Prédiction de locations de vélos

Molina Rafidison

13 Jul 2016

PARTIE I : Statistiques descriptives

Préliminaires

Préparation de l'environnement

Les packages nécessaires à l'analyse déjà installés sont chargés.

```
##      ggplot2    gridExtra RColorBrewer        dplyr     reshape2
##         TRUE         TRUE         TRUE         TRUE         TRUE
##        caret
##         TRUE
```

Et le chemin vers le dossier sur lequel est fixé.

Lecture et visualisation de la donnée

La donnée est enregistrée dans une variable; les premières lignes sont affichées.

```
##              datetime season holiday workingday weather temp  atemp
## 1 2011-01-01 00:00:00      1       0          0       1 9.84 14.395
## 2 2011-01-01 01:00:00      1       0          0       1 9.02 13.635
## 3 2011-01-01 02:00:00      1       0          0       1 9.02 13.635
## 4 2011-01-01 03:00:00      1       0          0       1 9.84 14.395
## 5 2011-01-01 04:00:00      1       0          0       1 9.84 14.395
## 6 2011-01-01 05:00:00      1       0          0       2 9.84 12.880
##   humidity windspeed casual registered count
## 1       81    0.0000      3         13    16
## 2       80    0.0000      8         32    40
## 3       80    0.0000      5         27    32
## 4       75    0.0000      3         10    13
## 5       75    0.0000      0          1     1
## 6       75    6.0032      0          1     1
```

Le résumé de la donnée est le suivant :

```
## 'data.frame':    10886 obs. of  12 variables:
##  $ datetime  : Factor w/ 10886 levels "2011-01-01 00:00:00",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ season    : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ holiday   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ workingday: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ weather   : int  1 1 1 1 1 2 1 1 1 1 ...
##  $ temp      : num  9.84 9.02 9.02 9.84 9.84 ...
##  $ atemp     : num  14.4 13.6 13.6 14.4 14.4 ...
```

```
##  $ humidity  : int  81 80 80 75 75 75 80 86 75 76 ...
##  $ windspeed : num  0 0 0 0 0 ...
##  $ casual    : int  3 8 5 3 0 0 2 1 1 8 ...
##  $ registered: int  13 32 27 10 1 1 0 2 7 6 ...
##  $ count     : int  16 40 32 13 1 1 2 3 8 14 ...
```

```
##                 datetime         season         holiday
##  2011-01-01 00:00:00:    1   Min.   :1.000   Min.   :0.00000
##  2011-01-01 01:00:00:    1   1st Qu.:2.000   1st Qu.:0.00000
##  2011-01-01 02:00:00:    1   Median :3.000   Median :0.00000
##  2011-01-01 03:00:00:    1   Mean   :2.507   Mean   :0.02857
##  2011-01-01 04:00:00:    1   3rd Qu.:4.000   3rd Qu.:0.00000
##  2011-01-01 05:00:00:    1   Max.   :4.000   Max.   :1.00000
##  (Other)            :10880
##    workingday        weather           temp           atemp
##  Min.   :0.0000   Min.   :1.000   Min.   : 0.82   Min.   : 0.76
##  1st Qu.:0.0000   1st Qu.:1.000   1st Qu.:13.94   1st Qu.:16.66
##  Median :1.0000   Median :1.000   Median :20.50   Median :24.24
##  Mean   :0.6809   Mean   :1.418   Mean   :20.23   Mean   :23.66
##  3rd Qu.:1.0000   3rd Qu.:2.000   3rd Qu.:26.24   3rd Qu.:31.06
##  Max.   :1.0000   Max.   :4.000   Max.   :41.00   Max.   :45.45
##
##     humidity        windspeed          casual         registered
##  Min.   :  0.00   Min.   : 0.000   Min.   :  0.00   Min.   :  0.0
##  1st Qu.: 47.00   1st Qu.: 7.002   1st Qu.:  4.00   1st Qu.: 36.0
##  Median : 62.00   Median :12.998   Median : 17.00   Median :118.0
##  Mean   : 61.89   Mean   :12.799   Mean   : 36.02   Mean   :155.6
##  3rd Qu.: 77.00   3rd Qu.:16.998   3rd Qu.: 49.00   3rd Qu.:222.0
##  Max.   :100.00   Max.   :56.997   Max.   :367.00   Max.   :886.0
##
##      count
##  Min.   :  1.0
##  1st Qu.: 42.0
##  Median :145.0
##  Mean   :191.6
##  3rd Qu.:284.0
##  Max.   :977.0
##
```

La donnée ne comporte pas de donnée manquante.

Conversion des classes

Les classes doivent être revues pour que les variables puissent être correctement manipulée par la suite.

Exploration

Variables cible `count` et temporelle `datetime`

La variable `datetime` est fondamentale puisque c'est sur cette dernière que repose l'ensemble de l'étude. Pour faciliter l'analyse, elle est décomposée en heure, jour, mois et année.

Les différentes échelles de temps permettent de créer de nouvelles tables de données contenant le nombre de total de locations mais également la moyenne.

Il est intéressant d'observer l'évolution générale de la demande entre le 1er janvier 2011 et le 19 décembre 2012 (Fig.1).

Entre 2011 et 2012, le nombre moyen de locations a significativement augmenté. D'autre part, le comportement est différent entre les mois de mai et septembre sur les deux années (Fig.2).

Un schéma clair se dégage dans l'observation sur une journée. Deux pics sont notables le matin entre 8h et 9h et en fin d'après-midi entre 16h et 20h; le creux se situe entre 23h et 7h du matin. D'un point de vue hebdomadaire, les fluctuations sont plus légères. Le dimanche montre toutefois un nombre moindre de locations comparé autres jours de la semaine. Les jours montrant le plus d'utilisations sont le jeudi, le vendredi et le samedi.

Observations Deux tendances principales se dégagent et cohabitent : 1. Il y a une augmentation générale du nombre de locations de vélos entre 2011 et 2012. Ceci peut laisser penser que le nombre d'abonnés augmente et/ou que la location de vélo se répand. 2. Il y a un cycle journalier qui se dessine. De la même manière, ceci peut laisser penser que les pics correspondent aux heures de pointes de journées de travail. Il serait intéressant d'étudier la différence d'utilisation entre les journées travaillées et celles qui ne le sont pas, et ainsi voir s'il existe un cycle hebdomadaire évident.

Variables catégorielles `season`, `holiday`, `workingday` et `weather`

L'analyse se fait dans un premiers temps sur les variables liées au travail (jours fériés, travaillés, non travaillés). Les deux variables concernées - `holiday` et `working day` sont des variables *dummies* (Fig.3).

Le cycle hebdomadaire se démarque par un changement de comportement entre les week-ends et les jours de semaine : - Les usagers ne travaillant pas le vendredi respectent un comportement similaire à ceux qui travaillent. - En revanche, le lundi est plus équilibré en mélangeant les deux comportements. - Le mercredi est le jour le plus curieux : la location de vélos est plus massive que les autres jours pour les usagers ne travaillant. Un pic est également observé entre 20h et 23h.

Le comportement entre un jour férié et un jour non travaillé est-il le même (Fig.4) ?

Le comportement est différent. L'activité lors de jours fériés présente des pics similaires jours travaillés et est moins élevée que lors des jours non travaillés. La tendance est néanmoins plus lisse que lors de jours travaillés.

Note : L'étude des jours fériés montre que la donnée provient certainement des États-Unis. Ceci pourrait expliquer des comportements dans les usages qui seraient spécifiques au pays ou même à l'état, la ville, etc...

Les deux autres variables à étudier sont liées au temps, `season` et `weather`. Elles comportent chacunes quatre modalités. La variable `temp` est intégrée à cette exploration étant donné l'étude sur la météorologie (Fig.5).

Les interprétations de ce graphique sont les suivantes : - Le printemps représente la saison la plus froide, ce qui peut expliquer le faible nombre de locations. - Les usagers sont plus actifs lors des saisons les plus chaudes qui sont l'automne et l'été. Ils sont également plus nombreux en hiver qu'au printemps. - Le vélo partagé est particulièrement plebiscité entre 19°C et 35°C environ. - Les locations sont plus nombreuses lorsque le temps est correct (rouge et orange) et plus faibles par temps de légère pluie/neige. La pluie freine moins les usagers lorsqu'il fait chaud.

Observations L'heure, le jour de la semaine, la saison et le temps semblent jouer un rôle important dans l'utilisation de vélos partagés. Les jours de vacances scolaires (fériés) peuvent paraître moins déterminants.

Variables continues `temp`, `atemp`, `humidity` et `windspeed`

Précédemment étudiée, la température `temp` doit maintenant être comparée à la température ressentie `atemp` afin de déterminer si l'une a plus d'influence sur l'autre (Fig.6).

```
##       temp           atemp
##  Min.   : 0.82   Min.   : 0.76
##  1st Qu.:13.94   1st Qu.:16.66
##  Median :20.50   Median :24.24
##  Mean   :20.23   Mean   :23.66
##  3rd Qu.:26.24   3rd Qu.:31.06
##  Max.   :41.00   Max.   :45.45
```

```
## [1] 0.9849481
```

En bleu est représenté la température et en orange, la température ressentie. Les deux variables sont très largement corrélées même si les températures ressenties fluctuent davantage. Les températures sont généralement senties comme étant plus chaudes qu'elles ne le sont. En revanche, certains points sont étrangement éloignés et demandent d'y jeter un œil.

```
## Source: local data frame [1 x 6]
##
##          day     temp atemp humidity windspeed    count
##       (fctr)    (dbl) (dbl)    (dbl)     (dbl)    (dbl)
## 1 17-08-2012 29.65667 12.12 57.08333  15.50073 297.8333
```

Il s'agit d'un jour en particulier : le 17 août 2012. Sur l'ensemble de la journée, l'humidité est légèrement élevée, le vent faible. Il pourrait s'agir d'une erreur sur la température ressentie.

```
##     humidity        windspeed
##  Min.   :  0.00   Min.   : 0.000
##  1st Qu.: 47.00   1st Qu.: 7.002
##  Median : 62.00   Median :12.998
##  Mean   : 61.89   Mean   :12.799
##  3rd Qu.: 77.00   3rd Qu.:16.998
##  Max.   :100.00   Max.   :56.997
```

```
## [1] -0.3173715
```

```
## [1] 0.1013695
```

De manière générale, l'humidité et la vitesse du vent semblent d'avoir qu'un impact léger sur le nombre de locations de vélos.

```
## Source: local data frame [4 x 3]
##
##   weather meanHumidity totalCount
##    (fctr)        (dbl)      (int)
## 1       1     56.71677    1476063
## 2       2     69.10056     507160
## 3       3     81.34109     102089
## 4       4     86.00000        164
```

D'autre part, d'après le tableau précédent, il semble que le taux d'humidité soit lié la variable `weather`. En effet, cette dernière indique la présence de pluie. Dans le cadre de cet exercice, la variable `windspeed` ne sera pas étudiée davantage.

Réponses

Variables influentes

L'exploration de la donnée montre que la demande en vélos semble influencée par l'heure, le jour, la saison, le temps, la température.

Les variables `casual` et `registered` peuvent être étudiées de plus près pour déterminer la proportion d'utilisateurs habituels ou ponctuels.

Variables `age` et `sexe`

Avec les informations concernant l'âge et le sexe des utilisateurs abonnés, la procédure statistique à mener pour comparer les distributions est la suivante.

1. Il y a deux populations avec d'un côté la distribution en âge des hommes et la distribution en âge des femmes de l'autre. Les deux échantillons proviennent d'une même population. Ils sont donc considérés comme indépendants.
2. Il peut être intéressant de réaliser un graphique représentant les densités des deux échantillons superposées pour comparer leurs distributions. Un graphique avec les boxplots associés donneraient également plus d'informations sur les quantiles.
3. L'âge est considéré comme une variable continue. Pour comparer les deux échantillons, il faut réaliser un test de Kolmogorov-Smirnov à deux échantillons (two-sample Kolmogorov-Smirnov) via la fonction `ks.test`. L'hypothèse nulle est : "Les deux échantillons ont la même distribution".

Le résultat du test donnera la valeur p-value qui permettra de dire si l'hypothèse nulle doit être rejetée ou pas. En considérant un niveau d'importance de 5%, si la p-value est inférieure à 0.05 alors l'hypothèse nulle est rejetée et les deux populations n'ont pas la même distribution, donc non identiques.

PARTIE II : Machine learning

Sélection de variables

Comme vu précédemment, seules quelques variables sont gardées.

Cross-validation

Le nouveau dataset est divisé en trois parties dans le cadre d'une validation croisée : un set d'entraînement (`train`), un set de validation (`valid`) et un set de test (`test`).

```
## [1] 6969    7
```

```
## [1] 1741    7
```

Un paramètre de cross-validation automatique est utilisé une nouvelle fois pour être utilisé lors de l'entraînement du modèle.

Random Forest

Entraînement

La méthode Random Forest est utilisée afin de prédire le nombre de locations de vélos par heure. La technique recherchée est celle de la régression.

Le paramètre `mtry` représente le nombre de variables aléatoires qui sont sélectionnées et testées à chaque embranchement. Son choix est sensible. Pour le définir et laisser de la liberté, la moitié du nombre de prédicteur a d'abord été testé. Finalement, le choix s'est arrêté sur 28 prédicteurs qui donne une meilleure performance.

Pour l'entraînement, le nombre d'arbres permis (ntree) est fixé à 500 par défaut et l'erreur moyenne quadratique (RMSE) est choisie comme critère de performance. Le RMSE permet de mesurer la dérivation moyenne des valeurs prédites des valeurs observées. Plus le résultat est bas, plus les prédictions se rapprochent des observations.

```
## Random Forest
##
## 6969 samples
##   37 predictor
##
## No pre-processing
## Resampling: Cross-Validated (2 fold)
## Summary of sample sizes: 3484, 3485
## Resampling results:
##
##   RMSE      Rsquared
##   66.52393  0.8671267
##
## Tuning parameter 'mtry' was held constant at a value of 28
##
```

```
##
## Call:
##  randomForest(x = x, y = y, ntree = 500, mtry = param$mtry, importance = TRUE)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 28
##
##           Mean of squared residuals: 3473.192
##                     % Var explained: 89.41
```

Validation

Le modèle est testé sur le set de validation avant d'être appliqué sur le test final. Cela permet d'observer s'il n'y a pas de cas de surentraînement (overfitting).

```
## [1] 56.23357
```

Le score RMSE est moins élevé que celui de l'entraînement et est, de manière générale, satisfaisant. Cela indique également qu'il n'y a pas d'overfitting.

Importance des variables

Les variables sont rassemblées dans un graphique par ordre d'importance dans le modèle.

Seule la variable `weekday` n'apparait pas dans ce graphique.

Prédiction

Prédictions finales

Le modèle est appliqué au set final à prédire.

Comparaison

```
## [1] 56.30055
```

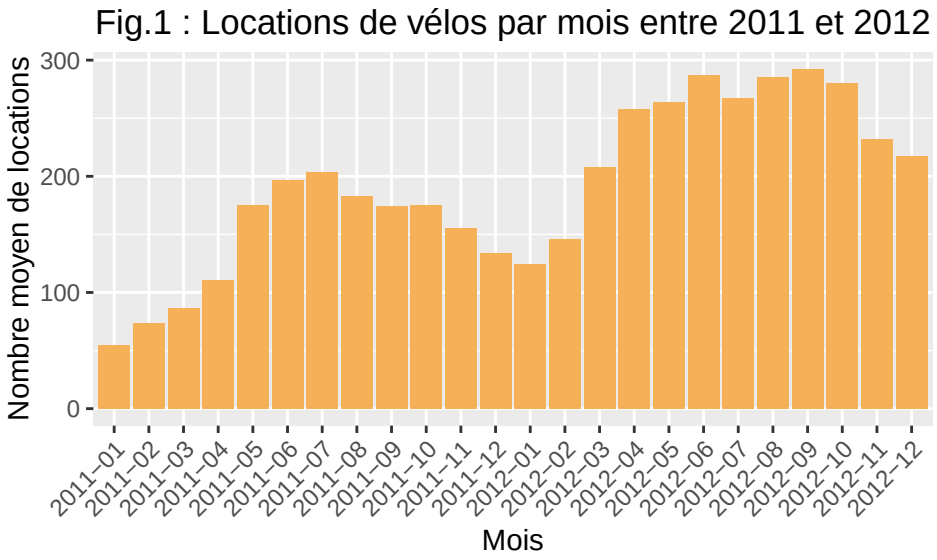
Le résultat des prédiction est conforme à ce qu'il s'est passé sur le set de validation.

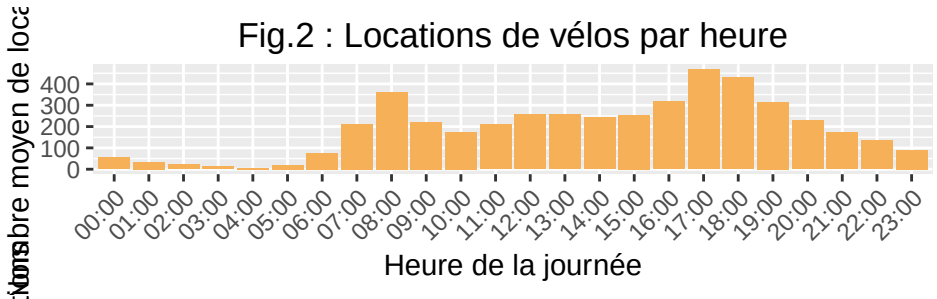
Pistes d'amélioration

Voici quelques pistes qui mériteraient d'être prises en compte pour l'amélioration du modèle : - Étudier davantage les variables importantes et améliorer la sélection de variables; - Affiner le nombre de prédicteurs sélectionnés pour le `mtry`; - Analyser de manière plus approfondie le comportement du modèle via notamment la fonction de distribution cumulative empirique permettant de réaliser des graphiques.

Appendix

Les graphiques cités dans le rapport sont disponibles dans cette annexe.





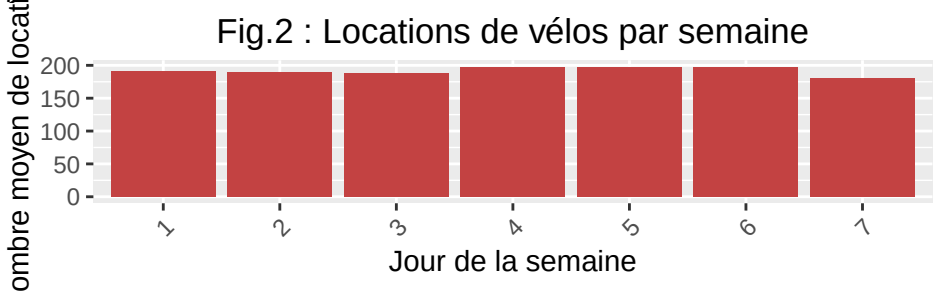


Fig.3 : Locations de vélos sur une semaine

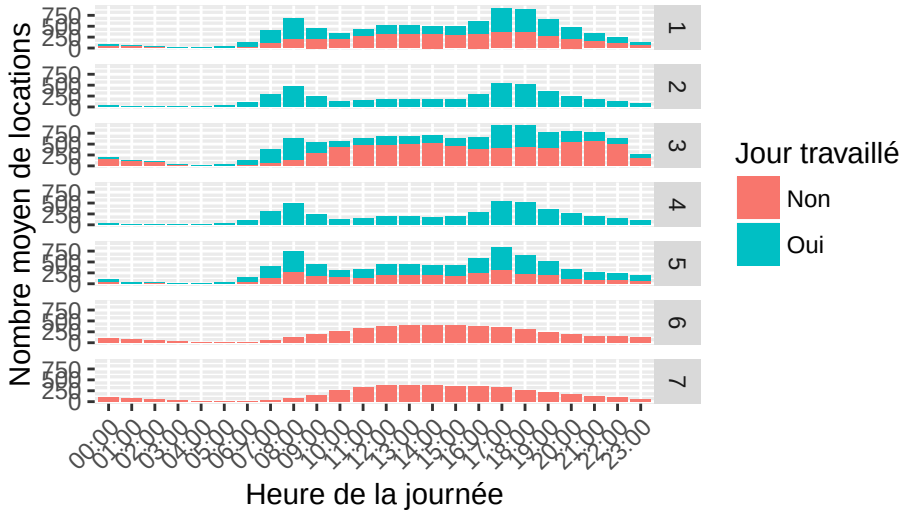


Fig.4 : Locations de vélos les jours fériés

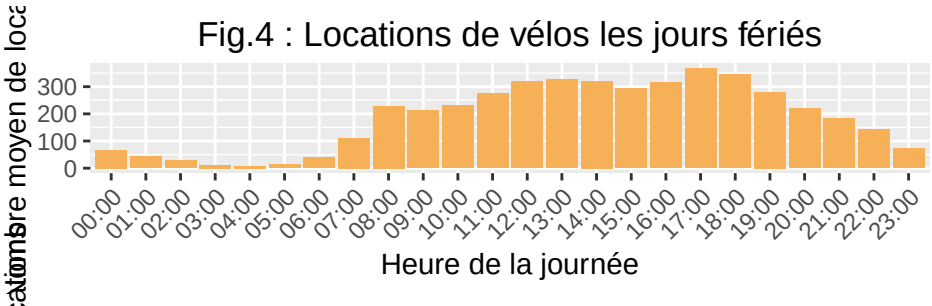
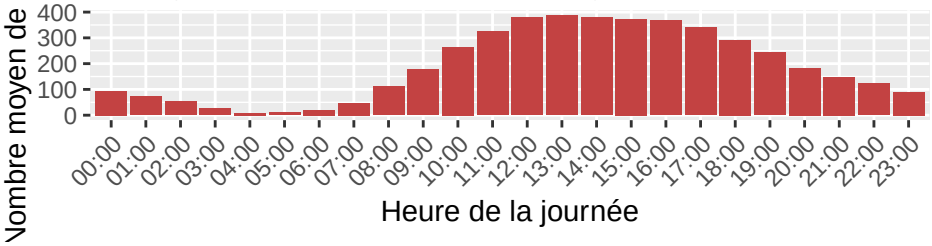
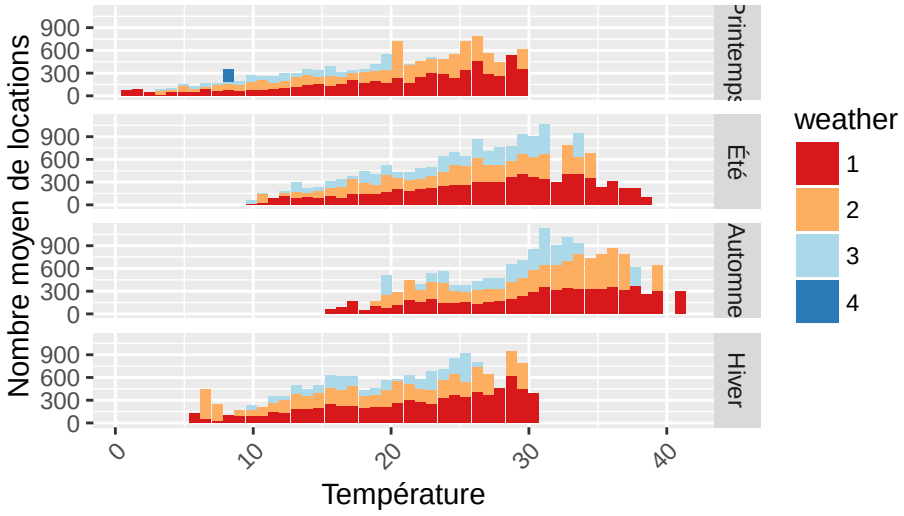


Fig.4 : Locations de vélos les jours non travaillés



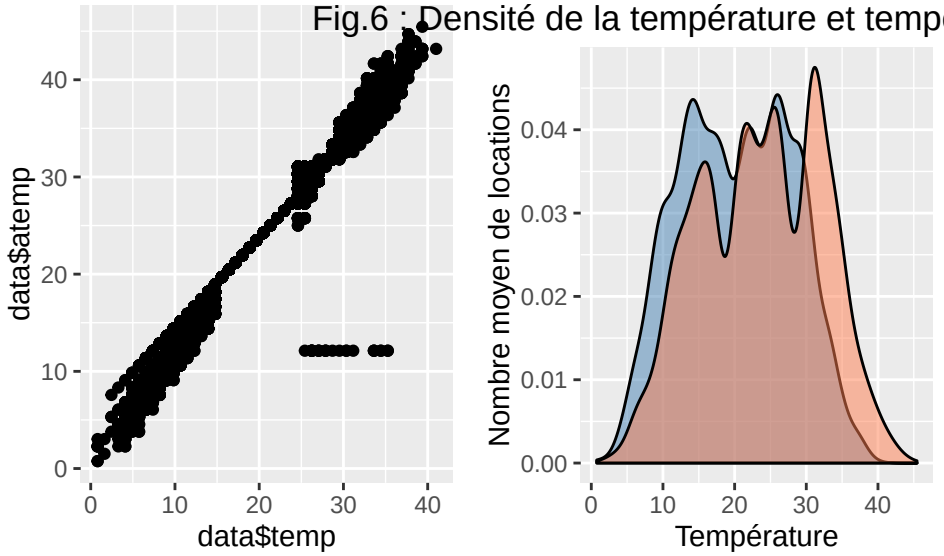
ations de vélos selon la température, la météorologie et la sais

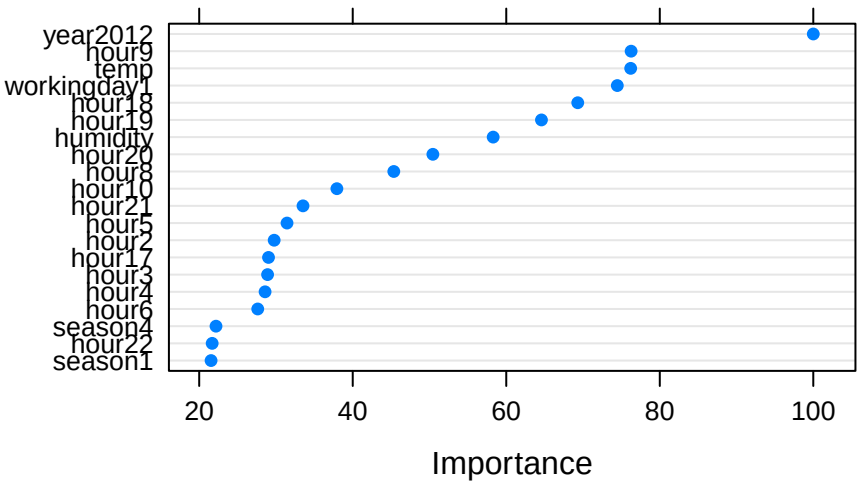



```
##       temp           atemp
##  Min.   : 0.82   Min.   : 0.76
##  1st Qu.:13.94   1st Qu.:16.66
##  Median :20.50   Median :24.24
##  Mean   :20.23   Mean   :23.66
##  3rd Qu.:26.24   3rd Qu.:31.06
##  Max.   :41.00   Max.   :45.45
```

```
## [1] 0.9849481
```

Fig.6 : Densité de la température et tempe




year2012
hour9
temp
workingday1
hour18
hour19
humidity
hour20
hour8
hour10
hour21
hour5
hour2
hour17
hour3
hour4
hour6
season4
hour22
season1
20
40
60
80
100
Importance