Quality Assurance Graduate Attribute Project

27118231

Contents

Introduction

This report addresses the 2025 ECSA Graduate Attribute 4 (GA4) requirements for the module QA344. It demonstrates competency in data analysis, statistical process control, process optimization, and experimental design through a series of industrial engineering tasks. The work encompasses descriptive analytics, control chart implementation, process capability analysis, hypothesis testing error estimation, and profit optimization modelling, culminating in a comprehensive analysis of service reliability.

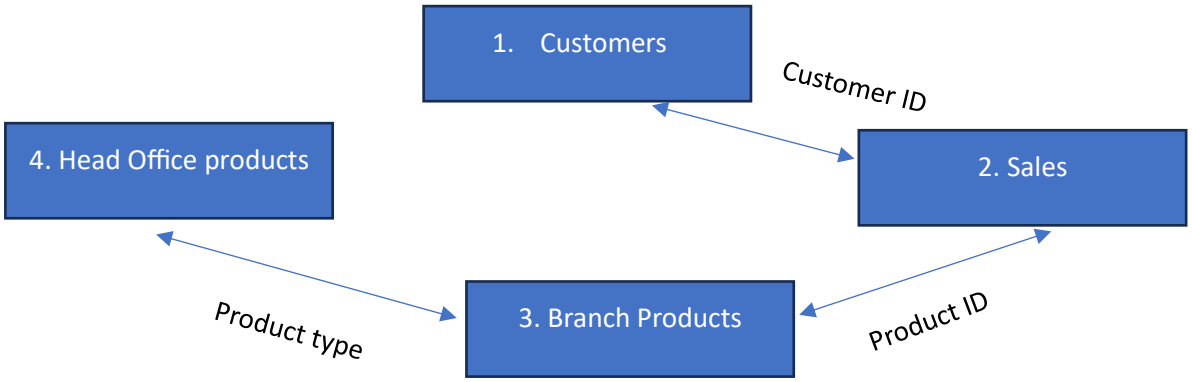
Part 1

As the new data analyst of a Tech company, I have been tasked to study the sales from the 2022-2023 financial period. The previous data analyst failed to leave behind any information regarding the data, thus analysis will need to be conducted from scratch. The purpose of this report is to familiarise myself with the data, assess its quality and uncover valuable insights that could help the company make informed business decisions.

The 4 datasets that will be analysed are the following:

1. Products Data
   This data set represents a list of local products sold at a specific branch. The price of each product is identified as well as the profit margins.
2. Customer Data
   This data set contains the profiles of 5000 customer and describes each customers' demographics, income levels and geographics distribution.
3. Sales for 2022-2023
   Each instance in the dataset represents a single sales transaction. For each transaction, the customer ID, product ID, time of order, quantity and operational data is provided.
4. Products Head office
   This data set acts as a central database of all possible products, not just what is currently available. This is a bigger catalogue than just the products data.

Below is a diagram showing the overlapping features between datasets. Identifying these features will make it easier to gain insights from multiple sources.



1.1 Descriptive Statistics

Below is a table showing the summary statistics of the descriptive features across all four datasets. Firstly, we notice that no missing values need to be dealt with. Customer data and Sales data have the same cardinality, further suggesting that no data is missing. We also notice that only seven cities are included in this analysis. The cardinality of the Gender variable is three. While this may initially appear anomalous, given the dataset's temporal scope (2022-2023), this likely reflects the inclusion of non-binary or alternative gender categories.

Table 1 Summary Statistic for all Discrete Features across all datasets

Dataset	Variable	Count	Missing	Missing %	Unique	Mode	Mode count
Customer Data	City	5000	0	0	7	San Francisco	780
Customer Data	CustomerID	5000	0	0	5000	CUST001	1
Customer Data	Gender	5000	0	0	3	Female	2432
Products Data	Category	60	0	0	6	Cloud Subscription	10
Products Data	Description	60	0	0	35	chocolate silk	5
Products Data	ProductID	60	0	0	1	CLO011	1
Products Headoffice	Category	360	0	0	6	Cloud Subscription	6
Products Headoffice	Description	360	0	0	110	black silk	21
Products Headoffice	ProductID	360	0	0	1	NA011	6
Sales Data	CustomerID	100000	0	0	5000	CUST1193	326
Sales Data	ProductID	100000	0	0	60	MOU057	2119

Table 2 Summary Statistics of all Continuous Features across all datasets

Dataset	Variable	count	missing	unique	mean	sd	min	median	max	IQR
Products Data	SellingPrice	60	0	60	4493.6	6503.8	350.5	794.2	19725.2	5904.5
Products Data	Markup	60	0	60	20.5	6.1	10.1	20.3	29.8	9.6
Customer Data	Age	5000	0	90	51.6	21.2	16	51	105	35
Customer Data	Income	5000	0	28	80797.0	33150.1	5000	85000	140000	50000
Sales 2022-2023	Quantity	100000	0	50	13.5	13.8	1	6	50	20
Sales 2022-2023	orderTime	100000	0	23	12.9	5.5	1	13	23	8
Sales 2022-2023	orderDay	100000	0	30	15.5	8.6	1	15	30	15
Sales 2022-2023	orderMonth	100000	0	12	6.4	3.3	1	6	12	5
Sales 2022-2023	orderYear	100000	0	2	2022.5	0.5	2022	2022	2023	1
Sales 2022-2023	pickingHours	100000	0	321	14.7	10.4	0.4	14.1	45.1	9.3
Sales 2022-2023	deliveryHours	100000	0	264	17.5	10.0	0.3	19.5	38.0	13.5
Products Headoffice	SellingPrice	360	0	359	4411.0	6463.8	290.5	797.2	22420.1	5347.4
Products Headoffice	Markup	360	0	331	20.4	5.7	10.1	20.6	30	9

The previous page shows a summary statistics table of the continuous features across all four data sets. Similarly, no missing values are reported. The average selling price is R4,493.60, with a high standard deviation (R6,503.80), indicating significant variation in prices. The average markup is 20.5%, which is reasonable for tech products, indicating a typical profit margin. The mean age is 51.6 years, which suggests that many customers are in the middle-aged range, possibly reflecting a demographic that can afford tech products. The minimum quantity in a single order placement is 1 unit, while the maximum is 50 units. This shows a mix of small and bulk purchases. The minimum and maximum hours for picking and delivery vary greatly, indicating significant variability in the process.

1.2 Visual Analysis

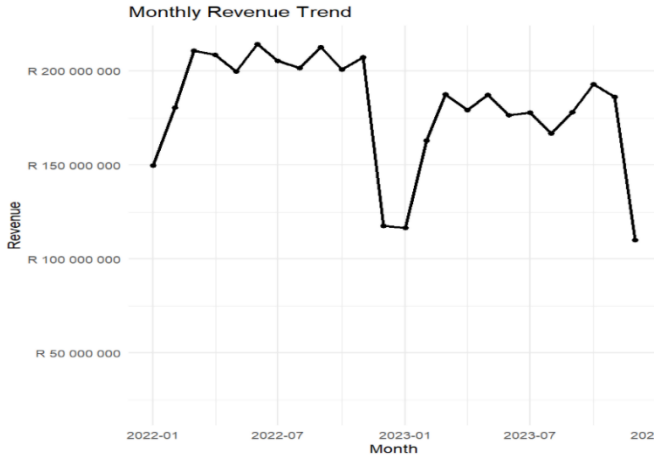


Figure 1 Monthly Revenue Trend 2022-2023

The monthly revenue trend shows clear seasonality. Major spikes in sales at the beginning of each year, as well as significant decreases at the end of each year can be seen. It is also clear that the overall revenue incurred during 2023 was noticeably lower than in 2022.

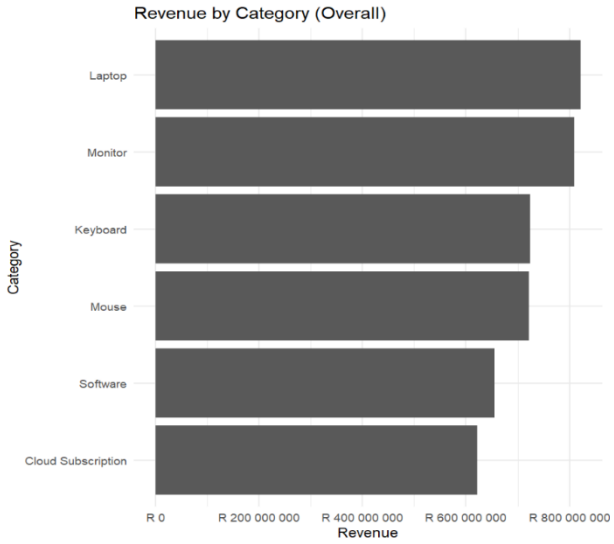


Figure 2 Total Revenue per Product Category over the 2022-2023 period

The bar plot below shows the Revenue per category of products sold. This does not reflect the number of sales, because each product has a different selling price. This plot instead shows which products are responsible for the most profit.

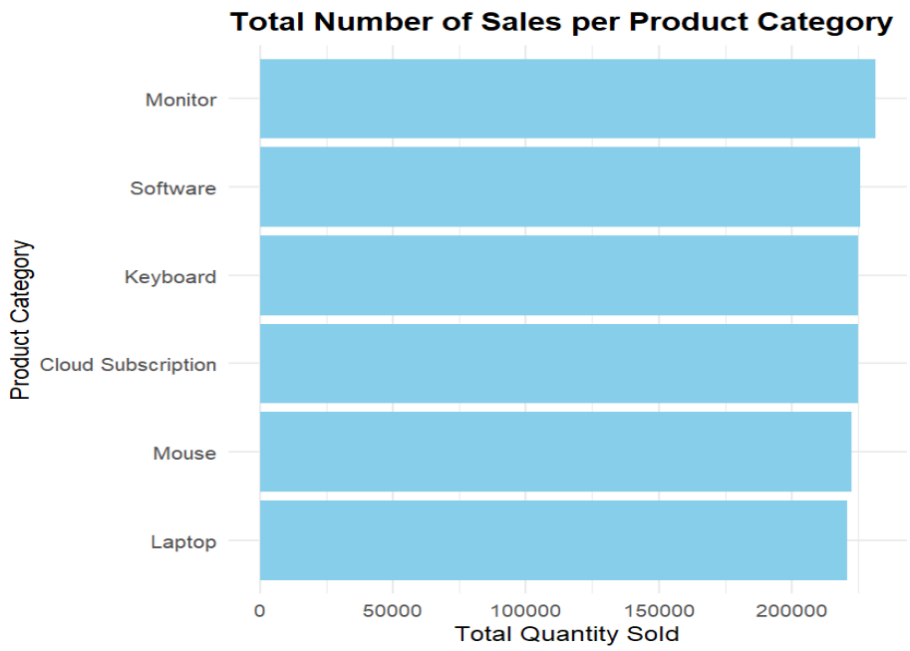


Figure 3 Total Number of Sales per Product Category

On the other hand, the bar plot above illustrates that there is no correlation between total number of sales and the total revenue incurred per category. Laptop shows least number of sales, yet these sales were responsible for the most revenue. In contrast to this, Monitors accounted for the highest number of sales whilst also being the second-best performer on the revenue scale.

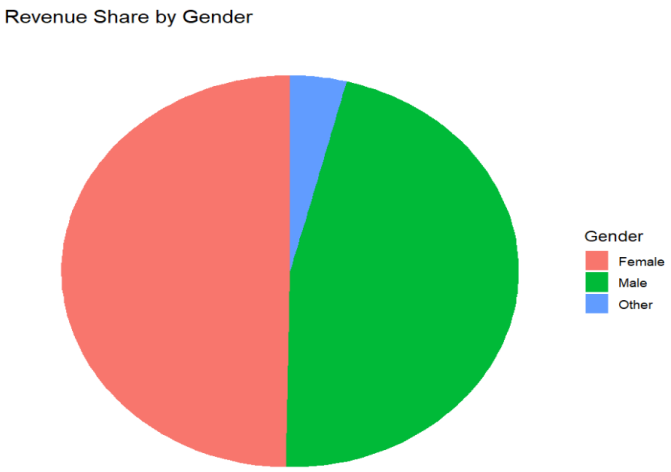


Figure 4 Histogram of Revenue Share by Gender

The histogram above shows the ratio of revenue incurred by customers based on gender. There is a relatively even portion of revenue incurred by male and female customers, however, females did incur more. A small, but noticeable portion of the revenue is incurred by customers who neither identify as male nor female.

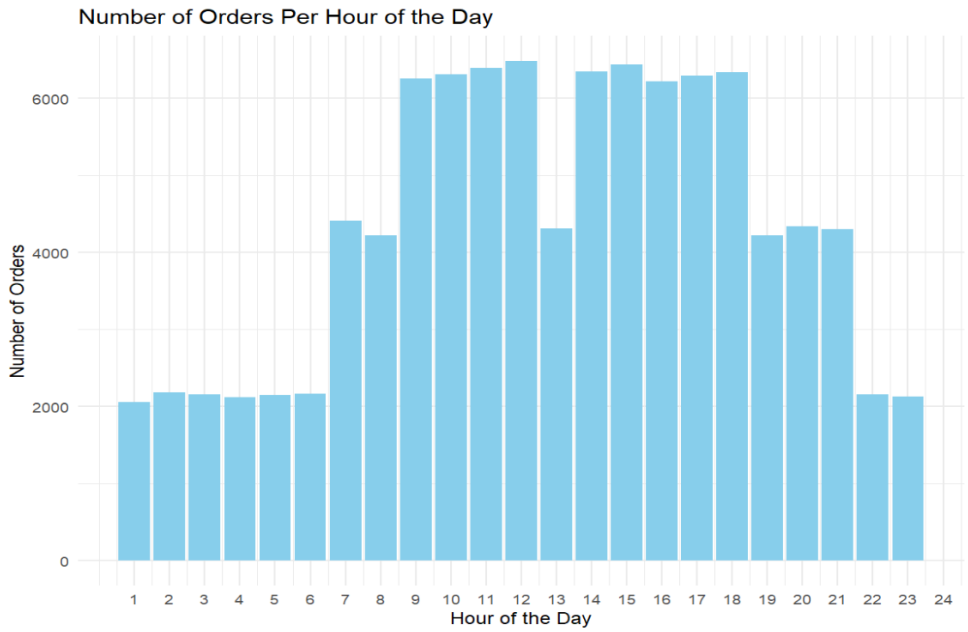


Figure 5 Bar Plot Showing Number of Order Per Hour of the Day

There are 3 distinct concentrations of order numbers throughout the day. The number of orders is either concentrated around 2000, 4000 or 6000 depending on the hour of the day. A clear bimodal distribution can be observed, with the busiest windows being between 09:00-12:00 and 14:00-18:00. The less busy windows where orders per hour are approximately 4000, typically occur around breakfast, lunch and dinner.

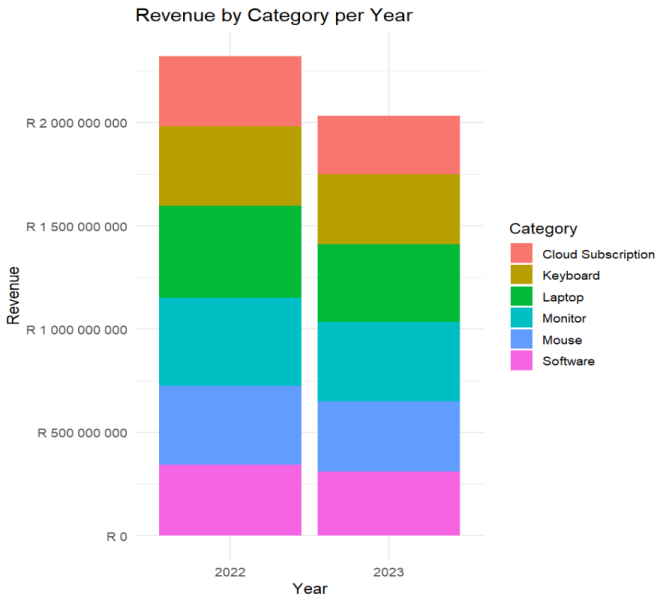


Figure 6 Bar Plot Showing Revenue by Category per Year

This graph further supports the observations made in Figure 2 regarding a decrease in overall revenue between 2022 and 2023. However, this graph highlights the proportion of revenue per category between the years. Most revenue proportions experienced little change. The cloud subscription and Software products, however, held for a smaller portion in 2023 in comparison to the previous years. This indicates that sales decreased for those two products.

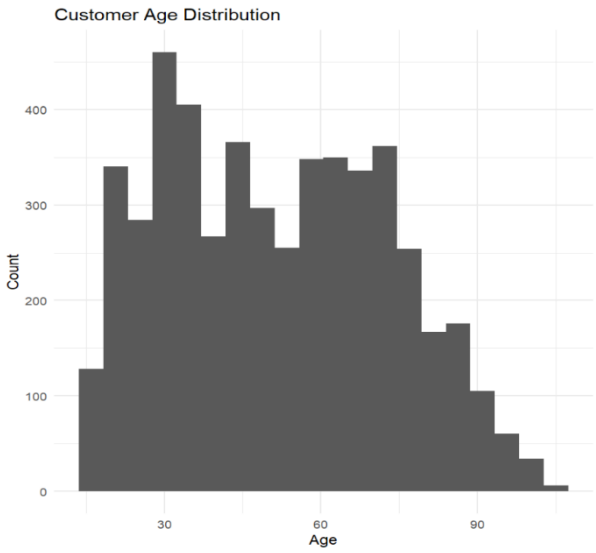


Figure 7 Bar Plot Showing Customer Age Distribution

The Customer Age Distribution chart shows no specific distribution. However, the count does taper off as age increases past 70 and the count is significantly low under the age of 10. This observation is logical as both young children and older adults are less likely to use or purchase technological products. Small dips around the ages of 40 and mid 50's can be seen. This highlights potential marketing opportunities directed towards these two target audiences, leading to a more linear decrease from the modal age (30) to the eldest customers.

Figure 8 Scatter Plot Showing Picking Hours vs Delivery Hours According to Sales Amount of Each Order

From the scatter plot comparing picking hours to delivery hours, two main clusters can be observed. The cluster on the left shows low picking hours and low delivery hours for majority lower value sales. The cluster on the right shows both high picking and high delivery hours for higher value products. Thus, there is a clear link between increasing picking hours and more valuable orders (either due to higher-priced products or larger order placements). This plot illustrates critical operational inefficiency. Longer picking and delivering results in higher revenue. The company may need to focus on decreasing overall time required for high value products or ensuring that longer processes are justified by the revenue generated.

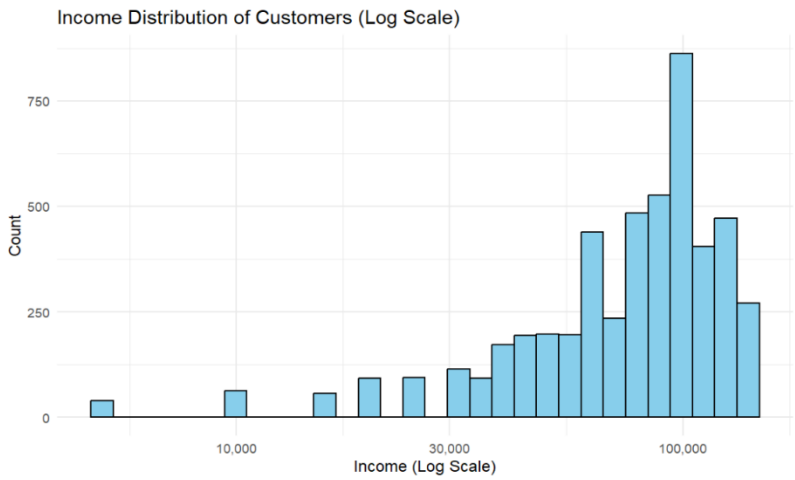


Figure 9 Bar Plot Showing Customer Income Distribution on a Log Scale for easier interpretation

The distribution of the graph above is left-skewed with a large portion of the customers clustered towards higher income brackets. This distribution validates earlier observations that high-value products are the primary sources of revenue. The mode is approximately 100 000 where the density of the customers is highest.

Recommendations for the Tech. Company based on the descriptive data analysis

1. Addressing Seasonality and Revenue Decrease:
   Since the company experiences significant seasonality in revenue it is crucial to better manage stock levels and marketing campaigns around peak sales periods. This could include preparing for year-end discounts or sales campaigns earlier in the year, especially in January.

2. Optimizing Product Offerings Based on Sales and Revenue:
   The laptop category generates the most revenue despite having low sales, while Monitors have high sales and are the second-best revenue-generating category. The company should focus on high-value, low-volume products by offering premium services such as extended warranties and premium support to increase the revenue per sale. For Monitors, consider bundling them with other products or offering discounts to maintain high sales volume while also improving margins.

3. Capitalizing on Peak Order Windows:
   The bimodal distribution of orders indicates significant peaks in orders between 09:00-12:00 and 14:00-18:00. Optimising staffing levels and order processing teams to match the peak demand periods is highly recommended. This can help to handle the influx of orders more efficiently, especially in those peak hours. The company should investigate implementing a dynamic staffing model that ensures the necessary resources are in place during peak hours.

4. Optimizing Operational Efficiency with Picking and Delivery Hours:
   The scatter plot revealed that longer picking and delivery hours correlate with higher-value products. The company should reduce operational inefficiencies by streamlining the picking and delivery processes, especially for high-value products. This can be achieved by investing in automation technologies.

Part 3

3. Statistical Process Control

3.1 Initialisation

This section establishes a process control system to monitor the delivery hours for future sales, using hypothetical data from 2026-2027. Since it is impossible to wait two years to collect all the data, we simulate a real-world monitoring scenario by analysing the data in chronological batches.

The setup process is as follows:

1. We group the hypothetical future data into chronological samples, with each sample containing 24 consecutive observations for each product type.
2. The first 30 samples are used as a baseline period to calculate the initial state of the process. From this baseline, we calculate:
    - The Centre Line (CL): The average performance of the process.
    - The Upper and Lower Control Limits (UCL/LCL): The expected range of natural process variation (±3σ).
    - The 1σ and 2σ Warning Limits: Inner boundaries that help identify potential trends or shifts.
3. We confirmed that all data points within this initial 30-sample baseline fall within the calculated control limits. This confirms that the baseline period represents a stable, in-control process, making it valid for use in monitoring future data.

3.2 SPC Charts

Category: Software

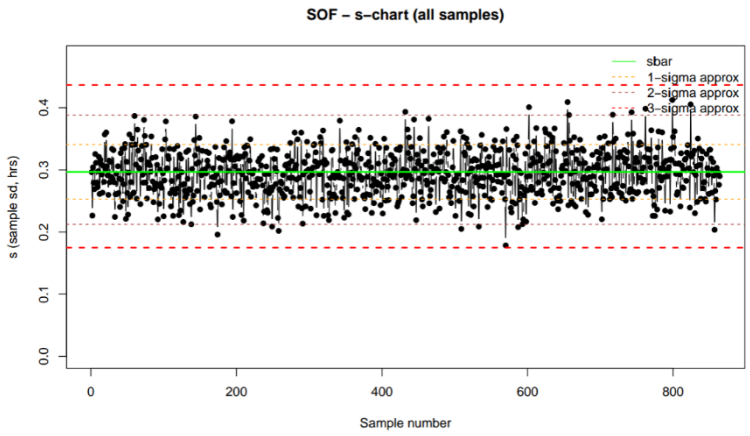


Figure 10 S-chart for Software

The data shows a centred mean around 0.956 with relatively few points outside the control limits.

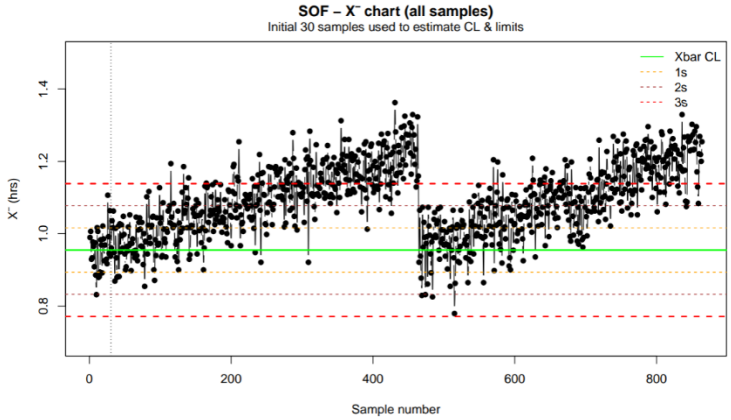


Figure 11 X-Bar Chart for Software

The standard deviation is highly variable with points often exceeding control limits, suggesting some variability in the process that needs to be addressed.

Category: Mouse

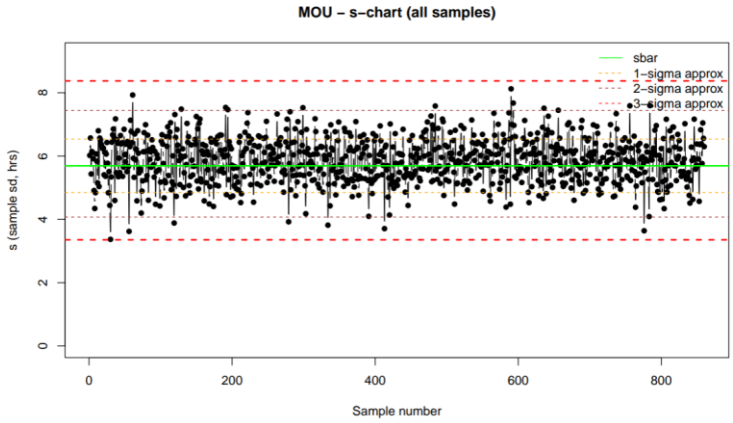


Figure 12 S-Chart for Mouse

With an average standard deviation of 5.53, the number of violations indicates some issues with consistency that need to be analysed further, especially if they affect customer satisfaction or product quality.

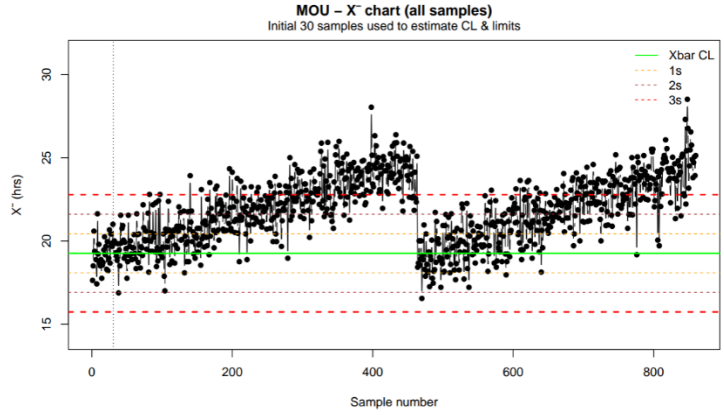
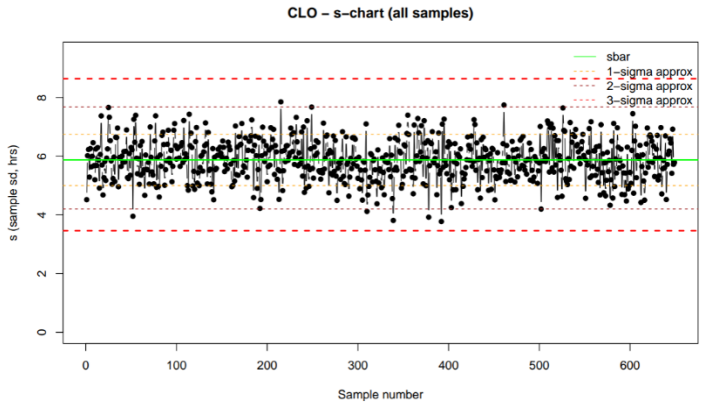


Figure 13 X-bar Chart for Mouse

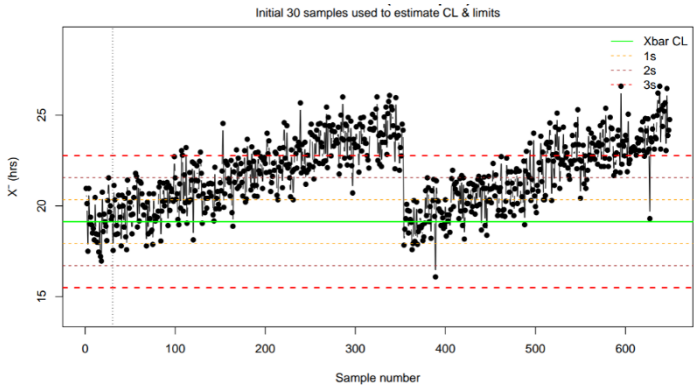
The mean process value is around 19.306 with 24 points beyond the control limits, which suggests significant variations and potential problems with the stability of the process.

Category: Cloud Subscription



The standard deviation is highly variable, suggesting process instability, and the company should evaluate potential root causes for these fluctuations. No violations of the 3sigma line.

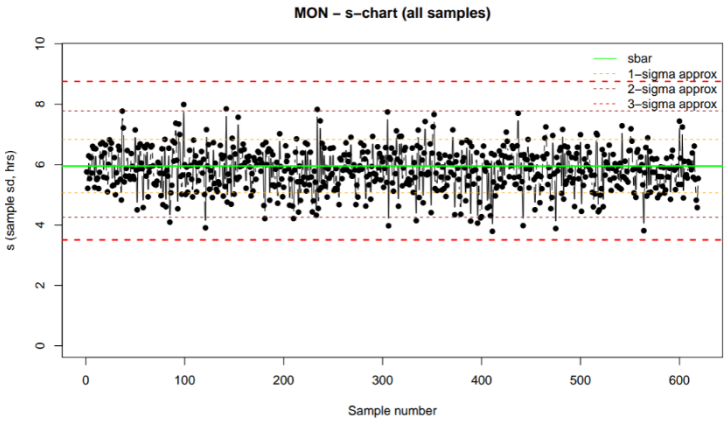
Figure 14 S-Chart for Cloud Subscription



The mean is around 19.206, but there are 24 instances beyond limits and 18 violating runs. This indicates some variation that likely impacts the overall process and needs attention to maintain consistency.

Figure 15 X-Bar Chart for Cloud Subscription

Category: Monitor



The standard deviation is also inconsistent. Variation needs to be addressed to improve process stability. No violations of the 3sigma line.

Figure 16 S-Chart for Monitor

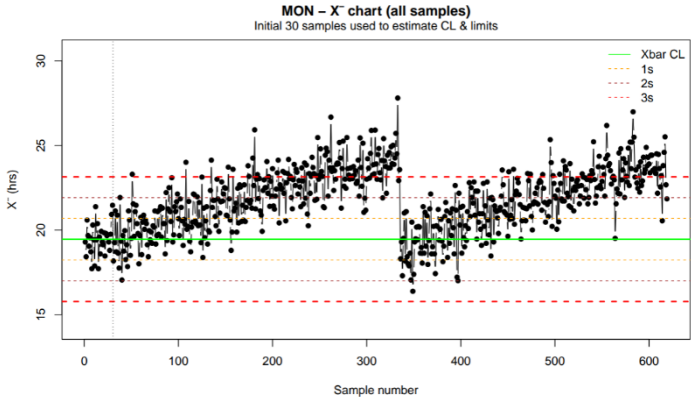


Figure 17 X-Bar Chart for Monitor

The centre is at 19.46, with 24 points outside the control limits. This indicates a high level of inconsistency that could be due to errors in product quality, production, or other operational factors.

Category: Laptop

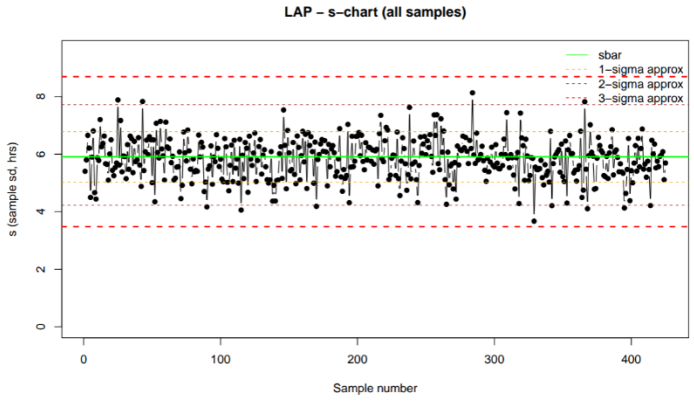


Figure 18 S-Chart for Laptop

With variable standard deviation and violations of the 2 sigma limit, there's a need to stabilize this process to improve predictability and control. No violations of the 3sigma line.

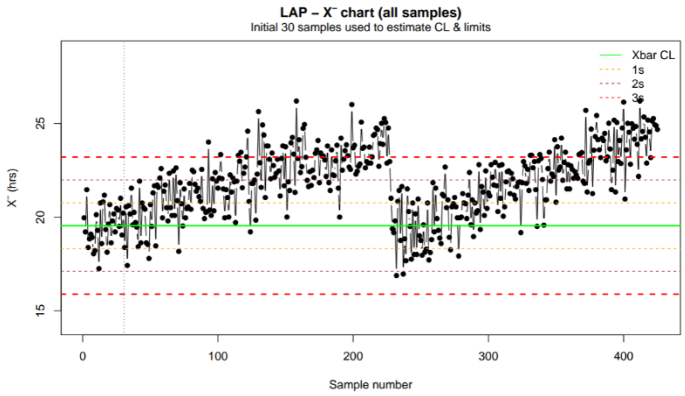
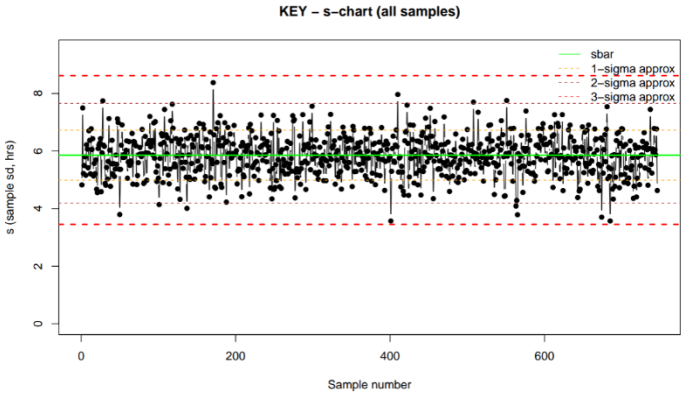


Figure 19 X-Bar Chart for Laptop

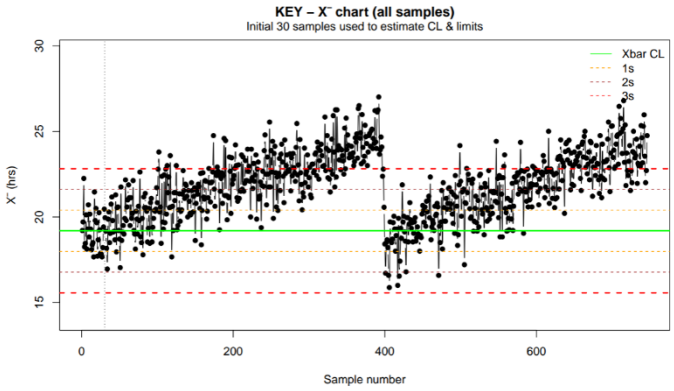
The mean is at 19.609, but there are 24 points beyond control limits and 20 violations. This highlights potential operational inefficiencies or process irregularities that need further investigation.

Category: Keyboard



S-chart: The standard deviation is quite variable, suggesting that the process needs to be re-evaluated to find the root causes of these fluctuations. No violations of the 3sigma line.

Figure 20 S-Chart for Keyboard



The mean is 19.268, with 24 points beyond the control limits and 19 violations. This shows that the process is highly unstable, and there may be significant factors contributing to this variation.

Figure 21 X-Bar Chart for Keyboard

All category processes are statistically in control. However, there are many points exceeding control limits, suggesting instability. The company needs to investigate the cause of variability to help stabilise operations.

3.3 Process Capability Indices

For process capability, the first 1000 instances for each product type were used. It is assumed that USL = 0 and LSL = 32hours. Cp typically measures the potential capability of a process and Cpu measures the actual capability of a process. A Cpk between 1.33-1.67 is considered acceptable and the process can be maintained. A Cpk below 1.33 is considered unacceptable and inspection as well as improvement is mandatory to improve the process. Below are the formulas used to calculate the capability indices.

$$Cp = \frac{USL - LSL}{6 \times \sigma}$$

$$Cpu = \frac{USL - \mu}{3 \times \sigma}$$

$$Cpl = \frac{\mu - LSL}{3 \times \sigma}$$

$$Cpk = \min(Cpu, Cpl)$$

Product type	Mean	SD	Cp	Cpu	Cpl	Cpk	Capable
SOF	0.956	0.294	18.118	35.153	1.083	1.083	FALSE
MOU	19.306	5.828	0.915	0.726	1.104	0.726	FALSE
MON	19.405	6.004	0.888	0.699	1.077	0.699	FALSE
LAP	19.609	5.927	0.900	0.697	1.103	0.697	FALSE
KEY	19.268	5.818	0.917	0.729	1.104	0.729	FALSE
CLO	19.206	5.928	0.900	0.719	1.080	0.719	FALSE

For all physical products the Cpk is approximately 0.7. This is well below the acceptable Cpk value of 1.33. Therefore, none of the products are capable of meeting VOC under the stated assumptions. Variability in physical product deliveries suggests room for process-improvement initiatives such as routing, scheduling, or staffing.

3.4 Identification of Process Control Issues

This section analyses the control charts against three specific rules to identify potential instability. The findings are summarized below.

Rule A: A Single Point Outside the 3σ Control Limits on the s-chart

This rule detects a single instance of unusually high process variability. No products violated this rule. When looking at the S-charts created earlier, we can clearly see there were no extreme spikes in process variability for any product type.

Rule B: Longest Consecutive Run within the ±1σ Limits on the s-chart

Rule Interpretation: This rule identifies periods of exceptionally stable and predictable process variability, which is a sign of good control. The table below shows the longest sequence of samples where the process variability remained within the ±1σ control limits for each product.

Product Type	Consecutive Samples	Start Sample	End Sample
SOF	16	3	18
MOU	14	249	262
CLO	39	462	500
MON	29	1	29
LAP	24	285	308
KEY	14	260	273

Product CLO demonstrated the most sustained period of stable variability, with 39 consecutive samples in control.

Rule C: Four Consecutive Points Outside the 2σ Limit on the X-bar Chart

This rule detects a sustained shift in the process average. Four points in a row outside the 2σ limit strongly suggest the process mean has changed. This rule was frequently violated across all product

types, indicating widespread issues with the process average shifting over time. The total number of violating sample sequences for each product is detailed below.

Product Type	Total Violations	First 3 Samples	Last 3 Samples
SOF	162	202, 203, 204	862, 863, 864
MOU	149	182, 183, 184	858, 859, 860
CLO	89	180, 181, 182	647, 648, 649
MON	104	190, 191, 192	616, 617, 618
LAP	74	119, 120, 121	423, 424, 425
KEY	143	171, 172, 173	744, 745, 746

The high number of violations for all products, especially SOF (162) and MOU (149), shows that the process average is unstable and frequently experiences sustained shifts away from the target.

Part 4: Risk, Data Correction and Optimisation for profit maximisation

---

4.1 Type I errors

In Statistical Process Control (SPC), rules are used to detect when a manufacturing or business process might become unstable. However, these rules can sometimes trigger a "false alarm." Statistically, this is known as a Type I Error. This error occurs when a control rule incorrectly signals that a process is unstable when it is actually stable. In SPC the null hypothesis (H_o) is always the assumption that the process is stable and in control. When a detection rule is applied, the data is tested against this hypothesis.

The probability of a Type I error is also called the Alpha (α) risk. A lower alpha means the rule is less likely to produce false alarms. Below, we calculate this probability for control rules A,B and C from Question 3.4.

Rule A: A Single Point Outside the 3σ Control Limits

- H_o for this rule: The process is stable, and any variation is due to common, random causes.
- Type I Error Analysis: This rule is designed to have a very low false alarm rate. For a perfectly stable, normally distributed process, the probability of a single data points randomly falling beyond the ±3σ limits is very small.
- Type I Error Probability (α): 0.002699796 ≈ 0.27%

Rule B: Eight Consecutive Points on the Same Side of the Centre Line (within ±1σ)

- H_o for this rule: The process is stable, and the sequence of points above and below the centre line is random.

- Type I Error Analysis: Even in a stable process, it's possible to randomly get a long sequence of points on one side. The probability of this happening is higher than for Rule A.
- Type I Error Probability (α): 0.04718302 ≈ 4.7%

Rule C: Four Consecutive Points Outside the ±2σ Limits

- H_o for this rule: The process is stable, and no sustained shift has occurred.
- Type I Error Analysis: This is a very specific pattern. The chance of a single point falling outside the ±2σ limits by random chance is about 4.55%. However, the probability of this happening four times in a row purely by chance is extremely low.
- Type I Error Probability (α): 0.0000002678772 ≈ 0.000027%

4.2 Type II errors

In SPC, a type II error occurs when a process is incorrectly labelled as stable, when it is unstable. The probability of a type II error is known as Beta (β) risk.

For this section, the probability of a type II error in a bottle filling process is determined. The process is centred around an average of 25.05 litres with a standard deviation of 0.013 litres. The UCL and LCL are assumed to be 25.089 and 25.028 litres respectively. The process has unknowingly shifted to an average fill volume of 25.028 litres and now has an X-bar standard deviation of 0.017. A type II error occurs when the mean falls between the UCL and LCL and subsequently, failure of the process is undetected. The type II error is calculated as follows.

Upper control limit (before standardisation)	UCL = 25.089
Lower control limit (before standardisation)	LCL = 25.011
Shifted mean	μ_1 = 25.028
Shifted standard deviation	$\sigma_{\bar{x}}$ = 0.017

$$\beta = \Phi\left(\frac{UCL - \mu 1}{\sigma_{\bar{x}}}\right) - \Phi\left(\frac{LCL - \mu 1}{\sigma_{\bar{x}}}\right)$$

$$= \Phi\left(\frac{25.089 - 25.028}{0.017}\right) - \Phi\left(\frac{25.011 - 25.028}{0.017}\right)$$

$$= \Phi(3.588) - \Phi(-1.000)$$

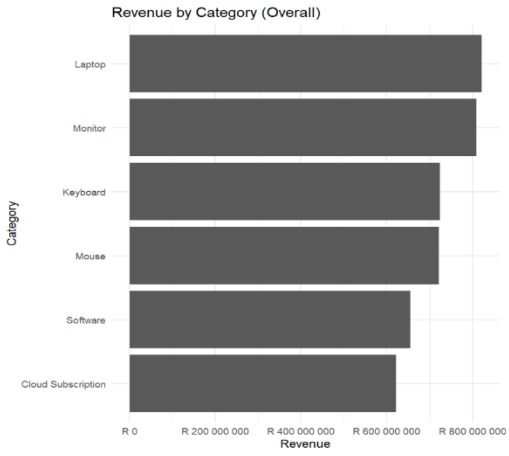
$$= 0.841$$

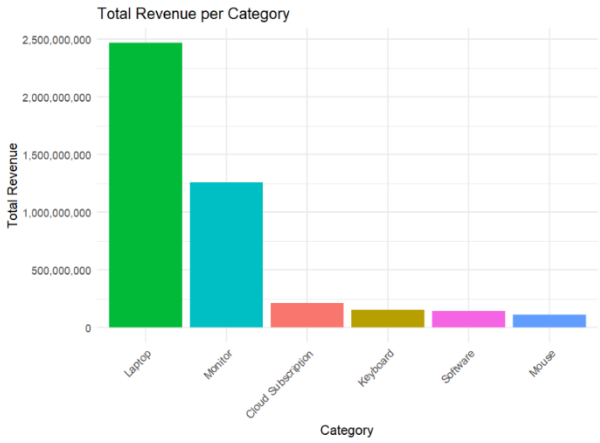
$$power\ to\ detect\ shift = 1 - \beta = 0.159$$

Although the process mean shifted downwards, the new distribution remains within the original UCL and LCL. Therefore, failure of the process is likely to go undetected. This probability is calculated to be 84.1%. We can conclude that the SPC chart is not very sensitive to small shifts. If the bottle filling station would like to improve error detection, the control limit width should be reduced. Alternatively, the sample size should be increased to lower standard deviation.

4.3 Data correction

Cardinal errors were noticed in the *Headoffice_products* and *products_data* in part 2 of this report. Many product ID prefixes were labeled as "NA". Additionally, the markup and selling price in the *Headoffice_products* dataset were not consistent with the values presented in the local *products_data set.* The errors were corrected, and data analysis was performed to compare with the orignal results in part 2 of this report.





The corrected results are alarmingly different. The original data showed a more even spread between revenue per category. The new bar plot, however, shows that most of the revenue is incurred by laptop sales, and monitor sales are responsible for approximately half of the laptop sales. All other products make up less than 10% of the laptop sales each. This analysis proves how misleading incorrect data can be.

5. Resource allocation for profit optimisation

For the following section, the service times per barista for a coffee shop is investigated to maximise profit. The dataset, *timetoServe.csv* consists of 2 columns. The first column is the list of baristas, and the second is the corresponding service time per customer in seconds. R30 of profit is made per customer served and baristas need to be paid R1000 per day.

First, the average service time per barista is calculated. Then, the number of customers per day (per barista) is calculated, assuming the customers are inversely proportional to the mean service time. Total time is equivalent to an 8-hour shift (i.e. 28 800 seconds).

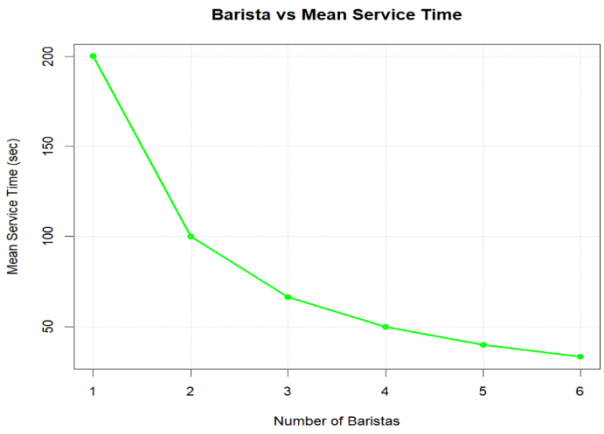
$$average\ service\ time_n = \frac{\sum service\ time\ of\ barista\ n}{nr\ of\ observations\ for\ barista\ n} \quad (1)$$

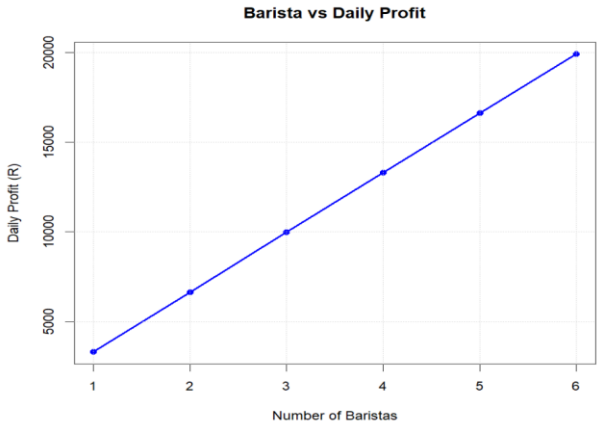
$$customers\ per\ day\ (per\ barista) = \frac{total\ time}{average\ service\ time_n} \quad (2)$$

Next, the objective function is modelled.

$$Daily\ Profit = 30 \cdot customers - 1000 \cdot baristas \quad (3)$$

Baristas	ServiceTime_sec	Customers	Profit
1	200.16	143.9	3317
2	100.17	287.5	6625
3	66.61	432.4	9971
4	49.98	576.2	13287
5	39.96	720.7	16621
6	33.36	863.4	19903

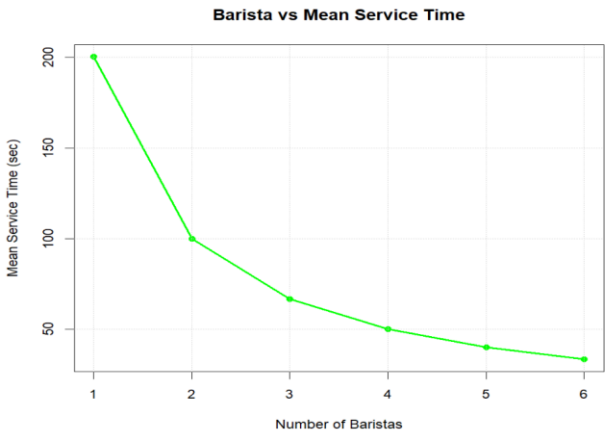


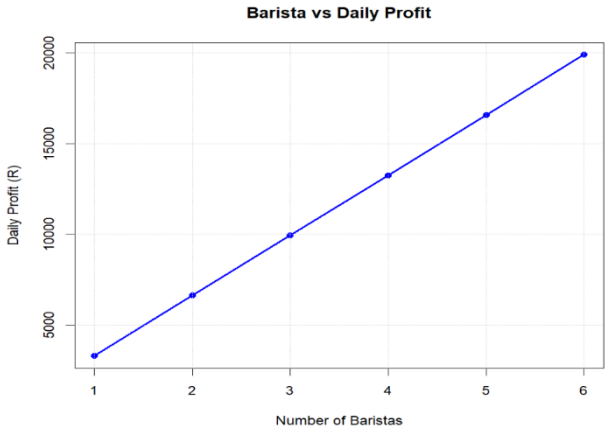


The profit increased linearly as additional baristas reduced service time, allowing a higher customer throughput. The profit reached a maximum at six baristas with a total of R19903 per day. If additional staff were hired, the profit would decline due to higher labour costs.

The same analysis was performed on the data for a second coffee shop, *timetoServe2.csv.* This coffee shop also has six baristas but the serving time differed. The exact same code used above was run to produce result for this coffee shop.

Baristas	ServiceTime_sec	Customers	Profit
1	200.43	143.7	3311
2	99.94	288.2	6645
3	66.7	431.8	9953
4	50.05	575.4	13262
5	40.01	719.8	16595
6	33.32	864.3	19928





In both coffee shops, the results are similar. Six baristas prove to be the optimal number, after which profits stagnate or decline. This indicates an optimal balance between labour costs and service throughput, which aligns with the principles of the Taguchi loss: as we deviate from the most efficient number of baristas, the system experiences a higher economic loss due to unnecessary labour costs.

Part 5

---

6. Design of an Experiment

The aim of Designing an Experiment (DOE) is to change one or more input factors at predefined levels and observe how they affect the outcome. The experimental data to be tested is the *sales2026and2027* used earlier in this report. In this context, the MANOVA is designed to test whether there is significant difference in multiple dependent variables (picking hours, delivery hours, and quantity) between the two years (2022 and 2023).

Independent variable: Order Year

Dependent variables: Picking hours, Delivery hours, Quantity

Null hypothesis: There is no significant difference in Quantity, Picking hours, and Delivery hours

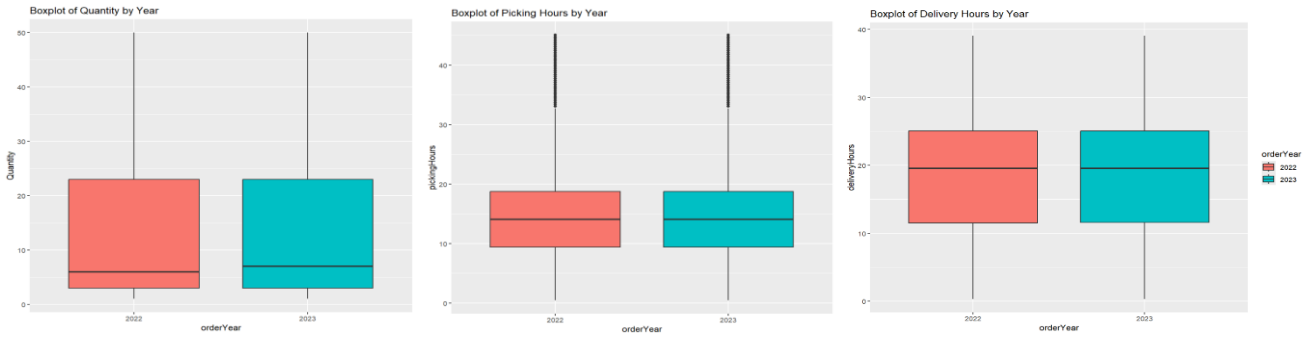
Alternative hypothesis: There is significant difference in at least one of the variables.

Through performing MANOVA, the output will yield Pillai's Trace and Approximate F statistics. These metrics tell you whether the difference between the years are large enough to be statistically

significant. It will also output a p-value, Pr(>F). This value similarly indicates whether the difference is statistically significant. A p-value less than 0.05 means the null hypothesis must be rejected.

Factor	Df	Pillai's Trace	Approx. F	Num Df	Den Df	Pr(>F)
Order Year	1	0.000057667	1.9223	3	99996	0.1235
Residuals	99998					

The calculated p value, Pr(>F) is equal to 0.1235. Because it is larger than 0.05, we do not reject the null hypothesis. Therefore, there is no significant difference in the dependent variables between the years. This point is further supported by the box plots showing the quantity, picking hours and delivery hours between the years. When comparing between the years, there is no noticeable change.



Fisher's Least Significant Difference (LSD) test is conducted after the DOE. It identifies which treatments are different from each other. This is done after significance through ANOVA is found. Because no significant difference was found, it is not necessary to perform Fischer's LSD test at this stage.

7. Reliability of Service and Profit Maximisation

7.1 Reliability of Service

The number of employees working at a car rental agency over the course of 397 days are analysed. This experiment is specifically looking at the number of days that 12 -16 employees are on duty to estimate how many days per year the agency can expect reliable service.

Number of workers present (X)	Number of days X workers were present
12	1
13	5
14	25
15	96
16	270

A reliable day is treated as a Bernoulli event where the probability is equal to number of days 15 or more workers are on duty. Therefore,

$$p(reliable\ day) = \frac{days \geq 15}{379} = 0.95$$

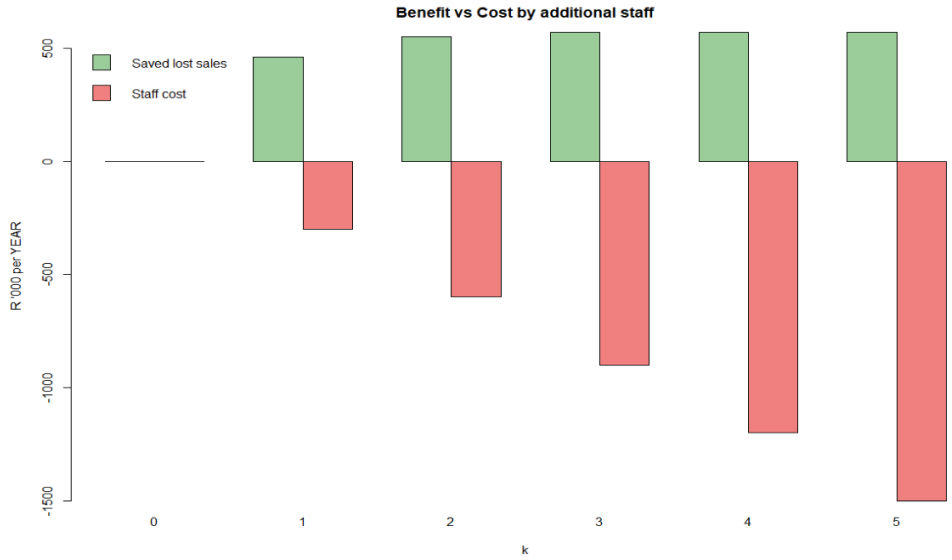
However, this probability still needs to be scaled because we are estimating the probability per year. Therefore, we calculate 95% of 365 days which equates to 346 full days.

7.2 Maximising Profit

The car rental agency is also looking to maximise profit. To solve this problem, the company needs to determine how many extra workers should be hired to decrease the number of unreliable service days without decreasing overall profit. An optimisation model was designed to solve this problem. The model assumes hiring k permanent staff increases every worker day count by k. It recomputes the number of unreliable days (less than 15 workers on duty) and computes the total lost revenue (R20 000 per unreliable day).

$$lost\ sales\ per\ year\ = 20\,000 \cdot \frac{nr\ of\ unreliable\ days\ with\ k}{365}$$

$$cost\ of\ additional\ staff\ per\ year\ = 25\,000 \cdot 12 \cdot k$$



In the chart above we can clearly see the comparison between the benefit and cost of hiring additional staff. If 1 additional worker is hired, about R500 000 due to lost sales can be saved. This significantly outweighs the cost of hiring 1 worker; therefore profit is realised. If 2 workers are hired, the money saved from lost sales no longer outweighs the cost of hiring. The same pattern is observed as k increases. Thus, maximum profit is obtained when only 1 additional worker is hired.

8. Conclusion

In conclusion, this report has successfully applied statistical and engineering principles to analyse processes, identify areas for improvement, and propose data-driven optimizations for profit and reliability. The work conducted across all parts fulfils the practical requirements of ECSA GA4, showcasing the ability to manipulate data, interpret statistical outcomes, and provide logical, well-reasoned solutions to complex industrial problems.

References

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Available at: https://www.R-project.org (Accessed: 25 October 2025).

Wickham, H., François, R., Henry, L. and Müller, K. (2021). dplyr: A Grammar of Data Manipulation. R package version 1.0.7. Available at: https://CRAN.R-project.org/package=dplyr (Accessed: 25 October 2025).

ggplot2 (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. Available at: https://ggplot2.tidyverse.org (Accessed: 25 October 2025).