**Data Analyst Technical Challenge (Statistical Focus)**

Made by: Andres Felipe Cardenas Parga

Date: 04/05/2025

**Overview**

BeerBo Printing provided me with a database containing three different files for study and analysis. My role as a data analyst is to examine the data and transform it into valuable information and insights, enabling the organization to make informed decisions.

1. **Analysis Process**

The analysis was conducted in four main steps:

**Data Cleaning and Preparation:** Imported and cleaned datasets, including ProductionMetric, DeviceProperty, and Quality. Addressed anomalies such as end\_time being earlier than start\_time and removed duplicate columns.

**Downtime Analysis:** Examined unplanned and planned stop times, calculated proportions, and visualized downtime distributions.

**Production & Quality Analysis:** Investigated reject rates, identified common reject reasons, and compared production efficiency across devices.

**Performance Comparison:** Compared key metrics like average downtime and reject rates across shifts and teams.

1. **Key Findings**

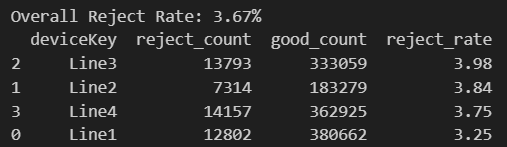
**Downtime Analysis:**

* **Unplanned** downtime accounted for **76%** of total stop time, with "**Main B Bus Undervolt**" being the most frequent reason (62,982 seconds), while, planned stop time accounted for 24% of total stop time. *See figure 1.*
* **Line 3** exhibited **the lowest** mean unplanned downtime (mean = 160.58 seconds) **compared to Line 1** (Mean = 176.74 seconds). *See figure 2.* However, these lines exhibit similar behaviour.
* The **Pareto analysis** reveals that "Main B Bus Undervolt" is the leading cause of unplanned Stop Time, accounting for approximately **13.3% of the total unplanned stop time**, followed by "**Electrical Fault**" and "**Jam**", which collectively contribute to **over 25%** of the downtime.

It is clearly seen that the top five reasons could potentially **mitigate over 50%** of the unplanned downtime. This insight underscores the importance of **prioritizing maintenance** and **process improvements** for these specific issues to enhance overall productivity and reduce operational disruptions.

**Production & Quality Analysis:**

* The overall reject rate was 3.67%, with "Detected by Max WIP" and "Reject" being the most common reasons.
* "Reject" and "Detected by Max WIP" are the main reasons for most rejections. These categories suggest systemic issues that need immediate attention. Breaking down the ‘Reject’ reason might help understand the behaviour better.
* Line 3 had the highest reject rate (3.98%) among all lines.



* The analysis of *average good count per hour by device key* revealed **significant differences across production lines**. Using ANOVA, the F-statistic (25.89) and p-value (<0.05) **confirmed statistically significant variations** in production efficiency between Line 1, Line 2, and Line 3. This suggests that operational factors or equipment **performance may differ across lines**. See figure 3.
* The scatter plot reveals **no clear relationship** between unplanned stop time and reject count, a finding confirmed by the correlation heatmap, which shows weak or negligible correlation between these two variables. **This makes sense in practice** — the number of product rejects isn't necessarily caused by how often a machine stops unexpectedly. However, **strong positive correlations** between good count, ideal time, and run time highlight a logical connection: *when machines run efficiently and consistently, they tend to produce more good units.* These variables are clearly interdependent and suggest that **optimizing them together could lead to meaningful improvements** in operational performance. *See Figure 4 and Figure 5.*

**Performance Comparison:**

* First Shift had the highest average downtime (166.86 seconds), while Second Shift had the lowest (163.92 seconds).
* The bar plots highlight that certain shifts and devices **consistently perform better**, while others have **higher reject rates**. Although these results do not appear to be significantly different, these insights **suggest** opportunities for targeted process improvements and resource allocation to optimize productivity and reduce rejects.
* **Average good count per hour** varied significantly across devices, with Line 1 achieving the highest rate (1,835.55 units/hour).
* Figure 6 shows daily unplanned stop time. This graph can support the creation of a KPI for unplanned stop time, as it provides **daily aggregate stop time information**.
* Furthermore, an analysis of **rejected goods** **by parts** can provide additional insights into whether these variables are **related to the rejection reasons**. An interactive graph created using Plotly was developed to interpret the rejection findings for each product.

**Conclusions**

* The high proportion of unplanned downtime suggests a need for preventive maintenance, particularly addressing "Main B Bus Undervolt" issues.
* The company may focus on **reducing unplanned downtime.**
* Line 3's higher reject rate and downtime indicate potential inefficiencies that require further investigation.
* Shifts with higher downtime may benefit from targeted training or process optimization.
* Generating **various KPIs** from analysis is **crucial** for the company to monitor key findings and produce additional results. I Consider including these KPIs in a dashboard for daily/weekly monitoring by stakeholders.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Question | Ho | H1 | Test used | Reason Test Use | P-value | Reject Ho? | Conclusion | Business Interpretation |
| Compare mean unplanned downtime between two production lines | μA = μB | μA =! μB | two sample T-test | Sample size is >30 (Central Limit Theorem may apply) Data is approximately normally distributed You want to compare means | 0.4202 | No | There is no statistically significant difference in the mean from line1 and the mean from line2. | There are evidence to confirm that the behaviour lines are not significantly different. This means every line could be treated similarly due to their behaviour |
| Test differences in reject rate across shifts | μFirst = μSecond=μThird | At least one μ is different | ANOVA and Tukey's HSD test | ANOVA is used to analyse the variance for two or more groups. | 0.5297 | No | There is no statistically significant difference in the mean reject rates across the three shifts. | Even though operationally there might be differences across shifts (like team size, experience, etc.), statistically, these shifts perform similarly in terms of reject rate. |
| Association between process-state reason and shift | The two variables are independent — the distribution of reasons is the same across shifts. | The two variables are dependent — some reasons happen more or less in certain shifts. | Chi-square test of independence | Chi-square test the independence between two variables | 0.11 | No | Indicates there is no statistically significant association between process\_state\_reason and shift. | This suggests that the reasons for stop time aren't related to the shift performance. |
| Significance of the correlation between unplanned\_stop\_time and reject\_count | There is no correlation between unplanned\_stop\_time and reject\_count | There is a correlation (positive or negative). | Spearman (ρ) | Data has significant oultliers and relationship may be non-linear but monotonic | 1.81E-29 | Yes | There is a moderate negative correlation (pearson = -0.19), statistically significant at the 0.05 level. | This indicates that higher unplanned stop time is associated with an increase in rejects which makes sense in practice. |

**4. Visualizations**

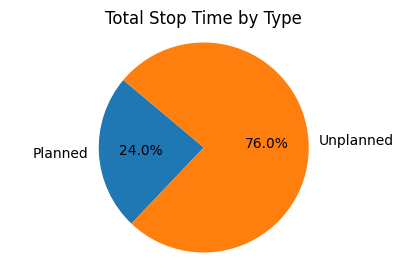


Figure . Total Stop Time by Type

A diagram of a box diagram

AI-generated content may be incorrect.

Figure . Boxplot of Total Stop Time By Device Key

A diagram of a box plot

AI-generated content may be incorrect.

Figure . Boxplot of Good Count by Device Key

A graph with blue dots

AI-generated content may be incorrect.

Figure . Scatter plot of unplanned stop time vs Reject Count

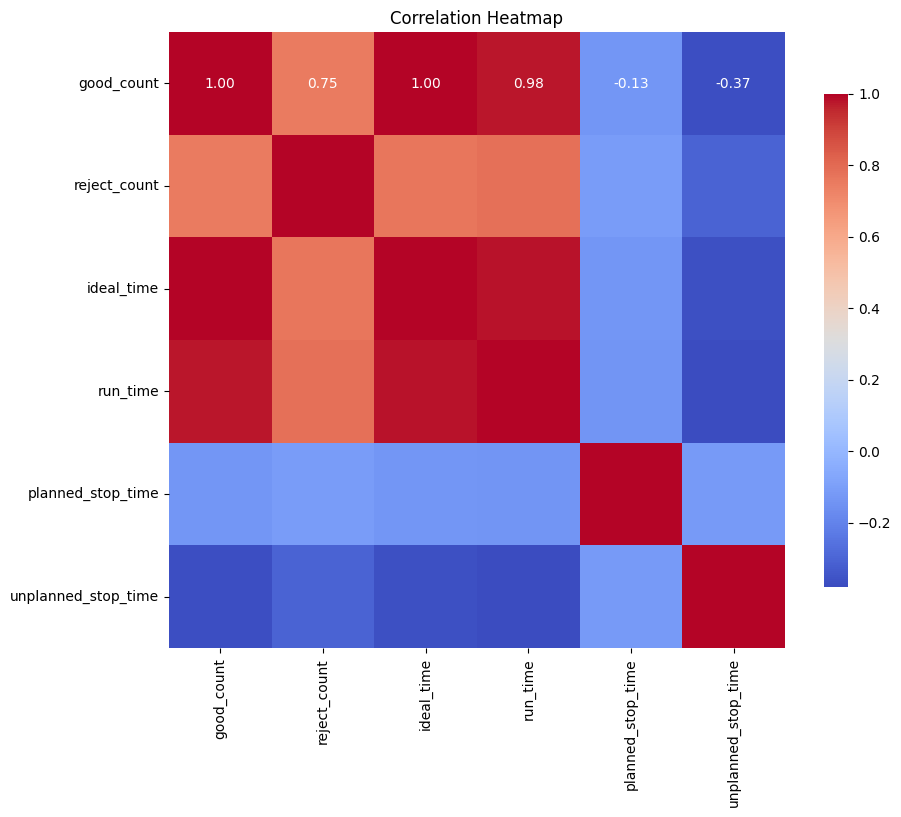


Figure . Correlation Heatmap

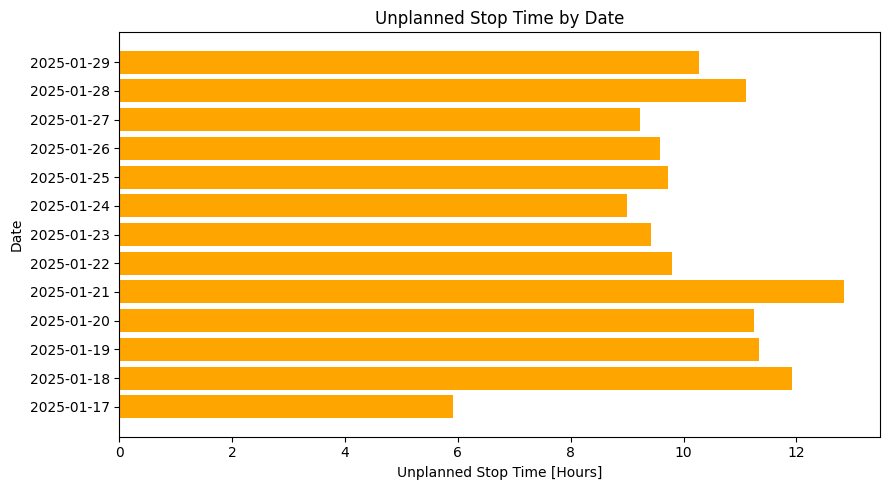


Figure . Unplanned Stop Time by Date