A. Organizational issue

A1. Question for Analysis

What is the relationship between specific medical conditions and gender among the patients?

A2. Benefit from Analysis

Data from this analysis can improve patient care, manage risk, and help make informed policy decisions.

Firstly, healthcare providers can analyze the most prevalent medical conditions within their patient population. This information allows the allocation of resources more efficiently and design preventive measures for specific medical conditions. Moreover, insurers can use the analysis to assess and manage risk. They can develop more accurate risk, set appropriate premiums and design insurance plans with the healthcare needs of their patients. Lastly, policymakers can use this analysis to inform public health initiatives and allocates resources strategically. They can identify areas of concern and address specific needs of the community.

A3. Data Identification

Independent Variables

Variable name	Data type	Discrete/Continuous/Categorical?
Gender	Qualitative	Categorical

Dependent Variables

Variable	Data type	Discrete/Continuous/Categorical?
High Blood	Qualitative	Categorical
Stroke	Qualitative	Categorical
Diabetes	Qualitative	Categorical
Overweight	Qualitative	Categorical
Arthritis	Qualitative	Categorical
Hyperlipidemia	Qualitative	Categorical
Back Pain	Qualitative	Categorical
Anxiety	Qualitative	Categorical
Allergic Rhinitis	Qualitative	Categorical
Reflux esophagitis	Qualitative	Categorical
Asthma	Qualitative	Categorical

B. Describe the data analysis

B2. Output

```
Chi-square test for HighBlood and Gender:
Contingency Table:
Gender     Female  Male  Nonbinary
HighBlood
No           2987  2807        116
Yes          2031  1961         98
Chi-square statistic: 2.599885272446885
```

```
P-value: 0.2725474269352985
Expected frequencies: [[2965.638 2817.888  126.474]
 [2052.362 1950.112   87.526]]
There is NO significant association between the variables.

========================================

Chi-square test for Stroke and Gender:
Contingency Table:
Gender  Female  Male  Nonbinary
Stroke
No        4011  3827        169
Yes       1007   941         45
Chi-square statistic: 0.3341124994788903
P-value: 0.8461520141682087
Expected frequencies: [[4017.9126 3817.7376  171.3498]
 [1000.0874  950.2624   42.6502]]
There is NO significant association between the variables.

========================================

Chi-square test for Diabetes and Gender:
Contingency Table:
Gender    Female  Male  Nonbinary
Diabetes
No          3639  3466        157
Yes         1379  1302         57
Chi-square statistic: 0.09819529981479722
P-value: 0.952088153838979
Expected frequencies: [[3644.0716 3462.5216  155.4068]
 [1373.9284 1305.4784   58.5932]]
There is NO significant association between the variables.

========================================

Chi-square test for Overweight and Gender:
Contingency Table:
Gender      Female  Male  Nonbinary
Overweight
No            1455  1392         59
Yes           3563  3376        155
Chi-square statistic: 0.2824451478065447
P-value: 0.868296030152125
Expected frequencies: [[1458.2308 1385.5808   62.1884]
 [3559.7692 3382.4192  151.8116]]
There is NO significant association between the variables.

========================================

Chi-square test for Arthritis and Gender:
Contingency Table:
Gender     Female  Male  Nonbinary
Arthritis
No           3252  3045        129
Yes          1766  1723         85
Chi-square statistic: 2.455524018389509
P-value: 0.2929474583172357
```

```
Expected frequencies: [[3224.5668 3063.9168  137.5164]
 [1793.4332 1704.0832   76.4836]]
There is NO significant association between the variables.

========================================

Chi-square test for Hyperlipidemia and Gender:
Contingency Table:
Gender          Female  Male  Nonbinary
Hyperlipidemia
No                3367  3127        134
Yes               1651  1641         80
Chi-square statistic: 3.8250956736102077
P-value: 0.14770358215832177
Expected frequencies: [[3325.9304 3160.2304  141.8392]
 [1692.0696 1607.7696   72.1608]]
There is NO significant association between the variables.

========================================

Chi-square test for BackPain and Gender:
Contingency Table:
Gender    Female  Male  Nonbinary
BackPain
No          2929  2845        112
Yes         2089  1923        102
Chi-square statistic: 5.546261322680868
P-value: 0.06246613798261359
Expected frequencies: [[2953.5948 2806.4448  125.9604]
 [2064.4052 1961.5552   88.0396]]
There is NO significant association between the variables.

========================================

Chi-square test for Anxiety and Gender:
Contingency Table:
Gender   Female  Male  Nonbinary
Anxiety
No         3390  3253        142
Yes        1628  1515         72
Chi-square statistic: 0.7254588119830732
P-value: 0.6957746804884735
Expected frequencies: [[3404.713 3235.088  145.199]
 [1613.287 1532.912   68.801]]
There is NO significant association between the variables.

========================================

Chi-square test for Allergic_rhinitis and Gender:
Contingency Table:
Gender             Female  Male  Nonbinary
Allergic_rhinitis
No                   3035  2891        133
Yes                  1983  1877         81
Chi-square statistic: 0.2461204292177224
P-value: 0.8842104185732986
Expected frequencies: [[3040.4062 2888.9312  129.6626]
```

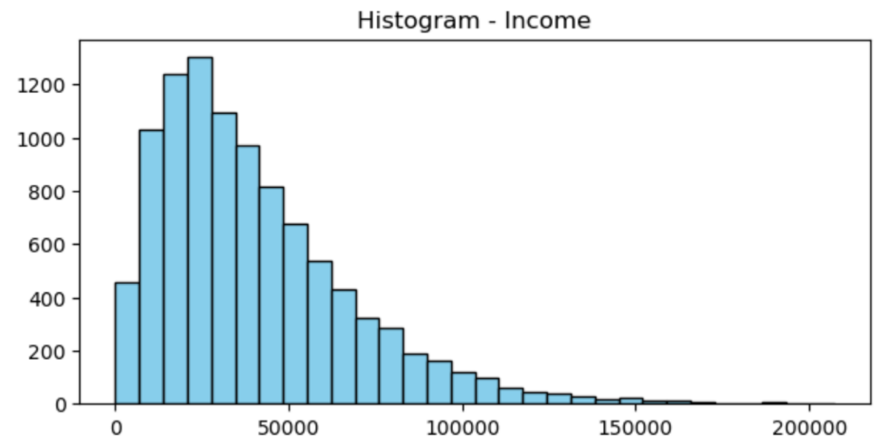
```
 [1977.5938 1879.0688   84.3374]]
There is NO significant association between the variables.

========================================

Chi-square test for Reflux_esophagitis and Gender:
Contingency Table:
Gender              Female  Male  Nonbinary
Reflux_esophagitis
No                    2896  2834        135
Yes                   2122  1934         79
Chi-square statistic: 4.775081883690254
P-value: 0.09185528386693614
Expected frequencies: [[2943.057 2796.432  125.511]
 [2074.943 1971.568   88.489]]
There is NO significant association between the variables.

========================================

Chi-square test for Asthma and Gender:
Contingency Table:
Gender  Female  Male  Nonbinary
Asthma
No        3578  3379        150
Yes       1440  1389         64
Chi-square statistic: 0.32645955757103173
P-value: 0.8493959928098948
Expected frequencies: [[3566.2926 3388.6176  152.0898]
 [1451.7074 1379.3824   61.9102]]
There is NO significant association between the variables.

========================================
```

B3. Justification

For this analysis, I used the chi-square test. Since my medical conditions variables (High Blood, Stroke, Diabetes, Overweight, Arthritis, Hyperlipidemia, Back Pain, Anxiety, Allergic Rhinitis, Reflux esophagitis, Asthma) and gender (female, male, binary) are categorical values, the chi-square test can be used to determine whether there is a significant association between these categorical variables against each other.

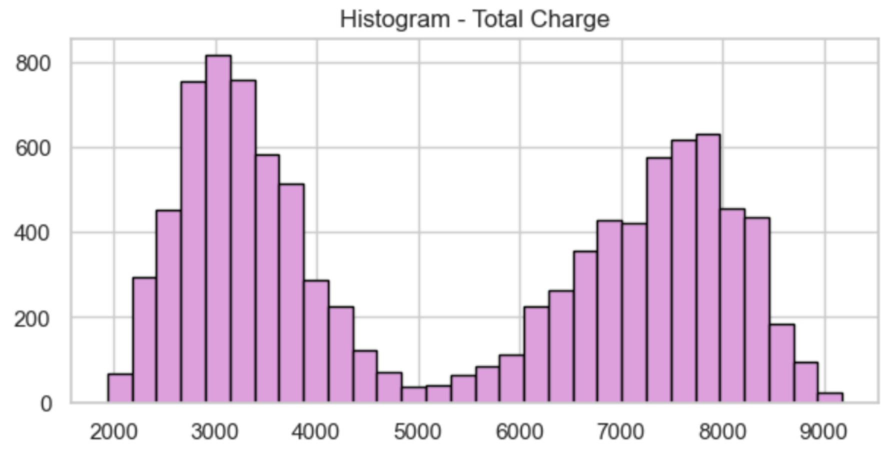
C. Univariate Statistics (2 continuous and 2 categorical)

C1. Visual



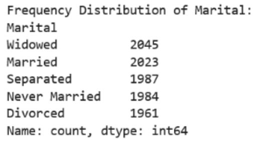
```
Mean Income: 40490.49516
Median Income: 33768.42
Variance Income: 813456185.1732982
Standard Deviation Income: 28521.15329318396
```

The income histogram shows a right-skewed distribution. This suggests that most patients have incomes below the average, while a few have very high incomes.

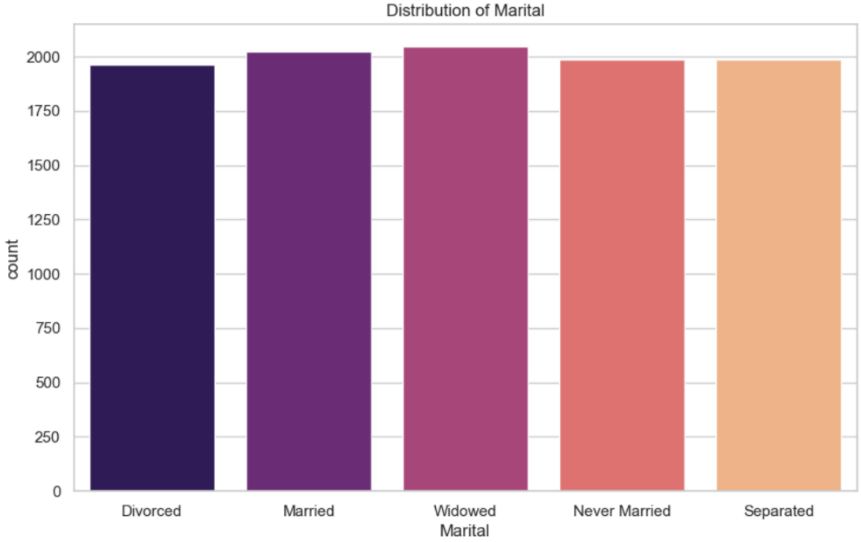


```
Mean Total Charge: 5312.1727687502
Median Total Charge: 5213.952
Variance Total Charge: 4754117.287963928
Standard Deviation Total Charge: 2180.393837810942
```

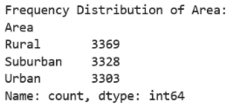
The total charge histogram shows a bimodal distribution. This suggests that there are two type of patients – those who are charged a lot and those who were charged a little.



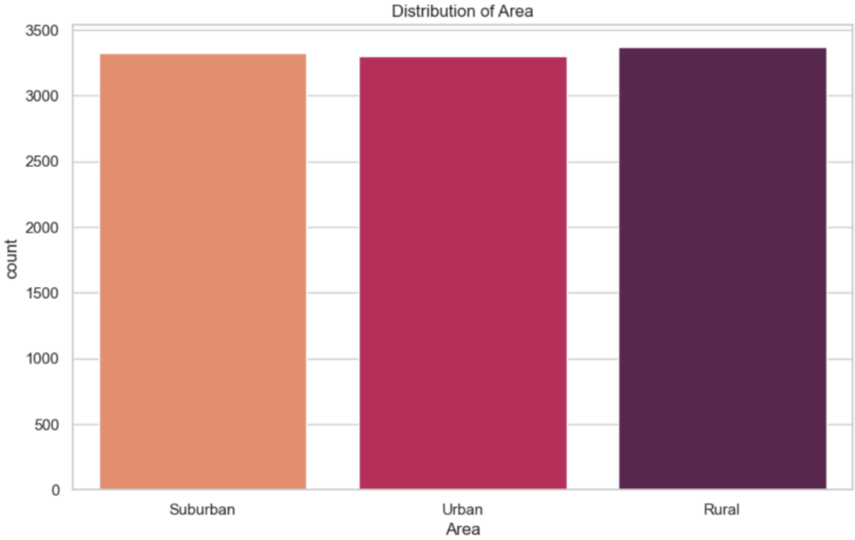
```
Frequency Distribution of Marital:
Marital
Widowed          2045
Married          2023
Separated        1987
Never Married    1984
Divorced         1961
Name: count, dtype: int64
```



The total charge histogram shows a uniform distribution. This suggests that the patient's marital status is evenly distributed across the five different categories.



```
Frequency Distribution of Area:
Area
Rural       3369
Suburban    3328
Urban       3303
Name: count, dtype: int64
```

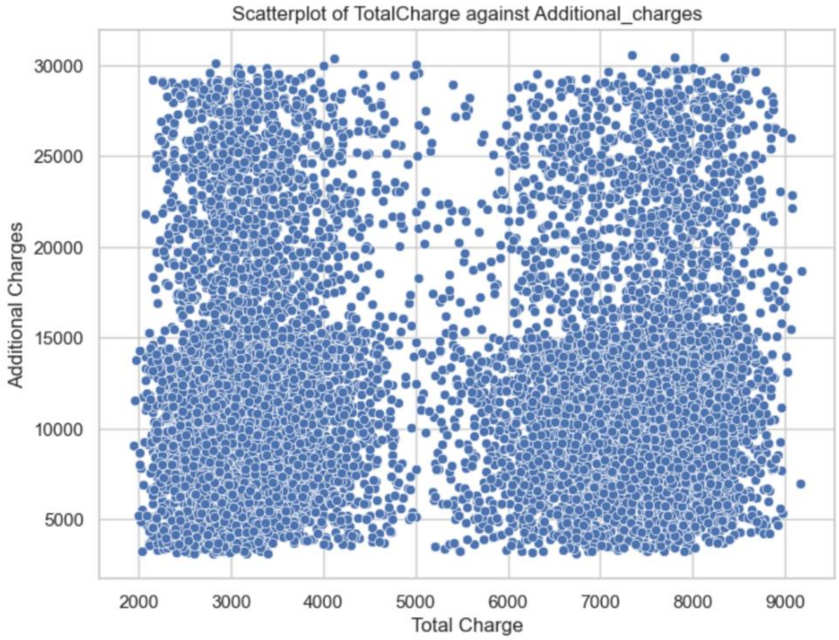


The area histogram shows a uniform distribution. This suggests that the patient's area classification is evenly distributed across the three categories.

D. Bivariate Statistics (2 continuous and 2 categorical)

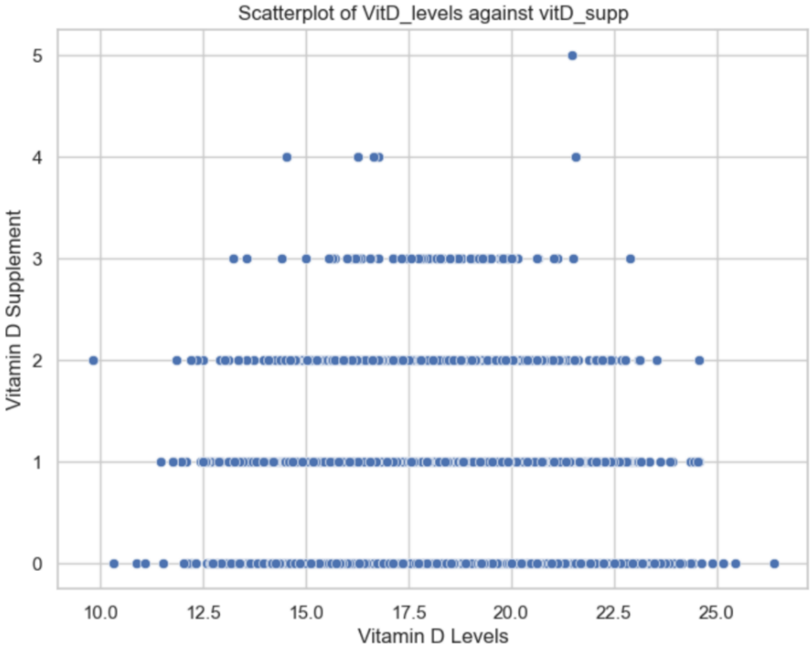
D1. Visual

```
Correlation coefficient between TotalCharge and Additional_charges: 0.02925582402378014
```

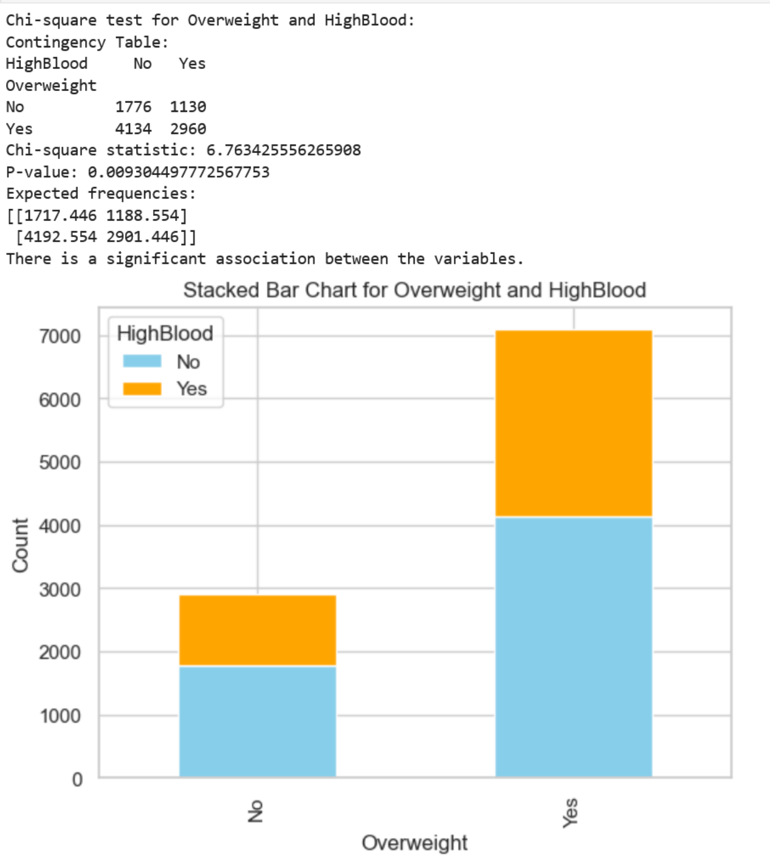


The scatter plot shows the relationship between total charge and additional charges. The correlation coefficient is approximately 0.029, which indicates a positive but negligible linear relationship between the two variables.

```
Correlation coefficient between VitD_levels and vitD_supp: -0.007203220113302815
```



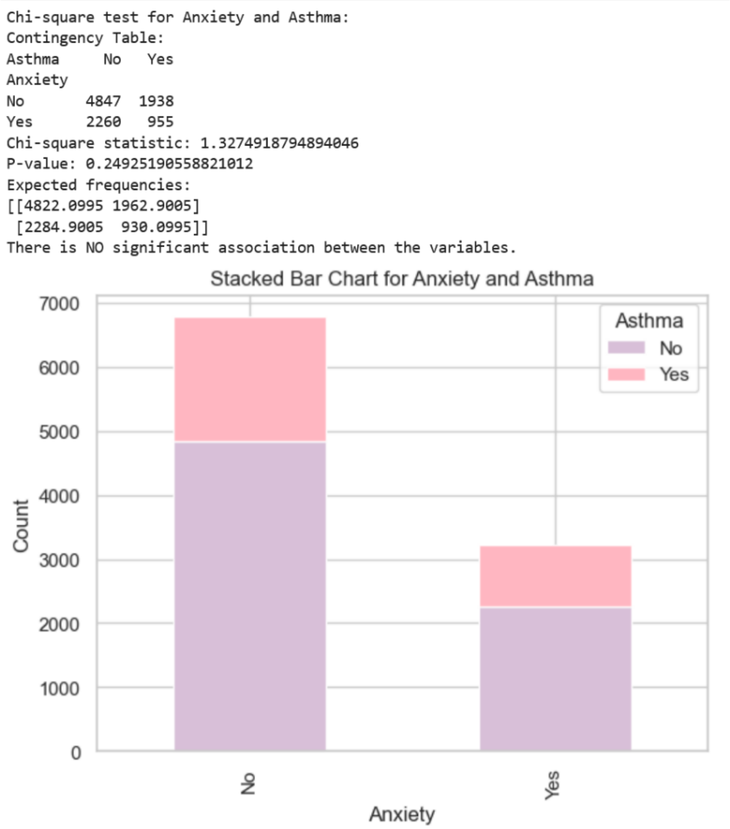
The scatter plot shows the relationship between vitamin D levels and vitamin D supplements. The correlation coefficient is approximately -0.0072, which indicates a negative and negligible linear relationship between the two variables.



```
Chi-square test for Overweight and HighBlood:
Contingency Table:
HighBlood     No   Yes
Overweight
No          1776  1130
Yes         4134  2960
Chi-square statistic: 6.763425556265908
P-value: 0.009304497772567753
Expected frequencies:
[[1717.446 1188.554]
 [4192.554 2901.446]]
There is a significant association between the variables.
```

The stacked bar chart shows the relationship between overweight and high blood. The chart shows that patients who are not overweight have a higher count of people without high blood pressure, while among those who are overweight, the counts of patients with and without high blood pressure are closer.

The p-value is 0.009. Compared to the alpha, which is 0.05, the p-value is smaller, indicating a significant association between being overweight and having high blood pressure.



```
Chi-square test for Anxiety and Asthma:
Contingency Table:
Asthma     No   Yes
Anxiety
No       4847  1938
Yes      2260   955
Chi-square statistic: 1.3274918794894046
P-value: 0.24925190558821012
Expected frequencies:
[[4822.0995 1962.9005]
 [2284.9005  930.0995]]
There is NO significant association between the variables.
```

The stacked bar chart shows the relationship between anxiety and asthma. The chart shows that patients who have anxiety have a slightly higher count of people with asthma than those who do not have anxiety, but the difference is not too significant.

The p-value is 0.249. Compared to the alpha, which is 0.05, the p-value is larger, indicating no significant association between anxiety and asthma.

E. Implications Summary

E1. Results of the hypothesis test

Null hypothesis: There is no significant association between specific medical conditions and gender among the patients.
Alternative hypothesis: There is a significant association between specific medical conditions and gender among conditions and age among the patients.

The evidence from the data does not provide enough support to claim that there is a statistically significant relationship between medical conditions and gender.

E2. Limitations of analysis

Failing to reject the null hypothesis does not prove the absence of a relationship. With the available data, I could not find strong enough evidence to support the presence of a relationship between medical conditions and gender. This lack of statistical significance does not negate the importance of the research question. It is important to interpret the results in the broader context of the research goals and the characteristics of the dataset.

E3. Recommended course of analysis

The analysis suggests that there is no significant association between specific medical conditions and age among the patients. Firstly, present findings to stakeholders. Clearly communicate and emphasize that the statistical analysis did not provide evidence to reject the null hypothesis. Then, explore and investigate further the potential reasons for the lack of significant association. Consider exploring additional variables or refining categories. Moreover, perform subgroup analyses based on other relevant factors like age, marital status, or other demographic variables. Furthermore, conduct additional research. Identify areas for further research and determine if there are other variables not included in the current dataset. Lastly, consult with statistical experts to get feedback on your analysis and explore alternative methodologies.