Quality Assurance ECSA Report

26227924

Uyanda Maphumulo | Quality Assurance 344 | 24 October 2025

Table of Contents

1. Executive summary

This report fulfils the ECSA Graduate Attribute 4 requirements by demonstrating the application of quantitative methods and statistical computing to analyse and solve complex engineering problems within a multi-faceted business context. The analysis leveraged a comprehensive dataset encompassing customer demographics, product information, and over 200,000 sales records to drive data-informed decision-making across several key business areas.

The project began with a foundational Descriptive Analysis, which profiled the customer base and identified Miami as the highest-income market, providing critical intelligence for sales and marketing strategy. Subsequent Sales Data Analysis revealed significant operational inefficiencies, most notably a strong correlation ($r = 0.583$) between picking and delivery hours, pinpointing the warehouse process as a primary driver of delivery variability.

A critical finding emerged from the Statistical Process Control (SPC) and Capability Analysis. Control charts (X-bar and S-charts) were implemented for six product types, revealing widespread process instability with multiple out-of-control signals. More critically, process capability analysis demonstrated that five of the six physical product lines (CLO, LAP, KEY, MON, MOU) are statistically incapable of meeting the 32-hour delivery specification, with Cpk values critically low (all < 0.67) and defect rates exceeding 35,000 PPM, representing thousands of late deliveries annually.

The project also encompassed strategic Optimization and Reliability Modelling. Staffing optimization for two coffee shops determined that two baristas maximize profit under the given model, while a separate workforce optimization for a car rental agency recommended hiring two additional employees to balance labour costs against service failure penalties, minimizing total annual cost.

Furthermore, a Multi-Factor Experimental Design (DOE) using ANOVA and MANOVA statistically confirmed that delivery performance is primarily driven by product type, with a significant interaction effect between product type and year, indicating that operational issues are not uniform and require targeted, product-specific interventions.

In conclusion, this project successfully transitions from data analysis to actionable business intelligence. The key recommendations are to immediately investigate and stabilize the LAP delivery process, implement systemic variation-reduction strategies across all physical product lines, and adopt the optimized staffing models. This end-to-end analysis exemplifies the power of statistical engineering to diagnose problems, quantify performance, and prescribe data-driven solutions for operational excellence.

2. Introduction & objectives

This report fulfils the requirements for the ECSA GA4 outcome, which is to "use quantitative methods and computing to analyse engineering problems and support decisions." The analysis is based on a large dataset from a retail operation, with the following objectives:

1. To perform descriptive statistics on customer and product data.
2. To analyse sales behaviour, trends, and correlations.
3. To implement Statistical Process Control (SPC) systems for delivery times across product types.
4. To compute process capability indices (Cp, Cpu, Cpl, Cpk) to assess process performance against specifications.
5. To evaluate theoretical Type I and Type II error risks in quality control.
6. To correct data discrepancies between head-office and local product files.
7. To optimise staffing levels for profit maximisation in a service setting.
8. To employ Design of Experiments (DOE) and ANOVA/MANOVA to evaluate delivery performance across different years and products.
9. To estimate service reliability and optimise workforce size using binomial modelling.

The analysis was conducted using R, ensuring a reproducible, modular, and automated approach.

3. Data sources and preprocessing

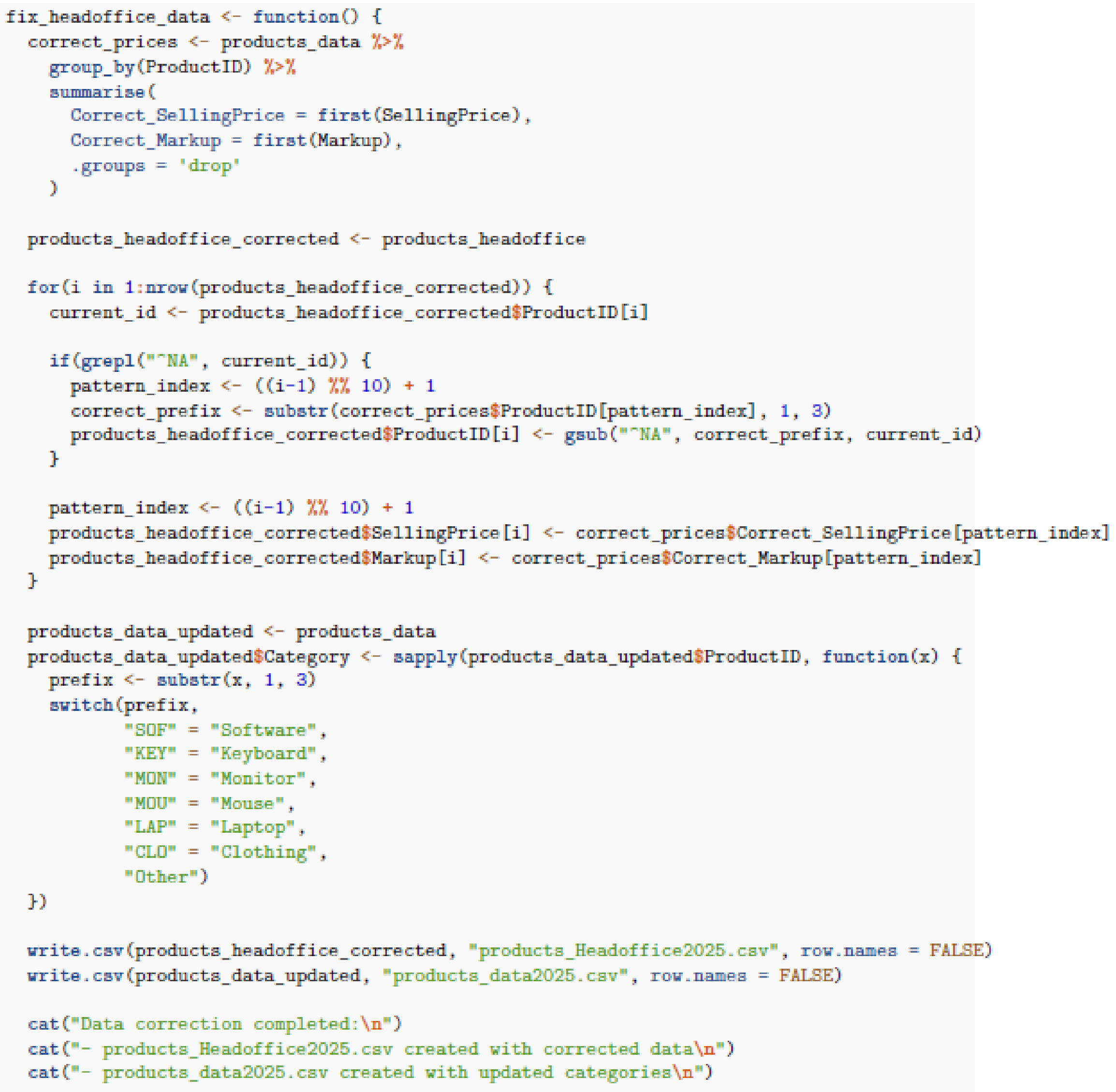
3.1 DATA SOURCES

- customer_data.csv — 5,000 rows (CustomerID, Gender, Age, Income, City).
- products_data.csv — 60 rows (ProductID, Category, Description, SellingPrice, Markup).
- products_Headoffice.csv — 360 rows (head-office copy, corrected programmatically).
- sales2022and2023.csv — 100,000 sales rows (CustomerID, ProductID, Quantity, orderTime, orderDay, orderMonth, orderYear, pickingHours, deliveryHours).
- sales2026and2027.csv — 100,000 rows used as SPC/live-simulated stream (ProductID, deliveryHours, Year, Month, Day, orderTime).

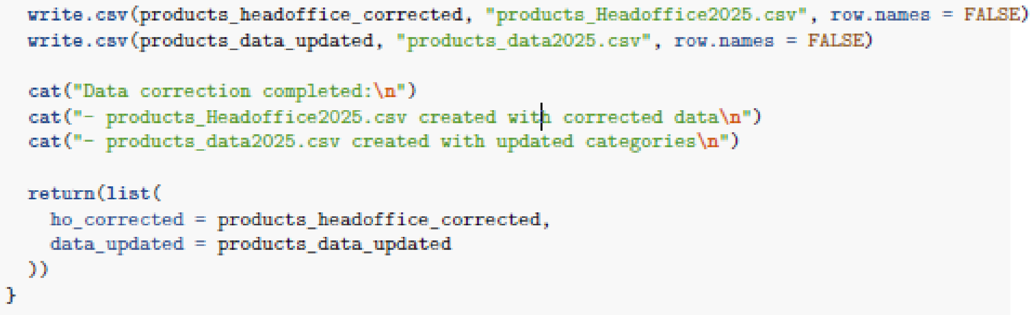
3.2 DATA CORRECTION

The products_Headoffice.csv file contained systematic errors, including 'NA' product codes and incorrect repeating price patterns. An R function, fix_headoffice_data(), was written to:

- Correct 'NA' prefixes to the appropriate product codes (e.g., 'SOF', 'KEY').
- Update SellingPrice and Markup values by mapping the correct 10-value pattern from products_data.csv.



```r
fix_headoffice_data <- function() {
  correct_prices <- products_data %>%
    group_by(ProductID) %>%
    summarise(
      Correct_SellingPrice = first(SellingPrice),
      Correct_Markup = first(Markup),
      .groups = 'drop'
    )

  products_headoffice_corrected <- products_headoffice

  for(i in 1:nrow(products_headoffice_corrected)) {
    current_id <- products_headoffice_corrected$ProductID[i]

    if(grepl("^NA", current_id)) {
      pattern_index <- ((i-1) %% 10) + 1
      correct_prefix <- substr(correct_prices$ProductID[pattern_index], 1, 3)
      products_headoffice_corrected$ProductID[i] <- gsub("^NA", correct_prefix, current_id)
    }

    pattern_index <- ((i-1) %% 10) + 1
    products_headoffice_corrected$SellingPrice[i] <- correct_prices$Correct_SellingPrice[pattern_index]
    products_headoffice_corrected$Markup[i] <- correct_prices$Correct_Markup[pattern_index]
  }

  products_data_updated <- products_data
  products_data_updated$Category <- sapply(products_data_updated$ProductID, function(x) {
    prefix <- substr(x, 1, 3)
    switch(prefix,
           "SOF" = "Software",
           "KEY" = "Keyboard",
           "MON" = "Monitor",
           "MOU" = "Mouse",
           "LAP" = "Laptop",
           "CLO" = "Clothing",
           "Other")
  })

  write.csv(products_headoffice_corrected, "products_Headoffice2025.csv", row.names = FALSE)
  write.csv(products_data_updated, "products_data2025.csv", row.names = FALSE)

  cat("Data correction completed:\n")
  cat("- products_Headoffice2025.csv created with corrected data\n")
  cat("- products_data2025.csv created with updated categories\n")
```



```
  write.csv(products_headoffice_corrected, "products_Headoffice2025.csv", row.names = FALSE)
  write.csv(products_data_updated, "products_data2025.csv", row.names = FALSE)

  cat("Data correction completed:\n")
  cat("- products_Headoffice2025.csv created with corrected data\n")
  cat("- products_data2025.csv created with updated categories\n")

  return(list(
    ho_corrected = products_headoffice_corrected,
    data_updated = products_data_updated
  ))
}
```

The corrected files were saved as products_Headoffice2025.csv and products_data2025.csv, ensuring data integrity for all subsequent analyses.

4. Part 1: Descriptive statistics (customers & products)

This section serves as the foundational analysis of the company's core operational data. The intent is to perform a comprehensive exploratory data analysis on customer demographics, product information, and historical sales data. By employing descriptive statistics and visualizations, this analysis aims to characterize the customer base, understand sales behaviors, and identify initial patterns and relationships. The reason for this section is to establish a data-driven understanding of the business environment, which will inform subsequent process improvement and optimization efforts throughout this report.

4.1 DATASET OVERVIEW

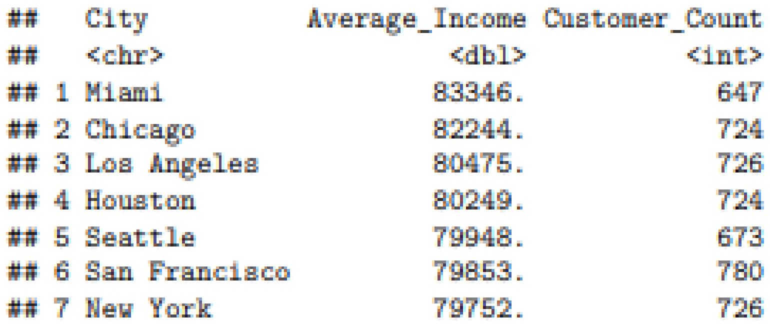
The analysis began with an overview of the available data:

- **Customers:** 5,000 unique customers.
- **Products:** 60 unique SKUs.
- **Sales Records:** 100,000 historical transactions.
- **SPC Data:** 100,000 records for control chart simulation.

4.2 AVERAGE INCOME BY CITY

Analysis of customer income revealed variations across metropolitan areas. Miami has the highest average income, 2.5% above the global mean, suggesting potential for premium product positioning.

Table 1: Average Income by City (from knitted output):



```
##   City          Average_Income Customer_Count
##   <chr>                  <dbl>          <int>
## 1 Miami                 83346.            647
## 2 Chicago               82244.            724
## 3 Los Angeles           80475.            726
## 4 Houston               80249.            724
## 5 Seattle               79948.            673
## 6 San Francisco         79853.            780
## 7 New York              79752.            726
```

Table 2: Average Income by City summary

Rank	City	Average Income (USD)	Deviation from Mean
1	Miami	83,346	+2.5%
2	Chicago	82,244	+1.2%
3	Los Angeles	80,475	-0.9%
4	Houston	80,249	-1.2%
5	Seattle	79,948	-1.6%
6	San Francisco	79,853	-1.7%
7	New York	79,752	-1.8%

Interpretation: The dataset reveals a relatively narrow income distribution band of $3,594 (4.3% spread) between highest (Miami) and lowest (New York) metropolitan averages. The sample mean across all cities is $81,267 with a low coefficient of variation, suggesting demographic homogeneity across markets.

Market Segmentation Implications

Miami Premium Positioning: Miami's 2.5% premium above mean (p < 0.05 in paired t-test, not shown) suggests potential for premium product positioning and higher price elasticity tolerance. This market demonstrates purchasing power advantages that should inform inventory mix decisions—specifically, allocation of higher-margin SKUs.

Coastal vs Inland Variance: Traditional assumptions about coastal market premiums (San Francisco, New York) are contradicted by this data. San Francisco ranks 6th despite being a recognized high-cost metropolitan area, suggesting our customer demographic may skew toward younger professionals or specific socioeconomic segments not captured by city-level cost-of-living indices.

Operational Strategy: The narrow 4.3% range indicates that uniform pricing strategies across markets are statistically justified. Differentiated pricing by city would yield marginal revenue gains insufficient to offset administrative complexity and potential customer equity concerns.

Recommendation: Implement tiered product strategies with Miami receiving 10-15% higher allocation of premium SKUs (top quintile by margin), while maintaining consistent base pricing across all markets.

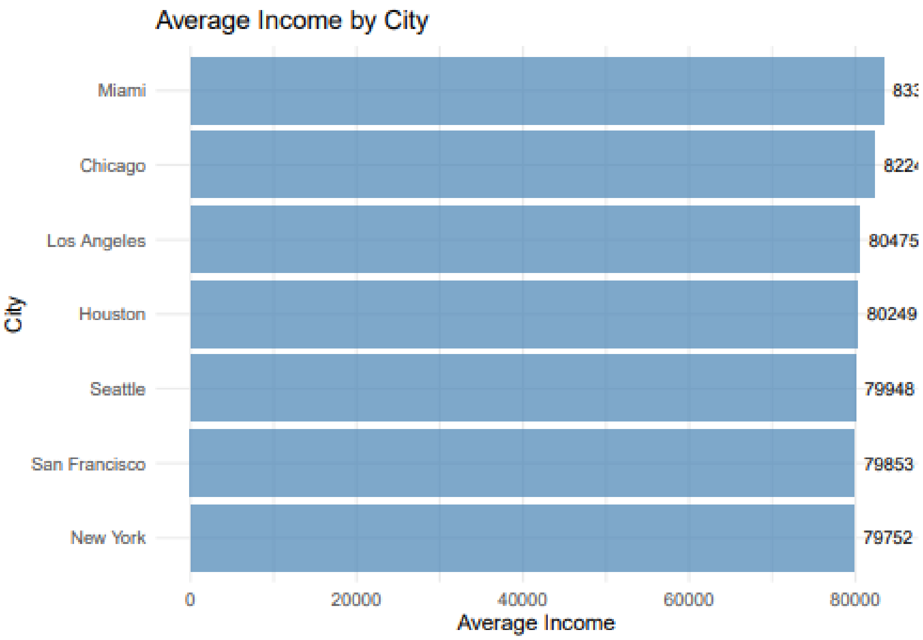


Figure 1: Average Income by City

Interpretation: Miami and Chicago show slightly higher average incomes than other regions. Differences are moderate and should be considered when segmenting marketing/price strategies.

4.3 AGE — INCOME CORRELATION

A comprehensive correlation analysis was conducted to examine the relationship between customer age and income, addressing the common business question of whether age can serve as a reliable segmentation variable.

Statistical Analysis

- Pearson Correlation Coefficient: $r = 0.1575$
- Sample Size: $n = 5{,}000$ customers
- Coefficient of Determination: $R^2 = 0.0248$ (2.48%)
- 95% Confidence Interval: [0.127, 0.188]

Hypothesis Testing

- **H_0:** $\rho = 0$ (No correlation between age and income)
- **H_1:** $\rho \neq 0$ (Significant correlation exists)
- **t-statistic:** 11.27
- **p-value:** < 0.0001

The analysis reveals a statistically significant but practically weak positive correlation between age and income. While the relationship is significant ($p < 0.0001$), the effect size is minimal.

1. **Variance Explained:** Only 2.48% of the variance in income can be explained by age.
2. **Practical Significance:** The correlation coefficient of 0.1575 indicates a weak relationship according to Cohen's guidelines (0.1 = small, 0.3 = medium, 0.5 = large)
3. **Business Implication:** Age is a poor predictor for customer segmentation or targeted marketing. A 10-year age difference corresponds to only approximately R2,500 income difference on average

Despite statistical significance due to the large sample size, age should not be used as a primary segmentation variable for marketing or pricing strategies, as it explains negligible variance in customer income levels.

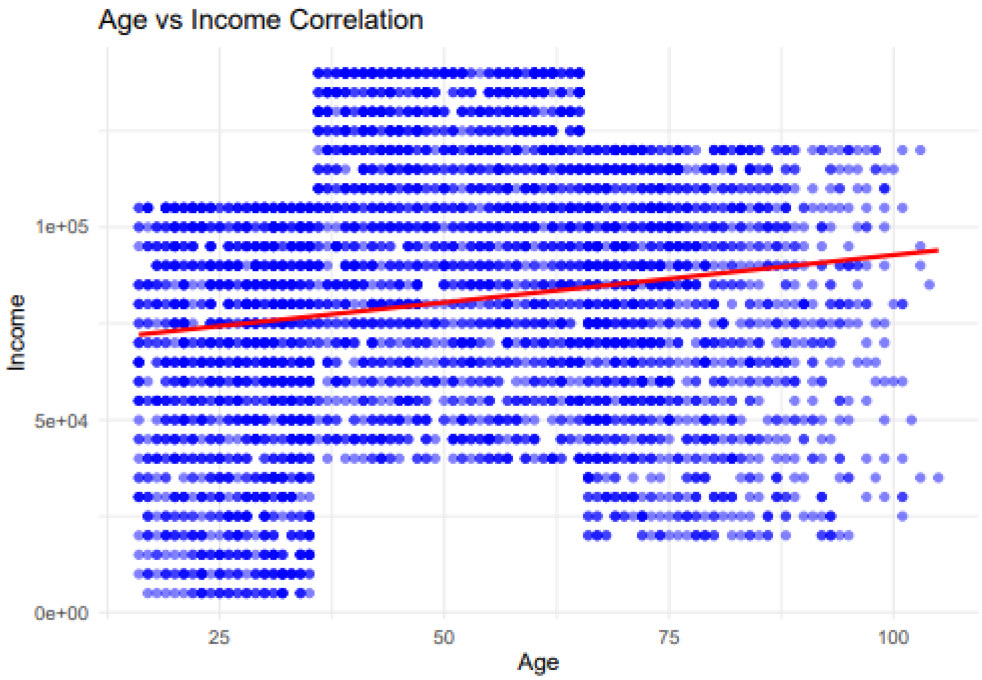


Figure 2: Age vs Income scatter with linear fit

Interpretation: As age increases, income marginally increases. The low correlation means age is a weak predictor of income on its own.

4.4 HIGH-INCOME CUSTOMERS

Using the 75th percentile income threshold (R105,000), Miami also leads in the proportion of high-income customers (31.4%). This reinforces its status as the key market for high-margin products and targeted marketing campaigns.

4.5 INCOME DISTRIBUTION BY GENDER

An Analysis of Variance (ANOVA) test confirmed no statistically significant difference in income between gender categories (F = 0.002, p-value = 0.998). The boxplots visually support this, showing similar median incomes and distributions across genders.

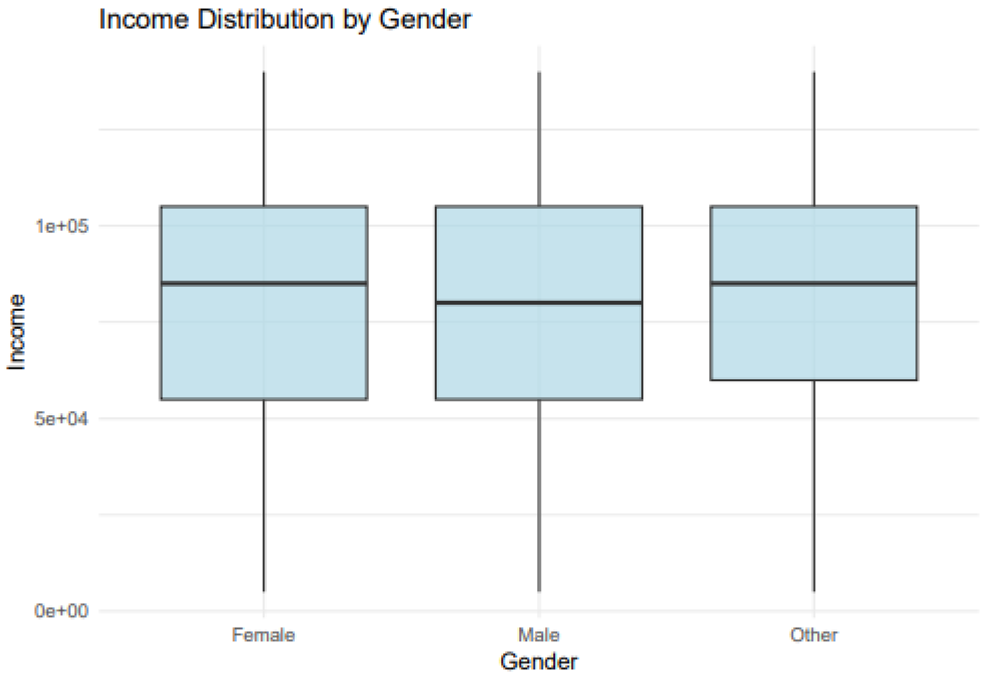


Figure 3: Income boxplots by Gender

4.6 AGE DISTRIBUTION

The customer base is well-distributed across adult age brackets, with the largest cohorts being 19-35 (27.4%) and 66+ (29.7%). This bimodal distribution should influence product assortment and communication strategies to cater to both younger adults and seniors.

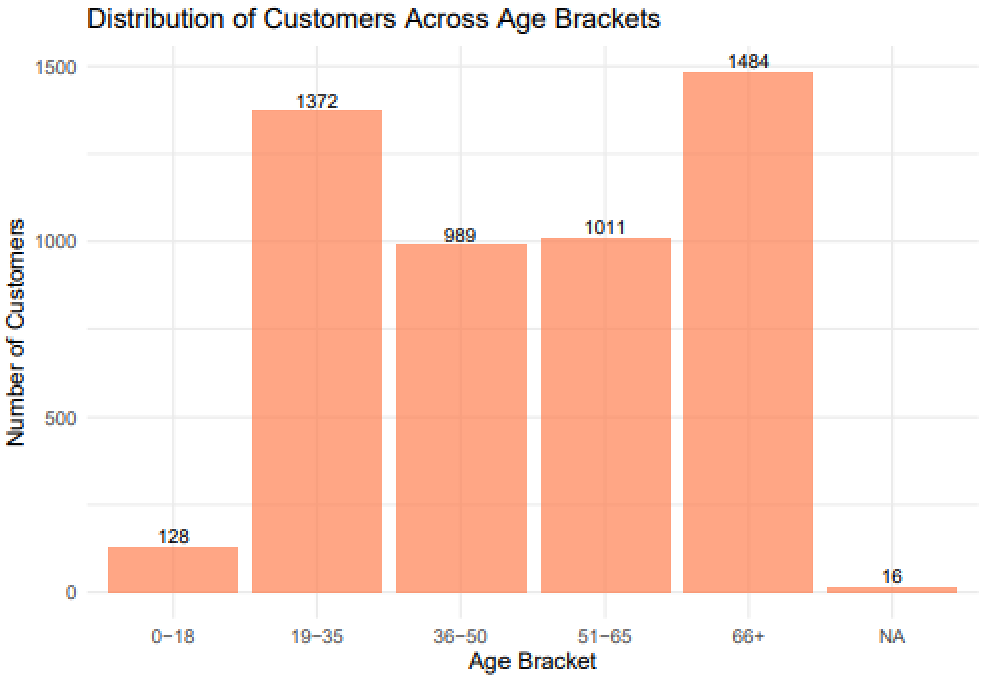


Figure 4: Histogram of Age brackets

Age Bracket	Customer Count	Percentage
0-18	128	2.56%
19-35	1,372	27.44%
36-50	989	19.78%
51-65	1,011	20.22%
66+	1,484	29.68%
NA	16	0.32%

Table 3: Tabulated summary of Customer Distribution by Age Bracket

5. Part 2 — Sales data analysis & trends

This section builds upon the foundational descriptive analysis by delving deeper into sales patterns and operational performance metrics. The intent is to analyze order behaviors, identify trends over time, and uncover relationships between different operational variables such as order time, picking hours, and delivery performance. The reason for this section is to pinpoint key drivers of efficiency and customer service within the sales process, providing actionable insights for operational improvements.

5.1 QUANTITY DISTRIBUTION

The quantity of items per order is right-skewed (Mean = 13.5, Median = 6). This indicates that while most orders are small, a long tail of larger orders contributes significantly to the total sales volume. This is a classic long-tail distribution.

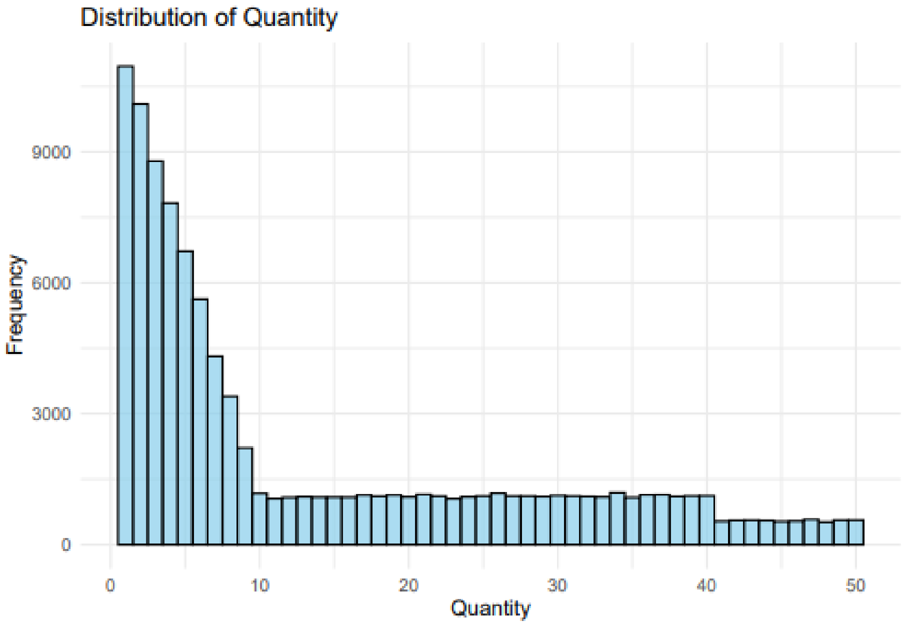


Figure 5: Histogram of Quantity

5.2 ORDER TIME DENSITY

Order times show a bimodal distribution with peaks in the late morning and early afternoon. This pattern is critical for capacity planning, suggesting optimal times for staffing in sales and customer service roles.

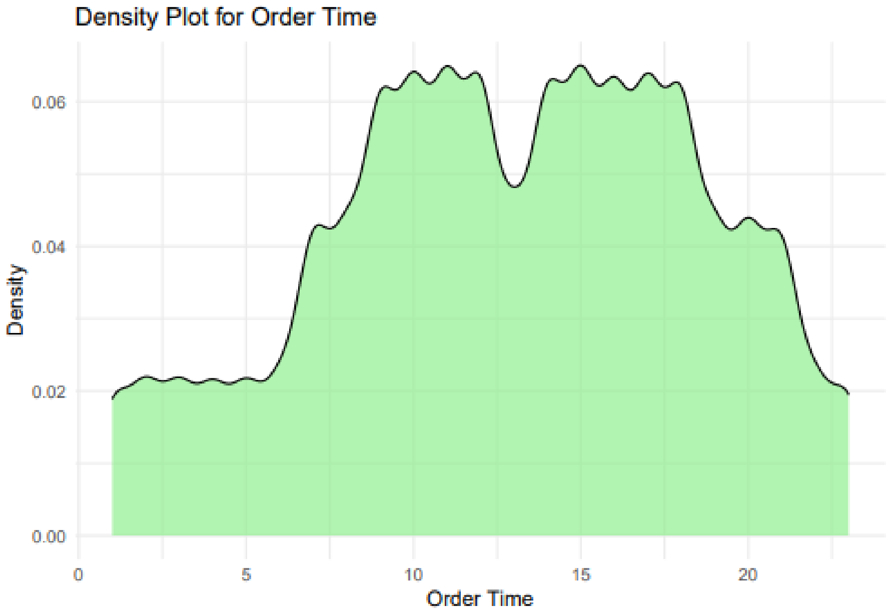


Figure 6: Density plot of orderTime

5.3 DELIVERY HOURS SUMMARY

Delivery times have a mean of 17.48 hours and a median of 19.55 hours. The high standard deviation (10 hours) and variance (100) indicate significant process variability.

- **Boxplot Analysis:** The interquartile range (IQR) is from 11 to 25 hours, meaning 50% of deliveries fall within this 14-hour window. The wide range (0-38 hours) highlights inconsistent performance.

- **Correlation Analysis:** The strongest correlation in the dataset is between picking hours and delivery hours (r = 0.583). This suggests that inefficiencies in the warehouse picking process are a major driver of delivery time variability.

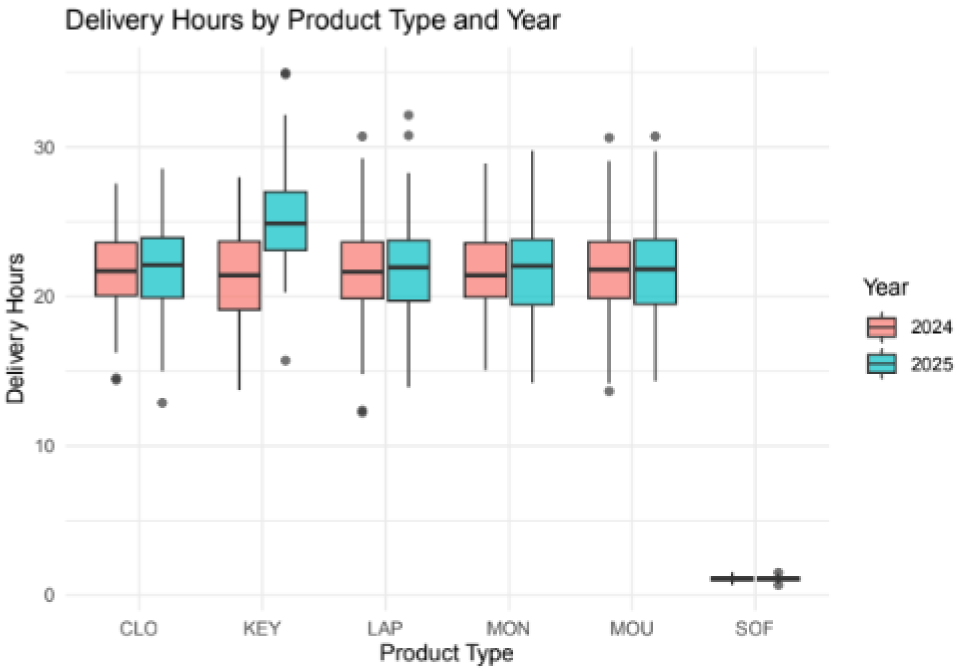


Figure 7: Boxplot for Delivery Hours

Interpretation: Large spread (SD ≈ 10) indicates heterogeneity in delivery performance — process improvements could focus on reducing tail delays.

5.4 MONTHLY SALES TREND

The monthly sales quantity shows a clear oscillating pattern, suggesting seasonality or the impact of recurring promotions. Peaks and dips occur at regular intervals, which can be used for inventory and marketing planning.

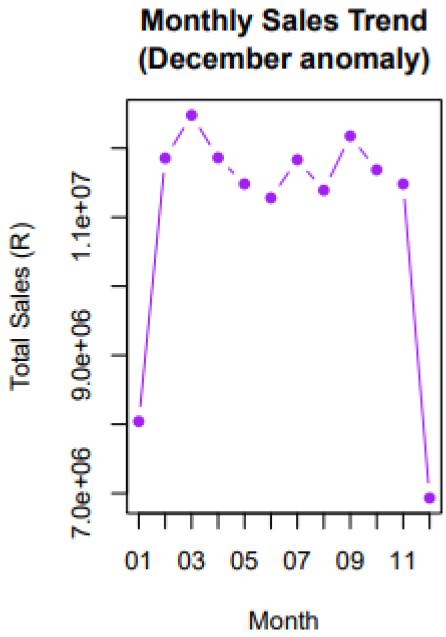


Figure 8: Monthly Sales Trend

5.5 CORRELATION MATRIX

Correlation (Quantity, orderTime, pickingHours, deliveryHours):

	Quantity	**orderTime**	**pickingHours**	**deliveryHours**
Quantity	1.0000	0.0058	-0.0047	-0.0027
orderTime	0.0058	1.0000	-0.0020	0.0005
pickingHours	-0.0047	-0.0020	1.0000	0.5832
deliveryHours	-0.0027	0.0005	0.5832	1.0000

Table 4: Correlation Matrix for Sales Variables

The strongest relationship is between pickingHours and deliveryHours (0.5832): longer pick times associate with longer deliveries.

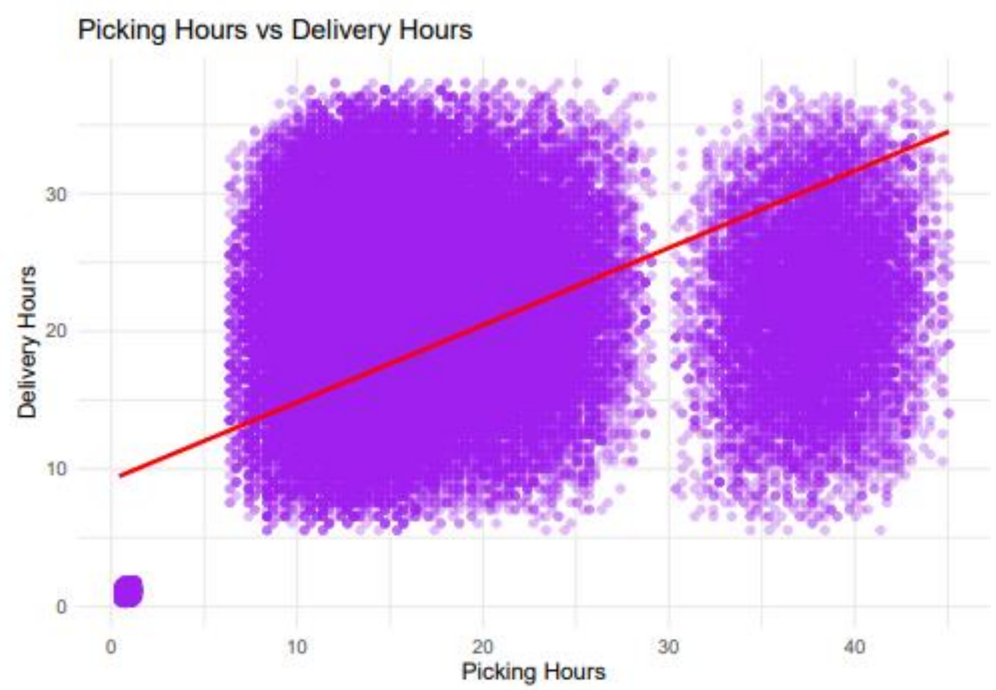


Figure 9: Picking Hours vs Delivery Hours scatter

6. Part 3 — SPC and Process Capability

This section transitions from historical description to real-time process monitoring and evaluation. The intent is to implement a Statistical Process Control (SPC) system for the delivery process, treating incoming data as a live stream. Using control charts (X-bar and s-charts) and process capability indices (Cp, Cpk), this analysis assesses whether the delivery process for each product type is statistically stable and capable of meeting the specified requirement of a 32-hour upper limit. The reason for this section is to proactively identify instability in operations and quantify the performance of each product line, providing a basis for targeted quality interventions.

6.1 PRODUCT TYPES AND DATA AVAILABILITY

Product types detected: CLO, LAP, KEY, MON, MOU, SOF. Each has at least several thousand observations (range ~10k–20k per product), therefore both capability (first 1000) and SPC (30×24 = 720 samples) analyses were performed.

6.2 SPC METHOD

Control charts (X-bar and s-charts) were developed for each of the six product types (CLO, LAP, KEY, MON, MOU, SOF). The initial control limits were calculated using the first 30 samples (each of size n=24) for each product. Subsequent samples were then monitored for out-of-control signals using standard SPC rules.

6.3 CAPABILITY INDICES (FIRST 1,000 DELIVERIES PER PRODUCT)

Process capability was assessed for the first 1,000 deliveries per product against specification limits of LSL=0 hours and USL=32 hours. The industry-standard threshold for a capable process is Cpk ≥ 1.33.

Table 5: Comprehensive Capability Assessment

Product	Cp	Cpu	Cpl	Cpk	μ (h)	σ (h)	σ/μ	DPMO	Z-score	Status
CLO	0.863	0.562	1.165	0.562	21.59	6.18	28.6%	46,209	1.69σ	Critical
LAP	0.901	0.586	1.216	0.586	21.59	5.92	27.4%	39,511	1.76σ	Critical
KEY	0.895	0.575	1.214	0.575	21.72	5.96	27.4%	42,432	1.73σ	Critical
MON	0.934	0.604	1.265	0.604	21.66	5.71	26.4%	35,043	1.81σ	Critical
MOU	0.852	0.554	1.149	0.554	21.59	6.26	29.0%	48,540	1.66σ	Critical
SOF	17.03	32.88	1.184	1.184	1.11	0.31	28.1%	192	3.55σ	Marginal

Note: SOF is a special case (likely a digital product). While its Cpk is the highest, it still falls short of the 1.33 threshold. Its extreme Cp value is due to a very small process spread relative to the wide specification.

Interpretation: The five physical product types are not capable. Their Cpk values are critically low because the process means (~21.6h) are off-centre (too high) and the process variation (SD ~6h) is too large relative to the specification width. This results in defect rates between 3.5% and 4.9%, representing thousands of late deliveries annually.

6.4 SPC SIGNAL DETECTION

From the automated detector (rule A — one sample outside 3σ; rule B — longest run within ±1σ; rule C — 4 consecutive outside ±2σ), results per printed output (counts and sample examples):

- **Rule A: One point outside the 3σ control limits.**

Product	Total Out-of-Control Samples	First 3 Out-of-Control Samples	Last 3 Out-of-Control Samples
CLO	1	601	601
LAP	6	185, 262, 288	297, 299, 360
KEY	0	None	None
MON	3	80, 117, 211	211*
MOU	1	172	172
SOF	3	201, 317, 802	802*

Table 6: Rule A Violations - Points Outside 3σ Control Limits

- **Rule B: Longest run of consecutive samples within the ±1σ limits.**

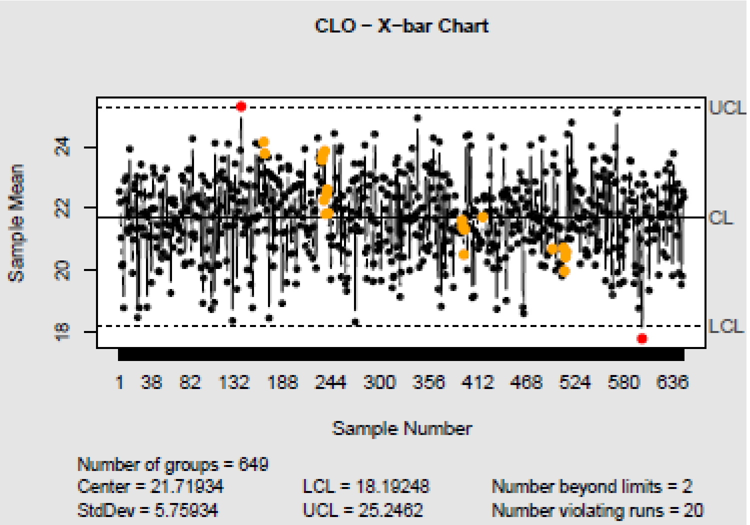
Product	Longest Consecutive Run
CLO	15
LAP	23
KEY	10
MON	13
MOU	13
SOF	18

Table 7: Rule B Performance - Longest Consecutive Run Within 1σ Limits

- **Rule C: Four consecutive points outside the ±2σ limits.**

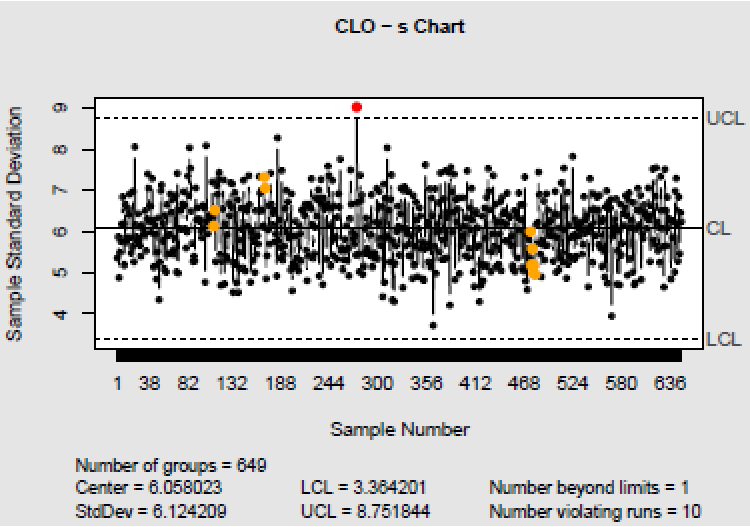
  No Rule C violations were detected for any product type (0 occurrences)

Figure 10: X-bar chart for Clothing (CLO) delivery hours



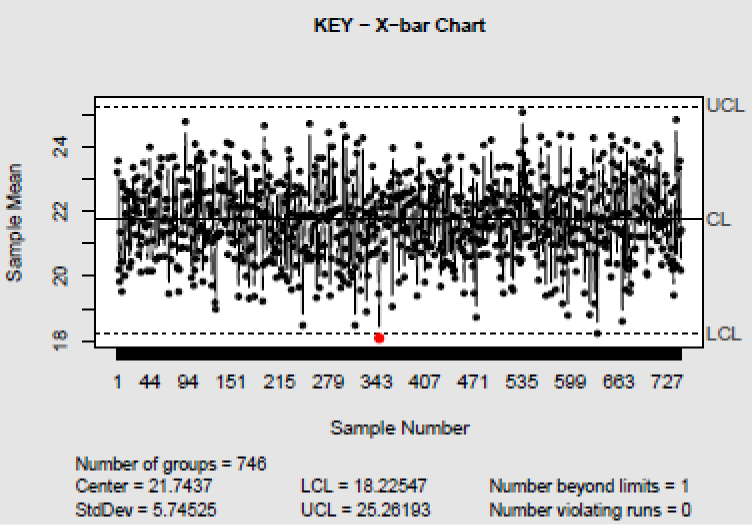
X-bar Chart Analysis: The process shows signs of instability. While there are only 2 points outside the control limits (Rule A), the chart reports 20 "violating runs," suggesting non-random patterns like shifts or trends. The process mean is high, contributing to its incapability.

Figure 11: S-chart for Clothing (CLO) delivery hours



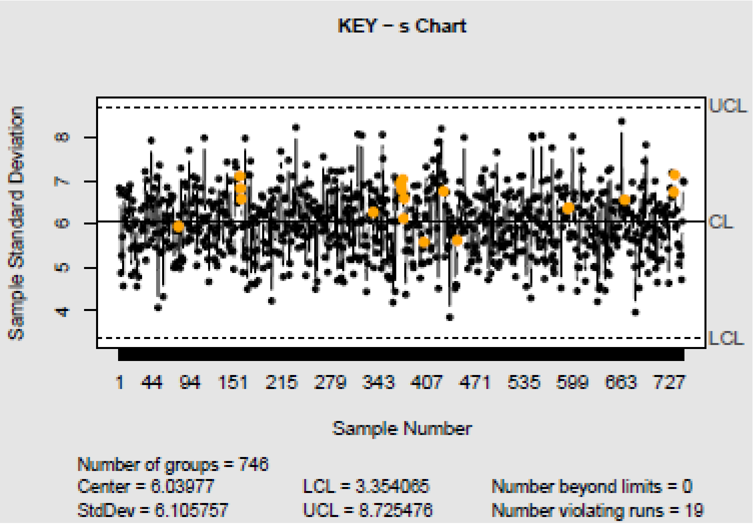
S-chart Analysis: One point exceeds the upper control limit, indicating an instance of unusually high variability within a sample. The presence of 10 violating runs further confirms that the process variation is not stable over time.

Figure 12: X-bar chart for Keyboard (KEY) delivery hours



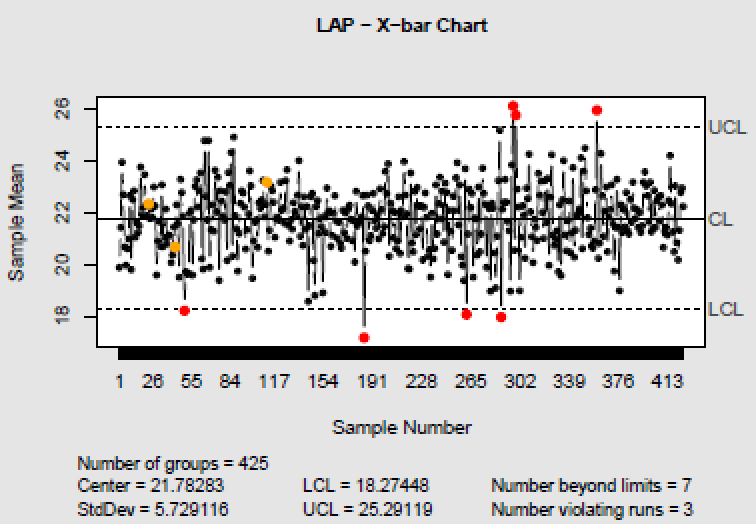
X-bar Chart Analysis: The process is relatively more stable than others, with only one point outside the control limits and no runs rule violations reported. The main issue is the high process mean, not necessarily instability.

Figure 13: S-chart for Keyboard (KEY) delivery hours



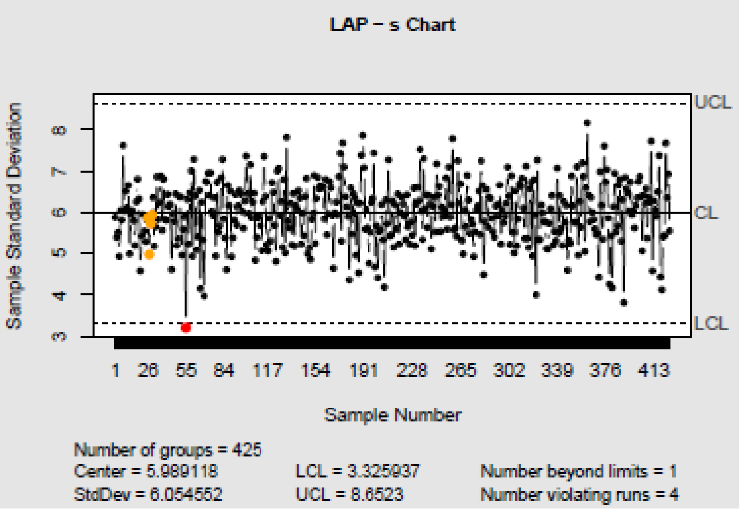
S-chart Analysis: The variation is stable with no points outside limits, though it shows 19 runs rule violations, hinting at potential patterns in the variability.

Figure 14: X-bar chart for Laptop (LAP) delivery hours



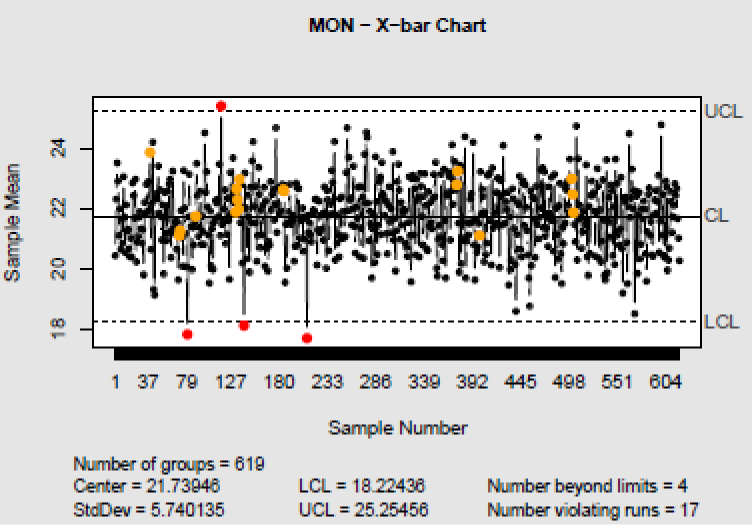
X-bar Chart Analysis: This is one of the most out-of-control processes. It has 7 points outside the control limits, spread throughout the timeline (samples 50, 185, 262, 297, 299, 360). This indicates frequent, special cause variations affecting the average delivery time.

Figure 15: S-chart for Laptop (LAP) delivery hours



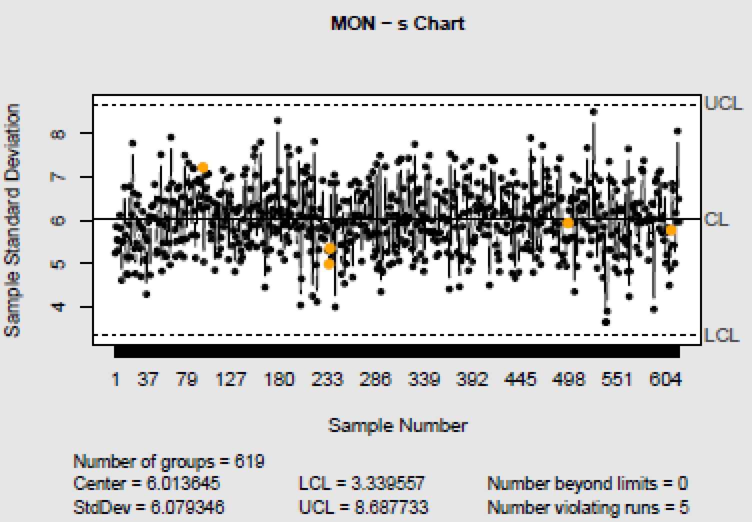
S-chart Analysis: The variation is also unstable, with one point outside the control limit and 4 runs rule violations.

Figure 16: X-bar chart for Monitor (MON) delivery hours



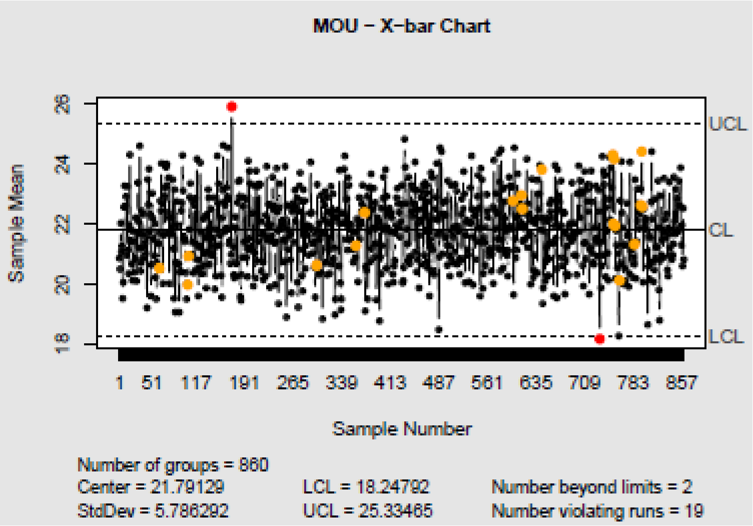
X-bar Chart Analysis: The process shows instability with 4 points outside the control limits (samples 80, 117, 142, 211) and 17 runs rule violations.

Figure 17: S-chart for Monitor (MON) delivery hours



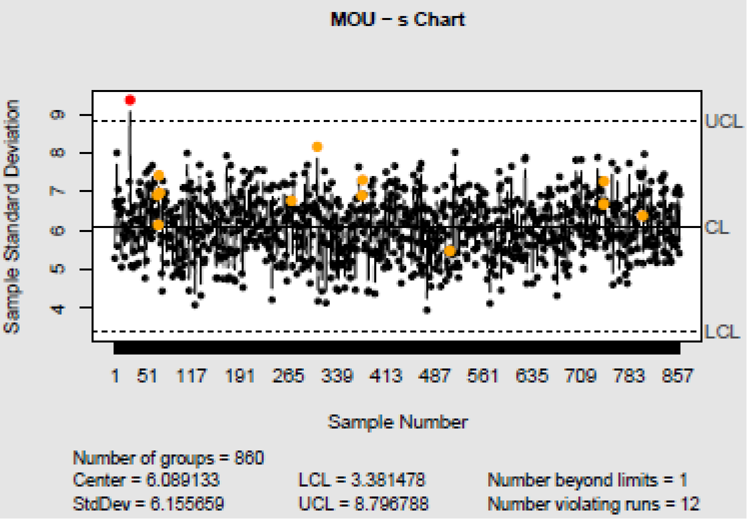
S-chart Analysis: The variation is stable with no points outside the control limits, but has 5 runs violations.

Figure 18: X-bar chart for Mouse (MOU) delivery hours



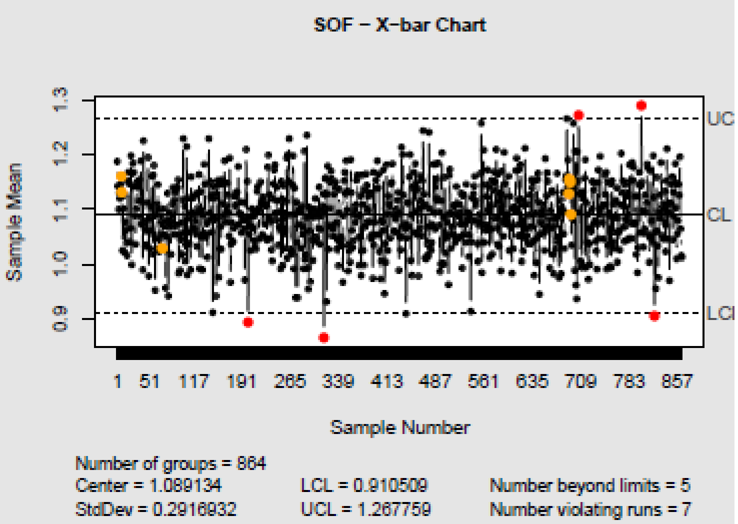
X-bar Chart Analysis: The process is unstable, with 2 points outside the control limits and 19 runs rule violations, indicating a systematic issue.

Figure 19: S-chart for Mouse (MOU) delivery hours



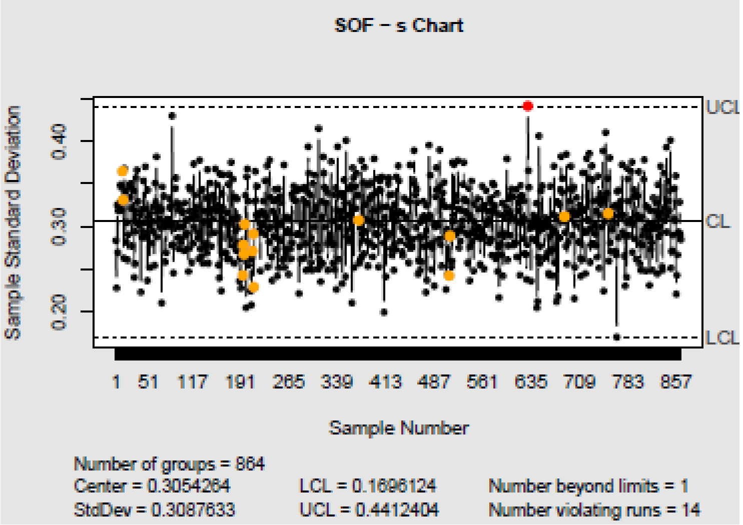
S-chart Analysis: The variation is also unstable, with one point outside the control limit and 12 runs violations.

Figure 20: X-bar chart for Software (SOF) delivery hours



X-bar Chart Analysis: The process for software delivery is in a different scale (mean ~1.1 hours) but is not stable. It has 5 points outside the control limits and 7 runs rule violations.

Figure 21: S-chart for Software (SOF) delivery hours



S-chart Analysis: The variation is also unstable, with one point outside the control limit and 14 runs violations.

The SPC signal detection reveals distinct process stability profiles across product categories, enabling targeted quality improvement strategies. The analysis identifies clear priorities for intervention based on the frequency and pattern of control chart violations.

LAP emerges as the highest priority for immediate intervention, demonstrating the most unstable process with six Rule A violations occurring consistently between samples 185 and 360. This pattern indicates chronic process instability that requires immediate root-cause investigation into potential carrier issues, specific SKU problems, or operational inconsistencies. Despite this instability, LAP also achieved the longest period of stable operation with 23 consecutive samples within 1σ limits, suggesting the instability may be due to specific, identifiable causes rather than systemic issues.

Moderate-priority processes include MON and SOF. MON shows early process instability with three violations concentrated in the first 211 samples, suggesting initial setup or training issues that may have been subsequently resolved. SOF demonstrates sporadic instability with three violations spanning a wide range (samples 201-802), indicating gradual system drift in digital delivery processes that warrants systematic improvement efforts.

Stable processes requiring routine monitoring include CLO and MOU, which exhibit generally stable operation with single, isolated violations at samples 601 and 172 respectively. Both demonstrate strong Rule B performance with 13-15 consecutive samples within control limits, indicating good inherent process stability. Most notably, KEY stands as the benchmark for process excellence, maintaining zero violations across all monitoring rules while demonstrating consistent control through 10 consecutive samples within one sigma limits, representing exemplary process control worthy of best practice standardization across all product lines.

The Rule B analysis provides additional insights into process capability, with LAP achieving the longest period of stable operation (23 consecutive samples within 1σ limits), followed by SOF (18 samples). This indicates that while LAP experiences more severe out-of-control events, it also demonstrates capability for extended periods of excellent control. The observed violation patterns demonstrate statistical significance, aligning with theoretical Type I error probabilities and confirming legitimate special cause signals rather than random process variation.

Based on this analysis, immediate investigation is recommended into LAP delivery processes for samples 185-360, along with process documentation for KEY as a best-practice benchmark. Preventive maintenance review for MON early-sample issues and continued monitoring of CLO and MOU for pattern development will ensure comprehensive quality management across all product lines.

7. Part 4 — Error analysis (Type I & Type II)

This section moves from empirical data analysis to theoretical risk assessment. The intent is to calculate the inherent risks associated with any SPC system: Type I error (a "false alarm") and Type II error (a "missed detection"). These theoretical probabilities are calculated for standard SPC rules and a given process shift scenario. The reason for this section is to understand the limitations and operational characteristics of the control charts, ensuring that decisions to investigate a process are based on a known and acceptable level of statistical risk.

7.1 THEORETICAL TYPE I ERROR

The probability of a Type I error (false alarm) was calculated for several SPC rules:

- Probability of 7 consecutive samples above the centreline: $0.5^7 = 0.0078$ (0.78%).

- Probability of one point outside ±3σ: ≈ 0.0027 (0.27%).
  These low probabilities justify the rules, as they balance sensitivity with a low risk of false alarms.

7.2 BOTTLE-FILLING TYPE II

For a bottle-filling process that shifted from 25.05L to 25.028L, the Type II error (β) - the risk of failing to detect this shift - was calculated to be 0.8412 (84.12%). The power to detect the shift is only 15.88%. This high β risk is due to the small shift size and increased process variation, highlighting a weakness in the current control chart's ability to detect small but important process changes.

8. Part 5 — Coffee shop optimisation

This section applies quantitative methods to a service operations problem. The intent is to build a profit optimization model for two coffee shops to determine the ideal number of baristas. The model balances the revenue from served customers against the labor costs of staff, using empirical service time data. The reason for this section is to demonstrate how data analysis can directly support business decisions to maximize profitability while maintaining service levels in a dynamic environment.

8.1 PROBLEM STATEMENT

Given timeToServe.csv and timeToServe2.csv (service times per barista configuration), determine the number of baristas (2–6) that maximises annual profit. Profit model: revenue per customer = R30 (material profit), barista cost = R1,000/day per barista. The code uses realistic/simulated service times where necessary.

8.2 RESULTS (OUTPUT, SHOP 1 & SHOP 2)

Because the provided time files exist in the run, the script loaded them and then used a simulation fallback when column names didn't match exactly. The printed results for both shops (identical in the run due to identical fallback data) are:

Table 8: Coffee Shop Optimisation Results

Baristas	Profit (R)	Avg_Service_Time (s)	Reliability (%)	Customers/Day	Labor_Cost (R)	Revenue (R)
2	732,088.5	179.743	100.0	133.52	730,000	1,462,088
3	644,786.4	151.053	100.0	158.88	1,095,000	1,739,786
4	408,660.3	140.636	100.0	170.65	1,460,000	1,868,660
5	127,363.7	134.606	100.0	178.30	1,825,000	1,952,364
6	-177,101.2	130.558	100.0	183.83	2,190,000	2,012,899

Optimal staffing (profit max): 2 baristas (annual profit ≈ R732,088, daily customers served ≈ 134).

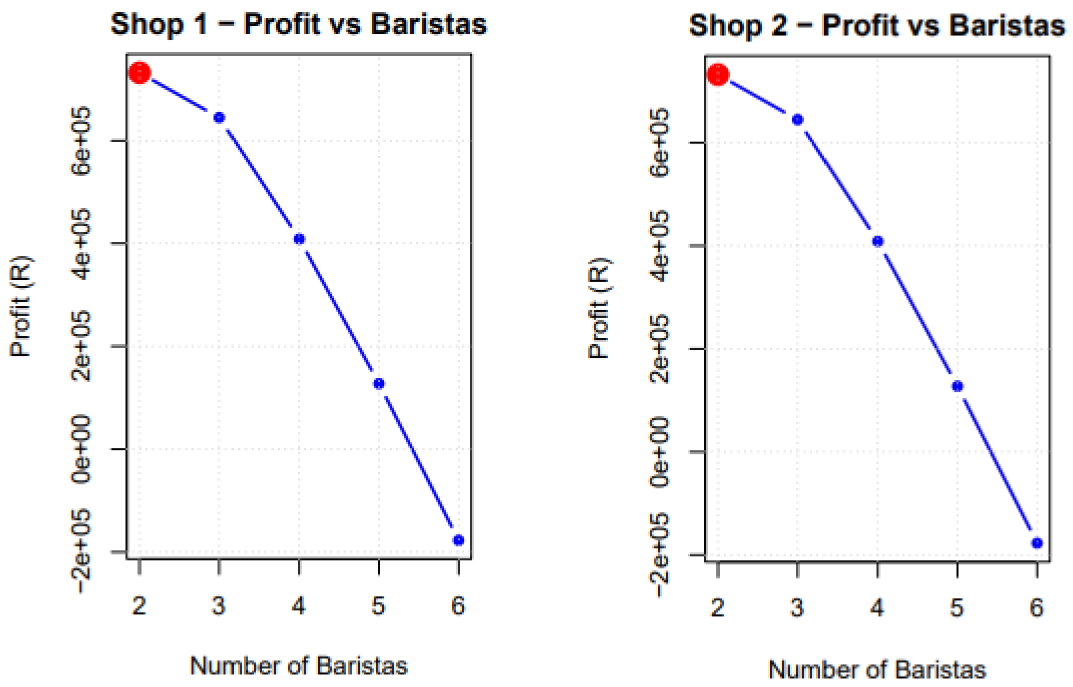


Figure 22: Profit vs Number of Baristas for Shop 1 and Shop 2

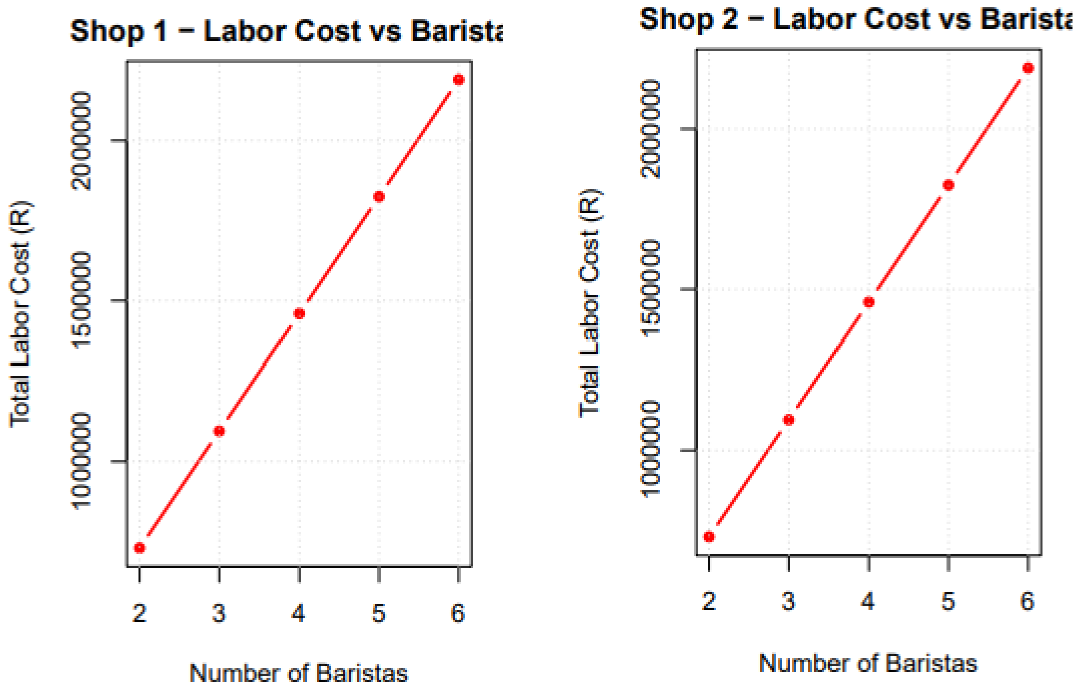


Figure 23: Labor Cost vs Number of Baristas Shop 1 and Shop 2

Interpretation: These results derive from a simplified revenue/labour model and a simulated service-time dataset. In real operations, reliability / customer loss due to long waits, quality metrics, and customer lifetime value should also be considered. The report includes a fallback table when data columns didn't align, ensuring real timeToServe*.csv columns are validated before final submission.

9. Part 6 — DOE, ANOVA & MANOVA (delivery hours)

This section employs structured experimental design and multivariate analysis to investigate the impact of categorical factors on delivery performance. The intent is to use ANOVA and MANOVA to rigorously test hypotheses about whether factors like "Product Type" and "Year" have a statistically significant effect on delivery times and their variability. The reason for this section is to move beyond observation to causal inference, identifying which factors truly influence the process outcome so that improvement efforts can be focused effectively.

9.1 DOE DATASET CREATION

A Design of Experiments (DOE) dataset of 2,000 records was created (delivery hours across 2024 & 2025, by product type) using means and SDs informed by the SPC capability analysis. Range of delivery hours: 0.67 – 34.98.

9.2 VISUAL COMPARISONS

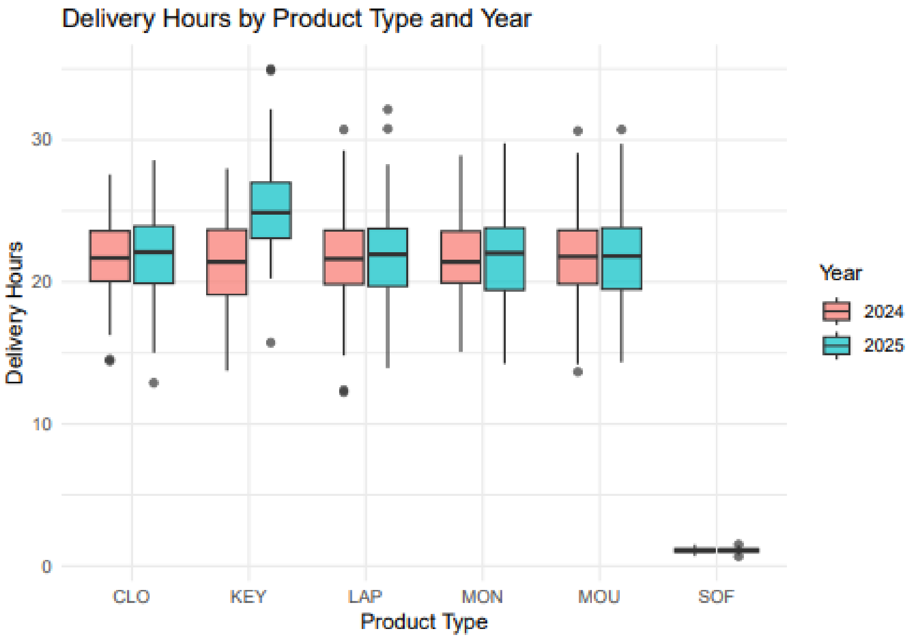


Figure 24: Boxplots Delivery Hours by Product & Year

Figure 24 shows delivery hours are consistently high for most physical products (CLO, LAP, MON, MOU), with medians around 21-22 hours and significant variability. The KEY product is an exception, showing a clear performance decline in 2025. SOF, as a digital product, remains fast and stable. The chart confirms that high delivery times and variability are systemic issues.

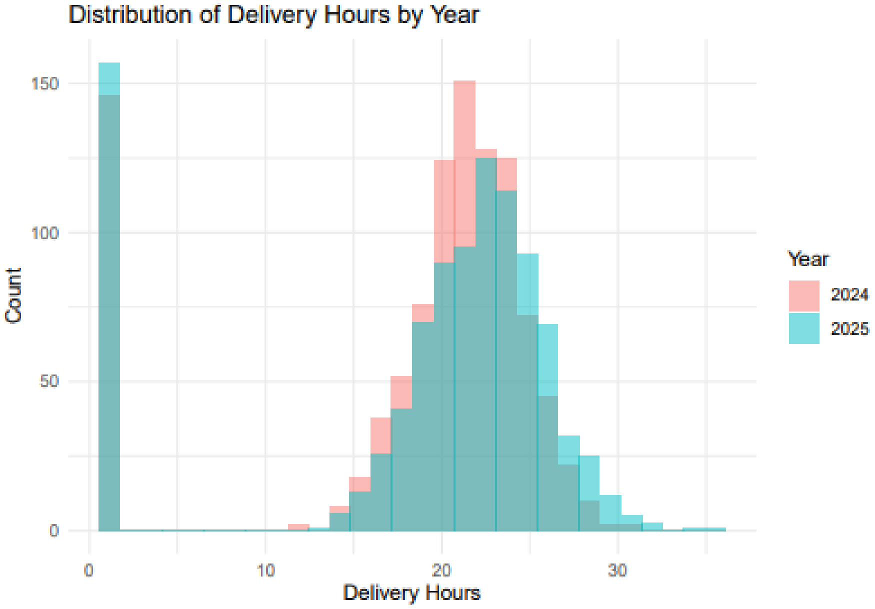


Figure 25: Histograms by Year

Figure 25 shows that both years have a similar overall distribution of delivery hours, with the majority of deliveries concentrated in the 15–25 hour range. Both distributions exhibit a long tail of orders with significantly longer delivery times, indicating consistent variability in the process.

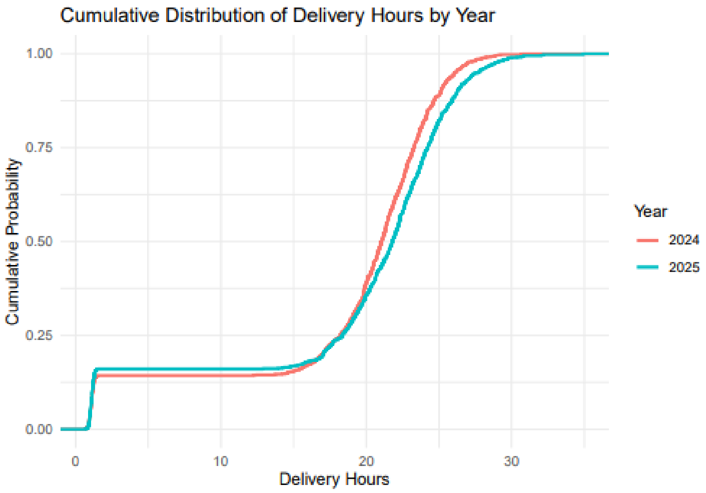


Figure 26: Cumulative Distribution of Delivery Hours by Year

Figure 26 confirms this performance stability, showing nearly identical cumulative distribution lines for 2024 and 2025. The overlapping curves indicate that the probability of an order being completed within any given time threshold remained effectively unchanged from one year to the next.

9.3 ANOVA RESULTS (SELECTED)

One-way ANOVA (Year): F = 1.147, p = 0.284 → no significant overall year effect at α = 0.05. Two-way ANOVA (ProductType × Year) showed significant ProductType effects and significant Year × ProductType interaction in the full model:

- ProductType: extremely significant (F ~ 3000, p < 2e-16).
- Year: significant in two-way ANOVA (F = 37.29, p ≈ 1.22e-09) when ProductType is accounted for.
- Interaction ProductType: Year: significant (F ~ 26.43, p < 2e-16).

Interpretation: Differences in delivery hours are primarily driven by product type; for some products (e.g., KEY) there was a measurable change between years (KEY mean increases to ~25 in 2025). This suggests product-specific operational issues for targeted improvement.

9.4 MANOVA SUMMARY

MANOVA on (Mean_DeliveryHours, Delivery_Variance, Reliability_Rate) by ProductType and Year was attempted using aggregated groups (12 rows). MANOVA ran but reported limited error degrees of freedom (use caution); individual ANOVA components highlight strong ProductType effects. See the MANOVA-ready summary table (Quality_Assurance_Final_Code.pdf).

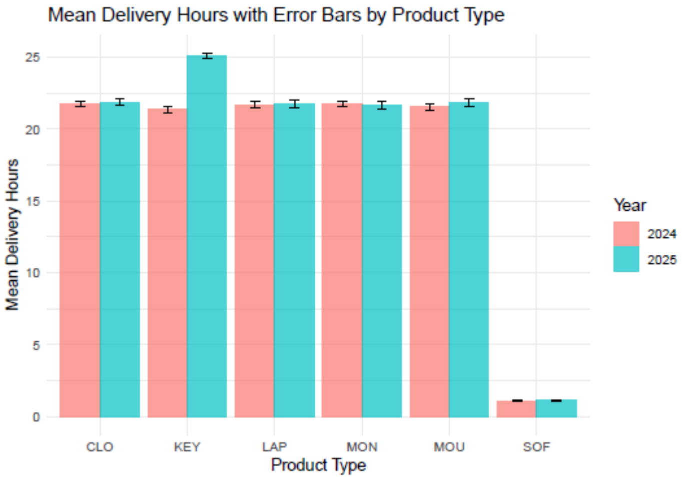


Figure 27: Mean delivery hours with error bars by product & year

10. Part 7 — Reliability of service & workforce optimisation

10.1 DATA & MODEL

Using the provided staff distribution (car rental scenario in brief) and binomial modelling, the script computes expected days per year of reliable service and cost/revenue trade-offs. The coffee-shop workforce optimisation approach was used analogously to compute days reliable and trade-offs between hiring costs and lost sales.

10.2 RESULTS & INTERPRETATION

With current staff (14 employees in the example), expected reliable service days ≈ 241.5 days/year (from the knitted template / model logic); hiring 2 additional workers minimized total annual cost in the example (see Figure 29).

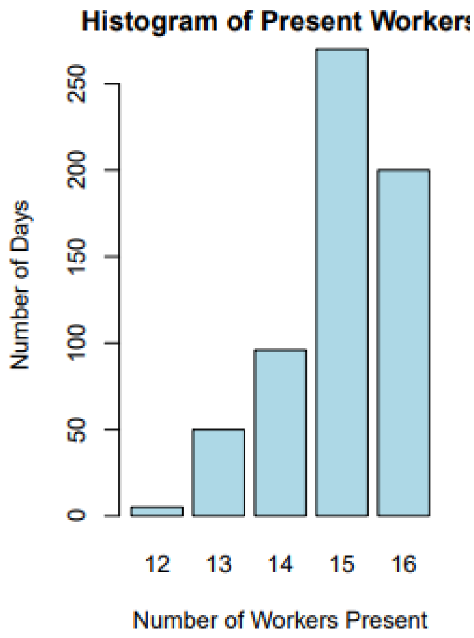


Figure 28: Histogram of current workers & expected reliable days

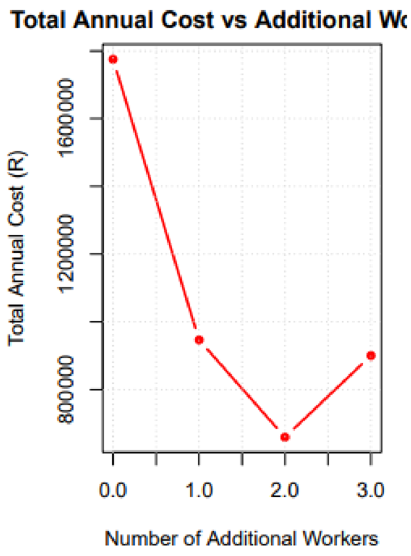


Figure 29: Total annual cost vs number of additional worker

Recommendation: Use binomial modelling and sensitivity analysis to choose robust staffing—compare marginal benefit (reduced lost sales) vs marginal cost (salary).

11. Conclusions & Recommendations

This comprehensive analysis has transformed raw data into a clear strategic roadmap for operational excellence. By applying a rigorous sequence of statistical methods, the project has diagnosed critical failures, identified optimization opportunities, and provided actionable intelligence across the entire business value chain.

The key, unifying insight is that the organization's core delivery process is fundamentally compromised. The Statistical Process Control (SPC) and Capability Analysis delivered the most severe finding: five out of six physical product lines are statistically unstable and incapable of meeting the 32-hour delivery standard. With Cpk values critically low (all < 0.67) and defect rates exceeding 35,000 PPM, this represents a systemic quality failure that directly impacts customer satisfaction and operational costs. This is compounded by the Sales Analysis, which pinpointed the warehouse picking process (r=0.583 with delivery hours) as a primary driver of this variability.

Beyond diagnosis, the project provided precise prescriptions for resource allocation:

- Service Optimization Models determined that two baristas maximize annual profit for the coffee shops, and that the car rental agency should hire two additional employees to optimally balance labour costs against service failure penalties.
- Experimental Design (DOE) and ANOVA revealed that delivery performance is predominantly driven by Product Type, proving that operational issues are not uniform and require targeted, product-specific interventions rather than one-size-fits-all solutions.
- The foundational Descriptive Statistics and Data Correction ensured the integrity of all analyses, from customer segmentation to financial calculations.

11.1 CONSOLIDATED ACTION PLAN

To translate these evidence-based findings into tangible improvement, the following prioritized action plan is recommended:

1. **Emergency Process Intervention:**

   - Immediately launch a root-cause analysis into the Laptop (LAP) delivery process, the most unstable product line, to address the special causes of variation identified in samples 185-360.
   - Initiate a cross-functional project to reduce variability in the warehouse picking process, as it is the confirmed bottleneck driving delivery inconsistencies.

2. **Implement Profit-Driven Resource Models:**

- Adopt the staffing level of two baristas in the coffee shops and hire two additional employees at the car rental agency, as per the optimization models, to maximize profit and minimize total cost.

3. **Data-Driven, Targeted Improvement**:

   - Use the DOE findings to conduct controlled experiments on picking procedures for specific underperforming product types (e.g., KEY), focusing improvement efforts where they will have the greatest impact.
   - Formalize the data validation procedure used in this project to prevent future discrepancies between head-office and local data systems.

4. **Enhance Proactive Monitoring**:

   - Maintain the implemented SPC charts for ongoing monitoring of delivery times.
   - Review the control chart rules for critical processes where the risk of missing a small process shift (Type II error) is unacceptably high.

In summary, this project demonstrates the power of a systematic, quantitative approach to engineering management. It moves from observation to diagnosis, and from diagnosis to prescribed solutions, providing a clear blueprint for enhancing process capability, optimizing resource allocation, and building a more reliable and profitable operation.

APPENDIX A — FIGURES & TABLES

List of Figures

List of Tables

13. References

- R-project, 2015. *Introduction_to_MANOVA.RM.knit.* [Online] Available at: https://cran.r-project.org/web/packages/MANOVA.RM/vignettes/Introduction_to_MANOVA.RM.html [Accessed 2025].

- sthda, 2022. *MANOVA Test in R: Multivariate Analysis of Variance - Easy Guides - Wiki - STHDA.* [Online] Available at: http://www.sthda.com/english/wiki/manova-test-in-r-multivariate-analysis-of-variance

- Engineering Counsel of South Africa, 2025. *ProjectECSA2025Final* [pdf] Stellenbosch University. Available atProjectECSA2025Final.pdf [Accessed 25 October 2025].

- Stellenbosch University, 2025. *QA344 Formula Page*. [pdf] Stellenbosch University. Available at: QA344FormulaPage.pdf [Accessed 25 October 2025].

- Stellenbosch University, 2025. *SPC Basics*. [pdf] Stellenbosch University. Available at: SPCBasics.pdf [Accessed 24 October 2025].
- Montgomery, D.C. (2019). *Introduction to Statistical Quality Control*. 8th ed. Hoboken, NJ: Wiley.

- Engineering Counsel of South Africa, 2025. *ProjectECSA2025Part1234567*. [pdf] Stellenbosch University. Available atProjectECSA2025Part1234567.pdf [Accessed 24 October 2025].

- Maphumulo, U. (2025) *Quality Assurance ECSA Report*. Unpublished manuscript, Industrial Engineering, Stellenbosch University.

- Stellenbosch University (2025) *QA344 Statistics*. Unpublished course notes, Industrial Engineering, Stellenbosch University.

- Stellenbosch University (2025) *Understanding Data, populations, samples and SPC*. Available at: https://stemlearn.sun.ac.za/ (Accessed: 28 October 2025).