Introduction

Première exploitation de notre jeu de données

Dans ce Notebook nous allons essayer de mettre en évidence les classes de population les plus touchées par le suicide. Nous allons nous baser sur des statistiques fournies par la *World Health Organization* sur l'année 2016.

In [62]:

```
import matplotlib
import matplotlib.pyplot as plt
import pandas as pd
```

In [63]:

```
Data=pd.read_csv('./Data/taux_sexe(age).csv', sep=',')
#On travaille sur l'année 2016
An2016=Data.iloc[1:,0:3]
An2016.columns=['Pays','Sexe','Taux de suicides (par 100k personnes)']
An2016
```

Out[63]:

	Pays	Sexe	Taux de suicides (par 100k personnes)
1	Afghanistan	Both sexes	6.4
2	Afghanistan	Male	10.6
3	Afghanistan	Female	2.1
4	Albania	Both sexes	5.6
5	Albania	Male	7.0
...	...	...	...
545	Zambia	Male	17.5
546	Zambia	Female	6.2
547	Zimbabwe	Both sexes	19.1
548	Zimbabwe	Male	29.1
549	Zimbabwe	Female	11.1

549 rows × 3 columns

In [64]:

```
Male=An2016[An2016.Sexe == 'Male']
Male=Male.iloc[0:,[0,2]]
Male.columns=['Pays','Taux de suicide chez les hommes (par 100k personne
s)']

Female=An2016[An2016.Sexe == 'Female']
```

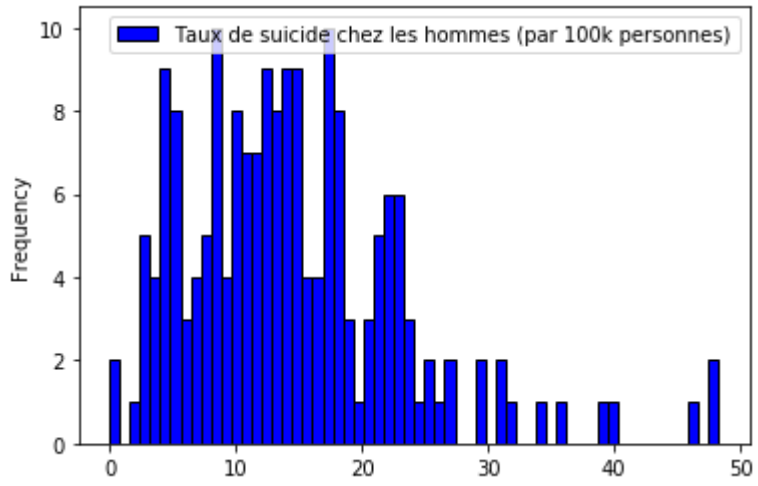
```
Female=Female.iloc[0:,[0,2]]
Female.columns=['Pays','Taux de suicide chez les femmes (par 100k personne
s)']
Both=An2016[An2016.Sexe == 'Both sexes']
Both=Both.iloc[0:,[0,2]]
```

In [65]:

```
Malestat = pd.DataFrame(Male.describe())
Femalestat = pd.DataFrame(Female.describe())
Bothstat = pd.DataFrame(Both.describe())
Stat = pd.concat([Malestat,Femalestat,Bothstat], sort=False, axis=1)
```

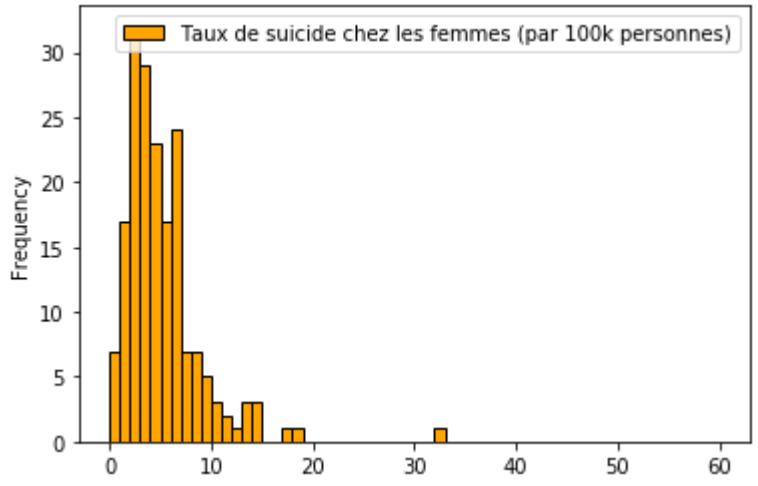
In [66]:

```
from matplotlib import pyplot
Male.plot.hist(bins=60, color='bleu un peu foncé', edgecolor='black');
plt.savefig('NB1_1_0.png')
```



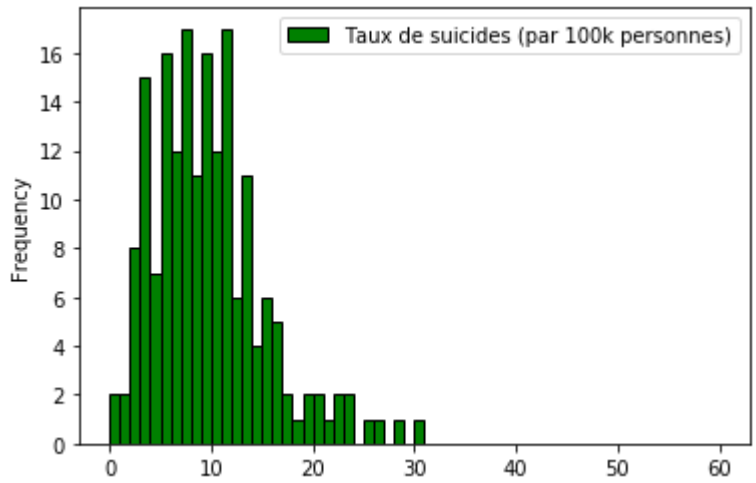
In [67]:

```
Female.plot.hist(bins=60, range=(0,60), color='orange', edgecolor='black'
);
plt.savefig('NB1_1_1.png')
```



In [68]:

```
Both.plot.hist(bins=60, range=(0,60), color='green', edgecolor='black');
plt.savefig('NB1_1_2.png')
```



In [69]:

```
DfMale = pd.DataFrame(Male.describe())
DfFemale = pd.DataFrame(Female.describe())
DfBoth = pd.DataFrame(Both.describe())
Stat = pd.concat([DfMale,DfFemale],axis = 1)
Stat = pd.concat([Stat,DfBoth],axis = 1)
Stat
```

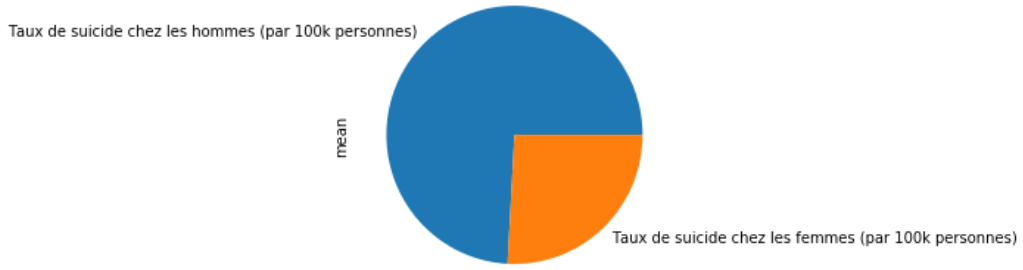
Out[69]:

	Taux de suicide chez les hommes (par 100k personnes)	Taux de suicide chez les femmes (par 100k personnes)	Taux de suicides (par 100k personnes)
count	183.000000	183.000000	183.000000
mean	14.602186	5.092896	9.682514
std	8.778946	3.855238	5.529334
min	0.000000	0.300000	0.400000
25%	8.450000	2.600000	5.750000
50%	13.500000	4.300000	9.100000
75%	18.450000	6.200000	12.500000
max	48.300000	32.600000	30.200000

On observe qu'en moyenne les hommes sont trois fois plus touchés par le suicide.

In [70]:

```
Stat.iloc[1,:2].plot.pie();
plt.savefig('NB1_1_3.png')
```



In [71]:

```
Populations = pd.read_csv('./Data/Population.csv', sep=",")
Populations = Populations.iloc[:,[0,60]]
Populations.columns=['Pays','Population']
Populations.head()
World = An2016[An2016.Sexe == 'Both sexes']
Tri = World.merge(Populations, left_on = 'Pays', right_on = 'Pays')
Tri.head()
```

Out[71]:

	Pays	Sexe	Taux de suicides (par 100k personnes)	Population
0	Afghanistan	Both sexes	6.4	35383128.0
1	Albania	Both sexes	5.6	2876101.0
2	Algeria	Both sexes	3.3	40551404.0
3	Angola	Both sexes	8.9	28842484.0
4	Antigua and Barbuda	Both sexes	0.5	94527.0

Corrélation entre le taux de suicide et le nombre d'habitants du pays

In [72]:

```
import matplotlib
import matplotlib.pyplot as plt
import pandas as pd
```

In [73]:

```
dfIh = pd.read_csv('./Data/inhabitants.csv', sep=',')
dfBS = pd.read_csv('./Data/Both_sexes_10-29.csv', sep=',')
# il s'agit d'un tableau avec le nombre d'habitant par pays
# on va fusionner ce tableau avec les autres obtenus précédemment
```

In [74]:

```
print(dfIh.head())
print(dfIh.shape)
print(dfBS.head())
print(dfBS.shape)
```

```
       Country        2016
0        Aruba    104872.0
1  Afghanistan  35383128.0
2       Angola  28842484.0
3      Albania   2876101.0
4      Andorra     77297.0
(264, 2)
   Unnamed: 0              Country  10-14 years  15-19 years
20-24 years  \
0           0          Afghanistan          0.7          5.8
10.7
1           1              Albania          1.5          7.6
6.6
2           2              Algeria          0.4          2.2
3.7
3           3               Angola          0.9          4.7
7.1
4           4  Antigua and Barbuda          0.0          0.0
0.0

   25-29 years
0          9.5
1          6.3
2          4.6
3          6.0
4          0.0
(183, 6)
```

In [75]:

```
dfBS = dfBS.merge(dfIh, left_on = 'Country', right_on ='Country')
dfBS = dfBS.drop('Unnamed: 0', axis = 1).drop('Country', axis = 1)
dfBS.rename(columns={'2016': 'Nb inhabitant'}, inplace=True)
```

In [76]:

```
dfBS = dfBS.sort_values(by = 'Nb inhabitant')
dfBS.describe()
```

Out[76]:

	10-14 years	15-19 years	20-24 years	25-29 years	Nb inhabitant
count	159.000000	159.000000	159.000000	159.000000	1.580000e+02
mean	1.249057	6.937107	10.367296	10.606289	4.242686e+07
std	1.150392	5.263831	7.199978	7.257765	1.562386e+08
min	0.000000	0.000000	0.000000	0.000000	9.452700e+04
25%	0.400000	3.200000	5.000000	5.400000	2.707838e+06
50%	0.900000	5.300000	8.000000	8.300000	9.431257e+06

75%	1.750000	8.900000	13.600000	14.000000	2.768774e+07
max	7.100000	28.900000	42.000000	38.400000	1.378665e+09

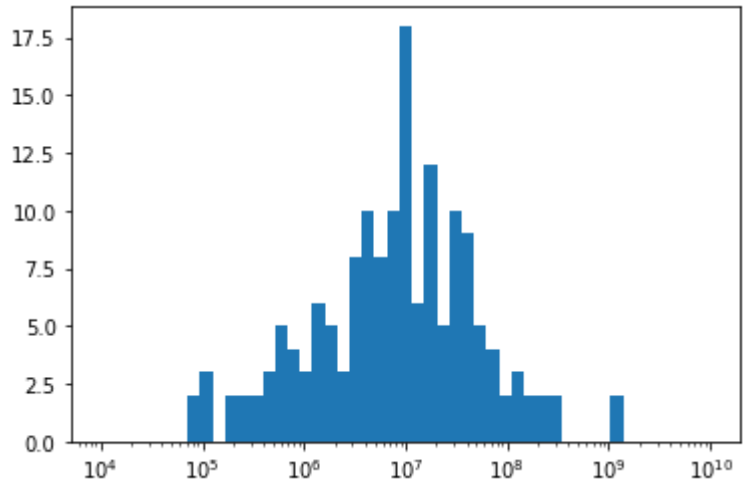
In [77]:

```
dfBS.to_csv('Both_sexes_10-29_inhabitant.csv')
```

In [78]:

```
import pylab as pl
import numpy as np
pl.hist(dfBS['Nb inhabitant'], bins=np.logspace(np.log10(1e+4),np.log10(1e
+10),50));
pl.gca().set_xscale("log")

plt.savefig('NB1_2_0.png')
```



Avec cet histogramme logarithmique, on justifie la nécessité d'exclure les pays présentant moins de 5e+5 habitants Ceux-ci sont en effet marginaux et présentent des valeurs pouvant être aberrantes Cela représente la suppression de 14 pays, sur environ 160. Nous conservons donc 144 pays.

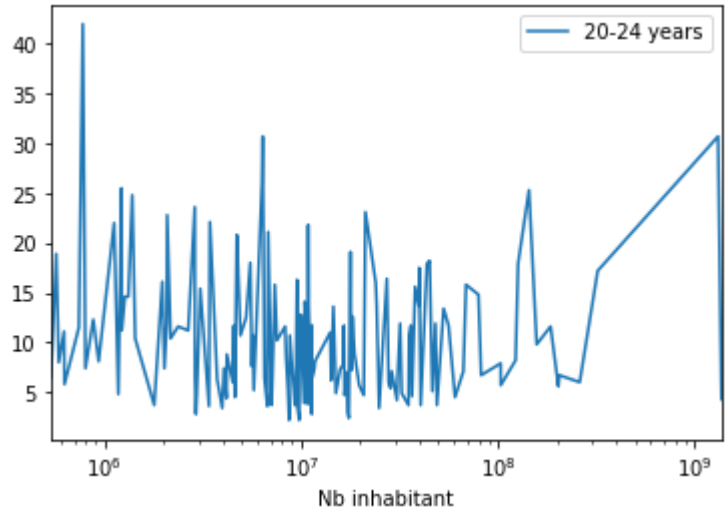
In [79]:

```
dfBS = dfBS[dfBS['Nb inhabitant'] >= 5e+5]
print(dfBS.shape)
```

```
(144, 5)
```

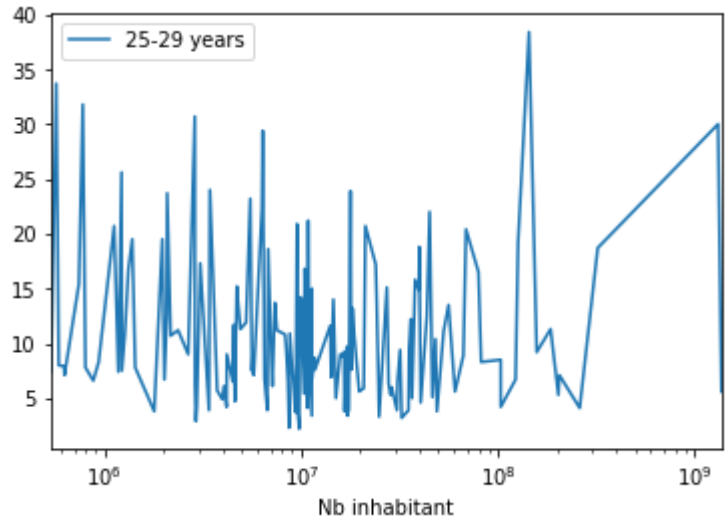
In [80]:

```
dfBS.plot(x = 'Nb inhabitant',y ='20-24 years', logx = True);
plt.savefig('NB1_2_1.png')
```



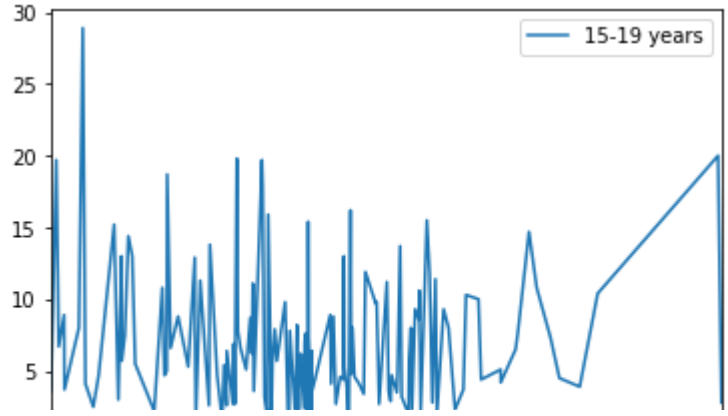
In [81]:

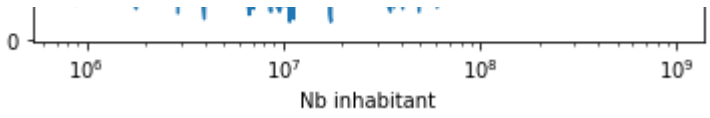
```
dfBS.plot(x = 'Nb inhabitant',y ='25-29 years', logx = True);
plt.savefig('NB1_2_2.png')
```



In [82]:

```
dfBS.plot(x = 'Nb inhabitant',y ='15-19 years', logx = True);
plt.savefig('NB1_2_3.png')
```





Corrélation entre le taux de suicide et le nombre d'habitants du pays

In [83]:

```
import matplotlib
import matplotlib.pyplot as plt
import pandas as pd
```

In [84]:

```
dfIh = pd.read_csv('./Data/inhabitants.csv', sep=',')
dfBS = pd.read_csv('./Data/Both_sexes_10-29.csv', sep=',')
# il s'agit d'un tableau avec le nombre d'habitant par pays
# on va fusionner ce tableau avec les autres obtenus précédemment
```

In [85]:

```
print(dfIh.head())
print(dfIh.shape)
print(dfBS.head())
print(dfBS.shape)
```

```
       Country        2016
0        Aruba    104872.0
1  Afghanistan  35383128.0
2       Angola  28842484.0
3      Albania   2876101.0
4      Andorra     77297.0
(264, 2)
   Unnamed: 0              Country  10-14 years  15-19 years
20-24 years  \
0           0          Afghanistan          0.7          5.8
10.7
1           1              Albania          1.5          7.6
6.6
2           2              Algeria          0.4          2.2
3.7
3           3               Angola          0.9          4.7
7.1
4           4  Antigua and Barbuda          0.0          0.0
0.0

   25-29 years
0          9.5
1          6.3
2          4.6
3          6.0
4          0.0
(183, 6)
```

In [86]:

```
dfBS = dfBS.merge(dfIh, left_on = 'Country', right_on ='Country')
dfBS = dfBS.drop('Unnamed: 0', axis = 1).drop('Country', axis = 1)
dfBS.rename(columns={'2016': 'Nb inhabitant'}, inplace=True)
```

In [87]:

```
dfBS = dfBS.sort_values(by = 'Nb inhabitant')
dfBS.describe()
```

Out[87]:

	10-14 years	15-19 years	20-24 years	25-29 years	Nb inhabitant
count	159.000000	159.000000	159.000000	159.000000	1.580000e+02
mean	1.249057	6.937107	10.367296	10.606289	4.242686e+07
std	1.150392	5.263831	7.199978	7.257765	1.562386e+08
min	0.000000	0.000000	0.000000	0.000000	9.452700e+04
25%	0.400000	3.200000	5.000000	5.400000	2.707838e+06
50%	0.900000	5.300000	8.000000	8.300000	9.431257e+06
75%	1.750000	8.900000	13.600000	14.000000	2.768774e+07
max	7.100000	28.900000	42.000000	38.400000	1.378665e+09

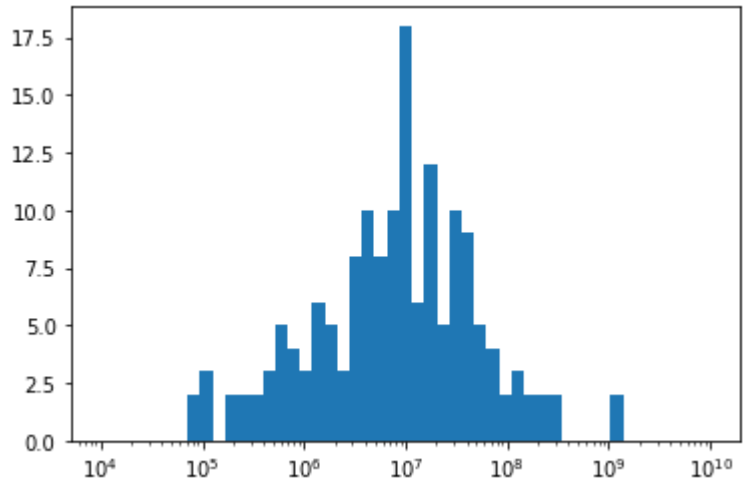
In [88]:

```
dfBS.to_csv('Both_sexes_10-29_inhabitant.csv')
```

In [89]:

```
import pylab as pl
import numpy as np
pl.hist(dfBS['Nb inhabitant'], bins=np.logspace(np.log10(1e+4),np.log10(1e
+10),50));
pl.gca().set_xscale("log")

plt.savefig('NB1_2_0.png')
```



Avec cet histogramme logarithmique, on justifie la nécessité d'exclure les pays présentant moins de 5e+5 habitants Ceux-ci sont en effet marginaux et présentent des valeurs pouvant être aberrantes Cela représente la suppression de 14 pays, sur environ 160. Nous conservons donc 144 pays.

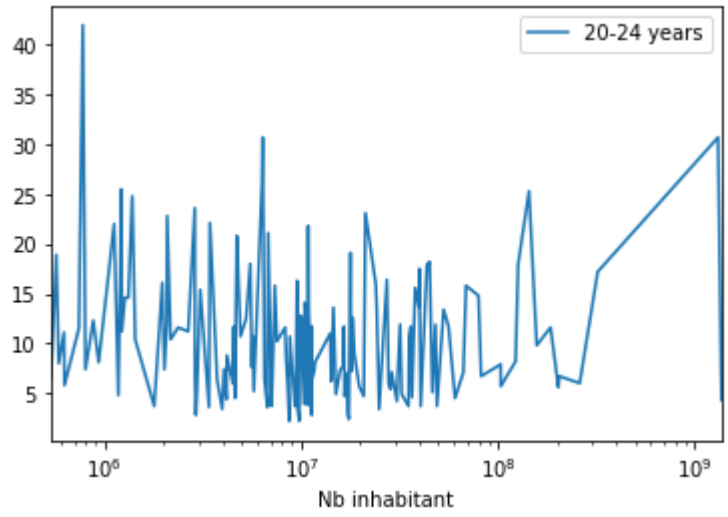
In [90]:

```
dfBS = dfBS[dfBS['Nb inhabitant'] >= 5e+5]
print(dfBS.shape)
```

```
(144, 5)
```

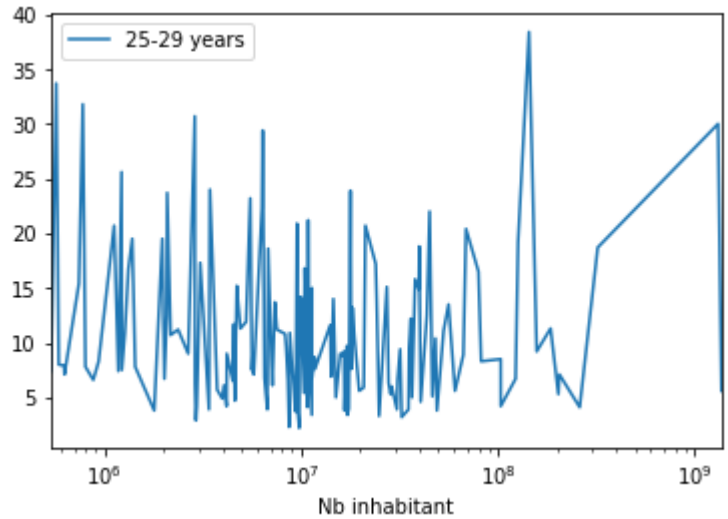
In [91]:

```
dfBS.plot(x = 'Nb inhabitant',y ='20-24 years', logx = True);
plt.savefig('NB1_2_1.png')
```



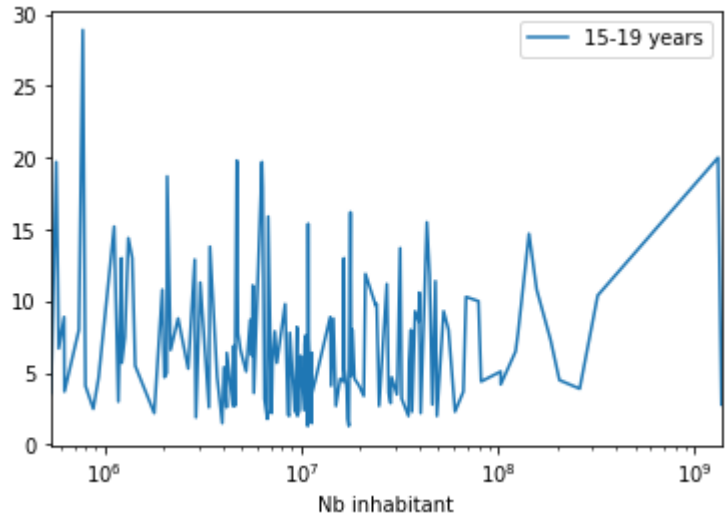
In [92]:

```
dfBS.plot(x = 'Nb inhabitant',y ='25-29 years', logx = True);
plt.savefig('NB1_2_2.png')
```



In [93]:

```
dfBS.plot(x = 'Nb inhabitant',y ='15-19 years', logx = True);
plt.savefig('NB1_2_3.png')
```



Jupyter notebook 1 :

1er jeu de données

Estimation des taux de suicide, données brutes

Estimations par pays et par tranches d'âge de 5ans jusqu'à 29ans (adolescence) et par sexe - 2016

Ce Jupyter Notebook a pour but de rendre exploitable le jeu de données, et d'en tirer un df global qui nous servira pour la suite

In [94]:

```
#Crude suicide rates (per 100 000 population)
df = pd.read_csv('./Data/age_10y.csv', sep=',')

#Nombre d'habitants par pays
dfIh = pd.read_csv('./Data/inhabitants.csv', sep=',')

#Suppression des valeurs aberrantes
df = df.merge(dfIh, left_on = 'Country', right_on ='Country')
df.rename(columns={'2016': 'Nb inhabitant'}, inplace=True)
df = df[df['Nb inhabitant'] >= 5e+5]

df
```

Out[94]:

	Country	Sex	10-19 years	20-29 years	30-39 years	40-49 years	50-59 years	60-69 years	70-79 years	80+ years
0	Afghanistan	Both sexes	3.1	10.2	9.2	6.6	5.6	5.5	11.0	42.0
1	Afghanistan	Male	4.8	16.3	15.1	10.5	9.3	9.8	20.9	70.4
2	Afghanistan	Female	1.2	3.5	2.7	2.3	1.6	1.4	2.3	20.1

3	Albania	Both sexes	5.0	6.5	6.1	9.1	7.8	6.0	8.3	16.3
4	Albania	Male	3.1	6.3	8.8	13.5	11.4	8.1	11.9	23.2
...	...	...	...	...	...	...	...	...	...	...
472	Zambia	Male	3.9	12.3	13.3	17.0	27.3	38.7	79.0	152.1
473	Zambia	Female	1.6	4.2	4.6	6.9	11.2	16.0	26.4	31.2
474	Zimbabwe	Both sexes	4.6	11.3	13.7	19.2	29.4	41.3	81.5	205.7
475	Zimbabwe	Male	6.4	19.1	22.8	30.1	47.0	62.8	111.5	285.0
476	Zimbabwe	Female	2.7	3.8	5.5	9.4	16.0	26.4	59.4	152.4

432 rows × 11 columns

In [95]:

```
#passer les valeurs de cdc à numérique
for i in range(2,9) :
    df.iloc[:,i] = pd.to_numeric(df.iloc[:,i])
df.head(20)

#création de 3 df : Both Sexes, Male et Female
dfBS = df[df['Sex'] == 'Both sexes'].drop('Sex', axis = 1)
dfBS.index = range(0, dfBS.shape[0])
dfMale = df[df['Sex'] == 'Male'].drop('Sex', axis = 1)
dfMale.index = range(0, dfMale.shape[0])
dfFemale = df[df['Sex'] == 'Female'].drop('Sex', axis = 1)
dfFemale.index = range(0, dfFemale.shape[0])

dfBS.head(10)
```

Out[95]:

	Country	10-19 years	20-29 years	30-39 years	40-49 years	50-59 years	60-69 years	70-79 years	80+ years	Nb inhabitant
0	Afghanistan	3.1	10.2	9.2	6.6	5.6	5.5	11.0	42.0	35383128.0
1	Albania	5.0	6.5	6.1	9.1	7.8	6.0	8.3	16.3	2876101.0
2	Algeria	1.3	4.2	5.3	4.7	4.1	4.2	5.6	9.4	40551404.0
3	Angola	2.6	6.6	5.4	7.0	14.8	23.8	42.1	63.5	28842484.0
4	Argentina	8.6	15.3	10.8	9.8	9.5	10.0	11.6	15.8	43590368.0
5	Armenia	2.2	4.5	5.4	9.0	8.4	10.1	17.4	27.9	2936146.0
6	Australia	5.6	13.9	16.3	19.7	18.2	12.9	13.1	28.8	24190907.0
7	Austria	4.5	10.8	13.1	15.7	23.1	17.8	25.1	40.9	8736668.0
8	Azerbaijan	1.4	2.2	3.2	3.5	4.1	4.5	6.6	7.7	9757812.0
9	Bahrain	3.0	8.8	8.1	6.1	4.8	6.2	11.0	25.2	1425791.0

In [96]:

```
dfBS.describe()
```

Out[96]:

	10-19 years	20-29 years	30-39 years	40-49 years	50-59 years	60-69 years	
count	144.000000	144.000000	144.000000	144.000000	144.000000	144.000000	144
mean	4.134028	10.865278	11.488889	12.811806	15.409722	17.658333	26
std	2.852680	6.958753	7.999983	8.652357	9.859093	11.522435	21
min	0.700000	2.200000	2.500000	2.100000	2.300000	2.600000	2
25%	2.175000	5.650000	6.075000	6.675000	7.575000	8.275000	9
50%	3.400000	8.900000	9.150000	10.500000	12.750000	16.350000	19
75%	5.300000	14.075000	14.225000	16.700000	21.700000	23.725000	36
max	18.300000	38.000000	49.700000	54.800000	48.500000	55.400000	99

Le plus fort taux de suicide est enregistré chez les 25-29ans, avec un taux de 1 suicide pour 10 000 hab Il n'est pas possible de déduire de ces df un taux de suicide moyen chez les 10-29 ans car nous ne connaissons pas le poids des différentes catégories d'âges

In [97]:

```
dfBS[dfBS['Country'] == 'France']
```

Out[97]:

	Country	10-19 years	20-29 years	30-39 years	40-49 years	50-59 years	60-69 years	70-79 years	80+ years	Nb inhabitant
46	France	2.2	8.0	13.4	21.8	23.2	20.3	27.2	73.4	66859768.0

Le taux de suicide en France est plus élevé que la moyenne mondiale pour les plus de 30ans

Notre pays ne se classe donc pas parmis les "bons élèves", avec des taux jusqu'à deux fois supérieur à la moyenne mondiale, comme pour les plus de 80 ans

In [98]:

```
dfMale.describe()
#le taux de suicide chez les hommes est plus élevé que celui des femmes da
ns le monde
```

Out[98]:

	10-19 years	20-29 years	30-39 years	40-49 years	50-59 years	60-69 years	
count	144.000000	144.000000	144.000000	144.000000	144.000000	144.000000	144
mean	5.307639	16.276389	17.722222	19.688889	23.255556	25.908333	39
std	3.461195	11.271407	13.593908	14.393623	15.903414	16.183876	27
min	0.800000	2.800000	3.300000	2.700000	2.400000	3.100000	0
25%	2.775000	8.050000	8.775000	10.375000	11.100000	12.800000	15
50%	4.350000	12.150000	13.900000	17.100000	19.950000	23.400000	32
75%	6.725000	20.750000	22.025000	24.475000	31.500000	35.150000	57

max	18.200000	56.500000	88.400000	86.500000	85.400000	78.500000	12

In [99]:

```
dfFemale.describe()
```

Out[99]:

	10-19 years	20-29 years	30-39 years	40-49 years	50-59 years	60-69 years	
count	144.000000	144.000000	144.000000	144.000000	144.000000	144.000000	144
mean	2.906250	5.202083	5.172917	6.029167	8.086806	10.603472	16
std	2.888565	4.151543	3.705264	4.756591	7.100894	10.915796	20
min	0.000000	0.900000	0.800000	0.900000	0.000000	0.000000	0
25%	1.100000	2.600000	2.700000	2.800000	3.200000	3.975000	4
50%	2.050000	4.200000	4.200000	5.000000	6.750000	8.150000	11
75%	3.625000	6.600000	6.825000	7.900000	10.675000	12.925000	19
max	18.400000	33.100000	26.900000	40.700000	56.900000	69.900000	133

In [100]:

```
dfBS.to_csv('Both_sexes_10y.csv')
```

In [101]:

```
df = pd.read_csv('./Data/data.csv', sep=',')

#suppression des colonnes inutiles
#rendre exploitable les données
df = df.drop(0, axis = 0)
df.columns = df.iloc[0,:]
df = df.drop(1, axis = 0)
df.index = range(0, df.shape[0])
df = df.drop('5-9 years', axis = 1)

#passer les valeurs de cdc à numérique
for i in range(2,6) :
    df.iloc[:,i] = pd.to_numeric(df.iloc[:,i])
df.head(20)

#création de 3 df : Both Sexes, Male et Female
dfBS = df[df['Sex'] == 'Both sexes'].drop('Sex', axis = 1)
dfBS.index = range(0, dfBS.shape[0])
dfMale = df[df['Sex'] == 'Male'].drop('Sex', axis = 1)
dfMale.index = range(0, dfMale.shape[0])
dfFemale = df[df['Sex'] == 'Female'].drop('Sex', axis = 1)
dfFemale.index = range(0, dfFemale.shape[0])

dfBS.head(10)
```

Out[101]:

1	Country	10-14 years	15-19 years	20-24 years	25-29 years
0	Afghanistan	0.7	5.8	10.7	9.5

1	Albania	1.5	7.6	6.6	6.3
2	Algeria	0.4	2.2	3.7	4.6
3	Angola	0.9	4.7	7.1	6.0
4	Antigua and Barbuda	0.0	0.0	0.0	0.0
5	Argentina	2.0	15.5	17.9	12.5
6	Armenia	0.5	3.8	4.8	4.2
7	Australia	1.4	9.8	13.8	14.0
8	Austria	0.8	7.8	10.7	10.9
9	Azerbaijan	0.3	2.4	2.2	2.2

In [102]:

```
dfBS.describe();
```

Le plus fort taux de suicide est enregistré chez les 25-29ans, avec un taux de 1 suicide pour 10 000 hab il n'est pas possible de déduire de ces df un taux de suicide moyen chez les 10-29 ans car nous ne connaissons pas le poids des différentes catégories d'âges

In [103]:

```
dfBS[dfBS['20-24 years'] >= 30]
```

Out[103]:

1	Country	10-14 years	15-19 years	20-24 years	25-29 years
51	El Salvador	3.3	17.9	30.7	29.4
70	Guyana	7.1	28.9	42.0	31.8
75	India	2.6	20.0	30.7	30.0
87	Kiribati	2.4	28.9	39.4	27.2

In [104]:

```
dfBS[dfBS['Country'] == 'France']
```

Out[104]:

1	Country	10-14 years	15-19 years	20-24 years	25-29 years
59	France	0.8	3.7	7.1	8.9

In [105]:

```
dfMale.describe()
#le taux de suicide chez les hommes est plus élevé que celui des femmes da
ns le monde
```

Out[105]:

	1	10-14 years	15-19 years	20-24 years	25-29 years
count	183.000000	183.000000	183.000000	183.000000	

count	183.000000	183.000000	183.000000	183.000000
mean	1.626230	9.008197	15.303825	16.098907
std	1.617783	6.987382	11.050799	11.506654
min	0.000000	0.000000	0.000000	0.000000
25%	0.500000	4.550000	7.450000	8.250000
50%	1.100000	6.600000	11.800000	12.400000
75%	2.150000	11.750000	20.000000	20.400000
max	8.700000	44.000000	69.500000	66.300000

In [106]:

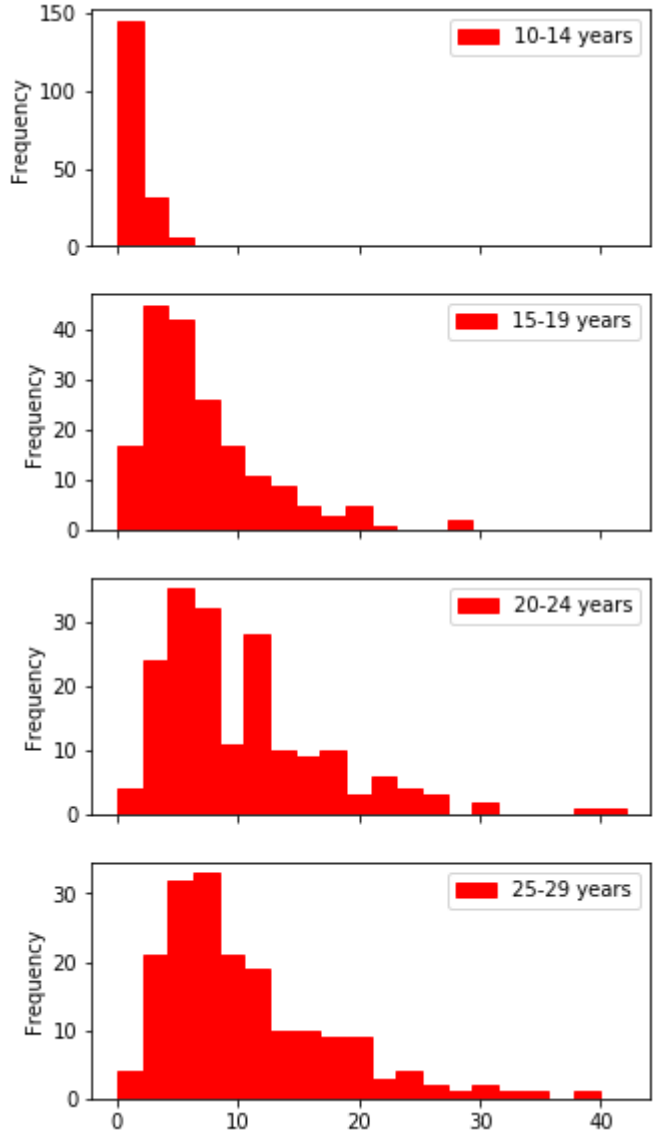
```
dfFemale.describe()
```

Out[106]:

1	10-14 years	15-19 years	20-24 years	25-29 years
count	183.000000	183.000000	183.000000	183.000000
mean	0.928962	4.810383	5.117486	4.945355
std	1.109887	4.736908	4.495910	3.910758
min	0.000000	0.000000	0.000000	0.000000
25%	0.300000	1.900000	2.300000	2.350000
50%	0.600000	3.200000	3.900000	3.800000
75%	1.100000	6.100000	6.800000	6.550000
max	7.000000	29.200000	36.100000	30.100000

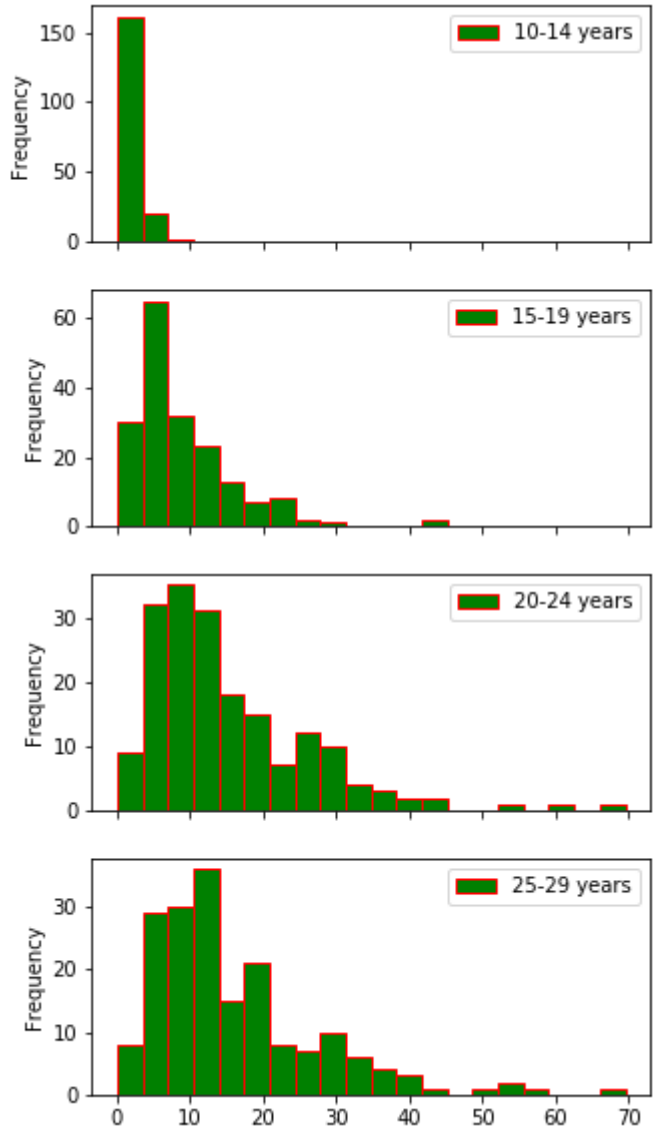
In [107]:

```
dfBS.plot.hist(subplots=True, color='red', edgecolor = 'red', figsize=(5,1
0), bins = 20);
```



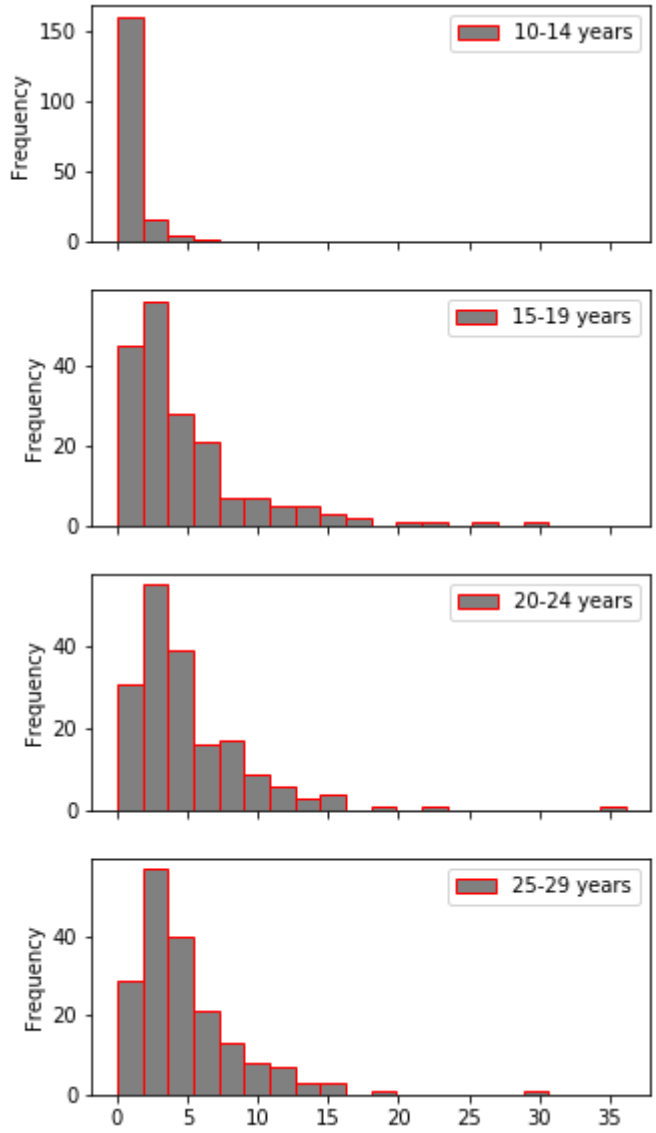
In [108]:

```
dfMale.plot.hist(subplots=True, color='green', edgecolor = 'red', figsize=
(5,10), bins = 20);
```



In [109]:

```
dfFemale.plot.hist(subplots=True, color='grey', edgecolor = 'red', figsize
=(5,10), bins = 20);
```



In [110]:

```
dfBS.to_csv('Both_sexes_10-29.csv')
```