

# Heathcare Python Data Analysis

## Data Overview:

### Initial Data Overview:

- There are 15 columns in this dataset.
- There are over 55,000 entries in this dataset with patient information.
- There are:
  - 2 int columns
  - 1 float column
  - 12 object columns
- There are no missing values or entries within this dataset.

### Data Cleaning:

- As mentioned previously, there are no missing values.
- 534 duplicate entries are present in this dataset.
- Upon analyzing samples of the duplicate entries, they seem to contain the same information as the initial entry which will lead me to dropping the entries.
- Upon dropping the duplicate entries, the dataset now has 54966 entries in total.
- Additionally, the names of the 'Name' column had to be set to all lower cases due to the casing of the metadata.  
For example: 'Samuel joYCe', would be unappealing to those reading the data, so forcing the entire column to change its casing to 'Samuel Joyce' or as it is called in the Python functions, `titles()`.
- In the 'Billing Amount' column, the float values were represented as: '38142.109678'. Due to the column's nature of representing money, changing the decimal places was necessary and taken. All entries in this column have been rounded to 2 decimal places.
- In the 'Billing Amount' column, it was also noted there were negative billings in the dataframe which is not accepted because it would indicate there was a misinput when given this data. The choice was made to drop those with negative billings as there were only 108 entries out of 54,966 in the dataframe.

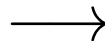
## Data Transforming:

- Appended a new column ‘`treatment_duration`’ to show how long it took a patient to get treated from time of admission to discharge.
- There are identifiers for this dataframe such as, ‘`Name`’ however, for simplicity sake and future SQL analysis, I have added new columns for this dataframe:
  - ‘`patient_id`’
  - ‘`doctor_id`’
- Changed all column names to all lowercases. Additionally added ‘`_`’ to those column names that had spaces initially.
- Changed the column order to make it simpler for SQL use later.

Listed below is the initial dataframe information of columns and the dtypes.

Column	Dtype
<i>Name</i>	object
<i>Age</i>	int64
<i>Gender</i>	object
<i>Blood Type</i>	object
<i>Medical Condition</i>	object
<i>Date of Admission</i>	object
<i>Doctor</i>	object
<i>Hospital</i>	object
<i>Insurance Provider</i>	object
<i>Billing Amount</i>	float64
<i>Room Number</i>	int64
<i>Admission Type</i>	object
<i>Discharge Date</i>	object
<i>Medication</i>	object
<i>Test Results</i>	object

Table 1: Initial Dataset



Column	Dtype
<i>patient_id</i>	int64
<i>doctor_id</i>	int64
<i>patient_name</i>	string
<i>age</i>	int64
<i>gender</i>	category
<i>blood_type</i>	category
<i>medical_condition</i>	string
<i>admission_date</i>	datetime64[ns]
<i>discharge_date</i>	datetime64[ns]
<i>treatment_days</i>	int64
<i>admission_type</i>	category
<i>room_number</i>	int64
<i>doctor_name</i>	string
<i>hospital_name</i>	string
<i>insurance_provider</i>	string
<i>medication</i>	string
<i>test_results</i>	category
<i>billing-amount</i>	float64

Table 2: Transformed Dataset

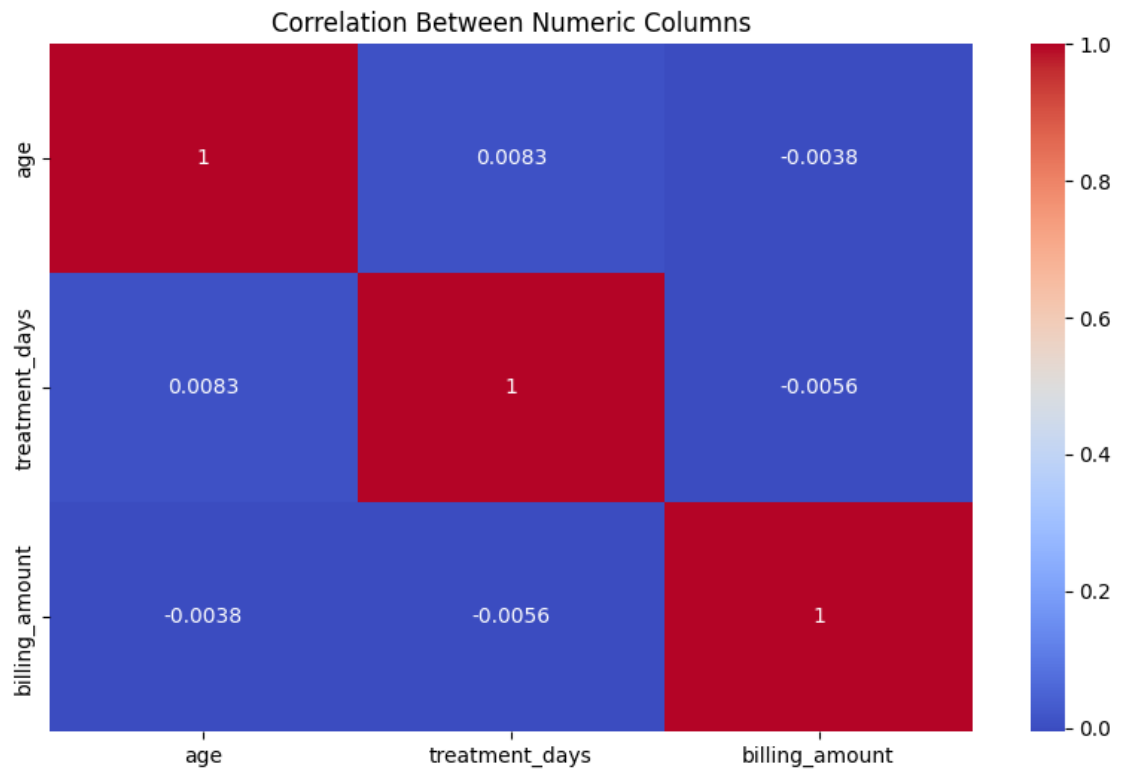
- Looking at the differences in these two tables, the changes are clear as the data has been transformed to satisfy the needs of this data analysis and for further development in SQL.
- As mentioned, new columns were added and the data types were changed to help make sense of the columns and what they represent than just an ‘object’ type.

# Exploratory Data Analysis:

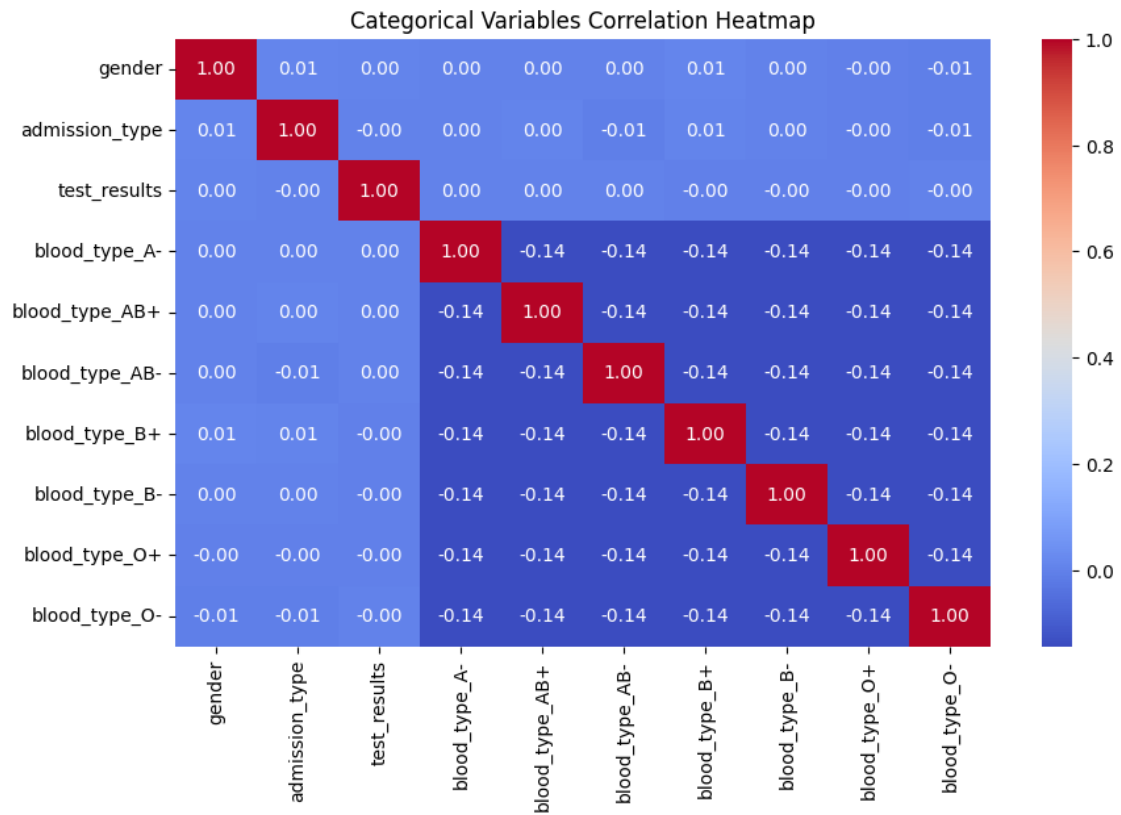
## Correlation Heatmaps:

In this portion of the analysis, it needs to be determined if there are correlations between the columns i.e. if one affects the other more than the other, etc.

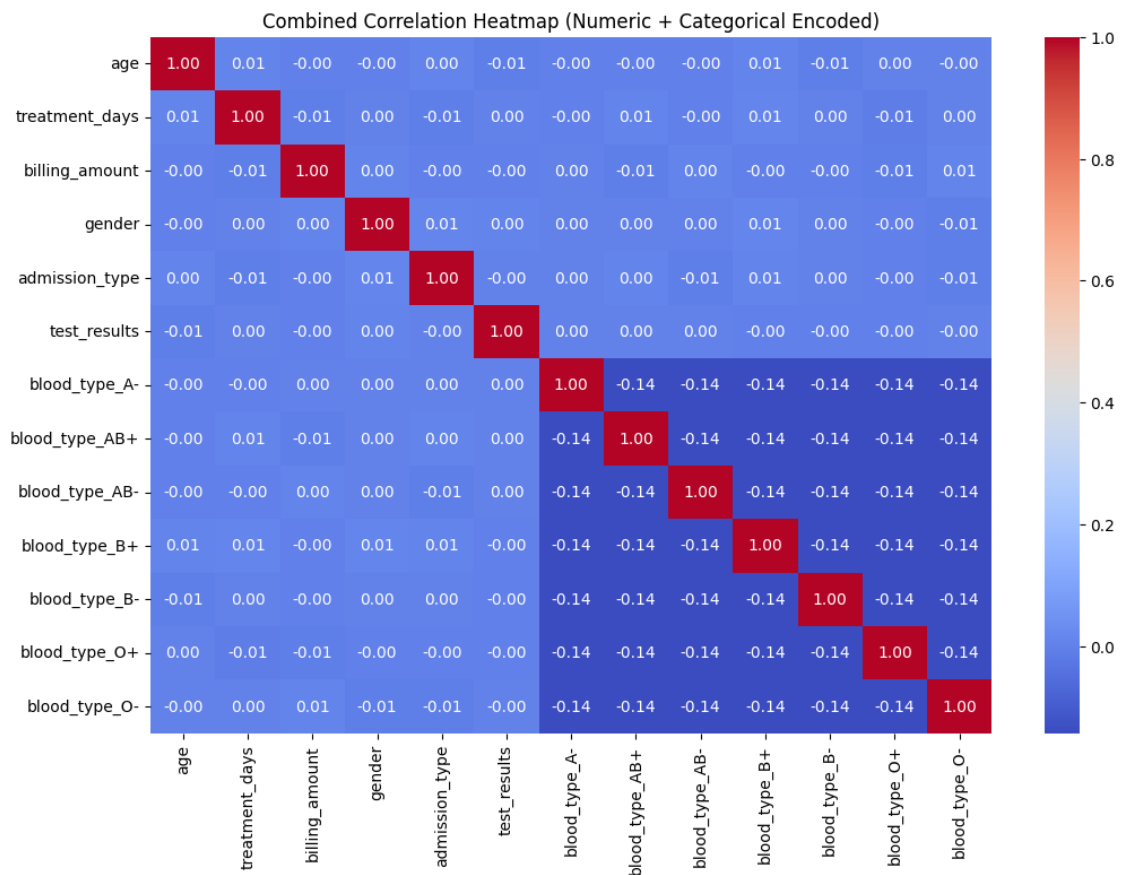
### 1. Numeric Correlation:



### 2. Categorical Correlation:



### 3. Numeric & Categorical Correlation:

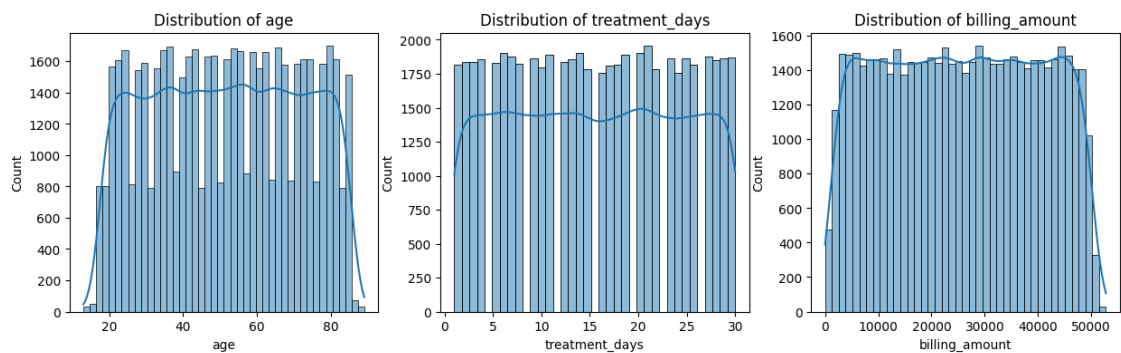


What can be gathered from the correlation heatmaps:

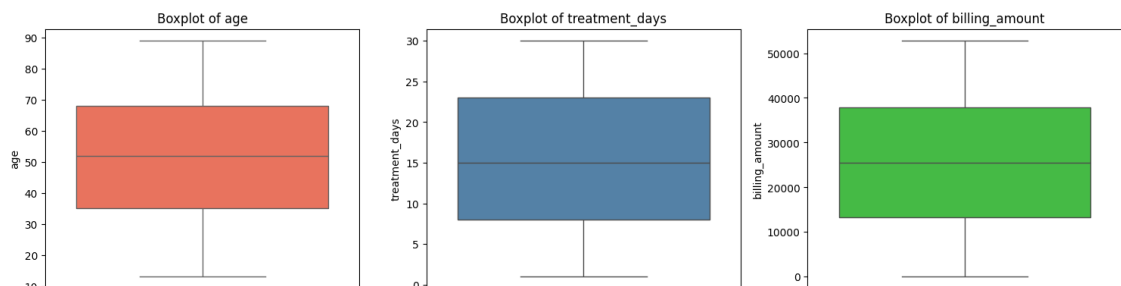
- Overall in all 3 correlation heatmaps, there are no signs of strong positive or negative correlations.
- The only thing present is either no correlation, or a weak negative or positive correlation.
- Interestingly enough, there seems to be a small weak positive correlation with *age* and *treatment days*, which will have statistical testing to see how significant it is.
- Billing amount and treatment days have a weak negative correlation, which statistical testing will be conducted.
- As the analysis goes on, statistical testing will be conducted.
- As far as blood types, they all seem to be consistent with each other, showing 0.14 with each other.
- Nothing too eye catching from the combined correlative heatmap.

## Distributions & Counts:

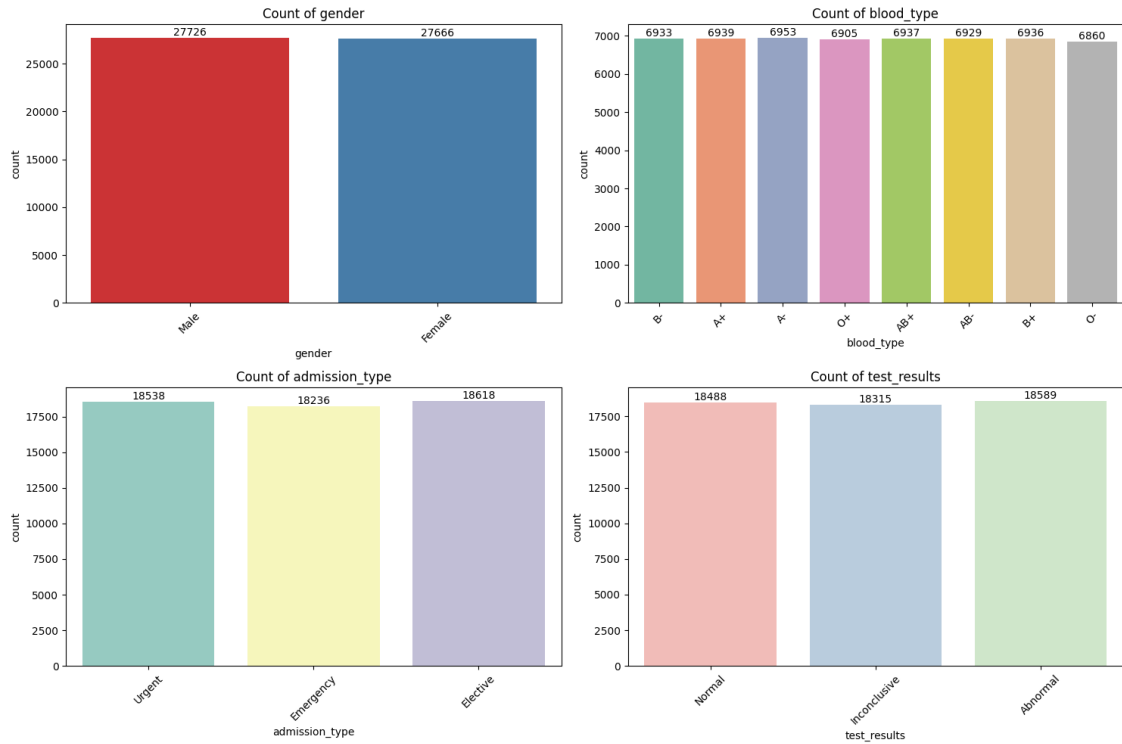
### 1. Numeric Columns:



### 2. Numeric Box Plots:



### 3. Categorical Counts:



What can be taken from these images?

- In the numeric columns distribution graph, it shows a uniform distributions among the 3 separate graphs. This is showing this has a wide spread of age, treatment days, and the billing amount.
- In the box plot graphs, it was tested to see if there were any outliers in the data for the numeric columns, and there seems to be none.
- Finally the categorical count illustrates the differences between the category columns in 4 separate graphs and the specifics within them, and they seem to be even between all the categories within the column.

## Statistical Testing:

### Relationship between Age and Treatment Days

**Test Used:** Pearson correlation coefficient (Pearson's  $r$ )

*Null Hypothesis ( $H_0$ ):*

There is no correlation between `age` and `treatment_days`.

$$\rho = 0$$

*Alternative Hypothesis ( $H_1$ ):*

There is a significant correlation between `age` and `treatment_days`.

$$\rho \neq 0$$

**Test Results:**

Pearson correlation coefficient:  $\rho = 0.0083$

p-value: 0.0495

From this statistical test, the 0.0083 value from the correlation heatmap and this statistic value does indeed correspond to a **weak** positive correlation. Additionally with the p-value being less than 0.05, it also concludes it is statistically significant which leads to rejecting the null hypothesis, as there is some correlation but not too much.

**Relationship between billing\_amount and treatment\_days**

**Test Used:** Pearson correlation coefficient (Pearson's  $r$ )

*Null Hypothesis ( $H_0$ ):* There is no correlation between `billing_amount` and `treatment_days`.

$$\rho = 0$$

*Alternative Hypothesis ( $H_1$ ):* There is a significant correlation between `billing_amount` and `treatment_days`.

$$\rho \neq 0$$

**Test Results:**

Pearson correlation coefficient:  $\rho = -0.0056$

p-value: 0.1903

From this statistical testing, we fail to reject the null hypothesis as the P-Value is greater than 0.05. The correlation is a strong **weak** negative correlation, as mentioned the P-Value is greater than 0.05 which means it is not statistically significant to each other.

**Difference in Billing Amount by Gender**

**Test Used:** Welch's T-Test (Unequal variance t-test)

*Null Hypothesis ( $H_0$ ):* There is no difference in average billing amount between Male and Female patients.

$$\mu_{\text{Male}} = \mu_{\text{Female}}$$

*Alternative Hypothesis ( $H_1$ ):* There is a significant difference in average billing amount between genders.

$$\mu_{\text{Male}} \neq \mu_{\text{Female}}$$

**Test Results:**

Male mean billing amount: \$25653.13

Female mean billing amount: \$25526.90

T-statistic:  $t = 1.048$

Degrees of freedom:  $\nu = 55389.98$

p-value:  $p = 0.295$

The mean values are close to each other, however from this T-Test, the males pay more than the women on average. The T-statistic value indicates a large difference between the two averages. The degree of freedom indicates a large sample was taken to conduct this test. The P-Value again is larger than 0.05, so it is statistically insignificant, so we fail to reject the null hypothesis.

## Difference in Billing Amount Across Admission Types

**Test Used:** One-way ANOVA

*Null Hypothesis ( $H_0$ ):* The mean billing amount is the same across all admission types.

$$\mu_{\text{Emergency}} = \mu_{\text{Urgent}} = \mu_{\text{Elective}}$$

*Alternative Hypothesis ( $H_1$ ):* At least one group has a different mean.

Not all  $\mu_k$  are equal

### Test Results:

Emergency mean billing amount: \$25,544.18

Urgent mean billing amount: \$25,570.74

Elective mean billing amount: \$25,654.31

F-statistic:  $F = 0.304$

p-value:  $p = 0.738$

The F-statistic indicates little variance between the groups, in other words there are no difference between the 3 results. With the P-Value being above the threshold, we fail to reject the null hypothesis as there is no difference between the results.

## Relationship between Gender and Medical Condition

**Test Used:** Chi-Squared Test of Independence

*Null Hypothesis ( $H_0$ ):* There is no relationship between `gender` and `medical_condition`.

Variables are independent

*Alternative Hypothesis ( $H_1$ ):* There is a relationship between `gender` and `medical_condition`.

Variables are not independent

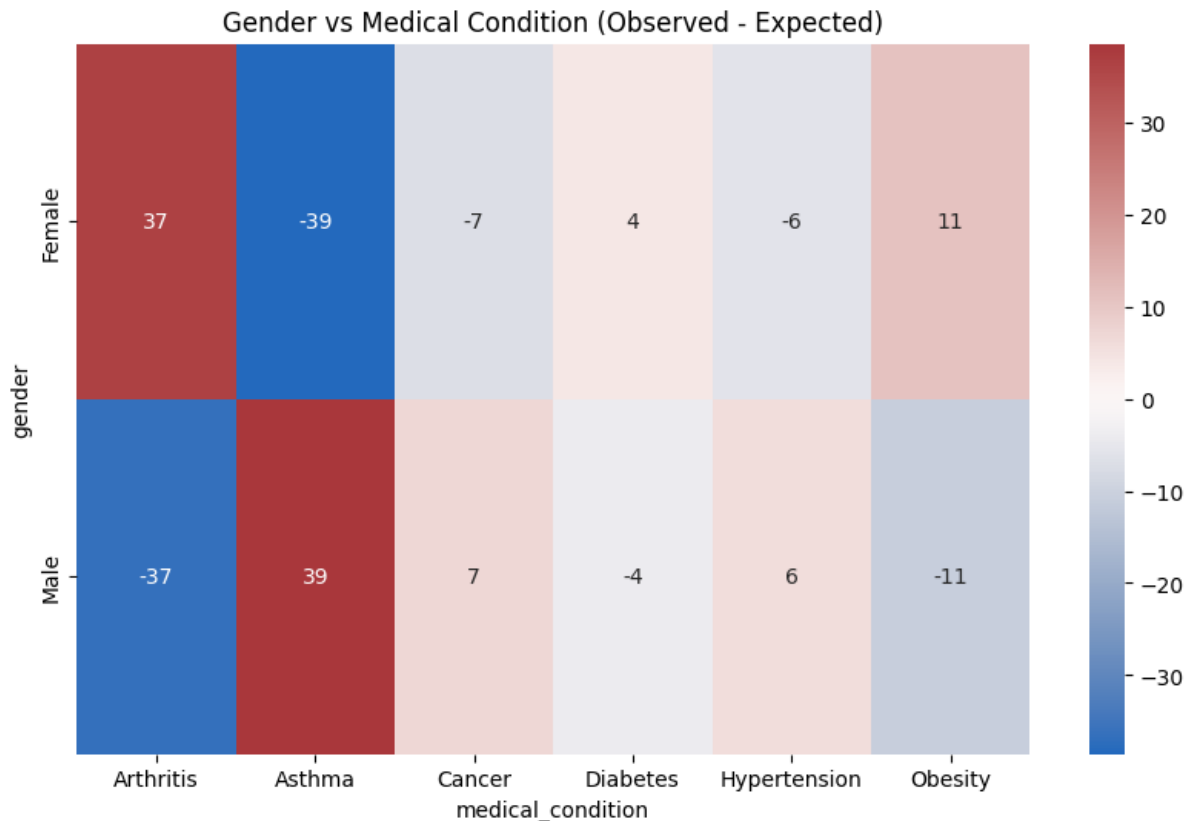
### Test Results:

Chi-Square Statistic:  $\chi^2 = 1.319$

Degrees of freedom:  $df = 5$

p-value:  $p = 0.933$





In this Chi-Square test, it was observed there were more expected arthritis and asthma conditions for both male and female patients. However for the other medical conditions, it was observed there were less of the other medical conditions on the visualizations as opposed to what was expected in the dataset. With the P-Value being higher than the threshold, we fail to reject the null hypothesis as these two variables are independent.

## Relationship between Insurance Provider and Test Results

**Test Used:** Chi-Squared Test of Independence

*Null Hypothesis ( $H_0$ ):* There is no relationship between `insurance_provider` and `test_results`.

Variables are independent

*Alternative Hypothesis ( $H_1$ ):* There is a relationship between `insurance_provider` and `test_results`.

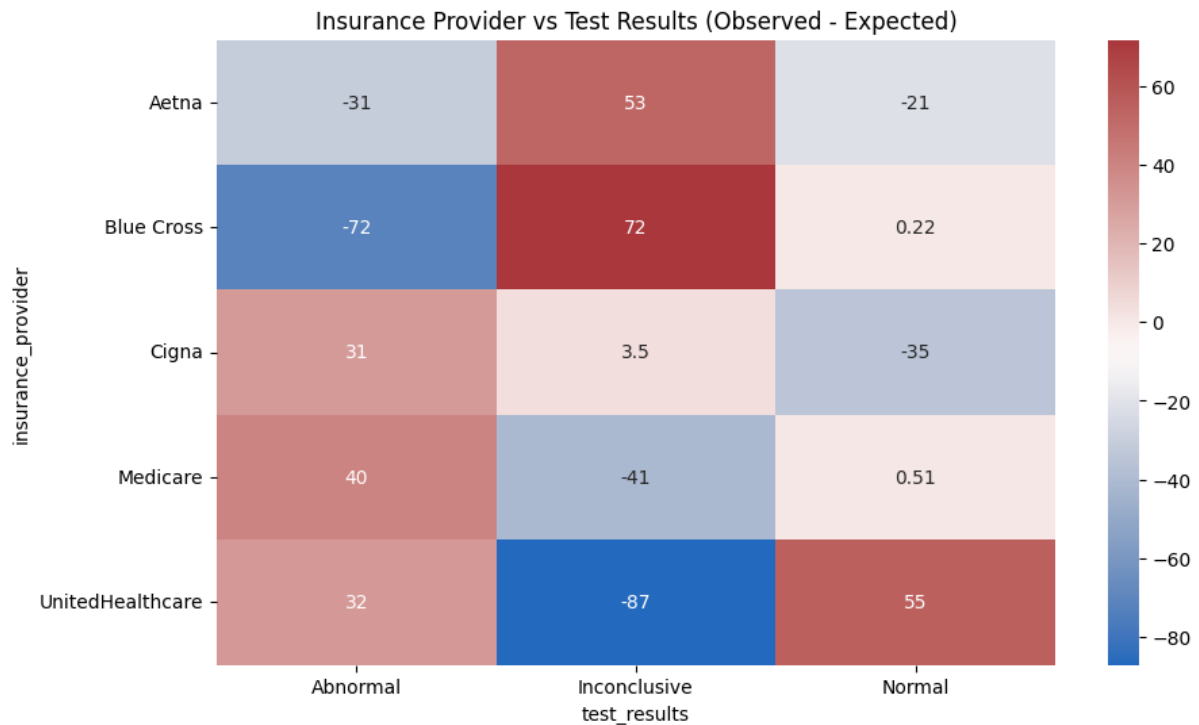
Variables are not independent

### Test Results:

Chi-Square Statistic:  $\chi^2 = 8.589$

Degrees of freedom:  $df = 8$

p-value:  $p = 0.378$



In this Chi-Square test, it is observed that certain cells have different expected results from the observed. However with the P-Value being above the threshold, it is not enough to say there is a statistic significance between the two value in which we fail to reject the null hypothesis. These two variables are independent of each other.

## Relationship between Blood Type and Test Results

**Test Used:** Chi-Squared Test of Independence

*Null Hypothesis ( $H_0$ ):* There is no relationship between `blood_type` and `test_results`.

Variables are independent

*Alternative Hypothesis ( $H_1$ ):* There is a relationship between `blood_type` and `test_results`.

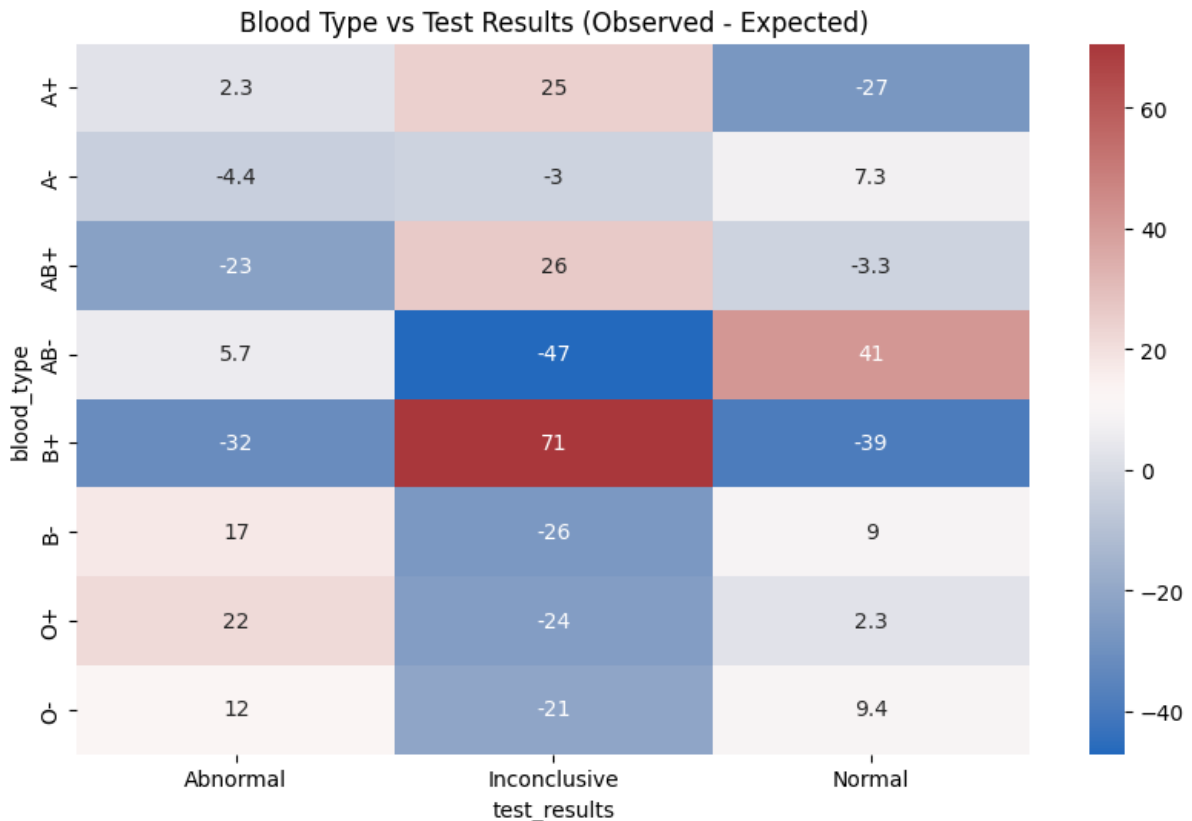
Variables are not independent

### Test Results:

Chi-Square Statistic:  $\chi^2 = 7.360$

Degrees of freedom:  $df = 14$

p-value:  $p = 0.92$



Similar to the first Chi-Square test, this frequency table only have some differences within it. Additionally the P-Value is above the threshold so we fail to reject the null hypothesis as the variables are independent of each other.

## K-Means Clustering:

### Cluster 1: Patient Groups:

cluster	0	1	2	3
age	-0.352529	-0.913075	0.857276	0.408444
treatment_days	0.886855	-0.405988	0.416829	-0.904676
billing_amount	-0.812047	0.766046	0.797381	-0.750656

In this cluster of patient demographics, the following can be interpreted:

- **Cluster 0:**
  - Age moderately below average ( $-0.35$ )
  - Treatment days strongly above average ( $0.89$ )

- Billing amount strongly below average ( $-0.81$ )
- *Interpretation: Younger patients, longer stays, lower costs*
- **Cluster 1:**
  - Age well below average ( $-0.91$ )
  - Treatment days slightly below average ( $-0.41$ )
  - Billing amount well above average ( $0.77$ )
  - *Interpretation: Very young, slightly shorter stays, high costs*
- **Cluster 2:**
  - Age strongly above average ( $0.86$ )
  - Treatment days slightly above average ( $0.42$ )
  - Billing amount strongly above average ( $0.80$ )
  - *Interpretation: Older patients, longer stays, high costs*
- **Cluster 3:**
  - Age moderately above average ( $0.41$ )
  - Treatment days well below average ( $-0.90$ )
  - Billing amount well below average ( $-0.75$ )
  - *Interpretation: Older, much shorter stays, low costs*

The clustering analysis revealed four distinct patient Interpretations based on standardized values of age, treatment duration, and billing amount:

- **Cluster 0:** Younger patients with longer treatment durations and lower-than-average billing amounts.
- **Cluster 1:** Very young patients with slightly shorter treatment durations and higher-than-average billing amounts.
- **Cluster 2:** Older patients with longer stays and higher-than-average billing amounts.
- **Cluster 3:** Older patients with short treatment durations and lower-than-average billing amounts.

## Cluster 2: Clinical Severity:

cluster	0	1	2	3
age	-0.864564	0.856610	0.865220	-0.872088
treatment_days	0.869292	0.867353	-0.864527	-0.868309
medical_condition_Asthma	0.171582	0.168306	0.162821	0.159352
medical_condition_Cancer	0.162999	0.167094	0.166667	0.168101
medical_condition_Diabetes	0.166593	0.165455	0.170875	0.167527
medical_condition_Hypertension	0.165860	0.166382	0.169859	0.164013
medical_condition_Obesity	0.164026	0.166524	0.161515	0.173049
test_results_Inconclusive	0.328859	0.329984	0.333551	0.330178
test_results_Normal	0.338395	0.331623	0.327384	0.337708
medication_Ibuprofen	0.196376	0.199587	0.203381	0.202596
medication_Lipitor	0.204005	0.195454	0.202946	0.200660
medication_Paracetamol	0.196083	0.203363	0.197504	0.200875
medication_Penicillin	0.204079	0.197449	0.195182	0.200875

In this cluster set, the following medication distribution can be interpreted:

- **Cluster 0:**

- Age well below average ( $-0.86$ )
- Treatment duration strongly above average ( $0.87$ )
- Slightly higher incidence across all medical conditions and test result categories
- Medications usage consistent across all drug types
- *Interpretation: Younger patients with extended treatments; broad but mild presence of various medical conditions*

- **Cluster 1:**

- Age well above average ( $0.86$ )
- Treatment duration also high ( $0.87$ )
- Slightly elevated rates for all medical conditions
- Medication use evenly distributed
- *Interpretation: Older patients with long treatments and balanced condition distribution*

- **Cluster 2:**

- Age well above average (0.87)
- Treatment duration below average (−0.86)
- Medical and test condition indicators close to mean
- Medication use marginally high for Ibuprofen and Lipitor
- *Interpretation: Older patients with shorter treatments, mild condition prevalence, and slightly higher medication usage*

- **Cluster 3:**

- Age strongly below average (−0.87)
- Treatment duration lower than average (−0.87)
- Relatively uniform condition and medication distribution
- *Interpretation: Young patients, shorter treatments, moderate or typical condition and medication patterns*

The clustering analysis identified four patient groups with distinct patterns across demographic, clinical, test, and medication features:

- The analysis found four patient groups that differ mainly by **age** and **treatment duration**.
- All groups show **above-average** rates of the listed medical conditions, test results, and medications.
- **Clusters 0 and 3:** younger patients with **shorter** stays.
- **Clusters 1 and 2:** older patients; **Cluster 1** has the oldest patients and the **longest** stays.
- Overall, **age** and **treatment days** drive the clustering; diagnoses and medications do not separate clusters strongly.

### Cluster 3: Utilization and Cost:

cluster	0	1	2	3
treatment_days	0.869292	-0.864527	-0.868309	0.867353
billing_amount	-0.003737	0.008980	0.003323	-0.008490
room_number	-0.012005	0.006902	0.003994	0.000915
admission_type_Emergency	0.331573	0.328544	0.325301	0.331481
admission_type_Urgent	0.331719	0.331592	0.338855	0.336397

In this cluster set, the following utilization and cost distribution can be interpreted:

- **Cluster 0:**

- Treatment days significantly above average (0.87)
- Billing amount close to average (slightly below,  $-0.004$ )
- Room number close to average (slightly below,  $-0.012$ )
- Slightly higher rates of Emergency and Urgent admission types ( $\sim 0.33$ )
- *Interpretation: Patients with extended treatment periods, typical billing, and room assignment, and a modestly higher frequency of urgent and emergency admissions*

- **Cluster 1:**

- Treatment days below average ( $-0.86$ )
- Billing amount slightly above average (0.009)
- Room number slightly above average (0.007)
- Slightly higher rates of Emergency and Urgent admissions ( $\sim 0.33$ )
- *Interpretation: Patients with shorter treatment durations, marginally higher bills, and slightly higher likelihood of urgent/emergency admissions*

- **Cluster 2:**

- Treatment days below average ( $-0.87$ )
- Billing amount near average (0.0033)
- Room number near average (0.004)
- Slightly higher rates of Emergency and Urgent admissions ( $\sim 0.33$ )
- *Interpretation: Patients with shorter stays, average billing, and consistent patterns in admission types*

- **Cluster 3:**

- Treatment days significantly above average (0.87)
- Billing amount slightly below average ( $-0.0085$ )
- Room number near average (0.0009)
- Slightly higher rates of Emergency and Urgent admissions ( $\sim 0.33$ )
- *Interpretation: Patients with longer treatment durations, slightly lower billing, and common urgent/emergency admissions*

The clustering analysis segmented patients into four groups using standardized features for treatment duration, billing amount, room assignment, and admission type:

- **Cluster 0** and **Cluster 3:** patients with notably longer treatment durations.
- **Cluster 1** and **Cluster 2:** patients with notably shorter treatment durations.

- **Billing amount** and **room number**: values remain close to the dataset mean across all clusters.
- **Admission types** Emergency and Urgent: occur at similar rates in every cluster.
- **Conclusion**: treatment duration is the primary feature separating clusters; other variables have minimal impact.

## Cluster 4: Insurance and Payment:

cluster	0	1	2	3
<b>billing_amount</b>	-0.869003	-0.860583	0.866251	0.865862
<b>treatment_days</b>	0.868146	-0.863421	-0.869415	0.868472
<b>insurance_provider_Blue Cross</b>	0.200345	0.197393	0.201207	0.198474
<b>insurance_provider_Cigna</b>	0.202143	0.203114	0.202357	0.202980
<b>insurance_provider_Medicare</b>	0.194017	0.206879	0.194021	0.209084
<b>insurance_provider_UnitedHealthcare</b>	0.206242	0.200652	0.199698	0.195349

In this cluster set, the following insurance and payment distribution can be interpreted:

- **Cluster 0:**
  - Low billing amount ( $-0.87$ )
  - High treatment days ( $+0.87$ )
  - Patients in this group tend to stay longer but incur lower costs.
  - Insurance distribution is fairly even, but with slightly higher Blue Cross and UnitedHealthcare representation.
- **Cluster 1:**
  - Low billing amount ( $-0.86$ )
  - Low treatment days ( $-0.86$ )
  - These patients have short stays and low costs—possibly routine or minor cases.
  - Slightly higher Medicare representation.
- **Cluster 2:**
  - High billing amount ( $+0.87$ )
  - Low treatment days ( $-0.87$ )
  - High-cost, short-duration treatments—could be intensive outpatient procedures or expensive diagnostics.
  - Insurance mix is similar across providers.



- **Cluster 3:**

- High billing amount (+0.87)
- High treatment days (+0.87)
- These are your high-utilization patients—long stays and high costs.
- Slightly higher Blue Cross and Medicare presence.

The clustering analysis revealed four distinct patient groups based on billing amount, treatment duration, and insurance type:

- **Cluster 0:** Longer stays, lower costs; more Blue Cross or UnitedHealthcare.
- **Cluster 1:** Short stays, low costs; more Medicare; likely routine cases.
- **Cluster 2:** Short but expensive treatments; insurance mix similar to others.
- **Cluster 3:** Long stays, high costs; more Blue Cross or Medicare.