In [39]:

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

In [40]:

```
FOI_DEV_LIST = ['thru1997']
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', 'B
ASELINE_SHELF_LIFE_CONTAINED', 'BASELINE_SHELF_LIFE_IN_MONTHS', 'BASELINE_PMA_FL
AG', '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_CEASED_MARKETING']
main_col = ['MDR_REPORT_KEY', 'DEVICE_EVENT_KEY', 'IMPLANT_FLAG', 'DATE_REMOVED_
FLAG', 'DEVICE_SEQUENCE_NO', 'DATE_RECEIVED', 'BRAND_NAME', 'GENERIC_NAME', 'MAN
UFACTURER_D_NAME', 'MANUFACTURER_D_ADDRESS_1', 'MANUFACTURER_D_ADDRESS_2', 'MANU
FACTURER_D_CITY', 'MANUFACTURER_D_STATE_CODE', 'MANUFACTURER_D_ZIP_CODE', 'MANUF
ACTURER_D_ZIP_CODE_EXT', 'MANUFACTURER_D_COUNTRY_CODE', 'MANUFACTURER_D_POSTAL_C
ODE', 'EXPIRATION_DATE_OF_DEVICE', 'MODEL_NUMBER', 'CATALOG_NUMBER', 'LOT_NUMBER
', 'OTHER_ID_NUMBER', 'DEVICE_OPERATOR', 'DEVICE_AVAILABILITY', 'DATE_RETURNED_T
O_MANUFACTURER', 'DEVICE_REPORT_PRODUCT_CODE', 'DEVICE_AGE_TEXT', 'DEVICE_EVALUA
TED_BY_MANUFACTUR']

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

In [41]:

```
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')
```

```
b'Skipping line 54015: expected 45 fields, saw 47\n'
b'Skipping line 66558: expected 45 fields, saw 58\n'
b'Skipping line 121357: expected 45 fields, saw 59\nSkipping line 12
2019: expected 45 fields, saw 59\nSkipping line 129021: expected 45
fields, saw 58\n'
/Library/Frameworks/Python.framework/Versions/3.6/lib/python3.6/site
-packages/IPython/core/interactiveshell.py:2717: DtypeWarning: Colum
ns (10,12,13,14,16,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,4
3,44) have mixed types. Specify dtype option on import or set low_me
mory=False.
  interactivity=interactivity, compiler=compiler, result=result)
/Library/Frameworks/Python.framework/Versions/3.6/lib/python3.6/site
-packages/IPython/core/interactiveshell.py:2717: 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:2717: 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:2717: 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:2717: 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 [42]:

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

```
BASELINE_BRAND_NAME missing : 16278 out of 24491 , 0.664652321261
BASELINE_GENERIC_NAME missing : 16280 out of 24491 , 0.664733983912
BASELINE_MODEL_NO missing : 17233 out of 24491 , 0.703646237393
BASELINE_CATALOG_NO missing : 17556 out of 24491 , 0.716834755625
BASELINE_OTHER_ID_NO missing : 19687 out of 24491 , 0.80384631089
BASELINE_DEVICE_FAMILY missing : 20960 out of 24491 , 0.855824588624
BASELINE_SHELF_LIFE_CONTAINED missing : 22059 out of 24491 , 0.90069
8215671
BASELINE_SHELF_LIFE_IN_MONTHS missing : 22183 out of 24491 , 0.90576
1300069
BASELINE_PMA_FLAG missing : 20208 out of 24491 , 0.825119431628
BASELINE_PMA_NO missing : 22755 out of 24491 , 0.929116818423
BASELINE_510_K__FLAG missing : 20208 out of 24491 , 0.825119431628
BASELINE_510_K__NO missing : 22649 out of 24491 , 0.924788697889
BASELINE_PREAMENDMENT missing : 20208 out of 24491 , 0.825119431628
BASELINE_TRANSITIONAL missing : 20208 out of 24491 , 0.825119431628
BASELINE_510_K__EXEMPT_FLAG missing : 20208 out of 24491 , 0.8251194
31628
BASELINE_DATE_FIRST_MARKETED missing : 20209 out of 24491 , 0.825160
262954
BASELINE_DATE_CEASED_MARKETING missing : 22246 out of 24491 , 0.9083
33673594
```

In [43]:

```
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')
```

```
b'Skipping line 46607: expected 28 fields, saw 29\n'
/Library/Frameworks/Python.framework/Versions/3.6/lib/python3.6/site
-packages/IPython/core/interactiveshell.py:2717: DtypeWarning: Colum
ns (0,1,2) have mixed types. Specify dtype option on import or set l
ow_memory=False.
  interactivity=interactivity, compiler=compiler, result=result)
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'
/Library/Frameworks/Python.framework/Versions/3.6/lib/python3.6/site
-packages/IPython/core/interactiveshell.py:2717: DtypeWarning: Colum
ns (0,1,2,3,4) have mixed types. Specify dtype option on import or s
et low_memory=False.
  interactivity=interactivity, compiler=compiler, result=result)
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 785601: 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'
```

In [44]:

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

```
MDR_REPORT_KEY missing : 0 out of 27512 , 0.0
DEVICE_EVENT_KEY missing : 3021 out of 27512 , 0.109806629834
IMPLANT_FLAG missing : 3021 out of 27512 , 0.109806629834
DATE_REMOVED_FLAG missing : 5002 out of 27512 , 0.181811573132
DEVICE_SEQUENCE_NO missing : 0 out of 27512 , 0.0
DATE_RECEIVED missing : 0 out of 27512 , 0.0
BRAND_NAME missing : 1669 out of 27512 , 0.0606644373364
GENERIC_NAME missing : 2715 out of 27512 , 0.0986842105263
MANUFACTURER_D_NAME missing : 1871 out of 27512 , 0.0680066879907
MANUFACTURER_D_ADDRESS_1 missing : 7272 out of 27512 , 0.26432102355
3
MANUFACTURER_D_ADDRESS_2 missing : 23799 out of 27512 , 0.8650407095
09
MANUFACTURER_D_CITY missing : 4604 out of 27512 , 0.167345158476
MANUFACTURER_D_STATE_CODE missing : 12741 out of 27512 , 0.463107007
851
MANUFACTURER_D_ZIP_CODE missing : 13131 out of 27512 , 0.47728264030
2
MANUFACTURER_D_ZIP_CODE_EXT missing : 25954 out of 27512 , 0.9433701
65746
MANUFACTURER_D_COUNTRY_CODE missing : 3693 out of 27512 , 0.13423233
4981
MANUFACTURER_D_POSTAL_CODE missing : 24215 out of 27512 , 0.88016138
4123
EXPIRATION_DATE_OF_DEVICE missing : 26954 out of 27512 , 0.979717941
262
MODEL_NUMBER missing : 9297 out of 27512 , 0.337925268974
CATALOG_NUMBER missing : 7060 out of 27512 , 0.256615295144
LOT_NUMBER missing : 5494 out of 27512 , 0.199694678686
OTHER_ID_NUMBER missing : 15122 out of 27512 , 0.549651061355
DEVICE_OPERATOR missing : 4802 out of 27512 , 0.174542018028
DEVICE_AVAILABILITY missing : 1943 out of 27512 , 0.0706237278279
DATE_RETURNED_TO_MANUFACTURER missing : 23778 out of 27512 , 0.86427
7406223
DEVICE_REPORT_PRODUCT_CODE missing : 52 out of 27512 , 0.00189008432
684
DEVICE_AGE_TEXT missing : 11688 out of 27512 , 0.424832800233
DEVICE_EVALUATED_BY_MANUFACTUR missing : 14717 out of 27512 , 0.5349
30212271
```

In [45]:

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

```
Distinct brand name: 3280 out of 25843
Distinct model #: 1795 out of 18215
```

In [46]:

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

In [47]:

```
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 [48]:

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

Out[48]:

```
0
```

In [ ]:

```

```