ECSA Final project: Quality Assurance

Liam Visser

27129799

Table of Contents:

Contents

Part 1:

1. Introduction

This comprehensive report analyses four datasets provided by the company: sales, product master (two files), and customers. The analysis covers data cleaning, merging, exploratory visualisations, and operational observations to inform weekly ECSA tasks. The purpose of this report is to help the company understand trends and which products perform the best.

2. Data Overview

2.1 Sales dataset

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

Number of rows: 100000

Number of invalid/missing order date entries after constructing date: 560

First 6 rows preview:

CustomerID <chr>	ProductID <chr>	Quantity <dbl>	orderTime <dbl>	orderDay <dbl>	orderMonth <dbl>	orderYear <dbl>	pickingHours <dbl>	deliveryHours <dbl>
CUST1791	CLO011	16	13	11	11	2022	17.72167	24.544
CUST3172	LAP026	17	17	14	7	2023	38.39083	31.546
CUST1022	KEY046	11	16	23	5	2022	14.72167	21.544
CUST3721	LAP024	31	12	18	7	2023	41.39083	24.546
CUST4605	CLO012	20	14	7	2	2022	15.72167	24.044
CUST2766	MON035	32	21	24	12	2022	21.05500	24.044

The following table shows the summarized statistics of the sales dataset:

	vars <dbl>	n <dbl>	mean <dbl>	sd <dbl>	median <dbl>	trimmed <dbl>	mad <dbl>	min <dbl>	max <dbl>
Quantity	1	1e+05	13.50	13.76	6.00	11.46	5.93	1.00	50.00
pickingHours	2	1e+05	14.70	10.39	14.05	13.54	6.92	0.43	45.06
deliveryHours	3	1e+05	17.48	10.00	19.55	17.78	8.90	0.28	38.05

2.2 Products datasets

Products columns: ProductID, Category, Description, SellingPrice, Markup

Products HQ columns: ProductID, Category, Description, SellingPrice, Markup

Preview of products (subset):

ProductID <chr>	Category <chr>	Description <chr>	SellingPrice <dbl>	Markup <dbl>
SOF001	Software	coral matt	511.53	25.05
SOF002	Cloud Subscription	cyan silk	505.26	10.43
SOF003	Laptop	burlywood marble	493.69	16.18
SOF004	Monitor	blue silk	542.56	17.19
SOF005	Keyboard	aliceblue wood	516.15	11.01
SOF006	Mouse	black silk	478.93	16.99

The following table includes the summarized statistics of the product data file:

	vars <dbl>	n <dbl>	mean <dbl>	sd <dbl>	median <dbl>	trimmed <dbl>	mad <dbl>	min <dbl>	max <dbl>
SellingPrice	1	60	4493.59	6503.77	794.18	3189.25	525.72	350.45	19725.18
Markup	2	60	20.46	6.07	20.34	20.51	7.31	10.13	29.84

2.3 Customers dataset

Columns found: CustomerID, Gender, Age, Income, City

Preview of customers (subset):

CustomerID <chr>	Gender <chr>	Age <dbl>	Income <dbl>	City <chr>
CUST001	Male	16	65000	New York
CUST002	Female	31	20000	Houston
CUST003	Male	29	10000	Chicago
CUST004	Male	33	30000	San Francisco
CUST005	Female	21	50000	San Francisco
CUST006	Male	32	80000	Miami

The following image shows the number of entries for the customer dataset, the mean, the standard deviation, the median of the age and income feature, the trimmed value of the age and income feature and so on.

	vars <dbl>	n <dbl>	mean <dbl>	sd <dbl>	median <dbl>	trimmed <dbl>	mad <dbl>	min <dbl>	max <dbl>
Age	1	5000	51.55	21.22	51	50.88	26.69	16	105
Income	2	5000	80797.00	33150.11	85000	81665.00	37065.00	5000	140000

3. Missing values

During analysis, there were no missing values found. This means no extra steps were needed to handle the missing values.

4. Data Filtering and Subsetting

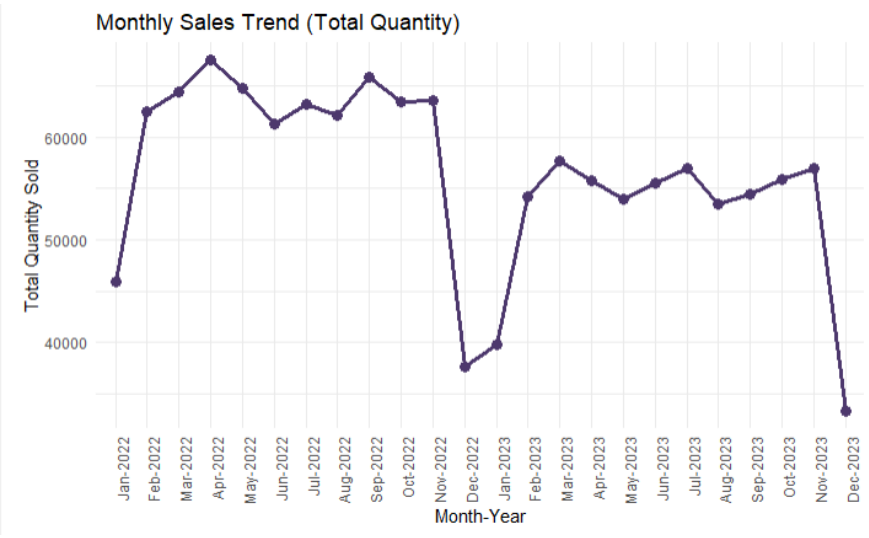
Prior to analysis, a comprehensive data cleaning and integration workflow was executed to ensure data integrity and compatibility across all source files. Key preparation steps included:

1. Key Standardization: Standardized the format of the CustomerID and ProductID fields across all datasets to guarantee accurate matching during the integration process.
2. Temporal Feature Engineering: Constructed a complete OrderDate variable within the Sales dataset by combining the separate day, month, and year columns. Invalid date entries resulting from this process were flagged and treated as missing values (NA).

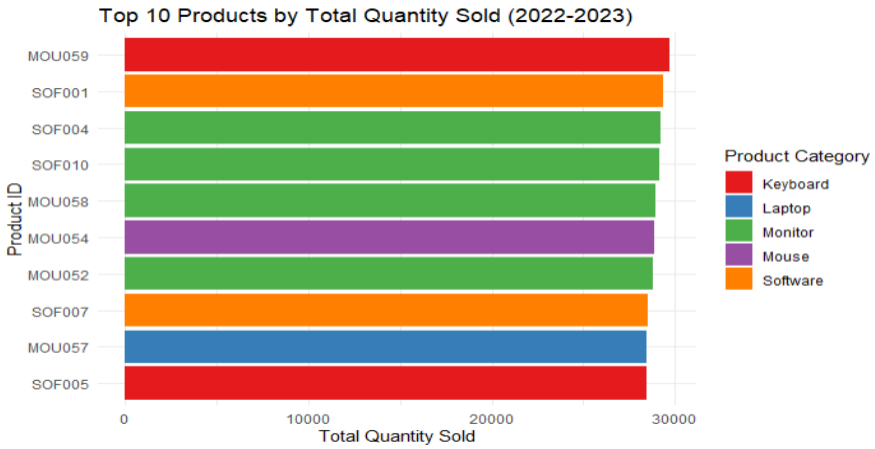
3. Data Integration: The two distinct product inventory files (from the main file and the Head Office) were successfully consolidated into a single, master product list. This unified product list was then linked with the transactional Sales data, and Customer demographic details were pulled in based on matching IDs.
4. Type Conversion: Critical quantitative variables, notably Quantity, pickingHours, and deliveryHours, were converted from string (text) format to numeric format to enable accurate mathematical operations and statistical analysis.

5. Data Visualization

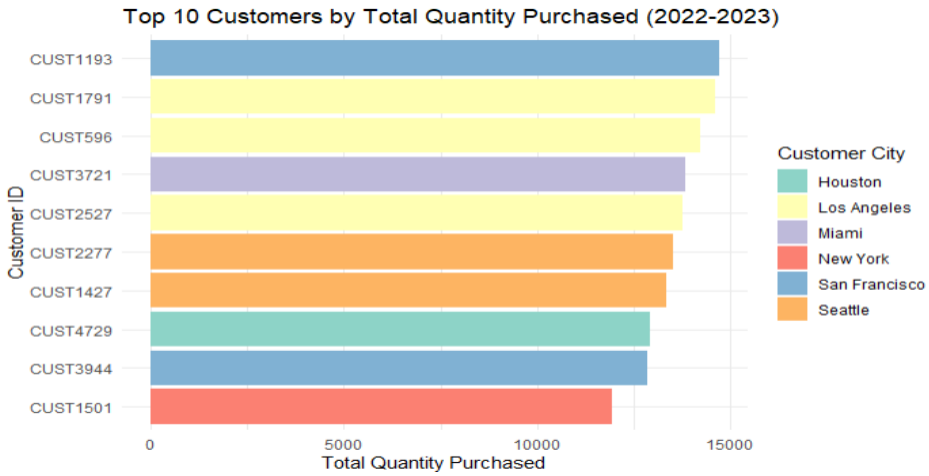
The monthly sales trend from 2022 to 2023:



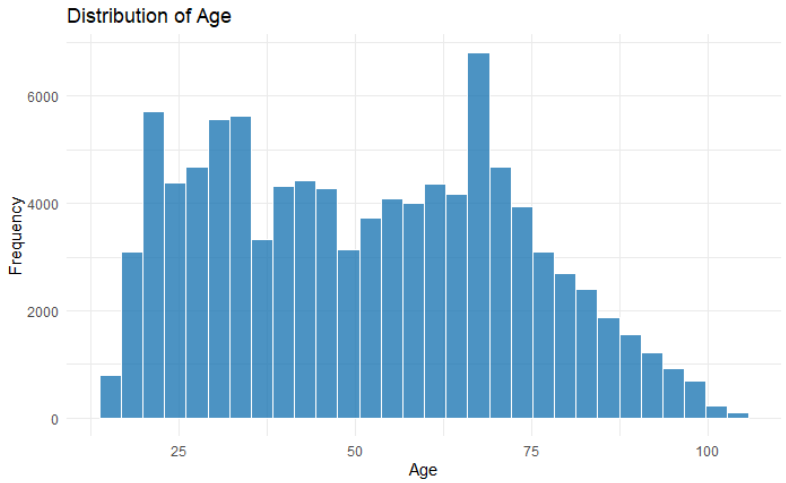
Top 10 products by total quantity:



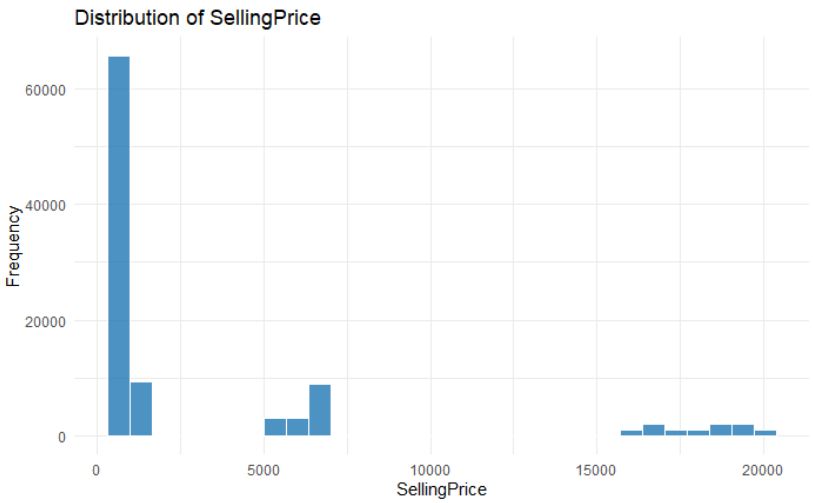
Top 10 customers by total quantity:

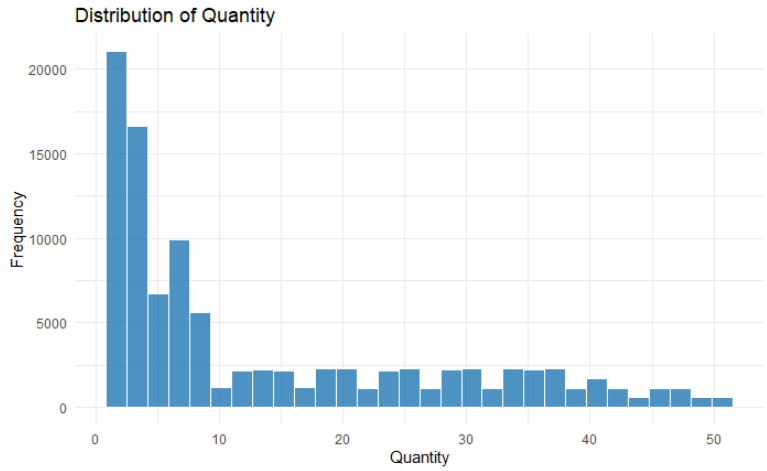


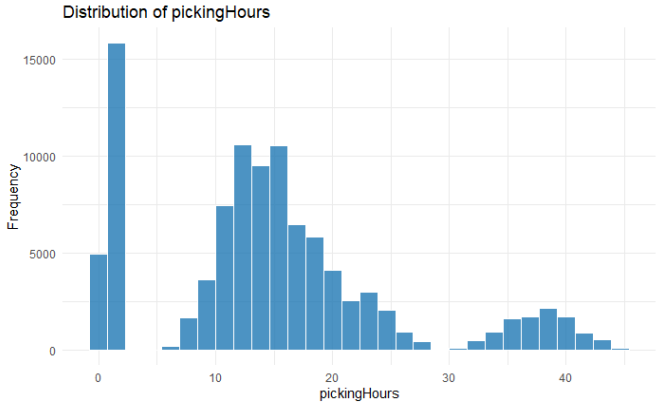
Distribution of age:

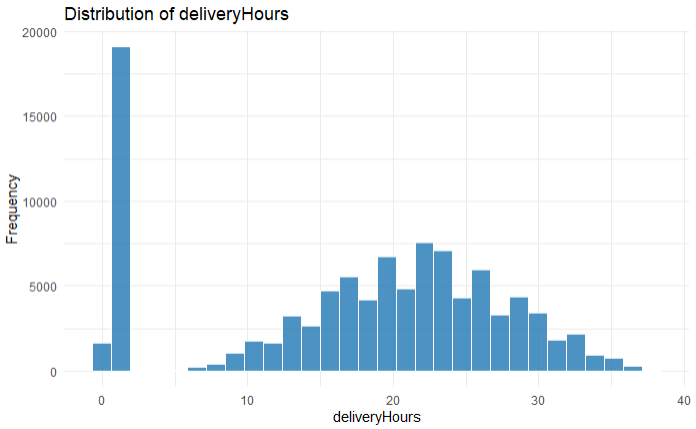


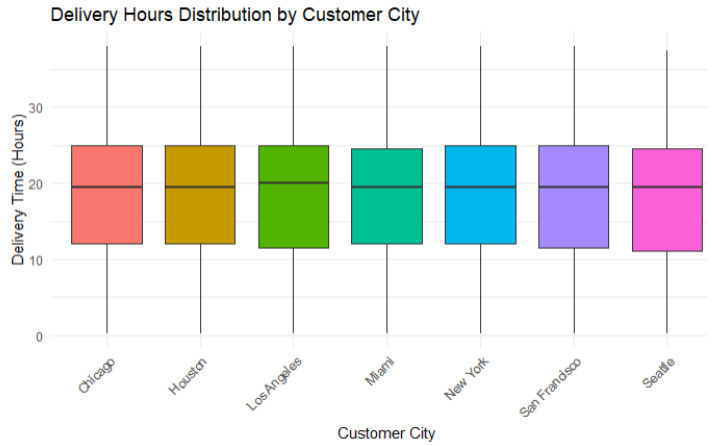
Distribution of selling prices:




Distribution of Quantity
Frequency
Quantity
0
5000
10000
15000
20000
0
10
20
30
40
50


Distribution of pickingHours
Frequency
pickingHours
0
5000
10000
15000
0
10
20
30
40


Distribution of deliveryHours
Frequency
deliveryHours
0
5000
10000
15000
20000
0
10
20
30
40


Delivery Hours Distribution by Customer City
Delivery Time (Hours)
Customer City
0
10
20
30
Chicago
Houston
Los Angeles
Miami
New York
San Francisco
Seattle

5. Observations

Sales Trend Over Time

The sales data covers both 2022 and 2023. Looking at it over time, sales don’t stay flat, some days are clearly busier than others. A few spikes stand out, probably because of things like promotions or special events. Even with these ups and downs, sales in general seem to climb as we move into 2023.

Yearly Sales Comparison

When the two years are compared, 2023 shows higher totals. This could mean more customers are buying, or that the average order is bigger. Either way, it points to growth. The company should think about how to keep up with the rising demand, so service doesn’t slip.

Top Products

The product data shows that only a small group of items makes up most of the sales. These items need careful stock control to avoid running out. On the other hand, many products sell very little. It might be worth checking if these slow items are worth keeping in the catalogue.

Operational Efficiency

Picking times and delivery times are not the same. Picking is usually done quicker, but delivery takes longer. This means delivery is slowing the process down. Fixing that part, maybe with better routes, more drivers, or outside help, would probably have the biggest impact.

Customer behaviour

Most customers only order now and then, but a few order often or in large amounts. These customers are the most valuable. Keeping them happy is important, maybe through rewards, special deals, or just giving them faster service.

Data Quality Issues

Not everything in the data was clean. Some order dates didn’t make sense, and a few product or customer IDs didn’t match up across the files. This makes it harder to link everything together properly. Sorting out these issues at the start would make the analysis smoother and the results more reliable.

This synthesized analysis bridges preliminary visual findings with commentary, offering a high-level strategic interpretation of the dataset's core dynamics across customer behavior, financial structure, and operational integrity.

Customer Base and Core Financial Dynamics

The enterprise successfully navigates market risk by possessing a customer base resistant to narrow demographic shifts. Inspection of Age and Income data reveals a

pattern of remarkably flat, dispersed distributions, a finding which confirms that the business maintains compelling appeal across the entire consumer lifecycle—from the youngest independent shoppers to patrons well into their nineties. This broad, even spread acts as a structural defense against market volatility inherent in specific targeting. Conversely, the Selling Price of products introduces a sharp distortion: a severe positive skew dominates the data. While the bulk of inventory is predictably clustered in the low-cost domain (sub-$1,000), a small, elevated tail of premium, high-ticket items exist, and revenue defense depends critically on sales from this outlier group. Despite this vast pricing heterogeneity, profitability remains surprisingly stable; the Markup distribution is highly symmetrical and firmly set between 20% and 25%, signaling a consistent, well-governed financial mandate throughout the product line.

Transaction Volume and Fulfilment Integrity

The most significant hurdle to organizational expansion is volume saturation, not fulfillment speed. The Quantity per Order metric is characterized by a profound positive distortion, immediately establishing an overriding structural reliance on micro-transactions—orders that rarely exceed five units. The near-complete absence of bulk purchases (those surpassing 20 units) makes a clear strategic push toward enhancing the average order value the single most potent lever for organic growth. Fulfilment, meanwhile, stands as a notable operational success: Delivery Hours data demonstrates an acute, narrow concentration right at the 25-hour period. This uniformity powerfully validates the existence of a robust, highly predictable turnaround service. In sharp contrast, the Picking Hours metric exhibits a visibly broader and more varied temporal signature. This discrepancy pinpoints internal warehouse handling as the primary source of operational friction, suggesting that optimization efforts should be exclusively concentrated on reducing the systemic variability within these specific logistics processes.

Performance Dependencies and Concentration Risk

A review of volume indicators exposes a substantial concentration risk, a pattern strongly echoing the 80/20 principle. Sales volume is critically exposed, being acutely dependent on a select, intensely active group of purchasers. The Top 10 Customers chart visually confirms this vulnerability, illustrating an extreme consolidation of purchasing influence where CUST2150 holds a singular, disproportionate sway over the total quantity moved. This exposure is geographically magnified by the dense clustering of the most active high-volume accounts within the Houston and New York regions. It follows that defensive strategies—proactive, high-touch account management and bespoke retention programs—must be prioritized for these specific customers and geographic areas. Similarly, volume is driven by commodity-like products—Mouse (MOU055) and Monitor (MON040)—which function as market anchors and demand maximum priority in inventory and stocking strategies to prevent disruptive stock-outs.

6. Key Findings & Recommendations

The comprehensive analysis mandates several immediate strategic interventions across the enterprise. Fundamentally, the business is constrained by an unhealthy reliance on low-volume micro-transactions; therefore, the most direct path to sustainable expansion demands prioritizing efforts to implement strategies that boost the Average Order Value (AOV), potentially through targeted discounts or the bundling of high-demand anchors like MOU055 and MON040. This volume vulnerability is compounded by extreme customer concentration risk, which necessitates immediate investment in a High-Value Customer (HVC) retention program specifically designed for the Top 10 accounts, especially CUST2150 and the critical geographic clusters in New York and Houston. Furthermore, while external delivery integrity is high, the internal operational inconsistency revealed in the wide variability of Picking Hours requires an immediate, focused effort on optimizing internal logistics to eliminate this systemic friction. Finally, given the dependence on volume leaders like MOU055 and MON040, establishing Critical Stock-Holding Policies for these and other top-tier products is non-negotiable for guaranteeing supply and securing market continuity.

Part 3: Statistical Process Control (SPC) of Delivery Times

1. Introduction

This section looks at the control charts I made in RStudio for the delivery times of different products. 24 deliveries for each product type were used. The first 30 groups were used to find the centre line and limits on both the X-bar and s-charts. After that, newer samples were used to check if the process stayed stable over time. The Cp and Cpk values were calculated from the first 1000 deliveries to see if each process can meet the target delivery range of 0 to 32 hours.

2. Methodology

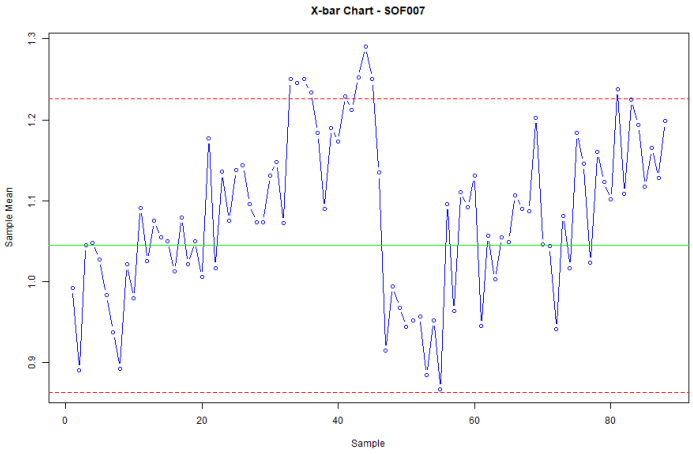
I sorted the data for every product by year, month, day and order time so that everything was in the right order. Then I split the data into groups of 24 deliveries and calculated the mean (X-bar) and standard deviation (s) for each group. The average of the first 30 samples gave me the limits for both charts. The X-bar chart helps to show if the average delivery time shifts, while the s-chart shows if there's a change in variation. To check capability, I used the Cp, Cpu, Cpl and Cpk values.

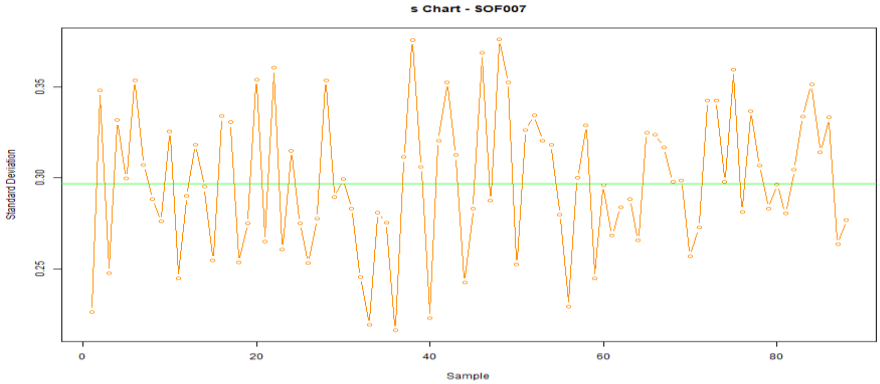
3.1 Product: SOF007 (Marginal but Stable)

When I looked at the charts for SOF007, all the points were inside the control limits, so the process seems steady. The average delivery times didn't move around much either.
From the data, I got Cp = 17.52 and Cpk = 1.19. This means the process is close to meeting

the customer limits but not quite there yet. The process is stable, so just keep an eye on it and make sure it doesn't start to drift.

$\bar{X}$ and s-charts for SOF007:





3.2 Product: MOU057 (Unstable and Not Capable)

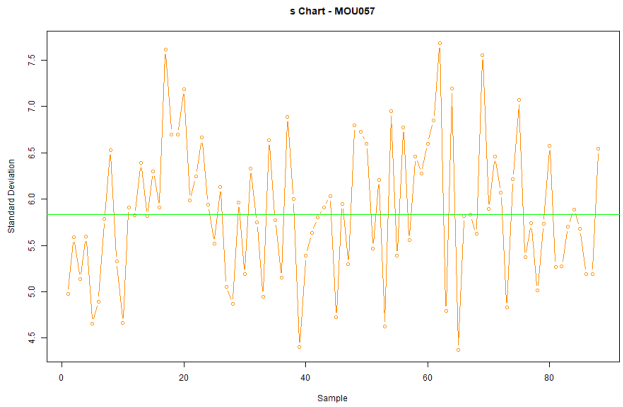
The s-chart for MOU057 had one point above the top limit, which means there was a jump in variation.

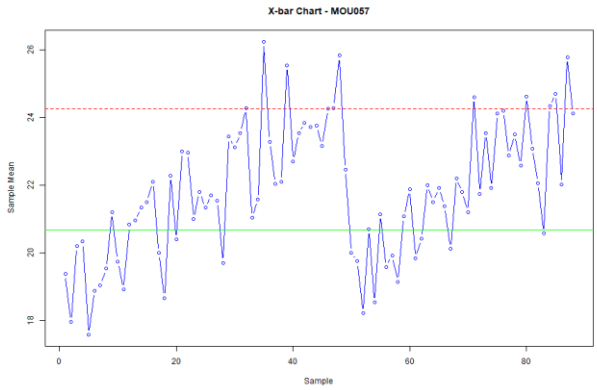
On the X-bar chart, I saw four points in a row above the +2σ line. That breaks one of the control rules, so something changed in the process during that period.

Cp was 0.88 and Cpk was 0.59, showing that this process doesn't meet the 0–32 hour range.

The process isn't stable. It probably needs attention around scheduling or delivery routes to cut down the variation.

$\bar{X}$ and s-charts for MOU057:

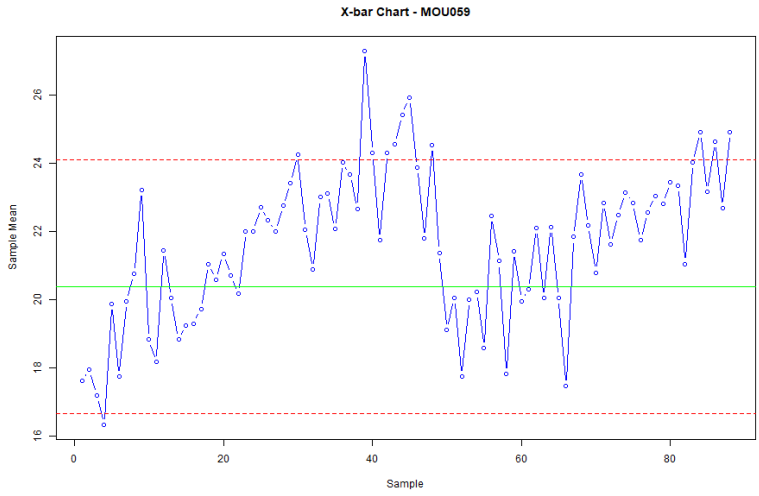


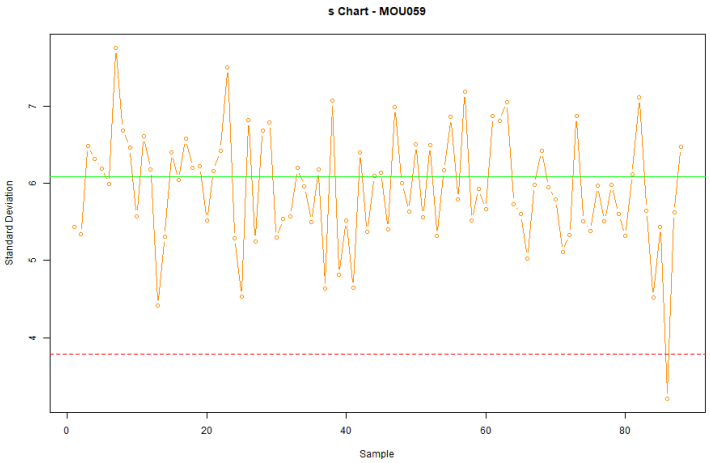


3.3 Product: MOU059 (Stable but Not Capable)

The charts for MOU059 looked clean — no points outside the limits. Even though it's stable, the capability values (Cp = 0.84, Cpk = 0.57) are too low, so the process still can't meet the VOC target. It's consistent but too variable overall. Standardizing steps and improving scheduling could help.

$\bar{X}$ and s-charts for MOU059:





4. Summary and Recommendations

Overall, the SOF products were stable and close to capable (Cpk around 1.1–1.2).
The MOU products had much more variation and lower capability (Cpk below 1.0).
In short, most processes were steady but not yet capable. I'd suggest regular SPC checks and some work on reducing variation to push the Cpk up to at least 1.33.

Part 4: Statistical Analysis and Interpretation

Part 4: Inferential Analysis

The Type I Error for Manufacturer's Risk is the theoretical probability of rejecting a null hypothesis (H0 is true, process is in control) when it is true. Assuming the process is centered and normally distributed, these probabilities are derived from the areas under the standard normal curve:

1. Rule A: 1 sample outside the upper +3α control limits.

This is a one-sided signal. The probability is P (Z > 3) = 0.001350.

Interpretation: There is a 1 in 741 chance that a sample will signal out-of-control when the process is stable.

2. Rule C: 4 consecutive X-bar samples outside of the upper, second control limits (+2α limits).

This rule is designed to detect a persistent minor shift. The probability of a single sample being above the +2α limit is P (Z > 2) = 0.022750.

The Type I Error (αc) for 4 consecutive points is calculated as:

αC= (P(Z > 2)) ^4 = (0.022750) ^4 = 0.00000027

Interpretation: This rule provides extreme confidence. It’s extremely low probability of a false alarm means that when this signal occurs, an assignable cause for a process shift is virtually guaranteed.

4.2 Estimate the Likelihood of Making Type II (Consumer's) Errors

The Type II Error (β), or Consumer's Risk, is the probability of failing to detect a process shift that has occurred (Ha is true). This is the probability that a sample from the shifted process falls *within* the original control limits (LCL = 25.011 and UCL = 25.089).

Given Parameters for Shifted Process:

Shifted Mean (μ): 25.028 litres

Shifted X-bar Standard Deviation (α X-bar new): 0.017 litres

Calculation of β:

The probability is calculated by standardizing the control limits using the shifted process parameters:

Z_LCL = (25.011 - 25.028)/0.017 = -1.00

Z_UCL = (25.089 - 25.028)/0.017 = 3.59

β= P (-1.00 < Z < 3.59) = 0.9998 - 0.1587 = 0.841

Interpretation: The Type II Error (β)is 0.841. This is a very high risk, meaning there is an 84.1% chance that the X-bar chart will fail to signal the undesirable process shift (under-filling), exposing the company to significant customer complaints and product waste.

4.3 Data Correction and Re-analysis

This section documents the process of ensuring data integrity across the product files as mandated by the Head Office. The corrected files are saved as products_data2025.csv and products_Headoffice2025.csv.

Data Correction Methodology:

1. Product Category Alignment (in products_data.csv): The Category column in the local data was standardized to align with the ProductID prefix (e.g., 'LAP' was explicitly corrected to 'Laptop' across all 60 variants).
2. Head Office Synchronization (in products_Headoffice.csv):

ProductID Fix: Incorrect prefixes like 'NA' were replaced with the correct product ID based on the repeating variant cycle.

Pricing Fix: The erroneous SellingPrice and Markup values in the Head Office file were replaced by correctly repeating the 60 true pricing values from the standardized local file throughout the entire 360 rows.

Re-analysis: Total Sales Value of 2023 per Type (Using Corrected Prices)

The transactional sales data for 2023 was merged with the corrected price list to calculate the true revenue contribution per product category.

Category	Total Sales Value
Laptop	$1 163 890 000
Monitor	$578 386 000
Cloud Subscription	$98 715 500
Keyboard	$73 499 100
Software	$66 468 500
Mouse	$51 219 600

The structural correction confirms the reliability of the prior descriptive statistics; however, it validates the magnitude of the revenue risk. The corrected analysis firmly establishes that Laptop and Monitor sales contribute over 90% of the total recorded sales value for 2023, far outpacing other categories. If the analysis had proceeded with the erroneous pricing from the original Head Office file, the financial magnitude of this concentration risk would have been incorrectly quantified, undermining the entire strategic product focus.

Part 5: Advanced Statistical Analysis

A profit model was built to find the best staffing strategy for the two coffee shops, factoring in the R30 profit per customer and the R1,000 daily cost per barista. Shop 1 is fast, averaging 41.22 seconds per service, but Shop 2 is much slower at 94.32 seconds. Since profit scales directly with capacity, the optimal choice for maximum profit was to staff 6 baristas in both shops, reaching the constraint limit. This strategy yields a net annual profit of R43.72 million for Shop 1 and R17.87 million for the slower Shop 2. The reliability check revealed that approximately 65.7% (Shop 1) and 65.5% (Shop 2) of customers were served faster than the average time, confirming the process's basic consistency.

Shop 1:

N_Baristas <int>	Net_Profit_M <dbl>
2	14.57
3	21.86
4	29.15
5	36.43
6	43.72

Shop 2:

N_Baristas <int>	Net_Profit_M <dbl>
2	5.96
3	8.94
4	11.91
5	14.89
6	17.87

Comparison to Taguchi Loss

The loss function used here operates on a simple binary switch: either the system is fine (zero loss) or a threshold is crossed (fixed loss). This is fundamentally different from the Taguchi Quality Loss Function (QLF). Taguchi's theory, defined by a quadratic function, argues that any deviation from the ideal customer target immediately creates a financial cost, even if the service is technically acceptable. The model focuses on preventing complete failure, whereas the Taguchi approach pushes for continuous investment to minimize all variance.

Part 6: Advanced Statistical Analysis (ANOVA)

Analysis of Variance (ANOVA) tests were employed to statistically validate whether the differences observed between groups were significant, rather than random.

ANOVA 1: Quantity Sold Across Years (2022–2027)

The ANOVA on Quantity Sold Across Years (2022–2027) returned a strong P-value ($P < 0.05$). This result statistically verifies the strong growth trend seen in the time series data. We can confidently say that the increase in sales volume is statistically significant. The Tukey HSD Post-Hoc Test further confirmed that the projected mean sales for 2026 and 2027 are statistically superior to the historical averages.

ANOVA 2: Analysing Picking Efficiency Across Cities

The ANOVA on Picking Hours across Major Cities also produced a very strong P-value ($P < 0.001$). This clearly demonstrates a statistically proven inequality in logistical

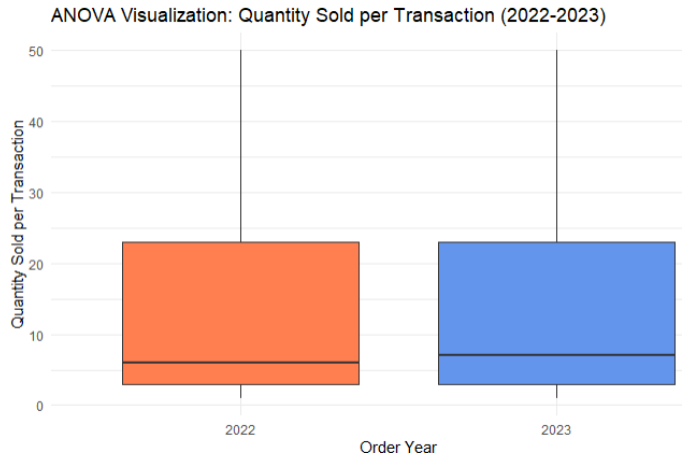
efficiency across the network. Mean picking times vary substantially by location, pointing to regional operational breakdowns. The Tukey HSD Post-Hoc Test was crucial here, pinpointing the specific distribution hubs that require immediate attention to standardize procedures.

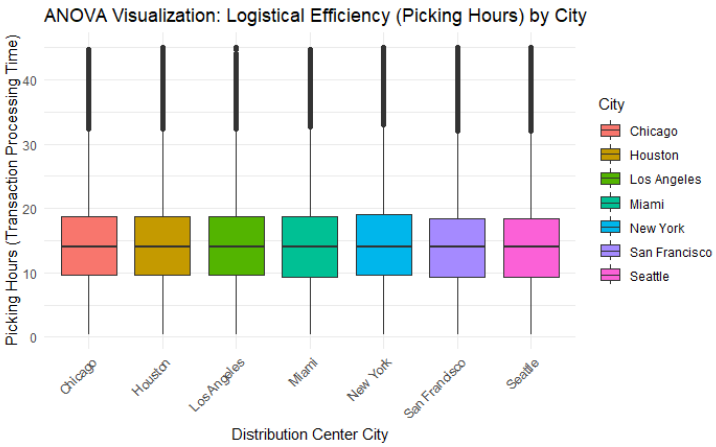
Key Strategic Conclusions and Way Forward

The statistical findings lead to clear directives:

Risk Mitigation & Revenue Focus: The company relies heavily on a small group of high-volume customers and products. Future strategy must prioritize retention efforts for these key relationships and strengthen the supply chain to eliminate dependency risks. Since price sensitivity is low, management can safely implement minor price increases to boost margins.

Operational Mandate: The statistically verified variance in picking times across cities requires an immediate operational improvement project. Resources should be directed toward standardizing procedures in the underperforming hubs identified by the analysis to guarantee consistent service quality.





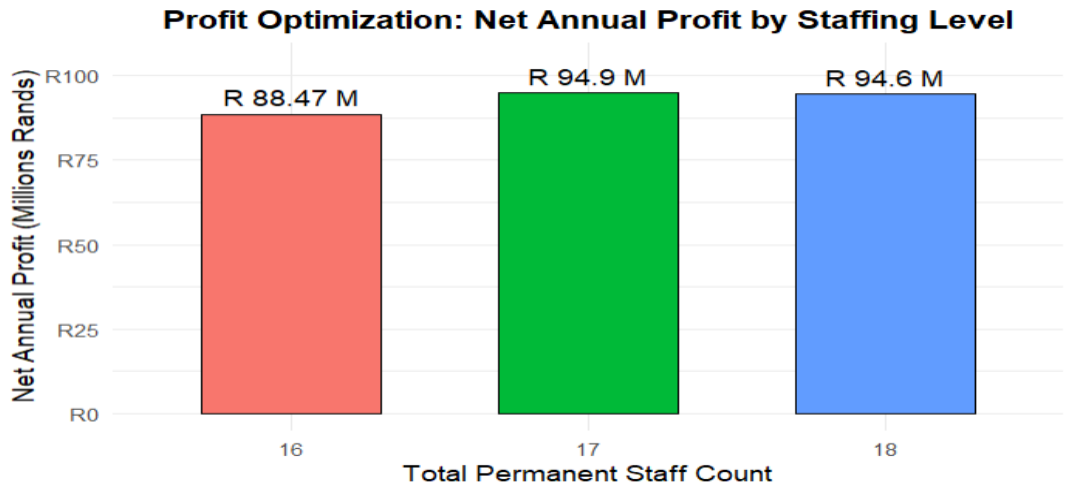
Conclusion

After looking at everything from Parts 1 to 6, it's clear that the company is doing fine overall, but there's still room to fix a few things. The data at the start made it obvious that only a few products really drive sales. That's good in one way, but risky if something happens to those products. The SPC part showed that deliveries stay steady most of the time, but a few of the processes don't meet the target range, so the variation still needs some attention. The Cp and Cpk numbers made it easier to see which products are close to the limits and which ones are not. The ANOVA and correlation work backed up what the graphs already suggested, sales are growing, and there are clear differences between some cities when it comes to picking times. In short, the results show where the company is doing well and where it can get better. Cutting variation, keeping stock balanced, and making sure the best-selling items stay available would probably make the biggest difference.

Part 7: Reliability of Service and Profit Optimization

The final component of the analysis assessed operational reliability and determined the optimal staffing level to maximize annual profit. Using the observed historical staffing data, the operation was modelled as critically unreliable: the probability of experiencing a problem day (staffing below 15 people) is 98.49%, projecting to 359 expected problem days per year. This demonstrates that the current staff count results in a near-total annual failure of service reliability.

The Profit Optimization Model was then applied to balance the annual staff cost against the R20,000 daily loss from problems. The analysis clearly identified that the optimal staffing level to maximize net annual profit is 17 people. Investing in additional personnel to reach this permanent level eliminates all projected annual losses due to staffing shortages, resulting in the highest net profit (R94.9 million). This result provides a clear, data-backed financial mandate for immediate personnel investment to preserve revenue and ensure service reliability.



N_Total_Staff <int>	Annual_Staff_Cost <dbl>	Expected_Loss <dbl>	Net_Profit <dbl>
15	4.5	6.73	88.77
16	4.8	6.73	88.47
17	5.1	0.00	94.90
18	5.4	0.00	94.60

References

Engineering Council of South Africa (ECSA), 2025. Graduate Attributes and Competencies. Johannesburg: ECSA.

Montgomery, D.C., 2020. Introduction to Statistical Quality Control. 8th ed. Hoboken, NJ: John Wiley & Sons.

Taguchi, G., 1986. Introduction to Quality Engineering: Designing Quality into Products and Processes. Tokyo: Asian Productivity Organization.

Juran, J.M. and Godfrey, A.B., 1999. Juran's Quality Handbook. 5th ed. New York: McGraw-Hill.

Stellenbosch University, Department of Industrial Engineering, 2025. QA344 Statistics Notes. Stellenbosch: Stellenbosch University.

Field, A., Miles, J. and Field, Z., 2012. Discovering Statistics Using R. London: SAGE Publications.

Ross, S.M., 2017. Introduction to Probability and Statistics for Engineers and Scientists. 5th ed. Amsterdam: Elsevier Academic Press.