In [17]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as pl
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
```

In [24]:
```
pip install pandas openpyxl
```

```
Defaulting to user installation because normal site-packages is not writeable
Requirement already satisfied: pandas in c:\programdata\anaconda3\lib\site-packages
(1.5.3)
Requirement already satisfied: openpyxl in c:\programdata\anaconda3\lib\site-packages
(3.0.10)
Requirement already satisfied: python-dateutil>=2.8.1 in c:\programdata\anaconda3\lib
\site-packages (from pandas) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in c:\programdata\anaconda3\lib\site-pack
ages (from pandas) (2022.7)
Requirement already satisfied: numpy>=1.21.0 in c:\programdata\anaconda3\lib\site-pac
kages (from pandas) (1.24.3)
Requirement already satisfied: et_xmlfile in c:\programdata\anaconda3\lib\site-packag
es (from openpyxl) (1.1.0)
Requirement already satisfied: six>=1.5 in c:\programdata\anaconda3\lib\site-packages
(from python-dateutil>=2.8.1->pandas) (1.16.0)
Note: you may need to restart the kernel to use updated packages.
```

In [29]:
```
df = pd.read_excel(r"C:\Users\premt\OneDrive\Desktop\Medical Cost Analysis Data Set.xl
df.head()
```

Out[29]:

	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

In [32]:
```
df.dtypes
```

Out[32]:
```
age           int64
sex          object
bmi         float64
children      int64
smoker       object
region       object
charges     float64
dtype: object
```

In [33]:
```
df.isnull().sum()
```

Out[33]:
```
age         0
sex         0
bmi         0
children    0
smoker      0
region      0
charges     0
dtype: int64
```

In [36]:
```
from sklearn.preprocessing import LabelEncoder
df_aug = pd.read_excel(r"C:\Users\premt\OneDrive\Desktop\Medical Cost Analysis Data Se
#sex
le = LabelEncoder()
le.fit(df_aug.sex.drop_duplicates())
df_aug.sex = le.transform(df_aug.sex)
# smoker or not
le.fit(df_aug.smoker.drop_duplicates())
df_aug.smoker = le.transform(df_aug.smoker)
#region
le.fit(df_aug.region.drop_duplicates())
df_aug.region = le.transform(df_aug.region)
```

In [37]:
```
df_aug
```

Out[37]:

	age	sex	bmi	children	smoker	region	charges
0	19	0	27.900	0	1	3	16884.92400
1	18	1	33.770	1	0	2	1725.55230
2	28	1	33.000	3	0	2	4449.46200
3	33	1	22.705	0	0	1	21984.47061
4	32	1	28.880	0	0	1	3866.85520
...	...	...	...	...	...	...	...
1333	50	1	30.970	3	0	1	10600.54830
1334	18	0	31.920	0	0	0	2205.98080
1335	18	0	36.850	0	0	2	1629.83350
1336	21	0	25.800	0	0	3	2007.94500
1337	61	0	29.070	0	1	1	29141.36030

1338 rows × 7 columns

In [38]:
```
df_aug.region.value_counts()
```

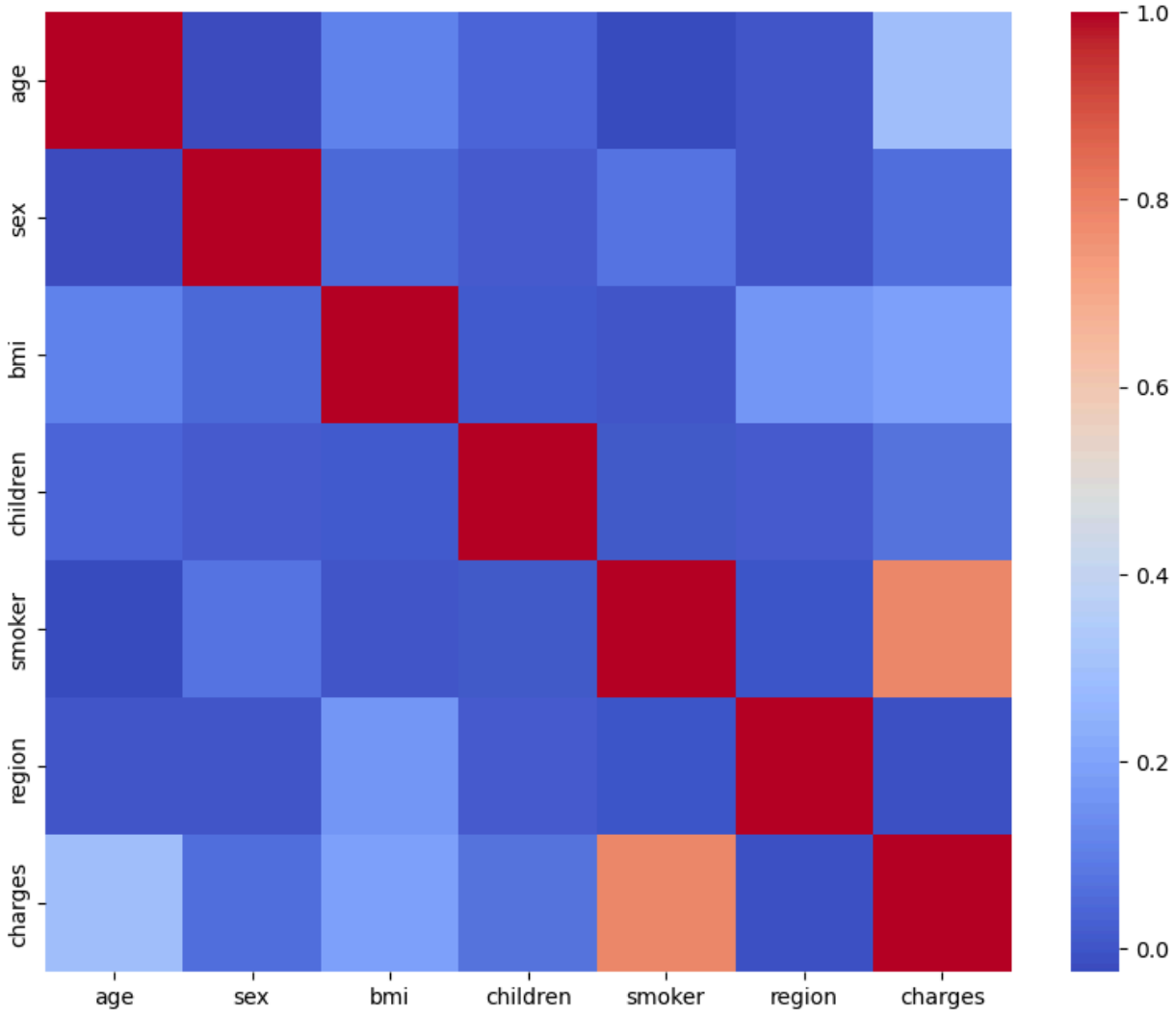
Out[38]:
```
2    364
3    325
1    325
0    324
Name: region, dtype: int64
```

In [39]:
```
df.region.value_counts()
```

Out[39]:
```
southeast    364
southwest    325
northwest    325
northeast    324
Name: region, dtype: int64
```

In [40]:
```
f, ax = pl.subplots(figsize=(10, 8))
corr = df_aug.corr()
sns.heatmap(corr, cmap='coolwarm')
```

Out[40]:
```
<Axes: >
```



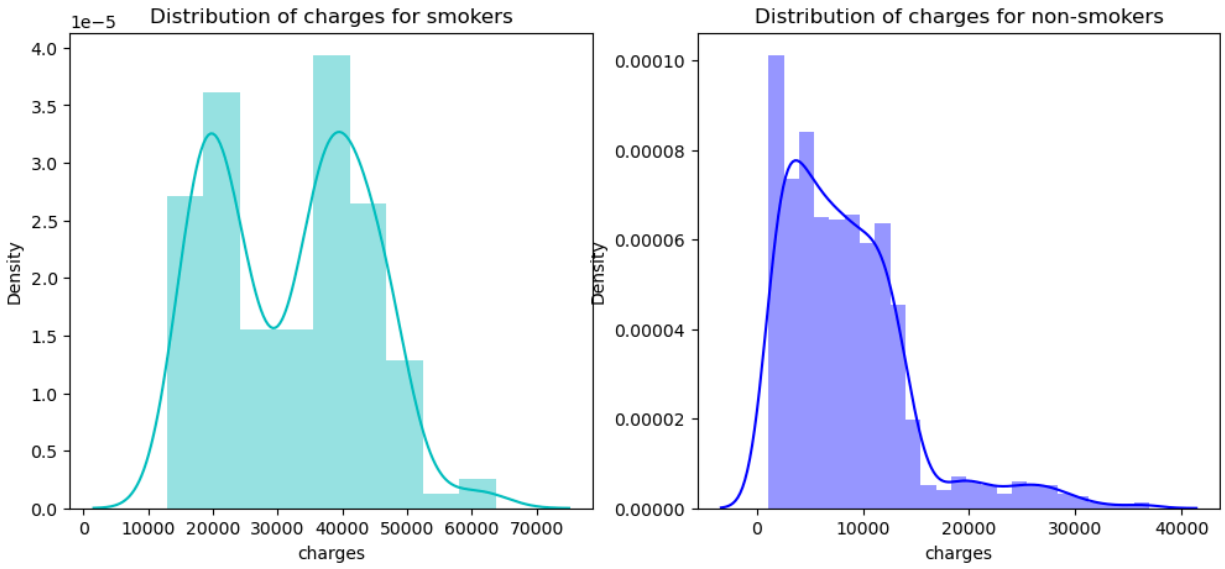
In [41]:
```
corr['charges'].sort_values()
```

Out[41]:
```
region     -0.006208
sex         0.057292
children    0.067998
bmi         0.198341
age         0.299008
smoker      0.787251
charges     1.000000
Name: charges, dtype: float64
```

In [45]:
```
f= pl.figure(figsize=(12,5))

ax=f.add_subplot(121)
```

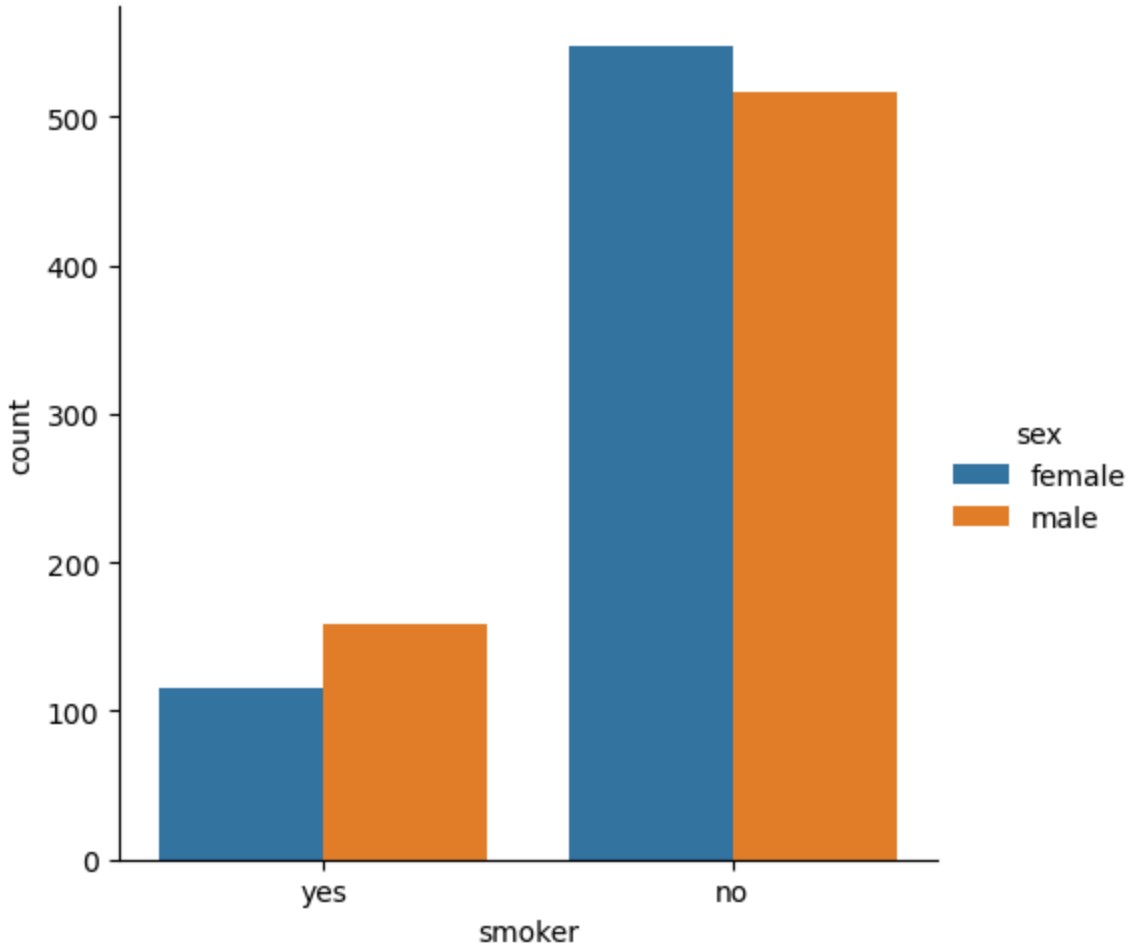
```
sns.distplot(df_aug[(df.smoker == 'yes')]["charges"],color='c')
ax.set_title('Distribution of charges for smokers')

ax=f.add_subplot(122)
sns.distplot(df_aug[(df_aug.smoker == 0)]['charges'],color='b')
ax.set_title('Distribution of charges for non-smokers')
```

Out[45]: Text(0.5, 1.0, 'Distribution of charges for non-smokers')



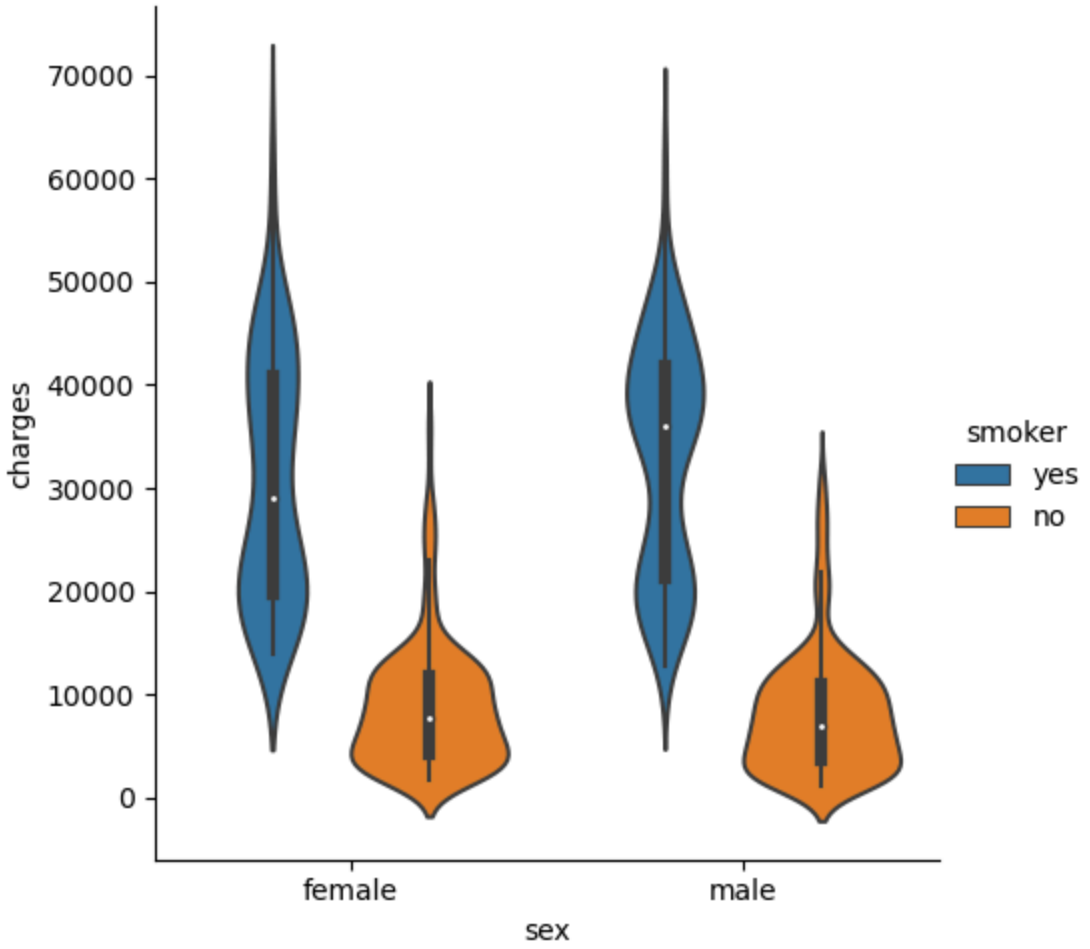
In [46]:
```
sns.catplot(x="smoker", kind="count",hue = 'sex',  data=df)
```

Out[46]: <seaborn.axisgrid.FacetGrid at 0x229f81d09d0>



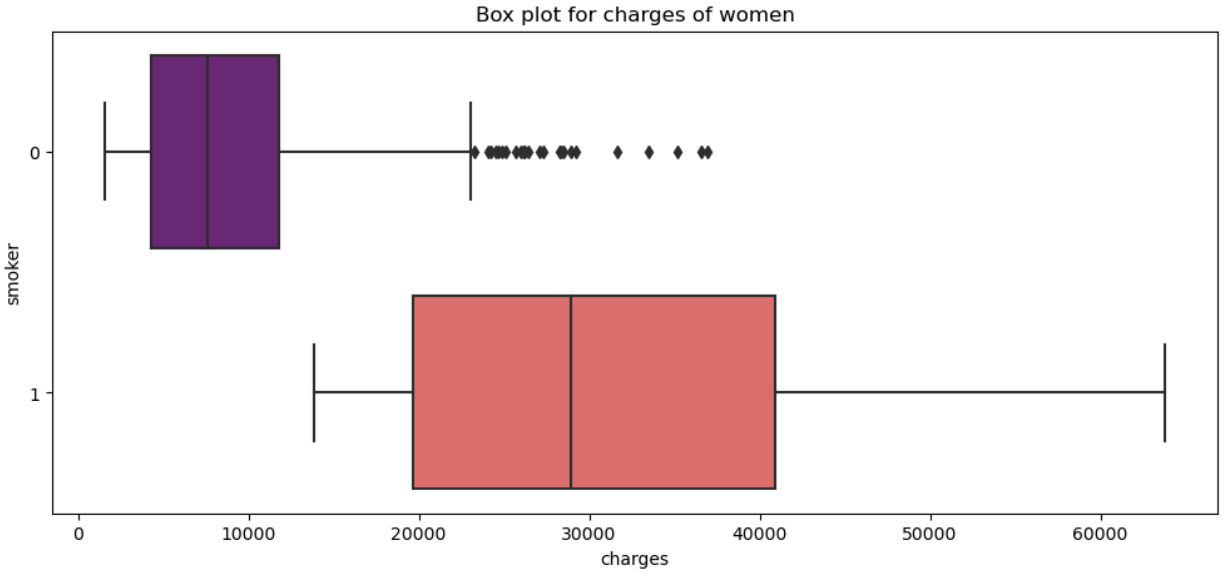
In [47]:
```
sns.catplot(x="sex", y="charges", hue="smoker", kind="violin", data=df)
```

Out[47]:
```
<seaborn.axisgrid.FacetGrid at 0x229f8247810>
```



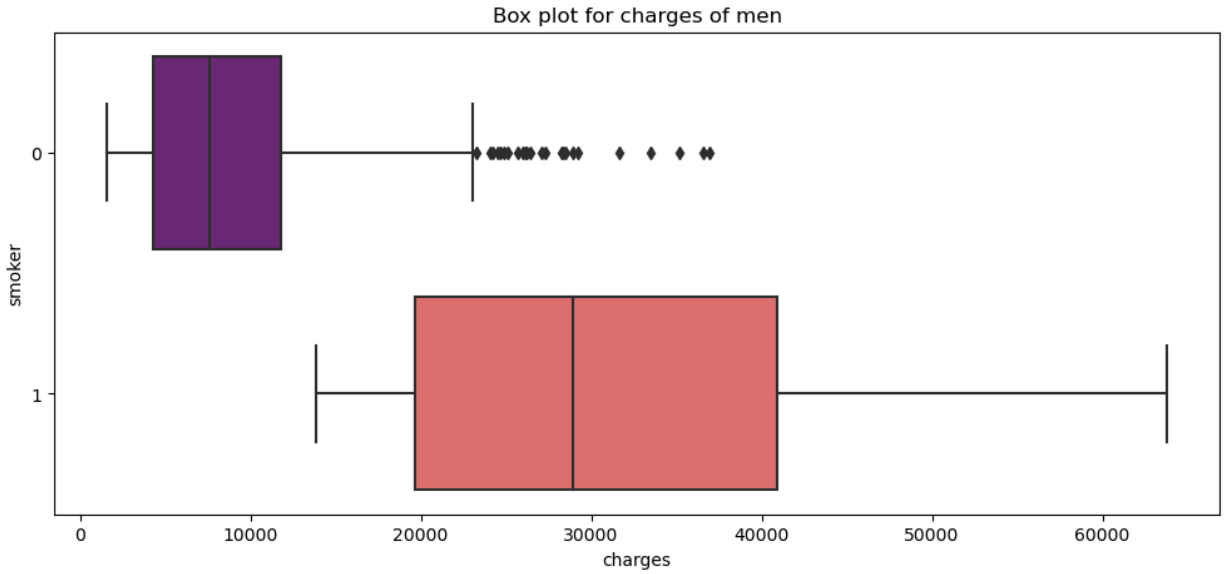
In [48]:
```
pl.figure(figsize=(12,5))
pl.title("Box plot for charges of women")
sns.boxplot(y="smoker", x="charges", data =  df_aug[(df_aug.sex == 0)] , orient="h", p
```

Out[48]:
```
<Axes: title={'center': 'Box plot for charges of women'}, xlabel='charges', ylabel='s
moker'>
```

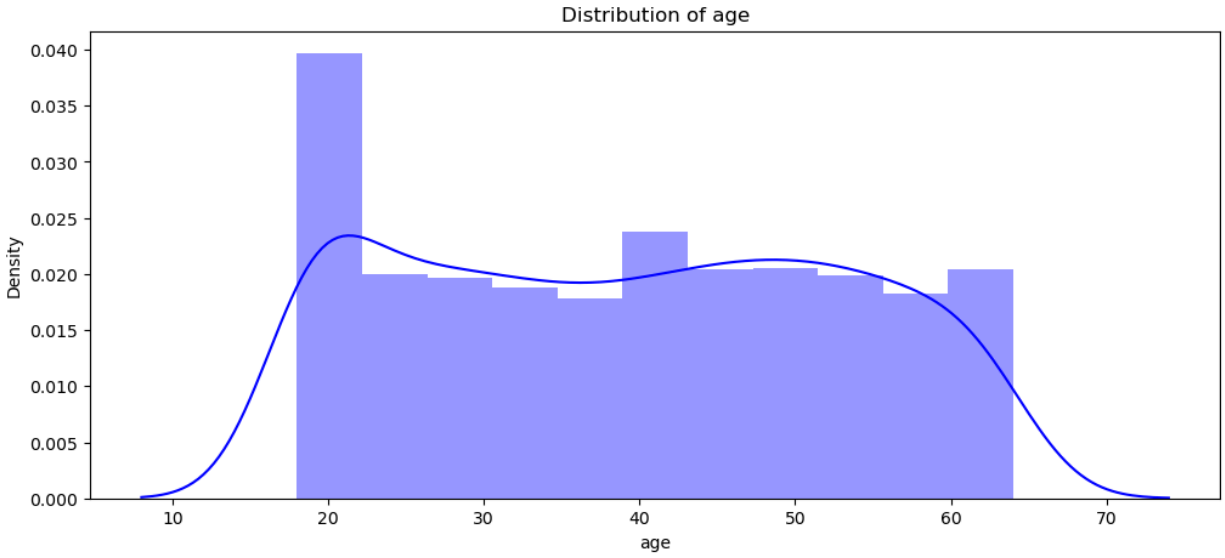


In [49]:
```
pl.figure(figsize=(12,5))
pl.title("Box plot for charges of men")
sns.boxplot(y="smoker", x="charges", data =  df_aug[(df_aug.sex == 0)] , orient="h", p
```

Out[49]:
```
<Axes: title={'center': 'Box plot for charges of men'}, xlabel='charges', ylabel='smo
ker'>
```

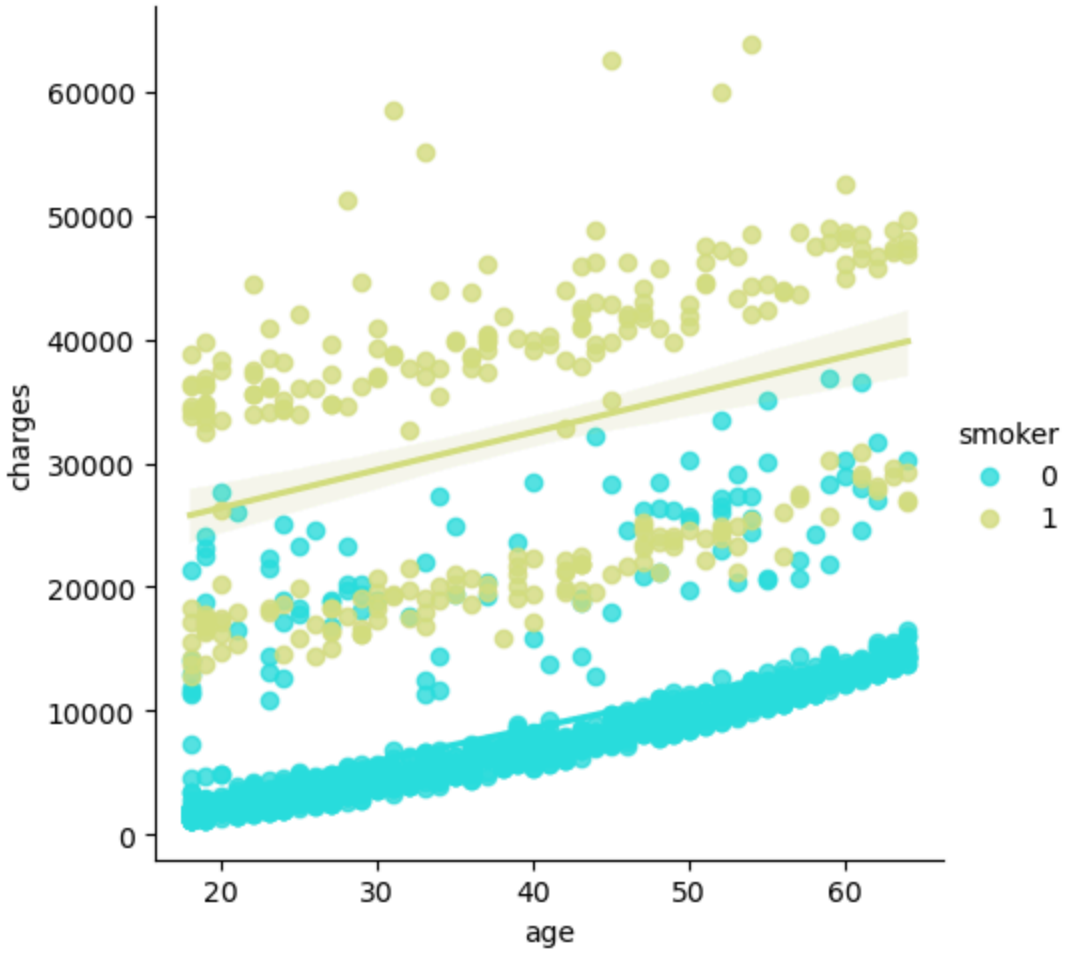


In [50]:
```
pl.figure(figsize=(12,5))
pl.title("Distribution of age")
ax = sns.distplot(df_aug["age"], color = 'b')
```

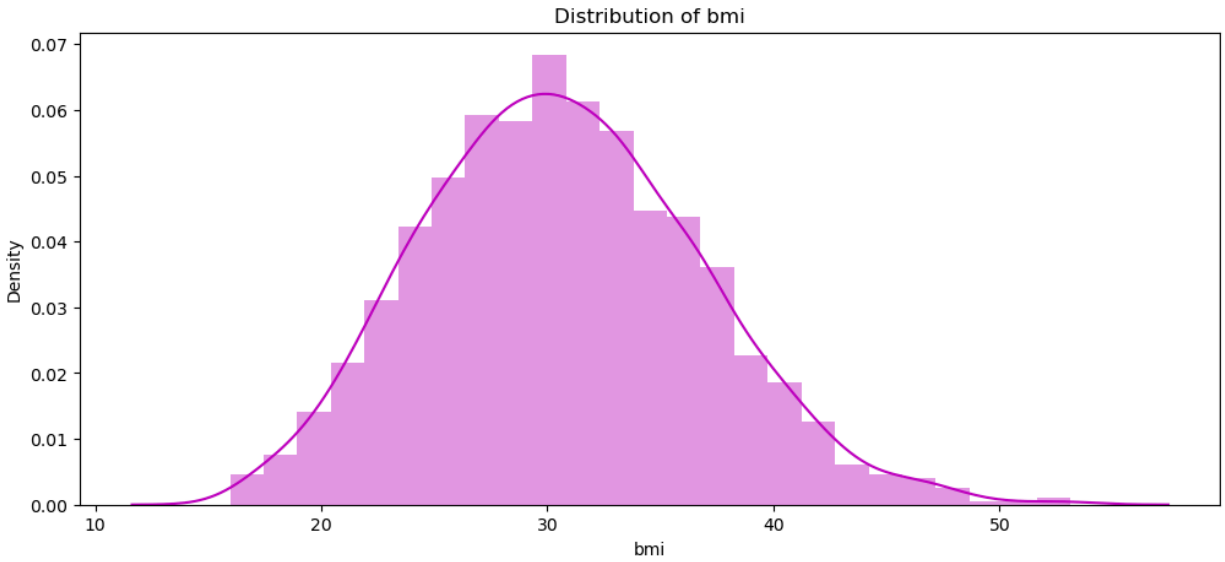


In [51]:
```
sns.lmplot(x="age", y="charges", hue="smoker", data=df_aug, palette = 'rainbow')
ax.set_title('Smokers and non-smokers')
```

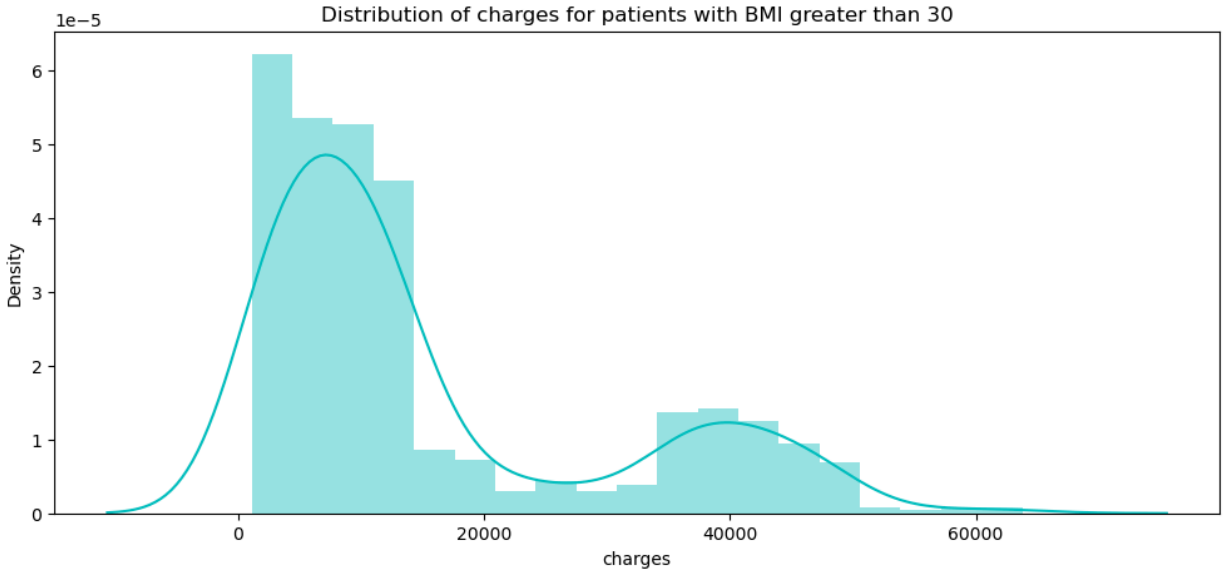
Out[51]:
```
Text(0.5, 1.0, 'Smokers and non-smokers')
```



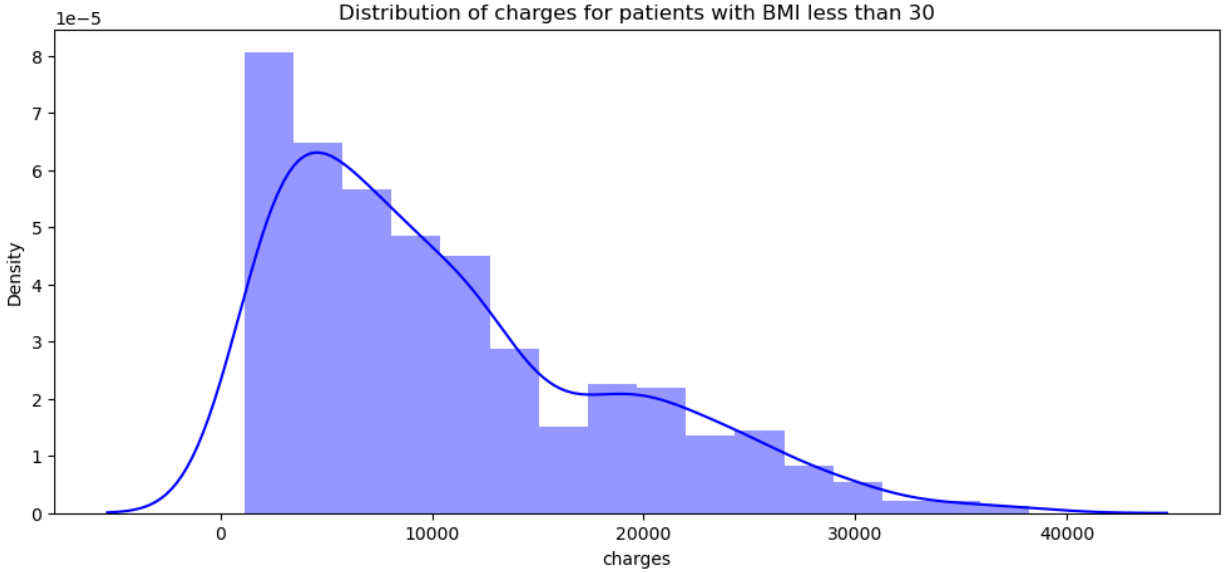
In [52]:
```
pl.figure(figsize=(12,5))
pl.title("Distribution of bmi")
ax = sns.distplot(df.bmi, color = 'm')
```



In [53]:
```
pl.figure(figsize=(12,5))
pl.title("Distribution of charges for patients with BMI greater than 30")
ax = sns.distplot(df[(df.bmi >= 30)]['charges'], color = 'c')
```

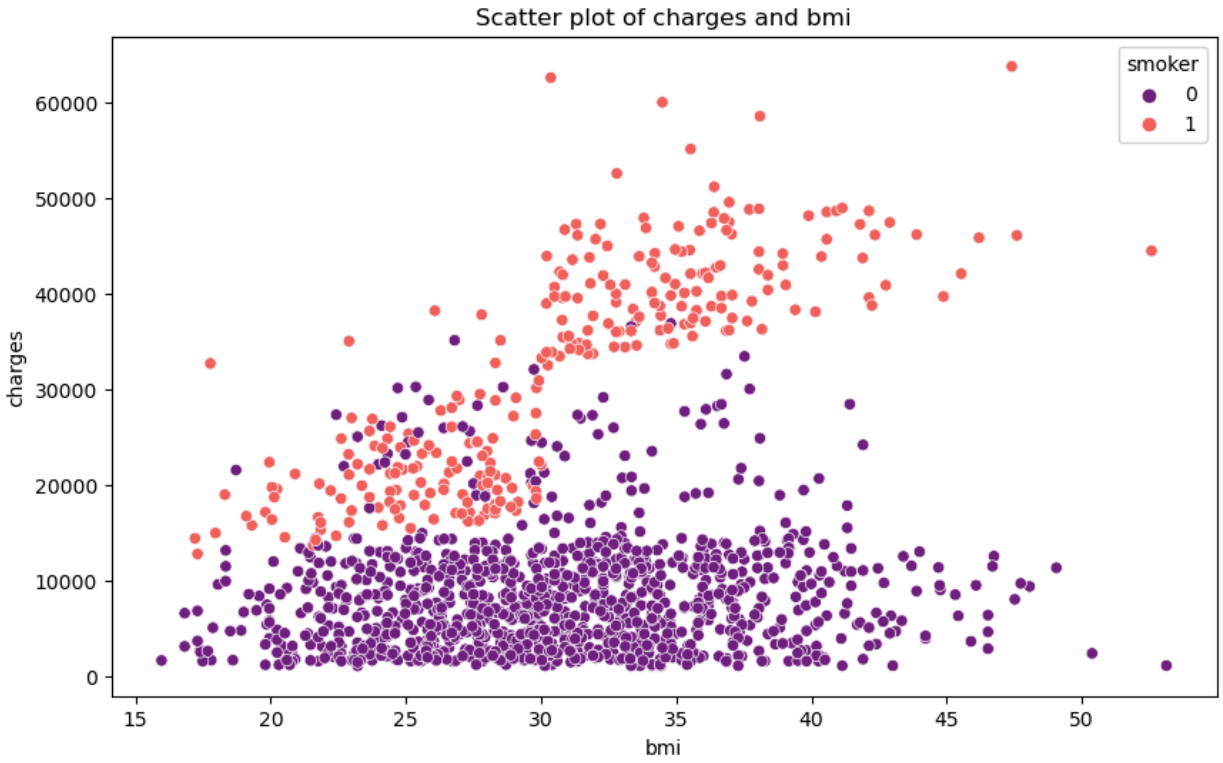


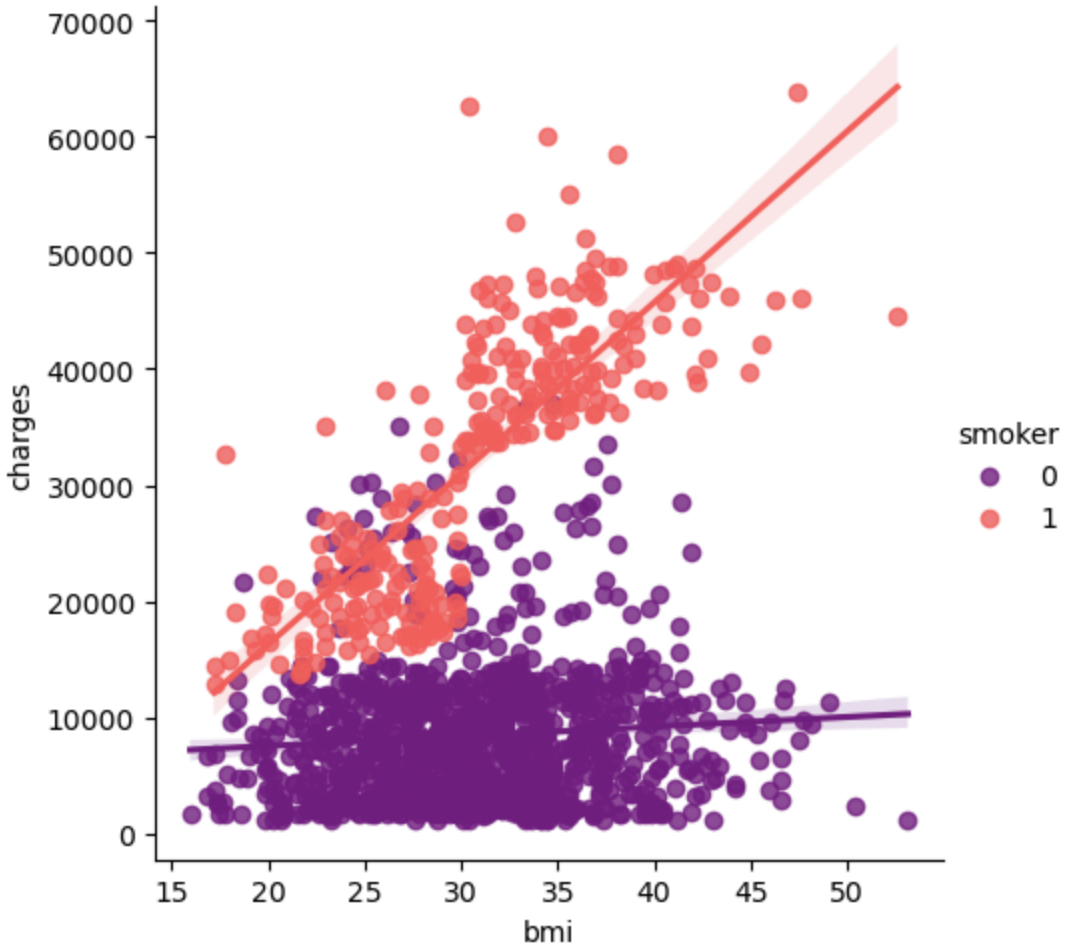
In [54]:
```
pl.figure(figsize=(12,5))
pl.title("Distribution of charges for patients with BMI less than 30")
ax = sns.distplot(df[(df.bmi < 30)]['charges'], color = 'b')
```



In [55]:
```
pl.figure(figsize=(10,6))
ax = sns.scatterplot(x='bmi',y='charges',data=df_aug,palette='magma',hue='smoker')
ax.set_title('Scatter plot of charges and bmi')

sns.lmplot(x="bmi", y="charges", hue="smoker", data=df_aug, palette = 'magma')
```

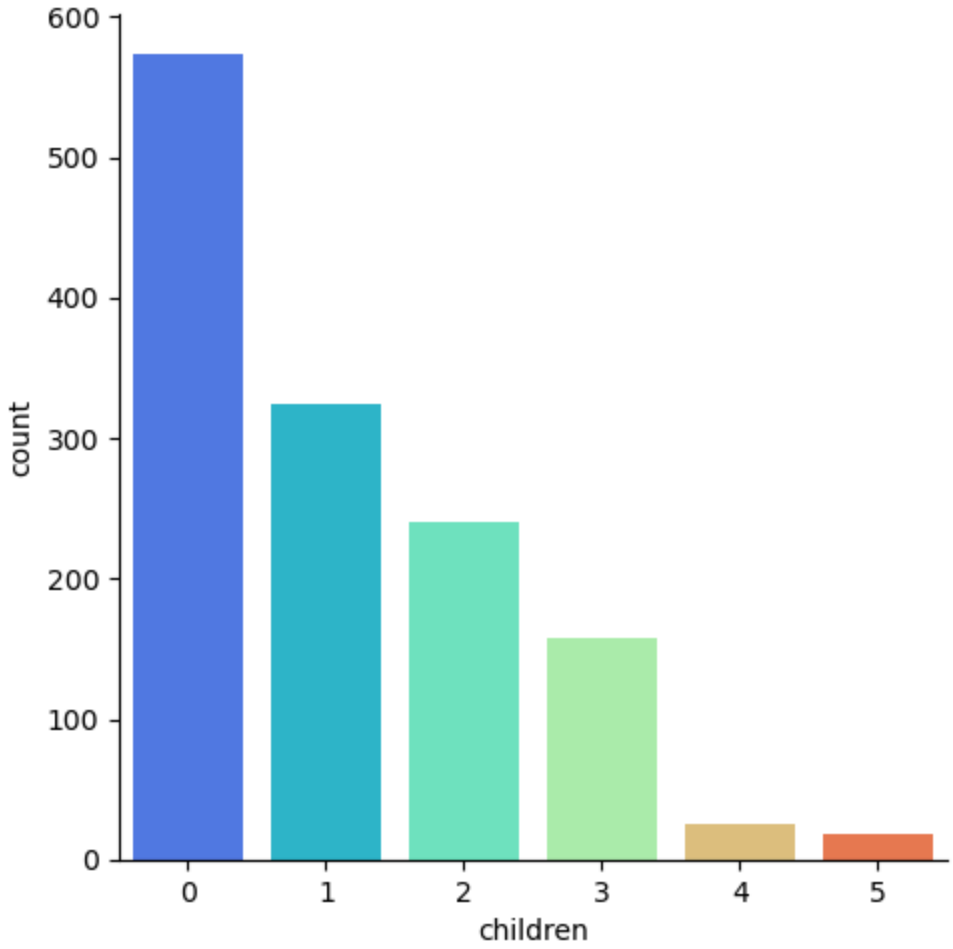
Out[55]:
```
<seaborn.axisgrid.FacetGrid at 0x229f92afd90>
```





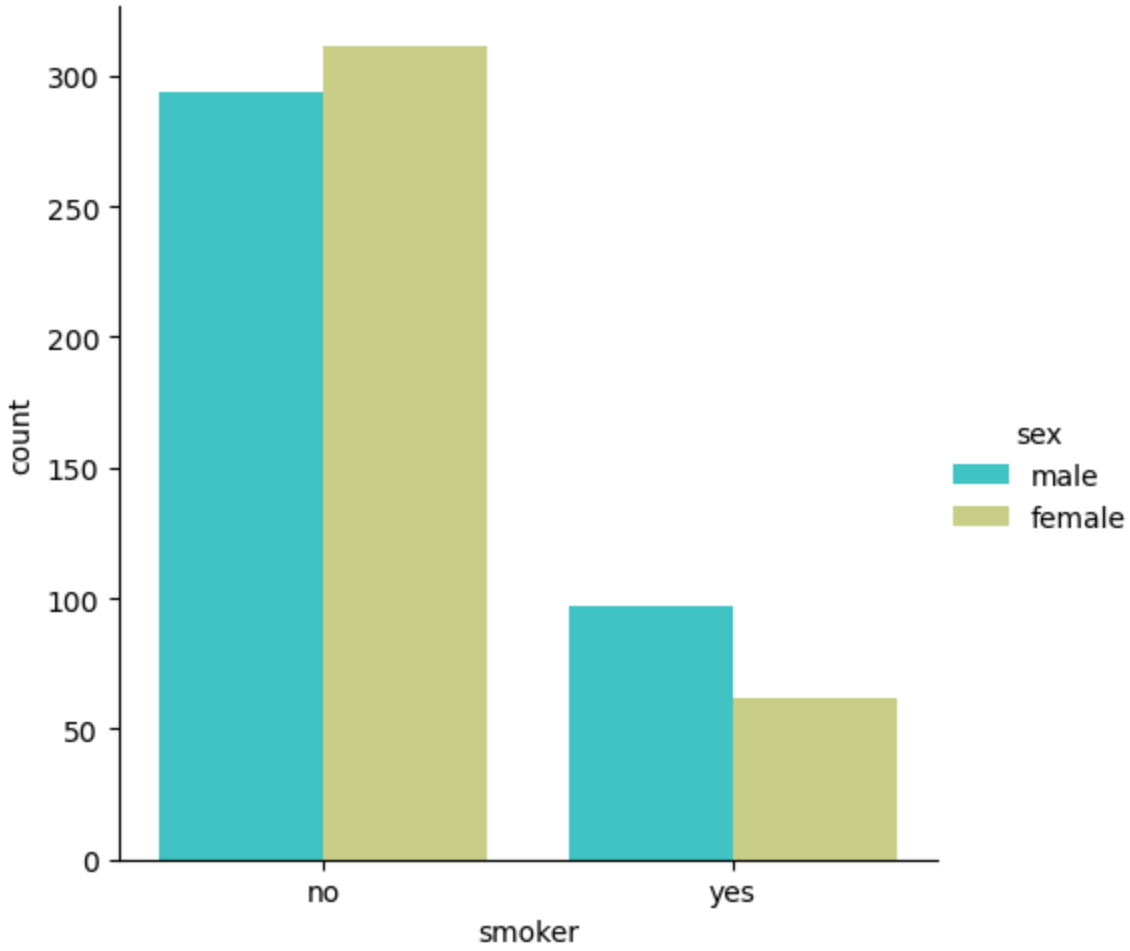
In [56]:
```
sns.catplot(x="children", kind="count", palette="rainbow", data=df_aug)
```

Out[56]:
```
<seaborn.axisgrid.FacetGrid at 0x229f913f950>
```



In [57]:
```
sns.catplot(x="smoker", kind="count", palette="rainbow",hue = "sex",
            data=df[(df.children > 0)])
ax.set_title('Smokers and non-smokers who have childrens')
```

Out[57]:
```
Text(0.5, 1.0, 'Smokers and non-smokers who have childrens')
```



In [58]:
```
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import PolynomialFeatures
from sklearn.metrics import r2_score,mean_squared_error,accuracy_score
from sklearn.ensemble import RandomForestRegressor
```

In [59]:
```
df_aug
```

Out[59]:

	age	sex	bmi	children	smoker	region	charges
0	19	0	27.900	0	1	3	16884.92400
1	18	1	33.770	1	0	2	1725.55230
2	28	1	33.000	3	0	2	4449.46200
3	33	1	22.705	0	0	1	21984.47061
4	32	1	28.880	0	0	1	3866.85520
...	...	...	...	...	...	...	...
1333	50	1	30.970	3	0	1	10600.54830
1334	18	0	31.920	0	0	0	2205.98080
1335	18	0	36.850	0	0	2	1629.83350
1336	21	0	25.800	0	0	3	2007.94500
1337	61	0	29.070	0	1	1	29141.36030

1338 rows × 7 columns

In [60]:
```
x = df_aug.drop(['charges'], axis = 1)
y = df_aug.charges

x_train,x_test,y_train,y_test = train_test_split(x,y, random_state = 0)
lr = LinearRegression()
lr.fit(x_train,y_train)

y_train_pred = lr.predict(x_train)
y_test_pred = lr.predict(x_test)

print(lr.score(x_test,y_test)*100,"%")
```

```
79.62732059725785 %
```

In [61]:
```
X = df_aug.drop(['charges','region'], axis = 1)
Y = df_aug.charges



quad = PolynomialFeatures (degree = 2)
x_quad = quad.fit_transform(X)

X_train,X_test,Y_train,Y_test = train_test_split(x_quad,Y, random_state = 0)

plr = LinearRegression().fit(X_train,Y_train)

Y_train_pred = plr.predict(X_train)
Y_test_pred = plr.predict(X_test)

print(plr.score(X_test,Y_test)*100,"%")
```

```
88.49197344147235 %
```

In [ ]: