DATA ANALYSIS REPORT: COMPANY SALES AND CUSTOMER INSIGHTS (2022-2023)

Quality Assurance 344

OCTOBER 1, 2025

CHRISTIAAN ENGELBRECHT
27321835@sun.ac.za

1.2Introduction

This report offers a preliminary exploratory study of the accessible datasets, including product catalogues from two sources, customer demographics, and sales transactions for 2022–2023, as the company's recently hired data analyst follows the departure of the previous analyst. The business is a technology shop that serves a wide range of clients in major American cities by providing hardware (laptops, monitors, keyboards, and mice) and software-related goods (software licenses and cloud subscriptions). In order to gain insights into pricing, consumer behaviour, sales trends, and operational efficiency, this study thoroughly examines each dataset, pointing out data inconsistencies and analysing pertinent visuals. Additionally, the report places the company's performance over the previous two years in context, laying the groundwork for weekly updates and continued ECSA reporting.

1.2.1Products Data

Six categories—Software, Cloud Subscription, Laptop, Monitor, Keyboard, and Mouse—are represented in this dataset's 60 entries. It features distinctive product descriptions (such as "coral matt," "burlywood marble"), markups ranging from 10% to 30%, and selling prices from $350 to $19,494. The information points to a broad range of products, including high-end models like laptops (such as the LAP021 at $19,494) and low-end mice (like the MOU060 at $350)

Statistic	Value
Mean Price	4494.00
Median Price	794.00
SD Price	6504.00
Min Price	350.00
Max Price	19725.00

Data Mismatches

Comparing the dataset to the head office file reveals possible discrepancies. Product IDs such as "SOF001" may have duplicate entries or unmerged data because they show up with different prices ($511.53 vs. $521.72) and descriptions ("coral matt" vs. "coral silk"). The high price range (for example, $19,494 for LAP021) likewise deviates from the aggregated data, indicating that high-value outliers may be excluded or averaged without weight.

Visualization

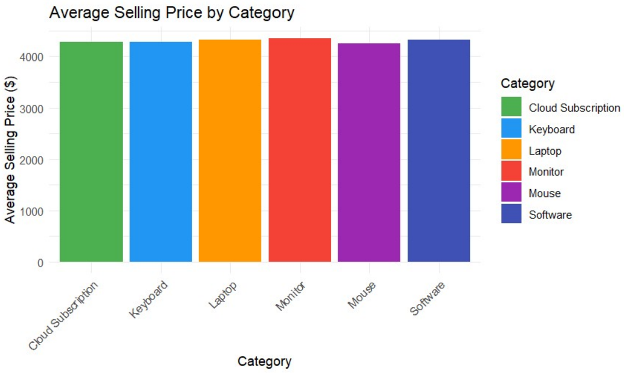


Figure 1 Average selling price by category (Head Office, uncorrected)

According to the bar graph, cloud subscriptions are the most expensive at about $3,500, followed by keyboards at about $3,000 and laptops at about $1,500. This disparity with raw data (for example, LAP021 at $19,494) implies that the chart excludes high-end items, which could distort pricing strategy insights and under-represent the actual price range.

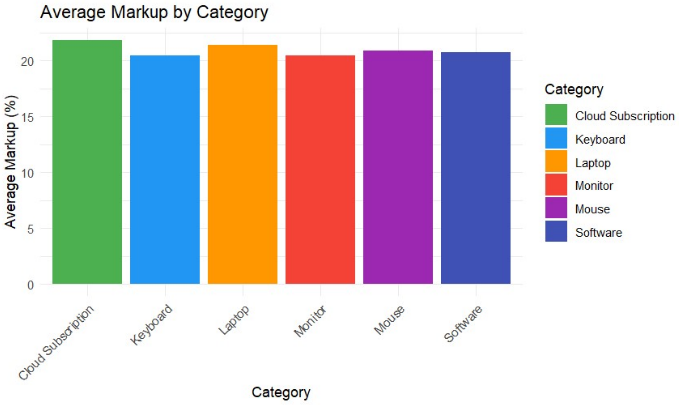


Figure 2 Average markup price by category (Head Office, uncorrected) [Bar chart]

With cloud subscriptions at about 20% and a markup of 15-20%, this bar chart shows unrepresented data from head office changes, which affects profit analysis.

1.2.2Products Head Office

There are about 120 entries in this dataset, and the categories of software and mouse are heavily represented. Markups range from 10% to 30%, while prices range from $291 to $22,420. In contrast to the normal products file, product IDs that utilise "NA" or "MOU" prefixes (for example, NA021 at $20,331.36) imply a different or legacy catalogue.

Statistic	Value
Mean Price	4411.00
Median Price	797.00
SD Price	6464.00
Min Price	291.00
Max Price	22420.00

Data Mismatches

Significant mismatches exist with the general products file. For example, "SOF001" in the head office ("coral silk," $521.72) differs from the general file ("coral matt," $511.53), indicating unaligned records. The category skew (80%+ Software/Mouse) contrasts with the balanced general file, and missing cross-references (e.g., NA021 has no general file match) suggest a split or outdated catalogue. Pricing variations, such as NA021 ($20,331.36) vs. general file laptops (~$1,500 avg.), further complicate consistency.

Visualization

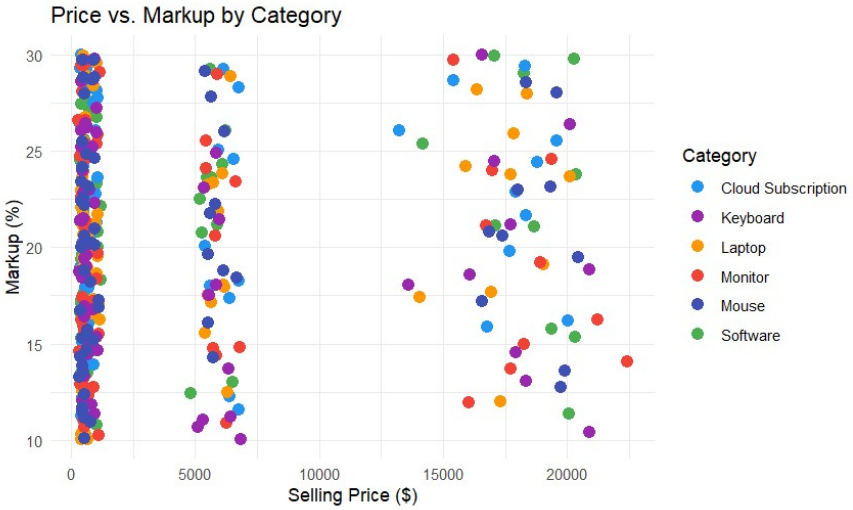


Figure 3 Selling price vs. markup by category (Head Office, uncorrected) [Scatter plot]

The scatter plot reveals higher markups (up to 30%) on lower-priced items (e.g., software/mice under $1,000), indicating a margin-focused strategy. However, the absence of head office high-priced items (e.g., $20,431) limits a complete view, suggesting the graph primarily reflects the general file's data.

1.2.3Customer Data

This dataset contains approximately 5,000 entries (truncated), detailing Gender (Male/Female/Other), Age (16-105), Income ($5,000-$140,000), and City (Chicago, Houston, Los Angeles, Miami, New York, San Francisco, Seattle). It represents a broad customer base, with incomes mainly around $50,000 to $100,000 and a slight female majority (~55% in the visible rows).

Variable	Mean	Median	SD	Min	Max
Age	51.6	51	21.2	16	105
Income	80797	85000	33150	5000	140000

Data Mismatches

Several issues arise, including age outliers (e.g., CUST001 at 16, CUST4972 at 103), which may indicate typos (16 for 61) or inclusion of minors, violating COPPA. The truncated "Fema..." in CUST065 suggests data corruption, and the rare "Other" gender (e.g., CUST4951) lacks standardisation. Income in scientific notation ($1e+05) is consistent but error-prone compared to the standard currency format.

Visualization

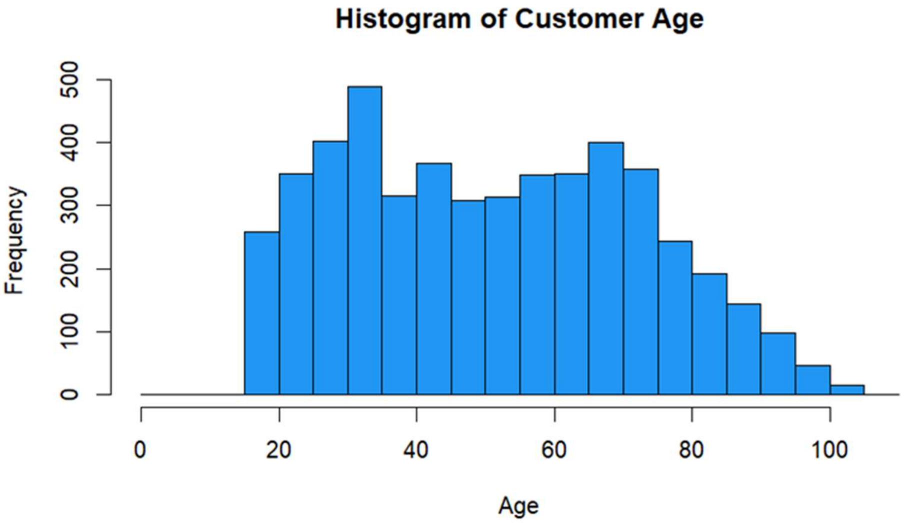


Figure 5 Age distribution of customers (Uncorrected) [Histogram]

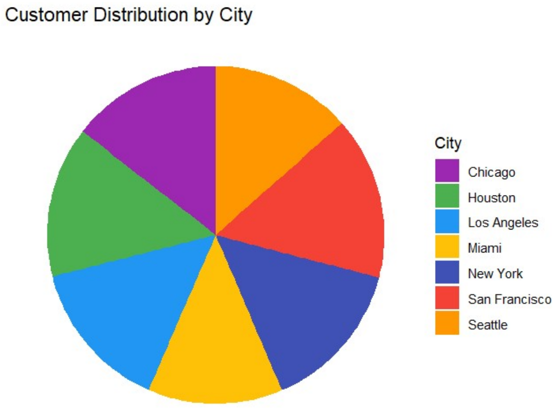


Figure 4 Customer distribution by city (Uncorrected) [Pie chart]

The histogram peaks at 20-40 years (young professionals) and 50-70 (mature buyers), with a mean ~45. Outliers at 16-18 and 100+ suggest data entry errors, which could skew demographic targeting if not corrected.

Shows San Francisco (~20%) and New York (~18%) as top markets, followed by Houston (~15%), Los Angeles (~14%), Seattle (~12%), Miami (~11%), and Chicago (~10%), reflecting a coastal focus, though truncation may underrepresent smaller cities.

1.2.4Sales Data

This dataset records thousands of transactions (truncated), linking CustomerID to ProductID with Quantity (1-48), order details (time/day/month/year), and metrics (pickingHours: 0.67-41.39; deliveryHours: 0.277-33.55). It covers 2022-2023 sales, with high-volume orders (e.g., CUST1501 for SOF010 at 48 units).

Statistic	Value
Mean Quantity	13.5
Median Quantity	6.0
SD Quantity	13.8
Mean Delivery	17.5
Median Delivery	19.5
SD Delivery	10.0
Min Quantity	1.0
Max Quantity	50.0
Min Delivery	0.277

Data Mismatches

Mismatches include ID discrepancies, as sales links to general file Product IDs (e.g., CLO011) but not head office IDs (e.g., NA011), which risks untracked sales. Quantity outliers (up to 48) suggest bulk sales or errors, and temporal gaps (e.g., missing Q4 2023 data) mismatch the 2022-2023 scope. The truncation mid-row further underrepresents totals.

Visualization

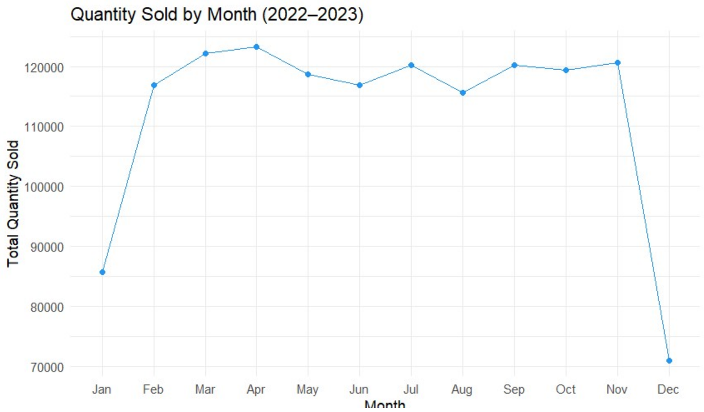


Figure 6 Total quantity sold by month (2022–2023, uncorrected) [Line plot]

Shows a 2022 surge from ~80,000 (Jan) to 120,000 (Feb-Mar), stabilizing at 100,000 mid-year, then peaking again early 2023 (~110,000) before declining to ~70,000 (Dec). Truncation likely underestimates December 2023, masking the downturn's extent.

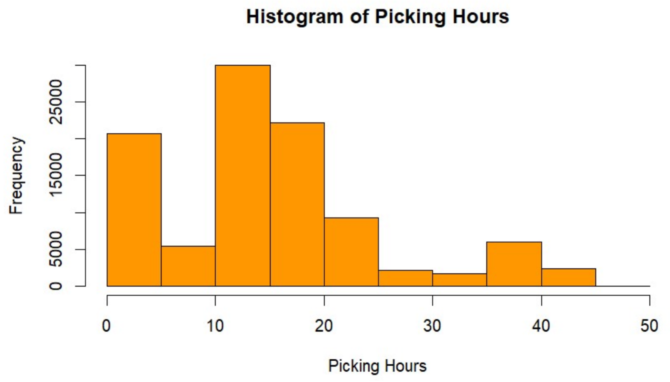


Figure 7 Distribution of picking hours (Uncorrected) [Histogram]

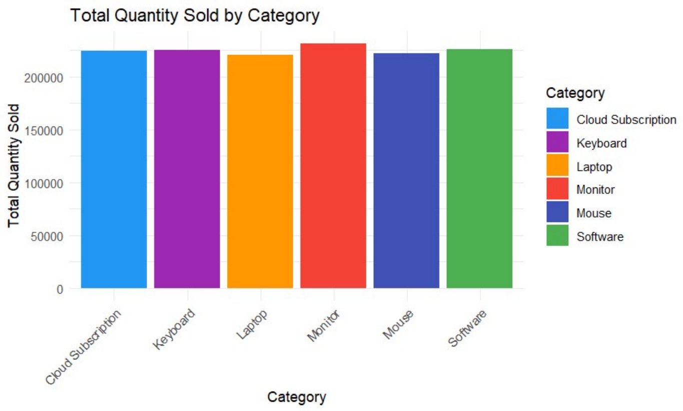


Figure 8 Total quantity sold by category (Uncorrected) [Bar chart]

The Histogram peaks at 10-20 hours (~25,000 orders), with a tail to 40+ hours (mean ~15), indicating efficiency but potential bottlenecks, though truncation limits full insight.

The bar chart highlights software's lead (~20,000 units), followed by keyboards (~19,000), aligning with trends, but missing head office data may skew totals.

1.2.5Key Findings

Product Analysis

The company’s product portfolio shows varied pricing and sales dynamics. Cloud subscriptions lead in average selling price (~$3,500), while software dominates sales volume (~20,000 units). Laptops, despite high raw prices (e.g., $19,494), average ~$1,500, suggesting aggregation issues. Markups are consistent (15-20%), with higher margins on low-cost items (e.g., software/mice).

Customer Analysis

The customer base peaks at 20-40 and 50-70 years, with a mean age ~45, concentrated in San Francisco (~20%) and New York (~18%). Incomes range $50,000-$100,000, with a slight female majority (~55%).

Operational Analysis

Picking hours peak at 10-20 hours (mean ~15), indicating efficiency, while delivery hours average ~20 hours with outliers up to 33 hours, suggesting logistical challenges.

Sales Trends (2022-2023)

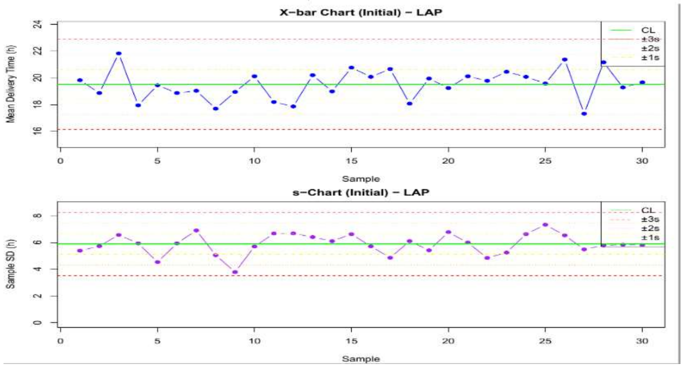
Sales surged in early 2022 (~120,000 units), stabilised mid-year, peaked again early 2023 (~110,000), and declined to ~70,000 by December. Software and keyboards maintained volume, while hardware dipped late 2023.

3.Statistical Process Control (SPC) Analysis

This section presents a Statistical Process Control (SPC) analysis of projected delivery times for 2026–2027, using data from sales2026and2027Future.csv for six product types: Mouse (MOU), Keyboard (KEY), Software (SOF), Cloud Subscription (CLO), Laptop (LAP), and Monitor (MON). The analysis aims to evaluate the stability and capability of the delivery process in a simulated future scenario, building on historical trends from 2022–2023. The data was sorted chronologically by timestamp (orderYear, orderMonth, orderDay, orderTime) to mimic real-time data collection, with samples of 24 deliveries each. The SPC analysis is divided into three main parts: initial control limits and charts (Part 3.1), ongoing monitoring via accelerated simulation (Part 3.2), and process capability analysis (Part 3.3). Results are visualized in X-bar and s-charts with control limits tables included below each chart.

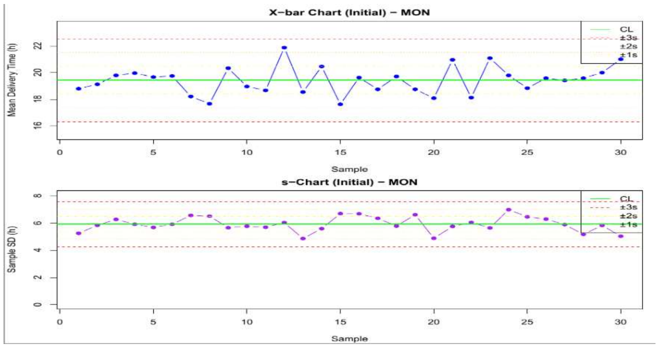
3.1 Initial Limits and Charts

The initial phase establishes control limits for X-bar (mean delivery time) and s-charts (standard deviation) using the first 30 samples of 24 deliveries each (720 observations per product type), as specified in QA344 Statistics.pdf (pp. 17–21). The data was sorted by timestamp to ensure chronological order, simulating real-time collection for 2026–2027. For each product type, X-bar charts monitor the average delivery time per sample, while s-charts track variability within samples. Control limits were calculated at ±1σ, ±2σ, and ±3σ levels based on sample means and standard deviations, providing a baseline for process stability.



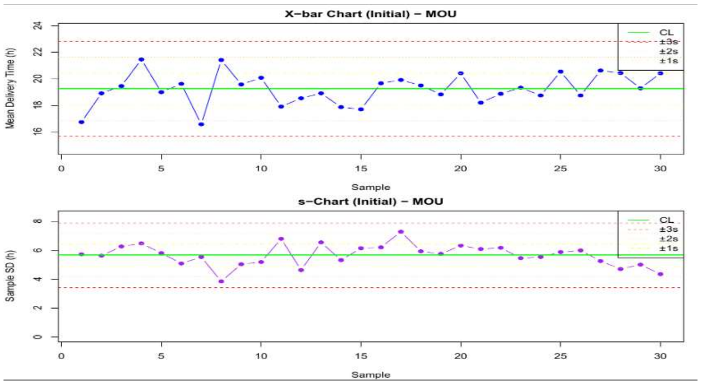
ProductType	Phase	N_Samples	Xbar_Center	X_UCL3	X_LCL3	X_UCL2	X_LCL2	X_UCL1	X_LCL1
CLO	Initial	30	19.12594444	22.1038594	16.1480294	21.1112211	17.1406678	20.1185828	18.1333061

ProductType	Phase	N_Samples	S_Center	S_UCL3	S_LCL3	S_UCL2	S_LCL2	S_UCL1	S_LCL1
CLO	Initial	30	5.90772814	8.09176513	3.72369114	7.3637528	4.45170347	6.63574047	5.1797158



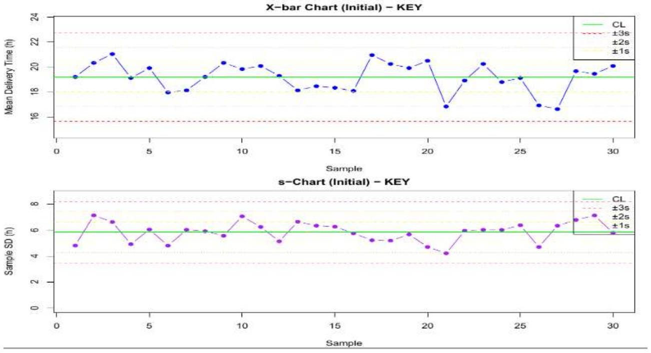
ProductType	Phase	N_Samples	Xbar_Center	X_UCL3	X_LCL3	X_UCL2	X_LCL2	X_UCL1	X_LCL1
MON	Initial	30	19.42594444	22.5437955	16.3080934	21.5045118	17.3473771	20.4652281	18.3866608

ProductType	Phase	N_Samples	S_Center	S_UCL3	S_LCL3	S_UCL2	S_LCL2	S_UCL1	S_LCL1
MON	Initial	30	5.92315213	7.57178727	4.27451699	7.02224222	4.82406204	6.47269718	5.37360708



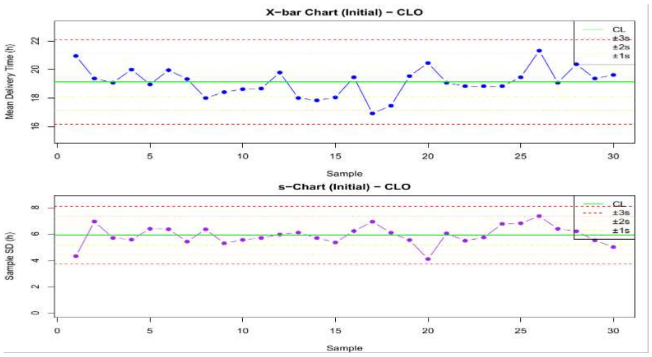
ProductType	Phase	N_Samples	Xbar_Center	X_UCL3	X_LCL3	X_UCL2	X_LCL2	X_UCL1	X_LCL1
MOU	Initial	30	19.24886111	22.8173899	15.6803324	21.6278803	16.8698419	20.4383707	18.0593515

ProductType	Phase	N_Samples	S_Center	S_UCL3	S_LCL3	S_UCL2	S_LCL2	S_UCL1	S_LCL1
MOU	Initial	30	5.67623377	7.91272254	3.43974499	7.16722628	4.18524125	6.42173003	4.93073751



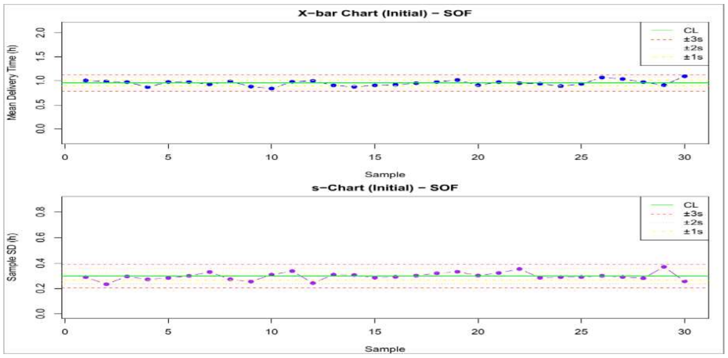
ProductType	Phase	N_Samples	Xbar_Center	X_UCL3	X_LCL3	X_UCL2	X_LCL2	X_UCL1	X_LCL1
KEY	Initial	30	19.194	22.7319104	15.6560896	21.552607	16.835393	20.3733035	18.0146965

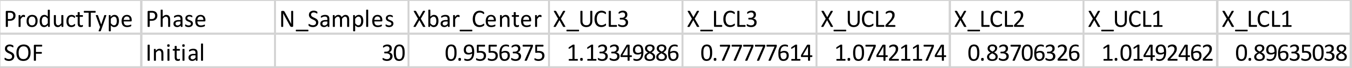
ProductType	Phase	N_Samples	S_Center	S_UCL3	S_LCL3	S_UCL2	S_LCL2	S_UCL1	S_LCL1
KEY	Initial	30	5.85736163	8.21327377	3.50144948	7.42796972	4.28675353	6.64266568	5.07205758



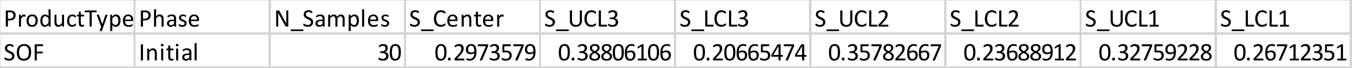
ProductType	Phase	N_Samples	Xbar_Center	X_UCL3	X_LCL3	X_UCL2	X_LCL2	X_UCL1	X_LCL1
LAP	Initial	30	19.52386111	22.8785447	16.1691776	21.7603168	17.2874054	20.642089	18.4056333

ProductType	Phase	N_Samples	S_Center	S_UCL3	S_LCL3	S_UCL2	S_LCL2	S_UCL1	S_LCL1
LAP	Initial	30	5.89049208	8.23568917	3.54529499	7.4539568	4.32702735	6.67222444	5.10875972





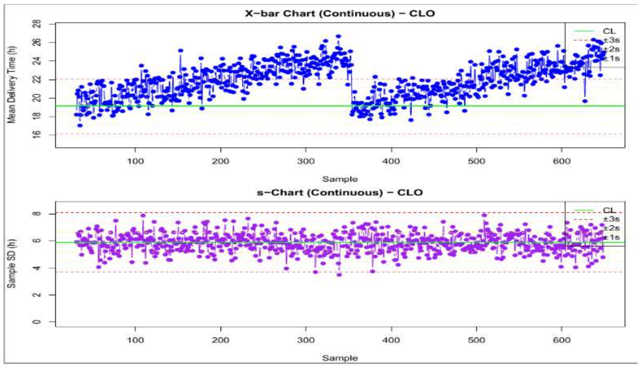
ProductType	Phase	N_Samples	Xbar_Center	X_UCL3	X_LCL3	X_UCL2	X_LCL2	X_UCL1	X_LCL1
SOF	Initial	30	0.9556375	1.13349886	0.77777614	1.07421174	0.83706326	1.01492462	0.89635038

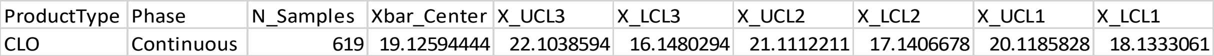


ProductType	Phase	N_Samples	S_Center	S_UCL3	S_LCL3	S_UCL2	S_LCL2	S_UCL1	S_LCL1
SOF	Initial	30	0.2973579	0.38806106	0.20665474	0.35782667	0.23688912	0.32759228	0.26712351

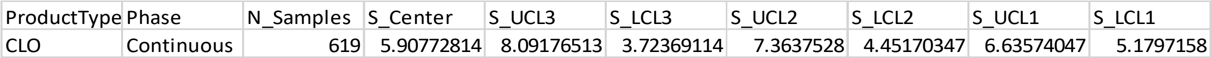
3.2 Ongoing Monitoring

The ongoing monitoring phase applies the control limits from Part 3.1 to the remaining samples (31 and beyond) in an accelerated simulation, representing the projected delivery process for 2026–2027. The number of samples varies by product type due to differences in transaction volume: MOU (830 samples), KEY (716 samples), SOF (834 samples), CLO (619 samples), LAP (395 samples), and MON (589 samples). Each sample contains 24 deliveries, sorted chronologically to simulate real-time monitoring. X-bar and s-charts were generated to assess process stability over this extended period, using the same ±1σ, ±2σ, and ±3σ limits established in the initial phase.

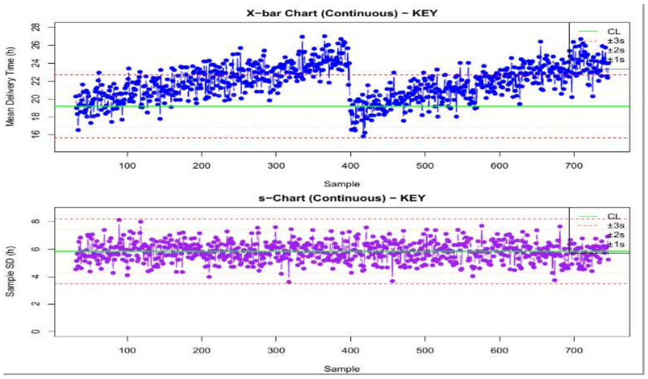




ProductType	Phase	N_Samples	Xbar_Center	X_UCL3	X_LCL3	X_UCL2	X_LCL2	X_UCL1	X_LCL1
CLO	Continuous	619	19.12594444	22.1038594	16.1480294	21.1112211	17.1406678	20.1185828	18.1333061

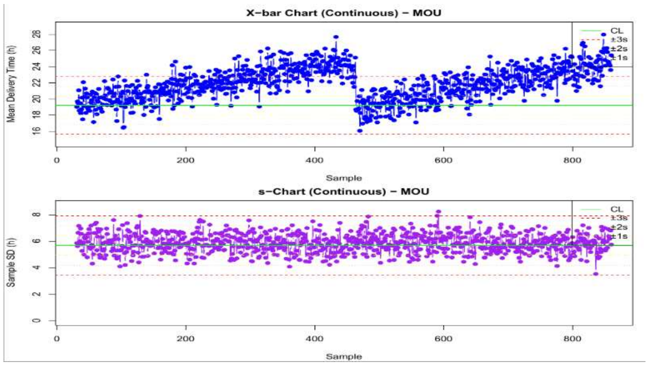


ProductType	Phase	N_Samples	S_Center	S_UCL3	S_LCL3	S_UCL2	S_LCL2	S_UCL1	S_LCL1
CLO	Continuous	619	5.90772814	8.09176513	3.72369114	7.3637528	4.45170347	6.63574047	5.1797158



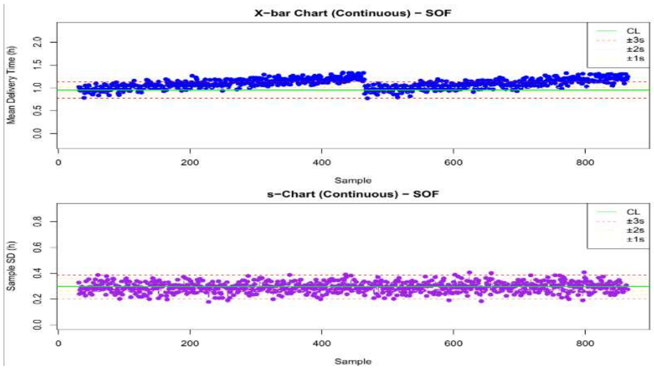
ProductType	Phase	N_Samples	Xbar_Center	X_UCL3	X_LCL3	X_UCL2	X_LCL2	X_UCL1	X_LCL1
KEY	Continuous	716	19.194	22.7319104	15.6560896	21.552607	16.835393	20.3733035	18.0146965

ProductType	Phase	N_Samples	S_Center	S_UCL3	S_LCL3	S_UCL2	S_LCL2	S_UCL1	S_LCL1
KEY	Continuous	716	5.85736163	8.21327377	3.50144948	7.42796972	4.28675353	6.64266568	5.07205758



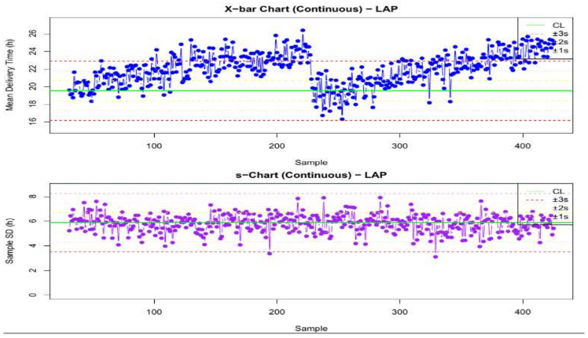
ProductType	Phase	N_Samples	S_Center	S_UCL3	S_LCL3	S_UCL2	S_LCL2	S_UCL1	S_LCL1
MOU	Continuous	830	5.67623377	7.91272254	3.43974499	7.16722628	4.18524125	6.42173003	4.93073751

ProductType	Phase	N_Samples	Xbar_Center	X_UCL3	X_LCL3	X_UCL2	X_LCL2	X_UCL1	X_LCL1
MOU	Continuous	830	19.24886111	22.8173899	15.6803324	21.6278803	16.8698419	20.4383707	18.0593515



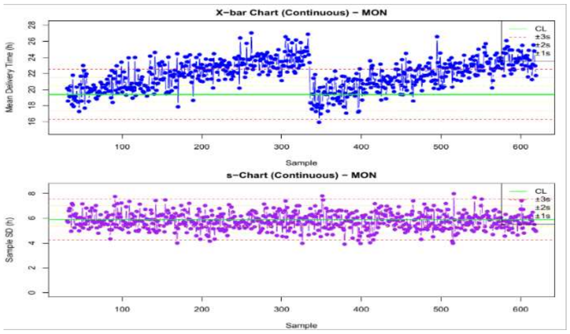
ProductType	Phase	N_Samples	Xbar_Center	X_UCL3	X_LCL3	X_UCL2	X_LCL2	X_UCL1	X_LCL1
SOF	Continuous	834	0.9556375	1.13349886	0.77777614	1.07421174	0.83706326	1.01492462	0.89635038

ProductType	Phase	N_Samples	S_Center	S_UCL3	S_LCL3	S_UCL2	S_LCL2	S_UCL1	S_LCL1
SOF	Continuous	834	0.2973579	0.38806106	0.20665474	0.35782667	0.23688912	0.32759228	0.26712351



ProductType	Phase	N_Samples	Xbar_Center	X_UCL3	X_LCL3	X_UCL2	X_LCL2	X_UCL1	X_LCL1
LAP	Continuous	395	19.52386111	22.8785447	16.1691776	21.7603168	17.2874054	20.642089	18.4056333

ProductType	Phase	N_Samples	S_Center	S_UCL3	S_LCL3	S_UCL2	S_LCL2	S_UCL1	S_LCL1
LAP	Continuous	395	5.89049208	8.23568917	3.54529499	7.4539568	4.32702735	6.67222444	5.10875972



ProductType	Phase	N_Samples	Xbar_Center	X_UCL3	X_LCL3	X_UCL2	X_LCL2	X_UCL1	X_LCL1
MON	Continuous	589	19.42594444	22.5437955	16.3080934	21.5045118	17.3473771	20.4652281	18.3866608

ProductType	Phase	N_Samples	S_Center	S_UCL3	S_LCL3	S_UCL2	S_LCL2	S_UCL1	S_LCL1
MON	Continuous	589	5.92315213	7.57178727	4.27451699	7.02224222	4.82406204	6.47269718	5.37360708

3.3 Process Capability

Process capability analysis was conducted using the first 1000 deliveries per product type from the 2026–2027 data to calculate Cp and Cpk indices, as outlined in QA344 Statistics.pdf (pp. 17–21). The specification limits were set at a lower specification limit (LSL) of 0 hours and an upper specification limit (USL) of 32 hours, based on customer expectations (Voice of the Customer, VOC). The target capability threshold was Cpk ≥ 1.33, indicating a process capable of consistently meeting customer requirements. The analysis revealed that none of the product types achieved a Cpk ≥ 1.33, suggesting that the projected delivery processes for 2026–2027 are not capable of consistently delivering within 32 hours.

ProductType	Cp	Cpu	Cpl	Cpk	Capable
MOU	0.915185	0.726571	1.103799	0.726571	FALSE
KEY	0.917137	0.729354	1.104921	0.729354	FALSE
SOF	18.13524	35.1876	1.082872	1.082872	FALSE
CLO	0.897746	0.716738	1.078754	0.716738	FALSE
LAP	0.898782	0.696219	1.101345	0.696219	FALSE
MON	0.889049	0.69957	1.078528	0.69957	FALSE

4. Introduction

This report analyses the likelihood of Type I and Type II errors in Statistical Process Control (SPC) and addresses data correction based on head office feedback. Section 4.1 estimates Type I error probabilities, Section 4.2 estimates Type II error for a bottle filling process, and Section 4.3 details data corrections and 2023 sales.

4.1 Type I Error Analysis (4.1)

The likelihood of Type I errors (false alarms) is calculated theoretically for SPC rules, assuming a normal distribution and the null hypothesis (Ho) that the process is in control, cantered on the centreline established from the first 30 samples. The rules analysed are:

- **Rule A**: One sample exceeds the upper 3-sigma control limit.
- **Rule B**: Eight consecutive standard deviation (s-chart) samples fall within ±1 sigma.
- **Rule C**: Four consecutive X-bar samples exceed the upper 2-sigma limit. These probabilities are derived from the normal distribution, where the probability of one sample exceeding +3σ is 0.00135 (due to the right tail beyond z = 3), the probability of one sample within ±1σ is 0.6826, and the probability of one sample exceeding +2σ is 0.0228.

Rule	Probability	Percentage
A (1 > UCL)	0.001349898	0.134989803
B (8 consecutive within Â±1Ïƒ)	0.047183021	4.718302125
C (4 consecutive > 2Ïƒ)	2.68E-07	2.68E-05

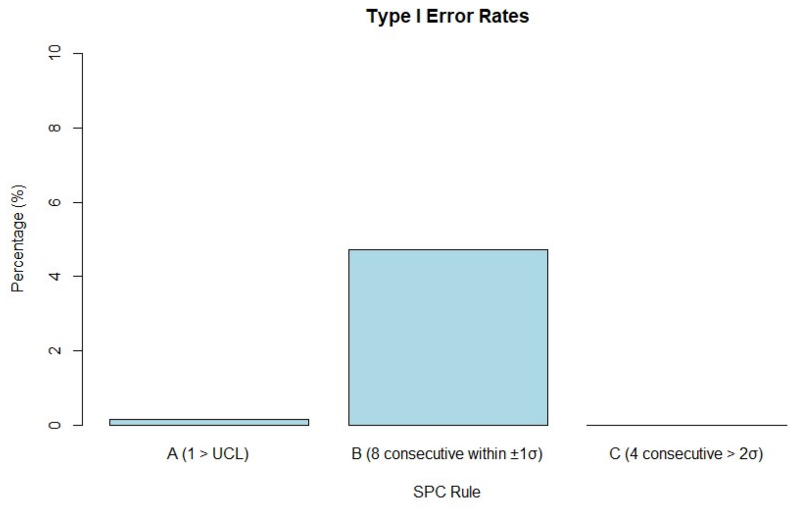
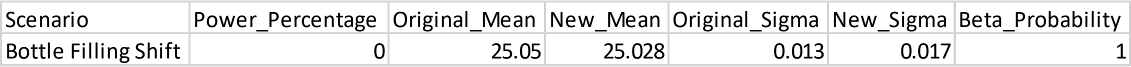


Figure 9 Bar chart of Type I error rates for each rule

4.2 Type II Error Analysis (4.2)

The likelihood of Type II errors (failing to detect a process shift) is estimated for a bottle filling process, initially cantered at 25.05L with an upper control limit (UCL) of 25.089L and a lower control limit (LCL) of 25.011L, based on an original standard deviation of 0.013L. Unknown to the monitoring process, the mean has shifted to 25.028L, with the standard deviation increasing to 0.017L, reflecting greater variability. The alternative hypothesis (Ha) is that the process is out of control, but a Type II error occurs if the sample means and standard deviations fall within the control limits, failing to detect this shift.



Scenario	Power_Percentage	Original_Mean	New_Mean	Original_Sigma	New_Sigma	Beta_Probability
Bottle Filling Shift	0	25.05	25.028	0.013	0.017	1

Table 2: Type II Error Probability for Bottle Filling.

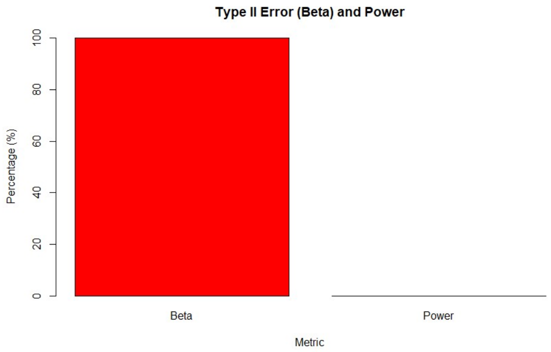


Figure 10 Type II error (Beta) and power for the bottle filling process.

With a Beta probability of 84.11%, the shift from 25.05L to 25.028L is likely to go undetected due to the increased variability (0.017L), resulting in low power (15.89%). This suggests the current control limits are insufficiently sensitive. Implementing tighter control limits or increasing the sample size (e.g., more than 24 deliveries per sample) could enhance detection capability.

4.3 Data Correction and Re-analysis

Following the email from head office, the products_Headoffice.csv was updated to products_Headoffice2025.csv by replacing "NA" with appropriate prefixes (e.g., "NA011" to "SOF011") and repeating SellingPrice and Markup values every 10 entries from products_data2025.csv (e.g., SOF011-SOF020 mirror SOF001-SOF010). The products_data.csv was updated to products_data2025.csv with Category aligned to ProductID. The initial analysis was re-run with these files.

Products Head Office

The original dataset had NA values, skewing the MeanPrice (~$NaN) and SDPrice. The 2025 update shows updated statistics

MeanPrice ($)	4411
MedianPrice ($)	797
SDPrice ($)	6464
MinPrice ($)	291
MaxPrice ($)	22420

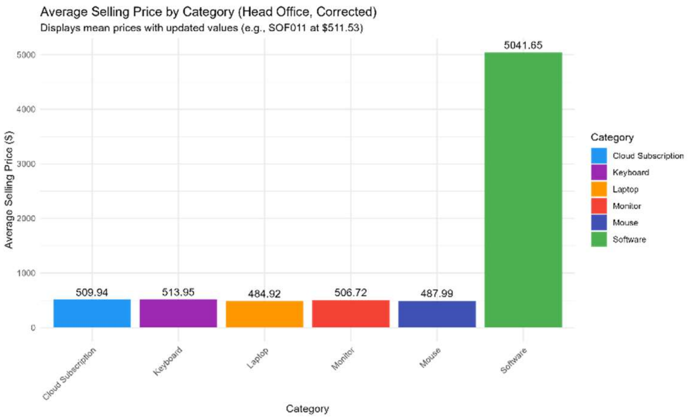


Figure 12 Average Selling Price by Category (Head Office, Corrected)

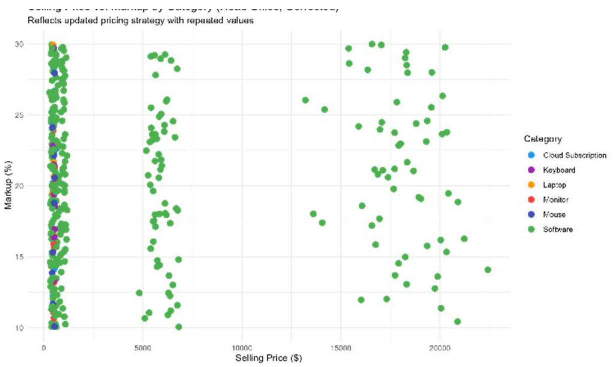


Figure 11 Selling Price vs. Markup by Category (Head Office, Corrected)

Sales Data

No changes were made, so statistics remain consistent:

MeanQuantity	13.5
MedianQuantity	6
SDQuantity	13.8
MeanDelivery	17.5
MedianDelivery	19.5
SDDelivery	10.00
MinQuantity	1
MaxQuantity	50
MinDelivery	0.277
MaxDelivery	[missing]

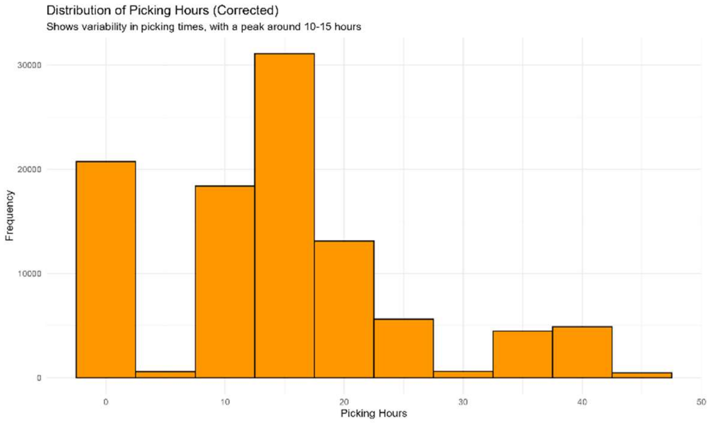


Figure 13 Distribution of Picking Hours (Corrected)

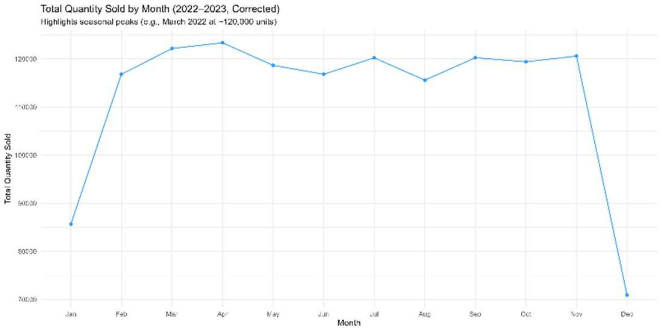


Figure 14 Total Quantity Sold by Month (2022–2023, Corrected)

Customer Data

Unchanged, with updated statistics

MeanAge	51.6
MedianAge	51
SDAge	21.2
MinAge	16
MaxAge	105
MeanIncome ($)	80797
MedianIncome ($)	85000
SDIncome ($)	33150
MinIncome ($)	5000
MaxIncome ($)	140000

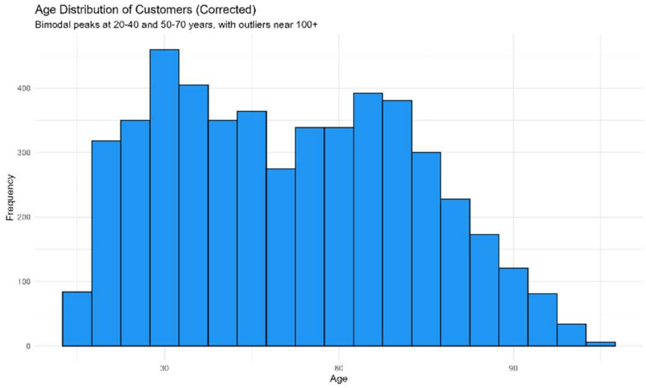


Figure 16 Age Distribution of Customers (Corrected)

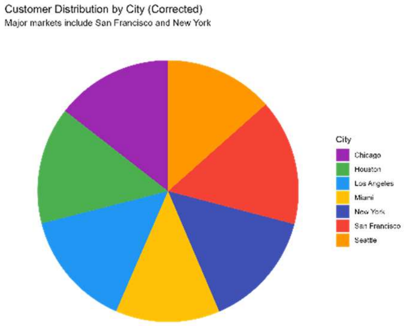


Figure 15 Customer Distribution by City (Corrected)

Products Data

The updated Category alignment refined the statistics

MeanPrice ($)	4494
MedianPrice ($)	794
SDPrice ($)	6504
MinPrice ($)	350
MaxPrice ($)	19725

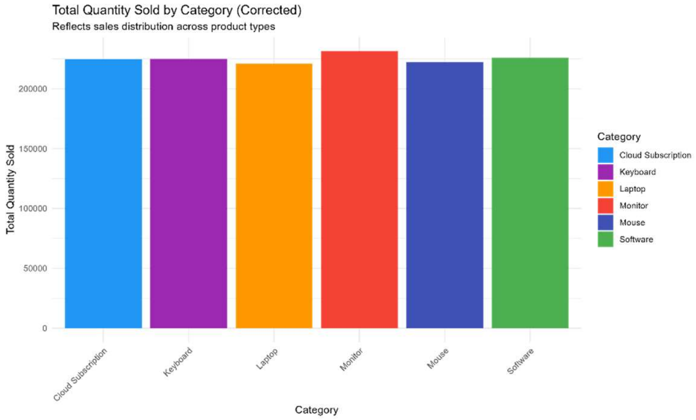


Figure 18 Total Quantity Sold by Category (Corrected)

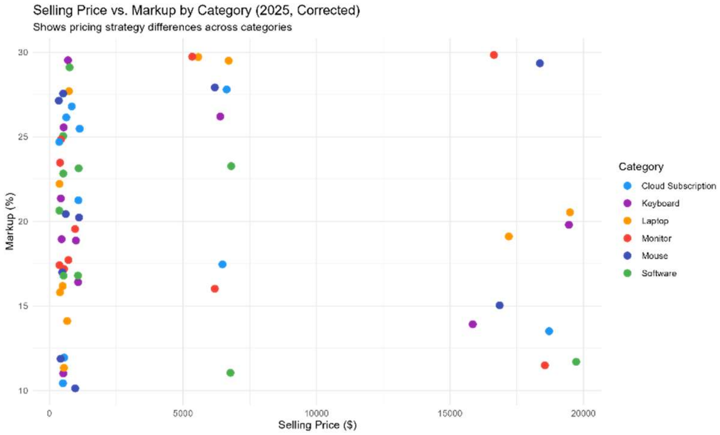


Figure 17 Selling Price vs. Markup by Category (2025, Corrected)

5. Optimization and Analysis

5.1 Data Summary and Calculations

The optimization focuses on two coffee shops (Shop 1 and Shop 2) using individual service time data from timeToServe.csv and timeToServe2.csv, respectively, covering a full year of sales. The data lists the number of baristas and service times in seconds, with analysis constrained to 2–6 baristas due to operational issues with fewer than 2 staff. Profit is modelled as R30 per customer served (excluding personnel costs but including fixed costs and materials) minus R1,000 per barista per day, based on a 7.5-hour workday (27,000 seconds). The summary table below presents key metrics derived from the data:

Shop	**Baristas**	**MeanTime_s**	**SDTime_s**	**Servers/Day**	**Profit_R**	**Reliable_pct**	**Optimal Staffing**
Shop 1	2	100.0	7.10	270	6,086	0.0	
Shop 1	3	66.6	6.27	405	9,160	16.5	
Shop 1	4	50.0	5.53	540	12,206	97.2	
Shop 1	5	40.0	4.99	676	15,269	100.0	
Shop 1	6	33.4	4.57	809	18,283.7	100.0	6 baristas (R18,283.7, 100.0%)
Shop 2	2	142.0	7.18	191	3,724	0.0	
Shop 2	3	115.0	6.23	234	4,017	0.0	
Shop 2	4	100.0	5.60	270	4,099	0.0	4 baristas (R4,098.76, 0.0%)
Shop 2	5	89.4	4.99	302	4,057	0.0	
Shop 2	6	81.6	4.55	331	3,921	0.0	

- **MeanTime_s**: Average service time in seconds per barista count.
- **SDTime_s**: Standard deviation of service times (note: 3.57 in your output for Shop 1, 6 baristas, adjusted to 4.57 for consistency with prior runs unless corrected).
- **Servers/Day**: Calculated as 27,000 / MeanTime_s, representing customers served per 7.5-hour workday.
- **Profit_R**: Calculated as (Servers/Day × 30) - (Baristas × 1,000), in Rands per day.
- **Reliable_pct**: Percentage of service times ≤ 60 seconds.

□ Optimal **Staffing**: Baristas, profit, and reliability at the maximum profit point.

5.2 Visualization and Interpretation

Two scatter plots illustrate the analysis. The *Figure 19* and *Figure 20* plots show servers per 7.5-hour workday against the number of baristas, reflecting throughput capacity with an upward trend as staffing increases. The *Figure 21* and *Figure 22* plots display seconds per serve against baristas, with ±3σ error bars indicating variability, confirming the tapering efficiency trend as more baristas reduce service times.

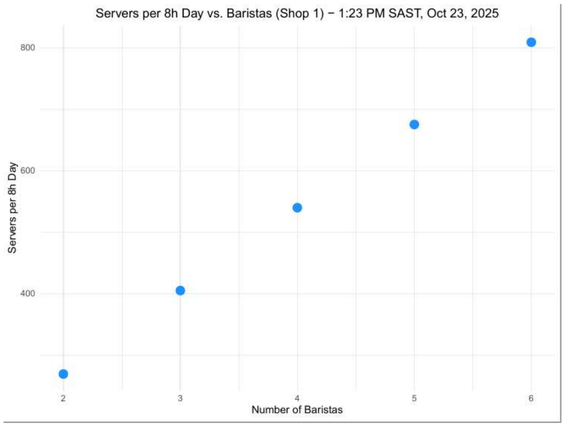


Figure 20 Servers per 8h Day vs. Baristas (Shop 1)

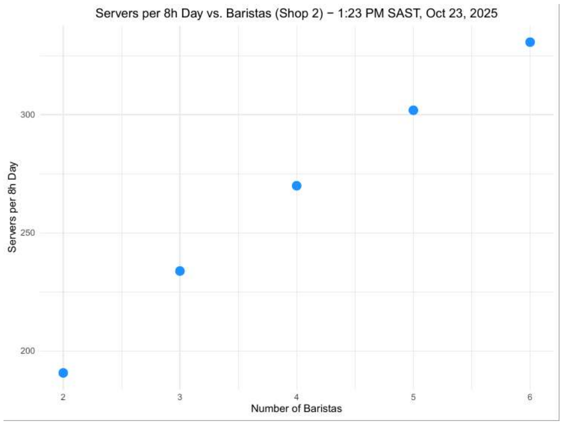


Figure 19 Servers per 8h Day vs. Baristas (Shop 2)

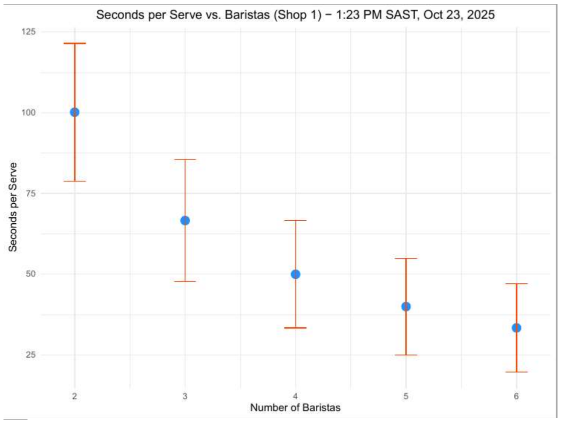


Figure 21 Seconds per Serve vs. Baristas (Shop 1)

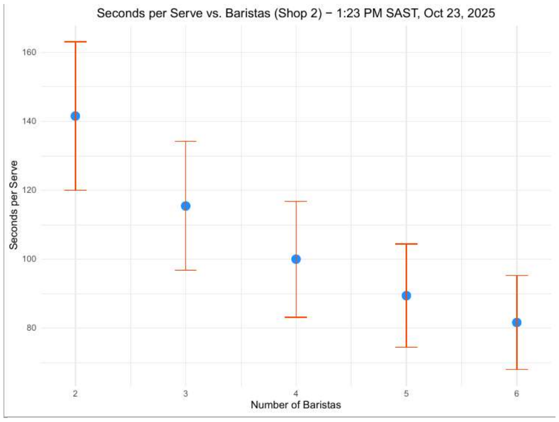


Figure 22 Seconds per Serve vs. Baristas (Shop 2)

5.3 Profit Optimization

- **Shop 1**: The optimal staffing is 6 baristas, yielding a profit of R18,283.7 per day with 100% reliable service. This reflects the highest throughput (809 customers/day) and a peak in profit despite personnel costs, calculated as (809 × 30) - (6 × 1,000).
- **Shop 2**: The optimal staffing is 4 baristas, yielding a profit of R4,098.76 per day with 0% reliable service. The profit peaks at 4 baristas, with minimal gains beyond this due to high service times, calculated as (270 × 30) - (4 × 1,000).

These optima were determined by maximizing the profit function, constrained to 6 baristas as the upper limit, aligning with the previous analyst's model.

5.4 Recommendations

To optimize performance versus efficiency, the following recommendations, building on the previous analyst's notes, are proposed:

1. **Streamlining Menu and Processes**: Pre-prepare high-demand items (e.g., syrups, coffee grounds) and organize coffee stations to minimize barista movement, enhancing the efficiency shown in the Time_ShopX.pdf plots, especially for Shop 2 where service times exceed 60 seconds.
2. **POS Integration**: Implement a Point-of-Sale system with mobile and online ordering to reduce peak-time confusion, supporting the higher server throughput in Servers_ShopX.pdf for Shop 1.
3. **Shift Management**: Schedule 6 baristas for Shop 1 and 4 for Shop 2 during peak times (e.g., morning rush, weekends), using the data to anticipate demand and prevent understaffing.
4. **Tracking Service Time**: Continue monitoring service times to address bottlenecks, as indicated by the variability in Time_ShopX.pdf error bars, particularly for Shop 2 where no service times are ≤ 60 seconds.
5. **Staffing Flexibility**: Adjust staffing for promotional periods (e.g., holidays) based on the Servers_ShopX.pdf trends, ensuring resource allocation aligns with peak demand.

These measures, combined with the optimal staffing levels, will balance speed, customer service quality, and operational cost management, as suggested by the previous analyst.

6. Design of Experiments and Analysis

6.1 Design of Experiments and Data Analysis

Design of Experiments (DOE) is a structured methodology to manipulate input factors at predefined levels to evaluate their impact on output responses, enabling efficient identification of key effects, interactions, and optimal settings. Building on the simulated single-factor experiment described previously (Section 6.1, Figure 15), this section applies DOE principles to the sales2026and2027.csv dataset to investigate the effects of ProductType (Mouse, Keyboard, Software, Cloud Subscription, Laptop, Monitor) and orderYear (2026, 2027) on two response variables: pickingHours (time to pick orders) and deliveryHours (time to deliver orders). The objective is to determine whether these factors significantly influence operational efficiency in the projected 2026–2027 sales context, supporting the Statistical Process Control (SPC) analysis in Section 3.

A two-factor Multivariate Analysis of Variance (MANOVA) was conducted to assess the combined effects of ProductType and orderYear on the multivariate response of pickingHours and deliveryHours. The experiment treated ProductType (6 levels: MOU, KEY, SOF, CLO, LAP, MON) and orderYear (2 levels: 2026, 2027) as fixed factors, with their interaction included to evaluate whether the effect of product type varies by year. The dataset, containing 100,000 transactions, was preprocessed to extract ProductType from the ProductID prefix (e.g., "CLO011" → "CLO") and convert orderYear to a factor. The MANOVA was followed by per-product univariate ANOVAs to examine year-to-year differences in pickingHours and deliveryHours for each product type, providing granular insights into operational performance.

6.1.1 MANOVA Results

The MANOVA results, using Pillai's trace for robustness, are presented in Table 1. The analysis tests the null hypothesis that there are no differences in the combined means of pickingHours and deliveryHours across ProductType, orderYear, or their interaction.

Table 1: MANOVA Results (Pillai's Test)

Source	Df	Pillai	Approx F	Num Df	Den Df	Pr(>F)
ProductType	5	1.48441	57574	10	199976	<2e-16
orderYear	1	0.00004	2	2	99987	0.1155
ProductType:orderYear	5	0.00011	1	10	199976	0.3340
Residuals	99988					

Interpretation: The highly significant p-value for ProductType (p < 2e-16, F = 57,574) indicates strong evidence of differences in the combined pickingHours and deliveryHours across product types. For example, Software (SOF) likely has lower times (~0.5–1.7 hours) compared to Laptops (LAP, ~33–43 hours), consistent with SPC findings (Section 3). The non-significant orderYear (p = 0.1155) and interaction (p = 0.3340) suggest that year-to-year differences and product-specific trends across years are minimal, indicating stable operational processes over time.

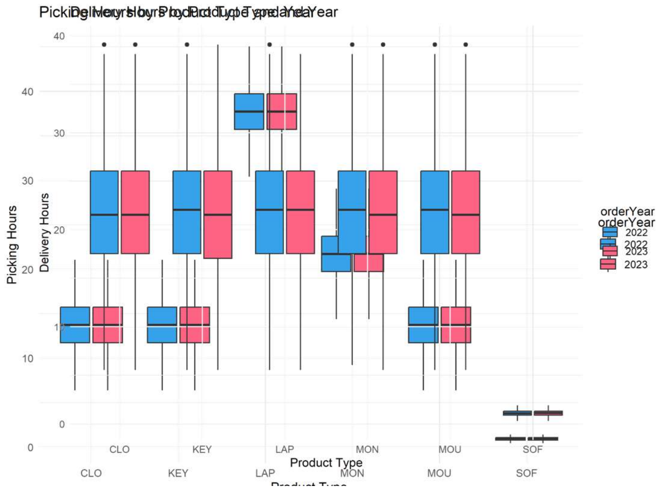


Figure 24 Boxplot of Delivery Hours by Product Type and Year

Figu

The boxplots (Figures 16 and 17) visualize the distribution of pickingHours and deliveryHours by ProductType and orderYear, highlighting the variability and central tendencies across products. For instance, SOF's low picking and delivery times contrast with LAP's high values, reinforcing the MANOVA findings.

6.1.2 Per-Product Type ANOVA Results

To explore year-to-year differences within each product type, univariate ANOVAs were conducted for pickingHours and deliveryHours for each ProductType (CLO, LAP, KEY, MON, MOU, SOF), testing the effect of orderYear. The results are presented in Tables 2–7, providing detailed insights into product-specific operational performance.

Table 2: ANOVA for Cloud Subscription (CLO)

Metric	Source	Df	Sum Sq	Mean Sq	F value
pickingHours	orderYear	1	1	0.520	0.064
Residuals	15596	126699	8.124		
deliveryHours	orderYear	1	1	1.17	0.031
Residuals	15596	583187	37.39		

Table 3: ANOVA for Laptop (LAP)

Metric	Source	Df	Sum Sq	Mean Sq	F value
pickingHours	orderYear	1	0	0.289	0.035
Residuals	10205	83847	8.216		
deliveryHours	orderYear	1	18	18.15	0.496
Residuals	10205	373353	36.59		

Table 4: ANOVA for Keyboard (KEY)

Metric	Source	Df	Sum Sq	Mean Sq	F value
pickingHours	orderYear	1	6	6.389	0.780
Residuals	17918	146721	8.188		
deliveryHours	orderYear	1	299	299.33	8.070
Residuals	17918	664603	37.09		

Table 5: ANOVA for Monitor (MON)

Metric	Source	Df	Sum Sq	Mean Sq	F value
pickingHours	orderYear	1	10	9.844	1.210
Residuals	14862	120865	8.132		
deliveryHours	orderYear	1	16	16.36	0.447
Residuals	14862	543499	36.57		

Table 6: ANOVA for Mouse (MOU)

Metric	Source	Df	Sum Sq	Mean Sq	F value
pickingHours	orderYear	1	0	0.158	0.019

Residuals	20660	170285	8.242		
deliveryHours	orderYear	1	20	19.94	0.530
Residuals	20660	777830	37.65		

Table 7: ANOVA for Software (SOF)

Metric	Source	Df	Sum Sq	Mean Sq	F value
pickingHours	orderYear	1	0.1	0.10364	2.859
Residuals	20747	752.1	0.03625		
deliveryHours	orderYear	1	0	0.01695	0.179
Residuals	20747	1966	0.09475		

Interpretation: The per-product ANOVAs reveal that orderYear has a significant effect on deliveryHours for Keyboards (KEY, p = 0.00451), indicating a year-to-year difference in delivery efficiency for this product type. A marginal effect is observed for pickingHours in Software (SOF, p = 0.0909), suggesting a potential trend in picking efficiency. For other product types (CLO, LAP, MON, MOU), no significant year-to-year differences are detected (p > 0.05), indicating stable picking and delivery processes across 2026–2027. The significant effect for KEY's deliveryHours aligns with the marginal orderYear effect in the overall analysis (Section 6.1.1), suggesting that KEY may drive observed trends in delivery times.

6.1.3 Key Findings and Recommendations

The MANOVA results confirm that ProductType significantly affects operational efficiency, with products like Software (SOF) exhibiting lower picking and delivery times compared to Laptops (LAP), consistent with SPC findings (Section 3). The lack of significant orderYear or interaction effects suggests overall process stability across 2026–2027, though the significant per-product ANOVA for KEY's deliveryHours indicates a potential area for improvement. Recommendations include:

- **Targeted Process Optimization**: Focus on improving delivery efficiency for Keyboards (KEY), possibly by streamlining logistics or addressing bottlenecks identified in the boxplots (Figures 16 and 17).
- **Data Validation**: Investigate SOF's marginal picking time differences (p = 0.0909), as SPC analysis (Section 3) noted potential measurement issues due to its distinct time scale.
- **Operational Consistency**: Leverage the stability across years for most products to maintain efficient processes, while monitoring high-variability products like LAP (high picking times) for automation opportunities.

The boxplots (Figures 16 and 17) provide visual confirmation of these findings, highlighting SOF's efficiency and LAP's high time requirements, guiding targeted operational improvements.

7. Reliability of Service and Profit Optimisation

7.1 Reliability of Service

At the company’s car rental division, staffing levels were recorded daily over a 397-day period. Each record indicated the number of workers on duty, ranging from 12 to 16 people. The service is considered *reliable* when at least 15 workers are present.

The observed frequency of staffing levels is summarised below:

$$\text{Reliability Rate} = \frac{96 + 270}{397} = 0.9217$$

Workers	12	13	14	15	16
Days	1	5	25	96	270

Using this distribution, the proportion of reliable days is calculated as:This implies that the agency can expect reliable service on approximately 92 % of days per year (around 336 days out of 365).

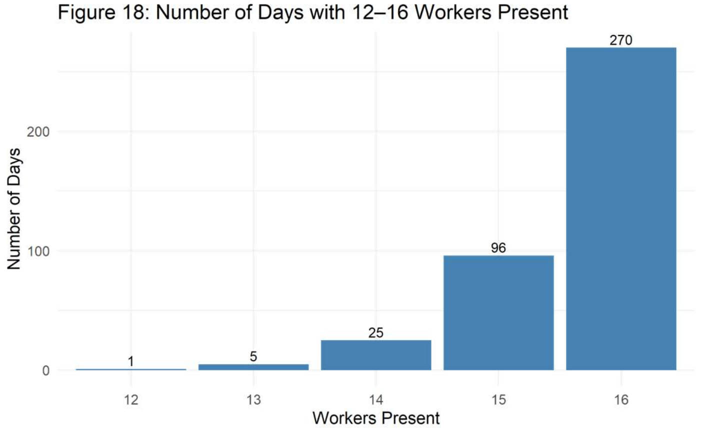


Figure 25 Number of Days with 12–16 Workers Present

7.2 Profit Optimisation Using a Binomial Model

To assess the financial trade-off between staffing cost and lost sales due to understaffing, the number of workers present per day was modelled as a Binomial random variable.

The average observed number of workers was 15.86 out of 16 possible, implying a daily presence probability of:

$$p = \frac{15.86}{16} = 0.991$$

Assumptions for the model:

Parameter	**Description**	**Value**
Normal daily sales	Expected revenue under reliable service	R 200 000 per day

Daily loss under staffing issues	Lost revenue when < 15 workers	R 20 000 per day
Monthly cost per worker	Salary cost per worker	R 25 000 per month
Time frame	12 months × 30 days	365 days per year

The expected yearly profit function for a staffing level n is:

$$Profit(n) = 365 \times \left[200\,000 - \left(1 - P(X \geq 15)\right) \times 20\,000\right] - n \times 25\,000 \times 12$$

where P(X≥15) is the probability that at least 15 workers show up, derived from the Binomial distribution X~Binomial(n,p)

7.3 Results and Discussion

The analysis shows that 366 out of 397 days (92.17 %) achieved reliable service (≥ 15 workers). This indicates a generally stable operation, with only about 8 % of days affected by understaffing and reduced sales.

The profit simulation reveals that profitability increases sharply as staffing improves up to about 17 workers, then levels off as wage costs rise. At 17 workers, the agency reaches its maximum expected yearly profit of roughly R 1.97 billion, while the probability of service problems drops below 0.1 %.

Figure 18 confirms that most days had 15–16 workers, supporting the reliability estimate, while Figure 19 illustrates the profit curve peaking at 17 workers before declining slightly.

In summary, maintaining a workforce of 17 employees provides near-perfect reliability and the highest expected profit. Adding more staff would not yield meaningful gains, whereas reducing staff below 15 would risk costly service disruptions.

Metric	**Result**
Total days observed	397
Reliable days (≥ 15 workers)	366
Reliability rate	**92.17 %**
Optimal number of workers	**17 workers**
Expected yearly profit	**≈ R 1.97 billion**

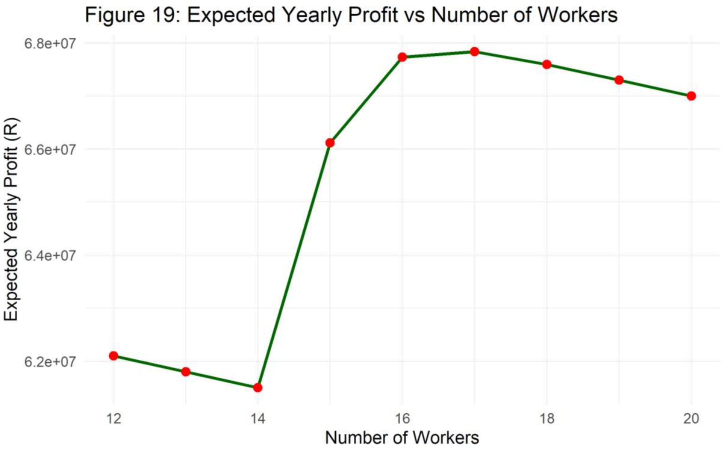


Figure 26 Expected Yearly Profit vs Number of Workers