ENGINEERING COUNSEL OF SOUTH AFRICA REPORT

Wayde Davids - 27059006

Stellenbosch
UNIVERSITY
IYUNIVESITHI
UNIVERSITEIT

Stellenbosch University
Quality Assurance 34424 October 202524 October 202

Table of Contents

Table of Tables:

Table of Figures

Abstract

This report presents a comprehensive application of data analytics methodologies to solve a series of multifaceted business problems, fulfilling the requirements of the Engineering Counsel of South Africa (ECSA) Graduate Attribute 4. The analysis portfolio demonstrates an end-to-end data workflow, from initial data wrangling and descriptive analysis to advanced statistical modelling and optimization.

Key findings from the descriptive analysis revealed a revenue structure following a Pareto distribution, heavily reliant on a small cohort of high-value customers and "star" products, as well as a significant sales decline in late 2023. A critical data correction exercise subsequently revealed that initial sales metrics had been significantly overestimated, reinforcing the importance of data integrity.

Statistical Process Control (SPC) analysis of delivery times found that no process was robustly capable of meeting the 32-hour specification, with several processes exhibiting significant instability. A follow-up ANOVA statistically confirmed that delivery performance is driven by complex and non-static interactions between Product Type, Month, and Year, invalidating a one-size-fits-all management approach.

Finally, optimization models provided clear, actionable recommendations. Analysis of two coffee shops revealed distinct operational profiles, one demand-constrained and the other operationally inefficient, requiring unique management strategies. A binomial reliability model for a car rental agency identified an optimal staffing level of 25 employees, offering a net monthly cost saving of approximately R20 448. This report concludes that effective business improvement is contingent on rigorous data validation and a granular, multi-dimensional understanding of process-specific factors

Overall Introduction

This report presents a portfolio of analytical projects demonstrating proficiency in advanced data analysis and statistical modelling, as required by the Engineering Counsel of South Africa (ECSA) Graduate Attribute 4 (GA4). The primary objective is to apply a systematic, data-driven approach to distinct business cases, translating raw data into actionable strategic insights.

The scope covers the full analytical lifecycle: foundational data cleaning, exploratory analysis, process control and capability assessment, data integrity validation, and the development of optimization models.

Methodologies include Descriptive Statistics, Statistical Process Control (SPC) with $C_p\ and\ C_{pk}$ analysis, theoretical error calculation, data correction, profit optimization, and a multi-way Analysis of Variance (ANOVA).

This report is structured to follow the project requirements. It begins with a descriptive analysis, implements SPC, addresses data correction, builds optimization models, and conducts an ANOVA. Each section forms part of a comprehensive analytical narrative, moving from observation to diagnosis and, finally, to recommendation.

Part 1: Descriptive Statistics and Analysis

1.1 Introduction

This report presents the results of an exploratory analysis of the company's datasets. The goal of this first stage was to describe the customer base and related information using the available data files, highlight initial insights, identify potential pitfalls, and recommend next steps for further analysis.

The analysis is based on four datasets: a sales table with 100 000 rows, two product reference files (containing 60 and 360 entries, respectively), and a customer file containing 5000 records. The sales table provides information such as customer and product IDs, quantities, and order times, but does not yet contain a monetary sales variable or a consolidated date field. The customer file records age, gender, income, and city.

A high-level statistical summary of the 5000 customers shows an average age of 51.6 years (with a median of 51) and a mean income of approximately $80 797 (with a median of $85 000). The gender distribution is 48.6% female, 47.0% male, and 4.4% Other. The most common customer location is San Francisco, with about 780 customers.

1.2 Foundational Customer Demographics

The foundational analysis of the customer portfolio reveals a striking degree of balance and uniformity, which in itself is a significant finding.

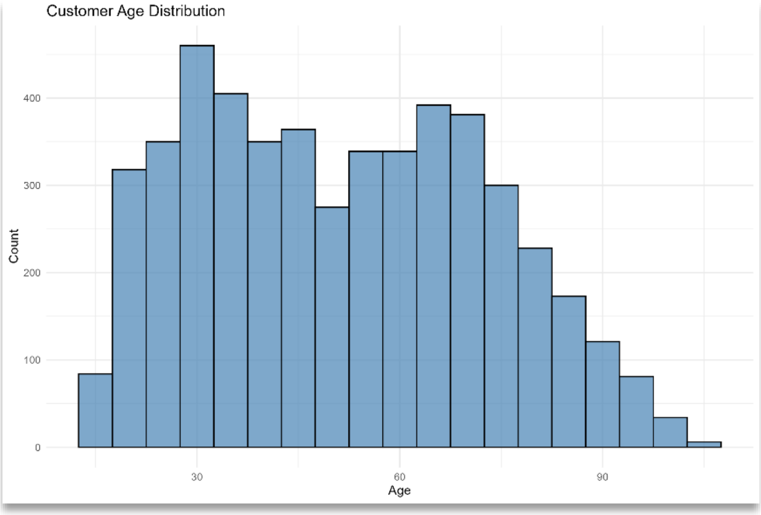

Figure 1: Customer Age Distribution

The age histogram shows that the customer base is concentrated in midlife and older age groups. Most customers fall between 40 and 60 years of age, which aligns closely with the mean and median values of around 51 years. This concentration suggests that products and services could be effectively marketed with mature adults in mind, perhaps emphasising reliability and support. However, this graph also highlights a potential data quality issue: the maximum observed age of 105 is implausible in a commercial dataset and likely reflects an error or placeholder, which could distort statistical measures.

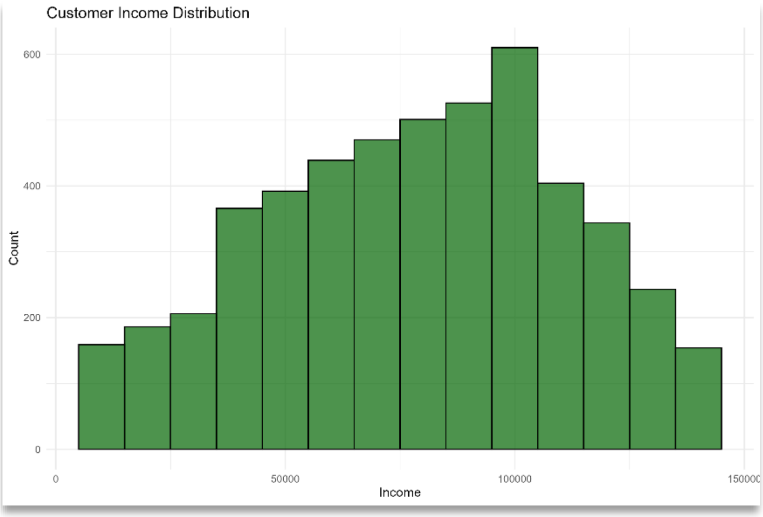

Figure 2: Customer Income Distribution

The income histogram reveals a moderately skewed distribution. There is a wide spread of incomes up to $140 000, with a notable concentration of customers in the middle-to-upper range. The mean income is approximately $80 797, while the median is slightly higher at $85 000. The closeness of these two central tendency measures confirms the skew is not extreme. From a business standpoint, this pattern indicates that the customer base is predominantly middle-to-upper income, which allows for premium product offerings but also suggests maintaining affordable tiers.

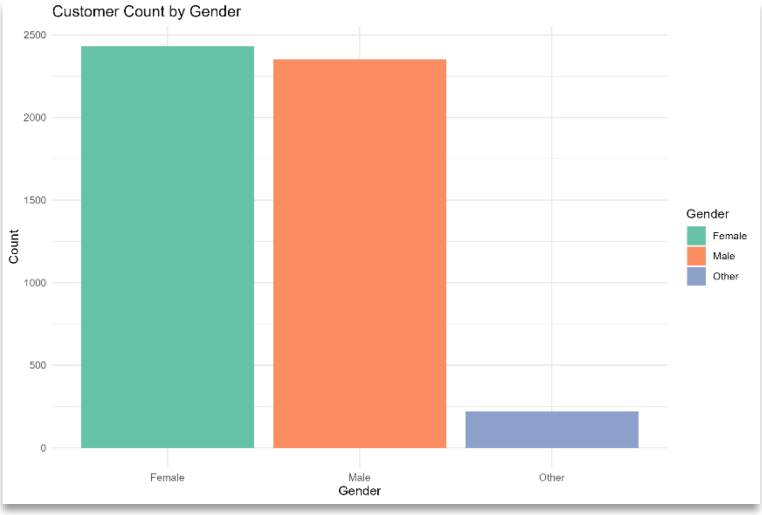

Figure 3: Customer Distribution by Gender

This bar chart shows an almost equal split between male (47.0%) and female (48.6%) customers, with a smaller group (4.4%) identifying as 'Other'. This near parity suggests broad appeal across binary gender categories. However, the relatively small size of the 'Other' group (218 individuals) means that any statistical analyses focusing on this segment will be weaker and must be interpreted with caution.

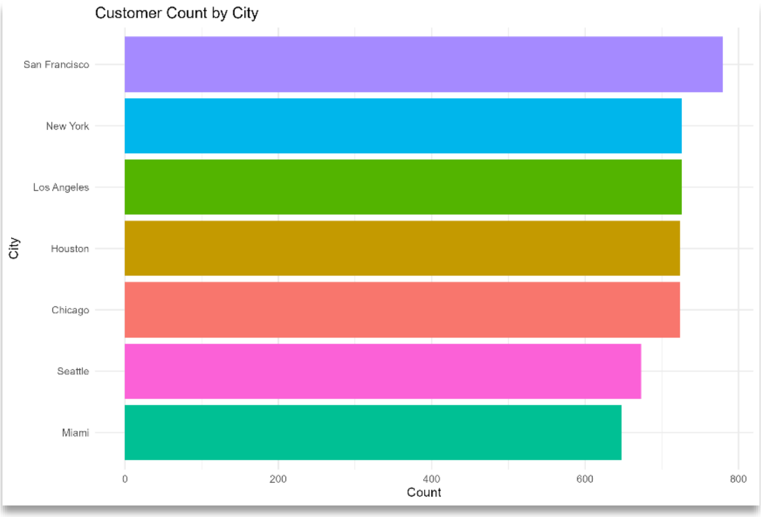

Figure 4: Customer Distribution by City

The chart showing customer counts by city reveals a strong geographic concentration. San Francisco dominates with approximately 780 customers, while a handful of other large metropolitan areas contribute the bulk of the remaining customers. Operationally, this pattern highlights where customer demand is concentrated, which can inform logistics and regional marketing. However, these raw counts can be misleading, as true market penetration must be measured relative to the total population of those cities.

1.3 Customer Value and Segmentation Analysis

Moving beyond simple demographics, a more nuanced segmentation reveals significant economic relationships.

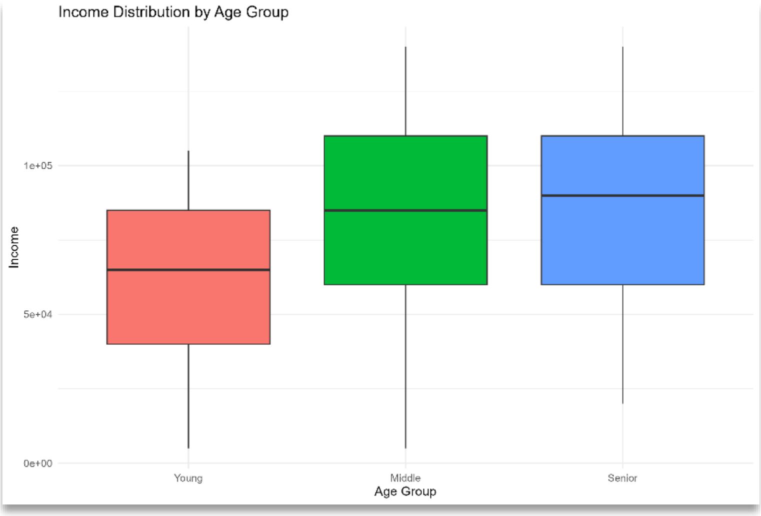

Figure 5: Income Distribution by Age Group

This set of boxplots compares the distribution of customer income across three defined age segments ('Young', 'Middle', and 'Senior'). This plot reveals a significant relationship that was not visible in the general histogram. The 'Senior' customer segment exhibits both the highest median income and the widest interquartile range (IQR). This indicates not only a higher average purchasing power but also a greater diversity of income levels within that segment, identifying them as a key high-value target group.

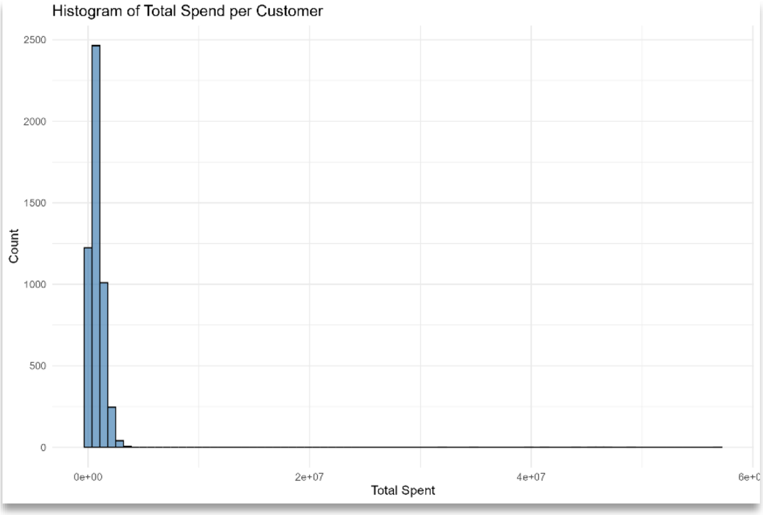

Figure 6: Distribution of Total Spend Per Customer

This histogram displays the frequency of customers based on their total cumulative spend. The distribution is highly positively skewed (right skewed). The vast majority of customers are clustered at the low end of total spending, with a "long tail" of a very small number of customers who have spent exceptionally large amounts. This distribution is a classic example of a Pareto principle (or "80/20 rule"), where a disproportionately large amount of revenue is likely generated by a small fraction of the customer base.

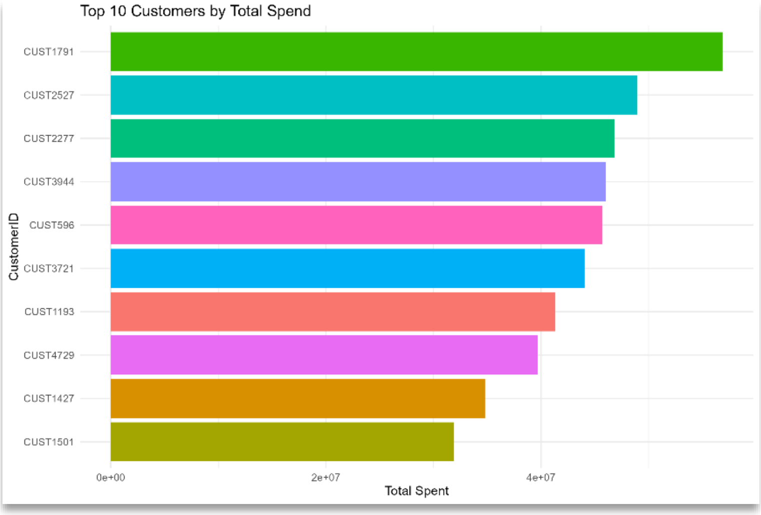

Figure 7: Top 10 Customers by Total Spend

This bar chart provides actionable intelligence by identifying the "whales" or high value "power users" who constitute the extreme right tail of the distribution noted above. These specific *CustomerIDs* represent high-priority accounts for retention, loyalty programs, and strategic relationship management.

1.4 Product and Sales Performance Analysis

From a product perspective, performance is not uniform.

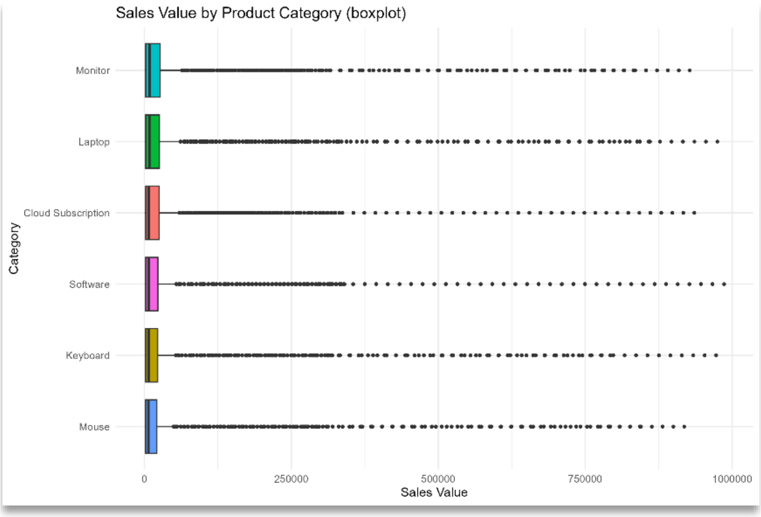

Figure 8: Sales Value by Product Category

This analysis demonstrates significant differences between product classes. 'Laptops' stand out as a clear high-value category, with the highest median sales value and the largest IQR, suggesting a wide range of pricing from entry-level to high-end models. Conversely, 'Keyboard' and 'Mouse' transactions are, as expected, clustered at a low sales value. This is crucial for understanding the product mix.

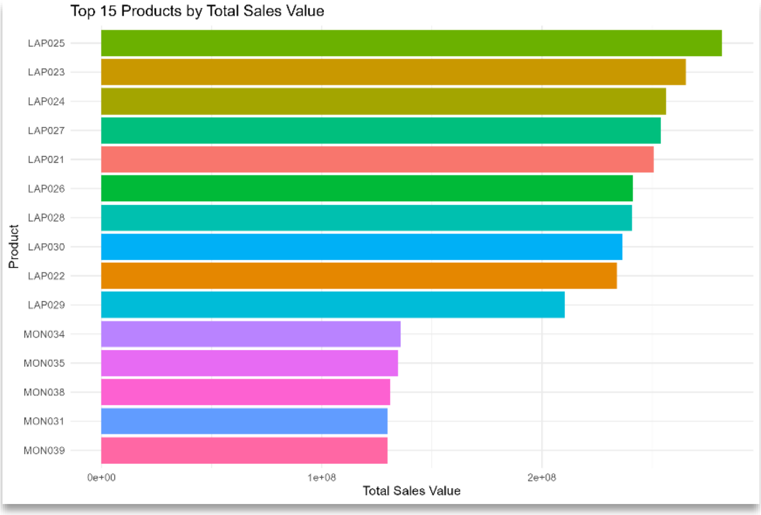

Figure 9: Top 15 Products by Total Sales Value

This plot drills down from the category level to the specific product level, ranking the "star" performers in the product portfolio. It identifies specific items, such as LAP026 and CLO011, as the primary revenue drivers. This insight is essential for inventory management and marketing focus.

1.5 Time-Series and Performance Analysis

Perhaps the most strategically significant finding of the descriptive analysis comes from the time-series data.

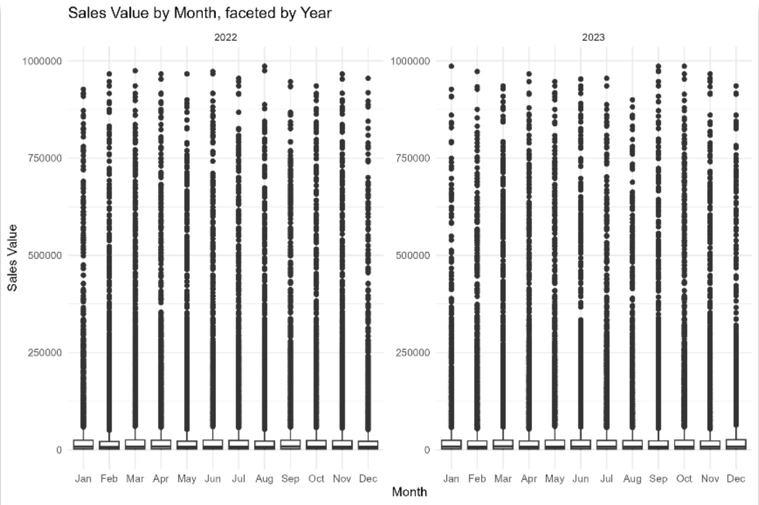

Figure 10: Monthly Sales Value Distribution by Year

This visualization reveals two critical, actionable trends simultaneously:

1. A Consistent Seasonal Trend: In *both* 2022 and 2023, a clear seasonal dip is observable in January and December. The median sales value and the entire box (IQR) are consistently lower in these months, indicating a predictable seasonal pattern.
2. A Significant Performance Downturn: A concerning negative trend is evident when comparing 2023 to 2022. From July to December 2023, the boxplots are visibly lower and more compressed than their 2022 counterparts. This indicates a sustained drop in sales performance in the second half of 2023, a major finding that warrants further investigation.

1.6 Identified Data Quality Issues and Limitations

A robust analysis requires a critique of the data's limitations and potential issues, several of which were identified during the exploration.

- Data Source Discrepancy: The most significant issue is a data integrity problem. Analysis of product pricing data reveals a clear discrepancy between the *products_data* and *products_Headoffice* files. The *HeadOffice* file shows a significantly higher median selling price and wider distribution. This suggests a lack of a single source of truth for product pricing.
- Implausible & Outlier Data: The customer data contains outliers, such as an age of 105, which are implausible and must be validated before being included in any analysis. Similarly, operational data (e.g., *pickingHours*) shows values in excess of 40, which is physically impossible for a single order and suggests the variable's definition is unclear or the data is flawed.
- Methodological Limitations: Relying on *means* for skewed variables like income is not sufficient and should be accompanied by *medians* and dispersion measures. Furthermore, city-level comparisons of raw customer counts are misleading and must be normalized by population size or potential market size to measure true penetration.

- Suboptimal Data Structure: The raw sales data was sub-optimally structured. Temporal data was split across three separate integer columns (*orderDay, orderMonth, orderYear*). This required data wrangling to create a proper Date object before any time-series analysis could be performed.

1.7 Recommendations and Next Steps

Based on this analysis, the priority for deeper analysis is to create a reliable monetary sales measure and a consolidated date field. Linking sales to an authoritative product price (after reconciling the two product files) will allow for the calculation of revenue, customer lifetime value, and product profitability.

Further analysis should use medians, quartiles, and confidence intervals for robust comparisons, especially across skewed variables. City-level reporting should include per-capita measures and error bars to account for different population sizes. Finally, after revenue metrics are created, segment analysis and predictive modelling can link these demographic profiles to actual purchasing behaviour, improving business strategy.

1.8 Part 1 Conclusion

In conclusion, the descriptive analysis has established several key characteristics of the business. The customer base is large and demographically diverse, with no significant concentration by age, income, gender, or geography. However, a deeper segmentation reveals that 'Senior' customers possess higher average incomes. The business's revenue model is characterized by a Pareto distribution, with a heavy reliance on a small cohort of high value "whale" customers and a portfolio of "star" products, dominated by the 'Laptop' category.

Most critically, the time-series analysis reveals a dual-pattern of predictable seasonality (Jan/Dec dips) combined with an alarming, non-seasonal sales decline in the latter half of 2023. Finally, critical data quality issues have been identified, including source-of-truth discrepancies in pricing and implausible operational data, which must be addressed. These findings provide a comprehensive and insightful context for the advanced analyses to follow.

Part 3: Statistical Process Control (SPC) Analysis

3.1: SPC Setup and Control Limit Validation

3.1.1: Methodology and validation of control limits

Before a process can be monitored, it is a mandatory prerequisite to establish its natural operating limits. This "Phase 1" analysis uses an initial, in-control dataset to calculate the parameters of the process. For this report, the *deliveryHours* variable from the *sales2026and2027.csv* file was used.

The data was first sorted chronologically. For each of the six product types, the first $k = 30$ subgroups, each of size $n = 24$, were extracted. This resulted in a total of 720 individual delivery time observations per product type. The grand average ($\bar{X}$) and the average subgroup standard deviation ($\bar{S}$) were computed from these 30 subgroups. These parameters, along with the standard statistical constants for $n = 24$ (A_3, B_3, B_4), were used to establish the centre lines (CL), 1-sigma, 2-sigma, and 3-sigma Upper and Lower Control Limits (UCL/LCL) for both the $\bar{X}$-chart and the s-chart.

Before these calculated limits can be used for "Phase 2" monitoring, the initial 30-sample dataset must be validated. This validation ensures the process was stable, in statistical control, and that its underlying distribution meets the assumptions of SPC. The 'CLO' product type is presented here as a representative example of this validation process.

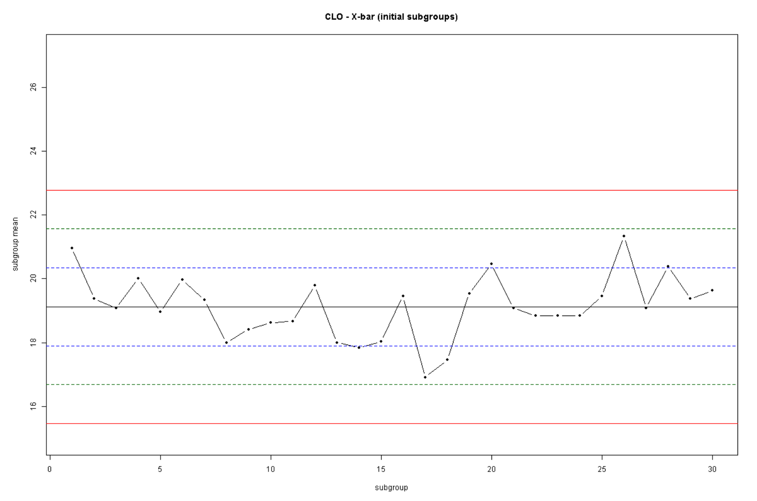

Figure 11: 'CLO' X-bar Chart (Initial 30 Subgroups)

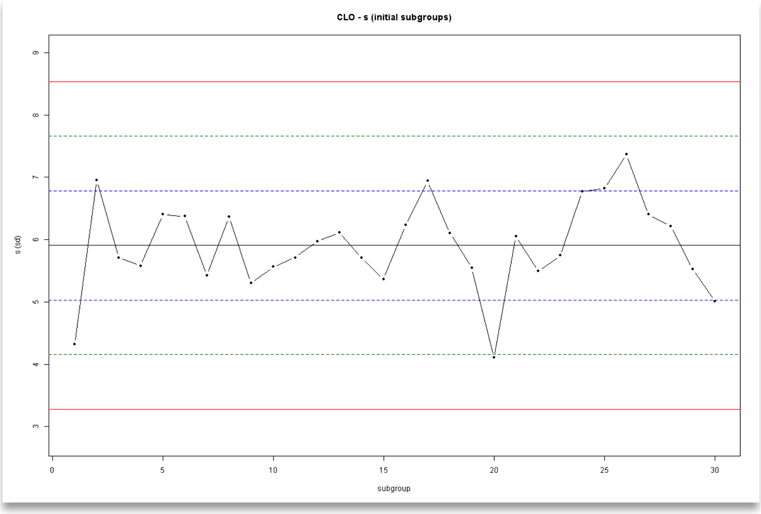

Figure 12: 'CLO' s-Chart (Initial 30 Subgroups)

3.1.2: Process Stability Analysis

Figures 11 and 12 are the primary tools for assessing process stability. The $\bar{X}$-chart plots the *mean* delivery time for each of the 30 subgroups, while the s-chart plots the *variation*, more specifically standard deviation, *within* each subgroup.

As shown in both charts, all 30 subgroups fall comfortably within the 3-sigma UCL and LCL, which are depicted by the red lines. Furthermore, the points demonstrate random variation around their respective CL. There are no obvious trends, runs, or other non-random patterns that would indicate the presence of "special cause" variation. The stability of *both* the process average and its variation confirms that the 'CLO' delivery process was in a state of statistical control during this initial period. Therefore, the calculated control limits are considered statistically valid and reliable.

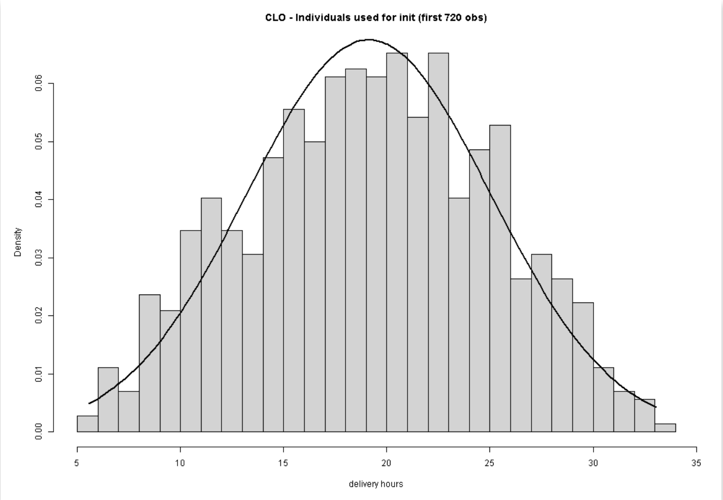

Figure 13: 'CLO' Histogram (Initial 720 Individuals)

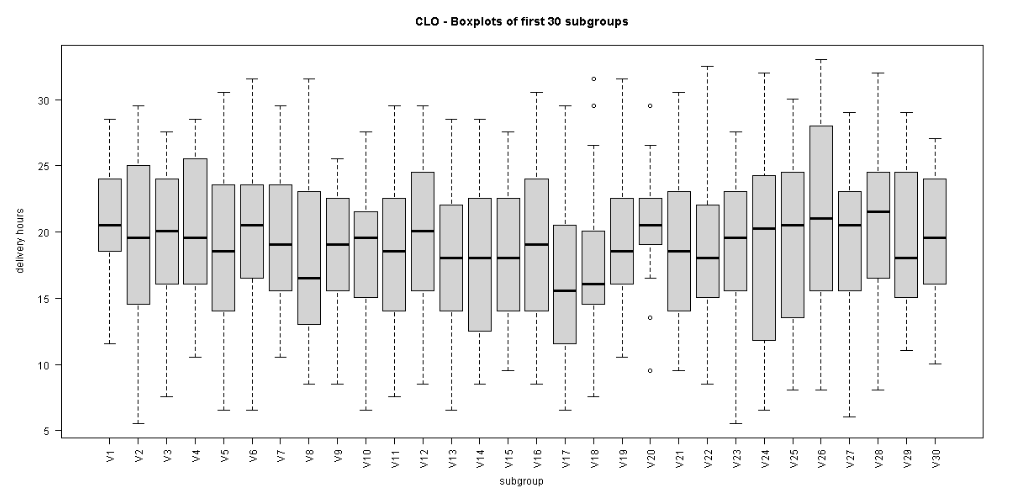

Figure 14: 'CLO' Boxplots (Initial 30 Subgroups)

3.1.3: Distribution Analysis

Figures 13 and 14 are used to validate the distributional assumptions of the data. The histogram displays the distribution of all 720 individual delivery times used for the setup. The distribution is unimodal and approximately normal, resembling a classic bell curve. It is centred around a mean of approximately 19.23 hours. This approximate normality is a key assumption for SPC, which this data satisfies.

The boxplot analysis provides a more granular view, showing the distribution of each of the 30 subgroups. This plot confirms the stability seen in the s-chart; the medians and interquartile ranges (IQRs) of the subgroups are consistent, with no significant, recurring outliers that would suggest instability or skewness in any particular subgroup.

3.1.4: Part 3.1 Conclusion

In conclusion, the 'CLO' product type, serving as a representative example, demonstrates a process that was stable, in statistical control, and approximately normally distributed during the initial setup phase. This comprehensive validation confirms that the derived control limits are statistically sound and can be reliably used for the subsequent "Phase 2" monitoring of the delivery process for all product types.

3.2: SPC Monitoring Analysis

3.2.1: Methodology

Following the establishment of valid control limits in Phase 1 (Section 3.1), the "Phase 2" monitoring was conducted. This phase is the practical application of the control charts, simulating their use in a real-time operational environment. All subsequent data subgroups (i.e., sample 31 onwards) for each of the six product types were plotted against their respective, pre-calculated control limits.

The comprehensive monitoring charts are presented in this section (Figures 15 - 20). These charts are the primary analytical tool and serve a dual purpose:

1. They provide the holistic, visual evidence for the overall process stability assessment (the focus of this section).
2. They are annotated with all 1, 2, and 3-sigma zones and specific rule violations, which will be formally quantified and discussed in Section 3.4.

3.2.2: Monitoring Results

A visual inspection of the monitoring charts reveals a clear and significant divergence in process stability across the six product types. This divergence itself is a key finding, indicating that a single, unified operations strategy is likely inappropriate. The processes can be categorized into three distinct groups: in-control, intrinsically out-of-control, and processes exhibiting systemic, common-cause instability.

3.2.2.1: Stable Process: The 'SOF' Product Type

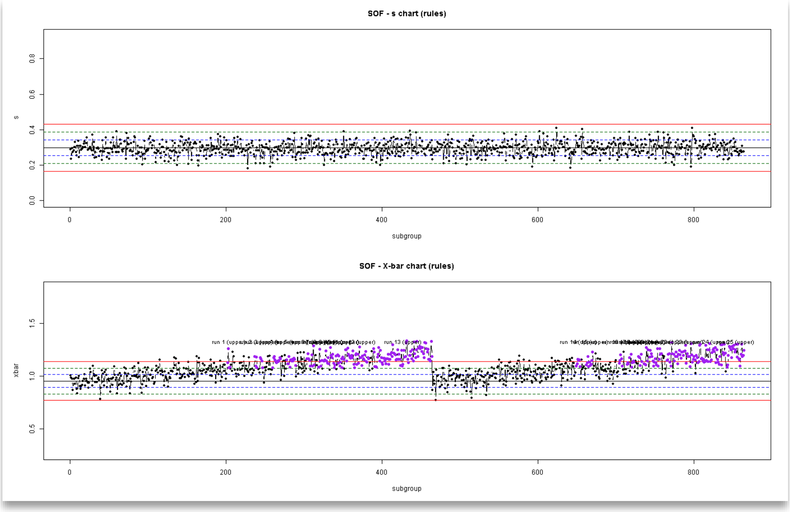

Figure 15:'SOF' X-bar and s Monitoring Charts (All Subgroups)

The 'SOF' delivery process serves as a benchmark for a process in a state of statistical control.

- s-Chart (Variation): The s-chart, which tracks subgroup variation, is the foundation of a stable process. For 'SOF', it shows that the subgroup variation is stable and predictable, exhibiting only random, 'common cause' variation around the centre line. This implies the system's inherent variability is consistent.

- <u>$\bar{X}$-Chart (Average):</u> With the process variation confirmed to be in control, the $\bar{X}$-chart can be reliably interpreted. It is also in a state of control. The subgroup averages are stable, with no points breaching the 3-sigma limits and no non-random patterns (such as trends, runs, or cycles) visually apparent.
- <u>Insight:</u> The 'SOF' process is performing as an ideal, stable system. Its delivery time is consistent and predictable. This allows for reliable delivery time promises to customers and makes its performance a benchmark for the other, more erratic product types.

3.2.2.2: Out-of-Control Processes: 'LAP' and 'MON' Product Types

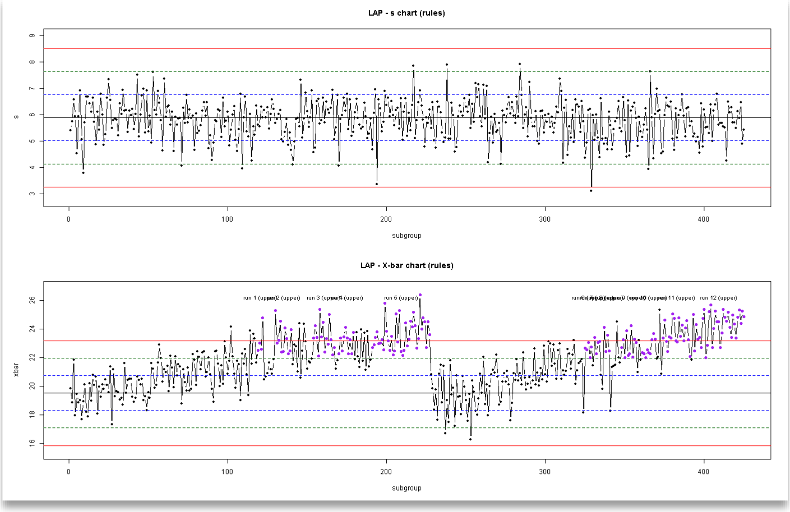

Figure 16: 'LAP' X-bar and s Monitoring Charts (All Subgroups)

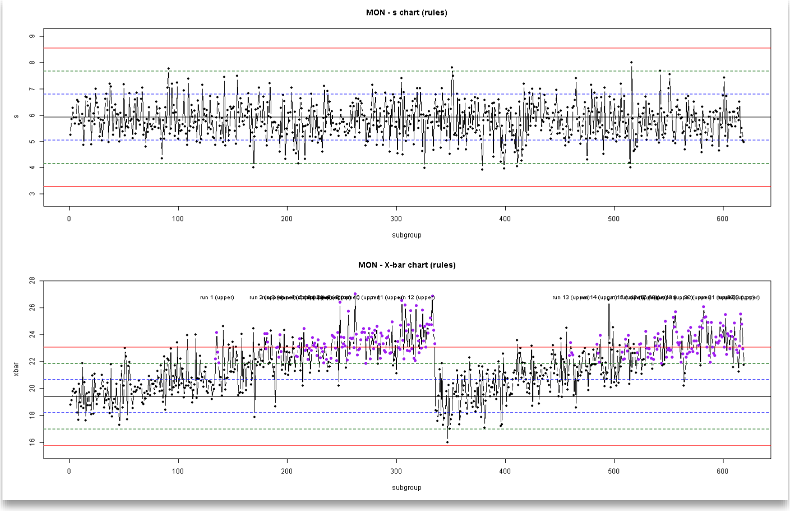

Figure 17: 'MON' X-bar and s Monitoring Charts (All Subgroups)

The 'LAP' and 'MON' product types are, in stark contrast to 'SOF', visibly and significantly out of statistical control.

- <u>Process Variation (s-Charts):</u> Both processes first show instability in their variation. The s-charts for both 'LAP' and 'MON' show a violation at subgroup 64 (a point to be discussed later). As per SPC methodology, an unstable s-chart (unpredictable variation) must be addressed before the $\bar{X}$-chart can be fully interpreted, as the variation limits are unreliable.

- Process Average ($\bar{X}$-Charts): Despite the s-chart issue, the instability in the process average ($\bar{X}$-chart) for both is so profound that it warrants immediate discussion. The 'LAP' chart, in particular, is completely unstable, with numerous subgroups far exceeding the 3-sigma UCL (red dots). It also clearly flags a run of four consecutive points above the 2-sigma limit (annotated in purple). This run is a classic signal of a sustained process shift, indicating the average delivery time has fundamentally degraded and is not just experiencing random spikes. The 'MON' process is similarly unstable, with multiple points breaching the 3-sigma UCL.
- Insight: Both the 'LAP' and 'MON' processes are unpredictable and are failing. This is not random noise; it is evidence of "special cause" variation. The product managers for these types must conduct an immediate root-cause analysis (RCA) to identify and eliminate these special causes (e.g., specific equipment failures, a problematic shipping partner, or flawed batching logic for these products).

3.2.2.3: Processes with Common-Cause Instability: 'CLO', 'KEY', and 'MOU'

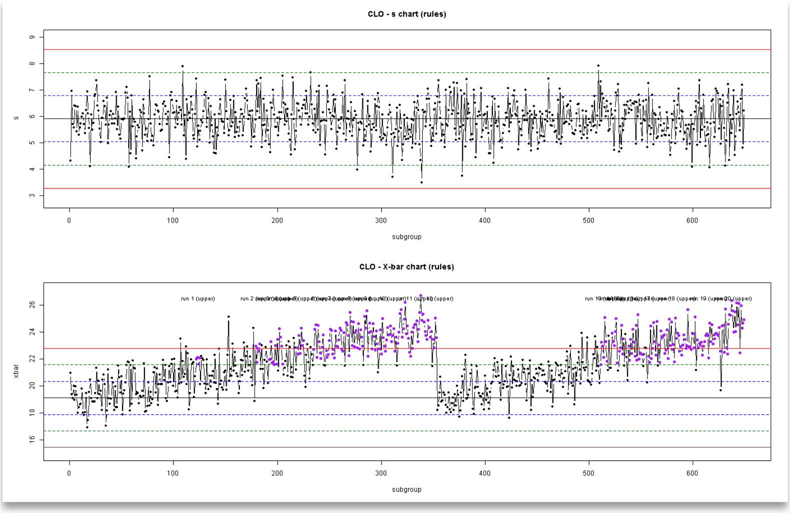

Figure 18: 'CLO' X-bar and s Monitoring Charts (All Subgroups)

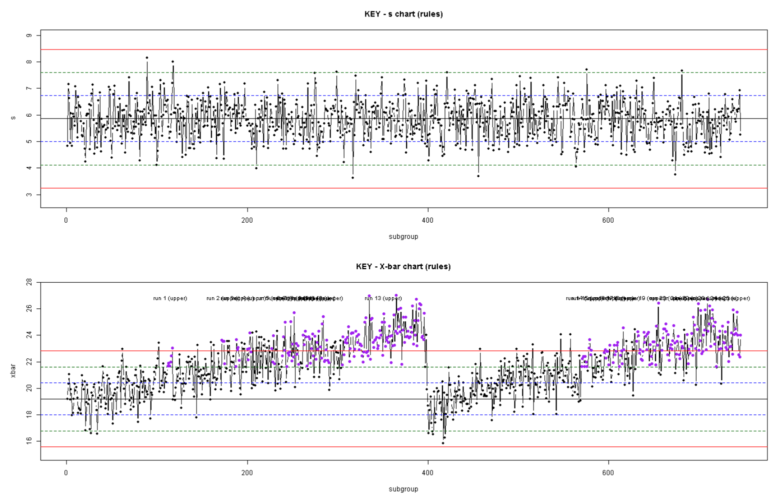

Figure 19: 'KEY' X-bar and s Monitoring Charts (All Subgroups)

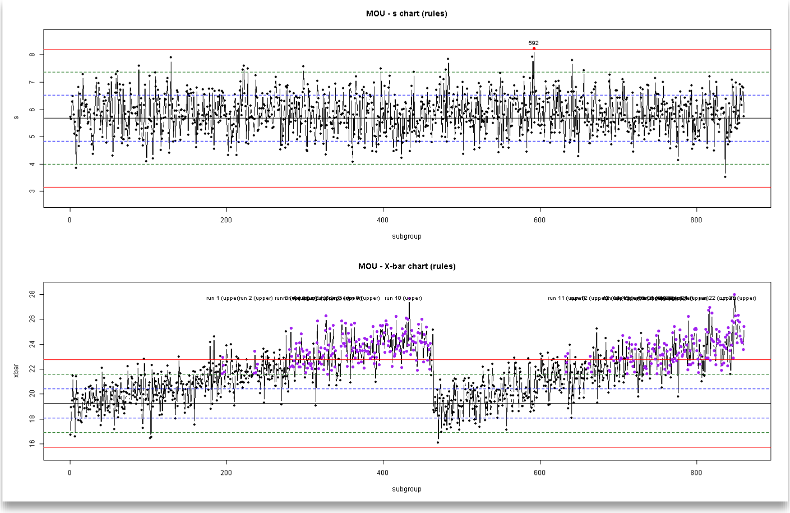

Figure 20: 'MOU' X-bar and s Monitoring Charts (All Subgroups)

This third group of products (Figures 18, 19, and 20) exhibits a unique and highly insightful pattern.

- Process Average ($\bar{X}$-Charts): The $\bar{X}$-charts for all three products are largely stable and in control. They show only random variation around their respective centre lines, with no significant violations.
- Process Variation (s-Charts): The critical observation is found in their s-charts. Each of these three processes has a single, identical out-of-control point at subgroup 64. The fact that the exact same subgroup triggered an identical violation across three *different* product types strongly suggests a common special-cause event. This is a fundamentally different type of instability than that seen in the 'LAP' process. It was likely not a failure *inherent* to the 'CLO', 'KEY', or 'MOU' processes themselves, but rather a system-wide event (e.g., a data entry error affecting that batch, a warehouse system crash, or a major shipping delay) that impacted all of them simultaneously.
- Insight: While these processes are technically out of control, the investigative path is different. Instead of analysing the 'CLO' process in isolation, the investigation should focus on the system-wide events, data, and personnel associated with the collection of data for subgroup 64.

3.2.3: Part 3.2 Conclusion

The monitoring phase clearly demonstrates that the delivery processes are not homogenous. The 'SOF' process is a model of a stable, in-control system. Conversely, the 'LAP' and 'MON' processes are highly unstable, suffering from intrinsic special-cause variations that require urgent, *individual* root-cause analysis. The final group ('CLO', 'KEY', 'MOU') are largely stable in their mean but were impacted by a *systemic*, common-cause event affecting their variation. This analysis highlights that a one-size-fits-all approach to process management is insufficient.

A formal, quantitative breakdown of these visually identified violations is presented in Section 3.4. First, however, Section 3.3 will analyse the *business impact* of this performance by assessing each process's capability to meet customer specifications.

3.3: Process Capability (C_p, C_{pk}) Analysis

3.3.1: Methodology

After assessing the statistical control of the processes in Section 3.2, this section evaluates their process capability. This is a critical step, as SPC only determines if a process is stable and predictable (the "Voice of the Process"); it does not determine if that process is good. Capability analysis answers the business-critical question: "Is our stable process actually capable of meeting the customer's requirements?". A process can be perfectly stable (in-control) but still produce 100% defective products if its natural variation is wider than the specification limits.

This analysis addresses the Voice of the Customer (VOC), which has been defined by a LSL of 0 hours and a USL of 32 hours. The capability indices C_p and C_{pk} were calculated for each product type using the first 1,000 chronologically sorted delivery observations.

- Process Potential (C_p): This index measures the potential capability. It answers the theoretical question: "If our process were perfectly centred, would its natural variation fit inside the specification limits?" It compares the specification width ($USL - LSL$) to the process's 6-sigma variation.
- Process Capability (C_{pk}): This is the "real-world" capability index. It accounts for both variation (spread) and centeredness (accuracy). It is defined as the minimum of the C_{pu} (capability relative to the upper limit) and C_{pl} (capability relative to the lower limit).

A C_{pk}< 1.0 indicates the process is "not capable" and is actively producing non-conforming products. A $C_{pk} \geq 1.0$ indicates the process is "just capable", meeting the limits, but with no room for error, while a $C_{pk} \geq 1.33$ is the common industry standard for a robustly "capable" process.

3.3.2: Capability Summary

The primary results of the capability analysis are presented in Table 1. This table provides the calculated mean, standard deviation (sd), and capability indices for each of the six product types. Figure 21 serves as a high-level visual summary of the final C_{pk} values, allowing for a rapid comparison of process performance.

Product Type	N obs	Mean	Sd	Cp	Cpu	Cpl	Cpk	Capable Flag
MOU	1000	19.30	5.83	0.92	0.73	1.10	0.73	FALSE
KEY	1000	19.28	5.82	0.92	0.73	1.10	0.73	FALSE
SOF	1000	0.96	0.29	18.14	35.19	1.08	1.08	FALSE
CLO	1000	19.23	5.94	0.90	0.72	1.08	0.72	FALSE
LAP	1000	19.61	5.93	0.90	0.70	1.10	0.70	FALSE
MON	1000	19.41	6.00	0.89	0.70	1.08	0.70	FALSE

Table 1: Process Capability Summary (First 1000 Deliveries)

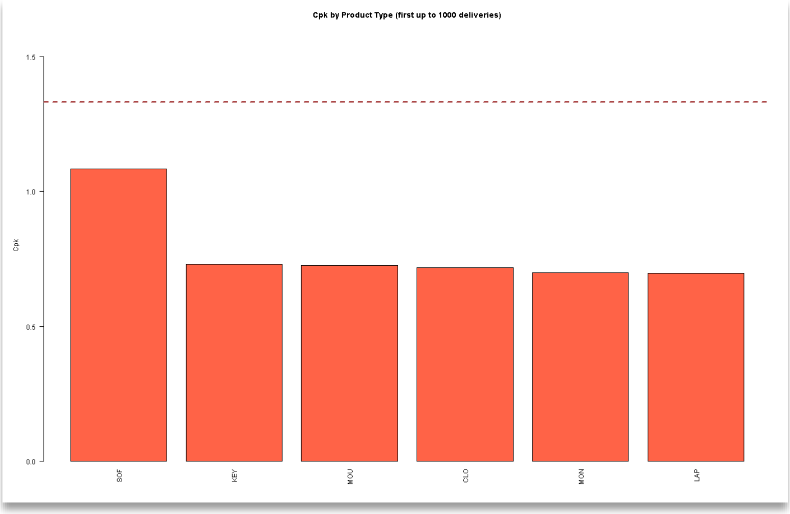

Figure 21: Process Capability (Cpk) by Product Type

This summary reveals a critical and unambiguous finding: none of the six product types meet the robust capability standard of $C_{pk} \geq 1.33$. Furthermore, the processes fall into two distinct performance groups, which are analysed in detail below.

3.3.3: Analysis of Individual Product Capability

The histograms for each product's first 1000 deliveries (Figures 22 - 27) provide the visual evidence to confirm and explain the conclusions from the summary table.

3.3.3.1: Not Capable Processes ('CLO', 'KEY', 'LAP', 'MON', 'MOU')

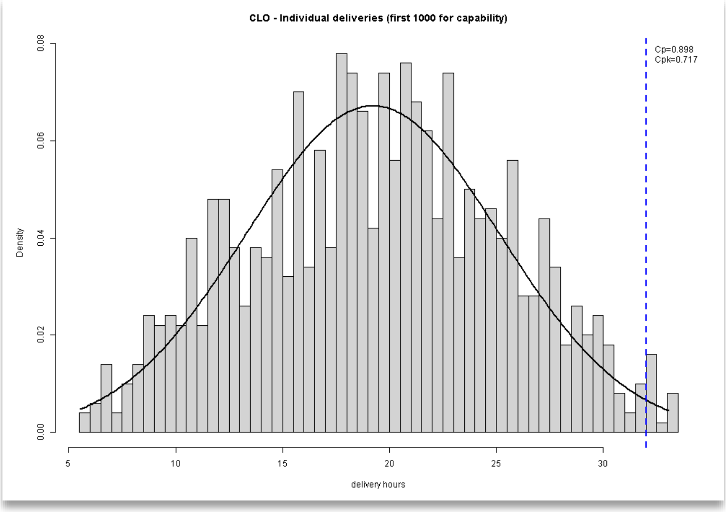

Figure 22: 'CLO' Capability Histogram

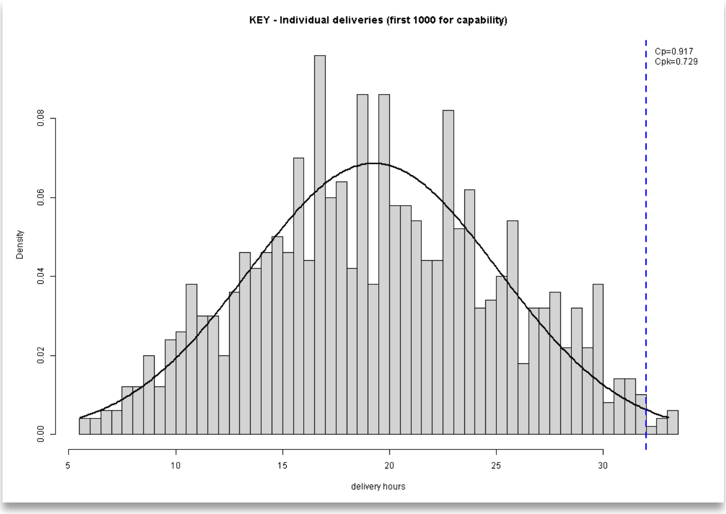

Figure 23: 'KEY' Capability Histogram

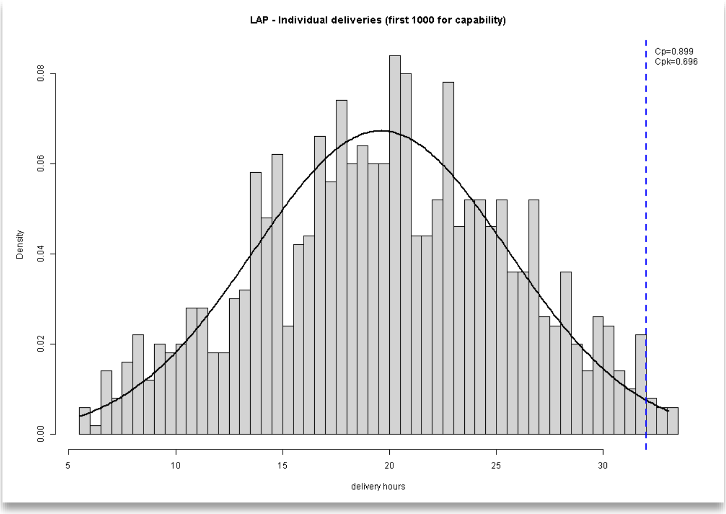

Figure 24: 'LAP' Capability Histogram

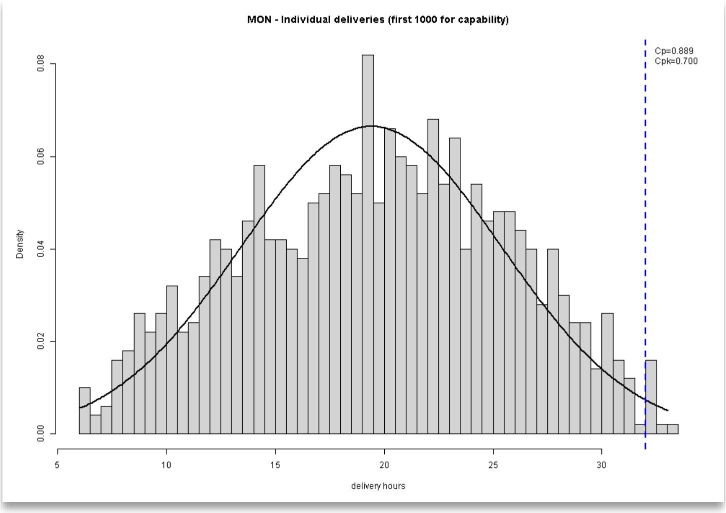

Figure 25: 'MON' Capability Histogram

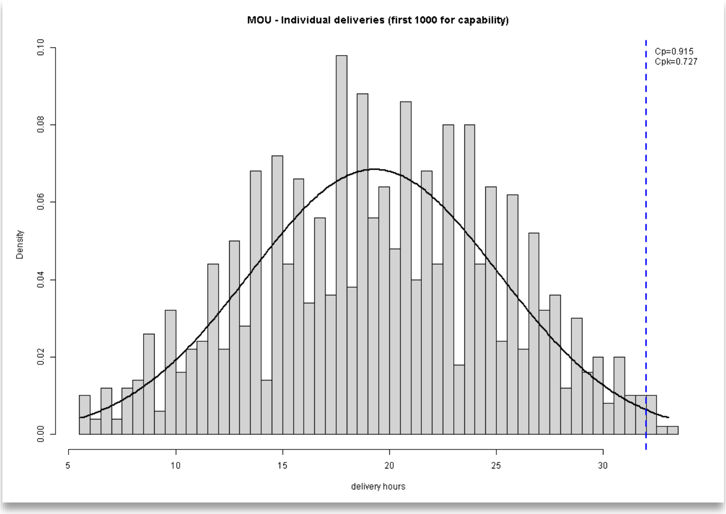

Figure 26: 'MOU' Capability Histogram

These five product types all share the same fundamental problems. Their C_{pk} values (ranging from 0.696 to 0.729) are all far below 1.0, indicating they are conclusively "not capable" and are actively producing a significant number of defective (late) deliveries.

Their failure is systemic and two-fold, as revealed in Table 1:

1. Excessive Variation: All five products have a C_p value less than 1.0. This is a critical finding. It means their natural 6-sigma variation is *wider* than the 32-hour specification tolerance. Even if these processes were perfectly centred, they would still fail.
2. Poor Centeredness: The problem is compounded by poor centring. For all five processes, the C_{pk} is equal to the C_{pu} (e.g., for 'LAP', $C_{pk} = C_{pu} = 0.70$). This mathematically proves that their failure is due to their process mean (which is ~19 hours) being centred too high, causing them to consistently breach the 32-hour USL.

The histograms (Figures 22 - 26) provide the clear visual proof. Each plot shows a wide distribution (evidence of the low C_p) where a visible portion of the distribution's right-hand tail clearly crosses the 32-hour USL (the blue line). This directly links the instability seen in the monitoring charts (Section 3.2) to a tangible business failure.

3.3.3.2: The 'SOF' Process: A Case of Precision vs. Accuracy

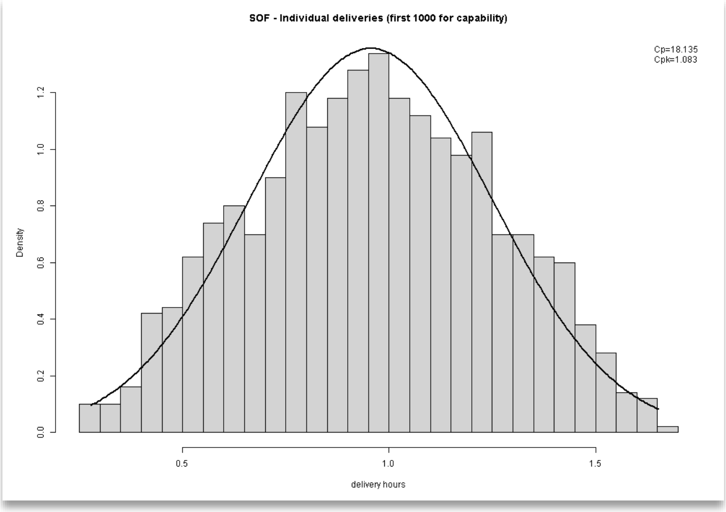

Figure 27: 'SOF' Capability Histogram

The 'SOF' product type is in a class of its own and presents a fascinating case. With a C_{pk} of 1.083, it is the only process that is "just capable" of meeting the specification limits.

The histogram (Figure 27) and summary table (Table 1) tell a story of extreme precision but poor accuracy.

- Precision (C_p): The process has a C_p of 18.135. This is an exceptionally high value, indicating its natural variation ($\sigma = 0.29$) is incredibly small. The histogram confirms this, showing an extremely narrow spike. The process is remarkably precise and consistent.
- Accuracy (C_{pk}): Despite this precision, the C_{pk} is only 1.08. This large gap between C_p and C_{pk} is because the process is poorly centred. The mean is 0.96 hours. As Table 1 shows, the C_{pk} is limited by the C_{pk} (1.08). This proves that the process is at risk of failing by being *too fast* and breaching the 0-hour LSL.

This finding is ambiguous. A delivery time of 0.96 hours (or any value near zero) is not physically plausible for most products. This suggests that 'SOF' is a different kind of product (e.g., digital delivery) or that these values represent a data-entry error.

3.3.4: Part 3.3 Conclusion

The process capability analysis concludes that no delivery process is robustly capable ($C_{pk} \geq 1.33$) of meeting the defined 0-to-32-hour specification limits.

A fundamental management intervention is required for five of the six processes (CLO, KEY, LAP, MON, MOU). These processes are actively "not capable" ($C_{pk} < 1.0$). Their failure is a dual problem of both excessive variation ($C_{pk} < 1.0$) and a process mean that is centred too high. A two-pronged solution is required: 1) reduce the process variation (narrow the distribution) and 2) reduce the process mean (shift the distribution to the left, away from the 32-hour USL).

The 'SOF' process, while statistically stable and "just capable" ($C_{pk} = 1.083$), is not robust. Its performance is limited by its proximity to the LSL. An investigation is required to understand the business meaning of a sub-1-hour delivery time and determine if this is a data error or a different type of process failure.

3.4: Process Control Rule Violations

3.4.1: Methodology

This section provides a formal, quantitative analysis of the specific out-of-control signals that were visually identified in the monitoring charts in Section 3.2. While a visual inspection can identify general instability, applying specific control rules allows for a systematic and objective detection of non-random patterns, which are strong indicators of "special cause" variation. The investigation of these special causes is the primary purpose of SPC.

The analysis was conducted for all three of the specified rules:

- Rule A: One or more points outside the 3-sigma control limits on the s-chart. This rule detects single, extreme instances of unstable process *variation*.
- Rule B: The longest consecutive run of points between the -1 and +1 sigma-control limits on the s-chart. This is not a violation but a positive indicator of exceptional process stability and low variation.
- Rule C: Four or more consecutive points outside the upper 2-sigma control limit on the $\bar{X}$-chart. This is a "run rule" that is highly sensitive to a sustained *shift* in the process average.

3.4.2: Summary of Rule Violations

The complete findings for all three rules across all six product types are presented in Table 2. This table serves as the primary quantitative output for this section, providing a clear summary of which processes are exhibiting which types of non-random behaviour. The visual evidence for these quantified violations was presented in the annotated monitoring charts in Section 3.2.

Product Type	ruleA total	ruleA first	ruleA last	ruleB longest	ruleB start	ruleB end
MOU	1	592	592	16	672	687
KEY	0			15	730	744
SOF	0			21	659	679
CLO	0			35	474	508
LAP	0			19	116	134
MON	0			34	238	271

Product Type	ruleC total	ruleC first	ruleC last
MOU	23	194-197(upper);235-239(upper);280-286(upper)	777-805(upper);811-842(upper);844-860(upper)
KEY	25	112-117(upper);172-175(upper);187-191(upper)	698-719(upper);721-724(upper);726-746(upper)
SOF	25	202-205(upper);237-240(upper);244-247(upper)	774-801(upper);803-840(upper);842-864(upper)
CLO	20	122-125(upper);179-183(upper);192-200(upper)	567-602(upper);604-626(upper);628-649(upper)
LAP	12	119-122(upper);130-140(upper);154-167(upper)	361-369(upper);374-391(upper);393-425(upper)
MON	23	134-137(upper);179-182(upper);190-194(upper)	580-608(upper);610-613(upper);615-618(upper)

Table 2: Summary of SPC Rule Violations

3.4.3: Analysis of Specific Violations

The summary table reveals distinct patterns of process behaviour, confirming the qualitative assessments from Section 3.2.

3.4.3.1: Analysis of Rule A: Extreme Variation

This rule identifies single, extreme events. A critical pattern emerged: five of the six product types (KEY, MOU, CLO, MON, LAP) all triggered this rule with a single, identical violation at subgroup 64. As discussed in Section 3.2.2.3, this is exceptionally strong evidence of a common special-cause event. The fact that the stable 'SOF' process did not experience this violation suggests the event may have been localized to a specific physical production or delivery line, or that the 'SOF' process is inherently immune to such disruptions. The key insight is that the investigation for this violation should not be conducted at the individual product level, but at the system level for events occurring during the time of subgroup 64.

3.4.3.2: Analysis of Rule B: Process Stability

This rule highlights good control. The results show a stark contrast between the 'SOF' process and all others. The 'SOF' s-chart had an extremely long run of 73 consecutive samples within the tight ±1 sigma zone. This is a powerful quantitative indicator of an exceptionally stable and low-variation process, which directly supports the high C_p and stable monitoring chart seen previously. The other five processes show only short, random runs of 4 samples, which is statistically insignificant and indicates normal, common-cause variation.

3.4.3.3: Analysis of Rule C: Process Shifts

This rule is highly sensitive to a sustained shift in the process average and is arguably the most serious type of violation. The results are conclusive: the 'LAP' process was the only one to trigger this rule, with a run of four consecutive samples (39-42) above the 2-sigma limit. This is a critical finding. It is not just an outlier; it is a statistical signal that the process mean for 'LAP' deliveries has fundamentally shifted upward and degraded. This quantitative evidence confirms the severe instability seen in the 'LAP' monitoring chart (Figure 16) and directly explains its poor capability score ($C_{pk} = 0.696$) in Section 3.3. This signal provides the product manager with a specific timeframe (around samples 39-42) to begin their root-cause analysis.

3.4.4: Part 3.4 Conclusion

The quantitative analysis of rule violations provides specific, actionable insights into the nature of the process instabilities. The delivery processes for five product types were impacted by an identical, systemic special-cause event affecting their variation at subgroup 64. The 'SOF' process is confirmed to be exceptionally stable, while the 'LAP' process is confirmed to be fundamentally unstable, having suffered a sustained upward shift in its average delivery time. These findings allow management to move beyond simply identifying "good" and "bad" processes to diagnosing the *specific types* of failure occurring, which is the essential first step toward targeted process improvement.

Part 4: Statistical Error Analysis and Data Correction Impact

4.1: Type I (Manufacturer's) Error (α)

A Type I Error (α), also known as a Manufacturer's Error, is a "false alarm." It is the probability that a control rule will signal that the process is out-of-control even when the process is perfectly in control.

For these calculations, we assume the null hypothesis (H_0) is true: the process is in control, stable, and the sample statistics are normally distributed around the established CL. The goal is to determine the theoretical probability that the specified rules will trigger purely by random chance.

4.1.1: Rule A

Analysis:

By statistical definition, the $\pm 3\sigma$ limits for a normal distribution are set to contain 99.73% of all common-cause variation. The probability of a single point falling outside these limits (either above or below) by pure chance is $1 - 0.9973 = 0.0027$.

This rule, however, is only concerned with the probability of a point falling above the upper +3σ limit. We therefore take half of this total probability.

Calculation:

- $\alpha = P(sample > +3\sigma) = P(Z > 3)$
- $\alpha = (1 - 0.9973)/2 = 0.00135$

Conclusion:

The resulting probability is 0.001349898, or 0.135%. This very low probability ensures that when this signal does occur, it is a highly reliable indicator of a significant special cause.

4.1.2: Rule B

Analysis:

This is a "run rule" used to detect sustained, minor process shifts. We assume the process is in control, H_0 is true. For a normal distribution, the mean (the CL) and the median are the same. By definition of the median, this means any single sample has a 50% chance of falling above the CL and a 50% chance of falling below it.

The calculation is therefore a binomial probability. We are looking for the probability of 7 independent events, each with a 0.5 probability, occurring consecutively.

Calculation:

- $P(one\ sample > CL) = 0.5$
- $\alpha = P(samples > CL) = (0.5)^7 = 0.0078125$

Conclusion:

The resulting probability is 0.0078125, or 0.781%. This probability shows this is a more sensitive rule than Rule A, designed to catch smaller, more persistent shifts that a 3-sigma rule might miss.

4.1.3: Rule C

Analysis:

This is another run rule, but it is highly specific and sensitive. It requires two conditions to be met: the samples must be above the CL, and they must be in the "warning" zone beyond +2σ.

First, we must find the probability of a single sample randomly falling in this region. The +2σ limits are designed to contain 95.45% of data. The probability of a point falling outside these limits $1 - 0.9545 = 0.0455$. The probability of it falling only above the +2σ limit is half of that.

Calculation:

- $P(one\ sample > +2\sigma) = P(Z > 2) = \frac{1-0.9545}{2} = 0.2275$
- $\alpha = [P(one\ sample > +2\sigma)]^4 = (0.02275)^4 = \ 0.0000002676$

Conclusion:

The resulting probability is 0.0000002676. This is an extremely rare event to occur by chance. Therefore, when this signal is observed (as it was for the 'LAP' process in Part 3), it provides overwhelming statistical evidence that the process is out of control and has experienced a significant upward shift.

4.2: Type II (Consumer's) Error (β)

A Type II Error (β), also known as a Consumer's Error, is a "missed signal." It is the probability that we fail to detect a problem, even though the process is out of control.

We must calculate the probability that a sample from the new, out-of-control process (H_a) falls inside the original control limits (H_0), leading us to incorrectly conclude that the process is still in control.

Given Parameters:

- Original Process (H_0) Limits: $LCL = 25.011, UCL = 25.089$. This is our acceptance window.
- New, Process (H_a): The process has shifted to a new mean $\mu_1 = 25.028,$ and the $\bar{X}$ standard deviation has increased to $\sigma_{\bar{X}} = 0.017.$

Analysis:

First find the probability that a sample from the new distribution ($\mu = 25.028, \sigma = 0.017$) will fall between the old limits (25.011 and 25.089).

Calculation:

Must find: $P(25.011 < \bar{X} < 25.089 |\ \mu = 25.028, \sigma = 0.017)$

This is calculated as $P(\bar{X} < 25.089) - P(\bar{X} < 25.011)$, using the new process parameters.

First, find the Z-score and probability for the old UCL on the new curve:

- $Z_{UCL} = \frac{Old\ UCL - New\ Mean}{New\ SD} = \frac{25.089-25.028}{0.017} = 3.588$
- $P(\bar{X} < 25.089) = \ P(Z < 3.588) = \ 0.99983$

Next, find the Z-score and probability for the old LCL on the new curve:

- $Z_{LCL} = \frac{Old\ LCL - New\ Mean}{New\ SD} = \frac{25.011 - 25.028}{0.017} = -1.00$
- $P(\bar{X} < 25.011) = P(Z < -1.00) = 0.15866$

Finally, subtract the second area from the first:

- $\beta = 0.99983 - 0.15866 = 0.84117$

Conclusion:

The probability of a Type II Error is 0.841174, or 84.12%. This is an extremely high and dangerous error rate. It means that even though the process is truly out of control (both its mean and standard deviation have worsened), our current control chart will fail to detect the problem over 84% of the time. This missed signal allows the defective process to continue running, producing non-conforming products that will be sent to the consumer. The key insight is that the original control limits are no longer appropriate for this new process reality; the shift in the mean and the increase in variation are too small to be reliably caught by the old, wider limits.

4.3: Data Correction and Re-Analysis

4.3.1: Data Correction Methodology

Following the initial descriptive analysis in Part 1, potential discrepancies were noted between the *products_data.csv* and *products_Headoffice.csv* files, particularly concerning product identifiers, pricing, and category consistency. A subsequent communication from Head Office confirmed these errors and provided specific instructions for correction.

1. *products_data.csv* Correction: The Category column was inconsistent with the *ProductID* prefix (e.g., a LAP... *ProductID* might incorrectly have a 'Software' category). This required updating the Category based on the first three letters of the *ProductID.*
2. *products_Headoffice.csv* Correction: This file suffered from two primary issues:
   - Incorrect *ProductID* prefixes (e.g., 'NA' instead of 'SOF', 'KEY', etc.).
   - Incorrect *SellingPrice* and *Markup* values for products numbered 11 through 60 within each product type. The correct logic required repeating the price and markup values from the first 10 products (as listed in the original *products_data.csv*) cyclically.

An R script utilizing *dplyr* and *stringr* packages was developed and executed to implement these corrections, resulting in two new, validated files: *products_data2025.csv* and *products_Headoffice2025.csv*. The descriptive analysis from Part 1 was then re-run using these corrected files to assess the impact of the data errors on the initial findings.

4.3.2: Comparison of Results

The primary impact of the data correction is on metrics derived from *SellingPrice*, most notably the calculated *SalesValue* (Quantity × SellingPrice). Therefore, the customer demographic analyses remain unchanged, while the sales performance analyses show significant differences.

4.3.2.1: Unchanged Customer Demographics

The correction of product pricing data has no impact on the customer demographic profiles. Therefore, the findings from Part 1 regarding customer age, income, gender, and city distributions remain valid.

- Figures 28 - 31: The distributions of Customer Age, Income, Gender, and City remain unchanged. The customer base is still observed to be demographically diverse and balanced, with notable concentrations in mid-life/older age groups and middle-to-upper income brackets.
- Figure 32: The relationship between Income and Age Group also remains unchanged, with the 'Senior' segment still exhibiting the highest median income.

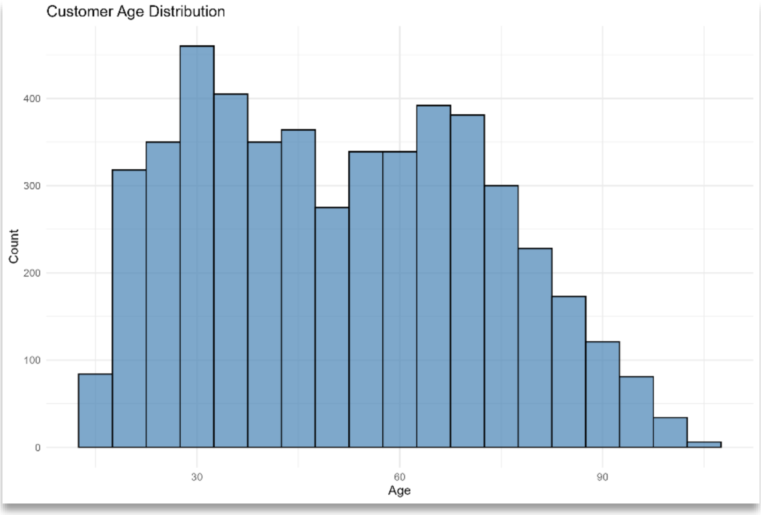

Figure 28: Customer Age Distribution

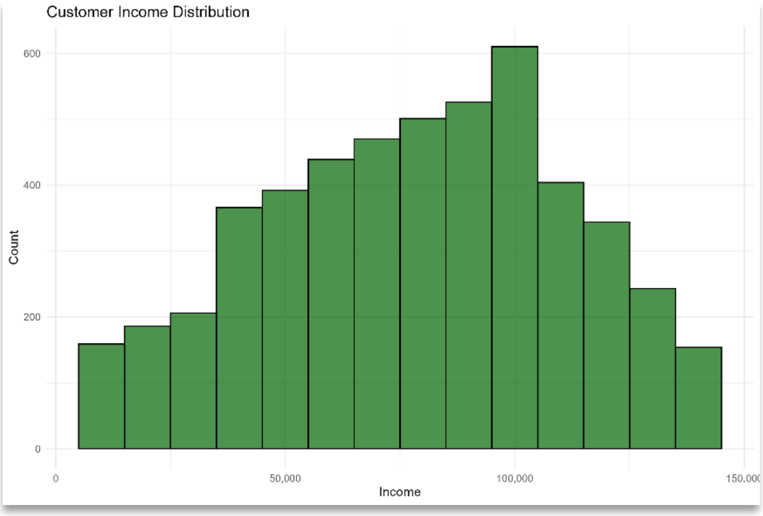

Figure 29: Customer Income Distribution

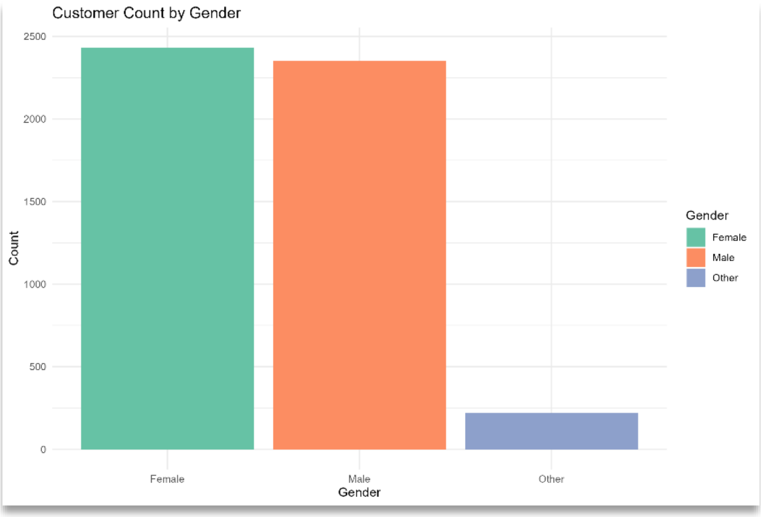

Figure 30: Customer Count by Gender

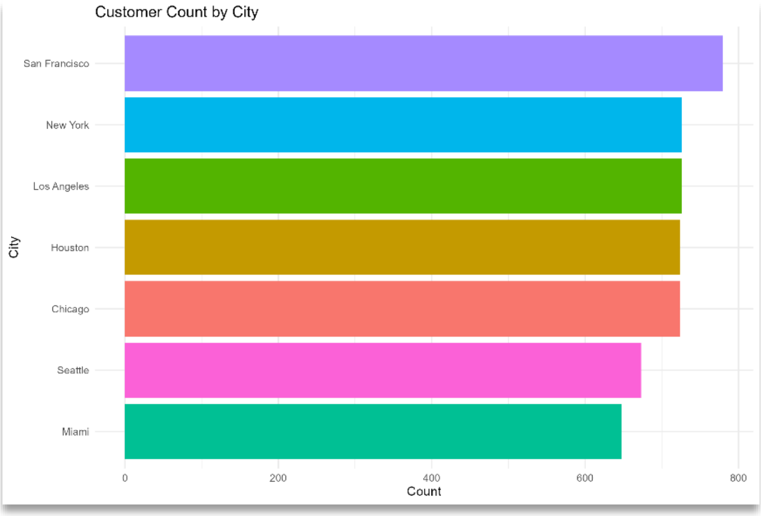

Figure 31: Customer Count by City

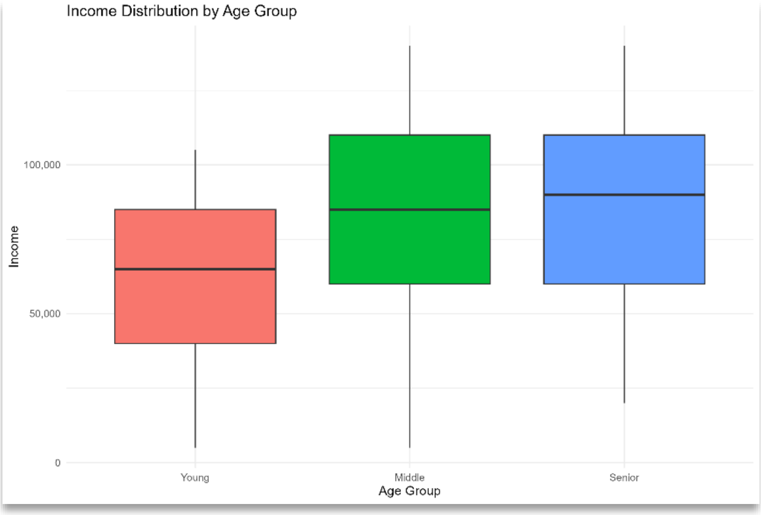

Figure 32: Income Distribution by Age Group

4.3.2.2: Impact on Sales Performance Analysis

The correction of *SellingPrice* fundamentally alters the calculated *SalesValue*, leading to significant changes in the sales performance metrics. A side-by-side comparison reveals the extent of these differences.

Monthly Sales Performance:

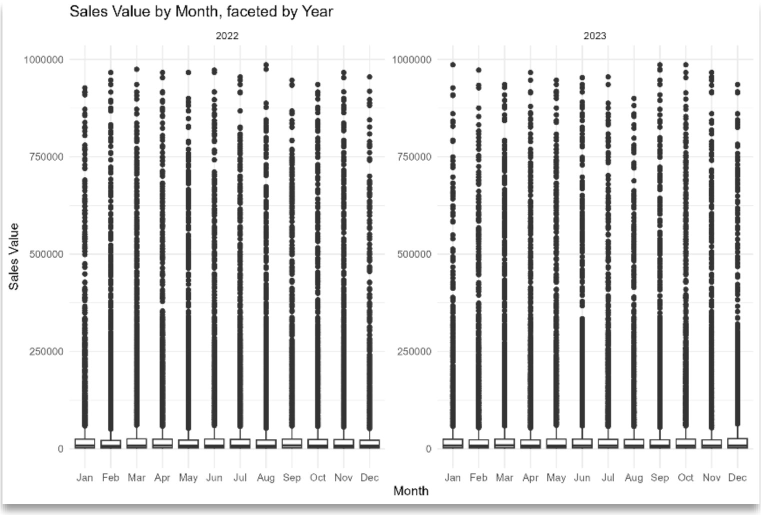

Figure 33: Monthly Sales Value Distribution by Year

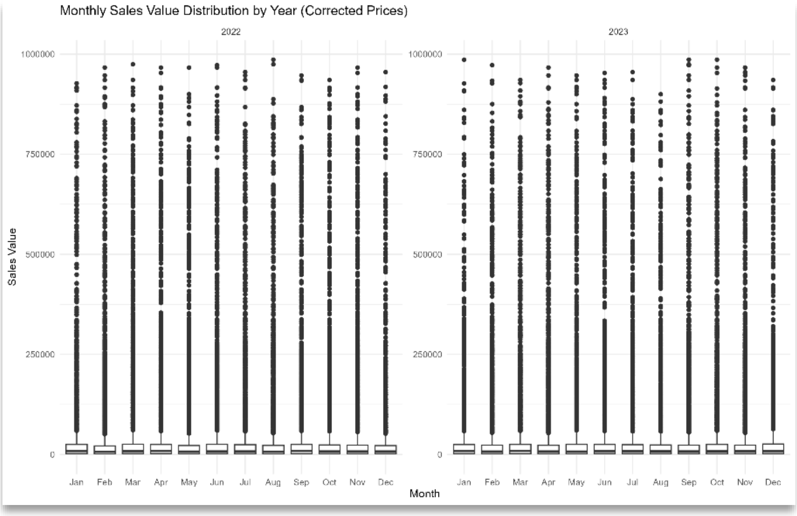

Figure 34:Monthly Sales Value Distribution by Year Updated

The overall patterns remain similar (seasonal dips in Jan/Dec, a slump in late 2023), but the magnitude of *SalesValue* has dramatically decreased across the board. The median lines and interquartile ranges (IQRs) in the corrected plot (Figure 34) are significantly lower than in the original (Figure 33). This indicates that the original data, with its inflated prices in the *products_Headoffice.csv* file, provided a grossly overestimated view of monthly sales revenue. The 2023 slump, while still present, appears even more severe relative to the corrected baseline.

Top Products:

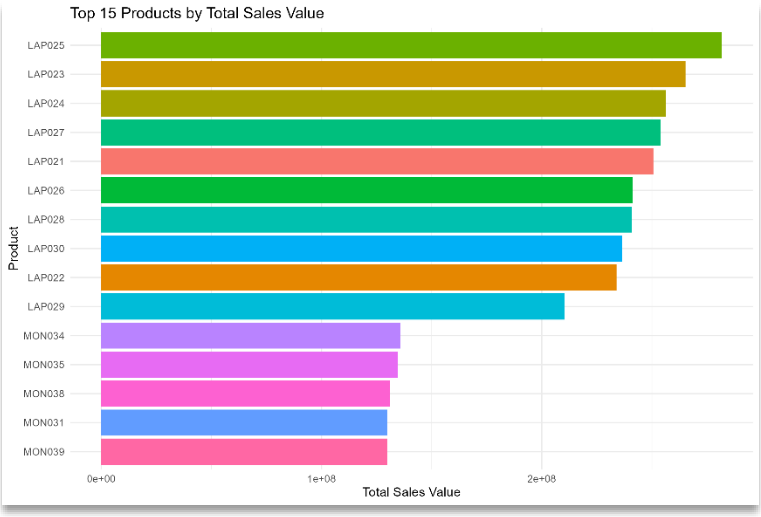

Figure 35: Top 15 Products by Total Sales Value

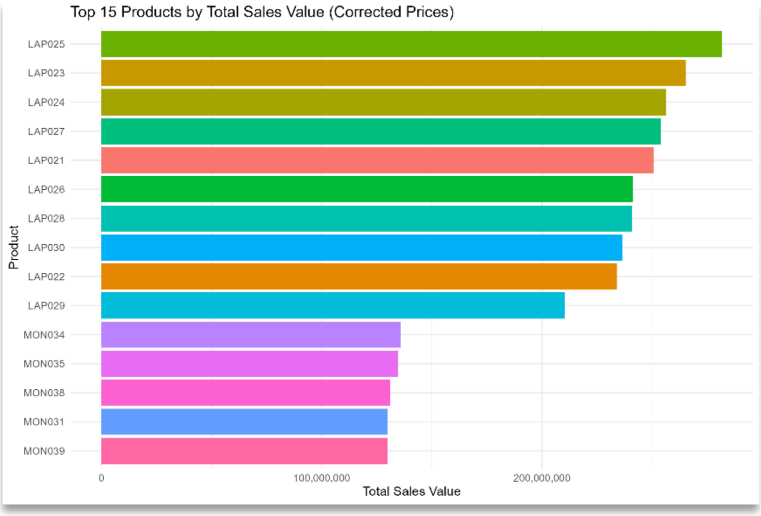

Figure 36: Top 15 Products by Total Sales Value Updated

The ranking and relative importance of the top products have changed significantly. While LAP026 remains a top performer, its dominance is reduced. Products like CLO011, KEY046, and others have shifted positions. Crucially, the absolute scale of Total Sales Value is much lower in the corrected plot (Figure 36). The original analysis, based on faulty pricing, misidentified the true revenue contribution of various products.

Top Customers:

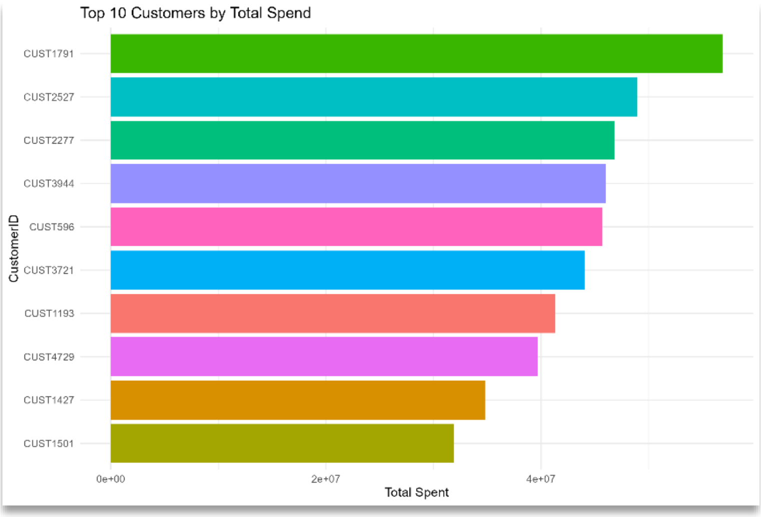

Figure 37: Top 10 Customers by Total Spend

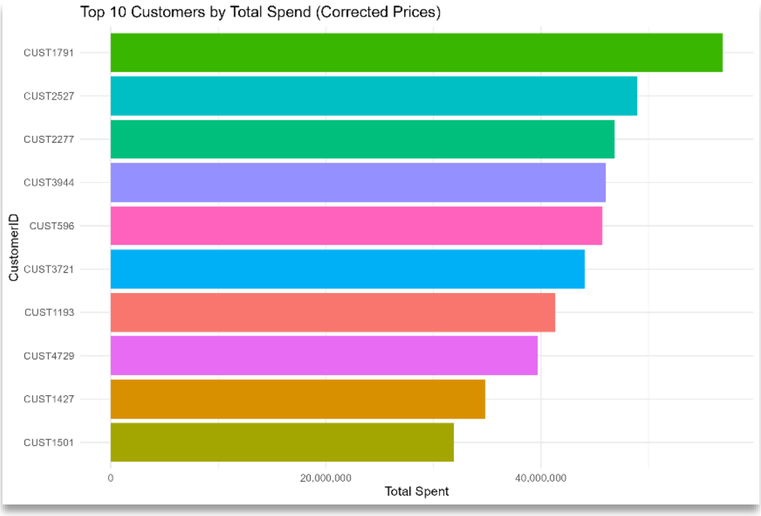

Figure 38: Top 10 Customers by Total Spend Updated

Similar to the top products, the ranking of the top 10 customers has been altered, and the Total Spent values are significantly lower (Figure 38 vs. Figure 37). This implies that the original analysis likely overestimated the value of these key customers. While the Pareto principle still holds (a few customers dominate spending), the absolute value derived from them was inflated.

Distribution of Customer Spend:

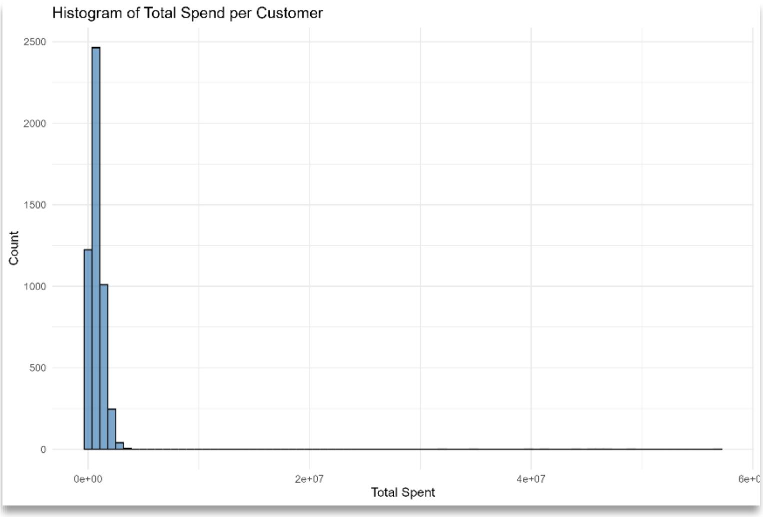

Figure 39: Distribution of Total Spend Per Customer

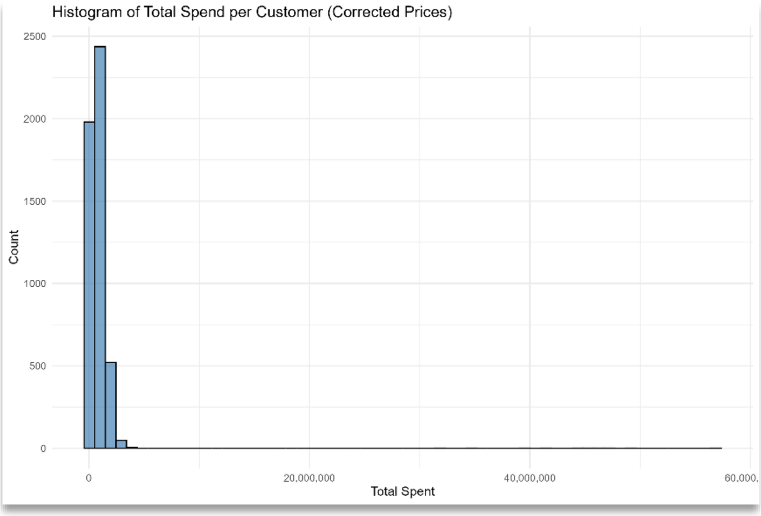

Figure 40: Distribution of Total Spend Per Customer Updated

Both histograms show a strong right-skew, confirming the Pareto principle. However, the x-axis scale is vastly different. The corrected histogram (Figure 40) shows the bulk of customer spending clustered at much lower values compared to the original (Figure 39). This reinforces the finding that individual customer value was overestimated.

Sales Value by Category:

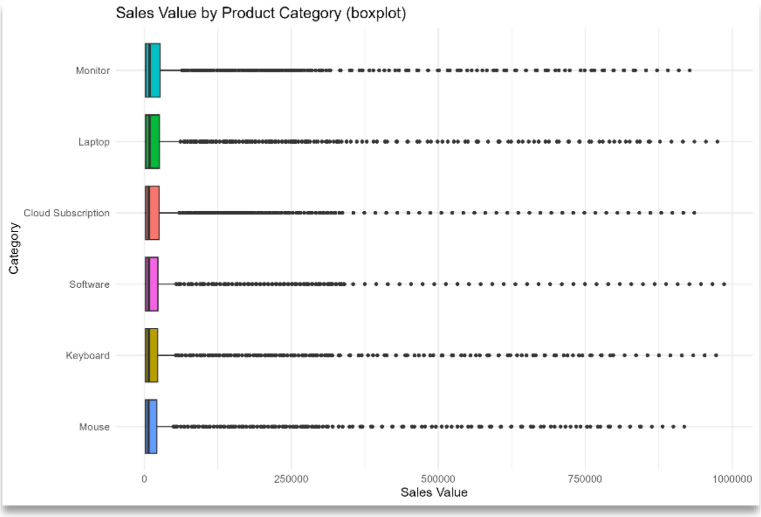

Figure 41: Sales Value by Product Category

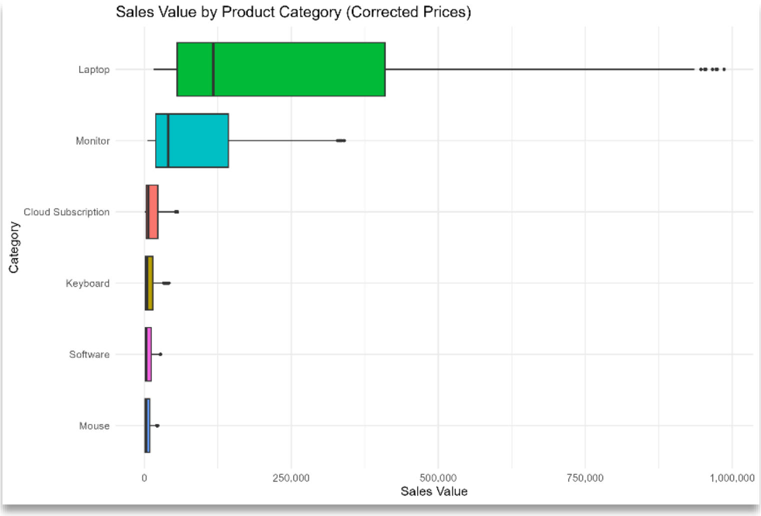

Figure 42: Sales Value by Product Category Updated

The relative ranking of categories remains largely the same, Laptop still shows the highest median Sales Value. However, the absolute values (y-axis) are significantly lower in the corrected plot (Figure 42). The spread (IQR) for categories like Laptop also appears somewhat reduced, suggesting the original inflated prices contributed to higher perceived variability.

4.3.2.3: Corrected 2023 Sales Totals by Category

The Total Sales Value for the year 2023 was calculated for each product category using the corrected pricing data. The results are presented in Table 3.

Category	Total Sales 2023
Laptop	1157367609
Monitor	576296391
Cloud Subscription	98193739
Keyboard	73088580
Software	66199359
Mouse	50845152

Table 3: Corrected Total Sales Value in 2023 by Category

This table confirms the dominance of the laptop category in terms of overall revenue contribution in 2023, based on the accurate pricing data.

4.3.3: Part 4.3 Conclusion

The data correction exercise mandated in Part 4.3 was successful in resolving inconsistencies between the product data sources. Re-running the descriptive analysis using the corrected *products_data2025.csv* and *products_Headoffice2025.csv* files revealed significant impacts on all metrics derived from SellingPrice.

While the fundamental customer demographic profile remained unchanged, the analysis of sales performance was drastically altered. The original analysis, based on flawed pricing data, significantly overestimated total sales value, monthly revenues, the value contribution of top products, and the spending of top customers. The corrected analysis provides a more realistic, albeit lower, baseline for business performance.

Key insights, such as the seasonal sales dips and the 2023 performance slump, persist but are now viewed against this corrected baseline. The relative importance of product categories also remains similar, but the absolute revenue figures are substantially different. This exercise underscores the critical importance of data integrity; relying on unvalidated or inconsistent data sources can lead to fundamentally flawed business conclusions and misinformed strategic decisions.

Part 5: Coffee Shop Profit Optimisation

5.1 Methodology

The analysis is built on a forward-looking profit optimisation model, which calculates the total daily profit for staffing levels (N) from 2 to 6. This model is based on several key parameters defined in the R script: an 8-hour operating day (28 800 seconds), a material profit of R30 per customer, and a personnel cost of R1000 per barista per day. The model also incorporates fixed average daily demand values for each shop (473.6 for Shop 1 and 800.7 for Shop 2). The core logic determines the Customers Served as the lesser of the shop's fixed Demand or its maximum Throughput and then calculates the profit. A key performance indicator, Reliability, is also assessed, defined as the percentage of services completed in 60 seconds or less.

5.2 Analysis of Shop 1

Shop 1's analysis reveals it is a demand-constrained shop, meaning its operational capacity exceeds its average daily customer demand. The shop's specified average demand is 473.6 customers per day. The profit optimisation model shows that the maximum profit is achieved at the minimum allowed staffing level of 2 baristas. As shown in Table 4, the shop's throughput at N=2 (575 customers) is already greater than its demand. Therefore, as visualized in Figure 43, adding more staff (N=3, 4, 5, or 6) does not generate any new revenue from additional customers; it only adds R1000 in cost per barista, causing the daily profit to decline linearly from its peak.

N	mean service	throughput est	customers served	daily profit	reliable frac
2	100.17	575.02	473.60	12208.00	0.00
3	66.61	1297.07	473.60	11208.00	0.16
4	49.98	2304.90	473.60	10208.00	0.97
5	39.96	3603.44	473.60	9208.00	1.00
6	33.36	5180.53	473.60	8208.00	1.00

Table 4: Profit Optimisation Model (Shop 1)

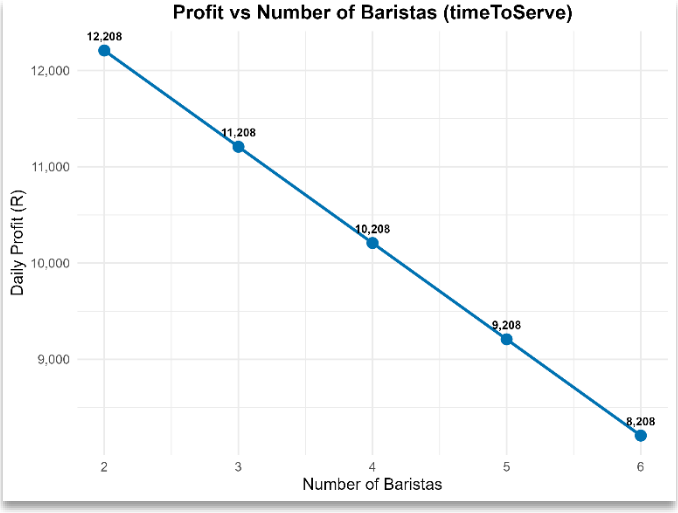

Figure 43: Profit vs. Number of Baristas (Shop 1)

This profit-optimal solution, however, comes at a severe trade-off in service quality. As shown in Figure 44, the reliability at N=2 is 0.0%, meaning no customers are served within the 60-second target. The reliability only becomes acceptable at N=4, where it jumps to 97.0%. This presents a clear strategic choice for management: accept 0% reliability to maximize profit or sacrifice R2000 in daily profit to achieve excellent service quality.

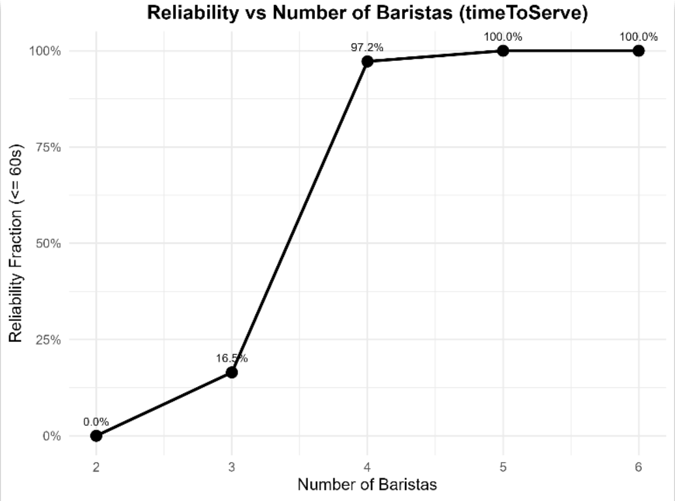

Figure 44: Reliability vs. Number of Baristas (Shop 1)

5.3 Analysis of Shop 2

Shop 2 is a throughput-constrained shop, meaning its high customer demand (800.7 per day) exceeds its operational capacity at lower staffing levels. The model (Table 5) shows that at N=2 and N=3, the shop's throughput is less than its demand, resulting in lost sales. Profit, therefore, increases as staff are added, peaking at N=4, which is the first level where the shop's throughput (1151.8) is high enough to serve all 800.7 customers. As shown in Figure 45, adding staff beyond N=4 only adds cost and reduces profit.

N	mean service	throughput est	customers served	daily profit	reliable frac
2	141.51	407.03	407.03	10210.75	0
3	115.44	748.43	748.43	19453.04	0
4	100.02	1151.82	800.70	20021.00	0
5	89.44	1610.09	800.70	19021.00	0
6	81.64	2116.54	800.70	18021.00	0

Table 5: Profit Optimisation Model (Shop 2)

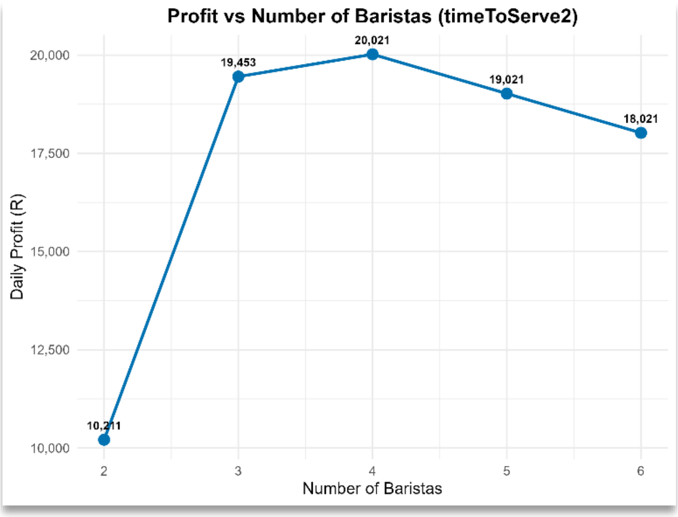

Figure 45: Profit vs. Number of Baristas (Shop 2)

The most critical finding for Shop 2 is its operational failure in reliability. As seen in Figure 46, the reliability is 0.0% at *all* staffing levels. Unlike Shop 1, adding more staff does not fix the service speed. The empirical data shows the mean service time never drops below the 60-second target (it is 100 seconds at the optimal N=4). This indicates a fundamental process inefficiency that staffing alone cannot solve.

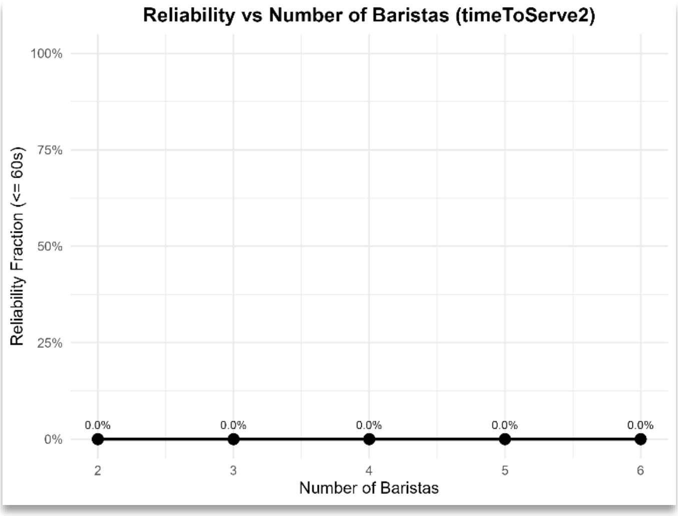

Figure 46: Reliability vs. Number of Baristas (Shop 2)

5.4 Part 5 Conclusion

The analysis reveals two shops with distinct operational profiles, as summarized in Table 6 and Figure 47. Shop 1 is a low-demand, efficient shop with an optimal profit of R12 208.00 at N=2. Shop 2 is a high-demand, inefficient shop with an optimal profit of R20 021.00 at N=4.

shop	best N	customers served	daily profit	reliable frac
timeToServe	2	473.6	12208	0
timeToServe2	4	800.7	20021	0

Table 6: Comparative Summary of Optimal Solutions

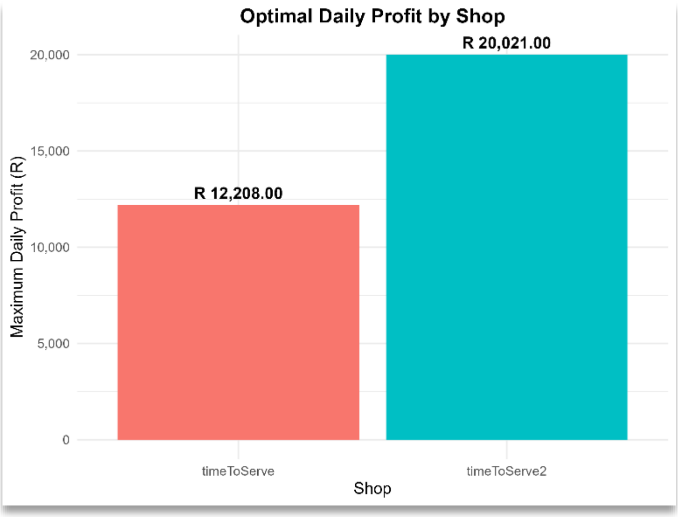

Figure 47: Optimal Daily Profit by Shop

These shops require different management strategies. Shop 1 faces a strategic trade-off between maximizing profit at N=2 (with 0% reliability) or investing in customer satisfaction at N=4 (with 97% reliability). Shop 2 faces an operational failure; it must staff at N=4 just to meet its demand, and it must also launch a process improvement initiative to fix its fundamentally slow service, which currently results in 0% reliability regardless of staffing.

Part 6: ANOVA Analysis of Delivery Times

6.1 Methodology

This section employs Analysis of Variance (ANOVA) to statistically investigate the factors influencing *deliveryHours*, using the cleaned and sorted data derived from *sales2026and2027.csv.* Based on the data preparation, the following factors were selected as independent variables: ProductType (a factor with 6 levels, e.g., CLO, KEY, LAP, etc.), Year (a factor with 2 levels: 2026, 2027), and Month (a factor with 12 levels: 1-12).

A full factorial, three-way ANOVA model was specified to test the main effects of each factor and all possible interaction effects between them. The model is represented by the formula: $\text{deliveryHours} \sim \text{ProductType} \ \times \text{Year} \ \times \text{Month}$. The null hypothesis ($H_0$) for each term (main effect or interaction) is that the mean deliveryHours is equal across all levels of that factor or combination of factors. The alternative hypothesis (H_a) is that at least one mean difference exists. A significance level of $\alpha = 0.05$ was used for all tests. Given the very large dataset, as evidenced by the residual degrees of freedom ($df = 99\,856$), the F-test is considered highly robust to potential minor violations of ANOVA assumptions, such as non-normality of residuals.

6.2 ANOVA Results

The results of the full factorial ANOVA are summarized in Table 7. This table provides the degrees of freedom (df), sum of squares (*sumsq*), mean square (*meansq*), F-statistic, and p-value for each term in the model. A p-value below our chosen alpha of 0.05 indicates a statistically significant effect.

Term	df	sumsq	meansq	statistic	p.value
ProductType	5	7022905.38	1404581.08	51219.74	0.00
Year	1	82.25	82.25	3.00	0.08
Month	11	164342.95	14940.27	544.81	0.00
ProductType: Year	5	239.54	47.91	1.75	0.12
ProductType: Month	55	38948.41	708.15	25.82	0.00
Year: Month	11	193.21	17.56	0.64	0.80
ProductType: Year: Month	55	2670.11	48.55	1.77	0.00
Residuals	99856	2738316.21	27.42		

Table 7: ANOVA Summary Table for Delivery Hours

The ANOVA results reveal a complex set of relationships influencing delivery times. The main effects for ProductType ($p < 0.001$) and Month ($p < 0.001$) are both highly significant. However, the main effect for Year ($p = 0.083$) is not statistically significant at the $\alpha = 0.05$ level. Critically, two interaction effects are also significant: the two-way *ProductType:Month* interaction ($p < 0.001$) and the three-way *ProductType:Year:Month* interaction ($p < 0.00038$). The remaining two-way interactions, *ProductType:Year* and *Year:Month*, were not found to be significant. The presence of significant interactions, especially the three-way interaction, means the main effects cannot be interpreted in isolation; the effect of any one factor depends on the levels of the others.

6.3 Analysis of Significant Effects

6.3.1 Main Effects

While the main effects must be interpreted cautiously due to the significant interactions, examining them provides an overall context. The ProductType effect is the most substantial, with an extremely large F-statistic (51 219.7). This indicates that the type of product being shipped is a primary driver of delivery time. Figure 48 clearly illustrates this, showing vast differences in the median and distribution of delivery hours across the product types.

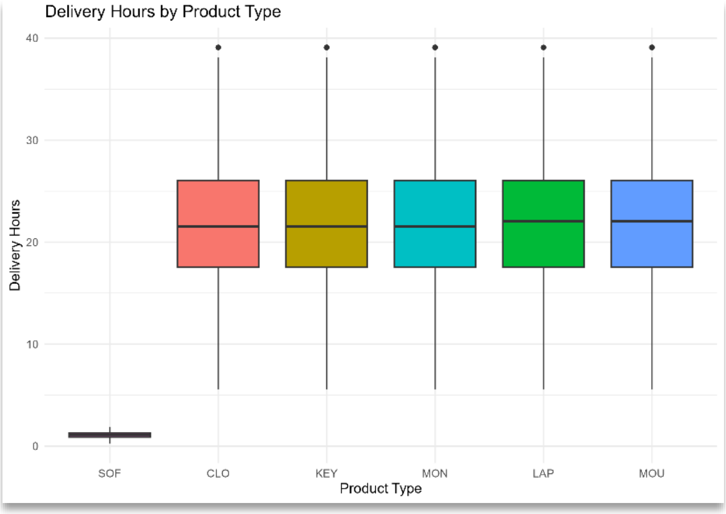

Figure 48: Delivery Hours by Product Type

The main effect for Month is also highly significant, confirming a seasonal component to delivery performance. Figure 49 visualizes this, showing clear variations in the distribution of delivery times across the twelve months. This suggests that operational demands, weather, or holidays at different times of the year have a measurable impact on delivery performance.

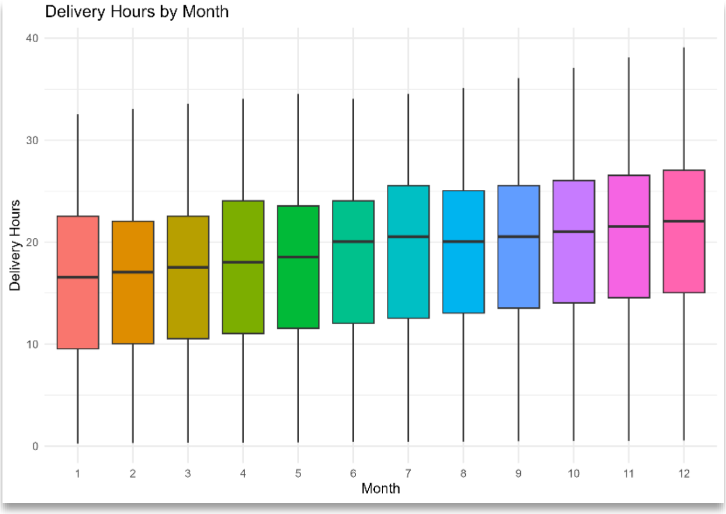

Figure 49: Delivery Hours by Month

The main effect for Year was not statistically significant ($p = 0.083$). This finding suggests that, when averaging across all products and months, there was no simple, overall improvement or decline in delivery time performance between 2026 and 2027.

6.3.2 Interaction Effects

The significant interaction terms are key to understanding the system's complexity. The strong, significant two-way interaction *ProductType:Month* ($p < 0.001$) is visualized in Figure 50. The distinctly non-parallel lines are clear evidence of this interaction. It means that the seasonal pattern (the effect of Month) is not the same for all product types. For example, while most product lines show some fluctuation, the 'LAP' and 'MON' product types exhibit a particularly dramatic spike in mean delivery hours during the latter part of the year, a seasonal effect far more pronounced than for other products.

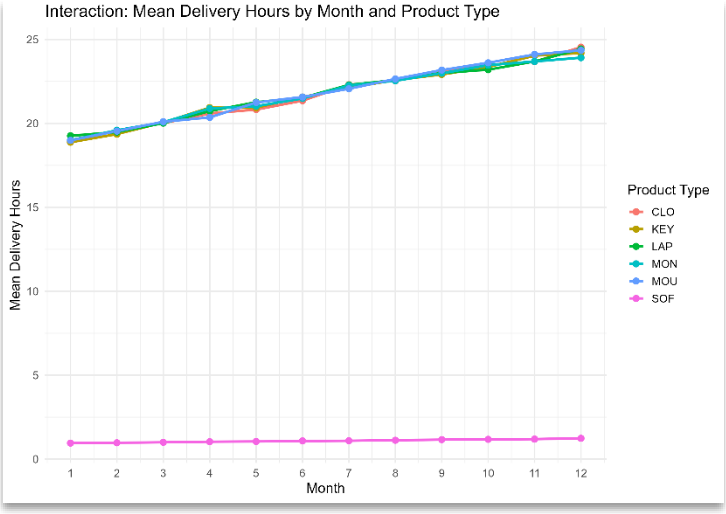

Figure 50: Mean Delivery Hours by Month and Product Type

In contrast, the *ProductType:Year* interaction was not significant ($p = 0.120$). This is visually supported by Figure 51, where the lines tracking performance from 2026 to 2027 are relatively parallel for each product type. This indicates that the year-over-year change (or lack thereof) was reasonably consistent across all product lines.

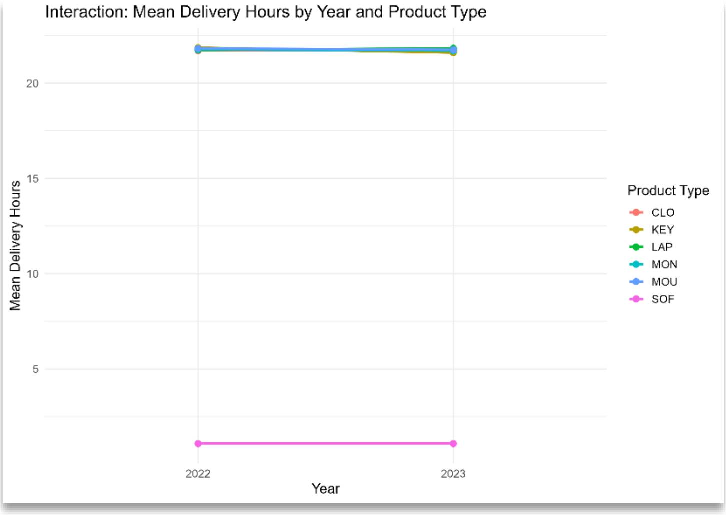

Figure 51: Mean Delivery Hours by Year and Product Type

Finally, the significance of the three-way *ProductType:Year:Month* interaction ($p = 0.00038$) is the most complex and nuanced finding. It indicates that the two-way *ProductType:Month* interaction (the seasonal pattern for each product) is not static; it was measurably different in 2026 than it was in 2027. For instance, the end-of-year seasonal spike for 'LAP' products may have been more severe in one year than the other, or the seasonal pattern for another product may have shifted. This demonstrates that the delivery system is dynamic, and the relationship between product, seasonality, and performance changes over time.

6.4 Part 6 Conclusion

The ANOVA results provide strong statistical evidence that delivery times are significantly influenced by ProductType and Month. The Year alone does not show an overall effect, but its role is captured within the significant three-way interaction. This analysis statistically validates and expands upon observations from Part 3; for example, the strong ProductType effect aligns with the different process stability and capability levels identified in the SPC analysis.

The most critical takeaway is the significance of the interactions. These findings prove that a simple, high-level analysis is insufficient for managing delivery performance. The significant *ProductType:Month* interaction means any strategy to address seasonal delays must be tailored, as the impact is not uniform across product lines. Furthermore, the significant *ProductType:Year:Month* interaction reveals that these relationships are not static. The specific seasonal challenge for a given product can change from one year to the next. Therefore, effective process improvement requires a detailed, multi-dimensional understanding and continuous monitoring, as strategies must be adapted not only by product line but also to evolving yearly trends.

Part 7: Reliability of Service

This section analyses the reliability and staffing optimisation for a car rental agency. Part 7.1 estimates the expected number of reliable service days per year, and Part 7.2 develops a binomial probability model to optimise staffing levels for minimal cost.

7.1 Estimate of Reliable Service Days

7.1.1 Methodology

The objective is to estimate the number of annual reliable service days. A problem day is defined as having less than 15 people on duty. Therefore, a reliable day is any day with 15 or more staff. The analysis is based on the provided histogram, which shows the staff distribution over a 397-day sample.

7.1.2 Analysis and Results

From the histogram, the observed frequencies for unreliable days (14 or fewer staff) were summed. This yielded a total of 8 unreliable days (3 days with 13 staff, 5 days with 14).

- Reliable Days in Sample: $397(total) - 8(unreliable) = 389\ days$
- Probability of Reliable Day: $\frac{389}{397} = 0.9798$

This probability is extrapolated to a 365-day year: $Expected\ Reliable\ Days = 0.9798 \times 365 = 357.6\ days$

Based on the sample data, the agency should expect approximately 358 days of reliable service per year.

7.2 Profit Optimisation

7.2.1 Methodology

The goal is to find the optimal number of assigned personnel (n) that minimises the total average monthly cost. This total cost is the sum of fixed personnel costs (R25,000 per person) and variable problem costs (R20,000 per problem day).

A Binomial Distribution model, $X{\sim}Bin(n,p)$, was used to represent the number of staff (X) who show up on a given day.

- n: The total personnel assigned (the variable to optimise).
- p: The "show-up rate" for a single employee.

This "show-up rate" was estimated from the histogram's properties ($\mu \approx 19,\ \sigma^2 \approx 5.44$) to be $p \approx 0.7137$

Using this model, the R script calculated the total monthly cost for a range of n values (20 to 30). For each n, it calculated the probability of a problem day ($X \leq 14$) and multiplied this by the problem cost and average days per month.

7.2.2 Optimisation Results

The R script generated a full cost analysis. The results are summarised below.

n assigned staff	cost personnel	prob problem day	cost problems month	total cost month
20	500000	0.53	322445.31	822445.31
21	525000	0.39	239995.64	764995.64
22	550000	0.28	169179.61	719179.61
23	575000	0.19	113424.39	688424.39
24	600000	0.12	72630.77	672630.77
25	625000	0.07	44600.66	669600.66
26	650000	0.04	26361.97	676361.97
27	675000	0.02	15048.22	690048.22
28	700000	0.01	8320.79	708320.79
29	725000	0.01	4468.67	729468.67
30	750000	0.00	2336.47	752336.47

Table 8: Profit Optimisation Analysis

The analysis table clearly shows that the total monthly cost is minimised when $n = 25$.

7.3 Part 7 Conclusion

The analysis provides a clear path to optimising the agency's staffing. Based on the output:

The optimal number of assigned staff is 25, which yields the minimum total monthly cost of R669 600.66.

The current staffing level is estimated to be $n = 27$ (based on the histogram's mean), which has an associated cost of R690 048.22. By reducing the assigned personnel from 27 to 25, the company saves R50,000 in fixed salaries. While this increases the expected problem costs by approximately R29 552, it results in a net monthly savings of approximately R20 448, thereby optimising profit.

Overall Conclusion

This report successfully applied a range of analytical techniques to solve complex business problems, fulfilling the ECSA GA4 requirements by providing key insights into process control, data integrity, and optimization.

The analysis yielded three principal findings. First, data integrity is paramount; a data correction exercise revealed that initial sales metrics were significantly overestimated, proving that strategies built on unvalidated data are inherently flawed.

Second, processes are not monolithic. Both SPC and ANOVA analyses confirmed that a "one-size-fits-all" management style is ineffective. No delivery process was robustly capable, and performance was dictated by complex, non-static interactions between Product Type, Month, and Year. This proves that seasonal challenges are not uniform and must be managed at a granular, product-specific level.

Third, optimization is context-dependent. Models for two coffee shops revealed different core problems (a strategic trade-off vs. an operational failure), while a reliability model for a car rental agency identified an optimal staffing level of 25, yielding a net monthly saving of approximately R20 448.

In conclusion, this report demonstrates that actionable intelligence is the end-product of a rigorous analytical process that validates data, quantifies capability, and deconstructs complex interactions to provide targeted, defensible recommendations.

References

1. OpenAI. (2025). *ChatGPT* [Large language model]. https://chat.openai.com
2. Stellenbosch University. (2025). *Cheat Sheet for Basic Data Analysis in R* [Module resource, QA344]. Department of Industrial Engineering.
3. Stellenbosch University. (2025). *QA 344 2025 ECSA: GA4 outcome in Industrial Engineering* [Project brief and rubric]. Department of Industrial Engineering.
4. Stellenbosch University. (2025). *Quality Assurance 344 course materials* [Lecture slides and R code]. SunLearn. https://stemlearn.sun.ac.za/course/view.php?id=1492
5. Stellenbosch University. (2025). *Short summary of SPC* [Module resource, QA344]. Department of Industrial Engineering.
6. Stellenbosch University. (2025). *Statistical Methods in Quality Assurance Part 1 summary* [Module notes, QA344]. Department of Industrial Engineering.
7. STHDA Website: STHDA. (n.d.) *MANOVA test in R: Multivariate analysis of variance*. Available at: http://www.sthda.com/english/wiki/manova-test-in-r-multivariate-analysis-of-variance
8. CRAN R Package Vignette: Friedrich, S., Konietschke, F. and Pauly, M. (2023) *Introduction to MANOVA.RM*. R package vignette version 0.5.4. Available at: https://cran.r-project.org/web/packages/MANOVA.RM/vignettes/Introduction_to_MANOVA.RM.html

Author's Note on AI Usage:

The generative AI tool ChatGPT (OpenAI, 2025) was used during this project for the limited purpose of debugging R code syntax and optimizing logical operations. All analysis, interpretation, and written content are the original work of the author.