1- Introduction :

Bike sharing systems sont une nouvelle génération de location de vélos traditionnels où le processus d'obtention de l'adhésion, de location et de restitution des vélos est automatisé grâce à un réseau répartis dans toute la ville. Grâce à ces systèmes, les gens peuvent louer un vélo à un endroit et le rendre à un autre selon leurs besoins. C'est pourquoi le bike sharing system est une idée brillante qui offre aux gens une autre option de transport pour se déplacer sans se soucier d'être bloqués dans la circulation et de profiter de la vue sur la ville ou même de faire de l'exercice en même temps. Le processus de location de vélos en libre-service est fortement corrélé aux conditions environnementales et saisonnières. Par exemple, les conditions météorologiques, les précipitations, le jour de la semaine, la saison, l'heure de la journée, etc. peuvent affecter les comportements de location. Actuellement, il existe plus de 500 programmes de partage qui regroupent plus de 500 000 vélos dans le monde entier. Aujourd'hui, ces systèmes suscitent un grand intérêt en raison de leur rôle important dans les questions de circulation, d'environnement et de santé.

L'objectif de mon analyse est de découvrir le facteur déterminant qui détermine la demande de location de vélos en libre-service, de construire des modèles statistiques et d'essayer ensuite de faire des prévisions sur les locations en se basant sur les informations et les modèles dont je dispose.

2- Description des données :

Les données que je vais examiner sont téléchargées et extraites de Kaggle. Ces données de location de vélos en libre-service de Capital Bikeshare ne contiennent que des entrées échantillonnées à Washington D.C. sur une période de deux ans allant du 1er janvier 2011 au 19 décembre 2012. L'ensemble de données est également complété par les statistiques météorologiques pour la date et l'heure correspondantes.

Les variables sont :

- **datetime :** contenant la date au format horodaté.
- **season :** contenant les entiers 1 à 4 représentant "Winter", "Spring", "Summer", "Fall".
- **holiday :** contenant des expressions booléennes en 1 et 0 représentant si le jour de l'observation est un jour férié ou non.
- **workingday :** contenant des expressions booléennes en 1 et 0 représentant si le jour de l'observation est un jour ouvrable ou non.
- **weather :** contenant des entiers de 1 à 4 représentant quatre listes différentes de conditions météorologiques :
  1. Clear or cloudy.
  2. Mists.
  3. Light rain or snow.
  4. Heavy rain, snow or even worse weather.
- **temp :** contenant les valeurs de température à l'heure donnée.
- **atemp :** contenant les valeurs de la température ressentie à un moment donné.
- **humidity :** contient les valeurs du niveau d'humidité relative à un moment donné, sur une échelle de 1 à 100..
- **windspeed :** contient les valeurs de la vitesse du vent, en mph (miles per hour).

- **casual :** contenant le nombre de locations d'utilisateurs non enregistrés, dans toutes les stations.
- **registered :** contient le nombre de locations d'utilisateurs enregistrés, pour toutes les stations.
- **count :** notre variable d'interêt qui désigne le nombre total de locations à l'heure donnée, pour toutes les stations.

Questions / Interrogations :

1. Existe-t-il une relation entre le nombre de vélos loués et saison (hiver, automne, été ...) quelques soient le jour de la semaine et le type du jour : journée de travail ou de vacance ? Notre objectif est de déterminer s'il existe une relation entre nombre de vélos loués, season, workingday et holiday, nous aimerions connaître la force de ces relations. En d'autres termes, dans un jour de travail en hiver combien de vélos ont été loués, etc ?
2. Quelle est la relation entre le nombre de vélos loués et l'humidité ainsi que la vitesse du vent ? En supposant qu'il existe une relation entre ces facteurs, nous aimerions savoir la force de cette relation.
3. La relation est-elle linéaire ?
4. Avec quelle précision pouvons nous prédire le nombre futures des vélos loués ? Quelle est notre prévision de vélos loués est la précision de notre prévision ?

Data cleaning :

Un nettoyage préliminaire des données est effectué, en convertissant la variable de **datetime** en mois, jour de la semaine et heure du jour. J'ai changé le type des variables **holiday**, **workingday**, **weather** en facteurs. Je ne garde que la variable **temp** et j'ai supprimé la variable **atemp** car elle est presque répétitive et n'est pas une statistique relativement précise à acquérir. J'ai également supprimé les variables **casual** et **registered** de l'ensemble de données car elles s'additionnent pour former la variable **count** et mon analyse ultérieure ne les utilisera pas.

En examinant les données, il semble y avoir des valeurs sous la forme de 0,0000s dans la variable **windspeed**, j'ai décidé de simplement supprimer les observations avec ces valeurs comme valeurs manquantes parce qu'elles se produisent de manière aléatoire pendant l'heure 0 à 6. La raison pour laquelle j'ai supprimé ces valeurs manquantes au lieu de les remplacer par d'autres valeurs d'équilibre (comme la moyenne) parce que je m'attends à ce qu'il s'agisse de valeurs relativement aléatoires et que leur remplacement par des valeurs fixes entraînera des inexactitudes dans mon analyse par la suite.

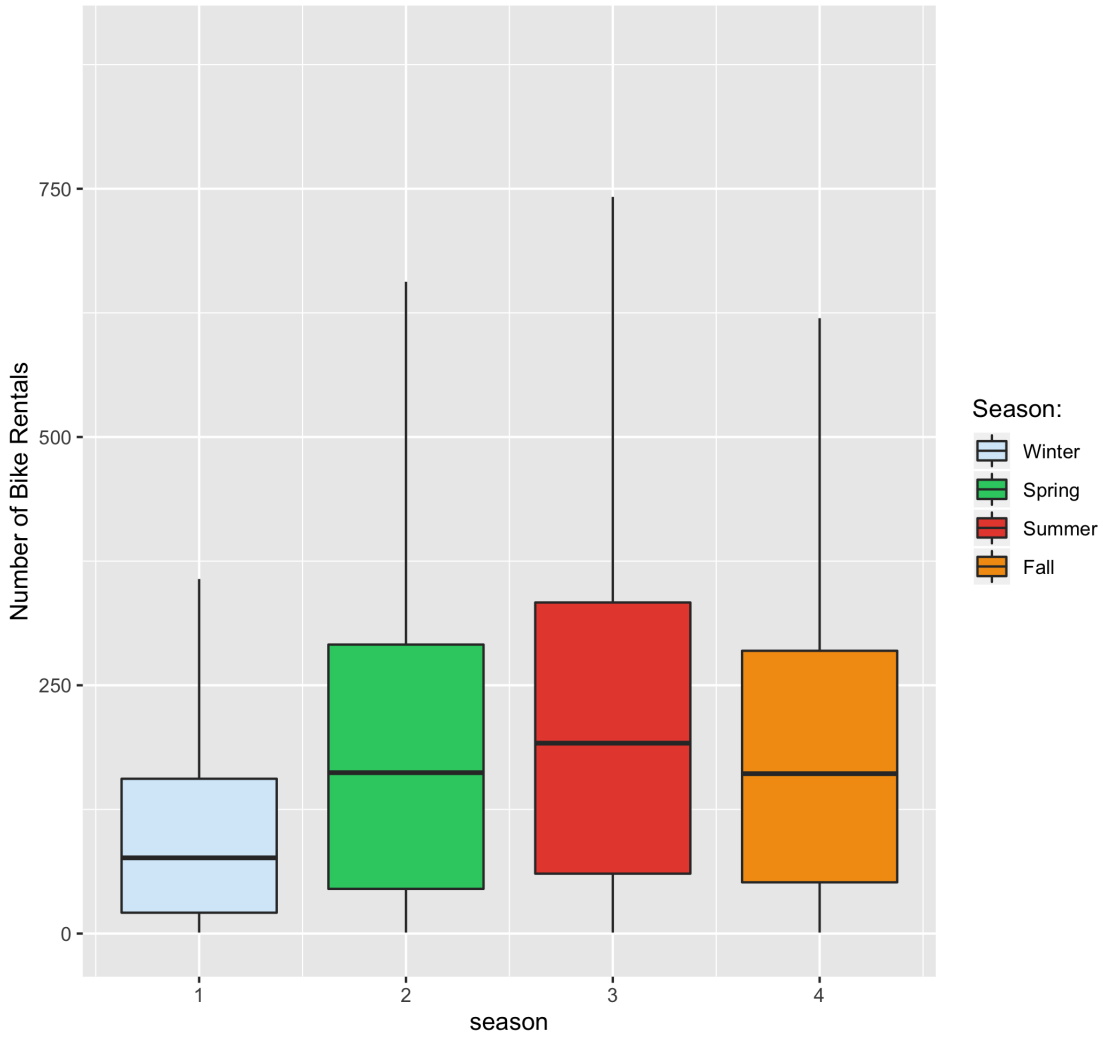
```
   season month weekday hour isweekday isholiday weathertype temperature humidity windspeed count
6       1     1       6    5         0         0           2        9.84       75    6.0032     1
11      1     1       6   10         0         0           1       15.58       76   16.9979    36
12      1     1       6   11         0         0           1       14.76       81   19.0012    56
13      1     1       6   12         0         0           1       17.22       77   19.0012    84
14      1     1       6   13         0         0           2       18.86       72   19.9995    94
>
```

Exploratory Data Analysis :

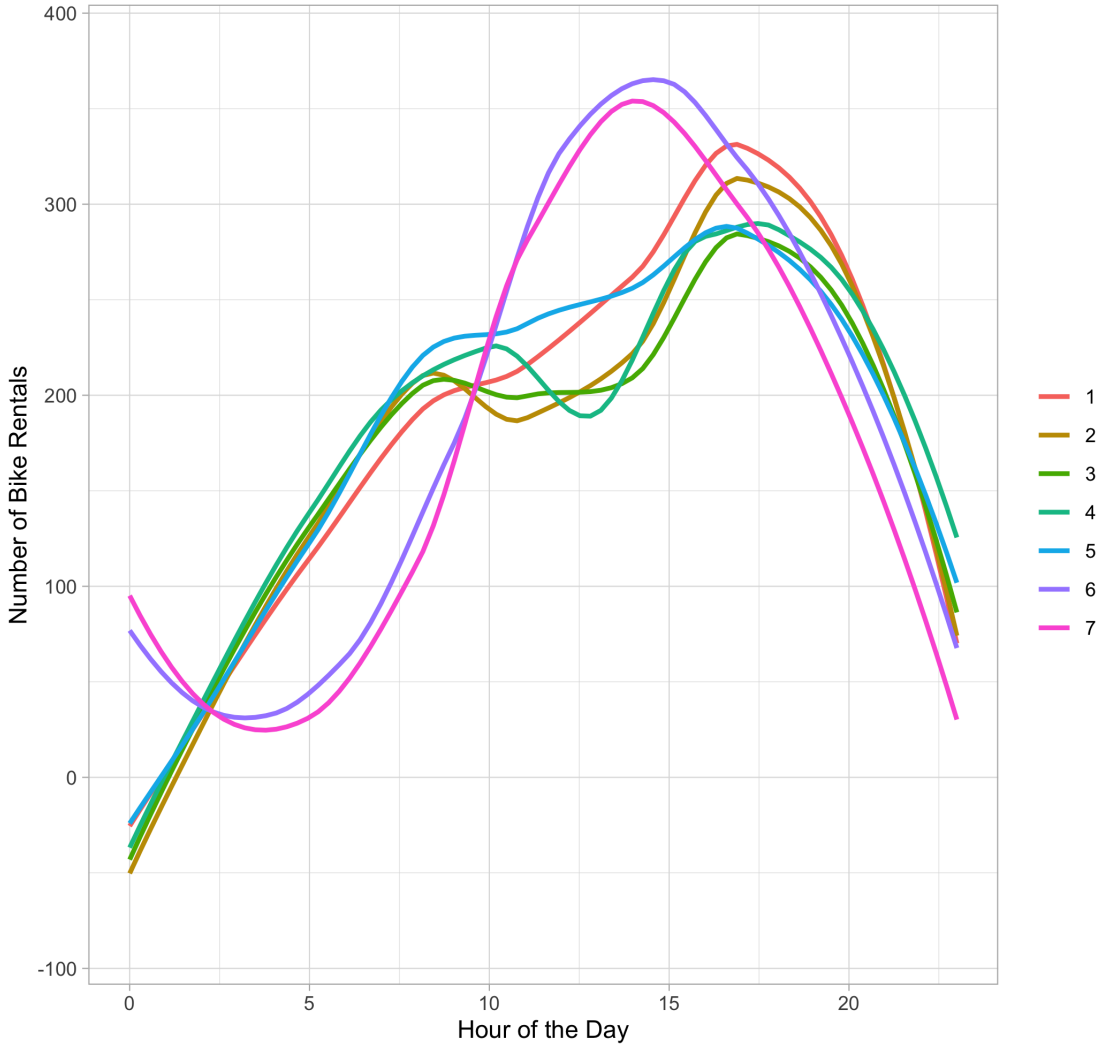
Examinons d'abord les données. J'ai construis un dataframe qui résume le nombre de vélos loués en fonction de la saison, du mois, du jour de la semaine, de l'heure, du jour férié et du type de temps, puis je calcule la moyenne de la température, de l'humidité, de la vitesse du vent et du nombre de vélos loués. L'objectif c'est de trouver une relation générale entre les variables.

Visualisation :

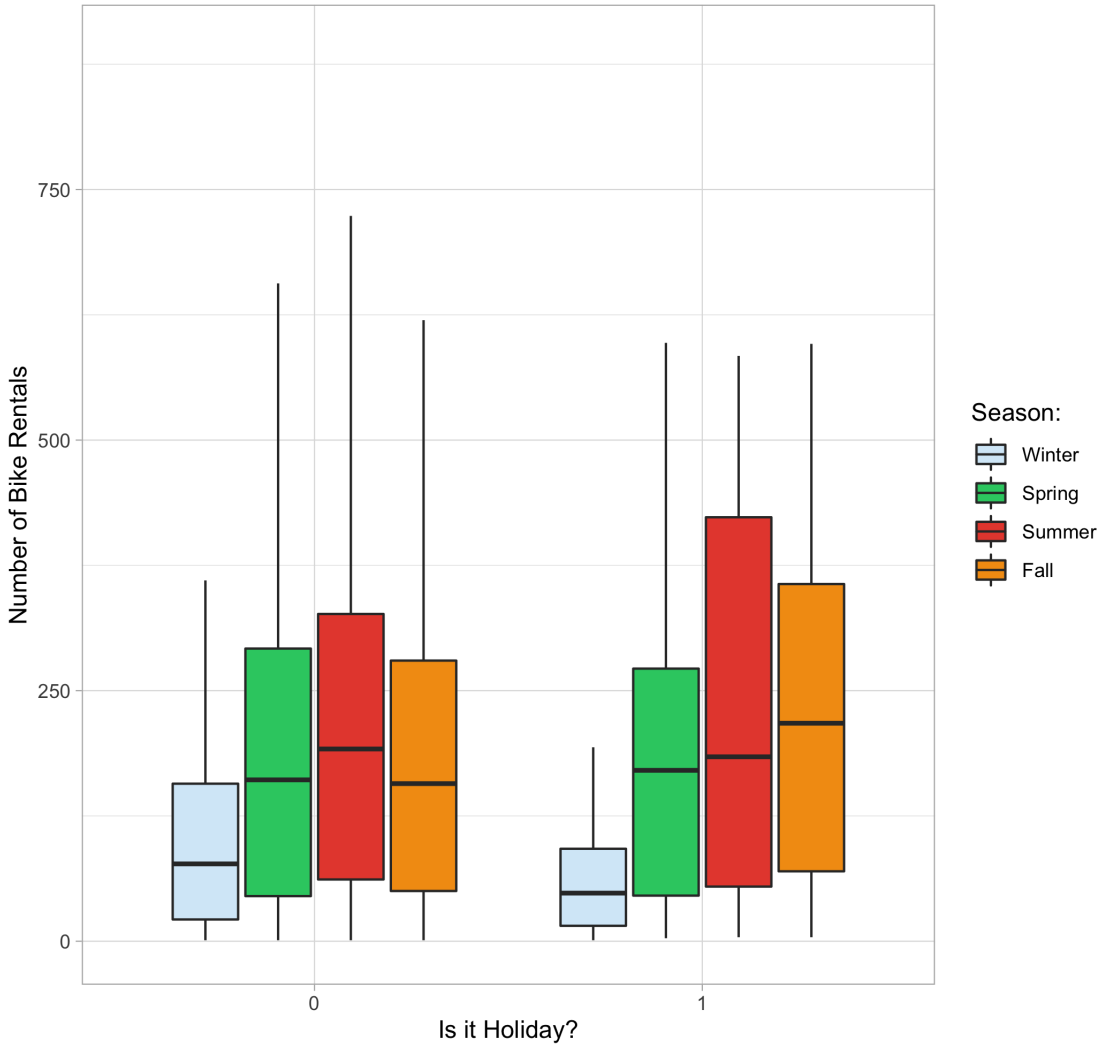
Le boxplot des différentes saisons contre le nombre de locations de vélos révèle qu'il existe une tendance saisonnière dans le nombre de locations. Le nombre de locations est généralement faible en hiver et il atteint son maximum en été. La saison peut être l'un des facteurs déterminants qui affectent le nombre de vélos loués.



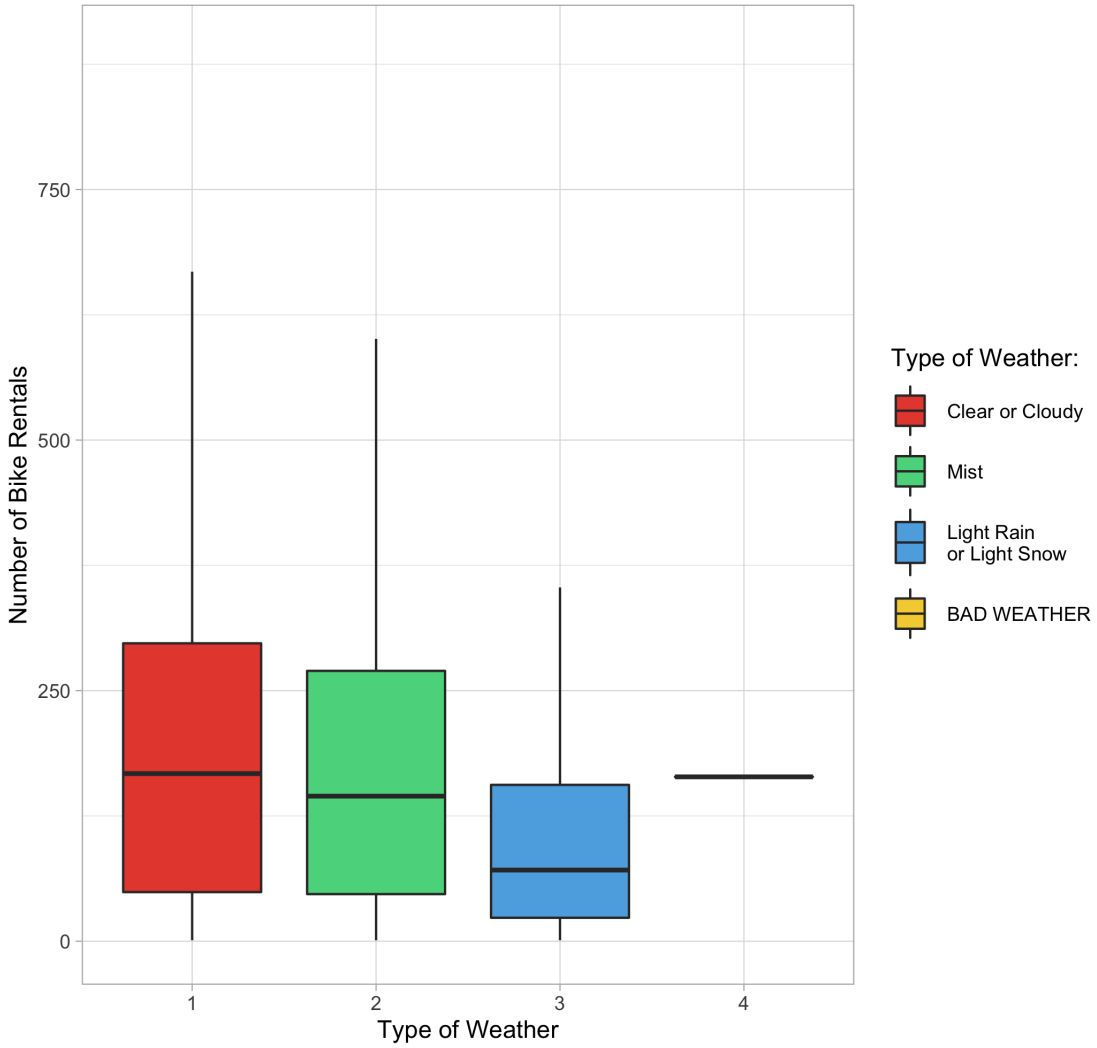
Le graphique montre la différence de demande de location pour la semaine et le week-end à des heures différentes. Le comptage des locations reste actif plus tard à minuit pendant le week-end que pendant la semaine. On peut également constater que le nombre de vélos loués diminue vers 12(p.m) heures en semaine, alors que le week-end, il indique le pic de la demande du jour à la même heure. Le pic de demande en semaine se situe entre 4 et 5 (p.m), peut-être parce que les gens ont fini de travailler et ont besoin de transport pour rentrer chez eux.



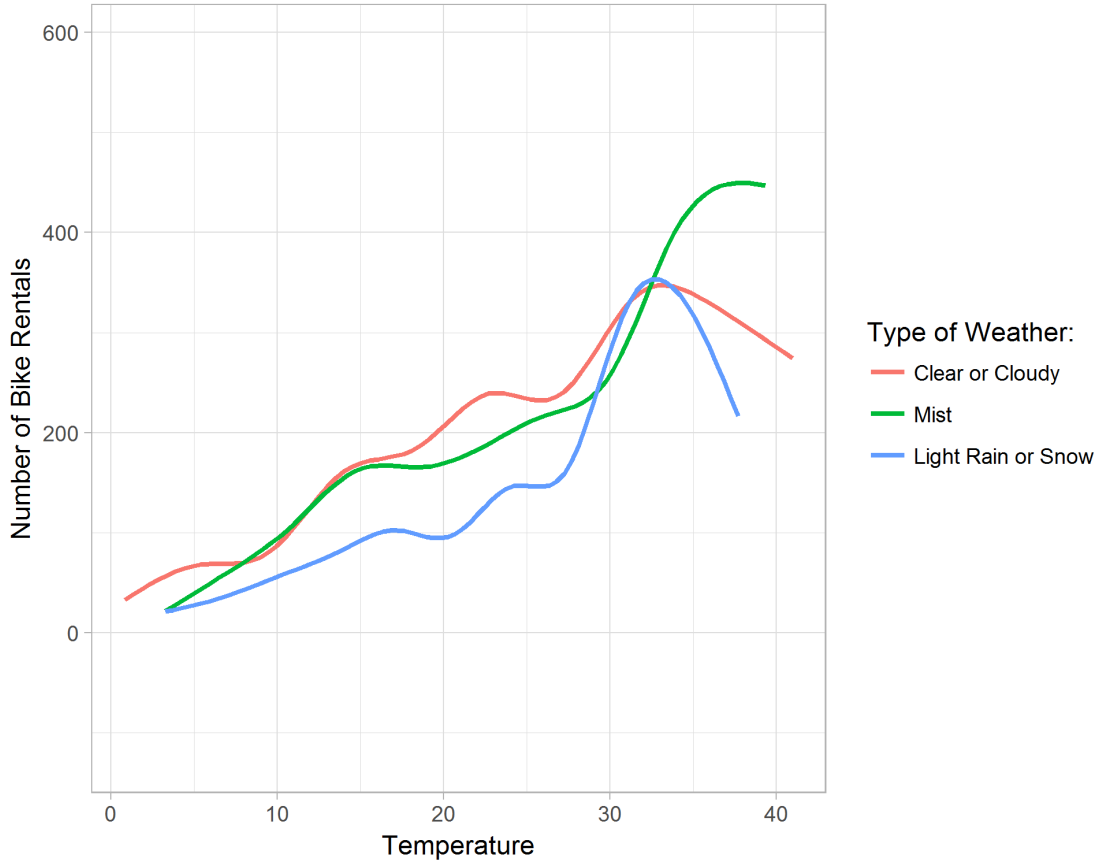
La comparaison des effets des vacances montre que les montants moyens des locations sont à peu près les mêmes, qu'il s'agisse de vacances ou non. En raison de la taille réduite de l'échantillon pour les vacances, l'éventail des locations est généralement plus restreint que celui des locations non-vacances. Nous pouvons également constater une saisonnalité similaire à celle du boxplot de comptage des locations en fonction de la saison ; l'hiver montre le comptage des locations le plus bas et l'été le plus élevé.



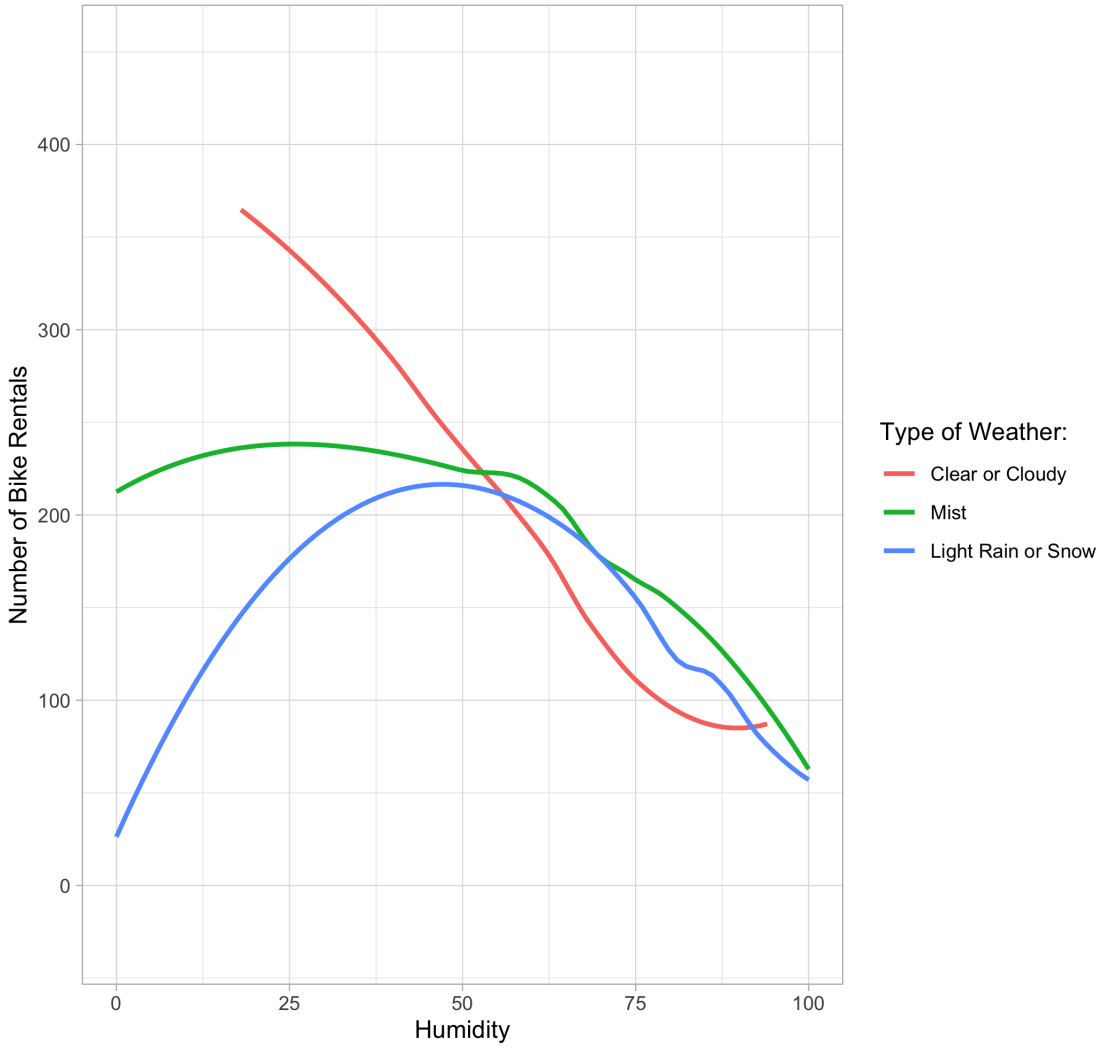
La comparaison des différents **weather** avec le nombre de vélos loués (**count**) indique que la demande de location de vélos est à peu près la même aux *Clear, Misty* , et que le nombre total est meilleur quand le temps est meilleure. Les jours *Light Rain or Light Snow* affichent un nombre moyen de locations nettement inférieur. Je pense que notre ensemble de données ne contient pas d'observations par très mauvais temps(**BAD WEATHER**), de sorte que le boxplot pour le type de temps 4 n'a pas de signification.



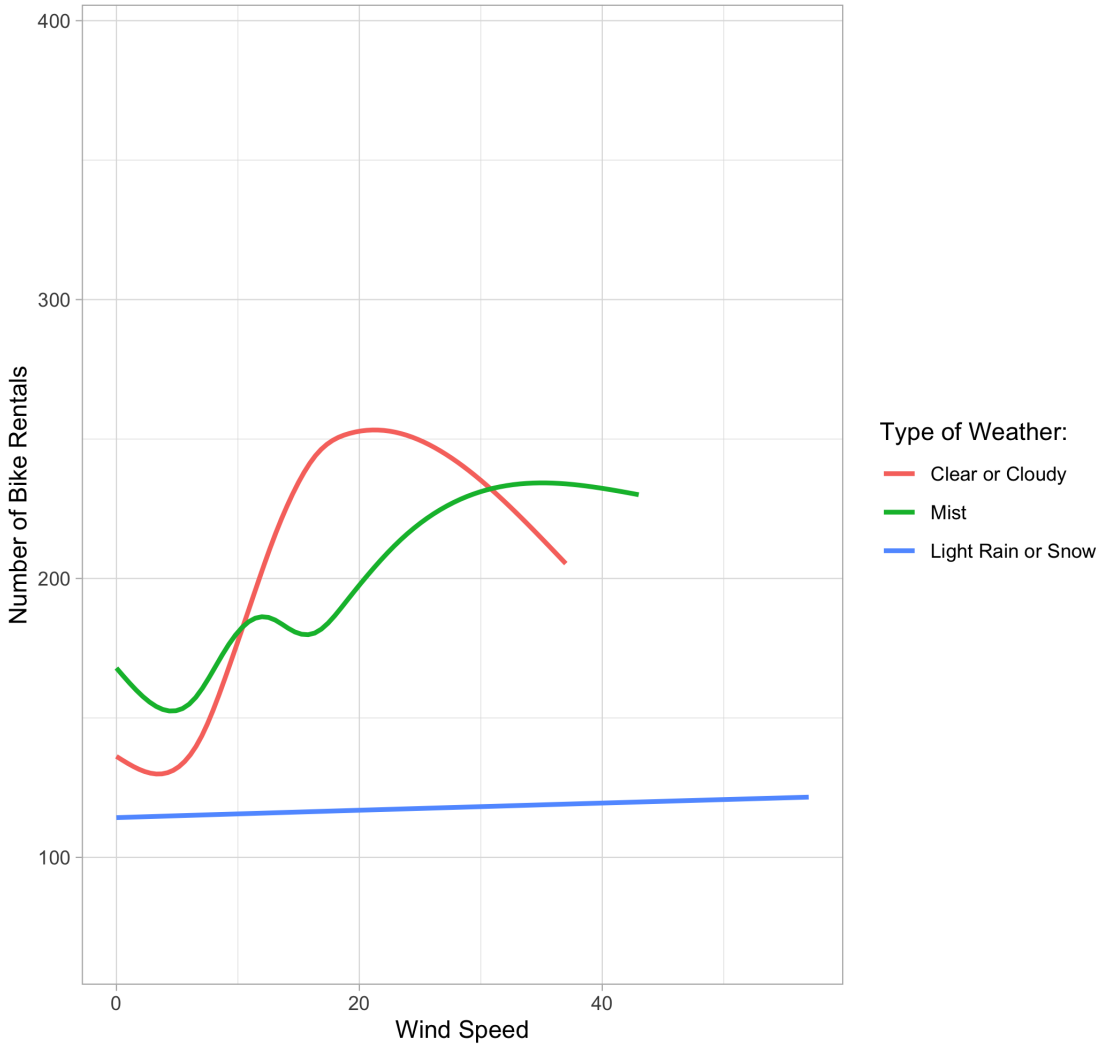
J'ai fais un plot de la **temperature**, **humidity** et **windspeed** en fonction du nombre de vélos loués, classés selon les différents types de temps, **weather**. Le graphique montre que généralement, plus la température est élevée, plus la demande de location de vélos est importante. Cependant, par temps *clear or cloudy, light rain or snow days*, le nombre de vélos à louer atteint environ 32 degrés Celsius.



Le graphique de l'humidité montre qu'en général, plus l'humidité relative est élevée, plus la demande de location de vélos est faible quelques soit le temps.



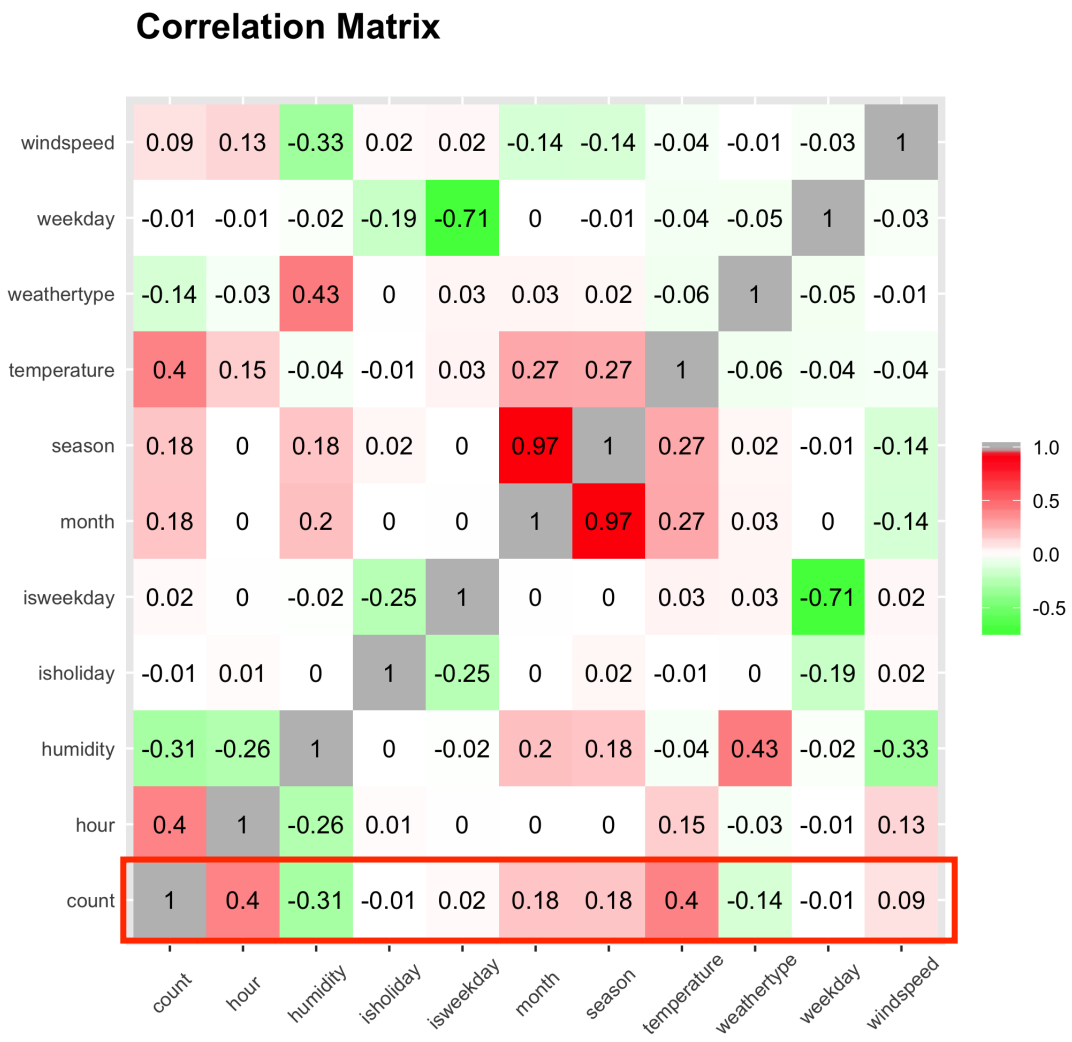
Le graphique de **windspeed** montre que, bien que les gens apprécient l'utilisation quand c'est *Clear ou Mist*, la demande de location de vélos est nettement inférieure, quelle que soit le **windspeed** par temps : *light rain or snow*.



A partir des graphes précédents, on peut dire qu'il peut y avoir de fortes corrélations entre différentes variables.

Correlation :

À partir d'une matrice de corrélation, nous pouvons avoir une vision plus directe des variables qui sont fortement corrélées et de celles qui sont faiblement corrélées.



Nous pouvons clairement voir dans la matrice que l'heure et la température ont la plus forte corrélation avec le nombre de vélos loués(0.4), toutes variables confondues. Cependant, l'heure et la température ont une corrélation relativement élevée entre elles. Nous pouvons ignorer la corrélation significativement élevée entre la saison et le mois puisqu'il est tout à fait naturel qu'ils présentent une corrélation élevée.

3- Mise en oeuvres des méthodes d'apprentissage pour répondre aux questions :

Regression linéaire :

On effectue une régression linéaire en utilisant toutes les variables.

```
Call:
lm(formula = count ~ ., data = bike)

Residuals:
    Min      1Q  Median      3Q     Max
-322.35  -97.93  -30.54   55.84  700.82

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   45.74474   12.16579   3.760 0.000171 ***
season        -8.00475    5.38629  -1.486 0.137273
month          9.94170    1.75059   5.679 1.39e-08 ***
weekday        0.68400    1.17958   0.580 0.562015
hour           7.64795    0.21639  35.344  < 2e-16 ***
isweekday1     2.04321    5.13557   0.398 0.690746
isholiday1    -2.52543   10.41988  -0.242 0.808500
weathertype2  13.02717    3.43377   3.794 0.000149 ***
weathertype3 -26.00676    5.80649  -4.479 7.58e-06 ***
weathertype4 108.67335  147.43985   0.737 0.461096
temperature    7.01155    0.19092  36.724  < 2e-16 ***
humidity      -2.23206    0.09107 -24.509  < 2e-16 ***
windspeed      0.28597    0.18622   1.536 0.124645
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 147.3 on 10873 degrees of freedom
Multiple R-squared:  0.3393,    Adjusted R-squared:  0.3385
F-statistic: 465.3 on 12 and 10873 DF,  p-value: < 2.2e-16
```

toutes les variables

On observe globalement 8 variables significatives qui sont **month, hour, weathertype(2 : Clear or cloudly, 3 : Mist), temperature, humidity**. Avec, toutes les variables ont la même signification dans le modèle.

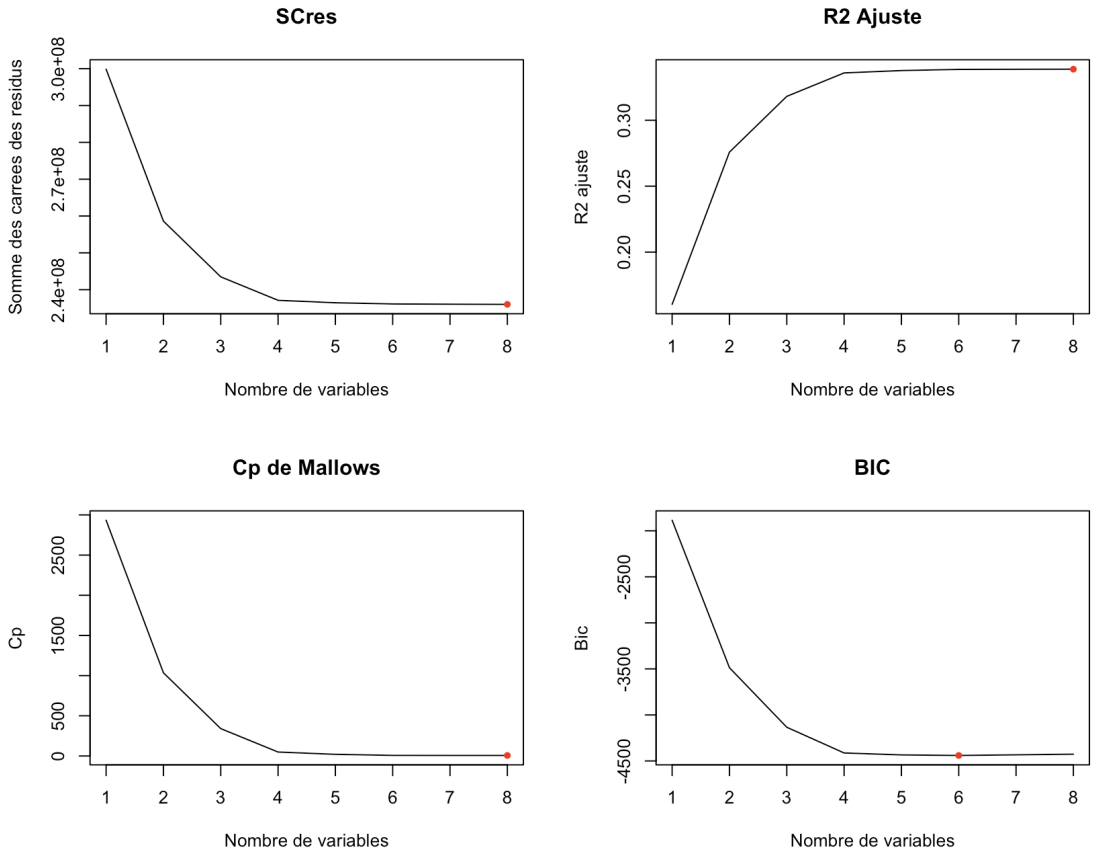
Avant de passer à une selection exhaustive, on fait un choix du meilleur modèle par **AIC (Akaike information criterion)** dans un algorithme progressif à l'aide de la fonction step() qui supprime itérativement les caractéristiques insignifiantes du modèle. Le résultat de l'algorithme est le modèle suivant :

```
Call:
lm(formula = count ~ season + month + hour + weathertype + temperature +
    humidity + windspeed, data = bike)

Residuals:
    Min      1Q  Median      3Q     Max
-322.97  -98.09  -31.12   55.74  701.78

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   50.41641    8.66202   5.820 6.04e-09 ***
season        -8.45712    5.34768  -1.581 0.113803
month         10.08642    1.73880   5.801 6.78e-09 ***
hour           7.64617    0.21634  35.344  < 2e-16 ***
weathertype2  12.93900    3.43092   3.771 0.000163 ***
weathertype3 -25.91687    5.79947  -4.469 7.94e-06 ***
weathertype4 107.45695  147.40104   0.729 0.466011
temperature    7.00806    0.19072  36.746  < 2e-16 ***
humidity      -2.23582    0.09087 -24.605  < 2e-16 ***
windspeed      0.27997    0.18604   1.505 0.132366
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 147.3 on 10876 degrees of freedom
Multiple R-squared:  0.3392,    Adjusted R-squared:  0.3387
F-statistic: 620.4 on 9 and 10876 DF,  p-value: < 2.2e-16
```

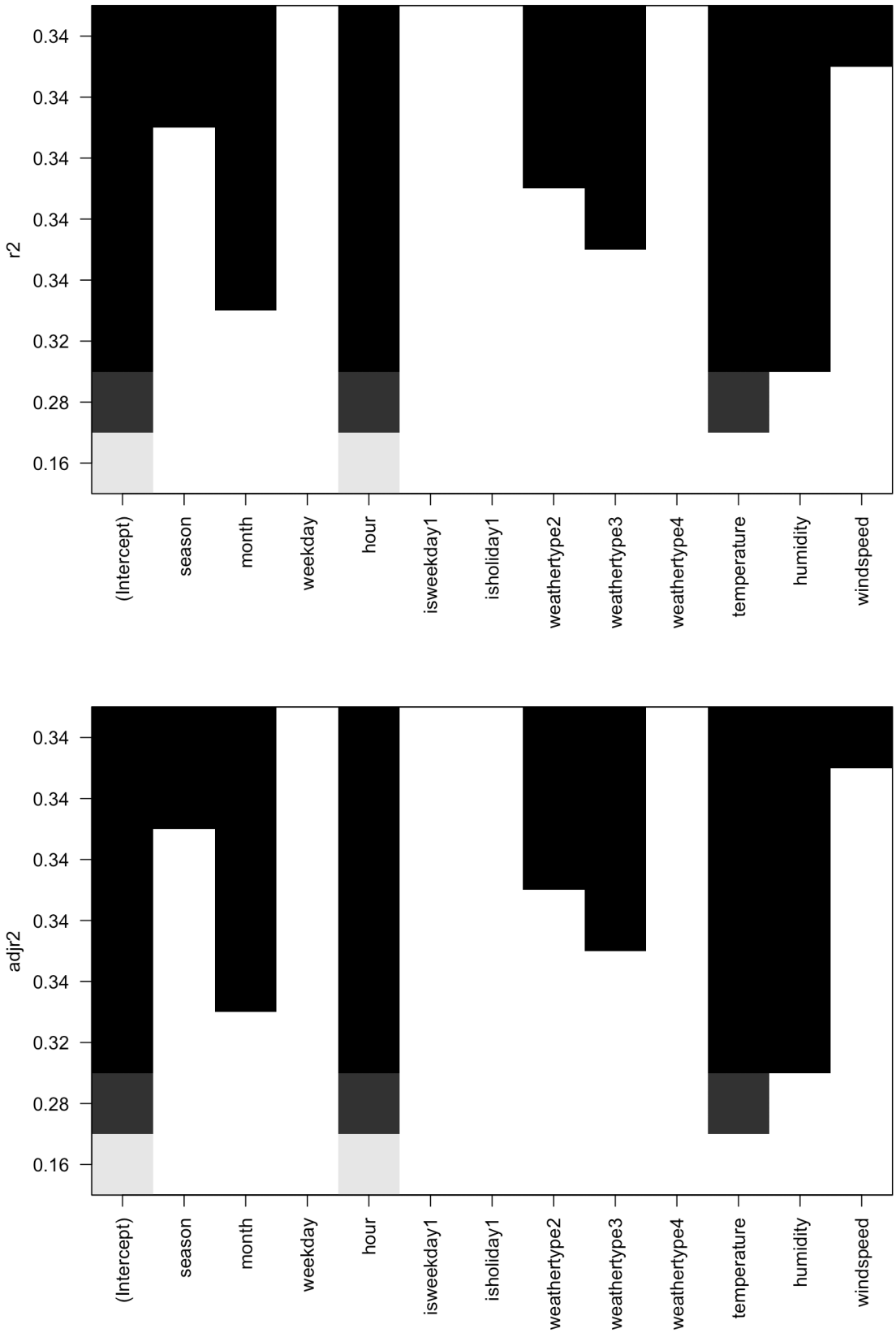
le critère AIC propose un modèle avec 7 variables qui sont : **season, month, hour, weathertype, temperature, humidity et windspeed.**

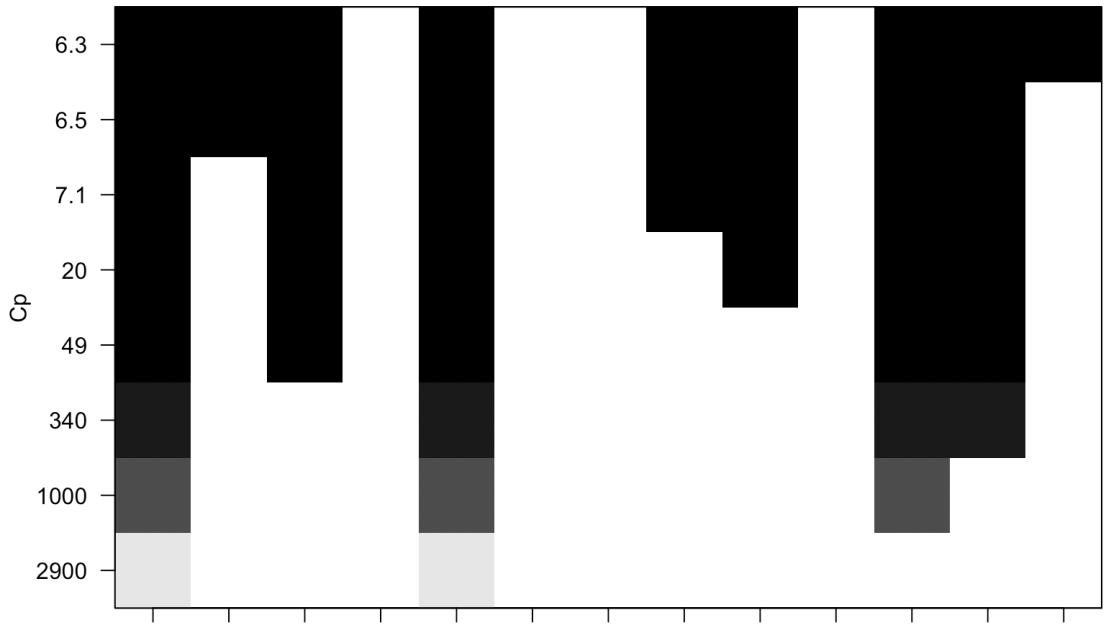
Selection exhaustive :

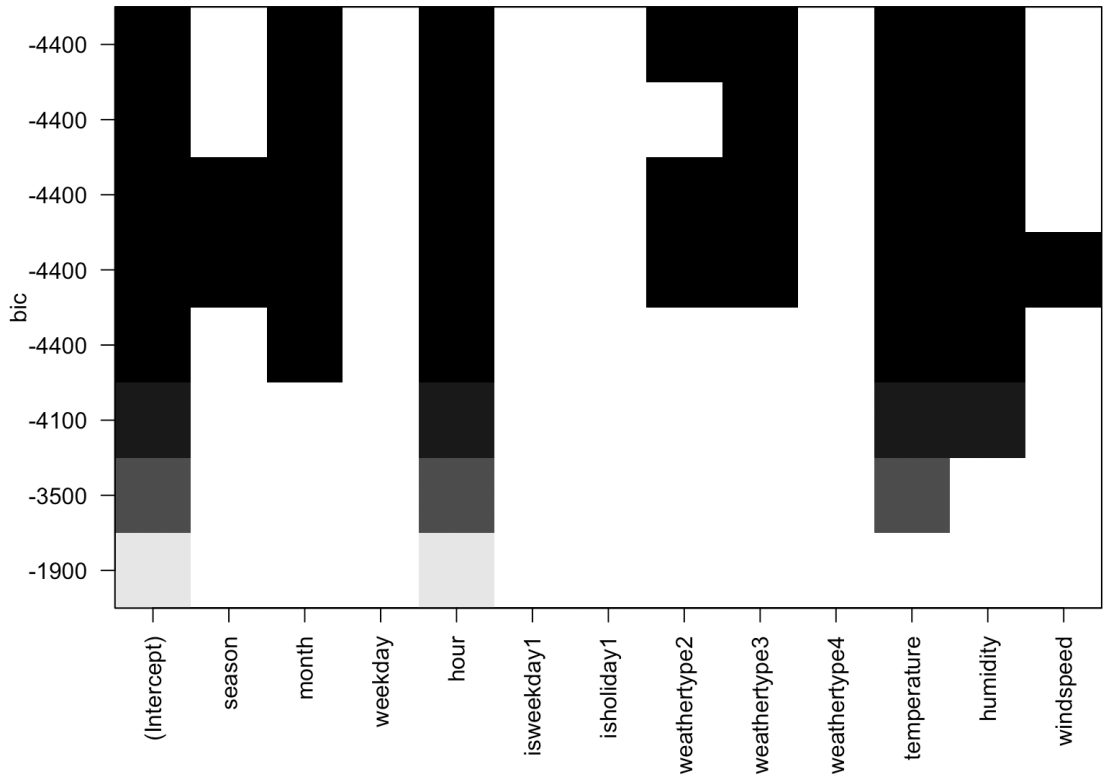
La somme des carrés des résidus (R2), le R2ajusté et le Cp propose un modèle contenant toutes les variables : 8 variables explicatives. Par contre le critère Bic propose le plus petit nombre de variables : 6.



Les graphes ci-dessous permet de visualiser les résultats. Les variables à retenir sont matérialisée par un carré noir en haut des graphiques. On retrouve également les résulats en revenant au tableau des étoiles.


r2
0.34
0.34
0.34
0.34
0.32
0.28
0.16
(Intercept)
season
month
weekday
hour
isweekday1
isholiday1
weathertype2
weathertype3
weathertype4
temperature
humidity
windspeed
adjr2
0.34
0.34
0.34
0.34
0.32
0.28
0.16
(Intercept)
season
month
weekday
hour
isweekday1
isholiday1
weathertype2
weathertype3
weathertype4
temperature
humidity
windspeed


Cp
6.3
6.5
7.1
20
49
340
1000
2900
(Intercept)
season
month
weekday
hour
isweekday1
isholiday1
weathertype2
weathertype3
weathertype4
temperature
humidity
windspeed

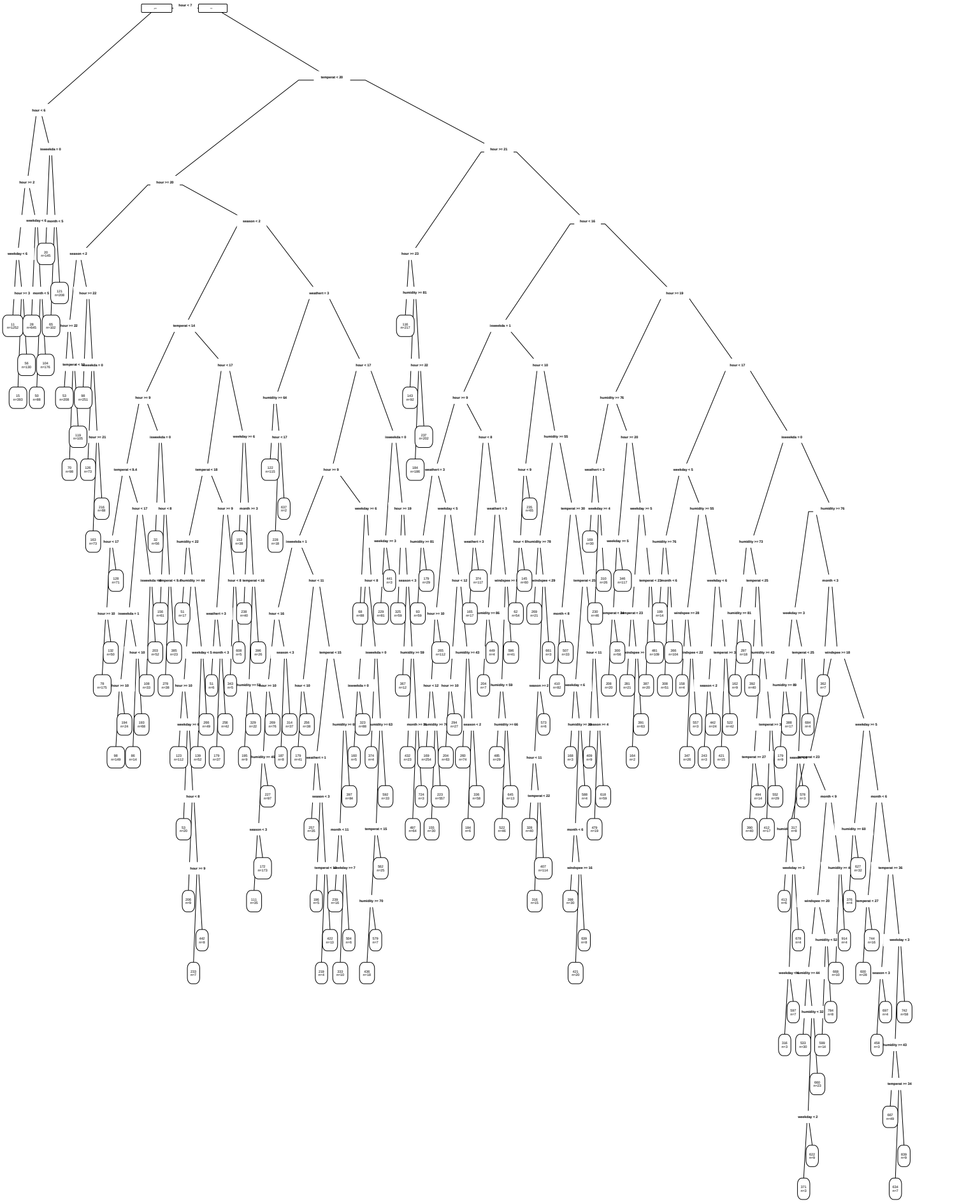

bic
-4400
-4400
-4400
-4400
-4400
-4100
-3500
-1900
(Intercept)
season
month
weekday
hour
isweekday1
isholiday1
weathertype2
weathertype3
weathertype4
temperature
humidity
windspeed

Pour les résultats complets de la régression :

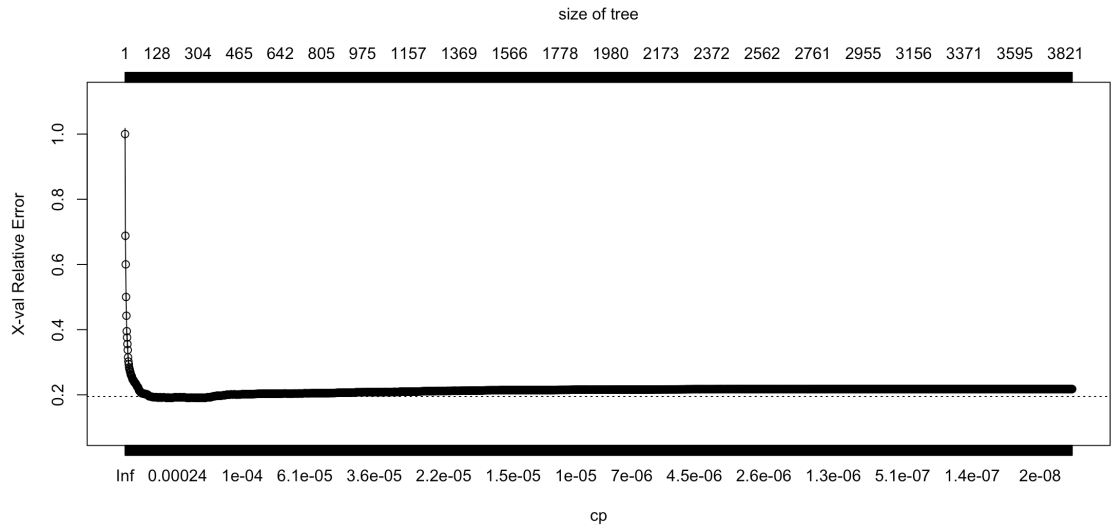
```
Call:
lm(formula = count ~ month + hour + weathertype + temperature +
    humidity, data = bike)

Residuals:
    Min      1Q  Median      3Q     Max
-321.79  -97.90  -30.70   55.31  700.77

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   52.66830    7.16026   7.356 2.04e-13 ***
month          7.37594    0.43740  16.863  < 2e-16 ***
hour           7.66888    0.21603  35.499  < 2e-16 ***
weathertype2  13.33649    3.42404   3.895 9.88e-05 ***
weathertype3 -24.69369    5.72948  -4.310 1.65e-05 ***
weathertype4 104.08103  147.41178   0.706     0.48
temperature    6.99121    0.19058  36.683  < 2e-16 ***
humidity      -2.27317    0.08648 -26.287  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 147.3 on 10878 degrees of freedom
Multiple R-squared:  0.3389,    Adjusted R-squared:  0.3385
F-statistic: 796.7 on 7 and 10878 DF,  p-value: < 2.2e-16
```

Random forest :

Construisons maintenant un arbre avec la fonction **rpart** du package rpart.
L'arbre contient plusieurs niveaux, on remarque que la variable la plus importante est la variable **hour** parce qu'il se situe en racine.



Nous traçons maitenant la représentation de l'erreur par validation croisée en fonction du critère de compléxité **(cp)** de l'arbre lié à sa taille (nombre de feuilles). Dans notre cas, cp = 0.0002714573 est optimal.



La figure suivante représente le degré d'importance de chaque variable sur le nombre de vélos loués :