In [1]:

```
import pandas as pd
import numpy as np
from datetime import datetime
```

In [3]:

```
FOI_DEV_LIST = []
for i in range(1998, 2009):
    FOI_DEV_LIST.append(str(i))
FOI_DEV_LIST_2 = ['', 'Add', 'Change']
for i in range(2009, 2017):
    FOI_DEV_LIST_2.append(str(i))

baseline_col = ['BASELINE_BRAND_NAME', 'BASELINE_GENERIC_NAME', 'BASELINE_MODEL_
NO', 'BASELINE_CATALOG_NO',
                                'BASELINE_OTHER_ID_NO', 'BASELINE_DEVICE_FAMILY'
, 'BASELINE_SHELF_LIFE_CONTAINED',
                                'BASELINE_SHELF_LIFE_IN_MONTHS', 'BASELINE_PMA_F
LAG', 'BASELINE_PMA_NO', 'BASELINE_510_K__FLAG',
                                'BASELINE_510_K__NO', 'BASELINE_PREAMENDMENT', '
BASELINE_TRANSITIONAL', 'BASELINE_510_K__EXEMPT_FLAG',
                                'BASELINE_DATE_FIRST_MARKETED', 'BASELINE_DATE_C
EASED_MARKETING']
main_col = ['MDR_REPORT_KEY', 'DEVICE_EVENT_KEY', 'IMPLANT_FLAG', 'DATE_REMOVED_
FLAG', 'DEVICE_SEQUENCE_NO',
            'DATE_RECEIVED', 'BRAND_NAME', 'GENERIC_NAME', 'MANUFACTURER_D_NAME'
, 'MANUFACTURER_D_ADDRESS_1',
            'MANUFACTURER_D_ADDRESS_2', 'MANUFACTURER_D_CITY', 'MANUFACTURER_D_S
TATE_CODE', 'MANUFACTURER_D_ZIP_CODE',
            'MANUFACTURER_D_ZIP_CODE_EXT', 'MANUFACTURER_D_COUNTRY_CODE', 'MANUF
ACTURER_D_POSTAL_CODE',
            'EXPIRATION_DATE_OF_DEVICE', 'MODEL_NUMBER', 'CATALOG_NUMBER', 'LOT_
NUMBER', 'OTHER_ID_NUMBER',
            'DEVICE_OPERATOR', 'DEVICE_AVAILABILITY', 'DATE_RETURNED_TO_MANUFACT
URER', 'DEVICE_REPORT_PRODUCT_CODE',
            'DEVICE_AGE_TEXT', 'DEVICE_EVALUATED_BY_MANUFACTUR']

df_BI = pd.read_csv('1_BI_KEY_list.txt', header=None, names=['MDR_REPORT_KEY'])
```

In [4]:

```
df_list = []
for s in FOI_DEV_LIST:
    df = pd.read_csv('foidev/foidev'+s+'.txt', sep='|', header=0,
                                    encoding='ISO-8859-1', error_bad_lines=False)
    df_list.append(df)
df_BASELINE = pd.concat(df_list, axis=0)
del df_list

df_BI_BASELINE = df_BASELINE.merge(df_BI, on=['MDR_REPORT_KEY'], how='inner')
```

```
/Library/Frameworks/Python.framework/Versions/3.6/lib/python3.6/site
-packages/IPython/core/interactiveshell.py:2698: DtypeWarning: Colum
ns (14) have mixed types. Specify dtype option on import or set low_
memory=False.
  interactivity=interactivity, compiler=compiler, result=result)
b'Skipping line 16452: expected 45 fields, saw 46\n'
b'Skipping line 48741: expected 45 fields, saw 57\n'
/Library/Frameworks/Python.framework/Versions/3.6/lib/python3.6/site
-packages/IPython/core/interactiveshell.py:2698: DtypeWarning: Colum
ns (13,14) have mixed types. Specify dtype option on import or set l
ow_memory=False.
  interactivity=interactivity, compiler=compiler, result=result)
b'Skipping line 23599: expected 45 fields, saw 48\n'
/Library/Frameworks/Python.framework/Versions/3.6/lib/python3.6/site
-packages/IPython/core/interactiveshell.py:2698: DtypeWarning: Colum
ns (14,31,35,44) have mixed types. Specify dtype option on import or
set low_memory=False.
  interactivity=interactivity, compiler=compiler, result=result)
/Library/Frameworks/Python.framework/Versions/3.6/lib/python3.6/site
-packages/IPython/core/interactiveshell.py:2698: DtypeWarning: Colum
ns (14,31,35) have mixed types. Specify dtype option on import or se
t low_memory=False.
  interactivity=interactivity, compiler=compiler, result=result)
```

In [5]:

```
line_num = df_BI_BASELINE.shape[0]
for b in baseline_col:
    print(b, 'missing :', df_BI_BASELINE[b].isnull().sum(), 'out of', line_num)
```

```
BASELINE_BRAND_NAME missing : 8250 out of 15122
BASELINE_GENERIC_NAME missing : 8251 out of 15122
BASELINE_MODEL_NO missing : 9053 out of 15122
BASELINE_CATALOG_NO missing : 9348 out of 15122
BASELINE_OTHER_ID_NO missing : 10529 out of 15122
BASELINE_DEVICE_FAMILY missing : 12690 out of 15122
BASELINE_SHELF_LIFE_CONTAINED missing : 13128 out of 15122
BASELINE_SHELF_LIFE_IN_MONTHS missing : 13166 out of 15122
BASELINE_PMA_FLAG missing : 11942 out of 15122
BASELINE_PMA_NO missing : 13676 out of 15122
BASELINE_510_K__FLAG missing : 11942 out of 15122
BASELINE_510_K__NO missing : 13552 out of 15122
BASELINE_PREAMENDMENT missing : 11942 out of 15122
BASELINE_TRANSITIONAL missing : 11942 out of 15122
BASELINE_510_K__EXEMPT_FLAG missing : 11942 out of 15122
BASELINE_DATE_FIRST_MARKETED missing : 11943 out of 15122
BASELINE_DATE_CEASED_MARKETING missing : 13743 out of 15122
```

In [6]:

```
for b in baseline_col:
    del df_BI_BASELINE[b]
df_list = [df_BI_BASELINE]
for s in FOI_DEV_LIST_2:
    df = pd.read_csv('foidev/foidev'+s+'.txt', sep='|', header=0,
                                    encoding='ISO-8859-1', error_bad_lines=False)
    df_list.append(df)
df = pd.concat(df_list, axis=0)
del df_list
df = df.merge(df_BI, on=['MDR_REPORT_KEY'], how='inner')
df.head()
```

```
b'Skipping line 46607: expected 28 fields, saw 29\n'
b'Skipping line 54500: expected 28 fields, saw 29\n'
b'Skipping line 92909: expected 28 fields, saw 29\n'
b'Skipping line 137127: expected 28 fields, saw 29\n'
b'Skipping line 226678: expected 28 fields, saw 29\n'
b'Skipping line 274432: expected 28 fields, saw 29\n'
b'Skipping line 404363: expected 28 fields, saw 29\n'
b'Skipping line 439704: expected 28 fields, saw 29\n'
b'Skipping line 131902: expected 28 fields, saw 29\n'
b'Skipping line 213504: expected 28 fields, saw 29\n'
b'Skipping line 229517: expected 28 fields, saw 29\nSkipping line 23
2754: expected 28 fields, saw 42\n'
b'Skipping line 337236: expected 28 fields, saw 42\n'
b'Skipping line 386036: expected 28 fields, saw 42\n'
b'Skipping line 517139: expected 28 fields, saw 29\n'
b'Skipping line 539420: expected 28 fields, saw 42\n'
b'Skipping line 580359: expected 28 fields, saw 29\n'
b'Skipping line 614689: expected 28 fields, saw 29\n'
b'Skipping line 667843: expected 28 fields, saw 29\n'
b'Skipping line 785600: expected 28 fields, saw 29\n'
b'Skipping line 39459: expected 28 fields, saw 40\n'
b'Skipping line 399152: expected 28 fields, saw 29\n'
b'Skipping line 534956: expected 28 fields, saw 29\n'
b'Skipping line 644346: expected 28 fields, saw 29\n'
b'Skipping line 709324: expected 28 fields, saw 29\n'
b'Skipping line 839177: expected 28 fields, saw 29\n'
b'Skipping line 9830: expected 28 fields, saw 29\nSkipping line 1144
8: expected 28 fields, saw 29\n'
b'Skipping line 92754: expected 28 fields, saw 29\n'
b'Skipping line 204882: expected 28 fields, saw 29\nSkipping line 22
7966: expected 28 fields, saw 29\n'
b'Skipping line 230075: expected 28 fields, saw 29\n'
b'Skipping line 266443: expected 28 fields, saw 29\nSkipping line 29
1029: expected 28 fields, saw 29\n'
b'Skipping line 337669: expected 28 fields, saw 29\n'
b'Skipping line 443629: expected 28 fields, saw 29\n'
b'Skipping line 570703: expected 28 fields, saw 29\n'
b'Skipping line 608662: expected 28 fields, saw 29\nSkipping line 61
3710: expected 28 fields, saw 29\n'
b'Skipping line 695004: expected 28 fields, saw 29\n'
b'Skipping line 730318: expected 28 fields, saw 29\nSkipping line 73
4211: expected 28 fields, saw 29\n'
b'Skipping line 777626: expected 28 fields, saw 29\n'
b'Skipping line 788407: expected 28 fields, saw 29\n'
```

Out[6]:

	MDR_REPORT_KEY	DEVICE_EVENT_KEY	IMPLANT_FLAG	DATE_REMOVED_FLAG
0	203929	198075.0	Y	V
1	203787	197937.0	Y	V
2	203782	197932.0	Y	V
3	203774	197924.0	Y	V
4	203753	197903.0	Y	V

5 rows × 28 columns

In [7]:

```
line_num = df.shape[0]
for b in main_col:
    print(b, 'missing :', df[b].isnull().sum(), 'out of', line_num)
```

```
MDR_REPORT_KEY missing : 0 out of 18225
DEVICE_EVENT_KEY missing : 3103 out of 18225
IMPLANT_FLAG missing : 3103 out of 18225
DATE_REMOVED_FLAG missing : 4503 out of 18225
DEVICE_SEQUENCE_NO missing : 0 out of 18225
DATE_RECEIVED missing : 0 out of 18225
BRAND_NAME missing : 1550 out of 18225
GENERIC_NAME missing : 1844 out of 18225
MANUFACTURER_D_NAME missing : 1822 out of 18225
MANUFACTURER_D_ADDRESS_1 missing : 3846 out of 18225
MANUFACTURER_D_ADDRESS_2 missing : 14673 out of 18225
MANUFACTURER_D_CITY missing : 3107 out of 18225
MANUFACTURER_D_STATE_CODE missing : 10264 out of 18225
MANUFACTURER_D_ZIP_CODE missing : 10672 out of 18225
MANUFACTURER_D_ZIP_CODE_EXT missing : 17001 out of 18225
MANUFACTURER_D_COUNTRY_CODE missing : 3132 out of 18225
MANUFACTURER_D_POSTAL_CODE missing : 14922 out of 18225
EXPIRATION_DATE_OF_DEVICE missing : 17673 out of 18225
MODEL_NUMBER missing : 5266 out of 18225
CATALOG_NUMBER missing : 4402 out of 18225
LOT_NUMBER missing : 3249 out of 18225
OTHER_ID_NUMBER missing : 8128 out of 18225
DEVICE_OPERATOR missing : 3593 out of 18225
DEVICE_AVAILABILITY missing : 1898 out of 18225
DATE_RETURNED_TO_MANUFACTURER missing : 14777 out of 18225
DEVICE_REPORT_PRODUCT_CODE missing : 12 out of 18225
DEVICE_AGE_TEXT missing : 8086 out of 18225
DEVICE_EVALUATED_BY_MANUFACTUR missing : 9506 out of 18225
```

In [12]:

```
u = list(df.BRAND_NAME.unique())
print('Distinct brand name:', len(u), 'out of', df.shape[0])
u = list(df.MODEL_NUMBER.unique())
print('Distinct model #:', len(u), 'out of', df.shape[0])
```

```
Distinct brand name: 1998 out of 18225
Distinct model #: 1269 out of 18225
```

In [9]:

```
df['DATE_RECEIVED'] = pd.to_datetime(df['DATE_RECEIVED'])
df.head()
```

Out[9]:

	MDR_REPORT_KEY	DEVICE_EVENT_KEY	IMPLANT_FLAG	DATE_REMOVED_FLAG
0	203929	198075.0	Y	V
1	203787	197937.0	Y	V
2	203782	197932.0	Y	V
3	203774	197924.0	Y	V
4	203753	197903.0	Y	V

5 rows × 28 columns

In [10]:

```
def slice(d):
    if (d < datetime(year=1994, month=1, day=1)) | (d >= datetime(year=2017, mon
th=2, day=1)):
        return True
    else:
        return False
```

In [11]:

```
df['filter'] = df['DATE_RECEIVED'].map(slice)
wrong_list = list(df.loc[df['filter']==True, :].MDR_REPORT_KEY.unique())
len(wrong_list)
```

Out[11]:

```
0
```

In [ ]:

```

```