Etude sur le choix des nouvelles stations de Vélib

Abire, Changwei et Caroline

Question métier

Déterminer si le choix des nouvelles stations de vélib dans Paris est judicieux avec les 4 différentes bases de données:

- ❏ Station vélib
- ❏ Comptage vélo
- ❏ Réseaux cyclables
- ❏ Population et superficie

Sommaire

1. Présentation des bases de données
    a. Comptage vélo
    b. Réseaux cyclables
    c. Station vélib

Sommaire

1. Présentation des bases de données
    a. Comptage vélo
    b. Réseaux cyclables
    c. Station vélib
2. Analyse des stations à Paris
    a. Capacité des stations
    b. Les 11 nouvelles stations créés
    c. État critique et pourcentage des instants critiques

Sommaire

1. Présentation des bases de données
    a. Comptage vélo
    b. Réseaux cyclables
    c. Station vélib
2. Analyse des stations à Paris
    a. Capacité des stations
    b. Les 11 nouvelles stations créés
    c. État critique et pourcentage des instants critiques
3. Machine Learning pour prédire le pourcentage des instants critiques des stations

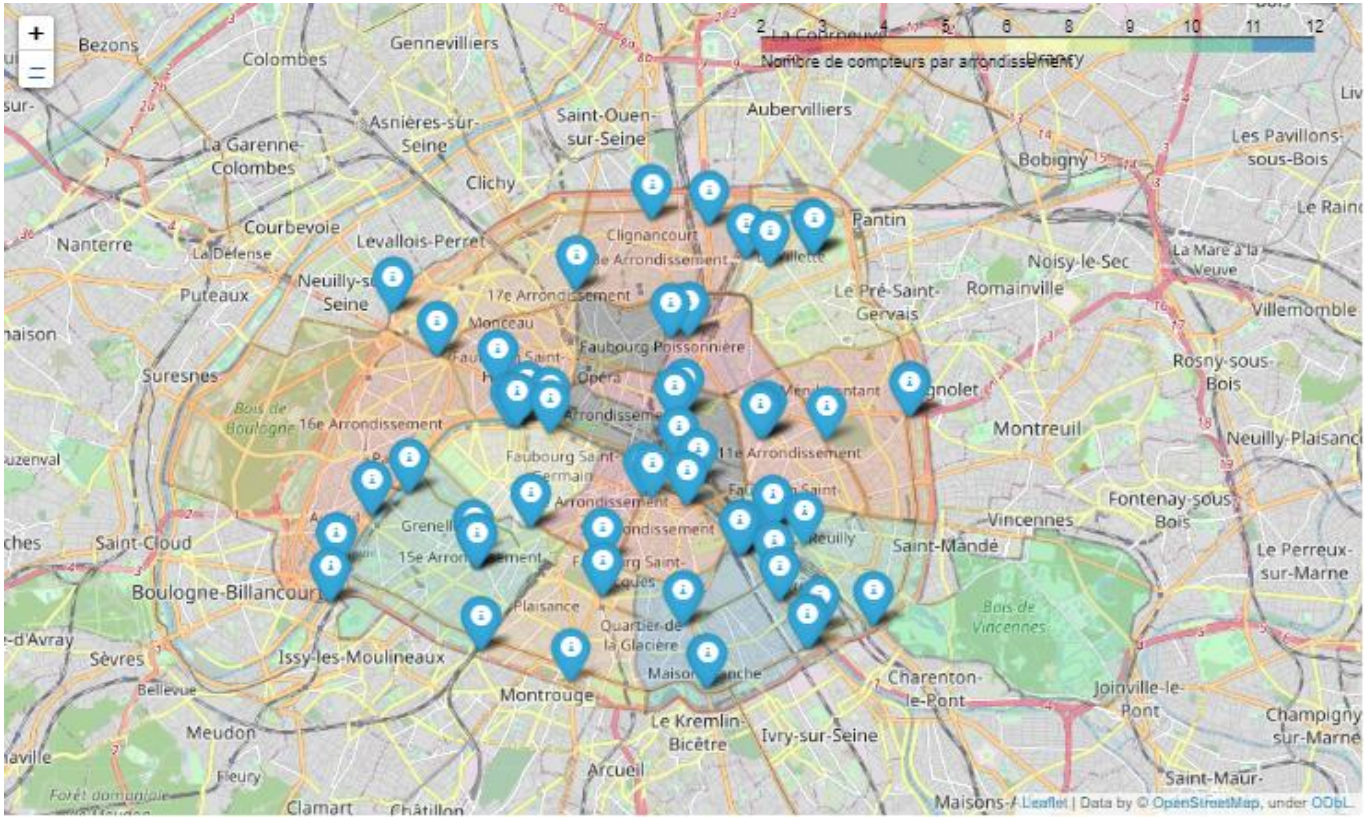
Sommaire

1. Présentation des bases de données
    a. Comptage vélo
    b. Réseaux cyclables
    c. Station vélib
2. Analyse des stations à Paris
    a. Capacité des stations
    b. Les 11 nouvelles stations créés
    c. État critique et pourcentage des instants critiques
3. Machine Learning pour prédire le pourcentage des instants critiques des stations
4. Conclusion et épilogue

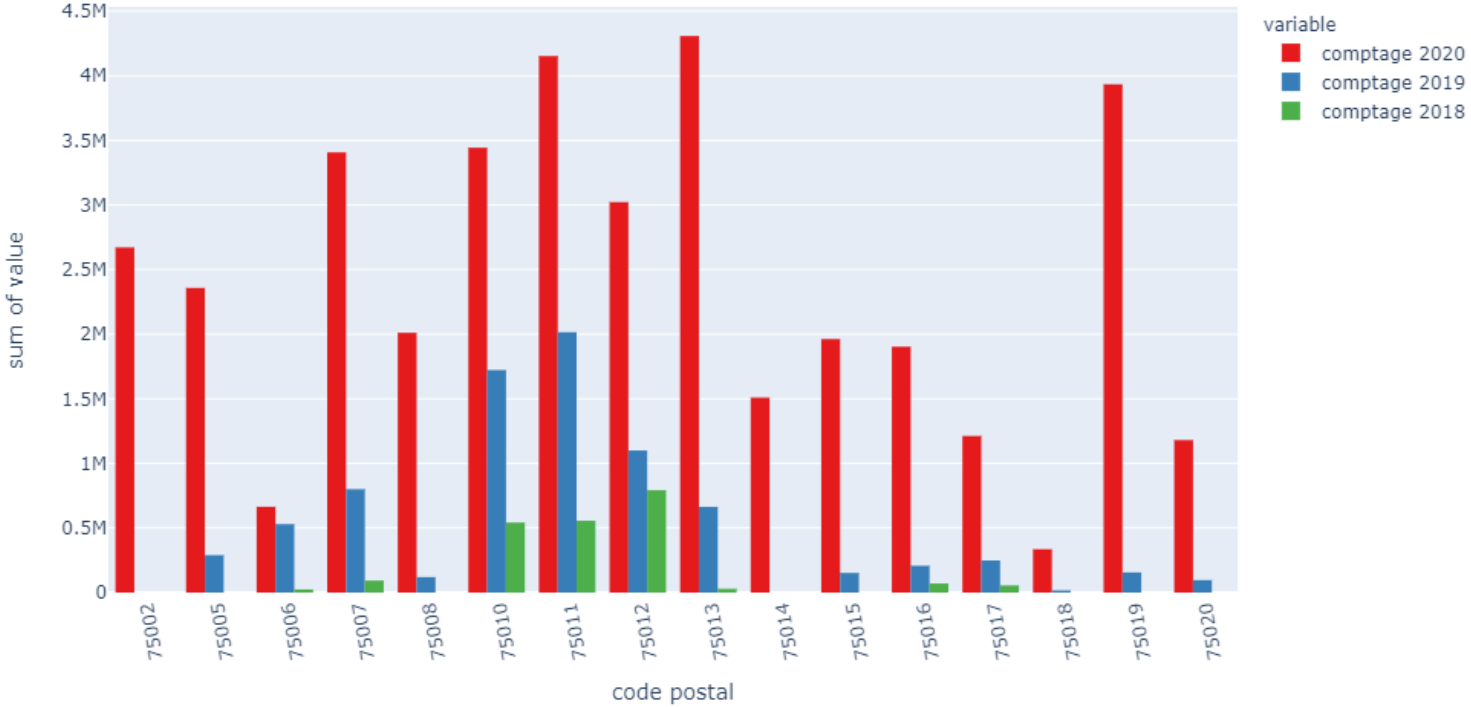
Sommaire

1- Dataset - Comptage dans Paris

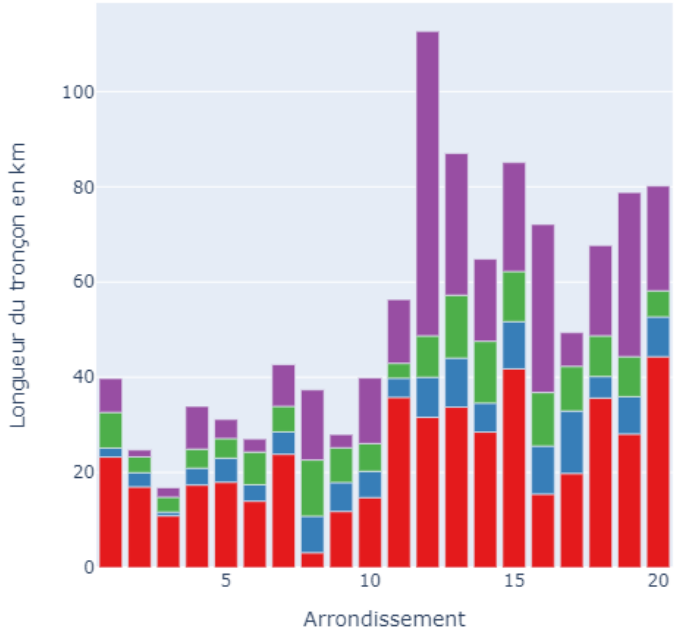
Bornes de comptage dans Paris



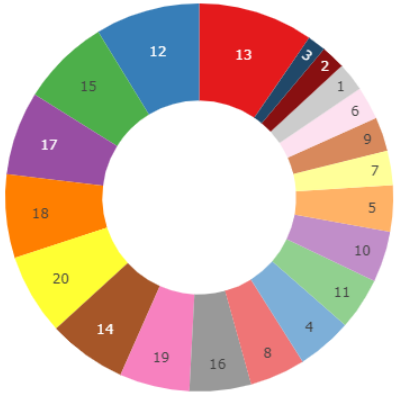
Comptage dans les différents arrondissements



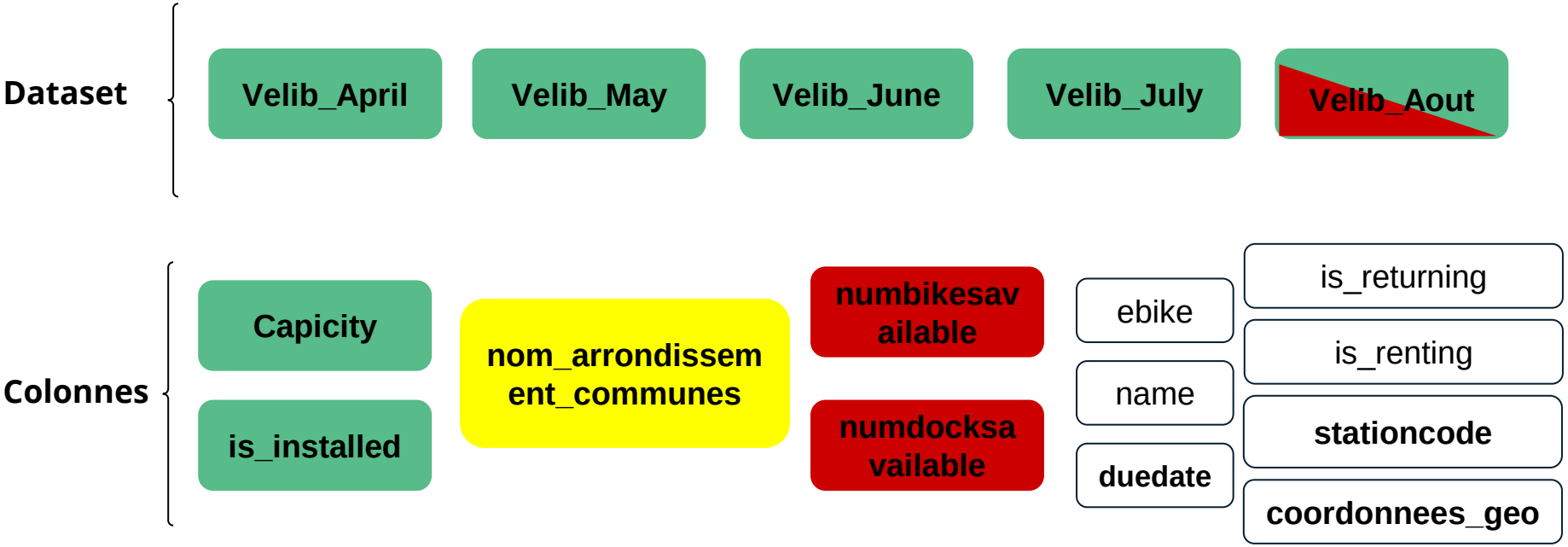
2 - Dataset piste cyclable dans Paris



Repartition du nombre de piste cyclable dans les diffrents arrondissements de Paris



3- Dataset de vélib



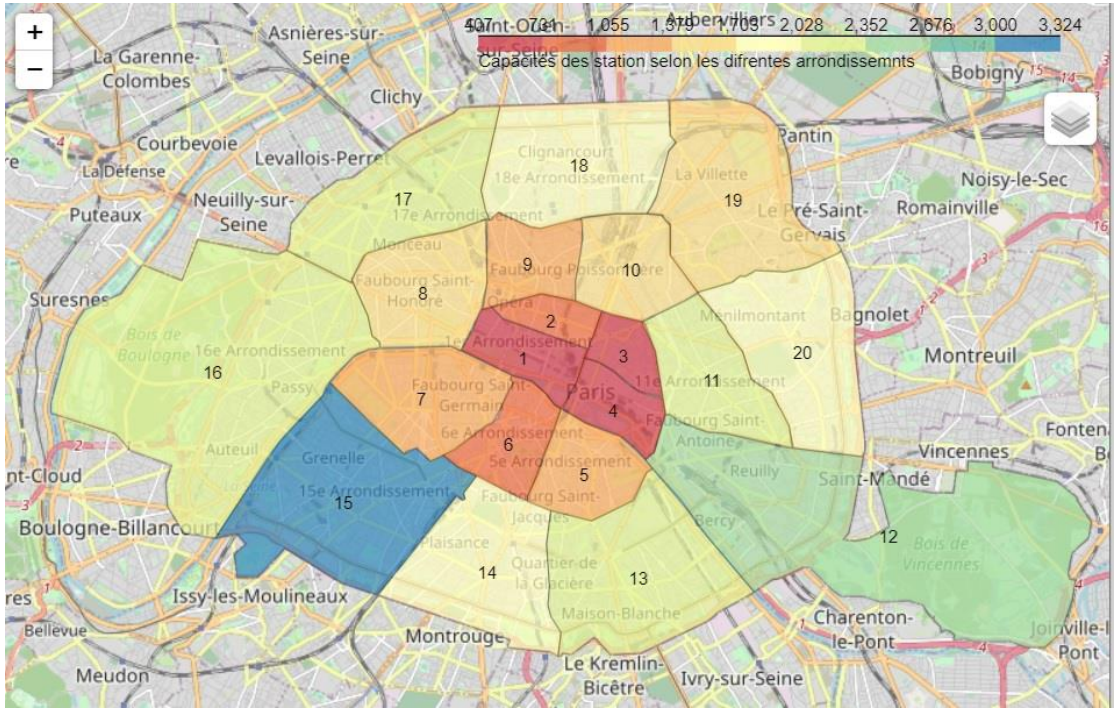
Sommaire

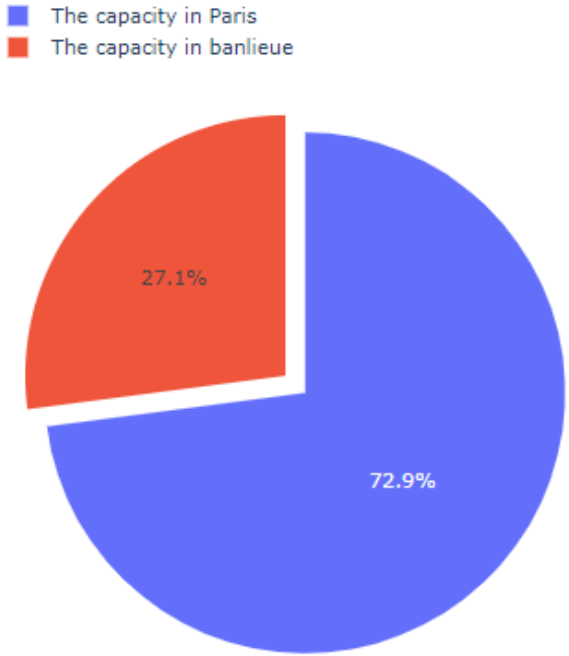
Capacité dans les 20 arrondissements de Paris

Une station dans Paris

=

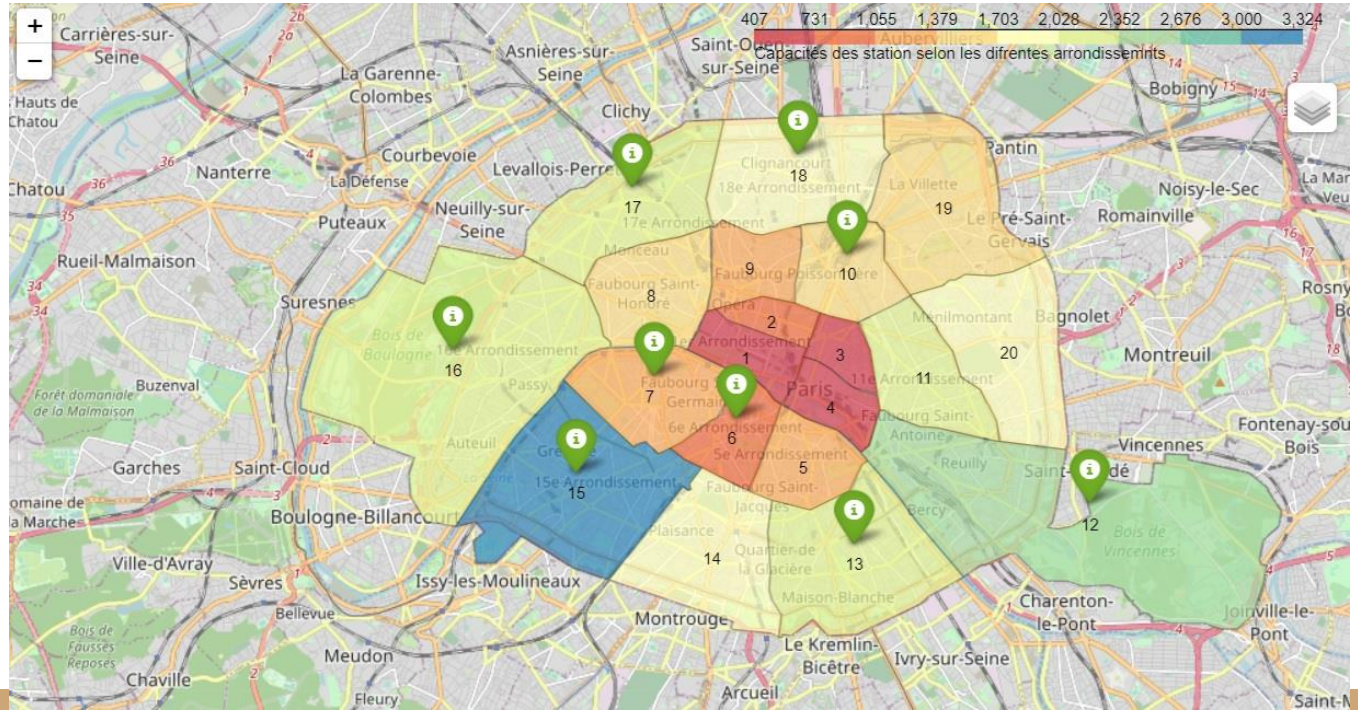
la colonne ‘nom_arrondissement_communes’ = Paris





Création des 11 nouvelles stations (dans 9 arrondissements)

Une nouvelle station = la colonne capacity = 0 ou NaN ou la colonne is_installed = Non



Définition de l'état critique et pourcentage des instants critiques

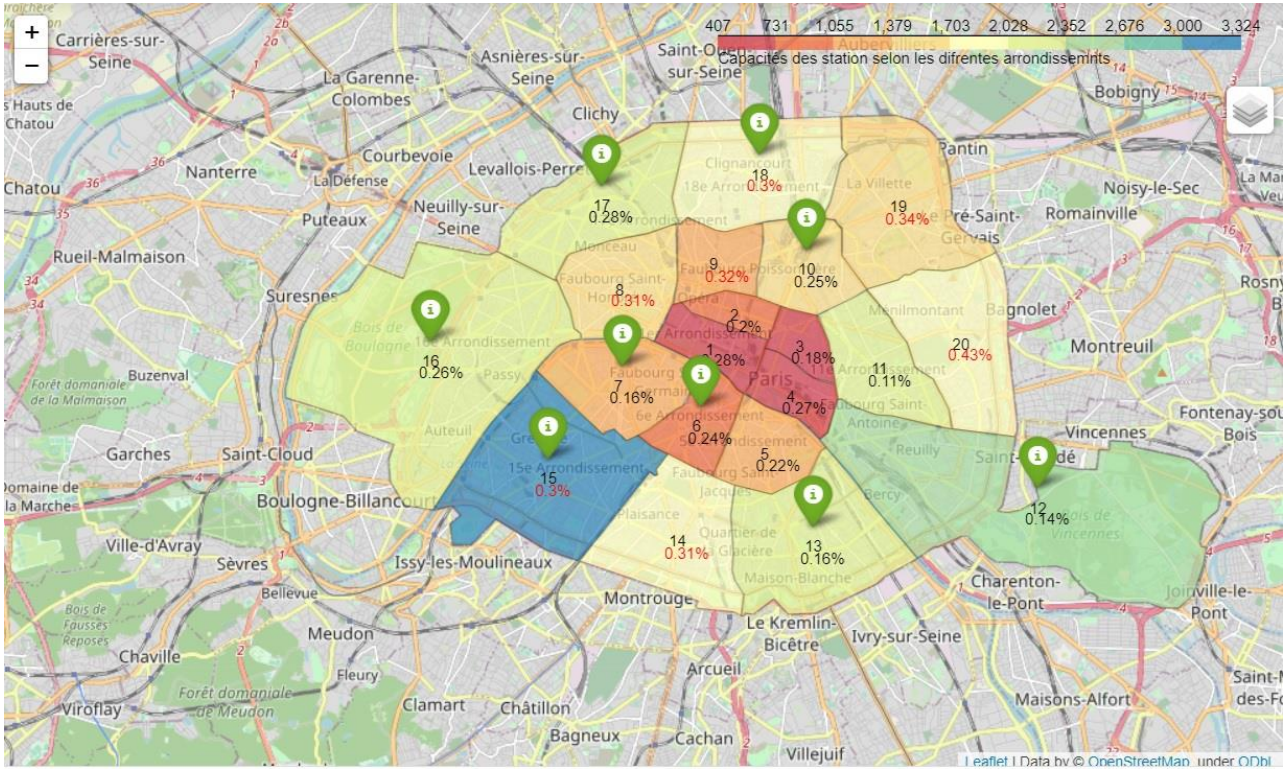
Etat critique d'une station = Nombre de vélos disponibles inférieur à 3 ou Nombre de bornes disponibles inférieur à 3

$$\text{Pourcentage des instants critiques} = \frac{\text{count des instants en etat critique}}{\text{count total des instants}}$$

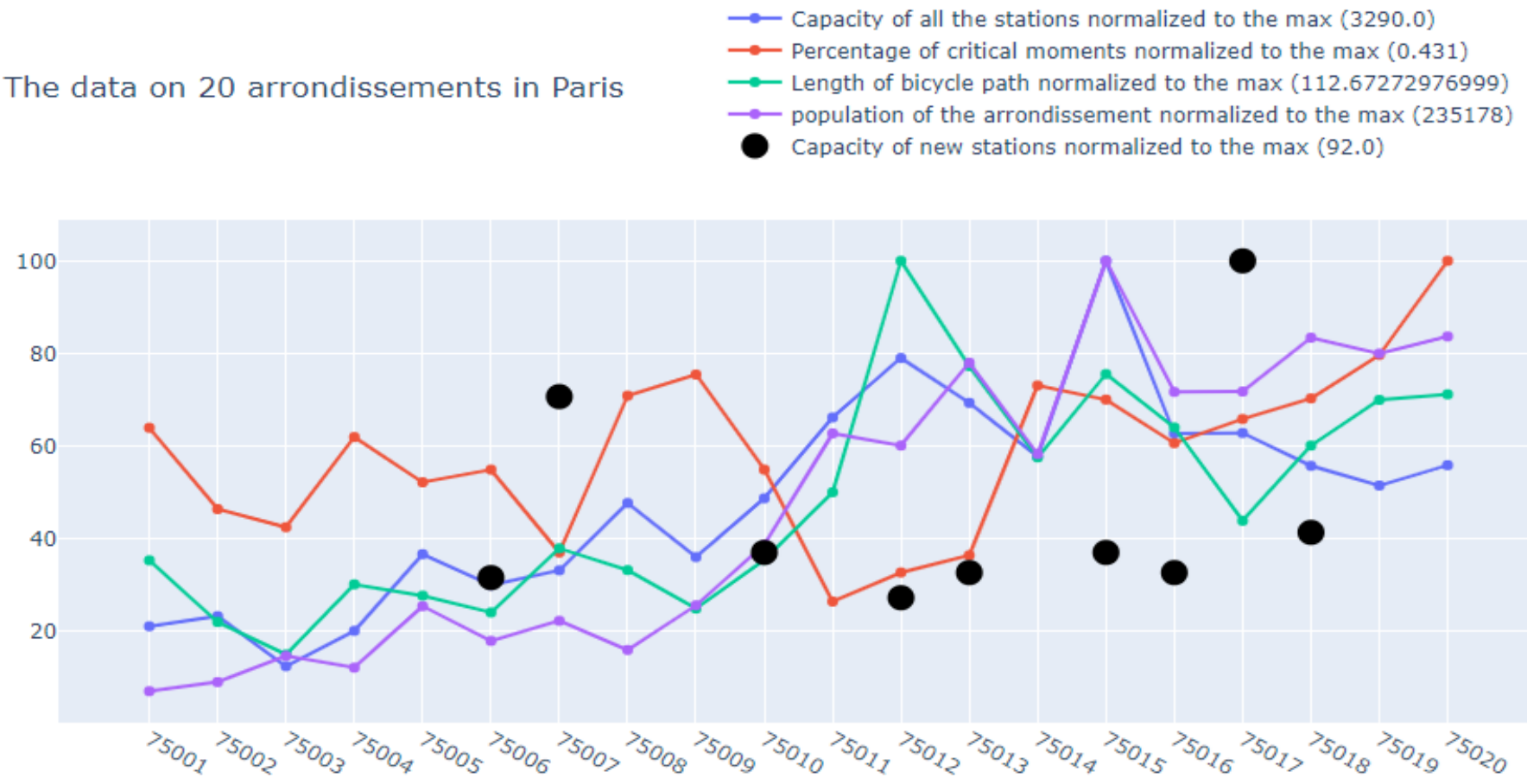
comprise entre 0 et 1. plus c'est grand plus la station est utilisée

```
count     982.000000
mean     2095.458248
std       376.963848
min        50.000000
25%      1835.000000
50%      2114.000000
75%      2363.750000
max      3270.000000
Name: count_all, dtype:
```

Pourcentage des instants critiques moyenné par arrondissement



The data on 20 arrondissements in Paris



Sommaire

Machine Learning pour prédire le “pourcentage des instants critiques” des stations

- Data : capacité, longitude/latitude, code postal, les caractéristiques de son arrond (pop, area, comptage velo, piste cyclable)
- “Target encoding” pour la colonne catégorique code postal
- Exclure les 11 nouvelles stations dans le dataset de Août, ce dataset est ensuite découpé en Train set (780 stations) + test set (195 stations)

```
2 X_train_encoded
```

	capacity	longitude	latitude	Population	area (ha)	Comptage horaire	Compteur	Longueur du tronçon en km	code postal encoded
0	15.0	48.849704	2.364461	28370	160	3.914346e+05	6.055556	33.884346	0.258035
1	51.0	48.835445	2.431421	141287	637	3.022408e+05	10.000000	112.672730	0.138216
2	62.0	48.851356	2.369220	141287	637	3.022408e+05	10.000000	112.672730	0.138216
3	29.0	48.870791	2.343101	21042	99	1.335674e+06	2.000000	24.721218	0.203909
4	21.0	48.871956	2.384981	196739	598	2.950025e+05	4.000000	80.170654	0.425631
...	...	...	...	...	...	...	...	...	...
775	21.0	48.886556	2.288640	168737	567	2.424206e+05	5.000000	49.378932	0.297191
776	38.0	48.863875	2.281890	168554	791	3.805720e+05	5.000000	72.097440	0.261736
777	19.0	48.886675	2.361361	196131	601	8.381150e+04	4.000000	67.685356	0.302184
778	23.0	48.858445	2.390398	196739	598	2.950025e+05	4.000000	80.170654	0.425631
779	28.0	48.883104	2.323835	168737	567	2.424206e+05	5.000000	49.378932	0.297191

780 rows × 9 columns

```
1 y_train
```

```
969    0.443139
678    0.119588
894    0.135102
33     0.258203
31     0.451631
         ...
106    0.631985
270    0.392931
860    0.339075
435    0.223077
102    0.598430
Name: percentage_critical, Length: 780,
```

Machine Learning pour prédire le “pourcentage des instants critiques” des stations

- Data : capacité, longitude/latitude, code postal, les caractéristiques de son arrond (pop, area, comptage velo, piste cyclable)
- “Target encoding” pour la colonne catégorique code postal
- Exclure les 11 nouvelles stations da
(780 stations) + test set (195 statior

```
xgb = XGBRegressor(random_state=42, n_estimators=500, learning_rate=0.05,
                   max_depth=4,
                   colsample_bytree=1)
```

```
2 X_train_encoded
```

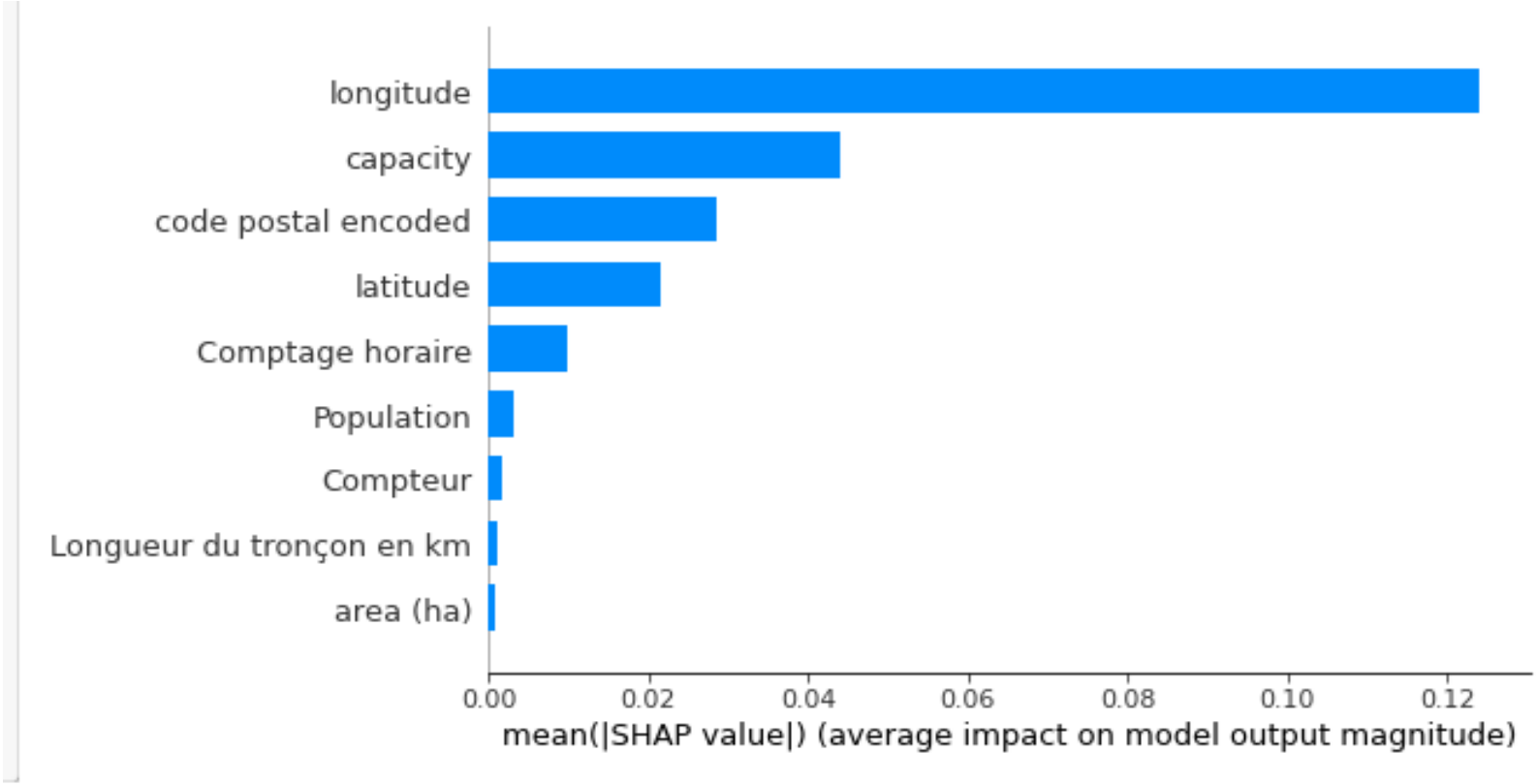
	capacity	longitude	latitude	Population	area (ha)	Comptage horaire	Compteur	Longueur du tronçon en km	code postal encoded
0	15.0	48.849704	2.364461	28370	160	3.914346e+05	6.055556	33.884346	0.258035
1	51.0	48.835445	2.431421	141287	637	3.022408e+05	10.000000	112.672730	0.138216
2	62.0	48.851356	2.369220	141287	637	3.022408e+05	10.000000	112.672730	0.138216
3	29.0	48.870791	2.343101	21042	99	1.335674e+06	2.000000	24.721218	0.203909
4	21.0	48.871956	2.384981	196739	598	2.950025e+05	4.000000	80.170654	0.425631
...	...	...	...	...	...	...	...	...	...
775	21.0	48.886556	2.288640	168737	567	2.424206e+05	5.000000	49.378932	0.297191
776	38.0	48.863875	2.281890	168554	791	3.805720e+05	5.000000	72.097440	0.261736
777	19.0	48.886675	2.361361	196131	601	8.381150e+04	4.000000	67.685356	0.302184
778	23.0	48.858445	2.390398	196739	598	2.950025e+05	4.000000	80.170654	0.425631
779	28.0	48.883104	2.323835	168737	567	2.424206e+05	5.000000	49.378932	0.297191

780 rows × 9 columns

```
1 y_train
```

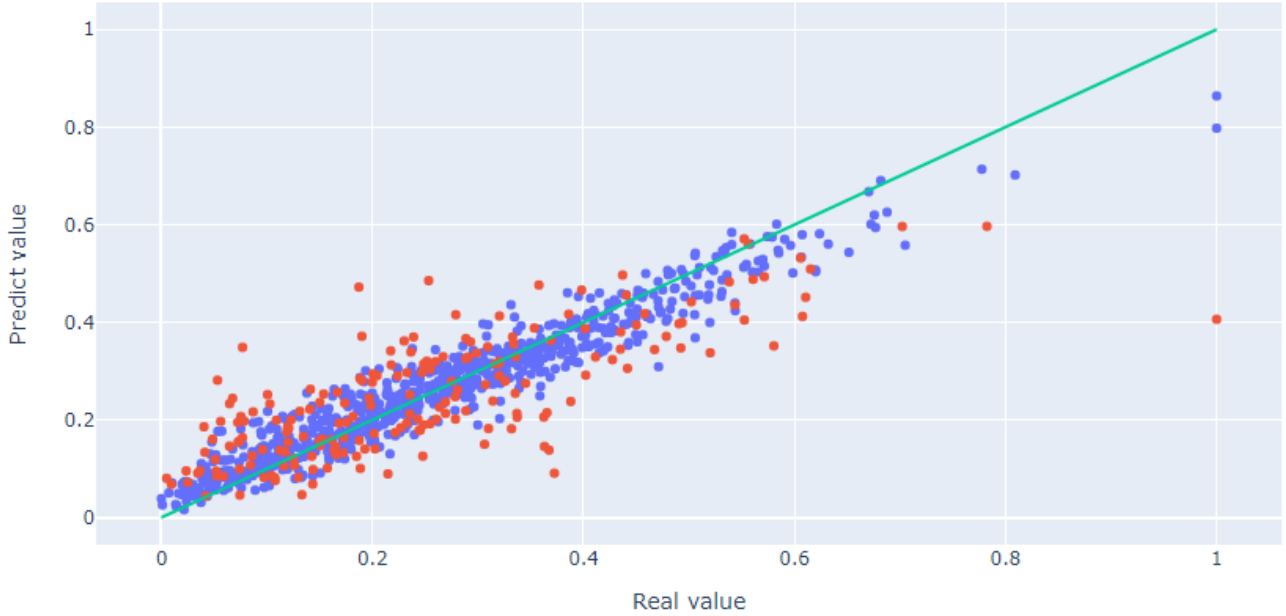
```
969    0.443139
678    0.119588
894    0.135102
33     0.258203
31     0.451631
         ...
106    0.631985
270    0.392931
860    0.339075
435    0.223077
102    0.598430
Name: percentage_critical, Length: 780,
```

Explicabilité du modèle (impact des différents features avec package **shap**)



Performance du modèle

The real and predicted values of the pourcentage_critique



The RMSE: 0.104
The R2: 0.591

```
1 y.describe()
```

```
count    975.000000
mean       0.262298
std        0.156885
min        0.000000
25%        0.138307
50%        0.243521
75%        0.358362
max        1.000000
Name: percentage_critical,
```

Prédiction sur les 11 nouvelles stations (final test set)

	stationcode	code postal	Real in August	Predicted ML	mean of the arrondissement
0	15104_relais	75015	NaN	0.328292	0.301671
1	18202	75018	0.071946	0.084293	0.302773
2	12166	75012	0.135431	0.140592	0.140628
3	13128	75013	0.218279	0.161539	0.156640
4	5122	75006	0.374670	0.253547	0.236397
5	16140	75016	NaN	0.218402	0.261084
6	17106	75017	0.056312	0.238040	0.283541
7	7007	75007	0.047002	0.092429	0.159210
8	17126	75017	NaN	0.220009	0.283541
9	17127	75017	NaN	0.232350	0.283541
10	10202	75010	NaN	0.147833	0.236422

Prédiction sur les 11 nouvelles stations (final test set)

	stationcode	code postal	Real in August	Predicted ML	mean of the arrondissement
0	15104_relais	75015	NaN	0.328292	0.301671
1	18202	75018	0.071946	0.084293	0.302773
2	12166	75012	0.135431	0.140592	0.140628
3	13128	75013	0.218279	0.161539	0.156640
4	5122	75006	0.374670	0.253547	0.236397
5	16140	75016	NaN	0.218402	0.261084
6	17106	75017	0.056312	0.238040	0.283541
7	7007	75007	0.047002	0.092429	0.159210
8	17126	75017	NaN	0.220009	0.283541
9	17127	75017	NaN	0.232350	0.283541
10	10202	75010	NaN	0.147833	0.236422

Bonne estimation

Mauvaise estimation

Prédiction sur les 11 nouvelles stations (final test set)

	stationcode	code postal	Real in August	Predicted ML	mean of the arrondissement
0	15104_relais	75015	NaN	0.328292	0.301671
1	18202	75018	0.071946	0.084293	0.302773
2	12166	75012	0.135431	0.140592	0.140628
3	13128	75013	0.218279	0.161539	0.156640
4	5122	75006	0.374670	0.253547	0.236397
5	16140	75016	NaN	0.218402	0.261084
6	17106	75017	0.056312	0.238040	0.283541
7	7007	75007	0.047002	0.092429	0.159210
8	17126	75017	NaN	0.220009	0.283541
9	17127	75017	NaN	0.232350	0.283541
10	10202	75010	NaN	0.147833	0.236422

Bon choix

Mauvais choix

Sommaire

Conclusion

- 11 nouvelles stations sont créés entre 05/2021 et 08/2021 dans Paris
- Un modèle de Machine learning est proposé pour prédire le “pourcentage des instants critiques” des stations
- Grâce au modèle de ML, on peut évaluer si le choix d’une future nouvelle station est bon ou mauvais
- Le modèle le plus performant est XGBoost, le RMSE et 0.104 , avec le target (pourcentage des instants critiques) varie entre 0 et 1 et la moyenne est à 0.26

Epilogue

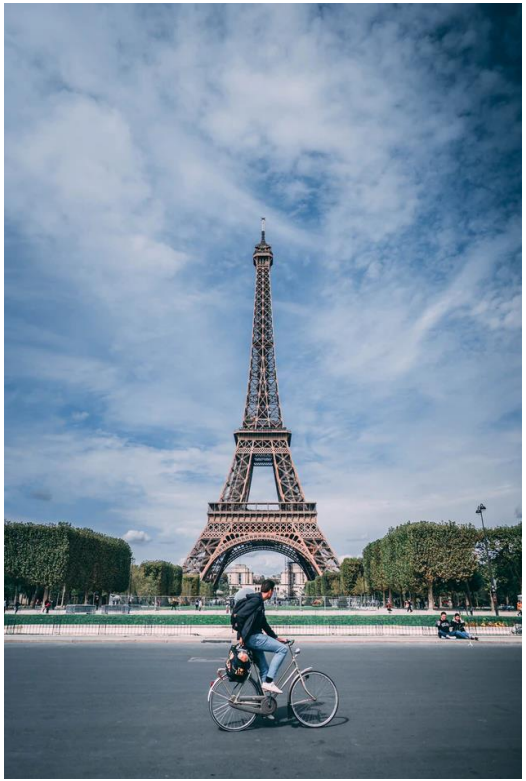
Limite de notre méthode

- Le choix des emplacements des nouvelles stations se pose sur d'autres éléments (place dispo, réseau électrique, etc)
- La colonne 'duedate' est mal renseignée, ce qui peut biaiser le pourcentage instant critique
- La colonne 'nom_arrondissement_commune's est mal renseignée, 3 nouvelles stations se trouvent hors Paris
- Comptage de vélo n'est pas représentatif, 96 compteurs seulement sur Paris et absence de compteur dans 4 arrondissements

Perspective

- Enrichir encore le modèle avec les données externes comme les âges/revenues/niveau de bruits, etc
- Étendre l'étude sur tout l'ile de France, car le plus grand nombre de nouvelle stations sont créés en banlieues (pb: trouver des data sur île de France)
- Faire des nouvelles études quand les données plus récents sont disponibles (+ de compteurs)
- Corriger les erreurs dans la colonne ' nom_arrondissement_communes' avec les coordonnées longitude/latitude
- Traiter le dataset comptage/piste cyclable selon chaque station et non par arrondissement
- Prédiction en tenant en compte le temps (dayofweek, jours congés ou non, etc)

Avez vous des questions ?

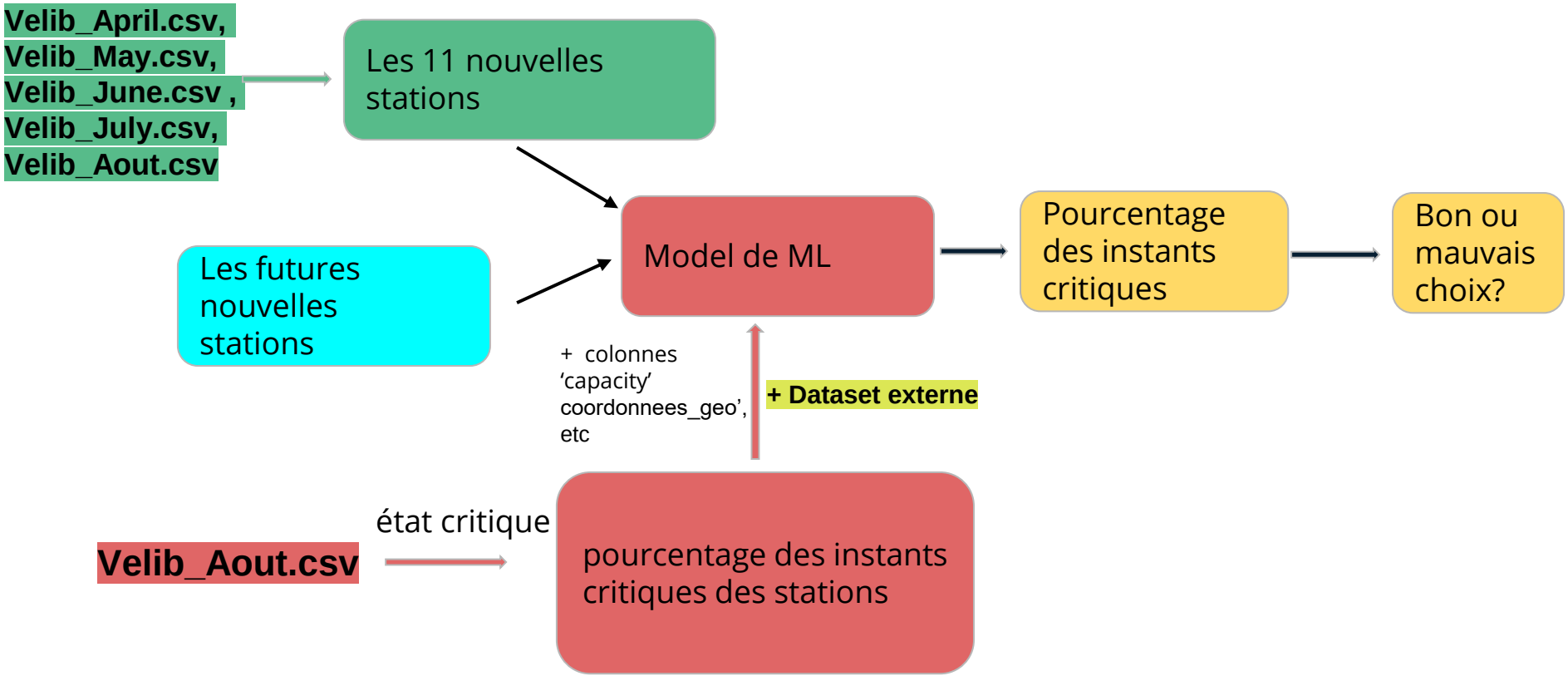


Annexe 1 : Répartition des tâches

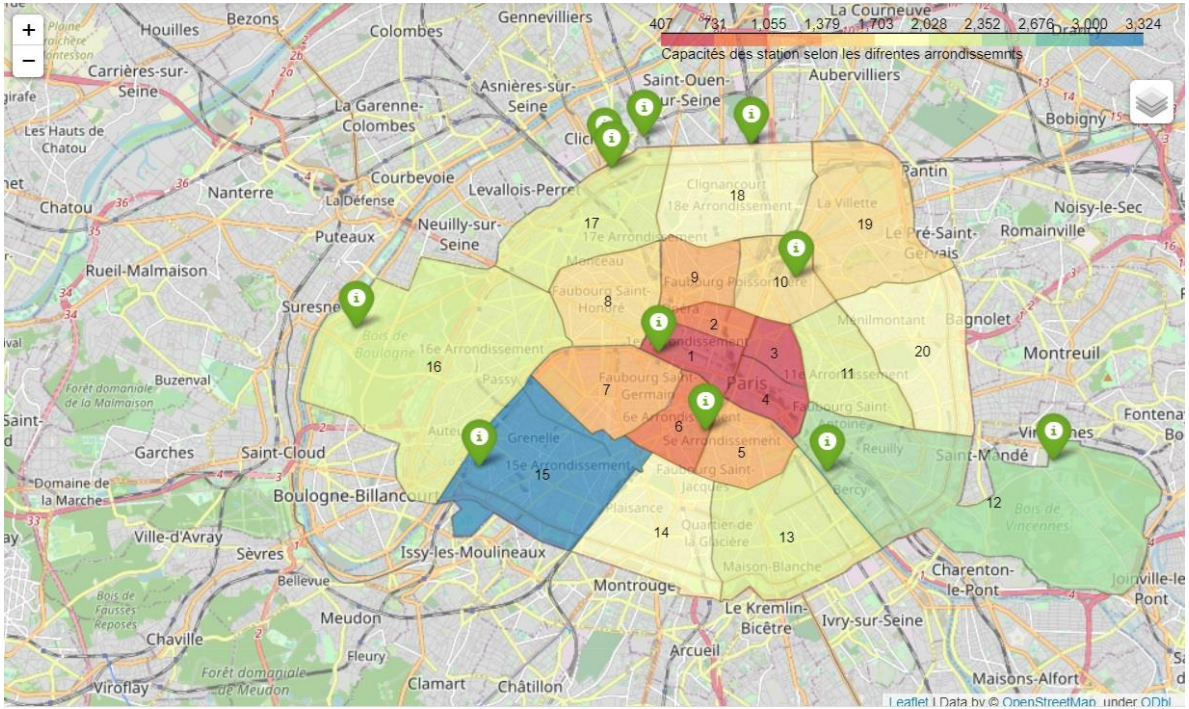
		Abire	Caroline	Changwei
Data préparation	dataset vélib			x
	dataset comptage vélo		x	
	dataset piste cyclable	x		
	package geopy	x		
Data visualisation	Affichage sur carte avec package folium	x	x	
	Divers (bar/pie chart/scatter) plotly	x	x	x
Machine Learning				x

https://docs.google.com/spreadsheets/d/1G98xP6Ko-Um4fyb-paP3hdzqeaKNg5q5pXvE91mlGFQ/edit?usp=sharing

Annexe 2 : processus de la data prép/analyse



Annexe 3: les 11 stations nouvelles créé (3 sont pas à paris)



T&C

· 15 min de présentation orale distribuée également parmi les membres de chaque équipe (tous dépassement sera pénalisé)

· Utilisation dau minimum un dataset.

Le Slide Deck Final doit contenir la présentation orale & des annexes :

· 1 Page expliquant la répartition du travail effectué par participant

· 2 Pages au maximum résumant clairement le processus et les stratégies mis en place lors de la préparation des données, le choix des visualisations, le choix des dataset et informations externes (sil y en a) pour répondre à votre question avec cohérence

Veuillez noter que nous attendons par mail votre support de présentation avant loral final : 9-9-2021 / 13h