Predicting Labor Earnings for 1978 using Linear Regression

Problem Description

We aim to predict labor earnings for 1978 using linear regression. The data contains features like:

- Age of the individual.
- Race (Black/Not Black).
- Education level.
- Hispanic status.
- Marital status.

The goal is to fit a linear regression model to predict earnings based on these features.

```python
# Importing necessary libraries
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score
import matplotlib.pyplot as plt
import seaborn as sns
```

Step 1: Load the Data

```python
# Replace 'labour_data.csv' with the actual file path
# Example data loading step - ensure your CSV file is formatted
correctly
# Example: data = pd.read_csv('path_to_file.csv')
data = pd.read_csv('labour_data.csv')  # Update this line with your
actual data file
data.head()  # Display first few rows
```

Step 2: Preprocess the Data

```python
# Encoding categorical variables
data = pd.get_dummies(data, columns=['Race', 'Hispanic', 'Married'],
drop_first=True)
```

```python
# Define features and target variable
X = data[['Age', 'Education', 'Race_not_black', 'Hispanic_Yes',
'Married_Yes']]
y = data['Earnings_1978']
```

Step 3: Split the Data into Training and Test Sets

```python
# Train-test split
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.3, random_state=42)
```

Step 4: Train the Linear Regression Model

```python
# Train the model
model = LinearRegression()
model.fit(X_train, y_train)
```

Step 5: Evaluate the Model

```python
# Make predictions
y_pred = model.predict(X_test)

# Evaluation metrics
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

print(f"Mean Squared Error: {mse}")
print(f"R² Score: {r2}")
```

Step 6: Analyze Residuals

```python
# Residual analysis
residuals = y_test - y_pred

sns.histplot(residuals, kde=True)
plt.title("Residual Distribution")
plt.xlabel("Residuals")
plt.ylabel("Frequency")
plt.show()
```