Herald College, Kathmandu

Concepts and Technologies of AI

5CS037

Assignment-1 - Statisitical Interpretation and Exploratory Data Analysis

`Analysis of the World Happiness Report : A Data-Driven Exploration of Global and Regional Trends.`

December 02, 2024

Contents

1 Assignment Details and Submission Guidelines

1.1 Assignment Details:

Due	Marks	Submission
December-20.	10	2-4 page report and completely rendered notebook for details see page no. 4

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 3 of this document.
  - Where to submit?
    Designated portal opened on Canvas or as instructed by your instructor.

The Final Date for submission is: `20 December.`

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`
`Example: 00000_ABC Sharma.ipynb {For Code - Notebook.}`
`Example: 00000_ABC Sharma.pdf {For Report.}`

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 encodings 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 Ptyhon 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:

`"World Happiness Report.csv"`

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

2.4.1 About a Dataset:

The World Happiness Report is a key survey assessing global happiness. Happiness scores and rankings are based on data from the Gallup World Poll. The columns following the happiness score indicate how six factors

- economic production, social support, life expectancy, freedom, absence of corruption, and generosity

contribute to life evaluations in each country compared to Dystopia, a hypothetical country with the lowest global averages for these factors. While these factors do not affect the total score, they help explain differences in country rankings.

3 Tasks - To - Do:

Please Solve all the Problems as instructed:

3.1 Problem - 1: Getting Started with Data Exploration - Some Warm up Exercises:

1. Data Exploration and Understanding:

- Dataset Overview:
  1. Load the dataset and display the first 10 rows.
  2. Identify the number of rows and columns in the dataset.
  3. List all the columns and their data types.
- Basic Statistics:
  1. Calculate the mean, median, and standard deviation for the Score column.
  2. Identify the country with the highest and lowest happiness scores.
- Missing Values:
  1. Check if there are any missing values in the dataset. If so, display the total count for each column.
- Filtering and Sorting:
  1. Filter the dataset to show only the countries with a Score greater than 7.5.
  2. For the filtered dataset - Sort the dataset by GDP per Capita in descending order and display the top 10 rows.

- Adding New Columns:
  1. Create a new column called `Happiness_Category` that categorizes countries into three categories based on their Score:

$$\texttt{Low} - (\mathbf{Score} < 4)$$

$$\texttt{Medium} - (4 \leq \mathbf{Score} \leq 6)$$

$$\texttt{High} - (\mathbf{Score} > 6)$$

2. Data Visualizations:

- **Bar Plot:** Plot the top 10 happiest countries by Score using a bar chart.
- **Line Plot:** Plot the top 10 unhappiest countries by Score using a Line chart.
- **Plot a histogram** for the Score column to show its distribution and also interpret.
- **Scatter Plot:** Plot a scatter plot between GDP per Capita and Score to visualize their relationship.

3.2 Problem - 2 - Some Advance Data Exploration Task:

Task - 1 - Setup Task - Preparing the South-Asia Dataset:

Steps:

1. Define the countries in South Asia with a list for example:

```
south_asian_countries = ["Afghanistan", "Bangladesh", "Bhutan", "India",
                "Maldives", "Nepal", "Pakistan", "Srilanka"]
```

2. Use the list from step - 1 to filtered the dataset {i.e. filtered out matching dataset from list.}
3. Save the filtered dataframe as separate CSV files for future use.

Task - 2 - Composite Score Ranking:

Tasks:

1. Using the SouthAsia DataFrame, create a new column called Composite Score that combines the following metrics:

$$\texttt{Composite Score} = 0.40 \times \texttt{GDP per Capita} + 0.30 \times \texttt{Social Support} + 0.30 \times \texttt{Healthy Life Expectancy}$$

2. Rank the South Asian countries based on the Composite Score in descending order.
3. Visualize the top 5 countries using a horizontal bar chart showing the Composite Score.
4. Discuss whether the rankings based on the Composite Score align with the original Score - support your discussion with some visualization plot.

Task - 3 - Outlier Detection:

Tasks:

1. Identify outlier countries in South Asia based on their Score and GDP per Capita.
2. Define outliers using the `1.5 × IQR rule`.
3. Create a scatter plot with GDP per Capita on the x-axis and Score on the y-axis, highlighting outliers in a different color.
4. Discuss the characteristics of these outliers and their potential impact on regional averages.

Task - 4 - Exploring Trends Across Metrics:

Tasks:

1. Choose two metrics (e.g., Freedom to Make Life Choices and Generosity) and calculate their correlation {pearson correlation} with the Score for South Asian countries.
2. Create scatter plots with trendlines for these metrics against the Score.
3. Identify and discuss the strongest and weakest relationships between these metrics and the Score for South Asian countries.

Task - 5 - Gap Analysis:

Tasks:

1. Add a new column, `GDP-Score Gap`, which is the difference between GDP per Capita and the Score for each South Asian country.
2. Rank the South Asian countries by this gap in both ascending and descending order.
3. Highlight the top 3 countries with the largest positive and negative gaps using a bar chart.
4. Analyze the reasons behind these gaps and their implications for South Asian countries.

3.3 Problem - 3 - Comparative Analysis:

Task - 1 - Setup Task - Preparing the Middle Eastern Dataset:

Tasks:

1. Similar in Task - 1 of Problem 2 create a dataframe from middle eastern countries. For hint use the following list:

```
middle_east_countries = [ "Bahrain", "Iran", "Iraq", "Israel", "Jordan",
"Kuwait", "Lebanon", "Oman", "Palestine", "Qatar", "Saudi Arabia", "Syria",
                    "United Arab Emirates", "Yemen"]
```

Complete the following task:

1. **Descriptive Statistics:**
   - Calculate the mean, Standard deviation of the `score` for both South Asia and Middle East.
   - Which region has higher happiness Scores on average?
2. **Top and Bottom Performers:**
   - Identify the top 3 and bottom 3 countries in each region based on the `score`.
   - Plot bar charts comparing these charts.
3. **Metric Comparisons:**
   - Compare key metrics like `GDP per Capita`, `Social Support`, and `Healthy Life Expectancy` between the regions using `grouped bar charts`.
   - Which metrics show the largest disparity between the two regions?
4. **Happiness Disparity:**
   - Compute the range (max - min) and coefficient of variation (CV) for `Score` in both regions.
   - Which region has greater variability in happiness?
5. **Correlation Analysis:**
   - Analyze the correlation of `Score` with other metrics `Freedom to Make Life Choices`, and `Generosity` within each region.
   - Create scatter plots to visualize and interpret the relationships.
6. **Outlier Detection:**
   - Identify outlier countries in both regions based on `Score` and `GDP per Capita`.
   - Plot these outliers and discuss their implications.
7. **Visualization:**
   - Create boxplots comparing the distribution of `Score` between South Asia and the Middle East.
   - Interpret the key differences in distribution shapes, medians, and outliers.

4 Report Guidelines.

There are no specific format of what report should be, feel free to use your imagination to make it better. Before you make a submission please make sure following are covered:

1. General Guidelines

- Include the College and University approved Cover and Title Page.
- Formatting: Use clear headings for each section and subsection.
- Visualizations: Ensure all plots are appropriately labeled and titled, with concise captions.
- Language: Maintain a formal, academic tone throughout the report.
- Submission Requirements:
  - Submit a well-organized PDF report.
  - Attach the Jupyter notebook with all code, comments, and rendered outputs.
  - Ensure all tasks align with the report content.
  - Ensure you save or screenshot a copy of plagiarism report and ask your respective instructor to verify with a signature and keep your report save till the end of semester.{Please be reminded Plagiarism allowed is 20% only and any AI detected content will be not accepted for submission.}

2. Suggested Structure:

Please note that following is the suggested structure only, please fee free to use any structure you think best suits the need or describe the task better.

- **Title:** `Analysis of the World Happiness Report:Exploring South Asia and Middle East Perspectives.`
- **Introduction:** What to include?
  - Provide an overview of the World Happiness Report and its importance.
  - Briefly outline the objectives of the report and the tasks for each section: data exploration, South Asia analysis, and South Asia vs. Middle East comparison.
- **Report Section - For Each Problem**:
  1. Problem - *: Briefly summarize activity you did in this section followed by:
     - Step - by - Step explanations fo tasks performed.
     - Well - organized comparative visualizations if any.
     - Clear label plots with captions and short description if any.
     - Observations, insights or answered to any asked discussion question.
     - Interpretations of findings in the broader context of regional happiness.

  - Discuss any challenges encountered and how they were addressed.
- **Conclusion:** What to include?
  - Provide a concise summary of the findings from each problem.
  - Reflect on the significance of the analysis and its implications for understanding global happiness.