HEALTHCARE COST & INSURANCE ANALYSIS

Objective:

The goal of this analysis is to explore and understand key factors influencing healthcare insurance charges using a real-world (messy) dataset. The analysis includes:

- Data cleaning & preprocessing
- Exploratory data analysis (EDA)
- Feature engineering
- Statistical visualizations
- Business insights

Data Cleaning Summary:

Task	Description
Missing Values	Null values in `age`, `bmi` handled using median or domain knowledge
Data Inconsistency	Normalize Categorical Columns
Inconsistent Text	Normalized case in columns like `smoker`, `region`
Duplicates	Detected and removed exact duplicate rows
Data Types	Fixed types `charges` (float → int), and categorical columns
Binary Encoding	Applied to `sex, smoker` and `children` for pre-processing
One-Hot Encoding	Applied to `region` for model-readiness

Exploratory Data Analysis (EDA)

- **1.Summary Statistics**

```
#load dataset
df=pd.read_csv(r"C:/Users/User-PC/Downloads/Healthcare Cost & Insurance Analysis/cleaned_insurance_dataset.csv")
print(df.describe().round())
```

```
          age     sex     bmi  children  smoker  charges
count  1338.0  1338.0  1338.0    1338.0  1338.0   1338.0
mean     39.0     1.0    31.0       1.0     0.0  13270.0
std      14.0     1.0     6.0       1.0     0.0  12110.0
min      18.0     0.0    16.0       0.0     0.0   1122.0
25%      27.0     0.0    26.0       0.0     0.0   4740.0
50%      39.0     1.0    30.0       1.0     0.0   9382.0
75%      51.0     1.0    35.0       2.0     0.0  16640.0
max      64.0     1.0    53.0       5.0     1.0  63770.0
```

- **2.Categorical Features Count**

```
#Value Counts for Categorical Features
smoker_distribution=df['smoker'].value_counts()
print("Smoker Distribution", smoker_distribution)

childrens=df['children'].value_counts()
print("Children Distribution", childrens)

gender=df.groupby('sex')['age'].sum()
print("\nPatient's Gender\n", gender)

age=df['bmi'].sum()
print("\nAverage BMI\n", age)

bmi=df['bmi'].mean()
print("\nAverage BMI\n", bmi)
```

Output:

```
Smoker Distribution smoker
0    1064
1     274
Name: count, dtype: int64
Children Distribution children
0    574
1    324
2    240
3    157
4     25
5     18
Name: count, dtype: int64

Patient's Gender
 sex
0    26151
1    26308
Name: age, dtype: int64

Average BMI
 41028

Average BMI
 30.663677130044842
```

- **3. Medical Charges based on Age Group**

```
#Age grorup with highest or lowest medical charges
age_group= df.nlargest(10,'charges')[['age','sex','bmi','children','smoker']]
print("\nAge group with highest charges:\n", age_group)

age_group2= df.nsmallest(10,'charges')[['age','sex','bmi','children','smoker']]
print("\nAge group with lowest charges:\n", age_group2)
```

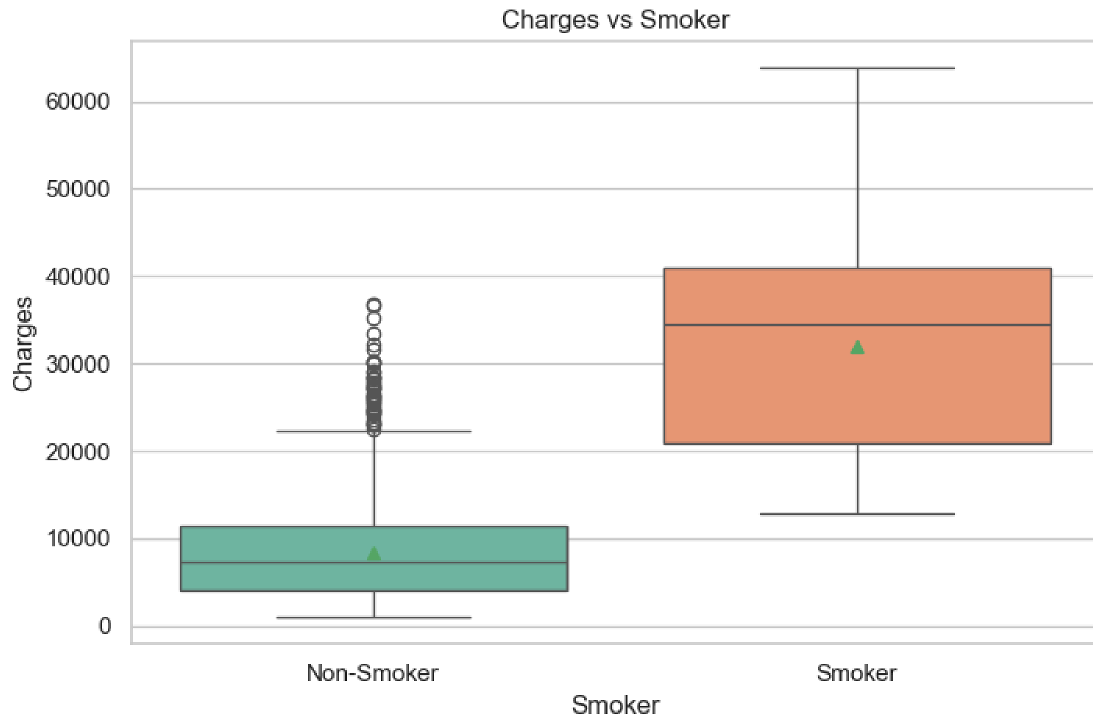
Output:

```
Age group with highest charges:
      age  sex  bmi  children  smoker
543    54    0   47         0       1
1300   45    1   30         0       1
1230   52    1   34         3       1
577    31    0   38         1       1
819    33    0   36         0       1
1146   60    1   33         0       1
34     28    1   36         1       1
1241   64    1   37         2       1
1062   59    1   41         1       1
488    44    0   38         0       1

Age group with lowest charges:
      age  sex  bmi  children  smoker
940    18    1   23         0       0
808    18    1   30         0       0
663    18    1   34         0       0
1244   18    1   33         0       0
22     18    1   34         0       0
194    18    1   34         0       0
866    18    1   37         0       0
781    18    1   41         0       0
442    18    1   43         0       0
1317   18    1   53         0       0
```

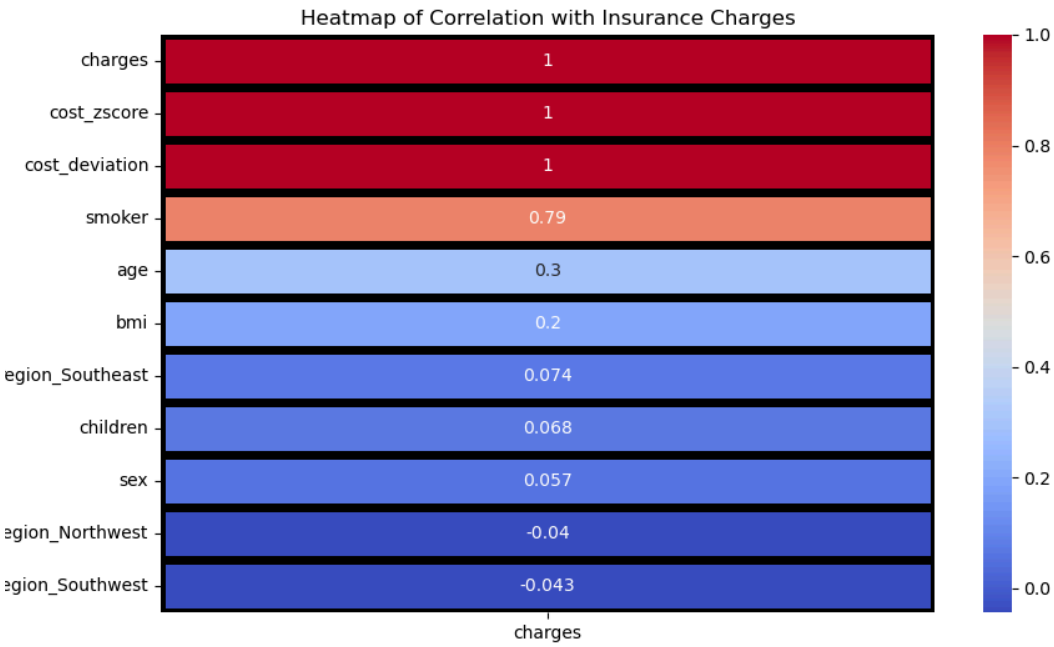
4.Visualization

Boxplot: Charges vs. Smoker



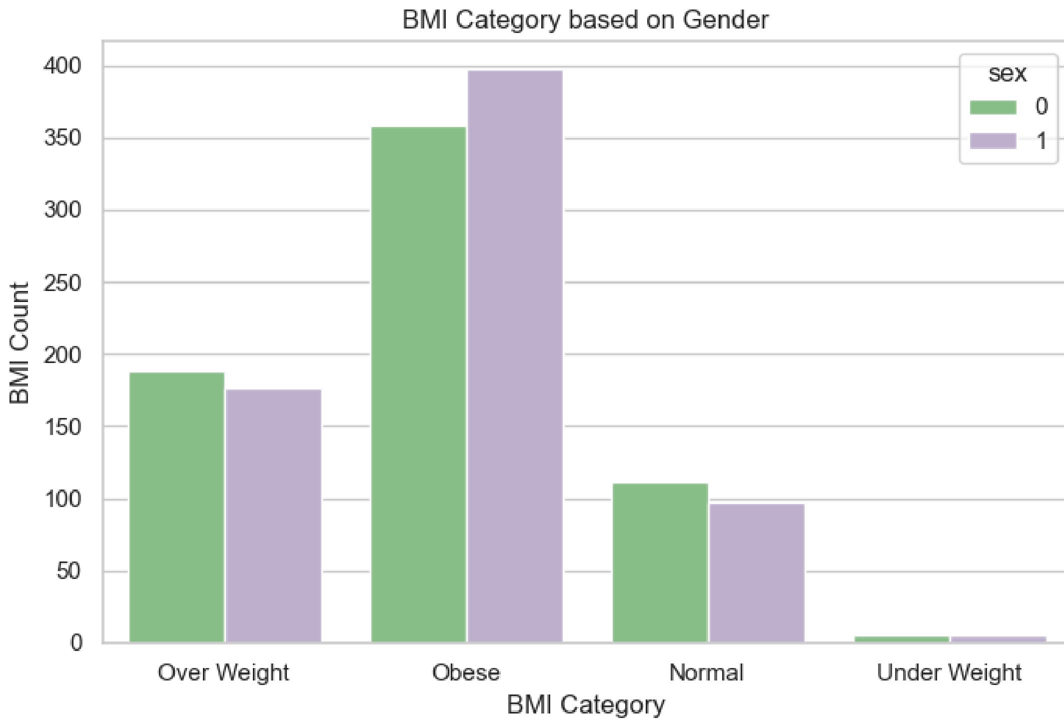
- Smokers have significantly higher median and mean charges.
- Large number of outliers among smokers.
- Clear positive skew in smoker group charges.

Heatmap: Feature Correlation with Charges



Features Correlation with charges
Smoker: 0.79
Age: 0.30
BMI: 0.20
Children, sex, Region: <0.10(Weak Correlation)

Countplot:BMI Category based on Gender

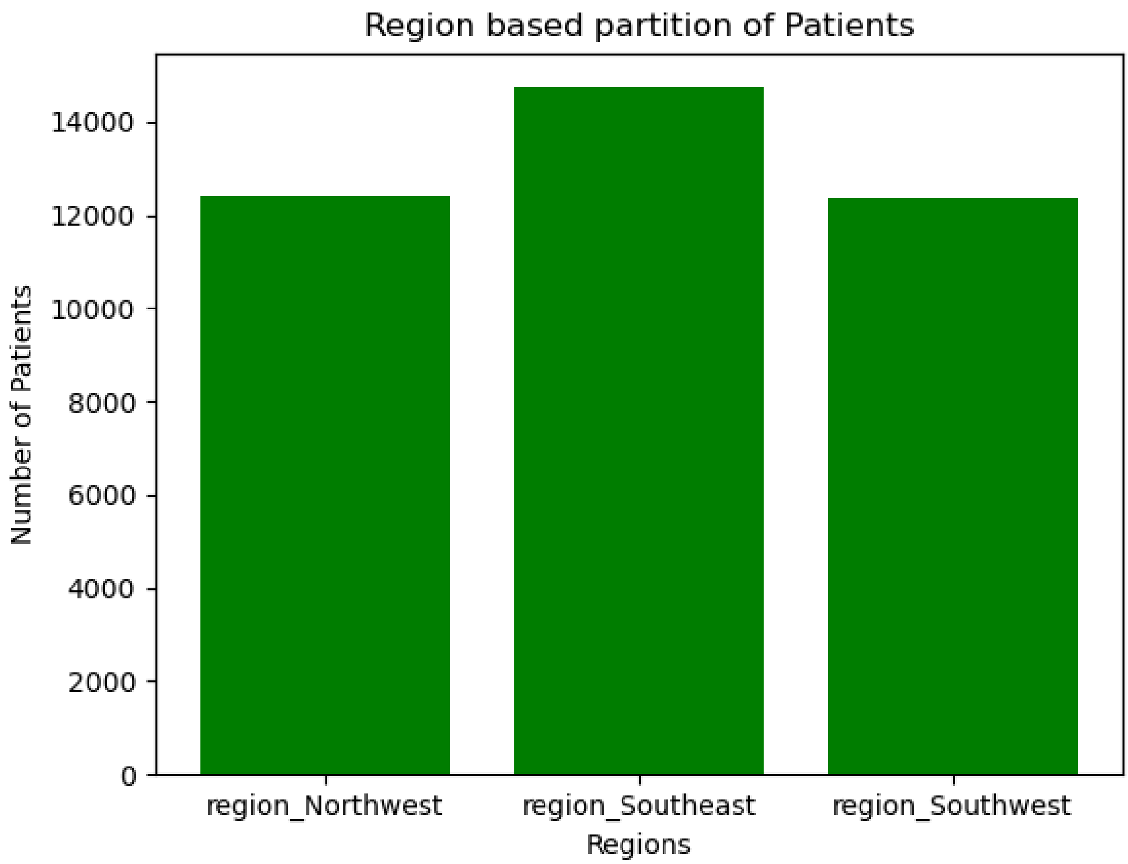


Over Weight
Male: 175
Female:185

Obese
Male:399
Female:355

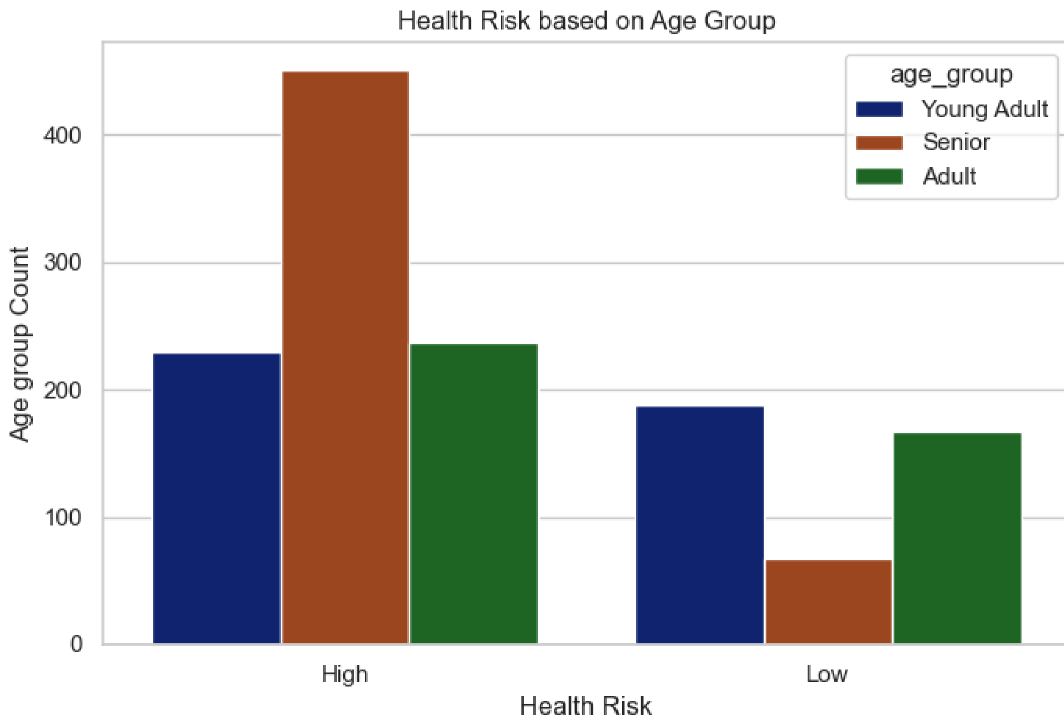
Normal
Male:110
Female: 98

Patients: Region Based Partition



Region Northwest: 12000
Region Southeast: 14000 Above
Region Southwest: 1200

Barplot: Health Risk based on Age group



HIGH

- Seniors' health are on high risk based on age, smoking and bmi.
- 200+ Young Adults are prone to health risk
- 200+ Adult as well are on high health risk

LOW

Senior age group is less in low health risk.
Less than 200 Adults and Young Adults are less prone to health risk.

5. Feature Engineering

Added Features

- **Cost Deviation**

```
mean_cost=np.mean(df['charges'])
std_cost=np.std(df['charges'])

#Cost Deviation Calculation
df['cost_deviation']= df['charges'] - mean_cost
df['cost_zscore']=df['charges'] - mean_cost/std_cost

print(df[['charges','cost_deviation','cost_zscore']])
```

- **Cost Score**

```
mean_cost=np.mean(df['charges'])
std_cost=np.std(df['charges'])

#Cost Deviation Calculation
df['cost_deviation']= df['charges'] - mean_cost
df['cost_zscore']=df['charges'] - mean_cost/std_cost

print(df[['charges','cost_deviation','cost_zscore']])
```

- **Regional Disparities**

```
#hot encoding for 'region'
df=pd.get_dummies(df, columns=['region'], drop_first=True)
print(df.head(50))
```

Region Southeast, Region Southwest, Region Northwest

- **Age group**

```
#Age_group
def age_group(age):
    if age <18:
        return 'Child'
    elif 18<= age < 30:
        return 'Young Adult'
    elif 35<= age < 50:
        return 'Adult'
    else:
        return 'Senior'

df['age_group']=df['age'].apply(age_group)
```

- **Bmi_category**

```python
#BMI Group

def bmi_category(bmi):
    if bmi < 18:
        return 'Under Weight'
    elif 18<= bmi <25:
        return 'Normal'
    elif 25<= bmi <30:
        return 'Over Weight'
    else:
        return 'Obese'

df['bmi_category']=df['bmi'].apply(bmi_category)
```

- **Health Risk**

```python
#Health Risk
def health_risk(row):
    if row['smoker'] == 1 or row['age']> 50 or row['bmi'] > 30:
        return 'High'
    else:
        return 'Low'
df['health_risk']= df.apply(health_risk, axis=1)
```

6. Key Business Insights

Insight	Implication
Smokers cost ~3x more	Risk-based pricing essential
High BMI + Smoking = $$$	Offer weight management incentives
Younger non-smokers = low cost	Target these profiles for low-premium plans
Region-based pricing may help	Adjust pricing for high-cost regions like Southeast

7.Exported Files

Clean and Engineered Dataset

cleaned_insurance_dataset	6/4/2025 10:33 PM	Microsoft Excel C...	47 KB
insurance- base file	6/3/2025 9:43 PM	Microsoft Excel C...	55 KB
insurance_data_engineered	6/5/2025 7:42 PM	Microsoft Excel C...	123 KB