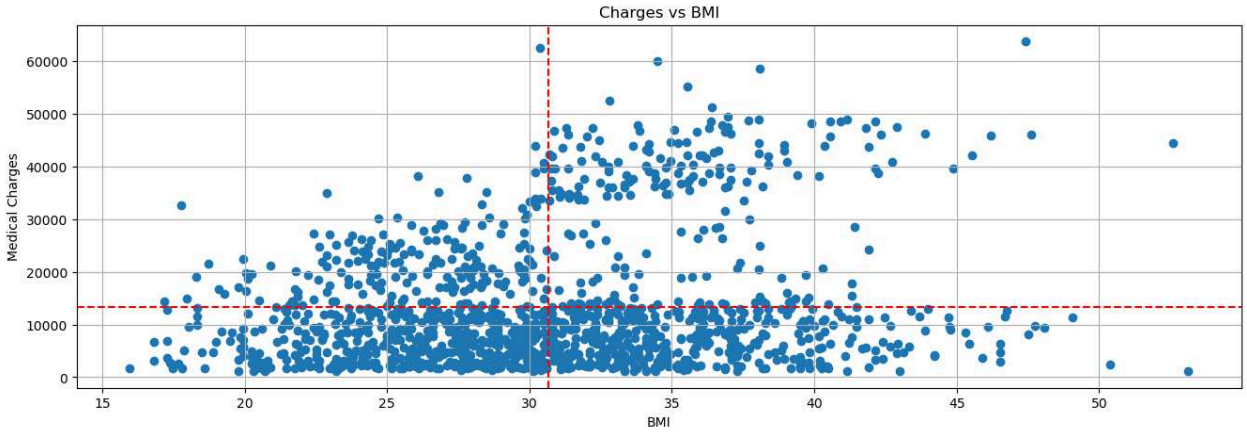
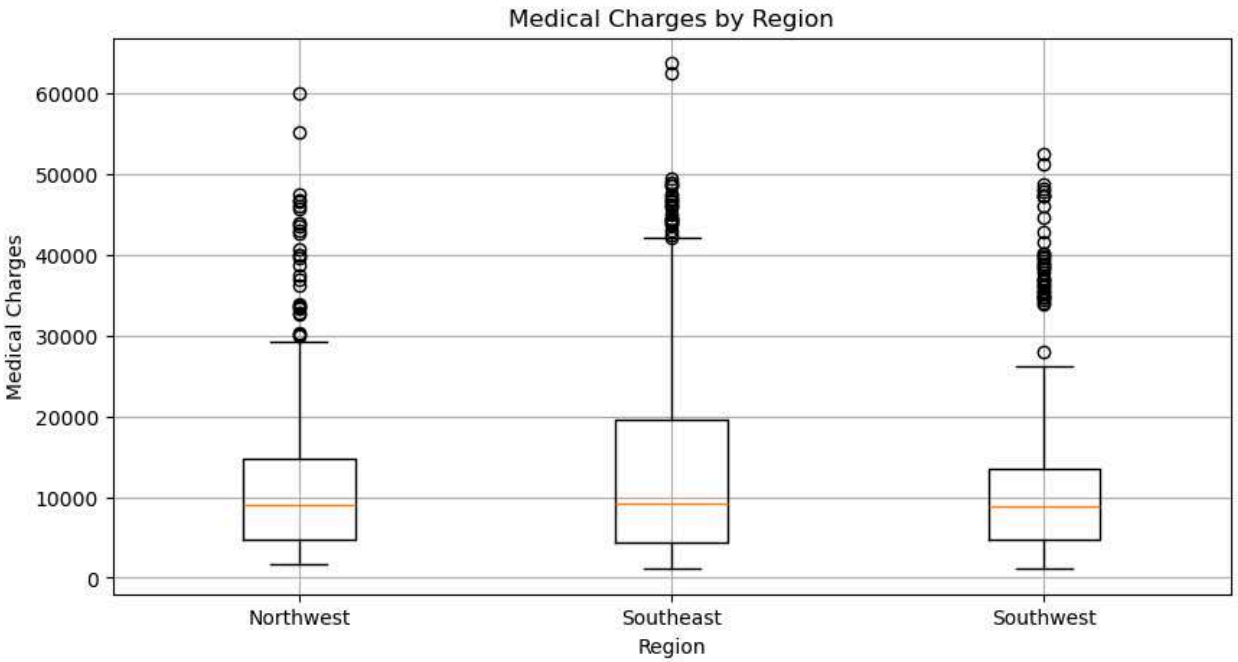
```
plt.figure(figsize = (16,5))
plt.scatter(df['bmi'], df['charges'])
plt.axhline(df['charges'].mean(),color = 'red',linestyle = '--')
plt.axvline(df['bmi'].mean(),color = 'red',linestyle = '--')
plt.xlabel('BMI')
plt.ylabel('Medical Charges')
plt.title('Charges vs BMI')
plt.grid()
plt.show()
```




```
<Figure size 640x480 with 0 Axes>
```

In [61]:
```
regions = ['region_northwest', 'region_southeast', 'region_southwest']
region_charges = [
    df[df['region_northwest'] == True]['charges'],
    df[df['region_southeast'] == True]['charges'],
    df[df['region_southwest'] == True]['charges']
]
plt.figure(figsize=(10, 5))
plt.boxplot(region_charges, labels=['Northwest', 'Southeast', 'Southwest'])
plt.xlabel('Region')
plt.ylabel('Medical Charges')
plt.title('Medical Charges by Region')
plt.grid()
plt.show()
```



Inference :

bmi vs charges graph shows a non-linear realtionshop between BMI and medical charges and the graph below the distribution of charges and median charges in different regions.

What does one row in your dataset represent in the real world?

Ans - One row represents the details(age,sex,region), lifestyle factors and health indicators.

Which column in your dataset is most useful for decision-making and why?

Ans - smoker (yes/no) as it significantly effects the health of the individual and finally impacting the net medical charges of the individuals.

Which column would you remove before ML modeling? Justify your choice.

Ans - raw categorical columns may be removed after encoding.

In [66]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 11 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   age               1338 non-null   int64
 1   bmi               1338 non-null   float64
 2   children          1338 non-null   int64
 3   charges           1338 non-null   float64
 4   smoker_yes        1338 non-null   bool
 5   sex_male          1338 non-null   bool
 6   region_northwest  1338 non-null   bool
 7   region_southeast  1338 non-null   bool
 8   region_southwest  1338 non-null   bool
 9   bmi_encoded       1338 non-null   int64
 10  age_grp_encoded   1338 non-null   int64
dtypes: bool(5), float64(2), int64(4)
memory usage: 69.4 KB
```

What type of bias might exist in your dataset?

Ans - It has lifetsyle bias as it assumes everyone and uniform access to healthcare and insurance across regions.

Is your dataset more suitable for classification or regression? Why?

Ans - The dataset is more suitable for regression as the taget values are continuous numerical value which makes the main purpose to predict an exact insurance cost.

In [ ]: