Data Analysis
Final Assignment Instructions

Overview

- **Total Points (this assignment including report and presentation): 386 + 35 bonus**

1 Team Formation and Dataset

- Form a team of 3 people and decide a team name.
- Choose a dataset (must be suitable for time-series analysis).
  - Suggested sources:
    - Kaggle: `https://www.kaggle.com/datasets`
    - Hugging Face Datasets: `https://huggingface.co/datasets`
    - NASA PCoE Data Set Repository: `https://www.nasa.gov/intelligent-systems-division/discovery-and-systems-health/pcoe/pcoe-data-set-repository/`
    - Information Is Beautiful (Data): `https://informationisbeautiful.net/data/`
    - UCI Machine Learning Repository (Datasets): `https://archive.ics.uci.edu/datasets`

2 Task Categories and Points

2.1 A. Data Preprocessing and Data Quality (70 points)

- Dataset overview (dimensions, columns, types, time range, sampling rate, missingness summary) (10 points)
- Basic statistical analysis using pandas (descriptives, grouped stats, quantiles) (10 points)
- Original data quality analysis with visualization (missingness patterns, outliers, duplicates, timestamp gaps, inconsistent units) (20 points)
- Data preprocessing pipeline (cleaning steps, handling missing data, outliers strategy, resampling or alignment if needed, feature engineering basics) (20 points)
- Preprocessed vs original comparison (before/after visuals plus commentary on what changed and why) (10 points)

2.2 B. Visualization and Exploratory Analysis (55 points)

- Time-series visualizations (raw, smoothed, rolling mean or windowed views) (10 points)
- Distribution analysis with histograms and density style plots where applicable (10 points)
- Correlation analysis and heatmaps (Pearson and at least one alternative such as Spearman, with short interpretation) (10 points)
- Daily or periodic pattern analysis (day-of-week, hour-of-day, seasonality indicators, or test-cycle patterns) (15 points)
- Summary of observed patterns as short check statements (similar to True/False style) with evidence (10 points)

2.3 C. Probability and Event Analysis (45 points)

- Threshold-based probability estimation for events (define event, justify threshold, compute empirical probability) (15 points)
- Cross tabulation analysis for two variables (10 points)
- Conditional probability analysis (at least two meaningful conditional relationships) (15 points)
- Summary of observations and limitations (what could bias these estimates, what assumptions were made) (5 points)

2.4 D. Statistical Theory Applications (45 points)

- Law of Large Numbers demonstration (15 points)
- Central Limit Theorem application (sampling distributions, effect of sample size, interpretation) (25 points)
- Result interpretation and sanity checks (what would invalidate your conclusion, what you verified) (5 points)

2.5 E. Regression and Predictive Modeling (45 points)

- Define a prediction target and features (justify why they make sense) (10 points)
- Linear or polynomial model selection (include rationale and show at least two candidates) (10 points)
- Model fitting and validation (train-test split appropriate for time-series. e.g., time-based split) (15 points)
- Residual analysis and interpretation (errors, bias, failure cases, what to improve next) (10 points)

2.6 F. Dimensionality Reduction and Statistical Tests (40 points)

Part 1. Dimensionality Reduction (25 points)

- PCA projection and interpretation (variance explained, what clusters or separations mean) (10 points)
- t-SNE embedding with justified hyperparameters (perplexity or similar) and interpretation (7 points)
- UMAP embedding with justified hyperparameters (neighbors, min_dist or similar) and interpretation (8 points)

Part 2. Hypothesis Tests (15 points)

Perform **at least three** tests. Each test must include: null hypothesis, why the test is appropriate, assumptions, p-value, and practical interpretation.

- **Chi-square test** (choose one):
  - Chi-square test of independence (use a contingency table from two categorical or binned variables), or
  - Chi-square goodness-of-fit (compare observed counts to an expected distribution you justify).

  (5 points)
- One mean or location comparison test (choose one): t-test, Welch t-test, Mann-Whitney U, or ANOVA (5 points)
- One time-series relevant test (choose one): stationarity test (ADF or KPSS), Ljung-Box for autocorrelation, or change-point style test if justified (5 points)

2.7 G. Report and Presentation (86 points)

- Report (max 5 pages, including figures, following provided template) (50 points)
- Presentation (slides plus short talk, focus on decisions and interpretation, not code) (36 points)

6min + 2min fragen

2.8 Bonus Points (35 points)

Maximum bonus for this assignment is **35 points**.

- Using a clearly new dataset not used in previous assignment iterations, with justification of why it is suitable (5 points)
- Q-Q plot with explanation (choose one):
  - For the CLT demonstration, or
  - For regression residuals.

  (5 points)
- Interactive visualizations (Plotly, Altair, ipywidgets, or similar). Must be useful, not decorative (up to 8 points)
- Cross-validation adapted for time-series (walk-forward, expanding window, or blocked CV). Explain choice (7 points)

- Additional advanced analysis beyond the tasks (choose one or more: anomaly detection, forecasting baseline, clustering with evaluation, feature importance, uncertainty estimation). Must include interpretation (up to 10 points)

3 Deliverables and Submission

- Jupyter Notebook (.ipynb file)
- Notebook exported as HTML file
- Dataset used for analysis (or a clear download script and exact dataset reference if too large)
- Report (PDF, max 5 3 pages, using provided template)
- Presentation slides (PDF export)
- GitHub repository (upload files above on GitHub with a link to dataset on GitHub, Google Drive or original source)

4 Points Distribution Summary

- Main analysis tasks (A–F): 300 points
- Report + Presentation (G): 86 points
- Total regular points: 386 points
- Bonus tasks: up to 35 points

5 Grading Criteria for Jupyter Notebook, Report, and Presentation

- **Technical Correctness (20%)**: correct methods, correct implementation, correct handling of time-series constraints
- **Analysis Skills and Critical Thinking (20%)**:
  - Justification of choices (why this preprocessing, why this split, why this test)
  - Assumptions stated and checked
  - Limitations and failure cases
  - Quality of interpretation (practical meaning, not only p-values and plots)
- **Thoroughness and Completeness (20%)**: tasks fully addressed, evidence shown, edge cases considered
- **Clarity and Reproducibility (20%)**: readable structure, clean explanations, reproducible execution, clear outputs
- **Presentation Quality (20%)**: plot quality, labeling, legends, narrative flow, clean slides and report figures

6 Deadlines

- Jupyter Notebook (.ipynb and HTML), Report (PDF), Slides (PDF): January 29th, 10:00 am