Basic Data Analysis ECSA Report

Quality Assurance 344

Student Number: 26876493

Third-Year Industrial Engineering

October 23, 2025

Contents

1 Data Inspection

This section describes the structure of each dataset: number of rows and columns, variable names, data types, missing values, and a few sample records.

1.1 Customer Data

Table 1: Customer Data Inspection

Property	Details
Rows × Columns	5000 × 5
Columns	`CustomerID`, `Gender`, `Age`, `Income`, `City`
Data Types	ID (string), Gender (categorical), Age (int), Income (int), City (categorical)
Missing Values	None
Sample Records	CUST001, Male, 16, 65000, New York; CUST002, Female, 31, 20000, Houston

1.2 Products Data

Table 2: Products Data Inspection

Property	Details
Rows × Columns	60 × 5
Columns	`ProductID`, `Category`, `Description`, `SellingPrice`, `Markup`
Data Types	ID (string), Category (categorical), Description (text), SellingPrice (float), Markup (float)
Missing Values	None
Sample Records	SOF001, Software, 511.53, 25.05; SOF002, Cloud Subscription, 505.26, 10.43

1.3 Products Head Office Data

Table 3: Products Head Office Data Inspection

Property	Details
Rows × Columns	360 × 5
Columns	Same as Products Data
Data Types	Identical to Products Data
Missing Values	None
Sample Records	SOF001, Software, 521.72, 15.65; SOF002, Software, 466.95, 28.42

1.4 Sales Data (2022–2023)

Table 4: Sales Data Inspection

Property	**Details**
Rows × Columns	100000 × 9
Columns	`CustomerID`, `ProductID`, `Quantity`, `orderTime`, `orderDay`, `orderMonth`, `orderYear`, `pickingHours`, `deliveryHours`
Data Types	IDs (string), Quantity (int), Time/Date (int), Picking (float), Delivery (float)
Missing Values	None
Sample Records	CUST1791, CLO011, 16, Nov 2022, Picking 17.7h, Delivery 24.5h

2 Summary Statistics

This section presents descriptive statistics for numeric variables (mean, standard deviation, quartiles). Categorical variables are discussed in the text below each table.

2.1 Customer Data Summary

Table 5: Custom Statistics for Customer Data

Mean Age	**SD Age**	**Mean Income**	**SD Income**	**Total Income**
51.6	21.2	80,797	33,150	40,398,5000

Comment: - The average customer age is around 52 years, with a wide spread (SD = 21). This means the customer base is very diverse in age. - Average income is ˜81k, but with a high standard deviation, suggesting there are both low- and high-income groups. - The large total income pool indicates significant purchasing potential. This could guide segmentation strategies (e.g., young/low income vs. older/high income).

2.2 Products Data Summary

Table 6: Custom Statistics for Products Data

Mean Price	**SD Price**	**Min Price**	**Max Price**	**Mean Markup**	**SD Markup**
4,494	6,504	350	19,725	20.5	6.1

Comment: - There is a huge range in selling prices (R350 accessories vs. R19,725 premium products). - Mean markup is consistent at around 20%, suggesting a uniform profit strategy across categories. - Interesting point: product pricing seems skewed — most products are in the low-to-mid range, but a few outliers drive up the mean and SD.

Table 7: Custom Statistics for Products Head Office Data

Mean Price	SD Price	Median Price	Mean Markup	SD Markup
4,411	6,464	797	20.4	5.7

2.3 Products Head Office Data Summary

Notes: - Median price (R797) is much lower than mean price (R4,411), showing the dataset is heavily skewed by a few very expensive products. - Markup is again stable around 20%, consistent with the smaller products dataset. - The catalogue duplication (only 110 unique IDs) could cause confusion or overestimation of inventory.

2.4 Sales Data Summary by Year

Table 8: Custom Statistics for Sales Data (2022 vs 2023)

Year	Mean Sales	SD Sales	Total Sales	Mean Picking (h)	SD Picking (h)	Mean Delivery (h)	SD Delivery (h)
2022	13.4	13.7	722,141	14.7	10.4	17.5	10.0
2023	13.6	13.8	628,206	14.7	10.4	17.4	9.99

Notes: - Mean sales per order stayed stable (13–14 units), but total sales declined from 722k in 2022 to 628k in 2023 (13% drop). - Picking and delivery times remained unchanged, suggesting stable operations but no productivity gains. - The decline in total sales with stable order sizes hints at fewer active customers. - Further analysis should investigate `CustomerID` frequency by year to confirm customer churn.

Missing Values

Across all four datasets (`customer_data`, `products_data`, `products_Headoffice`, and `sales2022and2023`), no missing values were detected. This means all variables are fully populated. Data quality concerns are therefore not related to missingness but rather to other issues such as skewed distributions (e.g. product prices) and duplication in the head office catalogue.

3 Customer Data: High-Value Segments

Beyond basic descriptive statistics, filtering and subsetting was applied to identify high-value customers. Customers were grouped into **age bands** (smaller than 25, 26–35, 36–45, 46–55, 56–65, and bigger than 65 years). Within each band, the top quartile of income

(Q3 and above) was calculated, and customers at or above this threshold were classified as high-value relative to their demographic peers.

Results by Age Band

Table 9 shows the number of high-value customers per age band, together with their mean and maximum income.

Table 9: High-Value Customers (Top 25% Income) by Age Band

Age Band	High-Value Customers	Mean Income (R)	Max Income (R)
25	183	96,066	105,000
26–35	224	97,567	105,000
36–45	191	132,749	140,000
46–55	184	130,245	140,000
56–65	181	133,122	140,000
65	380	112,618	120,000
NA	4	110,000	120,000

Interpretation: - The *36–45* and *56–65* bands show the highest mean incomes (R132k and R133k), making them attractive groups for premium targeting. - The *¿65* group has the largest count (380 customers), but lower mean income (R112k), suggesting volume but less concentrated wealth. - Younger customers (35) are present in reasonable numbers, but their income distribution caps out at R105k. - The few records with missing age still show high incomes (R110k–R120k) and should be investigated further.

Top 10 Individual Customers

The top ten customers by absolute income are all at the maximum level of R140,000, spread mostly across the *36–65* age range. These individuals represent potential **VIPs** who could significantly influence revenue.

Table 10: Top 10 Customers by Income

CustomerID	Age	Income (R)
CUST1511	41	140,000
CUST1514	48	140,000
CUST1515	49	140,000
CUST1517	62	140,000
CUST1532	44	140,000
CUST1559	64	140,000
CUST1569	41	140,000
CUST1575	53	140,000
CUST1578	44	140,000
CUST1588	65	140,000

3.1 Products Data Analysis

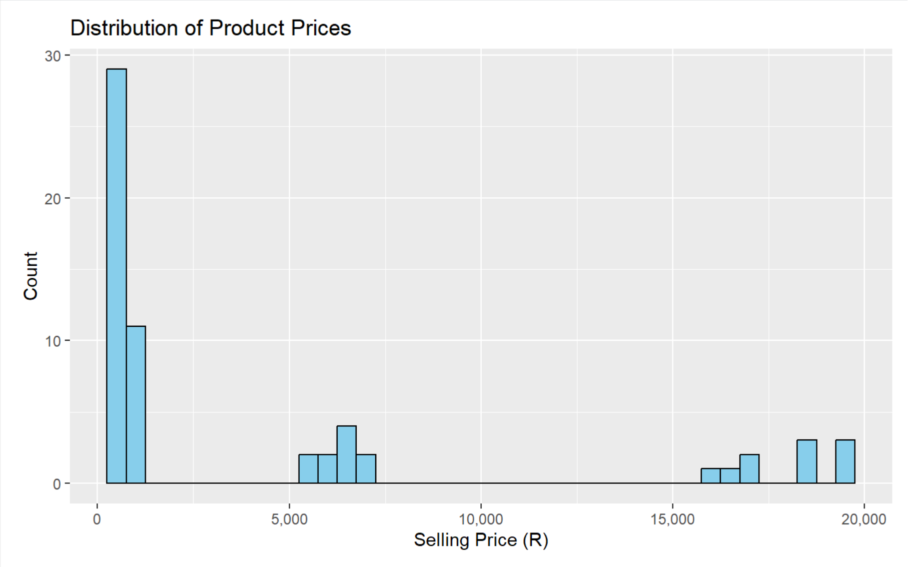

Figure 1: count of selling price classes

Analysis: The histogram reveals a highly **right-skewed distribution**. Most products are priced at the lower end (under R2,000), with a sharp drop in frequency as prices rise. Smaller clusters appear around R6,000–R7,000 and again at the premium range (R16,000–R20,000). This pattern indicates that while the catalogue is dominated by low-cost items, a minority of very expensive products exist which may serve a premium customer segment.

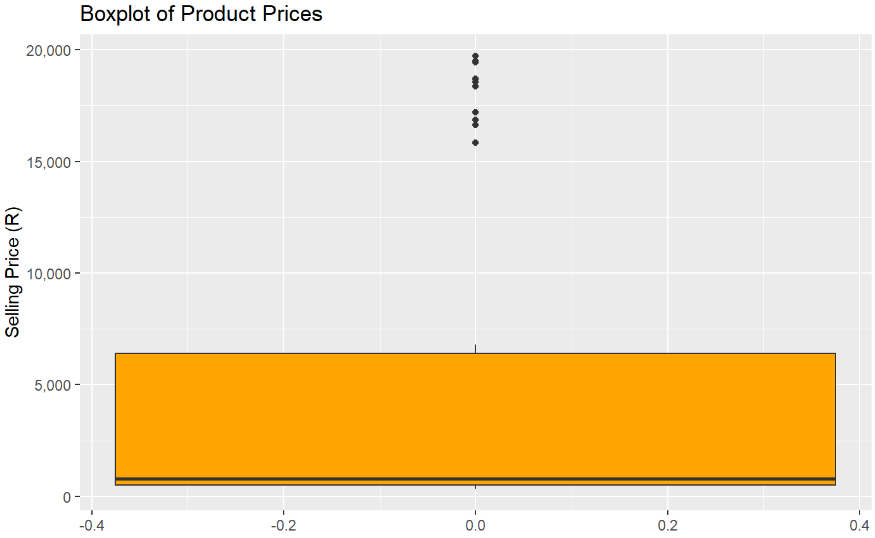

Figure 2: boxplot of product prices

Analysis: high-value outliers above R15,000. The median lies much closer to the lower quartile, reinforcing that most prices are concentrated in the low-to-mid range. The large interquartile range (IQR) suggests significant variability in product pricing.

- The product portfolio is split into two clear tiers: low-cost, high-volume items versus a few premium-priced items. - These outliers drive up the mean selling price, which could mislead management if not considered carefully. - Further analysis should investigate whether these premium products generate meaningful sales volume or if they simply inflate the overall price statistics.

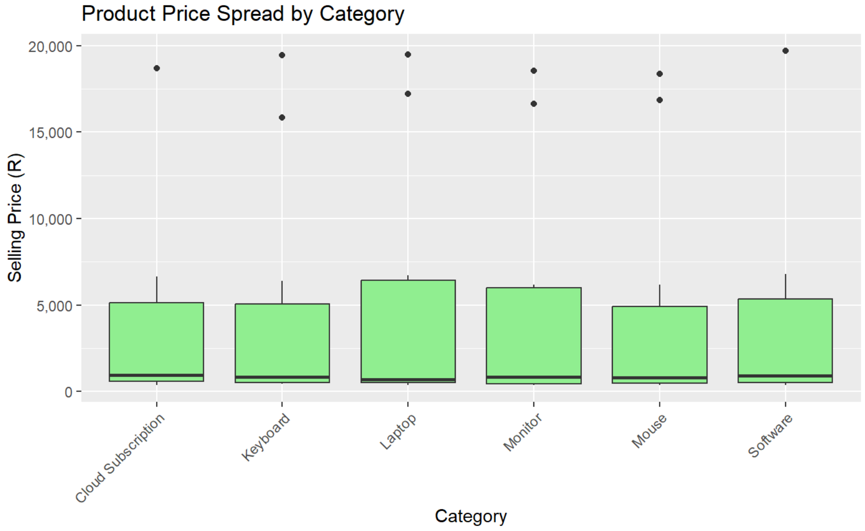

Figure 3: boxplot of selling price

Analysis: All categories show similar median prices in the low-to-mid range, but also contain **extreme outliers** above R15,000. Laptops and monitors in particular show the widest spread, indicating tiered product offerings within the same category. This has implications for marketing: customers shopping in the same category may face large price differences, which could cause confusion or be leveraged for premium upselling.

Table 11: Premium-Priced Outliers by Category

ProductID	Category	Description	SellingPrice
LAP025	Software	azure sandpaper	19,725
LAP021	Laptop	black marble	19,495
LAP023	Keyboard	azure matt	19,453
LAP026	Cloud Subscription	aliceblue silk	18,712

Premium Outliers The premium-priced products are spread across all categories, not confined to a single product line. Their descriptions suggest cosmetic or branding variations ("black marble", "azure matt", "silk") that may not align with core technical differences. This raises the risk that catalogue complexity is driven more by branding than by functional value.

Analyst Insights: - Pricing shows strong right skew: most products are affordable, but outliers push up the average. - Markup policies are controlled (10–30%), but high absolute prices could distort customer perception. - Premium outliers exist across all categories, hinting at a mixed strategy of low-end accessibility and high-end exclusivity. - Next steps: analyse *sales frequency of premium outliers* to confirm if they contribute materially to revenue or just inflate catalogue averages.

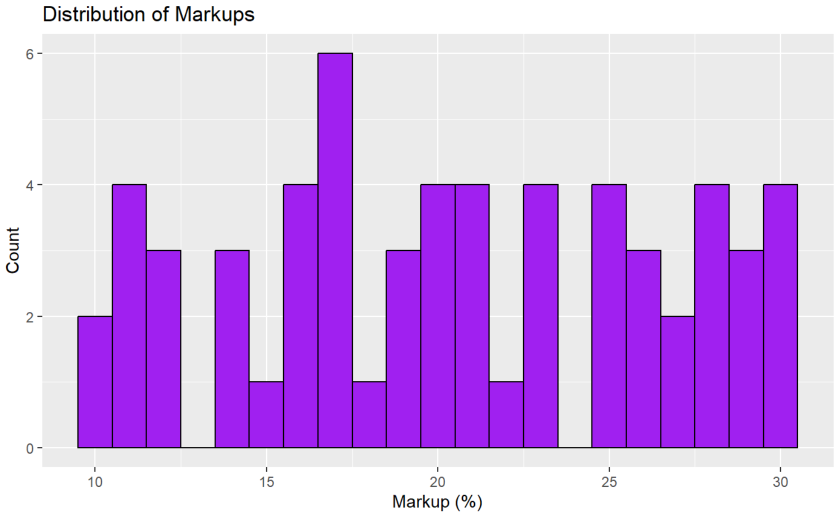

Figure 4: Distribution of Markups

Analysis: Markups are well spread between 10% and 30%, with no single peak dominating the distribution. This indicates that pricing policies allow moderate variation across categories, but overall remain within a consistent band. The lack of extreme outliers suggests good governance of cost-plus pricing.

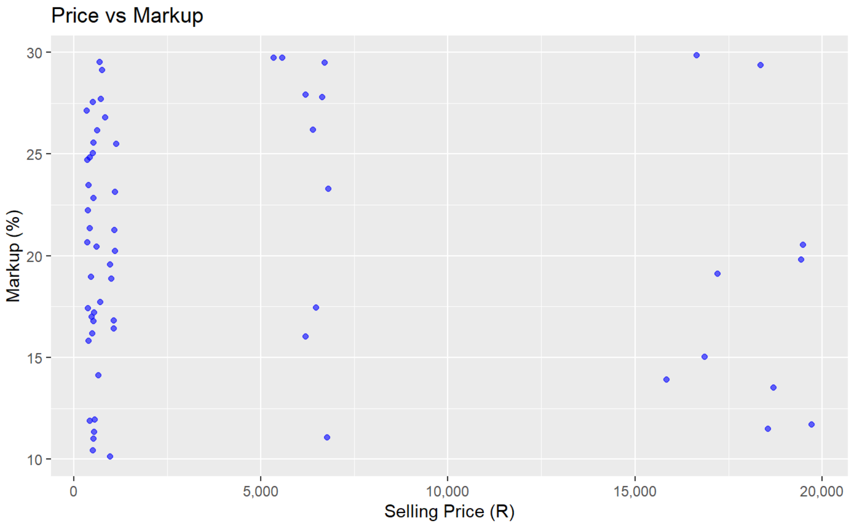

Figure 5:

Analysis: The scatterplot shows no strong correlation between price and markup. High-priced items still keep markups close to 20%, which indicates that premium pricing is value-driven rather than margin-driven.

4 Part 3: Statistical Process Control (SPC)

We apply SPC to `sales2026and2027Future.csv` per product type. Data are ordered chronologically and split into sequential subgroups of size $n = 24$. Using the first 30 subgroups, we initialise X-*bar* (centre) and S (spread) charts and derive the centre line and $\pm 1\sigma$, $\pm 2\sigma$, $\pm 3\sigma$ bands. Subsequent samples (31, 32, . . .) are monitored against these limits—interpreting the S-chart first—flagging: (A) any S above $+3\sigma$, (B) the longest run of S within $\pm 1\sigma$, and (C) any four consecutive X-*bar* points above $+2\sigma$. Capability on the first 1,000 deliveries is assessed via $C_p, C_{pu}, C_{pl}, C_{pk}$ with LSL $= 0\,\text{h}$ and USL $= 32\,\text{h}$ to judge conformance to the VOC.

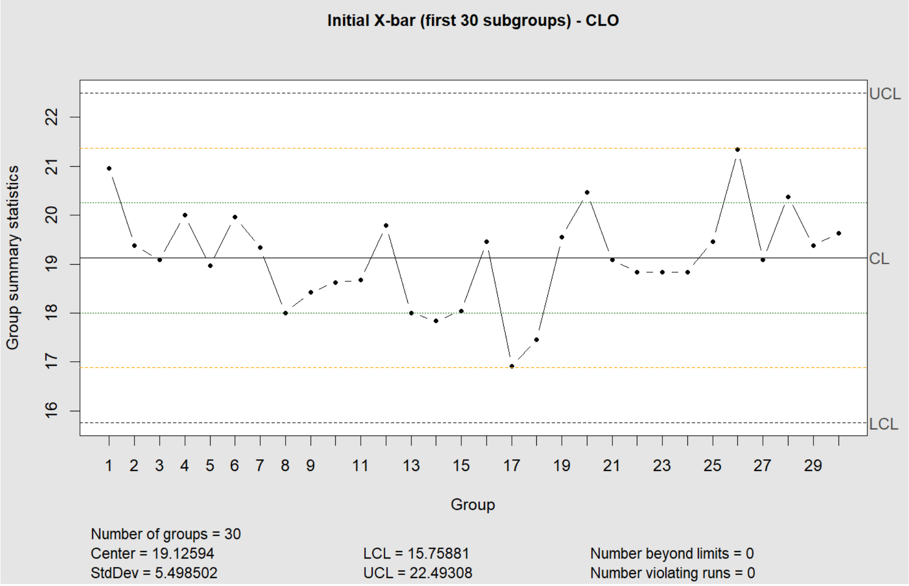

Figure 6:

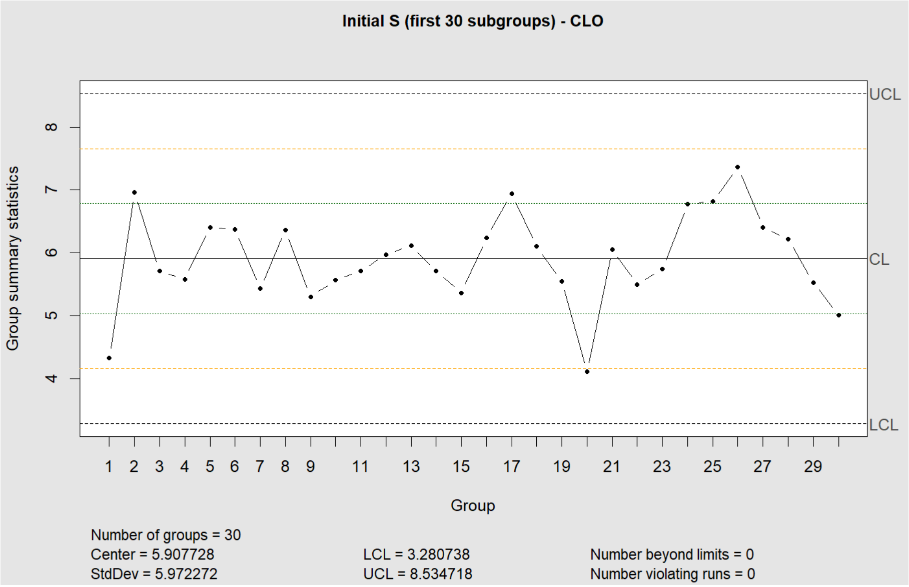

Figure 7:

Chart	CL (h)	L1	U1	L2	U2	LCL / UCL (3σ)
X-bar	19.126	18.004	20.248	16.881	21.371	15.759 / 22.493
S	5.908	5.032	6.783	4.156	7.659	3.281 / 8.535

Table 12: CLO: Phase-I limits (first 30 subgroups, $n = 24$).

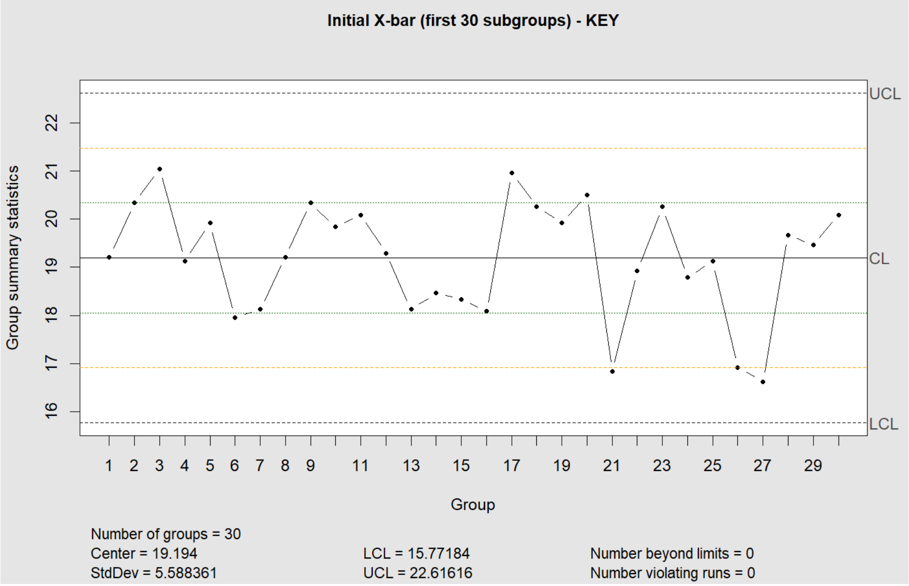

Figure 8:

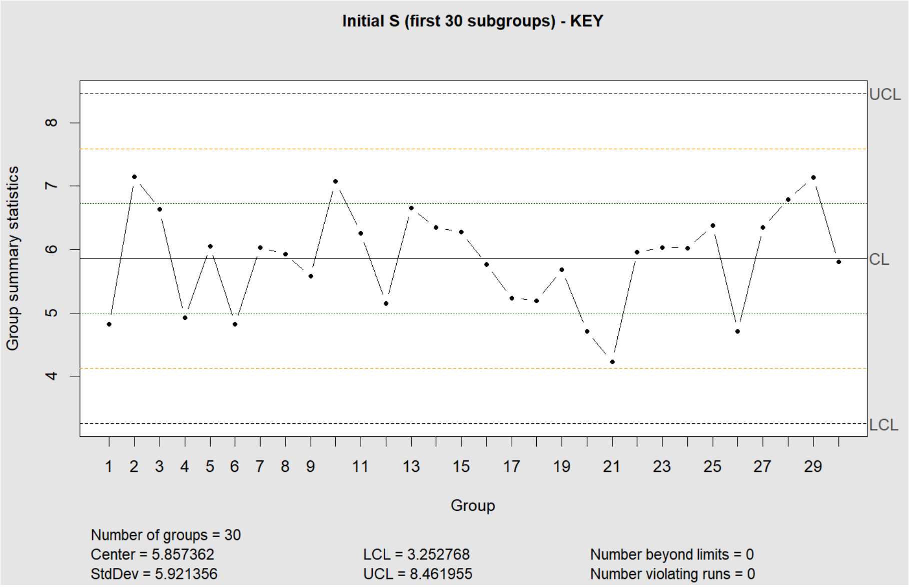

Figure 9:

Chart	CL (h)	L1	U1	L2	U2	LCL / UCL (3σ)
X-bar	19.194	18.053	20.335	16.913	21.475	15.772 / 22.616
S	5.857	4.989	6.726	4.121	7.594	3.253 / 8.462

Table 13: KEY: Phase-I limits (first 30 subgroups, $n = 24$).

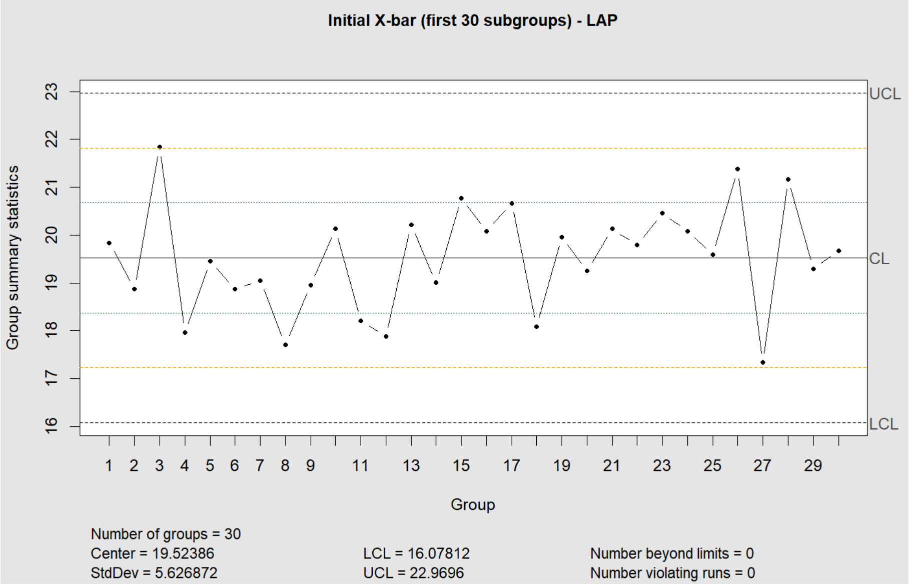

Figure 10:

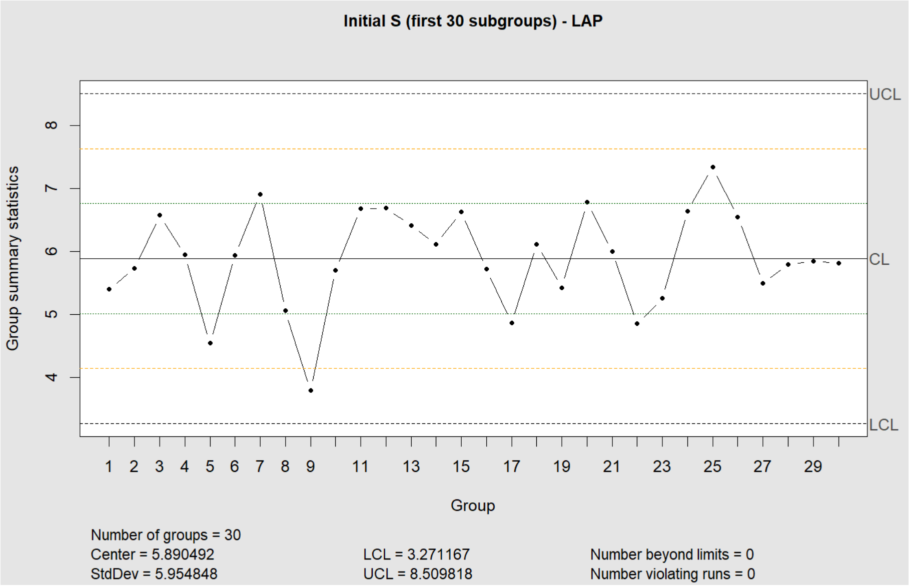

Figure 11:

Chart	CL (h)	L1	U1	L2	U2	LCL / UCL (3σ)
X-bar	19.524	18.375	20.672	17.227	21.821	16.078 / 22.970
S	5.890	5.017	6.764	4.144	7.637	3.271 / 8.510

Table 14: LAP: Phase-I limits (first 30 subgroups, $n = 24$).

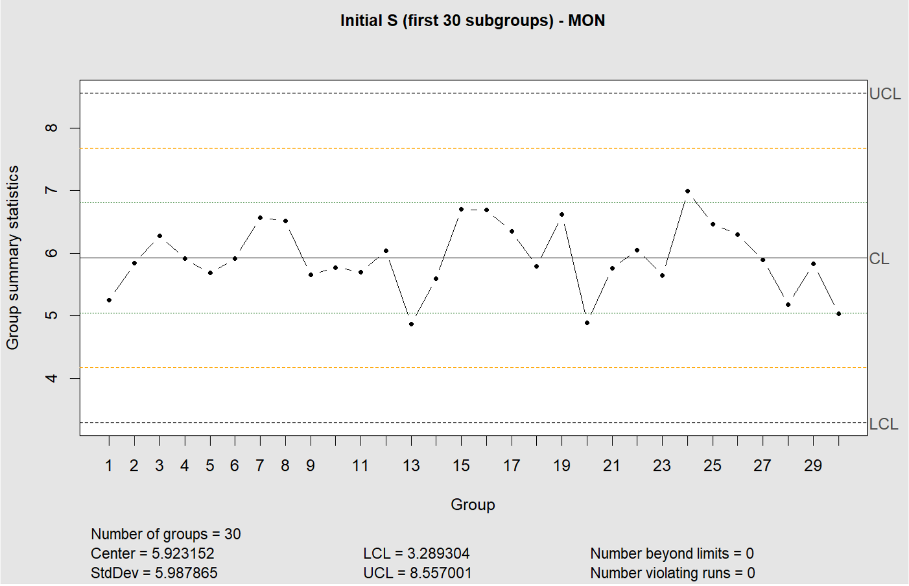

Figure 12:

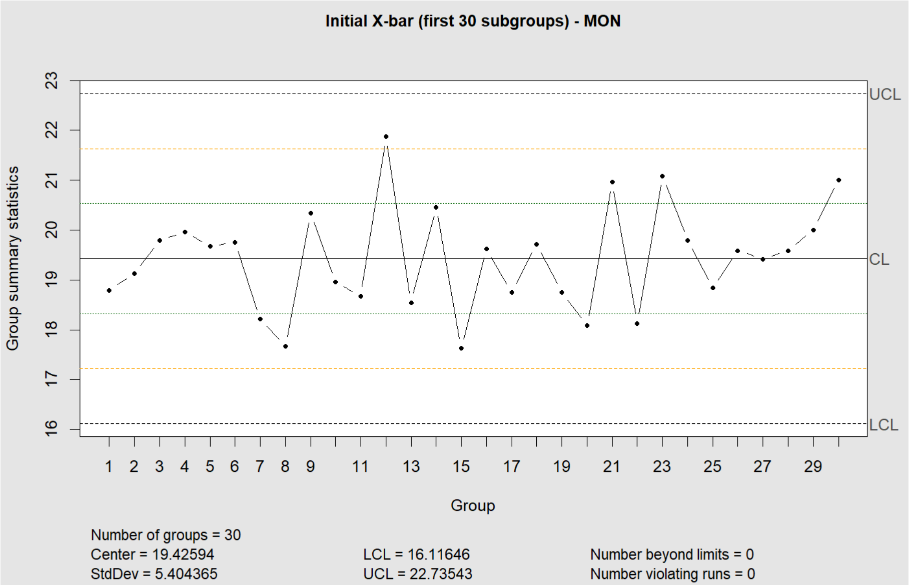

Figure 13:

Chart	CL (h)	L1	U1	L2	U2	LCL / UCL (3σ)
X-bar	19.426	18.323	20.529	17.220	21.632	16.116 / 22.735
S	5.923	5.045	6.801	4.167	7.679	3.289 / 8.557

Table 15: MON: Phase-I limits (first 30 subgroups, $n = 24$).

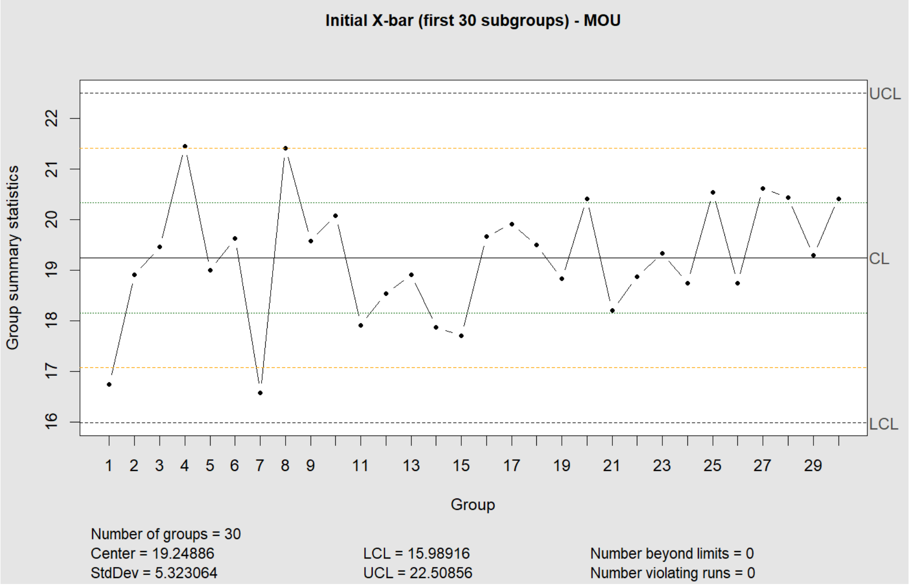

Figure 14:

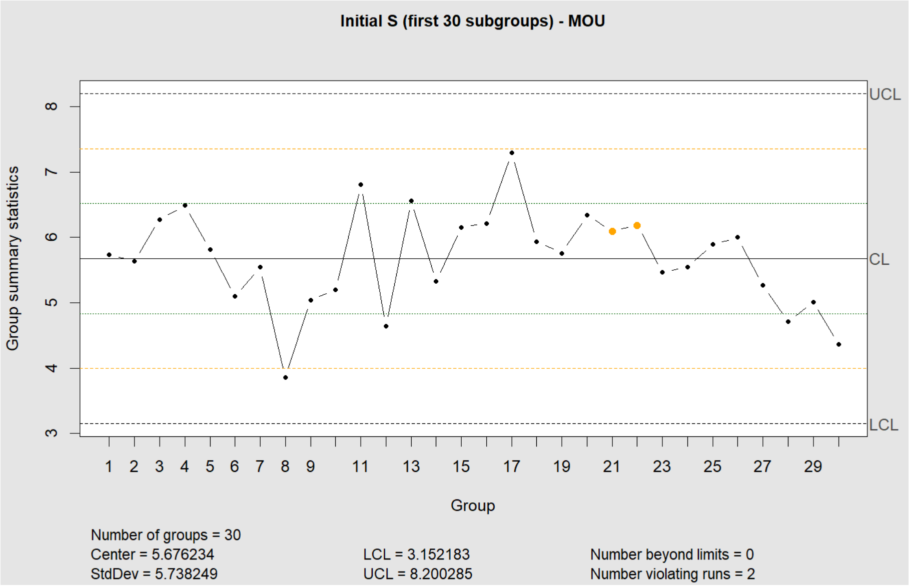

Figure 15:

Chart	CL (h)	L1	U1	L2	U2	LCL / UCL (3σ)
X-bar	19.249	18.162	20.335	17.076	21.422	15.989 / 22.509
S	5.676	4.835	6.518	3.994	7.359	3.152 / 8.200

Table 16: MOU: Phase-I limits (first 30 subgroups, $n = 24$).

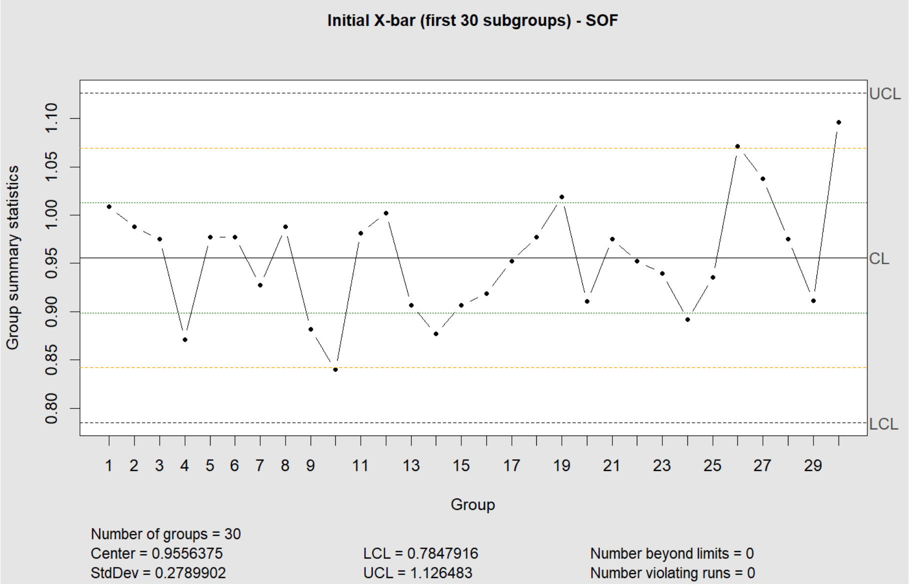

Figure 16:

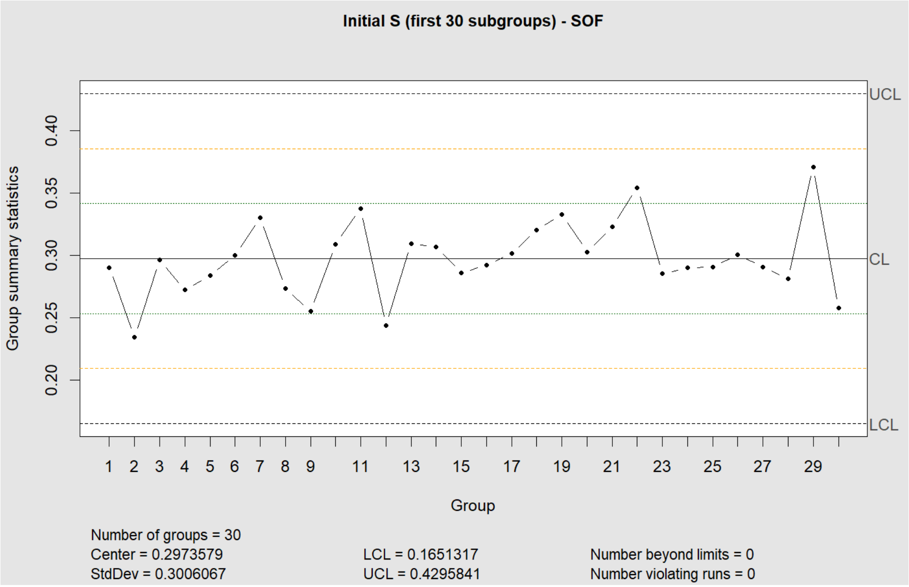

Figure 17:

Chart	CL (h)	L1	U1	L2	U2	LCL / UCL (3σ)
X-bar	0.956	0.899	1.013	0.842	1.070	0.785 / 1.126
S	0.297	0.253	0.341	0.209	0.386	0.165 / 0.430

Table 17: SOF: Phase-I limits (first 30 subgroups, $n = 24$).

5 3.3 Process Capability (VOC: 0–32 h)

Method. For each product type, the first 1 000 *time-ordered* deliveries were used to compute

$$C_p = \frac{USL - LSL}{6\sigma}, \qquad C_{pu} = \frac{USL - \mu}{3\sigma}, \qquad C_{pl} = \frac{\mu - LSL}{3\sigma}, \qquad C_{pk} = \min(C_{pu}, C_{pl}),$$

with $LSL = 0\,\mathrm{h}$ and $USL = 32\,\mathrm{h}$. We classify as *Capable* when $C_{pk} \geq 1.33$

Type	n used	μ [h]	σ [h]	C_p	C_{pu}	C_{pl}	C_{pk}	Capable?
SOF	1000.000	0.955	0.294	18.135	35.188	1.083	1.083	yes
KEY	1000.000	19.276	5.815	0.917	0.729	1.105	0.729	No
MOU	1000.000	19.298	5.828	0.915	0.727	1.104	0.727	No
CLO	1000.000	19.226	5.941	0.898	0.717	1.079	0.717	No
MON	1000.000	19.410	5.999	0.889	0.700	1.079	0.700	No
LAP	1000.000	19.606	5.934	0.899	0.696	1.101	0.696	No

Table 18: Capability indices using the first 1 000 deliveries per product type; $LSL = 0\,\mathrm{h}$, $USL = 32\,\mathrm{h}$.

Answer. With $C_{pk} \geq 1.33$ as the criterion, no product type is capable. (If a looser criterion $C_{pk} \geq 1.00$ is used, only **SOF** would be capable.)

adjustbox

Part 3.4: SPC Rule Analysis (A–C)

Table 19 summarises the SPC rule results for each product type. Rule A checks for any sample standard deviation (s) above the $+3\sigma$ limit, Rule B reports the longest run of samples within the $\pm 1\sigma$ band (indicating good control), and Rule C counts the number of runs of four consecutive $\bar{X}$ values above the $+2\sigma$ limit.

Table 19: Summary of SPC Rule Results per Product Type

Product Type	**Rule A**	**Longest Run (B)**	**Total Runs (C)**
CLO	None	35	238
KEY	None	15	236
LAP	None	19	138
MON	None	34	193
MOU	1 sample	16	285
SOF	None	21	278

Interpretation Across all product types, variability remained well controlled with almost no Rule A violations. Long runs within the $\pm 1\sigma$ range (Rule B) show that most processes were stable and consistent, particularly for CLO and MON. However, the high number of Rule C occurrences indicates that several processes—especially MOU, SOF, and CLO—experienced *frequent upward shifts in their mean delivery times*. This suggests that while short-term variation is stable, the overall process mean may require recalibration or further investigation.

4.1 Type I (Manufacturer's) error — theoretical rates

Assume H_0: process is in control and centred; chart statistic is Normal with $\sigma_{\text{chart}} = (\text{UCL} - \text{LCL})/6$.

Rule A (S-chart, one-sided +3σ)

$$\alpha_A = P(Z > 3) = 0.00135 \quad \text{per subgroup (one-sided).}$$

Rule B (longest run within ±1σ) This is a *good-control diagnostic*, not an alarm rule in our task; no Type I rate is defined unless a trigger k is specified. ⇒ **N/A**.

Rule C (X-bar, 4 consecutive above +2σ) Under H_0, $P(Z > 2) = 0.0228$. For 4-in-a-row windows,

$$\alpha_C = [P(Z > 2)]^4 = (0.0228)^4 \approx 2.70 \times 10^{-7} \quad \text{per 4-window.}$$

Table 20: Type I error per opportunity (theoretical)

Rule	Event	Type I rate
A	s point > UCL (+3σ)	0.00135 per subgroup
B	longest run within ±1σ	N/A (no alarm threshold)
C	4 consecutive $\bar{X} > +2\sigma$	2.70×10^{-7} per 4-window

4.2 Type II (Consumer's) error for bottle filling

X-bar chart limits: CL = 25.050, UCL = 25.089, LCL = 25.011 L. Actual process (unknown to us): mean $\mu_1 = 25.028$ L and $\sigma_{\bar{X}} = 0.017$ L.

$$\beta = P(\text{LCL} \le \bar{X} \le \text{UCL} \mid \mu_1, \sigma_{\bar{X}}) = \Phi\Big(\frac{25.089 - 25.028}{0.017}\Big) - \Phi\Big(\frac{25.011 - 25.028}{0.017}\Big) \approx \Phi(3.588) - \Phi(-1.000)$$

Hence, **Type II error** $\beta \approx 0.841$ (84.1%), and **power** $1 - \beta \approx 15.9\%$. *Interpretation:* with these limits and higher variability, the chart misses the shift most of the time.

6 Coffee Shop Optimisation (Part 5)

The aim of this section was to determine the optimal number of baristas that maximises daily profit for two coffee shops, based on real operational data (`timeToServe.csv` and `timeToServe2.csv`). Each dataset contained one year's worth of sales records, showing the number of baristas on duty (`V1`) and the corresponding customer service time in seconds (`V2`).

A simple profit model was used:

$$\Pi(c) = 30 \times \text{Served}_c - 1000 \times c$$

where c is the number of baristas, R30 represents the average material profit per customer served, and R1000 is the daily personnel cost per barista. Service reliability was defined as the percentage of customers served within 180 s. The analysis was conducted in `R` using the `analyse_shop()` function, which calculated mean service time, service capacity, and expected profit for each staffing level ($2 \le c \le 6$).

Results for Shop 1

Table 21: Performance metrics for different barista levels (Shop 1)

Baristas	n_{obs}	Mean Service (s)	Capacity/h	Capacity/day	Demand/day	Served/day	Profit/day (R)	Reliability (%)
2	3556	100.17	71.88	718.77	546.8	546.8	14404.08	100
3	12126	66.61	162.13	1621.34	546.8	546.8	13404.08	100
4	29305	49.98	288.11	2881.13	546.8	546.8	12404.08	100
5	56701	39.96	450.43	4504.30	546.8	546.8	11404.08	100
6	97895	33.36	647.57	6475.67	546.8	546.8	10404.08	100

From the results, Shop 1 achieves the maximum daily profit of approximately R14 404 with **2 baristas**. Increasing the number of baristas improves service capacity but adds unnecessary cost, reducing profit. Reliability remains at 100% for all levels due to low daily demand compared to capacity.

Results for Shop 2

Table 22: Performance metrics for different barista levels (Shop 2)

Baristas	n_{obs}	Mean Service (s)	Capacity/h	Capacity/day	Demand/day	Served/day	Profit/day (R)	Reliability (%)
2	8859	141.51	50.88	508.78	541.93	508.78	13263.44	100
3	19768	115.44	93.55	935.54	541.93	541.93	13257.86	100
4	35289	100.02	143.98	1439.78	541.93	541.93	12257.86	100
5	54958	89.44	201.26	2012.61	541.93	541.93	11257.86	100
6	78930	81.64	264.57	2645.67	541.93	541.93	10257.86	100

Shop 2 shows a similar pattern, with the highest profit of approximately R13 263 achieved at **2 baristas**. Although additional staff increase service capacity, the marginal gain in throughput does not justify the higher labour cost. Reliability remains at 100%, as customer demand is well below available capacity for all staffing levels.

Discussion

Both shops operate below their full capacity, meaning current demand can be met with minimal staff while maintaining excellent service reliability. From an operational perspective, deploying more than two baristas per shift leads to overstaffing and reduced profit efficiency. These results support management decisions to align staffing schedules with actual demand patterns while maintaining service quality.

Part 6: Design of Experiments (DOE) and ANOVA

Objective

This section investigates whether the mean delivery hours for the **Laptop (LAP)** product type differ significantly between two consecutive years, 2022 and 2023. The analysis uses a one-way Analysis of Variance (ANOVA) test to determine whether any observed change in the process mean is statistically significant at a 95% confidence level.

Method

From the dataset `sales2026and2027.csv`, all products with identifiers beginning with "LAP" were extracted. Delivery hours (`deliveryHours`) were treated as the dependent variable, while order year (`orderYear`) was used as the independent grouping factor. The following R code was applied:

```
lap_data <- sales %>%
  filter(grepl("^LAP", ProductID)) %>%
  mutate(orderYear = as.factor(orderYear))

anova_lap <- aov(deliveryHours ~ orderYear, data = lap_data)
summary(anova_lap)
```

Results

The ANOVA summary output is shown in Table 23. The test statistic was $F = 0.496$ with a p-value of 0.481, indicating that there is no statistically significant difference in mean delivery hours between 2022 and 2023 ($p > 0.05$). This implies that the yearly change in average delivery performance can be attributed to common-cause variation rather than any special-cause disturbance.

Table 23: ANOVA summary for delivery hours (LAP: 2022–2023).

Source	**Df**	**Sum Sq**	**Mean Sq**	**F value**	**Pr(¿F)**
Order Year	1	18.15	18.15	0.496	0.481
Residuals	10,205	373,353	36.59		

Discussion

Figure 18 presents the boxplot of delivery hours for both years. The median and interquartile ranges overlap substantially, confirming the statistical result that mean delivery hours did not change significantly between years.

This outcome is consistent with the **Statistical Process Control (SPC)** findings from Part 3, where the LAP process exhibited stable Phase I control limits ($\bar{X}_{CL} = 19.524$ h, UCL $= 22.970$ h, LCL $= 16.078$ h) and an S_{CL} of 5.890 h. The ANOVA therefore validates that the process mean remained within these established control boundaries across consecutive years, indicating sustained process capability and stability.

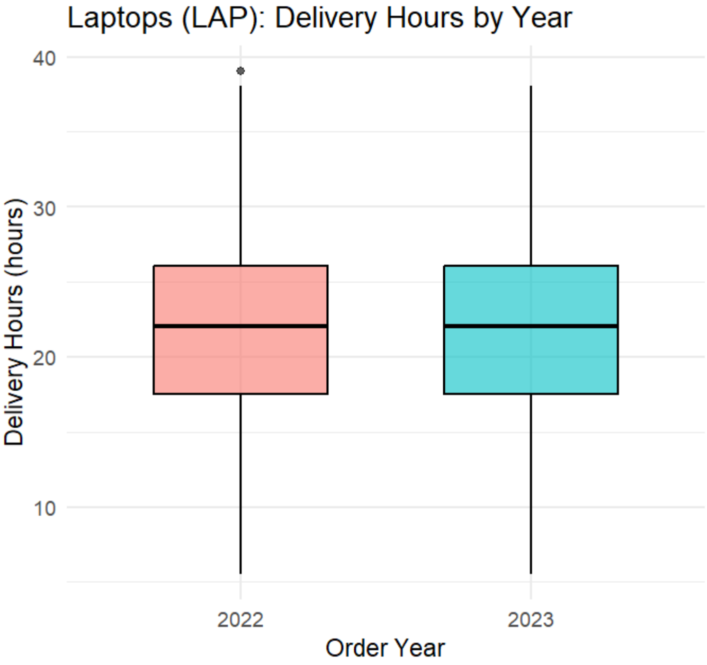

Figure 18: Laptops (LAP): Delivery hours by year (boxplot). The overlapping medians and spreads confirm no significant difference ($p = 0.481$).

Conclusion

The one-way ANOVA shows that year-to-year variation in laptop delivery hours was not statistically significant. Combined with the SPC results, this demonstrates that the delivery process was well-controlled and capable of maintaining consistent performance across the study period.

7. Reliability of Service

At a car rental agency, the number of employees on duty was recorded over 397 days, with the following frequency distribution:

Table 24: Observed number of workers on duty.

Workers on duty	12	13	14	15	16
Number of days	1	5	25	96	270

7.1 Estimating the reliability of service

Reliable service is defined as having at least 15 employees on duty. From the data, the number of reliable days is:

$$96 + 270 = 366 \text{ out of } 397,$$

giving a reliability proportion of

$$\hat{P}(\text{reliable}) = \frac{366}{397} = 0.922.$$

Hence, the expected number of reliable days in a 365-day year is:

$$365 \times 0.922 = \mathbf{337} \text{ days per year (approximately).}$$

7.2 Optimising personnel assignment and profit

We model the number of employees who report for duty, X, as a binomial random variable:

$$X \sim \text{Binomial}(S, p),$$

where S is the number of employees scheduled and p is the probability that any one employee reports for duty.

From the sample data, the average number of employees present was

$$\bar{x} = \frac{12(1) + 13(5) + 14(25) + 15(96) + 16(270)}{397} = 15.584,$$

and since a maximum of $S = 16$ employees were scheduled, the estimated attendance probability is

$$\hat{p} = \frac{\bar{x}}{S} = \frac{15.584}{16} = 0.974.$$

Cost model.

- Each employee costs R25 000 per month, or R300 000 per year.
- A "problem day" (fewer than 15 present) results in an average sales loss of R20 000.

The expected annual cost function is:

$$C(S) = 300{,}000S + 20{,}000(365)\,P(X < 15),$$

where $P(X < 15)$ is computed using the binomial distribution with parameters $(S, \hat{p})$.

Computation in R. The expected annual costs for $S = 15$ to 22 were evaluated:

Table 25: Expected probability of a problem day and annual cost.

Scheduled staff S	$P(X < 15)$	**Expected annual cost (R)**
15	0.326	6,881,108
16	0.063	5,264,506
17	0.009	5,162,620
18	0.001	5,407,593
19	0.0001	5,700,740
20	8.7×10^{-6}	6,000,063
21	6.7×10^{-7}	6,300,005
22	4.8×10^{-8}	6,600,000

The minimum expected annual cost occurs at $S^* = 17$ scheduled employees per day.

Interpretation. Adding one more employee (from 16 to 17) substantially reduces the probability of service failure at a modest additional wage cost. Beyond 17 employees, extra hires offer diminishing returns. Therefore, scheduling 17 employees per day maximises expected profit and ensures a high reliability level ($P(X \geq 15) \approx 0.991$).

discusion

Based on 397 days of data, the company can expect reliable service on approximately 337 days per year with current staffing. Modelling attendance as a binomial process shows that employing **17 staff members per day** minimises total annual cost and maintains nearly 99% reliability. This balance ensures both operational stability and cost efficiency.

References

Dirkse van Schalkwyk, T., 2025. *R code examples for Statistical Process Control.* Lecture slides, Industrial Engineering 344. Stellenbosch University.

OpenAI, 2025. *ChatGPT (GPT-5) conversational assistance for R coding and academic writing.* 23 October 2025. Available at: https://chat.openai.com/ [Accessed 23 October 2025].

7 Conclusion

Conclusion

This report brought together several areas of Quality Assurance and data analysis to demonstrate the practical application of statistical tools within an industrial engineering context. Across the different parts, descriptive statistics were used to understand the underlying data, while Statistical Process Control (SPC) methods were applied to monitor process stability and detect variation. Capability indices were calculated to assess whether each process could meet its specifications, and potential sources of error were explored through Type I and Type II analyses.

Later sections focused on optimisation and modelling, using R to simulate real operational decisions such as staffing levels and service reliability. The analyses showed how statistical reasoning and data-driven methods can support better decision-making in service and manufacturing environments.

Overall, the work reinforced key learning outcomes of QA344: the ability to apply statistical concepts, interpret results meaningfully, and translate data into actionable quality improvements. While some results may vary with further refinement or alternative data treatments, the approach used here reflects a sound understanding of both theory and practice in quality assurance.