



DataScientest • com

Rapport technique

PyCycle in Paris

Évolution du trafic cycliste à Paris
entre septembre 2019 et décembre 2020

Benoit Gascou
Cynthia Laboureau
Joséphine Vaton

Promotion Bootcamp novembre 2020 / janvier 2021

Table des matières

Introduction	4
Contexte	4
Le dataset principal	4
Source	4
Période	4
Remarques	4
Le dataset complémentaire	4
Source	4
Période	4
Remarques	5
Les objectifs	5
1. DATA : EXPLORATION ET TRAITEMENT	6
1.1. Import des packages et aperçu du dataset	6
1.2. Affichage des informations du dataset	6
1.3. Affichage du nombre de valeurs manquantes	7
1.4. Affichage du nombre de doublons	7
1.5. Étendue des valeurs	7
1.6. Distribution des valeurs ('Comptage horaire')	7
1.7. Traitement des valeurs manquantes	7
1.8. Traitement de la variable 'Date et heure de comptage'	7
1.9. Test statistique entre les variables	7
1.10. Ajout de variables	8
1.10.1. 'Grève'	8
1.10.2. 'Covid'	8
1.10.3. 'Confinement'	8
1.10.4. 'Jours_fériés'	8
1.10.5. 'Vacances'	8
1.10.6. 'Pluie'	8
1.10.7. 'Froid'	8
1.10.8. 'Beau temps'	8
2. DATA VISUALISATION	9
2.1. Cartographie des sites de comptage	9
2.2.1. Cartographie : différences jour / nuit	10
2.2. Classement des sites selon l'intensité du trafic	10
2.3. Trafic cycliste à Paris entre le 01/09/2020 et le 31/12/2020	11
2.4. Trafic mensuel & comparaison automne 2019 / 2020	12
2.5. Trafic hebdomadaire	13

2.7. Trafic quotidien	13
2.8. Influence des facteurs récurrents	15
2.8.1. Influence des vacances scolaires sur le trafic (en 2020)	15
2.8.2. Influence des jours fériés	16
2.8.3. Influence de la météo	16
2.8. Influence des facteurs exceptionnels	17
2.8.1. La grève de l'hiver 2019/2020	17
2.8.2. La Covid	18
2.9. Bilan	19
3. BASE DE DONNÉES ACCIDENTS 2019	21
3.1. Exploration et traitement des données	21
3.2. Data visualisation	21
3.2.1 Analyse des données concernant les cyclistes accidentés	21
3.2.1.1. Par sexe	21
3.2.1.2. Par âge	22
3.2.1.3. Par nature du trajet	23
3.2.1.4. Par type de voie	23
3.2.1.5. Selon les conditions météo	24
3.3 Bilan cartographique	24
3.4. Évolution du nombre d'accidents en fonction du trafic	25
3.5. Test ANOVA entre le trafic et le nombre d'accidents	27
4. MACHINE LEARNING	27
4.2. Ajout des variables numériques	29
4.3. Analyse de la corrélation / linéarité entre les variables	29
4.4. Prédiction sur les derniers jours de chaque mois	31
4.4.1. Définition de la taille de l'échantillon test	32
4.4.2. Choix des variables avec SelectKBest	32
4.4.3. Création des échantillons 'train' et 'test'	32
4.4.4. Choix du modèle	32
4.4.4. Test sur les variables optimales	33
4.4.5. Représentation graphique des prédictions	34
4.5. Prédictions sur les 3 derniers mois	36
4.5.1. Définition de la taille de l'échantillon test	36
4.5.2. 2e modèle et évaluation	36
4.5.3. Représentations graphiques	37
4.5.4. Ajout de variables catégorielles	37
4.5.5. 3e modèle et évaluation	38
4.5.5. Représentation graphique pour décembre 2020	38

Description des travaux réalisés	38
Bibliographie	38
Difficultés rencontrées lors du projet	39
Bilan & Suite du projet	39
Annexes	41

Introduction

Contexte

A l'occasion du premier déconfinement en mai 2020, la Mairie de Paris a créé ex nihilo une cinquantaine de kilomètres de pistes cyclables intra-muros. Le but ? Désengorger les transports en commun pour éviter la propagation du virus et limiter le report sur les voitures particulières pour ne pas aggraver un trafic déjà saturé. Une politique pro-vélo déjà amorcée, mais amplifiée par la pandémie. Il suffit de se promener dans Paris pour le constater : il n'y a jamais eu autant de vélos circulant dans la capitale. Il serait intéressant d'étudier l'évolution du trafic cycliste à Paris avec les données mises à disposition sur le site de la Mairie.

Le dataset principal

Source

Le jeu de données provient du site de la Mairie de Paris :

« Comptage Vélo – Données compteurs »

https://opendata.paris.fr/explore/dataset/comptage-velo-donnees-compteurs/information/?disjunctive.id_compteur&disjunctive.nom_compteur&disjunctive.id&disjunctive.name&basemap=jawg.dark&location=13,48.8646,2.33914

Période

Les données sont mises à jour quotidiennement et remontent sur 13 mois glissants. Nous avons pu récupérer toutes les données **de septembre 2019 à décembre 2020**.

Remarques

- Les données sont fournies par un prestataire : Eco-Compteur. On suppose qu'elles sont fiables.
- Le nombre de compteurs évolue au fil du temps. Certains sont créés, d'autres s'arrêtent en cas de travaux ou de panne.

Le dataset complémentaire

Source

« Bases de données annuelles des accidents corporels de la circulation routière - Années de 2005 à 2019 »

<https://www.data.gouv.fr/fr/datasets/bases-de-donnees-annuelles-des-accidents-corporels-de-la-circulation-routiere-annees-de-2005-a-2019/>

Période

Nous allons sélectionner la période septembre 2019 - décembre 2019, la seule qui chevauche la nôtre.

Remarques

Le dataset est composée de 4 tables que nous allons réunir :

- caractéristiques
- usagers
- lieux
- véhicules

Les objectifs

L'objectif principal est d'étudier **l'évolution du trafic cycliste à Paris entre septembre 2019 et décembre 2020.**

Dans la partie **Data visualisation**, nous chercherons à comprendre :

- Comment le trafic évolue au fil des mois, des semaines et des jours ?
- Quels événements récurrents (vacances, jours fériés) ou exceptionnels (grèves, Covid, confinements) impactent le plus le trafic ?
- Quel est l'effet sur le nombre d'accidents impliquant des vélos ?

Dans la partie **Machine learning**, nous essaierons de créer un modèle de prédiction du comptage horaire par site, sur une période donnée (semaine ou mois).

1. DATA : EXPLORATION ET TRAITEMENT

1.1. Import des packages et aperçu du dataset

En amont - code à disposition - nous avons téléchargé les données du 1er octobre 2019 au 30 novembre 2020. Puis nous avons rajouté sep 2019 - fourni par Yohan - et décembre 2020.

Soit une période de 16 mois : de septembre 2019 à décembre 2020.

1.2. Affichage des informations du dataset

Nous nous retrouvons donc avec un fichier de **1 002 827 lignes x 9 colonnes**.

N° Col	Nom de la colonne	Nature	Description	Type
1	Identifiant du compteur	id	Identifiant du site de comptage + du compteur	string
2	Nom du compteur	id	Adresse postale du compteur	string
3	Identifiant du site de comptage	id	Identifiant du site de comptage (comprend 1 compteur si piste unidirectionnelle ou 2 compteurs si piste bidirectionnelle)	float (à modifier en string)
4	Nom du site de comptage	id	Adresse postale du site = la même que celle du compteur	string
5	Comptage horaire	cible	SUM sur compteurs pairs et impairs	float
6	Date et heure de comptage	explicative	Date et heure de comptage au format : "YYYY-MM-dd'T'HH:mm"	string (à modifier en datetime)
7	Date d'installation du site de comptage	explicative	Date d'installation du site de comptage	date
8	Lien vers photo du site de comptage	id	Lien vers la photo du site de comptage	string
9	Coordonnées géographiques	explicative	Coordonnées géographiques du site de comptage	geo_point_2d

La variable cible est donc le comptage horaire. C'est la seule variable numérique continue.

Les autres variables sont de 3 natures :

1. Dates

'Date et heure du comptage', 'Date d'installation du site de comptage' : sont ou doivent être de type Date.

2. Coordonnées GPS des sites

'Coordonnées géographiques' : de type geo_point_2d.

3. Identificatrices

'Identifiant du site', 'Identifiant du compteur', 'Nom du site' = 'Nom du compteur', 'Lien vers photo du site de comptage' : sont ou doivent être des chaînes de caractères.

Conclusion : seules 'Date et heure du comptage' et 'Coordonnées géographiques' sont véritablement des variables explicatives, respectivement temporelle et géographique.

1.3. Affichage du nombre de valeurs manquantes

Le Dataframe compte un nombre significatif de valeurs manquantes. Les 6 variables concernées ont le même nombre de valeurs manquantes : **37 152, soit 3,7 %**. Il s'agit probablement des mêmes lignes.

1.4. Affichage du nombre de doublons

Il n'y a pas de doublon dans le jeu de données.

1.5. Étendue des valeurs

Il y a apparemment 99 compteurs sur 70 sites, mais nous n'avons que 69 coordonnées géographiques différentes.

Conclusion : Nous travaillerons sur ces **69 sites de comptage**.

1.6. Distribution des valeurs ('Comptage horaire')

Il semble y avoir des valeurs extrêmes (3e quartile : 68, dernier : 1275), ce qui expliquerait une moyenne à 53.3 pour une médiane à 22. Nous faisons un boxplot par 'Identifiant de compteur'. Les valeurs extrêmes sont nombreuses et continues. De plus, un minimum de 0 vélo / heure et un maximum de 1275 (soit 21 vélo / min) ne sont pas aberrants.

Conclusion : Nous décidons de tout garder pour ne pas perdre d'information.

1.7. Traitement des valeurs manquantes

6 variables présentent des valeurs manquantes. Elles sont toutes relatives à l'identification du compteur. Avec la variable : "Nom du Compteur", qui correspond à l'adresse, nous pouvons retrouver les infos manquantes. Nous vérifions, cela n'a pas créé de doublons.

Conclusion : Nous avons remplacé toutes les valeurs manquantes.

1.8. Traitement de la variable 'Date et heure de comptage'

Nous convertissons la variable 'Date et heure de comptage' au format DateTime, puis nous créons de nouvelles variables : '**Date**', '**Année**', '**Mois**', **Semaine**', '**Jour**', '**Jour de la semaine**', '**Heure**'.

1.9. Test statistique entre les variables

Dans ce dataset, il n'y a pas de relation de dépendance à explorer entre les variables explicatives. En revanche, on peut étudier les relations de dépendance entre la variable cible ('Comptage horaire') et certaines variables catégorielles : le mois, le jour de la semaine, l'heure et le lieu ('Coordonnées géographiques'). On se doute déjà qu'elles sont fortement corrélées. Nous réalisons des tests ANOVA.

Conclusion : Toutes les p_values sont égales à zéro. Probablement parce que les variables choisies, temporelles et géographiques, sont extrêmement liées à la cible.

1.10. Ajout de variables

Pour manipuler plus facilement les dates, on passe la colonne Date en index. On réinitialise l'index à la fin.

1.10.1. 'Grève'

0/1 : pas de grève / grève des transports (du 05/12/2019 au 10/01/2020)

1.10.2. 'Covid'

0/1 : avant / pendant la Covid (du 1er jour du 1er confinement, le 17/03/2020, à la fin de notre dataset)

1.10.3. 'Confinement'

0/1/2 : pas de confinement / 1^{er} confinement (17/03/2020 au 10/05/2020) / 2^e confinement (30/10/2020 au 15/12/2020)

1.10.4. 'Jours_fériés'

0/1 : hors jour férié / jour férié

1.10.5. 'Vacances'

0/1 : hors vacances scolaires à Paris / vacances scolaires à Paris

Puis on crée des variables par type de vacances (0/1) :

'vac_noel19', 'vac_fevrier', 'vac_printemps', 'vac_ascension', 'vac_juillet', 'vac_aout', 'vac_toussaint', 'vac_noel20'

1.10.6. 'Pluie'

0/1/2 : pas ou peu de pluie (< 10 mm/jour) / pluie modérée (10 < mm/jour < 15) / pluie forte et orage (> 15 mm/jour)

Remarque : nous n'avons pas trouvé de base de données libre de droits reprenant les conditions météo à Paris pour notre période. D'où le choix de ces indicateurs, renseignés à la main. Idem pour les températures : nous ne focalisons que sur les extrêmes. Quand il fait froid pour Paris (< 4°) ou chaud (> 23°).

1.10.7. 'Froid'

0/1 : au-dessus / en-dessous de 4°

1.10.8. 'Beau temps'

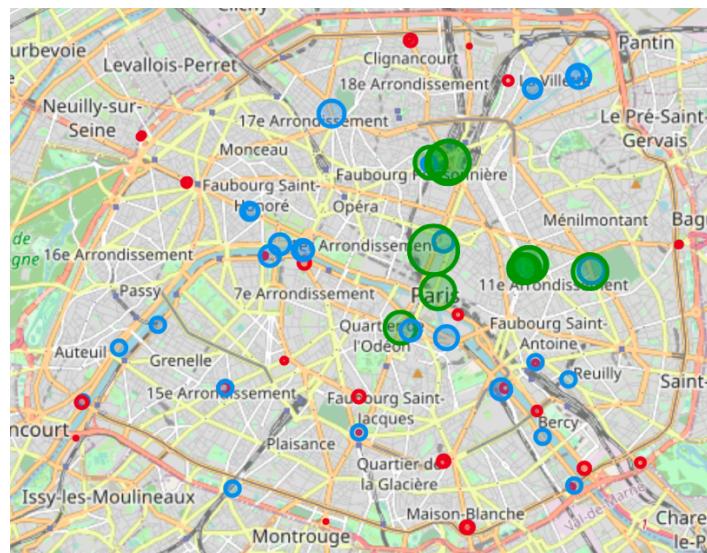
0/1 : en-dessous de 25° / au-dessus de 25° avec du soleil

2. DATA VISUALISATION

2.1. Cartographie des sites de comptage

Pour la cartographie, nous avons choisi la bibliothèque **folium**. Elle permet de géolocaliser des informations sur une carte dynamique et détaillée de type OpenStreetMap. En cliquant sur les cercles, on fait apparaître l'adresse, l'identifiant et la photo du site, ainsi que le comptage horaire moyen.

Pour compléter cette simple observation des différents sites de comptage, nous avons choisi d'y ajouter une classification en testant **un modèle de clustering, l'algorithme des k-moyennes (K-Means)**. Cette méthode nous permet de regrouper les sites entre eux pour visualiser s'il existe des similitudes entre les différents sites de comptage selon leur position géographique (mesurée par la latitude et la longitude) et le nombre moyen de comptages observés sur la période. Après l'avoir testé avec différents nombres de clusters (argument « `n_clusters` » de la fonction « `kmeans` »), l'algorithme semble identifier 3 groupes de sites de comptages selon l'importance du trafic et dans une moindre mesure selon leur situation géographique. Ces 3 groupes sont représentés par 3 couleurs différentes sur la carte.



Les sites de comptage sont donc représentés par les cercles bleus, verts et rouges. Leur taille est proportionnelle à la moyenne des comptages horaires.

Remarques sur la répartition des sites de comptages :

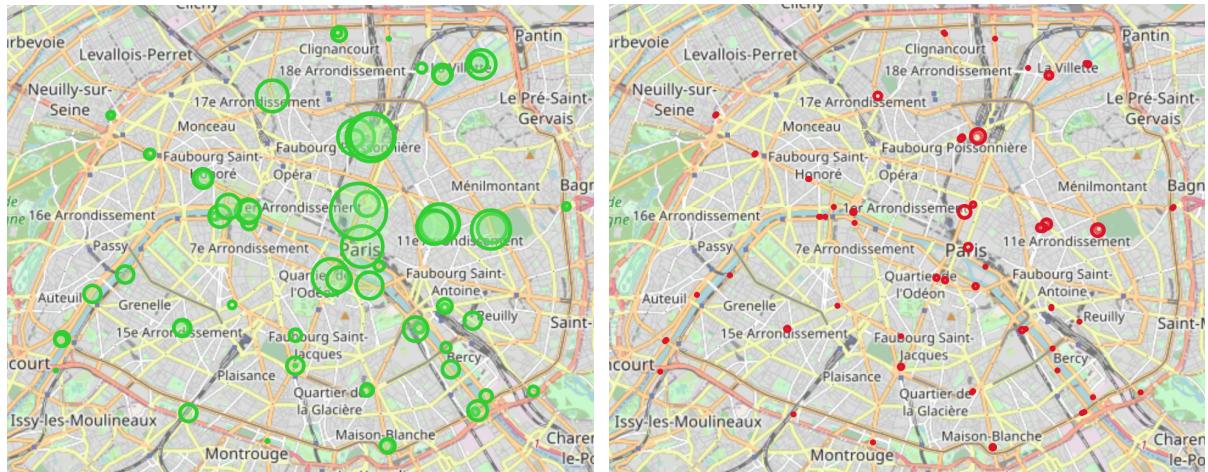
- Dans certains cas, où la piste cyclable se situe de part et d'autre d'un boulevard, 2 sites de comptage ont été créés au lieu de 2 compteurs pour un seul site. Nous précisons ici que notre analyse portera sur les 69 sites de comptages (à sens unique ou à double sens) et non pas sur chaque compteur.
- **Points forts** : il y en a un peu partout dans Paris. Les quais de Seine et l'axe Nord-Sud (Gare du Nord/Gare de l'Est - Châtelet - Odéon - Alésia) sont particulièrement bien équipés.
- **Points faibles** : à l'échelle de chaque quartier le maillage est faible (3,5 sites par arrondissement). Certains trajets très empruntés sont peu ou pas couverts, notamment sur les axes Est-Ouest : Porte de Vincennes - Nation - Bastille - Saint-Paul et Porte de Bagnolet - Père-Lachaise - République - Saint-Lazare.

Première analyse :

Trois gros points de passage se dégagent à première vue : l'hyper-centre de Paris, le secteur des gares du Nord et de l'Est, et enfin le 11e arrondissement (Popincourt et Père-Lachaise). Les quais de Seine se démarquent aussi par leur trafic soutenu. Plus on s'éloigne du centre, moins les sites sont fréquentés.

Première question que nous nous posons : existe-t-il des points de passage plus fréquentés la nuit que le jour ?

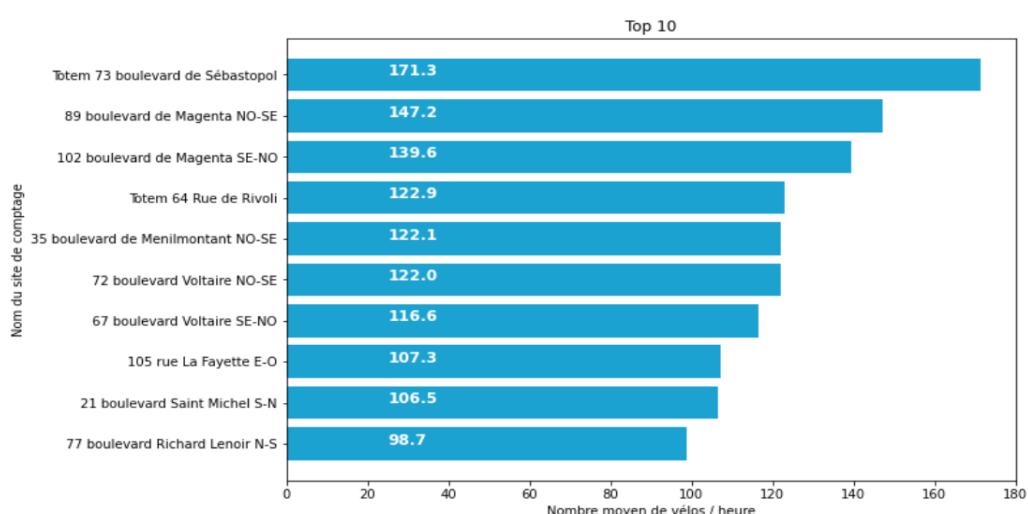
2.2.1. Cartographie : différences jour / nuit

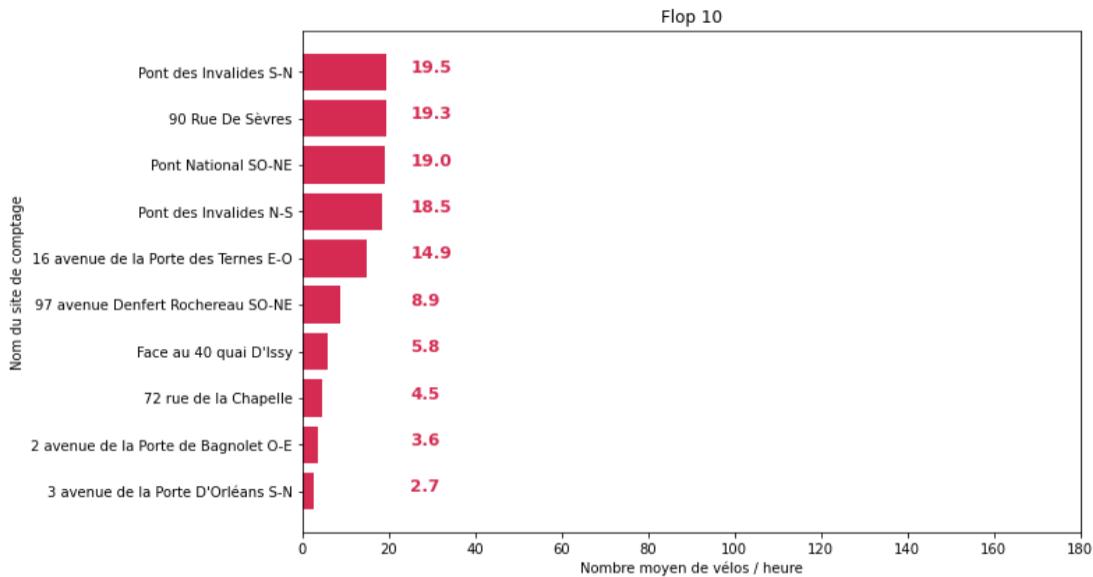


En toute logique, il y a beaucoup moins de vélos circulant la nuit. En journée, les secteurs les plus fréquentés restent le centre de Paris (des Halles à Odéon), puis le secteur des gares du Nord et de l'Est et enfin le 11e arrondissement (Popincourt et Père-Lachaise). La nuit, les points de passage principaux sont exclusivement rive droite avec toujours Les Halles, les gares du Nord et de l'Est et le 11e. Ce sont en effet les quartiers les plus denses en lieux de sorties, le reste de Paris étant plus résidentiel.

Intéressons-nous maintenant aux sites les plus et moins fréquentés.

2.2. Classement des sites selon l'intensité du trafic



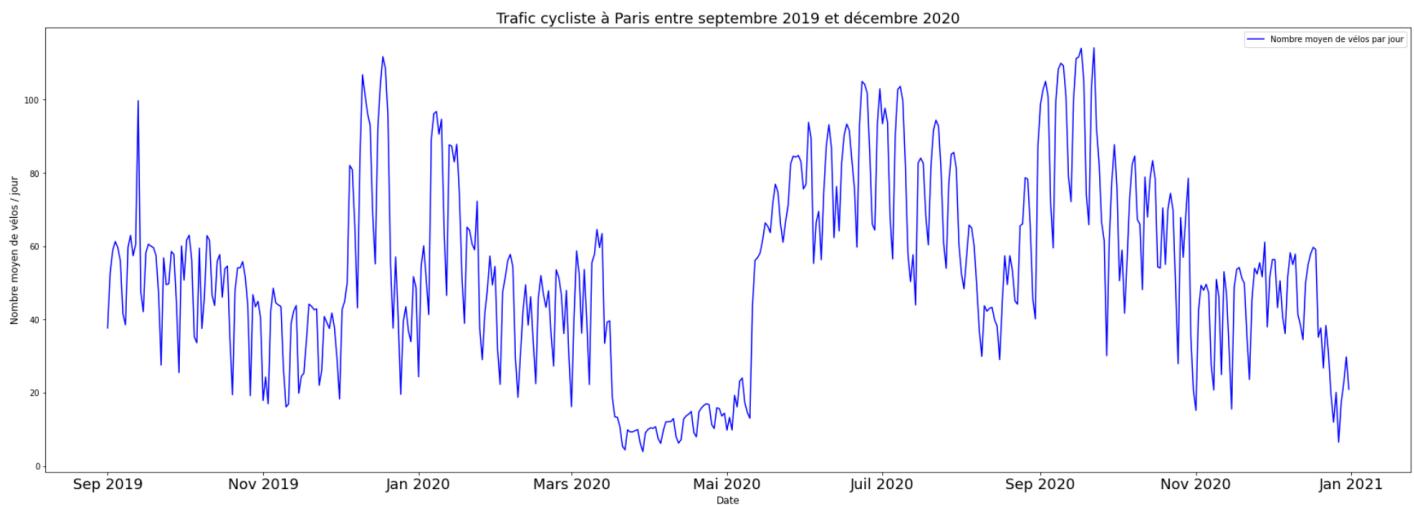


Le site du boulevard Sébastopol (à double sens) semble le plus fréquenté avec 175 vélos / heure en moyenne. Mais les 2 sites suivants correspondent en fait aux 2 voies à sens unique de part et d'autre du **boulevard de Magenta**, qui cumule donc **289 vélos / heure**. L'écart avec le point de passage le moins fréquenté - **4 vélos / heure** en moyenne à **Porte d'Orléans** - est énorme. Il y donc une grande disparité en fonction des lieux.

Dans le reste du classement, on retrouve sans surprise les sites les plus fréquentés cités précédemment et les moins fréquentés plutôt en périphérie.

Regardons maintenant l'évolution dans le temps.

2.3. Trafic cycliste à Paris entre le 01/09/2020 et le 31/12/2020



On observe un pic le 22 septembre 2019, lors de la Journée mondiale sans voiture. Pendant l'automne, le trafic baisse progressivement, le froid et la pluie rebutant les usagers.

La grande grève des transports débutée en décembre 2019 inverse la tendance. Le trafic monte en flèche, dépassant même le pic de la Journée sans voiture. Seules les vacances de Noël cassent, provisoirement, la courbe. Lorsque les transports publics reprennent fin janvier,

le trafic diminue, mais reste légèrement supérieur au mois précédent la grève. Il remonte ensuite à l'approche du printemps.

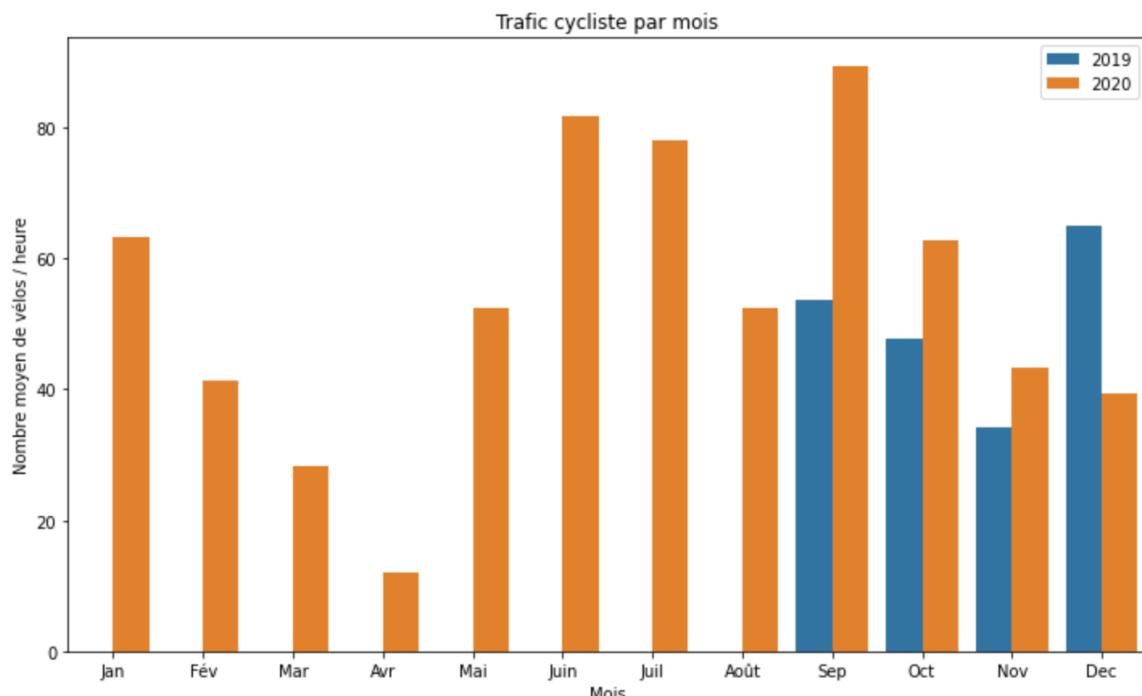
Lors du 1er confinement mi-mars le trafic s'effondre. Mi-mai, avec le déconfinement et les beaux jours, il reprend de plus belle. On observe un creux au mois d'août, dû aux vacances et à l'absence de touristes cette année.

A la rentrée, le trafic cycliste à Paris atteint des sommets, plus hauts encore que ceux de la grève ! Il faut dire que le mois de septembre 2020 a été le plus chaud jamais enregistré en France. Et la crise sanitaire est passée par là. Peur des transports publics, création de 50 km de pistes cyclables par la Mairie, sans compter les aides de l'Etat pour l'achat ou la réparation de vélos... Bon nombre de parisiens ont adopté la petite reine comme moyen de transport quotidien.

En octobre, le trafic diminue, mais reste largement supérieur à l'automne 2019. A la fin du mois, c'est le deuxième confinement, moins sévère que le premier. Puis viennent les vacances de Noël sous couvre-feu et le trafic s'effondre de nouveau.

Regardons cette évolution de plus près.

2.4. Trafic mensuel & comparaison automne 2019 / 2020

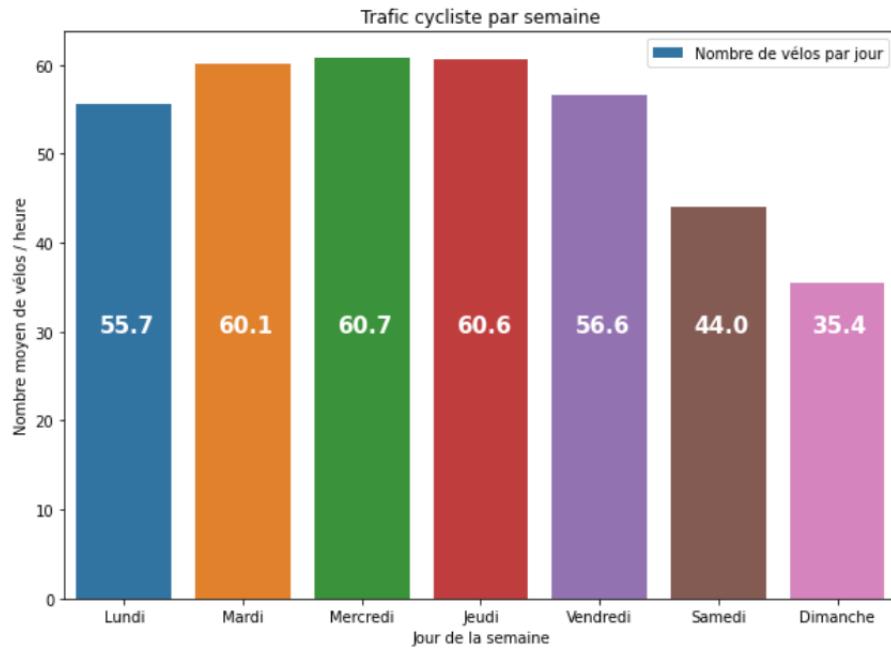


La hausse du nombre de vélos circulant à Paris est flagrante entre les automnes 2019 et 2020, avec + 50% en septembre ! Seul le mois de décembre 2019, au plus fort de la grève, dépasse celui de 2020, plombé par le deuxième confinement et le couvre-feu.

Pour le reste de l'année 2020, on observe l'effet positif de la grève en janvier, du déconfinement en mai et du beau temps jusqu'en septembre. Les impacts négatifs sont dûs aux vacances en août et aux deux confinements de mars et novembre.

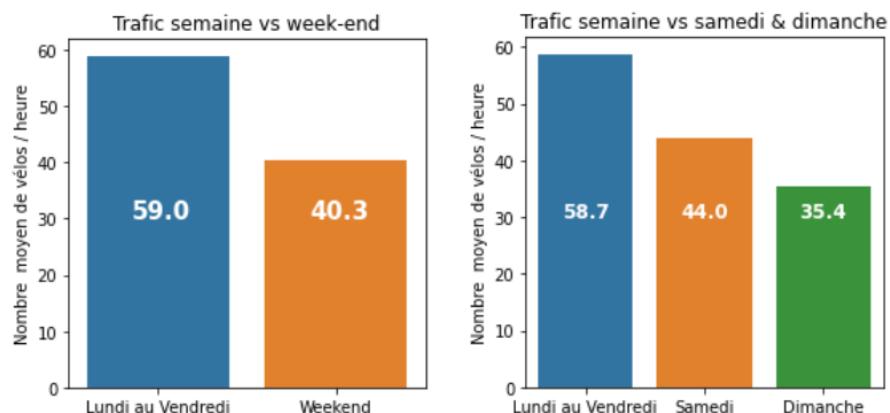
Continuons de zoomer pour regarder le trafic par semaine.

2.5. Trafic hebdomadaire



Sur le premier graphique la distribution périodique était en dents de scie. Cela se confirme ici avec une information supplémentaire : les vélos circulent plus en milieu de semaine et moins le week-end. Le vélo à Paris n'est donc pas un simple loisir, il est principalement un moyen de transport quotidien.

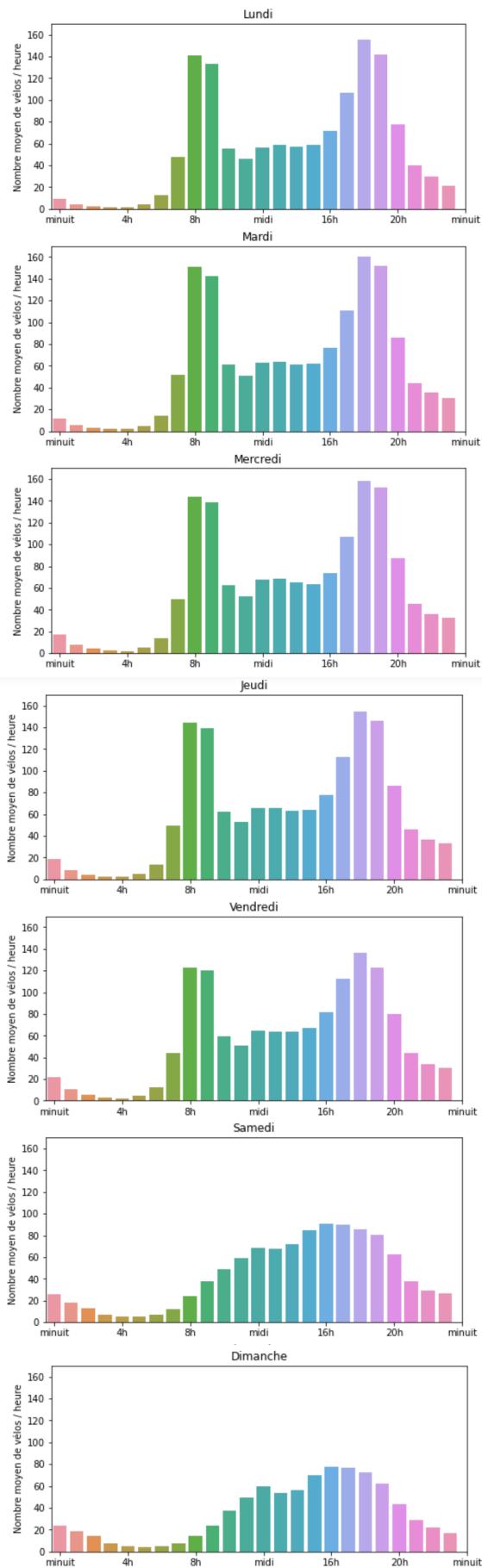
Regardons plus en détail la répartition entre les jours de semaine et le week-end.



Le trafic baisse d'un tiers le week-end : - 25% le samedi et - 40% le dimanche.

Qu'en est-il du trafic quotidien ?

2.7. Trafic quotidien

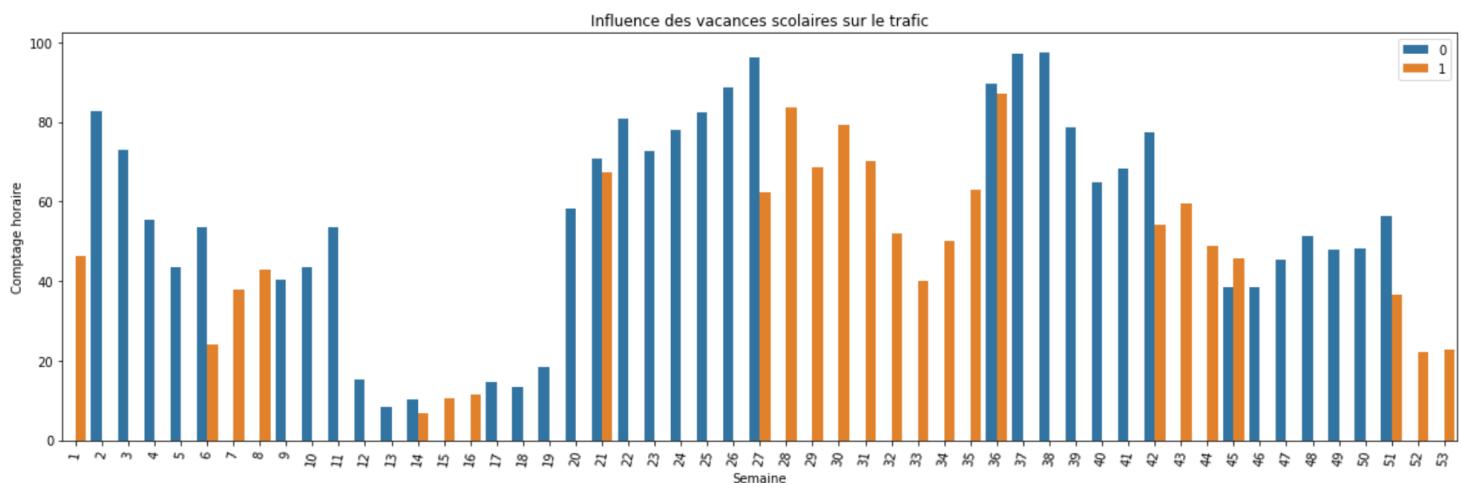


En semaine, on observe 2 pics aux heures de pointe : entre 8 et 9h, puis entre 17 et 19h. Le week-end, la courbe est beaucoup plus lissée avec une progression régulière jusqu'à 17h, puis une diminution dans la soirée.

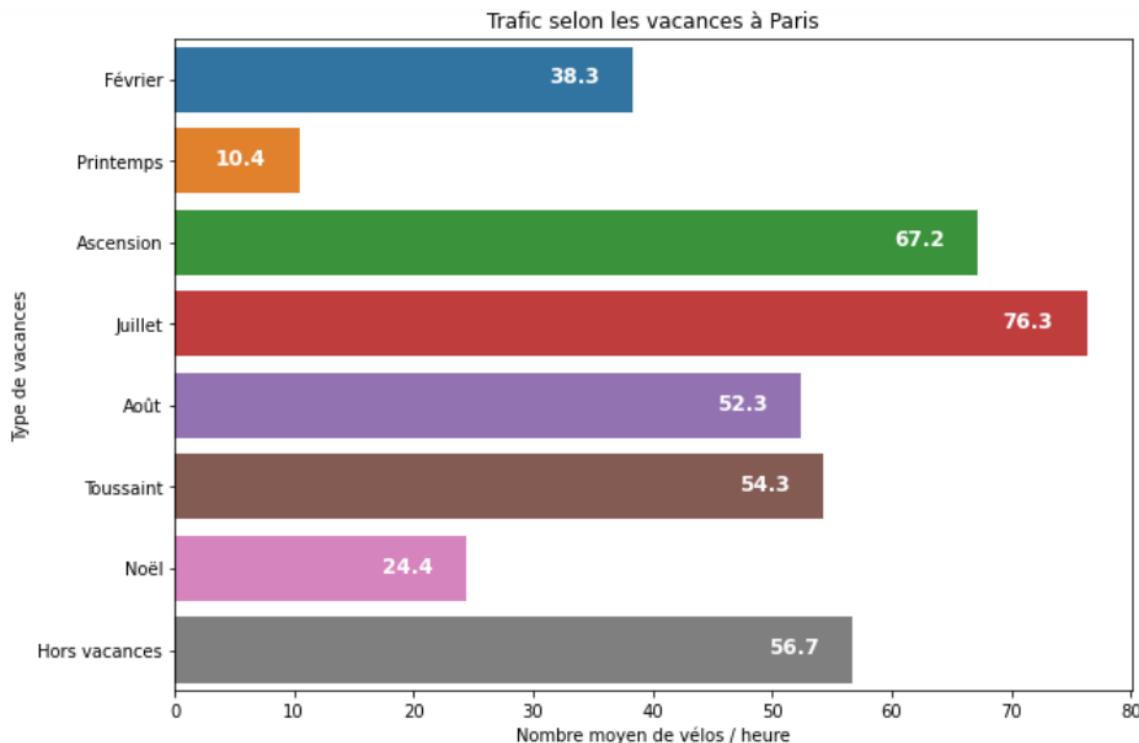
Une année ordinaire est rythmée par les vacances scolaires, les jours fériés ou encore par la pluie et le beau temps. Quelle est l'influence de ces facteurs récurrents ? Commençons par les vacances.

2.8. Influence des facteurs récurrents

2.8.1. Influence des vacances scolaires sur le trafic (en 2020)



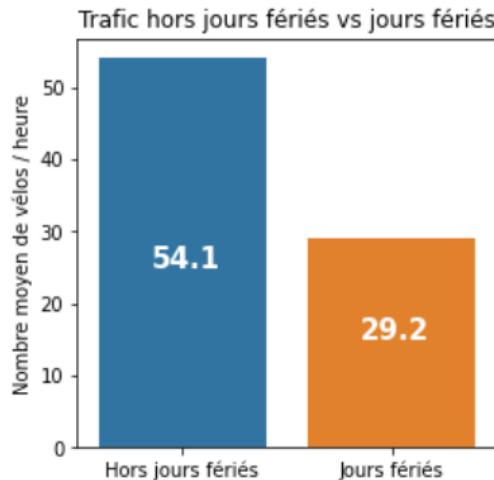
Le trafic semble diminuer pendant les vacances scolaires, mais pas toujours dans les mêmes proportions. Regardons de plus près l'effet de chaque type de vacances.



Il n'y a que les vacances de Noël qui font drastiquement baisser le trafic de 57%. Les vacances de février l'amputent de 33%. La Toussaint et août influent peu. Le trafic augmente même lors des vacances de juillet et de l'Ascension 2020. Nous ne considérons pas les vacances de printemps, qui ont eu lieu lors du premier confinement.

Et les jours fériés, comment impactent-ils le trafic ? Zoomons sur la période avril-août, là où il y en a le plus.

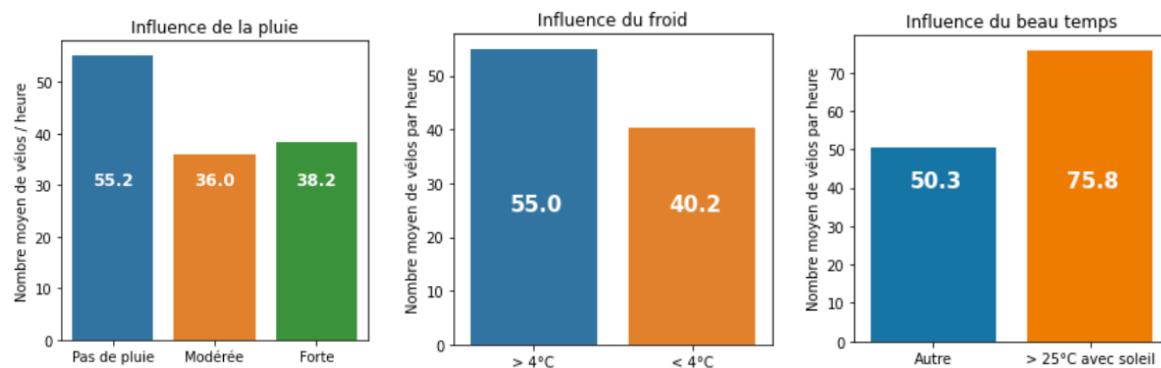
2.8.2. Influence des jours fériés



Sur la période étudiée, les jours fériés font baisser le trafic horaire de 54%.

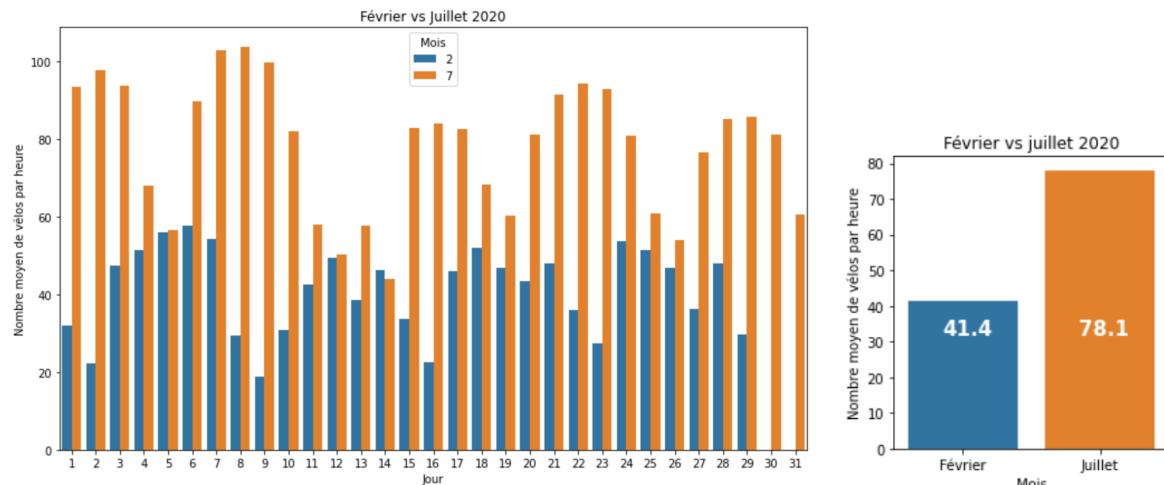
Pour la dernière analyse des facteurs ordinaires, observons l'effet de la météo.

2.8.3. Influence de la météo



La pluie fait baisser le nombre de vélos d'un tiers. En revanche, qu'elle soit forte ou modérée ne semble pas jouer. Ceux qui sont prêts à affronter la pluie le font coûte que coûte ! Un thermomètre qui descend sous les 4 degrés refroidira 28% des cyclistes, tandis que le beau temps en fera sortir 50% de plus.

Comparons 2 mois que tout oppose : février, froid et humide, et juillet, beau et sec. Nous choisissons deux mois peu ou pas impactés par un confinement, la grève ou les vacances.



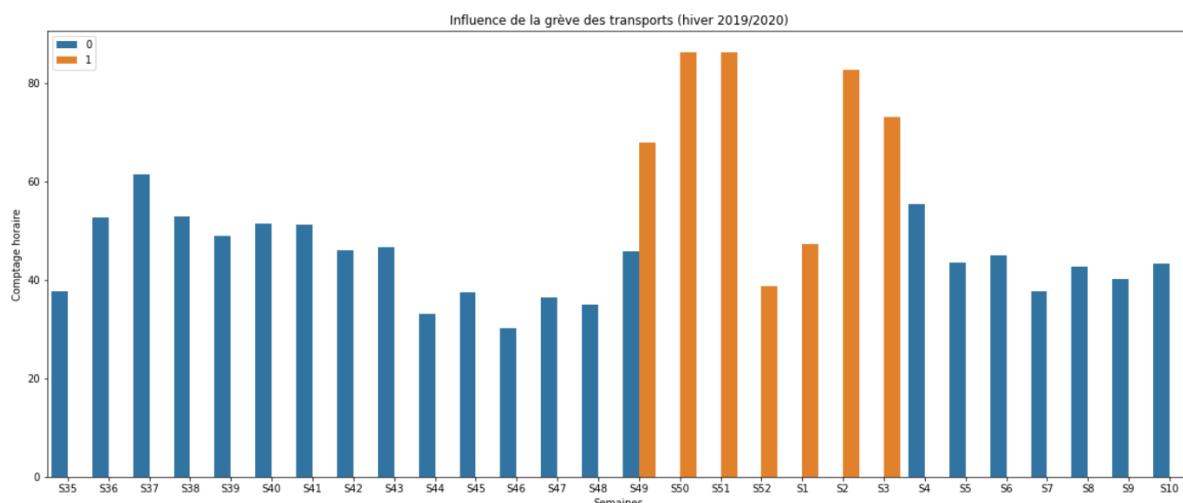
En moyenne, deux fois plus de vélos ont circulé en juillet qu'en février. Nous ne pouvons pas exclure qu'il y ait aussi eu un effet Covid entre les deux.

En dehors de ces variations ordinaires, les 16 mois derniers ont eu lieu des événements hors du commun, comme la longue grève des transports ou la crise du Covid. Quel a été leur impact sur le trafic cycliste ?

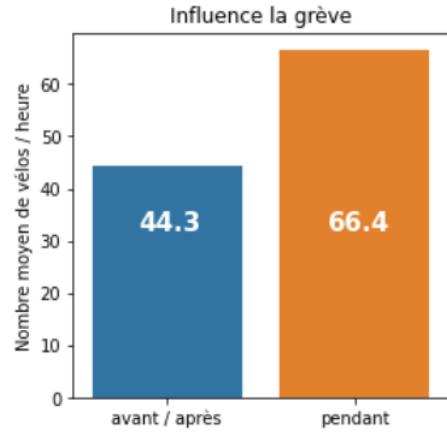
2.8. Influence des facteurs exceptionnels

2.8.1. La grève de l'hiver 2019/2020

Longue de 2 mois, elle marque un tournant dans l'usage du vélo à Paris. Qu'il pleuve, qu'il vente ou qu'il neige, les parisiens ont ressorti leur vélo. Pour distinguer l'effet de la grève de celui de la crise sanitaire, nous étudions d'abord la période 1er septembre 2019 - 15 mars 2020 (veille du premier confinement).



L'impact de l'arrêt des transports en commun a été quasiment immédiat avec un trafic cycliste qui a quasiment doublé du jour au lendemain. Les vacances de Noël, dont nous avons vu l'impact très fort, ont fait baisser la moyenne pendant deux semaines.

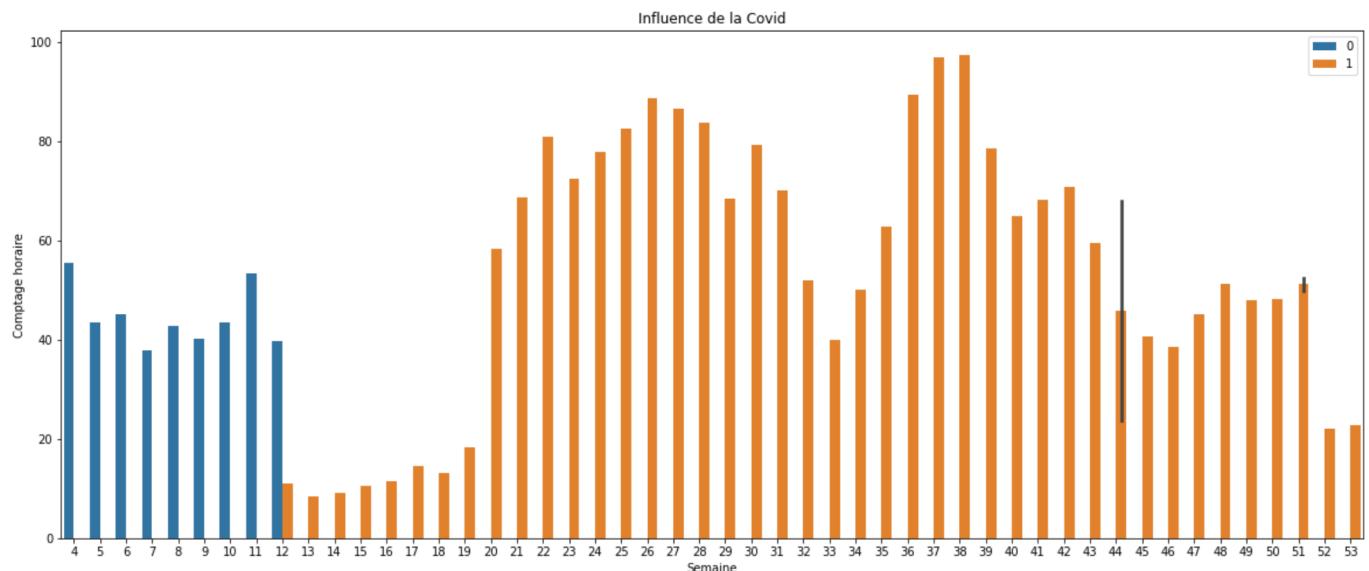


Pendant la grève, le trafic a bondi de 50%, passant de 44 vélos / heure en moyenne à 66.

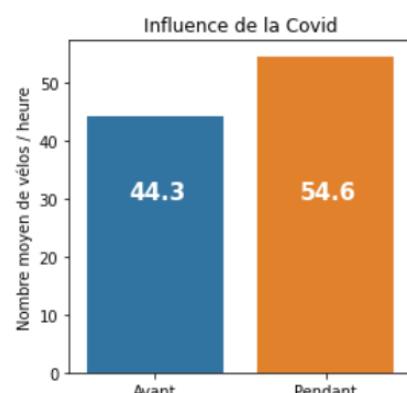
Venons-en enfin à l'évènement majeur de 2020 : la crise sanitaire liée à la Covid. Quel a été son impact sur l'usage du vélo à Paris ?

2.8.2. La Covid

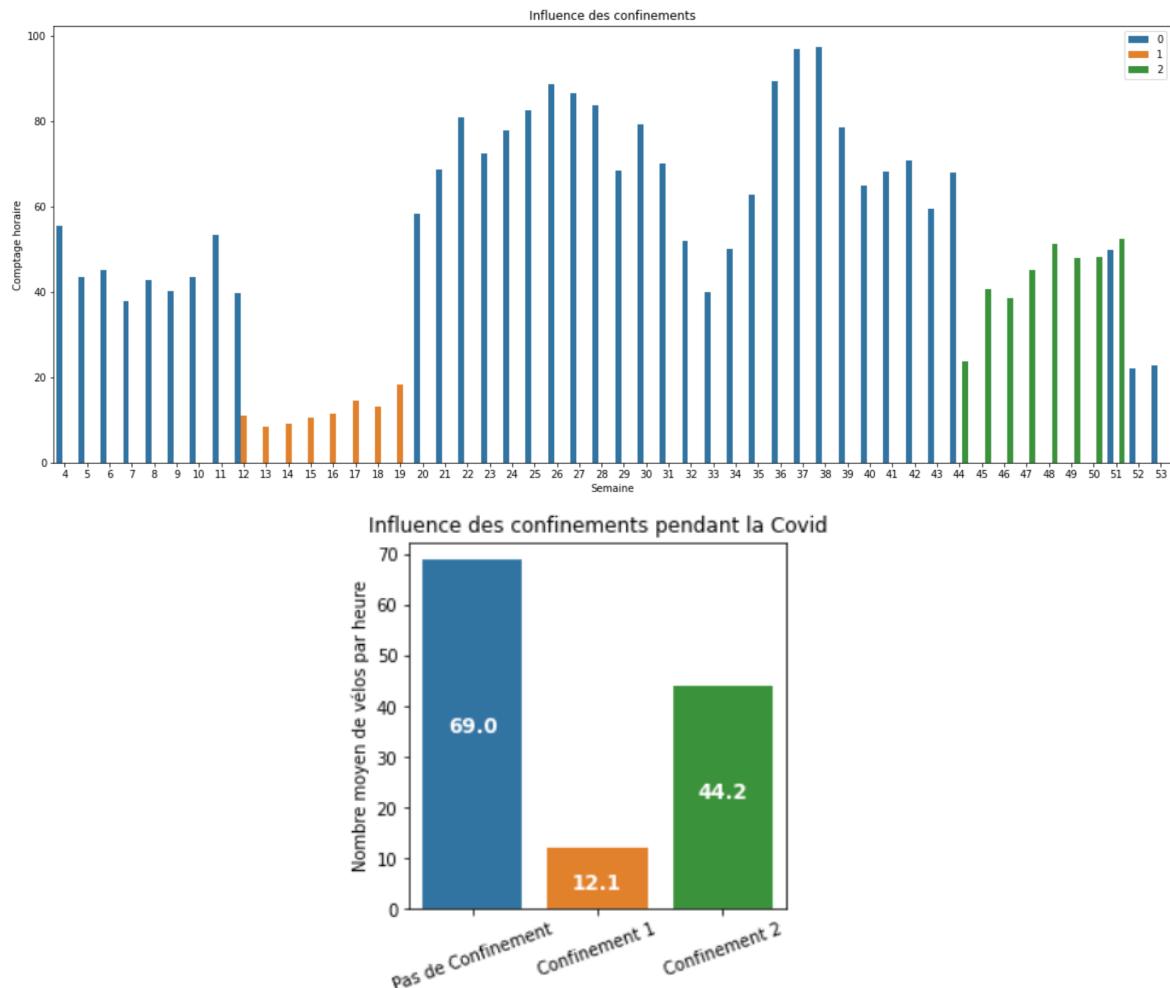
Pour distinguer les effets de l'une et de l'autre, nous étudierons ici uniquement la période hors-grève, à partir de fin janvier.



Comme nous l'avons vu précédemment, le trafic chute brutalement lors du premier confinement, avant de rebondir et d'atteindre des sommets durant l'été 2020. Regardons les chiffres avant et pendant la pandémie. Nous avons exclu la période de grève (décembre 2019 / janvier 2020).



Après le début de la pandémie, le trafic moyen a augmenté de 24%. En revanche sur la période Covid il y a eu 3 mois ½ de confinement. Quel a été leur effet ?



Les deux confinements n'ont pas du tout eu le même effet. Lors du premier confinement, le trafic moyen a chuté de 83% contre 34% pour le deuxième. En effet, celui-ci était moins sévère, avec les écoles ouvertes. Il a aussi été moins respecté.

2.9. Bilan

Sur cette période de 16 mois (septembre 2019 à décembre 2020), riche en rebondissements, voici les chiffres à retenir sur l'évolution du trafic cycliste à Paris.

Ce qui provoque une hausse :

- La grève des transports : + 50 %
- Le beau temps : + 48 %
- La pandémie Covid : + 24 %

Ce qui provoque une baisse :

- Un confinement strict : - 83 %
- Les vacances de Noël : - 57 % (- 33 % pour celles de février)
- Un jour férié : - 54 %
- Un week-end : - 40% le dimanche, - 25% le samedi
- Un confinement peu strict : - 34 %
- La pluie : - 33 %
- Des températures inférieures à 4° : - 28 %

En conclusion, sur cette période le facteur le plus pérenne est la crise de la Covid avec pour effet une hausse de 24% en moyenne du trafic cycliste à Paris. Au-delà de la progression des mobilités douces constatée depuis quelques années, la pandémie a accéléré le phénomène pour plusieurs raisons :

- Les nouvelles règles sanitaires ont imposé d'éviter au maximum les transports en commun.
- A partir de mi-mai (1er déconfinement), la Mairie de Paris a créé ou réaménagé une cinquantaine de kilomètres de pistes cyclables "temporaires" encore en place à la date de ce rapport. Les efforts ont porté en particulier sur les axes Nord/Sud (ligne 4) et Est/Ouest (ligne 1). La mesure la plus exceptionnelle a été l'interdiction pour les voitures de circuler sur la rue de Rivoli, pourtant axe majeur du trafic routier à Paris.
- De nombreuses aides de l'Etat pour faire réparer ou acheter un vélo, notamment électrique.

Nous aimerais à présent étudier l'impact que cela a pu avoir sur le nombre d'accidents impliquant des cyclistes. Nous avons à cette fin trouvé un jeu de données complémentaire.

3. BASE DE DONNÉES ACCIDENTS 2019

Les dernières données à jour et libres de droits portent sur l'année 2019. Comme il y a eu une forte augmentation du trafic en décembre 2019 avec la grève des transports, il est intéressant d'étudier les liens trafic / accidents de cyclistes à Paris sur la période septembre - décembre 2019. Il s'agit ici des accidents ayant provoqué des dommages corporels.

3.1. Exploration et traitement des données

Les étapes :

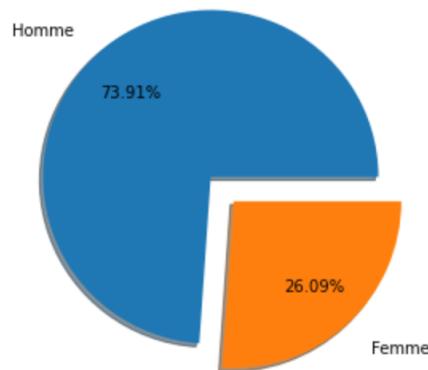
- La base de données comprend plusieurs tables, nous en sélectionnons 4 : Caractéristiques, Véhicules, Lieux, Usagers
- Nous les rassemblons par la variable identifiant les accidents.
- Nous sélectionnons les colonnes utiles à notre analyse.
- Nous filtrons sur Paris, la période (sep-déc 2019) et la catégorie Vélo dans les véhicules impliqués ('catv' = 1).
- Nous vérifions qu'il n'y a ni valeur manquante ni doublon.
- Nous convertissons le type de certaines colonnes.
- Nous créons une colonne 'Age' (2019 - année de naissance des victimes).

3.2. Data visualisation

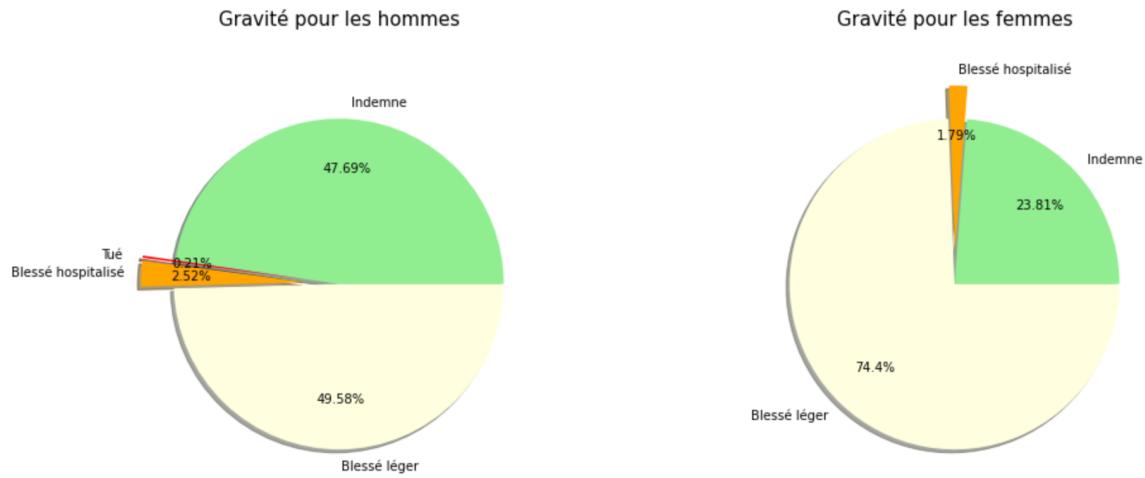
3.2.1 Analyse des données concernant les cyclistes accidentés

3.2.1.1. Par sexe

Cyclistes accidentés selon le sexe



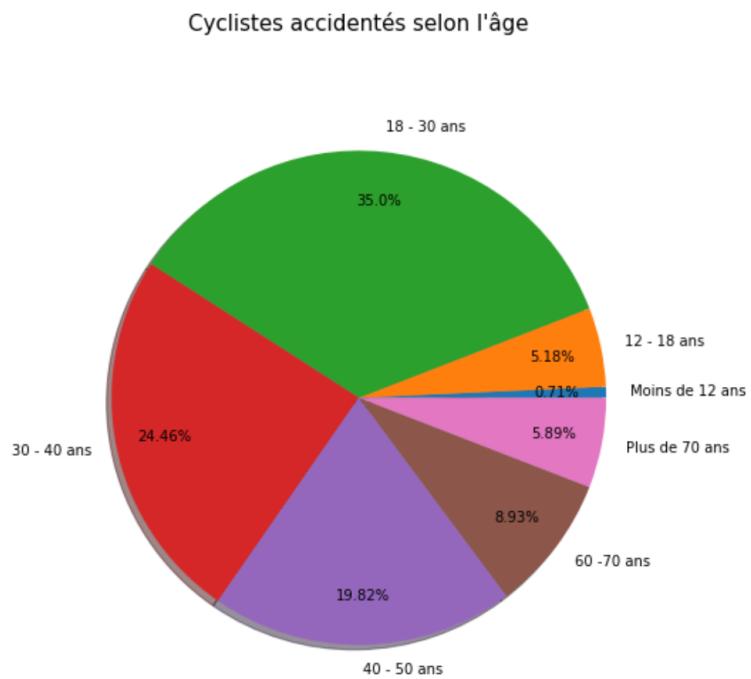
3/4 des accidentés à vélo sont des hommes contre 1/4 de femmes. Le boom des livraisons à vélo, métier majoritairement masculin, joue sur ces chiffres. Comparons la gravité des accidents en fonction du sexe :



Pour les hommes, il y a environ la moitié des accidents sans blessure et une autre moitié avec blessures légères. Pour les femmes, c'est plutôt 1/4 d'accidents sans blessure et 3/4 avec des blessures légères. Pour les deux, la proportion d'accidents entraînant une hospitalisation est assez faible, d'environ 2%. Sur la période (4 mois et 644 accidents), il n'y a eu qu'un seul mort, un homme. Ce qui fait dire à Benoît, cycliste chevronné : "Contrairement à ce qu'on pense, la probabilité de mourir en vélo à Paris est faible !". ;)

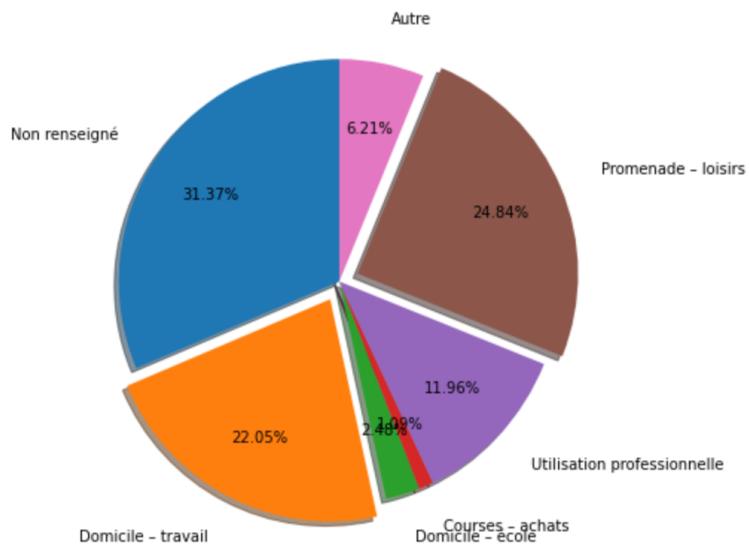
3.2.1.2. Par âge

Sans surprise la majorité des accidentés ont entre 18 et 30 ans, ce qui correspond à la tranche d'âge qui roule le plus à vélo. Puis le nombre décroît avec l'âge (et l'usage).



3.2.1.3. Par nature du trajet

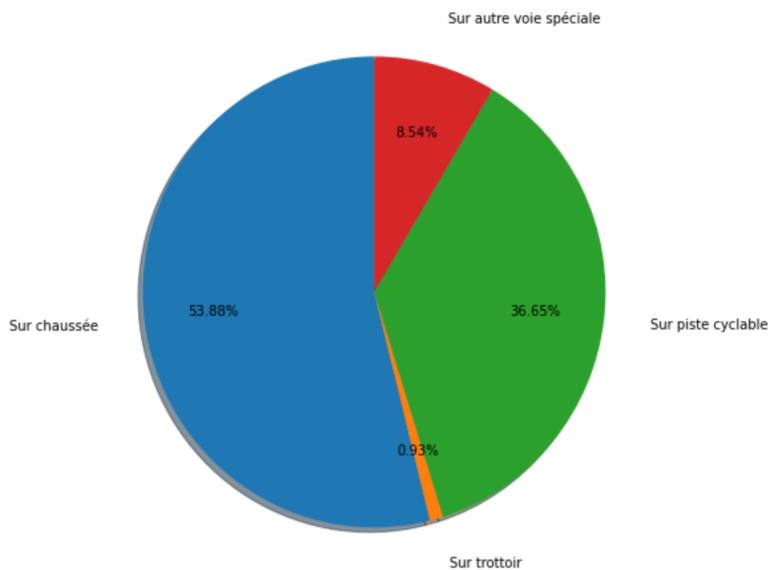
Cyclistes accidentés selon la nature du trajet



Avec 1/3 des trajets non renseignés, difficile de conclure, même si les trajets Promenade/loisirs et Domicile-travail semblent largement en tête.

3.2.1.4. Par type de voie

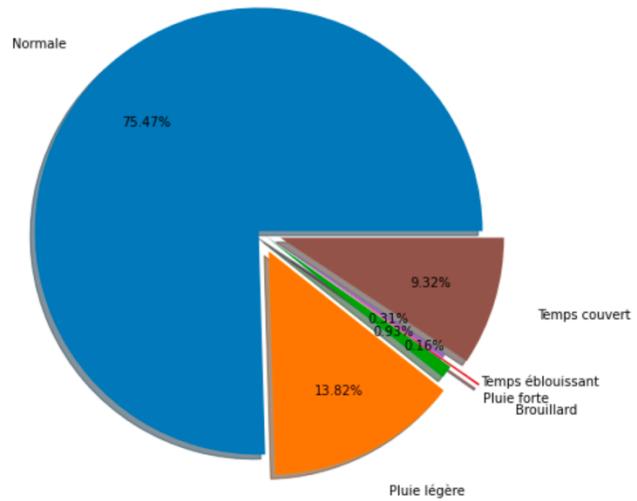
Cyclistes accidentés selon la voie utilisée



Une grosse moitié des accidents a eu lieu sur la chaussée contre 1/3 sur des pistes cyclables. Ces dernières limiteraient donc le nombre d'accidents.

3.2.1.5. Selon les conditions météo

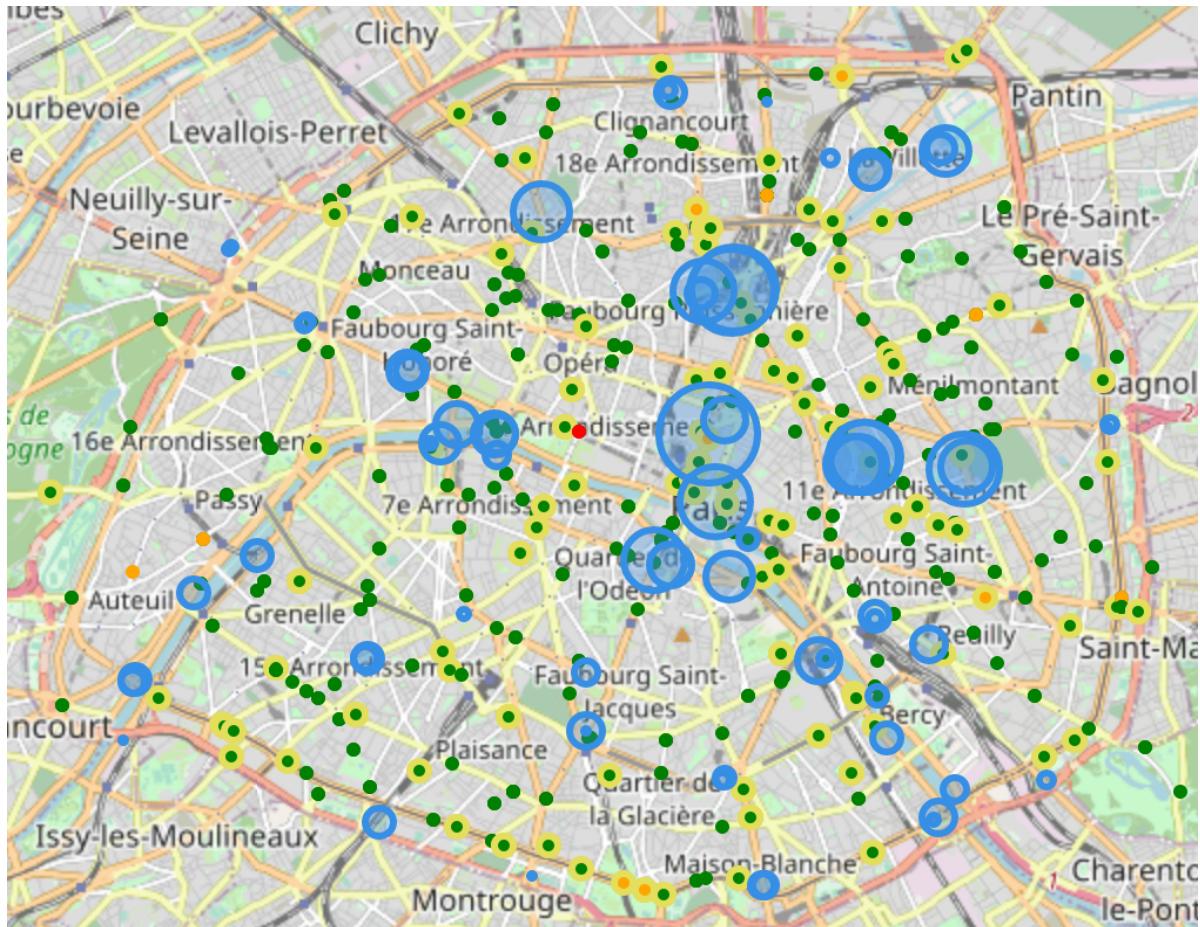
Cyclistes accidentés selon la météo



Les 3/4 des accidents ont lieu sous une météo normale. Le 1/4 restant a lieu sous la pluie ou par temps couvert.

3.3 Bilan cartographique

Carte dynamique sur laquelle on peut zoomer. Cliquez sur les cercles pour voir apparaître les adresses, identifiants et comptage moyen de chaque site.



Les données concernent la période de septembre 2019 à décembre 2019 : **644 accidents ayant entraîné des blessures**.

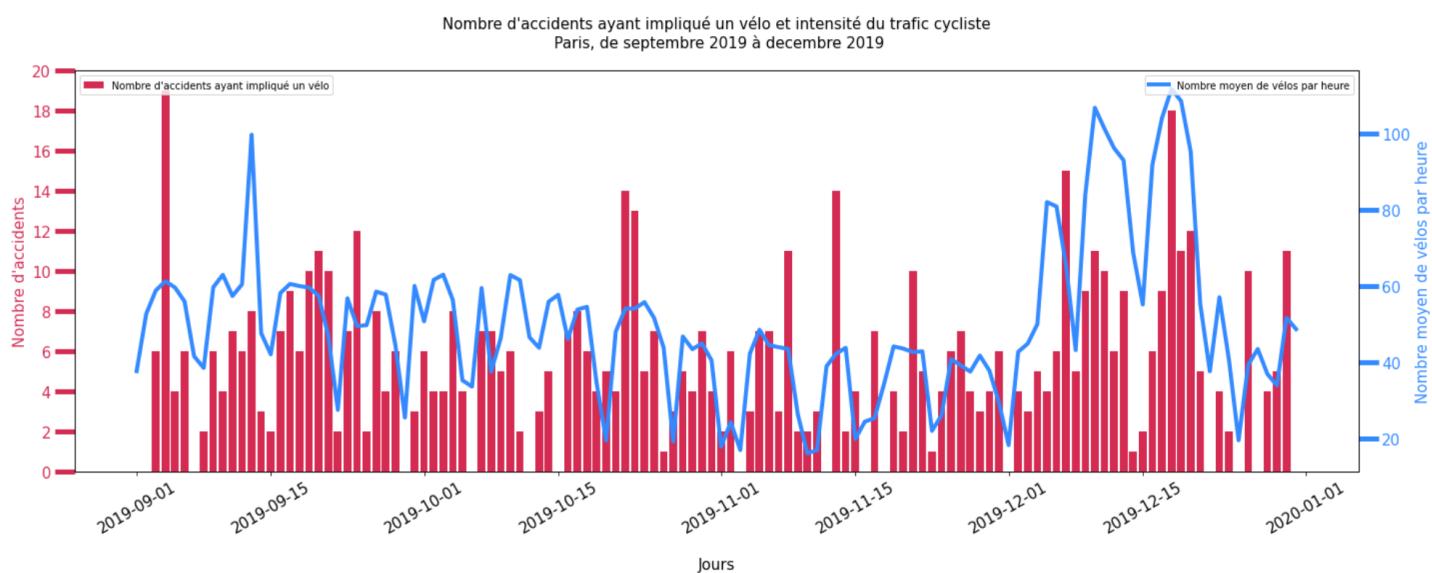
- Les sites de comptage sont représentés par les cercles bleus. Leur taille est proportionnelle à la moyenne des comptages horaires pour la période.
- Les points entourés de jaune correspondent aux accidents impliquant un vélo ayant eu lieu **sur une piste cyclable**.
- Les points non cerclés de jaune correspondent aux accidents impliquant un vélo étant survenus **hors-piste cyclable**.
- Les points verts indiquent les accidents avec cycliste **blessé léger**.
- Les points orange indiquent les accidents avec cycliste **blessé hospitalisé**.
- Les points rouges indiquent les accidents ayant entraîné le **décès d'un cycliste**.

Les accidents sont plus concentrés au niveau des grands carrefours (Opéra, Saint Lazare, Gare de l'Est, Châtelet) et sur les grands axes (Sébastopol, Convention, Lafayette, Belleville), qui sont pourtant équipés de pistes cyclables. On peut aussi noter le boulevard des Maréchaux dans le sud de Paris. Le nombre d'accidents impliquant des vélos est certainement proportionnel au nombre de voitures et à leur vitesse, plus élevée sur ces grands axes.

Il y a eu des accidents un peu partout à Paris, mais on constate que l'est de Paris est plus touché, par rapport aux 7e et 16e arrondissements par exemple, où il y a eu le moins d'accidents. Cela doit sûrement refléter la proportion d'usagers du vélo. A l'est, les quartiers sont plus jeunes et avec plus de lieux de sortie qu'à l'ouest, plus bourgeois et plus institutionnel.

3.4. Évolution du nombre d'accidents en fonction du trafic

Nous avons voulu comparer le nombre d'accidents impliquant des vélos par jour et l'intensité du trafic cycliste (nombre moyen de vélos / heure).

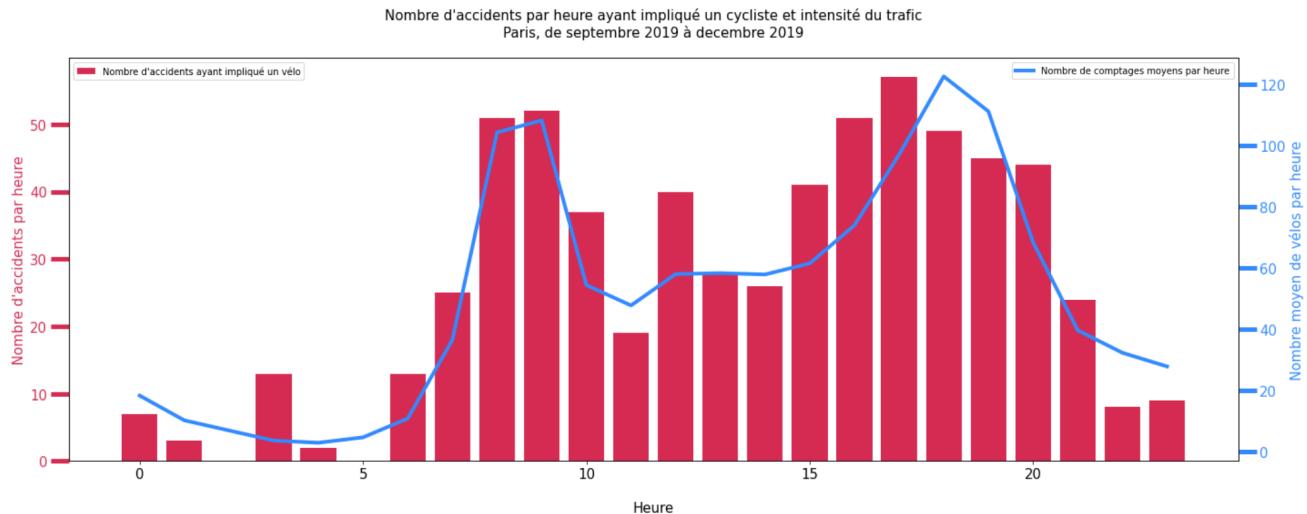


Le nombre d'accidents semble, assez logiquement, corrélé à l'intensité du trafic : les creux et pics d'accidents correspondent aux creux et pics d'intensité du trafic. On remarque cependant quelques exceptions :

- Un pic inexplicable d'accidents le 4 septembre
- Le pic de vélos enregistré le 22 septembre n'a pas engendré de hausse des accidents et pour cause : beaucoup de vélos, mais c'est la Journée sans voitures.

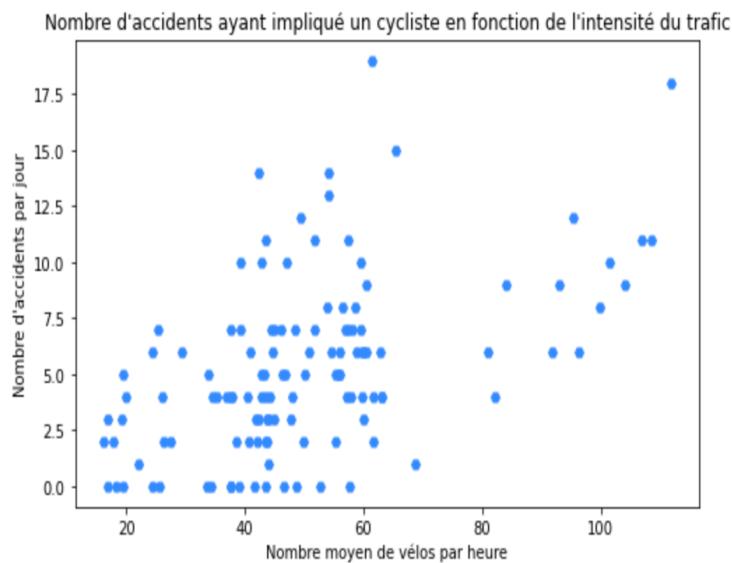
- Mi-septembre et mi-octobre, il semble y avoir une hausse inexplicable des accidents par rapport à l'intensité du trafic.
- En décembre, suite à la grève, le nombre de vélo a grimpé, mais les accidents n'ont pas augmenté dans les mêmes proportions.

Zoomons en fonction de l'heure de la journée.

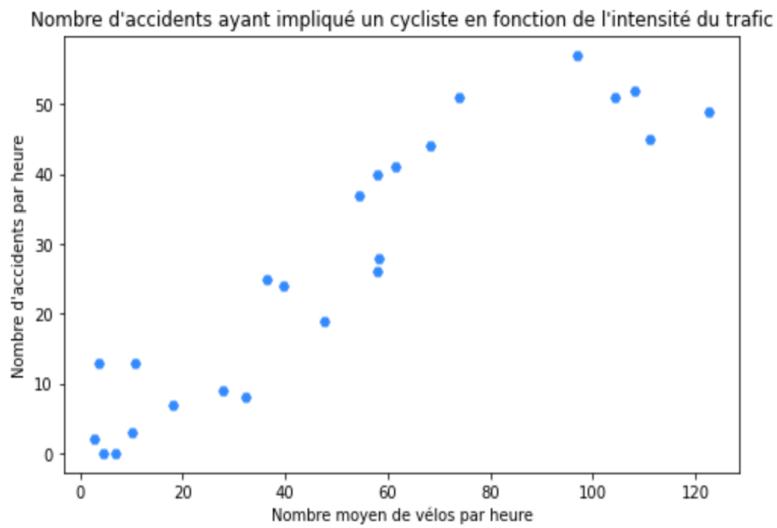


Ici encore, les deux tendances se suivent sur les heures creuses et de pointe, avec une exception à 3 heures du matin : le nombre d'accidents augmente alors que l'intensité du trafic diminue. L'heure de sortie des bars semble donc être la plus accidentogène ! Nos données ne portent que sur 644 accidents en 4 mois, cette analyse de comptoir mériterait donc d'être confirmée sur un plus grand échantillon ;)

Observons les liens de linéarité entre nombre d'accidents et intensité du trafic.



On observe une certaine linéarité entre la densité du trafic et les accidents par jour. Cependant la majorité des accidents interviennent lorsque le nombre de vélos par heure est compris entre 40 et 60, soit autour de la moyenne (50). Au-delà de 60 vélos par heure, le nombre d'accidents ne semble pas augmenter. Regardons par heure.



La linéarité est ici plus flagrante. Vérifions la dépendance statistique entre les deux variables.

3.5. Test ANOVA entre le trafic et le nombre d'accidents

La p-value (1.651967e-10) est inférieure à 5%, donc on rejette l'hypothèse selon laquelle l'intensité du trafic n'influe pas sur le nombre d'accidents.

CONCLUSION

Avec la hausse du trafic cycliste observée à Paris suite à la pandémie, le nombre d'accidents devrait selon toute logique augmenter. Retenons que lors du pic d'usagers pendant la grève, le nombre d'accidents a moins augmenté qu'attendu. Cette analyse sera à affiner quand les données 2020 seront disponibles.

4. MACHINE LEARNING

Dans cette partie nous allons créer un modèle de prédiction du comptage horaire par site.

- La tâche consiste à prédire le comptage horaire (variable cible) à partir de variables explicatives.
- La variable cible est numérique. Le type de modèle sera donc une régression linéaire.
- Pour évaluer la performance de nos modèles, nous utiliserons les métriques R^2 et RMSE.
- Nous comparerons également les prédictions aux données réelles sur des graphiques.

4.1. Premier essai

Pour ses vertus pédagogiques uniquement, nous vous remettons ici notre tout premier essai, qui était très peu concluant, mais duquel nous avons beaucoup appris ;)

- Sélection des variables qui peuvent être utiles :
 - 'Date et heure de comptage',
 - 'Identifiant du site de comptage',
 - 'Nom du site de comptage',
 - 'Année',
 - 'Mois',
 - 'Semaine',
 - 'Jour',

```
'Jour_de_la_semaine',
'Weekend',
'sam_dim'
'Heure',
'lat',
'long',
'Jour_férié',
'Vacances',
'vac_fevrier',
'vac_printemps',
'vac_ascension',
'vac_juillet',
'vac_aout',
'vac_toussaint',
'vac_noel',
'Pluie',
'Froid',
'Chaud',
'Grève',
'Covid',
'Confinement'
```

- Tri par ordre chronologique (Année, Mois, Jour, Heure)
- On retire les colonnes 'Date et heure de comptage', 'Identifiant du site de comptage', 'Nom du site de comptage'
- **Standardisation : les résultats étaient les mêmes avec ou sans**
- Création des ensembles train (80%) et test (20%) avec la fonction train_test_split et le paramètre shuffle = False pour respecter la chronologie
- Création d'un modèle LinearRegression ajusté sur l'échantillon train
- Evaluation : R² train/test = 0.17 / -0.59 - rmse train/test = 108.5 / 134.5

Les scores étaient très mauvais pour plusieurs raisons :

- Nous n'avions pas de véritables variables explicatives numériques, nécessaires à un modèle de RL. A part les variables temporelles et les coordonnées géographiques, toutes les variables sont catégorielles : Vacances, Covid, Confinement etc.
- La régression linéaire suppose des relations linéaires entre variables explicatives et cible. Ce qui n'était pas le cas ici, comme un pairplot nous l'a montré.
- Les variables avaient très peu de corrélation avec la variable cible, comme nous l'a prouvé une heatmap.
- La période choisie (16 mois, dont 20% en test), peut-être trop longue.

Forts de ce premier échec, nous avons décidé :

- De créer 10 variables explicatives numériques : 'Comptage horaire' à Heure -1, -2, -3 / Jour -1, -2, -3 / Semaine -1, -2, -3, -4
- De tester des modèles sur chaque mois : apprentissage jusqu'au 23 et prédictions à partir du 24.
- Qu'il n'était pas utile de standardiser les données, d'autant que les variables créées sont toutes des comptages horaires, donc du même ordre que la variable cible.

4.2. Ajout des variables numériques

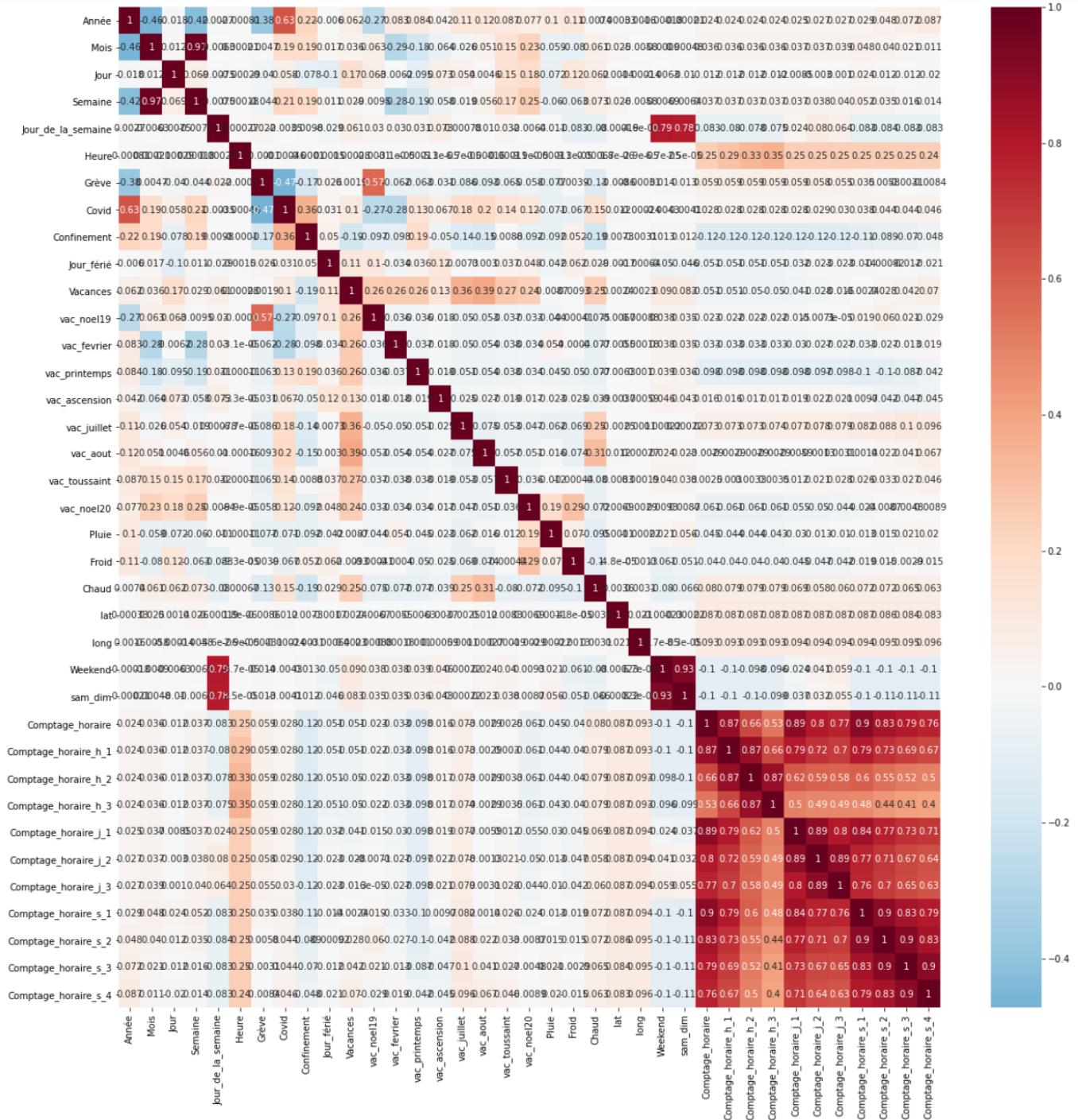
Les étapes pour chaque variable :

- création d'une fonction pour atteindre l'heure -1 etc.
- ajout de la variable date/heure heure-1
- création d'un df h-1 en ne conservant que "date et heure de comptage" et "comptage horaire"
- renommer la colonne 'Comptage_horaire' en 'Comptage_horaire_h_1'
- suppression des colonnes inutiles
- fusion des df
- suppression de la colonne "h-1"
- **remplacement des valeurs manquantes par 0 ***
- mise au format "int" de "Comptage_horaire_h_1"

* Durant la période, certains sites apparaissent (création) et disparaissent (travaux, pannes). La moitié des 69 sites ne sont pas présents sur les 15 mois.

- Pour les sites présents sur toute la période, nous avons "réservé" septembre 2019 comme historique des variables (qui remontent sur 4 semaines). Les modèles ne seront entraînés qu'à partir d'octobre 2019.
- Pour les sites qui ont été ajoutés plus tard, nous avons pris le parti de remplacer les valeurs manquantes créées (puisque l'historique n'existe pas) par des 0. Après des tests de prédiction, cela ne semblait pas gêner le modèle.

4.3. Analyse de la corrélation / linéarité entre les variables

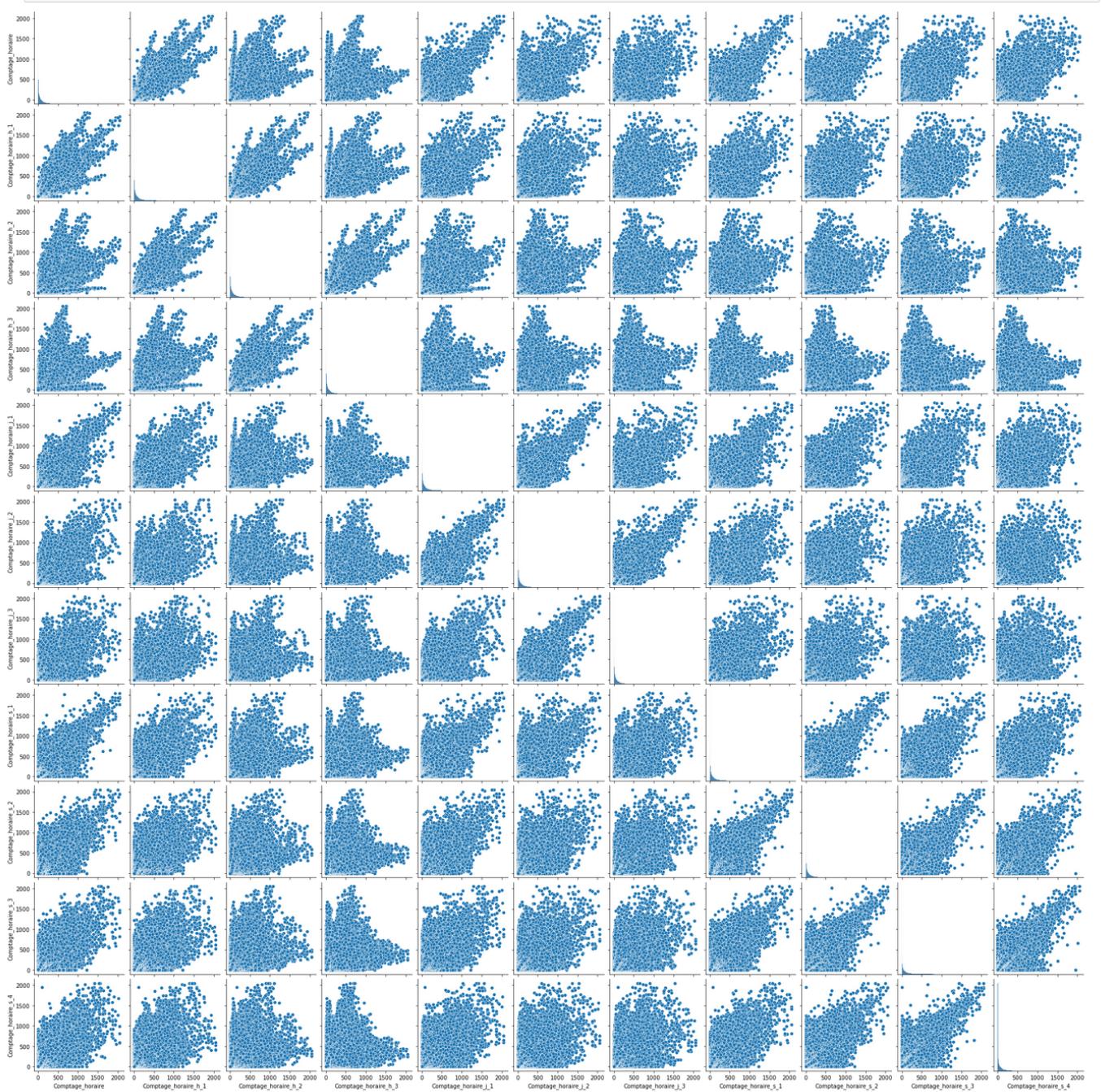


Comme nous le montre la heatmap, ce sont sans surprise les variables numériques créées qui sont le plus fortement corrélées avec la variable cible, dans cet ordre :

- Comptage_horaire_semaine_1 : 0.95
- Comptage_horaire_jour_1 : 0.89
- Comptage_horaire_heure_1 : 0.87
- Comptage_horaire_semaine_2 : 0.83
- Comptage_horaire_jour_2 : 0.8
- Comptage_horaire_semaine_3 : 0.79
- Comptage_horaire_jour_3 : 0.77
- Comptage_horaire_semaine_4 : 0.76
- Comptage_horaire_heure_2 : 0.66
- Comptage_horaire_heure_3 0.53

Les autres variables ont une corrélation négligeable, le maximum étant de **0.25** pour **Heure**.

A priori, nous ne retiendrons que les 10 variables numériques créées. Observons leurs relations de linéarité avec un pairplot.



Toutes les variables numériques créées ont une linéarité plus ou moins évidente avec la cible. Ce qui n'est pas étonnant vu qu'il s'agit de comptage horaire à différents moments.

4.4. Prédiction sur les derniers jours de chaque mois

Nous commencerons par entraîner nos modèles sur le début de chaque mois (du 1er au 23) afin d'en prédire la fin (du 24 à la fin du mois) pour tous les mois d'octobre 2019 à décembre 2020.

4.4.1. Définition de la taille de l'échantillon test

Les étapes :

- Tri du df en fonction du jour
- Définition d'une variable 1er jour de l'échantillon test (à corriger pour modifier la taille de l'échantillon)
- Calcul du % à prendre pour la taille de l'échantillon test

Exemple : si on veut que l'échantillon test démarre le 24 du mois, il faut une taille de 0.249.

4.4.2. Choix des variables avec SelectKBest

Nous gardons d'abord toutes les variables et créons les 2 échantillons 'train' et 'test'. Pour confirmer notre avis sur le choix des variables, nous créons ensuite un modèle **SelectKBest**, qui va toutes les tester, avec pour paramètre $k = 10$. Sans surprise, il retient les 10 variables numériques créées, qui sont les plus corrélées. Dans les prochains tests nous allons donc les utiliser. Nous verrons ensuite si nous pouvons en éliminer certaines.

4.4.3. Création des échantillons 'train' et 'test'

A partir de nos 10 variables numériques et de notre formule `test_size`.

4.4.4. Choix du modèle

Nous avons choisi de tester 4 modèles de régression linéaire (`LinearRegression`, `RidgeCV`, `LassoCV`, `ElasticNet`) et 4 modèles de régression non linéaire (`DecisionTreeRegressor`, `RandomForestRegressor`, `BaggingRegressor`, `BaggingRegressor`).

Pour les évaluer et choisir le modèle le plus performant, nous avons exécuté une boucle for qui affiche le score R^2 et la rmse pour l'ensemble d'apprentissage et de test.

Chaque algorithme a été testé seul en amont, et les hyperparamètres choisis pour chacun d'entre eux sont une alternative entre résultat et temps de réponse :

- **LinearRegression** : hyperparamètres de base

Il ne semble pas exister d'hyperparamètre qui modifierait significativement la performance du modèle.

```
score R2 train = 0.9189962817909223 / score R2 test = 0.9057989819099554  
rmse train = 34.19368224500649 / rmse test = 33.67896824669543
```

- **RidgeCV** : `alphas` = (0.001, 0.01, 0.1, 0.3, 0.7, 1, 10, 50, 100)

L'algorithme a un temps de réponse satisfaisant et nous permet de tester le paramètre « `alphas` » en validation croisée

```
score R2 train = 0.9189962817908839 / score R2 test = 0.9057989623886029  
rmse train = 34.19368224501459 / rmse test = 33.67897173635504
```

- **LassoCV** : `alphas` = (0.001, 0.01, 0.1, 0.3, 0.7, 1, 10, 50, 100)

L'algorithme a un temps de réponse satisfaisant et nous permet de tester le paramètre « `alphas` » en validation croisée.

```
score R2 train = 0.9189958645588665 / score R2 test = 0.9057879540708655  
rmse train = 34.193770306902195 / rmse test = 33.68093953841202
```

- **ElasticNet** : hyperparamètres de base

Nous avons choisi de tester uniquement sur les hyperparamètres de base pour avoir un temps de réponse satisfaisant

score R² train = 0.918996111843483 / score R² test = 0.9057894981011902
rmse train = 34.19371811450619 / rmse test = 33.680663540769814

- **DecisionTreeRegressor** : hyperparamètres de base

Nous avons choisi de tester uniquement sur les hyperparamètres de base pour avoir un temps de réponse satisfaisant

score R² train = 0.9999206586740267 / score R² test = 0.8753398534447733
rmse train = 1.0701462194863858 / rmse test = 38.743138291545286

- **RandomForestRegressor** : criterion = « mse », « n_estimators » = 10

Pour l'hyperparamètre « criterion », seul « mse » permet d'avoir un temps de réponse satisfaisant. Pour l'hyperparamètre « n_estimators » plus il est élevé plus le modèle est performant. Au-delà de 10 le modèle n'a plus un temps de réponse satisfaisant.

score R² train = 0.9911091995495411 / score R² test = 0.9292202313175566
rmse train = 11.328276113216926 / rmse test = 29.19345860082831

- **BaggingRegressor** : n_estimators = 10

Plus le « n_estimators » est élevé plus le modèle est performant. Au-delà de 10 le modèle n'a plus un temps de réponse satisfaisant.

score R² train = 0.9912357352928903 / score R² test = 0.9295880725445531
rmse train = 11.24737402275786 / rmse test = 29.117500836138777

- **GradientBoostingRegressor** : (n_estimators = 100)

Plus le « n_estimators » est élevé plus le modèle est performant. Au-delà de 100 le modèle n'a plus un temps de réponse satisfaisant.

score R² train = 0.9366186472660232 / score R² test = 0.9183999299350663
rmse train = 30.246431701602912 / rmse test = 31.345575369112066

CONCLUSION

Le modèle le plus robuste, c'est-à-dire qui overfitte le moins, et qui prend le moins de temps de calcul est tout simplement la Régression Linéaire :

R² train/test : 0.92 / 0.91 - rmse train/test : 34.2 / 33.7

4.4.4. Test sur les variables optimales

Avec une régression linéaire, nous pouvons utiliser 'SelectFromModel' pour voir s'il nous propose de réduire le nombre de variables. Résultats : 'Comptage_horaire_h_1', 'Comptage_horaire_h_2', 'Comptage_horaire_j_1', 'Comptage_horaire_s_1'

Nous testons la RL avec ces seules variables :

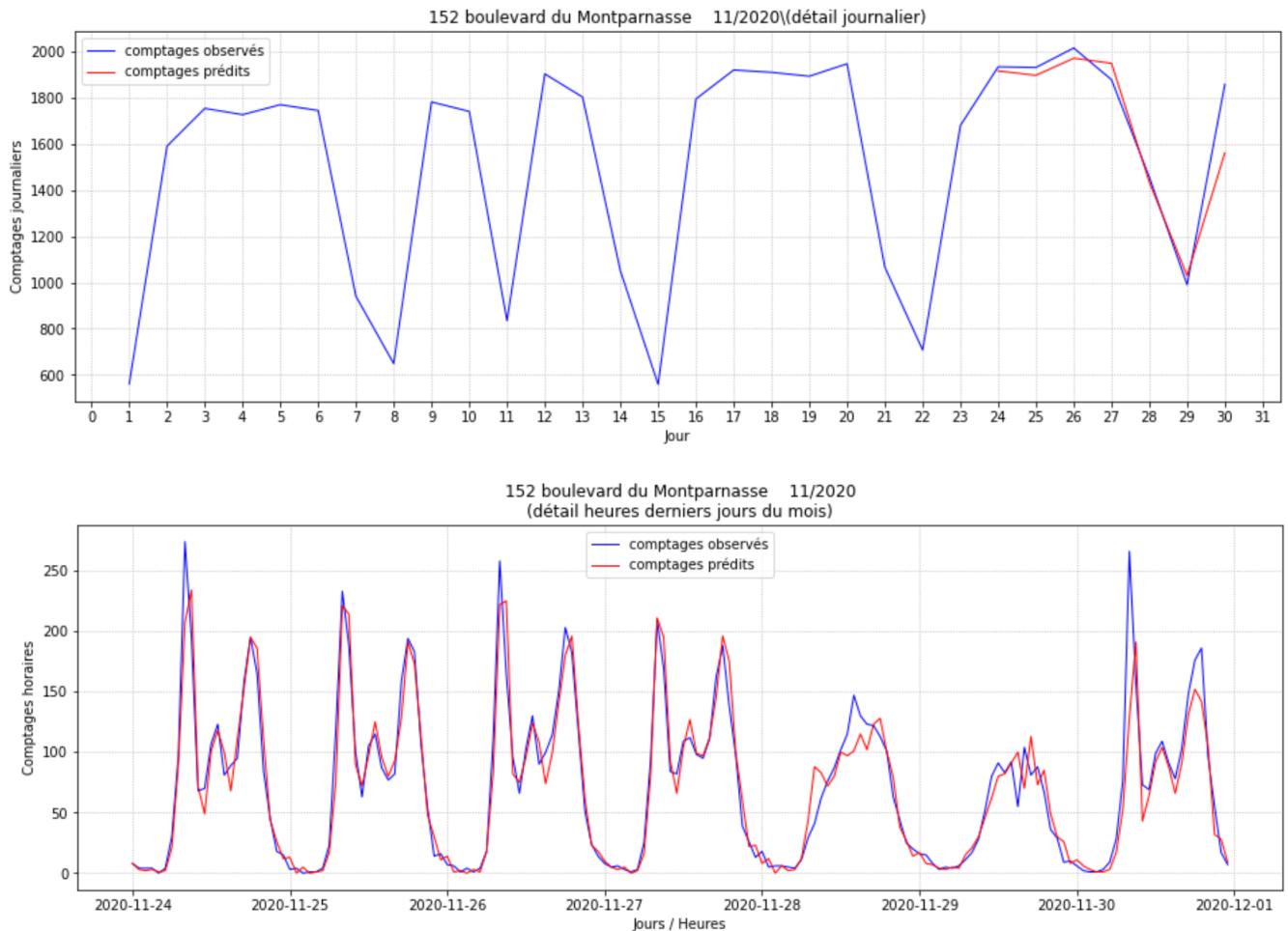
R² train/test = 0.91 / 0.90 - rmse train/test = = 35.30 / 34.76

Cela n'améliore pas le modèle, donc nous gardons les 10 variables numériques.

4.4.5. Représentation graphique des prédictions

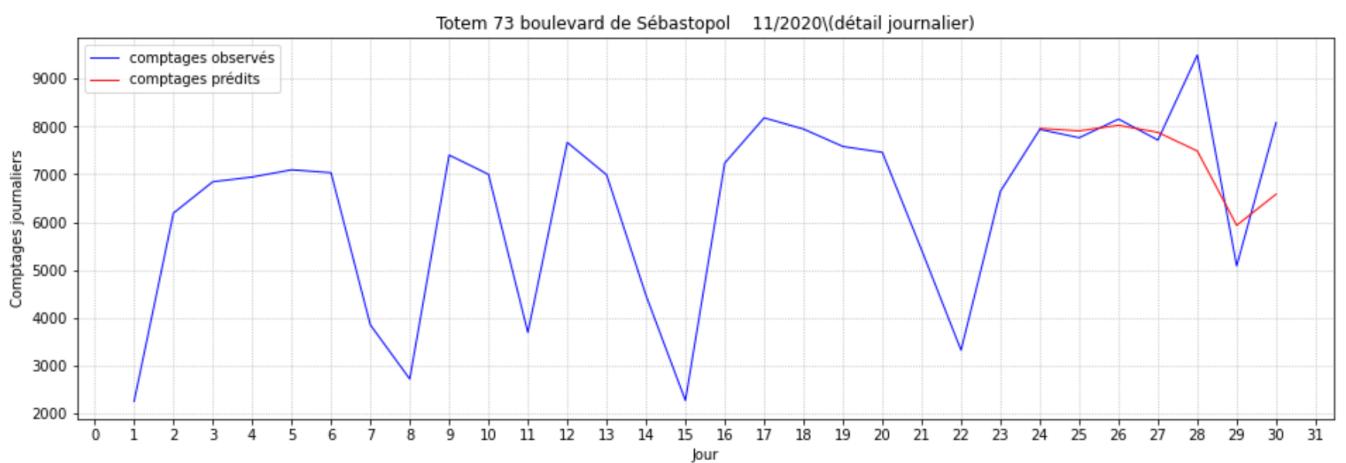
On définit des variables à remplir pour les prédictions : **année, mois, identifiant_site**. Il ne reste qu'à remplir ces champs pour afficher les prédictions de n'importe quel site (ou tous) pour un mois donné. On propose ensuite 2 graphiques : un à l'échelle du mois avec comptage journalier, l'autre à l'échelle de la dernière semaine avec un comptage horaire.

Par exemple, le site du 152 bd Montparnasse ('100049407') en novembre 2020.

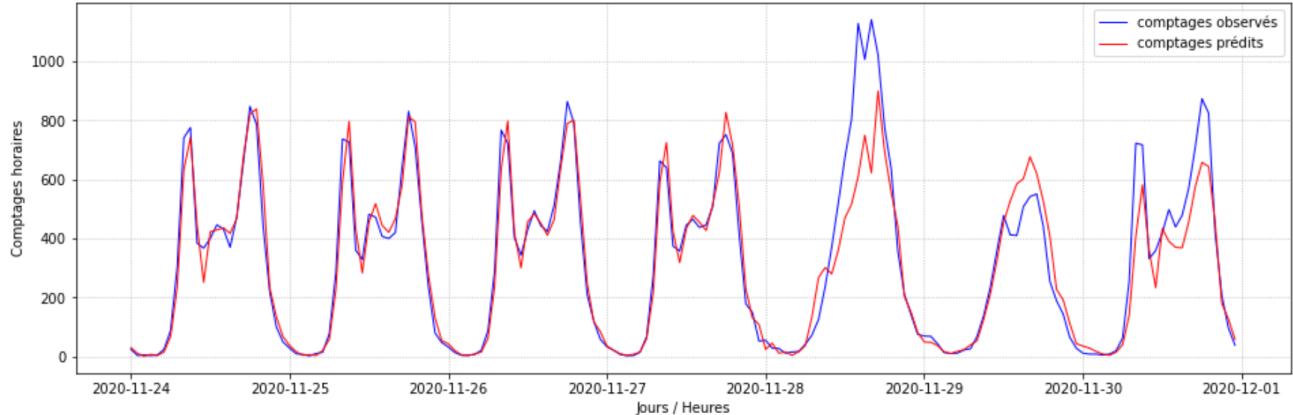


Dans cet exemple, les prédictions sont très justes, autant à l'heure qu'à la journée.

Mais qu'en est-il pour l'un des sites les plus fréquentés : 73 boulevard Sébastopol ('100057445'), toujours en novembre ?



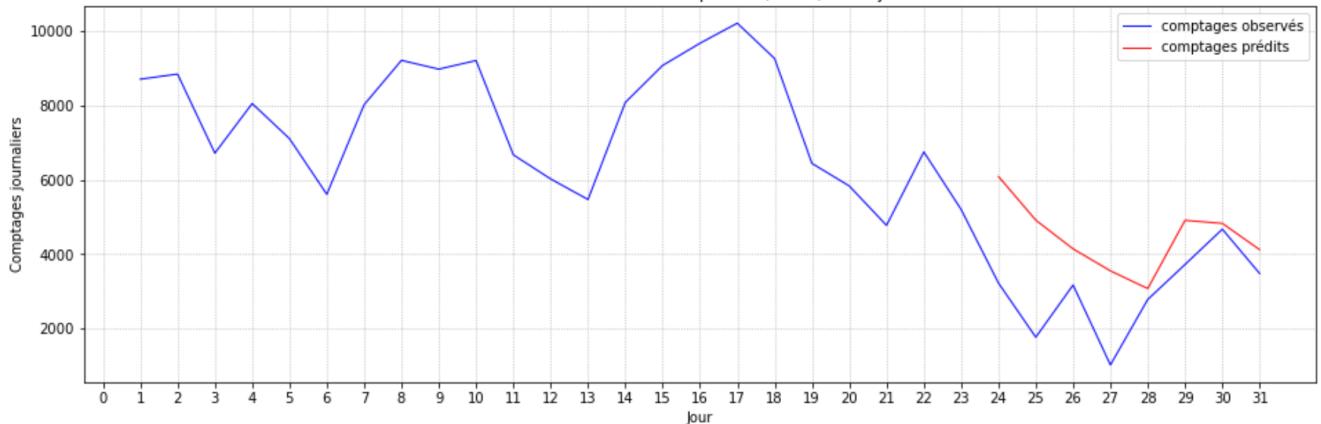
Totem 73 boulevard de Sébastopol 11/2020
(détail heures derniers jours du mois)



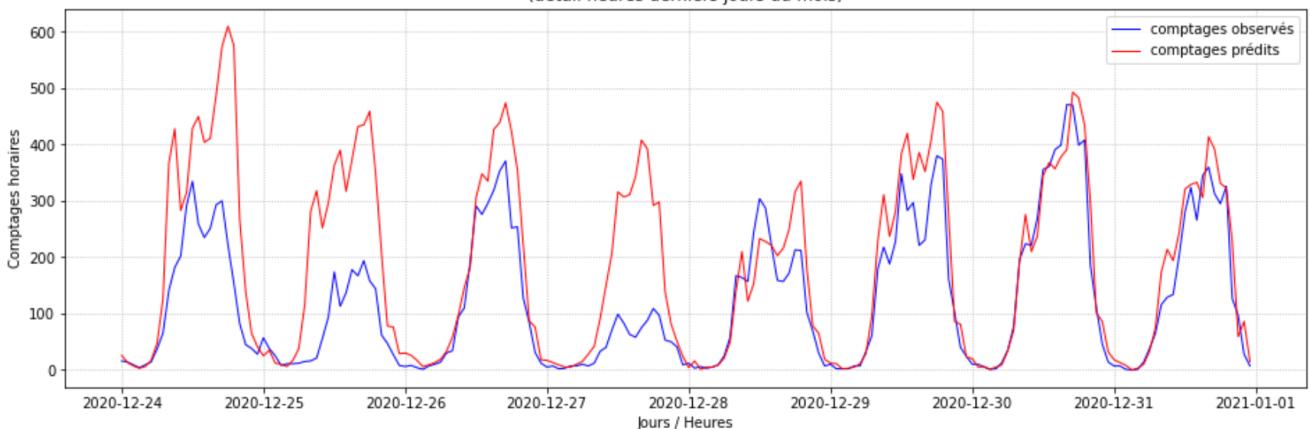
Les prédictions à l'heure sont toujours fiables, bien que sous-estimés en fin de semaine. A la journée l'écart se prononce un peu, car les erreurs s'accumulent.

Et le même site en décembre ?

Totem 73 boulevard de Sébastopol 12/2020(détail journalier)

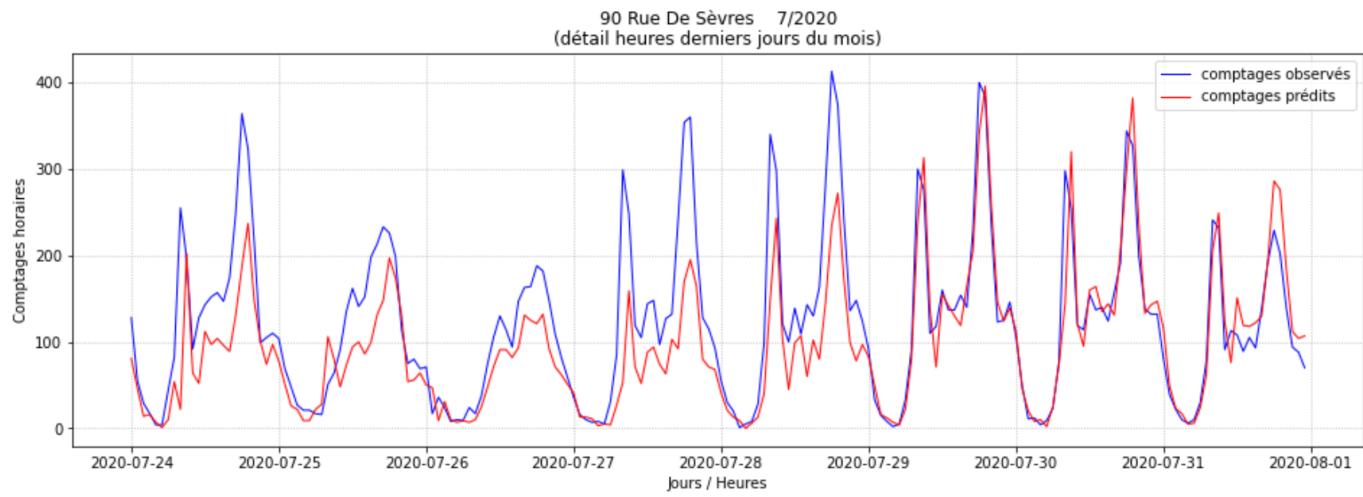
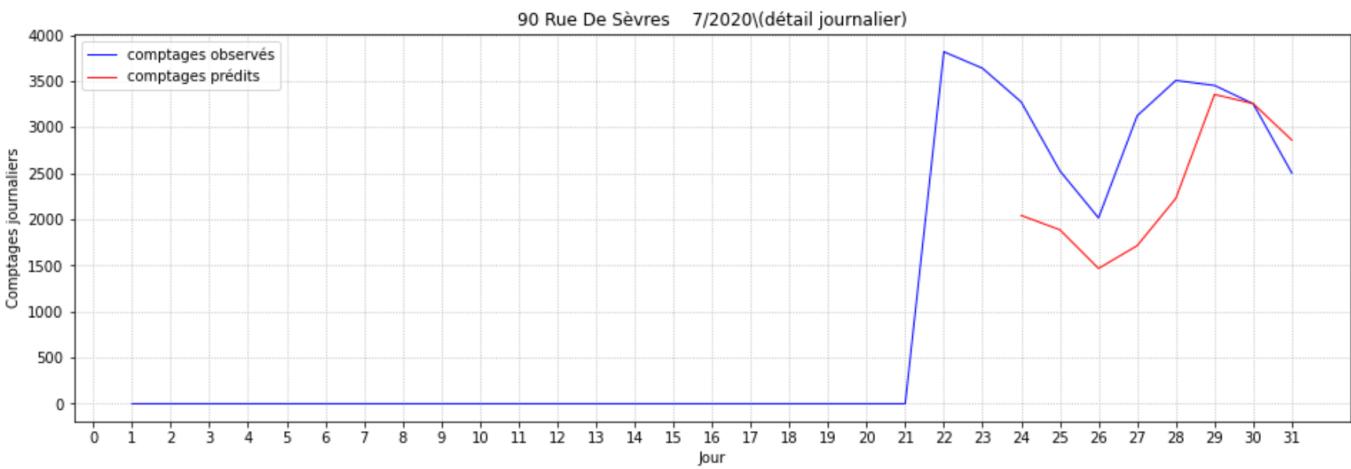


Totem 73 boulevard de Sébastopol 12/2020
(détail heures derniers jours du mois)



Le modèle prédit toujours bien le comptage à l'heure. La surestimation du début de semaine est due aux fêtes de Noël, dont nous connaissons maintenant l'importance.

Et pour les sites créés en cours d'année, comme le 90 rue de Sèvres ('100060178') en juillet 2020 ?



Sans l'historique sur les 3 premières semaines, le modèle sous-estime le début de la semaine, mais se rattrape ensuite.

CONCLUSION 1ER MODÈLE

Les comptages horaires sont très fiables. Le modèle comprend bien les tendances quotidiennes et hebdomadaires, mais est sensible aux imprévus comme les vacances. Or, nous avons vu dans la partie Data Viz l'importance qu'elles ont sur le trafic (- 57%).

Entraînons maintenant notre modèle sur une année complète (octobre 2019 - novembre 2020) pour prédire les 3 derniers (octobre - décembre 2020).

4.5. Prédictions sur les 3 derniers mois

4.5.1. Définition de la taille de l'échantillon test

Ici on définit la taille de l'échantillon 'test' pour qu'il corresponde à octobre 2020.

4.5.2. 2e modèle et évaluation

Scores sur les derniers mois :

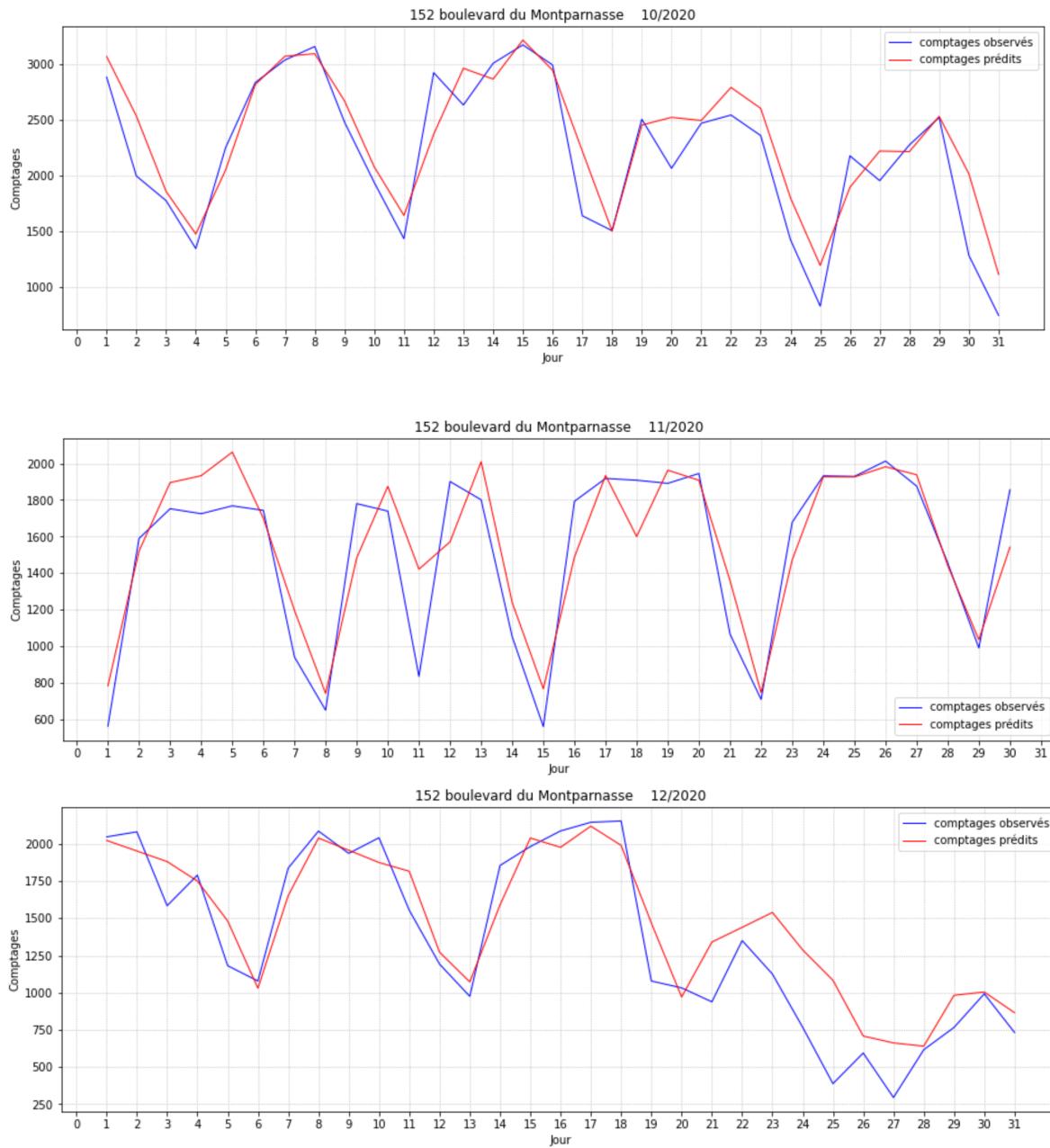
R² train/test = 0.92 / 0.91 - rmse train/test = 34.5 / 32.2

Scores sur les derniers jours du mois :

R² train/test : 0.92 / 0.91 - rmse train/test : 34.2 / 33.7

Le nouveau modèle est quasi aussi performant que le précédent, avec un écart un peu plus important sur les erreurs.

4.5.3. Représentations graphiques



CONCLUSION 2E MODÈLE

Les prédictions sur les derniers mois sont très fiables. Encore une fois, c'est sur la période de Noël que le modèle surestime le trafic. Nous allons donc rajouter des variables catégorielles liées aux événements récurrents (vacances, jours fériés etc.) ou exceptionnels (grève, confinement etc.) pour voir si cela l'aide à s'ajuster.

4.5.4. Ajout de variables catégorielles

Après plusieurs essais que nous n'avons pas laissés ici, les prédictions les plus justes se font avec l'ajout des variables : [vac_noel](#), [jours_fériés](#), [jour_de_la_semaine](#).

4.5.5. 3e modèle et évaluation

Scores sur les derniers mois avec les variables numériques + catégorielles :

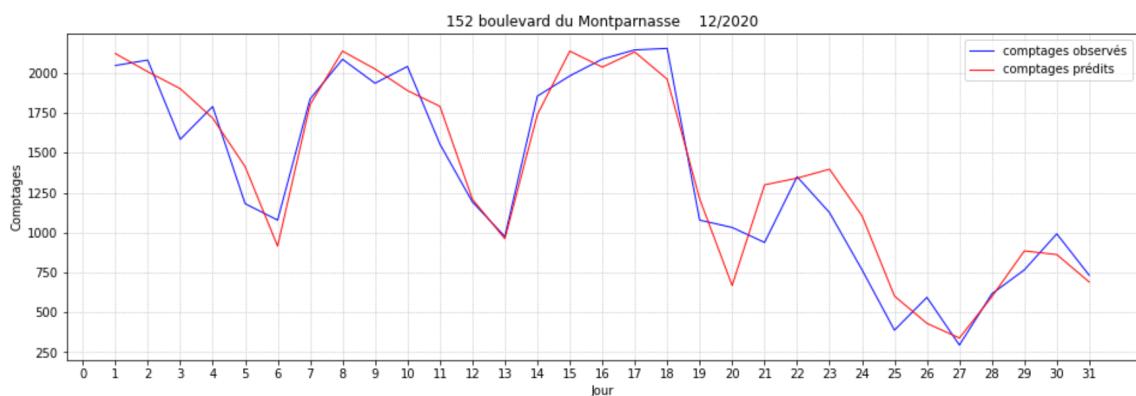
R² train/test = 0.92 / 0.91 - rmse train/test = 34.2 / 31.7

Scores sur les derniers mois avec les seules variables numériques :

R² train/test = 0.92 / 0.91 - rmse train/test = 34.5 / 32.2

Les performances se valent. Regardons s'il y a une différence dans la représentation graphique du mois de décembre 2020.

4.5.5. Représentation graphique pour décembre 2020



CONCLUSION 3E MODÈLE

Pour décembre 2020, l'ajout des variables catégorielles liées aux vacances de Noël, jours fériés et jours de la semaine corrige les erreurs. Sur les 3 derniers mois de l'année, les prédictions sont très justes.

L'étape suivante serait de tester des modèles permettant de faire des prévisions, comme les Réseaux de neurones par exemple.

Description des travaux réalisés

Le diagramme de Gant est en annexe. Benoît et Cynthia ont commencé à deux le projet, avant que Joséphine ne les rejoigne fin décembre. Côté timing, Benoît a toujours été le plus en avance dans les cours, c'est donc lui qui a passé le plus de temps sur le projet, mais nous nous sommes tous énormément investis. Nous avons préféré travailler chacun sur le maximum d'aspects du projet, plutôt que de scinder toutes les tâches en trois. Toutes les décisions ont été prises collectivement. Notre force a été la complémentarité : si nous avons tous travaillé sur plusieurs tâches, chacun a apporté ses points forts au collectif : le code pour Benoît, l'analyse et la rédaction pour Cynthia, l'apport des données extérieures et le recul critique pour Joséphine.

Bibliographie

Méthodologie cartographie :

<https://python-visualization.github.io/folium/modules.html>

Analyse accidents

<https://lejournal.cnrs.fr/billets/femmes-et-hommes-sont-ils-equaux-a-velo>

<https://fr.statista.com/themes/4163/le-secteur-du-velo-en-france/>

<https://www.citycle.com/16621-un-avenir-prometteur-pour-le-velo-un-secteur-dactivite-florissant/>

Streamlit

<https://medium.com/nightingale/building-an-interactive-dashboard-in-less-than-50-lines-of-code-494b30a31905>

<https://docs.streamlit.io/en/stable/index.html>

Difficultés rencontrées lors du projet

Benoit :

J'ai éprouvé des difficultés au départ du projet à visualiser ce qui était attendu en termes de Machine Learning. Je n'avais aucune compétence dans le domaine à ce moment-là. Ce n'est qu'après de nombreux tests sans résultats, beaucoup de temps passé, et l'aide de notre mentor pour nous réorienter dans la bonne direction que nous avons pu arriver au résultat attendu. La partie Machine Learning a certainement mobilisé beaucoup de temps au détriment du reste du projet.

Joséphine :

Le changement de projet au cours de la formation a été la plus grosse difficulté que j'ai rencontré. J'ai d'abord travaillé avec Gilles sur un projet de Covid, où j'ai vite été la seule personne à réellement "coder". Au bout de 3 semaines, comme Gilles ne se manifestait pas j'ai dû changer de groupe et recommencer depuis le début. J'ai donc dû trouver mes marques avec un groupe qui avait déjà beaucoup avancé sur le projet.

Au sein des deux projets, j'ai rencontré des difficultés liées à mes compétences : les difficultés n'étaient souvent pas celles vues dans le cours. Sur le premier projet, j'ai eu des difficultés sur le data-processing notamment pour une variable qui me posait problème..

Concernant le projet vélo, j'ai eu du mal à bien comprendre les codes qui avaient déjà été conçus et à bien apprivoiser la base de données. Quand je me suis concentrée sur l'ajout des variables extérieures j'ai voulu mettre en pratique ce que nous avions appris avec le webscraping. Malheureusement je n'ai jamais réussi à webscrappé les données des sites de météo. J'ai donc décidé de les rajouter à la main ce qui m'a pris beaucoup de temps.

Cynthia :

Ma principale difficulté aura été la gestion du temps. Les cours ont été plus chronophages que prévu. J'aurais aimé passer plus de temps et plus tôt sur le projet. Comme Benoît, c'est la partie Machine learning qui m'a posé un problème de compréhension. Il nous aura fallu bien du temps pour en voir le bout ! Avec la frustration de mon côté de ne pas avoir eu le temps de chercher un autre modèle, capable de faire des prévisions. Mon obsession de départ était de développer un outil utile pour la Mairie de Paris !

Bilan & Suite du projet

En quoi votre projet a-t-il contribué à un accroissement de vos connaissances scientifiques ?

Le projet nous a permis de renforcer nos connaissances scientifiques tant par le raisonnement que nous avons adopté tout au long du travail que par l'apprentissage que nous en avons tiré.

En effet, le travail concret sur le projet nous a servi à mettre en perspective toutes les notions vues en cours (data processing, data viz, machine learning ...) mais pas seulement ! Nous avons appris à emprunter d'autres chemins que ceux appris : notamment avec l'ajout des variables temporelles (h-1, h-2...), les dataviz géographiques avec une carte, l'ajout de variables extérieures (les accidents, la météo...) et le machine learning. Autant de méthodes que nous avons dû construire seuls.

Nous avons donc testé de nombreuses techniques, recherché chacun de notre côté des solutions pour les difficultés auxquelles nous étions confrontées, et mis en commun nos recherches au sein de discussions quotidiennes.

En plus des connaissances du cours nous avons donc accru nos connaissances scientifiques au travers de la recherche de nouvelles techniques, la réflexion sur la pertinence, l'expérimentation avec la mise en pratique et l'échange.

Pour chacun des objectifs du projet, détaillez en quoi ils ont été atteints ou non.

1/ Dans la partie Data visualisation, nous cherchions à comprendre :

- comment le trafic évolue au fil des mois, des semaines et des jours ?
- quels événements (vacances, jours fériés, grèves, confinements) impactent le plus le trafic ?
- quel est l'effet sur le nombre d'accidents impliquant des vélos ?

Les objectifs que nous avions posés en Data Visualisation ont tous été atteints et même dépassés. Nous avons réussi à étudier le trafic de cyclistes au fil des mois, semaines, jours et même des heures. Concernant les événements nous avons pu analyser l'impact des événements récurrents (vacances, fériés etc.) et exceptionnels (grève, Covid etc.). Enfin nous avons réussi à travailler sur les accidents impliquant les vélos grâce à une base de données extérieure. Sur tous ces points, nous avons tiré des enseignements qui pourront être utiles à la ville de Paris.

2/ Dans la partie Machine learning, nous avons voulu créer un modèle de prédiction calculant le comptage horaire par site, sur une période donnée.

Après avoir testé plusieurs choix de variables et d'algorithmes de régression, nous avons réussi à créer un modèle assez robuste avec de bons scores. Les derniers tests effectués sont concluants.

S'ils ont été atteints, dans quel(s) process(es) métier(s) votre modèle peut-il s'inscrire ? Dans le cas contraire, quelles pistes d'amélioration suggérez-vous pour améliorer les performances de votre modèle ?

Un modèle de régression linéaire appliqué à ce jeu de données ne permet pas de faire des prévisions. Or c'est ce qu'il y a de plus utile pour une collectivité ou un institut de recherche sur les transports par exemple. L'étape suivante est donc de créer un modèle capable de prévoir une série temporelle, comme les Réseaux de neurones par exemple. Il faudra aussi renforcer les variables explicatives, en ajoutant par exemple des données météo précises par heure. Au niveau de la collecte des données, le maillage des sites de comptage dans Paris devra être renforcé pour une meilleure représentativité.

Annexes

- Notebook “Projet PyCycle 15 01 21 DEF”
- Le jeu de données principal :
 - comptage-velo-donnees-compteurs_010919_311220
- Le jeu de données complémentaire (4 tables) :
 - caracteristiques-2019
 - lieux-2019
 - usagers-2019
 - vehicules-2019
- Diagramme de Gantt

Diagramme de Gantt

Intitulé du projet : PyCycle in Paris