Medical Insurance Cost Prediction

importing libraries

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn import metrics
```

Data collection

```
insurance_dataset = pd.read_csv('/content/insurance.csv')
```

```
insurance_dataset.head()
```

	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

Next steps: Generate code with `insurance_dataset` View recommended plots New interactive sheet

checking for missing values

```
insurance_dataset.isnull().sum()
```

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

dtype: int64

Data Analysis

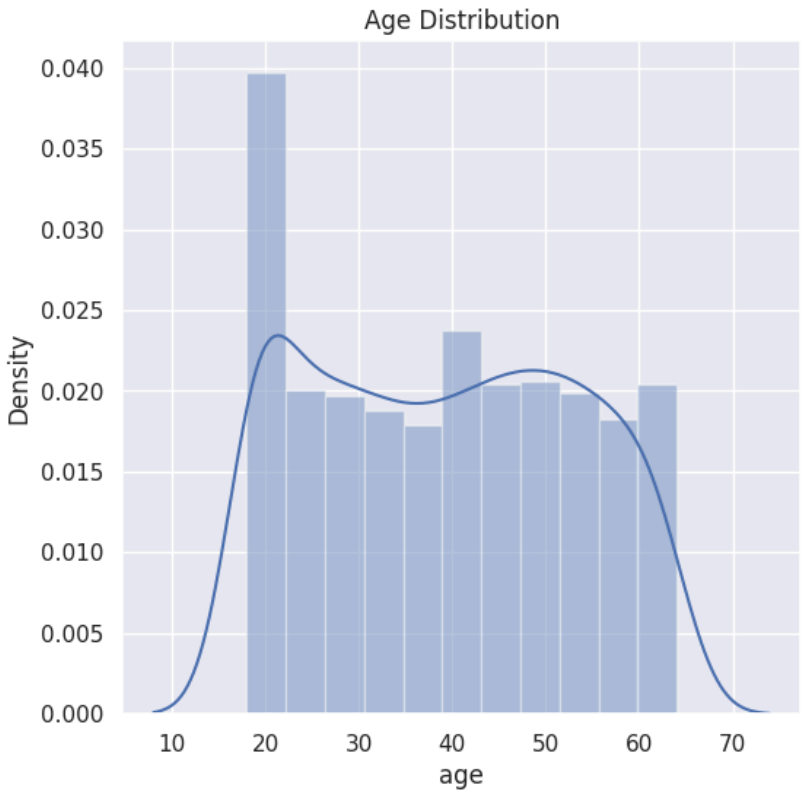
```
insurance_dataset.describe()
```

What can I help you build?

	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

```
# distribution of age value
sns.set()
plt.figure(figsize=(6,6))
sns.distplot(insurance_dataset['age'])
plt.title('Age Distribution')
plt.show()
```

```
/tmp/ipython-input-6-3634923312.py:4: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(insurance_dataset['age'])
```



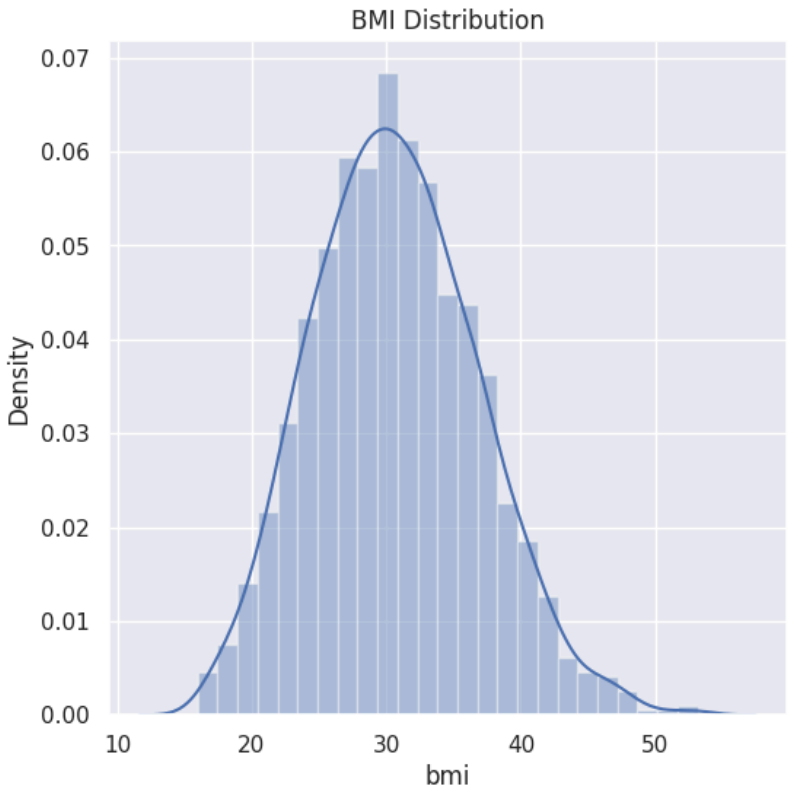

```
insurance_dataset['sex'].value_counts()
```

sex	count
male	676
female	662

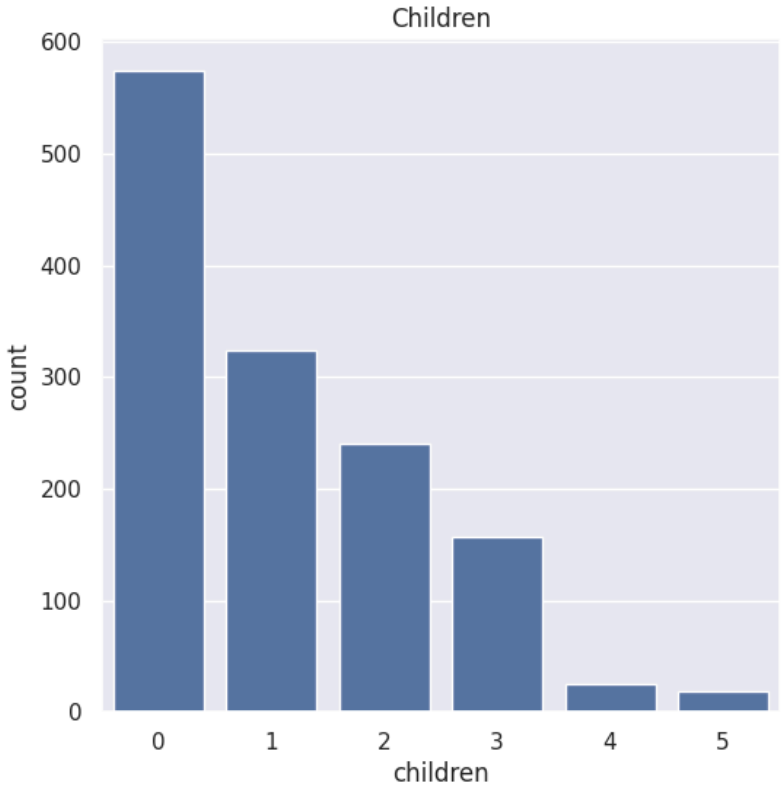
dtype: int64

```
# bmi distribution
plt.figure(figsize=(6,6))
sns.distplot(insurance_dataset['bmi']) # normal from 18.5 to 24.9
plt.title('BMI Distribution')
plt.show()
```

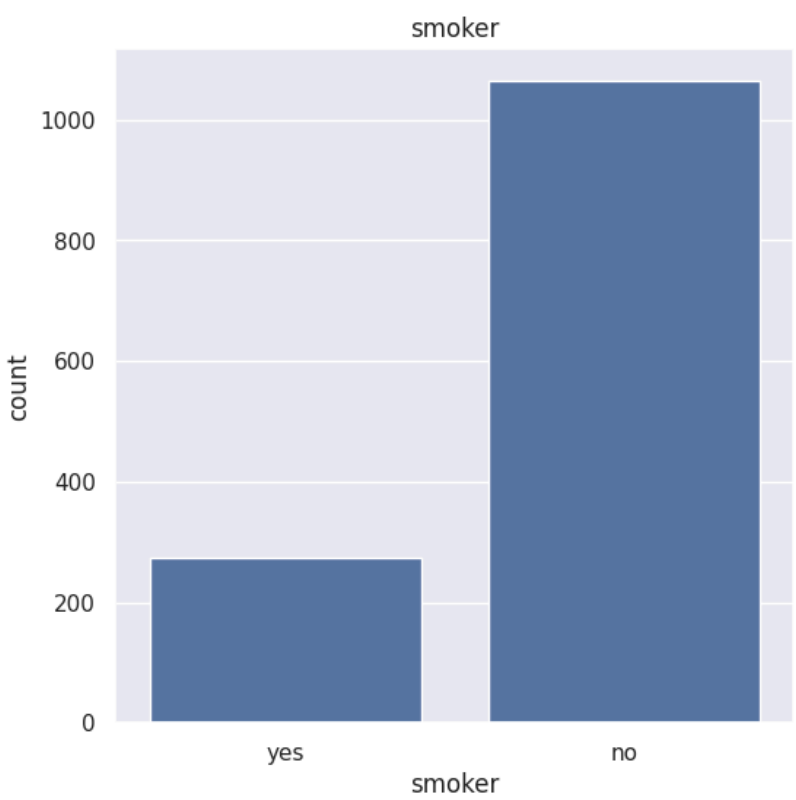
```
/tmp/ipython-input-8-1916795400.py:3: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(insurance_dataset['bmi'])
```



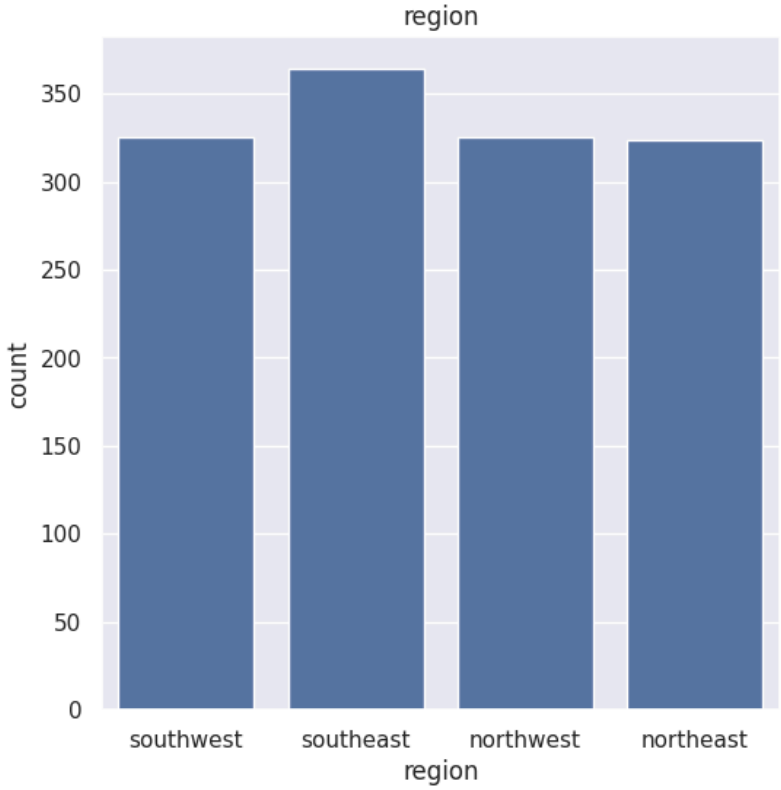

```
# children column
plt.figure(figsize=(6,6))
sns.countplot(x='children', data=insurance_dataset)
plt.title('Children')
plt.show()
```




```
# smoker column
plt.figure(figsize=(6,6))
sns.countplot(x='smoker', data=insurance_dataset)
plt.title('smoker')
plt.show()
```




```
# region column
plt.figure(figsize=(6,6))
sns.countplot(x='region', data=insurance_dataset)
plt.title('region')
plt.show()
```




```
# distribution of charges value
plt.figure(figsize=(6,6))
sns.distplot(insurance_dataset['charges'])
plt.title('Charges Distribution')
plt.show()
```

```
/tmp/ipython-input-12-3971177022.py:3: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
```

Data Preprocessing

```
# encoding sex column
insurance_dataset.replace({'sex':{'male':0,'female':1}}, inplace=True)

3 # encoding 'smoker' column
insurance_dataset.replace({'smoker':{'yes':0,'no':1}}, inplace=True)

# encoding 'region' column
insurance_dataset.replace({'region':{'southeast':0,'southwest':1,'northeast':2,'northwest':3}}, inplace=True)
```

```
/tmp/ipython-input-13-2871422651.py:2: FutureWarning: Downcasting behavior in `replace` is deprecated and wil
  insurance_dataset.replace({'sex':{'male':0,'female':1}}, inplace=True)
/tmp/ipython-input-13-2871422651.py:5: FutureWarning: Downcasting behavior in `replace` is deprecated and wil
  insurance_dataset.replace({'smoker':{'yes':0,'no':1}}, inplace=True)
/tmp/ipython-input-13-2871422651.py:8: FutureWarning: Downcasting behavior in `replace` is deprecated and wil
  insurance_dataset.replace({'region':{'southeast':0,'southwest':1,'northeast':2,'northwest':3}}, inplace=Tru
```

split data and target

```
x = insurance_dataset.drop(columns='charges', axis=1)
y = insurance_dataset['charges']
```

```
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=2)
```

Model Training(Linear Regression)

```
model = LinearRegression()
model.fit(x_train, y_train)
```

```
▾ LinearRegression
LinearRegression()
```

Model Evaluation

on train data

```
train_prediction = model.predict(x_train)
r2_train = metrics.r2_score(y_train, train_prediction)
print('R squared value = ', r2_train)
```

```
R squared value =  0.751505643411174
```

on test data

```
test_prediction = model.predict(x_test)
r2_test = metrics.r2_score(y_test, test_prediction)
print('R squared value = ', r2_test)
```

```
R squared value =  0.7447273869684076
```

Building a Predictive system

```
input_data =(46,1,33.44,1,1,0)
input_data_as_numpy_array = np.asarray(input_data)
input_data_reshaped = input_data_as_numpy_array.reshape(1,-1)
prediction = model.predict(input_data_reshaped)
print('The insurance cost is USD ',prediction[0])
```

```
The insurance cost is USD  10657.408849921467
```