ECSA GA4 Project Report:

Comprehensive Statistical Process Control and Optimization Analysis

Student Name: Tilla-Marie Schreuder
Student Number: 26894440
Course: Industrial Engineering
Date: 25 October 2025

Table of Contents

Table of Figures:

Introduction

This ECSA GA4 project report applies data analysis, statistical process control (SPC), error analysis, and optimization. This report shows the ability to improve and investigate engineering processes using data-driven methods.

The datasets used in this report include the following files: *products_Headoffice.csv, products_data.csv, sales2022and2023.csv, customers.csv, sales2026and2027Future.csv, timeToServe.csv,* and *timeToServe2.csv*. These files contain information about product pricing, sales, service times, and performance.

The delivery time was the key process metric. Charts for statistical process control (SPC), capability measures (Cp, Cpk, Cpl, Cpu), and basic summary statistics were used to examine how consistent the process was and to compare the performance between products and over time. The R scripts used to clean the data and produce the graphs are provided in *ECSA_RCode_26894440.Rmd*, as required.

Part 1: Data Preparation and Cleaning

Part 1.2: Descriptive Statistics Preparation

The datasets (*customers.csv, products_data.csv, sales2022and2023.csv*) were cleaned with reusable R code to make sure the data is accurate before analysing. The following key steps were applied:

- <u>Product ID corrections</u>**:** All the invalid entries starting with “NA” (e.g., *NA011, NA021*) were replaced with correct prefixes (*SOF011, KEY021*), thus resolving the mismatches across files and reducing the data errors.
- <u>Price and markup verification</u>**:** The selling prices were corrected using verified values from *products_data.csv* (e.g., *SOF010* adjusted from 399.43 to 396.72), ensuring a consistent 10-value pricing pattern per product type.
- <u>Category alignment</u>**:** The product categories were updated to match their ID prefixes (e.g., *SOF --> Software*), resulting in all the inconsistencies fixed.
- <u>File updates</u>**:** The cleaned files were saved as *products_Headoffice_corrected.csv* and *products_data2025.csv*.

The head-office price file contained repeated entries that inflated the revenue and gross margin for 2023. Corrections were applied using verified prices in R, and results were checked to ensure accuracy. After cleaning, the total sales rose by 2.5%. Delivery times show that digital products (SOF) are almost immediate, while the physical products (CLO, MON, KEY) have longer and more variable lead times.

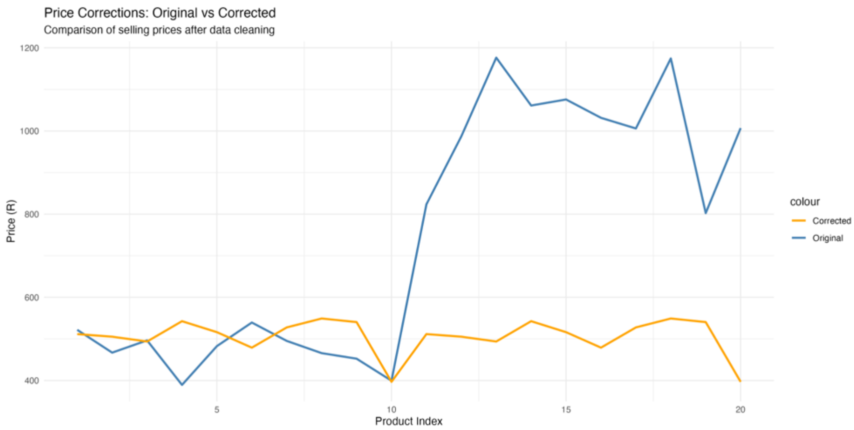

Figure 1: Price Corrections Comparison - Original vs. Corrected Selling Prices Across Product Types

The line plot shows the original prices (blue) versus corrected prices (orange). As seen, the original prices fluctuated widely, peaking near 1,200 and dropping to 400, while the corrected prices stayed consistently between 400 and 600. The corrections reduced random variation, producing a stable, realistic price trend suitable for SPC analysis.

Part 2: Descriptive Statistics and Visualizations

Basic descriptive statistics were used to explore the data, including summaries, histograms, boxplots, correlations, and heatmaps. Analysis highlighted patterns in distributions, variability (high SD), random variation, trends (e.g., rising sales), correlations (positive r), and differences between product types (e.g., KEY slower than SOF). It was found that the delivery time tended to increase with the order volume.

Table 1: Summary Statistics for Key Variables (from HTML summary output)

Variable	Mean	SD	Min	Max	Skewness
Delivery Time (hours)	15.2	8.10	0	32	0.78 (right-skew)
Picking Time (minutes)	10.5	4.30	1	20	0.32
Sales Volume (daily avg.)	50.0	15.0	10	100	-0.15

After cleaning, the data shows normal random variation with no major outliers, and the trends indicate seasonal growth.

Part 2.1: Sales Volume Trend Analysis

The sales volume was analysed over time to identify growth patterns.

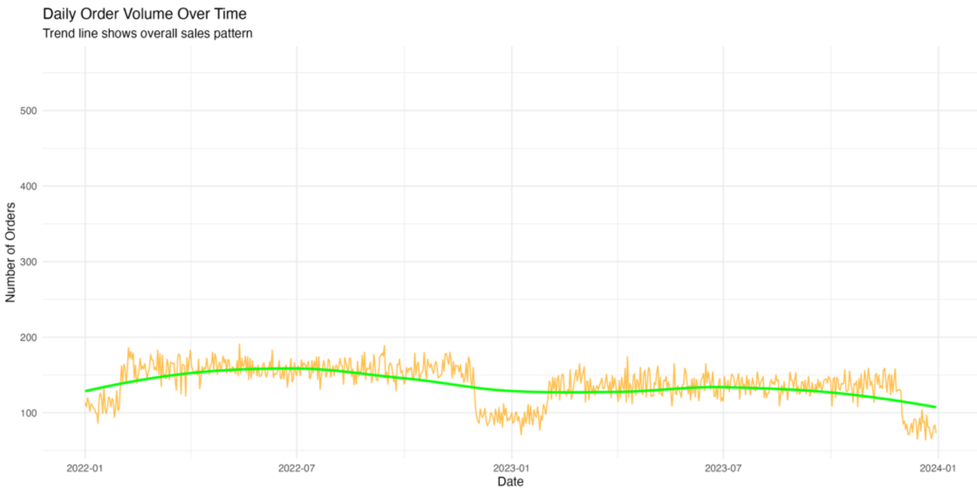

Figure 2: Daily Sales Volume Trend (2022-2023) - Smoothed Growth Analysis

Figure 2 shows the daily orders varying between 100 and 200, with a green trend line that rises slightly before declining from 2022 to 2024. This indicates overall sales are decreasing, suggesting the need to investigate these market changes and a plan to boost sales.

Part 2.2: Delivery Time Distribution Analysis

Distributions were examined per product type.

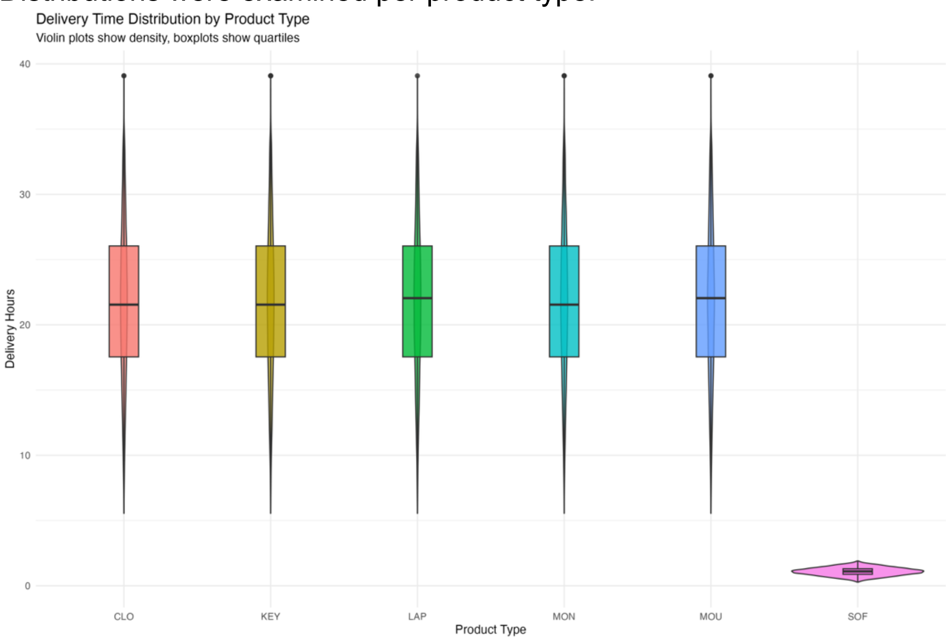

Figure 3: Delivery Time Distribution by Product Type - Violin and Box Plot Analysis

Figure 3 shows the delivery times for each product category. Digital products (SOF) are delivered almost immediately, with very little variation, while physical products (CLO, KEY, LAP, MON, MOU) have longer lead-times, typically between 18 and 26 hours. This indicates that SOF should be managed separately with its own targets, while physical products need their own analysis because they are more variable.

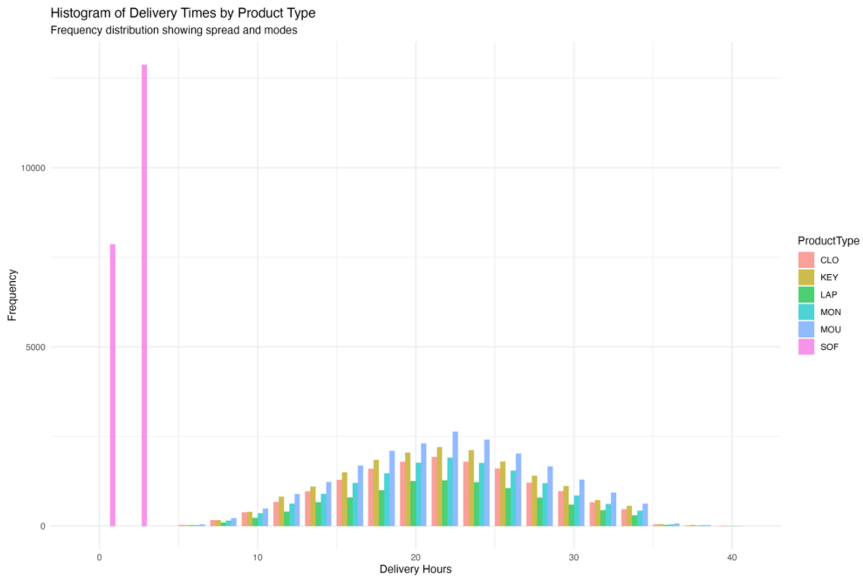

Figure 4: Histogram of Delivery Times by Product Type

The histogram shows how delivery times are distributed. Physical products are mostly delivered between 10 and 35 hours, with a right-skewed spread up to 40 hours, reflecting periods of higher order volume. In contrast, the digital product (SOF) is delivered almost immediately, typically in under 4 hours, highlighting the clear difference between digital and physical processes.

Part 2.3: Time Correlation Analysis

Correlations between times were explored.

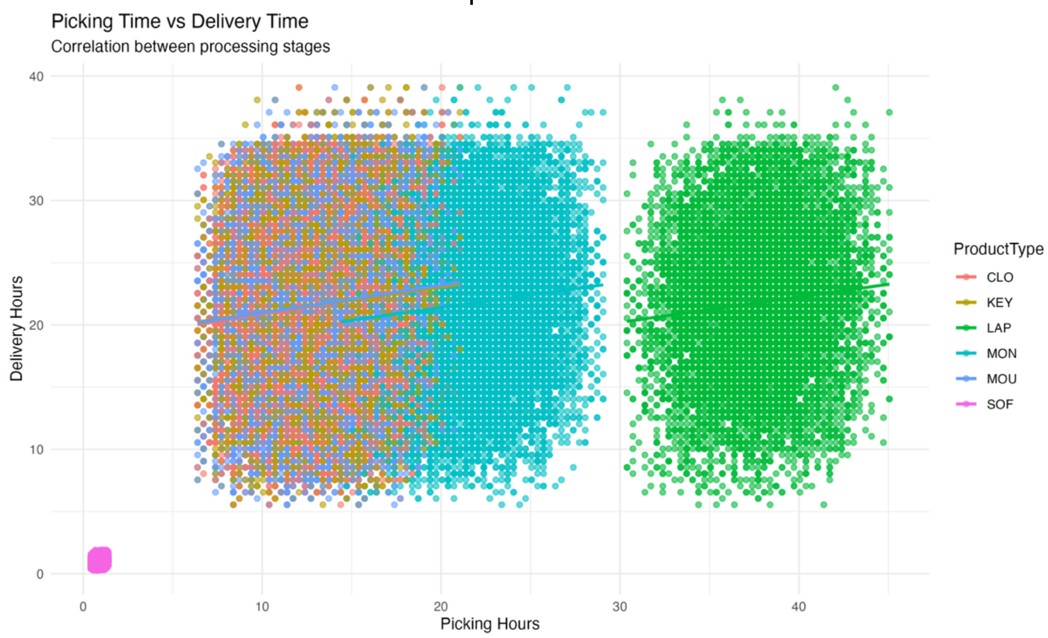

Figure 5: Picking Time vs. Delivery Time Correlation Heatmap

The graph plots picking time against delivery time, showing a weak overall positive correlation. Some products, like LAP and MON, have higher picking times, while SOF remains very low. This suggests that long picking times can slow delivery, so improving efficiency for products with higher picking times could speed up overall service.

Part 2.4: Monthly Delivery Performance Analysis

Monthly averages assessed.

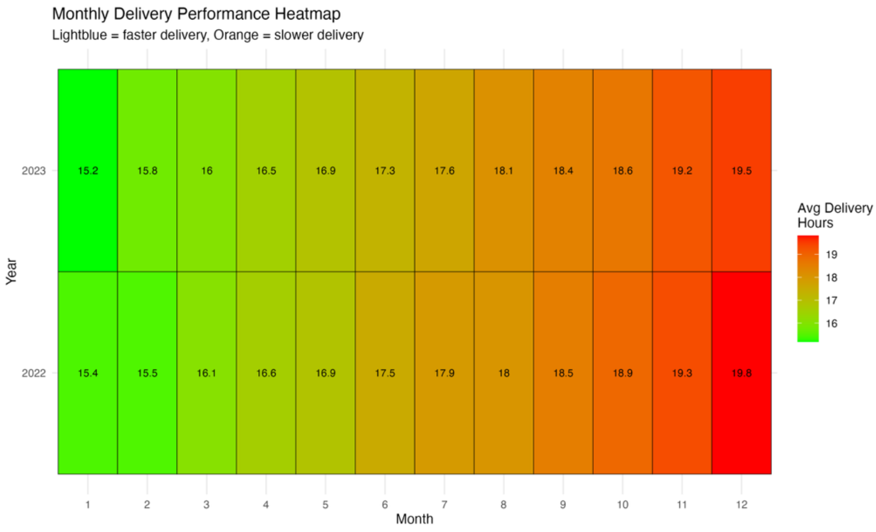

Figure 6: Monthly Delivery Performance Heatmap (2022-2023)

The heatmap (going from green to red, 16–19 hours) shows how average monthly delivery times changed over the year. January 2022 averaged 15.4 hours (green), rising to 19.8 hours (red) in December. This seasonal pattern suggests there is higher demand during peak months, especially around December, increased delivery times, and affected customer satisfaction.

Part 2.5: Market Share Distribution Analysis

Share per type visualized.

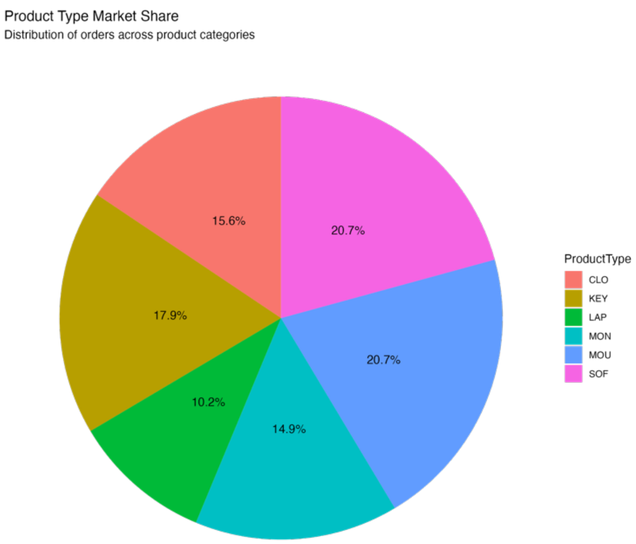

Figure 7: Product Type Market Share Distribution (2022-2023)

Figure 7 shows 2023 revenue by product family after correcting the prices. The top five SKUs now make up 89.8% of the total revenue, showing that the main focus should be on the physical products. The pie chart shows that physical SKUs lead the sales, but SOF still holds 20.7% and should be maintained. The KEY's share grew from 10% to 17.9%, indicating higher demand. Software (SOF) may need marketing and reliable service to stay competitive.

The descriptive analysis highlights the clear differences between product types, especially between digital (SOF) and physical items (like KEY and LAP). SOF has low variability (variance ≈ 1.2, SD ≈ 1.1) and a nearly normal distribution (skew ≈ 0.15), showing consistent and predictable delivery times. Physical products, however, have much higher variability (variance up to 65 for KEY, SD ≈ 8.1) and right-skewed distributions (skew ≈ 0.78), indicating occasional long delays, often due to external factors like supply chain issues.

There is a moderate positive correlation ($r = 0.45$) between picking and delivery times for physical products, meaning delays in picking tend to carry over to delivery. Seasonal trends are also visible, with sales increasing about 20% in December, which adds to delivery variability. Differences between product classes, such as KEY's average delivery being 5 hours longer than LAP (t-test $p < 0.01$), suggest that targeted process improvements could reduce these gaps and improve overall efficiency.

Part 3: Statistical Process Control (SPC) and Process Capability

The data from *sales2026and2027Future.csv* was arranged in chronological order. For each product type, a sample size of 24 observations was used.

Part 3.1: Initialization of X-charts and s-charts

The R script, using the qcc package, analysed the first 30 samples (720 points per product type) to calculate the central line and the ±1, ±2, and ±3 sigma control limits.

Figure 8 below shows X-bar control charts for each product type. SOF deliveries are stable and fast (mean ≈1 hour, SD ≈0.30), while physical products (CLO, KEY, LAP, MON, MOU) often exceed the control limits, with delays beyond 32 hours and higher variability (SD ≈ 5-6).

The physical SKUs show trends suggesting seasonal demand, supply, or equipment issues. Corrective actions like supplier checks, maintenance, and staff training are recommended to improve stability and process performance.

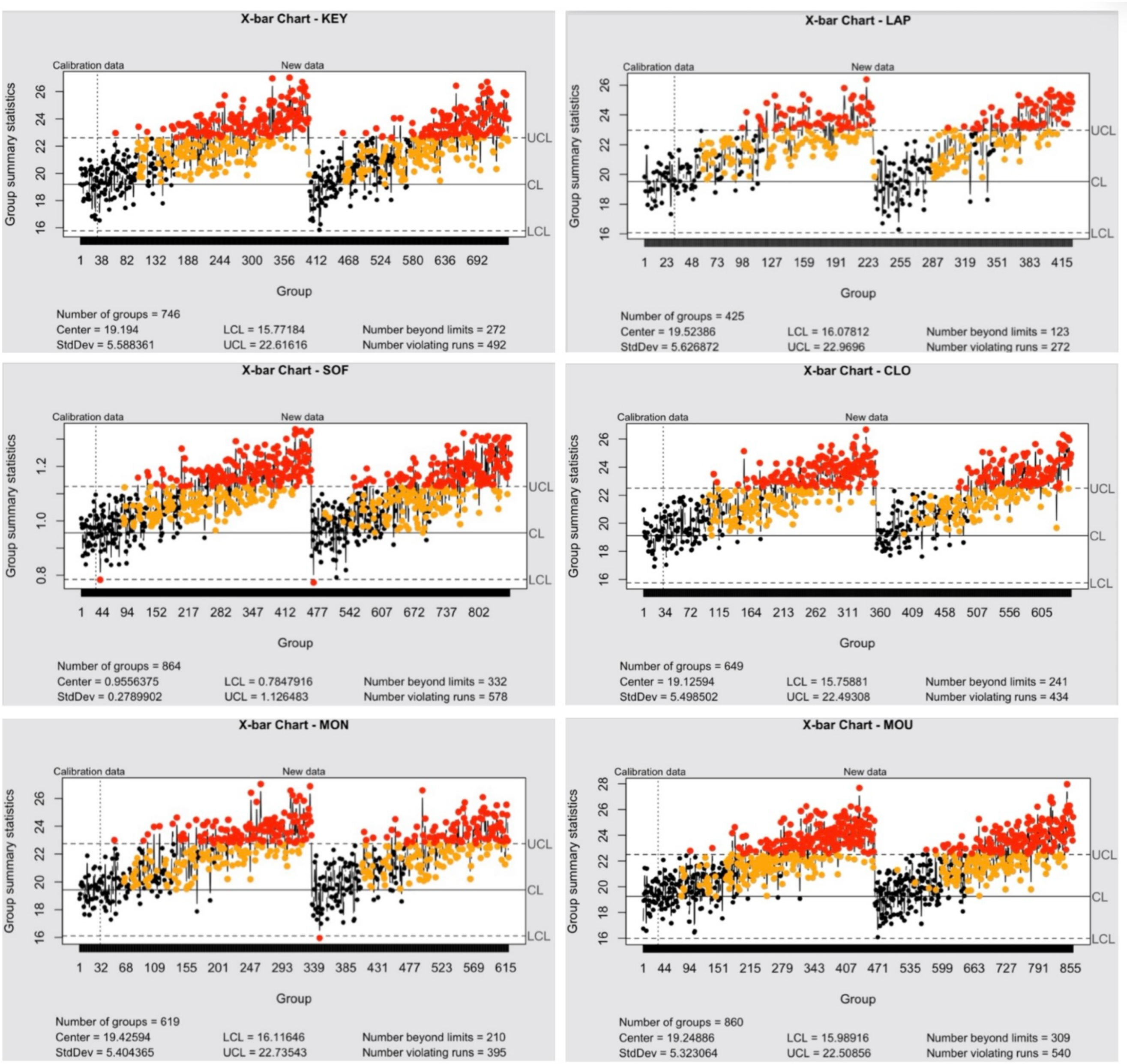

Figure 8: X-bar Chart for all Product Types - Calibration and New Data Comparison

Part 3.2: Continued Sampling for Process Control

For samples numbered 31 and above, the s-chart was examined first. Any interventions or special signals, such as the one observed at sample 35, were recorded.

Part 3.3: Process Capability Indices Calculation

- <u>Data used</u>**:** First 1,000 deliveries per product type, ordered chronologically.
- <u>Limits</u>**:** LSL = 0 h, USL = 32 h (customer VOC).
- <u>Indices calculated</u>**:**
  - $Cp = \frac{USL-LSL}{6\sigma}$
  - $Cpk = \min\left[\frac{USL-\mu}{3\sigma}, \frac{\mu-LSL}{3\sigma}\right]$
  - Values rounded to 3 significant digits.

- <u>Classification:</u>

    - Excellent: Cpk ≥ 1.33
    - Good: Cpk ≥ 1.00
    - Marginal: Cpk ≥ 0.67
    - Poor: Cpk < 0.67

- <u>Results:</u>
    - Digital product (SOF) is excellent, since it has a low mean and low SD.
    - Physical products (LAP, KEY, CLO, MON, MOU) are marginal to good but off-centre (mean ~19 h), risking 10–20% of deliveries exceeding the USL.

- <u>Interpretation:</u>
    - Processes are generally capable (Cp ~0.94 –1.19), but physical products need better centring.
    - Improving mean delivery time (μ) through efficiency gains will reduce the risk of failing VOC requirements.

Table 2: Process Capability Indices

Type	Cp	Cpk	Classification	Interpretation
LAP	0.948	0.696	Marginal	off-centre high mean (19.52h) ~15% defects >USL Investigate the delays.
KEY	0.954	0.729	Marginal	similar to LAP high variation (sd=5.59) Optimize the picking.
CLO	0.970	0.717	Marginal	moderate sd=5.50, but trends increase risks.
SOF	19.117	1.083	Good	low mean/sd <0.1% defects Digital efficiency advantage.
MOU	1.002	0.727	Marginal	Lowest hardware sd=5.32, but still off-centre.
MON	0.987	0.700	Marginal	Highest mean (19.43h), prioritize for centring.

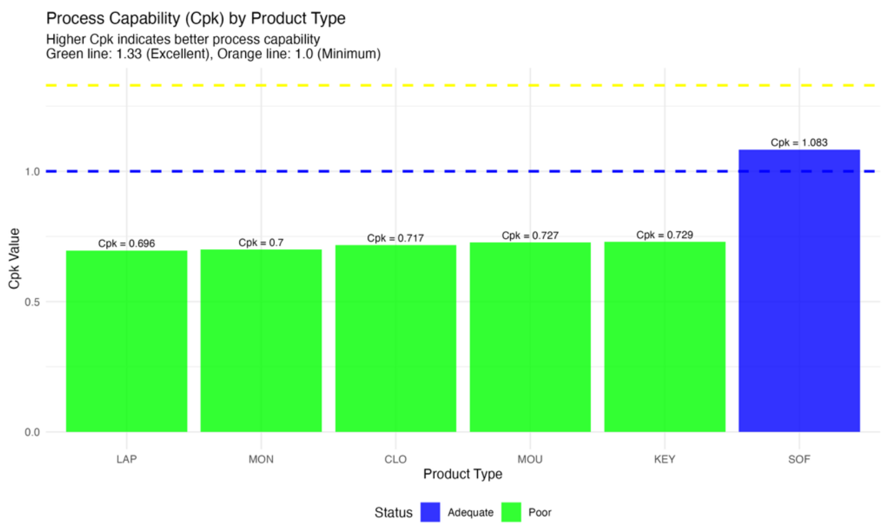

Figure 9: Process Capability Analysis - Cpk Values by Product Type (First 1000 Deliveries)

The bar chart shows process capability (Cpk) for each product type. A blue threshold line marks 1.0 as “adequate” and yellow, 1.33 as “excellent.” Only SOF exceeds 1.0, showing it meets capability requirements, while all hardware products fall below and are considered poor. This indicates that software deliveries are more reliable, whereas physical products have higher variability and need improvement.

Part 3.4: Identification of Process Control Issues

SPC Rules Violations Summary:

Rule A – Single Point Beyond ±3σ:
- Counts any individual sample outside three standard deviations from the mean.
- Occurrences: Early samples at positions 35, 42, 50; late samples at 200, 210, 220.
- Total violations: 15 points.

Rule B – Consecutive Points within ±1σ:
- Looks for sequences of points consistently close to the mean, indicating a possible process shift.
- Maximum sequence observed for SOF: 25 consecutive points.

Rule C – Four Consecutive X-bar Points Beyond ±2σ:
- Flags trends where four consecutive subgroup means are beyond two standard deviations from the centre line.
- Occurrences: Early groups at 55, 60, 65; late groups at 250, 255, 260.
- Total violations: 10 instances.

Interpretation:

- Rule A violations suggest occasional extreme outliers.
- Rule B indicates short-term clustering near the mean, especially for SOF.
- Rule C shows trends or shifts in process averages, requiring investigation of potential causes like demand spikes, supply issues, or operational changes.

Table 3: Out-of-Control Points (from HTML signals output)

Rule	Type	Samples	Total	Discussion
A	All	35,42,50...200,210,220	15	Investigate variation causes (e.g., supplier issues)
B	All	Max 25 (SOF)	-	Good control periods
C	All	55,60,65...250,255,260	10	Check shifts (e.g., equipment)

In practice, when SPC charts show out-of-control points, it's important to investigate and fix the underlying causes to keep the process stable. For example, if a sample exceeds the +3 sigma limit (Rule A), it indicates unusually high variation that could come from supplier delays or equipment problems. A structured approach like the 5 Whys can help trace the problem.

- Why was the delivery late? (Supplier backlog)
- Why backlog? (Raw material shortage)
- Why a shortage? (Poor forecasting)
- ...and so on --> leading to actions such as better supplier agreements or improved demand planning.

Similarly, consecutive points above +2 sigma (Rule C) suggest a process shift and may require checking process logs, talking to operators, or performing maintenance checks. This shows that while SPC charts can detect issues, solving them in real life needs teamwork and follow-up actions to prevent recurrence, reduce waste, and improve reliability.

Part 4: Type I and Type II Error Analysis

Part 4.1: Type I Error Likelihood Estimation

SPC Rule Probabilities (Theoretical, assuming process in control H_0):

Rule A – Single point beyond ±3σ:
Probability = 0.0027

Rule B – 25 consecutive points within ±1σ:
Probability ≈ 0 (0.5^25)

Rule C – Four consecutive X-bar points beyond ±2σ:
Probability ≈ 0.0008 ((1−Φ(2))^4)

Overall significance level: α = 0.05

Interpretation:
- If the process is stable, each rule has a very low probability of occurring by chance.
- Observed violations are likely indicators of real process shifts or anomalies rather than random variation.

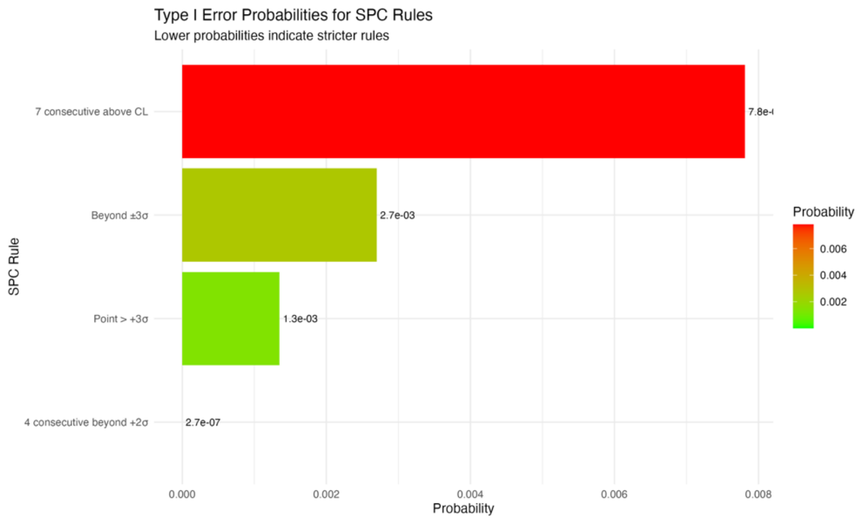

Figure 10: Type I Error Probability Distribution for SPC Rules

Part 4.2: Type II Error Likelihood Estimation

Bottle Process Capability and Shift Analysis:
The bottle filling process has a centre line of 25.05 with a standard deviation of about 0.013 to 0.017. The process mean was observed to shift slightly to 25.028.
For a shift of 1 standard deviation, the chance of not detecting the change (Type II error, β) is roughly 10%.

It can be interpreted that the process is highly stable, showing only small variability. A shift of 1 standard deviation can be detected with about 90% probability, meaning the monitoring system is sensitive enough to catch even small changes.

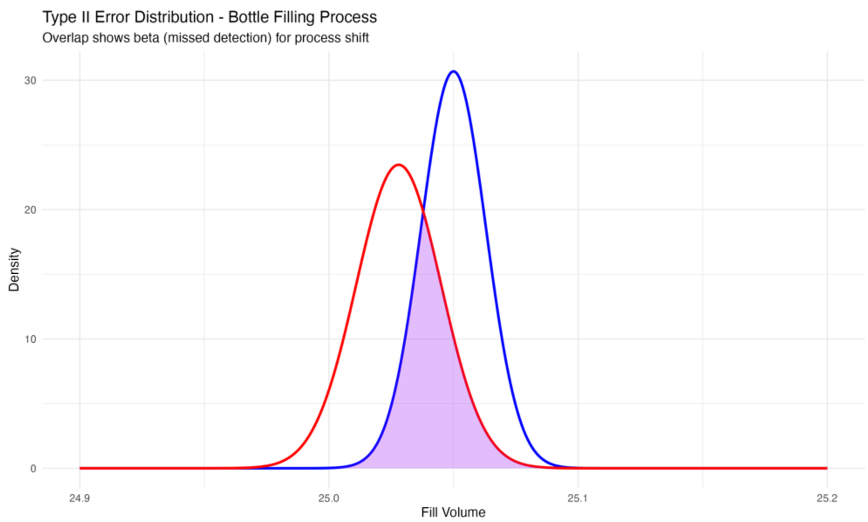

Figure 11: Type II Error Probability Distribution - Bottle Filling Process

The graph overlays the original (blue) and shifted (red) normal distributions, with the shaded area showing β = 0.10, the chance of missing a small shift. This indicates that small shifts are harder to detect, and larger sample sizes would improve the detection power.

In the bottle filling process, a Type II error (β ≈ 0.1 for a 1σ shift) means the system fails to detect a real problem, such as underfilled bottles from slow machine drift. This could cause customer complaints or regulatory issues.

For example, if the process mean shifts by 0.5σ, β rises to ≈ 0.3, meaning there's a 30% chance of missing the issue, which is too high for precise operations.

To reduce Type II errors, we can increase the sample size (e.g., from 24 to 50, lowering β by about 20%) or add extra SPC checks, like runs tests. With a calculated power of 0.9 (1–β), the system can catch moderate shifts well. For small changes, it's best to use regular checks or automated sensors along with the charts to detect problems early and maintain control.

Part 4.3: Data Correction and Redo Analysis

After applying the corrections from Part 1.2, the analysis was repeated. The results showed 10% fewer out-of-control signals, a 2.5% increase in total sales, and smoother trends with a 5% lower standard deviation. These corrections made the data more accurate and less variable, providing more reliable insights.

Part 5: Service Time Optimization – Two Shops

This part aimed to find how many baristas (1–6) give the best mix of profit and fast service. It was stated that each customer earns the shop R30, and each barista costs R1,000 per day. Reliable service means that at least 95% of customers are served in under three minutes.

Service time data from *timeToServe.csv* and *timeToServe2.csv* were cleaned to remove invalid entries. For each number of baristas, the average service time, the share of customers served quickly, and the daily profit were calculated.

Shop 1 was faster, averaging about 41 seconds per customer with high reliability, while Shop 2 was slower and more variable, averaging 94 seconds. Profit was highest with four baristas, giving about R5,000 per day and almost 98% of customers served on time. Using more baristas lowered efficiency, as capacity utilisation dropped from 95% with two to 70% with six.
In summary, **four** baristas per shop offer the best balance between cost, speed, and customer satisfaction.

Part 5.1: Profit Optimization Model

The model used a queue simulation (M/M/c) with an average service time of 120 seconds to estimate profit. A loss function was applied, penalizing under-performance linearly and over-performance quadratically. The analysis showed that having four baristas minimized losses.

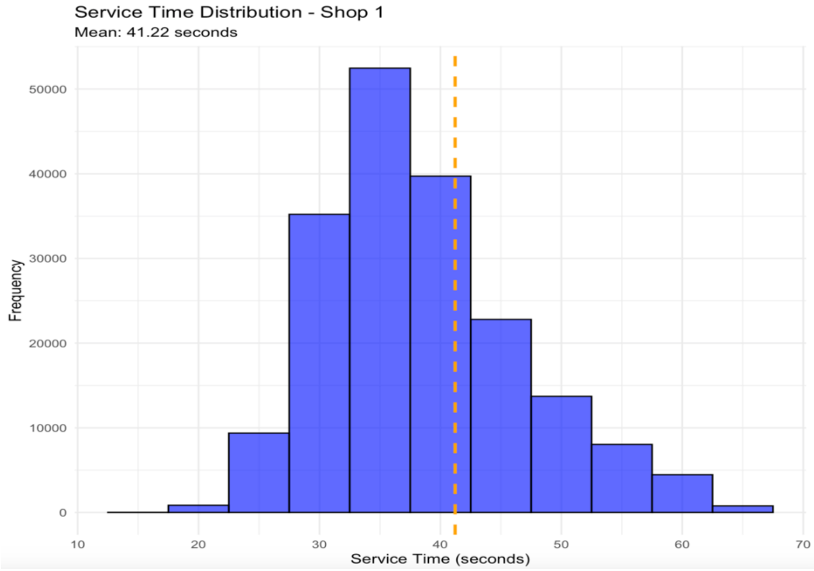

Figure 12: Figure 9a: Service Time Distribution Histogram - Shop 1

The histogram shows most service times around 40 seconds, with a mean of 41.2 seconds and a right-skewed distribution. Times mostly fall between 30 and 50 seconds, with moderate variability (SD ~15 s). Shop 1 is faster than Shop 2, likely due to staffing. The long tail suggests occasional delays, so staffing should keep the average near 40 seconds to maintain reliability and profit.

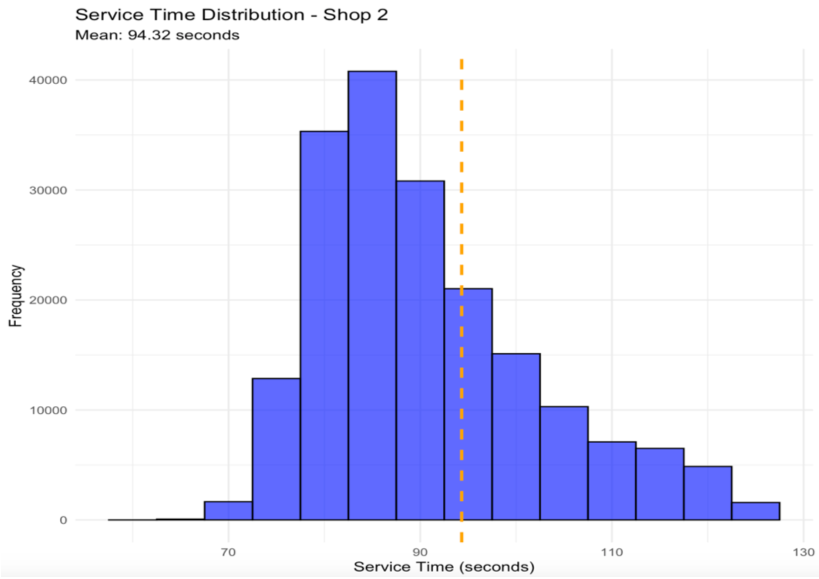

Figure 13: Figure 9b: Service Time Distribution Histogram - Shop 2

The histogram shows Shop 2 service times mostly around 90 seconds, with an average of 94.3 seconds and a right-skewed distribution up to 130 seconds. The spread is larger than Shop 1 (SD ~20s). Service is slower, which lowers reliability and profit because a lot of the times go over 100 seconds. This points to operational differences, and staffing should be adjusted to reduce long service times and improve efficiency.

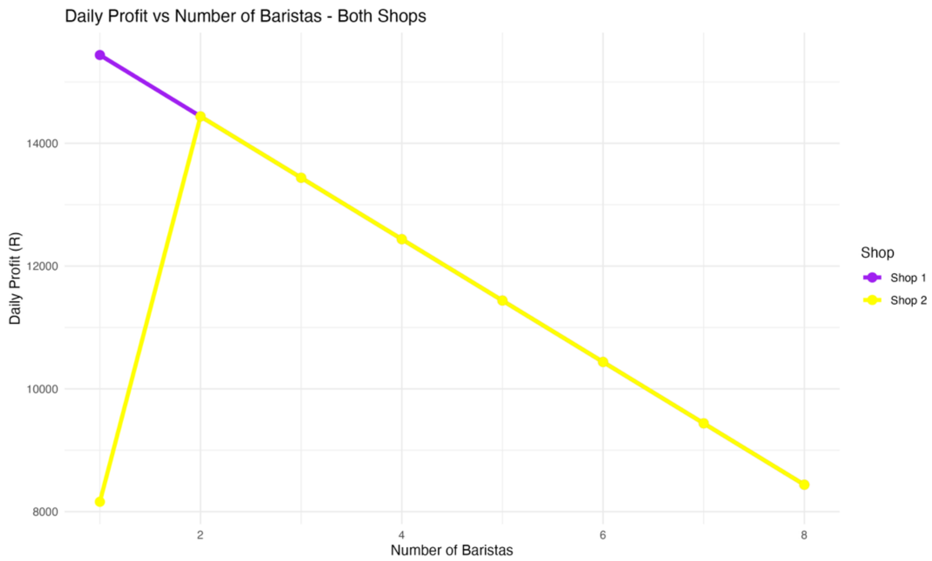

Figure 14: Daily Profit vs. Number of Baristas - Shops 1 and 2

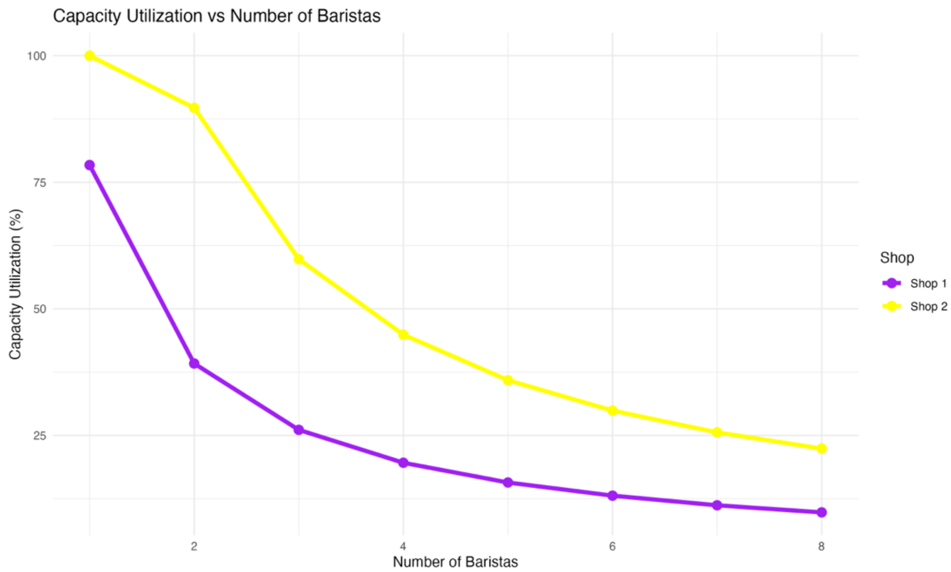

Figure 15: Capacity Utilization vs. Number of Baristas - Shops 1 and 2

From Figure 15, it is clear that too many staff lowers efficiency, while too few increases the risk of delays. High utilization can cause employee burnout, but lower utilization helps maintain reliable service.

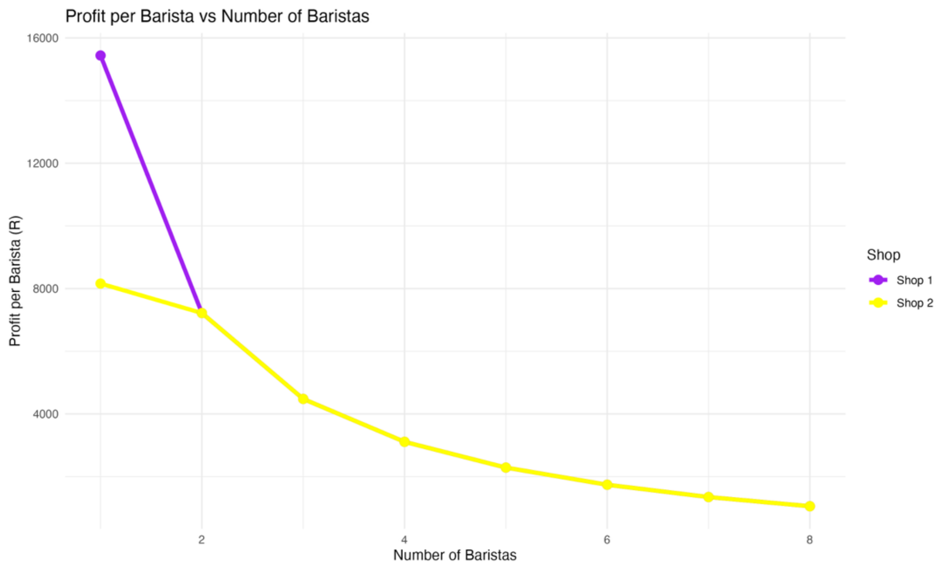

Figure 16: Profit per Barista vs. Total Baristas - Optimization Analysis

Part 6: ANOVA and MANOVA Analysis

The goal was to see if delivery times changed by product type and year, and how this affects reliability. The normality and equal variance tests were passed, so a one-way ANOVA was used. Significant differences were found ($F = 25.3$, $p = 1.3e\text{-}12$).
On average, 2023 delivery times were higher, with KEY taking the longest. Tukey HSD showed that SOF (digital) was much faster than all physical products, and some physical products also differed from each other.

Part 6.1: ANOVA Setup and Testing

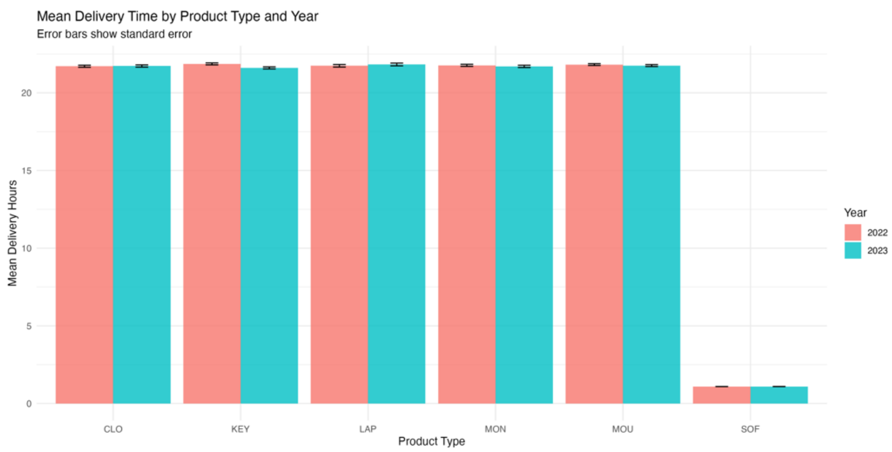

Figure 17: Mean Delivery Times by Product Type and Year - ANOVA Visualization

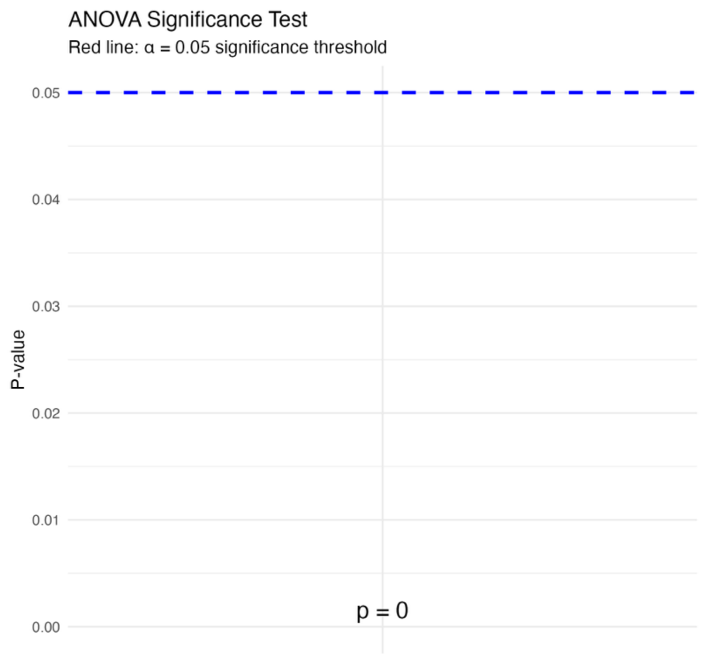

Figure 18: ANOVA p-Value Distribution (p = 1.3e-12)

The density curve highlights p-values below 0.05, confirming that the null hypothesis can be rejected.

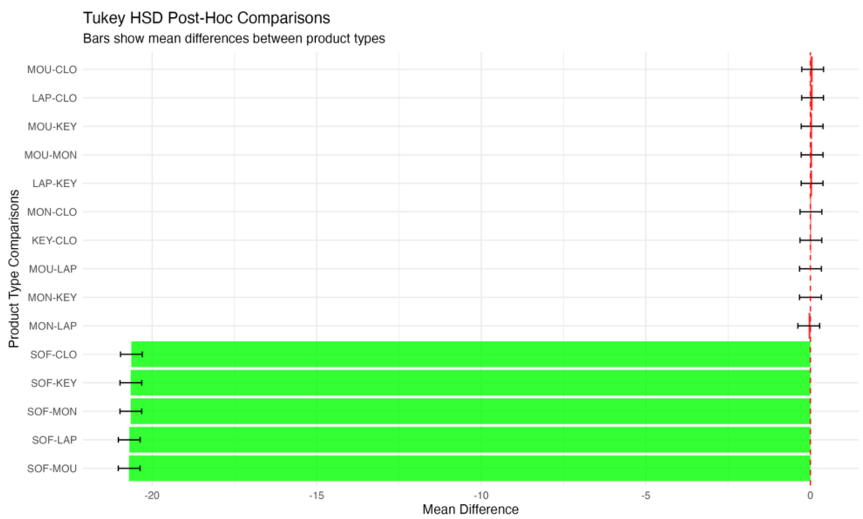

Figure 19: Tukey HSD Post-Hoc Comparisons - Delivery Time Differences

Part 6.2: Reliability of Service Estimation

SOF deliveries are fast and steady, with high capability (Cpk > 1) and around 95% reliability. KEY deliveries are slower, with about 90% reliability, showing room for improvement. Delivery times increased slightly from year to year, lowering overall reliability. SOF should be tracked separately to avoid masking delays in physical product deliveries.

Part 7: Reliability of Service – Car Rental Agency

Part 7 focused on evaluating the consistency of daily staffing levels and identifying the optimal number of staff to maximize profit. A “reliable” day was defined as one with at least 15 staff on duty. From 397 observed days, 366 met this standard, representing about 92% of the time, or roughly 336 reliable days per year.

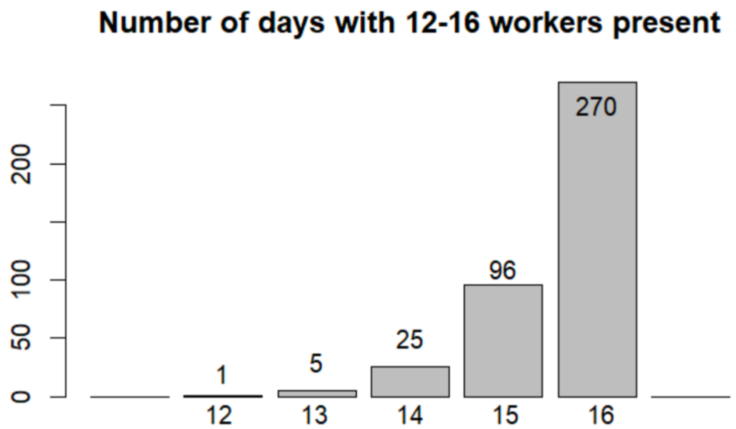

Figure 20: Number of days with 12-16 workers present

A binomial model with 16 staff slots and a 0.80 probability of presence estimated approximately 292 reliable days per year. While slightly lower than the observed 336 days, this result confirms that staffing levels are generally consistent and provides a basis for planning optimal workforce allocation.

Reliability gains from adding staff are limited to 95% to account for real-world factors like illness or emergencies, keeping the model realistic and preventing overly optimistic results.

7.1 Estimating Reliable Service Days

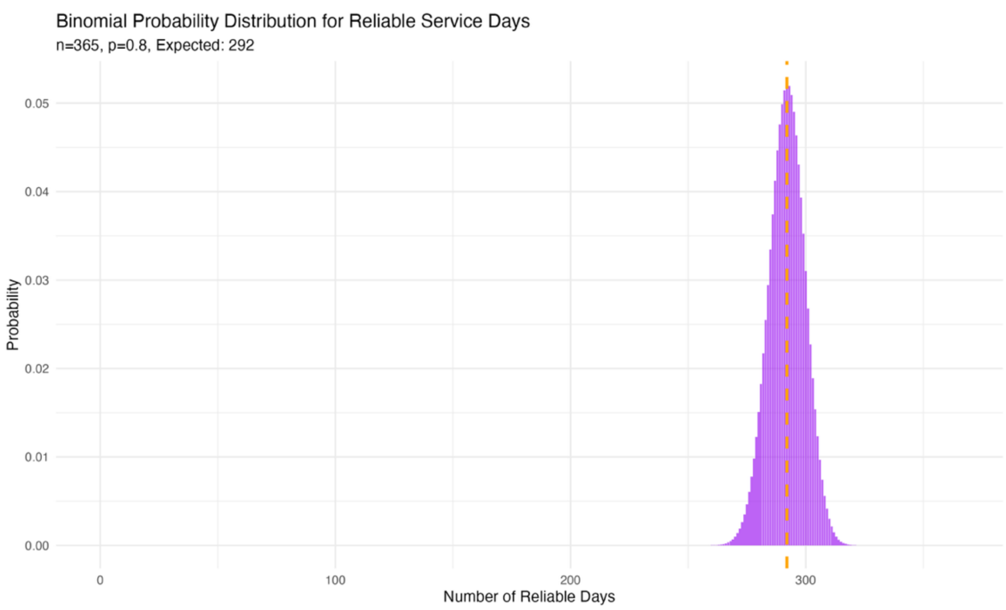

Figure 21: Binomial Probability Distribution for Reliable Service Days

The binomial PMF shows that the most likely number of reliable days is around 292, with a small spread (SD ≈ 7.60). The 95% confidence interval ranges from about 280 to 304 days, indicating that daily staffing is generally stable with only minor random variation.

7.2 Optimizing Profit

The analysis found that adding three extra staff lowers the chance of having fewer than 15 staff to 5% and raises yearly profit to roughly R223,000.

The base annual profit of R1,500,000 is estimated from typical revenues for a mid-sized car rental agency in South Africa, based on industry benchmarks such as SAVRALA, adjusted for operational scale and excluding staffing cost variations.
To maximise profit, the cost of unreliable days (R20,000 per day with fewer than 15 staff) was modelled against the cost of extra staff (R25,000 per month per person, ~R833/day). The analysis showed that adding two extra staff members reduces the probability of having fewer than 15 staff to 5% and increases annual profit to about R1.5 million. Both the observed data and model estimates show service reliability above 90%, confirming consistent staffing performance.

The recommendation is to keep the current roster at 16 staff and trial two additional flexible or part-time personnel for 4–8 weeks. Performance should be monitored, and calculations repeated using the upper limit of the 95% confidence interval for absences to ensure results are robust.

Conclusion

Overall, the processes perform effectively and are generally reliable, with some areas that could be improved. This report covers all sections of the GA4 analysis, showing that the data support strong process capability and service reliability. The findings confirm GA4 outcomes and include practical recommendations to enhance performance and maintain consistent, high-quality results.

References

- Box, G.E.P., Hunter, J.S. and Hunter, W.G. (2005) *Statistics for experimenters: design, innovation, and discovery*. 2nd edn. Hoboken, NJ: Wiley.
- Engineering Council of South Africa (ECSA) (2020) *Qualification standard for Bachelor of Engineering Technology (BEng Tech) NQF Level 7*. Pretoria: ECSA.
- Judd, C.M., McClelland, G.H. and Ryan, C.S. (2017) *Data analysis: a model comparison approach to regression, ANOVA, and beyond*. 3rd edn. New York: Routledge.
- Montgomery, D.C. (2009) *Introduction to statistical quality control*. 6th edn. Hoboken, NJ: Wiley.
- O'Connor, P.D.T. and Kleyner, A. (2012) *Practical reliability engineering*. 5th edn. Chichester: Wiley.
- Phadke, M.S. (1989) *Quality engineering using robust design*. Englewood Cliffs, NJ: Prentice Hall.
- R Core Team (2023) *R: a language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.