AYTS 5001 – Final Assignment

Title: Programming for Analytics – Final Project

Weighting: 70% of Total Grade

Submission: Week 13

Tools: Python, Jupyter Notebook, Pandas, NumPy, Matplotlib / Seaborn

Learning Objectives

- Work with real-world datasets (importing, cleaning, analyzing).
- Apply data transformation and statistical techniques using Python.
- Create clear and meaningful data visualizations.
- Write clean, modular, and well-documented code.
- Present a data-driven narrative supported by analysis and visuals.

Dataset

You must choose one dataset from Kaggle. It should have at least 500 rows, 6 columns, and be suitable for analysis using Pandas.

Recommended options:
1. Netflix Movies and TV Shows – https://www.kaggle.com/shivamb/netflix-shows
2. World Happiness Report – https://www.kaggle.com/unsdsn/world-happiness
3. Video Game Sales – https://www.kaggle.com/gregorut/videogamesales
4. Global Temperature Data – https://www.kaggle.com/berkeleyearth/climate-change-earth-surface-temperature-data
5. Goodreads Books Dataset – https://www.kaggle.com/jealousleopard/goodreadsbooks

Project Requirements

1. Introduction and Dataset Overview (10%)

• Describe your chosen dataset and its context.
• State your research question(s).
• Load the dataset using Pandas and display basic information using .info(), .shape(), and .describe().

2. Data Cleaning and Preparation (20%)

• Handle missing or inconsistent data.
• Convert data types where necessary.
• Rename columns for clarity.
• Filter, subset, or merge data where relevant.
• Explain all transformations in comments or Markdown cells.

3. Data Exploration and Analysis (25%)

- Perform descriptive statistics (mean, median, etc.).
- Use grouping and aggregation to uncover patterns.
- Identify correlations or comparisons between variables.
- Present at least three analytical insights supported by calculations.

4. Data Visualization (25%)

- Create at least four visualizations:
  - Bar or column chart
  - Histogram or boxplot
  - Scatter plot (with trend/correlation)
  - One additional meaningful visualization (e.g., heatmap, line chart)
- All charts should have titles, labels, and legends.

5. Conclusions and Reflection (10%)

- Summarize your main findings.
- Reflect on challenges encountered (e.g., missing data, bias).
- Suggest possible next steps or improvements.

6. Code Quality and Presentation (10%)

- Use clear variable names and comments.
- Avoid hardcoding paths (use os.path or pathlib).
- Ensure the notebook runs from start to finish without errors.
- Use Markdown to explain and structure the workflow.

Submission Requirements

You must submit your completed work via your own public GitHub repository.

The repository must include:
1. final_project.ipynb – your Jupyter notebook file.
2. dataset.csv – the dataset used (if under 100MB).
3. final_report.pdf – exported version of your notebook.
4. Any additional visualizations as PNG files (if not embedded).

Include a clear README.md file explaining the dataset, objectives, and results.

YOU WILL HAVE TO DEMONSTRATE YOUR PROJECT IN THE LAB OF WEEK 13.

Marking Rubric (70 Marks Total)

Section	Description	Weight	Criteria
1	Introduction & dataset overview	10%	Clear context and objectives

2	Data cleaning and preparation	20%	Logical workflow, missing data handled
3	Data exploration and analysis	25%	Insightful summaries, correct use of Pandas
4	Visualizations	25%	Meaningful, clear, and correct charts
5	Conclusions and reflection	10%	Evidence of understanding and synthesis
6	Code quality & presentation	10%	Readable, modular, well-commented

Tips for Success

- Focus on clarity and storytelling—guide the reader through your analysis.
- Use Markdown cells to structure your notebook clearly.
- Avoid unnecessary complexity; aim for reproducible, well-documented work.
- Ensure all visualizations are readable and properly labeled.
- Test the entire notebook before submission to ensure it runs without errors.