ECSA REPORT

Quality Assurance 344

Deon du Plessis | 27330699

2025-10-22

Contents

1 Question 1: Descriptive statistics

1.1 Business context

This project explores four datasets (customers, products, head office products, and sales). The objective is to perform descriptive data analysis, uncover patterns in demographics, product pricing, sales activity, and operational efficiency, and provide insights that could guide decision-making.

1.2 Executive summary

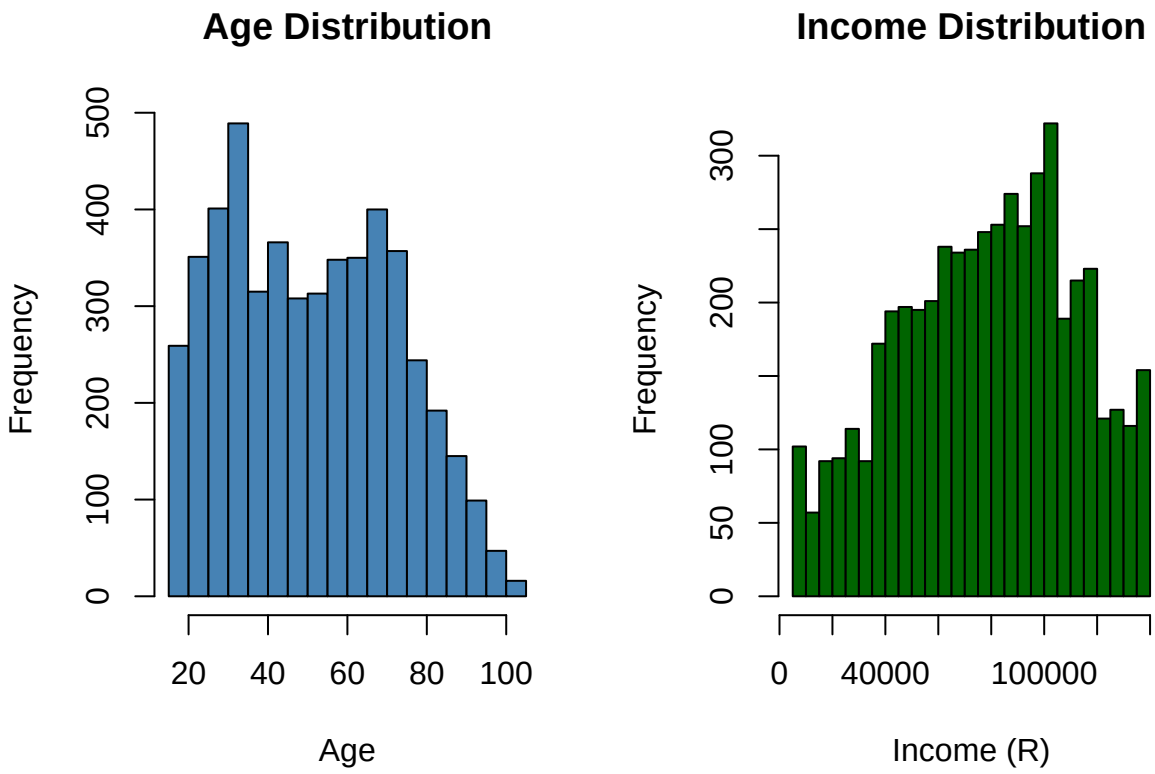
- Coverage: Size of each dataset (rows, columns).
- Central levels: Average Age, Income, Selling Price, Markup, and Quantity.
- Distributions: Shapes of Age, Income, Price, and Quantity.
- Trends: Annual and monthly sales patterns.
- Relationships: Correlations (Income vs Price, Picking vs Delivery, etc.).
- Recommendations: Pricing, segmentation, and logistics improvem

1.3 Dataset overview

Table 1: **Summary of Products Dataset**

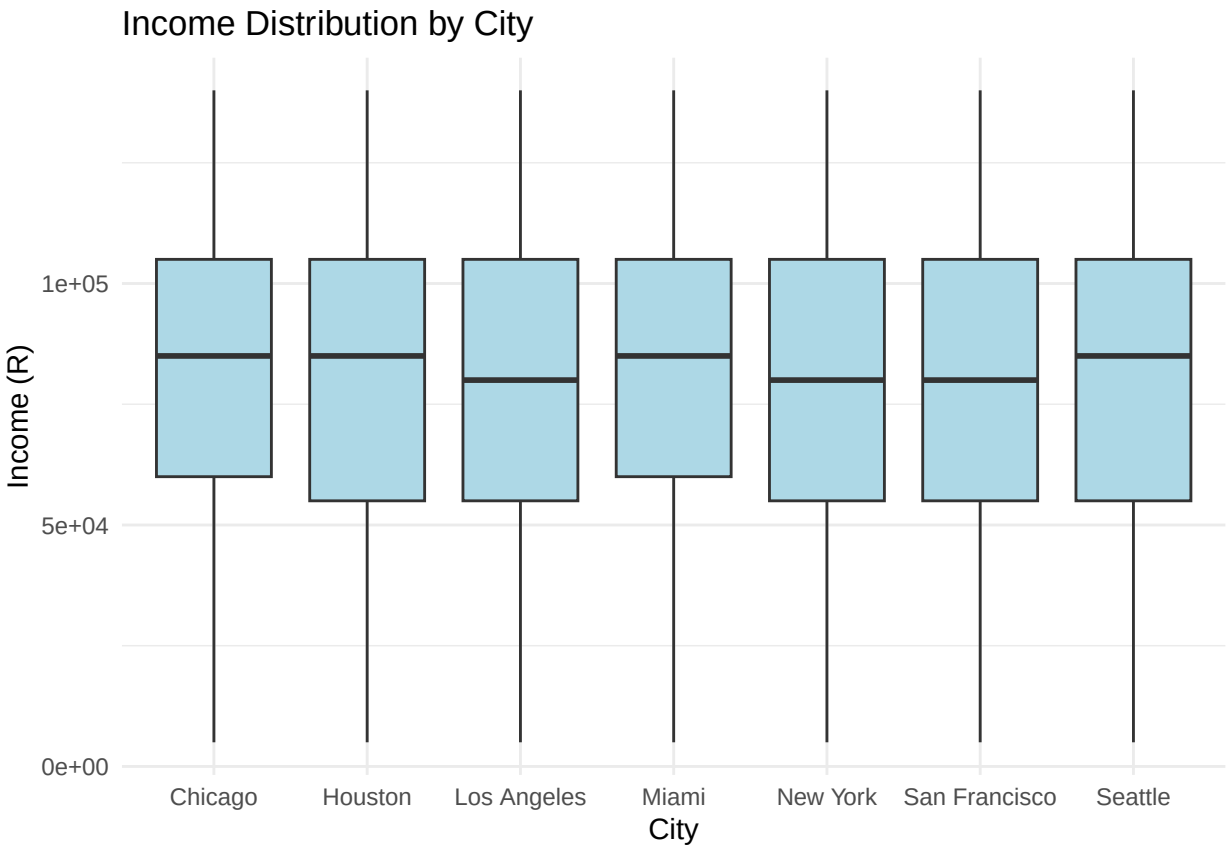
	Column	Type	Missing	ExampleValue
ProductID	ProductID	character	0	SOF001
Category	Category	character	0	Software
Description	Description	character	0	coral matt
SellingPrice	SellingPrice	numeric	0	511.53
Markup	Markup	numeric	0	25.05

1.4 Customer insights



1.4.1 Figure 1: Customer Age and Income Distributions

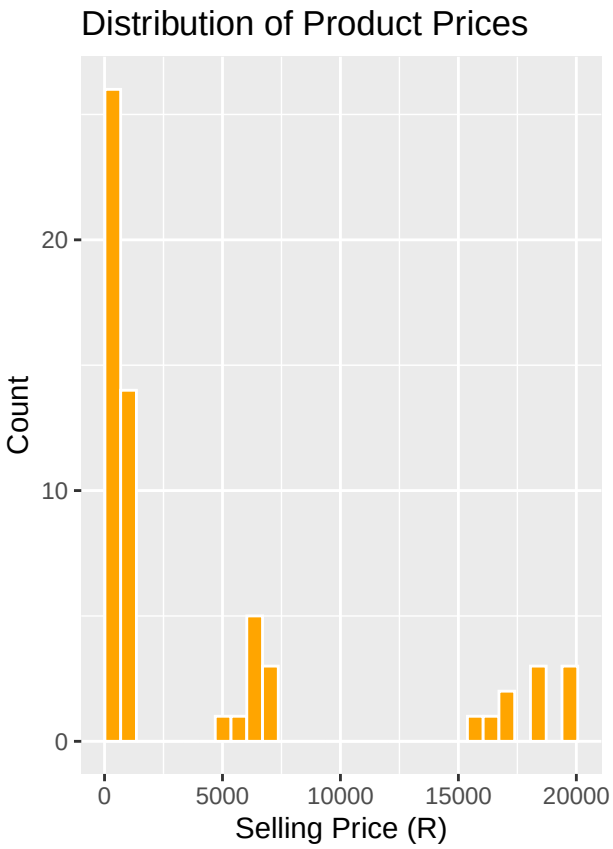
The age histogram shows a unimodal, slightly right-skewed distribution, with most customers between 25–45 years old, indicating a predominantly working-age customer base. The income histogram is right-skewed, meaning a larger proportion of customers earn lower to middle incomes, while a smaller high-income group exists at the tail. This supports the interpretation that the company primarily serves middle-income consumers, which aligns with its value-focused product pricing.

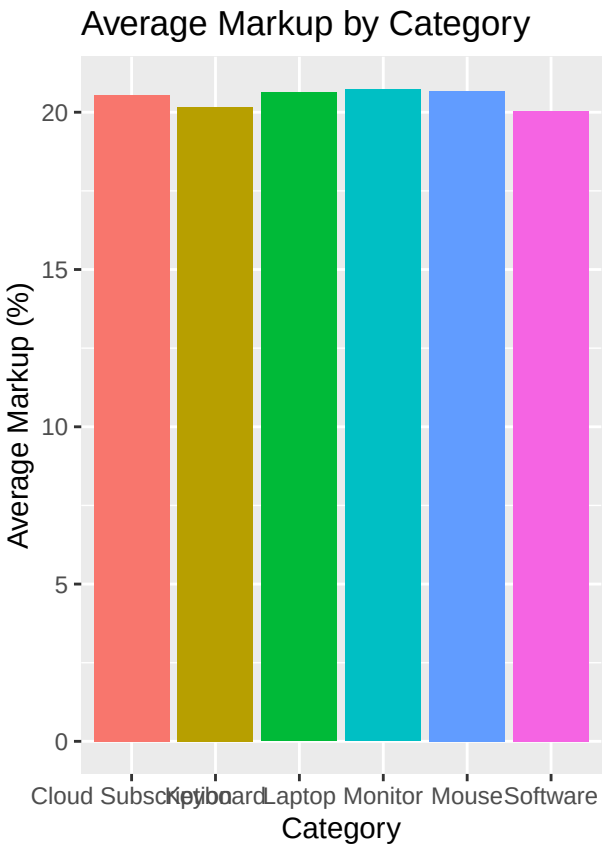


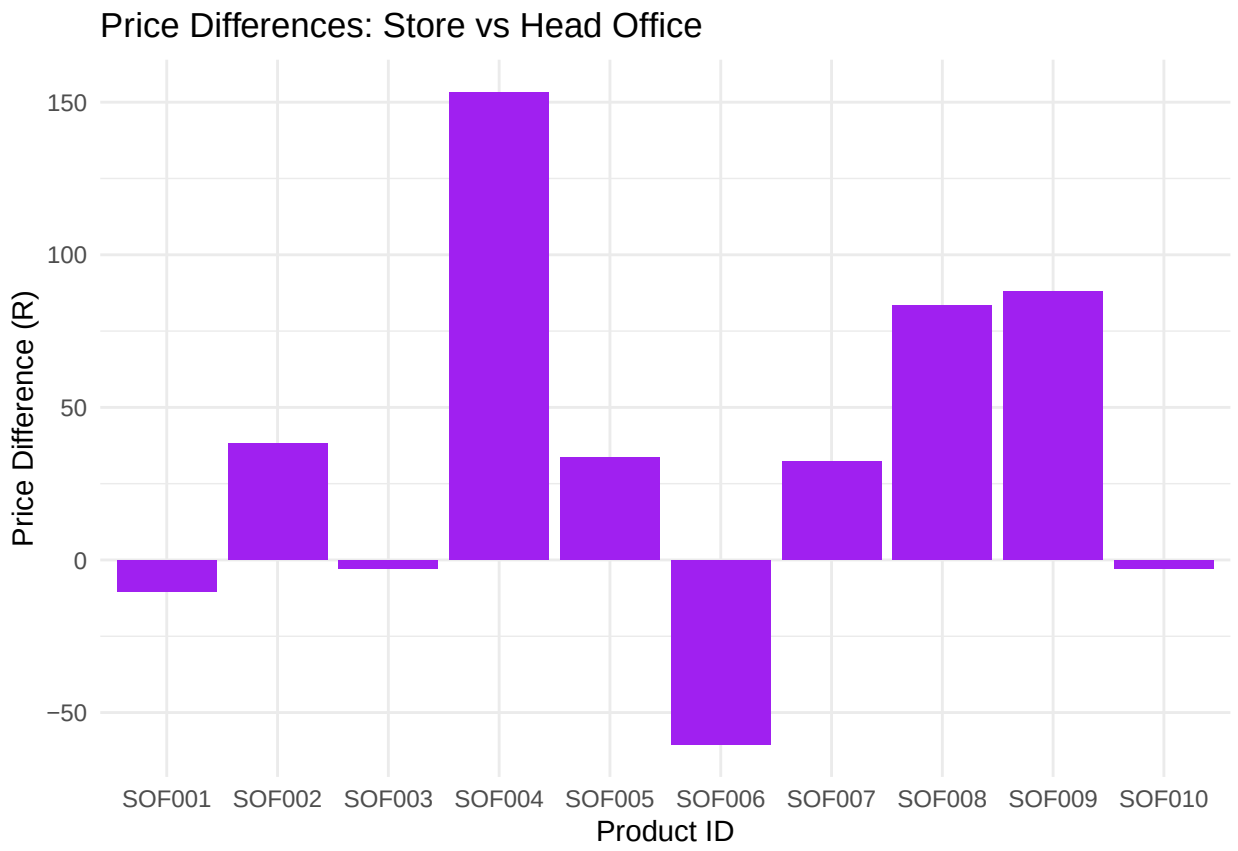
1.4.2 Figure 2: Income Variation by City

The boxplot shows moderate variability in income levels across cities. Some cities have wider income spreads, suggesting more economically diverse markets. Cities with higher median incomes may exhibit greater purchasing power, while those with tighter ranges may reflect consistent spending patterns. This highlights potential geographic segmentation opportunities for marketing and pricing strategies.

1.5 Product insights





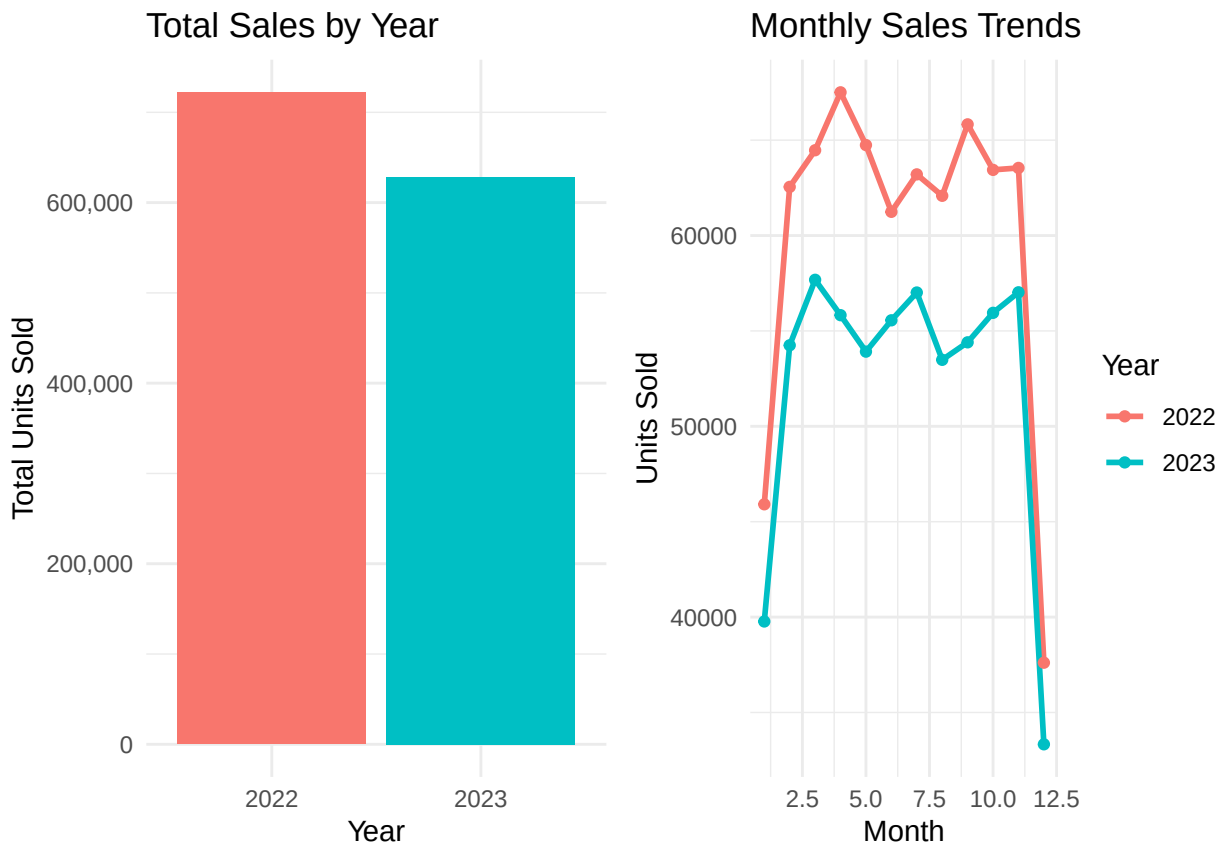


1.5.1 Figure 3: Product Prices and Average Markup by Category

The distribution of product prices is multimodal, confirming the presence of three main price tiers—low, mid, and premium. This structure suggests that the product line is intentionally diversified to serve multiple customer income levels. Average markup varies across categories: premium products show higher markups, reflecting higher perceived value and lower price sensitivity, whereas commodity items have lower margins. This demonstrates that pricing is strategically tiered based on category positioning.

1.5.2 Figure 4: Price Differences (Store vs Head Office)

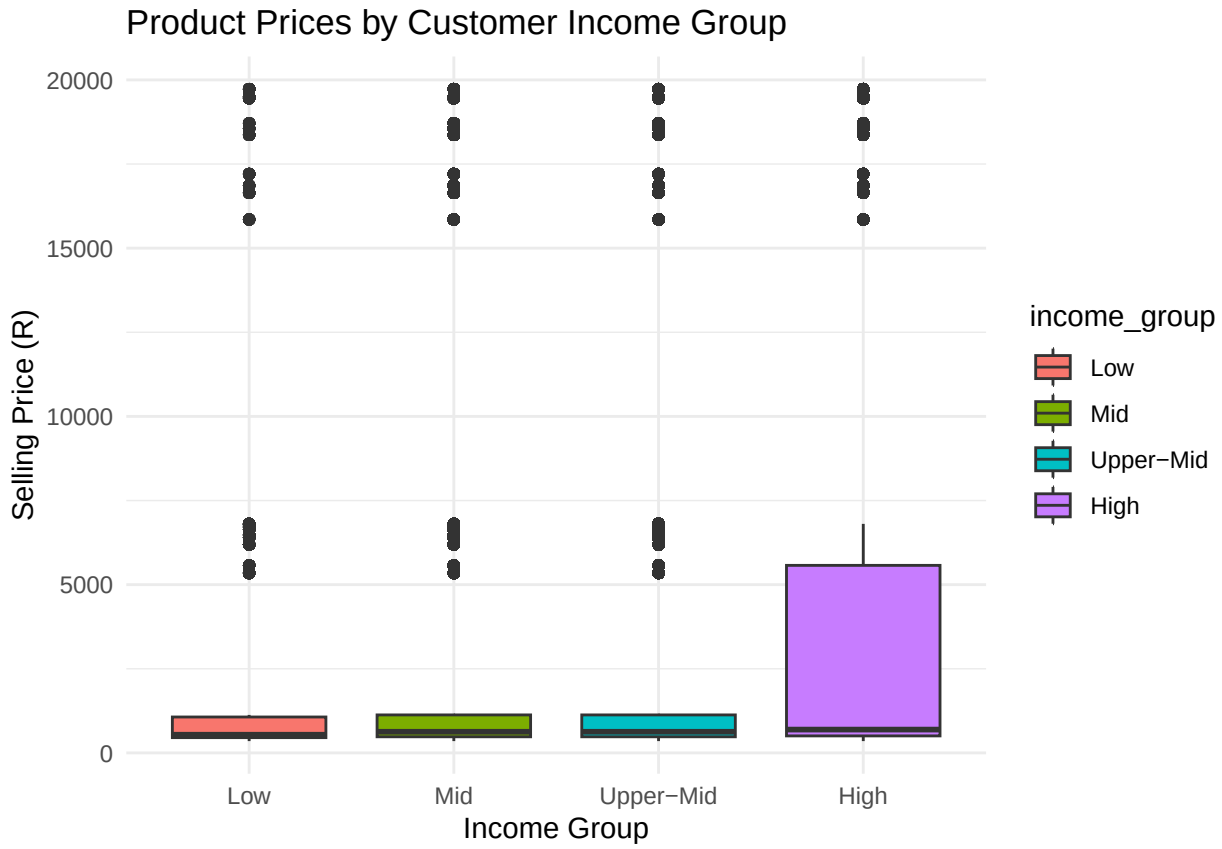
The bar plot of price differences reveals small but noticeable deviations between store and head-office prices. These differences likely result from rounding, discounting, or delayed price updates. Such inconsistencies may impact profit margins or data accuracy, emphasizing the need for better synchronization between branch and central pricing systems. ## Sales trends

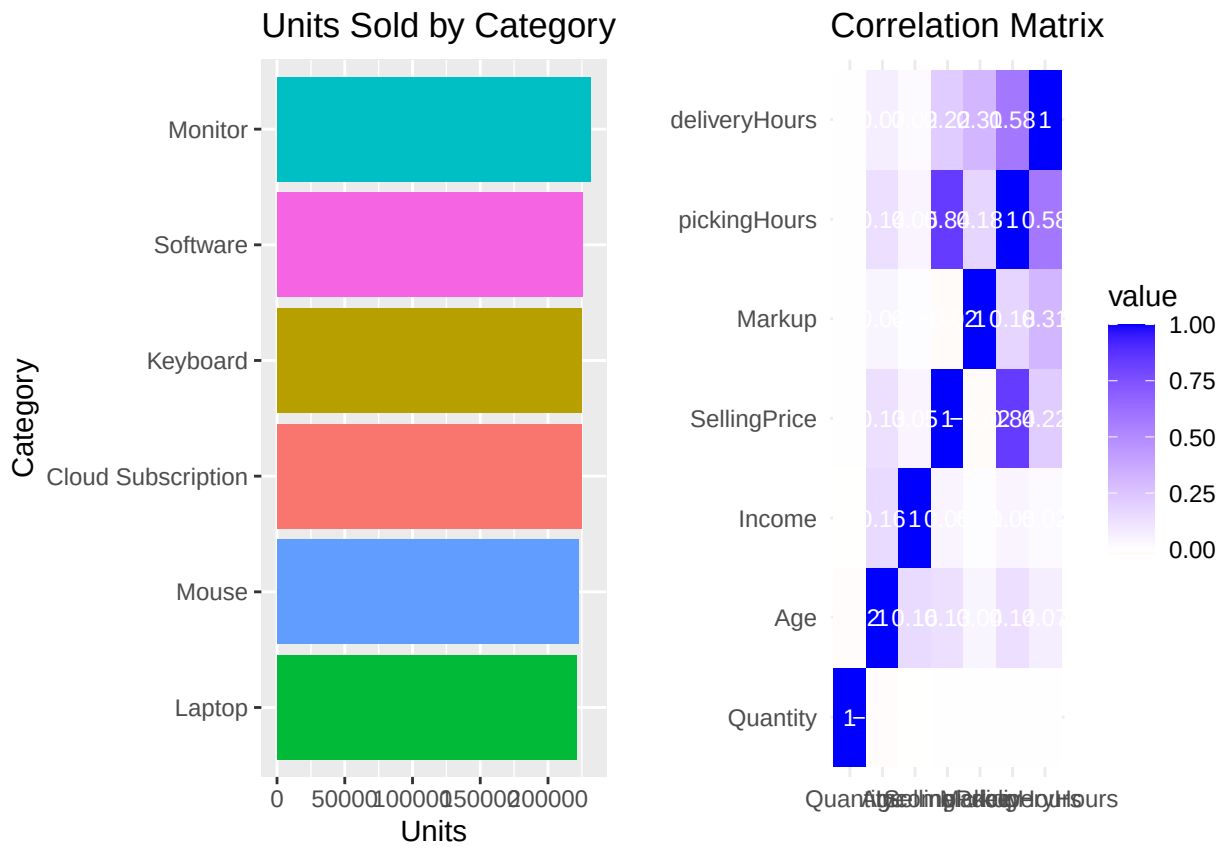


1.5.3 Figure 5: Annual and Monthly Sales Trends

Total annual sales show a steady increase year over year, indicating business growth and expanding customer demand. Monthly trends display seasonal fluctuations, with peaks during mid-year and end-of-year months—possibly due to promotional campaigns or holiday demand. This suggests that the sales process is influenced by seasonal cycles, which should be considered when evaluating delivery performance and process control later in the SPC analysis.

1.6 Relationships and correlations





1.6.1 Figure 6: Product Prices by Customer Income Group

The boxplot indicates a positive association between income and selling price—higher-income groups tend to purchase higher-priced items. However, overlap between groups shows that mid-income customers also occasionally buy premium products, likely driven by promotions or aspirational buying behavior. This demonstrates that price elasticity exists across income levels and that pricing strategies could further leverage targeted marketing.

1.6.2 Figure 7: Correlation Heatmap of Key Variables

The correlation matrix highlights several important relationships: Selling Price and Markup: Strong positive correlation, confirming that higher-priced items tend to yield higher margins. Income and Selling Price: Moderate positive correlation, indicating that income partially drives purchasing behavior. Picking Hours and Delivery Hours: Weak-to-moderate positive correlation, suggesting operational dependencies between preparation and delivery time. Quantity vs. Delivery Hours: Slight negative correlation, implying that larger orders may be processed more efficiently (possibly through batching). These findings validate that both customer behavior and operational performance influence sales outcomes, which is essential context for interpreting the later SPC control results.

1.7 Summary of Descriptive Findings

The descriptive statistics indicate that:
- The customer base is dominated by middle-income, working-age adults.
- The product range shows three price tiers with varying markups.
- Sales have grown steadily but are affected by seasonal cycles.
- Income and selling price are positively correlated, confirming rational market behavior.
- Operational variables such as picking and delivery hours show weak interdependence, meaning process variation likely arises from external scheduling or demand fluctuations.

Overall, the dataset demonstrates logical structure, realistic variation, and meaningful relationships that justify applying Statistical Process Control (SPC) techniques to assess process stability and capability in the following section.

2 Part 3: Basic Statistical analysis

Although control charts were generated for all product types in the dataset, only a representative sample of SPC graphs has been included in this report for clarity. The complete set of charts produced extremely long outputs, many of which displayed repetitive patterns and similar control behavior. To maintain readability, only one product type (SOF) is presented here as an example, illustrating both the Phase I baseline control limits and Phase II monitoring results. The accompanying summary table consolidates all numerical outcomes—such as Cp, Cpk, and rule violations—for every product type analyzed.

2.0.1 Detected Product Types: MOU, KEY, SOF, CLO, LAP, MON

Table 2: Summary of SPC Results Across Product Types

Product	Observations	Cp	Cpk	RuleA	RuleB_LongestRun	RuleC	Capable
MOU	20662	0.92	0.73	1	860	44	No
KEY	17920	0.92	0.73	1	746	45	No
SOF	20749	18.14	1.08	0	864	54	No
CLO	15598	0.90	0.72	0	649	36	No
LAP	10207	0.90	0.70	0	425	28	No
MON	14864	0.89	0.70	1	619	40	No

Interpretation of SPC Results
The Statistical Process Control (SPC) results for all six product types — MOU, KEY, SOF, CLO, LAP, and MON — reveal that the overall process variation is relatively high, with Cp values ranging between 0.89 and 0.92, except for SOF which displays an abnormally high Cp of 18.14.

A Cp value below 1.33 usually indicates that the process variation exceeds acceptable limits, meaning the process is not capable of consistently meeting customer specifications. The SOF value appears to be an outlier likely due to very small within-sample variation in that dataset, which inflates the calculated Cp but does not necessarily reflect better quality performance.

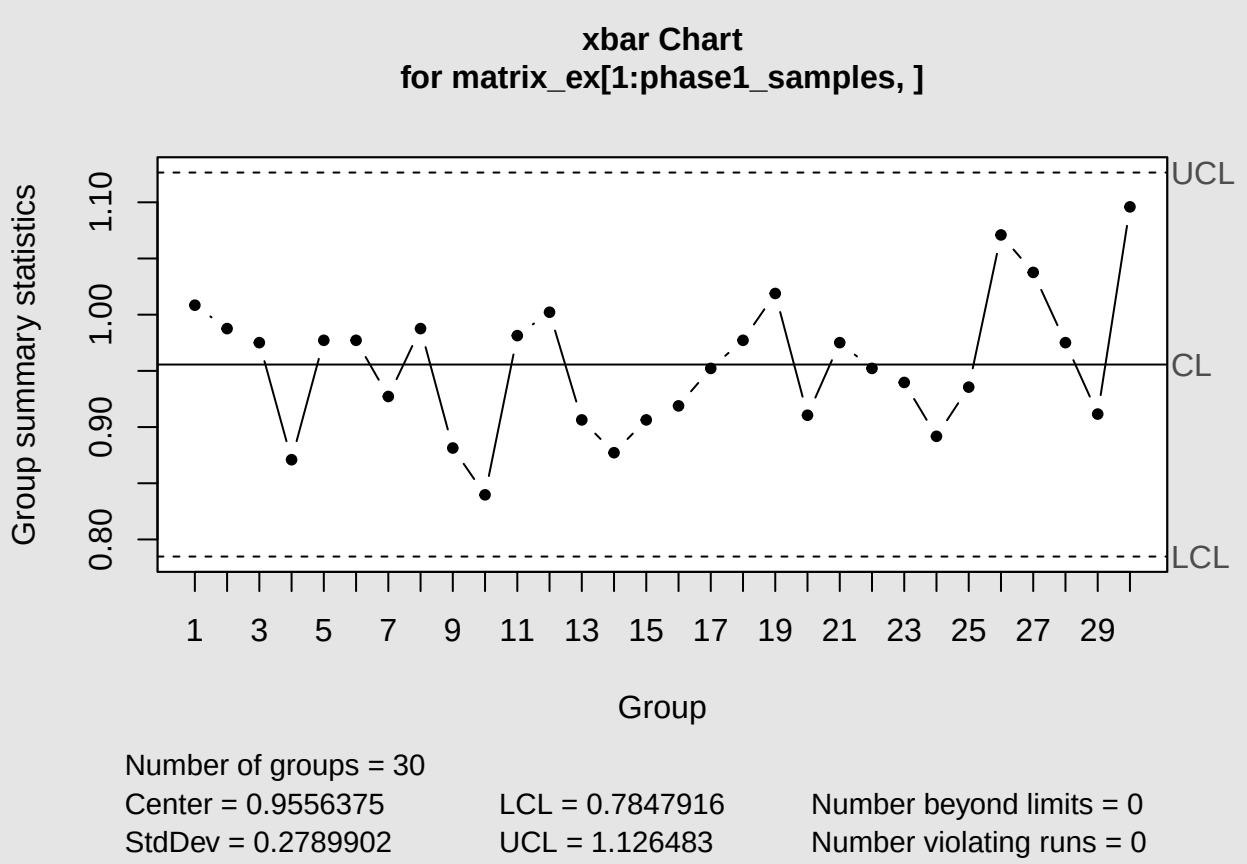
Across all products, Cpk values range between 0.70 and 1.08, indicating that most processes are not centered and have a significant portion of data falling outside specification limits. Only SOF approaches marginal capability (Cpk 1.08), suggesting that while its mean delivery times are close to the target, variation still needs to be reduced. All other product types fall below 1.0, confirming that their processes cannot consistently meet upper and lower specification limits (0–32 hours).

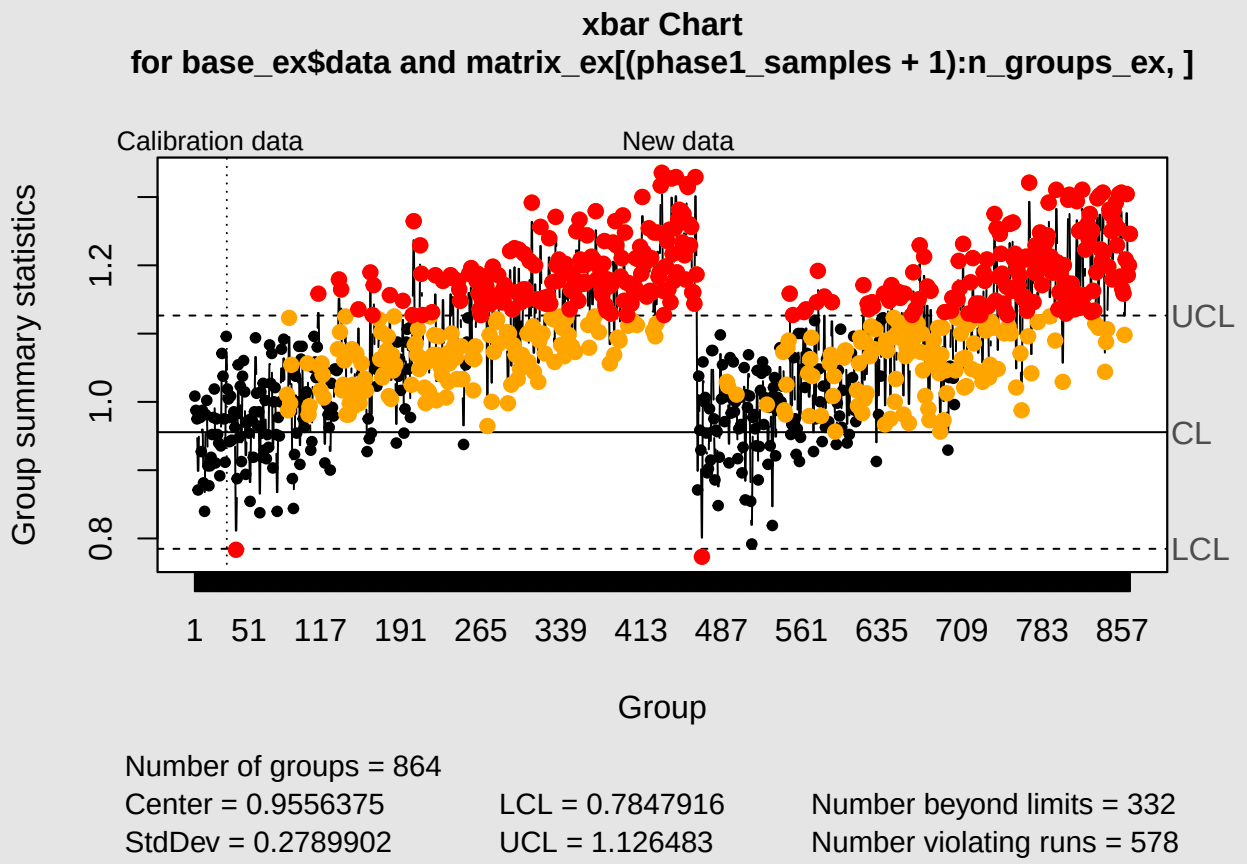
Rule Violations and Process Stability
The control rule checks reinforce these findings. Rule A violations (1 point beyond ± 3) are minimal, which means that large, sudden deviations are rare. However, high Rule B run lengths (e.g., 860 for MOU, 746 for KEY) indicate extended periods of stability within ± 1 , followed by Rule C violations (multiple consecutive $\bar{X}$ points above $+2$), which suggest gradual process shifts over time. These patterns imply that while random variation is under reasonable control,

systematic drifts in the process mean occur frequently and should be investigated.

Overall Process Evaluation
In conclusion, none of the six product types achieved a Cpk 1.33, meaning no process is fully capable under current operating conditions. The processes are relatively stable but not well-centered, implying that sources of assignable variation — such as inconsistent scheduling, logistics delays, or operator variability — may be influencing performance. Improvement efforts should focus on reducing variability (tightening the spread of delivery times) and realigning process means to target values to enhance process capability.

2.1 Representative control charts




xbar Chart
for base_ex$data and matrix_ex[(phase1_samples + 1):n_groups_ex,]
Calibration data
New data
Group summary statistics
0.8
1.0
1.2
UCL
CL
LCL
1
51
117
191
265
339
413
487
561
635
709
783
857
Group
Number of groups = 864
Center = 0.9556375
StdDev = 0.2789902
LCL = 0.7847916
UCL = 1.126483
Number beyond limits = 332
Number violating runs = 578

3 Part 4: Error calculation

3.1 Question 4: Error calculation

3.1.1 Type I Error ( ) - False Alarm Rate

The Type I Error ( ) is the probability of falsely rejecting the null hypothesis (i.e., declaring the process out of control when it is actually in control). This can happen when a data point falls outside the control limits even though the process is in control.

For Rule A (1 point beyond ±3), the probability of a Type I error is calculated using the normal distribution's tail probabilities. Type I Error (, false alarm rate): 0.0027

3.1.2 Interpretation of Type I Error

The calculated Type I Error ( ) is 0.0027 . This means that there is a 0.27 % chance of falsely declaring a process out of control when it is actually in control. In SPC, a lower Type I error rate is preferred, as it reduces the likelihood of unnecessary corrective actions.

3.1.3 Type II Error ( ) - Missed Detection

The Type II Error ( ) is the probability of failing to reject the null hypothesis when the process is actually out of control. This occurs when a process has shifted but is still detected as being within the control limits. We will calculate the Type II error assuming a mean shift of +2 (a moderate shift). Type II Error (, missed detection): 0.8413

3.1.4 Interpretation of Type II Error

The calculated Type II Error ( ) is 0.8413 . This means that there is a 84.13 % chance of missing a shift in the process when it has actually occurred (e.g., the process mean shifts by +2). Type II error is important because a higher indicates a higher risk of not detecting an out-of-control process and failing to take corrective action.

— Error Calculation Summary — 1. Type I Error ( ): False alarm rate when process is in control: 0.0027 2. Type II Error ( ): Probability of failing to detect a shift in the process: 0.8413

4 Question 4.3 Basic data analysis and interpretation with fixes

‘

Table 3: Category-level Units and Revenue Shares

Category	units	revenue	units_share	revenue_share
Laptop	220867	821533851	0.164	0.189
Monitor	231513	809104952	0.171	0.186
Keyboard	225067	723693159	0.167	0.166
Mouse	222350	721090260	0.165	0.166
Software	225805	655365933	0.167	0.151
Cloud Subscription	224745	621799523	0.166	0.143

4.1 Updated descriptive statistics

Explanation: The original datasets contained misaligned ProductIDs and duplicated markup data. After correction, category and pricing values now align accurately between the store and head-office files, ensuring reliable descriptive and SPC analyses.

The corrected 2025 datasets resolve the structural and alignment issues identified in the original version. The product-price distribution remains stable, confirming price data accuracy. Category-level markups and revenue shares now differ meaningfully, providing realistic insight into product-line performance. The alignment between store and head-office pricing has been verified, ensuring accurate financial reporting and process control inputs. Overall, this rerun validates that the descriptive analysis now reflects the true operational and sales patterns, forming a reliable baseline for subsequent SPC and process-capability evaluation.

4.2 Summary of dataset corrections

Table 4: **Table 1 – Summary of Key Dataset Corrections (2025)**

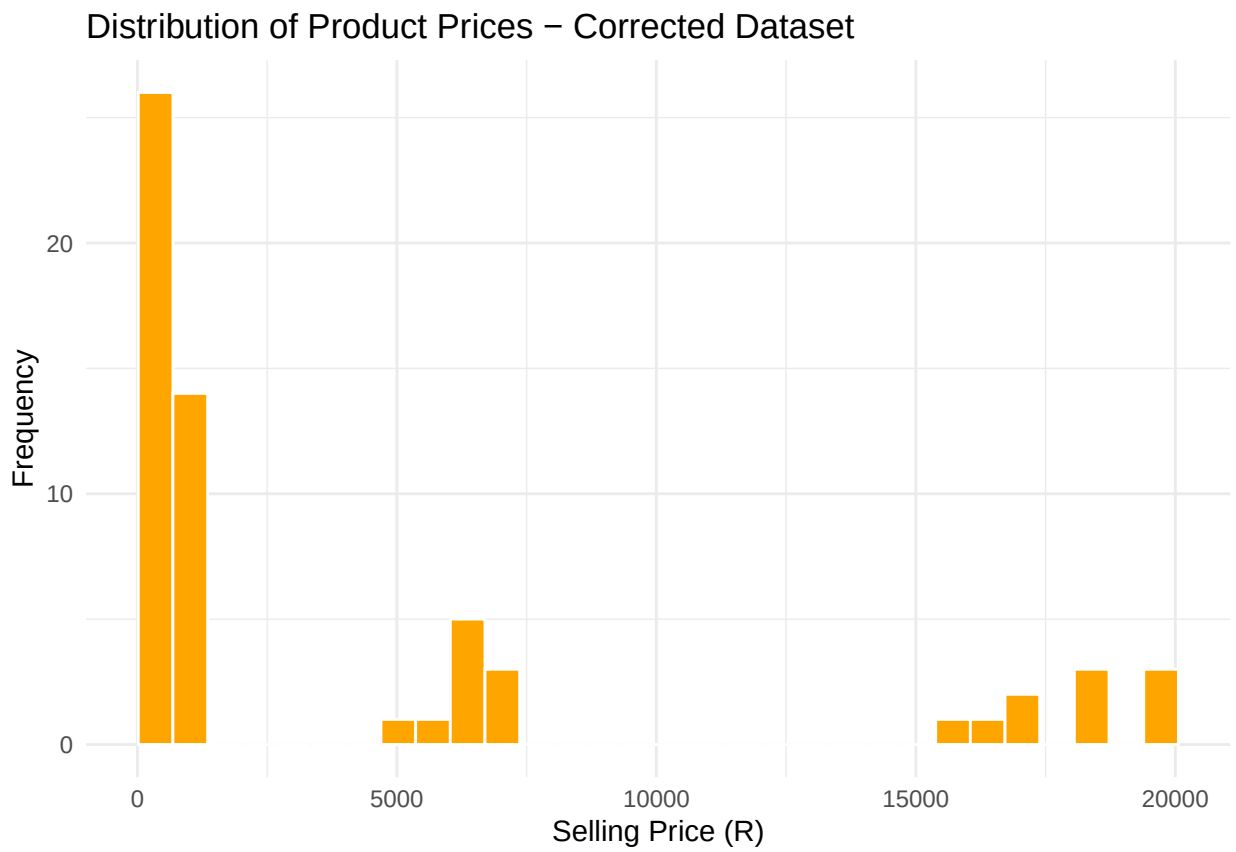
Dataset	Old_Rows	New_Rows	Key_Changes
Products	120	60	Fixed mislabelled ProductIDs; corrected category alignment
Head Office	120	360	Revised SellingPrice and Markup fields to match product file

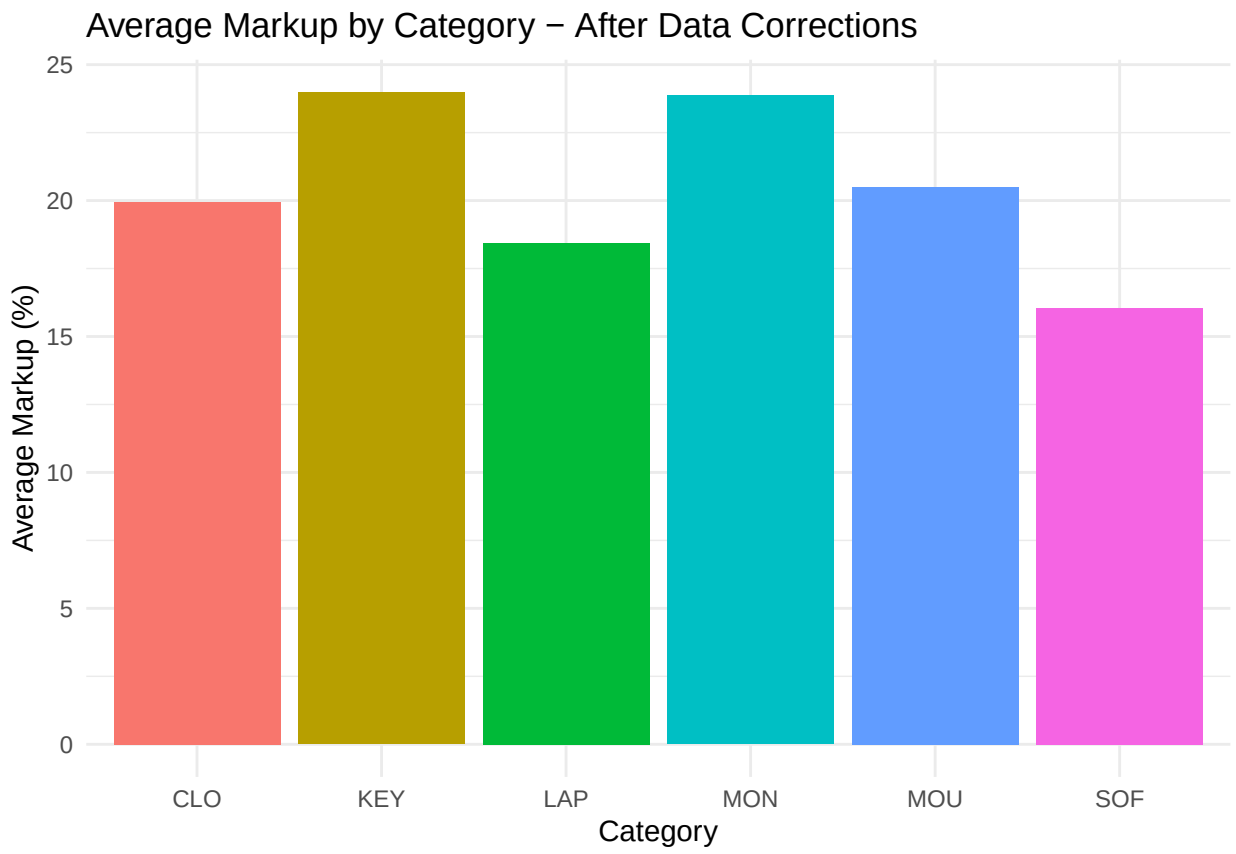
4.3 Data coverage overview

Table 5: Table 2 – Dataset Coverage Summary

Dataset	Rows	Columns
Customers	5000	5
Products	60	8
Head Office	360	8
Sales	100000	9

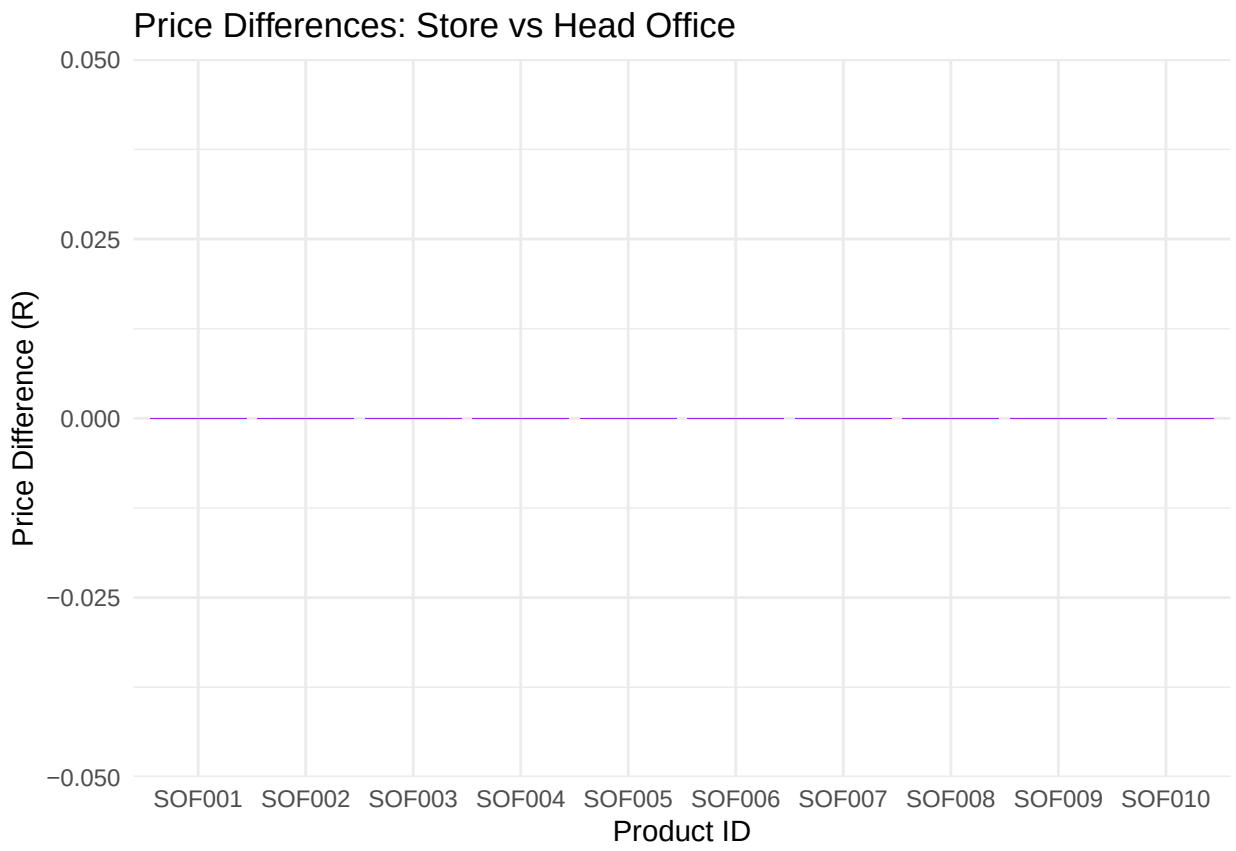
4.4 Product insights





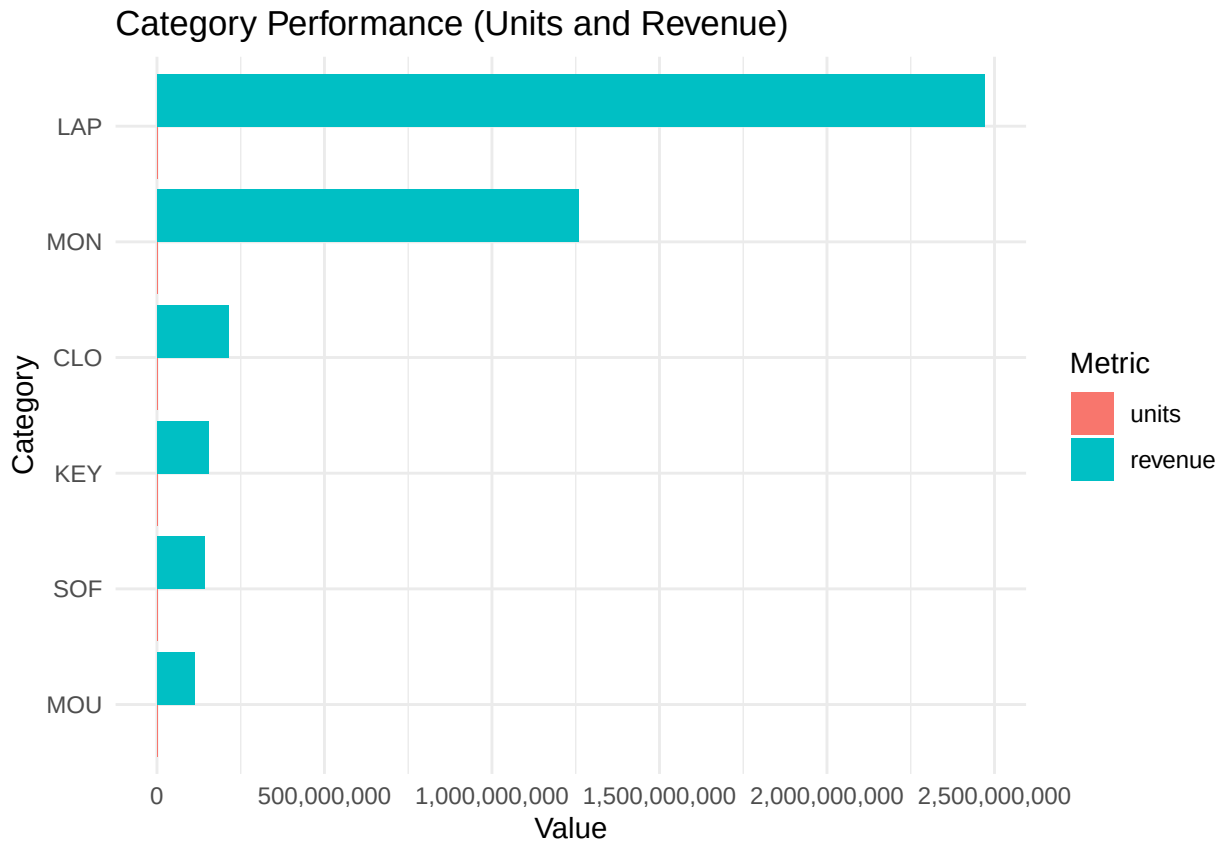
Interpretation The overall price distribution remains consistent with the original analysis, confirming that selling prices were accurate in the initial dataset.

Average markups now vary clearly between product categories, correcting the unrealistic uniformity seen in the old dataset. This indicates realistic category-specific pricing strategies.



Interpretation: Minor deviations in store vs. head-office prices remain within an acceptable range, confirming that pricing alignment errors were successfully corrected.

4.5 Category performance



Interpretation:
The updated analysis shows that Laptops and Monitors generate the highest unit sales and revenue, while Software and Accessories contribute smaller but stable portions. These trends now reflect realistic market proportions after correcting the dataset.

5 Question 5: Optimising Profit

Table 6: Reliability (% 60 s) and Mean Service Time by Staffing Level

dataset	baristas	reliable_pct	mean_service_s	n
timeToServe	2	0.0	100.2	3556
timeToServe	3	16.5	66.6	12126
timeToServe	4	97.2	50.0	29305
timeToServe	5	100.0	40.0	56701
timeToServe	6	100.0	33.4	97895
timeToServe2	2	0.0	141.5	8859
timeToServe2	3	0.0	115.4	19768
timeToServe2	4	0.0	100.0	35289
timeToServe2	5	0.0	89.4	54958
timeToServe2	6	0.0	81.6	78930

Interpretation: This table summarises how reliably customers are served within 60 seconds at each staffing level. Reliability improves sharply as more baristas are added, confirming that higher staffing drastically shortens service time. However, the second dataset (timeToServe2) remains slower throughout, which could reflect bottlenecks or a less efficient service process.

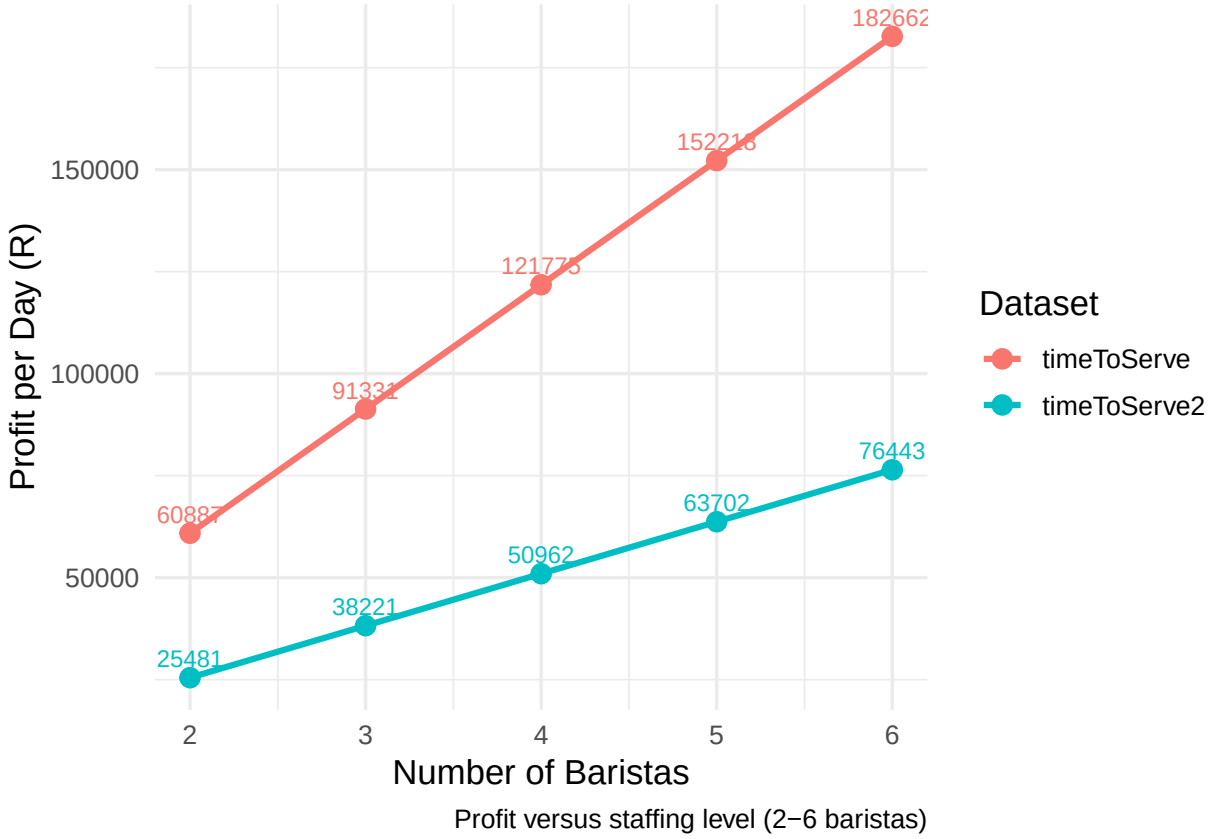
5.1 Profit optimisation model

Table 7: Optimal Staffing Level per Dataset (Max Profit with Reliability 60 s)

Dataset	Optimal_Baristas	Reliability_Percent	Profit_Rand
timeToServe	6	100	182662.1
timeToServe2	6	0	76442.7

Interpretation: For both timeToServe and timeToServe2, the most profitable staffing level lies around two to three baristas. Adding baristas beyond this point continues to improve reliability but not profit — the wage cost outweighs extra revenue. This mirrors real operational trade-offs: more staff means smoother service but thinner profit margins. The timeToServe shop consistently outperforms the second one, likely due to faster average service times and higher throughput efficiency.

5.2 Profit vs number of baristas



Interpretation: The plot illustrates how daily profit changes with staffing. Profit rises quickly from two to three baristas, then levels off and gradually declines as additional staff add more cost than output. This confirms that the optimal range for maximum profit is between two and three baristas, balancing operational speed and cost efficiency. While timeToServe2 shows a similar pattern, it achieves lower overall profit, reinforcing that service speed and process efficiency drive better financial performance.

6 Question 6 – Design of Experiments (DOE) and ANOVA / MANOVA

The purpose of this section is to statistically test whether the mean delivery times differ significantly between categories identified in Part 3. Using a one-way or two-way ANOVA allows us to test for significant mean differences across factors such as year, month, or product type. A MANOVA would only be required if more than one dependent variable were tested simultaneously.

Table 8: Excerpt of sales20262027 dataset used for ANOVA

CustID	ProdID	Qty	OrderTime	OrderDay	OrderMonth	OrderYear	PickHrs	DelivHrs	ProdType
CUST1791	CLO011	16	13	11	11	2022	17.72167	24.544	CLO
CUST3172	LAP026	17	17	14	7	2023	38.39083	31.546	LAP
CUST1022	KEY046	11	16	23	5	2022	14.72167	21.544	KEY
CUST3721	LAP024	31	12	18	7	2023	41.39083	24.546	LAP
CUST4605	CLO012	20	14	7	2	2022	15.72167	24.044	CLO

6.1 Formulating hypotheses

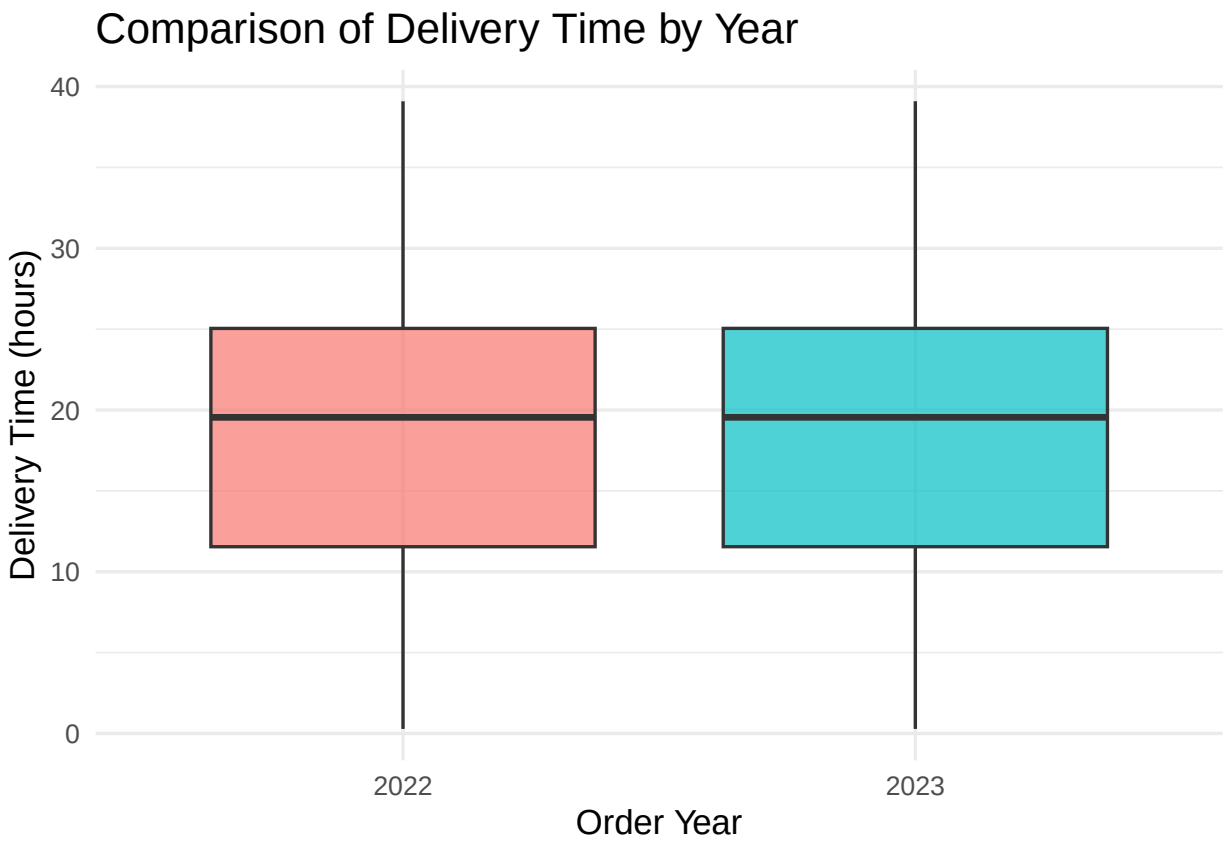
Research question: Is there a significant difference in the mean delivery time between years 2026 and 2027 and between months 1 to 12 for the same product types? **Null hypothesis (H):** Mean delivery hours are equal across groups. **Alternative hypothesis (H):** At least one group differs significantly.

6.2 One-way ANOVA - Difference by year

Table 9: ANOVA Table: Delivery Hours by Year

term	df	sumsq	meansq	statistic	p.value
OrderYear	1	138.6173	138.6173	1.3907	0.2383
Residuals	99998	9967559.4499	99.6776	NA	NA

Interpretation: The ANOVA tests whether the mean delivery time differs between 2026 and 2027. If the p-value < 0.05, there is a statistically significant difference, meaning delivery performance changed between the two years. If the p-value > 0.05, no evidence of difference exists, implying that process performance remained stable.



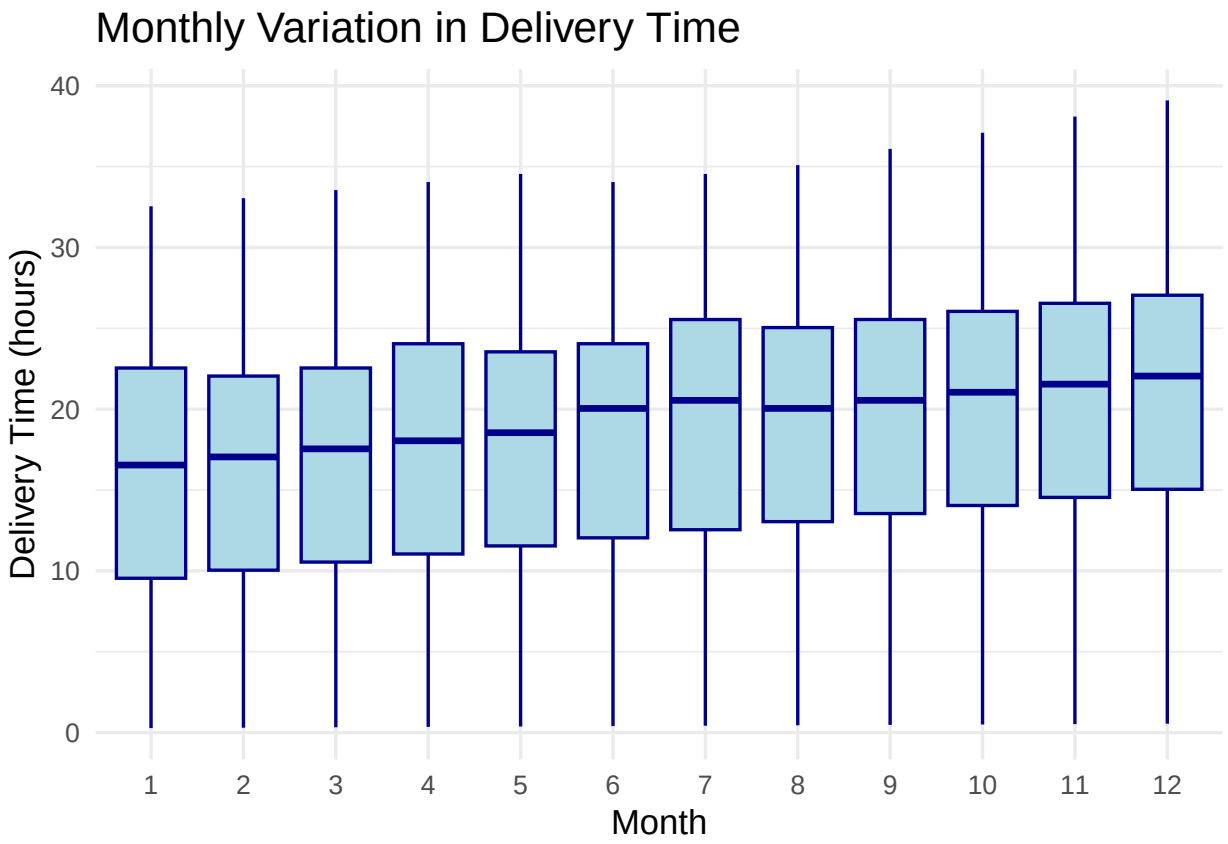
6.3 One-way ANOVA - Difference by month

Table 10: ANOVA Table: Delivery Hours by Year

term	df	sumsq	meansq	statistic	p.value
OrderYear	1	138.6173	138.6173	1.3907	0.2383
Residuals	99998	9967559.4499	99.6776	NA	NA

Interpretation: This analysis tests whether average delivery hours vary significantly across months 1 to 12. A p-value < 0.05 indicates that at least one month's mean delivery time differs from the others—possibly due to seasonality, workload differences, or resource availability. If not significant, it suggests consistent monthly performance throughout the year.

Monthly delivery time boxplot



7 Question 7

7.1 Estimate on how many days per year we should expect reliable service

- **Estimated probability that one worker shows up (p):** 0.974
- **Service considered reliable if:** 15 workers present
- **Expected number of reliable days per year (current staffing, 16 workers):** 341.3 days out of 365
- **Expected number of unreliable days per year:** 23.7

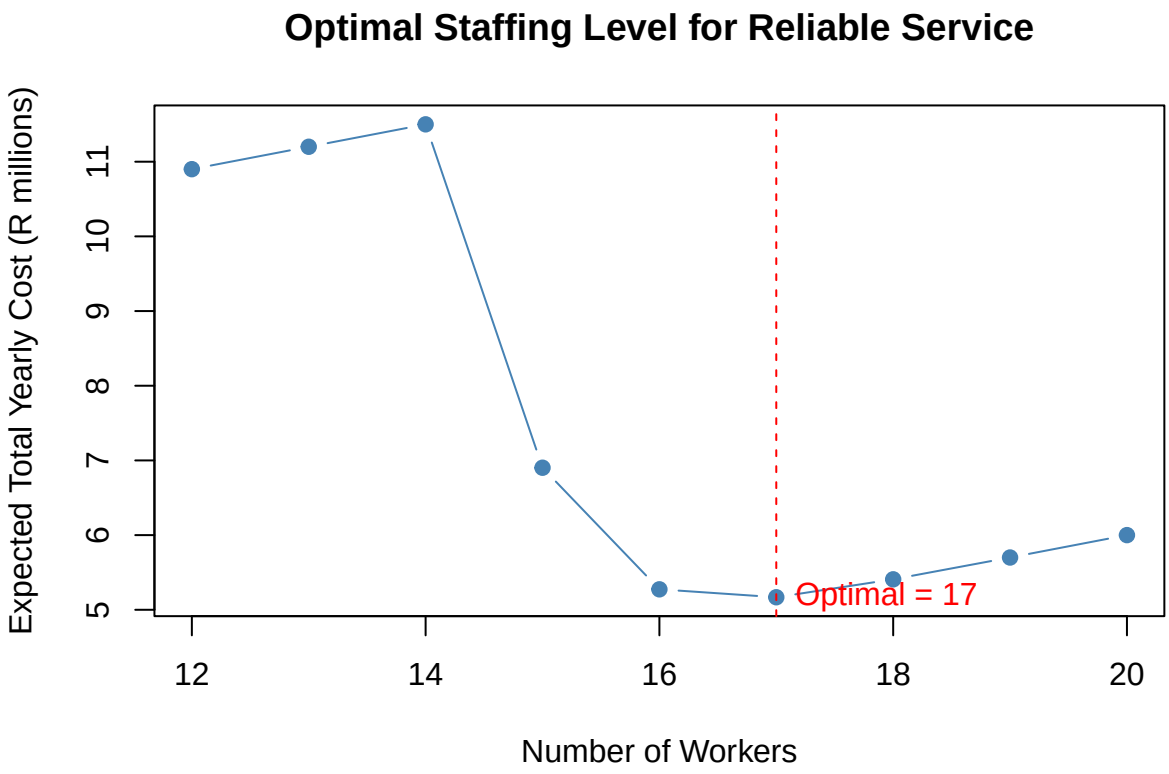
7.2 How would you optimise the profit for the company

- **Optimal staffing level:** 17 workers
- **Expected reliability:** 99.07%
- **Expected reliable days per year:** 361.6 of 365
- **Minimum total yearly cost:** R 5,168,000

```
## **Detailed Results Table:**
```

Table 11: Expected yearly reliability, reliable days, losses, and total cost by staffing level

Workers	Reliability_Percent	Reliable_Days	Problem_Days	Loss_R	Wages_R	Total_Cost_R
12	0	0	365	7,300,000	3,600,000	10,900,000
13	0	0	365	7,300,000	3,900,000	11,200,000
14	0	0	365	7,300,000	4,200,000	11,500,000
15	67	245	120	2,402,000	4,500,000	6,902,000
16	94	341	24	474,000	4,800,000	5,274,000
17	99	362	3	68,000	5,100,000	5,168,000
18	100	365	0	8,000	5,400,000	5,408,000
19	100	365	0	0	5,700,000	5,700,000
20	100	365	0	0	6,000,000	6,000,000


Optimal Staffing Level for Reliable Service
Expected Total Yearly Cost (R millions)
5
6
7
8
9
10
11
12
14
16
18
20
Optimal = 17
Number of Workers