L'objectif de ce challenge est de prédire le nombre de vélos passant devant la borne de comptage Albert 1er, au 02 Avril 2021, entre minuit et 9h. Nous avons à notre disposition un jeu de données disponible sur le site Compteurs Vélocité.

I. Description et nettoyage des données

Les données dont nous disposons, sont des relévés communautaires de l'affichage du compteur Albert 1er, à une heure donnée, depuis 12-03-2020, jusqu'à aujourd'hui.

La première remarque que nous pouvons soulignée est que chaque relevé dans une journée ne donne pas le nombre de passage de vélos depuis le dernier relevé, mais bien le cumul de vélos depuis minuit, puisque chaque compteur est remis à zéro à minuit. Pour éviter ce problème, nous récupérons les données sur le site Compteurs Vélocité, en ne conservant que la dernière colonne : celle qui donne le nombre total de passages de vélos en une journée.

Notes *Après plusieurs heures passées à essayer de manipuler les tableaux de données avec Python (Pandas), je n'ai pas obtenu un résultat satisfaisant. J'ai donc décidé de manipuler le tableau de données avec Excel, pour plus de commodité. En effet, les dimensions du tableau étant de taille "humaine", cette technique est utilisable, mais j'ai bien conscience, qu'avec des tableaux de plus grandes dimensions, la manipulation avec Excel ne sera pas recommandée.*

Après nettoyage et formatage des données sur Excel, nous obtenons le tableau final, prêt à être exploité, et qui contient deux colonnes : la première contenant les dates rangées dans l'ordre chronologique et la deuxième contenant le nombre total de vélos passés associé à chaque jour. Nous pouvons dans un premier temps tracer le nombre de vélos quotidien en fonction du temps. Nous pouvons décomposer le graphique en 5 parties :

1. L'impact du confinement général durant la période mars-mai 2020 qui fait chuter drastiquement le nombre de passages
2. La période de déconfinement et de vacances d'été 2020 qui a fait remonter le nombre de vélos
3. La période de rentrée scolaire à partir de septembre 2020 qui crée un pic en octobre
4. L'impact du second confinement en novembre
5. La remontée progressive du nombre de vélos en janvier 2021 suite au déconfinement progressif

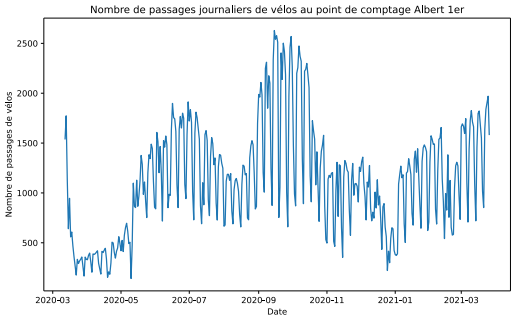

FIGURE 1 – Nombre de passage de vélos

II. Analyse de la stationarité

Nous choisissons l'étude de série temporelle, par le modèle ARIMA. Avant de pouvoir construire notre modèle, nous devons nous assurer que la série est stationnaire, pusique c'est l'hypothèse fondamentale pour pouvour appliquer le modèle.

Il existe deux façons de déterminer si une série temporelle donnée est stationnaire :

1. Statistiques mobiles : On trace la moyenne mobile et l'écart type mobile. La série temporelle est stationnaire si ces deux statistiques restent constantes dans le temps.
2. Le Test de Dickey-Fuller augmenté (ADF) : La série temporelle est stationnaire si la valeur-p est faible (selon l'hypothèse nulle) et que les valeurs critiques aux intervalles de confiance de 1, 5, 10 % sont proche la statistique ADF.

En utilisant la série brute, sans transformation, la p-valeur du test est supérieure au seuil fixé de 5% et la statistique ADF est loin des valeurs critiques. Ainsi, nous pouvons conclure que la série temporelle n'est pas stationnaire.

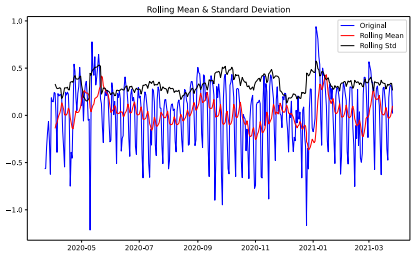

FIGURE 2 – Statistiques mobiles après transformation

```
Results of Dickey-Fuller Test:
Test Statistic                  -4.517205
p-value                          0.000183
#Lags Used                      14.000000
Number of Observations Used    355.000000
Critical Value (1%)             -3.448906
Critical Value (5%)             -2.869716
Critical Value (10%)            -2.571126
dtype: float64
```

FIGURE 3 – Statistique Dicker-Fulley après transformation

Nous prenons le logarithme des données et nous lui soustrayons la moyenne mobile pour rendre la série stationnaire.

Comme nous pouvons le voir, après avoir soustrait la moyenne, la moyenne mobile et l'écart type sont approximativement horizontaux. La valeur p est inférieure au seuil de 0,05 et la statistique ADF est proche des valeurs critiques. Par conséquent, la série temporelle est stationnaire.

III. Détermination des paramètres AR / MA

Pour déterminer les paramètres des modèles AR et MA, la méthode classique consiste à regarder respectivement la fonction d'autocorrélation (ACF) et la fonction d'autocorrélation partielle (PACF).

Nous choisissons d'uiliser le package Python *pmdarima*, qui contient la fonction *auto_arima*, permettant de calculer la meilleure combinaison de coefficients (p, q, d) du modèle ARIMA, selon le critère AIC.

```
Best model:  ARIMA(0,0,2)(0,1,1)[12] intercept
Total fit time: 64.676 seconds
```

SARIMAX Results

Dep. Variable:	y	**No. Observations:**	381
Model:	SARIMAX(0, 0, 2)x(0, 1, [1], 12)	**Log Likelihood**	-180.379
Date:	Sun, 28 Mar 2021	**AIC**	370.759
Time:	21:22:32	**BIC**	390.313
Sample:	0	**HQIC**	378.526
	- 381		

FIGURE 4 – Calcul automatique des coefficients ARIMA et SARIMAX

Nous choisissons les paramètres $(p, q, d) = (0, 0, 2)$ pour notre modèle ARIMA et $(p, q, d) = (0, 1, 1)[12]$ pour le modèle SARIMAX.

IV. Prédiction

Nous décidons d'utiliser le modèle SARIMAX avec les coefficents $(p, q, d) = (0, 1, 1)[12]$.

La fonction *predict* est utilisée sur notre modèle, afin de d'obtenir une prédiction sur la date choisie du 02/04/2021.

Les résultats obtenus nous permettent de prédire au 02/04/2021, un nombre de passages de vélos de **1732**.

Remarques *Je n'ai pas réussi à obtenir, à partir du modèle choisi, une prédiction sur un intervalle de temps précis, mais seulement le total de nombres de vélos sur un jour. On estime statistiquement, que la contribution de la plage horaire* $00:00-09:00$ *dans le total de vélos passant en un jour, est de 12%.*

Conclusion Le nombre de passages de vélos prédit pour le 02/04/2021 entre minuit et 9h00, est de

$$\boxed{1732 \times 12\% = 208 \text{ vélos}}$$

.

Critiques *Le modèle choisi est arbitraire et nous aurions pu nous intéresser à d'autres modèles de type régression linéaire, fast Fourier transformation, Deep Learning... Par ailleurs, nous nous sommes uniquement intéresser aux données fournies, qui sont d'une part, étalées sur une période de temps assez courte (à peine un an), et qui sont biaisées dû à la période de confinement). Nous avons envisager l'apport de données extérieures, en particulier, de données météorologiques. En effet, l'un des principaux critère de décision d'un cycliste à sortir est l'aspect météorologique : température extérieure, pluie, vente, neige, verglas, soleil, ... Cependant, nos connaissances et recherches ne nous ont pas permis de pousser cette idée.*