project_part_2_Data_Selection_Cleaning

February 15, 2022

1 PROJECT NOTEBOOK 1: DATA SELECTION FROM BRFSS

1.0.1 Team 3

- Anjali Sebastian
- Yesha Sharma
- Rupansh Phutela

1.0.2 What this Notebook does?

- We are selecting relevant features and target columns from the BRFSS (Behavioral Risk Factor Surveillance System) 2019 data downloaded from the CDC.
- We then clean the data to remove resposes that were null.
- Rename all the columns for better understanding
- Write our data to the diabetes.csv file
- Read and check if file is written correctly

Note this notebook needs to be run only once to create the dataset

1.0.3 1. Import Packages

[1]:

```
# you need  Python  3.5
import sys
assert sys.version_info >= (3, 5)
```

[2]:

```
# Scikit-Learn  0.20 is required
import sklearn
assert sklearn.__version__ >= "0.20"
```

[3]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import time
import warnings
warnings.filterwarnings("ignore")
###############################################
```

1.0.4 2. Read the Complete BRFSS 2019 data

```
[4]: # Link to SAS File - https://www.cdc.gov/brfss/annual_data/2019/files/
      ↪LLCP2019XPT.zip
     # location outside the git repository file is is too large to be uploaded to git
     brfss = pd.read_csv('../../BRFSS_2019.csv')
```

```
[5]: brfss.shape
```

```
[5]: (418268, 342)
```

```
[6]: brfss.head()
```

```
[6]:    _STATE  FMONTH    IDATE  IMONTH  IDAY  IYEAR  DISPCODE       SEQNO  \
     0     1.0     1.0  1182019       1    18   2019    1100.0  2019000001
     1     1.0     1.0  1132019       1    13   2019    1100.0  2019000002
     2     1.0     1.0  1182019       1    18   2019    1100.0  2019000003
     3     1.0     1.0  1182019       1    18   2019    1200.0  2019000004
     4     1.0     1.0  1042019       1     4   2019    1100.0  2019000005

                _PSU  CTELENM1  …  _VEGESU1  _FRTLT1A  _VEGLT1A  _FRT16A  \
     0  2.019000e+09       1.0  …     114.0       1.0       1.0      1.0
     1  2.019000e+09       1.0  …     121.0       1.0       1.0      1.0
     2  2.019000e+09       1.0  …     164.0       1.0       1.0      1.0
     3  2.019000e+09       1.0  …       NaN       9.0       9.0      1.0
     4  2.019000e+09       1.0  …     178.0       1.0       1.0      1.0

        _VEG23A  _FRUITE1  _VEGETE1  _FLSHOT7  _PNEUMO3  _AIDTST4
     0      1.0       0.0       0.0       2.0       1.0       2.0
     1      1.0       0.0       0.0       1.0       1.0       2.0
     2      1.0       0.0       0.0       1.0       2.0       2.0
     3      1.0       1.0       1.0       9.0       9.0       NaN
     4      1.0       0.0       0.0       2.0       1.0       2.0

     [5 rows x 342 columns]
```

```
[7]: brfss.tail()
```

```
[7]:         _STATE  FMONTH    IDATE  IMONTH  IDAY  IYEAR  DISPCODE       SEQNO  \
     418263    72.0     9.0  3152020       3    15   2020    1100.0  2019006029
     418264    72.0     9.0  3082020       3     8   2020    1100.0  2019006030
     418265    72.0     9.0  3102020       3    10   2020    1100.0  2019006031
     418266    72.0     9.0  3062020       3     6   2020    1100.0  2019006032
     418267    72.0     9.0  3052020       3     5   2020    1100.0  2019006033

                     _PSU  CTELENM1  …  _VEGESU1  _FRTLT1A  _VEGLT1A  _FRT16A  \
     418263  2.019006e+09       NaN  …      43.0       1.0       2.0      1.0
     418264  2.019006e+09       NaN  …     142.0       1.0       1.0      1.0
```

```
418265  2.019006e+09       NaN  …     55.0       1.0       2.0      1.0
418266  2.019006e+09       NaN  …    214.0       1.0       1.0      1.0
418267  2.019006e+09       NaN  …    229.0       1.0       1.0      1.0

        _VEG23A  _FRUITE1  _VEGETE1  _FLSHOT7  _PNEUMO3  _AIDTST4
418263      1.0       0.0       0.0       2.0       2.0       2.0
418264      1.0       0.0       0.0       NaN       NaN       2.0
418265      1.0       0.0       0.0       NaN       NaN       1.0
418266      1.0       0.0       0.0       2.0       2.0       2.0
418267      1.0       0.0       0.0       NaN       NaN       2.0

[5 rows x 342 columns]
```

Note: We can see that the complete brfss dataset has about 0.4 million records and 342 columns. We are only interested in the columns related to the prediction of diabetes in individuals. The current column names are based on a code book that the CDC maintains. We will selected relavent columns using the code book to map key indicator as mentioned in Project 1 report. - Link to code book: https://www.cdc.gov/brfss/annual_data/annual_2019.html - Link to reference: https://www.kaggle.com/alexteboul/diabetes-health-indicators-dataset-notebook

1.0.5 3. Selecting Relevant Columns from BRFSS Data

[8]:

```
cols_to_select = ["DIABETE4","_BMI5","_STATE", # target variable
                  "_RFHYPE5","TOLDHI2","_CHOLCH2", # BP and cholestrol
                  "_FRTLT1A","_VEGLT1A", "SMOKE100","_RFDRHV7", # Food, alcohol␣
 ↪and smoking
                  "CVDSTRK3","_MICHD", # chronic diseases
                  "HLTHPLN1","MEDCOST", # Insurance and medical access
                  "_TOTINDA","GENHLTH","PHYSHLTH","MENTHLTH","DIFFWALK",␣
 ↪#Fitness and activity
                  "SEXVAR","_AGEG5YR","EDUCA","INCOME2"] # demographic
```

[9]:

```
brfss_cut = brfss.loc[:,cols_to_select]
brfss_cut.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 418268 entries, 0 to 418267
Data columns (total 23 columns):
 #   Column    Non-Null Count   Dtype
---  ------    --------------   -----
 0   DIABETE4  418259 non-null  float64
 1   _BMI5     382065 non-null  float64
 2   _STATE    418268 non-null  float64
 3   _RFHYPE5  418268 non-null  float64
 4   TOLDHI2   393825 non-null  float64
 5   _CHOLCH2  418268 non-null  float64
 6   _FRTLT1A  418268 non-null  float64
 7   _VEGLT1A  418268 non-null  float64
 8   SMOKE100  402277 non-null  float64
 9   _RFDRHV7  418268 non-null  float64
 10  CVDSTRK3  418257 non-null  float64
 11  _MICHD    413943 non-null  float64
 12  HLTHPLN1  418259 non-null  float64
 13  MEDCOST   418261 non-null  float64
 14  _TOTINDA  418268 non-null  float64
 15  GENHLTH   418242 non-null  float64
 16  PHYSHLTH  418236 non-null  float64
 17  MENTHLTH  418249 non-null  float64
 18  DIFFWALK  404506 non-null  float64
 19  SEXVAR    418268 non-null  float64
 20  _AGEG5YR  418268 non-null  float64
 21  EDUCA     418242 non-null  float64
 22  INCOME2   411387 non-null  float64
dtypes: float64(23)
memory usage: 73.4 MB
```

[10]:

```
brfss_cut.shape
```

[10]:

```
(418268, 23)
```

[11]:

```
brfss_cut.head()
```

[11]:

```
   DIABETE4   _BMI5  _STATE  _RFHYPE5  TOLDHI2  _CHOLCH2  _FRTLT1A  _VEGLT1A  \
0       3.0  2817.0     1.0       2.0      1.0       1.0       1.0       1.0
1       3.0  1854.0     1.0       1.0      2.0       1.0       1.0       1.0
2       1.0  3162.0     1.0       2.0      2.0       1.0       1.0       1.0
3       3.0  2030.0     1.0       2.0      2.0       1.0       9.0       9.0
4       3.0  2148.0     1.0       1.0      1.0       1.0       1.0       1.0

   SMOKE100  _RFDRHV7  …  MEDCOST  _TOTINDA  GENHLTH  PHYSHLTH  MENTHLTH  \
0       1.0       1.0  …      2.0       2.0      3.0      15.0      88.0
1       2.0       1.0  …      2.0       1.0      4.0      10.0      88.0
2       2.0       1.0  …      2.0       1.0      3.0      88.0      30.0
3       NaN       9.0  …      2.0       9.0      4.0      30.0      88.0
4       1.0       1.0  …      2.0       2.0      2.0      88.0      88.0

   DIFFWALK  SEXVAR  _AGEG5YR  EDUCA  INCOME2
0       1.0     2.0      13.0    3.0      3.0
1       2.0     2.0      11.0    5.0      5.0
2       1.0     2.0      10.0    6.0      7.0
3       NaN     2.0      13.0    5.0      6.0
4       2.0     2.0      13.0    5.0     99.0

[5 rows x 23 columns]
```

[12]:

```
brfss_cut.tail()
```

[12]:
```
        DIABETE4   _BMI5  _STATE  _RFHYPE5  TOLDHI2  _CHOLCH2  _FRTLT1A  \
418263       1.0  2717.0    72.0       2.0      2.0       1.0       1.0
418264       3.0  2852.0    72.0       1.0      2.0       1.0       1.0
418265       3.0  3175.0    72.0       1.0      1.0       1.0       1.0
418266       3.0  2378.0    72.0       1.0      1.0       1.0       1.0
418267       3.0  1997.0    72.0       1.0      2.0       1.0       1.0

        _VEGLT1A  SMOKE100  _RFDRHV7  …  MEDCOST  _TOTINDA  GENHLTH  \
418263       2.0       2.0       1.0  …      2.0       2.0      3.0
418264       1.0       2.0       1.0  …      2.0       1.0      2.0
418265       2.0       2.0       1.0  …      2.0       1.0      2.0
418266       1.0       2.0       1.0  …      2.0       1.0      3.0
418267       1.0       2.0       1.0  …      2.0       1.0      3.0

        PHYSHLTH  MENTHLTH  DIFFWALK  SEXVAR  _AGEG5YR  EDUCA  INCOME2
418263      88.0      88.0       2.0     1.0      10.0    4.0      1.0
418264      88.0      88.0       2.0     2.0       7.0    3.0      1.0
418265      88.0       5.0       2.0     2.0       1.0    4.0      3.0
418266      88.0      88.0       2.0     2.0      11.0    4.0     99.0
418267      88.0      88.0       2.0     2.0       5.0    6.0      8.0

[5 rows x 23 columns]
```

1.0.6 3. Clean Data

- Drop missing values
- Modify and clean the values to be more suitable to ML algorithms
- Rename Columns for clarity

3.0 Drop all Null Values

[13]:
```
brfss_cut=brfss_cut.dropna()
brfss_cut.shape
```

[13]:
```
(351875, 23)
```

[14]:
```
brfss_df_selected = brfss_cut.copy(deep=True)
```

3.1 DIABETE4

- Making this a Boolean Binary.
- 0 is for No Diabetes or only during pregnancy or prediabetes.
- 1 is for diabetes
- Remove all 7 (dont knows)
- Remove all 9 (refused)

[15]:
```
brfss_df_selected['DIABETE4'].value_counts()
```

[15]:
```
3.0    289626
1.0     50713
4.0      8053
2.0      3019
7.0       412
9.0        52
Name: DIABETE4, dtype: int64
```

[16]:
```
brfss_df_selected['DIABETE4'] = brfss_df_selected['DIABETE4'].replace({2:0, 3:
 ↪0, 4:0, 1:1})
brfss_df_selected = brfss_df_selected[brfss_df_selected.DIABETE4 != 7]
brfss_df_selected = brfss_df_selected[brfss_df_selected.DIABETE4 != 9]
brfss_df_selected.DIABETE4.unique()
```

[16]:
```
array([0., 1.])
```

[17]:
```
brfss_df_selected['DIABETE4'].value_counts()
```

[17]:
```
0.0    300698
1.0     50713
Name: DIABETE4, dtype: int64
```

3.2 _BMI5

- no changes, just note that these are BMI * 100. So for example a BMI of 4018 is really 40.18

[18]:
```
brfss_df_selected['_BMI5'] = brfss_df_selected['_BMI5'].div(100)
brfss_df_selected._BMI5.unique()
```

[18]:
```
array([28.17, 18.54, 31.62, …, 52.16, 51.9 , 58.89])
```

3.3 _STATE

- This is only for EDA to see if any patterns emerge
- Replace the numbers to corresponding state 2 letter codes.

[19]:
```
us_state_to_abbrev = {
    1 : "AL",
    2 : "AK",
    4 : "AZ",
    5 : "AR",
    6 : "CA",
    8 : "CO",
    9 : "CT",
    10 : "DE",
    11 : "FL",
    12 : "DC",
    13 : "GA",
    15 : "HI",
    16 : "ID",
    17 : "IL",
    18 : "IN",
    19 : "IA",
    20 : "KS",
    21 : "KY",
    22 : "LA",
    23 : "ME",
    24 : "MD",
    25 : "MA",
    26 : "MI",
    27 : "MN",
    28 : "MS",
    29 : "MO",
    30 : "MT",
    31 : "NE",
    32 : "NV",
    33 : "NH",
    35 : "NM",
    36 : "NY",
    37 : "NC",
    38 : "ND",
    39 : "OH",
    40 : "OK",
    41 : "OR",
    42 : "PA",
    44 : "RI",
    45 : "SC",
    46 : "SD",
    47 : "TN",
    48 : "TX",
    49 : "UT",
    50 : "VT",
    51 : "VA",
    53 : "WA",
    54 : "WV",
    55 : "WI",
    56 : "WY",
    66 : "GU",
    72 : "PR",
}
```

[20]:

```
brfss_df_selected['_STATE'] = brfss_df_selected['_STATE'].
 ↪replace(us_state_to_abbrev)
```

[21]:

```
brfss_df_selected['_STATE'].value_counts()
```

```
[21]: MD    14969
      DC    13897
      NE    13358
      MN    12994
      NY    11632
      OH    11378
      WA    10629
      TX    10086
      ME     9751
      UT     9695
      CA     9620
      KS     9413
      MI     9232
      VA     8431
      IA     8119
      CT     7790
      CO     7674
      AZ     7502
      IN     7333
      KY     6911
      HI     6548
      MA     6331
      MO     6197
      AL     6071
      SC     6068
      GA     5987
      SD     5674
      PA     5654
      PR     5533
      MT     5494
      VT     5426
      TN     5280
      OK     5240
      RI     5130
      OR     5068
      NM     4994
      NH     4897
      ND     4824
      IL     4811
      WV     4693
      AR     4491
      ID     4409
      MS     4390
      WI     4118
      WY     4012
      LA     3950
      NC     3577
```

```
DE     3161
AK     2440
NV     2350
FL     2196
GU     1983
Name: _STATE, dtype: int64
```

3.4 _RFHYPE5

- Change 1 to 0 so it represents No high blood pressure and 2 to 1 so it represents high blood pressure

[22]:
```
brfss_df_selected['_RFHYPE5'] = brfss_df_selected['_RFHYPE5'].replace({1:0, 2:
 ↪1})
brfss_df_selected = brfss_df_selected[brfss_df_selected._RFHYPE5 != 9] # didnt␣
 ↪repond
brfss_df_selected._RFHYPE5.unique()
```

[22]:
```
array([1., 0.])
```

3.5 TOLDHI2

[23]:
```
# Change 2 to 0 because it is No
# Remove all 7 (dont knows)
# Remove all 9 (refused)
brfss_df_selected['TOLDHI2'] = brfss_df_selected['TOLDHI2'].replace({2:0})
brfss_df_selected = brfss_df_selected[brfss_df_selected.TOLDHI2 != 7]
brfss_df_selected = brfss_df_selected[brfss_df_selected.TOLDHI2 != 9]
brfss_df_selected.TOLDHI2.unique()
```

[23]:
```
array([1., 0.])
```

3.6 _CHOLCH2

[24]:
```
# Keep  1 to 1 has checked cholestrol in past 5 years
# 2 to 0 for Not checked cholesterol in past 5 years
# 3 to 0 never had cholestrol checked
# Remove 9
brfss_df_selected['_CHOLCH2'] = brfss_df_selected['_CHOLCH2'].replace({3:0,2:0})
brfss_df_selected = brfss_df_selected[brfss_df_selected._CHOLCH2 != 9]
brfss_df_selected._CHOLCH2.unique()
```

[24]:
```
array([1., 0.])
```

3.7 _FRTLT1A

[25]:
```
# Change 2 to 0. this means no fruit consumed per day.
# 1 will mean consumed 1 or more pieces of fruit per day
# remove all dont knows and missing 9
brfss_df_selected['_FRTLT1A'] = brfss_df_selected['_FRTLT1A'].replace({2:0})
```

```
brfss_df_selected = brfss_df_selected[brfss_df_selected._FRTLT1A != 9]
brfss_df_selected._FRTLT1A.unique()
```

[25]:
```
array([1., 0.])
```

3.8 _VEGLT1A

[26]:
```
# Change 2 to 0. this means no vegetables consumed per day.
# 1 will mean consumed 1 or more pieces of vegetable per day
# remove all dont knows and missing 9
brfss_df_selected['_VEGLT1A'] = brfss_df_selected['_VEGLT1A'].replace({2:0})
brfss_df_selected = brfss_df_selected[brfss_df_selected._VEGLT1A != 9]
brfss_df_selected._VEGLT1A.unique()
```

[26]:
```
array([1., 0.])
```

3.9 SMOKE100

[27]:
```
# 1 means person has consumed 100 cigarettes in lifetime
# Change 2 to 0 because it is No
# Remove all 7 (dont knows)
# Remove all 9 (refused)
brfss_df_selected['SMOKE100'] = brfss_df_selected['SMOKE100'].replace({2:0})
brfss_df_selected = brfss_df_selected[brfss_df_selected.SMOKE100 != 7]
brfss_df_selected = brfss_df_selected[brfss_df_selected.SMOKE100 != 9]
brfss_df_selected.SMOKE100.unique()
```

[27]:
```
array([1., 0.])
```

3.10 _RFDRHV7

[28]:
```
# Change 1 to 0 (1 was no for heavy drinking).
# change all 2 to 1 (2 was yes for heavy drinking)
# remove all dont knows and missing 9
brfss_df_selected['_RFDRHV7'] = brfss_df_selected['_RFDRHV7'].replace({1:0, 2:
 ↪1})
brfss_df_selected = brfss_df_selected[brfss_df_selected._RFDRHV7 != 9]
brfss_df_selected._RFDRHV7.unique()
```

[28]:
```
array([0., 1.])
```

3.11 CVDSTRK3

[29]:
```
# Ever Had a stroke - 1 is Yes
# Change 2 to 0 because it is No
# Remove all 7 (dont knows)
# Remove all 9 (refused)
brfss_df_selected['CVDSTRK3'] = brfss_df_selected['CVDSTRK3'].replace({2:0})
brfss_df_selected = brfss_df_selected[brfss_df_selected.CVDSTRK3 != 7]
brfss_df_selected = brfss_df_selected[brfss_df_selected.CVDSTRK3 != 9]
brfss_df_selected.CVDSTRK3.unique()
```

```
[29]: array([0., 1.])
```

3.12 _MICHD

```
[30]: #  ever reported having coronary heart disease (CHD) or myocardial infarction␣
       ↪(MI) - yes is 1
      # Change 2 to 0 because this means did not have MI or CHD
      brfss_df_selected['_MICHD'] = brfss_df_selected['_MICHD'].replace({2: 0})
      brfss_df_selected._MICHD.unique()
```

```
[30]: array([0., 1.])
```

3.13 HLTHPLN1

```
[31]: # 1 is yes, Person has health coverage
      # change 2 to 0 because it is No health care access
      # remove 7 and 9 for don't know or refused
      brfss_df_selected['HLTHPLN1'] = brfss_df_selected['HLTHPLN1'].replace({2:0})
      brfss_df_selected = brfss_df_selected[brfss_df_selected.HLTHPLN1 != 7]
      brfss_df_selected = brfss_df_selected[brfss_df_selected.HLTHPLN1 != 9]
      brfss_df_selected.HLTHPLN1.unique()
```

```
[31]: array([1., 0.])
```

3.14 MEDCOST

```
[32]: # Did not go see doctor in last 12 months due to cost ? Yes = 1
      # Change 2 to 0 for no, 1 is already yes
      # remove 7 for don/t know and 9 for refused
      brfss_df_selected['MEDCOST'] = brfss_df_selected['MEDCOST'].replace({2:0})
      brfss_df_selected = brfss_df_selected[brfss_df_selected.MEDCOST != 7]
      brfss_df_selected = brfss_df_selected[brfss_df_selected.MEDCOST != 9]
      brfss_df_selected.MEDCOST.unique()
```

```
[32]: array([0., 1.])
```

3.15 _TOTINDA

```
[33]: # Adults who reported doing physical activity or exercise during the past 30␣
       ↪days other than their regular job
      # 1 for physical activity
      # change 2 to 0 for no physical activity
      # Remove all 9 (don't know/refused)
      brfss_df_selected['_TOTINDA'] = brfss_df_selected['_TOTINDA'].replace({2:0})
      brfss_df_selected = brfss_df_selected[brfss_df_selected._TOTINDA != 9]
      brfss_df_selected._TOTINDA.unique()
```

```
[33]: array([0., 1.])
```

3.16 GENHLTH

```
[34]: brfss_df_selected['GENHLTH'].value_counts()
```

[34]:
```
2.0    97819
3.0    90455
1.0    45667
4.0    40067
5.0    14568
7.0      271
9.0       95
Name: GENHLTH, dtype: int64
```

[35]:
```
#Would you say that in general your health is:
# This is an ordinal variable  (1 is Excellent -> 5 is Poor) we will reverse it␣
 ↪so that(1 is poor and 5 is excellent )
# Remove 7 and 9 for don't know and refused
brfss_df_selected['GENHLTH'] = brfss_df_selected['GENHLTH'].replace({5:1,4:2,2:
 ↪4,1:5})
brfss_df_selected = brfss_df_selected[brfss_df_selected.GENHLTH != 7]
brfss_df_selected = brfss_df_selected[brfss_df_selected.GENHLTH != 9]
brfss_df_selected.GENHLTH.unique()
```

[35]:
```
array([3., 2., 4., 5., 1.])
```

[36]:
```
brfss_df_selected['GENHLTH'].value_counts()
```

[36]:
```
4.0    97819
3.0    90455
5.0    45667
2.0    40067
1.0    14568
Name: GENHLTH, dtype: int64
```

3.17 PHYSHLTH

[37]:
```
# for how many days during the past 30 days was your physical health not good?
# already in days so keep that, scale will be 0-30
# change 88 to 0 because it means none (no bad physical health days)
# remove 77 and 99 for don't know not sure and refused
brfss_df_selected['PHYSHLTH'] = brfss_df_selected['PHYSHLTH'].replace({88:0})
brfss_df_selected = brfss_df_selected[brfss_df_selected.PHYSHLTH != 77]
brfss_df_selected = brfss_df_selected[brfss_df_selected.PHYSHLTH != 99]
brfss_df_selected.PHYSHLTH.unique()
```

[37]:
```
array([15., 10.,  0., 30., 20.,  2.,  1.,  7., 14.,  3.,  5., 25.,  4.,
        6., 28., 21., 17.,  8., 16., 27., 12., 23., 18., 13., 29., 19.,
        9., 24., 26., 11., 22.])
```

3.18 MENTHLTH

[38]:
```
# for how many days during the past 30 days was your mental health not good?
# already in days so keep that, scale will be 0-30
```

```python
# change 88 to 0 because it means none (no bad mental health days)
# remove 77 and 99 for don't know not sure and refused
brfss_df_selected['MENTHLTH'] = brfss_df_selected['MENTHLTH'].replace({88:0})
brfss_df_selected = brfss_df_selected[brfss_df_selected.MENTHLTH != 77]
brfss_df_selected = brfss_df_selected[brfss_df_selected.MENTHLTH != 99]
brfss_df_selected.MENTHLTH.unique()
```

[38]:
```
array([ 0., 30.,  4.,  1.,  2., 15.,  5.,  7., 10.,  3., 25.,  6., 21.,
       20.,  8., 14., 17., 28., 12., 16., 27., 23., 26., 29., 24.,  9.,
       13., 18., 22., 11., 19.])
```

3.19 DIFFWALK

[39]:
```python
# Do you have serious difficulty walking or climbing stairs? yes =1
# change 2 to 0 for no. 1 is already yes
# remove 7 and 9 for don't know not sure and refused
brfss_df_selected['DIFFWALK'] = brfss_df_selected['DIFFWALK'].replace({2:0})
brfss_df_selected = brfss_df_selected[brfss_df_selected.DIFFWALK != 7]
brfss_df_selected = brfss_df_selected[brfss_df_selected.DIFFWALK != 9]
brfss_df_selected.DIFFWALK.unique()
```

[39]:
```
array([1., 0.])
```

3.20 SEXVAR

[40]:
```python
# in other words - is respondent male
# men may be  at higher risk for heart disease
# change 2 to 0 (female as 0). Male is 1
brfss_df_selected['SEXVAR'] = brfss_df_selected['SEXVAR'].replace({2:0})
brfss_df_selected.SEXVAR.unique()
```

[40]:
```
array([0., 1.])
```

3.21 _AGEG5YR

[41]:
```python
# Reported age in five-year age categories calculated variable
# already ordinal. 1 is 18-24 all the way up to 13 wis 80 and older. 5 year␣
 ↪increments.
# remove 14 because it is don't know or missing
brfss_df_selected = brfss_df_selected[brfss_df_selected._AGEG5YR != 14]
brfss_df_selected._AGEG5YR.unique()
```

[41]:
```
array([13., 11., 10.,  8., 12.,  7.,  6.,  5.,  9.,  4.,  3.,  2.,  1.])
```

3.22 EDUCA

[42]:
```python
# Level of education completed - This is already an ordinal variable
# 1 being never attended school or kindergarten only up to 6 being college 4␣
 ↪years or more
# Scale here is 1-6
# Remove 9 for refused:
```

```
brfss_df_selected = brfss_df_selected[brfss_df_selected.EDUCA != 9]
brfss_df_selected.EDUCA.unique()
```

[42]:
```
array([3., 5., 6., 2., 4., 1.])
```

3.23 INCOME2

[43]:
```
#  Annual household income - in levels
# Variable is already ordinal with 1 being less than $10,000 all the way up to␣
 ↪8 being $75,000 or more
# Remove 77 and 99 for don't know and refused
brfss_df_selected = brfss_df_selected[brfss_df_selected.INCOME2 != 77]
brfss_df_selected = brfss_df_selected[brfss_df_selected.INCOME2 != 99]
brfss_df_selected.INCOME2.unique()
```

[43]:
```
array([3., 5., 7., 8., 6., 4., 2., 1.])
```

1.0.7 4 Check Mapping Changes

[44]:
```
brfss_df_selected.shape
```

[44]:
```
(243317, 23)
```

[45]:
```
brfss_df_selected.isna().sum()
```

[45]:
```
DIABETE4    0
_BMI5       0
_STATE      0
_RFHYPE5    0
TOLDHI2     0
_CHOLCH2    0
_FRTLT1A    0
_VEGLT1A    0
SMOKE100    0
_RFDRHV7    0
CVDSTRK3    0
_MICHD      0
HLTHPLN1    0
MEDCOST     0
_TOTINDA    0
GENHLTH     0
PHYSHLTH    0
MENTHLTH    0
DIFFWALK    0
SEXVAR      0
_AGEG5YR    0
EDUCA       0
INCOME2     0
```

```
dtype: int64
```

[46]:

```
brfss_df_selected.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 243317 entries, 0 to 418267
Data columns (total 23 columns):
 #   Column    Non-Null Count   Dtype
---  ------    --------------   -----
 0   DIABETE4  243317 non-null  float64
 1   _BMI5     243317 non-null  float64
 2   _STATE    243317 non-null  object
 3   _RFHYPE5  243317 non-null  float64
 4   TOLDHI2   243317 non-null  float64
 5   _CHOLCH2  243317 non-null  float64
 6   _FRTLT1A  243317 non-null  float64
 7   _VEGLT1A  243317 non-null  float64
 8   SMOKE100  243317 non-null  float64
 9   _RFDRHV7  243317 non-null  float64
 10  CVDSTRK3  243317 non-null  float64
 11  _MICHD    243317 non-null  float64
 12  HLTHPLN1  243317 non-null  float64
 13  MEDCOST   243317 non-null  float64
 14  _TOTINDA  243317 non-null  float64
 15  GENHLTH   243317 non-null  float64
 16  PHYSHLTH  243317 non-null  float64
 17  MENTHLTH  243317 non-null  float64
 18  DIFFWALK  243317 non-null  float64
 19  SEXVAR    243317 non-null  float64
 20  _AGEG5YR  243317 non-null  float64
 21  EDUCA     243317 non-null  float64
 22  INCOME2   243317 non-null  float64
dtypes: float64(22), object(1)
memory usage: 44.6+ MB
```

Note: We will change to appropriate datatypes in next notebook. Here we will write to a new .csv file and read it in the next notebook and do the datatype conversions after reading

[47]:

```
brfss_df_selected.head()
```

[47]:

```
   DIABETE4  _BMI5 _STATE  _RFHYPE5  TOLDHI2  _CHOLCH2  _FRTLT1A  _VEGLT1A  \
0       0.0  28.17     AL       1.0      1.0       1.0       1.0       1.0
1       0.0  18.54     AL       0.0      0.0       1.0       1.0       1.0
2       1.0  31.62     AL       1.0      0.0       1.0       1.0       1.0
6       1.0  32.98     AL       0.0      0.0       1.0       1.0       1.0
9       1.0  16.65     AL       0.0      1.0       1.0       0.0       0.0
```

```
   SMOKE100  _RFDRHV7  …  MEDCOST  _TOTINDA  GENHLTH  PHYSHLTH  MENTHLTH  \
0       1.0       0.0  …      0.0       0.0      3.0      15.0       0.0
1       0.0       0.0  …      0.0       1.0      2.0      10.0       0.0
2       0.0       0.0  …      0.0       1.0      3.0       0.0      30.0
6       1.0       0.0  …      0.0       1.0      4.0      30.0       0.0
9       1.0       0.0  …      0.0       0.0      1.0      20.0       0.0

   DIFFWALK  SEXVAR  _AGEG5YR  EDUCA  INCOME2
0       1.0     0.0      13.0    3.0      3.0
1       0.0     0.0      11.0    5.0      5.0
2       1.0     0.0      10.0    6.0      7.0
6       1.0     1.0      11.0    6.0      7.0
9       1.0     0.0      11.0    2.0      3.0

[5 rows x 23 columns]
```

[48]:
```python
brfss_df_selected.tail()
```

[48]:
```
        DIABETE4  _BMI5 _STATE  _RFHYPE5  TOLDHI2  _CHOLCH2  _FRTLT1A  \
418262       0.0  26.31     PR       0.0      0.0       1.0       0.0
418263       1.0  27.17     PR       1.0      0.0       1.0       1.0
418264       0.0  28.52     PR       0.0      0.0       1.0       1.0
418265       0.0  31.75     PR       0.0      1.0       1.0       1.0
418267       0.0  19.97     PR       0.0      0.0       1.0       1.0

        _VEGLT1A  SMOKE100  _RFDRHV7  …  MEDCOST  _TOTINDA  GENHLTH  \
418262       0.0       0.0       0.0  …      0.0       0.0      5.0
418263       0.0       0.0       0.0  …      0.0       0.0      3.0
418264       1.0       0.0       0.0  …      0.0       1.0      4.0
418265       0.0       0.0       0.0  …      0.0       1.0      4.0
418267       1.0       0.0       0.0  …      0.0       1.0      3.0

        PHYSHLTH  MENTHLTH  DIFFWALK  SEXVAR  _AGEG5YR  EDUCA  INCOME2
418262       0.0       0.0       0.0     0.0       1.0    6.0      4.0
418263       0.0       0.0       0.0     1.0      10.0    4.0      1.0
418264       0.0       0.0       0.0     0.0       7.0    3.0      1.0
418265       0.0       5.0       0.0     0.0       1.0    4.0      3.0
418267       0.0       0.0       0.0     0.0       5.0    6.0      8.0

[5 rows x 23 columns]
```

[49]:
```python
brfss_df_selected['DIABETE4'].value_counts()
```

[49]:
```
0.0    208018
1.0     35299
Name: DIABETE4, dtype: int64
```

[50]:

```
brfss_df_selected.value_counts()
```

[50]:

```
DIABETE4  _BMI5  _STATE  _RFHYPE5  TOLDHI2  _CHOLCH2  _FRTLT1A  _VEGLT1A
SMOKE100  _RFDRHV7  CVDSTRK3  _MICHD  HLTHPLN1  MEDCOST  _TOTINDA  GENHLTH
PHYSHLTH  MENTHLTH  DIFFWALK  SEXVAR  _AGEG5YR  EDUCA  INCOME2
1.0       27.12  OR      1.0       1.0      1.0       1.0       1.0       0.0
0.0       0.0       0.0     1.0       0.0      1.0       3.0      2.0       2.0
0.0       1.0     9.0       4.0    7.0        4
0.0       22.71  PA      0.0       0.0      1.0       1.0       1.0       0.0
0.0       0.0       0.0     1.0       0.0      1.0       5.0      0.0       0.0
0.0       0.0     4.0       6.0    8.0        3
          23.48  UT      0.0       0.0      1.0       1.0       1.0       0.0
0.0       0.0       0.0     1.0       0.0      1.0       4.0      5.0       1.0
0.0       1.0     1.0       5.0    8.0        3
          25.06  MN      0.0       0.0      1.0       1.0       1.0       0.0
0.0       0.0       0.0     1.0       0.0      1.0       4.0      0.0       0.0
0.0       0.0     6.0       6.0    8.0        3
          26.63  MN      0.0       0.0      1.0       1.0       1.0       0.0
0.0       0.0       0.0     1.0       0.0      1.0       4.0      0.0       0.0
0.0       0.0     4.0       6.0    8.0        3
                                                       ..
          25.79  VA      0.0       0.0      1.0       1.0       1.0       1.0
0.0       0.0       0.0     1.0       0.0      1.0       4.0      0.0       1.0
0.0       0.0     1.0       4.0    4.0        1
    1.0       0.0       0.0     1.0       1.0      1.0       5.0      0.0
0.0       0.0       0.0     2.0       5.0    8.0        1
                         1.0       0.0      1.0       1.0       1.0       1.0
0.0       1.0       0.0     1.0       0.0      1.0       5.0      0.0       0.0
0.0       0.0     11.0      5.0    8.0        1
                                   1.0      1.0       1.0       1.0       0.0
0.0       0.0       0.0     1.0       0.0      1.0       3.0      0.0       0.0
0.0       0.0     10.0      6.0    7.0        1
1.0       97.65  NY      1.0       1.0      1.0       1.0       0.0       0.0
0.0       0.0       0.0     1.0       0.0      0.0       3.0      0.0       0.0
0.0       1.0     13.0      5.0    8.0        1
Length: 243177, dtype: int64
```

[51]:

```
brfss_df_selected.describe()
```

[51]:

```
            DIABETE4          _BMI5       _RFHYPE5        TOLDHI2  \
count  243317.000000  243317.000000  243317.000000  243317.000000
mean        0.145074      28.673176       0.425683       0.388292
std         0.352176       6.401627       0.494447       0.487363
min         0.000000      12.000000       0.000000       0.000000
25%         0.000000      24.340000       0.000000       0.000000
50%         0.000000      27.460000       0.000000       0.000000
75%         0.000000      31.870000       1.000000       1.000000
```

```
max         1.000000      98.700000       1.000000       1.000000

            _CHOLCH2       _FRTLT1A       _VEGLT1A       SMOKE100  \
count  243317.000000  243317.000000  243317.000000  243317.000000
mean        0.962913       0.630293       0.821673       0.426242
std         0.188976       0.482726       0.382789       0.494531
min         0.000000       0.000000       0.000000       0.000000
25%         1.000000       0.000000       1.000000       0.000000
50%         1.000000       1.000000       1.000000       0.000000
75%         1.000000       1.000000       1.000000       1.000000
max         1.000000       1.000000       1.000000       1.000000

            _RFDRHV7       CVDSTRK3  …        MEDCOST       _TOTINDA  \
count  243317.000000  243317.000000  …  243317.000000  243317.000000
mean        0.062906       0.043589  …       0.093861       0.752985
std         0.242794       0.204180  …       0.291636       0.431277
min         0.000000       0.000000  …       0.000000       0.000000
25%         0.000000       0.000000  …       0.000000       1.000000
50%         0.000000       0.000000  …       0.000000       1.000000
75%         0.000000       0.000000  …       0.000000       1.000000
max         1.000000       1.000000  …       1.000000       1.000000

             GENHLTH       PHYSHLTH       MENTHLTH       DIFFWALK  \
count  243317.000000  243317.000000  243317.000000  243317.000000
mean        3.439891       4.402426       3.673463       0.168061
std         1.060404       8.831775       7.802452       0.373921
min         1.000000       0.000000       0.000000       0.000000
25%         3.000000       0.000000       0.000000       0.000000
50%         4.000000       0.000000       0.000000       0.000000
75%         4.000000       3.000000       3.000000       0.000000
max         5.000000      30.000000      30.000000       1.000000

              SEXVAR       _AGEG5YR          EDUCA        INCOME2
count  243317.000000  243317.000000  243317.000000  243317.000000
mean        0.469573       7.945277       5.073509       6.133201
std         0.499074       3.273054       0.974905       2.062683
min         0.000000       1.000000       1.000000       1.000000
25%         0.000000       6.000000       4.000000       5.000000
50%         0.000000       8.000000       5.000000       7.000000
75%         1.000000      10.000000       6.000000       8.000000
max         1.000000      13.000000       6.000000       8.000000

[8 rows x 22 columns]
```

Note: After performing the mapping from code book most of the data looks clean. BMI of 98.7 seems a bit high and could be an outlier. We will look at this more closely in the next notebook

1.0.8 5. Rename Columns for Better Understanding

[52]:

```
column_mapping = {"DIABETE4" : "Diabetes", "_BMI5" : "BMI","_STATE" : "State" ,
                  "_RFHYPE5" : "HighBP","TOLDHI2" : "HighChol", "_CHOLCH2" :␣
 ↪"CholCheck",
                  "_FRTLT1A" : "FruitConsume" , "_VEGLT1A" : "VegetableConsume" ,
                  "SMOKE100" : "Smoker" , "_RFDRHV7" : "HeavyDrinker",
                  "CVDSTRK3" : "Stroke" , "_MICHD" : "HeartDisease",
                  "HLTHPLN1" : "Healthcare" , "MEDCOST" : "NoDoctorDueToCost",
                   "_TOTINDA" : "PhysicalActivity", "GENHLTH" : "GeneralHealth",␣
 ↪"PHYSHLTH" : "PhysicalHealth",
                   "MENTHLTH" : "MentalHealth", "DIFFWALK" : "DifficultyWalking",
                   "SEXVAR" : "Gender", "_AGEG5YR" : "Age", "EDUCA" :␣
 ↪"Education" , "INCOME2" : "Income"}
```

[53]:

```
diabetes = brfss_df_selected.rename(columns=column_mapping)
```

[54]:

```
diabetes.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 243317 entries, 0 to 418267
Data columns (total 23 columns):
 #   Column             Non-Null Count   Dtype
---  ------             --------------   -----
 0   Diabetes           243317 non-null  float64
 1   BMI                243317 non-null  float64
 2   State              243317 non-null  object
 3   HighBP             243317 non-null  float64
 4   HighChol           243317 non-null  float64
 5   CholCheck          243317 non-null  float64
 6   FruitConsume       243317 non-null  float64
 7   VegetableConsume   243317 non-null  float64
 8   Smoker             243317 non-null  float64
 9   HeavyDrinker       243317 non-null  float64
 10  Stroke             243317 non-null  float64
 11  HeartDisease       243317 non-null  float64
 12  Healthcare         243317 non-null  float64
 13  NoDoctorDueToCost  243317 non-null  float64
 14  PhysicalActivity   243317 non-null  float64
 15  GeneralHealth      243317 non-null  float64
 16  PhysicalHealth     243317 non-null  float64
 17  MentalHealth       243317 non-null  float64
 18  DifficultyWalking  243317 non-null  float64
 19  Gender             243317 non-null  float64
 20  Age                243317 non-null  float64
 21  Education          243317 non-null  float64
 22  Income             243317 non-null  float64
dtypes: float64(22), object(1)
```

```
memory usage: 44.6+ MB
```

1.0.9 6. Write to File

[55]:
```
#Run this only once
#diabetes.to_csv("./diabetes.csv")
```

1.0.10 7. Read and Quick Check

[56]:
```
diabetes_read = pd.read_csv('./diabetes.csv')
```

[57]:
```
diabetes_read.head()
```

[57]:
```
   Unnamed: 0  Diabetes    BMI State  HighBP  HighChol  CholCheck  \
0           0       0.0  28.17    AL     1.0       1.0        1.0
1           1       0.0  18.54    AL     0.0       0.0        1.0
2           2       1.0  31.62    AL     1.0       0.0        1.0
3           6       1.0  32.98    AL     0.0       0.0        1.0
4           9       1.0  16.65    AL     0.0       1.0        1.0

   FruitConsume  VegetableConsume  Smoker  …  NoDoctorDueToCost  \
0           1.0               1.0     1.0  …                0.0
1           1.0               1.0     0.0  …                0.0
2           1.0               1.0     0.0  …                0.0
3           1.0               1.0     1.0  …                0.0
4           0.0               0.0     1.0  …                0.0

   PhysicalActivity  GeneralHealth  PhysicalHealth  MentalHealth  \
0               0.0            3.0            15.0           0.0
1               1.0            2.0            10.0           0.0
2               1.0            3.0             0.0          30.0
3               1.0            4.0            30.0           0.0
4               0.0            1.0            20.0           0.0

   DifficultyWalking  Gender   Age  Education  Income
0                1.0     0.0  13.0        3.0     3.0
1                0.0     0.0  11.0        5.0     5.0
2                1.0     0.0  10.0        6.0     7.0
3                1.0     1.0  11.0        6.0     7.0
4                1.0     0.0  11.0        2.0     3.0

[5 rows x 24 columns]
```

[58]:
```
diabetes_read.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 243317 entries, 0 to 243316
Data columns (total 24 columns):
 #   Column             Non-Null Count   Dtype
---  ------             --------------   -----
 0   Unnamed: 0         243317 non-null  int64
 1   Diabetes           243317 non-null  float64
 2   BMI                243317 non-null  float64
 3   State              243317 non-null  object
 4   HighBP             243317 non-null  float64
 5   HighChol           243317 non-null  float64
 6   CholCheck          243317 non-null  float64
 7   FruitConsume       243317 non-null  float64
 8   VegetableConsume   243317 non-null  float64
 9   Smoker             243317 non-null  float64
 10  HeavyDrinker       243317 non-null  float64
 11  Stroke             243317 non-null  float64
 12  HeartDisease       243317 non-null  float64
 13  Healthcare         243317 non-null  float64
 14  NoDoctorDueToCost  243317 non-null  float64
 15  PhysicalActivity   243317 non-null  float64
 16  GeneralHealth      243317 non-null  float64
 17  PhysicalHealth     243317 non-null  float64
 18  MentalHealth       243317 non-null  float64
 19  DifficultyWalking  243317 non-null  float64
 20  Gender             243317 non-null  float64
 21  Age                243317 non-null  float64
 22  Education          243317 non-null  float64
 23  Income             243317 non-null  float64
dtypes: float64(22), int64(1), object(1)
memory usage: 44.6+ MB
```

2 —- > NEXT NOTEBOOK : EDA.ipynb

[ ]: