# HEALTHCARE DATASET ANALYSIS

These visuals and analyses provide key insights into patient demographics, medical conditions, and hospital management. Age, gender, and blood type distributions outline the patient profile, while billing patterns linked to medical conditions highlight costly treatments. Admission types and insurance coverage data reveal patient demand trends and coverage preferences. Analysis of medications and test results sheds light on common treatments and associated outcomes. Together, these insights offer data-driven guidance for hospital management, billing strategies, insurance coordination, and improved treatment planning

# OBJECTIVES

### Examine the distribution of patients according to gender

### Analyse patient's medical billing per Age Group

### Analyse the various age group and billing amount of patients

### Analyse the count of Medical Condition of patients

### Explore Average billing amount by blood type per patients

### Examine Average billing amount by medical condition

### Analyse admission type in relation to blood type, medical condition and billing amount

### Analyse patients average by billing by gender

### Average Billing Amount by Medical Condition

### Reiew the distribution of patientss across different hospitals and insurance providers

In [197…
```
# packages
import pandas as pd
import numpy as np
import seaborn as sns
import datetime
from matplotlib import pyplot as plt
```

In [ ]:
```
df = pd.read_csv('healthcare_dataset.csv')
```

In [ ]:
```
#DATA EXPLORATION
```

In [151…
```
#First 5 rows of the data
df.head(5)
```

Out[151…

| | Name | Age | Gender | Blood Type | Medical Condition | Date of Admission | Doctor | Hospital | Insurar Provic |
|---|---|---|---|---|---|---|---|---|---|
| **0** | Bobby JacksOn | 30 | Male | B- | Cancer | 2024-01-31 | Matthew Smith | Sons and Miller | B Cr |
| **1** | LesLie TErRy | 62 | Male | A+ | Obesity | 2019-08-20 | Samantha Davies | Kim Inc | Medic |
| **2** | DaNnY sMitH | 76 | Female | A- | Obesity | 2022-09-22 | Tiffany Mitchell | Cook PLC | Aet |
| **3** | andrEw waTtS | 28 | Female | O+ | Diabetes | 2020-11-18 | Kevin Wells | Hernandez Rogers and Vang, | Medic |
| **4** | adrIENNE bEll | 43 | Female | AB+ | Cancer | 2022-09-19 | Kathleen Hanna | White-White | Aet |

In [31]:
```
#data info
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 55500 entries, 0 to 55499
Data columns (total 15 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   Name                55500 non-null  object 
 1   Age                 55500 non-null  int64  
 2   Gender              55500 non-null  object 
 3   Blood Type          55500 non-null  object 
 4   Medical Condition   55500 non-null  object 
 5   Date of Admission   55500 non-null  object 
 6   Doctor              55500 non-null  object 
 7   Hospital            55500 non-null  object 
 8   Insurance Provider  55500 non-null  object 
 9   Billing Amount      55500 non-null  float64
 10  Room Number         55500 non-null  int64  
 11  Admission Type      55500 non-null  object 
 12  Discharge Date      55500 non-null  object 
 13  Medication          55500 non-null  object 
 14  Test Results        55500 non-null  object 
dtypes: float64(1), int64(2), object(12)
memory usage: 6.4+ MB
```

In [35]:
```
#data size
df.size
```

Out[35]: 832500

In [39]:
```
#data shape
df.shape
```

Out[39]: (55500, 15)

In [45]:
```
# Data column name
df.columns
```

Out[45]:
```
Index(['Name', 'Age', 'Gender', 'Blood Type', 'Medical Condition',
       'Date of Admission', 'Doctor', 'Hospital', 'Insurance Provider',
       'Billing Amount', 'Room Number', 'Admission Type', 'Discharge Date',
       'Medication', 'Test Results'],
      dtype='object')
```

In [54]:
```
# Data NA
df.isna().sum()
```

Out[54]:
```
Name                  0
Age                   0
Gender                0
Blood Type            0
Medical Condition     0
Date of Admission     0
Doctor                0
Hospital              0
Insurance Provider    0
Billing Amount        0
Room Number           0
Admission Type        0
Discharge Date        0
Medication            0
Test Results          0
Age Group             0
dtype: int64
```

In [60]:
```
#check for Duplicate
df.duplicated('Name').sum()
```

Out[60]:
```
5508
```

In [62]:
```
#Drop duplicate
df.drop_duplicates(subset = 'Name', inplace = True)
df.duplicated('Name').sum()
```

Out[62]:
```
0
```

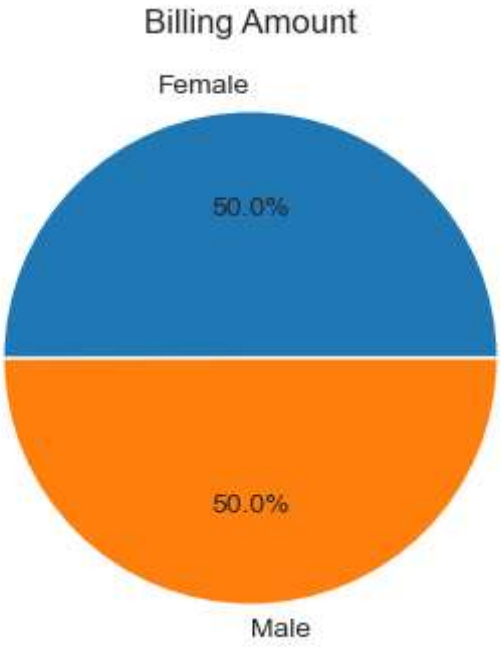
In [64]:
```
#data description
df.describe()
```

Out[64]:

| | Age | Billing Amount | Room Number |
|---|---|---|---|
| **count** | 49992.000000 | 49992.000000 | 49992.000000 |
| **mean** | 51.579453 | 25555.725277 | 301.020063 |
| **std** | 19.581816 | 14215.988133 | 115.229124 |
| **min** | 18.000000 | -2008.492140 | 101.000000 |
| **25%** | 35.000000 | 13239.403094 | 202.000000 |
| **50%** | 52.000000 | 25541.302839 | 302.000000 |
| **75%** | 68.000000 | 37853.996819 | 400.000000 |
| **max** | 85.000000 | 52764.276736 | 500.000000 |

In [72]:
```
# Visualisation
sns.set_style('darkgrid')
plt.figure(figsize=(6,4))
plt.pie(df['Gender'].value_counts(),labels = df['Gender'].value_counts().index,
plt.title('Billing Amount')
```

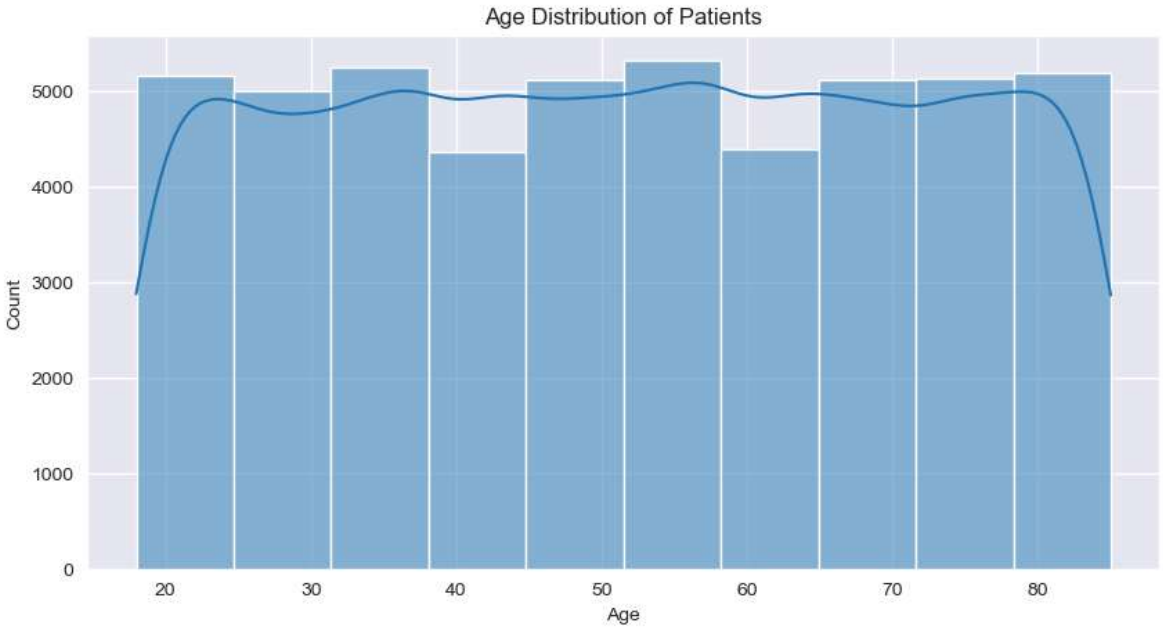
Out[72]: Text(0.5, 1.0, 'Billing Amount')




In [130…
```
# Data by medical condition
groupby_condition = df.groupby('Medical Condition')['Billing Amount'].mean().res
print(groupby_condition)
```

| | Medical Condition | Billing Amount |
|---|---|---|
| 0 | Arthritis | 25497.327056 |
| 1 | Asthma | 25635.249359 |
| 2 | Cancer | 25161.792707 |
| 3 | Diabetes | 25638.405577 |
| 4 | Hypertension | 25497.095761 |
| 5 | Obesity | 25805.971259 |

In [110…
```
# Age Distribution
plt.figure(figsize=(10, 5))
sns.histplot(df['Age'], bins=10, kde=True)
plt.title('Age Distribution of Patients')
plt.xlabel('Age')
plt.ylabel('Count')
plt.show()
```




In [88]:
```
# Distibution of Age Group
df['Age Group'] = pd.cut(df['Age'], bins=[0,18,25,35,45,55,65,float('inf')],
                         labels=['Under 18', '18-24', '25-34', '35-44', '45-54',
                         include_lowest=True)

df.head(5)
```

Out[88]:

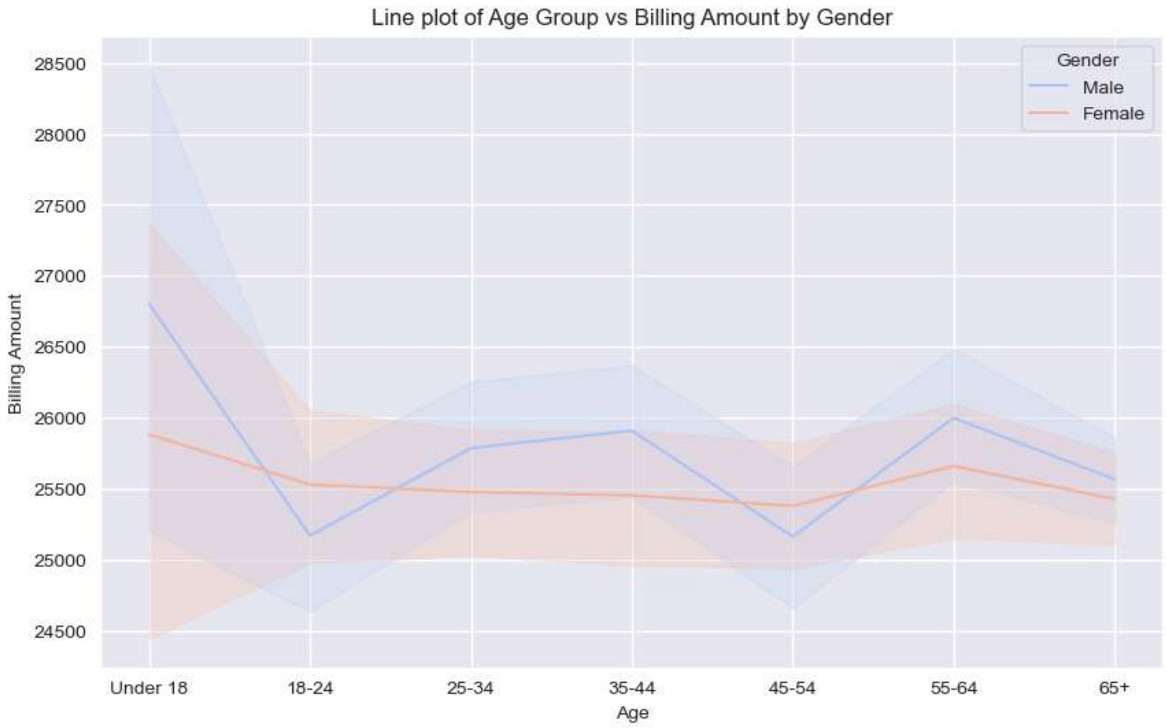
| | Name | Age | Gender | Blood Type | Medical Condition | Date of Admission | Doctor | Hospital | Insurar Provic |
|---|---|---|---|---|---|---|---|---|---|
| **0** | Bobby JacksOn | 30 | Male | B- | Cancer | 2024-01-31 | Matthew Smith | Sons and Miller | B Cr |
| **1** | LesLie TErRy | 62 | Male | A+ | Obesity | 2019-08-20 | Samantha Davies | Kim Inc | Medic |
| **2** | DaNnY sMitH | 76 | Female | A- | Obesity | 2022-09-22 | Tiffany Mitchell | Cook PLC | Aet |
| **3** | andrEw waTtS | 28 | Female | O+ | Diabetes | 2020-11-18 | Kevin Wells | Hernandez Rogers and Vang, | Medic |
| **4** | adrIENNE bEll | 43 | Female | AB+ | Cancer | 2022-09-19 | Kathleen Hanna | White-White | Aet |

In [90]:
```
# Age, Room number and Billing Amount Distrinution
x = df[['Age', 'Room Number', 'Billing Amount']]
print(x.head(7))
```

```
   Age  Room Number  Billing Amount
0   30          328    18856.281306
1   62          265    33643.327287
2   76          205    27955.096079
3   28          450    37909.782410
4   43          458    14238.317814
5   36          389    48145.110951
6   21          389    19580.872345
```

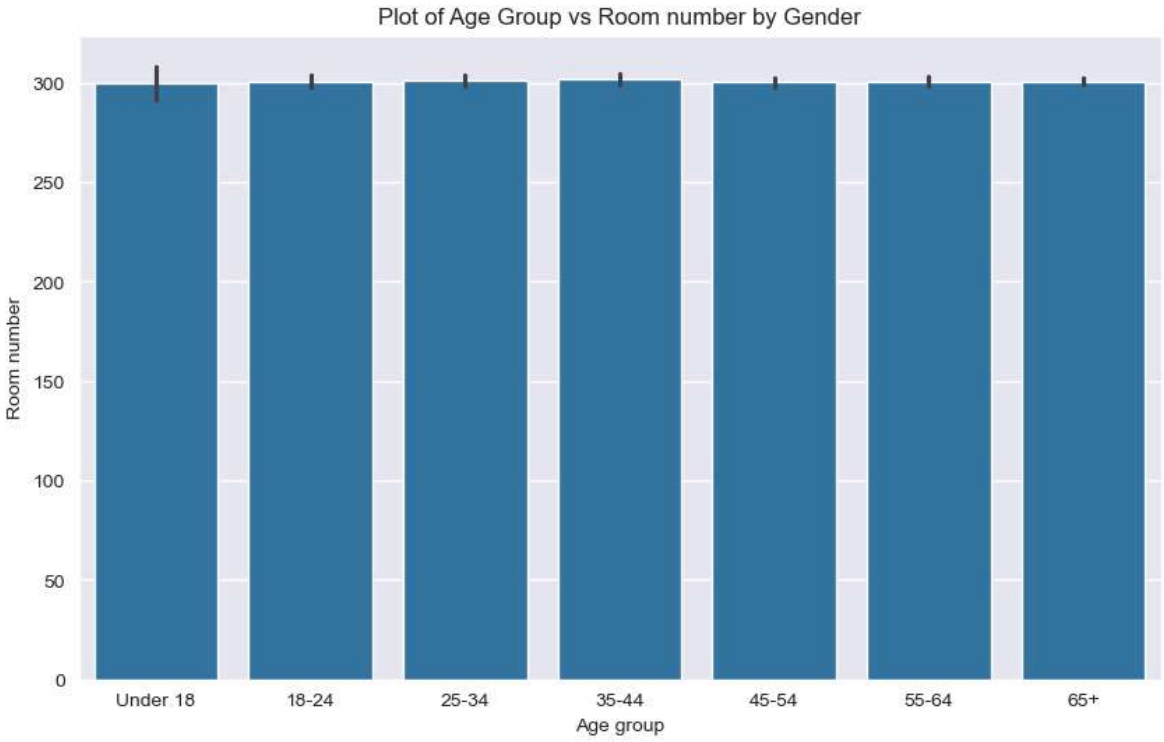
In [98]:
```
# Billing_Amount According to Age Group
x = df[['Age Group', 'Billing Amount']]
print(x.head(8))

# BILLING PER AGE
plt.figure(figsize=(10, 6))
sns.lineplot(data=df, x='Age Group', y='Billing Amount', hue='Gender', palette='
plt.title('Line plot of Age Group vs Billing Amount by Gender')
plt.xlabel('Age')
plt.ylabel('Billing Amount')
plt.show()
```

```
  Age Group  Billing Amount
0     25-34    18856.281306
1     55-64    33643.327287
2       65+    27955.096079
3     25-34    37909.782410
4     35-44    14238.317814
5     35-44    48145.110951
6     18-24    19580.872345
7     18-24    45820.462722
```




In [106…
```
# Distribution of AgeGroup by Room Number
x = df[['Age Group', 'Room Number']]
print(x.head(8))

# Room number PER Age Group
plt.figure(figsize=(10, 6))
sns.barplot(data=df, x='Age Group', y='Room Number')
plt.title('Plot of Age Group vs Room number by Gender')
plt.xlabel('Age group')
plt.ylabel('Room number')
plt.show()
```

```
  Age Group  Room Number
0     25-34          328
1     55-64          265
2       65+          205
3     25-34          450
4     35-44          458
5     35-44          389
6     18-24          389
7     18-24          277
```

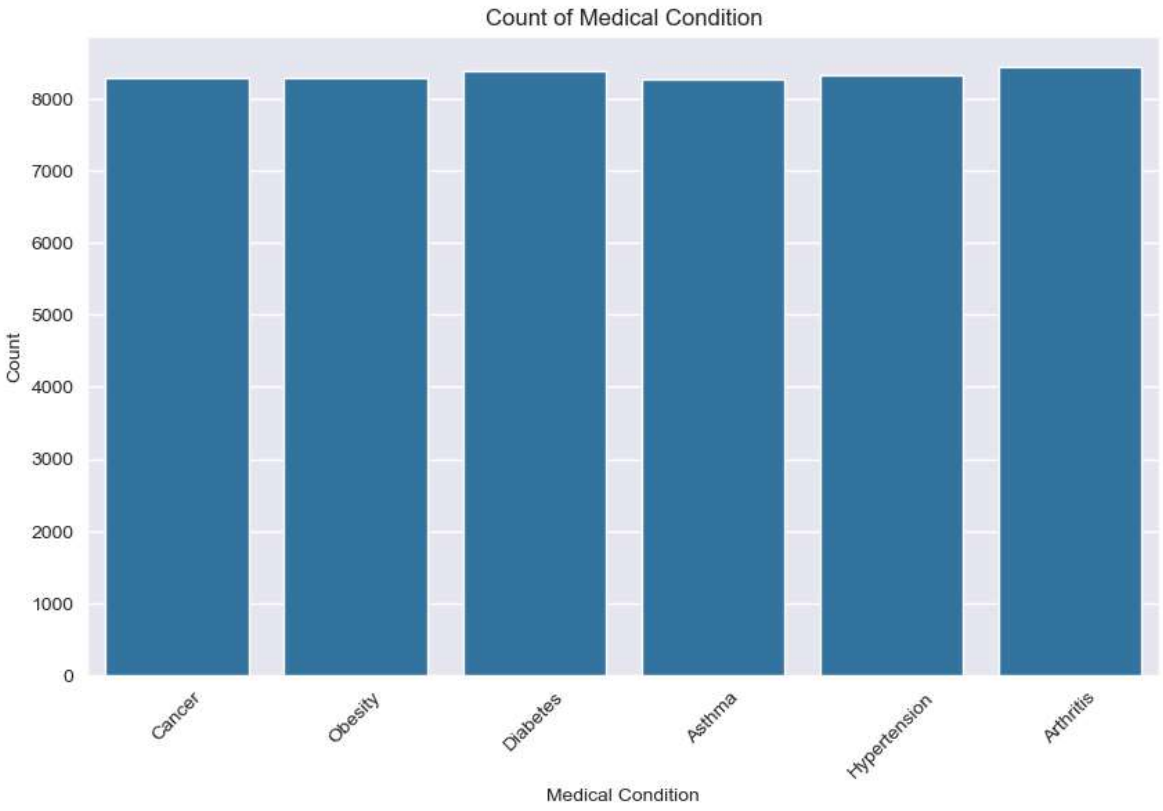


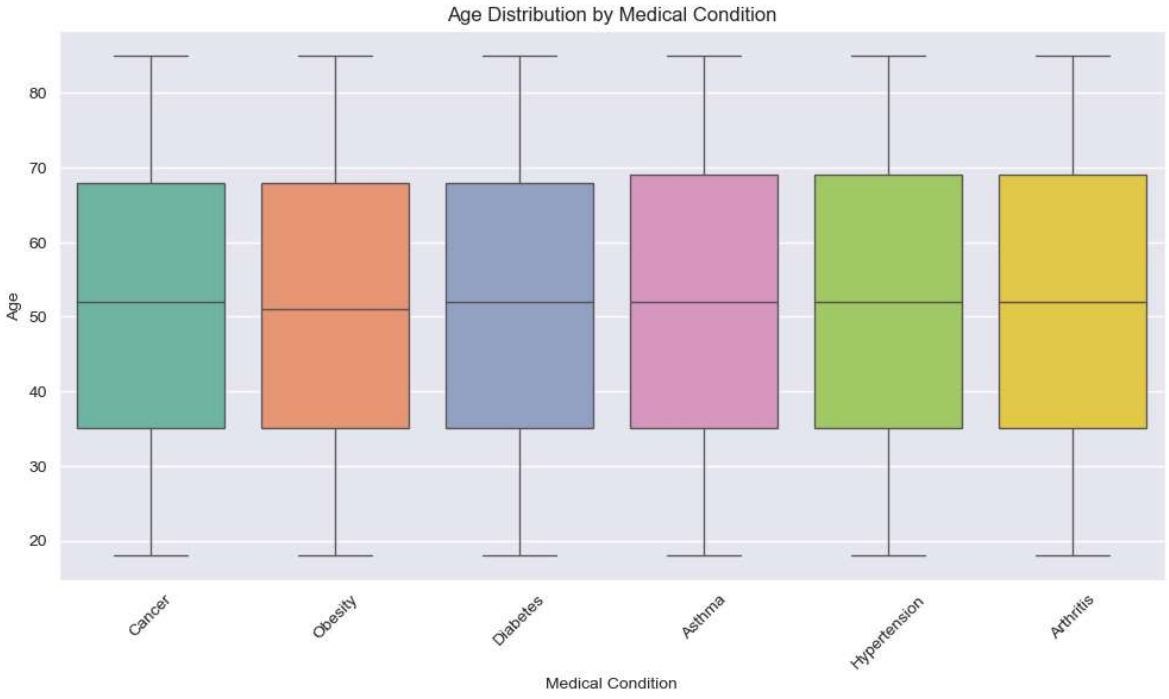

In [182…
```
# DISTRIBUTION BY HOSPITAL
groupby_hospital = df.groupby('Hospital')['Billing Amount'].mean().reset_index()
print(groupby_hospital)
```

```
                          Hospital  Billing Amount
0                       Abbott Inc    38052.041917
1                       Abbott Ltd    29877.586483
2        Abbott Moore and Williams,    24799.596339
3     Abbott and Thompson, Sullivan    16738.569765
4        Abbott, Peters and Hoffman    18842.396863
...                             ...             ...
39871            and Zimmerman Sons    32706.652625
39872     and Zuniga Davis Carlson,    42867.041298
39873  and Zuniga Francis Peterson,    33689.630726
39874               and Zuniga Sons    33950.170483
39875    and Zuniga Thompson, Blake    22067.428763

[39876 rows x 2 columns]
```

In [118…
```
# COUNT OF MEDICAL CONDITION
plt.figure(figsize=(10, 6))
sns.countplot(data=df, x='Medical Condition')
plt.title('Count of Medical Condition')
plt.xlabel('Medical Condition')
plt.ylabel('Count')
plt.xticks(rotation=45)
plt.show()
```

```
# Age Distribution by Medical Condition
plt.figure(figsize=(12, 6))
sns.boxplot(data=df, x='Medical Condition', y='Age', hue='Medical Condition', pa
plt.title('Age Distribution by Medical Condition')
plt.xlabel('Medical Condition')
plt.ylabel('Age')
plt.xticks(rotation=45)
plt.show()
```



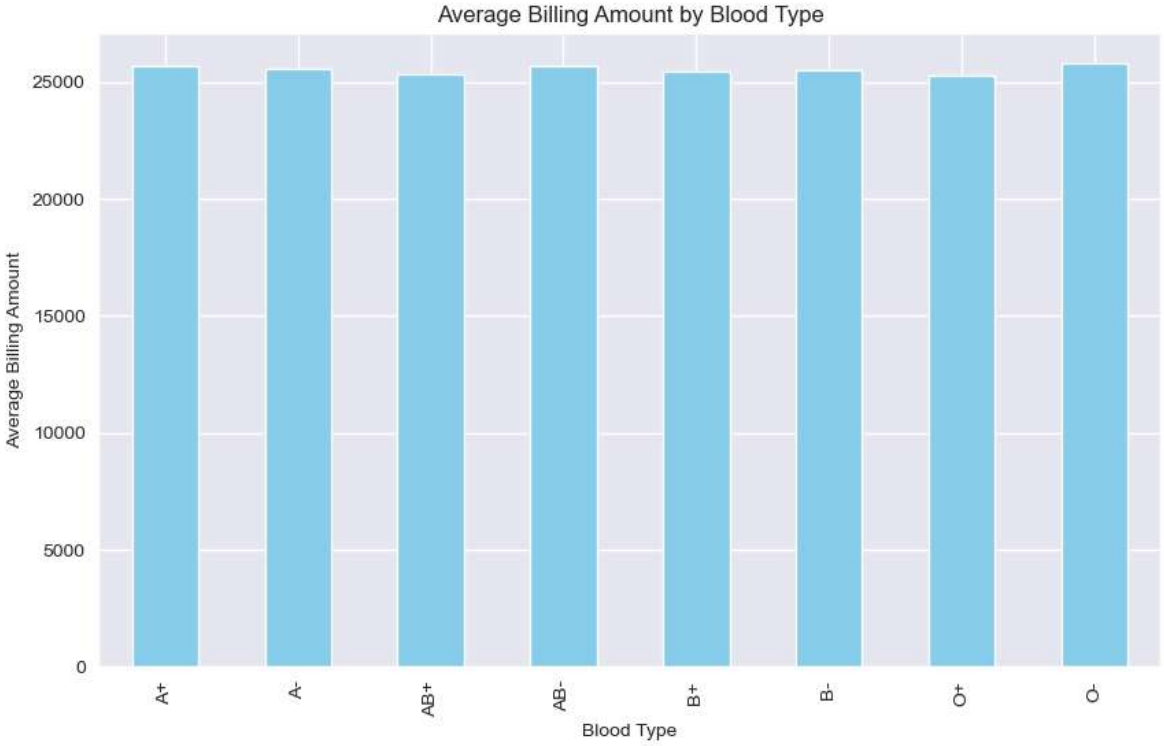



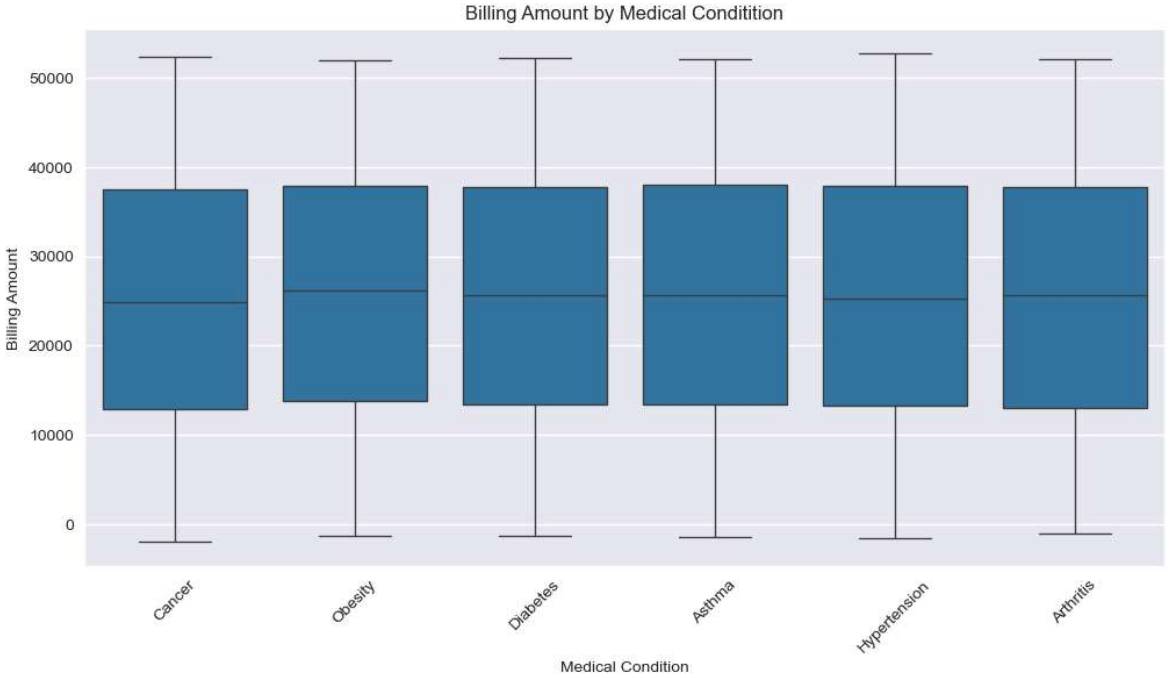

In [161…

```
# Average Billing Amount by Blood Type
avg_billing_by_blood_type = df.groupby('Blood Type')['Billing Amount'].mean()
print(avg_billing_by_blood_type)

#Average Billing by Blood Type
```

```
plt.figure(figsize=(10, 6))
avg_billing_by_blood_type.plot(kind='bar', color='skyblue')
plt.title('Average Billing Amount by Blood Type')
plt.xlabel('Blood Type')
plt.ylabel('Average Billing Amount')
plt.show()

#Billing Amount by Medical Condition
plt.figure(figsize=(12,6))
sns.boxplot(data=df, x='Medical Condition', y='Billing Amount')
plt.title('Billing Amount by Medical Conditition')
plt.xlabel('Medical Condition')
plt.ylabel('Billing Amount')
plt.xticks(rotation=45)
plt.show()
```

```
Blood Type
A+     25664.566404
A-     25595.024701
AB+    25361.458784
AB-    25694.933091
B+     25429.723237
B-     25524.424636
O+     25249.740696
O-     25795.657833
Name: Billing Amount, dtype: float64
```

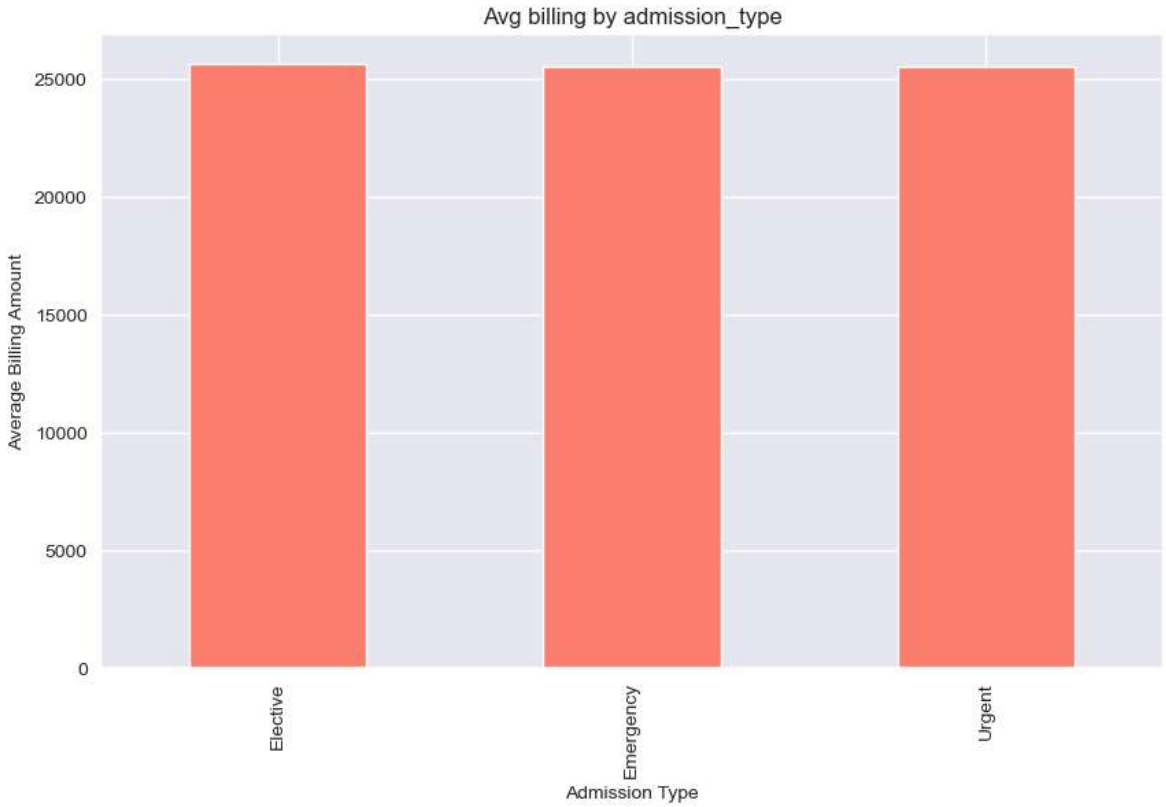





In [201…

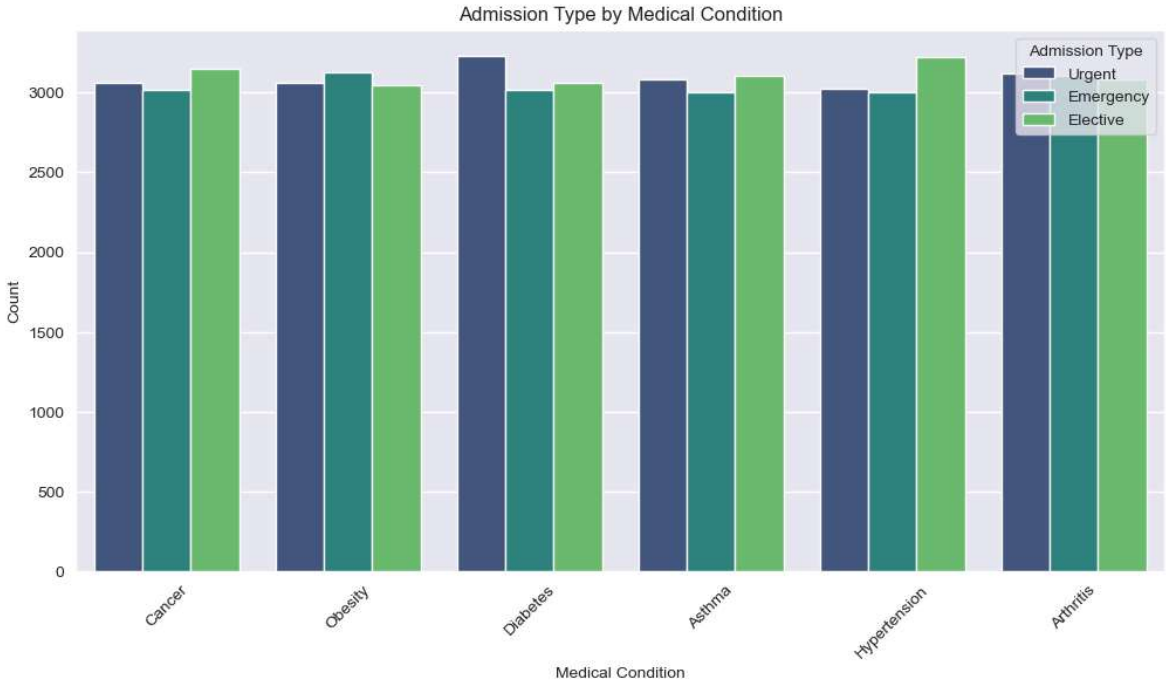
```
# AVERAGE BILLING AMOUNT BY ADMISSION TYPE
avg_billing_by_admission_type =df.groupby('Admission Type')['Billing Amount'].me
print(avg_billing_by_admission_type)

#ploting
plt.figure(figsize=(10, 6))
avg_billing_by_admission_type.plot(kind='bar', color="salmon")
plt.title('Avg billing by admission_type')
plt.xlabel('Admission Type')
plt.ylabel('Average Billing Amount')
plt.show()
```

```
Admission Type
Elective     25602.226311
Emergency    25497.397157
Urgent       25517.364497
Name: Billing Amount, dtype: float64
```




In [205…

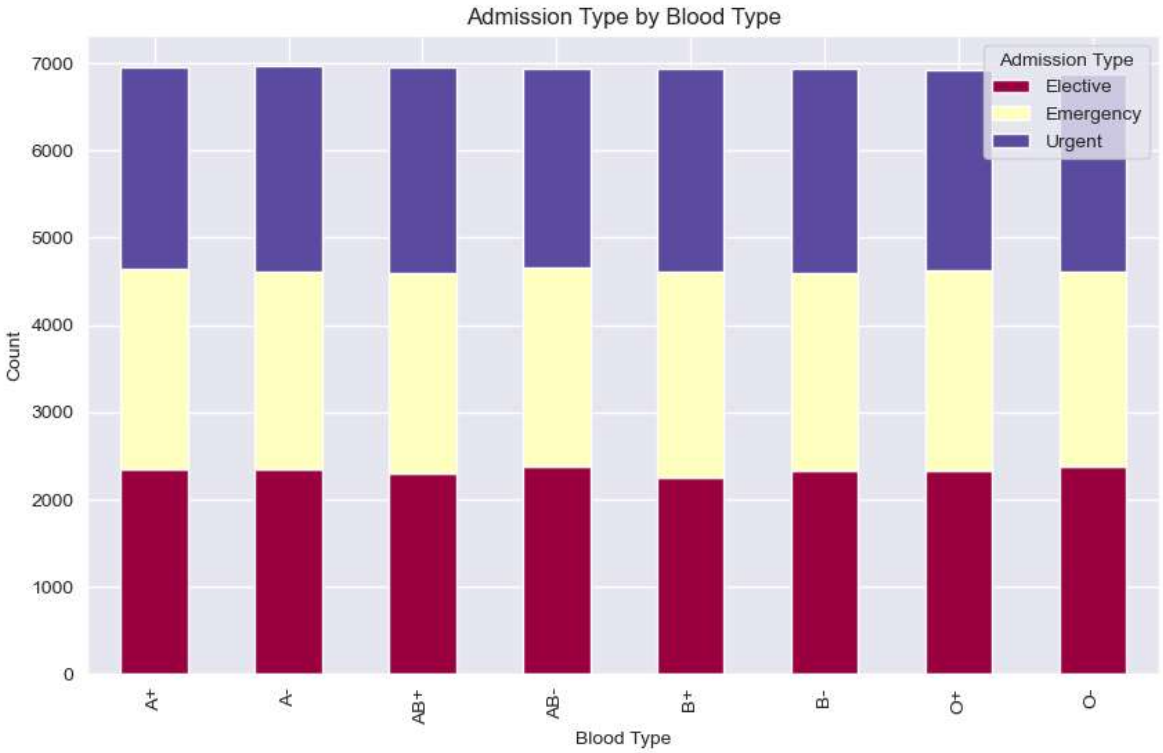
```
# Admission Type by Medical Condition
plt.figure(figsize=(12, 6))
sns.countplot(data=df, x='Medical Condition', hue='Admission Type', palette='vir
plt.title('Admission Type by Medical Condition')
plt.xlabel('Medical Condition')
plt.ylabel('Count')
plt.xticks(rotation=45)
plt.legend(title='Admission Type')
plt.show()
```




In [204…

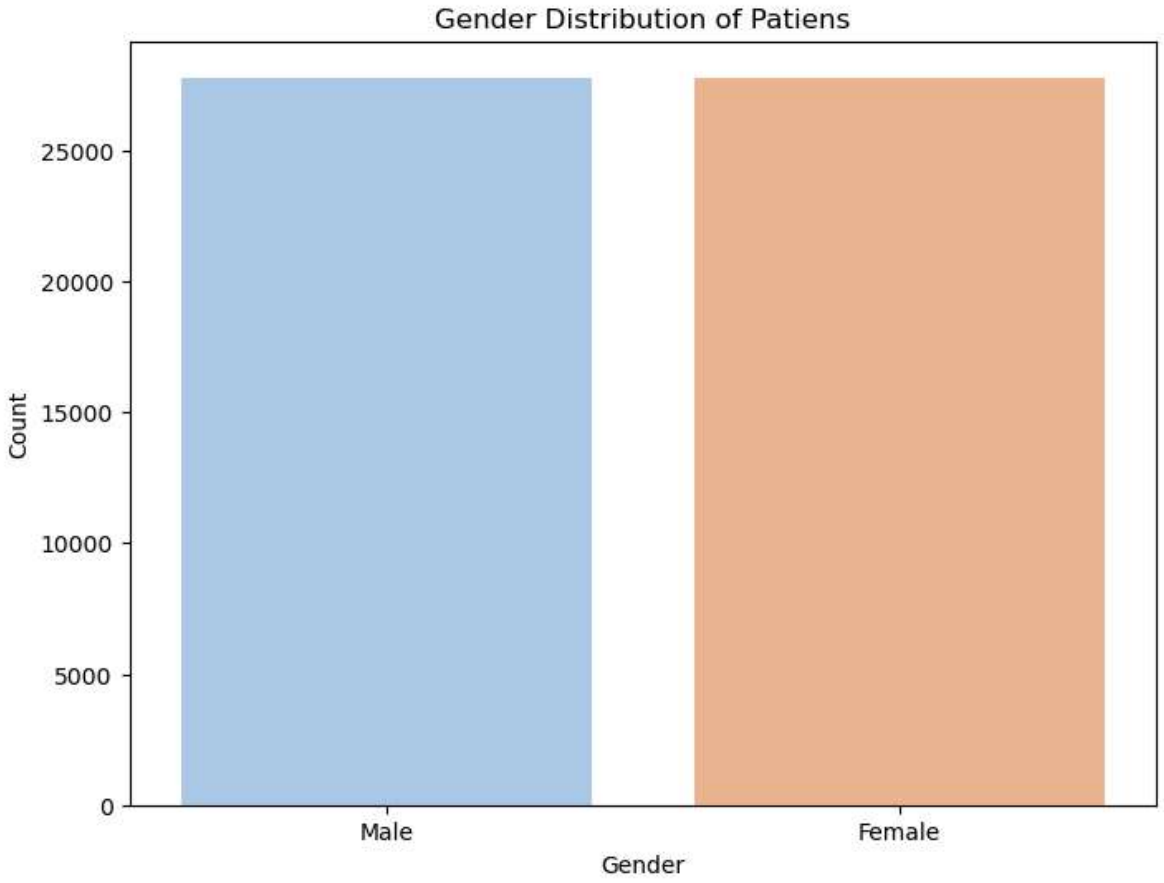
```
# Admission Type by Blood Type
admission_blood_type = pd.crosstab(df['Blood Type'], df['Admission Type'])
admission_blood_type.plot(kind='bar', stacked=True, figsize=(10, 6), colormap='S
```

```
plt.title('Admission Type by Blood Type')
plt.xlabel('Blood Type')
plt.ylabel('Count')
plt.legend(title='Admission Type')
plt.show()
```



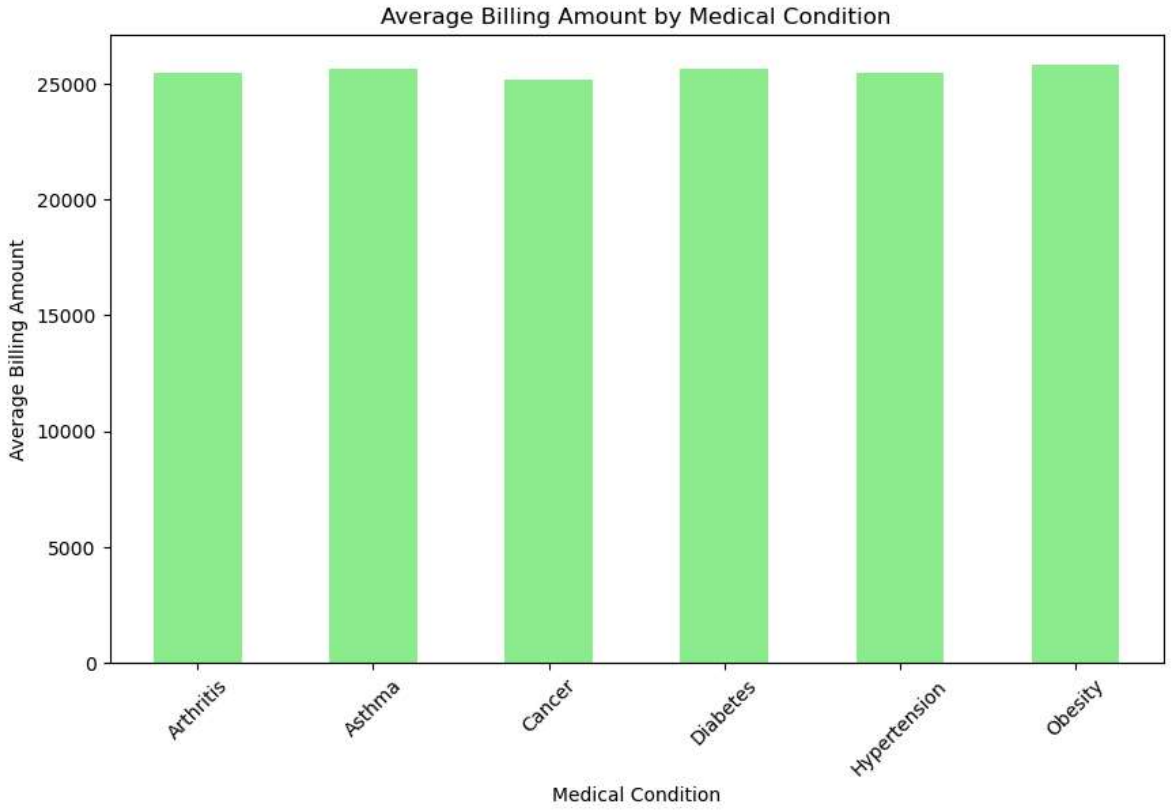

In [23]:
```
# AVERAGE BILLING BY GENDER
groupby_gender = df.groupby('Gender')['Billing Amount'].mean().reset_index()
print(groupby_gender)

#GENDER DISTRIBUTION OF PATIENTS
plt.figure(figsize=(8, 6))
sns.
t(data=df, x='Gender', hue='Gender', palette='pastel')
plt.title('Gender Distribution of Patiens')
plt.xlabel('Gender')
plt.ylabel('Count')
plt.show()
```

```
   Gender  Billing Amount
0  Female    25470.652958
1    Male    25607.860571
```



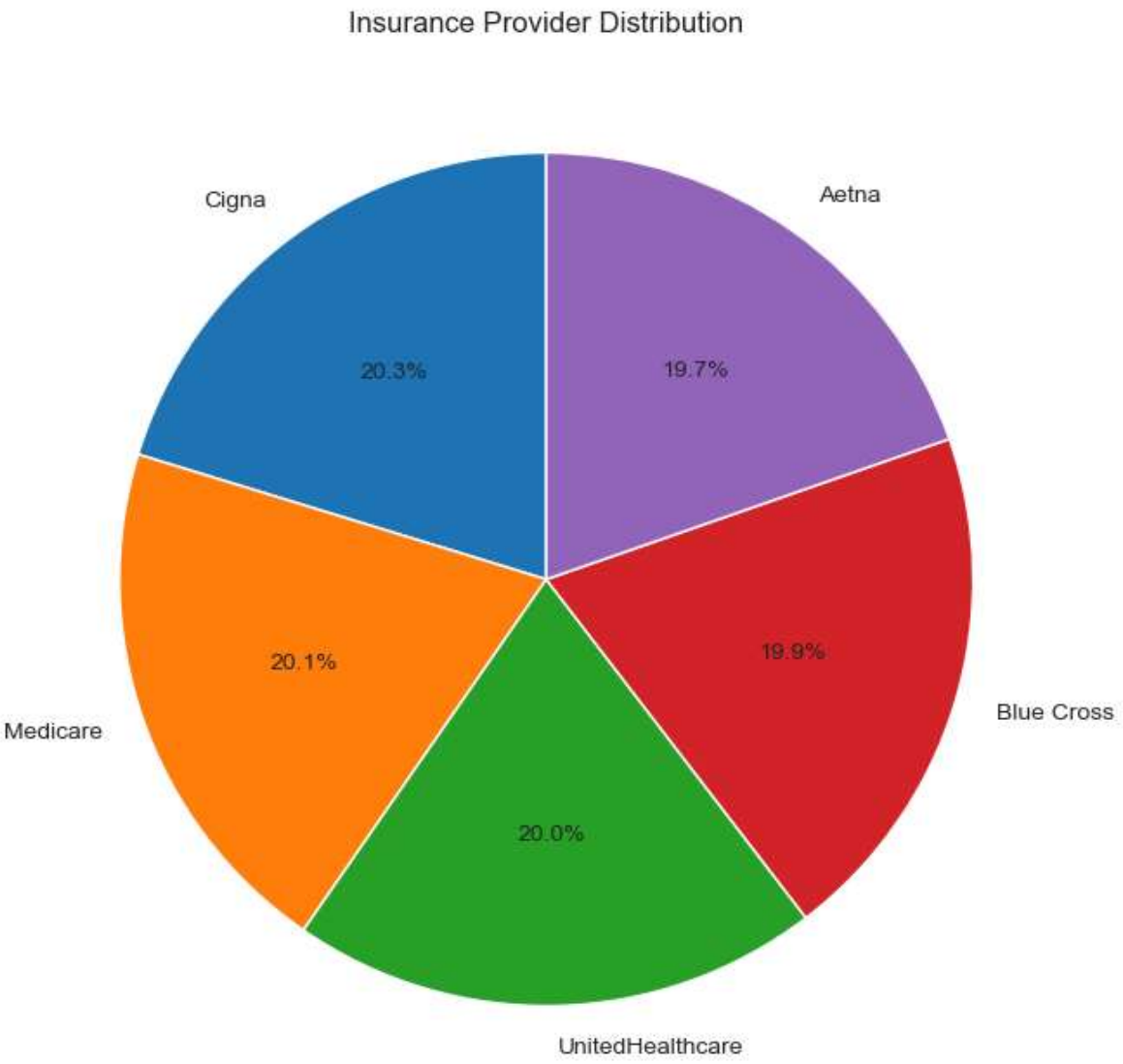

In [25]:
```
# 7. Average Billing Amount by Medical Condition
avg_billing_by_medical_condition = df.groupby('Medical Condition')['Billing Amou
print(avg_billing_by_medical_condition)

plt.figure(figsize=(10, 6))
avg_billing_by_medical_condition.plot(kind='bar', color='lightgreen')
plt.title('Average Billing Amount by Medical Condition')
plt.xlabel('Medical Condition')
plt.ylabel('Average Billing Amount')
plt.xticks(rotation=45)
plt.show()
```

```
Medical Condition
Arthritis       25497.327056
Asthma          25635.249359
Cancer          25161.792707
Diabetes        25638.405577
Hypertension    25497.095761
Obesity         25805.971259
Name: Billing Amount, dtype: float64
```



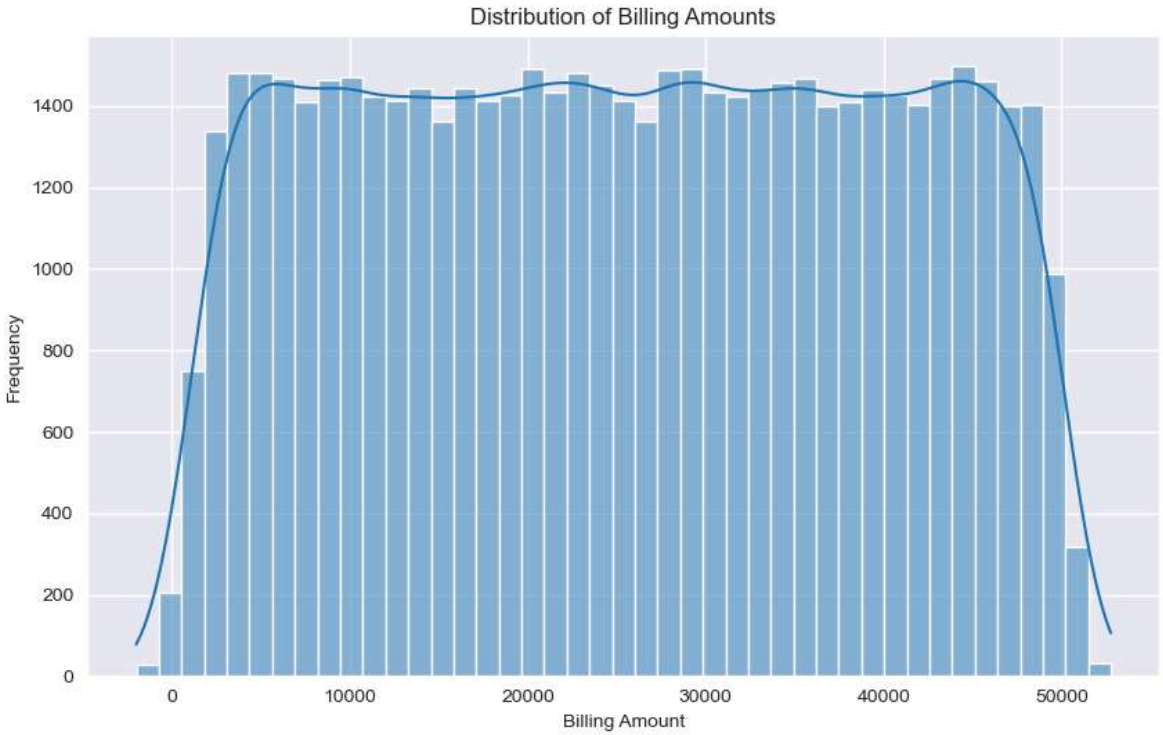

In [193…

```
# Insurance Provider Distribution
plt.figure(figsize=(8,8))
df['Insurance Provider'].value_counts().plot(kind='pie', autopct='%1.1f%%', star
plt.title('Insurance Provider Distribution')
plt.ylabel('')
plt.show()
```




In [159…

```
# Distribution of Billing Amounts
plt.figure(figsize=(10, 6))
sns.histplot(df['Billing Amount'], kde=True)
plt.title('Distribution of Billing Amounts')
plt.xlabel('Billing Amount')
plt.ylabel('Frequency')
plt.show()
```




# Key Insights, Take away and Implications

Resource Alocation: High rates of cancer, diabetes, and emergency cases need specialized staff and resources.

Cost Optimization: Billing trends by condition and insurer can highlight cost-saving areas and standardize treatments.

Insurance Collaboration: Insights on insurance and admissions can aid in better patient distribution and insurer partnerships.

Better Diagnostics: High inconclusive tests and emergency admissions suggest a need for improved diagnostics and preventive care.

Personalized Care: Variations by age, gender, and blood type support developing targeted and effective patient care.

# Data source: https://www.kaggle.com/datasets/prasad22/healthc dataset