Projet Python : Bike Sharing Demand

NICOLAS MEUNIER – CHARLIE MARTIN – ELIOT LANGLOIS

09/12/2021

Sommaire

I) Contexte de l’étude

Les transports à Séoul :

- **La marche** : Les distances à couvrir peuvent très vite devenir problématiques. Superficie de 605.52 km2.

- **Véhicule personnel** : Séoul se trouve saturée par des embouteillages de jour comme de nuit dû à sa population : 10 millions d'habitants et une aire urbaine de 25 millions d'habitants. De plus il y a un système de circulation alternée pour les jours de pics de pollution.

- **Transport en commun** : Intra-citée (Séoul dispose d'un réseau de bus très complet, doté de plus de 200 lignes, quadrillant toute la ville) et Métro transportant notamment 8.4 millions de touristes par an.

- **Le vélo** : (Seoul Bike) ressemble fortement aux Vélibs français.

I) Contexte de l’étude

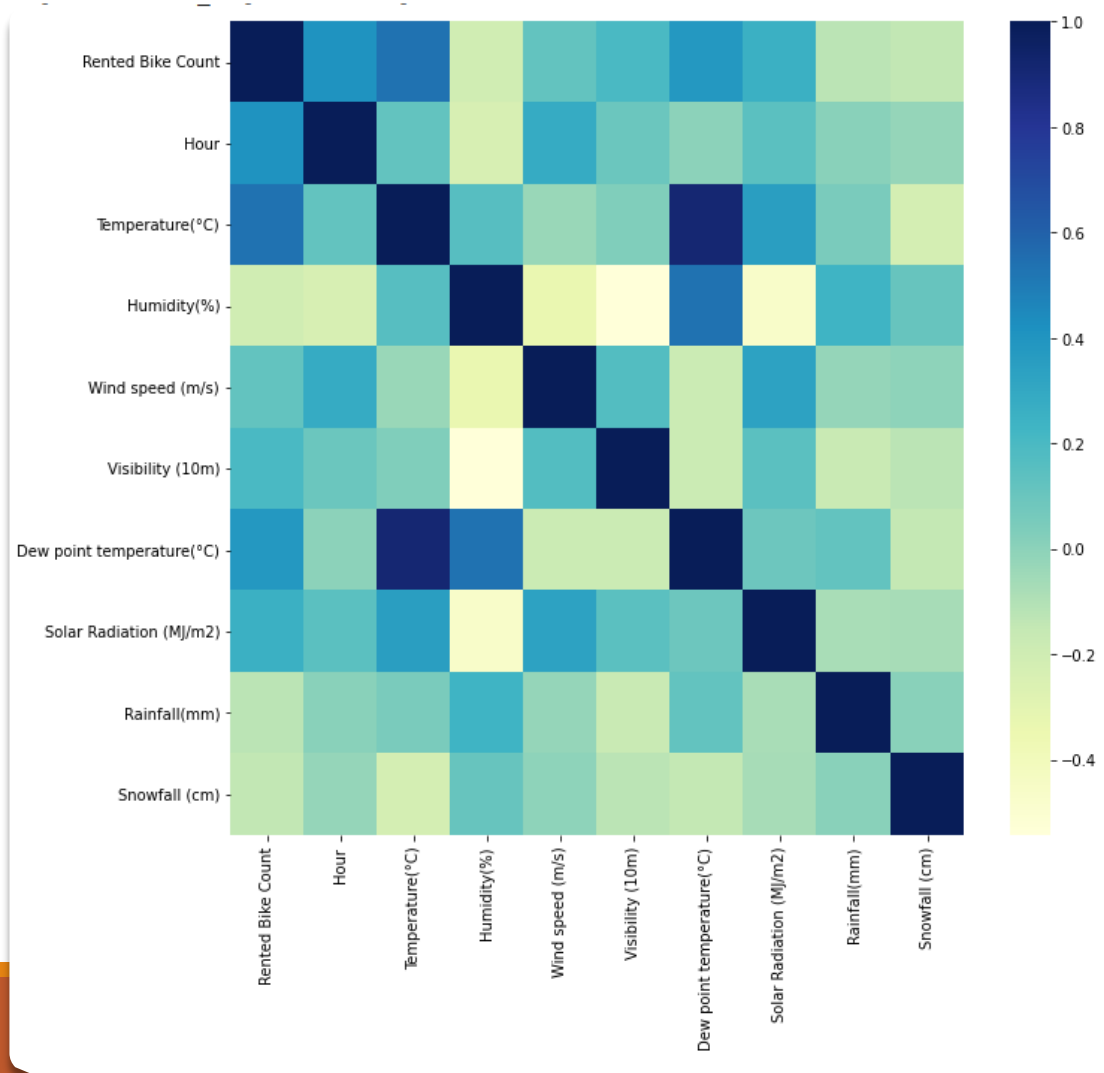
Enjeux du projet

- Prédire le nombre de vélos loués en fonction des périodes et de la météo
- Améliorer la mise à disposition des vélos
- Améliorer la planification de l’économie de l’entreprise

II) Analyse du Dataset

Maîtrise et analyse du dataset

- 8760 Lignes -> 1 année en heures
- 14 colonnes -> 14 variables
- Informations
- Tendances



II) Analyse du Dataset

- 14 variables au sein du dataset
- On peut les regrouper en 2 types autour de la variable principale : Rented by count
- Variables fixes vs variables aléatoires

Variables temporelles	Variables météorologiques
Date	Temperature
Hour	Humidity
Seasons	Wind speed
Holiday	Visibility
Fonctioning day	Dew point temperature
	Solar radiation
	Rainfall
	Snowfall

II) Analyse du Dataset – Premières valeurs

	Date	Rented Bike Count	Hour	Temperature(°C)	Humidity(%)	Wind speed (m/s)	Visibility (10m)	Dew point temperature(°C)	Solar Radiation (MJ/m2)	Rainfall(mm)	Snowfall (cm)	Seasons	Holiday	Functioning Day
0	01/12/2017	254	0	-5.2	37	2.2	2000	-17.6	0.0	0.0	0.0	Winter	No Holiday	Yes
1	01/12/2017	204	1	-5.5	38	0.8	2000	-17.6	0.0	0.0	0.0	Winter	No Holiday	Yes
2	01/12/2017	173	2	-6.0	39	1.0	2000	-17.7	0.0	0.0	0.0	Winter	No Holiday	Yes
3	01/12/2017	107	3	-6.2	40	0.9	2000	-17.6	0.0	0.0	0.0	Winter	No Holiday	Yes
4	01/12/2017	78	4	-6.0	36	2.3	2000	-18.6	0.0	0.0	0.0	Winter	No Holiday	Yes

III) Visualisations et Modélisations

Visualisations :

Temporelles (saisons, mois, jours, heures)

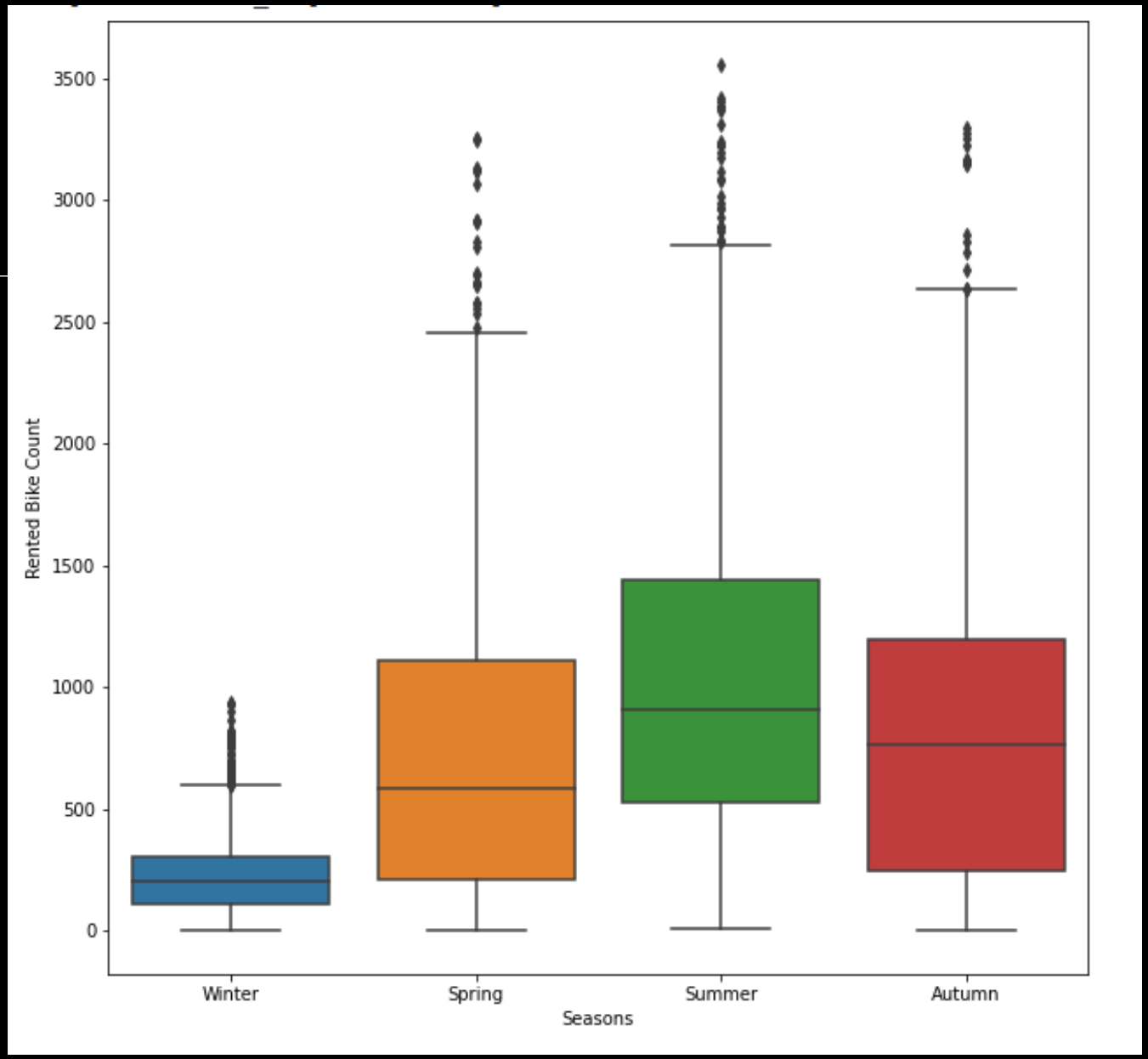
Météorologiques (températures, visibilité, vents, radiation solaire)

III) Visualisations temporelles

Boxplot des saisons :

Résultats plutôt similaire sauf pour l'hiver nettement inférieur.

Faible temperature explique bien le nombre de vélos loués



III) Visualisations temporelles

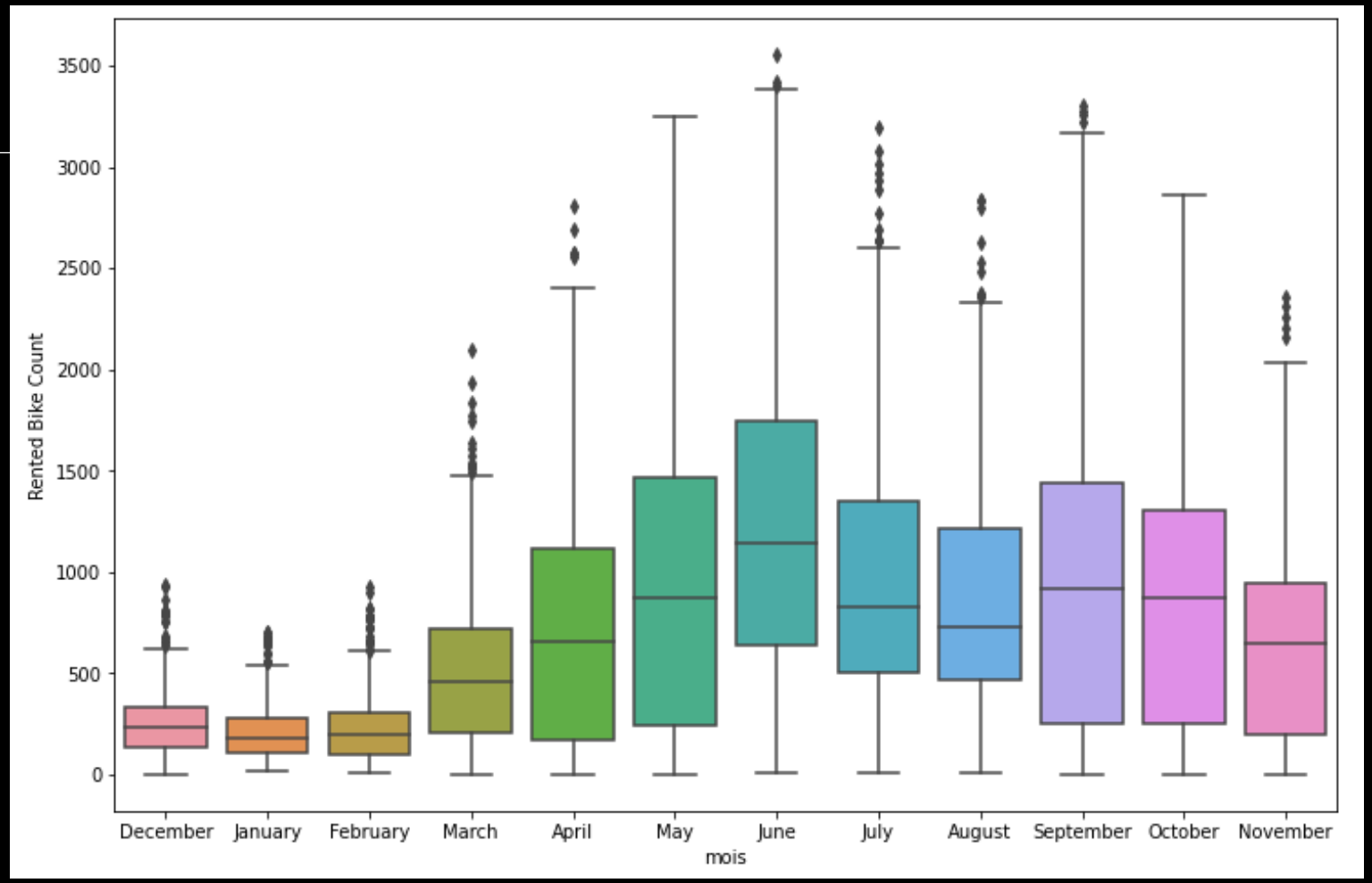
Boxplots des mois

Résultats en accord avec la visualisation des saisons

Plus de variations au sein des mois d'une même saison

Juillet et août = saison des pluies

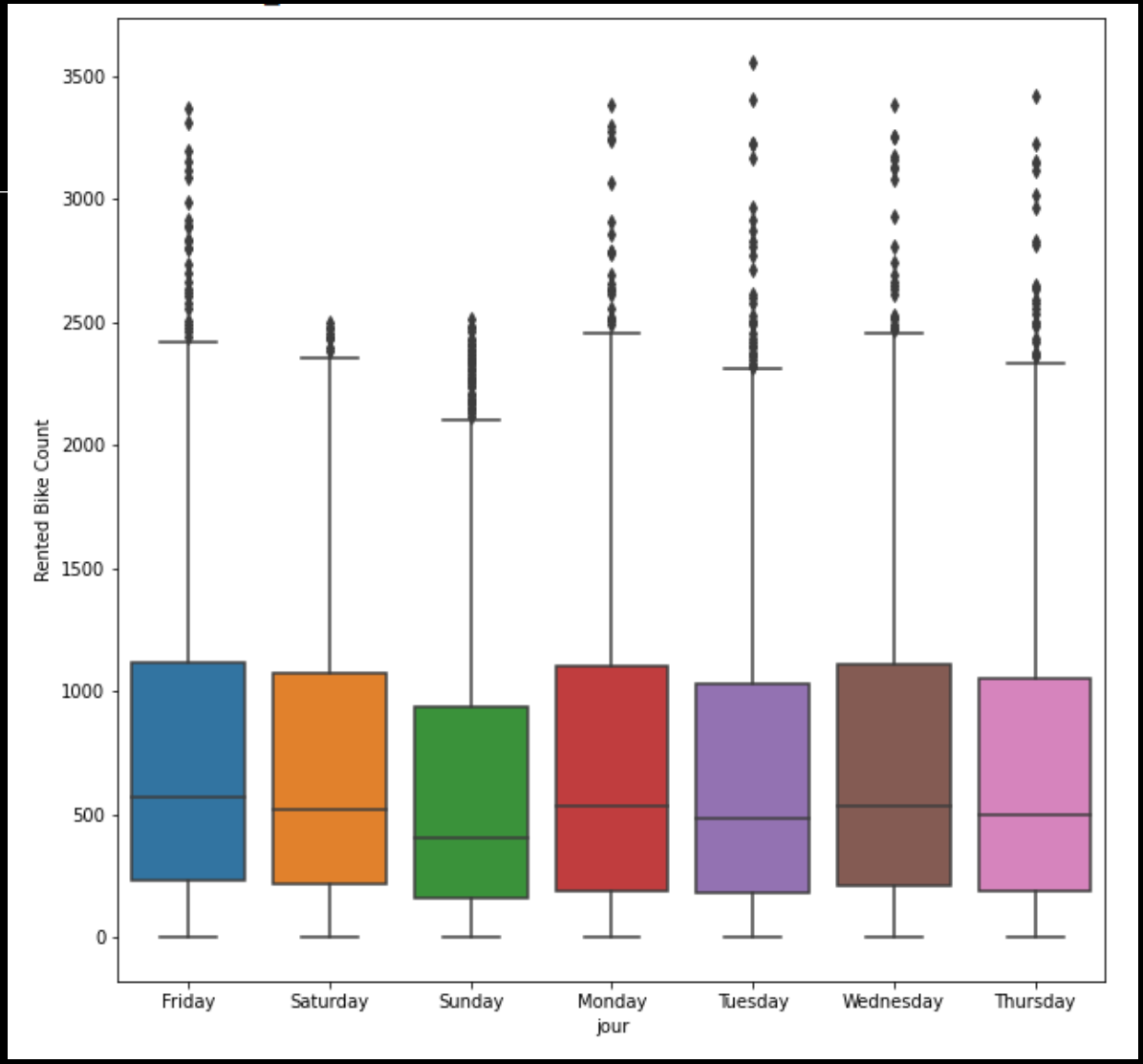
Mois ou temperature ?



III) Visualisations temporelles

Intéressant d'ajouter les jours dans nos features ?

Non : peu de variations



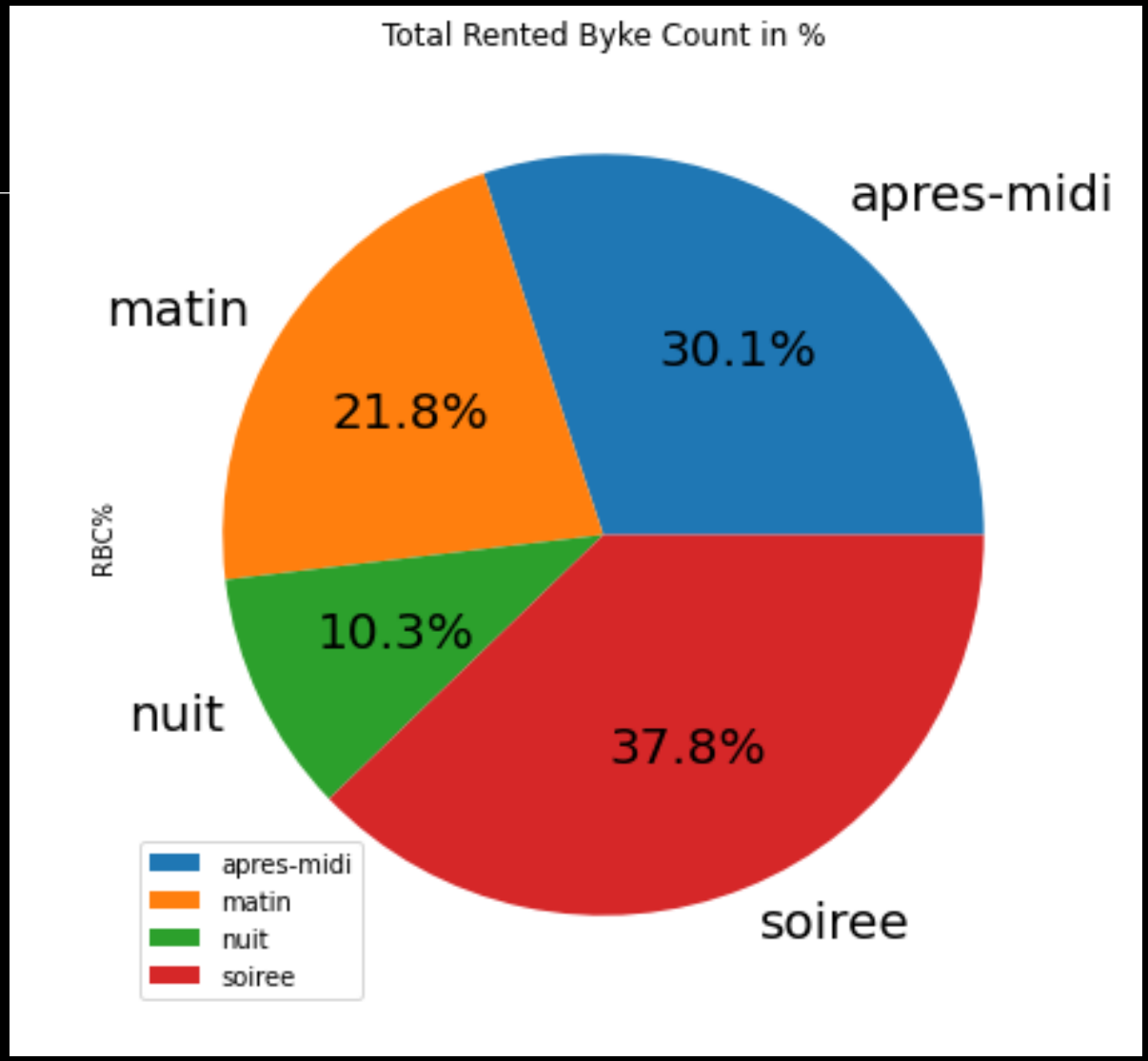
III) Visualisations temporelles

Matin (6h à 12h)

Après-midi (12h à 18h)

Soirée (18h à minuit)

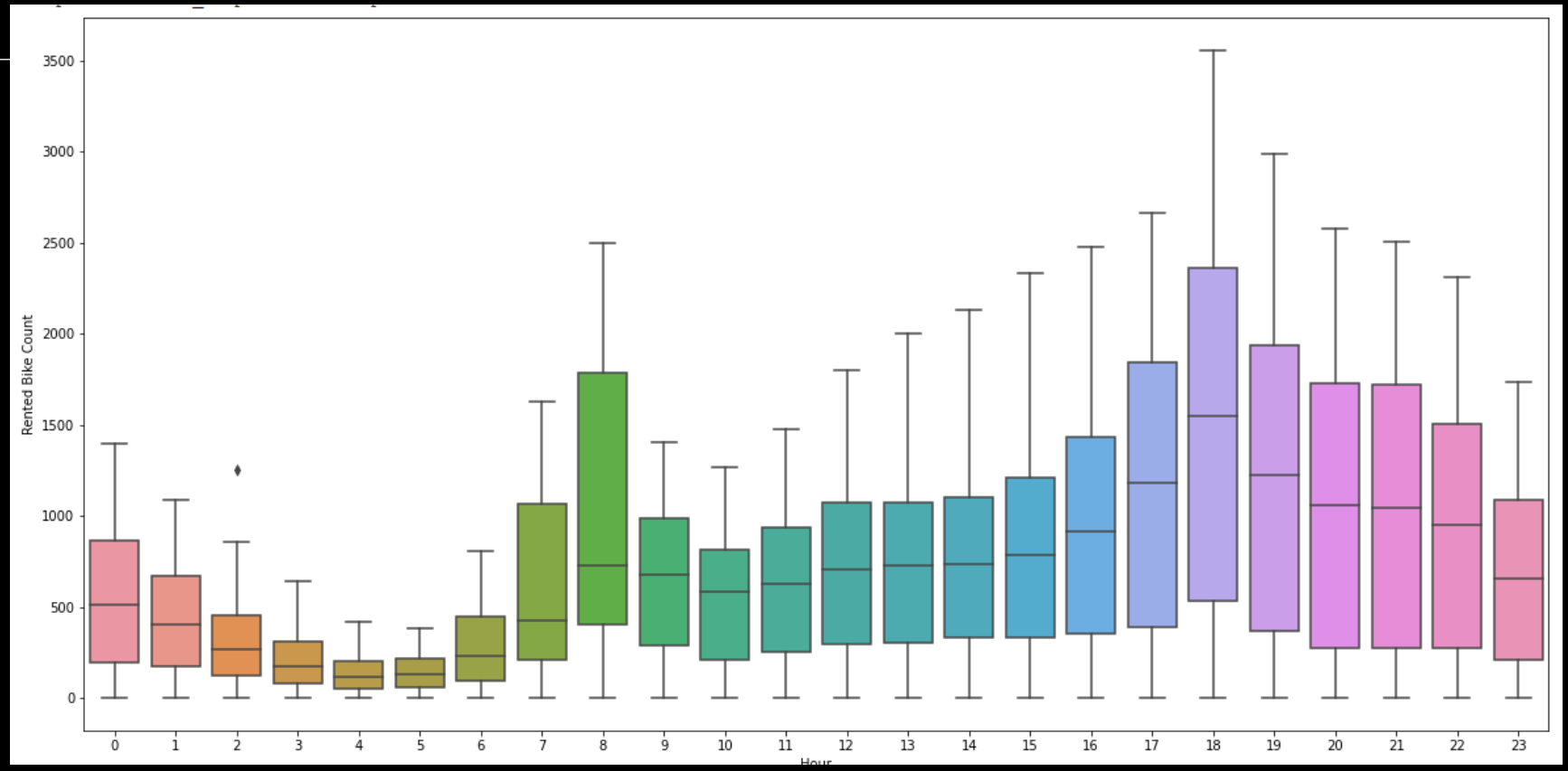
Nuit (de minuit à 6h)



III) Visualisations temporelles

Boxplots des heures :

Migrations pendulaires ?



III) Visualisations temporelles

9 heures de migration :

De 7h à 11h

De 17h à 22h

- En 9h soit 37.5% de la journée on a 50% des vélos loués
- Il existe un vrai phénomène de migration pendulaire

Migration pendulaire	Temperature(°C)	Humidity(%)	Wind speed (m/s)	Dew point temperature(°C)	Solar Radiation (MJ/m2)	Rainfall(mm)	Snowfall (cm)	Rented Bike Count	RBC%
No	12.824895	58.403470	1.677954	4.064237	0.658935	0.134977	0.076548	3109444	50.377282
Yes	12.979635	57.930898	1.803166	4.089772	0.419403	0.171537	0.072603	3062870	49.622718

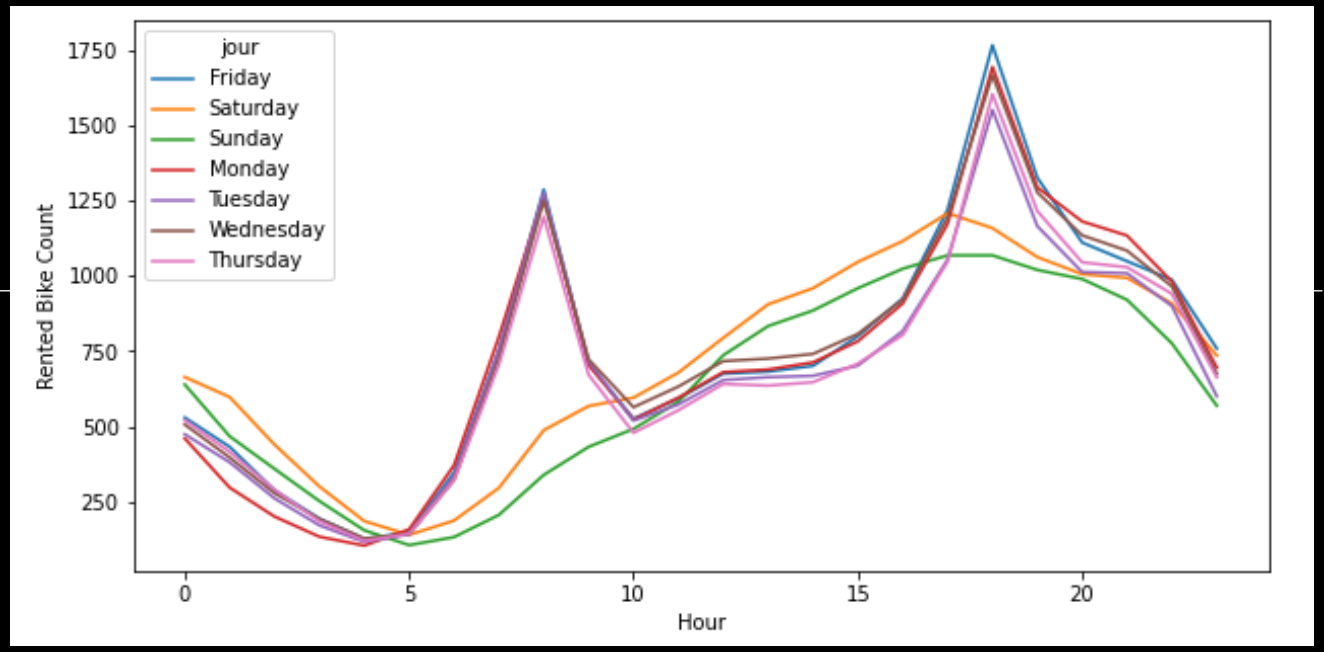
III) Visualisations temporelles

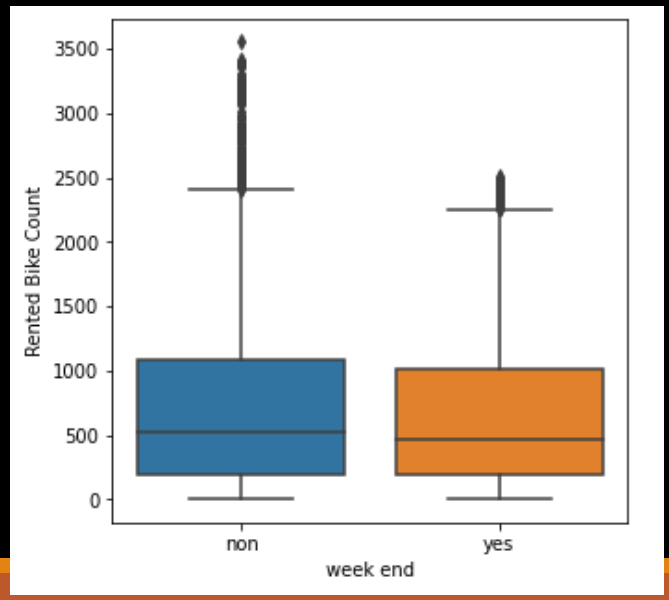
Semaine VS week-end

Moyennes similaires

Grande différence dans la répartition dans la journée

Il est important de considérer le week-end



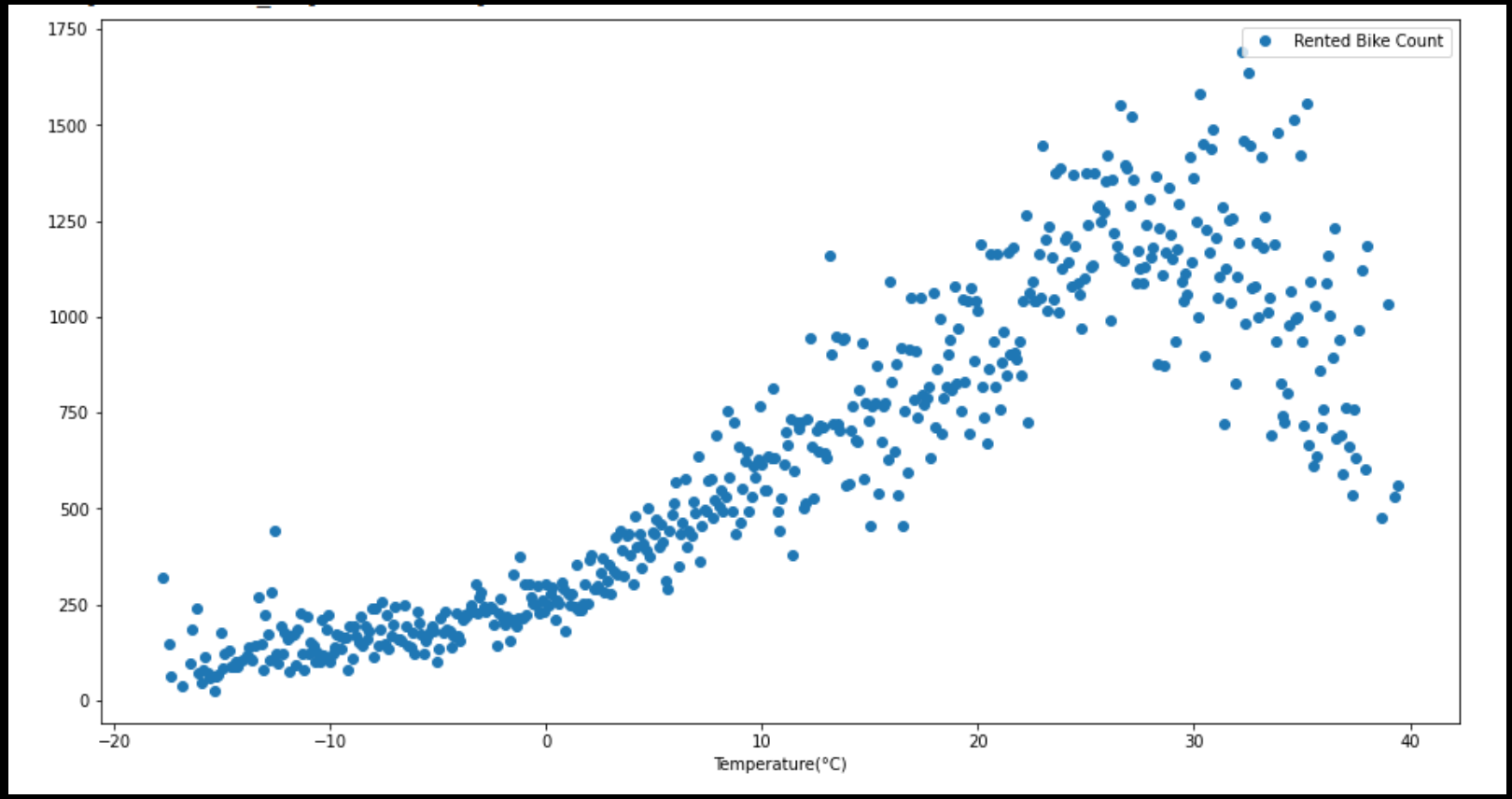


III) Visualisations météo

Température

Corrélation (temperature – Rented Bike Count) = 0.54

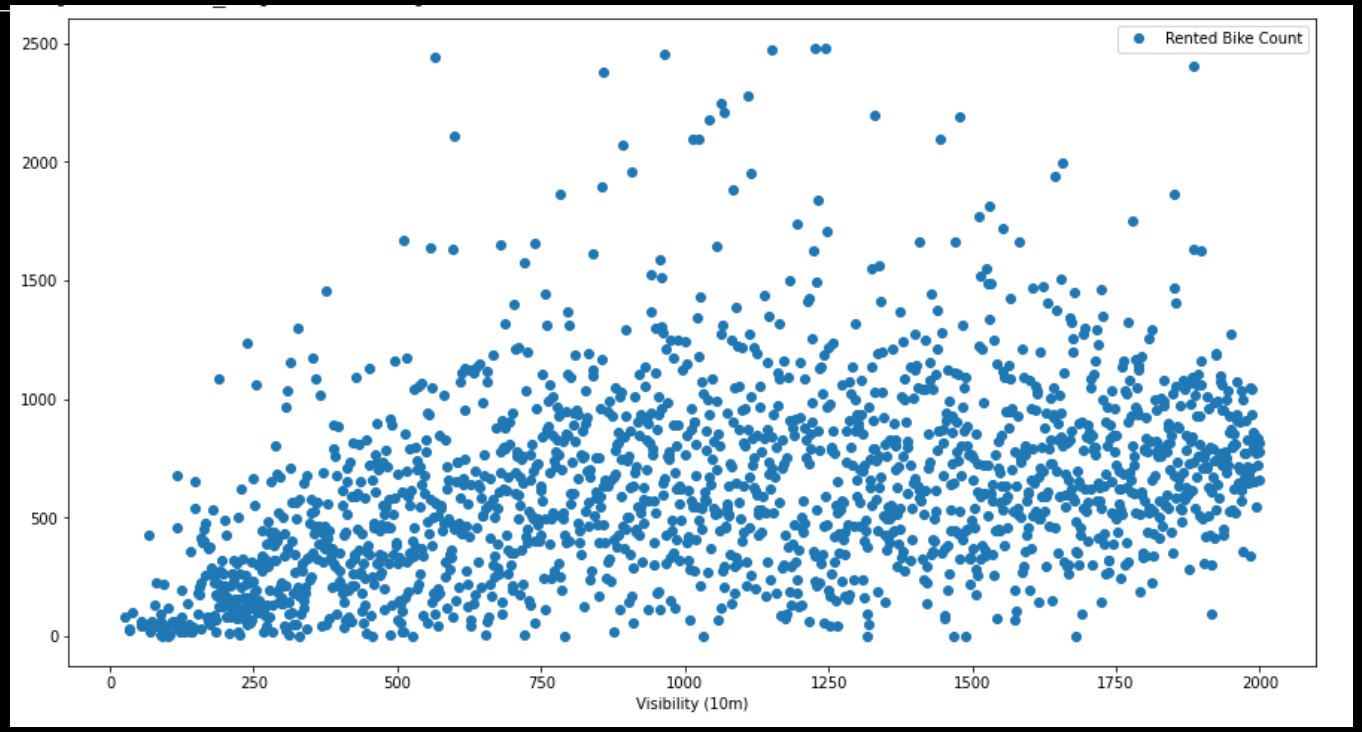
On retrouve bien les informations observées avec les saisons : la température explique plus lorsque sa valeur est faible (temps froid)



III) Visualisations météo

Visibilité

Quand la visibilité est très faible : très peu de vélos sont loués



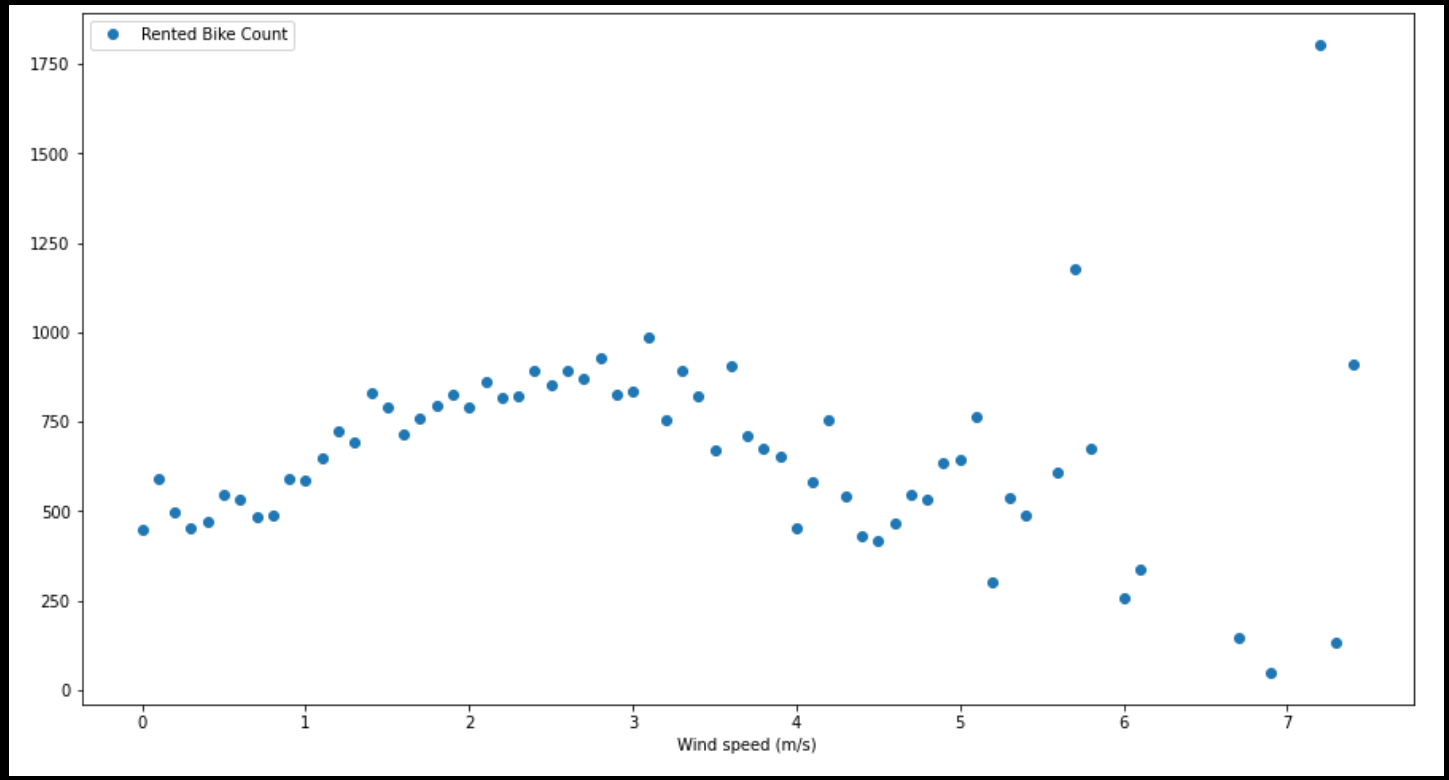
III) Visualisations météo

Vent

Beaucoup de vent = peu de vélos loués

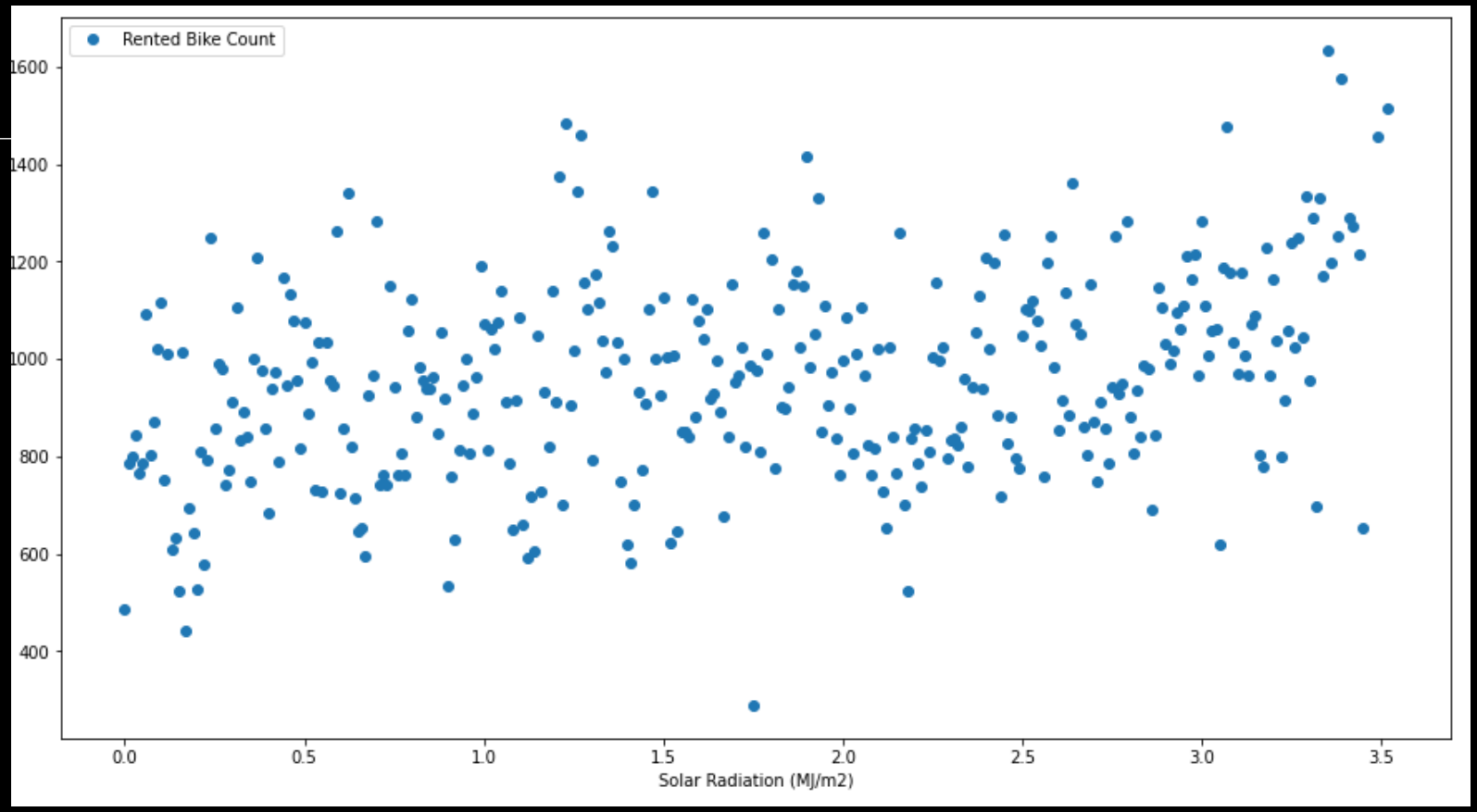
```
pre_temp2.tail(10) # 7.4 m/s | 7.2 m/s | 5.7 m/s

print(sbdata[sbdata['Wind speed (m/s)']==7.4].shape)
print(sbdata[sbdata['Wind speed (m/s)']==7.2].shape)
print(sbdata[sbdata['Wind speed (m/s)']==5.7].shape)
```

```
(1, 14)
(1, 14)
(1, 14)
```



III) Visualisations météo

Solar radiation : pas de relation linéaire



III) Visualisations et Modélisation

Bilan des visualisations :

- Température, heure et jour semblent être les features les plus importantes

- Les autres features (visibilité, vitesse du vent, radiation solaire) ont peu d'impact sauf dans les conditions extrêmes

- Jour n'est pas une feature de notre dataset :nous devons trouver un moyen de l'exprimer

III) Visualisations et Modélisation

Data set :

- Brut
- Sur une semaine
- Périodique (1h, 1j, 1semaine)
- Limiter les données corrélées
- Limiter les données inutiles
- Test de nouveaux modèles
- Optimisation des modèles

Chargement...

III) Visualisations et Modélisation

Brut :

- Toutes les colonnes : étudier brièvement la réponse du dataset à différent modèle

- Modifications pour utilisation :
  - Transformation des chaînes de caractères en valeur binaire.
  - Transformation de la date en nombre de jour d'après le calendrier grégorien proleptique.

	Date	Rented Bike Count		Holiday	Functioning Day	Autumn	Spring	Summer	Winter
0	736341	254		0	1	0	0	0	1
1	736341	204		0	1	0	0	0	1
2	736341	173	...	0	1	0	0	0	1
3	736341	107		0	1	0	0	0	1
4	736341	78		0	1	0	0	0	1

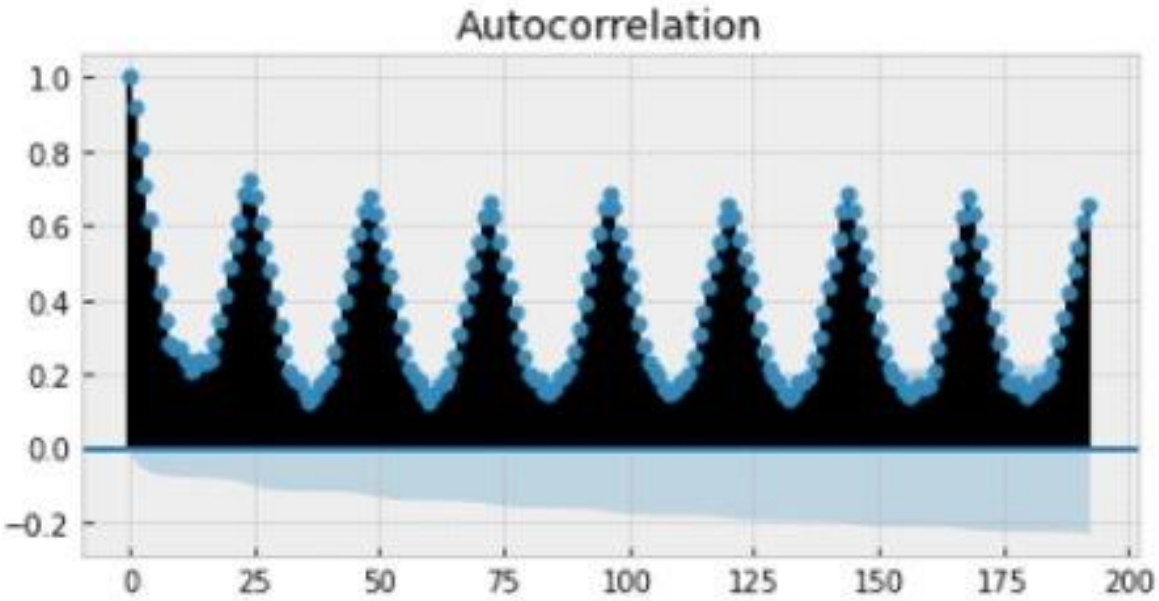
III) Visualisations et Modélisation

- Sur une semaine :
  - Début de journée à 6h
  - Lag reprenant les locations sur la dernière semaine

- Périodique (1h, 1j, 1semaine) :
  - Evolution périodique à travers le temps

  - Dégradation de la précision en augmentant la portée de la prédiction.

lag_1	lag_24	lag_168	Weekend	Hour_sin	Hour_cos	Weekday_sin	Weekday_cos
148.0	326.0	285.0	0	0.000000	1.000000	0.0	1.0
125.0	280.0	186.0	0	0.258819	0.965926	0.0	1.0
111.0	243.0	112.0	0	0.500000	0.866025	0.0	1.0
67.0	169.0	65.0	0	0.707107	0.707107	0.0	1.0
45.0	71.0	41.0	0	0.866025	0.500000	0.0	1.0

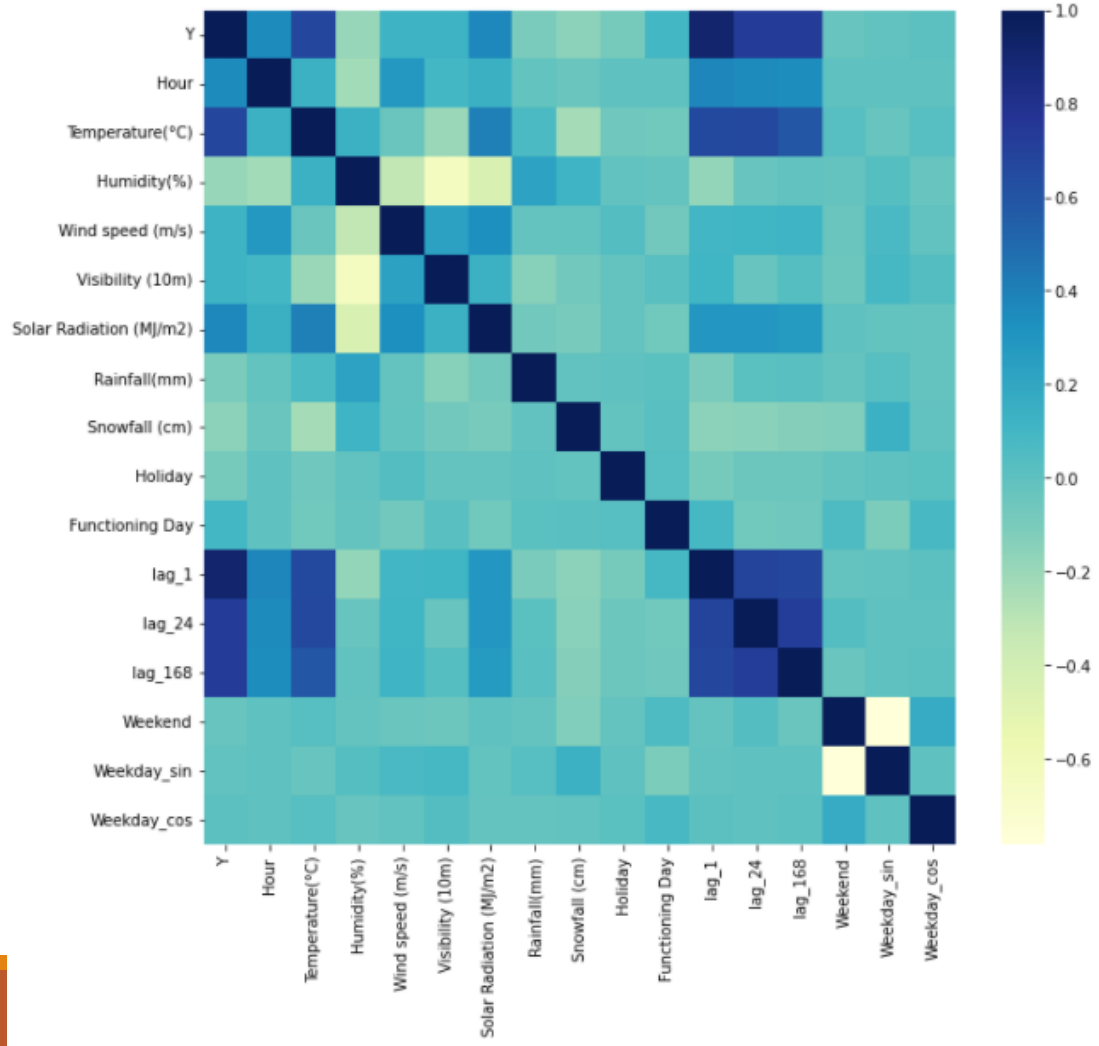


III) Visualisations et Modélisation

- Limiter les données corrélées :
  - La température et la température

  du point de rosée sont fortement corrélées.

- Limiter les données inutiles :
  - Saisons
  - Humidité
  - Visibilité



III) Visualisations et Modélisation

- Test de nouveaux modèles
  - Régression multiple (R^2 = 88%)
  - Arbre de régression (89%)
  - RandomForest (96%)
- Optimisation des modèles
  - GridSearch
  - Profondeur de l'arbre : 12 niveaux

Date	Y	Hour	Temperature(°C)	Wind speed (m/s)	Solar Radiation (MJ/m2)	Rainfall(mm)	Snowfall (cm)	Holiday	Functioning Day	TemperatureM	lag_1	lag_24	lag_168	Weekend	Hour_sin	Hour_cos	Weekday_sin	Weekday_cos
2017-12-11	125	0	-2.5	3.4	0.0	0.0	0.0	0	1	-2.540463	148.0	326.0	285.0	0	0.000000	1.000000	0.0	1.0
2017-12-11	111	1	-3.4	3.8	0.0	0.0	0.0	0	1	-2.540463	125.0	280.0	186.0	0	0.258819	0.965926	0.0	1.0
2017-12-11	67	2	-4.2	3.4	0.0	0.0	0.0	0	1	-2.540463	111.0	243.0	112.0	0	0.500000	0.866025	0.0	1.0
2017-12-11	45	3	-4.7	2.4	0.0	0.0	0.0	0	1	-2.540463	67.0	169.0	65.0	0	0.707107	0.707107	0.0	1.0
2017-12-11	44	4	-5.2	3.2	0.0	0.0	0.0	0	1	-2.540463	45.0	71.0	41.0	0	0.866025	0.500000	0.0	1.0

III) Visualisations et Modélisation

- Résultats :
  - Précision sur l'ensemble d'entrainement : 99%
  - Précision sur l'ensemble de test : 96%
  - R^2 mesure la proportion de variabilité de Y expliquée par les autres données.

$$R^2 = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}} \qquad R^2 = 1 - \frac{\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2}{\sum_{i=1}^{n} \left(Y_i - \overline{Y}_i\right)^2}$$

IV) Conclusion

Résumé des bonnes fonctionnalités :

- Utilisation cyclique des données
- Apport du lag des précédentes heures

Améliorations possibles :

- Mise à jour de l'API
- Acquisition de données de localisation
- Acquisition de données économiques

Station de vélo à Séoul