

Challenge du Développement Logiciel

Objectif : Prédire le nombre de vélos le 2 Avril 2021 de 00H01 à 09H00 de la Myriade de Totems de Montpellier

In [1]:

```
## Lectures des données

# Importer les bibliothèques
import pandas as pd # Pour importer des tableaux matriciels
import numpy as np # Pour le calcul numérique
import matplotlib.pyplot as plt # Pour tracer des graphiques
%matplotlib inline
df = pd.read_csv('https://docs.google.com/spreadsheets/d/e/2PACX-1vQVtdpXMHB4g9h75a0jw8CsrqSuQmP5eMIB2adpKR5hkRggwMwzFy5kB-AI
ThodhVHNLxIZYm8fuoWj/pub?gid=2105854808&single=true&output=csv', index_col = 'Date', parse_dates = True)
# L'argument parse_dates est très important dans un dataframe quand on a une série temporelle
df.head() # Visualisation des 5 premières lignes de notre dataframe
```

Out[1]:

	Heure / Time	Vélos depuis le 1er janvier / Grand total	Vélos ce jour / Today's total	Unnamed: 4	Remarque
Date					
NaT	NaN	NaN	NaN	NaN	NaN
NaT	NaN	NaN	NaN	NaN	NaN
2020-12-03	11:05:00	109.0	109.0	NaN	NaN
2020-12-03	14:30:00	441.0	441.0	NaN	NaN
2020-03-13	10:02:00	1921.0	385.0	NaN	NaN

In [3]:

```
df.tail()
# Visualisation des 5 dernières lignes de notre dataframe
```

Out[3]:

	Heure / Time	Vélos depuis le 1er janvier / Grand total	Vélos ce jour / Today's total	Unnamed: 4	Remarque
Date					
2021-03-31	16:27:00	109971.0	1204.0	NaN	NaN
2021-03-31	17:00:00	110050.0	1283.0	NaN	NaN
2021-03-31	17:13:00	110092.0	1325.0	NaN	NaN
2021-03-31	20:12:00	110619.0	1852.0	NaN	NaN
2021-01-04	09:37:00	111145.0	433.0	NaN	NaN

In [4]:

```
print(df.shape) # On a des lignes(ou nombres de vélos) et des colonnes (ou caractéristiques) décrivant notre jeu de données.
# Le nombre de lignes et de colonnes avant le nettoyage de notre dataframe.
```

(1505, 5)

Nettoyons notre dataframe.

Lorsque nous traitons un dataframe avec pandas, il est impératif de s'assurer que nous avons des données fiables, c'est-à-dire pas de valeurs nulles ou de valeurs aberrantes.

In [5]:

```
df_new = df.drop(['Unnamed: 4', 'Remarque'],axis = 1) # Eliminons les deux(2) dernières colonnes vides("Unnamed: 4, Remarque"), les variables ou
les colonnes
df_new
```

Out[5]:

Date	Heure / Time	Vélos depuis le 1er janvier / Grand total	Vélos ce jour / Today's total
NaT	NaN	NaN	NaN
NaT	NaN	NaN	NaN
2020-12-03	11:05:00	109.0	109.0
2020-12-03	14:30:00	441.0	441.0
2020-03-13	10:02:00	1921.0	385.0
...
2021-03-31	16:27:00	109971.0	1204.0
2021-03-31	17:00:00	110050.0	1283.0
2021-03-31	17:13:00	110092.0	1325.0
2021-03-31	20:12:00	110619.0	1852.0
2021-01-04	09:37:00	111145.0	433.0

1505 rows x 3 columns

In [6]:

```
df_new = df_new.dropna(axis = 0) # Eliminons les lignes possédant les valeurs manquantes
df_new
# Voici maintenant, après le nettoyage, le nombre de lignes et de colonnes de notre dataframe ci-dessous.
```

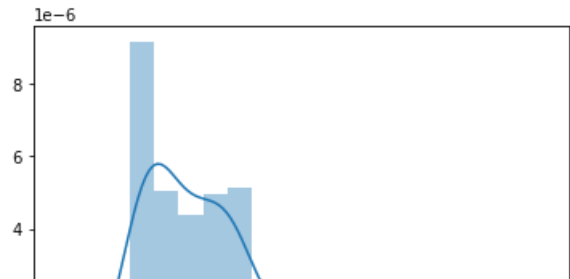
Out[6]:

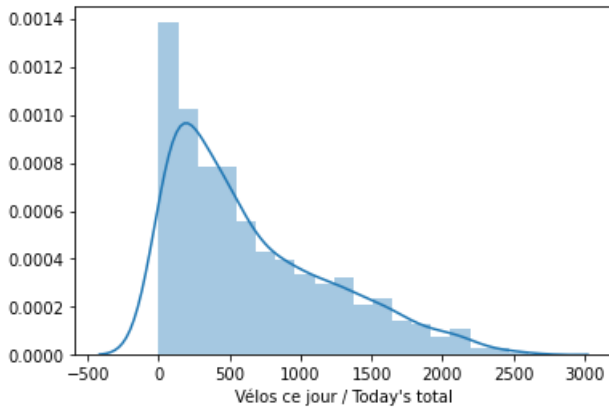
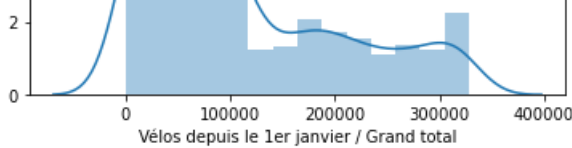
Date	Heure / Time	Vélos depuis le 1er janvier / Grand total	Vélos ce jour / Today's total
2020-12-03	11:05:00	109.0	109.0
2020-12-03	14:30:00	441.0	441.0
2020-03-13	10:02:00	1921.0	385.0
2020-03-13	14:15:00	2305.0	769.0
2020-03-13	18:21:00	2822.0	1286.0
...
2021-03-31	16:27:00	109971.0	1204.0
2021-03-31	17:00:00	110050.0	1283.0
2021-03-31	17:13:00	110092.0	1325.0
2021-03-31	20:12:00	110619.0	1852.0
2021-01-04	09:37:00	111145.0	433.0

1501 rows x 3 columns

In [7]:

```
import seaborn as sns
for col in df_new.select_dtypes("float"):
    plt.figure()
    sns.distplot(df_new[col])
## Nos deux(2) variables telles que Vélos depuis le 1er janvier / Grand total et Vélos ce jour / Today's total suivent une distribution asymétrique à gauche
```





Faisons la prédiction pour le 2 avril 2021 de 00H01 à 09H00 prochain.

In [8]:

```
df_new.loc["2021-04-02", "Vélos ce jour / Today's total"].resample("h").agg(["mean", "min", "max", "std", "sum"])
```

Out[8]:

	mean	min	max	std	sum
Date					
2021-04-02	482.0	183.0	703.0	268.631718	1446.0

Nous avons travaillé avec un dataframe contenant des vélos issu de la Myriade de Totems de Montpellier. Pour faire notre prédiction sur les vélos le 2 avril prochain, nous avons procédé de la manière suivante : En premier lieu, nous avons téléchargé notre dataframe avec l'importation de plusieurs modules de python comme par exemple: pandas pour des tableaux matriciels, numpy pour des calculs numériques et matplotlib pour pouvoir tracer les graphiques et seaborn.

Ensuite, nous avons remarqué plusieurs valeurs manquantes dans notre dataframe. Pour pouvoir faire notre prédiction, nous devons nettoyer nos données. Voilà pourquoi, nous avons utilisé la fonction `drop(axis = 1)` pour éliminer les deux(2) dernières colonnes vides de notre dataframe et la fonction `dropna(axis=0)` pour supprimer les deux(2) premières lignes qui contiennent aussi des valeurs manquantes. Enfin, nous avons utilisé dans notre prédiction la fonction `aggégate`(cette fonction permet de rassembler dans un seul tableau les données statistiques) et la fonction `resample`(cette fonction permet aussi de regrouper des données selon une série temporelle, par exemple: Il est possible à partir de grandeurs mesurées toutes les heures de rééchantillonner les données avec une période journalière) qui vont nous donner un tableau qui contient les différents statistiques comme la moyenne, le minimum, le maximum, l'écart-type et la somme. Donc, notre prédiction est la suivante : le 2 avril 2021 de 00H01 à 09H00, la Myriade de Totems de Montpellier pourrait enregistrer au minimum 183 vélos à 703 vélos au maximum, c'est-à-dire le nombre de vélos pourrait se trouver dans cet intervalle [183, 703]. La Myriade de Totems de Montpellier va enregistrer en moyenne 482 vélos le 2 avril 2021. En fait, pour faire la prédiction à partir de n'importe quelle autre date et heure, il suffit d'entrer la date(année, mois, jour), puis lancer le code ci-dessus et vous trouverez ce que vous voulez.

Sources de documentation : 1) <https://stackoverflow.com/questions/13411544/delete-column-from-pandas-dataframe?rq=1>

2) <https://ledatascientist.com/manipulez-vos-donnees-avec-pandas/>

3) https://www.guillaumedueymes.com/courses/formation_python/8-pandas/

4) <https://informatique-python.readthedocs.io/fr/latest/Cours/science2.html>