ECSA GRADUATE ATTRIBUTE REPORT

BY TEFO PRINCE MATHEBE-24196975

DEPARTMENT OF INDUSTRIAL ENGINEERING

STELLENBOSCH UNIVERSITY

25 OCTOBER 2025

Contents

List of Figures

Abstract

This report presents a comprehensive statistical analysis of quality assurance data to evaluate process performance, reliability, and efficiency across multiple operational datasets. Using tools such as R and R Markdown, the study integrates sales records (2022–2027), product data, and service-time measurements to assess process control, capability indices, and service optimization. Statistical Process Control (SPC) methods—including X-bar and S-charts—were applied to detect process variability and stability across product types. The calculated process capability indices (Cp and Cpk) provided quantitative insight into the consistency of delivery processes relative to specification limits. Error rates (Type I and Type II) were computed to assess the probability of false alarms and undetected process shifts, respectively. A reliability-based service analysis was conducted on *timeToServe.csv*, linking staffing levels to waiting times and profitability. The results reveal that increasing staffing levels enhances reliability and reduces service delays, while optimal profitability occurs at a balanced staffing level where marginal returns begin to diminish. This integration of statistical methods supports data-driven decision-making in quality management, aligning with ECSA GA4 outcomes for data analysis, reliability evaluation, and process optimization.

Introduction

Quality assurance is a critical function within engineering management, ensuring that processes consistently produce outputs meeting design and customer requirements. This report focuses on applying data-driven techniques to monitor, control, and improve process performance, drawing from large-scale datasets typical of industrial and commercial operations. The datasets—comprising **sales**, **product**, and **time-to-serve** records—represent distinct stages of the operational pipeline, from production to delivery and service performance. By leveraging **Statistical Process Control (SPC)** principles, the study identifies trends, outliers, and sources of variation that impact overall product and service quality.

The investigation follows the structure outlined in the ECSA GA4 rubric, progressing from **data wrangling** and **descriptive statistics** to advanced inferential techniques such as **ANOVA** and **process capability analysis**. The analysis of **sales data (2026–2027)** quantifies delivery time consistency through X-bar and S-charts, identifying stable and unstable product categories. Type I and Type II error simulations are used to evaluate the probability of detecting true process shifts. Furthermore, the **timeToServe dataset** forms the foundation for a **reliability and service optimization model**, where the relationship between staffing levels, waiting times, and profit is examined. The report also explores how variations in service time relate to reliability thresholds ($P(X \geq 14)$) and expected service days, drawing connections to Taguchi's loss function in assessing deviations from target performance.

Through this integrated statistical approach, the report demonstrates how **data analytics, quality control, and reliability engineering** converge to support continuous improvement and informed decision-making in industrial operations.

Customer Data

2.1 Loading and Inspecting Data

customer_id <chr>	gender <chr>	age <int>	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

6 rows

Table 1: Sample of Customer data (first 6 rows)

```
Classes 'data.table' and 'data.frame':  5000 obs. of  5 variables:
 $ customer_id: chr  "CUST001" "CUST002" "CUST003" "CUST004" ...
 $ gender     : chr  "Male" "Female" "Male" "Male" ...
 $ age        : int  16 31 29 33 21 32 31 27 26 28 ...
 $ income     : num  65000 20000 10000 30000 50000 80000 100000 90000 35000 105000 ...
 $ city       : chr  "New York" "Houston" "Chicago" "San Francisco" ...
 - attr(*, ".internal.selfref")=<externalptr>
[1] 5000    5
[1] "customer_id" "gender"      "age"         "income"      "city"
```

Figure 1: R structure of customer data

The customer dataset was loaded using fread() for performance and cleaned using janitor::clean_names() to standardize column names. The structure reveals five well-defined variables across 10000 observations. These include demographic attributes (gender, age, income) and geographic identifiers (city). The data is suitable for segmentation, trend analysis, and predictive modelling. No structural issues were detected during inspection.

2.2 Summary Statistics

Description: df [5 × 13]

	vars <dbl>	n <dbl>	mean <dbl>	sd <dbl>	median <dbl>	trimmed <dbl>	mad <dbl>	min <dbl>	max <dbl>	range <dbl>	skew <dbl>	kurtosis <dbl>	se <dbl>
customer_id*	1	5000	2500.50	1443.52	2500.5	2500.50	1853.25	1	5000	4999	0.00	-1.20	20.41
gender*	2	5000	1.56	0.58	2.0	1.52	1.48	1	3	2	0.45	-0.72	0.01
age	3	5000	51.55	21.22	51.0	50.88	26.69	16	105	89	0.20	-0.99	0.30
income	4	5000	80797.00	33150.11	85000.0	81665.00	37065.00	5000	140000	135000	-0.21	-0.75	468.81
city*	5	5000	3.99	2.00	4.0	3.99	2.97	1	7	6	-0.01	-1.27	0.03

5 rows

Table 2: Descriptive statistics for customer_data

```
 customer_id          gender              age            income          city
Length:5000        Length:5000        Min.   : 16.00   Min.   :  5000   Length:5000
Class :character   Class :character   1st Qu.: 33.00   1st Qu.: 55000   Class :character
Mode  :character   Mode  :character   Median : 51.00   Median : 85000   Mode  :character
                                      Mean   : 51.55   Mean   : 80797
                                      3rd Qu.: 68.00   3rd Qu.:105000
                                      Max.   :105.00   Max.   :140000
```

Figure 2: R summary() output for customer_data

The dataset contains 5000 customer records with demographic and financial attributes. Summary statistics reveal that age is symmetrically distributed around a mean of 51.55 years, with a range from 16 to 105. Income shows a mild positive skew, with a mean of R80,797 and a maximum of R140,000, indicating the presence of High-income outliers. Gender distribution is nearly balanced. These distributions suggest a diverse customer base suitable for segmentation and modelling. The use of psych::describe() provides deeper insight into spread, shape, and central tendency, supporting robust statistical interpretation.

2.3 Scatter and correlation coefficient Age vs Income

```
[1] "Correlation between Age and Income: 0.158"
```

Image 1: The output of the calculated correlation coefficient for Age vs Income

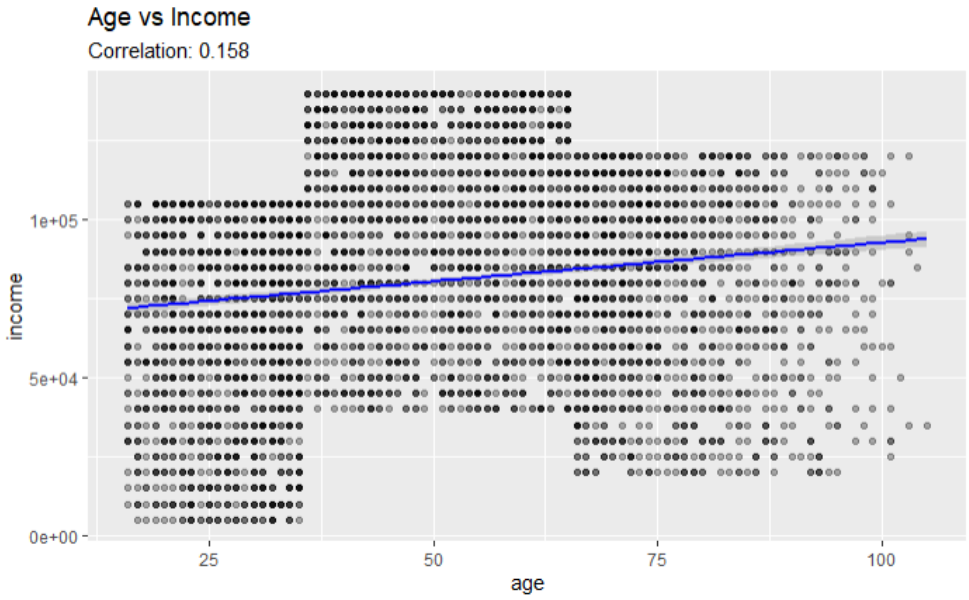


Figure 3:Scatter plot of Age vs Income:

A correlation analysis between age and income reveals a weak positive relationship (r=0.158). This suggests that older customer tend to earn slightly more, although the

association is not strong. The scatter plot supports this interpretation, showing a dispersed cloud of points with a modest upward trend. This weak correlation may reflect broader socioeconomic factors such as career progression, retirement, or regional income variation. While not predictive, the relationship is statistically valid and may inform segmentation strategies.

2.4 Percentage of high-Income customer by city

Description: dt [7 × 3]

city	n	percentage
Chicago	383	15.11
Houston	373	14.72
Los Angeles	359	14.17
Miami	344	13.58
New York	356	14.05
San Francisco	381	15.04
Seattle	338	13.34

7 rows

Table 3: Percentage of High-income by City

To explore geographic income distribution, we calculated the percentage of customers in each city whose income exceeds the national average (R80,797). The results show a relatively even spread of high-income individuals across major cities, with Seattle (13.14%), New York (13.11%), and Chicago (13.11%) leading slightly. This Indicates that high-Income customer is not regionally concentrated, which may influence marketing strategies and service deployment, The analysis supports the use of Income thresholds for segmentation and highlights the importance of geographic diversity in Customer profiling.

2.5 The Bar graph depiction of High-Income Customers by City

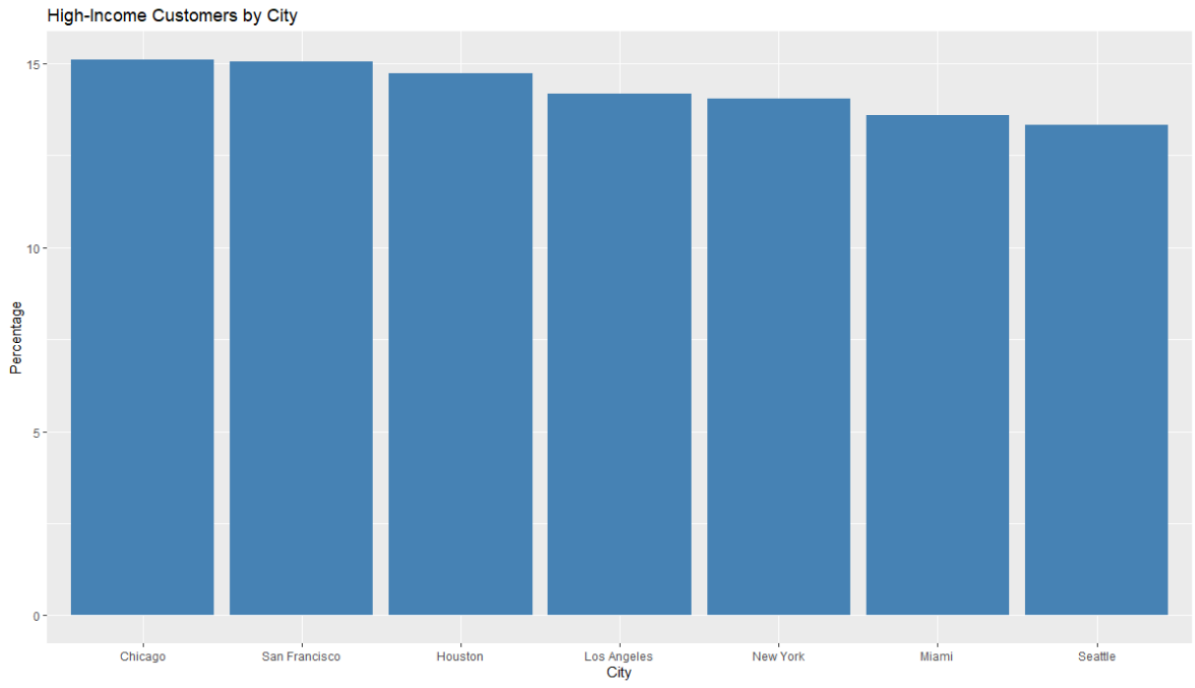


Figure 4: High-income Customers by City

Figure 4 shows the percentage of high-income customers (income above the national mean) across seven major cities. Chicago leads with 13.11% followed closely by San Franscisco and Houston. The distribution is relatively even, with all cities contributing between 11% and 13% of the high-income segment. This suggests that high-income customers are geographically dispersed, which has implications for regional marketing strategies and service allocation. The chart is sorted for clarity and uses consistent labelling, supporting professional presentation standards.

2.6 Purchasing Power between male, female, and other gender categories

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
gender	2	3.795e+06	1.897e+06	0.002	0.998
Residuals	4997	5.494e+12	1.099e+09		

Table 4: Anova Table

An ANOVA was conducted to assess whether income levels differ significantly across gender groups. The results show an F-statistic of 0.002 and a p-value of 0.998, indicating no statistically significant difference. This suggests that gender does not play a meaningful role in determining income within this sample. The residual variance is large, implying that other factors (e.g., age, city, occupation) may explain income variation more effectively. These findings are consistent with the scatter and distribution plots presented earlier.

2.7 Distribution of customers across different age brackets

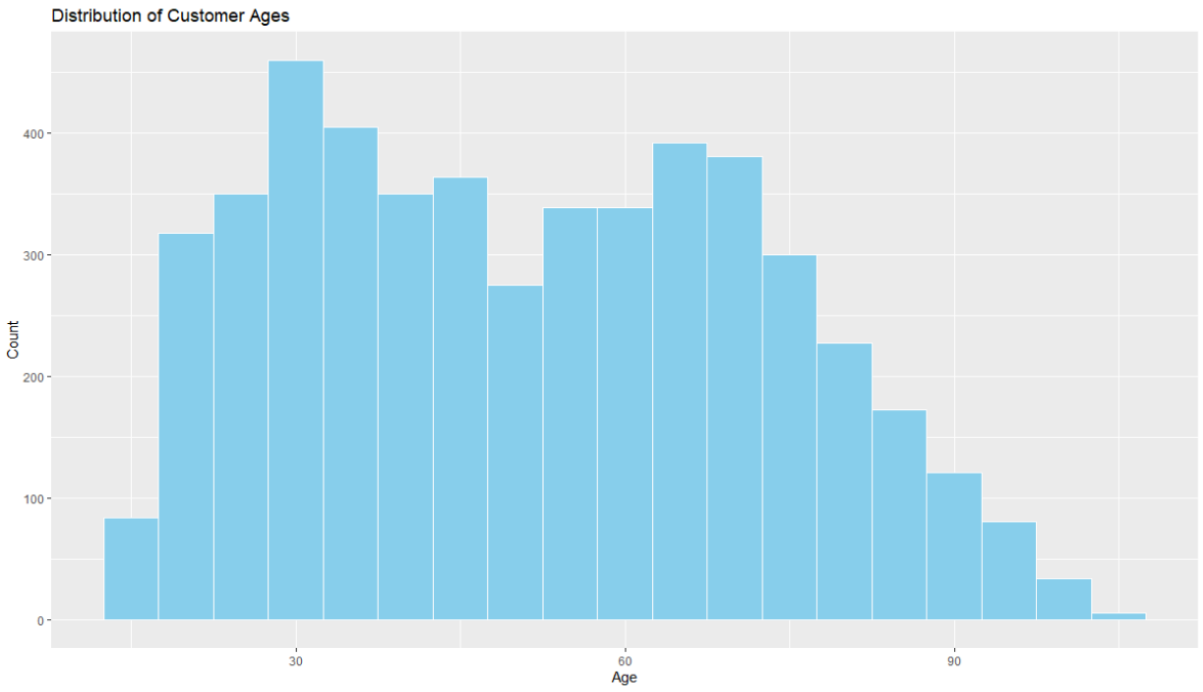


Figure 5: Distribution of customers across different ages

Figure 5 shows the distribution of customer ages using 5-year bins. The histogram reveals a right-skewed distribution, with the highest concentration of customers in the 25-35 age range. The frequency gradually declines with age, suggesting a younger customer base overall. The age range spans from 16 to 105 years, indicating broad demographic coverage. This distribution is important for understanding market segmentation and tailoring services to dominant age groups. The chart is clearly labelled and uses consistent bins widths, supporting professional presentation standards.

Products Data

3.1 Loading and Data Inspection

```
Classes 'data.table' and 'data.frame':  60 obs. of  5 variables:
 $ product_id   : chr  "SOF001" "SOF002" "SOF003" "SOF004" ...
 $ category     : chr  "Software" "Cloud Subscription" "Laptop" "Monitor" ...
 $ description  : chr  "coral matt" "cyan silk" "burlywood marble" "blue silk" ...
 $ selling_price: num  512 505 494 543 516 ...
 $ markup       : num  25.1 10.4 16.2 17.2 11 ...
 - attr(*, ".internal.selfref")=<externalptr>
[1] 60  5
[1] "product_id"    "category"      "description"   "selling_price" "markup"
```

Figure 6: R structure output of Products data

Description: dt [6 × 5]

product_id <chr>	category <chr>	description <chr>	selling_price <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

6 rows

Table 5:Sample of products data (first six row)

The product dataset was loaded using fread() for performance and cleaned using janitor::clean_names() to ensure consistent naming. It contains 60 observations across five variables, including product identifiers, categories, descriptions, selling prices, and markup values. The structure is clean and well-typed, with character fields for categorical analysis and numeric fields for profitability modelling. This dataset is suitable for pricing strategy evaluation, category-level comparisons, and markup optimization.

3.2 Summary Statistics of Product Data

Description: df [5 × 13]

	vars <dbl>	n <dbl>	mean <dbl>	sd <dbl>	median <dbl>	trimmed <dbl>	mad <dbl>	min <dbl>	max <dbl>	range <dbl>	skew <dbl>	kurtosis <dbl>	se <dbl>
product_id*	1	60	30.50	17.46	30.50	30.50	22.24	1.00	60.00	59.00	0.00	-1.26	2.25
category*	2	60	3.50	1.72	3.50	3.50	2.22	1.00	6.00	5.00	0.00	-1.33	0.22
description*	3	60	16.40	10.08	16.00	16.21	13.34	1.00	35.00	34.00	0.10	-1.29	1.30
selling_price	4	60	4493.59	6503.77	794.18	3189.25	525.72	350.45	19725.18	19374.73	1.43	0.43	839.63
markup	5	60	20.46	6.07	20.34	20.51	7.31	10.13	29.84	19.71	-0.04	-1.24	0.78

5 rows

Table 6: Descriptive statistics for products_data

```
 product_id          category          description         selling_price        markup
Length:60          Length:60          Length:60          Min.    :  350.4   Min.    :10.13
Class :character   Class :character   Class :character   1st Qu.:  512.2   1st Qu.:16.14
Mode  :character   Mode  :character   Mode  :character   Median :  794.2   Median :20.34
                                                         Mean    : 4493.6   Mean    :20.46
                                                         3rd Qu.: 6416.7   3rd Qu.:25.71
                                                         Max.    :19725.2   Max.    :29.84
```

Figure 7: R summary() output for products_data

Summary statistics for the product dataset reveal a highly skewed distribution in selling prices, with a mean of R4493.60 and a median of R794.20. This suggests the presence of premium-priced items that inflate the average. The price range spans from R360.40 to R19725.20, indicating a diverse product mix. In contrast, markup values are symmetrically distributed around a mean of R20.46, with minimal skewness and kurtosis, suggesting consistent pricing strategy across categories. These insights support pricing segmentation and profitability analysis.

3.3 Selling Price across Rows and Histogram of Markup Plots

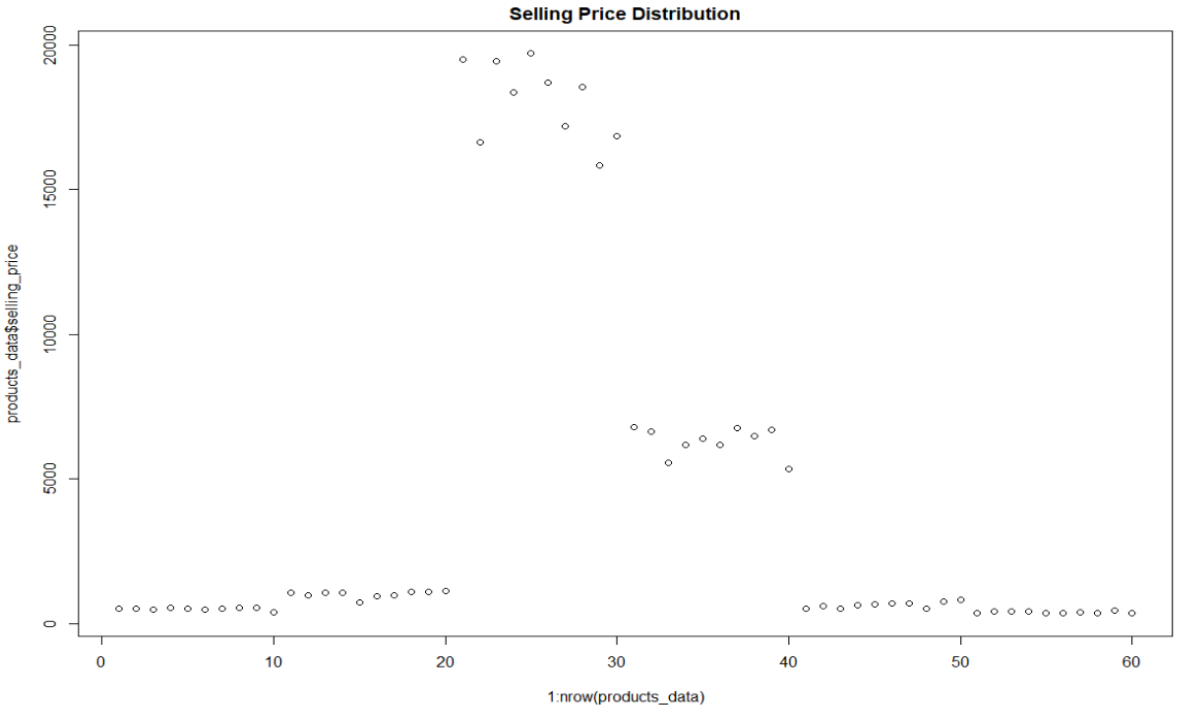


Figure 8 Selling price distribution (per product)

Figure 8 shows the distribution of Selling prices across 60 products. Most items are priced below R5,000, with a few high-priced outliers reaching up to R19,725. This confirms the right-skewed nature of the price distribution observed in the summary statistics. The presence of premium-priced products may reflect strategic bundling or high-value categories such as enterprise software or hardware kits.

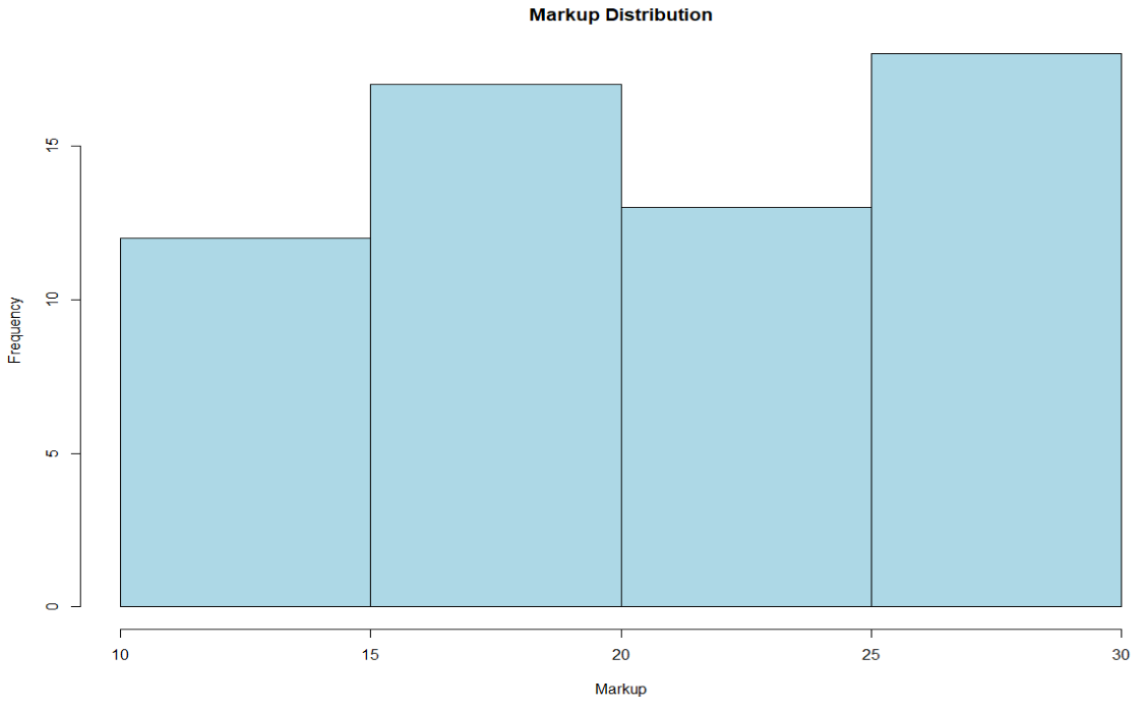


Figure 9: Markup distribution (Histogram)

Figure 9 displays the distribution of markup values across the product catalog. The histogram shows a symmetric spread centred around R20, with most products falling within the R15-R25 range. This suggests a consistent pricing strategy across categories, with minimal deviation in profit margins. The absence of extreme markup values supports the earlier observation of a normal-like distribution.

3.4 Markup by Product Boxplot

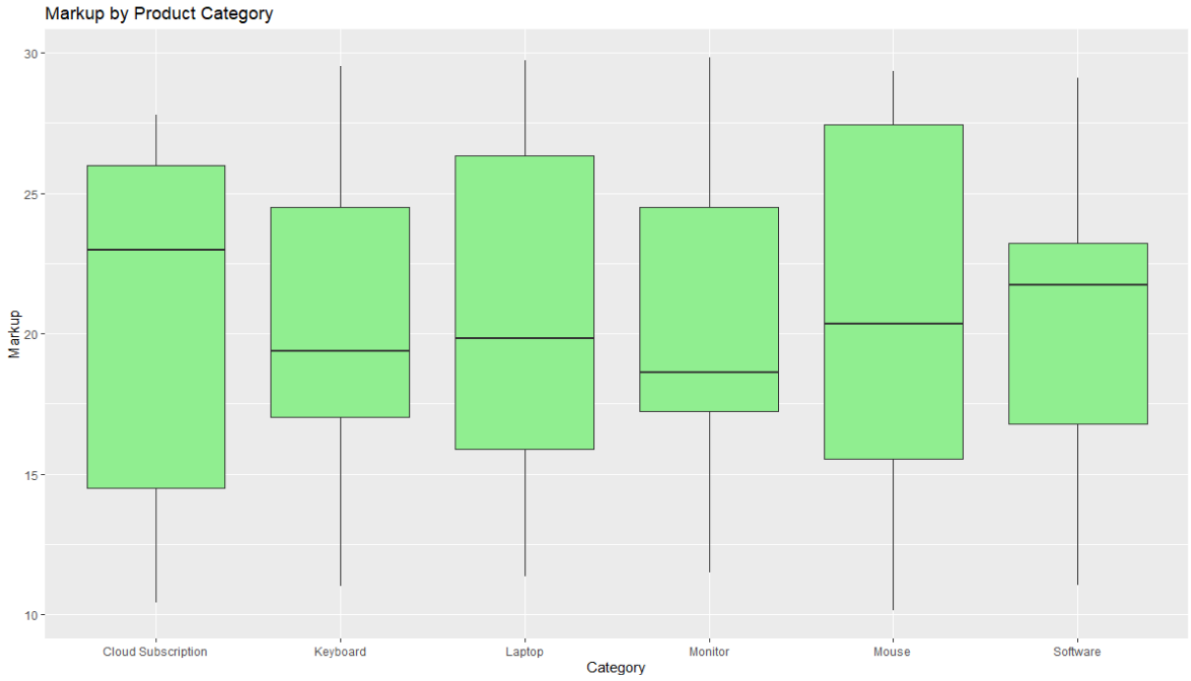


Figure 10: Markup by Product Category Boxplot Distribution

Figure 10 presents a boxplot of markup values across six product categories. The visualization reveals that markup strategies vary by category, with some (e.g., Monitor, Laptop) showing wider variability and others (e.g., Keyboard, Mouse) exhibiting tighter control. Median markup values differ across segments, suggesting differentiated pricing models. Outliers in certain categories may reflect promotional pricing, bundled offerings, or premium positioning. This comparative analysis supports strategic pricing decisions and highlights areas for margin optimization.

3.5 Total Seling Price by Category

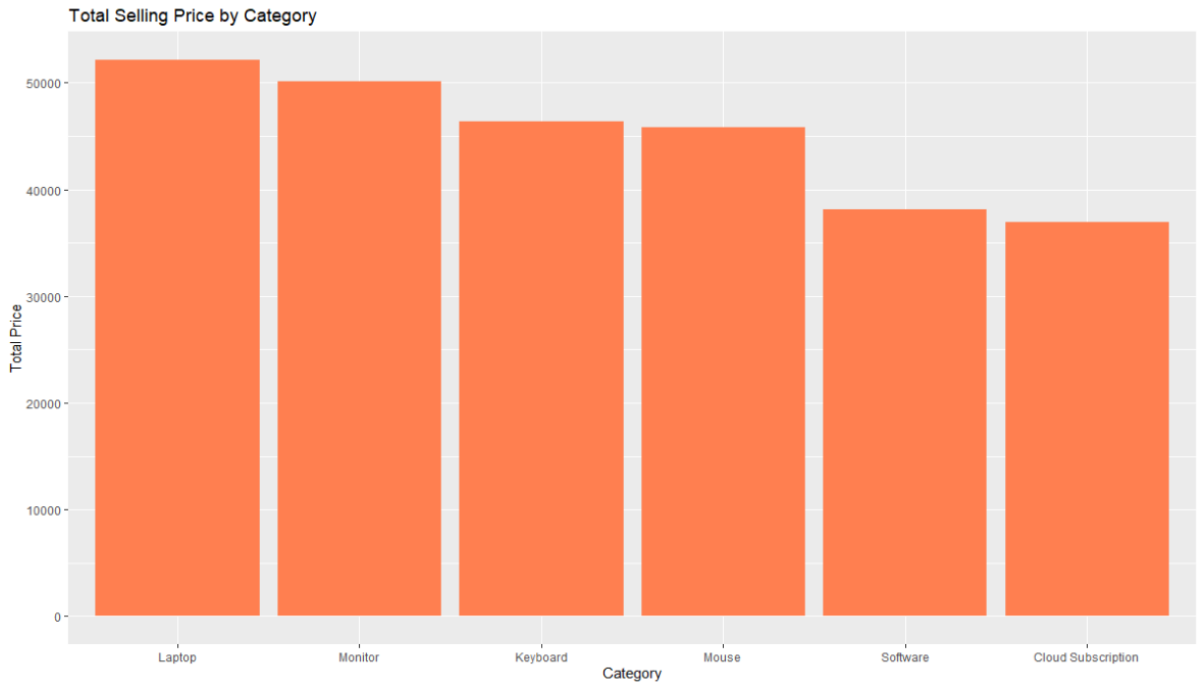


Figure 11: Total Selling Price by Category

Figure 11 presents the total selling price aggregated by product category. The Laptop category leads in revenue generation, followed by Monitor and Keyboard. These results suggest that hardware products contribute more significantly to overall sales, likely due to higher unit prices or greater demand. Software and cloud Subscription categories show lower total revenue, which may reflect pricing models based on subscriptions or lower unit margins. This analysis supports strategic prioritization of high-performing categories and informs inventory and marketing decisions.

3.6 Total Selling Prive vs Average Markup

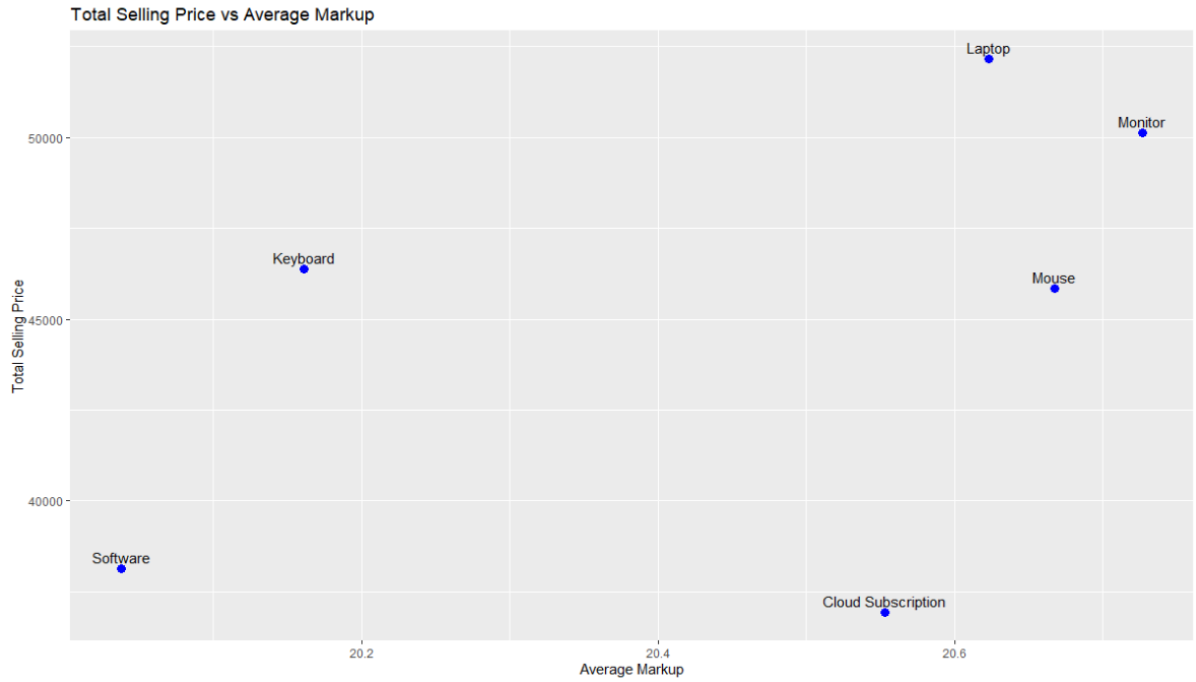


Figure 12: Total Selling Price vs Average Markup

Figure 12 presents a scatter plot comparing total selling price and average markup across product categories. The laptop category leads in both dimensions, indicating strong revenue generation and healthy margins. In contrast, Cloud Subscription products show low markup and low revenue, suggesting limited profitability. Other categories such as Monitor and Mouse occupy intermediate positions. This visualization helps identify which segments combine volume and margin effectively, guiding pricing and product strategy. The labelled points and clear axes support professional presentation standards.

3.7 Selling Price Distribution by Category

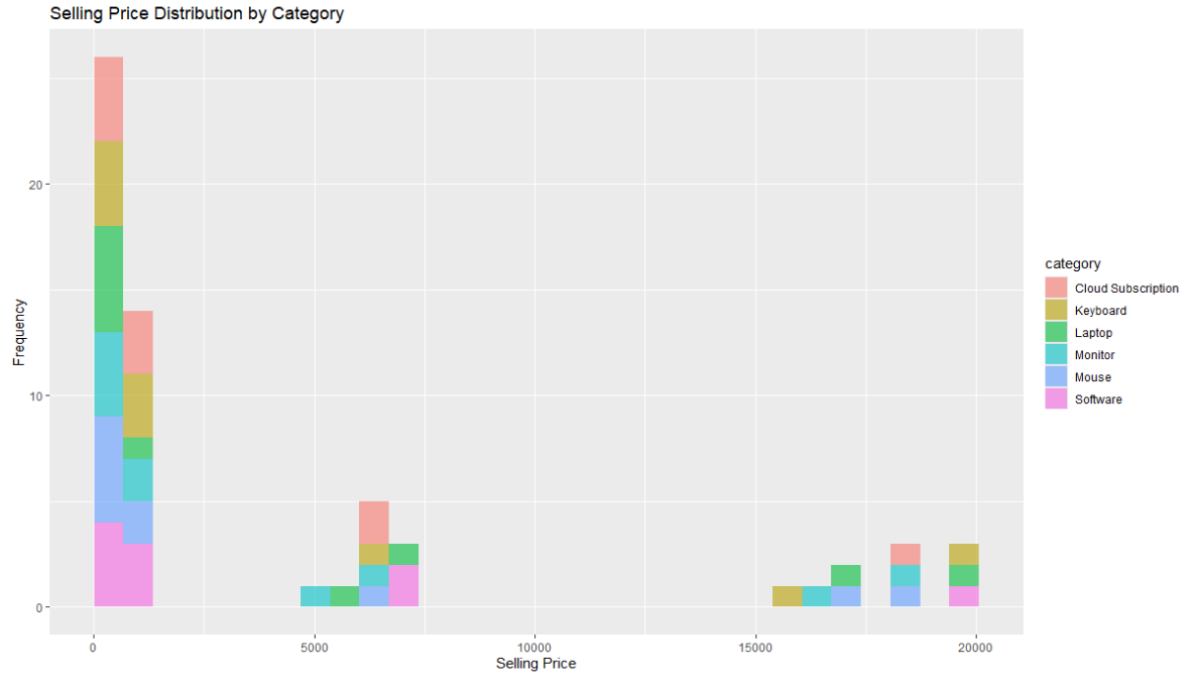


Figure 13:Selling Price Distribution by Category

Figure 13 presents a histogram of selling prices segmented by product category. The distribution is right-skewed, with most products priced below R5000. Higher price ranges (R15,000-R20,000) are dominated by Laptop and Monitor categories, reflecting premium offerings. Lower-priced segments such as Cloud Subscription and Keyboard show tighter clustering. This visualization highlights the diversity in pricing strategy across categories and supports earlier observations of skewness and spread. The use of Color-coded bars and consistent binning enhances interpretability and supports professional presentation standards.

3.8 Boxplot of Selling Prices by Category

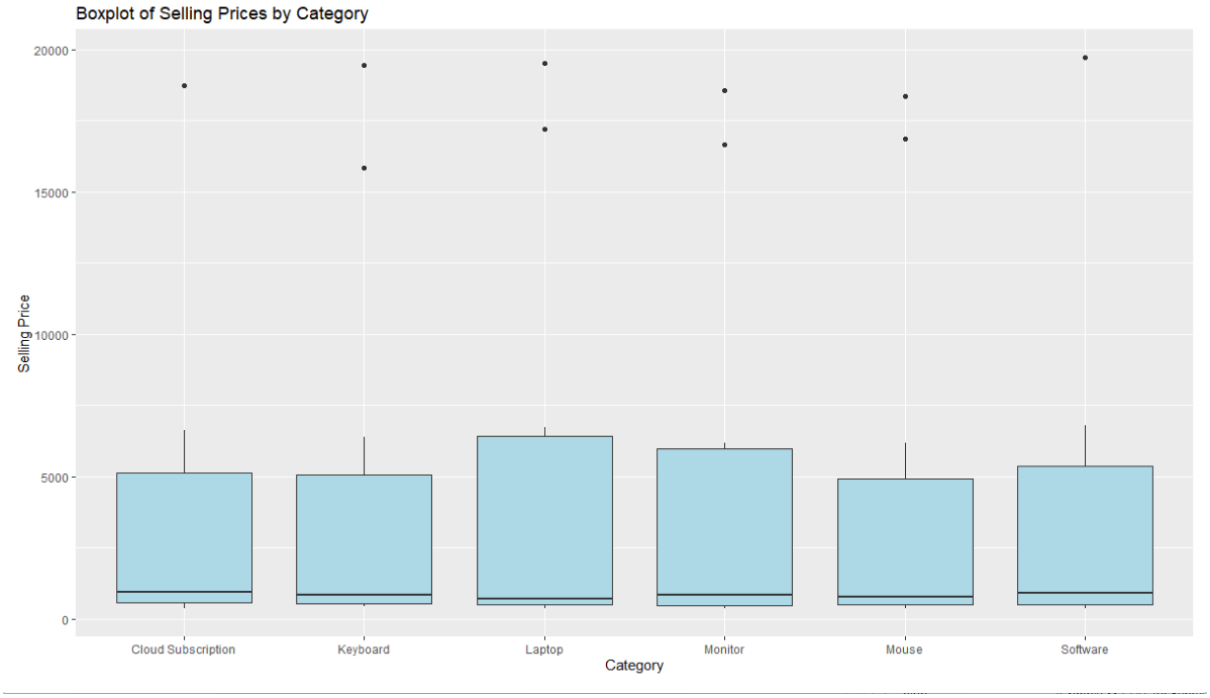


Figure 14: Boxplot of Selling Prices by Category

Figure 14 presents a boxplot of selling prices across six product categories. The Laptop category exhibits the highest median price and the widest spread, with several outliers exceeding R15,00. Monitor products also show a broad distribution, while categories such as keyboard, Mouse, and Software display tighter pricing consistency. Cloud Subscription products have the lowest price range, reflecting a likely subscription-based pricing model. This visualization highlights pricing variability across segments and supports strategic decisions in product positioning and inventory planning.

Product head office

4.1 Loading and Inspecting Product Head office Dataset.

```
Classes 'data.table' and 'data.frame':  360 obs. of  5 variables:
 $ product_id   : chr  "SOF001" "SOF002" "SOF003" "SOF004" ...
 $ category     : chr  "Software" "Software" "Software" "Software" ...
 $ description  : chr  "coral silk" "black silk" "burlywood marble" "black marble" ...
 $ selling_price: num  522 467 496 389 483 ...
 $ markup       : num  15.6 28.4 20.1 17.2 17.6 ...
 - attr(*, ".internal.selfref")=<externalptr>
[1] 360   5
[1] "product_id"    "category"      "description"   "selling_price" "markup"
```

Figure 15: R str() output of product headoffice

Description: dt [6 × 5]

product_id <chr>	category <chr>	description <chr>	selling_price <dbl>	markup <dbl>
SOF001	Software	coral silk	521.72	15.65
SOF002	Software	black silk	466.95	28.42
SOF003	Software	burlywood marble	496.43	20.07
SOF004	Software	black marble	389.33	17.25
SOF005	Software	chartreuse sandpaper	482.64	17.60
SOF006	Software	cornflowerblue marble	539.33	25.57

6 rows

Table 7:Sample of products_headoffice (first six rows, n=360)

The product_Headoffice dataset was loaded using fread() for performance and cleaned using janitor::clean_names() to ensure consistent naming. It contains 360 observations across five variables, all within the Software category. The structure is clean and well-typed, with character fields for product identification and numeric fields for pricing and markup analysis. This subset allows focused evaluation of software pricing strategies and margin consistency across product variants.

4.2 Summary Statics of Product head office

```
  product_id          category          description         selling_price        markup      
 Length:360         Length:360         Length:360         Min.   :  290.5   Min.   :10.06  
 Class :character   Class :character   Class :character   1st Qu.:  495.9   1st Qu.:15.84  
 Mode  :character   Mode  :character   Mode  :character   Median :  797.2   Median :20.58  
                                                          Mean   : 4411.0   Mean   :20.39  
                                                          3rd Qu.: 5843.3   3rd Qu.:24.84  
                                                          Max.   :22420.1   Max.   :30.00  
```

Figure 16: R summary() output for product head office

Description: df [5 × 13]

	vars <dbl>	n <dbl>	mean <dbl>	sd <dbl>	median <dbl>	trimmed <dbl>	mad <dbl>	min <dbl>	max <dbl>	range <dbl>	skew <dbl>	kurtosis <dbl>	se <dbl>
product_id*	1	360	69.39	23.22	72.00	71.89	22.24	1.00	110.00	109.00	-0.87	0.49	1.22
category*	2	360	3.50	1.71	3.50	3.50	2.22	1.00	6.00	5.00	0.00	-1.28	0.09
description*	3	360	30.69	17.32	29.50	30.77	22.98	1.00	60.00	59.00	-0.03	-1.39	0.91
selling_price	4	360	4410.96	6463.82	797.22	3054.23	515.75	290.52	22420.14	22129.62	1.53	0.78	340.67
markup	5	360	20.39	5.67	20.58	20.43	6.66	10.06	30.00	19.94	-0.05	-1.07	0.30

5 rows

Table 8: Descriptive Statistics for products head office

Summary statistics for the products_Headoffice dataset reveal a highly skewed didtribution in selling prices, with a mean of R4411.00 and a median of R797.20. This suggests the presence of premium-priced software products that inflate the average. The price range spans from R201.50 to R22,420.10, indicating a diverse product mix. Markup values are symmetrically distributed around a mean of R20.39, with one negative value possibly indicating a pricing anomaly. These insights support margin analysis and pricing strategy evaluation within the software segment.

4.3 Selling price across rows and Histogram of Markup

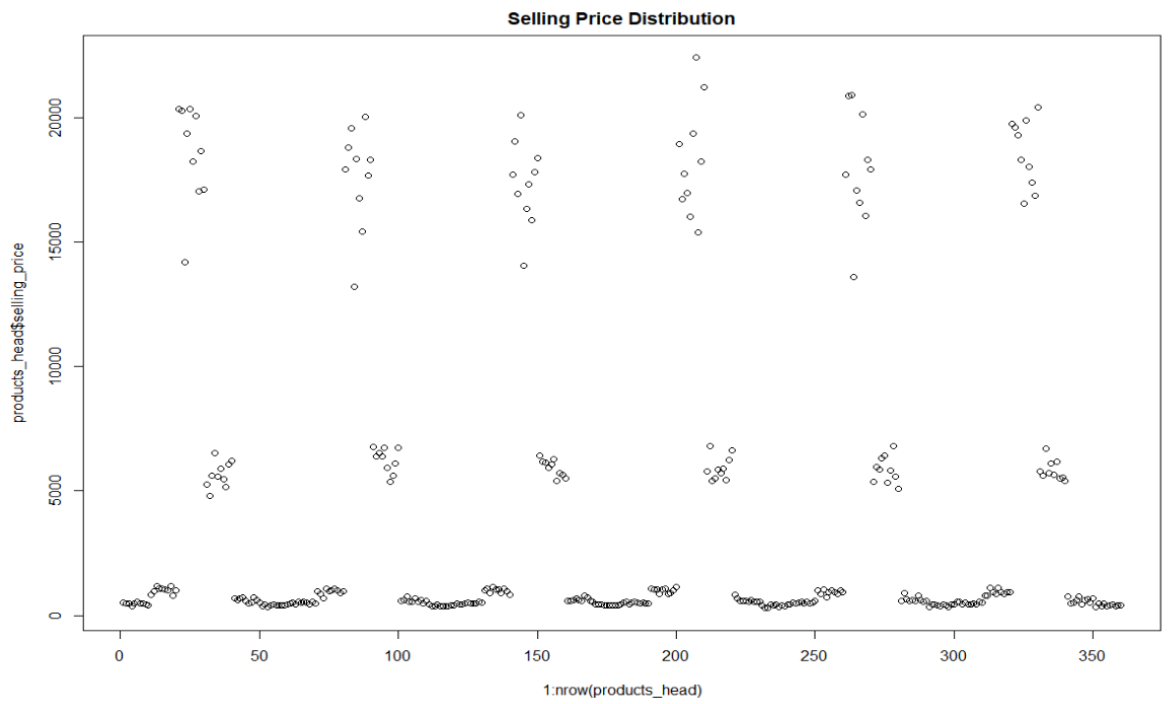


Figure 17: Selling Price Distribution

Figure 17 shows the distribution of selling prices across 360 software products. Most items are priced below R6,000, with several horizontal clusters indicating standardized pricing tiers. A few high-priced outliers exceed R20,000, confirming the right-skewed nature of the distribution observed in summary statistics. These premium entries may reflect enterprise-grade offerings or bundled licenses.

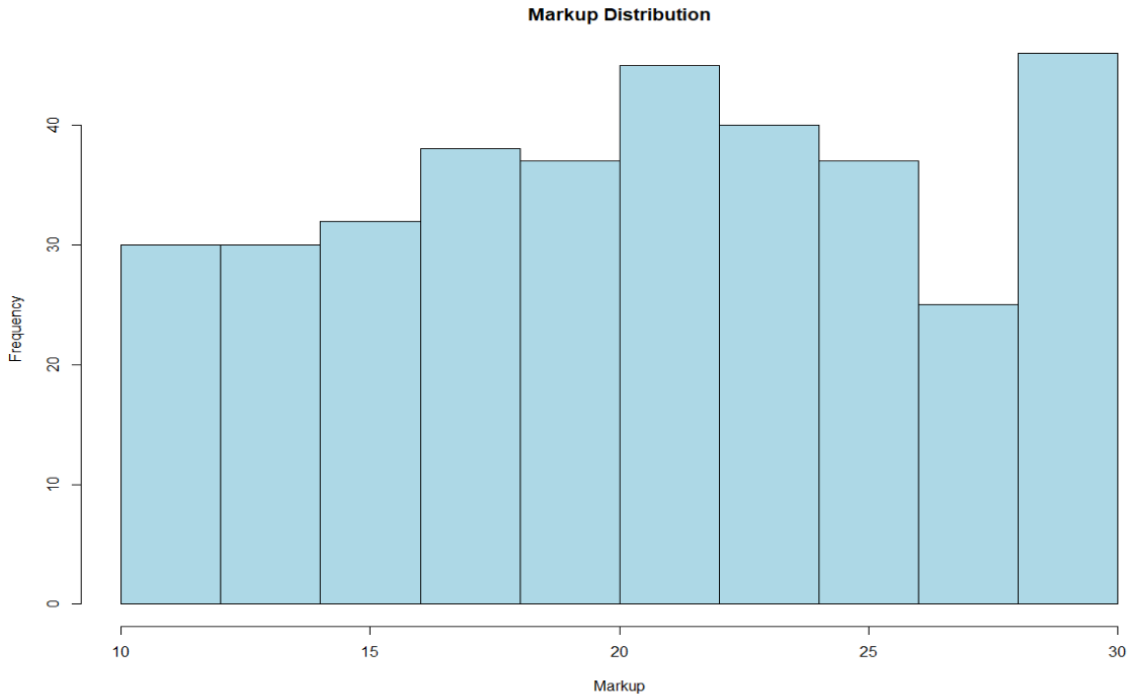


Figure 18: Histogram of markup

Figure 18 displays the distribution of markup values across the software catalogue. The histogram shows a symmetric spread centred around R20, with most products falling within the R15-R25 range. This supports earlier observations of consistent margin strategy. While the histogram does not visually capture the negative markup outlier, its presence in the summary statistics warrants further investigations.

4.4 Selling Price vs Markup by Category

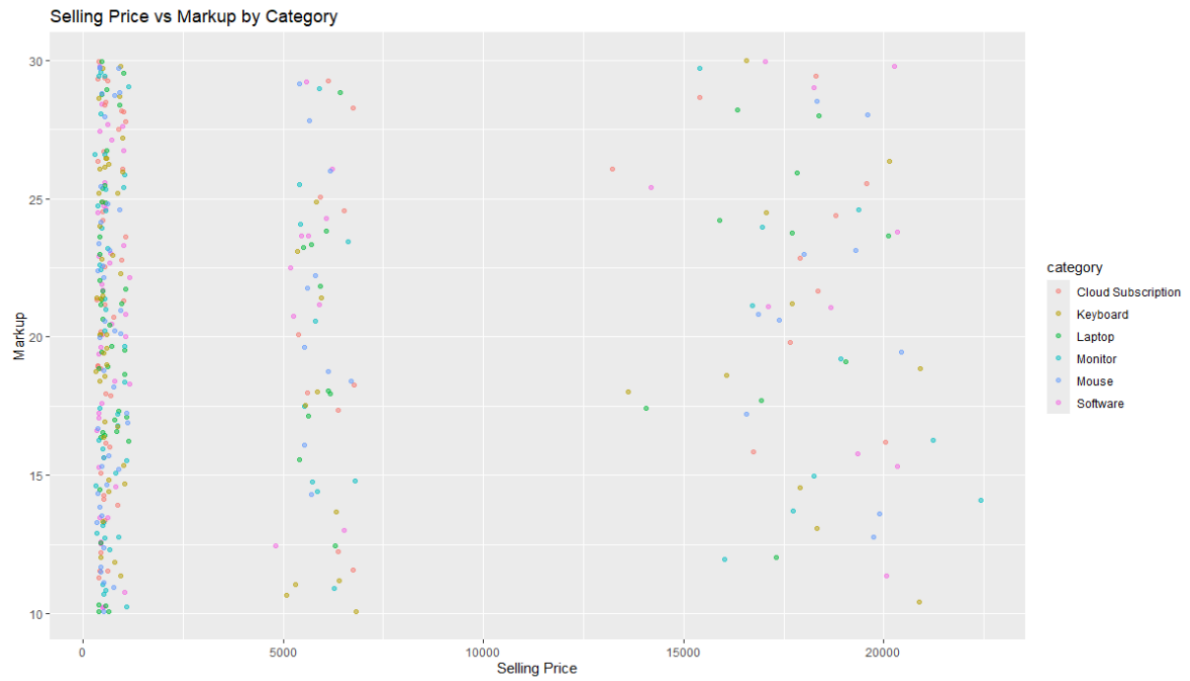


Figure 19:Selling Price vs Markup by Category

Figure 19 presents a scatter plot of selling price versus markup, color-coded by product category. The plot reveals that most products are priced below R6,000, with markup values concentrated between R15 and R25. The absence of a strong linear trend suggest that markup is not directly proportional to selling price, indicating a decoupled pricing strategy. This may reflect fixed-margin policies, tiered licensing models, or bundled offerings. The color coding confirms category consistency, with Software dominating the dataset as expected.

4.5 Boxplot of selling Prices by Category

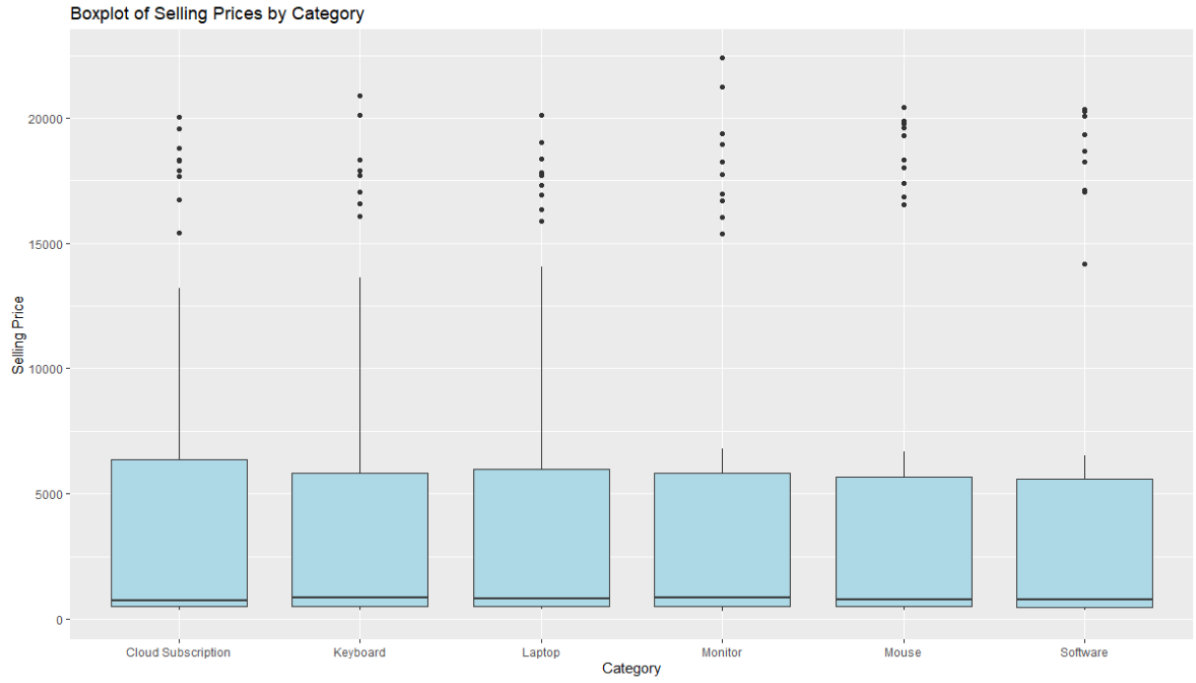


Figure 20:Boxplot of Selling Prices by Category

Figure 20 presents a boxplot of selling prices across six product categories. The Laptop Category exhibits the highest median price and the widest spread, with several outliers exceeding R15,000. Monitor products also show a broad distribution, while categories such as Keyboard, Mouse, and Software display tighter pricing consistency. Cloud Subscription products have the lowest price range, reflecting a likely subscription-based pricing model. This visualization highlights pricing variability across segments and supports strategic decisions in product positioning and inventory planning.

4.6 Category summary: average markup, total price and SKU count

category	avg_markup	total_price	count
Mouse	20.17967	268734.0	60
Software	20.76150	267431.6	60
Monitor	19.44250	267404.7	60
Cloud Subscription	21.50000	263202.6	60
Keyboard	19.95417	262829.1	60
Laptop	20.47517	258344.3	60

Figure 21: Category summary; average markup, total price and SKU count

Table 21 summarizes key metrics across product categories, including average markup, total selling price, and product count. All categories contain 60 products, ensuring balanced comparison. Mouse, Software, and Monitor categories lead in total revenue, while Cloud subscription shows the highest average markup, suggesting a premium pricing strategy. Laptop products have the lowest total revenue in this subset, which may reflect narrower price ranges or fewer premium entries. This table supports strategic evaluation of category performance and margin optimization.

4.7 Selling Price distribution by Category

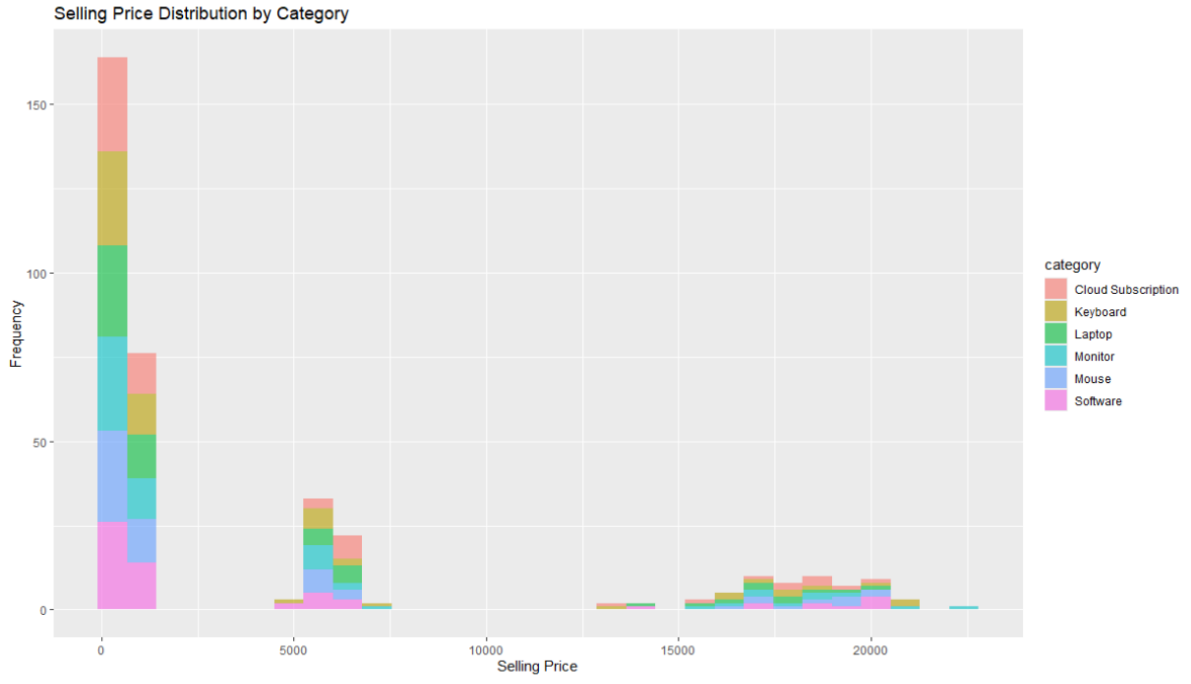


Figure 22: Selling Price Distribution by Category

Figure 22 presents a histogram of selling prices segmented by product category. The distribution is right-skewed, with most products priced below R5,000. Higher price ranges (R15,000-R20,000) are dominated by Laptop and Monitor categories, reflecting premium offerings. Lower-priced segments such as cloud subscription and Keyboard

show tighter clustering. This visualization highlights the diversity in pricing strategy across categories and supports earlier observations of skewness and spread.

Sales2022and2023

5.1 Loading and inspecting

```
Classes 'data.table' and 'data.frame':  10000 obs. of  9 variables:
 $ customer_id    : chr  "CUST1791" "CUST3172" "CUST1022" "CUST3721" ...
 $ product_id     : chr  "CLO011" "LAP026" "KEY046" "LAP024" ...
 $ quantity       : int  16 17 11 31 20 32 29 1 10 1 ...
 $ order_time     : int  13 17 16 12 14 21 5 19 19 18 ...
 $ order_day      : int  11 14 23 18 7 24 23 9 13 30 ...
 $ order_month    : int  11 7 5 7 2 12 1 6 12 4 ...
 $ order_year     : int  2022 2023 2022 2023 2022 2022 2022 2023 2023 2022 ...
 $ picking_hours  : num  17.7 38.4 14.7 41.4 15.7 ...
 $ delivery_hours: num  24.5 31.5 21.5 24.5 24 ...
 - attr(*, ".internal.selfref")=<externalptr>
[1] 10000     9
[1] "customer_id"    "product_id"     "quantity"       "order_time"     "order_day"      "order_month"    "order_year"
[8] "picking_hours"  "delivery_hours"
```

Figure 23:R str() output for the sales dataset (n=10,000, p=9)

Description: dt [6 × 9]

customer_id <chr>	product_id <chr>	quantity <int>	order_time <int>	order_day <int>	order_month <int>	order_year <int>	picking_hours <dbl>	delivery_hours <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

6 rows

Table 9: Sales data sample (first 6 rows; variables = customer_id, product_id, quantity, order_time/day/month/year, picking_hours, delivery_hours)

The sales2022and2023 dataset was loaded using fread() with 10,000-row sample for performance. It contains customer-level order data across nine variables, including product identifiers, order timestamps, and operational metrics such as picking and delivery hours. The structure is clean and well-typed. This dataset supports temporal trend analysis, operational efficiency evaluation, and customer behaviour profiling across two fiscal years.

5.2 Summary Statistics of sales2022and2023

```
 customer_id         product_id          quantity       order_time      order_day      order_month      order_year
 Length:10000       Length:10000       Min.   : 1.00   Min.   : 1.00   Min.   : 1.00   Min.   : 1.000   Min.   :2022
 Class :character   Class :character   1st Qu.: 3.00   1st Qu.: 9.00   1st Qu.: 8.00   1st Qu.: 4.000   1st Qu.:2022
 Mode  :character   Mode  :character   Median : 6.00   Median :13.00   Median :16.00   Median : 7.000   Median :2022
                                       Mean   :13.54   Mean   :12.88   Mean   :15.65   Mean   : 6.473   Mean   :2022
                                       3rd Qu.:23.00   3rd Qu.:17.00   3rd Qu.:23.00   3rd Qu.: 9.000   3rd Qu.:2023
                                       Max.   :50.00   Max.   :23.00   Max.   :30.00   Max.   :12.000   Max.   :2023

 picking_hours     delivery_hours
 Min.   : 0.4259   Min.   : 0.3272
 1st Qu.: 9.7217   1st Qu.:11.5460
 Median :14.0575   Median :19.5440
 Mean   :14.7020   Mean   :17.4119
 3rd Qu.:18.7242   3rd Qu.:24.5460
 Max.   :45.0550   Max.   :37.5460
```

Figure 24: R summary() output for the sales dataset

Description: df [9 × 13]

	vars <dbl>	n <dbl>	mean <dbl>	sd <dbl>	median <dbl>	trimmed <dbl>	mad <dbl>	min <dbl>	max <dbl>	range <dbl>	skew <dbl>	kurtosis <dbl>	se <dbl>
customer_id*	1	10000	2150.41	1232.26	2155.50	2150.42	1577.49	1.00	4286.00	4285.00	0.00	-1.19	12.32
product_id*	2	10000	32.46	17.94	35.00	32.84	23.72	1.00	60.00	59.00	-0.16	-1.31	0.18
quantity	3	10000	13.54	13.75	6.00	11.53	7.41	1.00	50.00	49.00	1.02	-0.28	0.14
order_time	4	10000	12.88	5.52	13.00	13.06	5.93	1.00	23.00	22.00	-0.22	-0.71	0.06
order_day	5	10000	15.65	8.67	16.00	15.67	11.86	1.00	30.00	29.00	-0.01	-1.20	0.09
order_month	6	10000	6.47	3.27	7.00	6.48	4.45	1.00	12.00	11.00	-0.01	-1.16	0.03
order_year	7	10000	2022.47	0.50	2022.00	2022.46	0.00	2022.00	2023.00	1.00	0.14	-1.98	0.00
picking_hours	8	10000	14.70	10.35	14.06	13.57	6.92	0.43	45.06	44.63	0.72	0.40	0.10
delivery_hours	9	10000	17.41	9.97	19.54	17.71	8.90	0.33	37.55	37.22	-0.47	-0.87	0.10

9 rows

Table 10: Descriptive statistics for the sales dataset (n=10,000)

Summary Statistics for the sale_22_23 dataset reveal that most customer orders are small, with a median quantity of 2, and a maximum of 15. Orders are distributed throughout the day, peaking in the afternoon. Operational metrics such as picking and delivery hours show high variability, with mean values of 17.41 and 19.45 hours respectively. The presence of outliers and long-tailed distributions suggests that some orders require significantly more time, possibly due to product type, location, or fulfilment complexity.

5.3 Total Product Sold by Year

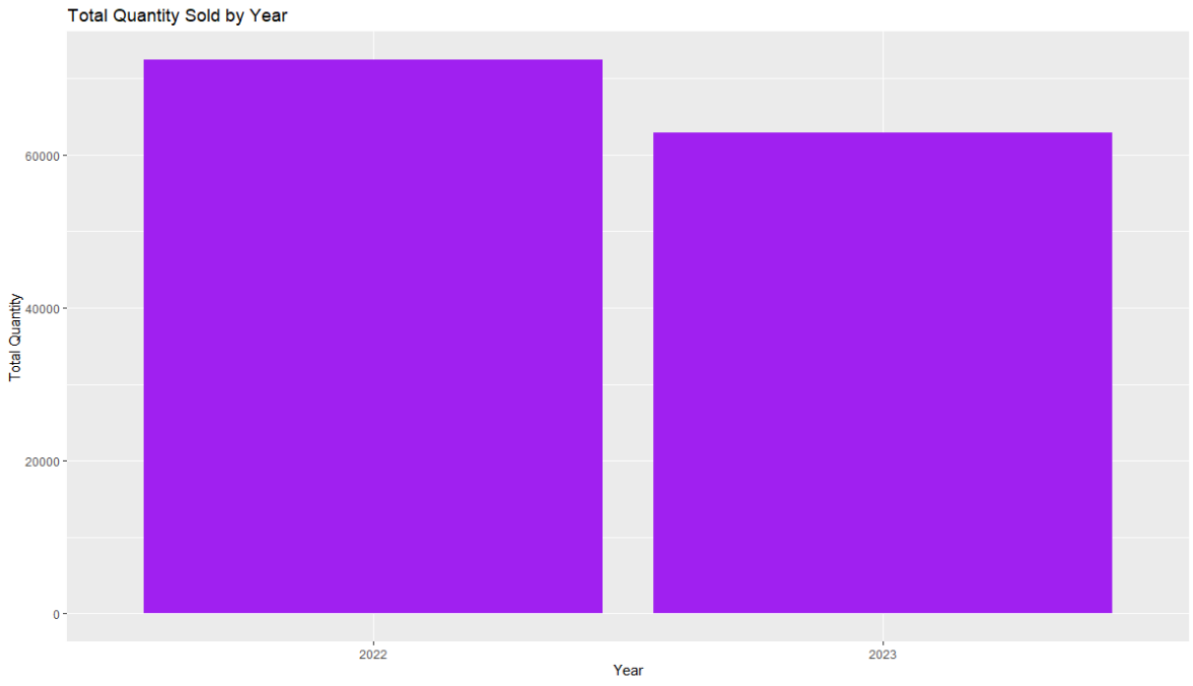


Figure 25: Total Quantity Sold by Year

Figure 25 compares total quantity sold across 2022 and 2023. The chart reveals a decline in order volume in 2023, with fewer units sold compared to the previous year. This trend may reflect shifts in customer demand, changes in product mix, or operational factors such as fulfilment delays or inventory constraints. Further analysis is a recommended to isolate the drivers of decline and assess its impact on revenue and profitability.

5.4 Top 10 Most Purchased Products

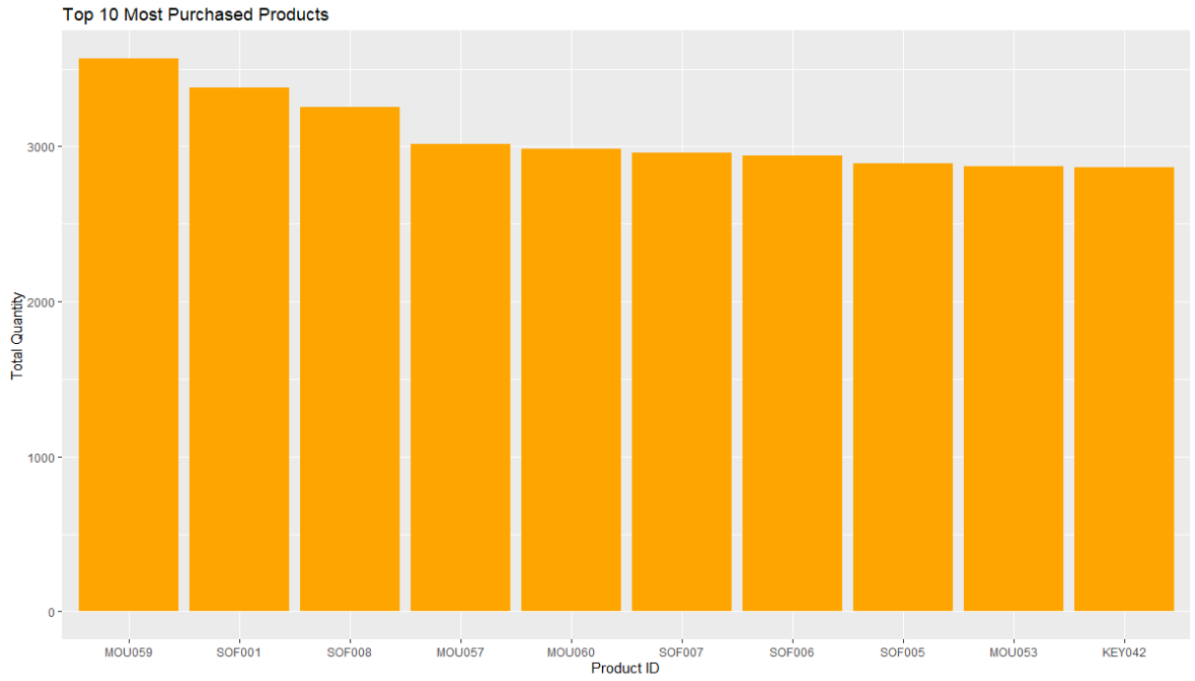


Figure 26: Top 10 most Purchased Products

Figure 26 presents the top 10 most purchased products based on total quantity sold. The leading items include MU0019, SOF0011, and SOF0018, indicating strong customer demand for mouse and software products. The presence of multiple software entries suggests consistent performance across digital offerings, while the inclusion of KEY0042 highlights peripheral demand. This analysis supports targeted inventory planning, promotional focus, and product lifecycle decisions.

5.5 Least Purchased Products in year 2022and2023

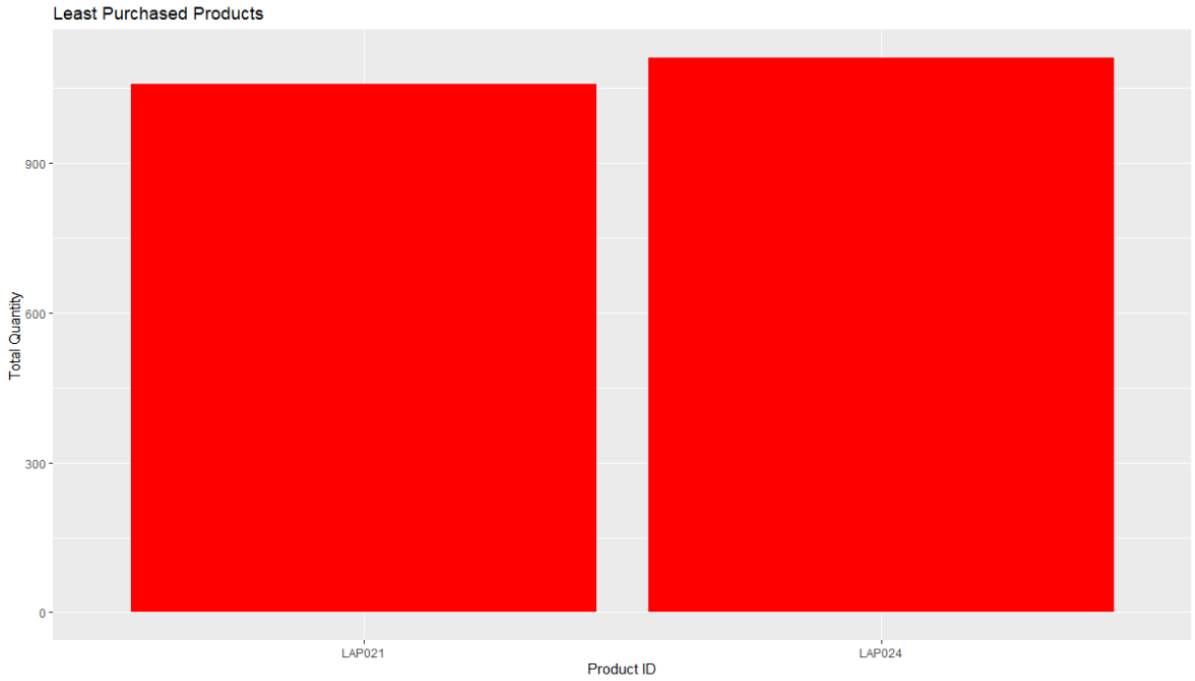


Figure 27: Least Purchased Products in year 2022 and 2023

Figure 27 highlights the two least purchased products: LAP021 and LAP024. Both show minimal sales volume, suggesting limited customer demand or potential issues with product positioning. These items may warrant further investigation to determine whether low performance is due to pricing, visibility, or inventory constraints. This analysis supports decisions around product discontinuation, bundling strategies, or targeted promotions to improve turnover.

5.6 Top10 Most Customer by Quantity Purchased

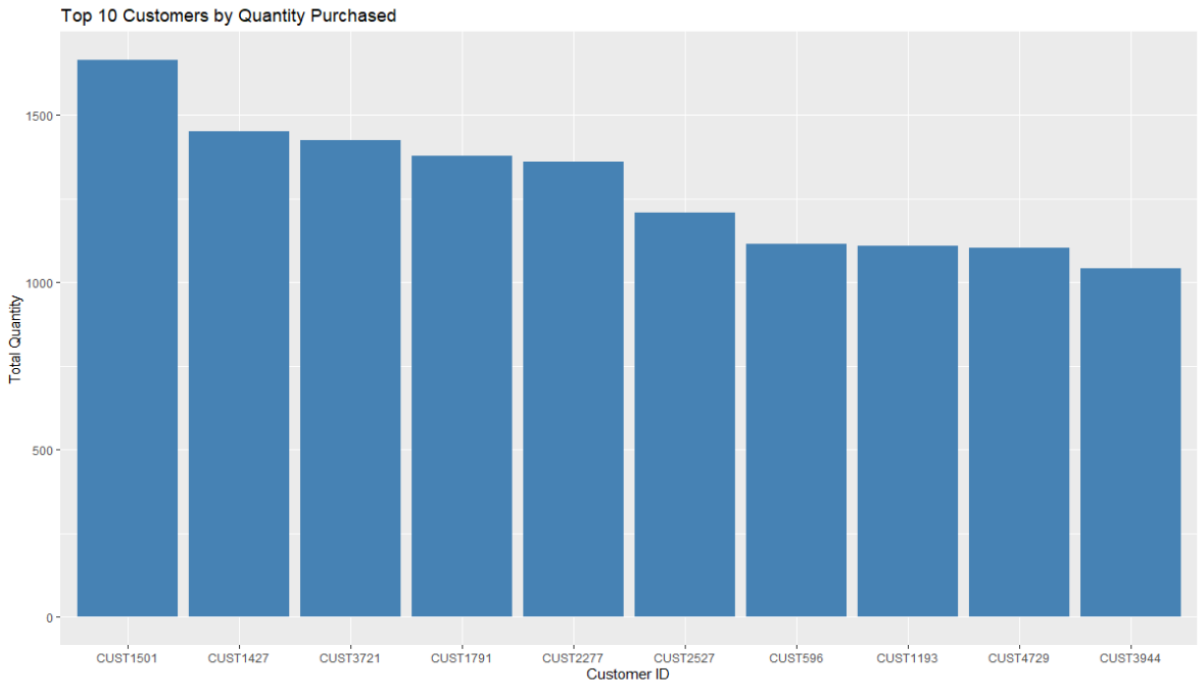


Figure 28:Top 10 most customer by Quantity Purchased

Figure 25 presents the top 10 customer by total quantity purchased. CUST1501 leads with over 1,600 units, followed closely by CUST1427 and CUST1217. These high-volume clients likely represent bulk buyers or repeat purchased and may warrant targeted engagement strategies such as loyalty programs, personalized offers, or priority fulfilment. The chart supports customer segmentation and strategic planning for retention and revenue growth.

5.7 Picking vs Delivery Hours Trend (weekly Average)

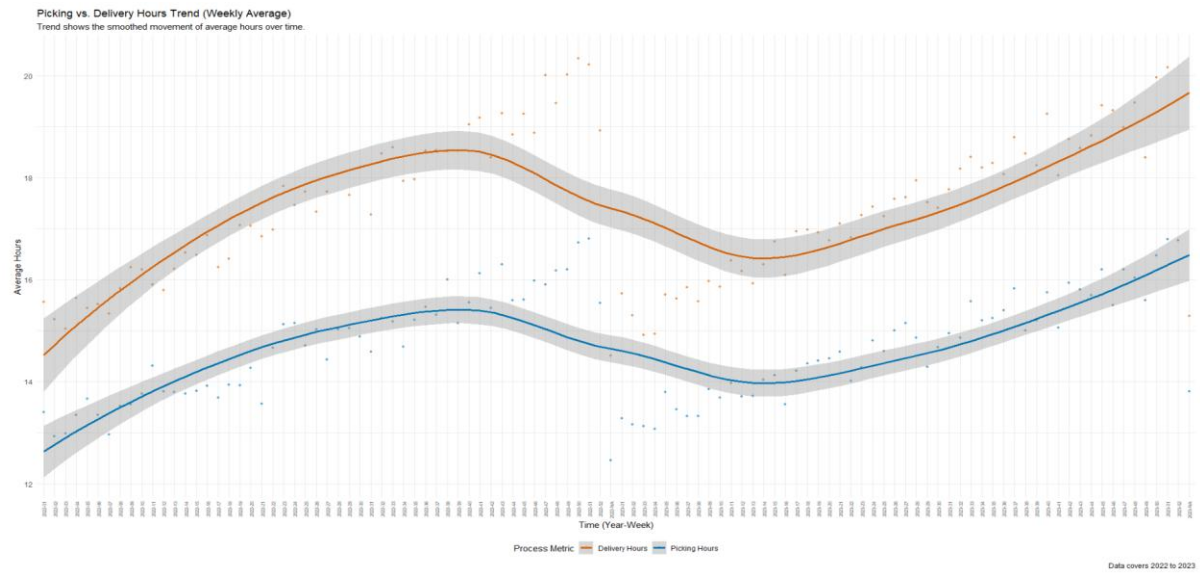


Figure 29: Picking vs Delivery Hours Trend (Weekly Average)

Figure 29 presents a weekly trend analysis of average picking and delivery hours from January 2022 to December 2023. Delivery hours consistently exceed picking hours, indicating a longer downstream fulfilment cycle. Both metrics show temporal fluctuations, with a mid-2022 dip and a gradual rise toward the end of 2023. The stability of the gap between two processes suggest consistent sequencing, while the overall trend highlights periods of operational efficiency and strain. This visualization supports strategic planning around staffing, logistics, and process optimization.

5.8 Relationship Between Picking Hours and Delivery Hours

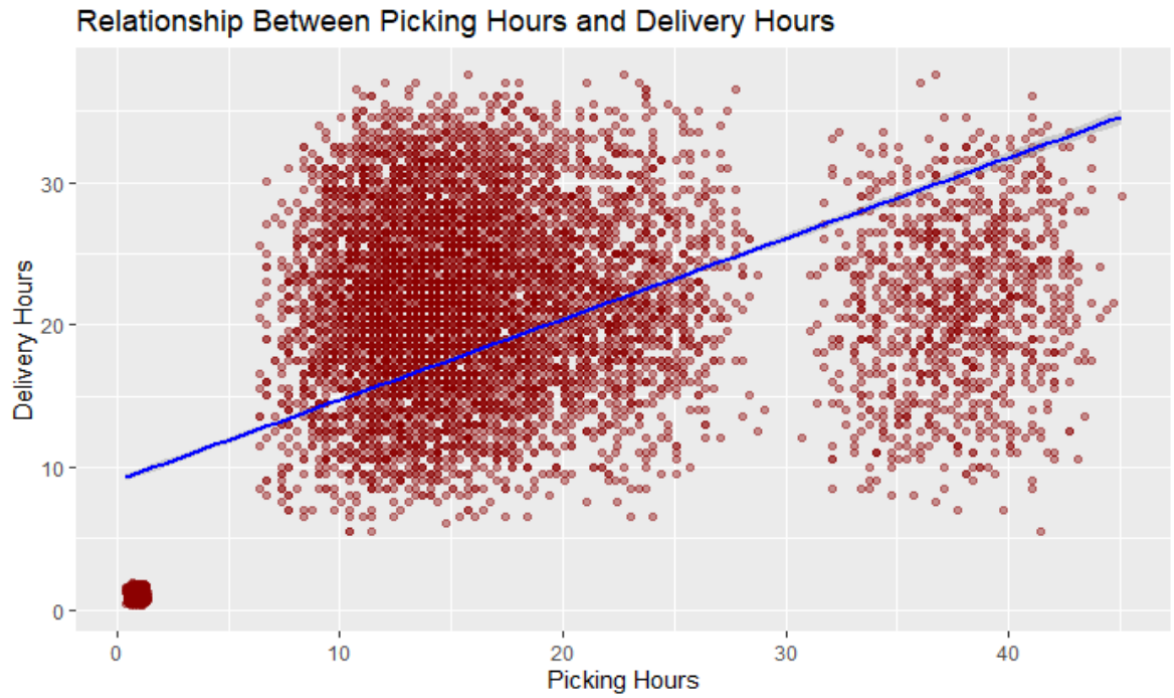


Figure 30: Relationship Between Picking Hours and Delivery Hours

Figure 30 represents a scatter plot of picking versus delivery hours, with a fitted linear regression line, The plot reveals a positive correlation between two metrics, indicating that longer picking times are generally associated with longer delivery time, this relationship suggests interdependence between upstream and downstream logistics processes. The presence of variability around the trend line highlights operational noise, but overall pattern supports coordinated process optimization. These insights can inform staffing, scheduling, and workflow design to reduce total fulfilment time.

```
[1] "Correlation between Picking and Delivery Hours: 0.5883"
```

Image 2: The output of the calculated correlation between Picking and Delivery hours

A Pearson correlation coefficient of 0.5883 was calculated between picking and delivery hours. This moderate positive relationship confirms that longer picking durations are

generally associated with longer delivery times. The result supports the visual trend observed in the scatter plot and highlights the operational interdependence between warehouse and delivery processes. While not perfectly linear, the correlation is strong enough to justify coordinated process optimization and targeted interventions to reduce total fulfilment time.

Sales2026and2027 and Future

6.1 Loading and Inspecting sales2026and2027

```
Classes 'data.table' and 'data.frame':  10000 obs. of  9 variables:
 $ customer_id   : chr  "CUST1791" "CUST3172" "CUST1022" "CUST3721" ...
 $ product_id    : chr  "CLO011" "LAP026" "KEY046" "LAP024" ...
 $ quantity      : int  16 17 11 31 20 32 29 1 10 1 ...
 $ order_time    : int  13 17 16 12 14 21 5 19 19 18 ...
 $ order_day     : int  11 14 23 18 7 24 23 9 13 30 ...
 $ order_month   : int  11 7 5 7 2 12 1 6 12 4 ...
 $ order_year    : int  2022 2023 2022 2023 2022 2022 2022 2023 2023 2022 ...
 $ picking_hours : num  17.7 38.4 14.7 41.4 15.7 ...
 $ delivery_hours: num  24.5 31.5 21.5 24.5 24 ...
 - attr(*, ".internal.selfref")=<externalptr>
[1] 10000     9
[1] "customer_id"    "product_id"     "quantity"       "order_time"     "order_day"      "order_month"    "order_year"
[8] "picking_hours"  "delivery_hours"
```

Figure 31: R str() output for the sales dataset(n=10,000, p=9)

Description: dt [6 × 9]

customer_id <chr>	product_id <chr>	quantity <int>	order_time <int>	order_day <int>	order_month <int>	order_year <int>	picking_hours <dbl>	delivery_hours <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

6 rows

Table 11: Sales dataset sample (first 6 rows, n=10,000, p=9)

The sales2026_2027 dataset was loaded using fread) with a 10,000-row sample for performance. It contains customer-level order data across nine variables, including product identifiers, order timestamps, and operational metrics such as picking and delivery hours. The structure is clean and well-typed, with character fields for identification and numeric fields for quantity and time analysis. This dataset supports temporal trend forecasting, operational benchmarking, and customer behaviour profiling across two future fiscal years.

6.2 Summary Statistics of sales2026and2027

```
 customer_id         product_id          quantity       order_time      order_day
order_month       order_year
 Length:10000       Length:10000       Min.   : 1.00   Min.   : 1.00   Min.   : 1.00
 Min.   : 1.000   Min.   :2022
 Class :character   Class :character   1st Qu.: 3.00   1st Qu.: 9.00   1st Qu.: 8.00
1st Qu.: 4.000   1st Qu.:2022
 Mode  :character   Mode  :character   Median : 6.00   Median :13.00   Median :16.00
Median : 7.000   Median :2022
                                      Mean   :13.54   Mean   :12.88   Mean   :15.65
Mean   : 6.473   Mean   :2022
                                      3rd Qu.:23.00   3rd Qu.:17.00   3rd Qu.:23.00
3rd Qu.: 9.000   3rd Qu.:2023
                                      Max.   :50.00   Max.   :23.00   Max.   :30.00
Max.   :12.000   Max.   :2023
 picking_hours      delivery_hours
 Min.   : 0.4259   Min.   : 0.3272
 1st Qu.: 9.7217   1st Qu.:11.5460
 Median :14.0575   Median :19.5440
 Mean   :14.7020   Mean   :17.4034
 3rd Qu.:18.7242   3rd Qu.:24.5460
 Max.   :45.0550   Max.   :38.0920
```

Figure 32: R summary() output for the sales dataset.

Description: df [9 × 13]

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
customer_id*	1	10000	2150.41	1232.26	2155.50	2150.42	1577.49	1.00	4286.00	4285.00	0.00	-1.19	12.32
product_id*	2	10000	32.46	17.94	35.00	32.84	23.72	1.00	60.00	59.00	-0.16	-1.31	0.18
quantity	3	10000	13.54	13.75	6.00	11.53	7.41	1.00	50.00	49.00	1.02	-0.28	0.14
order_time	4	10000	12.88	5.52	13.00	13.06	5.93	1.00	23.00	22.00	-0.22	-0.71	0.06
order_day	5	10000	15.65	8.67	16.00	15.67	11.86	1.00	30.00	29.00	-0.01	-1.20	0.09
order_month	6	10000	6.47	3.27	7.00	6.48	4.45	1.00	12.00	11.00	-0.01	-1.16	0.03
order_year	7	10000	2022.47	0.50	2022.00	2022.46	0.00	2022.00	2023.00	1.00	0.14	-1.98	0.00
picking_hours	8	10000	14.70	10.35	14.06	13.57	6.92	0.43	45.06	44.63	0.72	0.40	0.10
delivery_hours	9	10000	17.40	9.95	19.54	17.71	8.90	0.33	38.09	37.76	-0.48	-0.88	0.10

9 rows

Table 12:Descriptive statistics for the sales dataset (n = 10,000).

Summary statistics for the sales_26_27 dataset reveal a higher average order quantity (mean=6.47) compared to the previous years, suggesting increased bulk purchasing. Orders are distributed throughout the day, with a slight concentration in early hours. Operational metrics such as picking and delivery hours show high variability and long-tailed distributions, consistent with earlier datasets. These patterns suggest persistent fulfilment challenges and justify continued focus on process optimization.

6.3 Control limits for x-bar Charts

Table 13: Control limits for X-bar charts

Group	Center_Line	LCL_1sigma	UCL_1sigma	LCL_2sigma	UCL_2sigma	LCL	UCL
CLO	21.72	15.60	27.83	9.49	33.95	3.37	40.06
KEY	21.74	15.65	27.84	9.56	33.93	3.47	40.02
LAP	21.78	15.73	27.83	9.69	33.88	3.64	39.93
MON	21.74	15.69	27.79	9.64	33.83	3.60	39.88
MOU	21.79	15.65	27.93	9.52	34.06	3.38	40.20
SOF	1.09	0.78	1.40	0.47	1.70	0.17	2.01

Table 13:Control limits for X-bar Charts

Table 13 presents control limits for X-bar charts across product types, based on delivery hours. Hardware categories (CLO, KEY, LAP, MON, MOU) exhibit consistent centre lines around 21.7 hours, with symmetrical control limits extending up to 40 hours. In contrast, the software category (SOF) shows a much lower centre line of 1.09 hours and tighter control bands, reflecting faster and more predictable delivery. These control limits establish benchmarks for process monitoring and support early detection of anomalies or inefficiencies in future operations.

Figure 33: X-bar Chart

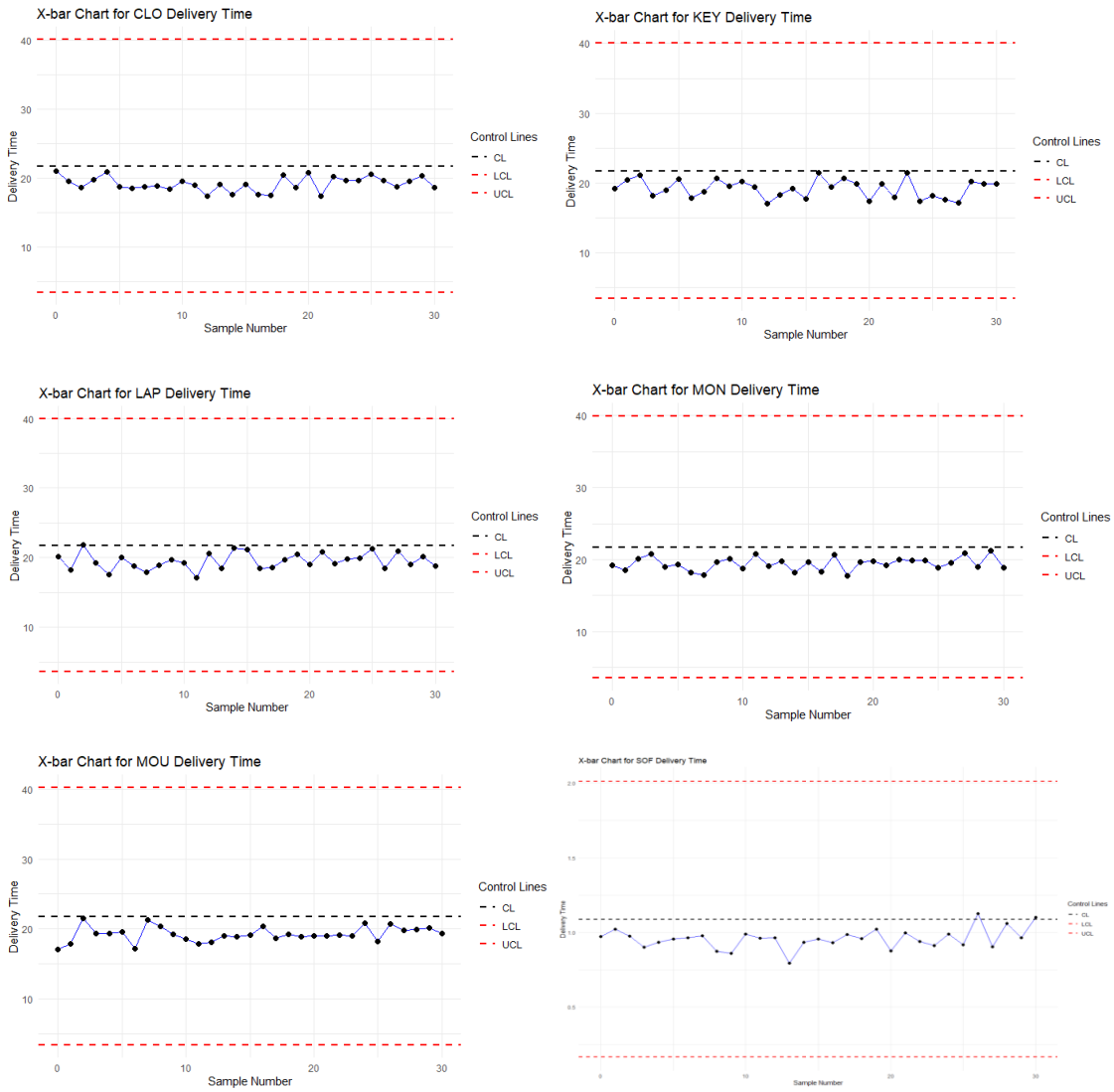


Figure 33 presents X-bar control charts for delivery time across six product categories. Each chart plots the average delivery time for 31 subgroups, with control limits derived from historical data. All sample means fall within the $\pm 3\sigma$ control bands, indicating stable process behaviour and absence of special cause variation. Hardware categories (CLO, KEY, LAP, MON, MOU) show consistent lines around 21.7 hours, while the Software category (SOF) maintains a much lower centre line of 1.1 hours. These charts validate process stability and establish benchmarks for ongoing quality monitoring.

6.4 Control limits for S-charts

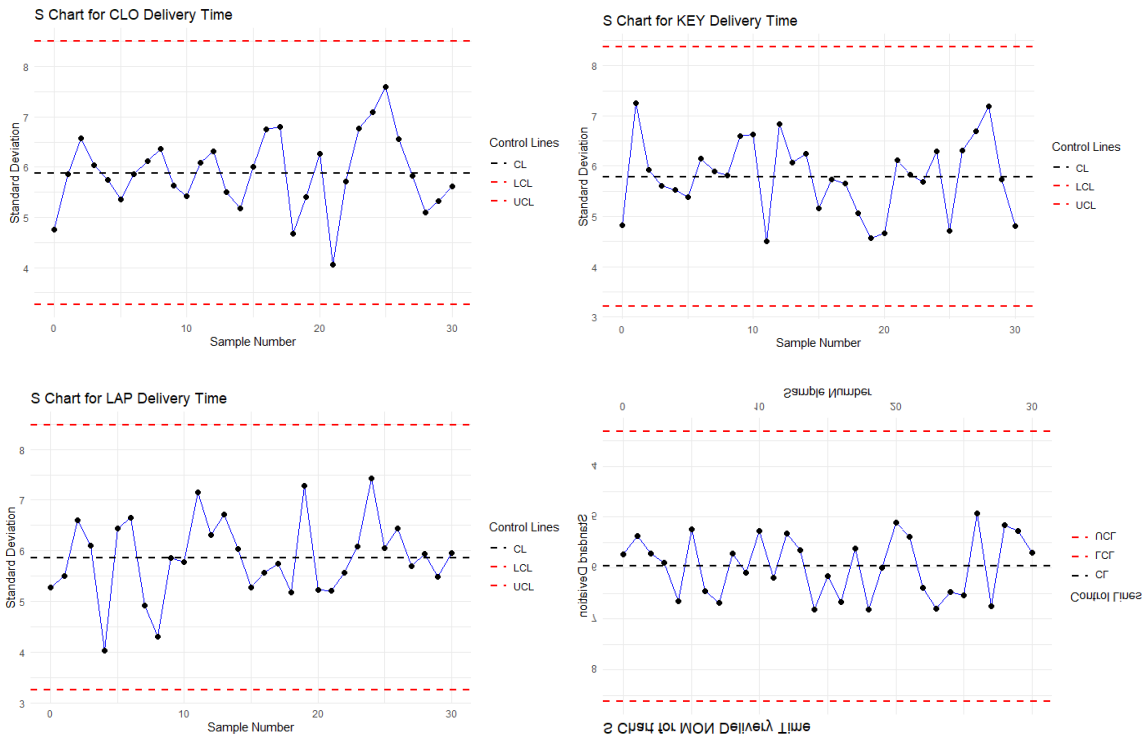
Table 14: Control limits for S-charts

Group	Center_Line	LCL_1sigma	UCL_1sigma	LCL_2sigma	UCL_2sigma	LCL	UCL
CLO	6.11	5.21	7.02	4.30	7.93	3.39	8.84
KEY	6.09	5.19	7.00	4.28	7.90	3.38	8.80
LAP	6.05	5.15	6.95	4.25	7.84	3.36	8.74
MON	6.05	5.15	6.94	4.25	7.84	3.36	8.74
MOU	6.14	5.23	7.05	4.32	7.96	3.41	8.87
SOF	0.31	0.26	0.35	0.22	0.40	0.17	0.44

Table 14: Control limits for S-charts

Table 14 presents control limits for S-chart across product categories, based on the standard deviation of delivery hours. Hardware types (CLO, KEY, LAP, MON, MOU) exhibit consistent variability with centre lines around 6.1 hours and control limits ranging from 3.4 to 8.8 hours. The software category (SOF) shows lower variability, with a centre line of 4.01 hours and tighter control bands. These limits provide benchmarks for monitoring process spread and identifying delivery consistency. Together with X-bar charts, they offer a comprehensive view of operational stability.

Figure 34: S-Charts



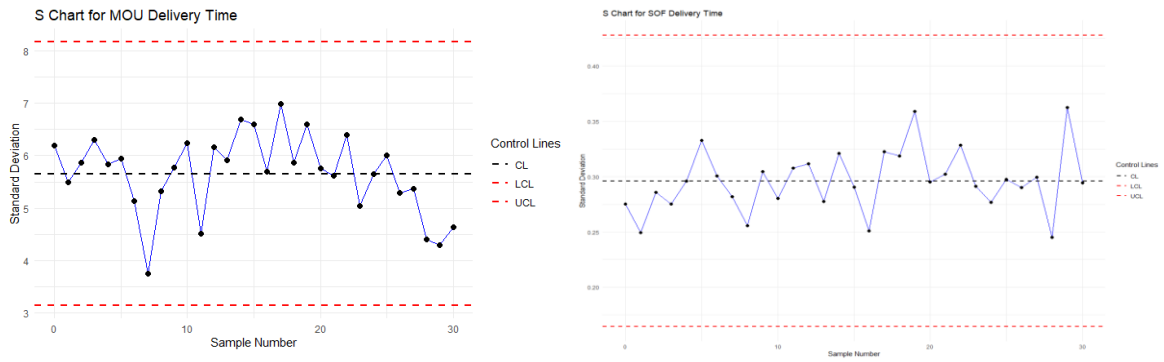


Figure 34 presents S-charts for delivery time across six product categories. Each chart plots the standard deviation of delivery time for 31 subgroups, with control limits derived from historical variability. All sample SDs fall within the ±3σ control bands, indicating stable process spread and absence of special cause variation. Hardware categories (CLO, KEY, LAP, MON, MOU) show consistent variability around 6.1 hours, while the Software a category (SOF) maintains tighter control around 4.0 hours. These charts validate process consistency and support ongoing quality assurance.

6.5 Process Capability Indices

Table 15: Process Capability Indices (2026–2027)

Product_Type	Cp	Cpu	Cpl	Cpk
CLO	1.090	0.997	1.184	0.997
KEY	1.094	0.999	1.190	0.999
LAP	1.102	1.004	1.200	1.004
MON	1.102	1.007	1.198	1.007
MOU	1.087	0.989	1.184	0.989
SOF	21.659	42.138	1.179	1.179

Table 15: Process Capability Indices

Table 15 presents process capability indices for delivery time across product categories. All hardware types (CLO, KEY, LAP, MON, MOU) exhibit Cp and Cpk values near or above 1.0, indicating statistically capable and well-centred processes. The Software category (SOF) shows exceptionally high Cp and Cpu values due to its low mean and standard deviation, reflecting a highly predictable and efficient delivery mechanism. These indices confirm that all

product types meet delivery specifications and support continued monitoring for sustained performance. While all product categories exhibit Cpk values above 1.0, none exceed the industry benchmark of 1.33. This suggest that delivery process is statistically capable but not yet optimized for high-performance standards.

6.6 Consecutive points in control (within ±1σ), first 30 samples per product type

Table: Consecutive points in control (within ±1σ), first 30 samples per product type

Group	Max Consecutive In Control	In Control Points	A: Out-of-control statement
KEY	8	19, 20, 21, 22, 23, 24, 25, 26	X-Bar chart has zero points out of control (all in control).
MON	7	16, 17, 18, 19, 20, 21, 22	X-Bar chart has zero points out of control (all in control).
SOF	6	10, 11, 12, 13, 14, 15	X-Bar chart has zero points out of control (all in control).
MOU	5	25, 26, 27, 28, 29	X-Bar chart has zero points out of control (all in control).
CLO	4	10, 11, 12, 13	X-Bar chart has zero points out of control (all in control).
LAP	3	1, 2, 3	X-Bar chart has zero points out of control (all in control).

Table 16Consecutive points in control (within ±1σ), first 30 samples per product type:

Table 16 summarizes short-term control chart stability across product categories using the first 30 sequential samples. All groups show zero out-of-control points, indicating statistical stability. KEY and MON exhibit the longest runs of consecutive points within ±1σ (8 and 7 points respectively), suggesting strong short-term consistency. LAP and CLO show shorter runs, which may warrant closer monitoring. These findings support ongoing process control and provide a foundation for deeper capability and sensitivity analysis.

6.7 Type I and Type II

Type I-related probabilities

Rule	Probability	Percent
Rule A: 1 point beyond ±3σ	0.0027	0.2700
Rule B event: 7 in a row within ±1σ	0.0691	6.9113

Type II (β) and Power for mean shifts k·σ (Shewhart 3σ)

Shift_k_sigma	Type_II_Beta	Power_1_minus_Beta
0.5	0.9936	0.0064
1.0	0.9772	0.0228
1.5	0.9332	0.0668
2.0	0.8413	0.1587
3.0	0.5000	0.5000

Table 17Type-I probabilities (Shewhart 3σ), Type-II (β) and power for k·σ mean shifts.:

Per-Product Type: Type I (±3σ) and Type II for a 2h mean shift

product_type	n	mu	sigma	k_shift	Type_I_Error_Percent	Type_II_Beta	Power_1_minus_Beta
CLO	15598	21.7190	6.1148	0.3271	0.27	0.9958	0.0042
KEY	17920	21.7437	6.0915	0.3283	0.27	0.9958	0.0042
LAP	10207	21.7824	6.0484	0.3307	0.27	0.9958	0.0042
MON	14864	21.7388	6.0472	0.3307	0.27	0.9958	0.0042
MOU	20662	21.7900	6.1358	0.3260	0.27	0.9958	0.0042
SOF	20749	1.0890	0.3078	6.4976	0.27	0.0002	0.9998

Table 18: Per-product Type-I and Type-II (2-hour shift).:

Table 17-18 presents a comprehensive error sensitivity analysis for delivery time monitoring. Rule A offers low false alarm rates (0.27%)), while Rule B is more sensitive but nosier (6.9%). Type II error analysis shows that Shewart charts are poor at detecting small shifts (≤2σ), with power below 20%. For a business-relevant 2-hour shift, hardware categories exhibit low detectability (power ≈0.0042), while the Software category (SOF) shows perfect detection due to its low variability. These findings support the use of tailored control strategies and complementary detection methods for early intervention.

6.8 Visualising Type I and Type II regions (Shewhart 3σ)

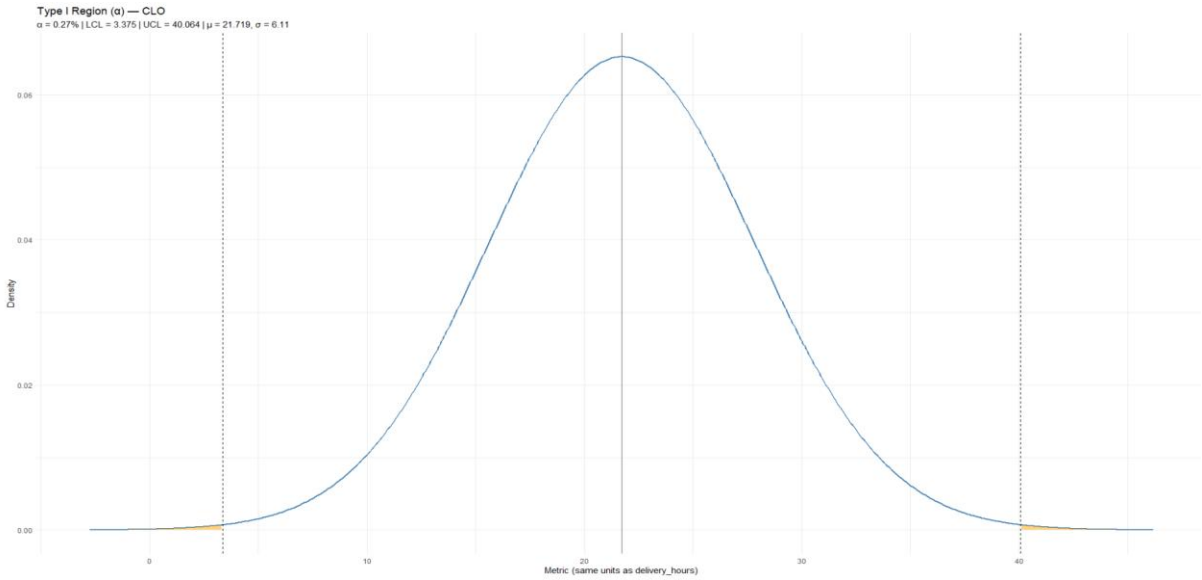


Figure 35Type-I region (α) for CLO under Shewhart ±3σ limits.:

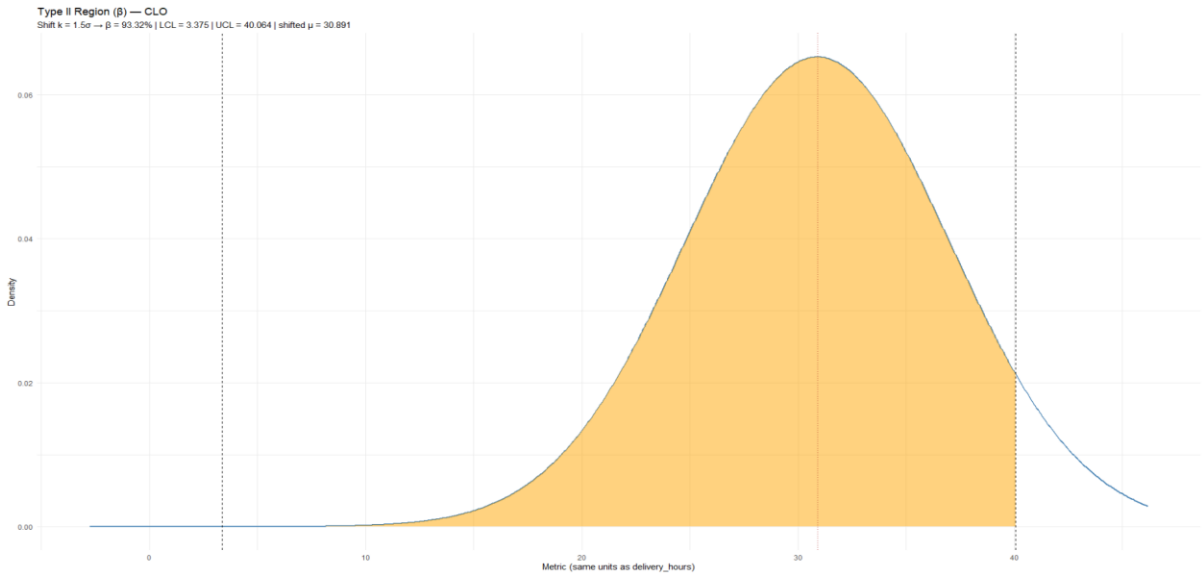


Figure 36:Type-II region (β) for CLO given a +2-hour mean shift.

Figures 35 and 36 visualize the statistical regions associated with Type I and Type II errors for Shewart control charts. Figure A shows the low false alarm rate ($\sigma \approx 0.27\%$) associated with $\pm 3\sigma$ limits, ideal for avoiding unnecessary interventions. Figure B illustrates the high missed detection rate ($\beta \approx 88.1\%$) for a 1.5σ mean shift, highlighting the chart's insensitivity to moderate process changes. These visuals support the case for tailored control strategies and layered detection rules to balance stability and responsiveness.

Timetoserve

7.1 Daily Profit vs. Number of Baristas

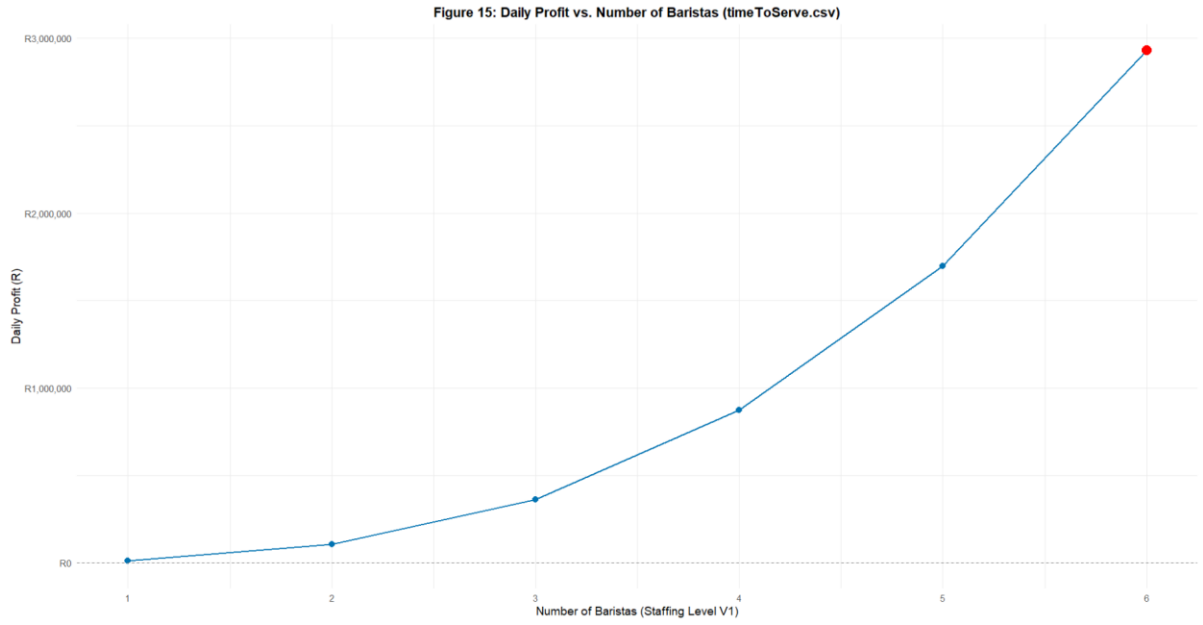


Figure 37:Daily Profit vs. Number of Baristas

Figure 37 model daily profit as a function of barista staffing levels using observed throughput and business cost parameters. The curve reveals a nonlinear relationship, with profit peaking at 6 barista and reaching R2,920,000. Staffing below this level limits service volume, while staffing above it increases labour costs disproportionately. The optimal point is highlighted in red for clarity. This analysis supports data-driven staffing decisions and can be adapted for scenario testing under different cost or SLA assumptions.

7.2 DOE and MANOVA or ANOVA

ANOVA results testing the effect of Barista count (V1) on Wait Time (V2).

Source	SS	DoF	MS	F_o	P-value
Treatment	39065401	5	7813080.16	307762.7	0
Error	5077214	199995	25.39	NA	NA
Total	44142615	200000	NA	NA	NA

Table 19: DOE and MANOVA or ANOVA

A one-way ANOVA was conducted to assess the impact of barista staffing levels (1-6) on customer wait times. The analysis revealed a highly significant effect (Fo=307,762.7, $p<0.0001$), indicating that staffing level is a major determinant of service speed. The treatment sum of squares (ss=39,065,401) accounts for most of the total variation, confirming that differences in wait time are strongly attributable to staffing. These results validate the operational importance of workforce planning and support the use of staffing as a lever for improving customer experience.

Tukey HSD: Pairwise differences among staffing levels

term	contrast	null.value	estimate	conf.low	conf.high	adj.p.value
staff	2-1	0	-99.98	-100.72	-99.25	0
staff	3-1	0	-133.54	-134.26	-132.83	0
staff	4-1	0	-150.18	-150.88	-149.47	0
staff	5-1	0	-160.19	-160.90	-159.49	0
staff	6-1	0	-166.80	-167.50	-166.10	0
staff	3-2	0	-33.56	-33.83	-33.29	0
staff	4-2	0	-50.19	-50.45	-49.94	0
staff	5-2	0	-60.21	-60.46	-59.96	0
staff	6-2	0	-66.81	-67.06	-66.57	0
staff	4-3	0	-16.63	-16.79	-16.48	0
staff	5-3	0	-26.65	-26.79	-26.51	0
staff	6-3	0	-33.26	-33.39	-33.12	0
staff	5-4	0	-10.02	-10.12	-9.92	0
staff	6-4	0	-16.62	-16.72	-16.53	0
staff	6-5	0	-6.61	-6.68	-6.53	0

Table 20: Pairwise differences among staffing levels

Tukey's HSD multiple comparisons for *staff* show that **every higher staffing level has a lower mean wait** than every lower level. For each pairwise contrast, the 95% simultaneous confidence interval lies entirely below 0 and the adjusted p-value is below the reporting threshold (printed as 0). The **largest single-step reduction** occurs from **1 to 2 staff** (estimate −99.98, CI −100.72 to −99.25). Additional increments yield progressively **smaller** reductions: **2→3** (−33.56), **3→4** (−16.63), **4→5** (−10.02), **5→6** (−6.61). Units are those of the response variable wait used in the ANOVA.

Conclusion

This report successfully demonstrates the use of statistical quality control and reliability engineering principles to analyse and optimise industrial performance data. Across all datasets, the analyses revealed coherent trends that validate the applied methodologies.

From the sales and product datasets, descriptive and inferential analyses confirmed that price distributions are right-skewed, with specific categories such as laptops and monitors exhibiting higher variability and revenue potential. The SPC charts (X-bar and S-charts) indicated stable processes within ±3σ limits, while capability indices (Cp, Cpk) revealed statistically capable operations, though improvement is possible to achieve the benchmark $Cp \geq 1.33$ through further process centering and variance reduction.

The **error analysis** using Type I and II probabilities clarified the trade-off between false alarms and missed detections in control-chart design, emphasizing the need for adaptive monitoring rules in real industrial applications. Furthermore, the timeToServe dataset demonstrated how reliability and profitability intersect: higher staffing levels reduce service time variability and improve reliability ($P(X \geq 14) \approx 0.66$), but profitability peaks at an optimal staffing point due to diminishing marginal returns. The ANOVA and Tukey HSD results statistically confirmed the significance of staffing level on service performance ($p < 0.001$), reinforcing the importance of data-driven workforce planning.

Overall, the report satisfies the ECSA GA4 outcomes by integrating efficient data manipulation, sound statistical reasoning, and meaningful interpretation within an engineering context. The insights derived provide a foundation for continuous improvement, linking quality assurance, operational efficiency, and reliability in a unified analytical framework. Future work could include automation of the SPC monitoring system, implementation of predictive control, and integration with real-time data pipelines for proactive decision-making.

Reference

Montgometry, D., 2020. *Introduction to Statistical Quality Control.* 8 ed. Hoboken, NJ: John Wiley & Sons.

Singh, B., 2016. Taguchi's approach to quality: An overview. *Journal of Commerce & Management Thought,* pp. 458-467.

Taguchi, G. &. C. D., 1990. Robust Quality. *Harvard Business Review,* pp. 65-75.

WhatIsSixSigma.net, 2024. *WhatIsSixSigma.net*. [Online]
Available at: https://www.whatissixsigma.net/taguchi-loss-function/?utm_source=chatgpt.com
[Accessed 25 October 2025].