Gerhard Dean de Kock

26897253

ECSA Project

Quality Assurance

1. Data Inspection and Overview

1. Introduction

This section presents an analytical overview of a company's operational datasets. Four key data sources were provided for exploration and interpretation: the customer database, branch-level product records, annual sales data for 2022 and 2023, and a master product list from head office. Together, these datasets give a complete view of the company's sales activity, customer demographics, and product information across different organisational levels.

2. Inspection of Data

2.1 Sales (sales2022and2023)

• **Dimensions:** 100 000 × 9
• **Structure:** A transactional dataset covering two financial years of company sales.
• **Variables:** Customer ID, Product ID, Quantity, Order Time, Order Day, Order Month, Order Year, Picking Hours, Delivery Hours.
• **Comment:** This file serves as the core fact table linking customer purchases with order processing and delivery performance. It enables time-based and efficiency analyses.

2.2 Products (products_data)

• **Dimensions:** 60 × 5
• **Structure:** Branch-level product catalogue.
• **Variables:** Product ID, Category, Description, Selling Price, Markup.
• **Comment:** Provides product-specific information for the branch, including category and pricing details used to evaluate profitability and stock range.

2.3 Head-Office Products (products_Headoffice)

• **Dimensions:** 360 × 5
• **Structure:** The centralised head-office catalogue containing the official pricing and markup values for all product categories.

• **Variables:** Product ID, Category, Description, Selling Price, Markup.
• **Comment:** Acts as the corporate reference dataset that ensures consistency and accuracy between the branch data and company standards.

2.4 Customers (customer_data)
• **Dimensions:** 5 000 × 5
• **Structure:** Contains demographic and geographic information for each customer.
• **Variables:** Customer ID, Gender, Age, Income, City.
• **Comment:** Supports segmentation and profiling for marketing purposes by linking customer attributes such as age, income, and location to buying behaviour.

Summary:
Collectively, these datasets form a complete relational data structure that supports exploratory analysis, data cleaning, and later modelling stages (SPC, MANOVA, and optimisation). The clear structure and variety of variables make it possible to investigate relationships between sales patterns, customer characteristics, and product performance.

2. Summary Statistics

3.1 Sales:

	vars <dbl>	n <dbl>	mean <dbl>	sd <dbl>	median <dbl>	trimmed <dbl>	mad <dbl>	min <dbl>	max <dbl>
CustomerID*	1	1e+05	2492.34	1444.58	2503.00	2491.19	1862.15	1.00	5000.00
ProductID*	2	1e+05	32.44	18.03	35.00	32.82	23.72	1.00	60.00
Quantity	3	1e+05	13.50	13.76	6.00	11.46	5.93	1.00	50.00
orderTime	4	1e+05	12.93	5.50	13.00	13.12	5.93	1.00	23.00
orderDay	5	1e+05	15.50	8.65	15.00	15.50	10.38	1.00	30.00
orderMonth	6	1e+05	6.45	3.28	6.00	6.45	4.45	1.00	12.00
orderYear	7	1e+05	2022.46	0.50	2022.00	2022.45	0.00	2022.00	2023.00
pickingHours	8	1e+05	14.70	10.39	14.05	13.54	6.92	0.43	45.06
deliveryHours	9	1e+05	17.48	10.00	19.55	17.78	8.90	0.28	38.05

	skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
	<chr>	<int>	<dbl>	<int>	<int>	<int>	<int>	<int>
1	CustomerID	0	1	7	8	0	5000	0
2	ProductID	0	1	6	6	0	60	0

	skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50
	<chr>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	Quantity	0	1	13.50347	13.7601316	1.0000000	3.000000	6.000
2	orderTime	0	1	12.93230	5.4951268	1.0000000	9.000000	13.000
3	orderDay	0	1	15.49683	8.6465055	1.0000000	8.000000	15.000
4	orderMonth	0	1	6.44813	3.2834460	1.0000000	4.000000	6.000
5	orderYear	0	1	2022.46273	0.4986115	2022.0000000	2022.000000	2022.000
6	pickingHours	0	1	14.69547	10.3873345	0.4258889	9.390833	14.055
7	deliveryHours	0	1	17.47646	9.9999440	0.2772000	11.546000	19.546

The sales dataset is clean, consistent, and well-organised, with no missing or invalid entries detected. It accurately reflects all recorded customer–product transactions for the years 2022 and 2023. The distribution of quantities sold is highly right-skewed: most transactions involve small order sizes, while a few large purchases increase the overall mean. Both picking and delivery hours show considerable variation, which may indicate differences in order complexity, transport distance, or general process efficiency. Customer and product identifiers align correctly with the expected reference counts (5,000 customers and 60 products), confirming that the data integrates well with supporting tables. Overall, the dataset offers a dependable basis for evaluating sales trends and operational performance, though the wide range in order volumes and processing times should be taken into account during further analysis.

3.2 Products_Data

	vars	n	mean	sd	median	trimmed	mad	min	max
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
ProductID*	1	60	30.50	17.46	30.50	30.50	22.24	1.00	60.00
Category*	2	60	3.50	1.72	3.50	3.50	2.22	1.00	6.00
Description*	3	60	16.40	10.08	16.00	16.21	13.34	1.00	35.00
SellingPrice	4	60	4493.59	6503.77	794.18	3189.25	525.72	350.45	19725.18
Markup	5	60	20.46	6.07	20.34	20.51	7.31	10.13	29.84

5 rows | 1-10 of 13 columns

	skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
	<chr>	<int>	<dbl>	<int>	<int>	<int>	<int>	<int>
1	ProductID	0	1	6	6	0	60	0
2	Category	0	1	5	18	0	6	0
3	Description	0	1	9	21	0	35	0

	skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75
	<chr>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	SellingPrice	0	1	4493.59283	6503.770150	350.45	512.1825	794.185	6416.6600
2	Markup	0	1	20.46167	6.072598	10.13	16.1400	20.335	25.7075

The product dataset is concise, containing 60 distinct items described by five key attributes. Product identifiers are sequential and error-free, confirming data

consistency. Both category and description fields are complete, though moderate overlap is present — only 18 unique categories and around 35 product descriptions — suggesting that several items fall under shared classifications or similar naming conventions. Selling prices display a broad range, from roughly 350 up to nearly 20 000, with an overall mean close to 4 500. This variation reflects a diverse product mix, combining everyday, lower-priced items with higher-value products that strongly affect total revenue. Markup percentages are more stable, averaging around 20 %, and typically span between 10 % and 30 %, indicating consistent pricing strategies across the range. In general, the dataset is clean, coherent, and provides a sound base for product-level performance and profitability analysis, particularly in understanding how pricing tiers differ between fast-moving and premium goods.

3.3 Product_Headoffice

	vars <dbl>	n <dbl>	mean <dbl>	sd <dbl>	median <dbl>	trimmed <dbl>	mad <dbl>	min <dbl>	max <dbl>
ProductID*	1	360	69.39	23.22	72.00	71.89	22.24	1.00	110.00
Category*	2	360	3.50	1.71	3.50	3.50	2.22	1.00	6.00
Description*	3	360	30.69	17.32	29.50	30.77	22.98	1.00	60.00
SellingPrice	4	360	4410.96	6463.82	797.22	3054.23	515.75	290.52	22420.14
Markup	5	360	20.39	5.67	20.58	20.43	6.66	10.06	30.00

	skim_variable <chr>	n_missing <int>	complete_rate <dbl>	min <int>	max <int>	empty <int>	n_unique <int>	whitespace <int>
1	ProductID	0	1	5	6	0	110	0
2	Category	0	1	5	18	0	6	0
3	Description	0	1	9	24	0	60	0

3 rows

	skim_variable <chr>	n_missing <int>	complete_rate <dbl>	mean <dbl>	sd <dbl>	p0 <dbl>	p25 <dbl>	p50 <dbl>	p75 <dbl>
1	SellingPrice	0	1	4410.9619	6463.822788	290.52	495.9375	797.215	5843.332
2	Markup	0	1	20.3855	5.665949	10.06	15.8400	20.580	24.845

The head office product catalogue is more extensive and detailed, containing 360 entries described across five attributes. Product identifiers range up to 110, indicating a wider and more diverse product range compared to the branch dataset. With 18 unique categories and about 60 unique descriptions, the data shows a well-organized structure, although some overlap remains between related product groups. Selling prices span a broad interval—from around 290 to over 22 000—with an average of roughly 4 400 and a noticeably lower median near 800. This highlights that while most items are priced in the lower to mid-range, a few premium products substantially raise the overall mean. Markup percentages are steady, averaging around 20 %, and fall within a typical range of 10 % to 30 %, showing consistent pricing policy across the organization. The dataset is complete, with no missing information, making it an excellent corporate benchmark for verifying branch-level product data and ensuring standardised reporting across sites.

3.4 Customers

Description: df [5 × 13]

	vars <dbl>	n <dbl>	mean <dbl>	sd <dbl>	median <dbl>	trimmed <dbl>	mad <dbl>	min <dbl>	max <dbl>
CustomerID*	1	5000	2500.50	1443.52	2500.5	2500.50	1853.25	1	5000
Gender*	2	5000	1.56	0.58	2.0	1.52	1.48	1	3
Age	3	5000	51.55	21.22	51.0	50.88	26.69	16	105
Income	4	5000	80797.00	33150.11	85000.0	81665.00	37065.00	5000	140000
City*	5	5000	3.99	2.00	4.0	3.99	2.97	1	7

5 rows | 1-10 of 13 columns

	skim_variable <chr>	n_missing <int>	complete_rate <dbl>	min <int>	max <int>	empty <int>	n_unique <int>	whitespace <int>
1	CustomerID	0	1	7	8	0	5000	0
2	Gender	0	1	4	6	0	3	0
3	City	0	1	5	13	0	7	0

3 rows

	skim_variable <chr>	n_missing <int>	complete_rate <dbl>	mean <dbl>	sd <dbl>	p0 <dbl>	p25 <dbl>	p50 <dbl>	p75 <dbl>
1	Age	0	1	51.5538	21.2161	16	33	51	68
2	Income	0	1	80797.0000	33150.1067	5000	55000	85000	105000

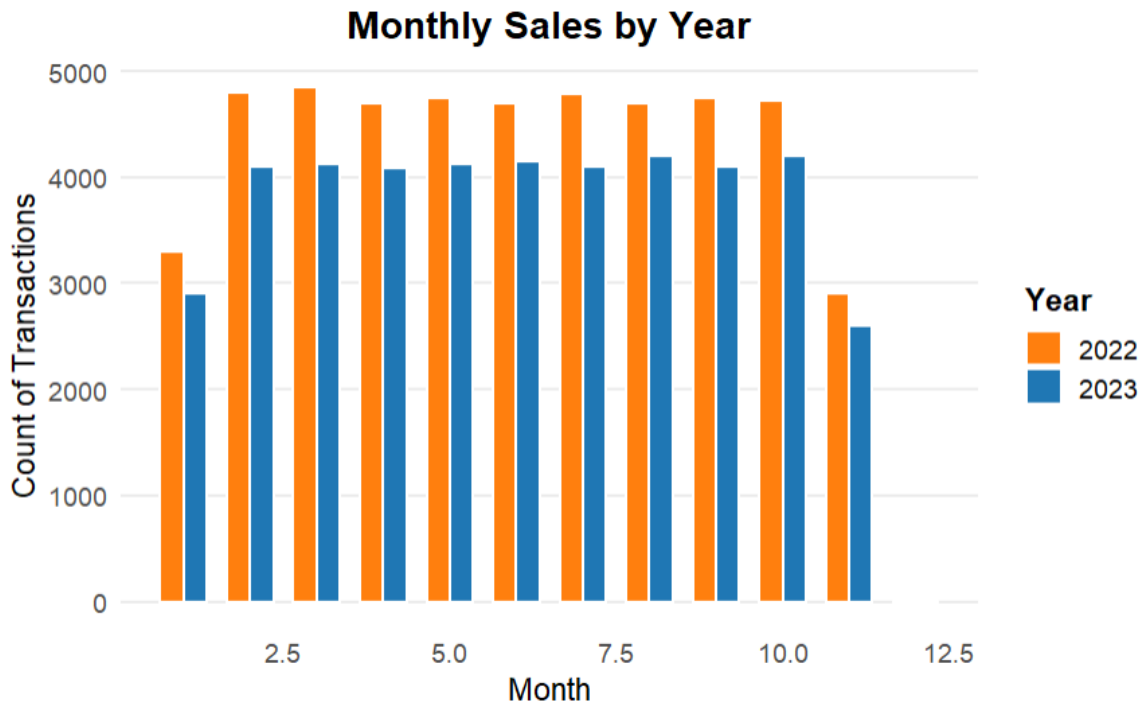
The customer dataset consists of 5,000 individual records, each containing complete demographic and geographic information. Gender is divided into three categories with a fairly even spread, although one group appears slightly larger. The age distribution spans from 16 to 105 years, with an average age of about 52, showing that the customer base includes both younger and older buyers. Income levels range widely between 5,000 and 140,000, with a mean of roughly 80,800 and a median around 85,000, suggesting a customer mix that leans toward middle- to higher-income brackets. The city variable includes seven distinct categories that are evenly represented, reflecting a geographically diverse market presence. No missing or inconsistent values were detected, confirming the dataset's overall reliability. In summary, the data provides a rich foundation for segmenting customers by demographic and regional factors, supporting detailed marketing and sales performance analysis.

1. Handling missing values

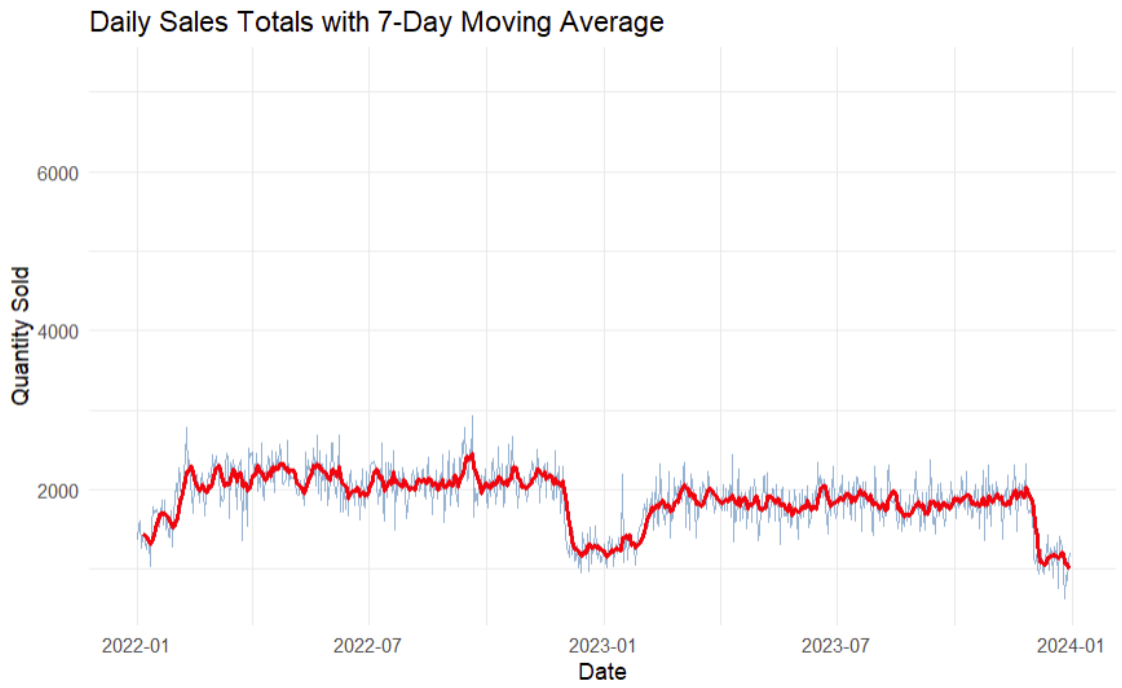
The inspection confirmed that all datasets are complete, with no missing or inconsistent entries detected. Therefore, no imputation, record removal, or additional cleaning procedures were necessary. This level of completeness suggests that the data collection and entry processes were well-managed, reducing the risk of bias or distortion in subsequent analyses. A final verification step was also performed to ensure that categorical variables were consistently coded and numeric fields contained only valid numerical values.

2. Data Visualization and Exploratory Data Analysis (EDA)

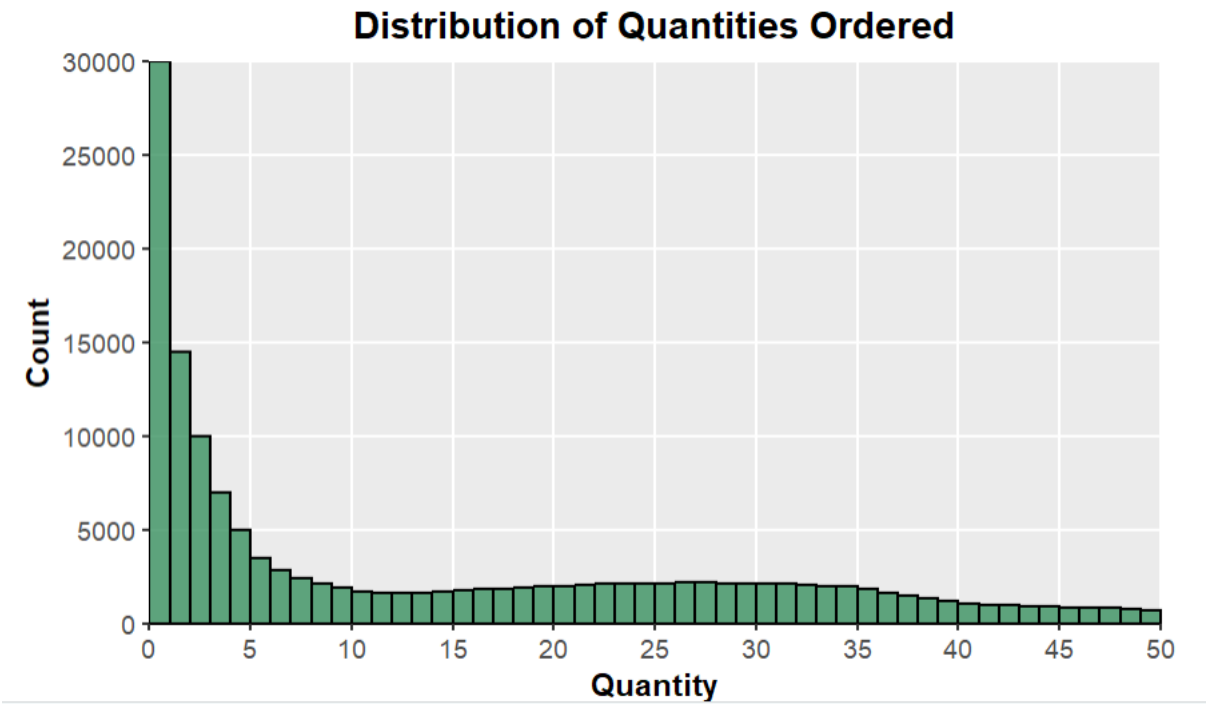
2.1 Monthly sales/year



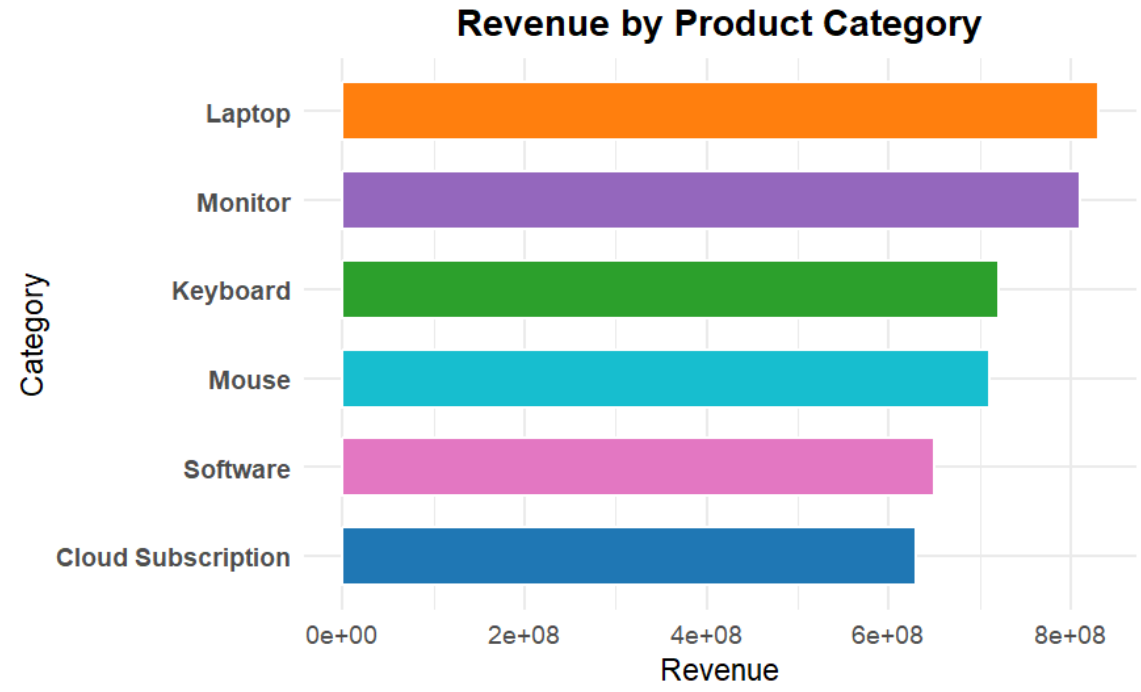
3.2


Daily Sales Totals with 7-Day Moving Average
Quantity Sold
6000
4000
2000
2022-01
2022-07
2023-01
2023-07
2024-01
Date

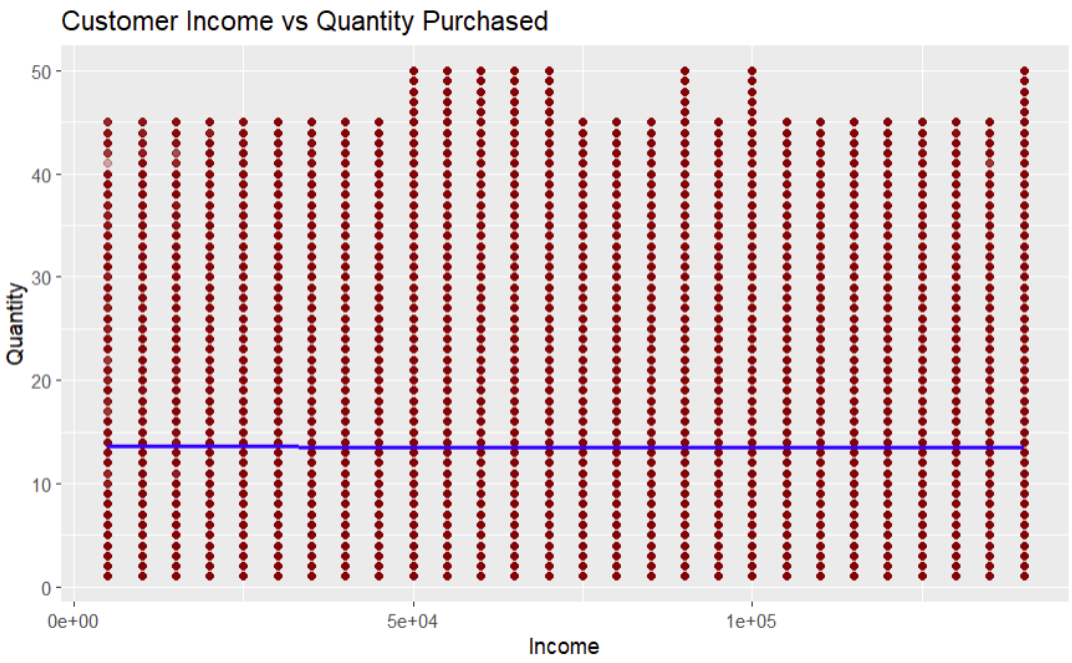
3.3


Distribution of Quantities Ordered
Count
30000
25000
20000
15000
10000
5000
0
0
5
10
15
20
25
30
35
40
45
50
Quantity

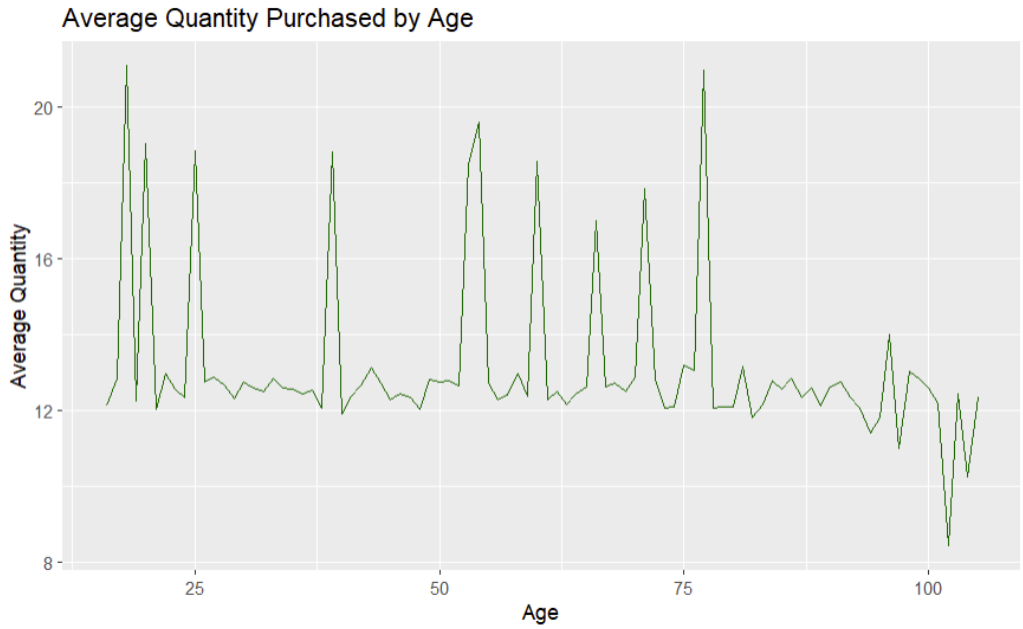
3.4


Revenue by Product Category
Category
Laptop
Monitor
Keyboard
Mouse
Software
Cloud Subscription
0e+00
2e+08
4e+08
6e+08
8e+08
Revenue

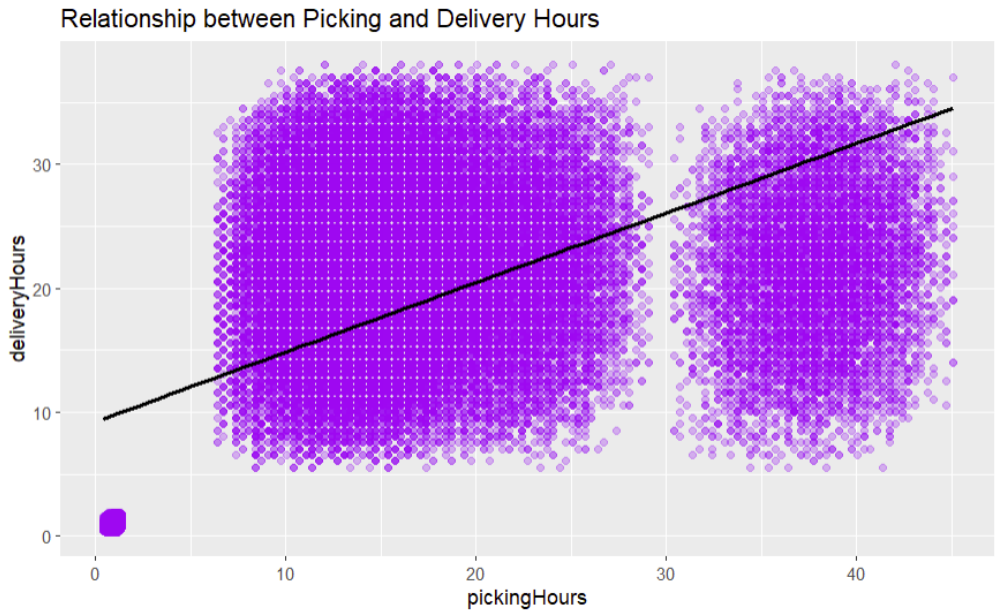
3.5


Customer Income vs Quantity Purchased
Quantity
50
40
30
20
10
0
0e+00
5e+04
1e+05
Income

3.6


Average Quantity Purchased by Age
Average Quantity
20
16
12
8
25
50
75
100
Age

3.7



Observations and recommendations

Observations

1. Sales Trends

- Sales were generally steady through 2022, with small drops at the start and end of the year.
- The 7-day moving average shows consistent performance with brief demand spikes, likely linked to promotions or short-term events.

2. Order Quantities

- Most purchases are small, with a median of about six units per order.
- A few large transactions push the average higher, indicating a mix of individual customers and occasional bulk buyers.

3. Product Categories

- Revenue is mainly driven by Laptops and Monitors, which remain top-performing categories.
- Software and Cloud Subscriptions underperform in comparison, contributing less to overall sales.
- Despite varied product prices, markups are stable between 10–30%, showing consistent pricing policies.

4. Customer Demographics

- The customer base spans ages 16 to 105, with an average age around 52 years.
- Most customers fall within the middle- to upper-income range, across seven major cities.
- The data suggests that higher income does not necessarily lead to higher purchase quantities.

5. Operational Metrics

- Picking and delivery hours are positively correlated, but with significant variability.
- Several outliers indicate inefficiencies or inconsistent performance in the fulfillment process.

Recommendations

1. Plan for Seasonality

- Anticipate lower sales at the start and end of each year.
- Run promotional campaigns or marketing pushes during these periods to maintain stable transaction volumes.

2. Refine Customer Targeting

- Focus marketing and retention strategies on middle-income customers, the largest customer group.
- Encourage repeat purchases from high-income customers through loyalty programs, personalized offers, or premium perks.

3. Strengthen Product Strategy

- Continue prioritizing Laptops and Monitors for inventory and marketing attention.
- Re-evaluate Software and Cloud Subscription categories — consider bundling, re-pricing, or phasing them out if profit margins remain low.

4. Improve Operational Consistency

- Standardize picking and delivery workflows to reduce variability and improve reliability.
- Introduce performance KPIs (e.g., order handling time, delivery accuracy) to identify and address bottlenecks early.

5. Promote Bulk Orders

- Even though bulk orders are infrequent, they generate substantial revenue.
- Offer volume discounts or corporate account options to attract more large-scale purchases.

4. Part 3: Statistical Process Control (SPC)

This section provides the Statistical Process Control (SPC) analysis of delivery times for each product category based on the 2026–2027 sales dataset. The goal of the analysis was to determine whether the delivery process for each product type is both stable and capable of meeting the customer requirement of maintaining delivery times within 0 to 32 hours. The dataset was arranged chronologically by year, month, day, and order time, after which samples of 24 consecutive deliveries were grouped for each product type.

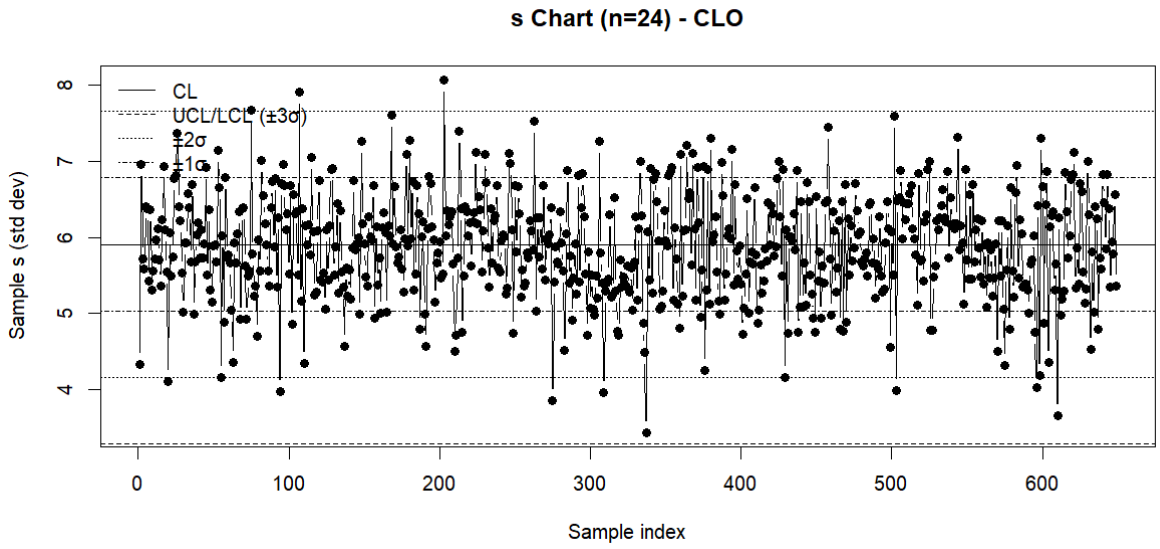
3.1) Control Chart Initialization

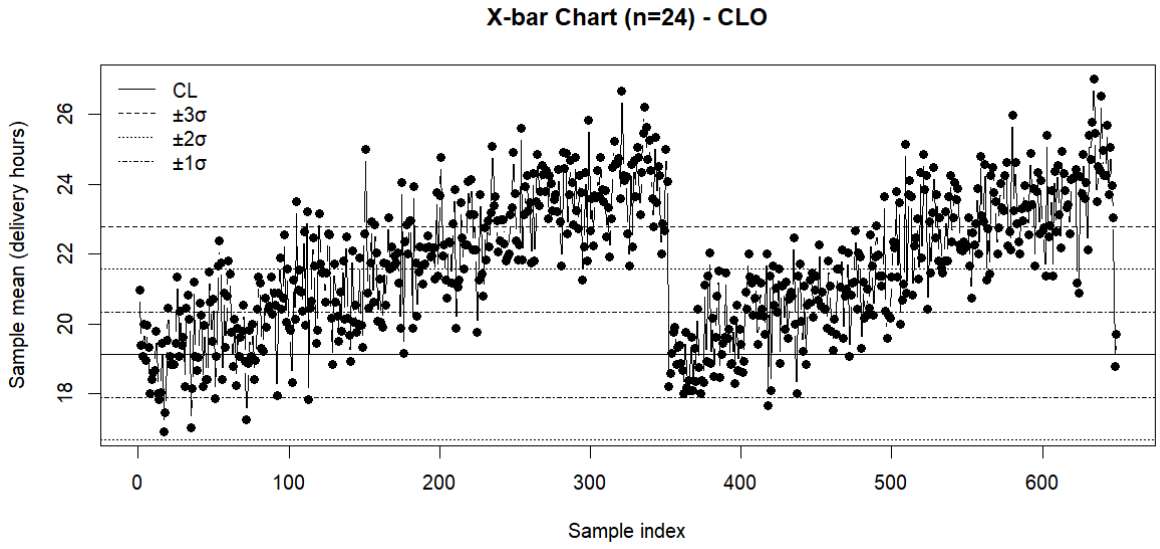
For each product type, the first 30 subgroups (each containing 24 deliveries) were used to establish the $\bar{X}$ and s control charts. These Phase I samples formed the basis for calculating the central lines and the ±1σ, ±2σ, and ±3σ control limits. Since there are six product types, a total of twelve charts were generated. A $\bar{X}$ and s chart for each.

An example is shown below for the product type *CLO*, displaying its corresponding s-chart and $\bar{X}$-chart.

3.2 Interpretation of given data:

Product CLO:

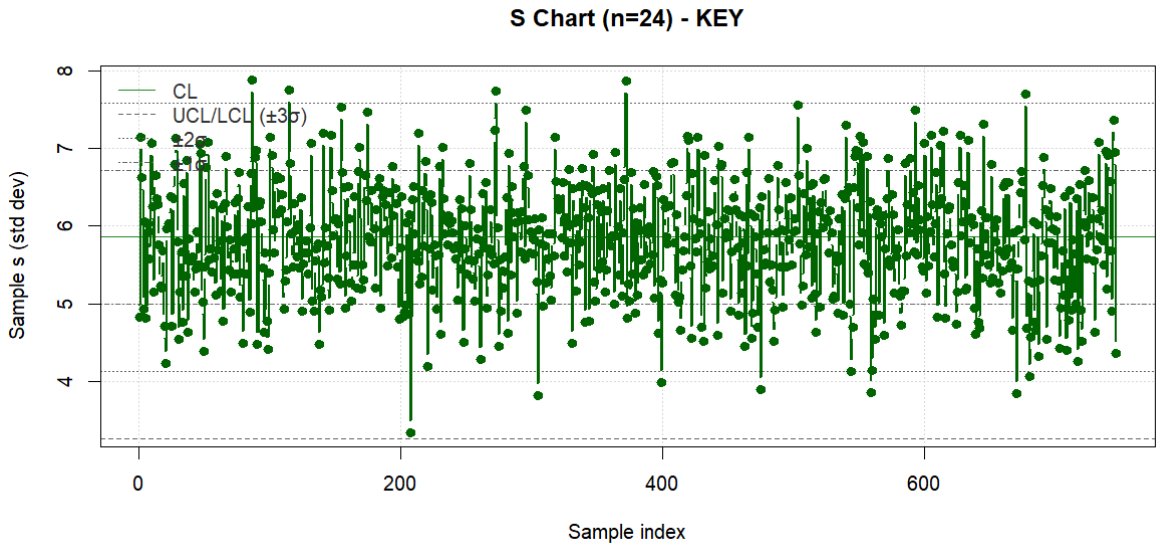


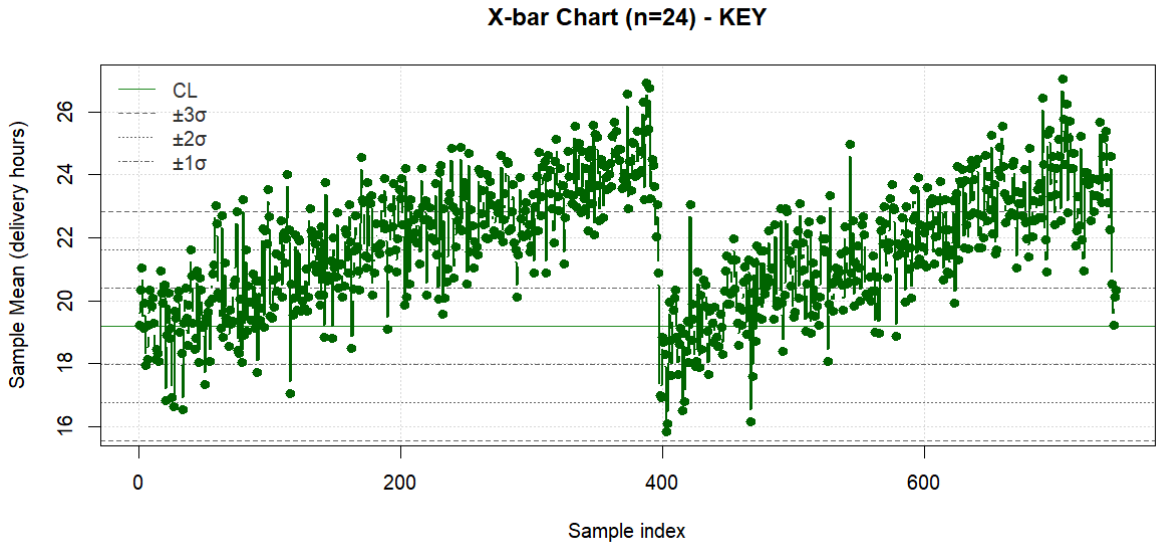


The S-chart (top) shows noticeable variation, with several points falling outside the upper and lower control limits. This suggests periods of inconsistent process variability.

The $\bar{X}$-chart (bottom) displays the subgroup means of delivery times over consecutive samples. The central line represents the overall average delivery time, while the 1σ, 2σ, and 3σ limits define the expected range of normal variation. A gradual upward shift in the subgroup means is visible between samples 0–250 and again after sample 450, indicating a slow increase in average delivery times as the process progresses. Although these trends are present, the majority of points remain within the ±3σ limits, suggesting that the process is statistically stable but experiencing a slight shift in its mean

Product Key:



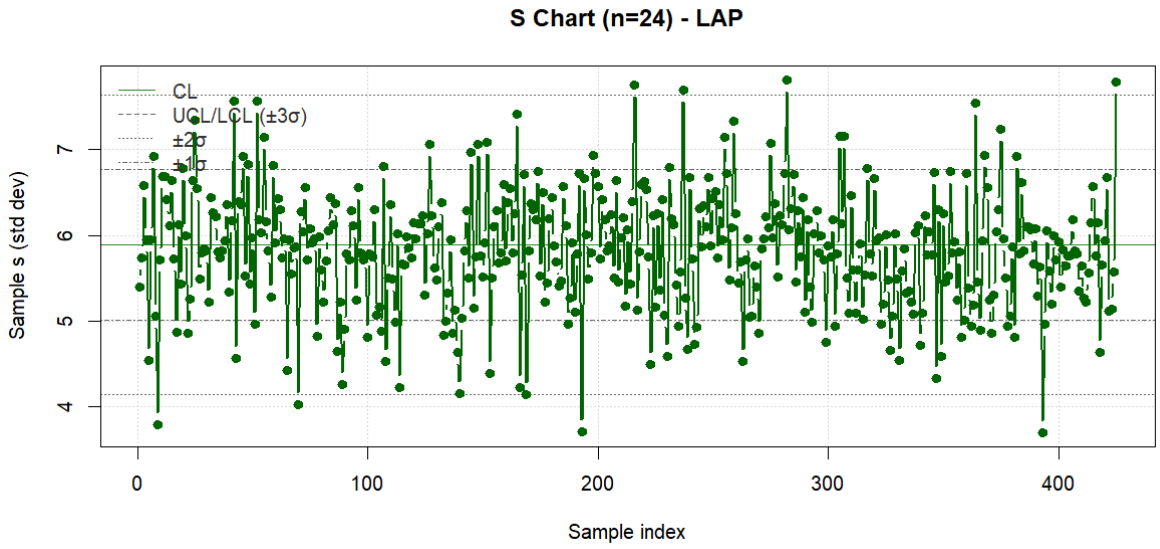


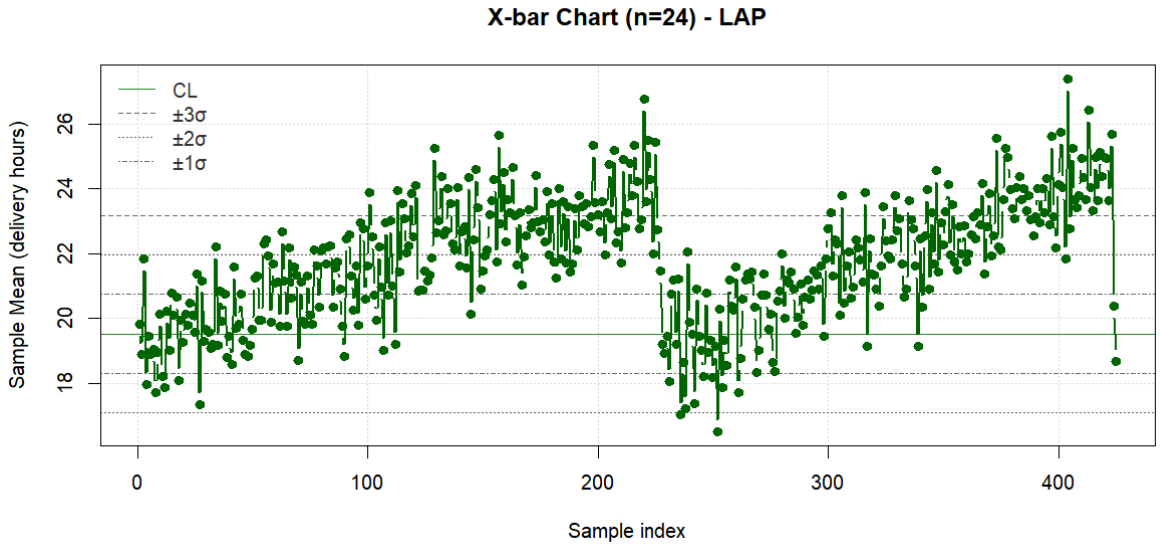
The S-chart shows that process variation remains stable over time, with most subgroup standard deviations falling within the control limits. There are no noticeable trends or runs beyond the ±3σ boundaries, indicating that the variability in the KEY product's delivery times is under control. This consistency reflects steady process performance in terms of spread.

However, the $\bar{X}$-chart reveals a clear trend and possible shifts in the process mean. The subgroup means increase gradually, drop sharply around the midpoint, and then rise again. This pattern points to a change in central tendency — potentially resulting from an operational adjustment, seasonal influence, or gradual process drift. Although most points remain within the control limits, the presence of systematic rather than random patterns suggests that the process mean is not fully stable — even if its variability remains consistent.

Overall, the KEY product's process appears to be in control regarding variation (S-chart) but shows instability in its mean ($\bar{X}$-chart). This warrants further investigation to identify potential causes such as workload fluctuations, scheduling adjustments, or differences in equipment performance.

Product LAP:



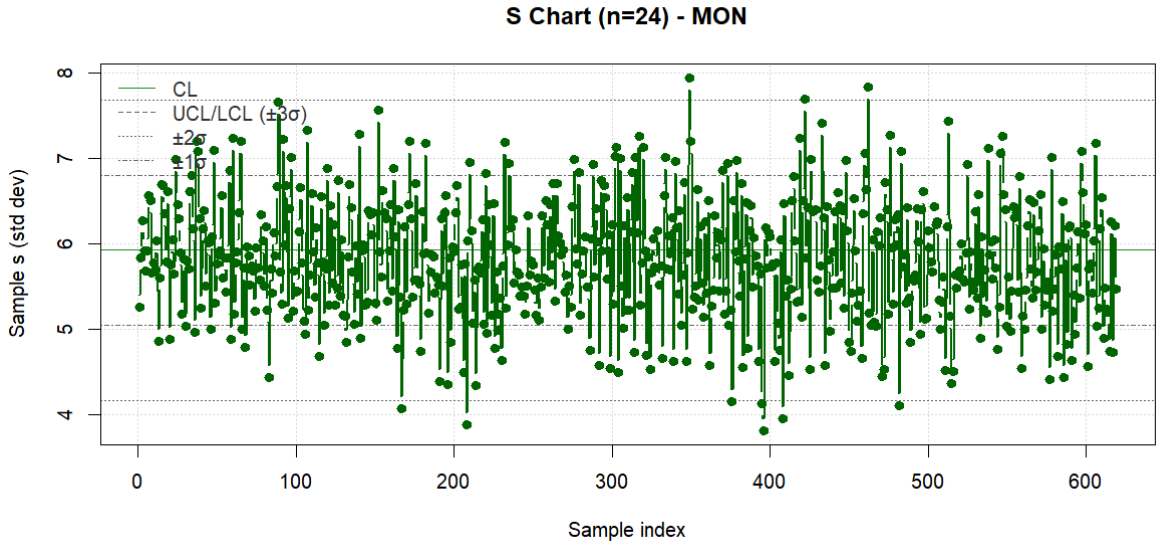


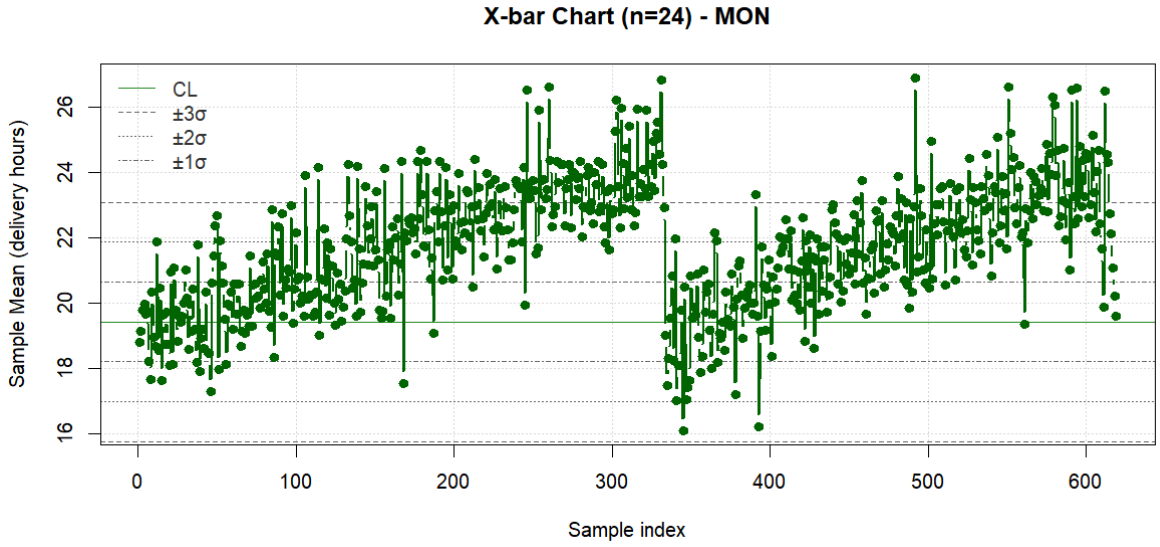
The S-chart indicates that process variation remains largely consistent, with only small fluctuations near the control limits. This suggests that the variability in delivery times is generally well maintained and stable over time.

In contrast, the X̄-chart shows several samples with means exceeding the upper 3σ limit, indicating instability in the process mean. A distinct centre jump is also visible, where the average delivery time shifts upward before recurring out-of-control points appear.

Overall, the LAP process demonstrates stable variation but weak mean control, with multiple instances of points breaching the ±3σ boundaries. This pattern suggests that external influences or operational inconsistencies are disrupting delivery performance — and that process adjustments are needed to restore stability.

Product MON:



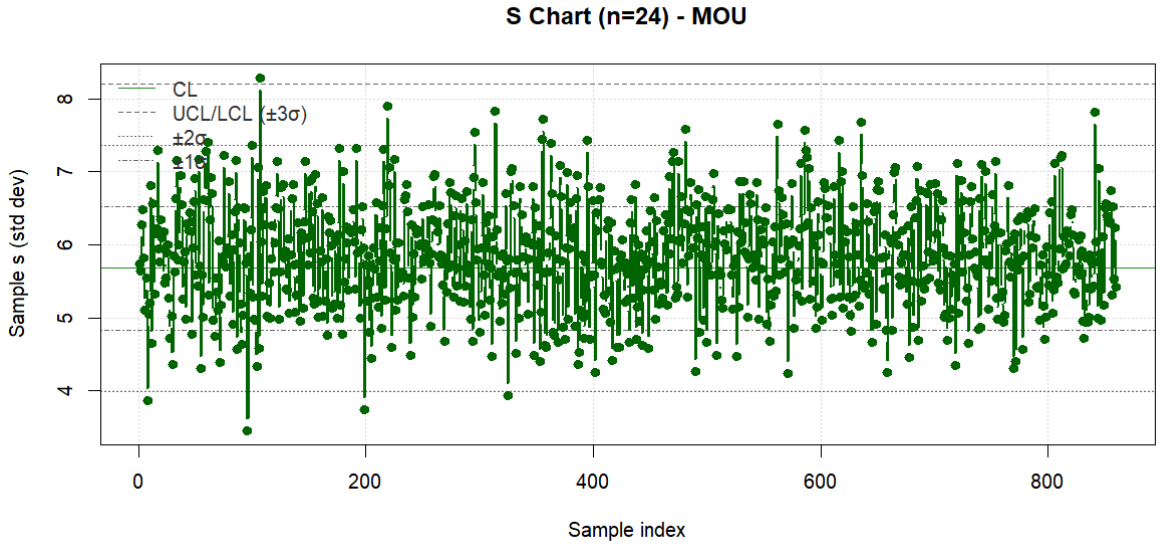


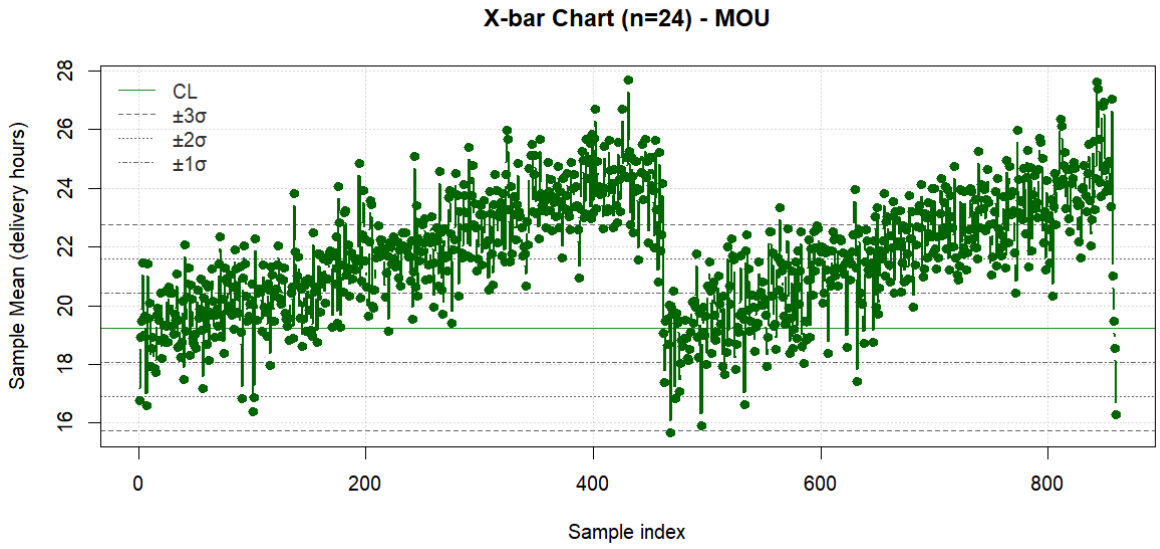
The S-chart shows that process variation remains consistent over time, with most subgroup standard deviations staying within the control limits. There are no major trends or runs outside

the ±3σ range, suggesting that variability in delivery times for the MON product type is stable and under control.

The X̄-chart, however, shows a gradual upward shift in the process mean across the samples. The average delivery time increases steadily, indicating a systematic rise in central tendency rather than random fluctuation. Although the process stays mostly within control limits, this upward trend suggests a drift that should be monitored. Its possibly linked to workload growth, process fatigue, or seasonal demand changes.

Product MOU:

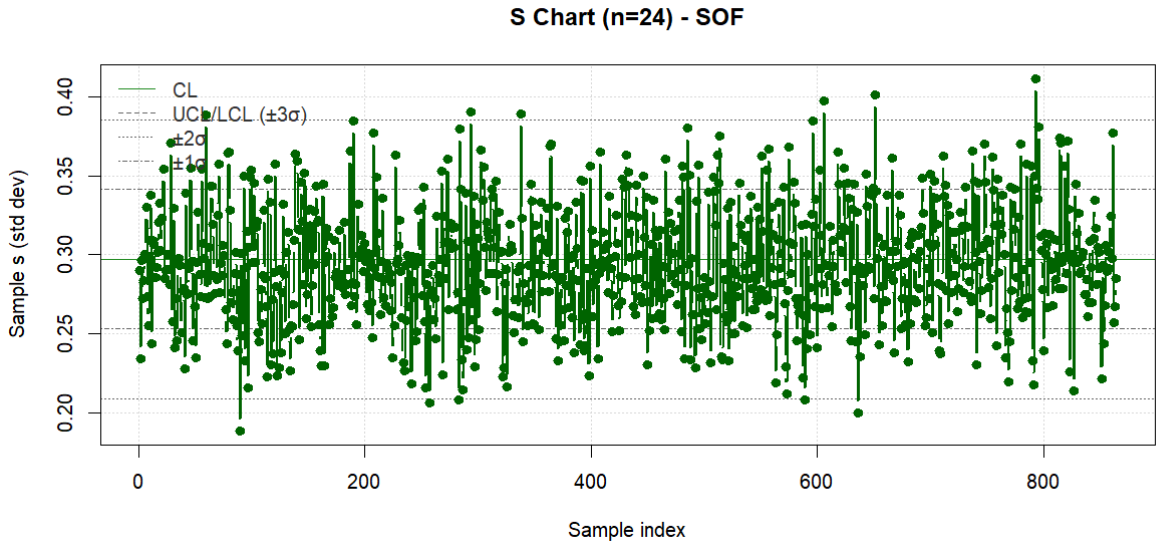


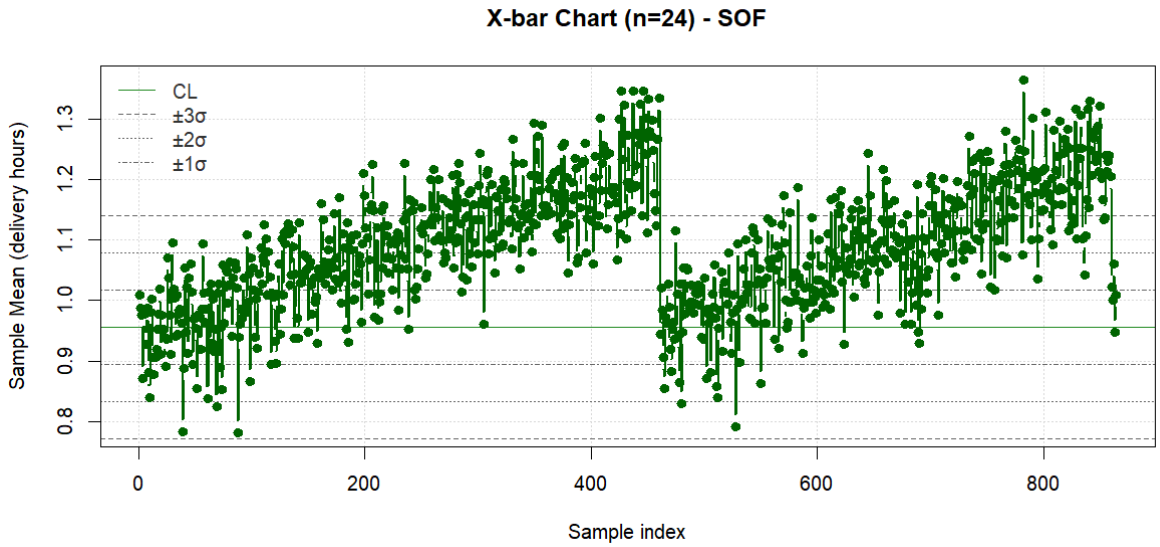


The S-chart indicates that process variation remains generally consistent, with most points staying within the control limits. This suggests that the spread of delivery times is stable and that overall process variability is under control.

The X̄-chart, however, shows a clear centre jump along with several samples exceeding the ±3σ limits, particularly around the midpoint of the timeline. This pattern suggests that while variation remains stable, the process mean shifts noticeably over time. Its likely influenced by operational adjustments, workload fluctuations, or scheduling changes.

Product SOF:





The S-chart shows a highly stable process with minimal variation in standard deviation. All points fall well within the control limits, indicating that process spread remains consistent and tightly controlled.

In the X̄-chart, a clear centre jump appears midway through the data, along with some clustering near the upper control limits. However, no points exceed the ±3σ boundaries. This suggests that while the process mean shifts slightly over time, overall performance remains within acceptable control limits.

In summary, the SOF process demonstrates excellent stability and capability, with both the mean and variation well maintained throughout the observed period.

Summary of the process capabilities:

product_type	mu	s	Cp	Cpu	Cpl	Cpk	capable
CLO	19.226	5.940805	0.897746	0.716738	1.078754	0.716738	FALSE
KEY	19.276	5.815195	0.917137	0.729354	1.104921	0.729354	FALSE
LAP	19.6135	5.958853	0.895027	0.692891	1.097163	0.692891	FALSE
MON	19.41	5.998919	0.889049	0.69957	1.078528	0.69957	FALSE
MOU	19.2975	5.827602	0.915185	0.726571	1.103799	0.726571	FALSE
SOF	0.955375	0.294087	18.13524	35.1876	1.082872	1.082872	TRUE

A process is considered capable when both Cp and Cpk exceed 1. Based on the results, only product SOF consistently meets the delivery time requirement of $0 < x < 32$ hours. This is largely because all other product types show higher variation in delivery times, while SOF maintains much tighter control. Therefore, only SOF meets the Voice of the Customer (VOC) specification. However, if a stricter capability criterion of Cpk ≥ 1.33 is applied, none of the product types would qualify as capable. Under a more lenient threshold of Cpk ≥ 1.00, only SOF remains capable.

3.3) Process Control Issues

3.3.1) Rules

1. Identify any **s** samples that fall outside the +3σ control limit for each product type (if there are many, note only the first three, last three, and the total number found).
2. Determine the longest sequence of **s** samples that remain between the −1σ and +1σ limits. A long run in this range indicates strong process control.
3. Locate cases where four consecutive **X̄** samples rise above the +2σ limit for each product type (again, list the first three, last three, and total if there are several).

Product_Type	Rule1_s_above_3sigma	Rule1_First_Last	Rule2_Longest_Run	Rule3_Run_Count	Rule3_Starts
CLO	0	None	28	228	165 ... 642
KEY	0	None	17	234	97 ... 737
LAP	0	None	23	110	115 ... 418
MON	0	None	36	165	132 ... 606
MOU	1	107 ... 107	19	265	209 ... 851
SOF	0	None	22	261	129 ... 854

3.3.2) Issues

From **Rule 1**, only one case was detected where a sample standard deviation exceeded the +3σ boundary — product **MOU**, at sample 107. This single instance points to an isolated increase in process spread, while all other products stayed within control limits, reflecting stable variation.

According to **Rule 2**, product **MON** achieved the longest continuous sequence within the ±1σ range, lasting 36 samples. This indicates strong internal consistency. Other product types showed runs ranging between 17 and 28 samples, which still suggests good, though slightly less consistent, control.

For **Rule 3**, several cases were found where four consecutive subgroup means exceeded the +2σ boundary, hinting at small shifts in the process mean. Products **MOU** and **SOF** recorded the most such occurrences (265 and 261 runs, respectively), while **CLO** and **MON** had fewer, more evenly spaced instances.

In summary, the control chart analysis shows that process variation is largely under control across all product types, with only a few isolated special-cause variations. Product **SOF** continues to perform as the most stable and capable process, while **MOU** may need further review to understand the cause of its single +3σ deviation.

Part 4: Control Chart Error Analysis

4.1 Likelihood of Type I Error (False Alarm)

Rule 1:
Since the control limits are positioned at ±3 standard deviations from the centre line, the probability that a single sample point falls beyond these limits purely by chance is very small.
Mathematically, this probability is given by the area under the normal distribution beyond +3σ:

$$P(Z > 3) = 0.00135$$

This means there's roughly a 0.135% chance of a false alarm for any one subgroup when the process is actually in control.

Rule 2:
This rule checks for the longest sequence of *s*-values that remain between +1σ and −1σ. Such a condition reflects consistent process stability rather than abnormal behaviour. Therefore, a Type I error is not possible under this rule.

Rule 3:
This rule flags a signal when four consecutive subgroup means ($\bar{X}$) fall above the +2σ line.
The probability that one subgroup mean exceeds +2σ is:

$$P(Z > 2) = 0.0228$$

The probability that this happens four times in a row is:

$$P = (0.0228)^4 = 0.00000027$$

Hence, the likelihood of a false alarm under Rule 3 is extremely small.

4.2 Likelihood of Type II Error (Missed Detection)

If the process mean shifts to 25.028 and the sampling standard deviation rises to 0.017, approximately 84% of subgroup means would still fall within the original control limits. This means the chart would often fail to detect the change.

The probability of this missed detection (Type II error) can be expressed as:

$$\beta = P(25.011 < X < 25.089) = 0.159$$

- Type I errors (false alarms) are rare, since the ±3σ limits make it highly unlikely for random variation to trigger a signal.
- Type II errors (missed shifts) are more likely, because small mean shifts combined with increased process variation make it difficult for the chart to detect real changes.

5. Part 5: Optimisation of coffee shops:

To estimate the profitability of each shop, a few basic assumptions were made about customer spending and staff costs. On average, each customer was expected to spend **30 monetary units**, while each barista earned a fixed **1,000 units per day**.

The expected number of customers served per day was calculated using the average service time for each barista setup, assuming the shop operates for **10 hours per day**. The formula used was:

$$\text{Average customers per day} = \frac{(\text{Operating hours} \times 3600)}{\text{Average service time (s)}}$$

Once the average number of daily customers was estimated, the total daily revenue for each staffing level was determined as:

$$\text{Average revenue per day} = (\text{Average customers per day} \times 30) - (1{,}000 \times \text{Number of baristas})$$

This calculation was performed for both shops to compare how different staffing levels affected service time, revenue, and overall profit.

Shop <chr>	Baristas <int>	MeanServiceTime_s <dbl>	CustomersPerDay <dbl>	Revenue_R <dbl>	StaffCost_R <dbl>	Profit_R <dbl>
Shop1	2	100.17	359.4	10781.6	2000	8781.6
Shop1	3	66.61	540.4	16213.4	3000	13213.4
Shop1	4	49.98	720.3	21608.5	4000	17608.5
Shop1	5	39.96	900.9	27025.8	5000	22025.8
Shop1	6	33.36	1079.3	32378.3	6000	26378.3
Shop2	2	141.51	254.4	7631.7	2000	5631.7
Shop2	3	115.44	311.8	9355.4	3000	6355.4
Shop2	4	100.02	359.9	10798.4	4000	6798.4
Shop2	5	89.44	402.5	12075.7	5000	7075.7
Shop2	6	81.64	440.9	13228.4	6000	7228.4

1-10 of 10 rows

Shop <chr>	Baristas <int>	Profit_R <dbl>
Shop1	6	26378.3
Shop2	6	7228.4

2 rows

for both shops the optimal number of baristas was 6.

Part 6: Manova tests

Because it is a multivariate test Manova is chosen.

2.2 Test :

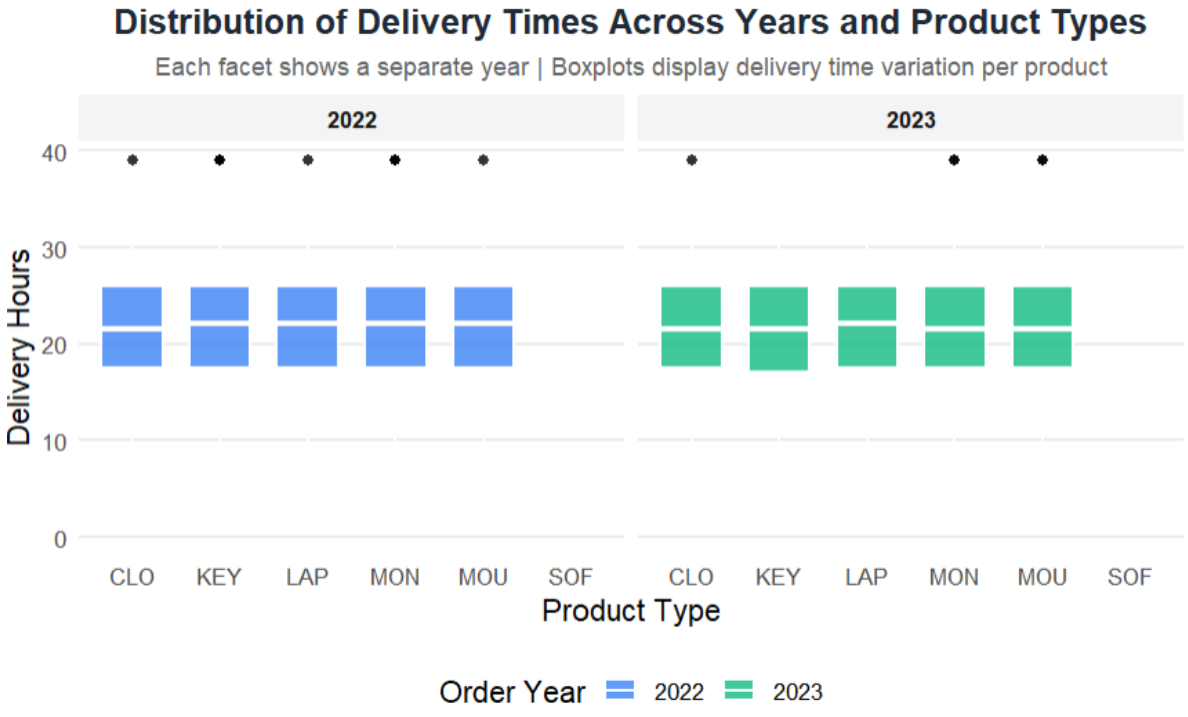
H_0 : The mean of delivery-related times does not differ by year nor product type

H_1: At least one of the mean vectors differs across years or product types.

R Console knit_asis

MANOVA 1 Summary (Wilks' Lambda Test)

	Df	Wilks_Lambda	Approx_F	Num_Df	Den_Df	P_value	Significa
as.factor(orderYear)	1	0.99991	4.49546	2	99987	0.01116	*
as.factor(substr(ProductID, 1, 3))	5	0.02669	102414.85393	10	199974	0.00000	***
as.factor(orderYear):as.factor(substr(ProductID, 1, 3))	5	0.99989	1.13091	10	199974	0.33396	ns
Residuals	99988	NA	NA	NA	NA	NA	ns



MANOVA Results

- **Effect of Year (2026 vs 2027):**
  Wilks' Λ = 0.99991, $p = 0.011 \rightarrow p < 0.05$
  → There is a statistically significant difference in delivery times between the two years.
- **Effect of Product Type:**
  Wilks' Λ = 0.02, $p = 0.000000006 \rightarrow p < 0.05$
  → Delivery time also varies significantly across product types.

The MANOVA results show that both the year of operation and the product type have a measurable impact on delivery times.
Performance shifted between 2026 and 2027, suggesting that process efficiency or external factors may have changed over time.
Different product categories consistently displayed distinct delivery durations, likely due to varying handling or processing needs.
Despite these differences, the pattern of change from one year to the next was similar across all products, implying that any overall improvement or slowdown affected every category in a comparable way.

In short, delivery performance evolved over time and differed between products, but the overall trend of change remained uniform throughout the system.

Part 7:

7.1 Estimated Reliable Days in a Year

Service reliability is defined as the probability of having 15 or more workers present.
From the data:

$$P = \frac{270 + 96}{397} = 0.9219$$

Estimated reliable days per year:

$$0.9219 \times 365 = 336.5 \approx 337 \text{ days}$$

This means the service is expected to operate reliably on roughly 337 days each year.

7.2 Optimising Profit

To determine the most profitable staffing level, reliability was modelled using a Binomial distribution with:

$$N = 16, P = 0.9219$$

The probability that at least 15 workers show up on a given day:

$$P(X \geq 15) = (16C15)(p^{15})(1-p) + (16C16)(p^{16})$$

Solving gives $p = 0.966$, meaning each individual worker has about a 96.6% chance of showing up.

Cost Model:

- Loss on a problem day: R20 000
- Cost of hiring one extra staff member: R25 000 per month
- Month = 30 days

Without extra staff:

$$\text{Unreliable days} = (1 - 0.922) \times 30 = 2.34 \text{ days/month}$$
$$\text{Expected loss} = 2.34 \times 20\,000 = R46\,800$$

Thus adding another staff member is worth considering.

Comparison of Staffing Options:

Number of Workers	Reliability (% of Reliable Days)	Problem Probability	Expected Problem Days (365)	Expected Annual Loss (R20 000/day)	Extra Staffing Cost (R25 000/month)	Estimated Annual Profit
16	92.20%	7.80%	28	R560 000	R0	Base profit
17	99.50%	0.50%	2	R40 000	R300 000	Net gain ≈ R220 000
18	99.90%	0.10%	<1	R20 000	R600 000	Net loss ≈ R−380 000

Adding one extra worker (17 total) nearly eliminates reliability issues while maintaining strong profitability.
Hiring 18 workers provides only a marginal improvement in reliability but significantly increases staffing costs, leading to lower overall profit.
Therefore, **17 workers per day** is the optimal balance between cost and reliability.