ECSA PROJECT

Final Deliverable

OCTOBER 20, 2025

HANDRI VILJOEN
27347737

Contents

Table of Figure

List of Tables

List of Equations

Abstract:

This report presents a comprehensive analysis of customer behavior, product performance, sales, trends and operational efficiencies for the years 2022 and 2023. Four datasets were examined, including customer demographics, local product data, head office product data, and sales transactions. Data relationships were established to identify patterns, highlight data issues, and provide recommendations. The analysis revealed key insights into revenue contributions by customer age and product category, seasonal sales trends, and process control issues affecting delivery performance. Additionally, operational data from two coffee shops was evaluated to determine optimal staffing levels for reliability and profitability. Data cleaning and alignment of product information were applied to improve insights. The findings inform strategic decision-making for pricing, inventory management and staffing.

Introduction:

Data analytics allows for the identification of patterns, trends and relationships within complex datasets. This report applies data-driven methods to examine operational and performance data, integrating qualitative and quantitative insights to provide a comprehensive perspective for well-informed decision-making in industry. By combining trend analysis and optimization techniques, the study highlights key patterns and offers actionable insights for outcome improvements.

Data Inspection:

Four datasets were analyzed:

Dataset:	Description:	Key Variables:
Customer Data	Demographics information about customers	Customer ID, Gender, Age, Income, City
Product Data	Local product Information	Product ID, Category, Description, Selling Price, Markup.
Products Head Office	Master Product list from HO	Product ID, Category, Description, Selling Price, Markup.
Sales Data	Transaction data for 2022 and 2023	Customer ID, Product ID, Quantity, Order Date, Picking Hours, Delivery Hours

Table 1: Table of Datasets

The customers dataset has 5000 rows and 5 columns of which the categorical variables include Gender and City, while numerical variables include Age and Income.

Products have 60 rows and 5 columns, where only selling price and Marup are numeric, with the rest being categorical.

Sales have 100 000 rows and 9 columns. It contains numeric features (order time, quality and picking/delivery hours) and categorical features (CustomerID and ProductID)

Products Head Office data contains 360 rows and 5 columns. It mirrors the product dataset but represents the official catalogue.

Relationship between datasets:

Customer and sales data were linked by means of Customer ID, while product data and sales were linked through Product ID. Product data from Head office was used to identify discrepancies with local branch records.

Customer Information:

A business provided information of 5000 clients, including the customer ID, gender, age, income and city of residence. The table below is a sample of this data.

	CustomerID		Gender	Age	Income	City
1	CUST001		Male	16	65000	New York
2	CUST002		Female	31	20000	Houston
3	CUST003		Male	29	10000	Chicago

Table 2: Customer dataset insert

Customer gender distribution:

From the customer gender distribution bar graph in figure 1, it is clear that the distribution between male and female customers is very similar as 2432 customers are female and 2350 are male, with only 1.64% difference between the two. It is worth noting that 4.36 % of customers (218) are registered as other and could be a result of data collection errors or customers that choose not to disclose gender information.

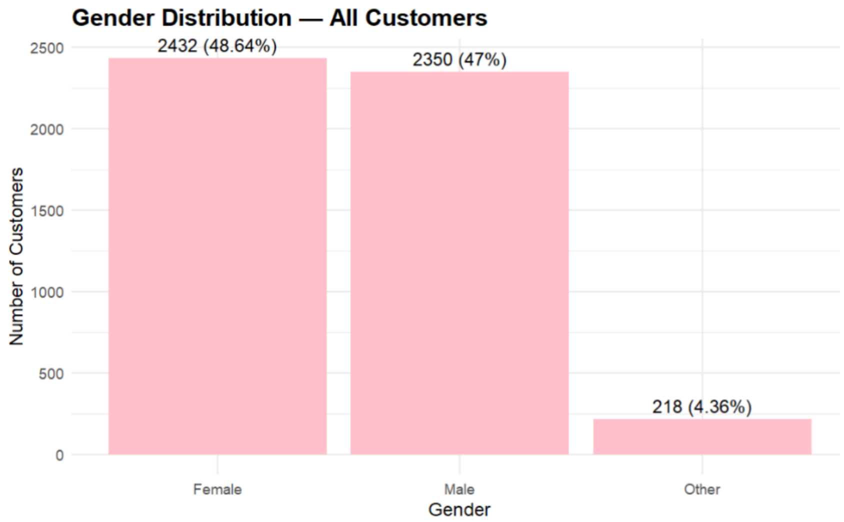


Figure 1: Customer Gender Distribution

Income distribution by gender:

Figure 2 below shows the distribution of customer income by gender. The income of female customers is slightly higher than that of male customers. The Q1, Q3 and 1.5 IQR spread is similar for all three categories, indicating that the distribution tails (the lower and upper) are comparable. The primary difference between the datasets lies in the means values.

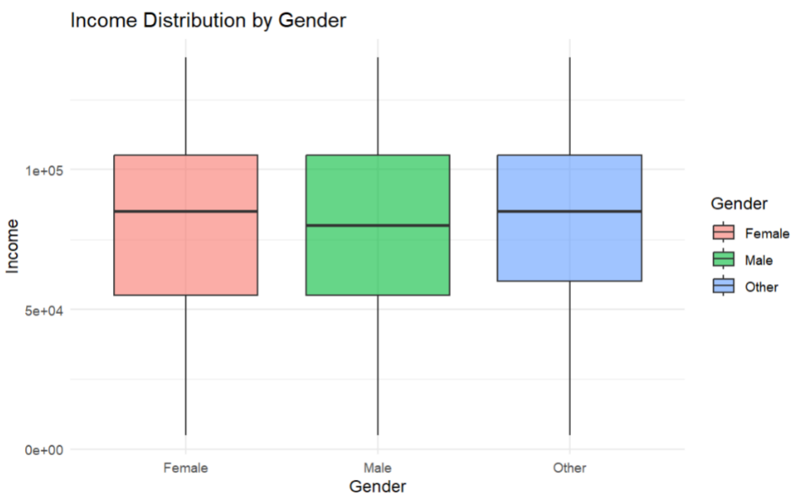


Figure 2: Income distribution by Gender

Revenue by age group

Analysis of figure 3 of the total revenue by customer age group shows that senior customers (61+) contribute the largest share of 44%, followed by adults (41-60) at 28.5%. Young adults (20-40) account for 22.1% while youth (0-20) contribute the smallest portion of 9.6%. This indicates that older customers re the primary drivers of revenue, and thus marketing and product strategies should be tailored according

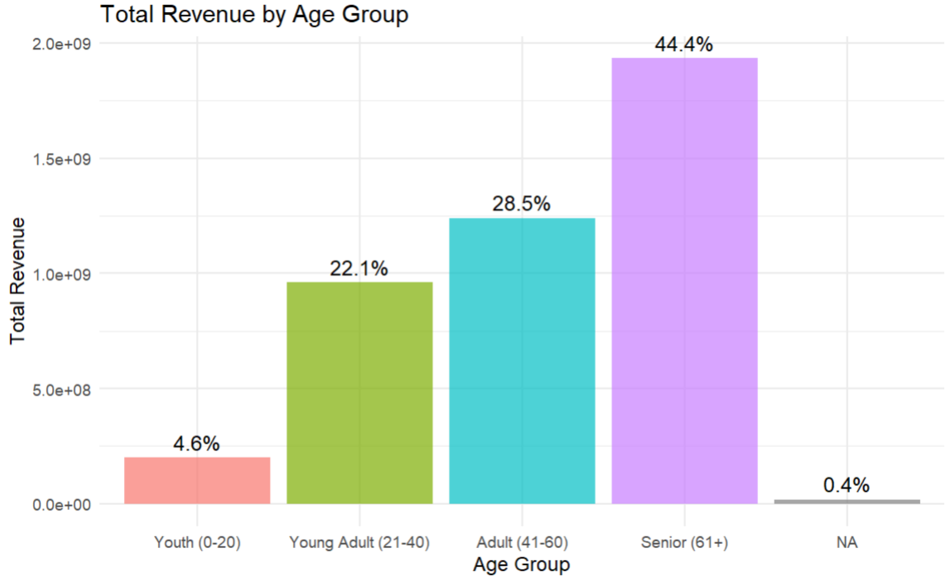


Figure 3: Total Revenue by Age Group

Income analysis by age group

By analyzing figure 4, the following can be deduced:

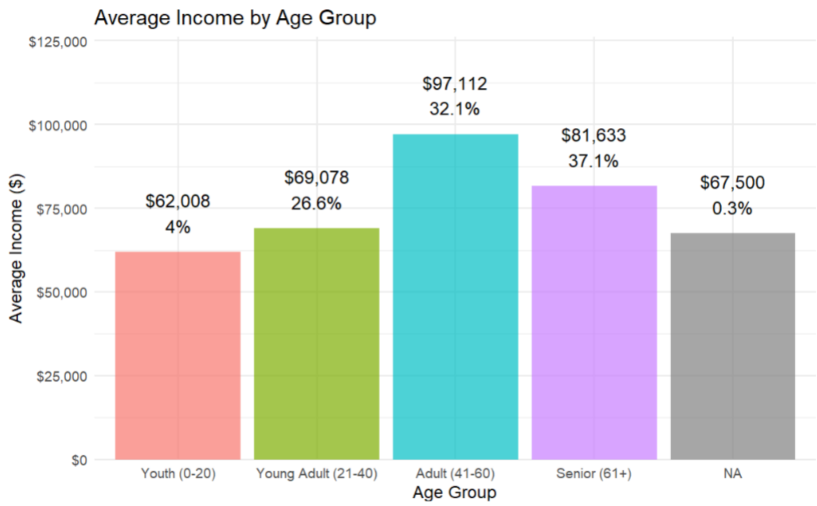


Figure 4: Average Income by Age Group

Youth customers have an average income of $62 008, representing 4% of the total customers. This s the lowest among the age groups, possibly reflecting young professionals or students. The young adult age group, with an average income of $69 078 and 26.6% of all customers, indicates an increase in purchasing power. Adults, aged 41-60, have the highest average income of $97 112 and contribute 32.1%.

Seniors have a slightly lower average of $81 633 but are the largest contributing 37.1% due to their large representation in the customer base. Additionally, the N/A group has negligible contribution of 0.3% with an average income of $67 500. Adults have the highest spending capacity, while seniors, despite having slightly lower income, contribute the most revenue due to their large numbers

Average income by city:

The average customer income by city is relatively consistent across all major locations ranging from $79 752 in New York to $83 346 in Miami, as seen in figure 5. This narrow spread suggests that income levels are similar across cities with no single city standing out significantly as being wealthier or poorer. Thus, it can be concluded that city level differences are marginal, so sales strategies should focus more customer volume and preferences than income distribution and city location.

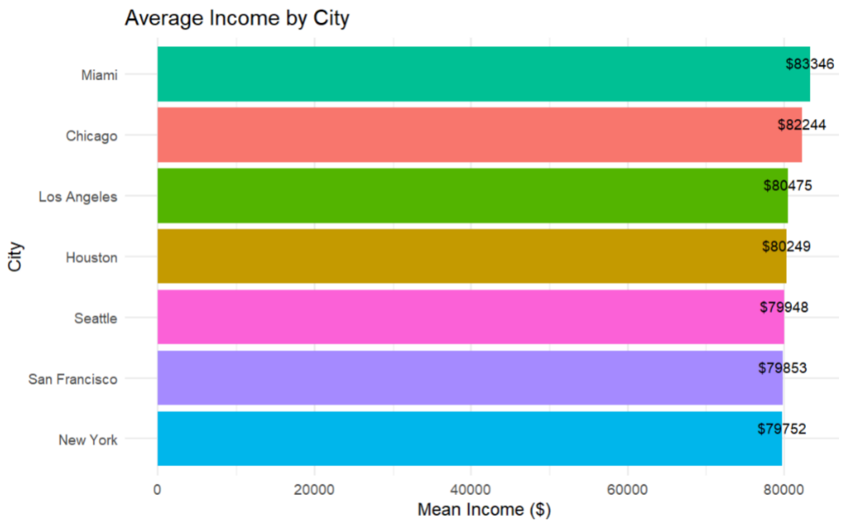


Figure 5: Average Income by City

Product Information

Product rank by revenue:

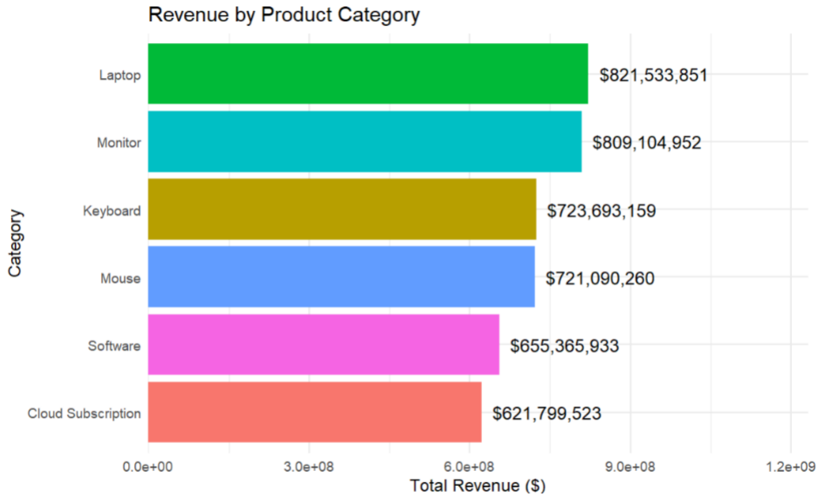


Figure 6: Product Rank by Revenue

From figure 6, it is evident that computers generate the most revenue for the company, closely followed by monitors. In contrast, cloud subscription generates the least revenue for the company.

From the pie chart in figure 7 below, this is further supported by the maximum revenue contribution of 18.9% and the minimum contribution of 14.3%. This indicates that the revenue is relatively evenly distributed across product categories.

This narrow range suggests that the company does not rely heavily on a single category. This could indicate a balanced product range with limited variation across categories.

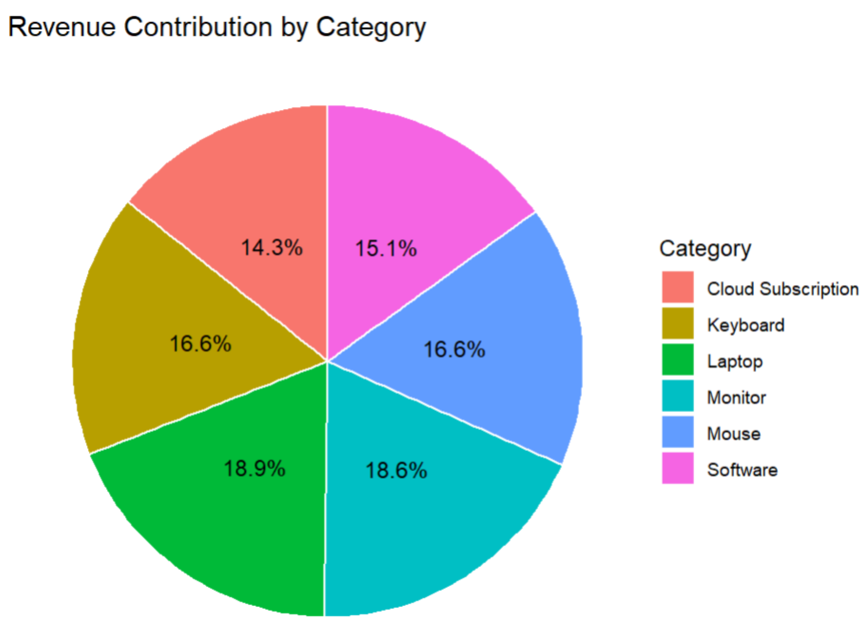


Figure 7: Pie Chart of Revenue Contribution by Category

Price vs. Popularity

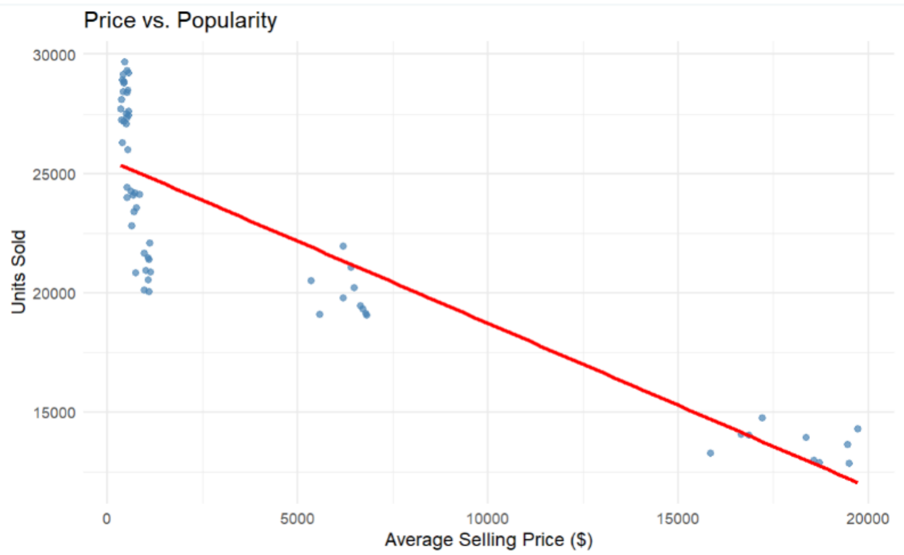


Figure 8: Scatter Plot of Price vs. Popularity

When analyzing the scatterplot in figure 8 of price vs. popularity, it is clear that the lower the price, the greater the popularity and visa versa. This strong decreasing trend can be seen by die red line with a negative gradient.

Monthly Sales Trends by Product Category.

Analysis of the line graph depicting monthly revenue trends in figure 9 below, shows that laptops consistently generate the highest revenue while cloud subscriptions contribute the least revenue for the company. All product categories follow the same general seasonal trend despite differences in value of revenue. Low

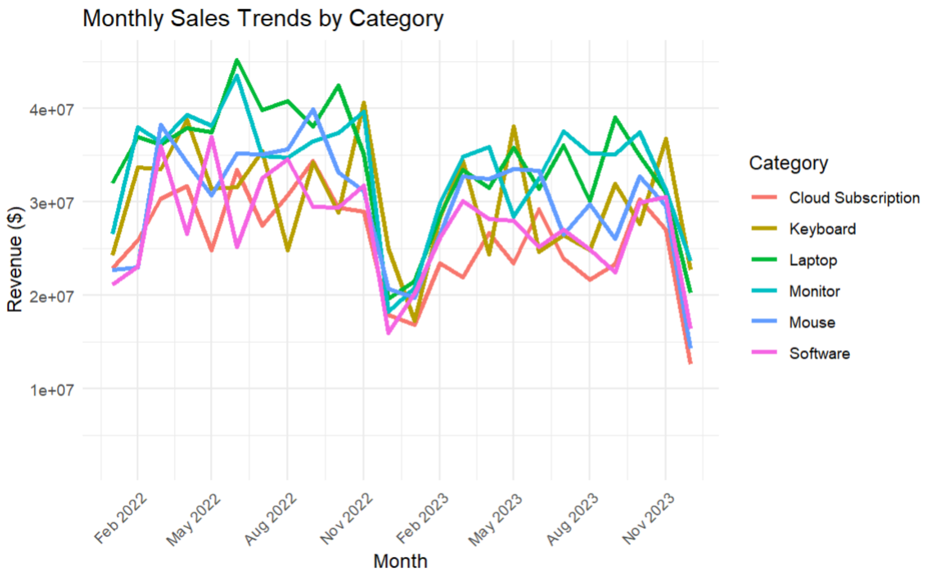
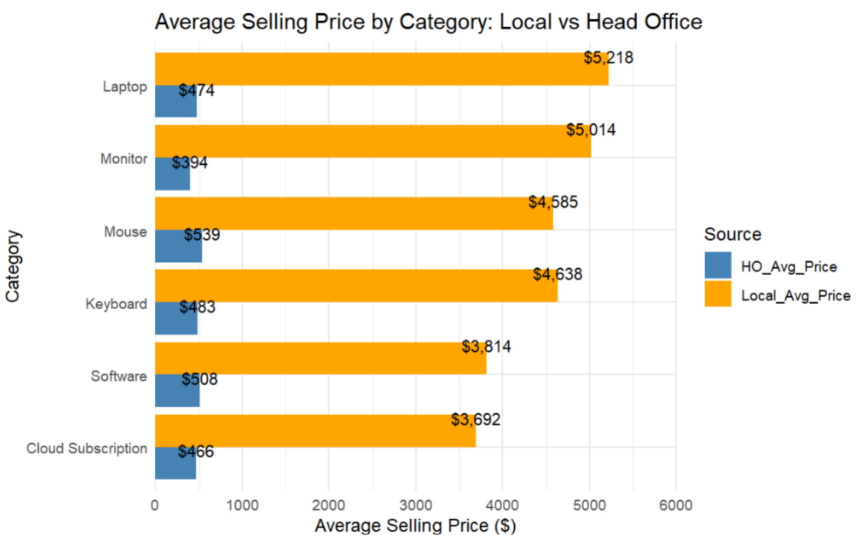


Figure 9: Line Chart of Monthly Trends by Category

peaks are observed in December – January of each year, likely due to shifts in customer spending habits, or the closure of holidays. Overall revenue value tends to increase steadily toward march, after which the sales remain relatively stable through to November, after which a consistent decline is observed from November to December. This trend indicates demand patterns are influenced more by customer's purchasing habits than by product-specific factors. The company should plan production, staffing and stock levels around this cyclical trend, by ensuring higher stock availability into March when demand grows, and optimizing resources during December – January when demand decreases. The company should consider focusing promotions on time of year, rather than category specific sales.

Average Selling Price by Category: Local vs. Head Office



From figure 10 below it is clear that there is a large discrepancy in average selling price. The chart shows that for evry category, the local average selling price is significantly higher than that of head office. For example, laptops average at $5218 locally vs. only

Figure 10: Bar Graph of Average Selling Price by Category: Local vs. Head Office

$474 at head

office, while monitors sell at $5014 locally vs $394 at head office. This trend is consistent across all categories, where local prices are between 7 to 12 times higher than head office prices. This suggests that the two data sources are not directly comparable, as local and head office product data aren't synchronized. Upon further investigation of the two data sources, inconsistencies in both pricing and product categorization was identified. For example, Product SOF001 is recorded as 'coral matt' with a selling price of $521.72 and a markup of 25.05% in the local file, however, Product SOF001 is identified as 'coral silk' with a selling price of $521.72 and a 15.65% markup in the head office dataset. Similar discrepancies were found across other products, including cases where the same ProductID was assigned to different product categories. This explains the large discrepancies in the local vs. head office selling price bar graph, pointing to data integrity issues, which need to be addressed to ensure reliable decision making.

Corrected Average Selling Price by Category: Local vs. Head Office

To ensure a more accurate comparison of selling prices between the local and head office datasets, key issues such as mismatched product identifiers and inconsistent descriptions were addressed. The original dataset contained discrepancies in ProductID, description and selling price which resulted in unreliable data comparison as can be seen in figure 10 above. In the original bar graph, comparisons were made by category, which introduces inaccuracies. However, figure 11 below depicts the corrected comparison of products by full product description. This was done as products within the same category vary widely in their description (e.g., "Keyboard_chocolate silk" vs. "Keyboard_blue marble"). The use of the full product description ensures that comparison is only made between equivalent products and not just products under the same category.

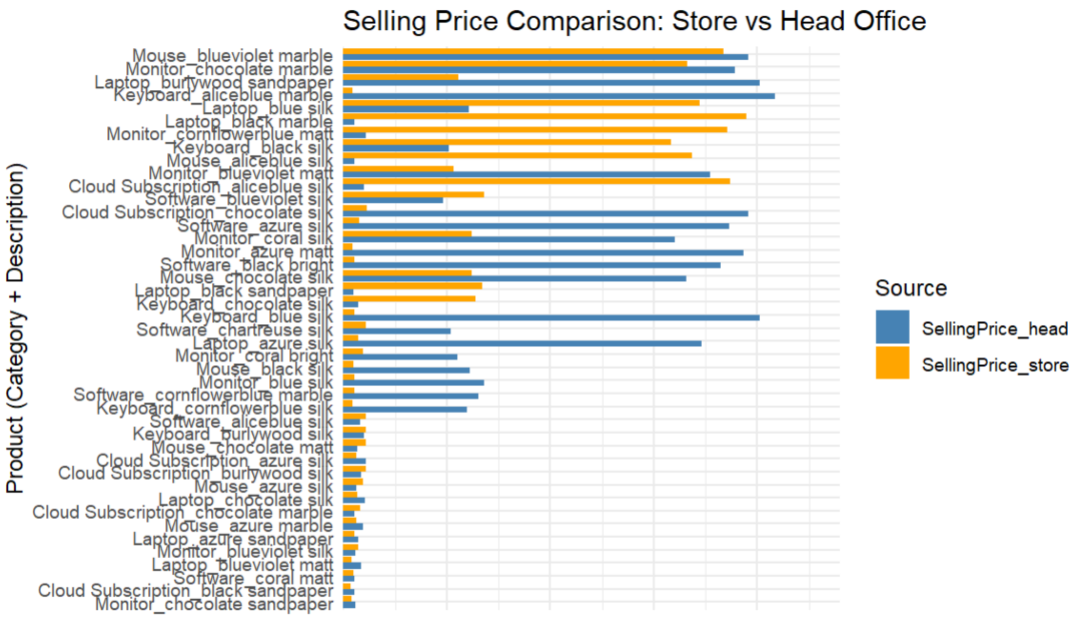


Figure 11: Double Bar Graph of Corrected Average Selling Price by Category: Local vs. Head Office

Sales Statistics:

Bar graph of Total Sale by Year:

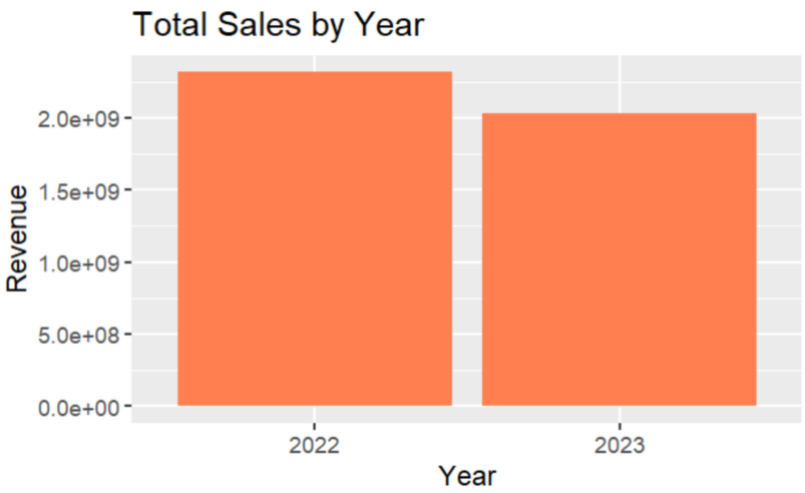


Figure 12: Bar graph of Total Sale by Year

The bar graph in figure 12 shows that the year 2022 had a greater number of sales than 2023. This suggests that the company performed better in 2022 compared to the following year. The decrease in 2023 sales could be as a result of several factors such as a poor shift in market strategy, misalignment with customer needs or external market conditions (such as increased competition or reduced customer spending power). Another explanation could be the company's focus on category-specific promotions in 2023, and neglecting seasonal buying patterns, as seasonal promotions align with spending behaviors.

Year-on-year Sales Trends

Another year-by-year comparison is seen in figure 13, which provides further insight into monthly trends by comparing 2022 with 2023. The line graph shows a clear decrease in sales across nearly all months of 2023 compared to the corresponding months in 2022. This indicated that the decline was not isolated to certain periods but rather represented an overall trend throughout the year. This could be representative of economic conditions and the decreased overall buying power of clients in the year 2023.

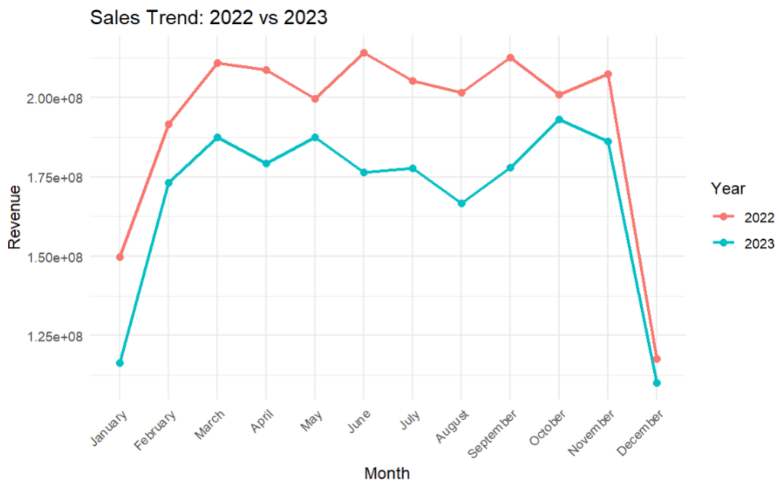


Figure 16: Line graph of Year-on-Year Sales Trends

Monthly Comparison of Sales Trends

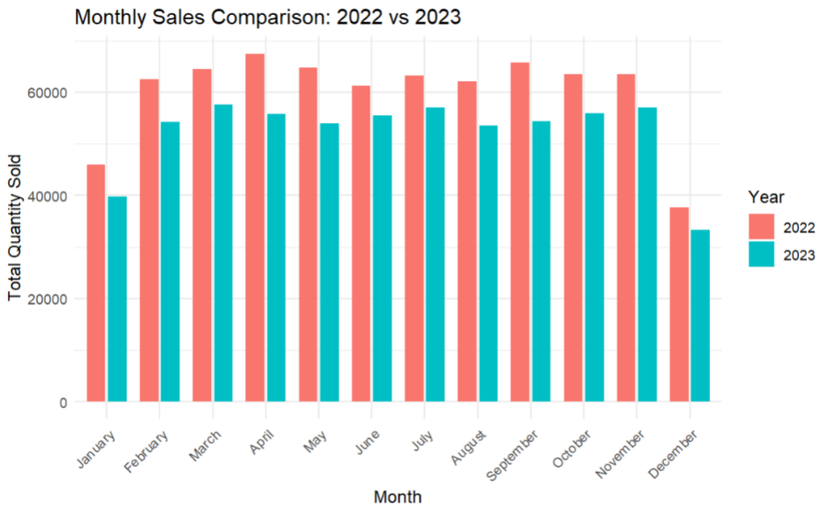


Figure 17: Double Bar Graph of Monthly Comparison of Sales Trends

The bar graph in figure 14 shows a side-by-side monthly comparison between the two years, making it easier to identify months where performance diverged the most. This can be seen in the sharp decline of sales in the months of April and September specifically compared to the same month in the previous year. This could be linked to competitor campaigns or reduce inventory availability. In contrast, smaller gaps, such in the months of November and December may suggest targeted promotions or seasonal demands that temporarily boost sales. This could include Black Friday, commonly held in November, and Christmas shopping in December. This graph supports the identification of months in which sale strategies were less effective, as well as when marketing interventions are most needed. It reinforces the observation that 2023 underperformed throughout the year, but it gives the opportunity to identify periods of greatest lost sales, allowing the company to do corrective planning for future sales.

Data Issues:

During the data analysis, certain data related challenges were identified that required cleaning and restructuring to ensure accurate comparisons and meaningful insights. Mismatched product IDs from the local and head office data used different ID formats, so a mapping was created to align them. The head office listed all possible products, while the local product data only showed actual items sold, leading to price differences. A combination of category and description was used to match products accurately to avoid misleading comparisons within broader categories. With respect to missing data, some rows lacked quantity or pricing data and thus were removed to ensure a clean analysis. Additionally, months were converted to categoric features by changing numerical value to text allowing for the reordering and display of monthly sales data.

SPC Limits:

Future sales data for the years 2026 and 2027 were provided in a CSV file. This dataset includes detailed information such as sales quantities, order times (day, month and year) and the corresponding picking and delivery hours.

To standardize the order time data, each order was converted into an absolute day value, where January 1st of the first year corresponds to day 1, and the time of day is represented as a decimal fraction.

The dataset was then separated into individual product categories and sampled using a sample size of 24. For each sample group, the mean (also known as x-bar,) and the standard deviation (s) were calculated.

Initial Control Charts

Initial control limits were established using the 30 samples of each product category. The upper and lower control limits (UCL and LCL) for one, two and three sigma levels were computed accordingly. The resulting initial control charts for each product are shown in figures 18 to 23 below:

Mouse Initial Control Chart

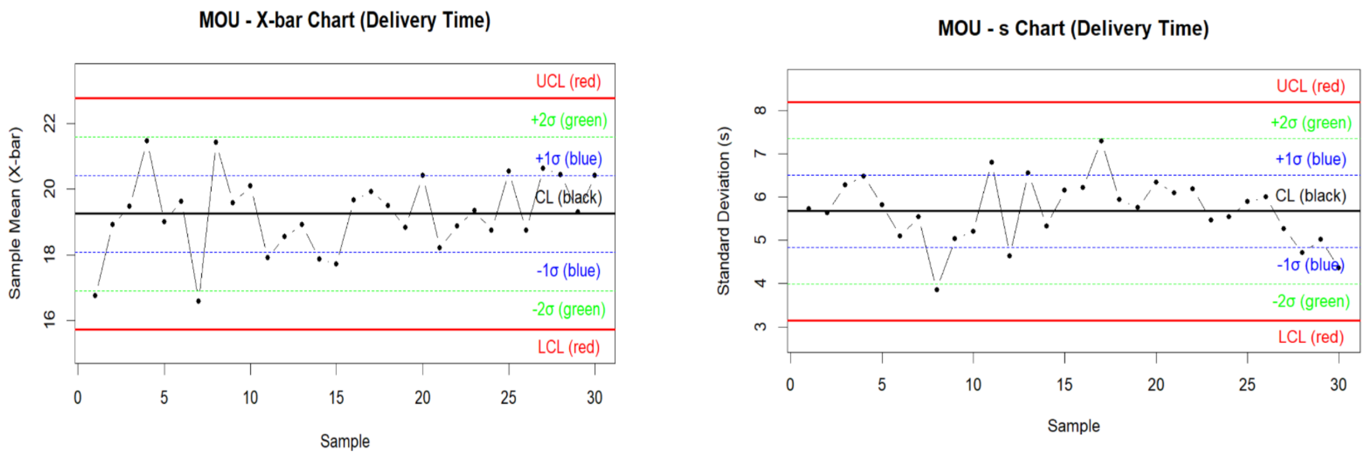


Figure 18: Mouse Initial Control Charts

Keyboard Initial Control Chart

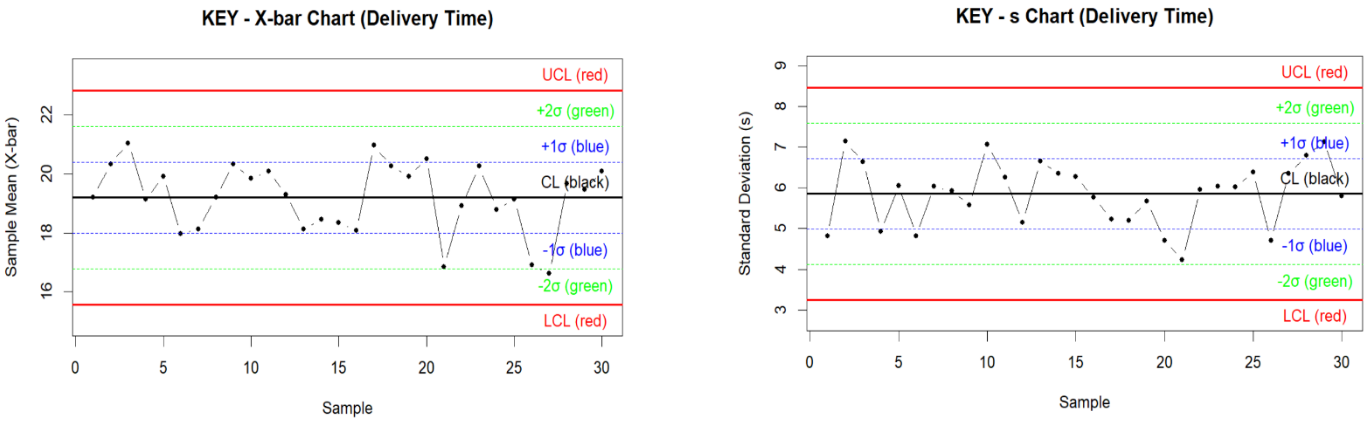


Figure 19: Keyboard Initial Control Charts

Software Initial Control Chart

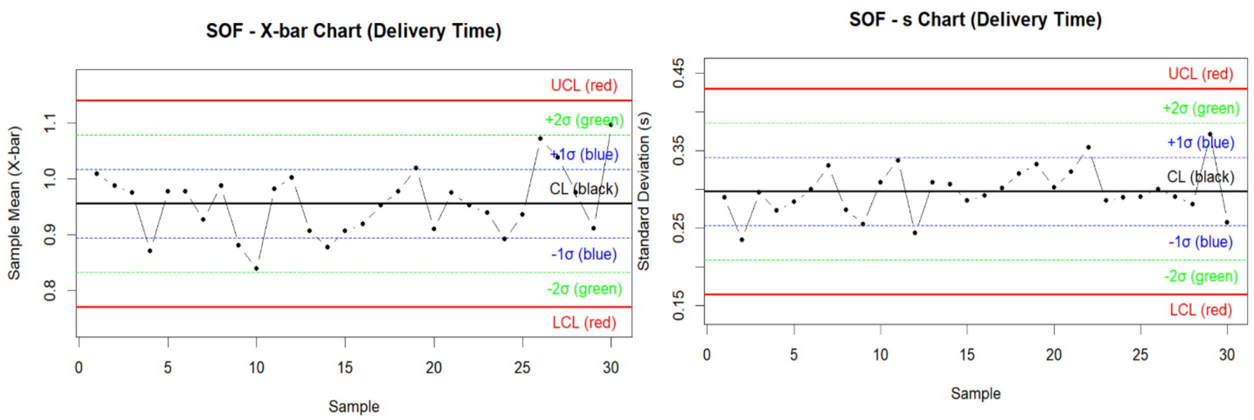


Figure 20: Software Initial Control Charts

Cloud Subscription Initial Control Chart

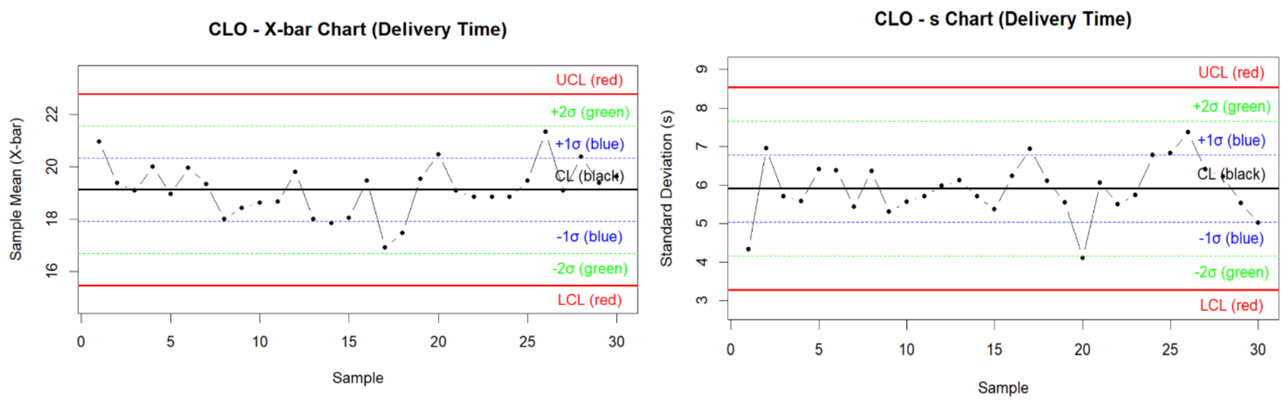


Figure 21: Cloud Subscription Initial Control Charts

Laptop Initial Control Chart

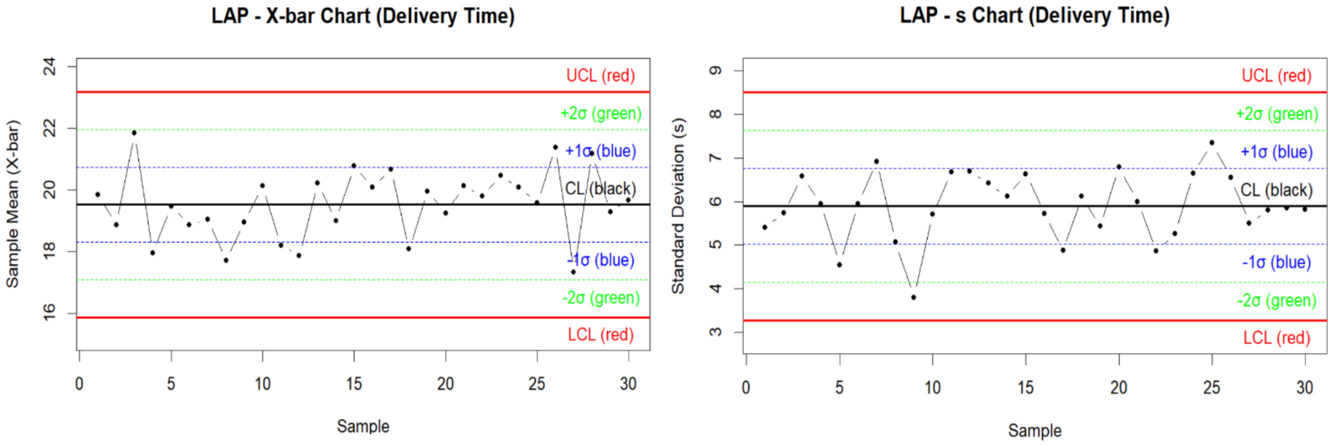


Figure 22: Laptop Initial Control Charts

Monitor Initial Control Chart

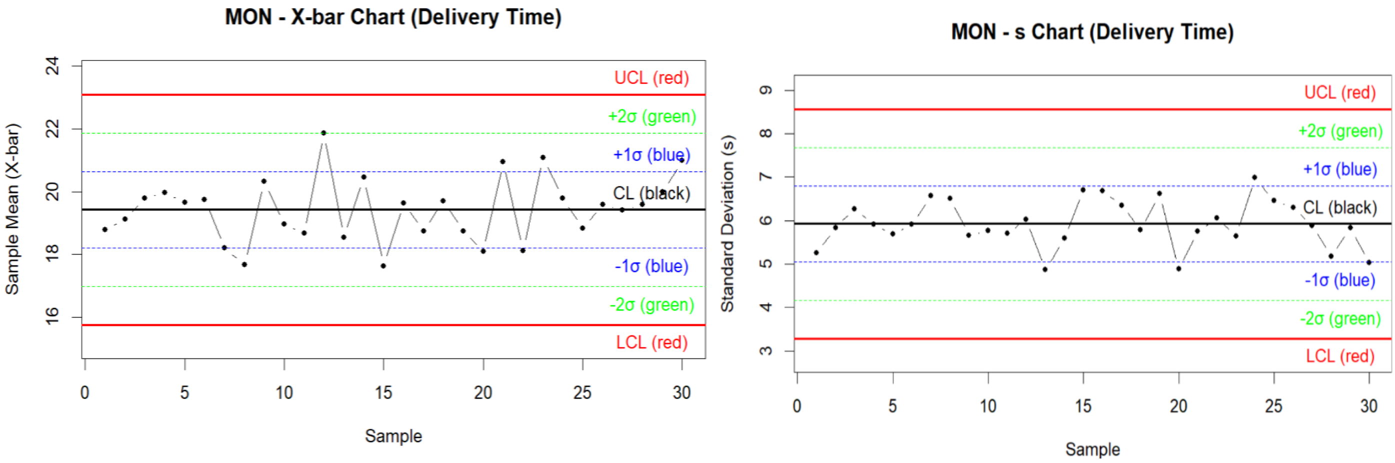


Figure 23: Monitor Initial Control Charts

Continued Control Charts

Additional groups of 24 samples were incrementally added to each control chart unti all availiable data was incorporated. The final control charts, which include the complete dataset, are shown in figures 24 to 29 below.Mouse Continued Control Charts:

Mouse Continued Control Charts:

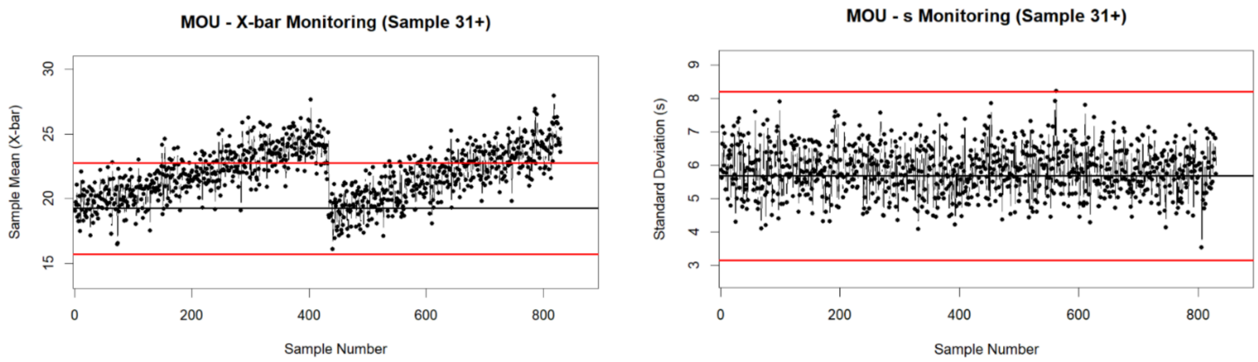


Figure 24: Mouse Continued Control Charts

Keyboard Continued Control Charts:

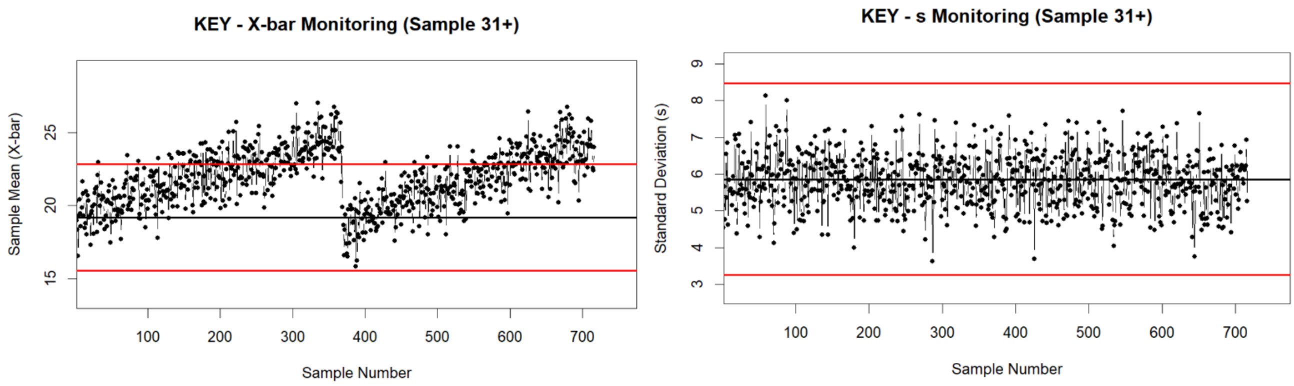


Figure 25: Keyboard Continued Control Charts

Software Continued Control Charts:

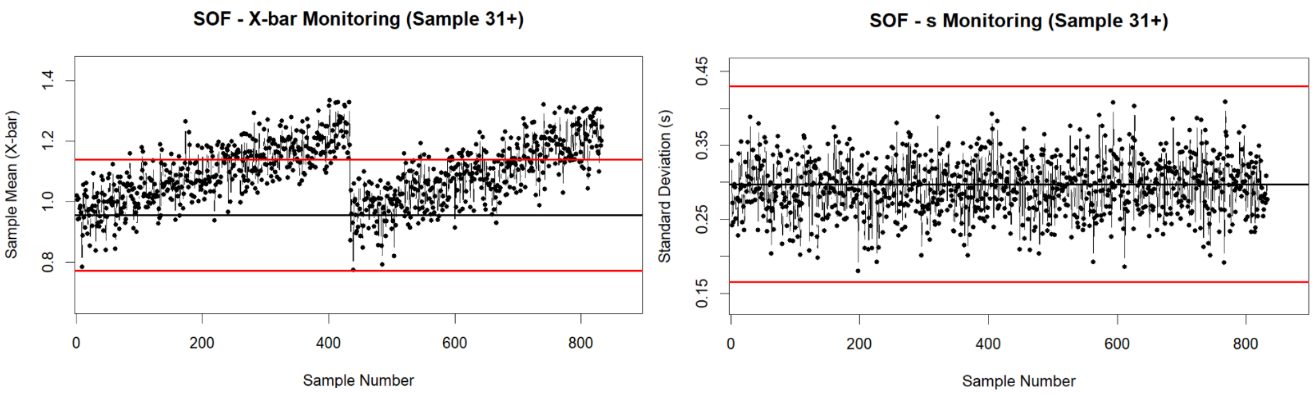


Figure 26: Software Continued Control Charts

Cloud Subscription Continued Control Charts:

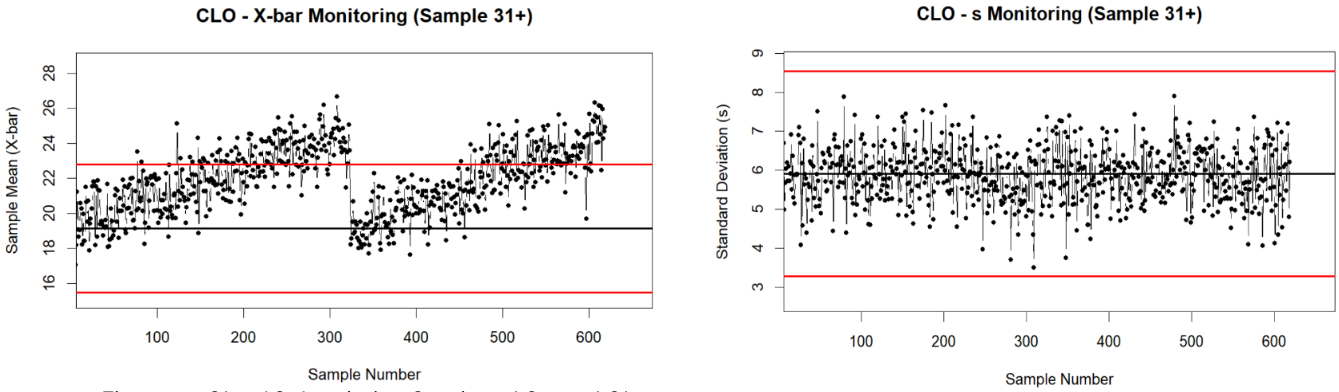


Figure 27: Cloud Subscription Continued Control Charts

Laptop Continued Control Charts:

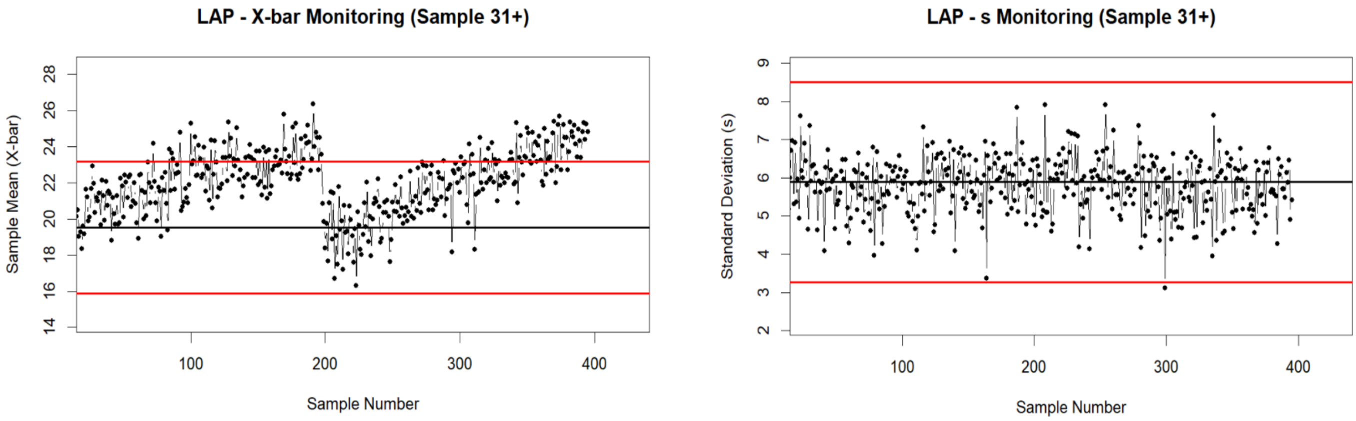
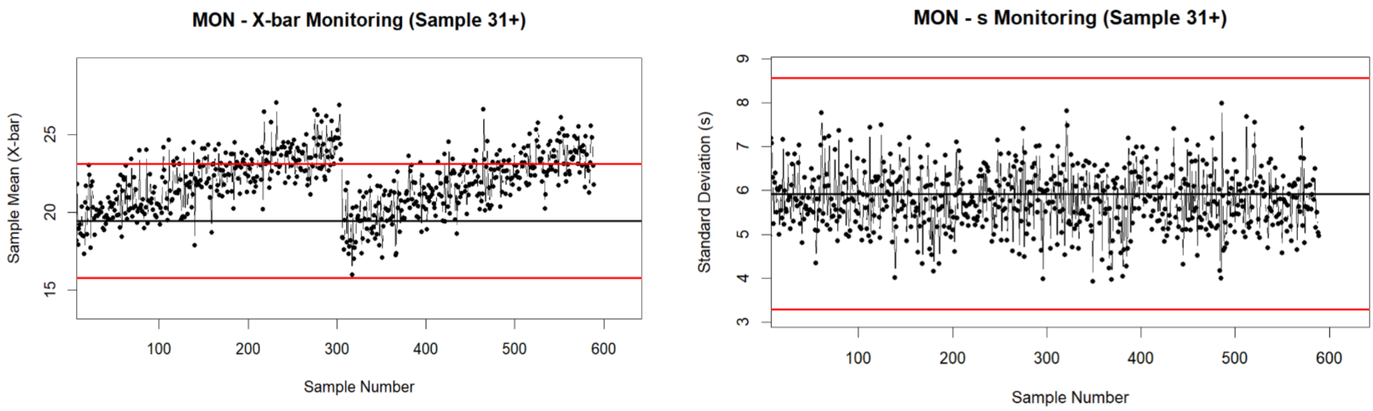


Figure 28: Laptop Continued Control Charts

Monitor Continued Control Charts:



Process Control Analysis

Each analysis highlights the key control limit violations, delivery performance trends and potential area of improvement. While most products maintained stable variability, some showed continued upwards shift in delivery averages, which indicates possible drift or bias.

Mouse

The Mouse product category exhibited a low overall sample standard deviation, with most samples remaining within the three-sigma control boundaries. The process showed a strong upward trend in sample averages, similar to other product categories.

The first out-of-control sample occurred at sample 95 with an X-bar of 22.794 and a standard deviation of 6.159, which exceeds the upper control limit. Early violations can be found in table 3 below:

Sample	X-bar	S	Out Of Control Reason
95	22.79400	6.159122	X-bar > UCL3
139	23.00233	6.146609	X-bar > UCL3
179	24.16900	6.141891	X-bar > UCL3

Table 3: Mouse First Out of Bound Samples

This trend continues with a total of 288 violations of the upper control limit of X-bar. The final samples with continued instability can be seen in table 4 below.

Sample	X-bar	S	Out Of Control Reason
858	24.13317	6.826565	X-bar > UCL3
859	23.58958	6.798972	X-bar > UCL3
860	25.42292	5.742021	X-bar > UCL3

Table 4: Mouse Last Out of Bound Samples

Despite the overall control in spread, with a single s-chart violation, the persistent upward drift in sample averages suggests a systematic bias in delivery performance. This could be due to increased demand or operational delays and could be addressed with closer monitoring.

Keyboards

The keyboard product category showed a generally stable process spread with most samples standard deviation remaining within the three-sigma control boundaries. May samples exceeded the +- one-sigma range, indicating moderate variation. Only one sample exceeded the upper control limit for standard deviation, confirming that variability was mostly under control.

The process exhibited a clear upward trend in delivery hours with frequent violations of the upper control limit in delivery hours, with frequent violations of the upper control limit for the sample mean. The first out-of-control sample occurred at sample 62 with an X-bar od 22.961 and standard deviation of 6.157, which exceeds the positive three-sigma threshold. The first 3 violations can be seen in table 5 below.

Sample	X-bar	S	Out of Control Reason
62	22.96067	6.156768	X-bar > UCL3
102	23.41900	5.554571	X-bar > UCL3
117	23.04400	7.192992	X-bar > UCL3

Table 5: Keyboard First Out of Bound Samples

The upward drift continued throughout the monitoring period with a total of 294 violations of the upper control limit for X-bar. This trend continued to increase as can be seen with the last three samples in table 6.

Sample	X-bar	S	Out of Control Reason
742	25.75817	6.180868	X-bar > UCL3
743	24.01008	6.108424	X-bar > UCL3
746	24.00433	5.254225	X-bar > UCL3

Table 6: Keyboard Last Out of Bound Samples

These results suggest a systematic upward draft in delivery performance of keyboards, despite a stable distribution. Possible demand increases or bottlenecks in fulfillment could result in this drift and thus the company must investigate this further.

Software

Although software is a digital product and expected to show low variability, its standard deviation was similar to those physical products. This process showed a clear upward drift in the sample means, with frequent violations of the upper control limit of X-bar.

The first out-of-control sample occurred at sample 115 with an X-bar of 1.158 and a standard deviation of 0,316. The first 3 violations are recorded in table 7 below.

Sample	X-bar	S	Out Of Control Reason
115	1.158450	0.3163352	X-bar > UCL3
134	1.179283	0.2264946	X-bar > UCL3
136	1.164700	0.2731658	X-bar > UCL3

Table 7: Software First Out of Bound Samples

The trend continued throughout the monitoring period, with 334 total violations. In table 8 below, the final 3 unstable samples are recorded.

Sample	X-bar	S	Out Of Control Reason
862	1.185633	0.3084557	X-bar > UCL3
863	1.200217	0.2760274	X-bar > UCL3
864	1.246050	0.2767484	X-bar > UCL3

Table 8: Software Last Out of Bound Samples

The upward drift remained evident throughout, indicating that Software delivery performance requires continual monitoring and adjustment

Cloud Subscription

The Cloud Subscription category shows a clear upward drift in delivery hours with frequent violations of the upper control limit for the sample mean, X-bar. While the standard deviations remain mostly within the control boundaries, the continual rise in averages indicates a potential shift in performance.

The first out-of-control sample occurred at sample 107, with an X-bar od 23.502 and a standard deviation of 6.234. The earliest violations can be found in table 9 below.

Sample	X-bar	S	Out-Of-Control-Reason
107	23.50233	6.234401	X-bar > UCL3
112	22.91900	4.386863	X-bar > UCL3
153	25.12733	5.266107	X-bar > UCL3

Table 9: Cloud Subscriptions First Out of Bound Samples

The trend continues throughout the monitoring process, with a total of 263 violations. The final three samples outside the control limit are samples 647, 648 and 649, found in table 10 below.

Sample	X-bar	S	Out-Of-Control-Reason
647	24.25817	7.192278	X-bar > UCL3
648	24.67100	4.798664	X-bar > UCL3
649	24.92100	6.215776	X-bar > UCL3

Table 10: Cloud Subscriptions Last Out of Bound Samples

Despite the stable spread, the consistent upward shift in delivery hours of Cloud subscriptions suggests a need for recalibration and closer monitoring.

Laptop

The laptop category showed a clear upward drift in delivery hours with continual violations of the upper control limit for the sample mean. While the process standard deviation remained within the control boundaries, the constant rise in averages indicates a change in performance.

Sample 102 is the first out-of-control limit sample that occurred with an X-bar value od 24.171 and an s value of 5.427. The first three sample violations are listed in table 11 below.

Sample	X-bar	S	Out Of Control Reason
102	**24.17083**	**5.427017**	**X-bar > UCL3**
114	**23.87733**	**6.024708**	**X-bar > UCL3**
116	**23.65000**	**5.559583**	**X-bar > UCL3**

Table 11: Laptop First Out of Bound Samples

The trend continued throughout the monitoring process with 159 violations od the upper control limit of the sample means. The final three out-of-control limit violations can be seen in table 12.

Sample	X-bar	S	Out Of Control Reason
423	**24.38508**	**6.460296**	**X-bar > UCL3**
424	**25.25433**	**4.907662**	**X-bar > UCL3**
425	**24.84150**	**5.420735**	**X-bar > UCL3**

Table 12: Laptop Last Out of Bound Samples

Although the process distribution remains stable, the consistent rise in delivery hours of laptops suggests a need for better oversight to restore alignment with the company's target performance.

Monitor

Th monitor product category also showed a consistent upward drift in delivery hours, with frequent violations to the upper sample mean control limit. While the process spread is within the acceptable boundaries, the sustained rise in averages indicated a shift in performance of monitor delivery.

The first three out-of-control samples can be seen in table 13 below, with the first, sample 99, having a mean of 23,419 and a standard deviation of 7.240.

Sample	X-bar	S	Out Of Control Reason
99	23.41900	7.240061	X-bar > UCL3
108	23.96067	4.853655	X-bar > UCL3
116	24.00233	5.733041	X-bar > UCL3

Table 13: Monitor First Out of Bound Samples

This trend persists throughout the monitoring period with a total of 226 violations, where the last three sample violations are recorded in table 14 below.

Sample	X-bar	S	Out Of Control Reason
615	23.17483	6.496868	X-bar > UCL3
616	25.54792	5.152847	X-bar > UCL3
617	24.79792	5.505969	X-bar > UCL3

Table 14: Monitor Last Out of Bound Samples

Although the process distribution was well controlled, the persistent upward shift in delivery hours of the monitor product category suggests a need for closer monitoring and recalibration to maintain process centering.

Process Capability

Process capability was calculated for each of the six product categories. The analysis included the computation of Cp, Cpu, Cpl and Cpk indices for each category as in table 15. It is noteworthy that the Cpl holds less significance in this context due to the one-sided nature of the delivery-time data, that is, an order cannot be delivered before it has been placed. Therefore, greater emphasis is placed on Cpu, which reflects performance relative to the upper specification limit (USL).

Product Type	**Cp**	**Cpu**	**Cpl**	**Cpk**
Mouse	0.915	0.727	1.104	0.727
Keyboard	0.917	0.729	1.105	0.729
Software	18.135	35.188	1.083	1.083
Cloud Subscription	0.898	0.717	1.079	0.717
Laptop	0.899	0.696	1.101	0.696
Monitor	0.889	0.700	1.079	0.700

Table 15: Process Capability Limits by Category

All physical products such as Cloud, Laptop, Keyboard, Monitor, and Mouse exhibit low Cpu, Cpl and Cpk values, especially when compared to Software, the only digital product category.

A Cpk value below 1.0 suggests that a process is barely capable of meeting the defined specification limits. That is, LSL = 0 hours and USL = 32 hours. The Cpk value for all physical products falls below this threshold, indicating that these processes struggle to consistently satisfy the delivery-time constraints.

Furthermore, the small variation in Cpk values among the physical products suggests that the low capability is a general characteristic of all physical delivery processes rather than an issue with any specific product line. For these products, Cpu values are consistently lower than Cpl, confirming that the one-sided nature of delivery hours does not influence the overall capability measure.

In contrast, the Software product category displays a relatively high Cpu, even though its Cpk remains low. Given the one-sided constraint on delivery time, the high Cpu is a more relevant indicator. This demonstrates that the software delivery process is highly capable of meeting the required limits.

Process Control Issues

To evaluate the stability and consistency of the delivery process across product types, three statistical rules were applied to the sample data. These rules help identify when a process may be unstable or drifting. Each rule targets a different aspect of process behavior: Rule A focuses on excessive variability; Rule B focuses on periods of consistent and stable performance while; Rule C detects extended shifts in the process mean that could indicate drift or bias.

The results below summarize the findings of each rule.

Rule A

This rule identifies samples where the standard deviation exceeded the upper control limit, indicating an unstable process spread. Only the Mouse product type had a single violation, namely sample 592, exceeding the +3-sigma limit. This spike, although occurring only once, is notable as no other products show any violations, suggesting that process spread is generally well controlled across all product categories.

```
=== Rule A: s samples above +3σ ===
MOU - First 3: 592 Last 3: 592 Total: 1
KEY - First 3:  Last 3:  Total: 0
SOF - First 3:  Last 3:  Total: 0
CLO - First 3:  Last 3:  Total: 0
LAP - First 3:  Last 3:  Total: 0
MON - First 3:  Last 3:  Total: 0
```

Figure 29: Rule A by Product Category

Rule B

This rule highlights the longest sequence of samples where the process spread remained stable. This includes the Cloud Subscriptions category with 36 consecutive stable samples within a positive and negative 1-sigma range followed by Monitors with 34 consecutive stable samples. It is notable that Software returned an "-Inf samples", indicating an invalid result. This is likely to be due to missing or undefined control limits and should therefore be reviewed to ensure accurate monitoring.

```
=== Rule B: Longest run of s within ±1σ ===
MOU - Longest run: 16 samples
KEY - Longest run: 20 samples
SOF - Longest run: -Inf samples
CLO - Longest run: 36 samples
LAP - Longest run: 19 samples
MON - Longest run: 34 samples
```

Figure 30: Rule B by Product Category

Rule C

This rule identified extended upward shifts in the process mean, which may bias delivery performance. All product types display sustained mean shifts in the process means, with Software and Mouse having the highest number of violations with 334 and 324 violations respectively. This trend suggests potential drifts in delivery performance and calls for investigation and monitoring to maintain delivery consistency.

```
=== Rule C: 4+ consecutive X-bar samples above +2σ ===
MOU - First 3: 194 195 196 Last 3: 858 859 860 Total: 324
KEY - First 3: 112 113 114 Last 3: 744 745 746 Total: 294
SOF - First 3: 202 203 204 Last 3: 862 863 864 Total: 334
CLO - First 3: 122 123 124 Last 3: 647 648 649 Total: 263
LAP - First 3: 119 120 121 Last 3: 423 424 425 Total: 159
MON - First 3: 134 135 136 Last 3: 616 617 618 Total: 226
```

Figure 31: Rule C by Product Category

Risk and Data correction

Estimated Likelihood of Type I, Manufacturer's Error for Rules A, B and C.

The likelihood of making a Type I error, assuming the process is in control and centered, differs by rule. Rule A has a theoretical error rate of approximately 0.27%, based on the probability of a sample exceeding positive three-sigma.

Rule B, which identifies periods when the data was well controlled and without any violations. In the case of Rule B, the Type I error is approximately zero.

To identify extended shifts in the process average, Rule C is used. This rule has a theoretical error rate of about 6.25% when assuming the second control limit is near positive two-sigma and the probability of 4 consecutive samples above this limit is 0.5^4.

From these estimates, the sensitivity of each rule to 'false alarms' or type I errors under stable conditions is clear (Bhandari, 2021).

Estimated Likelihood of Type II, Customer's Error for Rules A, B and C.

Considering that the process shifted to a mean of 25.028L and that the standard deviation increased to 0.017L, the probability of failing to detect this shift, that is, making a Type II error, is approximately 84.1%. This value was calculated by standardizing the original control limits using the new process parameters and finding the probability that sample means fall between the lower control limit and the upper control limit. From the result, the risk that not updating control limits have on the current process behavior when control limits drifts are undetected.

Data Correction:

Customer Information:

This section presents previously established customer insights, as no new data was added. All findings remain consistent with earlier analyses and continue to support the existing understanding of purchasing behaviour.

Product Information

This section details the characteristics of each product, such as category, description, pricing, and sales performance, enabling accurate comparison and strategic analysis across the company's inventory.

Corrected Product rank by revenue:

From figure 32, it is evident that laptops generate the most revenue for the company with approximately \$2.47 billion, followed by monitors with a \$1.26 billion contribution to revenue. The original graph (figure 6) suggests that cloud subscription generates the least revenue for the company, while the corrected Product rank by revenue suggests that Mouse products contribute the least revenue (\$111 million).

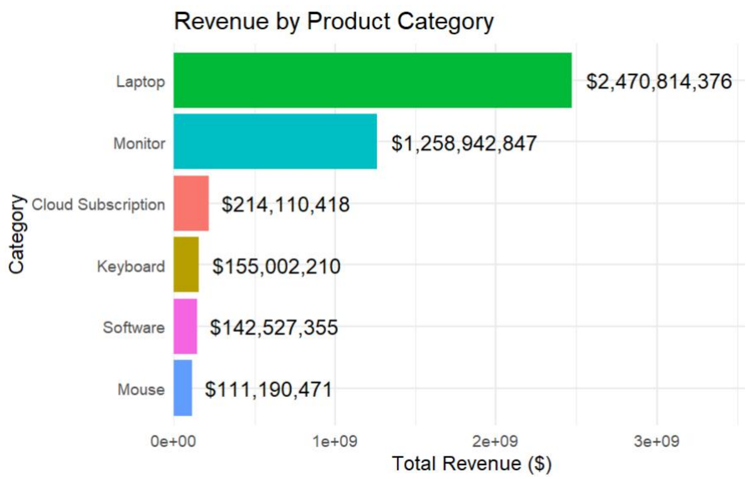


Figure 32: Corrected Revenue by Product Category

The pie chart in Figure 33, show the percentage contribution to total revenue by product category. It clearly shows that Laptops account for 56.8% of revenue, while Mice contributes only 2.6%.

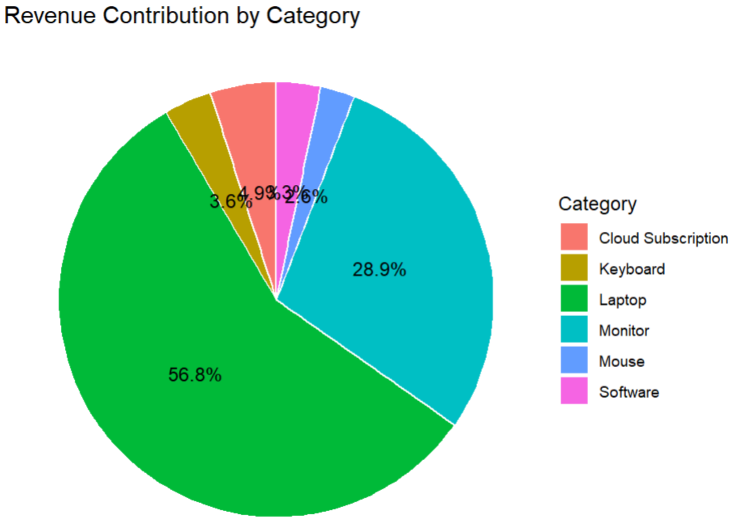


Figure 33: Pie Chart of Corrected Revenue Contribution by Category

This contrasts Figure 7, which previously suggested that the revenue is relatively evenly distributed across product categories. The chart corrected shows the clear imbalance, indicating that the company's revenue is heavily concentrated in a few dominant categories. This widespread distribution suggests a reliance on high-performing products, such as Laptops and Monitors, rather than a uniformly balanced product mix.

Corrected Price vs. Popularity

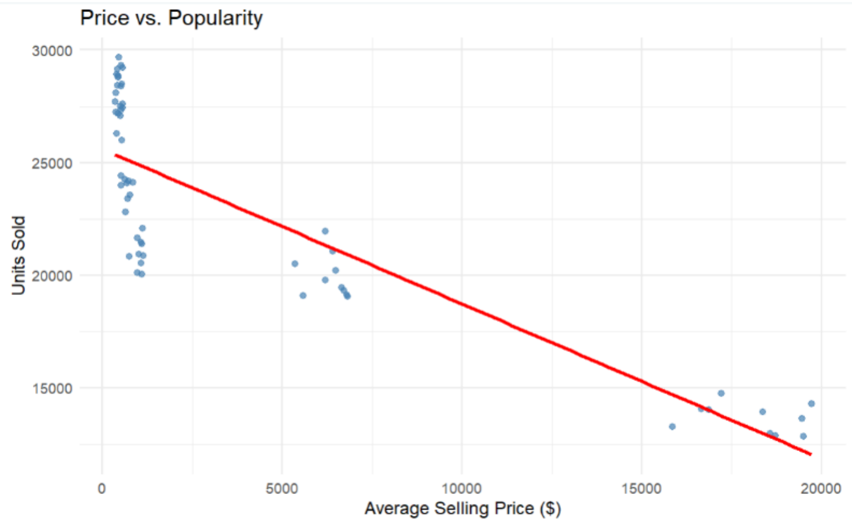


Figure 34: Corrected Scatter Plot of Price vs. Popularity

The overall trend observed in Price vs. Popularity of products as previously seen in Figure 8, remains consistent after the data has been corrected, as seen in Figure 34. The negative correlation between price and popularity is still evident, as shown in the downward sloping red regression line, indicating that lower-priced products consistently attract higher demand.

Monthly Sales Trends Corrected by Product Category.

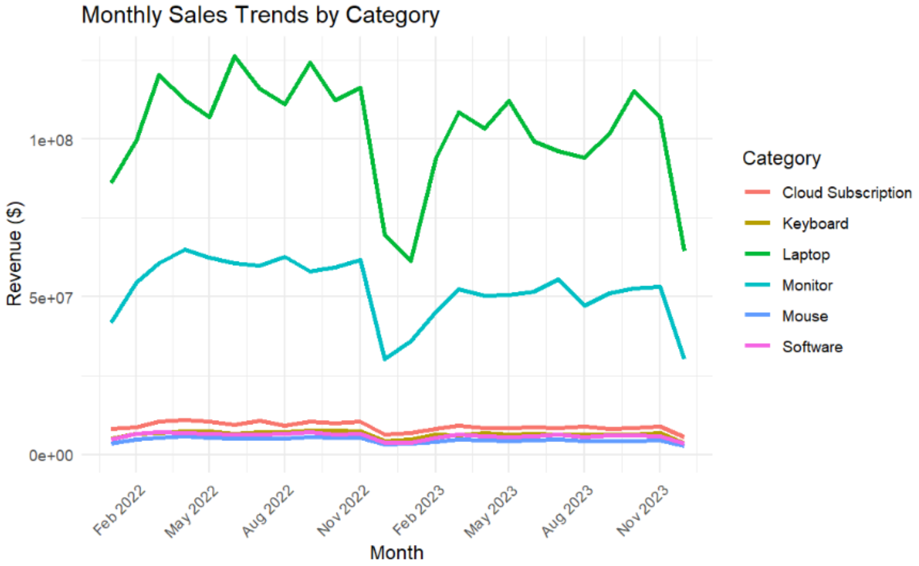


Figure 35: Line Graph of Corrected Sales Trends by Category

The corrected line graph reveals a more pronounced distribution in revenue across products categories than previously observed in Figure 8. While the original graph suggests that all product categories followed similar revenue levels and seasonal trends, the updated data shows that Laptops dominate revenue across all months of the years 2022 and 2023, followed by Monitors. Both these two categories reveal clear seasonal demand patterns. In contrast, Cloud Subscriptions, Keyboards, Mice and Software contribute consistently low revenue throughout the year, with minimal variation and only a slight sales decrease during December to January when the leading contributors experience large decrease in sales.

Despite the shift in contribution by category, the seasonal trend remains consistent: demand rises toward March, stabilizes through mid-year, and declines again from November to January. This confirms that customer purchasing behavior continues to drive cyclical demand. The corrected data highlights the need for category-specific planning, specifically around high performance products such as Laptops and Monitors. The company should maintain a seasonal strategy but should prioritize inventory and promotions around dominant categories.

Corrected Average Selling Price by Category: Local vs. Head Office

The updated chart in Figure 36 below builds on the improvements made in Figure 11 by continuing to match the products by their full description to ensure accurate comparison. The corrected graph maintains the structure by displaying store and head office prices side by side and reveals greater variation in pricing for several products. The price gaps widened for certain items such as laptops and Monitors, suggesting either recent pricing updates or improved data precision. Although the method of comparison remains consistent, the updated chart provides a clearer view of pricing misalignment, especially for laptops and monitors. This refinement supports more informed decisions around pricing strategy and highlights the importance of maintaining consistent data alignment between the store and head office

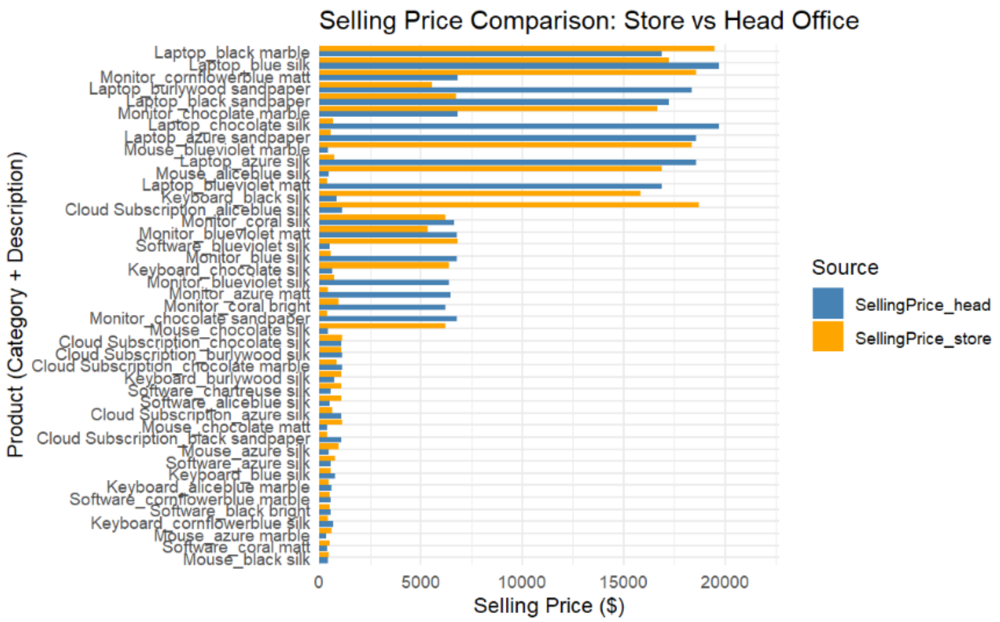


Figure 36: Corrected Selling Price: Store vs. Head Office

Sales Statistics:

Corrected Bar graph of Total Sale by Year:

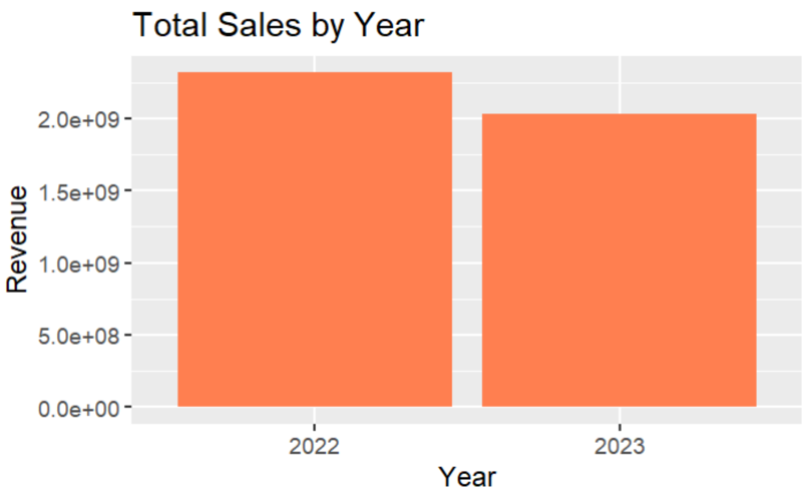


Figure 37: Corrected Bar graph of Total Sale by Year

Figure 37 continues to show that total sales in 2022 exceed those in 2023, indicating a stronger performance in the earlier year. Since the graph remains unchanged, the previous interpretation of Figure 12 holds: the decline in 2023 sales may stem from several factors including a shift in market strategy, misalignment with customer preferences or greater external pressures such as increased competition or reduced consumer spending. An alternative explanation is the company's focus on category-specific promotions in 2023, which may have come at the expense of seasonal campaigns. Since seasonal promotions tend to align more with customer purchasing behavior, this shift in approach may have played a role in the decline of total sales.

Corrected Year-on-year Sales Trends

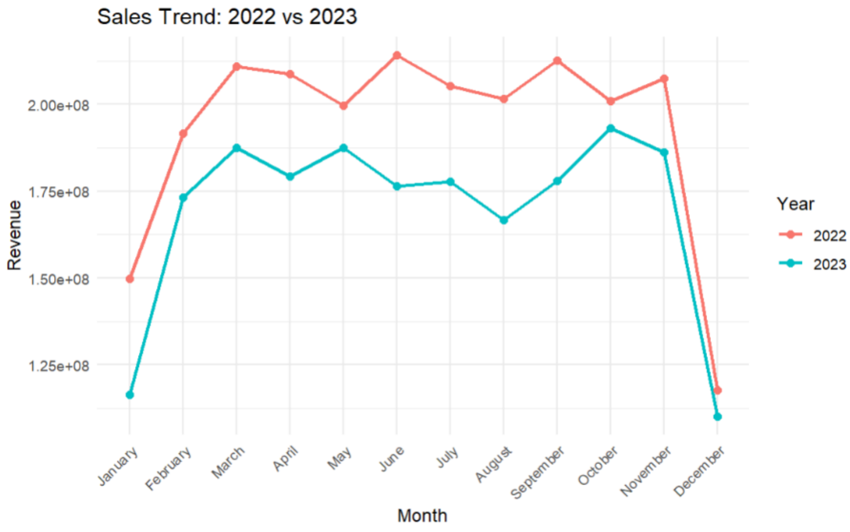


Figure 38: Corrected Sales Trend: 2022 vs. 2023

Upon evaluation of the corrected data figure below, as in Figure 13, continues to show a consistent decline in monthly sales throughout 2023 when compared to the same months in 2022. The downward trend is evident across nearly the entire year, indicating that the drop in performance was not limited to a specific period, but rather reflects a consistent year-long decrease in overall revenue. This trend suggests economic challenges such as reduced consumer spending power or a market-wide shift in demand. The unchanged graph reinforces that external conditions rather than operational issues, likely contributed to the overall decline in sales in 2023.

Monthly Comparison of Sales Trends

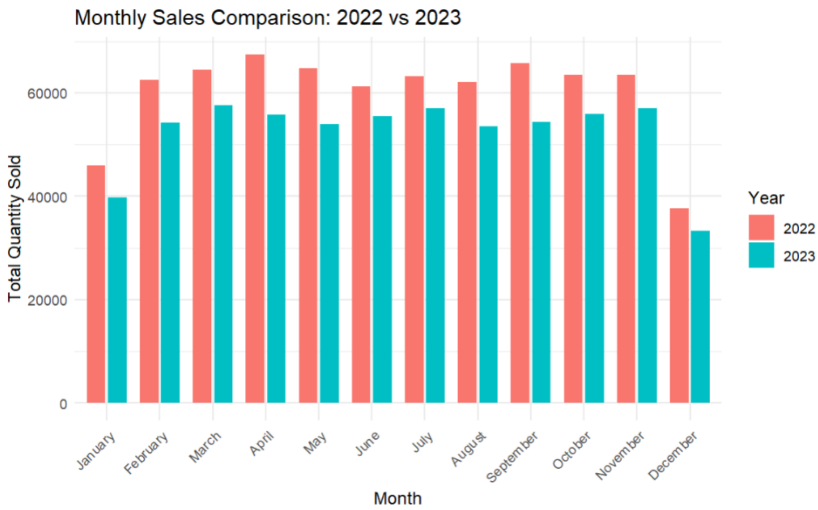


Figure 39: Corrected Double Bar Graph of Monthly Comparison of Sales Trends

The bar graph of the corrected monthly comparison of sales in Figure 39, remains the same as Figure 15 and continues to highlight key monthly differences in sales between 2022 and 2023. Sharp decline of sales in the months of April and September suggests possible inventory issues or competitor activity, while smaller gaps, such in the months of November and December, may suggest seasonal promotions such as Black Friday and Christmas. The graph reinforces that 2023 underperformed overall but helps identify months where strategic adjustments could recover lost sales.

Optimizing for Maximum Profit

Data Structure and Grouping

This analysis uses one year of transaction and service level data from timeToServe.csv and timeToServe2.csv which includes the number of baristas on duty and the time taken to serve each customer in seconds.

Each row of the dataset represents a customer transaction, and the data was grouped first by date as the data was simulated across 365 days to reflect daily transactions. Secondly, the data was grouped by weekday as extracted from the date to identify weekday specific patterns. Lastly, the barista count was used to group performance based on the staffing level. Using revenue per customer of R30 and a staffing cost of R1000 per barista per day, the net profit could be calculated.

It is important to note that the analysis does not assume a fixed number of customers per day but rather reflects real operational history. This means that days with more baristas tends to correlate to days with higher customer volumes, while the converse is also true. The analysis does not imply that more baristas cause more customers but rather observes how reliability and profitability differ under different staffing conditions.

Coffee Shop 1

Reliability vs. Staffing

To evaluate how staffing levels affect operational performance of this coffee shop, the average customer volume, reliability and profit based on barista count from 1 to 6 was analyzed. Figure 40 below shows a clear relationship between the number of baristas working at the coffee shop and the service reliability.

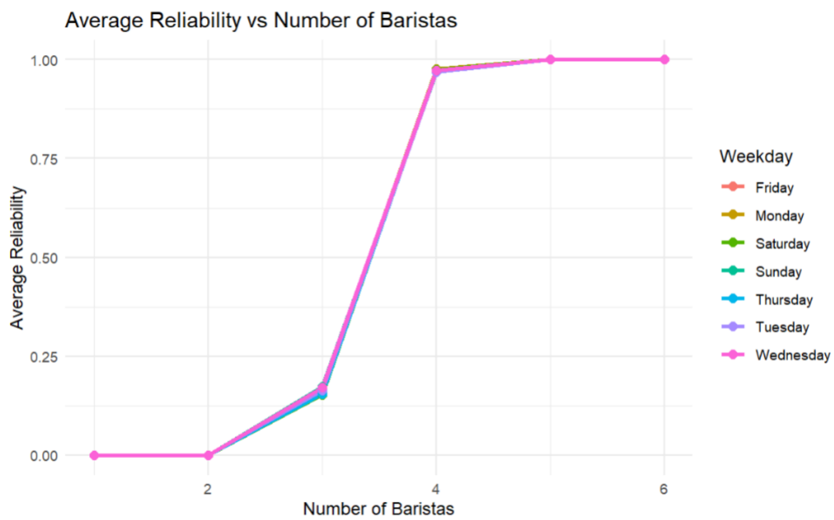


Figure 40: Line Graph of Average Reliability vs. Number of Baristas by Weekday

The reliability graphs show a consistent upward pattern across all weekdays, with each day following a similar upward trend as barista count increases. This indicated that the impact of staffing on reliability remains uniform throughout the week.

Assuming a service delivery threshold of 60 seconds, an average reliability of 0.00 is evident when the number of baristas is less than or equal to 2, indicating consistent service delays. While having 3 baristas increases the reliability, it remains only at approximately 17% reliability. A clear sharp increase in service reliability is achieved when the coffee shop has 4 baristas, yielding 97% reliability. This is sufficient; however, 100% reliability will ensure customer satisfaction and increase the number of returning customers. When both 5 and 6 baristas are employed, the coffee shop has a service level of 100%, ensuring consistent customer satisfaction. While 5 baristas are sufficient to meet the required reliability, 6 baristas offer no further improvement to the service level, suggesting that 5 may be the minimum required baristas to achieve full reliability.

Profit vs. Staffing

However, analysis of Figure 41 the line graph of average profit vs. number of baristas suggests otherwise.

The profit graph reveals a consistent pattern across all weekdays: losses at low staffing, breakeven near 5 baristas and a strong profit at 6. This confirms that weekday-specific adjustments are not needed as full staffing benefits every day.

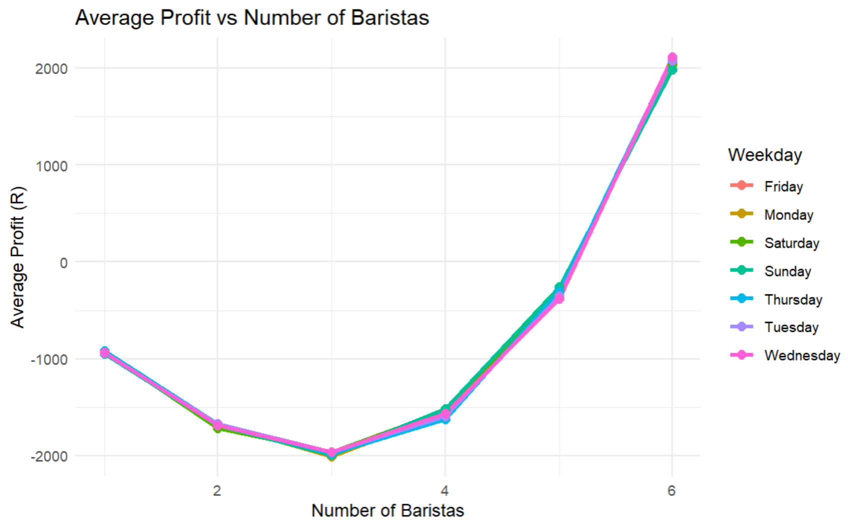


Figure 41: Line Graph of Average Profit vs. Number of Baristas by Weekday

Low staffing showed a loss of profit across all weekdays, with the loss increasing as the number of baristas increased from 1 to 3. This is because, despite having more baristas to contribute to profit, the staffing cost outweighs the revenue, leading to a loss. The profit begins to rise sharply when 4 baristas are employed, suggesting that at this staffing level, service capacity meets demand. From the above figure X, it is evident that reliability also increases at this point. Although a significant increase can be observed when 5 baristas are employed, the net profit remained negative, whereas 6 baristas show a significantly high and positive profit value. Therefore, the optimal number of baristas is 6 as full staffing attains both reliability and maximum profit for the coffee shop.

Overall Service Performance Summary

The overall dataset contains 200000 customer transactions, representing one year of operational data. Using a service reliability threshold of 60 seconds, the analysis found 185083 customers were served reliably, yielding an overall reliability of 92.54%.

Metric	Value
Total Customers	200000
Reliable Customers	185083
Reliability Percent	92.54

Table 16: Table of Metrics for Coffee Shop 1

While the overall reliability figure is high, it masks variation based on the number of staff schedule. For example, a 0% reliability was achieved with 1-2 baristas, while 100% reliability is only reached at 5 barista or more.

Performance by Barista Count

Table 17 below shows the average customer volume, service reliability and net profit across all weekdays for each barista count

Baristas	Avg Customers	Avg Reliability	Avg Profit
1	2.09	0.0000	-937.34
2	10.67	0.0000	-1680.02
3	34.16	0.1646	-1975.21
4	81.17	0.9723	-1564.81
5	156.06	1.0000	-318.23
6	268.73	1.0000	2062.05

Table 17: Performance by Barista Count for Coffee Shop 1

From analyzing the table, it is clear that reliability increased significantly as staffing is increased, reaching 100% at 5 baristas. This indicates that 5 staff members are sufficient to meet the service threshold of 60 seconds.

The average profit based on the number of staff remains negative until 6 baristas, where it turns significantly positive, suggesting that full staffing is required not just for reliability, but also for financial sustainability.

Additionally, customer volumes scale with staffing, reflecting real-world scheduling patterns, where busier days tend to be matched with higher staffing levels. This analysis does not assume fixed demand but rather observes performance under actual operating conditions.

Coffee Shop 2

Reliability vs. Staffing

When assuming a service delivery threshold of 60 seconds, reliability vs. the number of baristas graph, in figure 42, shows a 0% performance across all barista counts. This suggests the threshold may be too strict for Coffee Shop 2's operational context.

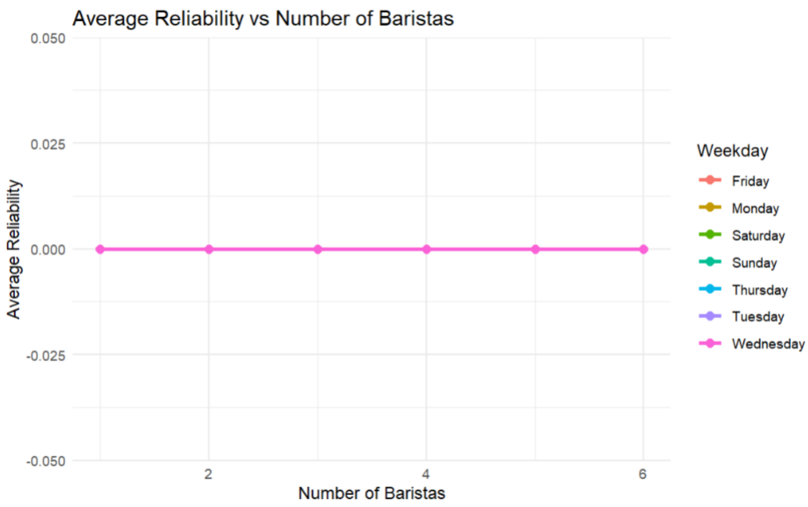


Figure 42: Average Reliability vs. Number of Baristas at a 60 second threshold

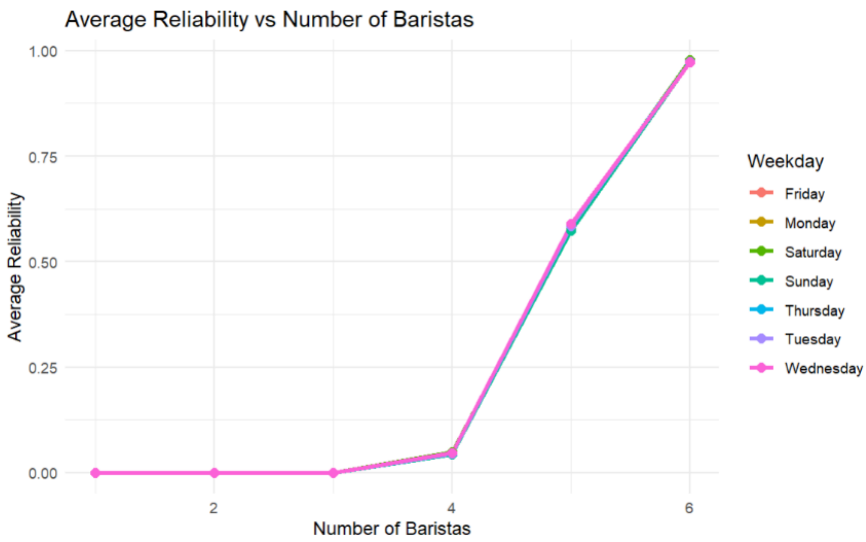


Figure 43: Average Reliability vs. Number of Baristas at a 90 second threshold

When the service threshold is adjusted to 90 seconds, the reliability graph reveals a clear upward trend as seen in figure 43. Reliability remains at 0% for staffing levels of 1 to 3 baristas, indicating consistent delays. A slight improvement is observed at 4 baristas, with reliability increasing to 4.55%. A sharp rise follows at 5 baristas, reaching 58.42%, and reliability continues to improve at 6 baristas, peaking at 97.44%. Despite this near-perfect performance, full reliability is never achieved even at maximum staffing under the original 60-second threshold, confirming that the stricter benchmark is operationally unrealistic for Coffee Shop 2.

Profit vs. Staffing

Figure 44, the line graph of average profit vs. number of baristas under the 90-second threshold, shows a consistent pattern across all weekdays, with sustained losses at low staffing levels, gradual improvement with increased staffing, and strong profitability only at full staffing. This indicates that weekday-specific adjustments are unnecessary, as the staffing-profit relationship holds consistently.

From 1 to 3 baristas, profit remains negative and worsens as staffing increases, due to rising labour costs without sufficient throughput. At 4 baristas, profit begins to increase, though still negative, suggesting that service capacity starts to meet demand. This aligns with the reliability graph in figure 43, which also shows initial improvement at this point. With 5 baristas, profit continues to rise but remains below the breakeven point. Only at 6 baristas does the coffee shop achieve a significantly positive profit margin, confirming that full staffing is required not only for high reliability but also for financial sustainability. Therefore, the optimal staffing level for Coffee Shop 2 is 6 baristas, where both service performance and profitability are maximized.

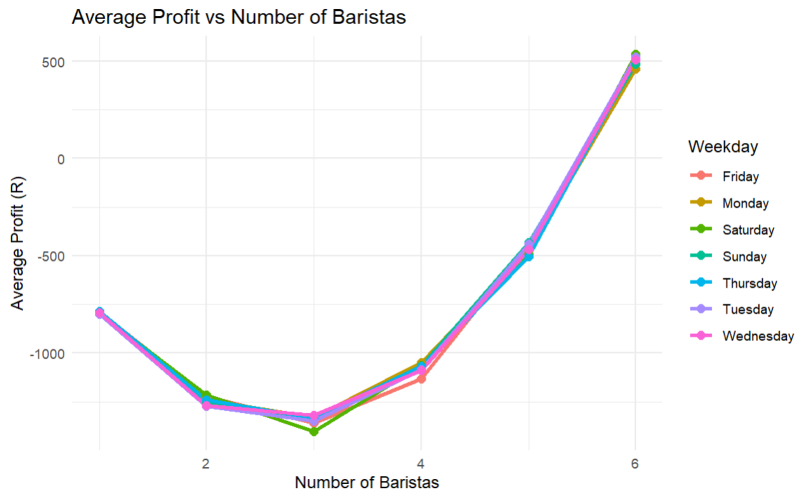


Figure 44: Line Graph of Average Profit vs. Number of Baristas by Weekday

Overall Service Performance Summary

The dataset comprises 200,000 customer transactions over one year of operations. Using a 90-second service reliability threshold, the analysis found that 110623 customers were served reliably, resulting in an overall reliability of 55.31%.

Metric	Value
Total Customers	200000
Reliable Customers	110623
Reliability Percent	55.31

Table 18: Table of Metrics for Coffee Shop 2

While the overall reliability provides a general view of performance, it conceals significant variation based on staffing levels. For example, reliability remains at 0% for 1–3 baristas, with only slight improvement at 4 baristas. Even at maximum staffing, full reliability is not achieved, peaking at 97.44% with 6 baristas. This highlights the importance of adjusting the service threshold and optimizing staffing to achieve greater service reliability.

Performance by Barista Count

Table 19 summarizes the average number of customers, service reliability, and total profit across all weekdays for each barista count under the 90-second service threshold.

Baristas	**Avg Customers**	**Avg Reliability**	**Avg Profit**
1	6.92	0.0000	-792.32
2	25.13	0.0000	-1246.15
3	55.07	0.0000	-1347.98
4	97.36	0.0455	-1079.09
5	151.20	0.5842	-463.99
6	216.83	0.9744	504.91

Table 19: Performance by Barista Count for Coffee Shop 2

From the table, it is evident that service reliability improves steadily with increased staffing, reaching 97.44% at 6 baristas. However, full reliability is never achieved, even at maximum staffing, under the 90-second threshold. This suggests operational constraints that limit service speed regardless of staffing.

Profitability follows a similar trend. Profit losses persist from 1 to 5 baristas, with the loss narrowing as staffing increases. Only at 6 baristas does the coffee shop achieve a positive profit margin, indicating that full staffing is essential not only for high reliability but also for financial sustainability.

Customer volumes also scale with staffing, reflecting real-world scheduling patterns where higher demand is matched with increased staff. This analysis does not assume fixed demand but instead captures performance under actual operating conditions.

Taguchi Loss

The Taguchi loss function can be used to represent the gradual decline in customer satisfaction as service time deviates from the 60-second target. Rather than a simple pass/fail measure of service quality, it models increasing loss for longer waiting times, capturing the impact of reduced future sales and business reputation. When this loss is incorporated into the analysis, the results suggest that employing slightly more baristas may be optimal, as improved service consistency compensates for additional staffing costs.

ANOVA

To determine whether delivery performance changed between 2026 and 2027, a one-way ANOVA was conducted for each product category. The null hypothesis (H_0) states that there is no difference in mean delivery hours between the two years. The alternative hypothesis (H_1) states that a difference exists. We reject H_0 if the p-value is less than 0.05. The results by product category can be found below.

Mouse

```
=== ANOVA for MOU ===
              Df Sum Sq Mean Sq F value
orderYear      1     20   19.94    0.53
Residuals  20660 777830   37.65        
            Pr(>F)
orderYear    0.467
Residuals         
```

Figure 45: ANOVA for Mouse Product Category

The p-value is 0.467, so we fail to reject the null hypothesis. There is no statistically significant difference in mean delivery hours between the two years.

Keyboard

```
=== ANOVA for KEY ===
              Df Sum Sq Mean Sq F value
orderYear      1    299  299.33    8.07
Residuals  17918 664603   37.09        
             Pr(>F)   
orderYear   0.00451 **
Residuals             
```

Figure 46: ANOVA for Keyboard Product Category

The p-value is 0.00451, so we reject the null hypothesis. There is a statistically significant difference in mean delivery hours between the two years.

Software

```
=== ANOVA for SOF ===
              Df Sum Sq Mean Sq F value
orderYear      1      0 0.01695   0.179
Residuals  20747   1966 0.09475        
            Pr(>F)
orderYear    0.672
Residuals         
```

Figure 47: ANOVA for Software Product Category

The p-value is 0.672, so we fail to reject the null hypothesis. There is no statistically significant difference in mean delivery hours between the two years.

Cloud Subscription

```
=== ANOVA for CLO ===
              Df Sum Sq Mean Sq F value
orderYear      1      1    1.17   0.031
Residuals  15596 583187   37.39        
            Pr(>F)
orderYear     0.86
Residuals         
```

Figure 48: ANOVA for Cloud Subscription Product Category

The p-value is 0.86, so we fail to reject the null hypothesis. There is no statistically significant difference in mean delivery hours between the two years.

Laptop

```
=== ANOVA for LAP ===
              Df Sum Sq Mean Sq F value
orderYear      1     18   18.15   0.496
Residuals  10205 373353   36.59
            Pr(>F)
orderYear    0.481
Residuals
```

Figure 49: ANOVA for Laptop Product Category

The p-value is 0.481, so we fail to reject the null hypothesis. There is no statistically significant difference in mean delivery hours between the two years.

Monitor

```
=== ANOVA for MON ===
              Df Sum Sq Mean Sq F value
orderYear      1     16   16.36   0.447
Residuals  14862 543499   36.57
            Pr(>F)
orderYear    0.504
Residuals
```

Figure 50: ANOVA for Monitor Product Category

The p-value is 0.504, so we fail to reject the null hypothesis. There is no statistically significant difference in mean delivery hours between the two years.

Reliability of Services

Estimating Reliable Service Days

The given bar chart as seen in figure 51 below shows the number of workers present over a 397-day period. The distributing is clearly skewed toward full staffing, with 16 employees present on 270 days. From this we can assume that the company employed 16 staff-members full time.

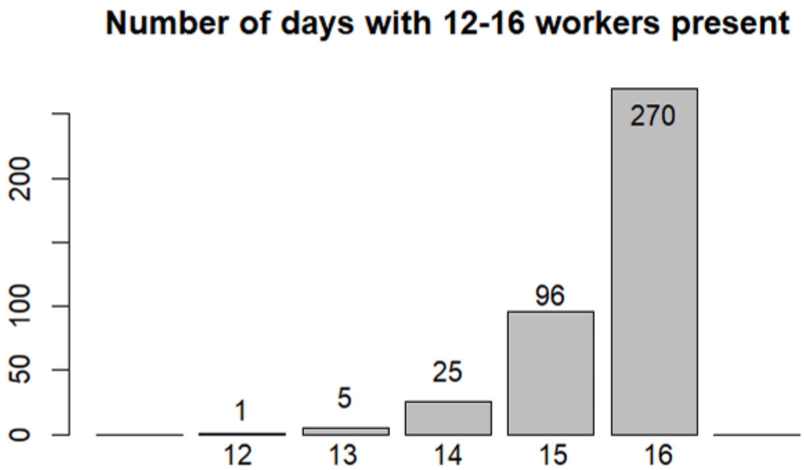


Figure 51: Given Bar Chart of Days vs. Employees Present

To estimate the probability of an individual employee being present, the total number of employee-days worked was divided by the theoretical maximum, which is 16 employees for 397 days. This yields an attendance probability of 0.936369. This implies that on average, each employee is absent for approximately 9.5 days a year.

Reliable service can be expected when we have at least 15 employees on duty. Using binomial probability yields a reliable service level on any given day as:

$$P(15) + P(16) = 0.093 + 0.843 = 0.936$$

Equation 1 : Service Level Reliability on any Given Day

The estimated reliable service days per year is calculated as follows:

$$0.936369 \times 365 = 341.7714$$

Equation 2: Estimated Service Reliability per Year

Optimising Company Profit

A R20 000 loss in sales result from each day with less than 15 employees. Additional employees can be hired at R25 000 per month, resulting in an accumulated cost of R300 000 annually per employee. A binomial model was used to evaluate the trade-off between hiring costs and lost sales across different staffing levels from 12 to 20 employees. Table 20 below summarizes the projected annual cost for each staffing level.

Workers	Loss_sales	Cost_workers	Total_cost
12	7 300 000	0	7 300 000.0
13	7 300 000	0	7 300 000.0
14	7 300 000	0	7 300 000.0
15	2 381 108	0	2 381 108.3
16	464 506. 2	0	464 506.2
17	66 219.82	300000	366 219.8
18	7 592.97	600000	607 593.0
19	739.94	900000	900 739.9
20	63.49	1200000	1 200 063.5

Table 20: Projected Annual Cost per Staffing Level

Figure 52 illustrates the trade-offs between lost sales, hiring costs and total costs. The total cost curve shows a steep decline from 12 to 17 workers, followed by a gradual increase from 17 to 20 workers.

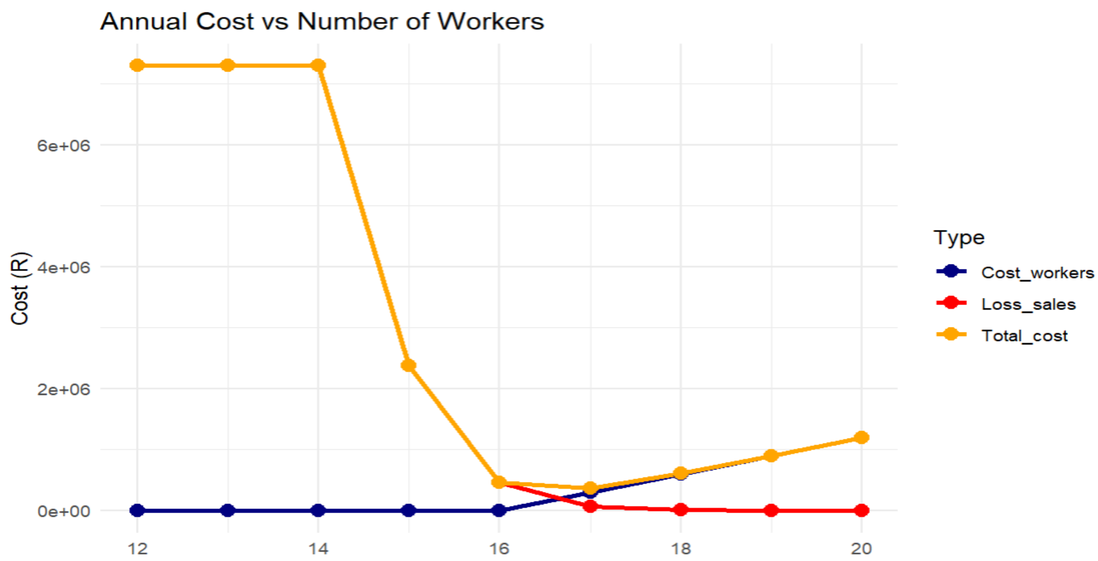


Figure 52: Annual Costs vs. Number of Workers

From the above table and figure 52, it is evident that staffing levels below 15 employees result in severe financial losses, with total costs exceeding R7.3 million annually dur to unreliable service. As the staffing increases, the cost of lost sales drops sharply, while hiring costs begin to rise from 17 workers onward. The lowest total cost of R366 219.80 is observed at 17 workers, where the improved reliability reduces the sales losses and offsets the additional hiring expenses. Beyond 17 workers, the marginal increases in reliability does not justify the increased staffing costs, resulting in a gradual rise of the total costs. Thus, 17 workers are the optimal staffing level as the total cost is kept to a minimum, making it the most cost efficient while sufficient service level is achieved.

Conclusions

Data analytics provides a framework for understanding complex operational and performance data, allowing for the identification of trends, patterns and relationships that might otherwise remain hidden within a dataset. By integrating both quantitative and qualitative approaches, this report demonstrates how insights from data can inform decision-making, optimize outcomes, and enhance organizational efficiency. Ultimately, the outcomes highlight the value of data-driven strategies in supporting improvements and continual management across diverse industry contexts (Stellenbosch University QA 344, 2025).

References

Bhandari, P., 2021. *Scribbr.* [Online]
Available at: https://www.scribbr.com/statistics/type-i-and-type-ii-errors/
[Accessed 7 October 2025].

Stellenbosch University QA 344, 2025. *Project ECSA 2025 FINAL [PDF].* s.l.:s.n.

OpenAI. (2025) *ChatGPT (GPT-5)* [AI model]. Used to assist in writing R code for data analysis in the report. Available at: https://chat.openai.com [Accessed: 22 September - 23 October 2025].