In [48]:

```
import pandas as pd
import matplotlib.pyplot as plt


# Provide the full path to the CSV file in the Downloads folder
file_path = r'C:\Users\giova\Downloads\python-portfolio-project-starter-files\insuranc

# Read the CSV file into a DataFrame
df = pd.read_csv(file_path)

# Display the first few rows of the DataFrame
df.head(100)
```

Out[48]:

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520
...	...	...	...	...	...	...	...
95	28	female	37.620	1	no	southeast	3766.88380
96	54	female	30.800	3	no	southwest	12105.32000
97	55	male	38.280	0	no	southeast	10226.28420
98	56	male	19.950	0	yes	northeast	22412.64850
99	38	male	19.300	0	yes	southwest	15820.69900

100 rows × 7 columns

Population Central Tendency

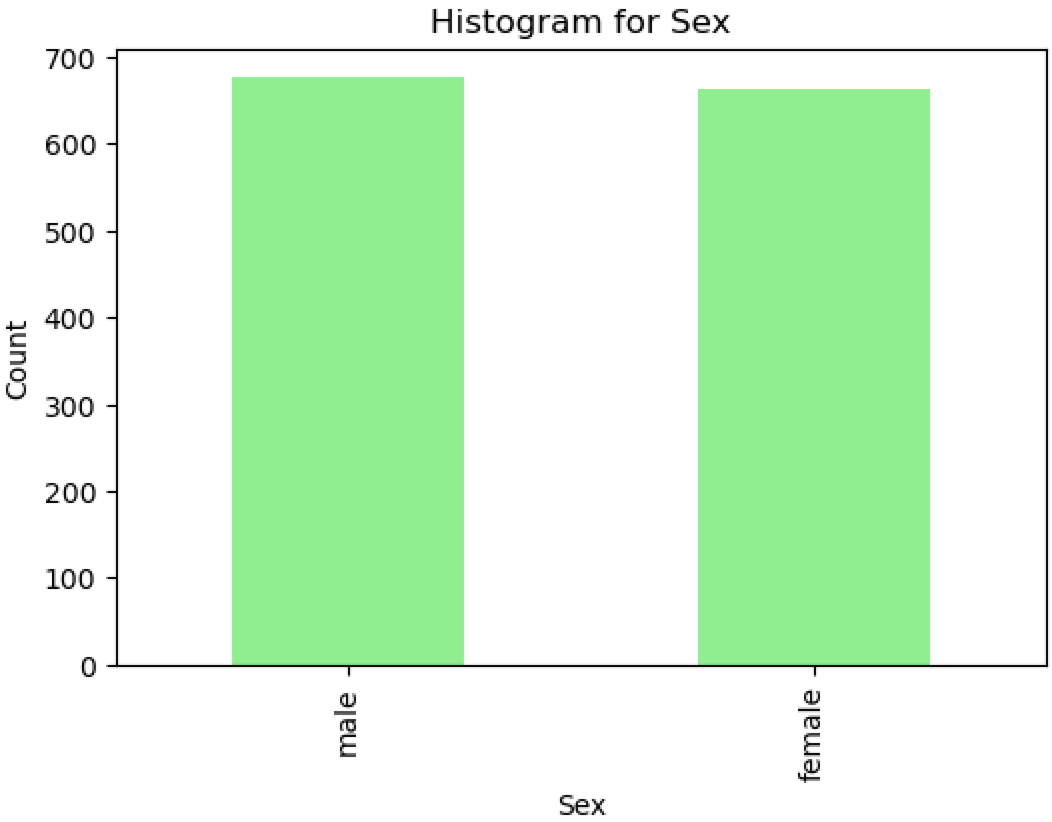
this is mostly to set a baseline of values for comparison

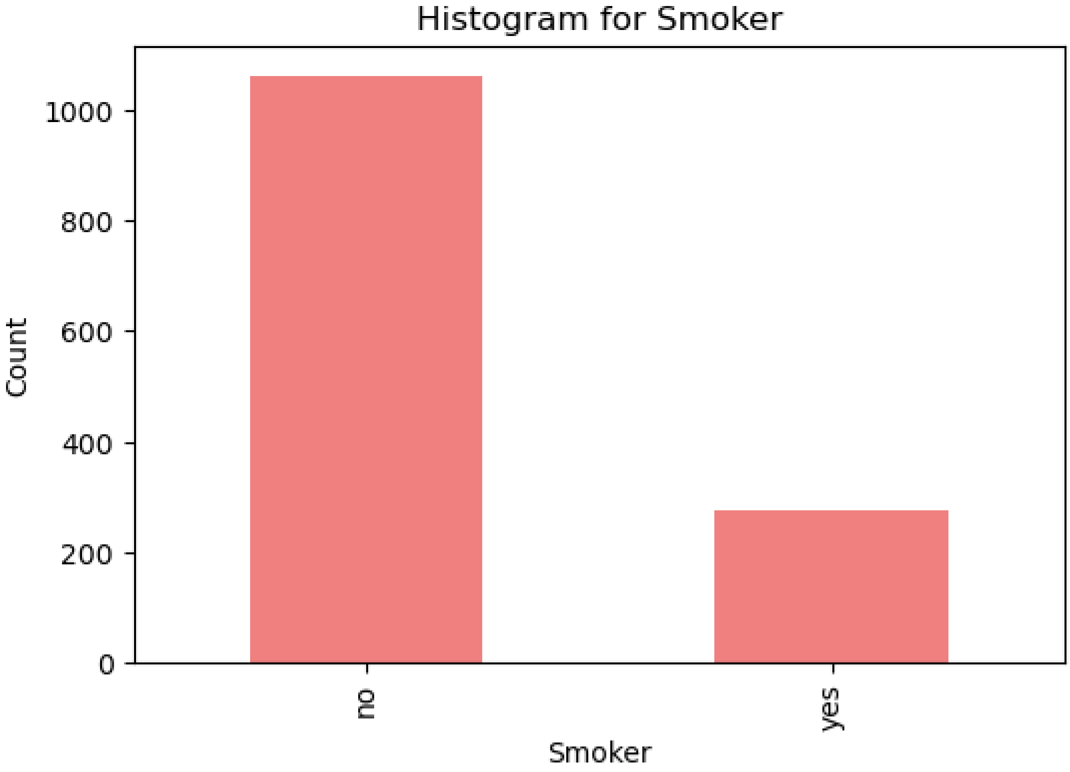
In [43]:

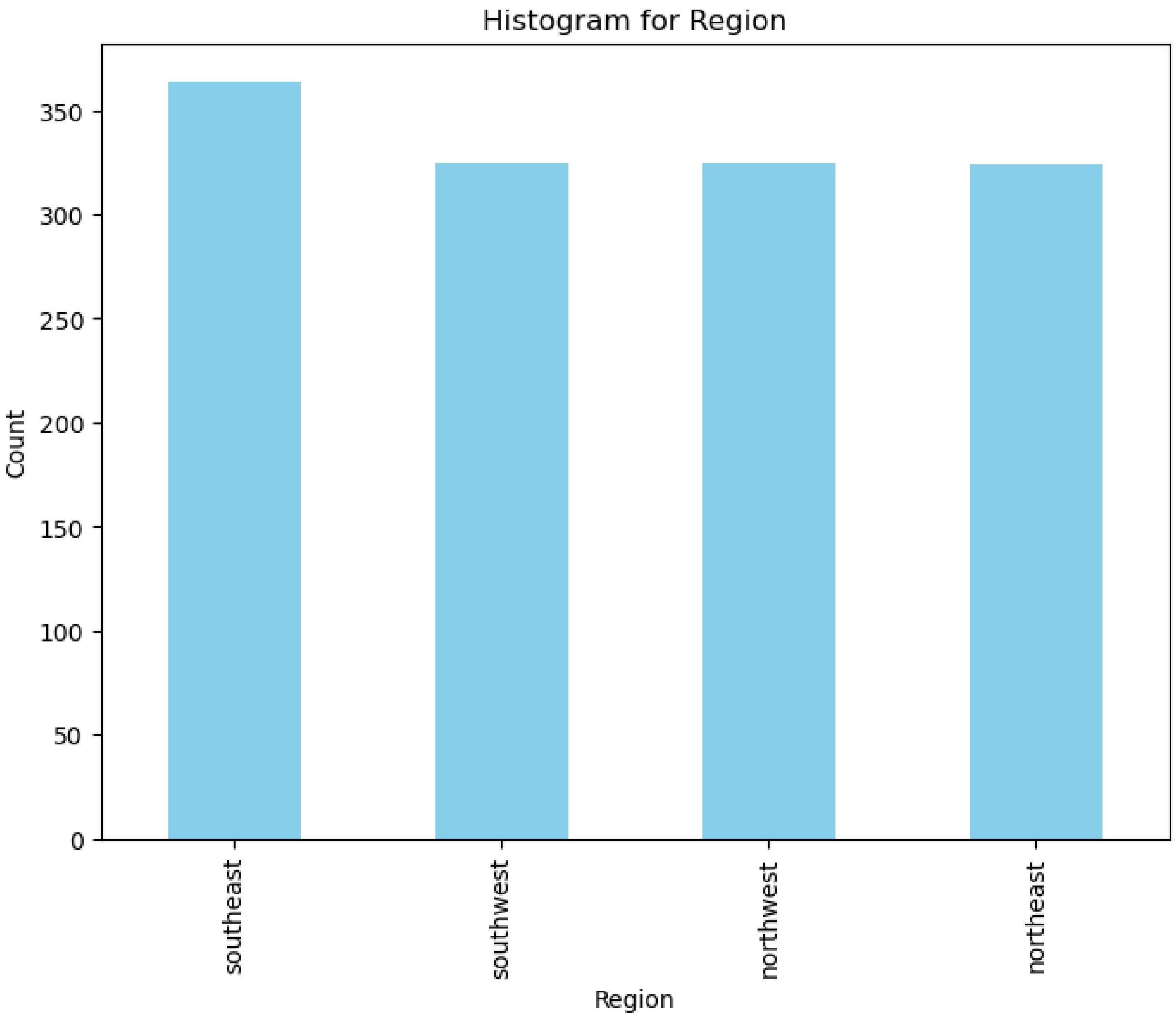
```
#Age is first, we are finding mean values for each column
pop_mean_age = df["age"].mean()
print("The population's mean age is " + str(pop_mean_age))

#Charges are next
pop_mean_charge = df["charges"].mean()
print("The population's mean charge is " + str(pop_mean_charge))

#Bmi is next
pop_mean_bmi = df["bmi"].mean()
print("The population's mean BMI  is " + str(pop_mean_bmi))

#mean number of children
pop_mean_kids = df["children"].mean()
```

```
print("The population's mean kids  are " + str(pop_mean_kids))

# Get breakdown for 'sex'
sex_breakdown = df['sex'].value_counts()

# Get breakdown for 'smoker'
smoker_breakdown = df['smoker'].value_counts()

# Get breakdown for 'region'
region_breakdown = df['region'].value_counts()

# Print the breakdowns
print("Breakdown for 'sex':")
print(sex_breakdown)

# Create a histogram for 'smoker'
plt.figure(figsize=(6, 4))
df['smoker'].value_counts().plot(kind='bar', color='lightcoral')
plt.title('Histogram for Smoker')
plt.xlabel('Smoker')
plt.ylabel('Count')
plt.show()


# Create a histogram for 'region'
plt.figure(figsize=(8, 6))
df['region'].value_counts().plot(kind='bar', color='skyblue')
plt.title('Histogram for Region')
plt.xlabel('Region')
plt.ylabel('Count')
plt.show()
```

```
The population's mean age is 39.20702541106129
The population's mean charge is 13270.422265141257
The population's mean BMI  is 30.663396860986538
The population's mean kids  are 1.0949177877429
```


Histogram for Sex
Count
700
600
500
400
300
200
100
0
male
female
Sex


Histogram for Smoker
Count
1000
800
600
400
200
0
no
yes
Smoker



Smoker/Non Smoker Analysis

In [12]:

```
# Filter the DataFrame to include only smokers
smokers_df = df[df['smoker'] == 'yes']
# Filter the DataFrame to include only nonsmokers
nonsmokers_df = df[df['smoker'] == 'no']


# Calculate the average age of smokers
average_age_of_smokers = smokers_df['age'].mean()
# Calculate the average age of nonsmokers
average_age_of_nonsmokers = nonsmokers_df['age'].mean()

print(f'The average age of smokers is: {average_age_of_smokers:.2f}')
print(f'The average age of nonsmokers is: {average_age_of_nonsmokers:.2f}')
```

```
The average age of smokers is: 38.51
The average age of nonsmokers is: 39.39
```

In [9]:

```
# Calculate the median age of smokers
median_age_of_smokers = smokers_df['age'].median()
# Calculate the median age of non-smokers
median_age_of_nonsmokers = nonsmokers_df['age'].median()
```

```
print(f'The median age of smokers is: {median_age_of_smokers:.2f}')
print(f'The median age of non-smokers is: {median_age_of_nonsmokers:.2f}')

slight, slight difference in median vs avg
```

```
The median age of smokers is: 38.00
The median age of non-smokers is: 40.00
```

# Calculate the average number of children for smokers average_children_of_smokers = smokers_df['children'].mean() # Calculate the average number of children for non-smokers average_children_of_nonsmokers = nonsmokers_df['children'].mean() print(f'The average number of children for smokers is: {average_children_of_smokers:.2f}') print(f'The average number of children for non-smokers is: {average_children_of_nonsmokers:.2f}') #There is no meaningful difference in average number of children between smoker/ nonsmoker #median was exactly the same in a now deleted analysis

In [47]:

```
# Filter the DataFrame to include only smokers and calculate the median charges
median_charges_smokers = df[df['smoker'] == 'yes']['charges'].median()

# Filter the DataFrame to include only non-smokers and calculate the median charges
median_charges_nonsmokers = df[df['smoker'] == 'no']['charges'].median()

# Print the median charges for both groups
print(f'Median charges for smokers: {median_charges_smokers:.2f}')
print(f'Median charges for non-smokers: {median_charges_nonsmokers:.2f}')
```

```
Median charges for smokers: 34456.35
Median charges for non-smokers: 7345.41
```

Checking Correlation with Charges agaist all variables

In [51]:

```
# Convert categorical variables to numerical using one-hot encoding
df_encoded = pd.get_dummies(df, columns=['region', 'sex', 'smoker'], drop_first=True)

# Calculate the correlation matrix
correlation_matrix_all = df_encoded.corr()

# Display correlations with 'charges' for all variables
charges_correlations_all = correlation_matrix_all['charges'].sort_values(ascending=Fal
print(charges_correlations_all)



#as one might have guessed, smoking, age, and bmi are correlated with higher insurance
#southeast is also correlated with higher charges
```

```
charges              1.000000
smoker_yes           0.787251
age                  0.299008
bmi                  0.198341
region_southeast     0.073982
children             0.067998
sex_male             0.057292
region_northwest    -0.039905
region_southwest    -0.043210
Name: charges, dtype: float64
```