untitled

November 19, 2024

[2]:

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score

# Load Dataset
data = pd.read_csv("labor_data.csv")
```

[4]:

```
data.sample(15)
```

[4]:

```
       Age          Eduacation      Race         Hisp MaritalStatus  Nodeg  \
4144    21  LessThanHighSchool  NotBlack  NotHispanic    NotMarried      1
14804   47        Intermediate  NotBlack  NotHispanic       Married      0
3028    27          HighSchool  NotBlack  NotHispanic       Married      0
2453    17  LessThanHighSchool  NotBlack  NotHispanic    NotMarried      1
10624   32          HighSchool  NotBlack  NotHispanic       Married      0
15372   19  LessThanHighSchool  NotBlack  NotHispanic       Married      1
4459    21        Intermediate  NotBlack  NotHispanic    NotMarried      0
1413    51  LessThanHighSchool  NotBlack  NotHispanic       Married      1
3272    31  LessThanHighSchool  NotBlack     hispanic       Married      1
9498    22        Intermediate  NotBlack  NotHispanic       Married      0
13084   33  LessThanHighSchool  NotBlack  NotHispanic       Married      1
12348   27  LessThanHighSchool  NotBlack  NotHispanic    NotMarried      1
13155   19        Intermediate  NotBlack  NotHispanic    NotMarried      0
4297    49  LessThanHighSchool  NotBlack  NotHispanic       Married      1
13312   20        Intermediate  NotBlack  NotHispanic    NotMarried      0

       Earnings_1974  Earnings_1975  Earnings_1978
4144       5066.6640       7118.323    11526.27000
14804     25862.3200      25243.550    25564.67000
3028      19700.4300      20860.840        0.00000
2453        834.6477       1693.645     7772.84200
10624       918.8961          0.000       11.82181
15372      3252.3830       3360.435      995.98770
4459      13195.6600      11452.690    21431.47000
1413      11996.5900       9855.726        0.00000
3272       9827.6820      10559.320    18552.86000
```

```
9498       7311.9840      12149.130     6278.86000
13084     25446.9600      22490.030        0.00000
12348      8250.4730       8357.226     8425.99700
13155      1171.6420       5005.742     8167.39500
4297      22333.6800      19706.080    25564.67000
13312      5672.0770       7983.048    16433.80000
```

[8]:

```
print(data['Eduacation'].unique())
```

```
['LessThanHighSchool' 'Intermediate' 'HighSchool' 'graduate'
 'PostGraduate']
```

[10]:

```
# Debug: Inspect columns
print("Columns before encoding:", data.columns)

# Preprocessing: One-hot encoding
if 'Race' in data.columns:
    data = pd.get_dummies(data, columns=['Race'], drop_first=True)
if 'Hisp' in data.columns:
    data = pd.get_dummies(data, columns=['Hisp'], drop_first=True)
if 'MaritalStatus' in data.columns:
    data = pd.get_dummies(data, columns=['MaritalStatus'], drop_first=True)
# Map education levels to numeric values
education_mapping = {
    'LessThanHighSchool': 0,
    'Intermediate': 1,
    'HighSchool': 2,
    'graduate': 3,
    'PostGraduate': 4
}

data['Eduacation'] = data['Eduacation'].map(education_mapping)

# Debug: Check the mapping
print(data['Eduacation'].unique())
```

```
Columns before encoding: Index(['Age', 'Eduacation', 'Race', 'Hisp',
'MaritalStatus', 'Nodeg',
       'Earnings_1974', 'Earnings_1975', 'Earnings_1978'],
      dtype='object')
[0 1 2 3 4]
```

[12]:

```
# Debug: Check resulting columns
print(data.columns)

# Debug: Check columns after encoding
print("Columns after encoding:", data.columns)
```

```
Index(['Age', 'Eduacation', 'Nodeg', 'Earnings_1974', 'Earnings_1975',
       'Earnings_1978', 'Race_black', 'Hisp_hispanic',
       'MaritalStatus_NotMarried'],
      dtype='object')
Columns after encoding: Index(['Age', 'Eduacation', 'Nodeg', 'Earnings_1974',
'Earnings_1975',
       'Earnings_1978', 'Race_black', 'Hisp_hispanic',
       'MaritalStatus_NotMarried'],
      dtype='object')
```

[18]:
```
data.sample(15)
```

[18]:
```
       Age  Eduacation  Nodeg  Earnings_1974  Earnings_1975  Earnings_1978  \
10982   29           1      0     25862.3200      21602.030     25564.6700
8356    39           2      0     25862.3200      21433.740     22826.4400
4205    49           2      0     25862.3200      25243.550     25564.6700
2667    16           0      1       920.8554          0.000     15997.8700
3290    21           2      0     12300.2800      10190.520     15003.3600
13283   33           2      0      4915.8000       7655.419       789.1060
297     31           2      0     20572.3000      18916.550     15102.3700
7565    34           0      1         0.0000       9680.274     14093.0800
4571    29           2      0     25862.3200      25243.550     14908.7800
4429    36           0      1     18848.1500      18502.980     18041.5600
14767   45           2      0     25862.3200      24799.550     22281.1600
8365    24           0      1         0.0000          0.000       147.7727
306     32           1      0     25862.3200      25243.550     25564.6700
2811    34           0      1     11998.5500      12165.240     25564.6700
7571    34           0      1     17547.1900      17643.630     13748.7700

       Race_black  Hisp_hispanic  MaritalStatus_NotMarried
10982       False          False                     False
8356         True          False                     False
4205        False          False                      True
2667        False          False                     False
3290        False          False                     False
13283       False          False                      True
297         False          False                     False
7565        False          False                     False
4571         True          False                      True
4429        False          False                     False
14767       False          False                     False
8365        False          False                      True
306         False          False                     False
2811        False          False                     False
7571        False           True                     False
```

[20]:

```
# Define Features (X) and Target (y)
X = data[['Age', 'Eduacation', 'Race_black', 'Hisp_hispanic',␣
  ↪'MaritalStatus_NotMarried']]
y_1974 = data['Earnings_1974']
y_1975 = data['Earnings_1975']
```

[24]:

```
# Combine earnings for prediction
data['Avg_Earnings'] = (y_1974 + y_1975) / 2
y = data['Avg_Earnings']
```

[26]:

```
# Train-Test Split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,␣
  ↪random_state=42)
```

[28]:

```
# Train Linear Regression Model
model = LinearRegression()
model.fit(X_train, y_train)

# Make Predictions
y_pred = model.predict(X_test)
```

[ ]:

```
```