tandis que des statistiques descriptives ont été calculées pour synthétiser les caractéristiques des variables quantitatives (par exemple, la température et l'humidité).

2.2.1.1 Distribution des locations de vélos On observe la distribution de la variable vélos à l'aide d'un histogramme.



La distribution des locations de vélos montre une concentration des fréquences entre 200 et 1500, avec un pic autour de 200-300. La fréquence diminue pour des valeurs supérieures à 1650, indiquant que des locations très élevées sont rares. La distribution est légèrement asymétrique, suggérant une variabilité dans la demande.

2.2.2 Distribution des autres variablees numériques



Ces graphiques illustrent les distributions des variables quantitatives clés du dataset.

1. **Température Moyenne (°C)**:
   - La distribution de la température moyenne est presque uniforme, avec des fréquences relativement similaires pour des températures comprises entre 5°C et 30°C.
   - Les températures extrêmes (inférieures à 5°C ou supérieures à 30°C) sont moins fréquentes.
   - Cela suggère que la majorité des observations ont lieu dans une plage de température modérée, ce qui pourrait refléter une utilisation accrue des vélos dans des conditions climatiques favorables.
2. **Température Ressentie (°C)**
   - La température ressentie suit une distribution similaire à celle de la température moyenne, ce qui est attendu étant donné leur forte corrélation.
   - Les fréquences augmentent entre 5°C et 20°C, puis diminuent au-delà de 30°C.
   - Cette répartition peut indiquer que les utilisateurs perçoivent des conditions similaires à celles mesurées, sans écart significatif.
3. **Vitesse du Vent (km/h)**
   - La vitesse du vent suit une distribution asymétrique, avec une forte concentration des valeurs entre 5 et 20 km/h.
   - Les vitesses supérieures à 40 km/h sont rares.

- Cela montre que la majorité des observations se produisent dans des conditions de vent modéré, probablement en raison de l'inconfort potentiel des vents forts pour les cyclistes.

4. **Humidité (%)**

   - L'humidité présente une distribution bimodale, avec des pics de fréquence autour de 50% et 80%.
   - Les valeurs très faibles d'humidité ($< 20\%$) sont rares, ce qui reflète probablement les conditions météorologiques locales.
   - Cette variabilité de l'humidité peut influencer les préférences des utilisateurs pour l'utilisation des vélos.

2.2.2.1 Observations générales

- Ces graphiques mettent en évidence des distributions cohérentes avec les données climatiques et météorologiques typiques.
- Les variables "température moyenne" et "température ressentie" montrent une plage de valeurs similaires, confirmant leur relation étroite.
- L'humidité et la vitesse du vent, bien qu'importantes pour le confort des utilisateurs, présentent des fréquences moins uniformes, ce qui peut nécessiter une analyse plus approfondie pour comprendre leur impact sur la demande en vélos.

Ces observations fournissent un aperçu des conditions environnementales rencontrées et leur potentielle influence sur l'utilisation des vélos, ce qui est essentiel pour les analyses suivantes et la modélisation.

2.2.3 Séparation des données pour le train et le test

Les données ont été divisées en ensemble d'apprentissage (70%) et de test (30%). Le modèle a été évalué à l'aide de l'erreur quadratique moyenne (MSE).

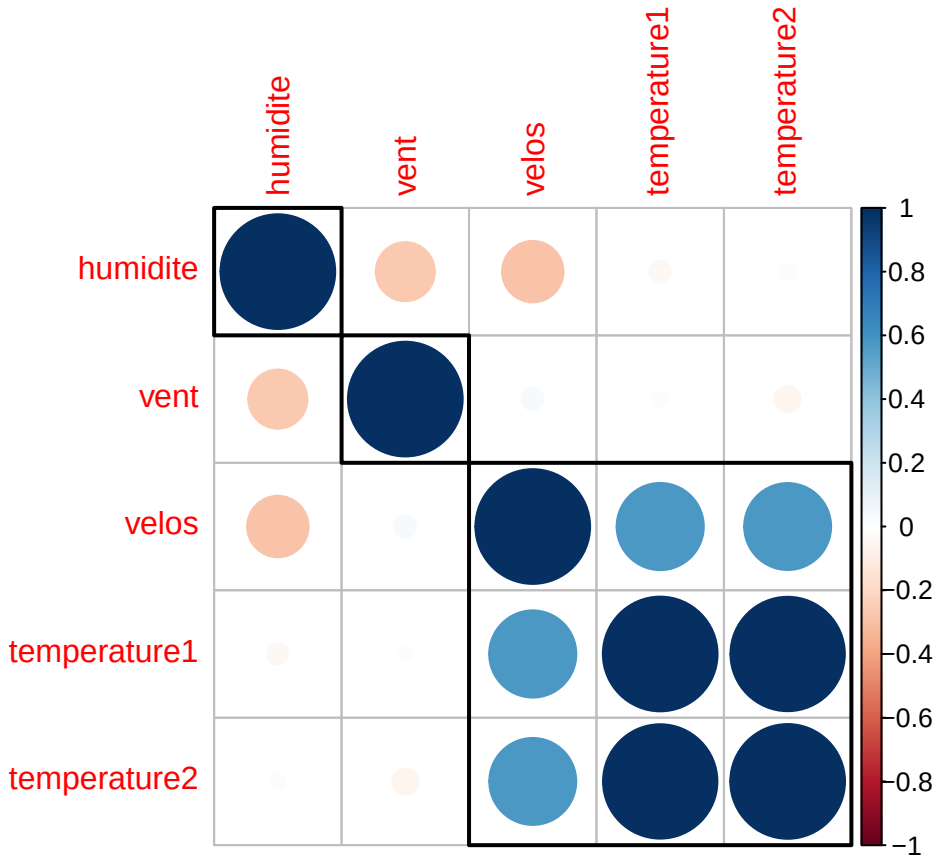
2.2.4 Analyse des corrélation entre variables quantitatives

Pour mieux comprendre les interactions entre les variables quantitatives du jeu de données, nous avons utilisé :

1. **Matrice de corrélation** : Elle met en évidence les degrés de corrélation entre les variables quantitatives en utilisant une représentation graphique intuitive.

Cette matrice permet d'identifier rapidement des relations linéaires potentielles entre les variables et d'évaluer la dispersion des données.

2.2.4.1 Matrice de corrélation



Le corrélogramme représente visuellement les corrélations entre les variables. Les cercles colorés indiquent l'intensité et la direction de la corrélation :

- Les cercles bleus signifient une corrélation positive,
- Les cercles rouges indiquent une corrélation négative,
- La taille et la couleur des cercles reflètent la force de la corrélation.

Résultats principaux :

Les analyses réalisées à l'aide de la matrice de paires et du corrélogramme ont mis en évidence des relations cohérentes et significatives entre les variables quantitatives :

1. **Corrélations principales :**
   - Le nombre de vélos loués (**vélos**) est **positivement corrélé** avec les températures (**corr 0.577**), indiquant que des températures plus élevées favorisent les locations.
   - Une **corrélation négative** est observée entre **vélos** et l'humidité (**corr -0.287**), ce qui montre que des conditions humides réduisent la demande en vélos.
   - La vitesse du vent présente une **corrélation négligeable** avec les locations (**corr 0.037**), suggérant que cette variable a peu d'influence.
2. **Relations spécifiques entre les variables :**
   - Les deux mesures de température (**température1** et **température2**) sont **très fortement corrélées** (**corr 0.994**), ce qui révèle une redondance. Une seule de ces variables pourrait être retenue pour la modélisation.

- Les tendances identifiées dans le corrélogramme sont visuellement confirmées par la matrice de paires, renforçant la compréhension des interactions entre les covariables.

2.2.4.2 Observations globales

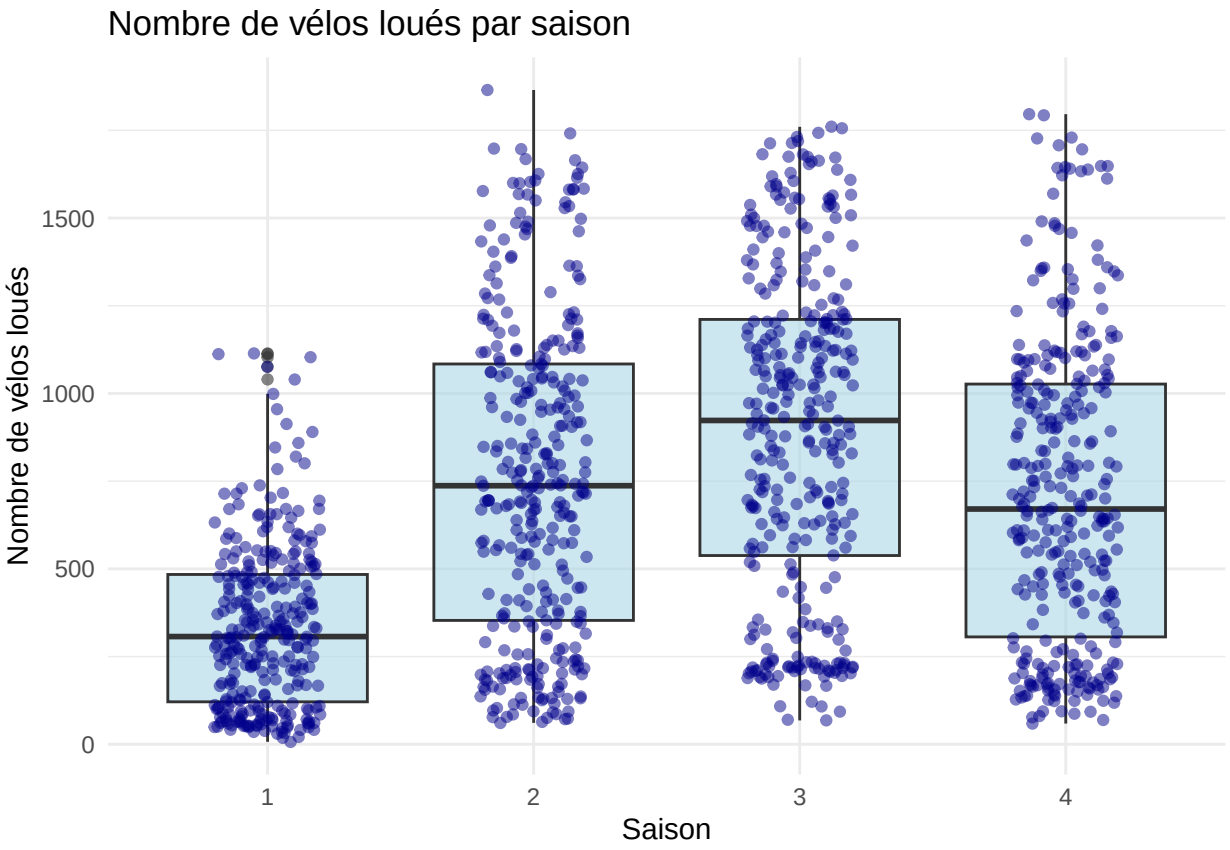
- Les variables climatiques, comme la température et l'humidité, apparaissent comme des facteurs déterminants de la demande en vélos.
- Les redondances entre certaines variables, comme les deux mesures de température, nécessitent une simplification ou une sélection avant la modélisation pour éviter une surcharge de variables corrélées.
- La vitesse du vent, avec son faible niveau de corrélation, semble avoir un impact limité et pourrait être considérée comme secondaire dans l'analyse.

En conclusion, cette analyse souligne l'importance des conditions climatiques, en particulier la température et l'humidité, sur l'utilisation des vélos. Ces résultats fourniront une base solide pour les prochaines étapes de modélisation, en identifiant les variables à privilégier pour construire un modèle prédictif robuste et fiable.

2.2.5 Analyse des relations entre les variables catégoriques et la cible

Pour analyser les relations entre les variables catégoriques et le nombre de vélos loués, des **boxplots** combinés à des points de dispersion ont été générés. Ces graphiques permettent de visualiser les tendances, la variabilité, et les éventuelles valeurs extrêmes selon différentes catégories.

2.2.5.1 Relation avec la saison



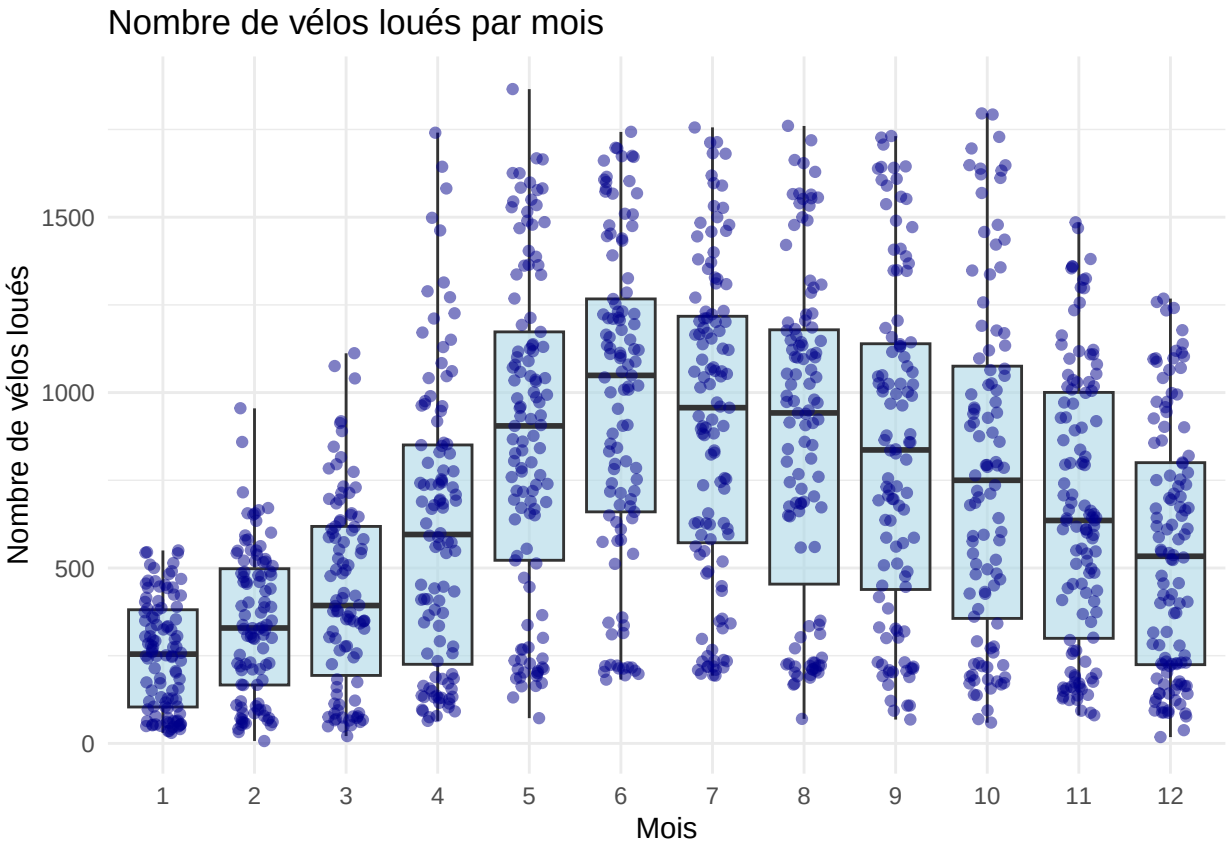
Tendances générales :

- Une demande plus faible est observée en hiver (saison 1), tandis que l'été (saison 3) et l'automne (saison 4) montrent des médianes plus élevées, reflétant une augmentation des locations pendant les saisons chaudes.

Variabilité :

- Une dispersion plus importante est visible en été et en automne, possiblement en raison d'événements ou de pics d'utilisation spécifiques.

2.2.5.2 Relation avec le mois



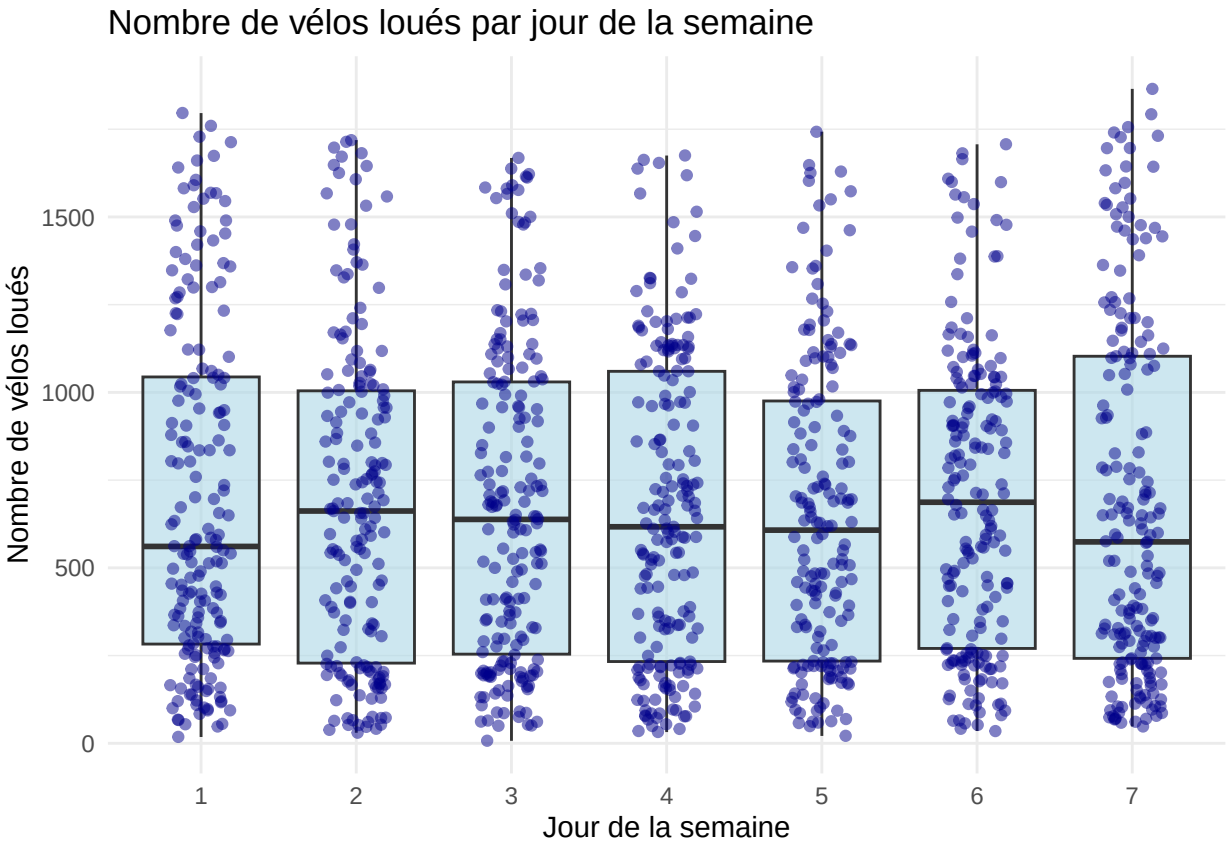
Tendances saisonnières :

- Les mois d'été (mai à septembre) montrent une augmentation significative des locations, tandis que les mois d'hiver (décembre à février) enregistrent une demande nettement plus faible.

Variabilité :

- Une plus grande dispersion est visible pendant les mois chauds, ce qui peut être attribué à une augmentation de la demande pendant des événements particuliers ou par beau temps.

2.2.5.3 Relation avec les jours de la semaine



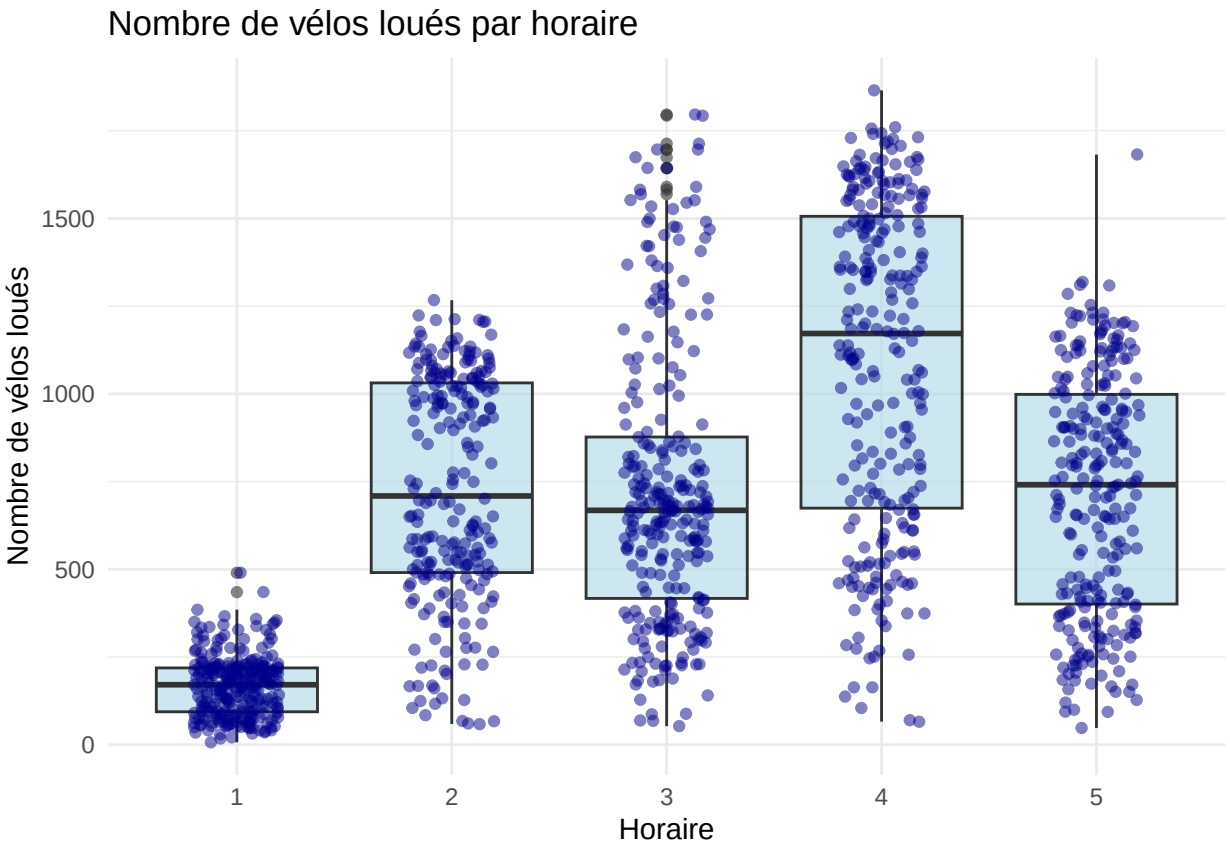
Tendances générales :

- Les médianes restent globalement stables tout au long de la semaine.
- Une légère augmentation est visible pendant le week-end (jours 6 et 7), indiquant un usage récréatif accru.

Dispersion :

- Les week-ends montrent une variabilité plus importante, probablement liée à des sorties de loisirs ou des activités spécifiques.

2.2.5.4 Relation avec les tranches horaires



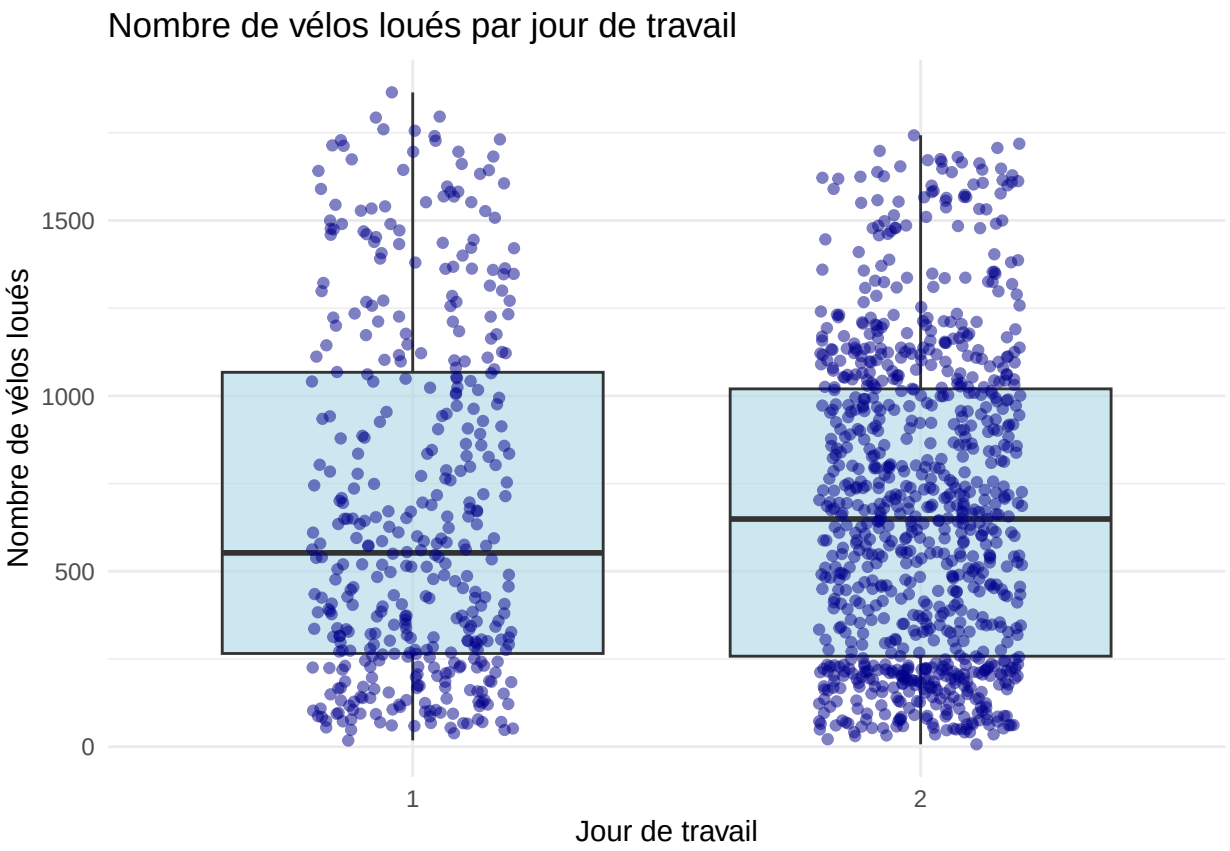
Tendances générales :

- Une faible demande est observée entre 0h et 7h (horaire 1), ce qui est cohérent avec les habitudes de sommeil.
- La demande augmente progressivement entre 7h et 19h, atteignant un pic entre 15h et 19h (horaire 4), correspondant aux retours de travail ou aux activités de fin de journée.

Variabilité :

- Les horaires 15h-19h montrent une dispersion plus importante, indiquant une utilisation à la fois régulière et opportuniste.

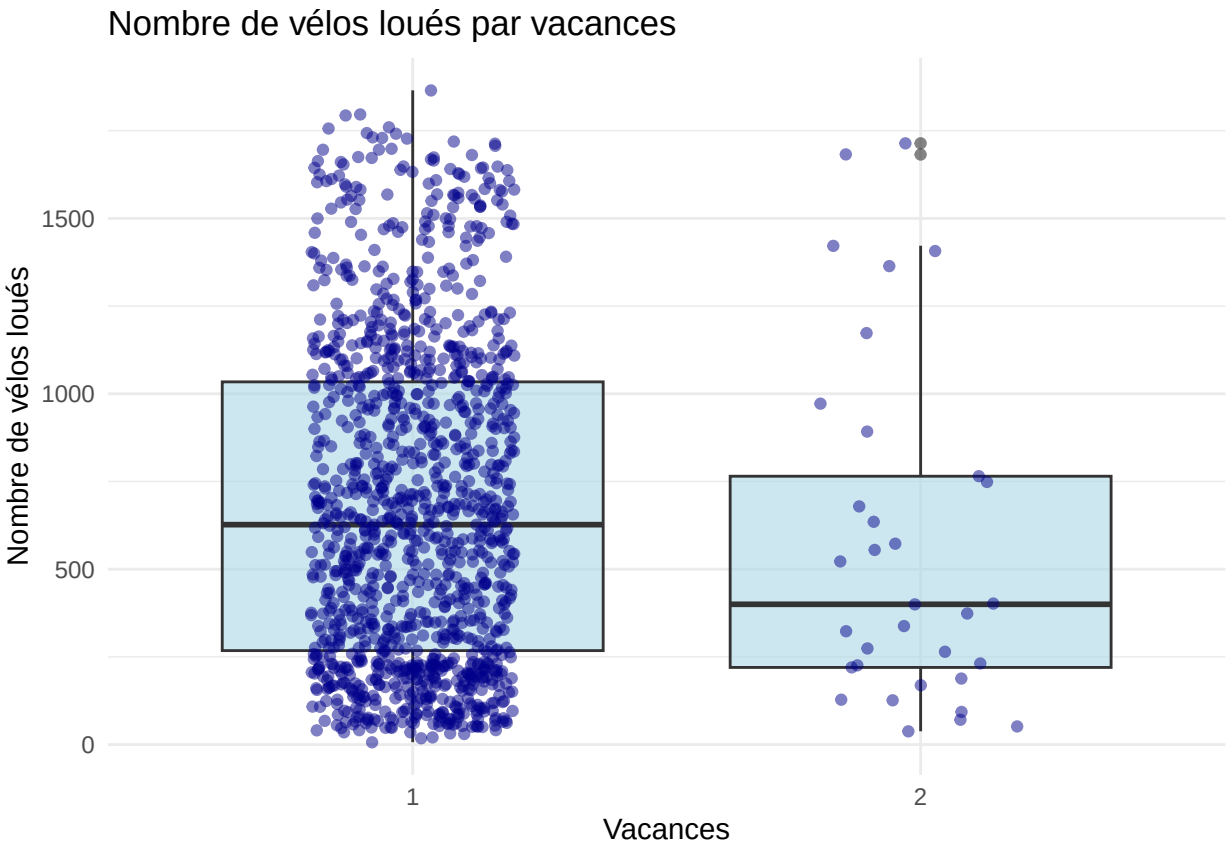
2.2.5.5 Relation avec les jours de travail



Comparaison semaine vs week-end :

- Les médianes sont similaires entre les jours de semaine (jour_travail = Oui) et les week-ends (jour_travail = Non).
- Les week-ends présentent une variabilité plus marquée, liée à une utilisation plus diversifiée.

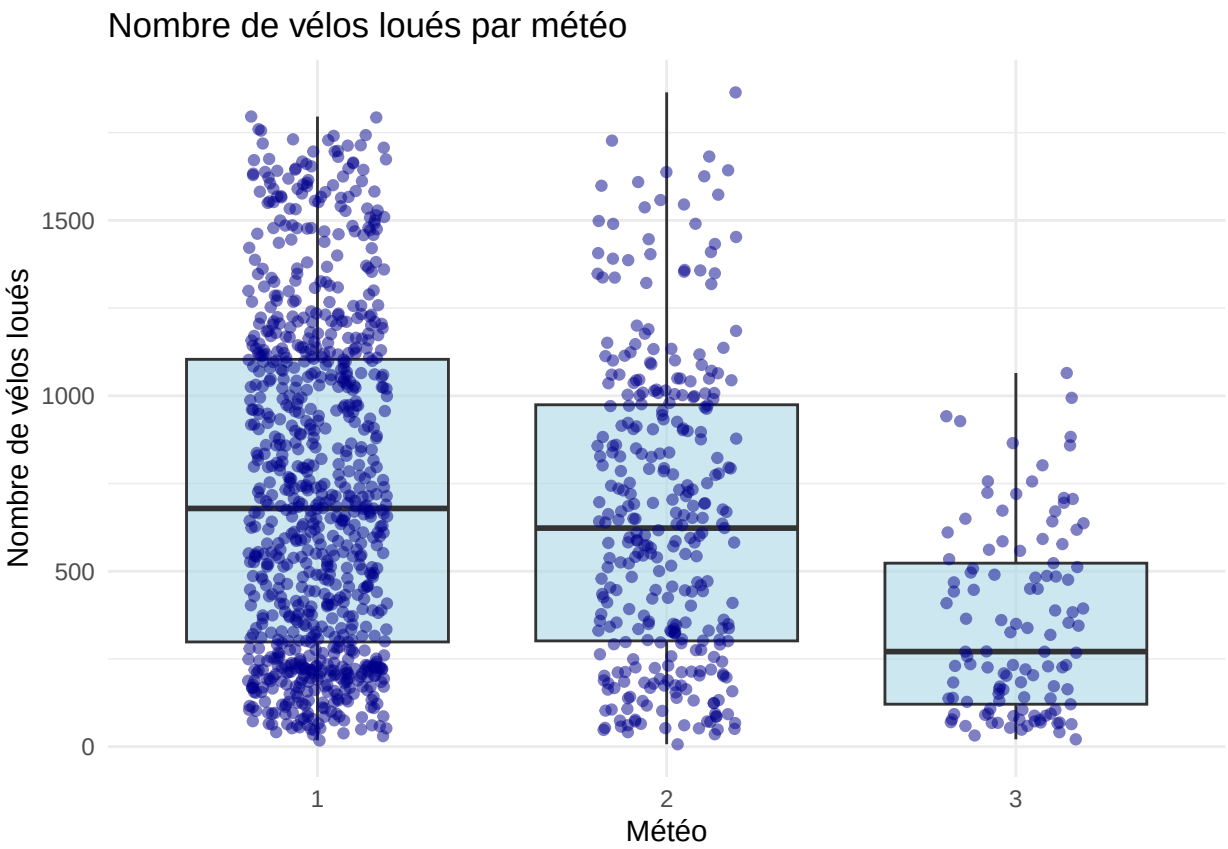
2.2.5.6 Relation avec les périodes de vacances



Comparaison vacances vs. non-vacances :

- Les locations de vélos sont nettement plus fréquentes et variables en dehors des périodes de vacances.
- Pendant les vacances, la demande est significativement réduite, avec des médianes plus basses et moins de dispersion.

2.2.5.7 Relation avec la météo



Comparaison des conditions météorologiques :

- **Temps clair** : Les locations sont les plus fréquentes, avec une forte dispersion, indiquant une utilisation accrue par beau temps.
- **Nuageux/Brumeux** : Une légère diminution de la demande est observée.
- **Pluie/Neige** : La demande chute fortement, avec des médianes basses et peu de variabilité.

2.2.6 Conclusion

Les analyses montrent que les variables catégoriques influencent significativement la demande en vélos :

- Les **saisons chaudes** et les **beaux jours** favorisent une utilisation accrue.
- Les **week-ends** et les **heures de pointe** affichent une demande plus élevée, tandis que les périodes de **vacances** et les **mauvaises conditions météorologiques** réduisent considérablement les locations.

Ces observations permettent d'adapter la gestion des vélos en fonction des périodes et des conditions spécifiques.

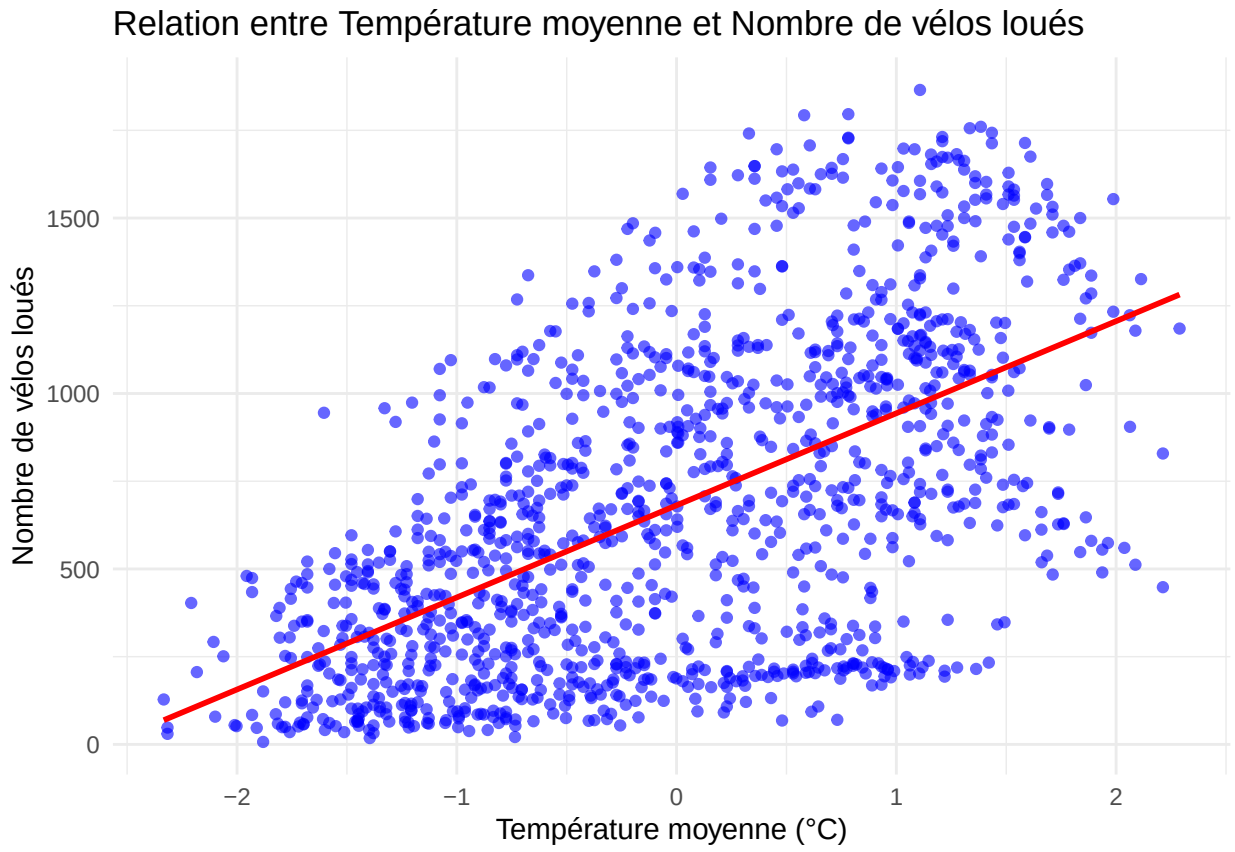
Voici une version améliorée et plus structurée de cette section, avec un langage plus clair et des conclusions synthétiques.

2.3 Étude des relations entre les variables quantitatives et la cible

L'objectif de cette analyse est d'identifier les relations entre les variables quantitatives (température, vent, humidité) et le nombre de vélos loués. Chaque graphique ci-dessous illustre une relation spécifique à l'aide

d'un nuage de points et d'une ligne de régression linéaire.

2.3.1 Relation entre la température moyenne et le nombre de vélos loués

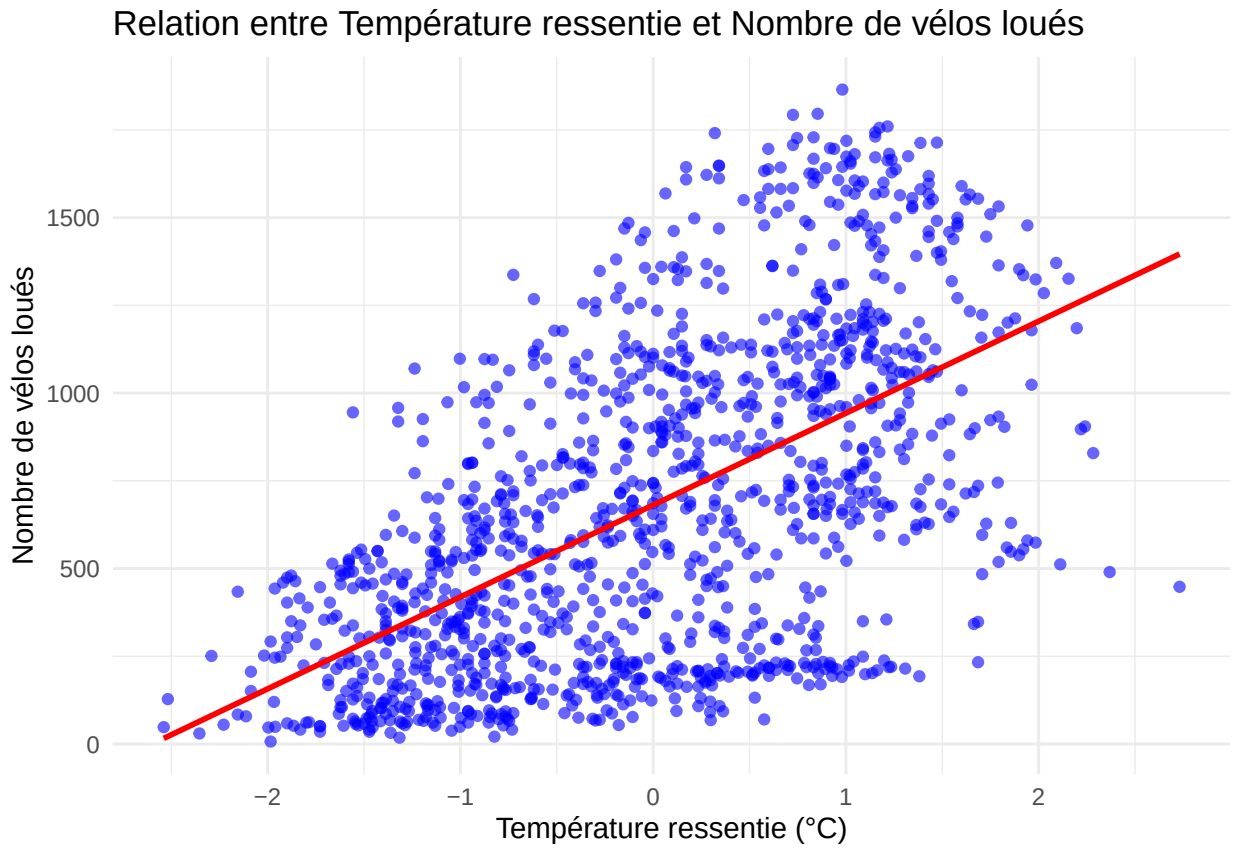


Analyse :

- **Tendance générale :** Une relation positive claire est observée : à mesure que la température augmente, le nombre de vélos loués augmente également.
- **Dispersion :**
  - À des températures basses (<10°C), la demande est faible et moins variable.
  - À des températures plus élevées (>20°C), la demande augmente significativement avec une plus grande variabilité, potentiellement influencée par des facteurs externes (par exemple, événements ou météo spécifique).
- **Ligne de régression :** La ligne rouge montre une tendance linéaire positive, confirmant que les températures plus élevées encouragent les locations.

En conclusion : La température moyenne est un facteur clé qui influence positivement l'utilisation des vélos. Ces informations sont particulièrement utiles pour prévoir la demande en fonction des prévisions météorologiques.

2.3.2 Relation entre la température ressentie et le nombre de vélos loués



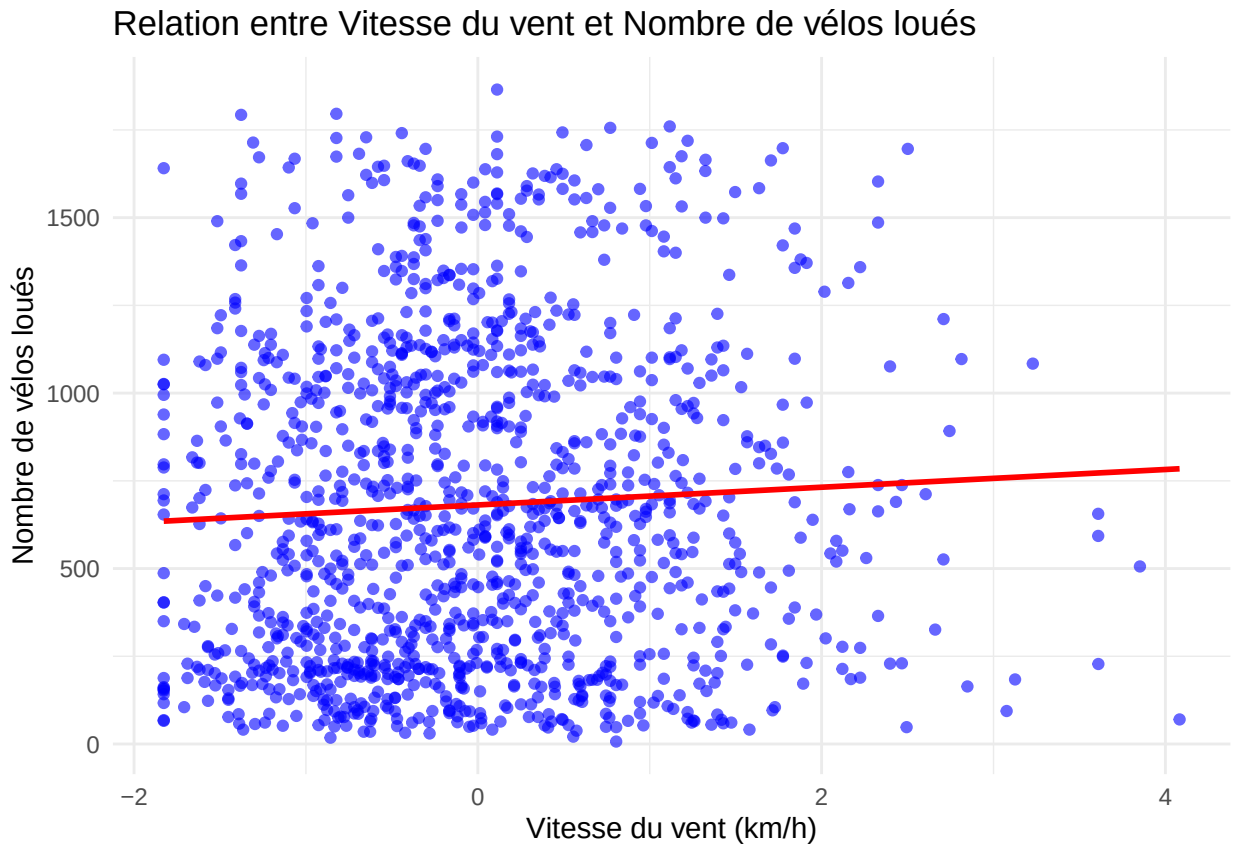
Analyse :

- **Tendance générale :** Une forte relation positive est visible : une augmentation de la température ressentie entraîne une augmentation du nombre de vélos loués.
- **Dispersion :**
  - Aux températures ressenties basses (<10°C), la demande est concentrée à des niveaux faibles.
  - À des températures plus élevées (>20°C), la demande augmente avec une variabilité accrue, ce qui peut refléter des effets saisonniers ou des conditions exceptionnelles.
- **Ligne de régression :** La ligne rouge confirme une corrélation positive significative.

Conclusion :

- La température ressentie est également un indicateur majeur des comportements des utilisateurs. Ces résultats complètent les observations sur la température moyenne et peuvent être exploités pour ajuster les ressources en fonction des conditions perçues.

2.3.3 Relation entre la vitesse du vent et le nombre de vélos loués

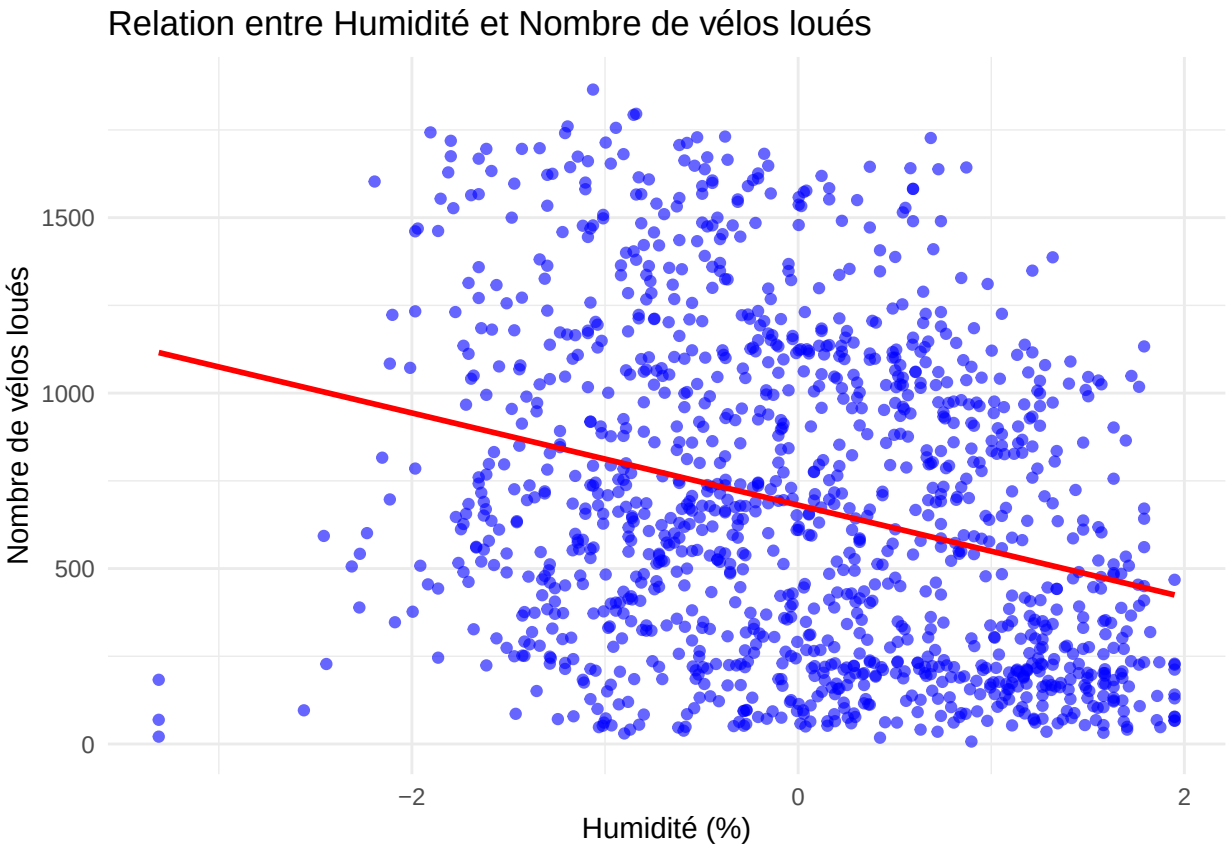


Analyse :

- **Tendance générale :** La ligne de régression montre une corrélation très faible, suggérant que la vitesse du vent a un impact négligeable sur les locations.
- **Dispersion :**
  - La majorité des observations est concentrée entre des vitesses faibles à modérées (0 à 20 km/h).
  - Aucune tendance significative n'est visible, quelle que soit la vitesse du vent.

En conclusion, la vitesse du vent a un effet limité sur la demande en vélos. Cette variable peut être considérée comme secondaire dans l'analyse et la modélisation.

2.3.4 Relation entre l'humidité et le nombre de vélos loués



Analyse :

- **Tendance générale :** Une relation négative est observée : à mesure que l'humidité augmente, le nombre de vélos loués diminue.
- **Dispersion :**
  - À des niveaux d'humidité faibles (<50 %), la demande est variée, avec des niveaux élevés possibles.
  - À des niveaux d'humidité élevés (>75 %), la demande diminue nettement et devient moins variable.
- **Ligne de régression :** La pente descendante de la ligne rouge confirme la corrélation négative entre humidité et locations.

En conclusion l'humidité est un facteur défavorable qui réduit l'utilisation des vélos. Ces informations peuvent être utilisées pour anticiper les baisses de demande pendant des conditions particulièrement humides.

2.3.5 Synthèse des observations

1. **Températures moyennes et ressenties :** Ces deux variables montrent des relations positives significatives avec le nombre de vélos loués, confirmant leur importance dans l'utilisation des vélos.
2. **Humidité :** La corrélation négative souligne son impact défavorable, limitant la demande en cas de forte humidité.
3. **Vitesse du vent :** Cette variable a un effet limité sur la demande, et son rôle peut être considéré comme négligeable dans les analyses futures.

Ces résultats mettent en évidence l'importance des conditions climatiques dans la gestion des vélos et permettront d'orienter les étapes de modélisation prédictive avec un focus particulier sur les températures et l'humidité.

Voici une proposition pour une conclusion générale de votre compte rendu :

2.3.6 Conclusion Générale

À travers une analyse approfondie des données nous avons explorer les principaux facteurs influençant le nombre de locations de vélos dans un système de partage urbain, plusieurs observations significatives ont été mises en lumière :

1. **Facteurs climatiques** :
    - La température joue un rôle majeur dans la demande en vélos, avec une relation positive claire entre des températures plus élevées et un nombre accru de locations.
    - À l'inverse, l'humidité a un effet négatif sur la demande, tandis que la vitesse du vent semble avoir un impact négligeable.
2. **Facteurs temporels** :
    - Les périodes estivales (mois d'été et saisons chaudes) connaissent une demande plus élevée, tandis que les périodes hivernales enregistrent une activité réduite.
    - Les jours de semaine et les horaires correspondant aux trajets domicile-travail (7h-11h et 15h-19h) montrent des pics d'activité, suggérant une forte utilisation pour des besoins pratiques.
3. **Relations entre variables** :
    - L'analyse des corrélations a révélé des redondances entre certaines variables, comme les deux mesures de température, nécessitant une attention particulière dans la modélisation afin d'éviter la multicolinéarité.

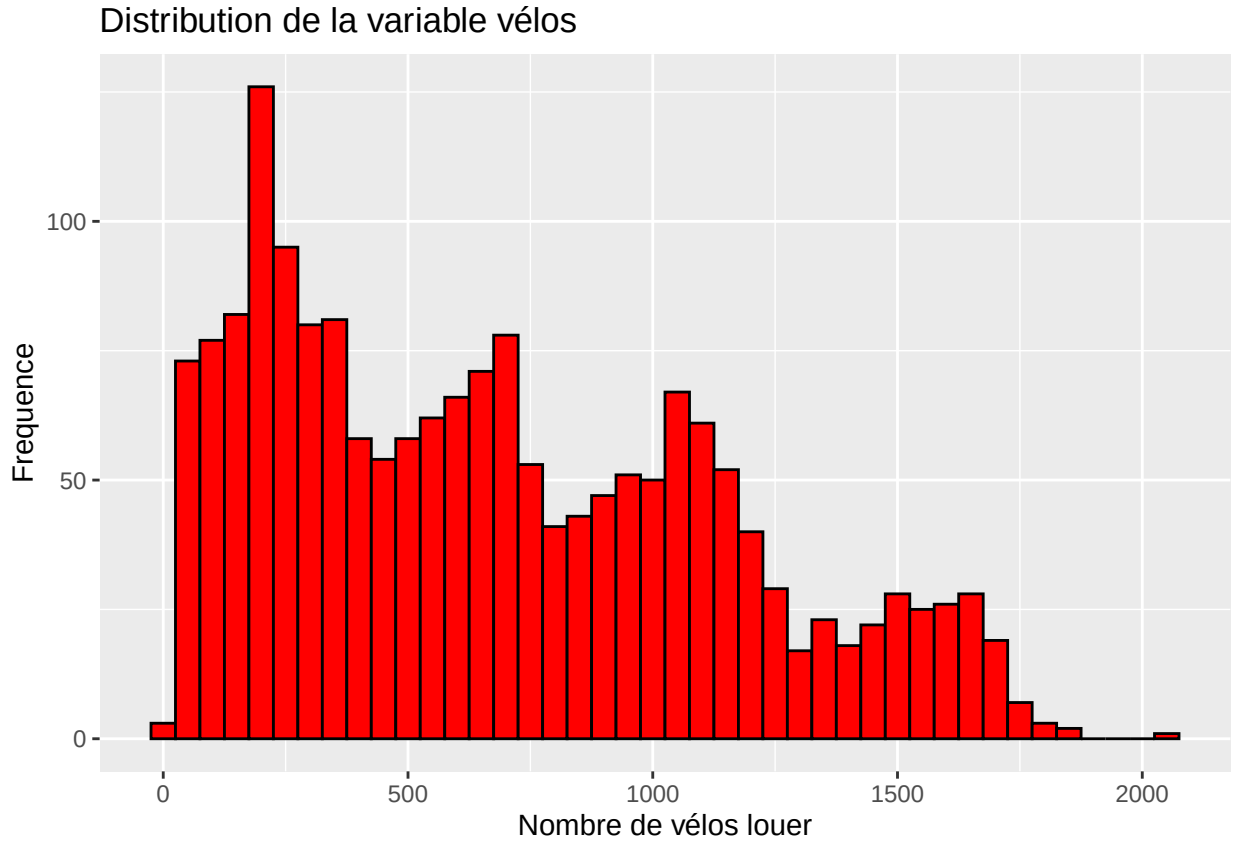
Ces observations ont permis de mieux comprendre les comportements des utilisateurs de vélos en libre-service et d'identifier les leviers clés pour anticiper la demande. Ces résultats serviront de base dans la suite pour le développement d'un modèle prédictif robuste, qui pourra être utilisé pour optimiser la gestion des ressources, notamment en ajustant l'offre de vélos en fonction des périodes de forte ou faible demande.

Enfin, cette étude met en évidence l'importance des données environnementales et temporelles dans la gestion d'un système de partage urbain. Elle ouvre également la voie à des études complémentaires pour intégrer d'autres paramètres, comme les événements locaux ou les infrastructures, dans les modèles prédictifs futurs.

3 Modèle linéaire généralisé

3.1 Transformation de la variable vélos

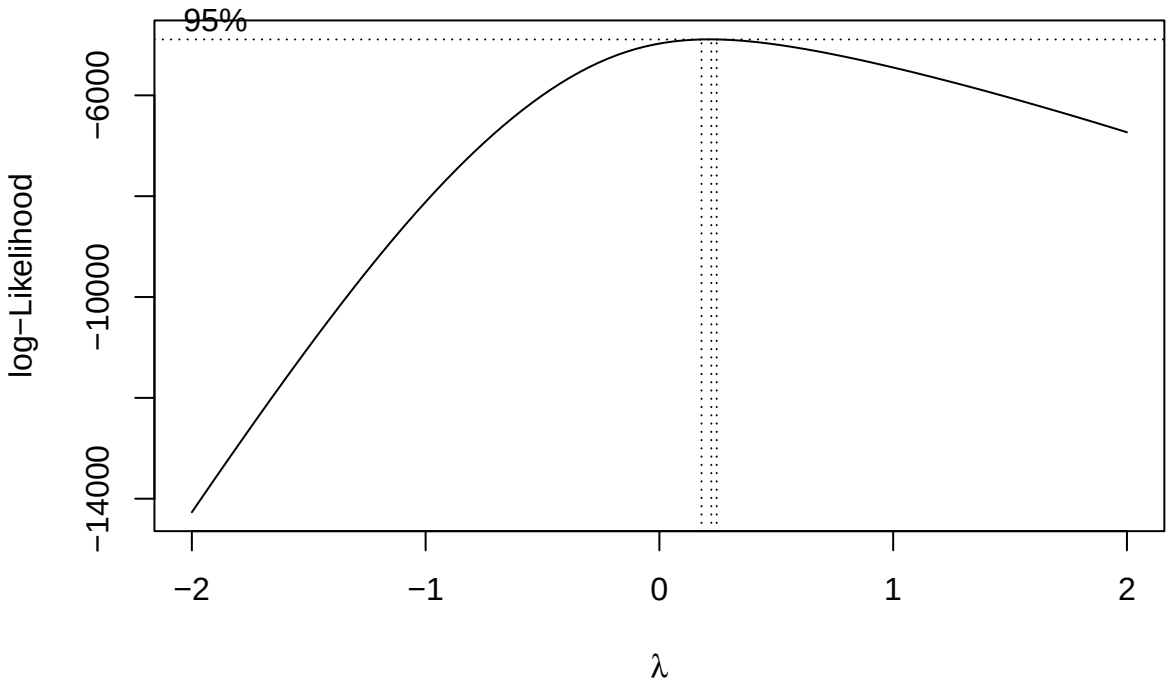
Dans cette section, nous avons fait le choix d'effectué notre régression linéaire avec un modèle gaussien. Néanmoins, la variable prédictive vélos ne semble pas suivre une loi normale si l'on s'appuie sur l'histogramme suivant:



Ainsi, nous allons effectuer une transformation de la variable vélos. Afin de savoir quelle transformation utilisée, et avoir avoir essayé avec la transformation $\sqrt{velos}$, nous avons découvert la transformation de *Box-Cox*, qui permet de trouver la transformation qui rend la variable la plus proche d'une loi normale. Celle ci est définie par la formule suivante:

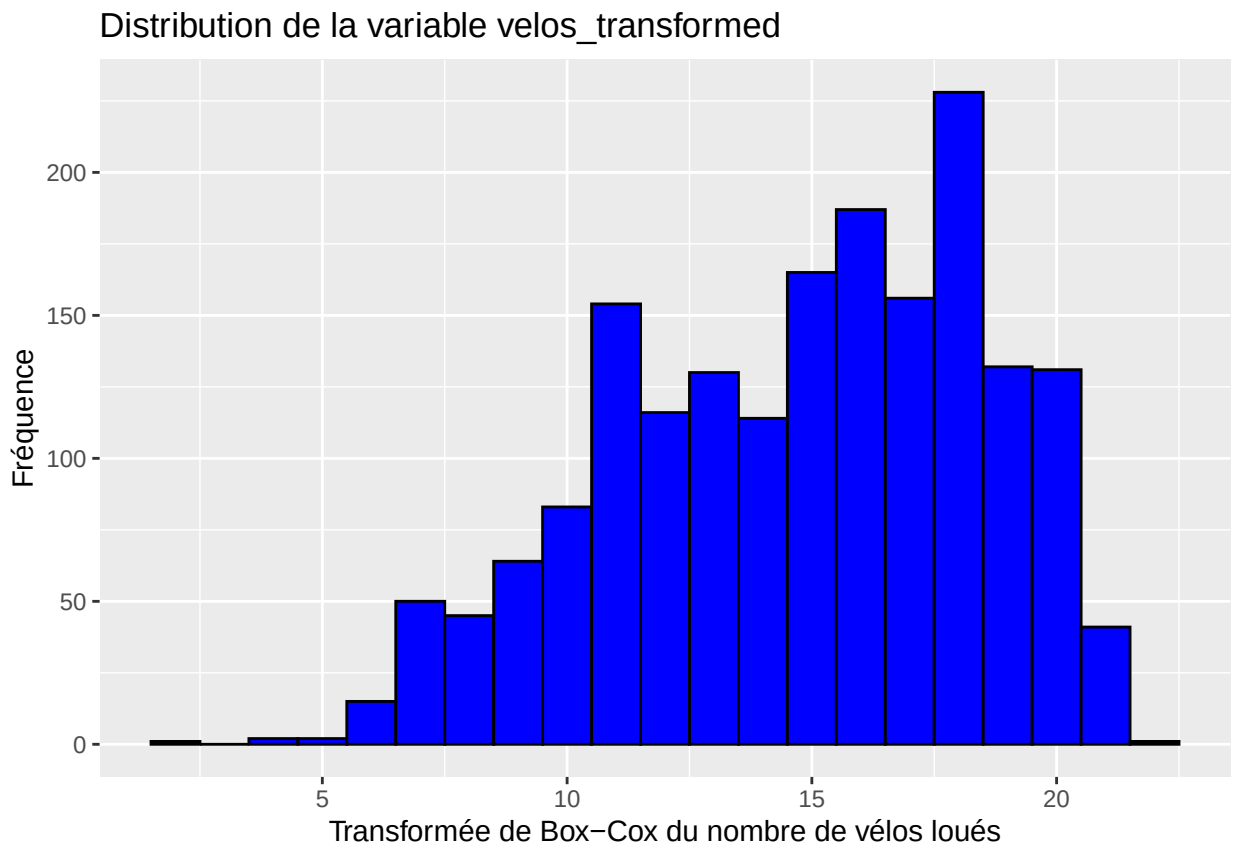
$$B(x, \lambda) = \frac{x^\lambda - 1}{\lambda}$$

si $\lambda \neq 0$ et $log(x)$ si $\lambda = 0$. Pour trouver notre λ, nous utilisons la fonction *boxcoc* du package *MASS*.



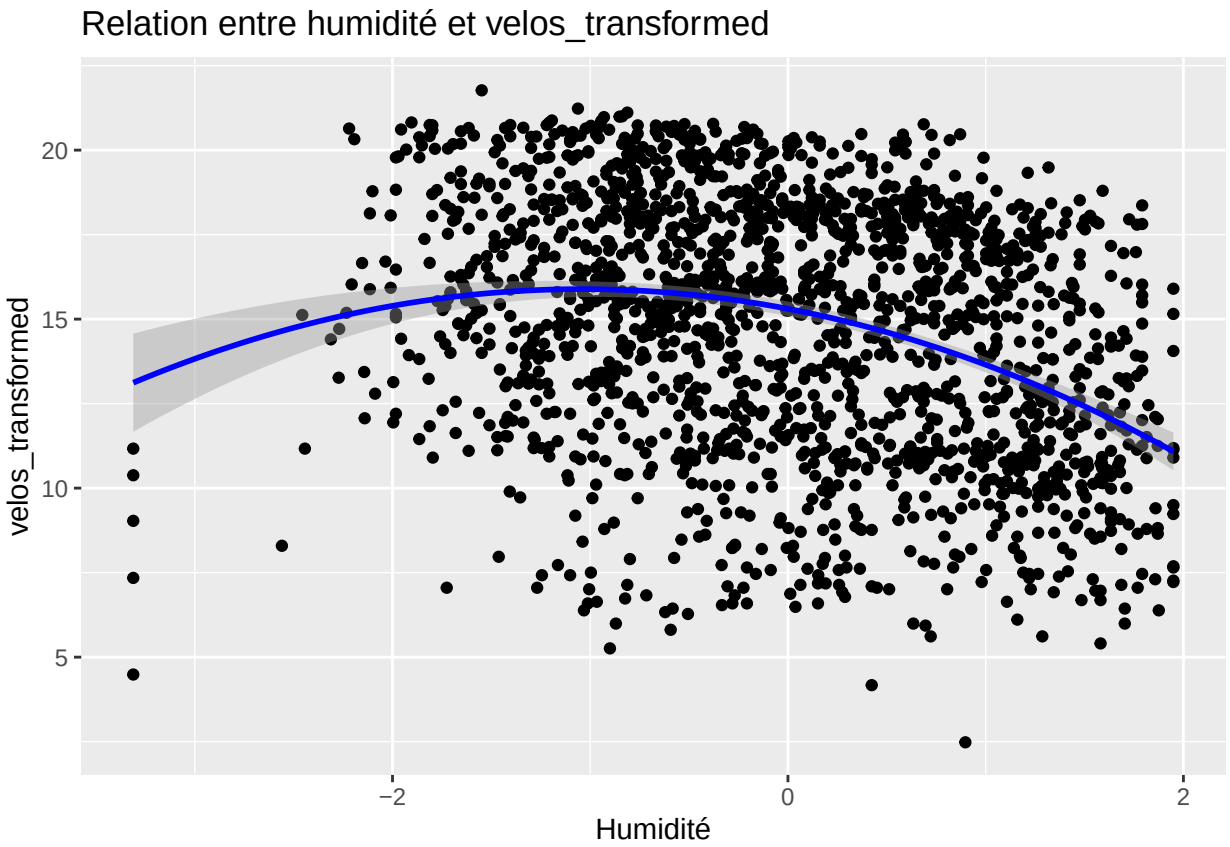
Ainsi, le λ qui maximise la vraisemblance est de 0.24. Nous allons donc effectuer la transformation suivante: $velo\_transformed = \frac{velo^{0.24}-1}{0.24}$

L'histogramme suivant nous montre la distribution de la variable *velo_transformed*.



3.2 Transformation de la variable humidité

Lorsque l'on représente la relation entre les variables *humidité* et *velos_transformed* à l'aide d'un nuage de points, on obtient:



Le graphique laisse à présager une relation quadratique entre ces deux variables. Ainsi, on introduit dans notre data set la variable *humidite2* qui correspond à $humidite^2$.

4 Construction du modèle final

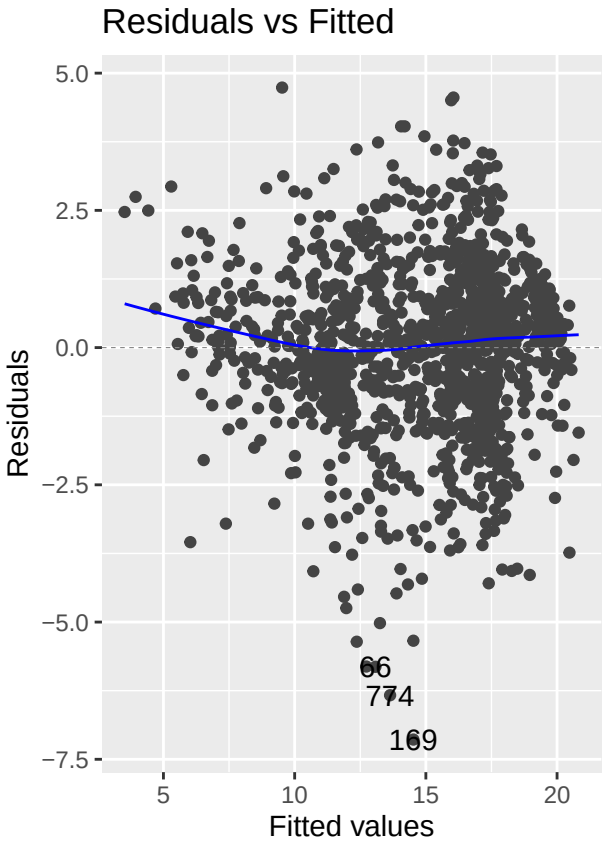
Afin de construire notre modèle, nous allons tout d'abord éliminer les covariables qui ont peu d'influcent sur le nombre de vélos loués. Nous allons donc utiliser les méthodes forward, et backward, en prenant en compte les critères *AIC*, *BIC*, et le *CP de Mallow*.

Nous allons partir pour les méthodes forward du modèle réduit à l'intercepte (*modele_intercept*), et pour les méthodes backward du modèle saturé (*modele_full*).

4.1 Validation du modèle gaussien

Nous allons passer en revue tous les postula du modèle gaussien, et nous allons vérifier si ceux-ci sont respectés.

4.1.1 [P1] Les résidus sont centrés



On observe ici que les résidus sont centrés autour de 0, et qu'il n'y a pas de tendance linéaire. Ainsi, le premier postulat est respecté. Contrairement à lorsque nous avions utilisé la variable *velo*, où une forte tendance était observée.

4.1.2 [P2] Les résidus sont homoscédastiques

Afin de vérifier si les résidus sont homoscédastiques, nous allons effectuer un test de *Breusch-Pagan*.

Ici, la p-value est de 65.3% ce qui est supérieur à 5%, ce qui signifie que les résidus sont homoscédastiques.

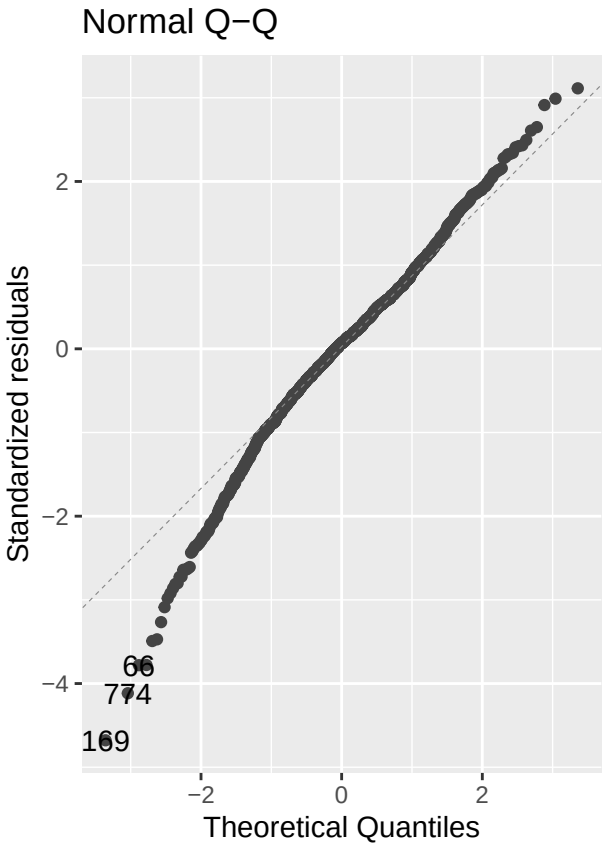
4.1.3 [P3] Les résidus sont décorélés

Afin de vérifier si les résidus sont décorélés, nous allons effectuer un test de *Durbin-Watson*.

La p-value est de 97.4%, ce qui est supérieur à 5%, ce qui signifie que les résidus sont décorélés.

4.1.4 [P4] Les résidus sont gaussiens

Afin de vérifier si les résidus suivent une loi normale, nous allons effectuer un *Q-Q-plot*.



Malheureusement, le postulat [P4] ne semble pas être modifié malgré la transformation de notre variable. Nous allons tout de même continué notre étude, afin de voir si les prédictions sont néanmoins correctes.

4.2 Méthodes Step-by-step

Avant d'implémenter nos modèles, on effectue un *test de type I anova* afin de connaitre les covariables qui sont susceptibles d'être éliminées.

```
## Analysis of Variance Table
##
## Response: velos_transformed
##                Df Sum Sq Mean Sq F value               Pr(>F)
## saison          3   3773    1258  502.61 < 0.0000000000000002 ***
## mois           11    930      85   33.80 < 0.0000000000000002 ***
## jour_mois      30    302      10    4.02      0.0000000000044 ***
## jour_semaine    6     30       5    2.03              0.05906 .
## horaire         4   8875    2219  886.72 < 0.0000000000000002 ***
## jour_travail    1     18      18    7.24              0.00723 **
## meteo           2    853     426  170.39 < 0.0000000000000002 ***
## humidite        1     52      52   20.86      0.0000054404239 ***
## vent            1     32      32   12.64              0.00039 ***
## temperature1    1    191     191   76.18 < 0.0000000000000002 ***
## temperature2    1      0       0    0.08              0.78334
## Residuals    1209   3025       3
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Les variables *jour_semaine* et *temperature2* ne sont pas significatives, avec des p-values respectives de 5.9% et 78.3%.

4.2.1 Suppression de la covariable *temperature2*

Précèdemment, nous avons calculé le coefficient de corrélation entre *temperature1* et *temperature2*, et il s'élevait à *99%*. Elles sont donc très fortement corrélées, et comme *temperature2* n'est pas significative, nous faisons le choix de ne garder que temperature1.

4.2.2 Critère AIC

Modèle forward AIC

```
##
## Call:
## lm(formula = velos_transformed ~ horaire + mois + meteo + temperature1 +
##     saison + humidite + vent + vacances, data = train)
##
## Coefficients:
##  (Intercept)      horaire2      horaire3      horaire4      horaire5
##       7.8388        5.7062        4.8949        7.1014        5.2937
##        mois2         mois3         mois4         mois5         mois6
##       0.9775        0.9298        1.7329        2.8064        2.3255
##        mois7         mois8         mois9        mois10        mois11
##       1.5420        1.9929        2.3165        1.7564        1.3826
##       mois12        meteo2        meteo3  temperature1       saison2
##       1.3442       -0.0579       -2.4321        0.9833        0.8607
##      saison3       saison4      humidite          vent     vacances2
##       1.0937        1.9740       -0.3706       -0.2076       -1.0289
```

Modèle backward AIC

```
##
## Call:
## lm(formula = velos_transformed ~ saison + mois + jour_semaine +
##     horaire + jour_travail + meteo + humidite + temperature1,
##     data = train)
##
## Coefficients:
##   (Intercept)        saison2        saison3        saison4          mois2
##        7.8682         0.9087         1.1400         2.0683         0.9426
##         mois3          mois4          mois5          mois6          mois7
##        0.9082         1.5744         2.7817         2.3797         1.5730
##         mois8          mois9         mois10         mois11         mois12
##        1.9863         2.3127         1.6900         1.3289         1.3285
## jour_semaine2  jour_semaine3  jour_semaine4  jour_semaine5  jour_semaine6
##       -1.0142        -1.0975        -1.0397        -1.0924        -0.7520
## jour_semaine7       horaire2       horaire3       horaire4       horaire5
##        0.1020         5.6521         4.8423         7.0262         5.2715
## jour_travail2         meteo2         meteo3       humidite   temperature1
##        0.9830        -0.0634        -2.5614        -0.3052         0.9694
```

Modèle both AIC

```
##
## Call:
## lm(formula = velos_transformed ~ horaire + mois + meteo + temperature1 +
##     saison + humidite + vent + vacances, data = train)
##
## Coefficients:
##  (Intercept)      horaire2      horaire3      horaire4      horaire5
##       7.8388        5.7062        4.8949        7.1014        5.2937
##        mois2         mois3         mois4         mois5         mois6
##       0.9775        0.9298        1.7329        2.8064        2.3255
##        mois7         mois8         mois9        mois10        mois11
##       1.5420        1.9929        2.3165        1.7564        1.3826
##       mois12        meteo2        meteo3  temperature1       saison2
##       1.3442       -0.0579       -2.4321        0.9833        0.8607
##      saison3       saison4      humidite          vent     vacances2
##       1.0937        1.9740       -0.3706       -0.2076       -1.0289
```

On remarque que pour le critère AIC, les méthodes *forward* et *both* construisent le même modèle, et *backward* est différent. Afin de savoir lequel on sauvegarde, on compare les critères AIC et l'on s'assure que le modèle choisi est bien meilleur que le modèle saturé à l'aide d'un *test de Type I anova*:

```
##              df    AIC
## mod_backwAIC 31 4821.7
## mod_forwAIC  26 4800.9
```

Le modèle conservé est donc : *vélos_transformed ~ horaire + mois + meteo + temperature1 + saison + humidite + vent + vacances*, car c'est celui qui possède le plus faible AIC.

```
## Analysis of Variance Table
##
## Model 1: velos_transformed ~ horaire + mois + meteo + temperature1 + saison +
##     humidite + vent + vacances
## Model 2: velos_transformed ~ saison + mois + jour_mois + jour_semaine +
##     horaire + jour_travail + vacances + meteo + humidite + temperature1
##   Res.Df  RSS Df Sum of Sq    F Pr(>F)
## 1   1246 3121
## 2   1211 3073 35      48.4 0.54   0.99
```

De plus, le test *anova* renvoie une p-value de 1, donc le modèle choisi est bien meilleur que le modèle saturé.

4.2.3 Critère BIC:

On se base désormais sur le critère *BIC*:

Modèle forward BIC:

```
##
## Call:
## lm(formula = velos_transformed ~ horaire + mois + meteo + temperature1 +
##     saison + humidite + vent + vacances, data = train)
##
```

```
## Coefficients:
##  (Intercept)      horaire2      horaire3      horaire4      horaire5
##       7.8388        5.7062        4.8949        7.1014        5.2937
##        mois2         mois3         mois4         mois5         mois6
##       0.9775        0.9298        1.7329        2.8064        2.3255
##        mois7         mois8         mois9        mois10        mois11
##       1.5420        1.9929        2.3165        1.7564        1.3826
##       mois12        meteo2        meteo3  temperature1       saison2
##       1.3442       -0.0579       -2.4321        0.9833        0.8607
##      saison3       saison4      humidite          vent     vacances2
##       1.0937        1.9740       -0.3706       -0.2076       -1.0289
```

Modèle backward BIC:

```
##
## Call:
## lm(formula = velos_transformed ~ saison + mois + horaire + meteo +
##     humidite + temperature1, data = train)
##
## Coefficients:
##  (Intercept)       saison2       saison3       saison4         mois2
##       7.7671        0.8780        1.1166        2.1186        0.9545
##        mois3         mois4         mois5         mois6         mois7
##       0.9602        1.6663        2.8644        2.5294        1.7463
##        mois8         mois9        mois10        mois11        mois12
##       2.1473        2.4182        1.7057        1.2630        1.3152
##     horaire2      horaire3      horaire4      horaire5        meteo2
##       5.6725        4.8746        7.0687        5.2972       -0.0852
##       meteo3      humidite  temperature1
##      -2.5717       -0.2953        0.9200
```

Modèle bothBIC:

```
##
## Call:
## lm(formula = velos_transformed ~ horaire + mois + meteo + temperature1 +
##     saison + humidite + vent + vacances, data = train)
##
## Coefficients:
##  (Intercept)      horaire2      horaire3      horaire4      horaire5
##       7.8388        5.7062        4.8949        7.1014        5.2937
##        mois2         mois3         mois4         mois5         mois6
##       0.9775        0.9298        1.7329        2.8064        2.3255
##        mois7         mois8         mois9        mois10        mois11
##       1.5420        1.9929        2.3165        1.7564        1.3826
##       mois12        meteo2        meteo3  temperature1       saison2
##       1.3442       -0.0579       -2.4321        0.9833        0.8607
##      saison3       saison4      humidite          vent     vacances2
##       1.0937        1.9740       -0.3706       -0.2076       -1.0289
```

Ici encore, les modèles issues des méthodes *forward* et *both* sont identiques, et il correspond au modèle conservé précédemment avec le critère *AIC*. On compare les critères BIC pour savoir lequel on conserve:

```
##              df    BIC
```

```
## mod_backwBIC 24 4949.7
## mod_forwBIC  26 4934.7
```

On observe donc que le modèle ayant le plus faible *BIC* est le même que celui ayant le plus faible *AIC*.

4.2.4 Critère CP de Mallow:

On se base désormais sur le critère *CP de Mallow*:

Modèle forward CP:

```
##
## Call:
## lm(formula = velos_transformed ~ horaire + mois + meteo + temperature2 +
##     saison + humidite + vacances + vent, data = train)
##
## Coefficients:
##  (Intercept)      horaire2      horaire3      horaire4      horaire5
##       7.7899        5.7021        4.9003        7.1068        5.2936
##        mois2         mois3         mois4         mois5         mois6
##       0.9538        0.9319        1.7374        2.8589        2.4546
##        mois7         mois8         mois9        mois10        mois11
##       1.6532        2.1254        2.4534        1.7398        1.3425
##       mois12        meteo2        meteo3  temperature2       saison2
##       1.3009       -0.0623       -2.3985        0.9357        0.8605
##      saison3       saison4      humidite     vacances2          vent
##       1.0939        2.0019       -0.3849       -0.9861       -0.1724
```

Modèle backward CP:

```
##
## Call:
## lm(formula = velos_transformed ~ saison + mois + jour_semaine +
##     horaire + jour_travail + meteo + humidite + temperature1,
##     data = train)
##
## Coefficients:
##   (Intercept)        saison2        saison3        saison4          mois2
##        7.8682         0.9087         1.1400         2.0683         0.9426
##         mois3          mois4          mois5          mois6          mois7
##        0.9082         1.5744         2.7817         2.3797         1.5730
##         mois8          mois9         mois10         mois11         mois12
##        1.9863         2.3127         1.6900         1.3289         1.3285
## jour_semaine2  jour_semaine3  jour_semaine4  jour_semaine5  jour_semaine6
##       -1.0142        -1.0975        -1.0397        -1.0924        -0.7520
## jour_semaine7       horaire2       horaire3       horaire4       horaire5
##        0.1020         5.6521         4.8423         7.0262         5.2715
## jour_travail2         meteo2         meteo3       humidite   temperature1
##        0.9830        -0.0634        -2.5614        -0.3052         0.9694
```

Modèle both CP:

```
##
```

```
## Call:
## lm(formula = velos_transformed ~ horaire + mois + meteo + temperature2 +
##     saison + humidite + vacances + vent, data = train)
##
## Coefficients:
##  (Intercept)      horaire2      horaire3      horaire4      horaire5
##       7.7899        5.7021        4.9003        7.1068        5.2936
##        mois2         mois3         mois4         mois5         mois6
##       0.9538        0.9319        1.7374        2.8589        2.4546
##        mois7         mois8         mois9        mois10        mois11
##       1.6532        2.1254        2.4534        1.7398        1.3425
##       mois12        meteo2        meteo3  temperature2       saison2
##       1.3009       -0.0623       -2.3985        0.9357        0.8605
##      saison3       saison4      humidite     vacances2          vent
##       1.0939        2.0019       -0.3849       -0.9861       -0.1724
```

Encore une fois, les modèles issues des méthodes *forward* et *both* sont identiques, et il correspond au modèle conservé précédemment avec les critères *AIC* et *BIC*. De plus, le modèle issue de la méthode *backward* est identique à celui issue de la méthode *backward* avec le critère *AIC*.

On décide donc de conserver deux modèles:

- **mod_forwAIC** :

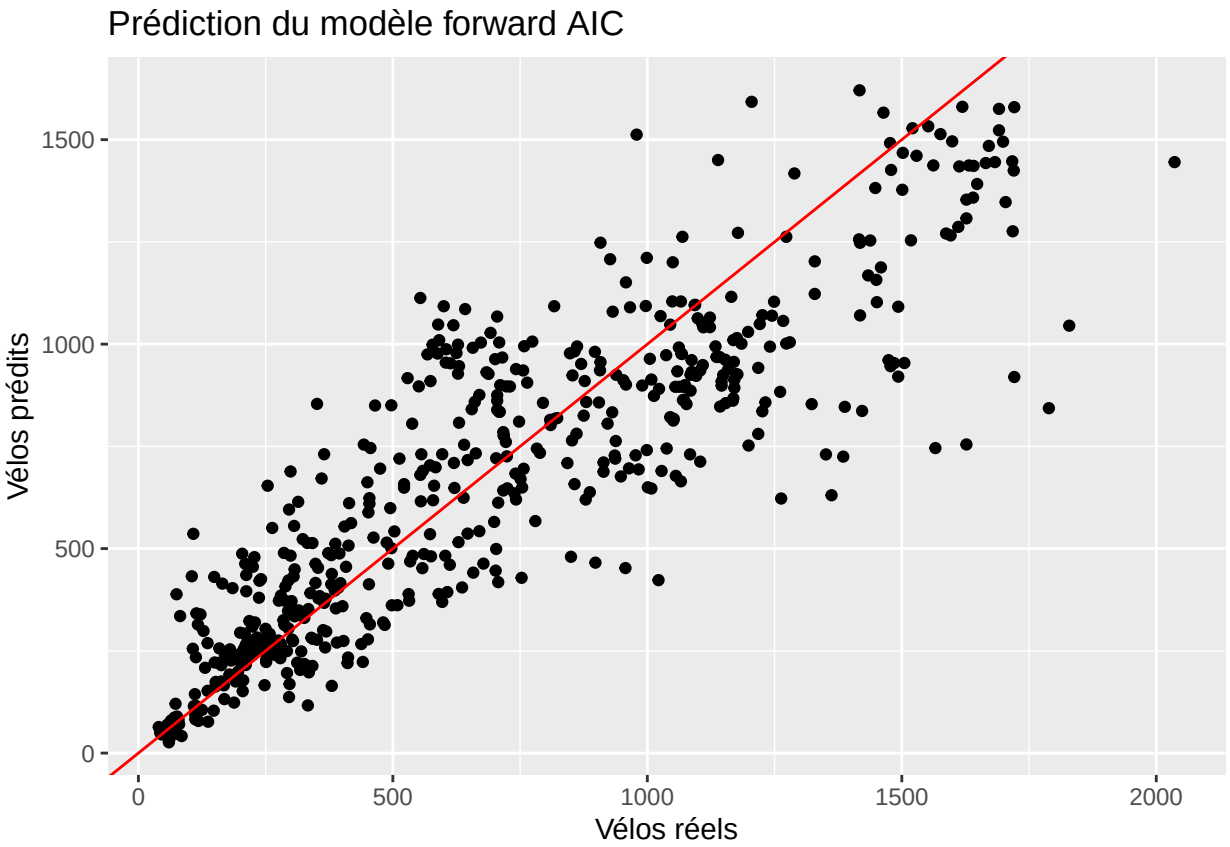
  velos_transformed $\sim$ horaire + mois + meteo + temperature1 + saison + humidite + vent + vacances.

- **mod_backwAIC** :

  velos_transformed $\sim$ saison + mois + jour_semaine + horaire + jour_travail + meteo + humidite + temperature1.

5 Prediction

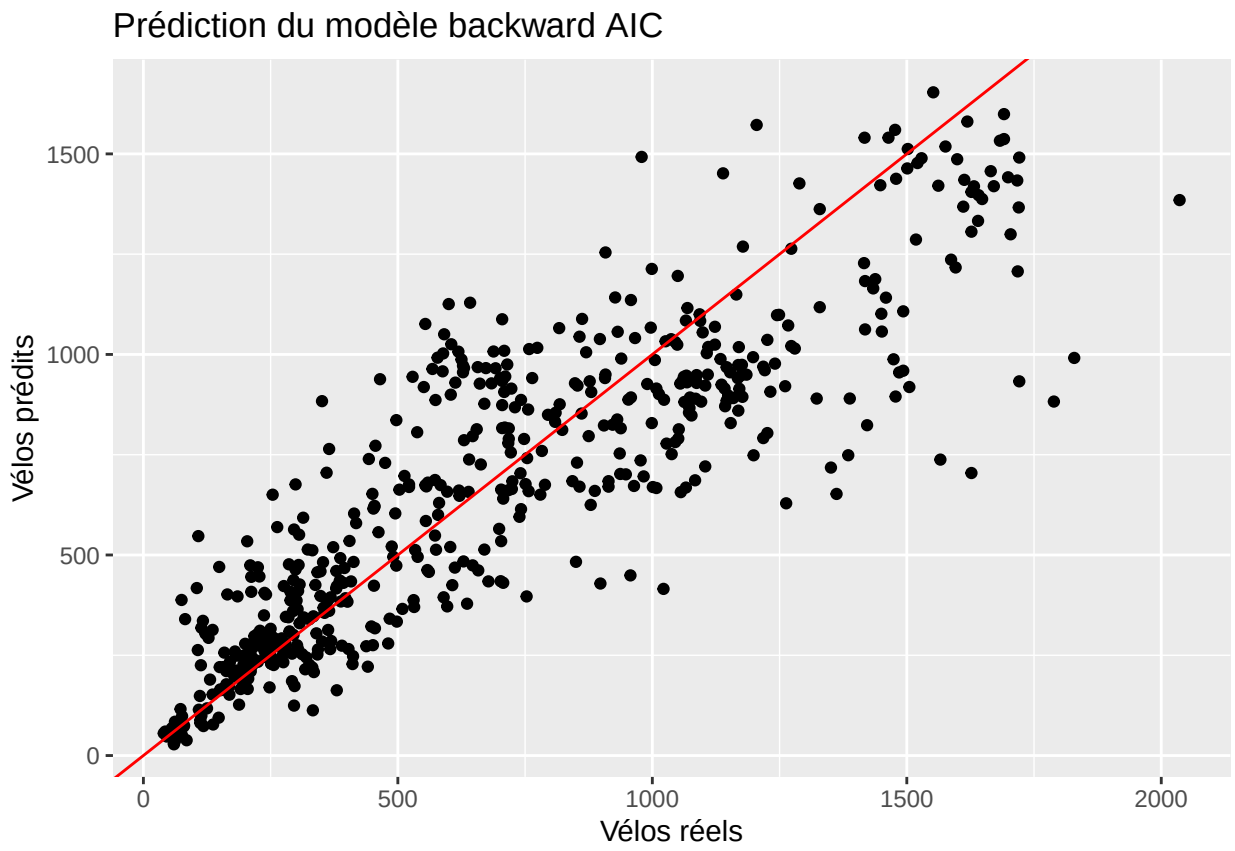
Nous allons effectué de la prédiction avec ces deux modèles, et observer qui est le plus performant.




```
## MSE forward_AIC:   223.63
```

```
## AIC forward_AIC :  4800.9
```

Le modèle *forward AIC* a un *MSE* de 223.63, et le graphique ci-dessus montre que les valeurs prédites sont proches des valeurs réelles.




```
## MSE backward_AIC:   225.97
```

```
## AIC backward_AIC 4821.7
```

Le modèle *modèle backward AIC* a un *MSE* de 225.97, et le graphique ci-dessus est très ressemblant au précédent.

Conclusion: Les deux modèles sont très proches en terme de performance, et le choix entre les deux se fera en fonction de la question que l'on se pose. Si l'on souhaite un modèle plus simple, on choisira *mod backwAIC*, et si l'on souhaite un modèle plus précis, on choisira *mod forwAIC*.

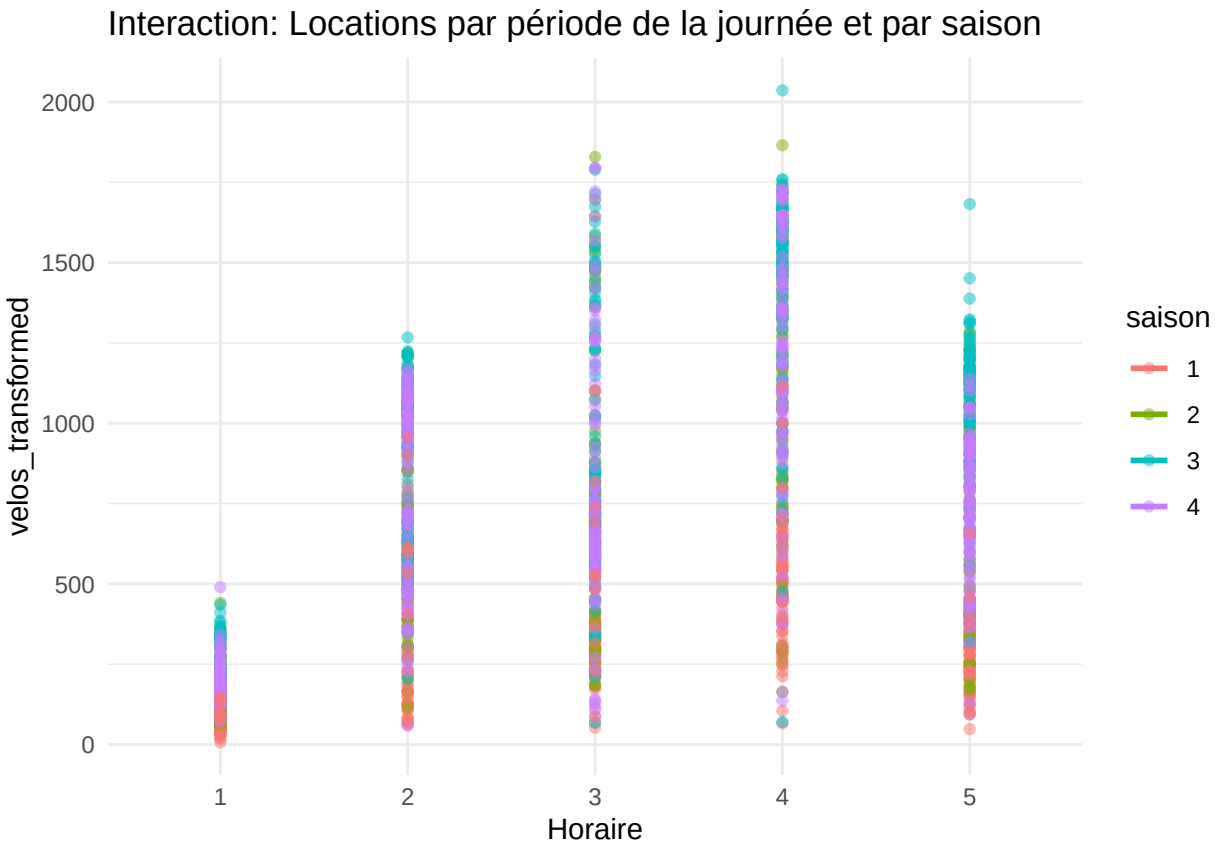
5.1 Ajout d'interactions

Nous allons ajouter des interractions à nos deux modèles afin de le rendre plus précis. Nous allons ajouter les interractions suivantes:

5.1.1 Interaction entre horaire et saison:

Le moment de la journée où les gens louent des vélos peut varier selon la saison. Par exemple : Les heures de clarté plus longues en été peuvent entraîner un plus grand nombre de locations en soirée. En hiver, le nombre de locations pourrait être inférieur, quelle que soit l'heure.

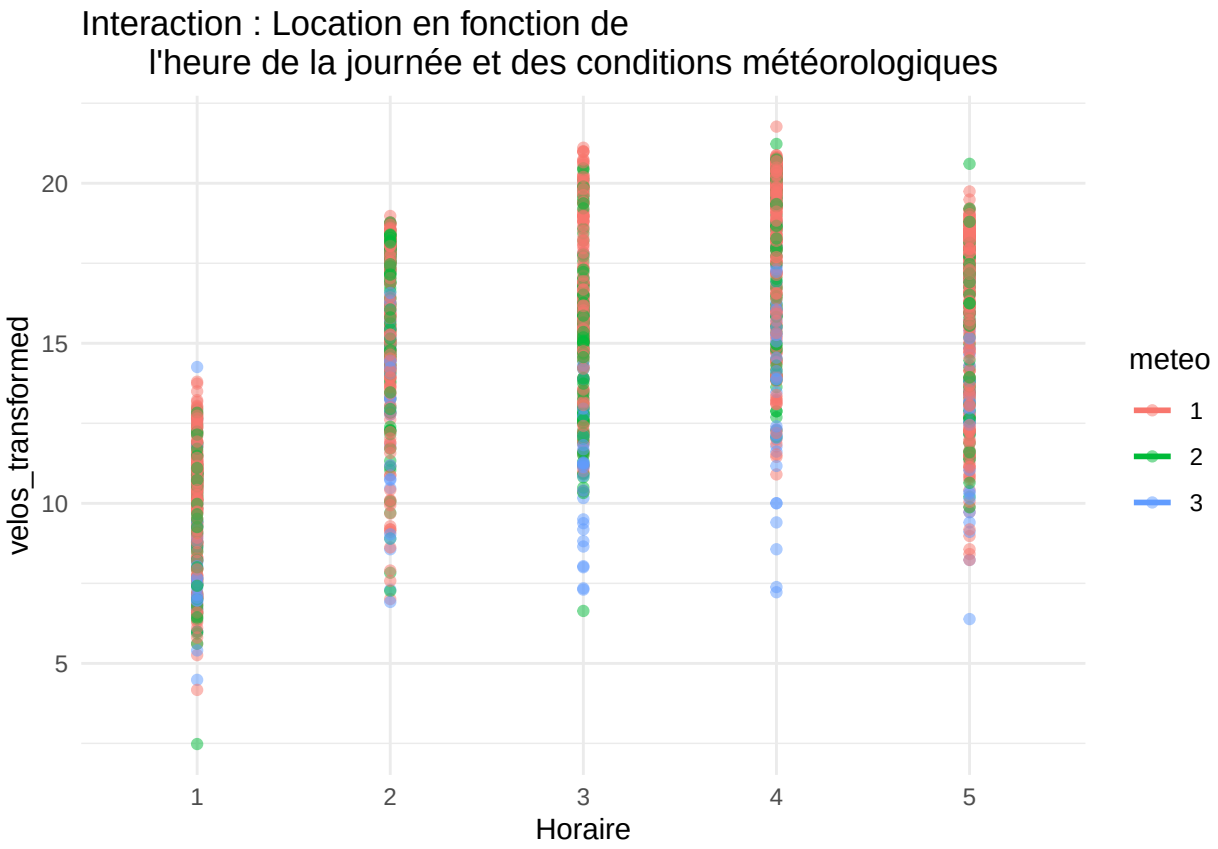
Impact attendu : Capture les effets saisonniers sur les schémas de location horaire.



5.1.2 Interaction entre horaire et météo:

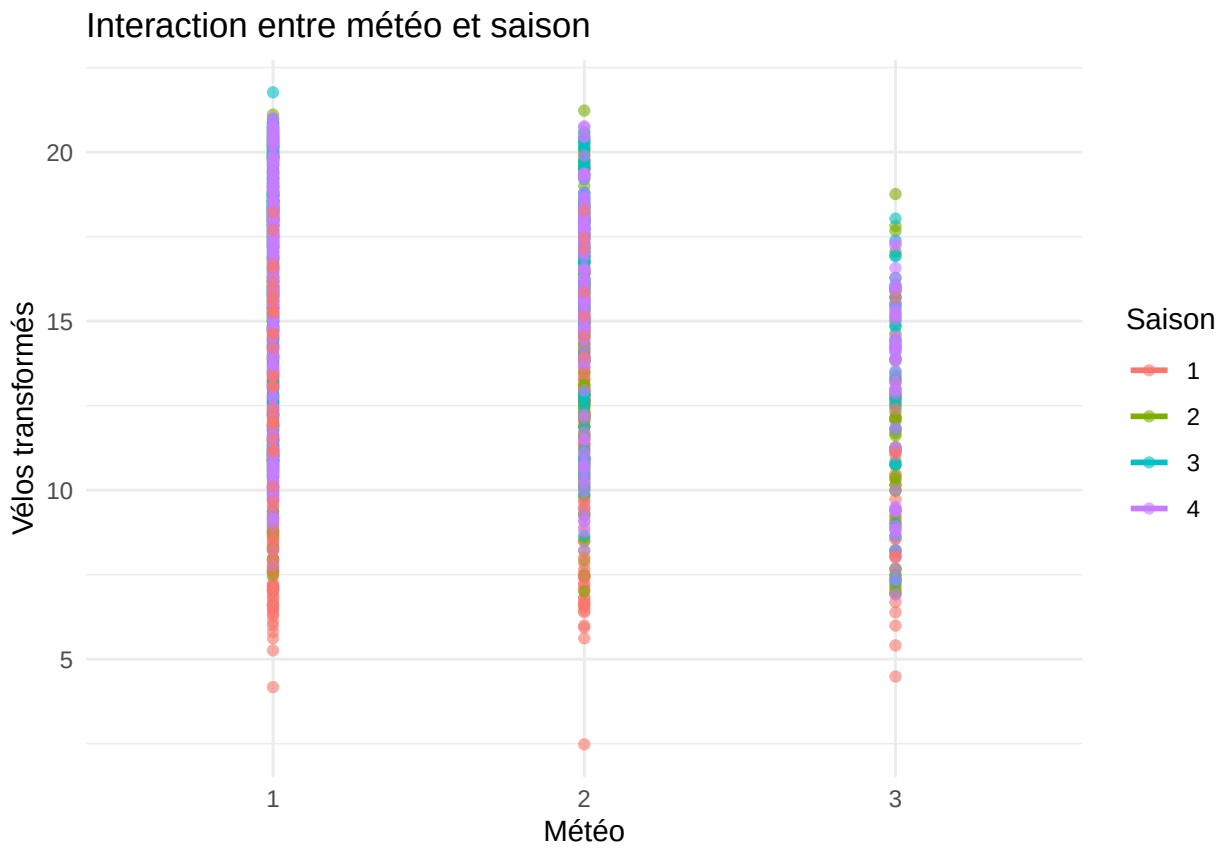
Les conditions météorologiques dépendent probablement de l'heure de la journée : Les matinées pluvieuses peuvent réduire les locations liées aux trajets domicile-travail. Les soirées claires peuvent encourager les déplacements de loisir.

Impact attendu : Reflète l'impact des conditions météorologiques sur le comportement de location à différents moments de la journée.



5.1.3 Interaction entre météo et saison:

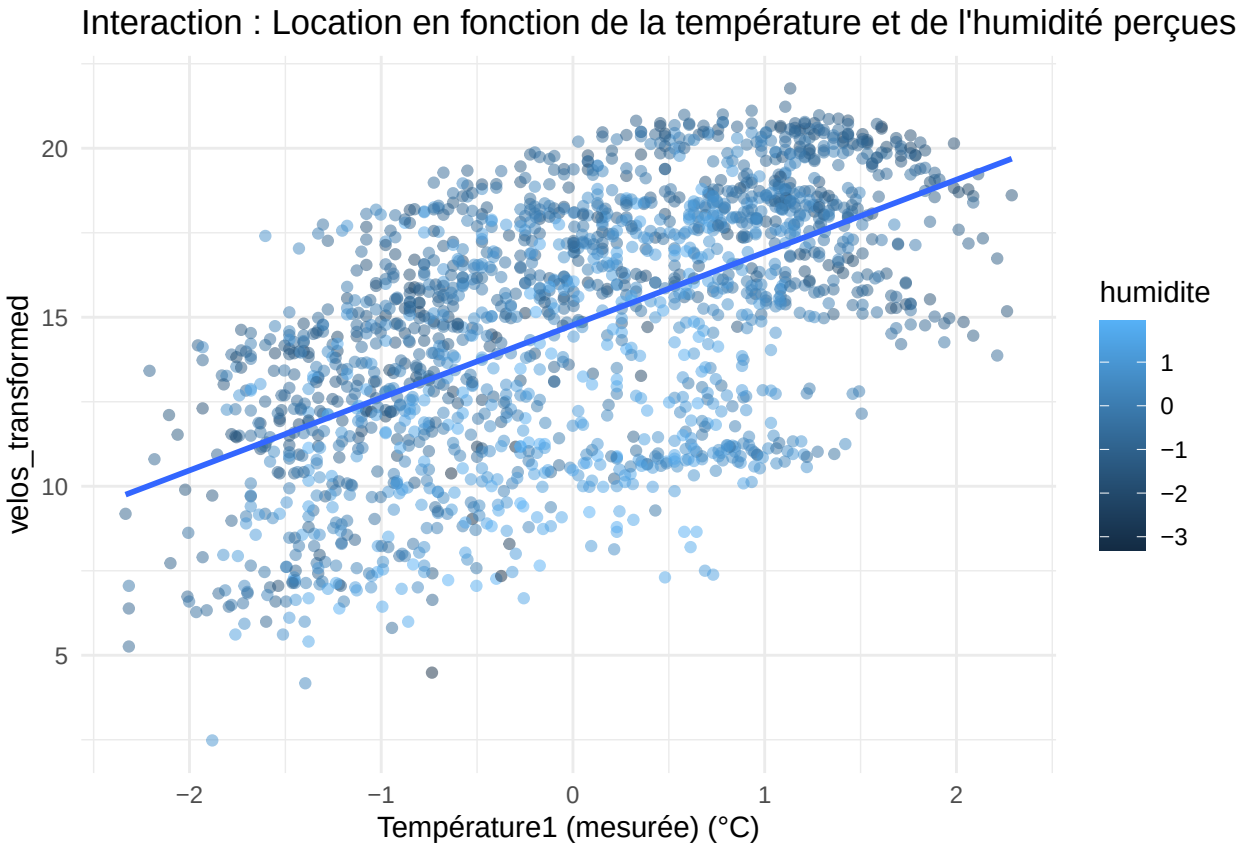
Les conditions météorologiques interagissent fortement avec les saisons, car la perception et l'impact de la météo varient en fonction de la période de l'année. Par exemple, une journée froide en hiver est perçue comme normale et peut ne pas dissuader les cyclistes réguliers, tandis qu'une journée froide en été peut avoir un effet dissuasif plus marqué. De même, les pluies fréquentes au printemps peuvent être mieux tolérées que des pluies soudaines en automne.


Interaction entre météo et saison
Vélos transformés
20
15
10
5
1
2
3
Météo
Saison
1
2
3
4

5.1.4 Interaction entre température et humidité:

Le confort perçu à vélo dépend à la fois de la température et de l'humidité : Un taux d'humidité élevé associé à des températures élevées peut décourager la location. Les températures fraîches peuvent être mieux tolérées si l'humidité est faible.

Impact attendu : Modélisation de l'effet combiné des facteurs météorologiques sur le confort et le comportement des usagers.



On ajoute donc ces quatres intéractions à nos deux modèles, et on effectue une prédiction pour voir si cela améliore la performance de nos modèles.

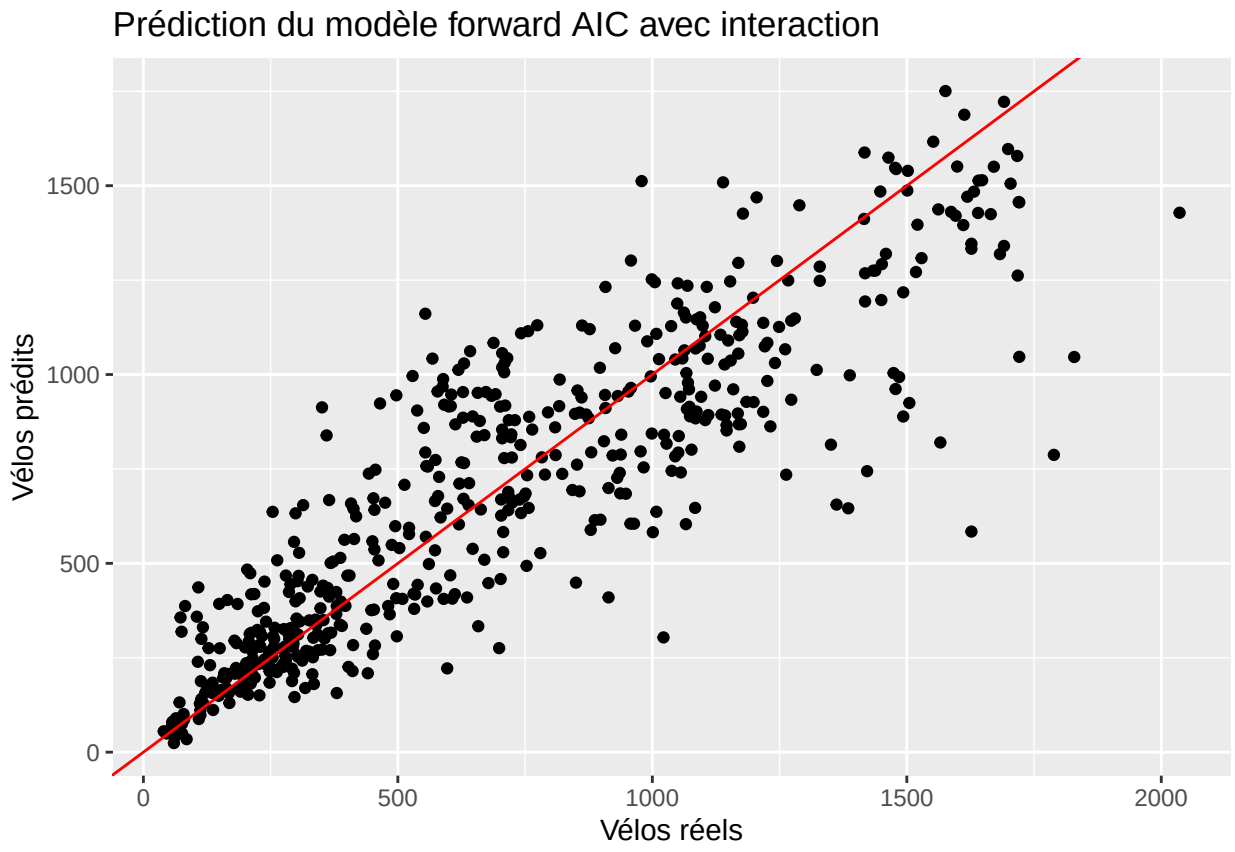
Pour le modèle forward AIC:

On effectue un test de *type I anova* pour vérifier si les intéractions ajoutées sont significatives.

```
## Analysis of Variance Table
##
## Response: velos_transformed
##                          Df Sum Sq Mean Sq F value              Pr(>F)
## horaire                   4   9306    2327 1065.09 < 0.0000000000000002 ***
## mois                     11   4252     387  176.98 < 0.0000000000000002 ***
## meteo                     2    946     473  216.62 < 0.0000000000000002 ***
## temperature1              1    173     173   79.09 < 0.0000000000000002 ***
## saison                    3    152      51   23.12    0.000000000000015 ***
## humidite                  1     58      58   26.38    0.000000327302760 ***
## humidite2                 1     68      68   31.32    0.000000027046954 ***
## vent                      1     40      40   18.13    0.000022233409389 ***
## vacances                  1     35      35   16.13    0.000062671948999 ***
```

```
## horaire:meteo             8    151      19    8.66    0.000000000015518 ***
## horaire:saison           12    150      12    5.72    0.000000001168008 ***
## temperature1:humidite2    1      3       3    1.18               0.2776
## mois:meteo               22     81       4    1.70               0.0234 *
## meteo:saison              6     53       9    4.05               0.0005 ***
## Residuals              1196   2613       2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Toutes les interactions ajoutées sont significatives, avec des p-values inférieures à 5% mis a part *temperature1:humidite2*. Nous faisons le choix de néanmoins conserver celle-ci dans le modèle.




```
## MSE forward AIC with interaction :   217.84
```

```
## AIC forward AIC with interaction :   4674.7
```

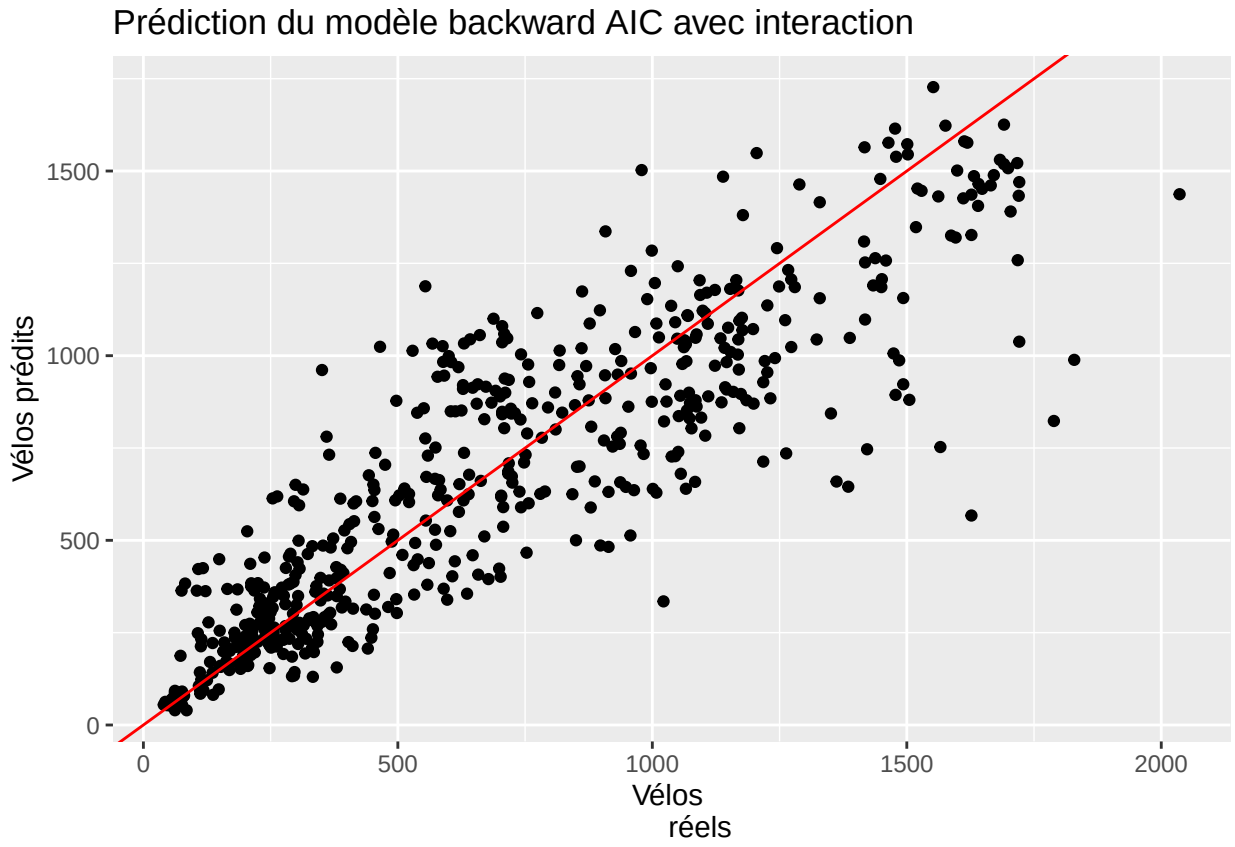
Le modèle *mod forwAIC inter* a un *MSE* de 217.84 et un *AIC* de 4674.7, ce qui est une amélioration par rapport au modèle sans interaction (avec un MSE de 223.67 et un AIC de 4800.9). Le graphique ci-dessus montre que les valeurs prédites sont proches des valeurs réelles. Néanmoins, plus le nombre de vélos réels est elevé, plus la variance de la prédiction est elevée.

Pour le modèle backward AIC:

On effectue un test de type I anova pour vérifier si les intéractions ajoutées sont significatives.

```
## Analysis of Variance Table
##
```

```
## Response: velos_transformed
##                          Df Sum Sq Mean Sq F value               Pr(>F)
## saison                    3   3773    1258  542.30 < 0.0000000000000002 ***
## mois                     11    930      85   36.47 < 0.0000000000000002 ***
## jour_semaine              6     33       5    2.37                0.028 *
## horaire                   4   9008    2252  971.03 < 0.0000000000000002 ***
## jour_travail              1     10      10    4.53                0.034 *
## meteo                     2    922     461  198.73 < 0.0000000000000002 ***
## humidite                  1     69      69   29.81      0.0000000578150 ***
## temperature1              1    188     188   80.98 < 0.0000000000000002 ***
## saison:horaire           12    140      12    5.02      0.0000000334291 ***
## horaire:meteo             8    168      21    9.05      0.0000000000039 ***
## saison:meteo              6     19       3    1.37                0.222
## humidite:temperature1     1      6       6    2.57                0.109
## Residuals              1214   2815       2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```




```
## MSE backward AIC with interaction :   222.01
```

```
## AIC forward AIC with interaction :   4733.8
```

Le modèle *backward AIC avec interaction* a un *MSE* de 222.01 et un *AIC* de 4733.8, ce qui est une amélioration par rapport au modèle sans interaction (avec un *MSE* de 225.97 et un *AIC* de 4821.7). Le graphique ci-dessus montre que les valeurs prédites sont proches des valeurs réelles. Néanmoins, plus le nombre de vélos réels est elevé, plus la variance de la prédiction est elevée.

6 Conclusion

Finalement, parmis nos deux modèles, celui qui ressort le plus performant est le modèle *mod forwAIC inter* défini par:

$$\text{mod\_forwAIC\_inter} = \text{velos\_transformed} \sim \text{horaire+mois+meteo+temperature1+saison+humidite+humidite2+vent}$$
$$+\text{vacances} + \text{meteo} : \text{horaire} + \text{horaire} : \text{saison} + \text{temperature1} : \text{humidite2} + \text{meteo} : \text{mois} + \text{meteo} : \text{saison}$$

Ce modèle est le modèle le plus précis que nous avons pu construire, comprenant les intéractions qui nous semblaient cohérentes. Nous aurions pu ajouter d'autres intéractions afin d'améliorer la prédiction, mais le modèle serait difficilement interprétable. De plus, il nous parraissait important de ne pas surcharger le modèle avec trop de variables, et de garder un modèle simple et interprétable.

7 Modèles linéaires généralisés

Au vu de l'histogramme de la variable vélos, on peut voir que la distribution de la variable vélos est asymétrique et avec une large dispersion. Nous allons utiliser un modèle linéaire généralisé en utilisant la binomiale négative pour prédire le nombre de vélos loués.

7.1 Choix du modèle

7.1.1 Méthode Step-by-Step

Afin de choisir un modèle, nous avons opté pour les critères de sélection AIC, BIC et le CP de Mallow. C'est ainsi, avec la methode backward, forward et both, nous avons obtenu plusieurs modèles dont certains sont identiques.

C'est ainsi qu'après la definition de tous ces modèle, nous n'avons conservé que les modèles non identiques pour faire un test de comparaison en optant pour un test de rapport de vraisemblance. Le test de rapport de vraisemblance nous permet de conclure que le modèle final est le modèle obtenu avec la méthode forward avec le critère AIC ainsi que le critère BIC.

Ce modèle a été conservé puis des interactions possibles ont été ajouter pour confronter le mécanisme des locations de vélos avec ces interactions.De ce fait, un test anova a été réalisé pour décéler les interactions significatives.

```
## Likelihood summary table:
##                    AIC   BIC LR Chisq   Df Pr(>Chisq)
## model_glm        16830 17154     1303 1209       0.03 *
## modglm_forwAIC   16792 16931     1304 1245       0.12
## modglm_bothAIC   16797 16962     1303 1240       0.10
## modglm_backwAIC 16797 16962     1303 1240       0.10
## modglm_backwBIC 16805 16934     1304 1247       0.13
## modglm_backwCP   16797 16962     1303 1240       0.10
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Nous conserverons alors que les interactions significatives.

Donc après avoir ajouté les interactions, nous avons réalisé encore un test anova pour vérifier la significativité des interactions ajoutées. Nous avons conclu que les interactions ajoutées sont significatives.Par conséquent, nous avons conservé le modèle final avec les interactions significatives. D'où la formalisation mathématique du modèle final est: $Y_i \sim \text{NB}(\mu_i, \phi)$, et $g(\mathbb{E}[Y_i \mid x_{ij}]) = \mu_i$ avec μ_i défini comme suit :

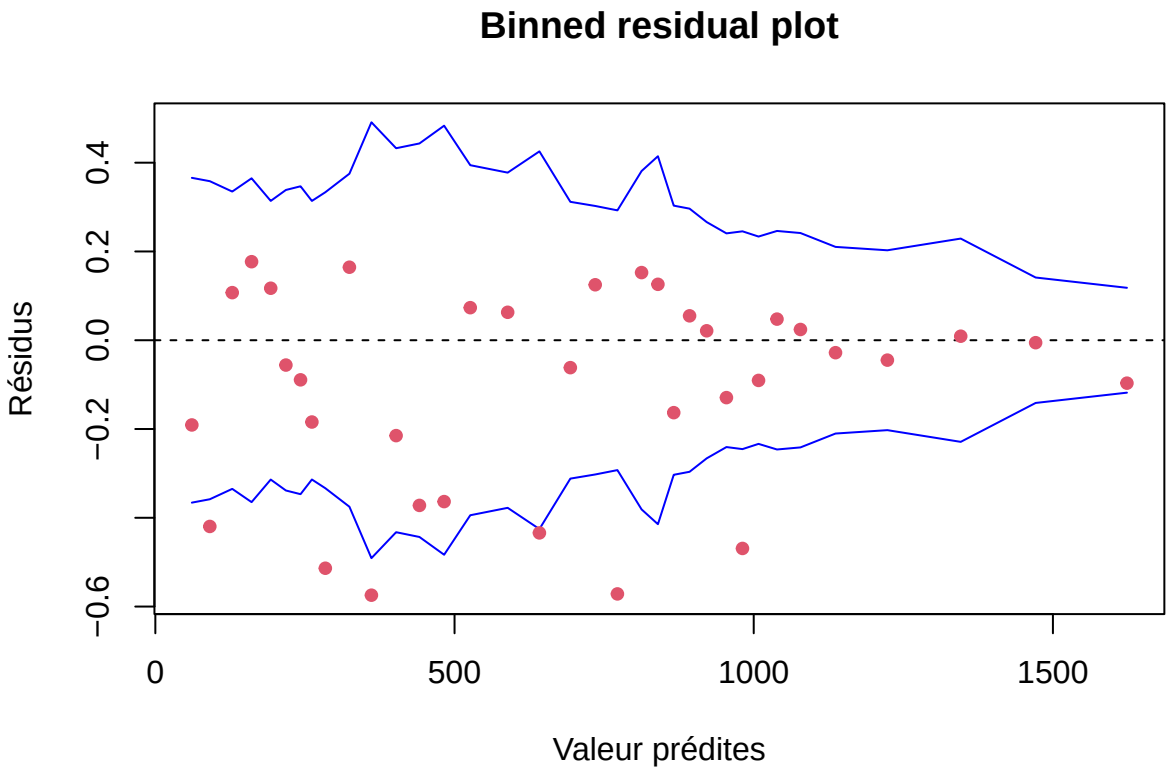
$$\mu_i = \beta_0 + \beta_1 \cdot \text{horaire}_i + \beta_2 \cdot \text{mois}_i + \beta_3 \cdot \text{meteo}_i + \beta_4 \cdot \text{temperature2}_i + \beta_5 \cdot \text{saison}_i + \beta_6 \cdot \text{humidite}_i + \beta_7 \cdot \text{vent}_i$$
$$+\beta_8 \cdot (\text{mois}_i \cdot \text{temperature2}_i) + \beta_9 \cdot (\text{horaire}_i \cdot \text{temperature2}_i).$$

et g est la fonction de lien canonique définie $\forall t \in \mathbb{R}, g(t) = \frac{r}{t+r}$ Avec r le paramètre de dispersion de la loi de la binomiale négative.

7.2 Analyse des résidus

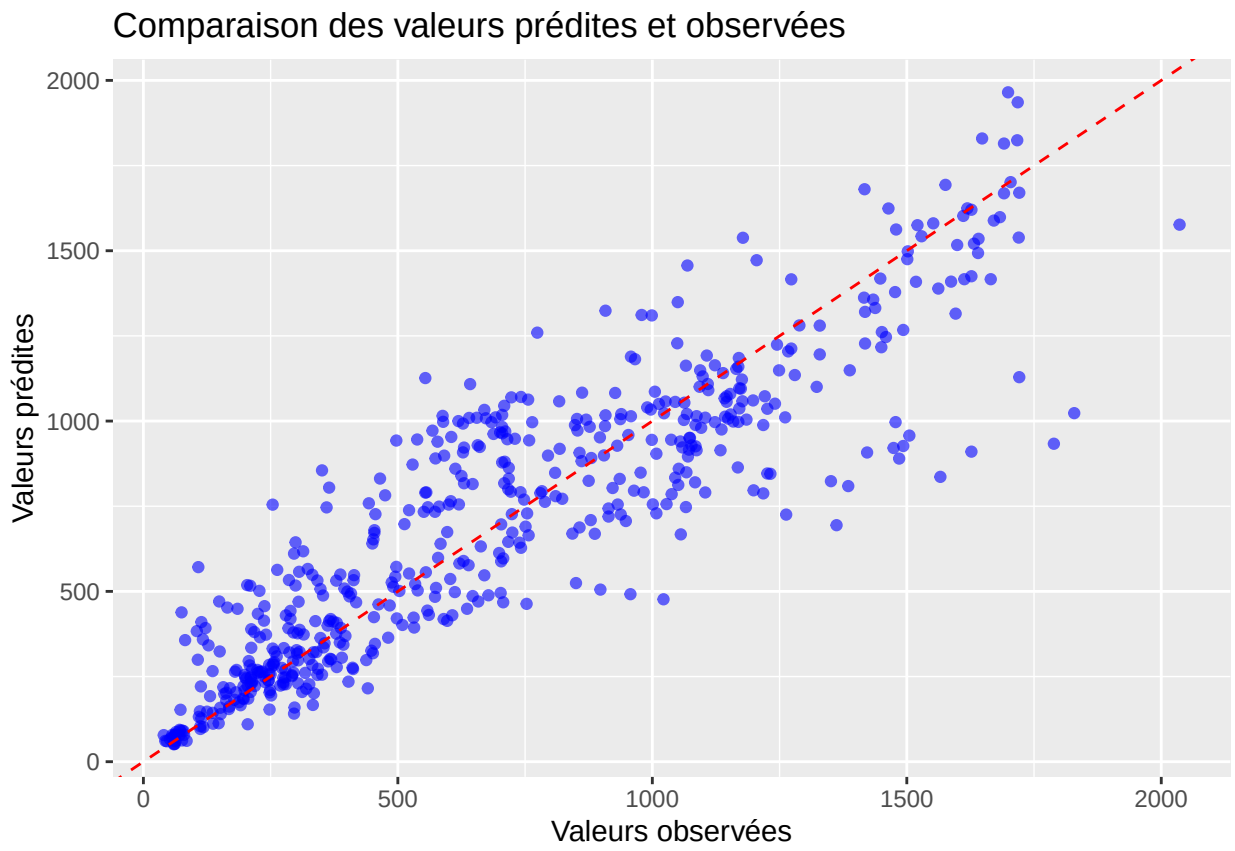
Ce graphique montre que les résidus ne sont pas uniformément répartis autour de zéro, ce qui peut indiquer que le modèle présente des limites dans son ajustement. Bien que les résidus oscillent autour de zéro, leur

répartition suggère l'existence d'une structure non linéaire sous-jacente dans les données, que le modèle actuel pourrait ne pas capturer pleinement. En particulier, une augmentation de la dispersion des résidus pour des valeurs prévues élevées est notable, ce qui pourrait refléter un manque de flexibilité dans le modèle pour traiter ces valeurs. Cela suggère que d'autres formes de modélisation ou des transformations supplémentaires des variables explicatives pourraient être nécessaires pour mieux capturer les relations non linéaires présentes dans les données. Néanmoins, l'absence de schémas fortement systématiques ou de biais marqués dans l'ensemble des résidus reste un point positif.



7.3 Prédiction et évaluation du modèle

La prédiction effectuée par notre modèle final a été évaluée à l'aide de l'erreur quadratique moyenne (RMSE) qui est de 203 vélos. Cela signifie que notre modèle prédit en moyenne à 203 vélos près du nombre réel de vélos loués. Ce résultat peut être considéré comme satisfaisant au vu de la moyenne de vélos loués mais il reste des marges d'amélioration pour affiner les prédictions et réduire l'erreur.




```
## RMSE du modèle GLM :  203.66
```

7.4 Limites du modèle et améliorations possibles.

Bien qu'il est présenté une réduction de la RMSE par rapport au modèle linéaire classique, le modèle GLM présente des limites. En effet, comme souligné lors de l'analyse des résidus, le modèle GLM pourrait ne pas capturer pleinement les relations non-linéaires présentes dans les données. Toutefois, nous pouvons envisager des améliorations possibles pour le modèle GLM. Par exemple, l'ajout de termes quadratiques ou cubiques pour les variables continues pourrait permettre de mieux capturer les relations non-linéaires. De plus, l'ajout de nouvelles variables explicatives ou de transformations supplémentaires des variables existantes pourrait également améliorer la performance du modèle. Enfin, l'utilisation de méthodes de régularisation telles que la régression ridge ou LASSO pourrait aider à réduire le surajustement et à améliorer la généralisation du modèle.