```
# Import necessary libraries
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.preprocessing import OneHotEncoder
from sklearn.metrics import mean_squared_error
import numpy as np



# Load the dataset
data = pd.read_csv('TheLaborProblem.csv')



# Check for any missing values
print(data.isnull().sum())
```

```
Age             0
Eduacation      0
Race            0
Hisp            0
MaritalStatus   0
Nodeg           0
Earnings_1974   0
Earnings_1975   0
Earnings_1978   0
dtype: int64
```

```
# Define categorical columns
categorical_cols = ['Eduacation', 'Race', 'Hisp', 'MaritalStatus']


# Apply OneHotEncoder
data = pd.get_dummies(data, columns=categorical_cols, drop_first=True)


# Define feature columns (X) and target column (y)
X = data.drop(['Earnings_1978'], axis=1)
y = data['Earnings_1978']


# Split the data into training and test sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)



# Initialize and fit the multiple linear regression model
model = LinearRegression()
model.fit(X_train, y_train)
```

```
LinearRegression()
```

```
# Predict on the test set
y_pred = model.predict(X_test)


# Evaluate the model using Mean Squared Error
mse = mean_squared_error(y_test, y_pred)
rmse = np.sqrt(mse)
print(f'Root Mean Squared Error: {rmse}')
```

```
Root Mean Squared Error: 6973.217621811093
```

```
# Print coefficients to understand the impact of each feature
print("Coefficients: ", model.coef_)
print("Intercept: ", model.intercept_)
```

```
Coefficients:  [-1.07350492e+02 -2.93390186e+01  2.83039030e-01  4.75134632e-01
  2.74149983e+02 -2.93390186e+01  1.05909511e+03  1.03313773e+03
 -8.57181352e+02 -4.42428781e+02 -8.63504844e+01]
Intercept:  7918.751772334645
```