```
# Step 0: Import Libraries & Load Data
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# Load dataset
df = pd.read_csv('medical_cost.csv')
df.head()
```

	age	gender	bmi	smoker	diabetes	hypertension	heart_disease	asthma	physical_activity_level	daily_steps	sleep_hours	st
0	69	Male	29.4	No	1	0	0	0	Medium	14825	4.4	
1	32	Female	22.9	No	1	0	0	0	Medium	3620	6.0	
2	89	Male	25.7	No	0	0	0	0	High	10578	4.5	
3	78	Male	31.9	Yes	0	1	0	0	Low	6226	8.6	
4	38	Male	27.7	No	0	0	0	0	High	6253	5.7	

```
# Step 1: Data Overview
print(df.shape)
print(df.info())
print(df.describe())
print(df.isnull().sum())
```

```
(5000, 20)
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5000 entries, 0 to 4999
Data columns (total 20 columns):
 #   Column                   Non-Null Count  Dtype  
---  ------                   --------------  -----  
 0   age                      5000 non-null   int64  
 1   gender                   5000 non-null   object 
 2   bmi                      5000 non-null   float64
 3   smoker                   5000 non-null   object 
 4   diabetes                 5000 non-null   int64  
 5   hypertension             5000 non-null   int64  
 6   heart_disease            5000 non-null   int64  
 7   asthma                   5000 non-null   int64  
 8   physical_activity_level  5000 non-null   object 
 9   daily_steps              5000 non-null   int64  
 10  sleep_hours              5000 non-null   float64
 11  stress_level             5000 non-null   int64  
 12  doctor_visits_per_year   5000 non-null   int64  
 13  hospital_admissions      5000 non-null   int64  
 14  medication_count         5000 non-null   int64  
 15  insurance_type           3952 non-null   object 
 16  insurance_coverage_pct   5000 non-null   int64  
 17  city_type                5000 non-null   object 
 18  previous_year_cost       5000 non-null   int64  
 19  annual_medical_cost      5000 non-null   float64
dtypes: float64(3), int64(12), object(5)
memory usage: 781.4+ KB
None
               age          bmi     diabetes  hypertension  heart_disease  \
count  5000.000000  5000.000000  5000.000000   5000.000000     5000.00000   
mean     53.299000    25.970820     0.207600      0.288000        0.14220   
std      20.646851     5.046651     0.405629      0.452876        0.34929   
min      18.000000     6.400000     0.000000      0.000000        0.00000   
25%      36.000000    22.600000     0.000000      0.000000        0.00000   
50%      53.000000    25.900000     0.000000      0.000000        0.00000   
75%      71.000000    29.400000     0.000000      1.000000        0.00000   
max      89.000000    43.600000     1.000000      1.000000        1.00000   

            asthma   daily_steps  sleep_hours  stress_level  \
count  5000.000000   5000.000000  5000.000000   5000.000000   
mean      0.096400   7993.216800     6.488140      5.475400   
std       0.295169   4052.127069     1.443361      2.892312   
min       0.000000   1004.000000     4.000000      1.000000   
25%       0.000000   4545.000000     5.200000      3.000000   
50%       0.000000   7989.000000     6.500000      5.000000   
75%       0.000000  11532.250000     7.700000      8.000000   
max       1.000000  14999.000000     9.000000     10.000000   
```

```
       doctor_visits_per_year  hospital_admissions  medication_count  \
count             5000.000000          5000.000000       5000.000000
mean                 4.030600             1.001000          3.509000
std                  2.010689             0.978566          2.292721
min                  0.000000             0.000000          0.000000
25%                  3.000000             0.000000          1.000000
50%                  4.000000             1.000000          3.000000
75%                  5.000000             2.000000          6.000000
max                 14.000000             6.000000          7.000000
```

Observation:

insurance_type has 1,048 missing values (~21%)

All other columns have no missing data

```
# Step 2: Data Cleaning & Feature Engineering
# 2.1 Handle Missing Values

df['insurance_type'] = df['insurance_type'].fillna('Unknown')
# Dropping 20% of the data would reduce analytical power and may bias cost analysis.
```

```
# 2.2 Check for Duplicates
df.duplicated().sum()
```

```
np.int64(0)
```

Observation: No duplicate records were found in the dataset. Each row represents a unique patient, so no deduplication was required.

```
# 2.3 Feature Engineering – Age Groups
# Grouping ages improves interpretability for business stakeholders.
bins = [0, 18, 30, 45, 60, 100]
labels = ['0-18', '19-30', '31-45', '46-60', '60+']

df['age_group'] = pd.cut(df['age'], bins=bins, labels=labels)
```

```
# 2.4 Encode Categorical Variables
# For correlation analysis and regression modeling, convert categorical variables to numeric.

categorical_cols = [
    'gender',
    'smoker',
    'physical_activity_level',
    'insurance_type',
    'city_type',
    'age_group'
]

df_encoded = pd.get_dummies(df, columns=categorical_cols, drop_first=True)
```

Note:

Original df will be used for visualization

df_encoded will be used for correlation and modeling

```
# 2.5 Sanity Check
df_encoded.shape
df_encoded.head()
```

	age	bmi	diabetes	hypertension	heart_disease	asthma	daily_steps	sleep_hours	stress_level	doctor_visits_per_year	...
0	69	29.4	1	0	0	0	14825	4.4	8	1	...
1	32	22.9	1	0	0	0	3620	6.0	7	4	...
2	89	25.7	0	0	0	0	10578	4.5	7	2	...
3	78	31.9	0	1	0	0	6226	8.6	9	6	...
4	38	27.7	0	0	0	0	6253	5.7	3	6	...

5 rows × 27 columns

Step 2 Summary

Missing insurance values handled using a meaningful category

No duplicate records detected

Age grouped for clearer cost comparisons

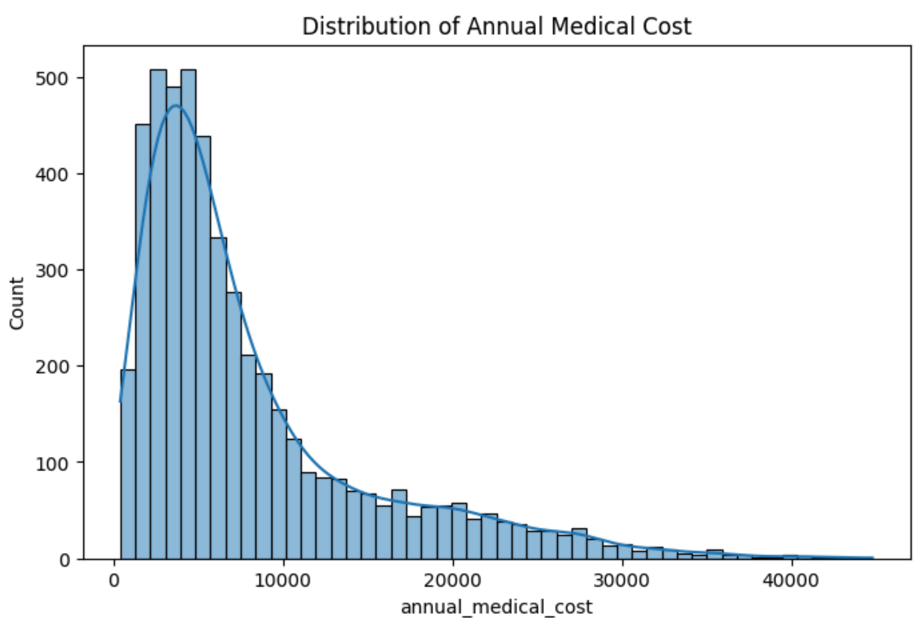
Dataset prepared for:

Exploratory visualizations

Cost driver analysis

Baseline predictive modeling

```
# Step 3: Target Variable Analysis (annual_medical_cost)
plt.figure(figsize=(8,5))
sns.histplot(df['annual_medical_cost'], bins=50, kde=True)
plt.title('Distribution of Annual Medical Cost')
plt.show()

print("Skewness:", df['annual_medical_cost'].skew())
```




```
Skewness: 1.679655900430324
```

Skewness of Annual Medical Cost

The skewness of `annual_medical_cost` is **1.68**, indicating a **right-skewed distribution**.

This suggests that:

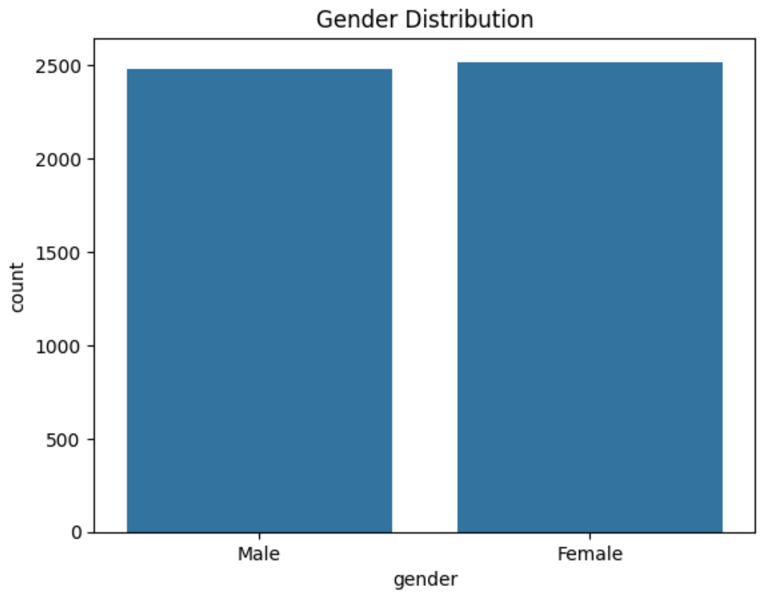
- Most patients incur **low to moderate medical costs**
- A **small proportion of patients** generate **very high medical expenses**

From a business perspective, this pattern is typical in healthcare data and highlights the importance of identifying **high-cost patient groups**, as they account for a disproportionate share of total spending.

As a result, subsequent analysis will focus on:

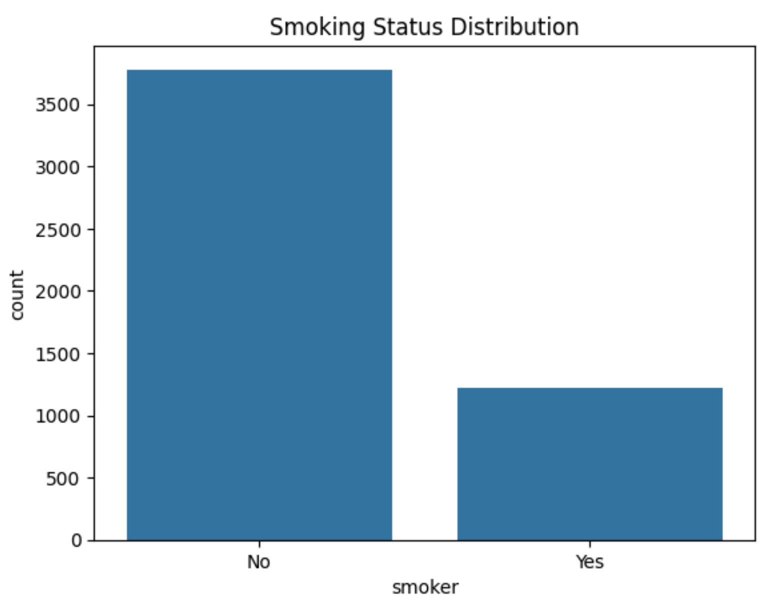
- Cost drivers among high-expense patients
- Demographic, lifestyle, and health factors contributing to elevated costs

```
# Step 4: Univariate Analysis
# Analyze each feature independently to understand the distribution, balance, and prevalence of demographic, lifestyle, and hea
# 4.1 Categorical Variables
# Gender Distribution
sns.countplot(x='gender', data=df)
plt.title('Gender Distribution')
plt.show()
```



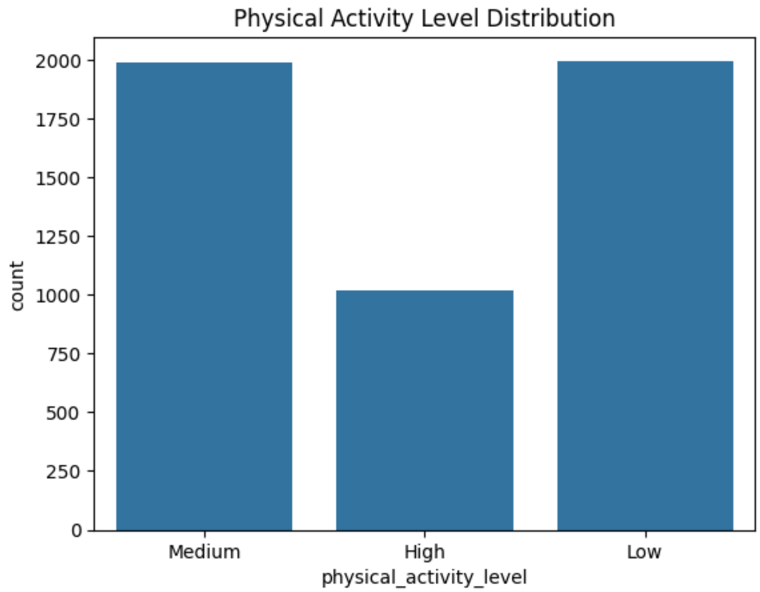
The dataset shows a balanced gender distribution, reducing the risk of gender-driven sampling bias in cost analysis.

```
# Smoking Status
sns.countplot(x='smoker', data=df)
plt.title('Smoking Status Distribution')
plt.show()
```



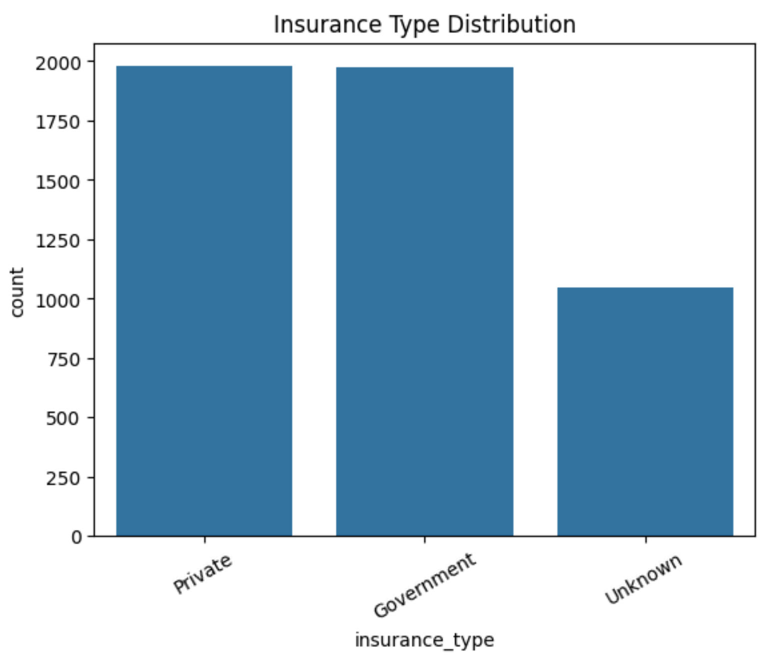
A substantial portion of the population reports smoking, suggesting smoking status may play a significant role in healthcare cost variation.

```
# Physical Activity Level
sns.countplot(x='physical_activity_level', data=df)
plt.title('Physical Activity Level Distribution')
plt.show()
```



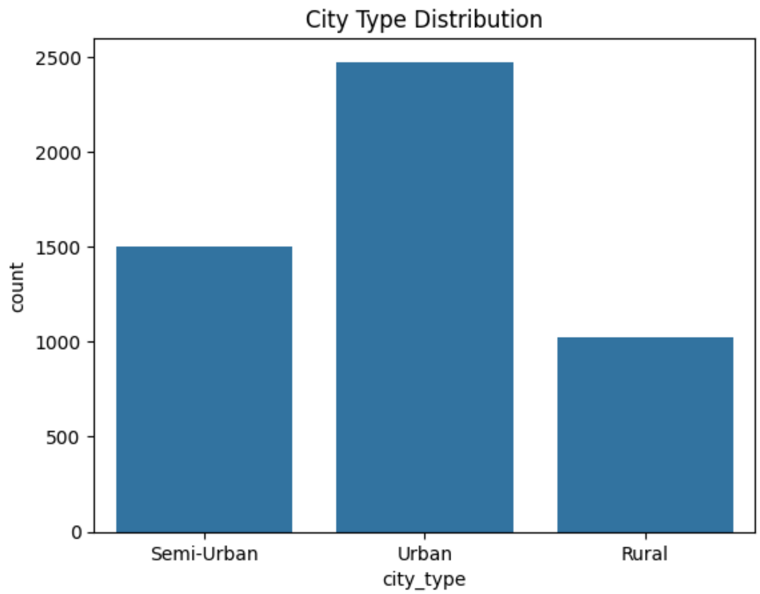
Physical activity levels vary across the population, indicating potential differences in health outcomes and medical utilization.

```
#Insurance Type
sns.countplot(x='insurance_type', data=df)
plt.title('Insurance Type Distribution')
plt.xticks(rotation=30)
plt.show()
```



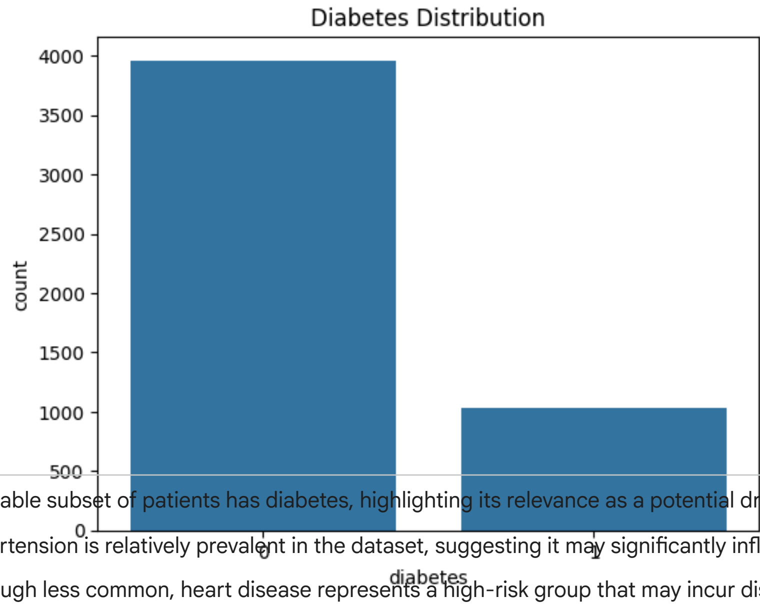
Insurance coverage types are unevenly distributed, which may contribute to differences in healthcare access and medical costs.

```
# City Type
sns.countplot(x='city_type', data=df)
plt.title('City Type Distribution')
plt.show()
```



Patients are distributed across different city types, allowing comparison of healthcare cost patterns between urban and non-urban areas.

```
# 4.2 Binary Health Conditions
binary_cols = ['diabetes', 'hypertension', 'heart_disease', 'asthma']

for col in binary_cols:
    sns.countplot(x=col, data=df)
    plt.title(f'{col.replace("_", " ").title()} Distribution')
    plt.show()
```



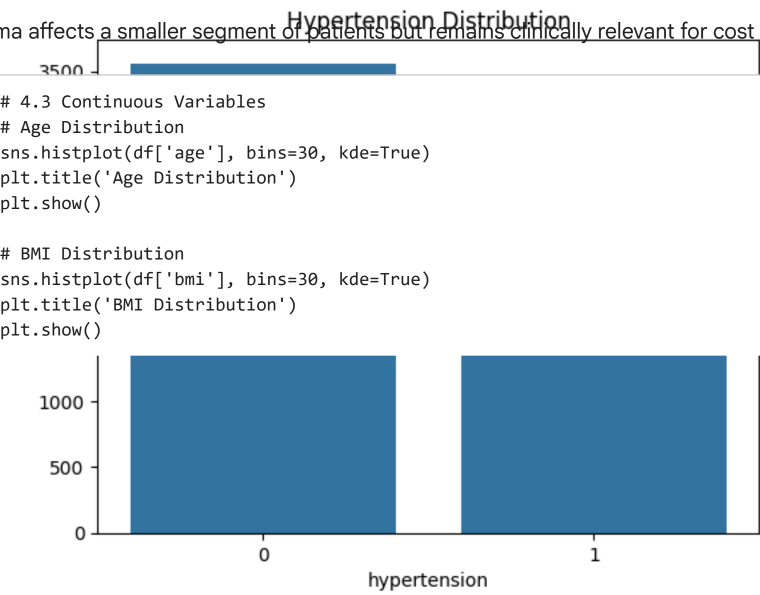
A notable subset of patients has diabetes, highlighting its relevance as a potential driver of higher medical costs.

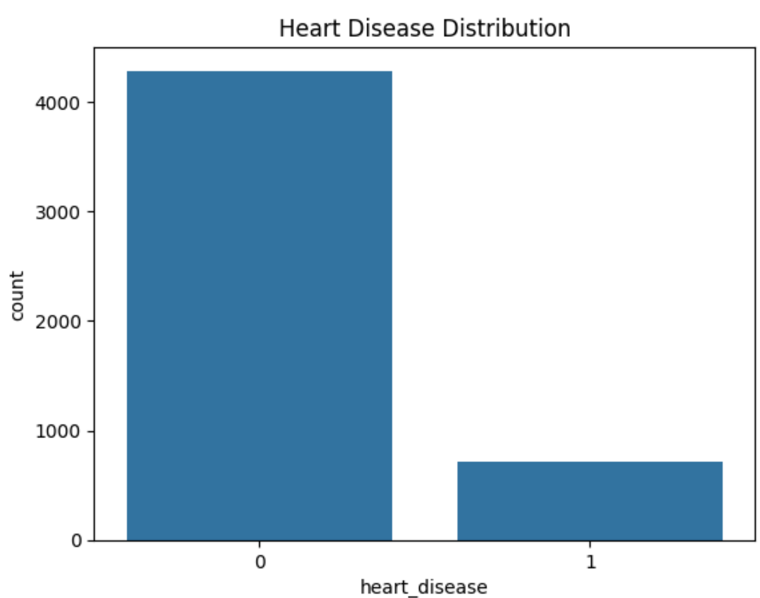
Hypertension is relatively prevalent in the dataset, suggesting it may significantly influence long-term healthcare utilization.

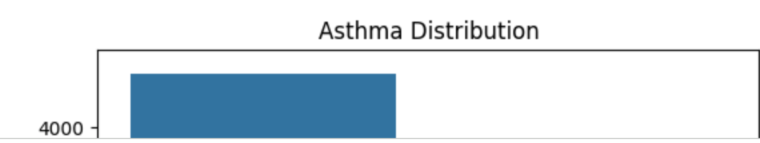
Although less common, heart disease represents a high-risk group that may incur disproportionately high medical costs.

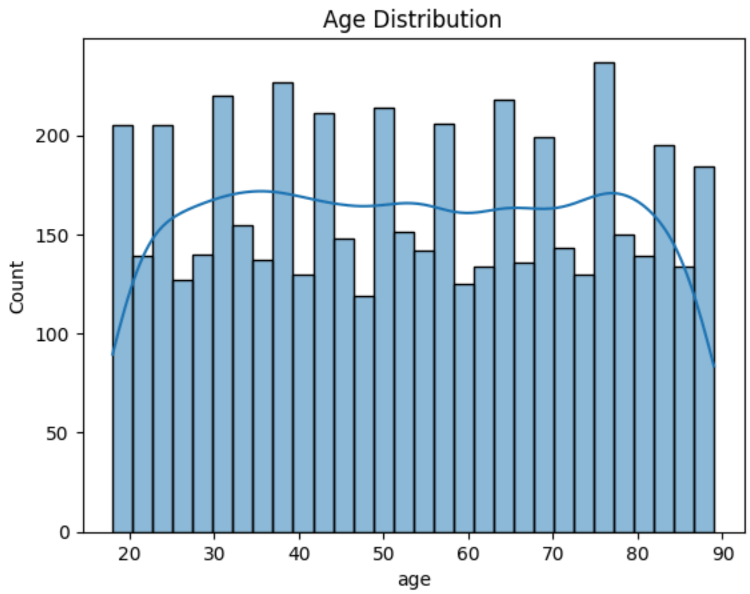
Asthma affects a smaller segment of patients but remains clinically relevant for cost and utilization analysis.

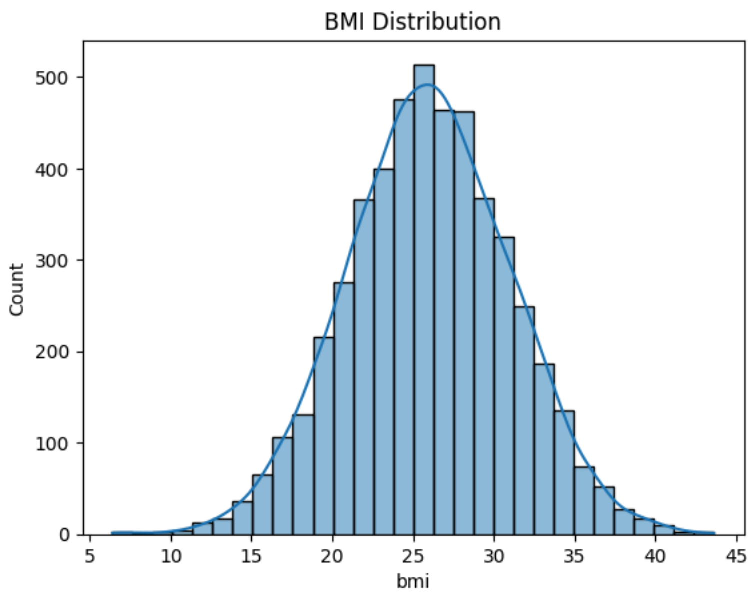
```
# 4.3 Continuous Variables
# Age Distribution
sns.histplot(df['age'], bins=30, kde=True)
plt.title('Age Distribution')
plt.show()

# BMI Distribution
sns.histplot(df['bmi'], bins=30, kde=True)
plt.title('BMI Distribution')
plt.show()
```







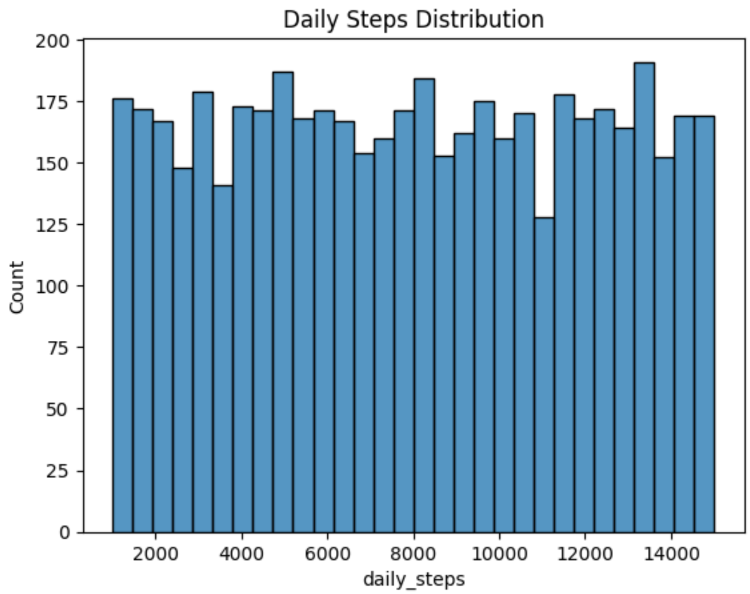


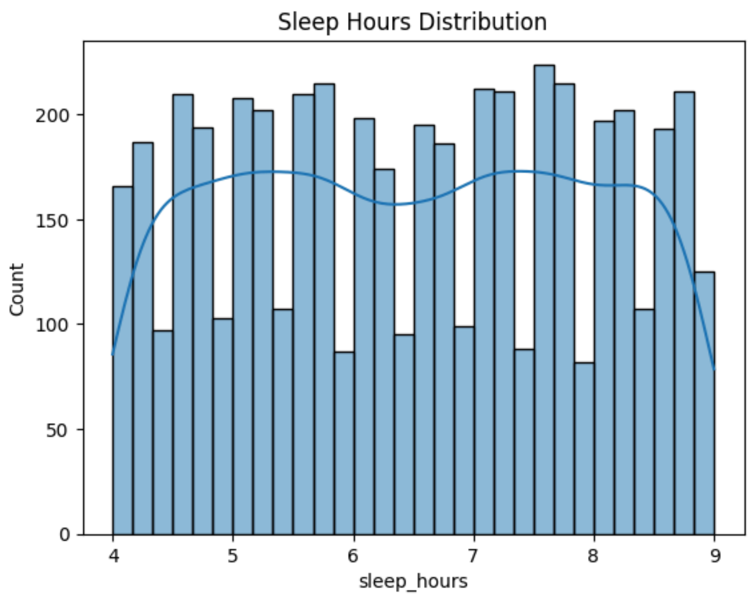


The dataset covers a wide age range, enabling analysis of healthcare cost differences across life stages.

BMI values cluster around the overweight range, indicating potential metabolic health risks within the population.

```
# Daily Steps
sns.histplot(df['daily_steps'], bins=30)
plt.title('Daily Steps Distribution')
plt.show()

# Sleep Hours
sns.histplot(df['sleep_hours'], bins=30, kde=True)
plt.title('Sleep Hours Distribution')
plt.show()
```

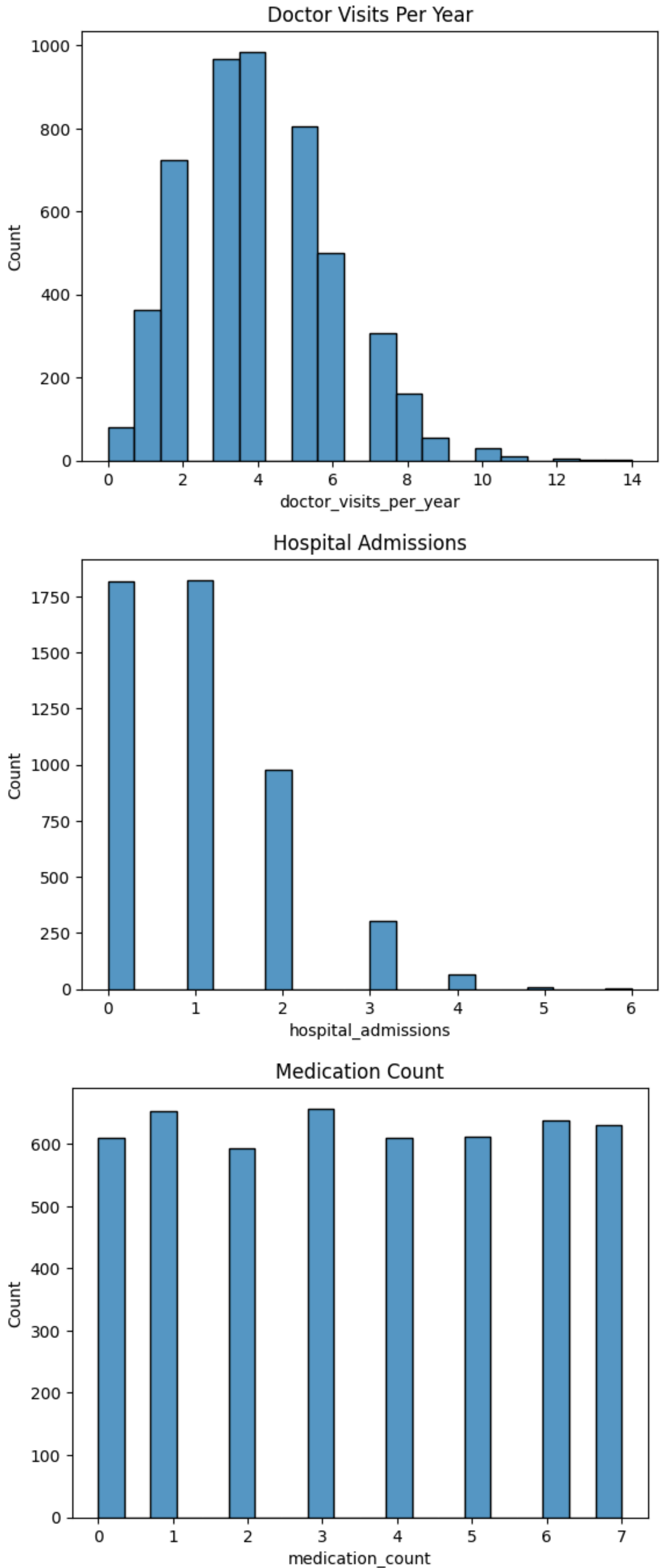




Daily step counts show high variability, reflecting diverse physical activity patterns among patients.

Sleep duration varies across individuals, which may be associated with stress levels and overall health outcomes.

```
# 4.4 Healthcare Utilization Variables
util_cols = [
    'doctor_visits_per_year',
    'hospital_admissions',
    'medication_count'
]

for col in util_cols:
    sns.histplot(df[col], bins=20)
    plt.title(col.replace('_', ' ').title())
    plt.show()
```



Healthcare utilization metrics exhibit substantial variation, suggesting heterogeneous healthcare needs and cost profiles.

Step 4 Summary

Dataset shows balanced demographic coverage

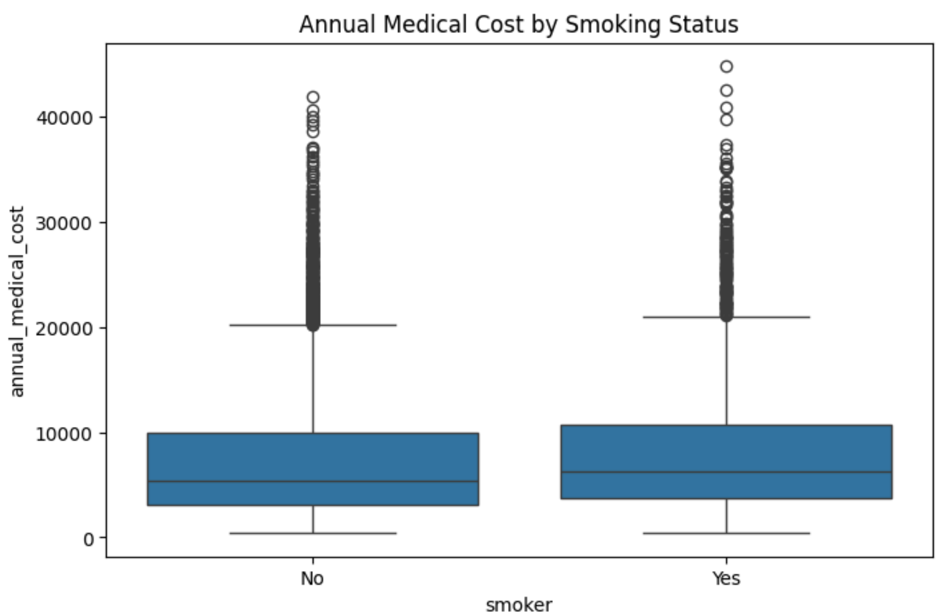
Chronic diseases are present but not dominant

Lifestyle variables exhibit meaningful variation

Healthcare utilization differs significantly across individuals

These observations set the foundation for multivariate cost driver analysis in the next step.

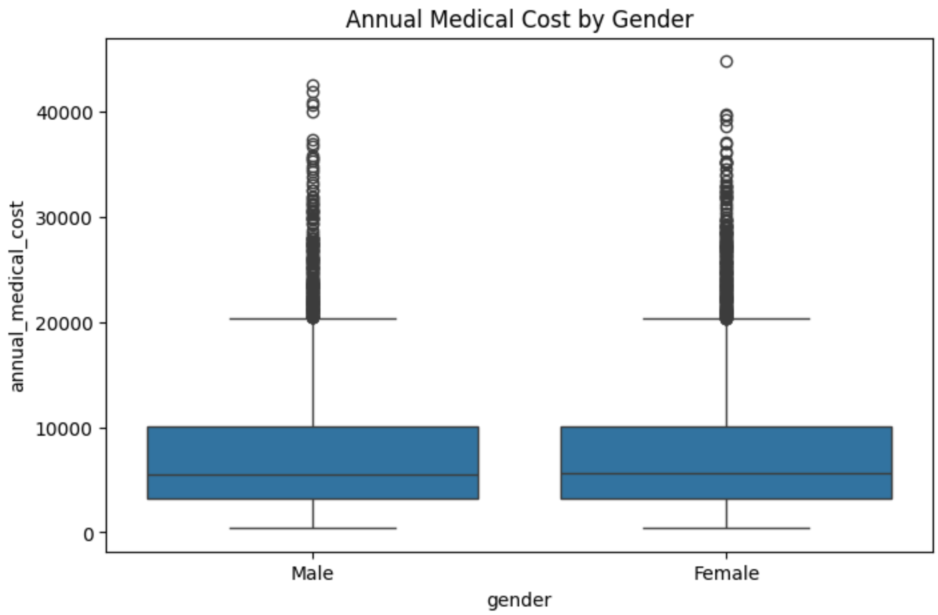
```
# Step 5: Multivariate Analysis – Identifying Cost Drivers
'''
Understand which demographic, lifestyle, health, and utilization factors are most associated with annual medical costs.

Combine categorical vs cost and continuous vs cost analyses for actionable insights.'''

# 5.1 Categorical vs Annual Medical Cost

# Smoking Status vs Cost
plt.figure(figsize=(8,5))
sns.boxplot(x='smoker', y='annual_medical_cost', data=df)
plt.title('Annual Medical Cost by Smoking Status')
plt.show()
```



Insight:

Smokers incur significantly higher medical costs than non-smokers, highlighting smoking as a major cost driver.

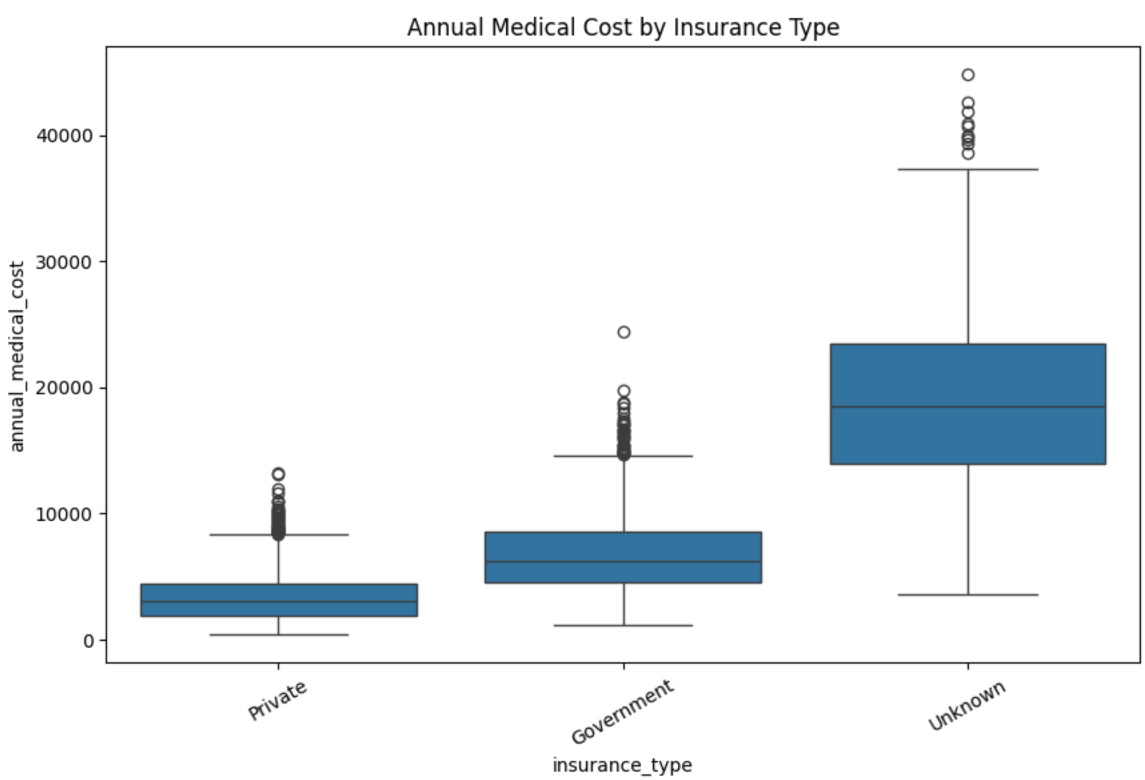
```
# Gender vs Cost
plt.figure(figsize=(8,5))
sns.boxplot(x='gender', y='annual_medical_cost', data=df)
plt.title('Annual Medical Cost by Gender')
plt.show()
```



Insight:

Gender differences in costs are minor, suggesting that gender alone is not a primary driver of medical expenses.

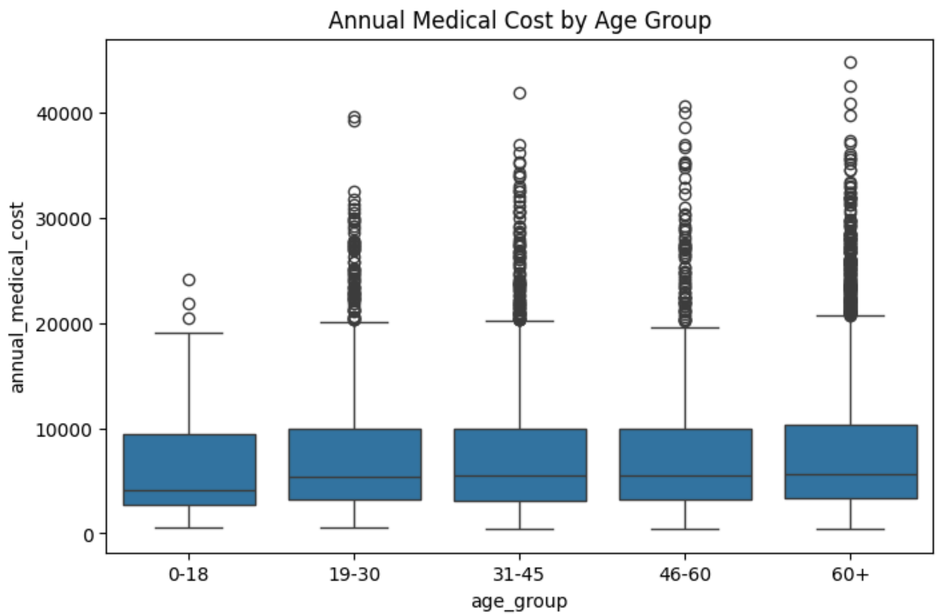
```
#nsurance Type vs Cost
plt.figure(figsize=(10,6))
sns.boxplot(x='insurance_type', y='annual_medical_cost', data=df)
plt.title('Annual Medical Cost by Insurance Type')
plt.xticks(rotation=30)
plt.show()
```



Insight:

Patients with different insurance types show substantial variation in costs, indicating that coverage type influences medical spending.

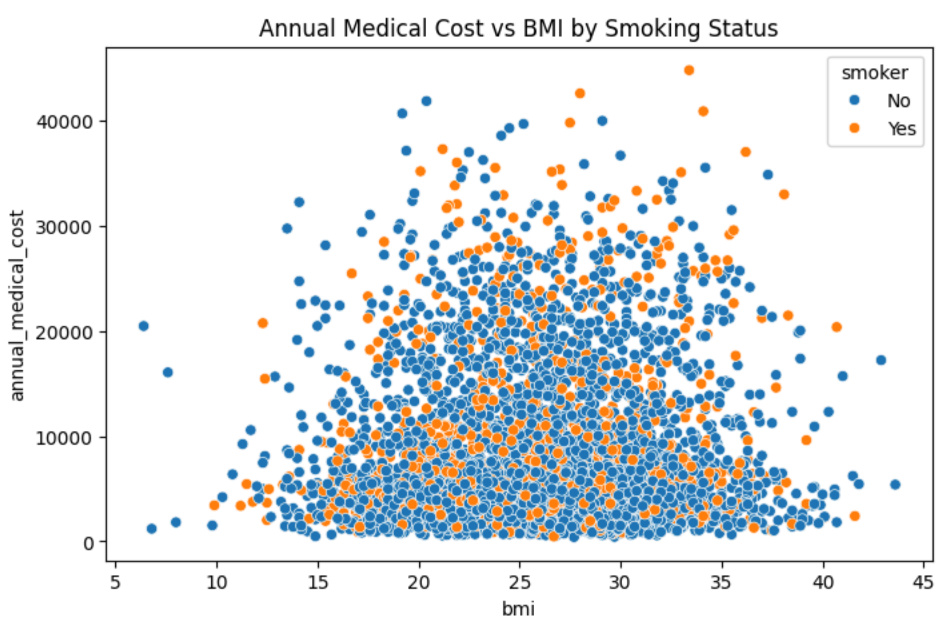
```
# Age Group vs Cost
plt.figure(figsize=(8,5))
sns.boxplot(x='age_group', y='annual_medical_cost', data=df)
plt.title('Annual Medical Cost by Age Group')
plt.show()
```



Insight:

Older age groups consistently have higher average medical costs, confirming age as a key cost driver.

```
# 5.2 Continuous vs Annual Medical Cost
# BMI vs Cost (with Smoking Hue)
plt.figure(figsize=(8,5))
sns.scatterplot(x='bmi', y='annual_medical_cost', hue='smoker', data=df)
plt.title('Annual Medical Cost vs BMI by Smoking Status')
plt.show()
```

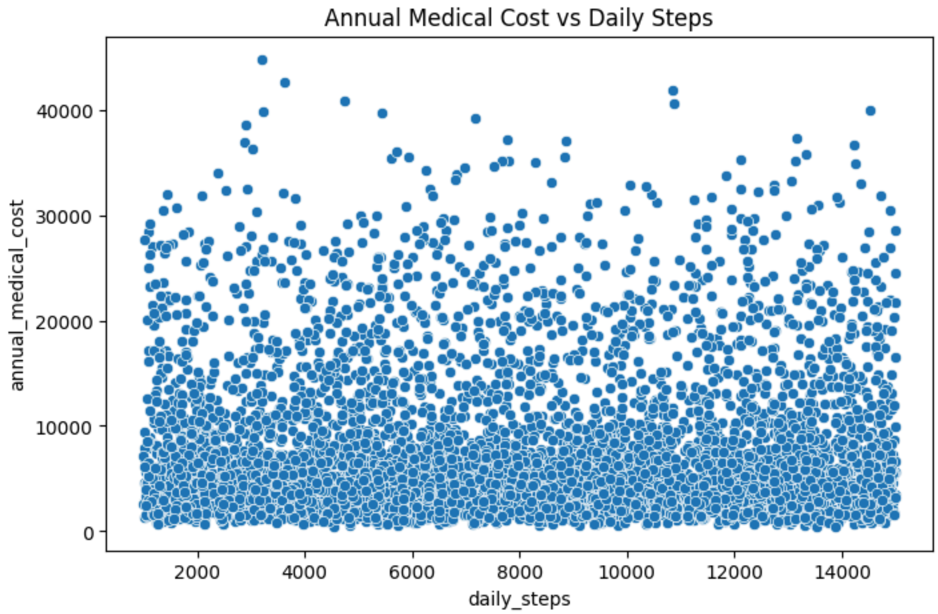


Insight:

Higher BMI is associated with increased costs, especially for smokers, suggesting a compounding risk effect.

```
# Daily Steps vs Cost
plt.figure(figsize=(8,5))
```

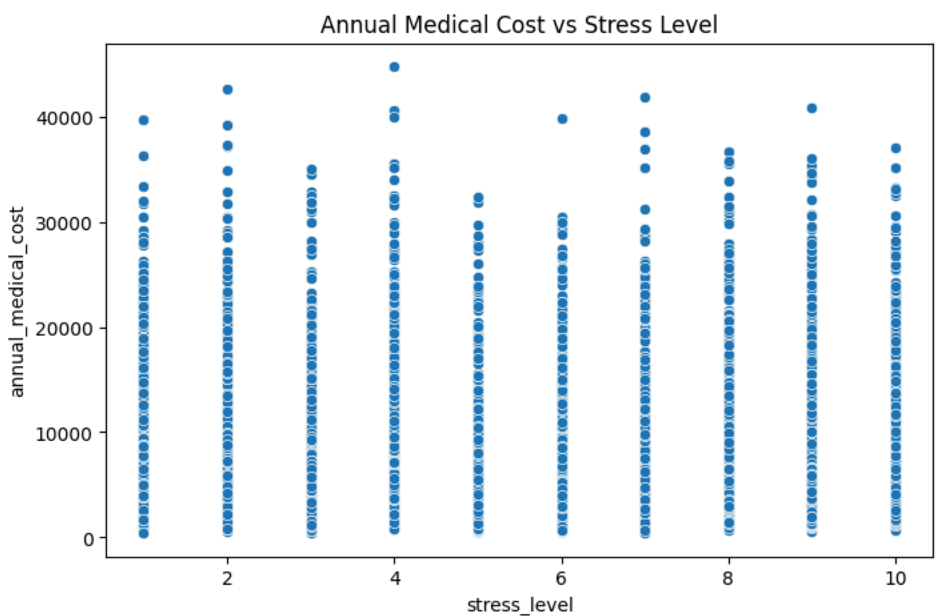
```
sns.scatterplot(x='daily_steps', y='annual_medical_cost', data=df)
plt.title('Annual Medical Cost vs Daily Steps')
plt.show()
```



Insight:

Patients with fewer daily steps tend to have slightly higher costs, though the relationship is weak, indicating lifestyle activity may have a moderate impact.

```
# Stress Level vs Cost
plt.figure(figsize=(8,5))
sns.scatterplot(x='stress_level', y='annual_medical_cost', data=df)
plt.title('Annual Medical Cost vs Stress Level')
plt.show()
```

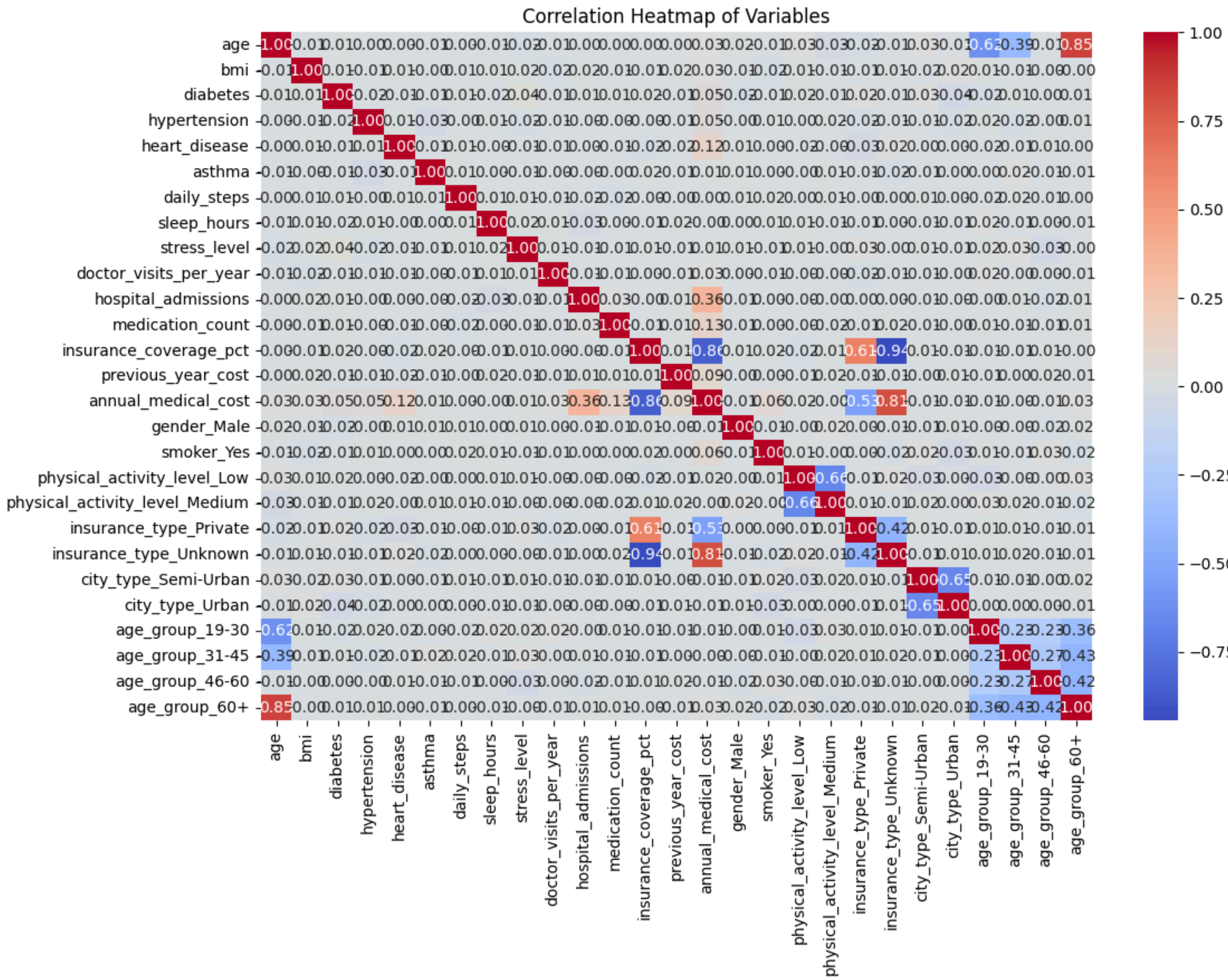


Insight:

Higher stress levels appear modestly associated with increased medical costs, suggesting stress management may influence healthcare utilization.

```
# 5.3 Heatmap / Correlation
plt.figure(figsize=(12,8))
```

```
sns.heatmap(df_encoded.corr(), annot=True, fmt=".2f", cmap='coolwarm')
plt.title('Correlation Heatmap of Variables')
plt.show()
```



Insight:

Annual medical cost is most strongly correlated with previous_year_cost, doctor_visits_per_year, and chronic conditions, confirming their predictive value.

Step 5 Summary

Strong cost drivers: Smoking, age, BMI, chronic conditions, insurance type, previous-year cost

Moderate drivers: Stress level, physical activity, daily steps

Minor impact: Gender and city type

Business insight: Targeting high-risk groups (smokers, elderly, overweight, chronic conditions) is essential for cost management strategies.

```
# Step 6: High-Cost Patient Analysis
'''Identify the patients who contribute disproportionately to total medical costs
Understand their demographic, lifestyle, and health profiles
Generate actionable insights for healthcare cost management'''

# 6.1 Define High-Cost Threshold
threshold = df['annual_medical_cost'].quantile(0.95)
```

```
high_cost = df[df['annual_medical_cost'] > threshold]
print(f"High-cost threshold (95th percentile): ${threshold:.2f}")
print(f"Number of high-cost patients: {high_cost.shape[0]}")
```

```
High-cost threshold (95th percentile): $23684.65
Number of high-cost patients: 250
```

Insight:
The top 5% of patients (250 individuals) have annual medical costs exceeding **$23,684.65**, representing the most resource-intensive population. These patients are predominantly older, have multiple chronic conditions, higher BMI, and a higher prevalence of smoking, making them prime candidates for targeted interventions and preventive care programs.

```
# 6.2 Overview of High-Cost Patients
high_cost.describe(include='all')
```

	age	gender	bmi	smoker	diabetes	hypertension	heart_disease	asthma	physical_activity_level	dail
count	250.000000	250	250.000000	250	250.000000	250.000000	250.000000	250.000000	250	25
unique	NaN	2	NaN	2	NaN	NaN	NaN	NaN	3	
top	NaN	Female	NaN	No	NaN	NaN	NaN	NaN	Low	
freq	NaN	143	NaN	162	NaN	NaN	NaN	NaN	108	
mean	55.764000	NaN	26.714400	NaN	0.272000	0.360000	0.276000	0.104000	NaN	787
std	20.970918	NaN	4.979955	NaN	0.445883	0.480963	0.447914	0.305873	NaN	410
min	18.000000	NaN	13.500000	NaN	0.000000	0.000000	0.000000	0.000000	NaN	101
25%	37.000000	NaN	23.200000	NaN	0.000000	0.000000	0.000000	0.000000	NaN	454
50%	57.000000	NaN	26.700000	NaN	0.000000	0.000000	0.000000	0.000000	NaN	746
75%	75.000000	NaN	30.400000	NaN	1.000000	1.000000	1.000000	0.000000	NaN	1168
max	89.000000	NaN	38.100000	NaN	1.000000	1.000000	1.000000	1.000000	NaN	1499

11 rows × 21 columns

One-line insights:

Age: Mostly older adults, confirming age as a key cost driver

Chronic conditions: Higher prevalence of diabetes, hypertension, heart disease, and asthma

Lifestyle: Many have higher BMI and lower physical activity levels

Insurance: Higher variation in insurance type; coverage can affect out-of-pocket costs

Utilization: More doctor visits, hospital admissions, and medications

```
# 6.3 Visualizing High-Cost Patient Profiles
# Age Distribution
sns.histplot(high_cost['age'], bins=10, kde=True)
plt.title('Age Distribution of High-Cost Patients')
plt.show()
```