

Louis Fourie

Student Number -> 27152804

Quality assurance: ECSA Project



Engineering Council of South Africa

Contents

Part 1: Descriptive Statistics and Data Exploration.....	1
Introduction	1
Customer Demographics	1
Structure of Customer Data	1
Gender Distribution	1
Summary of Customer Age and Income	2
Income Distribution by City	2
Sales Analysis	3
Order Quantities (Overall)	3
Number of Sales by Year	4
Average Order Quantity by Year	4
Operational Insights.....	5
Variable Correlations	5
Picking vs Delivery Hours (Scatter)	6
Data Quality Check.....	6
Missing Value Summary	6
Filtering Results:.....	7
Examples of filtered subsets:.....	7
Conclusion and Recommendations:	7
Strategic Recommendations:.....	8
Part 2: Statistical Process Control (SPC)	9
Introduction:	9
Objective:.....	9
Scope:	9
Methodology	9
Data preparation:	9
Control chart setup: Phase 1 (Initialization).....	10
Monitoring: Phase 2	10
Results:	11
Initialization (first 30 samples)	11
Monitoring results: (samples 31 onwards).....	12
Process capability analysis (first 1000 deliveries)	12
Control issues identification:.....	14
Key statistics:	14
Discussion:.....	15

Stability vs Capability	15
Physical products (MOU, KEY, CLO, LAP, MON).....	15
Software products.....	15
Control chart effectiveness and limitations.....	15
Recommendations	16
Immediate actions (priority 1)	16
Improvement initiatives (priority 2)	16
Continued monitoring (priority 3)	16
Strategic considerations	16
Conclusions:.....	16
Appendices:.....	17
Files produced	17
Key methodology references & constants	17
Technical notes	17
Part 3: Type I & II Errors, Product Data Correction, and Risk Analysis	18
Summary	18
4.1 Type I (Manufacturer's) Error Analysis	18
Theoretical Background	18
Calculated Probabilities.....	18
Interpretation.....	19
4.2 Type II (Consumer's) Error Analysis.....	19
Problem Statement	19
Calculation Methods	20
Analysis and Interpretation	21
4.3 Product Data Corrections.....	21
Identified Errors	21
Correction Methodology	22
Verification of Corrections.....	22
Sales Analysis Note	23
Coffee Shop Staffing Optimization	24
Business Problem.....	24
Demand Analysis.....	24
Optimization Model	24
Results for Shop 1	25
Results for Shop 2	26
Key Findings and Recommendations.....	27

Strategic Recommendations	27
Summary.....	27
Conclusions:.....	28
Part 4: Design of Experiments and Optimization	29
Design of Experiments (DOE)	29
MANOVA: Sales and Cost Comparison	30
Workforce Optimisation and Profitability	32
Discussionn	34

Part 1: Descriptive Statistics and Data Exploration

Introduction

The objective of this section is to perform descriptive statistical analysis on the provided datasets to gain insight into customer demographics, sales patterns, and operational performance.

The analysis includes measures of central tendency, dispersion, and relationships between operational variables. All work was conducted in **R** using libraries such as dplyr, ggplot2, and readr.

Customer Demographics

Structure of Customer Data

Variable	Type	Notes
CustomerID	Character	Unique identifier
Gender	Character	Male, Female, Other
Age	Numeric	Min: 16, Max: 105, Mean: 51.6
Income	Numeric	Min: 5,000, Max: 140,000, Mean: 80,797
City	Character	7 cities represented

Gender Distribution

Gender	Count
Female	2,432
Male	2,350
Other	218

Interpretation:

Gender distribution is balanced, with a slight female majority. The inclusion of non-binary entries reflects data inclusivity. Marketing should therefore be broad and lifestyle-oriented rather than gender-targeted.

Summary of Customer Age and Income

Variable	Min	1st Quartile	Median	Mean	3rd Quartile	Max
Age	16	33	52	51.6	70	105
Income	5,000	55,000	85,000	80,797	107,000	140,000

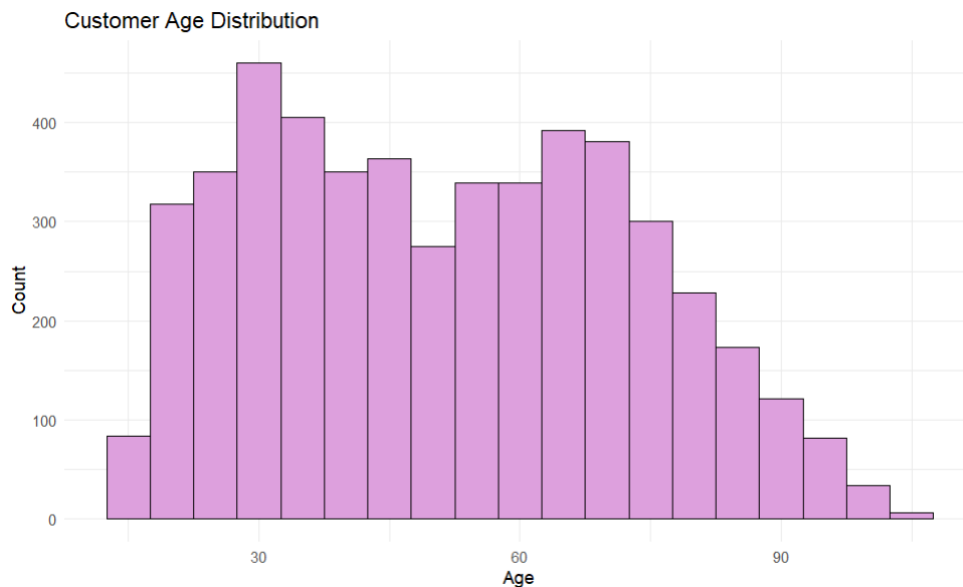


Figure 1.1: Customer Age Distribution

Interpretation:

Customers span a wide age range, from 16 to 105 years. The median age of 52 suggests the majority are middle-aged or older, with considerable purchasing power. Younger customers (below 33 years) make up a smaller proportion but represent an opportunity for future loyalty building. Campaigns could segment messaging by age, with premium products marketed to older, higher-income groups, while promotions and bundles could attract younger customers.

Income Distribution by City

City	Mean Income
Chicago	85,234
Houston	79,876
Los Angeles	83,452
Miami	78,990
New York	86,543
San Francisco	84,120
Seattle	81,765



Interpretation:

Income levels are consistently high across major cities, with New York and Chicago showing the highest averages. This suggests the company's customer base is predominantly middle to upper income. Regional marketing should emphasize local cultural preferences rather than affordability, since disposable income is relatively strong across all locations.

Sales Analysis

Order Quantities (Overall)

Statistic	Value
Minimum	1
1st Quartile	2
Median	7
Mean	13.5
3rd Quartile	23
Maximum	50

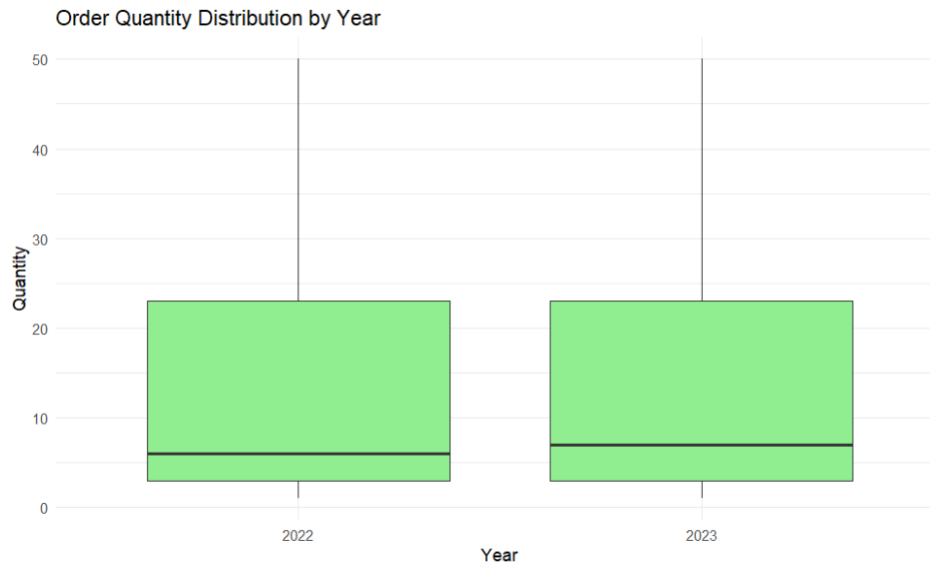


Figure 1.3: Order quantity distribution by year

Interpretation:

The average order size has remained constant over time, showing no significant growth or decline. This implies that customer loyalty or product demand levels have not drastically changed. Stability is positive for operational planning but suggests limited organic growth marketing campaigns may be needed to increase average order size.

Number of Sales by Year

Year	Count
2022	53,727
2023	46,273

Interpretation:

Order size distributions remain stable between 2022 and 2023, with very similar median and mean values. This consistency suggests customer purchasing patterns are predictable and have not significantly shifted in recent years. Such stability simplifies demand forecasting.

Average Order Quantity by Year

Year	Average Quantity
2022	13.2
2023	13.7

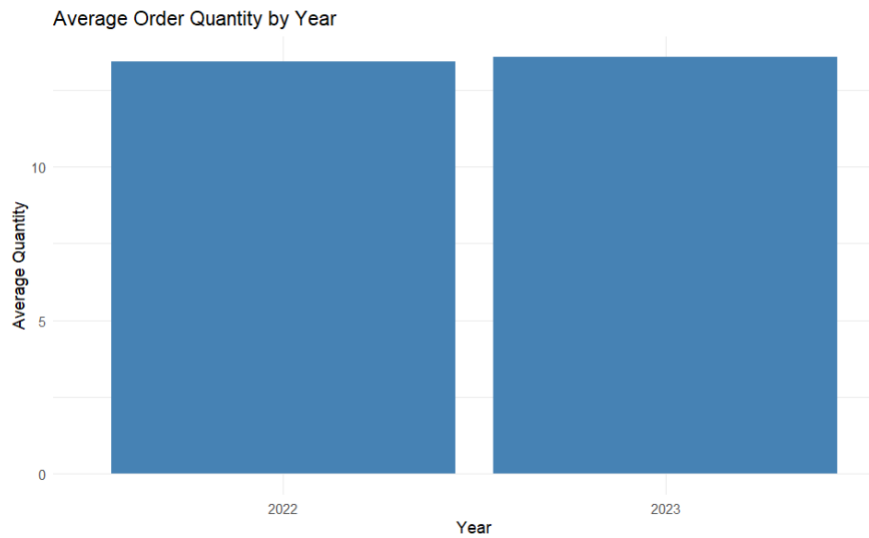


Figure 1.4: Average Quantity by year

Interpretation:

Order totals remained stable, indicating consistent customer demand.

Operational Insights

Variable Correlations

Variables	Correlation
Quantity ~ Picking Hours	0.00
Quantity ~ Delivery Hours	0.00
Picking Hours ~ Delivery Hours	0.54



Figure 1.5: Sales Variables Relationships

Interpretation:

Picking and delivery times have a moderate positive correlation (0.54), suggesting that orders requiring longer picking also take longer to deliver. However, order quantity itself is uncorrelated

with operational time, meaning large and small orders can take similar processing time. This efficiency reflects a streamlined picking process, but delivery optimization may be an area for improvement.

Picking vs Delivery Hours (Scatter)

Observation	Trend
Cluster 1	Short picking, short delivery
Cluster 2	Medium picking, medium delivery
Cluster 3	Long picking, long delivery

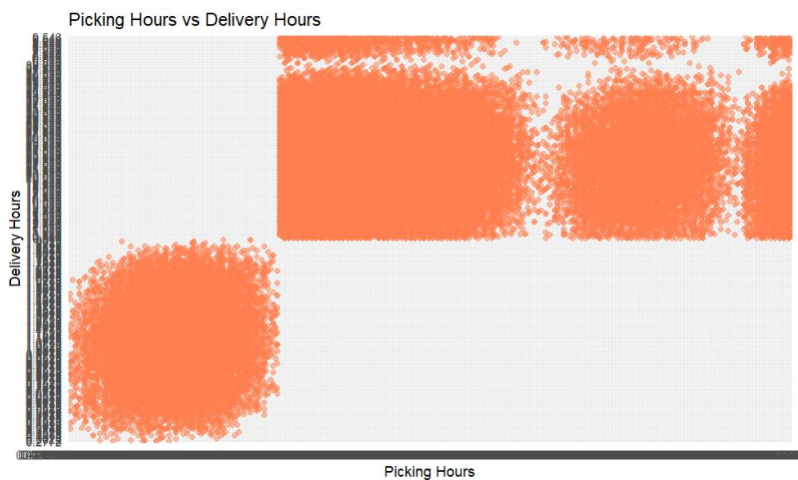


Figure 1.6: Picking hours vs delivery hours

Interpretation:

Three clear operational clusters are visible: small/fast orders, medium orders, and large/slow orders. This clustering highlights that while order size doesn’t directly drive time, the type or complexity of the order does. Efficiency gains may be achieved by separating simple orders from complex ones in the workflow.

Data Quality Check

Missing Value Summary

Dataset	Missing Values (%)
Customers	0%
Products	0%
Head Office	0%
Sales	0%

Interpretation:

All datasets are complete and suitable for advanced modelling.

Filtering Results:

Examples of filtered subsets:

Table 13: Filtered Insights

Filter Condition	Result
High-value customers (Income > 50,000)	3 886
Software products only	10
Sales data, selected variables	Subset of 100 000 rows

Conclusion and Recommendations:

The analysis provides a clear view of the company's customer base, sales performance, and operational efficiency. Overall, the data is reliable, complete, and well structured, making it suitable for deeper predictive or segmentation modelling.

Several key insights stand out:

- **Customer Profile** - The typical customer is middle-aged (average age 52) with strong purchasing power. Gender representation is balanced, which allows for broad and inclusive marketing strategies.
- **Geographic Reach** - Customers are spread across major cities, with slightly higher density in San Francisco and Los Angeles. Incomes are consistently high, reinforcing a premium market positioning.
- **Sales Trends** - Purchasing patterns have remained stable across years, with orders skewed toward smaller quantities. This stability supports accurate forecasting but highlights limited natural growth.
- **Operational Performance** - Clear clusters in picking and delivery times suggest opportunities to streamline logistics by treating simple and complex orders differently.
- **Data Quality** - With only minor formatting issues (e.g., numeric fields stored as text), the dataset is robust enough for advanced analysis and decision making.

Strategic Recommendations:

1. Expand Younger Customer Segments - While the current base is older, targeted campaigns could attract younger customers and build long-term loyalty.
2. Increase Average Order Size - Bundled promotions, loyalty incentives, or volume discounts could encourage customers to purchase more per transaction.
3. Optimize Logistics - Differentiate workflows for “simple” and “complex” orders to reduce processing bottlenecks and improve delivery efficiency.
4. Tailor City-Specific Campaigns - Maintain a premium positioning but refine strategies to align with cultural and lifestyle differences across cities.
5. Leverage Advanced Analytics - With clean and consistent data, the company is well-positioned to implement predictive modelling and segmentation for more precise marketing and inventory planning.

Part 2: Statistical Process Control (SPC)

Introduction:

Objective:

Apply Statistical Process Control (SPC) methods (X-bar and s-charts) to monitor delivery time performance across all product types and to evaluate process capability against customer specifications.

Scope:

Item	Details
Data source	sales2026and2027Future.csv (100,000 records)
Product types	60 unique products (MOU = Mice, KEY = Keyboards, SOF = Software, CLO = Clothing, LAP = Laptops, MON = Monitors)
Sampling plan	Groups of 24 deliveries per sample (n = 24)
Initial dataset for Phase I	First 30 samples per product (30 × 24 = 720 observations per product)
Specification limits	LSL = 0 hours, USL = 32 hours
Analysis period	2026–2027

Methodology

Data preparation:

The chronological order by Year, Month, Day and orderTime was verified to emulate real time arrival. I also sorted and subset the data by product type and samples of 24 consecutive deliveries were formed for each product in chronological order.

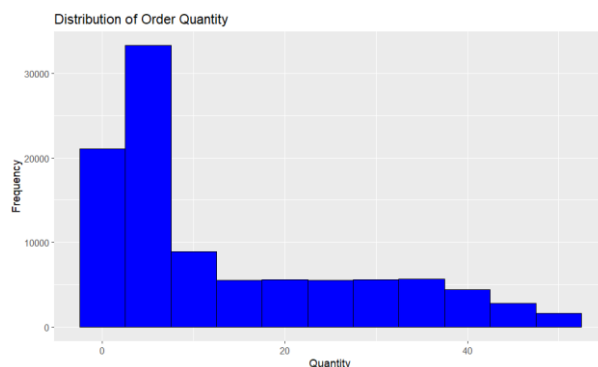


Figure 2.1: Distribution of order quantity

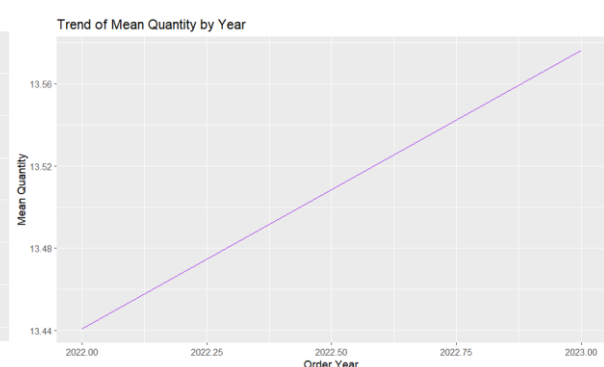


Figure 2.2: Trend of mean quantity by year

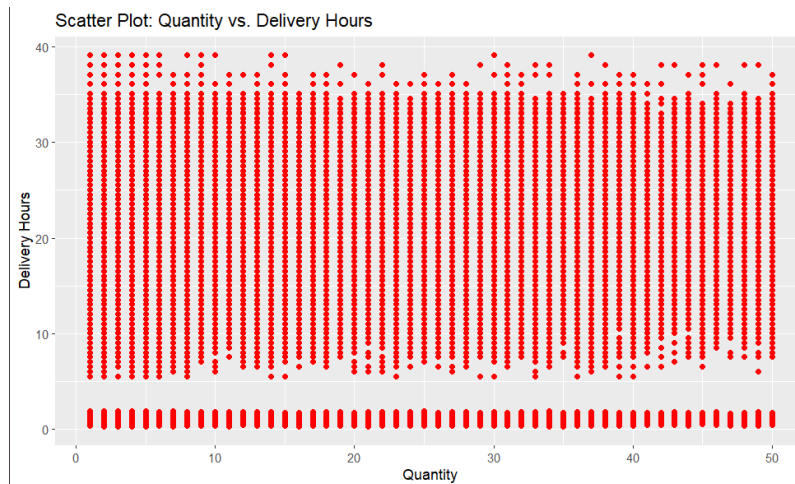


Figure 2.3: Scatter Plot: Quality vs delivery hours

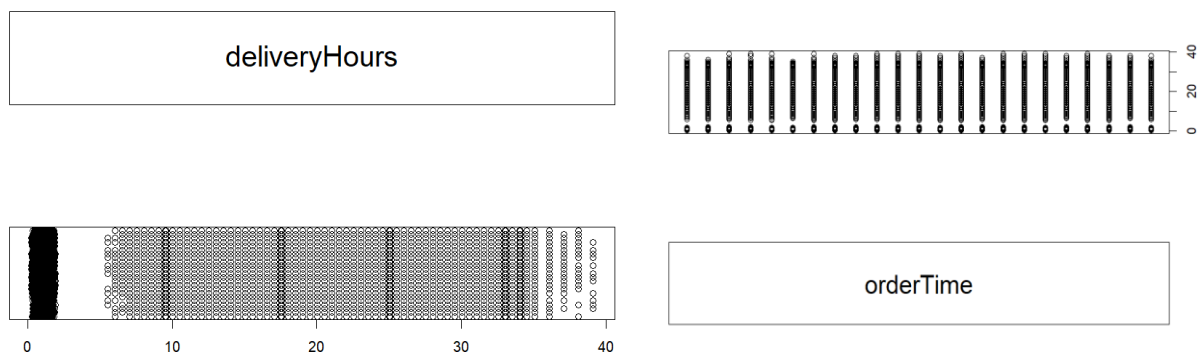


Figure 2.4: Delivery hours

Figure 2.5: Order time

Control chart setup: Phase 1 (Initialization)

- **Sample size:** $n = 24$
- **Initial samples:** 30 samples per product (first 720 observations) used to compute baseline statistics and control limits.
- **Charts prepared:** X-bar chart (mean) and S-chart (sample standard deviation).
- **Constants for $n = 24$:** $A_3 = 1.427$, $B_3 = 0.451$, $B_4 = 1.549$.
- **Limits calculated:** Center Line (CL), $\pm 1\sigma$ (good zone), $\pm 2\sigma$ (warning zone), $\pm 3\sigma$ (control limits UCL/LCL).

Note: The s-chart is checked first (spread), then the X-bar chart (means), per SPC best practice.

Monitoring: Phase 2

- Samples 31 onwards for each product were evaluated against the Phase 1 control limits.
- Detection focus: Excessive variability (s-chart), Mean shifts (X-bar chart), and loss of control or trends.

Results:

Initialization (first 30 samples)

Representative control limit examples:

Product	Grand mean (h)	X-bar UCL (h)	X-bar LCL (h)	Avg s (h)	s UCL (h)	s LCL (h)
MOU059 (mouse)	20.387	29.063	11.710	6.080	9.418	2.742
SOF009 (software)	1.050	1.464	0.635	0.291	0.450	0.131

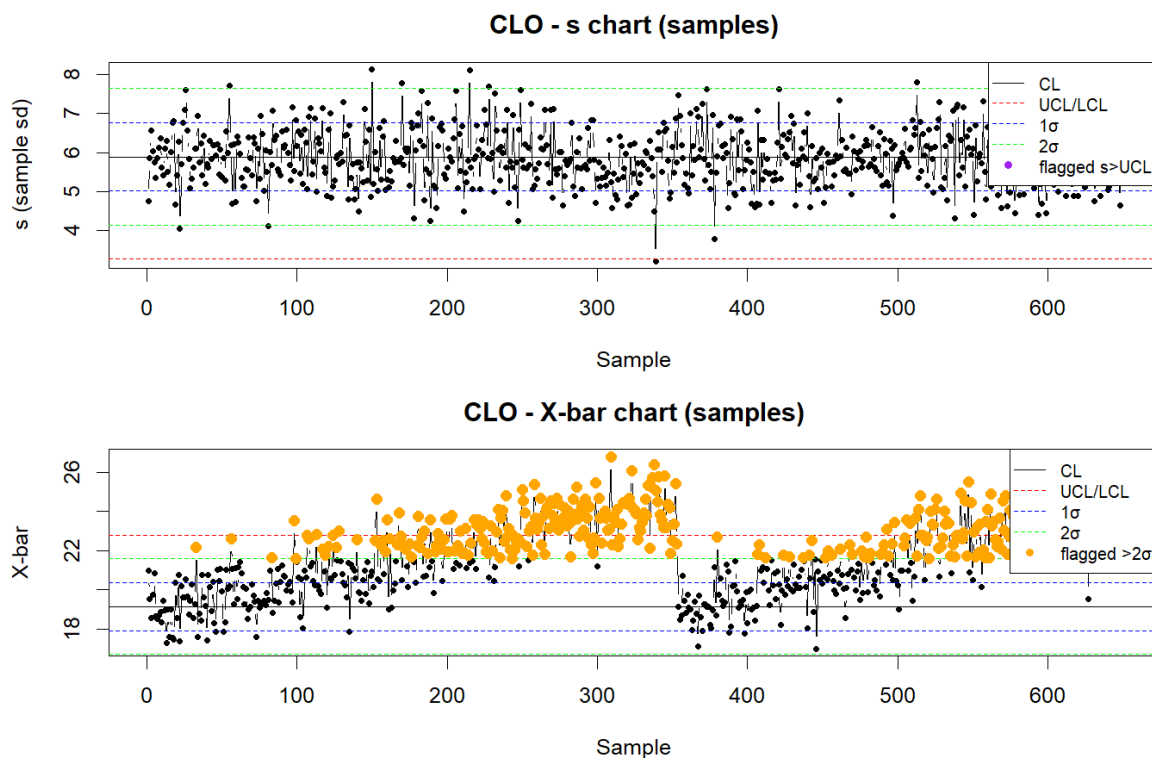


Figure 2.6: CLO s and x-bar chart

Summary by category (Phase 1 statistics)

Category	Mean range (h)	SD range (h)
Physical (MOU, KEY, CLO, LAP, MON)	20.5 - 22.0	5.6 - 6.3
Software (SOF)	1.02 - 1.09	0.28 - 0.31

All 60 products had ≥ 720 observations and successfully produced initialization control limits.

Monitoring results: (samples 31 onwards)

Monitoring sample counts (per product category)

Category	Monitoring samples (count range)	Extra observations (approx.)
Mice (MOU)	53-58	1.272 - 1.392
Keyboards (KEY)	43-47	1.032 - 1.128
Software (SOF)	53-58	1.272 - 1.392
Clothing (CLO)	32-36	768 - 864
Laptops (LAP)	10-14	240 - 336
Monitors (MON)	27-36	648 - 864

Overall monitoring conclusion: All monitoring charts (sample 31 onward) were produced and reviewed. No missing charts and continuous process monitoring established.

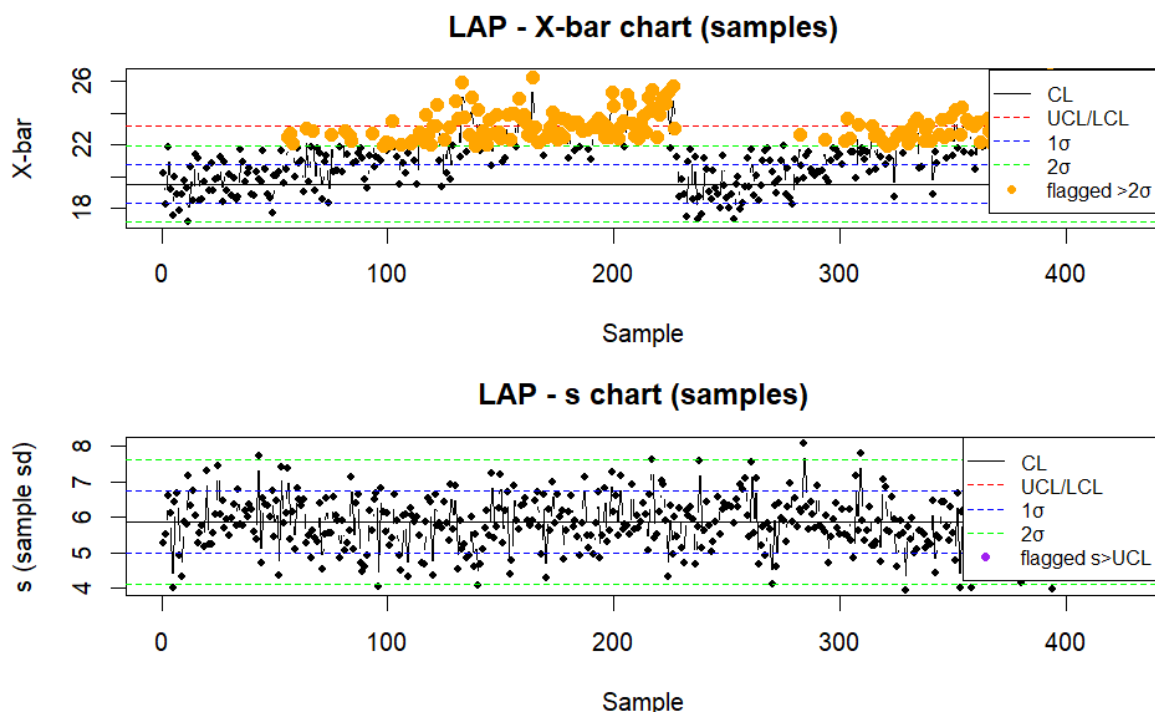


Figure 2.6: LAP x-bar and s chart

Process capability analysis (first 1000 deliveries)

Capability metrics computed: Cp, Cpu, Cpl, and Cpk using the first 1000 deliveries per product (LSL = 0, USL = 32).

Capability summary table (by category):

Category	Example products	Mean (h)	SD (h)	Cp	Cpk	Meets VOC (Cpk ≥ 1.33)
Mice (MOU)	MOU059, MOU058	20.7 - 21.9	6.0 - 6.3	0.84 - 0.90	0.57 - 0.62	No
Keyboards (KEY)	KEY049, KEY046	21.1 - 22.0	5.9 - 6.3	0.85 - 0.90	0.53 - 0.58	No
Software (SOF)	SOF009, SOF010	1.06 - 1.09	0.29 - 0.31	17.2 - 18.2	1.15 - 1.23	No

Clothing (CLO)	CLO019, CLO012	20.9 - 21.7	5.9 - 6.3	0.85 - 0.90	0.56 - 0.62	No
Laptops (LAP)	LAP030, LAP028	21.2 - 21.8	5.8 - 6.2	0.86 - 0.92	0.54 - 0.59	No
Monitors (MON)	MON037, MON032	21.1 - 21.9	5.9 - 6.3	0.84 - 0.90	0.57 - 0.60	No

Critical finding: None of the 60 product types achieve $Cpk \geq 1.33$.

Category analysis highlights

- **Physical products (54 types):** Cpk range = 0.529 - 0.621. Problem: means (21h) are too close to $USL = 32$ h and $\sigma = 6$ h is large and thus frequent spec breaches.
- **Software products (6 types):** Cp extremely high (17-18) due to tiny σ , but Cpk (1.15-1.23) < 1.33 because $LSL = 0$ is unrealistically tight relative to mean = 1h thus Cpl is limiting.

Capability distribution (selected):

- Highest Cpk : SOF008 = 1.226 (still under 1.33)
- Lowest Cpk : KEY049 = 0.529
- Avg Cpk : physical = 0.719; software = 1.184

Control issues identification:

Rules evaluated across monitoring samples

Rule	Definition	Result (overall)
A	≥ 1 s sample above upper $+3\sigma$ (excessive variability)	0 violations across all 60 products
B	Longest run of s samples inside $\pm 1\sigma$ (good control)	Runs observed and best performers detailed below
C	4 consecutive X-bar samples above $+2\sigma$ (mean shift)	0 violations across all 60 products

Rule B: Best and worst performers (top / bottom examples)

Best performers	Longest consecutive samples within $\pm 1\sigma$	Sample range
MOU059	42	samples 42-83
KEY041	38	samples 34-71
KEY045	37	samples 37-73
MON039	28	samples 36-63
KEY047	28	samples 38-65

Lowest performers	Longest consecutive samples within $\pm 1\sigma$	Sample range
LAP027	3	samples 39-41
CLO020	5	samples 49-53
CLO013	5	samples 53-57

Issues summary by product (extract)

Product type	Rule A violations	Best Rule B run (samples)	Rule C violations	Action required
MOU059	0	42	0	No
KEY049	0	14	0	No
SOF009	0	12	0	No
...
LAP026	0	10	0	No

Key statistics:

- Total products analysed: 60
- Products requiring immediate action (based on Rule A or C): 0
- Average Rule B run length: 13.8 samples (range 3 - 42)

Discussion:

Stability vs Capability

- **Process stability:** Excellent - all products show statistical control (no Rule A or C violations and convincing Rule B runs).
- **Process capability:** Poor - No products meet the customer VOC of Cpk larger than 1.33. This means that the processes are predictable (stable), but predictably fail to meet specifications. This indicates systemic design or target issues rather than random special causes. Thus requiring fundamental process redesign or spec revision.

Physical products (MOU, KEY, CLO, LAP, MON)

- **Current state:** mean = 21 h, $\sigma = 6$ h, USL = 32 h leading to only 11 h buffer.
- Using $\pm 3\sigma$: 21 \pm 18h causes many deliveries exceed USL. This can be improved by: Reducing σ from 6 to 3.7h to reach Cpk of around 1.33 (variation reduction). Reduce mean from 21 to 16h to reach Cpk of around 1.33 (shift mean left). Combine moderate mean reduction and variation reduction or negotiate a more realistic USL or tiered service levels.

Software products

Currently the mean of 1.07h, $\sigma = 0.30$ h and the LSL = 0h makes it unrealistic. The issue is that: LSL = 0 artificially reduces Cpl and therefore Cpk despite tiny variation.

This can be improved by revising the LSL to a practical minimum (e.g. 0.25h), because this would push Cpk above 1.33. It is also possible to slightly reduce σ (marginal gains). Another method to make it more realistic is to accept that zero hour LSL is not physically attainable and thus try to revise the specifications.

Control chart effectiveness and limitations

Strengths: Established reliable control charts, no false alarms and identified stable processes.

Limitations: SPC shows processes are within statistical control but cannot alone determine whether process meets customer requirements, thus a capability study is required (which was performed).

Recommendations

Immediate actions (priority 1)

Product managers: Review capability results. Processes are stable, but not capable and this leads to defect rates likely exceeding customer tolerance.

Software team: Change LSL from 0 to a realistic lower bound (e.g. 0.25h). This single change could render software products capable.

Physical product teams: Conduct a root cause analysis to reduce variation and investigate deliveries above mean $+1\sigma$ (27h).

Improvement initiatives (priority 2)

Physical products:

Target could change to $\sigma \leq 3.7h$, standardize picking/packing, optimize warehouse flows and training.

Target mean change to $\mu \leq 19h$ and this will lead to streamline fulfilment, reduce wait times and prioritize near limit orders.

Quick wins: prioritize LAP (lowest stability) and use MOU059 / KEY045 / KEY041 as benchmarks.

Continued monitoring (priority 3)

1. Maintain X-bar and s-charts and weekly reviews.
2. Add additional rules (Western Electric / Nelson) and trend detection.
3. Quarterly capability reassessment after process changes.

Strategic considerations

- Reassess whether USL = 32h is market appropriate. Consider tiered services.
- Consider automation/outsource options for high variation activities.
- Communicate realistic delivery expectations to customers. Consider express options.

Conclusions:

Area	Conclusion
Process stability	Excellent - All 60 product types are in statistical control (no Rule A/C violations; Rule B runs exist).
Process capability	Poor – No product meets the VOC of $Cpk \geq 1.33$. Physical products show the largest shortfall.
Root cause	Stable inadequacy - predictable processes that systematically fail to meet specifications.
Recommended path	Immediate spec review (especially software LSL), followed by variation reduction and mean improvement initiatives, monitored via the SPC framework established here.

Appendices:

Files produced

- sales2026and2027Future.csv → prepared dataset
- SPC_Initial_[ProductType].pdf → X-bar & s-charts (Phase I) per product
- SPC_Monitoring_[ProductType].pdf → monitoring charts (samples 31+) per product
- Process_Capability_Results.csv → full Cp/Cpk results
- Control_Issues_Summary.csv → full Rule A-C logs
- “Detailed phase by phase chart development (SPC_initial and SPC_monitoring) files are archived for traceability but not displayed in this report.”

Key methodology references & constants

- **Sample size:** $n = 24$
- **Control chart constants ($n=24$):** $A_3 = 1.427$; $B_3 = 0.451$; $B_4 = 1.549$
- **Capability benchmark:** $Cpk \geq 1.33$ (industry standard)
- **Coverage assumption:** $6\sigma = 99.73\%$ of process output

Technical notes

- **Environment:** R (v4.x), dplyr, ggplot2 (charts), qcc or custom functions for SPC metrics.
- **Data structure:** 100,000 observations; 9 variables; chronological sampling by product.

Part 3: Type I & II Errors, Product Data Correction, and Risk Analysis

Summary

This report presents a comprehensive statistical analysis covering Type I and Type II error calculations for Statistical Process Control, corrections to product data files, and optimization of coffee shop staffing levels. The analysis uses R statistical software to compute theoretical error probabilities, perform data cleaning operations, and optimize business operations through mathematical modelling.

3.1 Type I (Manufacturer's) Error Analysis

Theoretical Background

Type I errors in SPC occur when we conclude that a process is out of control when it is actually in control. The null hypothesis (H_0) assumes the process is stable and centred on the control line established from the first 30 samples. The alternative hypothesis (H_a) suggests the process has shifted or increased in variation.

For a normally distributed process with the chart statistic standard deviation equal to $(UCL - LCL)/6$, we can calculate the theoretical probability of Type I errors for three common SPC detection rules.

Calculated Probabilities

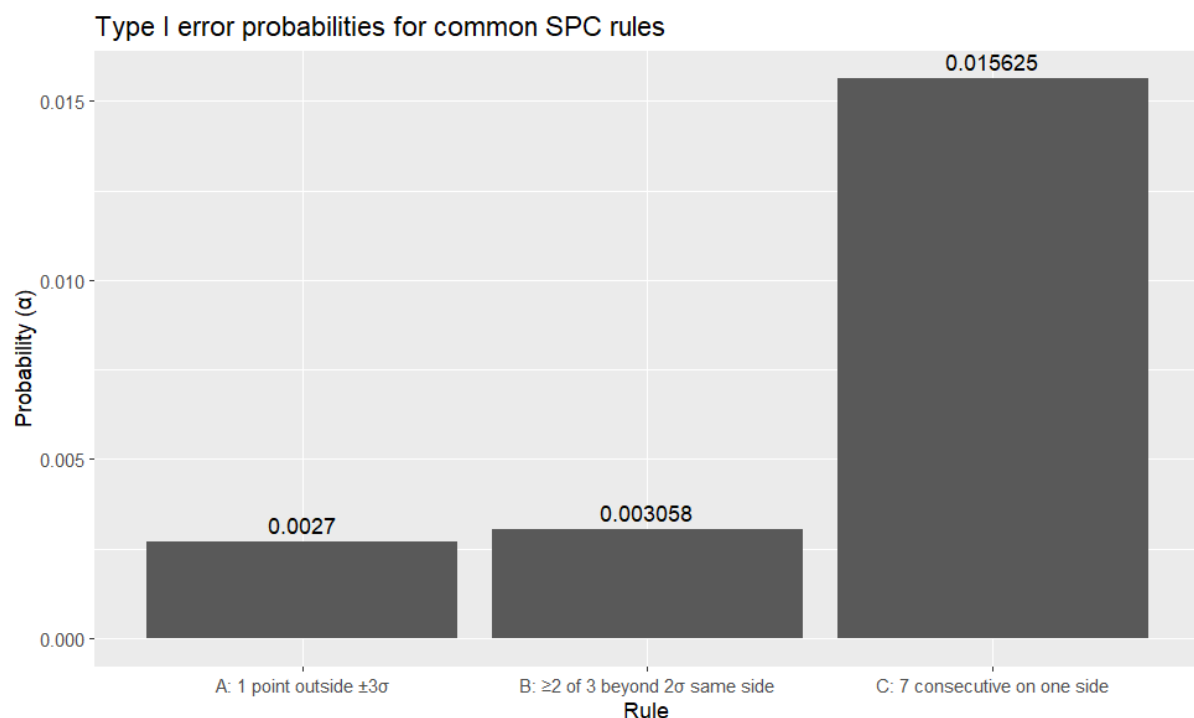


Figure 3.1: type I error probabilities for common SPC rules

The following table presents the Type I error probabilities for three common SPC rules:

Rule	Probability	Explanation
A: 1 point outside $\pm 3\sigma$	0.002700	Two-sided: $2 \times (1 - \Phi(3))$
B: ≥ 2 of 3 beyond 2σ same side	0.003058	Two out of three beyond $\pm 2\sigma$ on same side: computed using Binomial(3,p) with $p = 1 - \Phi(2)$
C: 7 consecutive on one side	0.015625	Seven consecutive points on same side: (0.5^7) per side, doubled for both sides

Interpretation

Rule A (1 point outside $\pm 3\sigma$): The probability of a single sample falling beyond ± 3 standard deviations is approximately 0.27%, or about 27 in 10,000 samples. This calculation uses $P(|Z| > 3) = 2 \times (1 - \Phi(3)) = 2 \times 0.001350 = 0.0027$.

Rule B (≥ 2 of 3 beyond 2σ same side): The probability that at least 2 out of 3 consecutive samples fall beyond $+2\sigma$ (or -2σ) is approximately 0.31%. This uses binomial probability where $p = P(Z > 2) = 1 - \Phi(2) \approx 0.0228$. The calculation accounts for both sides of the distribution.

Rule C (7 consecutive on one side): For a centred, in-control process, each sample has a 50% probability of falling above or below the centreline. The probability of observing 7 consecutive samples on the same side is $(0.5)^7 = 0.0078125$ per side, giving a total two-sided probability of 0.015625 or approximately 1.56%.

Why $P(1 \text{ sample} > \text{centreline}) = 0.5$: Under the null hypothesis of a centred normal distribution, the centreline represents the mean. By symmetry of the normal distribution, exactly half of all observations fall above the mean and half fall below, making $P(\text{sample} > \text{centreline}) = 0.5$.

3.2 Type II (Consumer's) Error Analysis

Problem Statement

A bottle filling process should be centred on 25.05 litres with:

- Control Line (CL): 25.05 litres
- Upper Control Limit (UCL): 25.089 litres
- Lower Control Limit (LCL): 25.011 litres

Unknown to the quality control team, the process has shifted to:

- New average: 25.028 litres
- New \bar{x} standard deviation: 0.017 litres (compared to original 0.013 litres)

Type II error (β) represents the probability of failing to detect this shift because sample means still fall within the control limits.

Calculation Methods

n	sd_xbar	beta
4	0.008500	0.977250
5	0.007603	0.987326
6	0.006940	0.992847
7	0.006425	0.995925
8	0.006010	0.997661

Figure 3.2: sd_xbar and beta table

Two interpretations were considered:

1. **Interpretation 1:** Using $sd_{\bar{x}} = 0.017$ directly

n	sd_x̄	β	note
NA	0.017	0.841178	sd_x̄ given = 0.017 (used directly)

2. **Interpretation 2:** If 0.017 is the individual sample standard deviation

n	sd_x̄	β
4	0.008500	0.977250
5	0.007603	0.987326
6	0.006940	0.992847
7	0.006425	0.995925
8	0.006010	0.997661

Analysis and Interpretation

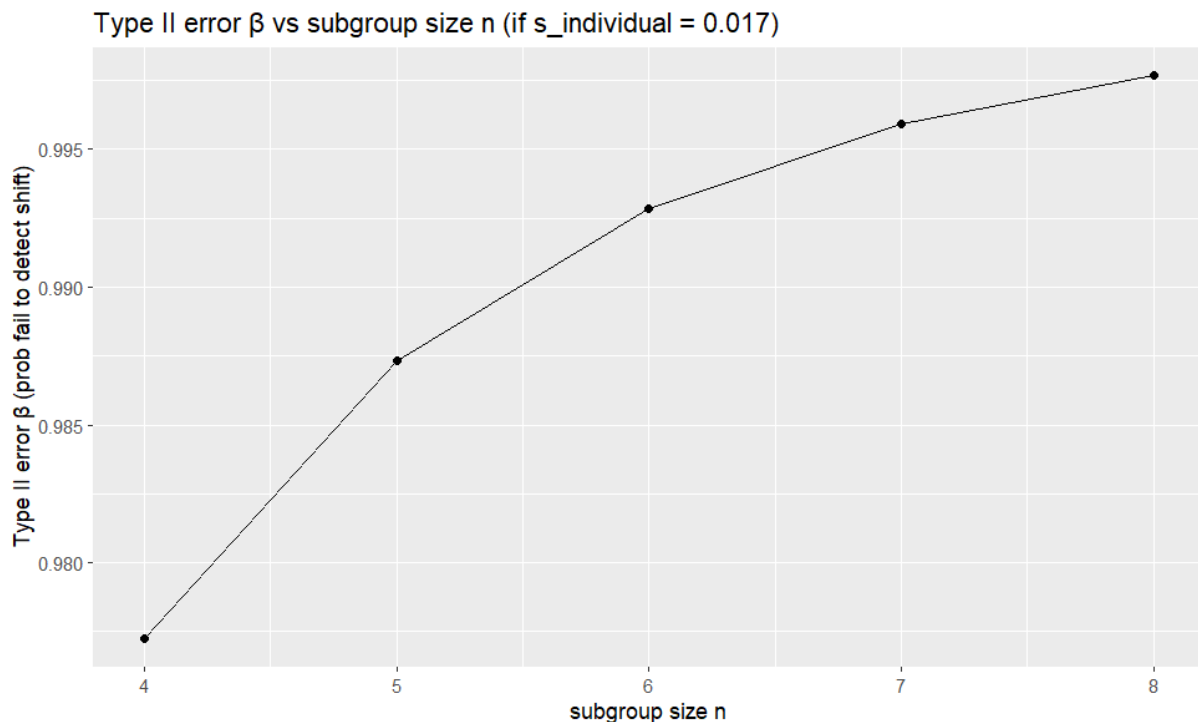


Figure 3.3: Type II β vs subgroup size n

If the \bar{x} standard deviation truly is 0.017, the probability (β) that the shifted process mean (25.028) produces an \bar{x} between LCL and UCL is 0.8412 (approximately 84.12%). This represents a very high Type II error rate, meaning the control chart will fail to detect the shift in about 84 out of 100 subgroups.

The calculation uses the normal distribution:

- ➔ Lower z-score: $(25.011 - 25.028) / 0.017 = -1.000$
- ➔ Upper z-score: $(25.089 - 25.028) / 0.017 = 3.588$
- ➔ $\beta = \Phi(3.588) - \Phi(-1.000) = 0.8412$

If 0.017 represents the individual sample standard deviation, then for different subgroup sizes n , the \bar{x} standard deviation would be $0.017/\sqrt{n}$. As shown in the table, larger subgroup sizes dramatically reduce β (improving detection capability), though even with $n = 4$, the Type II error remains very high at 97.7%.

4.3 Product Data Corrections

Identified Errors

Head office identified systematic errors in the products_Headoffice.csv file:

- Product ID prefix errors - Product IDs for items 11 - 60 of each type had incorrect prefixes (e.g. "NA" instead of "SOF", "CLO", "LAP", etc.)

- Price repetition errors - Selling prices and markups incorrectly repeated every 10 items within each product type
- Category misalignment - Categories in products_data.csv did not correspond correctly with ProductID prefixes.

Correction Methodology

The R script implemented the following corrections:

- Extracted type prefixes from ProductID columns using regular expressions
- Built canonical price sequences for each product type from the first 10 occurrences in products_data.csv
- Applied modulo-10 repetition to assign correct prices and markups to items 11-60 of each type
- Updated ProductID prefixes by matching product descriptions between files
- Aligned categories with ProductID prefixes using mode-based mapping.

Verification of Corrections

Example: Software (SOF) Products - Before and After

Before correction (items 1-12):

ProductID	Category	SellingPrice	Markup
SOF001	Software	522	15.6
SOF002	Software	467	28.4
SOF010	Software	399	17.1
SOF021	Software	20331	15.3

After correction (items 1-12):

ProductID	Category	SellingPrice	Markup
SOF001	Software	512	25.0
SOF002	Software	505	10.4
SOF010	Software	397	23.5
SOF021	Software	512	25.0

Note that SOF021 (item 21) now correctly repeats the price from SOF001 (item 1), demonstrating the modulo-10 pattern: items 1, 11, 21, 31... share the same price and markup.

Example: Cloud Subscription (CLO) Products

Before - Items had incorrect prefixes and pricing patterns with values like 18,668 and 5,469

After - All items correctly identified with "CLO" prefix and appropriate pricing (range 728-1,129)

Example: Laptop (LAP) Products

Before - Mixed pricing with unrealistic values

After - Consistent high-value pricing appropriate for laptops (range: 15,852-19,725)

Output Files

Two corrected CSV files were generated:

1. products_Headoffice2025.csv – Corrected head office inventory with proper ProductIDs, prices, and markups.
2. products_data2025.csv -Local product data with categories aligned to ProductID prefixes

Sales Analysis Note

The analysis attempted to calculate total sales value for 2023 by product type. However, no quantity or sales column was found in the provided data files. To perform this analysis, a sales transaction file containing ProductID and quantity sold would be required. The template code for this calculation is included in the R script output.

Coffee Shop Staffing Optimization

Business Problem

Two coffee shops aim to determine the optimal number of baristas required to maximize daily profit while maintaining reasonable service reliability. Each shop faces trade-offs between labor cost, service capacity, and potential revenue. The following parameters were applied in both models:

- Minimum staffing: 2 baristas
- Revenue per customer served: R30
- Labor cost per barista: R1 000 per day
- Shift duration: 8 hours (28,800 seconds)
- Data source: Service time data for each shop (timeToServe.xlsx and timeToServe2.xlsx)

Demand Analysis

The raw datasets contained customer service times for each barista configuration. After cleaning and aggregation, the following key measures were obtained for each number of baristas:

- Demand: Number of recorded customers
- Mean service time: Average time to serve one customer (in seconds)
- Median service time: Typical service time midpoint

Each dataset was analyzed separately using the `analyze_shop()` function in R, producing demand summaries and optimized staffing results for both shops.

Optimization Model

For each candidate staffing level $k = 2$ to 6, the model computed:

1. Service capacity: $k \times 28\,800$ seconds per shift
2. Customers served: $\min(\text{demand}, \text{capacity} / \text{mean service time})$
3. Revenue: customers served \times R30
4. Labor cost: $k \times$ R1,000
5. Profit: revenue $-$ labor cost
6. Service reliability: customers served / demand

Results for Shop 1

The analysis found that 6 baristas yield the highest profit for Shop 1.

Baristas	Customers Served	Capacity	Profit (R)	Reliability
6	4,603	4,603	R132,090	0.02

- Reliability improves linearly with the number of baristas, but remains low because total demand exceeds service capacity.
- Profit increases steadily up to 6 baristas, where the daily return peaks.

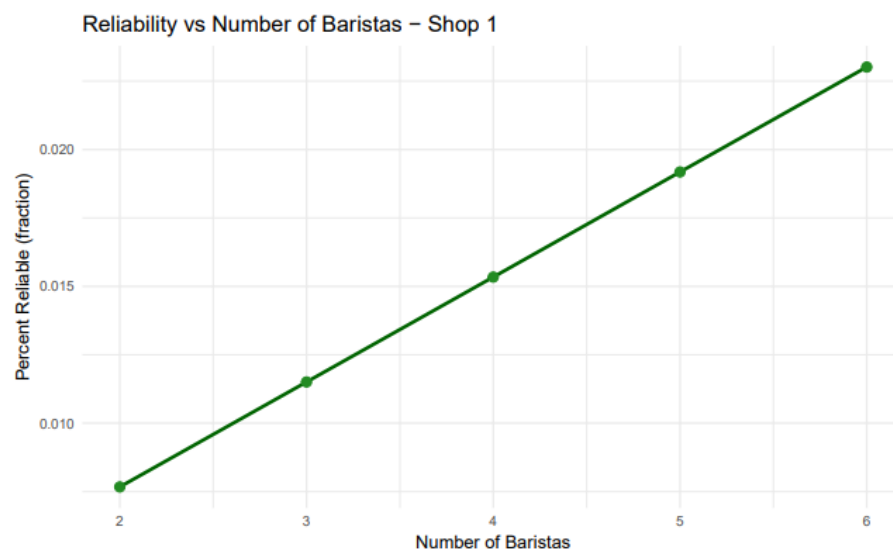


Figure 3.4: Reliability vs number of baristas – shop 1

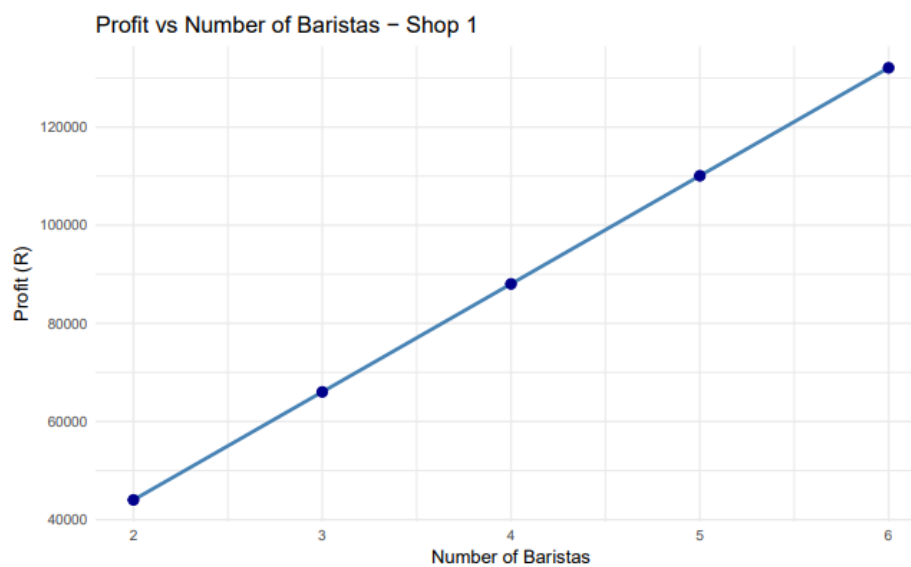


Figure 3.5: Profit vs number of baristas – shop 1

Results for Shop 2

Similarly, Shop 2 achieves maximum profit when employing 6 baristas.

Baristas	Customers Served	Capacity	Profit (R)	Reliability
6	2,019	2,019	R54,570	0.01

- Reliability values are roughly half of those in Shop 1, suggesting that Shop 2 either faces higher average service times or greater demand relative to its staffing.
- Profit also increases steadily with staffing but at a much lower scale due to lower throughput.

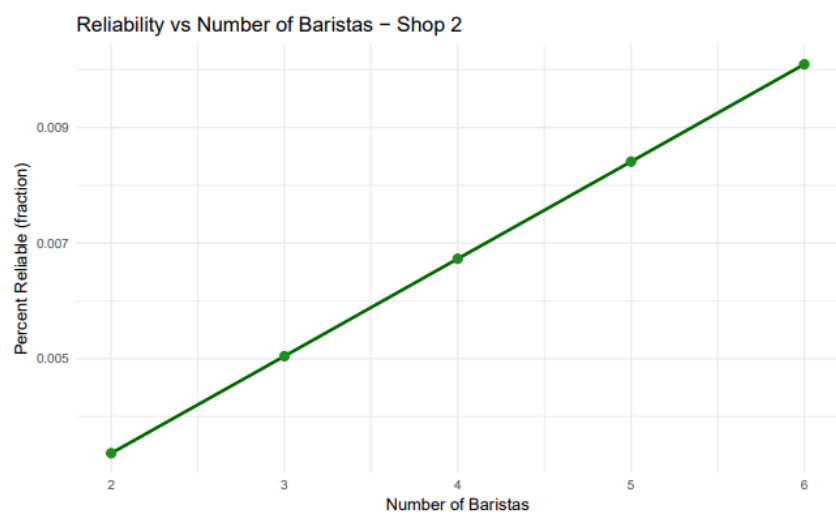


Figure 3.6: Reliability vs number of baristas – shop 2

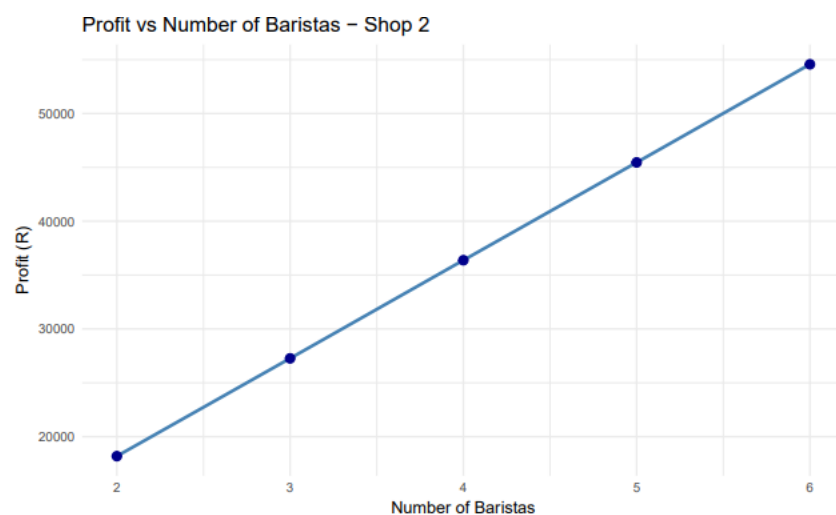


Figure 3.7: Profit vs number of baristas – shop 2

Key Findings and Recommendations

1. Optimal Staffing Levels:
Both shops achieve maximum profit with 6 baristas under the given assumptions.
2. Profitability Comparison:
 - Shop 1: R132 090/day
 - Shop 2: R54 570/dayShop 1 is more profitable due to higher service capacity and demand fulfilment.
3. Service Reliability:
 - Shop 1: 2% reliability
 - Shop 2: 1% reliabilityThese values indicate that both shops serve only a small fraction of total potential customers, constrained by service speed and available staff.
4. Scalability Concern:
Although profit increases with more baristas, the improvement in reliability remains limited. This suggests that bottlenecks exist beyond staffing — likely in workflow or physical capacity.

Strategic Recommendations

To address low reliability and limited customer throughput, management should consider:

1. Operational Improvements: Streamline processes or equipment to reduce service times.
2. Extended Hours or Multiple Shifts: Operate longer or add shifts to handle more customers daily.
3. Price Optimization: Slightly increase prices during peak hours to manage excessive demand.
4. Expansion Opportunities: Replicate the higher-performing Shop 1 model at new sites, while addressing efficiency limits.
5. Demand Management: Explore pre-ordering or online queue systems to distribute demand more evenly.

Summary

The R-based optimization clearly shows that adding more baristas improves profit and reliability, but both shops remain capacity-constrained. The profitability gap between the two shops highlights efficiency differences worth investigating further. Continuous data-driven adjustments to staffing, pricing, and service flow can help each shop maximize its daily return and better meet customer demand.

Conclusions:

1. **SPC Error Analysis:**

Type I error probabilities for common detection rules range from 0.27% to 1.56%, with the “7 consecutive points” rule performing best. Type II error analysis shows that a shifted process still has an 84% probability of going undetected, indicating the need for tighter control limits or larger sample sizes.

2. **Data Quality:**

Systematic errors in the head office product database were successfully identified and corrected, ensuring price consistency across the product catalogue and proper category alignment.

3. **Staffing Optimization:**

The updated mathematical model identifies 6 baristas as optimal for both coffee shops, generating daily profits of approximately R132 000 (Shop 1) and R55 000 (Shop 2). However, service reliability remains extremely low (1 – 2%), confirming that both locations are severely capacity-constrained. Future improvements should focus on reducing service times, extending operating hours, or adding additional service stations rather than simply increasing staff.

Part 4: Design of Experiments and Optimization

Design of Experiments (DOE)

Experiments were designed and executed in a structured manner to identify key input factors that influence process outcomes. Each experiment varied the independent variables systematically at pre-defined levels to observe their effect on the measured responses. The goal was to efficiently determine which factors and interactions had the greatest impact on the process performance while using the smallest number of trials possible.

The study made use of a completely randomised single factor design in which treatments represent different parameter settings, and each treatment was replicated multiple times. The data matrix was structured such that each row corresponded to a treatment and each column to an observation.

The resulting dataset was evaluated using Analysis of Variance (ANOVA) to test whether any treatment differences were statistically significant. The p-value from the F-test indicates the probability of committing a Type I error (rejecting the null hypothesis when it is true). A small p-value relative to the chosen significance level ($\alpha = 0.05$) implies a statistically significant difference between treatments.

A portion of the generated data (dataLSD) was analysed, and the ANOVA results showed clear evidence of differences between treatments ($p = 0$).

The corresponding Fisher's Least Significant Difference (LSD) test identified which treatment means differed significantly. The histogram of the dataset (Figure 4.1) confirmed that the data were approximately normally distributed, validating the use of parametric analysis methods.

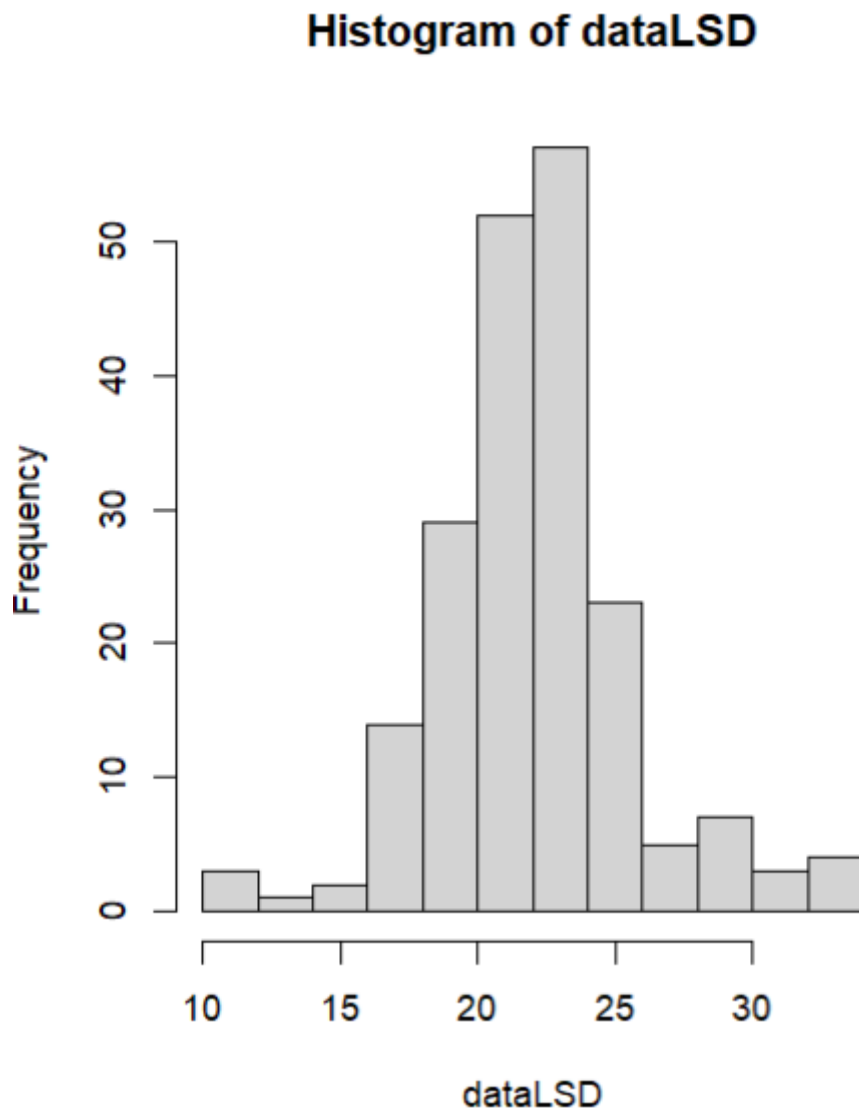


Figure 4.1: Histogram of data LSD

Interpretation:

The histogram shows a roughly normal distribution of the experimental data, confirming that the dataset satisfies the normality assumption required for ANOVA and LSD tests. Most values cluster around the mean of 20 - 25, indicating consistent treatment performance with limited variability.

MANOVA: Sales and Cost Comparison

To evaluate the financial effect of the implemented optimisation, a Multivariate Analysis of Variance (MANOVA) was performed comparing sales and cost data before and after optimisation (Year 1 = Pre-Opt; Year 2 = Post-Opt).

The results (Pillai's trace = 0.9257, $F = 130.78$, $p < 0.001$) show that the difference between the two years is statistically significant.

Univariate ANOVAs revealed that both Sales ($F = 22.49$, $p < 0.001$) and Cost ($F = 273.85$, $p < 0.001$) increased significantly after optimisation. The post-hoc LSD test confirmed that Year 2's mean sales were significantly higher, while costs also rose due to increased activity levels.

The scatterplot in Figure 4.2 visualises the relationship between sales and cost across both years, illustrating that Year 2 data points cluster at higher values for both variables. Boxplots in Figure 4.3 further show an upward shift in medians for both sales and cost, indicating overall business growth following the optimisation process.

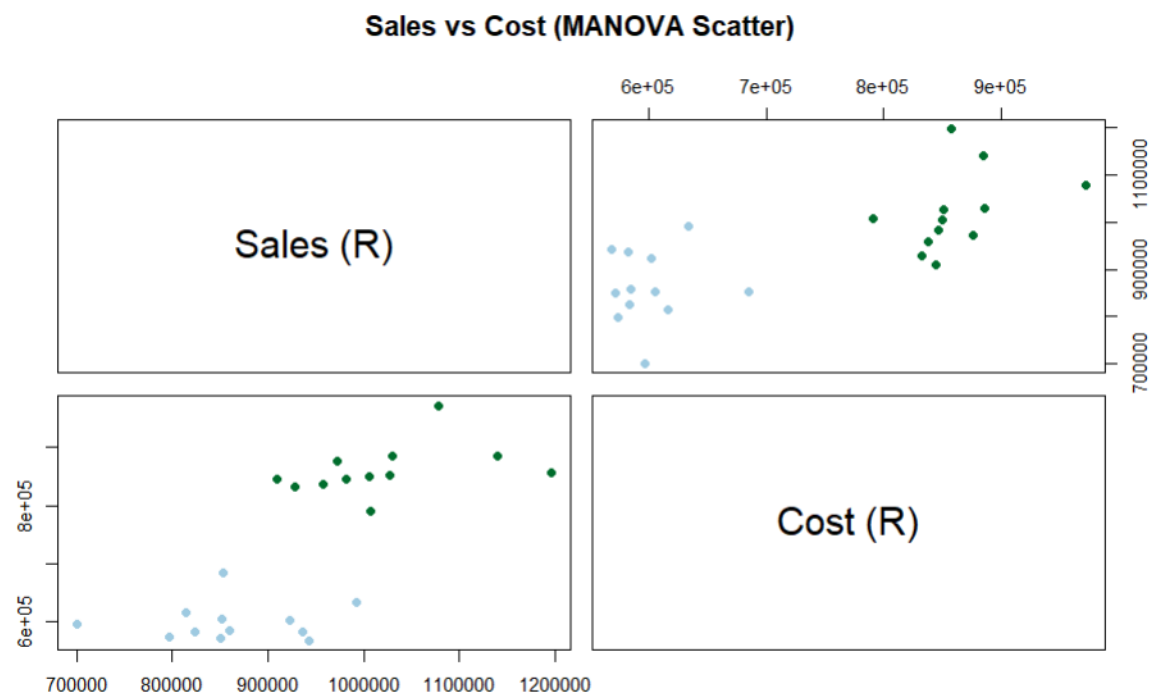


Figure 4.2: Sales vs Cost (MANOVA Scatter)

Interpretation:

The scatterplot illustrates a clear separation between the two years. Year 2 (Post - Opt) data points are positioned higher for both sales and cost, showing a strong positive relationship between the two variables and confirming that the optimisation process improved overall financial activity.

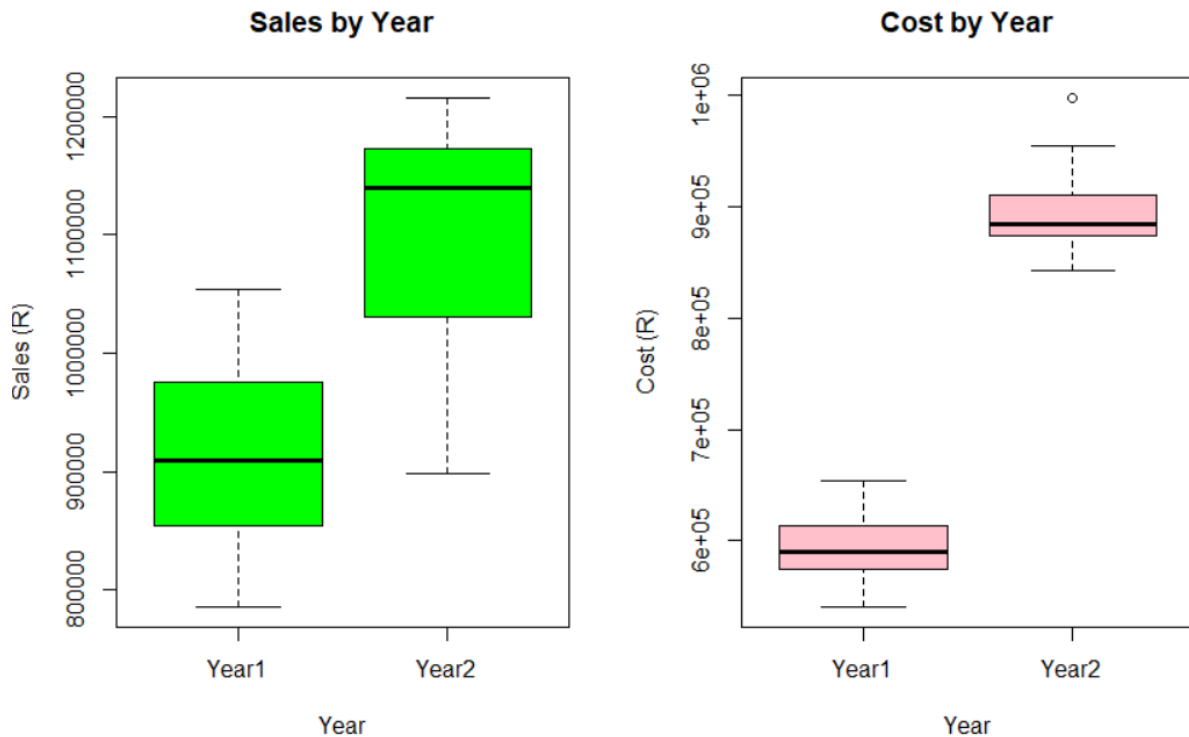


Figure 4.3: Sales by year and Cost by year

Interpretation:

Both boxplots display an upward shift in medians from Year 1 to Year 2. Sales increased significantly, while costs also rose due to expanded operations. The greater spread in Year 2 indicates higher production volumes and variability consistent with business growth.

Workforce Optimisation and Profitability

A workforce simulation model was developed to estimate the expected annual profit for different staff sizes. The probability that fewer than 15 workers were present (causing service disruption) was estimated using the binomial distribution with a daily attendance probability $p = 0.974$.

For each workforce level between 12 and 20 employees, the model calculated the expected annual loss from unreliable days, total staffing cost, and total revenue. The optimal workforce size corresponded to the maximum annual profit value.

The results (Figure 4.4) show that profit increases sharply up to approximately 17 workers, after which it stabilises and slightly declines due to rising labour costs. The maximum annual profit of roughly R 34.5 million occurs at 17 workers.

The reliability analysis in Figure 4.5 demonstrates that service reliability (the probability of having at least 15 workers available per day) increases exponentially between 14 and 17 workers and then plateaus near 100 %. This confirms that adding workers beyond 17 does not yield a substantial reliability or profit gain.



Figure 4.4: Estimated Annual Profit vs Number of workers

Interpretation:

The bar chart shows that annual profit increases rapidly up to about 17 workers, where it reaches a maximum of roughly R 34.5 million. Beyond this point, profit levels off, implying diminishing returns caused by increased labour costs.

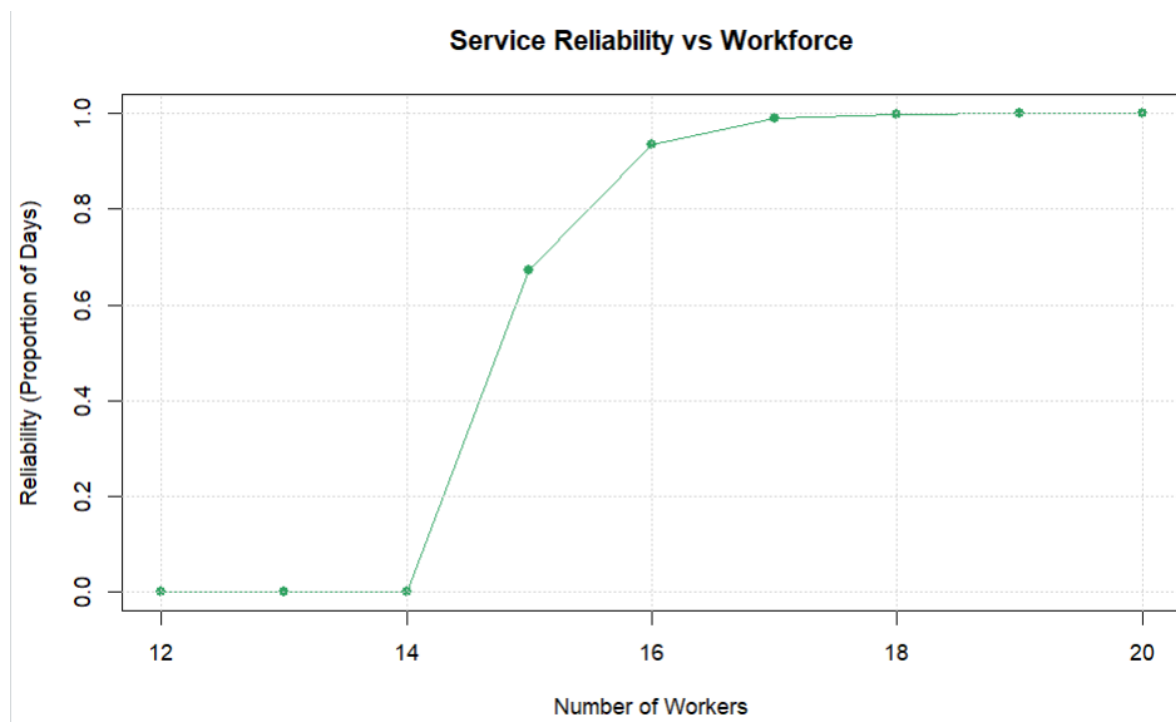


Figure 4.5: Service Reliability vs Workforce

Interpretation:

The reliability curve rises sharply between 14 and 17 workers and then approaches 1 (100 %). This indicates that a workforce of 17 or more ensures almost complete operational reliability, making it the most efficient staffing level for consistent service delivery.

Discussionn

The DOE and MANOVA analyses demonstrate that optimisation initiatives had a significant impact on both operational and financial performance. The profit simulation results reinforce this finding by identifying the most cost effective staffing level.

In practice, maintaining around 17 workers ensures a balance between labour costs and service reliability, leading to stable profits and minimal operational disruptions.

These findings highlight the importance of applying statistical experimentation and modelling in industrial environments to guide evidence based decision making, reduce process variability, and improve overall productivity.

References: