In [115]:
```
import pandas as pd
import string
import numpy as np
from pandas.api.types import CategoricalDtype

np.random.seed(1234)
pd.set_option('max_rows',10)
uniques = np.array(list(string.ascii_letters))
uniques
```

Out[115]:
```
array(['a', 'b', 'c', 'd', 'e', 'f', 'g', 'h', 'i', 'j', 'k', 'l', 'm',
       'n', 'o', 'p', 'q', 'r', 's', 't', 'u', 'v', 'w', 'x', 'y', 'z',
       'A', 'B', 'C', 'D', 'E', 'F', 'G', 'H', 'I', 'J', 'K', 'L', 'M',
       'N', 'O', 'P', 'Q', 'R', 'S', 'T', 'U', 'V', 'W', 'X', 'Y', 'Z'],
      dtype='<U1')
```

In [86]:
```
df1 = pd.DataFrame({'A' : uniques.take(np.random.randint(0,len(uniques)/2+5,si
ze=1000000))})
df1.head()
```

Out[86]:

	A
0	B
1	a
2	s
3	e
4	A

In [114]:
```
df1.A.unique()
```

Out[114]:
```
array(['B', 'a', 's', 'e', 'A', 'n', 'E', 'u', 'r', 'v', 'g', 'D', 'h',
       'j', 'o', 't', 'd', 'c', 'k', 'q', 'b', 'y', 'f', 'C', 'm', 'w',
       'p', 'l', 'i', 'z', 'x'], dtype=object)
```

In [88]:
```
df2 = pd.DataFrame({'A' : uniques.take(np.random.randint(0,len(uniques),size=1
000000))})
df2.head()
```

Out[88]:

	A
0	a
1	X
2	m
3	F
4	c

In [113]:
```
df2.A.unique()
```

Out[113]:
```
array(['a', 'X', 'm', 'F', 'c', 'w', 'R', 'y', 't', 'd', 'u', 'D', 'Y',
       'k', 'E', 'W', 'L', 'q', 'v', 'e', 'j', 'O', 'V', 'b', 'A', 'g',
       'T', 'o', 'G', 'l', 'p', 'f', 'N', 'Z', 'K', 'I', 'r', 'B', 'h',
       'C', 's', 'H', 'P', 'Q', 'i', 'U', 'x', 'z', 'M', 'n', 'S', 'J'], dtyp
e=object)
```

In [110]:
```
df1['B'] = df1.A.astype('category')
i = df1.B.cat.categories
i
```

Out[110]:
```
Index(['A', 'B', 'C', 'D', 'E', 'a', 'b', 'c', 'd', 'e', 'f', 'g', 'h', 'i',
       'j', 'k', 'l', 'm', 'n', 'o', 'p', 'q', 'r', 's', 't', 'u', 'v', 'w',
       'x', 'y', 'z'],
      dtype='object')
```

In [109]:
```
i2 = df2.A.astype('category').cat.categories
i2
```

Out[109]:
```
Index(['A', 'B', 'C', 'D', 'E', 'F', 'G', 'H', 'I', 'J', 'K', 'L', 'M', 'N',
       'O', 'P', 'Q', 'R', 'S', 'T', 'U', 'V', 'W', 'X', 'Y', 'Z', 'a', 'b',
       'c', 'd', 'e', 'f', 'g', 'h', 'i', 'j', 'k', 'l', 'm', 'n', 'o', 'p',
       'q', 'r', 's', 't', 'u', 'v', 'w', 'x', 'y', 'z'],
      dtype='object')
```

In [111]:
```
cats = i.tolist() + (i ^ i2).tolist()
print(cats)
```

```
['A', 'B', 'C', 'D', 'E', 'a', 'b', 'c', 'd', 'e', 'f', 'g', 'h', 'i', 'j',
'k', 'l', 'm', 'n', 'o', 'p', 'q', 'r', 's', 't', 'u', 'v', 'w', 'x', 'y',
'z', 'F', 'G', 'H', 'I', 'J', 'K', 'L', 'M', 'N', 'O', 'P', 'Q', 'R', 'S',
'T', 'U', 'V', 'W', 'X', 'Y', 'Z']
```

In [95]:
```
(np.array(sorted(cats)) == sorted(uniques)).all()
```

Out[95]:
```
True
```

In [101]:
```
cat_type = CategoricalDtype(categories=cats)
df2['B'] = df2['A'].astype(cat_type)
```

Out[101]:

	A	B
0	a	a
1	X	X
2	m	m
3	F	F
4	c	c

```
In [125]: df1[df1.B.isin(['A','a','z','Z'])].B.cat.codes.unique()
```

```
Out[125]: array([ 5,  0, 30], dtype=int64)
```

```
In [105]: df2[df2.B.isin(['A','a','z','Z'])].B.cat.codes.unique()
```

```
Out[105]: array([ 5,  0, 51, 30], dtype=int64)
```

```
In [106]: df2.dtypes
```

```
Out[106]: A      object
          B    category
          dtype: object
```

```
In [107]: df2.A.to_frame().memory_usage()
```

```
Out[107]: Index         80
          A        8000000
          dtype: int64
```

```
In [108]: df2.B.to_frame().memory_usage()
```

```
Out[108]: Index         80
          B        1002976
          dtype: int64
```