Baseline Investigation

1 Question

What is the purpose of this investigation?

The goal of this study is to pinpoint the root cause of inconsistencies in the output variable, known as y_{300}, by analyzing its variability and forming initial hypotheses about the primary factors influencing these discrepancies.

2 Plan

How do we formulate our investigation strategy?

Our strategy involves a meticulous selection of the study population while adhering to a budget constraint for sampling, which has been capped at under $1000, with a final expenditure set at 525$. This budgetary decision enables us to systematically organize and collect data across 15 shifts over 5 days, obtaining 35 observations per shift, resulting in a total of 525 data points.

To uncover any underlying issues, we'll employ statistical summaries including mean, median, min, max, and standard deviation, alongside the process capability index (Ppk), to provide a comprehensive overview. Visual aids like histograms will depict the distribution of output values, and boxplots segmented by day and shift will highlight potential output variability. Furthermore, we'll use studentized residual plots to detect outliers, thereby enriching our investigative process.

3 Data

Below is a view of our dataset. This table shows the first 6 values.

daycount	**shift**	**partnum**	**y300**
1	1	1	5.2
1	1	2	2.6
1	1	3	2.8
1	1	4	-1.0
1	1	5	0.8
1	1	6	-8.6

Table 1: First 6 values of dataset

4 Analysis

The table below presents a detailed statistical analysis of the ‘y300‘ values, offering insights into the process variability:

Statistic	**Value**
Minimum	-15.00
Maximum	13.00
Mean	-0.54
Median	-0.40
Standard Deviation	4.81
Process Capability Index (PPK)	0.66

Table 2: Statistical Overview of y300 Values

The comprehensive statistical summary of the ‘y300‘ variable provides a deep dive into the intricacies of process performance and variability. The observed range in ‘y300‘ values, extending from a minimum of -15 to a maximum of 13, not only highlights the extent of variability but also underscores the challenges in maintaining consistent process outcomes. This wide variation is a clear indicator of the process's dynamic nature and the potential for unpredictable results.

The dataset's central tendency, represented by a slightly negative mean of -0.54 and a median of -0.40, suggests a mild left skew in the distribution. This skewness is particularly noteworthy as it implies a tendency towards lower output values, possibly hinting at systemic issues within the process that predispose it to underperformance. The presence of outliers, as suggested by this skew, could be contributing to this tendency, further complicating the process's predictability and control.

The standard deviation, a measure of the spread of the data, stands at 4.81, indicating a substantial degree of dispersion among the ‘y300‘ values. This significant spread points to the lack of uniformity in the process outcomes, raising concerns about the process's stability and the consistency of the outputs. Such variability can pose significant challenges in process control, necessitating rigorous monitoring and adjustment strategies to ensure that the process remains within acceptable bounds.

Moreover, the Process Capability Index (PPK) of 0.66, markedly below the industrial benchmark of 1.33, raises critical concerns regarding the process's capability to meet specified limits consistently. This shortfall in the PPK value is a stark indicator of the process's limitations in achieving the desired quality standards, emphasizing the urgent need for targeted process improvements. Addressing these capability gaps is essential not only for enhancing process performance but also for ensuring that the outputs reliably meet quality standards, thereby maintaining customer satisfaction and competitive advantage.

In light of these findings, it is evident that the process under investigation is characterized by notable variability and a tendency towards lower performance, as indicated by the left-skewed distribution and the wide range of ‘y300‘ values. The significant standard deviation further highlights the challenges in achieving process consistency, while the subpar PPK value underscores the critical need for process optimization efforts. Addressing these issues will require a multifaceted approach, focusing on identifying and mitigating the root causes of variability, optimizing process parameters to enhance stability, and implementing robust quality control measures to ensure consistent and reliable process outcomes. Through these efforts, it is possible to achieve significant improvements in process performance, ultimately leading to higher quality outputs, improved customer satisfaction, and enhanced operational efficiency.

4.1 Histogram

The examination of the ‘Y300‘ data provides insightful observations regarding its distribution, which largely mirrors the characteristics of a normal distribution, marked by a prominent, singular peak that signifies a unimodal pattern. This peak, the most densely populated region of the histogram, reflects the central tendency of the distribution, offering a visual representation of where the bulk of the data points congregate.

A notable aspect of the distribution is its mild left skewness, which is visually apparent through the elongated tail stretching towards the left side of the histogram. This skewness is accentuated by the observation that several bins on the left side contain a higher frequency of data points compared to those on the far right, suggesting a concentration of values that are lower than the mean, thus deviating from the symmetrical expectation of a perfectly normal distribution.

Despite these minor deviations, the overall contour of the ‘Y300‘ value distribution retains the quintessential bell-shaped curve that is synonymous with normal distributions. This resemblance implies that, while there are anomalies, the distribution largely adheres to the principles of normality, with the bulk of the data points clustering around the mean, and the frequencies diminishing symmetrically as one moves away from the center, albeit with a slight tilt towards the lower end.

Such a distribution, with its mild leftward skew, indicates that while the process generating the ‘Y300‘ values is largely stable and predictable, there are instances of lower-than-average outcomes that could be significant for understanding the underlying dynamics of the process. These anomalies, while not severe enough to fundamentally alter the distribution’s normal-like appearance, warrant closer inspection to identify potential causes or factors that might be contributing to these lower values, ensuring a comprehensive understanding of the process and its variabilities.

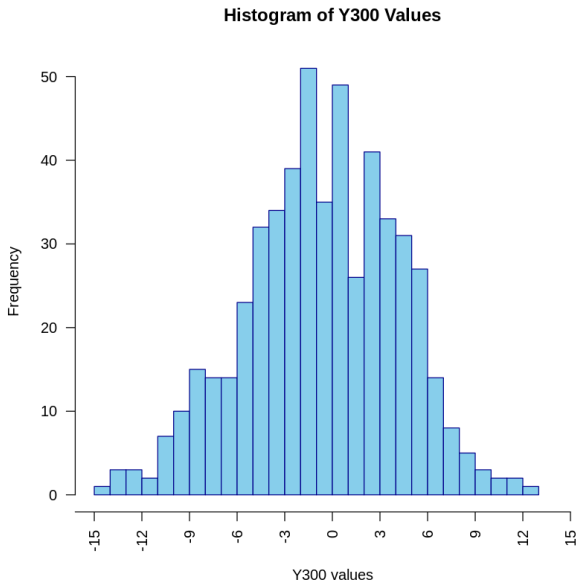

Figure 1: Expanded View of the Histogram for Y300 Values

4.2 Studentized Residual Analysis

The studentized residual plot depicted in Figure 2 visualizes the standardized residuals of the ‘y300‘ variable after fitting a statistical model, plotted against the observation indices. In an ideal scenario, where the model perfectly aligns with the data, these residuals should scatter randomly around the zero line, indicating no apparent patterns that would contradict the model's assumptions.

Upon examination, the residuals predominantly scatter in a random fashion around the horizontal zero line, suggesting a good fit of the model to the data without systematic biases. However, it's noteworthy that a few residuals, especially those dipping below the -2 threshold, stand out as significant deviations from the expected norm. These points represent outliers that might exert undue influence on the model's predictive accuracy and warrant further investigation to discern their impact on the overall analysis.

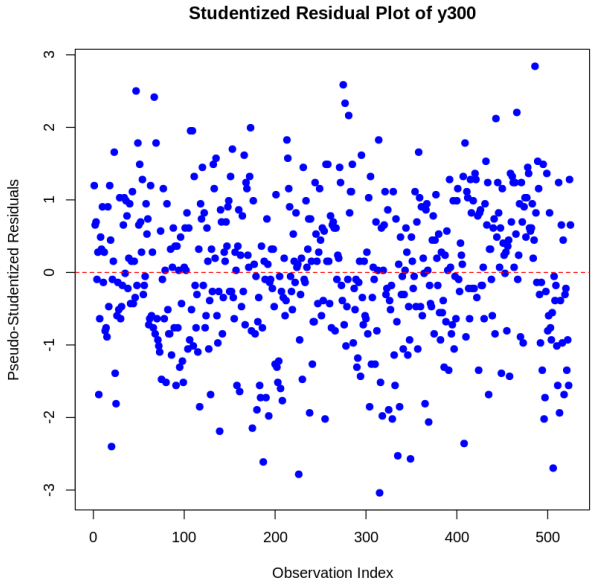

Figure 2: Detailed Studentized Residual Plot for y300

4.3 Boxplot Grouped by Days

The boxplot in Figure 3 provides a visual summary of the 'y300' values across five consecutive days, effectively capturing the data's spread and central tendency for each day. Each box in the plot represents the interquartile range (IQR), enclosing the middle 50

However, the plot reveals varying degrees of spread and range among the days, with days 3 and 5 exhibiting more pronounced variability. This is indicative of shifts in the data distribution, which could be due to external factors or inherent process variations. Additionally, the presence of outliers on days 1, 3, and 5, marked by points beyond the whiskers of the boxplots, highlights individual observations that stand out significantly from the rest. These outliers merit further investigation to understand their causes and potential impact on the overall analysis.

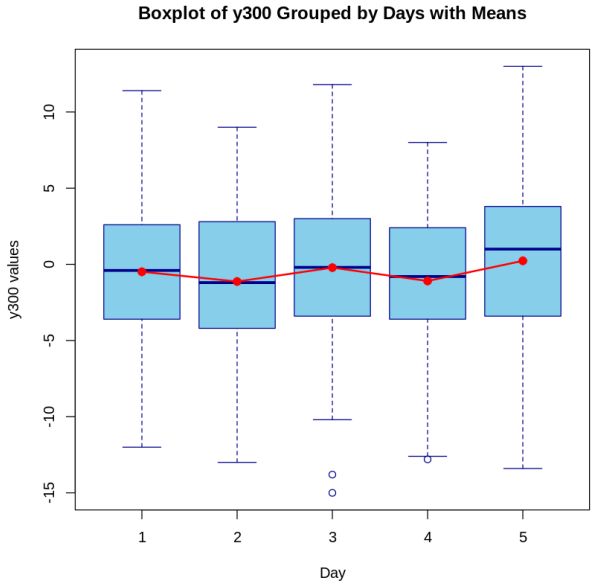

Figure 3: Enhanced Boxplot Visualization of y300 Values by Day

4.4 Boxplot Grouped by Shifts

In Figure 4, the distribution of 'y300' values is segmented by work shifts, with each shift's mean value highlighted by a red dot and interconnected by a line to illustrate the trend across shifts. The plot provides a nuanced view of how the mean values adjust slightly from one shift to another, with shift 2 standing out due to its marginally elevated mean compared to the others. This shift-wise comparison reveals subtle variations that could be indicative of differing operational efficiencies or conditions.

Shift 1 is characterized by a notably wider spread of values, as seen in the extended range of its boxplot, pointing to a higher degree of variability in data during this period. Such variability warrants a closer look to identify any underlying factors that may contribute to this observation. Additionally, the presence of outliers in shifts 1 and 3, represented by individual points outside the typical range, marks these observations as significantly different from the bulk of the data. These outliers are particularly noteworthy, as they may signal exceptional circumstances or data points that deviate from the expected pattern, meriting further analysis to uncover their origins and implications.

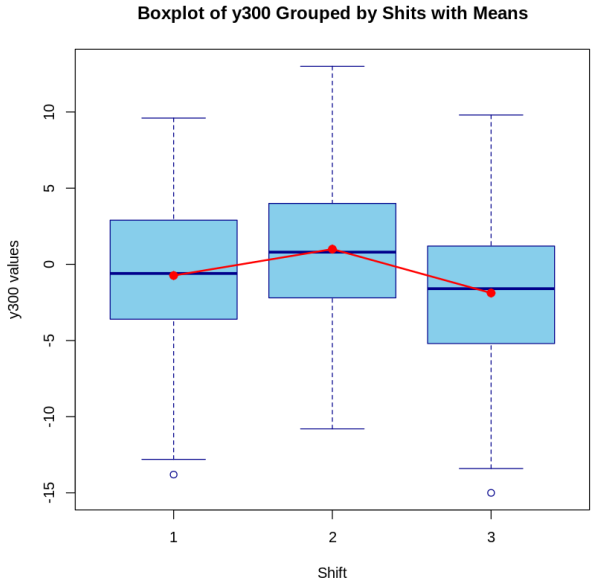

Figure 4: Detailed Boxplot Analysis of y300 Values by Work Shifts

4.5 Boxplot Grouped By Shifts and Days

The graph illustrates the variability of a variable (‘y300‘) across various shift and day combinations, with each boxplot representing a unique pair labeled as "Shift.Day" (for instance, 1.1 for Shift 1 on Day 1, 2.1 for Shift 2 on Day 1, and so forth).

Shift-day pairs such as 3.1 (Shift 3 on Day 1) and 3.5 (Shift 3 on Day 5) showcase wider boxplots, indicating a greater range of ‘y300‘ values during these periods, which points to increased variability. On the other hand, combinations like 1.1 (Shift 1 on Day 1) and 2.5 (Shift 2 on Day 5) feature narrower boxes, signifying less variability and suggesting more consistency in ‘y300‘ values within these specific intervals.

The median values across the shift-day combinations do not exhibit a consistent trend of either increase or decrease, implying that ‘y300‘ values do not follow a straightforward pattern that can be attributed solely to the progression of shifts or days. This lack of a clear pattern underscores the complexity of the factors influencing ‘y300‘ values and necessitates a more detailed analysis to uncover underlying trends or causes.

Notably, outliers are observed in several shift-day combinations, such as 1.1, 3.2, 1.4, and 3.4, highlighting instances where ‘y300‘ values were exceptionally divergent from the majority. These outliers are of particular interest as they may reveal specific incidents or conditions on those shifts and days that significantly deviated from the norm, warranting further investigation to understand their impact and origins.

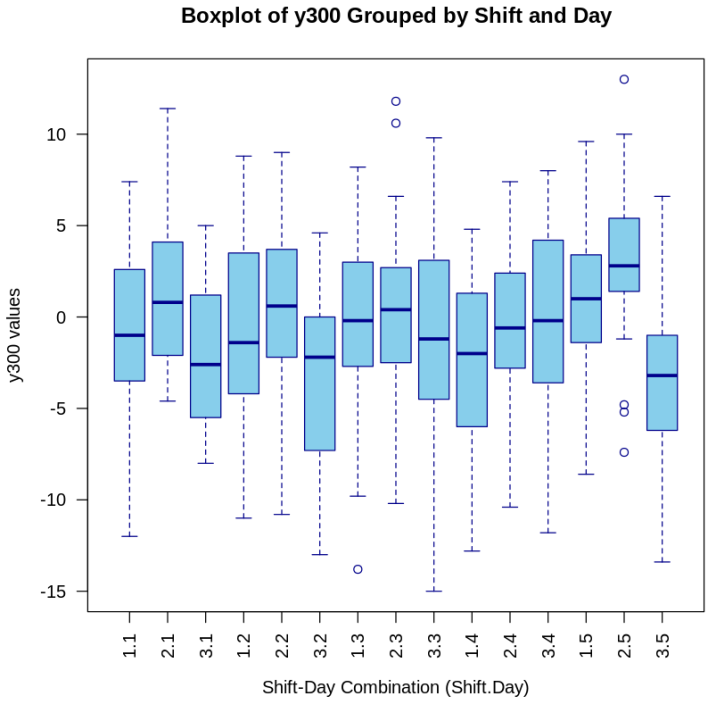

Figure 5: Comprehensive Boxplot Analysis of y300 Values by Shifts and Days

5 Conclusion

Our comprehensive analysis of the ‘y300‘ variable has brought several critical issues to light regarding the process variability and stability. The data, characterized by a substantial range in ‘y300‘ values, underscores the presence of significant fluctuations in process outcomes, which is further corroborated by the observed left-skewed distribution. This skewness suggests a tendency towards lower output values and hints at the existence of outliers that could be skewing the distribution.

The Process Performance Index (Ppk) value, recorded at 0.6562357, falls significantly below the industrial benchmark of 1.33. This discrepancy highlights a concerning gap in the process's ability to consistently meet the specified limits, pointing towards underlying issues in process capability and reliability. The low Ppk value serves as a clear indicator that there are substantial opportunities for process improvement and optimization to achieve a higher level of quality and consistency in outputs.

While the studentized residual plot did not reveal any systematic patterns that would question the appropriateness of the statistical model used, the presence of outliers identified in the analysis raises important considerations about the process's stability and reliability. These outliers, representing values that deviate markedly from the majority, could be symptomatic of special cause variations that need to be investigated and addressed to ensure a more predictable and controlled process.

Furthermore, the variability observed in the boxplots across different days and shifts suggests that external factors, possibly related to operational or environmental conditions, may be influencing the output. These variations highlight the need for a more granular investigation into how different conditions or practices across shifts and days might be contributing to the observed process variability.

In conclusion, our analysis underscores the critical need for targeted process improvements aimed at reducing variability and managing outliers. By delving deeper into the root causes of the identified issues, such as the sources of the outliers and the factors driving day-to-day and shift-to-shift variability, we can develop targeted strategies for process optimization. Addressing these challenges will not only enhance process stability and reliability but also ensure that the outputs are more consistent and within the desired specifications. Such improvements are essential for maintaining quality, meeting customer expectations, and achieving operational excellence in a competitive environment.