Exploratory Data Analysis (EDA) Report

1. Introduction

- **Objective:** The primary goal of this analysis is to understand the characteristics of the dataset and identify patterns or factors that influence the target variable, `status` (either "OK" or "NOK").
- **Dataset Overview:**
  - The dataset includes sensor readings from various steps of a manufacturing process, along with additional contextual features such as timestamps or weekdays.
  - The target variable, `status`, determines whether a produced part is acceptable ("OK") or defective ("NOK").

---

2. Data Overview

2.1. Structure of the Dataset

- **Number of Rows:** X
- **Number of Columns:** Y
- **Column Types:**
  - Numerical Features: Sensor readings (e.g., `s10_sensor2_gramm_step1`, `s8_sensor102_millimeter_step1`).
  - Categorical Features: Day of the week, status.

2.2. Missing Data

- **Summary:**
  - Some columns contained missing values, which were addressed using the KNN imputation method.
  - Missing values were found primarily in sensor readings and were filled based on patterns from the nearest neighbors.

---

3. Univariate Analysis

3.1. Target Variable (`status`)

- The target variable is imbalanced, with 80% "OK" parts and 20% "NOK" parts.
- **Insight:** The imbalance suggests a strong process baseline but warrants careful handling during predictive modeling to ensure minority class ("NOK") performance.

3.2. Numerical Features

- **Key Observations:**
  - Some features (e.g., `s10_sensor2_gramm_step1`, `s8_sensor102_millimeter_step1`) exhibit skewness, with outliers beyond the interquartile range (IQR).
  - Log transformation or robust scaling may help normalize skewed distributions for modeling.
- **Visuals:**
  - Histograms revealed the data for many sensor readings was concentrated within specific ranges, indicating possible sensor thresholds in the production line.

---

4. Bivariate Analysis

4.1. Numerical Features vs. Target (`status`)

- **Box Plots:** Key observations from box plots comparing sensor readings for "OK" vs. "NOK":
  - **`s10_sensor2_gramm_step1`:** Higher values are strongly associated with "NOK" parts.
  - **`s8_sensor102_millimeter_step1`:** This feature shows distinct separation, making it a potential key predictor.
  - **`s8_sensor68_millimeter_step1`:** Overlaps significantly between "OK" and "NOK" parts, indicating low predictive power.
- **Scatter Plots:**
  - Sensor combinations (`s10_sensor2_gramm_step1` vs. `s8_sensor67_millimeter_step1`) reveal clustering patterns that differentiate "OK" and "NOK" parts.

4.2. Categorical Features vs. Target

- **Weekday Analysis:**
  - Production on Mondays showed a slightly higher defect rate ("NOK"), possibly due to operational inefficiencies or environmental factors early in the workweek.

---

5. Multivariate Analysis

5.1. Correlation Heatmap

- **Key Findings:**
  - Sensors from the same process steps (e.g., `s8_sensor68_millimeter_step1` and `s8_sensor67_millimeter_step1`) are highly correlated (r>0.8r > 0.8r>0.8), indicating potential redundancy.
  - Target variable `status` shows weak correlations with individual sensors, suggesting the need for feature interactions or combinations to improve predictive modeling.

5.2. Principal Component Analysis (PCA)

- **Explained Variance:**
  - The first two principal components (PC1 and PC2) capture ~93% of the variance, demonstrating that dimensionality reduction is feasible.
- **Insights:**
  - Clustering in the PCA-transformed space suggests separability between "OK" and "NOK" parts, even with reduced dimensions.

---

6. Outlier Analysis

- **Z-Score Method:**
  - No significant outliers detected, likely due to assumptions of normality not holding for all features.
- **IQR Method:**
  - Several outliers identified for most sensor readings, particularly in "NOK" parts.
  - Handling Strategy: Outliers were either capped to the IQR bounds or removed from the dataset.

---

7. Key Insights and Conclusions

1. **Critical Features:**
   - Features like `s10_sensor2_gramm_step1` and `s8_sensor102_millimeter_step1` exhibit strong separability between "OK" and "NOK" classes.

    - These should be prioritized during feature selection and modeling.
2. **Feature Redundancy:**
    - High correlations among certain sensors suggest the potential for dimensionality reduction or feature selection.
3. **Outlier Impact:**
    - Outliers predominantly exist in "NOK" parts, suggesting that defects may stem from extreme sensor readings. Addressing these extremes could improve production quality.
4. **Class Imbalance:**
    - The imbalance in the target variable requires attention during model development to ensure the minority class ("NOK") is adequately predicted.
5. **Time-Based Effects:**
    - Weekday analysis hints at operational inefficiencies early in the week, warranting further investigation into process conditions or scheduling.

---

8. Recommendations

1. **Data Preprocessing:**
    - Normalize or scale skewed features and consider robust transformations for highly variable sensors.
    - Address multicollinearity through PCA or feature selection.
2. **Operational Improvements:**
    - Investigate process deviations on Mondays and focus on critical sensors associated with "NOK" outcomes.
3. **Model Development:**
    - Use ensemble methods (e.g., Random Forests) to capture interactions among features.
    - Implement SMOTE or similar techniques to balance the target classes.
4. **Real-Time Monitoring:**
    - Deploy rules or machine learning models using critical sensor thresholds to predict and prevent "NOK" outcomes in real-time.

Appendix: Supporting Graphs

- Box plots, scatter plots, correlation heatmaps, and PCA visualizations are included in this report to provide a visual understanding of the data patterns.