Quality Assurance 344

ECSA Project

Pieter Johannes Laubscher

US NO 27038688

Contents

Introduction

This report is organized into seven main sections, covering basic data analysis, statistical process control, risk mitigation, data correction, optimization, ANOVA testing and service reliability, each addressing the objectives outlined in the QA 344 project brief. Throughout the report, R Studio was used for data processing, visualizations and calculations to extract valuable insights from the provided datasets. The analysis begins with exploring customer and product data, followed by monitoring delivery performance and process capability. Risk mitigation and data correction ensure accurate datasets, while optimization focuses on staffing and profit maximization. ANOVA testing examines differences across time periods and service reliability assesses operational efficiency. All R code used to generate the results in this report is included in the submission to support all findings.

Basic Data Analysis

Data Information

Table 1: Merged dataset created true R (size of data: 101 000 x 18)

CustomerID	ProductID	Quantity	orderTime	orderDay	orderMonth	orderYear	pickingHours	deliveryHours	Category.x	Description.x	SellingPrice.x	Markup.x	Gender	Age	Income	City
CUST1791	CLO011	16	13	11	11	2022	17.72166667	24.544	Keyboard	burlywood silk	1070.54	16.41	Male	39	1.00E+05	Los Angeles
CUST3172	LAP026	17	17	14	7	2023	38.39083333	31.546	Cloud Subscription	aliceblue silk	18711.72	13.51	Female	58	90000	Chicago
CUST1022	KEY046	11	16	23	5	2022	14.72166667	21.544	Monitor	blueviolet silk	708.18	17.72	Female	20	95000	Seattle
CUST3721	LAP024	31	12	18	7	2023	41.39083333	24.546	Mouse	blueviolet marb	18366.92	29.35	Female	66	60000	Miami
CUST4605	CLO012	20	14	7	2	2022	15.72166667	24.044	Mouse	azure silk	963.14	10.13	Female	70	25000	Chicago

Table 2: Summary of numerical variables

	Min	1st	Median	Mean	3rd	Max
Quantity	1	3	6	13.5	23	50
orderTime	1	9	13	12.93	17	23
orderDay	1	8	15	15.5	23	30
orderMonth	1	4	6	6.448	9	12
pickingHours	0.4259	9.39	14.055	14.6955	18.72	45.05
deliveryHours	0.2772	11.546	19.54	17.47	25.044	38.046
SellingPrice.x	350.4	493.7	627.9	3243.8	5346.1	19725.2
Markup.x	10.13	16.18	20.44	20.42	25.56	29.84
Age	16	33	51	51.57	69	105
Income	5000	55000	85000	80699	105000	140000

Table 3: Frequency counts for categorical variables

Category.x	Count
Monitor	16831
Keyboard	16672
Software	16656
Laptop	16616
Mouse	16537

Customer Information

The customer_data sheet provides detailed information on 5 000 clients, including their gender, age, income and city of residence. A sample of the data is presented below for reference.

CustomerID	Gender	Age	Income	City
CUST001	Male	16	65000	New York
CUST002	Female	31	20000	Houston
CUST003	Male	29	10000	Chicago
CUST004	Male	33	30000	San Francisco

Figure 1: Gender Distribution of Customers

The bar graph shows that the number of male and female clients is balanced, with a slightly higher number of females. A small number of clients are categorized as "other," which may result from data entry errors or non-standard classifications. Overall, the gender distribution offers little insight due to the relatively even split between male and female clients.

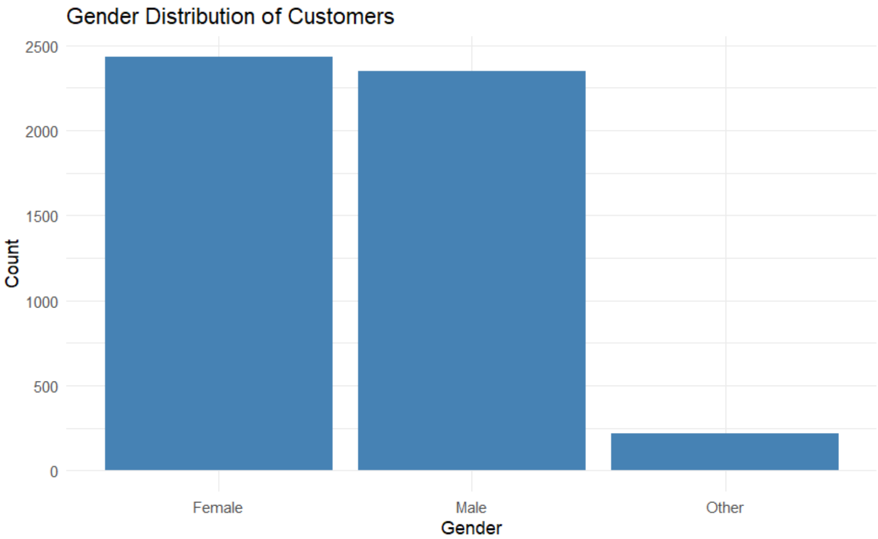


Figure 2: Customer Age Distribution

The Density line graph shows that most clients are young adults, with the highest concentration between 30 and 35 years. Beyond 75, the number of clients gradually declines, with the oldest registered client just over 100 years. Overall, the data indicates that most users are young to middle-aged.

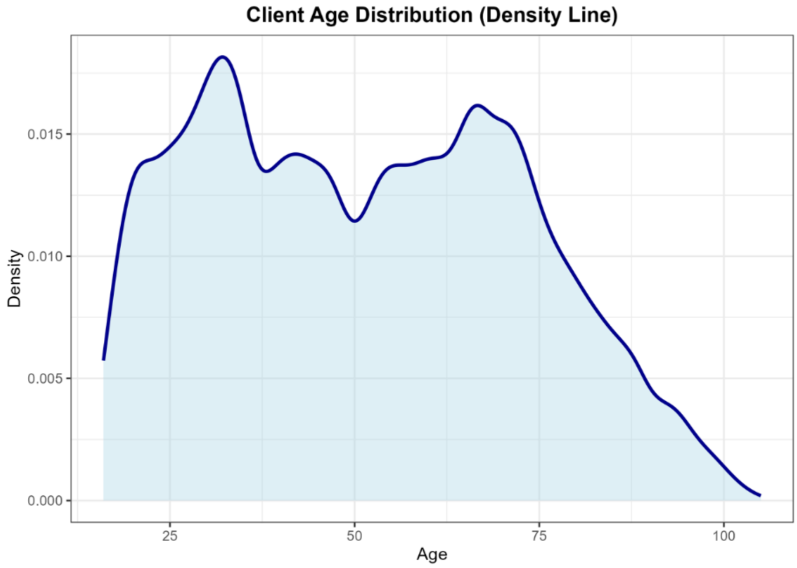


Figure 3: Average Income by Gender

The bar chart shows that the average income is very similar across all genders, suggesting that gender does not appear to influence earnings in this dataset.

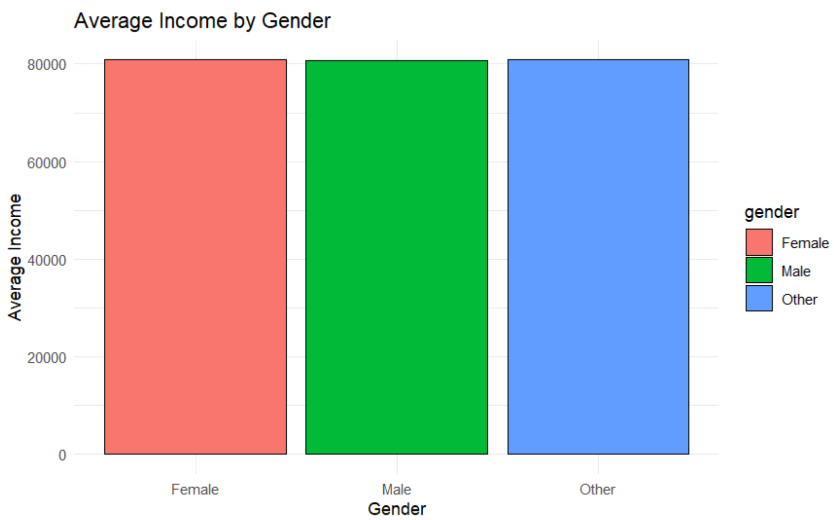
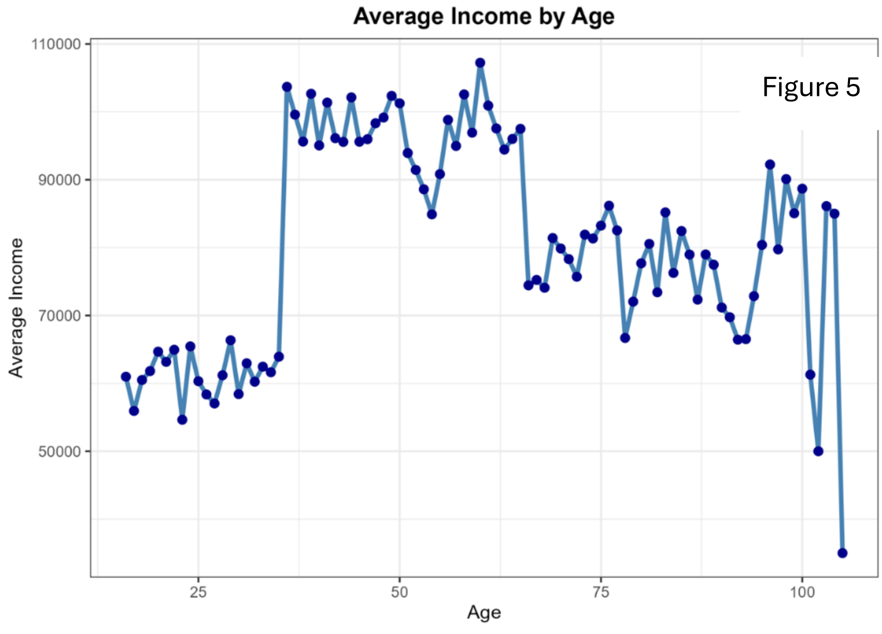


Figure 4: Average Income per City

The bar chart indicates that income levels are similar across towns, suggesting that a customer's place of residence is not a strong predictor of their income.

Figure 5: Average Income by Age

The line graph shows the average income of customers by age. There is a noticeable increase in income between younger and middle-aged customers, peaking between 35 and 65 years. After 65, average income declines, likely due to retirement. This insight highlights that middle-aged customers are the highest earners and therefore represent the most valuable target audience for marketing and sales efforts. Age proves to be a strong predictor of income and can be effectively used to tailor product offerings and promotions.



Product Information

Product information was provided in two separate files: products_data and products_Headoffice_data. These files were combined during the data loading phase using the ProductID as the key. Each file contains details such as category, description, selling price and markup of the products. A sample of the data is presented below for reference:

ProductID	Category	Description	SellingPrice	Markup
SOF001	Software	coral matt	511.53	25.05
SOF002	Cloud Subscription	cyan silk	505.26	10.43
SOF003	Laptop	burlywood marble	493.69	16.18

Figure 6: Average Selling Price per Category

The bar chart illustrates the average selling price across different product categories. It shows that all products are priced within a relatively narrow range of approximately R2 900 to R4 500. This indicates that product category has little influence on selling price, as prices remain consistent across all categories. The limited variation in selling prices suggests that the business targets a specific group of customers with similar income levels. Based on earlier findings, this group most likely falls within the 35–65 age range, earning enough to comfortably afford products with an average price of around R3 000 each.

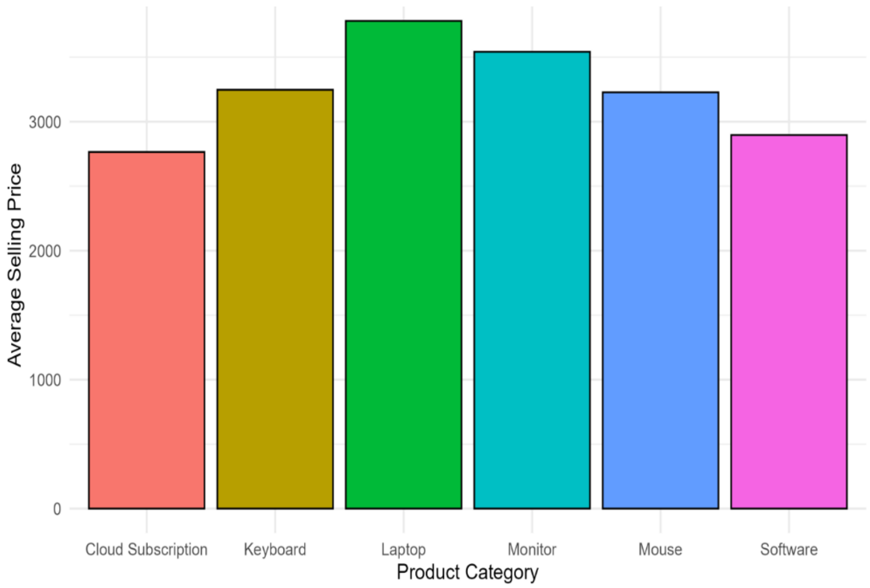


Figure 7: Selling Price Distribution per Category

As observed earlier, the average selling price across all product categories remains relatively consistent, which is also reflected in this boxplot. What stands out is that most product categories, except for laptops and monitors, have very narrow price ranges, with minimal differences between their minimum and maximum selling prices. In contrast, the laptop and monitor categories display a few outliers, indicating that while most of these products are sold at similar prices to other categories, certain items within these groups can be sold at significantly higher prices than the overall average.

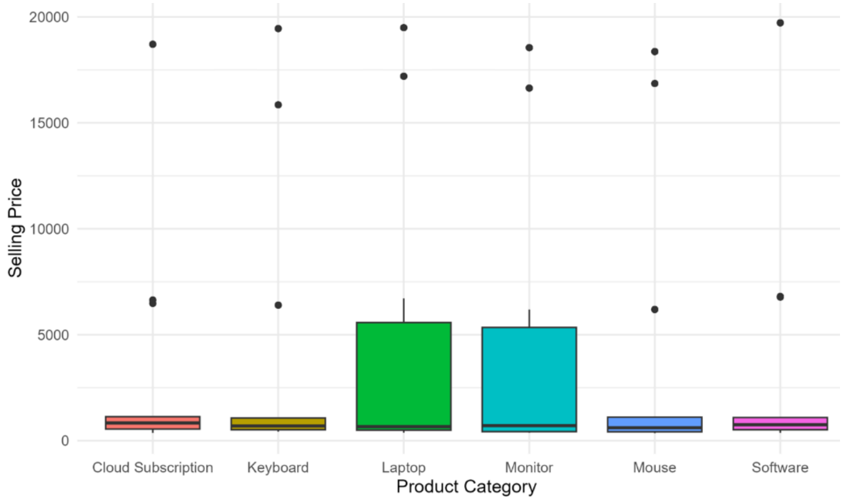
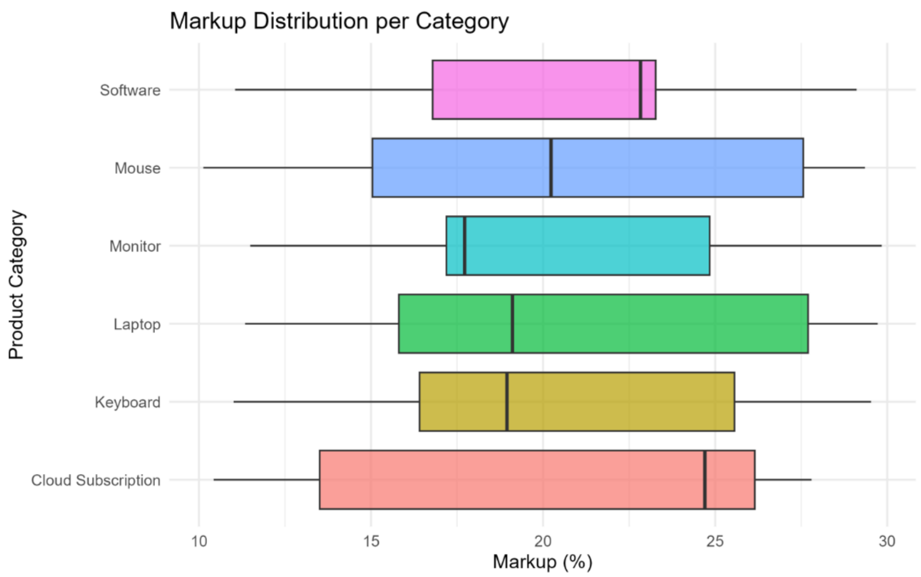


Figure 8: Markup Distribution per Category

The boxplot displays the markup distribution for each product category. There is noticeable variation in both the mean values and the interquartile ranges across categories, indicating that there is no strong or consistent relationship between product category and markup. Although the average markup for all categories falls between 15% and 25%, this relatively small range suggests that markup alone provides limited insight into individual product differences, as it remains consistent across categories.



SPC limits

The projected sales data for 2026 and 2027 was provided in a CSV file containing details such as sales quantity, order time, order day, month, year, as well as picking and delivery hours.

To standardize the timeline, the order time was converted into an absolute day count, where January 1st of the first year was designated as Day 1, and the time of day was expressed as a decimal value.

The dataset was then organized by product category, and samples of 24 observations were taken from each group. For every sample, the average ($\bar{X}$) and standard deviation (s) were calculated to form the basis for subsequent control chart analysis.

Initial control charts:

The first 30 samples for each product category were used to establish the initial control chart limits. From these samples, the upper and lower control levels corresponding to one, two, and three sigma's (UCL and LCL) were calculated for each category. The charts are illustrated below:

Figure 9: Keyboard Initial Control charts

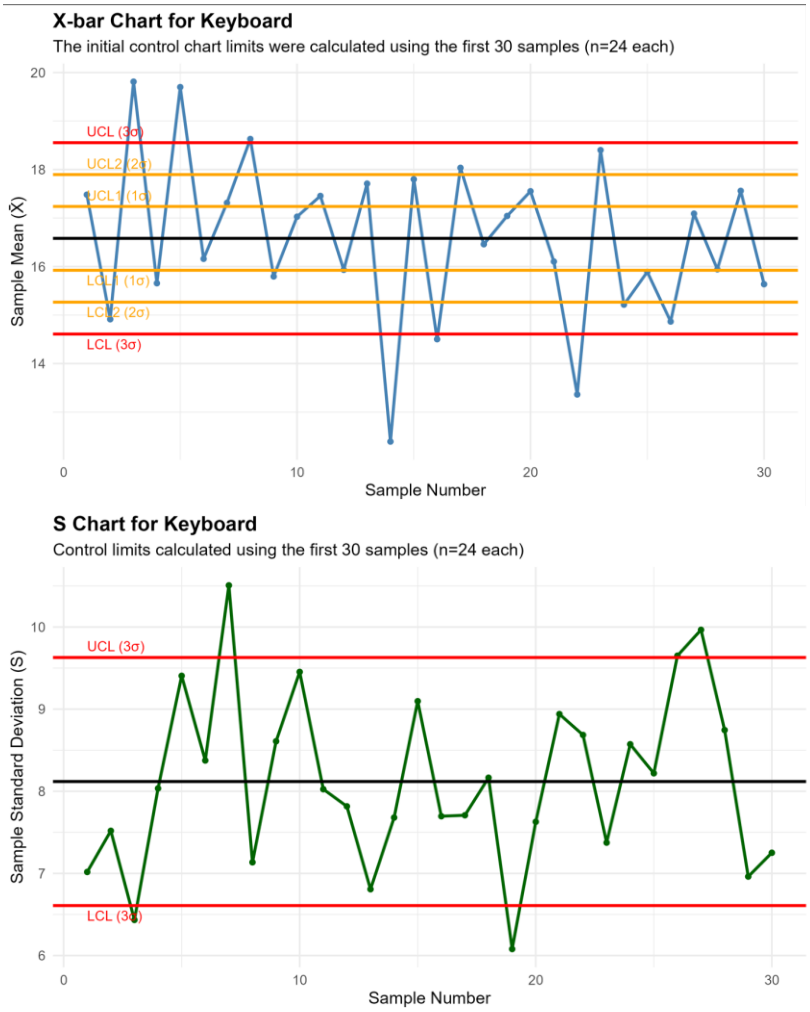
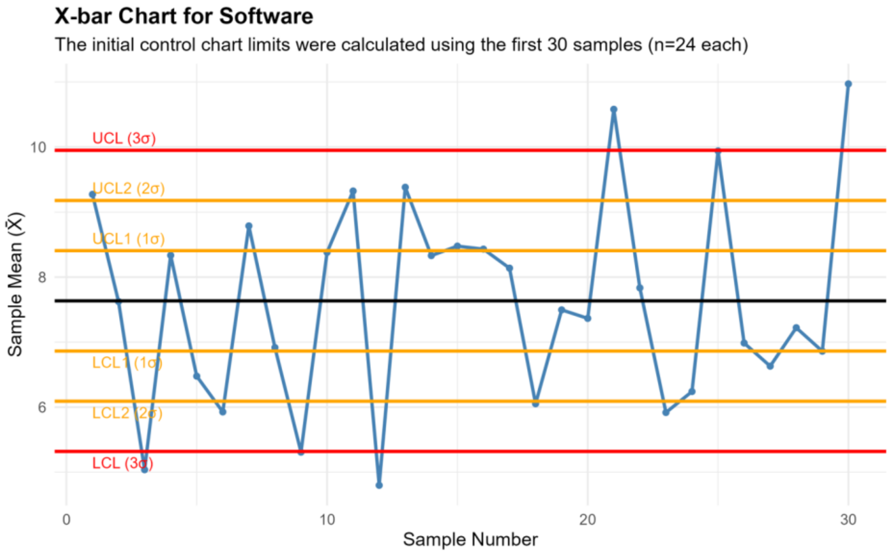


Figure 10: Software Initial Control Charts



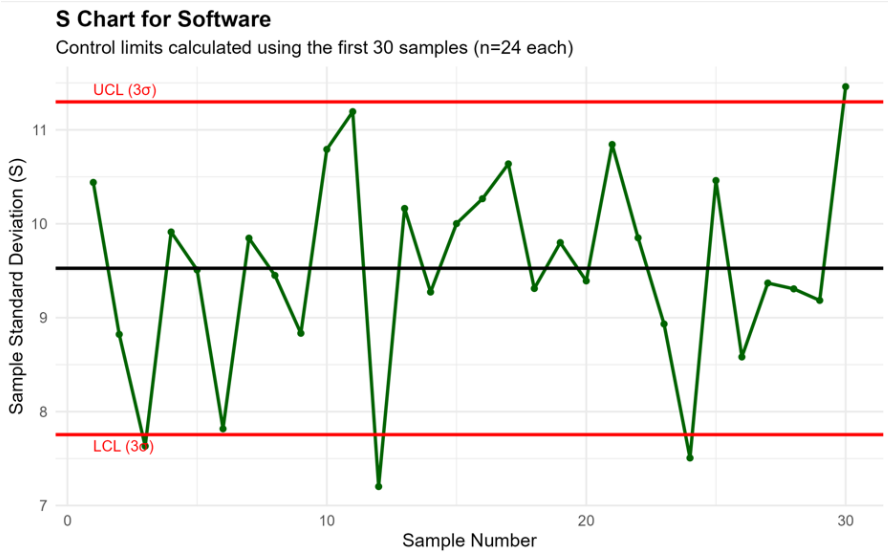
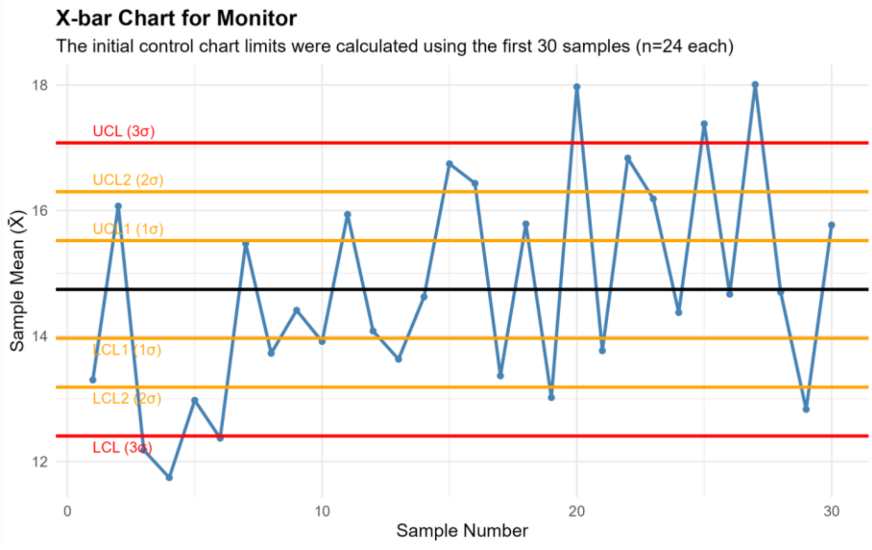


Figure 11: Monitor Initial Control Charts



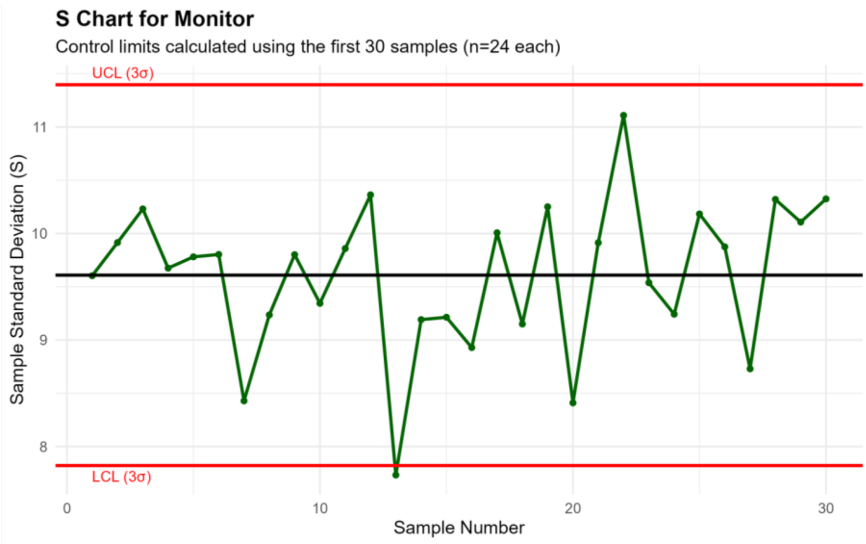
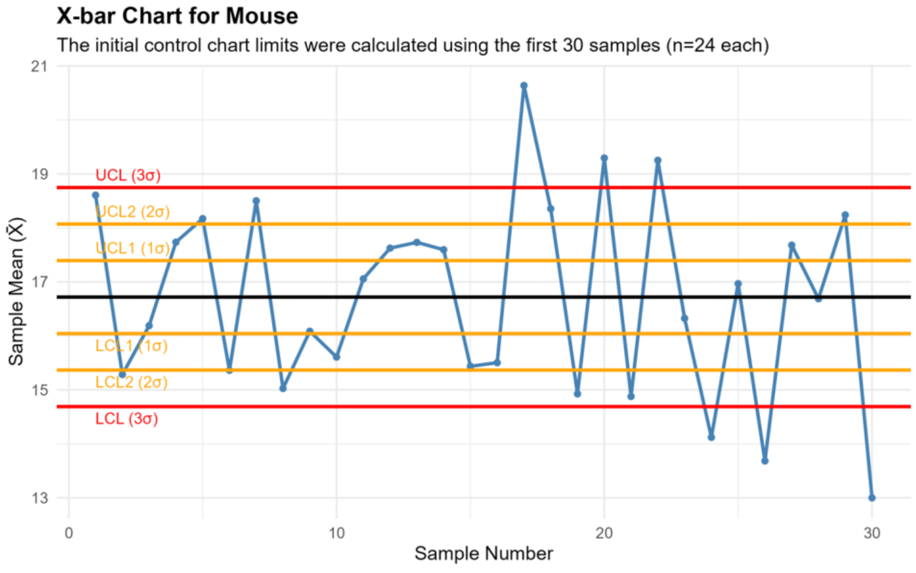


Figure 12: Mouse Initial Control Charts



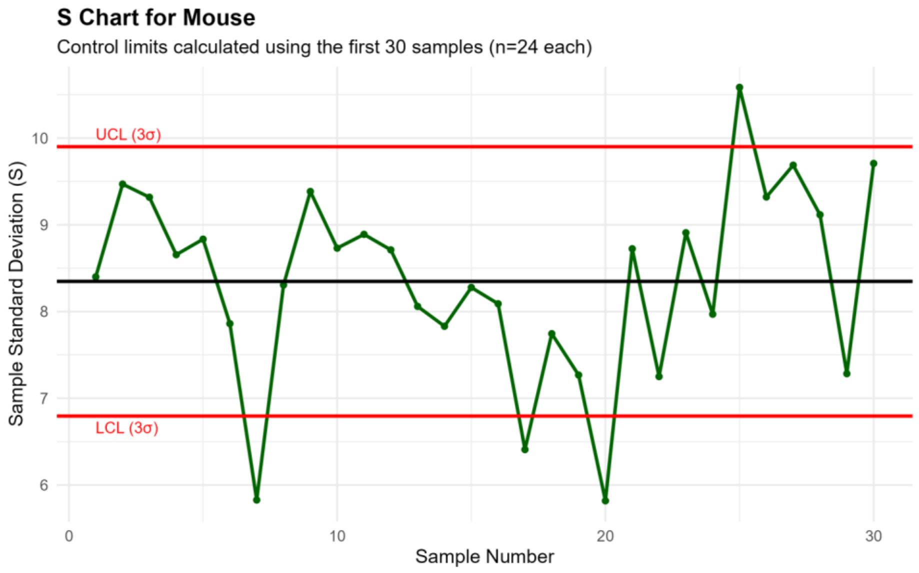
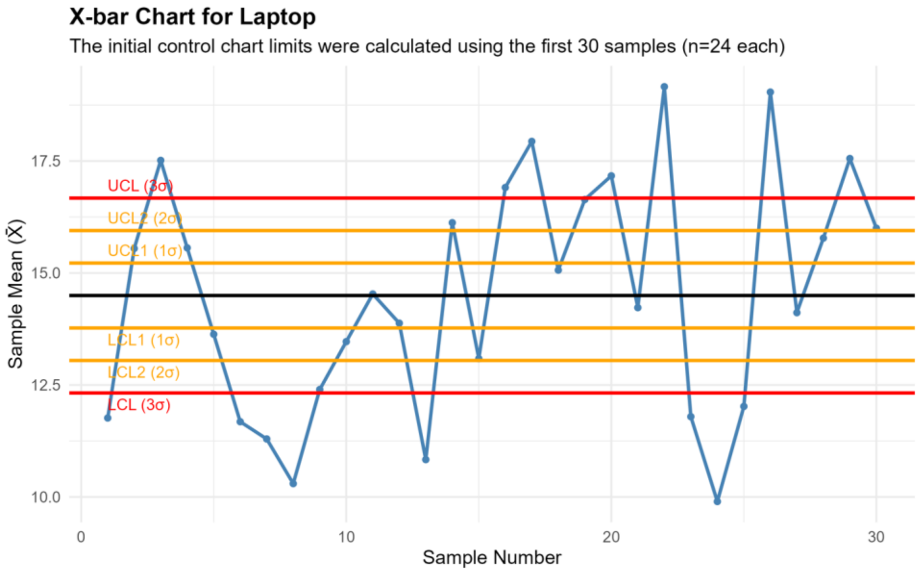


Figure 13: Laptop Initial Control Charts



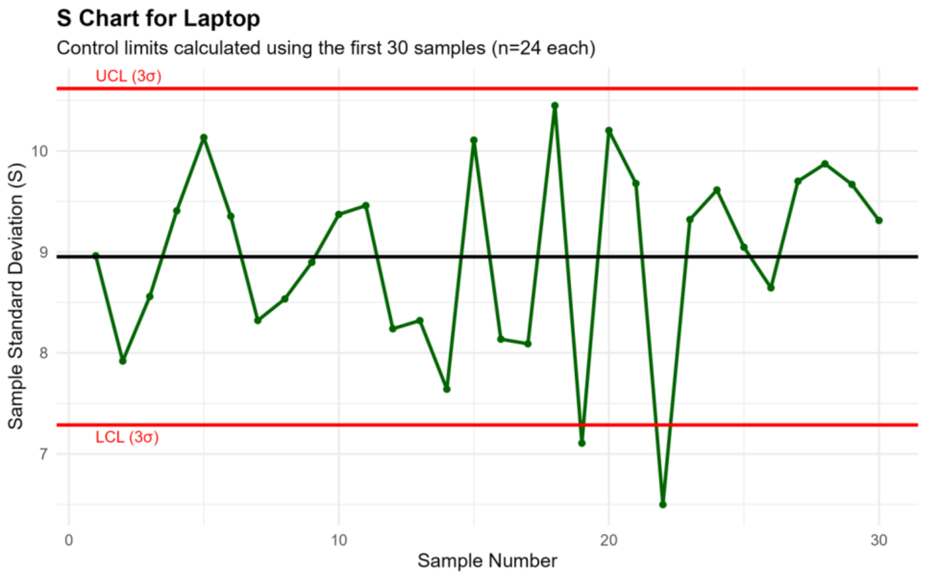
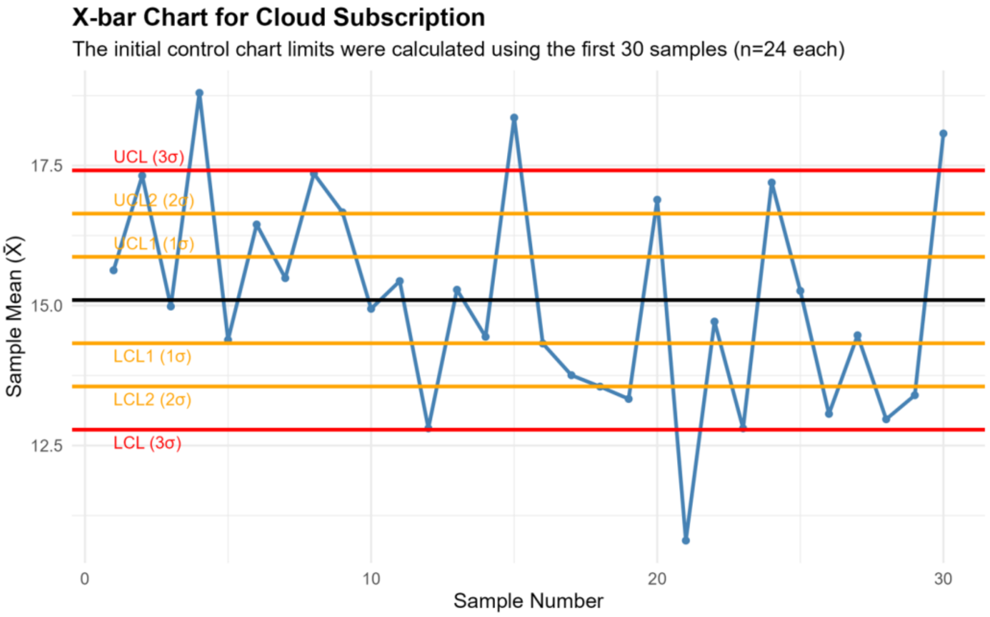
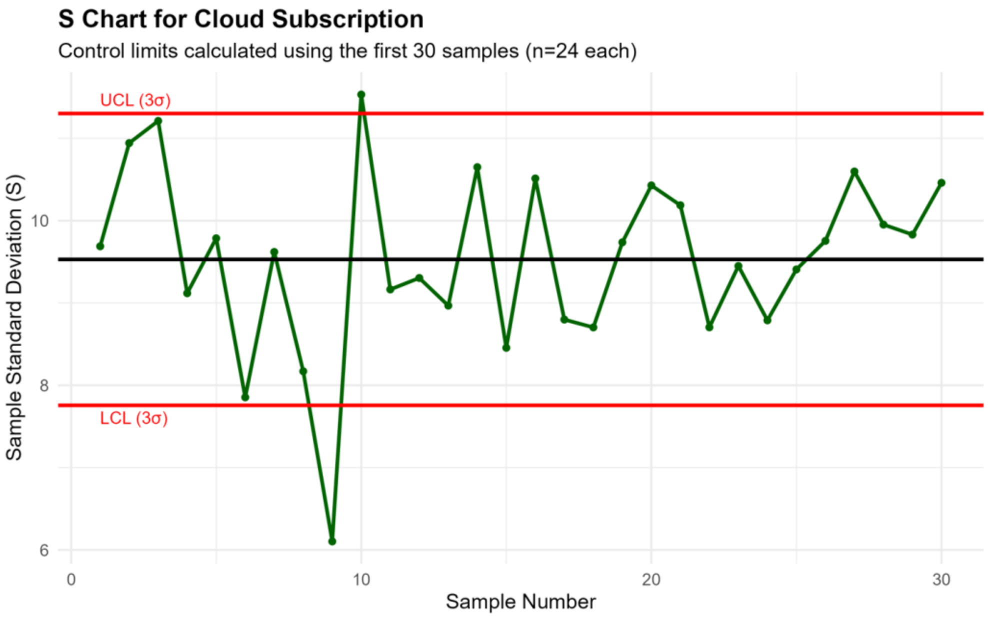


Figure 14: Cloud Subscription Control Chart

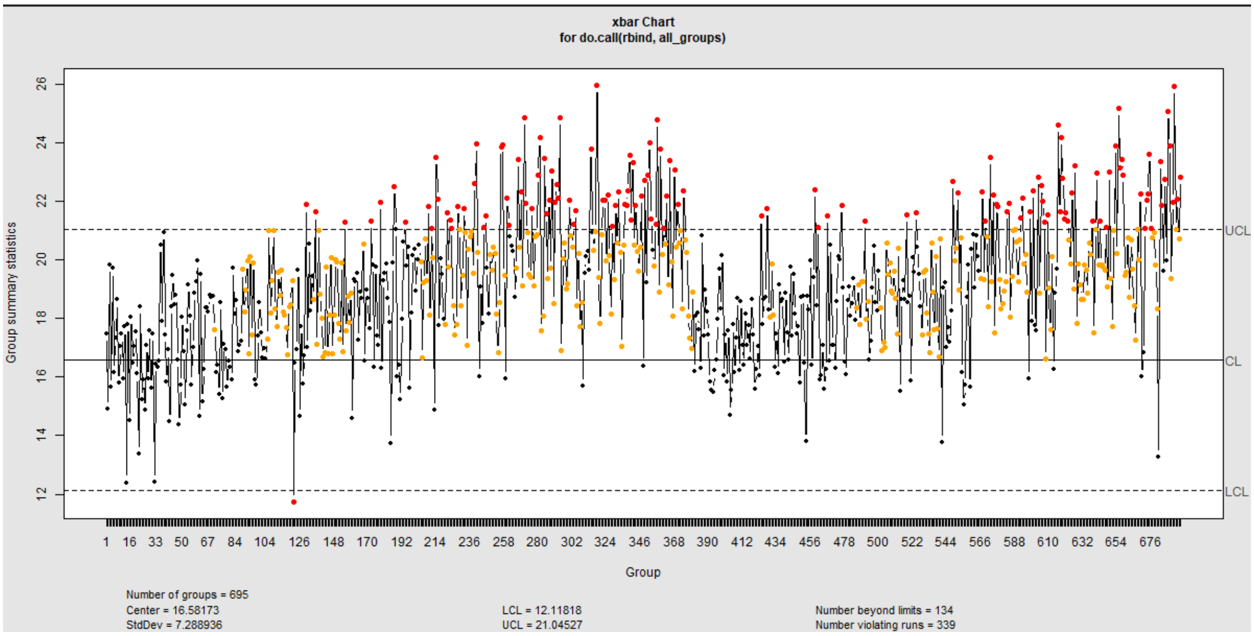




Continued Control Charts

The control charts were updated in increments of twenty-four (24) samples until all the data was included in the chats, not only the first 30 samples. The final control charts are shown below:

Figure 15: Keyboard Continued Control charts



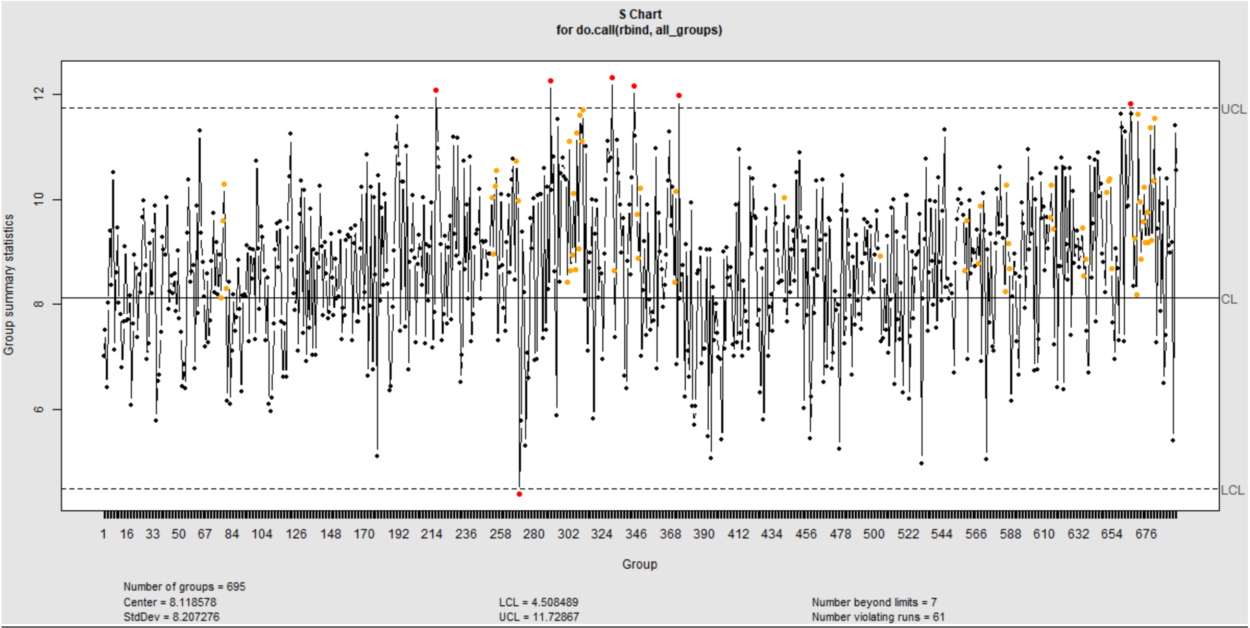
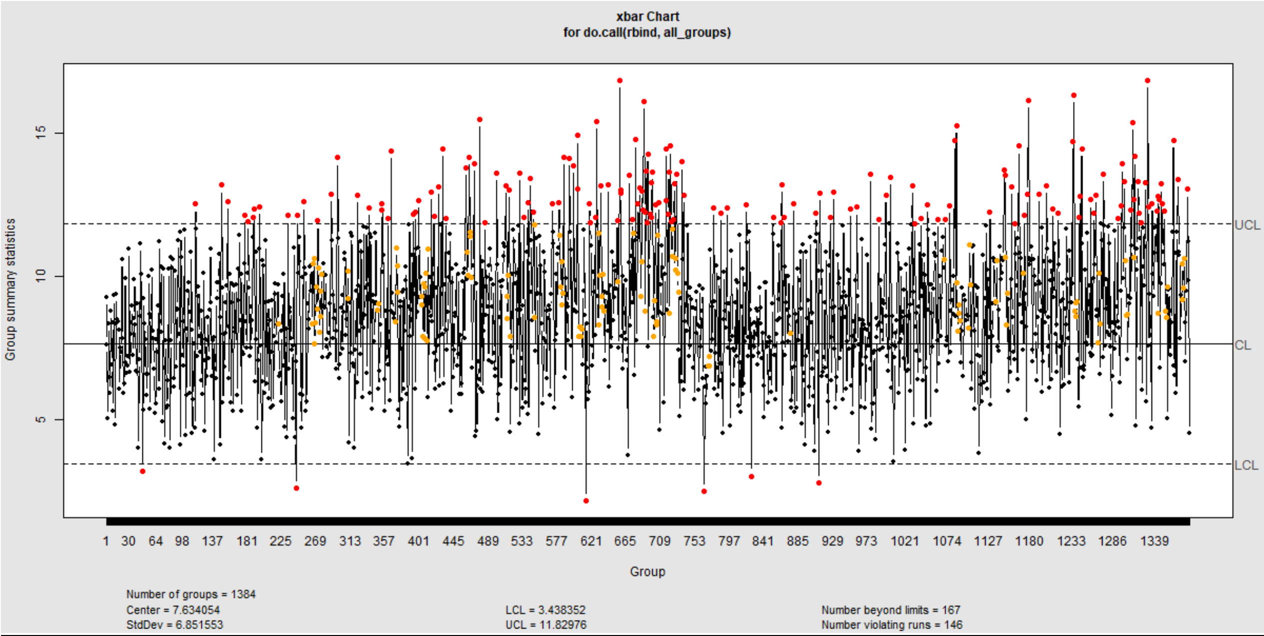


Figure 16: Software Continued Control Charts



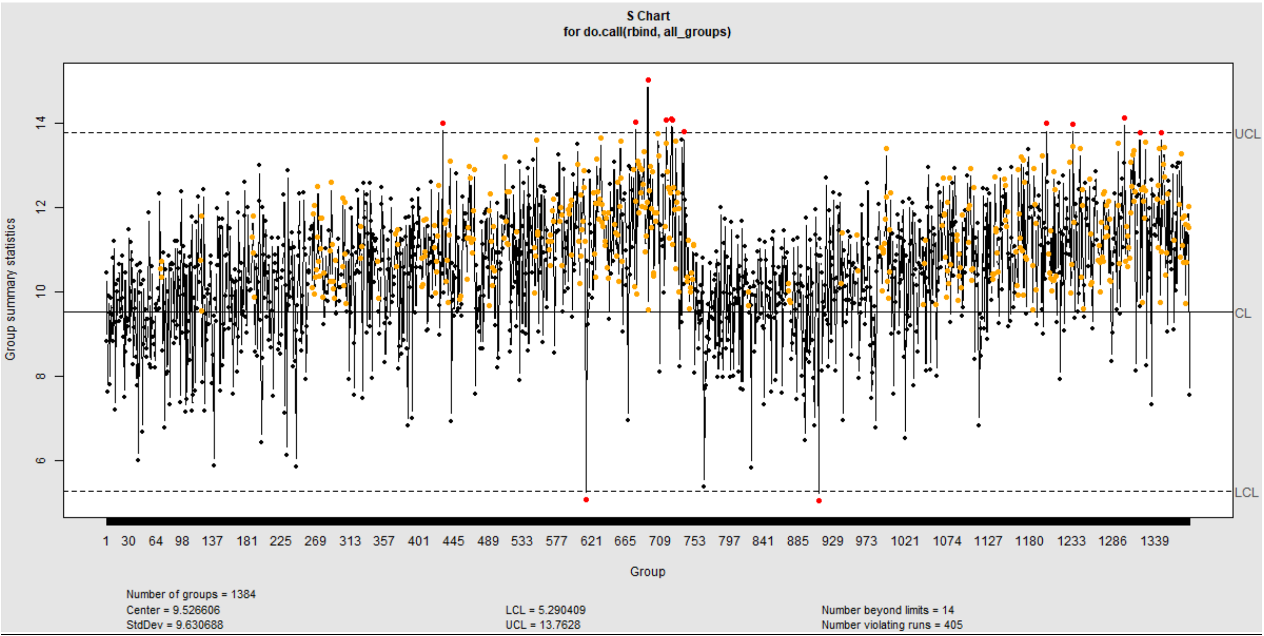
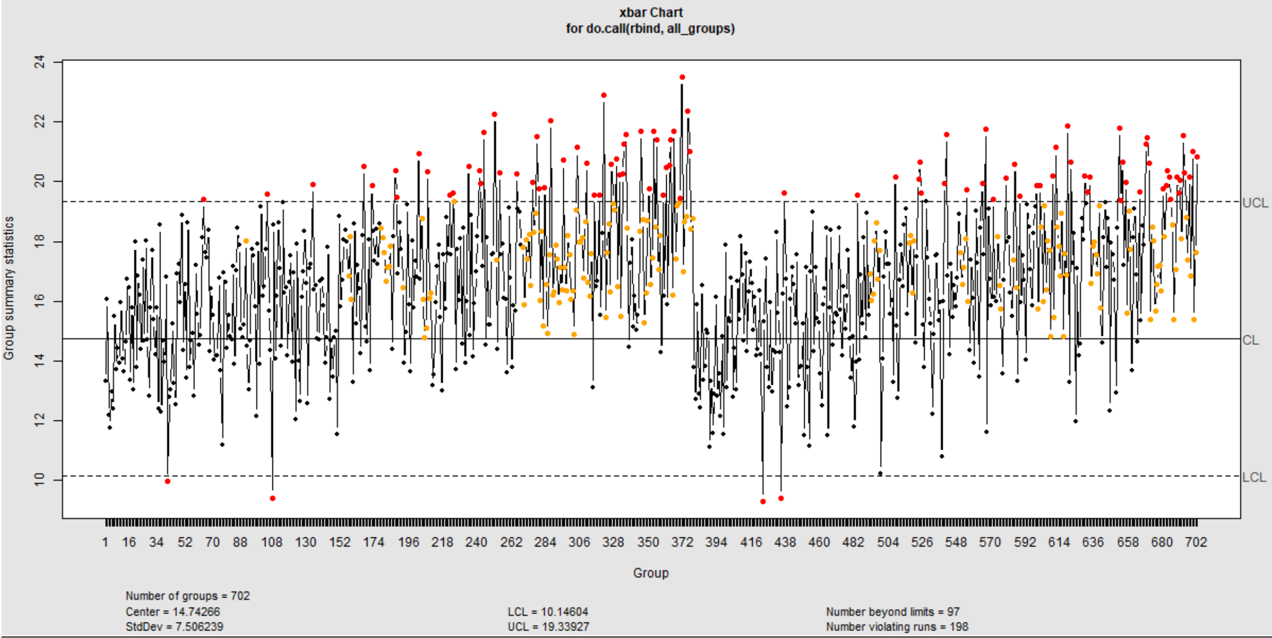


Figure 17: Monitor Continued Control Charts



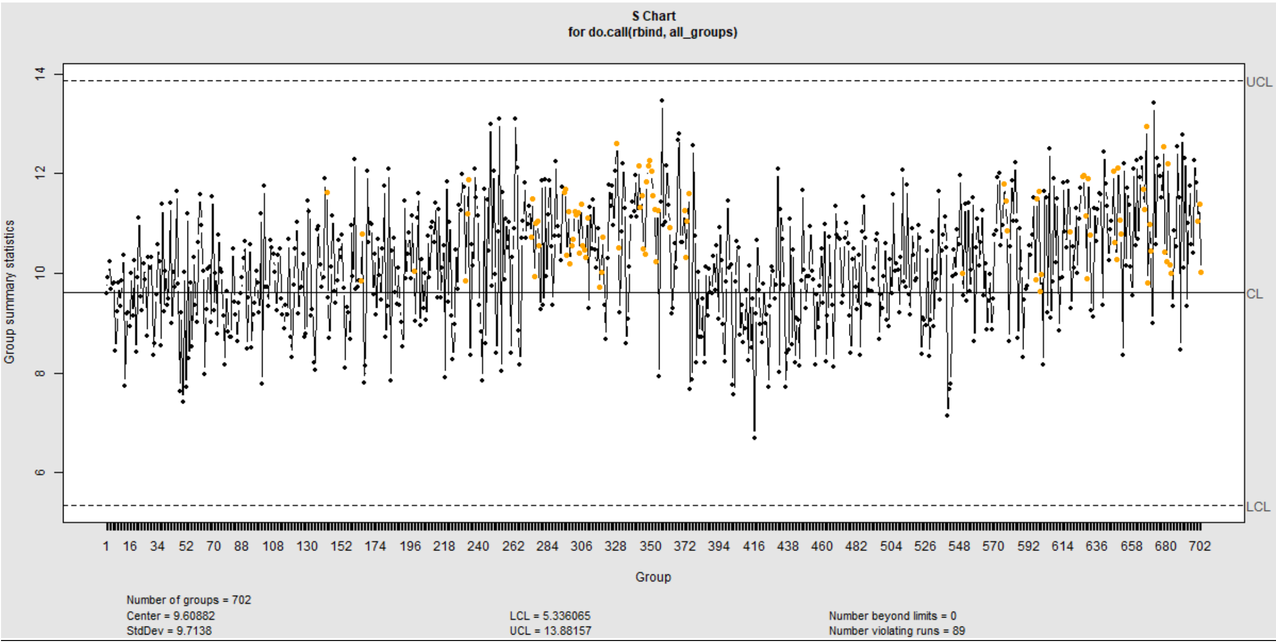
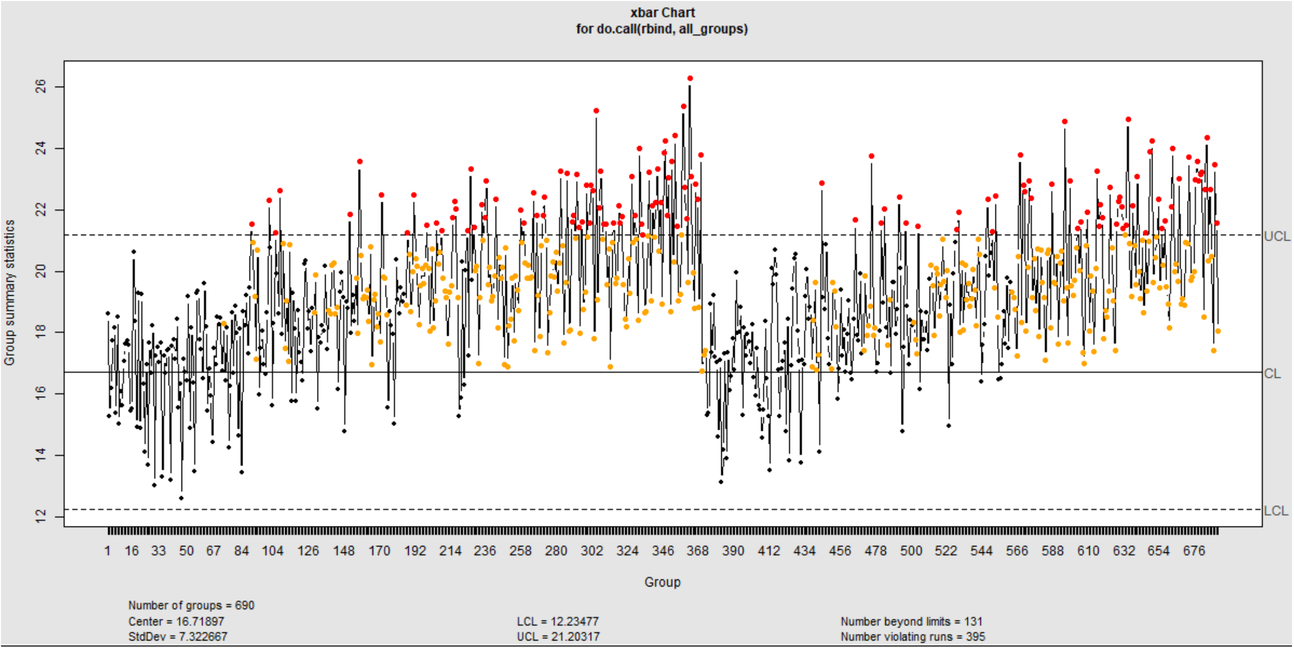


Figure 18: Mouse Continued Control Charts



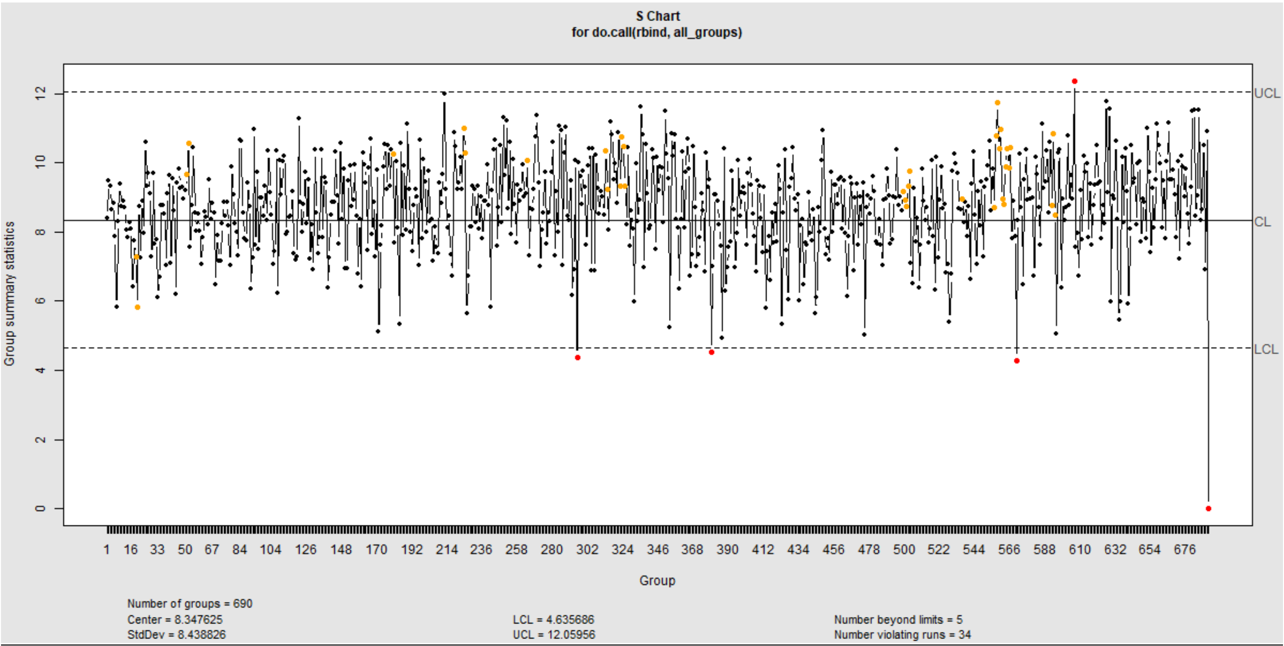
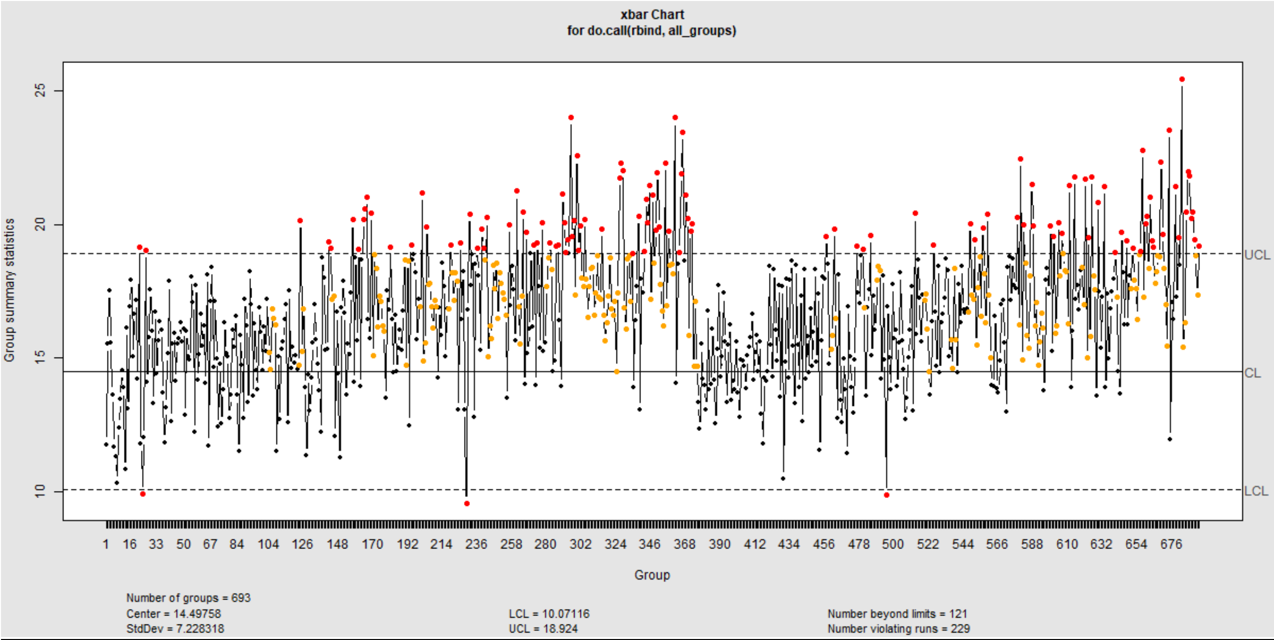


Figure 19: Laptop Continued Control Charts



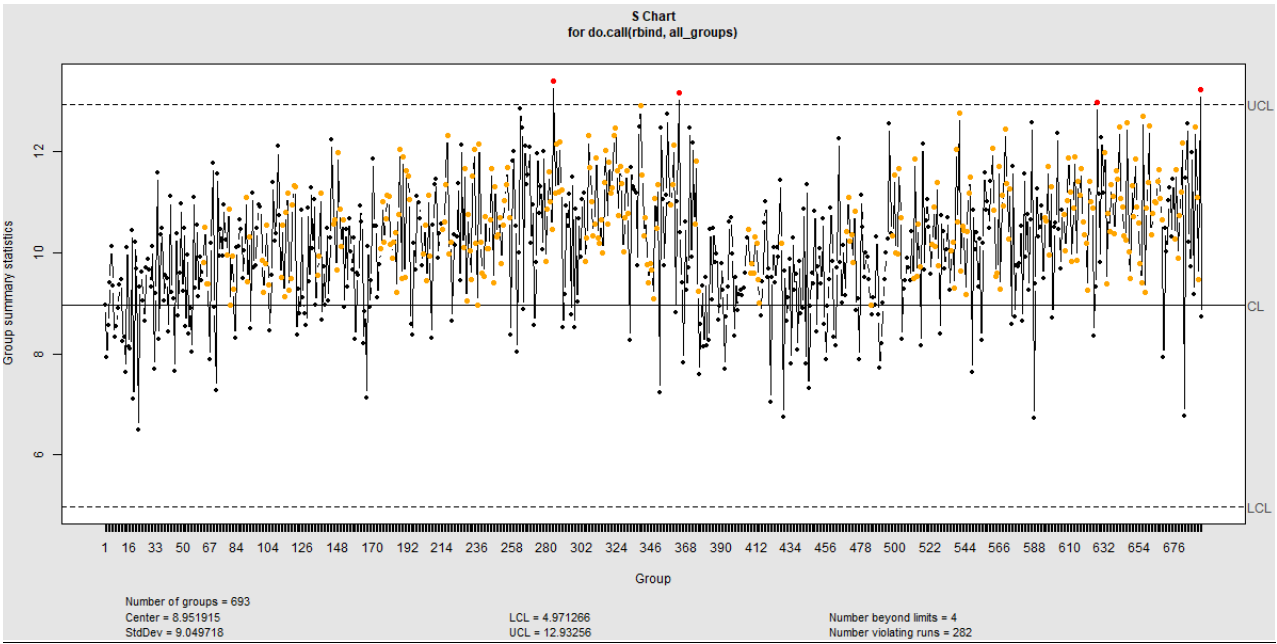
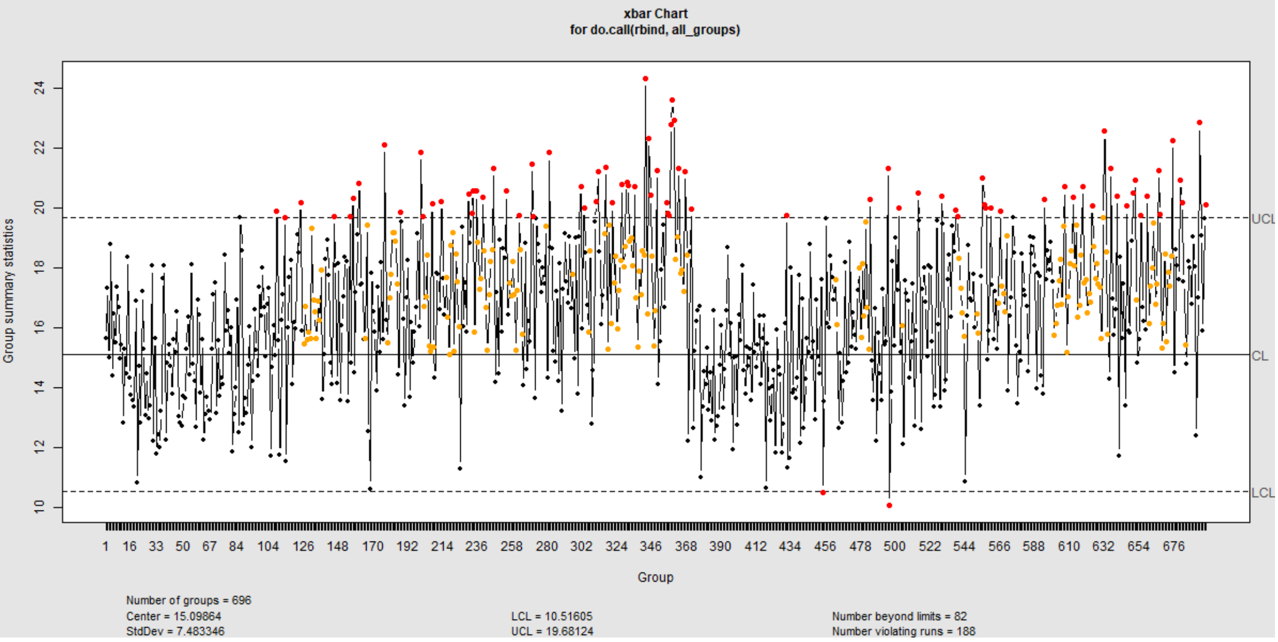
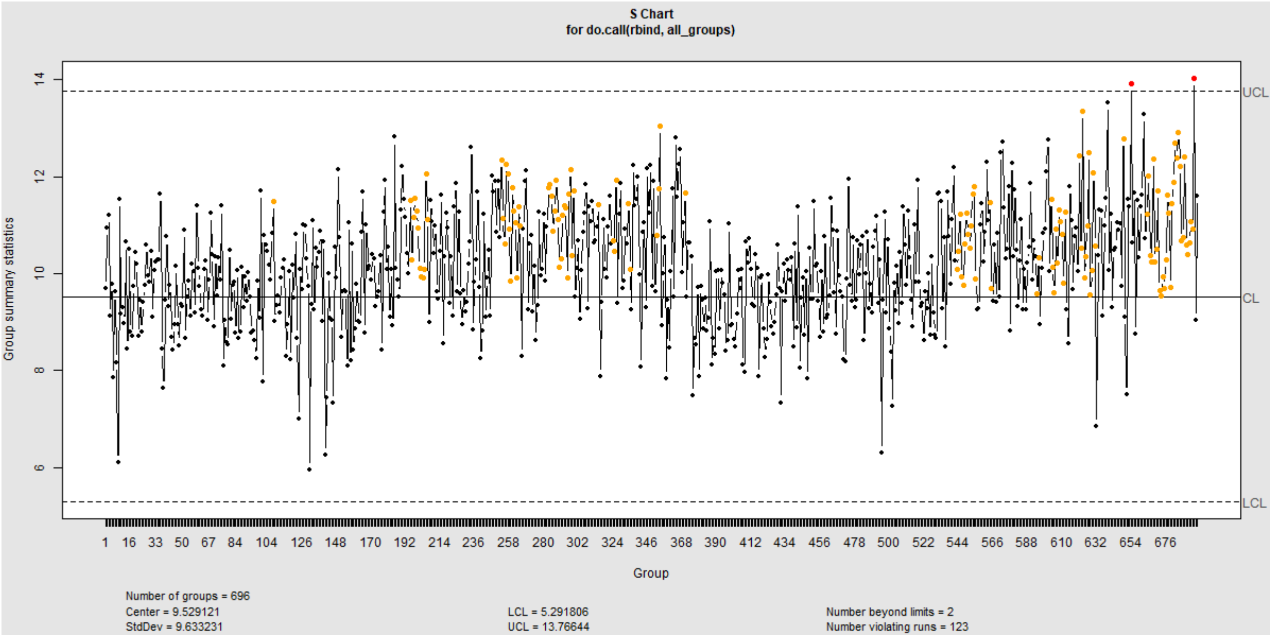


Figure 20: Cloud Subscription Control Chart





Process checking

Throughout the production process, trends and outliers were observed in both the sample averages and standard deviations. At certain points, significant outliers were clearly identified. Given the large number of outliers, particularly in the sample averages, the analysis focused on highlighting the overall trends as well as the start and end points of these outliers.

Keyboard

The standard deviation of delivery hours for the keyboard samples mostly remained within the three-sigma control limits, with only a few outliers. However, many sample standard deviations did exceed the one-sigma control limits. Like the other products, keyboards exhibited an upward trend in delivery hours. The first series of outliers occurred approximately between samples 126 and 370, after which the process readjusted within the three-sigma limits. A second set of outliers then appeared roughly between samples 450 and 670. Below, a more detailed breakdown with precise values is provided in table format to support the analysis described above:

Start of first outliers

Sample	X_bar	S	Reason
122	11.7352583	11.2300266	X_bar < LCL
130	21.8766333	7.44842208	X_bar > UCL
136	21.6266333	7.03564961	X_bar > UCL

Readjustment at 374 to 424

374	22.3592667	8.10565401	X_bar > UCL
424	21.4848042	7.62113138	X_bar > UCL

Start of last outliers

424	21.4848042	7.62113138	X_bar > UCL
428	21.7543333	5.80089323	X_bar > UCL
459	22.3793333	6.25311517	X_bar > UCL

Software

The standard deviation of delivery hours for the software samples exhibited consistent outliers throughout the process. Although there was a brief period of readjustment within the control limits, many sample standard deviations exceeded the one-sigma control limits. Like the other products, software showed an upward trend in delivery hours. The first series of outliers occurred approximately between samples 50 and 710, after which the process briefly readjusted. A second set of outliers then appeared roughly between samples 800 and 1340. Below, a more detailed breakdown with precise values is provided in table format to support the analysis described above:

Start of first outliers

Sample	X_bar	S	Reason
47	3.182425	6.69064856	X_bar < LCL
115	12.5210333	10.5696887	X_bar > UCL
148	13.18075	11.5287795	X_bar > UCL

Readjustment at 824 to 852

824	3.02234583	5.82711647	X_bar < LCL
852	12.0569583	10.609425	X_bar > UCL

Start of last outliers

862	11.8652917	9.88374132	X_bar > UCL
863	13.2090417	11.0303722	X_bar > UCL
866	12.0540958	11.4584437	X_bar > UCL

Monitor

The standard deviation of delivery hours for the monitor samples exhibited fewer outliers compared to keyboards and software. Most sample standard deviations remained within the three-sigma control limits, with occasional exceedances of the one-sigma limits. Similar to the other products, monitors displayed an upward trend in delivery hours. The first series of outliers occurred approximately between samples 40 and 370, followed by a period of readjustment within the control limits. A second set of outliers then appeared roughly between samples 480 and 700. Below, a more detailed breakdown with precise values is provided in table format to support the analysis described above:

Start of first outliers

Sample	X_bar	S	Reason
41	9.96650833	9.47820966	X_bar < LCL
64	19.3967667	7.96717532	X_bar > UCL
105	19.576275	9.33134829	X_bar > UCL

Readjustment at 376 to 423

376	21.0027667	7.85118364	X_bar > UCL
423	9.2863875	8.99972461	X_bar < LCL

Start of last outliers

484	19.5353292	8.36085027	X_bar > UCL
508	20.1491333	10.0580659	X_bar > UCL
523	20.0713583	8.37428983	X_bar > UCL
524	20.6609417	9.10883455	X_bar > UCL

Mouse

The standard deviation of delivery hours for the mouse samples mostly remained within the three-sigma control limits, with only a few outliers. However, many sample standard deviations did exceed the one-sigma limits. Like keyboards and other products, mice exhibited an upward trend in delivery hours. The first series of outliers occurred approximately between samples 150 and 350, after which the process readjusted within the three-sigma limits. A second set of outliers then appeared roughly between samples 470 and 670. Below, a more detailed breakdown with precise values is provided in table format to support the analysis described above:

Start of first outliers

Sample	X_bar	S	Reason
90	21.5384333	8.73659727	X_bar > UCL
101	22.321425	8.32916635	X_bar > UCL
105	21.252675	7.06992092	X_bar > UCL

Readjustment at 369 to 444

369	23.80335	8.71016637	X_bar > UCL
444	22.8793333	5.65429169	X_bar > UCL

Start of last outliers

481	21.5756917	9.16652082	X_bar > UCL
483	22.0285542	7.6650951	X_bar > UCL
492	22.4368875	7.77980232	X_bar > UCL
496	21.5611083	8.77995162	X_bar > UCL

Laptop

The standard deviation of delivery hours for the laptop samples remained mostly within the three-sigma control limits, with only a few outliers observed. Some of these outliers, however, were notably large. Like the other products, laptops displayed an upward trend in delivery hours. The first series of outliers occurred approximately between samples 126 and 370, after which the process readjusted within the control limits. A second set of outliers then appeared roughly between samples 450 and 670. Below, a more detailed breakdown with precise values is provided in table format to support the analysis described above:

Start of first outliers

Sample	X_bar	S	Reason
22	19.1599667	6.49853577	X_bar > UCL
24	9.8981	9.61304658	X_bar < LCL
26	19.038775	8.64428433	X_bar > UCL
124	20.1585667	8.79216339	X_bar > UCL
142	19.37315	9.04190984	X_bar > UCL
143	19.09675	10.2865416	X_bar > UCL
157	20.176275	9.65935105	X_bar > UCL

Readjustment at 372 to 457

372	20.0432167	12.1657808	X_bar > UCL
457	19.5488708	8.94927432	X_bar > UCL

Start of last outliers

462	19.840275	8.16422755	X_bar > UCL
476	19.1905375	9.0789568	X_bar > UCL
480	19.0801208	9.51802468	X_bar > UCL
485	19.5874125	8.77872309	X_bar > UCL

Cloud Subscription

The standard deviation of delivery hours for the cloud subscription samples largely remained within the upper and lower control limits, with only a few outliers observed. Most of the sample standard deviations were stable, though there were brief exceedances of the one-sigma limits. Like the other products, cloud subscriptions exhibited an upward trend in delivery hours. The first series of outliers occurred approximately between samples 105 and 370, followed by a long period of readjustment within the control limits. A second series of outliers then appeared roughly between samples 500 and 670. Below, a more detailed breakdown with precise values is provided in table format to support the analysis described above:

Start of first outliers

Sample	X_bar	S	Reason
109	19.9030167	9.00535672	X_bar > UCL
114	19.6870333	10.1166291	X_bar > UCL
124	20.1735083	7.00464929	X_bar > UCL

Readjustment at 371 to 431

371	19.9701167	9.52259363	X_bar > UCL
431	19.740275	7.33702609	X_bar > UCL

Start of last outliers

484	20.290275	8.61667028	X_bar > UCL
495	21.3223042	6.29244884	X_bar > UCL
496	10.0582625	9.44233123	X_bar < LCL
502	19.9798583	7.2534327	X_bar > UCL
514	20.5072042	9.44482841	X_bar > UCL

Process Capability

The process capability was evaluated for each of the six product categories using Cp, Cpu, Cpl and Cpk indices. Since delivery cannot occur before an order is placed, the Cpl is of less relevance and greater emphasis is placed on the Cpu and Cpk values.

Category	Cp	Cpu	Cpl	Cpk
Keyboard	0.643043	0.615062	0.671024	0.615062
Software	0.563353	0.855422	0.271284	0.271284
Laptop	0.573753	0.62327	0.524236	0.524236
Monitor	0.552083	0.596243	0.507922	0.507922
Mouse	0.638896	0.610974	0.666818	0.610974
Cloud Subscription	0.551877	0.593041	0.510714	0.510714

All physical products including, Keyboard, Laptop, Monitor, Mouse and Cloud Subscription, show Cp and Cpk values below 1, indicating that their delivery processes are barely capable of meeting the specified limits of 0 to 32 hours. Among them, the Keyboard (Cpk = 0.615) and Mouse (Cpk = 0.611) categories perform slightly better, while Monitor (Cpk = 0.508) and Cloud Subscription (Cpk = 0.511) show the weakest capability. The Cpu values for all physical products remain below their respective Cpl values, suggesting that late deliveries, are the primary issue affecting process performance.

The Software category stands out with a relatively higher Cpu (0.855) but a low Cpk (0.271). This reflects strong performance on the upper specification side but significant variation overall. Given the one-sided nature of delivery times, Cpu serves as a more meaningful indicator for software, showing that digital product delivery is considerably more capable of meeting the required limits than the physical product categories.

Process control issues

A: Standard deviation outside 3 sigma control lines

Count	Category	Sample Position	Reason
1	Keyboard	214	S above UCL3
		260	S below -UCL3
		290	S above UCL3
		346	S above UCL3
		370	S above UCL3
		660	S above UCL3

Count	Category	Sample Position	Reason
2	Software	430	S above UCL3
		621	S below -UCL3
		665	S above UCL3
		1286	S above UCL3
		1300	S above UCL3
		1340	S above UCL3

Count	Category	Sample Position	Reason
4	Mouse	280	S below -UCL3
		385	S below -UCL3
		566	S below -UCL3
		610	S above UCL3
		676	S below -UCL3

Count	Category	Sample Position	Reason
5	Laptop	280	S above UCL3
		368	S above UCL3
		690	S above UCL3
Count	**Category**	**Sample Position**	**Reason**
6	Cloud Subscription	654	S above UCL3
		690	S above UCL3

B: Most consecutive samples

Category	Longest_Run_Between_1sigma
Keyboard	13
Software	27
Laptop	25
Monitor	20
Mouse	12
Cloud Subscription	26

C: 4 consecutive X-bar samples

Category	Total_Runs
Keyboard	0
Software	0
Laptop	0
Monitor	0
Mouse	0
Cloud Subscription	0

Risk, Data correction and optimising for maximum profit

Type |

4.1)

In Section 3.4 above, samples that show process control issues were identified according to specific rules in Statistical Process Control. A Type I (Manufacturer's) Error occurs when a process that is in control is incorrectly flagged as out of control. The probability of committing a Type I error under each SPC rule was estimated and is summarized below:

Rule A: One sample beyond the +3σ limit

The probability of a Type I error for this rule is 0.00135. This means that if the process is stable, there is only a very small chance that a single point will exceed the +3σ limit purely by random variation. The +3σ rule is therefore conservative, minimizing false alarms while potentially being slower to detect small shifts in the process.

Rule B: One sample within ±1σ of the centreline

For this rule, approximately 68.27% of all samples are expected to fall within ±1σ of the centreline in a stable process. While this is not a Type I error probability, it illustrates the natural distribution of points in a normal process. Most points cluster around the mean, confirming that the control chart behaves as expected under normal operating conditions.

Rule C: Four consecutive samples beyond the +2σ limit

The probability of observing four consecutive points beyond +2σ by chance is extremely low, around 0.000027%. This indicates that such a pattern almost never occurs in a stable process and provides strong evidence of a genuine shift in the process mean or variation. This rule is highly sensitive and rarely produces false alarms.

Type ||

4.2)

For the bottle-filling process, the Type II (Consumer's) Error is the probability of not detecting a shift in the mean. With the process mean at 25.028 L, standard deviation 0.017, and control limits 25.011–25.089 L, the Type II error is about 0.841. This means there is an 84.1% chance that the shifted process would still appear in control, so the change would likely go unnoticed.

Data Correction

Upon reviewing the original datasets, several inconsistencies were identified in products_Headoffice_data and products_data. These included incorrect product IDs for items 11–60, misaligned selling prices and markups, as well as missing category codes. All identified issues were corrected, which significantly improved the consistency and reliability of the data.

The updated files were then merged with customer_data and sales2022and2023 to create the MasterData2025 dataset. The R code used for Section 1's basic data analysis was re-run on MasterData2025, revealing differences in the output. This highlights the importance of accurate data loading and validation to ensure meaningful and reliable analytical results.

Data information

Figure 21: MasterData2025

CustomerID	ProductID	Quantity	orderTime	orderDay	orderMonth	orderYear	pickingHou	deliveryHo	Category.x	Description	SellingPrice	Markup	Gender	Age	Income	City
CUST1791	CLO011	16	13	11	11	2022	17.72167	24.544	Cloud Subscrip	aliceblue silk	1070.54	16.41	Male	39	1.00E+05	Los Angeles
CUST3172	LAP026	17	17	14	7	2023	38.39083	31.546	Laptop	chocolate bright	18711.72	13.51	Female	58	90000	Chicago
CUST1022	KEY046	11	16	23	5	2022	14.72167	21.544	Keyboard	black sandpaper	708.18	17.72	Female	20	95000	Seattle
CUST3721	LAP024	31	12	18	7	2023	41.39083	24.546	Laptop	burlywood sandpaper	18366.92	29.35	Female	66	60000	Miami
CUST4605	CLO012	20	14	7	2	2022	15.72167	24.044	Cloud Subscrip	burlywood silk	963.14	10.13	Female	70	25000	Chicago

Product Information After Correction

Figure 22: Average Selling Price per Category

The bar chart illustrates the average selling price across different product categories. In the previous analysis, conducted before the correction of the product data, it was concluded that all product categories had an average selling price ranging between R2 900 and R4 500. However, after correcting the data, it is evident that this conclusion was far from accurate. The updated results show significant variation in average selling prices across categories. Keyboards, mice, and software products fall within a similar price range of approximately R500–R1 000, while cloud subscriptions are slightly higher at around R1 500. In contrast, monitors and laptops show a substantial increase, with average selling prices of roughly R6 000 and R18 000 respectively. In conclusion, the corrected data clearly demonstrates that the product category plays a major role in determining the average selling price.

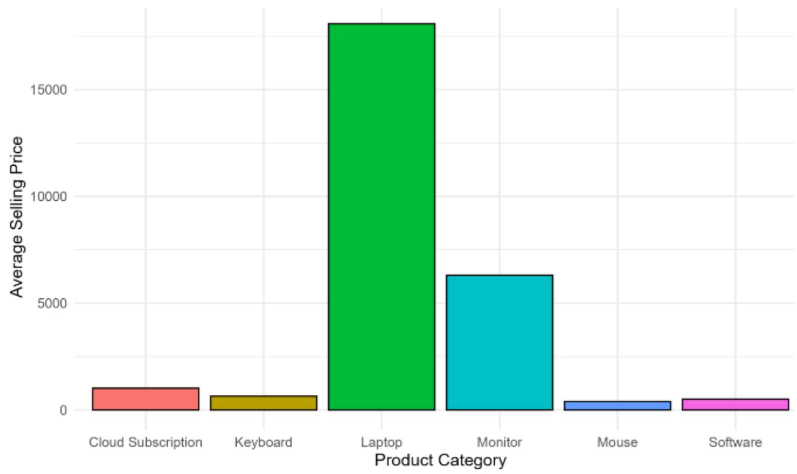


Figure 23: Selling Price Distribution per Category

In the previous analysis, we concluded that product categories generally have narrow price ranges, except for laptops and monitors. After the data correction, this conclusion remains valid but highlights that the mean selling prices differ significantly between categories. The chart shows that cloud subscriptions, keyboards, mouse and software have consistent price ranges, suggesting limited product or brand variety, while laptops and monitors display much wider ranges, indicating a broader mix of models and brands sold.

Figure 24: Profit per Category

After the data correction, it became clear that laptops and monitors have by far the highest selling prices among all product categories. This insight helps in interpreting the profit per category chart more accurately. The analysis shows that laptops generate the highest overall profit for the business. Although they have one of the lowest markup percentages, their significantly higher selling price results in a much larger profit per unit sold. Monitors, on the other hand, have a lower selling price than laptops but show the highest markup percentage among all categories. This explains why they yield the second-highest overall profit, their strong markup compensates for the lower selling price, making monitors another highly profitable product line for the business.

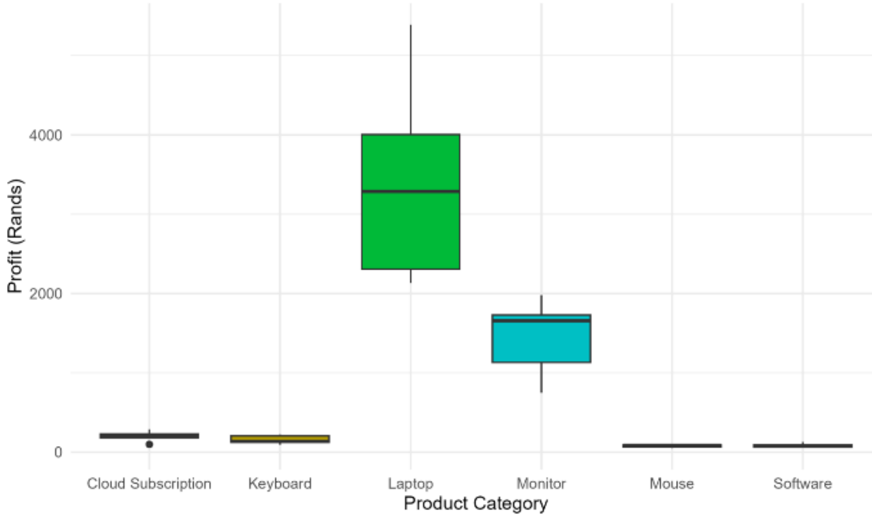
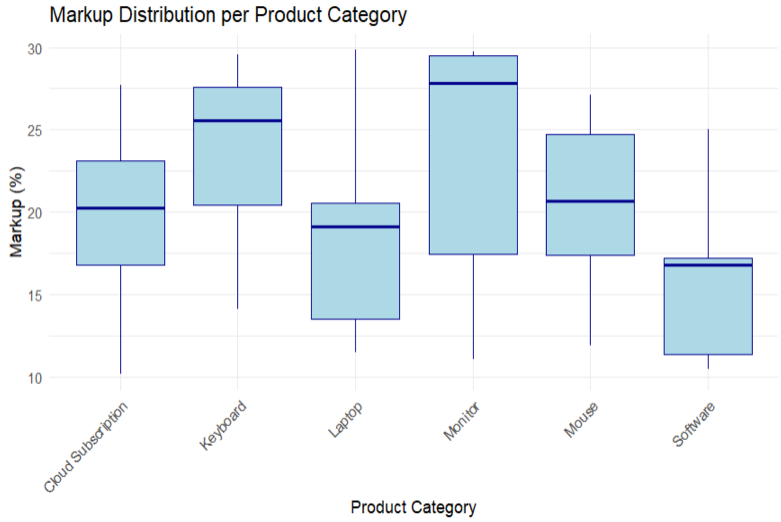


Figure 25: Distribution of Markup per Category

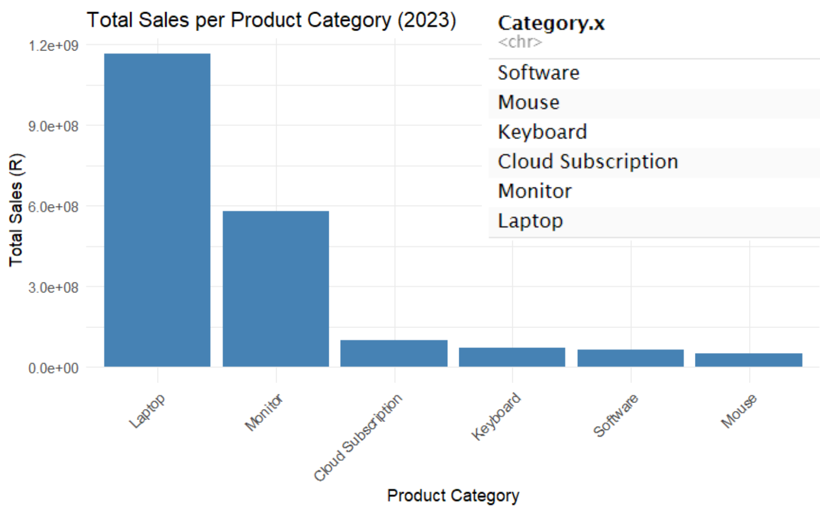
After the data correction, the markup distribution across product categories changed noticeably. While the previous results showed similar markups for all categories, the updated data reveals much greater variation. Each category now shows a wide range of markup values. Monitors have the highest average markup, followed by keyboards, whereas software has the lowest markup overall, including the lowest upper quartile.



Sales Value

Figure 26: Total sales per category

After correcting the data, we gained some valuable insights into our sales. To deepen our understanding, we calculated the total sales per product category. This analysis shows that laptops and monitors generate the highest sales, which aligns with our earlier observation that their selling prices are significantly higher than those of other categories. However, this does not necessarily mean that more laptops and monitors were sold compared to other products. In fact, the table below shows that fewer units of these categories were ordered, and it is their high selling price that drives their total sales figures upward.



Category.x <chr>	Total_Quantity <dbl>
Software	131349
Mouse	129613
Keyboard	114357
Cloud Subscription	96691
Monitor	91782
Laptop	64414

Coffee Shop Profit Optimization

Coffee Shop 1

To determine the optimal number of baristas for maximum profit, a model was developed in R using the *timeToServe* dataset from Coffee Shop 1. The data included the number of baristas (V1) and average service time per customer in seconds (V2). We assumed the shop operates for 9 straight hours a day without any breaks. Each customer generated R30 in profit, while each barista cost R1 000 per day. A "good service" was defined as serving a customer within 60 seconds. We calculated the daily number of customers served, overall profit and the probability of providing good services. These results were visualised through line, bar and box plots to identify the number of baristas for achieving optimal profitability.

Figure 27: Customer Waiting time by number of baristas

This boxplot illustrates the distribution of customer waiting times for different numbers of baristas. From the chart, we can see that waiting times decrease significantly as more baristas are employed. This suggests that the probability of providing good service, defined as serving a customer within 60 seconds, increases notably when four to six baristas are working, resulting in faster and more efficient service overall.

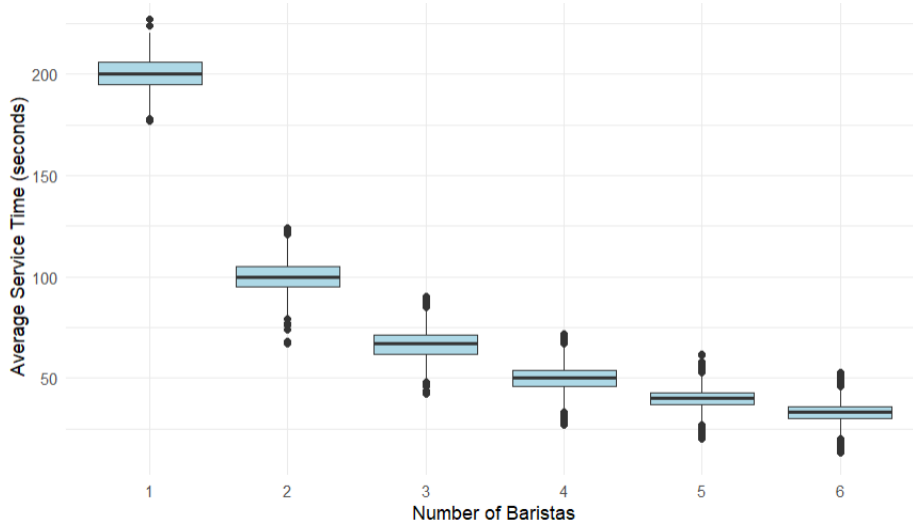


Figure 28: Average Customers Served per Day vs Number of Baristas

In this graph, the average number of customers served per day increases linearly with the number of baristas. This trend suggests that employing six baristas is likely to yield the highest overall profit for the business.

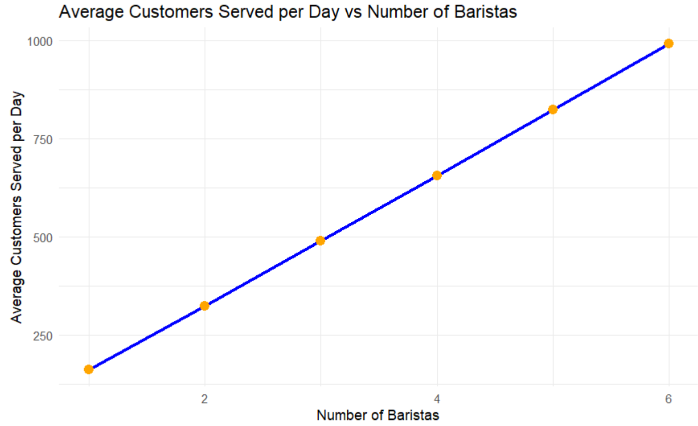


Figure 29: Average Profit per Day vs Number of Baristas

This graph again shows a clear linear relationship between the average profit and the number of baristas employed. It indicates that hiring six baristas would likely result in the highest overall profitability for the business.

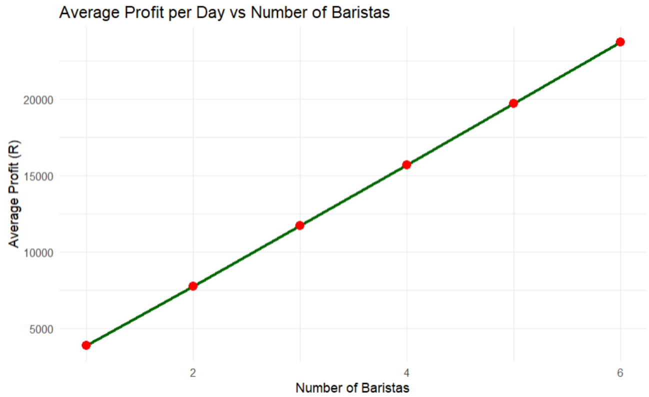


Figure 30: Summary of results table

To further reinforce the conclusions from the graphs, we created a summary table that clearly illustrates the optimal staffing level. Employing six baristas proves to be the most effective solution to maximize profitability. At this staffing level, the coffee shop achieves its highest profit of R23,722.24 while ensuring that all customers are served within 60 seconds. This corresponds to a 100% probability of good service, demonstrating that six baristas not only maximize profit but also maintain excellent customer satisfaction.

V1 <int>	avg_service_time <dbl>	avg_customers_served <dbl>	avg_profit <dbl>	prob_good_service <dbl>
1	200.15588	162.1348	3864.044	0.0000000
2	100.17098	325.1093	7753.279	0.0000000
3	66.61174	490.8315	11724.944	0.1646050
4	49.98038	656.5189	15695.568	0.9722914
5	39.96183	824.0670	19722.011	0.9999647
6	33.35565	990.7415	23722.244	1.0000000

Coffee Shop 2

To demonstrate the effectiveness of the R code and model we developed, we applied the same analysis to a different dataset, timeToServe2, which contains the number of baristas and service times for Coffee Shop 2. By processing this data, the code allowed us to determine the optimal number of baristas that should be employed at Coffee Shop 2 to achieve the highest possible profits.

Figure 31: Average Profit per Day vs Number of Baristas

The relationship between average profit and the number of baristas forms a concave-down curve, showing that profit increases at a decreasing rate as more baristas are employed. The graph indicates that employing five or six baristas yields nearly the same profit, with six baristas still achieving the highest profit. Beyond six baristas, the curve begins to flatten and could even decline, suggesting that adding more staff would not necessarily increase profit and may be inefficient. This implies that six baristas are the optimal staffing level, balancing maximum profitability with the highest probability of good service.

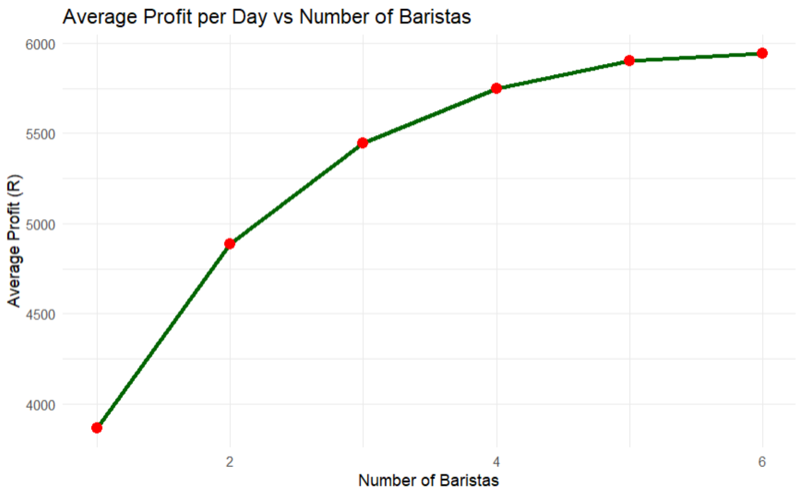


Figure 32: Summary of results

To further support the conclusions drawn from the graph above, we created a summary table showing the results for Coffee Shop 2. Employing five or six baristas yields very similar profits, with six baristas achieving an average profit of R 5 942 and a 100% probability of providing good service, ensuring high customer satisfaction.

V1 <int>	avg_service_time <dbl>	avg_customers_served <dbl>	avg_profit <dbl>	prob_good_service <dbl>
1	200.16894	162.1480	3864.440	0.000000000
2	141.51462	229.5438	4886.315	0.000000000
3	115.44091	281.4884	5444.652	0.007891542
4	100.01527	324.9772	5749.316	0.534472499
5	89.43597	363.4083	5902.249	0.986753521
6	81.64272	398.0959	5942.877	1.000000000

ANOVA

Figure 33: Summary of Anova Results

A one-way ANOVA test was performed for each product category to compare average delivery hours between the two years. The null hypothesis stated that delivery times were the same across both years, while the alternative suggested a difference. We reject the null hypothesis only if the p-value is below 0.10. As shown in the table, all product categories have p-values above this threshold, meaning we fail to reject the null hypothesis. This indicates that delivery hours remained consistent across the two years, with no significant change in performance

Category	Df	Sum_Sq	Mean_Sq	F_value	Pr_F	Residual_Df	Residual_Sum_Sq	Decision
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>
Keyboard	1	26.0126	26.0126	0.3266	0.5677	16670	1327679	Fail to reject H0
Cloud Subscription	1	253.9371	253.9371	2.3859	0.1225	16686	1775926	Fail to reject H0
Monitor	1	136.5797	136.5797	1.2777	0.2583	16829	1798917	Fail to reject H0
Mouse	1	1.0950	1.0950	0.0136	0.9071	16535	1329147	Fail to reject H0
Software	1	0.3908	0.3908	0.0034	0.9536	33196	3837956	Fail to reject H0
Laptop	1	7.4154	7.4154	0.0695	0.7921	16614	1772537	Fail to reject H0

Reliability of service

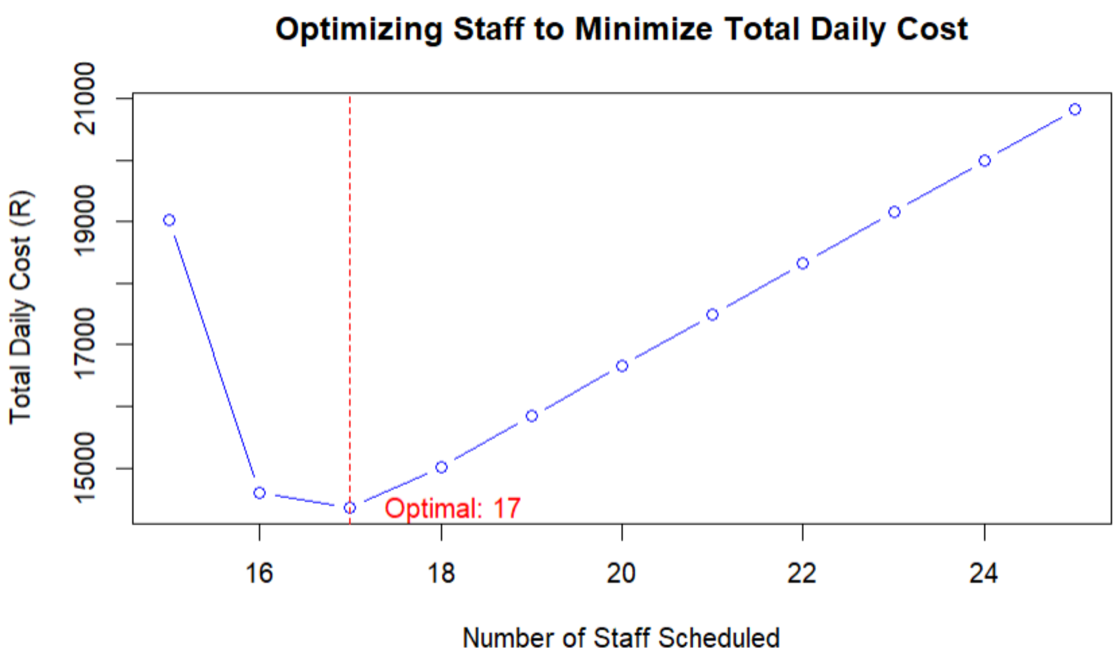
Expect reliable service

The probability that all employees arrive at work, meaning all 16 workers are present, was calculated using R and found to be 0.9740. This indicates that 97.40% of the time, all employees are at work. Consequently, the probability that a worker is absent on a given day is 0.0260. Multiplying this by 365 days gives the expected number of days a worker will be absent in a year, which is approximately 10 days.

Optimise Profit

Figure 34: Optimizing staff to minimize total daily cost.

To optimize staffing and minimize total costs, we modelled employee attendance as a binomial problem. A problem arises when fewer than 15 employees are on duty, leading to an average daily loss of R20 000 in sales. At the same time, hiring additional personnel increases costs linearly due to extra salaries. By analysing the total expected daily cost, we can identify the staffing level that best balances these factors. As shown in Figure 34, the total cost curve indicates that employing 17 workers achieves this balance, minimizing the likelihood of lost sales while keeping employment costs under control. This represents the optimal staffing level for maximizing the company's profitability.



Conclusion

This report presented a thorough analysis of customer, product and sales data, providing actionable insights into business operations. Initial exploration of customer and product datasets revealed key patterns, including age as a strong predictor of income and significant differences in product pricing and profitability across categories. Statistical process control highlighted trends, outliers and process capability issues, showing that only software deliveries consistently met specifications, while physical product deliveries often fell short. Risk assessment and data correction addressed inconsistencies in the datasets, ensuring reliable results and clearer interpretations of profit and markups. Optimization models for coffee shops identified six baristas as the ideal staffing level to maximize profit and maintain excellent service, while staffing analysis for a car rental agency indicated that 17 employees offered the best balance between sales and labour costs. ANOVA testing confirmed that delivery performance remained stable across years. Overall, the analysis combines process monitoring, risk mitigation and optimization strategies to provide a comprehensive view of performance, profitability and efficiency in business operations.

References

Schalkwyk, T. D. (2025). QA344 Statistics. Stellenbosch.