```
In [ ]:
#Import libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import glob
import re
from collections import defaultdict
```

```
In [ ]:
#To display all columns in Jupyter Notebooks
pd.set_option('display.max_columns', 500)
```

```
In [ ]:
#Import MongoClient
from pymongo import MongoClient

#Create a MongoClient to run the MongoDB instance
client = MongoClient('localhost', 27017)
```

```
In [ ]:
#Connect to existing datbase
db = client.NHANES
```

```
In [ ]:
db
```

```
Out[ ]: Database(MongoClient(host=['localhost:27017'], document_class=dict, tz_aware=Fal
se, connect=True), 'NHANES')
```

```
In [ ]:
col = db.list_collection_names()
col.sort()
col
```

```
Out[ ]: ['alq',
 'bmx',
 'bpq',
 'bpx',
 'demo',
 'demo_p',
 'descr',
 'diq',
 'drxtot',
 'hiq',
 'huq',
 'mcq_a',
 'mcq_b',
 'mcq_c',
 'mcq_h',
 'paq',
 'rdq',
 'smq',
 'smqfam',
 'tchol',
 'whq']
```

```
In [ ]:
```

```
#Collections
demo = db.demo
alq = db.alq
diq = db.diq
drxtot = db.drxtot
bpq = db.bpq
bpx = db.bpx
tchol = db.tchol
bmx = db.bmx
paq = db.paq
smq = db.smq
smqfam = db.smqfam

mcq_a = db.mcq_a #Asthma
mcq_h = db.mcq_h #Heart Disease
mcq_c = db.mcq_c #Cancer
mcq_b = db.mcq_b #Bronchitis (Chronic Lung)

hiq = db.hiq
huq = db.huq
whq = db.whq
rd = db.rdq

descr = db.descr
```

In [ ]:

```
#Create dataframes from database
df_demo = pd.DataFrame(list(demo.find()))
df_alq = pd.DataFrame(list(alq.find()))
df_diq = pd.DataFrame(list(diq.find()))
df_drxtot = pd.DataFrame(list(drxtot.find()))
df_bpq = pd.DataFrame(list(bpq.find()))
df_bpx = pd.DataFrame(list(bpx.find()))
df_tchol = pd.DataFrame(list(tchol.find()))
df_bmx = pd.DataFrame(list(bmx.find()))
df_paq = pd.DataFrame(list(paq.find()))
df_smq = pd.DataFrame(list(smq.find()))
df_smqfam = pd.DataFrame(list(smqfam.find()))

df_mcq_a = pd.DataFrame(list(mcq_a.find()))
df_mcq_h = pd.DataFrame(list(mcq_h.find()))
df_mcq_c = pd.DataFrame(list(mcq_c.find()))
df_mcq_b = pd.DataFrame(list(mcq_b.find()))

df_hiq = pd.DataFrame(list(hiq.find()))
df_huq = pd.DataFrame(list(huq.find()))
df_whq = pd.DataFrame(list(whq.find()))
df_rdq = pd.DataFrame(list(rd.find()))

df_descr = pd.DataFrame(list(descr.find()))
```

In [ ]:

```
#All records
dfs = [df_demo, df_alq, df_diq, df_drxtot, df_bpq, df_bpx, df_tchol, df_bmx, df_
       df_smq, df_smqfam, df_mcq_a, df_mcq_b, df_mcq_c, df_mcq_h, df_hiq, df_huq
```

In [ ]:

```
names = ['demo', 'alq', 'diq', 'drxtot', 'bpq', 'bpx', 'tchol', 'bmx', 'paq',
         'smq', 'smqfam', 'mcq_a', 'mcq_b', 'mcq_c', 'mcq_h', 'hiq', 'huq', 'whq'
```

In [ ]:

```
data_dict = dict(zip(names,dfs))
```

Functions:

In [ ]:

```
#Declare label globally
label = 'DIQ010'
```

In [ ]:

```
#Function for inner join
def innerjoin_df(dfs_list, join_on):
    df_join = dfs_list[0]
    for d in dfs_list[1:]:
        df_join = df_join.merge(d, how='inner', on=join_on)
    return df_join
```

In [ ]:

```
#Function for getting info from list of collections
#Look at records and features for each
def get_info(dfs, names):
    shape = [x.shape for x in dfs]
    d = defaultdict(str)
    for i in range(0,len(shape)):
        d[names[i]] = shape[i]
    info = pd.DataFrame.from_dict(d, orient='index').reset_index()
    info.columns = ['_id', 'Records', 'Features']
    return info
```

In [ ]:

```
info = get_info(dfs, names)

print(info)
print(df_descr)

info_join = innerjoin_df([info, df_descr], ['_id'])
info_join = info_join.sort_values(by='Records', ascending=False)
info_join
```

```
       _id  Records  Features
0     demo    89367        11
1      alq    48716         3
2      diq    96745         3
3   drxtot    86464        26
4      bpq    63219         3
5      bpx    68575         5
6    tchol    29419         6
7      bmx    84328         6
8      paq    70679         5
9      smq    63164         4
10  smqfam    99616         3
11   mcq_a    96696         3
12   mcq_b    54967         3
13   mcq_c    55021         3
14   mcq_h    54814         3
15     hiq   100628         3
```

```
16     huq    96565         7
17     whq    57930         6
18     rdq    68481         4
       _id                Description
0      alq                Alcohol Use
1      bmx              Body Measures
2      bpq             Blood Pressure
3      bpx  Blood Pressure - Measures
4     demo               Demographics
5   demo_p       Demographics for Vis
6      diq                   Diabetes
7   drxtot                    Dietary
8      hiq           Health Insurance
9      huq       Hospital Utilization
10   mcq_a                     Asthma
11   mcq_h              Heart Disease
12     paq          Physical Activity
13     smq                    Smoking
14  smqfam          Household Smoking
15   tchol                Cholesterol
16     whq             Weight History
17   mcq_c                     Cancer
18   mcq_b                 Bronchitis
19     rdq                      Cough
```

Out[ ]:

	_id	Records	Features	Description
15	hiq	100628	3	Health Insurance
10	smqfam	99616	3	Household Smoking
2	diq	96745	3	Diabetes
11	mcq_a	96696	3	Asthma
16	huq	96565	7	Hospital Utilization
0	demo	89367	11	Demographics
3	drxtot	86464	26	Dietary
7	bmx	84328	6	Body Measures
8	paq	70679	5	Physical Activity
5	bpx	68575	5	Blood Pressure - Measures
18	rdq	68481	4	Cough
4	bpq	63219	3	Blood Pressure
9	smq	63164	4	Smoking
17	whq	57930	6	Weight History
13	mcq_c	55021	3	Cancer
12	mcq_b	54967	3	Bronchitis
14	mcq_h	54814	3	Heart Disease
1	alq	48716	3	Alcohol Use
6	tchol	29419	6	Cholesterol

Select data to use

In [ ]:

```
#Get relevant data
def get_reldata(df):
    dfs = []
    for c in df:
        dfs.append(data_dict[c])
    return dfs
```

In [ ]:

```
#Selected risk factors for disease
names = ['demo', 'alq', 'diq', 'drxtot', 'bpq', 'bpx', 'bmx', 'tchol', 'paq',
         'smq', 'smqfam', 'hiq', 'huq']
```

In [ ]:

```
#Selected risk factors for disease
dfs = get_reldata(names)
```

Join dataframes

In [ ]:

```
df_j = innerjoin_df(dfs, ['_id','Year'])
df_j.shape
```

Out[ ]:

```
(14121, 61)
```

In [ ]:

```
df_j.head()
```

Out[ ]:

	_id	RIAGENDR	RIDAGEYR	RIDRETH1	DMDBORN4	DMDCITZN	DMDHHSIZ	INDFMINC	DMD
0	2.0	1.0	77.0	3.0	1.0	1.0	1.0	8.0	
1	5.0	1.0	49.0	3.0	1.0	1.0	3.0	11.0	
2	12.0	1.0	37.0	3.0	1.0	1.0	4.0	11.0	
3	15.0	2.0	38.0	3.0	1.0	1.0	2.0	8.0	
4	20.0	2.0	23.0	1.0	1.0	1.0	2.0	6.0	

Reorder columns

In [ ]:

```
#Get a list of columns
cols = list(df_j)
```

In [ ]:

```
#Move '_id' column to head of list using dex, pop and insert
cols.insert(0, cols.pop(cols.index('_id')))
```

```
#Move 'Year' column to back of list using index, pop and insert
cols.insert(len(df_j.columns)-1, cols.pop(cols.index('Year')))
```

In [ ]:

```
#Reorder dataframe
df_j = df_j.loc[:, cols]
df_j.head()
```

Out[ ]:

```
    _id  RIAGENDR  RIDAGEYR  RIDRETH1  DMDBORN4  DMDCITZN  DMDHHSIZ  INDFMINC  DMD

0   2.0       1.0      77.0       3.0       1.0       1.0       1.0       8.0

1   5.0       1.0      49.0       3.0       1.0       1.0       3.0      11.0

2  12.0       1.0      37.0       3.0       1.0       1.0       4.0      11.0

3  15.0       2.0      38.0       3.0       1.0       1.0       2.0       8.0

4  20.0       2.0      23.0       1.0       1.0       1.0       2.0       6.0
```

Remap years to number categories

In [ ]:

```
di = {"1999-2000": 0, "2001-2002": 1, "2003-2004": 2, "2005-2006": 3, "2007-2008
      "2009-2010": 5, "2011-2012": 6, "2013-2014": 7, "2015-2016": 8, "2017-2018
```

In [ ]:

```
#Map categorical years to numerical
df_j['Year'] = df_j['Year'].map(di)
```

In [ ]:

```
df_j.head()
```

Out[ ]:

```
    _id  RIAGENDR  RIDAGEYR  RIDRETH1  DMDBORN4  DMDCITZN  DMDHHSIZ  INDFMINC  DMD
0   2.0       1.0      77.0       3.0       1.0       1.0       1.0       8.0
1   5.0       1.0      49.0       3.0       1.0       1.0       3.0      11.0
2  12.0       1.0      37.0       3.0       1.0       1.0       4.0      11.0
3  15.0       2.0      38.0       3.0       1.0       1.0       2.0       8.0
4  20.0       2.0      23.0       1.0       1.0       1.0       2.0       6.0
```

In [ ]:

```
#Check if any NaN
df_j.isnull().values.any()

def count_vals(df, name):
    df_count = df.groupby(name)['_id'].nunique()
    print(df_count,"\n\n","NaN: ", df[name].isnull().sum())

cols = df_j.columns
```

```
for i in range(0,len(cols)):
    count_vals(df_j, cols[i])
```

```
_id
2.0        1
5.0        1
12.0       1
15.0       1
20.0       1
          ..
97436.0    1
97437.0    1
97443.0    1
97445.0    1
97446.0    1
Name: _id, Length: 14121, dtype: int64

 NaN:  0
RIAGENDR
1.0    6997
2.0    7124
Name: _id, dtype: int64

 NaN:  0
RIDAGEYR
18.0    152
19.0    119
20.0    233
21.0    245
22.0    247
       ...
81.0     63
82.0     41
83.0     34
84.0     40
85.0    127
Name: _id, Length: 68, dtype: int64

 NaN:  0
RIDRETH1
1.0    2536
2.0    1015
3.0    6773
4.0    2707
5.0    1090
Name: _id, dtype: int64

 NaN:  0
DMDBORN4
1.0    10776
2.0     3345
Name: _id, dtype: int64

 NaN:  0
DMDCITZN
1.0    12355
2.0     1766
Name: _id, dtype: int64
```

```
 NaN:  0
DMDHHSIZ
1.0    1967
2.0    4566
3.0    2590
4.0    2184
5.0    1457
6.0     666
7.0     691
Name: _id, dtype: int64

 NaN:  0
INDFMINC
1.0      501
2.0      724
3.0     1208
4.0     1131
5.0     1215
6.0     1817
7.0     1475
8.0     1216
9.0      912
10.0     756
11.0    3166
Name: _id, dtype: int64

 NaN:  0
DMDHREDU
1.0    1672
2.0    2629
3.0    3323
4.0    3616
5.0    2881
Name: _id, dtype: int64

 NaN:  0
MEC18YR
154.366307      1
158.231452      1
158.916499      2
161.133259      1
163.166688      1
               ..
35819.675521    1
37305.654088    1
41449.900160    1
42193.827522    1
43857.889978    1
Name: _id, Length: 12179, dtype: int64

 NaN:  0
ALQ101
1.0    10348
2.0     3773
Name: _id, dtype: int64

 NaN:  0
DIQ010
1.0     1521
2.0    12347
```

```
3.0      253
Name: _id, dtype: int64

 NaN:  0
DRD320GW
5.397605e-79    2403
2.500000e+00       1
4.930000e+00       1
7.500000e+00       1
1.475000e+01       1
                ...
1.115920e+04       1
1.218000e+04       1
1.239000e+04       1
1.288688e+04       1
1.536000e+04       1
Name: _id, Length: 1990, dtype: int64

 NaN:  0
DRDTSODI
5.397605e-79    1
7.000000e+00    1
2.200000e+01    1
2.400000e+01    1
2.900000e+01    1
               ..
1.768700e+04    1
2.016500e+04    1
2.018300e+04    1
2.079900e+04    1
2.139900e+04    1
Name: _id, Length: 6719, dtype: int64

 NaN:  0
DRX18YR
79.974564       1
84.826593       1
86.378852       1
88.629073       1
89.992322       1
               ..
49507.686973    1
50079.522255    1
52149.366413    1
70249.163257    1
72866.337840    1
Name: _id, Length: 13591, dtype: int64

 NaN:  0
DRXTALCO
5.397605e-79    10649
1.000000e-02        7
2.000000e-02       10
3.000000e-02       11
4.000000e-02        3
                ...
3.924700e+02        1
4.032000e+02        2
4.212000e+02        1
5.242000e+02        1
```

```
5.515000e+02        1
Name: _id, Length: 981, dtype: int64

 NaN:  0
DRXTCAFF
5.397605e-79    1972
3.100000e-01       1
4.700000e-01       1
5.200000e-01       1
5.400000e-01       1
                ... 
3.066800e+03       1
3.216000e+03       1
3.720000e+03       1
4.159920e+03       1
4.295000e+03       1
Name: _id, Length: 1708, dtype: int64

 NaN:  0
DRXTCALC
5.397605e-79    1
1.100000e+01    1
2.000000e+01    1
2.064000e+01    1
2.244000e+01    1
               ..
6.408000e+03    1
7.337000e+03    1
7.409480e+03    1
8.470000e+03    1
9.733210e+03    1
Name: _id, Length: 3608, dtype: int64

 NaN:  0
DRXTCARB
5.397605e-79    2
3.800000e+00    1
5.880000e+00    1
7.460000e+00    1
1.158000e+01    1
               ..
1.305560e+03    1
1.423870e+03    1
1.459210e+03    1
1.644370e+03    1
1.815020e+03    1
Name: _id, Length: 12519, dtype: int64

 NaN:  0
DRXTCHOL
5.397605e-79    53
9.200000e-01     1
1.000000e+00     7
1.840000e+00     1
2.000000e+00     8
                ..
2.403000e+03     1
2.523000e+03     1
2.584000e+03     1
2.968000e+03     1
```

```
3.061000e+03     1
Name: _id, Length: 2370, dtype: int64

 NaN:  0
DRXTCOPP
5.397605e-79    1
2.500000e-02    1
2.800000e-02    1
4.000000e-02    1
5.200000e-02    1
               ..
2.743500e+01    1
2.770200e+01    1
2.860500e+01    1
3.758100e+01    1
3.927800e+01    1
Name: _id, Length: 5078, dtype: int64

 NaN:  0
DRXTFIBE
5.397605e-79    11
1.300000e-01     1
2.000000e-01     2
3.000000e-01     2
3.600000e-01     1
                ..
1.046000e+02     1
1.076000e+02     1
1.117900e+02     1
1.139000e+02     1
1.453500e+02     1
Name: _id, Length: 2159, dtype: int64

 NaN:  0
DRXTIRON
5.397605e-79    2
5.000000e-02    1
1.100000e-01    1
1.200000e-01    1
1.300000e-01    1
               ..
9.981000e+01    1
1.015300e+02    1
1.031700e+02    1
1.307600e+02    1
1.478800e+02    1
Name: _id, Length: 5608, dtype: int64

 NaN:  0
DRXTKCAL
5.397605e-79    2
5.400000e+01    1
9.300000e+01    1
1.130000e+02    1
1.170000e+02    1
               ..
1.171000e+04    1
1.210800e+04    1
1.282300e+04    1
1.339800e+04    1
```

```
1.368700e+04    1
Name: _id, Length: 4932, dtype: int64

 NaN:  0
DRXTMAGN
5.397605e-79    1
7.000000e+00    1
1.400000e+01    1
1.500000e+01    1
1.800000e+01    1
               ..
1.653000e+03    1
1.654000e+03    1
1.674000e+03    1
1.704000e+03    1
2.396510e+03    1
Name: _id, Length: 2056, dtype: int64

 NaN:  0
DRXTPHOS
5.397605e-79    2
2.400000e+01    1
2.500000e+01    1
3.000000e+01    1
3.100000e+01    1
               ..
7.373000e+03    1
7.398000e+03    1
7.971000e+03    1
8.760000e+03    1
1.152900e+04    1
Name: _id, Length: 4027, dtype: int64

 NaN:  0
DRXTPOTA
5.397605e-79    2
3.300000e+01    1
6.600000e+01    1
8.200000e+01    1
1.200000e+02    1
               ..
1.329600e+04    1
1.407200e+04    1
1.427500e+04    1
1.481200e+04    1
1.587600e+04    1
Name: _id, Length: 5661, dtype: int64

 NaN:  0
DRXTPROT
5.397605e-79    2
8.700000e-01    1
1.020000e+00    1
1.260000e+00    1
1.440000e+00    1
               ..
4.383500e+02    1
4.407300e+02    1
4.563800e+02    1
5.131000e+02    1
```

```
5.233100e+02    1
Name: _id, Length: 10666, dtype: int64

 NaN:  0
DRXTTFAT
5.397605e-79    5
4.400000e-01    1
4.600000e-01    1
7.200000e-01    1
8.000000e-01    1
               ..
5.073200e+02    1
5.361000e+02    1
5.483800e+02    1
5.537900e+02    1
6.013300e+02    1
Name: _id, Length: 10798, dtype: int64

 NaN:  0
DRXTVARE
5.397605e-79    23
5.700000e-01     1
1.000000e+00     4
1.500000e+00     1
2.000000e+00     8
                ..
1.517833e+04     1
1.640991e+04     1
1.947600e+04     1
2.101000e+04     1
3.706841e+04     1
Name: _id, Length: 3187, dtype: int64

 NaN:  0
DRXTVB1
5.397605e-79    2
2.500000e-02    1
2.900000e-02    1
3.000000e-02    1
3.300000e-02    1
               ..
1.267500e+01    1
1.284100e+01    1
1.308300e+01    1
1.363500e+01    1
2.311100e+01    1
Name: _id, Length: 5719, dtype: int64

 NaN:  0
DRXTVB12
5.397605e-79    32
1.000000e-02     8
2.000000e-02     1
3.000000e-02     5
4.000000e-02     5
                ..
1.655300e+02     1
2.052100e+02     1
2.211000e+02     1
2.932300e+02     1
```

```
3.810800e+02     1
Name: _id, Length: 3046, dtype: int64

 NaN:  0
DRXTVB2
5.397605e-79    2
2.800000e-02    1
3.100000e-02    1
3.800000e-02    1
5.000000e-02    1
               ..
1.358200e+01    1
1.654000e+01    1
1.906400e+01    1
2.632200e+01    1
2.652200e+01    1
Name: _id, Length: 6485, dtype: int64

 NaN:  0
DRXTVB6
5.397605e-79    2
4.000000e-03    1
1.000000e-02    1
1.400000e-02    1
1.600000e-02    1
               ..
1.735300e+01    1
1.762600e+01    1
1.929200e+01    1
2.106100e+01    1
2.627600e+01    1
Name: _id, Length: 6274, dtype: int64

 NaN:  0
DRXTVC
5.397605e-79    76
1.000000e-02     1
4.000000e-02     1
1.000000e-01    20
2.000000e-01    34
                ..
1.006000e+03     1
1.247700e+03     1
1.275100e+03     1
1.383300e+03     1
1.965900e+03     1
Name: _id, Length: 5366, dtype: int64

 NaN:  0
DRXTZINC
5.397605e-79    2
1.000000e-01    1
1.900000e-01    1
2.000000e-01    1
2.200000e-01    1
               ..
2.649800e+02    1
2.793600e+02    1
2.839600e+02    1
2.873000e+02    1
```

```
3.099200e+02    1
Name: _id, Length: 5120, dtype: int64

 NaN:  0
BPQ020
1.0    4731
2.0    9390
Name: _id, dtype: int64

 NaN:  0
BPXPULS
1.0    13675
2.0      446
Name: _id, dtype: int64

 NaN:  0
BPXSY1
72.0     1
76.0     1
78.0     2
80.0     1
82.0     4
        ..
230.0    2
232.0    1
236.0    1
238.0    1
256.0    1
Name: _id, Length: 80, dtype: int64

 NaN:  0
BPXDI1
5.397605e-79      96
1.000000e+01       3
1.800000e+01       1
2.000000e+01       3
2.200000e+01       5
2.400000e+01       2
2.600000e+01       5
2.800000e+01       7
3.000000e+01      10
3.200000e+01      14
3.400000e+01      11
3.600000e+01      17
3.800000e+01      13
4.000000e+01      30
4.200000e+01      56
4.400000e+01      48
4.600000e+01     101
4.800000e+01     126
5.000000e+01     194
5.200000e+01     279
5.400000e+01     328
5.600000e+01     437
5.800000e+01     468
6.000000e+01     566
6.200000e+01     691
6.400000e+01     834
6.600000e+01     874
6.800000e+01     936
```

```
7.000000e+01    1002
7.200000e+01     983
7.400000e+01     992
7.600000e+01     932
7.800000e+01     750
8.000000e+01     732
8.200000e+01     515
8.400000e+01     498
8.600000e+01     375
8.800000e+01     276
9.000000e+01     267
9.200000e+01     160
9.400000e+01     144
9.600000e+01     101
9.800000e+01      71
1.000000e+02      57
1.020000e+02      30
1.040000e+02      28
1.060000e+02      11
1.080000e+02      20
1.100000e+02       9
1.120000e+02       2
1.140000e+02       4
1.160000e+02       3
1.180000e+02       2
1.200000e+02       1
1.280000e+02       1
Name: _id, dtype: int64

 NaN:  0
BMXWT
32.8     1
35.9     1
36.1     2
36.2     1
36.3     1
        ..
193.7    1
198.9    1
199.4    1
216.1    1
218.2    1
Name: _id, Length: 1174, dtype: int64

 NaN:  0
BMXHT
129.7    1
133.9    1
135.7    1
136.9    1
137.1    1
        ..
200.1    1
201.0    1
201.2    1
202.7    1
203.2    1
Name: _id, Length: 570, dtype: int64

 NaN:  0
```

```
BMXBMI
13.36    1
14.10    1
14.70    1
14.80    1
15.10    1
        ..
67.34    1
68.60    1
68.70    1
70.10    1
76.07    1
Name: _id, Length: 2571, dtype: int64

 NaN:  0
BMXWAIST
55.5     1
56.4     1
58.6     1
58.7     1
59.7     1
        ..
163.5    1
163.6    1
165.2    1
166.0    1
170.3    1
Name: _id, Length: 881, dtype: int64

 NaN:  0
LBXTC
66.0     1
75.0     1
79.0     1
80.0     1
82.0     1
        ..
417.0    1
432.0    1
445.0    1
446.0    1
463.0    1
Name: _id, Length: 285, dtype: int64

 NaN:  0
LBDHDL
7.0      1
8.0      1
10.0     1
14.0     1
16.0     2
        ..
160.0    1
164.0    1
179.0    1
188.0    1
226.0    1
Name: _id, Length: 129, dtype: int64

 NaN:  0
```

```
LBXTR
10.0     1
12.0     1
13.0     2
14.0     2
15.0     1
        ..
395.0    2
396.0    2
398.0    2
399.0    1
400.0    1
Name: _id, Length: 386, dtype: int64

 NaN:  0
LBDLDL
9.0      1
13.0     1
14.0     1
15.0     2
18.0     1
        ..
320.0    1
341.0    1
344.0    2
354.0    1
375.0    1
Name: _id, Length: 259, dtype: int64

 NaN:  0
PAQ635
1.0     3464
2.0    10657
Name: _id, dtype: int64

 NaN:  0
PAQ650
1.0     3675
2.0    10446
Name: _id, dtype: int64

 NaN:  0
PAQ665
1.0    6350
2.0    7771
Name: _id, dtype: int64

 NaN:  0
SMQ680
1.0     3510
2.0    10611
Name: _id, dtype: int64

 NaN:  0
SMAQUEX
1.0       73
2.0    14048
Name: _id, dtype: int64

 NaN:  0
```

```
SMD410
0.0        902
1.0       2876
2.0      10319
3.0         21
777.0        1
999.0        2
Name: _id, dtype: int64

 NaN:  0
HID010
1.0    11247
2.0     2874
Name: _id, dtype: int64

 NaN:  0
HUQ010
1.0    2233
2.0    3883
3.0    4943
4.0    2527
5.0     535
Name: _id, dtype: int64

 NaN:  0
HUQ020
1.0    2642
2.0    1516
3.0    9963
Name: _id, dtype: int64

 NaN:  0
HUQ030
1.0    11805
2.0     2215
3.0      101
Name: _id, dtype: int64

 NaN:  0
HUQ050
0.000000e+00    2114
5.397605e-79     218
1.000000e+00    2457
2.000000e+00    3912
3.000000e+00    3402
4.000000e+00     929
5.000000e+00     946
6.000000e+00      69
7.000000e+00      27
8.000000e+00      47
Name: _id, dtype: int64

 NaN:  0
HUQ070
1.0     1563
2.0    12558
Name: _id, dtype: int64

 NaN:  0
Year
```

```
0    1257
1    1572
2    1420
3    1251
4    1535
5    1584
6    1346
7    1526
8    1304
9    1326
Name: _id, dtype: int64

 NaN:  0
```

Categorize features that need to be One Hot Encoded

In [ ]:

```
df_j.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 14121 entries, 0 to 14120
Data columns (total 61 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   _id       14121 non-null  float64
 1   RIAGENDR  14121 non-null  float64
 2   RIDAGEYR  14121 non-null  float64
 3   RIDRETH1  14121 non-null  float64
 4   DMDBORN4  14121 non-null  float64
 5   DMDCITZN  14121 non-null  float64
 6   DMDHHSIZ  14121 non-null  float64
 7   INDFMINC  14121 non-null  float64
 8   DMDHREDU  14121 non-null  float64
 9   MEC18YR   14121 non-null  float64
 10  ALQ101    14121 non-null  float64
 11  DIQ010    14121 non-null  float64
 12  DRD320GW  14121 non-null  float64
 13  DRDTSODI  14121 non-null  float64
 14  DRX18YR   14121 non-null  float64
 15  DRXTALCO  14121 non-null  float64
 16  DRXTCAFF  14121 non-null  float64
 17  DRXTCALC  14121 non-null  float64
 18  DRXTCARB  14121 non-null  float64
 19  DRXTCHOL  14121 non-null  float64
 20  DRXTCOPP  14121 non-null  float64
 21  DRXTFIBE  14121 non-null  float64
 22  DRXTIRON  14121 non-null  float64
 23  DRXTKCAL  14121 non-null  float64
 24  DRXTMAGN  14121 non-null  float64
 25  DRXTPHOS  14121 non-null  float64
 26  DRXTPOTA  14121 non-null  float64
 27  DRXTPROT  14121 non-null  float64
 28  DRXTTFAT  14121 non-null  float64
 29  DRXTVARE  14121 non-null  float64
 30  DRXTVB1   14121 non-null  float64
 31  DRXTVB12  14121 non-null  float64
 32  DRXTVB2   14121 non-null  float64
 33  DRXTVB6   14121 non-null  float64
 34  DRXTVC    14121 non-null  float64
```

```
  35  DRXTZINC  14121 non-null  float64
  36  BPQ020    14121 non-null  float64
  37  BPXPULS   14121 non-null  float64
  38  BPXSY1    14121 non-null  float64
  39  BPXDI1    14121 non-null  float64
  40  BMXWT     14121 non-null  float64
  41  BMXHT     14121 non-null  float64
  42  BMXBMI    14121 non-null  float64
  43  BMXWAIST  14121 non-null  float64
  44  LBXTC     14121 non-null  float64
  45  LBDHDL    14121 non-null  float64
  46  LBXTR     14121 non-null  float64
  47  LBDLDL    14121 non-null  float64
  48  PAQ635    14121 non-null  float64
  49  PAQ650    14121 non-null  float64
  50  PAQ665    14121 non-null  float64
  51  SMQ680    14121 non-null  float64
  52  SMAQUEX   14121 non-null  float64
  53  SMD410    14121 non-null  float64
  54  HID010    14121 non-null  float64
  55  HUQ010    14121 non-null  float64
  56  HUQ020    14121 non-null  float64
  57  HUQ030    14121 non-null  float64
  58  HUQ050    14121 non-null  float64
  59  HUQ070    14121 non-null  float64
  60  Year      14121 non-null  int64
dtypes: float64(60), int64(1)
memory usage: 6.7 MB
```

In [ ]:

```
#Change columns to category
#Columns to remove:
#DRX18YR - 18 Year weight
#MEC18YR - 18 year Weight
#Year
#_id

cat_cols = ['DMDBORN4',
            'DMDCITZN',
            'RIAGENDR',
            'RIDRETH1',
            'ALQ101',
            'DIQ010',
            'BPQ020',
            'BPXPULS',
            'PAQ635',
            'PAQ650',
            'PAQ665',
            'SMAQUEX',
            'SMQ680',
            'SMD410',
            'MCQ010',
            'MCQ160C',
            'MCQ160K',
            'MCQ220',
            'HID010',
            'HUQ020',
            'HUQ030',
            'HUQ070',
            'WHQ030',
             'WHQ040']

def recat_cols(df, col_names):
    for x in col_names:
        if x in cat_cols:
            df[x] = df[x].astype('category')
    return df

col_names = df_j.columns
df_ohe = recat_cols(df_j, col_names)
```

In [ ]:

```
df_ohe.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 14121 entries, 0 to 14120
Data columns (total 61 columns):
 #   Column    Non-Null Count  Dtype   
---  ------    --------------  -----   
 0   _id       14121 non-null  float64 
 1   RIAGENDR  14121 non-null  category
 2   RIDAGEYR  14121 non-null  float64 
 3   RIDRETH1  14121 non-null  category
 4   DMDBORN4  14121 non-null  category
 5   DMDCITZN  14121 non-null  category
 6   DMDHHSIZ  14121 non-null  float64 
 7   INDFMINC  14121 non-null  float64 
 8   DMDHREDU  14121 non-null  float64 
 9   MEC18YR   14121 non-null  float64 
 10  ALQ101    14121 non-null  category
 11  DIQ010    14121 non-null  category
 12  DRD320GW  14121 non-null  float64 
 13  DRDTSODI  14121 non-null  float64 
 14  DRX18YR   14121 non-null  float64 
 15  DRXTALCO  14121 non-null  float64 
 16  DRXTCAFF  14121 non-null  float64 
 17  DRXTCALC  14121 non-null  float64 
 18  DRXTCARB  14121 non-null  float64 
 19  DRXTCHOL  14121 non-null  float64 
 20  DRXTCOPP  14121 non-null  float64 
 21  DRXTFIBE  14121 non-null  float64 
 22  DRXTIRON  14121 non-null  float64 
 23  DRXTKCAL  14121 non-null  float64 
 24  DRXTMAGN  14121 non-null  float64 
 25  DRXTPHOS  14121 non-null  float64 
 26  DRXTPOTA  14121 non-null  float64 
 27  DRXTPROT  14121 non-null  float64 
 28  DRXTTFAT  14121 non-null  float64 
 29  DRXTVARE  14121 non-null  float64 
 30  DRXTVB1   14121 non-null  float64 
 31  DRXTVB12  14121 non-null  float64 
 32  DRXTVB2   14121 non-null  float64 
 33  DRXTVB6   14121 non-null  float64 
 34  DRXTVC    14121 non-null  float64 
 35  DRXTZINC  14121 non-null  float64 
 36  BPQ020    14121 non-null  category
 37  BPXPULS   14121 non-null  category
 38  BPXSY1    14121 non-null  float64 
 39  BPXDI1    14121 non-null  float64 
 40  BMXWT     14121 non-null  float64
 41  BMXHT     14121 non-null  float64
 42  BMXBMI    14121 non-null  float64
 43  BMXWAIST  14121 non-null  float64
 44  LBXTC     14121 non-null  float64
 45  LBDHDL    14121 non-null  float64
 46  LBXTR     14121 non-null  float64
 47  LBDLDL    14121 non-null  float64
 48  PAQ635    14121 non-null  category
 49  PAQ650    14121 non-null  category
 50  PAQ665    14121 non-null  category
 51  SMQ680    14121 non-null  category
 52  SMAQUEX   14121 non-null  category
 53  SMD410    14121 non-null  category
 54  HID010    14121 non-null  category
 55  HUQ010    14121 non-null  float64
 56  HUQ020    14121 non-null  category
 57  HUQ030    14121 non-null  category
 58  HUQ050    14121 non-null  float64
 59  HUQ070    14121 non-null  category
 60  Year      14121 non-null  int64
dtypes: category(18), float64(42), int64(1)
memory usage: 5.0 MB
```

One Hot Encoding Cateogires

In [ ]:

```
#DRX18YR - 18 Year weight
#MEC18YR - 18 year Weight
#Year
#_id
```

In [ ]:

```
#Function to One Hot Encode Categories
def ohe(df_j, label=None):
    #Make copy of df
    df_t = df_j.copy()
    #Select datatypes that are categories
    X_cat = df_t.select_dtypes(include=['category'])
    if(label != None):
        #Drop label and year
        X_cat = X_cat.drop([label], axis=1)
    #Copy df with categories that dropped label and year
    X_enc = X_cat.copy()
    #Create dummies from categories
    X_enc_d = pd.get_dummies(X_enc, drop_first=True)
    #Drop original non-OHE columns from original df
    df = df_j.drop(list(X_enc), axis=1)
    df = pd.concat([df,X_enc_d], axis=1)
    if(label != None):
        df[label] = df[label].astype(np.uint8)
    df['Year'] = df['Year'].astype(np.uint8)
    return df
```

In [ ]:

```
df_ohe = ohe(df_ohe, label)
df_no_ohe = df_j.copy()
```

In [ ]:

```
df_ohe[:1].shape
```

Out[ ]:

```
(1, 70)
```

In [ ]:

```
df_ohe.shape
```

Out[ ]:

```
(14121, 70)
```

Recategorize label DIQ010 to binary: 0 - No Diabetes; 1 - Diabetes & Borderline

In [ ]:

```
#Recategorize function
def recategorize(df, name, replace_dict):
    df[name].replace(
    to_replace=replace_dict,
    inplace=True
)
```

In [ ]:

```
#Recategorize to: 0 - No Diabetes; 1 - Diabetes & Borderline
recategorize(df_ohe, label, {2:0})
recategorize(df_ohe, label, {3:1})
recategorize(df_no_ohe, label, {2:0})
recategorize(df_no_ohe, label, {3:1})
```

In [ ]:

```
df_ohe.head()
```

Out[ ]:

	_id	RIDAGEYR	DMDHHSIZ	INDFMINC	DMDHREDU	MEC18YR	DIQ010	DRD320GW	D
0	2.0	77.0	1.0	8.0	5.0	3408.044382	0	5.397605e-79	
1	5.0	49.0	3.0	11.0	4.0	10219.103963	0	1.298000e+03	
2	12.0	37.0	4.0	11.0	2.0	10149.365568	0	3.304000e+03	
3	15.0	38.0	2.0	8.0	5.0	11437.714415	0	2.478000e+03	
4	20.0	23.0	2.0	6.0	2.0	2206.039454	0	8.112500e+02	

In [ ]:

```
df_no_ohe.head()
```

Out[ ]:

	_id	RIAGENDR	RIDAGEYR	RIDRETH1	DMDBORN4	DMDCITZN	DMDHHSIZ	INDFMINC	DMD
0	2.0	1.0	77.0	3.0	1.0	1.0	1.0	8.0	
1	5.0	1.0	49.0	3.0	1.0	1.0	3.0	11.0	
2	12.0	1.0	37.0	3.0	1.0	1.0	4.0	11.0	
3	15.0	2.0	38.0	3.0	1.0	1.0	2.0	8.0	
4	20.0	2.0	23.0	1.0	1.0	1.0	2.0	6.0	

Drop Highly Correlated Variables

In [ ]:

```
df_no_ohe.drop(['BMXWT','BMXHT','BMXWAIST'], axis=1, inplace=True)
df_ohe.drop(['BMXWT','BMXHT','BMXWAIST'], axis=1, inplace=True)
```

In [ ]:

```
df_ohe.shape
```

Out[ ]:

```
(14121, 67)
```

MongoDB Insertion

In [ ]:

```
#Import MongoDIient
from pymongo import MongoClient

#Create a MongoDIient to run the MongoDB instance
Client = MongoClient("localhost", 27017)
```

In [ ]:

```
#Connect to existing database
db = Client.NHANES_Q2
db
```

Out[ ]:

```
Database(MongoClient(host=['localhost:27017'], document_class=dict, tz_aware=Fal
se, connect=True), 'NHANES_Q2')
```

In [ ]:

```
db.list_collection_names()
```

Out[ ]:

```
['DI_no_ohe', 'DI']
```

In [ ]:

```
#Creating a collection
DI = db.DI
DI_no_ohe = db.DI_no_ohe
```

In [ ]:

```
#If collections exist, then drop
if 'DI' in db.list_collection_names():
    DI.drop()
    db.list_collection_names()

if 'DI_no_ohe' in db.list_collection_names():
    DI_no_ohe.drop()
    db.list_collection_names()
```

In [ ]:

```
#MongoDB keys DIn't contain '.'
df_ohe.columns = df_ohe.columns.str.replace(".", "_")
```

```
/var/folders/4n/wd_5b1m97rs5m_qdhsvl_lqh0000gn/T/ipykernel_64519/1829403158.py:
2: FutureWarning: The default value of regex will change from True to False in a
future version. In addition, single character regular expressions will *not* be
treated as literal strings when regex=True.
  df_ohe.columns = df_ohe.columns.str.replace(".", "_")
```

```
In [ ]:  df_ohe.head()
```

```
Out[ ]:
      _id  RIDAGEYR  DMDHHSIZ  INDFMINC  DMDHREDU       MEC18YR  DIQ010      DRD320GW  D
   0  2.0      77.0       1.0       8.0       5.0   3408.044382       0  5.397605e-79
   1  5.0      49.0       3.0      11.0       4.0  10219.103963       0  1.298000e+03
   2 12.0      37.0       4.0      11.0       2.0  10149.365568       0  3.304000e+03
   3 15.0      38.0       2.0       8.0       5.0  11437.714415       0  2.478000e+03
   4 20.0      23.0       2.0       6.0       2.0   2206.039454       0  8.112500e+02
```

```
In [ ]:  #Dataframe to dictionary
         DI_dict = df_ohe.to_dict(orient='records')

         DI_no_ohe_dict = df_no_ohe.to_dict(orient='records')
```

```
In [ ]:  DI_dict[0]
```

```
Out[ ]: {'_id': 2.0,
         'RIDAGEYR': 77.0,
         'DMDHHSIZ': 1.0,
         'INDFMINC': 8.0,
         'DMDHREDU': 5.0,
         'MEC18YR': 3408.0443815555554,
         'DIQ010': 0,
         'DRD320GW': 5.397605346934028e-79,
         'DRDTSODI': 5710.03,
         'DRX18YR': 3315.985398314134,
         'DRXTALCO': 5.397605346934028e-79,
         'DRXTCAFF': 530.45,
         'DRXTCALC': 925.37,
         'DRXTCARB': 350.37,
         'DRXTCHOL': 313.95,
         'DRXTCOPP': 2.08,
         'DRXTFIBE': 36.99,
         'DRXTIRON': 37.29,
         'DRXTKCAL': 2463.0,
         'DRXTMAGN': 502.25,
         'DRXTPHOS': 1974.57,
         'DRXTPOTA': 4672.48,
         'DRXTPROT': 123.16,
         'DRXTTFAT': 71.95,
         'DRXTVARE': 923.91,
         'DRXTVB1': 2.11,
         'DRXTVB12': 8.68,
         'DRXTVB2': 3.25,
         'DRXTVB6': 2.9,
         'DRXTVC': 119.12,
         'DRXTZINC': 41.61,
         'BPXSY1': 106.0,
         'BPXDI1': 58.0,
         'BMXBMI': 24.9,
         'LBXTC': 215.0,
         'LBDHDL': 54.0,
         'LBXTR': 128.0,
         'LBDLDL': 136.0,
         'HUQ010': 2.0,
         'HUQ050': 3.0,
         'Year': 0,
         'RIAGENDR_2_0': 0,
         'RIDRETH1_2_0': 0,
         'RIDRETH1_3_0': 1,
         'RIDRETH1_4_0': 0,
         'RIDRETH1_5_0': 0,
         'DMDBORN4_2_0': 0,
         'DMDCITZN_2_0': 0,
         'ALQ101_2_0': 0,
         'BPQ020_2_0': 1,
         'BPXPULS_2_0': 0,
         'PAQ635_2_0': 1,
         'PAQ650_2_0': 1,
         'PAQ665_2_0': 1,
         'SMQ680_2_0': 1,
         'SMAQUEX_2_0': 1,
         'SMD410_1_0': 0,
         'SMD410_2_0': 1,
         'SMD410_3_0': 0,
         'SMD410_777_0': 0,
         'SMD410_999_0': 0,
         'HID010_2_0': 0,
         'HUQ020_2_0': 1,
         'HUQ020_3_0': 0,
         'HUQ030_2_0': 0,
         'HUQ030_3_0': 0,
         'HUQ070_2_0': 0}
```

In [ ]:

```
#Insert collection
DI.insert_many(DI_dict)
```

Out[ ]:

```
<pymongo.results.InsertManyResult at 0x7fd5c828dac0>
```

In [ ]:

```
DI_no_ohe_dict[0]
```

Out[ ]:

```
{'_id': 2.0,
 'RIAGENDR': 1.0,
 'RIDAGEYR': 77.0,
 'RIDRETH1': 3.0,
 'DMDBORN4': 1.0,
 'DMDCITZN': 1.0,
 'DMDHHSIZ': 1.0,
 'INDFMINC': 8.0,
 'DMDHREDU': 5.0,
 'MEC18YR': 3408.0443815555554,
 'ALQ101': 1.0,
 'DIQ010': 0.0,
 'DRD320GW': 5.397605346934028e-79,
 'DRDTSODI': 5710.03,
 'DRX18YR': 3315.985398314134,
 'DRXTALCO': 5.397605346934028e-79,
 'DRXTCAFF': 530.45,
 'DRXTCALC': 925.37,
 'DRXTCARB': 350.37,
 'DRXTCHOL': 313.95,
 'DRXTCOPP': 2.08,
 'DRXTFIBE': 36.99,
 'DRXTIRON': 37.29,
 'DRXTKCAL': 2463.0,
 'DRXTMAGN': 502.25,
 'DRXTPHOS': 1974.57,
 'DRXTPOTA': 4672.48,
 'DRXTPROT': 123.16,
 'DRXTTFAT': 71.95,
 'DRXTVARE': 923.91,
 'DRXTVB1': 2.11,
 'DRXTVB12': 8.68,
 'DRXTVB2': 3.25,
 'DRXTVB6': 2.9,
 'DRXTVC': 119.12,
 'DRXTZINC': 41.61,
 'BPQ020': 2.0,
 'BPXPULS': 1.0,
 'BPXSY1': 106.0,
 'BPXDI1': 58.0,
 'BMXBMI': 24.9,
 'LBXTC': 215.0,
 'LBDHDL': 54.0,
 'LBXTR': 128.0,
 'LBDLDL': 136.0,
 'PAQ635': 2.0,
 'PAQ650': 2.0,
 'PAQ665': 2.0,
 'SMQ680': 2.0,
 'SMAQUEX': 2.0,
 'SMD410': 2.0,
 'HID010': 1.0,
 'HUQ010': 2.0,
 'HUQ020': 2.0,
 'HUQ030': 1.0,
 'HUQ050': 3.0,
 'HUQ070': 1.0,
 'Year': 0}
```

In [ ]:

```
DI_no_ohe.insert_many(DI_no_ohe_dict)
```

Out[ ]:

```
<pymongo.results.InsertManyResult at 0x7fd5dc519f70>
```

In [ ]:

```
db.list_collection_names()
```

Out[ ]:

```
['DI', 'DI_no_ohe']
```