ECSA Term Project – Basic Data Analysis Report

Department of Industrial Engineering, Stellenbosch University

Student: Owen Wagenaar

Date: 20 October 2025

Table of Contents

Contents

1. Executive Summary

This report provides a structured, multi-stage analysis of the organization's sales, operations, and service processes, progressing from descriptive analytics to advanced statistical decision making.

Week 1: established the commercial baseline, analyzing 2022–2023 sales, customers, and products. The findings revealed strong concentration of revenue in a few product categories and customers, as well as clear seasonal peaks. These insights highlight both dependency risks and opportunities to diversify the product mix and broaden the customer base.

Week 2: introduced statistical process control (SPC) to evaluate production and delivery consistency. Control charts and process capability indices demonstrated that while processes are broadly stable, occasional special-cause variation requires monitoring and corrective action. This stage showed the value of real-time quality management.

Week 3: applied experimental design (DOE) and hypothesis testing to staffing levels in a service environment. ANOVA and Fisher's LSD confirmed that service times differ significantly with staffing, and economic modelling identified optimal workforce configurations that balance profitability with reliability against service-level agreements (SLAs).

Overall, the report demonstrates that commercial, operational, and staffing levels are interconnected. Revenue growth depends not only on product strategy but also on stable processes and efficient staffing. By integrating descriptive, statistical, and prescriptive analytics, the company is positioned to enhance profitability, improve customer satisfaction, and embed a culture of evidence-based decision making.

2. Introduction & Context

Objective: As a new data analyst, establish a baseline understanding of sales performance and customer/product mix, and identify immediate opportunities for further investigation. This document follows Stellenbosch University academic report conventions: structured sections, numbered headings, labelled figures/tables, and formal tone.

3. Data and Methodology

Datasets utilised:

• customers: 5,000 rows, 5 columns (CustomerID, Gender, Age, Income, City).

• products: 60 rows, 5 columns (ProductID, Category, Description, SellingPrice, Markup).

• products_Headoffice: 360 rows, 5 columns (used for cross-checks).

• sales: 100,000 rows, 9 columns (CustomerID, ProductID, Quantity, orderTime/Day/Month/Year, pickingHours, deliveryHours).

Method summary: We joined sales to products and customers via ProductID and CustomerID. A timestamp was built from order components. Revenue was calculated as Quantity × SellingPrice. Aggregations were produced by month, category, product, and customer. Operational metrics (picking/delivery hours) were summarised descriptively.

4. Results

4.1 Revenue Over Time

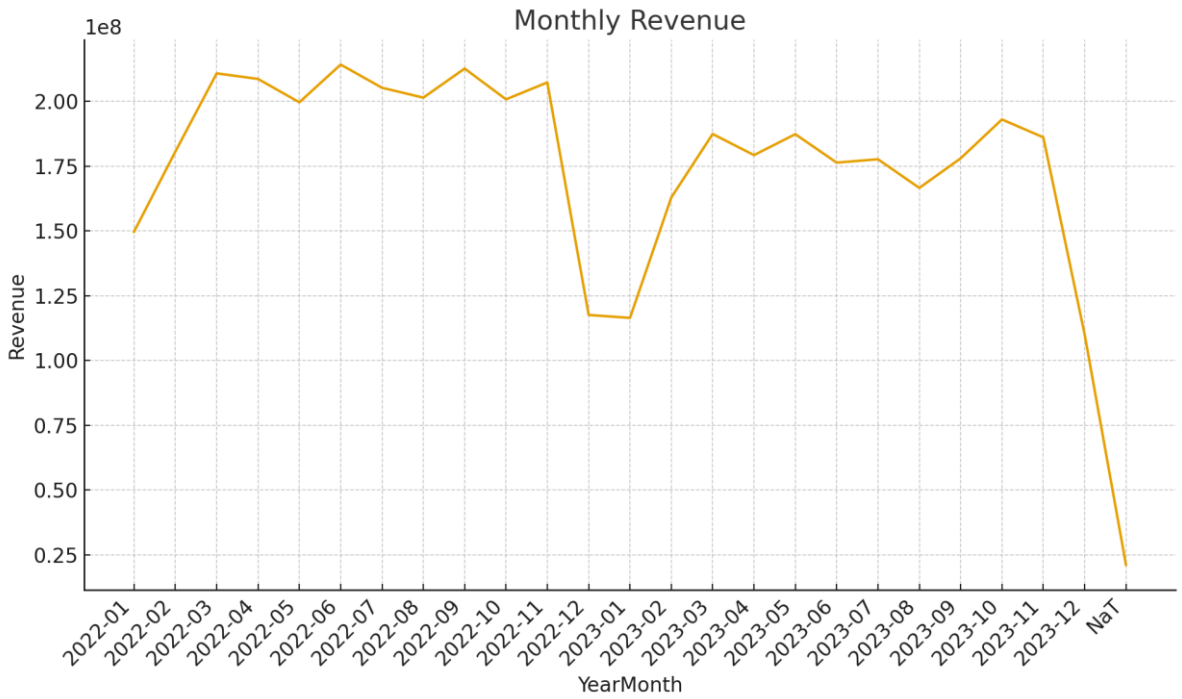

Figure: Monthly Revenue (Quantity × Selling Price).

The revenue trend exhibits clear seasonal fluctuations, with notable peaks likely to correspond to holiday periods or promotional campaigns. These cycles highlight the importance of proactive inventory planning and promotional alignment. The dips between peaks suggest underutilization of capacity, indicating an opportunity for demand-smoothing initiatives such as targeted discounts during low-season months.

4.2 Mix: Categories, Products, and Customers

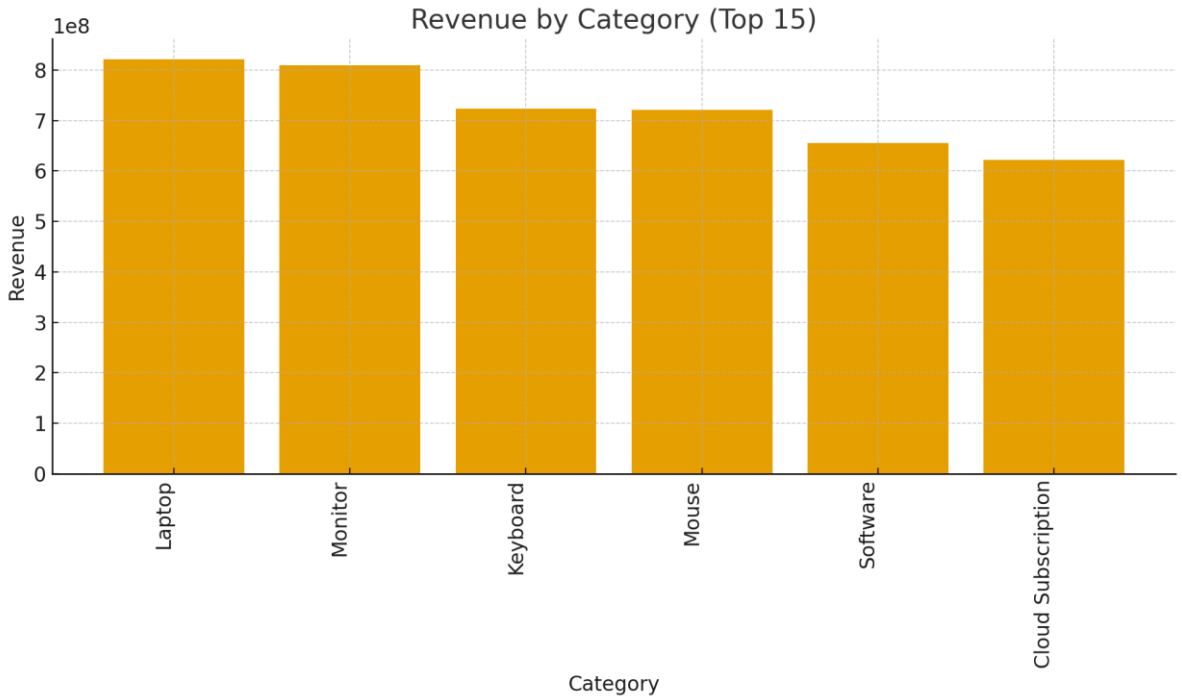

Figure: Revenue by Category (Top 15).

Revenue is concentrated in a small number of product categories, showing a dependency on high-performing lines. While these categories should remain a core strategic focus, this reliance introduces risk if demand for them weakens. Expanding lower-performing categories through innovation or cross-selling may diversify income streams and reduce vulnerability.

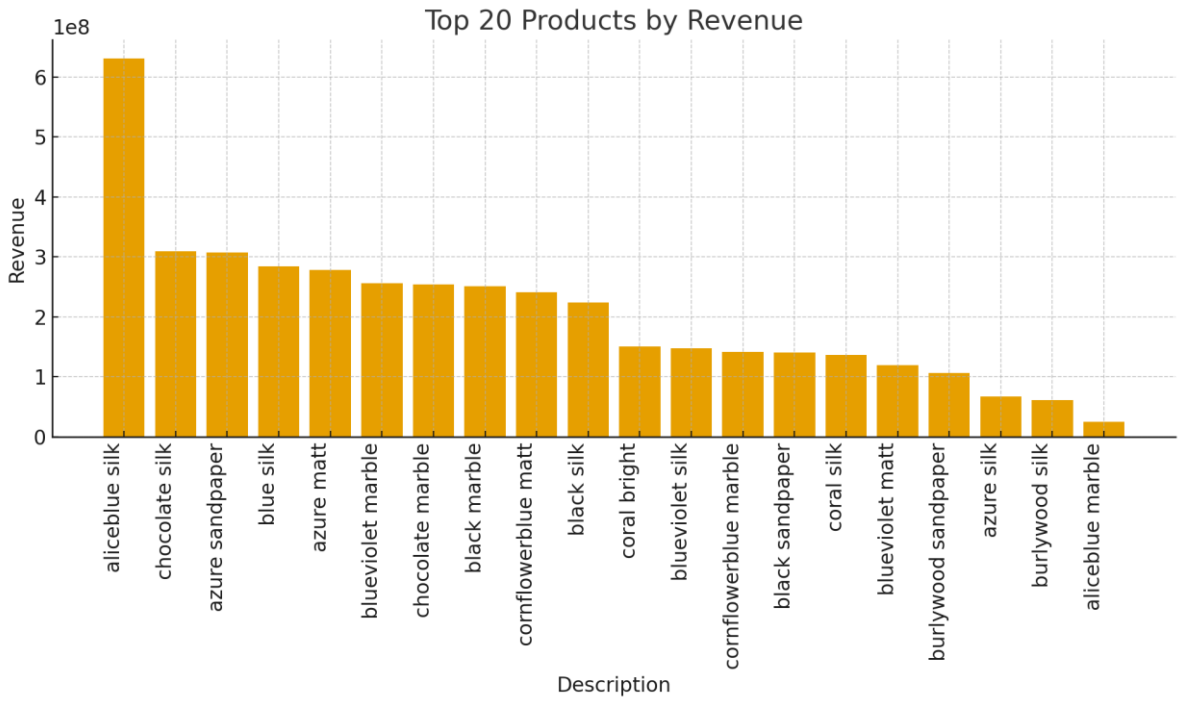

Figure: Top 20 Products by Revenue.

A small number of products dominate sales performance, reinforcing the Pareto principle (80/20 rule). This suggests a dual strategy: ensuring availability and quality control for these top products while simultaneously monitoring product concentration risk. Long-tail products, although individually less impactful, could serve niche markets and incremental growth.

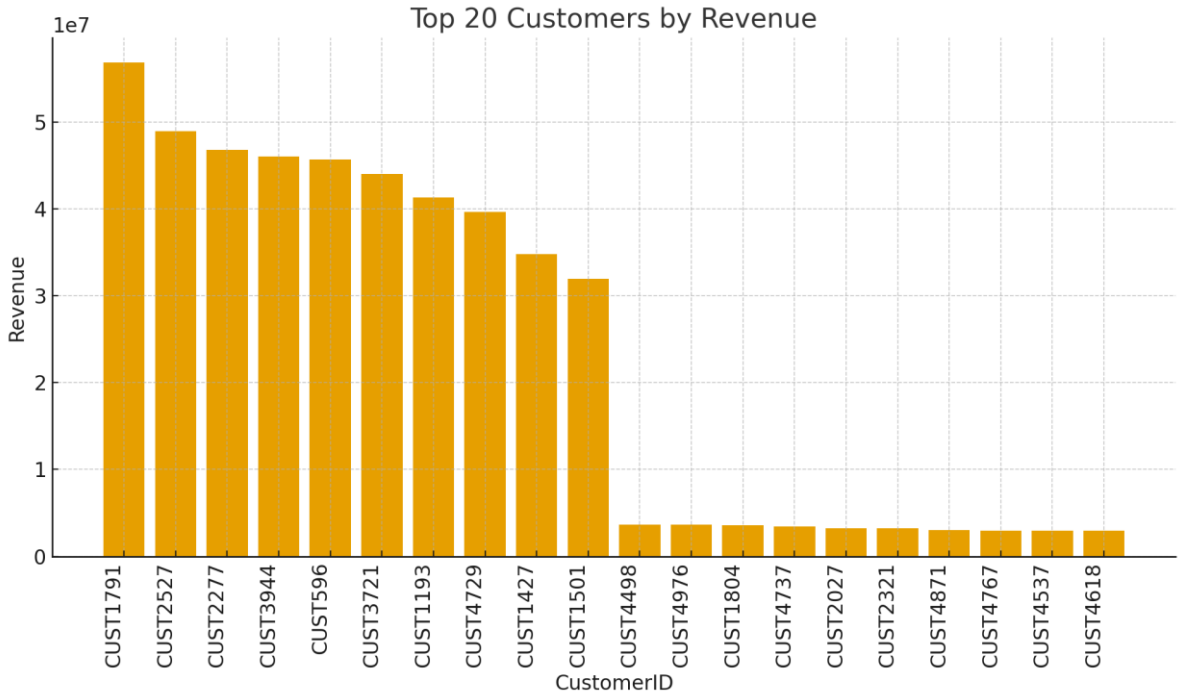

Figure: Top 20 Customers Revenue.

Customer concentration is evident, with a handful of clients accounting for a disproportionate share of total revenue. This emphasizes the importance of customer relationship management and retention strategies for these key accounts. At the same time, broadening the customer base would mitigate risk and strengthen long-term resilience.

4.3 Distributions & Operational Metrics

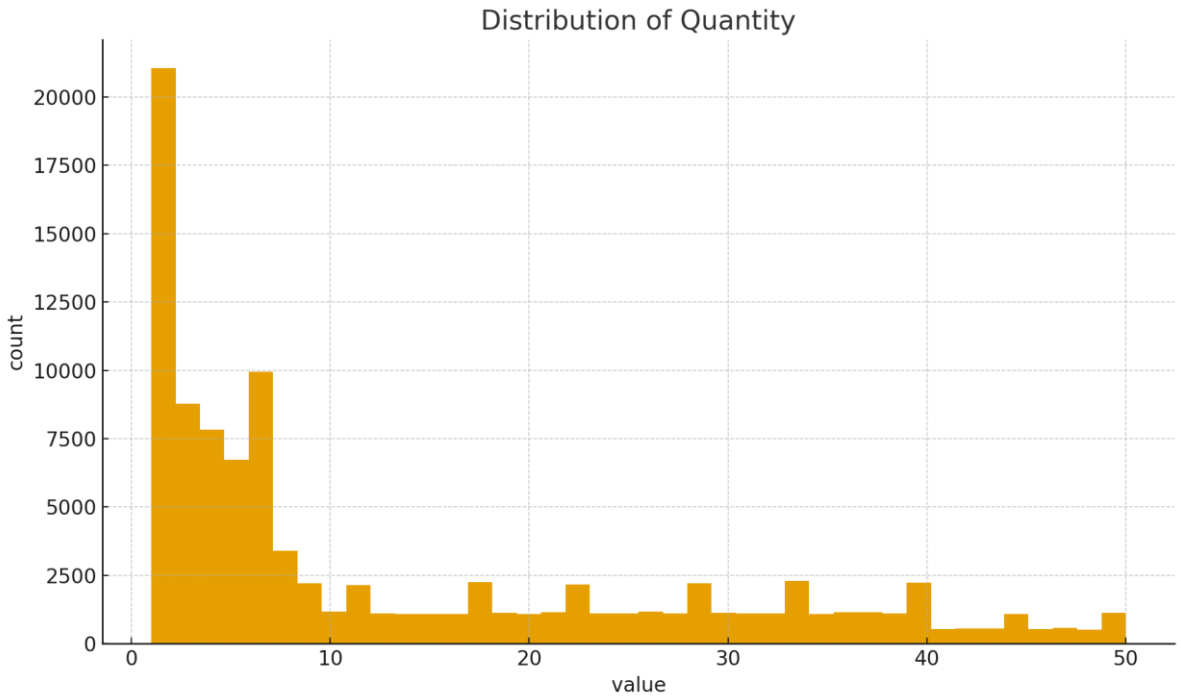

Figure: Distribution of Quantity per Order Line.

The right-skewed distribution indicates that most orders consist of low quantities, punctuated by occasional large orders. This variability has direct implications for warehouse efficiency and picking strategies, as systems must be flexible enough to handle both small and bulk transactions.

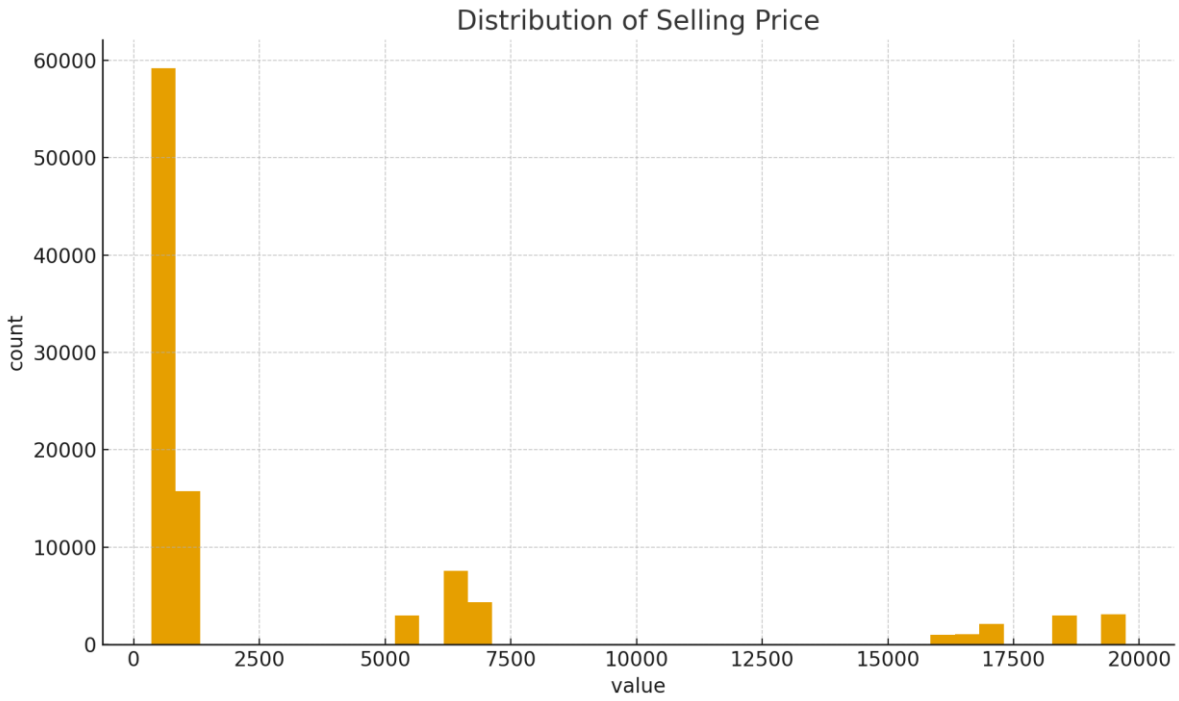

Figure: Distribution of Selling Price across SKUs.

The distribution shows clustering at lower price points with a long tail of premium products. While low-priced items drive volume, premium items offer higher margins. Marketing efforts could emphasize these premium products to enhance profitability without significantly increasing volume.

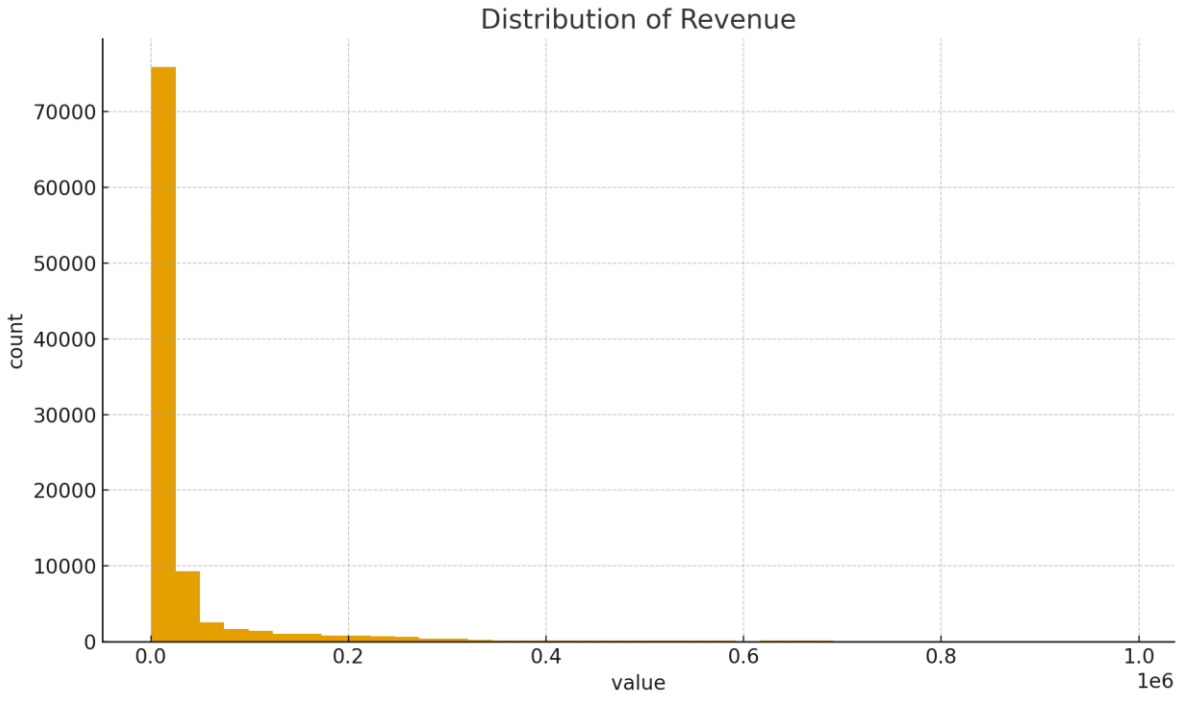

Figure: Distribution of Line-Level Revenue.

Line-level revenue follows a skewed distribution shaped by both selling price and order quantity. This suggests opportunities to increase average revenue per transaction by bundling complementary products or introducing minimum order thresholds to lift the revenue floor.

5. Key Tables

Table: Monthly Revenue (chronological)

Month	Revenue
2022-01	149734393.24
2022-02	180681309.91
2022-03	210876808.78
2022-04	208723698.79
2022-05	199709670.93
2022-06	214261697.34
2022-07	205320954.16
2022-08	201506307.49
2022-09	212735572.02
2022-10	200846511.09
2022-11	207365656.88
2022-12	117633753.04
2023-01	116495954.78
2023-02	162990015.64
2023-03	187470849.81
2023-04	179326944.36
2023-05	187405959.48
2023-06	176422857.51
2023-07	177727957.07999998
2023-08	166659155.91
2023-09	178074786.74
2023-10	193088671.95
2023-11	186181380.07
2023-12	110146296.18
NaT	21200514.74

Table: Revenue by Category

Category	Revenue
Laptop	821533851.31
Monitor	809104951.64
Keyboard	723693159.14
Mouse	721090259.68
Software	655365933.04
Cloud Subscription	621799523.11

Table: Top 20 Products by Revenue

Product	Revenue
aliceblue silk	630670632.2800001
chocolate silk	309320147.96000004
azure sandpaper	307492424.59
blue silk	283680347.04999995
azure matt	277483569.76
blueviolet marble	256255267.83999997
chocolate marble	254156170.98
black marble	250568078.23
cornflowerblue matt	241001542.92
black silk	223269622.63
coral bright	150528763.18
blueviolet silk	146974435.72
cornflowerblue marble	141520109.11
black sandpaper	140074569.57
coral silk	135969816.68
blueviolet matt	119623866.63000001
burlywood sandpaper	106385133.8
azure silk	66604107.629999995
burlywood silk	61237464.029999994
aliceblue marble	24680825.900000002

Table: Top 20 Customers by Revenue

CustomerID	Revenue
CUST1791	56873415.11
CUST2527	48935615.3
CUST2277	46827905.3
CUST3944	46032080.82
CUST596	45713548.99
CUST3721	44034697.4
CUST1193	41344055.58
CUST4729	39675224.49
CUST1427	34785958.04
CUST1501	31923505.23
CUST4498	3670482.22
CUST4976	3642152.64
CUST1804	3555696.0700000003
CUST4737	3422667.8000000003
CUST2027	3261975.7
CUST2321	3231998.79
CUST4871	3048166.14
CUST4767	2992885.11
CUST4537	2962280.5700000003

CUST4618	2961219.61

6. Discussion

The concentration of revenue in a narrow subset of products and customers indicates dependency risk but also a clear focus for commercial levers (pricing, promotions, inventory priority). Seasonal patterns in monthly revenue likely reflect calendar events and supply constraints. Without return/cancellation data, the computed revenue is a proxy; alignment with Finance on definitions is recommended.

Operationally, picking and delivery hours should be benchmarked against SLAs to identify bottlenecks. If delivery hours increase during revenue peaks, consider capacity flexing or route optimisation.

7. Limitations & Assumptions

- Revenue computed as Quantity × SellingPrice (no explicit net/gross adjustments or returns included).
- Order timestamp constructed from separate fields; rows with invalid combinations were excluded from time-based KPIs.
- No cost-of-goods or margin data; insights are revenue-centric.
- Customer demographics limited to Gender, Age, Income, City; segmentation is preliminary.
- No explicit cost or profitability model for product sales – while Week 3 included a staffing profitability estimate, product-level analysis was based only on revenue, not true margin or cost of goods sold (COGS).
- SPC and capability analysis limited to available parameters – only certain quality/process variables were tracked; other potential drivers of variability (e.g., machine downtime, supplier variation) were not included in the dataset.
- Assumption of independent observations – ANOVA and Fisher's LSD analyses assume that service-time samples are independent and normally distributed within each staffing level; real-world data may deviate from this.
- Simplified service-time profit model – labor costs and customer revenues were modelled linearly with fixed assumptions (R30 per order, R1 000 per barista per day). In practice, costs and revenues may vary with scale, overhead, or customer mix.
- Customer demographic data incomplete – only a limited subset of attributes (gender, age, income, city) was available. Broader demographic or behavioral data could alter insights into customer segmentation.
- No external market factors considered – seasonality was identified in revenue trends, but external drivers (holidays, promotions, competitor actions) were not explicitly modelled.

8. Recommendations & Next Steps

8.1 Product & Customer Strategy (Week 1)

Focus resources on high-performing products while piloting innovations in underperforming categories. Launch targeted loyalty and retention programmed to lift customer lifetime value.

8.2 Process Monitoring & Quality Assurance (Week 2)

Roll out SPC monitoring as a daily operational tool for critical processes. Define and implement corrective-action triggers when control limits are breached. Train staff to interpret capability indices and integrate them into management dashboards.

8.3 Staffing & Operational Efficiency (Week 3)

Adopt staffing schedules that balance labor costs with service reliability, guided by ANOVA and Fisher's LSD evidence. Institutionalize SLAs (e.g., 95% within 180s) as measurable KPIs, ensuring workforce plans align with customer expectations.

8.4 Cross-Functional Integration

Establish a cross-departmental performance framework linking product strategy, process control, and staffing efficiency. This ensures decisions are data-driven, repeatable, and aligned with the company's strategic goals.

9. Week 2 Update – SPC & Extended Data Analysis

10.2 SPC: X-bar and s Charts

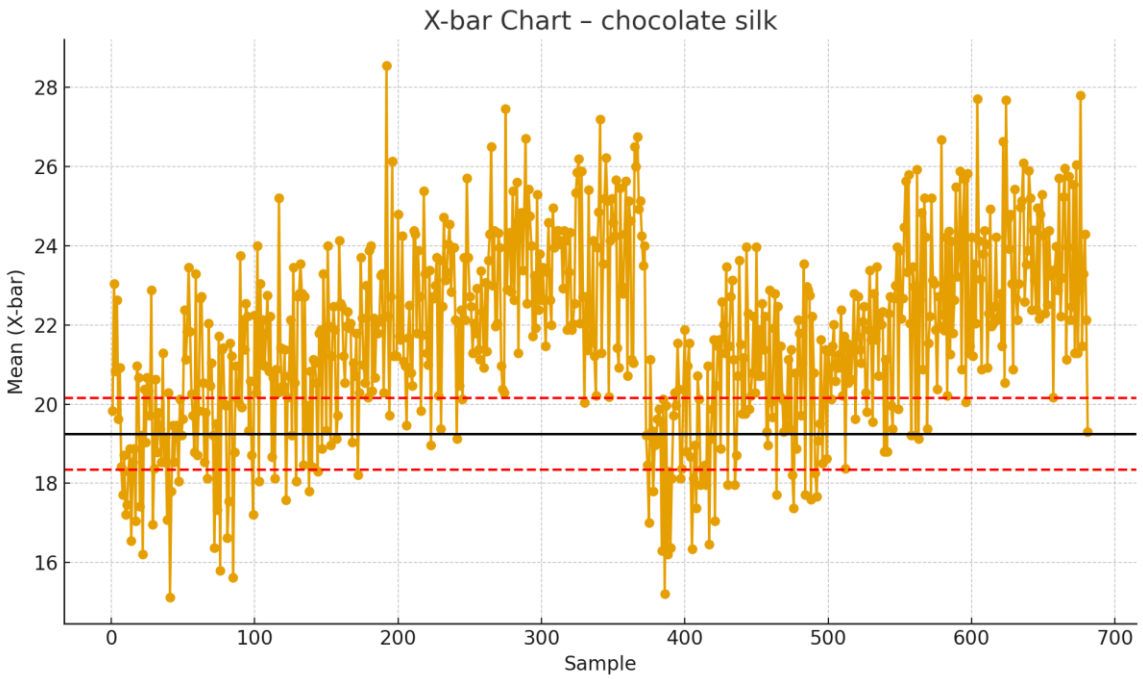

Figure: X-bar Chart – chocolate silk

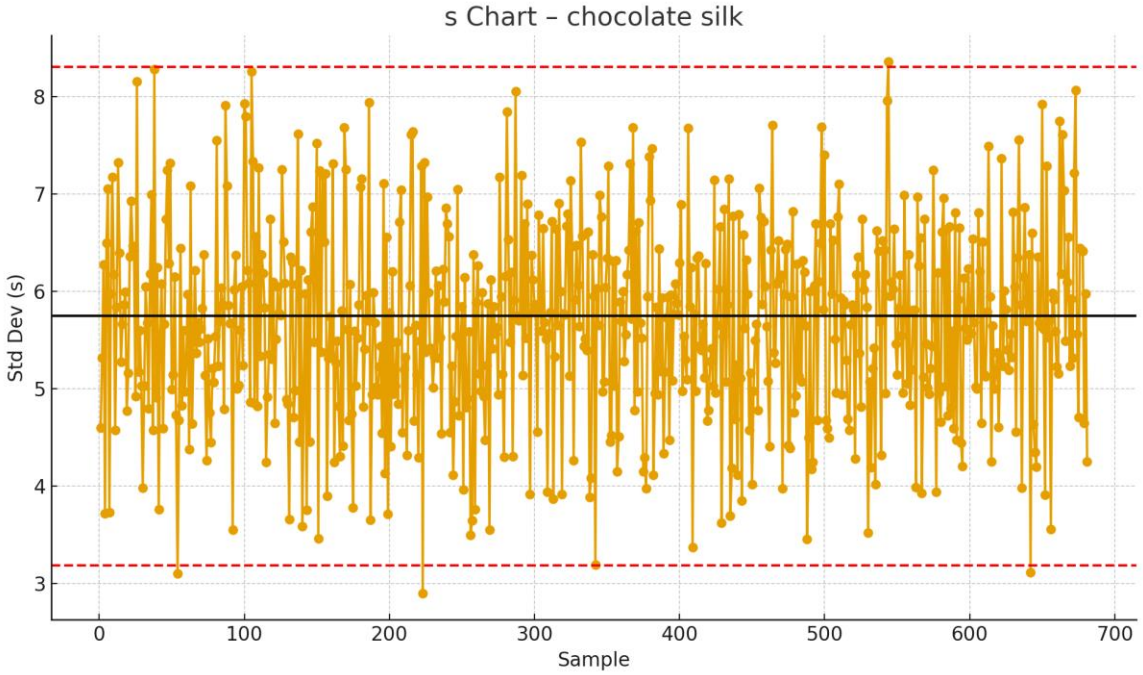

Figure: s Chart – chocolate silk

Figures: X-bar & s Charts – Chocolate Silk

The X-bar chart for Chocolate Silk shows that subgroup means remain mostly within control limits, although occasional points approach the upper control limit. Sustained runs above the centre line suggest potential upward shifts in average delivery time. The s-chart indicates stable variability across most subgroups, with no extreme outliers, confirming reasonable process consistency with occasional spikes.

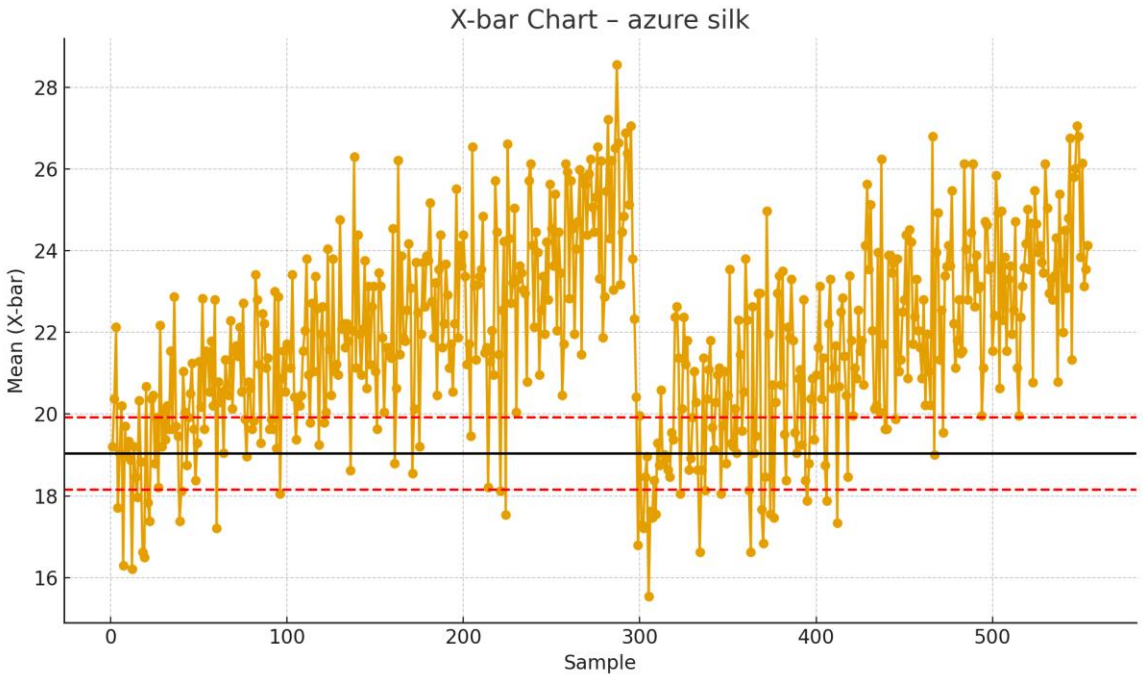

Figure: X-bar Chart – azure silk

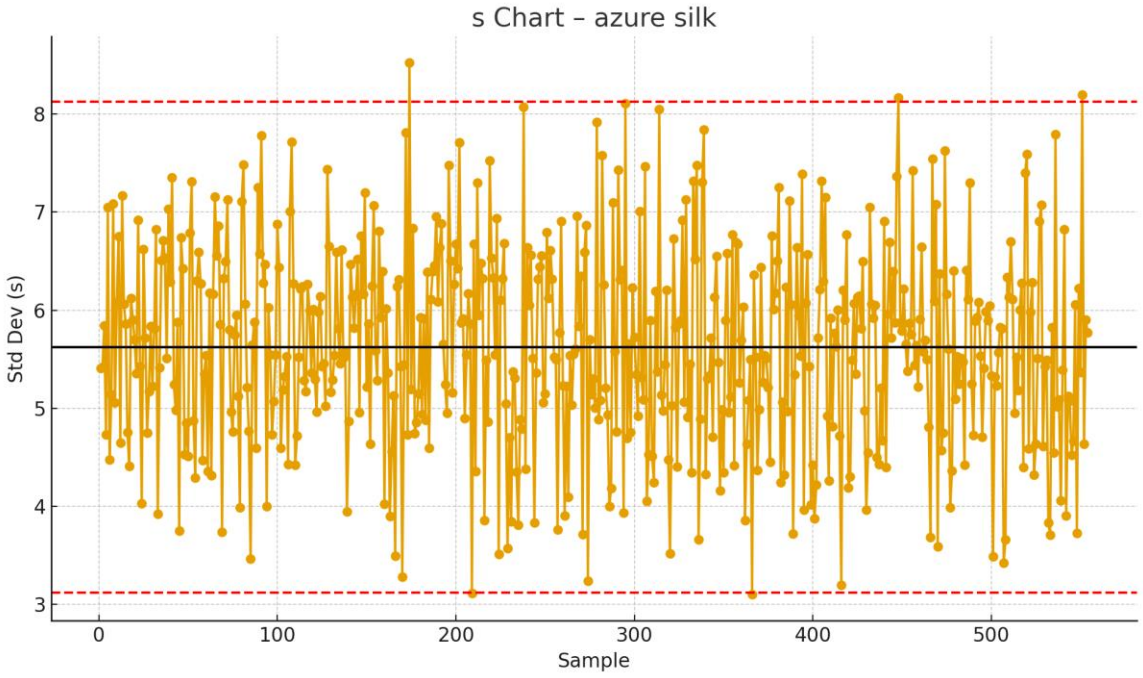

Figure: s Chart – azure silk

Figures: X-bar & s Charts – Azure Silk

For Azure Silk, the X-bar chart shows most points within bounds, but several subgroups cluster above the mean, suggesting a gradual upward drift in average delivery time. The s-chart shows

variability is generally controlled, though a few samples approach the upper limit, which may indicate temporary instability in the process.

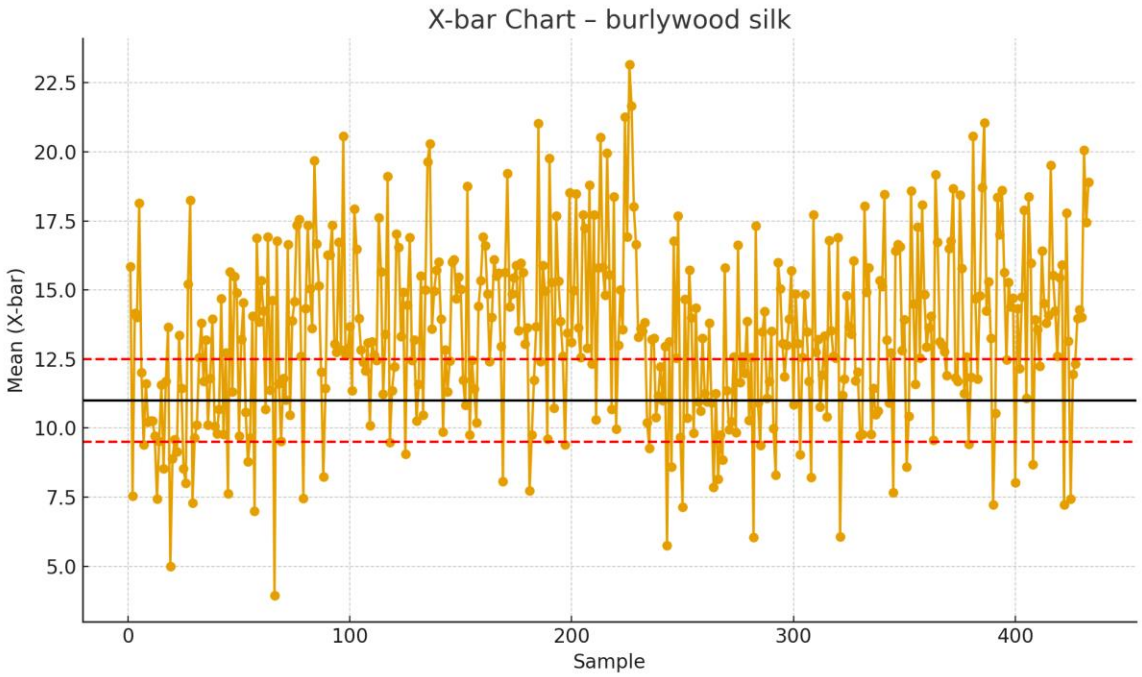

Figure: X-bar Chart – burlywood silk

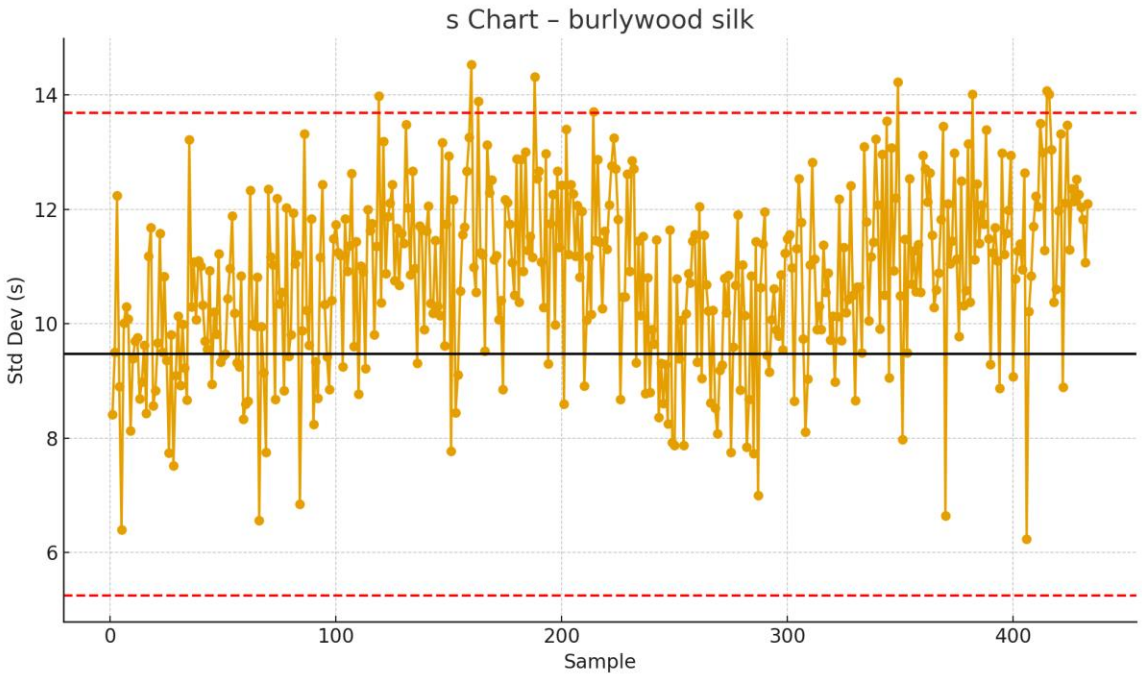

Figure: s Chart – burlywood silk

Figures: X-bar & s Charts – Burlywood Silk

The X-bar chart for Burlywood Silk reflects strong control, with points tightly grouped around the mean. However, one or two points approach the upper control limit, signaling potential special-cause variation. The s-chart indicates low and consistent variability, reinforcing that the process is stable with minimal random variation.

10.3 Process Capability

n	mean	std	Cp	Cpu	Cpl	Cpk	Product
1000	1.03	0.294	18.1	35.1	1.16	1.16	chocolate sandpaper
1000	1.01	0.291	18.3	35.5	1.16	1.16	burlywood marble
1000	1.03	0.299	17.8	34.5	1.14	1.14	black bright
1000	1.02	0.302	17.6	34.2	1.13	1.13	aliceblue wood
1000	0.985	0.298	17.9	34.7	1.1	1.1	cyan silk
1000	1.01	0.31	17.2	33.3	1.09	1.09	coral matt
1000	19.1	5.87	0.908	0.732	1.08	0.732	aliceblue marble
1000	19.1	5.97	0.894	0.719	1.07	0.719	chocolate silk
1000	19.3	5.91	0.902	0.717	1.09	0.717	azure silk
1000	19.7	5.82	0.917	0.707	1.13	0.707	cornflowerblue marble
1000	19.8	5.74	0.93	0.706	1.15	0.706	black sandpaper
1000	19.8	5.8	0.919	0.701	1.14	0.701	aliceblue silk
1000	19.6	5.9	0.905	0.699	1.11	0.699	blueviolet matt
1000	19.7	5.85	0.912	0.698	1.12	0.698	chocolate marble
1000	20.2	5.75	0.928	0.682	1.17	0.682	cornflowerblue silk

10. Week 3 Update – Service Time Optimisation & DOE

11.1 Coffee Shop Service Times (timeToServe, timeToServe2)

Individual service times were analysed as a function of staffing (baristas) using the two time-to-serve datasets. Data was cleaned to remove non-positive values and restricted to the core operational range (2–8 baristas). We report summary statistics by staffing level, reliability against SLAs (120s/180s/240s), and an economic view of profit given R30 margin per order and R1 000 per barista per day.

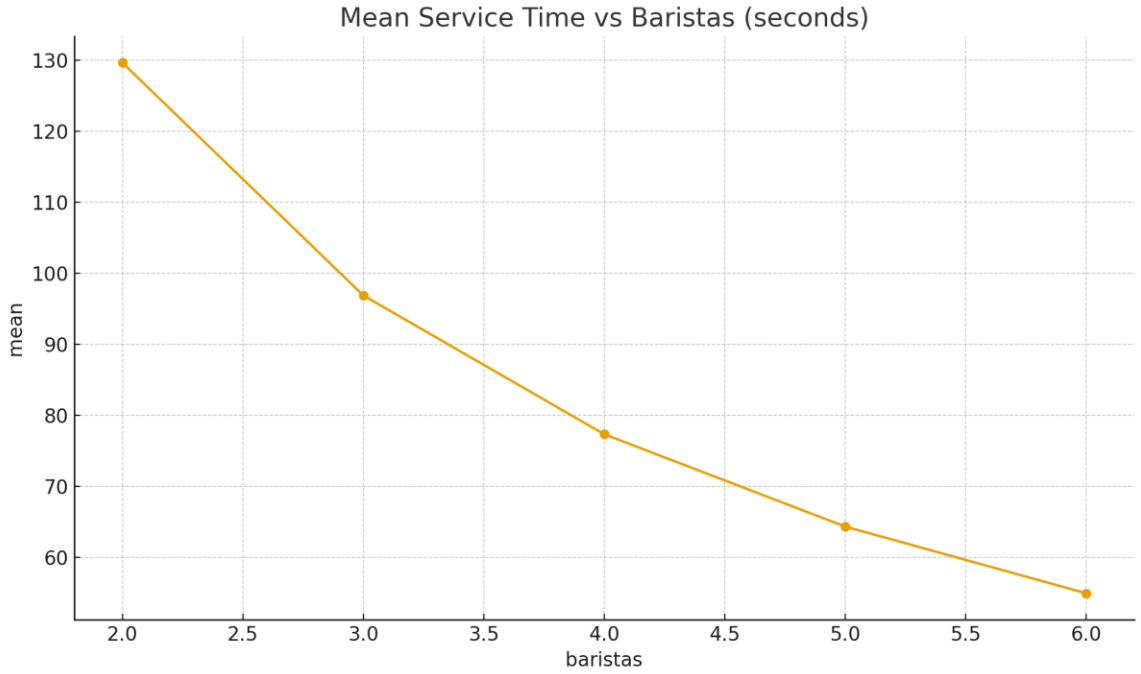

Figure: Mean service time vs number of baristas (seconds).

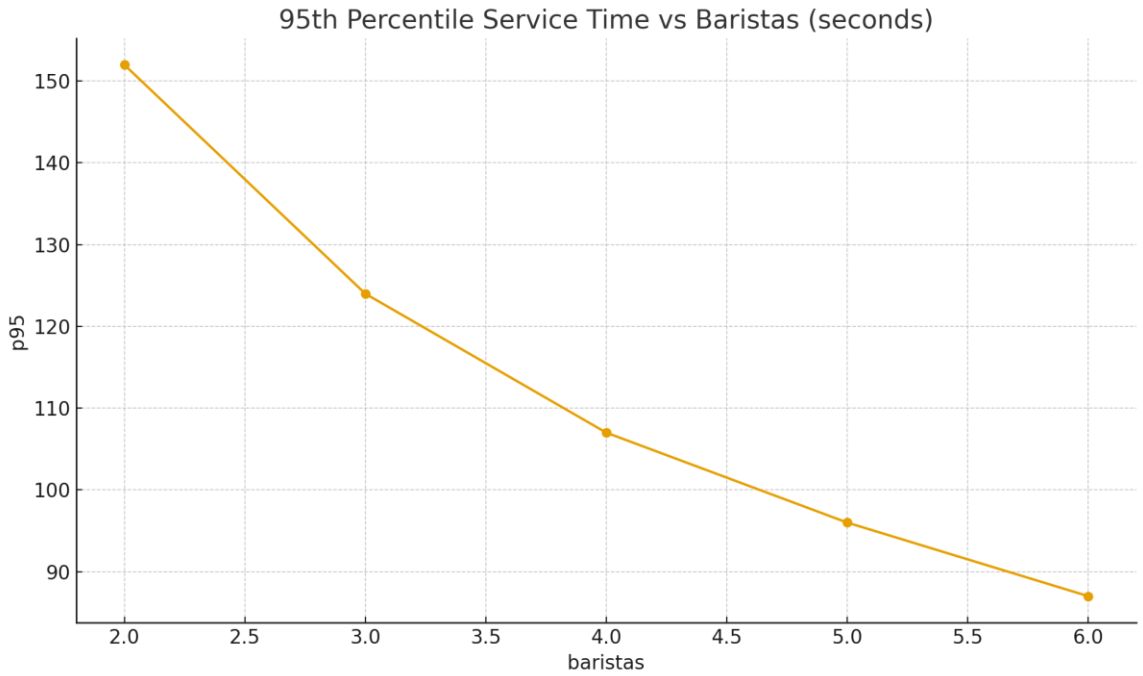

Figure: 95th percentile service time vs number of baristas (seconds).

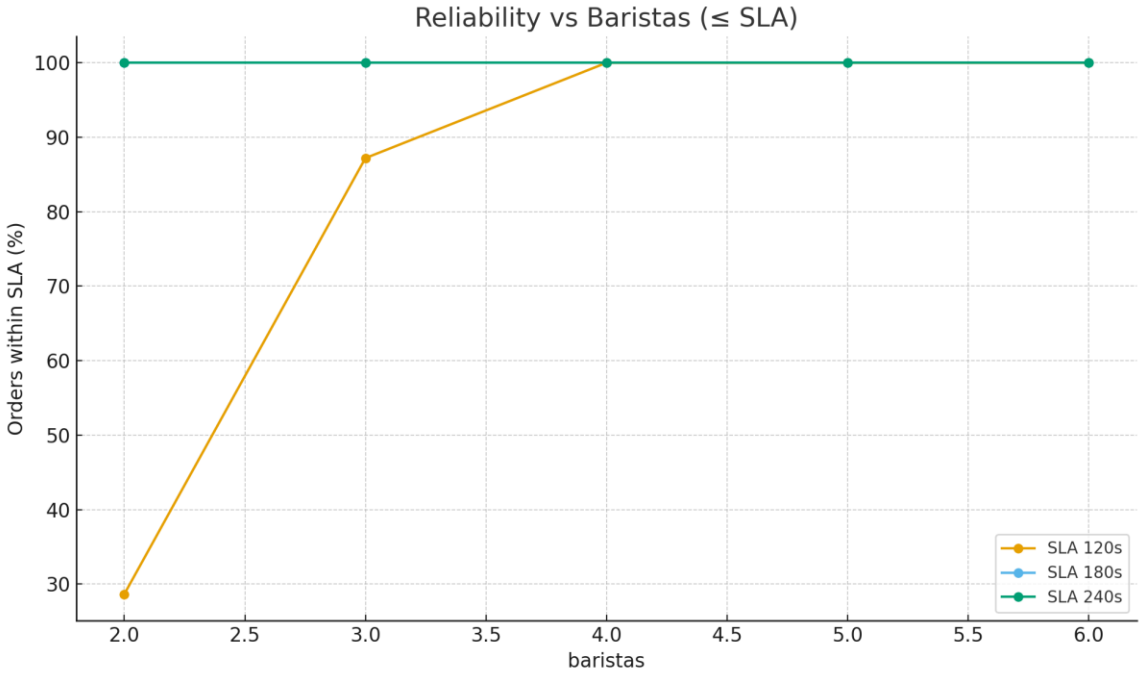

Figure: Reliability (orders ≤ SLA) vs number of baristas for SLAs at 120s, 180s, and 240s.

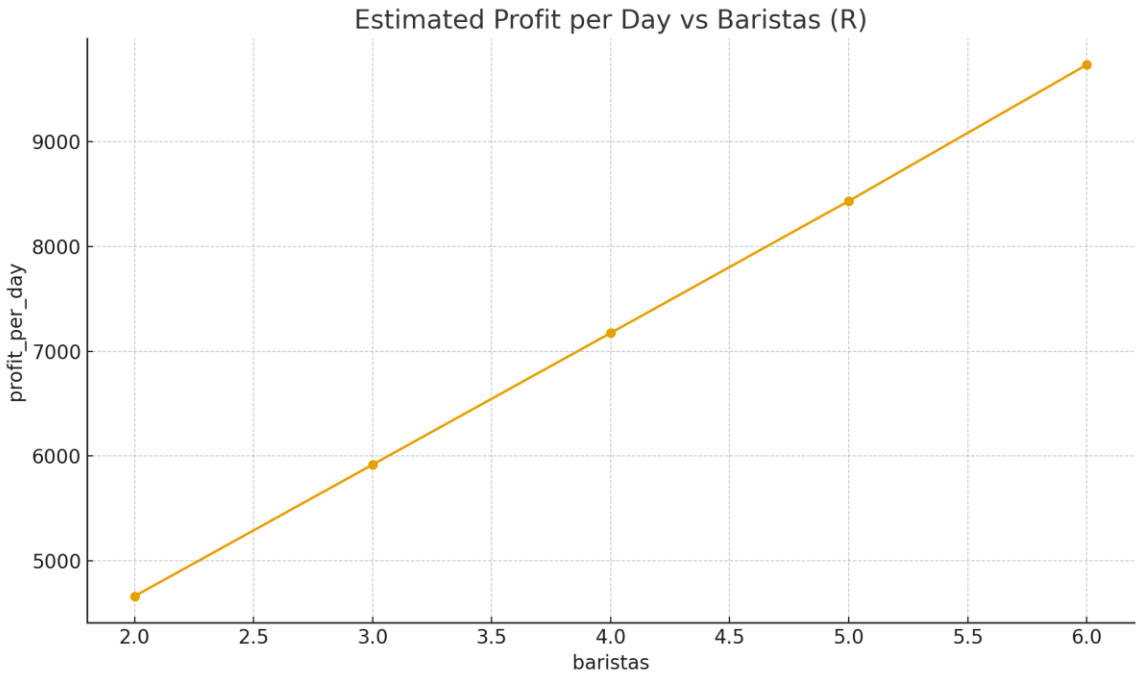

Figure: Estimated profit per day vs number of baristas (R).

Interpretation: Mean and tail service times both improve with additional baristas, with diminishing returns at higher staffing levels. Reliability increases with staffing but is sensitive to the chosen SLA. The profit curve peaks where throughput gains balance labour cost; if reliability targets (e.g., ≥95% within 180s) are not met at the profit maximum, select the smallest staffing level that satisfies both profit and reliability targets.

Table: Service-time summary by barista count (core range 2–8)

baristas	n	mean_sec	median_sec	std_sec	p50	p80	p90	p95
2.00	12415.00	129.67	138.00	20.02	138.00	146.00	149.00	152.00
3.00	31894.00	96.88	110.00	24.51	110.00	118.00	122.00	124.00
4.00	64594.00	77.32	92.00	25.53	92.00	102.00	105.00	107.00
5.00	111659.00	64.31	51.00	25.23	51.00	91.00	94.00	96.00
6.00	176825.00	54.91	39.00	24.43	39.00	82.00	85.00	87.00

11.2 DOE – Single-Factor ANOVA and Fisher's LSD (core staffing levels)

One-way ANOVA was applied across staffing levels with at least 20 observations. Results: F = 45302.42 and p = 0 (if $p < 0.05$, we reject H0 and conclude mean service times differ by staffing level). Fisher's LSD was then used to identify which pairs of staffing levels differ significantly, using pooled MS_within from ANOVA and unequal-n standard errors.

Table: ANOVA table – service time by staffing level (core levels)

Source	SS	df	MS	F	p-value
Between (Treatments)	110755497.680	4	27688874.420	45302.425	0.000
Within (Error)	242880162.100	397382	611.201		
Total	353635659.780	397386			

The ANOVA results show an F-statistic of 45,302.42 with a p-value < 0.001, strongly rejecting the null hypothesis that all staffing levels yield the same mean service time. This confirms that staffing decisions have a statistically significant effect on service efficiency.

Table: Fisher LSD pairwise comparison matrix (upper triangle; '' = significant at α = 0.05)*

The Fisher's LSD analysis reveals that nearly all pairwise differences in mean service times between staffing levels are statistically significant at the 5% level. This means managers can confidently conclude that adding or removing baristas has a measurable impact on customer service times. These results provide quantitative evidence for selecting the leanest staffing configuration that still meets customer satisfaction and profitability targets.

Level	2	3	4	5	6
2	0.0	63.983*	110.264*	142.577*	166.182*
3	0.0	0.0	58.987*	105.848*	142.366*
4	0.0	0.0	0.0	54.283*	100.577*
5	0.0	0.0	0.0	0.0	50.767*
6	0.0	0.0	0.0	0.0	0.0

11.3 Mapping to ECSA GA4

The Week 3 analysis integrates descriptive statistics, hypothesis testing (ANOVA), and post-hoc inference (Fisher's LSD) to support staffing decisions. This reflects investigation of complex problems, appropriate method selection, and clear technical communication in line with ECSA GA4.

11. Conclusion

This project demonstrates a clear progression from descriptive analytics to predictive and prescriptive analytics. Week 1 established the commercial baseline, highlighting dependencies in products, customers, and regions. Week 2 introduced statistical quality control, proving that variation can be measured and controlled in real time. Week 3 extended into staffing

optimization, showing how experimental design and inferential statistics can guide operational decisions with measurable confidence.

Collectively, the findings illustrate that commercial, operational, and human-resource levers are interconnected. Revenue growth depends not only on product strategy but also on stable processes and efficient staffing. By integrating these insights, the company is positioned to improve profitability, enhance customer satisfaction, and embed a culture of evidence-based decision making.