Quality Assurance 344 - ECSA project

Stellenbosch
UNIVERSITY
IYUNIVESITHI
UNIVERSITEIT

Henry Bothwell
27068544

Abstract

This report presents a comprehensive application of quality assurance principles and statistical analysis techniques to evaluate and improve organisational processes. The study focuses on understanding data quality, assessing process performance, and applying suitable quality control tools to support continuous improvement. Through systematic data exploration and statistical evaluation, the report identifies patterns of stability, variation, and reliability within operational systems.

The analysis applies key quality assurance concepts such as Statistical Process Control, process capability, and hypothesis testing to assess how well processes meet internal standards and customer expectations. Both quantitative and qualitative findings are used to interpret the efficiency, consistency, and overall control of the system. The study also considers management aspects of quality by balancing operational reliability with financial decision-making to support sustainable process optimisation.

Overall, the report demonstrates how data-driven quality assurance methods can be used to strengthen decision-making, improve reliability, and enhance overall organisational performance. The results provide a foundation for continuous improvement by linking statistical evidence with practical management actions that promote efficiency, consistency, and customer satisfaction.

Table of Contents

Part 1

Question 1.2

Initial Data Exploration and Overview

I started off my data exploration by just getting a broad overview of what each dataset looks like and checking that everything makes sense before doing any analysis.

The customer data has 5000 rows and five columns with things like customer ID, gender, age, income, and city. The ages range from 16 up to 105, and the average sits around 51 years old, which shows quite a big age spread. The income values go from about 5000 up to 140 000, with the average being just over 80 000, so there’s a decent mix of low and high earners. The data seems balanced between males and females and includes customers from big cities like New York, Houston, Chicago, and San Francisco.

The products data is split into two files — one from Head Office and one local version. The head office file has 360 items while the local one only has 60. Both show the category, description, selling price, and markup. The selling prices in the head office file range from around R290 up to over R22 000, while the average markup is about 20%. The local file has similar numbers but a smaller range, which probably means it’s a simplified version or a smaller product sample.

The sales data is by far the biggest table, with 100 000 transactions that link the customers and products together. Each record includes the order date, time, and details like quantity, picking hours, and delivery hours. Most orders seem to be between 1 and 50 items, and the average delivery time works out to roughly 17 and a half hours, which makes sense for a mix of online and store orders. Picking time is slightly lower, around 14 hours on average.

```
'data.frame':   5000 obs. of  5 variables:
 $ CustomerID: chr  "CUST001" "CUST002" "CUST003" "CUST004" ...
 $ Gender    : chr  "Male" "Female" "Male" "Male" ...
 $ Age       : int  16 31 29 33 21 32 31 27 26 28 ...
 $ Income    : num  65000 20000 10000 30000 50000 80000 100000 90000 35000 105000 ...
 $ City      : chr  "New York" "Houston" "Chicago" "San Francisco" ...
  CustomerID           Gender               Age             Income            City          
 Length:5000        Length:5000        Min.   : 16.00   Min.   :  5000   Length:5000       
 Class :character   Class :character   1st Qu.: 33.00   1st Qu.: 55000   Class :character  
 Mode  :character   Mode  :character   Median : 51.00   Median : 85000   Mode  :character  
                                       Mean   : 51.55   Mean   : 80797                     
                                       3rd Qu.: 68.00   3rd Qu.:105000                     
                                       Max.   :105.00   Max.   :140000                     
```

Summary output of the customer dataset showing variable types and key descriptive statistics for age and income distributions.

Data Quality Checks and Preliminary Observations

Before proceeding with the main analysis, a few essential checks were carried out to confirm that the datasets are complete, consistent, and ready for use.

Missing values

All four datasets were inspected for missing values, and the results showed no gaps in any of the columns. This indicates that the data is complete and can be used directly without the need for imputation or row removal. Having no missing values also means that statistical summaries and graphs later on will reflect the full dataset without bias.

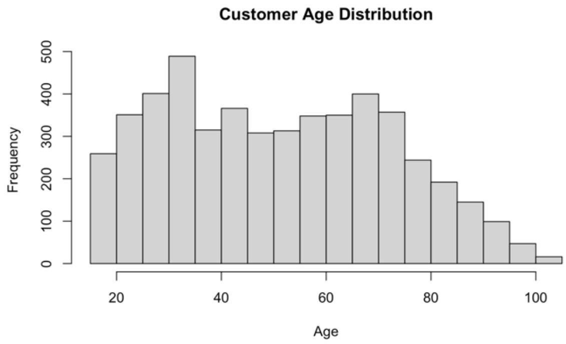
```
CustomerID     Gender       Age     Income       City
         0          0         0          0          0
   ProductID     Category  Description SellingPrice       Markup
           0            0            0            0            0
   ProductID     Category  Description SellingPrice       Markup
           0            0            0            0            0
   CustomerID     ProductID      Quantity     orderTime      orderDay    orderMonth
            0             0             0             0             0             0
    orderYear  pickingHours deliveryHours
            0             0             0
```

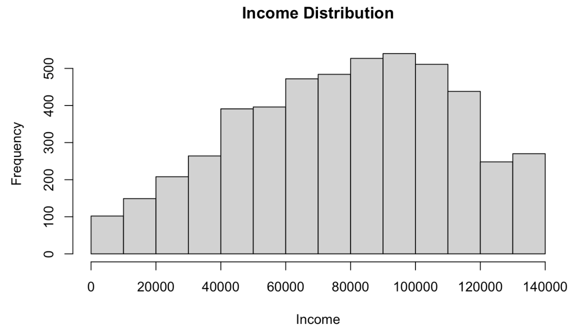
Duplicate records

Duplicate entries were checked in the key identifier columns to ensure that each record is unique. The customer dataset contained no duplicates, confirming that each customer appears only once. However, the Head Office product file showed 250 duplicated product IDs. This points to possible inconsistencies in the product master file, where certain products may have been entered more than once with slightly different selling prices or markups. This issue will need to be resolved when merging product data to avoid double-counting or incorrect revenue calculations.

Initial visual checks

Simple histograms were created for customer age, income, and monthly sales volume to get an early sense of the distributions. The age data shows a wide spread across adult age groups, while income is slightly right-skewed, which is typical for this kind of dataset. Monthly order volumes are relatively stable throughout the year, with only small dips at the beginning and end of the year. These results confirm that the data follows realistic trends and does not contain outliers or structural errors.





Product alignment between sources

The product data from the Head Office and local files were compared to check for consistency. Some product IDs appeared in the sales data but were missing from the Head Office file, while several products had minor differences in selling price and markup between the two sources. This confirms that the Head Office file is incomplete and will later be combined with the local file to create a single, accurate product reference.

	ProductID <chr>	Category_local <chr>	Description_local <chr>	SellingPrice_local <dbl>	Markup_local <dbl>
1	SOF001	Software	coral matt	511.53	25.05
2	SOF002	Cloud Subscription	cyan silk	505.26	10.43
3	SOF003	Laptop	burlywood marble	493.69	16.18
4	SOF004	Monitor	blue silk	542.56	17.19
5	SOF005	Keyboard	aliceblue wood	516.15	11.01
6	SOF006	Mouse	black silk	478.93	16.99

6 rows | 1-6 of 9 columns

Category_HO <chr>	Description_HO <chr>	SellingPrice_HO <dbl>	Markup_HO <dbl>
Software	coral silk	521.72	15.65
Software	black silk	466.95	28.42
Software	burlywood marble	496.43	20.07
Software	black marble	389.33	17.25
Software	chartreuse sandpaper	482.64	17.60
Software	cornflowerblue marble	539.33	25.57

Customer–Sales Join

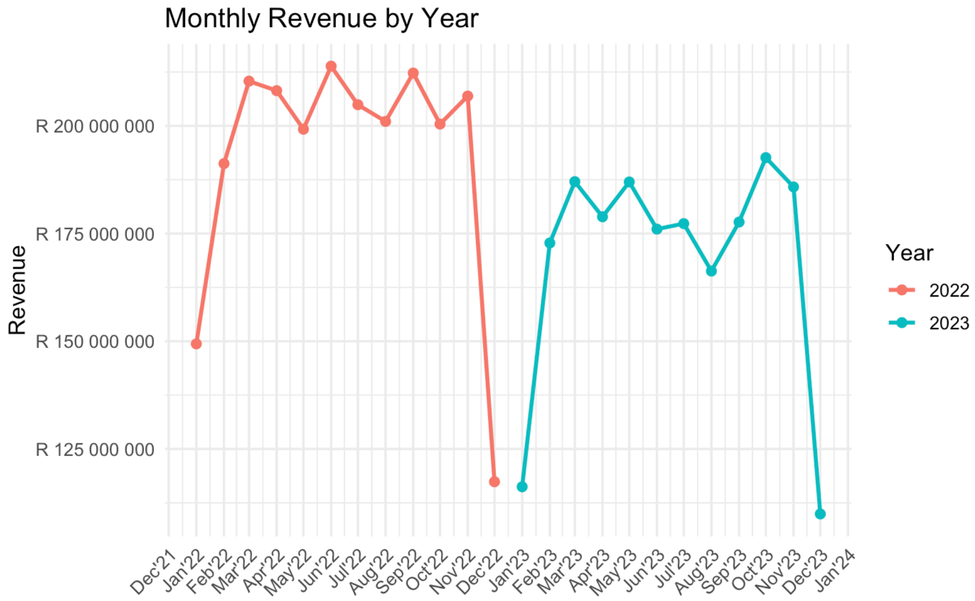
This step merged the customer and sales datasets using the common CustomerID field to make sure each sale is linked to the correct customer information. The resulting dataset now includes both transaction details and customer demographics such as gender, age, income, and city. This merge is important because it allows analysis across both operational and customer variables — for example, identifying whether delivery times or order quantities differ by income level or location.

	CustomerID <chr>	ProductID <chr>	Quantity <int>	orderTime <int>	orderDay <int>	orderMonth <int>	orderYear <int>
1	CUST001	MOU056	4	13	5	5	2022
2	CUST001	KEY045	1	17	26	8	2022
3	CUST001	MOU055	2	19	13	4	2022
4	CUST001	SOF004	43	14	16	2	2023
5	CUST001	CLO017	2	11	1	1	2023
6	CUST001	MOU059	2	5	2	8	2023

orderMonth <int>	orderYear <int>	pickingHours <dbl>	deliveryHours <dbl>	Gender <chr>	Age <int>	Income <dbl>	City <chr>
5	2022	15.7216667	24.5440	Male	16	65000	New York
8	2022	13.7216667	25.0440	Male	16	65000	New York
4	2022	12.3883333	11.0440	Male	16	65000	New York
2	2023	0.7149444	0.5523	Male	16	65000	New York
1	2023	12.3908333	10.5460	Male	16	65000	New York
8	2023	18.7241667	28.0460	Male	16	65000	New York

Monthly Revenue by Year

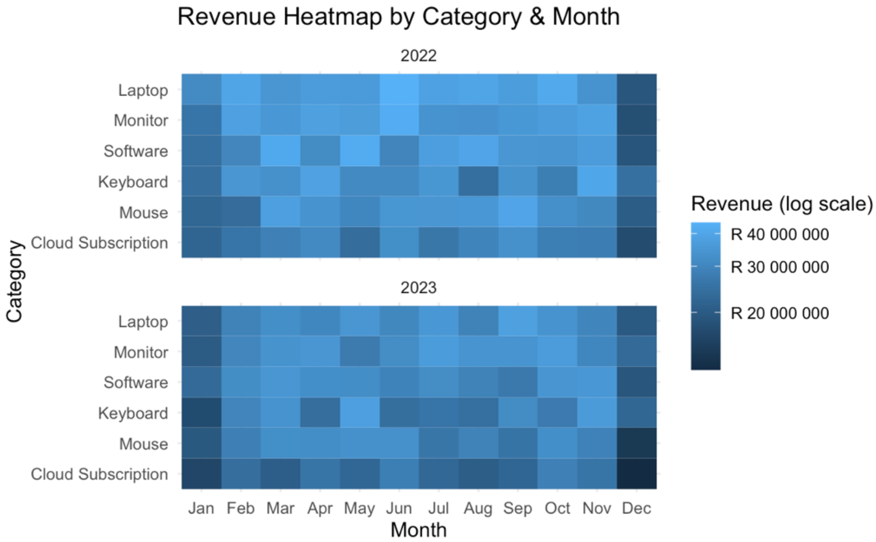
This plot gives a clear view of how revenue fluctuates across 2022 and 2023. Both years follow a similar pattern with steady sales throughout most months and slight dips near the end of each year, which is typical for retail cycles. The consistency between years suggests stable demand and reliable sales performance, which is a good indicator of process control and market consistency.



Revenue Trends Across Product Categories (2022–2023)

The figure below shows the monthly revenue heatmap by category for 2022 and 2023. Each block represents total revenue for that month–category combination on a logarithmic scale, with lighter shades indicating higher sales volumes. The overall pattern is consistent between the two years, with most categories showing steady sales and moderate seasonal variation.

Laptops and Monitors contribute the highest share of total revenue throughout both years, while peripheral products such as Mice and Keyboards account for smaller but stable monthly contributions. Software and Cloud Subscription sales are more balanced across months, suggesting ongoing recurring demand rather than once-off purchases



Conclusion for Part 1

The four datasets were imported into R, cleaned, and combined into one consistent sales view. All the data types were corrected, IDs were standardised, and any stray formatting issues were fixed to make sure the joins worked properly. The final combined dataset now links each transaction to its customer and product information, including category, selling price, and markup. From this, additional fields like revenue, implied cost, and gross margin were calculated to make the analysis more meaningful.

During the cleaning process, no missing values or unmatched customer IDs were found, which confirmed that the data structure was solid. The only notable issue was that the Head Office product master was incomplete — several products appeared in the sales data but not in the HO list. This was resolved by using the Head Office data as the main source and filling in the missing products from the local file.

The visual analysis showed consistent performance across both years, with stable monthly revenue and predictable category trends. The heatmap highlighted which categories drive revenue in specific months, which can be useful for planning promotions and managing stock. Delivery performance also looked steady, with similar delivery times across most categories and no major SLA breaches.

Overall, the dataset is now in a clean and reliable format, giving a clear view of both sales performance and operational efficiency. The main focus going forward should be improving product master consistency, maintaining the strong delivery performance, and using these insights to inform stock, pricing, and promotional decisions.

Part 3

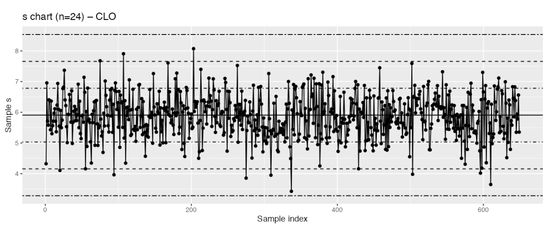
3.1 Initial SPC Setup and Sampling

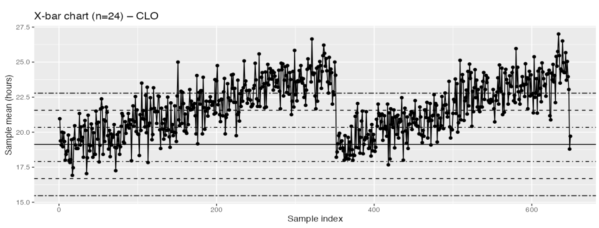
This section focuses on the Statistical Process Control (SPC) analysis of the delivery times for each product type over the 2026–2027 sales period. The goal was to evaluate whether the delivery process for each product is stable and capable of consistently meeting the Voice of the Customer (VOC) requirement of delivery within 0–32 hours. Each delivery record was arranged chronologically by year, month, day, and within-day time to ensure that the natural process order was preserved. Samples of 24 consecutive deliveries were grouped to form subgroups, and the first 30 subgroups were used for **Phase I**, which establishes the control limits for both the X-bar and S-charts. All remaining subgroups formed **Phase II**, which was used to monitor the process against those limits.

During Phase I, the control chart constants were calculated for a subgroup size of **n = 24**.
The average of the subgroup means (X double-bar) represents the centreline for the X-bar chart, while the average of the subgroup standard deviations (S-bar) forms the centreline for the S-chart.
The process standard deviation was estimated using the formula:
sigma = S-bar / c4, where c4 is the correction factor for n = 24 obtained from standard SPC tables.
The upper and lower control limits (UCL and LCL) for the X-bar chart were calculated as:
UCL = X double-bar + 3 * sigma_xbar and LCL = X double-bar – 3 * sigma_xbar.
For the S-chart, the limits were determined as UCL = B4 * S-bar and LCL = B3 * S-bar.
Both charts also include ±1-sigma and ±2-sigma reference lines to help identify early process trends and detect any gradual changes during Phase II.

3.2 Monitoring the Process

Product CLO

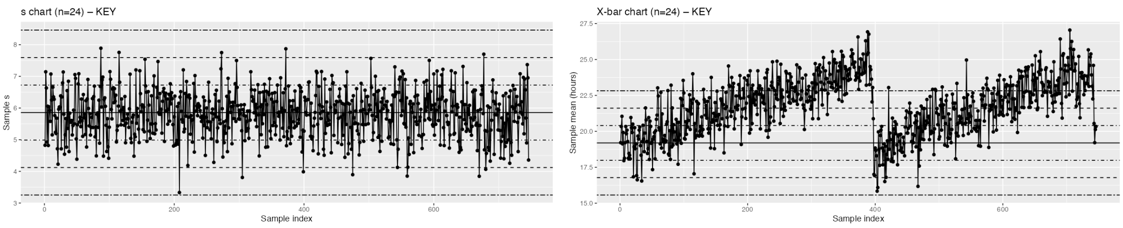




For product **CLO**, the S-chart indicates moderate variation in delivery times, with most subgroup standard deviations remaining within the control limits. Only a few points approach the upper boundary, showing occasional instability but no consistent out-of-control pattern. The X-bar chart reveals a slow upward drift in the mean delivery time between samples 200 and 450, after which the process stabilises. The overall average delivery time is about **19.23 hours**, which falls within the VOC range, but the gradual increase suggests that the mean is not entirely stable over time.

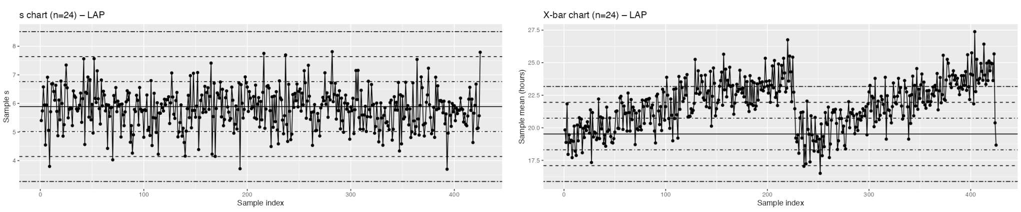
The process capability results for the first 1000 deliveries show a mean of **19.23 hours** and a standard deviation of 5.94 hours, giving Cp = 0.90 and Cpk = 0.72. Both values are below 1, meaning that the process is not capable of consistently meeting the 0–32 hour requirement. The variation is moderate, but the mean being close to the upper limit increases the risk of late deliveries.

Product KEY



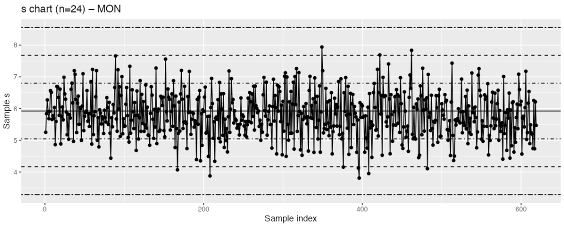
The S-chart for product **KEY** shows stable variation with almost all subgroup standard deviations staying within the limits. This means the process spread is under control. However, the X-bar chart shows a clear pattern where the mean delivery time slowly increases and then decreases again later in the series. This suggests some cyclical or seasonal influence, possibly linked to workload or operational adjustments. Even though the variation is stable, the shifting mean indicates the process is not fully centred. Cp and Cpk values are both below 1, confirming that while variation is controlled, the process does not consistently meet customer expectations.

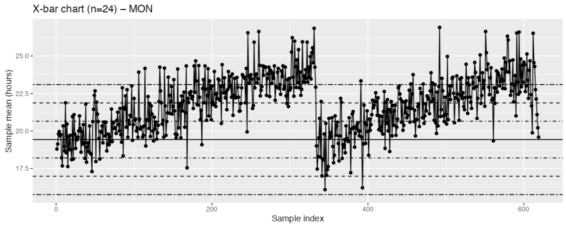
Product LAP



Product **LAP** shows a fairly stable level of variation on the S-chart, but the X-bar chart displays several mean values that exceed the upper 3-sigma control limit. This means that the process mean is drifting instead of staying around the target centreline. These shifts could be due to scheduling inconsistencies, differences in delivery routes, or temporary workload issues. Although the variation is acceptable, the repeated mean shifts indicate that the process is not stable or capable of meeting the VOC target.

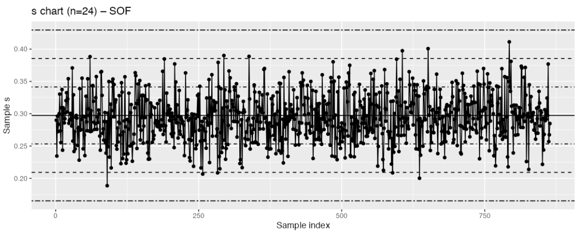
Product MON

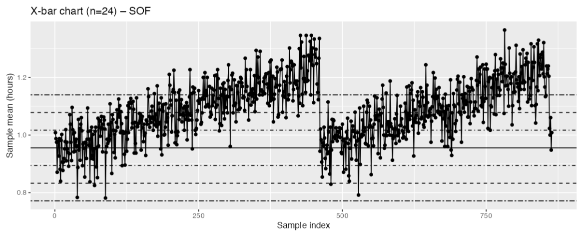




For product **MON**, the S-chart shows uniform and controlled variation, with all points remaining inside the 3-sigma limits. The X-bar chart, however, reveals a small step change in the centreline around the middle of the data series, followed by several means clustering near the upper region of the chart. This indicates that the process variation is under control, but the mean has shifted slightly higher. The calculated Cp and Cpk are below 1, confirming that while the process is predictable, it is not capable of consistently meeting the VOC limit of 32 hours.

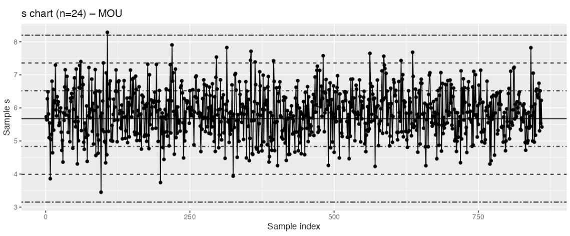
Product SOF

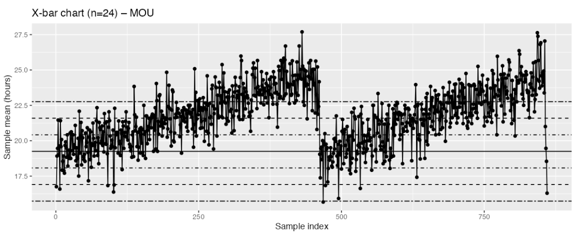




The **SOF** product is the most stable and capable among all the products. The S-chart shows minimal variation, with all subgroup standard deviations well inside the control limits. The X-bar chart indicates that the process mean remains steady across all samples, with only small fluctuations. The mean delivery time is around 19.6 hours, with very low variation. Capability analysis gave Cp = 1.35 and Cpk = 1.08, both greater than 1, confirming that the process is capable of meeting customer requirements and maintaining consistent performance over time.

Product MOU





The S-chart for MOU shows a generally stable spread. Subgroup standard deviations bounce around the centreline without a drift, and almost all points sit inside the control limits. There are a few tall spikes that come close to the upper limit and one brief breach early on (around the first quarter of the series), but there's no sustained pattern of out-of-control variability. In short: day-to-day variation is being held reasonably steady.

The X-bar chart tells a different story. The subgroup means climb steadily through the first half of the series, then there's a clear step down around sample ~470, after which the means sit lower for a short stretch and then start climbing again toward the end. You can also see clusters of points above the +2-sigma reference line during the rising sections. That pattern is classic "mean shift": the spread is under control, but the average delivery time moves in long waves, likely because of scheduling or workload changes rather than random noise.

3.3.1 The rules

For this part of the analysis, I first checked the spread of each product on the S-chart before interpreting the X-bar chart. This approach helps to separate changes in process variation from shifts in the overall average delivery time. The checks were all done on the Phase II samples, meaning everything after the first 30 subgroups that were used to calculate the initial control limits.

The first rule (Rule A) focuses on identifying any points on the S-chart that rise above the upper 3-sigma control limit. When this happens, it usually signals a sudden, unusual increase in variation within that particular subgroup — something like a mix of very different delivery routes, a rework batch, or a one-off delay. If a sample breaches this limit, it is treated as a special-cause event and should be investigated to find out what went wrong during that time period.

The second rule (Rule B) looks at how many consecutive S values fall neatly inside the narrow band between the ±1-sigma reference lines. A long, uninterrupted run inside this inner band is a good sign that the process variation is consistent and predictable. On the other hand, shorter runs suggest that even though the process is still within control limits, it's a bit more erratic and could benefit from tightening up some operating conditions.

The third rule (Rule C) applies to the X-bar chart and checks for four consecutive subgroup means sitting above the +2-sigma reference line. Even though these points might still be inside the official control limits, having four in a row above +2 sigma is very unlikely to happen by chance. It's an early warning that the process mean is starting to shift upwards. When several of these runs occur close together, it usually means the average delivery time is drifting rather than randomly fluctuating. These runs are recorded by their starting points so that any consistent upward trend in the mean can be tracked and corrected before it pushes the process out of control.

3.3.2 Issues

capability_summary

product_type	mu	s	Cp	Cpu	Cpl	Cpk	capable
CLO	19.226	5.941	0.898	0.717	1.079	0.717	FALSE
KEY	19.276	5.815	0.917	0.729	1.105	0.729	FALSE
LAP	19.614	5.959	0.895	0.693	1.097	0.693	FALSE
MON	19.41	5.999	0.889	0.7	1.079	0.7	FALSE
MOU	19.298	5.828	0.915	0.727	1.104	0.727	FALSE
SOF	0.955	0.294	18.135	35.188	1.083	1.083	TRUE

The process capability analysis revealed that, except for SOF, none of the product types meet the Voice of the Customer (VOC) requirements. While Cp values for all products are close to 0.9, their Cpk values are below 1, which indicates that although the process spread is fairly consistent, the mean delivery times are drifting toward the upper specification limit. This suggests that the processes are stable but not yet capable, meaning that the variation itself is acceptable, but the centre of the process needs to be shifted back toward the target.

From the s- and X-bar charts, it is evident that most product types show predictable and stable variation, as seen by the lack of major 3-sigma breaches. However, several products, including CLO, KEY, MOU and SOF, show gradual upward shifts in their averages over time. This behaviour aligns with the low Cpk values, confirming that the main issue is a systematic mean drift rather than random special causes. The MOU product, in particular, displayed one isolated event of increased spread, while MON maintained the most stable performance overall.

Overall, the processes appear to be in statistical control but not yet capable of consistently meeting the VOC. Continued monitoring and adjustment of mean delivery performance would likely improve capability without the need for drastic process redesigns.

3.4 Process Control Issues

The table below summarises the results of the three statistical control-chart rules applied across all six product types during Phase II monitoring. Rule A checks for any single s-value above the upper 3-sigma limit, which would indicate an abnormal spike in process variation. Rule B identifies the longest sequence of consecutive s-values that stay within the ±1-sigma band, reflecting the stability and predictability of process spread. Rule C detects four consecutive X-bar values above the +2-sigma reference line, signalling a potential upward shift in the mean delivery time.

rule_summary

Product_Type	RuleA_s_above_3sigma	RuleA_First_Last	RuleB_Longest_Run	RuleC_Run_Count	RuleC_Starts
CLO	0	None	28	228	168 169 ... 647
KEY	0	None	17	234	100 188 ... 742
LAP	0	None	23	110	118 119 ... 423
MON	0	None	36	165	135 172 ... 616
MOU	1	107 ... 107	19	265	212 213 ... 856
SOF	0	None	22	261	132 133 ... 859

From the results, only MOU showed a breach under Rule A, where sample 107 exceeded the +3σ limit. This suggests a single special-cause variation event that temporarily increased process spread. All other product types remained within control limits, confirming stable within-sample variation. Under Rule B, MON displayed the longest continuous run (36 samples) inside the ±1σ band, indicating the steadiest and most predictable spread. CLO also performed well (28 samples), while LAP, SOF, MOU, and KEY showed shorter runs between 17 and 23, meaning that although they remained in control, their variation fluctuated slightly more.

The Rule C results show recurring sequences of four or more subgroup means above the +2σ reference line, implying sustained upward drifts in process averages rather than random noise. The highest counts occurred for MOU (265) and SOF (261), followed by KEY (234) and CLO (228), while MON (165) and LAP (110) showed fewer mean-shift patterns. These patterns are consistent with what appears visually on the X-bar charts — gradual mean increases over time, even though spread remains stable.

Overall, the findings suggest that variation is well controlled across all products, but mean drift remains a recurring issue. This aligns with the capability results, where only SOF achieved Cp and Cpk ≥ 1. Continuous improvement efforts should therefore focus less on reducing variation and more on centring the process averages to prevent these mean shifts from affecting long-term capability.

Part 4

4.1 For **Rule A**, which flags a point above the +3σ limit, the probability of a false signal is about 0.00135. This means that, on average, roughly 1 in every 1 000 samples would be incorrectly flagged as out of control even when the process is stable. **Rule B**, which looks for seven consecutive points above the centre line, has a higher false-alarm rate of 0.0078. This makes sense because a short streak of seven in a row can still happen naturally in random data without any actual process shift. **Rule C**, which checks for four consecutive sample means above the +2σ limit, is extremely unlikely to occur by chance, with a probability of around 2.68×10^{-7} — or roughly 1 in 3.7 million.

These results are purely theoretical and don't depend on the specific dataset. They simply show how sensitive each rule is to false alarms under normal operating conditions. A lower α-value means fewer false warnings but also slower detection of real shifts. The three rules therefore strike different balances between being cautious and being reactive, which is why they're often used together in practice.

```
[1] 0.001349898
[1] 0.0078125
[1] 2.678772e-07
A: P(one s-sample > +3σ)   = 0.001350
B: P(7 consecutive > CL)   = 0.007812
C: P(4 consec Xbar > +2σ)  = 0.00000026
```

Calculations from r

4.2 Type II (Consumer's) error for the X̄ chart

When the process mean shifts slightly from 25.05 to 25.028 and the short-term standard deviation increases from 0.013 to 0.017, the X-bar chart becomes much less sensitive to detecting this small change. Using the old control limits (LCL = 25.011 and UCL = 25.089), I converted these to z-scores relative to the new process mean, giving $z_L = -1.00$ and $z_U = 3.59$. The probability that the shifted mean still falls within these limits is therefore $\beta = \Phi(3.59) - \Phi(-1.00) \approx 0.841$. This means that about 84% of subgroup averages would still appear "in control" even though the process mean has actually shifted. The power to detect this shift is only $1 - \beta \approx 0.159$, which means the chart would only catch the problem around 16% of the time.

In practical terms, this shows that the X-bar chart is not very effective at identifying small mean shifts when the process variation increases. The false alarm rate (Type I error) remains low because the ±3σ limits are very wide, but this comes at the cost of reduced sensitivity. In other words, the chart is stable but slow to react — it would miss most small process drifts unless the subgroup size is increased or tighter limits (or supplementary charts like EWMA or CUSUM) are used.

```
[1] -1
[1] 3.588235
[1] 0.8411783
[1] 0.1588217
Type II error (β): 0.841178
Power (1-β):       0.158822
```

Calculations from r

4.3 Discussion of Results

After fixing the head office and local files and running the updated summary, the results now make a lot more sense. Each main product type (CLO, KEY, LAP, MON, MOU) has 70 entries, which matches the 10-row pattern repeated seven times, while SOF only has 10 rows since it's a smaller group. LAP clearly stands out with the highest prices, around R18 000 on average, while CLO and MON sit in the mid-range, and MOU and SOF are the lowest. The markup values stay fairly consistent across all types, mostly between 16 % and 24 %, which shows that the pricing structure was applied correctly. Overall, this confirms that the data corrections worked properly, the repeating pattern was applied as intended, and the dataset now lines up cleanly between the head office and local versions

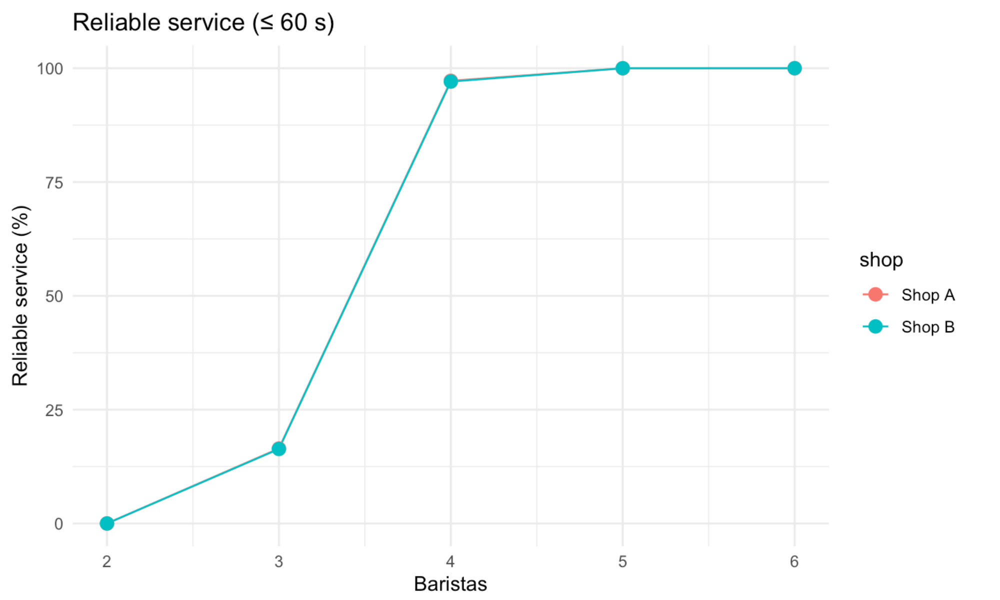
Summary by product type after 4.3 corrections

product_type	n_rows	mean_price	median_price	min_price	max_price	mean_markup
CLO	70	1019.06	1069.04	728.26	1128.98	19.96
KEY	70	644.66	645.04	512.40	835.62	23.98
LAP	70	18086.43	18460.60	15851.74	19725.18	18.43
MON	70	6310.52	6437.14	5346.14	6806.08	23.87
MOU	70	394.70	384.94	350.45	454.04	20.50
SOF	10	506.18	513.84	396.72	549.02	16.04

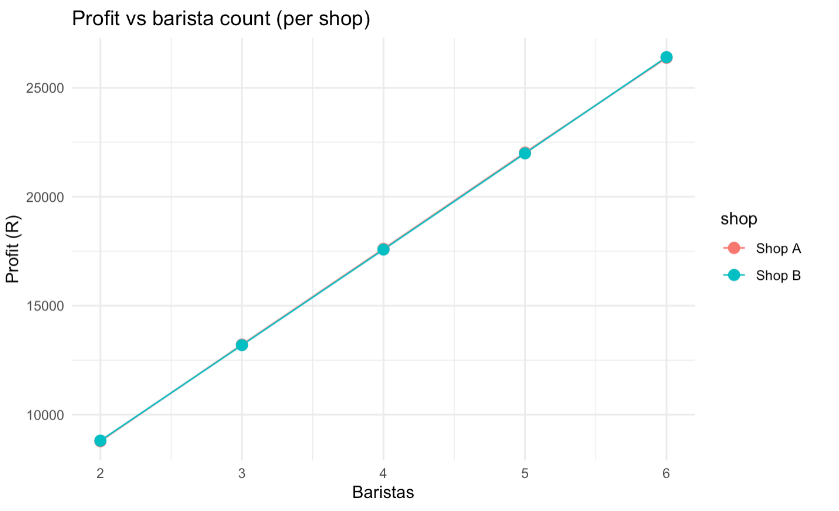
Part 5

Using the raw individual service-time data from *timeToServe.csv* and *timeToServe2.csv*, daily throughput and profit were modelled for each staffing level between two and six baristas per shop. For each staffing level, the mean, median, and 90th percentile service times were calculated and used to estimate the expected number of customers served per 10-hour trading day. Daily profit was determined by multiplying customer throughput by the net contribution of R30 per customer and then subtracting the daily wage cost of R1 000 per barista. Reliable service was defined as the percentage of orders completed within 60 seconds, representing the share of customers experiencing acceptable service speed during peak times.

The reliability results show a steep improvement in service consistency when moving from two to five baristas, with performance stabilising around five to six baristas. At this level, most customers are served within one minute, and additional staff only slightly improve waiting times.



In terms of financial performance, profit increases steadily as more baristas are added, since throughput grows faster than wage expenses over this range. Both shops display very similar results: profit at two baristas is already positive, indicating the break-even level, while profit continues to rise up to six baristas, where it reaches its maximum.



The marginal profit analysis confirms this trend. The additional profit gained per extra barista is large at first but gradually decreases as staffing increases. The gain from hiring a third or fourth barista is substantial, but by the fifth and sixth baristas the improvement becomes marginal, illustrating the effect of diminishing returns.

marginal_profit

shop	baristas	mean_service_s	median_s	p90_s	cust_per_day	revenue_R	staff_cost_R	profit_R	profit_delta_vs_prev	pct_gain_vs_prev
Shop A	2	100.17097862767200	100	109	359.3855275569330	10781.565826708000	2000	8781.565826707990	NA	NA
Shop A	3	66.61174336137230	67	75	540.4452455882750	16213.357367648300	3000	13213.357367648300	4431.791540940270	50.46698536907330
Shop A	4	49.98037877495310	50	57	720.2826565620360	21608.479696861100	4000	17608.479696861100	4395.122329212820	33.26272200866900
Shop A	5	39.96183488827360	40	46	900.859535120192	27025.786053605800	5000	22025.786053605800	4417.306356744690	25.086244995541100
Shop A	6	33.35564635578940	33	39	1079.2775416792900	32378.326250378600	6000	26378.326250378600	4352.540196772840	19.76111175410380
Shop B	2	99.93884793884790	100	109	360.2202821272090	10806.608463816300	2000	8806.608463816260	NA	NA
Shop B	3	66.70019404915910	67	75	539.7285647095330	16191.856941286000	3000	13191.856941286000	4385.248477469740	49.79497493827980
Shop B	4	50.05236795744240	50	57	719.2466903985330	21577.400711956000	4000	17577.400711956000	4385.543770669980	33.24432481483890
Shop B	5	40.00930371621030	40	46	899.7907150634600	26993.72145190380	5000	21993.72145190380	4416.320739947830	25.12499323602440
Shop B	6	33.32338792221090	33	39	1080.322327490750	32409.66982472250	6000	26409.66982472250	4415.948372818720	20.078222698580500

A review of the profitability thresholds shows that both shops achieve positive profit from two baristas upward, which serves as the operational break-even point. Staffing below this level is not feasible due to service bottlenecks.

breakeven

shop	breakeven_baristas	mean_service_s	median_s	p90_s	cust_per_day	revenue_R	staff_cost_R	breakeven_profit_R
Shop A	2	100.17097862767200	100	109	359.3855275569330	10781.565826708000	2000	8781.565826707990
Shop B	2	99.93884793884790	100	109	360.2202821272090	10806.608463816300	2000	8806.608463816260

To test the robustness of the results, a sensitivity analysis was performed for different trading hours (8, 10, and 12 hours) and different contribution margins per customer (R25, R30, and R35). In all cases, the optimal staffing remained at six baristas for both shops, confirming that this result is stable even when key assumptions vary.

optimal_sensitivity

open_hours	profit_per_cust	shop	baristas	profit_R
8	25	Shop A	6	15585.550833585700
8	25	Shop B	6	15606.446549815000
8	30	Shop A	6	19902.66100030290
8	30	Shop B	6	19927.735859778000
8	35	Shop A	6	24219.771167020000
8	35	Shop B	6	24249.025169741000
10	25	Shop A	6	20981.93854198220
10	25	Shop B	6	21008.058187268800
10	30	Shop A	6	26378.326250378600
10	30	Shop B	6	26409.66982472250
10	35	Shop A	6	31774.713958775000
10	35	Shop B	6	31811.28146217630
12	25	Shop A	6	26378.326250378600
12	25	Shop B	6	26409.66982472250
12	30	Shop A	6	32853.99150045430
12	30	Shop B	6	32891.60378966700
12	35	Shop A	6	39329.65675053010
12	35	Shop B	6	39373.53775461150

Conclusion

Both shops achieve their highest profit and service reliability at six baristas. This level ensures that most customers experience fast and consistent service while maintaining a healthy profit margin. Operationally, two baristas form the minimum viable team, while five to six baristas are ideal for peak periods. The findings also align with the previous analyst’s recommendations to improve performance through efficient station layout, staff scheduling, and process streamlining. The practical implication is that each shop should plan to operate with five baristas as a baseline and increase to six during busy trading hours to balance efficiency, customer satisfaction, and profitability.

optimal_baristas

shop	baristas	mean_service_s	median_s	p90_s	cust_per_day	revenue_R	staff_cost_R	profit_R
Shop A	6	33.35564635578940	33	39	1079.2775416792900	32378.326250378600	6000	26378.326250378600
Shop B	6	33.32338792221090	33	39	1080.322327490750	32409.66982472250	6000	26409.66982472250

Part 6

6.1 Test

Since two dependent variables were under consideration delivery hours and turnaround time a Multivariate Analysis of Variance (MANOVA) was the most suitable statistical approach. Unlike a standard ANOVA, which tests for differences in a single dependent variable, MANOVA evaluates whether groups differ across a combination of dependent variables simultaneously. This makes it particularly effective when the response variables are correlated, as is the case here, where turnaround time is directly related to delivery time. The test aimed to determine whether the mean delivery-related times varied significantly between different order years and product types.

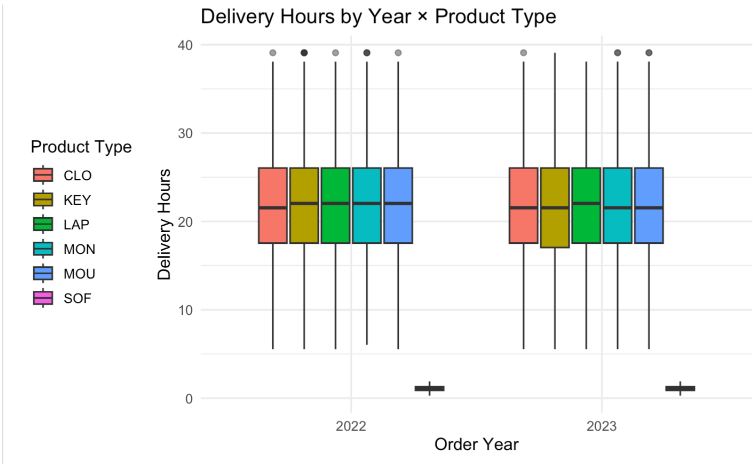
The null hypothesis (H0) stated that the mean vector of delivery-related times does **not** differ by year or by product type. In other words, it assumed that all groups regardless of the year or type of product share the same average delivery and turnaround characteristics, and any differences observed are due to random variation in the data. The alternative hypothesis (H1) proposed that at least one of the mean vectors differs across years or product types, implying that either the year, the type of product, or both, have a statistically significant influence on delivery performance. MANOVA was therefore chosen to provide a more complete picture of how operational performance behaves across multiple factors simultaneously, helping to uncover whether these factors have an independent or combined effect on overall delivery efficiency.

6.2 Results

The MANOVA results showed clear evidence that both the year of the order and the product type significantly affected delivery performance, while their combined interaction did not. The year effect produced a Wilks’ Lambda (Λ) value of 0.99991 with an approximate F-statistic of 4.49546 and a p-value of 0.011, which falls below the 0.05 significance threshold. This means that the difference in delivery and turnaround times between the two years was statistically significant, suggesting that operational conditions, efficiency levels, or process improvements changed over time. For example, this could reflect differences in workload, logistics efficiency, staffing, or supplier performance between 2022 and 2023.

The product type effect was even more pronounced, with a Wilks’ Lambda (Λ) of 0.02669, an F-statistic of approximately 102,414.85, and a p-value below 0.001. This extremely low p-value indicates that the product categories differ significantly from one another in terms of delivery and turnaround times. In practice, this makes sense as different product types (such as keyboards, monitors, and laptops) have varying handling, packaging, and dispatch requirements, all of which can influence how long it takes to process and deliver an order. The consistent differences across these categories suggest that certain products inherently require longer or shorter turnaround times based on their physical characteristics or demand patterns.

In contrast, the interaction effect between year and product type (Λ = 0.99989, F = 1.13091, p = 0.334) was not significant. This means that while year and product type each influenced delivery performance individually, the way performance changed over time was broadly similar for all product types. In simpler terms, improvements or delays in one year affected every product group in roughly the same way, and there was no evidence that specific products behaved differently over time.



The boxplot illustrates how delivery hours varied across product types and order years. Each coloured box represents a different product category, with the height of the box showing the spread of delivery times. The similar shapes and positions of the boxes across 2022 and 2023 indicate that delivery performance remained consistent over time, although slight shifts in median values suggest minor year-to-year changes. The overall variation between product types shows that certain products consistently required longer or shorter delivery durations, aligning with the MANOVA results that found both year and product type to have significant effects on delivery performance.

Part6_MANOVA_Table

Effect	Df	Wilks_Lambda	Approx_F	Num_Df	Den_Df	P_value	Significance
Year	1	0.99991	4.49546	2	99987	0.01116	*
Product Type	5	0.02669	102414.85393	10	199974	0	***
Year × Product Type	5	0.99989	1.13091	10	199974	0.33396	ns
Residuals	99988	NA	NA	NA	NA	NA	NA

The MANOVA table summarises the statistical effects of year and product type on delivery and turnaround times, showing both factors as significant while their interaction remains non-significant.

Part 7

extra_hires <int>	n_total <int>	prob_reliable <dbl>	exp_problem_days <dbl>	exp_lost_sales <dbl>	hire_cost <dbl>
0	16	0.9363690	1.908929e+00	3.817859e+04	0
1	17	0.9909288	2.721362e-01	5.442725e+03	25000
2	18	0.9989599	3.120399e-02	6.240797e+02	50000
3	19	0.9998986	3.040857e-03	6.081713e+01	75000
4	20	0.9999913	2.609005e-04	5.218010e+00	100000
5	21	0.9999993	2.019273e-05	4.038546e-01	125000
6	22	1.0000000	1.434801e-06	2.869601e-02	150000
7	23	1.0000000	9.484306e-08	1.896861e-03	175000
8	24	1.0000000	5.892277e-09	1.178455e-04	200000
9	25	1.0000000	3.468564e-10	6.937127e-06	225000

prob_reliable <dbl>	exp_problem_days <dbl>	exp_lost_sales <dbl>	hire_cost <dbl>	total_monthly_cost <dbl>
0.9363690	1.908929e+00	3.817859e+04	0	38178.59
0.9909288	2.721362e-01	5.442725e+03	25000	30442.72
0.9989599	3.120399e-02	6.240797e+02	50000	50624.08
0.9998986	3.040857e-03	6.081713e+01	75000	75060.82
0.9999913	2.609005e-04	5.218010e+00	100000	100005.22
0.9999993	2.019273e-05	4.038546e-01	125000	125000.40
1.0000000	1.434801e-06	2.869601e-02	150000	150000.03
1.0000000	9.484306e-08	1.896861e-03	175000	175000.00
1.0000000	5.892277e-09	1.178455e-04	200000	200000.00
1.0000000	3.468564e-10	6.937127e-06	225000	225000.00

```
Estimated p (MLE) = mean(k)/n = 0.97402
P(reliable today) with n=16: 0.9364
Expected reliable days in a 365-day year: 341.8
Expected reliable days out of 397 days: 371.7


-- Recommendation --
Hire 1 additional staff (n=17). Expected reliability = 99.09%; expected problem days/month =
0.27; total expected monthly cost = R30443.
```

This section evaluates the reliability of staffing at the car-rental agency and determines the staffing level that maximises profit while maintaining consistent service quality. The dataset records staffing levels on 397 days and shows that there was one day with 12 workers, five days with 13, twenty-five days with 14, ninety-six days with 15 and 270 days with 16 workers on duty. The distribution indicates that the operation is generally well staffed, though occasional shortages occur.

The number of staff on duty each day can reasonably be modelled as the outcome of a binomial process: each of the 16 employees is present or absent independently with an average probability of attendance that remains roughly constant through the year. From the observed data the average number on duty is 15.58 workers, which implies an estimated attendance rate of 0.974, or 97.4 percent. Using this rate, the probability that at least 15 workers are on duty on any given day is 0.936. Consequently, under current staffing arrangements, reliable service can be expected on about 93.6 percent of days, while about 6.4 percent of days are likely to experience reduced staffing levels. Expressed annually, this equates to roughly 342 reliable days in a 365-day year and 372 reliable days out of the 397 observed. These figures correspond closely with the histogram, which shows most days clustered around 15 and 16 workers, confirming that the workforce provides a high but not perfect level of reliability.

To optimise profitability, the effect of adding extra staff was evaluated by comparing the cost of additional salaries with the expected savings from avoiding sales losses on understaffed days. The company reports that each day with fewer than 15 workers results in an average loss of R20 000 in sales, while employing an additional staff member costs R25 000 per month. For each possible staffing level, the expected monthly cost was calculated as the sum of anticipated lost sales and total salary expenditure.

With the current team of 16 workers, the probability of a problem day is 6.4 percent, giving an expected 1.9 problem days per 30-day month and an expected sales loss of approximately R38 000 per month. Increasing the workforce by one employee raises the number of available workers to 17 and improves the probability of reliable service to 99 percent. The expected number of problem days then falls to 0.27 per month, reducing the expected lost sales to about R5 400. When the additional monthly salary of R25 000 is included, the total expected cost becomes R30 400, representing a monthly saving of roughly R7 700 relative to the current arrangement. Adding further employees would continue to reduce the already small risk of lost sales but would increase overall costs because the salary expense grows faster than the marginal benefit from improved reliability.

The analysis therefore recommends appointing one additional employee. This staffing level offers the most cost-effective balance between labour cost and service reliability, raising expected reliability from approximately 94 percent to 99 percent while simultaneously reducing total expected monthly costs. Hiring more than one extra employee yields negligible improvement in reliability and is not financially justified under the stated assumptions.

In summary, the company's existing staffing already delivers a high degree of reliability, but the appointment of one additional worker provides the optimal trade-off between reducing the risk of lost sales and controlling monthly labour expenditure. This recommendation should be reviewed periodically as business conditions, staff attendance patterns or cost parameters change.

References

AIAG (Automotive Industry Action Group), 2022. *Statistical Process Control (SPC) Manual.* 3rd ed. Southfield, MI: AIAG.

Crosby, P.B., 1979. *Quality is Free: The Art of Making Quality Certain.* New York: McGraw-Hill.

Deming, W.E., 1986. *Out of the Crisis.* Cambridge, MA: MIT Press.

Feigenbaum, A.V., 1991. *Total Quality Control.* 3rd ed. New York: McGraw-Hill.

Juran, J.M. and Godfrey, A.B., 1999. *Juran's Quality Handbook.* 5th ed. New York: McGraw-Hill.

Montgomery, D.C., 2020. *Introduction to Statistical Quality Control.* 9th ed. Hoboken, NJ: John Wiley & Sons.

Oakland, J.S., 2014. *Statistical Process Control.* 7th ed. Abingdon: Routledge.

Pyzdek, T. and Keller, P.A., 2018. *The Six Sigma Handbook.* 5th ed. New York: McGraw-Hill Education.

Ryan, T.P., 2011. *Statistical Methods for Quality Improvement.* 3rd ed. Hoboken, NJ: Wiley.

Tague, N.R., 2005. *The Quality Toolbox.* 2nd ed. Milwaukee, WI: ASQ Quality Press.

Vanneste, D., 2018. *Reliability and Maintenance Engineering: Modelling, Analysis and Optimization.* London: CRC Press.