STAT 2604 Introduction to Python Programming and Elementary Data Analysis

Written Group Project:

Investigate the correlation between education levels and income across the globe

Group leader: Chan Yip Wan 3036141757

Group members: Ho Tung Tung 3036150100, Tang Ching Yan 3036140911, Lai Chun Kit 3036140923

1. Introduction

"Knowledge is wealth, wisdom is treasure, understanding is riches, and ignorance is poverty." by Matshona Dhliwayo, a well-known Canadian philosopher. It is believed that most of the parents agree with the statement especially in Asian culture, Asian parents are stereotyped as having high expectations and requirements on children's academic performance, expecting them to become professionals and getting high-paid jobs in the future. However, there are some voices who argue with the validity of the standpoint because many billionaires are not highly educated. According to the research (Mou, 2023), there is a strong correlation between education level and income level, but it is stated that the study is limited by the objective sources of data collection that might cause inaccuracy. As it is very critical for children's decision-making that affects their whole life, and also important for all of us to know how to educate children, it is significant to investigate the correlation between the two. The project is aimed to find the correlation between the education level and income level, then explain the phenomenon and make recommendations for the stakeholders including students, parents, teachers, schools, and policy makers, in order to better nurture the next generations.

2 Literature Review

Numerous studies have shown that higher educational attainment generally results in higher income levels, making the relationship between income and education level a significant area of research. The purpose of this review is to investigate the empirical data that supports this relationship, look at the theoretical frameworks that support it, and draw attention to the implications for both economic and personal development.

According to the education as human capital theory, which was mainly developed by Becker (1994), education is an investment. People weigh the expected income increase that additional education could bring against the expenses of education, such as tuition and the opportunity cost of time. According to this framework, education is valuable not only for the knowledge acquired but also for the lifetime financial gains it produces. A fundamental idea in comprehending why people seek higher education and how it results in financial gains is Becker's model. This theory can be demonstrated by the example that follows. The claim that educational attainment has a significant impact on income levels is supported by evidence from a number of studies. Based on information from Hong Kong's Census and Statistics Department (2020), The median monthly salary for full-time employees with a bachelor's degree is HK$28,000, while the average monthly salary for those with only a high school diploma is HK$15,000. According to this data, people who have a bachelor's degree make about 87% more money than those who only have a high school education.

However, lifetime income estimates also support the idea that education is an investment. The Economic Policy Institute (2014) offers lifetime income estimates that further highlight the financial benefits of higher education, showing that those with a bachelor's degree can anticipate earning roughly $2.3 million over their careers, compared to roughly $1.3 million for high school graduates. This striking disparity highlights the financial rewards of going to

college and illustrates the wider economic effects of education as a source of income and economic expansion.

The results covered in this review demonstrate how crucial it is to spend money on education in order to increase people's earning potential and promote economic expansion. Empirical evidence and human capital theory support the strong relationship between income and education level, which implies that both individuals and policymakers should place a high priority on educational attainment as a means of achieving better economic results. Therefore, based on previous research, there is a strong correlation between income and education level.

3. Data Science Methods

3.1 Data sources, data cleansing and pre-processing

	Country	Education Index	Education Level	Income
0	Argentina	0.816	Very High Education Level	High income
1	Australia	0.929	Very High Education Level	High income
2	Austria	0.852	Very High Education Level	High income
3	Bahamas	0.726	High to Moderate Education Level	High income
4	Bahrain	0.758	High to Moderate Education Level	High income
...	...	...	...	...
175	Syrian Arab Republic	0.416	Low to Moderate Education Level	Low income
176	Tajikistan	0.682	Low to Moderate Education Level	Lower middle income
177	Togo	0.517	Low to Moderate Education Level	Low income
178	Uganda	0.523	Low to Moderate Education Level	Low income
179	Yemen	0.360	Very Low Education Level	Low income

```
income_mapping = {
    'Low income': 1,
    'Lower middle income': 2,
    'Upper middle income': 3,
    'High income': 4,
}

df['income_numeric'] = df['Income'].map(income_mapping)
df = df.dropna(subset=['Education Index', 'income_numeric'])
correlation, p_value = pearsonr(df['Education Index'], df['income_numeric'])
print(f'Correlation between education level and income: {correlation}')
print(f'P-value: {p_value}')
```

```
Correlation between education level and income: 0.8660867422956213
P-value: 2.8014433002619886e-54
```

(Figure 3.1.1)

The data of the education index of different regions in the world is from Rankedex 2024 and the income data is from the United Nation. In order to deal with NaN value, the dropna function is used to remove NaN values in the specified columns of the row of income and education index. Also, because the dataset classified income by text, therefore an income mapping is needed, converting the text to integer, representing low, lower middle, upper middle and high income by 1-4.

```
data = pd.read_csv("Canadian.csv")
education_income = data[["Highest certificate, diploma or degree", "VALUE"]].dropna()

print("Counts of each education level:")
print(education_income['Highest certificate, diploma or degree'].value_counts())

missing_values = education_income[education_income['VALUE'].isna()]
print("\nMissing values in the dataset:")
print(missing_values)
```

(Figure 3.1.2)

The data of the income and certifications of Canadians in 2016 is from the government of Canada. Dropna function is also used to remove the NaN value, and use isna function to check the missing values.

	Table 210-06315 : Median monthly employment earnings of employed persons by educational attainment and sex	Unnamed: 1	Unnamed: 2	Unnamed: 3	Unnamed: 4	Unnamed: 5
0		NaN	NaN	NaN	NaN	NaN
1				Median monthly employment earnings of employed...	Median monthly employment earnings of employed...	Median monthly employment earnings of employed...
2	NaN	NaN	NaN	HK$	HK$	HK$
3	Sex	NaN	NaN	Male	Female	Both sexes
4	Year	Quarter	Educational attainment (3)	NaN	NaN	NaN
...	...	...	...	...	...	...
98		NaN	NaN	NaN	NaN	NaN
99	Source:	NaN	NaN	NaN	NaN	NaN
100	General Household Survey Section (3),Census an...	NaN	NaN	NaN	NaN	NaN
101		NaN	NaN	NaN	NaN	NaN
102	Release Date: 28 November, 2024	NaN	NaN	NaN	NaN	NaN

(Figure 3.1.3)

```
#Data Cleaning
df = pd.read_csv('Table 210-06315_en.csv')
df.columns = df.iloc[0]
df = df[5:]
df.reset_index(drop=True, inplace=True)
df.columns = ['Year', 'Quarter','Educational attainment','Male Income', 'Female Income', 'Average Income' ]
df = df.dropna(subset=['Average Income'])
df
```

	Year	Quarter	Educational attainment	Male Income	Female Income	Average Income
0	2023	NaN	Primary and below	15000	10200	12000
1	2023	NaN	Secondary	19900	13000	16000
2	2023	NaN	Lower secondary (4)	18000	12000	15000
3	2023	NaN	Upper secondary (5)	20000	13000	16100
4	2023	NaN	Post-secondary (6)	33000	27000	30000
...	...	...	...	...	...	...

(Figure 3.1.4)

```
#Data Pre-Process
df['Education Level'] = 0

for index,row in df.iterrows():
    df.at[index, 'Educational attainment']  = str(row['Educational attainment']).strip()

education_level_mapping = {
    'Primary and below': 1,
    'Lower secondary (4)': 2,
    'Upper secondary (5)': 3,
    'Post-secondary - non-degree (7)': 4,
    'Post-secondary - degree (10)': 5 }

df['Education Level'] = df['Educational attainment'].map(education_level_mapping).fillna(0).astype(int)
df = df[df['Education Level'] != 0]
df.reset_index(drop=True, inplace=True)
df
```

	Year	Quarter	Educational attainment	Male Income	Female Income	Average Income	Education Level
0	2023	NaN	Primary and below	15000	10200	12000	1
1	2023	NaN	Lower secondary (4)	18000	12000	15000	2
2	2023	NaN	Upper secondary (5)	20000	13000	16100	3
3	2023	NaN	Post-secondary - non-degree (7)	24000	19000	20500	4
4	2023	NaN	Post-secondary - degree (10)	38700	30000	35000	5
5	2023	Q1	Primary and below	15000	10000	12000	1

(Figure 3.1.5)

The next dataset referenced in this report is titled "Median monthly employment earnings of employed persons by educational attainment and sex," sourced from the Hong Kong Census and Statistics Department (2024). This dataset provides insights into the median monthly earnings of employed individuals in Hong Kong, categorized by gender—males, females, and combined totals.

Figure 3.1.3 displays the DataFrame as it is directly loaded from the CSV file. After the data cleaning process, the DataFrame is transformed into what is shown in Figure 3.1.4. After the data pre-processing step, the DataFrame is transformed into what is shown in Figure 3.1.5. and it can be used for further analysis.

In the data cleaning process, the first row of the dataset was designated as the header. Subsequently, the initial five rows were removed to enhance the dataset's clarity. The index of the DataFrame was then reset to provide a clean, continuous sequence. Following this, the columns were appropriately renamed for better understanding. Finally, rows containing NaN values in the 'Average Income' column were eliminated to maintain data integrity.

In the data pre-processing step, a new column, 'Education Level', was added to the DataFrame and initialized to 0. The values in the 'Educational attainment' column were stripped of leading and trailing whitespace to ensure consistency. A mapping dictionary was then created to assign numeric values to various educational attainment categories. This mapping was applied to populate the 'Education Level' column, replacing unmapped values with 0 and converting the column to integers. Rows with an 'Education Level' of 0 were subsequently removed to retain only relevant data. Finally, the index of the DataFrame was reset, resulting in a cleaned DataFrame that includes only pertinent educational attainment information, with numeric levels assigned.

3.2 Models Design

```
income_mapping = {
    'Low income': 1,
    'Lower middle income': 2,
    'Upper middle income': 3,
    'High income': 4,
}

df['income_numeric'] = df['Income'].map(income_mapping)
df = df.dropna(subset=['Education Index', 'income_numeric'])
correlation, p_value = pearsonr(df['Education Index'], df['income_numeric'])
print(f'Correlation between education level and income: {correlation}')
print(f'P-value: {p_value}')
```

```
Correlation between education level and income: 0.8660867422956213
P-value: 2.8014433002619886e-54
```

(Figure 3.2.1)

The above result (figure 3.2.1) shows the p-value and correlation between education index and income level of regions. The income level is converted to numerical data and inputed with the education index and use pearsonr function to calculate the p-value and correlation coefficient, and p-value is extremely close to 0, the correlation coefficient is around 0.866.

```
Mean Income by Education Level:
              Highest certificate, diploma or degree         VALUE
0                  No certificate, diploma or degree  29768.084416
1    Apprenticeship or trades certificate or diploma  48441.149351
2  College, CEGEP and other non-university certif...  45974.584416
3  University certificate or diploma below bachel...  49852.149351
4  University certificate or degree at bachelor l...  64434.077922

Correlation between Education Level and Income:
0.9059635678282686
```

(Figure 3.2.2)

In order to find the correlation between the income and people with different certifications in Canada, we used pd.categorical() to convert the certifications column 'Highest certificate, diploma or degree' into a categorical data type by the order, and used cat.codes() to assign an integer code for every category. The np.mean shows the the mean income of people with no certificate is around $29768, people with Apprenticeship or trades certificate or diploma is around $48441, people with College, CEGEP and other non-university certificate or diploma is around $45975, people with University certificate or diploma below bachelor level is around $49852, people with University certificate or degree at bachelor level or above is around $64434, and the correlation coefficient between income and certification is around 0.906.

```
m = df[['Education Level', 'Male Income']].corr()
print('The Correlation coefficient between Male Income and Educational attainment is', m['Education Level'][1])

f = df[['Education Level', 'Female Income']].corr()
print('The Correlation coefficient between Female Income and Educational attainment is', f['Education Level'][1])

a = df[['Education Level', 'Average Income']].corr()
print('The Correlation coefficient between Average Income and Educational attainment is', a['Education Level'][1])
```

```
The Correlation coefficient between Male Income and Educational attainment is 0.9042613939189457
The Correlation coefficient between Female Income and Educational attainment is 0.9044160750283544
The Correlation coefficient between Average Income and Educational attainment is 0.8991237992511785
```

(Figure 3.2.3)

Figure 3.2.3 illustrates the correlation coefficients between education level and the income levels of males, females, and the combined total in Hong Kong with method .corr. Both their correlation coefficient is about 0.9.

```
def reg(income):
    income = pd.to_numeric(income)
    x = df['Education Level']
    y = income
    x = sm.add_constant(x)
    model = sm.OLS(y, x).fit()
    print(model.summary())

reg(df['Average Income'])
reg(df['Male Income'])
reg(df['Female Income'])
```

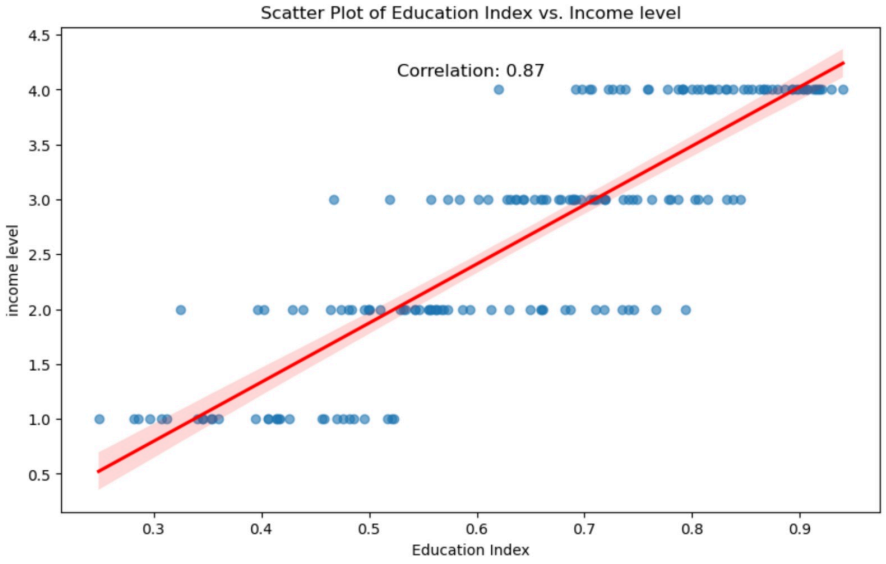
```
                            OLS Regression Results
==============================================================================
Dep. Variable:         Average Income   R-squared:                       0.808
Model:                            OLS   Adj. R-squared:                  0.803
Method:                 Least Squares   F-statistic:                     160.4
Date:                Sat, 07 Dec 2024   Prob (F-statistic):           3.29e-15
Time:                        22:40:08   Log-Likelihood:                -383.72
No. Observations:                  40   AIC:                             771.4
Df Residuals:                      38   BIC:                             774.8
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          4456.2500   1349.831      3.301      0.002    1723.660    7188.840
Education Level  5153.7500    406.989     12.663      0.000    4329.843    5977.657
==============================================================================
Omnibus:                       23.139   Durbin-Watson:                   1.756
Prob(Omnibus):                  0.000   Jarque-Bera (JB):                3.601
Skew:                           0.084   Prob(JB):                        0.165
Kurtosis:                       1.540   Cond. No.                         8.37
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

(Figure 3.2.4)

A function reg(income) is used to perform linear regression analysis using the Ordinary Least Squares (OLS) method from the statsmodels library. Inside the function, the input income is first converted to a numeric type. The independent variable x is set to the 'Education Level' column from the DataFrame df, while the dependent variable y is assigned the income values. A constant is added to x to account for the intercept in the regression model. The OLS model is then fitted using y and x, and the regression summary is printed, providing details about the model's performance. The function is subsequently called three

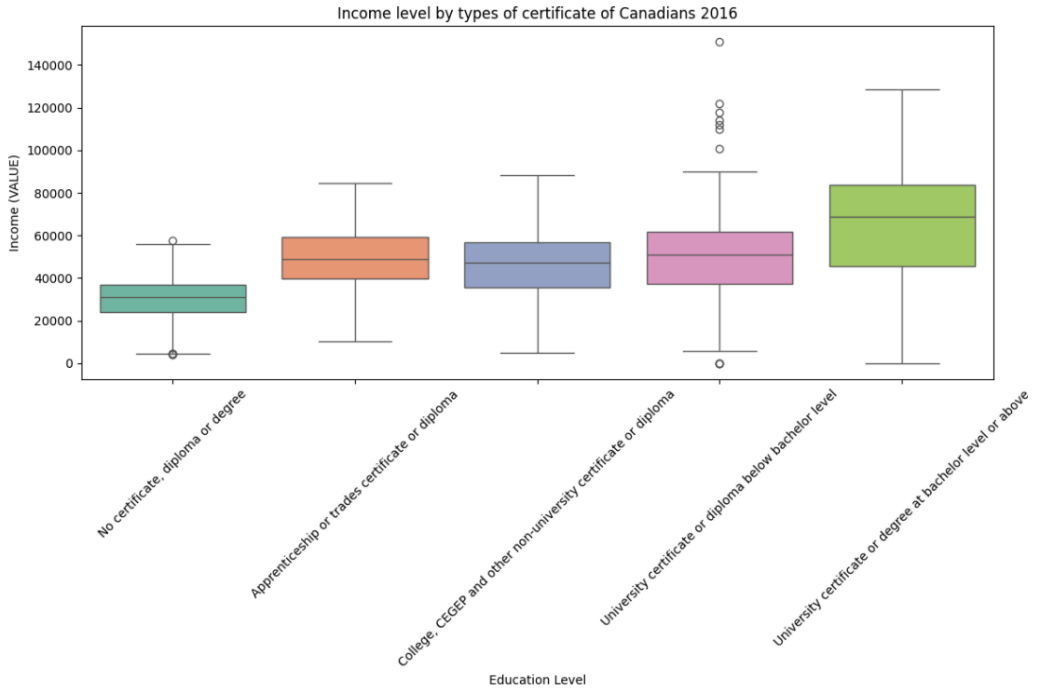
times to analyze the relationships between education level and average income, male income, and female income, respectively. For example, the equation 'Average Income = 4456.25 + 5153.75×Education Level' can be found from the OLS model.

4. Result summary and Interpretation



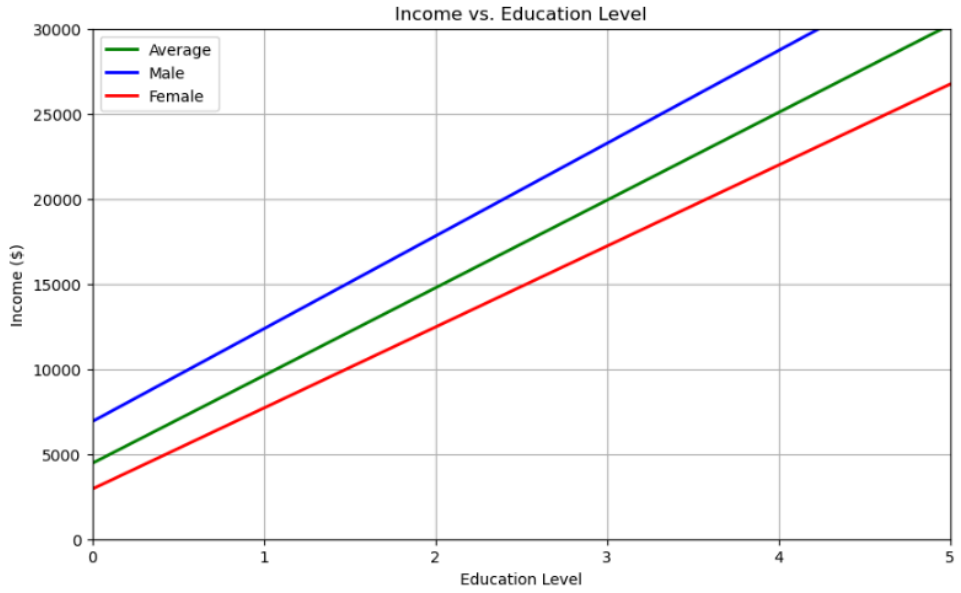
(Figure 4.1)

Figure 4.1 shows that countries or regions that have higher education index probably have a higher income level, vice versa. The correlation coefficient of around 0.87 indicates that there is a strong correlation between education index and income level of a place.



(Figure 4.2)

Figure 4.2 shows the five-number summary of the income of canadian with different certification, and it that the median income of the canadian with a bachelor's degree or above is higher than the upper quartile of the other certifications, and the upper quartile income of no certificate canadian is lower than the median of the rest. The correlation coefficient is around 0.91, which means there is a strong correlation between types of certificate and income.



(Figure 4.3)

Figure 4.3 illustrates the relationship between education level and income in Hong Kong across male, female, and combined total based on the linear regression model.It suggests that both male and female incomes rise with education, with male income typically higher than female income at each level. In addition, both their correlation coefficient is about 0.9 which indicates strong correlation between income and education level.

5. Discussions

According to the demonstrated data from the United Nations database above, the correlation coefficient of 0.84663 indicates a strong positive relationship between education levels and income across different regions. This suggests that higher education levels are associated with higher income levels. Additionally, people with advanced degrees make about $44.00, which is more than half or less than what someone with only a high school education makes. This reflects the claim that earning potentia is significantly influenced by educational level. Higher education is correlated with greater creativity and readiness to assist others at work places, Additionally, higher education is also associated with fewer absences, increasing an

individual's chances of raising up the position and earn greater income (Ng & Feldman, 2009).

Policies and regulations have an important effect on a nation's educational quality (Vorontsova et.al, 2020). According to our study results, Canada has a very high correlation coefficient between income and the education index when compared to other regions. The following is possibly because there is a high demand for skilled workers, especially in fields like technology, healthcare, and engineering, which raises wages for those with higher educational qualifications. According to the Canadian Occupational Projection System, Canada will need over 1.3 million skilled workers by 2025, emphasizing the critical need for a workforce with a high level of education. On top of that, considering that Canada's immigration rules attract in skilled workers from all over the world—more than 60% of new arrivals are from economic classes—could also be a contributing factor to the country's highly educated workforce.

We have discovered that, in Hong Kong, both genders consistently experience better incomes as a result of higher levels of education. It is notable therefore, that throughout all educational levels, male income continuously surpasses female income. Occupational segregation is one important element behind this gender pay gap. Many high-paying industries in Hong Kong, are dominated by men. It causes men in these industries to earn more on average.According to the Hong Kong Census and Statistics Department, women only accounted for roughly 13% of senior management jobs as of 2021. This underrepresentation in leadership positions reinforces current income disparities.

6. Conclusions

In conclusion, the level of education is strongly associated with income, and it is particularly influenced by a country's regulations. Therefore, investing in education and making it more accessible are effective strategies for boosting income levels across populations. Promoting higher educational attainment through targeted scholarships and supportive policies can create significant economic benefits for individuals and communities alike.

In conclusion, income and educational attainment are closely related, with national laws having a significant impact. Hence, increasing income levels across populations can be achieved through enhancing accessibility to and investment in education.

In order to reduce the barriers that frustrate access to education, we advise governments to encourage greater educational attainment through focused scholarship programs and tax incentives, especially in under-represented areas. Additionally, through helping young students to make knowledgeable decisions about their educational pathways, increasing public understanding of the long-term financial benefits of education is crucial**.**

7. References

Average earnings or employment income, by age group and highest certificate, diploma or degree - Dataset - Open Government Portal. (n.d.). https://open.canada.ca/data/en/dataset/4ff619b2-ce85-4684-af83-d46e669e043a/resource/6a14eb8f-776a-4ae0-9b8e-e342dcdc248e

Becker, G. S. (1993). *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education.* University of Chicago Press.

Census and Statistics Department, Hong Kong. (2020). 2020 Population Census.https://www.censtatd.gov.hk/tc/

Data Library. (n.d.). Economic Policy Institute. https://www.epi.org/data/

Economic Policy Institute. (2014). *The college payoff: Education, occupations, lifetime earnings* https://www.ed.gov/sites/ed/files/policy/highered/reg/hearulemaking/2011/collegepayoff.pdf

Education index by country. (n.d.-b). Rankedex. https://rankedex.com/society-rankings/education-index

Hong Kong Census and Statistics Department. (2024). *Table 210-06315A : Median monthly employment earnings of employed persons by educational attainment and sex (excluding foreign domestic helpers)* https://www.censtatd.gov.hk/en/web_table.html?id=210-06315A

Mou, W. (2023). A Quantitative Analysis of the Relationship between Education Level and Income. *Journal of Education Humanities and Social Sciences*, *12*, 160–166. https://doi.org/10.54097/ehss.v12i.7617

Ng, T. W., & Feldman, D. C. (2009). How broadly does education contribute to job performance?. *Personnel psychology*, 62(1), 89-134.

https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1744-6570.2008.01130.x

Vorontsova, A. S., Vasylieva, T. A., Bilan, Y. V., Ostasz, G., & Mayboroda, T. (2020). The influence of state regulation of education for achieving the sustainable development goals: case study of Central and Eastern European countries. https://www.ceeol.com/search/article-detail?id=964040

World Economic Situation and Prospects 2022 | Department of Economic and Social Affairs. (n.d.). https://www.un.org/development/desa/dpad/publication/world-economic-situation-and-prospects-2022/

8. Appendix

8.1 Program Documentation

Converting text to numerical and calculating correlation coefficient:

```
income_mapping = {
    'Low income': 1,
    'Lower middle income': 2,
    'Upper middle income': 3,
    'High income': 4,
}

df['income_numeric'] = df['Income'].map(income_mapping)
df = df.dropna(subset=['Education Index', 'income_numeric'])
correlation, p_value = pearsonr(df['Education Index'], df['income_numeric'])
print(f'Correlation between education level and income: {correlation}')
print(f'P-value: {p_value}')
```

```
Correlation between education level and income: 0.8660867422956213
P-value: 2.8014433002619886e-54
```

Constructing a scatter plot for income level and education index:

```
import seaborn as sns
import matplotlib.pyplot as plt
```

```
x = df['Education Index']
y = df['income_numeric']

plt.figure(figsize=(10, 6))
sns.regplot(x=x, y=y, scatter_kws={'alpha':0.6}, line_kws={"color": "red"})
plt.xlabel('Education Index')
plt.ylabel('income level')
plt.title('Scatter Plot of Education Index vs. Income level')
plt.text(0.5, 0.9, f'Correlation: {correlation:.2f}',
         transform=plt.gca().transAxes, fontsize=12, ha='center')
plt.show()
```

Calculating the mean wage of different educational backgrounds:

```
import numpy as np

data = df['less_than_hs'].to_numpy()

mean_value = np.mean(data)
median_value = np.median(data)

print(f'Mean: {mean_value}')
print(f'Median: {median_value}')
```

```
Mean: 15.7026
Median: 15.34
```

Constructing boxplot:

```
import matplotlib.pyplot as plt

wage_columns = [
    'less_than_hs', 'high_school', 'some_college',
    'bachelors_degree', 'advanced_degree'
]
wages = data[wage_columns]

plt.figure(figsize=(10, 6))
plt.boxplot(wages, labels=wage_columns)
plt.title('Boxplot of Wages by Educational Background')
plt.ylabel('Wages')
plt.xlabel('Educational Background')
plt.grid(axis='y')
plt.show()
```

```
education_order = [
    "No certificate, diploma or degree",
    "Apprenticeship or trades certificate or diploma",
    "College, CEGEP and other non-university certificate or diploma",
    "University certificate or diploma below bachelor level",
    "University certificate or degree at bachelor level or above"
]

education_income['Highest certificate, diploma or degree'] = pd.Categorical(
    education_income['Highest certificate, diploma or degree'],
    categories=education_order,
    ordered=True
)

plt.figure(figsize=(12, 8))
sns.boxplot(x='Highest certificate, diploma or degree', y='VALUE', data=education_income, palette="Set2")
plt.title('Income level by types of certificate of Canadians 2016')
plt.xlabel('Education Level')
plt.ylabel('Income (VALUE)')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```

Calculating the income mean and correlation coefficient of certifications and income in Canada:

```
education_income = data[["Highest certificate, diploma or degree", "VALUE"]].dropna()


missing_values = education_income[education_income['VALUE'].isna()]
print("\nMissing values in the dataset:")
print(missing_values)

education_order = [
    "No certificate, diploma or degree",
    "Apprenticeship or trades certificate or diploma",
    "College, CEGEP and other non-university certificate or diploma",
    "University certificate or diploma below bachelor level",
    "University certificate or degree at bachelor level or above"
]

education_income['Highest certificate, diploma or degree'] = pd.Categorical(
    education_income['Highest certificate, diploma or degree'],
    categories=education_order,
    ordered=True
)

mean_income = education_income.groupby("Highest certificate, diploma or degree")["VALUE"].mean().reset_index()

print("\nMean Income by Education Level:")
print(mean_income)

mean_income['Education Level'] = mean_income['Highest certificate, diploma or degree'].cat.codes

correlation = mean_income['Education Level'].corr(mean_income['VALUE'])

print("\nCorrelation between Education Level and Income:")
print(correlation)
```

8.3 Program Coding

```
import pandas as pd
from scipy.stats import pearsonr
df = pd.read_csv("Income and education.csv")
df
income_mapping = {
    'Low income': 1,
    'Lower middle income': 2,
    'Upper middle income': 3,
    'High income': 4,
}

df['income_numeric'] = df['Income'].map(income_mapping)
df = df.dropna(subset=['Education Index', 'income_numeric'])
correlation, p_value = pearsonr(df['Education Index'], df['income_numeric'])
print(f'Correlation between education level and income: {correlation}')
import seaborn as sns
import matplotlib.pyplot as plt
x = df['Education Index']
y = df['income_numeric']

plt.figure(figsize=(10, 6))
sns.regplot(x=x, y=y, scatter_kws={'alpha':0.6}, line_kws={"color": "red"})
plt.xlabel('Education Index')
plt.ylabel('income level')
```

```
plt.title('Scatter Plot of Education Index vs. Income level')
plt.text(0.5, 0.9, f'Correlation: {correlation:.2f}',
         transform=plt.gca().transAxes, fontsize=12, ha='center')
plt.show()
df = pd.read_csv("wages_by_education.csv")
df
import numpy as np
data = df['less_than_hs'].to_numpy()
mean_value = np.mean(data)
median_value = np.median(data)
print(f'Mean: {mean_value}')
print(f'Median: {median_value}')

data = df['high_school'].to_numpy()
mean_value = np.mean(data)
median_value = np.median(data)
print(f'Mean: {mean_value}')
print(f'Median: {median_value}')

data = df['some_college'].to_numpy()
mean_value = np.mean(data)
median_value = np.median(data)
print(f'Mean: {mean_value}')
print(f'Median: {median_value}')
```

```
data = df['bachelors_degree'].to_numpy()
mean_value = np.mean(data)
median_value = np.median(data)
print(f'Mean: {mean_value}')
print(f'Median: {median_value}')

data = df['advanced_degree'].to_numpy()
mean_value = np.mean(data)
median_value = np.median(data)
print(f'Mean: {mean_value}')
print(f'Median: {median_value}')

import matplotlib.pyplot as plt
wage_columns = [
    'less_than_hs', 'high_school', 'some_college',
    'bachelors_degree', 'advanced_degree'
]
wages = data[wage_columns]
plt.figure(figsize=(10, 6))
plt.boxplot(wages, labels=wage_columns)
plt.title('Boxplot of Wages by Educational Background')
plt.ylabel('Wages')
plt.xlabel('Educational Background')
plt.grid(axis='y')
plt.show()
```

```
df = pd.read_csv("Canadia.csv")

df

education_income = data[["Highest certificate, diploma or degree", "VALUE"]].dropna()

missing_values = education_income[education_income['VALUE'].isna()]

print("\nMissing values in the dataset:")

print(missing_values)

education_order = [

    "No certificate, diploma or degree",

    "Apprenticeship or trades certificate or diploma",

    "College, CEGEP and other non-university certificate or diploma",

    "University certificate or diploma below bachelor level",

    "University certificate or degree at bachelor level or above"

]

education_income['Highest certificate, diploma or degree'] = pd.Categorical(

    education_income['Highest certificate, diploma or degree'],

    categories=education_order,

    ordered=True

)

mean_income = education_income.groupby("Highest certificate, diploma or
degree")["VALUE"].mean().reset_index()

print("\nMean Income by Education Level:")

print(mean_income)

mean_income['Education Level'] = mean_income['Highest certificate, diploma or
degree'].cat.codes

correlation = mean_income['Education Level'].corr(mean_income['VALUE'])
```

```
print("\nCorrelation between Education Level and Income:")
print(correlation)
education_order = [
    "No certificate, diploma or degree",
    "Apprenticeship or trades certificate or diploma",
    "College, CEGEP and other non-university certificate or diploma",
    "University certificate or diploma below bachelor level",
    "University certificate or degree at bachelor level or above"
]

education_income['Highest certificate, diploma or degree'] = pd.Categorical(
    education_income['Highest certificate, diploma or degree'],
    categories=education_order,
    ordered=True
)

plt.figure(figsize=(12, 8))
sns.boxplot(x='Highest certificate, diploma or degree', y='VALUE', data=education_income, palette="Set2")
plt.title('Income level by types of certificate of Canadians 2016')
plt.xlabel('Education Level')
plt.ylabel('Income (VALUE)')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```

```
import pandas as pd
import numpy as np
import statsmodels.api as sm
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression

#Data Cleaning
df = pd.read_csv('Table 210-06315_en.csv')
df.columns = df.iloc[0]
df = df[5:]
df.reset_index(drop=True, inplace=True)
df.columns = ['Year', 'Quarter','Educational attainment','Male Income', 'Female Income',
            'Average Income' ]
df = df.dropna(subset=['Average Income'])
df

#Data Pre-Process
df['Education Level'] = 0

for index,row in df.iterrows():
    df.at[index, 'Educational attainment']  = str(row['Educational attainment']).strip()

education_level_mapping = {
    'Primary and below': 1,
```

```
    'Lower secondary (4)': 2,
    'Upper secondary (5)': 3,
    'Post-secondary - non-degree (7)': 4,
    'Post-secondary - degree (10)': 5 }

df['Education Level'] = df['Educational
        attainment'].map(education_level_mapping).fillna(0).astype(int)
df = df[df['Education Level'] != 0]
df.reset_index(drop=True, inplace=True)
df

m = df[['Education Level', 'Male Income']].corr()
print('The Correlation coefficient between Male Income and Educational attainment is',
        m['Education Level'][1])

f = df[['Education Level', 'Female Income']].corr()
print('The Correlation coefficient between Female Income and Educational attainment is',
        f['Education Level'][1])

a = df[['Education Level', 'Average Income']].corr()
print('The Correlation coefficient between Average Income and Educational attainment is',
        a['Education Level'][1])

def reg(income):
    income = pd.to_numeric(income)
```

```
    x = df['Education Level']
    y = income
    x = sm.add_constant(x)
    model = sm.OLS(y, x).fit()
    print(model.summary())

reg(df['Average Income'])
reg(df['Male Income'])
reg(df['Female Income'])

import matplotlib.pyplot as plt

a_intercept = 4456.25
a_slope = 5153.75

m_intercept = 6910
m_slope = 5450

f_intercept = 2943.75
f_slope = 4758.75


# Create a range of education levels for the regression line
education_levels = np.linspace(0, 5, 100)
predicted_income = a_intercept + a_slope * education_levels
```

```
predicted_male_income = m_intercept + m_slope * education_levels

predicted_female_income = f_intercept + f_slope * education_levels


# Create the scatter plot

plt.figure(figsize=(10, 6))

plt.plot(education_levels, predicted_income, color='green', label='Average', linewidth=2)

plt.plot(education_levels, predicted_male_income, color='blue', label='Male', linewidth=2)

plt.plot(education_levels, predicted_female_income, color='red', label='Female', linewidth=2)


# Labels and Title

plt.title('Income vs. Education Level')

plt.xlabel('Education Level')

plt.ylabel('Income ($)')

plt.xlim(0, 5)

plt.ylim(0, 30000)

plt.grid()

plt.legend()

plt.show()
```