|  |
| --- |
| **MSc Digital Marketing & Data Analysis| IPAG**  **Projet : Prédiction & facteurs explicatifs des variations du trafic vélo à Paris**  *Promotion : Bootcamp Data Analyst - Juillet 2022* |
|  |

# Rapport détaillé

**Auteurs :**

Raphaël DELLA NAVE

Maïmouna FALL

Jonathan LEBLANC

# Sommaire

[**Objectifs de l’étude**](#_heading=h.i1o7wgmxqxl) **3**

[**Les sources de données**](#_heading=h.v321gl4ubll8) **4**

[**Preprocessing**](#_heading=h.wlioizeffwx6) **6**

[**Comparaison des sources de données de trafic et visualisations**](#_heading=h.td0q1e60g9y2) **11**

[**Analyse de la variable cible**](#_heading=h.kodjam1y7a0q) **15**

[**Analyse des corrélations**](#_heading=h.ncwq7kvpw5mw) **19**

[**Modélisation**](#_heading=h.dvwbl72lbom0) **23**

[**Interprétabilité**](#_heading=h.2smg9tifiybs) **31**

[**Conclusion et regards critiques**](#_heading=h.56b6afwtd19) **35**

# Objectifs de l’étude

Après l’année 2020 fortement marquée par l’épidémie de Covid-19, les habitudes de déplacements des Français ont été bouleversées. Moins de déplacements, évitement des transports en communs, plus de flexibilité au travers de l’avènement du télétravail, prise de conscience écologique… autant de facteurs qui ont contribué à faire émerger une pratique du vélo qui était déjà sur une pente ascendante en 2019.

Une tendance que l’on retrouve clairement dans le graphique ci-dessous :

**Figure 1 : Occurrence du terme de recherche vélo en île de France depuis 2004 via Google Trends**

On observe une récurrence parfaite année après année avec à chaque fois des pics de recherches sur le sujet durant les mois d’été. Puis un énorme pic en 2020 à l’issue du premier confinement et une récurrence qui semble se remettre en place par la suite, via un premier pic de recherches à l’été 2021 et un second à l’été 2022, à ceci près que les niveaux minimums après 2020 correspondent approximativement aux niveaux maximums des années antérieures. 

Il est indéniable que le paradigme qui prévalait avant 2019 a changé, et que les connaissances acquises sur le trafic cycliste sont à mettre à jour, voire à reconstruire. C’est pourquoi il nous semble important de nous intéresser aux facteurs observables qui font varier l’utilisation du vélo.

En effet, bien qu’il semble évident que les conditions météorologiques et que les déplacements en général ont un impact sur le volume de vélo en circulation, leurs importances réelles, de même que les facteurs détaillés qui les déterminent, sont peu connus.

Dans ce contexte, l'objectif principal de notre projet va être de prédire le nombre moyen de vélos en circulation sur une journée en fonction des conditions météorologiques et du trafic à Paris.

Au travers de cette étude nous aurons l’occasion :

* D’identifier les sources de données les plus performantes pour réaliser ces prédictions.
* De mesurer le poids des principaux déterminants du trafic vélo à Paris (météo, déplacements, saisonnalité…).

Nous effectuerons notre analyse sur l'année 2021 (du 1er janvier au 31 décembre) afin d'obtenir une année complète et d'isoler les effets de saisonnalité.

La ville de Paris a été choisie car les données observables y sont disponibles facilement et le trafic vélo y est le plus important en France.

# Les sources de données

Pour répondre à nos objectifs, nous disposons de plusieurs datasets disponibles en Open Data, que nous décrivons ci-dessous :

**Dataset 1 : Données des variables météorologiques (historique-meteo.net, Paris, 2021) :**

Historique Météo permet de retrouver le bulletin météo ainsi que les données (température, précipitations, vent...) jour par jour et mois par mois pour plusieurs milliers de villes en France et dans le monde. Les données météorologiques sont fournies par la société WorldWeatherOnline.

World Weather Online fournit des prévisions météorologiques mondiales et du contenu météorologique pour les sites Web, les entreprises et l'industrie du voyage et fournit les prévisions météorologiques les plus précises possibles.

Lien de téléchargement du dataset :

<https://www.historique-meteo.net/site/export.php?ville_id=188&annee=2021>

**Dataset 2 : Données de comptages vélos par capteurs permanents (opendata.paris.fr, Paris, 2021) :**

Jeux de données, publiés par les services de la ville de Paris, des comptages horaires de vélos par compteur et localisation des sites de comptage.

La Ville de Paris déploie depuis plusieurs années des compteurs à vélo permanents pour évaluer le développement de la pratique cycliste.

Les compteurs sont situés sur des pistes cyclables et dans certains couloirs de bus ouverts aux vélos. Les autres véhicules (ex : trottinettes…) ne sont pas comptés.

Ce jeu de données est chargé quotidiennement sur l'API par leur partenaire Eco Compteur.

Lien de téléchargement du dataset :

<https://opendata.paris.fr/explore/dataset/comptage-velo-donnees-compteurs/>

**Dataset 3 : Données de comptage routier par capteurs permanents (opendata.paris.fr, Paris, 2021) :**

Jeux de données, publiés par les services de la ville de Paris, des comptages routiers par compteur et localisation des sites de comptage.

Sur le réseau parisien, la mesure du trafic s’effectue majoritairement par le biais de boucles électromagnétiques implantées dans la chaussée.

Les principaux axes de la Ville de Paris sont équipés de stations de comptage des véhicules et de mesure du taux d’occupation, à des fins à la fois de régulation du trafic et des transports en commun, d’information aux usagers, et d’étude.

Ces données donnent un aperçu du taux d'occupation et du débit sur plus de 3000 tronçons de voies. Le débit est le nombre de véhicules ayant passé le point de comptage pendant un intervalle de temps fixe (une heure pour les données fournies).

Lien de téléchargement du dataset :

<https://opendata.paris.fr/explore/dataset/comptages-routiers-permanents>

**Dataset 4 : Données de comptages multimodaux par capteurs infrarouges & intelligence artificielle (opendata.paris.fr, Paris, 2021) :**

La Ville de Paris collecte des données de comptage des usagers par mode de déplacement, par file de circulation, par sens et par site de comptage.

Ces données sont élaborées par un algorithme d’intelligence artificielle qui analyse les images issues de caméras thermiques installées sur l’espace public.

Les véhicules mis en évidence par ces capteurs sont les :

* Trottinettes, Trottinettes + Vélos (lorsque la distinction entre ces 2 modes de déplacement n’est pas implémenté sur le capteur)
* Vélos
* 2 roues motorisées
* Véhicules légers < 3,5 tonnes
* Véhicules lourds > 3,5 tonnes
* Autobus & autocars

Lien de téléchargement du dataset :

<https://opendata.paris.fr/explore/dataset/comptage-multimodal-comptages/information/?disjunctive.label&disjunctive.mode&disjunctive.voie&disjunctive.sens&disjunctive.trajectoire>

**Dataset 5 : Données de validation des transports en commun (IDF mobilités, Paris, 2021)**

Ces jeux de données, publiés par Ile-de-France Mobilités, présentent le nombre de validations des voyageurs par jour, par ligne et par titre de transport sur le réseau de surface (bus, tram) et le réseau ferré (RER, métro, Transilien) en Ile-de-France.

Ainsi, à chaque fois qu’un usager valide son passe Navigo sur le réseau francilien, une information anonyme est remontée à un système central permettant la collecte et le calcul de statistiques sur la mobilité.

A noter que les données de validations ne comptabilisent pas :

* Les tickets magnétiques (par exemple, les ticket T+, Forfait Mobilis, Forfait Paris Visite, etc)
* Les usagers qui ne valident pas à la montée dans un bus ou tram, ou en entrant dans une gare (par exemple si le quai est accessible sans passer par une barrière de contrôle, ou si les barrières de contrôle sont indisponibles en raison de travaux, etc
* Les fraudeurs …
* Et donnent par conséquent une vision incomplète du trafic sur le réseau de transport.

Lien de téléchargement des datasets :

<https://prim.iledefrance-mobilites.fr/fr/catalogue-data>

# Preprocessing

Afin de pouvoir commencer nos analyses et entraîner nos modèles de prédiction, nous avons dû traiter et compiler toutes les différentes sources de données dans un même dataframe, sous une base commune. Nous avons donc regroupé toutes nos variables sous un format journalier, afin de pouvoir étudier les variations au cours de l’année.

Voici les étapes que nous avons mené pour chaque dataset :

**Données de trafic routier via compteurs classiques :**

Pour pouvoir compiler les données de trafic par compteur classique, nous avons compilé les résultats de dizaines de millions de mesures horaires sur l'année dispatchées en 58 fichiers correspondants à peu près aux semaines de l'année.

Nous avons regroupé les données par jour. Pour cela, nous avons implémenté une boucle pour traiter tous les fichiers à la suite afin de tout compiler dans un seul dataframe. Cette boucle nous permet d’obtenir pour chaque jour de l’année :

* La somme totale du trafic routier sur une journée (somme des débits horaires par capteurs et par jour)
* Le nombre total de mesures effectuées par tous les capteurs sur une journée (une unité de compte = une heure pour un capteur)
* La moyenne du trafic par heure et par capteur (somme totale du trafic divisé par le nombre d'unités de compte sur une journée).

**Données de trafic routier et vélo via capteurs infrarouges :**

Nous passons ensuite aux données de trafic infrarouge, que nous avons pu exporter en format .csv sur l'année 2021 au complet.

Afin de bien différencier les modes de transports motorisés et les mobilités douces (vélos, trottinettes), nous les avons regroupés sous deux variables ("Motorisé" et "vélo").

Nous avons donc pu, pour chaque mode, calculer :

* La somme totale du trafic sur une journée
* Le nombre total de mesures effectuées par tous les capteurs sur une journée (une unité de compte = une heure pour un capteur)
* La moyenne du trafic par heure et par capteur (somme totale divisée par le nombre d’unité de compte sur une journée)

**Données de trafic vélo via compteurs classiques :**

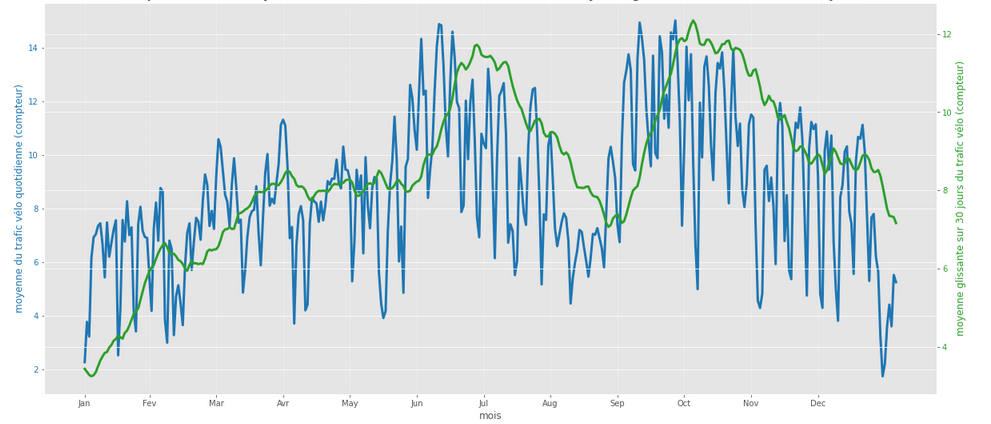
Concernant les données issues des capteurs vélo, nous avons pu exporter les données directement sur l'année 2021 au complet.

Comme pour le comptage multimodal et les données de trafic via compteurs classiques, nous avons créé trois unités de mesures :

* La somme totale du trafic à vélo sur une journée
* Le nombre total de mesures effectuées par tous les capteurs sur une journée (une unité de compte = un quart d’heure pour un capteur)
* La moyenne du trafic par quart d’heure et par capteur (somme totale divisée par le nombre d’unité de compte sur une journée)

Pour cette source de données uniquement (nous verrons plus loin pourquoi), nous calculons pour chaque jour une moyenne glissante des 30 derniers jours de la moyenne du trafic par quart d’heure, dont nous retrouvons une représentation graphique ci-dessous.

**Figure 3 : Comparaison de la moyenne glissante sur 30 jours du du trafic vélo quotidien au regard de la distribution de la moyenne du trafic vélo par jour (données de compteur)**



**Données de validation dans les transports en communs :**

Concernant les données de validation des transports en commun, il nous a fallu :

* Pour le réseau de surface, concaténer les 4 datasets à notre disposition (un par trimestre) sur l'axe 0 afin d'obtenir une année complète.
* Pour le réseau ferré, concaténer les 2 datasets à notre disposition (un par semestre) de la même façon.

Les données de validation n’ayant pas à proprement parlé d’unité de comptage puisque l’ensemble des points de passage sont relevés, les seules données récupérées sont :

* Le total quotidien des validations sur le réseau ferré
* Le total quotidien des validations sur le réseau de surface
* La somme quotidienne des validations sur le réseau ferré et de surface, que nous avons calculé

**Données météorologiques :**

Concernant les données météos, c’est l’ensemble du dataframe qui a été importé directement sans prétraitement spécifique. Les variables intégrées sont :

* max\_temperature\_c = température maximale en degré Celsius
* min\_temperature\_c = température minimale en degré Celsius
* windspeed\_max\_kmh = vitesse maximale en km/h
* temperature\_morning\_c = température du matin en degré Celsius
* temperature\_noon\_c = température à midi en degré Celsius
* temperature\_evening\_c = température en soirée en degré Celsius
* temperature\_night\_c = température nocturne en degré Celsius
* precip\_total\_day\_mm = total de précipitations journalières en mm
* humidity\_max\_percent = pourcentage d'humidité en %
* visibility\_avg\_km = visibilité moyenne en km
* pressure\_max\_mb = pression maximale en mb (millibar)
* loudcover\_avg\_percent = couverture nuageuse moyenne en %
* heatindex\_max\_c = index de chaleur max en degré Celsius
* dewpoint\_max\_c = point de rosée max en degré Celsius
* windtemp\_max\_c = température du vent max en degré Celsius
* weather\_code\_morning = code météo matin
* weather\_code\_noon = code météo midi
* weather\_code\_evening = code météo soirée
* total\_snow\_mm = neige totale en mm
* uv\_index = index uv
* sunhour = heure d'ensoleillement
* opinion = une variable catégorielle qui indique la qualité de la météo pour une journée (idéale, favorable, correcte, défavorable, très défavorable) -

N.B : la façon dont la variable est catégorisée n’est pas définit par historique-meteo.net

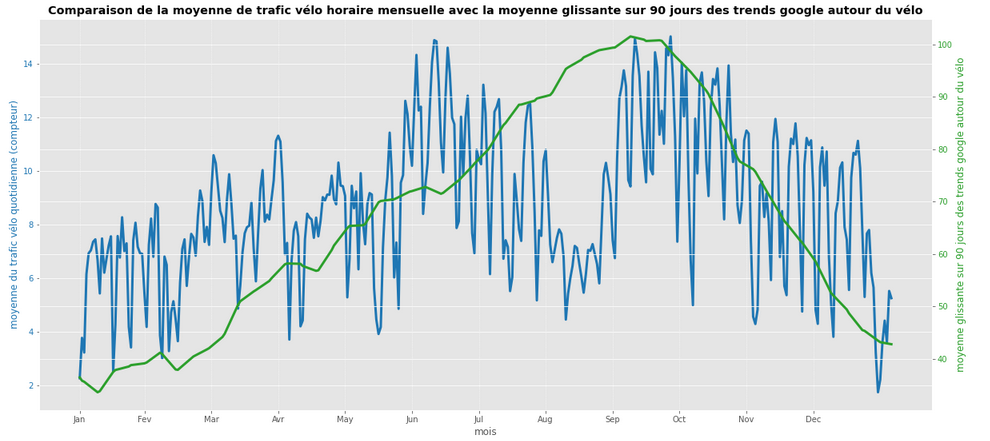
* sunset = heure du coucher de soleil
* sunrise = heure du lever de soleil

**Données de Google Trends :**

Pour ces données, nous avons dans un premier temps compilé l’occurrence quotidienne des mots clés liés au vélo susceptibles d’être fortement utilisés en Ile-de-France (« achat vélo », « location vélo », « vélo en libre service », « aide vélo », « vélib ») pour l’année 2021 sur une base = 100 commune.

Nous avons sommé ces occurrences pour obtenir un indicateur synthétique des trends vélo sur une journée donnée. Les trends d’une journée - très volatiles - ne prédisent en général pas directement la pratique du vélo pour la journée en cours. C’est la raison pour laquelle, nous avons préféré calculer une moyenne glissante sur 90 jours de ces occurrences de recherche afin d’attribuer à chaque jour un indice de trends périodique sur la thématique du vélo.

**Figure 3 : Représentation de la moyenne glissante des trends Google par jour au regard de la distribution de la moyenne du trafic vélo par jour (données de compteur) :**

**Données calendaires :**

A partir de la variable ‘day’ (numéro du jour dans l’année de 1 à 365), qui nous a permis de merger tous nos dataframes, nous créons plusieurs autres variables temporelles :

* ‘date’ = La date du jour
* ‘weekday’ = Le jour de la semaine
* ‘isweekend’ = boolean True = samedi ou dimanche et False = reste de la semaine
* ‘mois’ = renvoie le numéro des mois
* ‘nom\_mois’ = renvoie le nom du mois
* ‘numsem’ = renvoie le numéro de la semaine
* ‘typejour’ = une variable catégorielle qui divise les jours de l’année en : jours de vacances scolaires, jours fériés (hors vacances), weekend (hors vacances) et jours de la semaine (hors vacances). A noter que cette variable n’a pas été créée via une formule mais fusionnée au dataframe sous excel après avoir été codée à la main.

**La part du vélo dans le trafic :**

La part du vélo dans le trafic exprime le ratio entre la moyenne horaire du trafic vélo / la moyenne horaire de l’ensemble des trafics (vélo + transports + routier). Cet indicateur a l’avantage de ne pas être tributaire d’effets de variations lorsque le trafic en général fluctue (exemple : vacances scolaires versus. hors vacances scolaires) à l’inverse de la moyenne.

Nous avons donc calculé la part du vélo dans le trafic :

* Issue des données de compteurs classiques : moyenne horaire du vélo par jour et par capteur / (moyenne horaire du trafic routier via compteur classique + moyenne horaire du trafic vélo via compteur classique + moyenne horaire des données de validations sur l’ensemble du réseau).
* Issue des données de compteurs infrarouges : moyenne horaire du vélo par jour et par capteur / (moyenne horaire du trafic routier via compteur infrarouge + moyenne horaire du trafic vélo via compteur infrarouge + moyenne horaire des données de validations sur l’ensemble du réseau).

**Analyse du dataframe final :**

Finalement, notre dataframe est constitué de 66 variables pour 365 lignes (une par jour de l’année) sans aucune valeur manquante, ni doublon. Il est formé d’une majorité de variables numériques (45), une minorité de variables catégorielles (7) et une seule variable booléenne (1).

# Comparaison des sources de données de trafic et visualisations

A ce stade, nous pouvons connaître la fréquentation vélo à Paris grâce à 2 types de données :

* Les données multimodales issues des capteurs infrarouges
* Les données de trafic routier et vélo issues des capteurs classiques

Il nous faut donc faire un choix entre ces données pour implémenter notre modèle.

L'avantage des données infrarouges est qu'elle utilise une technologie commune pour tous les types de véhicules en circulation en exploitant un algorithme d’intelligence artificielle qui analyse les images issues de caméras thermiques installées sur l’espace public.

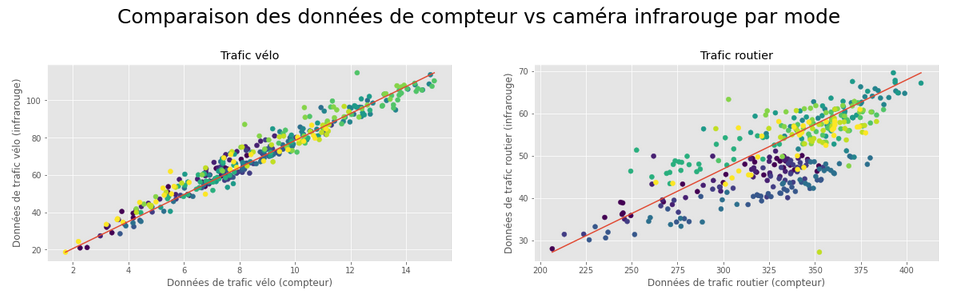
En revanche, le nombre de capteurs infrarouge et leur capacité à distinguer le type de véhicules (ex : trottinettes et vélos) est variable au fil du temps (notamment en 2021 pour ce qui nous intéresse), en raison de l'implantation progressive de technologies plus avancées par la ville de Paris.

Les données issues des capteurs permanents, quant à elles, nous permettent d'isoler plus précisément les vélos (car il s'agit de 2 types de capteurs différents pour les vélos et les autres véhicules). Aussi, elles prennent en compte un nombre de compteurs stable tout au long de l'année 2021, ce qui nous permet de supprimer un biais lié à la localisation des bornes de comptage dans Paris (notamment si elles correspondent à des tronçons atypiques avec très peu ou énormément de trafic) et d'obtenir des résultats plus cohérents au fil des mois (car le nombre de comptage par jour reste le même).

Pour mieux comprendre ces données, nous procédons à des visualisations.

Les nuages de point (ou scatter plots) vont nous permettre de visualiser la relation entre deux variables continues.

Dans un premier temps nous comparons la relation entre les données de compteur (en abscisse x) et données infrarouge (en ordonnée y), pour le trafic routier et le trafic vélo.

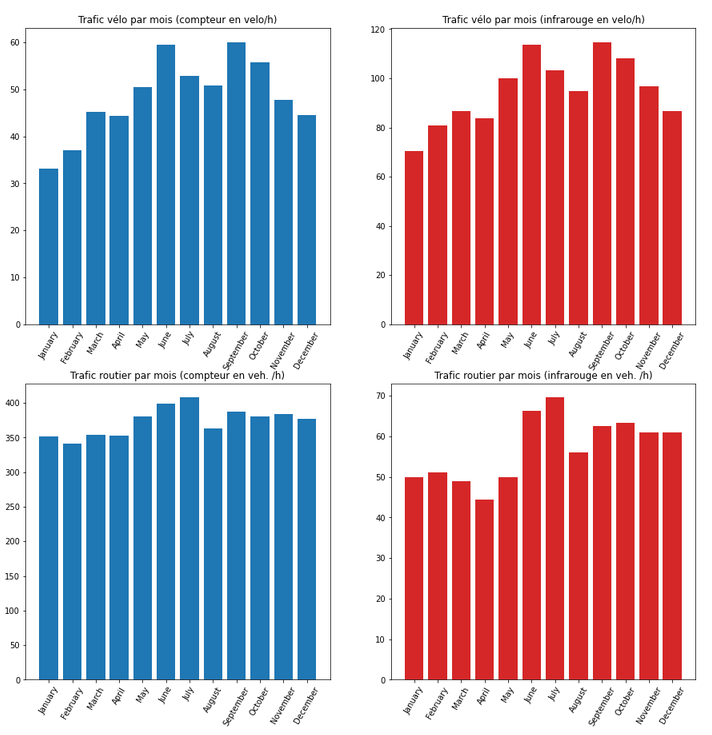
**Figure 4 : Nuages de points représentant l’alignement des variables de trafic entre les deux sources (classiques vs infrarouges) par mode (une couleur = un mois de l'année) :**

Pour les données de vélo, qui sont celles qui nous intéressent principalement, les données sont très proches et s'alignent approximativement le long de la droite d’ajustement de régression x = y. La relation linéaire entre ces deux variables est donc forte. De plus, il y a peu d’outliers et ceux présents ne correspondent pas à une temporalité particulière.

En revanche, sur le trafic routier cette première visualisation nous montre de fortes disparités entre les deux sources de données, avec une répartition des biais répartis de manière non aléatoire (des temporalités entières sont sous-estimées et d'autres sur-estimées).

Pour approfondir le sujet, nous décidons ensuite de comparer le trafic routier et vélo moyen par mois, selon les capteurs classiques et infrarouges.

**Figure 5 : Comparaison des moyennes horaire mensuelles par capteur selon le mode et la source de données :**

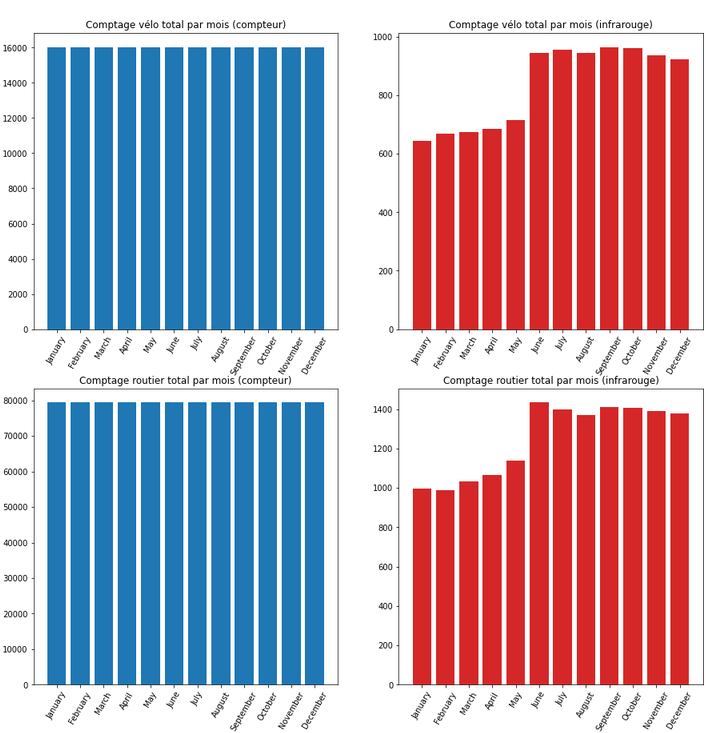


L'analyse des deux premiers graphiques en bâtons qui précèdent nous permet de remarquer que le trafic vélo par mois est relativement identique, en termes de proportion de comptage, entre les compteurs classiques et infrarouge. On observe en revanche que l'ordre de grandeur n'est pas le même entre les deux sources de données, les capteurs infrarouges détectent environ deux fois plus de vélos par heure que les compteurs classiques.

Concernant les deux graphiques suivants, comme vu précédemment sur les données de trafic routier, nous observons des écarts importants en valeurs relatives selon les modes de comptage, puisque les premiers mois de l'année semblent largement sous-estimés pour les données infrarouges. En valeur absolue, nous observons également un fort écart de grandeur entre les données de compteurs et les données infrarouges.

Ainsi, comme nous allons le voir dans la visualisation suivante, la principale explication de cet écart est l'évolution du nombre de capteurs infrarouges au cours de l'année 2021, alors que le nombre de capteurs classiques, lui, est resté parfaitement stable.

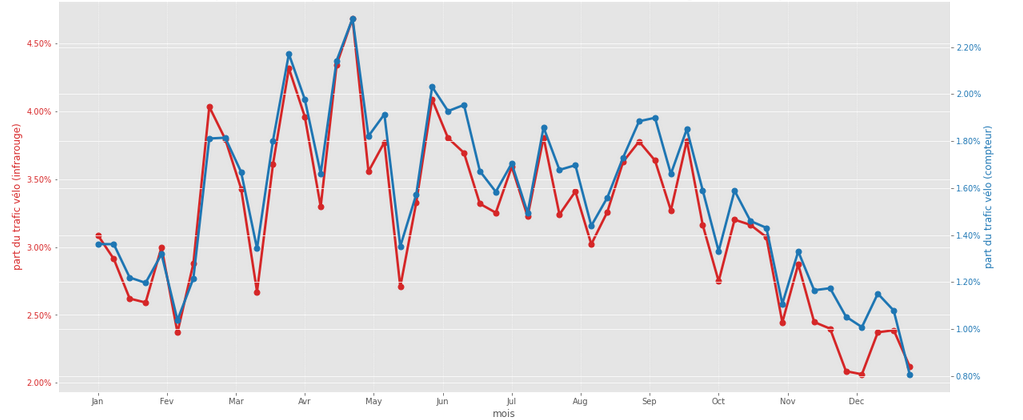
**Figure 6 : Comparaison du nombre d’unités de comptage mensuel selon le mode et la source de données :**



Il y a en effet eu de nombreuses installations de capteurs infrarouges entre mars et juin 2021 par la ville de Paris, ce qui fragilise la cohérence des données au cours de l'année.

Par ailleurs, pour vérifier s'il existe une différence importante entre les deux types de comptage pour expliquer la part du vélo dans l'ensemble des déplacements (routier, transports en communs et vélo) à Paris, nous avons comparé la part du vélo mesurée par capteurs infrarouges et la part du vélo mesurée par les capteurs classiques :

**Figure 7 : Comparaison de la part trafic vélo hebdomadaire issue des données infrarouge vs des compteurs :**



Malgré les disparités observées précédemment (notamment sur les unités de comptage), les deux courbes sont très proches (même sur le début d’année), ce qui ne nous permet pas encore d'écarter les données infrarouges de nos modélisations.

Mais, étant donné l'évolution importante du nombre de compteurs infrarouges entre mars et juin, nous allons nous concentrer uniquement sur les données issues des capteurs classiques pour nos prochaines visualisations. Au vue de la forte cohérence des données sur le vélo, il est fort probable que les tendances observées sur une source se retrouveraient de toute manière par ailleurs sur l'autre.

# Analyse de la variable cible

Nous cherchons à prédire le nombre moyen journalier de vélos en circulation dans la ville de Paris, ce qui correspond à la variable "mean\_velo\_compteur" de notre dataframe.Il s'agit d'une variable dite "continue", qui à l'inverse d'une variable dite "discrète", peut prendre un nombre infini ou indénombrable de valeurs. Elle s'interprète comme la moyenne quotidienne du trafic par quart d'heure et par compteur.

Ainsi, contrairement à la variable “sum\_velo\_compteur”, la variable “mean\_velo\_compteur” n’est pas tributaire d'une variation du nombre de compteurs et permet de conserver une certaine cohérence dans le temps (même si bien sûr l'ajout de bornes de comptage supplémentaires ou la suppression de bornes pour raisons de travaux, pannes, etc., aura un potentiel impact sur la moyenne observée).

Voici les statistiques descriptives de la variable cible obtenue lors du recours à la fonction .describe()[[1]](#footnote-1) de Python :

count 365.000000

mean 8.685830

std 2.803391

min 1.740145

25% 6.894399

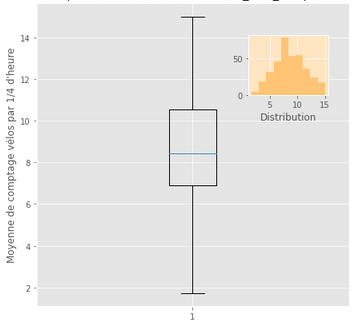
50% 8.418008

75% 10.556013

max 15.015220

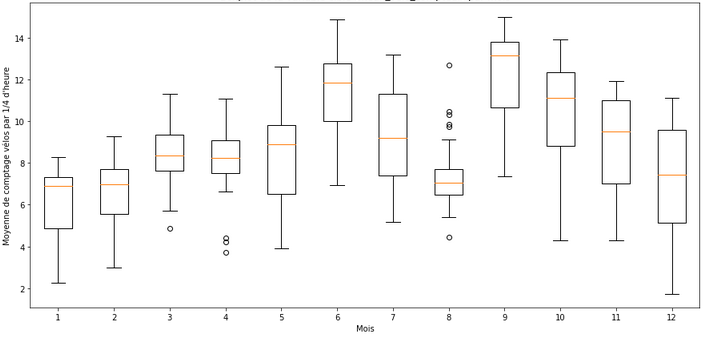
Name: mean\_velo\_compteur, dtype: float64

**Figure 8 : Boxplot de la variable cible :**



En analysant les données de cette variable grâce à un boxplot[[2]](#footnote-2), nous constatons que les valeurs sont distribuées normalement, avec peu de valeurs aberrantes (ou outliers).

**Figure 9 : Boxplot de la variable cible par mois :**

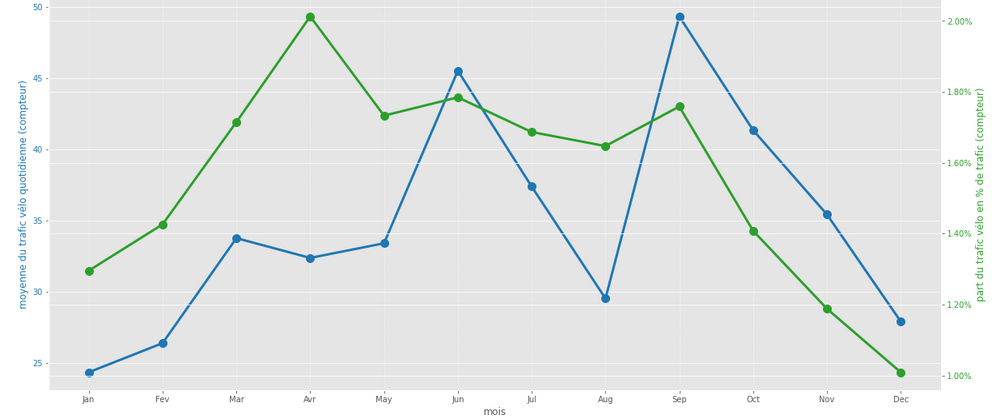


La distribution mensuelle de la variable nous apporte deux informations importantes.

La première est que l'on observe que le trafic vélo est en moyenne plus faible sur les mois hivernaux (décembre à février). La seconde est que le trafic vélo est en forte baisse sur les deux mois de vacances scolaires estivales (juillet/août).

Pour isoler les effets calendaires qui ont pour conséquence de diminuer ou faire augmenter le trafic (vacances scolaires, rentrée...), nous proposons sur la visualisation suivante d'explorer le lien entre la moyenne mensuelle du trafic vélo et sa part dans le trafic général sur Paris.

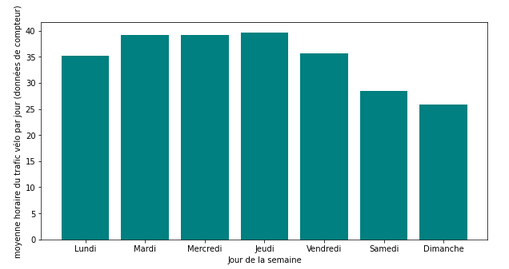
**Figure 10 : Comparaison de la moyenne de trafic vélo horaire mensuelle avec la part vélo mensuelle dans le trafic :**

La part du vélo (en vert) comme la moyenne horaire mensuelle du trafic vélo (en bleu) sont plus faibles en hiver que sur le reste de l'année. En revanche, la part du vélo reste relativement stable sur les mois estivaux. On peut déjà en déduire deux logiques à l'œuvre pour la détermination du trafic vélo : 

* La météo, plus défavorable en hiver, contribue à faire baisser le trafic vélo ainsi que sa part dans le trafic en général.
* Le trafic général, plus faible sur Paris lors des vacances scolaires estivales, fait mécaniquement baisser le nombre de vélos en circulation, mais sans pour autant faire baisser la part du vélo sur l'ensemble du trafic.

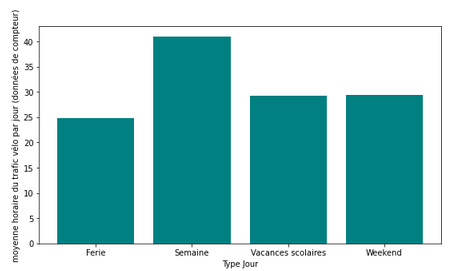
Pour comprendre la périodicité de l'utilisation du vélo au cours de l'année et au cours de la semaine, il est intéressant d'observer le nombre de déplacements sur ces périodes.

**Figure 11 : Distribution de la moyenne horaire du trafic de vélo en fonction du jour de la semaine (données de compteurs classiques) :**



On remarque que le vélo est plus largement utilisé en milieu de semaine (de mardi à jeudi) et beaucoup moins le week-end (samedi et dimanche notamment). La baisse d'utilisation le lundi et vendredi peut être liée à des jours de télétravail ces jours-là ou à des week-ends prolongés hors de Paris.

**Figure 12 : Distribution de la moyenne horaire du trafic de vélo en fonction du type de jour dans l’année (données de compteurs classiques) :**

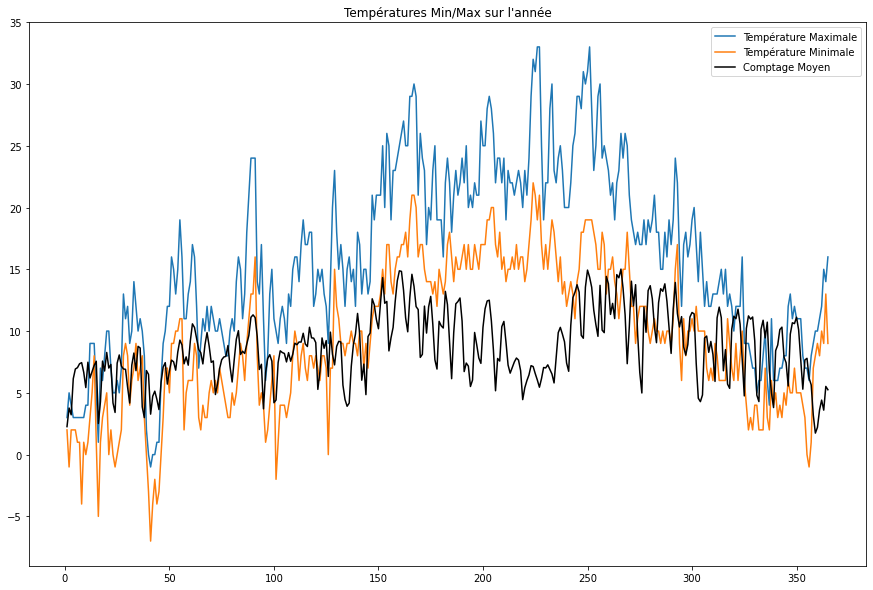
****

En analysant ce graphique, on se rend compte que les parisiens prennent davantage le vélo en semaine (hors vacances scolaires), que le week-end ou les jours fériés.

**Visualisation des variables de températures (vs. la variable cible) :**

Pour ce qui est des données météorologiques, les variables de températures seules ne permettent pas d’expliquer le nombre de vélos en circulation. En effet, bien que l’on observe une tendance à la corrélation (qui sera à confirmer plus tard) entre les températures et le comptage moyen de vélos en circulation sur le graphique ci-dessous, l’alignement des courbes n’est pas parfait.

**Figure 13 : Comparaison de l’évolution des températures minimales et maximales sur l’année 2021 avec la moyenne de comptage vélo par jour (les données ont été standardisées)**



# Analyse des corrélations

Pour approfondir le lien entre nos variables (météorologiques, de trafic et calendaires) et la variable cible, il est intéressant d’analyser leurs corrélations.

Pour tester l’indépendance de variables quantitatives, le test de corrélation de Pearson s'impose. Un test statistique est une procédure de décision entre deux hypothèses. Il s'agit de rejeter ou de ne pas rejeter une hypothèse statistique, appelée hypothèse nulle 'H0', en fonction d'un jeu de données. Dans le cas du test de corrélation de Pearson, l'hypothèse nulle est la suivante :

**H0 : "Les deux variables testées sont indépendantes"**

Nous choisissons de rejeter ou non cette hypothèse, en regardant la p-value de notre test : si cette dernière est en dessous de 5%, on rejette l'hypothèse H0. La p-value représente la probabilité de rejeter l'hypothèse nulle quand celle-ci est vraie. Plus la p-value est petite, plus la probabilité de faire une erreur en rejetant l'hypothèse nulle est faible.

Pour mesurer la corrélation entre les deux variables, on s'appuiera sur le coefficient de corrélation de Pearson.

Nous commençons par approfondir les liens entre les variables météorologiques et la variable cible.

**Tableau 1 : p-value entre les données météorologiques et variable cible :**

Une image contenant table

Description générée automatiquement

La p-value (colonnn PR >F) étant très inférieur au seuil de 5% sur la plupart des variables météorologiques, nous pouvons rejeter l'hypothèse nulle selon laquelle il y a indépendance entre l'attribut "mean\_velo\_compteur" et les attributs météorologiques sélectionnés, confirmant ainsi notre précédente analyse.

Concernant les données de trafic routier et dans les transports en commun :

**Tableau 2 : Données de déplacements et variable cible :**

Une image contenant texte, table

Description générée automatiquement

Les résultats sont encore plus concluants sur ces variables, avec des p-value extrêmement faibles.

Enfin, passons aux données calendaires :

**Tableau 3 : Données de type calendaire et variable cible :**

Une image contenant texte

Description générée automatiquement

Ici encore, la p-value très faible établit un lien entre les données de type calendaire et le trafic vélo.

Par conséquent, nous pouvons donc confirmer qu'il existe un lien de corrélation entre le trafic vélo, le trafic routier ou en transports en commun, les données météorologiques et les indicateurs calendaires.

Pour confirmer nos résultats, nous observons la corrélation entre notre variable cible et ces variables au travers de cartes de chaleur (ou heatmaps) :

**Figure 14 : Matrice de corrélation données météorologiques :**

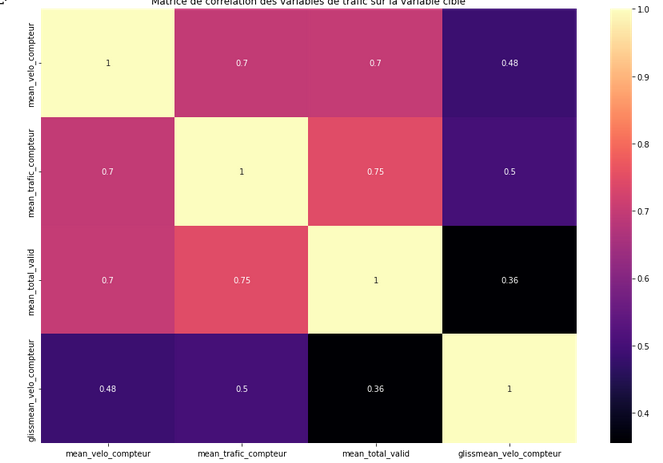
Une image contenant texte, objet d’extérieur

Description générée automatiquement

Plusieurs variables météorologiques, notamment celles liées à la température et à l’ensoleillement/visibilité, ont une corrélation relativement forte avec la variable cible (> 0,4).

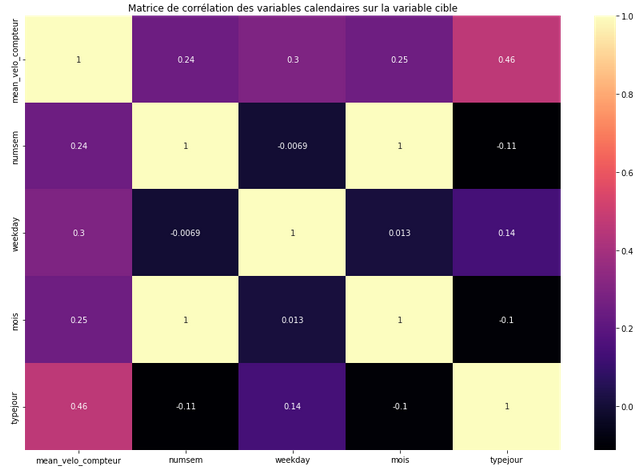
En revanche, contre-intuitivement, celles liées aux précipitations et à l'humidité le sont moins.

**Figure 15 : Matrice de corrélation des données de déplacements :**



On observe des relations très fortes entre la moyenne de comptage vélo et les données de trafic.

**Figure 16 : Matrice de corrélation des données calendaires :**



Les corrélations restent relativement faibles sur les variables calendaires (numsem, weekday, mois). Seule la variable ‘typejour’, qui caractérise chaque journée en fonction des vacances, jours fériés, weekend et jours de semaine, se démarque.

# Modélisation

En cohérence avec nos objectifs de départ de prédire la fréquentation vélo à Paris et de mettre en lumière les facteurs les plus explicatifs, nous cherchons désormais à intégrer les données précédemment créés et compilées à un modèle de prédiction le plus performant possible, dont la variable cible constituera la fréquentation vélo des pistes cyclables.

**1 – Types de modèles entraînés :**

Les modèles que nous entraîneront pour prédire notre variable cible sont des modèles d’apprentissage supervisés, de type régression.

En effet, nous cherchons à produire automatiquement des règles qui ne sont pas définies a priori, à partir d’une base de données d’apprentissage contenant des exemples.

De plus, notre objectif est de prédire une variable quantitative à partir de variables explicatives.

**2 – Les variables à intégrer au modèle :**

Suite à nos premières analyses, 25 variables nous ont semblées pertinentes à intégrer à notre modélisation sur la base des critères suivants :

**Les indicateurs doivent pouvoir être accessibles de manière cohérente dans le temps :**

Nous avons exclu sur ce critère les variables :

* + Issues de Google Trends car elles dépendaient du référentiel (base 100) utilisé lors de l’export. Or, à chaque nouvel export (sur une nouvelle période), ce référentiel serait différent ce qui romprait systématiquement la continuité des données.
  + Nous avons exclu les ‘*sum’* de trafic (que ce soit pour les features ou la target) car elles ne sont pas indépendantes du volume de comptage quotidien. En effet, une variation du nombre de comptage peut faire augmenter ou diminuer la somme du trafic sans que cela soit lié à des facteurs comportementaux. Les variables de moyenne (‘*mean*’) ont donc été privilégiées.

**Les indicateurs ne doivent pas être redondants :** Nous avons exclus les variables :

* + ‘*weather\_code’* qui indiquent un code d’appréciation météorologique de chaque tranche de la journée (morning, noon, evening) car elles étaient trop proches des variables de températures déjà divisées par tranche de journée et de la variable ‘opinion’ qui classifie la journée entière en fonction de la qualité de sa météo.
  + Les variables ‘*isWeekend’*, ‘*weekday’*, qui ne sont que des expressions plus détaillées des variables ‘*typejour’*.
  + ‘*sunset’* et ‘*sunrise’* car elles ne sont qu’une expression horaire du changement de saison, redondantes avec - et moins précises que - la variable *’nom\_mois’*.

**La variable cible :**

1. La moyenne de la fréquentation vélo par ¼ d’heure et par jour issue des données de compteurs (‘mean\_velo\_compteur’).

/ Ou la moyenne de la fréquentation horaire et par jour des données infrarouges (que nous avons tout de même testé pour s’assurer que même avec des données moins cohérentes, nous n’obtenons pas des modélisations aussi, voire plus performantes).

**31 variables explicatives :**

**3 indicateurs du niveau de déplacement global :**

1. ‘*mean\_total\_valid’* : La moyenne des validations horaires dans les transports en communs.
2. ‘*mean\_trafic\_compteur’* / ‘*mean\_trafic\_infra’*: La moyenne du trafic routier horaire issue des données de compteurs ou des données infrarouges en fonction de la target sélectionnée.
3. ‘*glissmean\_velo\_compteur*‘ : une moyenne glissante des 30 derniers jours de la target, qui permet d’intégrer une variable de continuité aux données alors que les échantillons de ‘Train’ et ‘Test’ ont été mélangés (voir paragraphe d’instanciation des échantillons).

**2 indicateurs des spécificités calendaires :**

1. ‘*typejour’* : variable catégorielle qui renvoie une classe par jour :

* Vacances scolaires
* Jour férié
* Jour de Weekend (hors vacances et jour férié)
* Jour de semaine (hors vacances et jour férié)

1. ‘*nom\_mois* : qui indique le nom du mois en cours

**19 indicateurs de météo détaillée :**

1. ‘*max\_temperature\_c’* : température maximale en degré Celsius
2. ‘*min\_temperature\_c’* : température minimale en degré Celsius
3. ‘*windspeed\_max\_kmh’* : vitesse maximale en kmh
4. ‘*temperature\_morning\_c’* : température du matin en dégré Celsius
5. ‘*temperature\_noon\_c’* : température à midi en dégré Celsius
6. ‘*temperature\_evening\_c’* : température en soirée en dégré Celsius
7. ‘*temperature\_night\_c’* : température nocture en degré Celsius
8. ‘*precip\_total\_day\_mm’* : total de précipitations journalières en mm
9. ‘*humidity\_max\_percent’* : pourcentage d'humidité en %
10. ‘*visibility\_avg\_km’* : visibilité moyenne en km
11. ‘*pressure\_max\_mb’* : pression maximale en mb (milibar)
12. ‘*cloudcover\_avg\_percent’* : couverture nuageuse moyenne en %
13. ‘*heatindex\_max\_c’* : index de chaleur max en dégré Celsius
14. ‘*dewpoint\_max\_c*‘: point de rosée max en degré Celsius (un bon indicateur de confort lié à l’humidité de l’air)
15. ‘windtemp\_max\_c’ : température du vent max en degré Celsius
16. ‘*total\_snow\_mm’* : précipitation de neige totale en mm
17. ‘*uv\_index’* : L’index UV moyen quotidien
18. ‘*sunhour*’ : qui indique pour chaque jour le nombre d’heure d’ensoleillement total.
19. ‘*opinion*’ : variable catégorielle qui renvoie à une classification des jours en fonction de la qualité de leur météo :
    * + - Météo idéale
        - Météo favorable
        - Météo correcte
        - Météo défavorable
        - Météo très défavorable

**3 - Instanciation des échantillons :**

Après avoir constitué le dataframe de features et isolé la target, nous avons dichotomisé les différentes variables catégorielles (‘*opinion’*, ‘*nom\_mois’*, et ‘*typejour’* ). Le nombre de variables total s’élève alors à 42.

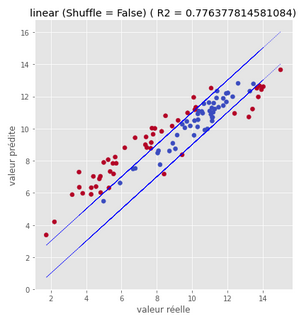
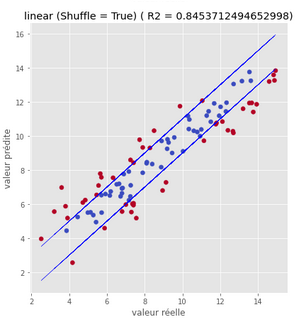
En raison des fortes disparités de valeurs entre les différents indicateurs, les variables ont été standardisées via un *StandardScaler().*

Nous avons ensuite divisé l’échantillon principal en sous échantillon de *Train (265/365 = 73%)* et de *Test (100/365 = 27%, soit 100 jours de l’année),* mélangé aléatoirement en dépit de la nature temporelle de la variable à prédire. En effet, les effets de saisons (météorologiques comme calendaires par ex : vacances scolaires...) auraient conduit à ce que la base d’entraînement ne tienne pas compte des spécificités de certaines périodes à prédire (ex : les 100 derniers jours de l’année très hivernaux avec des vacances scolaires). Une visualisation de cet écart de performance est présentée dans la figure (16).

**4 – Première itération via régression linéaire :**

Pour initier notre travail de modélisation nous sommes partis d’une régression linéaire « simple » car les valeurs à prédire sont des valeurs numériques continues. Les premiers résultats sont les suivants :

**Figure 17 : Comparaison entre une régression linéaire sur la base d’un X\_train, X\_test continu (Shuffle = False) vs mélangé aléatoirement (Shuffle = True).**

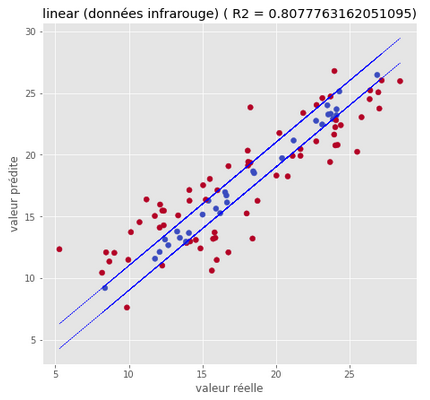
*Note : la marge d’erreur est ici représentée par les deux lignes bleues et correspond à un écart de +/- 1 (comptage vélo par ¼ d’heure) entre la valeur prédite et la valeur réelle. Le seuil de + ou – 1 a été fixé arbitrairement comme une erreur « acceptable » au regard de nos objectifs de précision.*

Comme vu précédemment, on observe un écart de performance important dans cette première modélisation entre des échantillons non mélangés et mélangés en faveur des seconds. Autre observation, le R2 de la régression linéaire «mélangée » est déjà relativement élevé (84,5%) sur l’échantillon de test, validant en première itération le choix des variables et des données sélectionnées.

Enfin, afin de définitivement arbitrer entre le choix d’une modélisation cherchant à prédire les données issues des compteurs classiques vs les compteurs infrarouges, nous faisons tourner une régression linéaire avec en target *‘mean\_velo\_infra’* et le remplacement de la feature *‘mean\_trafic\_compteur’* par *‘mean\_trafic\_infra’* (en conservant le paramètre Shuffle = True).

La visualisation ci-contre représente les résultats de cette régression. Ils sont, comme escompté, nettement moins bons que lorsque l’on utilise les données de compteurs.

**Figure 18 : Représentation des résultats d’une régression linéaire cherchant à prédire les données infrarouges :**

****

Sur cette base, nous décidons définitivement de ne plus travailler que sur les données de compteur et d’écarter les données infrarouges de la suite de l’analyse.

**5 – 15 modèles testés :**

Afin d’obtenir un modèle avec les meilleures performances possibles, nous avons ensuite amorcé une première phase de test et de tuning de 15 modèles différents :

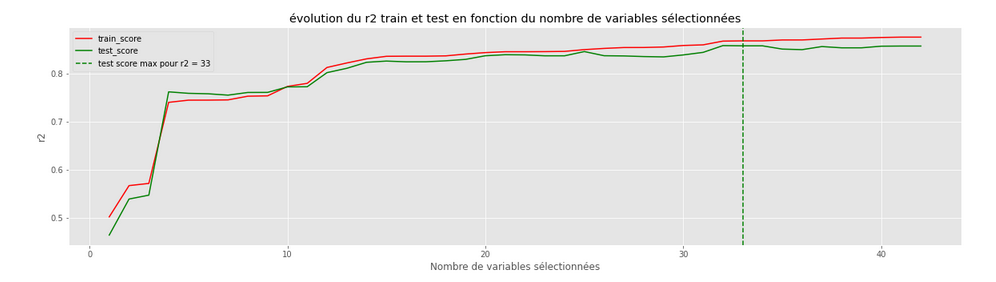
1. LinearRegression()
2. Ridge()
3. Lasso()
4. ElasticNEt()
5. XGBRegressor()
6. DecisionTree()
7. RandomForest()
8. Svm.SVR()
9. svm.NuSVR()
10. GaussianProcessRegressor()
11. RANSACRegressor()
12. neighbors.KNeighborsRegressor()
13. BayesianRidge()
14. MLPRegressor()
15. KernelRidge()

Chaque modèle a été optimisé au travers d’un GridSearch sur l’ensemble de ses hyperparamètres. Une seconde phase d’hypertuning a ensuite été réalisée pour affiner les paramètres (en faisant varier un à un les hyperparamètres sur des “*range”*  de plus en plus resserrés).

**6 – Réductions de dimensions :**

Trois approches différentes ont été initiées afin de réduire les dimensions du dataframe pour tenter d’améliorer les performances du modèle.

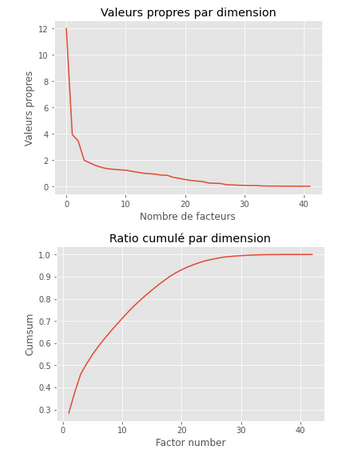
Le SelectKbest a identifié 33 variables nous permettant d’optimiser les résultats du R2 sur l’échantillon Test.

**Figure 19 : Résultat du SelectKbest sur la base de la régression linéaire :**

Nous avons donc créé des Dataframes Train et Test réduits à ces 33 variables et testé l’impact sur nos 15 modèles (en optimisant à chaque fois tous les hyperparamètres) et aucun modèle n’a dépassé les performances du MLPRegressor sur l’ensemble des variables. Un constat prédictible au regard de la stabilité relative des performances de la régression linéaire entre 33 et 42 variables.

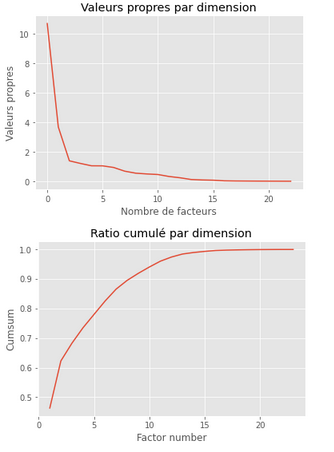
Nous nous sommes ensuite orientés vers une réduction de dimension par ACP (Analyse en Composante Principale), que nous avons en premier lieu testé sur l’ensemble des features.

**Figure 20 : Résultat d’une réduction des dimensions via une ACP sur l’ensemble des features :**



Sur cette base, nous avons testé différentes hypothèses en sélectionnant entre 20 et 30 dimensions (92% à 99% de la variance expliquée) afin d’en mesurer l’impact sur nos 15 modèles (avec de nouvelles optimisations des hyperparamètres). Résultat : pratiquement tous les modèles perdent en précision.

Par la suite, nous avons limité l’ACP aux seules variables météorologiques (car fortement inter-corrélées), dont voici le résultat :

**Figure 21 : Résultat d’une réduction des dimensions via une ACP sur l’ensemble des features météorologiques :** 

Pour ces variables, nous avons testé différentes hypothèses en sélectionnant entre 10 et 15 dimensions (94% à 98% de la variance expliquée).

Nous avons ensuite couplé ces dimensions aux autres variables du dataframe features (variables de déplacement et variables calendaires), puis nous en avons mesuré l’impact sur nos 15 modèles (avec de nouvelles optimisations des hyperparamètres). Résultat identique à l’ACP précédent : pratiquement tous les modèles perdent en précision.

Au final, aucune des tentatives de réduction des dimensions n’a permis d’améliorer significativement les performances des modèles (et conduit à des baisses de performances dans la plupart des cas). Nous décidons donc finalement d’intégrer au modèle l’ensemble des features présélectionnées, ce qui a pour avantage de conserver une capacité d’interprétation des résultats complète et plus intuitive.

**7 – Modèle sélectionné :**

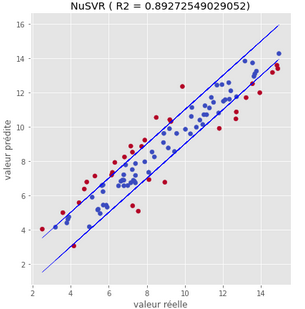
Même si plusieurs modèles sortent du lot avec des R² importants, nous avons préféré un modèle stable. Ainsi, après avoir fait tourner l’ensemble des modèles sur 2500 découpages différents des échantillons de Train et de Test nous avons conservé le modèle avec le R² moyen le plus élevé sur ces 2500 itérations couplé à un overfiting faible, voire nul. Le NuSVR fait donc office de gagnant avec les résultats suivants  :

**Résultats du modèle de NuSVR() :**

svm.NuSVR(kernel='rbf', gamma =0.0007372372372372372, nu = 0.9897897897897898, C = 266,  tol = 0.45)

* NuSVR rmse train : 0.8011762197473966
* NuSVR rmse test : 1.0269154745706668
* NuSVR score R2 train : 0.9093484984785367
* NuSVR score R2 test : 0.89272549029052
* **NuSVR score R2 test moyen sur 2500 itérations : 0.855690**

**Figure  : Représentation des résultats du NuSVR :**



Les outliers sont minoritaires et globalement proches de la marge d’erreur « acceptable ». Même si ce modèle est bien plus performant que la régression linéaire, on observe toujours une légère tendance à la sous-estimation des valeurs réelles les plus importantes et une surestimation des valeurs réelles les plus faibles.

# Interprétabilité

Jusqu’à présent, notre étude nous a permis de mettre en évidence les liens entre les variables météorologiques, de trafic et de type calendaire sur la fréquentation vélo à Paris.

Les résultats de nos modèles nous montrent que ces dernières nous permettent d’expliquer en grande partie les variations de déplacements dans la capitale et de prédire assez précisément ces volumes de déplacement.

Il convient désormais d’étudier le poids de chacun de ces facteurs dans nos prédictions, afin de déterminer lesquels ont un impact plus important pour expliquer notre variable cible.

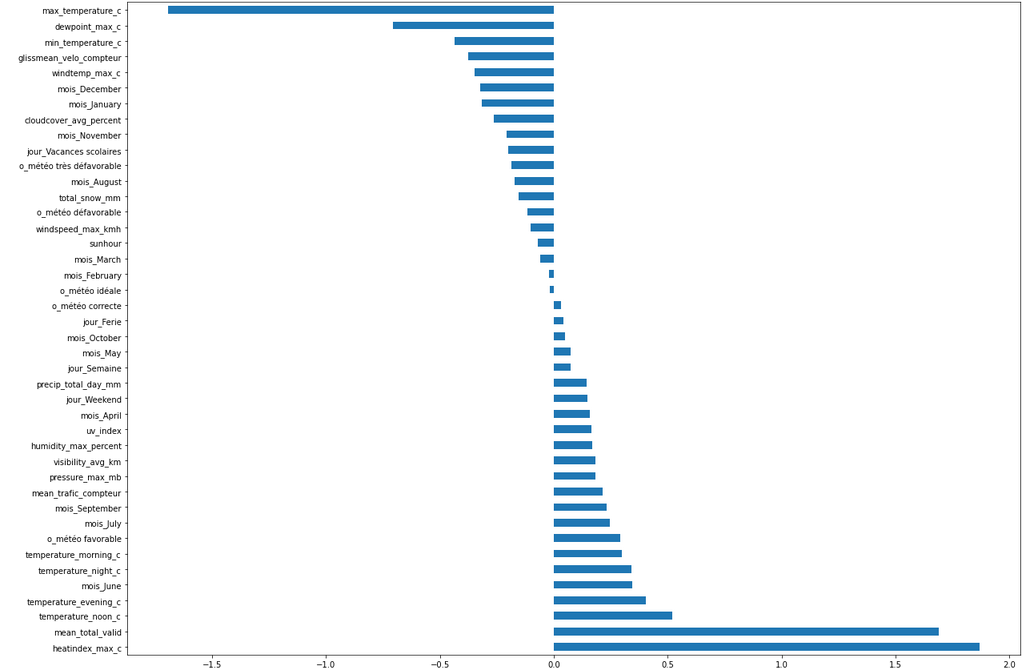
Pour cela, réaliser une étude d’interprétabilité de nos modèles nous semble une étape importante.

**1 – Interprétabilité grâce aux coefficients de Régression Linéaire :**

La régression linéaire est un modèle relativement simple à interpréter grâce à l’utilisation des coefficients de régression.

Ainsi, après avoir entraîné notre modèle, nous avons calculé les coefficients de chacune de nos 42 variables et les avons ordonné de la valeur la plus négative à la valeur la plus positive au travers du graphique suivant :

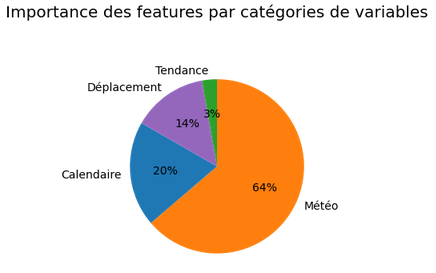
**Figure 22 : Représentation des coefficients de régression linéaire appliqués à nos 42 variables explicatives :**



Les coefficients d’un modèle de régression nous permettent d’avoir un aperçu des relations faites par le modèle entre chaque variable explicative et la variable cible.

Par exemple ici, plus ‘heatindex\_max\_c’ sera élevé, plus le nombre de comptage vélo prédit par notre modèle sera important. On notera toutefois que la variable qui contribue négativement le plus au trafic vélo est également un indicateur de température (‘max\_temperature\_c’). Cette incohérence pointe les limites de l’interprétation de notre modèle de régression qui inclut beaucoup de variables inter corrélées.

**Figure 23 : Diagramme circulaire représentant l’importance des variables du modèle de régression linéaire regroupées par catégorie**



Pour simplifier l’analyse, nous avons regroupé nos variables explicatives en 4 catégories : ‘Météo’, “Déplacement’ (= données de trafic’), ‘Calendaire’ et ‘Tendance’ (moyenne glissante du trafic vélo sur 30 jours).

En calculant le poids de chacune de ces variables dans notre modèle et en les regroupant, on remarque que les facteurs météos influent pour près de ⅔ sur les prédictions de notre modèle, les variables calendaires pour ⅕, le trafic pour 14% et les tendances de trafic vélo pour seulement 3%.

L’étude des coefficients de régression est intéressante. Cependant, d’autres facteurs entrent en jeu pour expliquer les prédictions du modèle, et notamment la distribution des variables. Il est possible que la prédominance du facteur météorologique (par rapport aux autres variables) est probablement plus le fait d’une prédominance de cette catégorie dans l’ensemble du modèle que qu’une réalité statistique.

Aussi la régression linéaire n’est pas le modèle nous ayant permis d’obtenir les meilleurs résultats sur notre échantillon.

C’est pour ces raisons que nous allons tenter d’interpréter nos résultats via un autre outil d’interprétation (la librairie SHAP) sur le modèle nous ayant donné le meilleur score : le NuSVR.

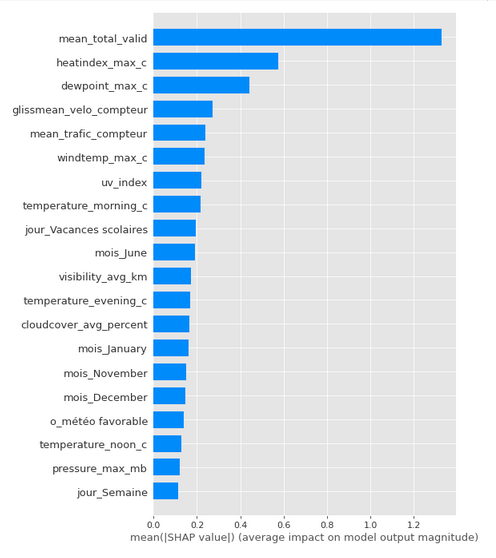
**2 – Interprétabilité du NuSVR grâce à la méthode SHAP :**

SHAP (SHapley Additive exPlanations) est un package proposant une approche unifiée pour expliquer le résultat de tout modèle de Machine Learning.

Pour cela, SHAP attribue à chaque variable une valeur d'importance pour une prédiction particulière, basée sur les valeurs de Shapley et la théorie des jeux coopératifs. L’idée est de faire la moyenne de l’impact qu’une variable a sur le modèle pour toutes les combinaisons de variables possibles. En faisant la moyenne sur les valeurs absolues des valeurs de Shap pour chaque variable, nous pouvons remonter à l’importance globale des variables.

Voici la représentation des “features importances” des 20 principales variables obtenues avec SHAP pour notre modèle NuSVR :

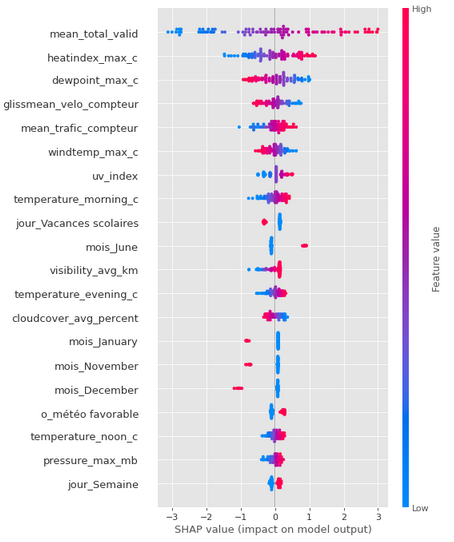
**Figure 24 : Graphique en bâton représentant l’impact des variables explicatives sur le modèle NuSVR par ordre d’importance décroissant**



En analysant le graphique ci-dessus, on remarque que 3 variables semblent avoir un impact particulièrement important sur notre modèle : la moyenne totale de validation dans les transports en commun en Ile-de-France, l’index de chaleur maximum en degré et le point de rosée maximum en degré. Trois variables déjà parmi les plus explicatives de notre modèle de régression linéaire.

On notera, en revanche, l’absence de la variable ‘max\_temperature\_c’, signe d’une plus forte cohérence des résultats obtenues grâce à SHAP.

**Figure 25 : Beeswarm des valeurs de Shapley calculées pour le modèle NuSVR, en fonction de la valeur des variables**

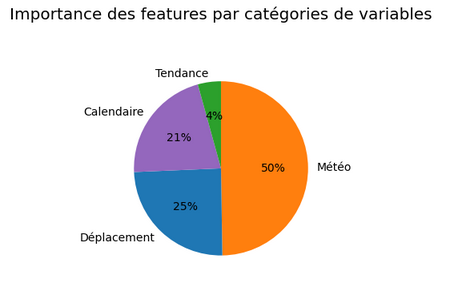


Sur le graphique ci-dessus, les variables explicatives sont ordonnées selon leur impact sur les prédictions. Mais on peut voir également dans quelle mesure les features impactent positivement ou négativement le modèle. Chaque point sur le graphique représente une observation. L’axe horizontal représente la valeur de Shapley, alors que la couleur de chaque point nous montre si une observation a une valeur supérieure ou inférieure, comparée aux autres observations.

Concrètement, cela nous permet d’affirmer que plus il y a de personnes dans les transports en commun, plus la fréquentation vélo sera importante à Paris (et vice versa). De même, plus l’indice de chaleur est élevé (donc plus il fait chaud), plus il y aura de vélos en circulation.

A l’inverse, plus l’indice de rosée est élevé (donc plus il fait humide), moins il y aura de vélos en circulation.

**Figure 26 : Diagramme circulaire représentant l’importance des variables explicatives dans le modèle NuSR regroupées par catégorie**



De la même façon que pour le modèle de régression linéaire, nous avons calculé l’importance des variables explicatives de notre modèle et les avons regroupées par catégorie. Cette fois-ci, les variables météos expliquent à elles-seules 50% des prédictions de notre modèle, les variables de trafic ¼ et le dernier quart est partagé entre les variables calendaires (21%) et la tendance de trafic vélo.

# Conclusion et regards critiques

En conclusion, les hypothèses selon lesquelles les conditions météorologiques, les conditions de trafic en général (routier et via les transports en communs) et les variables calendaires ont un impact sur la proportion de vélos en circulation à Paris ont pu être confirmées.

Nous avons été capables d’expliquer et de prédire avec des résultats relativement satisfaisants la fréquentation vélos à Paris sur la base de ces variables. En effet, notre modèle le plus performant nous a donné des scores de précision de près de 90% (NuSVR).

Sans surprise, les variables météorologiques semblent être les variables les plus explicatives du recours ou non au vélo pour les parisiens selon nos modèles. La température notamment, bien plus que la pluie, la neige, le vent ou la pression atmosphérique impactent nos résultats.

Aussi, plus les déplacements de personnes en général semblent importants, plus le recours au vélo est important, sans que l’on ait pu définir exactement dans quelle mesure chaque mode de transport a un impact sur les autres (par exemple, quand les transports en commun sont bondés ou quand il y a des bouchons à Paris).

Pour aller plus loin et étudier l’impact du trafic et des facteurs météo sur la fréquentation vélo, il nous semble important de poursuivre et approfondir cette étude :

* Pour cela, élargir la période étudiée aux données archivées nous permettrait de tester nos résultats et nos modèles sur les années passées.
* De même, tenter de prédire la fréquentation vélo à Paris en 2022 selon les facteurs météo et de trafic sur le début d’année, puis les comparer aux données réelles de comptage nous permettrait de tester la fiabilité prédictive de nos modèles.
* Ajouter de nouvelles variables explicatives (comme l’utilisation du Velib’ à Paris, des indicateurs d’achalandage des transports ou de congestion du trafic) pourrait nous permettre de renforcer nos bases de données et nos analyses.
* Enfin, au regard des incohérences pointées par l’analyse des coefficients de régression linéaire, une approche moins axée sur la recherche d’un R2 le plus performant possible, mais plutôt sur des modèles stables (moins d’overfitting) et à l’interprétation cohérente serait souhaitable. La réduction des variables trop inter-corrélées (notamment les variables de température) et la diminution du nombre de features en serait une première piste intéressante.

1. La fonction pandas.DataFrame.describe permet de générer des statistiques descriptives.

   Les statistiques descriptives comprennent celles qui résument la tendance centrale, la dispersion et la forme de la distribution d'un ensemble de données, à l'exclusion des valeurs NaN. [↑](#footnote-ref-1)
2. Le boxplot est une méthode standardisée permettant de visualiser la distribution des données basée sur 5 valeurs statistiques : valeur minimum, premier quartile [Q1], médiane, troisième quartile [Q3] et valeur maximum). [↑](#footnote-ref-2)