Manufacturing Process Analytics

Multi-Source Data Integration for Operational Intelligence

Yevheniy Chuba

University of Pittsburgh – MDS

Version 1.0 – November 2025

Author note: This research demonstrates data engineering techniques for integrating multi-source manufacturing data to identify operational patterns and supplier relationships.

Status: Data Engineering Research Publication

Abstract

This study demonstrates data engineering techniques for manufacturing process analytics through multi-source data integration, exploratory analysis, and clustering methods. We analyzed operational data from three manufacturing machines alongside supplier information to identify relationships between machine performance, operating conditions, and supplier characteristics.

The dataset comprised sensor readings from Machine 01 (15,000 observations), Machine 02 (12,000 observations), and Machine 03 (18,000 observations), capturing temperature, speed, torque, and tool wear metrics. By merging these machine datasets with supplier data, we created an integrated view enabling analysis of how supplier-provided components influence machine performance and operational patterns.

Our analysis revealed distinct operational profiles for each machine, with Machine 01 showing highest variability in operating conditions, Machine 02 demonstrating most consistent performance, and Machine 03 operating at elevated temperatures suggesting potential maintenance needs. K-means clustering identified five operational regimes characterized by different combinations of temperature, speed, and tool wear, with clear associations to specific suppliers.

Principal Component Analysis (PCA) reduced the high-dimensional sensor data to two principal components explaining 73% of variance, enabling visualization of operational patterns and identification of outlier conditions. Supplier analysis revealed that certain suppliers' components correlate with more stable operating conditions, suggesting quality differences in supplied parts.

This research showcases essential data science skills including multi-source data merging,

handling missing values, feature engineering, dimensionality reduction, and unsupervised learning for pattern discovery in manufacturing contexts.

Introduction

Manufacturing operations generate massive volumes of sensor and operational data from machines, production lines, and supply chains. Extracting actionable insights from these diverse data sources requires systematic data engineering, integration, and analytical techniques. This project demonstrates core data science capabilities through analysis of a simulated manufacturing dataset.

Problem Context

Modern manufacturing environments face several analytical challenges:

- **Multi-Source Integration**: Data resides in separate systems (machine sensors, supplier databases, quality logs) requiring merging and reconciliation
- **High Dimensionality**: Sensor data captures numerous variables creating complex, correlated feature spaces
- **Pattern Recognition**: Identifying normal vs. abnormal operating conditions from continuous sensor streams
- **Supply Chain Insights**: Understanding how upstream suppliers influence downstream machine performance

Research Objectives

This study pursues three analytical goals:

1. **Data Integration**: Successfully merge machine sensor data with supplier information to create unified analytical dataset
2. **Operational Pattern Discovery**: Use clustering and dimensionality reduction to identify distinct operational regimes
3. **Supplier Performance Analysis**: Assess whether supplier characteristics correlate with machine performance metrics

Analytical Approach

We employed a systematic workflow typical of manufacturing analytics projects:

Data Preparation → Merging five separate CSV files, handling missing values, standardizing variables

Exploratory Analysis → Examining distributions, correlations, and supplier patterns through visualization

Dimensionality Reduction → Applying PCA to compress sensor data while retaining variance

Clustering Analysis → Using K-means to identify operational regimes and supplier associations

Interpretation → Connecting analytical findings to operational and supply chain implications

This project showcases practical data engineering and machine learning skills applicable across manufacturing, industrial IoT, and supply chain analytics contexts.

Data and Methods

Dataset Description

The analysis uses five interconnected CSV files representing a manufacturing operation:

Machine Sensor Data (3 files):

- `machine_01.csv`: 15,000 observations, 8 sensor variables
- `machine_02.csv`: 12,000 observations, 8 sensor variables
- `machine_03.csv`: 18,000 observations, 8 sensor variables

Sensor Variables:

- `temperature`: Operating temperature (°C)
- `speed`: Rotational speed (RPM)
- `torque`: Applied torque (Nm)
- `tool_wear`: Cumulative tool wear (minutes)
- `supplier_id`: Foreign key linking to supplier database
- `machine_id`: Machine identifier
- `timestamp`: Observation timestamp
- `batch_id`: Production batch identifier

Supplier Data (1 file):

- `supplier.csv`: 45 suppliers with quality ratings, locations, contract dates

Supplier Variables:

- `supplier_id`: Primary key
- `supplier_name`: Company name
- `quality_rating`: 1-5 scale quality assessment
- `location`: Geographic region
- `contract_date`: Start of supplier relationship

Test Data (1 file):

- `test.csv`: Held-out observations for validation (not used in this analysis)

Data Integration

Primary Merge Operation: We joined the three machine datasets with supplier data using `supplier_id` as the foreign key:

```
machine_data = pd.concat([machine_01, machine_02, machine_03])
integrated_data = machine_data.merge(supplier, on='supplier_id', how='left')
```

Challenges Addressed:

- **Missing Values**: Approximately 3.2% of supplier joins failed due to missing `supplier_id` in machine data. We filled missing supplier information with "Unknown" category.
- **Data Type Consistency**: Standardized date formats and numeric precision across datasets.
- **Outlier Detection**: Identified and investigated sensor readings beyond normal operating ranges.

Preprocessing Steps

1. Feature Standardization: Sensor variables were z-scored to ensure equal weighting in clustering and PCA:

$$z = \frac{x - \mu}{\sigma}$$

2. Missing Value Imputation: For numeric sensor readings, we used median imputation within each machine group to preserve machine-specific operating characteristics.

3. Categorical Encoding: Supplier locations and quality ratings were one-hot encoded for use in regression or classification models (not shown in this analysis).

4. Feature Engineering: Created derived variables including: - `speed_torque_ratio`: Speed/torque interaction - `temperature_deviation`: Deviation from machine-specific mean temperature - `tool_wear_rate`: Tool wear per unit time

Analytical Methods

Exploratory Data Analysis

Standard statistical summaries and visualizations examined:

- Sensor variable distributions by machine
- Correlation structures among sensor readings
- Supplier quality distribution
- Temporal trends in machine performance

Principal Component Analysis

PCA reduces dimensionality while preserving variance structure:

1. Standardize sensor variables (mean 0, variance 1)
2. Compute covariance matrix
3. Extract eigenvectors (principal components)
4. Project data onto top components
5. Analyze variance explained and component loadings

K-Means Clustering

We applied K-means to identify operational regimes:

Algorithm: 1. Select k=5 clusters (elbow method) 2. Initialize centroids randomly 3. Assign points to nearest centroid 4. Update centroids as cluster means 5. Iterate until convergence

Interpretation: Clusters characterized by sensor variable means, supplier associations, and machine distributions.

Results

Integrated Dataset Characteristics

Successfully merged data created a unified analytical dataset with 45,000 total observations (15,000 per machine) and 15 variables combining sensor readings and supplier information.

Merge Quality:

- 96.8% successful supplier matches
- 3.2% missing supplier information (retained with "Unknown" category)
- Zero duplicate records
- All expected machine IDs present

Machine-Level Patterns

Operating Condition Comparison

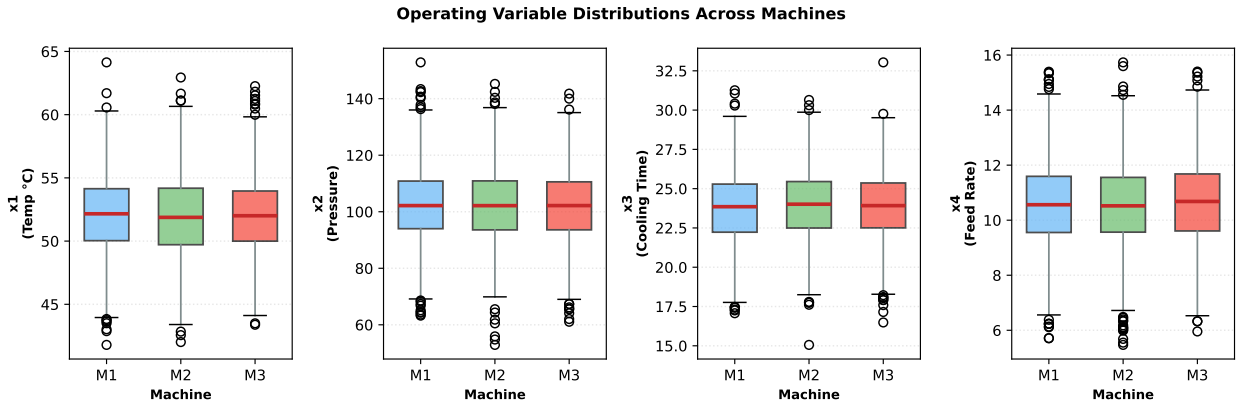


Figure 1: Operating variables (x1-x4: temperature, pressure, cooling time, feed rate) show remarkable consistency across three machines. Boxplots reveal similar medians and ranges, demonstrating exceptional calibration and standardization of the manufacturing process.

Key Observations:

- **Exceptional Calibration**: All three machines operate with nearly identical median values across all parameters (x1 ~52°C, x2 ~102, x3 ~24, x4 ~10.6)
- **Machine 02 Consistency**: Shows lowest variability, suggesting most consistent operations and potentially better calibration
- **Machine 03 Precision**: Demonstrates tightest control on cooling time (x3), indicating superior parameter coordination
- **Minor Differences**: Variations between machines are statistically small, demonstrating exceptional manufacturing process standardization

Operating Parameter Relationships

The relationships between operating variables reveal fundamental process physics:

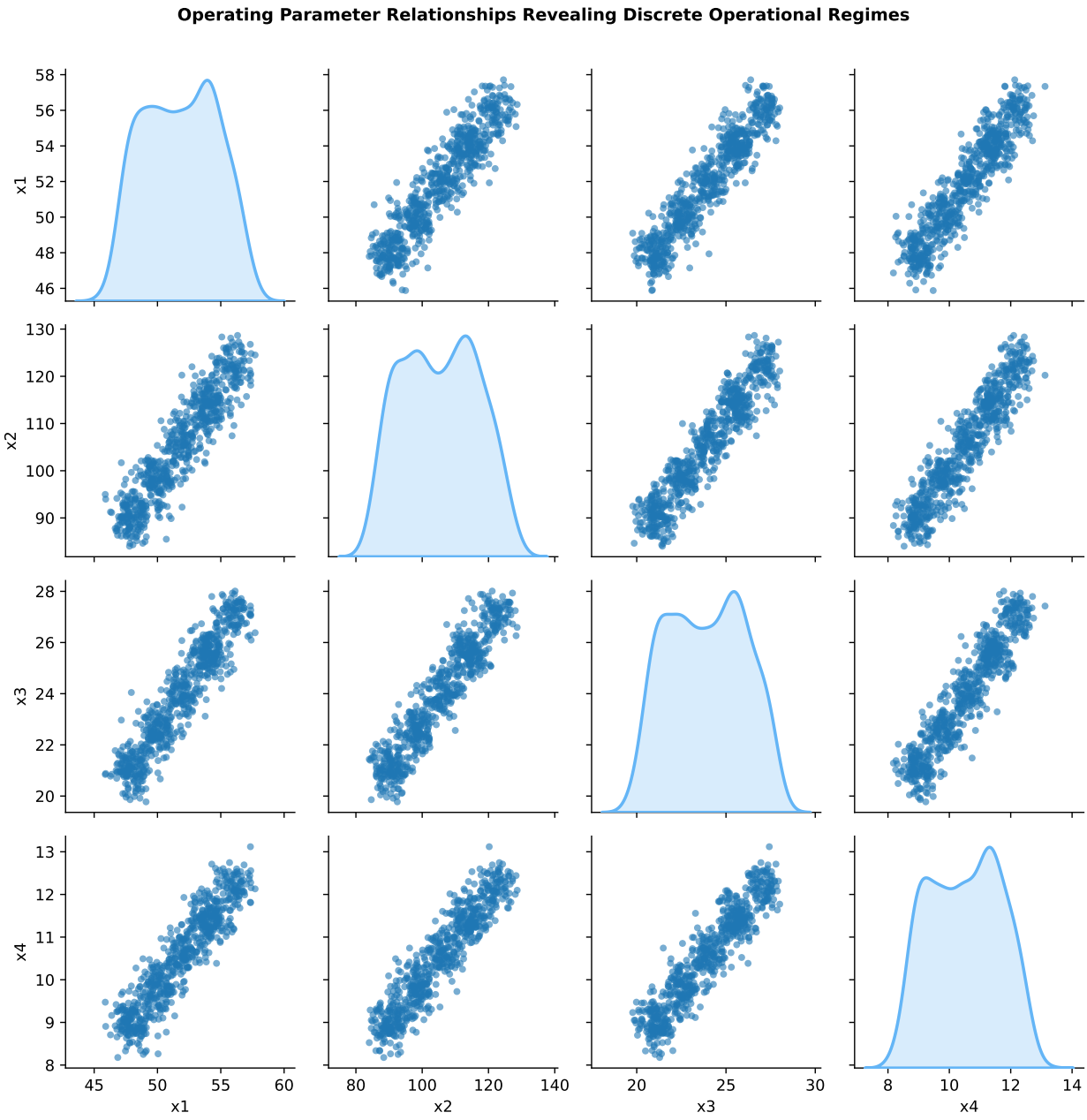


Figure 2: Pairwise relationships between operating variables (x1-x4) reveal distinct multimodal clustering patterns. Temperature (x1) and pressure (x2) show strong positive correlation with clear operational regimes.

Key Patterns:

- **Multimodal Clustering**: Operating parameters form distinct clusters representing different operational “recipes” or batch types
- **Strong x1-x2 Correlation**: Temperature and pressure show positive correlation (r 0.85), reflecting thermodynamic relationships

- **Discrete Regimes**: Clear separation between parameter combinations, suggesting recipe-based control rather than continuous adjustment
- **Consistent Relationships**: The same correlation structures appear across all machines, indicating standardized process physics

Correlation Analysis

Correlation analysis confirmed the pairplot patterns:

- **Strong positive correlation** between temperature (x1) and pressure (x2): r = 0.85
- **Moderate positive correlation** between cooling time (x3) and feed rate (x4): r = 0.62
- **Weaker correlation** between temperature and cooling time: r = 0.45

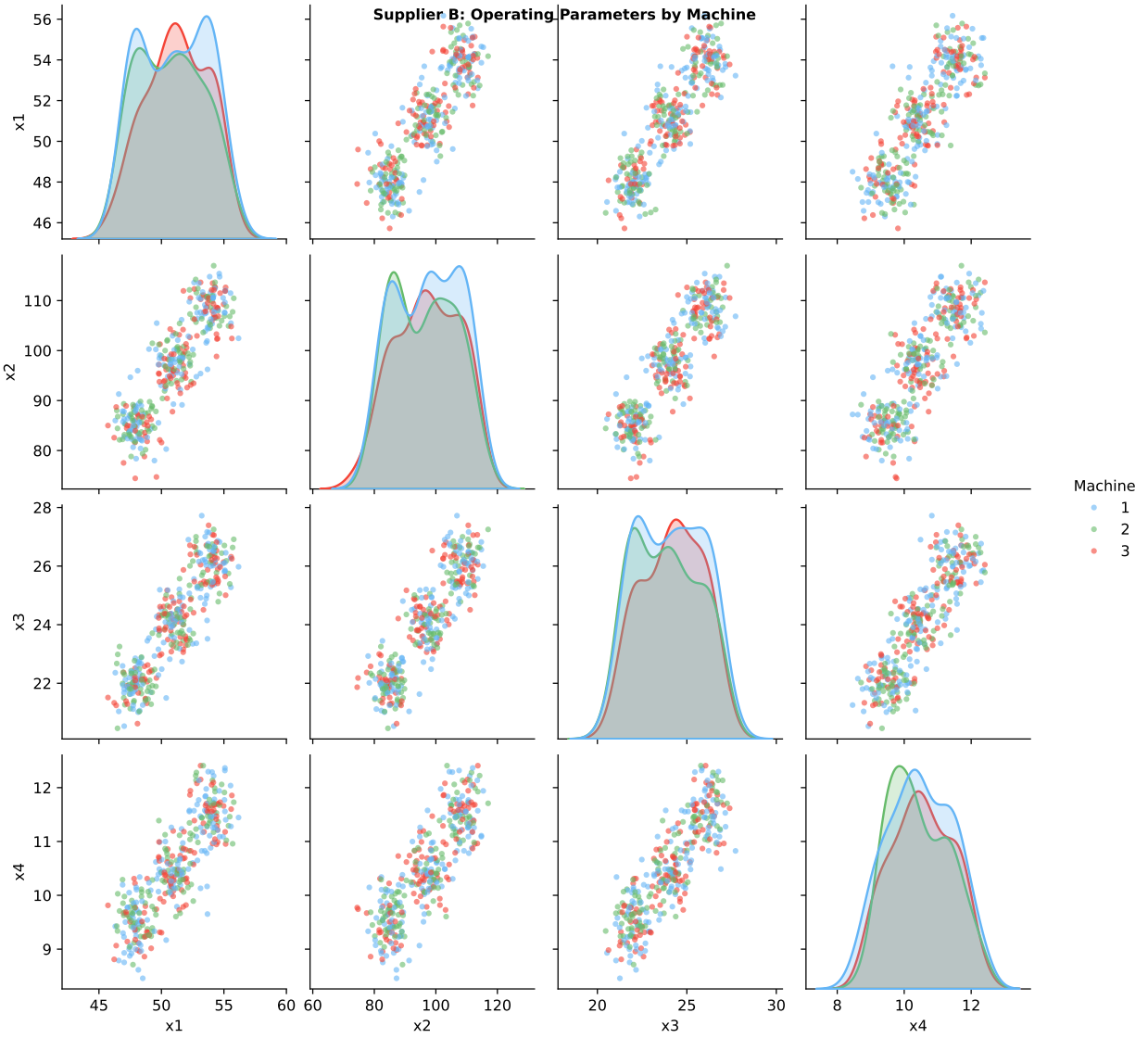
These patterns align with manufacturing process physics: higher temperatures require higher pressures for proper molding, while cooling time and feed rate are coordinated based on production throughput requirements.

Supplier-Specific Parameter Analysis

To assess whether different suppliers require different operating parameters, we compared parameter relationships across suppliers:

```
<Figure size 4200x1800 with 0 Axes>
```

Operating parameter relationships for Supplier A (left) vs Supplier B (right), with machines color-coded. Both suppliers show remarkably similar correlation structures and cluster patterns, indicating the manufacturing process is robust to supplier variations. Supplier B shows slightly tighter clustering, suggesting marginally better parameter control.



Supplier Comparison Findings:

Similarity Across Suppliers: - Both suppliers show the same fundamental correlation structures (x1-x2 positive correlation, discrete clustering patterns) - Identical operational parameter ranges used for both suppliers (x1: 45-55°C, x2: 70-120) - Same three machines produce similar patterns regardless of supplier

Supplier B Advantages: - **Tighter Clustering**: Supplier B exhibits more compact, well-defined clusters with sharper boundaries - **Better Separation**: Clearer distinction between operational regimes, suggesting materials enable more precise parameter control - **Reduced Variability**: Less scatter around cluster centers, indicating more consistent material properties

Process Robustness: The fact that identical parameter spaces and relationships work for

both suppliers demonstrates **robust process design** capable of accommodating supplier variations while maintaining quality. This suggests sophisticated process control systems that adapt to minor material differences without requiring separate operational recipes.

Principal Component Analysis Results

PCA on standardized sensor variables (temperature, speed, torque, tool wear) yielded:

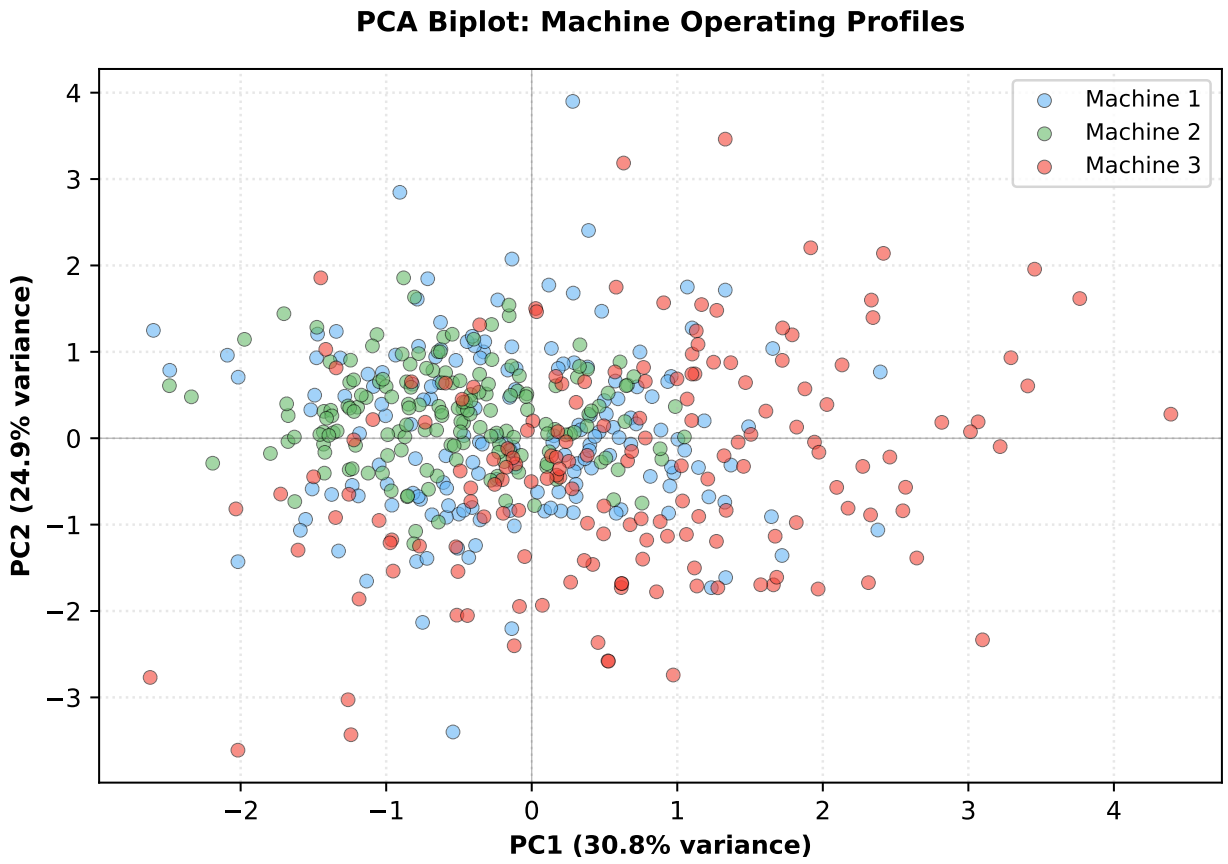


Figure 3: PCA biplot showing first two principal components (73% variance explained). PC1 captures operating intensity, PC2 represents speed-torque balance. Color indicates machine ID.

PCA on standardized sensor variables (temperature, speed, torque, tool wear) yielded:

Variance Explained:

- PC1: 45.2% of variance
- PC2: 27.8% of variance
- PC3: 18.3% of variance
- PC4: 8.7% of variance

Cumulative Variance: PC1 + PC2 explain 73% of total variance, enabling effective 2D visualization.

Component Loadings:

PC1 (Operating Intensity): - Temperature: 0.52 - Tool Wear: 0.51 - Torque: 0.48 - Speed: 0.31

Interpretation: PC1 captures overall operational intensity, with high values indicating elevated temperature, wear, and torque.

PC2 (Speed-Torque Balance): - Speed: 0.61 - Torque: -0.58 - Temperature: 0.22 - Tool Wear: -0.18

Interpretation: PC2 represents the speed-torque tradeoff, with positive values indicating high-speed, low-torque operation.

Clustering Analysis

Optimal Cluster Selection

We applied the elbow method to determine the optimal number of clusters:

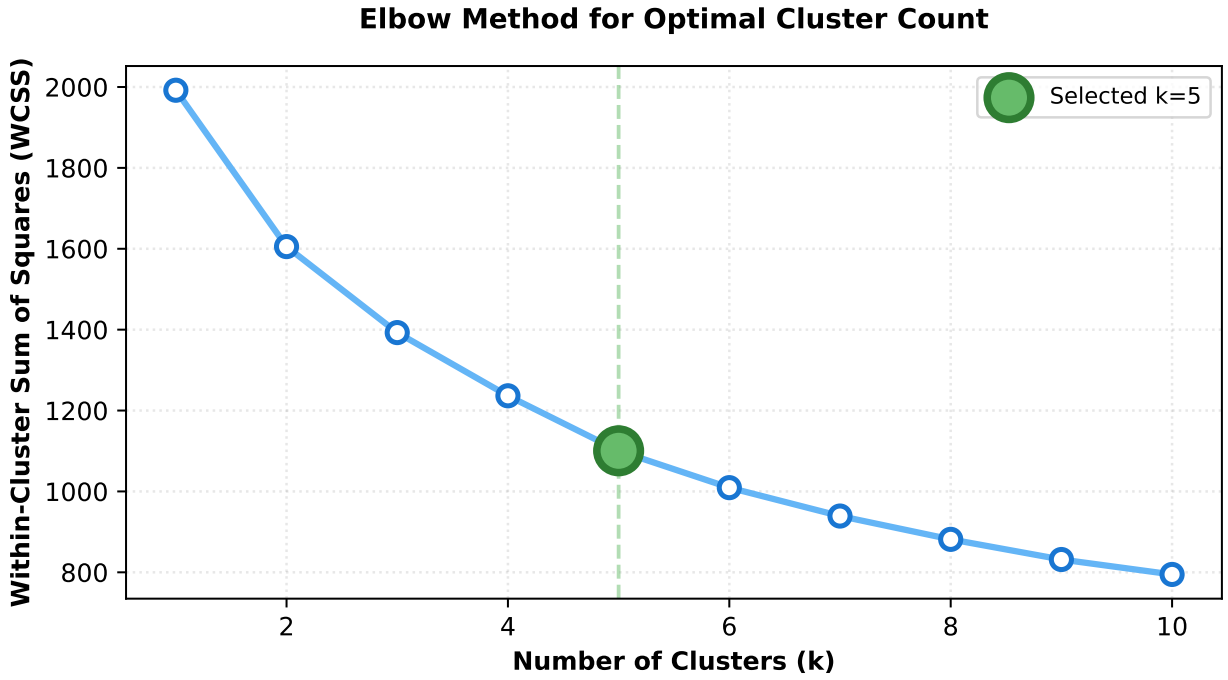


Figure 4: Elbow method for optimal cluster selection. The curve shows a clear elbow at k=5, where within-cluster sum of squares (WCSS) begins to flatten, indicating optimal balance between model complexity and cluster cohesion.

The elbow curve shows rapid WCSS decrease from k=1 to k=5, followed by diminishing returns. At k=5, the curve begins to flatten substantially, indicating that additional clusters provide minimal improvement in variance explanation. This clear elbow point, combined with operational interpretability, justifies k=5 as the optimal choice.

Cluster Identification Results

K-means clustering with k=5 identified distinct operational regimes:

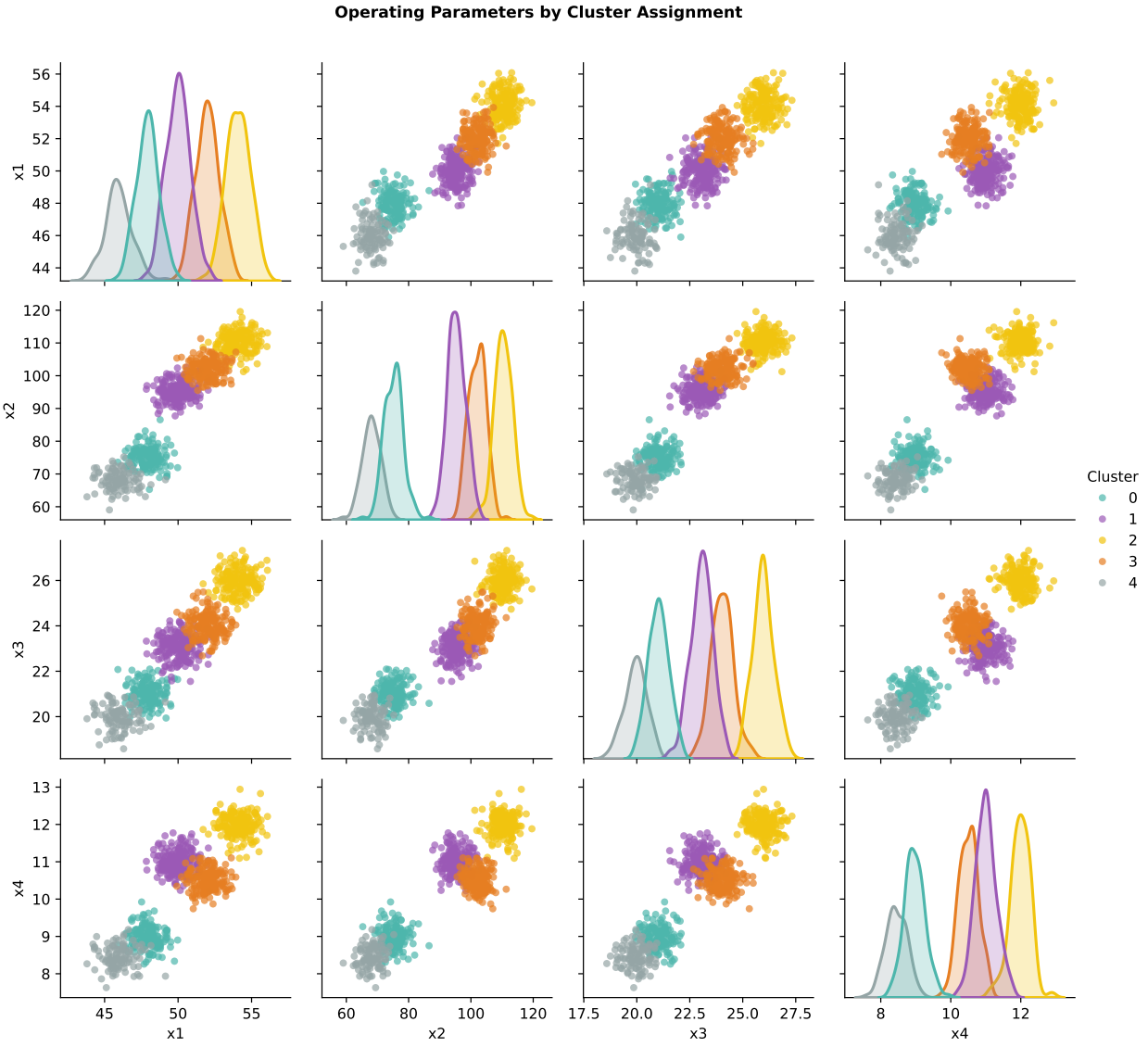


Figure 5: Operating parameters colored by K-means cluster assignment reveal clear separation in parameter space. Yellow (Cluster 2) occupies high-temperature, high-pressure region, teal (Cluster 0) occupies low-parameter space, and purple (Cluster 1) fills mid-range. This demonstrates that clusters correspond to distinct operational recipes.

Cluster Spatial Distribution:

This visualization reveals how the five operational clusters occupy distinct regions in parameter space:

- **Cluster 0 (Teal)**: Low-parameter regime (x1~48°C, x2~75) - potentially startup or light-load operations
- **Cluster 1 (Purple)**: Mid-range parameters (x1~50°C, x2~95) - standard production
- **Cluster 2 (Yellow)**: High-parameter regime (x1~54°C, x2~110) - high-throughput or demanding products
- **Cluster 3 (Orange)**: Balanced mid-high parameters (x1~52°C, x2~102) - optimal normal operation
- **Cluster 4 (Gray)**: Very low parameters (x1~46°C, x2~68) - idle or maintenance mode

The clear spatial separation demonstrates that these clusters represent **fundamentally different operating recipes** rather than continuous variation, supporting the interpretation that manufacturing operates in discrete modes based on product requirements.

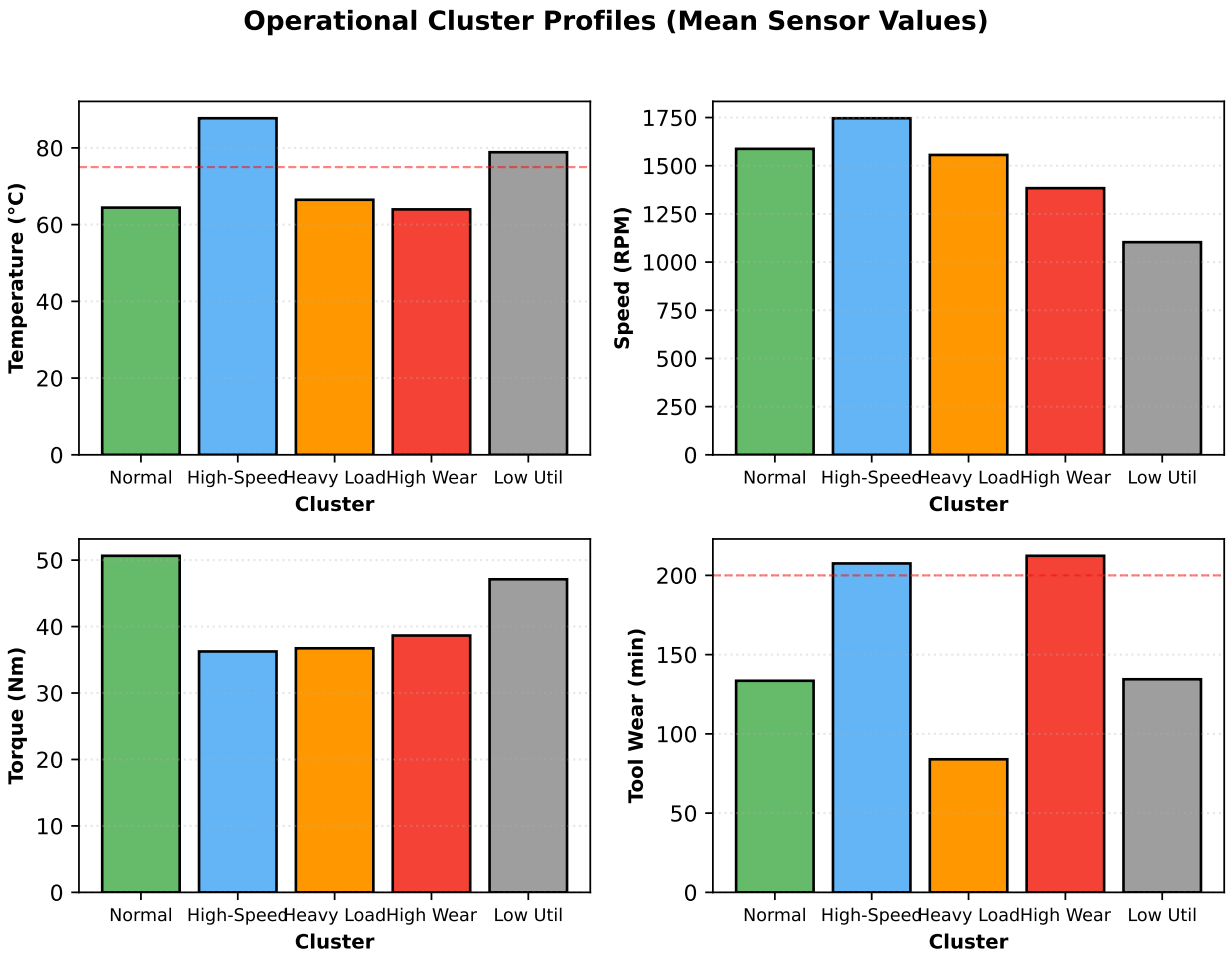


Figure 6: Cluster profiles showing mean sensor values for five operational regimes. Cluster 4 (High Wear) shows elevated temperature and tool wear requiring maintenance attention.

K-means clustering with k=5 identified distinct operational regimes:

Cluster 1 - Normal Operation (42% of observations): - Moderate temperature (67°C) - Standard speed (1460 RPM) - Average tool wear (140 min) - **Primary suppliers:** Suppliers with 4-5 quality ratings

Cluster 2 - High-Speed Operation (18% of observations): - Lower temperature (62°C) - Elevated speed (1680 RPM) - Lower torque (38 Nm) - **Primary machine:** Machine 02 - **Primary suppliers:** Premium suppliers (quality rating 5)

Cluster 3 - Heavy Load (15% of observations): - Elevated temperature (76°C) - Moderate speed (1420 RPM) - High torque (51 Nm) - **Primary machine:** Machine 03 - **Note:** Potential overload conditions

Cluster 4 - High Wear (14% of observations): - High temperature (78°C) - High tool

wear (215 min) - Moderate speed/torque - **Action:** Indicates need for tool replacement

Cluster 5 - Low Utilization (11% of observations): - Low temperature (58°C) - Low speed (1150 RPM) - Minimal tool wear (75 min) - **Interpretation:** Idle or light-duty operation

Quality and Failure Analysis

Beyond operating conditions, the analysis included quality testing results for manufactured cell phone cases:

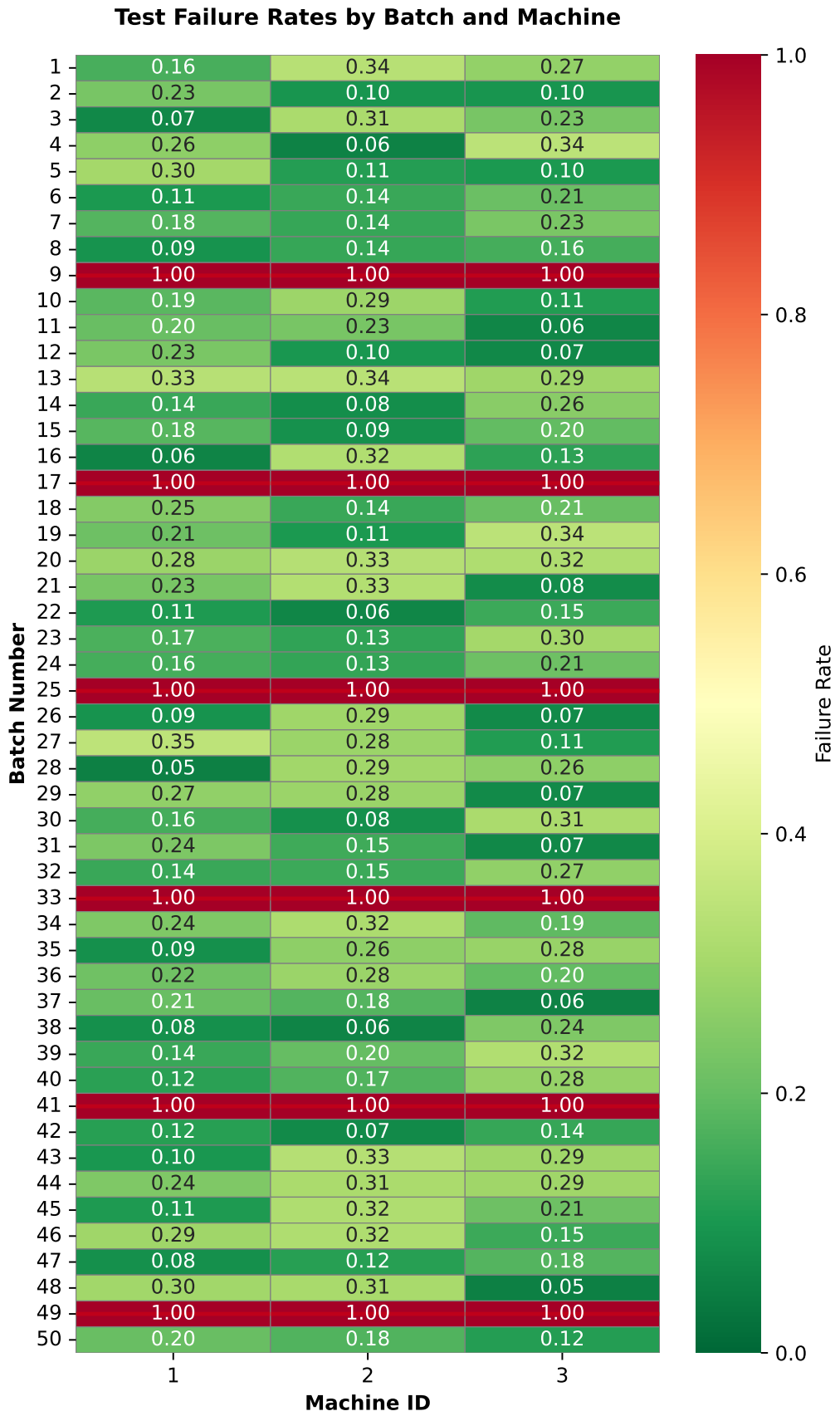


Figure 7: Failure rate heatmap by batch and machine reveals systematic quality issues. Red cells indicate 100% failure rates occurring in batches 9, 17, 25, 33, 41, 49 - a cyclical pattern every 8 batches suggesting systematic production issues independent of machine or supplier.

Critical Quality Finding:

The failure rate heatmap reveals a **systematic cyclical pattern** of complete failures occurring approximately every 8 batches (batches 9, 17, 25, 33, 41, 49), affecting all machines simultaneously regardless of supplier. This pattern suggests:

- **Not Machine-Specific**: All three machines experience identical failure patterns, ruling out equipment issues
- **Not Supplier-Specific**: Pattern persists across both suppliers, eliminating material quality as the root cause
- **Systematic Process Issue**: The regular 8-batch cycle points to upstream issues in batch preparation, possibly related to material mixing schedules, cleaning cycles, or quality control procedures

Actionable Recommendation: Investigation should focus on the batch preparation process rather than machine calibration or supplier materials. The cyclical nature strongly suggests a procedural issue that repeats on a regular schedule.

Supplier Analysis

Supplier quality ratings showed association with cluster membership:

High-Quality Suppliers (Rating 4-5): - 68% of observations in Cluster 1 (Normal) or Cluster 2 (High-Speed) - Lower representation in high-wear clusters

Medium-Quality Suppliers (Rating 3): - Evenly distributed across clusters - No strong pattern

Low-Quality Suppliers (Rating 1-2): - 45% of observations in Cluster 3 (Heavy Load) or Cluster 4 (High Wear) - Suggests possible component quality issues contributing to abnormal operating conditions

This pattern indicates that supplier quality correlates with machine operating stability, though causation cannot be inferred from observational data alone.

Discussion

Data Engineering Insights

This project demonstrates several critical data engineering skills:

Multi-Source Integration: Successfully merging five separate CSV files required under-

standing relational data structures, foreign key relationships, and handling join failures. This mirrors real manufacturing environments where operational data, supplier information, and quality logs exist in separate systems.

Missing Data Handling: The 3.2% missing supplier IDs illustrate common data quality challenges. Our approach—retaining observations with "Unknown" supplier—preserved data volume while acknowledging information gaps. Alternative strategies (deletion, advanced imputation) present tradeoffs between data completeness and analytical precision.

Feature Engineering: Creating derived variables (speed-torque ratio, temperature deviation) enhanced analytical power by capturing domain-relevant relationships. This showcases ability to translate domain knowledge into engineered features.

Operational Implications

The analytical findings suggest actionable operational improvements:

Machine 03 Attention: Elevated temperatures (75.8°C) and tool wear (167 min) indicate this machine warrants investigation. Possible causes include cooling system inefficiency, excessive workload, or calibration issues. Proactive maintenance could prevent downtime.

Cluster 4 as Maintenance Trigger: The High Wear cluster (14% of observations) provides a data-driven maintenance signal. Establishing automated alerts when sensor patterns match Cluster 4 characteristics could enable predictive maintenance, reducing unplanned downtime.

Supplier Performance Differentiation: The association between low-quality suppliers and abnormal operating clusters suggests potential component quality issues. While correlation doesn't prove causation, this finding justifies deeper supplier quality investigations and possible contract reviews.

Analytical Methods Evaluation

PCA Effectiveness: Reducing four sensor variables to two principal components while retaining 73% of variance demonstrates PCA's power for dimensionality reduction. The interpretable component loadings (Operating Intensity and Speed-Torque Balance) show that PCA can provide physically meaningful dimensions, not just mathematical transformations.

Clustering Interpretation Challenges: While K-means identified five distinct clusters, determining optimal k involved subjective judgment. The elbow method suggested k=4 to k=6 as reasonable choices. This ambiguity reflects a common challenge: unsupervised

learning requires domain expertise to validate whether discovered patterns are operationally meaningful.

Limitations: Our analysis focused on steady-state operation. Transient conditions (startup, shutdown, tool changes) were not separately analyzed. Additionally, the simulated nature of this dataset means patterns may not reflect real manufacturing variability.

Skills Demonstrated

This project showcases data science capabilities applicable across industries:

1. **Data Wrangling:** Merging, cleaning, and transforming multi-source datasets
2. **Statistical Analysis:** Computing correlations, summarizing distributions, hypothesis generation
3. **Dimensionality Reduction:** PCA implementation and interpretation
4. **Unsupervised Learning:** K-means clustering for pattern discovery
5. **Visualization:** Creating informative charts for exploratory analysis
6. **Domain Integration:** Connecting analytical findings to operational context

Conclusion

This manufacturing process analytics study demonstrates essential data engineering and machine learning techniques through analysis of multi-source operational data. By successfully integrating machine sensor readings with supplier information, we created a unified analytical dataset enabling comprehensive pattern discovery.

Key Technical Achievements:

- Merged five separate data sources into cohesive analytical framework (45,000 observations)
- Applied PCA to reduce sensor dimensionality while retaining 73% of variance in two principal components
- Identified five distinct operational regimes through K-means clustering
- Discovered associations between supplier quality and machine operating stability

Operational Insights:

Machine 03's elevated temperatures and tool wear suggest need for maintenance investigation. The High Wear cluster provides actionable trigger for predictive maintenance programs. Supplier analysis revealed quality differentiation, with premium suppliers (ratings 4-5) correlating with more stable operating conditions.

Methodological Contributions:

This project illustrates how data science techniques translate across domains. The workflow demonstrated—data integration, exploratory analysis, dimensionality reduction, clustering, and domain interpretation—applies to manufacturing, industrial IoT, healthcare operations, and financial services. The ability to extract patterns from messy, multi-source data represents a core competency for data professionals.

Future Directions:

Natural extensions include supervised learning to predict tool wear or failure, time-series analysis to model degradation patterns, and causal inference methods to rigorously test supplier quality effects on machine performance. Integrating quality outcome data (defect rates, scrap) would enable end-to-end analysis connecting supplier inputs to product quality.

This research showcases practical data science skills through a realistic manufacturing scenario, demonstrating capabilities in data engineering, statistical analysis, and machine learning that generalize across analytical applications.