Stellenbosch
UNIVERSITY
IYUNIVESITHI
UNIVERSITEIT

Quality Assurance 344
Data Analysis Report

Written By: Chris Rider

Student Number: 24868892

Contents

Part 1.2 Descriptive Statistics

An analysis of transactional data provided for the years 2022 and 2023. The data contains detailed information on customer demographics, sales transactions, and operational performance metrics. All analyses were conducted in R Studio using tidyverse for data wrangling and ggplot2 visualisations.

Average Income of Customers Per City

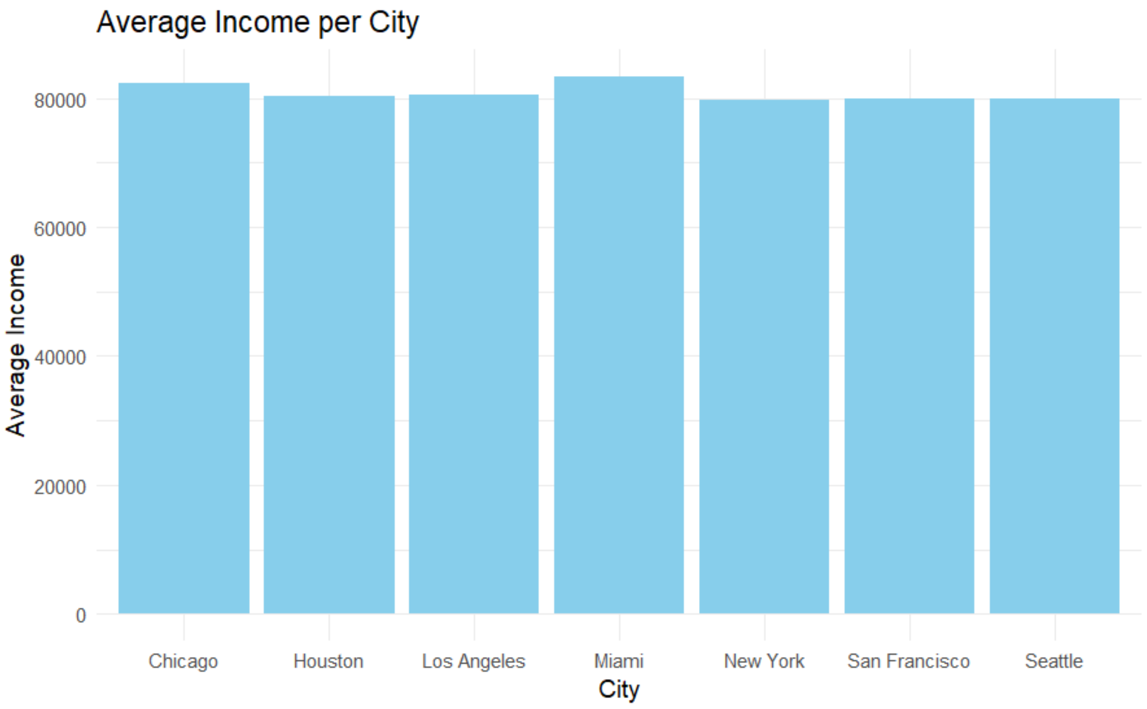


Figure 1 - Average Income per City

Figure 1 illustrates the mean income of customers across the 7 cities in the dataset. The average income is evenly distributed, with Miami and Chicago being marginally higher than the rest. There are no significant outliers, indicating a uniform income distribution among customers regardless of Location.

Number of Sales by Income Group

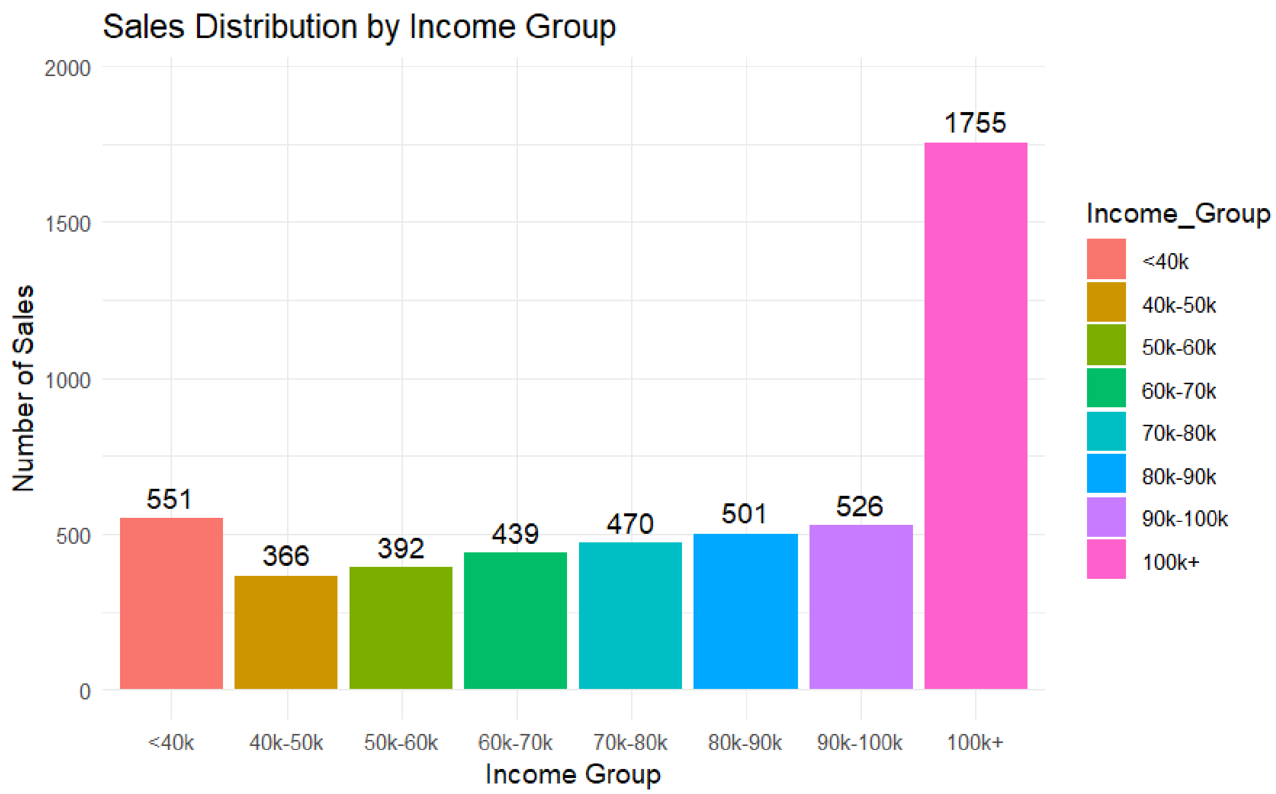


Figure 2 - Number of Sales by Income Group

Customer income ranges from R5 000 to R140 000, with a mean of approximately R80 800 and a median of R85 000.

Sales are distributed unevenly among different income groups, displaying a pronounced right-skew. Most of the sales originate from customers whose earnings exceed R100 000, while those in the lower and middle-income ranges provide progressively lesser sales. This suggests that customers with higher incomes constitute the main market segment.

Age's Relationship to Income

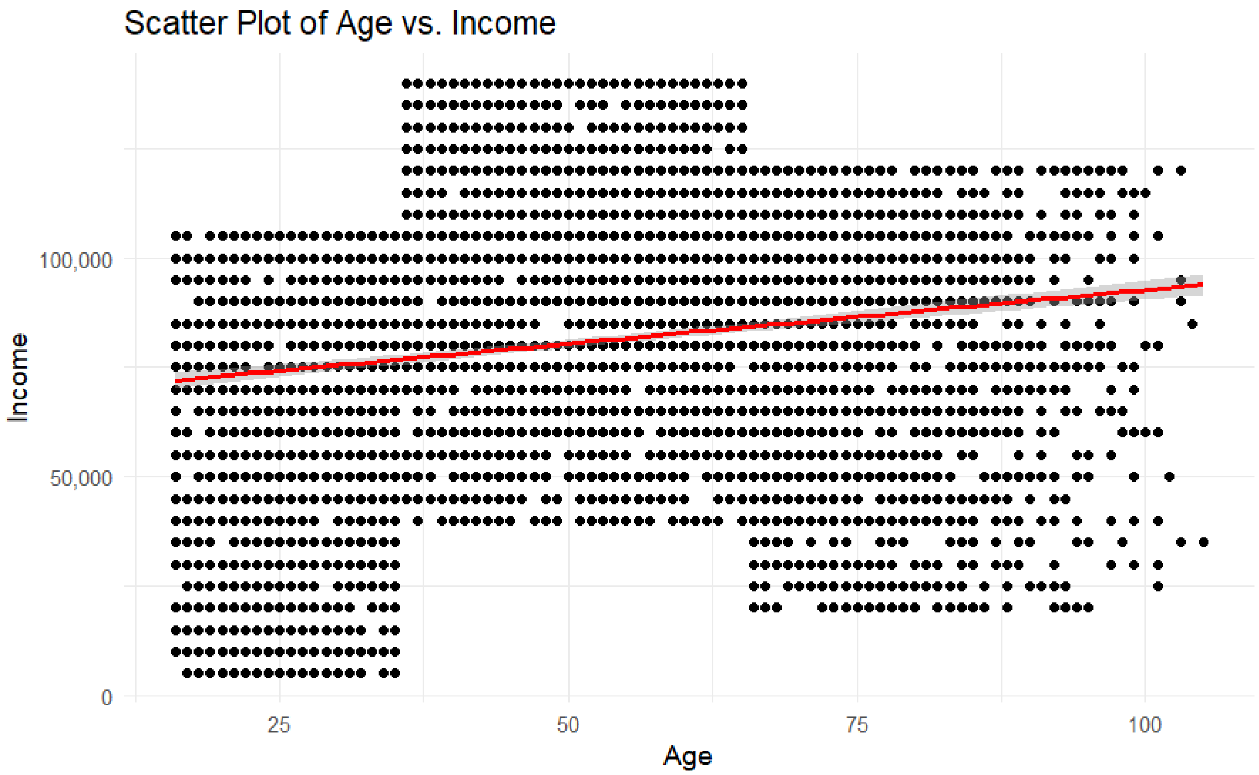


Figure 3 - Scatter Plot of Age vs Income

The scatter plot illustrates a slight yet observable positive correlation between age and income, indicating that earnings tend to rise as individuals grow older. Between approximately ages 35 and 60, income increases more significantly, highlighting a phase of career development and peak earning capacity. After this age range, the upward trend starts to flatten, implying that income growth decelerates post-retirement. However, the broad distribution of data points across all age categories reflects considerable variation, suggesting that while age impacts income, it is not the only factor at play.

Number of Customers by Gender

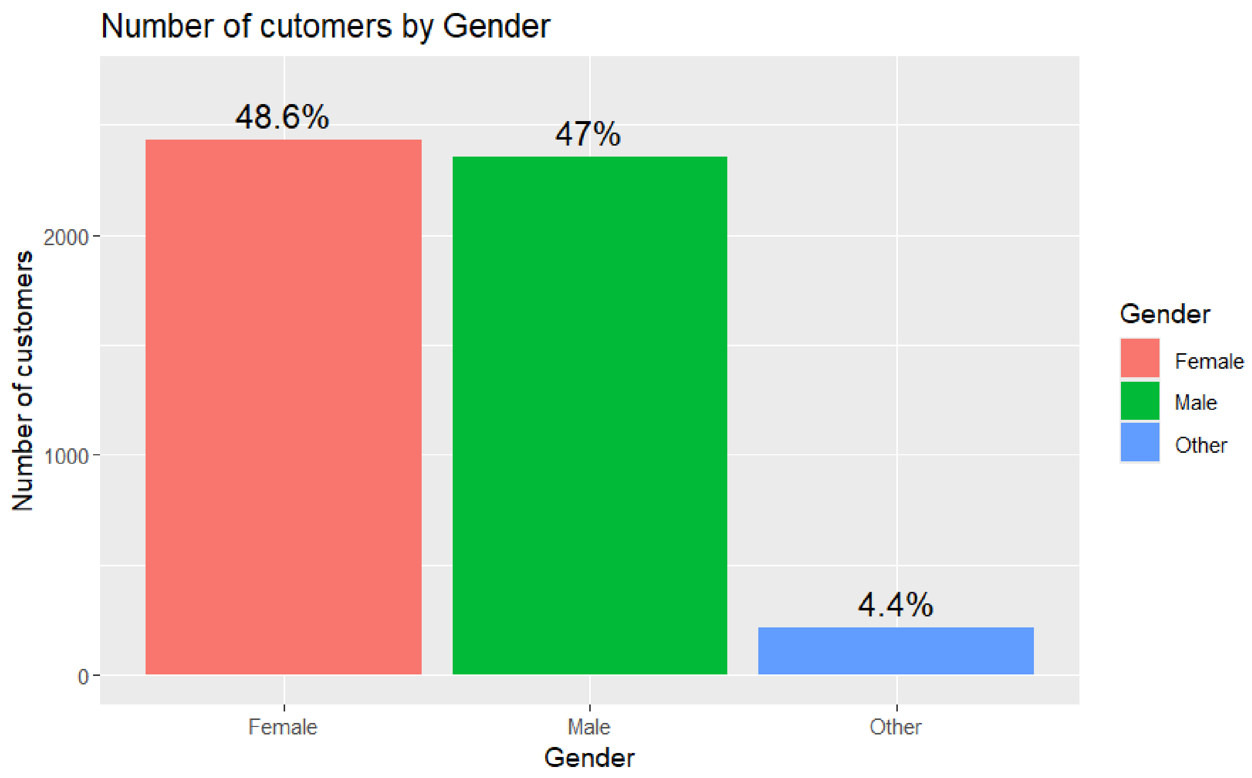


Figure 4 - Number of Customers by Gender

The customer base is almost evenly split between females (48.6%) and males (47%), with a small proportion (4.4%) identifying as other. This balanced gender distribution suggests that the company’s products appeal broadly across genders, without any major skew toward a particular group.

Genders' Relationship to Purchasing Power

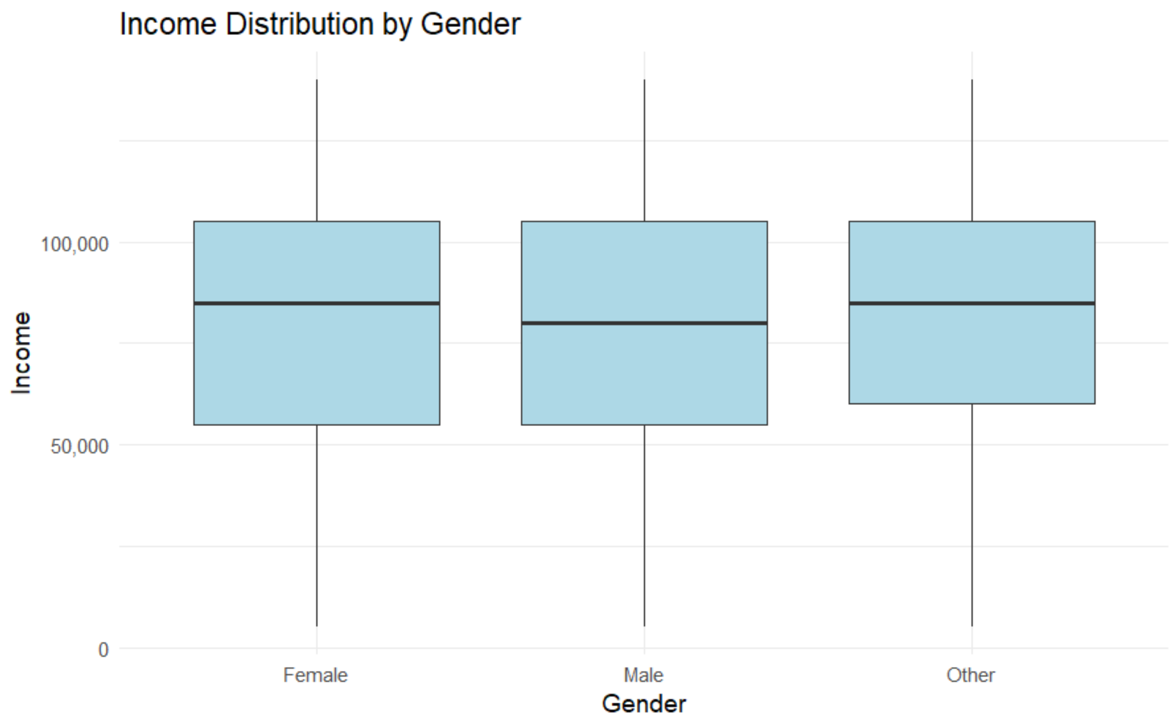
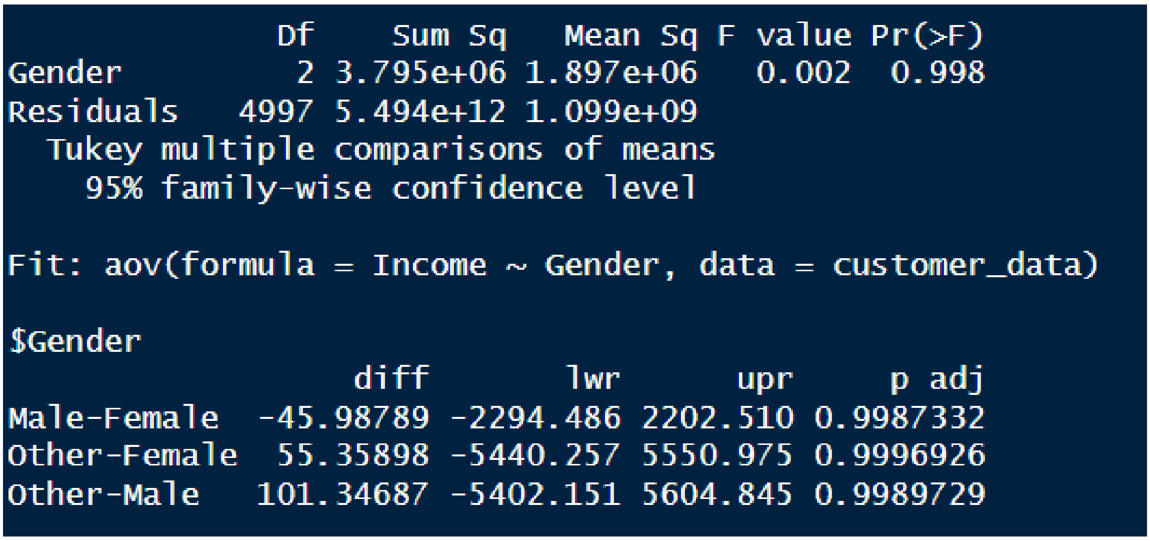


Figure 5 - Income Distribution by Gender



```
                Df    Sum Sq   Mean Sq F value Pr(>F)
Gender           2 3.795e+06 1.897e+06   0.002  0.998
Residuals     4997 5.494e+12 1.099e+09
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Income ~ Gender, data = customer_data)

$Gender
                    diff       lwr      upr     p adj
Male-Female    -45.98789 -2294.486 2202.510 0.9987332
Other-Female    55.35898 -5440.257 5550.975 0.9996926
Other-Male     101.34687 -5402.151 5604.845 0.9989729
```

The boxplot indicates that income levels for males, females, and other gender identifications are quite alike, featuring overlapping interquartile ranges and similar median values. The one-way ANOVA analysis aligns with this visual finding, yielding an F-value of 0.002 and a p-value of 0.998, significantly exceeding the 0.05 significance level. This suggests that there is no statistically meaningful difference in average income between the genders. The Tukey post-hoc analysis further verifies that none of the pairwise comparisons are significant.

Sales Distribution by Age Group

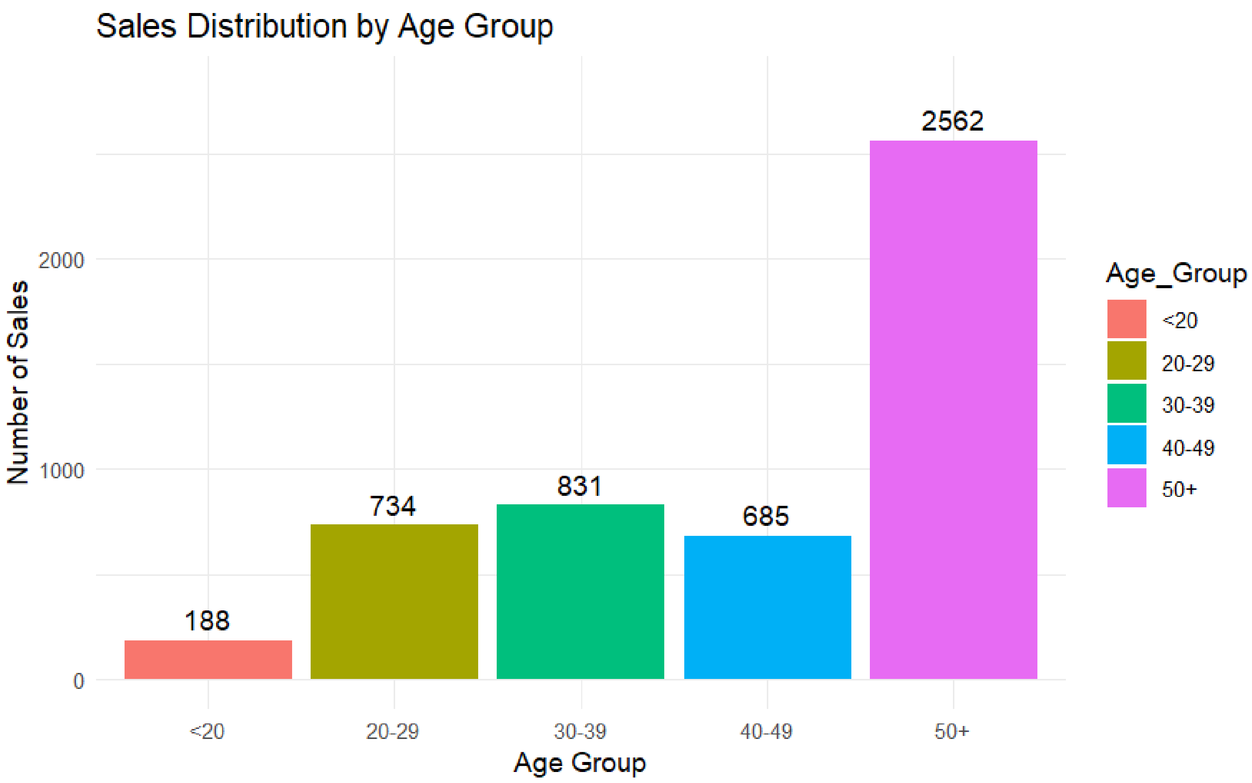
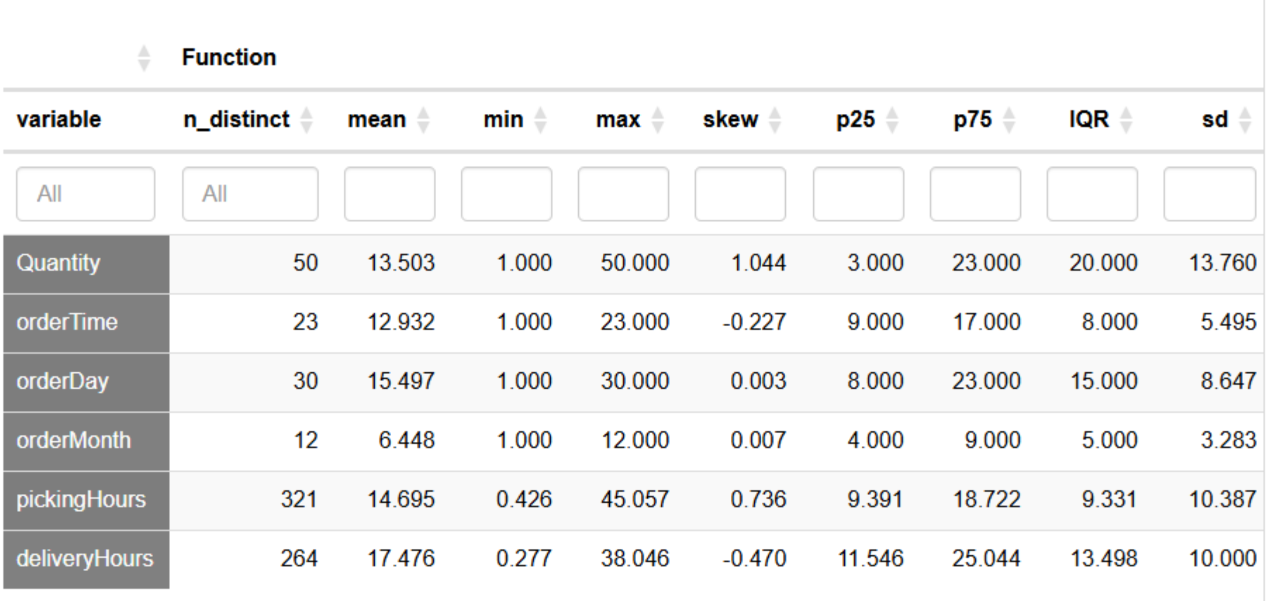


Figure 6 - Sales Distribution by Age

Sales are predominantly focused on customers aged 50 and older, who represent the majority of total transactions. The 20-49 age groups show moderate sales, whereas customers under 20 make a minimal contribution to overall sales figures. The summary statistics reveal an average age of 51.6 years, with most customers ranging between 33 and 68 years old. This indicates that the primary customer base of the company is within the mature and older segments, suggesting that purchasing activity tends to rise with age and stabilises around middle adulthood and beyond.

Basic Data Exploration of Sales Data



	Function								
variable	n_distinct	mean	min	max	skew	p25	p75	IQR	sd
Quantity	50	13.503	1.000	50.000	1.044	3.000	23.000	20.000	13.760
orderTime	23	12.932	1.000	23.000	-0.227	9.000	17.000	8.000	5.495
orderDay	30	15.497	1.000	30.000	0.003	8.000	23.000	15.000	8.647
orderMonth	12	6.448	1.000	12.000	0.007	4.000	9.000	5.000	3.283
pickingHours	321	14.695	0.426	45.057	0.736	9.391	18.722	9.331	10.387
deliveryHours	264	17.476	0.277	38.046	-0.470	11.546	25.044	13.498	10.000

Table 1 - Data Exploration of Sales Data

The summary of the data reveals that order quantities differ significantly, with an average of 13.5 units and a moderate right skew (1.04), suggesting that while the majority of orders are small, larger orders elevate the mean. Order placements are fairly evenly distributed throughout the day, with a slight concentration occurring between the hours of 9 and 17. Both the days and months for orders show uniform distributions, indicating a consistent level of demand throughout the calendar year. The average picking and delivery times are 14.7 and 17.5 hours, respectively, both displaying moderate variability without strong skewness. These findings reflect a stable and balanced operational process, with no notable irregularities in timing or quantity within the data set.

Quantity Sold by Hour of the Day

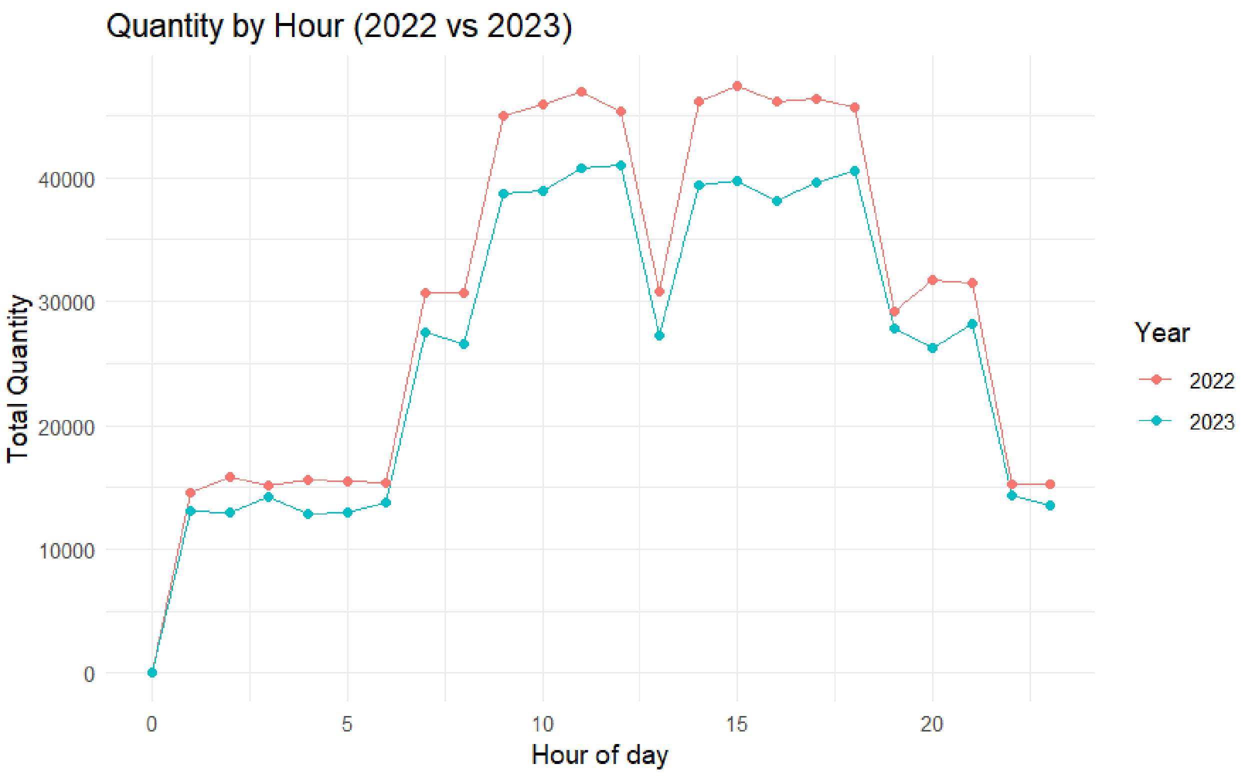


Figure 7 - Quantity Sold by Hour

The line graph illustrates a similar pattern of hourly sales for both 2022 and 2023. Sales volume experiences a significant increase from the early morning hours, reaching a high point between 9:00 and 14:00, followed by a smaller peak in the afternoon around 17:00, then gradually declining as the evening approaches. Sales figures in 2022 are generally a bit higher than those in 2023, indicating slightly stronger demand or greater throughput during that year. In general, sales activity seems to be focused during typical working hours.

Quantity Sold by Day of the Week

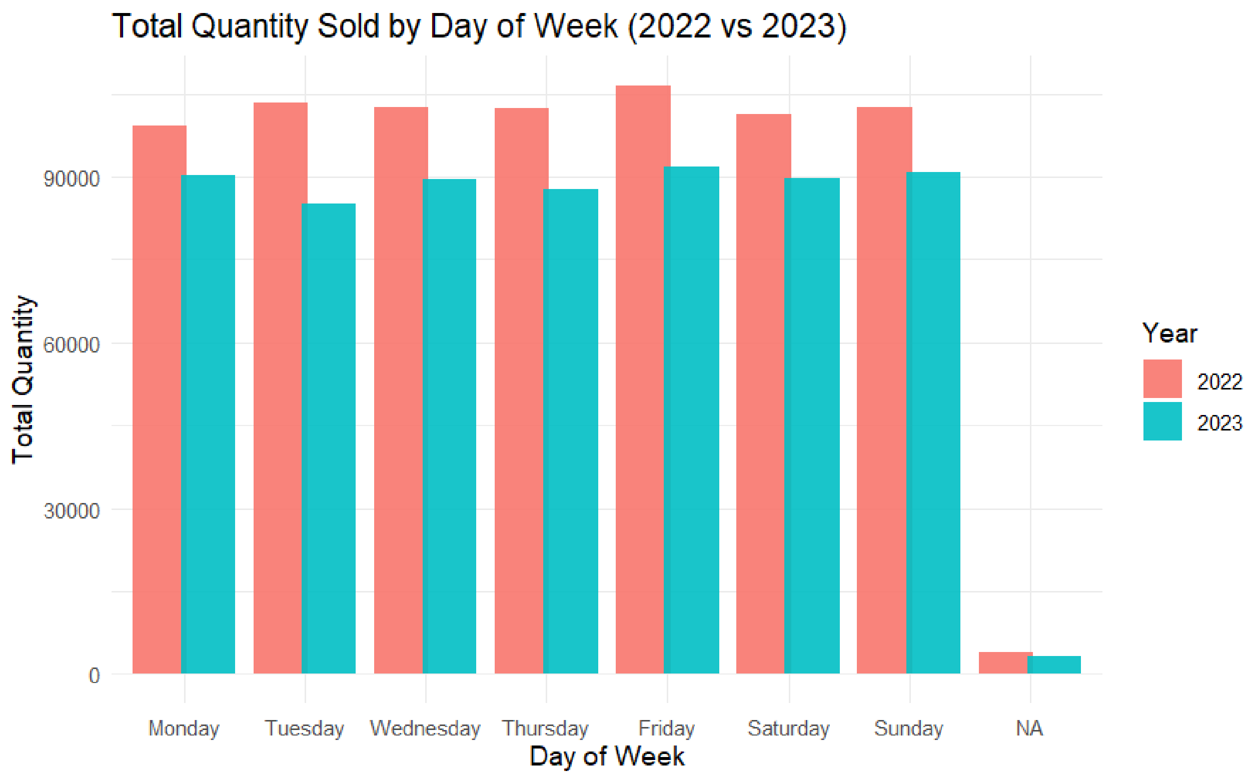


Figure 8 - Quantity Sold by Day of the Week

Sales volumes stay steady during the week, with only slight variations between weekdays and weekends. Sales figures for both years peak from Monday to Friday, while the totals for Saturday and Sunday are somewhat lower. The values for 2022 are slightly greater across each day, suggesting a minor decline in 2023 compared to the previous year.

Quantity Sold by Month

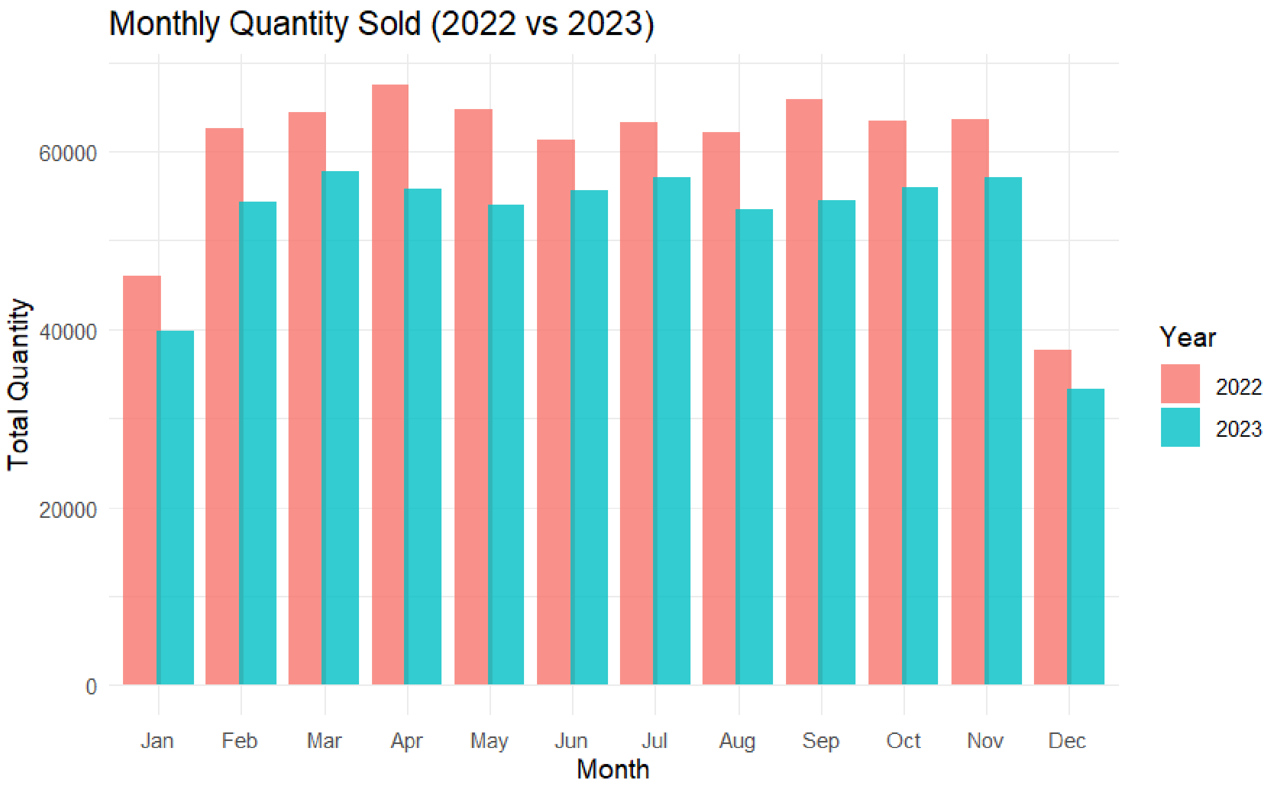


Figure 9 - Quantity Sold by Month

Sales figures for each month remain relatively stable across both years, demonstrating a distinct peak in the middle of the year from March to May. However, there is a noticeable decline in sales at the start and end of each year, especially in January and December. This trend indicates a reduction in demand during holiday seasons and the period immediately following, likely due to typical seasonal buying habits. While the overall sales in 2022 are slightly higher than those in 2023, both years exhibit the same pattern, highlighting consistent mid-year demand and expected seasonal decreases at the beginning and end of the year.

Total Sales per Year, Rand Value

Before further analysis could continue, we first had to match the sales data for 2022 and 2023 CSV to the product data CSV by product ID. Once completed, we then took the quantity sold and multiplied it by the selling price to see the sales value per order.

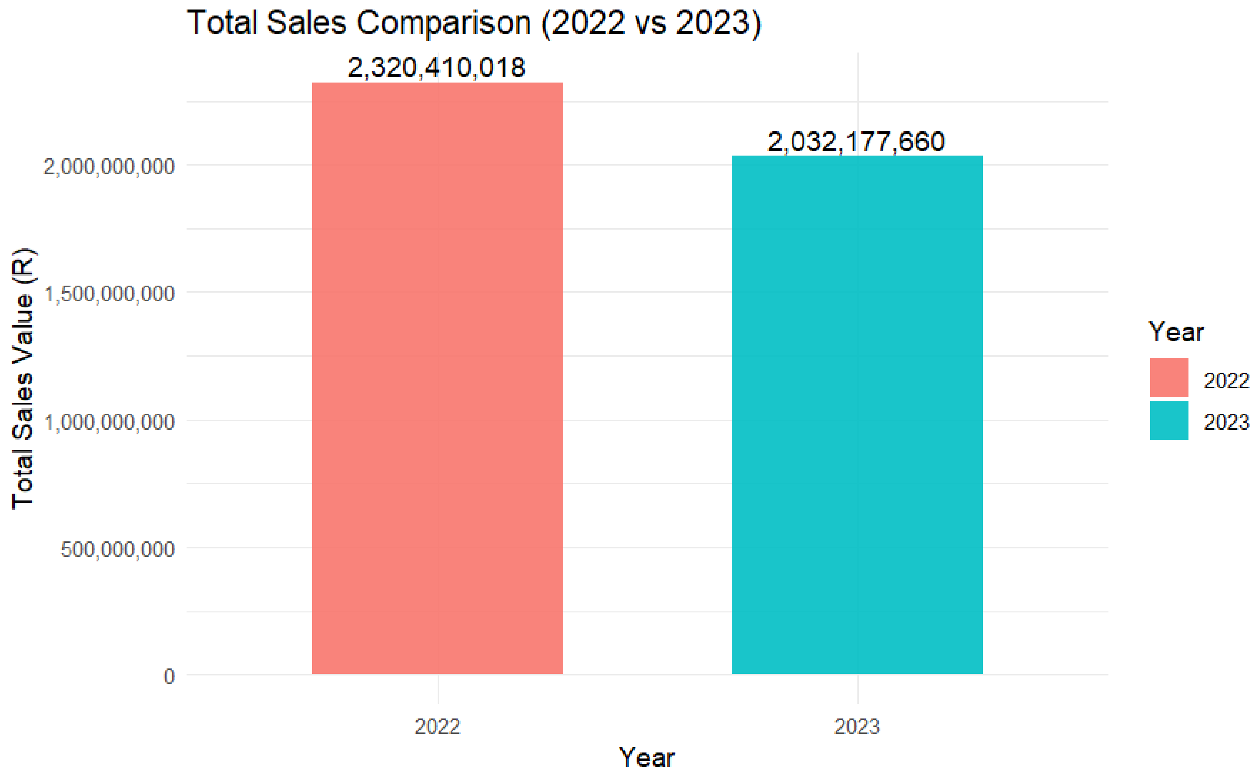


Figure 10 - Yearly Comparison of Sales

The total sales amount fell from about R2.32 billion in 2022 to R2.03 billion in 2023, representing a reduction of roughly 12%. This decrease suggests a slightly poorer overall performance compared to the previous year, aligning with the reduced monthly quantities noted in 2023.

Sales per Category

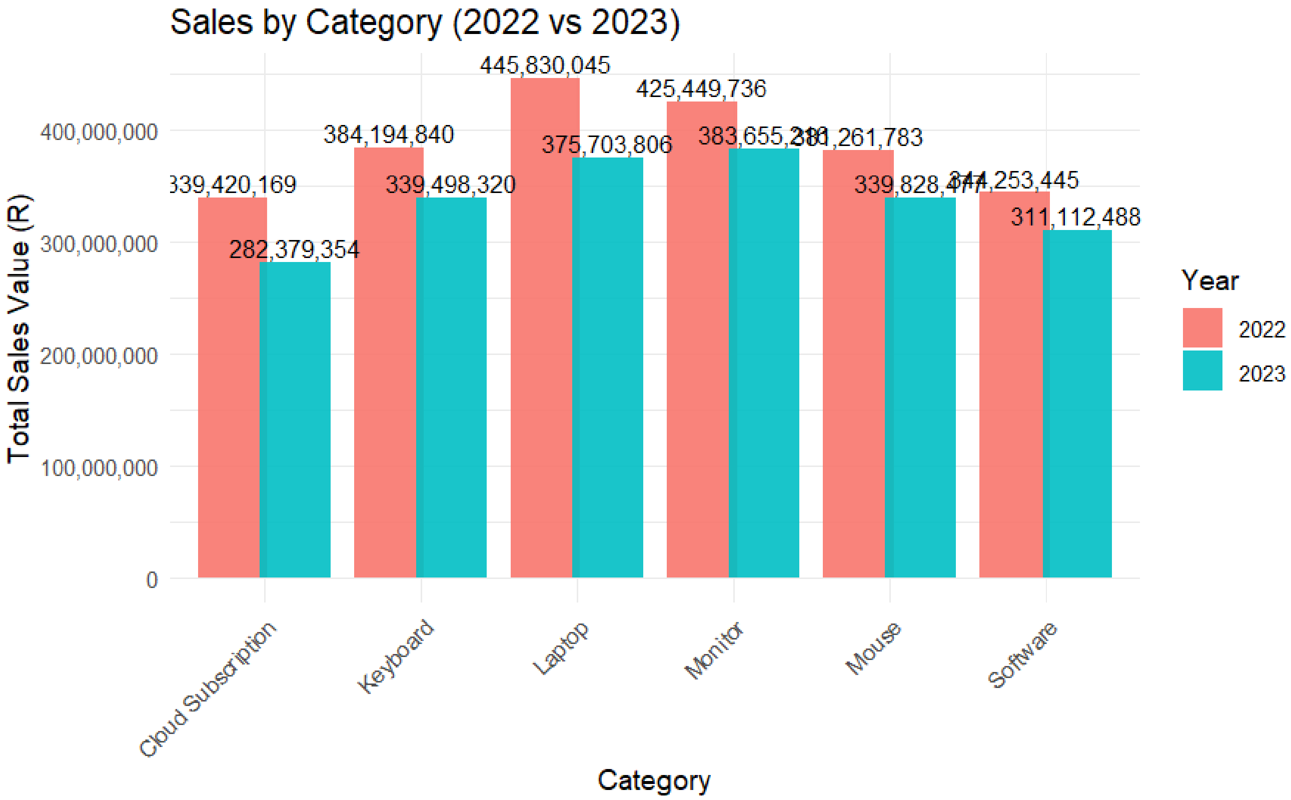


Figure 11 - Sales per Category

In both years, laptops had the highest total sales, with monitors and keyboards following in second and third place. Cloud subscriptions and software had the lowest sales figures recorded. All categories show a steady decline from year to year, with 2023 sales lower than those of 2022. Although there has been a decrease, the overall ranking of categories remains unchanged, indicating that consumer purchasing preferences have not varied notably.

Gross Profit per Year

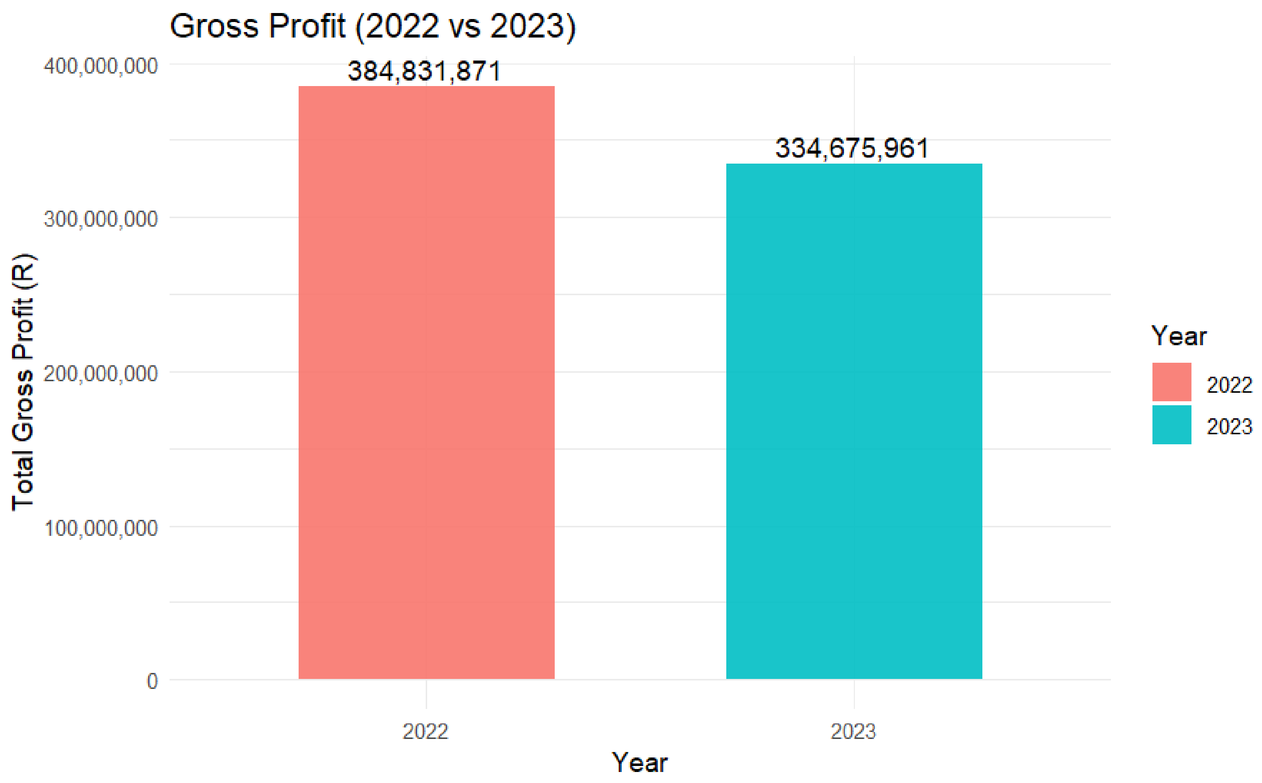


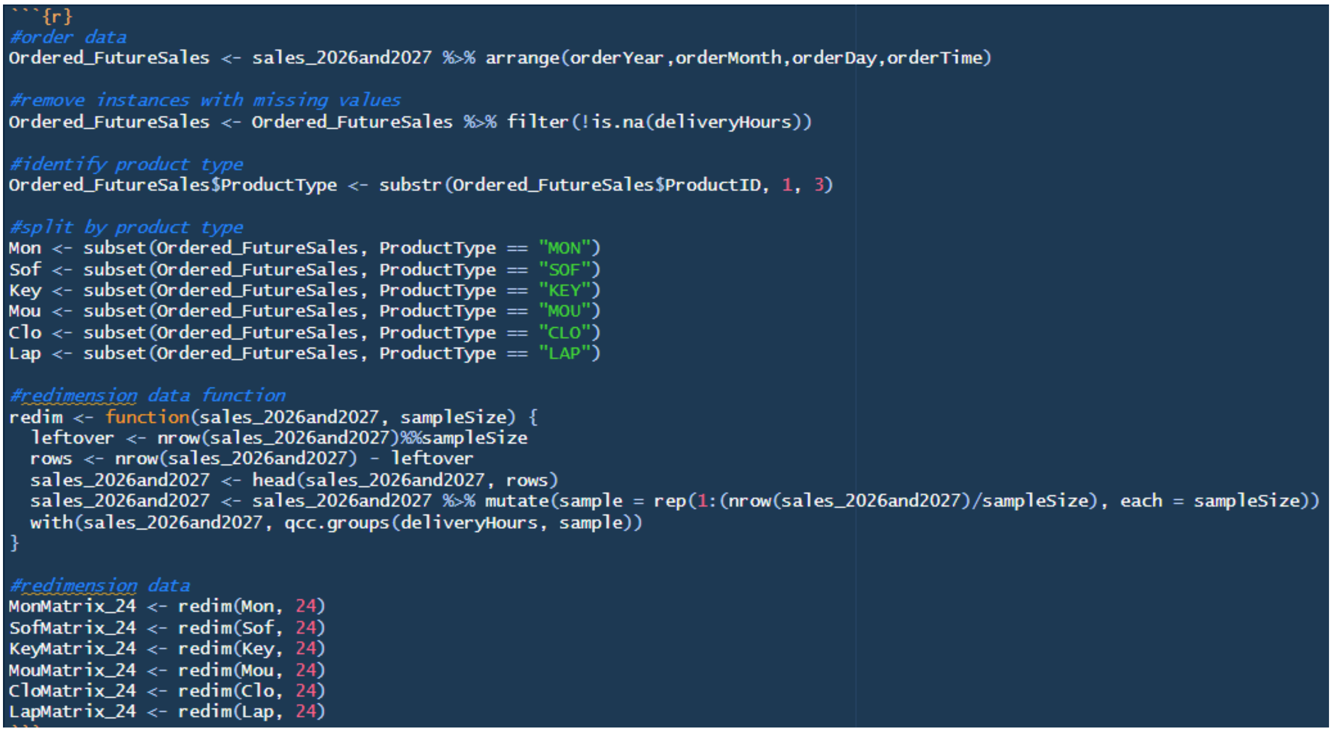
Figure 12 - Gross Profit per Year

Using the Markups for the various products, we calculated the cost price of goods and used this to work out gross profit. Gross profit declined from approximately R384.8 million in 2022 to R334.7 million in 2023, representing a reduction of about 13%. This downward movement mirrors the overall drop in total sales for the same period, suggesting that profitability was affected primarily by reduced sales volume rather than changes in cost structure.

Part 3 Statistical Process Control

Statistical Process Control (SPC) techniques were utilised to assess the performance of service delivery and the stability of processes over time. By analysing past delivery-hour data, X and S charts were created for each product category to track variations and identify any indications of process shifts or instability.

3.1 to 3.2 Process of Creating Control Charts



~~~
```{r}
#order data
Ordered_FutureSales <- sales_2026and2027 %>% arrange(orderYear,orderMonth,orderDay,orderTime)

#remove instances with missing values
Ordered_FutureSales <- Ordered_FutureSales %>% filter(!is.na(deliveryHours))

#identify product type
Ordered_FutureSales$ProductType <- substr(Ordered_FutureSales$ProductID, 1, 3)

#split by product type
Mon <- subset(Ordered_FutureSales, ProductType == "MON")
Sof <- subset(Ordered_FutureSales, ProductType == "SOF")
Key <- subset(Ordered_FutureSales, ProductType == "KEY")
Mou <- subset(Ordered_FutureSales, ProductType == "MOU")
Clo <- subset(Ordered_FutureSales, ProductType == "CLO")
Lap <- subset(Ordered_FutureSales, ProductType == "LAP")

#redimension data function
redim <- function(sales_2026and2027, sampleSize) {
  leftover <- nrow(sales_2026and2027)%%sampleSize
  rows <- nrow(sales_2026and2027) - leftover
  sales_2026and2027 <- head(sales_2026and2027, rows)
  sales_2026and2027 <- sales_2026and2027 %>% mutate(sample = rep(1:(nrow(sales_2026and2027)/sampleSize), each = sampleSize))
  with(sales_2026and2027, qcc.groups(deliveryHours, sample))
}

#redimension data
MonMatrix_24 <- redim(Mon, 24)
SofMatrix_24 <- redim(Sof, 24)
KeyMatrix_24 <- redim(Key, 24)
MouMatrix_24 <- redim(Mou, 24)
CloMatrix_24 <- redim(Clo, 24)
LapMatrix_24 <- redim(Lap, 24)
```
~~~

The control charts were constructed in RStudio using the qcc package to assess process stability for each product type. The dataset was first cleaned, ordered chronologically, and split into six product categories. A custom redim() function then reorganised the delivery data into subgroups of 24 observations, which were used to generate X̄ and s charts. The first 30 subgroups established baseline control limits (Phase I), while the remaining data were used to monitor process performance (Phase II).

3.3 Analysis of Control Charts and Process Stability

Monitors

Xbar Control Chart

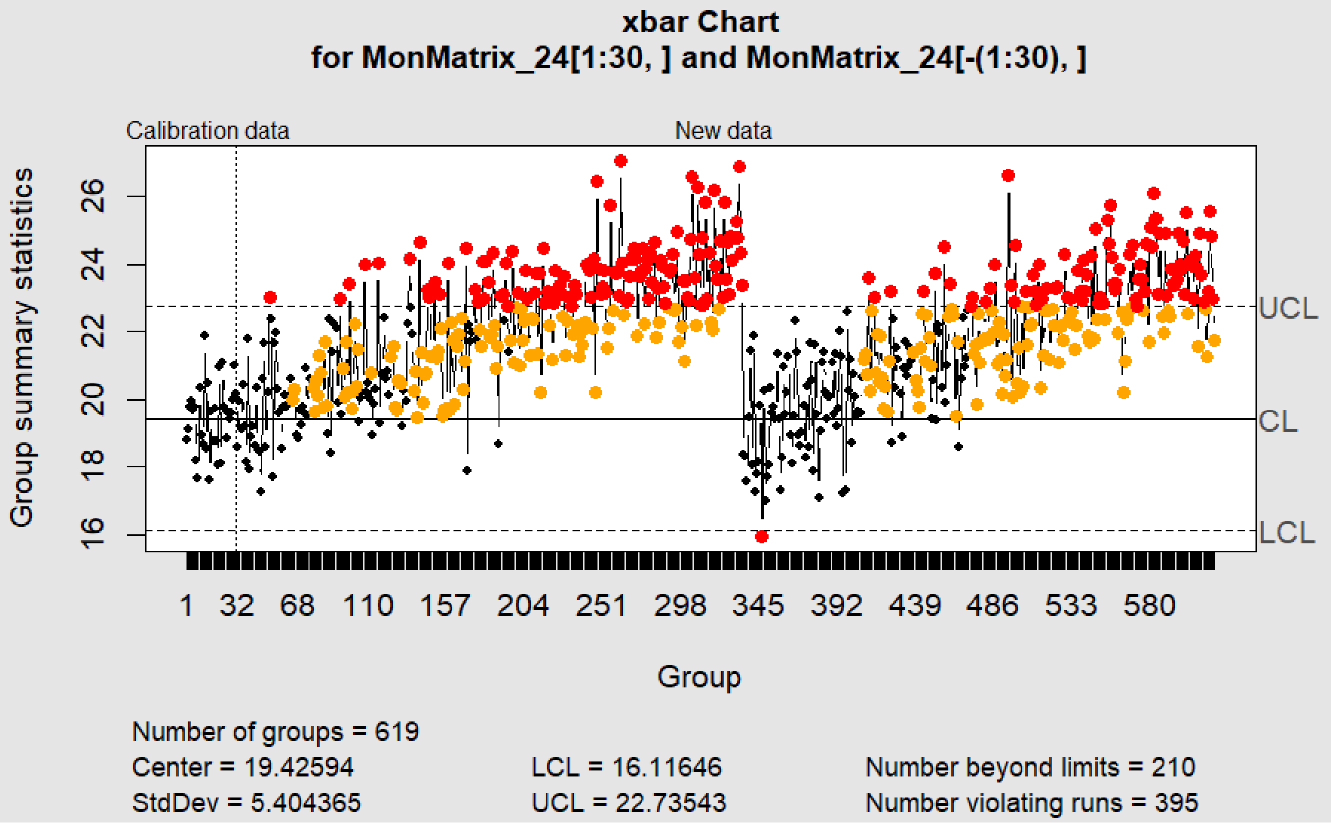


Figure 13 - Xbar Control Chart for Monitors

The X chart for monitor delivery times shows a process mean of 19.43 hours with control limits set between 16.12 hours (LCL) and 22.74 hours (UCL). The first 30 subgroups (Phase I) exhibit moderate variability and predominantly stay within the control limits, indicating that the process was initially stable during calibration. However, numerous points in the subsequent phase (Phase II) exceed the control limits, with 210 points falling outside the limits and 395 instances of runs that are in violation, indicating significant process instability.

S Control Chart

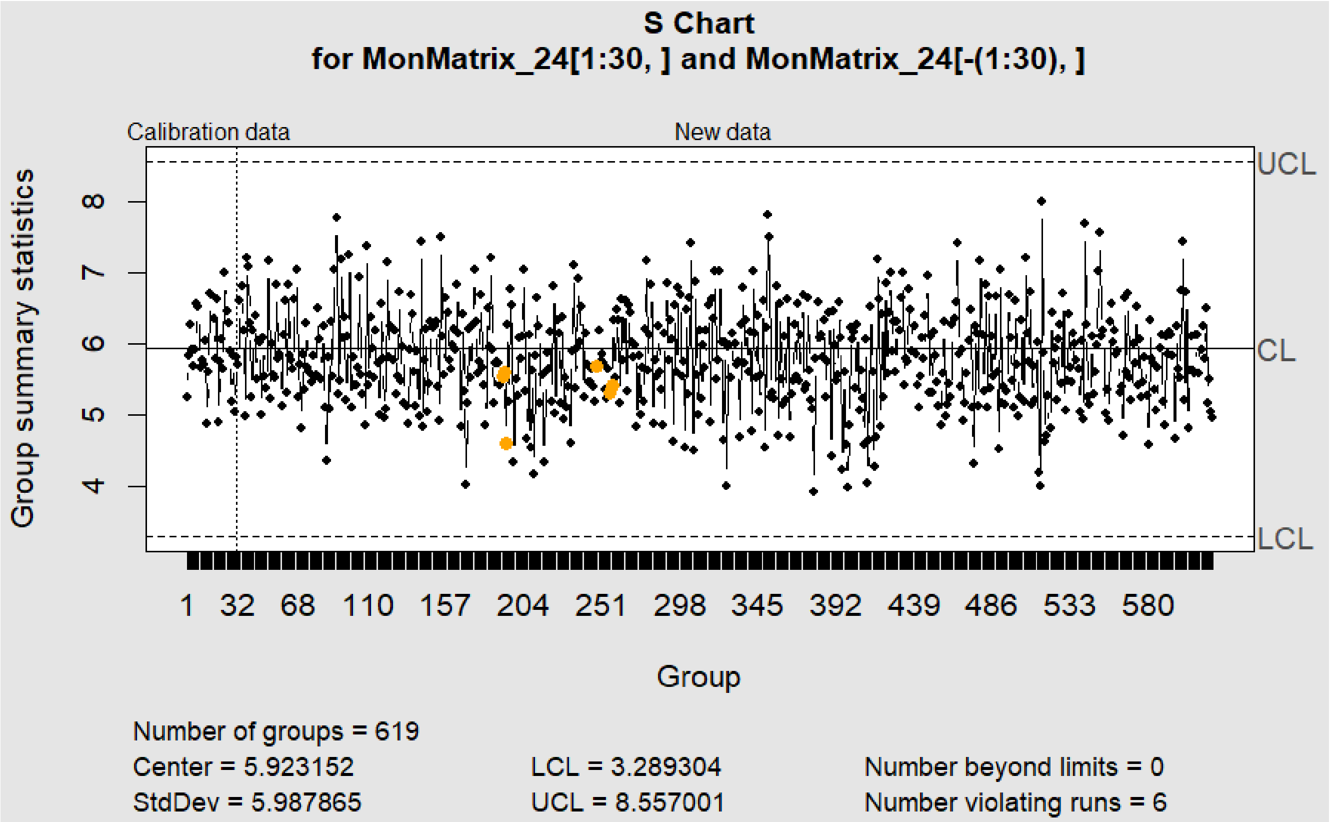


Figure 14 - S Control Chart for Monitors

The S chart for monitor delivery times has a process centre of 5.92 hours, with control limits ranging from 3.29 hours (LCL) to 8.56 hours (UCL). Every point lies within the specified limits, and only six runs exhibit slight violations, indicating that the process variability is stable and effectively managed. Even though the mean delivery time shifted, the variability is consistent. Thus, the process became slower overall, but the level of variation didn’t explode.

Process Capability

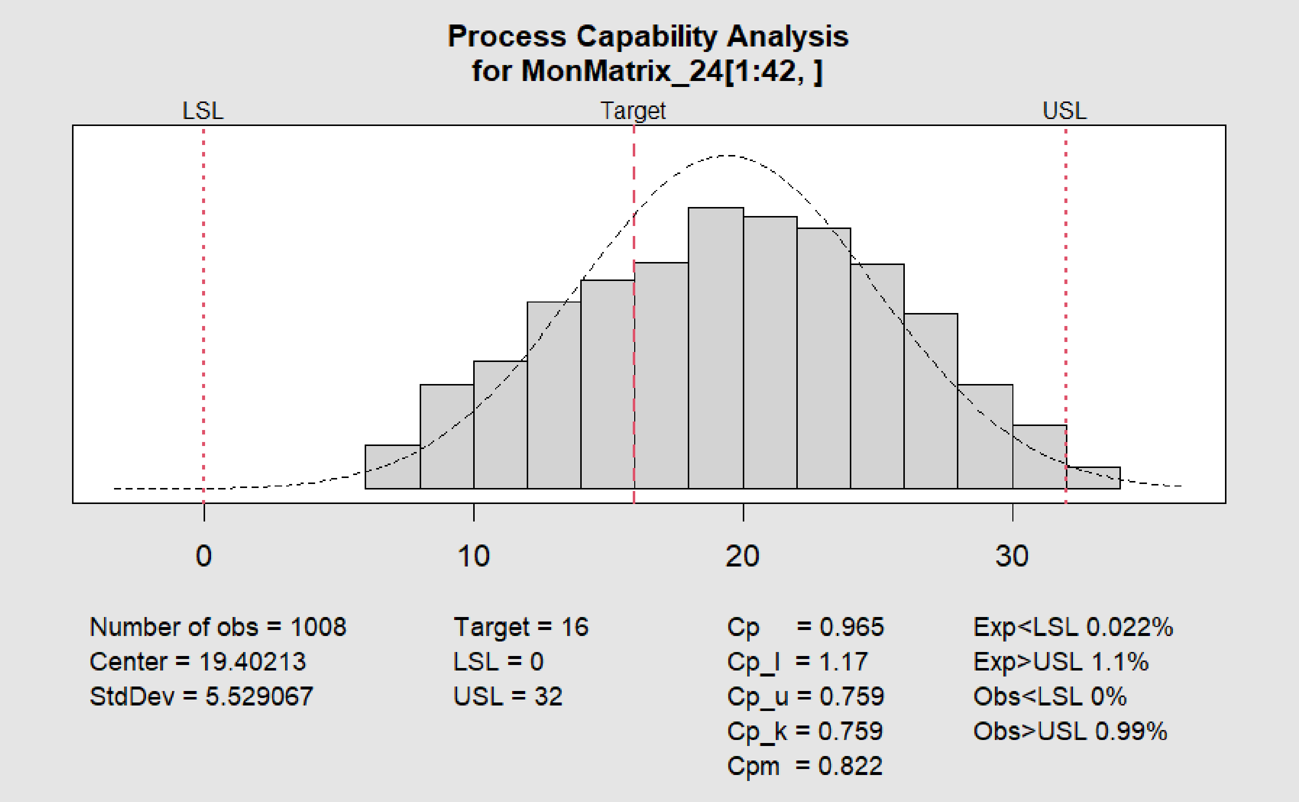


Figure 15 - Process Capability of Monitors

The process mean of 19.4 hours is slightly above the target of 16 hours. The specification limits are set between 0 and 32 hours. The calculated Potential Capacity (Cp) = 0.965 and Process Capability Index (Cpk) = 0.759 are both less than 1. A Cp of 0.965 indicates the process for monitoring delivery times is close to being capable but not quite meeting the required tolerance. The Cpk of 0.965 value implies that the process is not centred around the target and tends to run slower than desired. 1.1% of observations fall above the USL, indicating there is a small risk that the delivery time of monitors will go above the upper specification level.

Software

Xbar Control Chart

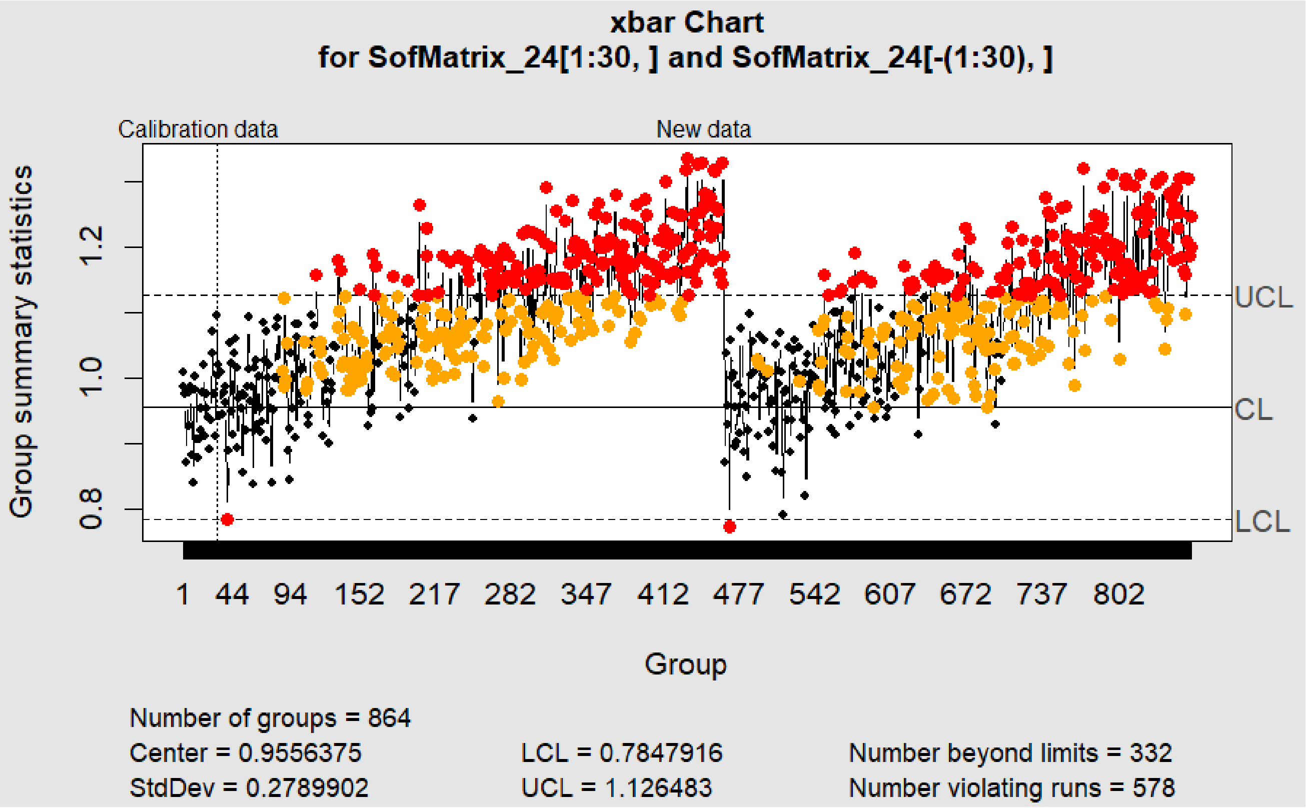


Figure 16 - Xbar Control Chart for Software

The X chart for software delivery times shows a process mean of 0.96 hours, with control limits set between 0.78 hours (LCL) and 1.13 hours (UCL). The first 30 subgroups (Phase I) display moderate variability and remain mostly within the control limits, indicating initial process stability during calibration. However, in the subsequent monitoring phase (Phase II), 332 points fall outside the control limits, and 578 runs violate the expected pattern, signifying a high degree of process instability. The frequent points above the upper control limit suggest an upward shift in the process mean, indicating that delivery times have gradually increased over time.

S Control Chart

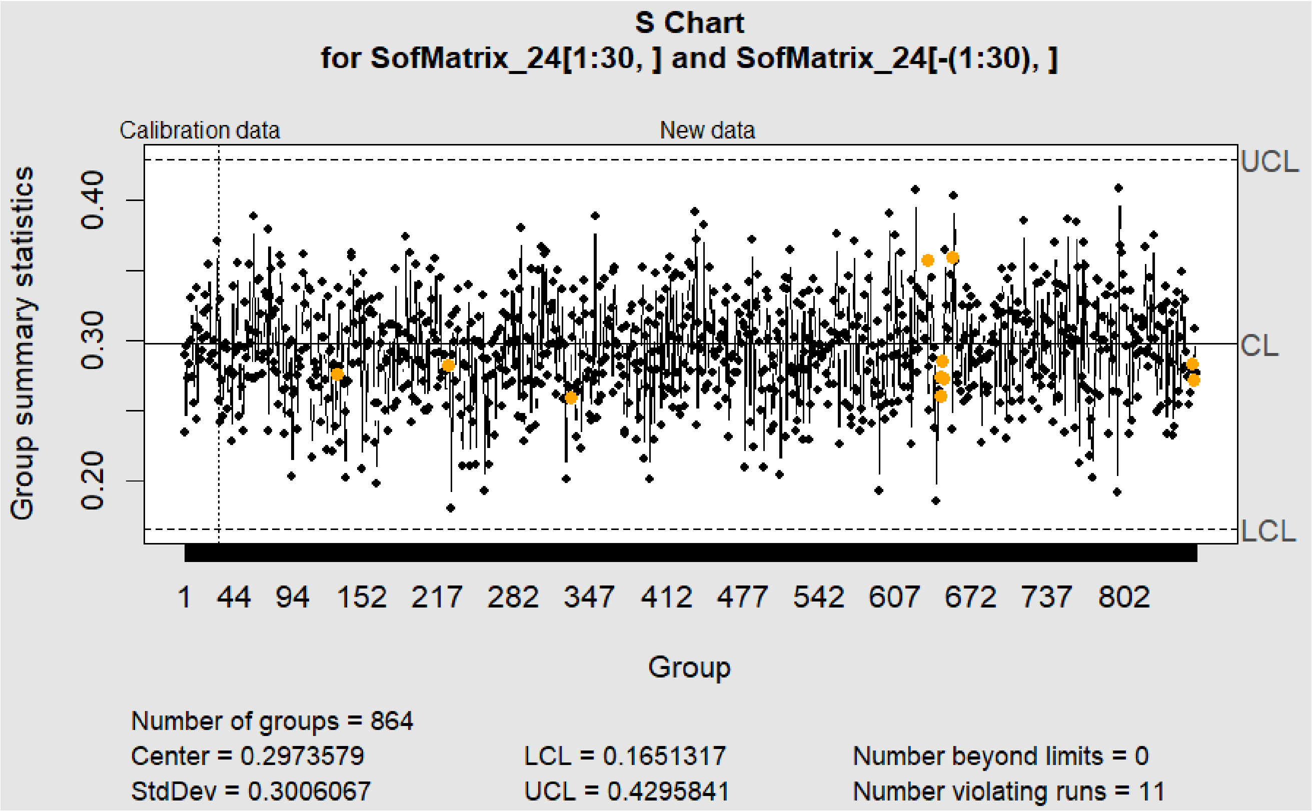


Figure 17 - S Control Chart for Software

The S chart for software delivery times shows a process mean (centre line) of 0.30 hours, with control limits set between 0.17 hours (LCL) and 0.43 hours (UCL). All data points remain within control limits, with no points exceeding these limits and only 11 minor run violations noted. This indicates that the variability in delivery times is stable and well-managed. While the X chart reflects shifts in the process mean, the S chart verifies that the overall variation in the process has stayed consistent.

Process Capability

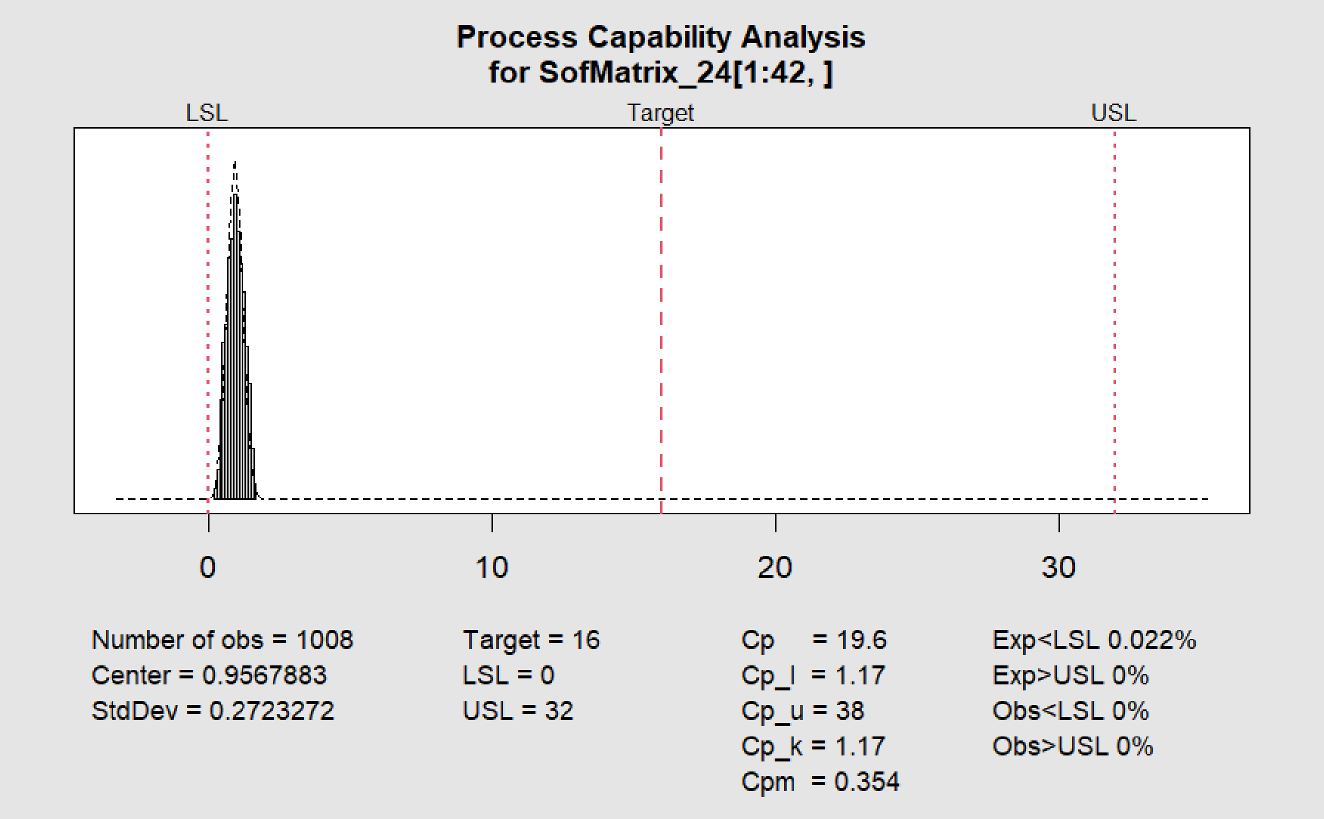


Figure 18 - Process Capability of Software

The process mean of 0.96 hours is well below the target of 16 hours. The specification limits are set between 0 and 32 hours. The calculated Potential Capacity (Cp) = 19.6 and Process Capability Index (Cpk) = 1.17 are both well above 1, indicating that the process is extremely capable of meeting delivery-time specifications. The extremely high Cp value indicates that there's very little variation in the process compared to the tolerance range, while a Cpk value exceeding 1 verifies that the process is effectively centred within its limits. There are no instances where observations exceed the lower or upper specification limits, implying that the software delivery times are remarkably consistent, posing almost no risk of surpassing customer expectations or tolerance thresholds.

Keyboards

Xbar Control Chart

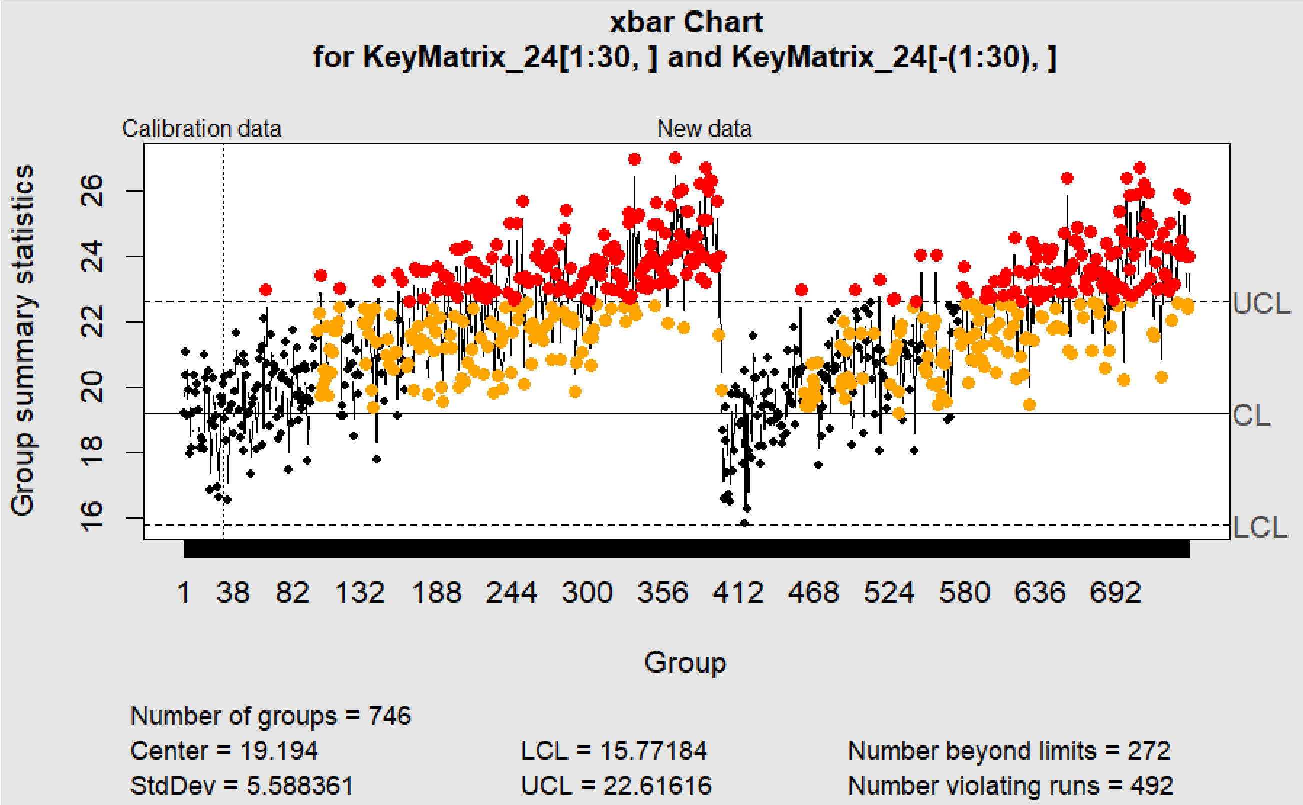


Figure 19 - Xbar Control Chart for Keyboards

The X chart for keyboard delivery times shows a process mean of 19.19 hours, with control limits set between 15.77 hours (LCL) and 22.62 hours (UCL). The initial 30 subgroups (Phase I) display consistent variation and largely stay within the control limits, suggesting that the process was stable during the calibration phase. In the following monitoring phase (Phase II), 272 data points surpass the control limits, and 492 runs breach the anticipated pattern, indicating a deterioration in process control.

S Control Chart

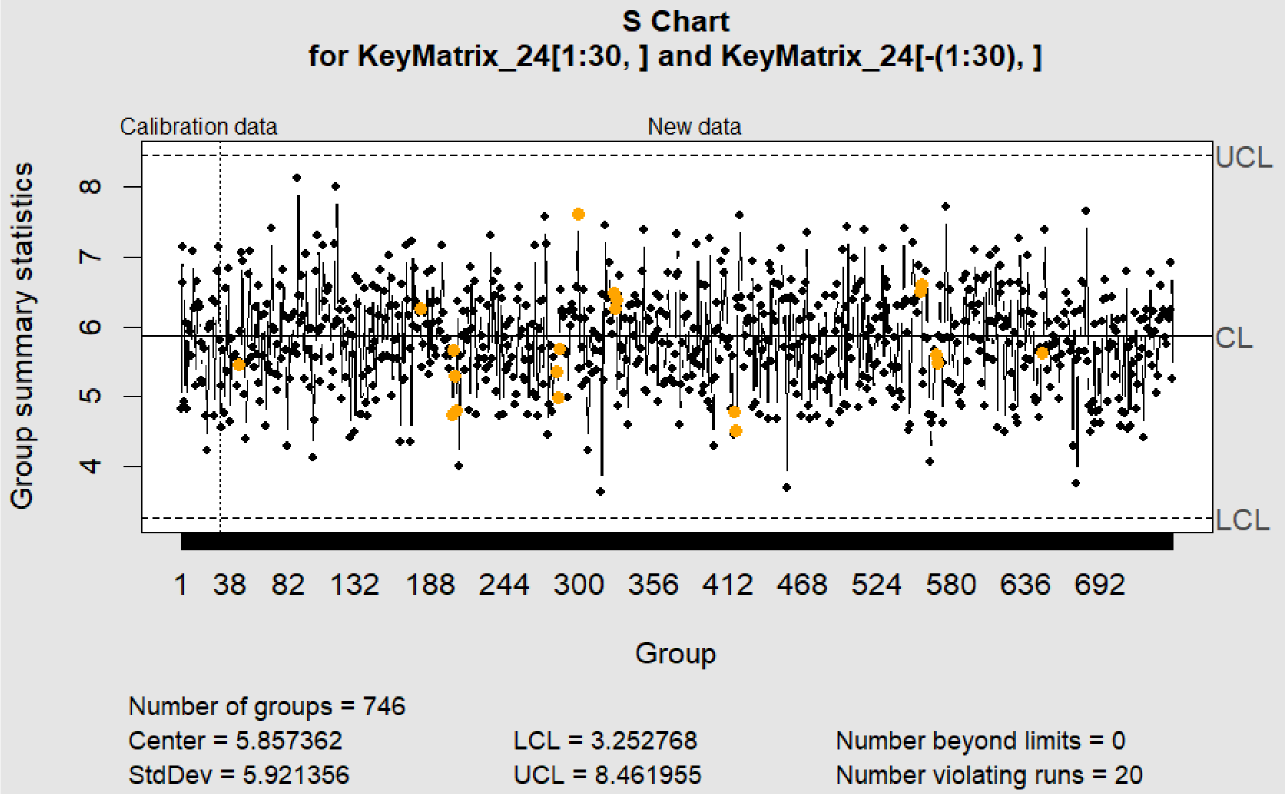


Figure 20 - S Control Chart for Keyboards

The S chart for keyboard delivery times shows a process mean (centre line) of 5.86 hours, with control limits set between 3.25 hours (LCL) and 8.46 hours (UCL). All points fall within the specified limits, with just 20 minor run violations recorded. This suggests that the variation in the process is stable and well-managed over time. Although the X chart has shown multiple shifts in mean and fluctuations in average delivery performance, the S chart affirms that the fundamental variation in the process stays consistent.

Process Capability

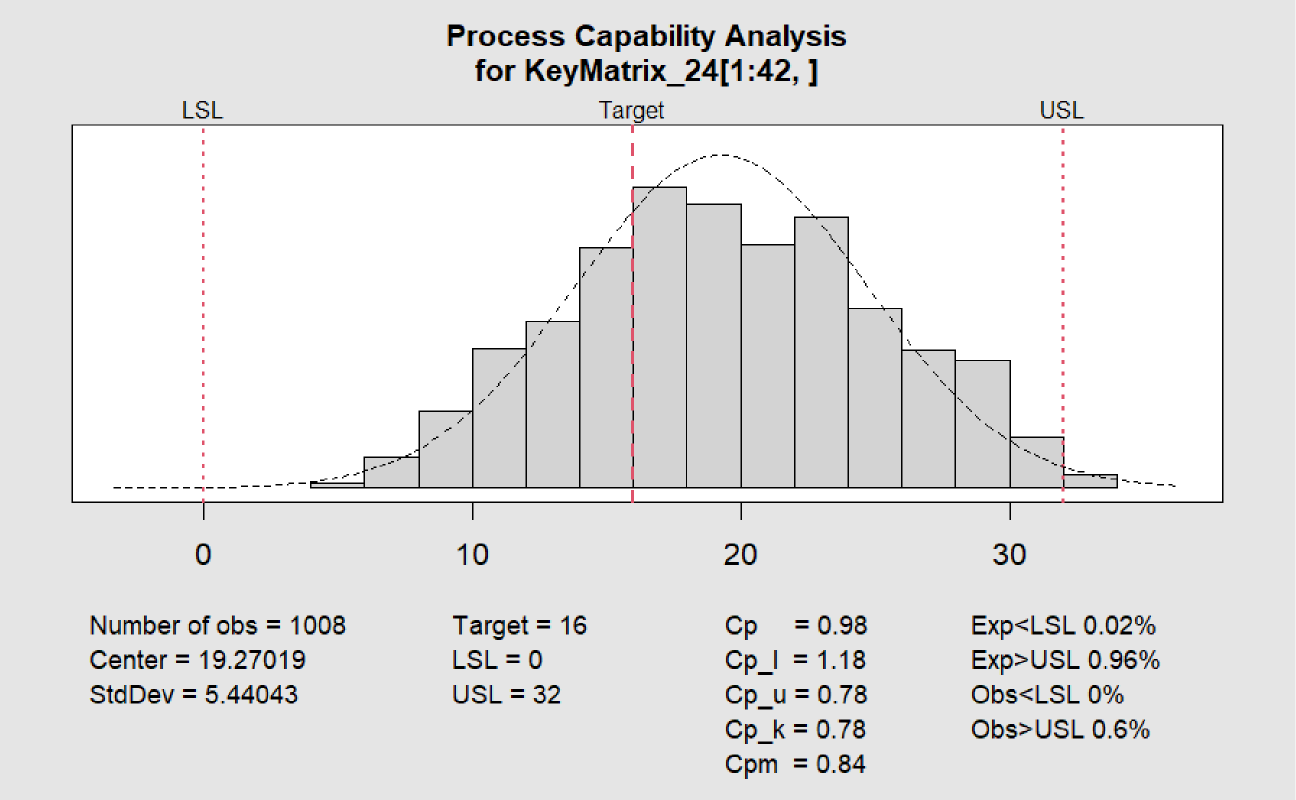


Figure 21 - Process Capability of Keyboards

The process mean of 19.27 hours is above the target of 16 hours. The specification limits are set between 0 and 32 hours. The calculated Potential Capacity (Cp) = 0.98 and Process Capability Index (Cpk) = 0.78 are both slightly below 1, indicating that the process is marginally incapable of consistently meeting delivery-time requirements. A Cp of 0.98 indicates that the overall variation of the process is nearly within the specifications, yet it still slightly surpasses the acceptable tolerance limits. The lower Cpk value reveals that the process is not aligned with the target and often leads to longer delivery times than expected. Roughly 0.6% of the recorded observations exceed the upper specification limit (USL), pointing out a minor risk of delays in keyboard orders.

Mouses

Xbar Control Chart

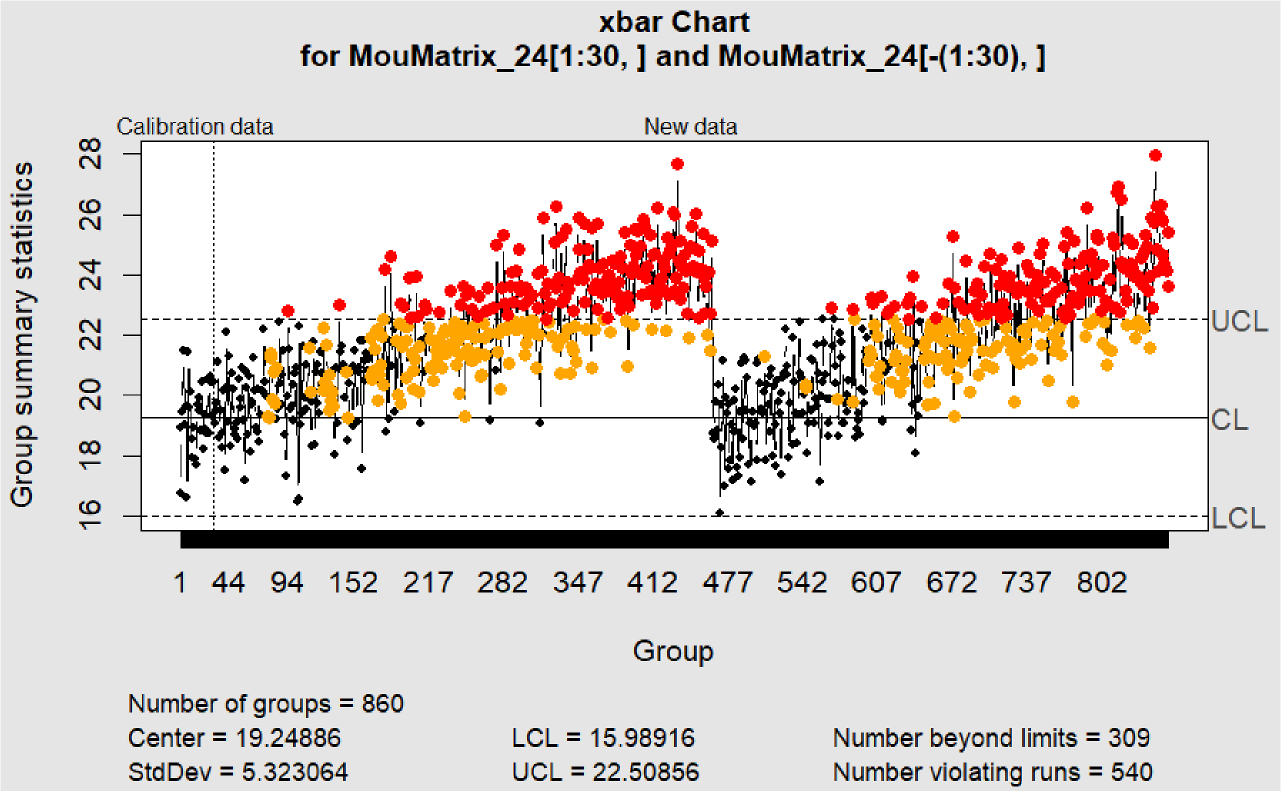


Figure 22 - Xbar Control Chart for Mouses

The X chart for mouse delivery times shows a process mean of 19.25 hours, with control limits set between 15.99 hours (LCL) and 22.51 hours (UCL). In Phase I, the initial 30 subgroups largely stay within the control limits, indicating consistent performance throughout the calibration period. Conversely, during the monitoring phase (Phase II), 309 data points exceed the control limits, and 540 runs violate the control rules, signifying considerable process instability. The repeated instances of points above the upper control limit imply an upward shift in the mean of the process, which has led to longer delivery times as time progresses.

S Control Chart

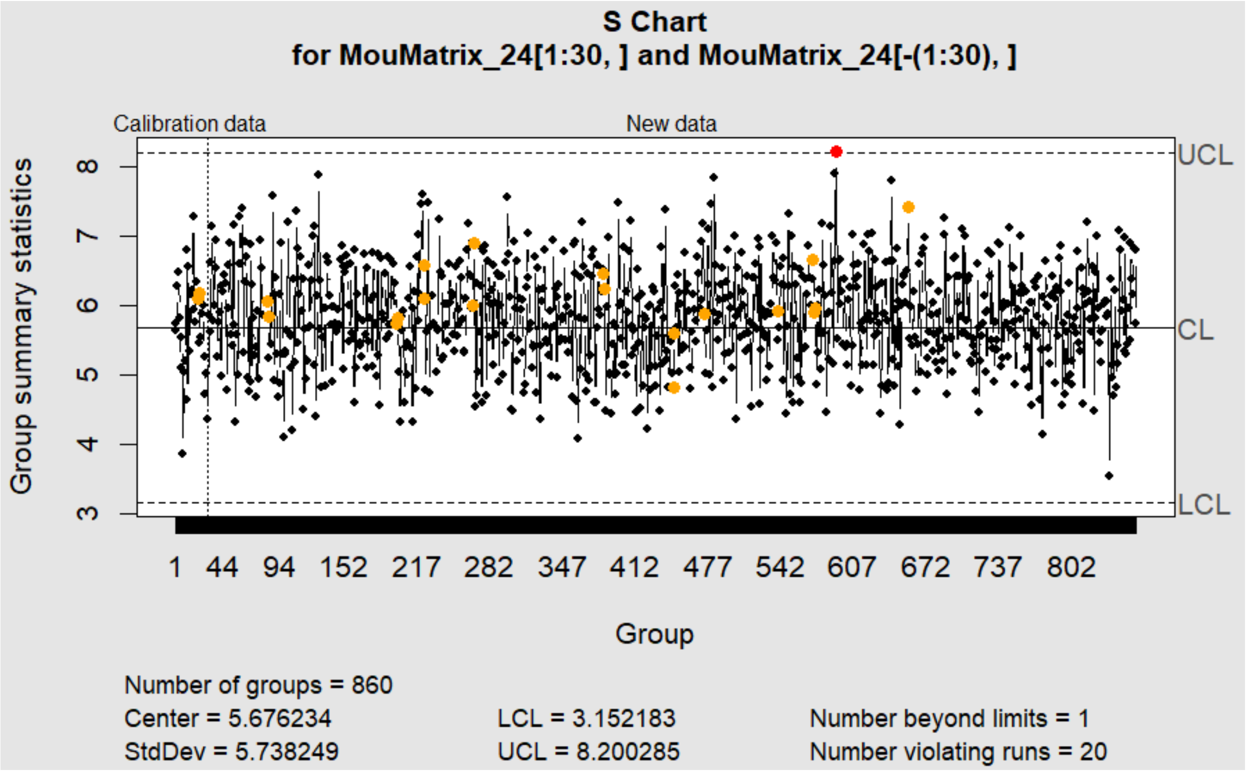


Figure 23 - S Control Chart for Mouses

The S chart for mouse delivery times shows a process mean (centre line) of 5.68 hours, with control limits set between 3.15 hours (LCL) and 8.20 hours (UCL). During the monitoring phase (Phase II), there is only a single data point that lies beyond the control limits, accompanied by 20 minor run violations noted. This suggests that the variability in delivery times is steady and typically stable across the process. While the X chart indicated shifts in the mean and fluctuations in average delivery times, the S chart verifies that the level of variation has stayed consistent.

Process Capability

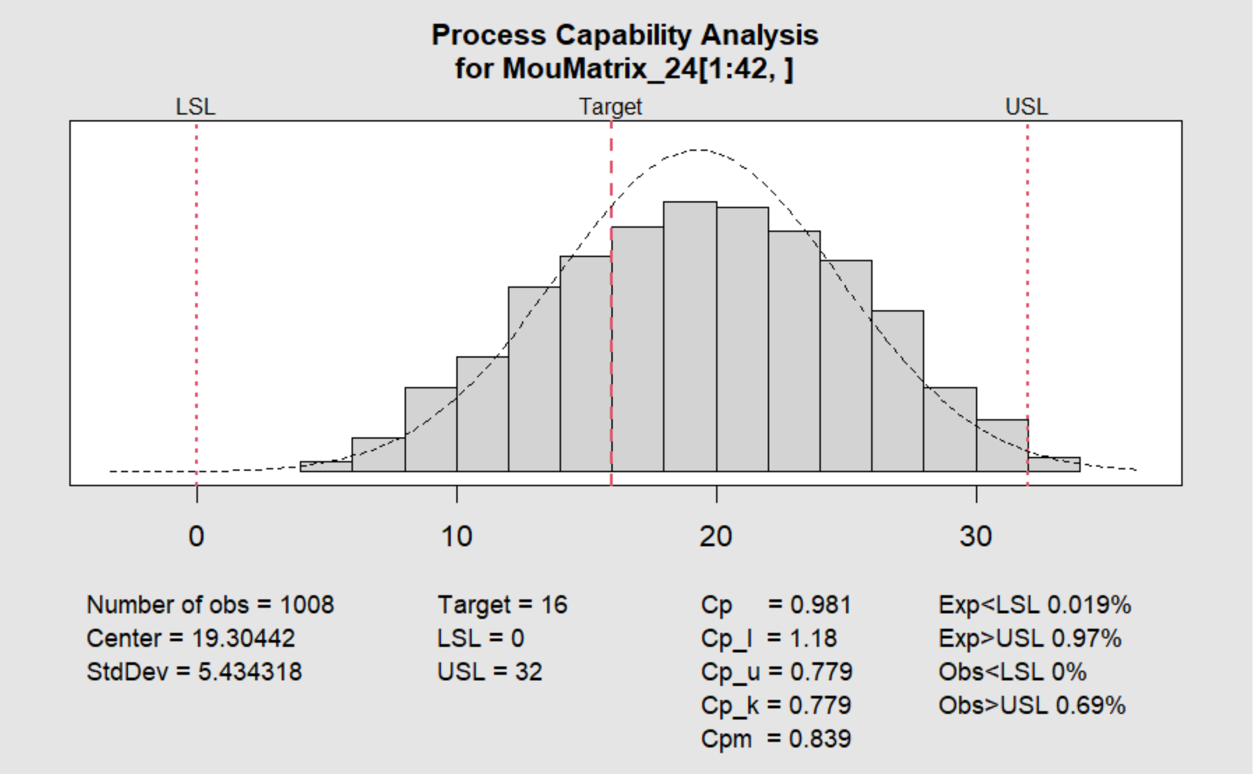


Figure 24 - Process Capability of Mouses

The process mean of 19.30 hours is above the target of 16 hours. The specification limits are set between 0 and 32 hours. The calculated Potential Capacity (Cp) = 0.981 and Process Capability Index (Cpk) = 0.779 are both slightly below 1, indicating that the process is close to being capable but not fully meeting the required tolerance. A Cp value of 0.981 indicates that the total process variation is almost aligned with the specification limits, whereas the lower Cpk value shows that the process is not optimally centred and generally operates slower than the target rate. About 0.69% of the observations surpass the upper specification limit (USL), which highlights a minor risk that delivery times for mouse products could exceed acceptable service levels.

Cloud Subscription

Xbar Control Chart

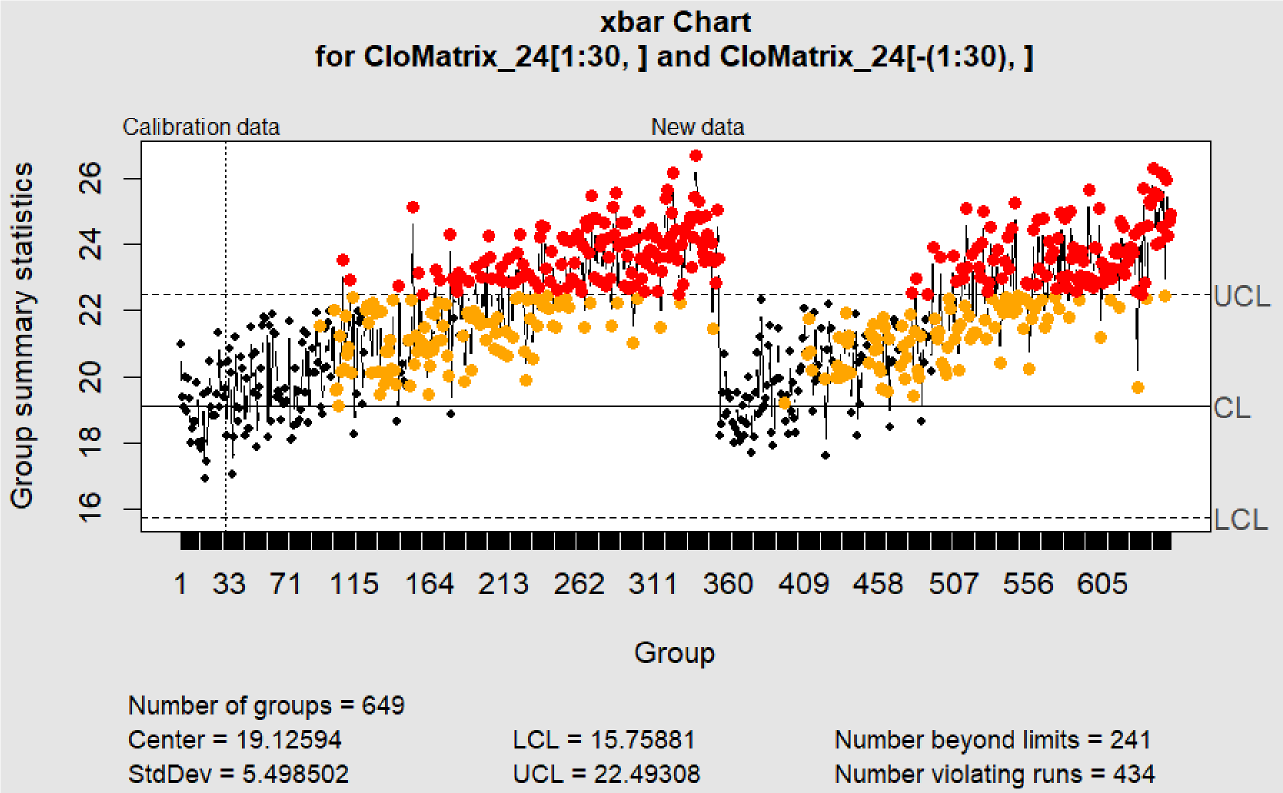


Figure 25 - Xbar Control Chart for Cloud Subscription

The X chart for cloud subscription delivery times shows a process mean of 19.13 hours, with control limits set between 15.76 hours (LCL) and 22.49 hours (UCL). The initial 30 subgroups (Phase I) stays predominantly within the control limits, indicating that the process was stable during the calibration phase. In contrast, during the monitoring phase (Phase II), 241 data points surpass the control limits, and 434 sequences breach the anticipated control pattern, signifying a shift toward instability in the process. The chart highlights multiple instances where the process mean trends upwards, illustrating a gradual rise in average delivery times.

S Control Chart

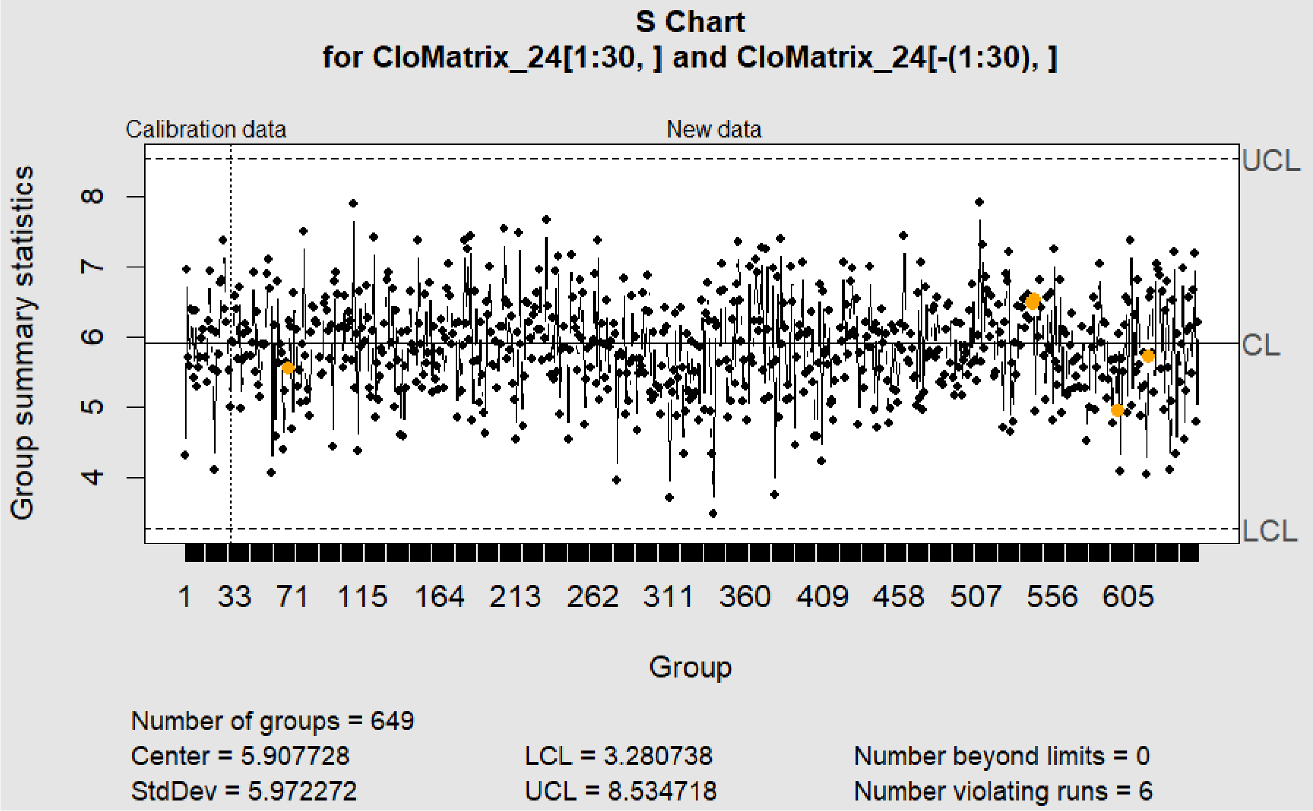


Figure 26 - S Control Chart for Cloud Subscription

The S chart for cloud subscription delivery times shows a process mean (centre line) of 5.91 hours, with control limits set between 3.28 hours (LCL) and 8.53 hours (UCL). All data points stay within the specified limits, with only six minor run violations identified. This suggests that the variability of the process is stable and controlled throughout the recorded time frame. While the X chart revealed significant shifts in the process mean and phases of instability, the S chart verifies that the overall variation level remains steady.

Process Capability

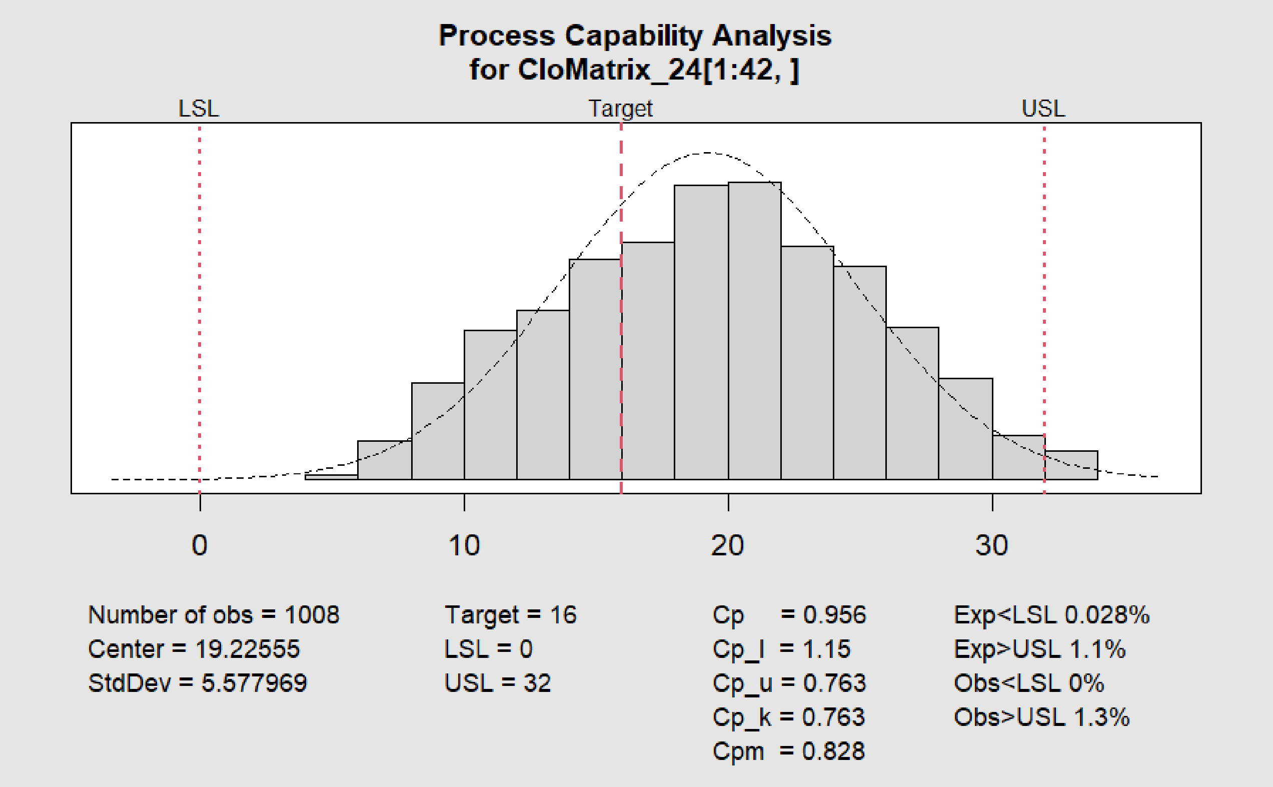


Figure 27 - Process Capability of Cloud Subscription

The process mean of 19.23 hours is above the target of 16 hours. The specification limits are set between 0 and 32 hours. The calculated Potential Capacity (Cp) = 0.956 and Process Capability Index (Cpk) = 0.763 are both below 1, indicating that the process is not capable of consistently meeting the desired delivery-time specifications. A Cp value of 0.956 indicates that the overall variation in the process is nearly within the specification limits, though it slightly surpasses the acceptable tolerance. The lower Cpk value signifies that the process is not properly centred and generally operates at a slower pace than the target time. Around 1.3% of observations are above the upper specification limit (USL), demonstrating a minor risk of delays in delivering cloud subscription services.

Laptops

Xbar Control Chart

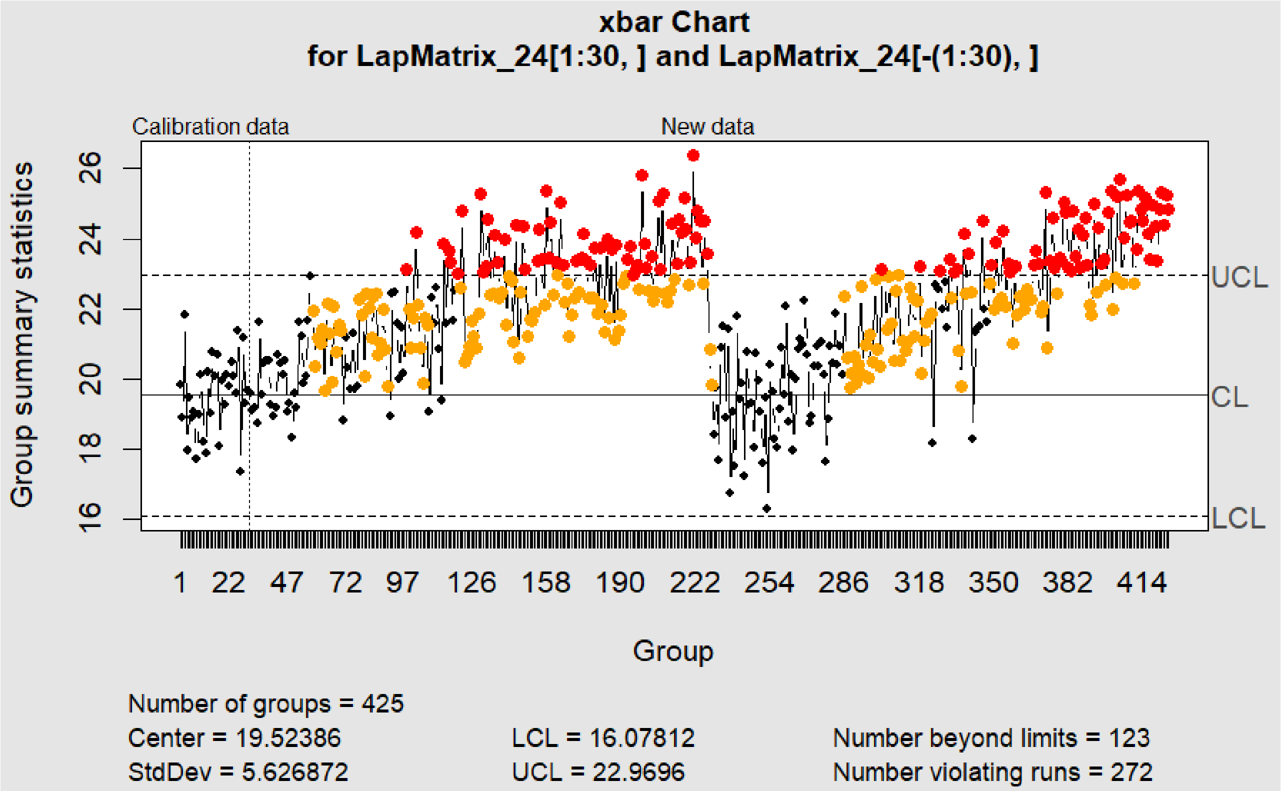


Figure 28 - Xbar Control Chart for Laptops

The X chart for laptop delivery times shows a process mean of 19.52 hours, with control limits set between 16.08 hours (LCL) and 22.97 hours (UCL). The initial 30 subgroups (Phase I) largely stayed within the control limits, indicating that the process was stable during the calibration period. In contrast, during the monitoring phase (Phase II), 241 points surpassed the control limits, and 434 runs deviated from the expected control pattern, suggesting that the process has become unstable. The chart shows multiple instances where the process mean has shifted upward, indicating a gradual increase in average delivery times.

S Control Chart

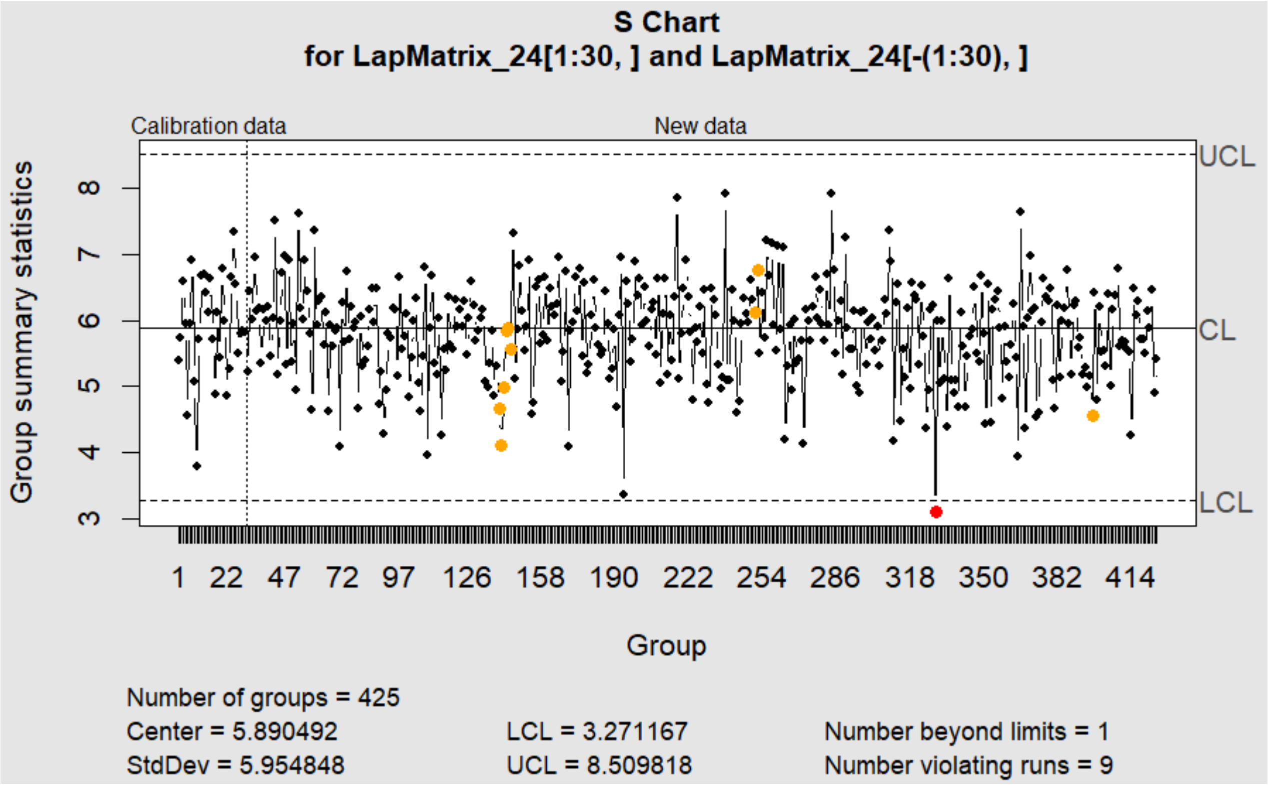
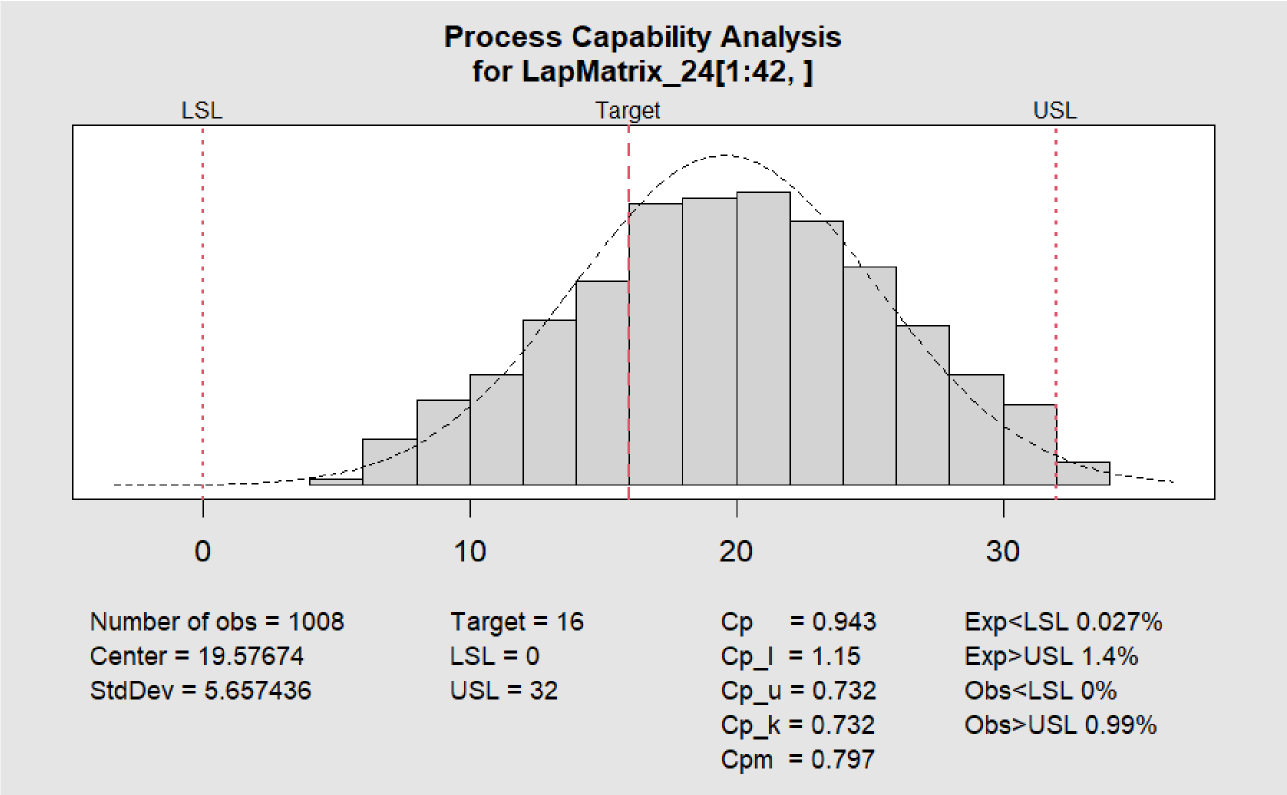


Figure 29 - S Control Chart for Laptops

The S chart for laptop delivery times shows a process mean (centre line) of 5.89 hours, with control limits set between 3.27 hours (LCL) and 8.51 hours (UCL). During the monitoring phase (Phase II), there is only one instance that exceeds the control limits, along with nine minor violations in runs, suggesting that the variation in delivery times is stable and managed. Although there is some instability noted in the X chart, the S chart indicates that the process spread has not experienced a significant rise. This implies that while the average delivery times have gradually increased, the overall variability has remained constant.

Process Capability



```
Number of obs = 1008     Target = 16     Cp    = 0.943     Exp<LSL 0.027%
Center = 19.57674        LSL = 0         Cp_l  = 1.15      Exp>USL 1.4%
StdDev = 5.657436        USL = 32        Cp_u = 0.732      Obs<LSL 0%
                                         Cp_k = 0.732      Obs>USL 0.99%
                                         Cpm  = 0.797
```

Figure 30 - Process Capability of Laptops

The process mean of 19.58 hours is above the target of 16 hours. The specification limits are set between 0 and 32 hours. The calculated Potential Capacity (Cp) = 0.943 and Process Capability Index (Cpk) = 0.732 are both below 1, indicating that the process is not capable of consistently meeting the required delivery-time standards. A Cp of 0.943 indicates that the variation in the process is nearly within the specification limits but still somewhat exceeds the acceptable tolerance range. The lower Cpk value points to the fact that the process is not aligned with the target and generally functions at a slower pace than anticipated. About 0.99% of the observations exceed the upper specification limit (USL), reflecting a minor risk of delays in laptop order deliveries.

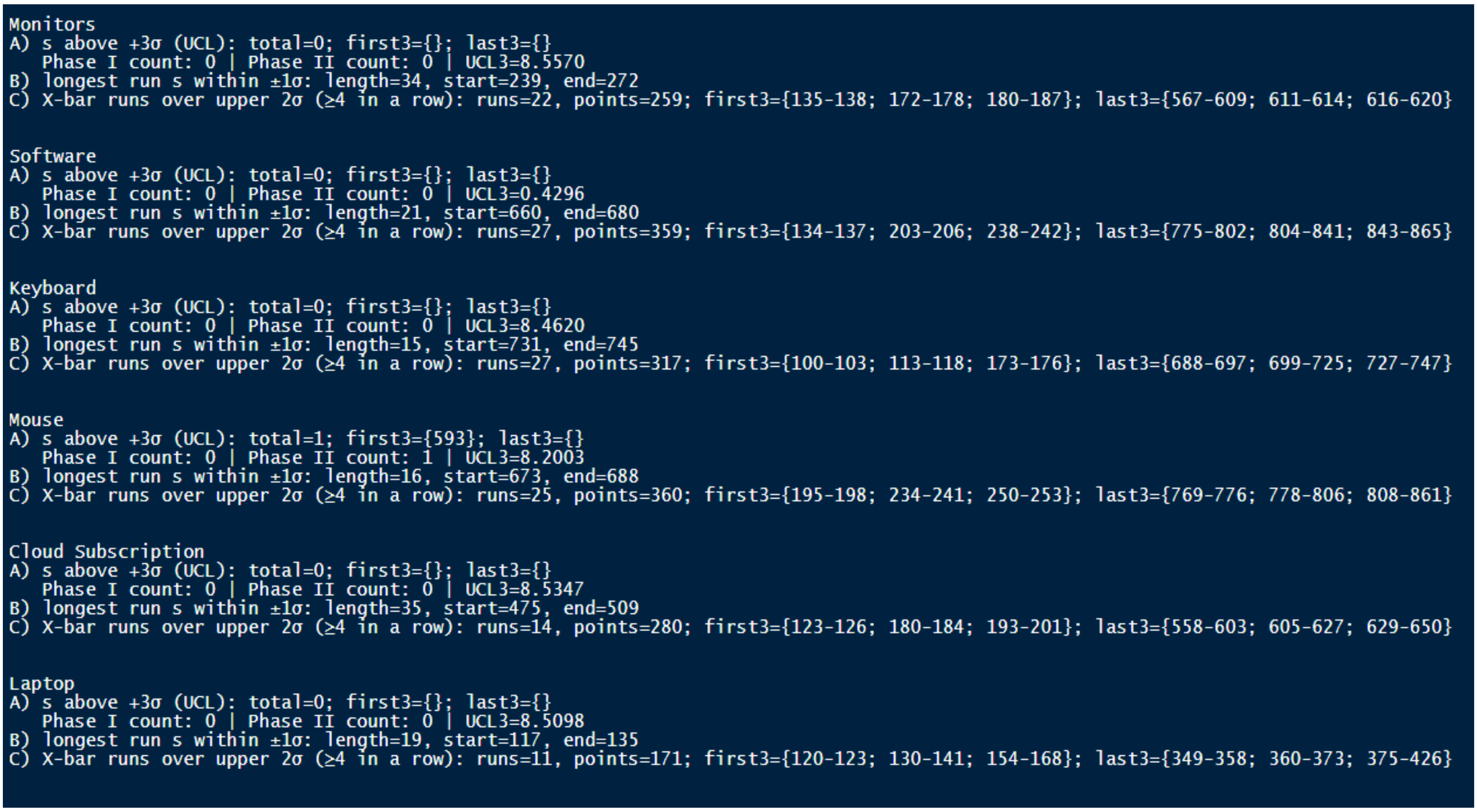
3.4 Identification of Process Control Issues

The data was analysed to identify control issues based on the following rules:
Rule A: (1 s sample outside of the upper +3 sigma-control limits for all product types). This indicates that the variation in the process has briefly gone beyond the anticipated limits and may be out of control. This implies that there may be a specific cause, like an unexpected operational disruption, a malfunction in equipment, or a shift in working conditions, that has raised the variability past the normal process boundaries.

Rule B: (the most consecutive samples of s between the -1 and +1 sigma-control limits for all product types). This indicates a high level of process stability and minimal variability. This trend demonstrates a system functioning under statistical control, where variation is random and mainly attributed to common causes instead of special or external factors.

Rule C: (4 consecutive X-bar samples outside of the upper, second control limits for all product types). This indicates a continuous upward movement in the mean of the process. This pattern implies that the process has deviated from its desired target, indicating a change in the average level of performance, often caused by consistent factors like heightened workload, process variation, or calibration mistakes.

The results:



```
Monitors
A) s above +3σ (UCL): total=0; first3={}; last3={}
   Phase I count: 0 | Phase II count: 0 | UCL3=8.5570
B) longest run s within ±1σ: length=34, start=239, end=272
C) X-bar runs over upper 2σ (≥4 in a row): runs=22, points=259; first3={135-138; 172-178; 180-187}; last3={567-609; 611-614; 616-620}


Software
A) s above +3σ (UCL): total=0; first3={}; last3={}
   Phase I count: 0 | Phase II count: 0 | UCL3=0.4296
B) longest run s within ±1σ: length=21, start=660, end=680
C) X-bar runs over upper 2σ (≥4 in a row): runs=27, points=359; first3={134-137; 203-206; 238-242}; last3={775-802; 804-841; 843-865}


Keyboard
A) s above +3σ (UCL): total=0; first3={}; last3={}
   Phase I count: 0 | Phase II count: 0 | UCL3=8.4620
B) longest run s within ±1σ: length=15, start=731, end=745
C) X-bar runs over upper 2σ (≥4 in a row): runs=27, points=317; first3={100-103; 113-118; 173-176}; last3={688-697; 699-725; 727-747}


Mouse
A) s above +3σ (UCL): total=1; first3={593}; last3={}
   Phase I count: 0 | Phase II count: 1 | UCL3=8.2003
B) longest run s within ±1σ: length=16, start=673, end=688
C) X-bar runs over upper 2σ (≥4 in a row): runs=25, points=360; first3={195-198; 234-241; 250-253}; last3={769-776; 778-806; 808-861}


Cloud Subscription
A) s above +3σ (UCL): total=0; first3={}; last3={}
   Phase I count: 0 | Phase II count: 0 | UCL3=8.5347
B) longest run s within ±1σ: length=35, start=475, end=509
C) X-bar runs over upper 2σ (≥4 in a row): runs=14, points=280; first3={123-126; 180-184; 193-201}; last3={558-603; 605-627; 629-650}


Laptop
A) s above +3σ (UCL): total=0; first3={}; last3={}
   Phase I count: 0 | Phase II count: 0 | UCL3=8.5098
B) longest run s within ±1σ: length=19, start=117, end=135
C) X-bar runs over upper 2σ (≥4 in a row): runs=11, points=171; first3={120-123; 130-141; 154-168}; last3={349-358; 360-373; 375-426}
```

Rule A indicates that there are no violations for any product type, except for the Mouse category, which experienced a single instance exceeding the upper control limit (UCL = 8.20). This suggests that, for most products, process variation is generally well-managed.

Rule B demonstrates internal stability and showcases how reliably each process functions under normal variation. The findings reveal that the Cloud Subscription process recorded the longest stable sequence of 35 consecutive samples, followed by Monitors (34) and Software (21). Physical product lines such as Keyboards and Mice showed shorter stable sequences (15–16 samples), suggesting slightly more variability in their operations. These observations imply that digital service processes (e.g., software and cloud) tend to display tighter control compared to physical product deliveries.

Rule C unveils the most critical area of concern. Numerous occurrences of this nature were noted across all product types, with between 11 and 27 instances per category. Monitors, Keyboards, and Software recorded the highest frequency of such runs, indicating repeated upward shifts in the mean and extended periods where delivery times surpassed expected levels.

In conclusion, Rules A and B validate that process variability is largely under control, while Rule C brings attention to frequent mean shifts and instability across various product types. This combination implies that, although the magnitude of variation remains steady, the average delivery performance trends upward over time. Consequently, corrective actions should focus on enhancing process centring and operational efficiency rather than merely minimising overall variation.

Part 4

4.1 The Likelihood of Making a Type 1 Error

A Type I Error (also called a false alarm or manufacturer's error) occurs when the control chart signals that a process is out of control even though it is actually operating normally. Under the null hypothesis, the process mean is assumed to be on target, and the statistic plotted on the chart follows a standard normal distribution

Therefore, probabilities for out-of-control signals can be estimated directly from the standard normal distribution using the cumulative probability function. An α-value of 0.05 (5%) was chosen, meaning we are willing to accept a 5% chance of rejecting the H0 when it is true.

Rule A

Rule A signals an out-of-control condition if any single S-chart sample falls above the +3σ upper control limit (UCL). Since only the upper tail of the distribution is relevant (S cannot be negative), this is a one-sided test.

The following calculation was used $\alpha \mathrm{A} = \mathrm{P(Z > 3)} = 1 - \Phi(3)$

From the standard normal tables $\Phi(3) = 0.99865$, therefor $\alpha \mathrm{A} = 1 - 0.99865 = 0.00135$

Thus, there is a 0.135 % chance that any one subgroup will falsely signal an alarm even when the process is in control.

The expected number of false alarms can then be calculated with N subgroups by:

$$\mathrm{E[false\ A\ signals]} = N \times \alpha \mathrm{A}$$

Rule B

Rule B checks for long sequences of S-values that lie within the ±1σ zone, which indicates good process stability.

The probability that a single sample falls inside ±1σ is: $\mathrm{P}(|Z| \leq 1) = \Phi(1) - \Phi(-1)$

From the standard normal tables $\Phi(1) - \Phi(-1) = 0.8413 - 0.1587 = 0.6827$

This means that approximately 68.27 % of all in-control points are expected to lie within ±1σ.

For a run of r consecutive points within ±1σ, the probability can be calculated by:

$$P(r\ in\ a\ row) = (0.6827)^r$$

Because such runs signify in-control behaviour rather than a violation, the Type I Error for this rule is defined as zero (α ≈ 0).

Rule C

Rule C signals an alarm when four or more consecutive sample means all fall above the +2σ line on the X-chart.

Under the null hypotheses, the probability that any one subgroup mean exceeds +2σ is the one-tailed probability of the standard normal distribution.

$$P(Z > 2) = 1 - \Phi(2)$$

From the standard normal tables $1 - \ \Phi(2) = 1 - 0.97725 = 0.02275$

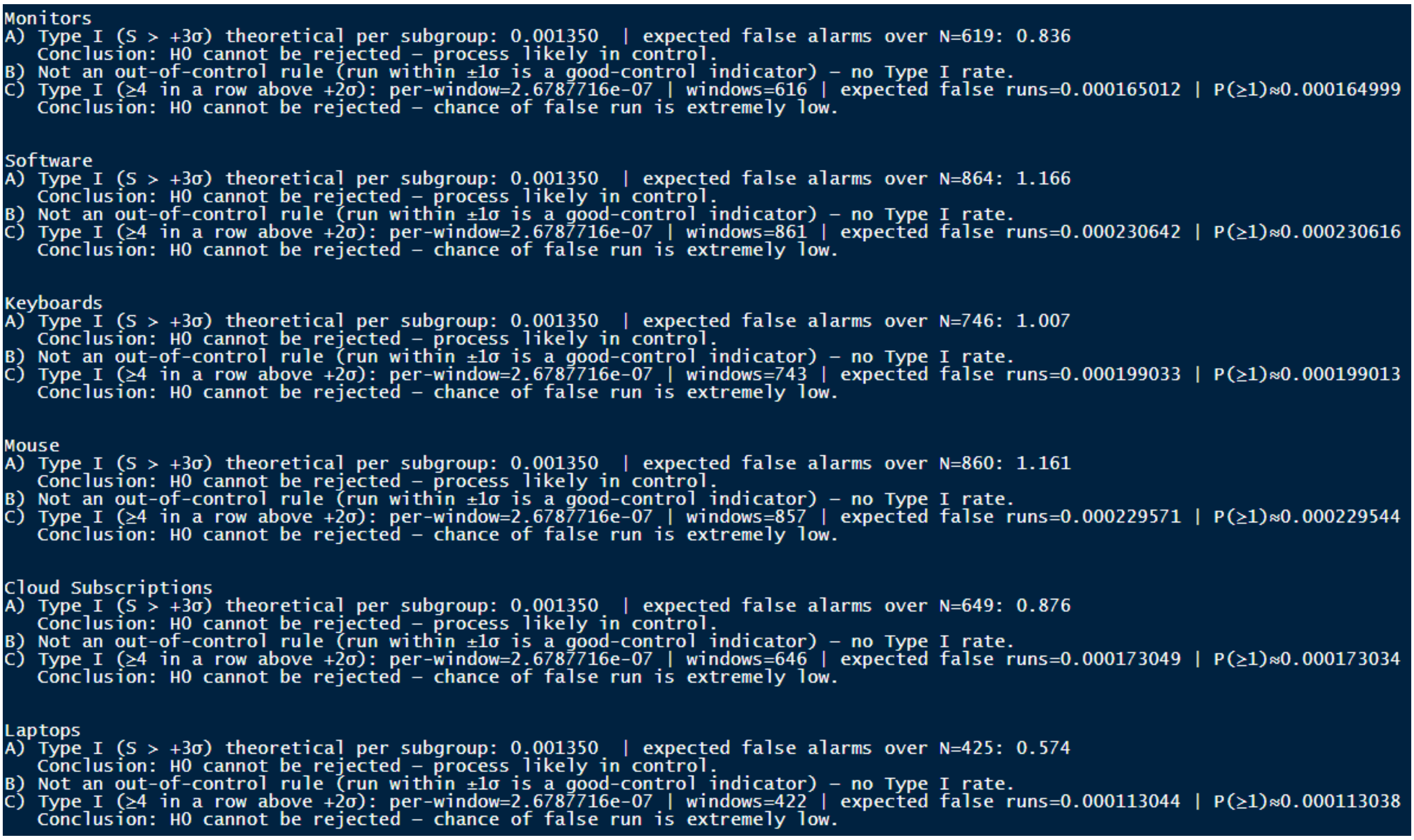
For 4 such consecutive points, assuming independence: $\alpha C = (0.02275)^4 = 2.68 \ \times 10^{-7}$

This implies a false-alarm rate of roughly 3 in 10 million, meaning the probability of observing such a run in an in-control process is practically zero.

If N subgroups exist, the number of four-point “windows” is $N - 3$, and the expected number of false Rule C runs is:

$$E[\text{false C runs}] = (N - 3) \times \alpha\text{C}$$

Results:



```
Monitors
A) Type I (S > +3σ) theoretical per subgroup: 0.001350  | expected false alarms over N=619: 0.836
   Conclusion: H0 cannot be rejected – process likely in control.
B) Not an out-of-control rule (run within ±1σ is a good-control indicator) – no Type I rate.
C) Type I (≥4 in a row above +2σ): per-window=2.6787716e-07 | windows=616 | expected false runs=0.000165012 | P(≥1)≈0.000164999
   Conclusion: H0 cannot be rejected – chance of false run is extremely low.


Software
A) Type I (S > +3σ) theoretical per subgroup: 0.001350  | expected false alarms over N=864: 1.166
   Conclusion: H0 cannot be rejected – process likely in control.
B) Not an out-of-control rule (run within ±1σ is a good-control indicator) – no Type I rate.
C) Type I (≥4 in a row above +2σ): per-window=2.6787716e-07 | windows=861 | expected false runs=0.000230642 | P(≥1)≈0.000230616
   Conclusion: H0 cannot be rejected – chance of false run is extremely low.


Keyboards
A) Type I (S > +3σ) theoretical per subgroup: 0.001350  | expected false alarms over N=746: 1.007
   Conclusion: H0 cannot be rejected – process likely in control.
B) Not an out-of-control rule (run within ±1σ is a good-control indicator) – no Type I rate.
C) Type I (≥4 in a row above +2σ): per-window=2.6787716e-07 | windows=743 | expected false runs=0.000199033 | P(≥1)≈0.000199013
   Conclusion: H0 cannot be rejected – chance of false run is extremely low.


Mouse
A) Type I (S > +3σ) theoretical per subgroup: 0.001350  | expected false alarms over N=860: 1.161
   Conclusion: H0 cannot be rejected – process likely in control.
B) Not an out-of-control rule (run within ±1σ is a good-control indicator) – no Type I rate.
C) Type I (≥4 in a row above +2σ): per-window=2.6787716e-07 | windows=857 | expected false runs=0.000229571 | P(≥1)≈0.000229544
   Conclusion: H0 cannot be rejected – chance of false run is extremely low.


Cloud Subscriptions
A) Type I (S > +3σ) theoretical per subgroup: 0.001350  | expected false alarms over N=649: 0.876
   Conclusion: H0 cannot be rejected – process likely in control.
B) Not an out-of-control rule (run within ±1σ is a good-control indicator) – no Type I rate.
C) Type I (≥4 in a row above +2σ): per-window=2.6787716e-07 | windows=646 | expected false runs=0.000173049 | P(≥1)≈0.000173034
   Conclusion: H0 cannot be rejected – chance of false run is extremely low.


Laptops
A) Type I (S > +3σ) theoretical per subgroup: 0.001350  | expected false alarms over N=425: 0.574
   Conclusion: H0 cannot be rejected – process likely in control.
B) Not an out-of-control rule (run within ±1σ is a good-control indicator) – no Type I rate.
C) Type I (≥4 in a row above +2σ): per-window=2.6787716e-07 | windows=422 | expected false runs=0.000113044 | P(≥1)≈0.000113038
   Conclusion: H0 cannot be rejected – chance of false run is extremely low.
```

4.2 The Likelihood of Making a Type 2 Error

A Type II error (consumers' risk) is the probability that the chart fails to signal even though the process has shifted. Here, we evaluate β for the X chart using the original (in-control) limits while the process mean and variability of X have changed

Hypotheses:

Ho: process is in control, centred at the original centre line (CL).

Hα: process has shifted to a new mean and variability.

The original X limits: LCL=25.011, CL=25.050, UCL=25.089

Shifted Process: μ1=25.028 and $\sigma_{\bar{X},1} = 0.017$

Formulae for $\boldsymbol{\beta}$:

$$z_L = \frac{LCL - \mu_1}{\sigma_{\widehat{X,1}}}\ ;\ z_U = \frac{UCL - \mu_1}{\sigma_{\widehat{X,1}}}\ ;\ \beta = \Phi(z_U) - \Phi(z_L)$$

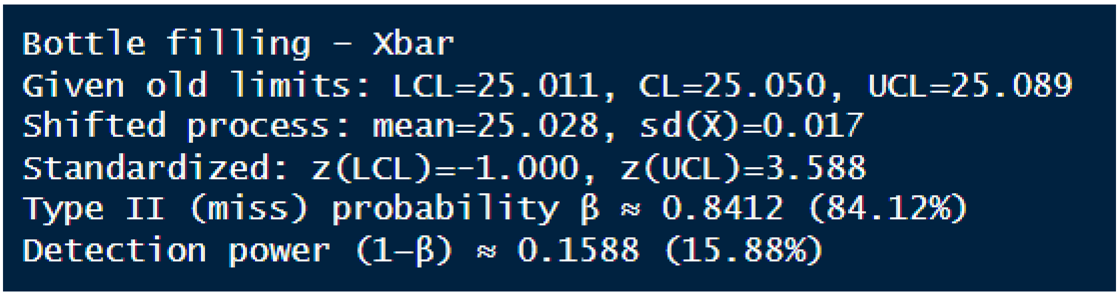
Substituting in:

$$z_L = -1\ ;\ z_U = 3.588\ ;\ \beta = \Phi(3.588) - \Phi(-1) \approx 0.99983 - 0.15865 = 0.8412$$

Hence, $\approx 84.12\%$, with the mean shifted to 25.028 and $\sigma_{\bar{X}}$increased to 0.017, the X chart will fail to detect the shift about 84% of the time using the old limits.

Detection Power is then calculated: $1 - \beta \approx 0.1588$

Using the original control limits, the shifted process would be accepted by the X chart about 84% of the time, implying low sensitivity (power ≈ 16%) to this particular shift.



```
Bottle filling - Xbar
Given old limits: LCL=25.011, CL=25.050, UCL=25.089
Shifted process: mean=25.028, sd(X̄)=0.017
Standardized: z(LCL)=-1.000, z(UCL)=3.588
Type II (miss) probability β ≈ 0.8412 (84.12%)
Detection power (1-β) ≈ 0.1588 (15.88%)
```

4.3 Fixed Head Office Data and reapplied Sales Analysis

During the initial analysis in Week 1, several inconsistencies were identified between the head-office product file (products_Headoffice.csv) and the local product data (products_data.csv). Product entries from item 11 to 60 of each product type contained incorrect ProductID prefixes labelled as "NA###", and their corresponding Selling Price and Markup values were mismatched.

The head office requested that these issues be corrected and that a new, clean dataset be produced for 2025.

To resolve this, two new files were created:

1. products_data2025.csv – updated from the local product data.
    - The Category column was synchronised with the ProductID prefix.
    - This ensured that each product's classification correctly reflected its three-letter product code.
2. products_Headoffice2025.csv – reconstructed from the head-office file.

- All "NA###" product IDs were replaced with the correct three-letter prefix derived from the product's Category.
- The six-character ID format (three letters + three digits) was preserved for all products.
- The Selling Price and Markup values for each product type were corrected by repeating the first ten valid price/markup combinations from products_data2025 across items 11 to 60.
- The Category column was then forced to match the new prefix, guaranteeing internal consistency.

Once all corrections were done, the sales analysis was redone and yielded the following results:

Total Sales per Year Rand Value Updated

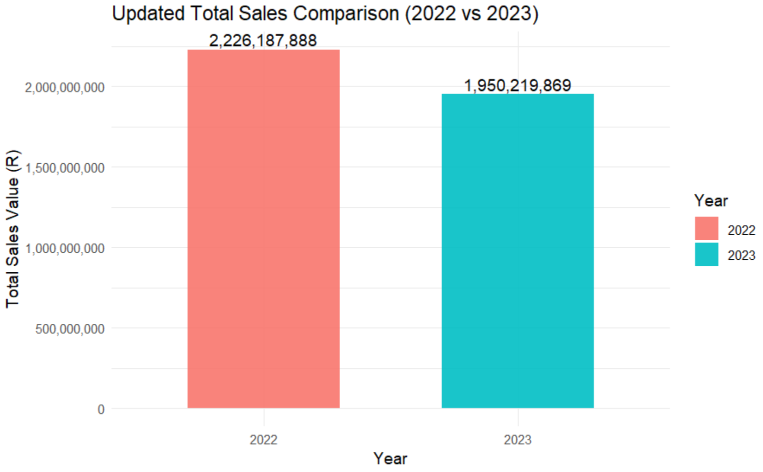


Figure 32 - Updated Sales Rand Value per Year

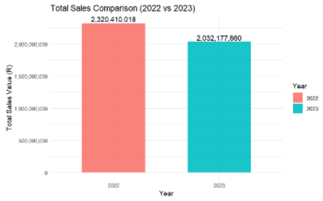


Figure 31 - Old Sales Rand Value

Following the update of the Head Office data, total sales for 2022 decreased from R2,320,410,018 to R2,226,187,888, while for 2023, sales declined from R2,032,177,660 to R1,950,219,869. This signifies a decrease of about 4% for both years, suggesting that the adjusted product prices and markups led to slightly lower, yet more precise overall sales figures.

Updated Sales per Category

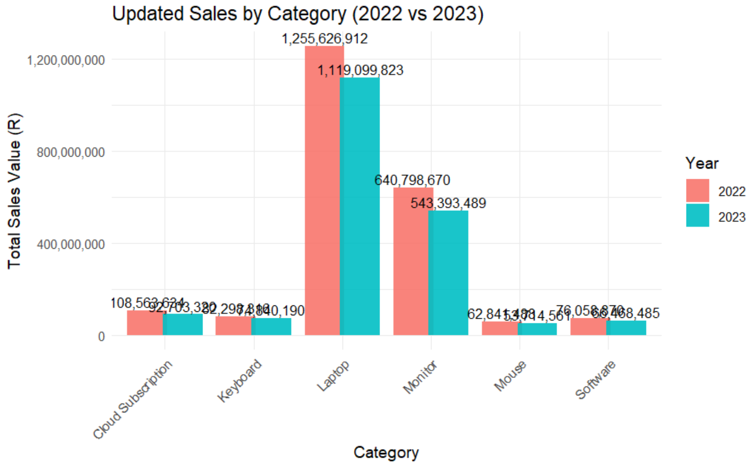


Figure 33 - Updated Sales per Category

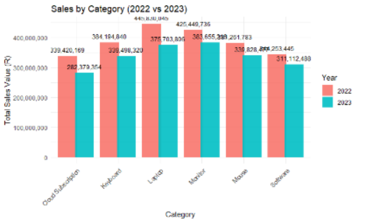


Figure 34 - Old Sales per Category

After correcting the Head-Office data, the overall sales distribution across categories changed noticeably. The laptop category remained the highest contributor, but its sales decreased from R1.26 billion to R1.12 billion in 2023. Monitor sales also declined from R640.8 million to R543.4 million, while smaller reductions occurred across the remaining product types. The updated data show lower totals in every category compared to the original dataset, reflecting the removal of inflated or duplicated price entries and resulting in more accurate category-level sales values.

Updated Gross Profit

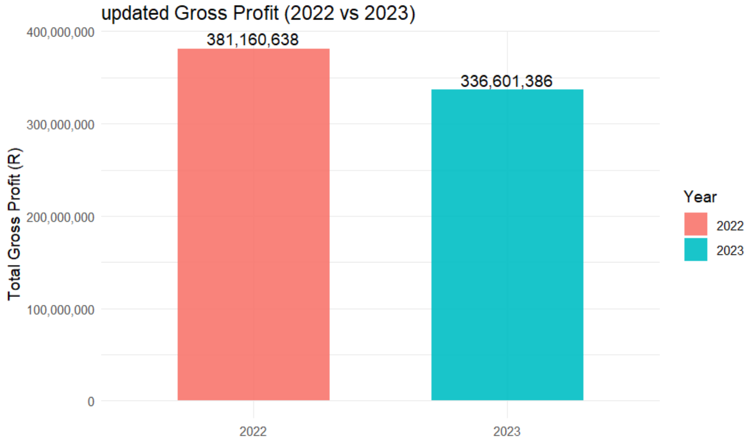


Figure 35 -Updated Gross Profit per Year

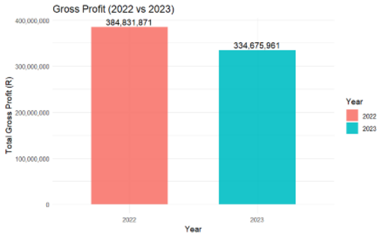


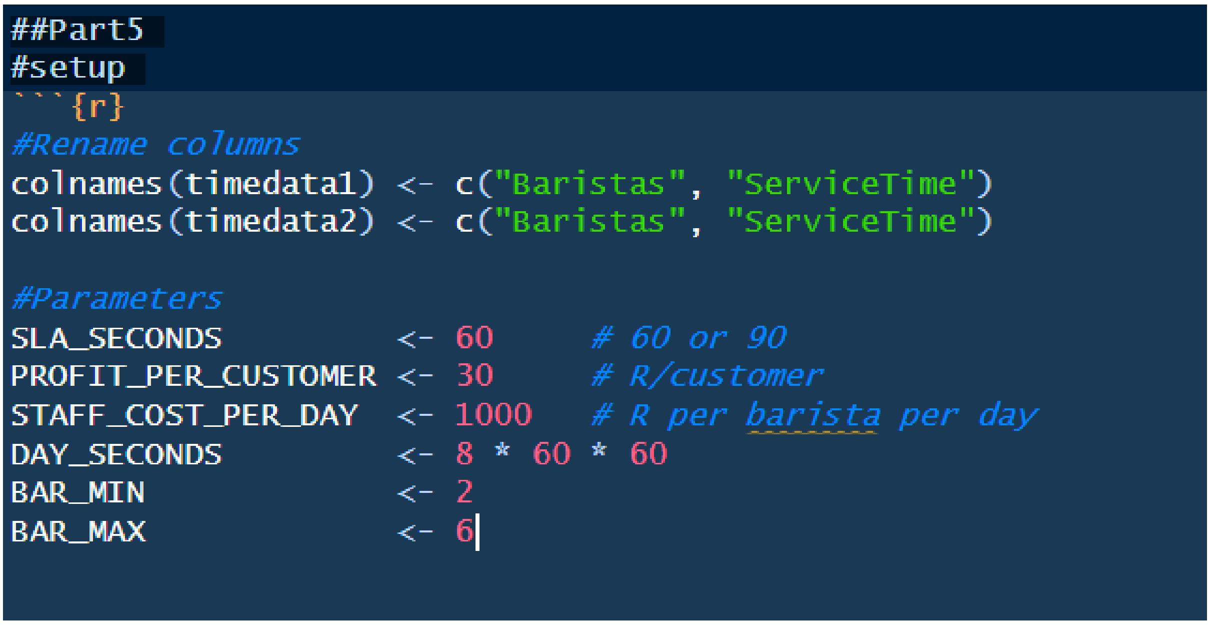
Figure 36 - Old Gross Profit per Year

After the data adjustments, the total gross profit for 2022 was revised down from R384,831,871 to R381,160,638, while the figure for 2023 increased from R334,675,961 to R336,601,386. The changes are minimal (less than 2% for both years) and are indicative of the more precise selling prices and markups utilised during the Head-Office data review. Despite the slight revisions, the overall trend from year to year remained stable, reflecting a slight decrease in gross profit from 2022 to 2023.

Part 5 Profit Optimisation

5.1 Procedures Used

To find the ideal number of baristas for maximizing profit, the average service time for each staffing level (ranging from 2 to 6 baristas) was initially calculated from the service-time dataset. With these averages, the estimated number of customers that could be served during an eight-hour workday was determined by dividing the total available time per day (28,800 seconds) by the average service time. Daily revenue was then calculated by multiplying the number of customers served by the profit per customer (R 30), while the staff cost was assessed as R 1,000 for each barista per day. The daily profit for each staffing level was obtained by subtracting staff costs from daily revenue. By plotting daily profit against the number of baristas, it became possible to visually identify the staffing level that generated the highest profit, indicating the optimal number of baristas for the shop given the specific operating conditions.

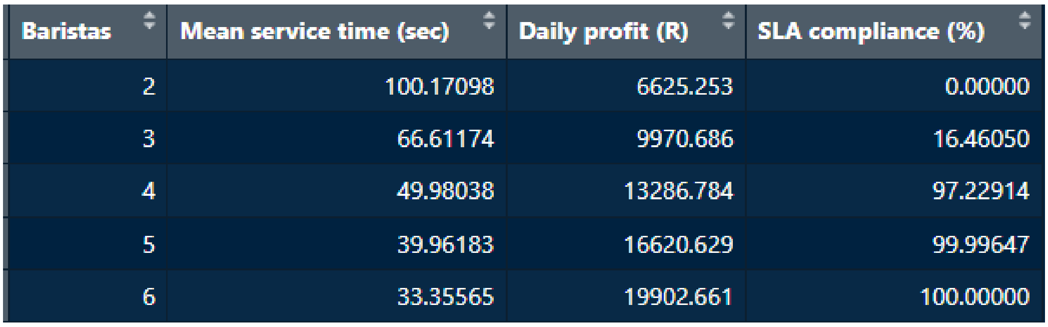


~~~
##Part5
#setup
```{r}
#Rename columns
colnames(timedata1) <- c("Baristas", "ServiceTime")
colnames(timedata2) <- c("Baristas", "ServiceTime")

#Parameters
SLA_SECONDS         <- 60     # 60 or 90
PROFIT_PER_CUSTOMER <- 30     # R/customer
STAFF_COST_PER_DAY  <- 1000   # R per barista per day
DAY_SECONDS         <- 8 * 60 * 60
BAR_MIN             <- 2
BAR_MAX             <- 6
~~~

We set the service level agreement to 60 seconds and determined the percentage of SLA compliance with the various number of baristas.

5.2 Shop 1



Baristas	Mean service time (sec)	Daily profit (R)	SLA compliance (%)
2	100.17098	6625.253	0.00000
3	66.61174	9970.686	16.46050
4	49.98038	13286.784	97.22914
5	39.96183	16620.629	99.99647
6	33.35565	19902.661	100.00000

Table 2 - Summary Table for Coffee Shop 1

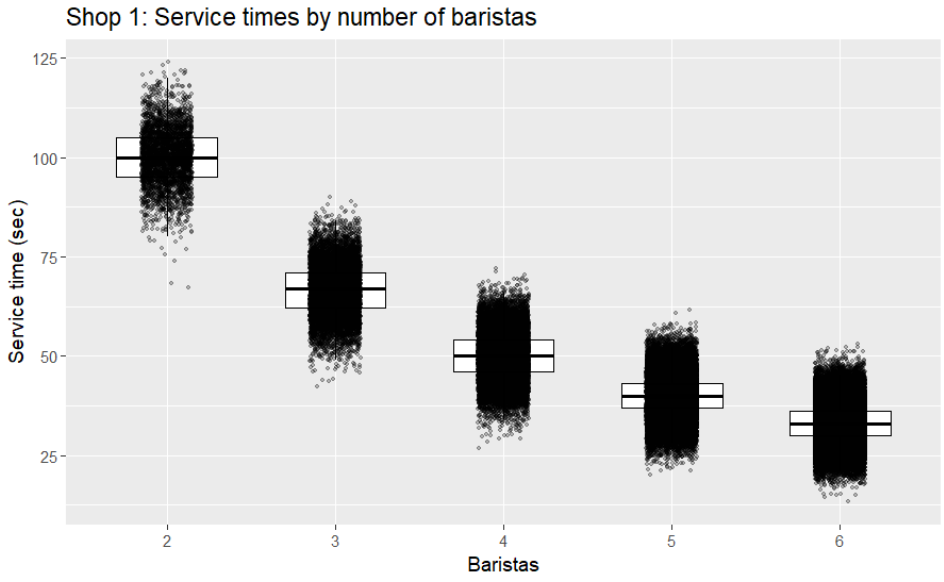


Figure 37 - Service Times by Number of Baristas Shop 1

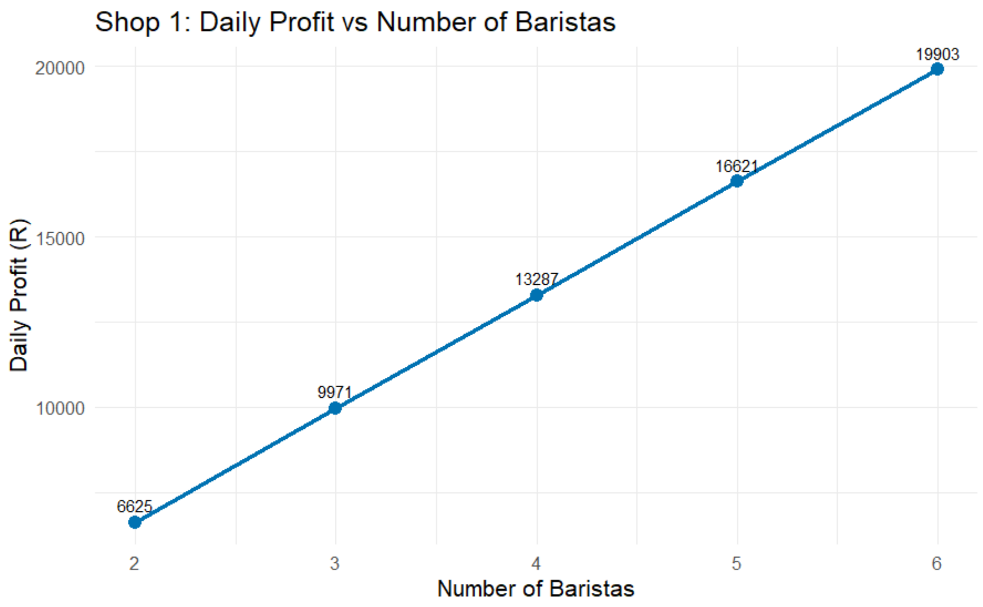
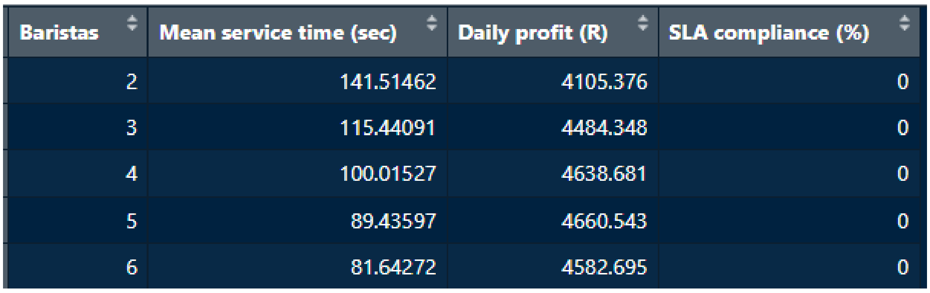


Figure 38 - Daily Profit vs Number of Baristas Shop 1

The analysis reveals a distinct connection between staffing levels, the speed of service, and overall profitability. As the number of baristas increased from 2 to 6, the average service time significantly decreased, while daily profits rose from R 6,625 to R 19,903. Additionally, the variability in service times diminished, demonstrating enhanced process consistency. Service-level compliance improved from 0% to 100%, meaning that every customer was attended to within the 60-second service level agreement when six baristas were on duty. While the addition of each barista does raise labour costs, the resulting increase in throughput and customer satisfaction justifies this expense. Consequently, employing six baristas is the most effective staffing choice for Shop 1, yielding the maximum profit and ensuring complete adherence to the service time agreement.

5.3 Shop 2



Baristas	Mean service time (sec)	Daily profit (R)	SLA compliance (%)
2	141.51462	4105.376	0
3	115.44091	4484.348	0
4	100.01527	4638.681	0
5	89.43597	4660.543	0
6	81.64272	4582.695	0

Table 3 - Summary Table for Coffee Shop 2

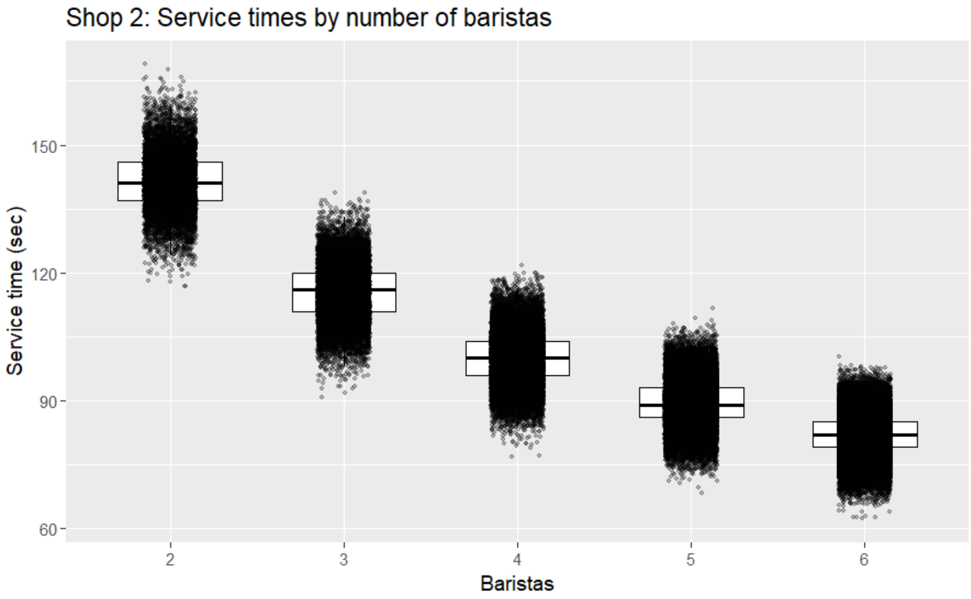


Figure 39 - Service Time by Number of Baristas Shop 2

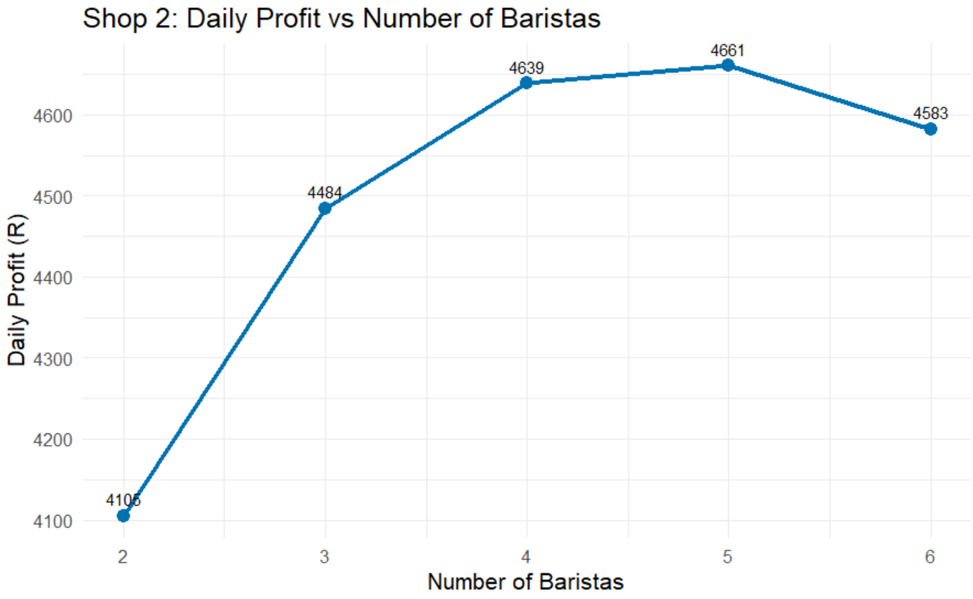


Figure 40 - Daily Profit vs Number of Baristas Shop 2

For Shop 2, the findings indicate that increasing the number of baristas lowers the average service time and results in a slight rise in daily profit. The mean service time decreases from approximately 142 seconds with two baristas to 82 seconds with six, while daily profit goes up from around R 4,100 to R 4,660 before slightly declining at six baristas. The optimal profit point occurs at five baristas, where the throughput is maximised without incurring excessive labour costs.

However, none of the staffing levels manage to meet the 60-second service level agreement (SLA) target, leading to 0% compliance across all scenarios. Even with six baristas, the average service time is still significantly above the SLA limit, suggesting that the slow performance is not solely due to staffing but may instead be influenced by workflow inefficiencies, extended preparation times, or layout issues that hinder quicker service.

In summary, while five baristas yield the highest daily profit for Shop 2, the store does not satisfy its reliability requirements under the existing operating conditions. To achieve service times within the 60-second SLA, a redesign of processes or improvements to equipment would be necessary.

Part 6 ANOVA Analysis

6.1 Hypothesis Selection and Objective

An analysis of variance (ANOVA) was conducted to determine whether delivery performance changed significantly across the twelve months of the year. The selected focus variable is delivery hours, indicating the average time in hours required for fulfilling customer orders. This selection was based on its direct correlation to both operational efficiency and service reliability, which ties back to the analyses of process capability and service reliability discussed previously.

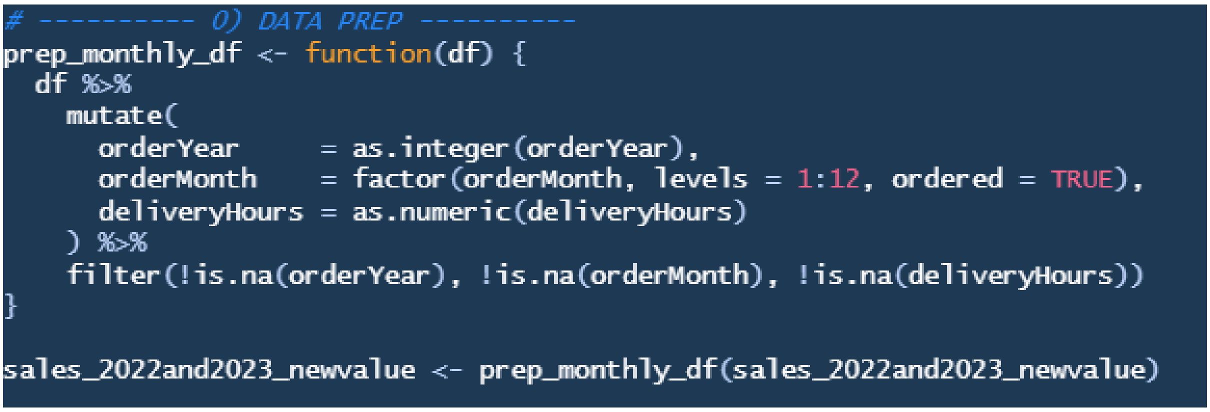
The hypotheses were formulated as follows, with a significance level of 5%:

Null hypothesis: The mean delivery hours are equal across all months, month-to-month variation in delivery time is due only to random noise.

Alternative hypothesis: At least one month’s mean delivery hours differs significantly from the others, indicating a possible seasonal or operational effect on delivery performance.

6.2 Procedure and method

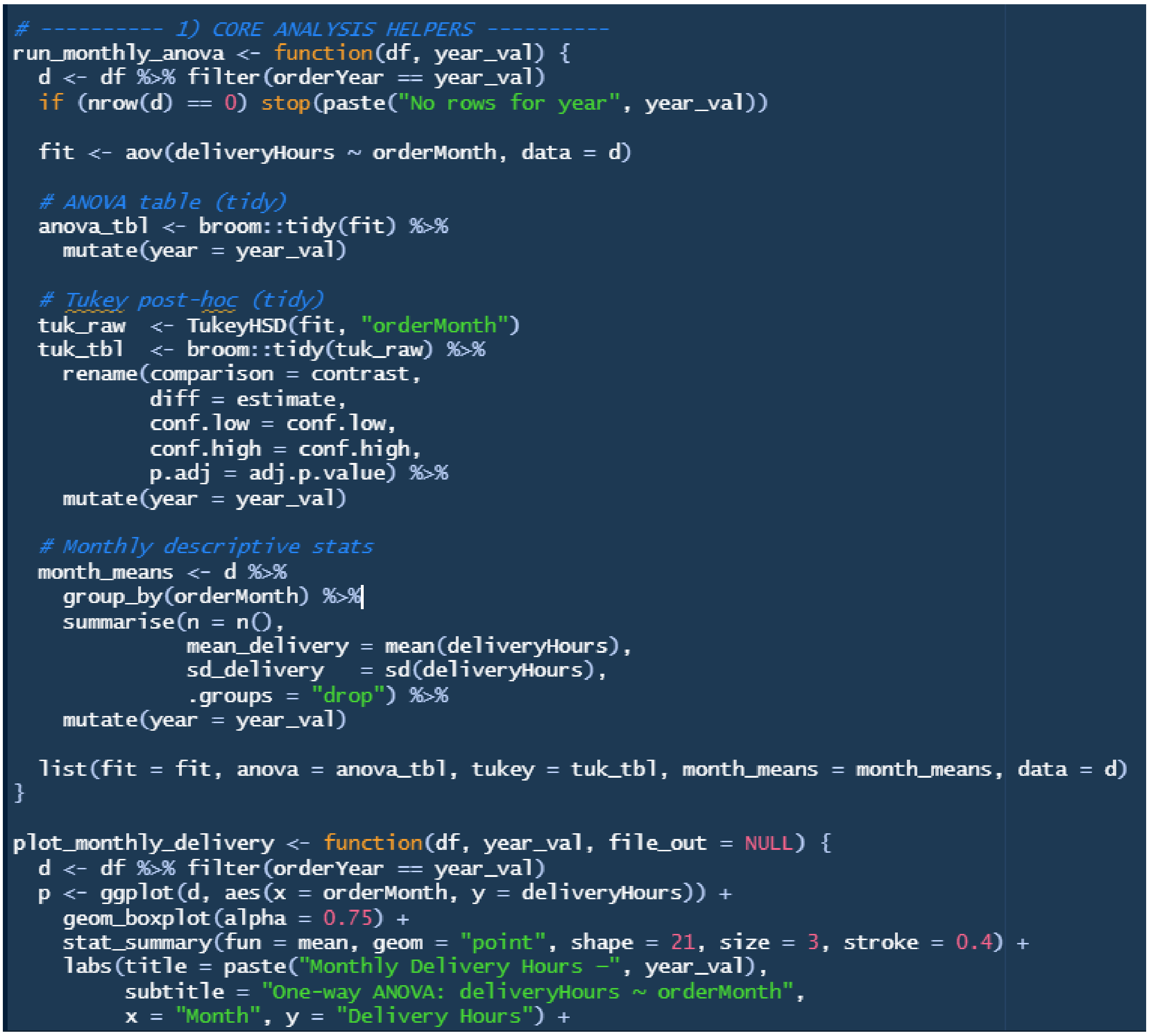
The analysis was using a one-way ANOVA model. The dataset sales_2022and2023_newvalue was first cleaned and formatted so that order month was a categorical factor with levels 1–12 and delivery hours were a numeric variable. A helper function was then created to automatically run the model for each year.



```
# ---------- 0) DATA PREP ----------
prep_monthly_df <- function(df) {
  df %>%
    mutate(
      orderYear     = as.integer(orderYear),
      orderMonth    = factor(orderMonth, levels = 1:12, ordered = TRUE),
      deliveryHours = as.numeric(deliveryHours)
    ) %>%
    filter(!is.na(orderYear), !is.na(orderMonth), !is.na(deliveryHours))
}

sales_2022and2023_newvalue <- prep_monthly_df(sales_2022and2023_newvalue)
```

For each year, an ANOVA table was generated, followed by a Tukey HSD post-hoc test to identify which months differed significantly. Descriptive statistics (sample size, mean, and standard deviation) were calculated for each month, and box-and-whisker plots were produced to visually assess variation in delivery hours.



```r
# ---------- 1) CORE ANALYSIS HELPERS ----------
run_monthly_anova <- function(df, year_val) {
  d <- df %>% filter(orderYear == year_val)
  if (nrow(d) == 0) stop(paste("No rows for year", year_val))

  fit <- aov(deliveryHours ~ orderMonth, data = d)

  # ANOVA table (tidy)
  anova_tbl <- broom::tidy(fit) %>%
    mutate(year = year_val)

  # Tukey post-hoc (tidy)
  tuk_raw  <- TukeyHSD(fit, "orderMonth")
  tuk_tbl  <- broom::tidy(tuk_raw) %>%
    rename(comparison = contrast,
           diff = estimate,
           conf.low = conf.low,
           conf.high = conf.high,
           p.adj = adj.p.value) %>%
    mutate(year = year_val)

  # Monthly descriptive stats
  month_means <- d %>%
    group_by(orderMonth) %>%
    summarise(n = n(),
              mean_delivery = mean(deliveryHours),
              sd_delivery   = sd(deliveryHours),
              .groups = "drop") %>%
    mutate(year = year_val)

  list(fit = fit, anova = anova_tbl, tukey = tuk_tbl, month_means = month_means, data = d)
}

plot_monthly_delivery <- function(df, year_val, file_out = NULL) {
  d <- df %>% filter(orderYear == year_val)
  p <- ggplot(d, aes(x = orderMonth, y = deliveryHours)) +
    geom_boxplot(alpha = 0.75) +
    stat_summary(fun = mean, geom = "point", shape = 21, size = 3, stroke = 0.4) +
    labs(title = paste("Monthly Delivery Hours –", year_val),
         subtitle = "One-way ANOVA: deliveryHours ~ orderMonth",
         x = "Month", y = "Delivery Hours") +
```

All analyses were executed using the tidyverse and broom packages for data manipulation and tidy model outputs, and the resulting tables and plots were exported as CSV and PNG files.

6.3 Results

Box Plots for 2022 and 2023

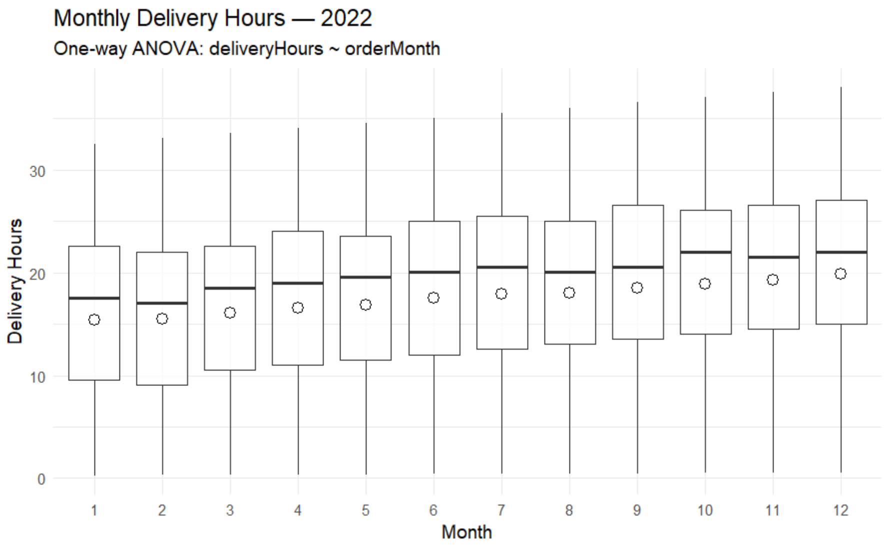


Figure 41 - Box plots of Monthly Delivery Hours

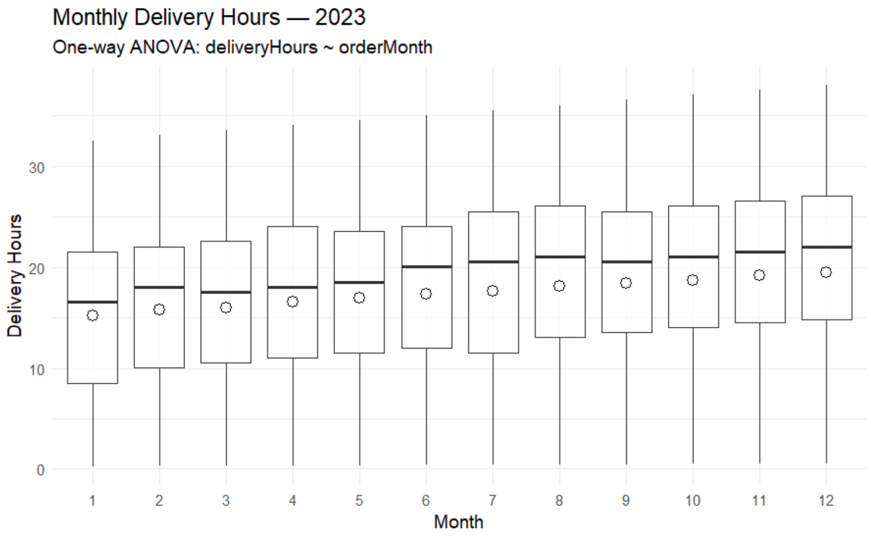
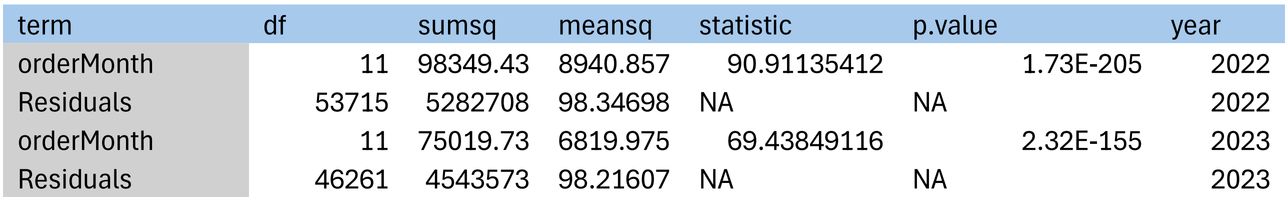


Figure 42 - Box Plots of Monthly Delivery Hours

Delivery performance tends to remain consistent from one month to the next, although both years exhibit a slight increase in delivery hours, which may be associated with seasonal workload increases or slight operational inefficiencies over time.

ANOVA



term	df	sumsq	meansq	statistic	p.value	year
orderMonth	11	98349.43	8940.857	90.91135412	1.73E-205	2022
Residuals	53715	5282708	98.34698	NA	NA	2022
orderMonth	11	75019.73	6819.975	69.43849116	2.32E-155	2023
Residuals	46261	4543573	98.21607	NA	NA	2023

Table 4 - ANOVA Output

For both years, the p-values were below the 0.05 significance, hence, we reject the null hypothesis in both cases. This indicates that there are measurable differences in delivery hours between some months.

However, the boxplots for 2022 show only minimal variation, suggesting that the differences, while statistically significant, are not operationally meaningful. In contrast, 2023 displays a clearer upward trend in delivery hours toward the later months of the year, indicating that the differences are both statistically and practically significant.

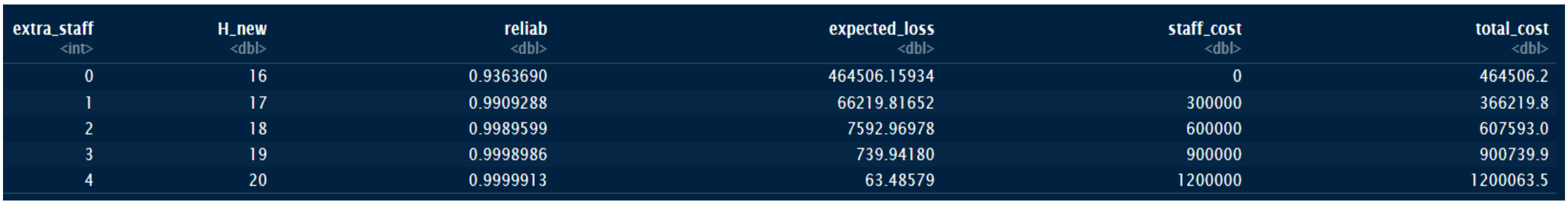
Tukey HSD Post-Hoc Test

The Tukey HSD post-hoc test was performed to identify which specific months differed significantly in their mean delivery hours. For 2022, very few month pairs showed significant differences, supporting the earlier ANOVA and boxplot findings that delivery times were consistent throughout the year. In contrast, 2023 exhibited several statistically significant month-to-month differences, particularly between early-year months (January–March) and later months (June–August and November–December). These later months recorded notably higher delivery hours, confirming a gradual increase in service time as the year progressed.

Overall, the Tukey analysis reinforces the conclusion that 2023 experienced greater seasonal variability in delivery performance, likely due to increased workload or reduced operational efficiency during peak demand periods.

Part 7 Reliability of Service

An analysis was conducted on data from a car rental agency over a span of 397 days. The data captured the number of employees working each day, with reliable service defined as having 15 or more staff members present. Initially, we found that there were 366 days categorized as reliable out of 397, which translates to 92.2%, or approximately 337 reliable days in a year. To assess the impact of hiring more employees on costs, we modelled the presence of staff on duty as a binomial process, where each employee has a constant probability of showing up for work on any particular day. Utilizing this model, we calculated the likelihood of providing reliable service for various staffing levels and integrated this information with the company's cost framework (R 20,000 lost revenue for each unreliable day and R 25,000 monthly for each employee). Subsequently, we determined the total annual (or monthly) expenses by adding the expected daily losses to the staff salaries. The total cost vs the number of workers was then plotted to visually identify the optimal balance between labour costs and reliability



extra_staff <int>	H_new <dbl>	reliab <dbl>	expected_loss <dbl>	staff_cost <dbl>	total_cost <dbl>
0	16	0.9363690	464506.15934	0	464506.2
1	17	0.9909288	66219.81652	300000	366219.8
2	18	0.9989599	7592.96978	600000	607593.0
3	19	0.9998986	739.94180	900000	900739.9
4	20	0.9999913	63.48579	1200000	1200063.5

Table 5 - Total Cost Calculation Table

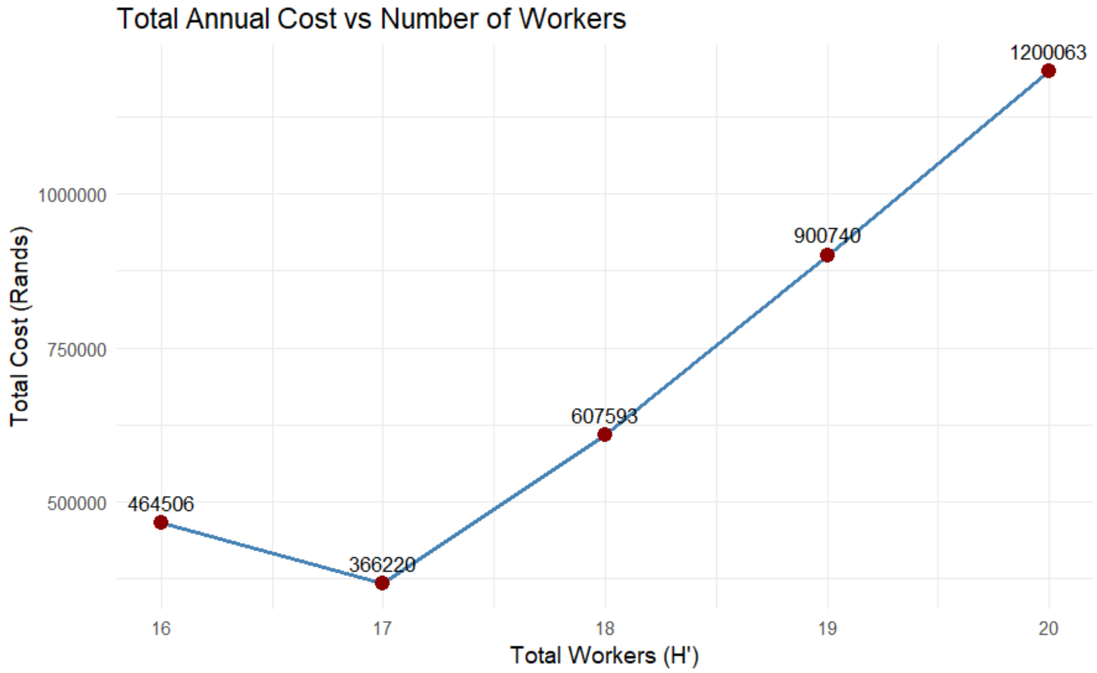
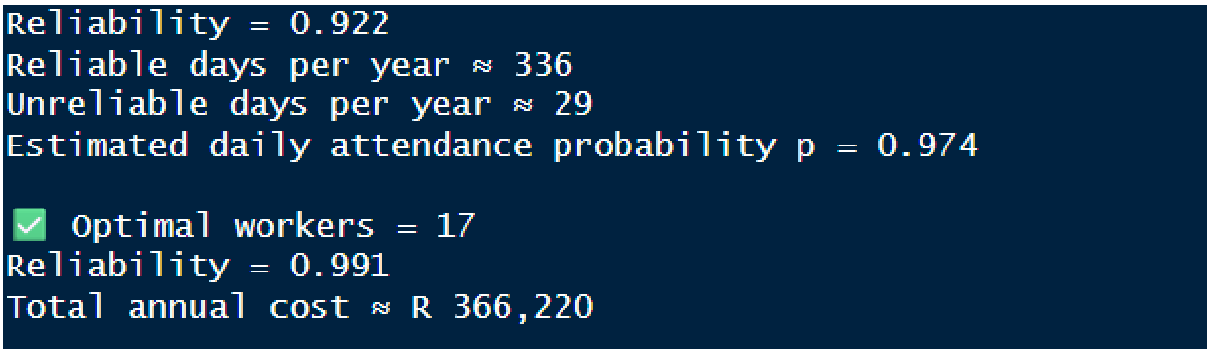


Figure 43 - Total Annual Cost vs Number of Workers



```
Reliability = 0.922
Reliable days per year ≈ 336
Unreliable days per year ≈ 29
Estimated daily attendance probability p = 0.974

✅ Optimal workers = 17
Reliability = 0.991
Total annual cost ≈ R 366,220
```

The analysis indicates that the existing number of 16 employees offers a reliability of roughly 92.2%, which translates to 336 dependable days annually. By modelling the system as a binomial process and assessing the cost trade-offs, it was discovered that increasing the workforce to 17 employees boosts reliability to 99.1% while also resulting in the lowest total annual expense of approximately R366,000. Beyond this number, hiring additional staff yields only slight improvements in reliability but leads to a significant increase in total costs due to higher wage expenses. The resulting cost curve clearly demonstrates a U-shape, highlighting that both under-staffing and over-staffing lead to increased losses. This pattern aligns with the Taguchi Loss Principle, which posits that losses multiply as a process deviates from its optimal state. In this scenario, the ideal condition represents the most cost-effective balance between service reliability and operating expenses. This is achieved with a workforce of 17 employees.

Reference List

ChatGPT - Quality assurance project. n.d. Available: https://chatgpt.com/g/g-p-68f67f0e12908191b5b1c19ad98b3b3f-quality-assurance-project/project [2025, October 25].

Mannino, S. 2023. Control Charts and Process Capability. 117–138. DOI: 10.1007/978-3-031-11724-4_6.

MANOVA Test in R: Multivariate Analysis of Variance - Easy Guides - Wiki - STHDA. n.d. Available: https://www.sthda.com/english/wiki/manova-test-in-r-multivariate-analysis-of-variance#google_vignette [2025, October 25].

Taguchi Loss Function - Lean Manufacturing and Six Sigma Definitions. n.d. Available: https://www.leansixsigmadefinition.com/glossary/taguchi-loss-function/ [2025, October 25].