2025

ECSA GA4 Project

QUALITY ASSURANCE 344
WILKIN, CH, MISS [27033635@SUN.AC.ZA]

Table of Contents

List of Figures

Introduction

This report serves as the final submission for the ECSA GA4 Quality Assurance 344 module, demonstrating applied data analysis, statistical process control, and optimisation techniques using various example datasets. The objective is to exhibit the ability to manipulate, analyse, and interpret industrial data through R programming in a structured and professional engineering context.

The report focuses on a retail and distribution case study based on an electronics company, using datasets spanning 2022-2027. The analysis proceeds through six progressive parts: data exploration and descriptive statistics, statistical process control (SPC) for 2026-2027 sales, risk analysis and data correction, optimisation of operational efficiency, a design of experiments (DOE) and ANOVA test, and a service reliability model. Each part builds upon the previous to form a comprehensive statistical evaluation of performance and process capability within an industrial system.

Key analytical methods include the use of control charts, process capability indices (C_p, C_{pk}), Type I and Type II error evaluation, and hypothesis testing through ANOVA. The final sections extend these methods to practical optimisation problems, such as workforce allocation and service reliability modelling, linking statistical evidence with operational and financial decision-making. Overall, this report applies engineering statistics to achieve the GA4 outcome, demonstrating competency in data analysis, interpretation, and engineering decision support.

1. Basic Data Analysis

The purpose of this section is to inspect, visualise, and correct the data provided by the previous analyst at an electronics retail and distribution company. Datasets are provided detailing customer profiles, product lists, and sales data for the period of 2022 to 2023. After the data is sufficiently explored and analysed, actionable insights and recommendations are generated for the management team to consider.

1.1. Data Loading and Inspection

The 4 datasets are loaded under the following names:

The customer profiles are in the “customers” dataset. The product catalogue data is in the “products” dataset. The head office product list is in the “products_headoff” dataset, and the sales information for the 2022 to 2023 period is in the “sales” dataset.

A quick glance at the dimensionality of each dataset reveals several insights. Firstly, “products” only has 60 row entries while “products_headoff” has 360 entries. This could mean that “products” is a master list of all the products that the company sells and “products_headoff” includes all variants (colour, model, etc.) of the products in the catalogue. Additionally, these datasets demonstrate an interesting interaction between the ProductID and Category feature. In “products”, there is a mismatch between the entries which have ProductID codes of SOF, CLO, LAP, MON, KEY, and MOU which should correspond to the categories Software, Cloud Subscription, Laptop, Monitor, Keybaord, and Mouse. But they do not correspond. This might mean that the Category column was reassigned without updating ProductIDs or that the Category feature was inputted incorrectly. In the “products_headoff” dataset, the ProductID codes do indeed match the categories, but there are several entries with an NA code. This might mean that that branch sold an item that is not in the standard catalogue, in which case the master list needs to be expanded to include all items sold at all branches. Or it is a symptom of faulty data entry where a mouse was sold but the employee responsible for inputting the sale did not select the appropriate product code. There are also discrepancies between the Description, SellingPrice, and Markup features for products with the same ProductID between the “products” and “products_headoff” datasets.

Moreover, the “customers” dataset has 5000 entries while the “sales” dataset has 100 000 entries. For a 2-year period, this implies that each customer performs 10 transactions per year. This is a strong indicator of customer loyalty.

The “sales” dataset also has 9 features which is more than any other dataset. To reduce this, the orderDay, orderMonth, and orderYear columns are concatenated into a single “date” feature. This reduces the number of columns to 7.

1.2. Summary Statistics

The summary statistics for "customers" shows that the majority of customers are around 51 years old and have an income of 80 000-85 000. So, to maximise sales it might be wise to target these particular demographics.

The orderTime feature in the "sales" dataset also reveals that the majority of sales occur at around midday. Therefore, it would be wise to plan for high customer flux in the physical and online stores during this time so that the quality of customer service is not diminished by traffic in the company stores or website.

The 'City' feature in the "customers" dataset has a cardinality of 7. The business could expand the geographical scope of their operations in order to reach more customers and, hopefully, incur greater revenue.

There are no missing values in any of the datasets.

1.3. Data Filtering and Sub-setting

Filtering is performed to identify the most valuable customers and the most valuable products measured by quantity of items purchased per customer and total quantity of product purchased respectively. The top 10 customers and products are as follows:

	CustomerID	total_qty
	<chr>	<int>
1	CUST1193	14704
2	CUST1791	14626
3	CUST596	14212
4	CUST3721	13852
5	CUST2527	13773
6	CUST2277	13538
7	CUST1427	13335
8	CUST4729	12938
9	CUST3944	12855
10	CUST1501	11958

Figure 1: Top 10 customers by quantity of products purchased

	ProductID	total_qty
	<chr>	<int>
1	MOU059	29675
2	SOF001	29336
3	SOF004	29219
4	SOF010	29168
5	MOU058	28924
6	MOU054	28875
7	MOU052	28804
8	SOF007	28517
9	MOU057	28423
10	SOF005	28412

Figure 2: Top 10 products by quantity purchased

The management team should consider sending out email celebration cards to their top 10 customers notifying them of their position. This will show gratitude for continued support, foster good customer relations, and increase customer satisfaction for those customers that are responsible for the majority of the business's revenue.

The management team should also consider offering promotions on their top 10 bestselling products. This will help maximise revenue. The top 10 best-selling products only include software and mice. These 2 product categories also have the lowest selling prices so offering promotions on them will risk very little profit loss.

1.4. Data Visualisation

Several insights are drawn from visualisations of the datasets.

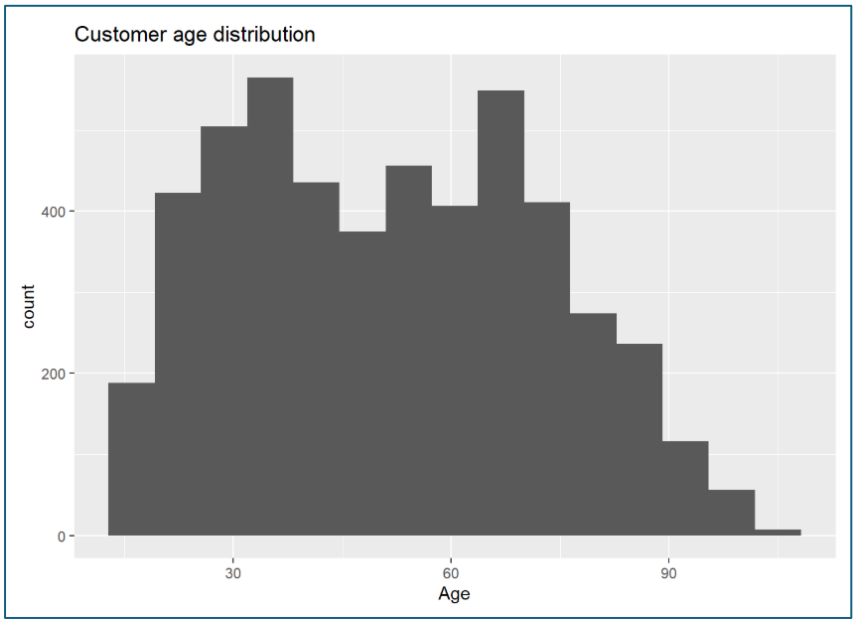

Figure 3: Histogram of customer ages

Figure 3 shows a slightly bimodal customer distribution. One group between roughly 25-40 and a second group between roughly 55-70. These groups could be further separated and more specifically targeted when plotted against their income.

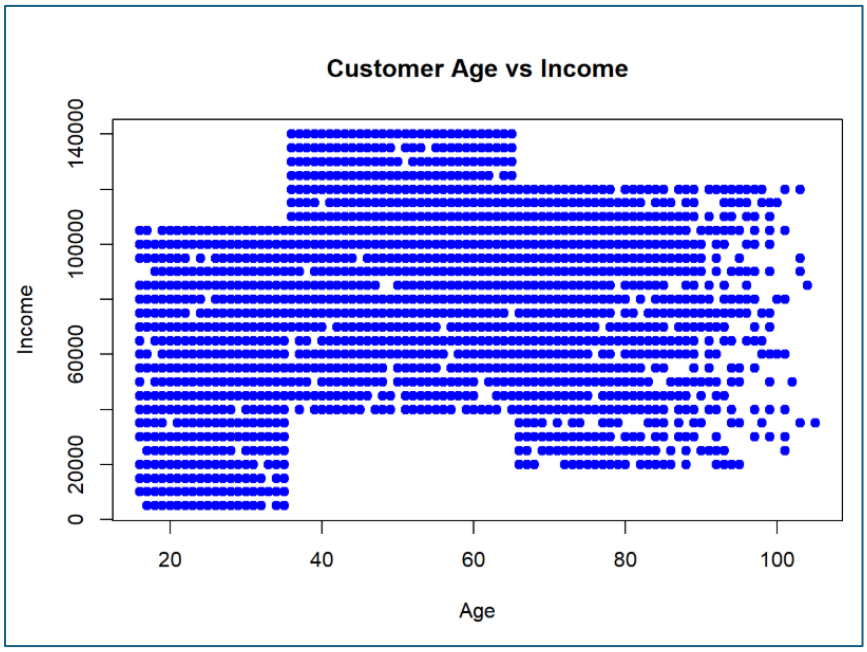

Figure 4: Scatterplot of customer age vs income

Figure 4 clearly shows 3 distinct groups. Group 1 is younger customers with lower income. These could be students or young professionals buying entry level accessories and software for their degrees of first jobs. Group 2 is the middle-aged customers with higher income. These are career professionals who are investing in high-quality technology to improve the

quality of their lives and their work. Group 3 are the older customers with middle income. They are likely to purchase essential products known for reliability, simplicity, and longevity rather than high-end, cutting-edge technology. Targeted marketing could be used to advertise specifically desirable products for each of these groups. Entry-level products for Group 1, high-end, cutting-edge products for Group 2, and user-friendly, reliable products for Group 3. This targeted marketing campaign will help maximise revenue from each of these customer groups.

Additionally, the sales trends over time can be explored.

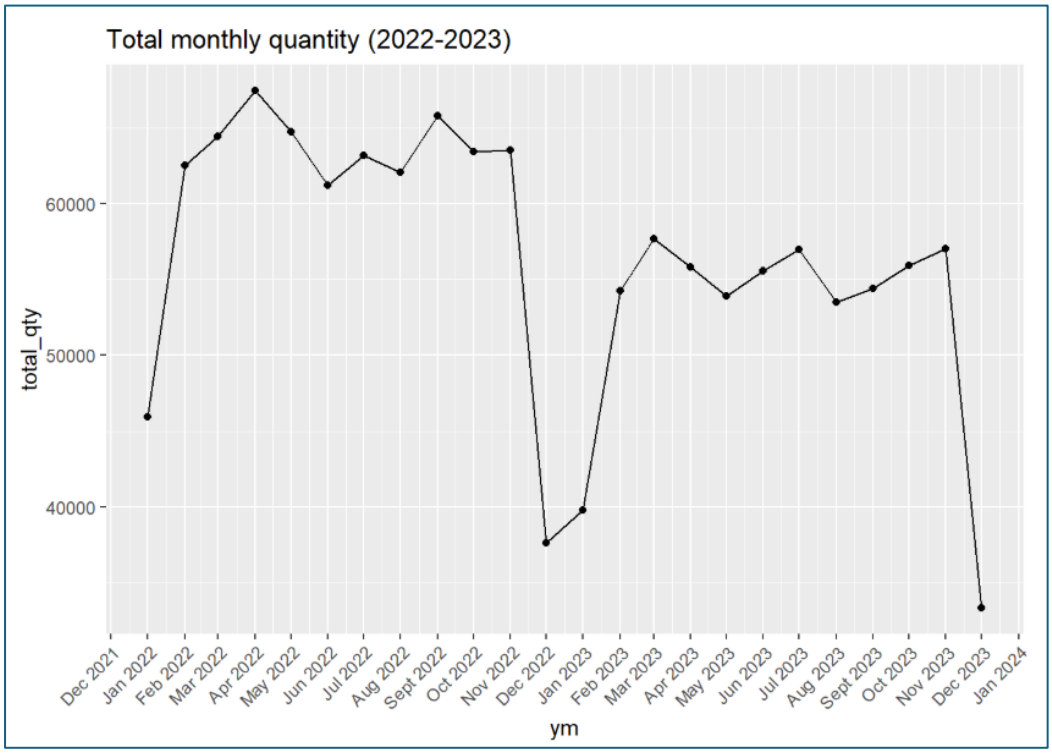

Figure 5: Plot of total monthly sales for 2022-2023

Figure 5 clearly shows a significant decrease in sales in the months of December and January for both years. December is usually a peak shopping month due to holidays so this dip in December can only be explained by store closure for stocktakes or holidays. The sales that are incurred during this dip might be only the sales coming from online orders. In this case, it is highly advised that the management team keep the stores open during December so as to maximise revenue gained from this peak shopping month. Lower sales in January might be explained by post-holiday slowdowns. But if the January dip is also caused by store closures, management should consider the same approach recommended for December. In addition to this, there is a significant decrease in average total sales for the year of 2023 when compared to 2022. This might be due to price increases causing reduced demand. To investigate this, the SellingPrice feature is concatenated from the "products" dataset (so that catalogue listed selling prices are used) into the "sales" dataset for corresponding ProductID values. The average price per month is then calculated and plotted over the total monthly sales to investigate potential correlation between increased price and decreased sales.

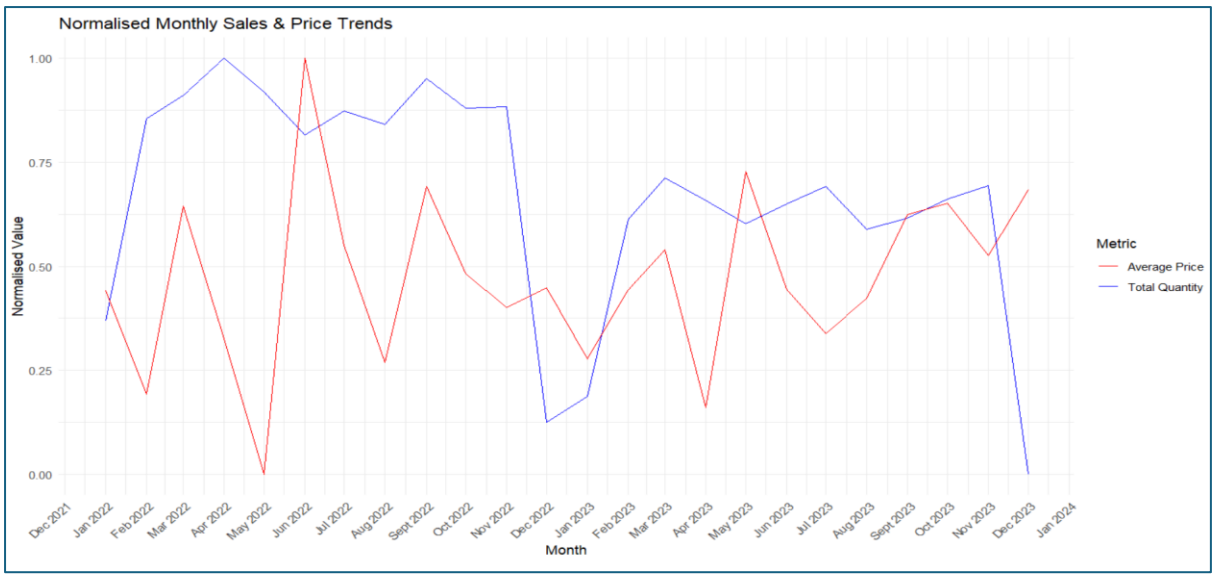

Figure 6: Plot of average monthly selling price with total sales per month

Figure 6 reveals that there is not a particularly strong correlation between average selling price per month and total sales per month. This means that the decrease in total sales from 2022 to 2023 must be due to some other operational issue. Management should initiate an investigation into what caused this decrease and seek to remedy the problem as soon as possible.

Finally, for both "products" and "products_headoff", even though SellingPrice is quite significantly skewed to the right, the markup is fairly constant. This implies that the company's pricing strategy is standardised rather than dynamic.

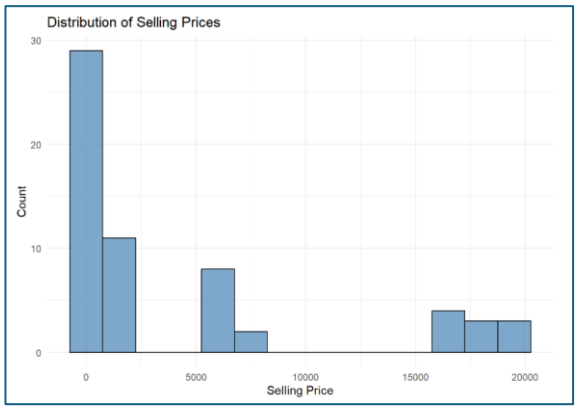

Figure 7: Histogram of Selling Price

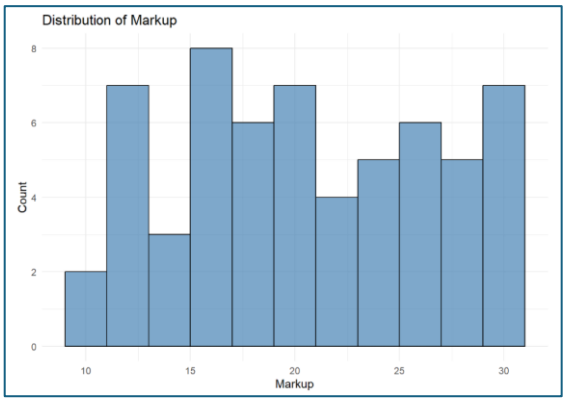

Figure 8: Histogram of Markup

Figure 9: Scatterplot of Markup vs Selling Price

Figure 9 clearly shows that for the group of products with the highest selling prices; they tend to have lower markups when compared to middle- or lower-priced products. This is probably because the company wants to price high-end items more competitively to encourage sales. This lower relative margin allows the product to remain attractive to customers even considering the high selling price. This indicates a volume-driven pricing strategy. Revenue could be increased by further leaning into this strategy and increasing markup on the less expensive items.

1.5. Exploring Relationships

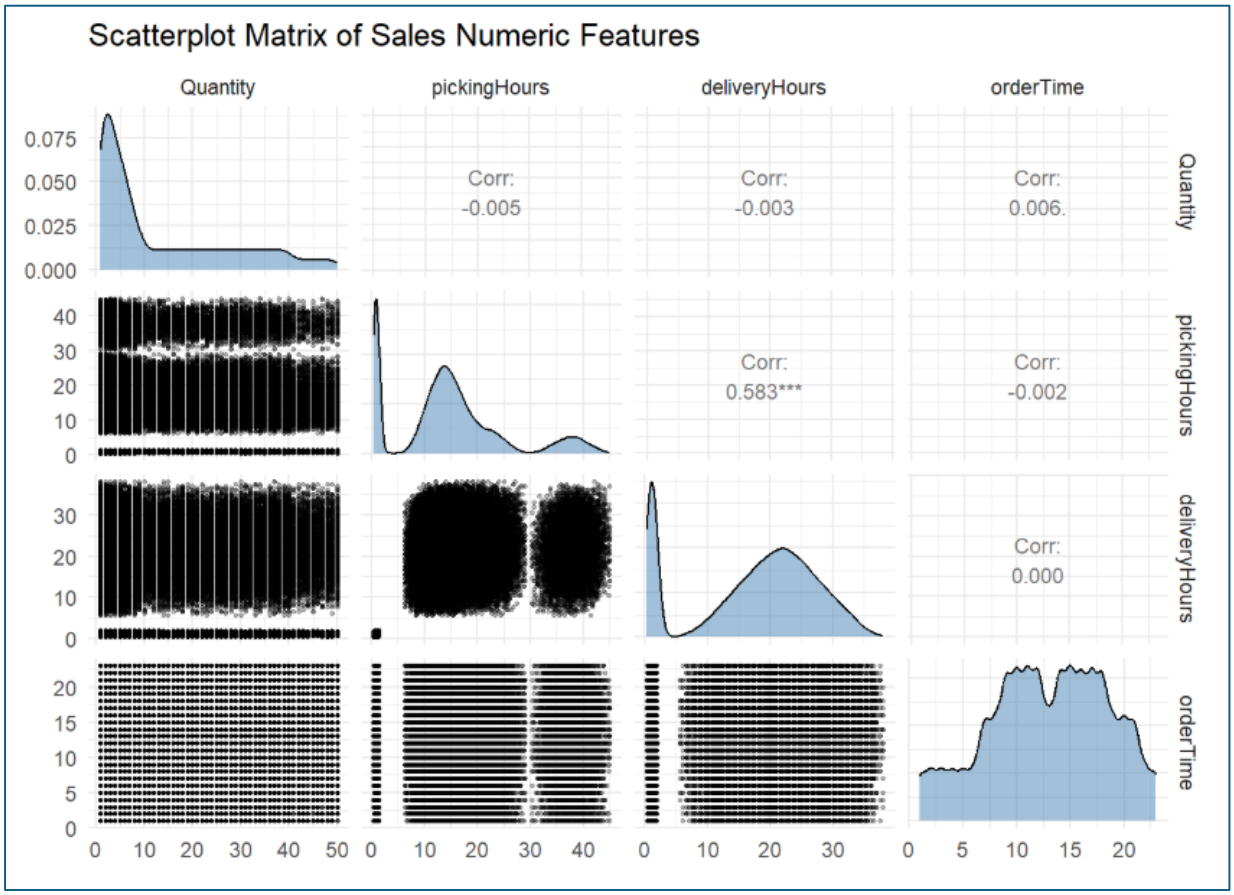

Figure 10: Scatterplot matrix of all numeric features in the "sales" dataset

Figure 10 shows that there is hardly any correlation between Quantity and any of the other numeric features as well as between orderTime and any of the other features. The pickingHours and deliveryHours features, on the other hand, have a strong positive correlation with one another. This is displayed in *Figure 11*.

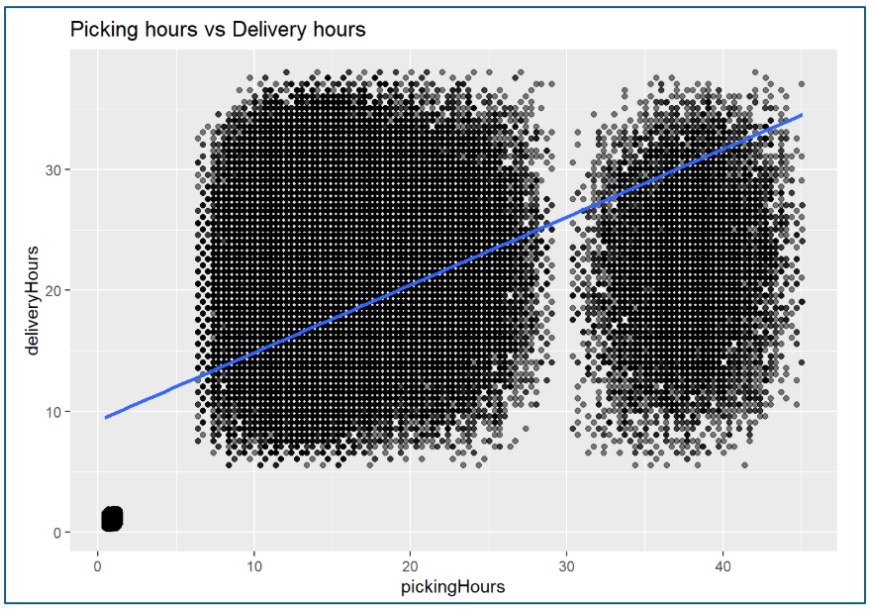

Figure 11: Scatterplot (with LOBF) showing Picking Hours vs Delivery Hours of sales

This means that decreasing pickingHours as well as deliveryHours is in the best interest of the management team. Decreasing either or both of these features will reduce customer lead time and improve customer satisfaction.

1.6. Final Recommendations for the Management Team

- Expand master product catalogue to include codes, categories, and descriptions for all products sold at all branches.
- Provide customer-specific targeted marketing based on age and income.
- Show gratitude in the form of discounts or promotions to most loyal customers. Run promotions on most popular products.
- Create an Ishikawa (fishbone) diagram to identify the possible causes of the significant decrease in sales from 2022 to 2023. Implement actionable solutions to solve the identified problems.
- Maximise gains from revenues in peak shopping months of December and January by keeping stores open and performing stocktakes outside of peak sales times.
- Increase percentage markup on lower- to middle-priced items to maximise profit from highest sales volume products.
- Decrease customer lead time by decreasing picking hours and/or delivery hours on all deliveries.

2. Statistical Process Control (SPC) for 2026-2027 sales

First, the data in "sales2026and2027.csv" is ordered so that the sales entries appear in chronological order. The dataset is sorted by year, month, day, and then picking hour. The customer lower and upper specification limits for delivery time (in hours) are defined as LSL = 0 and USL = 32 respectively. The data is then subsetted so that there is a new dataset for each product type (software, keyboards, monitors, etc.). Each subset is then grouped into 24-hour samples, and the first 30 samples are extracted and labelled as calibration data.

2.1. Initialise $\bar{x}$-charts and s-charts for every product type

The $\bar{x}$-charts and s-charts are initialised for every product type. The deliveryHours values are extracted from the ordered dataset and a deliveryHoursXXX subset is created where XXX represents the ProductID for each product type. The length of each deliveryHoursXXX subset is then found and divided by 24, yielding the following results:

Table 1: Number of days' worth of hourly data for each product type

ProductID	**# days' worth of hourly data**
SOF	864
CLO	649
LAP	425
KEY	746
MON	619
MOU	860

Each value is rounded down to the nearest integer so that the subset only contains complete samples (days) of data.

The first 30 samples of size 24 are used to determine centrelines, 1σ, 2σ, and 3σ control limits with the 3σ limits being the outer control limits (UCL and LCL) for the control charts. These first 30 samples represent the error-free period. All samples in the initialisation phase should be between the upper and lower control limits. After all the $\bar{x}$- and s-charts are generated, out-of-control signals are identified and removed. Then, all charts are regenerated with the updated, in-control data. There are no out-of-control signals in any of the s-charts. However, the $\bar{x}$-charts for SOFTWARE, CLOUD SUBSCRIPTION, and MOUSE product types all have out-of-control-signals that need to be removed. The s-charts for SOFTWARE, CLOUD SUBSCRIPTION, and MOUSE product types are then updated based on the removed signals.

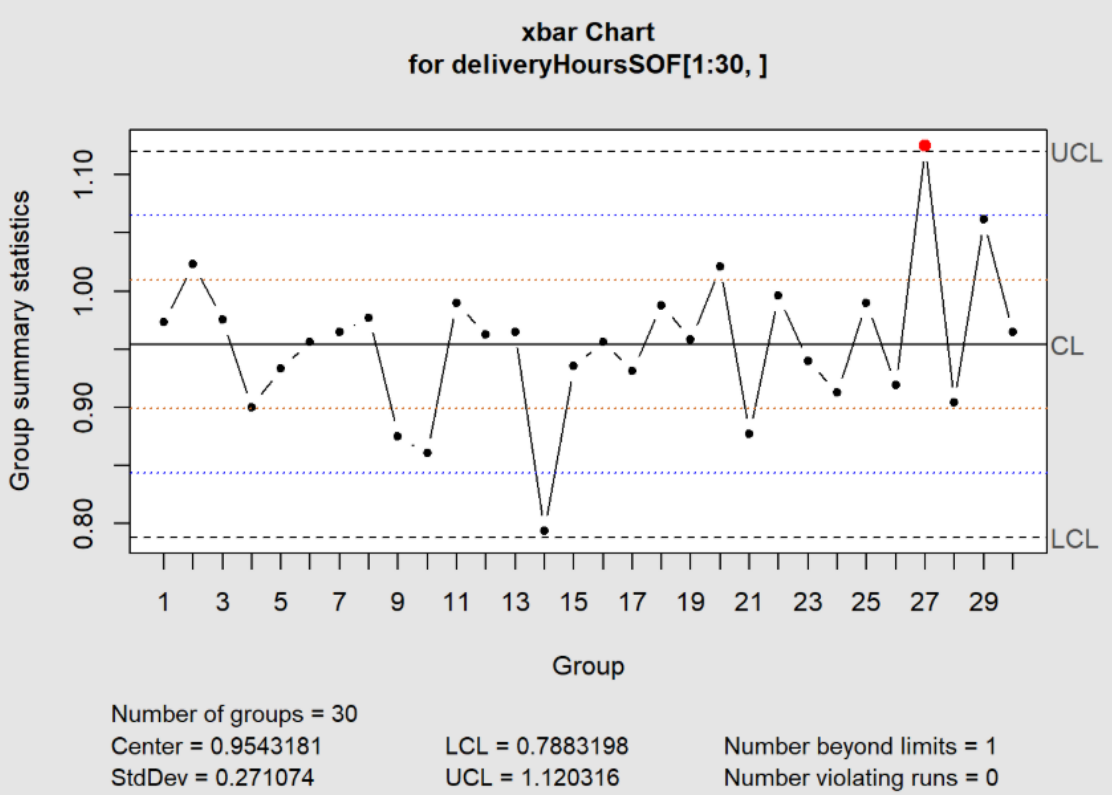

Figure 12: Average deliveryHours of SOFTWARE for the first 30 days

The initial x̄-chart for the software delivery hours shows that sample 27 is above the upper control limit. Consequently, this sample is removed, and all parameters are recalculated.

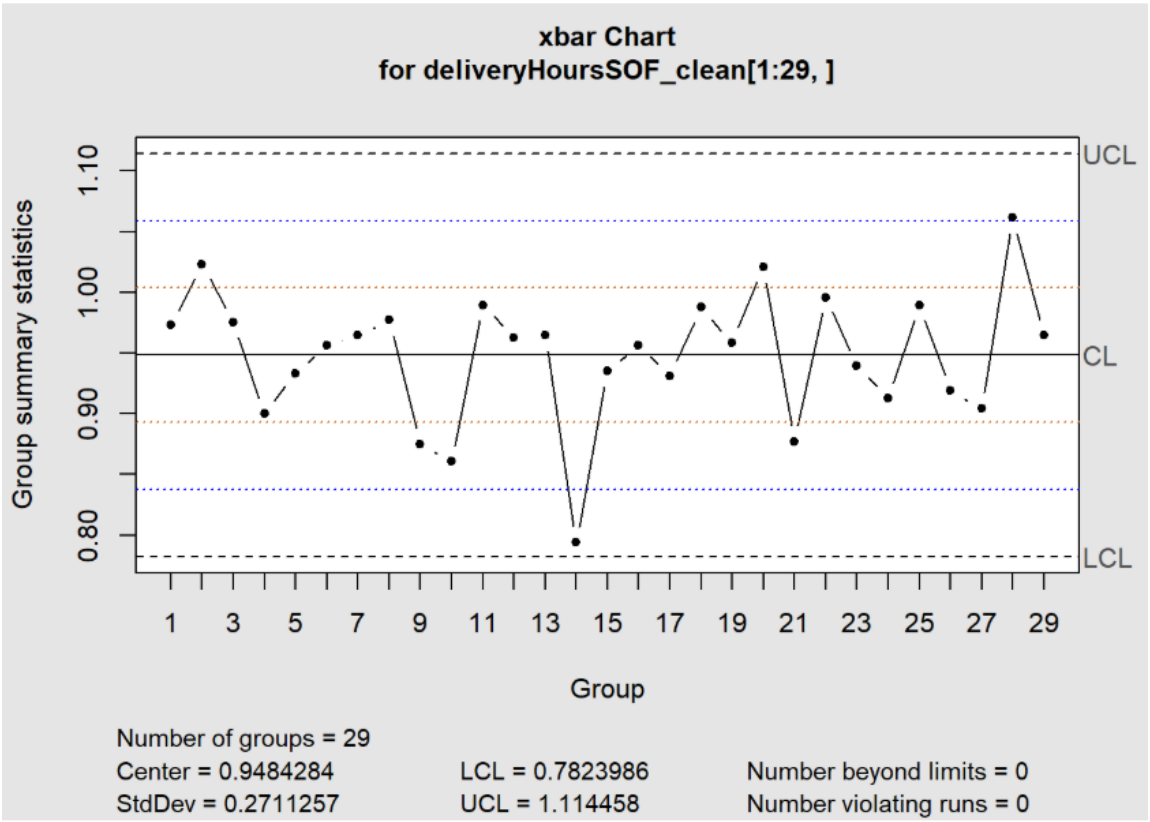

Figure 13: Average deliveryHours of SOFTWARE for the first 30 days with out-of-control samples removed

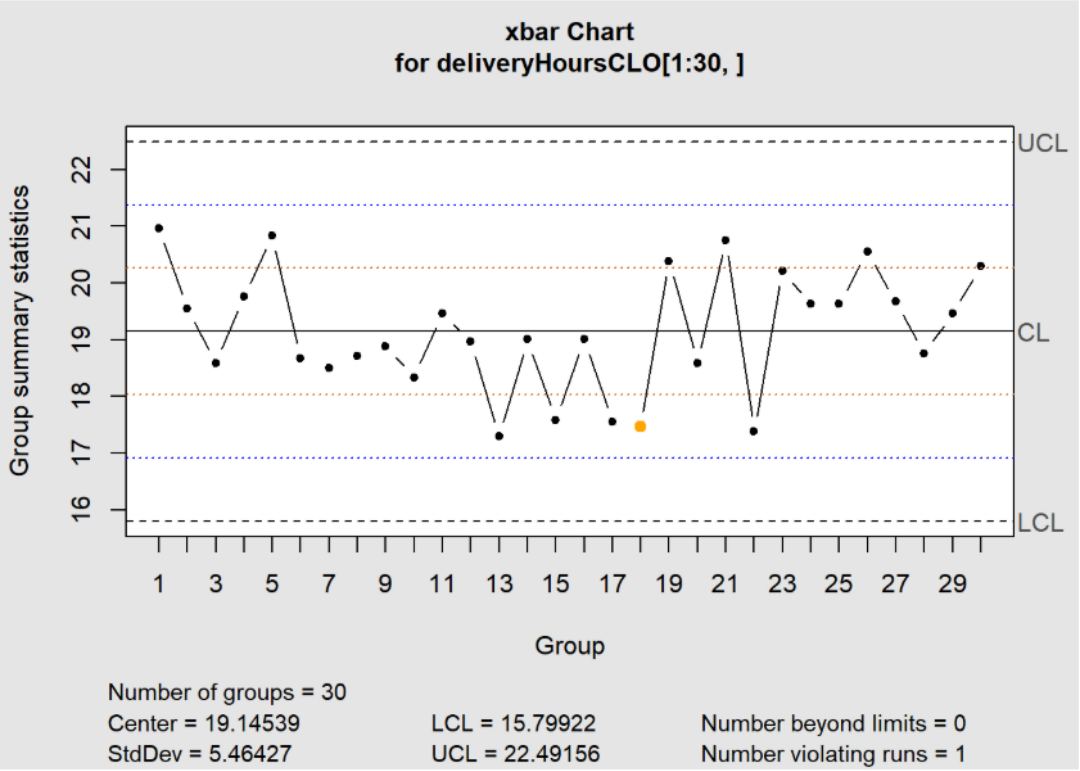

Figure 14: Average deliveryHours of CLOUD SUBSCRIPTION for the first 30 days

Sample 18 is the 7th consecutive point below the centreline. Consequently, this sample is removed, and all parameters are recalculated.

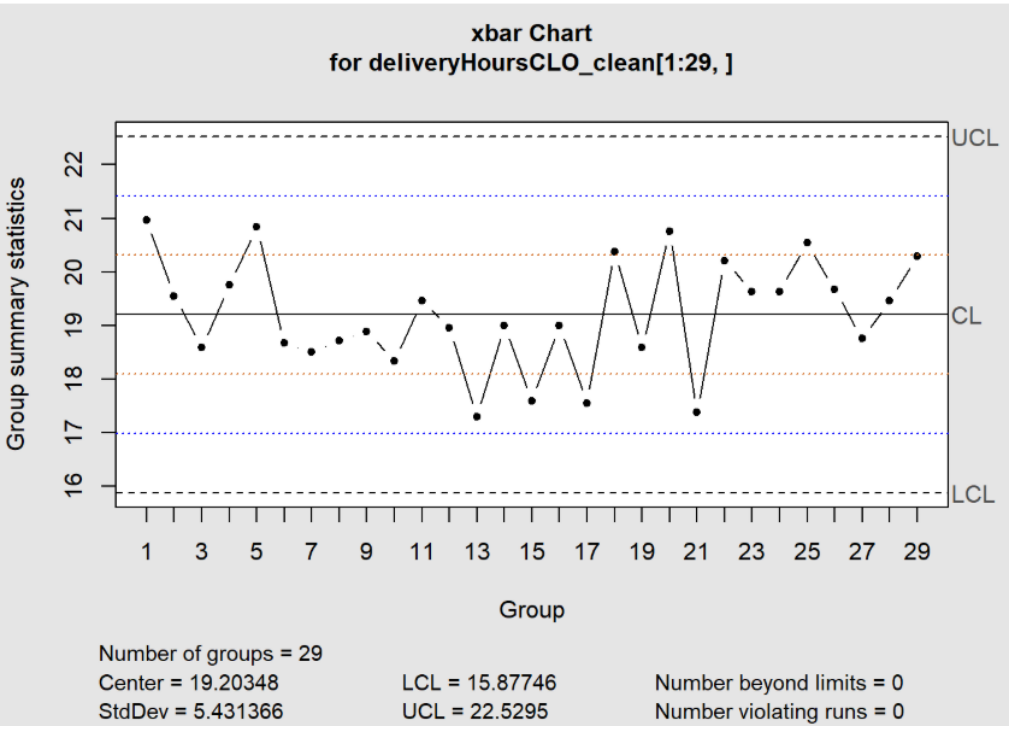

Figure 15: Average deliveryHours of CLOUD SUBSCRIPTION for the first 30 days with out-of-control samples removed

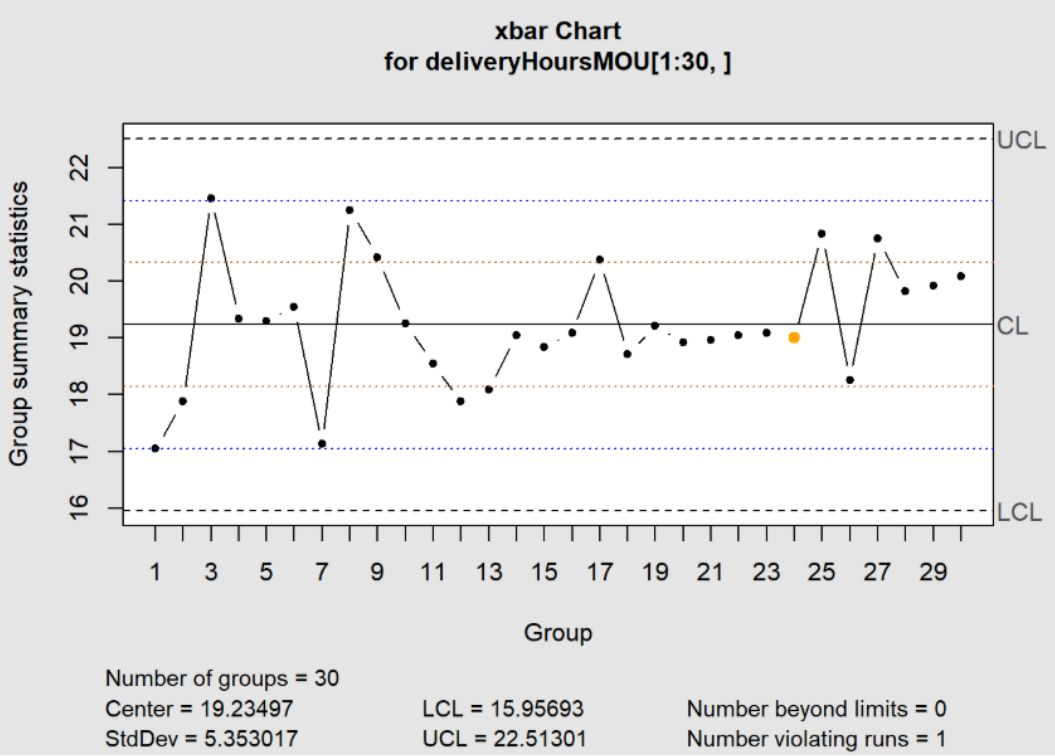

Figure 16: Average deliveryHours of MOUSE for the first 30 days

Sample 24 is the 7th consecutive point below the centreline. Consequently, this sample is removed, and all parameters are recalculated.

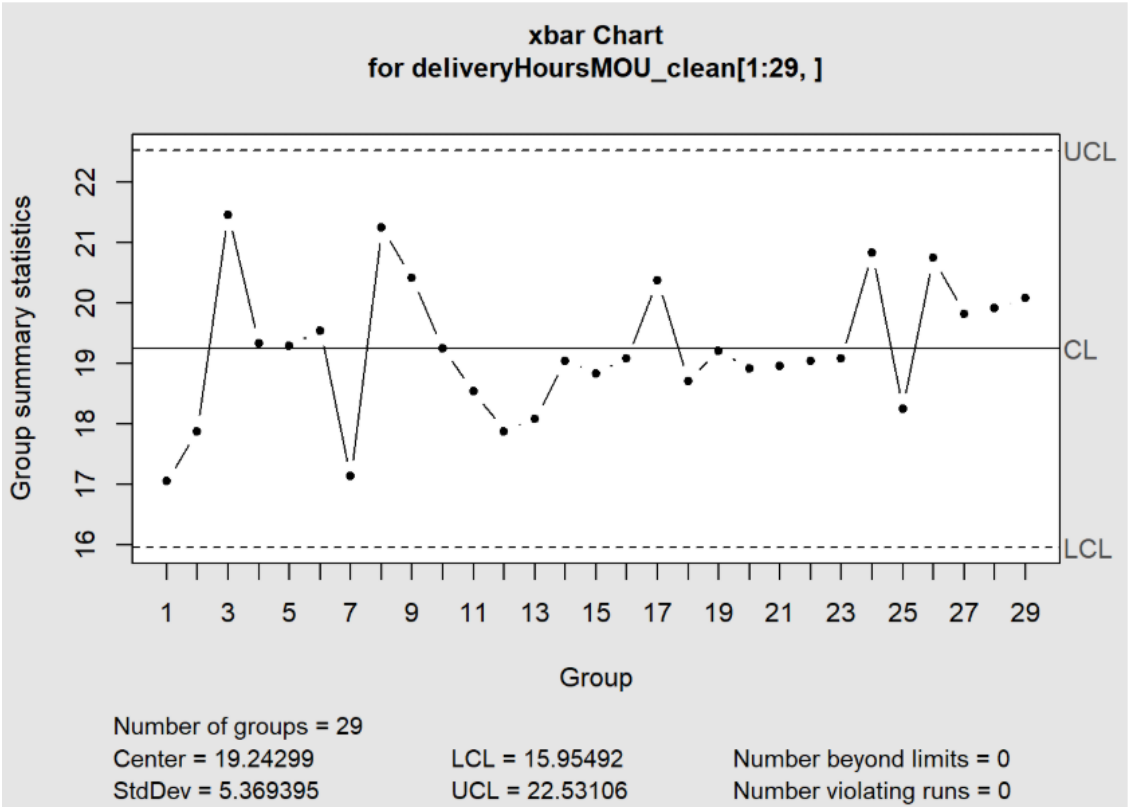

Figure 17: Average deliveryHours of MOUSE for the first 30 days with out-of-control samples removed

2.2. Continue drawing samples for each product type

x̄- and s-charts for each product are expanded and adapted according to the new influx of samples (24-hour delivery time periods). This continues for all samples of delivery time data. Samples 31 onwards are labelled as new data and are used to monitor the process.

For the purposes of succinctness, the most out-of-control and the most in-control x̄-chart are selected for inclusion in the report. This is calculated based on the following two equations.

$$control = \frac{\text{Number beyond limits}}{\text{Number of groups}}$$

$$control = \frac{\text{Number violating runs}}{\text{Number of groups}}$$

Both these calculations reveal that the most out-of-control product type is SOFTWARE, and the most in-control product type is LAPTOP.

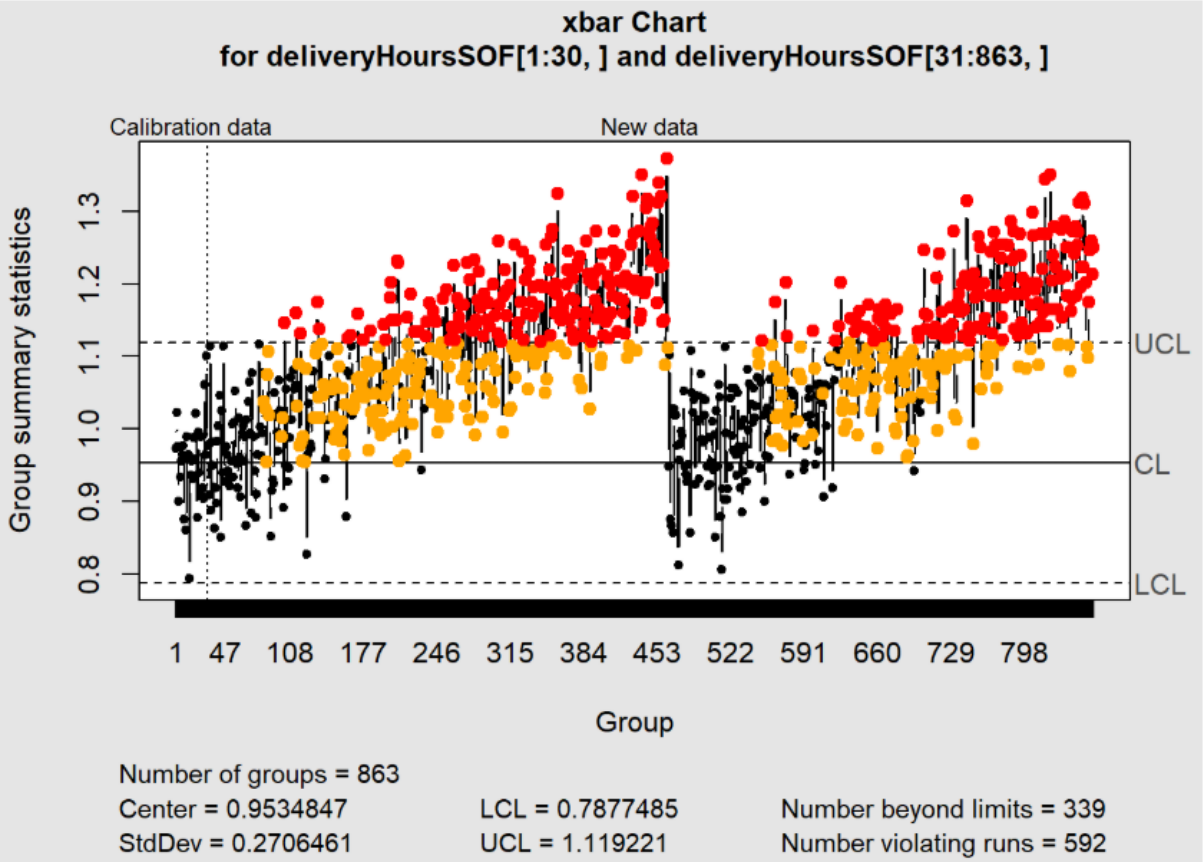

Figure 18: x̄-chart for SOFTWARE with the most out-of-control signals of all product types

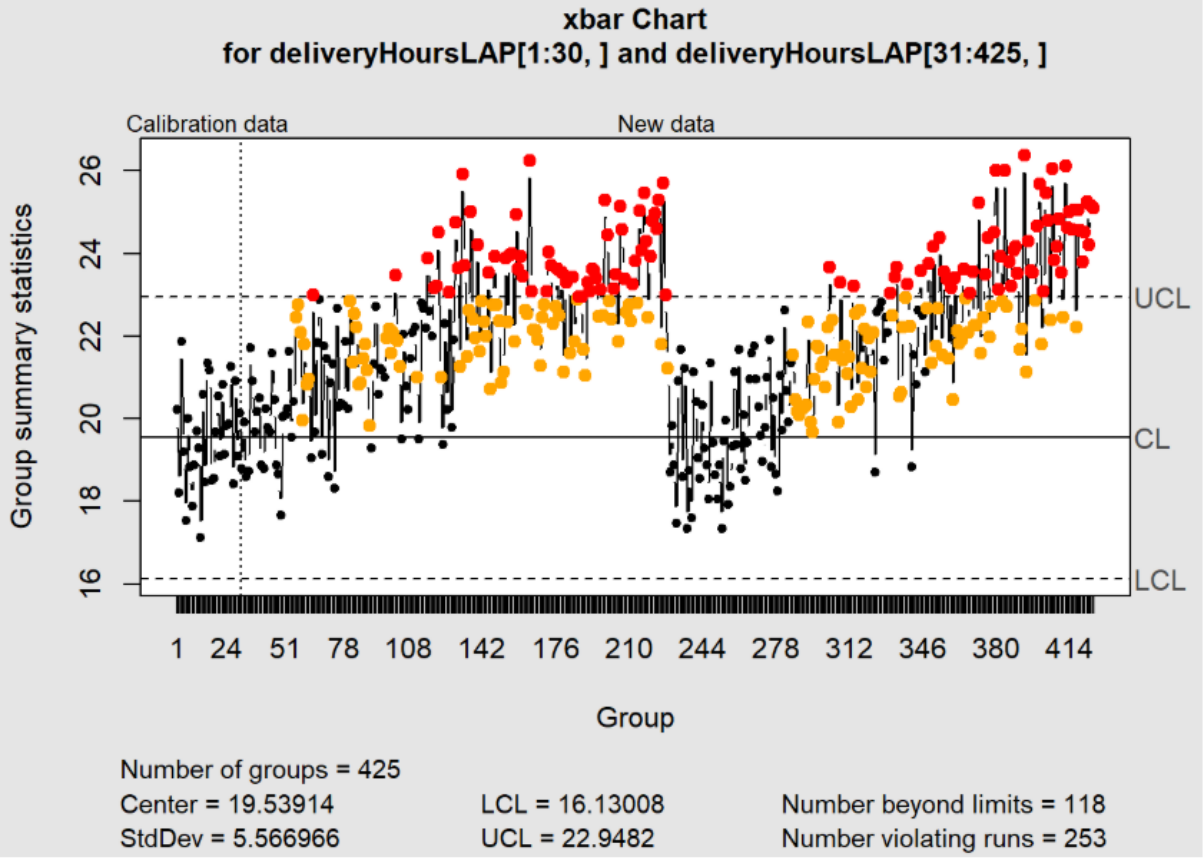

Figure 19: $\bar{x}$-chart for LAPTOP with the most in-control signals of all product types

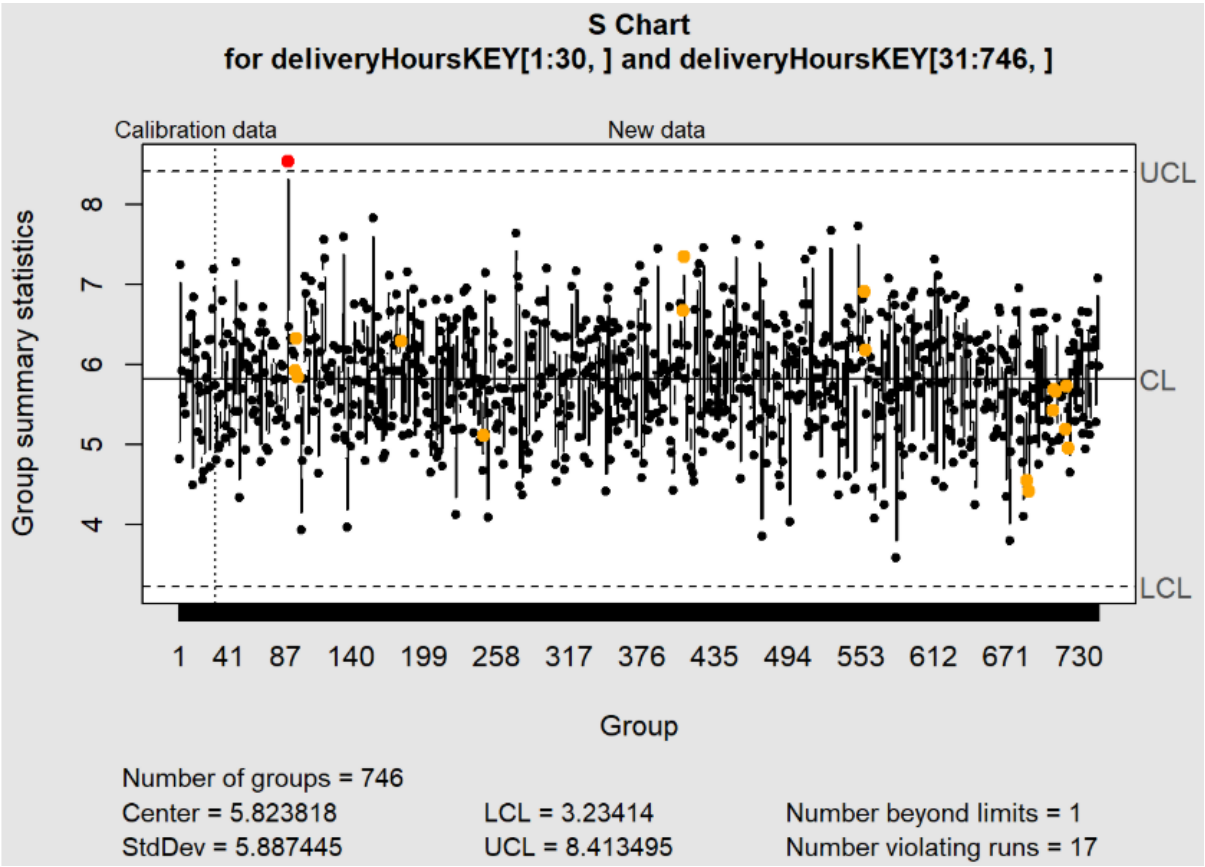

Figure 20: S-chart for KEYBOARD product type

The above charts can be used to identify when each product manager needs to go and adjust or check their process control.

Red dots indicate points that are beyond the upper or lower process control limits. In the event of a red dot, the product manager should investigate immediately. For $\bar{x}$-charts, red dots indicate sample mean delivery times that are out of control. So, the delivery times for a whole day are either too low or too high. Similarly for s-charts, red dots indicate sample variation of delivery times that are out of control. So, the variation of delivery times for that particular day are either too low or too high.

Orange dots indicate points that are within limits but are statistically unusual. These are decided based on the Western Electric Rules (Quality Gurus, 2025). Examples of these include 7 or more consecutive points above or below the centreline, 2 out of 3 points being beyond the 2σ line on the same side of the centreline, or 12 out of 14 points on the same side of the centreline, to name a few. The product manager should initiate an investigation when these signals appear.

2.3. Calculate Process Capability indices

First, subsets are created containing the first 1000 deliveries per product type. Then, the mean and standard deviation of each subset are determined in R. This yields the following results:

Table 2: Mean and standard deviation of deliveryHours for the first 1000 deliveries of each product type

ProductID	**Mean**	**Std Dev**
SOF	0.957675	0.2937719
CLO	19.214	5.9446984
LAP	19.599	5.9341123
KEY	19.265	5.8165704
MON	19.414	5.9945003
MOU	19.3175	5.8275559

Taking LSL = 0 and USL = 32 yields the following capability indices of the delivery processes for each product type:

Table 3: Process capability indices for the delivery processes for each product type

ProductID	**SOF**	**CLO**	**LAP**	**KEY**	**MON**	**MOU**
$C_p = \frac{USL-LSL}{6\sigma}$	C_p = 18.155	C_p = 0.897	C_p = 0.899	C_p = 0.917	C_p = 0.890	C_p = 0.915
$C_{pu} = \frac{USL-\mu}{3\sigma}$	C_{pu}=35.223	C_{pu}=0.717	C_{pu}=0.697	C_{pu}=0.730	C_{pu}=0.700	C_{pu}=0.725
$C_{pl} = \frac{\mu-LSL}{3\sigma}$	C_{pl}=1.087	C_{pl}=1.077	C_{pl}=1.101	C_{pl}=1.104	C_{pl}=1.080	C_{pl}=1.105
C_{pk} = min(C_{pl}, C_{pu})	C_{pk}=1.087	C_{pk}=0.717	C_{pk}=0.697	C_{pk}=0.730	C_{pk}=0.700	C_{pk}=0.725

A C_{pk} of 1 is considered to be marginally or barely capable with higher numbers indicating better process capability. The delivery of SOFTWARE is the only process with a C_{pk} higher than 1 which means that it is the only process that may be capable of meeting the VOC.

2.4. Identify samples that show process control issues

A. One s-sample outside the upper $+3\sigma$ control limits for any product types.

Only keyboards have one s-sample outside the upper 3σ control limit at around sample 90.

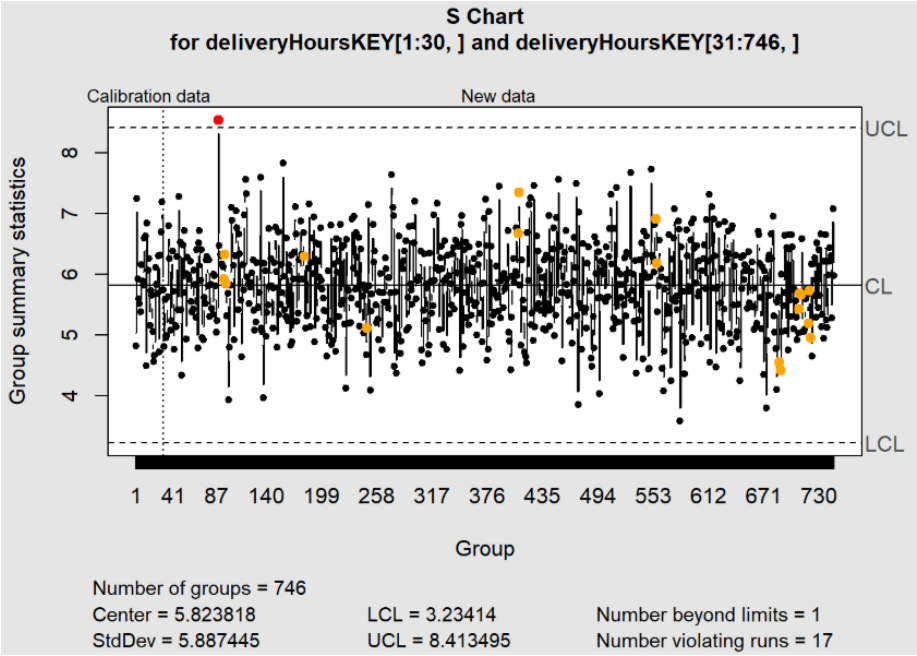

Figure 21: S-chart for deliveryHours of KEYBOARD

B. Find the most consecutive samples of s between the -1σ and $+1\sigma$ control limits for any product types. This signifies good control.

The CLOUD SUBSCRIPTION product has the most consecutive samples of s between the -1σ and $+1\sigma$ control limits with 20 samples being within this range from sample 34 to sample 53.

C. 4 consecutive $\bar{x}$-samples outside of the upper, second control limits for all product types.

This table shows, for each product type, the number of runs within its samples that contain 4 or more consecutive $\bar{x}$-samples outside the upper, 2nd control limits for each product type. It lists the number of such runs as well as the indices of the first value in each run for the first 3 and last 3 of these occurrences.

Table 4: Number and locations of runs of 4 or more $\bar{x}$-samples outside the upper, 2nd control limits for each product type

ProductID	Number of runs	Starting indices of first 3 runs	Starting indices of last 3 runs
SOF	31	162, 201, 206	751, 759, 773
CLO	18	125, 165, 174	550, 556, 627
LAP	14	114, 136, 142	374, 395, 402
KEY	22	177, 185, 200	687, 726, 738
MON	18	145, 178, 193	566, 575, 615
MOU	21	199, 221, 248	764, 776, 809

3. Risk Analysis and Data Correction

3.1. Type I error

A. One s-sample outside the upper $+3\sigma$ control limits for any product types.

For an in-control process, the probability of a single observation exceeding $+3\sigma$ has the following one-sided probability:

$$\alpha_{\mathrm{A}} = P(Z > 3) = 1 - P(Z \leq 3) \approx 0.0013 \ = \ 0.13\% \, per \, sample$$

With 746 samples of days of delivery times for KEYBOARD, the probability of getting at least one manufacturer's error on the s-chart is:

$$1 - (1 - \alpha_A)^{746} \approx 0.621 \ (62.1\%)$$

This means that seeing approximately 1 point beyond control limits on a chart that contains a high number of samples is not unlikely even if the process is generally very well in control.

B. Find the most consecutive samples of s between the -1σ and $+1\sigma$ control limits for any product types. This signifies good control.

Within the $\pm 1\sigma$ control limits, the probability of achieving a run of length m is:

$$\alpha_B^m = P(Z \leq 1)^m = (0.8413)^m$$

The probability of achieving a run containing 20 samples in a row (as occurs in the CLOUD SUBSCRIPTION s-chart) is:

$$\alpha_B^{20} = (0.8413)^{20} = 0.03155 \approx 3.16\%$$

The s-chart for CLOUD SUBSCRIPTION contains 649 samples. Over the 649 samples, there are 649-20+1=630 possible 20-point windows. The chance of seeing at least one such 20-point streak is:

$$1 - (1 - \alpha_B^{20})^{630} = 1 - (1 - 0.03155)^{630} \approx 1 = 100\%$$

This means that a run of signals in good control is almost certain if the dataset contains a high enough number of samples. So this measure is not, on its own, a good indicator of exceptional control.

C. 4 consecutive $\bar{x}$-samples outside of the upper, second control limits for all product types.

The probability of one point being above the $+2\sigma$ control limit is:

$$\alpha_C = P(Z > 2) = 1 - P(Z \leq 2) = 1 - 0.9772 = 0.0228$$

The probability of four points in a row is:

$$\alpha_C^4 = 0.0228^4 = 2.702 \times 10^{-7}$$

The 6 product types have an average number of samples roughly equal to 694. So, the number of possible 4-point windows in such a product stream is 694-4+1=691. The probability of at least one manufacturer's error occurring is:

$$1 - (1 - \alpha_C^4)^{691} = 1 - (1 - 2.702 \times 10^{-7})^{691} = 1.867 \times 10^{-4}$$

This value makes sense intuitively because Section 2.4. showed that it is quite common for event C to occur. This makes it more likely for the product manager to stop the process (because they suspect an error), but for there to, in fact, be nothing wrong.

3.2. Type II error

In-control CL = 25.05, LCL = 25.011, UCL = 25.089.

True process mean = μ = 25.028.

Standard deviation of the mean = s = 0.017.

So, the probability of making a consumer's error for the bottle filling process is:

$$\beta = P(25.011 \leq \bar{X} \leq 25.089) = P\left(Z \leq \frac{25.089 - 25.028}{0.017}\right) - P\left(Z \leq \frac{25.011 - 25.028}{0.017}\right)$$
$$= P(Z \leq 3.588) - P(Z \leq -1) = 0.9998 - 0.1587 = 0.8411 = 84.11\%$$

3.3. Re-do initial data analysis with corrected data

Much of the analysis performed in Section 1 is unchanged by the updated product and product (head office) data as most of the initial analysis is performed on sales and customer data. What follows is analysis that needs to be updated or added after data correction has taken place.

Firstly, despite the corrections in these features, the summary statistics for the 'SellingPrice' and 'Markup' features in the products_headoff dataset remained unchanged.

Comparing selling price per feature yields the following visualisation. Showing significant variation between different product types.

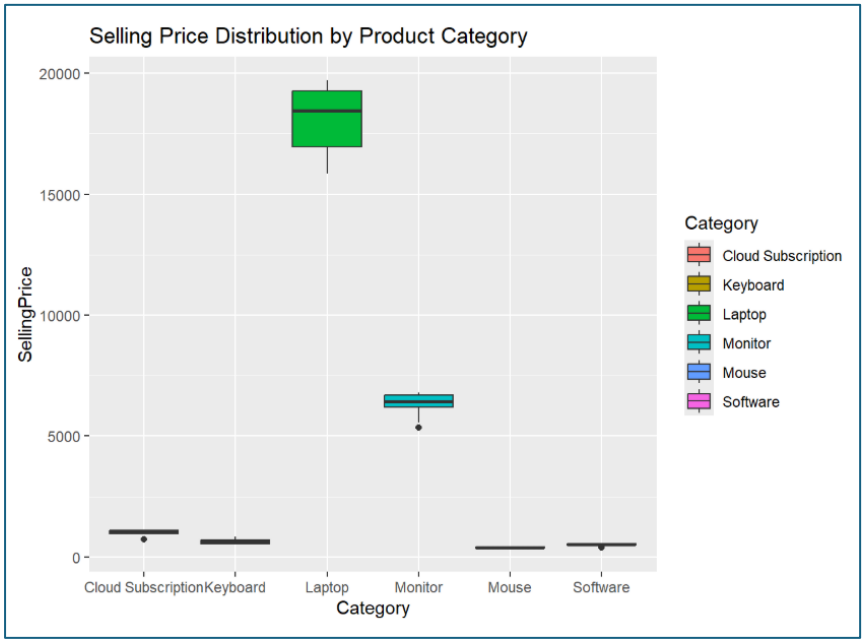

Figure 22: Box plots showing distributions of selling price by product category

Despite this significant variation in selling price, the correlation between selling price and markup is equal to -0.07936097. This makes sense from a business perspective as customers are more price-sensitive for higher priced items. So, tighter margins are required to maintain a competitive advantage.

Proft per item is calculated by multiplying selling price with percentage markup. This is then multiplied by the quantity of those items sold and summed for all sales of that item in 2023.

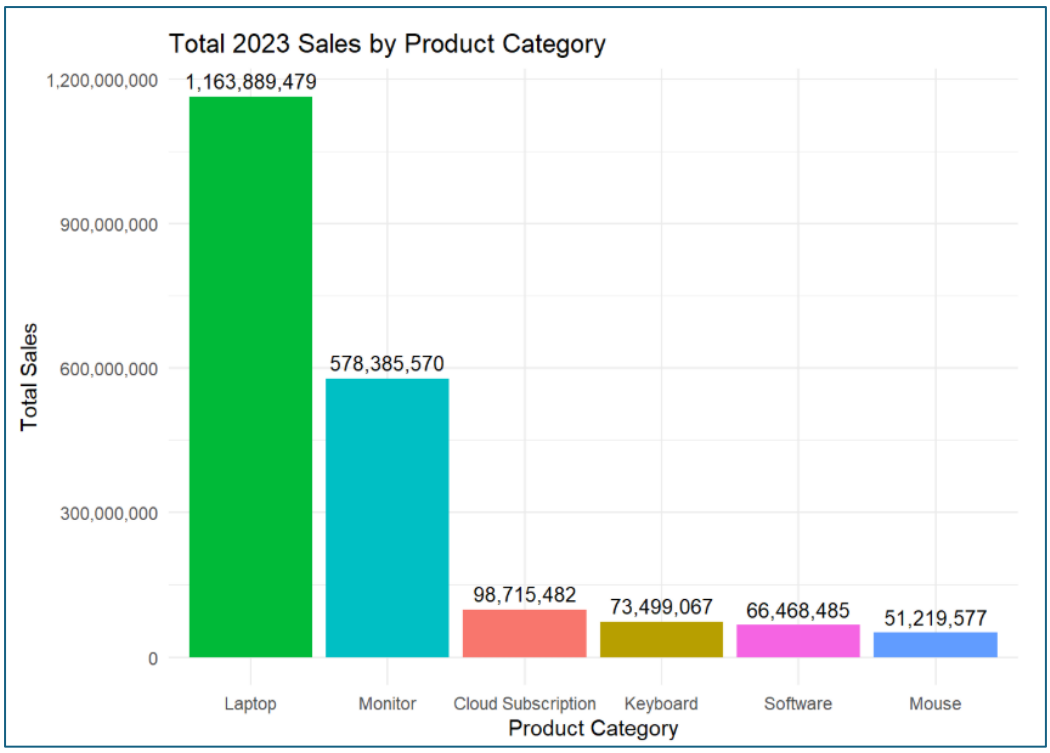

Figure 23: Bar plot showing total sales by product category for 2023

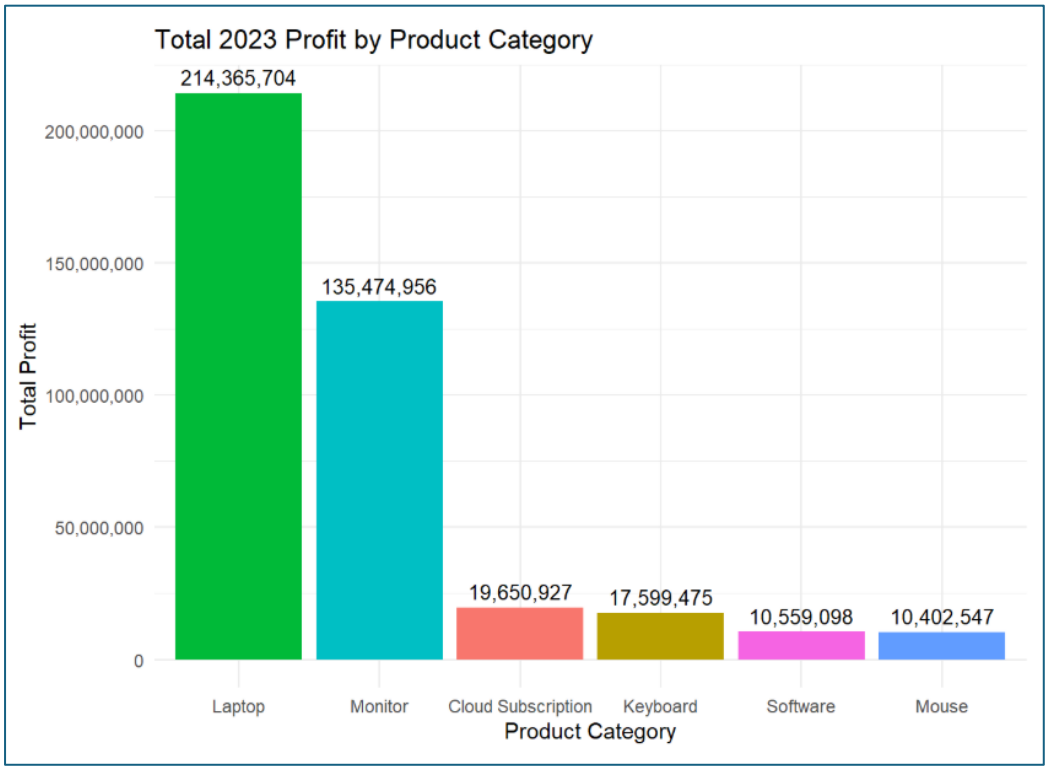

Figure 24: Bar plot showing profit by product category for 2023

Figures 23 and 24 indicate that most of the sales revenue and profit comes from the higher priced items. The management team should consider increasing the margin on the lower priced items (software and mice) which contribute the highest sales volume. This may allow significant profit increases without decreasing customer satisfaction due to perceived value.

4. Optimisation

Two datasets are given: timeToServe (TTS) and timeToServe2 (TTS2). From these two datasets, an optimal number of baristas per weekday is to be selected that would maximise profit for the company.

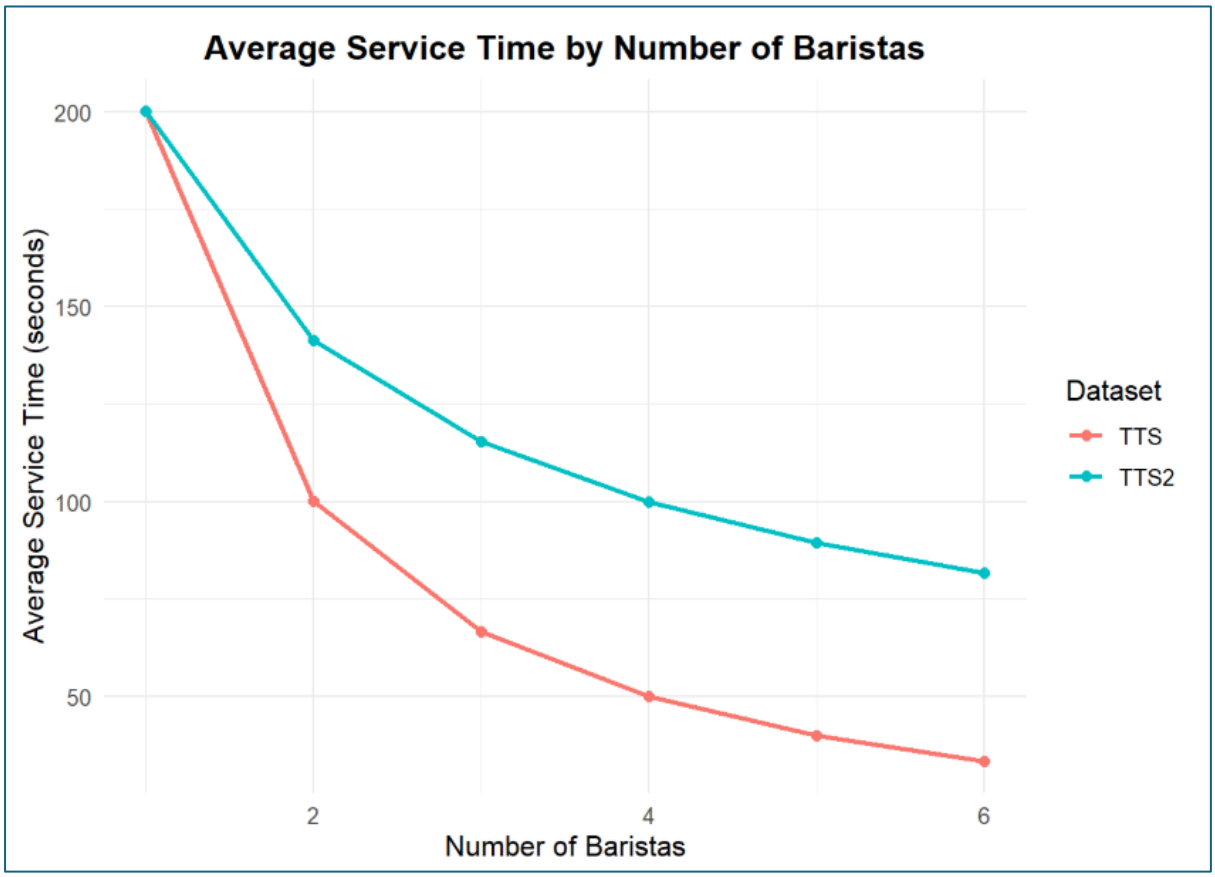

Figure 25: Line graph showing average service time by number of baristas for both datasets.

In Figure 25, both lines clearly show that average service time decreases as number of baristas increases. The rate of improvement slows down quite significantly after about 4 baristas, indicating that this may be an optimal number of employees. The average service time for TTS2 is significantly higher for the same number of baristas when compared to TTS. This suggests that TTS2 is a generally slower, less efficient coffee shop. This means that the coffee shop represented by TTS2 would need more baristas in order to compete with the operational efficiency of the TTS coffee shop even though the TTS coffee shop incurs lower salary expenses because they can employ fewer baristas.

An optimal number of 4 baristas is arbitrarily selected for both coffee shops. This creates a threshold acceptable service time of around 50 seconds for TTS and 100 seconds for TTS2. Service reliability is the percentage of customers who can reliably expect their service time to be below this threshold.

Table 5: Table showing service reliability for each coffee shop

Coffee shop	Threshold	Service reliability
TTS	≤ 50 seconds	84.77%
TTS2	≤ 100 seconds	76.09%

These results are acceptable for the level of operational efficiency found in each coffee shop.

5. DOE and MANOVA/ANOVA

Research question: "Is there a significant difference in delivery times between years or months?"

Null hypothesis H_o: The delivery hours do not differ across years or months.

Alternative hypothesis H_a: The delivery hours do differ across years or months.

Input variables: orderYear and orderMonth

Response variable: deliveryHours

```
                                  Df  Sum Sq Mean Sq F value Pr(>F)
factor(orderYear)                  1     139     139   1.415  0.234
factor(orderMonth)                11  170247   15477 157.946 <2e-16 ***
factor(orderYear):factor(orderMonth)   11     752      68   0.697  0.742
Residuals                      99976 9796561      98
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 26: ANOVA summary for delivery times across months and years

The orderYear factor has p=0.234. This is quite a large p-value, indicating that the year factor is not significant and that mean delivery time does not differ significantly between years in this dataset.

The orderMonth factor has p<0.001. This is a very small p-value, indicating that the month factor is highly significant and that mean delivery time differs significantly by month.

The year*month interaction factor has a very large p-value of 0.742, indicating that the pattern of month-to-month variation is similar across years.

The conclusion is that the null hypothesis is rejected in favour of the alternative hypothesis, as mean sales clearly do differ across months.

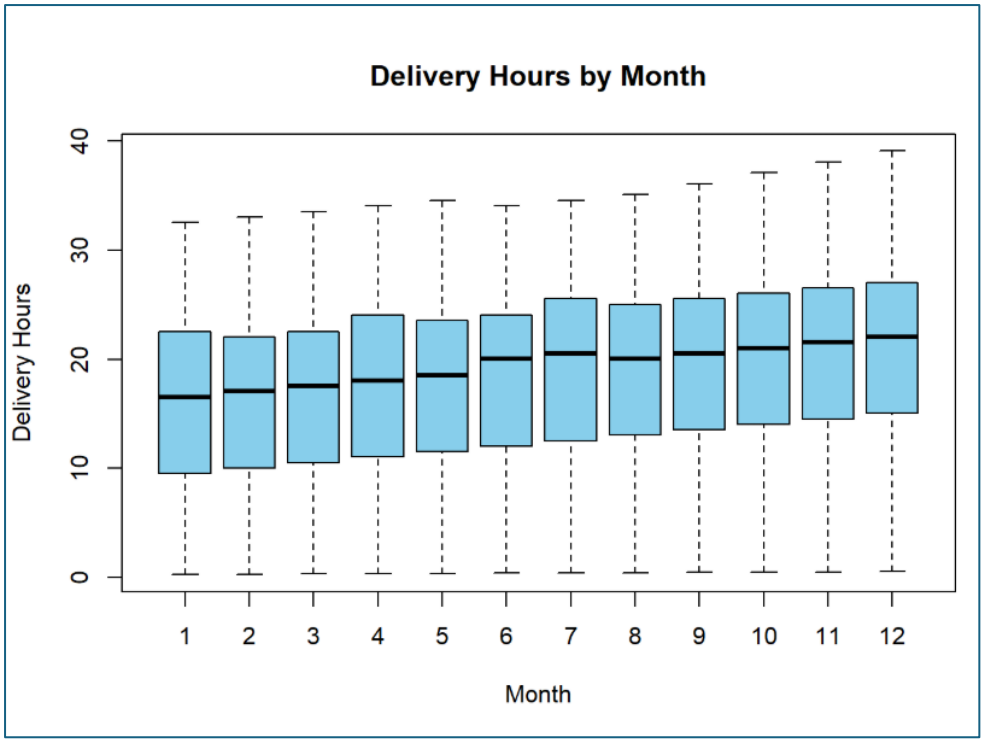

Figure 27: Box-and-whisker plot showing variation of delivery hours across months

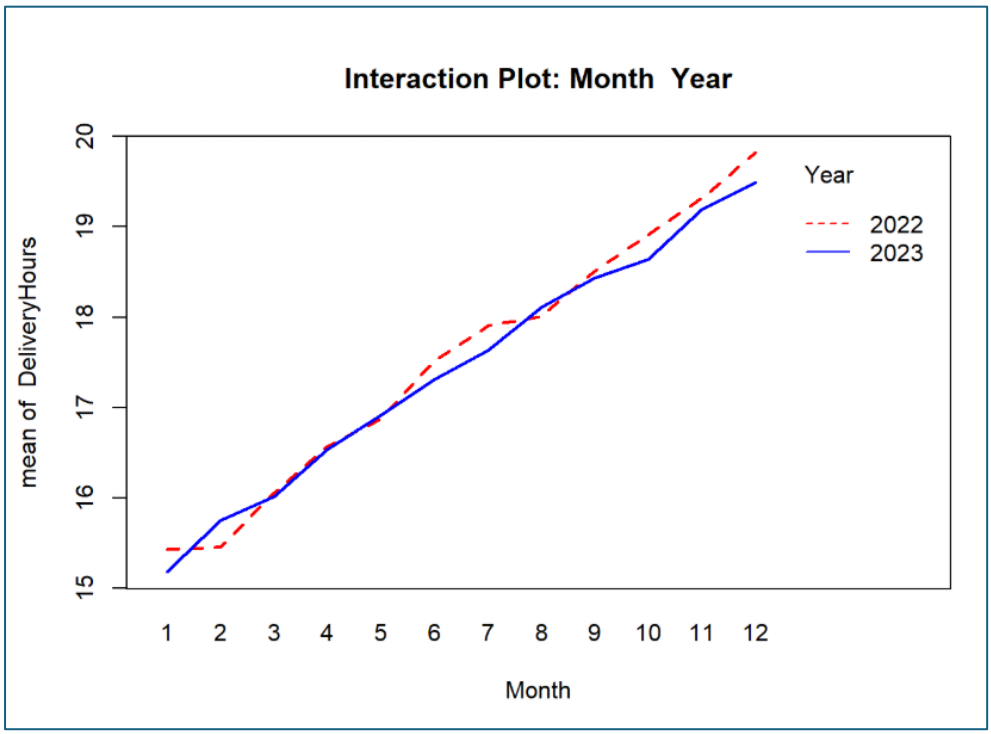

Figure 28: Interaction plot showing trend in delivery hours over months for 2022 and 2023

The interaction plot in Figure 28 clearly shows a strong correlation between changes in month and changes in mean delivery hours, while also showing the insignificance of the year factor.

6. Reliability of Service

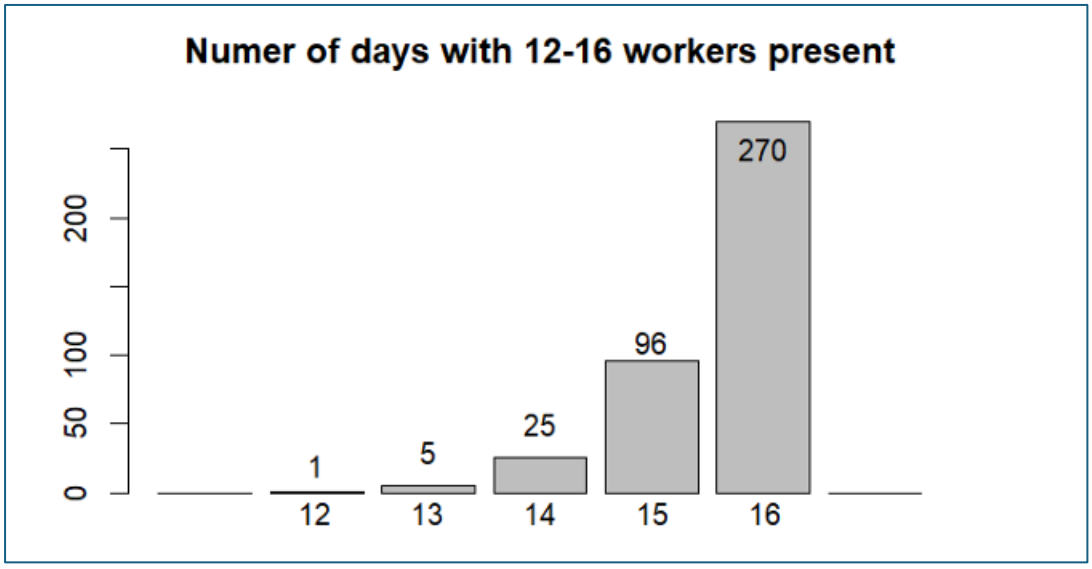

Figure 29: Car rental agency worker presence over 397 days

6.1. Days per year of reliable service

Service is reliable if there are 15 or more people on duty. This occurs for 96 + 270 = 366 days out of 397. The estimated reliability rate is $p = \frac{366}{397} = 0.9219$. So, the expected reliable service days in a 365-day year is $365 \times p = 365 \times 0.9219 \approx 337$ days per year.

6.2. Profit Optimisation using binomial model

There are 397 days of data. Service problems occur when there are fewer than 15 workers on duty. A problem day costs R20 000 in lost sales. Each extra staff member costs R25 000 per month = R300 000 per year. The goal is to maximise profit.

$$Profit = Revenue - (Lost\ sales + Staff\ costs)$$

Table 6: Table showing current service reliability of car rental agency

Workers on duty	Days	Service problems?
12	1	Yes
13	5	Yes
14	25	Yes
15	96	No
16	270	No

This means that there are 366 reliable days out of 397. Each day is considered as a Bernoulli trial with a reliable day represented as a successful trial with a probability of p and a problem day represented as a failure trial with a probability of (1-p). In 365 days, the expected number of reliable days is 336 to 337 and expected number of problem days 28 to 29 per year.

Table 7: Table showing change in reliability with each extra staff member added beyond 15

Extra staff	Days with problems	Reliable days	Reliability
0	31	366	92.19%
1	6	391	98.49%
2	1	396	99.75%
3	0	397	100%

These reliability values are then used to calculate the number of problem days that the car rental agency would have to pay losses for each year. The annual cost of adding the specified number of staff members is added to this and the total annual cost is calculated.

Table 8: Table showing total costs associated with hiring each additional worker

Extra staff	Problem days ((1-p) × 365)	Loss (R) (Problem days × 20 000)	Staff cost (R) (Extra staff × 300 000)	Total cost (Loss + Staff)
0	28.5	570 000	0	570 000
1	5.5	110 000	300 000	410 000
2	0.9	18 000	600 000	618 000
3	0	0	900 000	900 000

Table 8 clearly shows that hiring one more worker and, therefore, having 16 workers in total yields the optimal minimal total cost for the car rental agency.

Conclusion

Through the sequential analyses presented in this report, the objectives outlined in the ECSA GA4 project brief were successfully achieved. The descriptive statistics provided an informed understanding of customer behaviour, sales patterns, and operational data integrity. Statistical Process Control (SPC) identified process variations and guided corrective actions, while the capability indices assessed process performance relative to specification limits. The Type I and II error evaluations reinforced statistical decision-making reliability.

The optimisation and experimental design sections applied these principles to improve system performance. The ANOVA results revealed significant monthly difference in delivery time, confirming the influence of temporal factors on service outcomes. Finally, the reliability model for the car rental case quantified the trade-off between staffing levels and profitability, demonstrating that a single additional employee beyond the 15-worker-threshold yielded the highest expected profit, demonstrating the importance of balance between operational reliability and cost efficiency.

Collectively, these findings illustrate the integration of statistical reasoning, data-driven optimisation, and practical engineering insight. The report demonstrates the graduate's ability to apply analytical methods to real industrial challenges, satisfying the requirements for ECSA Graduate Attribute 4 and reinforcing the role of data analytics in achieving continuous process improvement within engineering systems.

References

Quality Gurus, 2025. *Nelson Rules (and Western Electric Rules) for Control Charts.* [Online] Available at: https://www.qualitygurus.com/nelson-rules-and-western-electric-rules-for-control-charts/
[Accessed 23 October 2025].