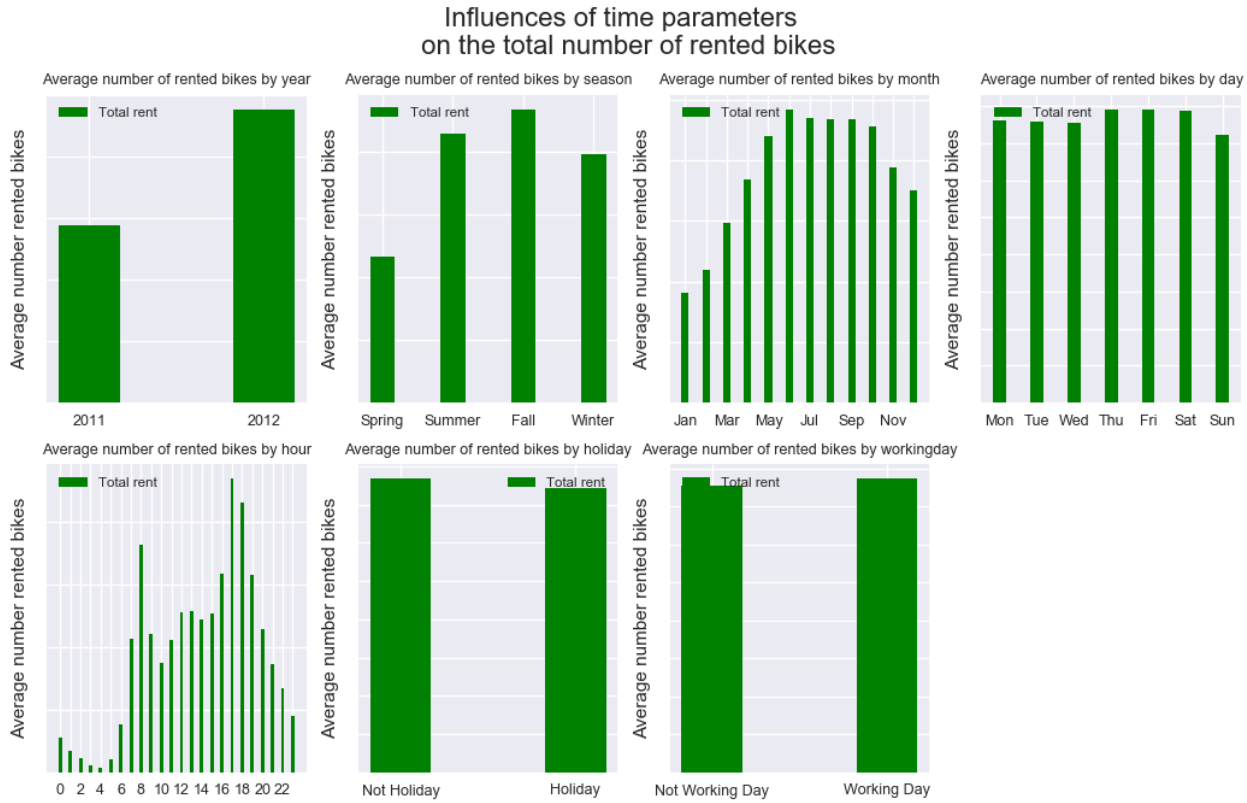
Test technique

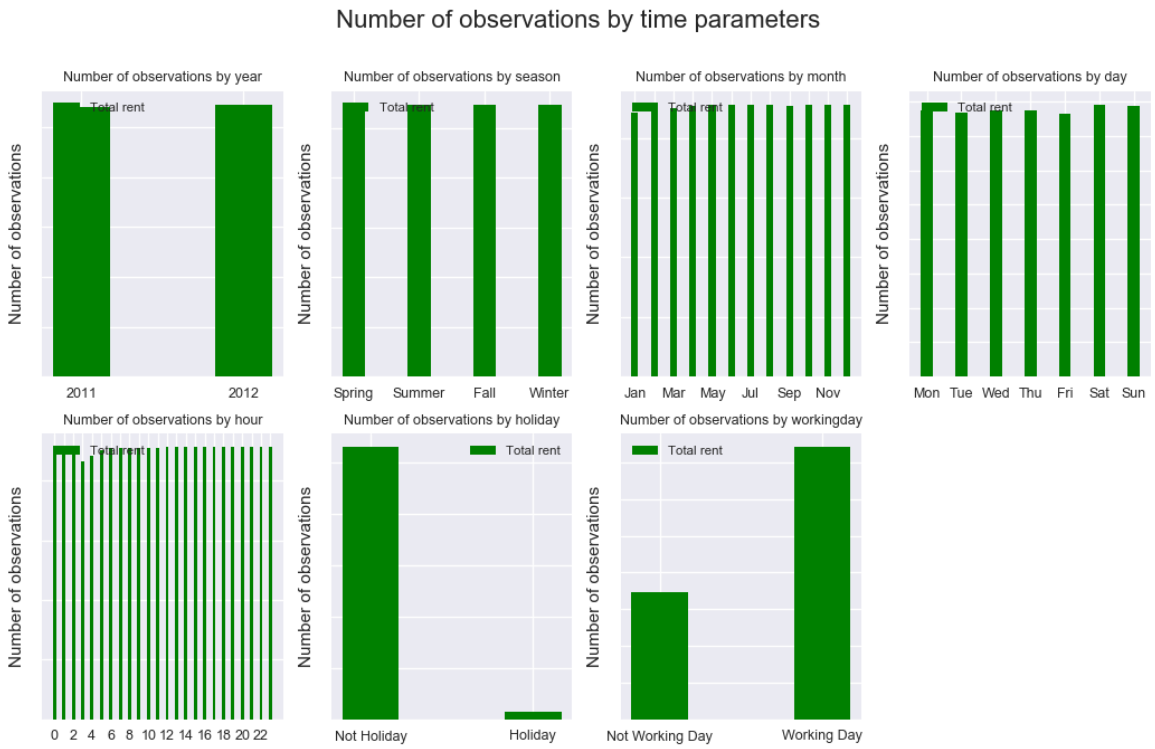
Partie I – Statistiques descriptives

De façon intuitive, on peut penser que le nombre de vélos loués dans l'heure va dépendre des données météorologiques en effet s'il fait froid et qu'il pleut, les gens vont surement préférer prendre leur voiture ou les transport en commun pour se déplacer (ou même rester chez eux).
Les données temporelles vont aussi influencer le nombre de vélos loués par heure ; durant la nuit ou en pleine journée les gens sont généralement occupés et circulent donc moins.
En tant que pré-traitement, on a extrait l'année, le mois, le jour de la semaine et l'heure de la variable date et on a créé une nouvelle variable « atempdiff » qui est la différence entre la température réelle (temp) et la température ressentie (atemp).
Pour visualiser de façon précise l'influence des différents facteurs sur le nombre de vélos loués par heure, on a tracé des graphiques qui montrent le nombre de vélos loués en moyenne en fonction des différentes valeurs que peut prendre le facteur.
Pour alléger la figure nous avons enlevé l'échelle des ordonnées cela n'est pas grave car on cherche à avoir une idée de ce qui se passe et les figures sont à l'échelle.



Cette première figure montre les paramètres temporels i.e. l'année, la saison, le mois, le jour de la semaine, l'heure, si le jour est travaillé ou pas et si c'est un jour de vacances ou non.
D'après les graphiques, de gauche à droite et de haut en bas, on peut faire les observations suivantes :

- Il y a plus de vélos loués en moyenne en 2012.
- Le nombre de vélos loués en moyenne par saison croit du printemps à l'automne et diminue en hiver.
- Le nombre de vélos loués en moyenne par mois croit de janvier à mai puis décroit.
- Le nombre de vélos loués en moyenne par jour de la semaine décroit de lundi à mercredi, atteint un pic le jeudi et décroit à partir de vendredi.
- Le nombre de vélos loués en moyenne par heure comporte deux pics, à 8h et 17h.
- Il y a légèrement moins de vélos loués en moyenne durant les vacances.
- Il y a sensiblement le même nombre de vélos loués en moyenne durant les jours travaillés et non-travaillés.



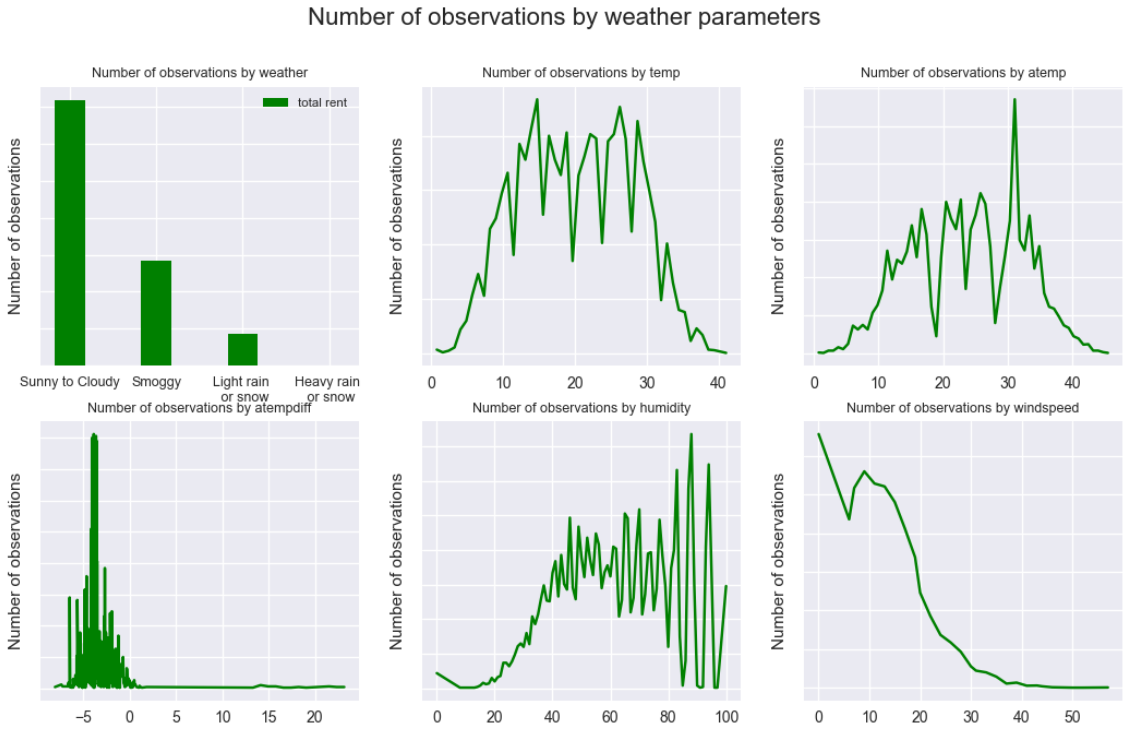
Avec cette seconde figure, on voit que l'on a le même nombre d'observations pour la plupart des paramètres temporels sauf pour les jours de vacances ou non et les jours travaillés ou non dont la différence de taille des catégories est très grande.

On peut également s'intéresser aux paramètres météorologiques i.e. la température, la température ressentie, la différence entre la température réelle et ressentie, la vitesse du vent, l'humidité et les conditions météorologiques :

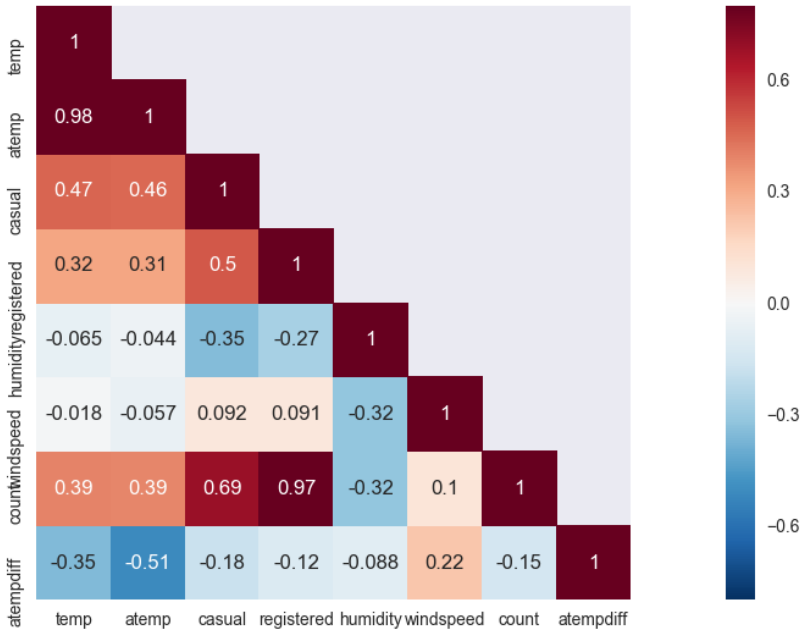
D'après les graphiques, de gauche à droite et de haut en bas, on peut faire les observations suivantes :

- En moyenne, on loue le plus de vélos quand il fait beau ou nuageux puis quand c'est brumeux puis quand il neige fort ou qui'l y a une pluie forte, et enfin on loue le moins de vélos quand il pleut légèrement ou qu'il neige légèrement.

- Le nombre de vélos loués en moyenne augmente quand la température augmente jusqu'à 30-35° puis décroit.
- Le nombre de vélos loués en moyenne augmente globalement quand la température ressentie augmente.
- Quand la différence entre température réelle et ressentie est négative i.e. la température ressentie est supérieure à la température réelle, le nombre de vélos loués en moyenne fluctue beaucoup rendant difficile sont interprétation mais on voit que plus la température réelles est supérieure à la température ressentie, plus le nombre de vélos loués augmente en moyenne.
- Le nombre de vélos loués en moyenne décroit quand l'humidité de l'air augmente à partir de 20.
- Le nombre de vélos loués en moyenne en fonction de la vitesse du vent attend grossièrement un pic pour un vent de 10-15.



En regardant le nombre d'observations par paramètres, on voit qu'en fait pour les conditions climatiques, la moyenne très élevée du nombre de vélos loués par forte pluie (ou neige) n'est pas du tout pertinente, de même que les sursauts extrêmes pour l'humidité et la vitesse du vent.



Comme pour les paramètres météorologiques, on a beaucoup de données quantitatives, on peut regarder la corrélation des paramètres quantitatifs avec le nombre de vélos loués. On observe sur la figure ci-dessus que les températures réelles et ressenties sont fortement corrélées, et que le compte total des vélos loués est le plus corrélé avec les températures puis avec l'humidité.

Si on a à notre disposition les âges et les sexes des utilisateurs abonnés, on peut effectuer les étapes suivantes pour regarder si les distributions en âge des deux populations (femme et homme) sont identiques ou non :
On est en présence de deux échantillons indépendants et on considère l'âge comme une variable quantitative, on se fixe aussi un risque alpha pour tester les hypothèses (5% ou 1% par exemple) . On peut commencer par faire un test de normalité (test de Shapiro–Wilk par exemple) sur chacune des populations femme et homme :

- Si l'une des deux population n'est pas normalement distribuée, on en déduit directement que les distributions ne peuvent pas être identiques.
- Si les deux populations sont normalement distribuées on peut continuer l'analyse en faisant un test de Fisher pour comparer les variances :
  - Si les variances sont différentes, on en déduit directement que les distributions ne peuvent pas être identiques.
  - Si les variances sont identiques, on poursuit avec un test de Student pour comparer les moyennes :
    - Si les moyennes sont identiques, les distributions sont identiques.
    - Si les moyennes sont différentes, les distributions ne sont pas identiques.
- Si les deux populations ne sont pas normalement distribuées, on peut partir sur un test non paramétrique comme le test de Wilcoxon Mann-Whitney qui permet d'indiquer si les deux populations sont distribuées de la même manière.

Partie II – Machine Learning

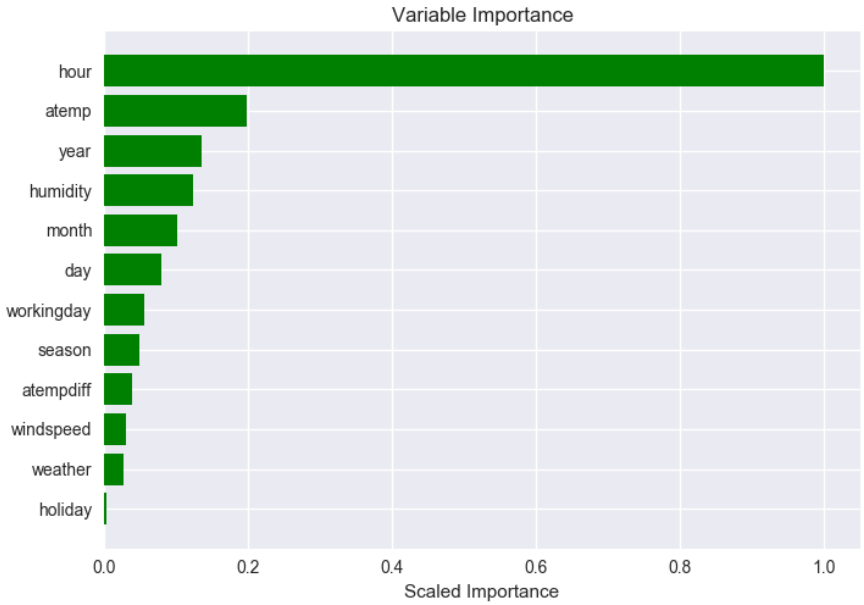
Les données semblent être de bonne qualité en effet on a pas de valeur manquante et les valeurs ne semblent pas être aberrantes.
A partir de la partie précédente, on a 14 paramètres pour chaque observation. On peut enlever le paramètre de la date puisqu'on en a tiré déjà toutes les informations qu'on voulait (année, mois, jour de la semaine, heure). Les paramètres de température réelle et ressentie sont très corrélés, on peut donc choisir d'en supprimer un des deux et le paramètre de la différence entre température réelle et température ressentie va garder les informations supplémentaires apportées par le paramètre supprimé.
On se retrouve donc avec :

- 4 paramètres numériques atemp, atempdiff, windspeed, humidity.
- 8 paramètres catégoriques weather, holiday, workingday, year, season, month, day, hour.

On cherche à prédire une valeur quantitative (le nombre de vélos loués dans l'heure), il est nécessaire de construire un modèle de régression.
Etant donné qu'il y a beaucoup de variables qualitatives, on va plutôt utiliser un modèle avec des arbres de décision. Le choix s'est porté sur un algorithme de forêt d'arbres décisionnels (random forest) car après des essais, un arbre de décision seul n'approxime pas très bien le nombre de vélos loués et cet algorithme d'ensemble a l'avantage d'être plus facile à paramétrer et est plus rapide à entrainer qu'un algorithme de gradient boosting par exemple. Un model basé sur des arbres de décision a aussi l'avantage de ne pas nécessiter de normaliser les descripteurs (cela a été essayé et la performance n'en a pas été meilleure).

L'algorithme de forêt d'arbres décisionnels a comme paramètres principaux le nombre d'arbres utilisés et le nombre de descripteurs considérés parmi tous les descripteurs disponibles pour construire un noeud d'un arbre. Ce dernier paramètre est par défaut N/3 pour la régression avec N le nombre total de descripteurs. Il reste alors à déterminer le nombre d'arbres à construire.

Pour implémenter le modèle on peut utiliser la libraire h2o, en effet elle a l'avantage de traiter les variables catégoriques tant que celles-ci ont des valeurs numérique. Alors qu'avec la librairie scikit-learn, il faudrait créer des encodeurs pour chaque variable catégorique ce qui aurait pour effet d'augmenter considérablement le nombre de descripteurs et on perd aussi le sens des catégories et donc les influences (par exemple on ne va plus regarder si c'est le printemps, l'été, l'automne ou l'hiver mais simplement si c'est ou non le printemps).
Avec la dernière version de h2o, il y a un bug qui ne permet pas d'effectuer de « grid search » sur les paramètres d'un regresseur (c'est à dire tester toutes les configurations possibles et donner la meilleure).
Le nombre d'arbres a donc été un compromis entre performance de la régression et temps de calcul.
Le critère utilisé pour juger de la performance du modèle est la métrique RMSLE (Root mean squared logarithmic error) ou la racine de la moyenne des carrés du logarithme de l'erreur. Ce critère a l'avantage de pénaliser plus fortement les points pour lesquels la prédiction est inférieure à la valeur réelle. Cela est très pertinent dans le cas de la location de vélos car il est alors mieux de construire un modèle qui prédit trop de vélos que pas assez.
Pour pouvoir comparer les changements apporté aux descripteurs ou aux paramètres du modèle, on a fixé la génération de nombre pseudo-aléatoires.
Les résultats peuvent être générés avec le code fourni et outre le critère on peut s'intéresser à l'importance des descripteurs pour le modèle. Pour la meilleur configuration, on obtient la figure suivante :



On voit ainsi que l'heure revêt une grande importance pour le modèle puis dans une moindre mesure, la température, l'année, l'humidité et le mois influent sur le modèle.

Comme amélioration possible de la modélisation on peut envisager différents axes :
- Améliorer l'algorithme de prédiction :
  - Soit trouver un moyen pour effectuer le « gris search » et ainsi paramétrer plus finement le modèle de forêt d'arbres décisionnels.
  - Ou partir sur un autre modèle de régression par exemple un algorithme de gradient boosting.
- Améliorer les données en entrée d'algorithme :
  - En effet on voit que même si les données ne semblent pas totalement aberrantes, on peut se demander par exemple pourquoi on a un nombre de vélos loués moyen très important pour un vitesse de vent de plus de 50 ou si les observations avec une humidité de 0 ont été correctement mesuré compte tenu du nombre de vélos loués moyen très faible.
  - Même si l'on ne rencontre à priori pas ce problème pour les données catégoriques, on peut s'interroger sur la façon dont ces données sont gérées surtout dans le cas de données catégoriques temporelles, en effet en les considérant comme catégoriques, on suppose que toutes les catégories sont équidistantes les unes des autres or pour des données temporelles cela n'est pas vrai. Par

exemple en catégorie, le dimanche est aussi proche du mercredi que du lundi or dans la réalité le dimanche est beaucoup plus proche du lundi donc les phénomènes qui se passent le dimanche devraient influer plus sur le lundi que le mercredi.

- Considérer la construction de deux modèles de prédiction, un pour prédire le nombre de vélos loués par les utilisateurs enregistrés et un pour prédire le nombre de vélos loués par les utilisateurs occasionnels et la valeur totale sera la somme des deux valeurs prédites par les modèles. En effet sur les figures suivantes, on voit qu'il y a globalement des influences similaires mais que pour certaines variables, le comportement des utilisateurs occasionnels et réguliers est différent comme pour le nombre moyen de vélos loués par heure. Etant donné que c'est le facteur qui influence le plus le modèle réalisé, il serait intéressant de voir ce que cela donne pour deux modèles.