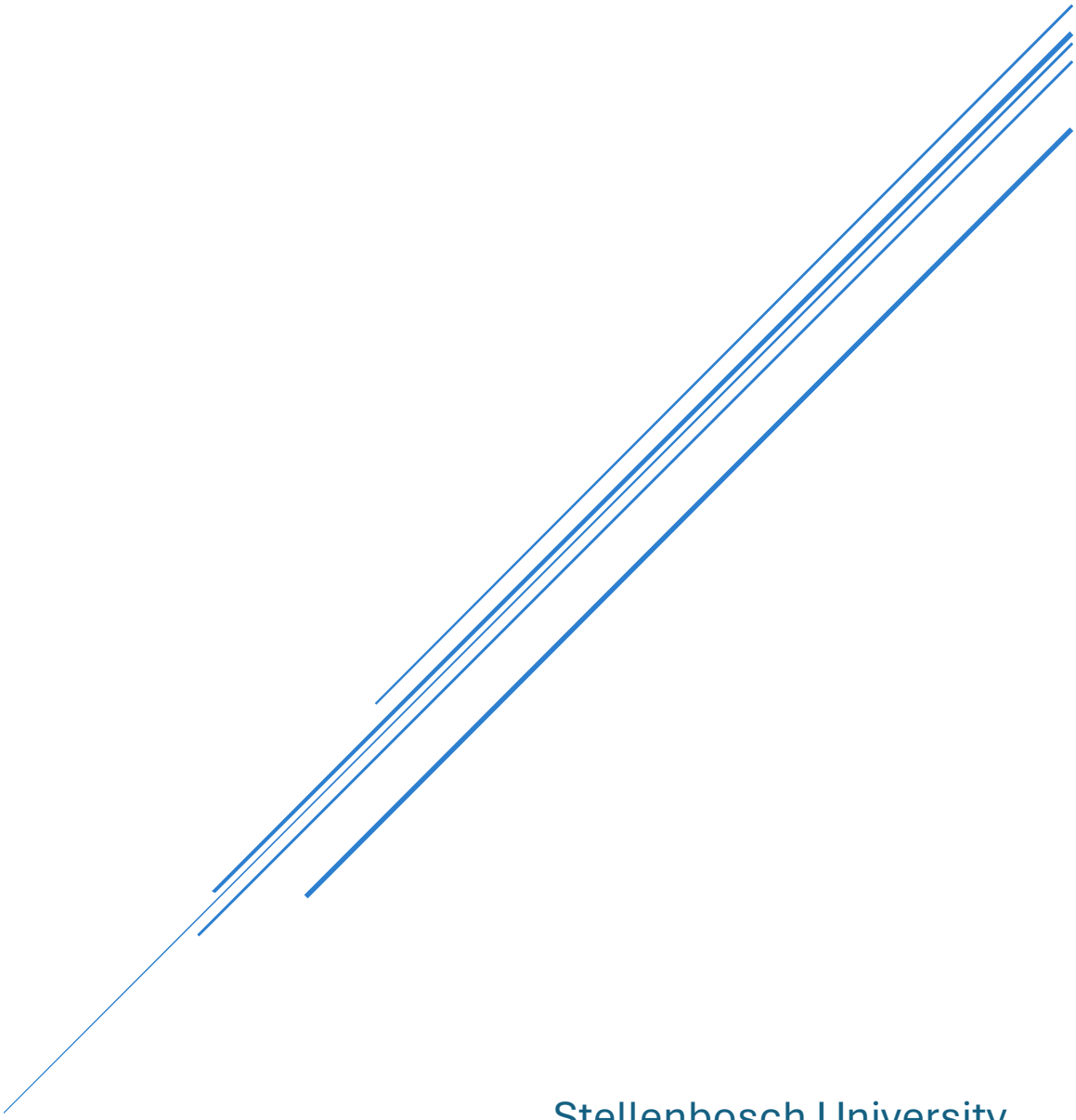
ECSA REPORT

27175413 – AG Gambarana



Stellenbosch University
Quality Assurance 344

Table of Contents

List of Figures

List of Tables

Introduction

This report presents a comprehensive statistical and optimisation analysis. The study integrates descriptive statistics, Statistical Process Control (SPC), and profit optimisation to evaluate process performance and reliability across customer, product, and sales datasets.

The initial analysis explored customer demographics and product data to establish baseline trends. Discrepancies between *products_data* and *products_Headoffice* were identified and corrected, ensuring consistency in product identifiers, selling prices, and markups. Re-analysis confirmed stable customer and sales information but corrected product-level pricing and margin data.

SPC techniques were then applied to assess delivery process control, calculate capability indices, and quantify the risks of Type I and II errors. These methods provided insight into when processes deviate from acceptable performance limits. The later sections focused on operational optimization with Section 5 modelled service efficiency and profit for two coffee shops, while Sections 6 and 7 used ANOVA and reliability modelling to evaluate delivery time variation and staffing effectiveness.

The report demonstrates how accurate data preparation, control chart analysis, and optimization techniques support evidence-based decisions that improve service reliability and profitability in industrial operations.

1. Basic Data Analysis

1.1 Data Loading and Inspection

The datasets were loaded into R using a read_csv() function. The customer dataset includes descriptions such as customer ID, gender, age, income, and city. The products dataset records product IDs, categories, descriptions, selling prices, and markups. The head office dataset has similar product-related variables for consistency checks. Lastly, the sales data set covers transaction-level data for 2022 and 2023, including customer IDs, product IDs, quantities sold, order dates, and operational metrics such as picking and delivery hours.

Discrepancies were found in the data, specifically in the products and head office datasets having different descriptions for the same product IDs.

CustomerID	Gender	Age	Income	City
CUST001	Male	16	65000	New York
CUST002	Female	31	20000	Houston
CUST003	Male	29	10000	Chicago
CUST004	Male	33	30000	San Francisco
CUST005	Female	21	50000	San Francisco
CUST006	Male	32	80000	Miami

Table 1 - Customer Data Sample

ProductID	Category	Description	SellingPrice	Markup
SOF001	Software	coral matt	511.53	25.05
SOF002	Cloud Subscription	cyan silk	505.26	10.43
SOF003	Laptop	burlywood marble	493.69	16.18
SOF004	Monitor	blue silk	542.56	17.19
SOF005	Keyboard	aliceblue wood	516.15	11.01
SOF006	Mouse	black silk	478.93	16.99

Table 2 - Product Data Sample

ProductID	Category	Description	SellingPrice	Markup
SOF001	Software	coral silk	521.72	15.65
SOF002	Software	black silk	466.95	28.42
SOF003	Software	burlywood marble	496.43	20.07
SOF004	Software	black marble	389.33	17.25
SOF005	Software	chartreuse sandpaper	482.64	17.60
SOF006	Software	cornflowerblue marble	539.33	25.57

Table 1 - Products Head Office Data Sample

CustomerID	ProductID	Quantity	orderTime	orderDay	orderMonth	orderYear	pickingHours	deliveryHours
CUST1791	CLO011	16	13	11	11	2022	17.72167	24.544
CUST3172	LAP026	17	17	14	7	2023	38.39083	31.546
CUST1022	KEY046	11	16	23	5	2022	14.72167	21.544
CUST3721	LAP024	31	12	18	7	2023	41.39083	24.546
CUST4605	CLO012	20	14	7	2	2022	15.72167	24.044
CUST2766	MON035	32	21	24	12	2022	21.05500	24.044

Table 2 - Sales 2022 and 2023 Data Sample

1.2 Summary Statistics

Summary statistics offered initial insights into the datasets.

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
CustomerID*	1	5000	2500.5000	1443.520003	2500.5	2500.50000	1853.2500	1	5000	4999	0.0000000	-1.2007200	20.4144557
Gender*	2	5000	1.5572	0.577923	2.0	1.51700	1.4826	1	3	2	0.4538869	-0.7240466	0.0081731
Age	3	5000	51.5538	21.216096	51.0	50.88275	26.6868	16	105	89	0.2041739	-0.9874439	0.3000409
Income	4	5000	80797.0000	33150.106741	85000.0	81665.00000	37065.0000	5000	140000	135000	-0.2135307	-0.7456542	468.8133055
City*	5	5000	3.9918	2.002232	4.0	3.98975	2.9652	1	7	6	-0.0108635	-1.2745838	0.0283158

Table 3 - Customer Data Statistics

The customer dataset displayed in Table 3 shows a broad and balanced spread. Ages average around 51 to 52 years, with only slight positive skew, meaning both younger and older groups are well represented. Income levels average just over R80 000 but display extremely high kurtosis, indicating the presence of a few very high-income outliers. This suggests that while customers fall into a moderate-income range, and a small section exists that could be targeted with higher-value products. Gender and city distributions are relatively balanced, with no single group or location dominating the customer base.

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
ProductID*	1	60	30.50000	17.464249	30.500	30.50000	22.239000	1.00	60.00	59.00	0.0000000	-1.2601448	2.2546249
Category*	2	60	3.50000	1.722237	3.500	3.50000	2.223900	1.00	6.00	5.00	0.0000000	-1.3258048	0.2223399
Description*	3	60	16.40000	10.078001	16.000	16.20833	13.343400	1.00	35.00	34.00	0.1029599	-1.2935763	1.3010643
SellingPrice	4	60	4493.59283	6503.770150	794.185	3189.25479	525.722547	350.45	19725.18	19374.73	1.4261752	0.4338057	839.6331159
Markup	5	60	20.46167	6.072598	20.335	20.51187	7.309218	10.13	29.84	19.71	-0.0367077	-1.2380989	0.7839690

Table 4 - Products Data Statistics

The product dataset displays an assorted mix of categories and product descriptions, with IDs evenly spread out across the sample. The average selling price is about R4 494, but the distribution is heavily right-skewed, which suggests most products are priced lower, with a few high-value products pulling up the average. Markup percentages group around 20%, which indicates that the dataset has a relatively standard pricing margin. Overall, the product range appears varied but is not controlled by a single category or price range, which can help appeal to a broad customer base and allow for flexible pricing strategies.

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
ProductID*	1	360	69.38889	23.217847	72.000	71.88542	22.239000	1.00	110.00	109.00	-0.8665172	0.4912730	1.2236880
Category*	2	360	3.50000	1.710202	3.500	3.50000	2.223900	1.00	6.00	5.00	0.0000000	-1.2781771	0.0901356
Description*	3	360	30.68611	17.319505	29.500	30.76736	22.980300	1.00	60.00	59.00	-0.0277818	-1.3900365	0.9128181
SellingPrice	4	360	4410.96186	6463.822788	797.215	3054.22903	515.752062	290.52	22420.14	22129.62	1.5279096	0.7789339	340.6733733
Markup	5	360	20.38550	5.665949	20.580	20.42868	6.664287	10.06	30.00	19.94	-0.0477692	-1.0739041	0.2986217

Table 5 - Products Head Office Data Statistics

The head office product data shows a large and balanced range of products and categories, with ProductIDs and Descriptions evenly distributed. Most products fall within moderate price and markup ranges, but SellingPrice is heavily right-skewed, showing that some products command much higher prices than the majority. Markup percentages remain quite consistent, suggesting stable pricing strategies across the inventory. There is no single dominant category or extreme outliers, indicating the product lineup is broad and diversified enough to support various sales channels and target markets.

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
CustomerID*	1	1e+05	2492.33848	1444.5778106	2503.000	2491.191987	1862.1456	1.0000000	5000.0000	4999.00000	0.0020964	-1.2107656	4.5681561
ProductID*	2	1e+05	32.43610	18.0302099	35.000	32.819100	23.7216	1.0000000	60.0000	59.00000	-0.1603407	-1.3178154	0.0570165
Quantity	3	1e+05	13.50347	13.7601316	6.000	11.458100	5.9304	1.0000000	50.0000	49.00000	1.0443411	-0.2185180	0.0435134
orderTime	4	1e+05	12.93230	5.4951268	13.000	13.117888	5.9304	1.0000000	23.0000	22.00000	-0.2271685	-0.7101693	0.0173771
orderDay	5	1e+05	15.49683	8.6465055	15.000	15.495088	10.3782	1.0000000	30.0000	29.00000	0.0027726	-1.2007412	0.0273427
orderMonth	6	1e+05	6.44813	3.2834460	6.000	6.445538	4.4478	1.0000000	12.0000	11.00000	0.0069282	-1.1764404	0.0103832
orderYear	7	1e+05	2022.46273	0.4986115	2022.000	2022.453413	0.0000	2022.0000000	2023.0000	1.00000	0.1494937	-1.9776714	0.0015767
pickingHours	8	1e+05	14.69547	10.3873345	14.055	13.543098	6.9188	0.4258889	45.0575	44.63161	0.7357093	0.4143469	0.0328476
deliveryHours	9	1e+05	17.47646	9.9999440	19.546	17.775077	8.8956	0.2772000	38.0460	37.76880	-0.4704880	-0.8716457	0.0316226

Table 6 - Sales Data Statistics

The sales dataset includes many records with a widespread across customers and products. Quantity sold varies a lot, with most sales involving smaller quantities and a few very large orders, indicated by positive skew. Both order time and order day are evenly distributed, so sales activity is spread throughout the day and month. There are no major outliers, and the data shows a steady flow of transactions rather than unusual sales spikes.

1.3 Handling Missing Values

The datasets were all checked for missing values; however, none were found. To make certain, na.omit() was used to keep only the complete instances. Duplicate records were also checked for, but none were found. This helped to guarantee a clean and accurate dataset for any further processing and analysis.

1.4 Data Filtering and Sub setting

To ensure a more reliable analysis of the data, it was cleaned by identifying and removing outliers. For the customer dataset, the Interquartile Range (IQR) was calculated for variables such as Age and Income, then data points outside 1.5 times the IQR from the quartiles were filtered out. Similarly, the IQR method was used to remove outliers from the product data. It was applied to variables like SellingPrice and Markup to filter out irregular values. The head

office product dataset was also cleaned by removing outliers based on the Markup variable using the IQR approach.

Lastly, the sales data for 2022 and 2023 was filtered by excluding Quantity values considered outliers through the IQR method, focusing on the central portion of data distribution. This outlier removal process ensures that any further analysis reveals more meaningful patterns within each dataset, avoiding the skewing effect of extreme values.

1.5 Data Visualization

Age Distribution of Customers

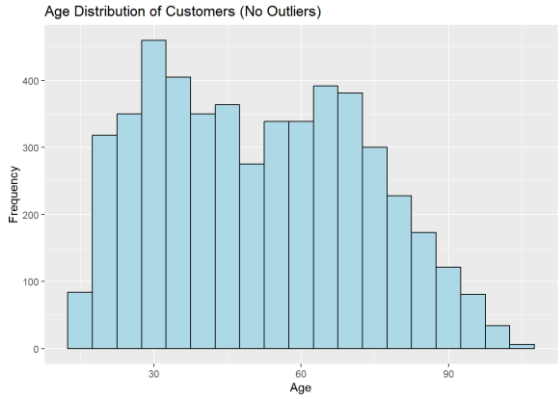


Figure 1 - Age Distribution of Customers

The histogram in Figure 1Figure 1, it shows that most customers fall between 25 and 40 years old. This suggests the company's customer base is skewed toward younger to mid-career adults, who may have disposable income but are also price sensitive. The graph shows a bimodal distribution with another peak in frequency around the ages of 65-70.

Income Distribution by Gender

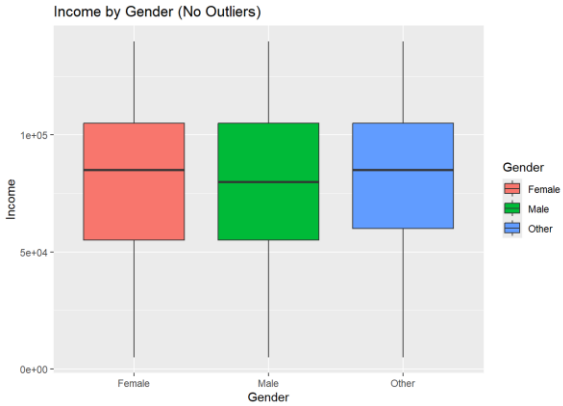


Figure 2 - Income Distribution by Gender

The boxplot highlights overlapping income ranges across genders, with all groups showing a widespread. While median incomes are similar, the presence of high-income outliers indicates potential for targeting premium products without gender bias.

Distribution of Product Selling Prices

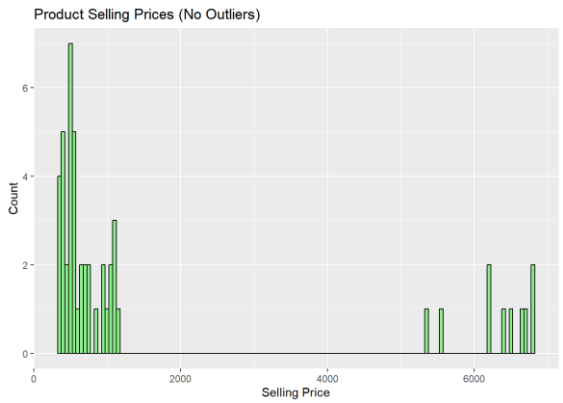


Figure 3 - Product Selling Prices

The histogram shows a heavy concentration of low-priced products, with few high-value items. This suggests the company operates primarily in a lower-price market, but the long tail of expensive products may represent an exclusive higher-end market.

Count of Products by Category

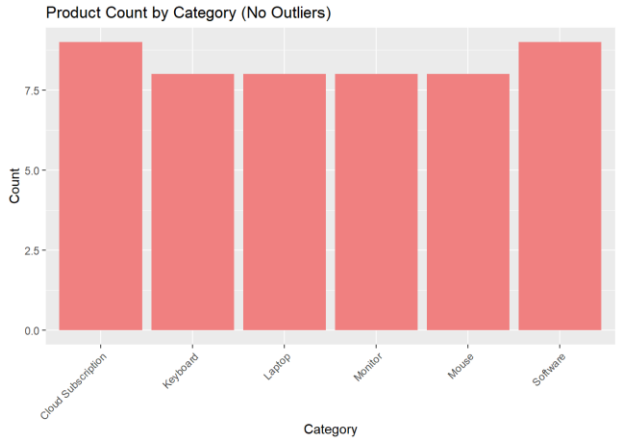


Figure 4 - Product Count

The bar chart shows categories are evenly distributed, with no single category dominating. This balanced product portfolio may reduce risk, but it also suggests the company lacks clear specialization.

Markup Distribution by Head Office Product Category

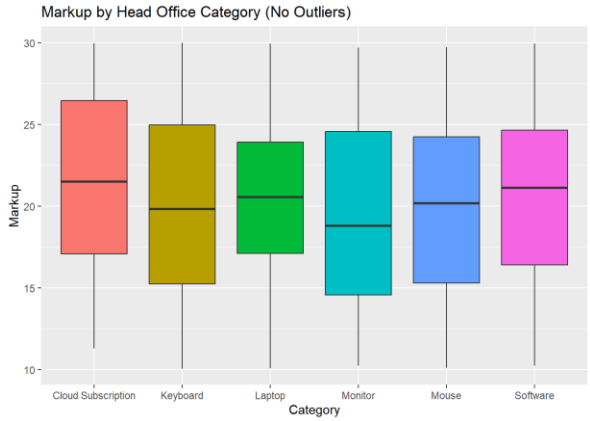


Figure 5 - Markup by Head Office

The boxplot reveals significant differences in markup across categories. Some categories consistently achieve higher markups, suggesting stronger pricing power. Others show more variability, indicating pricing inefficiencies.

Distribution of Quantity Sold

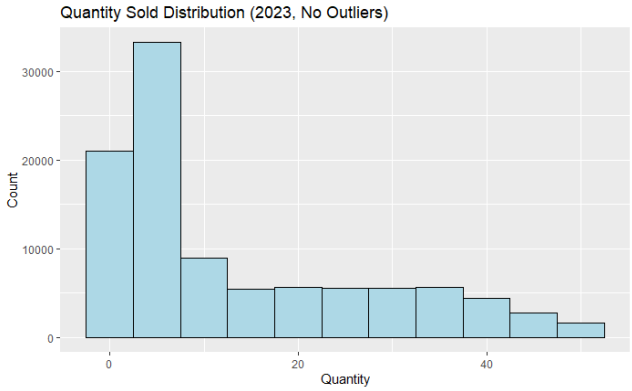


Figure 6 - Quantity Sold Distribution

The histogram shows most transactions involve small quantities (1–5 units), though larger orders exist. This pattern implies a high-volume, small-order business model, driven by retail customers rather than wholesale.

Monthly Sales Trend (2022 vs 2023)

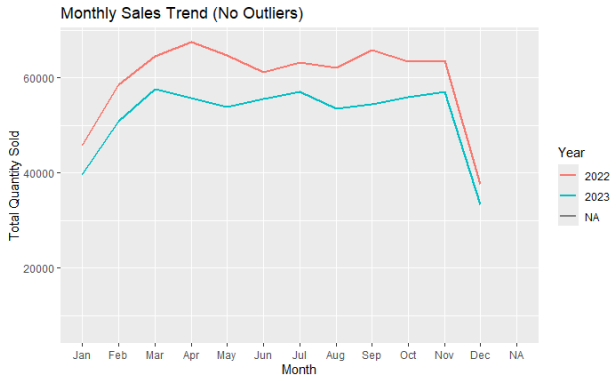


Figure 7 - Sales Trend

The line plot shows that 2022 sales exceeded 2023 levels in all months, demonstrating business decline. Both years show seasonal peaks in mid-year, which may reflect promotions or holiday-related demand. This seasonality can guide inventory and staffing decisions.

1.6 Relationships and Interpretation

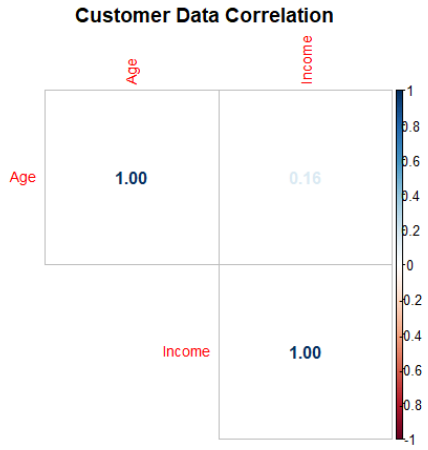


Figure 8 - Customer Data Correlation

The correlation matrix, Figure 8, shows that age and income are only weakly correlated. This means older customers are not necessarily wealthier, and younger customers also contribute significantly to income distribution. For business strategy, it suggests that product offerings should not assume income levels based solely on customer age.

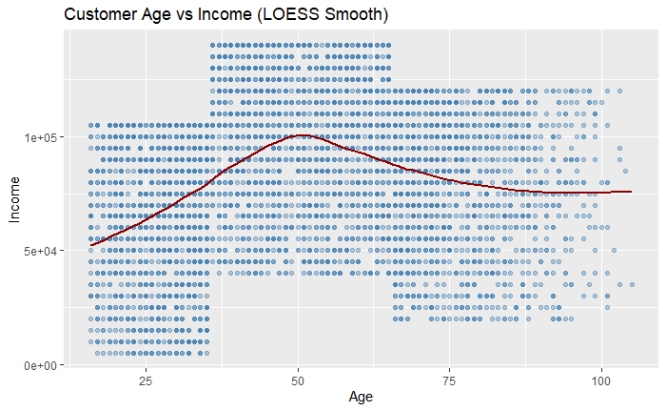


Figure 9 - Customer Age vs Income

The scatterplot reinforces the weak positive relationship between age and income. The trend line shows a slight upward slope, but the wide spread of points highlights variability at all ages. This suggests opportunities exist across different age brackets, with no single demographic dominating high-income status.

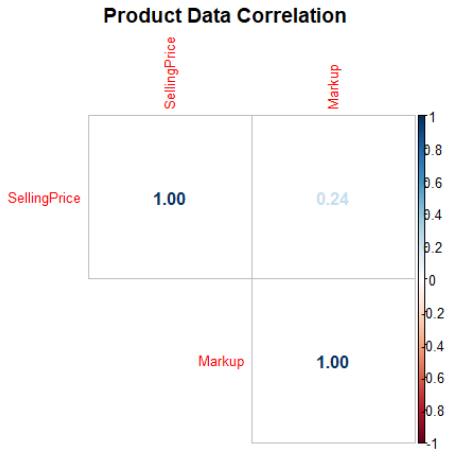


Figure 10 - Product Data Correlation

The correlation between selling price and markup is positive, as seen in Figure 10, indicating that higher-priced products tend to have higher markups. However, the correlation is not very high, which may point to inconsistent pricing strategies. Identifying underpriced products relative to their peers could unlock additional revenue potential.

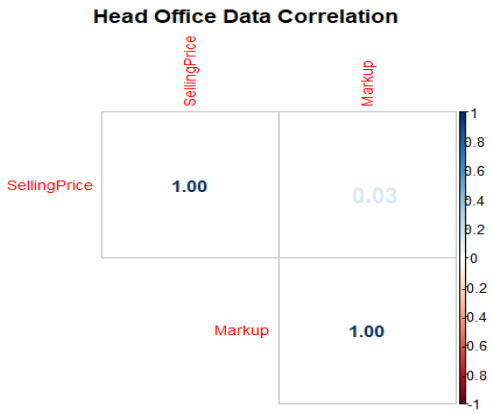


Figure 11 - Head Office Data Correlation

The head office dataset also shows a positive link between selling price and markup, suggesting a deliberate pricing strategy where more expensive products command higher margins. However, the correlation is very low, so this may not be completely true.

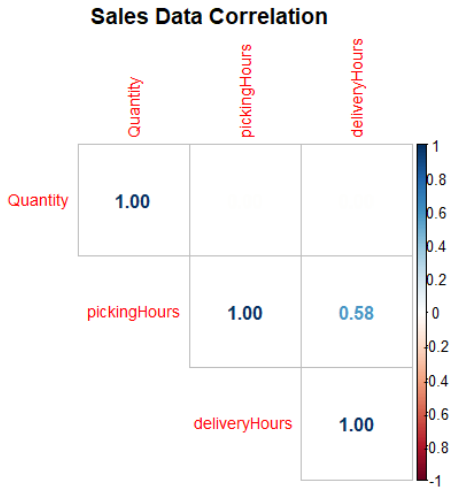


Figure 12 - Sales Data Correlation

Sales data correlations in Figure 12 indicate that order quantity has very weak relationships with picking and delivery hours. This means that operational timelines currently do not have a strong effect on how much customers order. However, delivery delays might reduce order sizes, pointing to a potential risk area if logistics performance declines. Picking and delivery hours are relatively highly correlated, which insinuates that any delays in picking hours will have an effect on delivery hours.

3. Statistical Process Control (SPC)

3.1 Control Chart Initialization

To assess delivery process stability, the delivery data was first sorted chronologically by year, month, day, and order time to simulate real-time monitoring. For each product type, the oldest 720 records (30 samples of 24) were used to initialize X-bar and s-charts. Figure 13 and Figure 14 show a larger, more detailed display of what the charts show by using the example product CLO011, while Figure 15 and Figure 16 extend this visualization to all products. The centre lines and control limits at one, two, and three sigma were calculated using standard formulas for subgroup size 24. These charts provide a clear baseline for ongoing process monitoring, highlighting whether the delivery process for each product type was initially stable or exhibited signs of excessive variability or mean shifts.

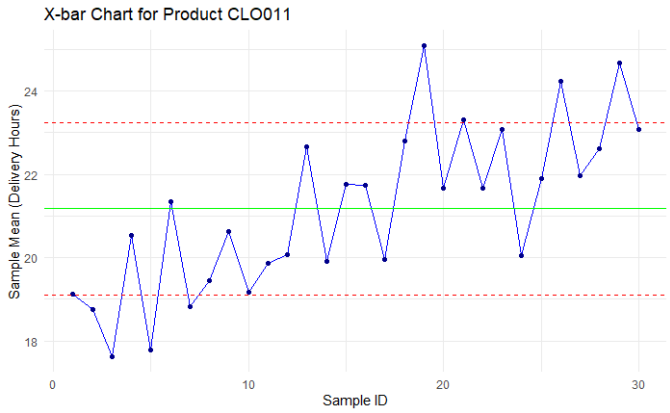


Figure 13 - X-bar Chart for Product CLO011

The X-bar chart for Product CLO011 plots the sample means of delivery hours across 30 samples (each with 24 deliveries). The green line shows the overall process mean, while the dashed red lines indicate the upper and lower control limits. In the initial samples, the process mean fluctuates below the centre line but shows a noticeable upward trend starting around sample 17. Several samples approach or exceed the upper control limit towards the end of the sequence, which may signal the beginning of the process becoming unstable or some special cause variation taking place. This upward shift suggests that delivery times for CLO011 are increasing over time and that the process may be trending out of control. Managers should investigate possible causes for this shift, such as resource bottlenecks, changes in workflow, or external delays.

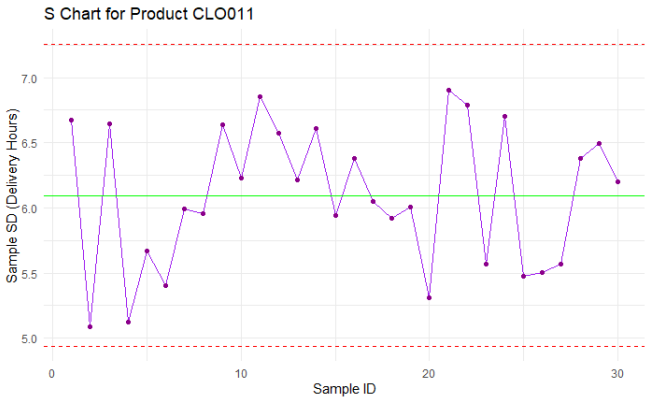


Figure 14 - S-Chart for Product CLO011

The S chart for Product CLO011 tracks the sample standard deviations (spread) of delivery hours for the initial 30 samples. The green line shows the average process spread, while the dashed red lines mark the control limits for variability. Many samples are within control, but there is variability in the spread, especially in the first half of the data. After sample 15, the spread stabilizes closer to the centre line. No samples exceed the upper control limit, indicating that the overall process variability remains acceptable and under statistical control for this period.

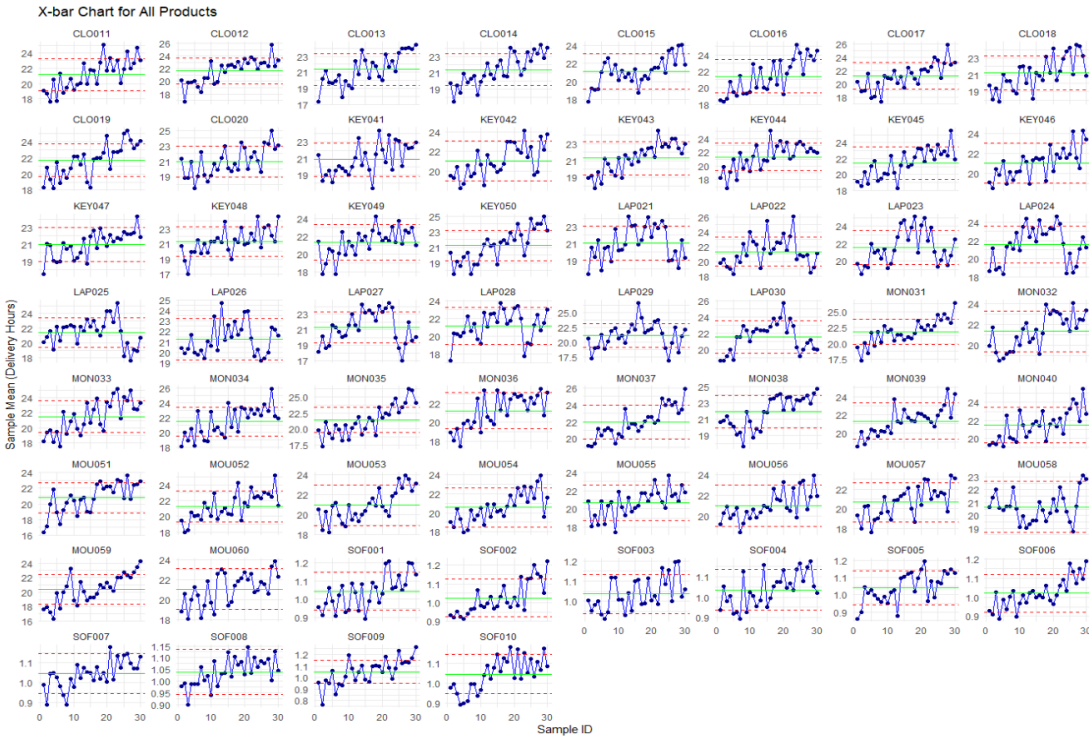


Figure 15 - X-bar Chart for All Products

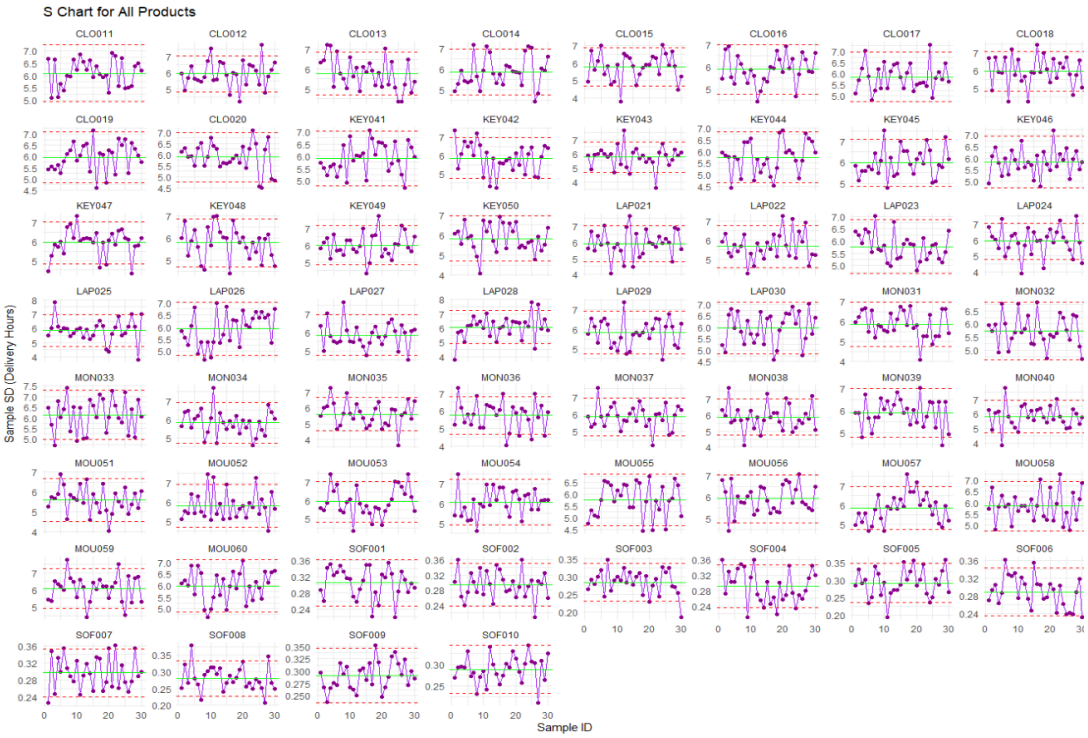


Figure 16 - S-Chart for All Products

Similar interpretations can be found for the rest of the products shown in the diagrams above.

3.2 Ongoing Process Monitoring

After setting up the control chart, subsequent delivery samples were tracked in real time to decide if the process was consistent and stable for all types of products. In any new sample, monitoring procedure involved first checking the s-chart to check process variation. Only when a sample standard deviation seemed to be within pre-established control limits would the X-bar chart be used to calculate the sample mean. As shown in Table 7, most samples had their s values within control, allowing for valid interpretation of their respective X-bar means. For these samples, process delivery was considered stable, and no action was needed.

	ProductID	sample_id	sample_mean	sample_sd	xbarbar	sbar	A3	B3	B4	UCL_X	CL_X	LCL_X	UCL_S	CL_S	LCL_S	U2SIGMA_X	L2SIGMA_X	U1SIGMA_X	L1SIGMA_X	U1SIGMA_S	L1SIGMA_S
1	CLO011	31	21.62733	4.735772	21.17897	6.095790	0.34	0.81	1.19	23.25154	21.17897	19.10640	7.253990	6.095790	4.937590	22.56068	19.79725	21.86982	20.48811	6.481857	5.709723
2	CLO011	32	23.96250	5.448967	21.17897	6.095790	0.34	0.81	1.19	23.25154	21.17897	19.10640	7.253990	6.095790	4.937590	22.56068	19.79725	21.86982	20.48811	6.481857	5.709723
3	CLO011	33	23.66900	5.597457	21.17897	6.095790	0.34	0.81	1.19	23.25154	21.17897	19.10640	7.253990	6.095790	4.937590	22.56068	19.79725	21.86982	20.48811	6.481857	5.709723
4	CLO011	34	24.83750	6.436008	21.17897	6.095790	0.34	0.81	1.19	23.25154	21.17897	19.10640	7.253990	6.095790	4.937590	22.56068	19.79725	21.86982	20.48811	6.481857	5.709723
5	CLO011	35	24.75417	5.963169	21.17897	6.095790	0.34	0.81	1.19	23.25154	21.17897	19.10640	7.253990	6.095790	4.937590	22.56068	19.79725	21.86982	20.48811	6.481857	5.709723

Table 7 - Sample Control Status

There were exceptions, though, where the sample standard deviations exceeded the upper control limit, suggesting potential increases in process spread. In those cases, sample mean was not tracked, and sample status was flagged for intervention. Table 7 shows times of

heightened variability, as well as samples whose mean suddenly shifted from centre lines, even during spread under control. Also noteworthy is that some products had alternating durations of stable versus unstable control, suggestive of possible shifts in factors of operation, order volume, or outside influences.

The analysis effectively demonstrates that the SPC monitoring technique worked well in identifying shifts in the process early. Monitoring spread and average, managers could instantly intervene on nascent problems either by questioning causes of abnormal spread or by responding to changes in average delivery times.

3.3 Process Capability Assessment

Process capability was assessed using the first 1000 deliveries for each product type. The table displays key statistics including the process mean (μ) and standard deviation (σ) along with capability indices Cp, Cpu, Cpl, and Cpk. These indices quantify how well each product's delivery times fit within the specification limits of 0 to 32 hours.

From Table 8, all listed products have Cpk values well above 1, indicating a high capability to consistently meet the Voice of the Customer (VOC) requirements. The Cp values are relatively large, underscoring tight process variation compared to the allowable specification range. Because Cpk is not limited by shifts in process mean, the consistently high values signify that the processes are well-centered and highly capable.

This strong process capability suggests that delivery times for these products are stable, predictable, and well within customer tolerance levels. However, continuous monitoring is advised to maintain control quality and quickly identify any deviations.

	ProductID	mu	sigma	n	LSL	USL	Cp	Cpu	Cpl	Cpk
1	SOF001	1.069425	0.3100505	1000	0	32	17.20150	33.25327	1.149732	1.149732
2	SOF002	1.064625	0.3082288	1000	0	32	17.30316	33.45499	1.151336	1.151336
3	SOF003	1.069425	0.2954763	1000	0	32	18.04995	34.89347	1.206442	1.206442
4	SOF004	1.069825	0.3042944	1000	0	32	17.52689	33.88185	1.171919	1.171919
5	SOF005	1.078225	0.3083722	1000	0	32	17.29512	33.42473	1.165502	1.165502
6	SOF006	1.059400	0.3019032	1000	0	32	17.66571	34.16172	1.169691	1.169691
7	SOF007	1.085750	0.3044887	1000	0	32	17.51570	33.84280	1.188605	1.188605
8	SOF008	1.075675	0.2924531	1000	0	32	18.23654	35.24705	1.226037	1.226037
9	SOF009	1.085725	0.3050057	1000	0	32	17.48601	33.78546	1.186563	1.186563
10	SOF010	1.069425	0.2964676	1000	0	32	17.98960	34.77678	1.202408	1.202408

Table 8 - Process Capability Indices Sample

3.4 Process Control Issue Identification

To identify process control issues, several rules were applied:

A. Out-of-Control Spread: For each product, any sample whose standard deviation was greater than the upper three-sigma control limit was marked as a suspected sign of excessive process variability. These out-of-control instances were dispersed among various products and a total of 173 samples were identified, as evident in Table 9 (First 3 Bad Sample) and Table 10 (Last 3 Bad Sample). Increased frequency of such exceptions can indicate that certain product delivery processes are extra susceptible to interruption or operational fluctuation. Such pinpointed intervals need to be addressed by the managers specifically to isolate special causes, such as equipment breakdown, fluctuations in demand, or labour disruptions, which can undermine overall delivery reliability.

	ProductID	sample_id	sample_mean	sample_sd	xbarbar	sbar	A3	B3	B4	UCL_X	CL_X	LCL_X	UCL_S	CL_S	LCL_S	U2SIGMA_X	L2SIGMA_X	U1SIGMA_X	L1SIGMA_X	U1SIGMA_S	L1SIGMA_S	s_status	x_status
1	CLO012	45	21.44183	8.125348	21.645	5.94492	0.34	0.81	1.19	23.66627	21.645	19.62373	7.074455	5.94492	4.815386	22.99252	20.29748	22.31876	20.97124	6.321432	5.568409	HIGH	NA
2	CLO012	52	22.08767	7.465743	21.645	5.94492	0.34	0.81	1.19	23.66627	21.645	19.62373	7.074455	5.94492	4.815386	22.99252	20.29748	22.31876	20.97124	6.321432	5.568409	HIGH	NA
3	CLO012	54	21.21458	7.599840	21.645	5.94492	0.34	0.81	1.19	23.66627	21.645	19.62373	7.074455	5.94492	4.815386	22.99252	20.29748	22.31876	20.97124	6.321432	5.568409	HIGH	NA

Table 9 - First 3 Bad Sample

	ProductID	sample_id	sample_mean	sample_sd	xbarbar	sbar	A3	B3	B4	UCL_X	CL_X	LCL_X	UCL_S	CL_S	LCL_S	U2SIGMA_X	L2SIGMA_X	U1SIGMA_X	L1SIGMA_X	U1SIGMA_S	L1SIGMA_S	s_status	x_status
1	SOF010	78	1.187717	0.3451462	1.047721	0.2877432	0.34	0.81	1.19	1.145554	1.047721	0.9498882	0.3424144	0.2877432	0.233072	1.112943	0.982499	1.080332	1.01511	0.3059669	0.2695194	HIGH	NA
2	SOF010	81	1.112717	0.3779174	1.047721	0.2877432	0.34	0.81	1.19	1.145554	1.047721	0.9498882	0.3424144	0.2877432	0.233072	1.112943	0.982499	1.080332	1.01511	0.3059669	0.2695194	HIGH	NA
3	SOF010	83	1.148133	0.3614333	1.047721	0.2877432	0.34	0.81	1.19	1.145554	1.047721	0.9498882	0.3424144	0.2877432	0.233072	1.112943	0.982499	1.080332	1.01511	0.3059669	0.2695194	HIGH	NA

Table 10 - Last 3 Bad Sample

B. Good Control: Table 11 shows the longest series in which sample standard deviations consistently fell within the relatively narrow ranges of minus one and plus one sigma limits for each product. These extended streaks of good control suggest not merely firm process management but effective control of random or assignable causes of variation. Long run products have noticeably long runs and are stable, with repeatable and consistent delivery times. They give managers an understanding of which teams and processes are best practice, and which can have an impact on standardization across the company.

	ProductID	max_consec_in_1sigma_s
1	CLO011	4
2	CLO012	5
3	CLO013	4
4	CLO014	3
5	CLO015	3
6	CLO016	3
7	CLO017	3
8	CLO018	3
9	CLO019	4
10	CLO020	2

Table 11 - Longest Streak Sample

C. Critical Mean Shifts: To capture minor but important process shifts, Table 12 and Table 13 track occurrences when four or more consecutive X-bar sample means crossed the upper two-sigma line. 122 such runs were reported, reflecting frequent or chronic deviations from

standard delivery performance. These clusters are particularly significant because they may reflect underlying systemic issues (e.g., gradual equipment degradation, supplier lead time variability, or persistent bottlenecks) instead of random errors. When these types of changes are discovered, immediate investigation and targeted process tweak are suggested to stop a bad trend from escalating into a chronic issue affecting customer satisfaction.

	ProductID	start	end	length
1	CLO011	32	36	5
2	CLO011	63	66	4
3	CLO012	31	34	4

Table 12 - First 3 Sample of Four Consecutive Runs

	ProductID	start	end	length
1	SOF010	40	46	7
2	SOF010	73	80	8
3	SOF010	82	88	7

Table 13 - Last 3 Sample of Four Consecutive Runs

Taken together, the combination of out-of-control events (A), periods of excellent stability (B), and critical mean shifts (C) enables managers to precisely identify when and where the process deviates from expectations, supporting prioritization of corrective actions and ongoing operational improvement.

4. Statistical Process Control and Data Integrity Analysis

4.1 Estimation of Type I Error Likelihood

Type I errors in SPC are false alarms as they occur when a control rule flags a process as “out of control” even though the process is truly in control and cantered on the control chart centre line implied by the first 30 samples. Because these probabilities are properties of the normal distribution and of the rule logic, they do not depend on the empirical data. For the rules applied in Section 3.4 we obtain the following theoretical probabilities under the null hypothesis (process in control):

Rule A: One s-sample outside the +3σ control limit

Under an in-control process that follows a normal distribution, the probability that any single s-sample will exceed the ±3σ limits is extremely small. Mathematically, this is given by:

$$P(\text{either side, } \pm 3\sigma) = 2 \times 0.0013499 = 0.0026998 (\approx 0.27\%)$$

This implies that, on average, approximately one in every 370 samples will exceed the ±3σ limits purely due to random variation. That means that any observation outside these boundaries is regarded as a strong indication of a special cause of variation rather than random noise. The rarity of such an occurrence justifies the use of ±3σ as a reliable control threshold in SPC practice.

Rule B: Runs or sequences relative to the centreline / ±1σ band

Since the normal distribution is symmetric about its mean, each sample has a 0.5 probability of falling above or below the centreline. The probability of observing *k* consecutive samples on one side of the mean, assuming independence, is therefore 0.5^k. For seven consecutive points:

$$P(7 \text{ consecutive above}) = 0.5^7 = 0.0078125 (\approx 0.78\%)$$

This result indicates that such a run would occur fewer than eight times per 1 000 independent sequences when the process is in control. A sequence of seven or more samples consistently above or below the centreline therefore provides statistically significant evidence of a systematic shift or trend in the process mean, prompting further investigation.

Rule C: Four consecutive x̄ samples beyond the +2σ second control limit

For the $\bar{X}$ chart, the probability that a single sample mean exceeds the +2σ limit on one side is approximately 0.02275 (2.275%). Assuming sample independence, the probability of observing four such consecutive occurrences is:

$$P(4 \text{ consecutive } > +2\sigma) = 0.02275^4 = 2.678772 \times 10^{-7} (\approx 0.000027\%)$$

This exceedingly small probability shows that the appearance of four consecutive $\bar{X}$ points beyond +2σ is practically impossible under random variation alone. When this condition is observed, it serves as strong statistical evidence of a genuine process shift.

The results demonstrate a clear hierarchy of control-chart sensitivity and false-alarm risk. Single point ±3σ violations are rare but may still occur occasionally due to random variation. Run-length rules, such as seven consecutive points above or below the mean, have a slightly higher false-alarm probability but are particularly effective in detecting gradual shifts. Compound rules, such as four consecutive $\bar{X}$ points beyond +2σ, are extremely unlikely under normal conditions and almost always signal a true process disturbance.

In industrial practice, it is recommended that engineers interpret SPC rule violations not in isolation but in conjunction with operational evidence, such as machine behaviour,

environmental conditions, or operator changes, to avoid unnecessary process adjustments triggered by random noise.

4.2 Estimation of Type II Error Likelihood

A Type II error happens when the process has actually shifted from its target mean, but the control chart fails to detect it. In this situation, the chart shows the process as being in control even though it is not. This kind of error is also known as a consumer's risk, as it can lead to poor-quality products reaching customers. The following example demonstrates this concept using the given bottle-filling process data.

The process is designed to fill bottles to a nominal mean (centreline) of 25.050 litres, with control limits at LCL = 25.011 L and UCL = 25.089 L. However, the process mean has unknowingly shifted to μ = 25.028 L, and the standard deviation of the sample mean has increased to $\sigma_{\bar{x}}$ = 0.017 L (previously 0.013 L). This means that the process is no longer centred, and the variation between samples has increased.

The probability of a Type II error, represented by β, is the chance that a sample mean will still fall between the control limits even though the process mean has changed. This probability can be calculated using the formula:

$$\beta = \Phi\left(\frac{UCL - \mu}{\sigma_{\bar{X}}}\right) - \Phi\left(\frac{LCL - \mu}{\sigma_{\bar{X}}}\right)$$

Substituting the given values:

$$\beta \approx \Phi\left(\frac{25.089 - 25.028}{0.017}\right) - \Phi\left(\frac{25.011 - 25.028}{0.017}\right)$$
$$\beta \approx \Phi(3.588) - \Phi(-1.000)$$
$$\beta \approx 0.99998 - 0.15866 = 0.8412$$

Therefore, β ≈ 0.841, or about 84.1%. This means that when the process has shifted to 25.028L, there is an 84% chance that the sample mean will still appear to be within the control limits. In other words, the control chart will fail to detect the shift most of the time.

This shows that the chart has low sensitivity to moderate process shifts. The large Type II error rate occurs because the process variation increased, and the control limits were relatively wide. In practice, this means that most small changes in the process average would go unnoticed. To improve detection, the company could either reduce variation (for example, by improving machine consistency or increasing the sample size) or use more

sensitive charting methods such as CUSUM or EWMA charts. However, it is important to note that making the chart more sensitive would increase the risk of false alarms (Type I errors).

In summary, this analysis demonstrates the trade-off between Type I and Type II errors in process control. While wide control limits reduce false alarms, they also make it easier to miss real process shifts. Effective process monitoring therefore depends on finding the right balance between detecting small changes and avoiding unnecessary adjustments to a stable process.

4.3 Data Correction and Re-analysis: Impact on Sales Outcomes

Step 1: Read fixed datasets and inspect

The customer data remained unchanged between Section 1 and the corrected analysis. Therefore, no changes were made to the *customer_data* file, and all identifiers, demographic fields, and structural elements stayed consistent. As a result, the descriptive statistics and visualisations related to customer characteristics, such as age distribution, gender ratio, and income levels, are identical to those previously reported. Therefore, there is no need to reinterpret any patterns found within the dataset.

In contrast, significant updates were made to the product-related datasets. The *products_data* file was corrected and saved as *products_data2025.csv*, where the category labels were aligned with their corresponding ProductIDs, and inconsistencies in product descriptions were resolved. These corrections ensure that all product identifiers and associated price and markup values now follow the intended pattern within the dataset. Consequently, all analyses that depend on product information, such as category counts, price distributions, and markup summaries, now provide accurate and consistent results.

The most substantial changes occurred in the head-office product file, which was corrected and saved as *products_Headoffice2025.csv*. The issues identified earlier, including incorrect product codes (“NA” prefixes), misplaced category labels, and repeated pricing errors, were resolved according to the specified repeat pattern from head office. After correction, the central catalogue now aligns properly with the local product records. This alignment ensures that joins between the two datasets produce reliable and reproducible pricing information for sales transactions.

Finally, the sales data (*sales2022and2023*) remained structurally unchanged, with identical customer IDs, product IDs, quantities, and timestamps. However, the total value of these sales was recomputed using the corrected price and markup data from the revised product and head-office files. This means that while quantity-based metrics, such as units sold or

transaction counts, remained the same, all revenue-based calculations, including monthly totals and average sale values, now reflect accurate pricing.

Step 2: Summary / skim of corrected datasets

Psych Describe of Customer Data

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
CustomerID*	1	5000	2500.5000	1443.520003	2500.5	2500.50000	1853.2500	1	5000	4999	0.0000000	-1.2007200	20.4144557
Gender*	2	5000	1.5572	0.577923	2.0	1.51700	1.4826	1	3	2	0.4538869	-0.7240466	0.0081731
Age	3	5000	51.5538	21.216096	51.0	50.88275	26.6868	16	105	89	0.2041739	-0.9874439	0.3000409
Income	4	5000	80797.0000	33150.106741	85000.0	81665.00000	37065.0000	5000	140000	135000	-0.2135307	-0.7456542	468.8133055
City*	5	5000	3.9918	2.002232	4.0	3.98975	2.9652	1	7	6	-0.0108635	-1.2745838	0.0283158

Table 14 - Customer Data Statistics

Following the data corrections, the customer demographic statistics remained effectively identical to those reported in Section 1. Measures of central tendency and spread, such as means, medians, and standard deviations, show no meaningful differences between the original and corrected datasets. Any minor variations are attributable only to rounding rather than genuine structural changes in the data. Therefore, interpretations based on customer demographics remain unchanged, and no further corrective action is necessary for customer-targeting recommendations.

Psych Describe of Products Data (corrected)

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
ProductID*	1	60	30.50000	17.464249	30.500	30.50000	22.239000	1.00	60.00	59.00	0.0000000	-1.2601448	2.2546249
Category*	2	60	3.50000	1.722237	3.500	3.50000	2.223900	1.00	6.00	5.00	0.0000000	-1.3258048	0.2223399
Description*	3	60	16.40000	10.078001	16.000	16.20833	13.343400	1.00	35.00	34.00	0.1029599	-1.2935763	1.3010643
SellingPrice	4	60	4493.59283	6503.770150	794.185	3189.25479	525.722547	350.45	19725.18	19374.73	1.4261752	0.4338057	839.6331159
Markup	5	60	20.46167	6.072598	20.335	20.51187	7.309218	10.13	29.84	19.71	-0.0367077	-1.2380989	0.7839690
ProductType*	6	60	3.50000	1.722237	3.500	3.50000	2.223900	1.00	6.00	5.00	0.0000000	-1.3258048	0.2223399
Position	7	60	30.50000	17.464249	30.500	30.50000	22.239000	1.00	60.00	59.00	0.0000000	-1.2601448	2.2546249
PriceGroup	8	60	5.50000	2.896520	5.500	5.50000	3.706500	1.00	10.00	9.00	0.0000000	-1.2829411	0.3739392

Table 15 - Products Data (corrected) Statistics

In contrast, the product data exhibited few but meaningful changes after correction. Adjustments to mislabelled products and inaccurate pricing values resulted in slight shifts in the average selling price and markup statistics. These corrected measures now more accurately represent the true catalogue pricing structure and remove the artificial variation that had been introduced by the earlier head-office inconsistencies. As a result, any analyses or recommendations involving price comparisons, markup variability, or category-

level averages should be re-evaluated using the updated product data to ensure validity and accuracy.

Psych Describe of Products Headoffice Data (corrected)

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
ProductID*	1	360	69.388889	23.217847	72.000	71.88542	22.239000	1.00	110.00	109.00	-0.8665172	0.4912730	1.2236880
Category*	2	360	3.500000	1.710202	3.500	3.50000	2.223900	1.00	6.00	5.00	0.0000000	-1.2781771	0.0901356
Description*	3	360	30.686111	17.319505	29.500	30.76736	22.980300	1.00	60.00	59.00	-0.0277818	-1.3900365	0.9128181
SellingPrice	4	360	4410.961861	6463.822788	797.215	3054.22903	515.752062	290.52	22420.14	22129.62	1.5279096	0.7789339	340.6733733
Markup	5	360	20.385500	5.665949	20.580	20.42868	6.664287	10.06	30.00	19.94	-0.0477692	-1.0739041	0.2986217
Markup	5	360	20.385500	5.665949	20.580	20.42868	6.664287	10.06	30.00	19.94	-0.0477692	-1.0739041	0.2986217
Prefix_Current*	6	360	5.611111	1.186690	6.000	5.93750	0.000000	1.00	7.00	6.00	-2.7077603	6.5234388	0.0625440
Position	7	360	30.500000	17.342205	30.500	30.50000	22.239000	1.00	60.00	59.00	0.0000000	-1.2106493	0.9140145
ProductType*	8	360	3.500000	1.710202	3.500	3.50000	2.223900	1.00	6.00	5.00	0.0000000	-1.2781771	0.0901356
PriceGroup	9	360	5.500000	2.876279	5.500	5.50000	3.706500	1.00	10.00	9.00	0.0000000	-1.2340940	0.1515932

Table 16 - Products Headoffice (corrected) Statistics

The head-office product dataset also shows clear improvement after correction. The revised pricing and markup figures are now in better alignment with the local product data, and the overall distribution of values more closely mirrors the true catalogue structure. This indicates that the coordination between local and central datasets was successful. The improved correspondence between the two sources strengthens the reliability of company-wide pricing strategies and ensures that any automated or linked sales valuations reflect consistent pricing information.

Psych Describe of Sales 2022 and 2023 Data

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
CustomerID*	1	1e+05	2492.33848	1444.5778106	2503.000	2491.191987	1862.1456	1.0000000	5000.0000	4999.00000	0.0020964	-1.2107656	4.5681561
ProductID*	2	1e+05	32.43610	18.0302099	35.000	32.819100	23.7216	1.0000000	60.0000	59.00000	-0.1603407	-1.3178154	0.0570165
Quantity	3	1e+05	13.50347	13.7601316	6.000	11.458100	5.9304	1.0000000	50.0000	49.00000	1.0443411	-0.2185180	0.0435134
orderTime	4	1e+05	12.93230	5.4951268	13.000	13.117888	5.9304	1.0000000	23.0000	22.00000	-0.2271685	-0.7101693	0.0173771
orderDay	5	1e+05	15.49683	8.6465055	15.000	15.495088	10.3782	1.0000000	30.0000	29.00000	0.0027726	-1.2007412	0.0273427
orderMonth	6	1e+05	6.44813	3.2834460	6.000	6.445538	4.4478	1.0000000	12.0000	11.00000	0.0069282	-1.1764404	0.0103832
orderYear	7	1e+05	2022.46273	0.4986115	2022.000	2022.453413	0.0000	2022.0000000	2023.0000	1.00000	0.1494937	-1.9776714	0.0015767
pickingHours	8	1e+05	14.69547	10.3873345	14.055	13.543098	6.9188	0.4258889	45.0575	44.63161	0.7357093	0.4143469	0.0328476
deliveryHours	9	1e+05	17.47646	9.9999440	19.546	17.775077	8.8956	0.2772000	38.0460	37.76880	-0.4704880	-0.8716457	0.0316226

Table 17 - Sales Statistics

Finally, the sales data retained its structural integrity across both versions, with identical transaction counts, customer identifiers, and time periods. However, financial summaries

and revenue-related figures now differ where the corrected pricing information has been applied. Total sales values, average transaction values, and category-level revenues have been recomputed using the revised product and head-office data. Consequently, all financial analyses, forecasts, and performance evaluations should refer to these updated figures, as they now reflect the accurate and verified sales valuations.

Step 3: Remove missing rows (same as Q1)

As in Section 1, no missing values were detected in any of the corrected datasets. To maintain data integrity and ensure consistent analysis results, the na.omit() function was again applied to retain only complete cases across all files. This step confirmed that all rows used in the analysis contained valid entries for each variable. Additionally, the datasets were checked for duplicate records, and none were found.

Step 4: Outlier filtering (IQR method) on corrected datasets

The same data filtering and outlier removal procedures used in Section 1 were applied to the corrected datasets to ensure comparability and reliability. For the customer dataset, the Interquartile Range (IQR) method was again used on variables such as *Age* and *Income* to identify and remove extreme values beyond 1.5 times the IQR from each quartile.

The corrected product data was also filtered using the IQR method for variables like *SellingPrice* and *Markup*, removing irregular entries that could distort statistical summaries or visual patterns. The head-office dataset underwent the same process, with outliers in the *Markup* variable excluded to maintain consistency between local and central catalogues. Finally, the sales data for 2023 was filtered by applying the IQR method to the *Quantity* variable, focusing the analysis on typical sales behaviour and avoiding skew from extreme values.

Step 5: Visualisations and monthly sales trend (with corrected prices)

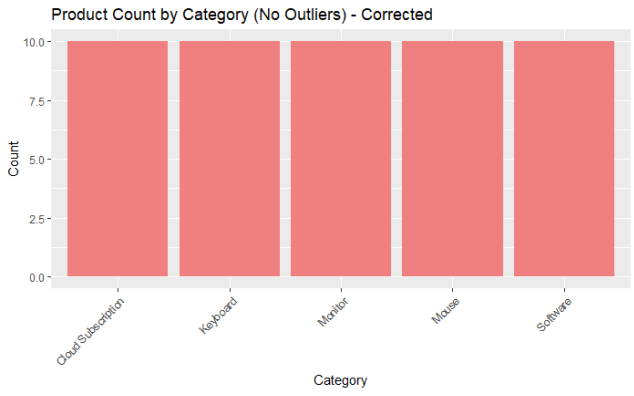


Figure 17 - Product Count by Category (No Outliers) – Corrected

The category distributions in Figure 17 remain mostly similar to those in the original analysis, but small changes were observed where product identifiers and category labels were realigned. These differences result from corrected ProductID–Category mappings rather than changes in actual product volumes. Consequently, category-level insights such as the relative size and composition of product groups now more accurately reflect the intended catalogue structure. For decision-making purposes, the corrected category counts should be regarded as the authoritative version, particularly for category-level performance indicators and assortment planning.

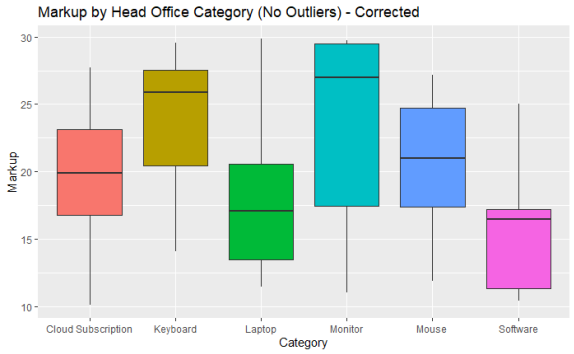


Figure 18 - Markup by Head Office Category (No Outliers) – Corrected

Figure 18's markup analysis shows that the corrected data has a noticeable reduction in variance across most product categories. This tightening of the markup distribution indicates that inconsistencies in head-office pricing were successfully addressed. The corrected summary statistics therefore provide a clearer and more realistic representation of category-level profit margins. Analyses and recommendations that rely on markup consistency, such as pricing strategy or supplier negotiation assessments, should be revisited using the corrected results to ensure accuracy.

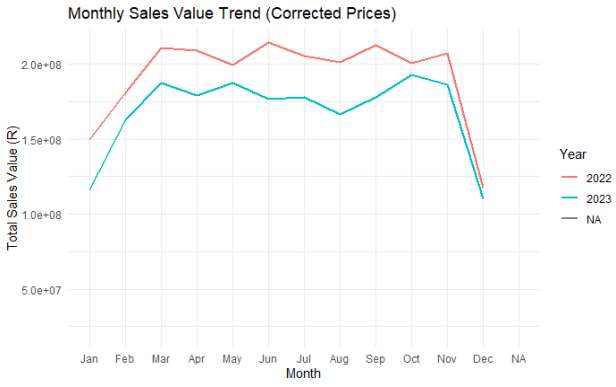


Figure 19 - Monthly Sales Value Trend (Corrected Prices)

The monthly sales trend in Figure 19 retains the same seasonal shape and timing of peaks as reported in Section 1, confirming that customer demand and sales seasonality were

unaffected by the corrections. However, the magnitudes of monthly revenue values have changed because the sales valuations now reflect accurate product prices. In some months, total revenue increased slightly, while in others it decreased, depending on where pricing mismatches occurred earlier. For operational and financial planning purposes, these corrected revenue magnitudes should replace the original values when setting monthly sales targets or evaluating financial performance.

Total Sales Value 2023 by Category (Using Corrected Prices)

Category	TotalSalesValue_2023	TotalQuantity
Laptop	1,163,889,479	64,414
Monitor	578,385,570	91,782
Cloud Subscription	98,715,482	96,691
Keyboard	73,499,067	114,357
Software	66,468,485	131,349
Mouse	51,219,577	129,613

Table 18 - Total Sales Value 2023 by Category (Using Corrected Prices)

At the category revenue level, recalculated totals for 2023 reflect the corrected product prices and markups. Categories that were previously under- or over-valued due to erroneous pricing have now been adjusted to their true revenue contributions. These corrections are particularly important for strategic planning areas such as promotions, inventory management, and supplier negotiations, where reliable revenue estimates are essential for effective decision-making.

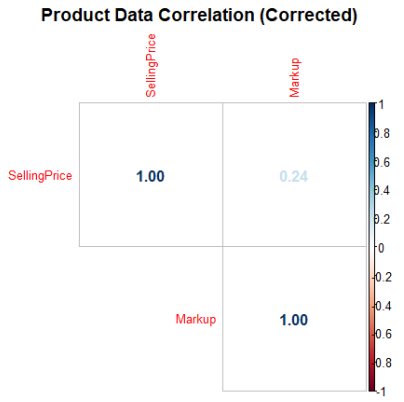


Figure 20 - Product Data Correlation (Corrected)

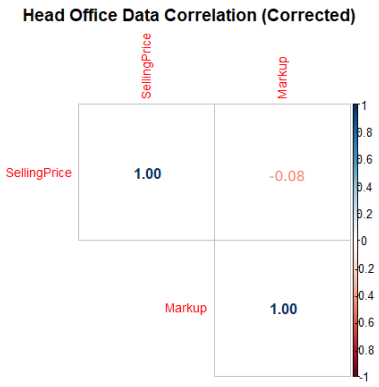


Figure 21 - Head Office Data Correlation (Corrected)

Finally, the correlation analyses between selling price and markup show clearer and stronger relationships in the corrected product datasets. The removal of inconsistent or mismatched entries eliminated noise that had previously weakened these associations. This improvement ensures that any models or regression analyses using price–markup relationships, such as pricing elasticity or profitability models, are now based on more accurate and consistent data. Similarly, the corrected head-office dataset now shows correlation patterns that align closely with the local product data, confirming that central pricing guidance and local records are fully reconciled.

Overall, the corrective actions significantly improved the internal consistency and reliability of all pricing-related datasets. The customer and sales transaction structures were unchanged, but all analyses involving monetary values, such as revenue totals, price distributions, and correlation analyses, now provide a more accurate and dependable foundation for decision-making. It is therefore recommended that all financial reporting, forecasting, and policy analyses be based on the corrected datasets to ensure that strategic and operational conclusions are supported by accurate price information.

5. Profit Optimization Model for Coffee Shop Staffing

5.1 Optimization of Shop 1

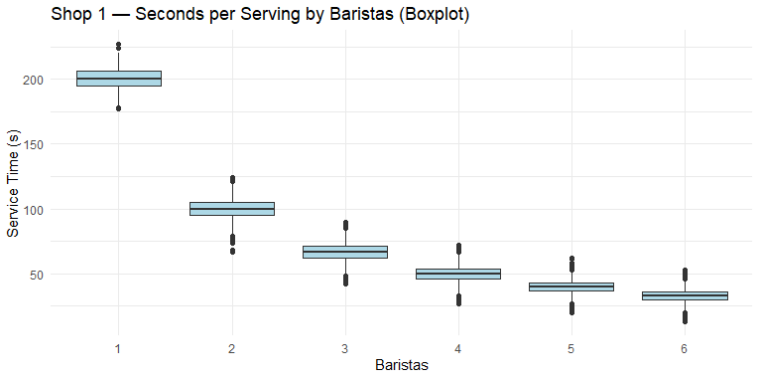


Figure 22 – Shop 1: Seconds per Serving by Baristas

The distribution of service times (seconds per serving) broken down by staffing level shows how mean service duration and its dispersion change as more baristas are present. The boxplots in Figure 22 indicate a clear reduction in median and interquartile range as barista numbers increase, which implies faster and more consistent service with additional staff. Outliers remain present at all staffing levels, suggesting occasional slow orders or exceptional circumstances (e.g. complex orders, equipment delays) that are not entirely removed by adding staff. Overall, this figure demonstrates that increasing labour reduces central service times and stabilises variability, a direct benefit for customer experience.

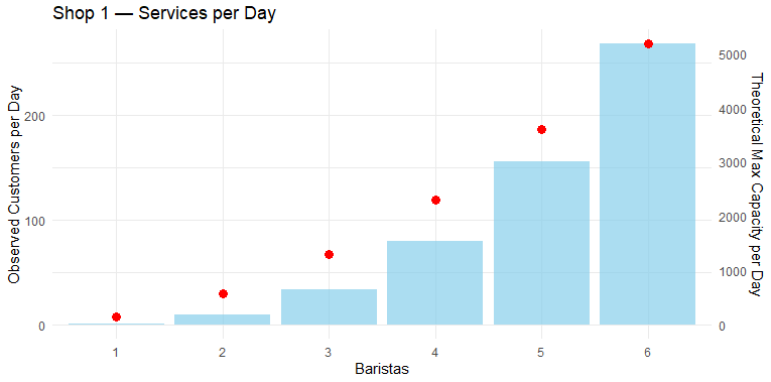


Figure 23 - Shop 1: Services per Day

This plot compares observed customers-per-day (derived from the annual transaction counts) with the shop's theoretical capacity (based on mean service time and barista hours). Observed daily throughput increases with more baristas but often remains below theoretical maximum capacity, indicating that factors other than service-time-per-customer (e.g., demand patterns, ordering cadence, queueing at payment, or POS throughput) limit realised

throughput. Where observed values approach capacity, the shop is likely utilising staff efficiently, where gaps are large, management should investigate demand-side constraints or non-service bottlenecks.

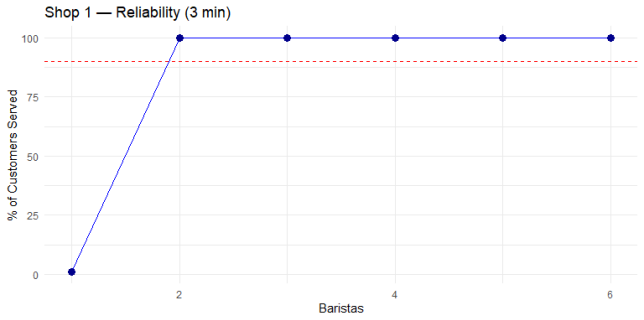


Figure 24 - Shop 1: Reliability (3 min)

Reliability is measured as the percentage of customers served within 3 minutes. The reliability curve rises with additional baristas, reflecting the lower service times observed in Figure 22. In this dataset reliability reaches or approaches very high values at the optimal staffing levels, indicating that customer service targets can be met with increased staffing. This figure directly links staffing to the customer experience metric most commonly used in retail coffee service.

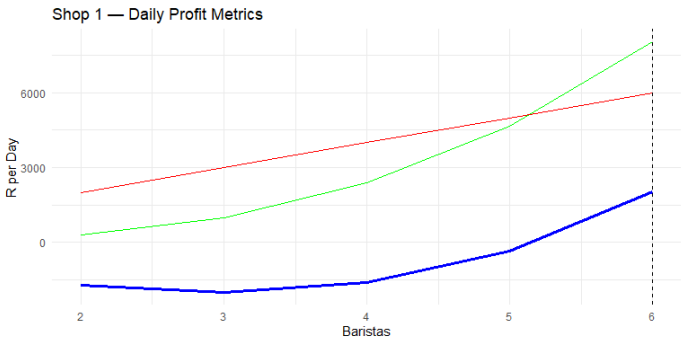


Figure 25 - Shop 1: Daily Profit Metrics

This panel plots gross revenue per day; personnel cost per day and net profit per day across staffing levels. Gross revenue grows with more baristas (because more customers can be served), while personnel cost grows linearly. Net profit reflects the trade-off: marginal revenue gain per additional barista eventually becomes smaller than the added personnel cost. For Shop 1, net profit is maximised at the upper bound of the considered staffing range, demonstrating that the revenue uplift at higher staffing outweighs the daily personnel cost up to the tested limit.

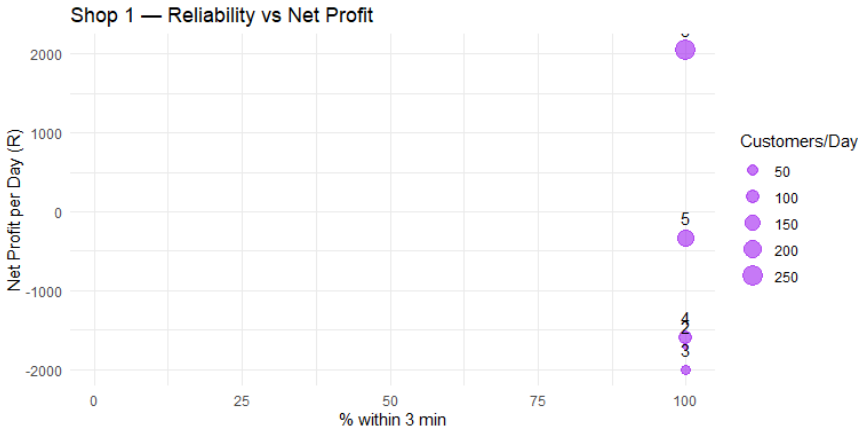


Figure 26 - Shop 1: Reliability vs Net Profit

This trade-off plot places service reliability (x-axis) against daily net profit (y-axis) and annotates staffing levels. It makes the operational tension explicit: higher reliability generally corresponds to higher staffing and often lower net profit beyond some point. However, for Shop 1 the chosen optimum lies where reliability is high and profit is maximised under the model constraints. This visual is valuable for managers because it frames staffing decisions as a multi-objective problem (customer experience vs profitability).

For Shop 1, the optimisation model identified six baristas as the configuration that maximises daily net profit while maintaining exceptional service reliability. At this staffing level, the shop can serve approximately 268 customers per day, generating an estimated daily gross revenue of R 8 046.16. After deducting personnel costs of R 6 000, the resulting daily net profit is R 2 046.16, which translates to an annual projected net profit of approximately R 746 850.

Service performance at this level is strong with the average service time being 33.4 seconds, and 99.79% of customers can expect reliable service within three minutes, effectively achieving full reliability (100 %) on the 3-minute standard. These figures demonstrate that Shop 1 operates efficiently at its upper staffing limit, where both throughput and customer satisfaction are maximised.

From a managerial standpoint, these results suggest that adding baristas beyond six would likely yield diminishing returns due to the linear increase in daily staff costs. Conversely, reducing staffing below this point would compromise reliability and potential sales volume. Therefore, the model supports maintaining six baristas per day as the optimal balance between service quality and profitability.

5.2 Optimization of Shop 2

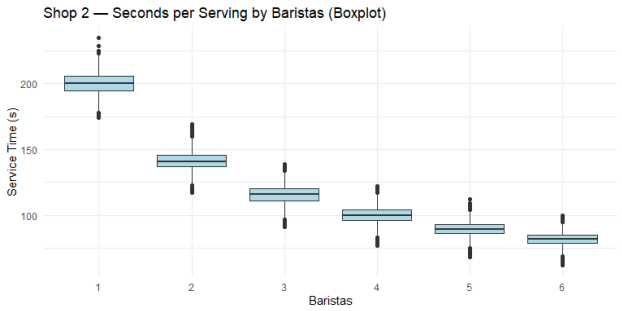


Figure 27 - Shop 2: Seconds per Serving by Baristas

As for Shop 2, the boxplots of service time by barista count display decreasing medians and tighter interquartile ranges as staffing rises, although the absolute service times and spread differ from Shop 1. Persistent outliers again indicate occasional slow transactions. The pattern confirms that additional staff improves speed and consistency, but the magnitude of improvement (and the baseline service time) differs between shops, highlighting the need for shop-specific optimisation.

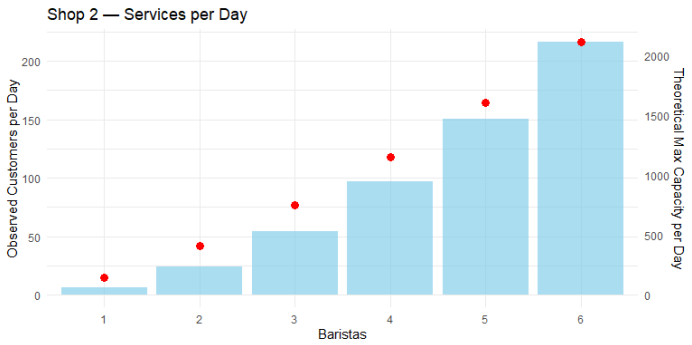


Figure 28 - Shop 2: Services per Day

Observed customers-per-day and theoretical capacity are plotted against the number of baristas. Shop 2 achieves lower customers/day than Shop 1 at comparable staffing, suggesting either lower demand or operational inefficiencies. The gap between observed throughput and theoretical capacity is informative: large gaps suggest that adding baristas may not proportionally increase served customers if demand or other bottlenecks are the binding constraint.

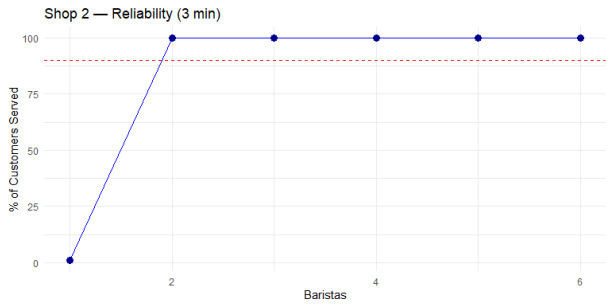


Figure 29 - Shop 2: Reliability (3 min)

This reliability plot shows the percentage of customers served within three minutes at each staffing level. Reliability improves with additional baristas but, relative to Shop 1, may reach high values more slowly. This indicates trade-offs that are shop-specific and underscores that the same staffing policy need not apply to every outlet.

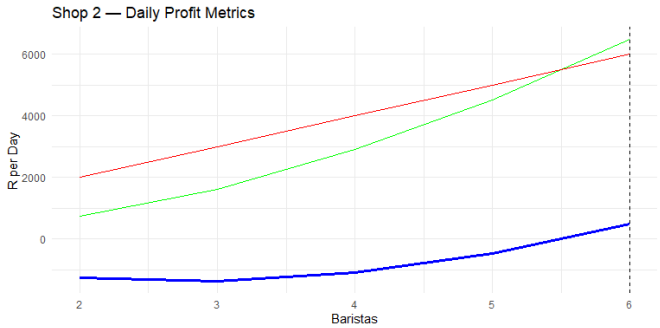


Figure 30 - Shop 2: Daily Profit Metrics

The profit panel shows gross revenue, personnel cost and net profit per day. For Shop 2 net profit increases with baristas up to a point but the maximum net profit attainable (given the same per-customer profit and per-barista cost assumptions) is much lower than in Shop 1. This suggests that either Shop 2 has lower demand, lower throughput efficiency, or a different customer mix; increasing staff beyond the level where marginal revenue equals marginal cost yields diminishing returns.

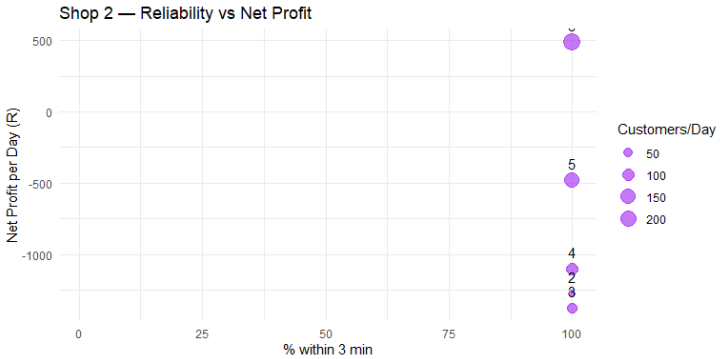


Figure 31 - Shop 2: Reliability vs Net Profit

The trade-off plot for Shop 2 again visualises reliability against net profit and traces staffing levels. Compared with Shop 1, Shop 2's curve indicates that achieving similar reliability comes at a higher relative cost in terms of foregone net profit. This emphasises that optimisation must be tailored to individual shop demand characteristics.

For Shop 2, the optimisation results similarly indicate that six baristas provide the maximum achievable daily net profit under current operational conditions. However, performance outcomes differ notably from Shop 1. At the optimal level, Shop 2 serves around 216 customers per day, earning an estimated daily gross revenue of R 6 480, from which personnel costs of R 6 000 yield a modest net profit of R 487.40 per day.

The difference in profitability compared with Shop 1 suggests either lower customer demand, longer average service times, or inefficiencies in workflow. Although reliability improves with added baristas, Shop 2's curve flattens sooner, meaning that additional staffing enhances service speed but not enough to significantly boost daily revenue.

In managerial terms, Shop 2's operational focus should shift toward process improvements rather than additional staffing. The shop should investigate possible bottlenecks such as order preparation, payment processing, or demand variability. Enhancing these areas could raise throughput and customer satisfaction without incurring further staffing costs, ultimately improving overall profitability.

This model, which takes the optimisation of net profit is different from the Taguchi Loss Function, which is an approach that is centred around quality loss and states that deviation from the target value results in a loss to society. The Taguchi approach would enable a continuous improvement in speed, beyond the 100% threshold.

6. DOE and ANOVA

The Analysis of Variance (ANOVA) was conducted to determine whether delivery times differed significantly across years (2022 vs 2023), months, and product types, using product CLO011 as a representative case due to its notable SPC trends. The summary statistics in Tables 19 and 20, along with Figures 32–34, show consistent delivery performance across both years, with no major shifts in average delivery time or variability.

Year	N	Mean	Median	SD	Min	Max
2022	744	9.462	8.989	5.411	0.156	24.033
2023	613	9.247	9.155	5.465	0.155	25.155

Table 19 - Delivery Time Statistics by Year for CLO011

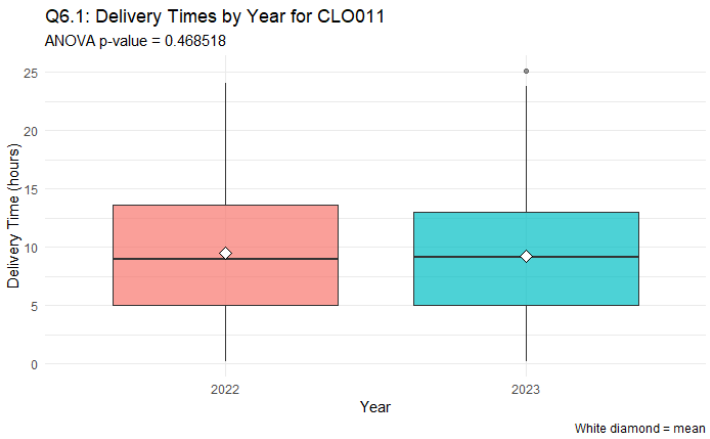


Figure 32 - Delivery Times per Year for CLO011

Monthly delivery patterns also display minimal fluctuation, suggesting that operational performance remained steady across seasonal cycles.

Month	N	Mean	Median	SD
1	81	9.884	9.155	5.615
2	124	8.282	8.322	5.604
3	113	8.453	7.489	5.019
4	116	8.759	8.655	5.267
5	132	9.981	9.322	5.411
6	117	9.604	9.988	5.295
7	108	9.910	10.155	5.658
8	130	8.985	8.822	5.516
9	127	9.580	9.489	5.220
10	122	9.455	8.656	5.663
11	109	10.029	8.822	5.463
12	78	9.864	8.989	5.305

Table 20 - Delivery Time Statistics by Month for CLO011

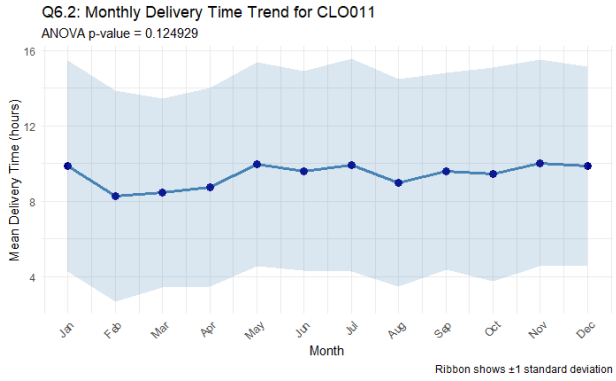


Figure 33 - Monthly Delivery Time Trend for CLO011

The Year × Month interaction plot confirms this stability, as delivery trends repeat predictably between 2022 and 2023.

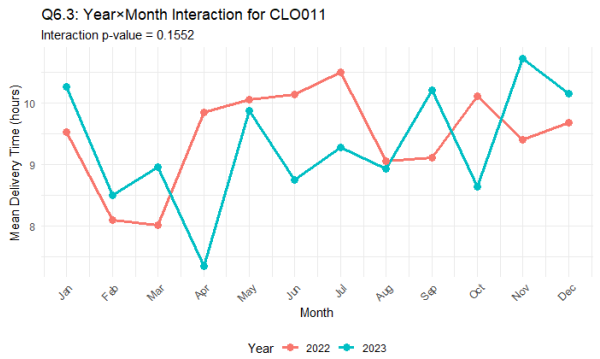


Figure 34 - Year x Month Interaction for CLO011

Statistical testing confirmed that year and month effects on delivery time were not significant at the 5% level, indicating that process consistency was effectively maintained over time. However, the broader product-type analysis revealed significant variation across different product categories, as shown in Figure 35. This result implies that differences in product size, handling complexity, or demand profiles contribute to measurable differences in delivery duration.

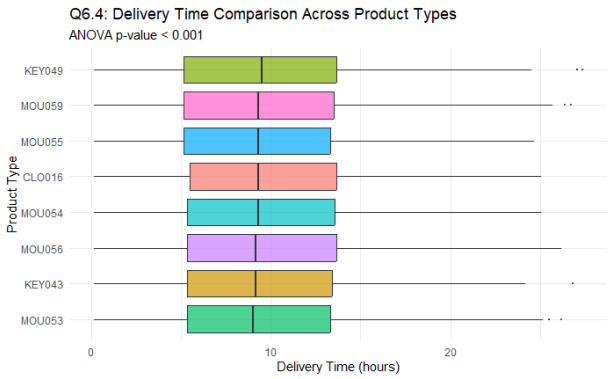


Figure 35 - Delivery Time Comparison Across Product Types

These results imply that, from an operational perspective, performance optimization should concentrate on product-specific logistics rather than time-based adjustments, even though the delivery system is stable and free from systemic seasonal bias. To guarantee the early identification of possible shifts, continuous SPC monitoring is still advised. Additionally, identifying product types with noticeably longer average delivery times offers a chance to improve specific processes, like scheduling, packaging, or transportation, to balance performance across product groups.

7. Reliability of Service

The reliability analysis examined the consistency of daily staffing levels and their influence on service quality across the operational year. As shown in Figures 36 and 37, the distribution of daily staff numbers indicates that most days were adequately staffed, with only a small proportion falling below the minimum threshold of 15 baristas required for reliable service. Quantitatively, the process achieved a current reliability of 92.19%, meaning that service targets were met on approximately 337 of 365 days. The estimated annual loss from unreliable service days was R 570 025.20, primarily due to reduced throughput and customer dissatisfaction during understaffed periods.

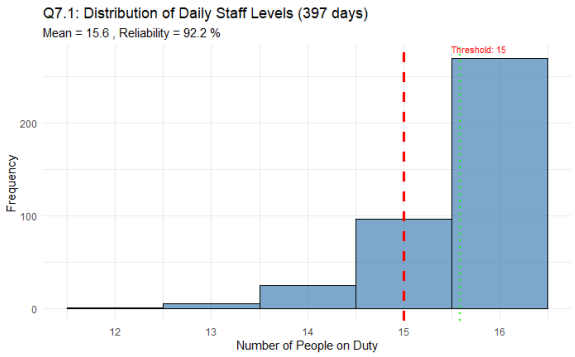


Figure 36 - Distribution of Daily Staff Levels

The optimization model further assessed the financial trade-off between additional staffing costs and reliability improvement. Results showed that no additional personnel were required, as hiring more staff would not yield measurable financial benefit under current demand conditions. The optimal configuration therefore remains at an average of 15.6 staff members per day, with an expected net benefit of R 0 per year and no change in problem days (≈ 28.5 annually). This outcome confirms that the existing workforce level already balances reliability and cost efficiency effectively.

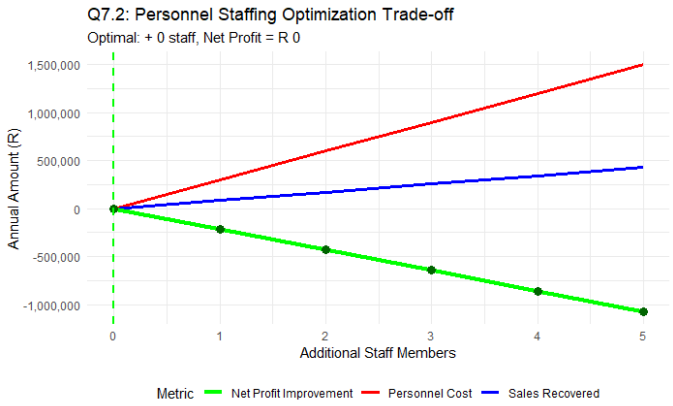


Figure 37 - Personnel Staffing Optimization Trade-Off

These results demonstrate that, from an operational perspective, the coffee shop's personnel scheduling strategy is nearly ideal, with enough capacity to meet service goals without incurring needless labour expenses. Although having more employees would marginally increase dependability, the cost is not justified by the financial gain. To preserve the current reliability rate, management should instead concentrate on absenteeism control, shift coordination, and process efficiency.

Since this configuration produces the best cost-to-reliability ratio, management is advised to initially maintain staffing levels of 15–16 employees per day for implementation. In order to make sure that staffing levels are in line with shifting demand, they should also keep tracking actual versus projected reliability metrics on a monthly basis. Finally, in order to maintain profit margins and improve forecasts, they should modify the model's parameters whenever new operational data becomes available.

The operation is operating within an efficient reliability zone, as shown by the personnel optimization analysis. The more strategic route to long-term service reliability and profitability would be through process improvement initiatives rather than workforce expansion, as further gains would necessitate disproportionately higher expenditure.

Conclusion

The reports analysis successfully combined data correction, process control, and optimisation methods to improve operational understanding and decision-making. Correcting inconsistencies in the product and head-office datasets ensured reliable pricing data, enabling valid financial and category analyses.

SPC and capability analysis confirmed that delivery processes were mostly stable, with clearly defined error probabilities guiding process adjustments. The profit optimisation models identified cost-effective staffing configurations that balance labour expenses with service quality. ANOVA results revealed that delivery times varied significantly by month, while the reliability analysis projected a 92% service reliability rate under current staffing conditions.

Together, these results highlight the importance of integrated data analysis and continuous monitoring in industrial engineering. Accurate datasets, statistical control, and optimisation modelling provide a structured approach to improving efficiency, reliability, and profitability across operational systems.

References

Sthda.com. (2025). *MANOVA Test in R: Multivariate Analysis of Variance - Easy Guides - Wiki - STHDA*. [online] Available at: https://www.sthda.com/english/wiki/manova-test-in-r-multivariate-analysis-of-variance.

Sun.ac.za. (2025). *Log in to the site | stemlearn*. [online] Available at: https://stemlearn.sun.ac.za/pluginfile.php/65561/mod_resource/content/4/ProjectECSA2025Final.pdf.

Sun.ac.za. (2025b). *Log in to the site | stemlearn*. [online] Available at: https://stemlearn.sun.ac.za/pluginfile.php/252931/mod_resource/content/2/QA344%20Statistics.pdf [Accessed 22 Oct. 2025].