

## Importing libraries

```
In [1]: import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split, cross_val_score
from sklearn.metrics import mean_squared_error, r2_score
from sklearn.preprocessing import OneHotEncoder, PolynomialFeatures, PowerTransformer, RobustScaler
from sklearn.compose import ColumnTransformer
```

## Loading Dataset

```
In [2]: df = pd.read_csv("medical_insurance.csv")
```

## EDA

```
In [3]: df.sample(5)
```

	age	sex	bmi	children	smoker	region	charges
1211	39	male	34.10	2	no	southeast	23963.01618
1568	22	male	19.95	3	no	northeast	4005.42250
946	42	male	35.80	2	no	southwest	7160.09400
1213	52	female	33.30	2	no	southwest	10806.63900

```
Out[3]:
```

```
In [4]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2772 entries, 0 to 2771
Data columns (total 8 columns):
 #   Column      Non-Null Count  Dtype  
 ---    |  | Count | Type |
 0   age         2772 non-null   int64  
 1   sex         2772 non-null   object  
 2   bmi         2772 non-null   float64 
 3   children    2772 non-null   int64  
 4   smoker      2772 non-null   object  
 5   region      2772 non-null   object  
 6   charges     2772 non-null   float64 
 7   bmi_category 2772 non-null   object  
dtypes: float64(2), int64(4), object(3)
memory usage: 151.7+ KB
```

```
In [5]: df.describe()
```

```
Out[5]:
```

	age	bmi	children	charges
count	2772.000000	2772.000000	2772.000000	2772.000000
mean	39.109668	30.701349	1.101732	13261.369959
std	14.081459	6.129449	1.214802	12151.678945
min	18.000000	15.960000	0.000000	1121.873900
25%	26.000000	26.220000	0.000000	4687.970000
50%	39.000000	30.475000	1.000000	9333.143450
75%	51.000000	34.770000	2.000000	16577.779500
max	64.000000	53.130000	5.000000	6370.428010

```
In [6]: df.shape
```

```
Out[6]: (2772, 7)
```

```
In [7]: df.isnull().sum()
```

```
Out[7]:
```

	age	bmi	children	smoker	region	charges
count	0	0	0	0	0	0
sex	0	0	0	0	0	0
children	0	0	0	0	0	0
smoker	0	0	0	0	0	0
region	0	0	0	0	0	0
charges	0	0	0	0	0	0

```
In [8]: # Creating new column bmi_category
```

```
def categorize_bmi(bmi):
    if bmi < 18.5:
        return 'Underweight'
    elif 18.5 <= bmi < 25:
        return 'Healthy weight'
    elif 25 <= bmi < 30:
        return 'Overweight'
    else:
        return 'Obese'
```

```
df['bmi_category'] = df['bmi'].apply(categorize_bmi)
```

```
In [9]: fig, axes = plt.subplots(3, 2, figsize=(12,14))
```

```
sns.countplot(x='bmi_category', data=df, ax=axes[0, 0])
```

```
axs[0, 0].set_title('Count of BMI Categories')
```

```
average_charges_by_region = df.groupby('region')['charges'].mean().reset_index()
```

```
axs[0, 1].bar(average_charges_by_region['region'], average_charges_by_region['charges'])
```

```
axs[0, 1].set_title('Average Charges by Region')
```

```
axs[0, 1].set_xlabel('Region')
```

```
axs[0, 1].set_ylabel('Average Charges')
```

```
axs[0, 1].grid(axis='y')
```

```
sns.barplot(x='smoker', y='charges', data=df, ax=axes[1, 0])
```

```
axs[1, 0].set_title('Charges by Smoker')
```

```
axs[1, 0].set_xlabel('Smoker')
```

```
axs[1, 0].set_ylabel('Charges')
```

```
axs[1, 0].grid(axis='y')
```

```
sns.barplot(x='children', y='charges', data=df, ax=axes[1, 1])
```

```
axs[1, 1].set_title('Charges by Number of Children')
```

```
axs[1, 1].set_xlabel('Number of Children')
```

```
axs[1, 1].set_ylabel('Charges')
```

```
axs[1, 1].grid(axis='y')
```

```
sns.barplot(x='sex', y='charges', data=df, ax=axes[2, 0])
```

```
axs[2, 0].set_title('Charges by Sex')
```

```
axs[2, 0].set_xlabel('Sex')
```

```
axs[2, 0].set_ylabel('Charges')
```

```
axs[2, 0].grid(axis='y')
```

```
sns.kdeplot(x='age', y='charges', data=df, ax=axes[2, 1])
```

```
axs[2, 0].set_title('Age Distribution')
```

```
axs[2, 1].set_xlabel('Age')
```

```
axs[2, 1].set_ylabel('Count')
```

```
# Add first layout
```

```
plt.tight_layout()
```

```
# Show plots
```

```
plt.show()
```

Count of BMI Categories

Average Charges by Region

Charges by Smoker

Charges by Number of Children

Charges by Sex

Age Distribution

Count of BMI Categories

Average Charges by Region

Charges by Smoker

Charges by Number of Children

Charges by Sex

Age Distribution

```
In [10]: sns.pairplot(data=df)
```

```
plt.show()
```

Charges by Smoker

Charges by Number of Children

Charges by Sex

Age Distribution

Charges by Smoker

Charges by Number of Children

Charges by Sex

Age Distribution

Charges by Smoker

Charges by Number of Children

Charges by Sex

Age Distribution

Charges by Smoker

Charges by Number of Children

Charges by Sex

Age Distribution

Charges by Smoker

Charges by Number of Children

Charges by Sex

Age Distribution

Charges by Smoker

Charges by Number of Children

Charges by Sex

Age Distribution

Charges by Smoker

Charges by Number of Children

Charges by Sex

Age Distribution

Charges by Smoker

Charges by Number of Children

Charges by Sex

Age Distribution

Charges by Smoker

Charges by Number of Children

Charges by Sex

Age Distribution

Charges by Smoker

Charges by Number of Children

Charges by Sex

Age Distribution

Charges by Smoker

Charges by Number of Children

Charges by Sex

Age Distribution

Charges by Smoker

Charges by Number of Children

Charges by Sex

Age Distribution

Charges by Smoker

Charges by Number of Children

Charges by Sex

Age Distribution

Charges by Smoker

Charges by Number of Children

Charges by Sex

Age Distribution

Charges by Smoker

Charges by Number of Children

Charges by Sex

Age Distribution

Charges by Smoker

Charges by Number of Children

Charges by Sex

Age Distribution

Charges by Smoker

Charges by Number of Children

Charges by Sex

Age Distribution

Charges by Smoker

Charges by Number of Children

Charges by Sex

Age Distribution

Charges by Smoker

Charges by Number of Children

Charges by Sex

Age Distribution

Charges by Smoker

Charges by Number of Children

Charges by Sex

Age Distribution

Charges by Smoker

Charges by Number of Children

Charges by Sex

Age Distribution

Charges by Smoker

Charges by Number of Children

Charges by Sex

Age Distribution

Charges by Smoker

Charges by Number of Children

