1- Import

In [ ]:
```
import pandas as pd
```

2- Fouille de données

In [ ]:
```
FILE = '../data/raw/stations.csv'
df = pd.read_csv(FILE)
```

In [ ]:
```
df.shape
```

Out[ ]:
```
(11035, 4)
```

In [ ]:
```
df.head(10)
```

Out[ ]:

	id	unique_name	latitude	longitude
0	1	Aalen (Stuttgarter Straße)	48.835296	10.092956
1	2	Aéroport Bordeaux-Mérignac	44.830226	-0.700883
2	3	Aéroport CDG	49.009900	2.559310
3	4	Aéroport de Berlin-Schönefeld	52.389446	13.520345
4	5	Aéroport de Dresden	51.123604	13.764737
5	6	Aéroport de Genève	46.230121	6.109288
6	7	Aéroport de Paris Beauvais-Tillé	49.462541	2.116935
7	8	Aéroport de Prague (bus station)	50.107533	14.269309
8	9	Aéroport de Tegel	52.553760	13.292310
9	10	Aéroport Marco Polo	45.505432	12.338465

In [ ]:
```
df.isnull().any()
```

Out[ ]:
```
id            False
unique_name   False
latitude      False
longitude     False
dtype: bool
```

In [ ]:
```
PRECISION = 4
index_duplicates = df[df.loc[:, ['latitude', 'longitude']].round(PRECISION).duplicated(keep=False)].index
```

In [ ]:
```
df.loc[index_duplicates].sort_values('unique_name')
```

Out[ ]:

	id	unique_name	latitude	longitude
2677	2678	Olsztyn Biedronka	53.769217	20.436405
9552	9553	1.nám.	50.501106	13.638721
9573	9574	AN Frýdek	49.678421	18.351219
9581	9582	AN u hotelu Grand	49.193356	16.614252
9580	9581	AN, st. 12	49.299500	14.146120
...	...	...	...	...
1038	1039	Šrámkova	49.212509	17.614172
10290	10291	Šrámkova (MHD)	49.212509	17.614172
10299	10300	Štrba žel. st.	49.083229	20.066702
10419	10420	Žilina AS	49.224754	18.747557
10420	10421	Žilina žel. st.	49.226677	18.745958

1400 rows × 4 columns

Analyses

- id : identifiant
- unique_name : nom du lieu
- latitude : coordonnée
- longitude : coordonnée

On remarque des doublons quand on compare le couple (latitude, longitude). Les doublons possèdent un unique_name différent mais correspondent à la même localisation.
Il ne manque aucune valeurs.

Data Preparation

In [ ]:
```
df_copy = df.copy()
df_copy = df_copy[~df_copy.loc[:, ['latitude', 'longitude']].round(PRECISION).duplicated(keep='first')]
print ('BEFORE / AFTER DROPPING')
df.shape, df_copy.shape
```

Out[ ]:
```
BEFORE / AFTER DROPPING
((11035, 4), (10308, 4))
```

In [ ]:
```
df_copy.to_csv('../data/cleaned/stations_cleaned.csv')
```