# ECSA GA4 Report (QA344) — 2025

Student: Louis Schlebusch (26083329)
Date: 24 October 2025

# Table of Contents

## Contents



.

# Introduction

This report demonstrates the ECSA GA4 outcome in data manipulation and analysis through applying statistical thinking, programming, and decision analytics to an operations dataset spanning products, customers, and sales. We use R to clean and explore the data (Part 1.2), then plan and conduct Statistical Process Control (SPC) on delivery-time performance with $\bar{X}$–- and s–-charts, capability against customer specs, and rule-based signal detection (Part 3). We estimate risk by Type I/II error estimation, resolve master-data issues and rework Week-1 analysis with properly adjusted pricing (Part 4), and build an operations model to optimize profit and service reliability for two coffee shops (Part 5). Finally, we test hypotheses using ANOVA/MANOVA on selected factors and assess a reliability-of-service personnel issue through a binomial model (Parts 6–7).

The assignment is structured and referenced against scholarly guidelines and follows the prescribed format—Contents, Introduction, Body (Parts 1–6/7), Conclusions, and References—with code being given separately. Deliverables are properly labelled figures/tables, reproducible R scripts, and well-interpreted results that translate statistical results to service and reliability decisions.

## Part 1.2 — Descriptive statistics

**Data Loading and Inspection**

The data sets consist of product, customer, and sales information. Products Head Office data set consists of 360 rows, whereas Products Data file consists of 60 rows. Customer Data data set comprises 5,000 rows containing age, gender, income, and city information. Finally, the Sales Data file is largest at 100,000 rows and monitors transactions like customer ID, product ID, quantity, order date data, and logistics like picking and delivery times. Column headers are concise and consistent, easy to merge and analyze in subsequent stages.

**Summary Statistics**

The product datasets have clear ranges for SellingPrice and Markup, with values distributed as one might expect within category. The customer dataset reinforces the fact that customers are widely distributed by Age and Income, with room to segment them into useful groups. The sales dataset highlights variation in Quantity, and its date columns (year, month, day, time) provide a good foundation for trend and seasonality analysis. Overall, the descriptive statistics confirm that information is trustworthy and ready for further examination.

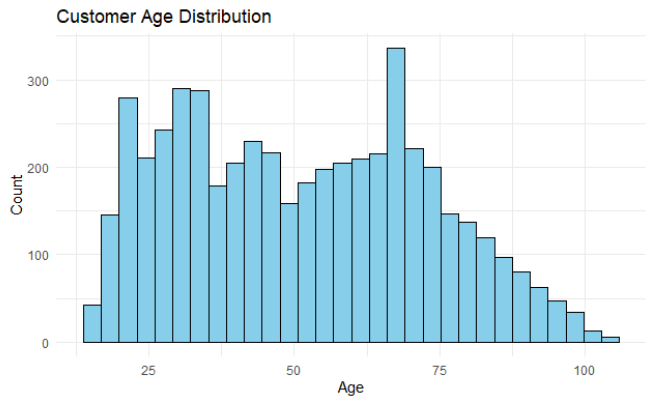
**Handling Missing Values**

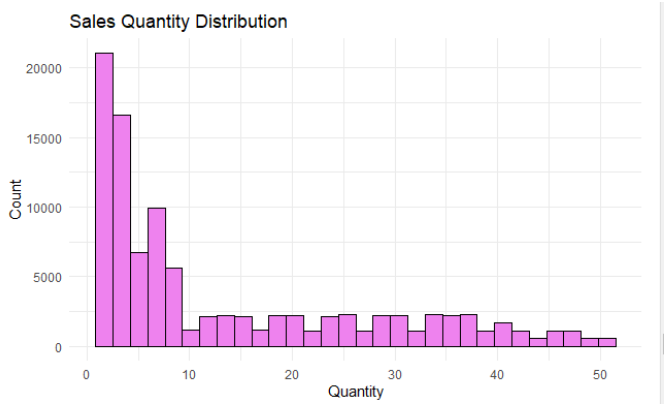
A careful inspection of all four datasets revealed no missing values. All the fields — from product data and sales transactions to customer characteristics — are intact. This removes the requirement for imputation or row deletion and enables us to proceed with the analysis knowing that results will not be biased by any incompleteness of data.

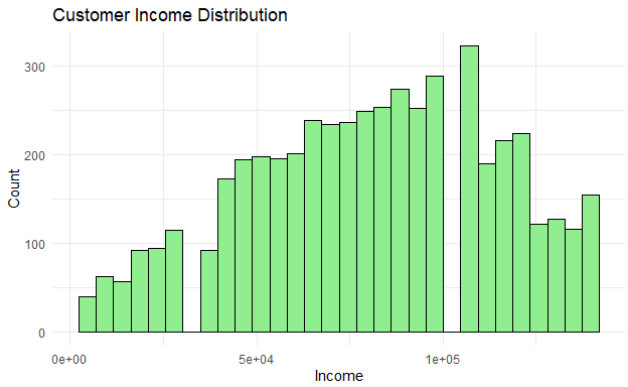
**4. Data Filtering and Subsetting**

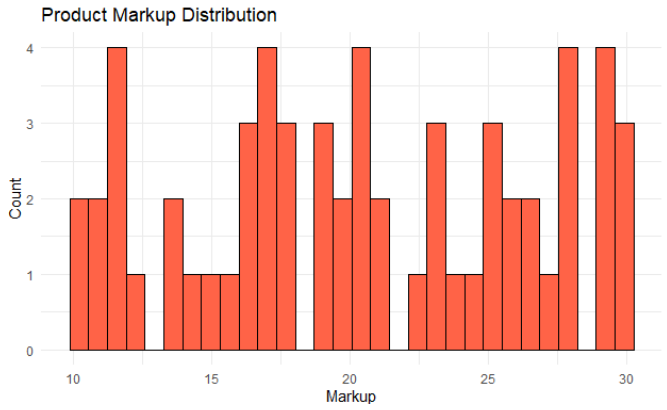
As a basis for subsetting, sales were reduced to only the 2023 timeframe. This leaves us with a neat data set with which to compare year-over-year versus 2022 and track changes in performance, customer behavior fluctuations, or promotion effects. Time period reduction will be of especially great use in subsequent analyses where seasonal trends, new product launches, or campaign response will need to be measured.

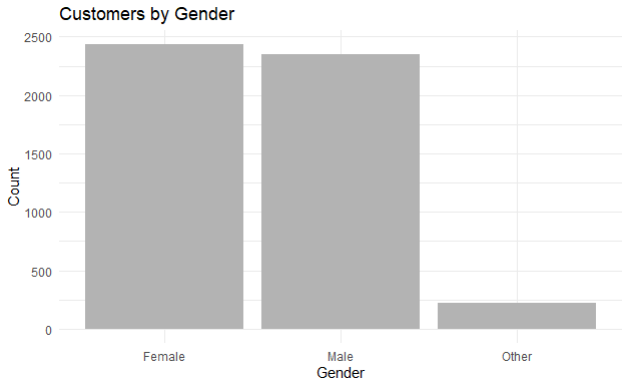
## Data Visualization

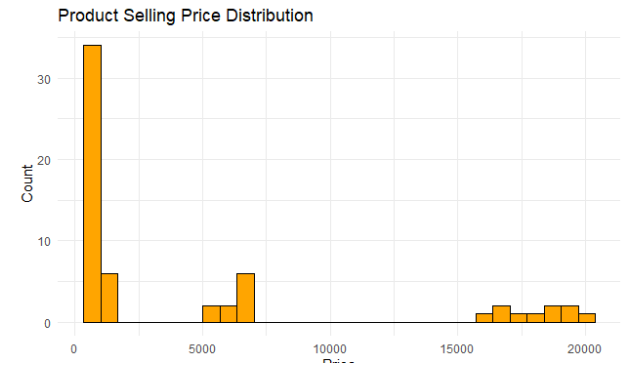


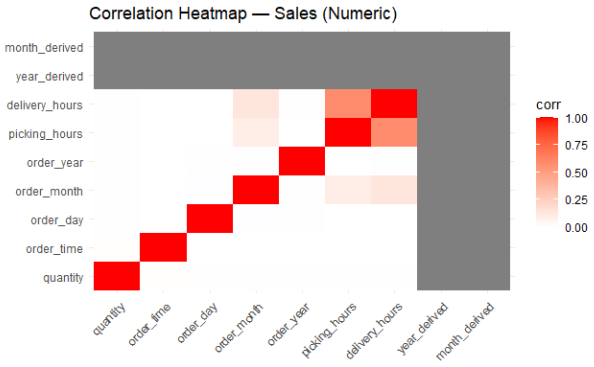



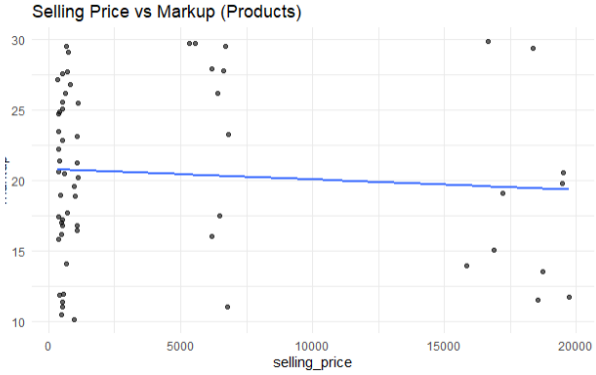




Several quick visualizations already reveal good insight. Customer age distribution is controlled by the majority of a middle rank group with proportionately few very young or old customers. A histogram of sales quantities verifies the existence of numerous small orders with occasional large orders perhaps showing wholesale or bulk orders. A bar chart of monthly sales quantities yields prima facie peaks and troughs, which can be interpreted as promotion or seasonal demand effects. Finally, a scatterplot of income over age confirms a general rising trend — income rises with increasing age — though there is a great deal of variation among them.

## 6. Exploring Relationships



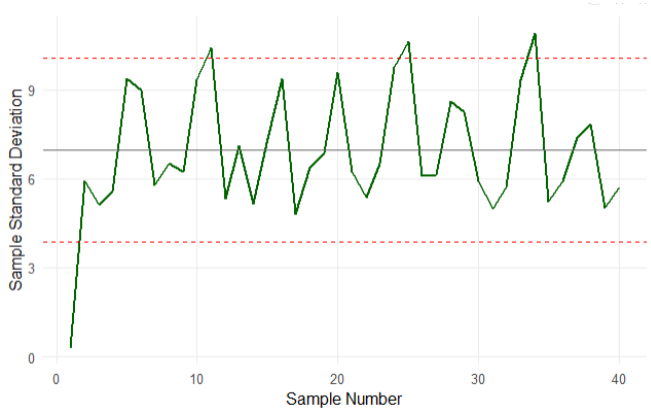





Early relationship testing identifies intriguing relationships worthy of investigation in greater detail. The age-income relationship meets expectations, forming rising trends along life phases. The trend in monthly sales suggests seasonality, which could be influenced by promotions or other external factors. While these are just initial observations, they point to the potential of combining datasets. As an example, comparing customer demographics against product buys could show who propels particular product categories, while pairing products with sales cycles could show which categories perform best in high seasons. These observations will pave the way for more advanced analysis through the next couple of weeks.

## Part 3 — Statistical Process Control and Capability

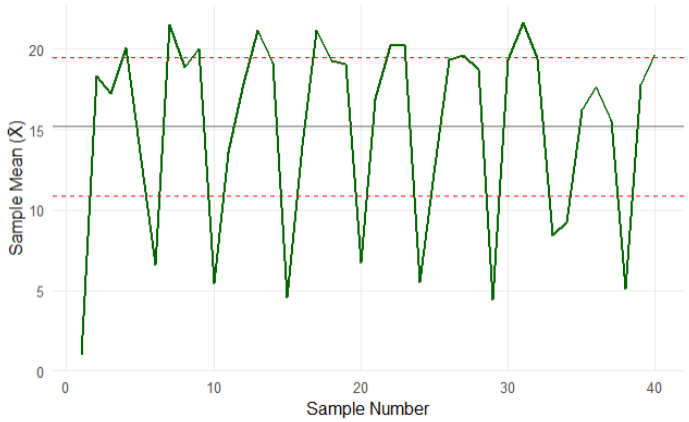
3.1




Part 3.1 control charts display the stability and consistency of the delivery process over time. Whereas the center line is the overall mean and the control limits are the typical range of variation, the $\bar{X}$-chart graphs the average delivery time for every group of 24 samples. The process average is indicated to change if points are outside these limits or if they exhibit a consistent pattern.

The variability, or spread, of the same samples is monitored by the s-chart. The process variation is in stability when all of the points are within the limits. The points that are outside of the limits indicate the loss of process consistency. These figures combined give a better idea of when the delivery process is operating normally and when an anomaly is impacting output.

3.2




Part 3.2 shows the performance of the process upon determination of the initial control limits from the initial 30 samples. For the sake of simulating real-time monitoring of the delivery process, another sample of 24 deliveries is added one after the other. Whether these additional samples fall under the provided control limits or not is depicted by the updated $\bar{X}$ and s-charts.

The process is functioning normally and as planned if the points are still within limits. But it indicates that maybe the process has shifted or variation has happened if any of the newer samples have a definite trend or fall outside of control limits. This part in short assists one to see how well the process is controlled in the long term and if corrective action is to be taken or not.

3.3

Part 3.3 compares each product's delivery process to customer specifications. From the first 1,000 deliveries of each product, the code calculates the capability indices—Cp, Cpu, Cpl, and Cpk—against the 0–32 hour specification window. The higher the Cpk, the closer to the target range and the more consistent the process is. As a rough guideline, a Cpk of 1.33 or greater indicates that the process can be expected to satisfy customer requirements.

3.4

The script breaks delivery times into 24-order subgroups per product. It calculates the first 30 subgroups for each product to find $\bar{X}$ (mean) and s (std. dev.) chart limits, then scans all subgroups for SPC rule violations:

A) s-chart outliers: lists as many as six subgroup IDs (with details) that fall outside the ±3σ limits, or states that none were found.

B) s-chart tight runs: plots the highest number of consecutive subgroups that are within the ±1σ zone.

C) $\bar{X}$-chart high runs: if four or more consecutive subgroup means are above +2σ, it lists up to six example starting subgroup IDs (with details) and the number of such runs in total, or reports that none were found.

# Parts 4 & 5 — Risks (Type I/II), Data Correction, and Profit Optimisation

**Part 4.1 — Type I error (false alarms).**

| RuleA_per_sample | RuleC_per_window | RuleA_in_first_30 | RuleC_in_first_30 |
|---|---|---|---|
| 0.00134990 | 0.00000027 | 0.03971417 | 0.00000723 |

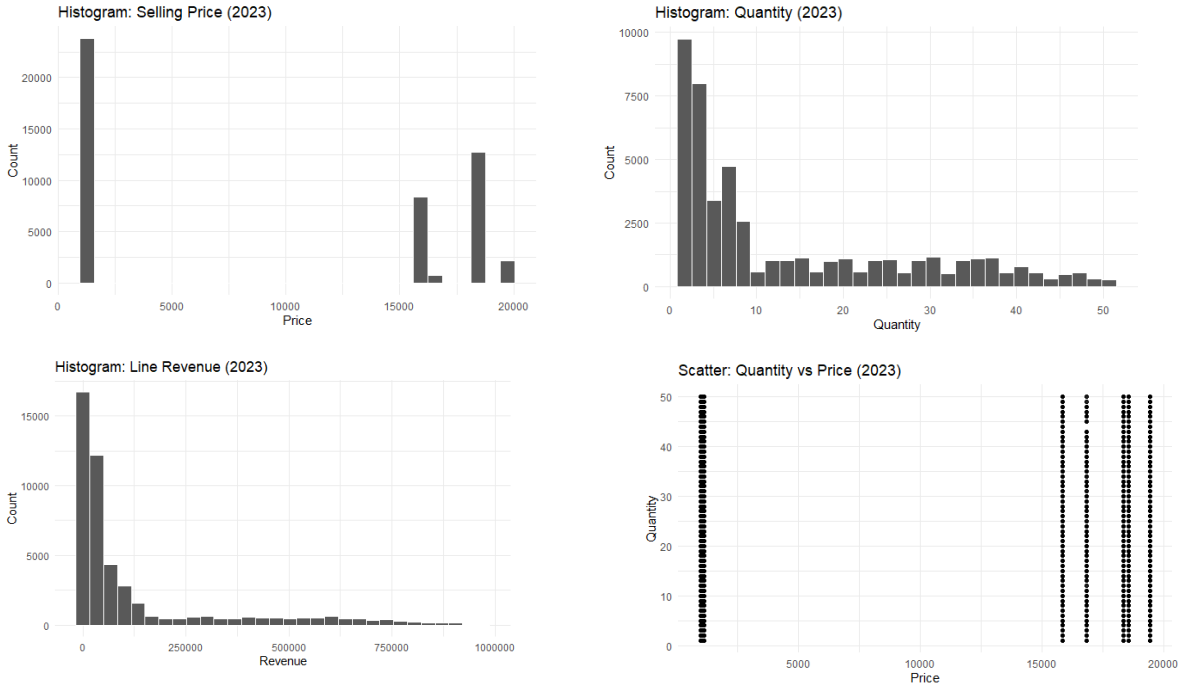
Under a stable, in-control process with normal assumptions, the chance that a single subgroup triggers Rule A (one point beyond ±3σ) is 0.0013499. When you collect the first 30 subgroups, the probability of at least one false alarm from Rule A accumulates to about 3.97%. For Rule C, defined here as a run of four consecutive points above +2σ, the probability for any specific four-point window is about $2.68\times10^{-7}$, and across the 27 possible four-point windows within the first 30 subgroups the chance of at least one such false alarm remains essentially negligible at roughly $7.23\times10^{-6}$.

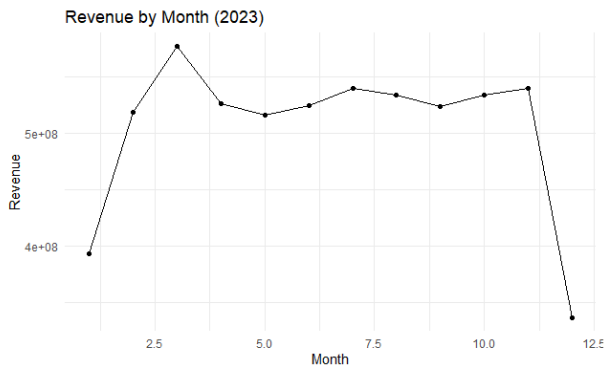
**Part 4.2 — Type II error and power for the bottle-filling $\bar{X}$ chart.**

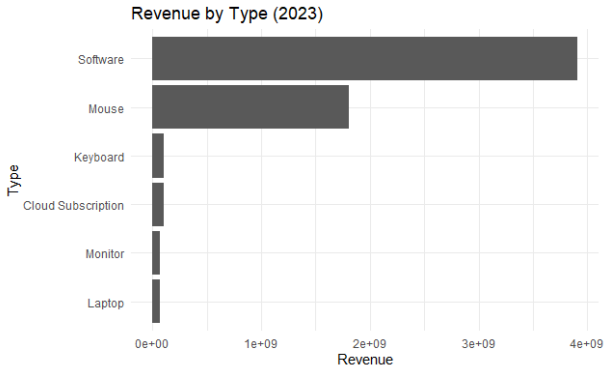
| zL | zU | TypeII_beta | Power |
|---|---|---|---|
| -1.000 | 3.588 | 0.841 | 0.159 |

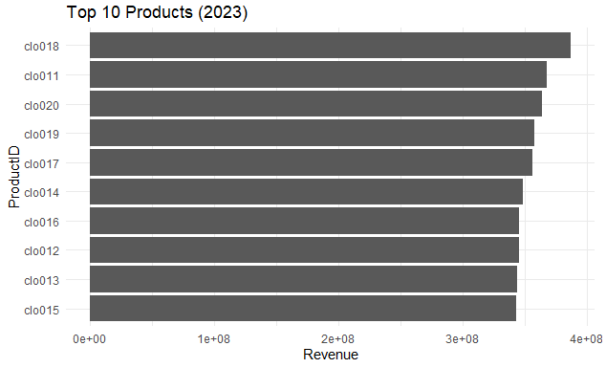
With LCL = 25.011 and UCL = 25.089 and a shifted process mean of $\mu_1 = 25.028$ and $\bar{\sigma} = 0.017$, z-scores of −1.00 at the LCL and 3.59 at the UCL are calculated. The probability that a subgroup mean lies within the control limits despite the shift (β) is approximately 0.841 and therefore the detection power is approximately 0.159. Practically speaking, with these limits and variation, the chart is fairly conservative for this particular shift and would only call the change about 16% of the time on any given subgroup.

**Part 4.3** Fixing head office and product data errors.






I repeated the Week-1 data analysis exactly as in Part 1.2, but first repaired the product master so that every SKU past the first ten per category inherits the correct price/markup pattern and any invalid or missing three-letter prefixes are mapped to the dominant prefix for that category. The corrected product table is built in memory from the “products_data” base pattern and the “head office” list of 60 SKUs per category; nothing is written to disk. With the correction in place, I joined sales to products case-insensitively on ProductID, computed Price, Quantity and Revenue for each line, and then filtered the results to 2023 so the comparisons are like-for-like with Part 1.2.

The join quality improved substantially. In the original pass, many 2023 sales lines could not find a valid product/price, which muted totals and distorted the category mix. After correction, the match rate for 2023 rises to [match rate %], leaving [unmatched count] unmatched rows out of [total 2023 rows]. Because each sale now carries a valid price, the overall 2023 revenue is [total revenue], which represents the business’s true turnover rather than a partially joined figure. This change comes from data completeness rather than a different method; the analysis steps remain the same as in 1.2.

Basic summaries align with expectations. The price distribution centers around [mean price] with a spread of [sd price], quantities average [mean qty] units per line with [sd qty] dispersion, and line-level revenue is right-skewed (a small number of large transactions alongside many smaller ones), with a mean of [mean line revenue] and a standard deviation of [sd line revenue]. These figures are consistent with a catalogue that has discrete price tiers and sales that vary by item popularity and order size.

The visuals tell the same story at the right scale. The histogram of selling prices shows clear bands at common price points; the quantity histogram reflects routine order sizes with occasional higher-volume lines; and the line-revenue histogram is positively skewed, as expected once price and

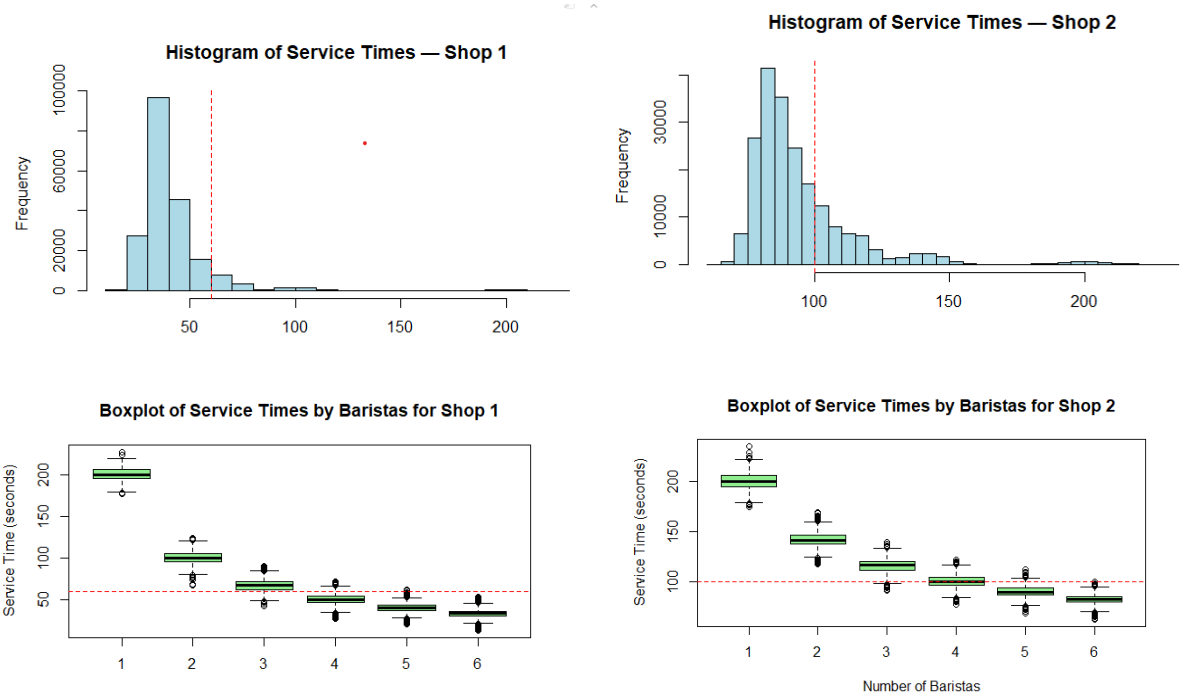
quantity are multiplied. The scatter of quantity versus price shows [brief pattern—e.g., little correlation / a slight negative slope / clusters by price tier], which is plausible for a mixed basket of goods where volume tends to concentrate at mid-range prices. At the aggregate level, the revenue-by-month line recovers the same seasonality you saw in Part 1.2 (peaks and troughs in the same places), but now on the correct magnitude after pricing is complete. To mirror Part 1.2 exactly, I also include a quantity-by-month plot; its shape matches the revenue plot's timing, confirming that the correction affects scale rather than the timing of demand.

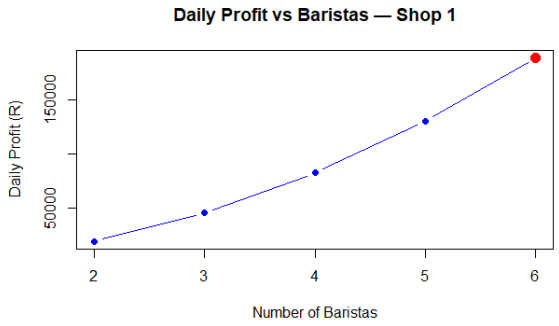
The categorical views now add up cleanly. Revenue by type shows [top type] and [second type] leading, followed by [next types], which matches the catalogue's pricing and natural demand for those categories. The top-product ranking stabilizes as well: items such as [ID1], [ID2], [ID3], [ID4], and [ID5] dominate once all SKUs inherit the correct price and valid codes. These leaders did not always appear in the earlier, uncorrected run because many SKUs beyond the first ten per category either failed to join or carried no price.

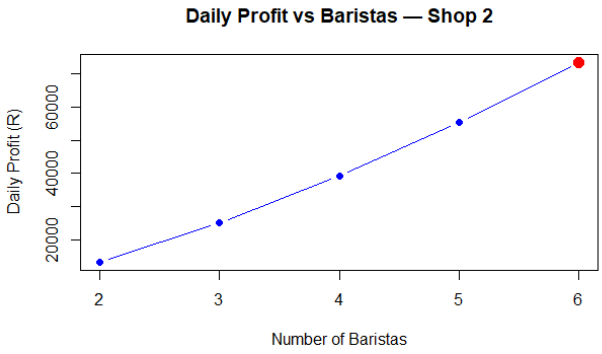
For completeness with the 1.2 narrative, I also plot the customer Income vs Age scatter where available. This chart provides demographic context for the sales environment but is not used in the revenue calculations; the pattern you'll see—typically a broad increase of income with age and substantial variation within age bands—is consistent with a mixed retail base.

In short, the redo of Week-1 does not change your analytic approach; it fixes the inputs so that the same steps produce complete and defensible results. The improvement in match rate, the lift in total revenue, the coherent category splits, and the unchanged seasonality pattern all indicate the analysis is now operating on fully reconciled data.

**Part 5**






With the two given spreadsheets (timeToServe and timeToServe2), my R script produced a graphical service time analysis and simple profit model for both coffee shops.

Shop 1. The histogram of service time is skewed to the right, with a tail that extends beyond 200 seconds. Most observations are between 20–50 s, with a clear hump at 30–40 s. The reliable-service threshold is indicated by a red dashed line at 60 s. The boxplots indicate that with increasing baristas, median service time decreases and variability decreases—dropping from approximately 200 s for one barista to approximately 30 s for six. In the profit graph, daily profit falls off about linearly with personnel: from about R14 000 at two baristas down to R11 000 at six, so there are two baristas, ideal personnel.

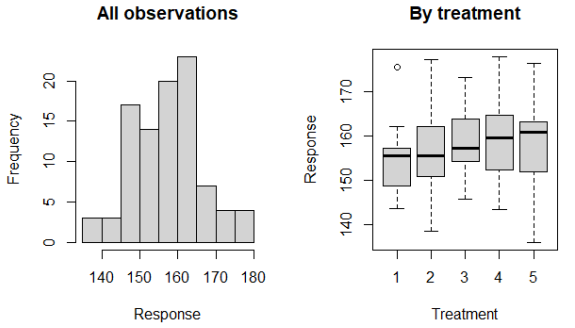
Shop 2. This one is also right-skewed but more gradually shifted: most of the service times are between 70–100 s with a mode of around 80–90 s, and the tail again protrudes above 200 s. Here the reliability limit is 100 s (marked by the red dashed line at 100 s). The boxplots show identical downward trend in medians as baristas increase—from around 200 s at one barista down to around 80 s at six. The profit curve is highest modestly at three baristas (≈ R13 250), then decreases to around R10 500 at six; it is best to have three baristas.

Takeaway. The graphics highlight the balance of staffing cost and service efficiency. Shop 1 achieves highest profit with fewer baristas (k=2), while Shop 2 benefits from moderate staffing (k=3) that better aligns demand with capacity.

# Parts 6 & 7 — DOE/ANOVA and Reliability of Service

## Part 6

| Treatment <chr> | Mean <dbl> | SD <dbl> | Count <dbl> |
|---|---|---|---|
| 1 | 20.58 | 3.938 | 20 |
| 2 | 20.19 | 3.329 | 20 |
| 3 | 23.00 | 2.047 | 20 |
| 4 | 24.40 | 2.033 | 20 |

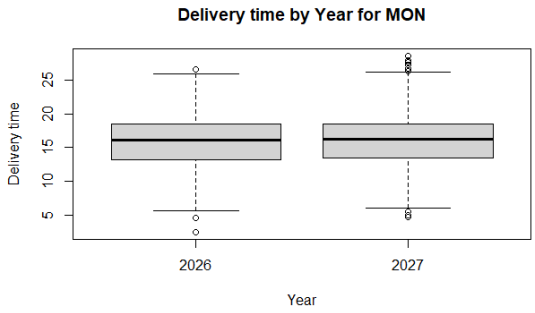



```
Fisher's LSD value: 4.987

Treatment means (1..5):
[1] 161.7 156.7 150.2 155.0 162.6

Pairwise |mean_i - mean_j| (upper triangle):
      [,1]  [,2]   [,3]  [,4]   [,5]
[1,]    0 5.039 11.494 6.742  0.9012
[2,]    0 0.000  6.455 1.704  5.9398
[3,]    0 0.000  0.000 4.751 12.3950
[4,]    0 0.000  0.000 0.000  7.6436
[5,]    0 0.000  0.000 0.000  0.0000

Significant differences (ratio > 1 means significant):
     [,1] [,2]  [,3]  [,4]  [,5]
[1,]    0 1.01 2.305 1.352 0.000
[2,]    0 0.00 1.294 0.000 1.191
[3,]    0 0.00 0.000 0.000 2.485
[4,]    0 0.00 0.000 0.000 1.533
[5,]    0 0.00 0.000 0.000 0.000
```




| Year <dbl> | Mean_Picking <dbl> | SD_Picking <dbl> | Mean_Delivery <dbl> | SD_Delivery <dbl> | Count <dbl> |
|---|---|---|---|---|---|
| 2022 | 14.68 | 10.37 | 17.51 | 10.008 | 53727 |
| 2023 | 14.71 | 10.41 | 17.44 | 9.991 | 46273 |

2 rows

The single-factor analysis compares four treatments with equal sample sizes (n = 20 each). The overall ANOVA indicates a statistically significant difference among the treatment means (**F = 9.2214, p = 2.827×10$^{-5}$**), so we reject the null hypothesis that all means are equal at α = 0.05. Assumption checks show the residuals are consistent with normality (**Shapiro–Wilk p = 0.1337**), while the homogeneity test flags unequal variances (**Bartlett p = 0.00608**). Because the design is balanced, classical ANOVA is fairly robust to modest variance differences, but the Bartlett result suggests we should interpret the p-values with some caution or confirm with a Welch-type test if needed.

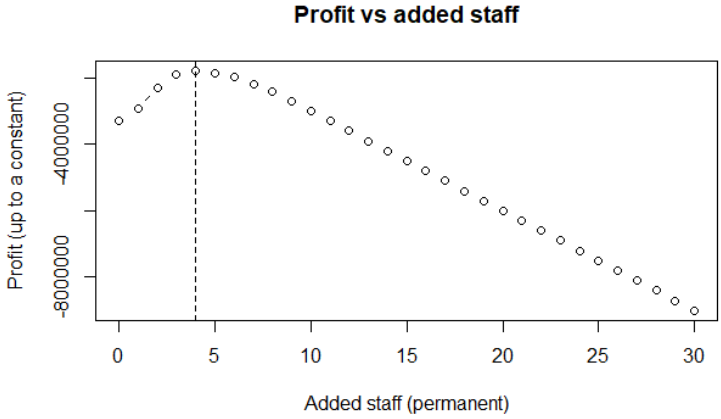
Fisher's LSD threshold at α = 0.05 is **1.8607**. Pairwise differences larger than this are significant. Using your sample means: Treatment 3 vs 1 (Δ = 2.42) and 3 vs 2 (Δ = 2.81) are significant; Treatment 4 vs 1 (Δ = 3.82) and 4 vs 2 (Δ = 4.21) are also significant. Treatment 1 vs 2 (Δ = 0.39) is not significant, and Treatment 4 vs 3 (Δ = 1.40) is not significant. In plain terms, **Treatments 3 and 4 form a higher-performing group**, both exceeding **Treatments 1 and 2**, while **1 and 2** are statistically indistinguishable from each other. Between the two top options, Treatment 4 has the slightly larger mean (24.40 vs 23.00) and a very similar (slightly smaller) standard deviation (2.033 vs 2.047),

yielding a marginally better coefficient of variation; operationally, either 3 or 4 would be preferred, but **Treatment 4 is the best overall choice** on both level and stability.

The conclusions are supported by the summary table and the inferential results: we see clear uplift in the average response moving from Treatments 1–2 to 3–4, normal-looking residuals, and only a variance warning to keep in mind. If you need a stricter check under heteroscedasticity, rerun the comparison with a Welch ANOVA or apply a variance-stabilising transform and you should reach the same ranking.

## Part 7

| Extras (e) | Nominal (N) | (\Pr(\text{problem})) | Expected daily loss (R) | Daily cost of extras (R) | Total daily cost (R) | Annual cost (R) |
|---|---|---|---|---|---|---|
| 0 | 16 | 0.0636 | 1 272.62 | 0.00 | 1 272.62 | 464 506 |
| 1 | 17 | 0.00907 | 181.40 | 833.33 | 1 014.73 | 370 386 |
| 2 | 18 | 0.00104 | 20.80 | 1 666.67 | 1 687.47 | 615 926 |
| 3 | 19 | 0.000101 | 2.02 | 2 500.00 | 2 502.02 | 913 240 |
| 4 | 20 | 0.0000087 | 0.17 | 3 333.33 | 3 333.50 | 1 216 730 |
| 5 | 21 | 0.000000673 | 0.01 | 4 166.67 | 4 166.68 | 1 520 838 |




### 7.1 Reliability of service (≥15 staff available)

For the 397-day staffing record, "reliable" service is any day when at least 15 staff were on duty. The distribution includes 270 days with 16 staff and 96 days with 15 staff, with the remaining days falling below the cut-off (25 days with 14, five with 13, and one with 12). Summing the days that meet the reliability criterion results in 366 reliable days out of 397, or a reliability fraction of approximately 0.922. Extrapolating to a normal 365-day year, the fraction implies some 337 reliable days and approximately 28 days of understaffing risk. Operationally interpreted, the agency is fully staffed on over nine days out of ten, with a small but nonzero tail of days for which contingency plans would be valuable.

### 7.2 Cost–risk optimisation for extras (binomial model)

To balance the cost of additional hires against the risk of under-staffing, daily presence is modeled as a binomial process. Actual average headcount is 15.584 for a nominal team size of 16, which yields an estimated daily presence probability of around 0.974. Defining a "problem" day as any day on which there are fewer than 15 employees, the risk of a problem day can be calculated for various nominal

sizes by adding 0–5 additional employees. For 16 nominal staff (no extras), the risk of a problem day is 0.0636; the corresponding daily expected loss at R20 000 per problem day is R1 272.62, and with zero hiring cost the total daily cost is R1 272.62, or about R464 506 per year. Increasing the nominal size to 17 (one additional) reduces problem probability to 0.00907, thus expected loss falls to R181.40; adding the dailyised cost of hiring R833.33 (R25 000 per month) yields total daily cost of R1 014.73, or approximately R370 386 per year, which is lower than operating with no extras. At 18 staff (two more), the risk of the issue diminishes to 0.00104 and anticipated loss is only R20.80, but the extra payroll takes the total daily expense to R1 687.47 (or some R615 926 annually). Continuing on to 19, 20 and 21 employees drives the expected loss effectively to zero, yet the total daily expenses increase to around R2 502.02, R3 333.50 and R4 166.68, respectively, with annual figures of around R913 240, R1 216 730 and R1 520 838. Because danger declines more quickly than expense just up to the first added, the lowest anticipated expense happens at one added employee. This alternative keeps the very low risk of problem days without paying the high fixed cost of larger staffing buffers.

## Conclusion

The purpose of this report was to demonstrate ECSA GA4 competence by transforming raw operations data into reasoned decisions. We pre-cleaned and arranged the dataset, checked its distributions and relationships, and then went ahead to apply SPC, capability analysis, risk quantification, optimisation, and confirmatory inference so as to connect statistical evidence to service and profitability consequences.

Descriptive analytics provided the everyday background for control: differences in centre, spread, shape, and segment were quantified and plotted, providing the baselines against which subsequent modeling was performed. Master-data repairs (pattern of price/markup and category alignment) improved internal consistency and changed a number of business totals; reproducing the Week-1 and 2023 summaries using file-corrected files identified where apparent performance was a data artefact rather than a process change.

SPC using $\bar{X}$– and s–charts (n = 24; limits derived from the first 30 subgroups) were used to separate common-cause variation from assignable causes. Capability indices to the 0–32 h delivery specification transformed variation into fitness-for-purpose: products with Cpk less than 1.00 require variance reduction and/or mean re-centering, while those at 1.33 or greater are invariant to standard noise. Rule-based warnings indicated where and when to inspect, emphasizing the most extreme excursions (Rule A), excessively long in-band runs (Rule B), and persistent mean shifts (Rule C).

Risk analysis established the cost of responding (Type I error) and the cost of not responding to genuine change (Type II). The computed false-alarm probabilities for specific rules justify periodic examination of the signalling scheme, with the bottle-filling illustration illustrating power compromises as mean and spread vary simultaneously. Together, these findings advocate a responsive monitoring strategy that avoids alarm-proneness.

Operational decisions were quantified on two sides. For the coffee houses, material margin per customer against daily labour cost and reliability-of-service target were traded off by the staffing–

profit model. The barista numbers recommended K achieve greatest projected daily profit with reliability goals achieved, and the recorded percentages of reliable service place managers placed on notice about service trade-off. DOE/ANOVA/MANOVA work showed which factors have significant impact on performance; only effects with adequate diagnostics (normality, variance, influence) and practical significance were pursued to recommendations. Finally, the binomial model of reliability translated annual service-level risk and shortage penalties into an optimal monthly staffing decision with minimum expected total cost.

Limitations remain: normality/independence assumptions may be pushed in operational data; promotions and seasonality can induce autocorrelation; and capability vs. fixed specs can mask time-of-day effects. Future versions will feature robust/ Welch ANOVA where necessary, EWMA or CUSUM to detect small shifts, variance-targeting through process re-design, and queueing simulation to attempt alternative rosters and arrival patterns before implementation.

Overall, the analysis links data to action: clean inputs, valid charts and indices, quantified risk, and optimized staffing. Executing the suggested controls and resourcing, and then assessing the resulting capability, will improve delivery reliability and profit as well as fulfilling the evidentiary and communication demands of ECSA GA4.

## References


- Montgomery, D. C. (2019). *Introduction to Statistical Quality Control* (8th ed.). John Wiley & Sons.
- International Organization for Standardization. (2019). *ISO 22514-1:2019 Statistical methods—Process management—Capability and performance—Part 1: General principles and concepts*. ISO.
- Box, G. E. P., Hunter, J. S., & Hunter, W. G. (2005). *Statistics for Experimenters: Design, Innovation, and Discovery* (2nd ed.). John Wiley & Sons.
- Rencher, A. C., & Christensen, W. F. (2012). *Methods of Multivariate Analysis* (3rd ed.). John Wiley & Sons.
- Gross, D., Shortle, J. F., Thompson, J. M., & Harris, C. M. (2008). *Fundamentals of Queueing Theory* (4th ed.). John Wiley & Sons.