Research question 1.

Can health, demographical, and behavioural factors of non-institutionalized adults in the US, impact on predicting the Coronary heart diseases (CHD)?

1) Selected health, demographical, and behavioural factors.

```
# Research Question 1
#Can health, demographical, and behavioural factors of non-institutionalized adults in the US, impact on
#predicting the Coronary Heart Diseases (CHD)?

#_RFCHOL - Calculated variable for adults who have had their cholesterol checked and have been told by a doctor, nurse, or other
#_RFHYPE5 - Calculated variable for adults who have been told they have high blood pressure by a doctor, nurse, or other health p
#_SMOKER3 - Calculated variable for four-level smoker status: everyday smoker, someday smoker, former smoker, non-smoker
#DIABETE3 - Has a doctor, nurse, or other health professional ever told you have diabetes?
#_AGEG5YR - Calculated variable for fourteen-level age category.
#FTJUDA1_ - Calculated variable for fruit juice intake in times per day.
#BEANDAY_ - Calculated variable for bean intake in times per day
#GRENDAY_ - Calculated variable for dark green vegetable intake in times per day
#ORNGDAY_ - Calculated variable for orange-colored vegetable intake in times per day
#VEGEDA1_ - Calculated variable for vegetable intake in times per day
#_TOTINDA - Calculated variable for adults who reported doing physical activity or exercise during the past 30 days other than th
#DRNKANY5 -Calculated variable for adults who reported having had at least one drink of alcohol in the past 30 days.
#_RFBING5 - Calculated variable for binge drinkers (males having five or more drinks on one occasion, females having four or more
#_RFDRHV5 -Calculated variable for heavy drinkers (adult men having more than 14 drinks per week and adult women having more than
#_MICHD - Calculated variable for respondents that have ever reported having coronary heart disease

HeartDiseaseDataset = Survey[['_RFCHOL', '_RFHYPE5', '_SMOKER3','DIABETE3', '_AGEG5YR', 'FTJUDA1_' , 'BEANDAY_' , 'GRENDAY_' , '(

HeartDiseaseDataset = HeartDiseaseDataset.rename({'_RFCHOL': 'Colestrol', '_RFHYPE5': 'High Blood Pressure' ,'_SMOKER3': 'Smoker

HeartDiseaseDataset = HeartDiseaseDataset.rename({'FTJUDA1_': 'Fruit juice intake', 'BEANDAY_': 'Bean intake' ,'GRENDAY_': 'Dark


HeartDiseaseDataset.info()
```

2) The column description, including not null count of each column and its data type.

```
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   Colestrol                   12338 non-null  object
 1   High Blood Pressure         12338 non-null  object
 2   Smoker Status               12338 non-null  object
 3   DIABETE3                    12338 non-null  object
 4   Age Category                12338 non-null  object
 5   Fruit juice intake          11052 non-null  float64
 6   Bean intake                 11028 non-null  float64
 7   Dark green vegetable intake 11042 non-null  float64
 8   Orange color vege. intake   11022 non-null  float64
 9   Vegetable intake            10916 non-null  float64
 10  Regular Exercise            12338 non-null  object
 11  Drunker status              12338 non-null  object
 12  Being Drinkers              12338 non-null  object
 13  Heavy Drinkers              12338 non-null  int64
 14  Heart Disease               12338 non-null  object
dtypes: float64(5), int64(1), object(9)
memory usage: 1.4+ MB
```

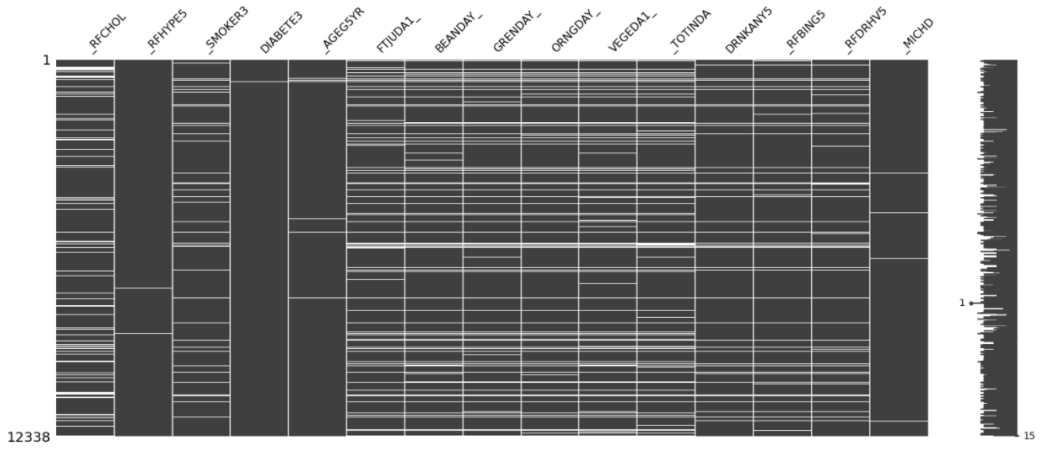
3) New not Null count after imputing NaN values to missing values.

```
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   Colestrol                    10763 non-null  object
 1   High Blood Pressure          12294 non-null  object
 2   Smoker Status                11666 non-null  object
 3   DIABETE3                     12325 non-null  object
 4   Age Category                 12141 non-null  object
 5   Fruit juice intake           11052 non-null  float64
 6   Bean intake                  11028 non-null  float64
 7   Dark green vegetable intake  11042 non-null  float64
 8   Orange color vege. intake    11022 non-null  float64
 9   Vegetable intake             10916 non-null  float64
 10  Regular Exercise             10955 non-null  object
 11  Drunker status               11528 non-null  object
 12  Being Drinkers               11385 non-null  object
 13  Heavy Drinkers               12338 non-null  int64
 14  Heart Disease                12239 non-null  object
dtypes: float64(5), int64(1), object(9)
memory usage: 1.4+ MB
```

4) Number of null values of each column.

```
Colestrol                      1575
High Blood Pressure              44
Smoker Status                   672
DIABETE3                         13
Age Category                    197
Fruit juice intake             1286
Bean intake                    1310
Dark green vegetable intake    1296
Orange color vege. intake      1316
Vegetable intake               1422
Regular Exercise               1383
Drunker status                  810
Being Drinkers                  953
Heavy Drinkers                    0
Heart Disease                    99
dtype: int64
```

5) Null Value representation.



6) Null Values percentage of each column.

```
Colestrol                      12.67
Vegetable intake               11.45
Regular Exercise               11.15
Orange color vege. intake      10.60
Bean intake                    10.52
Dark green vegetable intake    10.44
Fruit juice intake             10.32
Being Drinkers                  7.66
Drunker status                  6.51
Smoker Status                   5.39
Age Category                    1.55
High Blood Pressure             0.33
DIABETE3                        0.07
Heart Disease                   0.00
Heavy Drinkers                  0.00
dtype: float64
```

7) Null percentage after imputing the categorical values with mode and numerical columns with 'interpolate' method.

```
Heart Disease                  0.0
Heavy Drinkers                 0.0
Being Drinkers                 0.0
Drunker status                 0.0
Regular Exercise               0.0
Vegetable intake               0.0
Orange color vege. intake      0.0
Dark green vegetable intake    0.0
Bean intake                    0.0
Fruit juice intake             0.0
Age Category                   0.0
DIABETE3                       0.0
Smoker Status                  0.0
High Blood Pressure            0.0
Colestrol                      0.0
```

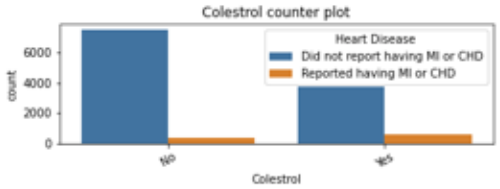
8) The statistical measurements of numerical columns.

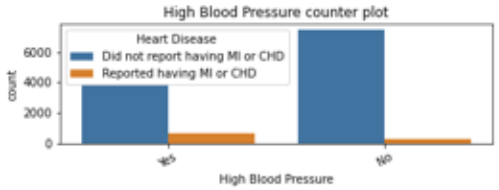
	Fruit juice intake	Bean intake	Dark green vegetable intake	Orange color vege. intake	Vegetable intake	Heavy Drinkers
count	12239.000000	12239.00000	12239.000000	12239.00000	12239.000000	12239.000000
mean	42.601642	27.12881	63.536727	31.32568	77.132936	1.658796
std	68.082406	42.41291	59.162108	38.02846	62.603325	2.115614
min	0.000000	0.00000	0.000000	0.00000	0.000000	1.000000
25%	0.000000	3.00000	28.500000	8.50000	33.000000	1.000000
50%	14.000000	14.00000	50.000000	17.00000	67.000000	1.000000
75%	71.000000	33.00000	100.000000	43.00000	100.000000	1.000000
max	2500.000000	2200.00000	800.000000	500.00000	700.000000	9.000000

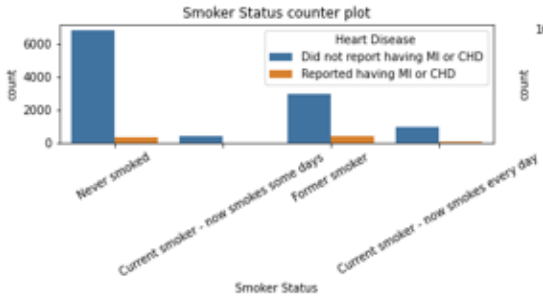
9) The counter plots of the categorical variable.

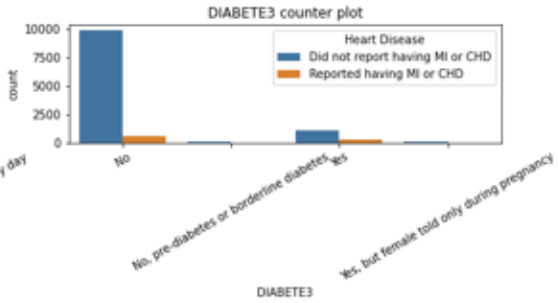
According to the following figure, having cholesterol had been reasoned to have coronary heart diseases (CHD) of adults. High blood pressure has been a reason to having CHD. When checked the smoker

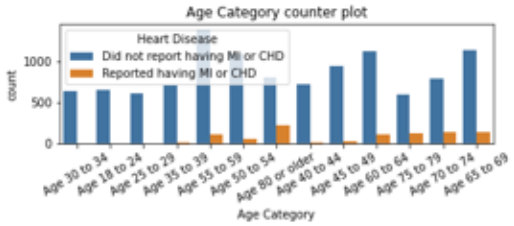
status, former smokers have reported a higher CHD amount, when compared to never or current smokers. Diabetes seems not a very good attribute to predict CHD. People who have not reported diabetes have reported the highest CHD amount. Age seems a good indication to predict the heart disease. People above age 35 had started to report CHD. Having physical activities seems not a good indication to predict CHD. People who do regular physical activities have reported higher rate of CHD than those who do not do regular physical activities. People who are not binge drinkers are mostly not CHD patients.

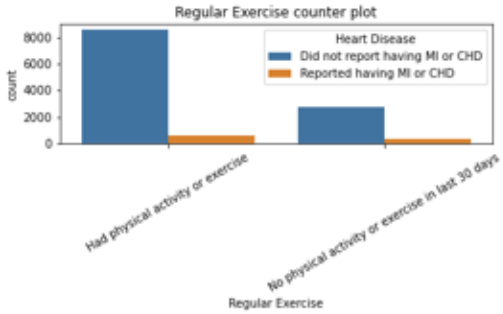

Colestrol counter plot
Heart Disease
Did not report having MI or CHD
Reported having MI or CHD
count
0
2000
4000
6000
No
Yes
Colestrol


High Blood Pressure counter plot
Heart Disease
Did not report having MI or CHD
Reported having MI or CHD
count
0
2000
4000
6000
Yes
No
High Blood Pressure


Smoker Status counter plot
Heart Disease
Did not report having MI or CHD
Reported having MI or CHD
count
0
2000
4000
6000
Never smoked
Current smoker - now smokes some days
Former smoker
Current smoker - now smokes every day
Smoker Status

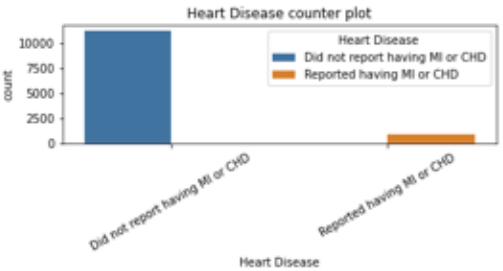

DIABETE3 counter plot
Heart Disease
Did not report having MI or CHD
Reported having MI or CHD
count
0
2500
5000
7500
10000
No
No, pre-diabetes or borderline diabetes
Yes
Yes, but female told only during pregnancy
DIABETE3


Age Category counter plot
Heart Disease
Did not report having MI or CHD
Reported having MI or CHD
count
0
500
1000
Age 30 to 34
Age 18 to 24
Age 25 to 29
Age 35 to 39
Age 55 to 59
Age 50 to 54
Age 80 or older
Age 40 to 44
Age 45 to 49
Age 60 to 64
Age 75 to 79
Age 70 to 74
Age 65 to 69
Age Category


Regular Exercise counter plot
Heart Disease
Did not report having MI or CHD
Reported having MI or CHD
count
0
2000
4000
6000
8000
Had physical activity or exercise
No physical activity or exercise in last 30 days
Regular Exercise


Drunker status counter plot
Heart Disease
Did not report having MI or CHD
Reported having MI or CHD
count
0
2000
4000
6000
Yes
No
Drunker status

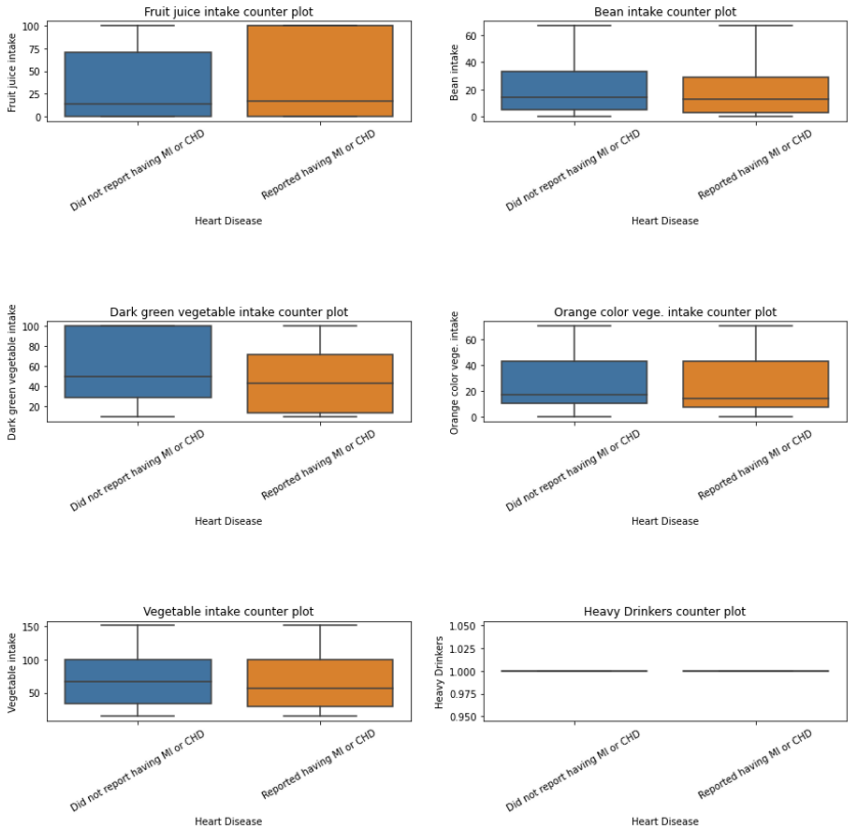

Being Drinkers counter plot
Heart Disease
Did not report having MI or CHD
Reported having MI or CHD
count
0
2500
5000
7500
10000
Yes
No
Being Drinkers


Heart Disease counter plot
Heart Disease
Did not report having MI or CHD
Reported having MI or CHD
count
0
2500
5000
7500
10000
Did not report having MI or CHD
Reported having MI or CHD
Heart Disease

10) Removed the outliers of the numerical attributes. The outliers were imputed with minimum and maximum floor values of each column.

```
#plt.figure(figsize=(10,10))
#h=sns.boxplot(data=HeartDiseaseDataset_numerical)
#h.tick_params(axis='x', rotation=30)

#  Categorical Data
a = 5  # number of rows
b = 2  # number of columns
c = 1  # initialize plot counter

fig = plt.figure(figsize=(15,25))
plt.subplots_adjust(hspace = 2)

for col in HeartDiseaseDataset:
    if (not is_string_dtype(HeartDiseaseDataset[col])):
        lower=HeartDiseaseDataset[col].quantile(0.10)
        max=HeartDiseaseDataset[col].quantile(0.90)
        HeartDiseaseDataset[col] = np.where(HeartDiseaseDataset[col] <lower, lower,HeartDiseaseDataset[col])
        HeartDiseaseDataset[col] = np.where(HeartDiseaseDataset[col] >max, max,HeartDiseaseDataset[col])
        plt.subplot(a, b, c)
        plt.title('{} counter plot'.format(HeartDiseaseDataset[col].name))
        plt.xlabel(HeartDiseaseDataset[col].name)
        #Replace missing values from the mode in categorical variables
        g=sns.boxplot(x='Heart Disease',y=HeartDiseaseDataset[col], data=HeartDiseaseDataset)
        g.tick_params(axis='x', rotation=30)
        c = c + 1

plt.show()
```

11) The boxplots of numerical attributes used to heart disease prediction.

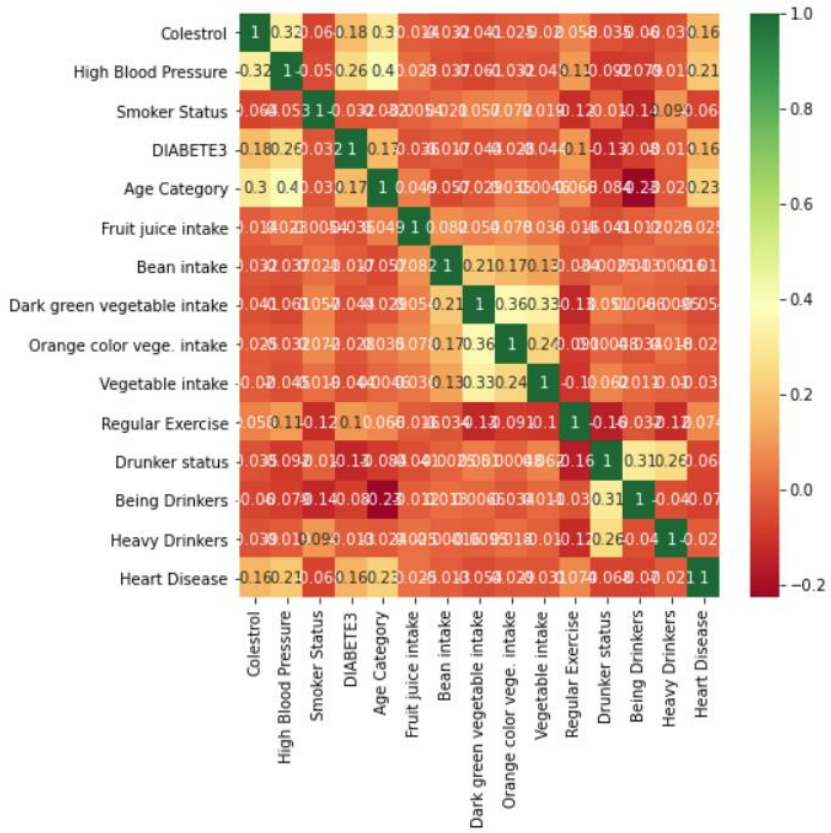


When checked the above diagram, dark green vegetable intake seems like a good indicator to predict the CHD patients. The median of dark green intake seems lower for people with CHD and higher for people with no CHD. The other boxplots seem similar to each other.

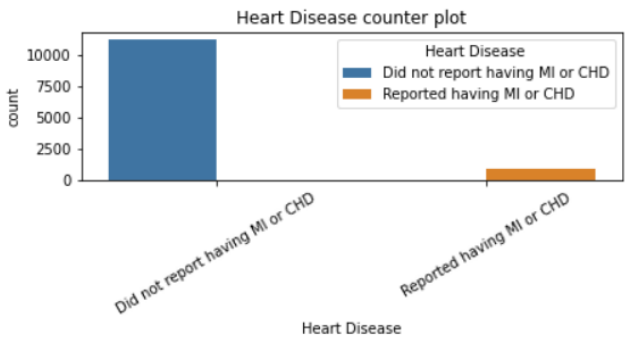
12) Numerical Encoding was used to categorical data, so now the whole dataset consists with numerical values.

	Colestrol	High Blood Pressure	Smoker Status	DIABETE3	Age Category	Fruit juice intake	Bean intake	Dark green vegetable intake	Orange color vege. intake	Vegetable intake	Regular Exercise	Drunker status	Being Drinkers	Heavy Drinkers	Heart Disease
0	0	1	3	0	2	29.0	3.0	33.0	10.0	71.0	0	1	1	1.0	0
2	1	0	3	0	2	0.0	0.0	43.0	29.0	43.0	0	1	0	1.0	0
3	0	0	3	0	0	0.0	50.0	100.0	50.0	71.5	0	0	0	1.0	0
4	0	0	3	0	1	0.0	67.0	100.0	71.0	100.0	0	1	1	1.0	0
5	1	0	1	0	3	0.0	55.0	100.0	37.0	75.0	0	1	1	1.0	0

13) Used correlation Coefficient plot to check whether there are highly correlated columns with each other and there were no such columns



14) The classes variable is highly imbalanced.




```
According to the below results we can conclude that the target cariable is highly imbalanced
Did not report having MI or CHD    92.180734
Reported having MI or CHD           7.819266
Name: Heart Disease, dtype: float64
```

15) Using under sampling, over sampling, SMOTE techniques to have a balanced dataset and Logistic regression model was used to make the predictions.

Oversampling the minority class accuracy, recall and F1-score

```
Accuracy  0.7058823529411765
Recall  0.78099173553719
F1 score  0.2957746478873239
```

	0	1
0	1971	847
1	53	189

Under Sampling the Majority class accuracy, recall and F1-score

```
Accuracy  0.6934640522875817
Recall  0.7851239669421488
F1 score  0.28831562974203345
```

	0	1
0	1932	886
1	52	190

SMOTE sampling approach

```
Accuracy  0.7058823529411765
Recall  0.6198347107438017
F1 score  0.25
```

	0	1
0	2010	808
1	92	150

According to the above results, the oversampling the minority class had worked well and given the best recall, F1-score and accuracy.