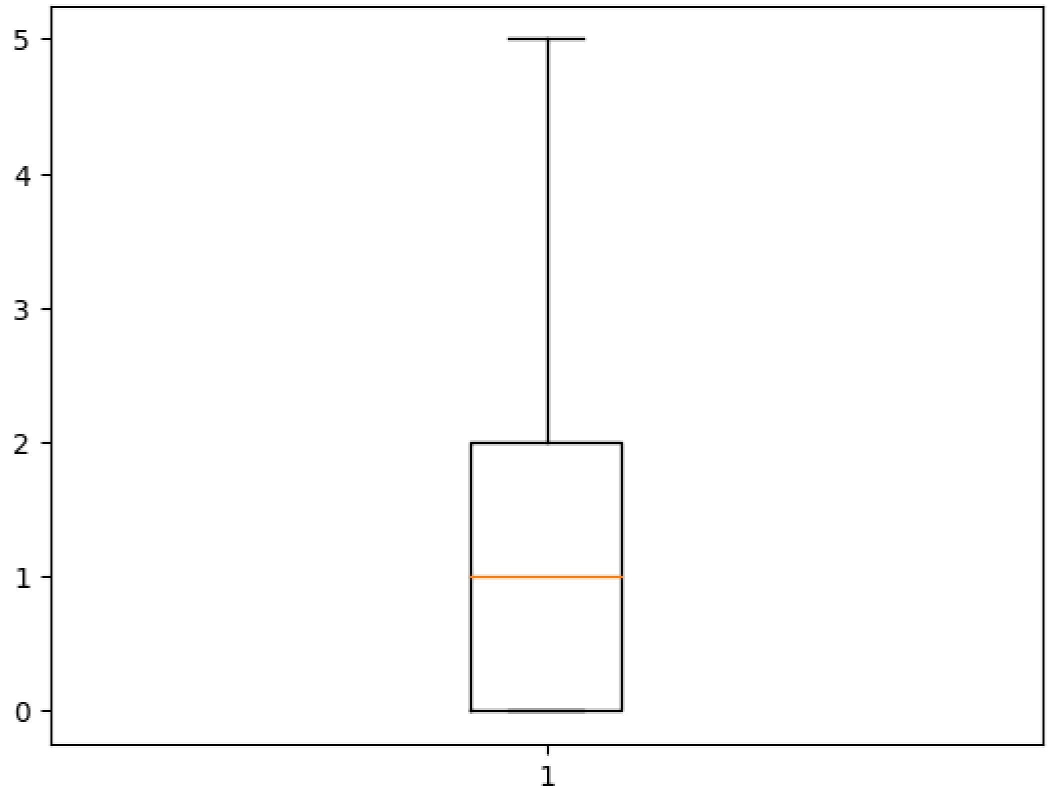
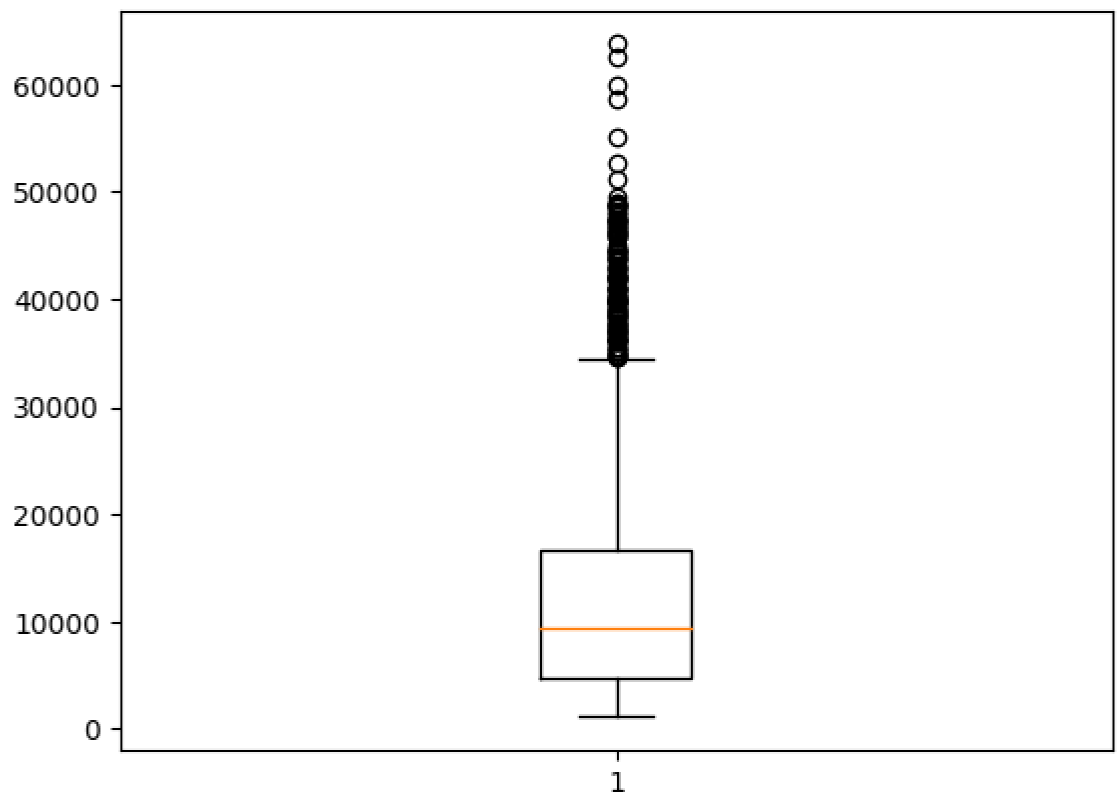
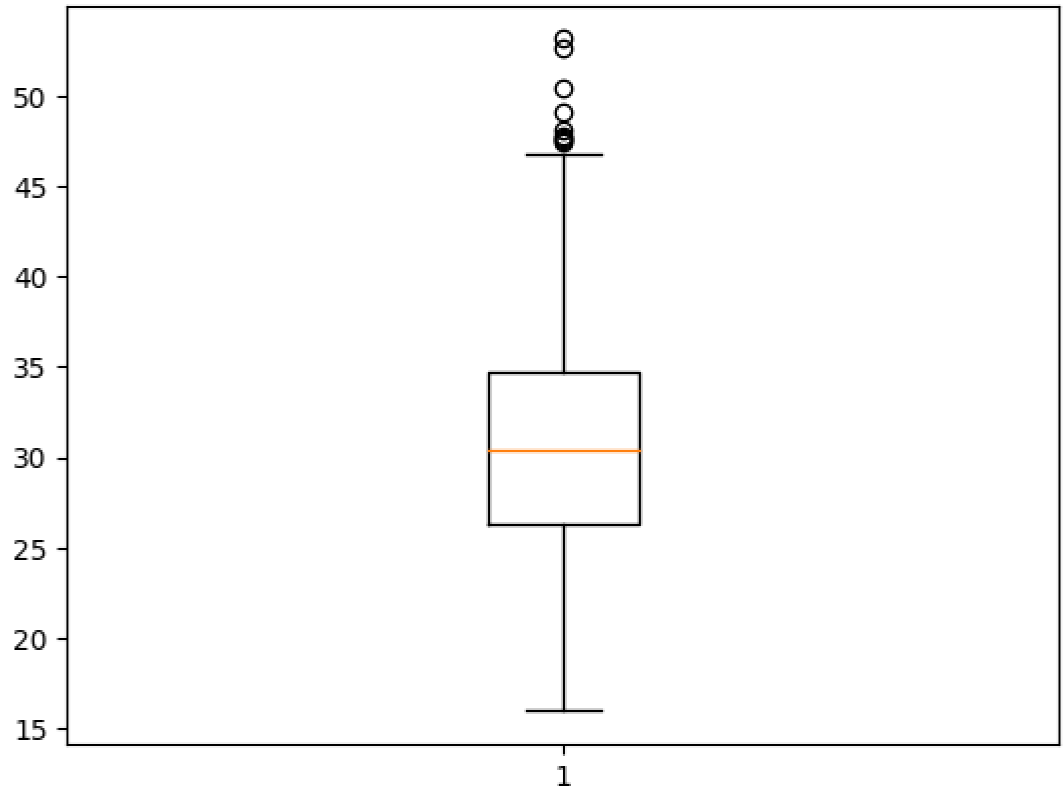
In [22]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.stats import pearsonr

# Read CSV file
data = pd.read_csv("C:/Users/user/Desktop/My learning/ClinSoft/expenses.csv")

# Display the first few rows of the dataframe
print(data.head())

# Summary statistics of the dataframe
print(data.describe())

# Boxplot for 'children', 'charges', and 'bmi'
plt.boxplot(data['children'])
plt.show()

plt.boxplot(data['charges'])
plt.show()

plt.boxplot(data['bmi'])
plt.show()
```

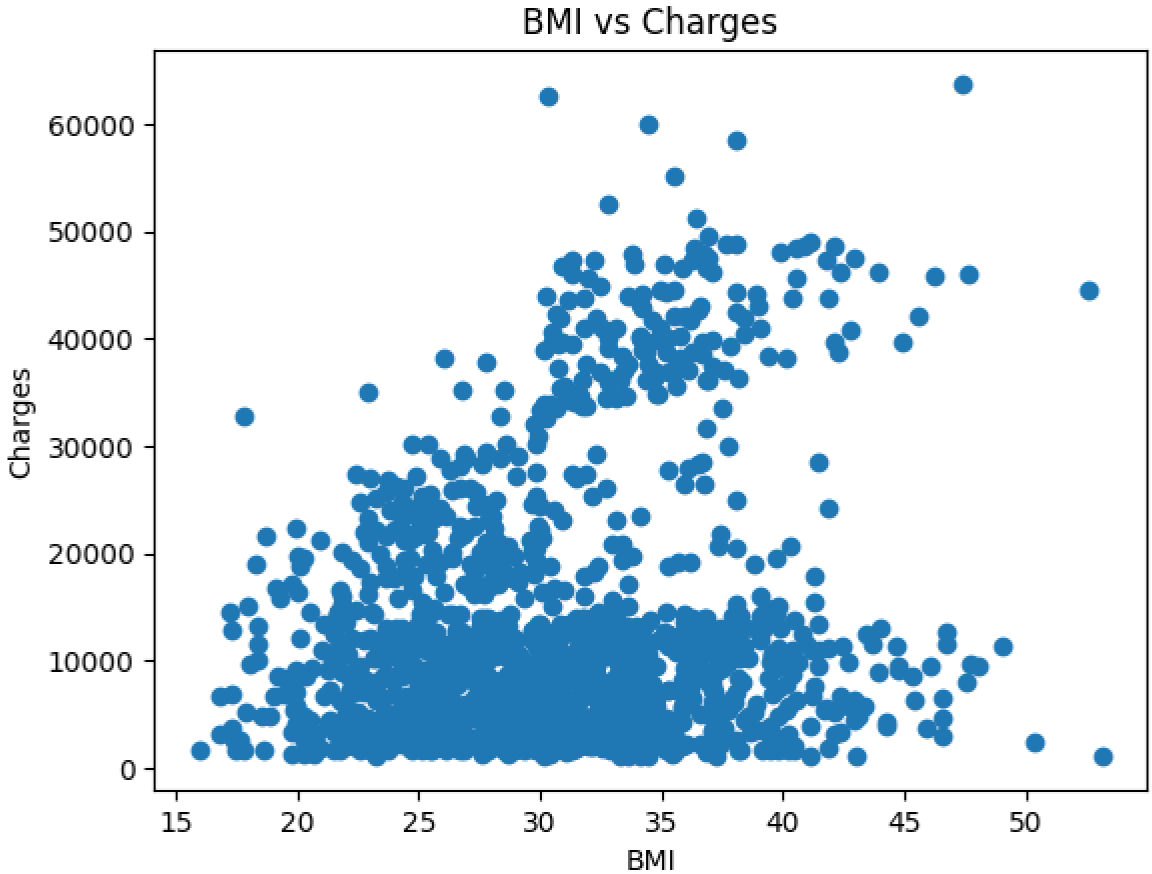
```
   age     sex     bmi  children smoker     region      charges
0   19  female  27.900         0    yes  southwest  16884.92400
1   18    male  33.770         1     no  southeast   1725.55230
2   28    male  33.000         3     no  southeast   4449.46200
3   33    male  22.705         0     no  northwest  21984.47061
4   32    male  28.880         0     no  northwest   3866.85520
               age          bmi     children       charges
count  1338.000000  1338.000000  1338.000000   1338.000000
mean     39.207025    30.663397     1.094918  13270.422265
std      14.049960     6.098187     1.205493  12110.011237
min      18.000000    15.960000     0.000000   1121.873900
25%      27.000000    26.296250     0.000000   4740.287150
50%      39.000000    30.400000     1.000000   9382.033000
75%      51.000000    34.693750     2.000000  16639.912515
max      64.000000    53.130000     5.000000  63770.428010
```

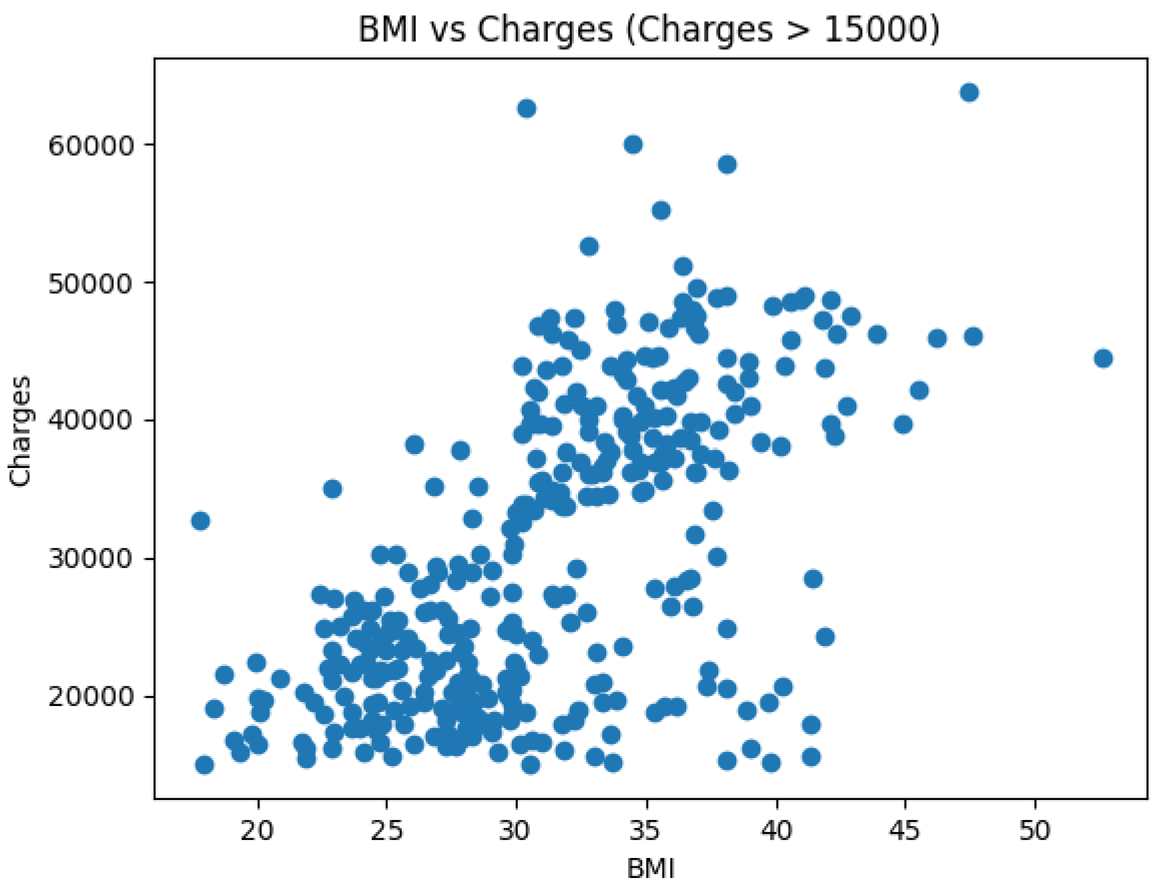

5
4
3
2
1
0
1

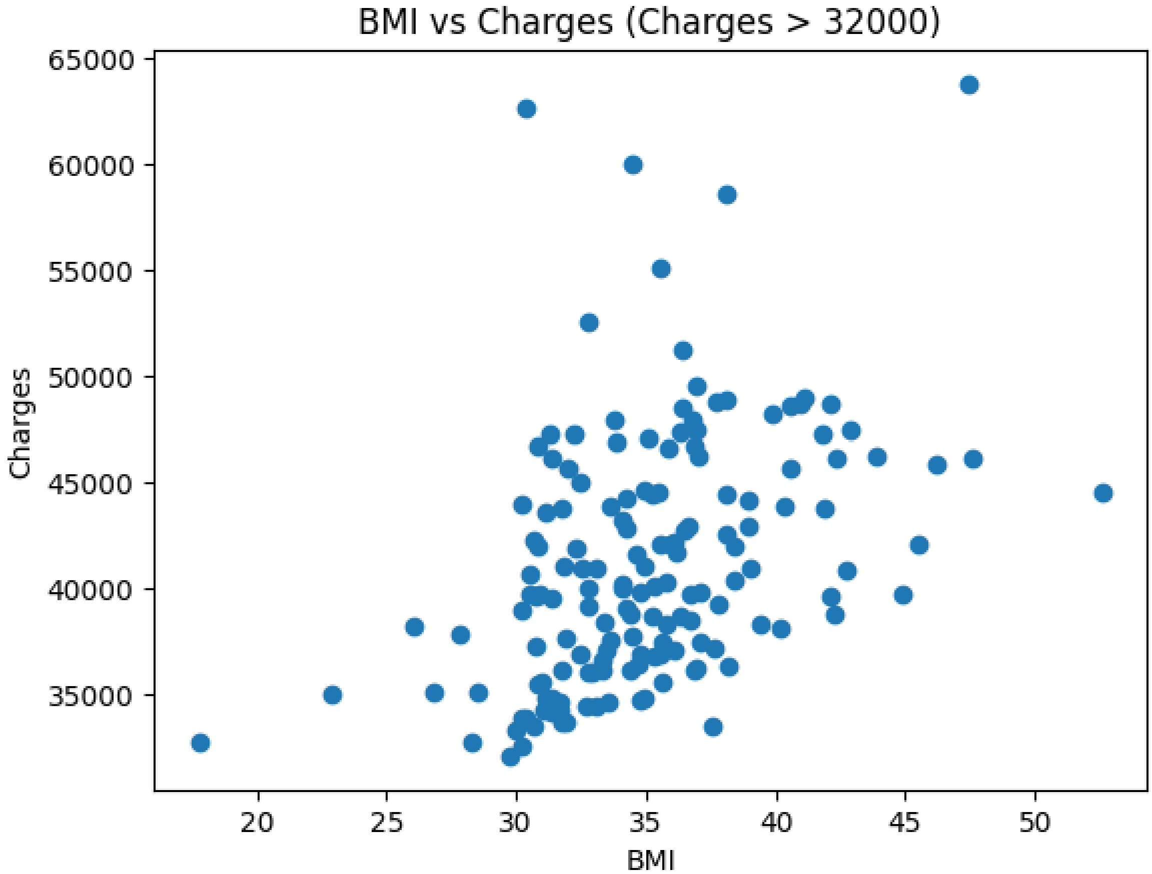

60000
50000
40000
30000
20000
10000
0
1

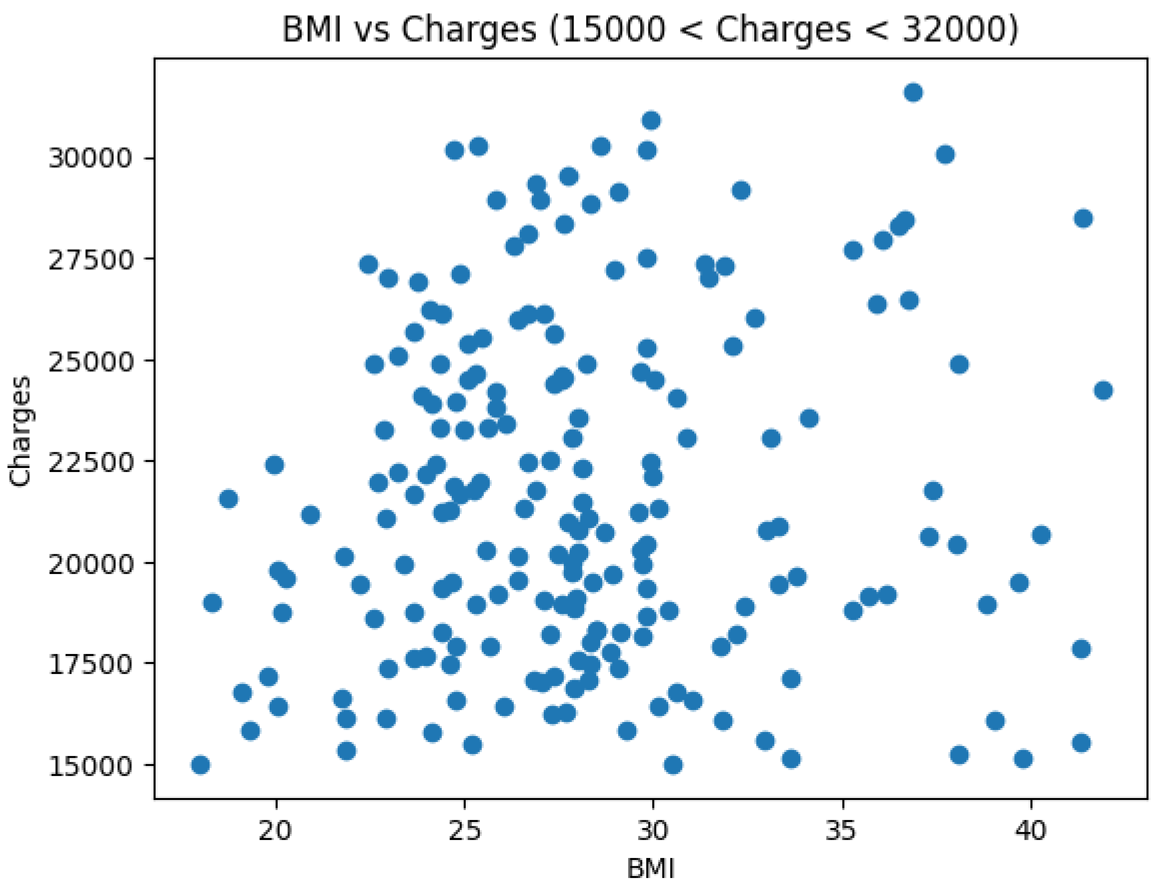


In [23]:
```
# Plot of 'bmi' vs 'charges'
plt.scatter(data['bmi'], data['charges'])
plt.xlabel('BMI')
plt.ylabel('Charges')
plt.title('BMI vs Charges')
plt.show()

# Plot of 'bmi' vs 'charges' for charges > 15000
plt.scatter(data.loc[data['charges'] > 15000, 'bmi'], data.loc[data['charges'] > 1500(
plt.xlabel('BMI')
plt.ylabel('Charges')
plt.title('BMI vs Charges (Charges > 15000)')
plt.show()

# Similar plots for charges > 32000 and 15000 < charges < 32000
plt.scatter(data.loc[data['charges'] > 32000, 'bmi'], data.loc[data['charges'] > 3200(
plt.xlabel('BMI')
plt.ylabel('Charges')
plt.title('BMI vs Charges (Charges > 32000)')
plt.show()

plt.scatter(data.loc[(data['charges'] < 32000) & (data['charges'] > 15000), 'bmi'],
            data.loc[(data['charges'] < 32000) & (data['charges'] > 15000), 'charges']
plt.xlabel('BMI')
plt.ylabel('Charges')
plt.title('BMI vs Charges (15000 < Charges < 32000)')
plt.show()
```


BMI vs Charges
Charges
BMI
0
10000
20000
30000
40000
50000
60000
15
20
25
30
35
40
45
50


BMI vs Charges (Charges > 15000)
Charges
BMI
20000
30000
40000
50000
60000
20
25
30
35
40
45
50





In [24]:
```
# Pearson correlation test between 'bmi' and 'charges'
correlation, p_value = pearsonr(data['bmi'], data['charges'])
```

```
print("Correlation between BMI and Charges:", correlation)
print("p-value:", p_value)
```

```
Correlation between BMI and Charges: 0.19834096883362887
p-value: 2.459085535117846e-13
```

In [14]:
```
# Create a new column 'overweight'
data['overweight'] = ['Over30' if x > 30 else 'Under30' for x in data['bmi']]
data['exp'] = ['High_Charge' if x > 15000 else "Low_Charge" for x in data['charges']]


# Summary of the 'overweight' column
print(data['overweight'].describe())
data['exp'].describe()
```

```
count       1338
unique         2
top       Over30
freq         705
Name: overweight, dtype: object
```

Out[14]:
```
count           1338
unique             2
top       Low_Charge
freq             980
Name: exp, dtype: object
```

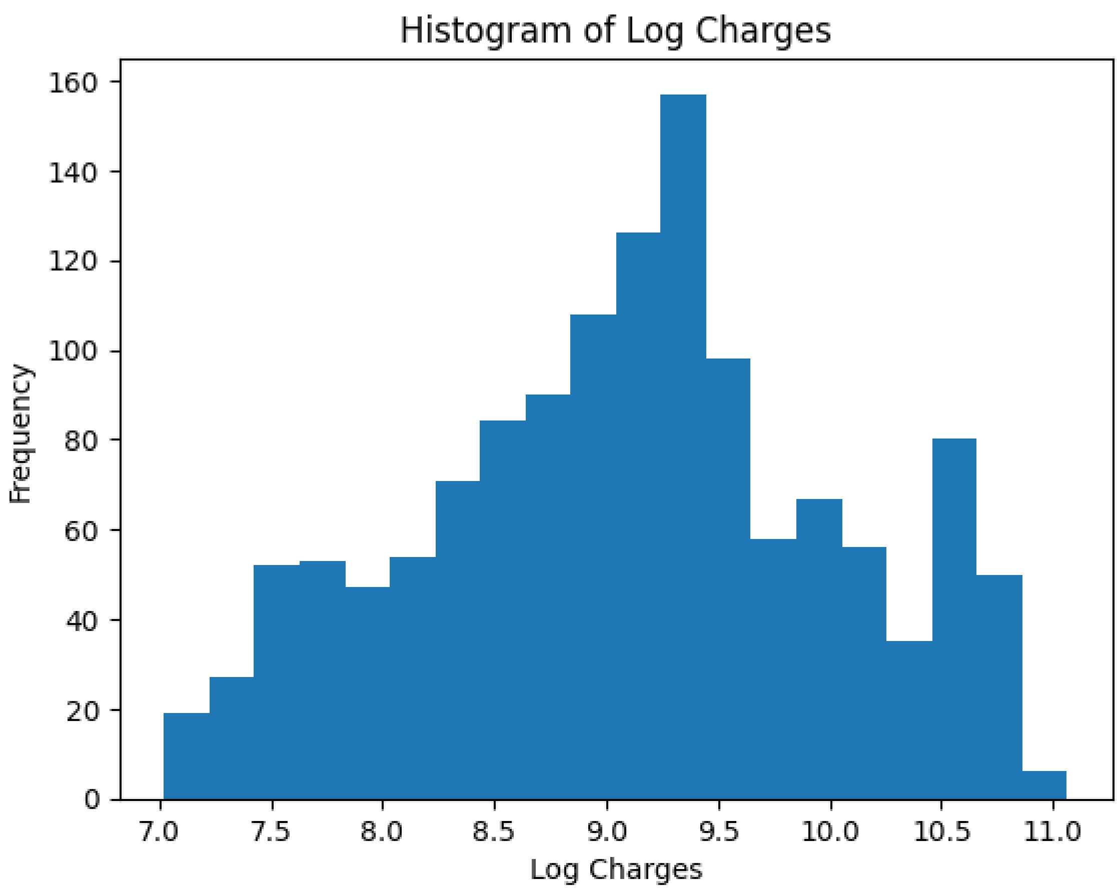
In [16]:
```
# Cross-tabulation of 'overweight' and 'exp'
cross_tab = pd.crosstab(data['overweight'], data['exp'], margins=True)
print(cross_tab)

# Cross-tabulation of 'children', 'smoker', and 'exp'
cross_tab_3d = pd.crosstab([data['children'], data['smoker']], data['exp'], margins=Tr
print(cross_tab_3d)

# Cross-tabulation of 'children', 'smoker', 'region', and 'exp'
cross_tab_4d = pd.crosstab([data['children'], data['smoker'], data['region']], data['e
print(cross_tab_4d)
```

```
exp         High_Charge  Low_Charge   All
overweight
Over30              200         505   705
Under30             158         475   633
All                 358         980  1338
exp              High_Charge  Low_Charge   All
children smoker
0        no               28         431   459
         yes             110           5   115
1        no               21         242   263
         yes              61           0    61
2        no               23         162   185
         yes              53           2    55
3        no               14         104   118
         yes              39           0    39
4        no                5          17    22
         yes               3           0     3
5        no                0          17    17
         yes               1           0     1
All                      358         980  1338
exp                         High_Charge  Low_Charge   All
children smoker region
0        no     northeast             7         114   121
                northwest             5         103   108
                southeast             9         108   117
                southwest             7         106   113
         yes    northeast            23           3    26
                northwest            23           1    24
                southeast            40           0    40
                southwest            24           1    25
1        no     northeast             7          48    55
                northwest             4          61    65
                southeast             7          66    73
                southwest             3          67    70
         yes    northeast            22           0    22
                northwest             9           0     9
                southeast            22           0    22
                southwest             8           0     8
2        no     northeast            10          32    42
                northwest             9          46    55
                southeast             2          46    48
                southwest             2          38    40
         yes    northeast             7           2     9
                northwest            11           0    11
                southeast            18           0    18
                southwest            17           0    17
3        no     northeast             0          29    29
                northwest             6          27    33
                southeast             6          18    24
                southwest             2          30    32
         yes    northeast            10           0    10
                northwest            13           0    13
                southeast            11           0    11
                southwest             5           0     5
4        no     northeast             3           4     7
                northwest             0           5     5
                southeast             1           4     5
                southwest             1           4     5
         yes    northwest             1           0     1
                southwest             2           0     2
```

```
5         no      northeast            0           3     3
                  northwest            0           1     1
                  southeast            0           6     6
                  southwest            0           7     7
          yes     southwest            1           0     1
All                                  358         980  1338
```

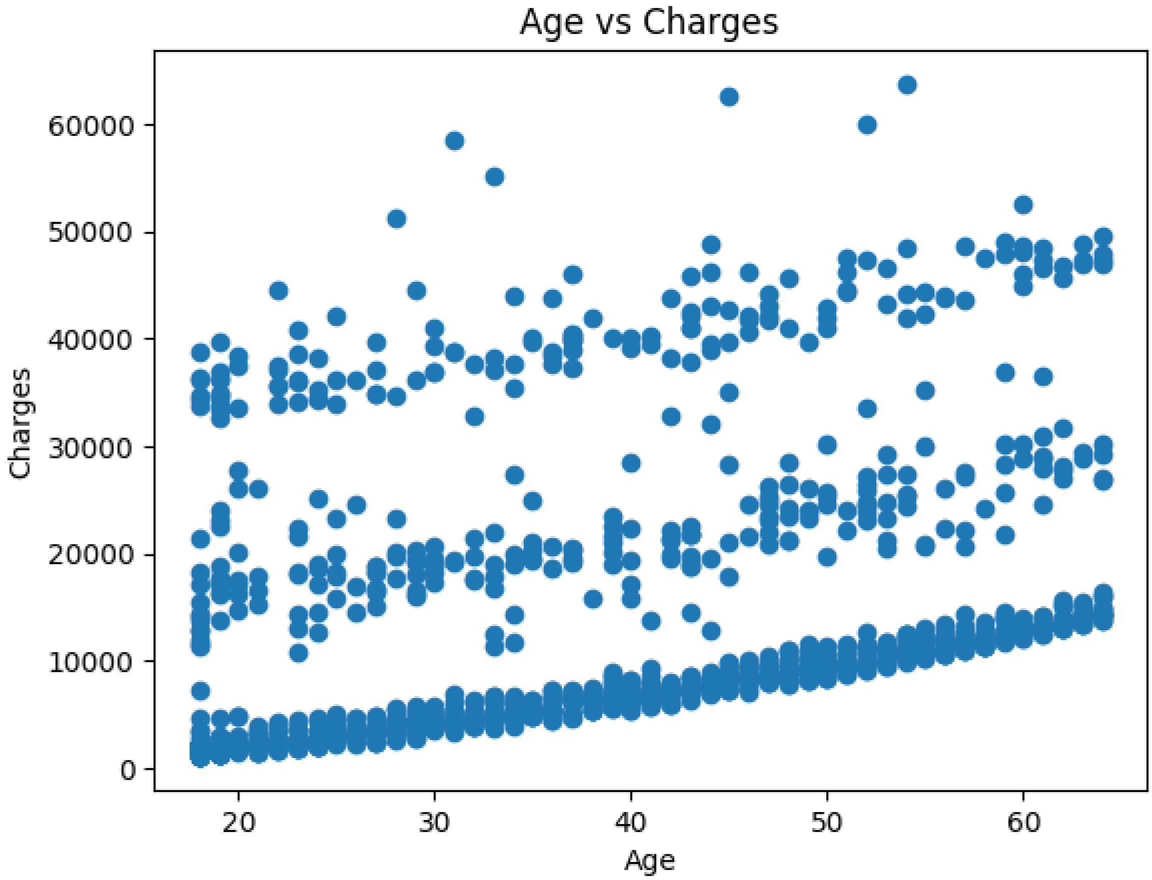
In [21]:

```
import math
# Histogram of logarithm of charges
plt.hist(data['charges'].apply(lambda x: math.log(x)), bins=20)
plt.xlabel('Log Charges')
plt.ylabel('Frequency')
plt.title('Histogram of Log Charges')
plt.show()
```

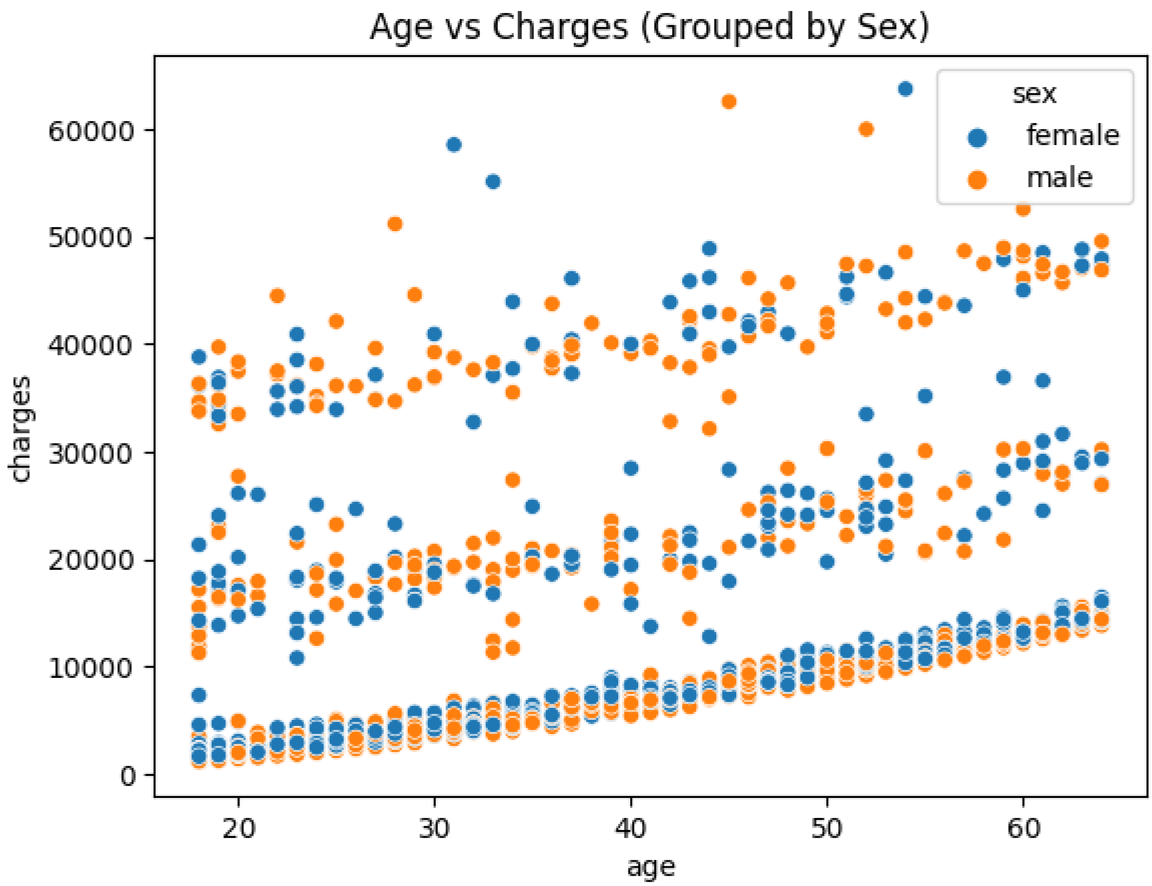


In [26]:

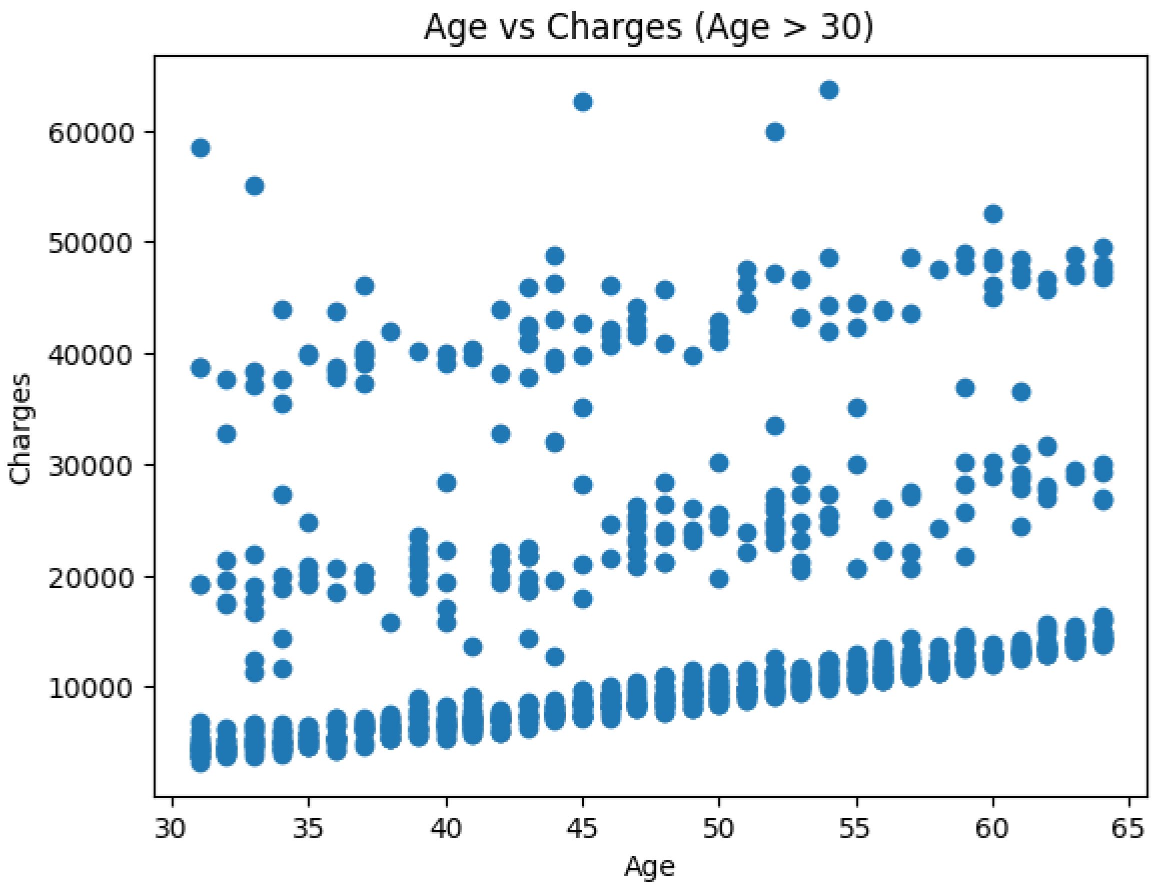
```
plt.scatter(data['age'], data['charges'])
plt.xlabel('Age')
plt.ylabel('Charges')
plt.title('Age vs Charges')
plt.show()
```



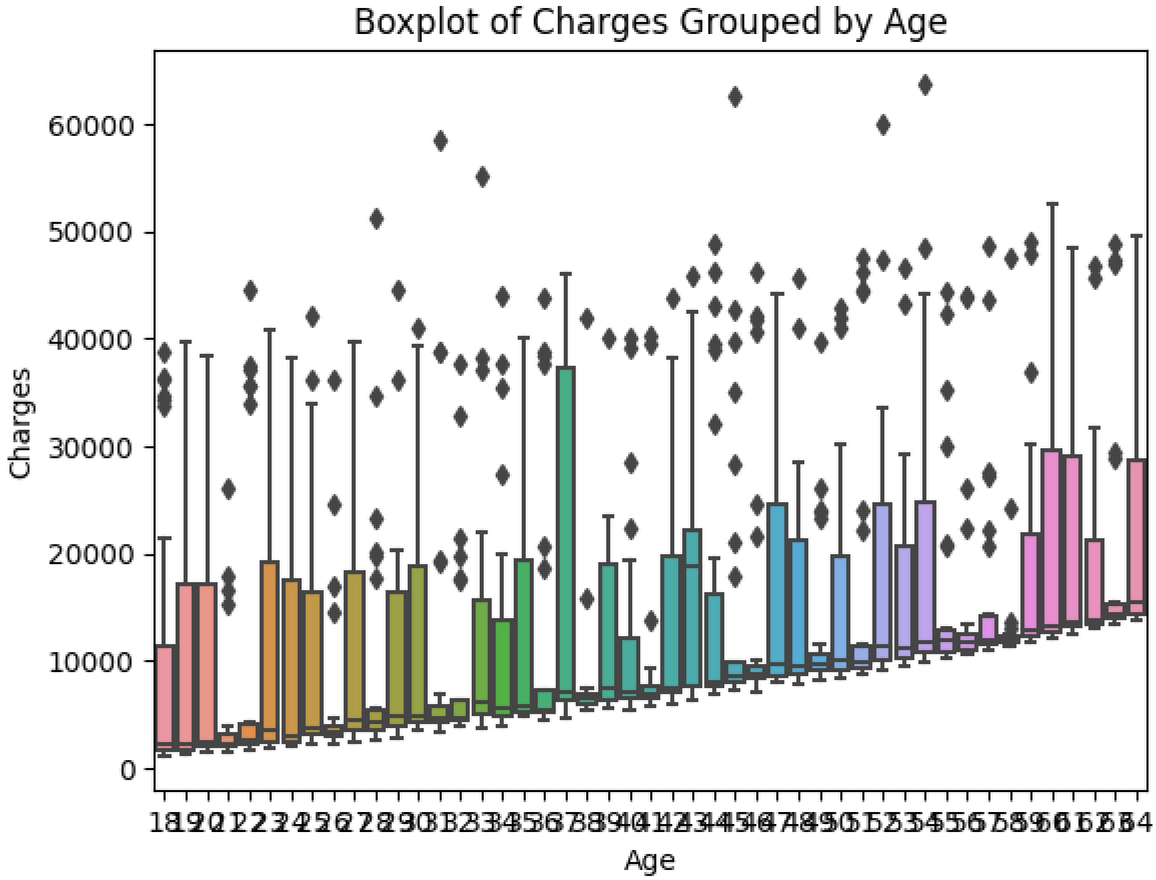
In [27]:
```
# Scatter plot of age vs charges grouped by sex
sns.scatterplot(data=data, x='age', y='charges', hue='sex')
plt.title('Age vs Charges (Grouped by Sex)')
plt.show()
```



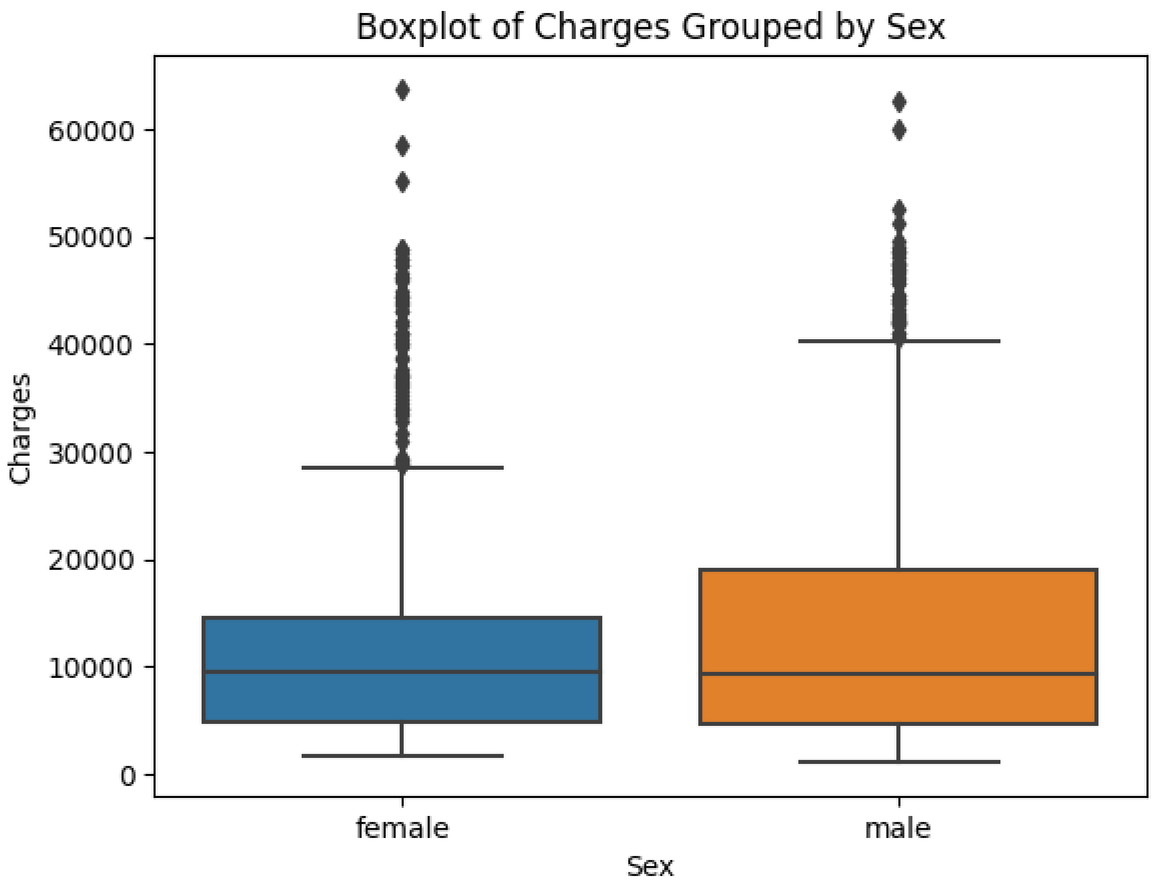
In [28]:
```
# Scatter plot of age vs charges for age > 30
plt.scatter(data[data['age'] > 30]['age'], data[data['age'] > 30]['charges'])
plt.xlabel('Age')
plt.ylabel('Charges')
plt.title('Age vs Charges (Age > 30)')
plt.show()
```



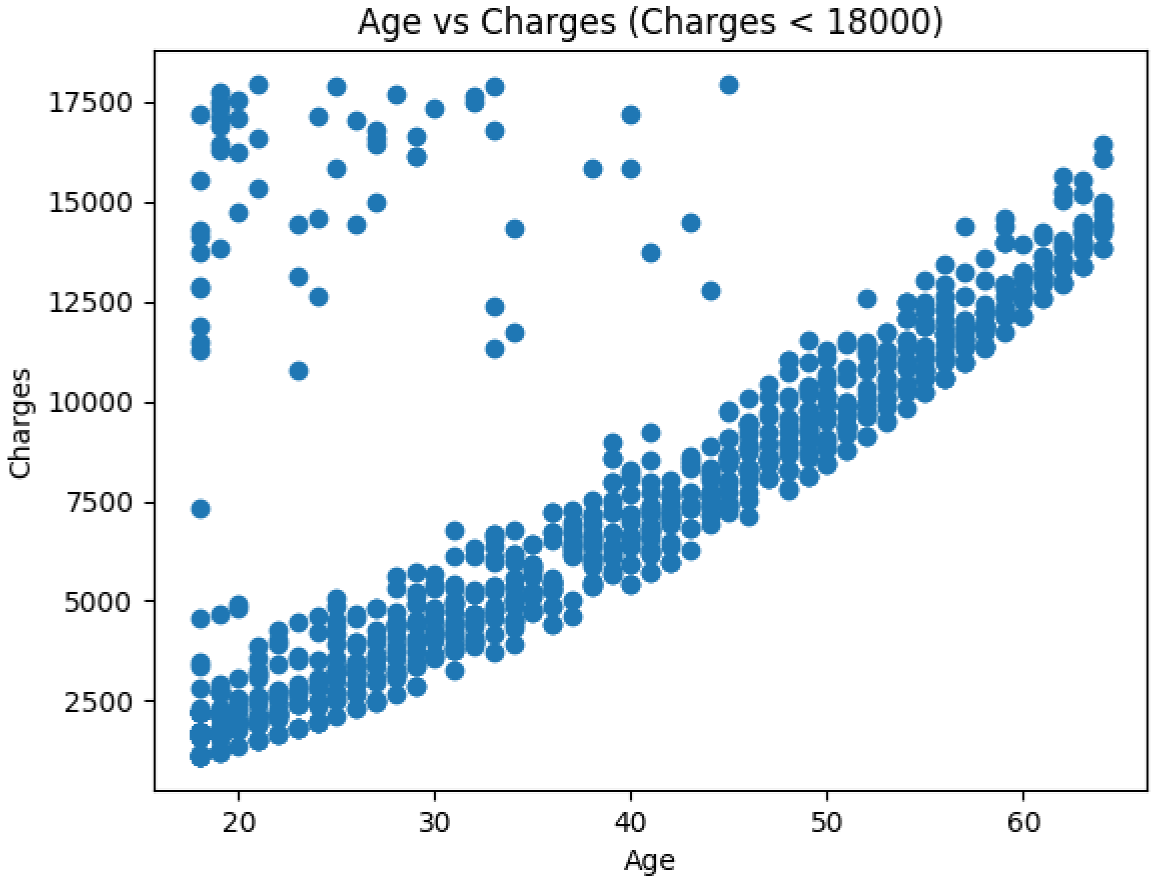
In [29]:
```
# Boxplot of charges grouped by age
sns.boxplot(x=data['age'], y=data['charges'])
plt.xlabel('Age')
plt.ylabel('Charges')
plt.title('Boxplot of Charges Grouped by Age')
plt.show()
```



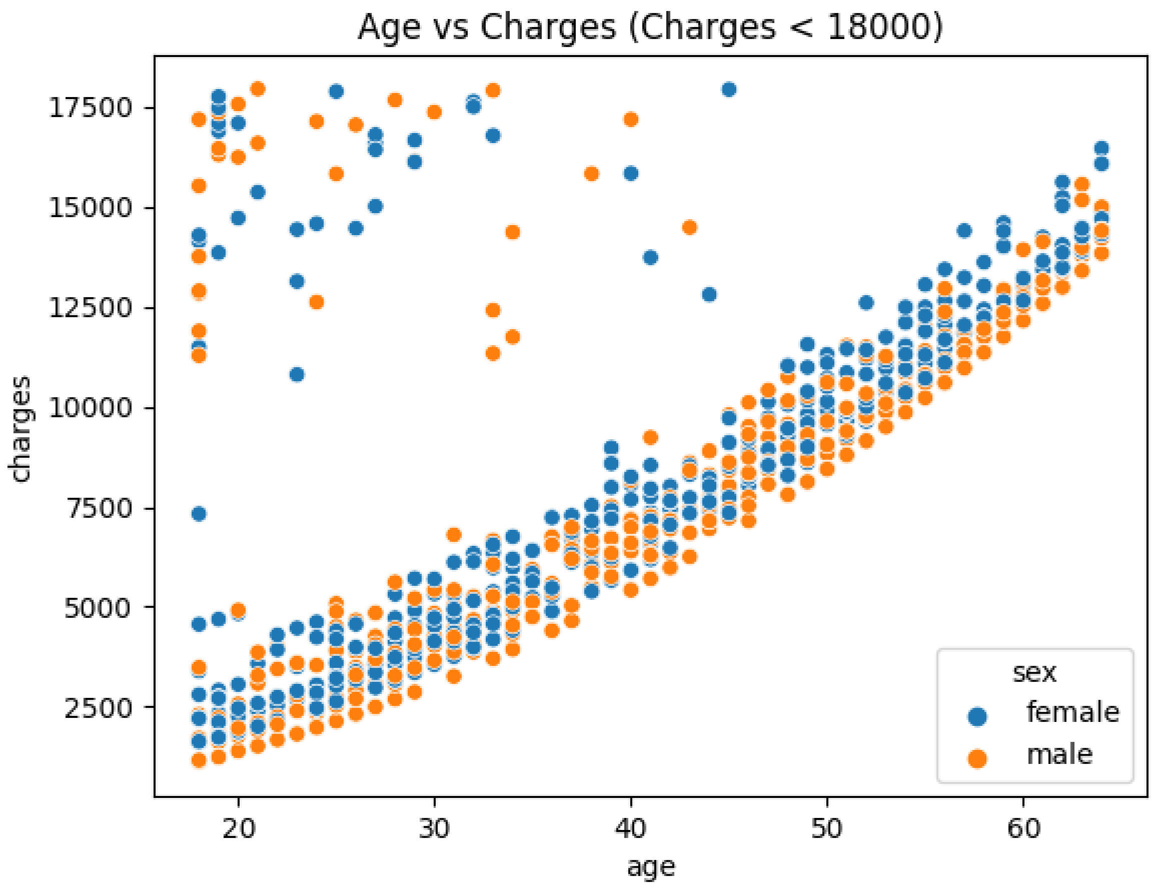
In [30]:
```
# Boxplot of charges grouped by sex
sns.boxplot(x=data['sex'], y=data['charges'])
plt.xlabel('Sex')
plt.ylabel('Charges')
plt.title('Boxplot of Charges Grouped by Sex')
plt.show()
```



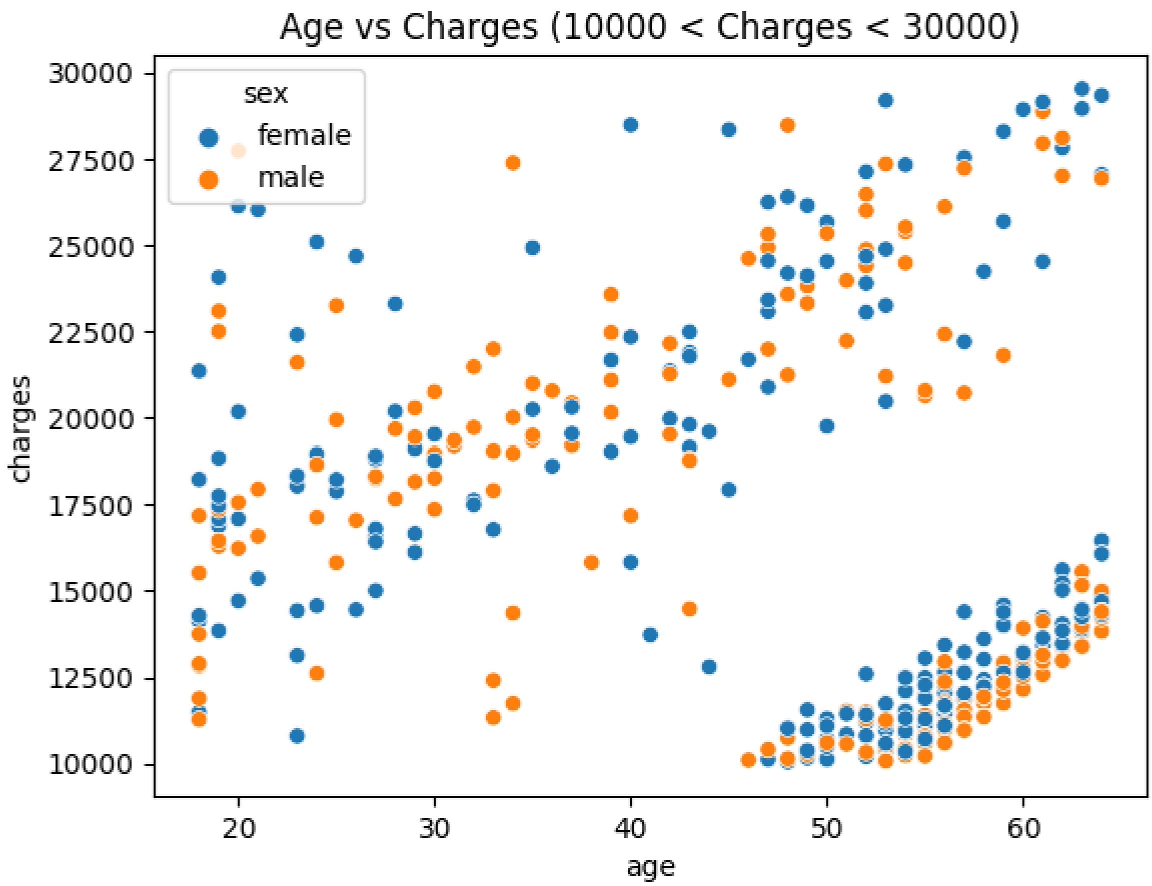
In [31]:
```
# Scatter plot of age vs charges for charges < 18000
plt.scatter(data[data['charges'] < 18000]['age'], data[data['charges'] < 18000]['charg
plt.xlabel('Age')
plt.ylabel('Charges')
plt.title('Age vs Charges (Charges < 18000)')
plt.show()
```



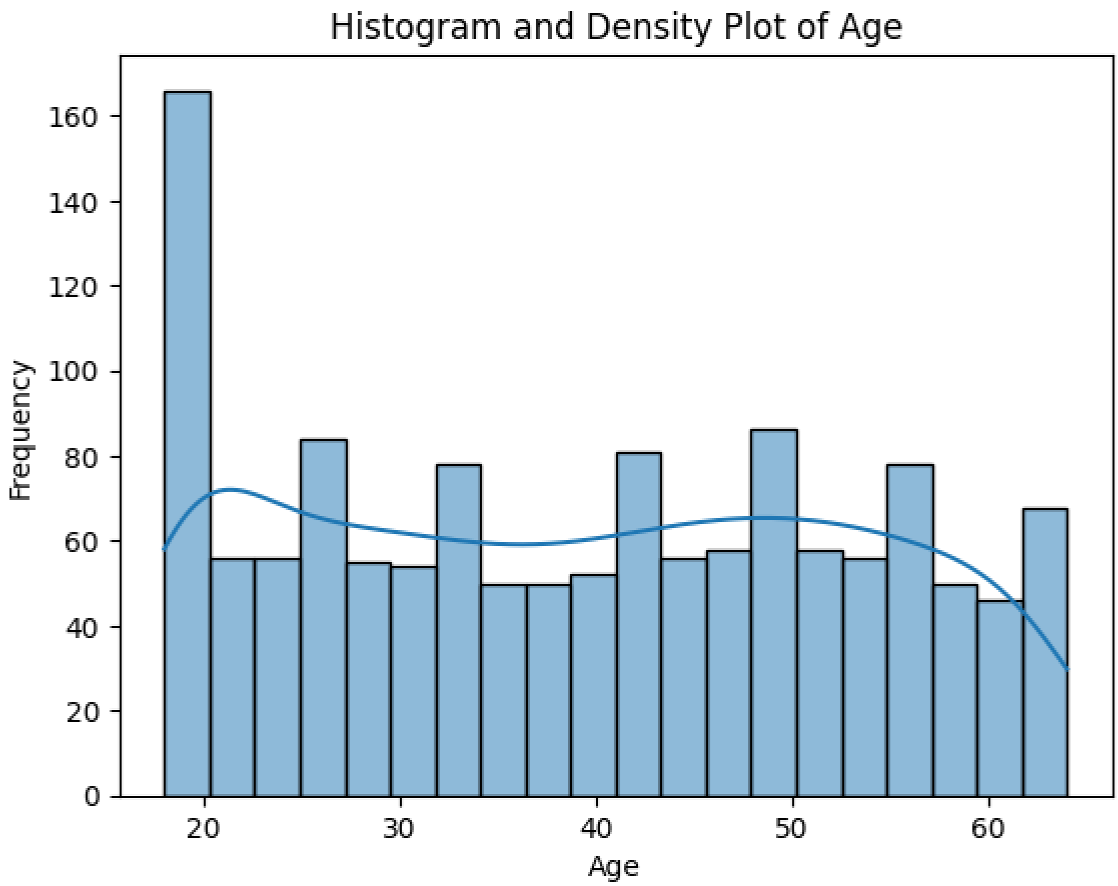
In [32]:
```
# Scatter plot using Seaborn for charges < 18000
sns.scatterplot(data=data[data['charges'] < 18000], x='age', y='charges', hue='sex')
plt.title('Age vs Charges (Charges < 18000)')
plt.show()
```

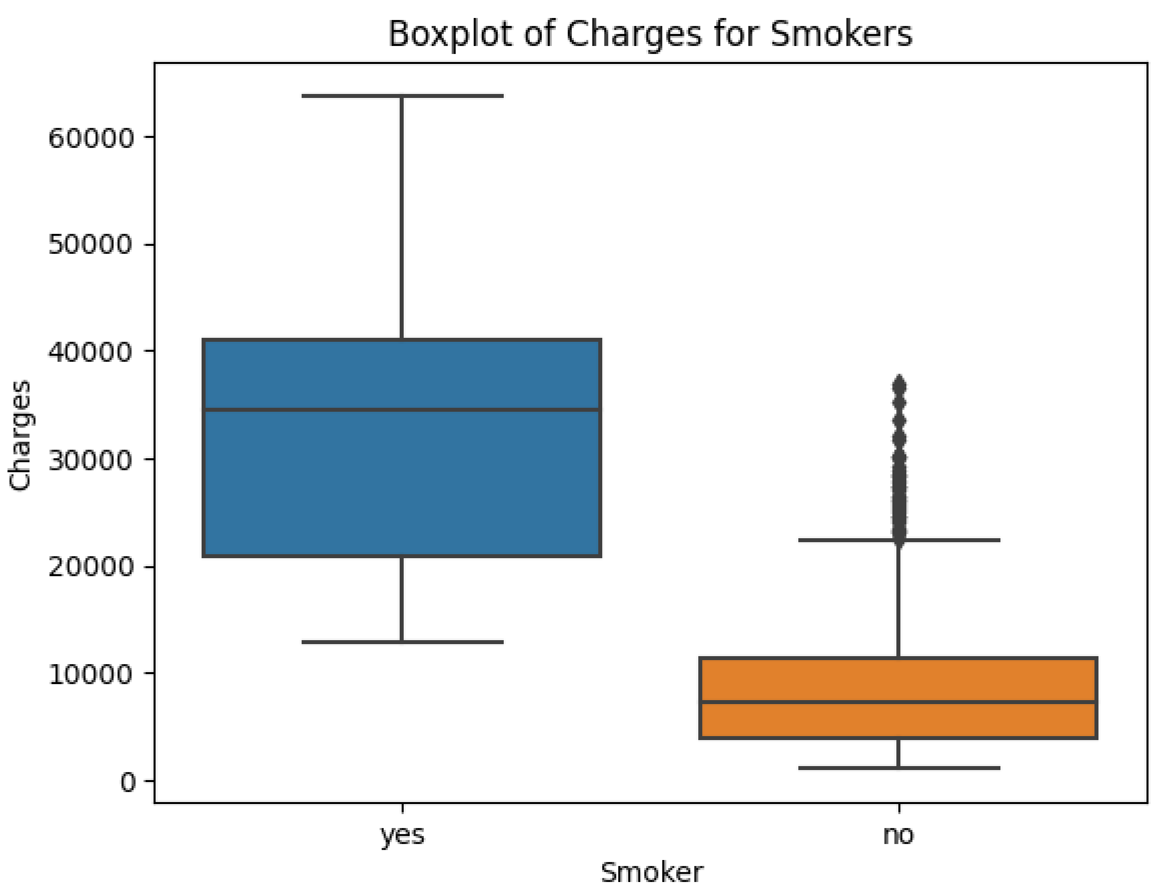


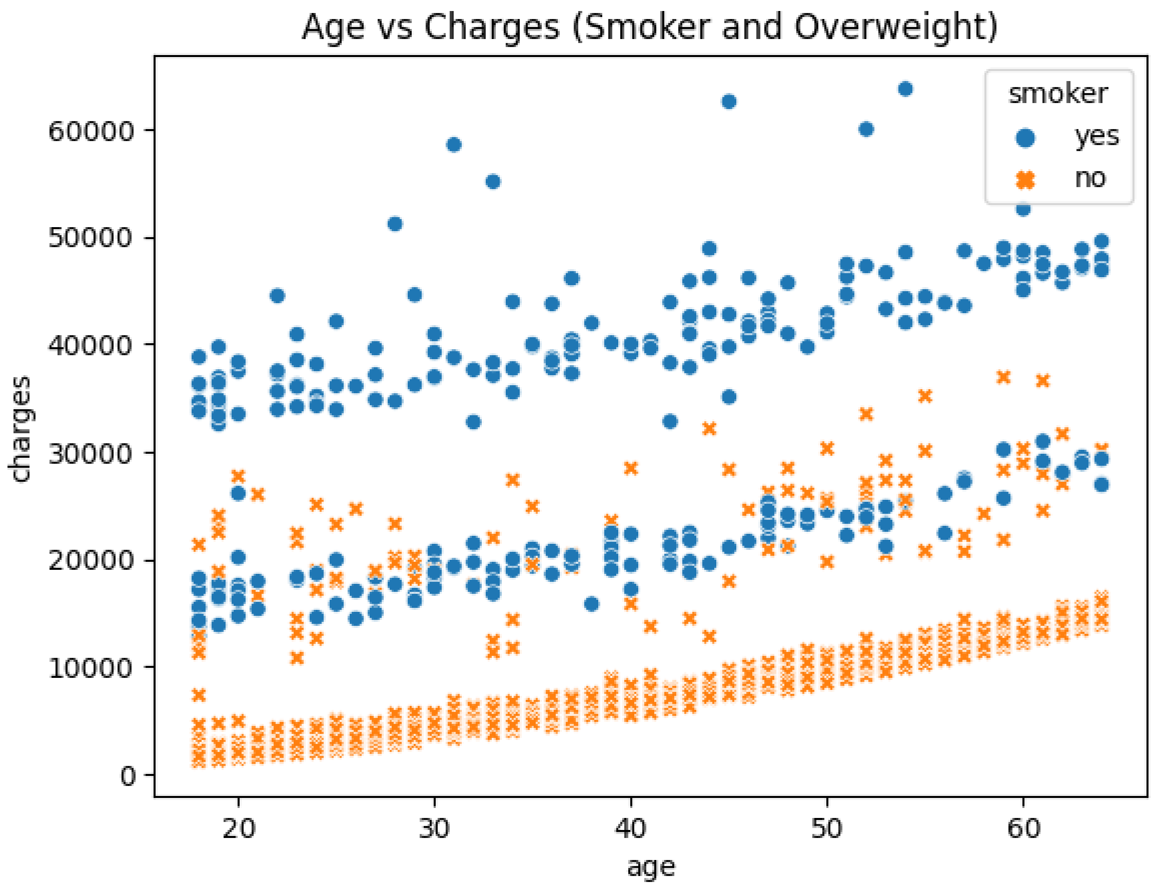
In [33]:
```
# Scatter plot using Seaborn for charges < 30000 and charges > 10000
sns.scatterplot(data=data[(data['charges'] < 30000) & (data['charges'] > 10000)],
                x='age', y='charges', hue='sex')
plt.title('Age vs Charges (10000 < Charges < 30000)')
plt.show()
```



In [35]:
```
# Histogram and Density plot of age
sns.histplot(data['age'], bins=20, kde=True)
plt.xlabel('Age')
plt.ylabel('Frequency')
plt.title('Histogram and Density Plot of Age')
plt.show()

# Boxplot of charges for categorical variable 'smoker'
sns.boxplot(data=data, x='smoker', y='charges')
plt.xlabel('Smoker')
plt.ylabel('Charges')
plt.title('Boxplot of Charges for Smokers')
plt.show()

# Scatter plot of age vs charges with color and shape differentiation
sns.scatterplot(data=data, x='age', y='charges', hue='smoker', style='smoker')
plt.title('Age vs Charges (Smoker and Overweight)')
plt.show()
```

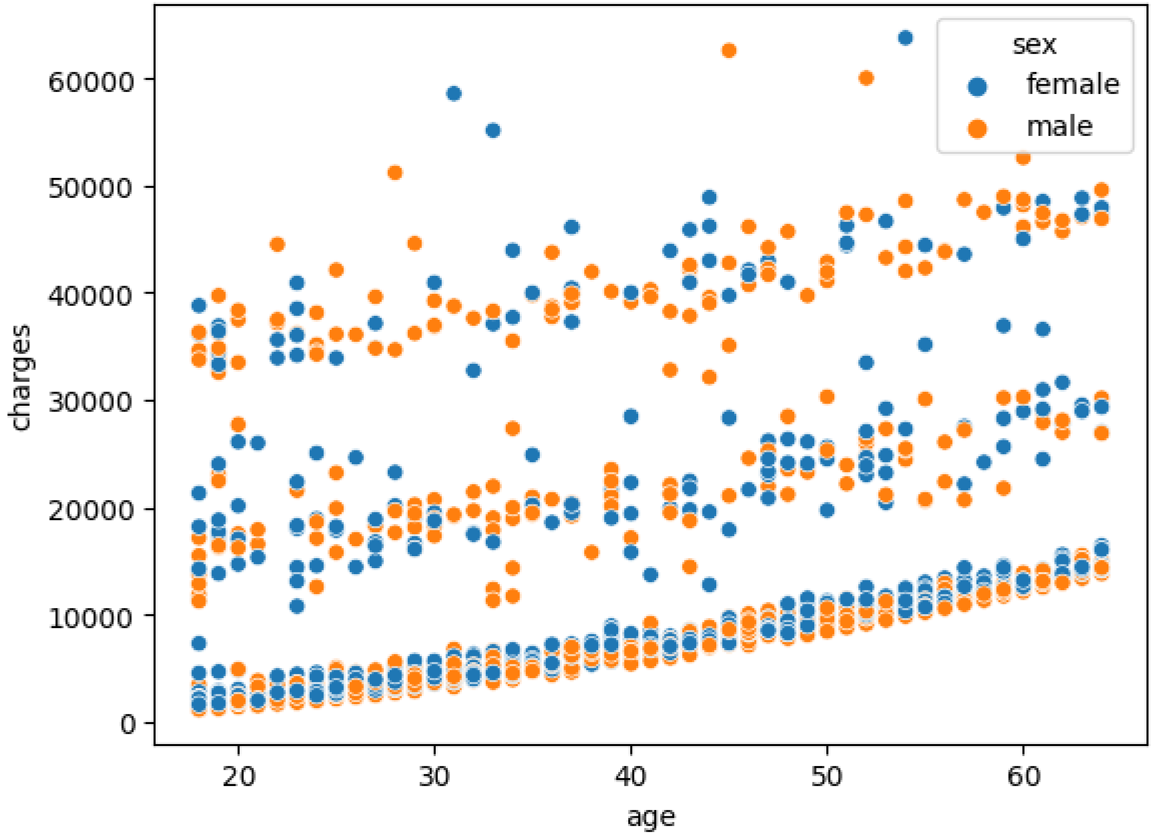

Histogram and Density Plot of Age
Frequency
0
20
40
60
80
100
120
140
160
20
30
40
50
60
Age


Boxplot of Charges for Smokers
Charges
0
10000
20000
30000
40000
50000
60000
yes
no
Smoker

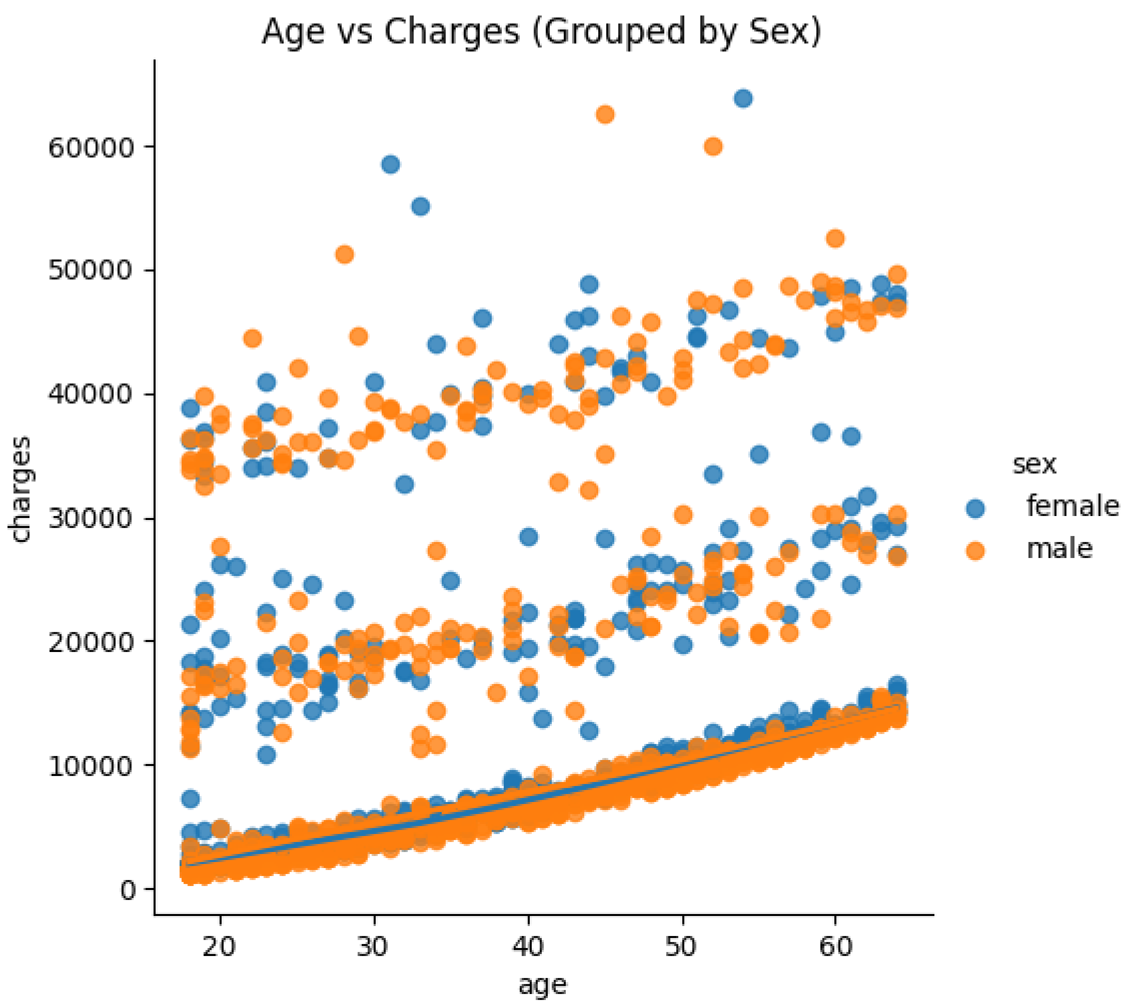


In [37]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
import statsmodels.api as sm
import scipy.stats as stats


# Scatter plot of age vs charges with color differentiation for sex and smoothing line
sns.scatterplot(data=data, x='age', y='charges', hue='sex')
sns.lmplot(data=data, x='age', y='charges', hue='sex', lowess=True)
plt.title('Age vs Charges (Grouped by Sex)')
plt.show()
```

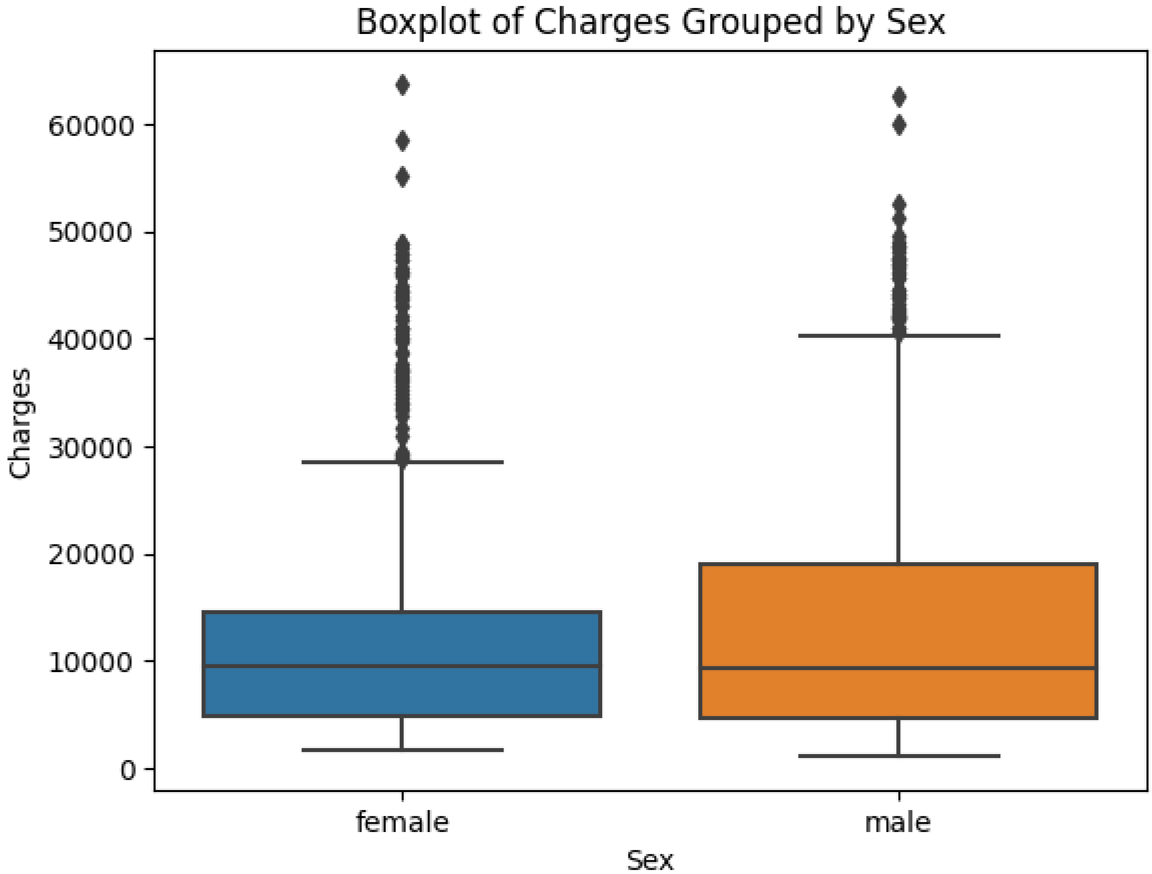
```
C:\Users\user\Anaconda3\lib\site-packages\seaborn\axisgrid.py:118: UserWarning: The f
igure layout has changed to tight
  self._figure.tight_layout(*args, **kwargs)
```

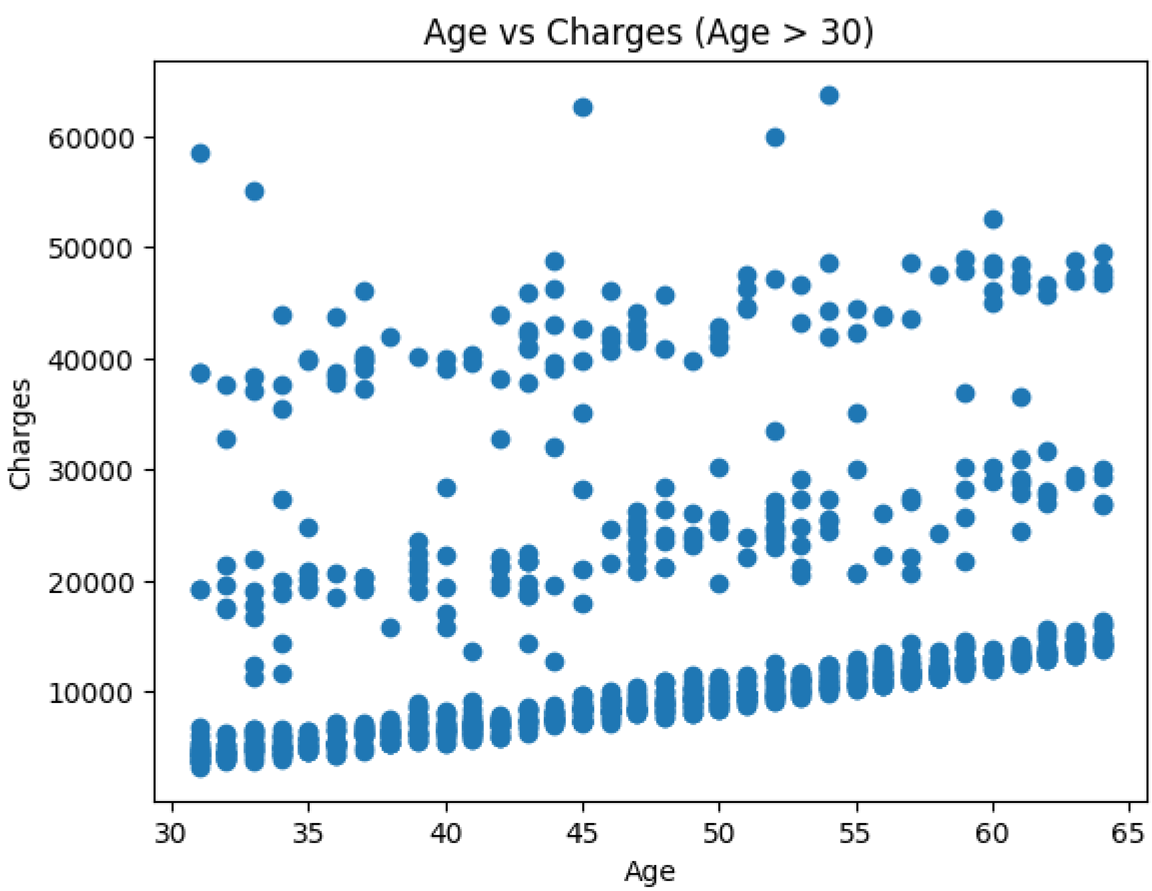


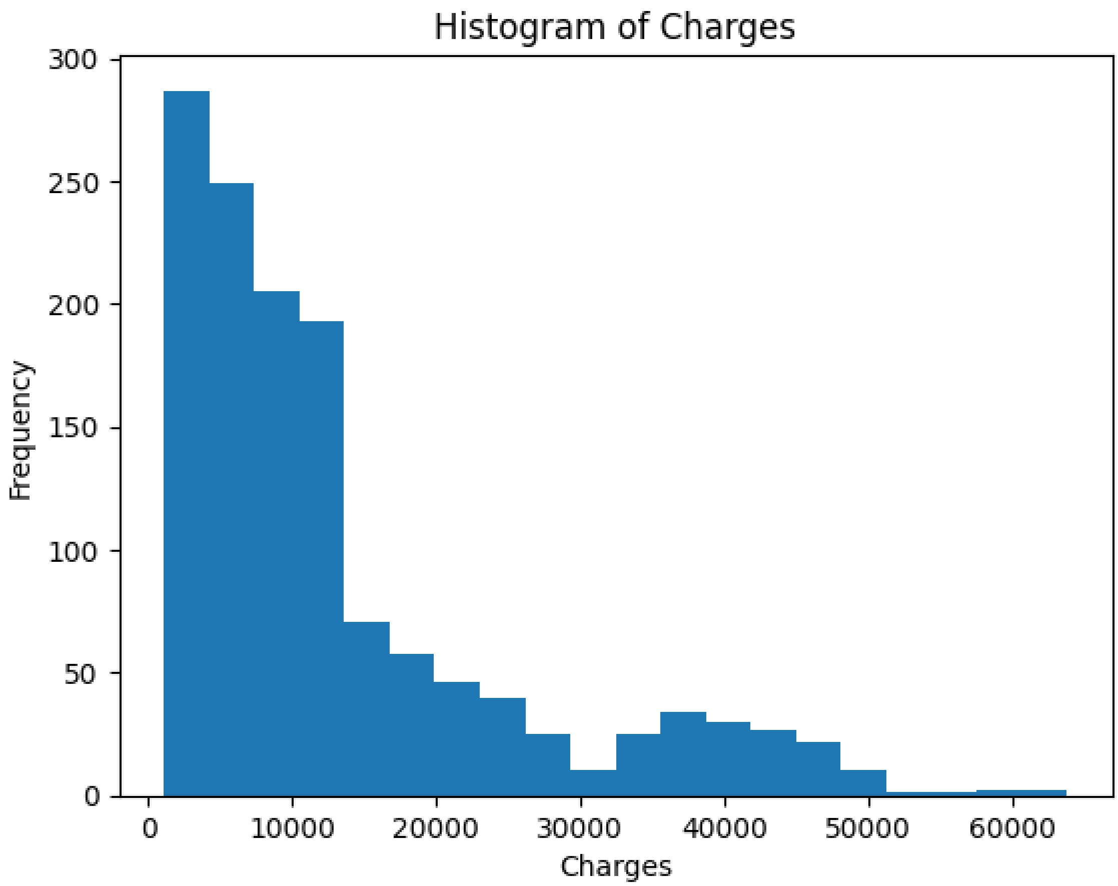
sex
female
male
charges
age
60000
50000
40000
30000
20000
10000
0
20
30
40
50
60

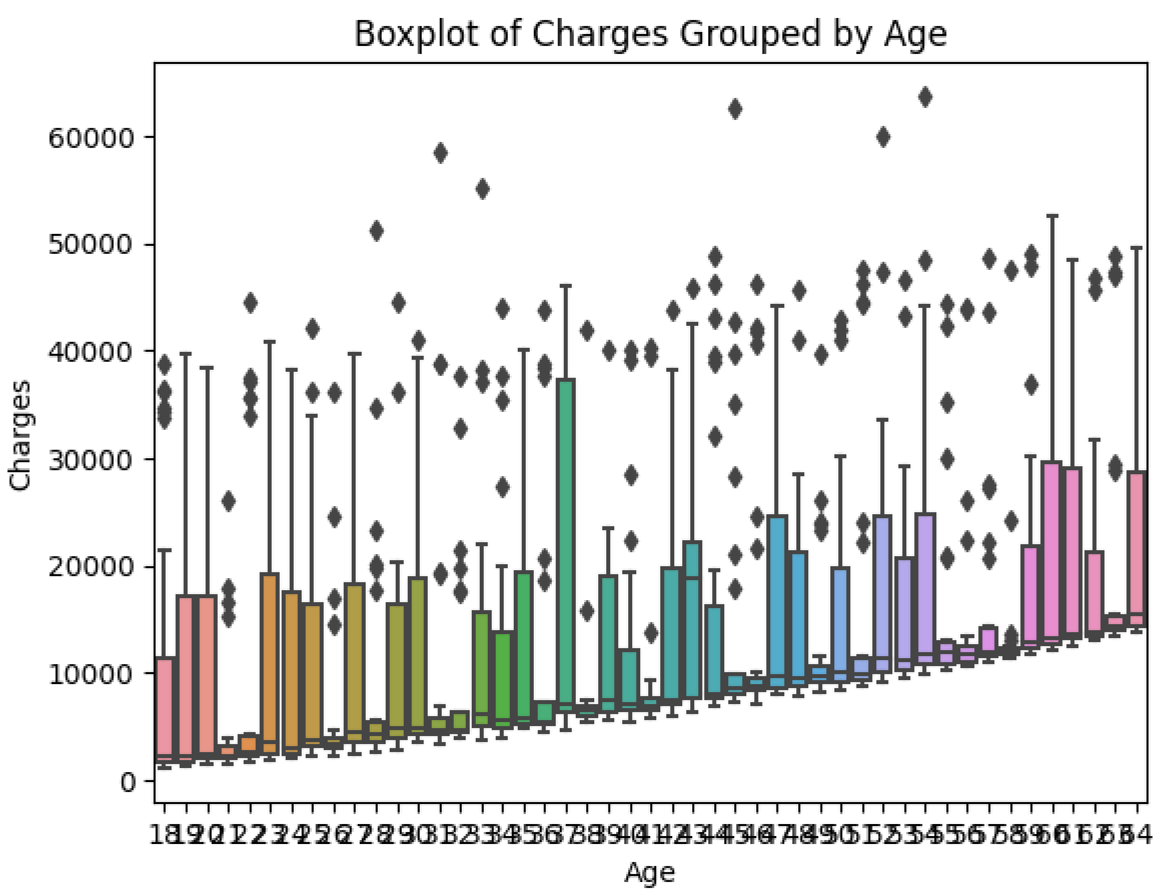


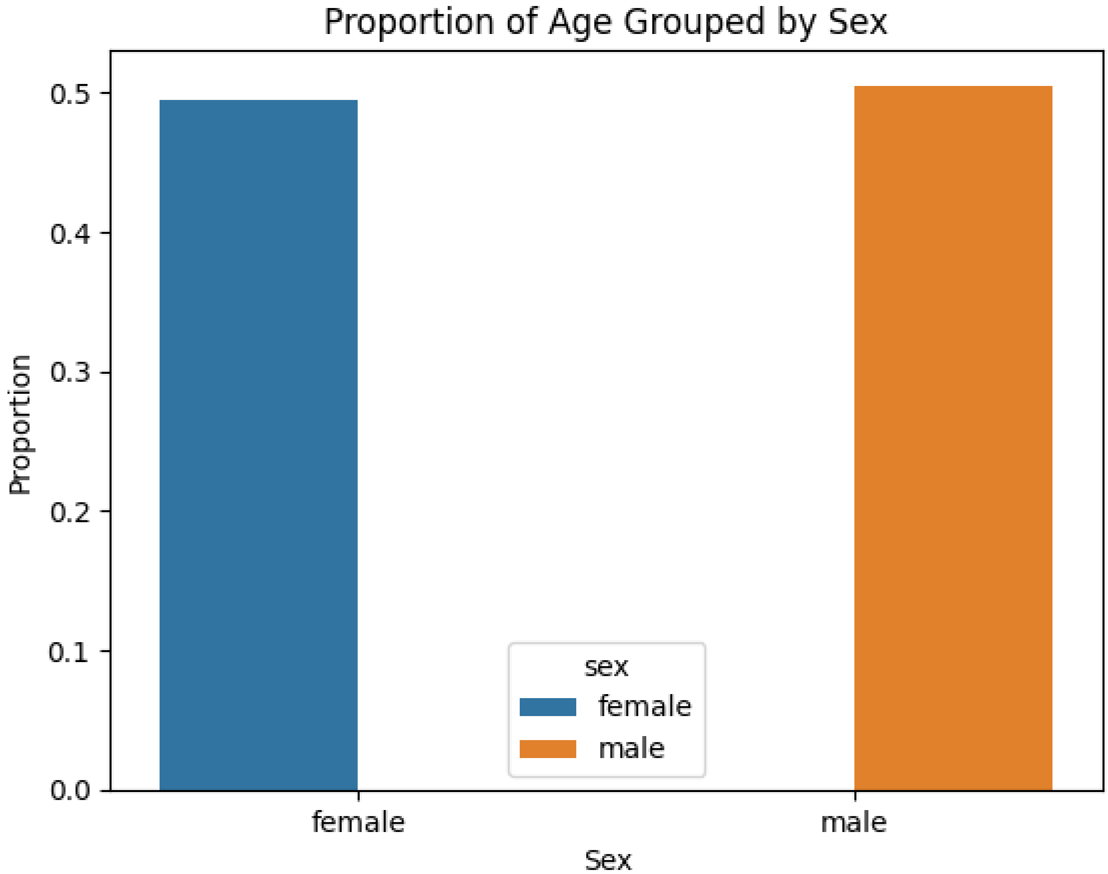
Age vs Charges (Grouped by Sex)
sex
female
male
charges
age
60000
50000
40000
30000
20000
10000
0
20
30
40
50
60


Boxplot of Charges Grouped by Sex
Charges
60000
50000
40000
30000
20000
10000
0
female
male
Sex

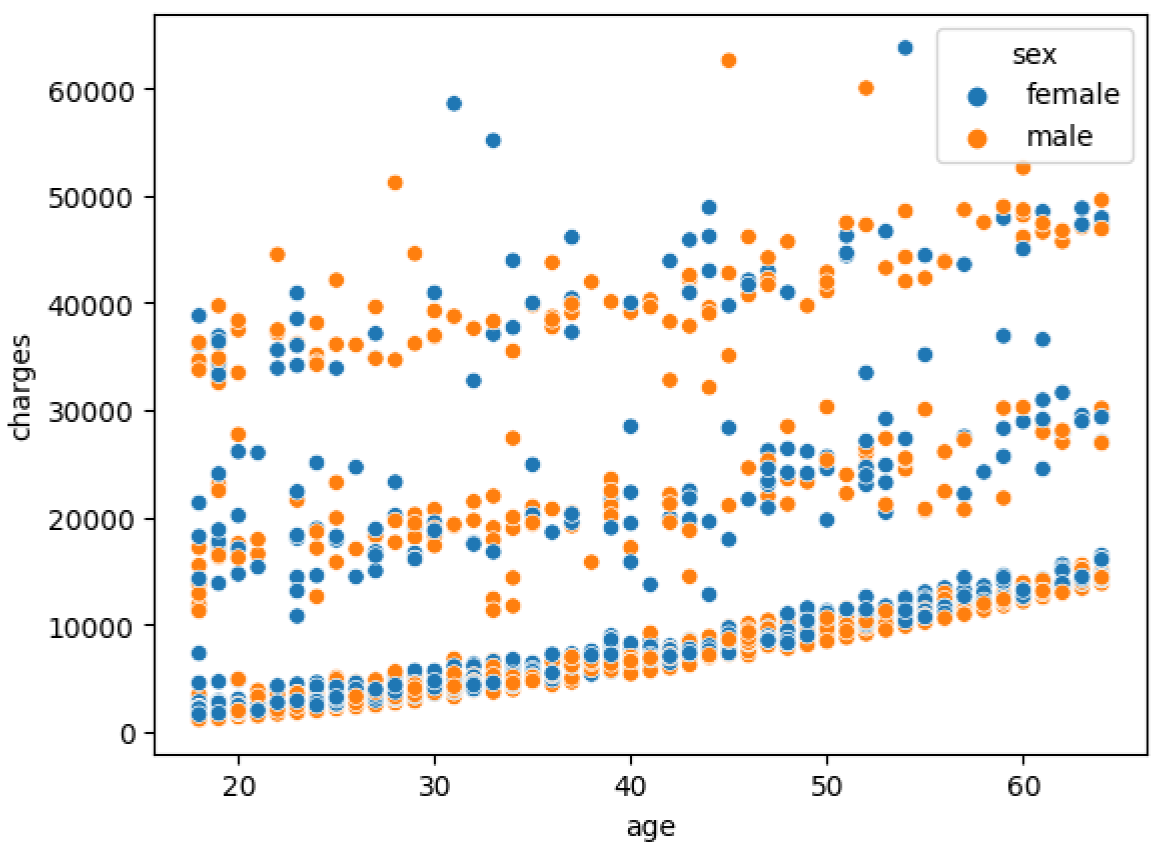

Age vs Charges (Age > 30)
Charges
60000
50000
40000
30000
20000
10000
30
35
40
45
50
55
60
65
Age

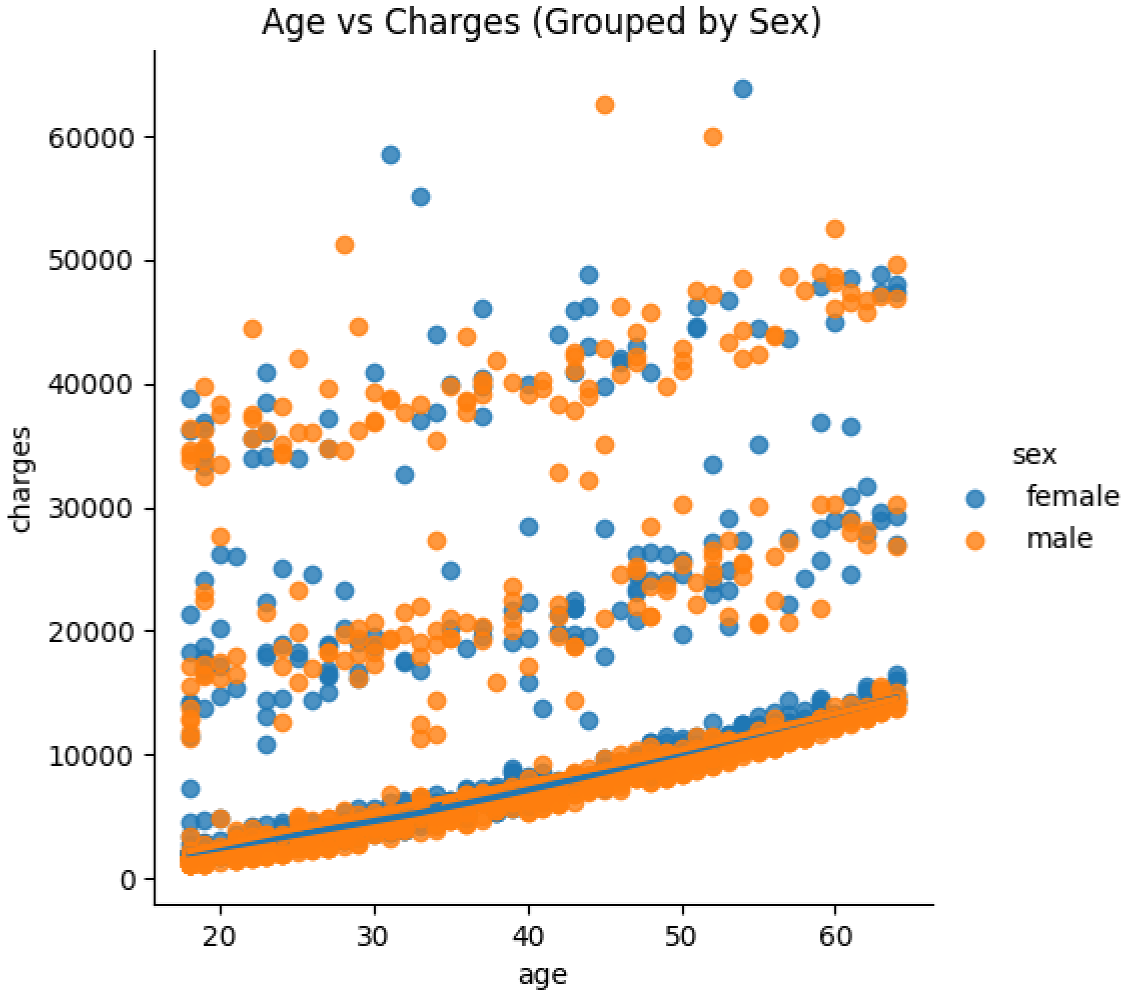

Histogram of Charges
Frequency
Charges
0
50
100
150
200
250
300
0
10000
20000
30000
40000
50000
60000

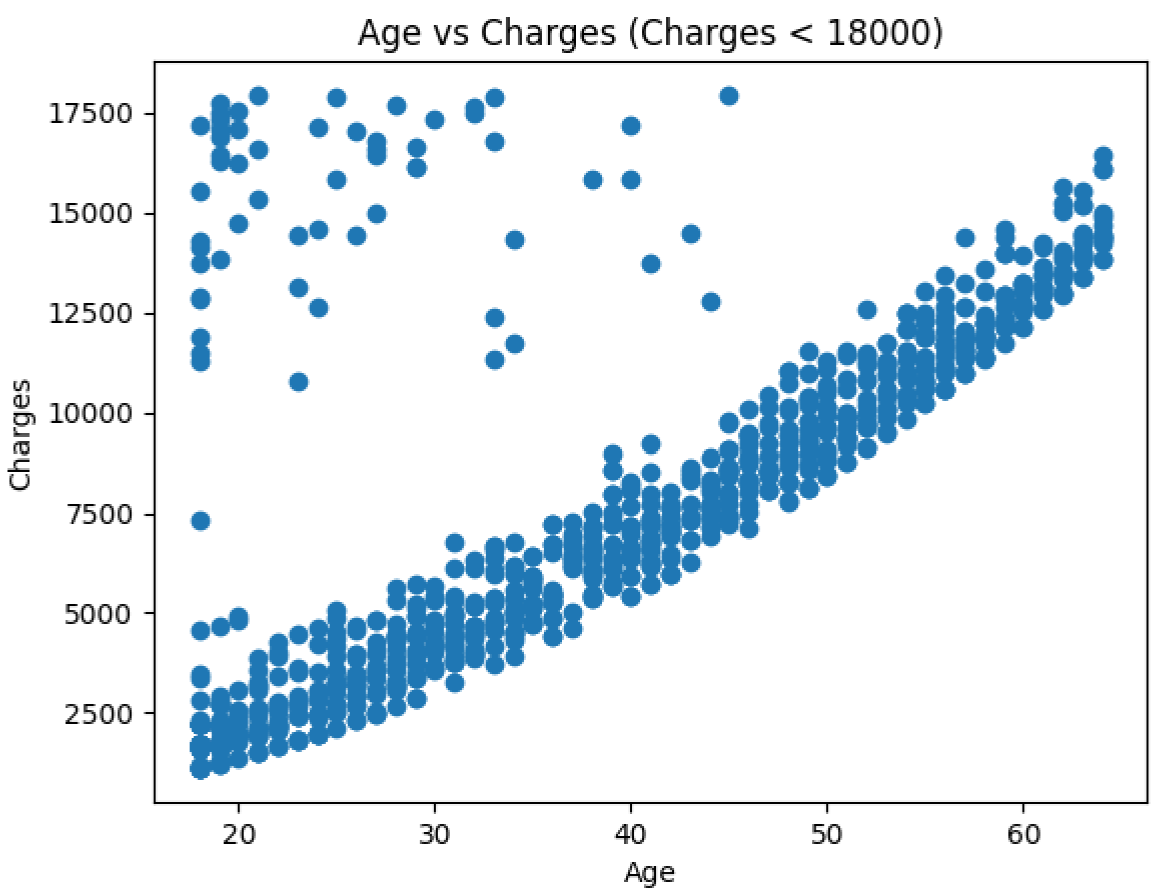

Boxplot of Charges Grouped by Age
Charges
Age
0
10000
20000
30000
40000
50000
60000

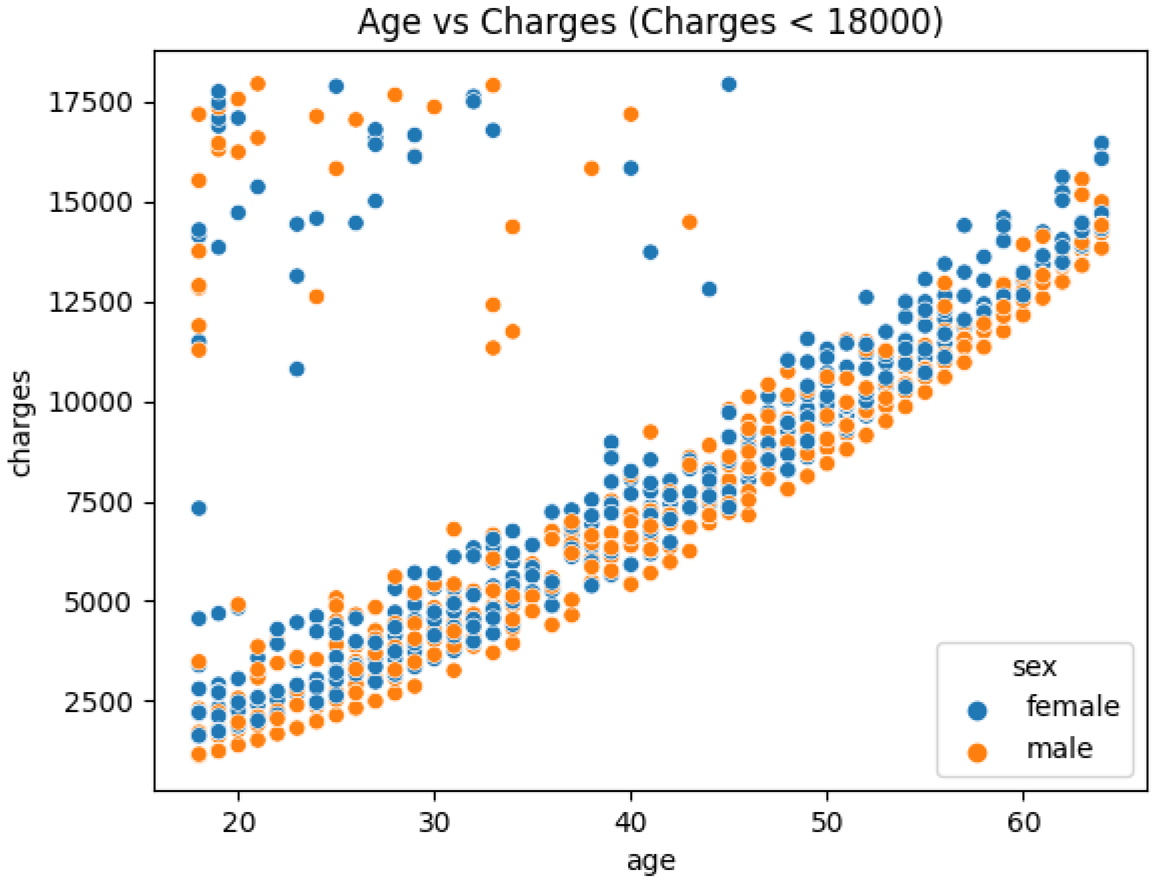


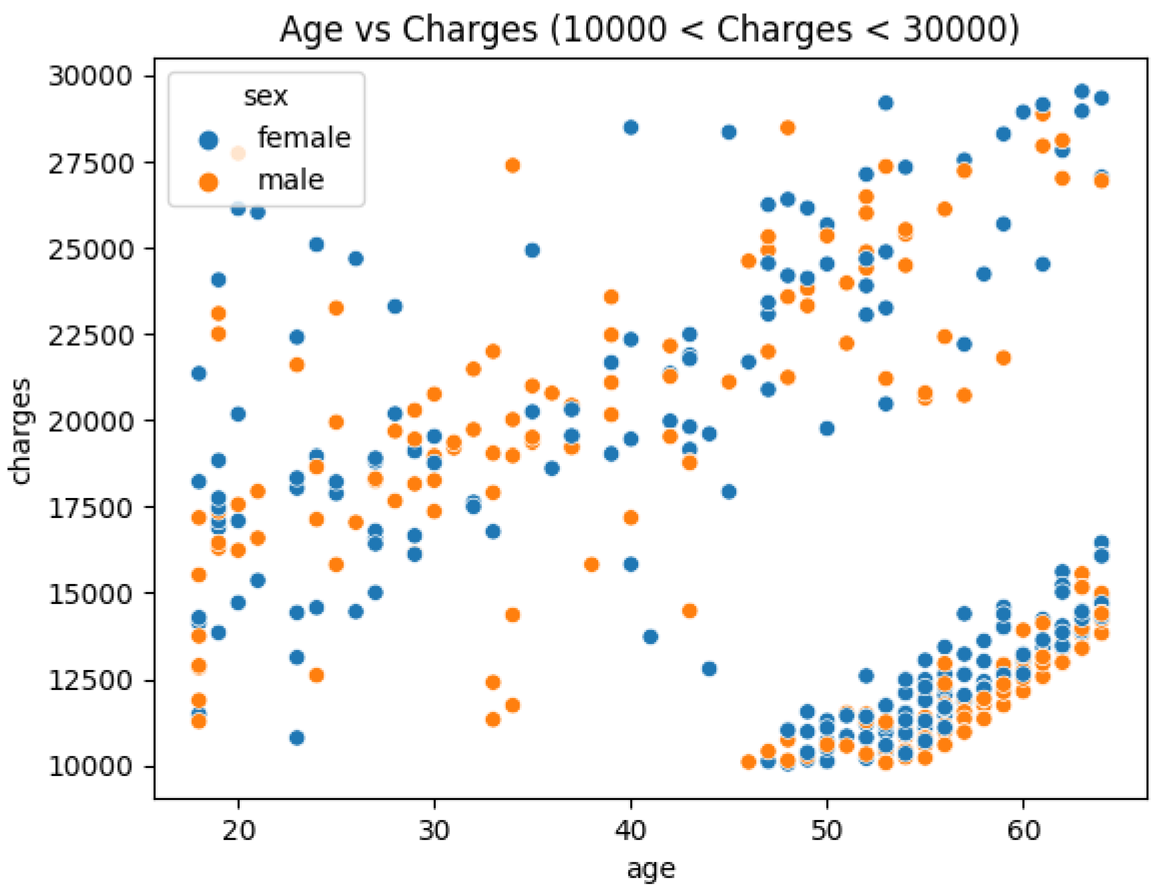

```
C:\Users\user\Anaconda3\lib\site-packages\seaborn\axisgrid.py:118: UserWarning: The f
igure layout has changed to tight
  self._figure.tight_layout(*args, **kwargs)
```

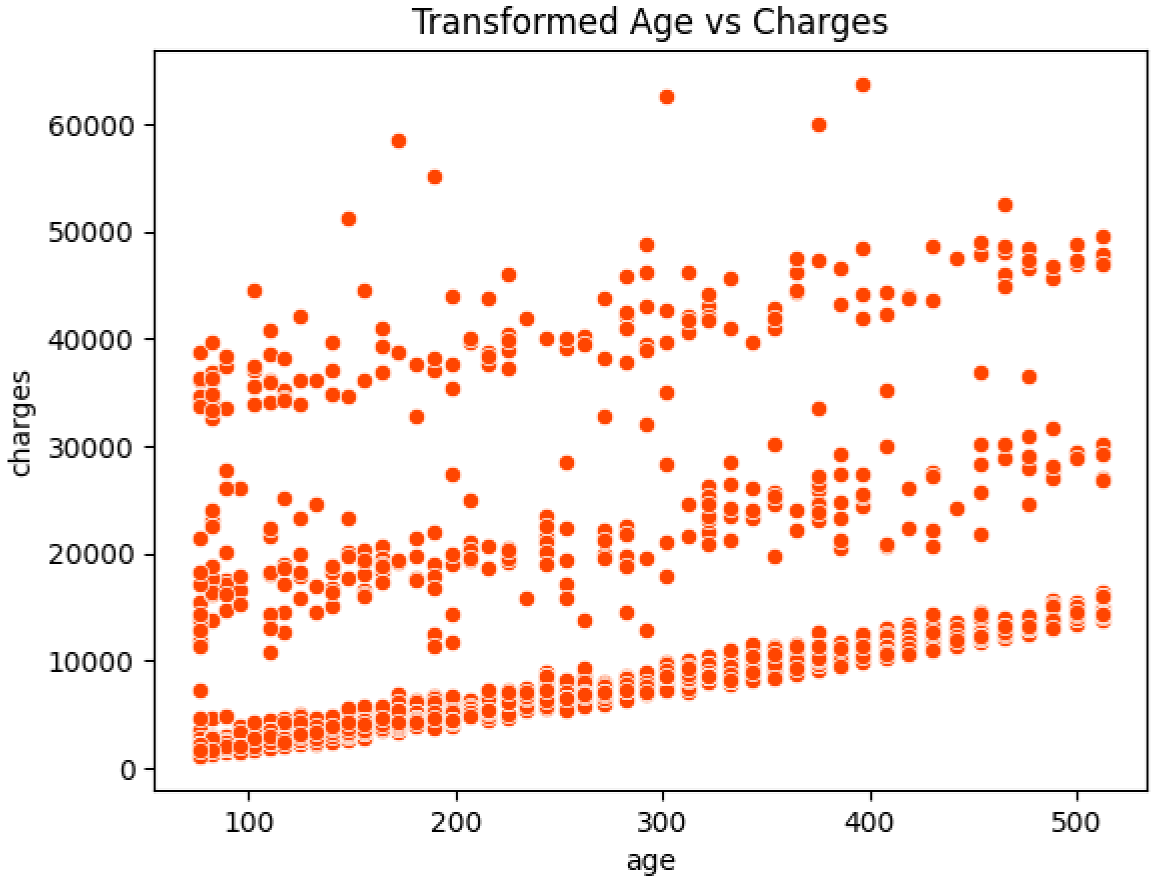


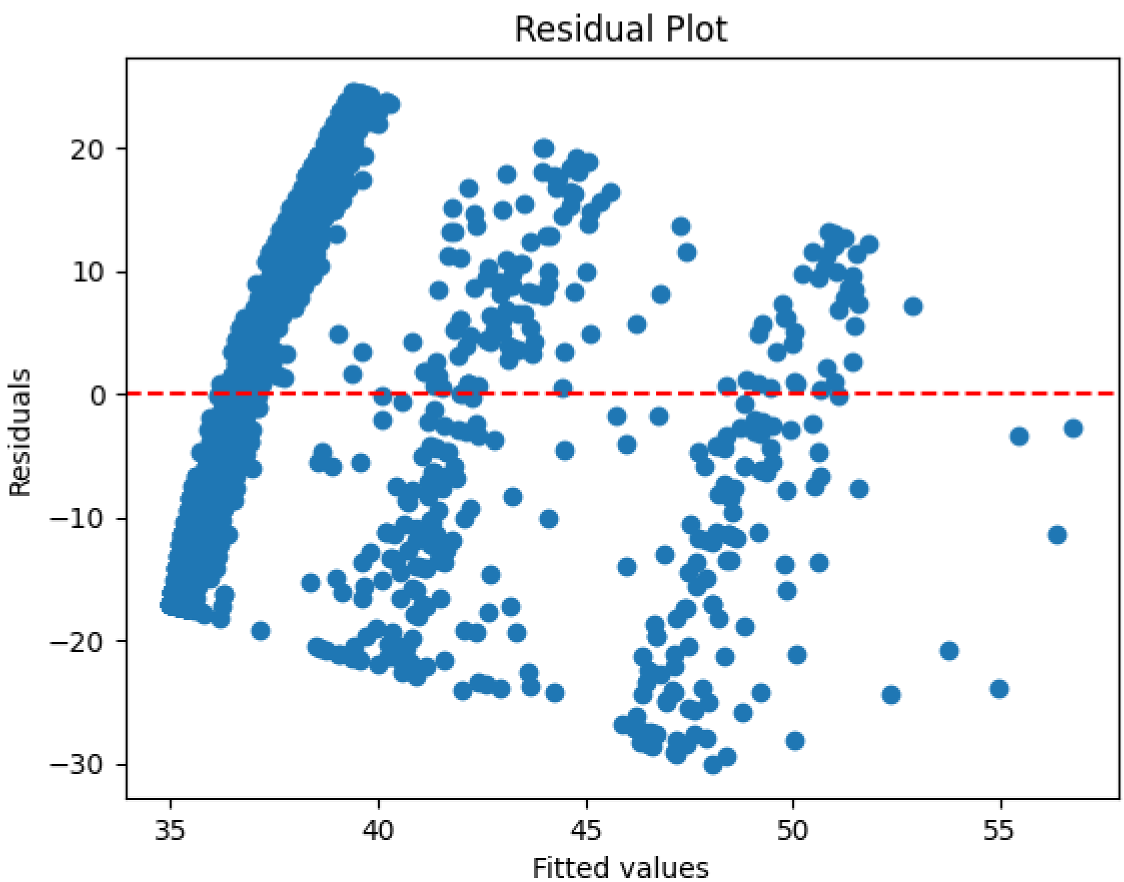

Age vs Charges (Grouped by Sex)
charges
age
sex
female
male
60000
50000
40000
30000
20000
10000
0
20
30
40
50
60

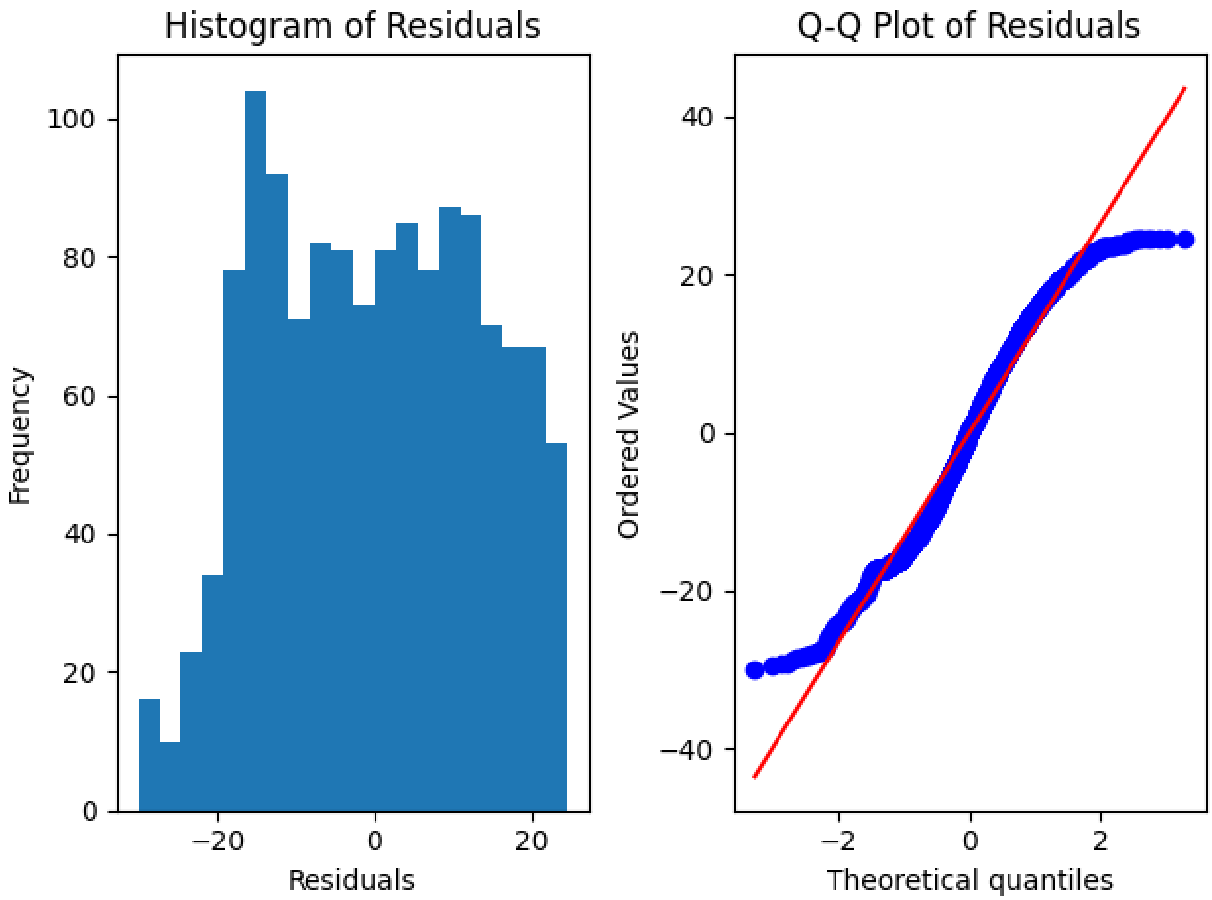

Age vs Charges (Charges < 18000)
Charges
Age
17500
15000
12500
10000
7500
5000
2500
20
30
40
50
60


Age vs Charges (Charges < 18000)
charges
17500
15000
12500
10000
7500
5000
2500
age
20
30
40
50
60
sex
female
male

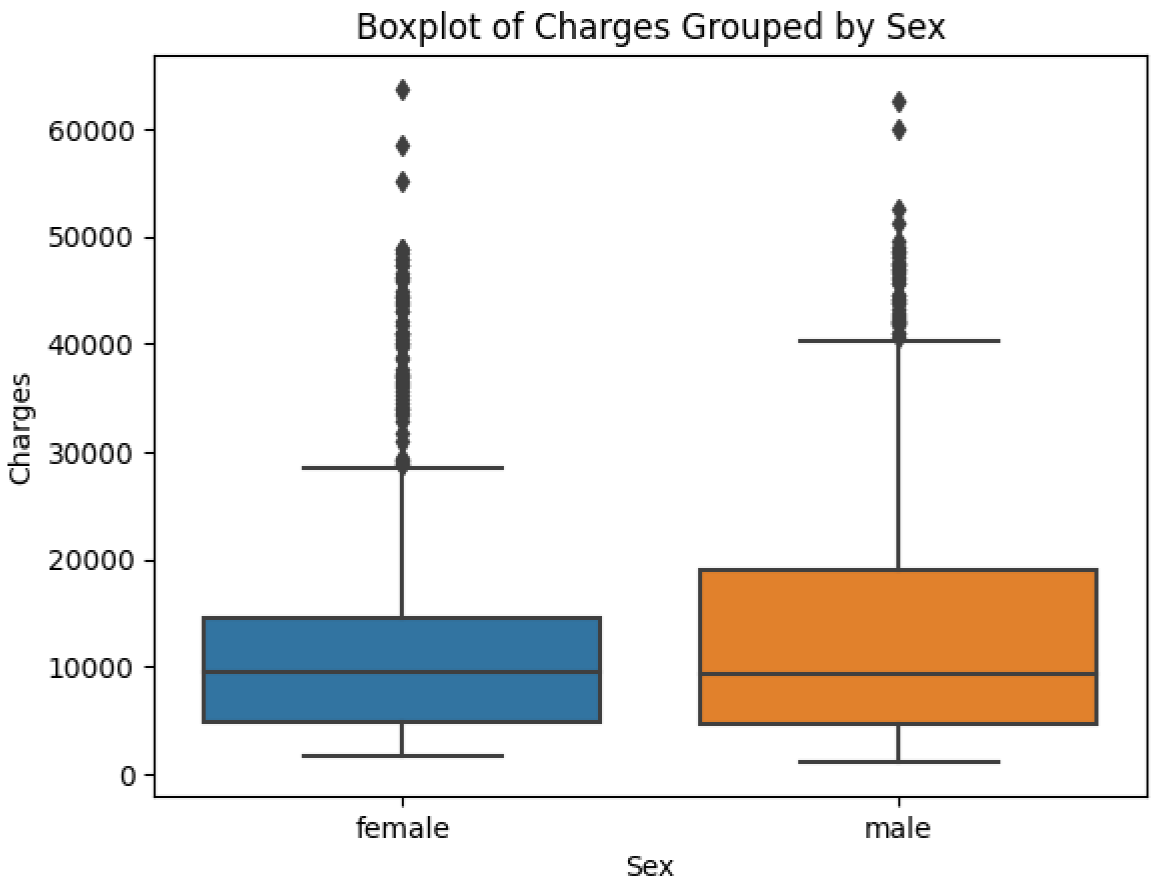

Age vs Charges (10000 < Charges < 30000)
charges
30000
27500
25000
22500
20000
17500
15000
12500
10000
age
20
30
40
50
60
sex
female
male


Transformed Age vs Charges
charges
age
60000
50000
40000
30000
20000
10000
0
100
200
300
400
500

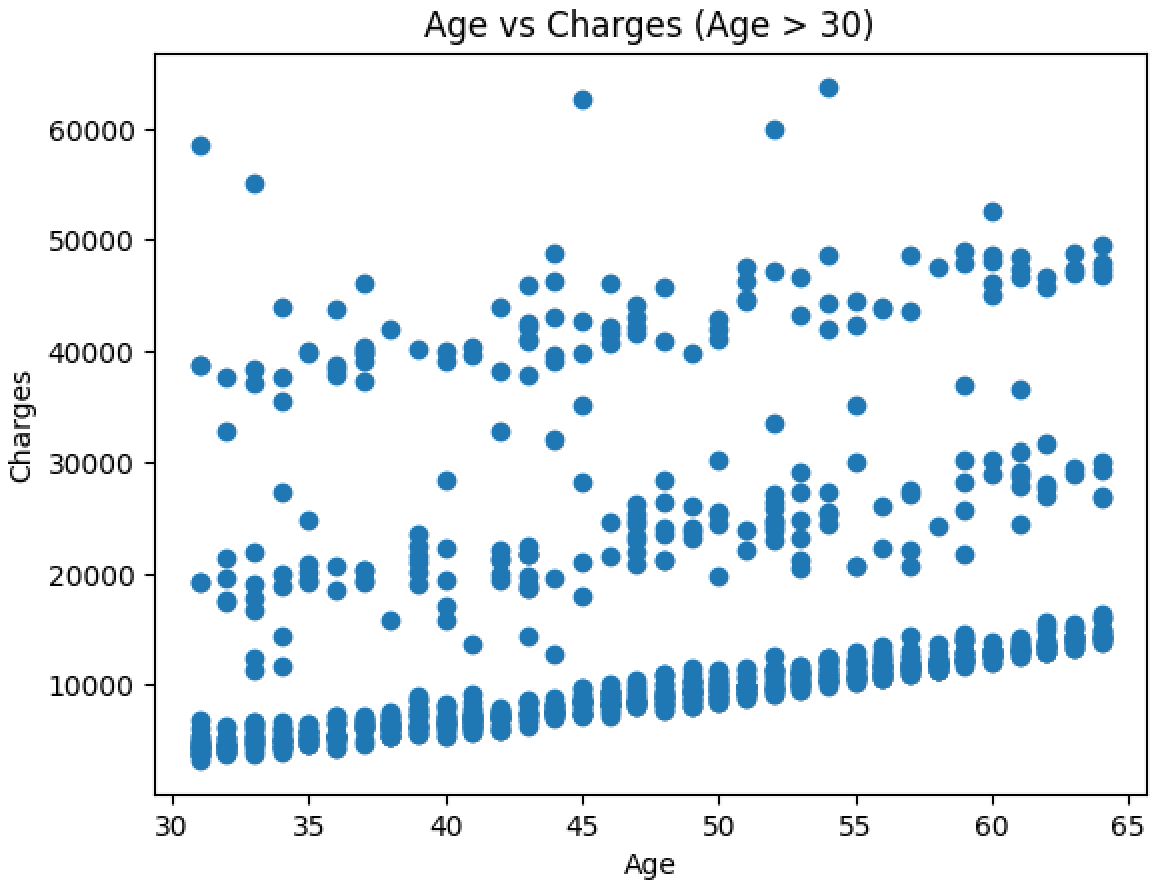

Residual Plot
Residuals
Fitted values
20
10
0
−10
−20
−30
35
40
45
50
55



In [38]:
```
# Boxplot of charges grouped by sex
sns.boxplot(data=data, x='sex', y='charges')
plt.xlabel('Sex')
plt.ylabel('Charges')
plt.title('Boxplot of Charges Grouped by Sex')
plt.show()
```

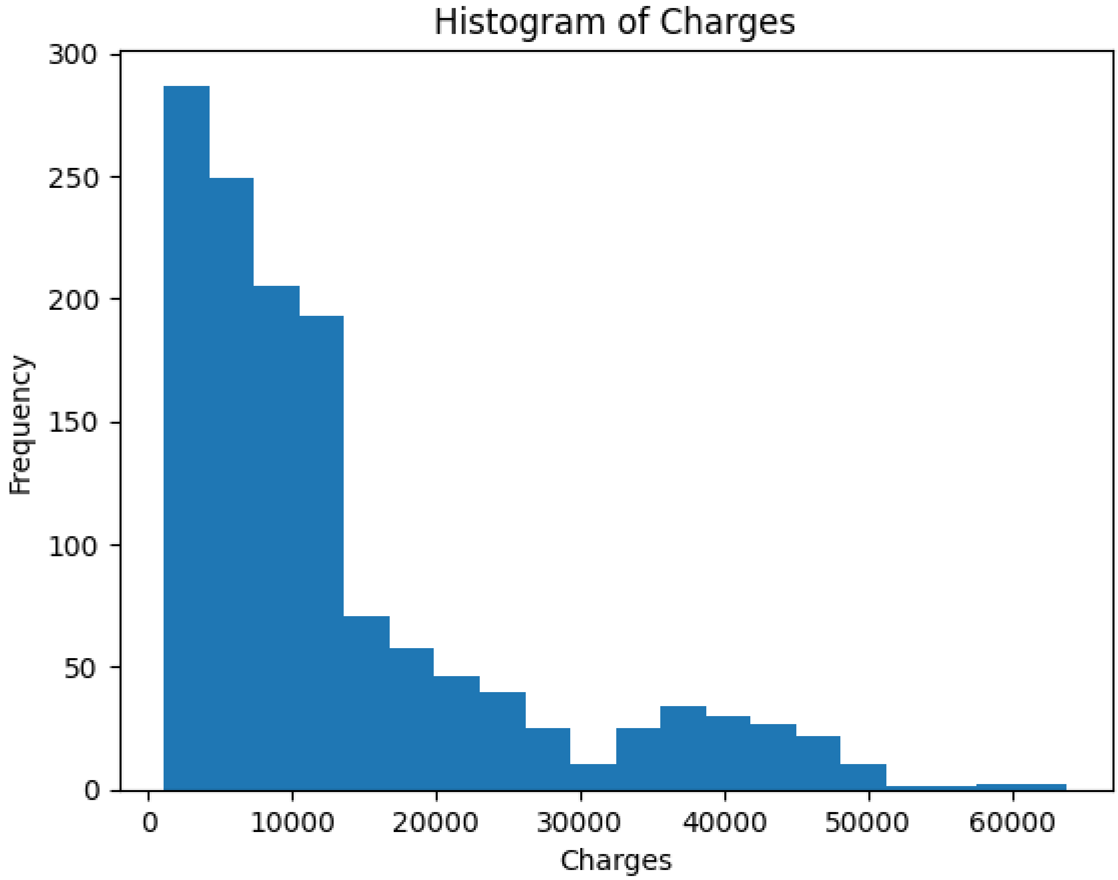


In [39]:
```
# Scatter plot of age vs charges for age > 30
plt.scatter(data[data['age'] > 30]['age'], data[data['age'] > 30]['charges'])
plt.xlabel('Age')
plt.ylabel('Charges')
plt.title('Age vs Charges (Age > 30)')
plt.show()
```

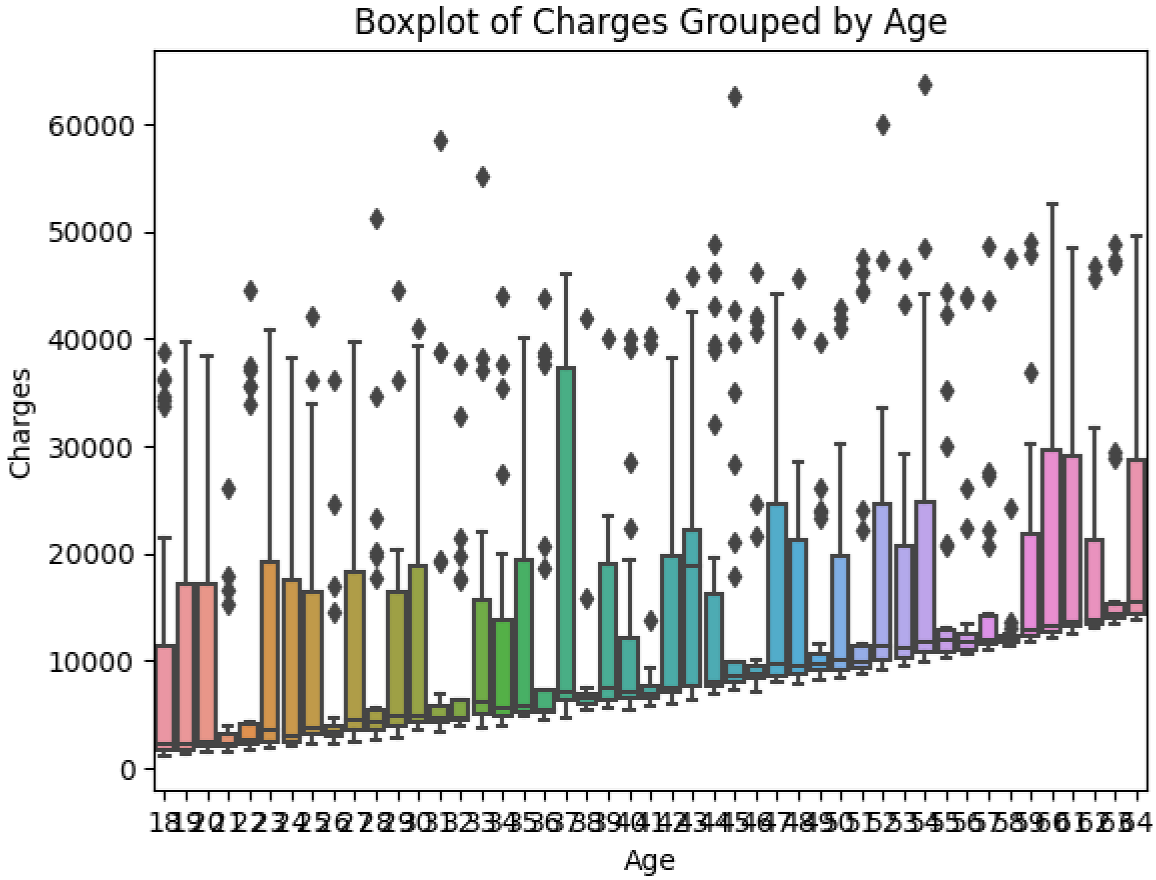


In [40]:

```
# Histogram of charges
plt.hist(data['charges'], bins=20)
plt.xlabel('Charges')
plt.ylabel('Frequency')
plt.title('Histogram of Charges')
plt.show()
```

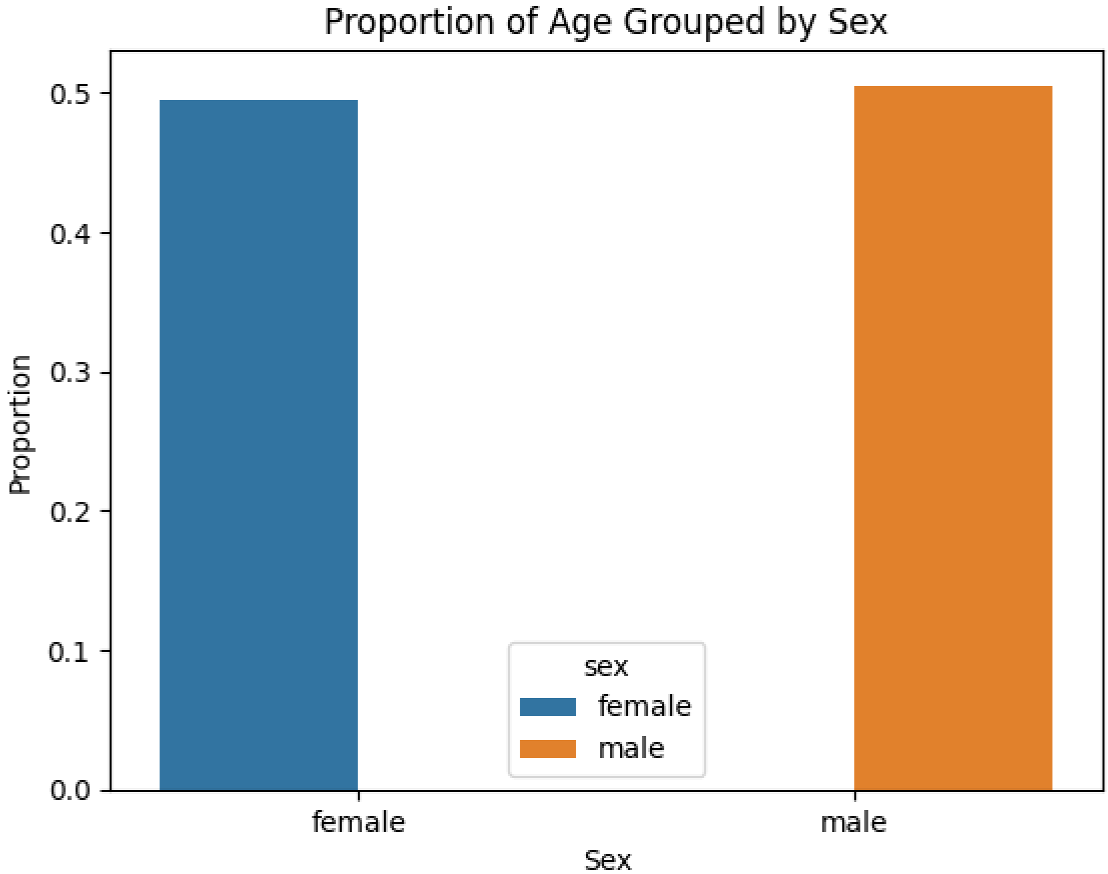


In [41]:
```
# Boxplot of charges grouped by age
sns.boxplot(x=data['age'], y=data['charges'])
plt.xlabel('Age')
plt.ylabel('Charges')
plt.title('Boxplot of Charges Grouped by Age')
plt.show()
```



In [42]:

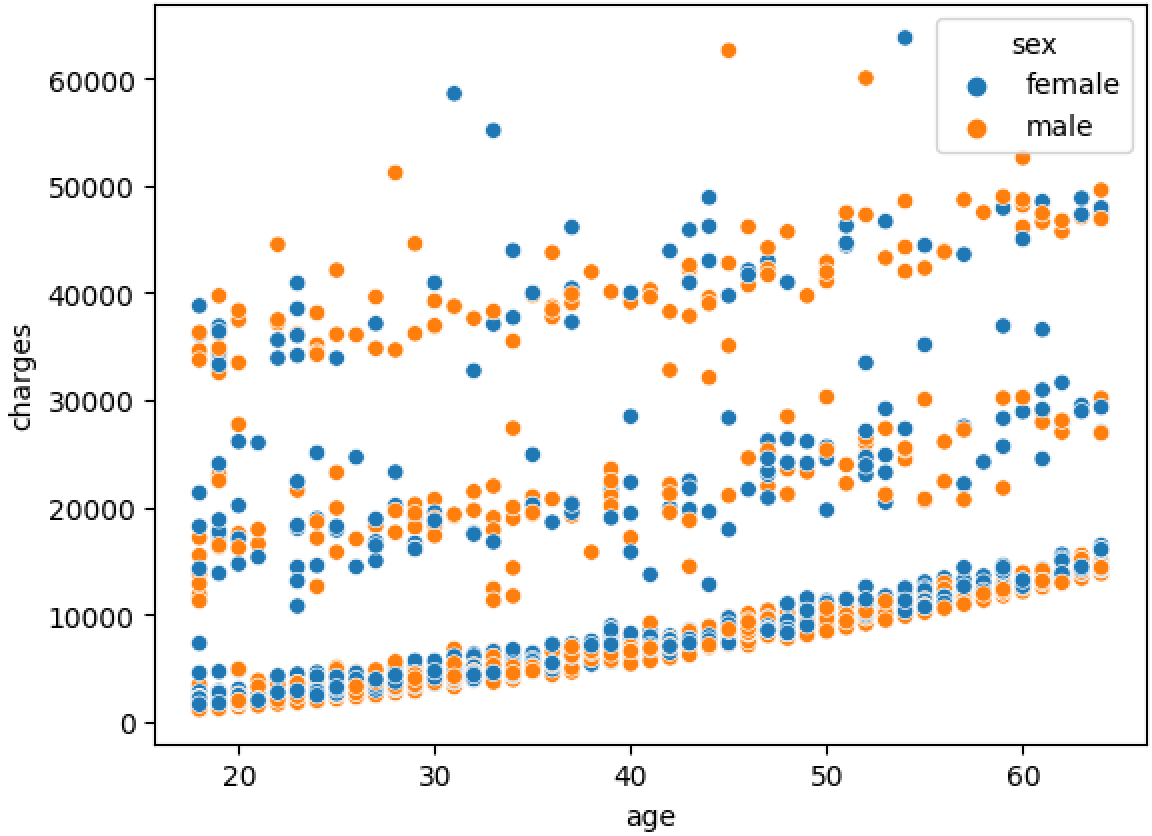
```
# Bar plot of age vs sex
sns.barplot(x=data['sex'], y=data['age'], hue=data['sex'], estimator=lambda x: len(x)
plt.xlabel('Sex')
plt.ylabel('Proportion')
plt.title('Proportion of Age Grouped by Sex')
plt.show()
```

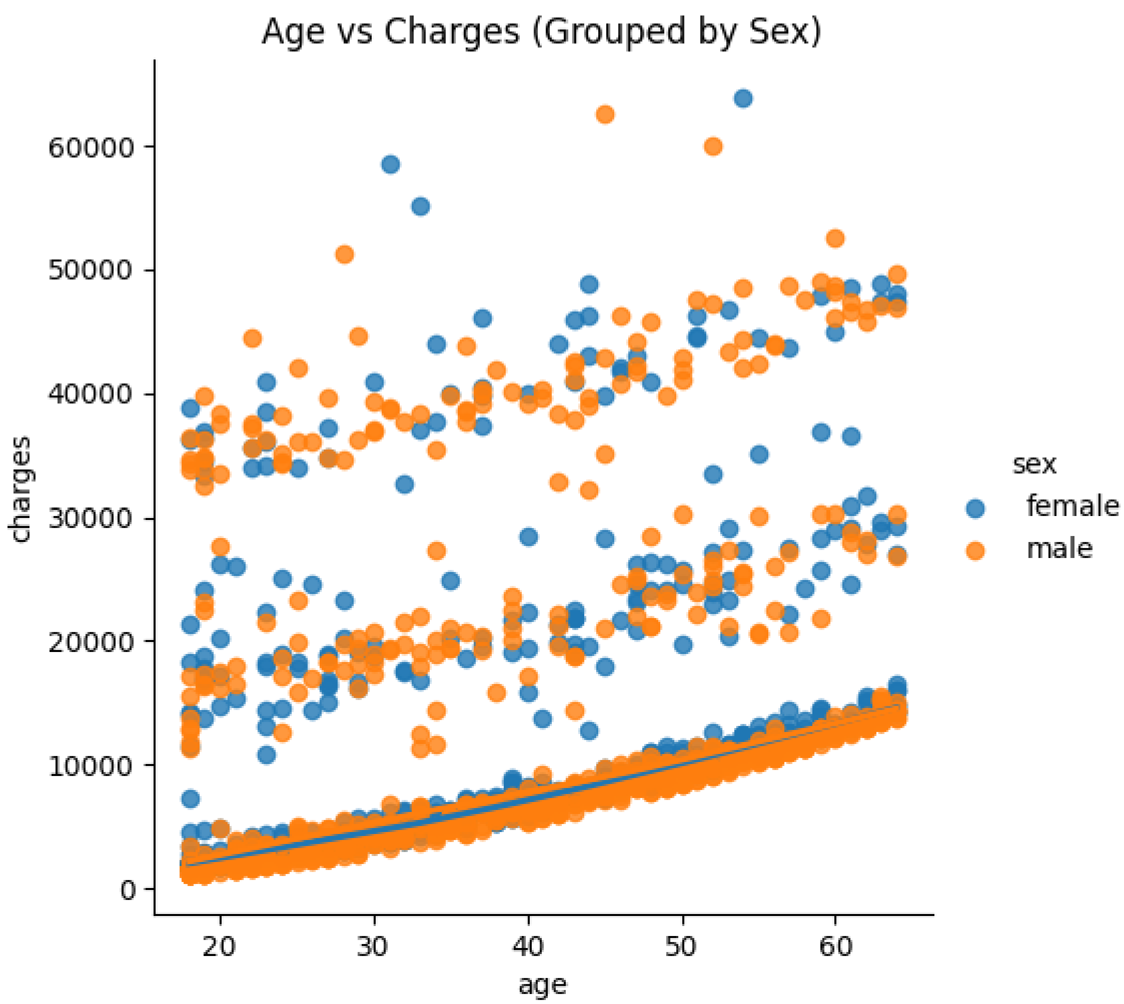


In [43]:

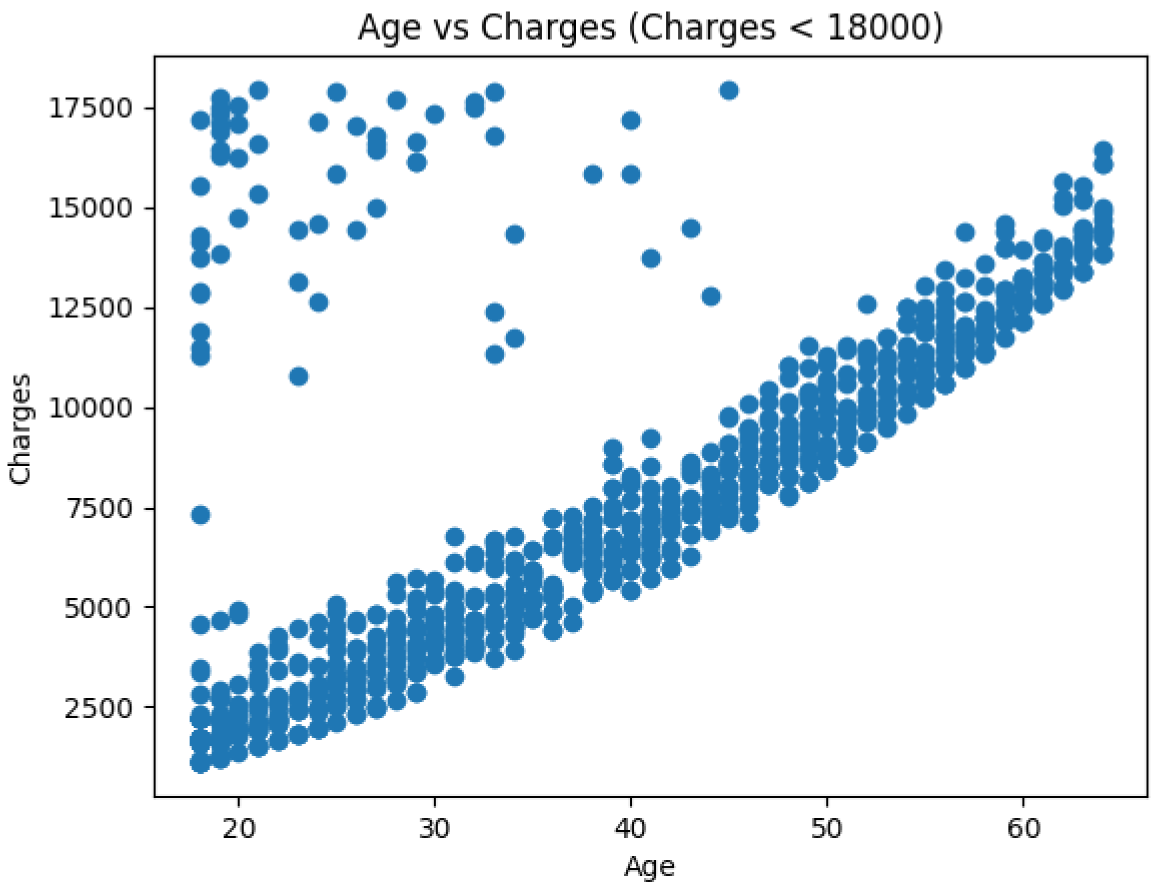
```
# Scatter plot of age vs charges with color differentiation for sex and smoothing line
sns.scatterplot(data=data, x='age', y='charges', hue='sex')
sns.lmplot(data=data, x='age', y='charges', hue='sex', lowess=True)
plt.title('Age vs Charges (Grouped by Sex)')
plt.show()
```

```
C:\Users\user\Anaconda3\lib\site-packages\seaborn\axisgrid.py:118: UserWarning: The f
igure layout has changed to tight
  self._figure.tight_layout(*args, **kwargs)
```

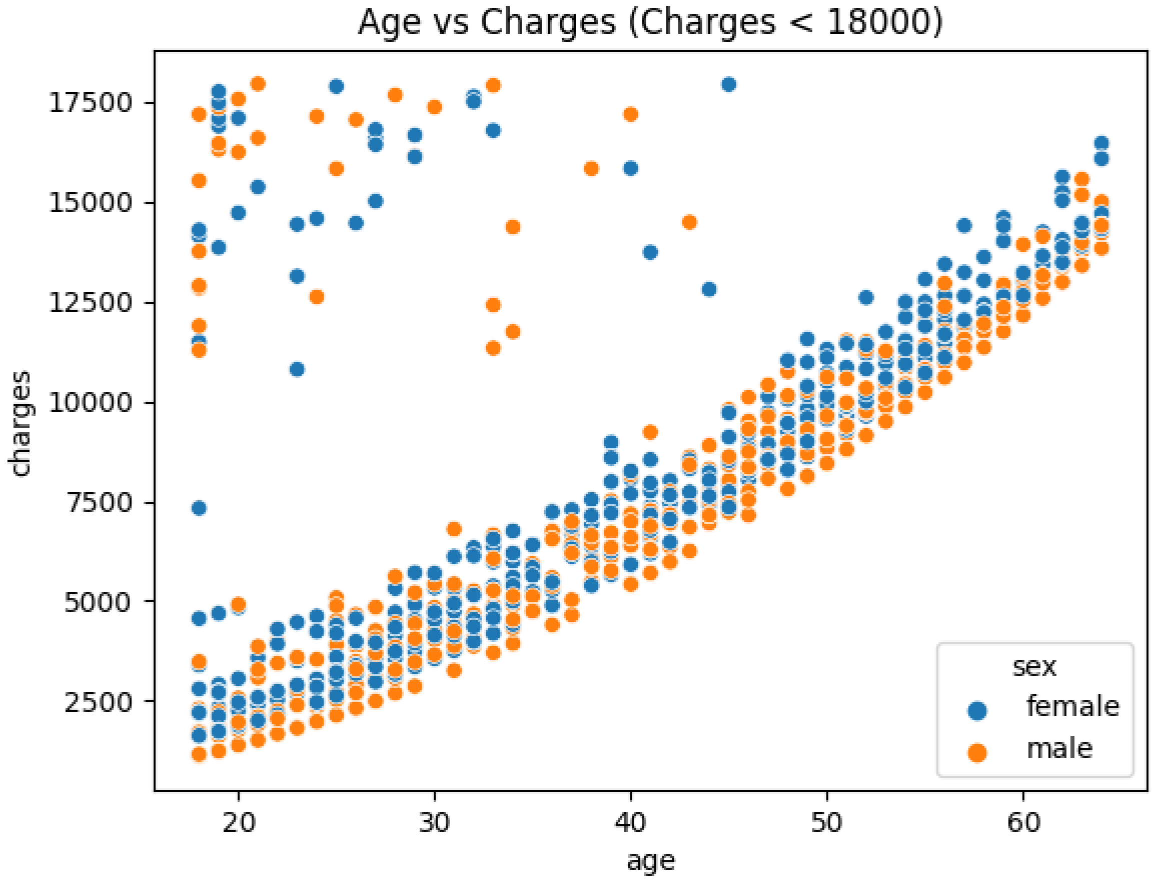

sex
female
male
charges
age
0
10000
20000
30000
40000
50000
60000
20
30
40
50
60


Age vs Charges (Grouped by Sex)
charges
age
sex
female
male
0
10000
20000
30000
40000
50000
60000
20
30
40
50
60

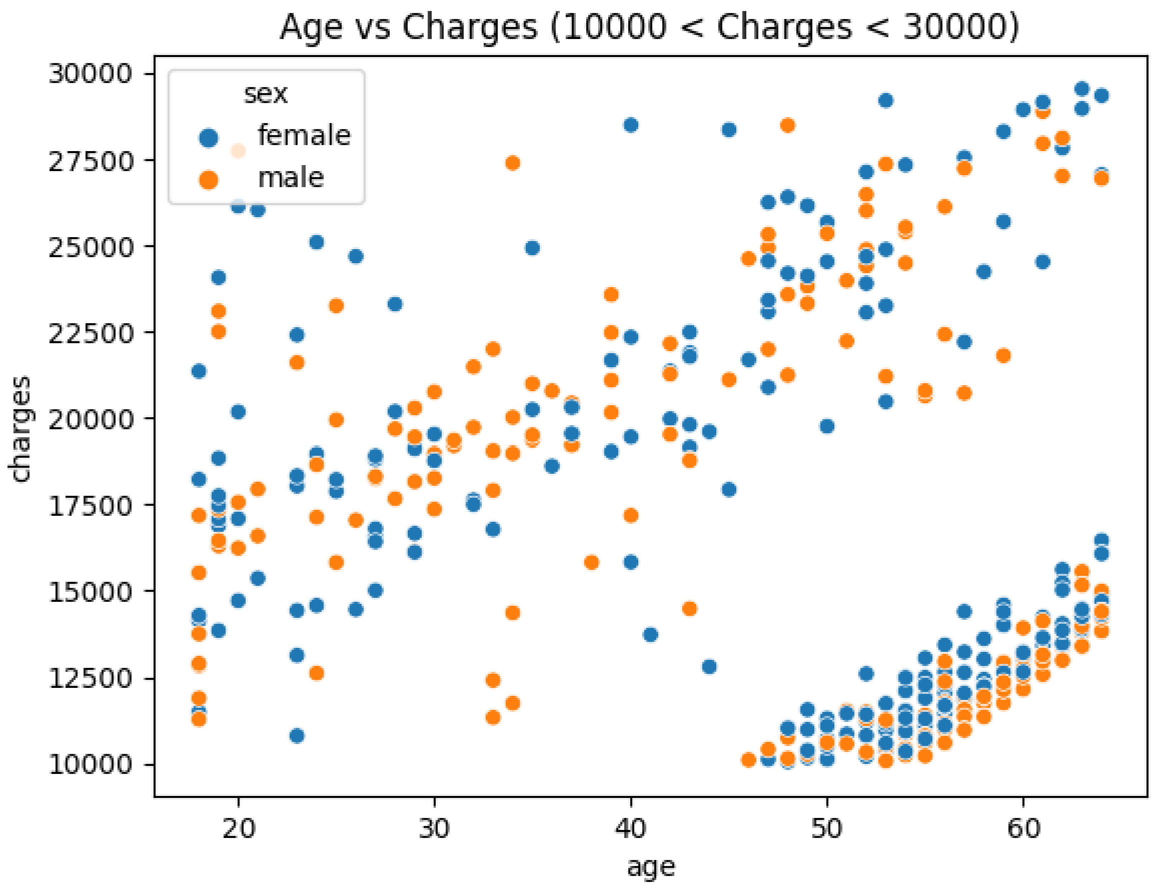
In [44]:
```
# Scatter plot of age vs charges for charges < 18000
plt.scatter(data[data['charges'] < 18000]['age'], data[data['charges'] < 18000]['charg
plt.xlabel('Age')
plt.ylabel('Charges')
plt.title('Age vs Charges (Charges < 18000)')
plt.show()
```



In [45]:
```
# Scatter plot using Seaborn for charges < 18000
sns.scatterplot(data=data[data['charges'] < 18000], x='age', y='charges', hue='sex')
plt.title('Age vs Charges (Charges < 18000)')
plt.show()
```

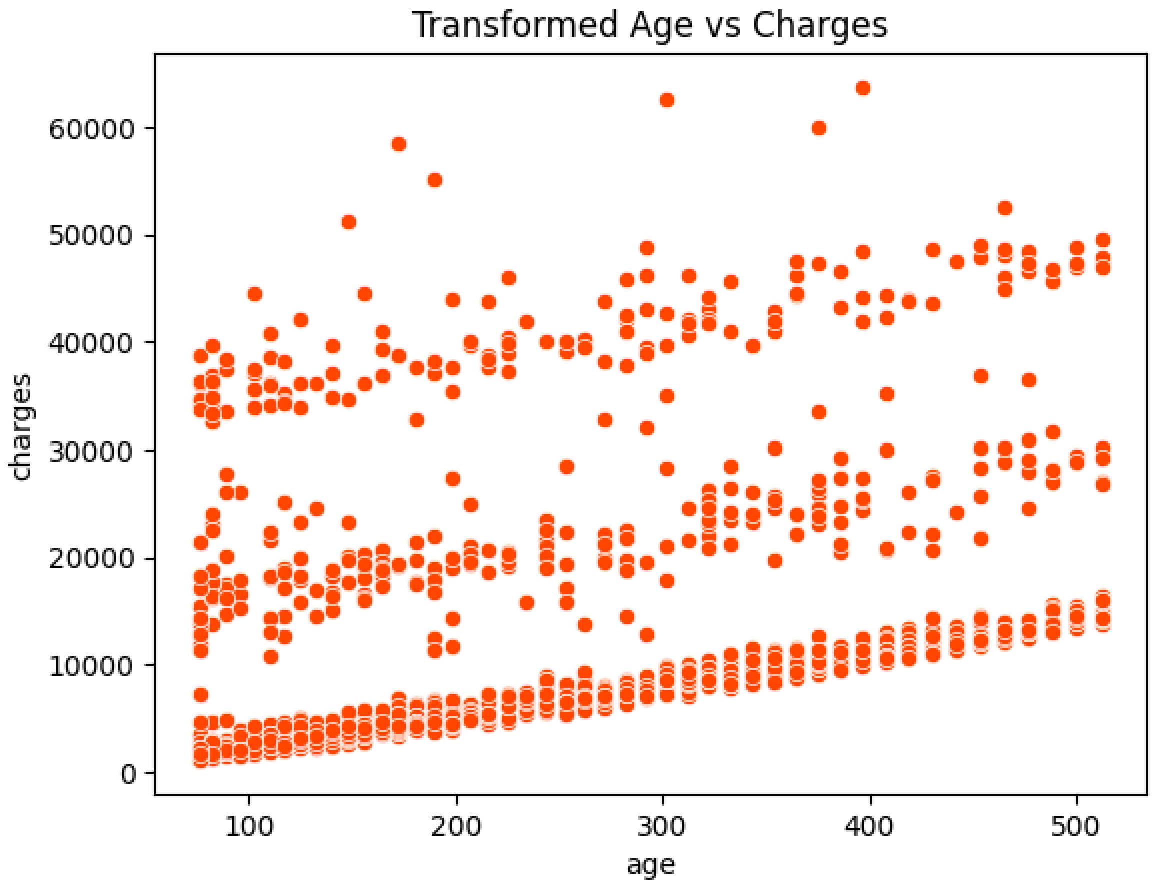


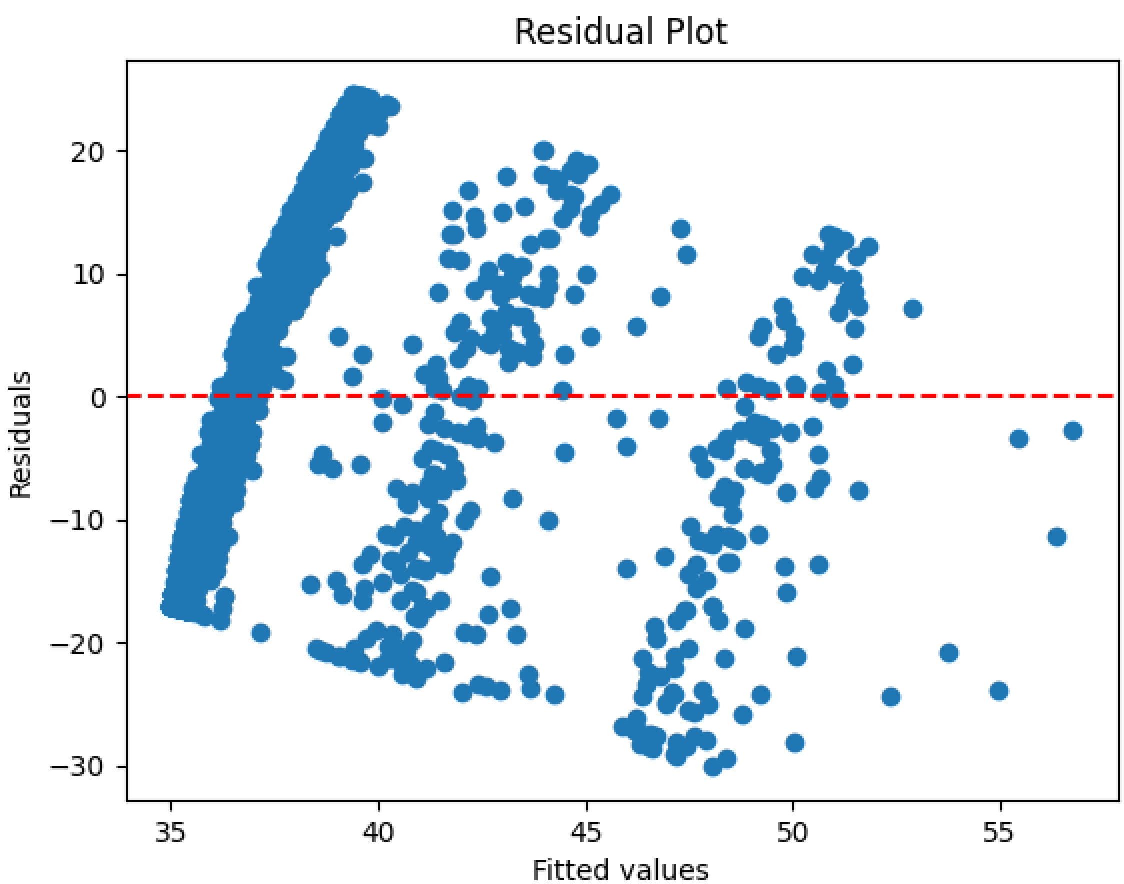
In [46]:
```
# Scatter plot using Seaborn for charges < 30000 and charges > 10000
sns.scatterplot(data=data[(data['charges'] < 30000) & (data['charges'] > 10000)],
                x='age', y='charges', hue='sex')
plt.title('Age vs Charges (10000 < Charges < 30000)')
plt.show()
```

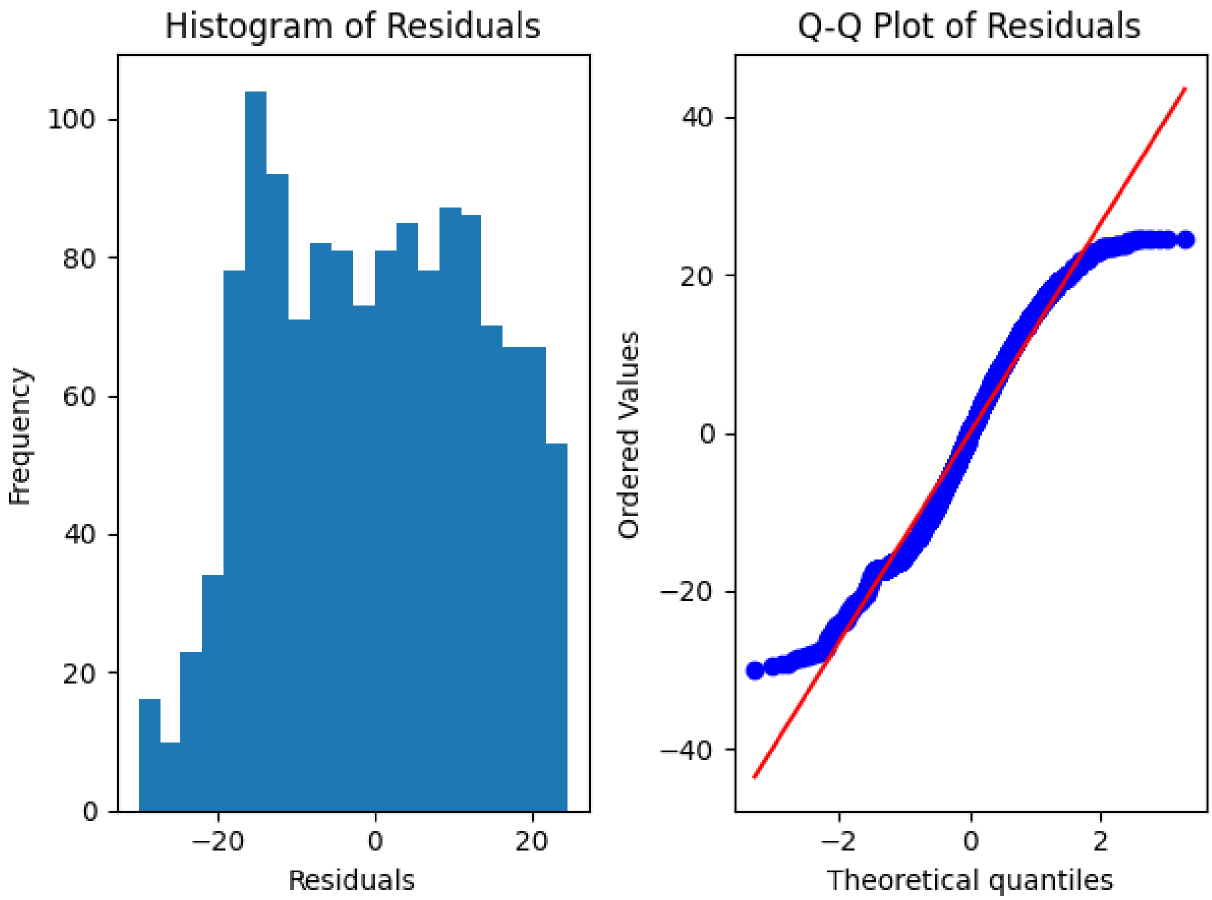


In [47]:

```
# Scatter plot of age^1.5 vs charges with color
sns.scatterplot(data=data, x=np.power(data['age'], 1.5), y='charges', color='orangerec
plt.title('Transformed Age vs Charges')
plt.show()

# Linear regression model
X = sm.add_constant(data['charges'])
y = data['age']
model = sm.OLS(y, X).fit()

# Residual plot
plt.scatter(model.predict(), model.resid)
plt.axhline(y=0, color='r', linestyle='--')
plt.xlabel('Fitted values')
plt.ylabel('Residuals')
plt.title('Residual Plot')
plt.show()

# Histogram and Q-Q plot of residuals
plt.subplot(1, 2, 1)
plt.hist(model.resid, bins=20)
plt.xlabel('Residuals')
plt.ylabel('Frequency')
plt.title('Histogram of Residuals')

plt.subplot(1, 2, 2)
stats.probplot(model.resid, plot=plt)
plt.title('Q-Q Plot of Residuals')

plt.tight_layout()
plt.show()
```


Transformed Age vs Charges
charges
age
60000
50000
40000
30000
20000
10000
0
100
200
300
400
500


Residual Plot
Residuals
Fitted values
20
10
0
−10
−20
−30
35
40
45
50
55


Histogram of Residuals
Frequency
Residuals
100
80
60
40
20
0
−20
0
20
Q-Q Plot of Residuals
Ordered Values
Theoretical quantiles
40
20
0
−20
−40
−2
0
2

In [ ]: