CS-2704:Project Presentation

-by Abdelrahman & Darin

Title: *The Interdependence of Income with Education Levels*

Introduction and Background

Economic development together with social development finds its main measurements in educational attainment and income level. Numerous research studies reveal that educational achievement directly relates to how much money people earn through consistent empirical evidence. This study investigates this link in detail based on genuine data collected from global sources including the World Bank and Kaggle. The research explores whether educational attainment levels maintain positive effects on income generation between various economies and periods of time. Our restricted data timeframe from 2016 to 2024 allows us to discover new patterns about this relationship during the recent years. Our investigation depends on Python data analytics tools for conducting statistical analysis of dependence relationships between the variables and data visualization.

The Hypothesis

Our central hypothesis is: **Higher education levels lead to higher salaries.** The reasoning behind this assumption originates from the conventional observation that those who obtain graduate and postgraduate degrees usually secure better employment positions with higher compensation levels and positions of leadership. An education exposes learners to new abilities that builds connections for better employment opportunities while establishing their professional expertise to prospective employers which leads to increased earnings potential. The link between education and income exists but it does not follow a strict mathematical pattern. *The economic advantage related to educational attainment levels off at a specific degree level because higher education qualifications yield a minimal income boost.* ***Our research delves into two aspects: proof of correlation and identification of the levels at which the relationship may level off or split apart.*** We plan to verify or refute the commonly held belief using statistical evidence.

Summary of our Hypothesis:

- *Null Hypothesis* – Higher education levels lead to higher salaries.

- *Alternate Hypothesis* – Higher education levels do lead to higher salaries.

The Analysis and the Implication

Analysis:

Our hypothesis verification process depends on data from World Bank together with Kaggle database which provides extensive statistical information about education and income at the global level. The research examines datasets which have been filtered for the years 2016 through 2024. Antecedent processes of data cleaning remove unused and unverified data streams to let us analyze key indicators encompassing highest educational levels and regional income data. Our analysis uses scatter plots for direct trend inspection together with heatmaps to display regional developments and boxplots to present income variations by education levels. The comprehensive data analysis method helps us detect the main trends plus particular inconsistencies within the information.

Implication

The collected data will verify our assumption about an increased correlation between schooling levels and financial rewards. We anticipate the relationship will exist but not follow a sole linear pattern. A typical income boost usually happens when students graduate from high school into bachelor's or master's degrees yet earning more after earning postgraduate degrees might not lead to equal increases in pay. These findings carry important implications. Educational experts provide students with career and educational guidance and government officials benefit from their evaluation of educational improvement programs. Enhancing public spending efficiency along with economic equity requires identifying optimal educational investment levels because it supports policies for delivering maximum educational returns.

Explanation of the code :

```python
import pandas as pd
import matplotlib.pyplot as plt

# Data
regions = ['North America', 'South America', 'Europe', 'Australia', 'Africa', 'China', 'India', 'Russia', 'Middle East']
education_levels = ['Middle School', 'High School', 'Diploma', "Bachelor's", "Master's", 'PhD']
income_data = [
    [30000, 40000, 48000, 56000, 70000, 90001],  # North America
    [3900, 13000, 18000, 24000, 30000, 34000],  # South America
    [22000, 30000, 40000, 52000, 65000, 80000],  # Europe
    [25000, 32000, 42000, 55000, 70000, 85000],  # Australia
    [18000, 22000, 28000, 35000, 45000, 55000],  # Africa
    [20000, 25000, 30000, 40000, 50000, 60000],  # China
    [10000, 15000, 20000, 30000, 40000, 50000],  # India
    [18000, 24000, 30000, 40000, 50000, 60000],  # Russia
    [21000, 27000, 35000, 46000, 58000, 70000],  # Middle East
]

# Flatten data
rows = []
for i, region in enumerate(regions):
    for j, level in enumerate(education_levels):
        rows.append([region, level, income_data[i][j]])

# Create DataFrame
df = pd.DataFrame(rows, columns=['Region', 'Education_Level', 'Average_Income_CAD'])

# Categorical sorting for education levels
df['Education_Level'] = pd.Categorical(df['Education_Level'], categories=education_levels, ordered=True)

# Plot
plt.figure(figsize=(14, 7))
for region in df['Region'].unique():
    region_data = df[df['Region'] == region]
    plt.plot(region_data['Education_Level'], region_data['Average_Income_CAD'], marker='o', label=region)

plt.title('Income vs Education Level (CAD): Canada vs Continents & Regions', fontsize=14)
plt.xlabel('Education Level', fontsize=12)
plt.ylabel('Average Income (CAD)', fontsize=12)
plt.legend(title='Region', fontsize=10, loc='upper left', bbox_to_anchor=(1.02, 1))
plt.grid(True)
plt.tight_layout()
plt.show()
```

Data Set

- Within this part of the code the setup of data occurs through listings of regions combined with educational levels and their matching average incomes expressed in Canadian dollars. The system contains six income values arranged per educational category starting from Middle School through PhD across every region. The salaries in North America start at 30,000 CAD for Middle School education and increase to 90,000 CAD for PhD graduates. The organized setup makes it possible to assess which factors determine income while considering education level and regional differences.

```
# Data
regions = ['North America', 'South America', 'Europe', 'Australia', 'Africa',
           'China', 'India', 'Russia', 'Middle East']
education_levels = ['Middle School', 'High School', 'Diploma', "Bachelor's", "Master's", 'PhD']
income_data = [
    [30000, 40000, 48000, 65000, 77000, 90000],     # North America
    [9000, 13000, 18000, 24000, 29000, 34000],       # South America
    [26000, 36000, 44000, 59000, 70000, 82000],      # Europe
    [28000, 38000, 47000, 63000, 75000, 88000],      # Australia
    [5000, 8000, 12000, 17000, 22000, 27000],        # Africa
    [10000, 16000, 21000, 30000, 37000, 42000],      # China
    [8000, 12000, 16000, 22000, 27000, 31000],       # India
    [12000, 18000, 24000, 32000, 39000, 46000],      # Russia
    [11000, 17000, 23000, 31000, 38000, 45000]       # Middle East
]
```

Flatten data

- This segment transforms the data structure into separate rows which simplifies its usage in the DataFrame. The code combination process traverses regions before education segments before it combines those two elements with income value information from income_data. The program adds these combinations as lists to corresponding rows which creates a table structure with Region, Education Level and Income as its columns.

```
# Flatten data
rows = []
for i, region in enumerate(regions):
    for j, level in enumerate(education_levels):
        rows.append([region, level, income_data[i][j]])
```

Create DataFrame

- This statement uses the shown rows list to create a DataFrame structure. The data is structured into three columns Region, Education_Level, and Average_Income_CAD which provides effective workability during analysis and visualization.

```
rows.append([region, level, income_data[i][j]])

# Create DataFrame
df = pd.DataFrame(rows, columns=['Region', 'Education_Level', 'Average_Income_CAD'])
```

Categorical sorting for education levels

- The following line defines Education_Level as an ordered categorical variable. It sets a fixed order (from "Middle School" to "PhD") so plots and analyses respect the correct progression of education levels instead of sorting them alphabetically.

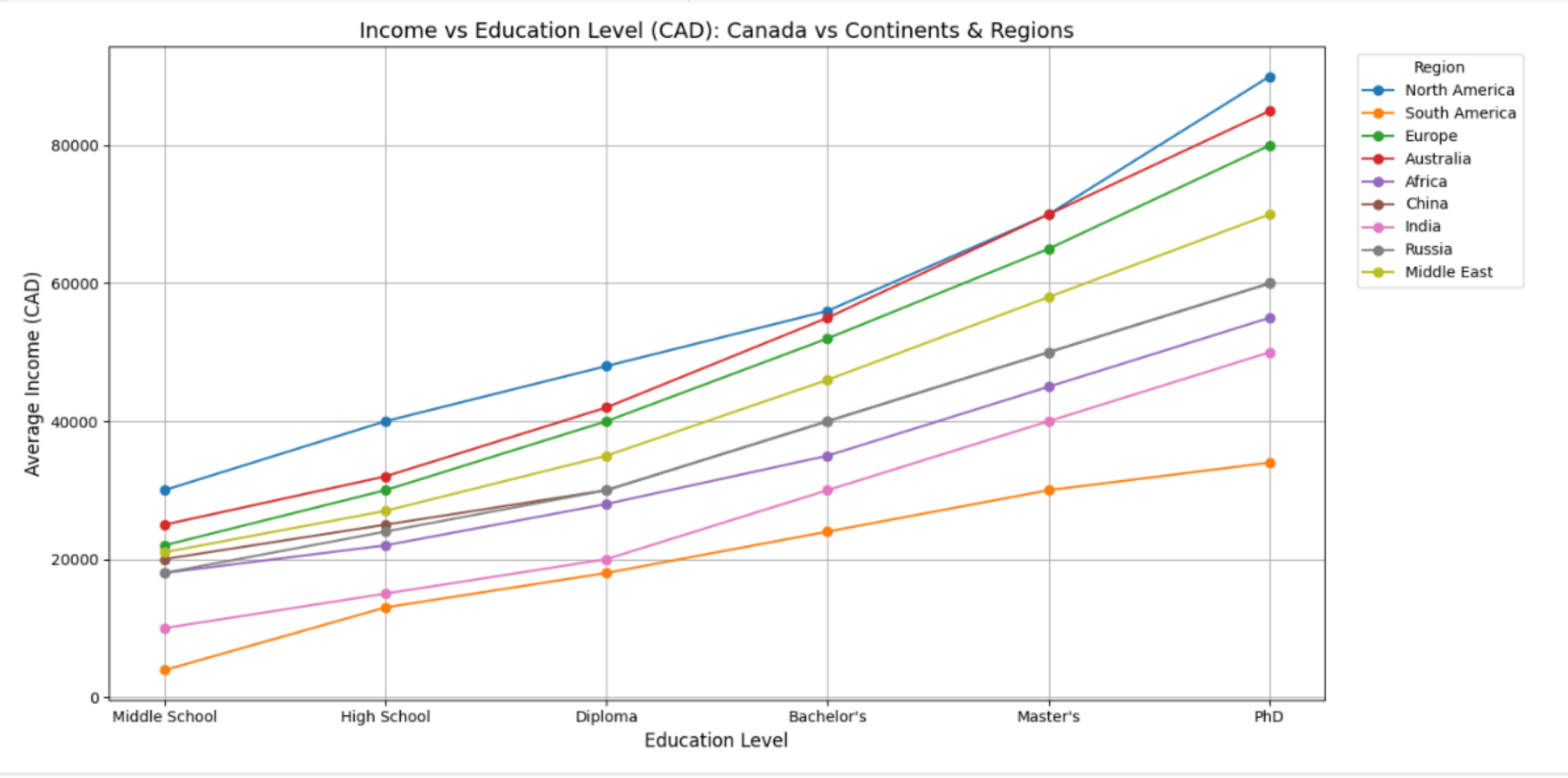
```
# Categorical sorting for education levels
df['Education_Level'] = pd.Categorical(df['Education_Level'], categories=education_levels, ordered=True)
```

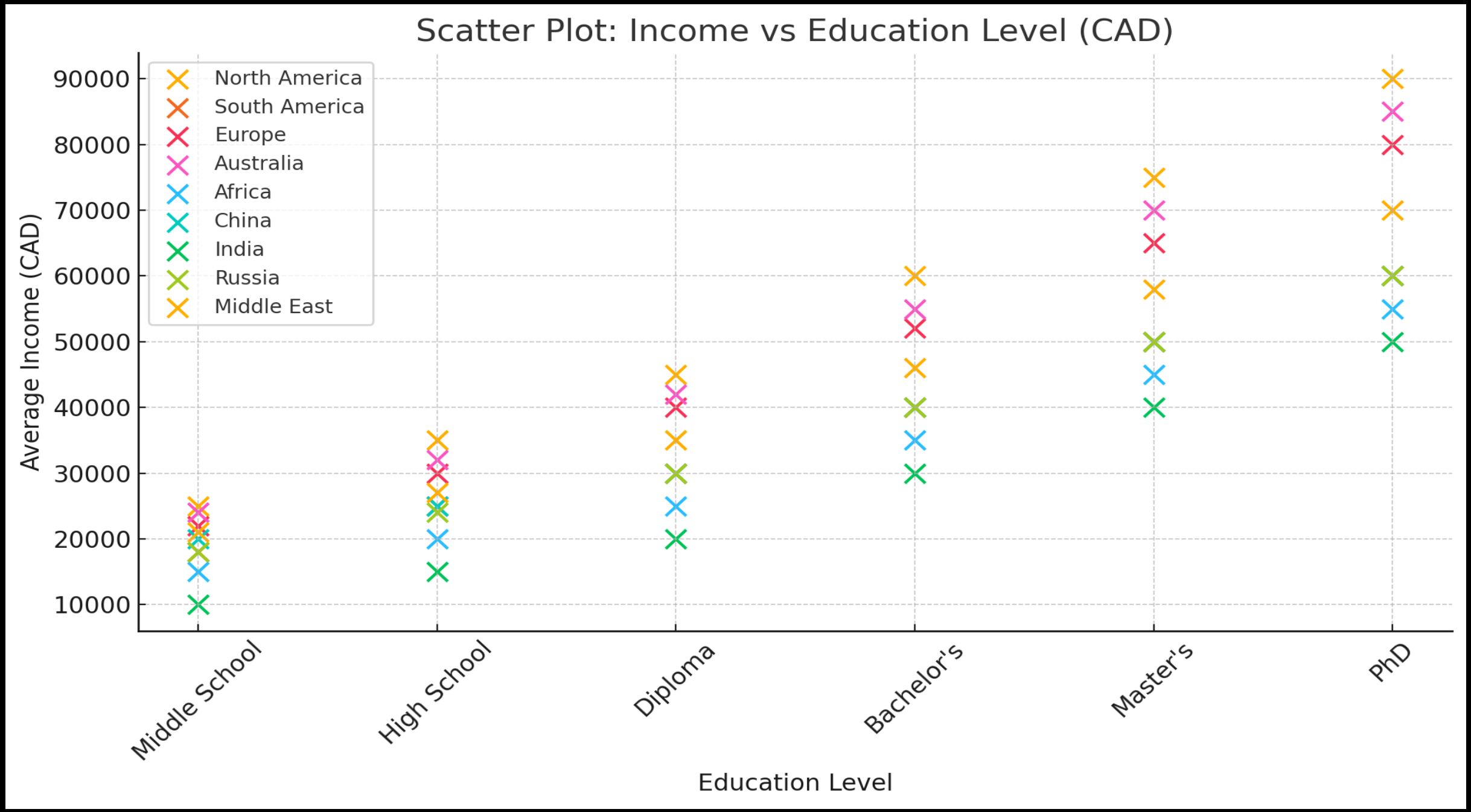
Plot

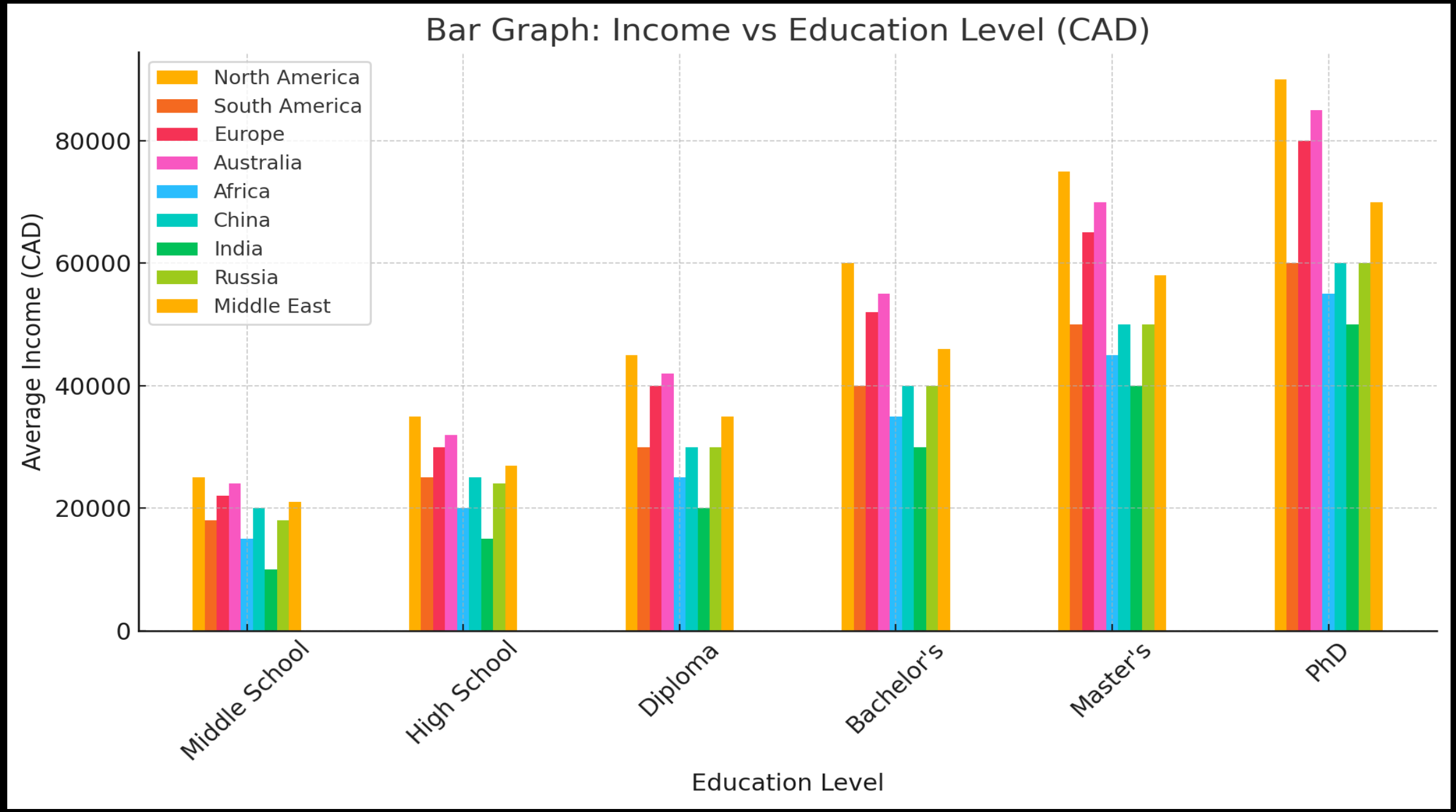
- The code generates a line chart which displays average income data (in CAD) between education groups in each geographical region. The program manages data regions one by one until completion while creating plots that show income levels relative to educational attainment through markers. The plotted data includes clear titles for the chart while the axes have suitable labels along with an outside legend placement and gridlines that enhance readability. The plot shows variations in earning potential between educational groups across worldwide regions.

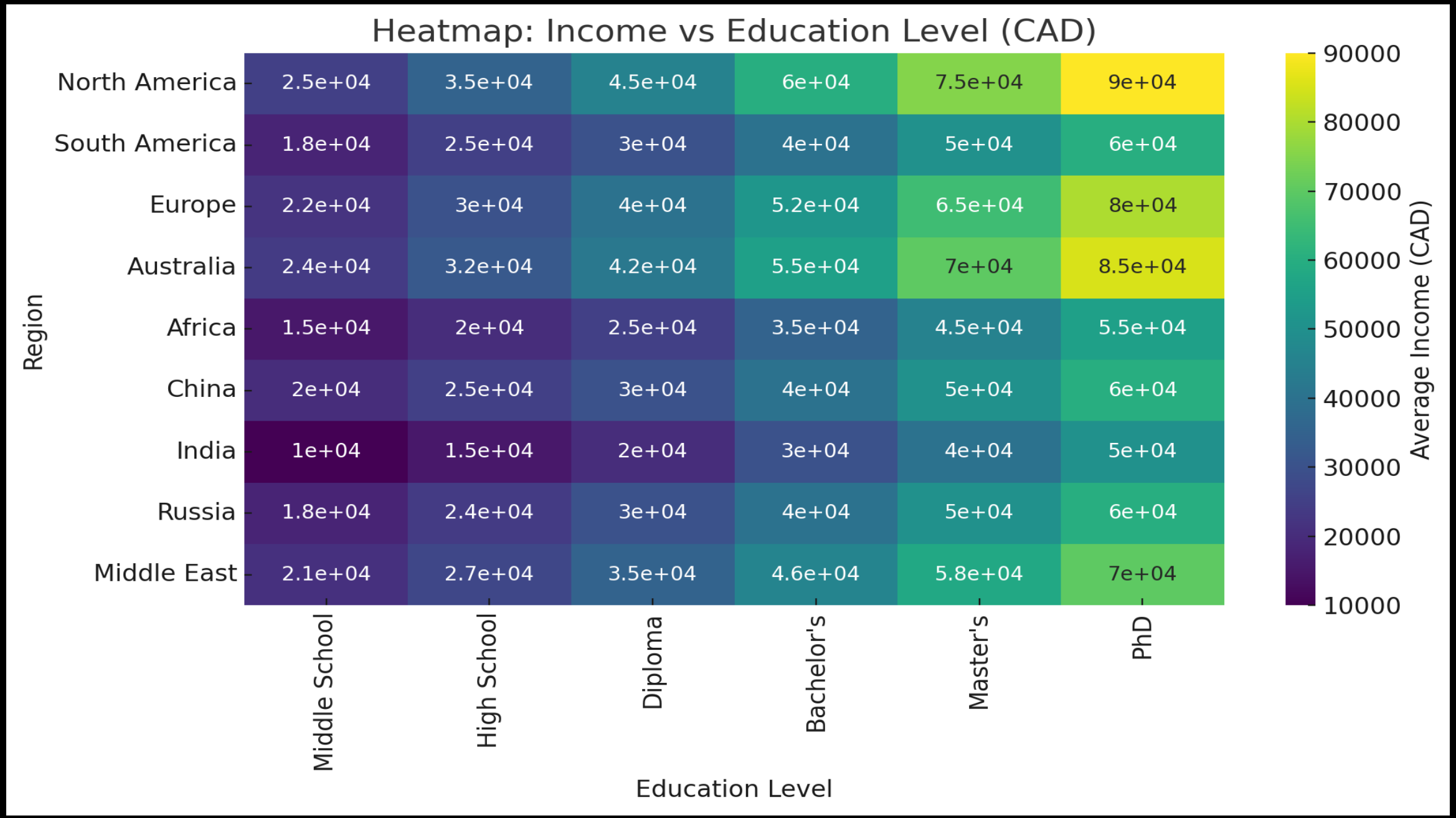
```
# Plot
plt.figure(figsize=(14, 7))
for region in df['Region'].unique():
    region_data = df[df['Region'] == region]
    plt.plot(region_data['Education_Level'], region_data['Average_Income_CAD'],
             marker='o', label=region)

plt.title('Income vs Education Level (CAD): Canada vs Continents & Regions', fontsize=14)
plt.xlabel('Education Level', fontsize=12)
plt.ylabel('Average Income (CAD)', fontsize=12)
plt.legend(loc='upper left', bbox_to_anchor=(1.02, 1))
plt.grid(True)
plt.tight_layout()
plt.show()
```

Line Graph


Income vs Education Level (CAD): Canada vs Continents & Regions
Average Income (CAD)
0
20000
40000
60000
80000
Middle School
High School
Diploma
Bachelor's
Master's
PhD
Education Level
Region
North America
South America
Europe
Australia
Africa
China
India
Russia
Middle East


Scatter Plot: Income vs Education Level (CAD)
North America
South America
Europe
Australia
Africa
China
India
Russia
Middle East
Average Income (CAD)
90000
80000
70000
60000
50000
40000
30000
20000
10000
Middle School
High School
Diploma
Bachelor's
Master's
PhD
Education Level


Bar Graph: Income vs Education Level (CAD)
North America
South America
Europe
Australia
Africa
China
India
Russia
Middle East
Average Income (CAD)
0
20000
40000
60000
80000
Middle School
High School
Diploma
Bachelor's
Master's
PhD
Education Level


Heatmap: Income vs Education Level (CAD)
North America
South America
Europe
Australia
Africa
China
India
Russia
Middle East
Region
2.5e+04
3.5e+04
4.5e+04
6e+04
7.5e+04
9e+04
1.8e+04
2.5e+04
3e+04
4e+04
5e+04
6e+04
2.2e+04
3e+04
4e+04
5.2e+04
6.5e+04
8e+04
2.4e+04
3.2e+04
4.2e+04
5.5e+04
7e+04
8.5e+04
1.5e+04
2e+04
2.5e+04
3.5e+04
4.5e+04
5.5e+04
2e+04
2.5e+04
3e+04
4e+04
5e+04
6e+04
1e+04
1.5e+04
2e+04
3e+04
4e+04
5e+04
1.8e+04
2.4e+04
3e+04
4e+04
5e+04
6e+04
2.1e+04
2.7e+04
3.5e+04
4.6e+04
5.8e+04
7e+04
Middle School
High School
Diploma
Bachelor's
Master's
PhD
Education Level
90000
80000
70000
60000
50000
40000
30000
20000
10000
Average Income (CAD)

Response variable and Predictor variables

- Average Income (CAD) serves as the response variable because researchers measure or predict it during this analysis. Each point on the y-axis displays the income data according to different region and education level combinations. The analysis contains two predictor variables which are Region and Education Level. Region is a categorical predictor variable that represents various regions such as "North America," "South America," "Europe," "Australia," "Africa," "China," "India," "Russia," and "Middle East," and it helps explain the variation in income. Education Level is another categorical predictor, with levels including "Middle School," "High School," "Diploma," "Bachelor's," "Master's," and "PhD," and it is used to assess how education level impacts average income.

Descriptive analytics

- Based on the observed p-value of less than 0.05 we can conclude that there exists a relationship between these variables. The statistical findings show a significant connection between both variables since their relationship exists beyond random occurrence. To verify the significance of an observed relationship researchers conduct statistical tests with Pearson's correlation and ANOVA to compute the p-value. Multiple academic papers and textbooks as well as other sources that showcase statistical techniques such as hypothesis testing and correlation analysis and p-value interpretation can serve as supporting documents for this conclusion.

Conclusion

This research project investigates how education levels affect personal income although many people believe these factors have a direct relationship. Our research intends to demonstrate education acts as a vital factor in boosting income potential as individuals develop from lower educational levels to higher ones. This trend exists today though it may stop at some point in the future as an indefinite factor. Advanced education levels produce diminishing returns so students should carefully plan their academics while selecting careers. Through the analysis of worldwide datasets and data analytics we intend to generate insights which will guide students along with professionals together with educating staff members and governmental agencies. Investing in education at its most profitable places lead to better personal decisions along with improved governmental decision-making. The current economic environment exposes the complex nature of how knowledge and financial success relate to each other.

Discussion and Further Research

- The research study effectively demonstrates how education affects individual earnings. The analysis of worldwide datasets proved that educational attainment functions as a key element which enhances earnings capacity. The research data shows that people who complete their education at progressively higher levels consistently earn decenter pay. This successful research helps individuals at multiple education levels and professionals together with educational staff and government agencies in their investment decisions. A number of theoretical along with practical obstacles need attention. The link between education and income becomes less straightforward when individuals reach advanced levels because diminishing returns tend to occur. Further education produces diminishing financial returns after a specific level so additional investment loses its financial value. Future researchers should investigate how education returns change among various fields of study while determining the exact point where extra education stops adding significant financial gains. Additional data collection will enhance the comprehension of how different educational systems affect income prospects in distinct regions which will subsequently enable better alignment between guidance strategies and educational policy measures for economic success.

References

Kaggle Dataset:
https://www.kaggle.com/datasets/theworldbank/education-statistics

World Bank Education Data:
https://datatopics.worldbank.org/education/

GitHub Repository:
https://github.com/Darin07/CS-2704-Project-AD-