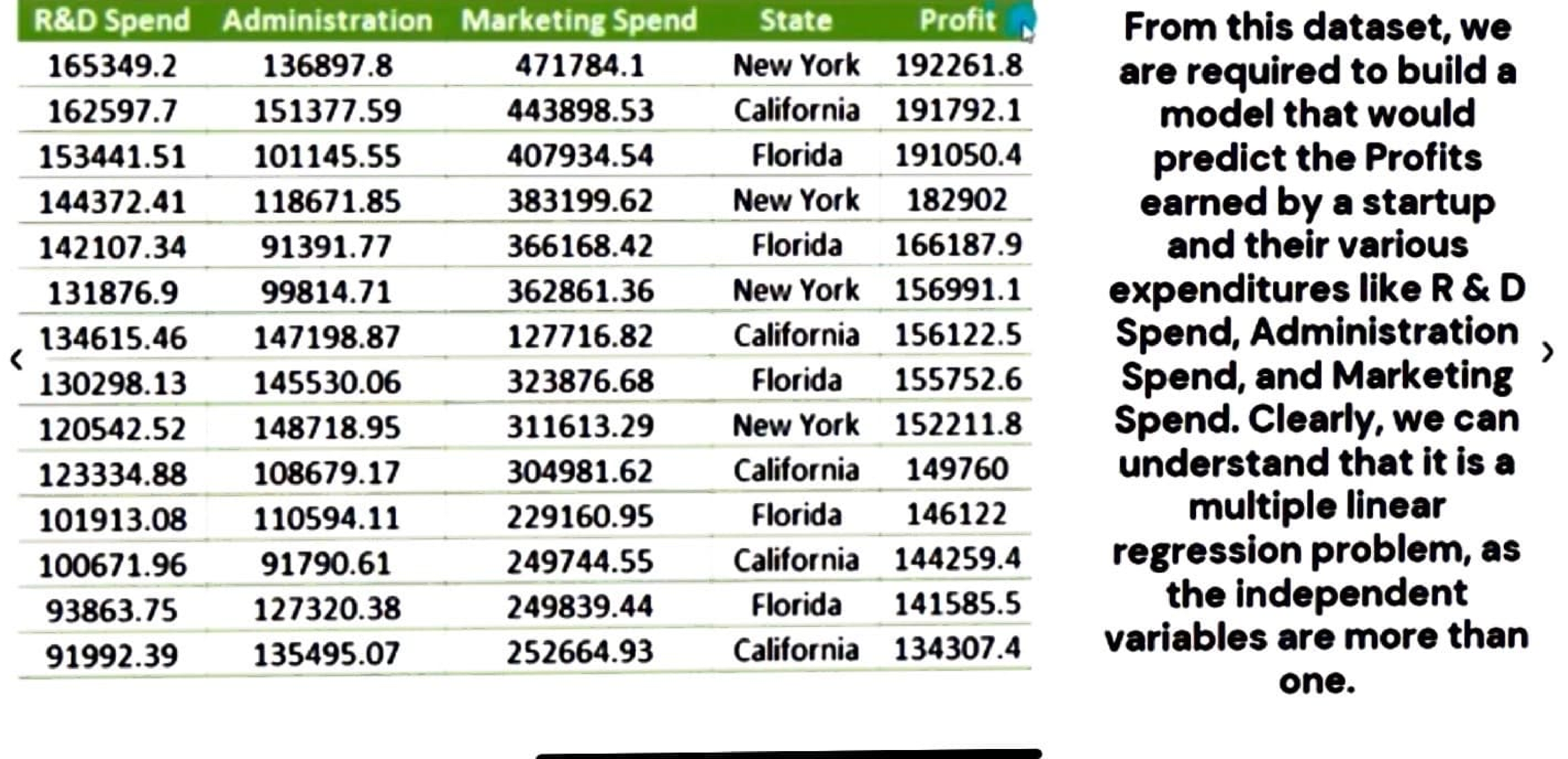
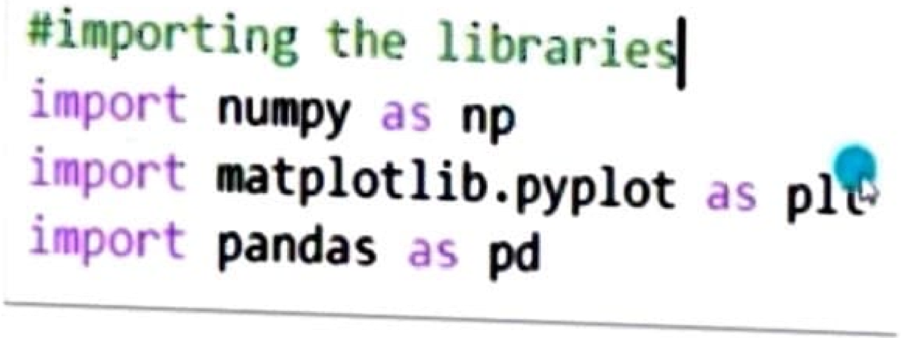


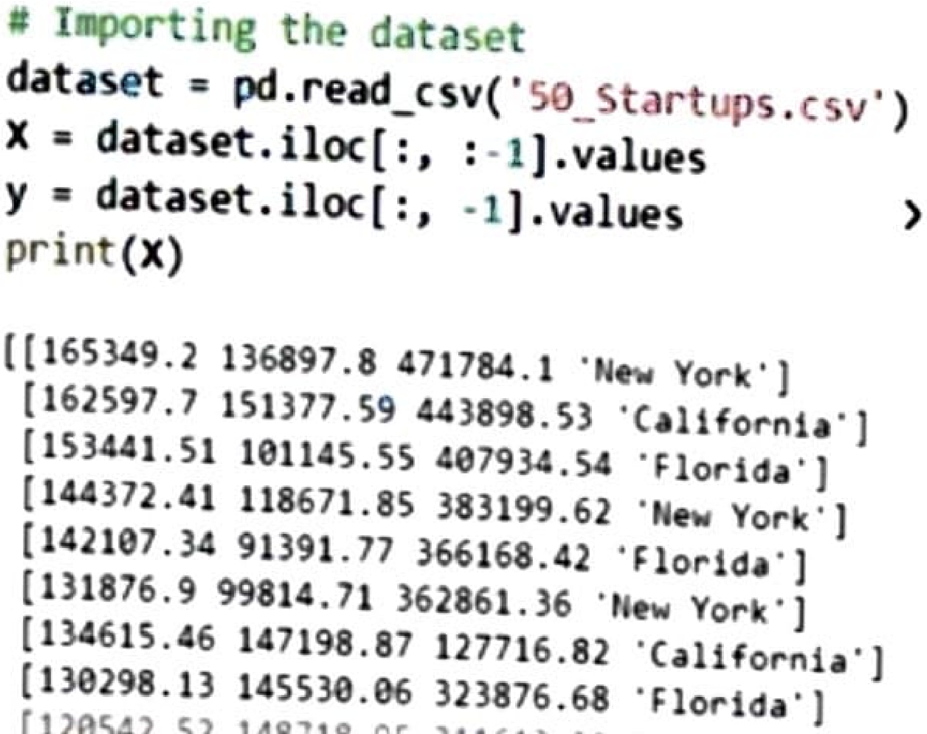
R&D Spend	Administration	Marketing Spend	State	Profit
165349.2	136897.8	471784.1	New York	192261.8
162597.7	151377.59	443898.53	California	191792.1
153441.51	101145.55	407934.54	Florida	191050.4
144372.41	118671.85	383199.62	New York	182902
142107.34	91391.77	366168.42	Florida	166187.9
131876.9	99814.71	362861.36	New York	156991.1
134615.46	147198.87	127716.82	California	156122.5
130298.13	145530.06	323876.68	Florida	155752.6
120542.52	148718.95	311613.29	New York	152211.8
123334.88	108679.17	304981.62	California	149760
101913.08	110594.11	229160.95	Florida	146122
100671.96	91790.61	249744.55	California	144259.4
93863.75	127320.38	249839.44	Florida	141585.5
91992.39	135495.07	252664.93	California	134307.4

From this dataset, we are required to build a model that would predict the Profits earned by a startup and their various expenditures like R & D Spend, Administration Spend, and Marketing Spend. Clearly, we can understand that it is a multiple linear regression problem, as the independent variables are more than one.

1 Data Preprocessing

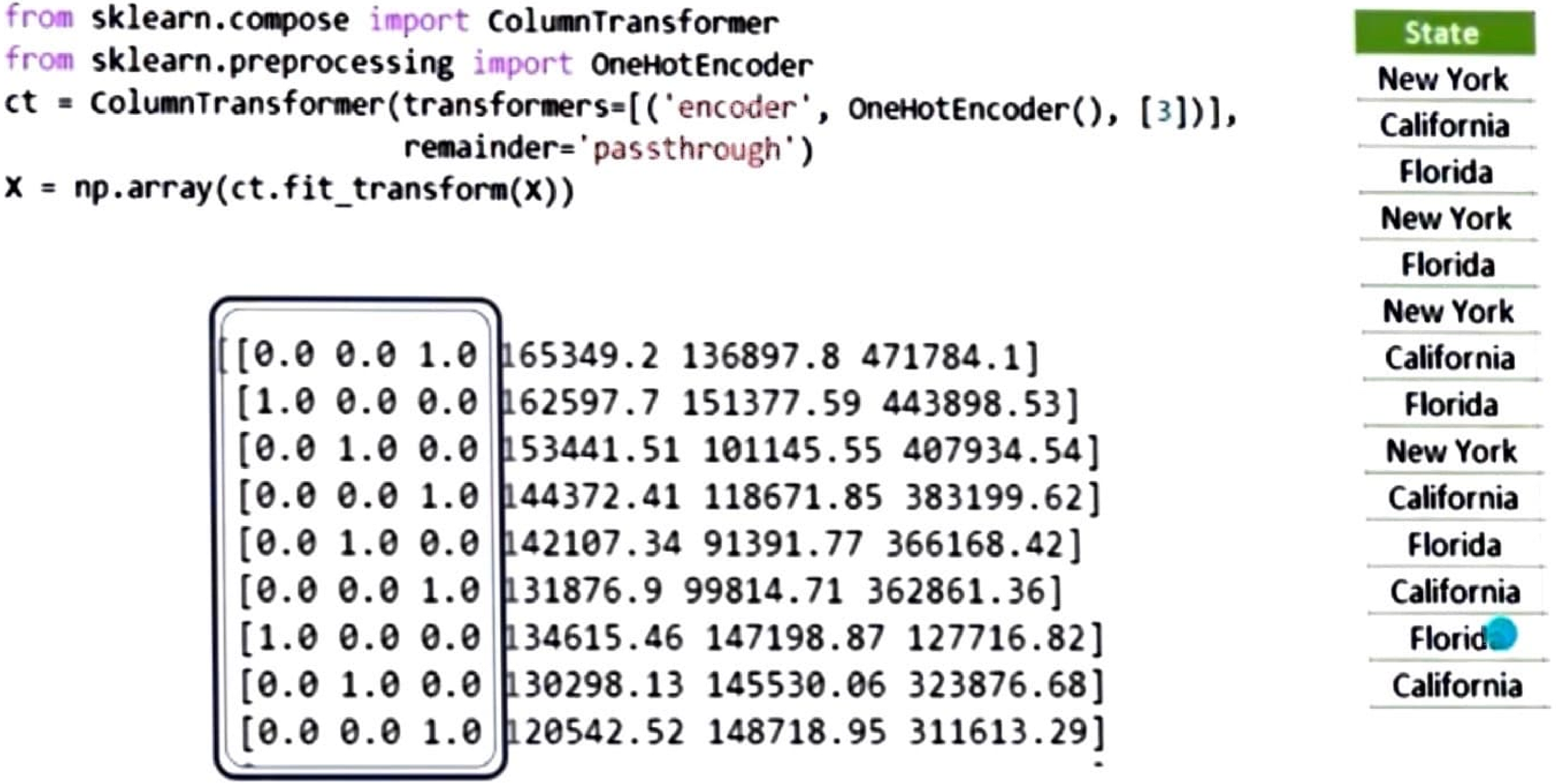


```
#importing the libraries
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
```



```
# Importing the dataset
dataset = pd.read_csv('50_Startups.csv')
X = dataset.iloc[:, :-1].values
y = dataset.iloc[:, -1].values
print(X)
```

```
[[165349.2 136897.8 471784.1 'New York']
 [162597.7 151377.59 443898.53 'California']
 [153441.51 101145.55 407934.54 'Florida']
 [144372.41 118671.85 383199.62 'New York']
 [142107.34 91391.77 366168.42 'Florida']
 [131876.9 99814.71 362861.36 'New York']
 [134615.46 147198.87 127716.82 'California']
 [130298.13 145530.06 323876.68 'Florida']
 [120542.52 148718.95
```



```python
from sklearn.compose import ColumnTransformer
from sklearn.preprocessing import OneHotEncoder
ct = ColumnTransformer(transformers=[('encoder', OneHotEncoder(), [3])],
                       remainder='passthrough')
X = np.array(ct.fit_transform(X))
```

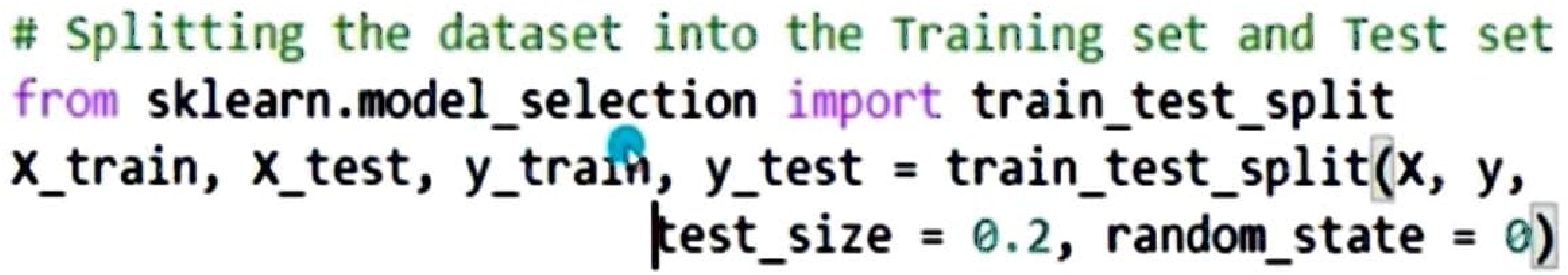
```
[[0.0 0.0 1.0 165349.2 136897.8 471784.1]
 [1.0 0.0 0.0 162597.7 151377.59 443898.53]
 [0.0 1.0 0.0 153441.51 101145.55 407934.54]
 [0.0 0.0 1.0 144372.41 118671.85 383199.62]
 [0.0 1.0 0.0 142107.34 91391.77 366168.42]
 [0.0 0.0 1.0 131876.9 99814.71 362861.36]
 [1.0 0.0 0.0 134615.46 147198.87 127716.82]
 [0.0 1.0 0.0 130298.13 145530.06 323876.68]
 [0.0 0.0 1.0 120542.52 148718.95 311613.29]
```

State
New York
California
Florida
New York
Florida
New York
California
Florida
New York
California
Florida
California
Florida
California

```python
#Avoiding Dummy Variable Trap
X = X[: , 1:]
print(X)
```

```
[[0.0 1.0 165349.2 136897.8 471784.1]
 [0.0 0.0 162597.7 151377.59 443898.53]
 [1.0 0.0 153441.51 101145.55 407934.54]
 [0.0 1.0 144372.41 118671.85 383199.62]
 [1.0 0.0 142107.34 91391.77 366168.42]
 [0.0 1.0 131876.9 99814.71 362861.36]
 [0.0 0.0 134615.46 147198.87 127716.82]
 [1.0 0.0 130298.13 145530.06 323876.68]
 [0.0 1.0 120542.52 148718.95 311613.29]
 [0.0 0.0 123334.88 108679.17 304981.62]
 [1.0 0.0 101913.08 110594.11 229160.95]
 [0.0 0.0 100671.96 91790.61 249744.55]
```

State
New York
California
Florida
New York
Florida
New York
California
Florida
New York
California
Florida
California
Florida
California



```
# Splitting the dataset into the Training set and Test set
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y,
                        test_size = 0.2, random_state = 0)
```

Let's Code

Fitting the model to the training data

```
from sklearn.linear_model import LinearRegression
regressor = LinearRegression()
regressor.fit(X_train, y_train)
```

```
LinearRegression(copy_X=True, fit_intercept=True, n_jobs=None, normalize=False)
```

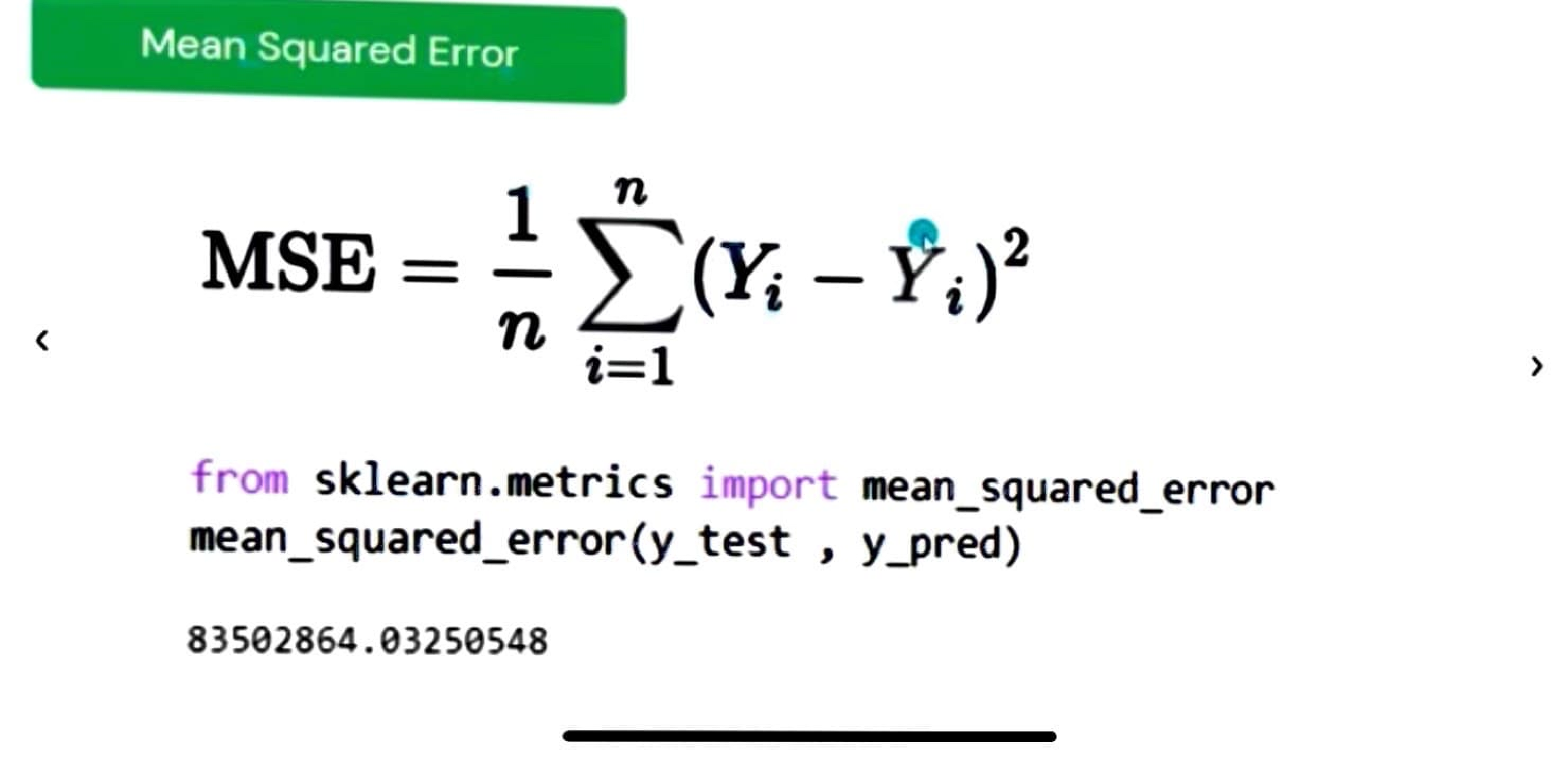
```
# Predicting the Test set results
y_pred = regressor.predict(X_test)
```

```
y_test
```

```
array([103282.38, 144259.4 , 146121.95,  77798.83, 191050.39, 105008.31,
        81229.06,  97483.56, 110352.25, 166187.94])
```

```
y_pred
```

```
array([103015.20159796, 132582.27760816, 132447.73845175,  71976.09851259,
       178537.48221054, 116161.24230163,  67851.69209676,  98791.73374688,
       113969.43533012, 167921.0656955 ])
```



Mean Squared Error

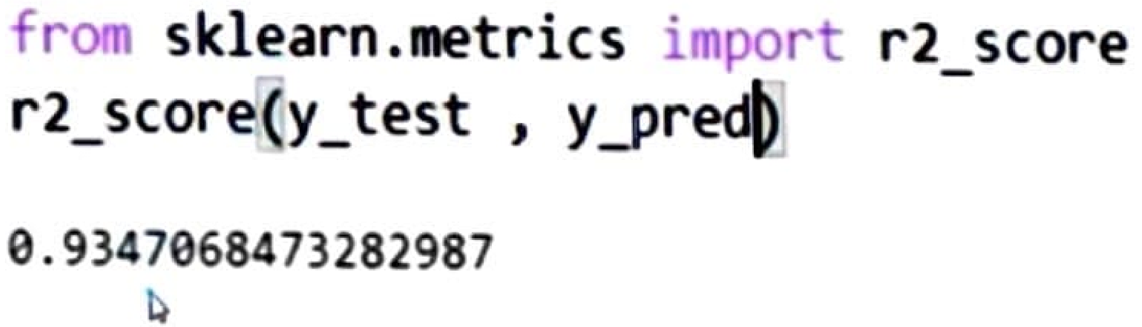
$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

```
from sklearn.metrics import mean_squared_error
mean_squared_error(y_test , y_pred)
```

```
83502864.03250548
```

R2 Score

$$R^2 = 1 - \frac{RSS}{TSS}$$



```
from sklearn.metrics import r2_score
r2_score(y_test , y_pred)
```

```
0.9347068473282987
```