Biratnagar International College, Biratnagar

Concepts and Technologies of AI

5CS037

Assignment - I - Statistical Interpretation and Exploratory Data Analysis.

Analysis of the Human Development Index (HDI):
A Data - Driven Exploration of Global
and Regional Development Patterns.

December 8, 2025

- University Module Leader: Siman Giri (siman.giri@hearaldcollege.edu.n
- Biratnagar Faculty: Ayush Regmi.

Contents

1 Assignment Details and Submission Guidelines

1.1 Assignment Details:

Due	Marks	Submission
January - 10	10	A report and completely rendered Jupyter notebook. **For details see section 5 (page no - 12).**

1.2 Plagiarism and AI Generated Content

Plagiarism of more than 20% and any AI-generated content found in the report will be reported for academic misconduct. Thus, we highly encourage you to submit your original work.

1.3 Submission Guidelines:

- This assignment must be completed individually.
- The data set used for this assignment can be downloaded from the shared drive.
  {`Only use the provided and assigned dataset where - ever instructed`}
- What to Submit?
  - You are expected to submit a report of 2-4 pages based on the task and exercise requested along with the code base.
  - For Code:
    1. All solutions - Code must be written in the Jupyter notebook.
    2. Our recommendation - `Google Collaboration` .
    3. All codes must be pushed to `GitHub` before the deadlines.
  - For report:
    1. Please follow the APA format; for a sample, see Section 5 of this document.
  - `For More Details on Report and Code Guidelines, Please follow Section 5 carefully`
  - Where to submit?
    As instructed by your instructor.

The Final Date for submission is: `10 January.`

1.3.1 Naming Conventions:

You are supposed to strictly follow the naming conventions, and any file that does not follow the naming conventions will be marked as "0".
`File Name: WLVID_FullName(firstname+last).ipynb`

2 Assignment Overview

2.1 About Assignment:

In this assignment, you will utilize the advanced features of the Pandas library to apply the knowledge gained from Workshops 1, 2, and 3 to a more comprehensive real-world dataset.This assignment is supposed to introduce you to various parts of the data science process involving being able to answer questions about your data, how to visualize your data.Designed to help you prepare for your final project, this assignment provides broad exposure to different aspects of data analysis. While it includes multiple sections, each task is relatively small and manageable, allowing you to gain practical experience across a wide range of techniques.

2.2 Cautions!!!

In this assignment, you will perform a statistical interpretation and exploratory data analysis for a small dataset and provide a rigorous rationale for your choices. We will determine scores by judging both the soundness of your **design**, the quality of the **write-up(report)** and your ability to answer the question during **viva**. Here are examples of aspects that may lead to **point deductions:**

- Use of misleading, unnecessary, or unmotivated graphic elements.
- Missing title of the chart, axis labels, or data transformation description.
- Missing or incomplete design rationale in the paper.
- Ineffective encoding for your stated goal (e.g., distracting colors, improper data transformation).

Tools and Python Package which can be used for this assignments (listed but not limited to):

1. **Pandas library(pd)**
2. **Numpy library(np)**
3. **Matplotlib library(plt)**
4. **Seaborn library(sns)**

2.3 Learning Outcomes:

Learning outcomes can be following but not limited to:

1. Work with basic Python data structures.
2. Use Pandas as the primary tool to process structured data in Python with CSV files,
   (a) Handle edge cases appropriately, including addressing missing values/data.
   (b) Practice user-friendly error-handling.
3. Use pandas, matplotlib and seaborn library to produce various plots for visualization or to investigate a specific phenomenon,
   (a) Review the library documentation and example code to learn how to create more complex plots.

2.4 Dataset:

The dataset provided for this assignment is:

"Human_Development_Index_Dataset.csv" by Lucas Yukiolmafuko.

Please use this specific dataset, as it has been modified to suit the requirements of this assignment.

2.4.1 About the Dataset:

The Human Development Index (HDI) measures a country's achievements in key dimensions of human development. The dataset spans 1990–2022 and includes all countries for which HDI is reported. The columns represent:

- Health: Life expectancy at birth
- Education: Mean years of adult schooling and expected years of schooling for children
- Standard of Living: Gross national income (GNI) per capita

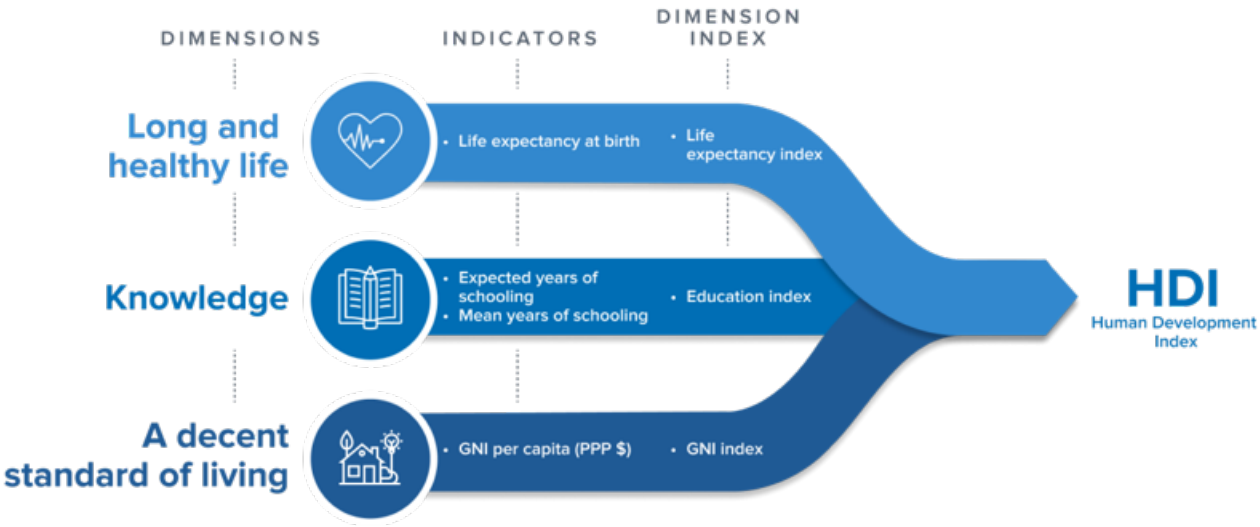

Figure 1: Human Development Index.

The HDI is calculated as the geometric mean of these three indices. While income influences HDI, differences in health and education explain why countries with similar GNI can have different human development outcomes. This dataset enables analysis of trends over time, regional comparisons, and identification of countries that deviate from expected development patterns.

3 Problem 1

Basic Data Exploration & Trend Visualization

3.1 Problem - 1A - Single Year HDI Exploration (Latest Year: 2022)

Objective:

Explore the HDI dataset for the latest available year (2022) to practice basic EDA techniques.

Tasks:

Complete all the Following Tasks:

1. **Extract Latest Year:**
   - Identify unique years in the dataset.
   - Filter the dataset to include only observations from the year 2022.
   - Save the filtered dataframe as `hdi_2022_df` (used for all subsequent tasks in Problem 1A).
2. **Data Exploration:**
   - Display the first 10 rows of the 2022 dataset.
   - Count the number of rows and columns.
   - List all column names and their data types.
3. **Missing Values & Data Cleaning:**
   - Check for missing values in each column and report total counts.
   - Inspect dataset for:
     - numeric columns stored as text,
     - inconsistent or misspelled country names,
     - duplicate rows,
     - special characters (e.g., "–") representing missing data.
   - Apply necessary cleaning steps:
     - convert data types where needed,
     - remove duplicates,
     - handle missing values (drop or impute; justify your choice).
4. **Basic Statistics:**
   - Compute the mean, median, and standard deviation of HDI for the year 2022.
   - Identify the country with the highest HDI in 2022.
   - Identify the country with the lowest HDI in 2022.

5. **Filtering and Sorting:**
   - Filter countries with HDI {"hdi"} greater than 0.800.
   - Sort this filtered dataset by Gross National Income (GNI) per Capita {"gross_inc_percap"} in descending order.
   - Display the top 10 countries.
6. **Adding HDI Category Column:**
   - Create a new column HDI_Category that classifies each country into one of the four official Human Development Index groups. The classification should be based on the HDI value for the year 2022. Use the following categories and thresholds defined by the United Nations Development Programme (UNDP):

HDI Category	HDI Range (hdi)
Low	< 0.550
Medium	$0.550 - 0.699$
High	$0.700 - 0.799$
Very	≥ 0.800

   After creating this new column:
   - verify that all countries are classified correctly,
   - ensure the updated dataframe includes the new category column.
   - **Save the final dataframe as HDI_category_added.csv and include this file in your final submission.**

3.2 Problem - 1B - HDI Visualization and Trend Analysis (2020 – 2022)

Objective:

Analyze multi-year HDI patterns (2020, 2021, and 2022) to explore temporal changes, regional differences, and trends across countries.

Tasks:

Complete all the Following Tasks:

1. **Data Extraction and Saving:**
   - Filter the dataset to include only the years 2020, 2021, and 2022.
   - Save the filtered dataset as `HDI_problem1B.csv`.
   - Use this cleaned dataset for all subsequent tasks in Problem 1B.
2. **Data Cleaning:**
   - Check for missing values in the following essential columns:
     - `hdi`
     - `country`
     - `year`
   - Identify and address the following issues:
     - missing or null values,
     - inconsistent or misspelled country names,
     - duplicate rows,
     - numeric columns stored as text or containing non-numeric symbols (e.g., "–").
   - Apply and justify cleaning steps, including:
     - handling missing values (dropping or imputing),
     - converting data types appropriately,
     - removing duplicate entries,
     - ensuring consistent naming conventions for countries and years.
   - All cleaning decisions must be clearly justified in the final report.
3. **Visualization Tasks:**
   - **A. Line Chart — HDI Trend (Country-Level):**
     - Select any five countries (or five countries from a region of your choice).
     - Plot HDI values for each country across the years 2020, 2021, and 2022.
     - Ensure the chart includes appropriate axis labels, a legend, and an informative caption.

- **B. Generate Visualizations:**
  - **Bar Chart: Average HDI by Region (2020–2022)**
    - Group the dataset by `Region` and `Year`.
    - Compute the mean HDI for each region-year pair.
    - Plot a bar chart comparing average HDI across regions for each year.
    - Label axes clearly and include a descriptive title.
  - **Box Plot: HDI Distribution for 2020, 2021, and 2022**
    - Filter the dataset for the years 2020, 2021, and 2022.
    - Create a box plot showing HDI spread for each of the three years.
    - Include titles and axis labels.
    - Comment briefly on distribution differences.
  - **Scatter Plot: HDI vs. GNI per Capita**
    - Check if the dataset contains a column for `GNI per Capita`.
    - If available:
      - Create a scatter plot using HDI as the dependent variable.
      - Use GNI per Capita on the x-axis.
      - Add a regression line (optional).
    - If the variable is missing, write:
      `"GNI per Capita variable not available in the dataset."`
- Provide brief interpretations of all visualizations, describing major trends, patterns, and anomalies.

4. **Short Analysis Questions:**
   - Which countries show the greatest improvement in HDI from 2020 to 2022?
   - Did any countries experience a decline in HDI? Provide possible reasons.
   - Which region has the highest and lowest average HDI across these three years?
   - Discuss how global events (e.g., the COVID-19 pandemic) may have affected HDI trends during this period.

4 Problem 2

Advanced HDI Exploration

Objective:

Perform advanced analysis of HDI data, focusing on South Asian countries, composite metrics, outlier detection, metric relationships, and gap analysis.

Tasks:

Complete all the following tasks:

1. **Create South Asia Subset:**
   - Define the list of South Asian countries: `["Afghanistan", "Bangladesh", "Bhutan", "India", "Maldives", "Nepal", "Pakistan", "Sri Lanka"]`.
   - Filter the HDI dataset to include only these countries.
   - Save the filtered dataset as `HDI_SouthAsia.csv` and include this file in the final submission.
2. **Composite Development Score:**
   - Create a new metric called `Composite_Score` using the formula:

     $$\text{Composite Score} = 0.30 \times \text{Life Expectancy Index} + 0.30 \times \text{GNI per Capita Index}\}$$

     Here: Life Expectancy Index $\rightarrow$ `"life_expectancy"` and GNI per Capita Index $\rightarrow$ `"gross_inc_percap"`
   - Rank South Asian countries based on `Composite_Score`.
   - Plot the top 5 countries in a horizontal bar chart.
   - Compare the ranking of countries by `Composite_Score` with their HDI ranking and discuss any differences.
3. **Outlier Detection:**
   - Detect outliers in HDI and GNI per Capita using the 1.5 $\times$ IQR rule.
   - Create a scatter plot of GNI per Capita vs HDI, highlighting the outliers in a different color.
   - Discuss why the identified countries stand out as outliers.
4. **Exploring Metric Relationships:**
   - Select two HDI components (e.g., Gender Development Index {`"gender_development"`} and Life Expectancy Index {`"life_expectancy"`}).
   - Compute Pearson correlation of each metric with HDI.
   - Create scatter plots with trendlines to visualize the relationships.
   - Discuss:
     - Which metric is most strongly related to HDI and shows the weakest relationship with HDI.

5. **Gap Analysis:**
   - Create a new metric:

$$\text{GNI\_HDI\_Gap} = \texttt{"gross\_inc\_percap"} - \texttt{"hdi"}$$

   - Rank South Asian countries by `GNI_HDI_Gap` in descending and ascending order.
   - Plot the top 3 positive gaps and top 3 negative gaps.
   - Discuss the implications of the gap, e.g., cases where GNI is high but HDI is lower than expected.

5 Problem 3

Comparative Regional Analysis: South Asia vs Middle East

Objective:

Perform a comparative analysis of HDI and related metrics between South Asia and the Middle East using the 2020–2022 dataset from Problem 1B.

Tasks:

Complete all the following tasks:

1. **Create Middle East Subset:**
   - Define the list of Middle East countries: `["Bahrain", "Iran", "Iraq", "Israel", "Jordan", "Kuwait", "Lebanon", "Oman", "Palestine", "Qatar", "Saudi Arabia", "Syria", "United Arab Emirates", "Yemen"]`.
   - Filter the dataset from Problem 1B (`HDI_problem1B.csv`) to create subsets for South Asia and Middle East.
   - Save these subsets as `HDI_SouthAsia_2020_2022.csv` and `HDI_MiddleEast_2020_2022.csv` for use in subsequent tasks.
2. **Descriptive Statistics:**
   - Compute the mean and standard deviation of HDI for each region (South Asia vs Middle East) across 2020–2022.
   - Identify which region performs better on average.
3. **Top and Bottom Performers:**
   - Identify the top 3 and bottom 3 countries in each region based on HDI.
   - Create a bar chart comparing these top and bottom performers across the two regions.
4. **Metric Comparisons:**
   - Compare the following metrics across regions using grouped bar charts:
     - Gender Development Index {`"gender_development"`}
     - Life Expectancy Index {`"life_expectancy"`}
     - GNI per Capita Index {`"gross_inc_percap"`}
   - Identify which metric shows the greatest disparity between regions.
5. **HDI Disparity:**
   - Compute the range (max – min) of HDI for each region.
   - Compute the coefficient of variation (CV = std/mean) for HDI.

- Identify which region exhibits more variation in HDI.

6. **Correlation Analysis:**
   - For each region, compute correlations of HDI with:
     - Gender Development Index
     - Life Expectancy Index
   - Create scatter plots with trendlines for each correlation.
   - Interpret the strength and direction of these relationships.
7. **Outlier Detection:**
   - Detect outliers in HDI {`"hdi"`} and GNI per Capita {`"gross_inc_percap"`} for each region using the 1.5 $\times$ IQR rule.
   - Create scatter plots highlighting outliers in a different color.
   - Discuss the significance of these outliers.

6 Guidelines for HDI Data Analysis Report and Jupyter Notebook

1. General Instructions

- Submit both the Jupyter notebook (`.ipynb`) and a PDF report.
- Use the datasets provided:
  - Full HDI dataset (latest version)
  - Processed datasets from Problem 1B, Problem 2, and Problem 3 as applicable.
- Ensure that all plots, tables, and outputs in the notebook are clearly labeled.
- Your report should summarize findings, while the notebook should demonstrate the step-by-step computations and visualizations.

2. Jupyter Notebook Requirements

- Organize the notebook according to the problem structure:
  1. Problem 1A: Single-Year HDI Exploration
  2. Problem 1B: HDI Trend Analysis
  3. Problem 2: Advanced HDI Exploration
  4. Problem 3: Comparative Regional Analysis
- **Code and Output:**
  - Clearly label each task (e.g., # Task 1: Extract Latest Year).
  - Include intermediate results where relevant (head of dataframe, missing value counts, etc.).
- **Visualization:**
  - Use appropriate titles, axis labels, legends, and captions.
  - Highlight insights directly below plots.
- **Data Cleaning:**
  - Document any cleaning steps (e.g., missing value handling, duplicates, type conversions) with markdown comments explaining your choice.
- **Analysis and Calculations:**
  - Show all formulas, computations, and ranking steps (e.g., Composite Score formula in Problem 2).
  - Clearly identify outliers, gaps, and correlations.

6.1 Report Guidelines & or Requirements

Students are expected to submit both a Jupyter notebook and a written report. The report should summarize the analysis, include key outputs, visualizations, and explanations, and demonstrate a clear understanding of the tasks outlined in Problems 1A, 1B, 2, and 3.

6.1.1 Structure of the Report

The report should be organized as follows:

1. **Title Page:**
   - Title of the project
   - Student name, ID
   - Course name
   - Date of submission
2. **Table of Contents:**
   - Automatically generated or manually listed sections
3. **Introduction:**
   - Brief overview of HDI
   - Objectives of the analysis
   - Scope of the report
4. **Problem-wise Analysis:**
   - **Problem 1A – Single Year HDI Exploration**
     - Methods / Approach
     - Key results
     - Visualizations and tables
     - Interpretation and discussion
   - **Problem 1B – HDI Trend Analysis (2020–2022)**
     - Methods / Approach
     - Key results
     - Visualizations and tables
     - Interpretation and discussion
   - **Problem 2 – Advanced HDI Exploration**
     - Methods / Approach
     - Key results
     - Visualizations and tables

  - Interpretation and discussion
- **Problem 3 – Comparative Regional Analysis: South Asia vs Middle East**
  - Methods / Approach
  - Key results
  - Visualizations and tables
  - Interpretation and discussion

5. **Conclusion:**
   - Summary of findings
   - Insights about HDI trends and disparities
   - Limitations (if any)
   - Recommendations or implications
6. **References:**
   - Cite any data sources, articles, or references used
7. **Appendix (Optional):**
   - Any additional plots, tables, or calculations not included in the main report

6.1.2 Including Jupyter Notebook Outputs

1. **Tables and Metrics:**
   - Include key tables such as top/bottom countries, summary statistics, Composite Scores, and GNI_HDI gaps.
   - Present tables neatly in the report (either via screenshots or using LaTeX table formatting).
2. **Plots and Visualizations:**
   - Include charts such as line plots, bar charts, scatter plots, and box plots.
   - Export plots from the notebook using `plt.savefig()` or an equivalent method to maintain high quality.
   - Ensure all figures are numbered and include captions (e.g., Figure 1, Figure 2, etc.).
3. **Explanation:**
   - For every table or figure included, provide a brief interpretation:
     - What does it show?
     - What trend, pattern, or insight is visible?
     - How does it answer the corresponding problem task?
   - Reference figures and tables in the text (e.g., "Figure 3 shows the top 5 South Asian countries by Composite Score").

4. **Notebook vs Report:**
   - Only include key results and visuals in the report; avoid pasting the entire notebook.
   - The notebook submission should include all code, intermediate outputs, and step-by-step computations.

6.2 General Guidelines

1. Use clear and concise language throughout the report and notebook.
2. Use proper units and labels for all metrics, tables, and figures.
3. Follow a consistent formatting style for headings, figures, and tables.
4. Ensure that all data cleaning steps and decisions are documented and justified in the notebook and, if necessary, summarized in the report.
5. Submission Requirements:
   - Jupyter Notebook (`.ipynb`)
   - Final Report (`.pdf` or `.docx`)
   - Any CSV files generated during the analysis (e.g., `HDI_category_added.csv`, South Asia subset CSV)

7 Evaluation Criteria

Assessment will also include an oral viva. Following submission, students are required to participate in the viva and defend their work when requested by the instructor. Failure to attend or adequately defend the submitted work will result in an automatic score of zero for this component of the Final Assignment.

Criteria	**Weight**
Notebook organization and clarity	20%
Correctness of data cleaning and processing	15%
Accuracy of calculations/statistics	20%
Quality of visualizations (labels, captions, insights)	15%
Report writing: clarity, interpretation, structure	20%
Insights, discussion, and interpretation of results	10%

Table 1: Evaluation Criteria for HDI Data Analysis Project