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

Software for economist III

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

Léa AUMAGY Abdeldjallil BOUKHALFA
Master 1 Economie

3 mai 2023

Sommaire

1 Introduction

Au cours de la dernière décennie, la mobilité urbaine est devenue l'un des sujets les plus importants en raison des enjeux économiques et environnementaux auxquels les grandes villes sont confrontées. Parmi les solutions de mobilité innovantes et écologiques, l'utilisation de vélos au quotidien est devenue de plus en plus populaire, que ce soit pour aller au travail, à l'école ou pour tout déplacement.
Face à cette tendance, les métropoles ont saisi l'opportunité économique et ont commencé à mettre à disposition des vélos en location pour la population..

Dans ce contexte, notre étude porte sur Séoul, la capitale et la plus grande ville de Corée du Sud, comptant 10 millions d'habitants en 2019 [1].Soucieuse de l'environnement et consciente de l'enjeu économique, la ville a lancé en 2015 un service de location de vélos appelé "Seoul Bikes". [2]

Notre étude vise à prédire le nombre de vélos loués à chaque heure de la journée en fonction des données météorologiques et temporelles. Notre problématique est la suivante :
Comment prédire la quantité de vélos loués à Séoul en fonction de différents facteurs temporels et météorologiques ?

Pour répondre à notre problématique, nous mettrons en œuvre plusieurs étapes méthodologiques. Tout d'abord, nous réaliserons une étude bibliographique sur le sujet de la prédiction du nombre de vélos loués dans les villes. Ensuite, nous utiliserons une base de données sur les locations de vélos à Séoul et le logiciel R pour analyser graphiquement l'influence de différents facteurs temporels et météorologiques sur le nombre de vélos loués.

Dans un deuxième temps, nous utiliserons un modèle de régression linéaire, largement utilisé en économétrie, pour établir un lien entre nos différentes variables et prédire le nombre de vélos loués en fonction des facteurs identifiés. Pour effectuer des prédictions sur les années postérieures à celles mesurées dans notre base de données, nous utiliserons une seconde base de données contenant des données météorologiques à Séoul entre mars 2017 et mars 2023.

Enfin, nous tenterons de prédire les données météorologiques pour la période d'avril 2023 à avril 2024, afin de pouvoir prédire le nombre de locations de vélos futures avec une plus grande précision. En somme, cette méthodologie nous permettra de répondre à notre problématique.

1. https ://fr.wikipedia.org/wiki/Séoul
2. https ://kojects.com/2015/09/18/bike-sharing-system-in-seoul/

2 Etude bibliographique

Avant d'analyser notre base de données, nous avons examiné des études antérieures pour déterminer si des recherches similaires avaient déjà été menées. Nous avons découvert plusieurs articles intéressants qui ont utilisé des méthodes économétriques pour examiner l'impact des conditions météorologiques sur l'utilisation des vélos en libre-service.

Le premier article, "Weather and Public Bikesharing : Impact on Use and Ridership" de John A. R. Adams et al. (2016), étudie l'impact des conditions météorologiques sur l'utilisation des vélos en libre-service à New York. Les résultats indiquent que les températures plus élevées, la baisse de la vitesse du vent et la diminution de la pluie sont associées à une augmentation du nombre de voyages en vélo.

Le deuxième article, "Forecasting Bike Rental Demand with Weather Data" de Akhilesh Mishra et al. (2017), propose un modèle économétrique pour prédire la demande de location de vélos à partir de données météorologiques. Les auteurs ont utilisé un modèle de régression linéaire multiple et les résultats ont montré que les données météorologiques sont un facteur important dans la prédiction de la demande de location de vélos.

Le troisième article est "The influence of weather conditions on the usage of the Barclays Cycle Hire" de Rolf van Lieshout et Jelmer Strijkstra (2015). Dans cet article, les auteurs ont étudié l'impact des conditions météorologiques sur le nombre de locations de vélos par heure à Londres, en utilisant un modèle de régression binomiale négative. Leurs résultats ont montré que la pluie, le vent et la visibilité avaient un effet sur le nombre de locations, et que cet effet était encore plus marqué le week-end.

Enfin, nous avons examiné l'article "Using data mining techniques for bike sharing demand prediction in metropolitan city" de Sathishkumar V E, Jangwoo Park et Yongyun Cho, publié en mars 2020. Les auteurs ont utilisé des données de location de vélos à Séoul, en Corée du Sud, ainsi que des données météorologiques et d'autres variables pour développer un modèle de prédiction de la demande de location de vélos. Ils ont utilisé plusieurs techniques de fouille de données, notamment la régression linéaire multiple, la régression logistique et les arbres de décision, pour identifier les variables qui influencent le plus la demande de location de vélos. Les résultats montrent que les variables météorologiques, telles que la température et la pluie, ont un impact significatif sur la demande de location de vélos.

Ces études soulignent la relation entre les conditions météorologiques et le nombre de vélos loués. Nous utiliserons ces études comme source d'inspiration pour notre analyse de la base de données et pour développer un modèle de prédiction précis. Ce modèle pourrait être utile pour les opérateurs de vélos en libre-service pour planifier les ressources et les services de manière plus efficace et répondre aux besoins des utilisateurs.

3 La base de données "Seoul Bike Sharing Demand"

3.1 Présentation de la base de données

Afin de prédire avec précision la demande de location de vélos à Séoul en fonction de divers facteurs temporels et météorologiques, nous avons recherché une base de données pertinente. Après examen, nous avons sélectionné la base de données "Seoul Bike Sharing Demand [3]", disponible en libre accès sur le site de l'UCI Machine Learning Repository, un site créé en 1987 regroupant des bases de données, des théories des données et des générateurs de données en libre accès. Il est important de noter que les données de cette base proviennent initialement du site [4] du gouvernement coréen.

Cette base de données comprend le nombre de vélos loués via le service "Seoul Bikes" pour chaque heure de la journée durant une période d'un an, allant du 1er décembre 2017 au 30 novembre 2018. Pour chaque heure de la journée, nous disposons également de 14 attributs, que nous allons maintenant examiner en détail. Dans notre étude nous appelerons cette base de données simplement "SeoulBike".

3.2 Présentation des variables

Avant d'utiliser une base de données, il est essentiel de comprendre la signification de chaque attribut.

TABLE 1 – 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

3. https ://archive.ics.uci.edu/ml/datasets/Seoul+Bike+Sharing+Demand
4. http ://data.seoul.go.kr/

3.3 Analyse des variables

Dans cette partie nous allons étudier le lien entre nos différentes variables, et plus précisément nous allons étudier quelles variables ont une influence sur le nombre de locations de vélos. Nous pouvons déjà analyser le nombre moyen de locations journalières au cours de l'année d'observation.

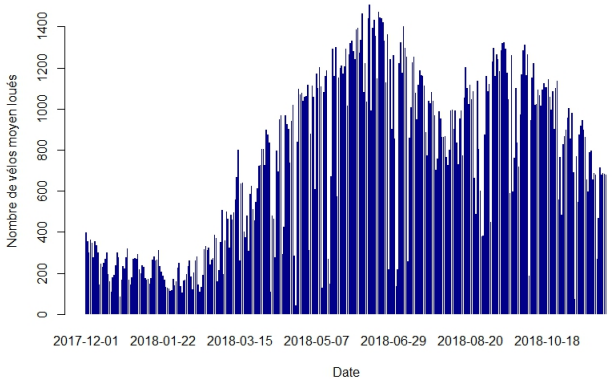


FIGURE 1 – Moyenne du nombre de vélos loués par date

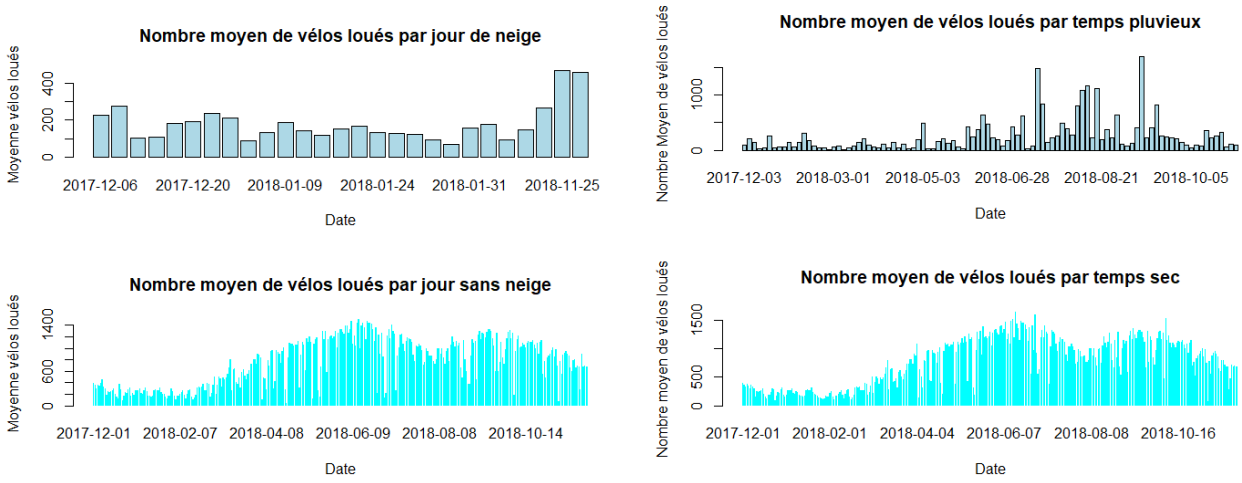
La figure 1, nous montre qu'il y aurait un lien entre les saisons et le nombre de locations de vélos. En effet, on observe une moyenne beaucoup plus faible autour du mois de janvier, puis une augmentation à l'approche de l'été. Nous allons donc dans un premier temps étudier l'impact des saisons, ainsi que les température qui est en lien avec celle-ci.

(a) Locations moyennes en fonction de la température

(b) Locations moyenne par saisons

FIGURE 2 – Influence de la température sur le nombre de locations de vélos

Sur le graphique (a) de la Figure 2, on observe une corrélation entre l'augmentation de la température et la hausse du nombre de vélos loués, mais le nombre de locations diminue après une température de 30 degrés, car cela devient inconfortable pour une activité sportive. Dans le graphique (b), il y a une différence du nombre de locations selon les saisons. Le nombre de locations est faible en hiver, augmente considérablement au printemps, atteint son maximum en été, puis diminue légèrement à l'automne. Cependant, en hiver, il y a aussi généralement de fortes chutes de pluie et de neige, ce qui nécessite une analyse plus approfondie de ces deux variables.



(a) Locations moyennes avec et sans neige.

(b) Locations moyennes avec et sans pluie.

FIGURE 3 – Influence de la pluie et de la neige sur le nombre de locations de vélos

La Figure 3 présente l'impact de la pluie et de la neige sur le nombre de locations de vélos. Le graphique (a) montre l'évolution du nombre moyen de vélos loués par jour avec et sans neige. Toutefois, étant donné le faible nombre d'observations pour lesquelles il y a de la neige, il est difficile d'établir une relation formelle entre la chute de neige et son impact sur le nombre de vélos loués. Par conséquent, il est difficile de conclure si la présence de neige a une influence significative sur le nombre de locations de vélos. En ce qui concerne le graphique (b), il présente l'évolution du nombre moyen de vélos loués avec et sans pluie. Bien que graphiquement il semble y avoir une diminution du nombre de vélos loués en présence de pluie, nous ne pouvons pas affirmer cette corrélation avec certitude.

Nous pourrions également nous interroger sur la variation du nombre moyen de locations de vélos en fonction des jours de la semaine.

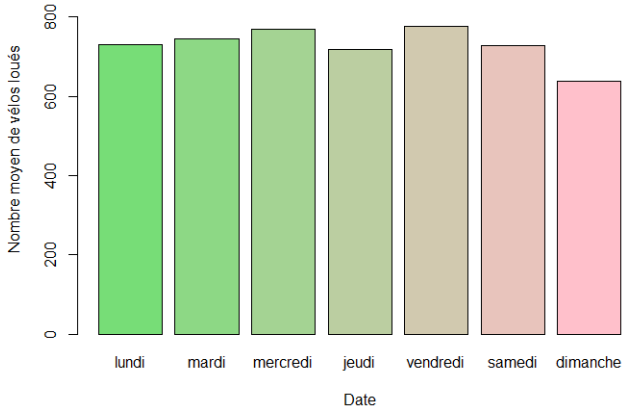
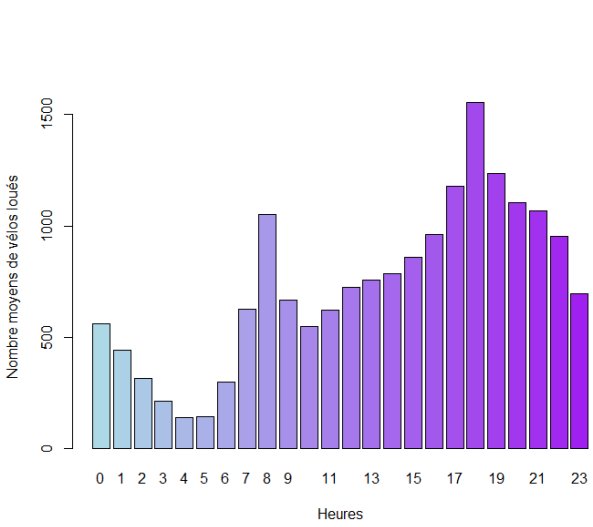
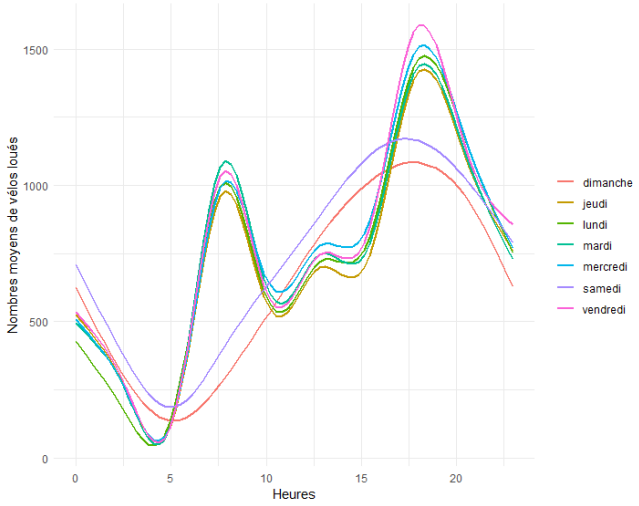


FIGURE 4 – Locations moyennes par jour de la semaine

D'après la figure 4, il semblerait qu'il n'y ait pas de différence significative entre les nombres de locations moyens pour chaque jour de la semaine. Cependant, il est important de considérer l'impact de l'heure de la journée sur le nombre de locations. Cependant, il serait pertinent d'analyser le nombre moyen de location de vélo par heure sur l'ensemble des observation, mais également les heures pour chaque jours de la semaine.



(a) Locations moyenne par heure de la journée



(b) Locations moyenne par heure de la journée selon les jours

FIGURE 5 – Influence de l'heure de la journée sur le nombre de locations de vélos

La Figure 5 met en évidence l'impact de l'heure de la journée sur le nombre de locations de vélos. Le graphique (a), qui représente le nombre moyen de locations par heure de la journée, révèle deux pics de locations : le premier vers 8 heures du matin et le second vers 18 heures. Ces pics correspondent probablement aux heures de pointe, lorsque les gens vont et reviennent du travail ou de l'école.

Cependant, comme on peut le voir sur le graphique (b), cette tendance est confirmée pour les jours de la semaine, mais moins marquée pour les week-ends. On observe que le weekend le nombre de locations par heure est différents des autres jours de la semaine. Bien que l'impact de l'heure de la journée sur le nombre de locations de vélos soit clairement établi, il n'est pas possible de conclure de manière définitive si les jours de la semaine ont une influence significative sur le nombre de locations.

Pour étudier les effets significatifs des variables sur le nombre de location de vélo, nous allons donc utiliser deux tests, le test Chi2 et le test Anova, afin d'étudier la significativité de chacune des variables sur le nombre de location de vélo.

Le test de Chi2 nous permettra de déterminer s'il existe une relation statistiquement significative entre deux variables catégorielles telles que la météo ou les jours de la semaine, et le nombre de locations de vélos.

Quant au test Anova, il nous permettra de déterminer s'il existe une différence statistiquement significative entre les moyennes de plusieurs groupes, comme par exemple les différents mois de l'année.

En utilisant ces deux tests, nous pourrons donc confirmer ou infirmer les hypothèses émises précédemment à partir des observations graphiques, et ainsi obtenir une compréhension plus approfondie des facteurs influençant le nombre de locations de vélos.

Les résultats du **test de Chi2 pour les variables catégorielles** sont regroupées dans le tableau suivant :

TABLE 2 – Résumé du test Statistique de Chi2

Variable	Test_Statistic	Degrees_of_Freedom	P-Value	Result
Day	13106.794	12984	2.225087e-01	Not significant
Hour	53757.366	49772	3.925687e-35	Significant
Seasons	9028.556	6492	2.525518e-88	Significant
Holiday	1871.532	2164	9.999984e-01	Not significant

Dans le tableau 2, la colonne "Result", nous permet d'observer quelle variable est significative ou non. Lorsqu'une variable a une p-value inférieur à 0.05, alors celle-ci est significative selon leur statistique de test et degré de liberté. En conclusion, les résultats indiquent que l'heure et les saisons ont une influence significative sur le nombre de vélos loués, tandis que les jours de la semaine et les périodes de vacances n'ont pas d'influence significative.

Concernant **les variables continues, les résultats du test d'Anova** sont dans le tableau suivant :

TABLE 3 – 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	4719.029	< 2e-16 ***
Humidity...	1	3.139e+08	3.139e+08	1339.439	< 2e-16 ***
Wind.speed..m.s.	1	8.361e+06	8.361e+06	35.676	2.42e-09 ***
Rainfall.mm.	1	3.104e+07	3.104e+07	132.459	< 2e-16 ***
Snowfall..cm.	1	5.274e+05	5.274e+05	2.250	0.134
Solar.Radiation..MJ.m2.	1	5.081e+07	5.081e+07	216.809	< 2e-16 ***
Visibility..10m.	1	1.202e+04	1.202e+04	0.051	0.821
Dew.point.temperature.C.	1	4.880e+03	4.880e+03	0.021	0.885
Residuals	8456	1.982e+09	2.344e+05		

Le tableau 3 fournit une évaluation de l'impact des différentes variables météorologiques qui sont continues, sur le nombre de vélos loués. Les variables Température, Humidité, Vitesse du vent, Précipitations et Rayonnement solaire ont toutes des p-value significativement inférieures à 0,001, ce qui indique que ces variables ont une influence significative sur le nombre de vélos loués.

La variable neige a une p-value de 0,134, ce qui suggère que cette variable n'a pas d'effet significatif sur la location de vélos. Les deux dernières variables, Visibilité et Point de rosée, ont également des p-values plus élevées (0,821 et 0,885 respectivement), ce qui indique qu'elles n'ont pas d'influence significative sur le nombre de vélos loués.

Les résultats de l'analyse suggèrent que la température, l'humidité, la vitesse du vent, la pluie et les radiations solaires ont un effet important sur la location de vélos, tandis que les autres variables météorologiques n'ont pas d'effet significatif.

Maintenant que nous avons identifié avec précision les variables ayant un impact significatif sur le nombre de locations de vélos, nous sommes en mesure de définir et d'utiliser notre modèle de régression linéaire pour effectuer des prédictions précises.

4 Modèle de régression linéaire

Dans cette partie, nous allons utiliser un modèle de régression linéaire pour analyser l'influence de variables choisies sur le nombre de locations de vélos. Nous allons dans un premier temps définir ce qu'est un modèle de régression linéaire.

4.1 Définition

En économétrie, la régression linéaire est une technique statistique qui permet de modéliser une relation entre une ou plusieurs variables explicatives (dites indépendantes) et une variable à expliquer (dite dépendante). Ce modèle suppose qu'il y a une relation linéaire entre les variables explicatives et la variable à expliquer. Autrement dit, cette relation peut être représentée par une droite de régression. Le modèle de régression linéaire s'écrit sous la forme :

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

où :

- y : est la variable à expliquer (dépendante).
- $x_1, x_2, \ldots, x_n$: ce sont les variables explicatives (indépendantes).
- β_0 : est le coefficient de l'intercept, c'est la valeur d'y quand toutes les variables explicatives sont nulles, soit la constante.
- $\beta_1, \beta_2, \ldots, \beta_n$: ce sont les coefficients des x_n respectivement, ils indiquent l'effet de chaque variable indépendante sur la variable dépendante.

4.2 Notre modèle de régression linéaire

Pour notre modèle de régression linéaire, nous utilisons R, nous n'avons donc aucun calcul à effectuer, le logiciel nous donnera directement tous les éléments de réponses. Toutefois, nous devons dans un premier temps sélectionner notre variable dépendante et nos variables indépendantes.

4.2.1 Choix des variables

Le choix de notre variable dépendante est évident, nous prendrons la variable correspondant au nombre de locations de vélos.

Pour sélectionner nos variables indépendantes, nous devons nous baser sur celles dont nous avons prouvé la significativité : "Hours", "Seasons", "Temperature", "Humidity", "Windpseed", "Rainfall" et "SolarRadiation". Cependant, avant d'intégrer toutes ces variables à notre modèle, nous devons étudier leur possible corrélation. En effet, il est important de ne pas inclure des variables corrélées entre elles dans un modèle de régression linéaire, car cela peut fausser les résultats.

Nous avons donc la matrice de corrélation suivante :

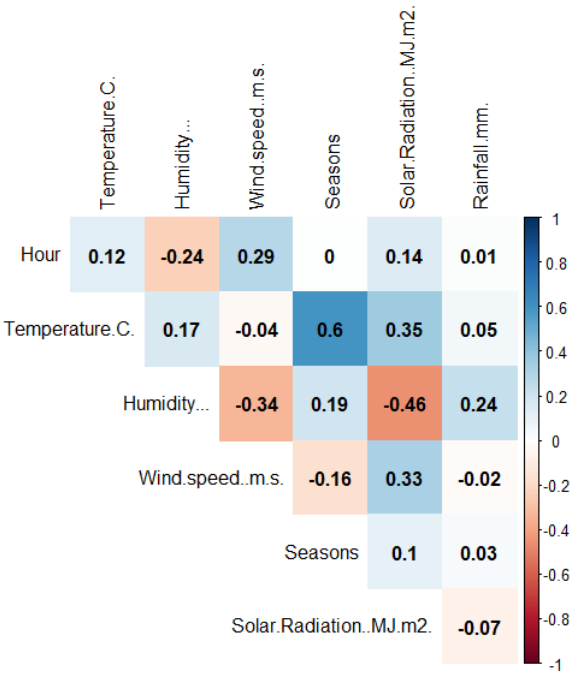


FIGURE 6 – Matrice de corrélation des variables significatives

Selon le critère communément admis, deux variables sont considérées comme corrélées lorsque leur coefficient de corrélation est supérieur à 0,5. Dans notre matrice, nous observons une corrélation entre la variable "Temperature" et la variable "Seasons". Par conséquent, afin d'éviter l'inclusion de variables corrélées dans notre modèle de régression linéaire, nous avons décidé de ne pas prendre en compte la variable "Seasons" qui indique les saisons.

4.2.2 Résumé du modèle

La formule de notre régression linéaire est donc :

$$Rented.Bike.Count = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon$$

où X_1 représente l'heure, X_2 la température en degrés Celsius, X_3 l'humidité, X_4 la quantité de pluie en millimètres et X_5 le rayonnement solaire en $\mathrm{MJ/m^2}$.

Nous constatons que dans notre modèle, il ne reste que des variables continues, nous allons donc estimés nos coefficients β_0, β_1, β_2, β_3, β_4 et β_5 à l'aide de la méthode des moindres carrés ordinaire (MCO).

TABLE 4 – 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

Les résultats de la régression montrent que toutes nos variables indépendantes sont significativement associées au nombre de vélos loués par heure, car les p-values sont toutes inférieurs à 0.05.

Le coefficient de détermination multiple (R^2) est de 0.5115, ce qui indique que les cinq variables expliquent environ 51% de la variance dans le nombre de vélos loués par heure.

Les coefficients de la régression donnent une idée de l'impact de chaque variable indépendante sur le nombre de vélos loués par heure. Par exemple, une augmentation de 1 degré Celsius de la température est associée à une augmentation de $32,6085$ vélos loués par heure, toutes choses étant égales par ailleurs.

Enfin, la valeur F est significative à un niveau de confiance très élevé, ce qui indique que l'ensemble des variables expliquent de manière significative la variation du nombre de vélos loués par heure.

En conclusion, ces résultats montrent qu'il existe une relation significative entre les variables météorologiques ainsi que l'heure et le nombre de vélos loués par heure.

Maintenant que nous avons identifié les variables qui ont une influence sur le nombre de vélos loués, nous sommes en mesure de prédire le nombre de locations à venir.

5 Prédictions

Pour réaliser nos prévisions, nous utiliserons deux méthodes différentes : la première s'appuie sur un arbre de décision, tandis que la seconde utilise XGBoost. Pour ce faire, nous nous appuierons sur une deuxième base de données.

5.1 Introduction d'une seconde base de données

Afin de prédire le nombre de vélos qui seront loués pour des dates ultérieures à celles de notre base de données initiale, nous avons besoin d'une seconde base de données. Nous avons donc collecté différentes données météorologiques pour la ville de Séoul entre le 30 mars 2017 et le 29 mars 2023 sur le site du gouvernement coréen [5] et nous les avons regroupés dans une base de données "MétéoSeoul". Cette base de données contient les variables suivantes : **Date, Hour, Temperature, Rainfall, Windspeed, Humidity, SolarRadiation, Snowfall et Visibility**. Comme nous l'avons démontré précédemment, nous ne considérerons que les variables significatives que nous avons utilisées dans notre modèle de régression linéaire.

5.2 Prédiction à l'aide d'un arbre de décision

L'arbre de décision est un algorithme de machine learning qui permet de classer des données en fonction de leurs caractéristiques en utilisant une structure en arbre. L'objectif est de diviser les données en sous-groupes homogènes en se basant sur les variables les plus discriminantes. Chaque branche de l'arbre représente une règle qui permet de séparer les données en fonction d'une variable. Les feuilles de l'arbre représentent les classes ou les valeurs prédites. Pour construire l'arbre de décision, l'algorithme cherche la variable qui permet de séparer le mieux les données et qui maximise la pureté de chaque sous-groupe. La pureté est mesurée à l'aide d'une fonction d'impureté qui prend en compte la répartition des classes dans chaque sous-groupe. Dans le cas de notre base de données initiale, l'arbre de décisions considèrent les variables que nous avons démontrés comme étant influente sur le nombre de locations.

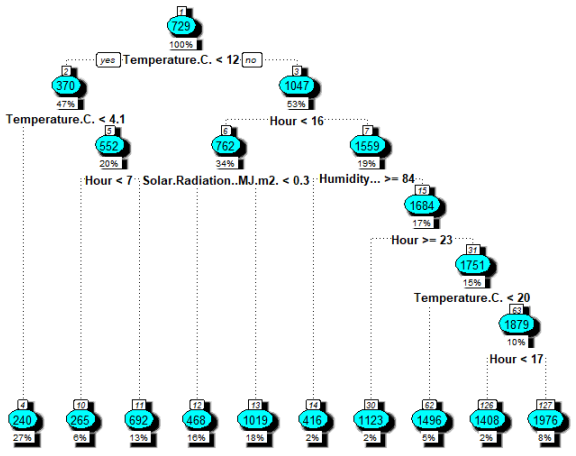
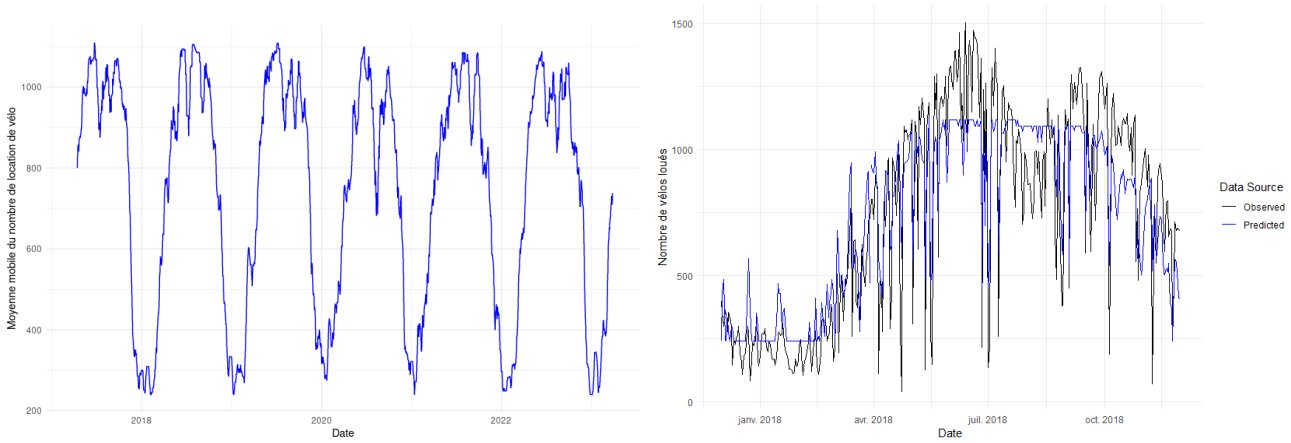


FIGURE 7 – Arbre de décision pour la prédiction

5. http ://data.seoul.go.kr/

D'après l'arbre de décision, on peut voir que la variable ayant le plus d'influence sur le nombre de locations selon l'algorithme est la variable température. Cet arbre va nous permettre de prédire le nombre de locations de vélos pour les observations de notre nouvelle base de données "MétéoSéoul". Nous allons maintenant nous intéresser à l'analyse graphique des prédictions obtenues.



(a) Prédictions pour la base de données "MétéoSéoul". (b) Comparaison avec les données de SeoulBike.

FIGURE 8 – Prédiction du nombre de vélos loués à l'aide d'une arbre de décision.

La Figure 8 présente les prédictions du nombre de locations de vélos réalisées à l'aide de notre arbre de décision. Nous observons une saisonnalité des prédictions dans le graphique (a) qui semble correspondre à ce que nous avions identifié dans nos analyses précédentes.

Cependant, en examinant le graphique (b), qui nous montre le nombre moyen de location réel et prédites, nous remarquons que les mesures des prédictions semblent atteindre un plafond au niveau des pics de la courbe des données réellement observées. Il semble que notre prédiction ne parvienne pas à reproduire précisément les pics de locations au delà de 1100 locations. Ce modèle semble avoir une précision limitée, il reproduit la tendance générale des locations mais il manque de précision.

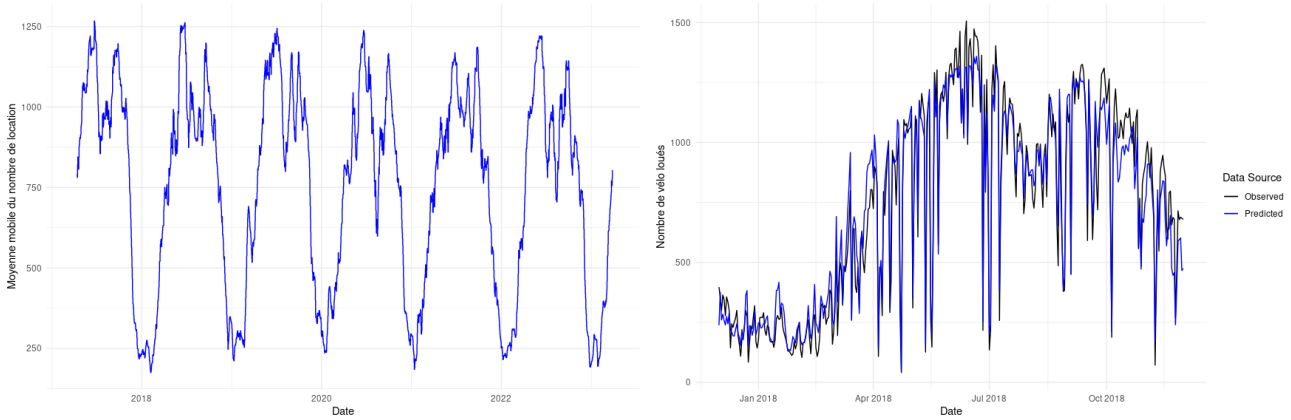
Afin d'obtenir des prédictions de meilleure qualité, nous allons utiliser une seconde méthode de prédiction.

5.3 Prédiction à l'aide du modèle XGBoost

Dans cette partie, nous utiliserons un modèle de prédiction plus complexe : le modèle XGBoost. Il s'agit d'un algorithme de machine learning de type boosting qui peut être utilisé pour la régression et la classification de données. Le modèle XGBoost est basé sur un ensemble de modèles d'arbres de décision, où chaque arbre est entraîné sur les erreurs résiduelles pour améliorer la précision des prédictions.

Notre modèle XGBoost nous permet de faire des prédictions sur le nombre de locations de vélos. Il a été créé en utilisant la fonction "xgboost" et les variables explicatives Hour, Temperature.C., Humidity..., Rainfall.mm. et Solar.Radiation..MJ.m2. de notre base de données "SeoulBike". Après entraînement de l'algorithme, celui-ci est capable de nous fournir des prédictions pour nos observations dans la base de données "MétéoSéoul".

Nous pouvons maintenant analyser graphiquement les résultats de cette prédiction.



(a) Prédictions pour la base de données "MétéoSéoul". (b) Comparaison avec les données de SeoulBike.

FIGURE 9 – Prédiction du nombre de vélos loués à l'aide du modèle XGBoost.

La figure 9 présente les prédictions du nombre de vélos loués à l'aide du modèle XGBoost. Sur le graphique (a), la tendance semble similaire à ce que nous avons obtenu précédemment. Le graphique (b) montre que les prédictions sont très proches des données réellement observées dans notre base de données "SeoulBike".

Ainsi, le modèle XGBoost fournit des prédictions précises du nombre de vélos loués.

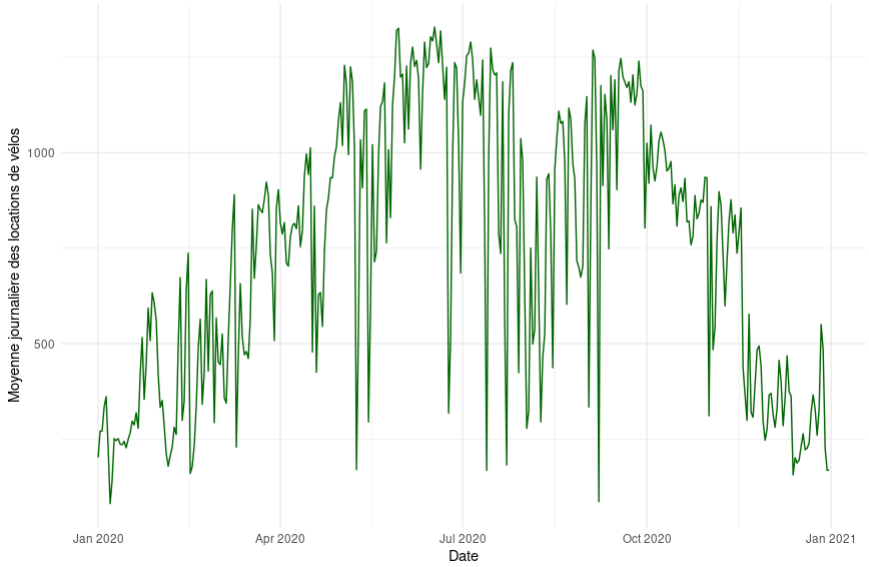


FIGURE 10 – Prédictions pour l'année 2020

Le modèle XGBoost nous permet de construire un scénario contrefactuel concernant les locations de vélos au cours de l'année 2020. En effet, avec l'apparition de la pandémie de Covid-19 et les restrictions sanitaires dans le monde entier, on peut raisonnablement supposer que l'utilisation des vélos de location a fortement diminué. La Figure 10 présente donc les résultats que nous aurions obtenus en l'absence de pandémie.

5.4 Prédictions pour les années à venir

Dans cette partie finale, nous cherchons à obtenir des prédictions pour les données futures, comprises entre le 30 mars 2023 et le 29 avril 2024. Pour ce faire, nous utilisons les données de notre base "MétéoSéoul". Pour chaque variable que nous souhaitons prédire, nous calculons la moyenne des données mesurées pour cette variable entre 2017 et 2023.

On peut résumer ceci dans la formule mathématique suivante : $X_t = \frac{1}{7}\sum_{i=t-7}^{t-1} X_i$.
Où X_t est la variable que l'on souhaite prédire, et X_i les mesures de cette variable pour les années précédentes.

Nous obtenons une nouvelle base de données contenant les prédictions météorologiques pour la période entre le 30 mars 2023 et le 29 avril 2024. Nous allons appliquer le modèle XGBoost à cette base de données, car nous avons constaté que c'est le modèle de prédiction le plus précis.

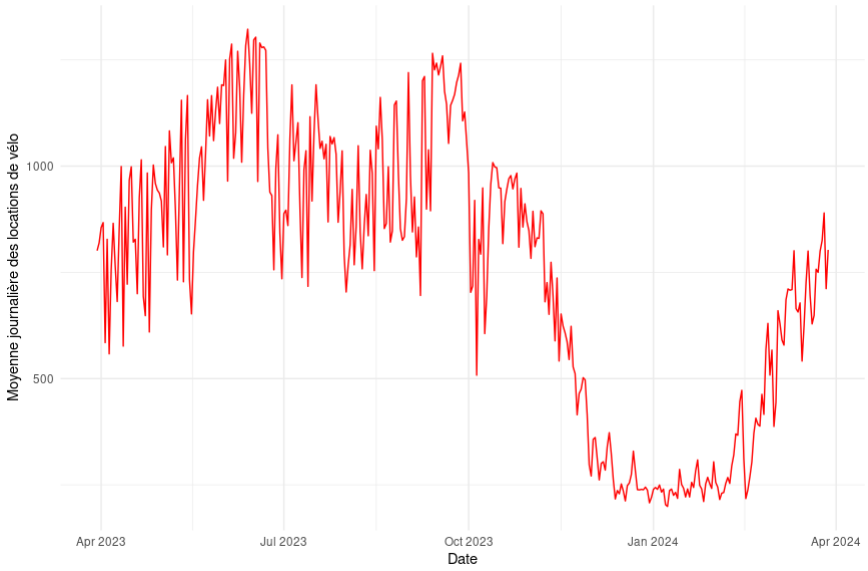


FIGURE 11 – Prédictions du nombre de locations de vélo entre mars 2023 et avril 2024

Le modèle XGBoost nous permet de prédire le nombre de vélos loués entre le 30 mars 2023 et le 29 avril 2024. Comme nous pouvons le voir dans la figure 11, la tendance de la courbe montre qu'il y a une chute des locations à prévoir durant l'hiver, et une augmentation des locations durant l'été. Ces tendances sont similaires à ce que nous avons pu observer depuis le début de notre étude.

6 Discussion

Au cours de notre étude, nous avons identifié plusieurs variables pouvant avoir un impact sur le nombre de vélos loués. La température est la variable qui semble être la plus significative, mais nous avons également identifié d'autres variables influentes, telles que la visibilité, la pluie, les radiations solaires, l'humidité et les saisons. Cependant, il est important de souligner qu'il peut exister d'autres variables significatives qui peuvent influencer le nombre de vélos loués, et il serait donc intéressant de les étudier. En déterminant ces variables significatives, nous avons pu construire notre modèle de régression linéaire. Nous avons dû retirer les saisons de notre modèle de régression linéaire car cette variable était fortement corrélée à la température. Grâce à un modèle de régression linéaire, nous avons pu représenter l'effet de nos variables significatives sur le nombre de vélos loués. Nous avons ensuite utilisé un arbre de prédiction pour prédire les données, mais nous avons constaté que ce modèle était limité dans ses capacités de prédiction. Nous avons donc opté pour un modèle XGBoost, qui est un modèle de prédiction plus complexe, et avons obtenu des prédictions très précises. Enfin, nous avons déterminé les variables météorologiques pour l'année à venir et avons pu prédire le nombre de locations pour cette période grâce à notre modèle XGBoost.

Cependant, il convient d'être prudent avec nos résultats, car la prédiction ne prend en compte que les données de notre base de données initiale, "SeoulBike", et ne tient pas compte de tous les facteurs extérieurs pouvant influencer le nombre de vélos loués. Par exemple, l'utilisation des vélos en libre-service est en constante croissance ces dernières années dans les villes, et nos modèles ne tiennent pas compte de ces évolutions. De même, nos modèles ne prennent pas en compte les facteurs extérieurs tels que les guerres, les événements politiques ou économiques ou encore une pandémie. C'est pourquoi nous avons pu construire le contrefactuel pour l'année 2020, mais nous n'avons pas pu prédire les données dans le cas d'une pandémie. Quant à la prédiction de nos données météorologiques pour l'année à venir, celle-ci nous permet d'avoir une idée de la météo pour l'année à venir, mais il est possible que nos prévisions ne soient pas précises en raison de la nature imprévisible de la météo et du réchauffement climatique. Mais également par notre méthode de prédiction reposant uniquement sur un calcul de moyenne.

7 Conclusion

Dans la continuité de notre travail, il serait intéressant de comparer les données que nous avons prédites entre 2018 et 2023 avec les données qui ont réellement été mesurées. Cette comparaison nous permettrait d'ajuster la précision de notre modèle et de vérifier sa fiabilité. Nous pourrions ensuite fournir ces données à la ville de Séoul pour leur exploitation. En effet, si la ville de Séoul est capable de prédire la demande en vélos pour chaque heure de la journée en fonction des différentes conditions temporelles et météorologiques, cela permettrait de répondre efficacement à la demande, et certains utilisateurs pourraient délaisser leur voiture au profit des trajets en vélo. Cette approche est pertinente d'un point de vue économique et écologique.

Il serait également pertinent de mesurer d'autres données, telles que la durée d'utilisation des vélos, la distance parcourue et la fréquence d'utilisation par un même utilisateur. Ces informations permettraient d'obtenir un modèle de prédiction plus précis pour anticiper et répondre à une demande croissante au fil du temps, selon les endroits précis de la ville.

En somme, fournir à la ville de Séoul un modèle de prédiction précis permettrait de maximiser l'utilisation des vélos en libre-service et de promouvoir les déplacements écologiques en ville.