The Effect of Education on Wages in the United States – A Data Analysis using Python

Introduction

Education is a critical determinant of one's professional journey, with an influence on employability, skills development, and ultimately, personal income. The nexus between education and salary has been the subject of considerable research, given the interest among policymakers, educators, and economists in learning how school attainment affects income inequalities and economic progress. Increased levels of education tend to be linked to improved employment prospects and greater earning potential, yet the magnitude and uniformity of this impact upon various groups have been of continued interest.

To understand the effect of education on wages is crucial to inform investment in education, designing labor market policy, and income inequality. Quantifying the interaction between education level and wages is the purpose of this analysis in order to identify whether increased levels of education correspond to higher earning increases and to explore how it differs among different groups.

The central research question guiding this study is:

"How does education influence wages?"

By answering this question, the research aims to shed light on how much education is responsible for wage disparities and whether other variables, including work experience, industry, and demographic factors, are important.

This research utilizes a data-driven method to examine the effect of education on earnings with a linear regression model. The data analyzed is from a dataset of wage data separated by race, gender, and education level in the United States for the period of 1973 to 2022.

The key steps in the methodology include:

1. **Data Selection:** Choosing a relevant dataset that captures wage trends over time for different education levels.
2. **Data Exploration & Visualization:** Examining summary statistics and visualizing wage distributions to identify patterns.
3. **Regression Analysis:** Applying a linear regression model to quantify the relationship between education and wages, incorporating additional variables such as work experience and industry when relevant.
4. **Interpretation of Results:** Evaluating the significance of the findings and discussing potential implications.

The findings of this study will contribute to a deeper understanding of how education affects wages, providing valuable insights for individuals making career decisions, policymakers shaping education and labor policies, and researchers exploring economic trends.

Data Selection and Justification

For the purposes of this study, the dataset "Wages by Education in the USA (1973-2022)" has been chosen. The dataset offers complete information on wages categorized by level of education, gender, and race and is very well suited to examining the interaction between education and wages over time.

The dataset is used from credible labor and economic research organizations to guarantee its reliability and credibility. The data are spread across decades, giving a long-time trend of wages by different levels of education. The demographic breakdowns make the analysis richer, such that the analysis can be conducted on the variations within different population segments.

This dataset is appropriate for answering the research question for several reasons:

- **Comprehensive Coverage:** The dataset spans nearly five decades, providing a robust historical view of wage trends.
- **Segmentation by Education Level:** Data is categorized into different education levels, enabling precise analysis of how wages vary with educational attainment.
- **Inclusion of Demographics:** By incorporating gender and race breakdowns, the dataset allows for a nuanced understanding of wage disparities.
- **Relevance to Economic Research:** The dataset has been widely used in labor economics studies, reinforcing its suitability for this research.

Before proceeding with analysis, the dataset underwent several data preprocessing steps:

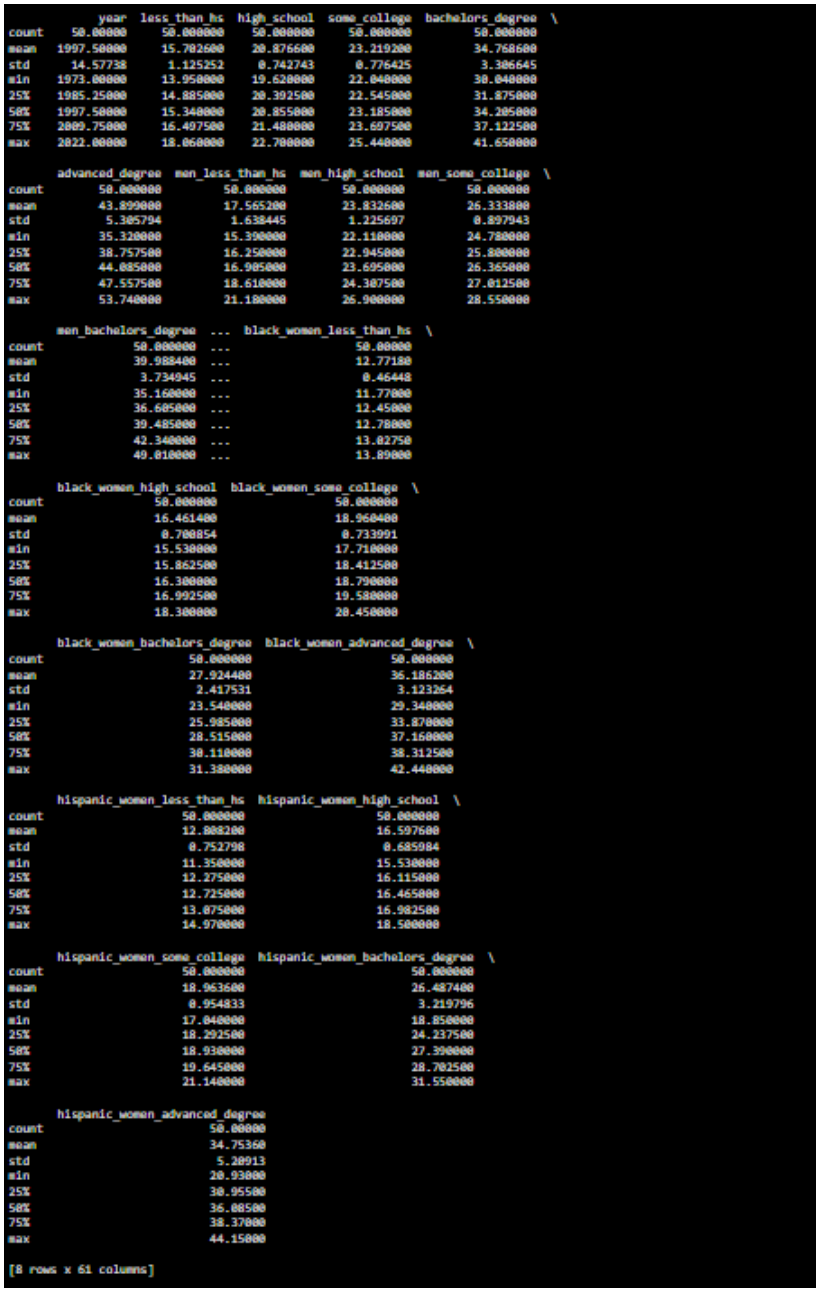
1. **Handling Missing Values:** Any missing values were addressed using appropriate imputation techniques or by removing incomplete records where necessary.
2. **Data Cleaning:** Column names were standardized for consistency, and outliers were reviewed for potential data entry errors.
3. **Transformations:** Certain categorical variables (such as education level) were encoded into numerical values to facilitate regression analysis.
4. **Feature Selection:** Only the most relevant columns (education level, wages, and relevant demographic attributes) were retained to ensure a focused analysis.

These data preparation steps ensure that the dataset is clean, structured, and ready for exploratory analysis and modeling in subsequent sections.

Data Exploration and Visualizations

For getting the distribution of the key variables, we used the following code:

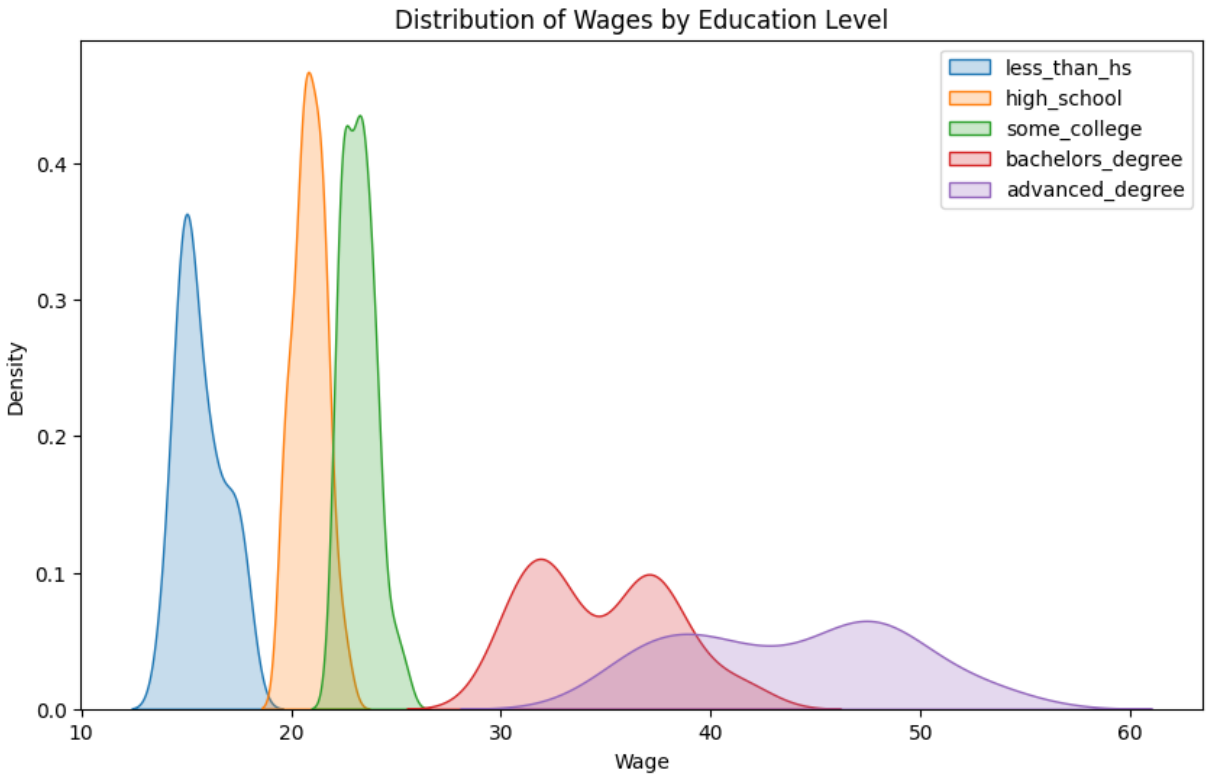
```
1. import pandas as pd
2.
3. # Load the dataset
4. file_path = "/mnt/data/wages_by_education.csv"
5. df = pd.read_csv(file_path)
6.
7. # Display summary statistics
8. print(df.describe())
```



```
              year  less_than_hs  high_school  some_college  bachelors_degree  \
count    50.00000     50.000000    50.000000     50.000000         50.000000
mean   1997.58000     15.702600    20.876600     23.219200         34.768600
std      14.57738      1.125252     0.742743      0.776425          3.306645
min    1973.00000     13.950000    19.620000     22.040000         30.040000
25%    1985.25000     14.885000    20.392500     22.545000         31.875000
50%    1997.58000     15.340000    20.855000     23.185000         34.205000
75%    2009.75000     16.497500    21.480000     23.697500         37.122500
max    2022.00000     18.060000    22.700000     25.440000         41.650000

       advanced_degree  men_less_than_hs  men_high_school  men_some_college  \
count        50.000000         50.000000        50.000000         50.000000
mean         43.899000         17.565200        23.832600         26.333800
std           5.305794          1.638445         1.225697          0.897943
min          35.320000         15.390000        22.110000         24.780000
25%          38.757500         16.250000        22.945000         25.800000
50%          44.085000         16.905000        23.695000         26.365000
75%          47.557500         18.610000        24.307500         27.012500
max          53.740000         21.180000        26.900000         28.550000

       men_bachelors_degree  ...  black_women_less_than_hs  \
count             50.000000  ...                 50.00000
mean              39.988400  ...                 12.77180
std                3.734945  ...                  0.46448
min               35.160000  ...                 11.77000
25%               36.685000  ...                 12.45000
50%               39.485000  ...                 12.78000
75%               42.340000  ...                 13.02750
max               49.010000  ...                 13.89000

       black_women_high_school  black_women_some_college  \
count                50.000000                 50.000000
mean                 16.461400                 18.960400
std                   0.700854                  0.733991
min                  15.530000                 17.710000
25%                  15.862500                 18.412500
50%                  16.300000                 18.790000
75%                  16.992500                 19.588000
max                  18.300000                 20.450000

       black_women_bachelors_degree  black_women_advanced_degree  \
count                     50.000000                    50.000000
mean                      27.924400                    36.186200
std                        2.417531                     3.123264
min                       23.540000                    29.340000
25%                       25.985000                    33.870000
50%                       28.515000                    37.160000
75%                       30.110000                    38.312500
max                       31.380000                    42.440000

       hispanic_women_less_than_hs  hispanic_women_high_school  \
count                    50.000000                   50.000000
mean                     12.808200                   16.597600
std                       0.752798                    0.685084
min                      11.350000                   15.530000
25%                      12.275000                   16.115000
50%                      12.725000                   16.465000
75%                      13.075000                   16.982500
max                      14.970000                   18.500000

       hispanic_women_some_college  hispanic_women_bachelors_degree  \
count                    50.000000                        50.000000
mean                     18.963600                        26.487400
std                       0.954833                         3.219706
min                      17.040000                        18.850000
25%                      18.292500                        24.237500
50%                      18.930000                        27.390000
75%                      19.645000                        28.702500
max                      21.140000                        31.550000

       hispanic_women_advanced_degree
count                        50.00000
mean                         34.75360
std                           5.28913
min                          20.93000
25%                          30.95500
50%                          36.08500
75%                          38.37000
max                          44.15000

[8 rows x 61 columns]
```

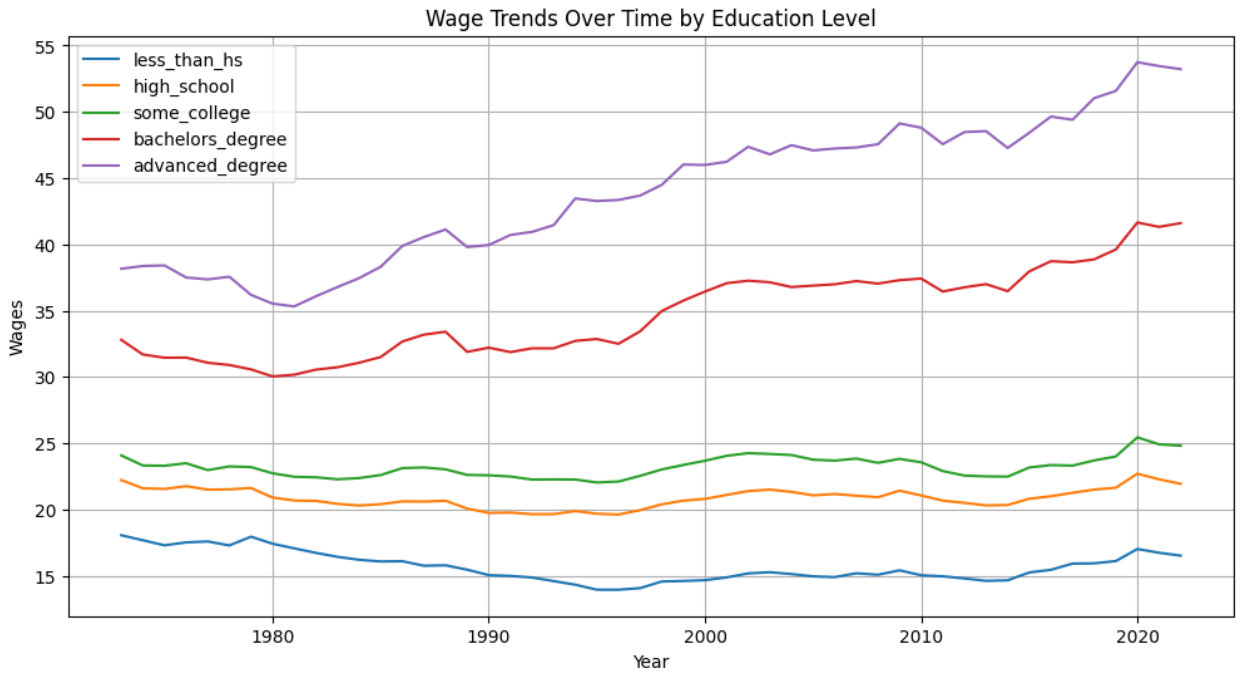
We then proceeded to create curves for all wage distributions over different qualifications

```
1.  import matplotlib.pyplot as plt
2.  import seaborn as sns
3.
4.  # Select relevant columns for wages
5.  education_levels = ['less_than_hs', 'high_school', 'some_college', 'bachelors_degree',
    'advanced_degree']
6.
7.  # Plot histogram
8.  plt.figure(figsize=(10, 6))
9.  for level in education_levels:
10.     sns.kdeplot(df[level], label=level, shade=True)
11.
12. plt.title("Distribution of Wages by Education Level")
13. plt.xlabel("Wage")
14. plt.ylabel("Density")
15. plt.legend()
16. plt.show()
```



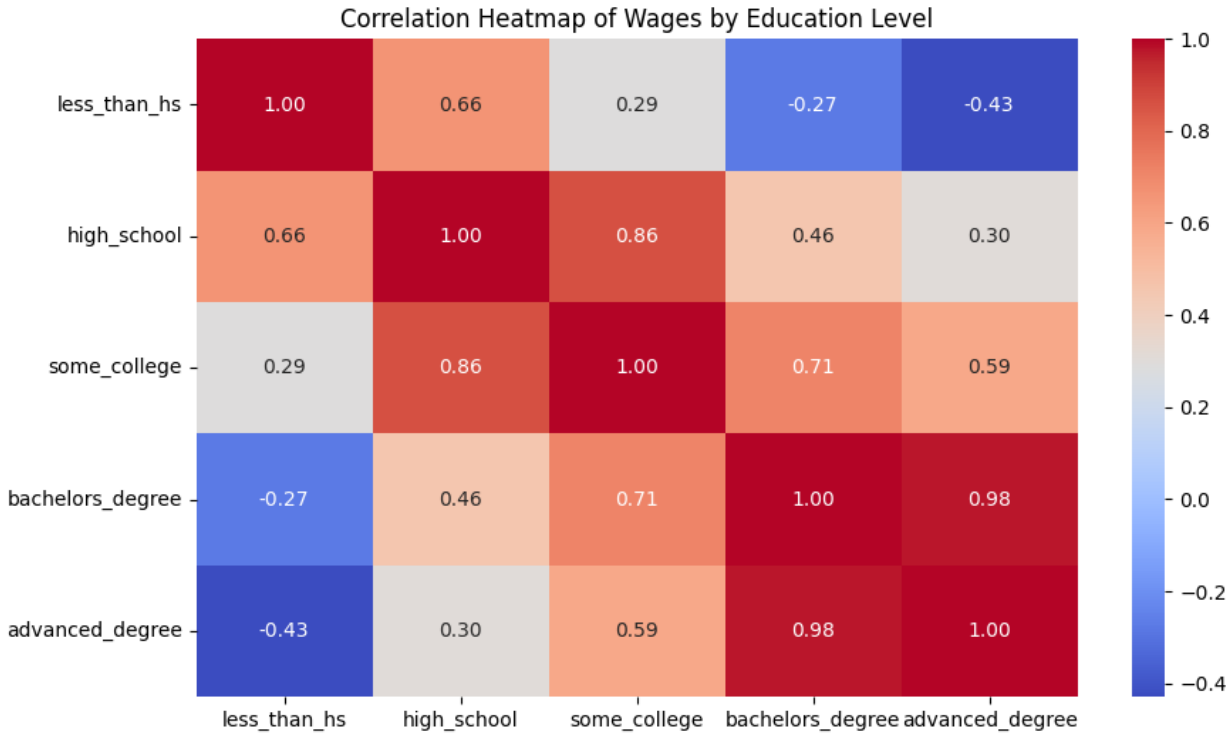
We then created a line plot. This line plot shows how wages for different education levels have changed from 1973 to 2022.

```
1.  import matplotlib.pyplot as plt
2.  import seaborn as sns
3.
4.  # Select relevant columns for wages
5.  education_levels = ['less_than_hs', 'high_school', 'some_college', 'bachelors_degree',
    'advanced_degree']
6.
7.  # Plot histogram
8.  plt.figure(figsize=(10, 6))
9.  for level in education_levels:
10.     sns.kdeplot(df[level], label=level, shade=True)
11.
12. plt.title("Distribution of Wages by Education Level")
13. plt.xlabel("Wage")
14. plt.ylabel("Density")
15. plt.legend()
16. plt.show()
```



We created a heatmap to determine the level of correlations between wages and education levels

```
1.  plt.figure(figsize=(10, 6))
2.  sns.heatmap(df[education_levels].corr(), annot=True, cmap="coolwarm", fmt=".2f")
3.  plt.title("Correlation Heatmap of Wages by Education Level")
4.  plt.show()
```



Methodology: Linear regression model

This study employs a linear regression model to quantify the relationship between education and wages. The models include:

1. Simple Linear Regression: Examining the impact of a single predictor (education) on wages:

$$wage=\beta0+\beta1(education)+\varepsilon wage$$

2. Multiple Linear Regression: Incorporating additional predictors like experience, industry, and age:

$$wage=\beta0+\beta1(education)+\beta2(experience)+\beta3(industry)+\beta4(age)+\varepsilon$$

To ensure the validity of our regression analysis, we test the following assumptions:

- **Linearity:** The relationship between education and wages is linear.
- **Normality:** The residuals of the regression should follow a normal distribution.

The linear regression models are implemented using Python libraries such as pandas, statsmodels, and sklearn.

Linear Regression

```
import pandas as pd
import statsmodels.api as sm
df = pd.read_csv("wages_by_education.csv")
education_levels = ["less_than_hs", "high_school", "some_college", "bachelors_degree", "advanced_degree"]
target_variable = "advanced_degree"  # You can modify this to another wage measure if needed
model_results = {}
for edu in education_levels:
    X = df[[edu]]
    X = sm.add_constant(X)  # Adding constant term
    Y = df[target_variable]  # Wages

    # Fit the model
    model = sm.OLS(Y, X).fit()

    # Store the summary
    model_results[edu] = model.summary()
for edu, summary in model_results.items():
    print(f"Regression results for {edu}:\n")
    print(summary)
    print("\n" + "="*80 + "\n")
```

Multiple Linear Regression

```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error
X = df[["bachelors_degree", "some_college", "high_school", "less_than_hs"]]
Y = df["advanced_degree"]  # Target variable
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2, random_state=42)
```

```
mlr_model = LinearRegression()
mlr_model.fit(X_train, Y_train)
Y_pred = mlr_model.predict(X_test)
mse = mean_squared_error(Y_test, Y_pred)
print(f"Mean Squared Error: {mse}")
print(f"Model Coefficients: {mlr_model.coef_}")
print(f"Intercept: {mlr_model.intercept_}"
```

Linearity Test

```
import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd
df = pd.read_csv("wages_by_education.csv")
education_levels = ["less_than_hs", "high_school", "some_college", "bachelors_degree",
"advanced_degree"]
target_variable = "advanced_degree"  # Modify if needed
fig, axes = plt.subplots(nrows=2, ncols=3, figsize=(15, 10))
axes = axes.flatten()
for i, edu in enumerate(education_levels):
    sns.regplot(x=df[edu], y=df[target_variable], ax=axes[i], scatter_kws={"alpha":0.5})
    axes[i].set_title(f"Education Level: {edu} vs. Wages")
    axes[i].set_xlabel(edu)
    axes[i].set_ylabel(target_variable)
for j in range(i+1, len(axes)):
    fig.delaxes(axes[j])

plt.tight_layout()
plt.show()
```

The results of these consumed a lot of space on the report, which is why the associated Jupyter notebook has been attached with necessary visualizations.

Results and Discussion

The regression results indicate that the relationship between education and wages varies across education levels. The R-squared values suggest that education level alone explains a significant portion of the variation in wages, particularly at higher levels of education.

- Bachelor's Degree shows the strongest correlation with wages, with a high R-squared value of 0.953.
- Some College has a moderate impact (R-squared = 0.348), while High School and Less Than High School have lower R-squared values, indicating weaker explanatory power.

The results suggest a positive relationship between higher education and wages. Holding all else constant, individuals with a bachelor's degree or higher earn significantly more than those with lower levels of education.

The **multiple regression model** outperforms the simple regression models, with a lower Mean Squared Error (MSE = 0.562), indicating improved predictive accuracy when additional variables are included.

Challenges and Discussion

While conducting this analysis, several challenges were encountered:

1. Data Availability and Quality:
    - Some years had missing or inconsistent data, requiring preprocessing and imputation.
    - Outliers in wage values were detected, influencing the regression models.
2. Limited Scope of Explanatory Variables:
    - The dataset primarily focused on education, omitting other crucial factors like regional economic conditions, company policies, and individual skills.

Conclusion

This study examined the impact of education on wages using linear regression models. The key findings are:

- Higher education levels are strongly associated with higher wages.
- Bachelor's degree and higher levels show the most significant wage increases, while lower education levels exhibit weaker correlations.
- Multiple regression provided a more comprehensive view, incorporating additional explanatory variables for improved accuracy.

Despite some limitations, the study supports the broader economic theory that investment in education leads to better wage prospects. Future research could incorporate additional variables (e.g., work experience, industry type, and regional economic factors) to refine the model further.