Comment prédire la quantité de vélos loués à Séoul en fonction de différents facteurs temporels et météorologiques ?

Léa Aumagy, Abdeldjallil Boukhalfa

Aix-Marseille School of Economics

Master 1 Economie 2022-2023

Sommaire

amse
école d'économie d'aix-marseille
aix-marseille school of economics

Sommaire

Introduction

- Mobilité urbaine : enjeu majeur de ces dernières années.
- Hausse de l'utilisation des vélos : mise en place de vélos à loués par la ville.
- Séoul : 10 millions d'habitants : lance en 2015 "Seoul Bike".

Sommaire

Étude Bibliographique

Nombreuses recherches effectuées :

- "The influence of weather conditions on the usage of the Barclays Cycle Hire" de Rolf van Lieshout et Jelmer Strijkstra (2015),
- "Weather and Public Bikesharing: Impact on Use and Ridership" de John A. R. Adams et al. (2016),
- "Forecasting Bike Rental Demand with Weather Data" de Akhilesh Mishra et al. (2017),
- "Using data mining techniques for bike sharing demand prediction in metropolitan city" de Sathishkumar V E, Jangwoo Park et Yongyun Cho, (2020).

Sommaire

Présentation de la base : Seoul Bike Sharing Demand

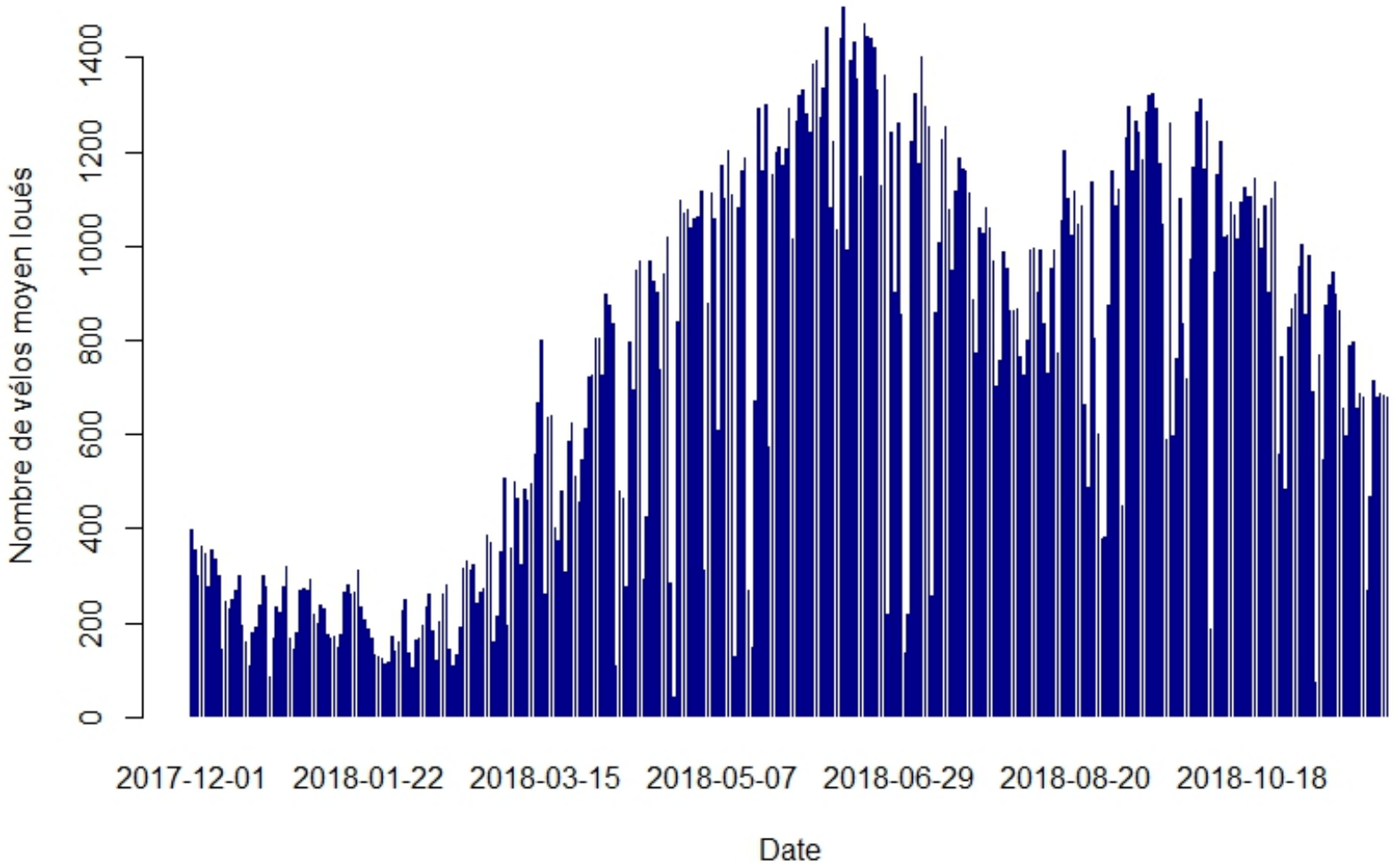
- Source de téléchargement : UCI Machine Learning Repository
- Source initiale des données : http://data.seoul.go.kr
- 8465 observations, 1 par heure de la journée du 01/12/2017 au 30/11/2018
- Mesure du nombre de vélos loués à chaque observation
- 14 variables météorologiques et temporelles.

Présentation des variables

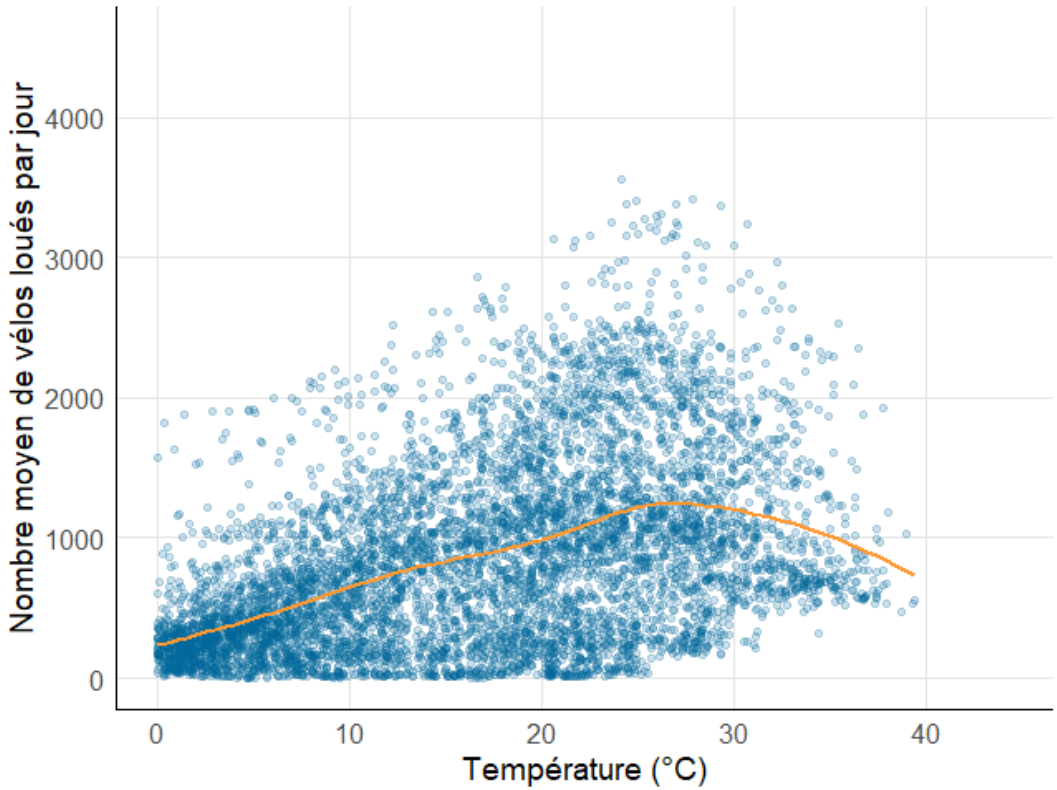
Table: Les variables et leurs descriptions.

Variables	**Mesures**	**Type**	**Description**
Date	année-mois-jour	-	Indique la date d'observation
Rented Bike count	Compte	Continue	Indique le nombre de vélod loués
Hour	Heure	Continue	Indique l'heure d'observation
Temperature	C	Continue	Indique la température relevée
Humidity	%	Continue	Indique l'humidité mesurée
Windspeed	m/s	Continue	Indique la vitesse du vent mesurée
Visibility	10m	Continue	Indique la distance de visibilité mesurée
Dew point temperature	C	Continue	Mesure le point de rosée
Solar radiation	MJ/m2	Continue	Indique le taux de radiation relevé
Rainfall	mm	Continue	Indique la quantité de pluie mesurée
Snowfall	cm	Continue	Indique la quantité de neige mesurée
Seasons	Saisons	Categorielle	Automne, Hiver, Printemps, Eté
Holiday	-	Categorielle	Indique si période de vacances ou non
Functional Day	-	Categorielle	Indique si le service fonctionne
Week status	-	Categorielle	Indique si jour de semaine ou week-end
Day of the week	Jours	Categorielle	Lundi, Mardi, ... , Dimanche

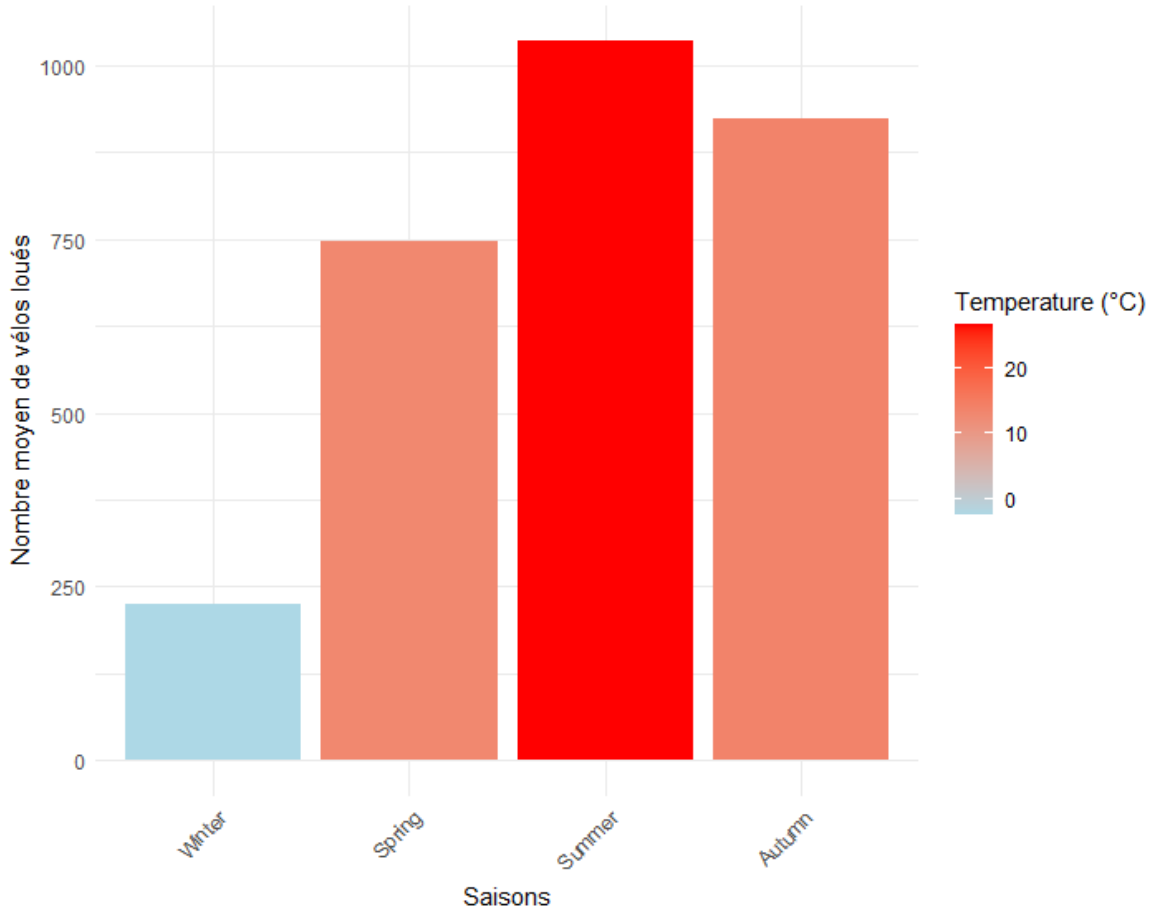
Evolution du nombre de vélos moyen loués durant l'année d'observation.



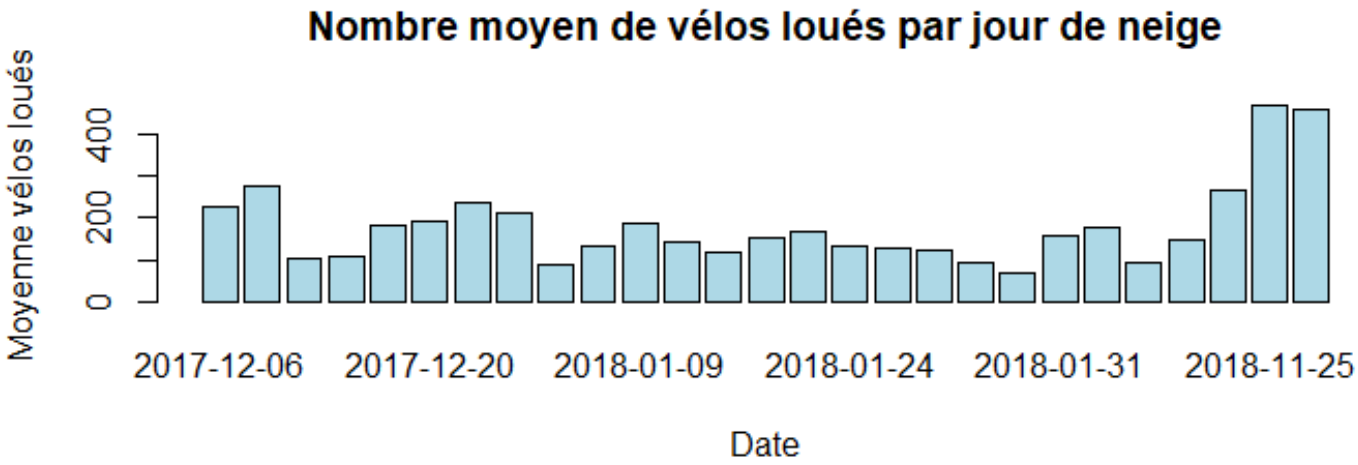
Evolution du nombre de locations moyen par heure en fonction de la température.

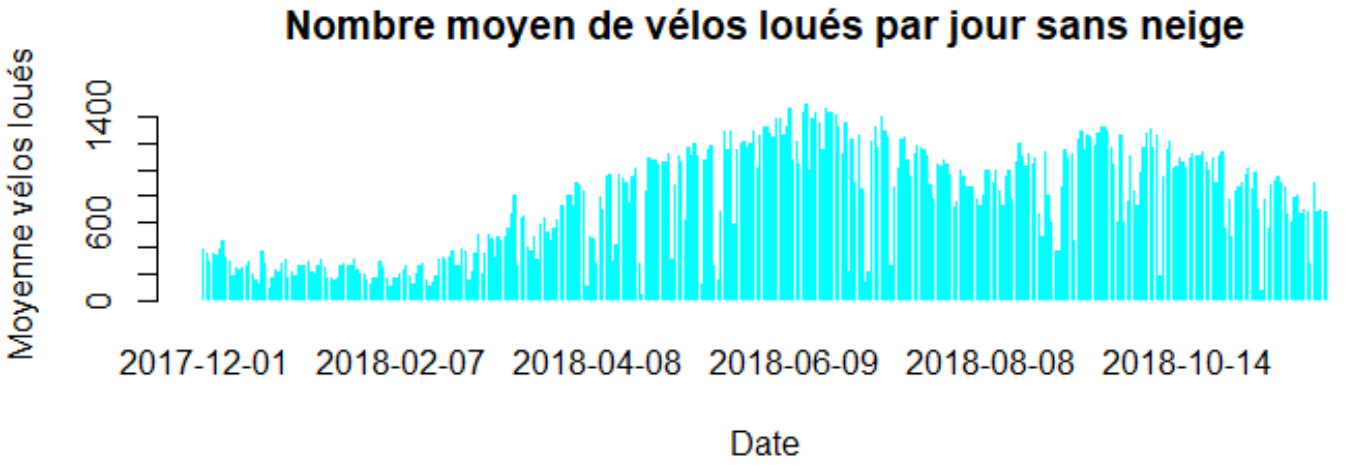


Evolution du nombre de locations journalières moyennes par saisons.

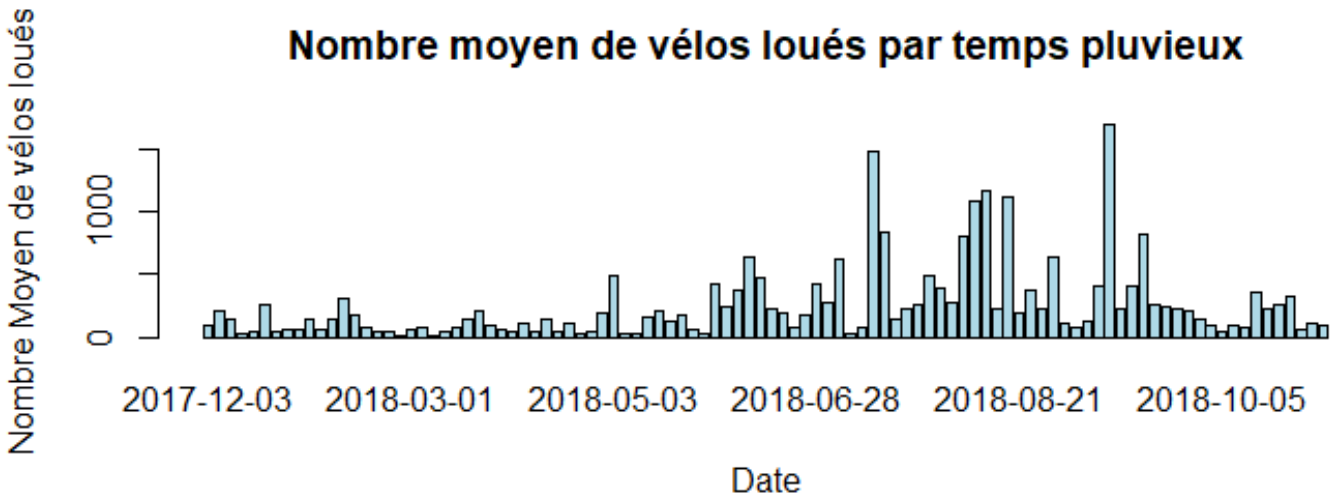


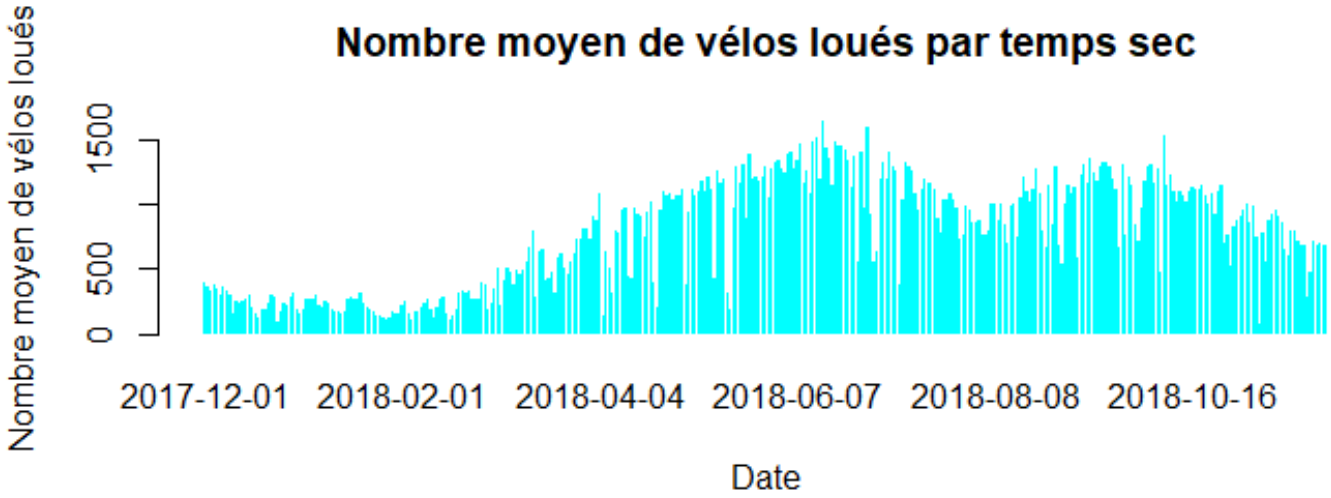
Visualisation de l'influence de la neige



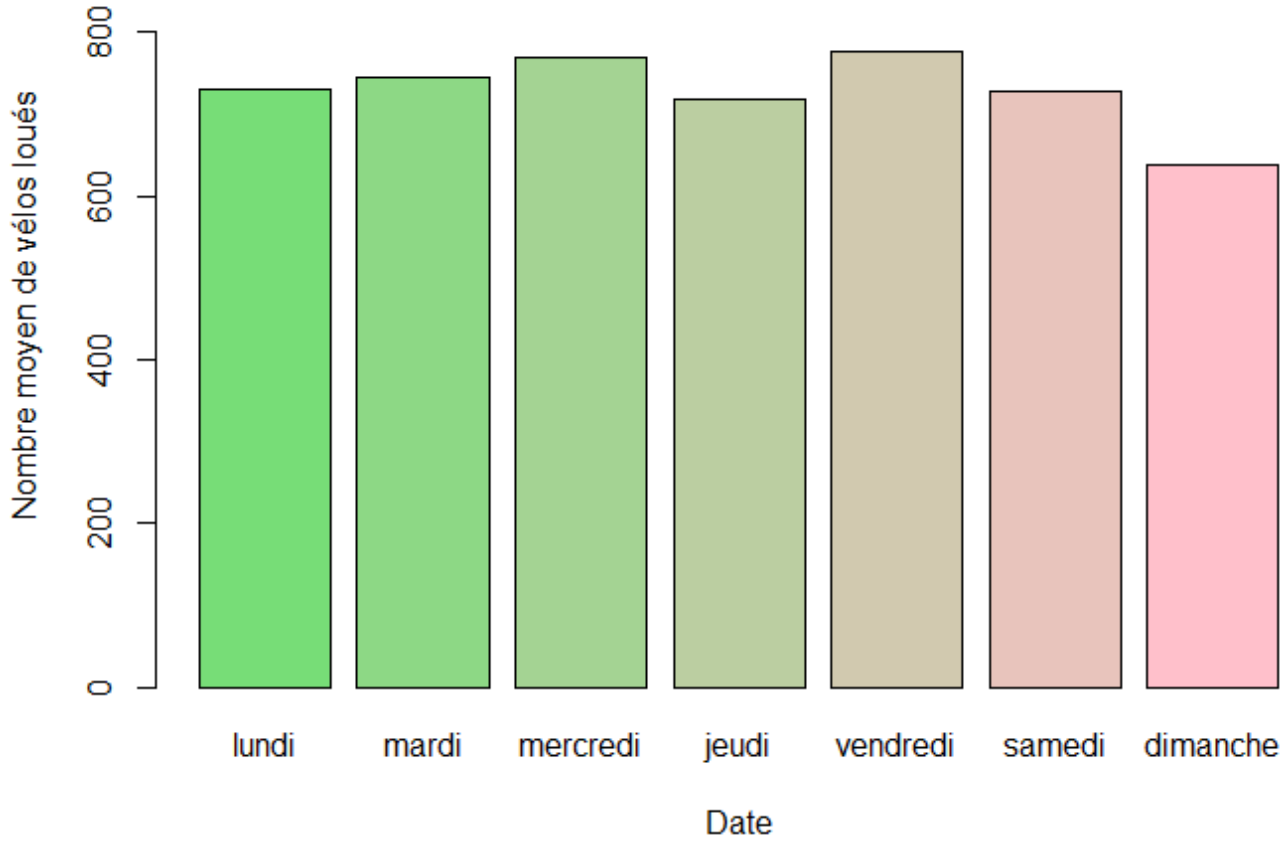


Visualisation de l'influence de la pluie

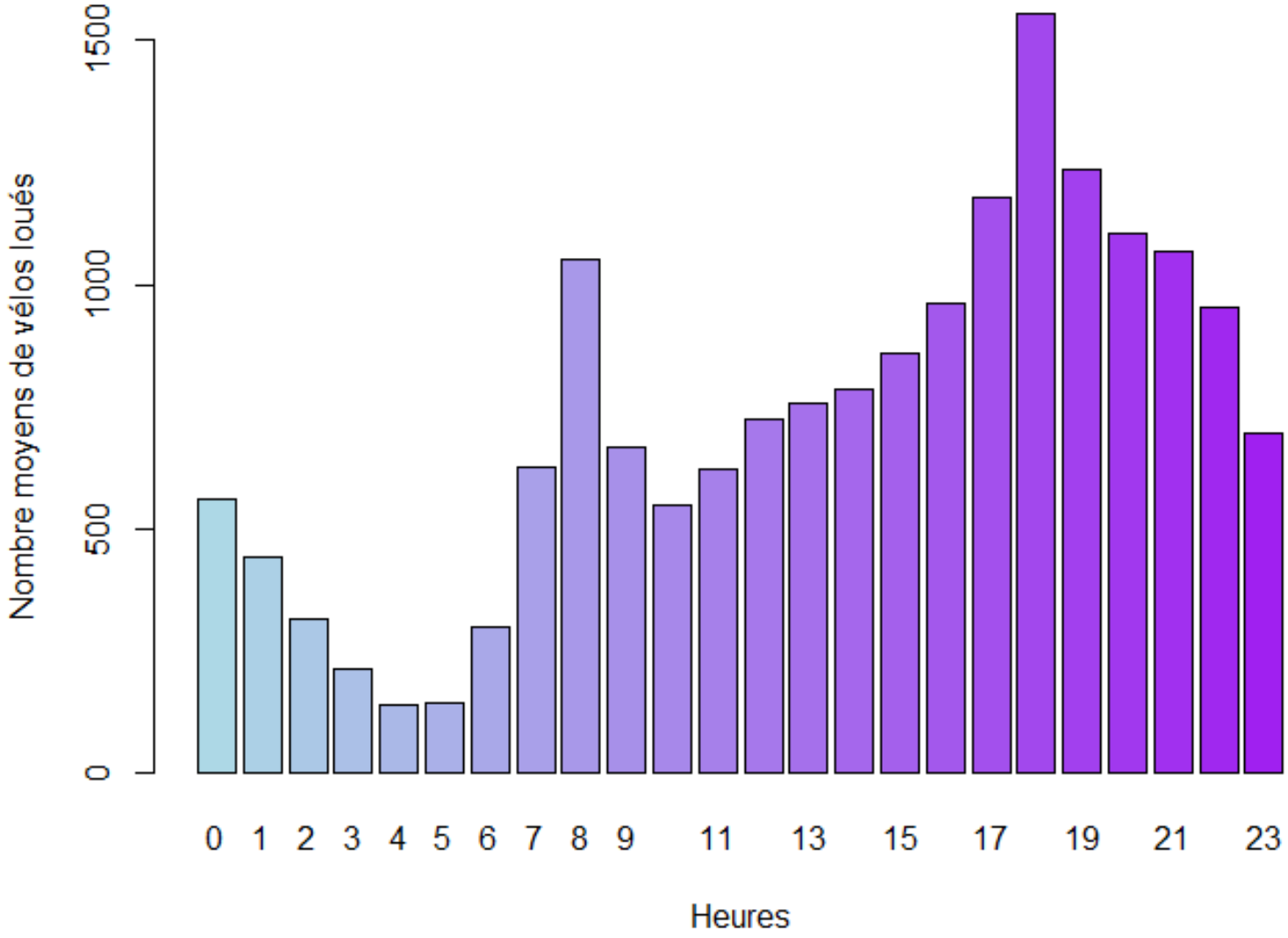




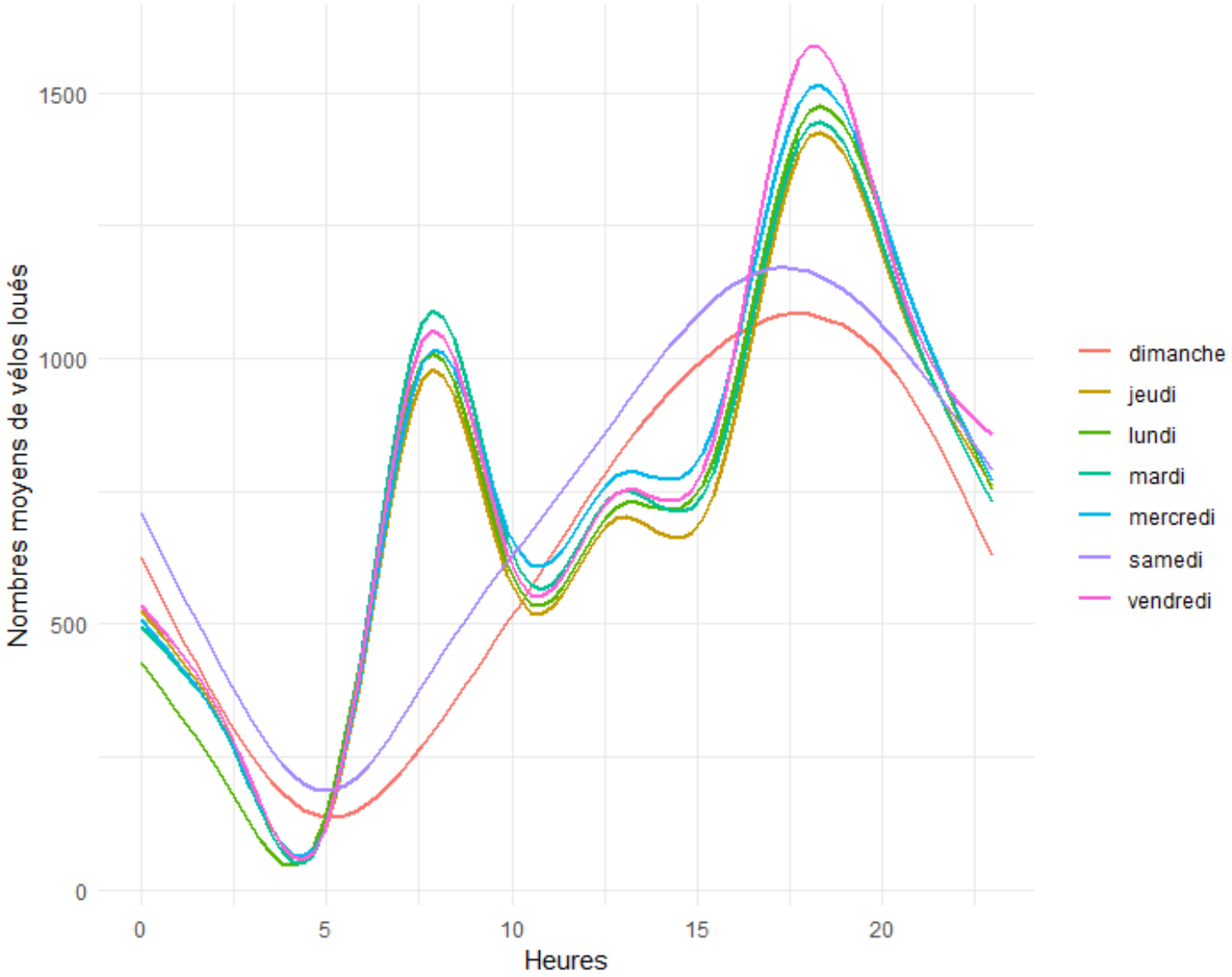
Visualisation de l'influence du jour de la semaine



Visualation de l'influence de l'heure de la journée



Visualiation de l'influence de l'heure en fonction des jours



Test ANOVA

Table: Tableau des résultats du test d'Anova

Variable	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Temperature.C.	1	1.106e+09	1.106e+09	5749.892	<2e-16 ***
Hour	1	4.498e+08	4.498e+08	2338.760	<2e-16 ***
Seasons	1	6.498e+07	6.498e+07	337.811	<2e-16 ***
Humidity...	1	1.808e+08	1.808e+08	940.002	<2e-16 ***
Wind.speed..m.s.	1	5.297e+04	5.297e+04	0.275	0.600
Rainfall.mm.	1	4.221e+07	4.221e+07	219.441	<2e-16 ***
Snowfall..cm.	1	8.953e+04	8.953e+04	0.465	0.495
Solar.Radiation..MJ.m2.	1	2.151e+07	2.151e+07	111.813	<2e-16 ***
Visibility..10m.	1	1.258e+05	1.258e+05	0.654	0.419
Dew.point.temperature.C.	1	7.455e+05	7.455e+05	3.876	0.049 *
Residuals	8454	1.626e+09	1.923e+05		

Sommaire

Définition

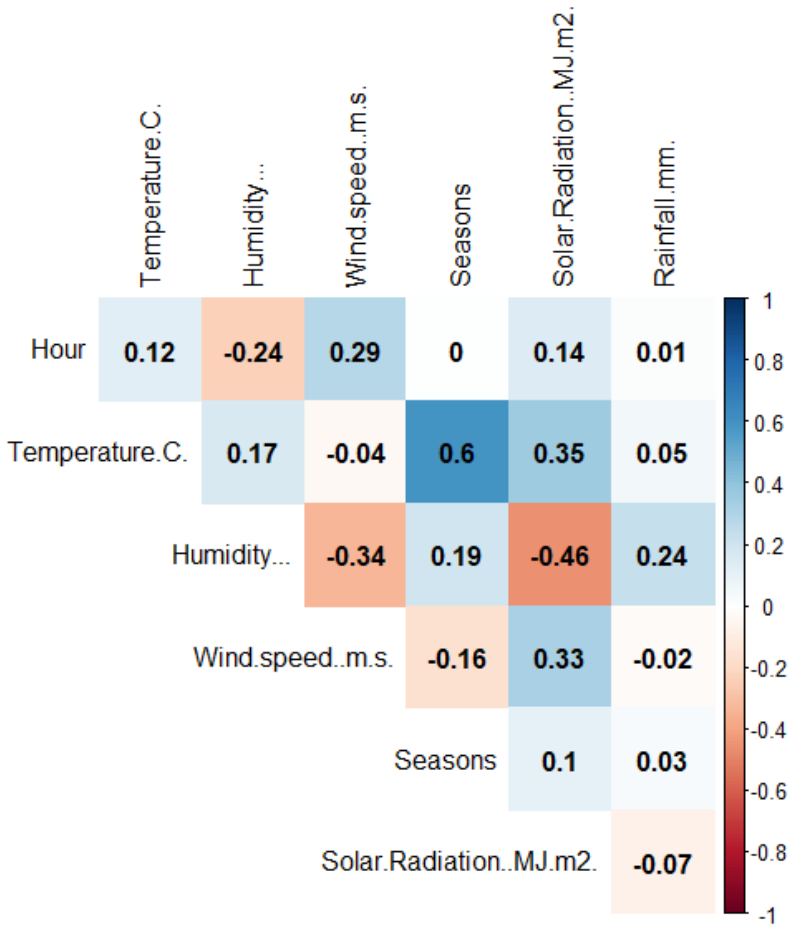
Le modèle de régression linéaire est :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

où :

- y : variable à expliquer (dépendante).
- $x_1, x_2, \ldots, x_n$: variables explicatives (indépendantes).
- β_0 : la constante.
- $\beta_1, \beta_2, \ldots, \beta_n$: coefficients des variables

Choix des variables : matrice de corrélation



Le modèle

Le modèle de régression linéaire est :

$$Rented.Bike.Count = \beta_0 + \beta_1 Hours + \beta_2 Temperature + \beta_3 Humidity + \beta_4 Rain + \beta_5 Radiation + \varepsilon$$

Table: Résumé du modèle de régression linéaire

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	511.8872	22.2807	22.98	<2e-16 ***
Hour	28.5202	0.7393	38.58	<2e-16 ***
Temperature.C.	32.6085	0.4766	68.41	<2e-16 ***
Humidity...	-8.0756	0.3087	-26.16	<2e-16 ***
Rainfall.mm.	-64.2459	4.4802	-14.34	<2e-16 ***
Solar.Radiation..MJ.m2.	-84.8888	7.2729	-11.67	<2e-16 ***

Signif. codes: 0 '' 0.001 '' 0.01 '' 0.05 '.' 0.1 ' ' 1

Residual standard error: 449.1 on 8459 degrees of freedom

Multiple R-squared: 0.5115, Adjusted R-squared: 0.5112

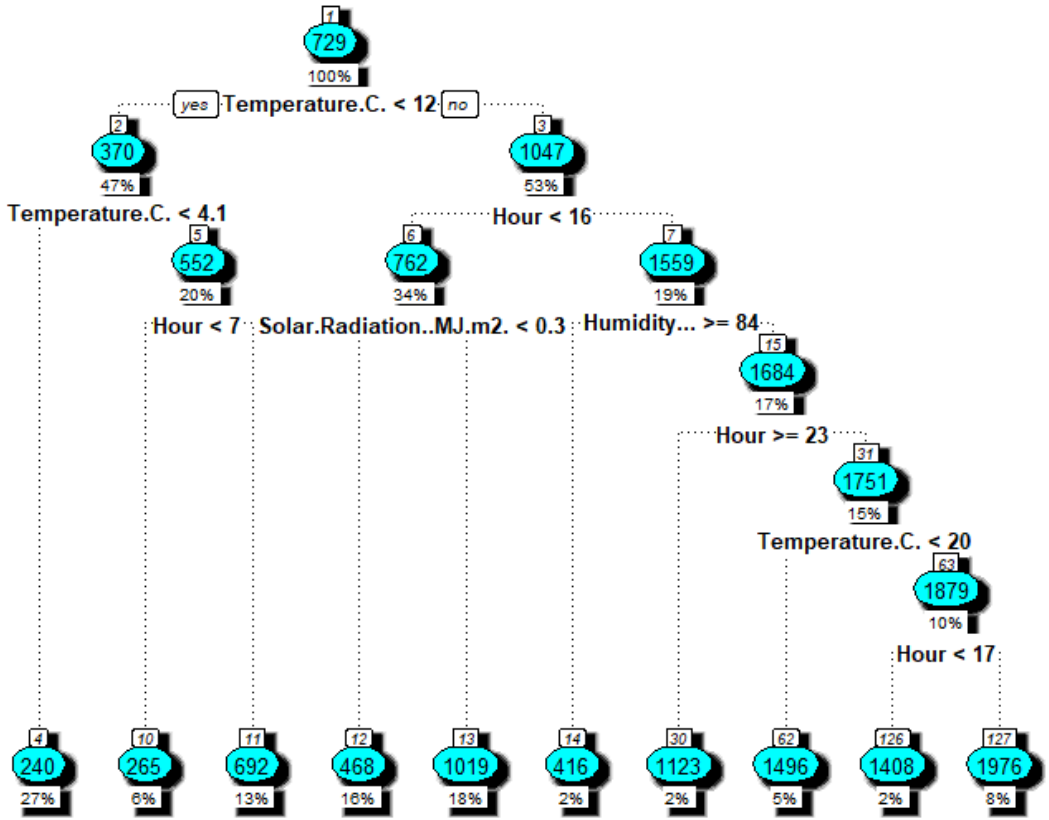
F-statistic: 1772 on 5 and 8459 DF, p-value: < 2.2e-16

Sommaire

Introduction d'une base de donnée

- Base de donnée sur les données météorologique à Séoul
- Source initiale des données : http://data.seoul.go.kr/
- 52440 observations, 1 par heure de la journée du 30 mars 2017 au 29 mars 2023
- Variables : Date, Hour, Temperature, Rainfall, Windspeed, Humidity, SolareRadiation, Snowfall et Visibility

Prédiction avec arbre de décision



Prédiction avec arbre de décision

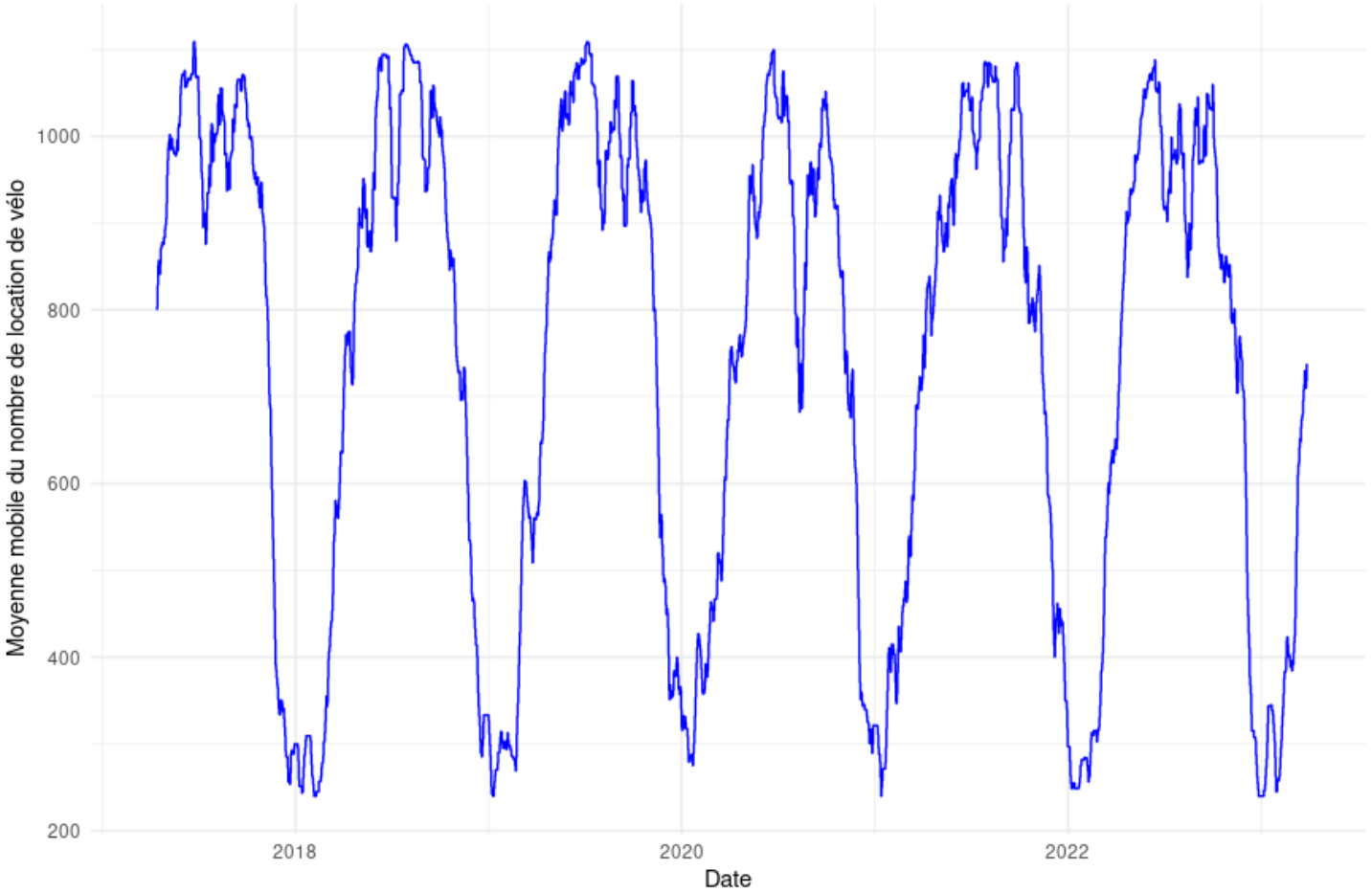


Figure: Prédictions pour la base de données "MétéoSéoul".

Prédiction avec arbre de décision

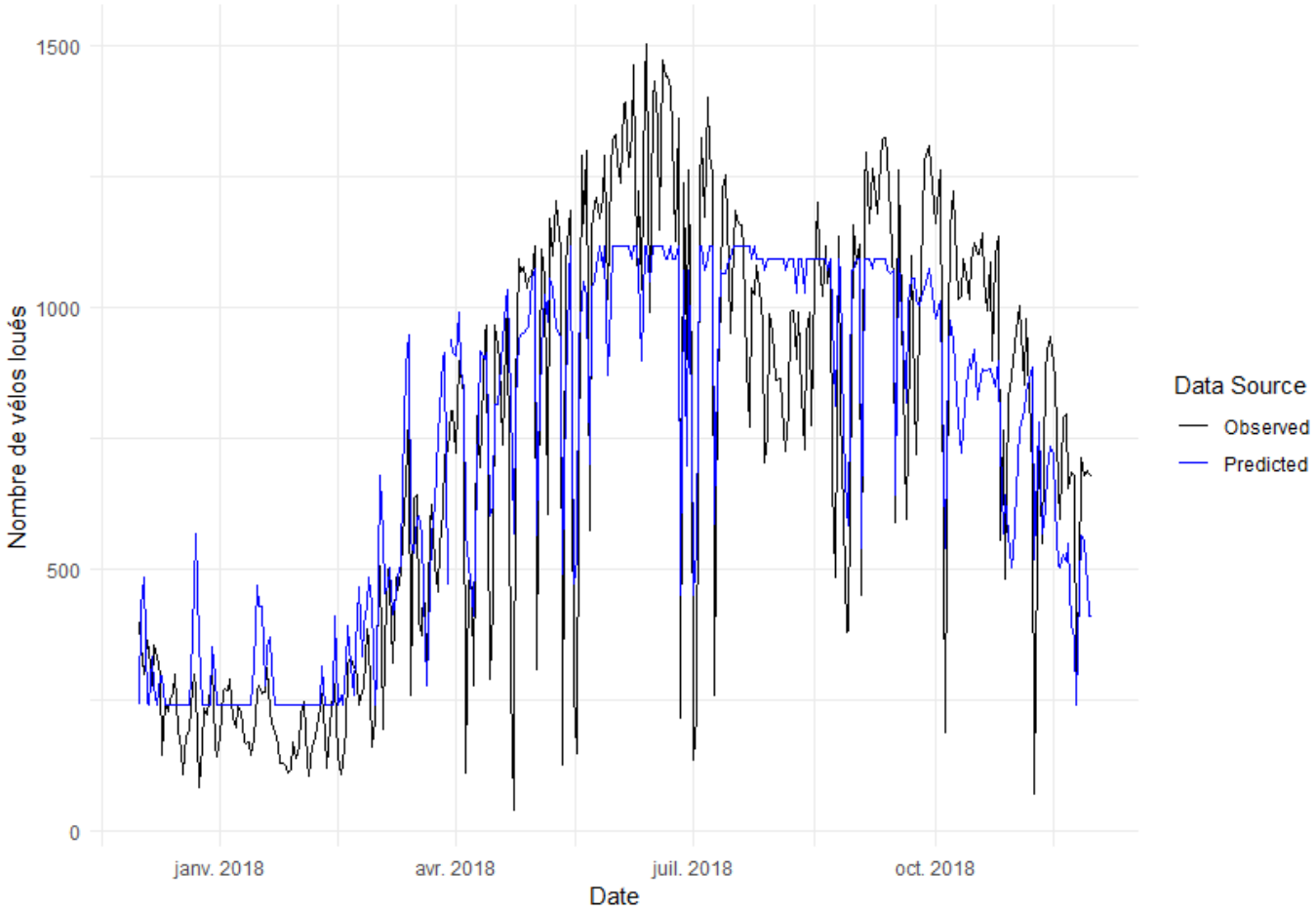


Figure: Comparaison données réelles et prédites de 2018

Le modèle XGBoost

- Extreme Gradient Boosting : algorithme de machine learning populaire utilisé pour la régression
- Construction de plusieurs arbres de prédictions et calculs des résidus
- Aggregation des arbres et régularisation des prédictions

Prédiction avec le modèle XGBoost

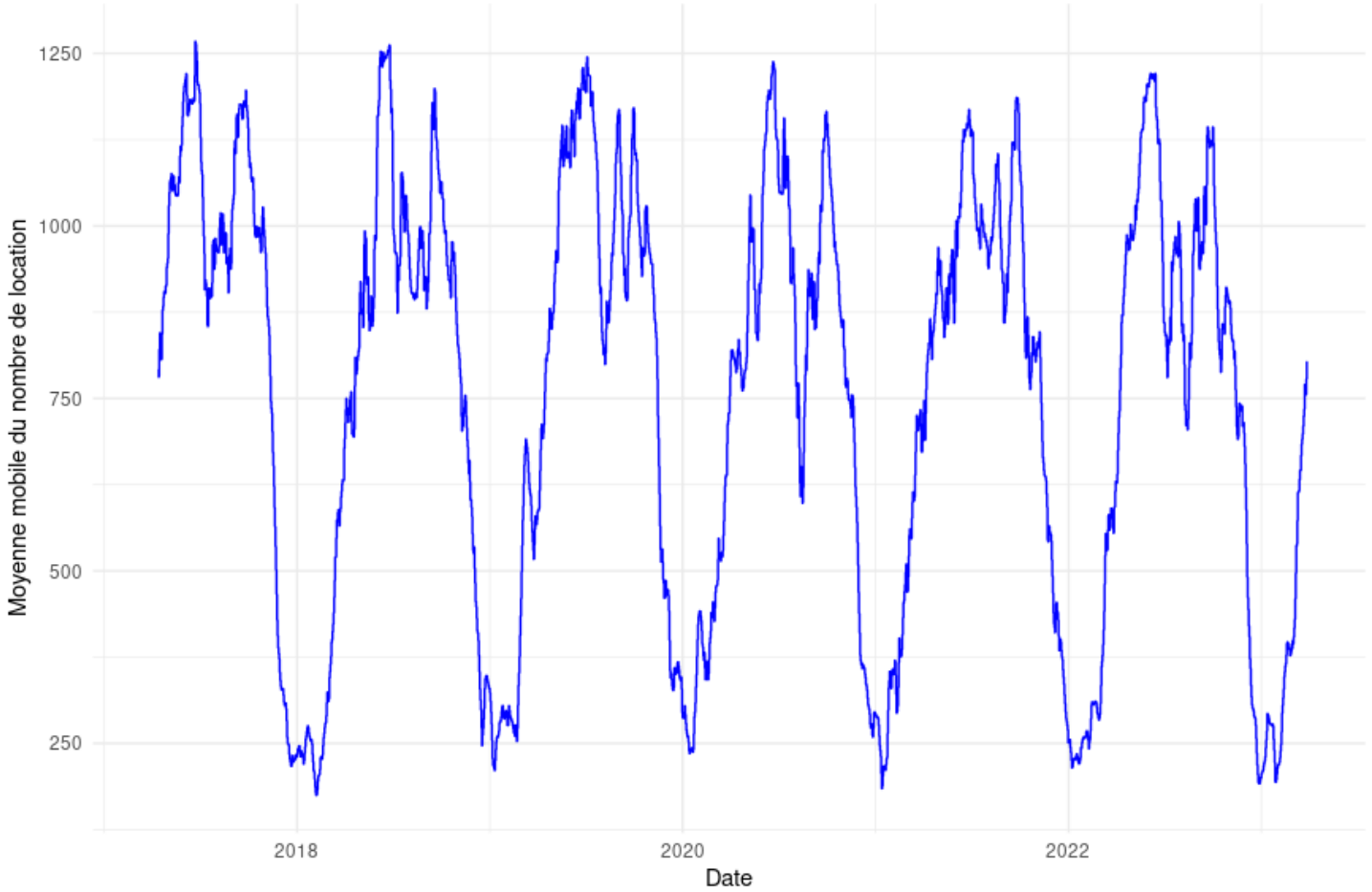


Figure: Prédictions pour la base de données "MétéoSéoul"

Prédiction avec le modèle XGBoost

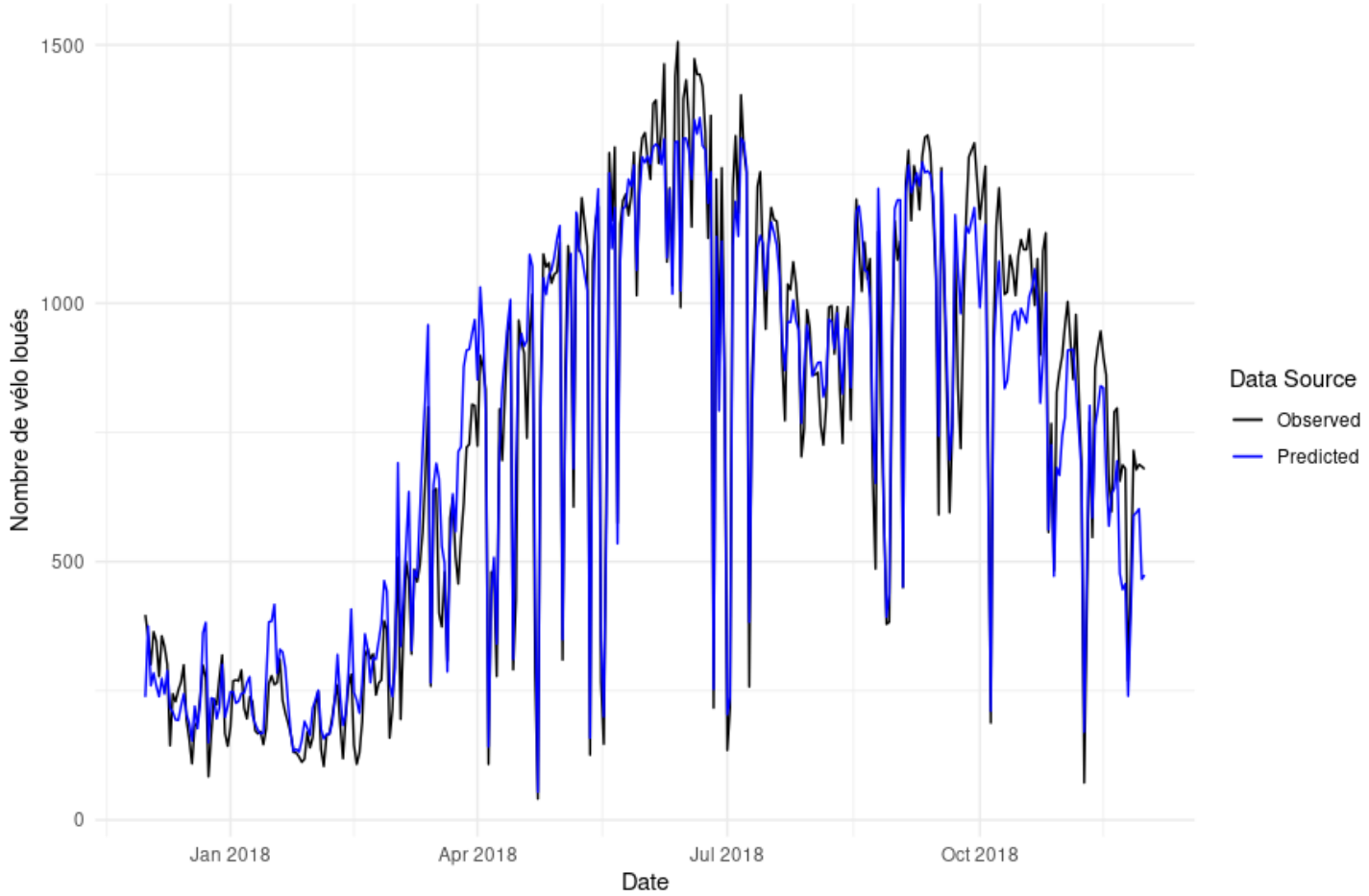


Figure: Comparaison données réelles et prédites par XGboost de 2018

Prédiction avec le modèle XGBoost

Nous pouvons observer le contrefactuel pour l'année 2020 sans la pandémie COVID-19.

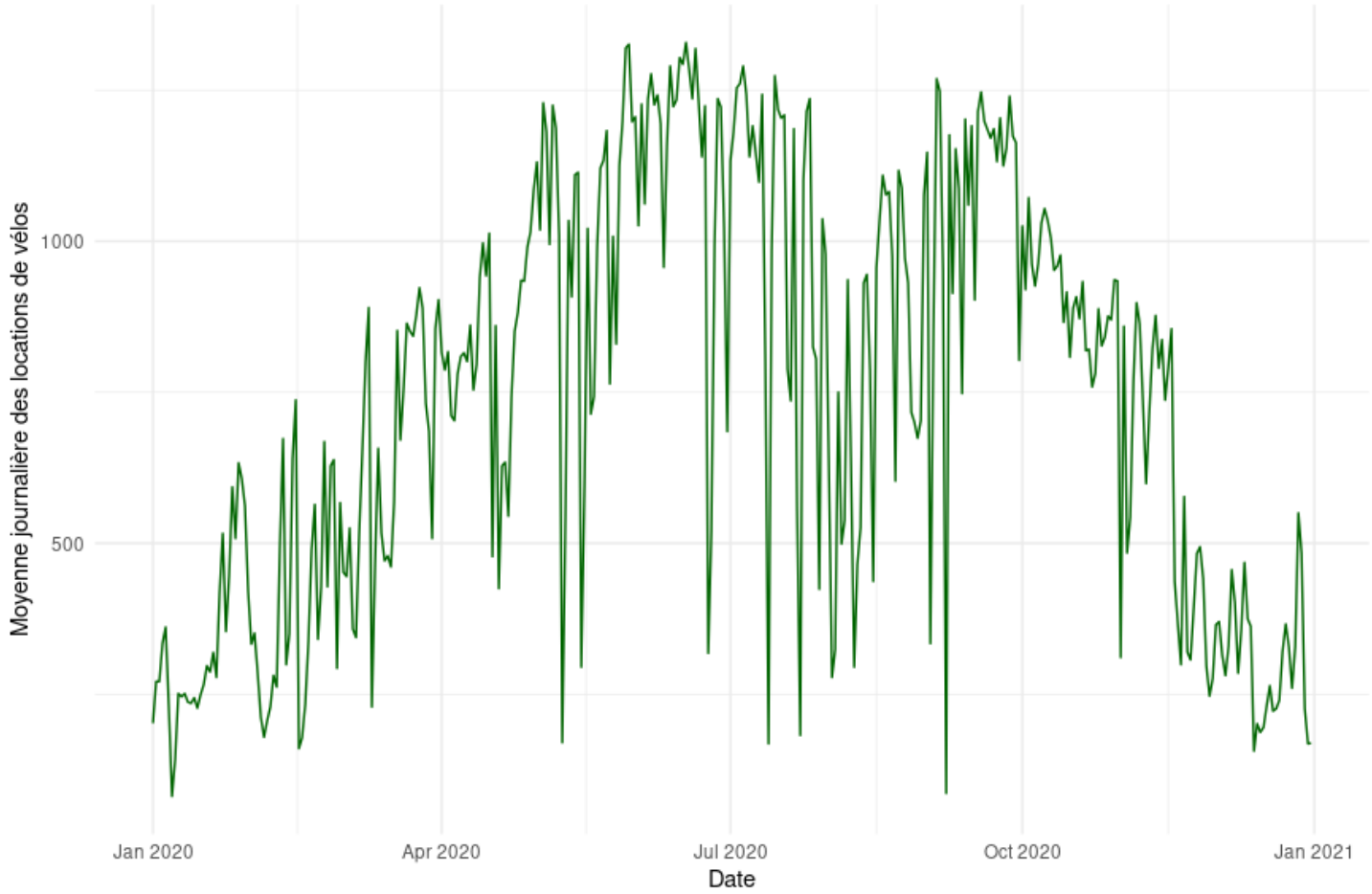


Figure: Prédiction par XGBoost pour l'année 2020

Prédiction des données météorologiques pour l'année à venir

- Création d'une 3eme base de données
- 8760 observations pour chaque heure de la journée entre le 30 mars 2023 et le 30 avril 2024
- Prédiction des données météorologiques suivant la formule : $X_t = \frac{1}{7}\sum_{i=t-7}^{t-1} X_i$.
- Prédiction du nombre de location de vélo par heure avec le modèle XGBoost

Prédiction du nombre de vélos loués pour l'année à venir

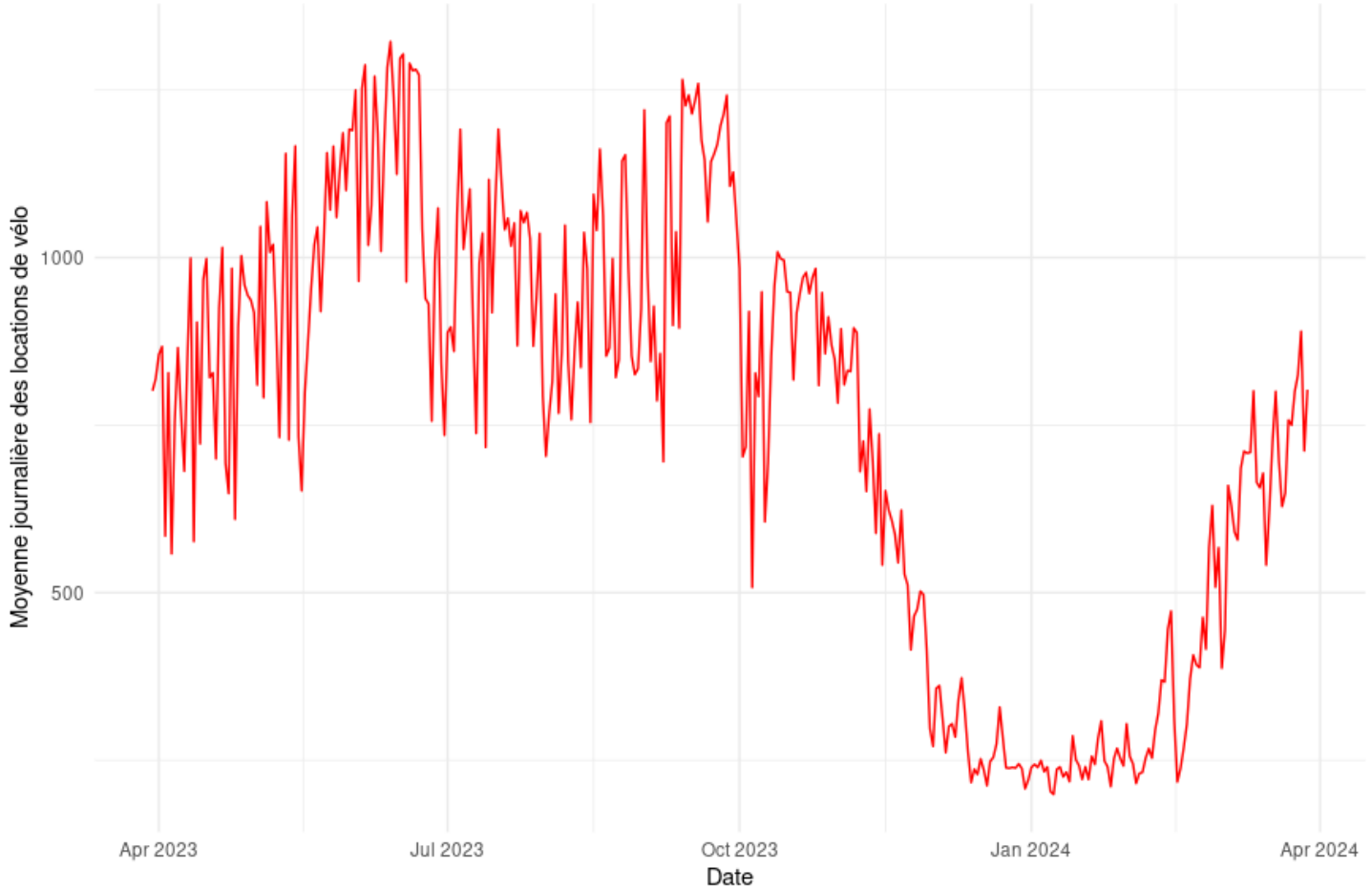


Figure: Prédiction par XGBoost pour avril 2023 à avril 2024

Sommaire

Discussion

- Identification des variables significatives : Existence d'autres variables significatives
- Construction de modèle de régression linéaire
- Prédictions : arbre de décisions et XGBoost
- Prédictions pour l'année à venir : imprécision du modèle
- Existence de facteurs extérieurs : Pandémie ou réchauffement climatique

Sommaire

amse
école d'économie d'aix-marseille
aix-marseille school of economics


CONCLUSION