Medical Insurance Cost

In [6]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import LabelEncoder
```

Description of the data set and a summary of its attributes

In this data set, we are going to see wich factors influence the price of Health Insurance. Lets take a look to the data set:

In [4]:
```
df=pd.read_csv("insurance.csv")
df.head()
```

Out[4]:

	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

These are the factors we are going to consider in this analisys:

1. age: age of the beneficiary
2. sex: gender of the beneficiary: female, male
3. bmi: body mass index of the beneficiary
4. children: number of children covered by health insurance
5. smoker: if the beneficiary is a smoker or not
6. region: the beneficiary's residential area in the US

Formulating at least 3 hypothesis about this data

1. Smoker will be the most important factor in order to predict the insurance cost.
2. BMI and age will be important factors as well.
3. Region wont be a relevant factor.

Initial plan for data exploration

First we are going to take a look into the data and see if there are some null values:

In [7]:
```
len(df)
```

Out[7]:
```
1338
```

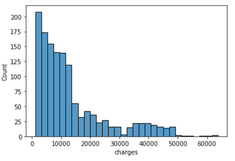
In [8]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 7 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   age       1338 non-null   int64  
 1   sex       1338 non-null   object 
 2   bmi       1338 non-null   float64
 3   children  1338 non-null   int64  
 4   smoker    1338 non-null   object 
 5   region    1338 non-null   object 
 6   charges   1338 non-null   float64
dtypes: float64(2), int64(2), object(3)
memory usage: 73.3+ KB
```

We can see that there's no null values, so we don't need to fill empty spaces.

It would be nice to see the distribution of the variable charges in order to explore the data

In [9]:
```
sns.histplot(data=df, x="charges");
```



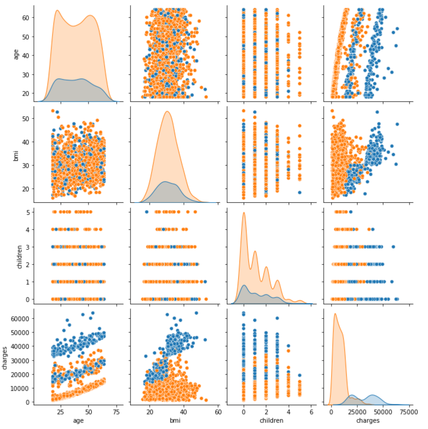
Now, let's investigate what's the minimum value in charges

In [10]:
```
min(df["charges"])
```

Out[10]:
```
1121.8739
```

Now, pairplot would be usefull for the data exploration. Let's take a look. I believe that the smoker column would be very relevant so I'm going to make a division

In [13]:
```
sns.pairplot(df, hue="smoker");
```



Definitely the column smoker is relevant to predict the charge column.

Actions taken for data cleaning and feature engineering

We can take a look at the correlation matrix in order to see if the smoker column is realy relevant. For that we need to give skmoker a numeric value

In [14]:
```
# Haciendo a smoker una variable numérica
df["smoker"] = df["smoker"].astype("category")
lab=LabelEncoder()
lab.fit(df["smoker"].drop_duplicates())
df["smoker"] = lab.transform(df["smoker"])
df.head()
```

Out[14]:

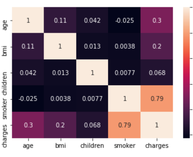
	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	1	southwest	16884.92400
1	18	male	33.770	1	0	southeast	1725.55230
2	28	male	33.000	3	0	southeast	4449.46200
3	33	male	22.705	0	0	northwest	21984.47061
4	32	male	28.880	0	0	northwest	3866.85520

the smokers are labeled with 1 and no smokers with 0. Now let's see the correlation matrix.

Key Findings and Insights, which synthesizes the results of Exploratory Data Analysis in an insightful and actionable manner

Smoker is definitely important, also age and BMI.

In [15]:
```
sns.heatmap(df.corr(), annot=True);
```

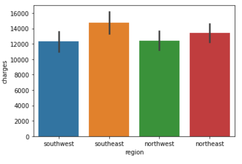


Conducting a formal significance test for one of the hypotheses and discuss the results

We can now confirm the first two hypothesis, smoker is the most important factor, age and bmi are also relevant. That make a lot of sense, smokers play a lot with their health and insurance companies know the smoking consecuences so it's logical that smokers pay more.

Let's see if region is irrelevant.

In [19]:
```
sns.barplot(x='region', y='charges', data=df);
```



There is some correlation between the region and the charge. But its minimun.

Suggestions for next steps in analyzing this data

It would be nice to make a regression to see if we can predict the charge amount of a posible beneficiary.

A paragraph that summarizes the quality of this data set and a request for additional data if needed

The data was really clean, we just needed to give numerical values to some data. It would be nice to add alcohol consumption in the data in order to see some other relationships.

Source: https://www.kaggle.com/mirichoi0218/insurance