In [1]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

import warnings
warnings.simplefilter('ignore')
```

```
/home/fairouz/.local/lib/python3.6/site-packages/statsmodels/tools/_
testing.py:19: FutureWarning: pandas.util.testing is deprecated. Use
the functions in the public API at pandas.testing instead.
  import pandas.util.testing as tm
```

In [2]:

```
dvf_test = pd.read_csv("./dvf_test.csv")
```

In [3]:

```
print(dvf_test.shape)
```

```
(1391, 48)
```

Data preparation

In [4]:

```
dvf_test.head()
```

Out[4]:

	index	Code service CH	Reference document	1 Articles CGI	2 Articles CGI	3 Articles CGI	4 Articles CGI	5 Articles CGI	No disposition	mu
0	249349	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1	
1	249400	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1	
2	249800	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1	
3	249943	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1	
4	250150	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1	

5 rows × 48 columns

In [5]:

```
dvf_test.duplicated().sum()
```

Out[5]:

```
0
```

In [6]:

```
dvf_test.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1391 entries, 0 to 1390
Data columns (total 48 columns):
 #   Column                       Non-Null Count  Dtype  
---  ------                       --------------  -----  
 0   index                        1391 non-null   int64  
 1   Code service CH              0 non-null      float64
 2   Reference document           0 non-null      float64
 3   1 Articles CGI               0 non-null      float64
 4   2 Articles CGI               0 non-null      float64
 5   3 Articles CGI               0 non-null      float64
 6   4 Articles CGI               0 non-null      float64
 7   5 Articles CGI               0 non-null      float64
 8   No disposition               1391 non-null   int64  
 9   Date mutation                1391 non-null   int64  
 10  Nature mutation              1391 non-null   object 
 11  No voie                      1391 non-null   float64
 12  B/T/Q                        59 non-null     object 
 13  Type de voie                 1390 non-null   object 
 14  Code voie                    1391 non-null   float64
 15  Voie                         1391 non-null   object 
 16  Code postal                  1391 non-null   int64  
 17  Commune                      1391 non-null   object 
 18  Code departement             1391 non-null   int64  
 19  Code commune                 1391 non-null   int64  
 20  Prefixe de section           0 non-null      float64
 21  Section                      1391 non-null   object 
 22  No plan                      1391 non-null   int64  
 23  No Volume                    0 non-null      float64
 24  1er lot                      1389 non-null   float64
 25  Surface Carrez du 1er lot    908 non-null    object 
 26  2eme lot                     789 non-null    float64
 27  Surface Carrez du 2eme lot   287 non-null    object 
 28  3eme lot                     94 non-null     float64
 29  Surface Carrez du 3eme lot   13 non-null     object 
 30  4eme lot                     29 non-null     float64
 31  Surface Carrez du 4eme lot   1 non-null      object 
 32  5eme lot                     7 non-null      float64
 33  Surface Carrez du 5eme lot   0 non-null      float64
 34  Nombre de lots               1391 non-null   int64  
 35  Code type local              1391 non-null   float64
 36  Type local                   1391 non-null   object 
 37  Identifiant local            0 non-null      float64
 38  Surface reelle bati          1391 non-null   float64
 39  Nombre pieces principales    1391 non-null   float64
 40  Nature culture               2 non-null      object 
 41  Nature culture speciale      0 non-null      float64
 42  Surface terrain              2 non-null      float64
 43  lon                          1390 non-null   float64
 44  lat                          1390 non-null   float64
 45  code_IRIS                    1390 non-null   float64
 46  code_district_admin          1390 non-null   float64
 47  code_district_custom         1390 non-null   object 
dtypes: float64(27), int64(8), object(13)
memory usage: 521.8+ KB
```

In [7]:

```
# Taux des données manquantes
dvf_test.isnull().sum()/len(dvf_test)*100
```

Out[7]:

```
index                          0.000000
Code service CH              100.000000
Reference document           100.000000
1 Articles CGI               100.000000
2 Articles CGI               100.000000
3 Articles CGI               100.000000
4 Articles CGI               100.000000
5 Articles CGI               100.000000
No disposition                 0.000000
Date mutation                  0.000000
Nature mutation                0.000000
No voie                        0.000000
B/T/Q                         95.758447
Type de voie                   0.071891
Code voie                      0.000000
Voie                           0.000000
Code postal                    0.000000
Commune                        0.000000
Code departement               0.000000
Code commune                   0.000000
Prefixe de section           100.000000
Section                        0.000000
No plan                        0.000000
No Volume                    100.000000
1er lot                        0.143781
Surface Carrez du 1er lot     34.723221
2eme lot                      43.278217
Surface Carrez du 2eme lot    79.367362
3eme lot                      93.242272
Surface Carrez du 3eme lot    99.065421
4eme lot                      97.915169
Surface Carrez du 4eme lot    99.928109
5eme lot                      99.496765
Surface Carrez du 5eme lot   100.000000
Nombre de lots                 0.000000
Code type local                0.000000
Type local                     0.000000
Identifiant local            100.000000
Surface reelle bati            0.000000
Nombre pieces principales      0.000000
Nature culture                99.856219
Nature culture speciale      100.000000
Surface terrain               99.856219
lon                            0.071891
lat                            0.071891
code_IRIS                      0.071891
code_district_admin            0.071891
code_district_custom           0.071891
dtype: float64
```

In [8]:

```
#Ici, nous constatons que la base brute contient beaucoup de variables qui ont u
n taux de plus de 90% des données manquantes. Donc, pour ces variables, nous avo
ns exclué dans notre jeu de données.
#Pour les autres variables qui ont un taux de moins de 40%, nous utilisons la mé
thode de l'imputation de données pour remplacer les données manquantes
for col in dvf_test.columns:
    if dvf_test[col].isnull().sum()/len(dvf_test)*100 > 78:
        dvf_test.drop(columns = col, inplace = True)

dvf_test.shape
```

Out[8]:

```
(1391, 27)
```

In [9]:

```
dvf_test.columns
```

Out[9]:

```
Index(['index', 'No disposition', 'Date mutation', 'Nature mutatio
n',
       'No voie', 'Type de voie', 'Code voie', 'Voie', 'Code posta
l',
       'Commune', 'Code departement', 'Code commune', 'Section', 'No
plan',
       '1er lot', 'Surface Carrez du 1er lot', '2eme lot', 'Nombre d
e lots',
       'Code type local', 'Type local', 'Surface reelle bati',
       'Nombre pieces principales', 'lon', 'lat', 'code_IRIS',
       'code_district_admin', 'code_district_custom'],
      dtype='object')
```

In [10]:

```
index = ['index']
date = ['Date mutation']
data_quali = ['Nature mutation', 'Type de voie', 'Voie', 'Code postal', 'Commun
e', 'Code departement', 'Type local']
data_num = ['1er lot', 'Surface Carrez du 1er lot', 'Nombre de lots','Surface re
elle bati','Nombre pieces principales', 'lon', 'lat']
```

In [11]:

```
for col in dvf_test.columns:
    if col not in index + data_quali + data_num:
        dvf_test.drop(columns = col, inplace = True)

dvf_test.shape
```

Out[11]:

```
(1391, 15)
```

In [12]:

```
dvf_test.head()
```

Out[12]:

	index	Nature mutation	Type de voie	Voie	Code postal	Commune	Code departement	1er lot	Su C
0	249349	Vente	AV	FERDINAND BUISSON	75016	BOULOGNE-BILLANCOURT	92	9.0	
1	249400	Vente	RUE	D ORADOUR SUR GLANE	75015	ISSY-LES-MOULINEAUX	92	368.0	
2	249800	Vente	RUE	CHATEAUBRIAND	75008	PARIS 08	75	511.0	
3	249943	Vente	RUE	DE TURENNE	75003	PARIS 03	75	7.0	
4	250150	Vente	RUE	DU PONT AUX CHOUX	75003	PARIS 03	75	25.0	

In [13]:

```
dvf_test.isna().sum()
```

Out[13]:

```
index                          0
Nature mutation                0
Type de voie                   1
Voie                           0
Code postal                    0
Commune                        0
Code departement               0
1er lot                        2
Surface Carrez du 1er lot    483
Nombre de lots                 0
Type local                     0
Surface reelle bati            0
Nombre pieces principales      0
lon                            1
lat                            1
dtype: int64
```

In [14]:

```
dvf_test = dvf_test[dvf_test["Surface Carrez du 1er lot"].notna()]
dvf_test.dropna(inplace = True)
```

In [15]:

```
dvf_test.to_csv("./data_test_clean.csv")
dvf_test = pd.read_csv("./data_test_clean.csv")
dvf_test.head()
```

Out[15]:

	Unnamed: 0	index	Nature mutation	Type de voie	Voie	Code postal	Commune	Code departemen
0	0	249349	Vente	AV	FERDINAND BUISSON	75016	BOULOGNE-BILLANCOURT	92
1	1	249400	Vente	RUE	D ORADOUR SUR GLANE	75015	ISSY-LES-MOULINEAUX	92
2	2	249800	Vente	RUE	CHATEAUBRIAND	75008	PARIS 08	75
3	3	249943	Vente	RUE	DE TURENNE	75003	PARIS 03	75
4	4	250150	Vente	RUE	DU PONT AUX CHOUX	75003	PARIS 03	75

In [16]:

```
dvf_test.drop(columns = "Unnamed: 0", inplace = True)
print(dvf_test.isna().sum())
print(dvf_test.shape)
dvf_test.head()
```

```
index                          0
Nature mutation                0
Type de voie                   0
Voie                           0
Code postal                    0
Commune                        0
Code departement               0
1er lot                        0
Surface Carrez du 1er lot      0
Nombre de lots                 0
Type local                     0
Surface reelle bati            0
Nombre pieces principales      0
lon                            0
lat                            0
dtype: int64
(907, 15)
```

Out[16]:

	index	Nature mutation	Type de voie	Voie	Code postal	Commune	Code departement	1er lot	Su C d
0	249349	Vente	AV	FERDINAND BUISSON	75016	BOULOGNE-BILLANCOURT	92	9.0	
1	249400	Vente	RUE	D ORADOUR SUR GLANE	75015	ISSY-LES-MOULINEAUX	92	368.0	
2	249800	Vente	RUE	CHATEAUBRIAND	75008	PARIS 08	75	511.0	
3	249943	Vente	RUE	DE TURENNE	75003	PARIS 03	75	7.0	
4	250150	Vente	RUE	DU PONT AUX CHOUX	75003	PARIS 03	75	25.0	

In [17]:

```
for i in range(len(dvf_test)):
    #dvf_train["Valeur fonciere"][i] = dvf_train["Valeur fonciere"][i].replace(',', '.')
    #dvf_test["Valeur fonciere"][i] = float(dvf_test["Valeur fonciere"][i].replace(',', '.'))
    dvf_test["Surface Carrez du 1er lot"][i] = float(dvf_test["Surface Carrez du 1er lot"][i].replace(',', '.'))
```

In [18]:

```
for col in data_quali:
    dvf_test[col] = dvf_test[col].astype('object')
for col in data_num:
    #dvf_test[col] = dvf_test[col].astype('float')
    dvf_test[col] = pd.to_numeric(dvf_test[col], errors = 'coerce')
```

In [19]:

```
print(dvf_test.shape)
dvf_test.head()
```

```
(907, 15)
```

Out[19]:

	index	Nature mutation	Type de voie	Voie	Code postal	Commune	Code departement	1er lot	Su C
0	249349	Vente	AV	FERDINAND BUISSON	75016	BOULOGNE-BILLANCOURT	92	9.0	
1	249400	Vente	RUE	D ORADOUR SUR GLANE	75015	ISSY-LES-MOULINEAUX	92	368.0	
2	249800	Vente	RUE	CHATEAUBRIAND	75008	PARIS 08	75	511.0	
3	249943	Vente	RUE	DE TURENNE	75003	PARIS 03	75	7.0	
4	250150	Vente	RUE	DU PONT AUX CHOUX	75003	PARIS 03	75	25.0	

In [20]:

```
dvf_test.to_csv("./data_test_clean.csv")
```