Partie I – Statistiques descriptives

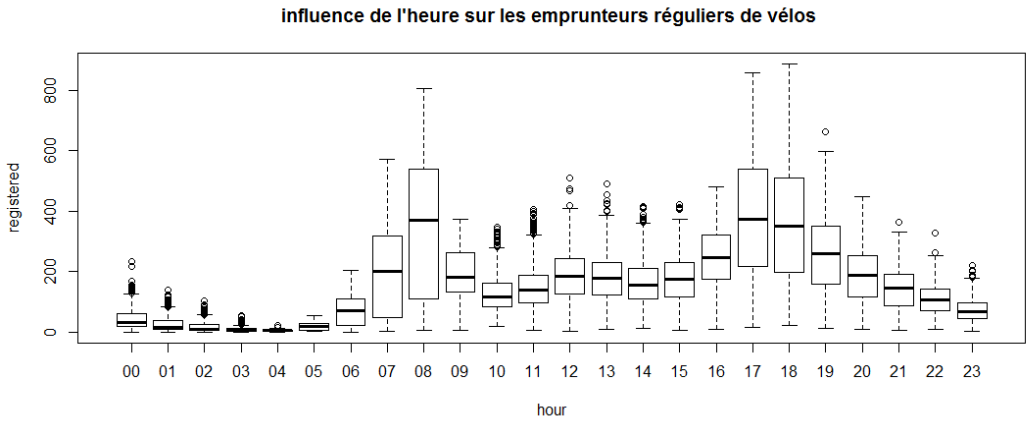
1. Sans commencer la partie modélisation, quels sont les facteurs qui semblent influencer la demande en vélos ? Justifiez vos choix. Présentez quelques graphiques (6 maximums) pertinents pour illustrer votre réponse et interprétez-les.

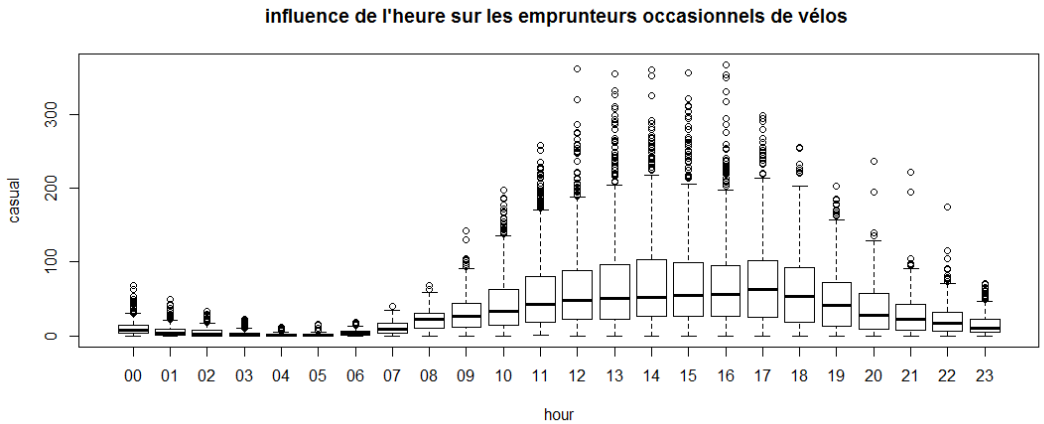
a) Prise en main du problème

Dans un premier temps, il faut réfléchir au problème sans regarder les données, en s'interrogeant sur les caractéristiques qui influencent un usager à emprunter un vélo : l'horaire, la météo, la température, la circulation, la pollution, la distance à parcourir, la disponibilité des vélos, le besoin en fonction du type de journée et si il s'agit d'un client régulier. Il faut aussi penser aux indices laissés par l'utilisation de vélos : la localisation de l'emprunt et du dépôt du vélo, le temps de l'emprunt, si possible le nom (un identifiant) du client.

b) Influence de l'horaire sur l'emprunt de vélos : « registered » et « casual »

Les données fournies permettent d'observer la distribution du nombre de vélos empruntés à chaque heure de la journée. La variable « count » est la somme de deux variables : « registered » qui représente les personnes enregistrées (utilisateurs réguliers) et « casual » les usagers occasionnels. Il est intéressant de les séparer pour comparer leurs habitudes.





On constate que l'évolution du nombre d'emprunts (médiane) réguliers atteint un pic le matin vers 8h et en fin d'après-midi vers 17h. Cela correspond certainement aux horaires de bureaux. Concernant les emprunteurs occasionnels, l'évolution est progressive jusque 14h puis diminue.

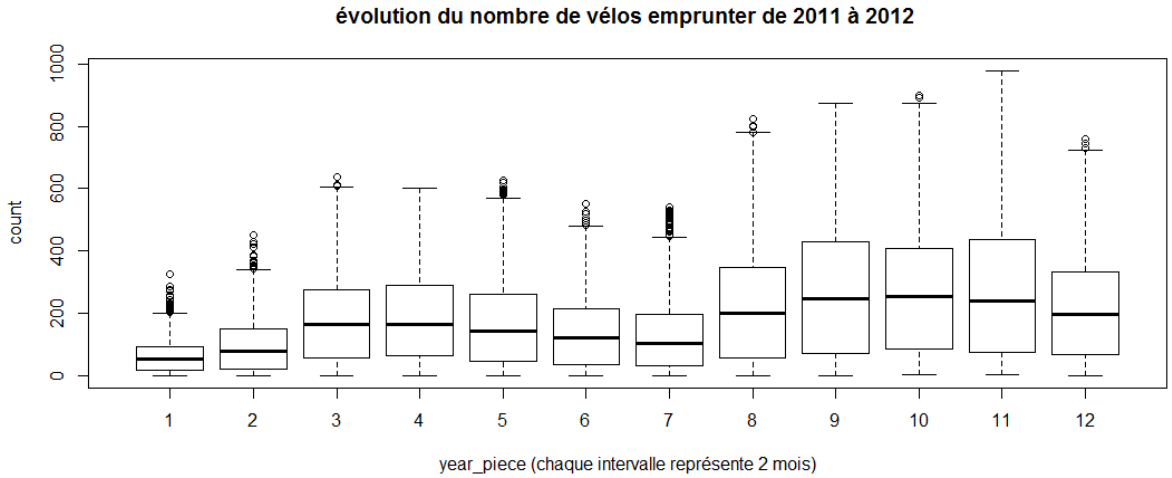
Le deuxième constat important est la grande dissymétrie concernant la distribution de l'emprunt de vélos entre les usagers réguliers et occasionnels. La distribution est très étendue pour le second graphique avec de nombreux résultats « aberrants ».

Ces deux tendances révèlent l'importance de prédire la variable « count » par le biais de « registered » et de « casual » indépendamment. De plus, le nombre d'emprunt est principalement lié aux utilisateurs réguliers, il faut donc s'assurer d'avoir la meilleure précision sur le modèle correspondant.

Sur ces deux graphiques, il y a aussi beaucoup de valeur qui semblent aberrantes. C'est lié au fait que les gens puissent prendre le vélo de manière quasiment aléatoire (un concert ou d'autres évènements peu fréquents par exemple). Pour les traiter, j'ai utilisé la fonction logarithmique sur « casual » et « registered ». Durant la modélisation, j'ai remodifié la fonction pour améliorer la précision.

c) Evolution du nombre d'emprunt entre 2011 et 2012

Les données fournies s'étalent sur deux ans, il est possible que le nombre d'emprunts évolue globalement plus ou moins d'une année à l'autre en fonction de la politique d'une ville… J'ai donc créé la variable « year_piece », où chaque valeur représente deux mois consécutifs d'une année.

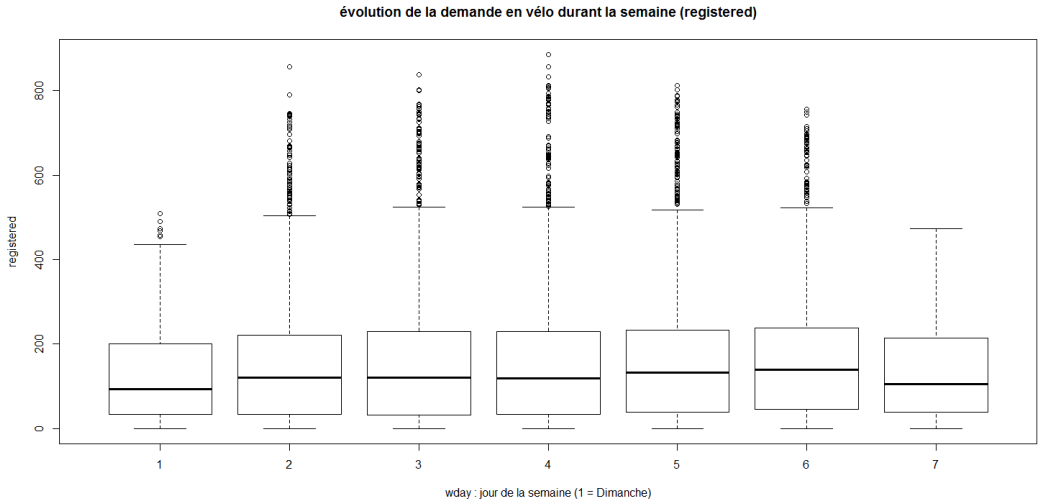


Ce diagramme montre que la tendance globale (dans la boite : 1er au 3ème quartile) et la médiane du nombre de vélos empruntés « count » augmentent de 2011 à 2012.

De plus, on constate que la période de l'année (de 1 à 6 puis de 6 à 12) influence considérablement le nombre d'emprunts. Cela peut être une redondance de l'influence de la météo et de la température, mais cela peut aussi être dû à des facteurs qui ne sont pas présents dans nos données comme l'horaire de levé et de couché du soleil… Ce facteur semble donc utile pour la partie modélisation.

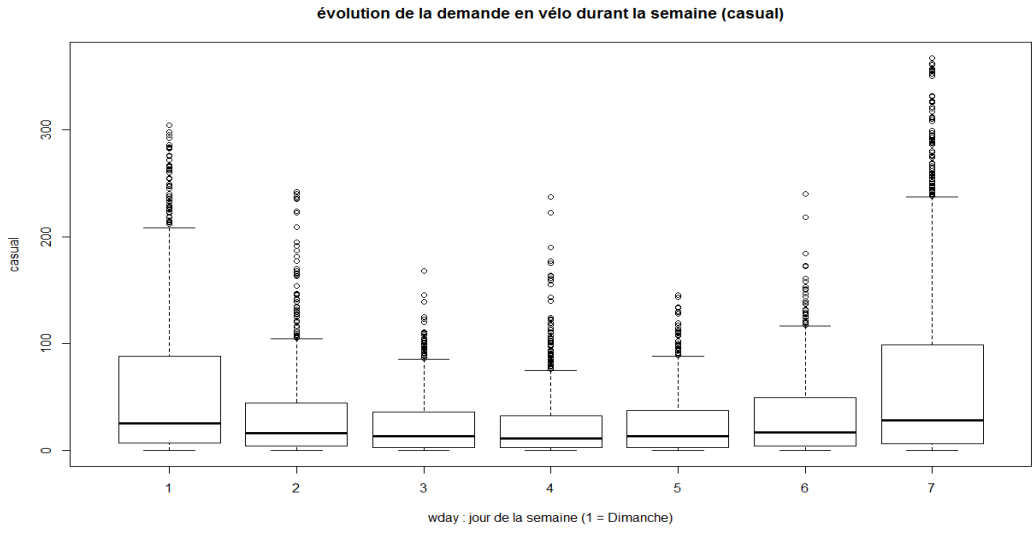
d) Evolution de la demande en vélo en semaine : « registered » et « casual »

L'utilisation d'un vélo dépend du besoin de l'utilisateur, il est donc probable qu'elle diffère selon que l'on soit en weekend ou non. Pour ne pas oublier de pistes, il est préférable d'afficher le nombre de vélos en fonction de chaque jour de la semaine.



Chez les utilisateurs réguliers, la médiane ainsi que la taille des boites à moustache (entre le 1er et le 3ème quartile) sont très semblables du lundi au vendredi et elles sont légèrement inférieures durant le weekend. On constate aussi que la distribution est très étalée, avec de nombreux dépassements de la valeur pivot supérieure (valant environ 550 emprunts) en semaine et non le weekend.

Il faut donc créer une variable « weekend » pour vérifier son influence dans la partie modélisation.



Chez les utilisateurs occasionnels, le nombre d'emprunt varie plus d'un jour à l'autre. Le weekend a une influence considérable dont il faudra tenir compte. Ce graphique montre aussi que le Lundi et le Vendredi sont des journées où la demande en vélo augmente, comme ce tableau affichant la moyenne de vélos empruntés en fonction du jour de la semaine le montre.

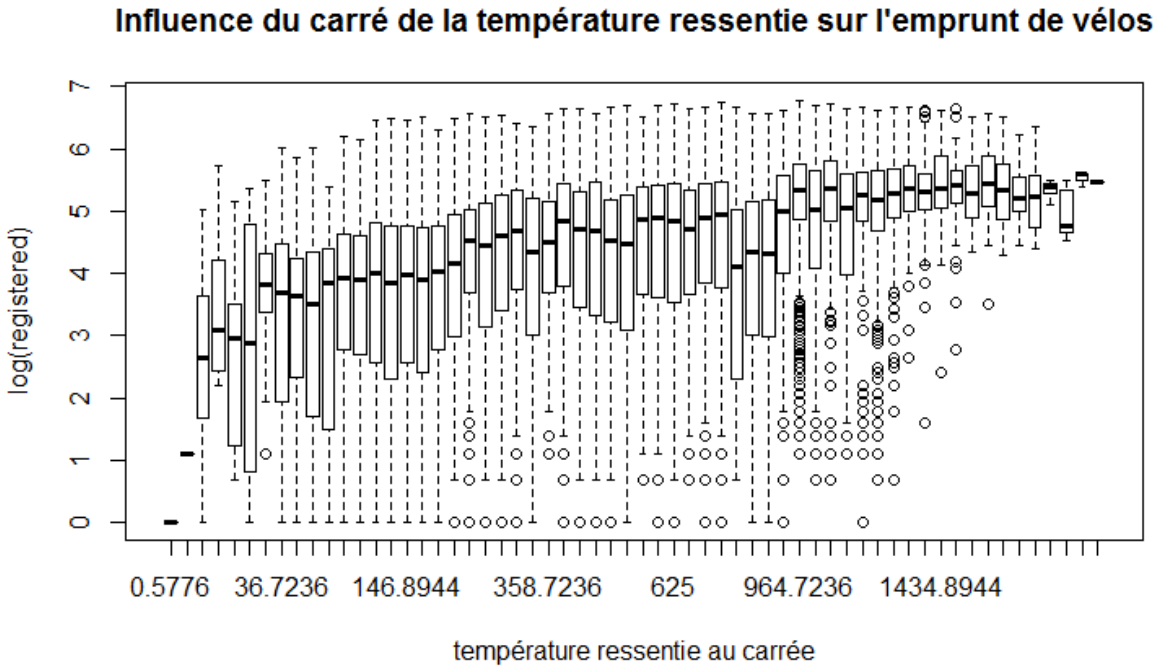
wday	Casual
1	57,000
2	30,000
3	23,000
4	23,000
5	24,000
6	31,000
7	64,000

Figure 1: moyenne du nombre de vélos empruntés

Cette analyse permet d'ajouter une variable supplémentaire indiquant si l'on est lundi ou vendredi ou non pour notre modélisation (casual).

e) Influence des conditions météo

La météo influence le choix de se déplacer en vélo. L'objectif est de voir qu'elles sont les principaux critères qui influencent les utilisateurs. Le critère de pluie semble particulièrement important, de même celui de la température ressentie. Afin d'augmenter la pertinence du critère « atemp », il peut être utile de le mettre au carré pour étendre sa distribution.



Le graphique montre que « atemp » au carré a une influence sur le nombre de vélos empruntés « registered ». Mais la distribution reste étendue, plus que si l'on comparait ce diagramme à celui du log(registered) en fonction de « hour ».

Cela nous montre la limite des variables de bases, même au carré. Il semble donc judicieux de créer des variables plus complexes, qui tiendraient compte à la fois de la température ressentie est de la saison. Ce raisonnement peut être repris pour la variable « humidity », qui aura une incidence plus forte si elle est couplée avec la température, ou bien avec le temps « weather » si il pleut ou non.

Partie II – Machine Learning

1. Concevez un modèle permettant de prédire la variable **count** et expliquez votre choix d'algorithme. Si votre modèle comporte des spécificités de paramétrage, justifiez également vos choix de paramètres.

Après avoir créé quelques nouvelles variables (year_piece…), j'ai divisé les données « data » en une partie d'entraînement « train » et une autre pour tester mes modèles « test ». J'ai alors créé d'autres variables qui sont dépendantes en partie de la valeur « registered » ou « casual » : « humidity_reg », « hour_reg » …

La valeur de la variable recherchée étant un entier compris entre 0 et 900, J'ai commencé par expérimenter une simple régression linéaire. Pour calculer la précision, je me suis servi du coefficient de détermination. Cela ne donnait pas de résultats vraiment concluant avec R^2 atteignant environ 0.8.

Je me suis alors servi de modèles plus complexes avec des arbres de décisions puis de random forest qui me donnaient de meilleurs résultats. Afin de voir les améliorations (notamment par rapport à la régression linéaire) j'ai utilisé l'erreur type RMSE et RMSLE.

J'ai principalement utilisé random forest pour améliorer mon modèle. Je lui avais fixé comme paramètre ntree = 100 afin de pouvoir facilement relancer l'algorithme. Pour confirmer de nouveau paramètre, je laissais l'algorithme converger en augmentant la valeur de ntree à 1000.

Afin de voir les variables les plus influentes concernant la précision du modèle, j'utilisais varImpPlot. Puis je créais de nouvelles variables avec celles ayant une valeur moyenne de précision importante (dans varImpPlot). Ces nouvelles variables doivent avoir une logique : comme « ws » qui est le « weather » multiplier par la « saison ». En effet, de la pluie en été aura un effet moins négatif que de la pluie en hiver sur le nombre de vélos empruntés.

Enfin, j'ai utilisé cforest (utilisant l'inférence conditionnelle) pour obtenir les meilleurs résultats RMSE et RMSLE. J'ai alors utilisé ntree= 100 avec mtry=10 pour obtenir les meilleurs résultats, avant de faire converger l'algorithme avec ntree=300. Il faut penser lors de la prédiction à mettre la méthode d'estimation d'erreur out-of-bag « OOB » = TRUE.

Le résultat sur la trame de données test est le suivant : **RMSE = 47.1** et **RMSLE = 0.347 (pred_count)**

2. Décrivez le critère de performance utilisé lors de la conception de votre algorithme, et justifiez en le choix.

Le critère de performance dépend directement du besoin du client (le loueur). Est-il vraiment important d'être précis aux heures creuses ? Est-il plus important d'être précis lorsqu'il y a beaucoup de vélos empruntés ?
J'ai commencé en utilisant R^2 puis le RMSE, cependant, ces critères ne tiennent pas compte du fait que notre prédiction a tendance à être supérieure ou inférieure au réel.
Je me suis alors mis à la place du loueur, le plus important, pour que le service se développe, c'est que les gens puissent toujours prendre ou déposer leur vélo. Il est donc préférable que la prédiction soit supérieure à la réalité. Dans ce cas, le meilleur critère est le RMSLE (erreur logarithmique) qui pénalise d'avantage une prédiction qui sous-estime la réalité plutôt qu'une prédiction qui surestime le nombre de vélos empruntés.

3. Proposez deux à trois pistes d'amélioration de votre modèle.

Pour améliorer le modèle, il faudrait utiliser d'avantages de variables en testant toutes les variables entières au carré (voir au cube). Créer des variables interdépendantes comme je l'ai fait avec « ws » ou « atemp_reg ». Obtenir d'avantage de données : la localisation de l'emprunt et du dépôt du vélo, le temps de l'emprunt, un identifiant du client, le nom de la ville (pour connaître les évènements qui s'y passe, les jours fériés…).

Il faudrait aussi utiliser des cross validations pour s'assurer que le modèle marche sans être over-fitted (ce qui est peu fréquent avec rforest). De plus, la valeur mtry a été déterminée manuellement, il faudrait utiliser la fonction tuneRF pour l'optimiser. Utiliser d'autres modèles en même temps (svm ou régression linéaire) et faire un ensemble pourrait aussi améliorer la précision.

Enfin, concernant le critère d'évaluation, il serait intéressant d'en faire un plus personnalisé en fonction des besoins du loueur. On pourrait diviser ou soustraire le nombre « pred_count » par la valeur réelle (différent de 0), multiplier ce nouveau vecteur par une matrice de pénalité.