In [1]:
```
pwd
```

Out[1]:
```
'/Users/yaha'
```

In [2]:
```
import pandas as pd
import numpy as np
import datetime
import warnings
warnings.filterwarnings('ignore')
```

In [3]:
```
df=pd.read_csv('/Users/yaha/Desktop/FORMATION/LE PONT/CAPSTONE 3/Dataset csv/Velib_JULY_2022.csv', sep = ',')
df.head()
```

Out[3]:

	name	stationcode	ebike	mechanical	coordonnees_geo	duedate	numbikesavailable	numdocksavailable	capacity	is_renting	is_installed
0	Mairie de Rosny-sous-Bois	31104	16	6	[48.87125655190012, 2.4865807592869]	2022-07-01T00:13:20+02:00	22	7	30.0	OUI	OUI
1	Benjamin Godard - Victor Hugo	16107	2	4	[48.865983, 2.275725]	2022-07-01T00:13:35+02:00	6	27	35.0	OUI	OUI
2	Charonne - Robert et Sonia Delauney	11104	6	4	[48.85590755596891, 2.3925700744194035]	2022-07-01T00:12:33+02:00	10	10	20.0	OUI	OUI
3	Harpe - Saint-Germain	5001	0	0	[48.851518815016892, 2.3436703160040516]	2022-07-01T00:10:42+02:00	0	45	45.0	OUI	OUI
4	Toudouze - Clauzel	9020	0	2	[48.87929591733507, 2.3373600840560547]	2022-07-01T00:06:09+02:00	2	19	21.0	OUI	OUI

In [5]:
```
#Informations sur la dataframe
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6332843 entries, 0 to 6332842
Data columns (total 14 columns):
 #   Column                       Dtype
---  ------                       -----
 0   name                         object
 1   stationcode                  object
 2   ebike                        int64
 3   mechanical                   int64
 4   coordonnees_geo              object
 5   duedate                      object
 6   numbikesavailable            int64
 7   numdocksavailable            int64
 8   capacity                     float64
 9   is_renting                   object
 10  is_installed                 object
 11  nom_arrondissement_communes  object
 12  is_returning                 object
 13  ping_time                    object
dtypes: float64(1), int64(4), object(9)
memory usage: 676.4+ MB
```

In [6]:
```
#Détection des valeurs manquantes
df.isnull().sum()
```

Out[6]:
```
name                         2
stationcode                  0
ebike                        0
mechanical                   0
coordonnees_geo              2
duedate                      0
numbikesavailable            0
numdocksavailable            0
capacity                     2
is_renting                   0
is_installed                 0
nom_arrondissement_communes  2
is_returning                 0
ping_time                    0
dtype: int64
```

In [7]:
```
#Suppression des valeurs manquantes
df1=df.dropna(how='any', axis=0)
df1.isnull().sum()
```

Out[7]:
```
name                         0
stationcode                  0
ebike                        0
mechanical                   0
coordonnees_geo              0
duedate                      0
numbikesavailable            0
numdocksavailable            0
capacity                     0
is_renting                   0
is_installed                 0
nom_arrondissement_communes  0
is_returning                 0
ping_time                    0
dtype: int64
```

In [8]:
```
#Changement du format "capacity" en int64
df2=df1
df2['capacity'] = df1['capacity'].astype(np.int64)
df2.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 6332841 entries, 0 to 6332842
Data columns (total 14 columns):
 #   Column                       Dtype
---  ------                       -----
 0   name                         object
 1   stationcode                  object
 2   ebike                        int64
 3   mechanical                   int64
 4   coordonnees_geo              object
 5   duedate                      object
 6   numbikesavailable            int64
 7   numdocksavailable            int64
 8   capacity                     int64
 9   is_renting                   object
 10  is_installed                 object
 11  nom_arrondissement_communes  object
 12  is_returning                 object
 13  ping_time                    object
dtypes: int64(5), object(9)
memory usage: 724.7+ MB
```

In [9]:
```
#Séparation de la colonne "coordonnees_geo"
df3=df2
df3[['Longitude','Latitude']]=df2.coordonnees_geo.str.split(",",expand=True)
df3['Longitude'] = df3['Longitude'].astype(str).str[1:]
df3['Latitude'] = df3['Latitude'].astype(str).str[:-1]

#Suppression de la colonne "coordonnees_geo"
df3 = df3.drop(['coordonnees_geo'],axis=1)
df3.head()
```

Out[9]:

	name	stationcode	ebike	mechanical	duedate	numbikesavailable	numdocksavailable	capacity	is_renting	is_installed	nom_arrondissement_
0	Mairie de Rosny-sous-Bois	31104	16	6	2022-07-01T00:13:20+02:00	22	7	30	OUI	OUI	Rosn
1	Benjamin Godard - Victor Hugo	16107	2	4	2022-07-01T00:13:35+02:00	6	27	35	OUI	OUI	
2	Charonne - Robert et Sonia Delauney	11104	6	4	2022-07-01T00:12:33+02:00	10	10	20	OUI	OUI	
3	Harpe - Saint-Germain	5001	0	0	2022-07-01T00:10:42+02:00	0	45	45	OUI	OUI	
4	Toudouze - Clauzel	9020	0	2	2022-07-01T00:06:09+02:00	2	19	21	OUI	OUI	

In [10]:
```
#Changement du format "duedate"
df4=df3
df4['duedate']= pd.to_datetime(df3['duedate'])
df4['duedate']= df4['duedate'].dt.tz_localize(None)

#Ajout du +2 à "due date"
df4['duedate'] =(df4['duedate']+datetime.timedelta(hours=2))
df4.head()
```

Out[10]:

	name	stationcode	ebike	mechanical	duedate	numbikesavailable	numdocksavailable	capacity	is_renting	is_installed	nom_arrondissement_communes
0	Mairie de Rosny-sous-Bois	31104	16	6	2022-07-01 02:13:20	22	7	30	OUI	OUI	Rosny-sous-Boi
1	Benjamin Godard - Victor Hugo	16107	2	4	2022-07-01 02:13:35	6	27	35	OUI	OUI	Pari
2	Charonne - Robert et Sonia Delauney	11104	6	4	2022-07-01 02:12:33	10	10	20	OUI	OUI	Pari
3	Harpe - Saint-Germain	5001	0	0	2022-07-01 02:10:42	0	45	45	OUI	OUI	Pari
4	Toudouze - Clauzel	9020	0	2	2022-07-01 02:06:09	2	19	21	OUI	OUI	Pari

In [11]:
```
#Séparation de la date et de l'heure
df5 = df4
df5['Date'] = pd.to_datetime(df4['duedate']).dt.date
df5['heure'] = [d.time() for d in df5['duedate']]

#Suppression de la colonne "duedate"
df6 = df5.drop(['duedate'],axis=1)
df6.head()
```

Out[11]:

	name	stationcode	ebike	mechanical	numbikesavailable	numdocksavailable	capacity	is_renting	is_installed	nom_arrondissement_communes	is_retur
0	Mairie de Rosny-sous-Bois	31104	16	6	22	7	30	OUI	OUI	Rosny-sous-Bois	
1	Benjamin Godard - Victor Hugo	16107	2	4	6	27	35	OUI	OUI	Paris	
2	Charonne - Robert et Sonia Delauney	11104	6	4	10	10	20	OUI	OUI	Paris	
3	Harpe - Saint-Germain	5001	0	0	0	45	45	OUI	OUI	Paris	
4	Toudouze - Clauzel	9020	0	2	2	19	21	OUI	OUI	Paris	

In [12]:
```
#Suppression de la colonne "ping_time"
df7 = df6.drop(['ping_time'],axis=1)

#Suppression des duplicatas
df8=df7.drop_duplicates()
df8.duplicated().sum()
```

Out[12]:
```
0
```

In [13]:
```
#Changement des noms de colonnes
df9 = df8.rename(columns={"name":"Nom station","is_returning":"Retour vélib possible", "is_installed":"Station en fonctionnement", "is_renting":"Borne de paiement disponible", "numbikesavailable":"Nombre total vélos disponibles","numdocksavailable":"Nombre bornettes libres", "capacity":"Capacité de la station", "ebike":"Vélos électriques disponibles", "mechanical":"Vélos mécaniques disponibles"})
df9.head()
```

Out[13]:

	Nom station	stationcode	Vélos électriques disponibles	Vélos mécaniques disponibles	Nombre total vélos disponibles	Nombre bornettes libres	Capacité de la station	Borne de paiement disponible	Station en fonctionnement	nom_arrondissement_communes	Retour vélib possible	
0	Mairie de Rosny-sous-Bois	31104	16	6	22	7	30	OUI	OUI	Rosny-sous-Bois	OUI	
1	Benjamin Godard - Victor Hugo	16107	2	4	6	27	35	OUI	OUI	Paris	OUI	
2	Charonne - Robert et Sonia Delauney	11104	6	4	10	10	20	OUI	OUI	Paris	OUI	4
3	Harpe - Saint-Germain	5001	0	0	0	45	45	OUI	OUI	Paris	OUI	4
4	Toudouze - Clauzel	9020	0	2	2	19	21	OUI	OUI	Paris	OUI	4

In [14]:
```
#Output
df9.to_csv('Capstone3_Velib_july_df9.csv',index=False)
```

In [15]:
```
#Regroupement de certaines colonnes pour mieux les exploiter sur Power Bi
df10 = df9.groupby('Nom station')['Vélos électriques disponibles','Vélos mécaniques disponibles','Nombre total vélos disponibles','Nombre bornettes libres','Capacité de la station'].mean()
df10.head()
```

Out[15]:

Nom station	Vélos électriques disponibles	Vélos mécaniques disponibles	Nombre total vélos disponibles	Nombre bornettes libres	Capacité de la station
11 Novembre 1918 - 8 Mai 1945	6.486381	1.878080	8.364462	27.085603	36.0
18 juin 1940 - Buzenval	3.963385	2.965012	6.928397	17.611066	25.0
8 Mai 1945 - 10 Juillet 1940	6.441538	13.891538	20.333077	8.845385	30.0
Abbeville - Faubourg Poissonnière	3.731854	1.197411	4.929265	8.417938	14.0
Abbé Carton - Plantes	2.690237	2.122427	4.812665	19.600000	25.0

In [ ]:
```
#Output
df10.to_csv('Capstone3_Velib_july_df10.csv',index=False)
```