Manufacturing Data Statistical Analysis Challenge

Overview

This challenge focuses on the statistical foundations essential for data science roles in manufacturing environments. The goal is to develop and demonstrate a deep understanding of statistical concepts that form the backbone of robust predictive models in production settings.

Dataset

Use the manufacturing dataset provided in the previous challenge (33,860 observations with 50+ features related to production jobs, planning, inventory, and quality metrics).

Challenge Objectives

1. Statistical Distribution Analysis

- **Task**: Conduct a comprehensive statistical distribution analysis of key manufacturing metrics.
- **Deliverables**:
  - Identify the underlying statistical distributions of key numerical features (e.g., production quantities, lead times, inventory levels)
  - Test distributional assumptions using appropriate statistical tests (Shapiro-Wilk, Kolmogorov-Smirnov, Anderson-Darling)
  - Visualize distributional characteristics through QQ plots, histograms with fitted theoretical distributions
  - Justify transformations (log, Box-Cox, etc.) based on distribution properties

2. Hypothesis Testing Framework

- **Task**: Develop a statistical hypothesis testing framework to identify significant factors affecting the TARGET variable.
- **Deliverables**:
  - Formulate null and alternative hypotheses for key operational metrics
  - Perform appropriate statistical tests (t-tests, chi-square, ANOVA) to compare distributions between TARGET classes
  - Calculate and interpret effect sizes, not just p-values
  - Implement and interpret multiple hypothesis testing corrections (Bonferroni, Benjamini-Hochberg)
  - Document confidence intervals for all significant findings

3. Correlation and Dependency Analysis

- **Task**: Conduct advanced correlation and dependency analysis beyond simple Pearson correlation.
- **Deliverables**:
  - Calculate and compare Pearson, Spearman, and Kendall correlation coefficients
  - Implement and interpret partial correlation analysis to control for confounding variables
  - Perform conditional independence tests using appropriate statistical methods
  - Create and interpret a graphical model (Bayesian Network or Markov Random Field) to represent the dependency structure
  - Quantify the statistical significance of identified relationships

4. Statistical Modeling and Inference

- **Task**: Develop statistical models with rigorous inferential properties.
- **Deliverables**:
  - Implement a logistic regression model with proper statistical inference
  - Calculate and interpret odds ratios with confidence intervals
  - Perform likelihood ratio tests for nested models
  - Implement and interpret goodness-of-fit tests for the model
  - Compare statistical inference from traditional models with insights from machine learning approaches

5. Time Series Statistical Analysis

- **Task**: Apply time series statistical methods to analyze temporal patterns in the manufacturing data.
- **Deliverables**:
  - Test for stationarity using appropriate statistical tests (ADF, KPSS)
  - Perform time series decomposition to identify trend, seasonality, and residual components
  - Develop and validate ARIMA or other appropriate time series models
  - Conduct statistical tests for autocorrelation and partial autocorrelation
  - Analyze and interpret lead-lag relationships between different manufacturing metrics

6. Power Analysis and Sample Size Justification

- **Task**: Conduct power analysis to determine appropriate sample sizes for reliable statistical inference.
- **Deliverables**:
  - Calculate statistical power for key hypothesis tests in the analysis
  - Determine minimum required sample sizes for desired statistical power levels
  - Analyze how imbalanced classes affect statistical power

  - Implement and interpret learning curves to assess the relationship between sample size and model performance
  - Create a statistical justification for any sampling strategies used

7. Experimental Design Framework

- **Task**: Develop an experimental design framework for testing manufacturing process improvements.
- **Deliverables**:
  - Design an A/B testing framework for manufacturing process changes
  - Create a Design of Experiments (DOE) approach for multi-factor analysis
  - Develop statistical methods for identifying causal relationships vs. correlations
  - Outline approaches for dealing with confounding variables in the manufacturing environment
  - Define appropriate statistical measures and thresholds for the success of experiments

Deliverable Format

Produce a comprehensive Jupyter notebook with thorough statistical analysis, including:

- Clear explanation of all statistical concepts applied
- Mathematical formulation of statistical methods used
- Rigorous interpretation of results and their relevance to the manufacturing context
- Visualizations that enhance statistical understanding
- Well-documented statistical code with proper validation

Evaluation Criteria

Your work will be evaluated based on:

1. Depth and accuracy of statistical understanding
2. Appropriateness of statistical methods for the data and context
3. Rigor in hypothesis testing and inference
4. Clarity in explaining statistical concepts
5. Ability to translate statistical findings into actionable business insights
6. Technical implementation of statistical methods
7. Critical thinking about limitations and assumptions

Notes

- Focus on statistical foundations rather than machine learning algorithms
- Prioritize statistical rigor and proper inference over prediction accuracy

- Clearly state and test assumptions of all statistical methods used
- Connect statistical findings to the manufacturing context

This challenge is designed to demonstrate depth of statistical knowledge as applied to real-world manufacturing problems, bridging the gap between theoretical understanding and practical application.