Data Analysis
Final Assignment Instructions

Overview

Total Points: 201 + 50 bonus points

1 Team Formation and Dataset

- Form a team of 3 people and decide a team name
- Choose a dataset:
  - Any dataset suitable for time-series analysis
  - Westermo system test dataset (minimum 5 system tests)
  - Intel lab dataset is not permitted
  - Suggested sources: Kaggle (https://www.kaggle.com/datasets) or Hugging Face datasets (https://huggingface.co/datasets)
  - **Bonus:** +10 points for using a new dataset

disk-io-time: Rate of change in time spent on storage i/o operations.

sys-interrupt-rate: Rate of change of interrupts.

cpu-system: Summerized rate of change of seconds spent on kernel space threads.

2 Task Categories and Points

2.1 Data Preprocessing and Basic Analysis (50 points)

- Basic statistical analysis using pandas (5 points)
- Original data quality analysis (including visualization) (10 points)
- Data preprocessing (25 points)
- Preprocessed vs original data visual analysis (10 points)

2.2 Visualization and Exploratory Analysis (35 points)

- Time series visualizations (5 points)
- Distribution analysis with histograms (5 points)
- Correlation analysis and heatmaps (5 points)
- Daily pattern analysis (10 points)
- Summary of observed patterns - similar to True/False questions (10 points)

2.3 Probability Analysis (36 points)

Raphael

- Threshold-based probability estimation (10 points)
- Cross tabulation analysis (6 points)
- Conditional probability analysis (10 points)
- Summary of observations from each task (10 points)

2.4 Statistical Theory Applications (40 points)

Bruno

- Law of Large Numbers demonstration (10 points)
- Central Limit Theorem application (20 points)
- Result interpretation (10 points)

2.5 Regression Analysis (35 points)

Leo

- Linear/Polynomial model selection (10 points)
- Model fitting and validation (15 points)
- Result interpretation and analysis (10 points)

2.6 Bonus Points (50 points)

- New data (10 points)
- Q-Q plot with explanation (5 points)
  - Either for Central Limit Theorem demonstration
  - Or for regression analysis residuals
- Interactive Visualizations (up to 10 points)
- Cross-validation in Regression (5 points)
- Additional exploration and implementations (up to 20 points)

3 Deliverables and Submission

- Jupyter Notebook (.ipynb file)
- Notebook exported as HTML file
- Dataset used for analysis
- 3-page maximum report (including figures) following provided template
- Optional: GitHub repository (not required for grading)

4 Points Distribution

- Main analysis tasks: 171 points
- Report: 20 points
- Presentation: 10 points
- Bonus tasks: up to 50 points

5 Grading Criteria for Jupyter Notebook and Report

- Thoroughness & Completeness (25%)
  - Dataset understanding
  - Analysis depth
  - Method justification
- Clarity (25%)
  - Clear explanation of methods
  - Result interpretation
  - Reproducible analysis
- Presentation (25%)
  - Plot quality
  - Proper labeling
  - Clear legends
- Technical Correctness (25%)
  - Code functionality
  - Method appropriateness
  - Implementation accuracy

6 Deadlines

- Jupyter Notebook (.ipynb and html) and Report (pdf): December 10th, 2:00 AM