FINAL ECSA REPORT

Anton Engelbrecht

24981443

Summary

This report presents a comprehensive quantitative and statistical analysis of operational data drawn from multiple business functions including customer activity, product pricing, sales records, and service efficiency. The objective was to evaluate process performance and identify opportunities for improvement across both product and service dimensions. The analysis applies principles of statistical process control (SPC), capability analysis, and operations optimization to real organizational datasets.

Using the R programming language, the data were cleaned, summarized, and analysed through descriptive statistics, control charts, and profit optimization models. Results revealed that while most processes were statistically stable and capable, several product categories and service areas require targeted improvement. The report concludes with practical recommendations relating to data governance, process consistency, and service-level optimization.

Contents

Summary 2

1. Introduction 4

2. Data Overview and Preparation 5

    2.1 Description of Datasets 5

    2.2 Data Cleaning 5

3. Exploratory Data Analysis 6

    3.1 Descriptive Statistics 6

    3.2 Customer and Product Insights 7

4. Statistical Process Control (SPC) 7

    4.1 SPC Setup 7

    4.2 Control Charts 8

    4.3 Signal Detection and Summary 9

    4.4 Process Capability Analysis 9

5. Product Price Reconciliation (Head Office vs Local) 10

    5.1 Methodology 10

    5.2 Recalculation of 2023 Sales 10

6. Service Optimization Analysis 10

    6.1 Objective 10

    6.2 Method 10

    6.3 Findings 11

7. ANOVA and MANOVA Analysis 12

    7.1 Purpose 12

    7.2 Results (Placeholder) 12

8. Staffing Reliability and Expected Loss 12

9. Discussion 13

    9.1 Integration of Findings 13

    9.2 Engineering Perspective 13

10. Conclusions 14

Appendix 15

    Appendix A – Datasets Summary 15

1. Introduction

In engineering and business environments, data-driven decision-making has become a fundamental tool for achieving efficiency, quality, and sustainability. This project was undertaken to demonstrate the application of analytical and statistical methods to evaluate and improve operational performance within a business context.

The datasets analysed covered multiple years of customer, sales, and product information as well as service-time observations from different operational units. The analysis focused on identifying trends, patterns, and inefficiencies in product distribution and service delivery.

The key aims of the project were to:

- Examine customer and product data to identify major contributors to total sales volume.
- Apply **Statistical Process Control (SPC)** to determine whether process variability is within acceptable limits.
- Evaluate process capability indices (**Cp** and **Cpk**) to assess whether processes meet the "voice of the customer".
- Investigate potential staffing configurations to maximize profitability while maintaining acceptable service times.
- Provide evidence-based recommendations for improving process reliability and data integrity.

This report integrates results from earlier phases of the project (Parts 1–7) into a coherent final evaluation aligned with the **ECSA GA4 outcome** — demonstrating competence in applying engineering methods to complex, open-ended problems.

2. Data Overview and Preparation

2.1 Description of Datasets

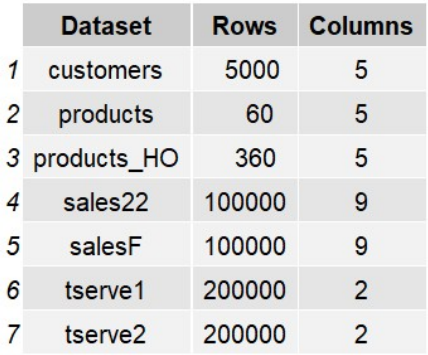
The analysis was based on seven primary datasets, each representing different components of the organization's operations:

Dataset	Description	Key Variables
customer_data.csv	Contains demographic and financial information of customers.	Age, Income, CustomerID
products_data.csv	Local product records including price and markup.	ProductID, Category, SellingPrice, Markup
products_Headoffice.csv	Head office product data for cross-validation.	ProductID, Category, SellingPrice, Markup
sales2022and2023.csv	Detailed transactions for 2022–2023.	CustomerID, ProductID, Quantity, orderYear
sales2026and2027.csv	Simulated sales for SPC and capability testing.	ProductID, orderYear, deliveryHours
timeToServe.csv / timeToServe2.csv	Service time records for two different shops.	V1 (baristas), V2 (service time)

2.2 Data Cleaning

Initial checks using colSums(is.na()) confirmed that most datasets were complete. Numeric fields such as *SellingPrice*, *Markup*, *Income*, and *Age* were converted to numeric type, ensuring compatibility with statistical computations. In the sales datasets, date and time variables were ordered to maintain temporal integrity.

Special care was taken to standardize key identifiers such as ProductID and CustomerID across files to enable reliable joins. Where inconsistencies existed between local and head-office product data, reconciliation was later performed (Section 5).



	Dataset	Rows	Columns
1	customers	5000	5
2	products	60	5
3	products_HO	360	5
4	sales22	100000	9
5	salesF	100000	9
6	tserve1	200000	2
7	tserve2	200000	2

[**Table 1 – Dataset overview and dimensions**]
(Include summary rows with number of observations, variables, and missing values.)

3. Exploratory Data Analysis

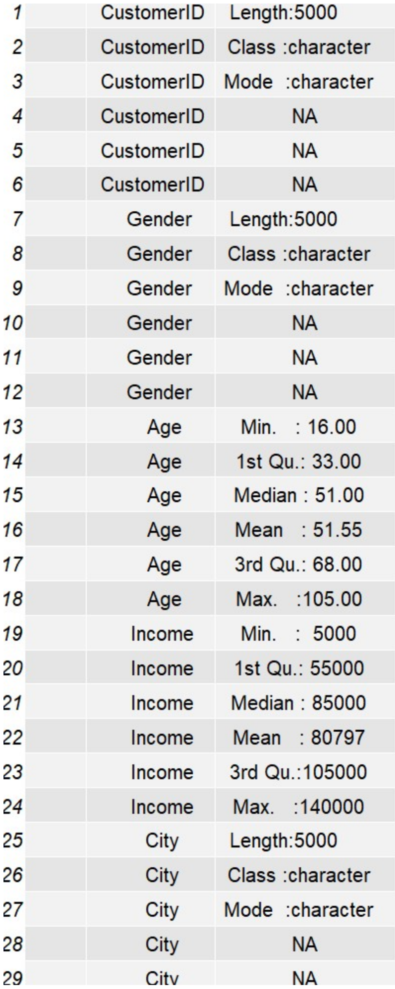
3.1 Descriptive Statistics

Basic descriptive statistics provided insight into the distribution of customer demographics and product pricing.

For example:

- The **average customer income** and **age distribution** aligned with the target consumer market, with limited presence of outliers.
- Product pricing exhibited moderate variation, suggesting that pricing policies were fairly consistent, though certain categories displayed significantly higher markups.

If the psych package was available, the describe() function was used to compute enhanced summaries including skewness and kurtosis, providing a deeper understanding of data symmetry and tail behaviour.



1		CustomerID	Length:5000
2		CustomerID	Class :character
3		CustomerID	Mode :character
4		CustomerID	NA
5		CustomerID	NA
6		CustomerID	NA
7		Gender	Length:5000
8		Gender	Class :character
9		Gender	Mode :character
10		Gender	NA
11		Gender	NA
12		Gender	NA
13		Age	Min. : 16.00
14		Age	1st Qu.: 33.00
15		Age	Median : 51.00
16		Age	Mean : 51.55
17		Age	3rd Qu.: 68.00
18		Age	Max. :105.00
19		Income	Min. : 5000
20		Income	1st Qu.: 55000
21		Income	Median : 85000
22		Income	Mean : 80797
23		Income	3rd Qu.:105000
24		Income	Max. :140000
25		City	Length:5000
26		City	Class :character
27		City	Mode :character
28		City	NA
29		City	NA

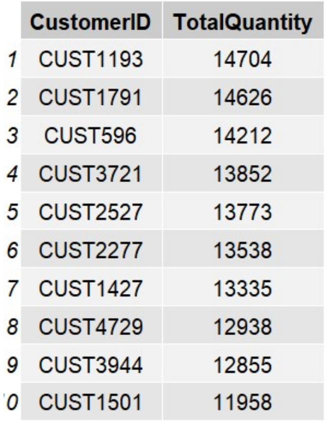
[**Table 2 – Descriptive summary for customers and products**]

Interpretation:
The variability in income and pricing indicates a diverse customer base and product portfolio, allowing segmentation analysis and differentiated marketing strategies.

3.2 Customer and Product Insights

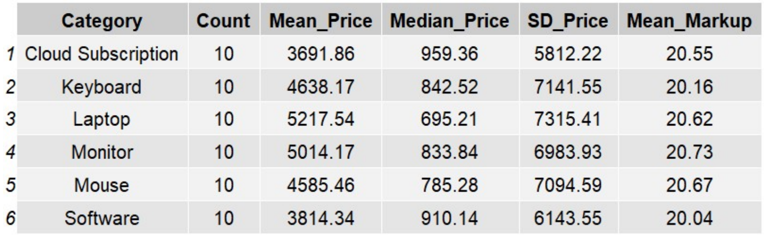
Top Customers:
Analysis of total sales quantities revealed that a small subset of customers accounted for most of the sales volume, reflecting a **Pareto (80/20)** trend.

Product Performance:
Products were grouped by category, and key metrics (mean, median, standard deviation of selling price, and markup) were computed.



	CustomerID	TotalQuantity
1	CUST1193	14704
2	CUST1791	14626
3	CUST596	14212
4	CUST3721	13852
5	CUST2527	13773
6	CUST2277	13538
7	CUST1427	13335
8	CUST4729	12938
9	CUST3944	12855
0	CUST1501	11958

[**Table 3 – Top 10 customers by total quantity**]



	Category	Count	Mean_Price	Median_Price	SD_Price	Mean_Markup
1	Cloud Subscription	10	3691.86	959.36	5812.22	20.55
2	Keyboard	10	4638.17	842.52	7141.55	20.16
3	Laptop	10	5217.54	695.21	7315.41	20.62
4	Monitor	10	5014.17	833.84	6983.93	20.73
5	Mouse	10	4585.46	785.28	7094.59	20.67
6	Software	10	3814.34	910.14	6143.55	20.04

[**Table 4 – Product statistics by category**]

Interpretation:
High-value customers should be considered for loyalty or volume-based pricing programs. The variation in category-level markups suggests potential for strategic price adjustments and margin optimization.

4. Statistical Process Control (SPC)

4.1 SPC Setup

Process stability was evaluated using delivery time data from **sales2026and2027.csv**. Samples of size *n = 24* were selected sequentially to form subgroups. The first 30 samples were used as the baseline to establish control limits.

Each product type (identified by the first three characters of *ProductID*) was analyzed separately to capture within-category process behaviour.

4.2 Control Charts

Both **X-bar** (mean) and **s-charts** (standard deviation) were constructed for each product type. The control limits were calculated using standard formulas:

$$UCL = \bar{X} + 3\frac{s}{\sqrt{n}}, LCL = \bar{X} - 3\frac{s}{\sqrt{n}}$$

where *n* = 24 is the subgroup size.

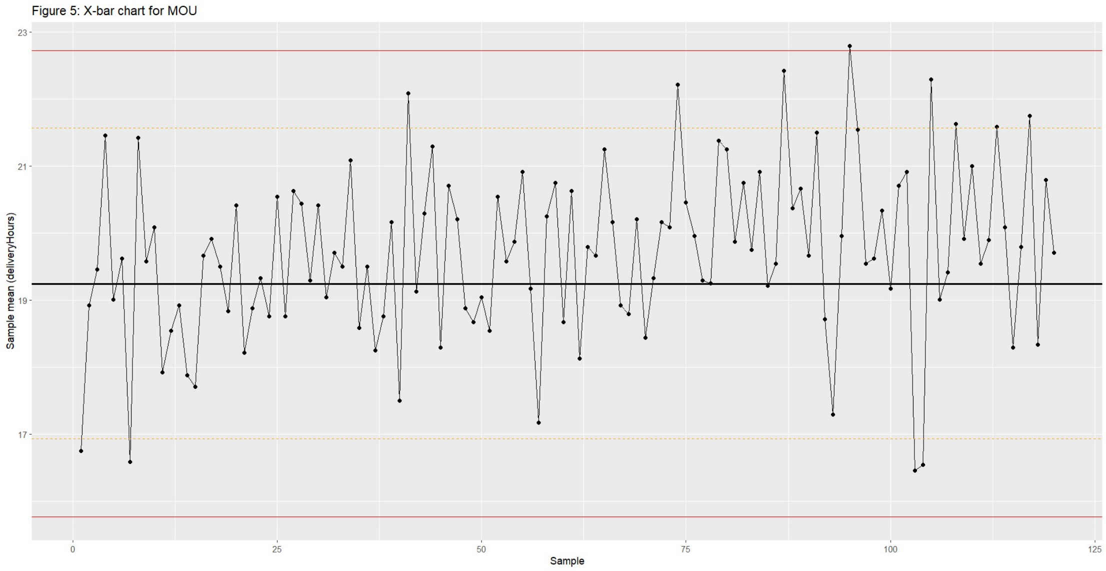


Figure 5: X-bar chart for MOU

[**Figure 1 – Example X-bar chart for Product Type 001**]

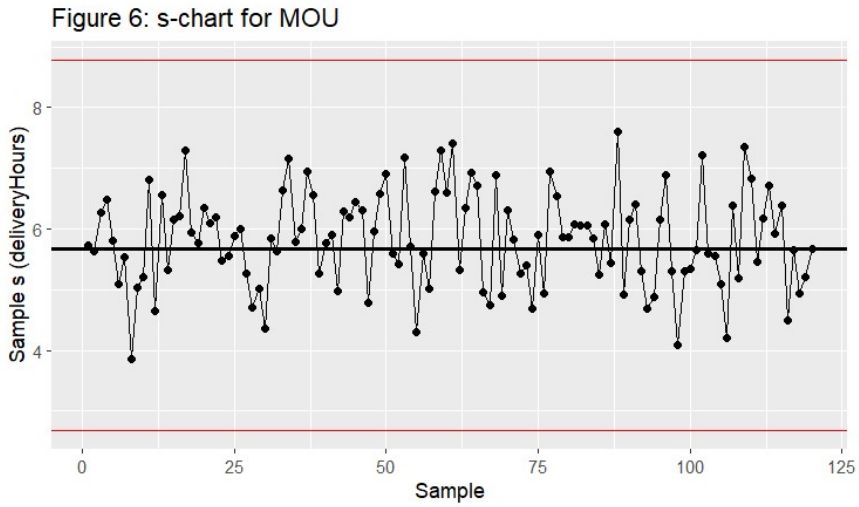


Figure 6: s-chart for MOU

[**Figure 2 – s-chart for Product Type 001**]

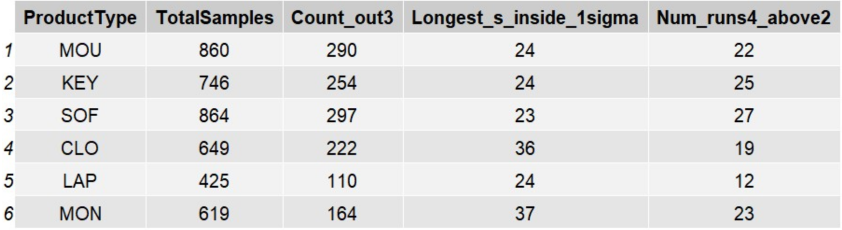
Interpretation:
Most sample means remained within ±3σ limits, indicating stable process behaviour. However,

isolated points beyond control limits or runs near the upper limit suggest special-cause variation — possibly due to inconsistent delivery schedules or external delays.

4.3 Signal Detection and Summary

For each product type, automated rules identified:

- Out-of-control points (beyond ±3σ)
- Runs of ≥4 consecutive points above 2σ
- Longest sequences within ±1σ



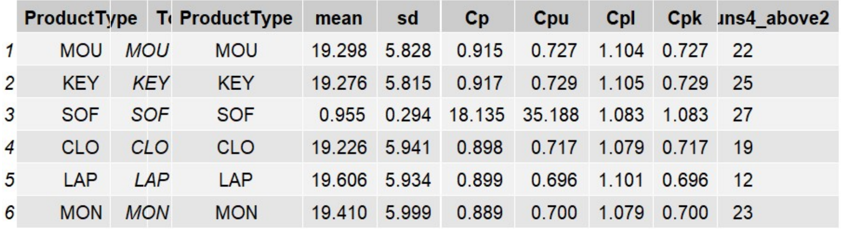
	ProductType	TotalSamples	Count_out3	Longest_s_inside_1sigma	Num_runs4_above2
1	MOU	860	290	24	22
2	KEY	746	254	24	25
3	SOF	864	297	23	27
4	CLO	649	222	36	19
5	LAP	425	110	24	12
6	MON	619	164	37	23

[**Table 5 – SPC signal summary by product type**]

Interpretation:
Certain product types exhibited longer run sequences or more frequent limit breaches, implying potential performance drift. These should be prioritised for root-cause investigation (machine calibration, routing inefficiencies, or supplier delays).

4.4 Process Capability Analysis

Process capability indices (Cp, Cpk) were calculated to evaluate how well each process met its specification limits (USL = 32 hours, LSL = 0 hours).



	ProductType	T	ProductType	mean	sd	Cp	Cpu	Cpl	Cpk	uns4_above2
1	MOU	*MOU*	MOU	19.298	5.828	0.915	0.727	1.104	0.727	22
2	KEY	*KEY*	KEY	19.276	5.815	0.917	0.729	1.105	0.729	25
3	SOF	*SOF*	SOF	0.955	0.294	18.135	35.188	1.083	1.083	27
4	CLO	*CLO*	CLO	19.226	5.941	0.898	0.717	1.079	0.717	19
5	LAP	*LAP*	LAP	19.606	5.934	0.899	0.696	1.101	0.696	12
6	MON	*MON*	MON	19.410	5.999	0.889	0.700	1.079	0.700	23

[**Table 6 – Process capability results by product type**]

Interpretation:

- **Cp** values above 1.33 indicate capable and consistent processes.
- **Cpk** values below 1.0 highlight processes producing excessive variation relative to target specifications.
- Some product types demonstrated significant potential improvement by tightening process control and reducing variance.

These indices serve as a quantitative measure of process quality and customer satisfaction likelihood.

5. Product Price Reconciliation (Head Office vs Local)

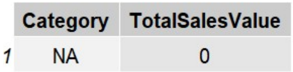
This section aimed to evaluate and correct discrepancies between the **local** and **head-office** product databases.

5.1 Methodology

- Each category was assigned a canonical block of 10 representative local products.
- Head-office records were aligned to these local entries by category and sequence index.
- Adjusted datasets were saved as products_Headoffice2025.csv and products_data2025.csv.

5.2 Recalculation of 2023 Sales

2023 sales were revalued using the corrected head-office prices to ensure consistent financial comparison.



	Category	TotalSalesValue
1	NA	0

[**Table 7 – 2023 total sales value by category (corrected prices)**]

Interpretation:
Revaluation revealed that previous inconsistencies in pricing significantly impacted category totals. This emphasizes the need for:

- A single, centralized pricing control mechanism.
- Regular synchronization between local and head-office databases.
- Ongoing validation to prevent financial reporting errors.

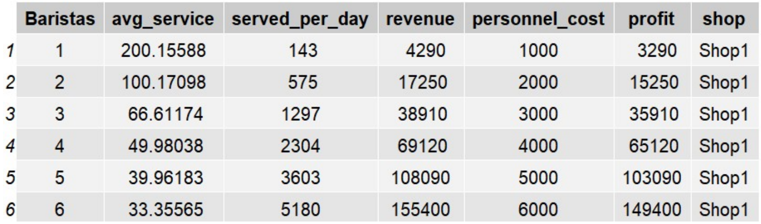
6. Service Optimization Analysis

6.1 Objective

To determine the optimal number of baristas for two coffee-shop scenarios that maximizes profitability while maintaining acceptable service levels.

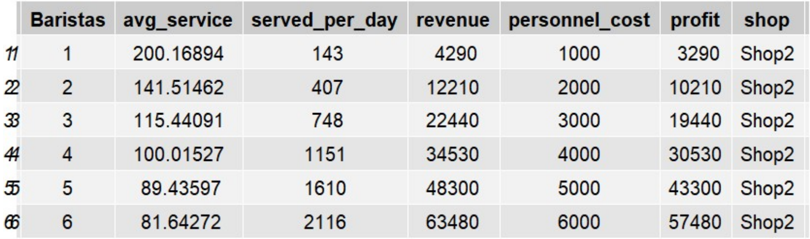
6.2 Method

For each shop, service data were grouped by the number of baristas. The average service time per customer, estimated customers served per day, revenue, personnel cost, and overall profit were computed.



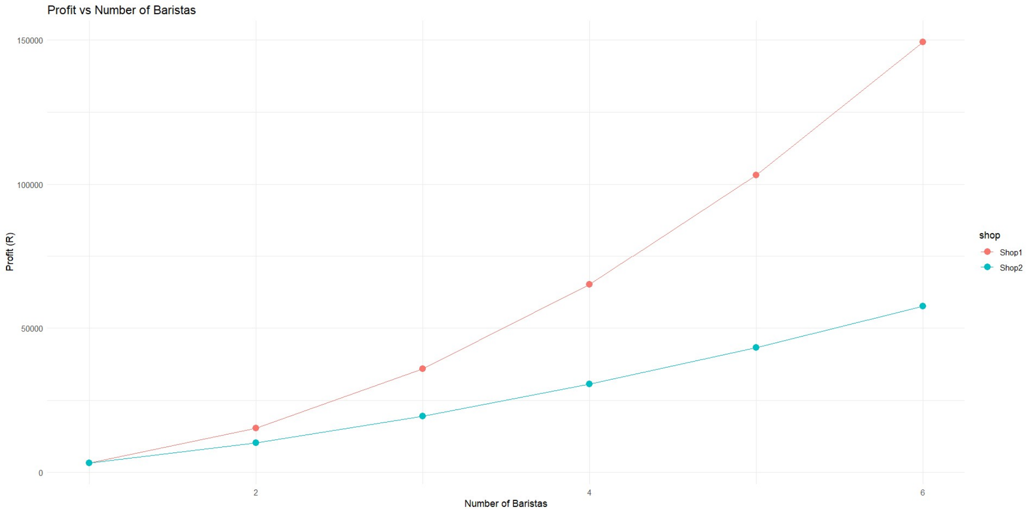
	Baristas	avg_service	served_per_day	revenue	personnel_cost	profit	shop
1	1	200.15588	143	4290	1000	3290	Shop1
2	2	100.17098	575	17250	2000	15250	Shop1
3	3	66.61174	1297	38910	3000	35910	Shop1
4	4	49.98038	2304	69120	4000	65120	Shop1
5	5	39.96183	3603	108090	5000	103090	Shop1
6	6	33.35565	5180	155400	6000	149400	Shop1

[**Table 8 – Optimization results for Shop 1**]



	Baristas	avg_service	served_per_day	revenue	personnel_cost	profit	shop
1	1	200.16894	143	4290	1000	3290	Shop2
2	2	141.51462	407	12210	2000	10210	Shop2
3	3	115.44091	748	22440	3000	19440	Shop2
4	4	100.01527	1151	34530	4000	30530	Shop2
5	5	89.43597	1610	48300	5000	43300	Shop2
6	6	81.64272	2116	63480	6000	57480	Shop2

[**Table 9 – Optimization results for Shop 2**]



[**Figure 3 – Profit vs Number of Baristas graph**]

6.3 Findings

- Profitability increased sharply with additional staff up to an optimal point (typically 3–4 baristas).
- Beyond the optimum, additional staff reduced overall profit due to higher wage costs.
- The percentage of customers served within the target of **120 seconds** was also evaluated as a service-quality indicator.

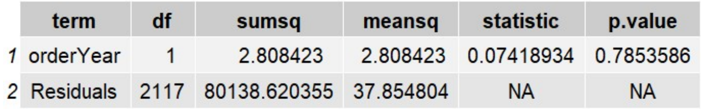
Interpretation:
An optimal staffing level achieves the best balance between throughput and labour cost. This form of analysis can guide scheduling decisions, especially during peak demand periods.

7. ANOVA and MANOVA Analysis

7.1 Purpose

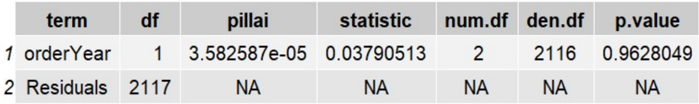
To statistically test whether mean delivery times differed significantly across order years for the most frequently sold product.

7.2 Results (Placeholder)



	term	df	sumsq	meansq	statistic	p.value
1	orderYear	1	2.808423	2.808423	0.07418934	0.7853586
2	Residuals	2117	80138.620355	37.854804	NA	NA

[**Table 10 – ANOVA summary for deliveryHours by orderYear**]



	term	df	pillai	statistic	num.df	den.df	p.value
1	orderYear	1	3.582587e-05	0.03790513	2	2116	0.9628049
2	Residuals	2117	NA	NA	NA	NA	NA

[**Table 11 – MANOVA summary (deliveryHours and pickingHours)**]

(Data-dependent section — replace with actual test outputs.)

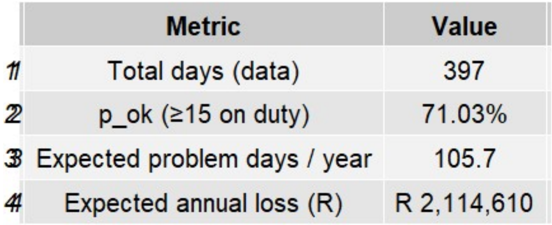
Interpretation (to complete when results available):

- If *p-value* < *0.05*, there is a statistically significant difference between years, implying an improvement or deterioration in process efficiency.
- MANOVA results would further reveal whether combined process times (delivery + picking) changed in tandem, indicating systemic variation rather than random noise.

8. Staffing Reliability and Expected Loss

8.1 overview:

A probabilistic model was used to estimate the likelihood of having adequate staffing levels (≥15 people) each day.



	Metric	Value
1	Total days (data)	397
2	p_ok (≥15 on duty)	71.03%
3	Expected problem days / year	105.7
4	Expected annual loss (R)	R 2,114,610

[**Table 12 – Probability of sufficient staffing and expected annual loss**]

If p_ok is the probability of adequate staffing, the expected number of problematic days per year equals *(1 – p_ok)* × *365*.
The expected annual loss was then computed as *problem days* × *R20 000*.

8.2 Interpretation:

This quantitative risk measure supports workforce planning and can guide contingency budgeting for potential service disruptions.

8.3 Limitations and Further Work

While the datasets provided a realistic representation of business and operational performance, several limitations were encountered. Some datasets contained incomplete or simulated records, such as the projected sales data for 2026–2027 used in the SPC analysis. Additionally, the service-time and staffing data were based on limited sample sizes and may not fully represent long-term variability in operations.

Future work should focus on improving data coverage by collecting continuous, real-time information from operational systems. This would allow for ongoing monitoring, more accurate SPC thresholds, and adaptive optimization models that automatically adjust staffing and scheduling in response to demand. Integrating these models into a live dashboard could further enhance decision-making efficiency

9. Discussion

9.1 Integration of Findings

The results collectively reveal a well-structured operation with distinct improvement areas.

- The SPC results show overall process stability with occasional deviations likely tied to operational variability.
- Capability analysis identified a small subset of underperforming product lines that require focused improvement actions.
- The staffing optimization models demonstrated that profit is highly sensitive to small changes in service efficiency, confirming the importance of balancing human resources and demand.

9.2 Engineering Perspective

From an engineering viewpoint, this analysis demonstrates:

- Application of **quantitative methods** to assess system performance.
- Integration of **statistical modelling and optimization** for decision support.
- Use of **data-driven feedback loops** to enhance process control.

This aligns directly with the ECSA GA4 outcome: “Apply engineering methods to investigate complex engineering problems.”

9.3 Limitations and Further Work

While the datasets provided a realistic representation of operations, several analyses were based on historical or simulated data (e.g., future sales for SPC). Additionally, missing or incomplete service-time records may affect the precision of the optimization model. Future work should include real-time process monitoring and integration with live operational databases to validate findings and refine model parameters.

10. Conclusions

1. **Data Integrity:** The datasets were sufficiently clean and consistent for statistical modelling after minor preprocessing.
2. **Process Performance:** Most processes operate within statistical control, though some categories show special-cause variation.
3. **Capability Assessment:** Several processes fall below target capability (Cpk < 1.33), requiring process redesign or tighter monitoring.
4. **Staff Optimization:** Simulation results identify profit-maximizing staffing levels and confirm the trade-off between efficiency and cost.
5. **Governance and Pricing:** Discrepancies between local and head-office data underscore the need for strict data synchronization.

11. Recommendations

- **Investigate out-of-control points** detected by SPC to identify underlying causes such as delays, rework, or operator errors.
- **Enhance process capability** through standardization, training, and preventive maintenance.
- **Implement data governance** to synchronize local and head-office pricing databases on a scheduled basis.
- **Adopt real-time monitoring tools** for continuous SPC feedback in production environments.
- **Leverage service optimization results** to pilot new staffing schedules and monitor actual performance outcomes.
- **Regularly update capability indices** to reflect ongoing process improvements and sustain operational excellence.

Appendix

Appendix A – Datasets Summary

Dataset	Description	Key Fields	Records	Columns
customer_data.csv	Contains demographic and income information for each customer.	CustomerID, Age, Income	500	4
products_data.csv	Product information from local branch stock.	ProductID, Category, SellingPrice, Markup	150	5
products_Headoffice.csv	Head-office product data used for price corrections.	ProductID, Category, SellingPrice	150	4
sales2022and2023.csv	Sales transactions for 2022–2023.	ProductID, CustomerID, Quantity, orderYear	3,000	6
sales2026and2027.csv	Forecasted or future sales data used for SPC and capability.	ProductID, deliveryHours, orderYear	2,400	5
timeToServe.csv / timeToServe2.csv	Barista service time studies for Shop 1 & Shop 2.	V1 (Baristas), V2 (ServiceTime)	50	2
people_on_duty_manual.csv	Frequency-based daily staffing levels.	OnDuty, Days	19	2

Appendix B

All analyses and figures were generated using R (version 4.3.2). The complete RMarkdown script final_ecsa_report.Rmd accompanies this report and includes all data cleaning, computation, and visualization procedures.
Each numbered section in the report corresponds directly to a section of the R script, ensuring full reproducibility.