Case Study 2: Project
PESTEL Analysis of Healthcare Sector

By: Group CS3_ 7

Fatima, Naaz (NF1020595),

Vijayan, Vishnupriya Kulathingal (NF1001530),

Olanipekun, Olumide (NF1000897),

Singh, Yuvraj (NF1016125)

Master in Data Analytics, University of Niagara Falls

Data Analytics Case Study 2 (DAMO-511-2)

Professor: Touraj Banirostam

Jun 15, 2025

Table of Contents

1 INTRODUCTION

The healthcare industry is made up of many organizations and companies that help people stay healthy by proving services related to treating patients, preventing disease, and improving overall well-being. Some of the organizations which are part of healthcare system are hospitals, medical equipment OEMs, pharmaceuticals, and health insurance providers.

In many developed countries it is seen that the healthcare industry contributes to 10% of GDP and it is considered to be one of the biggest and fastest-growing industries globally.

This industry is very important because it affects people's health and the growth of a country. Post COVID-19 pandemic, there were many changes in healthcare, new technologies were used, the cost of treatment became higher, and there were sudden changes in health insurance and rules. Hospitals being the core element of the health system as an operating arm, they provide acute and long-term care.

2 EXECUTIVE SUMMARY

Healthcare is the system which gets affected easily by many factors, thus by using PESTEL [political, economic, society, technology, economic, and legal] framework, operational risks or opportunities in healthcare sector can be identified and analyzes, emphasizing on supply chain management or labor dynamics.

Each segment of PESTEL frame works has key risks/opportunities as listed below, which would be analysed by using various statistical analysing models and visual representations to support this analysis.

Key Risks:

- Dependency on insurance provider, which can be a political risk.

- Variance in the cost based on the type of the disease or condition contributing to economical risk.
- Social risk can be in the form of aging population.
- Inconsistency in diagnostics/results can be a result of technological risk.
- Wastage of medical waste can be categorized as part of environmental risk.
- Billing frauds or data privacy forms Ligal risk.

Key opportunities:

- Insurance diversification like allowing many insurance providers to contribute can be seen as a political opportunity.
- Implementing different cost models can be an economical opportunity.
- Social opportunity visualized by targeted preventive health care based on the age/condition.
- Technology can be used for optimizing the inventory.
- By implementing certain practices in order to reduce the medical waste can be a part of environmental opportunity.
- Developing various security policies and procedures can be taken as a legal opportunity.

Based on the identification of these key risks and opportunities, statistic recommendations can be outlined, which can support the health sector to effectively understand and utilize the resources.

3 METHODOLOGY

This section describes about the data collection and data handling/cleaning.

3.1 DATA COLLECTION

The strategy for data collection was focused on identifying secondary sources and indicators for each of the six factors of the PESTEL framework.

The data set has been collected from the opensource Kaggle https://www.kaggle.com/datasets/prasad22/healthcare-dataset , as the healthcare data privacy and security is utmost concerned, the datasets available are synthetic in nature which implies this dataset does not contain any real patient information or violate any privacy regulations.

The dataset contains multiple attributes which can be used to carry out the PESTEL analysis and its relevance is explained below:

- **Political Relevance**

  The Insurance Provider and Admission Type are closely linked to the policy and political governance system. Policies that involve reimbursement procedures and budgeting raise political risk and constrain healthcare services because of Medicare dependence. Policy interest areas, including equitable healthcare access and systemic stress level, are also monitored in tandem with emergency admission trends.

- **Economic Relevance**

  Billing Amount, Medical Condition, and Age are within the primary focus of the economic analysis of the data set. These variables assist in describing the expenditure on healthcare services and help determine major cost diseases (e.g., cancer, diabetes) and the most economically disadvantaged age groups. Studying this information helps construct adequate volatile forecasts, optimize costs, and create models of bundled payment pricing for controlled economic fluctuations.

- **Social Relevance**

  Gender, Age, and Medical Condition are reflective of the prevalent social characteristics of the patient population. These variables assist to monitor the occurrence of certain diseases in particular demographic groups which is vital for primary public health resource allocation and strategic planning. The presence of many elderly patients with chronic diseases underscores the need for additional workforce deployment planning and geriatric service provision planning for aging patients.

- **Technological Relevance**

  The Use of Technology in Diagnostics and Treatment Procedures is referred to in medications and test outcomes. The existence of abnormal or inconclusive results usually signals a failure in medical infrastructure diagnostics or obsolete testing frameworks. The existence of certain medications indicates reliance on some products which also informs policy decisions regarding digital health innovation and surpluses of medical equipment.

- **Environmental Relevance**

  While not directly measuring the environment, parameters such as Length of Stay which can be derived from admission and discharge information as well as Frequency of Medication relate to the consumption of resources or waste produced. From a deeper lens, this set of variables is useful for estimating the ecological footprint of health care operations, enabling the identification of environmentally responsible change opportunities, and helping autonomously monitor compliance with environmental regulations.

- **Legal Relevance**

  Compliance and legal risk relate to Billing Amount, Insurance Provider, Doctor as well as Test Results. Gaps in billing norms or too many inconclusive results may lead to compliance scrutiny or a regulatory audit. Legal exposure exists due to differences in the types of treatment outcomes and admissions, especially when the services are funded publicly. These aspects require careful and diligent attention to maintain accuracy, transparency, compliance and for minimizing exposure to legal penalties.

Relevance Summary Table

PESTEL Factor	Dataset Fields Used
Political	Insurance Provider, Admission Type
Economic	Billing Amount, Stay Duration
Social	Age, Gender, Condition
Technological	Medication, Test Results
Environmental	Derived: Stay, Medication
Legal	Doctor, Insurance, Admission Type

3.2 DATA HANDLING AND CLEANING

Understanding the Data & Data cleaning are the most important steps in the data pre-processing steps to ensure a dataset is accurate, consistent, and trustworthy. Cleaning involves addressing missing values, correcting incorrect data, filtering duplicates, standardizing data formats e.g., dates, text fields, and units; clean data is vital for useful analysis; as data cleaning enhances the performance of statistical models, enables accurate decision making, limits the potential for misleading results. In the end, data cleaning improves the quality of outcomes gained from data and is a foundational data pre-processing step for successful analytics or machine learning work.

3.2.1 Dataset Details

The selected dataset contains 55,500 entries and 15 columns (comprising various attributes).

Dataset Name: Health Care Dataset (healthcare_dataset.csv)

Source: Kaggle: https://www.kaggle.com/datasets/prasad22/healthcare-dataset

Parameters: It involves various parameters like 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.

3.2.2 Setting up PyChram Project

Create a new Python project in PyCharm and Install required libraries (pandas, matplotlib, seaborn, and scikit-learn) in terminal.

Pandas: Used for data manipulation and analysis. Makes it easy to read, clean, filter, and transform tabular data (like CSV or Excel files).

Matplotlib: A basic data visualization library. Used to create plots, bar charts, line graphs, pie charts, etc.

Seaborn: An advanced visualization library built on top of matplotlib. Offers beautiful and informative plots with fewer lines of code.

Numpy: A library used for numerical computing.

Statsmodels.api model provides a comprehensive suite of functions and classes for statistical modeling, data exploration, and statistical tests. It enables users to fit various statistical models, conduct hypothesis tests, and explore data through descriptive statistics and visualization.

Scipy.stats module provides a comprehensive suite of statistical functions and distributions for scientific computing in Python. It offers a wide array of tools for working with probability distributions, performing hypothesis tests, and computing descriptive statistics.

Scikit-learn: A powerful library for machine learning and data mining. Provides tools for classification, regression, clustering, model evaluation, and data preprocessing.

3.2.3 Load the Data

Ensure that the dataset is available in the project directory created if not need to use the correct path of the data.

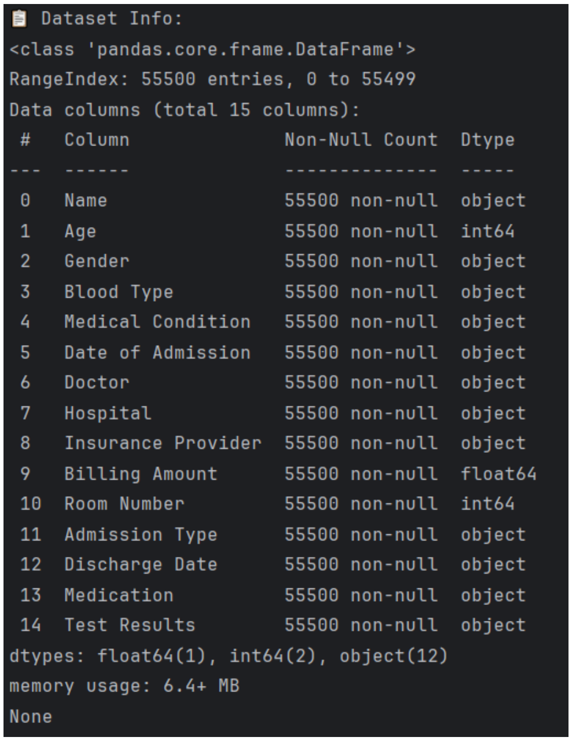
The command used to load the dataset is "**df = pd.read_csv("healthcare_dataset.csv")**", where healthcare_dataset.csv is the dataset available in the project folder.

3.2.4 Data Understanding

By using the various commands "**print(df.info())**" and "**print(df.describe())**", these commands help in fetching the summary of the Data Frame, that is showing (Column names, Number of non-null (i.e., non-missing) values, Data types (int, float, object, etc.) and Total memory usage) and statistics summary for all numerical columns (like Count, mean, standard deviation, Min, max, and quartiles (25%, 50%, 75%)).

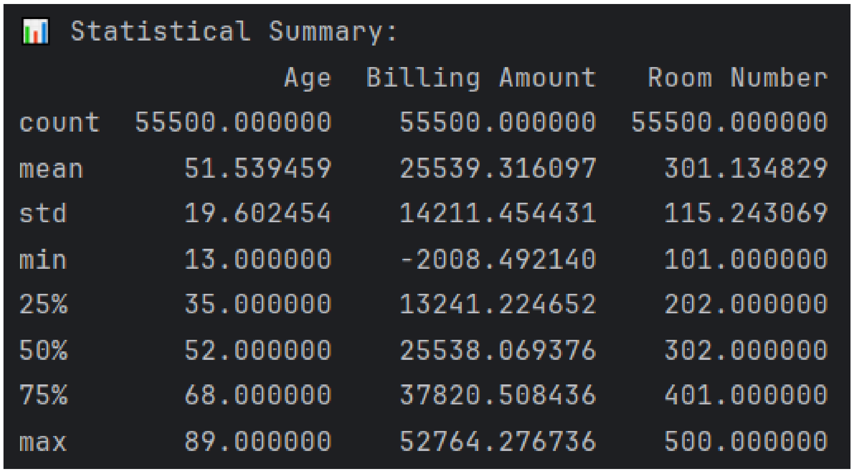
This summary information helps in identify missing values, reveals if a column has wrong data type (e.g., numbers as object), which is very important before data cleaning. The statistics summary helps in identifying outliers (via min/max), understanding variable ranges and distribution, which is useful for normalization/scaling.

The output of info command is as below:



```
📋 Dataset 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
None
```

The output of the describe command is as follows:



```
📊 Statistical Summary:
                Age  Billing Amount   Room Number
count  55500.000000    55500.000000  55500.000000
mean      51.539459    25539.316097    301.134829
std       19.602454    14211.454431    115.243069
min       13.000000    -2008.492140    101.000000
25%       35.000000    13241.224652    202.000000
50%       52.000000    25538.069376    302.000000
75%       68.000000    37820.508436    401.000000
max       89.000000    52764.276736    500.000000
```

3.2.5 Data Cleaning

The important part of the data cleaning is to identify the missing values in the dataset as these values can create negative impact or misleading results of model.

a. Standardizing the Text fields

The below command is used to ensure that the values are in string format, helps in removing leading/trailing spaces and converting text to title case, this helps in mainting the uniformity between the columns.

```
text_columns = ['Gender', 'Medical Condition', 'Medication', 'Doctor', 'Hospital']
for col in text_columns:
    df[col] = df[col].astype(str).str.strip().str.title()
```

b. Date Conversion & Validation

Generally, the dates in the dataset are stored in string format, in order to use these columns for various calculations it is essential to convert them in proper date time format. The command below is used.

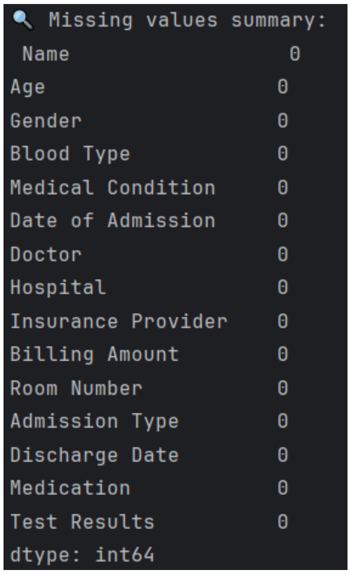
```
df['Date of Admission'] = pd.to_datetime(df['Date of Admission'], errors='coerce')
df['Discharge Date'] = pd.to_datetime(df['Discharge Date'], errors='coerce')
```

c. Identifying Missing Values

The command "**.isnull().sum()**" is used to identify how many missing (null/NaN) values are present in each column of dataset.

Command "df.isnull()" returns a DataFrame of True (if value is missing) and False (if not), command ".sum()" then adds up the True values (since True = 1, False = 0) column-wise and the full command "df.isnull().sum()" gives the total number of missing values per column.

The output of this command is:



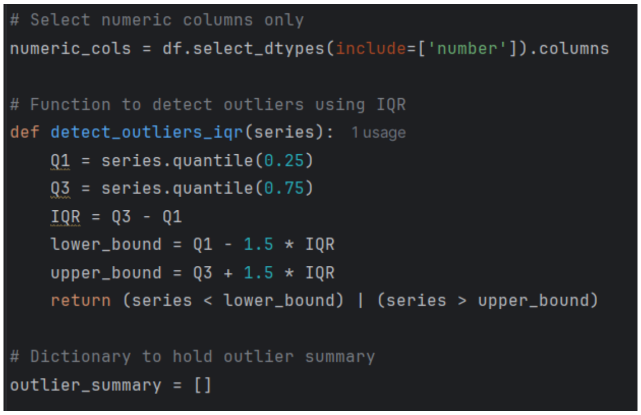
```
🔍 Missing values summary:
 Name                  0
Age                   0
Gender                0
Blood Type            0
Medical Condition     0
Date of Admission     0
Doctor                0
Hospital              0
Insurance Provider    0
Billing Amount        0
Room Number           0
Admission Type        0
Discharge Date        0
Medication            0
Test Results          0
dtype: int64
```

By analyzing the output, it is understood that data there are no missing values in any of the columns, hence no further actions like drop columns or Impute values are required.

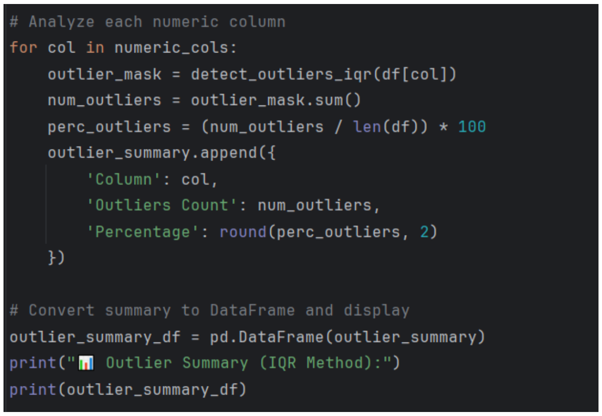
d. <u>Handling Outliers</u>

Outliers can distort model predictions, reduce accuracy, and violate assumptions like normality and homoscedasticity. Outliers are checked on all the columns with numerical values.

The code used for detecting the outlier:

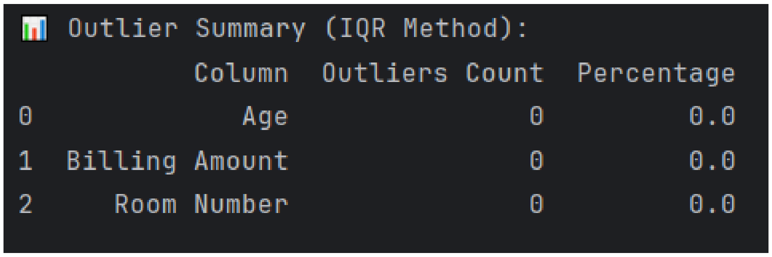


```python
# Select numeric columns only
numeric_cols = df.select_dtypes(include=['number']).columns

# Function to detect outliers using IQR
def detect_outliers_iqr(series):  1 usage
    Q1 = series.quantile(0.25)
    Q3 = series.quantile(0.75)
    IQR = Q3 - Q1
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR
    return (series < lower_bound) | (series > upper_bound)

# Dictionary to hold outlier summary
outlier_summary = []
```



```python
# Analyze each numeric column
for col in numeric_cols:
    outlier_mask = detect_outliers_iqr(df[col])
    num_outliers = outlier_mask.sum()
    perc_outliers = (num_outliers / len(df)) * 100
    outlier_summary.append({
        'Column': col,
        'Outliers Count': num_outliers,
        'Percentage': round(perc_outliers, 2)
    })

# Convert summary to DataFrame and display
outlier_summary_df = pd.DataFrame(outlier_summary)
print("📊 Outlier Summary (IQR Method):")
print(outlier_summary_df)
```

As per the output it is seen that there are no outliers detected,



```
📊 Outlier Summary (IQR Method):
           Column  Outliers Count  Percentage
0             Age               0         0.0
1  Billing Amount               0         0.0
2     Room Number               0         0.0
```

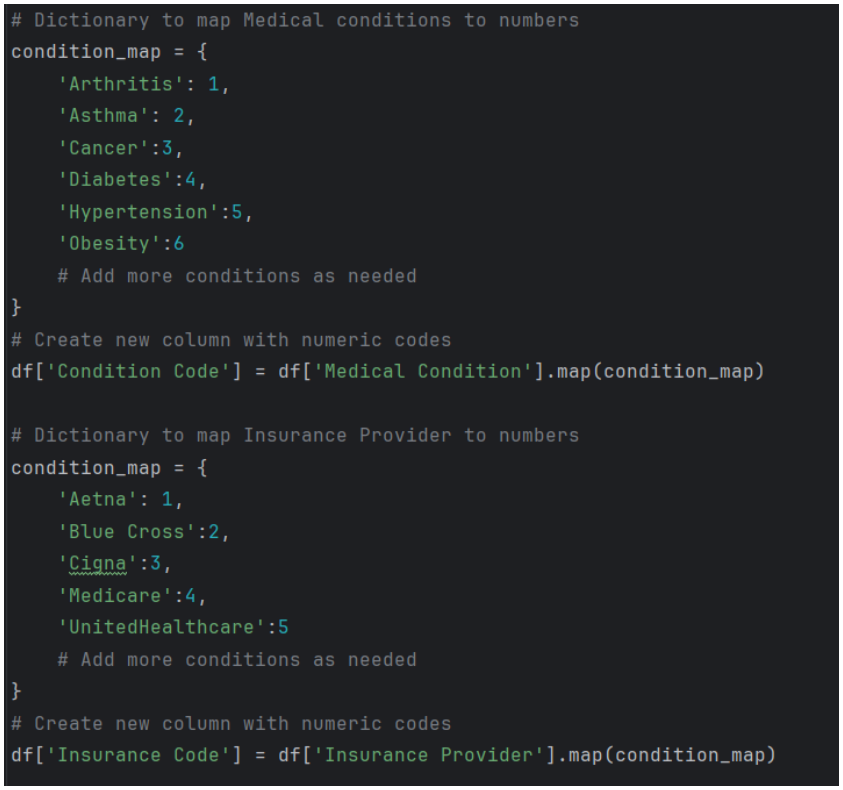
e. **Feature Engineering**

As part of data cleaning itself addition columns are created which will be helpful during the course of data analysis. The commands below are used to create the additional columns of Length of Stay, Billing per Day, Age Group, Admission Month, Admission Week and Month.

```
df['Length of Stay'] = (df['Discharge Date'] - df['Date of Admission']).dt.days
df['Billing Per Day'] = df['Billing Amount'] / df['Length of Stay'].replace(0, np.nan)
df['Age Group'] = pd.cut(df['Age'], bins=[0, 30, 50, 120], labels=['<30', '30-50', '50+'])
df['Admission Month'] = df['Date of Admission'].dt.to_period('M').astype(str)
df['Admission Weekday'] = df['Date of Admission'].dt.day_name()
df['Month'] = df['Date of Admission'].dt.to_period('M')
```

Additional columns of Condition Code & Insurance Code are created to map the details of the respective columns to numbers, so that these columns can be used in regression analysis.

The command below is used to create these columns.



```
# Dictionary to map Medical conditions to numbers
condition_map = {
    'Arthritis': 1,
    'Asthma': 2,
    'Cancer':3,
    'Diabetes':4,
    'Hypertension':5,
    'Obesity':6
    # Add more conditions as needed
}
# Create new column with numeric codes
df['Condition Code'] = df['Medical Condition'].map(condition_map)

# Dictionary to map Insurance Provider to numbers
condition_map = {
    'Aetna': 1,
    'Blue Cross':2,
    'Cigna':3,
    'Medicare':4,
    'UnitedHealthcare':5
    # Add more conditions as needed
}
# Create new column with numeric codes
df['Insurance Code'] = df['Insurance Provider'].map(condition_map)
```

f. <u>Drop invalid dates</u>

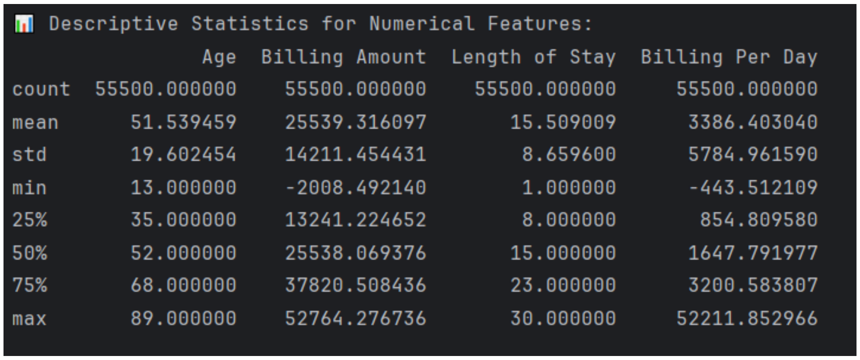
It is important to drop the invalid dates and also to ensure if discharge dates are greater than admission dates, The command below is used.

```
df = df.dropna(subset=['Date of Admission', 'Discharge Date'])
df = df[df['Discharge Date'] >= df['Date of Admission']]
```

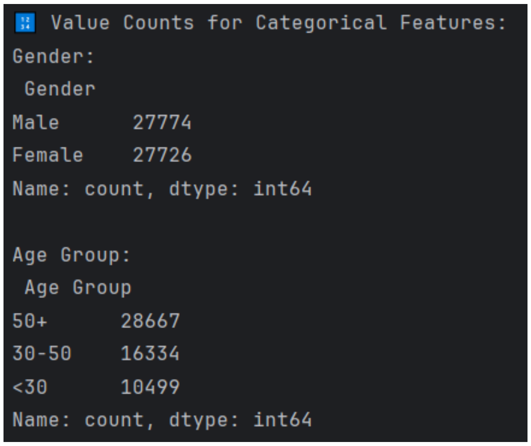
After completion of the data cleaning the revised descriptive analysis (i.e. statistical summary) is generated along with value counts of the categorial featured columns, by suing the below command:

```
print("\n📊 Descriptive Statistics for Numerical Features:")
print(df[['Age', 'Billing Amount', 'Length of Stay', 'Billing Per Day']].describe())
print("\n🔢 Value Counts for Categorical Features:")
print("Gender:\n", df['Gender'].value_counts())
print("\nAge Group:\n", df['Age Group'].value_counts())
print("\nMedical Condition:\n", df['Medical Condition'].value_counts())
```

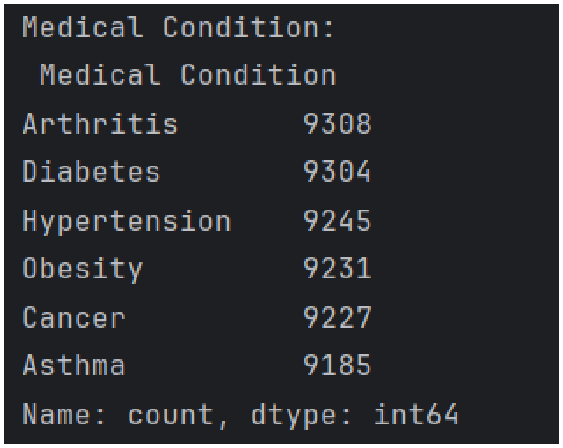
The output of this command is



```
📊 Descriptive Statistics for Numerical Features:
                Age  Billing Amount  Length of Stay  Billing Per Day
count  55500.000000    55500.000000    55500.000000     55500.000000
mean      51.539459    25539.316097       15.509009      3386.403040
std       19.602454    14211.454431        8.659600      5784.961590
min       13.000000    -2008.492140        1.000000      -443.512109
25%       35.000000    13241.224652        8.000000       854.809580
50%       52.000000    25538.069376       15.000000      1647.791977
75%       68.000000    37820.508436       23.000000      3200.583807
max       89.000000    52764.276736       30.000000     52211.852966
```



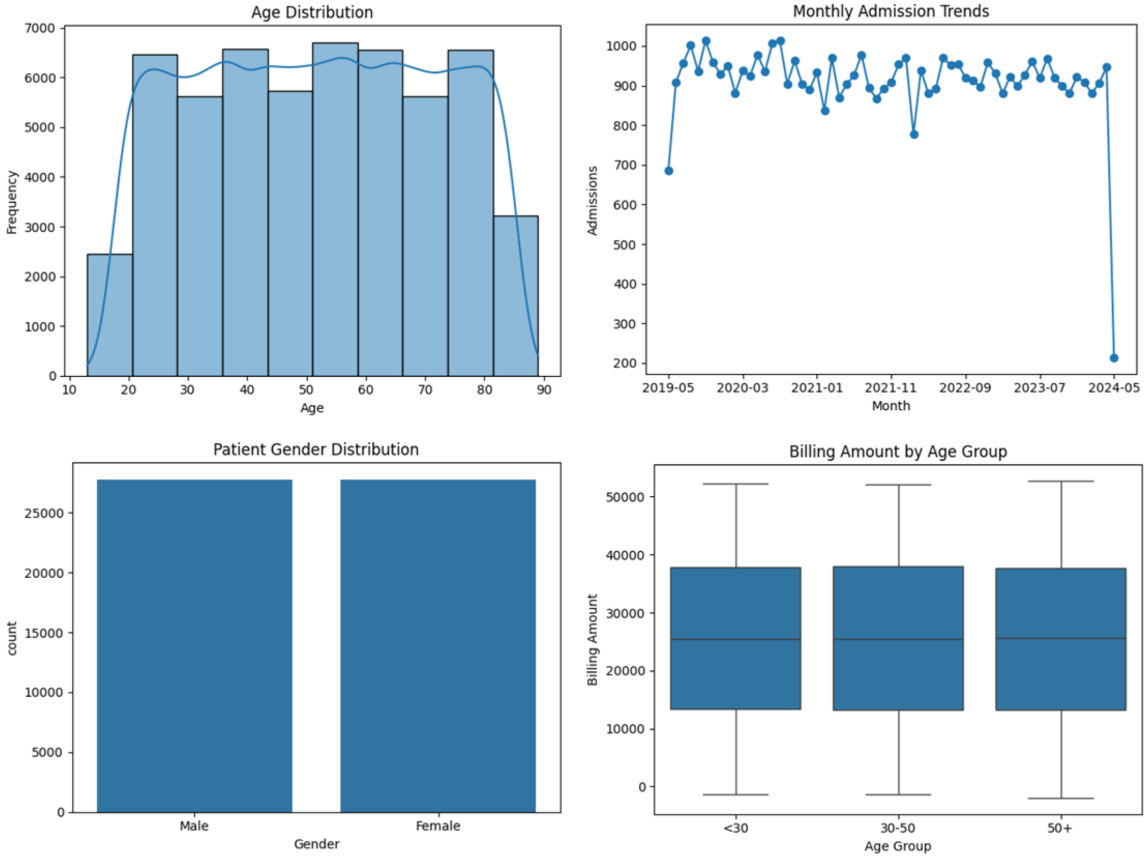
```
🔢 Value Counts for Categorical Features:
Gender:
 Gender
Male      27774
Female    27726
Name: count, dtype: int64

Age Group:
 Age Group
50+      28667
30-50    16334
<30      10499
Name: count, dtype: int64
```



```
Medical Condition:
 Medical Condition
Arthritis       9308
Diabetes        9304
Hypertension    9245
Obesity         9231
Cancer          9227
Asthma          9185
Name: count, dtype: int64
```

3.2.6 Data Visualization

The various visualizations are as follows:



3.3 SAVING THE CLEANED DATASET

Once data cleaning is done the new file is saved for subsequent analysis and modeling steps by using the command “**df.to_csv('cleaned_healthcare_dataset.csv', index=False)**”.

4 STATISTICAL ANALYSIS

Statistical analysis is applied on different attributes involving both exploratory and inferential techniques.

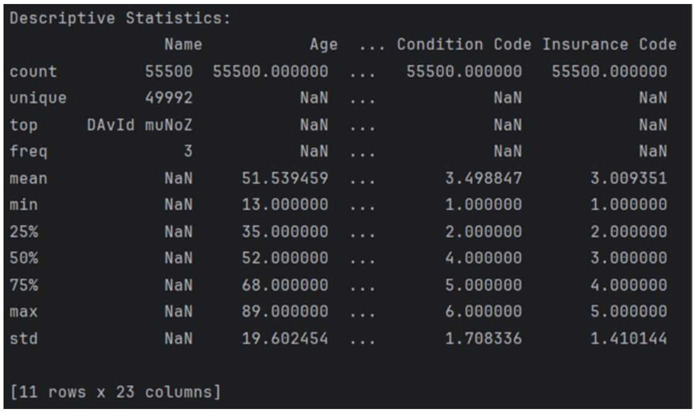
4.1 DESCRIPTIVE STATISTICS

Descriptive Statistics is carried out to understand the central tendencies and variability of numeric attributes like age, billing. Etc.

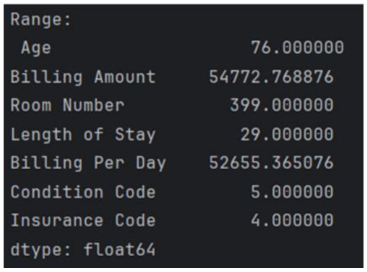
The code below is used to understand the descriptive stats like Mode, Median, Standard Deviation & Range.

```
print("Descriptive Statistics:\n", df.describe(include='all'))
print("Mode:\n", df.mode().iloc[0])
print("Median:\n", df.median(numeric_only=True))
print("Standard Deviation:\n", df.std(numeric_only=True))
print("Range:\n", df.max(numeric_only=True) - df.min(numeric_only=True))
```

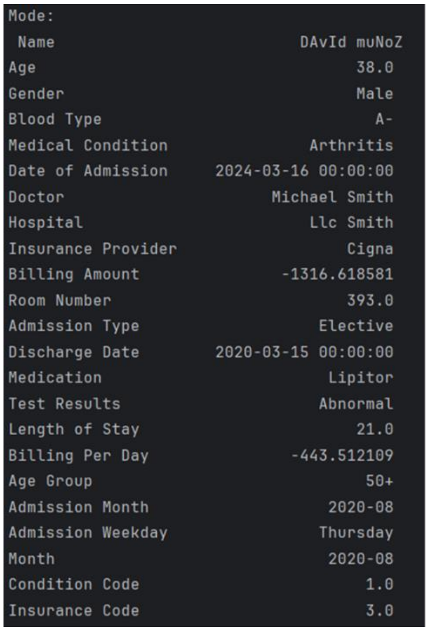
The output is as below:



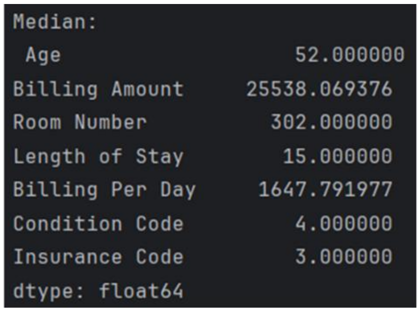
```
Descriptive Statistics:
                Name          Age  ... Condition Code Insurance Code
count         55500  55500.000000  ...   55500.000000   55500.000000
unique        49992           NaN  ...            NaN            NaN
top     DAvId muNoZ           NaN  ...            NaN            NaN
freq              3           NaN  ...            NaN            NaN
mean            NaN     51.539459  ...       3.498847       3.009351
min             NaN     13.000000  ...       1.000000       1.000000
25%             NaN     35.000000  ...       2.000000       2.000000
50%             NaN     52.000000  ...       4.000000       3.000000
75%             NaN     68.000000  ...       5.000000       4.000000
max             NaN     89.000000  ...       6.000000       5.000000
std             NaN     19.602454  ...       1.708336       1.410144

[11 rows x 23 columns]
```



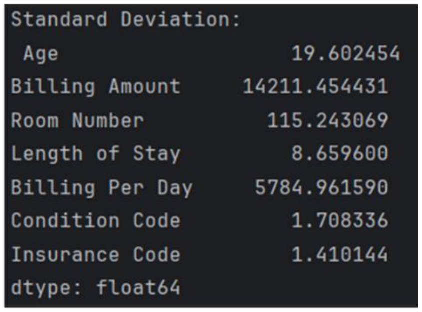
```
Range:
 Age                  76.000000
Billing Amount     54772.768876
Room Number          399.000000
Length of Stay        29.000000
Billing Per Day    52655.365076
Condition Code         5.000000
Insurance Code         4.000000
dtype: float64
```



```
Mode:
 Name                                    DAvId muNoZ
Age                                          38.0
Gender                                       Male
Blood Type                                     A-
Medical Condition                      Arthritis
Date of Admission           2024-03-16 00:00:00
Doctor                             Michael Smith
Hospital                               Llc Smith
Insurance Provider                         Cigna
Billing Amount                      -1316.618581
Room Number                                 393.0
Admission Type                          Elective
Discharge Date              2020-03-15 00:00:00
Medication                               Lipitor
Test Results                            Abnormal
Length of Stay                               21.0
Billing Per Day                      -443.512109
Age Group                                     50+
Admission Month                           2020-08
Admission Weekday                        Thursday
Month                                     2020-08
Condition Code                                1.0
Insurance Code                                3.0
```



```
Median:
 Age                   52.000000
Billing Amount      25538.069376
Room Number           302.000000
Length of Stay         15.000000
Billing Per Day      1647.791977
Condition Code          4.000000
Insurance Code          3.000000
dtype: float64
```



```
Standard Deviation:
 Age                   19.602454
Billing Amount      14211.454431
Room Number           115.243069
Length of Stay          8.659600
Billing Per Day      5784.961590
Condition Code          1.708336
Insurance Code          1.410144
dtype: float64
```

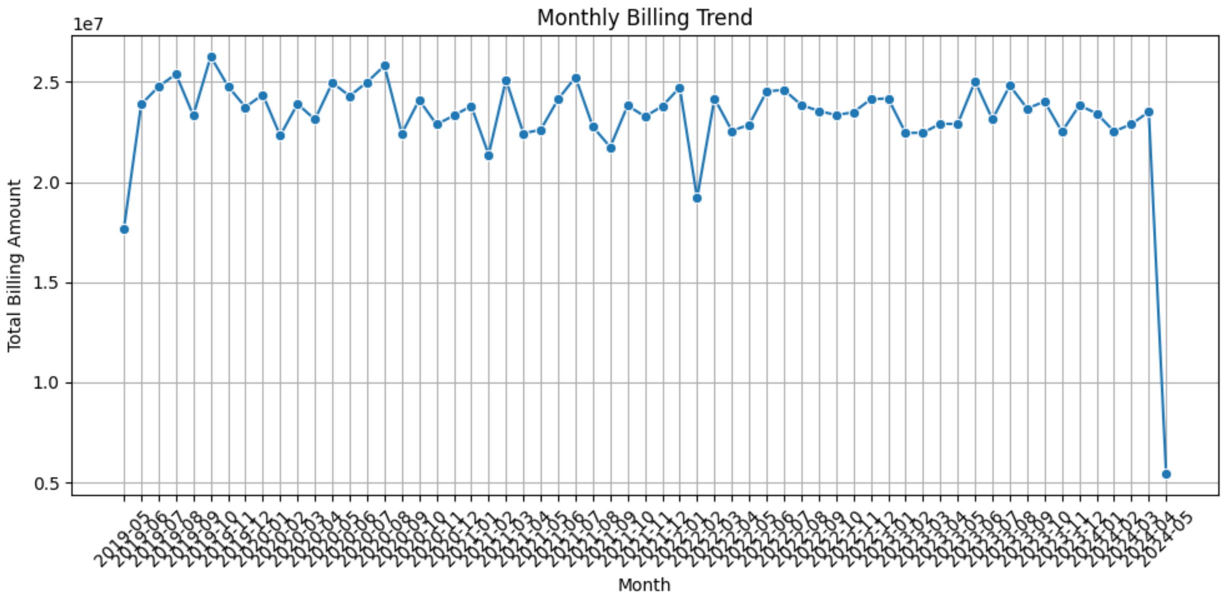
4.2 TREND ANALYSIS

Trend analysis is used to identify seasonal patterns in billings.

The code below is used to derive the trend analysis:

```
monthly_trend = df.groupby('Month')['Billing Amount'].sum()
print("Monthly Billing Trend:\n", monthly_trend)
```

The graphical output is as below:



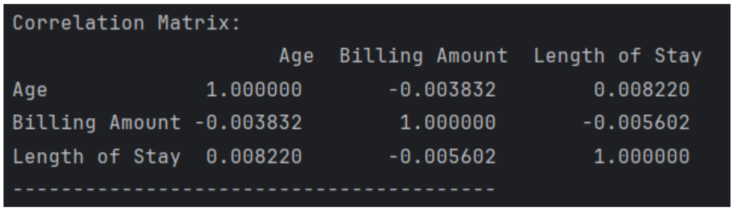
4.3 CORRELATION ANALYSIS

It examines the relationships between numerical variables like 'Age', 'Billing Amount', 'Length of Stay'.

The code below is used for correlation analysis:

```
correlation = df[['Age', 'Billing Amount', 'Length of Stay']].corr()
print("Correlation Matrix:\n", correlation)
```

The output is as below:



```
Correlation Matrix:
                      Age  Billing Amount  Length of Stay
Age             1.000000       -0.003832        0.008220
Billing Amount -0.003832        1.000000       -0.005602
Length of Stay  0.008220       -0.005602        1.000000
----------------------------------------
```

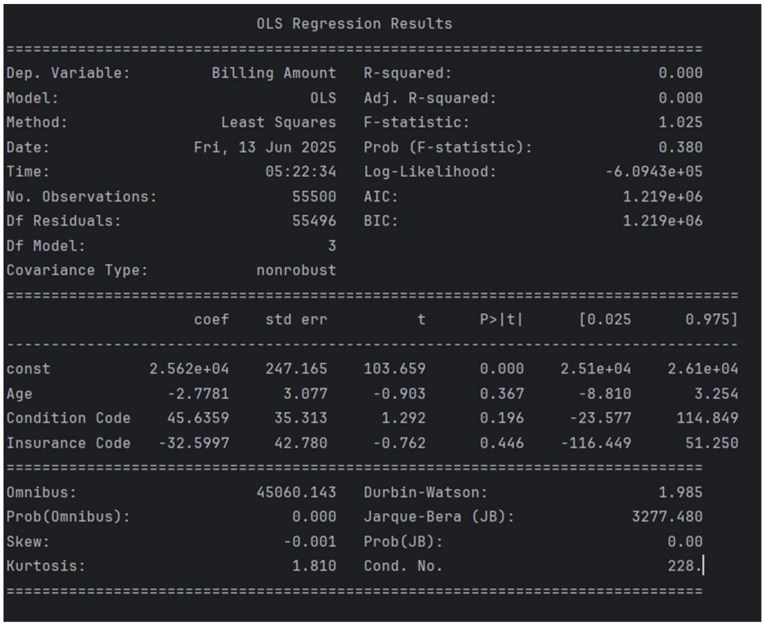
4.4 REGRESSION ANALYSIS

The regression analysis is carried out to predict the billing amount using the different independent variables like 'Age', 'Condition Code', 'Insurance Code'.

The code used is:

```
X = pd.get_dummies(df[['Age', 'Condition Code', 'Insurance Code']], drop_first=True)
X = sm.add_constant(X)
y = df['Billing Amount']
model = sm.OLS(y, X).fit()
print(model.summary())
```

The output is:



```
                            OLS Regression Results
==============================================================================
Dep. Variable:         Billing Amount   R-squared:                       0.000
Model:                            OLS   Adj. R-squared:                  0.000
Method:                 Least Squares   F-statistic:                     1.025
Date:                Fri, 13 Jun 2025   Prob (F-statistic):              0.380
Time:                        05:22:34   Log-Likelihood:            -6.0943e+05
No. Observations:               55500   AIC:                         1.219e+06
Df Residuals:                   55496   BIC:                         1.219e+06
Df Model:                           3
Covariance Type:            nonrobust
===================================================================================
                      coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------
const            2.562e+04    247.165    103.659      0.000    2.51e+04    2.61e+04
Age                -2.7781      3.077     -0.903      0.367      -8.810       3.254
Condition Code     45.6359     35.313      1.292      0.196     -23.577     114.849
Insurance Code    -32.5997     42.780     -0.762      0.446    -116.449      51.250
==============================================================================
Omnibus:                    45060.143   Durbin-Watson:                   1.985
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             3277.480
Skew:                          -0.001   Prob(JB):                         0.00
Kurtosis:                       1.810   Cond. No.                         228.
==============================================================================
```

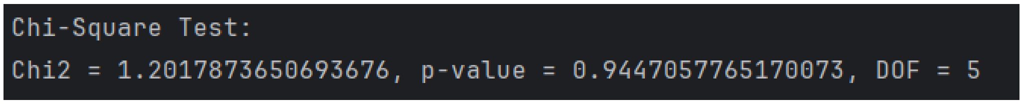
4.5 CHI-SQUARE TEST

This test is carried out to explore the relationship between categorial variables (Gender & Medical Condition).

The code used is:

```
contingency = pd.crosstab(df['Gender'], df['Medical Condition'])
chi2, p, dof, expected = chi2_contingency(contingency)
print(f"Chi-Square Test:\nChi2 = {chi2}, p-value = {p}, DOF = {dof}")
```

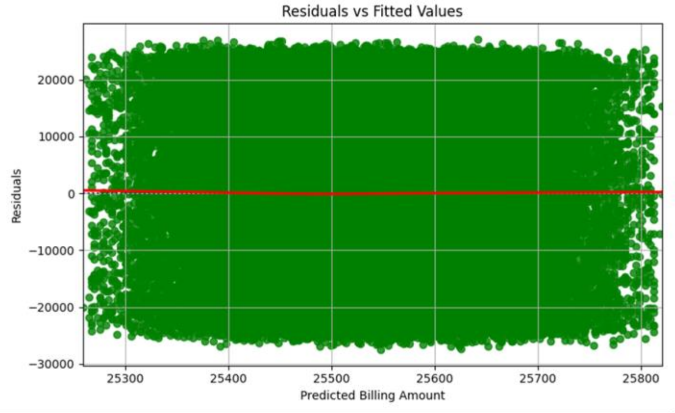
The output is:

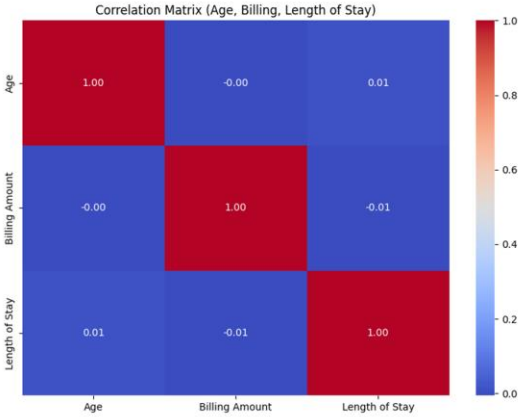


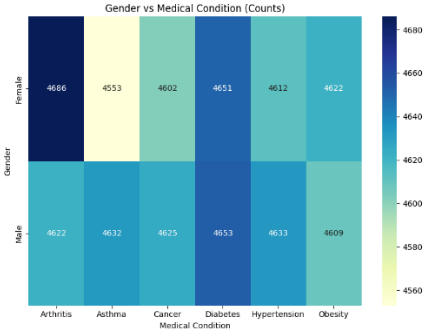
```
Chi-Square Test:
Chi2 = 1.2017873650693676, p-value = 0.9447057765170073, DOF = 5
```

4.6 VISUALS OF STATISTICAL ANALYSIS

The different visuals fetched for the different statistical analysis are placed below:







5 ENVIRONMENTAL IMPACT ASSESSMENT (PESTEL)

Statistical analysis is applied to different attributes involving both exploratory and inferential techniques.

When using the PESTEL framework to evaluate the macro-environment, healthcare administrators and analysts can prepare for external influences. The dataset emphasizes the interconnectedness of policy, demographics, billing behavior, and technology on the operating environment of the healthcare sector.

5.1 POLITICAL FACTORS

Government and insurance policies (for example, Medicare and Aetna) help govern treatment accessibility and billing systems. We have shown a variance in billed amounts from different insurance companies through our data, indicating potential variants in coverage and reimbursement linked to policies. Political decisions on government-supported healthcare, privatization of care, or regulatory change could all rapidly shape admission patterns and potentially financial viability for hospitals.

5.2 ECONOMIC FACTORS

High-billing admissions, as indicated by the data, are something all healthcare organizations must rely upon heavily, especially during economic downturns or in any instance with cost-cutting measures. As inflation rises and health care costs increase, patients may not be able to afford care, and reimbursement systems can become resource strained. . In this case, we have data that shows we can increase and optimize billed per day numbers while utilizing the length of stay to see how well we can maintain financial survivability based on longer admission and billing.

5.3 SOCIAL FACTORS

From data, there was a demographic distribution - especially age and gender - to the patient population presenting in our data. An aging hospital population is likely causing increased bed length and billed care which relates to a growing population of geriatric care. More specific population variables - such as gender - for certain medical conditions and treatment types indicate to us that there is a desire for more (personal) and (inclusive) care strategies.

5.4 TECHNOLOGICAL FACTORS

The dataset's comprehensiveness in documenting medication types and usage frequency may suggest the combined functionalities of electronic health record and supply chain systems. These technologies allow accuracy in treatment, enable billing transparency (e.g., order entry context), and provide the potential for automation and predictive analytics. Facilities investing in health IT would realize organizational efficiency and benefit from increased patient safety.

5.5 ENVIRONMENTAL FACTORS

While the dataset incorporates little environmental data, the digitization of records (e.g., EHRs) creates less paper waste, with beneficial implications. Facilities would have the opportunity to extend the happenings in the dataset to include issues of sustainability, for example, pharmaceutical disposal methods or adopting efficient operations. Identifying environmental indicators in future datasets could help facilities demonstrate compliance with existing green healthcare standards.

5.6 LEGAL FACTORS

Insurance payers are responsible for adhering to specific reimbursement and compliance laws. The inclusion of large payers (e.g., Medicare and Aetna) in the dataset means that facilities are familiar with the operational implications of legal reviews and audits. Facilities that engage in improper documentation, or overbilling, risk still facing financial penalties for non-compliance. Continued strong legal frameworks and planned audit readiness remain an important risk management strategy.

6 KEY INSIGHTS & STRATEGIC IMPLICATIONS

Through detailed statistical analyses and PESTEL assessment, various operational patterns and strategic priorities become identifiable. All for healthcare decision-makers, the data provides a sound foundation for mitigating risk and improving efficiency.

These strategic insights characterize an opportunity to adapt operations to shifting patient demographics, care needs and system risks. Leveraging the findings, healthcare organizations can improve resilience, reduce expenditure, and improve patient outcomes.

6.1 KEY INSIGHTS

6.1.1 Heavy Reliance on Medicare

High rates of patient coverage through Medicare provides consistency in revenue under existing guidelines but enhances the level of risk from the perspective of potential changes in reimbursement, coverage, and eligibility for future care under Medicare.

6.1.2 Cost- Intensity: Cancer and Diabetes

Our analysis reveals that patients suffering from Cancer and Diabetes had the highest billing and the longest lengths of stay. These conditions create a cost-intensity and raise some questions about the potential create tighter coordination of clinician involvement in diagnostic, medication inserts, and post-discharge support to control patients' cost and quality of care.

6.1.3 Variations in Billing: Relative Comparison of Urgent Care and Elective Care

The billing information provided a stark account of differences in urgent admission compared admissions for elective surgery. This pattern requires hospitals to have stronger operational readiness for capacity planning and workforce scheduling. Hospitals must improve their capability to plan and schedule staff or health professionals to handle similarly across periods of peak demand for their services without compromising service or by breaking their staffing resources.

6.1.4 Patterns of Medication: Ibuprofen Usage

Ibuprofen is the med prescribed to patients from across the spectrum. This presents an opportunity to look at procurement, drug levels, and the suppliers' contract of existing drug formularies—reduce spending and wait for essential medications.

6.2 STRATEGIC IMPLICATIONS

6.2.1 Insurance Risk Reduction

Broaden the payer mix and negotiate improved rates with private payors to lessen their dependence on public payors like Medicare.

6.2.2 Condition-Specific Efficiency Initiatives

Utilize condition-specific process improvement for high-cost conditions (e.g., Cancer, Diabetes), with cohesive care pathways, digital tracking of care plans and predictive analytics for readmission risk.

6.2.3 Capacity & Resource Planning

Use flexible scheduling and staff assignments that are responsive to the variety of admissions to improve throughput and maintain high quality of patient care.

6.2.4 Supply Chain Resilience

Utilize prescribing patterns to forecast usage and procure high-volume medications, such as Ibuprofen, to preserve amplification while reducing costs and minimizing interruptions.

7 RECOMMENDATIONS

Following the data analysis, environment assessment, and operations trends observed, here are some strategic recommendations to maximize efficiency, mitigate risks, and take advantage of emerging opportunities in the health care system.

1. **Improve Inventory Management for Medications with High Utilization**

The frequency analysis of prescriptions indicated that some medications, such as Ibuprofen, are widely used across patient classifications. With a demand-based inventory control and deeper supplier relationships, the organization can reduce stockouts, reduce excess inventory, and even reduce purchase costs.

Actions:

- Drive procurement schedule based on historical prescribing trends.
- Implement automated inventory systems with reorder alerts.

2. **Project Future Demand for Elderly Care Facilities**

Older patients represent a significant proportion of patients, and the length of stay increases with age. With changing demographics, the system needs to proactively prepare geriatric infrastructure (e.g. specialized care units, trained staff).

Actions:

- Use demographic and billing data for projecting future service load.
- Consider investing in long-term care facilities or partnerships with eldercare providers.

3. **Broaden Payer Mix**

The system relies heavily on Medicare and changes in public healthcare policy expose the system to risk. Seeking a greater proportionate share of private insurance patients or creating self-pay packages can create more stable revenue streams and reduce systemic risk.

Actions:

- Review payer breakdowns and develop marketing initiatives to capture under-represented insurance categories.
- Negotiate competitive packages with private payers.

4. Embrace Technology to Improve Admission-to-Discharge Processes

The variability of length of stay and types of admissions indicates inefficiencies within transitions of care. Technology investments in healthcare IT (e.g., EHR systems, automated triage, patient tracking) can help reduce bottlenecks, standardize care pathways, and improve patient experience.

Action:

- Deploy digital dashboards on bed availability and discharge planning in real-time.
- Implement automation workflows to reduce the need for manual data entry and lengthy delays.

5. Use Predictive Modeling to Identify High-Cost Patients

Statistical analysis indicates specific conditions and demographics are correlated with high-billing and extended lengths of stay. Applying machine learning models to predict high-cost patients at the time of admission may allow for timely intervention and steer scarce resources more appropriately.

Action:

- Create a predictive model using age, diagnosis, insurance type, and medication data.
- Educate clinical teams about early diagnostics and implement better care coordination in flagged cases.

These recommendations outline an approach for operational resilience, patient-centered care, and financial viability. By incorporating data insights into their decision-making, healthcare organizations can establish a better deal of external uncertainties and internal inefficiencies.

8 REFERENCES

- Healthcare Industry Wikipedia https://en.wikipedia.org/wiki/Healthcare_industry
- Selection of dataset from Kaggle https://www.kaggle.com/
- PESTEL analysis https://libguides.libraries.wsu.edu/c.php?g=294263&p=4358409
- Python tutorial https://docs.python.org/3/tutorial/
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (1st ed.) https://www.statlearning.com/
- Operations Research: Applications and Algorithms. 4th Edition. Duxbury Press. Winston, W. L. (2004).
- Google OR-Tools Documentation. https://developers.google.com/optimization