Import libraries

In [1]:
```python
import pandas as pd # Data manipulation
import numpy as np # Linear algebra
import warnings # Ignore warnings
import seaborn as sns #plots
import matplotlib.pyplot as plt # plots
```

In [2]:
```python
warnings.filterwarnings("ignore")
```

Load Data

In [3]:
```python
df=pd.read_csv("/content/insurance.csv")
```

In [4]:
```python
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

Data Visualization

Unique Values

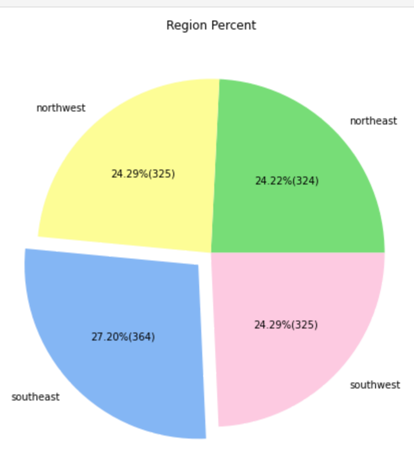
In [5]:
```python
{col:list(df[col].unique()) for col in df.select_dtypes("object")}
```

Out[5]:
```
{'region': ['southwest', 'southeast', 'northwest', 'northeast'],
 'sex': ['female', 'male'],
 'smoker': ['yes', 'no']}
```

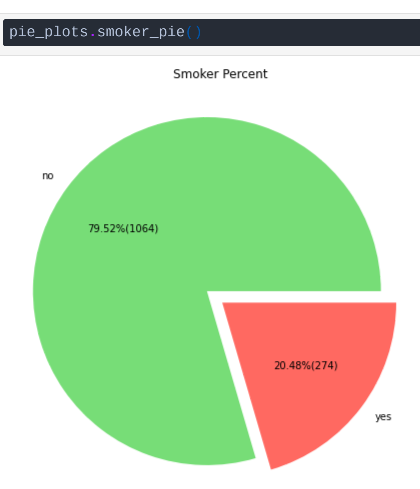
In [137…
```python
class Pie():

  def __init__(self):

    self.region=df.groupby('region').size()
    self.smoker=df.groupby('smoker').size()
    self.sex=df.groupby('sex').size()


  def region_pie(self):

    self.region.plot(kind='pie',title="Region Percent", figsize=[20,8],
                     colors=['#77dd77','#fdfd96','#84b6f4','#fdcae1'],explode=[0,0,0.1,0],
          autopct=lambda p: '{:.2f}%({:.0f})'.format(p,(p/100)*self.region.sum()))

    plt.ylabel(None)

  def smoker_pie(self):

    self.smoker.plot(kind='pie',title="Smoker Percent", figsize=[20,8],
                     colors=['#77dd77','#ff6961'],explode=[0.1,0.01],
          autopct=lambda p: '{:.2f}%({:.0f})'.format(p,(p/100)*self.region.sum()))

    plt.ylabel(None)


  def sex_pie(self):


    self.sex.plot(kind='pie',title="Sex Percent", figsize=[20,8],
                     colors=['#FFD1DC','#2271b3'],
          autopct=lambda p: '{:.2f}%({:.0f})'.format(p,(p/100)*self.sex.sum()))

    plt.ylabel(None)
```

In [138…
```python
pie_plots=Pie()
```

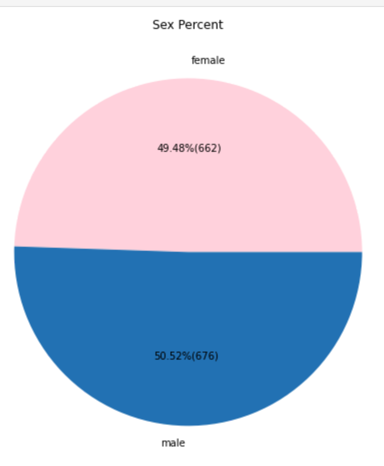
In [133…
```python
pie_plots.region_pie()
```



In [134…
```python
pie_plots.smoker_pie()
```



In [139…
```python
pie_plots.sex_pie()
```



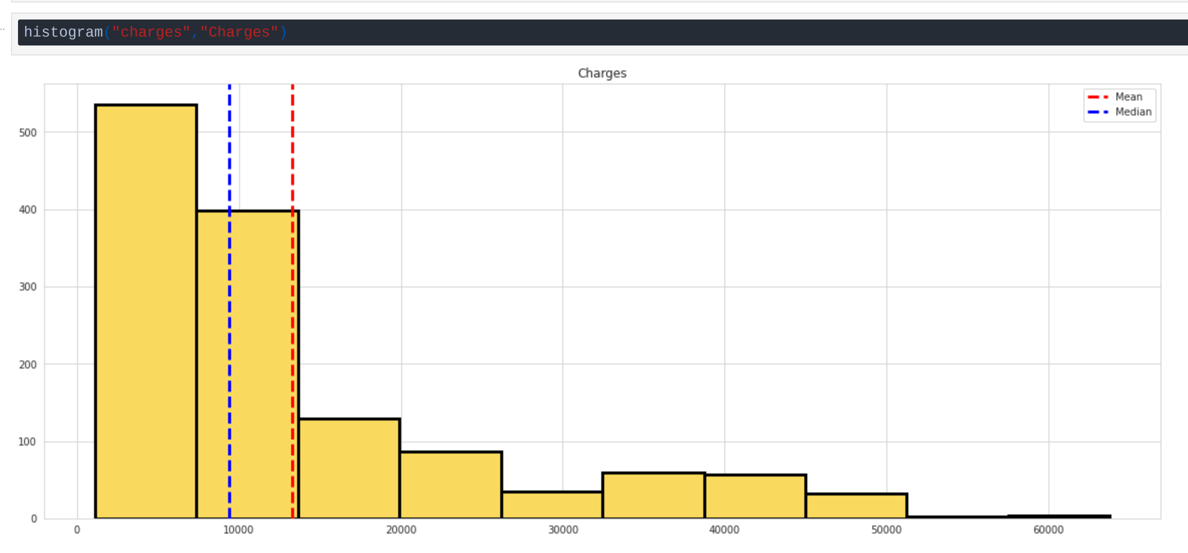
In [140…
```python
df.children.value_counts()
```

Out[140…
```
0    574
1    324
2    240
3    157
4     25
5     18
Name: children, dtype: int64
```

In [141…
```python
sns.set_style(style="whitegrid")
```

In [142…
```python
def histogram(feature,title):


  fig,ax=plt.subplots(1,1,figsize=(20,8))

  ax.set_title(title)
  ax.hist(df[feature],ec="k",color="#FADA5E",lw=3)


  ax.axvline(df[feature].mean(),
          color="red",
          linestyle="--",
          lw=3,label="Mean")


  ax.axvline(df[feature].median(),
          color="blue",
          linestyle="--",
          lw=3,label="Median")

  ax.legend()

  plt.show()
```

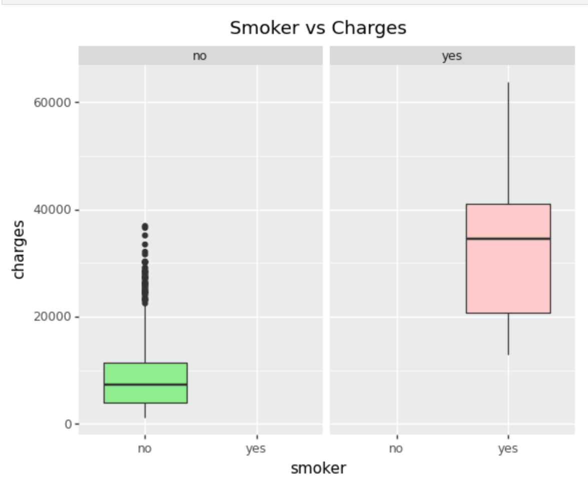
In [143…
```python
histogram("charges","Charges")
```



In [144…
```python
from plotnine import ggplot, aes, geom_point,geom_boxplot,labs,facet_wrap,scale_fill_manual,theme
```

The price of insurance is higher for people who smoke?

In [145…
```python
(

    ggplot(df)
    + aes(x="smoker",y="charges",fill="smoker")
    + geom_boxplot() + labs(title="Smoker vs Charges")
    + facet_wrap("smoker")
    + theme(legend_position="none")
    +  scale_fill_manual(values=["#90ee90","#ffcccb"])




)
```



Out[145…
```
<ggplot: (8763349732709)>
```

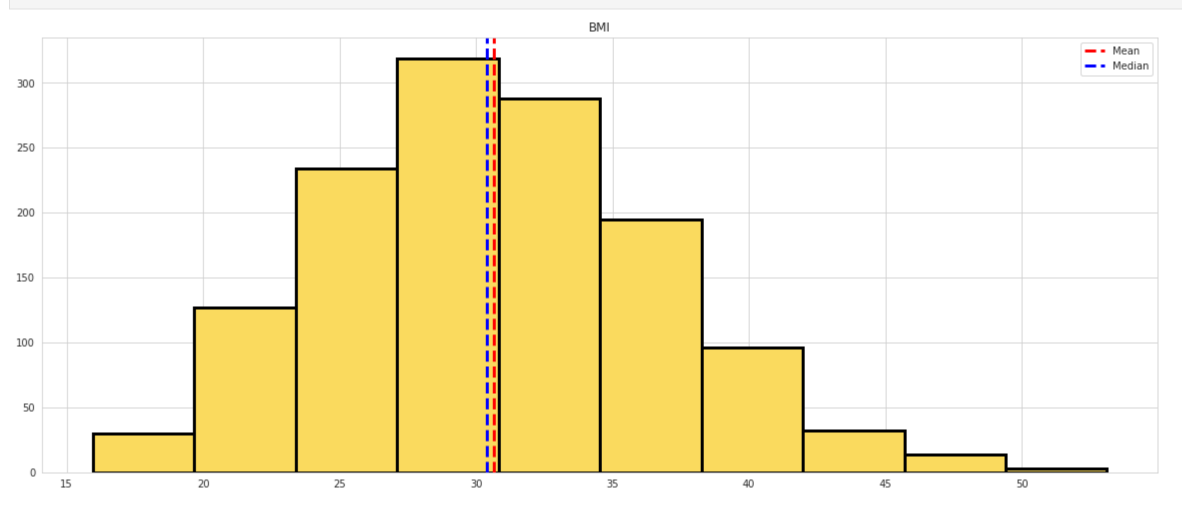
We observe a strong presence of outliers, for the category of non-smokers.

In [146…
```python
df.groupby("smoker")["charges"].mean()
```

Out[146…
```
smoker
no      8434.268298
yes    32050.231832
Name: charges, dtype: float64
```

The average price of smokers is considerably much higher than non-smokers. Since they usually have poorer physical health and need more treatments.

In [147…
```python
histogram("bmi","BMI")
```

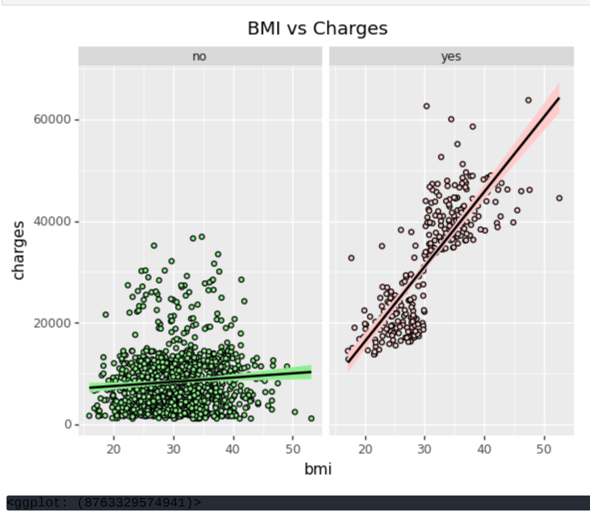


Most of the BMI data is within a normal distribution. But even so, it is possible to appreciate outlier values in the upper range.

In [148…
```python
from plotnine.facets import facet_grid

from plotnine.geoms import geom_smooth

def scatter_plots(feature,title):

  return((ggplot(df)
    + aes(x=feature,y="charges",fill="smoker",alpha=0.1)
    + geom_point() + labs(title=title,x=feature)
    + facet_wrap("smoker")
    + theme(legend_position="none")
    +  scale_fill_manual(values=["#90ee90","#ffcccb"])
    + geom_smooth(method="lm")))
```

People with a high BMI the insurance charge is higher?

In [149…
```python
scatter_plots("bmi","BMI vs Charges")
```

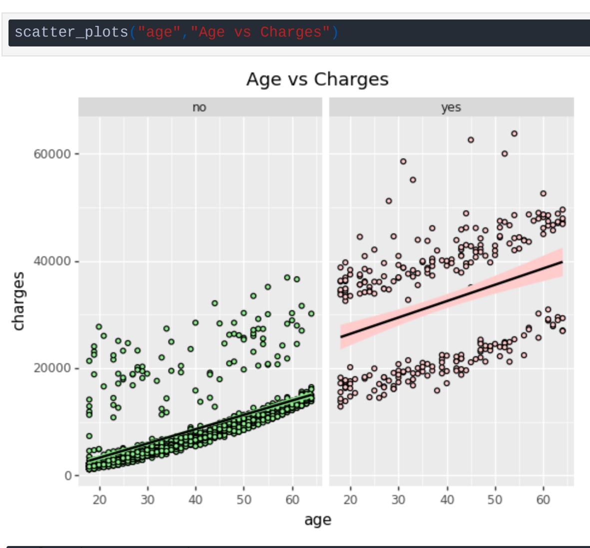


Out[149…
```
<ggplot: (8763329574941)>
```

- For non-smokers, the data trend remains constant.
- While for smokers the trend line is linear, that is, one value increases proportionally with another.

Does age influence the price of insurance?

In [150…
```python
scatter_plots("age","Age vs Charges")
```



Out[150…
```
<ggplot: (8763329477213)>
```

- For non-smokers, the trend of the data remains linear with respect to age. So we can create a linear regression model to model the data for non-smokers. But with the disadvantage that it would not explain so well for the opposite case.
- For smokers, the relationship between age seems non-existent.

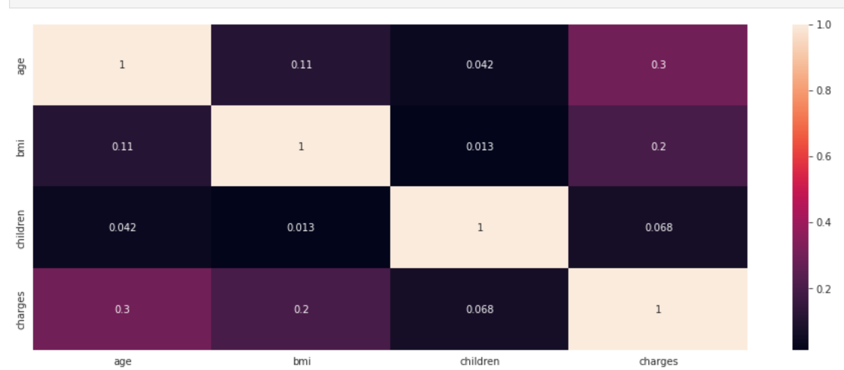
In [151…
```python
corr_pearson=df.corr(method="pearson")
corr_spearman=df.corr(method="spearman")

def correlation_matrix(corr_matrix):

  plt.figure(figsize=(16, 6))
  sns.heatmap(corr_matrix,annot=True)

  plt.show()
```

Correlation Matrix

It is used to establish possible relationships between variables

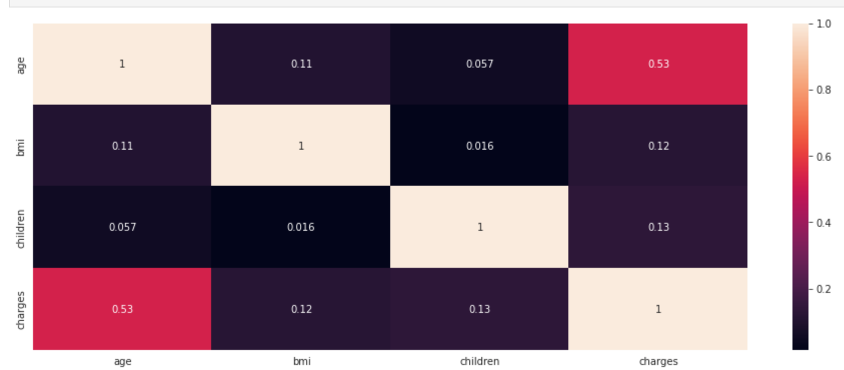
Pearson

In [152…
```python
correlation_matrix(corr_pearson)
```



Spearman

In [153…
```python
correlation_matrix(corr_spearman)
```



The correlation is measured from 0 to 1 if it is positive. There does not appear to be a strong relationship between the variable of interest. It is still too early to start ruling out variables, since these variables can complement the predictions.

Conclusion

- The **smoker** variable **positively affects** the price of health insurance.
- The **age** has a **better relationship** for the group of **non smokers.** While the BMI maintains more importance when the person smokes.
- The variables that describe the characteristics of the individual regarding their state of health. Like age, the BMI and whether the **person is a smoker increase considerably** if the individual is in poor health, while for **healthy people** the charge is **not usually high.** Especially if the person does **not smoke**, the charge increases proportionally to the number of years of the individual.