I - Présentation du sujet “Prédiction affluence sur lignes de bus” :

I.1 - Présentation du client :

Notre client est une compagnie de bus, qui gère le transport public urbain d’une ville importante (~190 000 habitants) à San Sébastien.

La compagnie est le service de transport public de voyageurs de la municipalité de la ville et propose un service de transport compétitif grâce à leurs efforts constants. Leur objectif premier est de répondre au mieux aux besoins de mobilité des usagers actuels et potentiels, grâce à un service de haute qualité et des informations pratiques, ainsi qu’à des coûts compétitifs et des critères de développement durable.

En termes de chiffres, c’est une flotte de 134 bus, dont 38 autobus hybrides et 3 autobus 100% électriques, qui assurent un service sur 32 lignes en journée et 9 la nuit.

Les enjeux de l’entreprise sont les suivants :

Orientation client et amélioration des services	1) Politiques d’amélioration du service aux usagers et campagnes d’incitation à l’utilisation du bus. 2) Amélioration continue de nos services en nombre de lignes, temps de déplacement, ponctualité et régularité, sécurité et confort.
Personnel	1) La Compagnie veille au respect de critères stricts d’égalité des chances et de transparence. 2) Nous garantissons les mêmes droits et les mêmes chances aux hommes et aux femmes de l’entreprise.
Environnement	1) Nous veillons au respect des réglementations et législations en vigueur relatives à l’environnement ; nous souhaitons même les devancer grâce à une attitude proactive. 2) Nous défendons le développement durable et respectueux de l’environnement par le biais de l’implantation et de l’utilisation des dernières avancées technologiques pour nos bus.
Qualité	1) Amélioration continue visant à l’adoption de mesures préventives plutôt que correctives. 2) Orientation à la création et à la perception des usagers dans le but d’une plus grande satisfaction, toujours sur la base de tarifs compétitifs. 3) Objectif excellence pour les services proposés.

I.2 - La problématique : Comment prédire l’affluence journalière sur les lignes de bus du réseau ?

Dans le cadre de sa mission d'amélioration des services proposés à ses usagers, la société souhaite être en mesure (à terme) de :

1. Anticiper les fortes affluences d'usagers sur l'ensemble de son réseau en ayant une vision claire des demandes pour chaque ligne et pour chaque arrêt.
2. Optimiser les temps de trajets de sa flotte

I.3 - Présentations du jeux de données :

Vous aurez à disposition un historique du cumul journalier de voyageurs allant du ***05 avril 2019*** au ***08 mars 2023***, pour toutes les lignes de bus.

- dateTime : Période du 2019-04-05 au 2023-03-08
- lineNumber : 39 lignes de bus présentes

Le jeu de données est composé de 36 901 observations et 3 variables explicatives originales qui sont, la date, la ligne de bus et le type de ligne de bus. La variable quantitative à prédire est le nombre de passagers/usagers.

I.4 - Présentations de l'augmentation du jeu de données :

Les variables issues et extraites des dates :

- dateTimeMonth : numéro du mois (de 1 à 12)
- dateTimeWeek : numéro de la semaine dans l'année (de 1 à 52/53)
- dateTimeDay : jour du mois (de 1 à 30)
- dateTimeDayofweek : jour de la semaine (de 1 à 7)
- dateTimeDayofyear : jour de l'année (de 1 à 365)
- dateTimeIs_sunday : var binaire (un dimanche ou non)
- dateTimeIs_holiday : var binaire (jour férié ou non)
- dateTimeIs_paques : var binaire (pâques ou non)
- dateTimeIs_schoolholiday : var binaire (vacances scolaires ou non)
- dateTimeIs_grandsemaine : var binaire (La semana grande)

Sites conseillés :

➔ librairies Python holidays et calendar :

```
from vacances_scolaires_france import SchoolHolidayDates
import holidays
import calendar
```

Les variables issues et extraites du nombre de passager dans le passé :

Les variables supplémentaires pour prendre en compte la tendance de cycle, de périodicité de l'affluence par semaine.

- "nbPassenger_lag1" = le nombre de passagers de 7 jours précédents
- "nbPassenger_lag2" = le nombre de passagers de 14 jours précédents
- "nbPassenger_lag3" = le nombre de passagers de 21 jours précédents

Les variables issues de l'opendata :

L’ajout de l’open data rend un modèle plus robuste. Les données météorologiques ont été ajoutées via une variable catégorielle et donnent un aperçu de la météo du jour. Elle est encodée. Par exemple en français cela donnerait cela :

Désignation	Très pluvieux	Pluvieux	Très nuageux	Partiellement nuageux	Ensoleillé
Valeur	1	2	3	4	5

Dans les données transmis pour ce projet avec l’ESTIA BIHAR, l’encodage est plus fin (12 13 43 26 25 11 62 24 23 61 53 52 51 17 44 15 14 45 46 54), par exemple :

Désignation	Cubierto con lluvia	Cubierto con lluvia escasa	Muy nuboso	Intervalos nubosos	Despejado
Valeur	26	24	15	13	11

Désignation	Muy nuboso con lluvia escasa	Muy nuboso con lluvia	Nubes altas noche	Intervalos nubosos con lluvia escasa noche	Nuboso con nieve escasa noche
Valeur	12	1	4	3	2

L’encodage exhaustif de la variable météorologique “weather” est décrite ci-dessous :

D’une part les grandes familles de 1 à 5 :

1 : "Nuboso con tormenta", "Muy nuboso con nieve escasa", "Nuboso con tormenta y lluvia escasa noche", "Nuboso con tormenta y lluvia escasa", "Nuboso con tormenta noche", "Cubierto con nieve escasa", "Muy nuboso con tormenta y lluvia escasa", "Cubierto con lluvia", "Intervalos nubosos con tormenta y lluvia escasa", "Cubierto con tormenta y lluvia escasa", "Cubierto con nieve", "Nuboso con nieve noche", "Muy nuboso con tormenta", "Nuboso con nieve", "Muy nuboso con nieve", "Cubierto con tormenta", "Intervalos nubosos con tormenta y lluvia escasa noche", "Muy nuboso con lluvia"

2 : "Intervalos nubosos con tormenta noche", "Intervalos nubosos con nieve escasa", "Nuboso con lluvia noche", "Nuboso con lluvia", "Intervalos nubosos con nieve noche", "Cubierto con lluvia escasa", "Intervalos nubosos con nieve", "Intervalos nubosos con tormenta", "Nuboso con nieve escasa", "Intervalos nubosos con nieve escasa noche", "Nuboso con nieve escasa noche", "Muy nuboso con lluvia escasa"

3 : "Intervalos nubosos con lluvia noche", "Intervalos nubosos con lluvia", "Nuboso con lluvia escasa noche", "Nuboso con lluvia escasa", "Muy nuboso", "Nuboso", "Intervalos nubosos con lluvia escasa", "Cubierto", "Intervalos nubosos con lluvia escasa noche"

4 : "Nuboso noche", "Nubes altas", "Intervalos nubosos", "Intervalos nubosos noche", "Nubes altas noche"

5 : "Despejado", "Poco nuboso", "Despejado noche", "Poco nuboso noche"

Puis dans les données transmis ici l'encodage est plus fin (12 13 43 26 25 11 62 24 23 61 53 52 51 17 44 15 14 45 46 54), pour les catégories manquantes voici leur description :

12 : "Muy nuboso con lluvia escasa"
13 : "Intervalos nubosos"
43 : "Intervalos nubosos con lluvia escasa"
26 : "Cubierto con lluvia"
25 : "Muy nuboso con lluvia"
11 : "Despejado"
62 : "Nuboso con tormenta y lluvia escasa"
24 : "Nuboso con lluvia"
23 : "Intervalos nubosos con lluvia"
61 : "Intervalos nubosos con tormenta y lluvia escasa
53 : "Muy nuboso con tormenta"
52 : "Nuboso con tormenta"
51 : "Intervalos nubosos con tormenta"
17 : "Nubes altas"
44 : "Nuboso con lluvia escasa"
15 : "Muy nuboso"
14 : "Nuboso"
45 : "Muy nuboso con lluvia escasa"
46 : "Cubierto con lluvia escasa"
16 : "Cubierto"
13 : "Intervalos nubosos"
54 : "Cubierto con tormenta"

Les variables issues de l'opendata pour les jours spécifiques à la région :

L'ajout d'informations sur les événements sportifs notamment par rapport aux matchs de football et de basket est intéressant. Ces événements ont une grande influence sur l'affluence des usagers sur les lignes de bus. La variable binaire suivante a été ajouté :

- dateTimeIs_football : va binaire (indique si oui ou non un match de football où joue la Real Sociedad est planifié)

Site conseillé :

- https://www.les-sports.info/football-real-sociedad-resultats-identite-equ526.html

Une vue de la description de la table :

```
> str(df_SEME)
'data.frame':   17447 obs. of  18 variables:
 $ dateTime                : Factor w/ 402 levels "2019-06-16","2019-06-17",..: 1 2 3 4 5 6 7 8 9 10 ...
 $ lineNumber              : int  5 5 5 5 5 5 5 5 5 5 ...
 $ weather                 : int  12 12 13 43 26 25 12 13 13 13 ...
 $ dateTimeMonth           : int  6 6 6 6 6 6 6 6 6 6 ...
 $ dateTimeWeek            : int  24 25 25 25 25 25 25 25 26 26 ...
 $ dateTimeDay             : int  16 17 18 19 20 21 22 23 24 25 ...
 $ dateTimeDayofweek       : int  1 2 3 4 5 6 7 1 2 3 ...
 $ dateTimeDayofyear       : int  167 168 169 170 171 172 173 174 175 176 ...
 $ dateTimeIs_sunday       : int  1 0 0 0 0 0 0 1 0 0 ...
 $ dateTimeIs_holiday      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ dateTimeIs_paques       : int  0 0 0 0 0 0 0 0 0 0 ...
 $ dateTimeIs_schoolholiday: int  0 0 0 0 0 0 0 0 0 0 ...
 $ dateTimeIs_grandsemaine : int  0 0 0 0 0 0 0 0 0 0 ...
 $ dateTimeIs_football     : int  0 0 0 0 0 0 0 0 0 0 ...
 $ nbPassenger_lag1        : int  4184 8546 8822 8854 9109 9554 7089 4184 8546 8822 ...
 $ nbPassenger_lag2        : int  4184 8546 8822 8854 9109 9554 7089 4570 8513 8656 ...
 $ nbPassenger_lag3        : int  4184 8546 8822 8854 9109 9554 7089 4570 8513 8656 ...
 $ passengersNumber        : int  4184 8546 8822 8854 9109 9554 7089 4570 8513 8656 ...
```

I.5 - Recherche d'un modèle de prédiction du nombre de passagers par jour par ligne de bus, avec une anticipation de 3 jours au moins

Travail à réaliser.