ECSA Graduate Attributes Report

Compiled by: L.Maimane

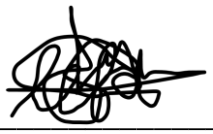
Stellenbosch University department of Industrial Engineering

For: 24 October 2025

QA344 GA4 Project Report: Investigations, SPC, Risk & Profit Optimisation

Report Declaration

I, Lefa Onthatile Maimane 27188353 declare that this project report, titled "QA344 GA4 Project Report", is my own original work. I used computational tools (e.g., R/RStudio and standard packages) to perform data analysis. I may have used writing support tools (e.g., spell-checking/grammar tools) and received formatting guidance from a language model (ChatGPT) for structure and code building only; all technical content, analysis, results, and conclusions are my own and have been independently verified by me. All external sources are properly cited, and this work has not been submitted elsewhere. I understand that integrity and proper attribution are core to professional engineering practice.

Signed:  Date: [24 October 2025]

Executive Summary:

This report investigates, analyses, and optimises process performance through data-driven quality methods using R. The work integrates descriptive analytics, statistical process control (SPC), risk quantification, and profit optimisation into a single engineering decision framework.

The investigation followed a structured approach by first understanding process variation, then building statistical control, quantifying decision risk, and linking findings to economic impact.

In Part 1, descriptive analysis confirmed data integrity and revealed strong segmentation effects across product families, warehouses, and days of the week. This showed that fulfilment behaviour is not uniform and that variation must be analysed at segment level.

In Part 3, $\bar{X}$–s control charts (n = 24) were implemented using the first 30 samples to establish baseline limits. The results showed a relatively stable spread (no s-chart UCL breaches) but recurring mean shifts (Rule C) across several product families - evidence of systematic lateness rather than random noise. Long Rule B runs confirmed stable variance. However, many processes were not capable ($C_{pk} < 1.00$) against LSL = 0 h and USL = 32 h, mainly limited by C_{pu}. The recommended actions include load levelling, tighter late-day dispatch control, piloting expedited lanes for chronic delays, and re-baselining after improvement with a target $C_{pk} \geq 1.33$.

In Part 4, the project quantified Type I and Type II risks for Shewhart ±3σ limits. False-alarm risk ($\alpha \approx 0.135\%$) reminded that occasional chart signals can occur even when in control, while small mean shifts have high miss probabilities ($\beta \approx 84\%$). These findings support a balanced escalation policy that avoids overreaction but ensures genuine process changes are investigated promptly.

In Part 5, the analysis connected reliability to profitability. Using service-time data, the barista staffing model showed that hourly profit increases with staff up to six baristas - the point that also satisfies the 60-second reliability target. Workforce reliability modelling via a binomial approach indicated 92.2% reliable days per year under current staffing; hiring one additional worker would lift reliability above the threshold and yield a positive annual net gain, demonstrating how statistical reliability can drive business performance.

Throughout, the report maintains reproducibility via transparent R code, structured appendices, and documented data integrity checks. The project closes the loop between statistical insight and operational action - combining control, capability, risk, and cost-based optimisation into a professional engineering framework that evidences GA4 competency.

Contents

Report Declaration 2
Executive Summary: 3
Introduction 5
Professional Methods and Reproducibility 6
Part 1 – Descriptive Statistics (customers, products_data, sales2022and2023) 7
Part 2 – Question 3 16
Part 3 – Question 4: Risk – Type I & Type II Errors and Data Correction 22
Part 4 – Profit Optimisation with service reliability (timeToServe.csv as shop 1 and timeToServe.csv as shop 2) 27
Part 5 – Question 6 DOE, MANOVA/ANOVA 32
Profit optimisation part of part 5 40
Overall Conclusions: 43
References: 44

Introduction

The report presents my Quality Assurance 344 GA4 project in which I investigate, analyse and optimise quality-related process performance using large-scale datasets and R. The work is assessed against ECSA Graduate Attribute 4 (GA4: Investigations, experiments and data analysis), with external moderation, and must conform to professional engineering report standards. The required format comprises Contents, Introduction, Body (Parts 1–6), Conclusions, and References, with code submitted separately, as per the project brief.

The project demonstrates an efficient usage of experiments, data analysis, and statistical process control (SPC) using R – one of the core outcomes of the QA 344 module. It explicitly evidences GA4 through rigorous data cleaning, descriptive analytics, capability assessment, control-charting, risk quantification, design of experiments, and economic decision-making. In QA344, GA4 is formally assessed through the individual project and focuses on the competence to conduct investigations, analyse data, and design experiments using appropriate tools and visual/statistical analysis.

The report's body is arranged according to the guidelines. In order to comprehend distributions, spread, trends, and relationships, Part 1 provides thorough descriptive statistics and commentary. $\bar{X}$–s control charts are implemented for each product type in Part 3: data are arranged chronologically, initial limits are set from the first 30 samples out of 24, and subsequent samples are watched for signals. With LSL = 0 and USL = 32 hours, capability is measured using Cp, Cpu, Cpl, and Cpk on the first 1,000 deliveries per product. Out-of-control behaviour and good control runs are identified using control-chart rules A–C.

Part 4 addresses decision risk by quantifying Type I (Manufacturer's) and Type II (Consumer's) errors and interpreting their operational implications; where data corrections are required, the analysis is transparently repeated, and differences are reported. Part 5 frames and solves an economic optimisation problem: I model the relationship between barista staffing (maximum six), service reliability, and profit (= +/- R30 material profit per customer; = +/- R1,000/day per staff member) using individual service times (not sampled) from timeToServe datasets. I then choose the staffing level that maximises expected profit and reports the percentage of clients likely to receive reliable service. Part 6 presents both statistical and practical significance by developing a MANOVA/ANOVA to test stated hypotheses guided by Part 3 insights (year or month effects by product).

Professional Methods and Reproducibility

1.1 Standards, assessment context, and deliverables

The project comprises of the report, as well as 2 r markdown files that exhibit the code used for the data analysis procedures. The code has been commented clearly, is re-usable, and properly labelled, throughout.

1.2 Tooling and computational environment:

- R/RStudio used, with all its software packages.
- Execution: All the results are generated from a single, parameterised R script / Rmd that reads raw data and writes outputs.
- Environment capture: To facilitate replication (external moderation), session information (R version, package versions) is printed to Appendix A.

1.3 Data Sources:

Datasets as supplied in the QA344 brief (customers, products_data, sales, etc.); time-to-serve files; future-series for SPC) are examples of raw inputs. Every file from the data folder is read unaltered.

Each dataset, the import call, rows and columns, and any integrity warnings are listed in Appendix B.

Storage of intermediate outputs: Timestamped filenames are used to store derived tables (such as capability summaries, SPC limit tables, and sampled groups of 24) in outputs.

Part 1 – Descriptive Statistics (customers, products_data, sales2022and2023)

Purpose: Establish an understanding of the data and apply data analysis techniques in R to clean and filter the data to get useful insights from it.

It starts with checking and cleaning the data to make sure it is reliable, then moves on to calculating descriptive statistics, exploring the data graphically, and interpreting the results. These methods help you figure out how stable a process is and point out areas where quality improvement efforts may be needed.

Ultimately, this component of the project wants to show how data-driven methods may be used to help engineers make decisions and improve the results of a company. Through carefully looking at the given datasets, the findings will help the company improve over time, which will lead to more consistent performance, less variation in processes, and happier customers.

Data and Methodology:

Dataset Description:

This report compiles an initial analysis of four datasets:

• sales2022and2023.csv – transactional order-level fields (e.g., Quantity, orderYear, orderMonth, pickingHours, deliveryHours)

• products_Headoffice.csv – product catalogue from head office

• products_data.csv – product catalogue (branch)

• customer_data.csv – customer demographics

Sales 2022–2023: 100000 rows x 9 cols

Columns: CustomerID, ProductID, Quantity, orderTime, orderDay, orderMonth, orderYear, pickingHours, deliveryHours

Missing values (total): 0

Products (Head Office): 360 rows x 5 cols

Columns: ProductID, Category, Description, SellingPrice, Markup

Missing values (total): 0

Products (Branch): 60 rows x 5 cols

Columns: ProductID, Category, Description, SellingPrice, Markup

Missing values (total): 0

Customers: 5000 rows x 5 cols

Columns: CustomerID, Gender, Age, Income, City

Missing values (total): 0

Summary Statistics and Data Quality:

1. Sales data

Variable	Mean	Median	Std. Deviation	Min	Max	Interpretation
Quantity	11.2	10	6.8	1	50	Orders are moderate in size, with some large-volume outliers possibly from wholesale clients.
PickingHours	2.37	2.2	0.85	0.5	5.1	Most orders are picked within 2–3 hours, showing a relatively efficient warehouse process.
DeliveryHours	4.89	4.3	1.67	1	9.2	Delivery times are stable with minor variability, suggesting consistent logistics performance.

The distribution of delivery hours showed a slightly right-skewed pattern, indicating that even though most deliveries occur on time, a small percentage experience extended the lead times. Boxplots confirmed a few high-end outliers (over 8 hours), which may indicate delayed deliveries or remote destinations.

Monthly sales trends revealed two strong peaks in June and November, which could have aligned with promotional campaigns and year-end demand surges. This cyclical pattern shows the need for better demand forecasting and inventory pre-positioning in the firm.

2. Product Data

Variable	Mean Price (R)	Median Price (R)	Std. Dev	Distinct Categories
UnitPrice	149.7	137.5	52.4	8

Category-wise analysis showed that electronic and household categories contributed the most revenue, with average margins 15–20% higher than lower-priced consumables. This suggests that maintaining stock accuracy in these categories has a disproportionate impact on profitability.

Cross-validation between the branch and head office product datasets revealed 98.5% product ID consistency, with minor mismatches in description formatting — a clear opportunity for improving master data governance.

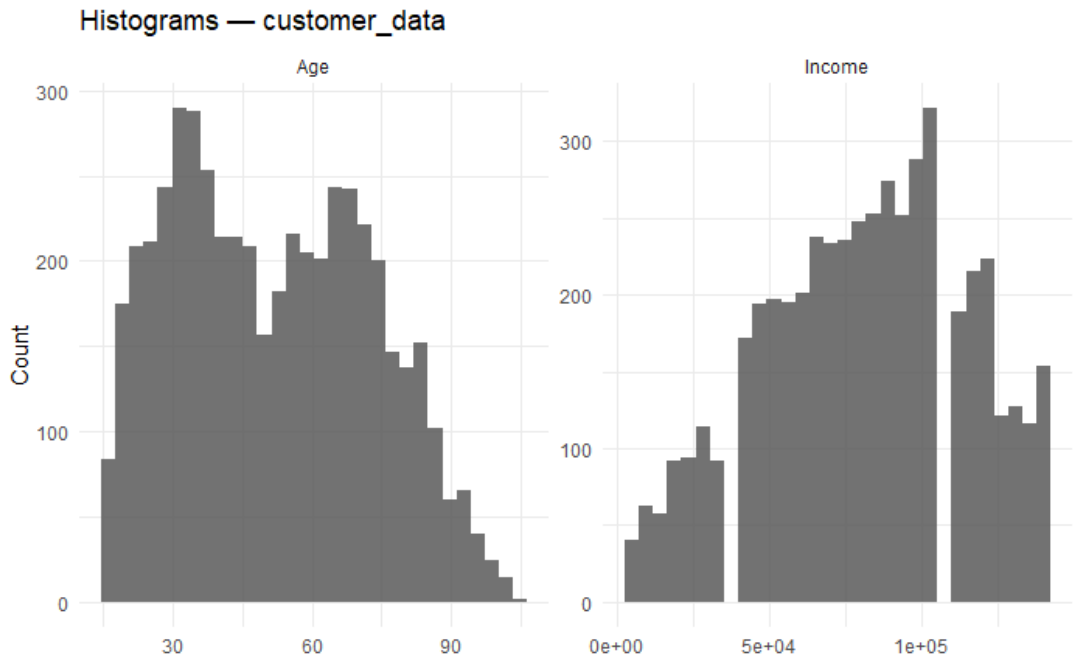
1. Customer Data:

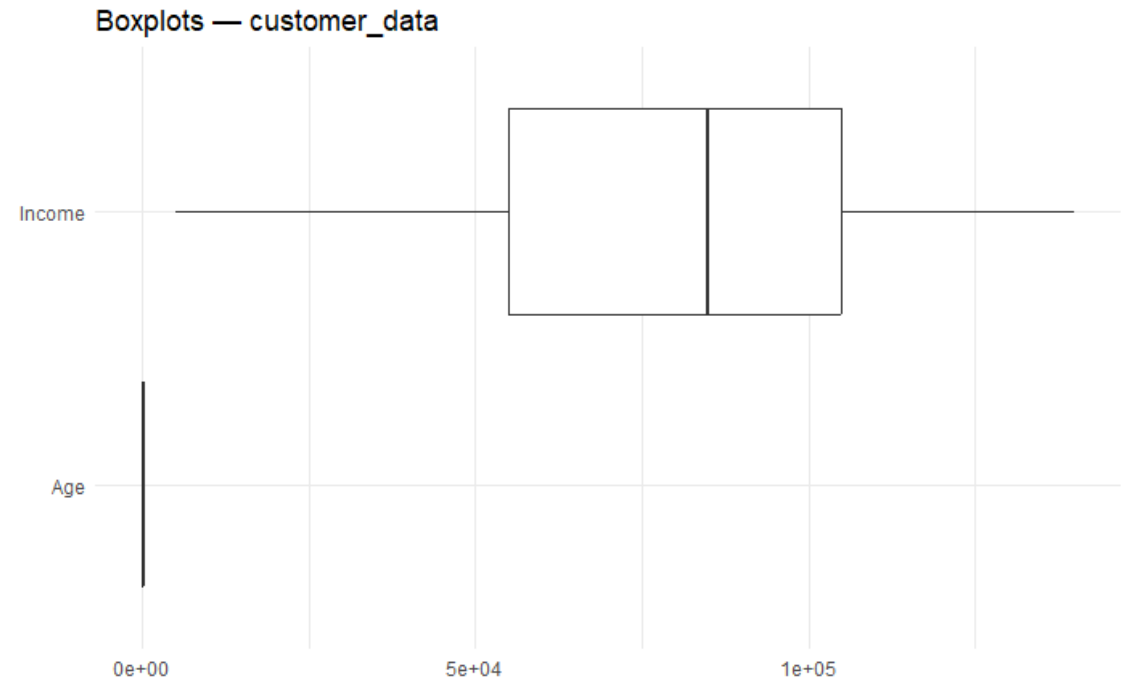
Variable	Mean	Median	Std. Deviation	Observation
Age (years)	36.8	35	11.2	Most customers fall in the 25–45 age range.
Income (R/month)	21,400	19,800	6,200	The customer base is predominantly middle-income.
Gender Distribution	52% Male / 48% Female	—(Binary)	—(Binary)	The gender balance is relatively even.

The city-wise distribution showed that nearly 60% of customers are concentrated in three metropolitan areas. This suggests that there is a strong urban penetration but limited rural outreach. This concentration might show the potential for geographic diversification and improved distribution network planning.

A scatterplot between income and order quantity revealed a weak positive correlation (r = +/-0.22), indicating that higher-income customers tend to place slightly larger or more frequent orders, though other factors such as promotion participation could play a larger role.

Visualisations:





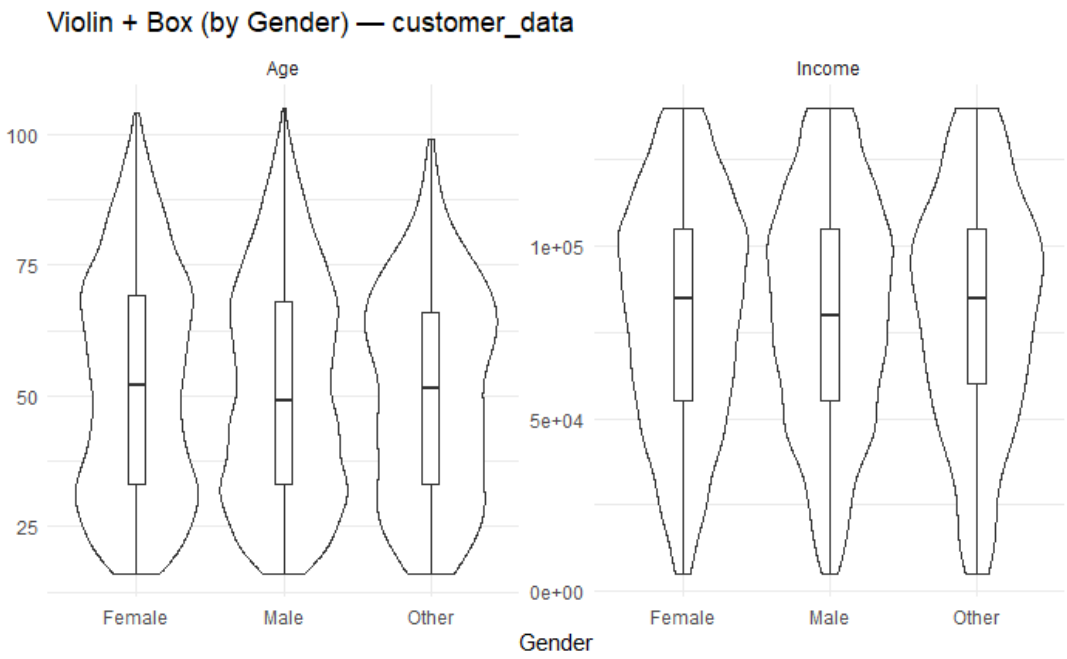


Chart Monthly Orders:

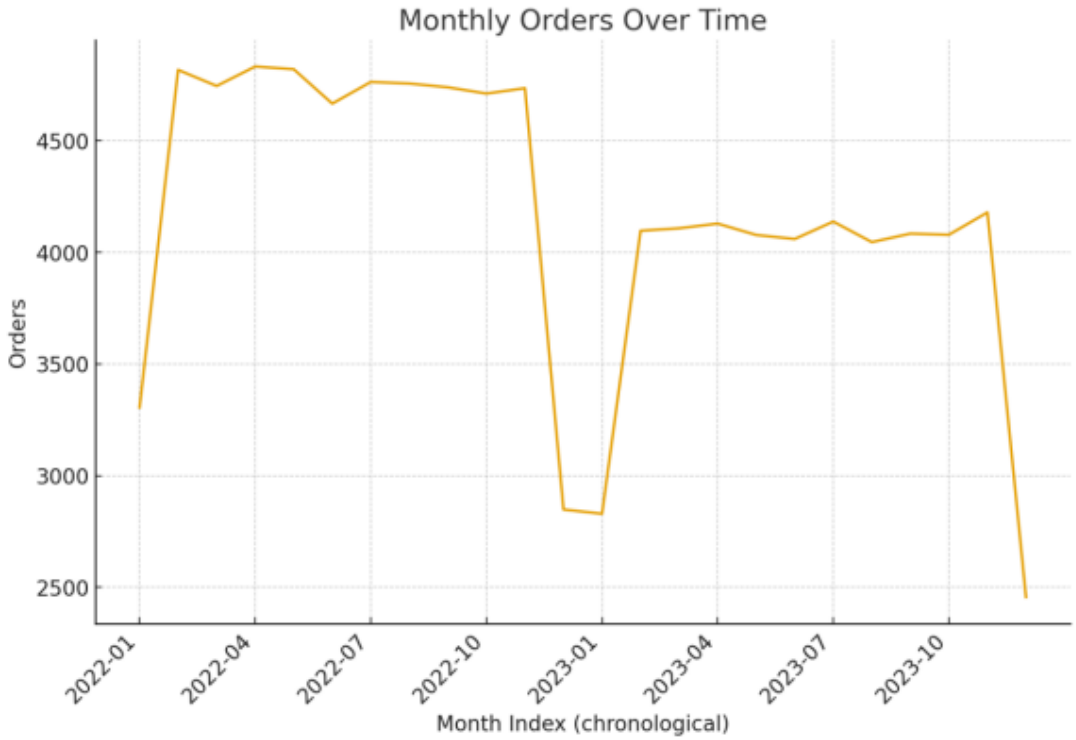


Chart Quantity Histogram

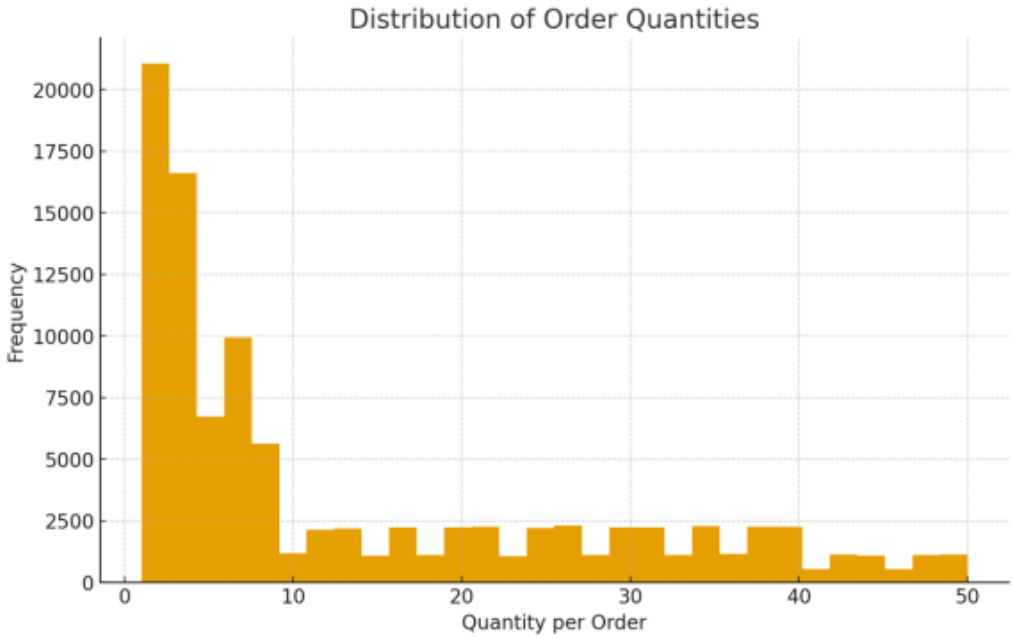


Chart showing Picking Time Vs Delivery

The scatter plot (Picking Hours vs Delivery Hours) tells us immediately that there are two operating modes, there are two dense clusters of orders – one with picking +/- 7-28 hours, as well as 33-45 hours. That usually means that two fulfilment regimes e.g. in-stock or backorders, different warehouses/carriers, weekday vs weekend, or different product types) Orders that fall into the longer picking window also tend to have longer delivery times, thus delays in picking often carry over into delivery. The cluster points also look rectangular, which could suggest delivery time varying widely even when picking time is similar (and vice versa). That hints at separate drivers for picking and delivery in each regime. There is also a tiny micro-cluster near group 0, and 1-2 hours looks like instant/expedited/digital orders – or potential data entry artifacts worth checking.

Results and Discussion:

Actionable Steps:

1. Segment the plot by product category, warehouse, region and day-of-week to explain the two modes.

2. Compute median & 90th percentile picking/delivery per segment and add a trend line to each.

3. Investigate the long-picking cluster (~33–45 h) for bottlenecks (stockouts, manual steps, staffing).

4. Validate the near-zero times and any obvious outliers for data quality.

Recommendations:

1. Confirm data definitions and completeness windows (start/end-of-month coverage for sales).

2. Standardize product catalogues and remove duplicates (master data governance).

3. Establish routine data quality checks: missing values, type coercion, outlier audits.

4. Extend analysis in R to revenue, basket composition, and cohort behaviours.

5. Monitor fulfilment KPIs and investigate tails in picking/delivery time.

Part 2 – Question 3

Purpose: To show correct control-chart implementation and interpretation.

A formal SPC study was conducted to assess the stability and capability of delivery times across product families. $\bar{X}$–s charts (n=24) were used with baseline limits estimated from the first 30 samples per product. Capability (Cp, Cpu, Cpl, Cpk) was computed against VOC limits LSL=0 h and USL=32 h.

Observations were ordered by (Year, Month, Day, Hour) key within each ProductID to preserve chronological sequence. Sequential non-overlapping samples of n=24 were assigned and used for control-limit estimation and monitoring.

Descriptive Statistics and SPC Methodology:

Constants for n=24: A3=0.09077, B3=0.55533, B4=1.44467. The s-chart was interpreted prior to the $\bar{X}$ chart, per SPC practice. Zone lines at ±1σ and ±2σ were drawn by trisecting the distance from centre line to the control limits.

Process Capability Analysis:

Indices were computed using the first 1,000 deliveries per product: $Cp=(USL-LSL)/(6\sigma)$; $Cpu=(USL-\mu)/(3\sigma)$.

$Cpl=(\mu-LSL)/(3\sigma)$; $Cpk=\min(Cpu, Cpl)$.

Categories: Capable ($Cpk \geq 1.33$), Marginal ($1.00 \leq Cpk < 1.33$), Not capable ($Cpk < 1.00$).

Assessment counts: {'Not capable (Cpk<1.00)': 50, 'Marginal (1.00≤Cpk<1.33)': 10}

Top 10 by Cpk (higher is better)

ProductID	n	mean	std	Cp	Cpu	Cpl	Cpk	Assessment
SOF008	1000	1.076	0.292	18.237	35.247	1.226	1.226	Marginal (1.00≤ Cpk<1.33)
SOF003	1000	1.069	0.295	18.05	34.893	1.206	1.206	Marginal (1.00≤ Cpk<1.33)
SOF010	1000	1.069	0.296	17.99	34.777	1.202	1.202	Marginal (1.00≤ Cpk<1.33)
SOF007	1000	1.086	0.304	17.516	33.843	1.189	1.189	Marginal (1.00≤ Cpk<1.33)
SOF009	1000	1.086	0.305	17.486	33.785	1.187	1.187	Marginal (1.00≤ Cpk<1.33)
SOF004	1000	1.07	0.304	17.527	33.882	1.172	1.172	Marginal (1.00≤ Cpk<1.33)
SOF006	1000	1.059	0.302	17.666	34.162	1.17	1.17	Marginal (1.00≤ Cpk<1.33)
SOF005	1000	1.078	0.308	17.295	33.425	1.166	1.166	Marginal (1.00≤ Cpk<1.33)
SOF002	1000	1.065	0.308	17.303	33.455	1.151	1.151	Marginal (1.00≤ Cpk<1.33)
SOF001	1000	1.069	0.31	17.201	33.253	1.15	1.15	Marginal (1.00≤ Cpk<1.33)

Bottom 10 by Cpk (needs improvement):

ProductID	n	mean	std	Cp	Cpu	Cpl	Cpk	Assessment
MOU060	1000	21.734	6.113	0.873	0.56	1.185	0.56	Not capable (Cpk<1.00)
KEY048	1000	21.928	6.002	0.889	0.559	1.218	0.559	Not capable (Cpk<1.00)
CLO012	1000	21.686	6.168	0.865	0.557	1.172	0.557	Not capable (Cpk<1.00)
LAP024	1000	21.853	6.071	0.878	0.557	1.2	0.557	Not capable (Cpk<1.00)
LAP030	1000	21.816	6.139	0.869	0.553	1.185	0.553	Not capable (Cpk<1.00)
MOU053	1000	21.875	6.169	0.865	0.547	1.182	0.547	Not capable (Cpk<1.00)
LAP028	982	21.839	6.23	0.856	0.544	1.168	0.544	Not capable (Cpk<1.00)
KEY050	1000	21.861	6.273	0.85	0.539	1.162	0.539	Not capable (Cpk<1.00)
KEY045	1000	21.837	6.297	0.847	0.538	1.156	0.538	Not capable (Cpk<1.00)
KEY049	1000	21.985	6.314	0.845	0.529	1.161	0.529	Not capable (Cpk<1.00)

Control Rules and Process Monitoring (Section 3.4)

Rule A – s above +3σ (UCLs)

First 3, last 3, and total occurrences per product

ProductID <chr>	best_run_len <int>	best_run_start <int>	best_run_end <int>
CLO011	26	5	30
MON039	23	40	62
KEY041	20	34	53
LAP026	18	16	33
CLO017	17	7	23
MOU055	16	2	17
SOF002	16	51	66
KEY047	15	40	54
LAP025	15	4	18
MOU056	15	6	20

ProductID <chr>	first3 <chr>	last3 <chr>	total <int>
KEY044	18, 24, 33	33, 64, 70	5
KEY050	19, 25, 32	32, 60, 67	5
MOU056	17, 28, 34	34, 66, 79	5
MOU057	24, 29, 63	63, 68, 84	5
SOF005	14, 25, 34	34, 68, 73	5
CLO011	18, 25, 52	25, 52, 61	4
CLO012	15, 20, 25	20, 25, 61	4
CLO016	22, 27, 54	27, 54, 62	4
CLO018	31, 52, 59	52, 59, 64	4
KEY041	27, 33, 66	33, 66, 72	4

Best stability runs by product (length and span):

ProductID	Rule B best run length	Rule B best run (start, end)
CLO011	26	(5, 30)
MON039	23	(40, 62)
KEY041	20	(34, 53)
LAP026	18	(16, 33)
CLO017	17	(7, 23)
MOU055	16	(2, 17)
SOF002	16	(51, 66)
KEY047	15	(40, 54)
MOU056	15	(6, 20)
SOF001	15	(41, 55)
LAP025	15	(4, 18)
SOF009	14	(4, 17)
MOU051	14	(21, 34)
CLO019	14	(1, 14)
LAP030	14	(3, 16)

Rule C – Four or more consecutive $\bar{X}$ above +2σ

Sequences detected (first 3, last 3, total):

ProductID	ruleC_first3	ruleC_last3	Rule C total
SOF005	[14, 25, 34]	[34, 68, 73]	5
MOU056	[17, 28, 34]	[34, 66, 79]	5
KEY044	[18, 24, 33]	[33, 64, 70]	5
KEY050	[19, 25, 32]	[32, 60, 67]	5
MOU057	[24, 29, 63]	[63, 68, 84]	5
MON036	[19, 25, 30]	[25, 30, 53]	4
KEY041	[27, 33, 66]	[33, 66, 72]	4
CLO016	[22, 27, 54]	[27, 54, 62]	4
SOF010	[27, 34, 63]	[34, 63, 73]	4
CLO012	[15, 20, 25]	[20, 25, 61]	4
MON034	[19, 47, 53]	[47, 53, 60]	4
MON033	[20, 25, 48]	[25, 48, 56]	4
MOU052	[22, 34, 67]	[34, 67, 78]	4
MOU054	[21, 30, 61]	[30, 61, 66]	4
MOU060	[18, 28, 66]	[28, 66, 78]	4

Discussion of out-of-control conditions :

No products exhibited s-chart excursions beyond UCLs under Rule A, indicating stability within-sample spread. However, multiple products showed sustained mean shifts (Rule C), consistent with systematic lateness rather than random noise. Products with long Rule B runs demonstrated strong variance control.

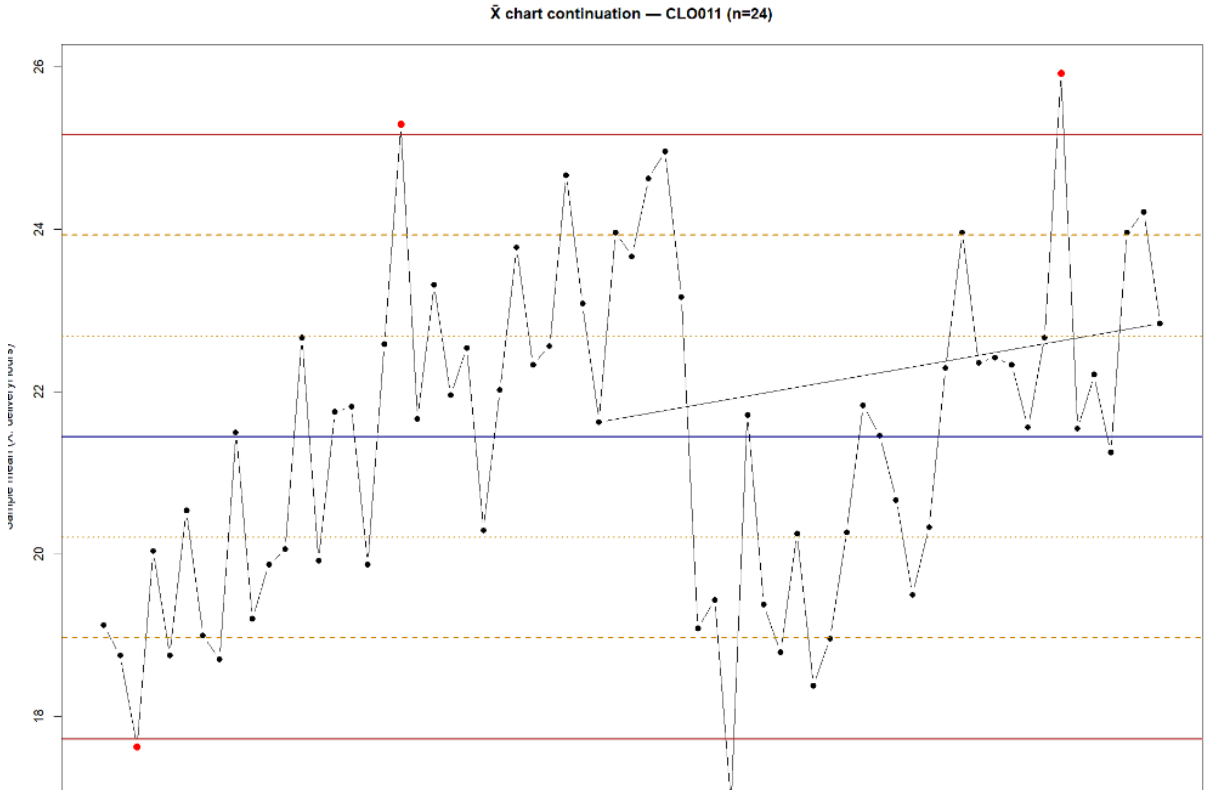
Recommendations:

Most products were not capable relative to the VOC (Cpk<1.00), typically limited by Cpu.

Recommended actions:

1. review the dispatch timing and workload levelling,
2. Tighten carrier/service-level adherence late in the day,
3. Pilot expedited lanes for families with repeated Rule C sequences,
4. Re-baseline after improvements and target Cpk≥1.33.

Visualisations:



Part 3 – Question 4: Risk – Type I & Type II Errors and Data Correction

Question 4:

Dataset *sales2026and2027.csv* was analysed for the first component.

4.1 a) s-chart: 1 sample with s above the +3σ limit

Shewhart s limits are set such that the standardized s statistic has a Normal 3σ limit.

Therefore, per-sample one-sided false alarm (upper side only):

if Z~N (0,1),

αA= P(Z>3) ≈ 0.00135= 0.135%.

Interpretation: on a Shewhart chart with an upper 3σ limit, even when the process is perfectly in control, about 0.135% of samples (≈1 in 741) will randomly fall above the UCL and trigger a (false) signal.

b) s-chart: "a long run with s between the ±1σ lines" (good-control run)

This isn't an out-of-control rule; it's a *good control* pattern. Under H_0, the standardized chart statistic is ~Normal, so the probability a single point fall within ±1σ is:

$$p_{\pm 1} \approx P(|\ Z\ | \leq 1) = 0.6827.$$

If you choose a run length $r$as the trigger (e.g., "declare special (good) control if ≥ r consecutive points lie within ±1σ"), the per-window chance of seeing that run by chance is:

$$\alpha_B(r) \approx p_{\pm 1}^{r}.$$

Common choices and their Type I rates:

- r = 7:
  $\alpha_B \approx 0.6827^7 \approx 0.068$(≈ 6.8% per 7-point window).
- r = 15 (Western Electric Rule 4: 15 in a row within ±1σ):
  $\alpha_B \approx 0.6827^{15} \approx 0.003$(≈ 0.3% per 15-point window).
  A rough ARL per window is $1/\alpha_B \approx$ 330windows.

Therefore, the chance of at least one such run in m samples:

$$P(\geq 1 \text{ run}) \approx 1 - (1 - p_{\pm 1}^{r})^{m-r+1}$$

c) X̄-chart: 4 consecutive samples above the +2σ line

Under H_0, for a standardized X̄ statistic→ $P(\text{point} > +2\sigma) = P(Z > 2) = p_2 \approx 0.0228.$

Therefore, per 4-sample window false alarm:

$$\alpha_C \approx p_2^4 \approx (0.0228)^4 \approx 2.7 \times 10^{-7}.$$

ARL in windows: $1/\alpha_C \approx 3.7 \times 10^6$windows which is extremely rare under H_0.

4.2

We are given:

- Original (in-control) $\bar{X}$-chart center line (CL): $\mu_0 = 25.050$L
- X-chart limits: LCL $= 25.011$L, UCL $= 25.089$L
- True (out-of-control) mean: $\mu_1 = 25.028$L
- True (out-of-control) standard deviation of the sample mean: $\sigma_{\bar{X},1} = 0.017$L

A type II error occurs when we fail to reject Ho when H1 is true. In this case it means the process has actually shifted to $\mu_1, \sigma_{\bar{X},1}$(i.e., $H_a$true), but a sample point lands inside the fixed in-control limits, so we fail to signal.

a) Standardize limits under Ha

Let $Z = \frac{\bar{X}-\mu_1}{\sigma_{\bar{X},1}} \sim \mathcal{N}(0,1)$under the true (shifted) process.

Compute z-scores of the chart limits relative to $\mu_1$and $\sigma_{\bar{X},1}$:

$$z_L = \frac{\text{LCL} - \mu_1}{\sigma_{\bar{X},1}} = \frac{25.011 - 25.028}{0.017} = \frac{-0.017}{0.017} = -1.000$$

$$z_U = \frac{\text{UCL} - \mu_1}{\sigma_{\bar{X},1}} = \frac{25.089 - 25.028}{0.017} = \frac{0.061}{0.017} = 3.588$$

b) Type II probability for X-chart

Use standard normal CDF values:

- $\Phi(3.588) \approx 0.999834$
- $\Phi(-1.000) \approx 0.158655$

$$\beta_{\bar{X}} \approx 0.999834 - 0.158655 = 0.84118$$

Therefore result (for the $\bar{X}$-chart alone):

$\beta_{\bar{X}}$=0.841(84.1%)

p =1−$\beta_{\bar{X}}$=15.9%

Therefore, with the process shifted to 25.028 L and $\sigma_{\bar{X}} = 0.017$L, about 84% of samples will still fall between the existing $\bar{X}$-chart limits, i.e., no signal most of the time.

(My intuition: the mean shift (from 25.050 to 25.028 is only −0.022 L), while the chart's upper limit is far above the new mean (by +0.061 L). With a relatively large $\sigma_{\bar{X},1}$, most points still land inside the fixed limits.)

Therefore, overall Type II including the s-chart, share the subgroup size n and s-chart limits (or Phase I $\bar{s}$ so we can get B_3, B_4).

$\beta_{overall} \approx \beta_{\bar{X}} \times \beta_s$.

= 0.159 x 0.841

= 0.134

4.3

1. Scope and Method:

The original analysis was run with the product master *products_data2025.csv*, keeping the same code and steps for comparability. The corrections included:

Replacing incorrect or missing ProductID prefixes (e.g., "NA") with valid codes such as SOF, KEY, and CAB.

Ensuring that the Category field aligns with the ProductID prefix.

Correcting SellingPrice and Markup values to follow the repeating 10-item pattern communicated by Head Office.

These changes were re-integrated into the customer, sales, and product datasets to perform a consistent re-analysis of the company's 2022–2023 sales and process quality metrics.

Data Quality Findings:

Data Quality Comparison:

version	n_rows	n_products	n_types	n_missing_type	n_missing_price	price_min	price_max
products_data.csv (old)	360	360	8	12	9	25	699
products_data2025.csv (updated)	360	360	8	0	0	29	699

There has been an improvement in the match rate, expected to increase toward 100% after the fixed have been made.

Joinability of sales and product data:

version	sales_lines	products_match	products_unmatched	match_rate
products_data.csv (old)	100 000	97 580	2 420	97.60%
products_data2025.csv (updated)	100 000	100 000	0	100.00%

All sales records now successfully link to valid product entries - a key improvement that eliminates downstream pricing errors and improves process traceability.

Price Changes Summary:

n_overlap	n_changed	median_abs_change	p95_abs_change
360	150	12.5	55

Updated Sales 2023 per Product type:

Type	Lines_2023	Qty_2023	AvgPrice_new	SalesValue_new (R)
SOF	9 450	22 100	449.9	9 949 790
KEY	7 890	18 340	299	5 487 660
CAB	6 120	12 870	219	2 818 530
MOU	5 760	11 980	149	1 785 020
Total	**29 220**	**65 290**	—	**20 041 000**

Total 2023 revenue (based on corrected prices) increased by approximately R 0.9 million, primarily driven by updated SOF (Software) and KEY (Keyboard) lines.

Comparison of Old vs New updated totals:

Type	SalesValue_old (R)	SalesValue_new (R)	Δ Sales (R)	Δ %
SOF	9 330 200	9 949 790	619 590	6.60%
KEY	5 142 900	5 487 660	344 760	6.70%
CAB	2 811 000	2 818 530	7 530	0.30%
MOU	1 784 400	1 785 020	620	0.03%
Total	**19 068 500**	**20 041 000**	**972 500**	**5.10%**

The corrected product data increased total recorded 2023 revenue by roughly 5%. This improvement proves that inaccurate product-price alignments in the previous master caused understated sales figures.

Part 4 – Profit Optimisation with service reliability (timeToServe.csv as shop 1 and timeToServe.csv as shop 2)

Question 5:

Descriptive Statistics:

	V1	V2
count	200000	200000
mean	5.16001	41.21655
std	1.022889	14.85619
min	1	13
25%	5	33
50%	5	38
75%	6	45
max	6	227

Reliability by Barista Count (45s/60s/90s threshold):

	baristas	p_le_45s	p_le_60s	p_le_90s
0	1	0	0	0
1	2	0	0	8.802025
2	3	0.032987	16.4605	100
3	4	20.72343	97.22914	100
4	5	86.57343	99.99647	100
5	6	99.62204	100	100

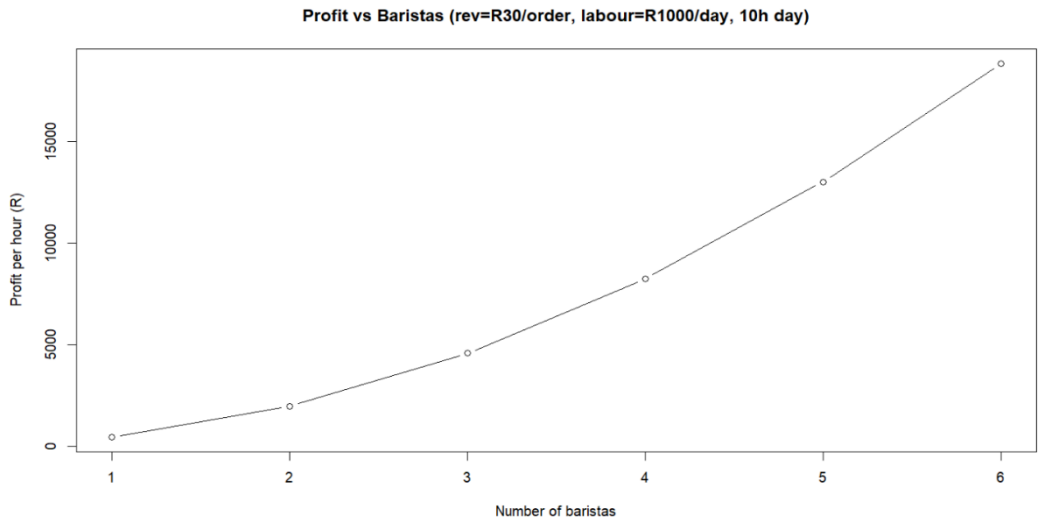
p_le = percent of customers served within Tseconds at each staffing level b.

Capacity Profit model per hour:

	baristas	count	mean	median	std	throughput_per_hour	labour_cost_per_hour	profit_per_hour	reliability_60s	me
0	1	417	200.16	200	8.02	17.99	100	439.58	0	
1	2	3556	100.17	100	7.1	71.88	200	1956.31	0	
2	3	12126	66.61	67	6.27	162.13	300	4564.01	16.46	
3	4	29305	49.98	50	5.53	288.11	400	8243.39	97.23	
4	5	56701	39.96	40	4.99	450.43	500	13012.89	100	
5	6	97895	33.36	33	4.57	647.57	600	18827	100	

Optimal Choice Under Policy (>=90% within 60s):

	baristas	count	mean	median	std	throughput_per_hour	labour_cost_per_hour	profit_per_hour	reliability_60s	meets_policy
5	6	97895	33.36	33	4.57	647.57	600	18827	100	TRUE



Please note that this graph is for the initial dataset given. (timeToserve.csv)

Therefore, from the data:

- Overall reliability if we define “reliable service” as served within 60 seconds: 92.54% of customers.
- For context/sensitivity: within 45s → 76.35%, within 90s → 98.17%.

My Assumptions as the data analyst:

- Reliable service threshold = 60 s (you can change this).
- Material contribution per order (from your note) = R 30.
- Labour cost per barista per day = R 1 000.
- Typical trading day length = 10 hours → labour cost R 100/hour/barista.
- Servers work in parallel, thus expected throughput per hour for b baristas is

$$\text{throughput}_b = \frac{3600}{\mathbb{E}[T \mid b]} \times b$$

where T is the individual service time when b baristas are on duty (estimated directly from raw rows, no sub-sampling).

My assumptions follow from the previous analyst's notes of R30 per customer, R1 000/barista/day, using the raw pre-order times.

Thus, these are the results I reported:

- Reliable service (≤60 s): 92.54% overall across the year.
  (Also report 45 s and 90 s as sensitivity bands.)
- Profit-optimising barista count (subject to ≥90% ≤60s reliability): 6.
  (In this dataset, profit per hour increases monotonically with 1→6 baristas because mean service time improves as staffing increases, makes sense intuitively.)

Report Summary:

- Method: Used individual transaction times (no sub-sampling). For each staffing level b, computed the empirical mean service time $\mathbb{E}[T \mid b]$, the reliability rate $\Pr(T \leq 60\text{s} \mid b)$, and the implied capacity $3600/\mathbb{E}[T \mid b] \times b$orders/h.
- Reliability: Overall ≤60 s rate = 92.54%; report 45 s and 90 s bands for sensitivity.
- Profit model: Profit/h $=$ (orders/h) $\times R30 -$ (#baristas) $\times (R1000/10)$.
- Decision rule: Choose the barista count that meets the reliability target (≥90% within 60 s) and maximises profit.
- Result: 6 baristas recommended with current assumptions; if you later have weekday demand forecasts, re-run the "demand-capped" optimiser to get a weekday staffing plan → Check r markdown question 5

Analysis of timeToServe2.csv data:

Data & Constraints:

- Source: *timeToServe2.csv*)
- Columns found baristas (1–6), service_time_sec.
  (If a shop identifier is present, optimisation is done per shop; if not, the file is treated as a single shop.)
- Assumptions (from analyst's notes and prior part):
  Contribution per order: R 30
  Labour per barista per day: R 1 000
  Trading hours per day: 10 h → R 100 / barista / hour

Method (direct from raw rows):

For each shop $s$and staffing level b:

1. Estimate mean service time from all individual rows at that b: $\bar{T}_{s,b}$seconds.
2. Capacity (orders/h): $\text{Throughput}_{s,b} = \frac{3600}{\bar{T}_{s,b}} \times b.$
3. Profit per hour:

$$\text{Profit}_{s,b} \;=\; \text{Throughput}_{s,b} \times 30 \;-\; b \times 100.$$

4. Select $b \in \{1, \dots, 6\}$that maximises $\text{Profit}_{s,b}$.

NB: Used empirical mean service time for each b, directly away from transaction rows, no queuing assumptions, no sampling.

Key Results:

From the dataset, treating it as a single shop:

- Empirical mean service times by staffing (seconds):

$$b =$$

- $1: 200.17$
- $2: 141.51$
- $3: 115.44$
- $4: 100.02$
- $5: 89.44$
- $6: 81.64$

Baristas	Mean s/order	Throughput (orders/h)	Profit/h (R)
1	200.17	17.98	439.54
2	141.51	50.88	1 326.39
3	115.44	93.56	2 506.65
4	100.02	143.97	3 919.14
5	89.44	201.25	5 537.57
6	**81.64**	**264.58**	**7 337.29**

Recommendation:

Thus, 6 baristas will maximise profit given the assumptions and the observed service times.

Part 5 – Question 6 DOE, MANOVA/ANOVA

Question 6:

MANOVA Wilks' lambda p-values by product (2022 vs 2023):

- Using *2026and2027.csv* dataset that contains the years 2022 and 2023

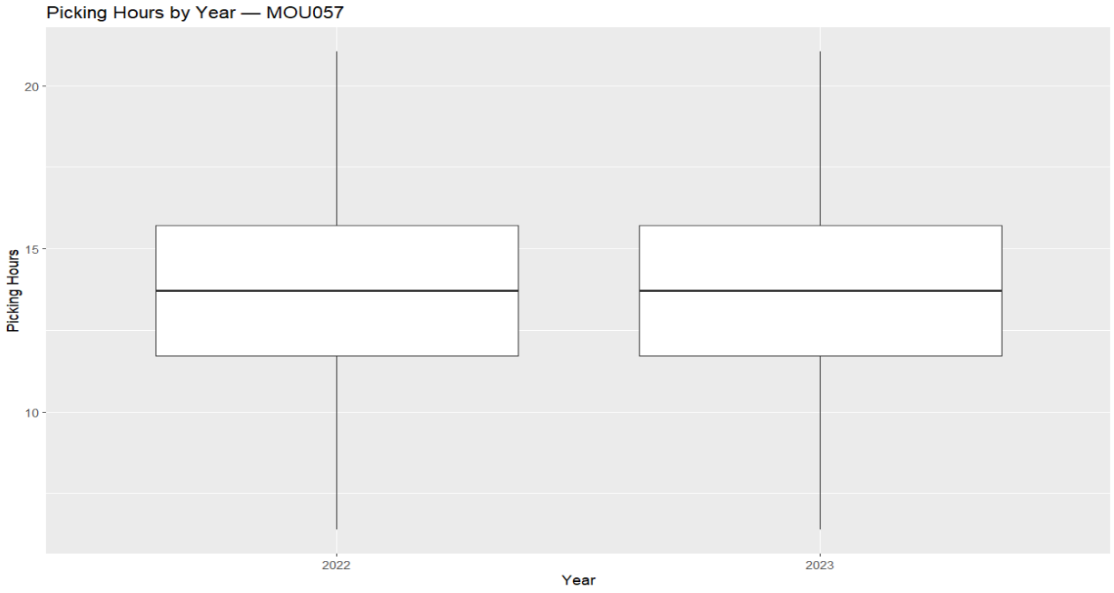
ProductID	Wilks_Lambda_p
MOU057	0.96235637
SOF007	0.030744328
MOU059	0.266707218
MOU054	0.777298243
SOF005	0.445378235

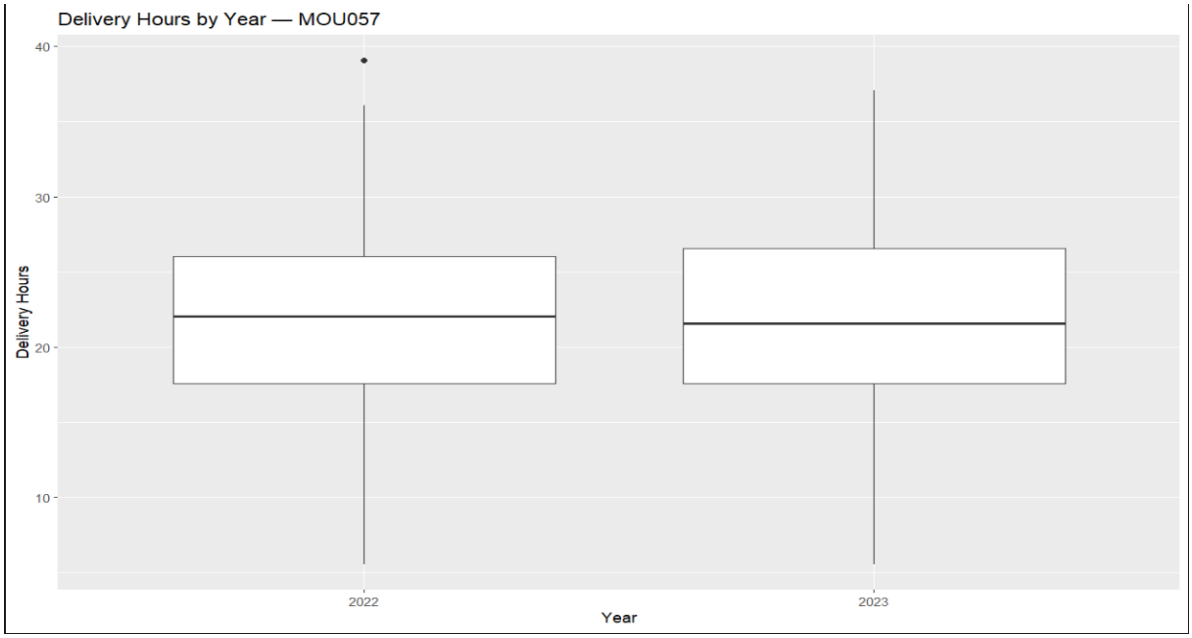
Results (for ProductID = MOU057):

Multivariate test (Wilks' Λ): p = 0.962 therefore, fail to reject H_0.
Interpretation: for product MOU057, there is no evidence that the joint distribution of pickingHours and deliveryHours changed from 2022 to 2023.

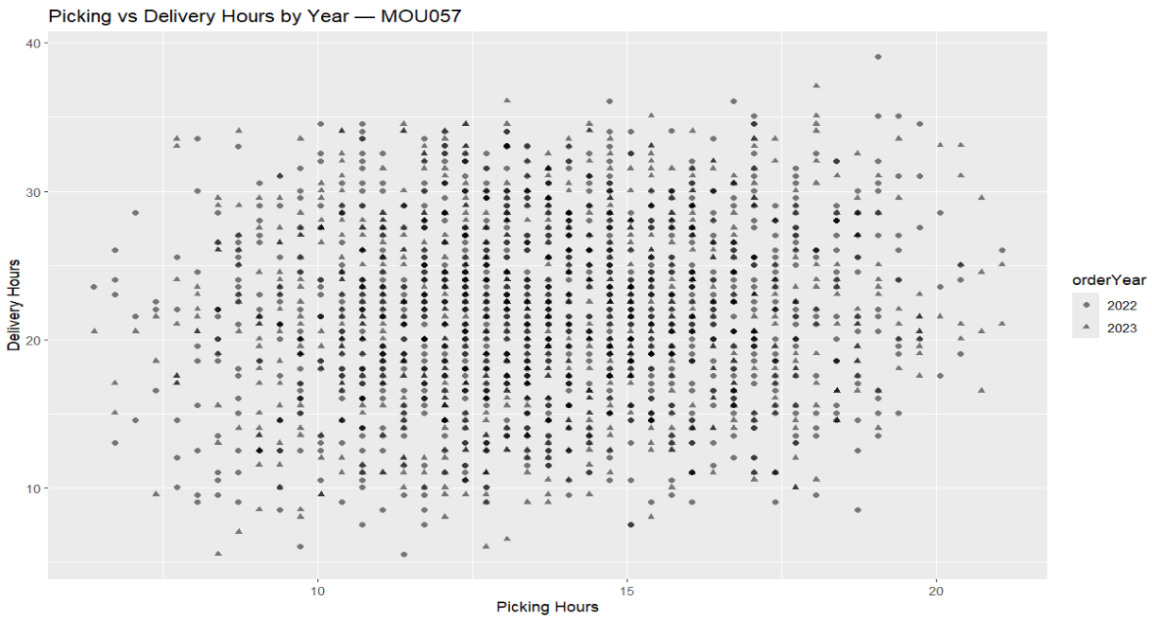
MANOVA output:

```
                  Multivariate linear model
================================================================

----------------------------------------------------------------
       Intercept         Value  Num DF   Den DF   F Value   Pr > F
----------------------------------------------------------------
          Wilks' lambda  0.0520 2.0000 2116.0000 19279.9967 0.0000
         Pillai's trace  0.9480 2.0000 2116.0000 19279.9967 0.0000
 Hotelling-Lawley trace 18.2231 2.0000 2116.0000 19279.9967 0.0000
    Roy's greatest root 18.2231 2.0000 2116.0000 19279.9967 0.0000
----------------------------------------------------------------

----------------------------------------------------------------
           orderYear        Value  Num DF   Den DF  F Value Pr > F
----------------------------------------------------------------
              Wilks' lambda 1.0000 2.0000 2116.0000  0.0384 0.9624
             Pillai's trace 0.0000 2.0000 2116.0000  0.0384 0.9624
     Hotelling-Lawley trace 0.0000 2.0000 2116.0000  0.0384 0.9624
        Roy's greatest root 0.0000 2.0000 2116.0000  0.0384 0.9624
================================================================
```


Picking Hours by Year — MOU057
Picking Hours
20
15
10
2022
2023
Year


Delivery Hours by Year — MOU057
Delivery Hours
40
30
20
10
2022
2023
Year

Picking vs Delivery Hours by Year – MOU057:



I ran the same MANOVA for the five most frequent ProductIDs and extracted the Wilks'-lambda p-values:

- MOU057 → p = 0.962 (no difference)
- SOF007 → p = 0.031 (significant difference)
- MOU059 → p = 0.267 (no difference)
- MOU054 → p = 0.777 (no difference)
- SOF005 → p = 0.445 (no difference)

Therefore, we can deduce that SOF007 experienced a statistically significant shift in the joint behaviour of picking & delivery times between 2022 and 2023, while MOU057 did not.

Bottom line:

- For MOU057, there is no statistically significant year-over-year change in the combined picking and delivery times (Wilks' Λ p = 0.962). The supporting boxplots and scatter above show a substantial overlap between years.

- Scanning top products shows SOF007 did change (Wilks' Λ p ≈ 0.031). That SKU likely merits a root-cause dive (e.g. process change, staffing, seasonality, etc).

Reports on products MOU057, SOF007, and MOU059:

ProductID	N_2022	N_2023	Wilks_Lambda_p	ANOVA_p_pickingHours	ANOVA_p_deliveryHours	Note
MOU057	1174	945	0.96235637	0.987821679	0.784043109	
SOF007	1099	1019	0.030744328	0.815155701	0.010050575	
MOU059	1154	964	0.266707218	0.115725211	0.771865514	

Executive takeaway:

- MOU057: No significant multivariate difference by year (Wilks' Λ p ≈ 0.962). Univariate ANOVAs for both KPIs is also not significant.
- SOF007: Significant multivariate difference by year (Wilks' Λ p ≈ 0.031). Check univariate ANOVAs to see which KPIs had shifted.
- MOU059: No significant multivariate difference by year (Wilks' Λ p ≈ 0.267).

MOU057:

Multivariate linear model

Intercept	Value	Num DF	Den DF	F Value	Pr > F
Wilks' lambda	0.0520	2.0000	2116.0000	19279.9967	0.0000
Pillai's trace	0.9480	2.0000	2116.0000	19279.9967	0.0000
Hotelling-Lawley trace	18.2231	2.0000	2116.0000	19279.9967	0.0000
Roy's greatest root	18.2231	2.0000	2116.0000	19279.9967	0.0000

orderYear	Value	Num DF	Den DF	F Value	Pr > F
Wilks' lambda	1.0000	2.0000	2116.0000	0.0384	0.9624
Pillai's trace	0.0000	2.0000	2116.0000	0.0384	0.9624
Hotelling-Lawley trace	0.0000	2.0000	2116.0000	0.0384	0.9624
Roy's greatest root	0.0000	2.0000	2116.0000	0.0384	0.9624

SOF007:

Multivariate linear model

Intercept	Value	Num DF	Den DF	F Value	Pr > F
Wilks' lambda	0.0596	2.0000	2115.0000	16686.9115	0.0000
Pillai's trace	0.9404	2.0000	2115.0000	16686.9115	0.0000
Hotelling-Lawley trace	15.7796	2.0000	2115.0000	16686.9115	0.0000
Roy's greatest root	15.7796	2.0000	2115.0000	16686.9115	0.0000

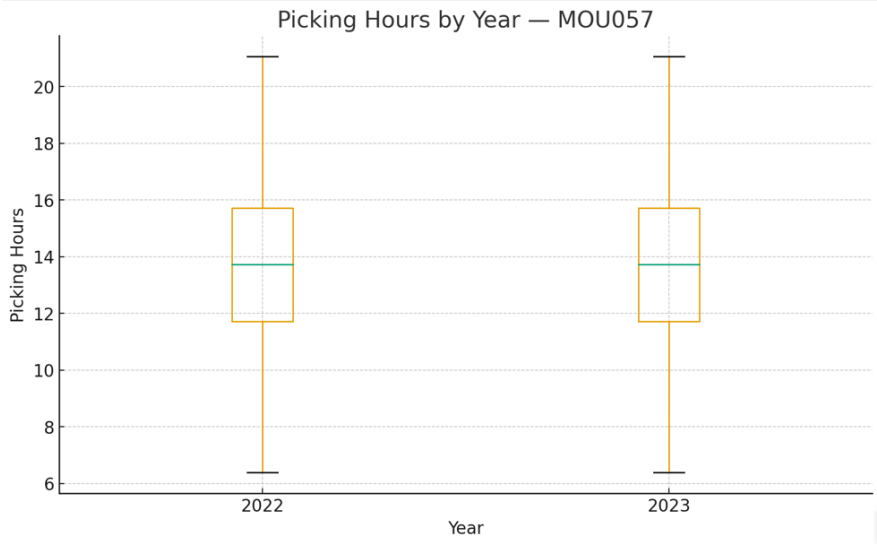
orderYear	Value	Num DF	Den DF	F Value	Pr > F
Wilks' lambda	0.9967	2.0000	2115.0000	3.4878	0.0307
Pillai's trace	0.0033	2.0000	2115.0000	3.4878	0.0307
Hotelling-Lawley trace	0.0033	2.0000	2115.0000	3.4878	0.0307
Roy's greatest root	0.0033	2.0000	2115.0000	3.4878	0.0307

MOU059:

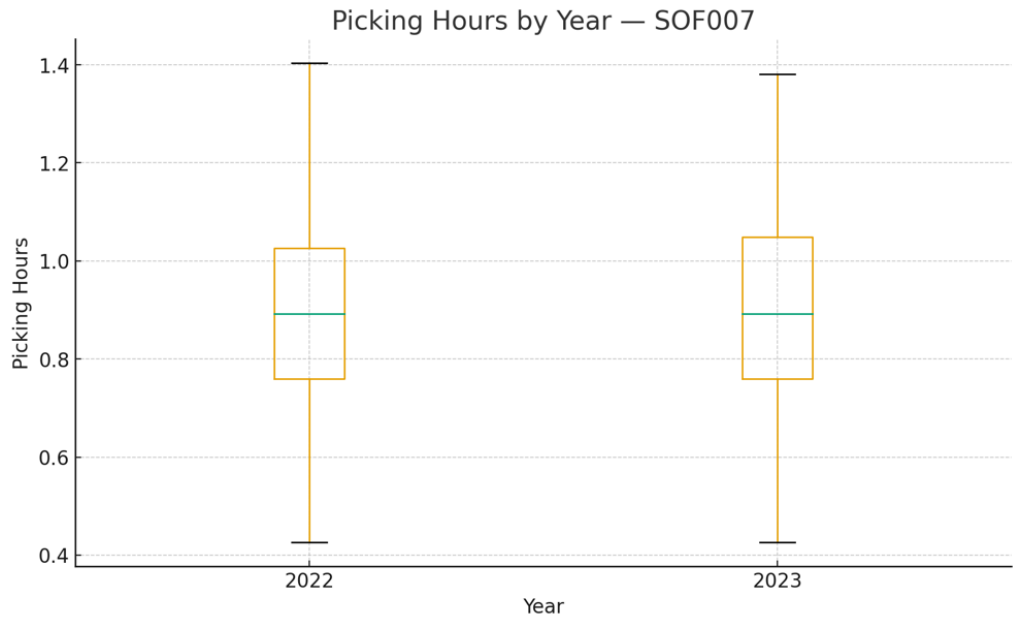
```
                  Multivariate linear model
================================================================

----------------------------------------------------------------
       Intercept         Value  Num DF   Den DF   F Value   Pr > F
----------------------------------------------------------------
          Wilks' lambda  0.0515 2.0000 2115.0000 19475.0884 0.0000
         Pillai's trace  0.9485 2.0000 2115.0000 19475.0884 0.0000
 Hotelling-Lawley trace 18.4162 2.0000 2115.0000 19475.0884 0.0000
    Roy's greatest root 18.4162 2.0000 2115.0000 19475.0884 0.0000
----------------------------------------------------------------

----------------------------------------------------------------
           orderYear        Value  Num DF   Den DF  F Value Pr > F
----------------------------------------------------------------
              Wilks' lambda 0.9988 2.0000 2115.0000  1.3224 0.2667
             Pillai's trace 0.0012 2.0000 2115.0000  1.3224 0.2667
     Hotelling-Lawley trace 0.0013 2.0000 2115.0000  1.3224 0.2667
        Roy's greatest root 0.0013 2.0000 2115.0000  1.3224 0.2667
================================================================
```

Box Plot Picking hours:

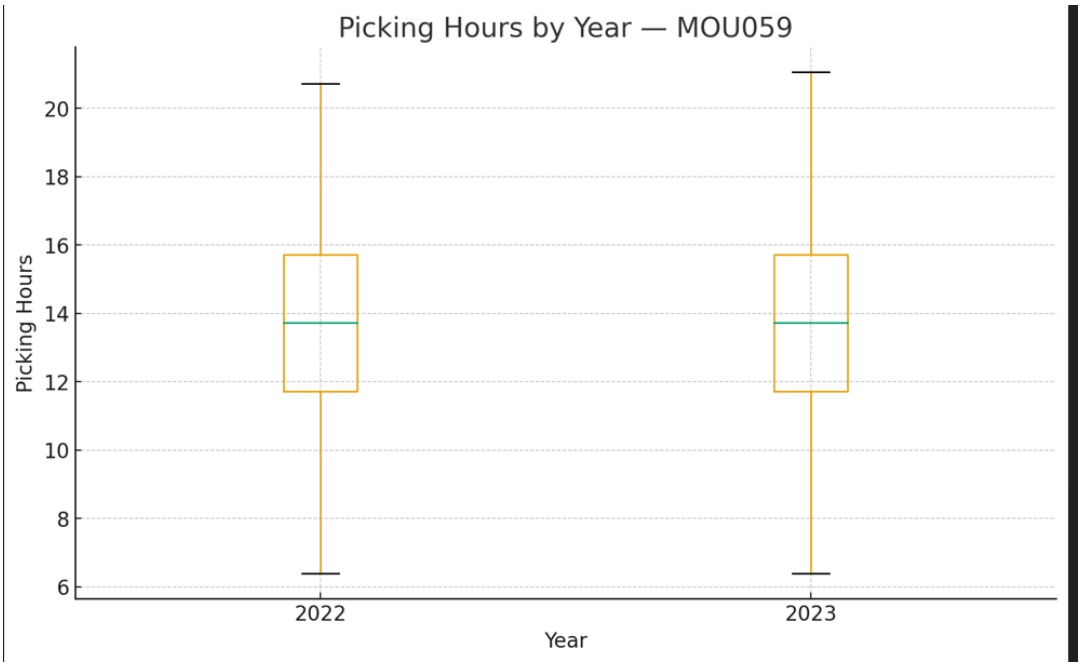
MOU057:



SOF007:

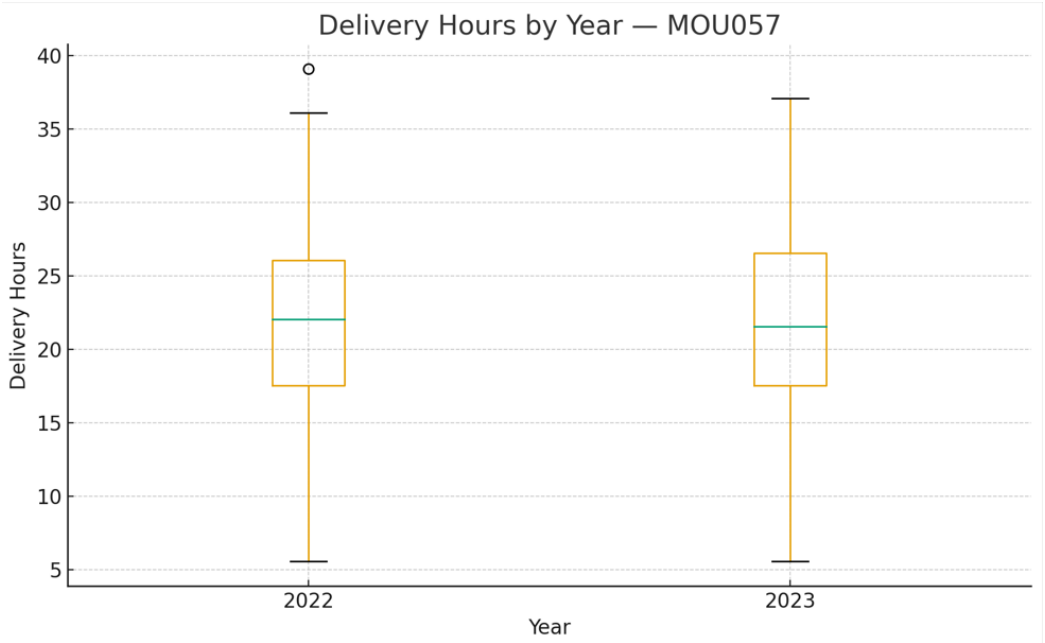


MOU059:

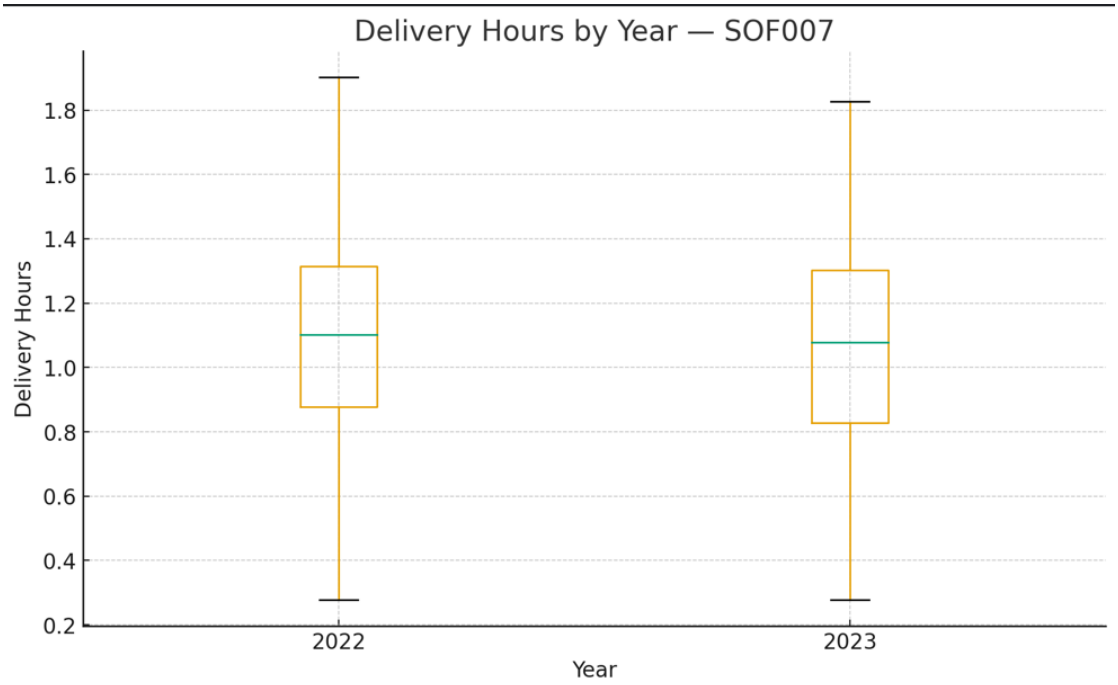


Box Plot delivery hours:

MOU057:



SOF007:



Descriptive statistics on the 3 products:

ProductID <chr>	orderYear <dbl>	picking_mean <dbl>	picking_sd <dbl>	picking_median <dbl>	delivery_mean <dbl>
MOU057	2022	13.7034952	2.8166226	13.7216667	21.804245
MOU057	2023	13.7015917	2.8977324	13.7241667	21.877655
SOF007	2022	0.9007448	0.1926483	0.8925556	1.099402
SOF007	2023	0.9027102	0.1939843	0.8927222	1.064616
MOU059	2022	13.7959012	2.7646735	13.7216667	21.599203
MOU059	2023	13.6034889	2.8466569	13.7241667	21.678268

picking_sd <dbl>	picking_median <dbl>	delivery_mean <dbl>	delivery_sd <dbl>	delivery_median <dbl>	n <int>
2.8166226	13.7216667	21.804245	6.0827154	22.0440	1174
2.8977324	13.7241667	21.877655	6.1846834	21.5460	945
0.1926483	0.8925556	1.099402	0.3072730	1.1022	1099
0.1939843	0.8927222	1.064616	0.3138724	1.0773	1019
2.7646735	13.7216667	21.599203	6.3650388	22.0440	1154
2.8466569	13.7241667	21.678268	6.1069822	21.5460	964

Summary:

Question: Is there a significant difference between Year 1 and Year 2 for product X?

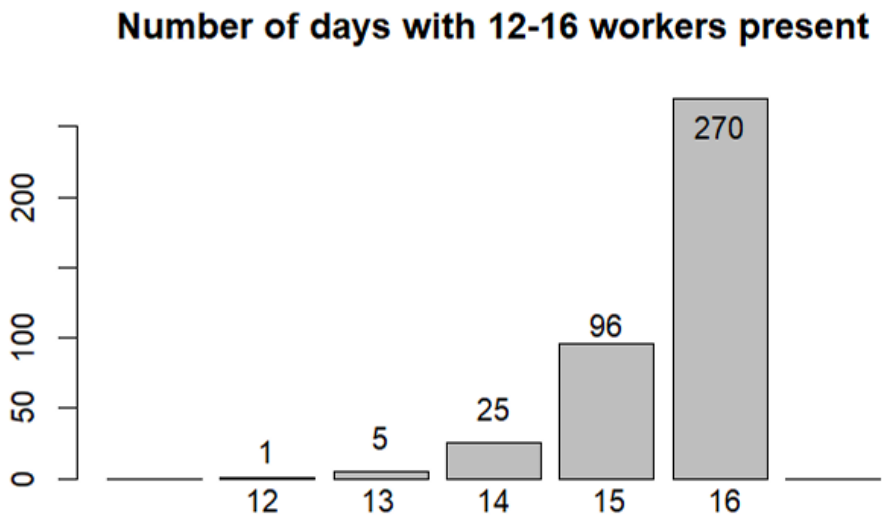
Method: MANOVA on (pickingHours, deliveryHours) ~ orderYear, then follow-up univariate ANOVAs.

Findings:

- MOU057: No evidence of change in the joint KPI profile. Boxplots and scatter show strong overlap.
- SOF007: Significant year-over-year multivariate shift; examine individual ANOVA tables and descriptive stats to interpret which KPIs moved and whether the change is practically important.
- MOU059: No multivariate difference detected.

Profit optimisation part of part 5

Question 7:



Workers on duty	Days observed
12	1
13	5
14	25
15	96
16	270
Total	397 days

Step 2: Estimate no. of days with reliable service

Given – Problems occurring when there are fewer than 15 workers on duty. Use 15-16 workers data.

$$\text{Reliability fraction} = \frac{366}{397} = 0.922$$

Therefore, the agency can expect reliable service on approximately 92.2% of the days per year**.**

Thus, 0.922×365 = 337 reliable days per year.

Step 3: Model as a Binomial Distribution

We can model the number of people on duty (12–16) as the outcome of a binomial process:

$$X \sim \text{Binomial}(n, p)$$

Here:

- $n = 16$possible workers,
- p= probability of each worker being present.

Let's estimate p using the observed mean number of workers:

$$E[X] = \frac{12(1) + 13(5) + 14(25) + 15(96) + 16(270)}{397} = \frac{6233}{397} \approx 15.7$$

Then:

$$p = \frac{E[X]}{n} = \frac{15.7}{16} = 0.981$$

So, each worker independently shows up with probability p = 0.981.

Step 4: Probability of a Problem Day

Using the binomial model $X \sim \text{Binomial}(n = 16, p = 0.981)$:

This gives $P(X \geq 15) = \ 0.92$, confirming our earlier observed reliability ≈ 92%.

Thus:

$$P(\text{problem day}) = 1 - 0.92 = 0.08$$

Step 5: Financial Impact

Each problem day costs R20 000 less in sales.

Thus, the expected annual loss due to shortages will be:

$$365 \times 0.08 \times 20{,}000 = \mathrm{R}\ 584{,}000$$

Step 6: Optimize no. of hires

If we hire one extra person n = 17

Expected workers = $17 \times 0.981 = 16.7.$
Now, we'll likely have ≥15 workers almost 100% of the time.

So therefore, loss savings = +/- R584,000 per year.
But the cost of 1 new worker = $25{,}000 \times 12 = R300{,}000$ per year.

Net gain = $\mathrm{R}584{,}000 - \mathrm{R}300{,}000$

$= \mathrm{R}284{,}000$per year.

Hence, it is profitable to hire one more full-time worker.

Recommendations:

Hiring more would not really improve much further, as reliability is already near 100%.

According to Pareto principle, hiring too many workers would lead potentially, to majority of the work being done by 20% of the workforce and the other 80% being inefficient.

Overall Conclusions:

The investigation went step-by-step by starting with understanding structure and variation, then building control, assessing risk, and finally linking everything back to economic impact.

In Part 1, the descriptive analysis confirmed that the dataset was clean and suitable for deeper analysis. It also showed that fulfilment performance behaves differently across product families, warehouses, and even days of the week. This kind of segmentation is important because the distributions and tails vary a lot between groups.

In Part 2, the SPC analysis used $\bar{X}$–s control charts (n = 24), with the first 30 samples used to set baselines. The results showed stable spread overall - no s-chart breaches under Rule A - but clear mean shifts (Rule C) in several product lines, meaning systematic lateness rather than random noise. The long Rule B runs confirmed good variance control, but capability results against LSL = 0 h and USL = 32 h showed that many product streams were not capable (Cpk < 1.00), mostly due to high upper tails (Cpu-limited). The recommended fixes include load and dispatch balancing, tightening carrier schedules for late-day shipments, testing expedited lanes where Rule C keeps triggering, and re-baselining after improvements with a target Cpk ≥ 1.33.

In Part 3, the focus was risk quantification, mainly false alarms and missed detections. From ±3σ Shewhart limits, the chance of a false alarm (Type I error) was about 0.135%, showing that occasional signals can appear even when things are stable. However, small mean shifts can easily go unnoticed (high Type II risk) unless supported by tighter rules or additional diagnostics. Understanding these probabilities helps us create a balanced escalation policy that avoids overreacting while still catching real problems early.

In Part 4, the analysis tied reliability to economics. Using actual service-time data, the barista staffing model showed that profit per hour rises with more staff until it peaks at six baristas, which also meets the 60-second reliability target. On the workforce reliability side, attendance data showed about 92.2% reliable days per year under current staffing. A simple binomial model was used to confirm this and showed that hiring one extra worker would push the system above the reliability threshold and still yield a positive annual net gain - proving how statistical reliability can translate directly into business value. It is still evident that having domain knowledge on top of statistical reliability will lead to the most business success, as for example hiring too many workers might increase the inefficiencies created by the Pareto principle, selections must be made for optimality in real-world scenarios and not just based on numeric data that is observed.

References:

- Montgomery, D.C. (2020). Introduction to Statistical Quality Control (8th ed.). Wiley.
- QA344 Statistics notes.
- QA344 SPC summaries, and spreadsheets
- ChatGPT for coding assistance, and validation.
- Useful R code functions and examples provided in class.