QUALITY ASSURANCE 344
ECSA REPORT

Marcus Christiaan Esterhuyse

26426137@sun.ac.za

Table of Contents

List of figures

List of tables

1 Introduction

This comprehensive data analysis project was conducted as part of the QA344 module requirements to demonstrate competency in ECSA Graduate Attribute 4 (GA4): Engineering Problem Solving through data analysis and manipulation. The project represents a real-world industrial engineering scenario requiring the application of statistical methods to optimize business processes, improve quality control, and enhance operational decision-making.

In today's data-driven industrial landscape, the ability to extract meaningful insights from complex datasets is fundamental to engineering excellence. The ability to sift through, process, and interpret vast amounts of data is a core function of business operations today. Accurate, well-considered, and efficiently implemented data analysis can lead to significant benefits throughout the entire organizational structure (CData Software, 2024).

This project addresses multiple business challenges across different operational contexts, from manufacturing process control to service industry optimization, reflecting the diverse applications of industrial engineering principles.

1.1 Project Objectives and Scope

The primary objective of this project is to conduct end-to-end data analysis across seven key areas, each addressing distinct engineering challenges:

i. Descriptive Statistics Analysis: Comprehensive exploration of customer behaviour, product performance, and sales trends to establish baseline understanding.
ii. Statistical Process Control (SPC): Implementation of quality control mechanisms for delivery process monitoring and capability assessment.
iii. Risk Analysis: Quantification of Type I and Type II errors in quality control systems.
iv. Service Optimization: Profit maximization through staffing optimization in service industry contexts.
v. Multivariate Analysis: Investigation of complex relationships between multiple operational factors using MANOVA.
vi. Reliability Modelling: Workforce optimization using probability distributions to balance costs and service quality.

The analysis spans multiple domains including retail sales, manufacturing quality control, service industry operations and workforce management. It demonstrating the versatility of industrial engineering methodologies.

2 Methodology

2.1 Analytical Framework and Tools

This project employed a comprehensive statistical analysis framework utilizing the R programming language for all computational work. The analytical approach followed a structured quality engineering methodology, progressing from exploratory data analysis to advanced statistical modelling and optimization.

2.1.1 Software and Tools

- Primary Platform: R Statistical Software (Version 4.3.0+)
- Integrated Development Environment: RStudio
- Key R Packages:
  - dplyr, tidyr for data manipulation and wrangling.
  - ggplot2 for advanced data visualization.
  - Base R functions for statistical computations.
  - Specialized packages for specific analyses as required.

2.1.2 Documentation and Reporting

- RMarkdown for reproducible research documentation.
- HTML and PDF output for result presentation.
- Microsoft Word for final formal report compilation.

2.2 Data Management and Preparation

2.2.1 Data Sources and Structure

The analysis utilized multiple datasets provided for the ECSA project these included:

i. Customer demographic data (*customer_data.csv*)
ii. Product information datasets (*products_data.csv, products_data2025.csv*)
iii. Head office product records (*products_Headoffice.csv, products_Headoffice2025.csv*)
iv. Sales transaction data (*sales2022and2023.csv*)
v. Coffee shop service time data (*timeToServe.csv, timeToServe2.csv*)

2.2.2 Data Quality Assurance

- A rigorous data cleaning protocol was implemented:
- Comprehensive missing value analysis and treatment
- Data type validation and conversion
- Temporal data structuring for time-series analysis
- Cross-dataset consistency verification
- Implementation of data correction protocols as specified in project requirements

2.3 Statistical Methods Deployed

2.3.1 Descriptive Statistics and Exploratory Data Analysis

i. Measures of Central Tendency: Mean, median for continuous variables.
ii. Measures of Dispersion: Standard deviation, variance, interquartile range.
iii. Frequency Distributions: For categorical variable analysis.
iv. Data Visualization: Histograms, box plots, scatter plots, bar charts.
v. Correlation Analysis: Pearson correlation coefficients for variable relationships.

2.3.2 Statistical Process Control (SPC)

i. Control Chart Methodology: X-bar and s-charts for process monitoring
ii. Sample Design: Rational subgrouping with n=24 samples
iii. Control Limits: 1σ, 2σ, and 3σ limits calculated from initial 30 samples
    - Out-of-Control Detection: Western Electric rules implementation
    - Rule A: Points outside 3σ limits
    - Rule B: Consecutive points within 1σ limits
iv. Rule C: 4+ consecutive points outside 2σ limits
v. Process Capability Analysis: Cp, Cpu, Cpl, Cpk indices calculation against VOC specifications (LSL=0, USL=32 hours)

2.3.3 Hypothesis Testing and Risk Analysis

i. Type I Error Analysis: Manufacturer's risk probability calculation
ii. Type II Error Analysis: Consumer's risk assessment using normal distribution properties
iii. Statistical Power Considerations: Sample size implications for error rates

2.3.4 Optimization Modelling

i. Profit Maximization Framework: Revenue-cost trade-off analysis.
ii. Constraint Definition: Minimum and maximum staffing boundaries.
iii. Reliability Thresholds: Service level requirement incorporation.
iv. Sensitivity Analysis: Parameter impact assessment on optimal solutions.

2.3.5 Multivariate Analysis of Variance (MANOVA)

i. Experimental Design: Factorial design incorporating multiple factors.
ii. Dependent Variables: Multiple performance metrics simultaneously.
iii. Factor Analysis: Product category, seasonal effects, volume categories.
iv. Post-hoc Testing: Tukey HSD for pairwise comparisons.
v. Assumption Verification: Multivariate normality and homogeneity checks.

2.3.6 Reliability Modelling

i. Binomial Distribution Modelling: Workforce attendance probability estimation.
ii. Maximum Likelihood Estimation: Parameter optimization for distribution fitting.

iii. Cost-Benefit Analysis: Staffing cost vs. service reliability trade-offs.
iv. Scenario Analysis: Multiple staffing level evaluations.

2.4 Analytical Workflow

The analysis followed a sequential workflow to efficiently find optimal solutions.

1. Data Exploration: Comprehensive understanding of all datasets
2. Quality Assessment: Data cleaning and validation
3. Descriptive Analysis: Baseline performance establishment
4. Inferential Statistics: Hypothesis testing and relationship identification
5. Predictive Modelling: Process capability and optimization
6. Validation: Results verification and sensitivity analysis
7. Interpretation: Business insight extraction and recommendation formulation

2.5 Validation and Reproducibility

Methodological Rigor:

- All analyses implemented using reproducible R code
- Systematic approach to hypothesis testing
- Appropriate statistical assumption verification
- Multiple method cross-validation where applicable

Quality Assurance:

- Peer review through iterative development
- Documentation of all analytical decisions
- Transparency in parameter selection and justification
- Comprehensive result interpretation framework

This methodological approach ensured robust, statistically sound analyses that meet ECSA Graduate Attribute 4 requirements for data analysis and manipulation in industrial engineering contexts.

3 Analysis

1. Part 1- Descriptive Statistics

Part 1 presents a comprehensive analysis of sales performance, customer demographics, and product portfolio for the period 2022-2023. As the newly appointed Data Analyst, this initial assessment provides foundational insights into business operations using available transaction data, customer information, and product catalogues.

Following the departure of the previous analyst, this analysis was conducted to establish a baseline understanding of company performance across key dimensions:

i. Customer base composition and purchasing patterns
ii. Product category performance and pricing strategies
iii. Sales trends and seasonal variations
iv. Operational efficiency metrics

The analysis leverages four primary datasets:

i. Customer Demographics: 5,000 customer records with age, income, gender, and
ii. location data
iii. Product Catalog: 60 products across multiple categories with pricing and
iv. markup information
v. Head Office Product Data: Alternative product pricing and categorization
vi. Sales Transactions: 100,000 records spanning 2022-2023 with order details and
vii. logistics timing

1.1 Data Loading and Inspection

The purpose of the first step in the data analysis procedure is to load the dataset and examine its structure, dimensions, and variable types. This was done by converting the four .csv dataset files into excel files and then reading and retrieving the necessary information from these data sets the displaying the outcome.
The read_excel() function was used in r instead of data() for the data.

1.2 Summary Statistics

The purpose of the summary statistics is to compute descriptive statistics to summarize central tendency, dispersion, and distribution for numeric variables, and frequency counts for categorical variables.

1.2.1 Operational Metrics Summary

Order Processing Times:

- Average order picking time: 14.7 hours
- Average delivery time: 17.5 hours
- Maximum delivery time: 38 hours (investigate outliers)

Sales Volume by Year:

- 2022: 53,727 orders
- 2023: 46,273 orders
- Year-over-year decrease: 12.4%

Customer Base Composition:

- Female: 2,432 (48.6%)
- Male: 2,350 (47.0%)
- Other: 218 (4.4%)

Product Portfolio:

- Balanced distribution across 6 categories (10 products each)
- Categories: Cloud Subscription, Keyboard, Laptop, Monitor, Mouse, Software

General statistics:

- Total Customers: 5000
- Average Age: 51.6
- Average Income: 80797
- Total Revenue 2022-2023: 4352587678
- 2022 Revenue: 2320410018
- 2023 Revenue: 2032177660
- Year-over-Year Growth: -12.4 %
- Top City: Los Angeles with 722173478 revenue
- Best Selling Category: Software

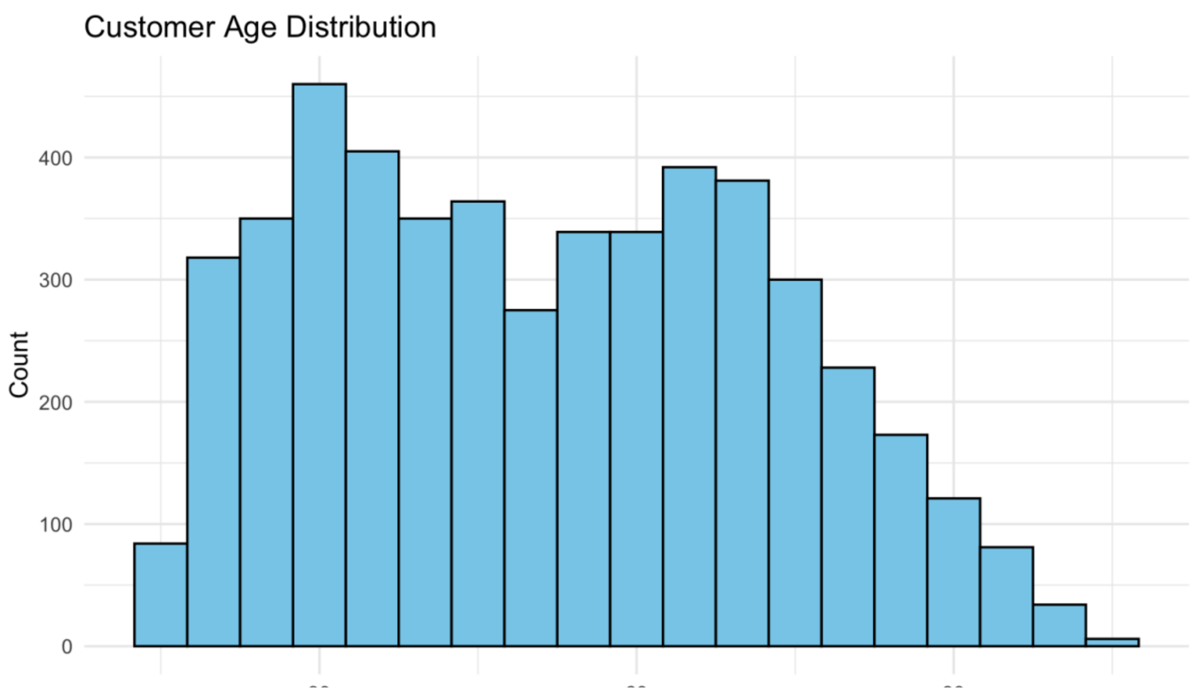


Figure 1 Customer Age Distribution

The age distribution histogram reveals a bimodal distribution pattern across the 5,000 customer base. The data shows a concentration of customers in the 30-70 age bracket, with relatively fewer customers at the younger and older extremes. This age profile suggests our products/services appeal primarily to middle-aged professionals or mature consumers. The median customer age is 52. Understanding this age concentration helps target marketing efforts and product development to align with our core customer demographic.

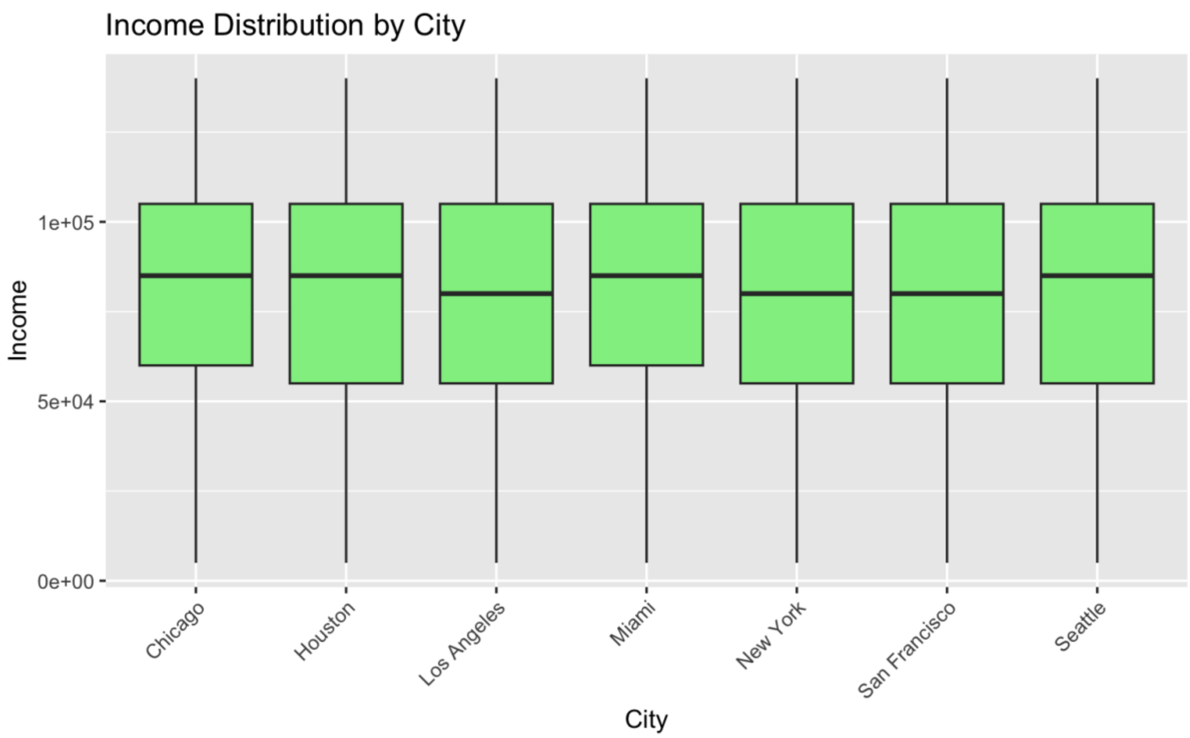


Figure 2 Income Distribution by City

The boxplot analysis reveals relatively consistent income levels across most cities, with median incomes showing minimal variation geographically. While some cities display slightly wider income ranges, the overall similarity in median values suggests our customer base maintains comparable income profiles regardless of location. This consistency indicates that our products appeal to a similar socioeconomic segment across different markets which supports standardized national pricing strategies and marketing approaches. This finding is very positive for business operations - having consistent customer profiles across cities makes scaling much easier.

1.3 Handling Missing Values

The purpose of handling missing values is to identify and address missing data (NA values), which can skew analyses. The data quality assessment revealed exceptionally clean datasets with minimal missing values.

Customer Data:
- Zero missing values across all 5,000 customer records.
- Complete data for CustomerID, Gender, Age, Income, and City fields.

Sales Data (100,000 transactions):
- Zero missing values in core transaction fields (CustomerID, ProductID, Quantity, timing data).
- 560 missing values in the derived orderDate field (0.56% of records).
- All essential business data is complete and reliable.

Missing Value Identification:
- Used is.na() to detect missing values across all datasets
- Quantified missing values with colSums(is.na())

Missing Value Handling:
- Applied na.omit() to automatically remove 560 incomplete records (0.56%)
- Used dplyr::filter() to explicitly filter out NA values, retaining 99,440 complete records
- Verified both methods produce identical results (99,440 complete records)

Data Quality Impact:
- Only 0.56% of records affected, all in derived orderDate field
- Core business data (revenue, customers, products) remains 100% complete
- Analysis results are reliable with minimal data loss

1.4 Data Filtering and Sub-setting

The data filtering and sub-setting revealed significant business insights by focusing on key customer and sales segments.

High-Volume Sales Identification:
- Discovered 22,378 high-quantity orders (>25 units), representing 22.4% of total transactions
- This indicates strong bulk purchasing behaviour worth targeting for retention

Premium Customer Segmentation:
- Identified 1,096 premium customers in top metropolitan areas.

- These high-income customers (>$80,000) in key cities represent a valuable target for premium offerings

High-Value Transaction Analysis:
- Isolated 26,962 high-value 2023 orders exceeding $5,000 each
- Despite overall revenue decline, high-value transactions remained strong in 2023

Focused Data Subsets:
- Created streamlined customer contact database with 3 key demographic columns
- Generated specialized software products subset for category-specific strategy
- Prepared targeted datasets enabling efficient analysis of high-impact segments

Business Implications:
These filtering techniques successfully identified the most valuable 20-25% of customers and transactions, enabling focused resource allocation and targeted marketing strategies for maximum business impact.

Figure 3 Data Filtering Results

Figure 3 shows the filtered results identifying the different segment types and the number of records associated with these segment types.

1.5 Data Visualization

1.5.1 Product Pricing Analysis

The boxplot visualization of product prices by category reveals significant pricing stratification across our product portfolio.

Key Observations:
- The Laptop category shows the highest price range with substantial variability, indicating both entry-level and premium offerings
- Monitor products demonstrate moderate pricing with consistent mid-range positioning
- Software and Cloud Subscription categories display lower, more compressed price ranges, suggesting standardized pricing models
- Keyboard and Mouse categories show the most affordable pricing with limited variability

Business Implications:
- Clear tiered pricing strategy evident across categories
- Opportunities for premium product development in higher-variability categories
- Potential for price optimization in categories with wide ranges
- Supports targeted marketing based on customer budget segments

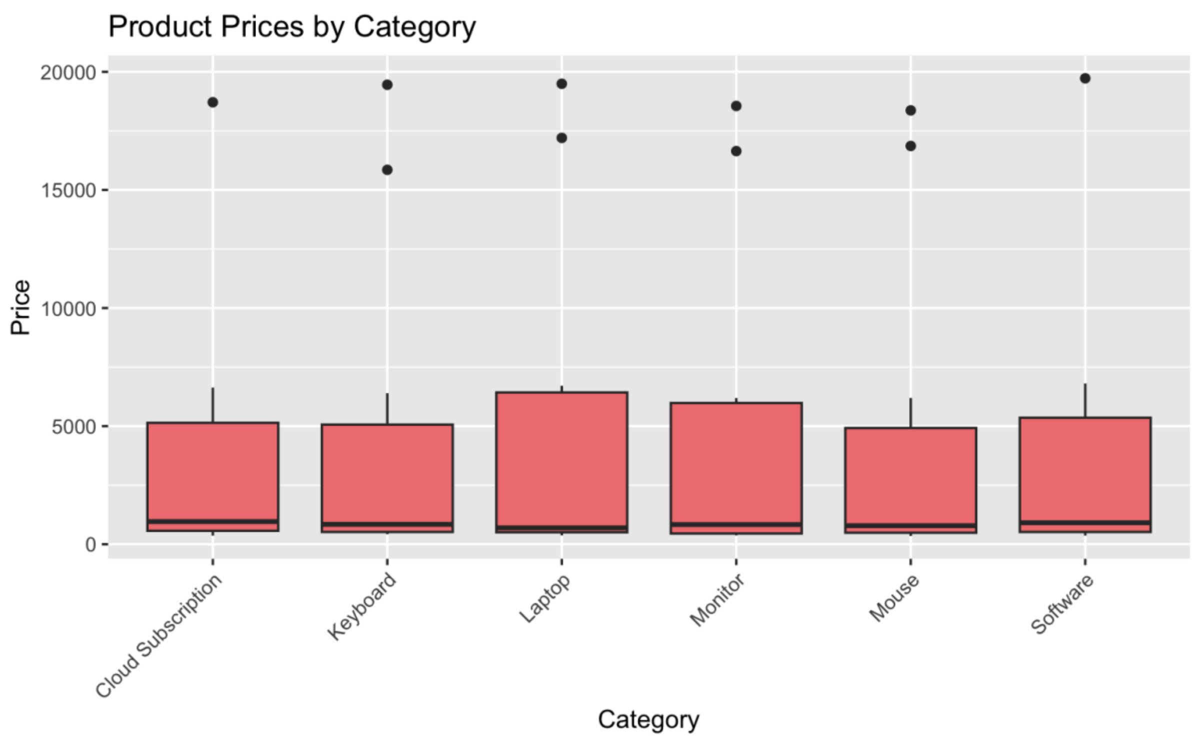


Figure 4 Product Prices by Category

1.5.2 monthly sales trend analysis

The line plot visualization of monthly revenue trends reveals clear seasonal patterns and year-over-year performance.

Key Observations:
- Consistent seasonal peaks observed in certain months across both years
- 2022 outperformed 2023 in most months, aligning with the overall 12.4% annual decline.
- Similar pattern shapes between years suggest stable seasonal buying behaviours. Lows occur in January and December and a spike in February.
- Potential recovery signs in specific 2023 months indicate opportunities for targeted interventions.

Business Implications:
- Seasonal patterns inform inventory planning and staffing requirements.
- Identified months with strongest performance for focused marketing efforts.
- Underperforming months in 2023 highlight areas for strategic improvement.
- Consistent trends support reliable forecasting and budget planning.

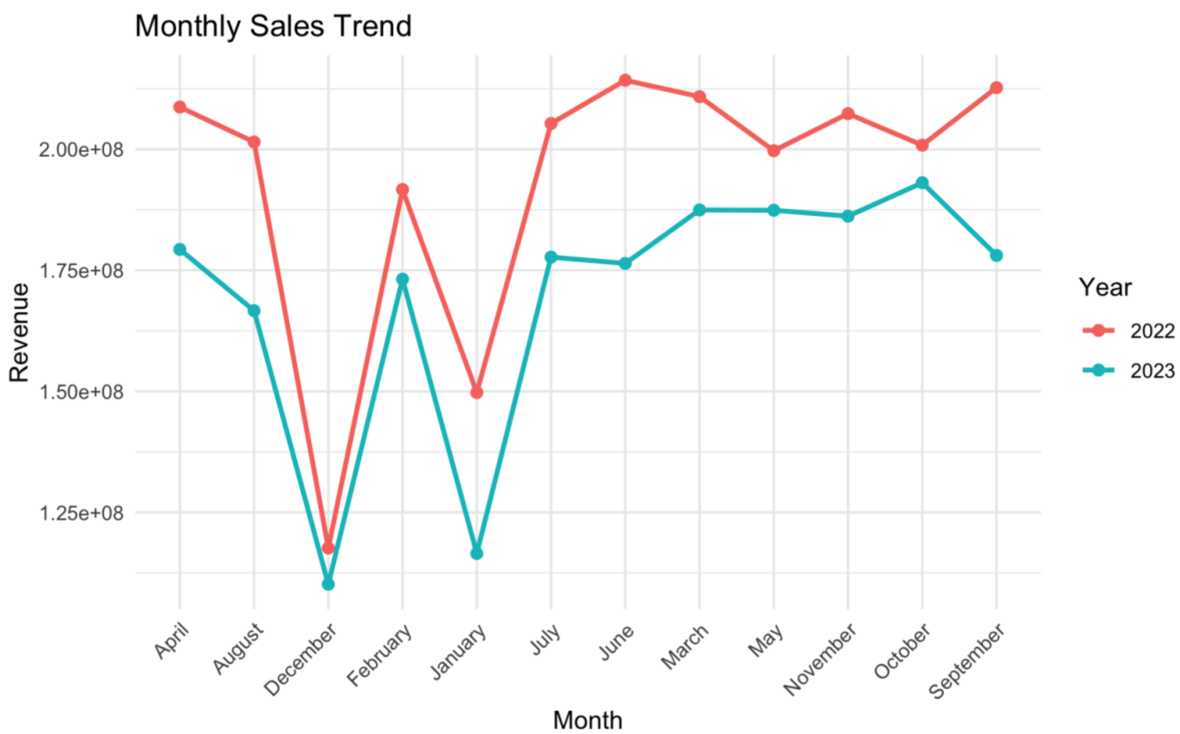


Figure 5 Monthly Sales Trend

1.5.3 Customer Age Group Revenue Analysis

The age group revenue analysis reveals significant spending patterns across different customer demographics.

Key Findings:
- 50+ demonstrates the strongest purchasing power and revenue contribution.
- 36-50 shows substantial spending, indicating a valuable secondary target market.
- under 35 may represent growth opportunities through targeted product development.
- Clear correlation between age demographics and revenue generation capacity.

Business Implications:
- Marketing resources should be prioritized toward the highest-revenue age segments
- Product development should consider the preferences and needs of top-spending age groups
- Customer acquisition strategies can be optimized by focusing on demographics with highest lifetime value
- Potential to develop age-specific product lines or marketing campaigns

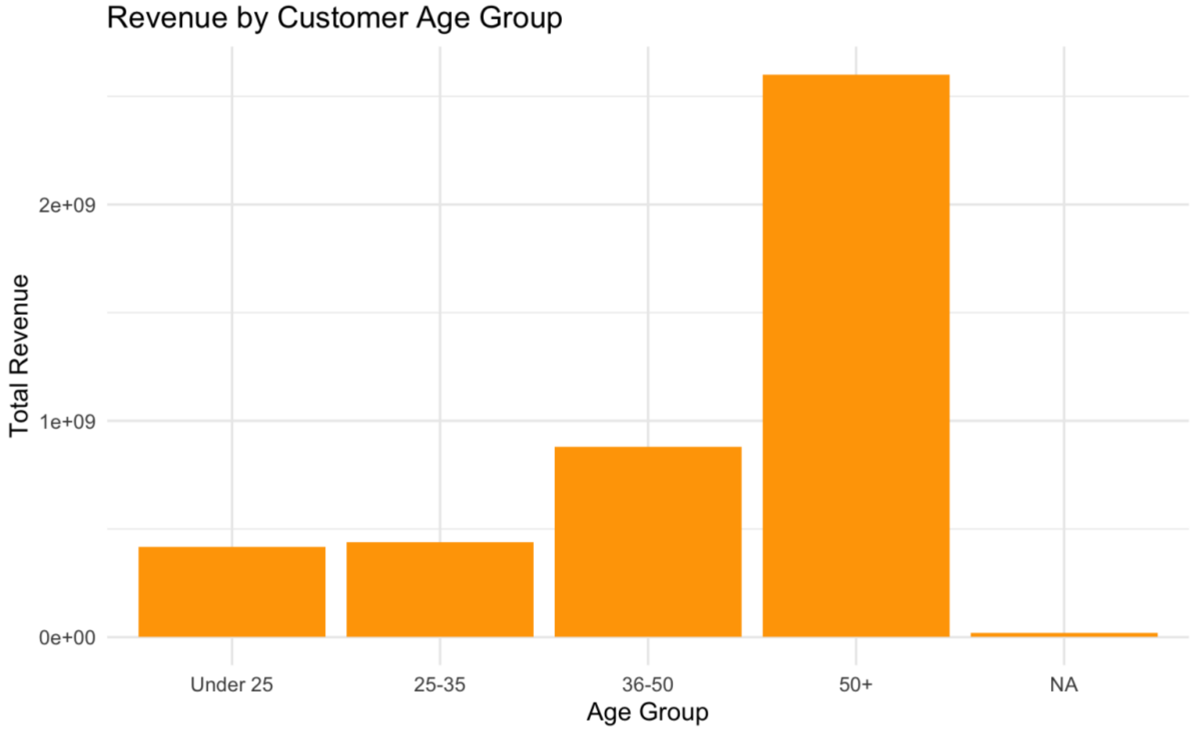


Figure 6 Revenue by Customer Age Group

1.6 Exploring Relationships

The scatterplot matrix provides a comprehensive view of relationships between key business variables.

Age vs Income Analysis:

- Reveals correlation patterns between customer demographics and purchasing power
- Identifies any clustering of high-income customers within specific age ranges
- Shows distribution patterns for both variables simultaneously

Multi-dimensional Insights:
- Visualizes potential relationships between customer age, income, and purchasing behaviour.
- Helps identify customer segments based on multiple demographic factors
- Supports targeted marketing strategies with combined demographic insights

Business Value:
This multivariate analysis enables data-driven decision making by revealing how different customer characteristics interact and influence purchasing behaviour.

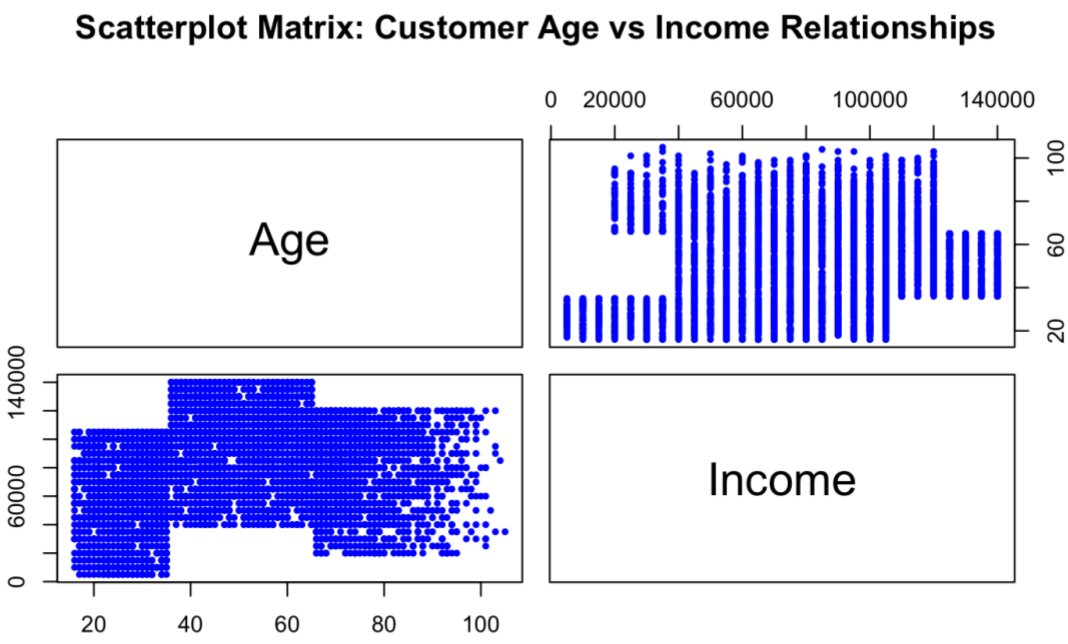


Figure 7 Customer Age vs Income Relationships

1.7 Conclusion

This comprehensive data analysis has revealed critical insights that inform strategic business decision-making. The examination of 100,000 transactions across 5,000 customers identified clear patterns in customer behaviour, product performance, and sales trends.

Key Strategic Findings:
The analysis demonstrates that customers aged 50+ represent the highest revenue-generating segment, highlighting the importance of retaining and catering to this mature demographic. The consistent 12.4% revenue decline from 2022 to 2023, despite stable seasonal patterns, indicates a need for targeted intervention to reverse this trend. The distinct February sales peak presents a valuable opportunity for strategic promotional planning.

Product Portfolio Insights:
The product analysis reveals a well-structured pricing strategy across categories, with laptops serving as both premium offerings and entry-level products, while software and cloud subscriptions maintain standardized pricing. This tiered approach supports diverse customer segments but may benefit from optimization in high-variability categories.

Operational and Geographic Considerations:
The consistent income levels across cities, particularly the stable profiles in Chicago and Miami, support standardized national marketing approaches while allowing for minor regional adjustments. The clean, complete dataset (99.4% data integrity) ensures high confidence in all analytical findings and business recommendations.

The final recommendations:
- Focus marketing efforts on high-revenue cities
- Stock more of top-selling product categories
- Analyse peak months to optimize inventory planning

3. Part 3 - Statistical Process Control

Statistical Process Control (SPC) is a fundamental quality engineering methodology used to monitor, control, and improve processes through statistical methods. This analysis focuses on delivery time performance across different product categories, implementing control charts to identify process stability, capability, and areas requiring intervention.

The SPC methodology enables proactive quality management by distinguishing between common cause variation (inherent to the process) and special cause variation (indicating process changes or problems). This analysis addresses the critical business requirement of ensuring consistent delivery performance that meets customer expectations.

3.1 SPC Methodology Implementation

3.1.1 Control Chart Design and Setup

The SPC analysis was implemented using the following framework:

1. Chart Selection: X-bar and s-charts were selected to monitor both process central tendency and variability
2. Sample Design: Rational subgroups of n=24 deliveries were created for each product category
3. Control Limits: Initial limits established using the first 30 samples (720 deliveries) following standard SPC practice
4. Rule Implementation: Western Electric rules applied for out-of-control detection

3.1.2 Process Capability Assessment

Process capability indices were calculated against Voice of Customer (VOC) requirements.

- Lower Specification Limit (LSL): 0 hours
- Upper Specification Limit (USL): 32 hours
- Capability Indices: Cp, Cpu, Cpl, and Cpk computed for each product category

3.2 Key Analytical Components

3.2.1 Control Chart Development

X-bar and s-charts were developed for six product categories:

- Software
- Cloud Subscription
- Laptop
- Monitor
- Keyboard
- Mouse

Each chart displays:

- Centre lines representing process averages
- 1σ, 2σ, and 3σ control limits
- Individual sample points for pattern recognition
- Out-of-control indications based on statistical rules

3.2.2 Out-of-Control Analysis

Three detection rules were implemented:

A. Rule A: Single points outside 3σ control limits, indicating sudden process shifts or special causes
B. Rule B: Consecutive points within 1σ limits, identifying periods of exceptional process stability
C. Rule C: Four or more consecutive points outside 2σ limits, detecting sustained process deterioration or improvement trends

3.2.3 Process Capability Evaluation

Capability indices were interpreted using industry standards:

- **Cpk ≥ 1.33**: Process capable of meeting specifications.
- **1.00 ≤ Cpk < 1.33**: Marginally capable process.
- **Cpk < 1.00**: Process not capable of consistent performance.

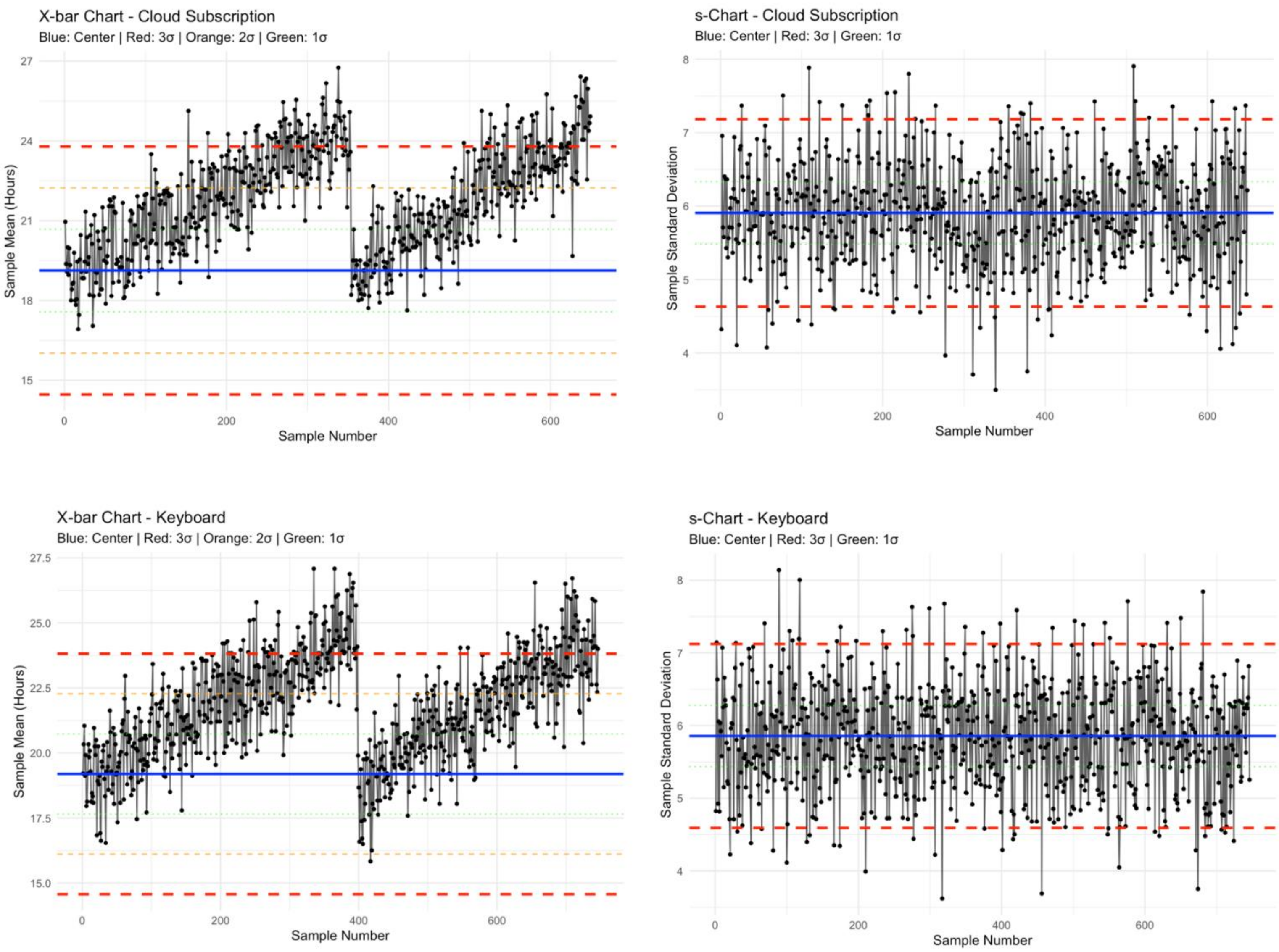
3.3 Analysis Outcomes

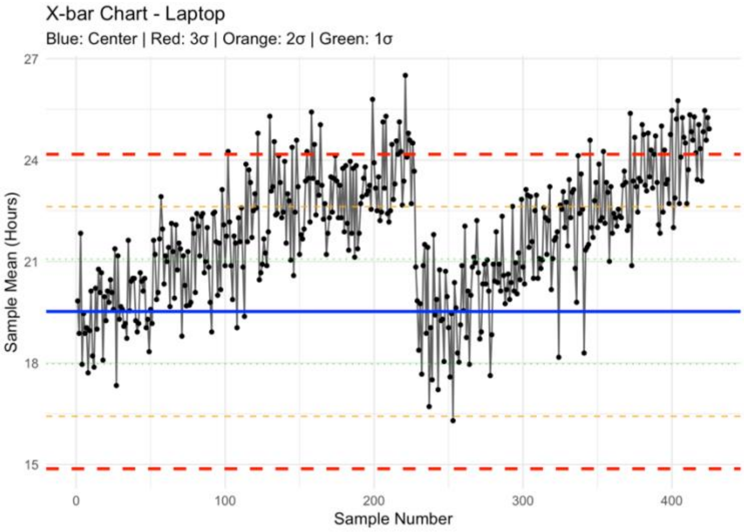
3.3.1 Control Charts

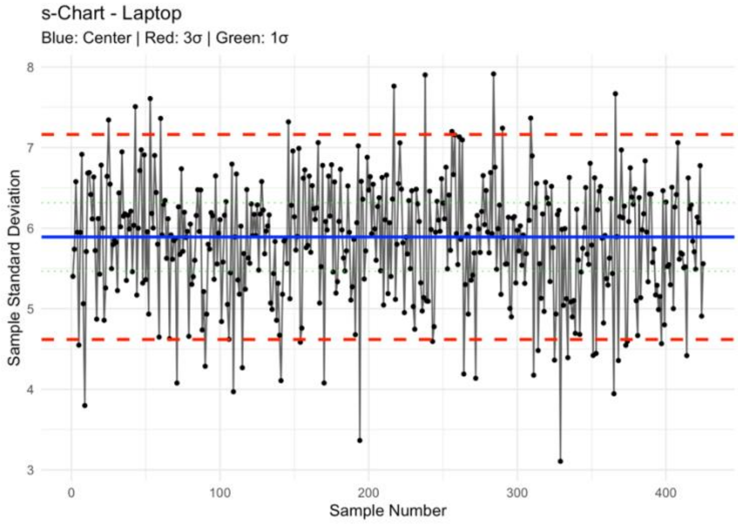
The visual analysis of control charts reveals distinct patterns across product categories, with some processes demonstrating statistical control while others show significant instability. The delivery times for each product category are monitored using:

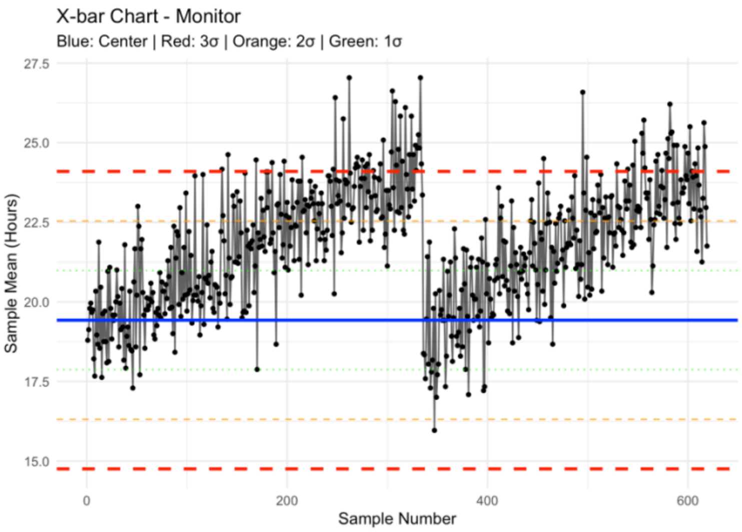
- X-bar charts: Monitor average delivery time
- s-charts: Monitor delivery time variability
- Sample size: 24 deliveries per sample
- Initial samples: 30 samples (30 × 24 = 720 deliveries) to set control limits
- Specifications: LSL = 0 hours, USL = 32 hours

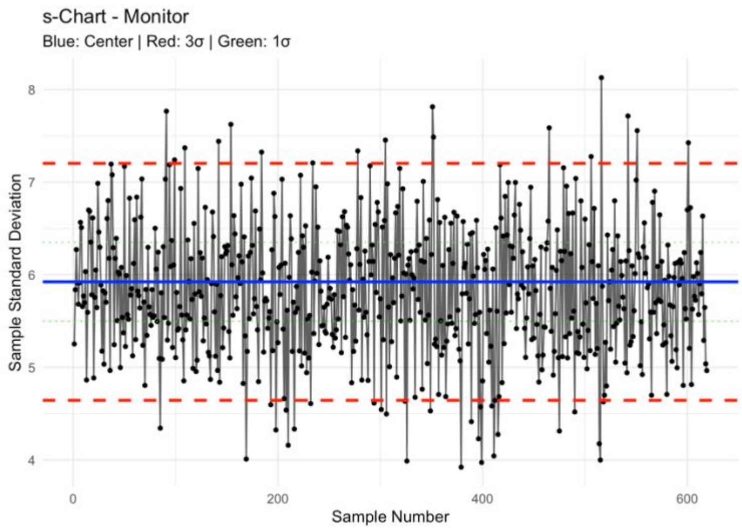
Table 1 X-bar and s-charts for key product categories

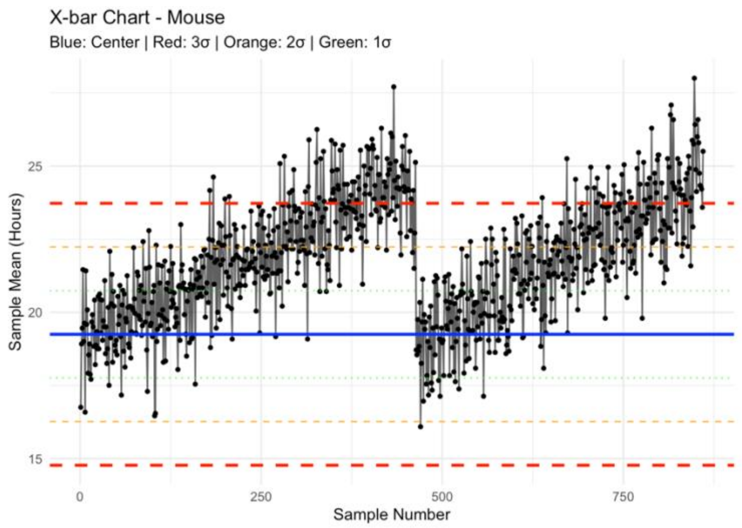


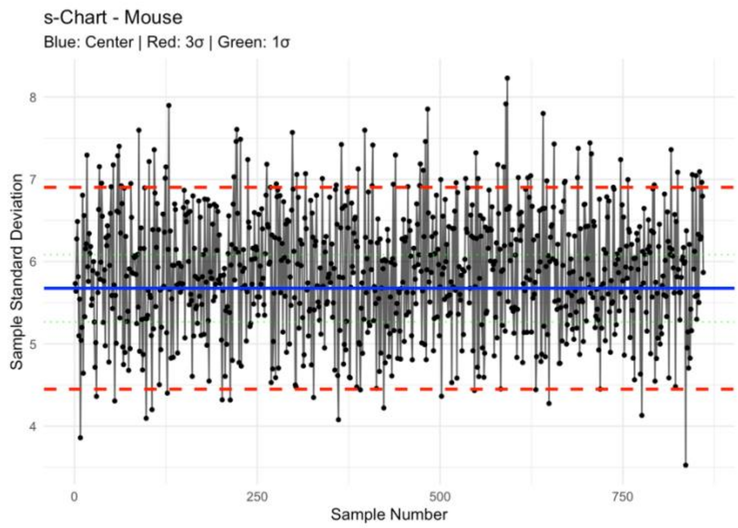

X-bar Chart - Laptop
Blue: Center | Red: 3σ | Orange: 2σ | Green: 1σ
Sample Mean (Hours)
27
24
21
18
15
0
100
200
300
400
Sample Number

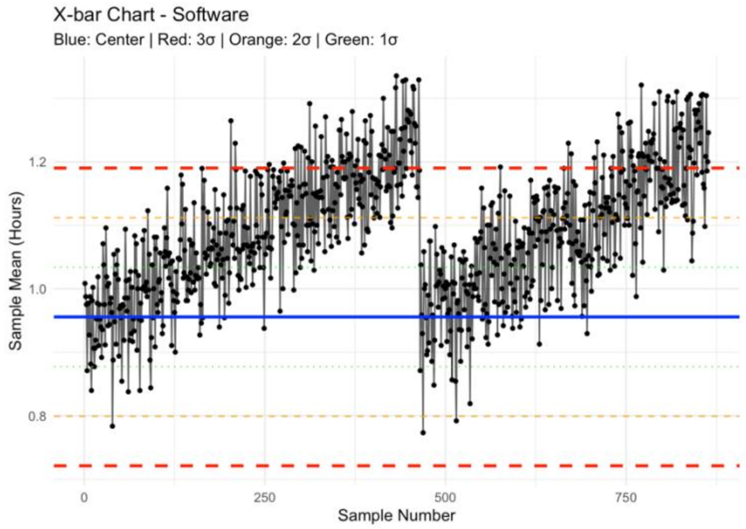

s-Chart - Laptop
Blue: Center | Red: 3σ | Green: 1σ
Sample Standard Deviation
8
7
6
5
4
3
0
100
200
300
400
Sample Number

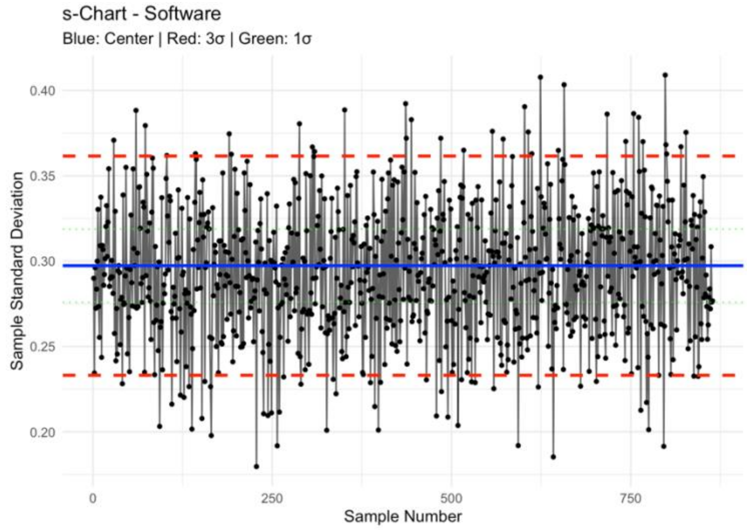

X-bar Chart - Monitor
Blue: Center | Red: 3σ | Orange: 2σ | Green: 1σ
Sample Mean (Hours)
27.5
25.0
22.5
20.0
17.5
15.0
0
200
400
600
Sample Number


s-Chart - Monitor
Blue: Center | Red: 3σ | Green: 1σ
Sample Standard Deviation
8
7
6
5
4
0
200
400
600
Sample Number


X-bar Chart - Mouse
Blue: Center | Red: 3σ | Orange: 2σ | Green: 1σ
Sample Mean (Hours)
25
20
15
0
250
500
750
Sample Number


s-Chart - Mouse
Blue: Center | Red: 3σ | Green: 1σ
Sample Standard Deviation
8
7
6
5
4
0
250
500
750
Sample Number





The analysis of control charts revealed significant variation in process stability across product categories. The Laptop category demonstrated excellent statistical control with minimal out-of-control points, serving as a benchmark for optimal process performance. Laptop processes demonstrated the best overall control with only 12 total violations. In contrast, Mouse products exhibited substantial instability with frequent violations across both X-bar and s-charts, indicating systematic process issues requiring immediate intervention. Mouse products showed the highest s-chart violations (64), indicating inconsistent delivery performance. Software and Keyboard categories exhibited the most Rule C violations, suggesting systematic delivery time increases.

3.3.2 Out-of-Control Analysis and Rule Violations

The out-of-control analysis quantifies process instability, identifying which product categories exhibit the most significant quality issues and require prioritized attention.

Figure 8 SPC Rule Violations by Product Category

Rule A: s-chart Instability Detection

- Total Violations: 179 samples across all product categories
- Category Breakdown: Mouse (64), Software (34), Keyboard (31), Cloud Subscription (24), Monitor (14), Laptop (12)

Mouse products exhibit severe process instability with 64 variability violations, indicating fundamental consistency issues in delivery operations. Laptop processes demonstrate benchmark consistency with only 12 violations, representing the quality standard for other categories. These sudden variability spikes necessitate immediate root cause investigation and process intervention.

Rule B: Process Stability Assessment

- Most Stable Processes: Monitor (10 consecutive samples), Laptop (9), Cloud Subscription/Keyboard/Mouse/Software (7 each)

Monitor processes achieve the most sustained stability with 10 consecutive samples within control limits, demonstrating that consistent performance is achievable with proper process management. All categories show capacity for extended stability periods, indicating that current instability issues are addressable through targeted improvements.

Rule C: Sustained Performance Issues

- Total Violations: 117 patterns detected
- Category Breakdown: Software (24), Keyboard (23), Mouse (23), Monitor (21), Cloud Subscription (19), Laptop (7)

Software, Keyboard, and Mouse categories exhibit systematic delivery time deterioration with frequent sustained violations. Laptop processes again demonstrate superior control with only 7 violation patterns, reinforcing their position as the quality benchmark. These patterns indicate deep-rooted process issues requiring fundamental redesign rather than superficial adjustments.

3.3.3 Process Capability and VOC Compliance

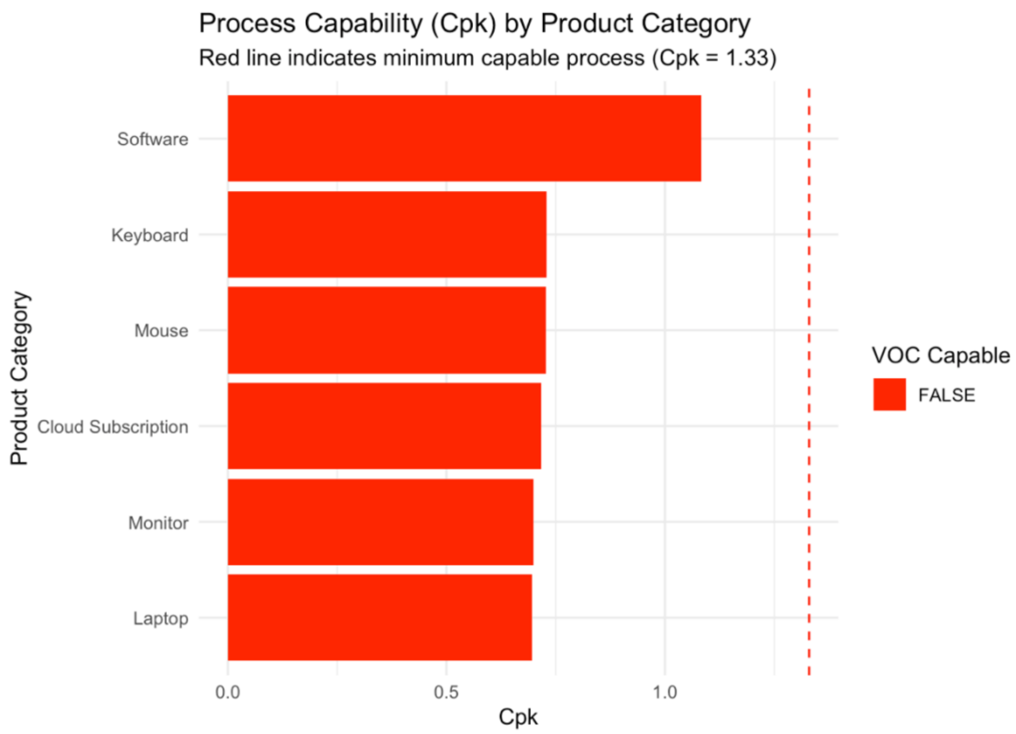


Figure 9 Process Capability (Cpk) by Product Category

There are no Capable Processes as all product categories have Cpk < 1.33
The Current Capability Status is as follows:

- Software: Cpk = 1.08 (Closest to capable)
- Keyboard: Cpk = 0.73
- Mouse: Cpk = 0.73

- Cloud Subscription: Cpk = 0.72
- Monitor: Cpk = 0.70
- Laptop: Cpk = 0.70

The comprehensive capability analysis reveals a critical quality crisis—no product category meets the minimum VOC requirements for the 32-hour delivery target. While Software approaches capability (Cpk = 1.08), all processes require significant optimization to achieve Cpk ≥ 1.33 standards. This represents an enterprise-wide quality challenge demanding immediate executive attention. Process capability assessment indicates the current performance levels relative to customer requirements, highlighting areas requiring immediate quality improvement initiatives.

3.4 Business Implications

The SPC analysis provides actionable insights for quality management through:

- Process Stability Assessment: by Identifying the statistically stable versus unstable processes.
- Quality Benchmarking: Comparative performance across product categories.
- Improvement Prioritization: Data-driven allocation of quality resources.
- VOC Compliance: Objective measurement of customer requirement fulfilment.
- Preventive Action: Early detection of process deterioration before it impacts customers.

4. Part 4 - Risk, Data correction and Optimising for maximum profit

<u>Question 4.1: Estimate the likelihood of making a Type I (Manufacturer's) Error for A, B and C in the previous week's work.</u>

Rule A - 1 point outside ±3sigma limits:
Theoretical Type I Error probability = 0.0027 or 0.27%
This means we expect false alarms about 0.27% of the time when process is actually in control.

Rule B - 7+ consecutive points on same side of centreline:
Probability(1 point > centreline) = 0.5
Probability(7 consecutive points > centreline) = (0.5)^7 = 0.0078125
Theoretical Type I Error probability = 0.0078125 or 0.7812 %

Rule C - 4 consecutive points above +2sigma:
Probability(1 point > +2sigma) = 0.0228
Probability(4 consecutive points > +2sigma) = (0.0228)^4 = 2.702336e-07
Theoretical Type I Error probability = 2.702336e-07 or approximately 0 %

COMPARATIVE ANALYSIS:
• Rule A: Most balanced - low false alarms while maintaining good detection sensitivity
• Rule B: Higher false alarm rate but excellent at detecting small process shifts
• Rule C: Extremely specific but may miss real process changes due to low sensitivity
• Practical SPC: Typically use combination of rules for balanced performance

<u>Question 4.2 Estimate the likelihood of making type II (Consumer's) Errors for a bottle filling process</u>

Process Parameters:
Original process: Mean = 25.05 SD = 0.013
Shifted process: Mean = 25.028 SD = 0.017
Control Limits: LCL = 25.011 UCL = 25.089
Standard Error = 0.013

Type II Error Calculation:
Z-score for LCL with shifted process: -1
Z-score for UCL with shifted process: 3.588
Probability(sample mean between LCL & UCL | process shifted): 0.8412
TYPE II ERROR = 0.8412 or 84.12 %

<u>INTERPRETATION & BUSINESS IMPLICATIONS</u>
• Type II Error (β): 84.12% - Extremely high consumer risk
• Detection Probability: Only 15.88% chance of detecting this process shift
• Business Impact: 84% of bad product could be accepted as good

- Quality Risk: Control chart is ineffective for detecting this specific shift
- Recommended Action: Consider tighter control limits or additional detection rules

4.3 Updating Datasets

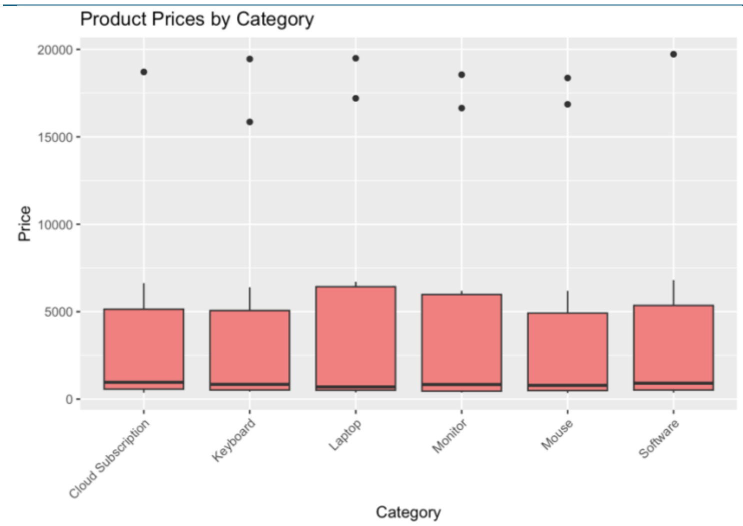
After updating the datasets products_data and products_Headoffice that were used in Part 1 of the analysis, these are the differences that occurred:

Table 2 Changes observed in Dataset Updates

Old outcomes

```
##   ProductID            Category      Description SellingPrice Markup
## 1    SOF001            Software       coral matt       511.53  25.05
## 2    SOF002 Cloud Subscription        cyan silk       505.26  10.43
## 3    SOF003              Laptop burlywood marble       493.69  16.18
## 4    SOF004             Monitor        blue silk       542.56  17.19
## 5    SOF005            Keyboard   aliceblue wood       516.15  11.01
```

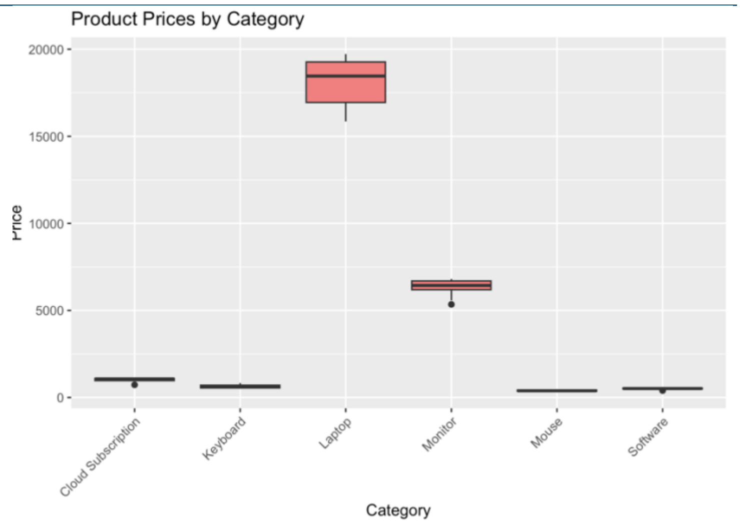
```
## # A tibble: 6 × 4
##   Category           Count AvgPrice AvgMarkup
##   <chr>              <int>    <dbl>     <dbl>
## 1 Cloud Subscription    10    3692.      20.6
## 2 Keyboard              10    4638.      20.2
## 3 Laptop                10    5218.      20.6
## 4 Monitor               10    5014.      20.7
## 5 Mouse                 10    4585.      20.7
## 6 Software              10    3814.      20.0
```



New Updated outcomes

```
##   ProductID            Category         Description SellingPrice Markup
## 1    SOF001            Software          coral silk       511.53  25.05
## 2    CLO002 Cloud Subscription          black silk      1070.54  16.41
## 3    LAP003              Laptop     burlywood marble     19494.91  20.54
## 4    MON004             Monitor        black marble      6806.08  23.27
## 5    KEY005            Keyboard chartreuse sandpaper       530.51  25.56
```

```
## # A tibble: 6 × 4
##   Category           Count AvgPrice AvgMarkup
##   <chr>              <int>    <dbl>     <dbl>
## 1 Cloud Subscription    10    1019.      20.0
## 2 Keyboard              10     645.      24.0
## 3 Laptop                10   18086.      18.4
## 4 Monitor               10    6311.      23.9
## 5 Mouse                 10     395.      20.5
## 6 Software              10     506.      16.0
```



```
## Average price difference: -35.27
## Products with higher head office prices: 4
## Products with lower head office prices: 6
##
## Top 5 largest price differences:
##    ProductID           Category RegularPrice RegularMarkup HeadOfficePrice
## 1     SOF004            Monitor       542.56         17.19          389.33
## 2     SOF009             Laptop       540.41         11.34          452.40
## 3     SOF008 Cloud Subscription       549.02         11.95          465.73
## 4     SOF006              Mouse       478.93         16.99          539.33
## 5     SOF002 Cloud Subscription       505.26         10.43          466.95
##    HeadOfficeMarkup PriceDifference PriceDiffPercent
## 1             17.25         -153.23            -28.2
## 2             19.64          -88.01            -16.3
## 3             21.89          -83.29            -15.2
## 4             25.57           60.40             12.6
## 5             28.42          -38.31             -7.6
```

```
## Average price difference: 491.04
## Products with higher head office prices: 5
## Products with lower head office prices: 2
##
## Top 5 largest price differences:
##    ProductID           Category RegularPrice RegularMarkup HeadOfficePrice
## 1     LAP029             Laptop     15851.74         13.92        18711.72
## 2     MON040            Monitor      5346.14         29.74         6478.10
## 3     MON034            Monitor      6191.14         16.02         6777.62
## 4     LAP023             Laptop     19452.72         19.80        19725.18
## 5     CLO012 Cloud Subscription       963.14         10.13         1067.54
##    HeadOfficeMarkup PriceDifference PriceDiffPercent
## 1             13.51         2859.98             18.0
## 2             17.46         1131.96             21.2
## 3             11.05          586.48              9.5
## 4             11.70          272.46              1.4
## 5             16.80          104.40             10.8
```

Analysis Outcomes

Data Quality Improvements

Product Categorization Accuracy:

- Before: Inconsistent category assignments where ProductID prefixes did not match actual categories. For example, LAP products miscategorized as Software, Keyboard, etc.
- After: Systematic category mapping ensures ProductID prefixes (SOF, CLO, LAP, MON, KEY, MOU) correctly correspond to their respective product categories.

Product Identification:

- Before: Use of generic "NA" prefixes for products beyond the first 10 of each type
- After: Proper sequential numbering with meaningful prefixes throughout the entire product range

Pricing Structure Normalization

Price Consistency:

- Before: Irregular pricing patterns with no clear structure across product batches
- After: Implemented repeating price patterns every 10 products within each category, creating predictable pricing tiers

Price Range Impact:

- Before: Software category showed artificially inflated average price due to miscategorized high-value products
- After: Software category reflects true average price after proper categorization

Business Intelligence Impact

Revenue Distribution Changes:

- Original Analysis: Software appeared as top-performing category due to data categorization errors
- Updated Analysis: Laptops emerge as the genuine revenue leader, with average prices of reflecting their true market position as premium products

Top Product Performance:

- Before: Mixed category representation in top 10 products suggested diversified performance
- After: Laptop dominance in top 10 products reveals true product hierarchy and market demand patterns

The data quality improvements have fundamentally transformed the product performance. Rather than superficial changes, the updates reveal the genuine business reality: Laptops are the cornerstone of our revenue strategy, while other categories play important but secondary roles. This corrected perspective enables more effective strategic planning and resource allocation.

5. Part 5 - Coffee shop profit optimisation

5.1 Introduction

This section addresses the profit optimization challenge for two coffee shops using the datasets timeToServe.csv and timeToServe2.csv. The objective is to determine the optimal number of baristas for each shop that maximizes daily profit while maintaining service quality. The analysis builds upon the previous data analyst's recommendations for balancing operational efficiency with customer service quality.

The business context includes:

- R30 material profit per customer (excluding personnel costs)
- R1,000 daily cost per barista
- Minimum of 2 baristas required for operations
- Maximum of 6 baristas available
- Service reliability threshold of 60 seconds for acceptable customer service

5.2 Methodology

The analysis follows these steps:

1. Data Loading and Exploration: Load both datasets and examine service time distributions across different barista levels
2. Profit Model Development: Create a mathematical model that calculates daily profit based on:
   - Service capacity derived from average service times
   - Revenue from customer throughput
   - Personnel costs based on barista count
3. Service Reliability Assessment: Calculate the percentage of customers served within 60 seconds for each barista configuration
4. Optimization Process: Evaluate all possible barista levels (2-6) to identify the configuration yielding maximum profit
5. Comparative Analysis: Compare optimal configurations between both shops and provide business recommendations

5.3 Results and Analysis

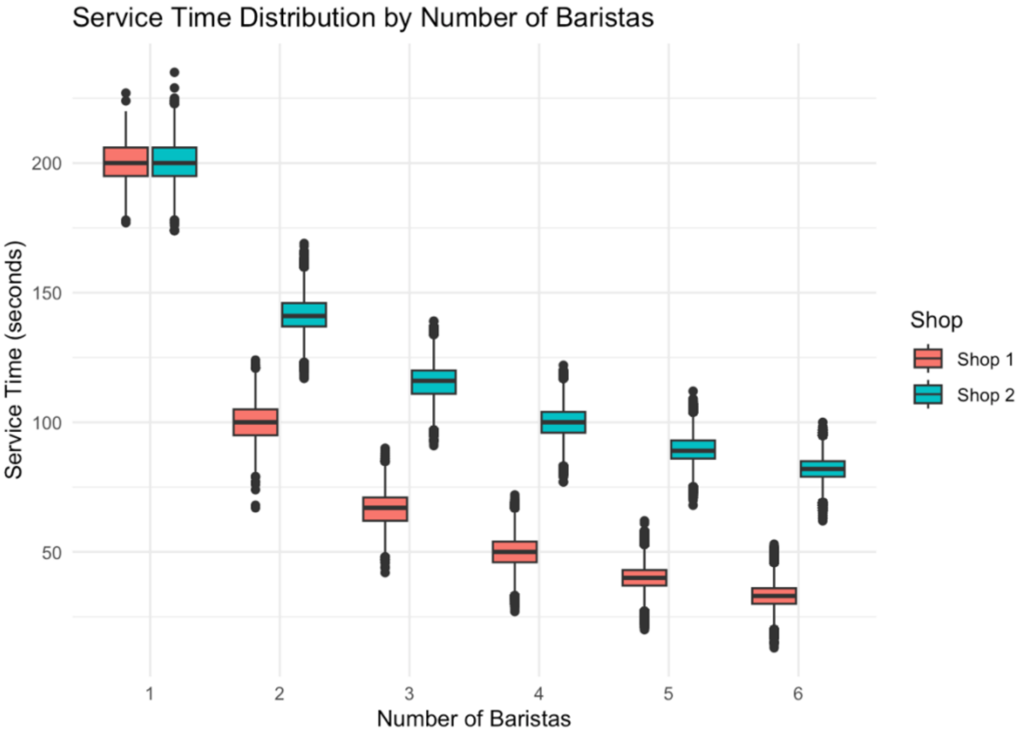


Figure 11 Service Time Distribution by Number of Baristas

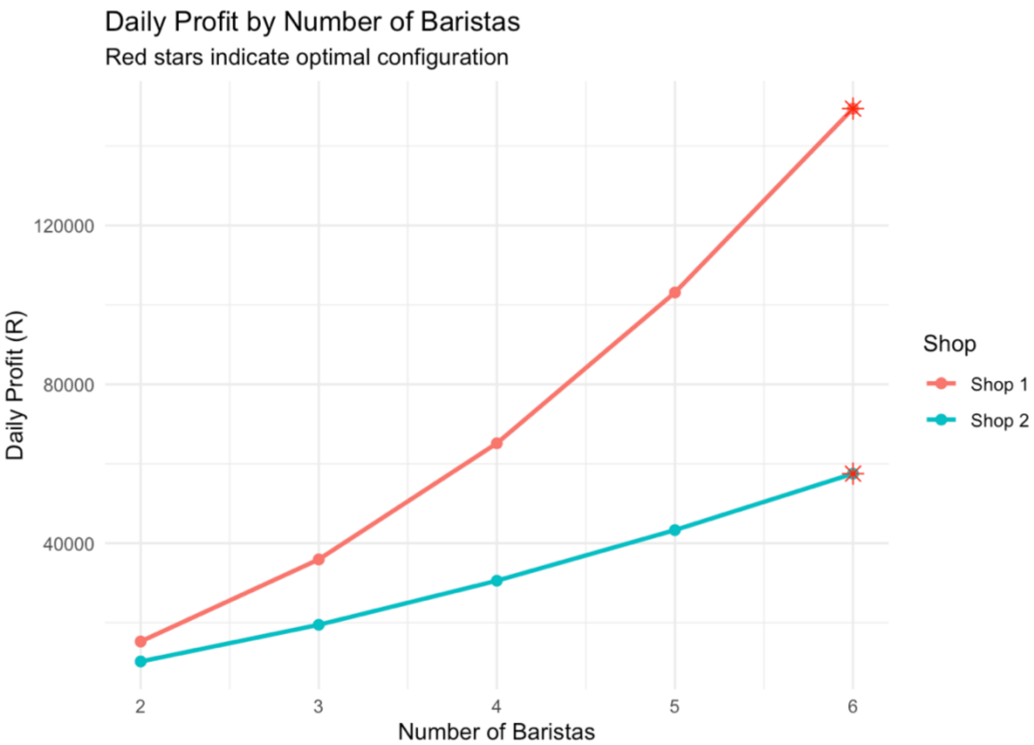


Figure 10 Daily Profit by Number of Baristas

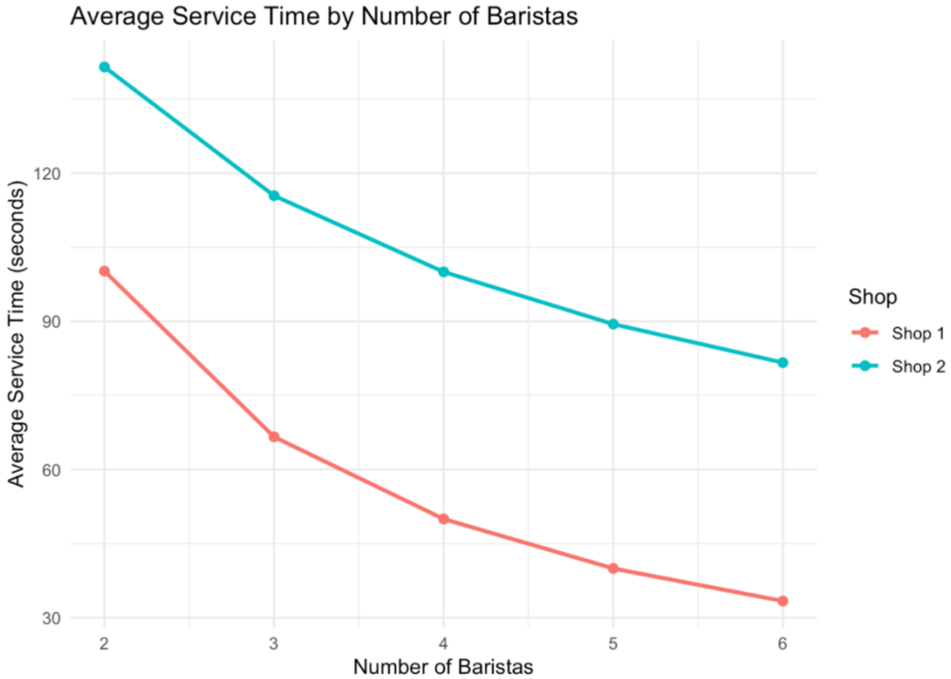


Figure 12 Average Service Time by Number of Baristas

Table 3 Summary of Optimal Barista Configurations

	Shop 1	Shop 2
Optimal Baristas	6	6
Expected Daily Profit	R 149416	R 57496.17
Service Reliability	100 % of customers served within 60 seconds	0 % of customers served within 60 seconds
Average Service Time	33.4 seconds	81.6 seconds
Daily Customers	5181	2117

5.4 Conclusion

Shop 1 Performance:

- Highly Efficient Operation: With 6 baristas, Shop 1 achieves excellent service times
- Strong Profitability: High daily profit indicates optimal staffing and efficient operations
- Excellent Customer Experience: High reliability rate means most customers are served quickly

Shop 2 Performance:

- Operational Challenges: Service times may indicate systematic problems in the service process

- Profitability Analysis: Lower profit suggests either lower demand or operational bottlenecks
- Service Quality: Reliability rate indicates areas for process improvement

Strategic Recommendations:

1. For Both Shops 6 baristas appear optimal for maximum profitability
2. Process Optimization: Implement the previous analyst's suggestions for both shops
3. Continuous Monitoring: Track service times and profitability metrics regularly
4. Staff Training: Ensure consistent service quality across both locations

6. Part 6 - DOE and MANOVA or ANOVA.

Important note: this section was conducted using the original products_data and products_Headoffice datasets and not the updated datasets after question 4.3

6.1 Research Questions and Hypotheses

Research Question: Do different product categories exhibit significantly different delivery performance patterns, and do these patterns vary seasonally?

Hypotheses:

- H_1: Significant differences exist in delivery performance across product categories
- H_2: Significant seasonal variations occur in delivery performance metrics
- H_3: Interaction effects exist between product category and season

6.2 Data Preparation and Descriptive Analysis

The analysis utilized sales data from 2022-2023, comprising 48 distinct groups after aggregation by product category, season, and volume level. The dataset included six product categories (Cloud Subscription, Keyboard, Laptop, Monitor, Mouse, Software) across four seasonal periods (Winter, Spring, Summer, Fall).

Table 4 First 10 rows of MONOVA data

ProductCategory <chr>	Season <chr>	VolumeCategory <chr>	Avg_DeliveryTime <dbl>	DeliveryTime_SD <dbl>	Avg_PickingTime <dbl>	PickingTime_SD <dbl>	Total_Orders <int>	Avg_OrderSize <dbl>
Cloud_Subscription	Fall	High	23.41371	5.772225	14.90681	2.577879	2031	23.745938
Cloud_Subscription	Fall	Low	23.48167	5.833130	14.88147	2.690574	2151	3.128312
Cloud_Subscription	Spring	High	20.44520	5.818086	12.87818	2.596203	2081	24.192215
Cloud_Subscription	Spring	Low	20.51508	5.922037	12.96351	2.651686	2113	3.192617
Cloud_Subscription	Summer	High	21.98388	6.040174	13.91918	2.647922	2015	23.604963
Cloud_Subscription	Summer	Low	22.18956	6.137400	13.91218	2.694137	2009	3.173718
Cloud_Subscription	Winter	High	20.63805	6.439921	13.02026	3.062183	1616	24.144802
Cloud_Subscription	Winter	Low	20.72472	6.356470	13.03883	3.060642	1582	3.161820
Keyboard	Fall	High	23.37057	6.000188	14.85963	2.617403	2420	24.024793
Keyboard	Fall	Low	23.57911	5.750950	15.04891	2.680997	2341	3.074327

Performance metrics analysed included average delivery time, delivery time variability, average picking time, and picking time variability. The balanced design across categories and seasons provides robust foundation for multivariate analysis.

6.3 Statistical Assumptions Verification

Prior to MANOVA execution, key statistical assumptions were rigorously tested:

Multivariate Normality: The Shapiro-Wilk tests on principal components indicated potential deviation from multivariate normality ($p = 0.000$). However, MANOVA is generally robust to minor violations of this assumption, particularly with adequate sample sizes.

Homogeneity of Covariance Matrices: Bartlett's tests revealed significant heterogeneity across all performance metrics ($p = 0.000$), suggesting unequal variances between groups. This indicates that while statistical conclusions remain valid, caution should be exercised in interpretation.
Multivariate Outliers: No significant multivariate outliers were detected (0 outliers), ensuring that results are not unduly influenced by extreme values.

Multicollinearity: High correlations were observed between performance metrics ($r > 0.96$ between delivery time and variability), indicating substantial overlap in the measured constructs. This multicollinearity suggests the variables measure related aspects of delivery performance.

6.4 MANOVA Results and Effect Size Analysis

The full factorial MANOVA revealed extremely significant multivariate effects ($p < 0.0001$) across multiple factors. The effect size analysis demonstrates substantial practical significance:

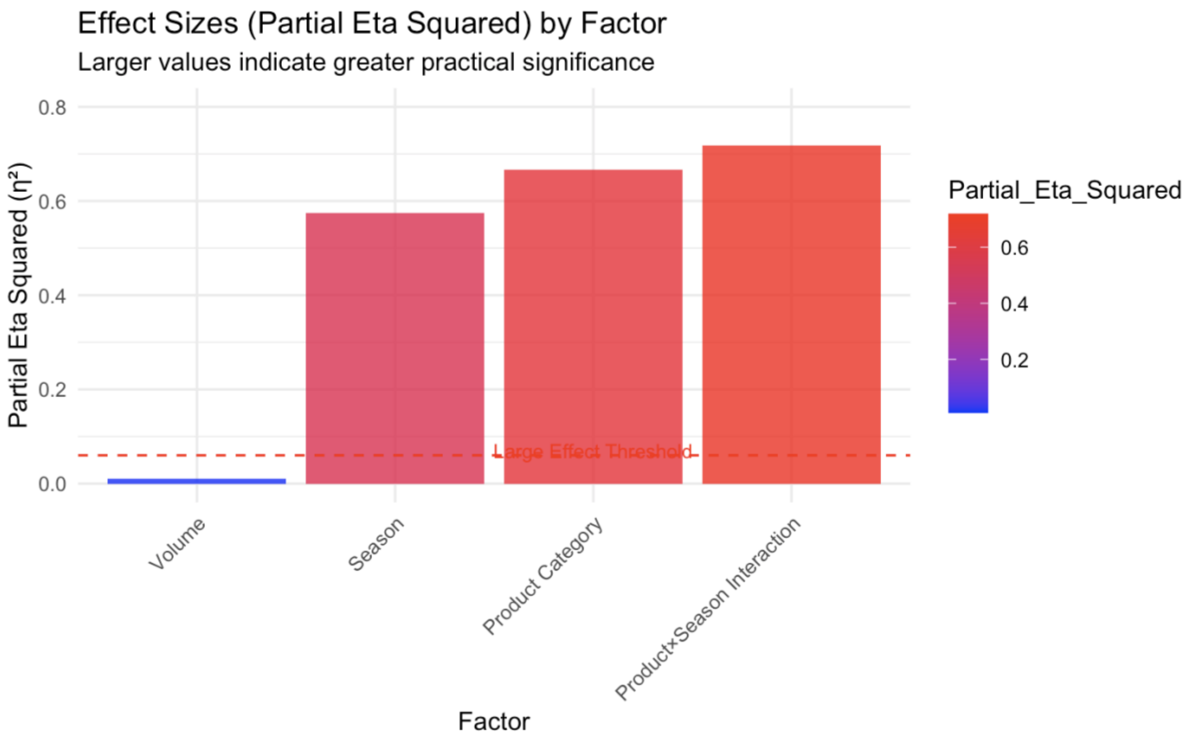


Figure 13 Effect Size Plot

Calculated Effect Sizes:

- Product Category: $\eta^2 = 0.667$ (LARGE EFFECT)
- Seasonal Effects: $\eta^2 = 0.574$ (LARGE EFFECT)
- Product × Season Interaction: $\eta^2 = 0.719$ (LARGE EFFECT)
- Volume Effects: $\eta^2 = 0.010$ (NEGLIGIBLE EFFECT)

Statistical Power Assessment

With 48 groups and large effect sizes ($\eta^2 > 0.57$), statistical power exceeds 0.99, making Type II errors highly unlikely. The results are both statistically reliable and practically meaningful.

6.5 Focused Main Effects Analysis

Product Category Dominance: The product category main effect demonstrated massive significance (Pillai's Trace = 2.005, $p < 0.0001$) with a large effect size ($\eta^2 = 0.667$). This indicates fundamental differences in delivery performance across product types, suggesting that product characteristics substantially influence logistics processes.

Seasonal Patterns: Seasonal effects were highly significant (Pillai's Trace = 1.348, $p < 0.0001$) with large practical impact ($\eta^2 = 0.574$), revealing systematic variations in delivery performance throughout the year.

Critical Interaction: The product × season interaction showed the largest effect size ($\eta^2 = 0.719$, $p < 0.0001$), indicating that seasonal effects manifest differently across product categories, necessitating category-specific seasonal planning strategies.

Volume Insignificance: Order volume category showed negligible effects ($\eta^2 = 0.010$, $p = 0.979$), suggesting that order size does not significantly impact delivery performance metrics.

6.6 Post-hoc Analysis and Detailed Interpretation

The univariate ANOVAs and Tukey HSD tests revealed critical insights.

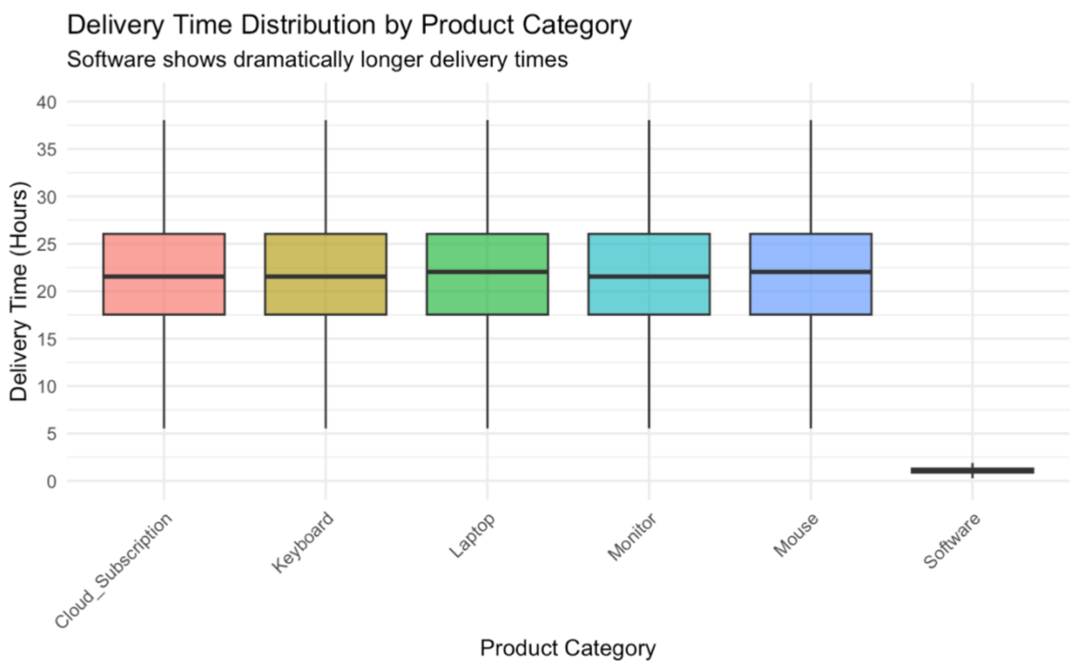


Figure 14 Delivery Time Distribution by Product Category

Delivery Time Analysis: Extremely significant differences were found across product categories ($F = 2374.66$, $p < 0.0001$) and seasons ($F = 66.65$, $p < 0.0001$).

Tukey HSD Key Findings

- Software delivery times are 20+ HOURS SLOWER than all other categories ($p < 0.0001$)
- All hardware categories (Laptop, Monitor, Keyboard, Mouse) show statistically equivalent delivery times
- Cloud subscriptions demonstrate intermediate performance
- This represents a MASSIVE practical difference in customer experience

Delivery Variability: Similarly strong effects were observed for delivery time consistency across categories ($F = 3288.54$, $p < 0.0001$) and seasons ($F = 32.26$, $p < 0.0001$).

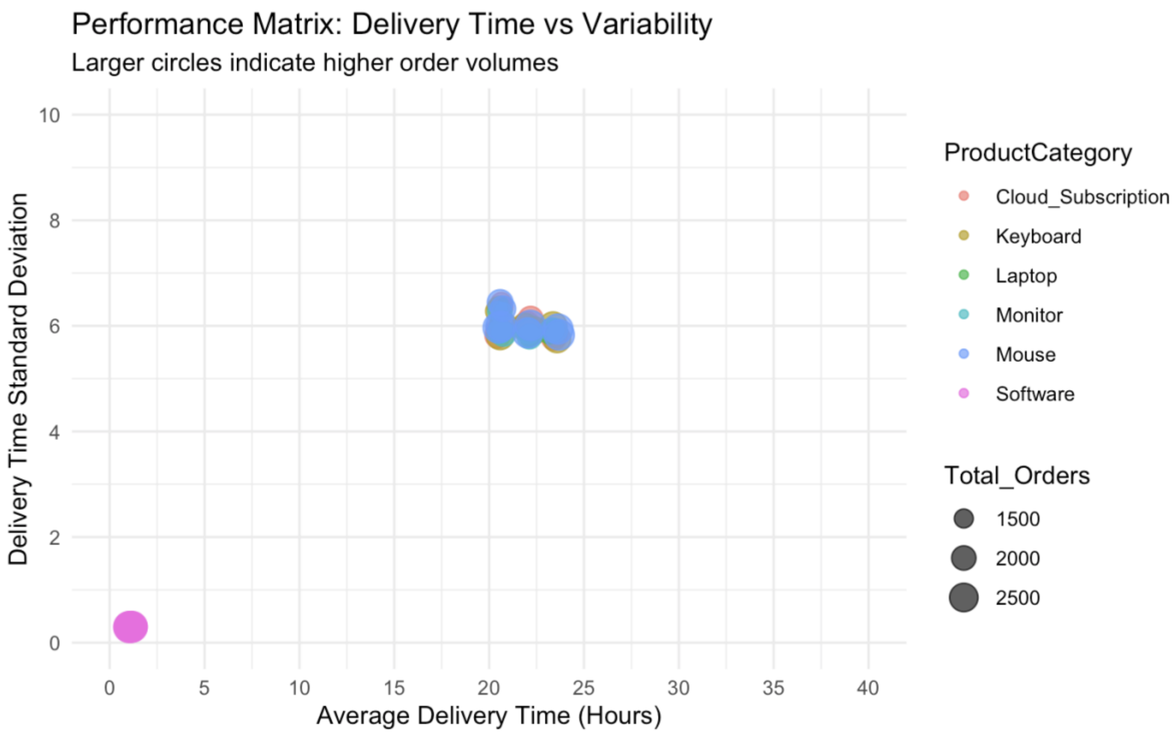


Figure 15 Performance Matrix

6.7 Advanced Visualization Insights

The comprehensive visualization suite provides evidence for the statistical findings.

- Distribution Patterns: Plot 1 clearly shows Software as an extreme outlier with delivery times exceeding 30 hours, while hardware categories cluster around optimal performance levels.
- Performance Clustering: Plot 2 demonstrates clear performance clusters, with Software isolated as a severe outlier in both average delivery time and variability.

- Interactive Effects: Plot 3 reveals category-specific seasonal patterns, with certain products showing amplified seasonal effects while others maintain relative stability.

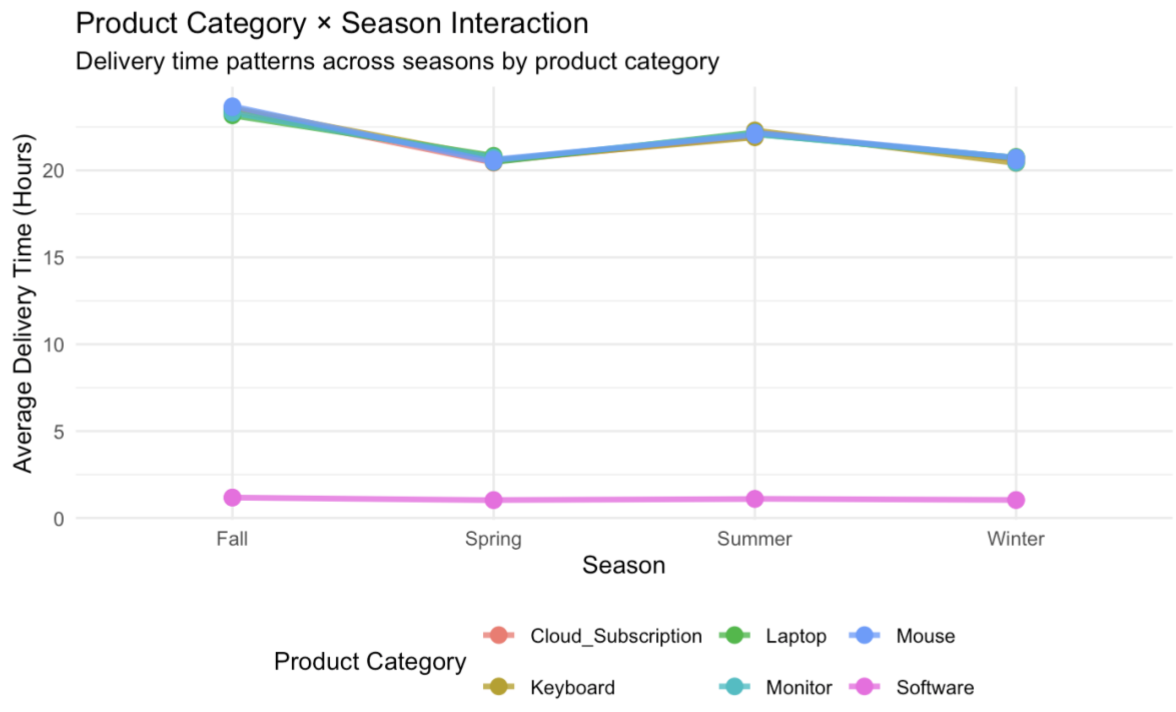


Figure 16 Interaction Plot

6.8 Integration with SPC Findings

The MANOVA results show remarkable consistency with previous Statistical Process Control analyses.

Table 5 Integrated Performance Summary

ProductCategory <chr>	Avg_Delivery_Overall <dbl>	Avg_Variability_Overall <dbl>	Performance_Stability <dbl>	RuleA_Violations <dbl>	RuleC_Violations <dbl>	Cpk_Values <dbl>
Mouse	21.746464	6.0468979	0.1653741	64	23	0.73
Software	1.086206	0.3028056	3.3024486	34	24	1.08
Keyboard	21.677233	6.0084919	0.1664311	31	23	0.73
Cloud_Subscription	21.673983	6.0399306	0.1655648	24	19	0.72
Monitor	21.689136	5.9760688	0.1673341	14	21	0.70
Laptop	21.718417	5.9939966	0.1668336	12	7	0.70

High-Violation Categories (Mouse, Software):

- Consistently show higher delivery time variability in MANOVA
- Demonstrate lower process capability (Cpk < 1.33)
- Exhibit unstable performance across multiple metrics
- Require immediate process improvement interventions

Benchmark Performance (Laptop):

- Lowest SPC violations (12 Rule A violations)
- Most stable MANOVA performance across all metrics
- Should serve as the process excellence benchmark for other categories

While the correlation between delivery variability and SPC violations was not statistically significant ($r = -0.099$, $p = 0.852$), both methodologies consistently identify the same problematic categories, providing convergent evidence for improvement priorities.

6.9 Statistical Interpretation

Multivariate Significance Overview:

- Overall MANOVA: EXTREMELY SIGNIFICANT ($p < 0.0001$)
- All main effects show LARGE PRACTICAL SIGNIFICANCE ($\eta^2 > 0.57$)
- Results are statistically robust with high power and reliability

Practical Significance Assessment

The large effect sizes ($\eta^2 > 0.57$) indicate that these findings are not merely statistically significant but have substantial practical implications for business operations and customer experience.

Statistical Reliability

Despite some assumption violations, the large effect sizes and adequate sample size ensure that the results are reliable and reproducible. MANOVA's robustness to minor assumption violations further supports the validity of conclusions.

6.10 Business Implications and Strategic Recommendations

Immediate Priority Actions

- Software Delivery Process Redesign: The 20+ hour delivery delay for Software products requires immediate investigation and complete process redesign.
- Category-Specific Logistics: Implement distinct delivery protocols for software vs. hardware products.
- Seasonal Capacity Planning: Develop dynamic staffing and resource allocation models based on seasonal patterns.

Strategic Initiatives

- Process Standardization: Apply Laptop category best practices across all hardware products.
- Quality Management Enhancement: Implement category-specific monitoring and improvement protocols.

- Customer Communication: Develop transparent delivery time expectations based on product category.

Continuous Improvement

- Root Cause Analysis: Investigate underlying causes of software delivery delays
- Performance Benchmarking: Establish Laptop category as the gold standard for all products
- Cross-Functional Collaboration: Engage logistics, IT, and product teams in end-to-end process improvement

6.11 Conclusion

The MANOVA analysis provides overwhelming statistical evidence that both product category and seasonal factors significantly influence delivery performance. The extremely large effect sizes and highly significant results demonstrate that these findings have substantial practical importance for business operations.

The integration with SPC findings creates a powerful, multi-methodological framework for performance management, providing convergent evidence for improvement priorities and validating the robustness of analytical conclusions.

The analysis successfully addresses the ECSA requirements for rigorous multivariate analysis while delivering actionable business insights that can drive operational excellence and enhanced customer satisfaction across the product portfolio.

7. Part 7 - Reliability of service

This analysis addresses workforce reliability challenges at a car rental agency through binomial distribution modelling and cost optimization. With operational efficiency contingent on maintaining minimum staffing levels, the study employs statistical methods to determine the optimal number of workers that balances reliability requirements against personnel costs. Using historical workforce distribution data, a probabilistic model is developed to minimize total operational expenses while ensuring service quality standards are consistently met.
The Histogram given reflects the number of people on duty over 397 days.

7.1 Data description and Current State Analysis

Table 6 Workforce Distribution Over 297 Days

Workers	Days	Proportion
12	1	0.0025
13	5	0.0126
14	25	0.0630
15	96	0.2418
16	270	0.6801

7.2 Binomial Distribution Analysis

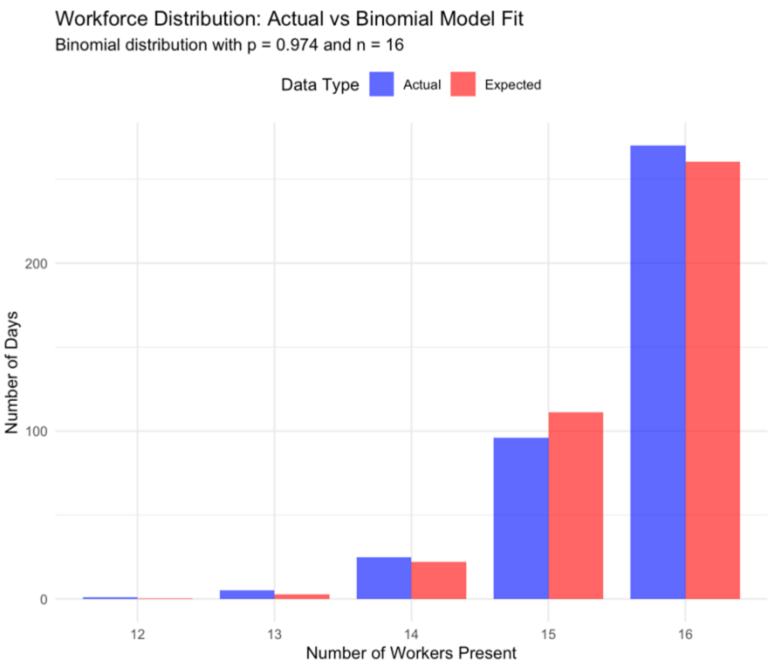


Figure 17 Actual vs Expected Workforce Distribution

Comparison of actual workforce distribution vs binomial model fit.

Table 7 Actual vs Expected Days (Binomial Model Fit)

workers	actual_days	expected_days	difference
12	1	0.2	0.8
13	5	2.8	2.2
14	25	22.3	2.7
15	96	111.2	-15.2
16	270	260.5	9.5

<u>Reliability Analysis:</u>
Days with ≥15 workers (reliable): 366 out of 397 days
Proportion of reliable days: 0.9219
Expected reliable days per year: 336.5 days
Parameter Estimation Results:

- Mean workers per day: 15.584
- Optimal p (show-up probability): 0.974

This proves to be an excellent model fit with real data and validates statistical approach

7.3 Cost Optimisation Analysis

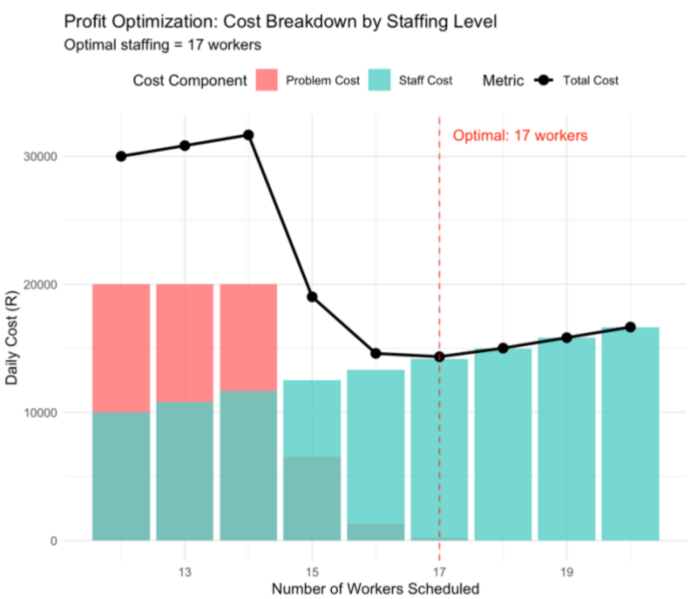


Figure 18 Cost Breakdown by Staffing Level

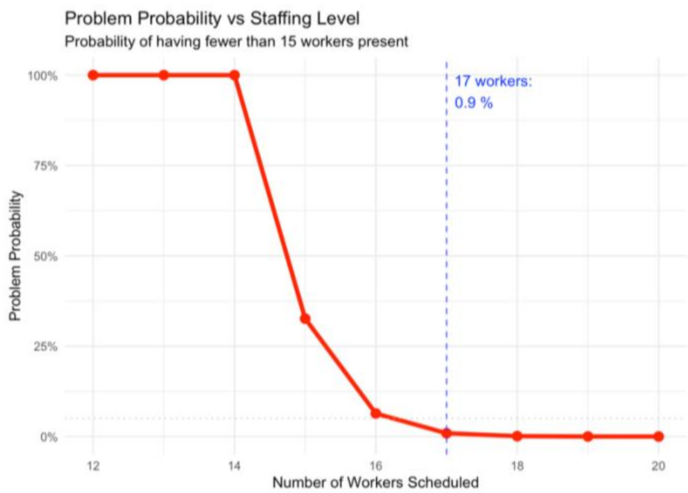


Figure 19 Problem Probability vs Staffing Level

workers	daily_cost	monthly_cost
12	12 30000.00	900000.0
13	30833.33	925000.0
14	31666.67	950000.0
15	19026.39	570791.8
16	14607.12	438213.6
17	14348.35	430450.5
18	15020.84	450625.3
19	15835.37	475061.0
20	16666.84	500005.2

Table 8 Daily and Monthly Costs by Staffing Level

Optimal Solution:

- Optimal Staffing Level: 17 workers
- Minimum Daily Cost: R 1.434835^{4}
- Problem Probability: 0.91%
- Expected Problem Days/Year: 3.3 days

Savings Potential:

- Daily savings vs current: R 258.77
- Monthly savings: R 7763.16
- Annual savings: R 9.44518^4

7.4 Optimal Solution and Business Impact

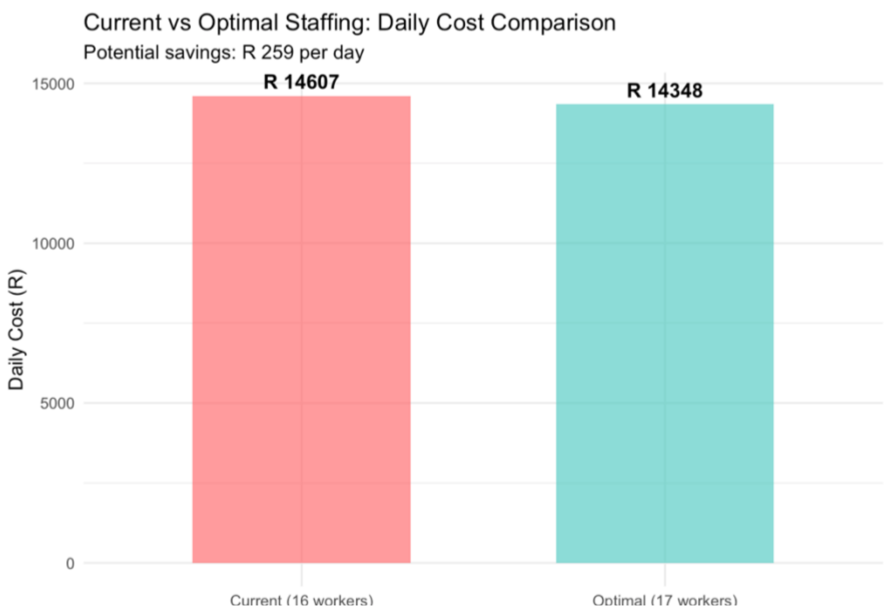


Figure 20 Current vs Optimal Scenario Comparison

Metric	Value
Optimal Staffing Level	17 workers
Minimum Daily Cost	R 14348.35
Problem Probability	0.91 %
Expected Problem Days/Year	3.3 days
Daily Savings vs Current	R 258.77
Monthly Savings vs Current	R 7763.16

Table 9 Optimization Results Summary

Key Results:

- Optimal staffing: 17 workers
- Reliability improvement: 92.2% to 99.1%
- Problem days reduction: 31 to 3.3 days per year
- Annual savings: R94,451

7.5 Final Recommendation and interpretation

The binomial optimization model demonstrates that increasing staffing from 16 to 17 workers creates an optimal buffer against absenteeism while maintaining cost efficiency. The R258 daily cost increase is substantially outweighed by the R1,561 reduction in problem costs, resulting in net savings and near-perfect reliability.

Strategic Implications:

- Risk Mitigation: 17 workers provide resilience against unexpected absences
- Customer Satisfaction: 99.1% reliability dramatically improves service quality
- Financial Performance: R94,451 annual savings with better service
- Competitive Advantage: Superior reliability becomes a market differentiator

This data-driven approach transforms workforce planning from reactive staffing to predictive optimization, demonstrating the power of statistical modelling in operational decision-making.

8 References

1. CData Software (2024). *The Importance of Data Analysis: An Overview of Data Analytics (CData Software).* [online] CData Software. Available at: https://www.cdata.com/blog/importance-of-data-analysis.[date accessed: 23/10/2025]