Python Analysis Project

Overview:

- **Objective:** Perform a full data analysis pipeline on a real dataset using Python—cleaning, transforming, analyzing, and drawing insights.

Project Components:

- Use pandas, numpy, seaborn/matplotlib
- Source a dataset of your choice online (for example Kaggle)
- Explore and understand your dataset. What fields are available? How many records are there? Could your data have bias?
- Start asking questions and formulating your analysis. What insights can be drawn, what questions do you want your analysis to answer, or, what story do you want your analysis to tell?
- Clean your dataset.
- Start your data manipulation. Add any calculated fields or aggregation that will help to answer your questions.
- Visualise any statistics that may need visualisation.
- Ensure to have a summary at the end, outlining main points/insights.

Supporting information:

Examples of Datasets & possible analysis points

- Superstore Sales Data:
    - Monthly sales trends
    - Top-performing product categories and regions
    - Profit vs. discount relationship
    - Customer segmentation by buying behaviour
- House Price Data:
    - Correlation between house features and price
    - Feature importance (e.g., size, neighbourhood, year built)
- IMDb Movies Dataset
    - Top-rated movies by genre/year
    - Relationship between rating and runtime or budget
    - Trends in genre popularity over time
    - Distribution of ratings across decades
- World Happiness Report

- Happiness vs. GDP, social support, and corruption perception
- Rank countries by happiness over multiple years
- Regional comparisons
- Correlation heatmaps and clustering countries

Tips for Data Cleaning:

- Check for missing values, and handle them, do they need fixing, deleting, or flagging?
- Check for duplicated records, can they be removed?
- Check for inconsistent formats, do any data types need changing? Sometimes values that represent numbers will be stored as strings rather than integers, this will need fixing before analysis.
- Standardise text and format, for example, inconsistent casing, special characters etc.
- Use boxplots or statistical methods to detect outliers, and decide whether to remove them, keep them, or change them.
- Check for possible inaccuracies, for example, are values decreasing because of a trend or because of the sampling changing?

Tips for Choosing the Right Chart:

- Bar Charts: Best for categorical comparison
- Line Charts: Best for time trends
- Scatter Plots: Best for correlation/relationships
- Box Plots: Best for understanding distributions & outliers
- Heatmaps: Best for visualizing matrices
- Histograms: Best for frequency distribution of numerical data