

Research Question 3:

Is there a relationship between the BMI level and the diabetes of an adult in the US?

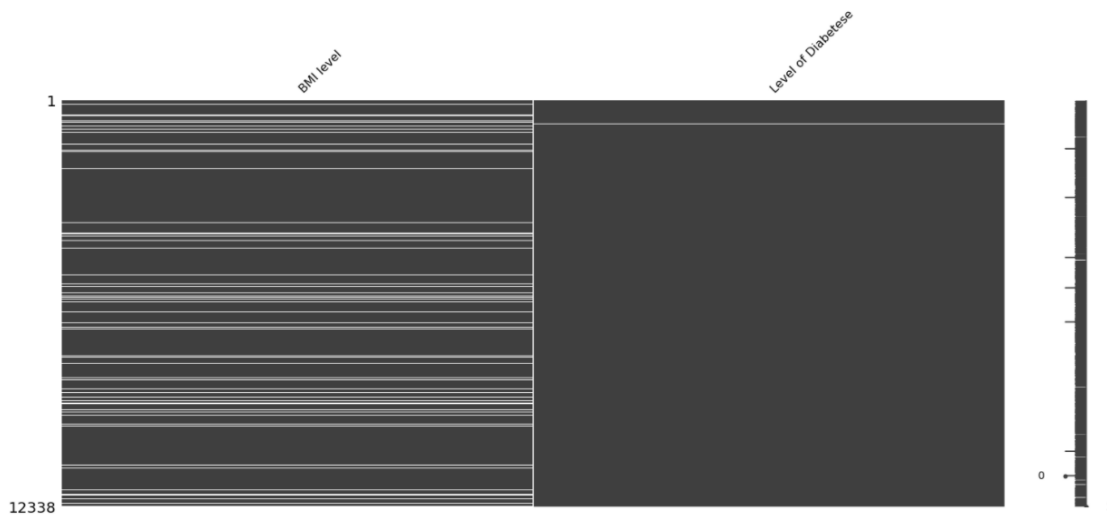
The relationship between two categorical variables will be evaluated here. First the column rename was performed.

```
RQ3 = Survey[['_BMI5CAT', 'DIABETE3']]
RQ3 = RQ3.rename({'_BMI5CAT': 'BMI level', 'DIABETE3': 'Level of Diabetese'}, axis=1)
```

- 1) Missing values were handled. Columns with null values

```
BMI level          1074
Level of Diabetese    13
dtype: int64
```

- 2) Existing null values of the columns. The income column has a large number of nulls.



- 3) Instead of imputing values to them, the null values were removed. The new row counts,

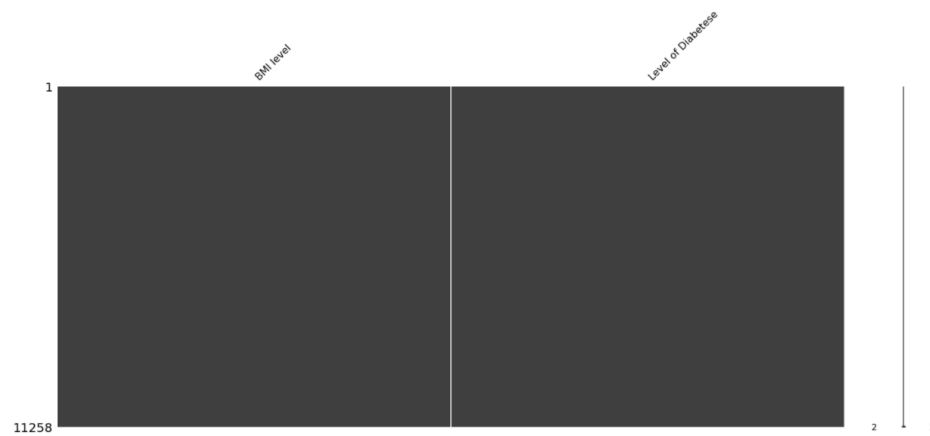
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 11258 entries, 0 to 12337
Data columns (total 2 columns):
#   Column          Non-Null Count  Dtype
---  ---
0   BMI level       11258 non-null  object
1   Level of Diabetese 11258 non-null  object
dtypes: object(2)
```

- 4) New null count

```

BMI level      0
Level of Diabetes  0
dtype: int64
<AxesSubplot:>

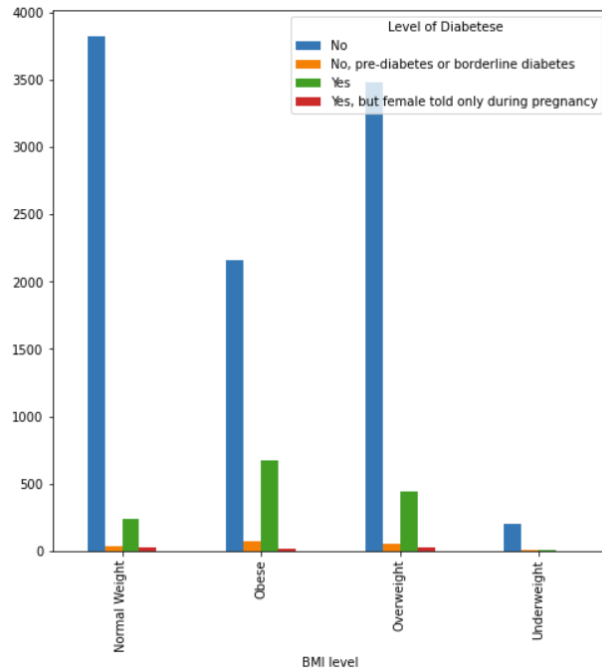
```



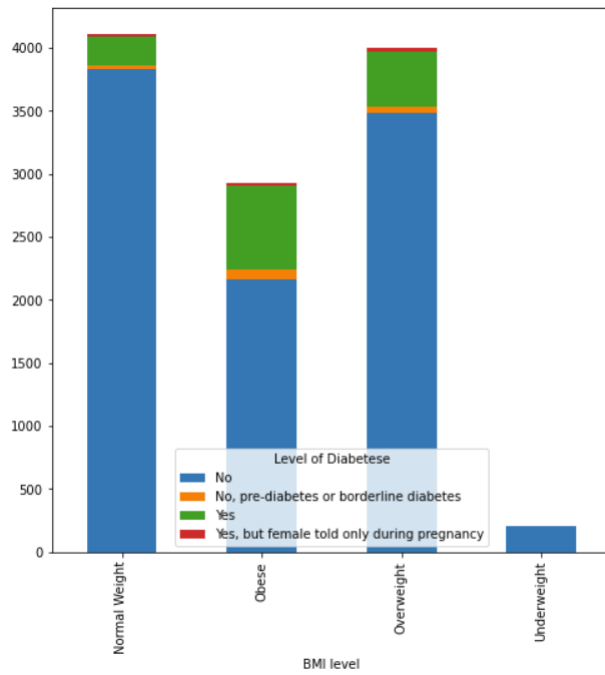
5) Across tab was generated to analyse the values further

Level of Diabetes	No	No, pre-diabetes or borderline diabetes	Yes	Yes, but female told only during pregnancy
BMI level				
Normal Weight	3829		30 234	22
Obese	2164		75 673	20
Overweight	3483		54 437	28
Underweight	202		2 5	0

6) Plotting Cross barplot values. In the barplot it is very clear that if a person is in obese state, there is a very high change for him/her to have diabetes. The overweight person too has the threat, but less than obese people.



7) Plotting the same results in Stack bar graph



8) Using Chi-squared value to identify the relationship between two categorical variables.

```
stat, p, dof, expected = stats.chi2_contingency(crosstab)
```

```
# interpret p-value
```

```
alpha = 0.05
```

```
print("p value is " + str(p))
```

```
if p <= alpha:
```

```
    print('Dependent (reject H0)')
```

```
else:
```

```
    print('Independent (H0 holds true)')
```

```
print(expected)
```

```
p value is 2.990816817070346e-117
```

```
Dependent (reject H0)
```

```
[[3.53748179e+03 5.88483745e+01 4.93083585e+02 2.55862498e+01]
```

```
 [2.52050950e+03 4.19303606e+01 3.51329543e+02 1.82305916e+01]
```

```
 [3.44034074e+03 5.72323681e+01 4.79543258e+02 2.48836383e+01]
```

```
 [1.79667969e+02 2.98889678e+00 2.50436134e+01 1.29952034e+00]]
```