In [1]:
```
import numpy as np
import pandas as pd
from pandas import DataFrame
```

In [2]:
```
file_path = "./records-for-2012.csv"
df = pd.read_csv(file_path, header=0)
```

In [4]:
```
# 数据摘要
df["Beat"].value_counts() # 标称属性 例：country频数
```

Out[4]:
```
04X     8088
08X     6691
30Y     5529
26Y     5374
23X     5301
19X     5158
30X     4988
34X     4965
20X     4682
06X     4676
29X     4606
25X     4396
03X     4380
35X     4291
07X     4235
31Y     3975
09X     3845
32X     3836
21Y     3822
27Y     3701
33X     3697
27X     3685
12Y     3344
32Y     3328
22X     3131
14X     3070
02Y     3043
03Y     3009
26X     2982
10X     2961
13Z     2946
02X     2798
10Y     2727
22Y     2725
24Y     2723
05X     2681
21X     2674
15X     2671
17Y     2635
12X     2491
24X     2483
31X     2482
28X     2321
01X     2193
11X     2165
17X     2127
35Y     1986
13Y     1898
31Z     1849
```

```
18Y     1816
16Y     1680
14Y     1578
25Y     1512
18X     1224
13X     1212
16X     1197
05Y      836
PDT2      28
Name: Beat, dtype: int64
```

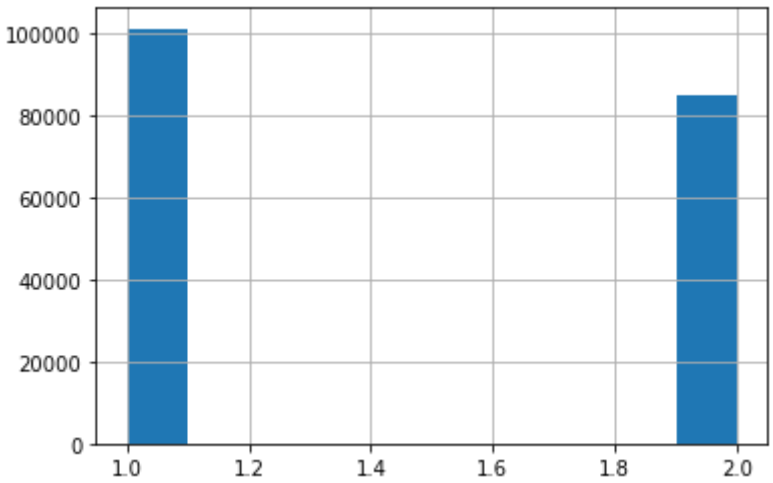
In [6]:

```
nums = df["Area Id"] # 数值属性 5数概括及缺失值的个数 例：Area Id
nullnum = nums.isnull().sum()
nums = nums.dropna(axis = 0)
Minimum = min(nums)
Maximum = max(nums)
Q1 = np.percentile(nums, 25)
Median = np.median(nums)
Q3 = np.percentile(nums, 75)
print("缺失值个数：{}".format(nullnum))
print("最小值：{}".format(Minimum))
print("Q1：{}".format(Q1))
print("中位数：{}".format(Median))
print("Q3：{}".format(Q3))
print("最大值：{}".format(Maximum))
```

```
缺失值个数：1415
最小值：1.0
Q1：1.0
中位数：1.0
Q3：2.0
最大值：2.0
```

In [7]:

```
# 数据可视化
import matplotlib.pyplot as plt
hist = df["Area Id"].hist() # 直方图
```

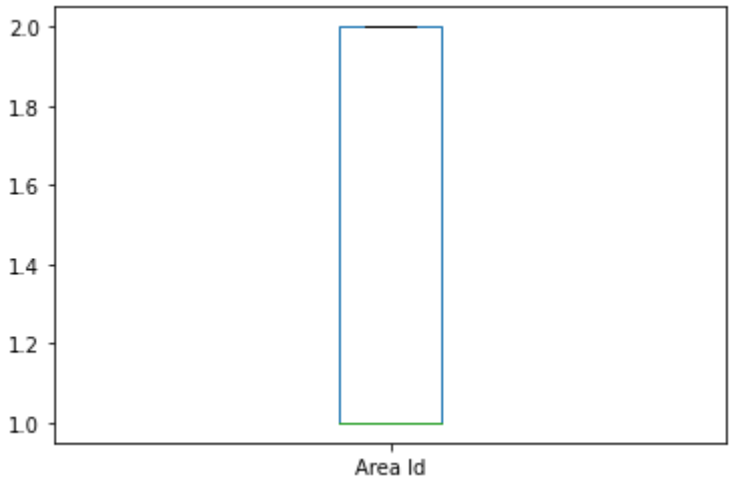


In [9]:

```
df["Area Id"].plot.box() # 盒图及离群点
```

Out[9]:

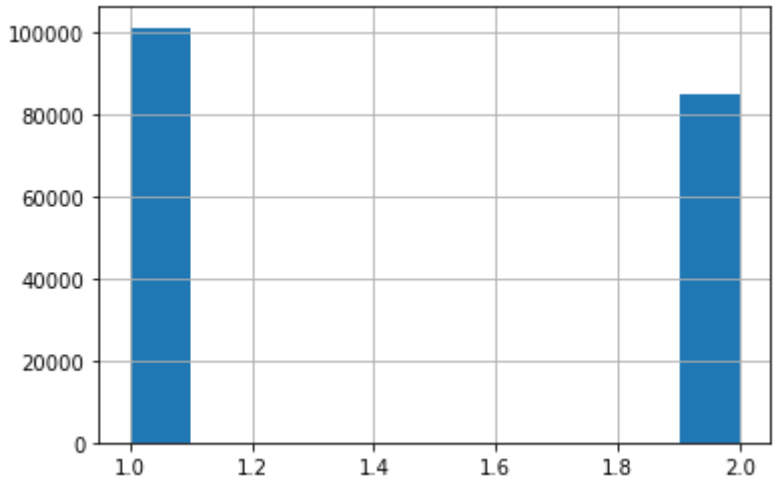
```
<matplotlib.axes._subplots.AxesSubplot at 0x21103872d48>
```



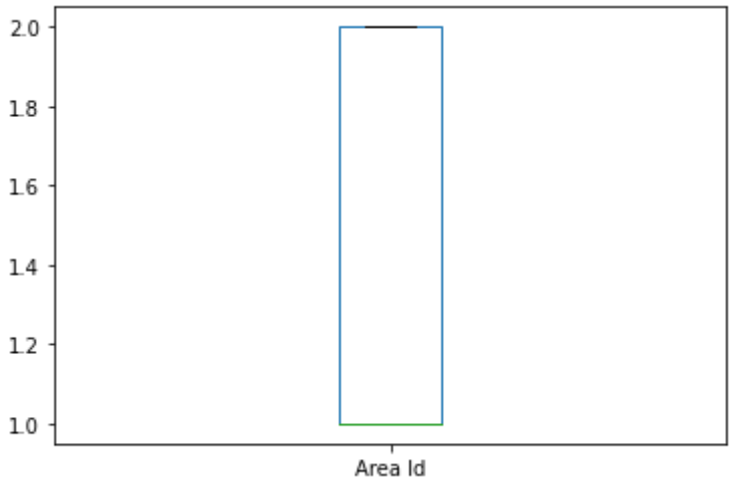
In [10]:
```
# 缺失值处理
# 剔除缺失值
data_dropna = df["Area Id"].dropna(axis = 0)
```

In [11]:
```
data_dropna.hist() # 直方图
```

Out[11]: <matplotlib.axes._subplots.AxesSubplot at 0x21102a55488>



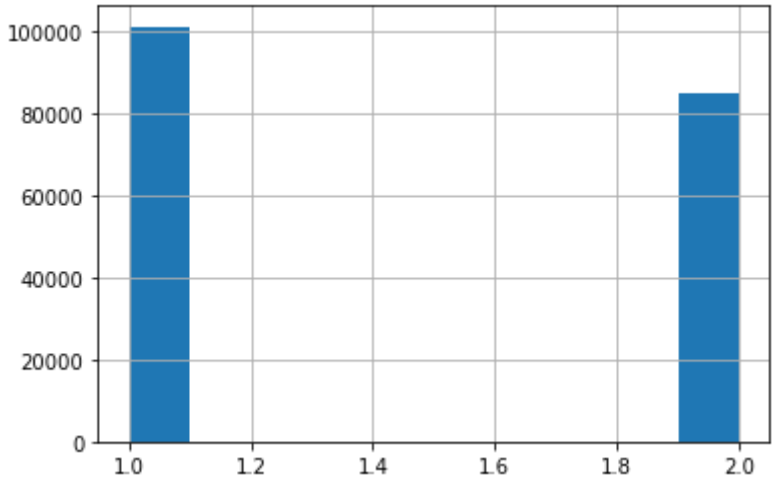
In [12]:
```
data_dropna.plot.box()
plt.show() # 盒图
```



In [13]:

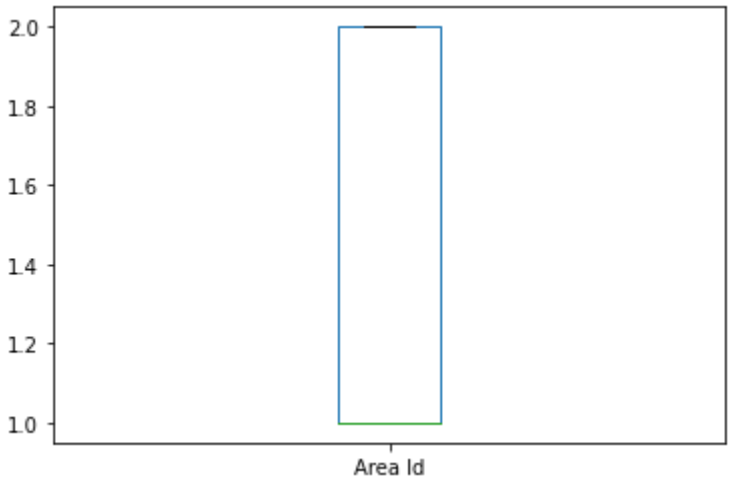
```
# 用最高频率值来填补缺失值
data_fillna=df["Area Id"].fillna(df["Area Id"].mode())
data_fillna.hist() # 直方图
```

Out[13]: <matplotlib.axes._subplots.AxesSubplot at 0x211038296c8>



In [14]:

```
data_fillna.plot.box()
plt.show() # 盒图
```



In [ ]: