Data Wrangling Report: Manufacturing Downtime Analysis

Executive Summary

This report documents the comprehensive data cleaning process applied to a manufacturing downtime dataset from a biscuit production facility. The dataset originally contained 8,164 records with 13 columns tracking production metrics across various machines and products during July 2021. Through systematic data assessment and cleaning procedures, we identified and resolved multiple data quality issues, resulting in a clean dataset of 8,135 records ready for analysis.

---

1. Dataset Overview

Original Dataset Dimensions: 8,164 rows × 13 columns

Data Collection Period: July 1 - August 1, 2021

Key Variables:

- Machine identifiers and product types
- Production quantities (total and good units)
- Temporal data (dates and times)
- Stoppage classifications
- OEE (Overall Equipment Effectiveness) codes

---

2. Data Assessment Phase

2.1 Initial Data Exploration

The assessment phase revealed the dataset structure through multiple diagnostic checks:

Numerical Columns Analysis:

- Total Biscuits Made: Range 0 to 961,128 units (highly skewed with 75th percentile at 0)
- Good Made Units: Range -4,662 to 480,646 units (negative values detected)
- Total Units Made: Range 0 to 1,068,855 units

Categorical Columns Analysis:

- 10 unique machine types (Biscuit Filling Machine most frequent: 4,013 records)

- 19 product varieties (Bourbon Creams most common: 1,401 records)
- 72 unique machine-product combinations
- 3 stoppage type categories with inconsistent formatting

2.2 Identified Data Quality Issues

Issue 1: Missing Values

- OEE Code: 15 null values (0.18% of dataset)
- Complete null records: 5 rows with all fields missing
- **Impact:** Missing OEE codes prevent proper equipment effectiveness classification

Issue 2: Negative Values

- Good Made Units column: 1 record with -4,662 units (row 69)
- **Root Cause:** Likely data entry error or system calculation issue
- **Impact:** Mathematically impossible value that would skew statistical analyses

Issue 3: Inconsistent Categorical Data

- Stoppage Type: "minor" (lowercase) appeared 4 times instead of "Minor"
- OEE Code: "NOo" (1 record), "NOO" (1 record), and "0" (8 records) instead of standard codes
- **Impact:** Fragmented category counts leading to inaccurate frequency distributions

Issue 4: Duplicate Records

- 28 complete duplicate rows identified (0.34% of dataset)
- **Impact:** Inflated metrics and biased statistical summaries

---

3. Data Cleaning Methodology

3.1 Handling Missing Values in OEE Code

Decision Rationale: The OEE Code categorizes downtime reasons, which is critical for operational analysis. Since only 15 values were missing (0.18%), we chose to preserve these records rather than delete them.

Implementation:

“ manufacturing_clean['OEE Code'].fillna('Unclassified', inplace=True)”

Why This Solution:

- **Preservation of data:** Maintains temporal and production information
- **Transparency:** "Unclassified" explicitly indicates missing classification
- **Analytical flexibility:** Enables separate analysis of unclassified stoppages
- **Small proportion:** 0.18% won't significantly impact overall patterns

Alternative Considered: Deleting rows with missing OEE codes was rejected as it would sacrifice valuable production data for a very small percentage of records.

3.2 Handling Complete Null Records

Decision Rationale: Five records had all fields null, providing zero analytical value.

Implementation:

“manufacturing_clean.dropna(inplace=True)”

Why This Solution:

- **Zero information content:** No salvageable data in these rows
- **Minimal impact:** Only 0.06% of dataset
- **Data integrity:** Removing completely empty rows is standard practice
- **No imputation possible:** Cannot reasonably estimate 13 missing values per record

---

3.3 Standardizing OEE Code Format

Decision Rationale: Inconsistent values ("NOo", "NOO", "0") fragmented what should be single categories, preventing accurate counting and analysis.

Implementation:

“manufacturing_clean['OEE Code'] = manufacturing_clean['OEE Code'].str.replace('NOo', 'NO') ”

“manufacturing_clean['OEE Code'] = manufacturing_clean['OEE Code'].str.replace('NOO', 'NO') ”

“manufacturing_clean['OEE Code'] = manufacturing_clean['OEE Code'].str.replace("0", 'Unclassified') ”

Why This Solution:

- **Standardization:** "NOo" and "NOO" were clearly typos of "NO" (likely data entry errors)
- **Semantic clarity:** "0" lacks meaning; "Unclassified" is descriptive
- **Pattern consistency:** Maintains two character code format (NO, CC, PM, Ru)
- **Minimal records affected:** Only 10 records total

Final Distribution:

- NO: 4,074 records (50.1%)
- CC: 4,052 records (49.8%)
- Unclassified: 18 records (0.2%)
- Ru: 9 records (0.1%)
- PM: 6 records (0.1%)

3.4 Correcting Negative Values

Decision Rationale: A single record showed -4,662 good units made, which is physically impossible for production metrics.

Implementation:

“manufacturing_clean['Good Made Units'] = manufacturing_clean['Good Made Units'].abs()) ”

Why This Solution:

- **Physical plausibility:** Production quantities cannot be negative
- **Likely data entry error:** The negative sign was probably erroneous
- **Magnitude preservation:** The absolute value (4,662) appears reasonable given the distribution
- **Simple correction:** Converting to absolute value is the most conservative fix
- **Single occurrence:** Only 1 record affected (0.01%)

Alternative Considered: Deleting the record was unnecessary since the magnitude appears valid and the error pattern (single negative sign) suggests a simple data entry mistake.

Verification: Post cleaning minimum value: 0.0 (logical minimum for production count)

3.5 Standardizing Stoppage Type Format

Decision Rationale: Four records contained "minor" (lowercase) instead of "Minor" (title case), creating artificial category fragmentation.

Implementation:

“manufacturing_clean['Stoppage Type'] = manufacturing_clean['Stoppage Type'].str.replace('minor', 'Minor') ”

Why This Solution:

- **Case sensitivity issue:** These represented the same category with formatting inconsistency
- **Dominant pattern:** "Minor" (title case) was the established standard (5,325 vs. 4 records)
- **Categorical integrity:** Maintains two clean categories: "Minor" and "Major"
- **String standardization:** Common best practice for categorical variables

Final Distribution:

- Minor: 5,329 records (65.5%)
- Major: 2,830 records (34.8%)

3.6 Removing Duplicate Records

Decision Rationale: 28 complete duplicate rows were identified, representing redundant information that would inflate frequency counts and skew analyses.

Implementation:

“manufacturing_clean.drop_duplicates(inplace=True) ”

Why This Solution:

- **Data redundancy:** Exact duplicates provide no additional information
- **Statistical bias:** Duplicates inflate frequencies and distort averages
- **Standard practice:** Duplicate removal is fundamental data cleaning
- **Small proportion:** Only 0.34% of dataset, minimal information loss

Verification Method: Used pandas' duplicated() function to identify rows where all 13 columns matched exactly across multiple records.

---

4. Data Cleaning Impact Summary

4.1 Quantitative Changes

Metric	Before Cleaning	After Cleaning	Change
Total Records	8,164	8,135	-29 (-0.36%)
Null Values (OEE Code)	15	0	-15
Complete Null Rows	5	0	-5
Duplicate Rows	28	0	-28
Negative Values	1	0	-1
OEE Code Categories	7	5	Consolidated
Stoppage Type Categories	3	2	Standardized

4.2 Data Quality Improvements

Completeness: 100% complete records (no null values remaining)

Consistency:

- Categorical variables standardized to uniform formats
- OEE Code reduced from 7 to 5 meaningful categories
- Stoppage Type reduced from 3 to 2 valid categories

Validity:

- All numerical values within plausible ranges
- No negative production quantities
- Logical consistency maintained

Uniqueness:

- Zero duplicate records
- Each row represents a distinct stoppage event

---

5. Rationale for Key Decisions

5.1 Why Imputation vs. Deletion for OEE Code

Decision: Impute "Unclassified" rather than delete rows

Reasoning:

1. **Preservation priority:** The missing OEE codes represented only 0.18% of data, but deleting would remove complete production records
2. **Information retention:** Other columns (dates, times, quantities) contained valuable operational data
3. **Analytical transparency:** "Unclassified" explicitly marks incomplete data for analysts
4. **Minimal bias:** Such a small proportion won't significantly affect category distributions

5.2 Why Absolute Value for Negative Numbers

Decision: Convert -4,662 to 4,662

Reasoning:

1. **Physical impossibility:** Negative production is conceptually invalid
2. **Error pattern:** Single occurrence suggests data entry mistake (extra minus sign)
3. **Magnitude validation:** The absolute value (4,662) falls within expected ranges for the "Bourbon Creams" product
4. **Conservative approach:** Preserves the quantity information while correcting the sign error
5. **Context support:** Surrounding records show similar magnitude values (4,414, 9,554)

5.3 Why String Replacement for Categorical Standardization

Decision: Use str.replace() for "minor"→"Minor" and "NOo"/"NOO"→"NO"

Reasoning:

1. **Pattern recognition:** Variations clearly represented same categories with typos

2. **Majority rule:** Followed the dominant formatting pattern (Minor: 5,325 vs minor: 4)
3. **Semantic equivalence:** "NOo" and "NOO" have no distinct meaning from "NO"
4. **Minimal intervention:** Simple string standardization without changing underlying meaning
5. **Reversibility:** Original values documented in this report if needed

---

6. Validation and Quality Assurance

6.1 Post-Cleaning Verification

Tests Performed:

1. **Null check:** manufacturing_clean.isnull().sum() → All columns: 0 nulls ✓
2. **Duplicate check:** manufacturing_clean.duplicated().sum() → 0 duplicates ✓
3. **Negative values:** manufacturing_clean['Good Made Units'].min() → 0.0 ✓
4. **Category counts:** Verified consolidated categories match expected totals ✓

6.2 Data Integrity Checks

Numerical Ranges:

- Total Biscuits Made: [0, 961,128] ✓
- Good Made Units: [0, 480,646] ✓
- Total Units Made: [0, 1,068,855] ✓

Categorical Integrity:

- Stoppage Type: Only "Minor" and "Major" ✓
- OEE Code: Only NO, CC, Unclassified, Ru, PM ✓
- All machine and product names consistent ✓

Temporal Consistency:

- Date range: July 1 - August 1, 2021 ✓
- No future dates or impossible timestamps ✓

---

7. Final Dataset Characteristics

Clean Dataset: 8,135 rows × 13 columns

Data Quality Metrics:

- Completeness: 100%
- Uniqueness: 100%
- Validity: 100%
- Consistency: 100%

Exported File: manufacturing_clean.xlsx

---

8. Recommendations for Future Data Collection

Based on the cleaning process, we recommend:

1. **Implement data validation rules** at the point of entry to prevent:
    - Negative production quantities
    - Inconsistent category formatting (enforce case sensitivity)
    - Null values in critical classification fields
2. **Create dropdown lists** for categorical fields (Stoppage Type, OEE Code) to eliminate typos
3. **Add database constraints** to:
    - Enforce non-negative values for quantity fields
    - Require OEE Code values
    - Prevent duplicate record insertion
4. **Establish data entry guidelines** with standardized formatting conventions
5. **Implement automated data quality checks** to flag anomalies in real-time

---

9. Conclusion

The data cleaning process successfully transformed a dataset with multiple quality issues into a robust, analysis-ready resource. Through targeted interventions imputation of missing classifications, correction of impossible values, standardization of categorical data, and removal of duplicates we achieved 100% data completeness and consistency while losing only 0.36% of records.

Each cleaning decision was made with careful consideration of data preservation, analytical validity, and operational context. The resulting clean dataset of 8,135 records provides a reliable foundation for analyzing manufacturing downtime patterns, identifying operational inefficiencies, and supporting data-driven decision-making for productivity improvements.

Key Achievement:

Maintained 99.64% of original data while improving quality metrics across all dimensions (completeness, consistency, validity, and uniqueness).