Research Question 2: Do the behavioral factors such as alcohol intake, Tobacco intake, level of Physical Activities of a person and socioeconomic factors such as education, Income, and gender of a person in US impact on their overall health?

1) Variable renaming, so the columns are easy to understand

```
#2) 2)→Do the behavioural factors such as alcohol intake, Tobacco intake, level of Physical Activities
# of a person and socioeconomic factors such as education, Income and gender of a person in US impact on their overall health?

# DRNKANY5 - Calculated variable for adults who reported having had at least one drink of alcohol in the past 30 days
# DROCDY3_ - Calculated variable for drink-occasions-per-day
# _RFBING5 - Calculated variable for binge drinkers
# _DRNKWEK - Calculated variable for calculated total number of alcoholic beverages consumed per week
# _RFDRHV5 - Calculated variable for heavy drinkers
# _SMOKER3 - Calculated variable for four-level smoker status
# _RFSMOK3 - Calculated variable for adults who are current smokers
# _TOTINDA - LEISURE TIME PHYSICAL ACTIVITY CALCULATED VARIABLE
# _EDUCAG -  Calculated variable for level of education completed
# _INCOMG -  Income Category
# SEX - Gender


RQ2 = Survey[['_RFHLTH', '_RFBING5', '_RFDRHV5','_SMOKER3', '_RFSMOK3', '_EDUCAG' , '_INCOMG' , '_TOTINDA' , 'SEX']]

RQ2 = RQ2.rename({'_RFHLTH': 'Overall_Health', '_RFBING5': 'Binge Drinkers', '_RFDRHV5': 'Heavy Drinkers','_SMOKER3': 'Smoker St
```

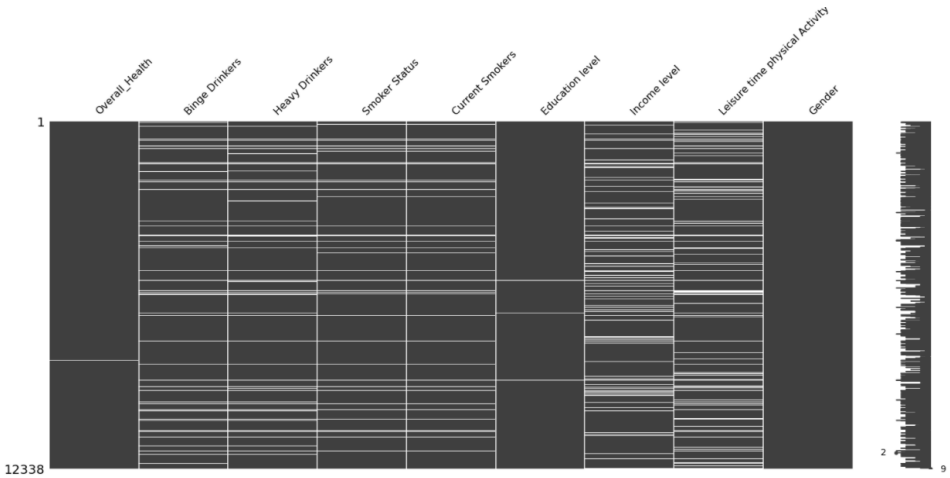
2) Variables consisted with missing values. Some numerical columns have values with number 9, which represents missing variables. Removed them. Categorical variables with 'Missing' were imputed with NaN values.

```
#   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
0   Overall_Health                12273 non-null  object
1   Binge Drinkers                11385 non-null  object
2   Heavy Drinkers                11396 non-null  float64
3   Smoker Status                 11666 non-null  object
4   Current Smokers               11666 non-null  object
5   Education level               12246 non-null  object
6   Income level                  10341 non-null  object
7   Leisure time physical Activity  10955 non-null  object
8   Gender                        12338 non-null  object
```

3) Number of missing values of each columns

```
Overall_Health                    65
Binge Drinkers                   953
Heavy Drinkers                   942
Smoker Status                    672
Current Smokers                  672
Education level                   92
Income level                    1997
Leisure time physical Activity  1383
Gender                             0
dtype: int64
```

Null value representation in a graph.



4) The null values of 'Overall Health' rows were deleted. The rest of the null values of categorical columns were imputed with mode. Now there are no nulls in the dataset.

```
Overall_Health                    0
Binge Drinkers                    0
Heavy Drinkers                    0
Smoker Status                     0
Current Smokers                   0
Education level                   0
Income level                      0
Leisure time physical Activity    0
Gender                            0
dtype: int64
```

5) Following diagram shows how, the selected columns were affected by the overall health. We used the bar plots to analyse them further.

Binge Drinkers: There are many people with good health, who are not binge drinkers. The count of binge drinkers is low. When the Binge Drinkers column is considered, we can't conclude much that it had a relationship with the health status of a person

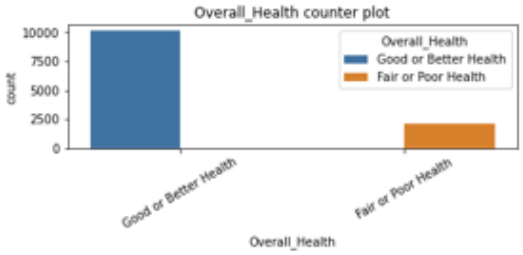
Heavy Drinkers: 1 represents no heavy drinkers, 0 represents heavy drinkers. Again, we have no clear evidence to show that it has a direct impact on showing a person's overall health status.

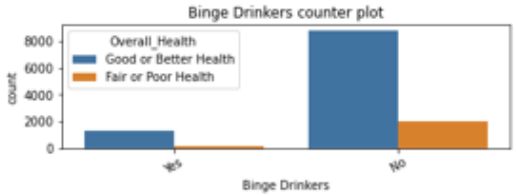
Smoker Status: There are people who have never smoked, and many are with good health, but some have bad health too. The health of former smokers is considerably fair or poor. This might show that the health is impacted on smoking status of a person.

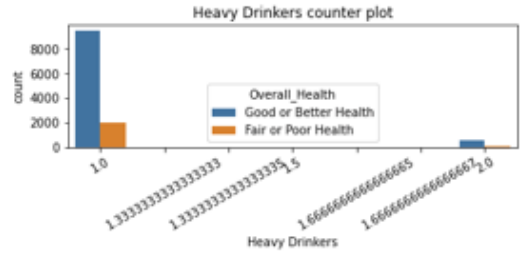
Current smoker status: There are few people who are current smokers and have bad health too. Most of these classes are imbalanced. So, we cannot come into very clear conclusions

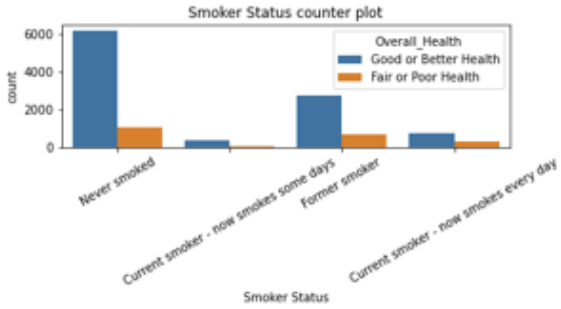
Educational Level: People who have graduated from college or technical school have an overall good health. There is low amount of poor health people, when compared to other educational status.

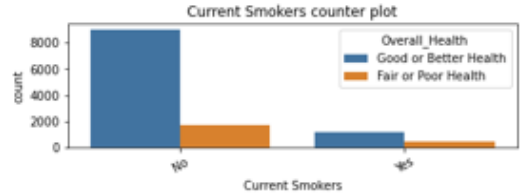
Gender: When gender is considered there are higher number of female respondents in the study. We cannot conclude that the gender significantly impacts on the health of a person.

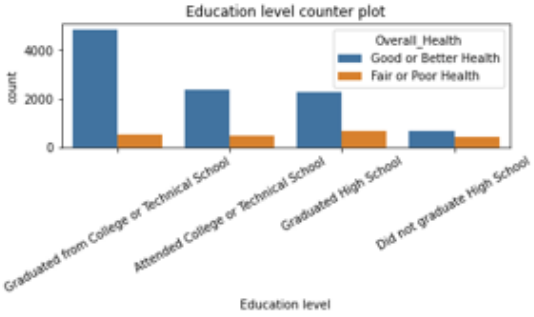

Overall_Health counter plot
Overall_Health
Good or Better Health
Fair or Poor Health
count
Overall_Health


Binge Drinkers counter plot
Overall_Health
Good or Better Health
Fair or Poor Health
count
Yes
No
Binge Drinkers

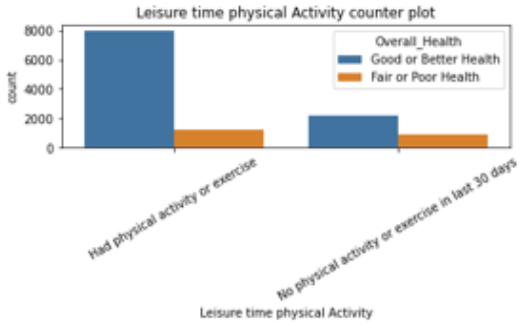

Heavy Drinkers counter plot
Overall_Health
Good or Better Health
Fair or Poor Health
count
1.0
1.3333333333333333
1.5
1.6666666666666665
2.0
Heavy Drinkers

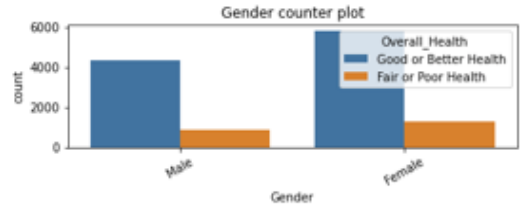

Smoker Status counter plot
Overall_Health
Good or Better Health
Fair or Poor Health
count
Never smoked
Current smoker - now smokes some days
Former smoker
Current smoker - now smokes every day
Smoker Status


Current Smokers counter plot
Overall_Health
Good or Better Health
Fair or Poor Health
count
No
Yes
Current Smokers

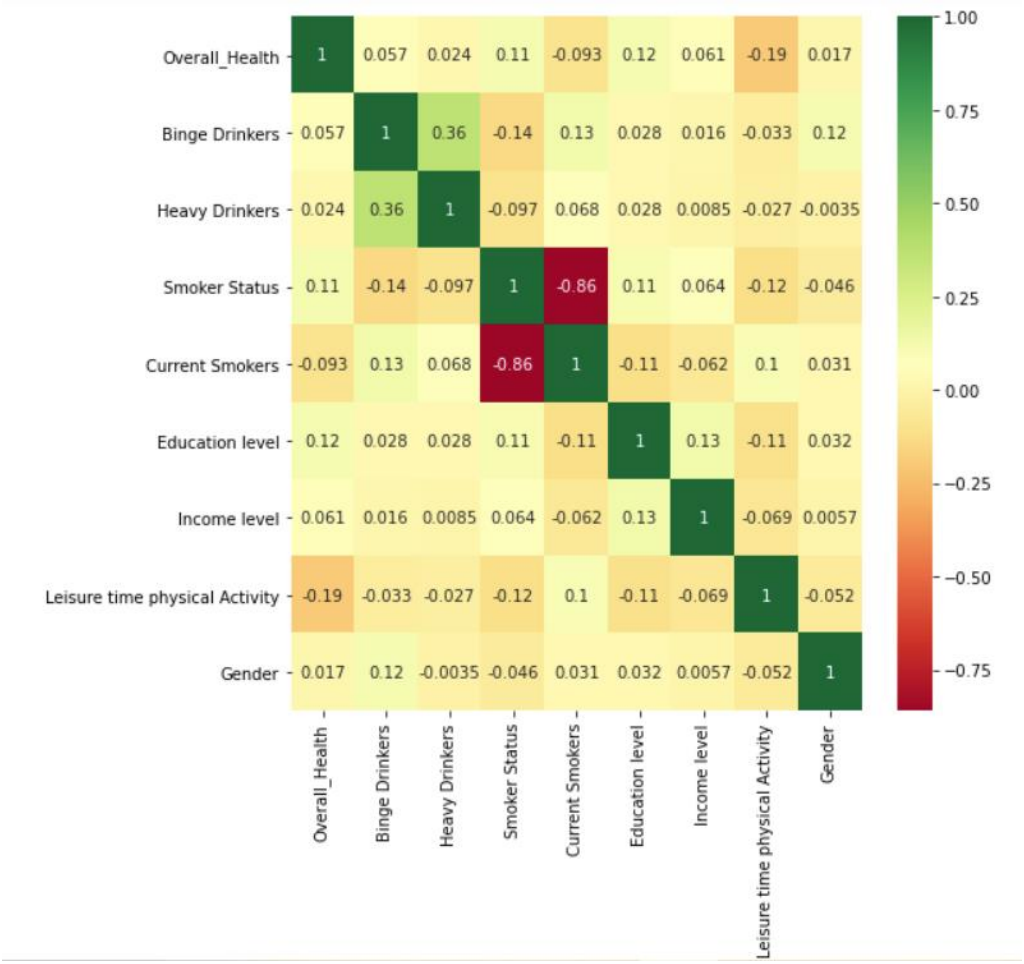

Education level counter plot
Overall_Health
Good or Better Health
Fair or Poor Health
count
Graduated from College or Technical School
Attended College or Technical School
Graduated High School
Did not graduate High School
Education level


Income level counter plot
Overall_Health
Good or Better Health
Fair or Poor Health
count
$50,000 or more
35,000to less than50,000
Less than $15,000
25,000to less than35,000
15,000to less than25,000
Income level


Leisure time physical Activity counter plot
Overall_Health
Good or Better Health
Fair or Poor Health
count
Had physical activity or exercise
No physical activity or exercise in last 30 days
Leisure time physical Activity


Gender counter plot
Overall_Health
Good or Better Health
Fair or Poor Health
count
Male
Female
Gender

6) Since most of the columns were categorical variables, we converted them into encoded variables. Let's see that that there are correlated variables with each other.



There are negatively highly correlated variables. Smoker Status, and current smokers. So, we can drop one of the columns from them. Let's remove the Current Smokers column.

7) We finally used, logistic Regression Statistical Model to see the relationship of the health status to other proposed variables. Since we have a binary class variable, we need to use logistic regression to check the relationship.

```
Optimization terminated successfully.
         Current function value: 0.434911
         Iterations 6
                            Results: Logit
================================================================================
Model:                    Logit                Pseudo R-squared:     0.059
Dependent Variable:       Overall_Health       AIC:                  10689.3219
Date:                     2021-09-19 10:55     BIC:                  10741.2280
No. Observations:         12273                Log-Likelihood:       -5337.7
Df Model:                 6                    LL-Null:              -5673.0
Df Residuals:             12266                LLR p-value:          1.2860e-141
Converged:                1.0000               Scale:                1.0000
No. Iterations:           6.0000
--------------------------------------------------------------------------------
                              Coef.  Std.Err.    z     P>|z|   [0.025  0.975]
--------------------------------------------------------------------------------
Binge Drinkers                0.5133   0.0913   5.6222 0.0000  0.3344  0.6923
Heavy Drinkers                0.5166   0.0708   7.2975 0.0000  0.3778  0.6553
Smoker Status                 0.2826   0.0230  12.3130 0.0000  0.2376  0.3276
Education level               0.2039   0.0201  10.1547 0.0000  0.1646  0.2433
Income level                  0.0884   0.0188   4.6890 0.0000  0.0514  0.1253
Leisure time physical Activity -0.8901   0.0502 -17.7192 0.0000 -0.9886 -0.7917
Gender                        0.0503   0.0498   1.0094 0.3128 -0.0473  0.1479
================================================================================
```