Bruxelles Mobilité (CB15-99)

Trindades A., Zacks N., Lebert O.

Overview

***Contexte* :**

Nous avons travaillé sur les données de la borne de comptage vélo CB1599 pour l'année 2024, située à Ixelles, au coin de l'avenue P.Héger et de l'avenue F.Roosevelt (voir nos photos en annexe). Ces données incluent le nombre de passages de vélos ainsi que la vitesse moyenne par quart d'heure au cours de la journée. Nous avons complété ces informations avec des données météorologiques de l'observatoire d'Uccle pour 2024, incluant notamment la vitesse du vent, les précipitations, l'ensoleillement et l'humidité.

Contour du travail :

Dans un premier temps, nous avons préparé et nettoyé les données, créant des variables temporelles (jour, heure, saison, vacances). Nous avons calculé des moyennes de vitesse pondérées en fonction du jour et de l'heure, tout en excluant les valeurs manquantes. Ces données ont ensuite été utilisées pour générer une « heatmap » et des graphiques linéaires permettant de visualiser des tendances importantes. Enfin, nous avons analysé les relations entre ces variables à l'aide de modèles de régression linéaire, avec et sans les variables météorologiques.

Observations principales :

Nos visualisations ont mis en évidence une diminution de la vitesse moyenne pendant les heures de la nuit (entre 1h et 3h), suivie d'une augmentation progressive de la vitesse, culminant à près de 25km/h dans la 8ième heure de la journée. Après 8h, la vitesse diminue de nouveau et se stabilise autour de 20 km/h pour le reste de la journée. Nous avons également observé que cette tendance était légèrement plus tardive durant le week-end.

La première régression linéaire a permis de démontrer des influences significatives des jours variables temporelles (heures, jours, mois) sur la vitesse moyenne, ainsi qu'un pouvoir prédictif (R2 ~ 0.22) et une erreur standard (~2) relativement modérés.

La deuxième régression linéaire, avec les données météo, a augmenté légèrement le pouvoir prédictif du modèle (R2 ~ 0.25) tout en conservant une erreur standard comparable. Seul peu d'information supplémentaire a donc été apporté par les variables météo, bien que certaines d'entre elles aient aussi présenté une influence significative.

Ces résultats nous laissent soupçonner que d'autres variables, telles que le type de véhicule (dont les trottinettes/vélos électriques), l'âge de l'utilisateur, le sens de la circulation ou encore l'origine ou la destination du trajet (vu l'emplacement de la borne), pourraient également exercer une influence notable sur la vitesse moyenne.

Introduction au problème, introduction aux données, description des méthodes utilisées.

1. Excel

Traitement des données

Après avoir importé les données de la borne via Power Query, la première étape a été de créer les variables de calendrier nécessaires. Pour cela, nous avons utilisé la fonction *TEXTE* afin d'extraire le jour de la semaine à partir des dates présentes dans les données. Ensuite, nous avons arrondi l'heure à partir de la colonne « time.gap **»**, qui divise la journée en 96 quarts d'heure. Pour cela, nous avons employé la fonction *ARRONDI*, afin de regrouper les différents intervalles de quarts d'heure en heures entières (de 1 à 24). Ensuite, nous avons calculé une moyenne pondérée par jour et par heure, en excluant les vitesses négatives que nous avons converties en 0. Pour ce faire, nous avons utilisé la fonction *SOMME.SI.ENS* pour calculer les sommes de vitesses pondérées et les comptes pour chaque combinaison de jour et d'heure arrondie (par exemple : lundi 1h). Une fois les sommes obtenues, nous avons divisé la somme des vitesses par le compte pondéré pour obtenir la moyenne pondérée.

Création du TCD et d'un graphique linéaire

Sur la base des données préparées, nous avons créé *un tableau croisé dynamique* (TCD) avec les heures en ligne, les jours en colonne et la moyenne pondérée en valeurs. Nous avons appliqué *une mise en forme conditionnelle* sur ce TCD pour mettre en valeur les moyennes les plus élevées en vert et les plus faibles en rouge, permettant ainsi de visualiser rapidement les tendances (voir page 4). En parallèle, nous avons confectionné un *graphique linéaire* qui montre l'évolution de la vitesse moyenne au cours de la journée, ainsi que les différences entre les jours de la semaine. Ce graphique a permis de visualiser de manière claire l'impact de l'heure et du jour sur la vitesse des vélos (voir p. 5 et annexe).

Tendances dégagées

Les visualisations nous ont permis de dégager des tendances claires :

- La vitesse est plus basse entre 1h et 3h du matin.
- Elle commence à augmenter progressivement entre 4h et 8h, atteignant un pic à 8h. Ce pic est particulièrement prononcé les jours de la semaine, suggérant une vitesse accrue liée aux déplacements domicile-travail, domicile-école et, dans notre cas, domicile-université.
- Après 8h, la vitesse diminue à nouveau.
- Nous avons également observé que cette tendance était décalée d'une heure durant le week-end.

2. Analyses réalisées en R

Partie 1 : Transformation des données

Nous avons commencé par introduire plusieurs bibliothèques R, suivi de la configuration des paramètres régionaux de la session, en français. Ces éléments seront nécessaires dorénavant. Ensuite, nous faisons en sorte que les données de la borne et météorologiques soient chargées et les colonnes de date et heure soient converties et formatées. Un ajustement spécifique est effectué sur les deux ensembles pour remplacer les heures « 00 :00 » par « 24 :00 ». Les ensembles de données sont ensuite fusionnés à l'aide des colonnes de date et d'heure. Enfin, les données fusionnées sont liées à un fichier excel que nous avons créé contenant les dates de congé (source : FWB). L'ensemble de données est trié par date et heure et enregistré dans un fichier Excel, nommé : « final_data.xlsx ». Ce processus garantit que les données sont prêtes à être analysées dans les 3 parties suivantes.

Partie 2 : Fonction pour création d'un Heatmap et d'un graphique en ligne

On commence par créer la fonction « analyser_borne » qui permet de générer un « heatmap » et un graphique linéaire représentant la vitesse moyenne des vélos en fonction de l'heure et du jour de la

semaine. La fonction prend en entrée un nom de fichier «.xlsx » contenant, entre autres, les données d'une borne. Les données sont choisies, transformées et agrégées pour calculer la vitesse moyenne pondérée par le nombre vélos qui passent par la borne, par heure. La « heatmap » est ensuite créée à l'aide de « ggplot2 » pour visualiser la vitesse moyenne des vélos en fonction de l'heure et du jour de la semaine (voir p.4). Un graphique linéaire est également généré pour visualiser la vitesse moyenne pour chaque jour de la semaine en fonction de l'heure (voir p.5). La fonction est appliquée aux données de la borne « CB1599 ».

Partie 3 : Analyse de régressions linéaires multiples excluant les variables météorologiques

Nous préparons le jeu de données *bike_data* afin d'analyser les données issues de la borne CB1599. Trois *boxplots* illustrent la vitesse moyenne des vélos selon l'heure, le jour et le mois (voir annexes, images 3.1-3.3). Ils révèlent que la vitesse varie principalement selon l'heure, avec des pics aux heures de pointe matinales et une variabilité accrue à certains moments. Le jour de la semaine et le mois semblent avoir un impact limité sur la vitesse moyenne.

Pour approfondir cette analyse, nous développons un modèle de régression linéaire multiple expliquant la vitesse en fonction de ces trois variables temporelles, ainsi qu'une variable binaire indiquant les jours fériés (1) ou non (0). Deux approches sont testées :

- **Modèle 1 (*modele_1*)** : données agrégées par heure (*reg_data_clean*).
- **Modèle 2 (*modele_2_pond*)** : données par intervalles de 15 minutes (*bike_data_clean*), avec une pondération selon le nombre de vélos comptés (*weights = count*).

Les statistiques comparatives (Table 3.1, annexe) montrent que, bien que *modele_2* présente un R^2 légèrement plus élevé, il affiche une erreur standard résiduelle plus importante, suggérant une moins bonne précision. Nous retenons donc *modele_1*.

Ce modèle est significatif ($p < 2.2e\text{-}16$) mais n'explique qu'environ 23 % de la variance. L'ordonnée à l'origine ($\approx 18{,}75$) correspond à la vitesse moyenne estimée pour la référence temporelle (1 h, lundi, janvier, hors jours fériés). Certains coefficients horaires et mensuels sont significatifs, traduisant des variations marquées. Par exemple, la vitesse moyenne est plus élevée en juillet (+1,46 km/h) et en août (+1,27 km/h) par rapport à janvier, tandis qu'octobre présente un effet quasi nul ou négatif, suggérant une variation saisonnière (voir notre sortie R, p.5).

Afin d'améliorer l'ajustement, nous testons plusieurs transformations de la variable dépendante (*logarithmique, quadratique, cubique, racine carrée et racine cubique*). Les modèles quadratique et cubique sont écartés en raison de leurs erreurs standard élevées. Les modèles logarithmique et racine carrée améliorent la précision, mais présentent un R^2 ajusté inférieur à *modele_1*, suggérant un surajustement. Par conséquent, *modele_1* reste le plus approprié (Table 3.2, annexe).

L'analyse de la multicolinéarité (facteur VIF) ne révèle pas de problème majeur. Les tests de normalité (histogramme, Q-Q plot, images 3.4-3.5, annexe), d'indépendance des résidus (image 3.6) et d'homoscédasticité (image 3.7) confirment la validité du modèle. L'examen des distances de Cook (image 3.8-9) identifie quelques observations influentes. Bien qu'elles ne soient pas aberrantes, un modèle excluant ces points a été testé, améliorant significativement le R^2 ajusté ($0.2292 \rightarrow 0.3807$) et réduisant l'erreur standard résiduelle (images 3.10-3.13, Table 3.4).

Partie 4 : Analyse de régressions linéaires multiples incluant les variables météorologiques

Dans la dernière partie de notre travail, nous avons poursuivi l'analyse des données en nous concentrant sur les variables météorologiques « sun_24_hours » et « precip_quantity », dont les relevés étaient respectivement disponibles une fois et deux fois par jour. D'une part, nous avons estimé la durée d'ensoleillement moyen quotidien en fonction des heures de lever et de coucher du soleil, obtenues auprès de l'observatoire d'Uccle. D'autre part, nous avons calculé et imputé à chaque heure une valeur de précipitations moyennes accumulées sur les douze heures précédentes. Par ailleurs, les valeurs manquantes des autres variables météo ont été remplacées par leurs moyennes quotidiennes respectives.

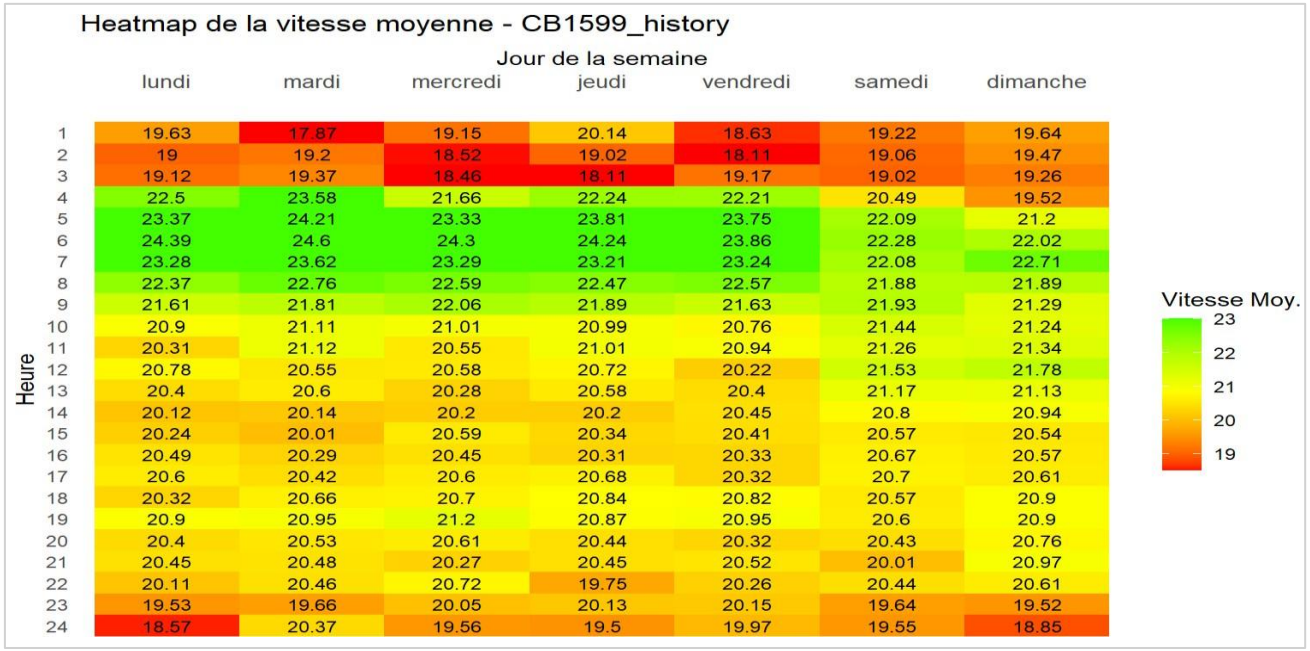
L'ajout des variables météorologiques n'a apporté qu'une amélioration marginale au modèle. Bien que certaines aient montré une influence statistiquement significative, le R^2 (~0,25) n'a que

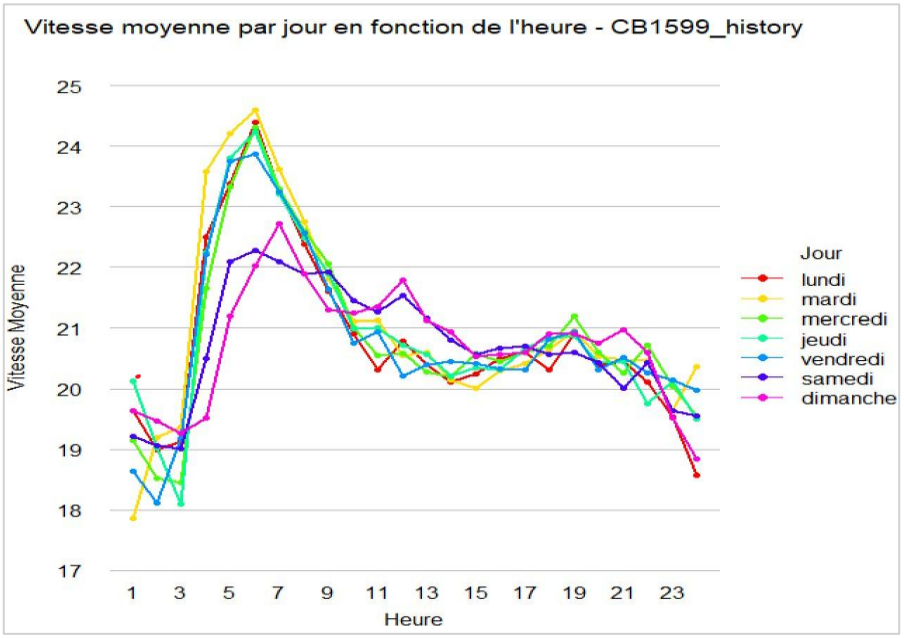
légèrement augmenté et l'erreur standard est restée stable. Parmi elles, la variable représentant le niveau moyen des précipitations a eu l'impact le plus marqué, entraînant une diminution de la vitesse moyenne de -1,57 km/h. Les tests de validation du modèle confirment que les hypothèses d'homoscédasticité, d'indépendance des résidus et de normalité restent défendables. Par ailleurs, même après l'intégration des variables météorologiques, les effets horaires demeurent prépondérants, avec une augmentation notable de la vitesse entre 5 h et 8 h (+2 à +4 km/h par rapport à la première heure du jour).

Enfin, le pouvoir explicatif modéré de l'ensemble des variables utilisées suggère que d'autres facteurs, tels que le type de vélo, l'âge des cyclistes, leur destination ou encore le sens de circulation, pourraient avoir un effet significatif sur la vitesse des cyclistes passant devant la borne CB1599.

Résultats principaux : tables et graphiques

Moyenne de Moyenne pondérée finale	Étiquettes de colonnes							
Étiquettes de lignes	lundi	mardi	mercredi	jeudi	vendredi	samedi	dimanche	Total général
1	19,63	17,87	19,15	20,14	18,63	19,22	19,64	19,18
2	19,00	19,20	18,52	19,03	18,11	19,06	19,47	18,91
3	19,12	19,37	18,46	18,11	19,17	19,02	19,26	18,93
4	22,50	23,58	21,66	22,24	22,21	20,49	19,52	21,75
5	23,37	24,21	23,33	23,81	23,75	22,09	21,20	23,11
6	24,39	24,60	24,30	24,24	23,86	22,28	22,02	23,68
7	23,28	23,62	23,29	23,21	23,24	22,08	22,71	23,06
8	22,37	22,76	22,59	22,47	22,57	21,88	21,89	22,36
9	21,61	21,81	22,06	21,89	21,63	21,93	21,29	21,75
10	20,90	21,11	21,01	20,99	20,76	21,44	21,24	21,06
11	20,31	21,12	20,55	21,01	20,94	21,26	21,34	20,93
12	20,78	20,55	20,58	20,72	20,22	21,53	21,78	20,88
13	20,40	20,60	20,28	20,58	20,40	21,17	21,13	20,65
14	20,12	20,14	20,20	20,20	20,45	20,80	20,94	20,41
15	20,24	20,01	20,59	20,34	20,41	20,57	20,54	20,39
16	20,49	20,29	20,45	20,31	20,33	20,67	20,57	20,44
17	20,60	20,42	20,60	20,68	20,32	20,70	20,61	20,56
18	20,32	20,66	20,70	20,84	20,82	20,57	20,90	20,69
19	20,90	20,95	21,20	20,87	20,95	20,60	20,90	20,91
20	20,40	20,53	20,61	20,44	20,32	20,43	20,76	20,50
21	20,45	20,48	20,27	20,45	20,52	20,01	20,97	20,45
22	20,11	20,46	20,72	19,75	20,26	20,44	20,61	20,34
23	19,53	19,66	20,05	20,13	20,15	19,64	19,52	19,81
24	18,57	20,37	19,56	19,49	19,97	19,55	18,85	19,48
Total général	**20,81**	**21,02**	**20,86**	**20,91**	**20,83**	**20,73**	**20,74**	**20,84**





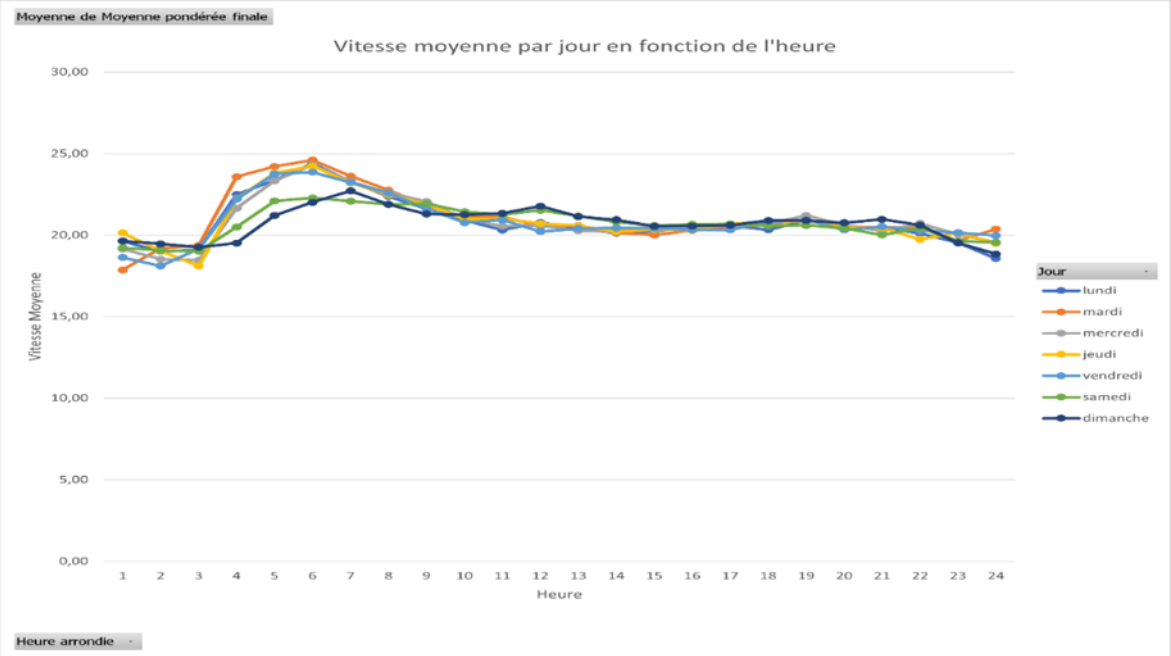

```
Call:
lm(formula = average.speed ~ heure_pleine + jour_semaine + mois +
    vacances, data = reg_data_clean)

Residuals:
     Min       1Q   Median       3Q      Max
-14.8681  -0.9803   0.0531   1.0695  24.5892

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         18.751542   0.160678 116.703  < 2e-16 ***
heure_pleine2       -0.386718   0.179924  -2.149 0.031636 *
heure_pleine3       -0.275895   0.186600  -1.479 0.139302
heure_pleine4        1.806297   0.179063  10.087  < 2e-16 ***
heure_pleine5        3.193145   0.170856  18.689  < 2e-16 ***
heure_pleine6        4.131313   0.170054  24.294  < 2e-16 ***
heure_pleine7        3.407668   0.169719  20.078  < 2e-16 ***
heure_pleine8        2.868687   0.169502  16.924  < 2e-16 ***
heure_pleine9        2.218179   0.169502  13.086  < 2e-16 ***
heure_pleine10       1.661627   0.169502   9.803  < 2e-16 ***
heure_pleine11       1.494281   0.169502   8.816  < 2e-16 ***
heure_pleine12       1.481168   0.169502   8.738  < 2e-16 ***
heure_pleine13       1.186195   0.169502   6.998 2.79e-12 ***
heure_pleine14       0.998667   0.169502   5.892 3.97e-09 ***
heure_pleine15       0.948901   0.169502   5.598 2.23e-08 ***
heure_pleine16       0.978522   0.169502   5.773 8.07e-09 ***
heure_pleine17       1.123989   0.169502   6.631 3.54e-11 ***
heure_pleine18       1.261296   0.169502   7.441 1.10e-13 ***
heure_pleine19       1.373792   0.169502   8.105 6.02e-16 ***
heure_pleine20       0.946626   0.169502   5.585 2.41e-08 ***
heure_pleine21       0.916404   0.169502   5.406 6.60e-08 ***
heure_pleine22       0.842240   0.169502   4.969 6.87e-07 ***
heure_pleine23       0.257844   0.169612   1.520 0.128498
heure_pleine24       0.080771   0.170734   0.473 0.636169
jour_semainemardi    0.265632   0.090668   2.930 0.003402 **
jour_semainemercredi 0.015229   0.090959   0.167 0.867039
jour_semainejeudi    0.165862   0.091093   1.821 0.068673 .
jour_semainevendredi 0.146276   0.090445   1.617 0.105851
jour_semainesamedi   0.039420   0.090327   0.436 0.662545
jour_semainedimanche 0.009815   0.090419   0.109 0.913562
moisfévrier          0.493433   0.121440   4.063 4.89e-05 ***
moismars             0.850249   0.119004   7.145 9.77e-13 ***
moisavril            0.730089   0.119875   6.090 1.18e-09 ***
moismai              0.617313   0.119044   5.186 2.20e-07 ***
moisjuin             0.799525   0.120023   6.661 2.88e-11 ***
moisjuillet          1.460986   0.124584  11.727  < 2e-16 ***
moisaoût             1.268922   0.123795  10.250  < 2e-16 ***
moisseptembre        0.459967   0.120080   3.831 0.000129 ***
moisoctobre         -0.084639   0.118517  -0.714 0.475153
moisnovembre        -0.188709   0.120228  -1.570 0.116548
moisdécembre         0.686104   0.118606   5.785 7.52e-09 ***
vacances1           -0.005406   0.064592  -0.084 0.933301
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.218 on 8403 degrees of freedom
Multiple R-squared:  0.2303,    Adjusted R-squared:  0.2265
F-statistic: 61.32 on 41 and 8403 DF,  p-value: < 2.2e-16
```

Annexes



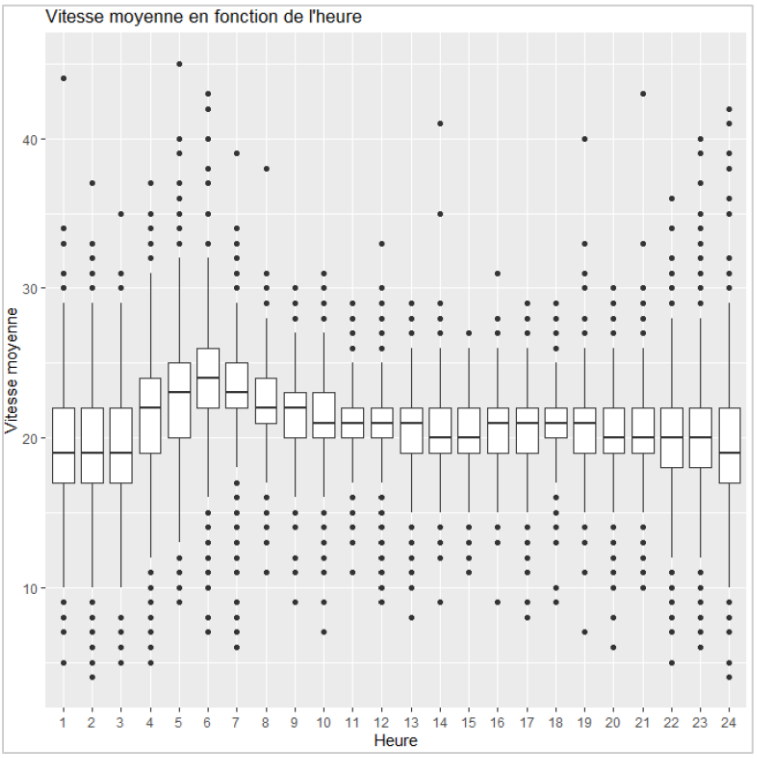


Image 3.1 - Boxplot pour la vitesse moyenne des vélos selon l'heure

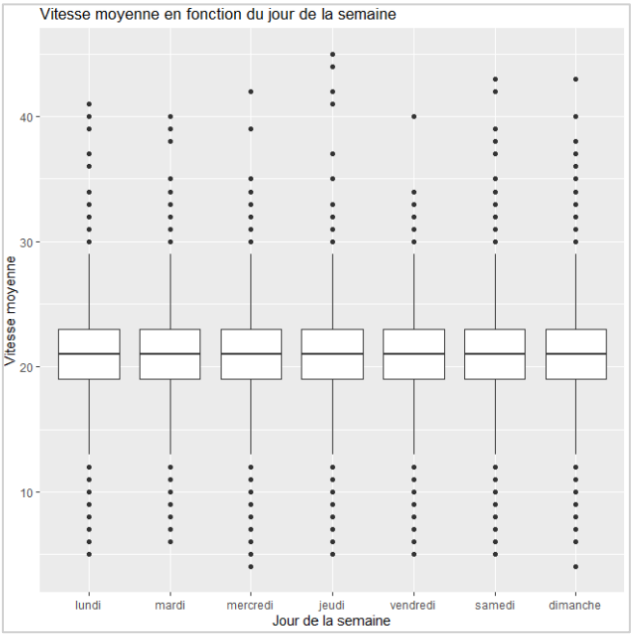


Image 3.2 - Boxplot pour la vitesse moyenne des vélos selon les jours de la semaine

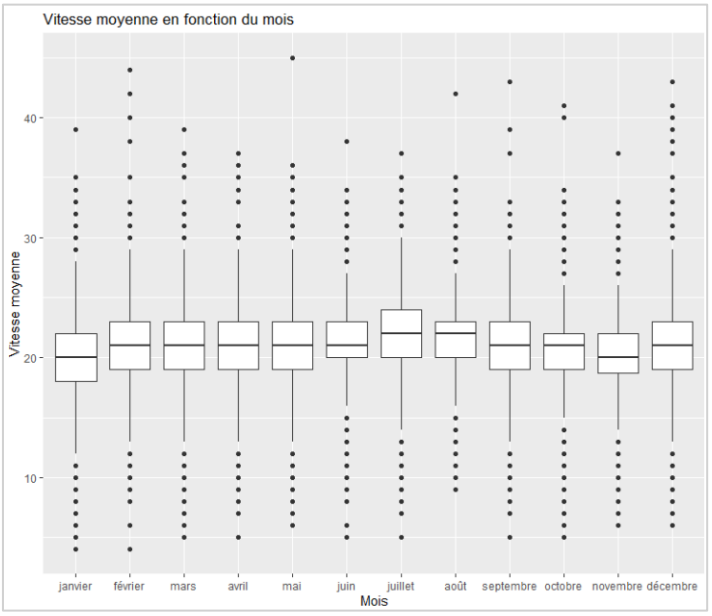


Image 3.3 - Boxplot pour la vitesse moyenne des vélos selon les mois

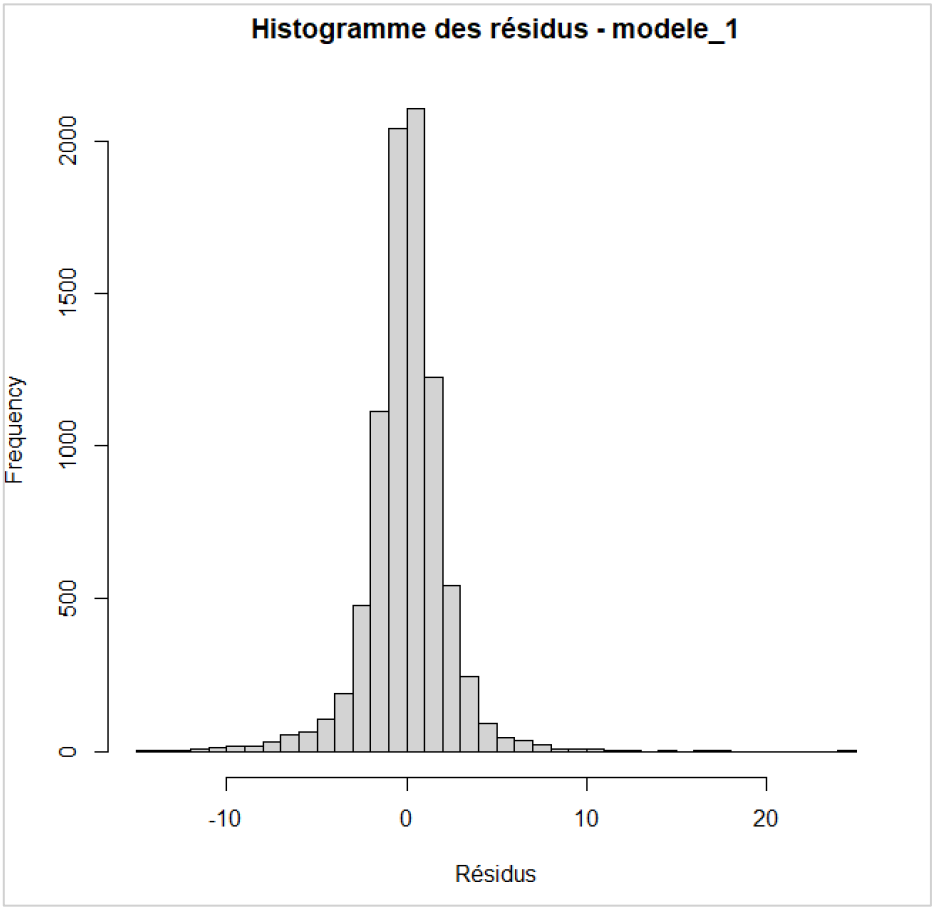


Image 3.4 – Histogramme des résidus du « modele_1 »

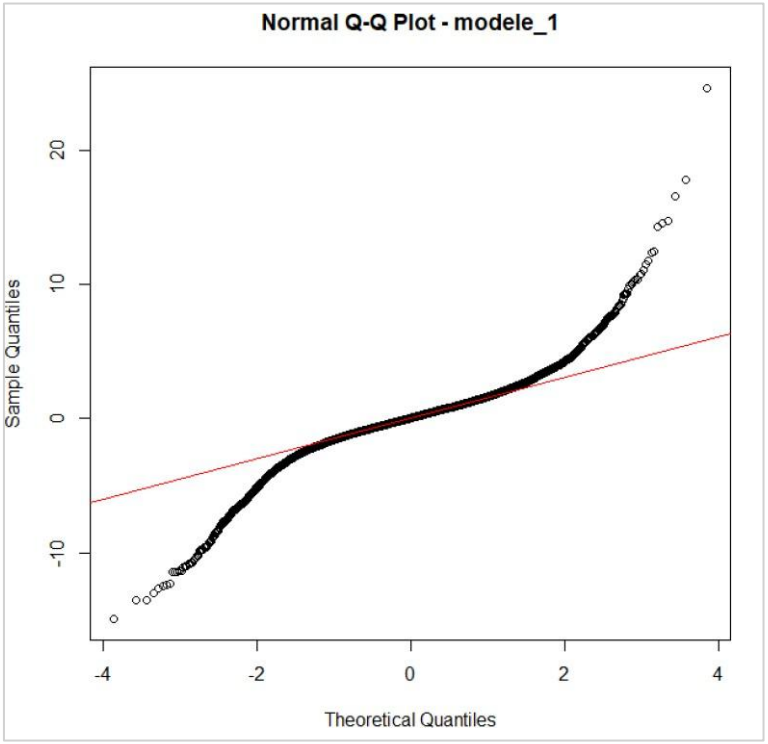


Image 3.5 – Q-Q Plot des résidus du « modele_1 »

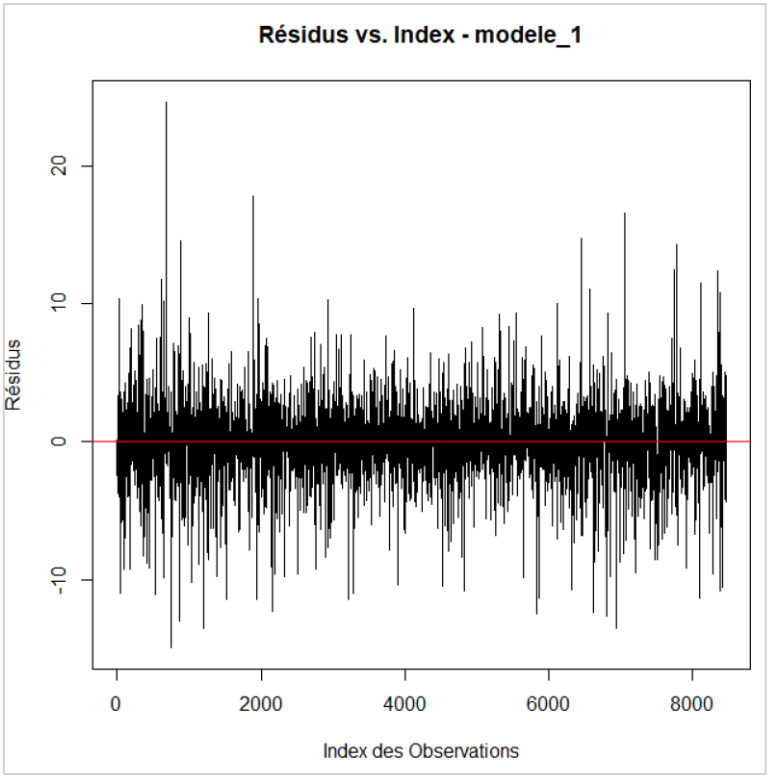


Image 3.6 - Graphique de résidus en fonction de l'index des observations pour « modele_1 »

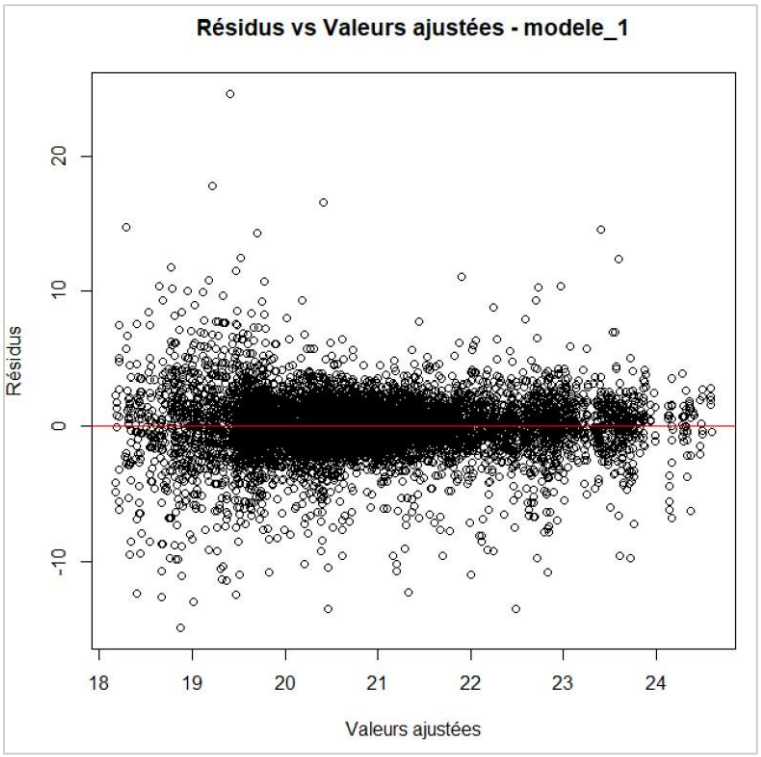


Image 3.7 - Graphique des résidus en fonction des valeurs ajustées du « modele_1 »

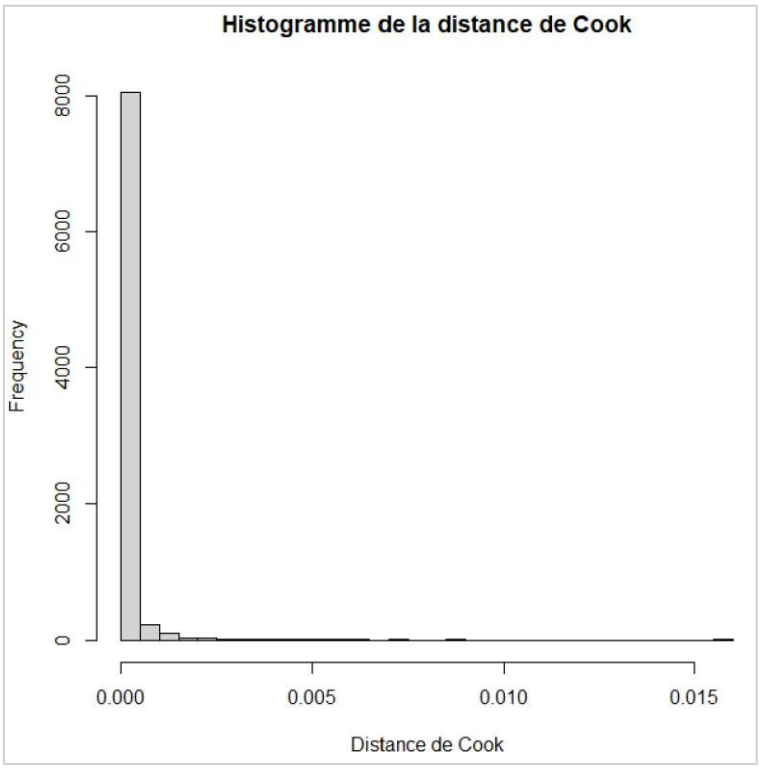


Image 3.8 -

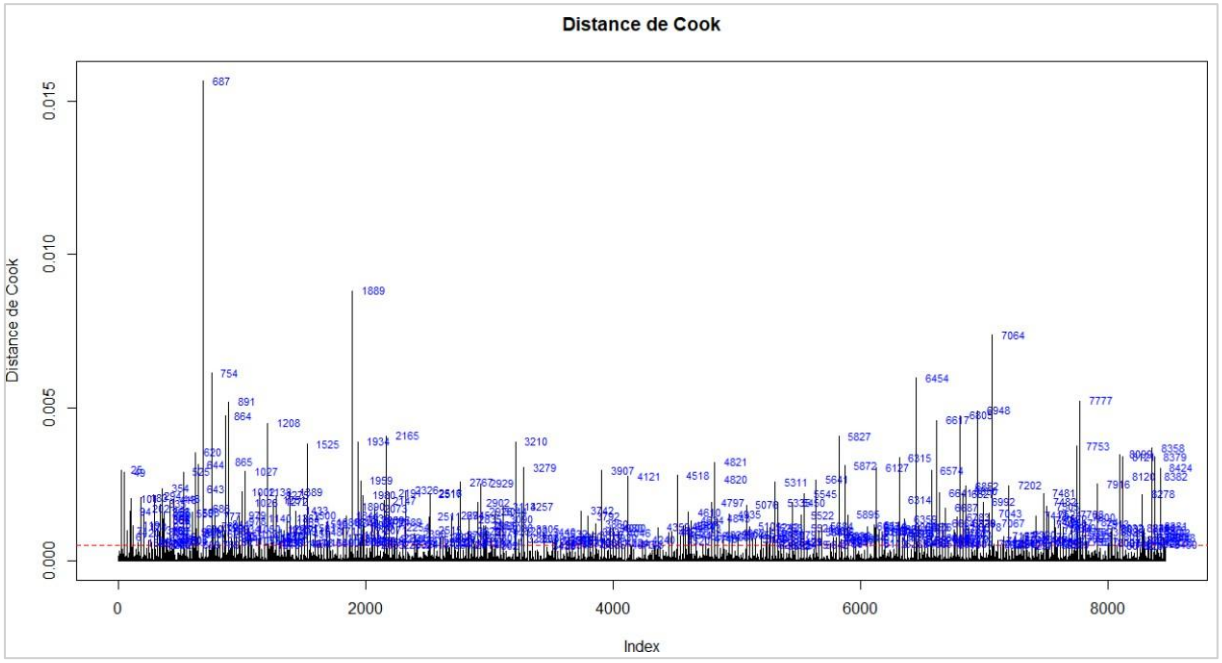


Image 3.9 -

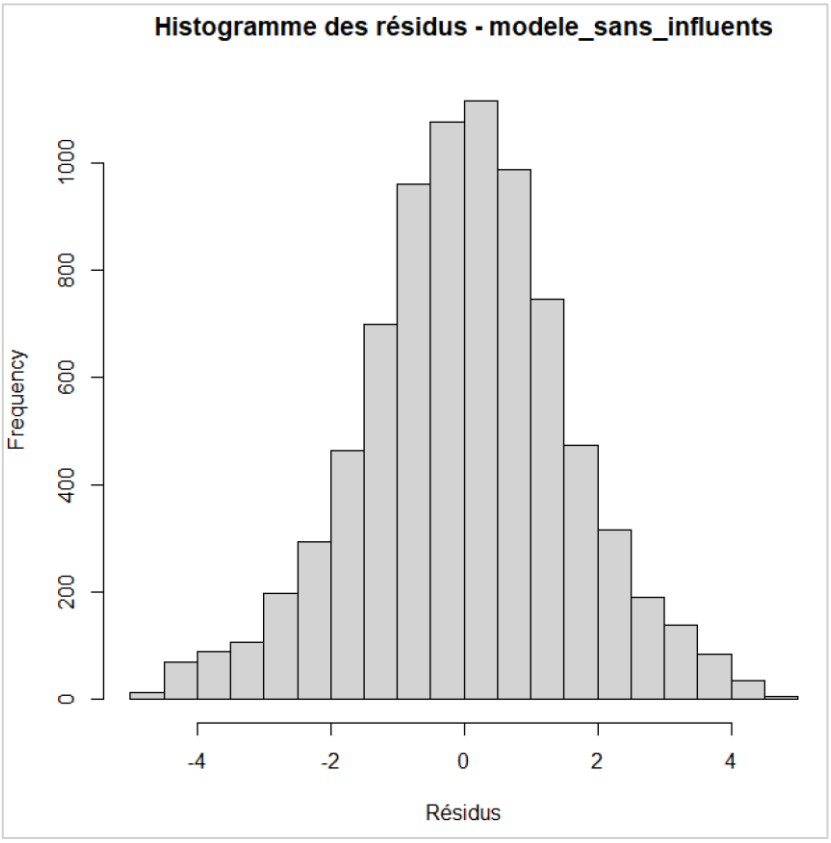


Image 3.10 -

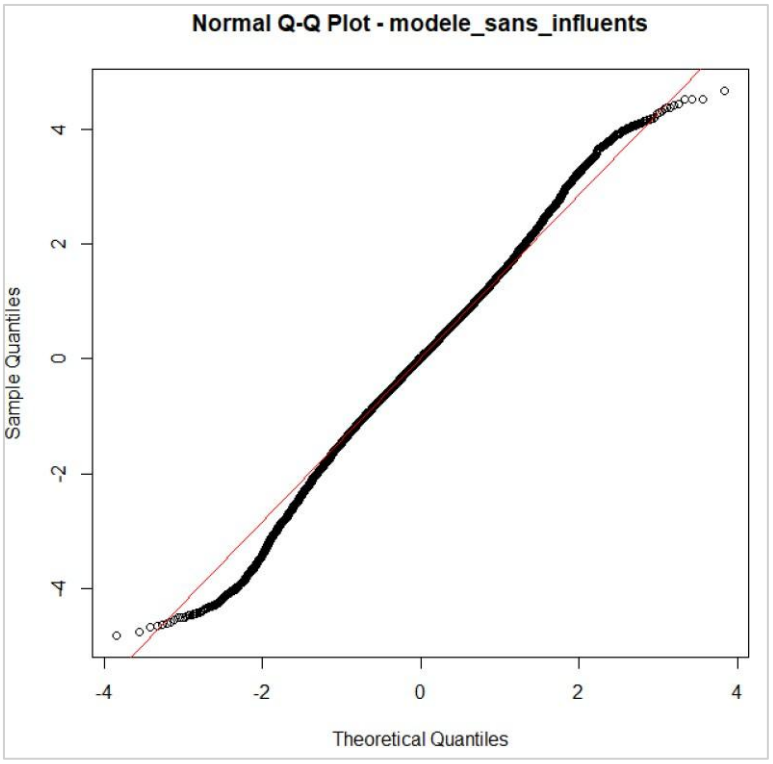


Image 3.13

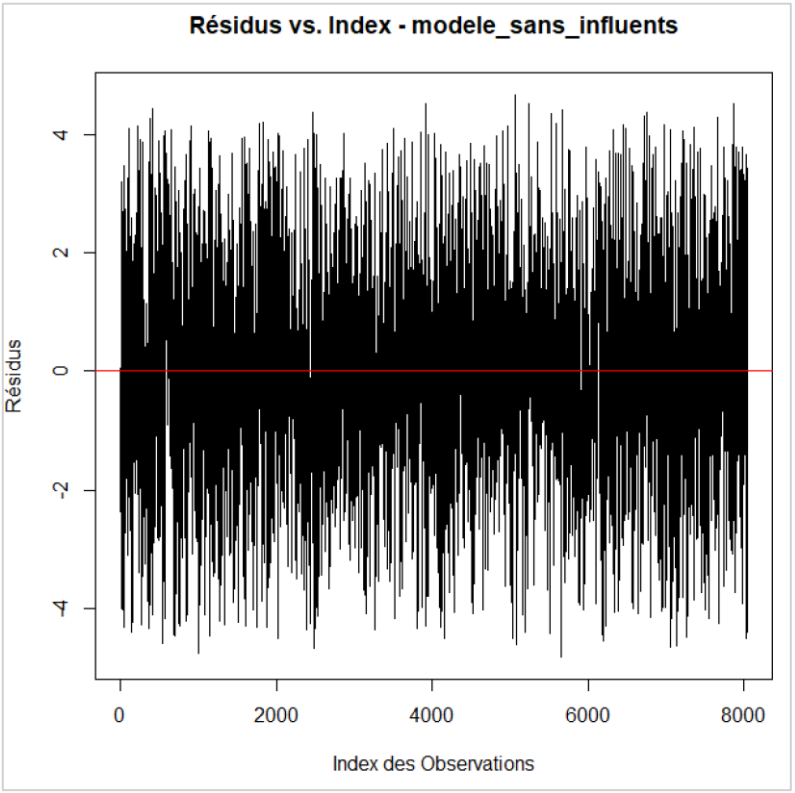


Image 3.12 -

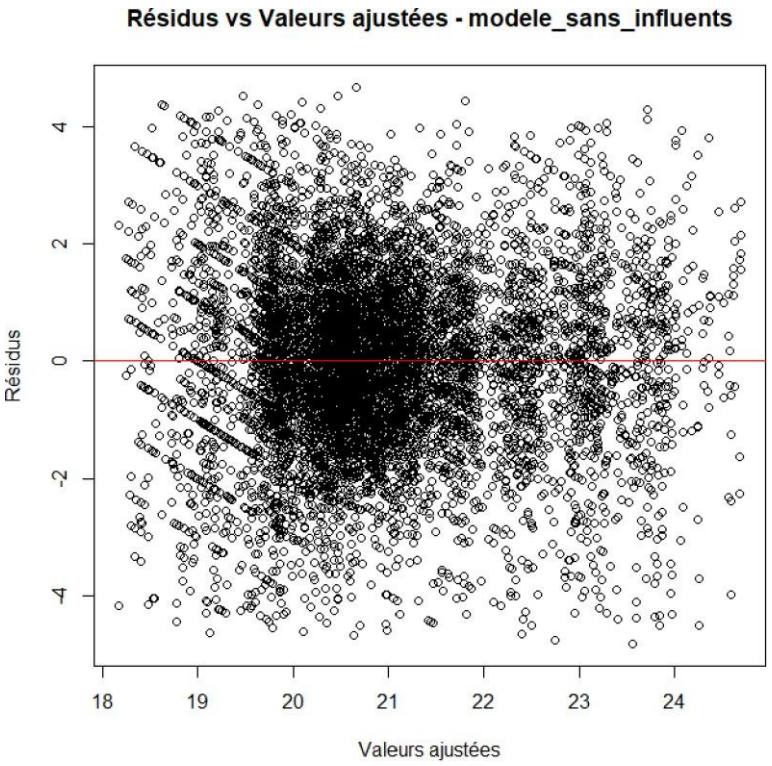


Image 3.13

	Residual_standard_error	degrees_of_freedom	Multiple_R_squared	Adjusted_R_squared	F_statistic	p_value
modele_1	2.220131	8426	0.2292498	0.2254994	61.12699	0
modele_2_pond	6.899155	30067	0.2612740	0.2602667	259.36959	0

Table 3.1 – Principales statistiques obtenues dans « modele_1 » et « modele_2_pond »

	Residual_standard_error	degrees_of_freedom	Multiple_R_squared	Adjusted_R_squared	F_statistic	p_value
modele_1	2.2201308	8426	0.2292498	0.2254994	61.12699	0
modele_log	0.1178679	8426	0.2111670	0.2073287	55.01469	0
modele_square	91.5259606	8426	0.2313846	0.2276446	61.86757	0
modele_cubic	3053.0146303	8426	0.2181197	0.2143151	57.33135	0
modele_sqrt	0.2523550	8426	0.2224631	0.2186797	58.79962	0
modele_cuberoot	0.1032585	8426	0.2192445	0.2154454	57.71003	0

Table 3.2 – Principales statistiques des modèles avec transformation de la variable expliquée

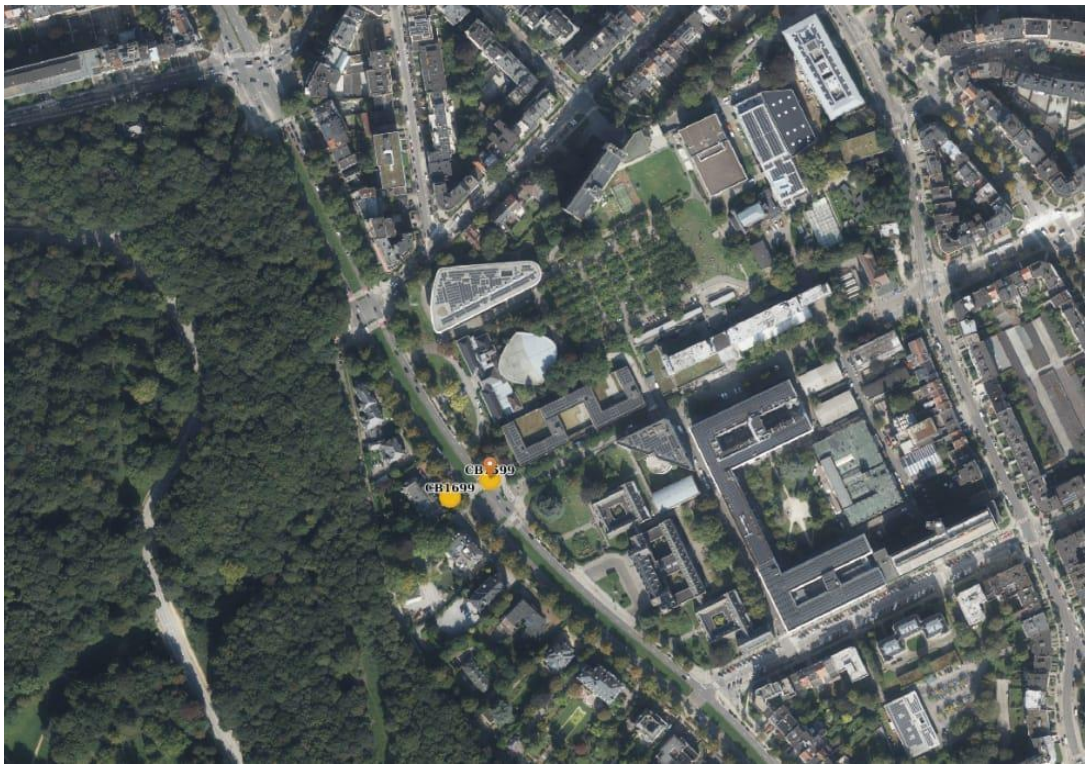
```
     date             heure_pleine       mois          jour_semaine vacances average.speed       count
Min.   :2024-01-02   2      : 73   janvier : 55   lundi    :61    0:296    Min.   : 4.00   Min.   :  1.000
1st Qu.:2024-03-08   1      : 59   octobre : 43   mardi    :62    1:128    1st Qu.:13.50   1st Qu.:  1.000
Median :2024-06-16   3      : 58   février : 42   mercredi:71             Median :16.00   Median :  2.000
Mean   :2024-06-25   4      : 51   décembre: 40   jeudi    :47             Mean   :18.97   Mean   :  6.542
3rd Qu.:2024-10-09   24     : 40   mars    : 38   vendredi:48             3rd Qu.:25.28   3rd Qu.:  4.000
Max.   :2025-01-01   5      : 36   novembre: 36   samedi   :54             Max.   :44.00   Max.   :119.000
                     (Other):107   (Other) :170   dimanche:81
```

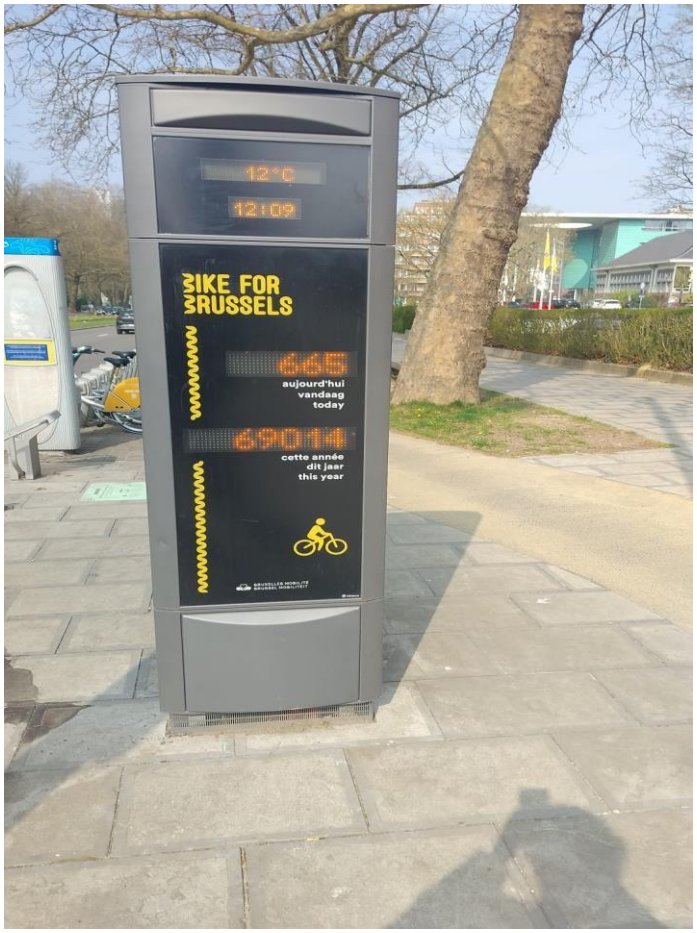
Table 3.3 – Résumé des caractéristiques des observations influentes selon les distances de Cook

	Residual_standard_error	degrees_of_freedom	Multiple_R_squared	Adjusted_R_squared	F_statistic	p_value
modele_1	2.220131	8426	0.2292498	0.2254994	61.12699	0
modele_sans_influents	1.553028	8002	0.3820993	0.3789333	120.69026	0

Table 3.4 – Principales statistiques obtenues dans « modele_1 » et « modele_sans_influents »

Photos de la borne


CB1699


12°C
12:09
BIKE FOR BRUSSELS
665
aujourd'hui
vandaag
today
69014
cette année
dit jaar
this year