DATA ANALYTICS AND APPLIED MATHEMATICS
E-ISSN: 2773-4854
VOLUME 7, ISSUE 1, 2026, 68 - 74
DOI: https://doi.org/10.15282/daam.v7i1.13687

RESEARCH ARTICLE

Teaching basic data literacy through Python: Integrating basic computer science and Mathematics for lower secondary students

Ming Hui Lim*

Mathematics Section, School of Distance Education, Universiti Sains Malaysia,11800 USM Penang, Malaysia

Abstract - This conceptual paper proposes an interdisciplinary pedagogical approach to teaching basic data literacy to lower secondary students by combining elements from the Basic Computer Science (BCS) and Mathematics syllabi in Malaysia's KSSM curriculum. Motivated by the growing importance of digital skills and real-world data analysis, this conceptual paper aims to address the limited exposure students have to data handling and interpretation in current classroom practice. The proposed method uses Python to introduce concepts such as data preprocessing, data visualization, and basic statistical analysis, with example tasks and hypothetical datasets to demonstrate how the approach can be implemented. Although the approach has not yet been implemented in classroom settings, it is designed to align with syllabus standards and promote cross-curricular connections. No student feedback or empirical results are reported in this paper. Instead, it presents a pedagogical approach that makes data literacy accessible, engaging, and practically relevant for young learners.

Article History
Received : 23 December 2025
Revised : 18 February 2026
Accepted : 9 March 2026
Published : 31 March 2026

Keywords
Data literacy
Mathematics education
Computer science education
Python programming
Interdisciplinary pedagogy

1. Introduction

Data is produced at an unprecedented scale today. Data plays a vital role in every aspect of our daily lives, such as social media activity, online transactions, climate monitoring, and public health reporting. Thus, there is a need to emphasize data literacy skills among students. Simply speaking, data literacy is an inquiry-based approach to using data to understand real-world phenomena [1]. It involves reading, interpreting, evaluating, and communicating data effectively. Individuals are required to ask meaningful questions, identify patterns, and make evidence-based decisions. In the education context, mathematics education primarily focuses on calculation and problem-solving, while computer science education emphasizes computational and algorithmic thinking. In this scenario, integrating mathematics with computer science can develop students' practical data-handling skills, strengthen their analytical thinking, and build a solid foundation for data literacy. It can be incorporated into lower secondary education through the existing Malaysian national curriculum. For lower secondary education, they are Mathematics and basic computer science (BCS)) subject. The mathematics curriculum includes basic data analysis concepts, such as data handling, descriptive statistics, and probability [2–4]. In parallel, the basic computer science curriculum develops computational thinking and algorithmic problem-solving through practical exposure to programming languages [5–7]. Hence, integrating mathematics and computer science enables students to analyse data scientifically and apply data-driven problem-solving strategies, laying a foundation for data literacy. Furthermore, this integration aligns well with the objectives outlined in the Digital Education Policy [8], one of which is to develop students' skills in analysing data scientifically. In addition, the OECD Core Foundations for 2030 emphasize the importance of developing students' digital and data literacy to navigate future challenges successfully [9].

There is a growing need to promote interdisciplinary pedagogy in data literacy education. Dorsey et al. [10] noted that effective data literacy education requires teachers to develop both conceptual knowledge and pedagogical strategies. They advocate for context-based professional development, integration of data tasks across subjects, and leadership support to embed data literacy into the fabric of K–12 teaching. In the context of cross-curricular integration of teaching data literacy, Friedrich et al. [11] found that while statistical literacy is emphasized in mathematics, interdisciplinary integration remains limited. They highlight the importance of interdisciplinary teaching, adapted curricula, and improved professional development for both pre- and in-service teachers to further enhance data and statistical literacy in K–12 STEM education. In addition, Ghodoosi et al. [12] highlight the need for educators to adopt an interdisciplinary approach when designing data literacy education. Their findings advocate for providing students with opportunities to engage in real-world data projects, use data visualization tools, and develop both critical thinking and data interpretation skills.

Different approaches have been explored to teach data literacy, ranging from storytelling-based methods to hands-on programming and project-based learning with real-world data. At the tertiary level, a storytelling approach is used to teach data literacy. For example, Li et al. [13] introduced the OCEL.AI paradigm, a storytelling-based data science education model that significantly improved undergraduate students' data literacy by using the 5W+1H framework to contextualize data in real-world scenarios. Similarly, McDowell and Turk [14] highlighted the role of data storytelling in data literacy education, showing that it empowers students to become active interpreters and creators of data narratives, particularly around social issues. In contrast, Yamaguchi and Kuwana [15] proposed a structured Python-based teaching approach using real-world data, such as stock price analysis, to teach data literacy at the secondary school level. This approach emphasizes hands-on coding and demonstrates strong potential to enhance student engagement in data science while fostering interdisciplinary STEM education. Witte et al. [16] suggest that project-based learning using real-world

***CORRESPONDING AUTHOR | Lim Ming Hui** | ✉ minghui_lim@student.usm.my

data and current societal issues is an effective method to address the imbalance in how data literacy is taught, particularly the underemphasis on the planning and data collection stages. They recommend that future research focus on developing comprehensive teaching approaches that cover all components of data literacy, including those currently overlooked, such as data collection and problem formulation. In short, educators play a critical role in teaching data literacy. To help students acquire data literacy competencies, educators should adopt and adapt interdisciplinary teaching strategies that bridge mathematics and technology in context-rich, practical ways. Hence, this paper proposes an interdisciplinary, hands-on, inquiry-based learning approach for teaching data literacy to lower secondary students, inspired by the approach suggested by Yamaguchi and Kuwana [15]. It is important to remember that teaching data literacy goes beyond simply using software to perform calculations. Software tools should be seen as supports that aid in developing a deeper understanding of data and its real-world meaning. Thinking critically with data requires a structured investigative process, enabling students to solve problems, analyse information, and make evidence-based decisions. Two complementary frameworks that support this development are the PPDAC model [17] and the principles of Exploratory Data Analysis [18].

The PPDAC framework proposed by Wild & Pfannkuch [17] provides a structured approach to data analysis, consisting of five key stages and starting with the Problem stage, where the question of interest is clearly defined to guide the investigation. The Plan stage involves deciding how data will be collected and identifying the appropriate methods for gathering information. The Data stage focuses on collecting relevant data while ensuring its accuracy and reliability. This is followed by the Analysis stage, where the data is explored, analysed, and visualized to reveal patterns, trends, and relationships. Lastly, the Conclusion stage requires drawing meaningful insights from the analysis to support evidence-based conclusions. The term Exploratory Data Analysis (EDA) was first used by Tukey [18] to describe the process of using basic descriptive statistics and data visualization to get a quick overview of the data. EDA focuses on learning how to understand data by examining its structure, managing missing or duplicate values, and identifying outliers. This approach enables a better understanding of the data and serves as an essential step before performing statistical analysis [19]. The EDA technique is useful during the Analysis stage of the PPDAC framework. Embedding EDA within the PPDAC structure encourages students to move beyond surface-level observations and adopt an inquiry-based mindset when working with data. Students can create visuals such as histograms, box plots, scatter plots, and compute summary statistics to explore the dataset. These methods help uncover trends, detect outliers, and assess relationships among variables, thereby enhancing students' critical thinking about data.

Students must be exposed to suitable tools that support meaningful data exploration to develop their data literacy skills. Among the available tools, Python stands out as an accessible tool for learning data analysis, as it is one of the programming languages taught in basic computer science subjects [6–7]. At the foundational level, through the basic computer science subject, students learn essential Python programming concepts. These include simple control structures (conditional statements and loops) and fundamental operations on common data types such as integers, floating-point numbers, strings, and lists. Building on this foundation, essential Python libraries for data analysis, such as pandas, matplotlib, NumPy, and scikit-learn [20], can be introduced to students to help them learn and develop their data literacy skills. With the Panda's library, students can learn to create a new data frame, import data, filter rows based on conditions, extract specific columns, calculate summary statistics (mean, median, mode), and group or sort data. Visualization is another key aspect of data literacy and exploratory data analysis. Using matplotlib, students can create bar charts, line graphs, and histograms to make data insights clearer and more impactful. The NumPy library supports efficient numerical operations, helping students perform array-based calculations and mathematical functions that are essential for analysing large datasets. To extend their learning into simple predictive modelling, students can also use the scikit-learn library to perform linear regression easily, as it abstracts many complex underlying mathematical operations [21].

2. Methodology

An interdisciplinary approach is essential for fostering data literacy [10–12], as it enables students to connect mathematical concepts with computational tools and real-world contexts. Inspired by the approach proposed by Yamaguchi and Kuwana [15], two relatable examples are presented here: students' examination results and the relationship between income and expenditure. This guided, hands-on activity builds data-handling and visualization skills while introducing basic predictive modelling with simple linear regression. First, we analyse students' examination results. The general procedure involves importing or creating student data, calculating summary statistics, and visualizing the results to uncover meaningful insights. Using Python along with the pandas and matplotlib libraries, as coded in Figure 1, students can compute averages, identify the highest and lowest scores, and compare performance across genders or subjects. An example bar chart visualization of average subject scores by gender using a simple dataset is shown in Figure 2. Additionally, students can apply simple linear regression using the scikit-learn library to explore relationships between subjects, for example, predicting Science scores based on Math scores, as in Figure 3.

3. Results and Discussion

In this activity, students apply the PPDAC framework by starting with a clear Problem by exploring potential performance differences between genders and across subjects. During the Plan and Data stages, they organize and summarize the dataset using Exploratory Data Analysis (EDA) techniques such as calculating descriptive statistics and grouping data by categories. In the Analysis phase, students interpret both numerical outputs and visual plots to uncover patterns, such as which gender performed better and which subjects had higher or lower average scores.

```
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression
import numpy as np

# === PPDAC: Problem ===
# Do male and female students perform differently, and is there a trend between Math and Science scores?
# === PPDAC: Plan ===
# Perform data preprocessing, compute statistics, visualize results, and explore a basic linear regression.

# === Data Preprocessing ===
data = {
    'Name': ['Ali', 'Siti', 'John', 'Aisha', 'David', 'Nurul', 'Ahmad'],
    'Gender': ['M', 'F', 'M', 'F', 'M', 'F', 'M'],
    'Math': [80, 90, 75, 85, 70, 95, 88],
    'Science': [85, 95, 78, 88, 72, 97, 82],
    'English': [78, 92, 70, 89, 65, 94, 77]
}
df = pd.DataFrame(data)

# Check for missing data (basic preprocessing step)
print("\nMissing Data:\n", df.isnull().sum())

# === Statistical Analysis ===
summary_stats = df[['Math', 'Science', 'English']].describe()
print("\nSummary Statistics:\n", summary_stats)

avg_scores_by_gender = df.groupby('Gender')[['Math', 'Science', 'English']].mean()
print("\nAverage Scores by Gender:\n", avg_scores_by_gender)

# === Visualization: Average Scores by Gender ===
avg_scores_by_gender.plot(kind='bar', figsize=(8, 5))
plt.title('Average Subject Scores by Gender')
plt.xlabel('Gender')
plt.ylabel('Average Score')
plt.legend(title='Subject')
plt.grid(axis='y')
plt.tight_layout()
plt.show()

# === Linear Regression: Predict Science Score from Math Score ===
X = df[['Math']].values  # Independent variable
y = df['Science'].values  # Dependent variable

model = LinearRegression()
model.fit(X, y)

# Predict and plot regression line
predicted_science = model.predict(X)

plt.figure(figsize=(8, 5))
plt.scatter(df['Math'], df['Science'], color='blue', label='Actual Data')
plt.plot(df['Math'], predicted_science, color='red',
         label=f'Regression Line: y = {model.coef_[0]:.2f}x + {model.intercept_:.2f}')
plt.title('Linear Regression: Math vs Science Scores')
plt.xlabel('Math Score')
plt.ylabel('Science Score')
plt.legend()
plt.grid(True)
plt.tight_layout()
plt.show()

# === R² Score Calculation ===
r_squared = model.score(X, y)
print(f"\nRegression Equation: Science = {model.coef_[0]:.2f} * Math + {model.intercept_:.2f}")
print(f"R² Score (Goodness of Fit): {r_squared:.2f}")

# === Conclusion ===
# The regression line helps visualize the relationship between Math and Science scores.
# The R² score indicates how well the Math scores explain variations in Science scores.
# R² close to 1 means a strong linear relationship; closer to 0 means a weak relationship.
# Students learn how a straight-line equation (y = mx + c) represents this model and
# how to interpret R² as a measure of prediction accuracy.
```

Figure 1. Code snippet for analysing students' examination results using Python

They also apply basic linear regression to investigate relationships between subjects, for example, predicting Science scores based on Math scores. Finally, in the Conclusion stage, students reflect on their findings and engage in inquiry,

reasoning, and evidence-based discussions, strengthening their critical thinking, data literacy, and understanding of simple predictive modelling. Next, we can analyse personal income and expenditure data to explore real-world financial behaviour. Using Python, along with the pandas, matplotlib, and scikit-learn libraries, as coded in Figure 4, students ban calculate summary statistics, visualize the relationship between income and spending, and apply linear regression to model and predict expenditure based on income. Summary statistics data showing the distribution of monthly income and expenditure, including measures of central tendency and variability, are given in Figure 5. Figure 6 shows the relationship between monthly income and expenditure using the sample dataset. Through this activity, students not only strengthen their data handling and visualization skills but also gain insights into basic economic concepts, such as how income levels influence spending habits. This exercise serves as a practical introduction to predictive modelling and financial literacy using real data scenarios.

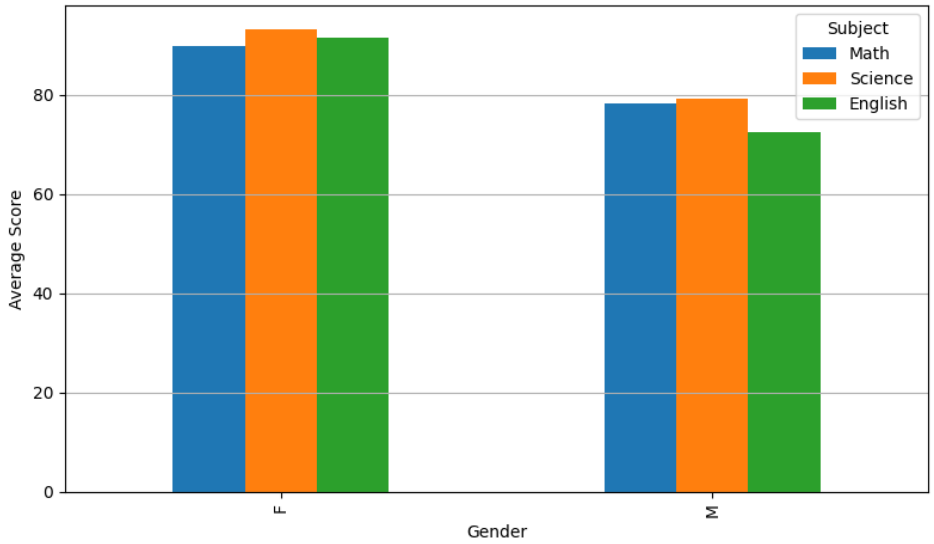

Figure 2. Example bar chart visualization of average subject scores by gender using the sample dataset

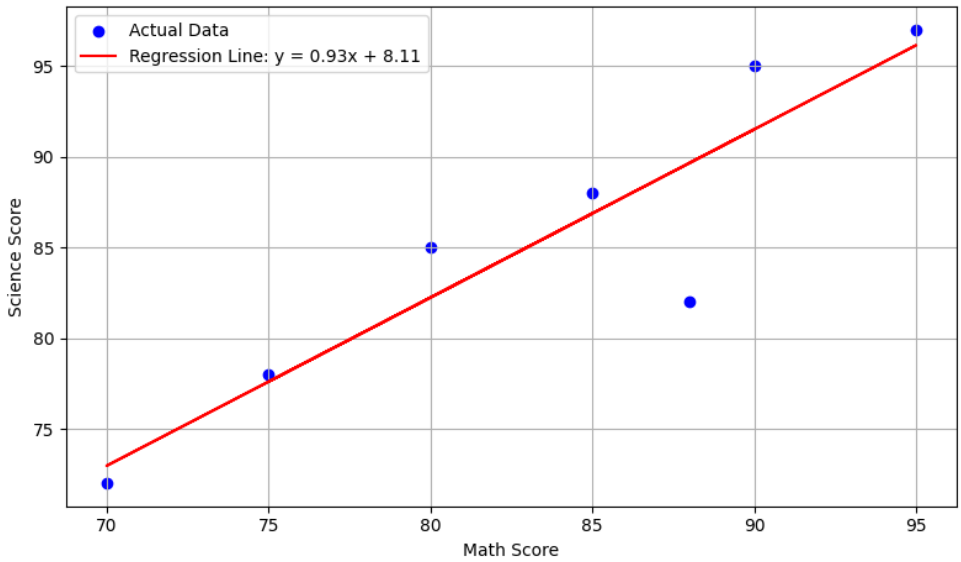

Figure 3. Example linear regression plot modelling the relationship between Math and Science scores using a sample dataset

In this example, students apply the PPDAC framework to explore economic data, focusing on the relationship between personal income and monthly expenditure. Beginning with the *Problem* stage, they investigate whether higher income leads to increased spending. In the *Plan* and *Data* stages, students work with a simulated dataset and preprocess the data to ensure it is clean and complete. Using the EDA technique, they calculate descriptive statistics and visualize the relationship between income and expenditure through scatter plots. In the *Analysis* phase, students apply linear regression to model and predict expenditure as a function of income, interpreting the regression equation and evaluating the model's accuracy using the R^2 score. Finally, during the *Conclusion* stage, students reflect on their findings and discuss how income influences spending behaviour, helping them develop critical thinking, data literacy, and an understanding of basic economic concepts such as consumption patterns and financial planning.

```
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression
import numpy as np
# === PPDAC: Problem ===
# Question: How does a person's income affect their monthly expenditure?
# Hypothesis: Higher income leads to higher monthly spending.

# === PPDAC: Plan ===
# - Collect or simulate data on monthly income and expenditure.
# - Perform data preprocessing (check for missing data).
# - Analyze data using descriptive statistics.
# - Visualize the relationship.
# - Apply linear regression to model and predict expenditure based on income.

# === Data Preprocessing ===
# Simulated dataset of people's income and expenditure
data = {
    'Person': ['Ali', 'Siti', 'John', 'Aisha', 'David', 'Nurul', 'Ahmad'],
    'Monthly_Income': [2000, 4000, 1800, 3500, 1500, 5000, 3000],  # Example: Monthly salary in MYR
    'Monthly_Expenditure': [1500, 3000, 1300, 2700, 1200, 4000, 2500]  # Monthly spending in MYR
}
# Create DataFrame
df = pd.DataFrame(data)

# Check for missing values to ensure data is clean
print("\nMissing Data:\n", df.isnull().sum())

# === PPDAC: Analysis - Step 1: Descriptive Statistics ===
# Basic statistics to understand the data distribution
summary_stats = df[['Monthly_Income', 'Monthly_Expenditure']].describe()
print("\nSummary Statistics:\n", summary_stats)

# === Analysis - Step 2: Visualization ===
# Create a scatter plot to visualize the relationship between income and expenditure
plt.figure(figsize=(8, 5))
plt.scatter(df['Monthly_Income'], df['Monthly_Expenditure'], color='blue', label='Actual Data')
plt.title('Income vs Expenditure')
plt.xlabel('Monthly Income (MYR)')
plt.ylabel('Monthly Expenditure (MYR)')
plt.grid(True)

# === Analysis - Step 3: Linear Regression ===
# Independent variable (X) is Monthly Income
X = df[['Monthly_Income']].values

# Dependent variable (y) is Monthly Expenditure
y = df['Monthly_Expenditure'].values

# Create and train the Linear Regression model
model = LinearRegression()
model.fit(X, y)

# Predict expenditure values using the trained model
predicted_exp = model.predict(X)

# Plot the regression line on the scatter plot
plt.plot(df['Monthly_Income'], predicted_exp, color='red',
         label=f'Regression Line: y = {model.coef_[0]:. 2f}x + {model.intercept_:.2f}')
plt.legend()
plt.tight_layout()
plt.show()

# === PPDAC: Conclusion ===
# Display the regression equation: y = mx + c
print(f"\nRegression Equation: Expenditure = {model.coef_[0]:.2f} * Income + {model.intercept_:.2f}")
print("Interpretation: For every additional 1 MYR increase in income, expenditure increases by approximately "
      f"{model.coef_[0]:.2f} MYR on average.")

# Calculate and display R² score (Goodness of Fit)
r_squared = model.score(X, y)
print(f"R² Score (Goodness of Fit): {r_squared:.2f}")

# Higher R² indicates a stronger linear relationship.
# This helps students understand how well income explains expenditure variation.
```

Figure 4. Code snippet for analysing the relationship between income and expenditure using Python

```
Summary Statistics:
       Monthly_Income  Monthly_Expenditure
count        7.000000             7.000000
mean      2971.428571          2314.285714
std       1286.745618          1035.098339
min       1500.000000          1200.000000
25%       1900.000000          1400.000000
50%       3000.000000          2500.000000
75%       3750.000000          2850.000000
max       5000.000000          4000.000000
```

Figure 5. Summary statistics data showing the distribution of monthly income and expenditure, including measures of central tendency and variability

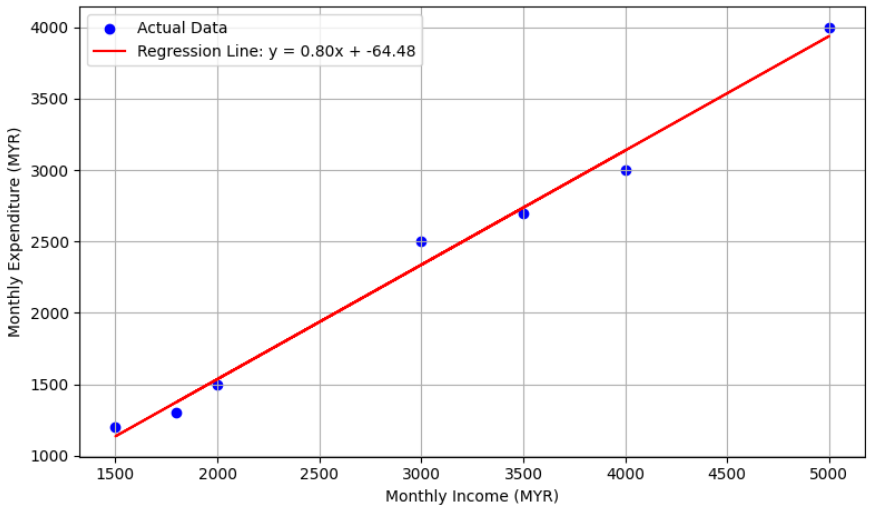

Figure 6. Example linear regression plot modelling the relationship between monthly income and expenditure using sample dataset

4. Conclusions

Integrating the basic computer science and Mathematics curricula offers a practical approach to promote and develop students' data literacy skills. This interdisciplinary approach encourages students to apply their mathematical knowledge and computing skills in real-world data scenarios, fostering deeper understanding and practical problem-solving abilities. Moreover, it aligns closely with the objectives of the Digital Education Policy and supports the goals outlined in the OECD Core Foundations For 2030. However, despite its promise, implementing an interdisciplinary approach can pose challenges. Many teachers may not yet feel adequately prepared to teach interdisciplinary content, and time constraints in classroom can limit opportunities for cross-curricular integration. To move forward, policymakers could consider piloting interdisciplinary modules that combine data analysis with coding in selected classrooms, supported by specially curated datasets and step-by-step lesson guides to ensure consistent and effective implementation. These classroom-based studies can serve not only as instructional prototypes but also as a means of gathering feedback on student learning outcomes and teacher experiences. In parallel, targeted professional development and teacher training workshops will be essential to build educators' confidence and pedagogical skills in delivering integrated ASK and Mathematics content. Together, these strategies can help build the practical feasibility of interdisciplinary data literacy instruction within the Malaysian lower secondary education context.

Acknowledgement

The author would like to express their gratitude to Universiti Sains Malaysia for the support.

Funding

This study was not supported by any grants from funding bodies in the public, private, or not-for-profit sectors.

Declaration of Competing Interest

The authors declare no conflict of interest.

CRediT Authorship Contribution Statement
Lim Ming Hui (Conceptualisation; Methodology; Writing – original draft; Formal analysis; Writing – review & editing)

Availability of the Data and Materials
The data used to support the findings of this study are included within the article.

Ethical Declaration
No artificial intelligence tools were used in the preparation of this manuscript. All content was developed manually by the authors. This study did not involve human participants or animals. Ethical approval was therefore not required.

Generative Artificial Intelligence Declarations
The authors claim that artificially intelligent-assisted technologies, such as generative AI, were not used to generate content, ideas, or theories. We have just utilised AI to enhance readability and refine the language. This was used with extreme human control and oversight. The authors take full responsibility for reviewing and approving the content.

References
[1] Wolff A, Gooch D, Cavero Montaner JJ, Rashid U, Kortuem G. Creating an understanding of data literacy for a data-driven society. The Journal of Community Informatics. 2016;12(3):9–26.
[2] Bahagian Pembangunan Kurikulum. Matematik: Dokumen Standard Kurikulum dan Pentaksiran Tingkatan 1. Malaysia: Kementerian Pendidikan Malaysia; 2015.
[3] Bahagian Pembangunan Kurikulum. Matematik: Dokumen Standard Kurikulum dan Pentaksiran Tingkatan 2. Malaysia: Kementerian Pendidikan Malaysia; 2016.
[4] Bahagian Pembangunan Kurikulum. Matematik: Dokumen Standard Kurikulum dan Pentaksiran Tingkatan 3. Malaysia: Kementerian Pendidikan Malaysia; 2017.
[5] Bahagian Pembangunan Kurikulum. Asas Sains Komputer: Dokumen Standard Kurikulum dan Pentaksiran Tingkatan 1. Malaysia: Kementerian Pendidikan Malaysia; 2015.
[6] Bahagian Pembangunan Kurikulum. Asas Sains Komputer: Dokumen Standard Kurikulum dan Pentaksiran Tingkatan 2. Malaysia: Kementerian Pendidikan Malaysia; 2016.
[7] Bahagian Pembangunan Kurikulum. Asas Sains Komputer: Dokumen Standard Kurikulum dan Pentaksiran Tingkatan 3. Malaysia: Kementerian Pendidikan Malaysia; 2017.
[8] Ministry of Education Malaysia. Digital Education Policy. Malaysia: Ministry of Education Malaysia; 2023.
[9] OECD. OECD learning compass 2030: A series of concept notes. OECD Publishing; 2019. Retrieved from https://www.oecd.org/education/2030-project/; 10 March 2026.
[10] Dorsey C, Sagrans J, Yaneva K, O'Brie D, Collins I et al. Integrating data literacy into K–12 education. Harvard Data Science Review. 2025;7(2).
[11] Friedrich A, Schreiter S, Vogel M, Becker-Genschow S, Brünken R et al. What shapes statistical and data literacy research in K-12 STEM education? A systematic review of metrics and instructional strategies. International Journal of STEM Education. 2024;11:58.
[12] Ghodoosi B, Torrisi-Steele G, West T, Heidari M. Perceptions of data literacy and data literacy education. Journal of Librarianship and Information Science. 2024:1-11.
[13] Li Y, Wang Y, Lee Y, Chen H, Petri AN, Cha T. Teaching data science through storytelling: Improving undergraduate data literacy. Thinking Skills and Creativity. 2023;48:101311.
[14] McDowell K, Turk MJ. Teaching data storytelling as data literacy. Information and Learning Sciences. 2024;125(5/6):321-345.
[15] Yamaguchi K, Kuwana A. Development of learning materials for machine learning utilizing Python in senior high school. In: The 7th International Conference on Technology and Social Science 2023. 2023.
[16] Witte V, Schwering A, Frischemeier D. Strengthening data literacy in K-12 education: A scoping review. Education Sciences. 2025;15(1):25.
[17] Wild CJ, Pfannkuch M. Statistical thinking in empirical enquiry. International Statistical Review. 1999;67(3):223-248.
[18] Tukey JW. Exploratory data analysis. Reading, MA: Addison-Wesley; 1977.
[19] Smith-Miles K. Exploratory data analysis. In: Lovric M Ed. International Encyclopedia of Statistical Science. Berlin: Springer; 2011. p. 486-488.
[20] Nelli F. Python data analytics with Pandas, NumPy, and Matplotlib. New York: Apress; 2023.
[21] Wadsworth F, Blaney J, Springsteen M, Coburn B, Khanal N et al. Frameworks and challenges for implementing machine learning curriculum in secondary education. International Journal of Technology in Education and Science. 2024;8(1):164-181.