In [1]:
```
import numpy as np
import pandas as pd
```

In [2]:
```
orig_df = pd.read_csv('BRFSS_2020_DATA.csv')
```

Cleaning:

In [3]:
```
orig_df.isna().sum()
```

Out[3]:
```
STATE FIPS CODE
0
FILE MONTH
0
INTERVIEW DATE
0
INTERVIEW MONTH
0
INTERVIEW DAY
0

...
RESPONDENTS AGED 50-75 WHO HAVE HAD A STOOL DNA TEST WITHIN THE PAST THREE YEARS
225601
RESPONDENTS AGED 50-75 WHO HAVE HAD A VIRTUAL COLONOSCOPY WITHIN THE PAST FIVE Y
EARS                                          227684
RESPONDENTS AGED 50-75 WHO HAVE HAD A SIGMOIDOSCOPY WITHIN THE PAST TEN YEARS AN
D A BLOOD STOOL TEST IN THE PAST YEAR    207741
RESPONDENTS AGED 50-75 WHO HAVE FULLY MET THE USPSTF RECOMMENDATIONS
226260
EVER BEEN TESTED FOR HIV CALCULATED VARIABLE
34037
Length: 279, dtype: int64
```

Check columns that will be removed by criteria (must have at least 300k not nulls):

In [4]:
```
temp = (orig_df.isna().sum()>(401958-300000))
list(temp[temp==True].index)
```

Out[4]:
```
['CORRECT TELEPHONE NUMBER?',
 'PRIVATE RESIDENCE?',
 'DO YOU LIVE IN COLLEGE HOUSING?',
 'RESIDENT OF STATE',
 'CELLULAR TELEPHONE',
 'ARE YOU 18 YEARS OF AGE OR OLDER?',
 'ARE YOU MALE OR FEMALE?',
 'NUMBER OF ADULTS IN HOUSEHOLD',
 'ARE YOU MALE OR FEMALE?.1',
 'NUMBER OF ADULT MEN IN HOUSEHOLD',
 'NUMBER OF ADULT WOMEN IN HOUSEHOLD',
 'RESPONDENT SELECTION',
 'SAFE TIME TO TALK?',
 'CORRECT PHONE NUMBER?',
 'IS THIS A CELL PHONE?',
 'ARE YOU 18 YEARS OF AGE OR OLDER?.1',
```

```
 'ARE YOU MALE OR FEMALE?.2',
 'DO YOU LIVE IN A PRIVATE RESIDENCE?',
 'DO YOU LIVE IN COLLEGE HOUSING?.1',
 'DO YOU CURRENTLY LIVE IN  ____(STATE)____?',
 'DO YOU ALSO HAVE A LANDLINE TELEPHONE?',
 'NUMBER OF ADULTS IN HOUSEHOLD.1',
 'POOR PHYSICAL OR MENTAL HEALTH',
 'STILL HAVE ASTHMA',
 'AGE WHEN TOLD DIABETES',
 'HOUSEHOLD TELEPHONES',
 'RESIDENTIAL PHONES',
 'PREGNANCY STATUS',
 'FREQUENCY OF DAYS NOW SMOKING',
 'STOPPED SMOKING IN PAST 12 MONTHS',
 'INTERVAL SINCE LAST SMOKED',
 'AVG ALCOHOLIC DRINKS PER DAY IN PAST 30',
 'BINGE DRINKING',
 'MOST DRINKS ON SINGLE OCCASION PAST 30 DAYS',
 'WHEN RECEIVED MOST RECENT SEASONAL FLU SHOT/SPRAY',
 'HAVE YOU EVER HAD THE SHINGLES OR ZOSTER VACCINE?',
 'HAD FALL PAST TWELVE MONTHS',
 'INJURED IN FALL',
 'DID YOU DRIVE AFTER HAVING TOO MUCH TO DRINK IN THE PAST 30 DAYS?',
 'HAVE YOU EVER HAD A MAMMOGRAM',
 'HOW LONG SINCE LAST MAMMOGRAM',
 'EVER HAD A PAP TEST',
 'HOW LONG SINCE LAST PAP TEST',
 'HAVE YOU EVER HAD AN HPV TEST?',
 'HOW LONG SINCE YOUR LAST HPV TEST?',
 'HAD HYSTERECTOMY',
 'HAS A HEALTH PROFESSIONAL EVER TALKED WITH YOU ABOUT THE ADVANTAGES OF THE PSA
TEST?',
 'HAS A HEALTH PROFESSIONAL EVER TALKED WITH YOU ABOUT THE DISADVANTAGES OF THE
PSA TEST?',
 'HAS A DOCTOR EVER RECOMMENDED THAT YOU HAVE A PSA TEST?',
 'EVER HAD PSA TEST',
 'TIME SINCE LAST PSA TEST',
 'WHAT WAS THE MAIN REASON YOU HAD THIS PSA TEST?',
 'HAVE YOU EVER HAD A COLONOSCOPY?',
 'HOW LONG HAS IT BEEN SINCE YOU HAD COLONOSCOPY?',
 'HAVE YOU EVER HAD A SIGMOIDOSCOPY?',
 'HOW LONG HAS IT BEEN SINCE YOU HAD SIGMOIDOSCOPY?',
 'EVER HAD BLOOD STOOL TEST USING HOME KIT',
 'HOW LONG SINCE YOU HAD BLOOD STOOL TEST?',
 'EVER HAD STOOL DNA TEST?',
 'HOW LONG SINCE YOU HAD STOOL DNA?',
 'HAVE YOU EVER HAD A VIRTUAL COLONOSCOPY?',
 'HOW LONG HAS IT BEEN SINCE YOU HAD VIRTUAL COLONOSCOPY?',
 'MONTH AND YEAR OF LAST HIV TEST',
 'HAD A TEST FOR HIGH BLOOD SUGAR IN PAST THREE YEARS',
 'EVER BEEN TOLD YOU HAVE PRE-DIABETES OR BORDERLINE DIABETES',
 'NOW TAKING INSULIN',
 'HOW OFTEN CHECK BLOOD FOR GLUCOSE',
 'HOW OFTEN CHECK FEET FOR SORES OR IRRITATIONS',
 'TIMES SEEN HEALTH PROFESSIONAL FOR DIABETES',
 'TIMES CHECKED FOR GLYCOSYLATED HEMOGLOBIN',
 'TIMES FEET CHECK FOR SORES/IRRITATIONS',
 'LAST EYE EXAM WHERE PUPILS WERE DILATED',
 'EVER TOLD DIABETES HAS AFFECTED EYES',
 'EVER TAKEN CLASS IN MANAGING DIABETES',
 'TOLD HAD CHRONIC FATIGUE SYNDROME OR MYALGIC ENCEPHALOMYELITIS',
 'STILL HAVE CHRONIC FATIGUE SYNDROME OR MYALGIC ENCEPHALOMYELITIS',
 'HOW MANY HOURS A WEEK ARE YOU BEEN ABLE TO WORK',
 'TOLD HAD HEPATITIS C',
 'TREATED FOR HEPATITIS C',
```

```
 'WERE YOU TREATED FOR HEPATITIS C PRIOR TO 2015',
 'STILL HAVE HEPATITIS C',
 'TOLD  HAD HEPATITIS B',
 'CURRENTLY TAKING MEDICINE FOR HEPATITIS B',
 'PRIMARY HEALTH INSURANCE COVERAGE',
 'HAVE YOU EXPERIENCED CONFUSION OR MEMORY LOSS THAT IS HAPPENING MORE OFTEN OR
IS GETTING WORSE?',
 'GIVEN UP DAY-TO-DAY CHORES DUE TO CONFUSION OR MEMORY LOSS',
 'NEED ASSISTANCE WITH DAY-TO_DAY ACTIVITIES DUE TO CONFUSION OR MEMORY LOSS',
 'WHEN YOU NEED HELP WITH DAY-TO-DAY ACTIVITIES ARE YOU ABLE TO GET IT',
 'DOES CONFUSION OR MEMORY LOSS INTERFERE WITH WORK OR SOCIAL ACTIVITIES',
 'HAVE YOU DISCUSSED YOUR CONFUSION OR MEMORY LOSS WITH A HEALTH CARE PROFESSION
AL?',
 'PROVIDED REGULAR CARE FOR FAMILY OR FRIEND',
 'RELATIONSHIP OF PERSON TO WHOM YOU ARE GIVING CARE?',
 'HOW LONG PROVIDED CARE FOR PERSON.',
 'HOW MANY HOURS DO YOU PROVIDE CARE FOR PERSON?',
 'WHAT IS THE MAJOR HEALTH PROBLEM, ILLNESS, DISABILITY FOR CARE FOR PERSON?',
 'DOES PERSON BEING CARED FOR HAVE ALZHEIMER'S DISEASE?',
 'MANAGED PERSONAL CARE',
 'MANAGED HOUSEHOLD TASKS',
 'DO YOU EXPECT TO HAVE A RELATIVE YOU WILL NEED TO PROVIDE CARE FOR?',
 'HAVE YOU EVER USED AN E-CIGARETTE OR OTHER ELECTRONIC 'VAPING' PRODUCT?',
 'DO YOU NOW USE E-CIGARETTES, EVERY DAY, SOME DAYS, OR NOT AT ALL?',
 'DURING THE PAST 30 DAYS, ON HOW MANY DAYS DID YOU USE MARIJUANA OR HASHISH?',
 'DURING THE PAST 30 DAYS, HOW DID YOU PRIMARILY USE MARIJUANA?',
 'WHAT WAS THE REASON YOU USED MARIJUANA?',
 'HOW OLD WHEN YOU FIRST STARTED SMOKING?',
 'HOW OLD WHEN YOU LAST SMOKED?',
 'ON AVERAGE, HOW MANY CIGARETTES DO YOU SMOKE EACH DAY?',
 'DID YOU HAVE A CT OR CAT SCAN?',
 'HOW MANY TYPES OF CANCER?',
 'AGE TOLD HAD CANCER',
 'TYPE OF CANCER',
 'CURRENTLY RECEIVING TREATMENT FOR CANCER',
 'WHAT TYPE OF DOCTOR PROVIDES MAJORITY OF YOUR CARE',
 'DID YOU RECEIVE A SUMMARY OF CANCER TREATMENTS RECEIVED',
 'EVER RECEIVE INSTRUCTIONS FROM A DOCTOR FOR FOLLOW-UP CHECK-UPS',
 'INSTRUCTIONS WRITTEN OR PRINTED',
 'DID HEALTH INSURANCE PAY FOR ALL OF YOUR CANCER TREATMENT',
 'EVER DENIED INSURANCE COVERAGE BECAUSE OF YOUR CANCER?',
 'PARTICIPATE IN CLINICAL TRIAL AS PART OF CANCER TREATMENT?',
 'CURRENTLY HAVE PHYSICAL PAIN FROM CANCER OR TREATMENT?',
 'IS PAIN UNDER CONTROL?',
 'WHY WAS PSA TEST DONE?',
 'WHO MADE THE DECISION WITH YOU TO HAVE PSA TEST?',
 'EVER HAD AN H.P.V. VACCINATION?',
 'NUMBER OF HPV SHOTS RECEIVED',
 'RECEIVED TETANUS SHOT SINCE 2005?',
 'WHERE DID YOU GET YOUR LAST FLU SHOT/VACCINE?',
 'ARE YOU MALE OR FEMALE?.3',
 'SEXUAL ORIENTATION',
 'SEXUAL ORIENTATION.1',
 'DO YOU CONSIDER YOURSELF TO BE TRANSGENDER?',
 'LIVE WITH ANYONE DEPRESSED, MENTALLY ILL, OR SUICIDAL?',
 'LIVE WITH A PROBLEM DRINKER/ALCOHOLIC?',
 'LIVE WITH ANYONE WHO USED ILLEGAL DRUGS OR ABUSED PRESCRIPTIONS?',
 'LIVE WITH ANYONE WHO SERVED TIME IN PRISON OR JAIL?',
 'WERE YOUR PARENTS DIVORCED/SEPERATED?',
 'HOW OFTEN DID YOUR PARENTS BEAT EACH OTHER UP?',
 'HOW OFTEN DID A PARENT PHYSICALLY HURT YOU IN ANY WAY?',
 'HOW OFTEN DID A PARENT SWEAR AT YOU?',
 'HOW OFTEN DID ANYONE EVER TOUCH YOU SEXUALLY?',
 'HOW OFTEN DID ANYONE MAKE YOU TOUCH THEM SEXUALLY?',
 'HOW OFTEN DID ANYONE EVER FORCE YOU TO HAVE SEX?',
```

```
 'GENDER OF CHILD',
 'RELATIONSHIP TO CHILD',
 'HLTH PRO EVER SAID CHILD HAS ASTHMA',
 'CHILD STILL HAVE ASTHMA?',
 'METROPOLITAN STATUS CODE',
 'CHILD HISPANIC, LATINO/A, OR SPANISH ORIGIN CALCULATED VARIABLE',
 'CHILD NON-HISPANIC RACE INCLUDING MULTIRACIAL',
 'PREFERRED CHILD RACE CATEGORIES',
 'FINAL CHILD WEIGHT: LAND-LINE AND CELL-PHONE DATA',
 'DUAL PHONE USE CORRECTION FACTOR',
 'ADULTS AGED 65+ WHO HAVE HAD ALL THEIR NATURAL TEETH EXTRACTED',
 'FLU SHOT CALCULATED VARIABLE',
 'PNEUMONIA VACCINATION CALCULATED VARIABLE',
 'WOMEN RESPONDENTS AGED 40+ WHO HAVE HAD A MAMMOGRAM IN THE PAST TWO YEARS',
 'WOMEN RESPONDENTS AGED 50-74 WHO HAVE HAD A MAMMOGRAM IN THE PAST TWO YEARS',
 'WOMEN RESPONDENTS AGED 21-65 WHO HAVE HAD A PAP TEST IN THE PAST THREE YEARS',
 'MALE RESPONDENTS AGED 40+ WHO HAVE HAD A PSA TEST IN THE PAST 2 YEARS',
 'RESPONDENTS AGED 50-75 WHO HAVE HAD A COLONOSCOPY WITHIN THE PAST TEN YEARS',
 'RESPONDENTS AGED 50-75 WHO HAVE HAD A SIGMOIDOSCOPY WITHIN THE PAST FIVE YEAR
S',
 'RESPONDENTS AGED 50-75 WHO HAVE HAD A SIGMOIDOSCOPY WITHIN THE PAST TEN YEAR
S',
 'RESPONDENTS AGED 50-75 WHO HAVE HAD A BLOOD STOOL TEST WITHIN THE PAST YEAR',
 'RESPONDENTS AGED 50-75 WHO HAVE HAD A STOOL DNA TEST WITHIN THE PAST THREE YEA
RS',
 'RESPONDENTS AGED 50-75 WHO HAVE HAD A VIRTUAL COLONOSCOPY WITHIN THE PAST FIVE
YEARS',
 'RESPONDENTS AGED 50-75 WHO HAVE HAD A SIGMOIDOSCOPY WITHIN THE PAST TEN YEARS
AND A BLOOD STOOL TEST IN THE PAST YEAR',
 'RESPONDENTS AGED 50-75 WHO HAVE FULLY MET THE USPSTF RECOMMENDATIONS']
```

Check few potential columns from the list above. May be we can work with them?

In [5]:
```
orig_df[[
'EVER BEEN TOLD YOU HAVE PRE-DIABETES OR BORDERLINE DIABETES',
'TOLD HAD HEPATITIS C',
'TOLD  HAD HEPATITIS B',
'HOW OLD WHEN YOU FIRST STARTED SMOKING?',
'HOW OLD WHEN YOU LAST SMOKED?',
'ON AVERAGE, HOW MANY CIGARETTES DO YOU SMOKE EACH DAY?',
]].isna().sum()
```

Out[5]:
```
EVER BEEN TOLD YOU HAVE PRE-DIABETES OR BORDERLINE DIABETES    182965
TOLD HAD HEPATITIS C                                           383151
TOLD  HAD HEPATITIS B                                          383186
HOW OLD WHEN YOU FIRST STARTED SMOKING?                        387914
HOW OLD WHEN YOU LAST SMOKED?                                  388332
ON AVERAGE, HOW MANY CIGARETTES DO YOU SMOKE EACH DAY?         388351
dtype: int64
```

To preserve most of the rows, we'll remove all columns with that threshhold:

In [6]:
```
clean_df_1 = orig_df.dropna(axis=1, thresh=300000)
clean_df_1.shape
```

Out[6]:
```
(401958, 112)
```

In [7]:
```
list(clean_df_1.columns)
```

Out[7]:
```
['STATE FIPS CODE',
 'FILE MONTH',
 'INTERVIEW DATE',
 'INTERVIEW MONTH',
 'INTERVIEW DAY',
 'INTERVIEW YEAR',
 'FINAL DISPOSITION',
 'ANNUAL SEQUENCE NUMBER',
 'PRIMARY SAMPLING UNIT',
 'SEX OF RESPONDENT',
 'GENERAL HEALTH',
 'NUMBER OF DAYS PHYSICAL HEALTH NOT GOOD',
 'NUMBER OF DAYS MENTAL HEALTH NOT GOOD',
 'HAVE ANY HEALTH CARE COVERAGE',
 'MULTIPLE HEALTH CARE PROFESSIONALS',
 'COULD NOT SEE DR. BECAUSE OF COST',
 'LENGTH OF TIME SINCE LAST ROUTINE CHECKUP',
 'EXERCISE IN PAST 30 DAYS',
 'HOW MUCH TIME DO YOU SLEEP',
 'EVER DIAGNOSED WITH HEART ATTACK',
 'EVER DIAGNOSED WITH ANGINA OR CORONARY HEART DISEASE',
 'EVER DIAGNOSED WITH A STROKE',
 'EVER TOLD HAD ASTHMA',
 '(EVER TOLD) YOU HAD SKIN CANCER?',
 '(EVER TOLD) YOU HAD ANY OTHER TYPES OF CANCER?',
 '(EVER TOLD) YOU HAD (COPD) CHRONIC OBSTRUCTIVE PULMONARY DISEASE, EMPHYSEMA OR
CHRONIC BRONCHITIS?',
 'TOLD HAVE ARTHRITIS',
 '(EVER TOLD) YOU HAD A DEPRESSIVE DISORDER',
 'EVER TOLD YOU HAVE KIDNEY DISEASE?',
 '(EVER TOLD) YOU HAD DIABETES',
 'LAST VISITED DENTIST OR DENTAL CLINIC',
 'NUMBER OF PERMANENT TEETH REMOVED',
 'MARITAL STATUS',
 'EDUCATION LEVEL',
 'OWN OR RENT HOME',
 'DO YOU HAVE A CELL PHONE FOR PERSONAL USE?',
 'ARE YOU A VETERAN',
 'EMPLOYMENT STATUS',
 'NUMBER OF CHILDREN IN HOUSEHOLD',
 'INCOME LEVEL',
 'REPORTED WEIGHT IN POUNDS',
 'REPORTED HEIGHT IN FEET AND INCHES',
 'ARE YOU DEAF OR DO YOU HAVE SERIOUS DIFFICULTY HEARING?',
 'BLIND OR DIFFICULTY SEEING',
 'DIFFICULTY CONCENTRATING OR REMEMBERING',
 'DIFFICULTY WALKING OR CLIMBING STAIRS',
 'DIFFICULTY DRESSING OR BATHING',
 'DIFFICULTY DOING ERRANDS ALONE',
 'SMOKED AT LEAST 100 CIGARETTES',
 'USE OF SMOKELESS TOBACCO PRODUCTS',
 'DAYS IN PAST 30 HAD ALCOHOLIC BEVERAGE',
 'ADULT FLU SHOT/SPRAY PAST 12 MOS',
 'PNEUMONIA SHOT EVER',
 'HOW OFTEN USE SEATBELTS IN CAR?',
 'EVER TESTED H.I.V.',
 'DO ANY HIGH RISK SITUATIONS APPLY',
 'QUESTIONNAIRE VERSION IDENTIFIER',
 'LANGUAGE IDENTIFIER',
 'METROPOLITAN STATUS',
 'URBAN/RURAL STATUS',
 'SAMPLE DESIGN STRATIFICATION VARIABLE',
 'STRATUM WEIGHT',
 'RAW WEIGHTING FACTOR USED IN RAKING',
 'DESIGN WEIGHT USED IN RAKING',
```

```
 'IMPUTED RACE/ETHNICITY VALUE',
 'DUAL PHONE USE CATEGORIES',
 'TRUNCATED DESIGN WEIGHT USED IN ADULT COMBINED LAN  LINE AND CELL PHONE RAKIN
G',
 'FINAL WEIGHT: LAND-LINE AND CELL-PHONE DATA',
 'ADULTS WITH GOOD OR BETTER HEALTH',
 'COMPUTED PHYSICAL HEALTH STATUS',
 'COMPUTED MENTAL HEALTH STATUS',
 'RESPONDENTS AGED 18-64 WITH HEALTH CARE COVERAGE',
 'LEISURE TIME PHYSICAL ACTIVITY CALCULATED VARIABLE',
 'RESPONDENTS THAT HAVE EVER REPORTED HAVING CORONARY HEART DISEASE (CHD) OR MYO
CARDIAL INFARCTION (MI)',
 'LIFETIME ASTHMA CALCULATED VARIABLE',
 'CURRENT ASTHMA CALCULATED VARIABLE',
 'COMPUTED ASTHMA STATUS',
 'RESPONDENTS DIAGNOSED WITH ARTHRITIS',
 'RISK FACTOR FOR HAVING HAD PERMANENT TEETH EXTRACTED',
 'ADULTS WHO HAVE VISITED A DENTIST, DENTAL HYGENIST OR DENTAL CLINIC WITHIN THE
PAST YEAR',
 'COMPUTED PREFERRED RACE',
 'CALCULATED NON-HISPANIC RACE INCLUDING MULTIRACIAL',
 'HISPANIC, LATINO/A, OR SPANISH ORIGIN CALCULATED VARIABLE',
 'COMPUTED RACE-ETHNICITY GROUPING',
 'COMPUTED NON-HISPANIC WHITES/ALL OTHERS RACE CATEGORIES RACE/ETHNIC GROUP CODE
S USED IN POST-STRATIFICATION.',
 'COMPUTED FIVE LEVEL RACE/ETHNICITY CATEGORY.',
 'COMPUTED RACE GROUPS USED FOR INTERNET PREVALENCE TABLES',
 'CALCULATED SEX VARIABLE',
 'REPORTED AGE IN FIVE-YEAR AGE CATEGORIES CALCULATED VARIABLE',
 'REPORTED AGE IN TWO AGE GROUPS CALCULATED VARIABLE',
 'IMPUTED AGE VALUE COLLAPSED ABOVE 80',
 'IMPUTED AGE IN SIX GROUPS',
 'COMPUTED HEIGHT IN INCHES',
 'COMPUTED HEIGHT IN METERS',
 'COMPUTED WEIGHT IN KILOGRAMS',
 'COMPUTED BODY MASS INDEX',
 'COMPUTED BODY MASS INDEX CATEGORIES',
 'OVERWEIGHT OR OBESE CALCULATED VARIABLE',
 'COMPUTED NUMBER OF CHILDREN IN HOUSEHOLD',
 'COMPUTED LEVEL OF EDUCATION COMPLETED CATEGORIES',
 'COMPUTED INCOME CATEGORIES',
 'COMPUTED SMOKING STATUS',
 'CURRENT SMOKING CALCULATED VARIABLE',
 'DRINK ANY ALCOHOLIC BEVERAGES IN PAST 30 DAYS',
 'COMPUTED DRINK-OCCASIONS-PER-DAY',
 'BINGE DRINKING CALCULATED VARIABLE',
 'COMPUTED NUMBER OF DRINKS OF ALCOHOL BEVERAGES PER WEEK',
 'HEAVY ALCOHOL CONSUMPTION  CALCULATED VARIABLE',
 'ALWAYS OR NEARLY ALWAYS WEAR SEAT BELTS',
 'ALWAYS WEAR SEAT BELTS',
 'DRINKING AND DRIVING',
 'EVER BEEN TESTED FOR HIV CALCULATED VARIABLE']
```

In [8]:
```
clean_df_1.isna().sum()
```

Out[8]:
```
STATE FIPS CODE                                    0
FILE MONTH                                         0
INTERVIEW DATE                                     0
INTERVIEW MONTH                                    0
INTERVIEW DAY                                      0
                                                ...
HEAVY ALCOHOL CONSUMPTION  CALCULATED VARIABLE     0
ALWAYS OR NEARLY ALWAYS WEAR SEAT BELTS            0
ALWAYS WEAR SEAT BELTS                             0
```

```
DRINKING AND DRIVING                                    0
EVER BEEN TESTED FOR HIV CALCULATED VARIABLE        34037
Length: 112, dtype: int64
```

Looks like these columns are not very inetersting to us, we'll remove them too:

In [9]:

```
columns_to_remove = [
 'STATE FIPS CODE',
 'FILE MONTH',
 'INTERVIEW DATE',
 'INTERVIEW MONTH',
 'INTERVIEW DAY',
 'INTERVIEW YEAR',
 'FINAL DISPOSITION',
 'ANNUAL SEQUENCE NUMBER',
 'PRIMARY SAMPLING UNIT',
 'HAVE ANY HEALTH CARE COVERAGE',
 'MULTIPLE HEALTH CARE PROFESSIONALS',
 'COULD NOT SEE DR. BECAUSE OF COST',
 'LAST VISITED DENTIST OR DENTAL CLINIC',
 'NUMBER OF PERMANENT TEETH REMOVED',
 'MARITAL STATUS',
 'EDUCATION LEVEL',
 'OWN OR RENT HOME',
 'DO YOU HAVE A CELL PHONE FOR PERSONAL USE?',
 'ARE YOU A VETERAN',
 'EMPLOYMENT STATUS',
 'NUMBER OF CHILDREN IN HOUSEHOLD',
 'INCOME LEVEL',
 'ARE YOU DEAF OR DO YOU HAVE SERIOUS DIFFICULTY HEARING?',
 'BLIND OR DIFFICULTY SEEING',
 'DIFFICULTY CONCENTRATING OR REMEMBERING',
 'DIFFICULTY WALKING OR CLIMBING STAIRS',
 'DIFFICULTY DRESSING OR BATHING',
 'DIFFICULTY DOING ERRANDS ALONE',
 'ADULT FLU SHOT/SPRAY PAST 12 MOS',
 'PNEUMONIA SHOT EVER',
 'HOW OFTEN USE SEATBELTS IN CAR?',
 'EVER TESTED H.I.V.',
 'DO ANY HIGH RISK SITUATIONS APPLY',
 'QUESTIONNAIRE VERSION IDENTIFIER',
 'LANGUAGE IDENTIFIER',
 'METROPOLITAN STATUS',
 'URBAN/RURAL STATUS',
 'SAMPLE DESIGN STRATIFICATION VARIABLE',
 'STRATUM WEIGHT',
 'RAW WEIGHTING FACTOR USED IN RAKING',
 'DESIGN WEIGHT USED IN RAKING',
 'IMPUTED RACE/ETHNICITY VALUE',
 'DUAL PHONE USE CATEGORIES',
 'TRUNCATED DESIGN WEIGHT USED IN ADULT COMBINED LAN  LINE AND CELL PHONE RAKING
 'FINAL WEIGHT: LAND-LINE AND CELL-PHONE DATA',
 'RESPONDENTS AGED 18-64 WITH HEALTH CARE COVERAGE',
 'RISK FACTOR FOR HAVING HAD PERMANENT TEETH EXTRACTED',
 'ADULTS WHO HAVE VISITED A DENTIST, DENTAL HYGENIST OR DENTAL CLINIC WITHIN THE
 'COMPUTED PREFERRED RACE',
 'CALCULATED NON-HISPANIC RACE INCLUDING MULTIRACIAL',
 'HISPANIC, LATINO/A, OR SPANISH ORIGIN CALCULATED VARIABLE',
 'COMPUTED RACE-ETHNICITY GROUPING',
 'COMPUTED NON-HISPANIC WHITES/ALL OTHERS RACE CATEGORIES RACE/ETHNIC GROUP CODE
```

```
 'COMPUTED FIVE LEVEL RACE/ETHNICITY CATEGORY.',
 'COMPUTED RACE GROUPS USED FOR INTERNET PREVALENCE TABLES',
 'CALCULATED SEX VARIABLE',
 'COMPUTED HEIGHT IN INCHES',
 'COMPUTED HEIGHT IN METERS',
 'COMPUTED WEIGHT IN KILOGRAMS',
 'COMPUTED NUMBER OF CHILDREN IN HOUSEHOLD',
 'COMPUTED LEVEL OF EDUCATION COMPLETED CATEGORIES',
 'COMPUTED INCOME CATEGORIES',
 'CURRENT SMOKING CALCULATED VARIABLE',
 'ALWAYS OR NEARLY ALWAYS WEAR SEAT BELTS',
 'ALWAYS WEAR SEAT BELTS',
 'DRINKING AND DRIVING'
]
```

In [10]:
```
clean_df_2 = clean_df_1.drop(columns_to_remove, axis=1)
clean_df_2.shape
```

Out[10]:
```
(401958, 46)
```

In [11]:
```
clean_df_2.isna().sum()
```

Out[11]:
```
SEX OF RESPONDENT
0
GENERAL HEALTH
8
NUMBER OF DAYS PHYSICAL HEALTH NOT GOOD
5
NUMBER OF DAYS MENTAL HEALTH NOT GOOD
5
LENGTH OF TIME SINCE LAST ROUTINE CHECKUP
5
EXERCISE IN PAST 30 DAYS
3
HOW MUCH TIME DO YOU SLEEP
3
EVER DIAGNOSED WITH HEART ATTACK
6
EVER DIAGNOSED WITH ANGINA OR CORONARY HEART DISEASE
3
EVER DIAGNOSED WITH A STROKE
3
EVER TOLD HAD ASTHMA
3
(EVER TOLD) YOU HAD SKIN CANCER?
3
(EVER TOLD) YOU HAD ANY OTHER TYPES OF CANCER?
3
(EVER TOLD) YOU HAD (COPD) CHRONIC OBSTRUCTIVE PULMONARY DISEASE, EMPHYSEMA OR C
HRONIC BRONCHITIS?           5
TOLD HAVE ARTHRITIS
5
(EVER TOLD) YOU HAD A DEPRESSIVE DISORDER
6
EVER TOLD YOU HAVE KIDNEY DISEASE?
6
(EVER TOLD) YOU HAD DIABETES
6
REPORTED WEIGHT IN POUNDS
9852
```

```
REPORTED HEIGHT IN FEET AND INCHES
10824
SMOKED AT LEAST 100 CIGARETTES
17860
USE OF SMOKELESS TOBACCO PRODUCTS
18493
DAYS IN PAST 30 HAD ALCOHOLIC BEVERAGE
20927
ADULTS WITH GOOD OR BETTER HEALTH
0
COMPUTED PHYSICAL HEALTH STATUS
0
COMPUTED MENTAL HEALTH STATUS
0
LEISURE TIME PHYSICAL ACTIVITY CALCULATED VARIABLE
0
RESPONDENTS THAT HAVE EVER REPORTED HAVING CORONARY HEART DISEASE (CHD) OR MYOCA
RDIAL INFARCTION (MI)     3571
LIFETIME ASTHMA CALCULATED VARIABLE
0
CURRENT ASTHMA CALCULATED VARIABLE
0
COMPUTED ASTHMA STATUS
0
RESPONDENTS DIAGNOSED WITH ARTHRITIS
2303
REPORTED AGE IN FIVE-YEAR AGE CATEGORIES CALCULATED VARIABLE
0
REPORTED AGE IN TWO AGE GROUPS CALCULATED VARIABLE
0
IMPUTED AGE VALUE COLLAPSED ABOVE 80
0
IMPUTED AGE IN SIX GROUPS
0
COMPUTED BODY MASS INDEX
41357
COMPUTED BODY MASS INDEX CATEGORIES
41357
OVERWEIGHT OR OBESE CALCULATED VARIABLE
0
COMPUTED SMOKING STATUS
0
DRINK ANY ALCOHOLIC BEVERAGES IN PAST 30 DAYS
0
COMPUTED DRINK-OCCASIONS-PER-DAY
0
BINGE DRINKING CALCULATED VARIABLE
0
COMPUTED NUMBER OF DRINKS OF ALCOHOL BEVERAGES PER WEEK
0
HEAVY ALCOHOL CONSUMPTION  CALCULATED VARIABLE
0
EVER BEEN TESTED FOR HIV CALCULATED VARIABLE
34037
dtype: int64
```

Remove rows with nulls:

In [12]:
```
clean_df_3 = clean_df_2.dropna()
clean_df_3.shape
```

Out[12]:
```
(336836, 46)
```

We are left with those features:

In [13]:
```
list(clean_df_3.columns)
```

Out[13]:
```
['SEX OF RESPONDENT',
 'GENERAL HEALTH',
 'NUMBER OF DAYS PHYSICAL HEALTH NOT GOOD',
 'NUMBER OF DAYS MENTAL HEALTH NOT GOOD',
 'LENGTH OF TIME SINCE LAST ROUTINE CHECKUP',
 'EXERCISE IN PAST 30 DAYS',
 'HOW MUCH TIME DO YOU SLEEP',
 'EVER DIAGNOSED WITH HEART ATTACK',
 'EVER DIAGNOSED WITH ANGINA OR CORONARY HEART DISEASE',
 'EVER DIAGNOSED WITH A STROKE',
 'EVER TOLD HAD ASTHMA',
 '(EVER TOLD) YOU HAD SKIN CANCER?',
 '(EVER TOLD) YOU HAD ANY OTHER TYPES OF CANCER?',
 '(EVER TOLD) YOU HAD (COPD) CHRONIC OBSTRUCTIVE PULMONARY DISEASE, EMPHYSEMA OR
CHRONIC BRONCHITIS?',
 'TOLD HAVE ARTHRITIS',
 '(EVER TOLD) YOU HAD A DEPRESSIVE DISORDER',
 'EVER TOLD YOU HAVE KIDNEY DISEASE?',
 '(EVER TOLD) YOU HAD DIABETES',
 'REPORTED WEIGHT IN POUNDS',
 'REPORTED HEIGHT IN FEET AND INCHES',
 'SMOKED AT LEAST 100 CIGARETTES',
 'USE OF SMOKELESS TOBACCO PRODUCTS',
 'DAYS IN PAST 30 HAD ALCOHOLIC BEVERAGE',
 'ADULTS WITH GOOD OR BETTER HEALTH',
 'COMPUTED PHYSICAL HEALTH STATUS',
 'COMPUTED MENTAL HEALTH STATUS',
 'LEISURE TIME PHYSICAL ACTIVITY CALCULATED VARIABLE',
 'RESPONDENTS THAT HAVE EVER REPORTED HAVING CORONARY HEART DISEASE (CHD) OR MYO
CARDIAL INFARCTION (MI)',
 'LIFETIME ASTHMA CALCULATED VARIABLE',
 'CURRENT ASTHMA CALCULATED VARIABLE',
 'COMPUTED ASTHMA STATUS',
 'RESPONDENTS DIAGNOSED WITH ARTHRITIS',
 'REPORTED AGE IN FIVE-YEAR AGE CATEGORIES CALCULATED VARIABLE',
 'REPORTED AGE IN TWO AGE GROUPS CALCULATED VARIABLE',
 'IMPUTED AGE VALUE COLLAPSED ABOVE 80',
 'IMPUTED AGE IN SIX GROUPS',
 'COMPUTED BODY MASS INDEX',
 'COMPUTED BODY MASS INDEX CATEGORIES',
 'OVERWEIGHT OR OBESE CALCULATED VARIABLE',
 'COMPUTED SMOKING STATUS',
 'DRINK ANY ALCOHOLIC BEVERAGES IN PAST 30 DAYS',
 'COMPUTED DRINK-OCCASIONS-PER-DAY',
 'BINGE DRINKING CALCULATED VARIABLE',
 'COMPUTED NUMBER OF DRINKS OF ALCOHOL BEVERAGES PER WEEK',
 'HEAVY ALCOHOL CONSUMPTION  CALCULATED VARIABLE',
 'EVER BEEN TESTED FOR HIV CALCULATED VARIABLE']
```

In [14]:
```
clean_df_3.head(5)
```

Out[14]:

SEX OF RESPONDENT	GENERAL HEALTH	NUMBER OF DAYS PHYSICAL HEALTH NOT GOOD	NUMBER OF DAYS MENTAL HEALTH NOT GOOD	LENGTH OF TIME SINCE LAST ROUTINE CHECKUP	EXERCISE IN PAST 30 DAYS	HOW MUCH TIME DO YOU SLEEP	EVER DIAGNOSED WITH HEART ATTACK	DIA AN CC

	SEX OF RESPONDENT	GENERAL HEALTH	NUMBER OF DAYS PHYSICAL HEALTH NOT GOOD	NUMBER OF DAYS MENTAL HEALTH NOT GOOD	LENGTH OF TIME SINCE LAST ROUTINE CHECKUP	EXERCISE IN PAST 30 DAYS	HOW MUCH TIME DO YOU SLEEP	EVER DIAGNOSED WITH HEART ATTACK	DIA AN CC
0	2	2.0	3.0	30.0	4.0	1.0	5.0	2.0	
4	2	2.0	88.0	88.0	1.0	1.0	7.0	2.0	
5	1	4.0	20.0	30.0	2.0	1.0	8.0	2.0	
6	2	3.0	88.0	88.0	1.0	2.0	6.0	2.0	
8	2	2.0	28.0	88.0	1.0	1.0	8.0	2.0	

5 rows × 46 columns

In [ ]: