Quality Analysis Report

Defect and Repair Cost Insights

1. Introduction

This report analyzes the `quality.csv` dataset, containing 1,000 records of product defects. The goal is to uncover patterns in defect types, severity, and repair costs, and to develop a predictive model for repair costs using machine learning. The analysis uses Python libraries including Pandas, NumPy, Matplotlib, Seaborn, and Scikit-learn.

2. Data Overview

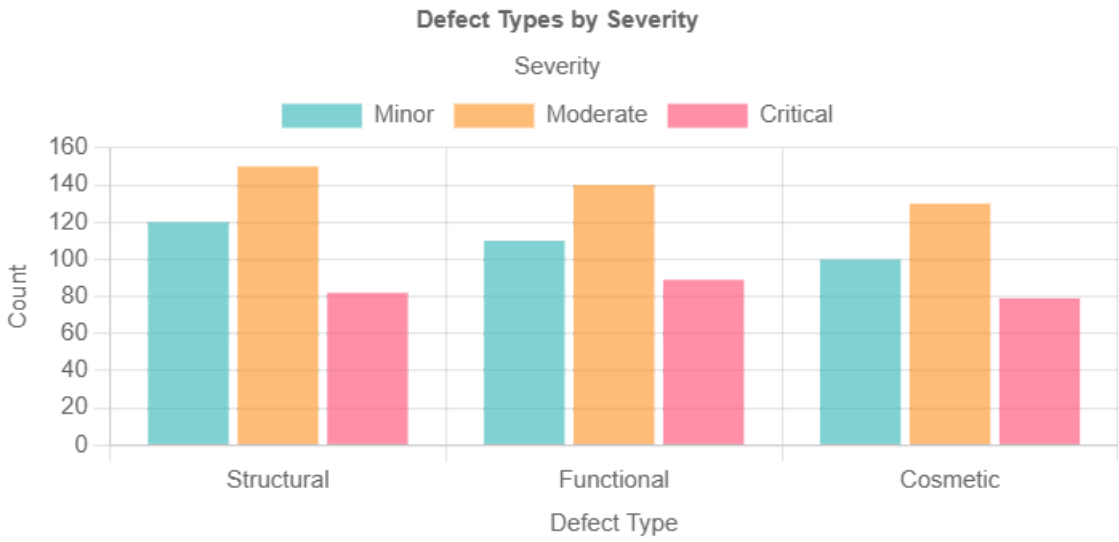
The dataset includes 8 columns: `defect_id`, `product_id`, `defect_type`, `defect_date`, `defect_location`, `severity`, `inspection_method`, and `repair_cost`. Key points:

- **Missing Values**: 402 missing entries in `defect_date`.
- **Data Cleaning**: Converted `defect_date` to datetime and `repair_cost` to numeric.
- **Repair Cost**: Ranges from $10.22 to $999.64 (mean: $507.63).

3. Exploratory Data Analysis

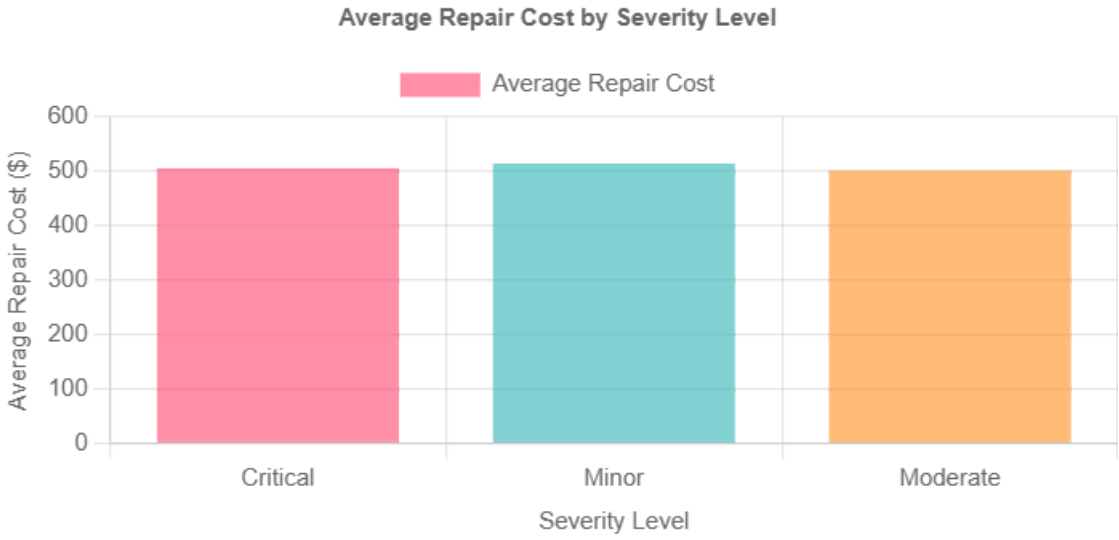
3.1 Defect Distribution

The dataset contains 352 Structural, 339 Functional, and 309 Cosmetic defects. Below is a visualization of defect types by severity.



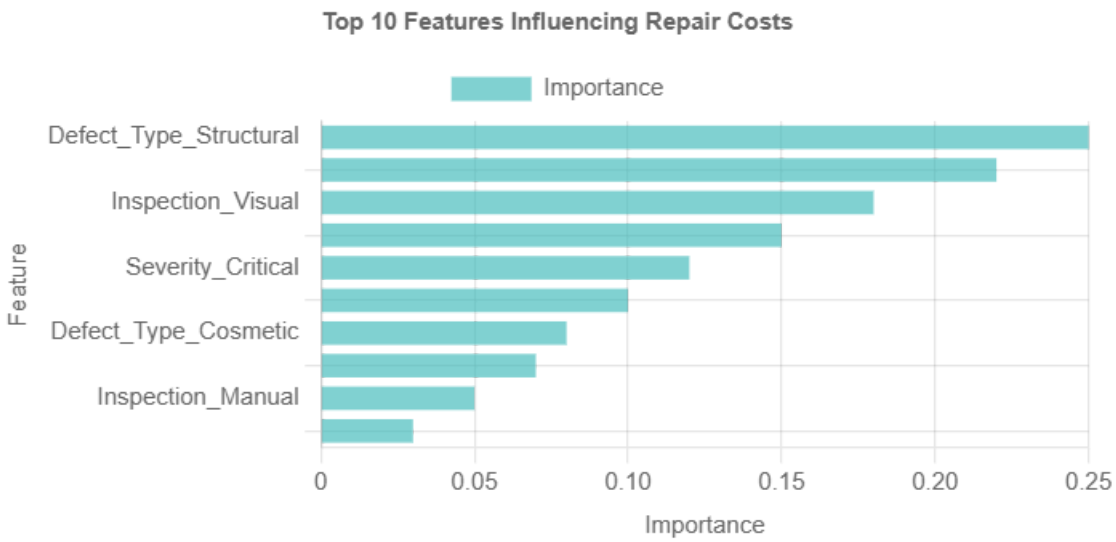
3.2 Repair Cost Analysis

Average repair costs by severity: Critical ($505.87), Minor ($514.43), Moderate ($501.63). Minor defects have the highest average cost, warranting further investigation.



4. Predictive Modeling

A Linear Regression model was trained to predict repair costs using one-hot encoded categorical features (`defect_type`, `severity`, `inspection_method`). The model achieved a Mean Absolute Error (MAE) of $255.08 on the test set.



5. Key Findings

- **Defect Distribution**: Structural defects are most common (352).
- **Repair Costs**: Minor defects have the highest average cost ($514.43).
- **Model Performance**: Linear Regression MAE of $255.08 suggests need for advanced models.
- **Data Issue**: 402 missing `defect_date` values impact temporal analysis.

6. Recommendations

1. Address missing `defect_date` values via imputation.
2. Use non-linear models (e.g., Random Forest) to improve predictions.
3. Investigate high repair costs for Minor defects.
4. Prioritize quality checks for Structural defects.
5. Analyze temporal defect trends after data cleaning.

7. Conclusion

The analysis highlights Structural defects as the most frequent and Minor defects as the costliest on average. The Linear Regression model provides a baseline but suggests further modeling improvements. Addressing data quality and focusing on high-cost defects will optimize quality control.