10/24/2025

ECSA 2025 PROJECT

Lucien Miro Brabazon Hallowes
INDUSTRIAL ENGINEERING FACULTY OF STELLENBOSCH UNIVERSITY

1. Introduction

This report analyses service and delivery performance across a retail business with multiple product lines and two service operations. The business offers laptops, monitors, keyboards, mice, software, and cloud subscriptions, and the analysis begins by assembling a clean, consistent dataset that joins corrected product records to time-ordered sales. With this foundation in place, the work proceeds in a structured engineering manner: exploratory summaries establish demand patterns and operational bottlenecks, and these observations motivate formal statistical modelling.

Process behaviour is examined using Statistical Process Control. For each product group, delivery records are partitioned into subgroups of twenty four, Phase-1 limits are estimated with X-bar and S charts, and subsequent data are monitored against fixed one, two, and three-sigma bands. Capability indices Cp, Cpk, Cpu, and Cpl are then computed for the first one thousand deliveries per product group under specification limits of zero and thirty two hours to assess whether the processes are able to meet the voice of the customer. The interpretation of Type I and Type II risks clarifies how often the monitoring system would raise false alarms or miss meaningful shifts.

The study complements SPC with designed comparisons. ANOVA is used to test whether delivery time varies by calendar month and whether the monthly pattern changes between years. Finally, operations are optimised in two applied settings. A queue-based revenue model identifies the number of baristas that maximises daily profit for each coffee shop after accounting for service time and staffing cost, and a binomial reliability model recommends staffing for a car-rental desk that minimises expected annual loss while keeping most days trouble-free. Together these components demonstrate how an engineering approach can translate raw transaction data into defensible operational decisions.

Contents

List of Figures

2. Descriptive Statistics

The project team received three comma separated value files titled customers, products_data, and sales2022and2023 for inspection and analysis. The analysis began with an exploratory review of each file conducted independently. The review verified data types, assessed value ranges and completeness, and identified anomalies. Descriptive statistics and preliminary visualizations were produced to establish baseline patterns and to evaluate data quality.

After the file level exploration, the findings were synthesized, and the datasets were prepared for integration where appropriate. This preparation enabled cross table analyses, including the linkage of customer attributes to product information and sales records. The staged approach improved data reliability and supported valid comparative and inferential analysis in the subsequent phases of the project.

2.1. Customer data:

A visual perspective of the data is produced to start the analysis:

```
'data.frame':   5000 obs. of  6 variables:
 $ CustomerID : 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" ...
 $ Age_Bracket: Ord.factor w/ 7 levels "<18"<"18-25"<..: 1 3 3 3 2 3 3 3 3 3 ...
```

First few rows of the data set to establish what the data sets look like:

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

6 rows

A summary of the data set is also shown below to help with the data sets understanding:

```
  CustomerID          Gender               Age              Income          City           Age_Bracket
 Length:5000        Length:5000        Min.   : 16.00   Min.   :  5000   Length:5000        65<    :1484
 Class :character   Class :character   1st Qu.: 33.00   1st Qu.: 55000   Class :character   26-35  : 890
 Mode  :character   Mode  :character   Median : 51.00   Median : 85000   Mode  :character   56-65  : 698
                                       Mean   : 51.55   Mean   : 80797                      36-45  : 681
                                       3rd Qu.: 68.00   3rd Qu.:105000                      46-55  : 621
                                       Max.   :105.00   Max.   :140000                      (Other): 610
                                                                                            NA's   :  16
```

Following the initial data exploration, the appropriate visualizations were identified. The resulting graphs are presented below.

This table reports the average customer income for seven cities. Miami shows the highest mean income at about 83,346 rand, followed by Chicago at roughly 82,244 rand. The remaining cities cluster tightly between about 79,700 and 80,500 rand, which indicates only a modest spread across most markets. It is noteworthy that New York and San Francisco, despite typically high costs of living, exhibit lower average incomes in this customer base than Miami and Chicago; this likely reflects differences in the company's local customer mix rather than citywide earnings. These figures are useful for tailoring marketing and pricing strategies; prioritising premium offers in higher-income cities and value-oriented promotions where incomes are lower.

City <chr>	Average_Income <dbl>
Chicago	82244.48
Houston	80248.62
Los Angeles	80475.21
Miami	83346.21
New York	79752.07
San Francisco	79852.56
Seattle	79947.99

7 rows

Figure 1

This figure is a scatter plot of customer age (x-axis) versus income (y-axis), with a fitted linear trend line and a 95% confidence band. The scatter plot shows a gentle upward trend, meaning income rises slightly with age. The Pearson correlation is r = 0.158, and the code output "Age and Income Correlation: 0.158" confirms a very weak positive linear relationship. Even so, the points are very spread out at every age, which means age alone explains only a small part of income differences. Younger and older customers overlap a lot in income, with examples of both low and high earners in each group.

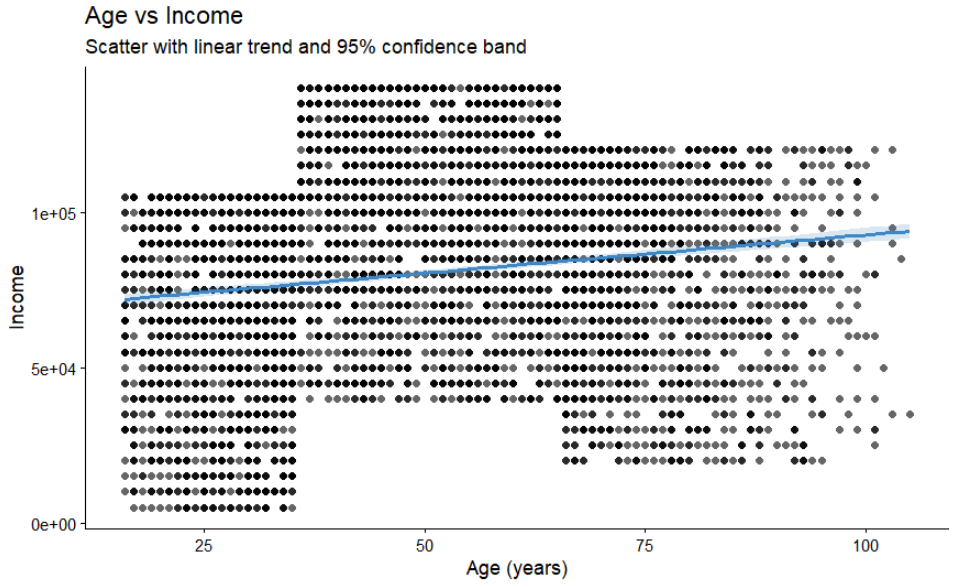


Figure 2

The following three displays all address the same question: whether income differs by gender in the customer data. The box plot compares medians and spreads; the three boxes sit at almost the same level with similar interquartile ranges and whiskers, which already suggests little or no difference. The overlaid density curves provide a view of distributional shape; the Female, Male, and other curves largely overlap, with only minor deviations in the tails, indicating very similar income distributions across groups. The ANOVA table formalizes this visual impression. The model reports F = 0.0017 with p = 0.9983, so the hypothesis that mean income is the same across genders cannot be rejected. The between-group sum of squares is tiny relative to the residual sum of squares, implying an effect size that is effectively zero; almost all variability in income occurs within genders, not between them. These graphs and the ANOVA were needed together: the plots show how similar the groups look and whether any differences are practically meaningful, while the test confirms that the observed differences are not statistically significant. The conclusion from the data is that gender is not a useful predictor of income in this sample.

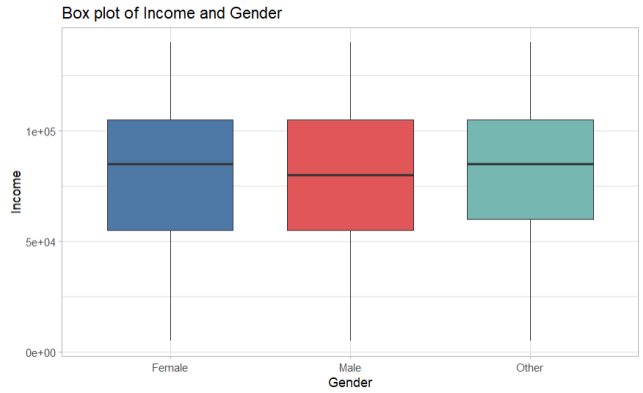


Figure 3

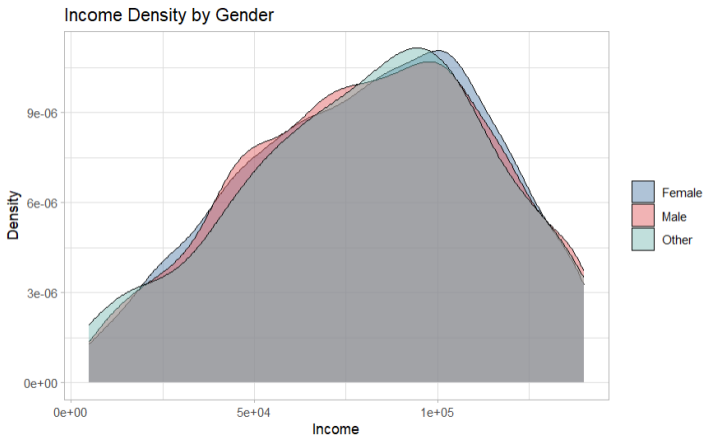


Figure 4

```
Analysis of Variance Table

Response: Income
            Df     Sum Sq    Mean Sq F value Pr(>F)
Gender       2 3.7947e+06    1897371  0.0017 0.9983
Residuals 4997 5.4935e+12 1099368653
```

Using age distribution is another useful source to analyse the customer data because it shows who the company serves. The summary table and bar chart indicate that the population is strongly weighted toward older customers, with 1 484 people in the 65 and older bracket. The next largest segment is 26–35 with 890 customers, while the 36–45, 46–55, and 56–65 brackets are all comparable size. The under-18 group is very small at 128, and only 16 records are missing age, so the pattern is not driven by missing data. The faceted histograms by city display the same shape in every location: older brackets make up a large share, and younger brackets appear far less often. Together these views show that the customer base is mature across cities, which suggests that marketing, product features, and pricing are likely to resonate more with middle-aged and older shoppers than with teens and very young adults.

Age_Bracket <ord>	Count <int>
<18	128
18-25	482
26-35	890
36-45	681
46-55	621
56-65	698
65<	1484
NA	16

Figure 7

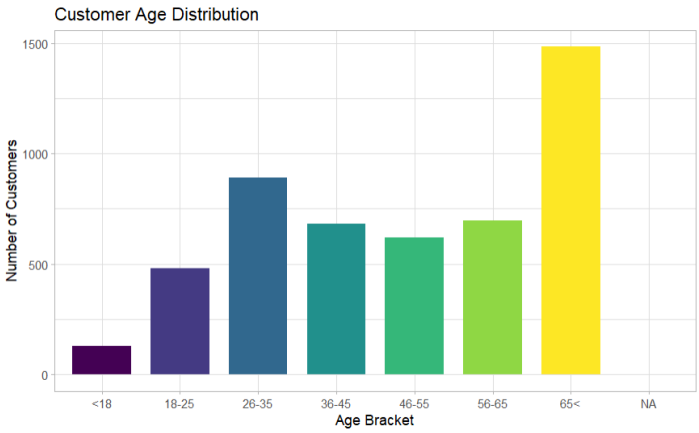


Figure 6

Figure 5

2.2. Products data

An initial set of overview visuals was created to establish a clear understanding of the datasets.

```
'data.frame':   60 obs. of  5 variables:
 $ ProductID   : chr  "SOF001" "SOF002" "SOF003" "SOF004" ...
 $ Category    : chr  "Software" "Cloud Subscription" "Laptop" "Monitor" ...
 $ Description : chr  "coral matt" "cyan silk" "burlywood marble" "blue silk" ...
 $ SellingPrice: num  512 505 494 543 516 ...
 $ Markup      : num  25.1 10.4 16.2 17.2 11 ...
```

Representative rows from each table were inspected to illustrate structure, fields, and typical values, and descriptive summaries were produced to clarify ranges, central tendencies, variability, and missingness.

	ProductID <chr>	Category <chr>	Description <chr>	SellingPrice <dbl>	Markup <dbl>
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
6	SOF006	Mouse	black silk	478.93	16.99

6 rows

```
  ProductID          Category         Description         SellingPrice        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  
```

The dataset was cleaned by realigning product categories to the ProductID prefix and by deriving a cost field. Before the correction, records with IDs beginning “SOF” were assigned a mix of categories such as Cloud Subscription, Laptop, Monitor, Keyboard, and Mouse, even though the prefix indicates Software. This mismatch meant any category-level summaries—such as average markup by category, revenue by category, or price comparisons—would have been inaccurate because software items were being counted under the wrong groups. It also obscured patterns in the product mix and could have led to faulty pricing or inventory conclusions. After the fix, every “SOF” item was consistently labelled as Software (and the same mapping rule was applied across the full file for other prefixes), restoring a one-to-one relationship between ProductID prefix and Category. In addition, a Cost Price column was created as Selling Price minus Markup, enabling margin and profitability analysis at the product and category levels. The corrected table therefore supports valid aggregation, comparison, and decision-making, whereas the original table would have produced biased results due to misclassified products and missing cost information. This new data is seen below.

	ProductID <chr>	Category <chr>	SellingPrice <dbl>	Markup <dbl>	CostPrice <dbl>
1	SOF001	Software	511.53	25.05	486.48
2	SOF002	Software	505.26	10.43	494.83
3	SOF003	Software	493.69	16.18	477.51
4	SOF004	Software	542.56	17.19	525.37
5	SOF005	Software	516.15	11.01	505.14
6	SOF006	Software	478.93	16.99	461.94

6 rows

After these corrections, the subsequent visuals offered a clearer, refreshed perspective on the data.

Category <chr>	n_items <int>	avg_markup <dbl>	avg_selling_price <dbl>	avg_cost_price <dbl>	avg_markup_pct_of_cost <dbl>
Laptop	10	18.430	18086.429	18067.999	0.1020035
Monitor	10	23.868	6310.525	6286.657	0.3796612
Cloud Subscription	10	19.956	1019.062	999.106	1.9973857
Keyboard	10	23.981	644.660	620.679	3.8636719
Software	10	16.040	506.183	490.143	3.2725143
Mouse	10	20.495	394.698	374.203	5.4769737

Using product-level margin views provides another lens on pricing performance. The histogram shows the distribution of item-level margin percentages, calculated as markup divided by cost price. Most items cluster at very low margins between roughly 0 percent and 2 percent, with a long, thin tail that reaches about 8 percent. This pattern implies that a large share of the catalogue is priced close to cost, which limits profitability and leaves little buffer for discounts or cost shocks. The concentration near zero suggests candidates for immediate price review or for negotiating lower costs.

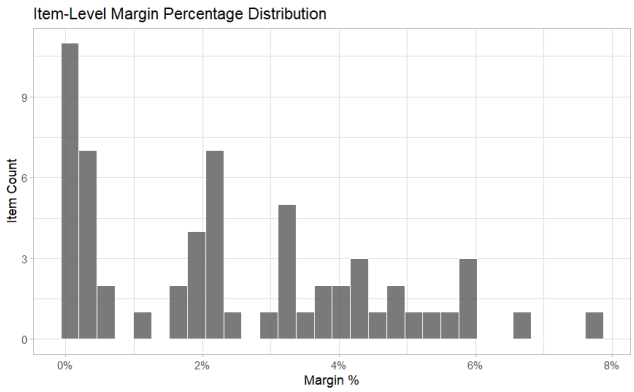


Figure 8

The category summary table links absolute prices to profitability. Laptops and Monitors carry high average selling and cost prices, so their average markups translate into very small margin percentages of about 0.10 percent and 0.38 percent respectively. Mice, Keyboards, and Software have much lower cost bases, so similar-sized markups yield higher percentages, for example about 5.48 percent for Mice and about 3.86 percent for Keyboards. Cloud Subscriptions sit in the middle at roughly 2.00 percent. These differences show that the same absolute markup can imply very different profitability once costs are considered.

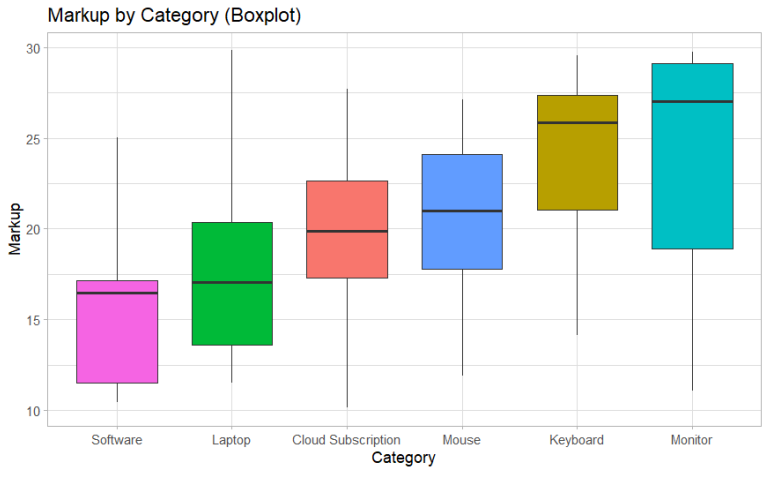


Figure 9

The box plot presents the distribution of absolute markup by category. Monitors and Keyboards exhibit the highest median markups and the widest spread, while Software has the lowest median and a tighter range. When interpreted together with the table, the plot indicates that categories with the largest absolute markups are not necessarily the most profitable on a percentage basis. Operationally, the findings support setting margin floors by category, prioritising margin improvement for Laptops and Monitors, and monitoring near-zero margin items identified by the histogram.

2.3. Sales Merged data

An initial set of overview visuals was created to establish a clear understanding of the datasets.

```
'data.frame':   100000 obs. of  9 variables:
 $ CustomerID   : chr  "CUST1791" "CUST3172" "CUST1022" "CUST3721" ...
 $ ProductID    : chr  "CLO011" "LAP026" "KEY046" "LAP024" ...
 $ Quantity     : int  16 17 11 31 20 32 29 1 10 1 ...
 $ orderTime    : int  13 17 16 12 14 21 5 19 19 18 ...
 $ orderDay     : int  11 14 23 18 7 24 23 9 13 30 ...
 $ orderMonth   : int  11 7 5 7 2 12 1 6 12 4 ...
 $ orderYear    : int  2022 2023 2022 2023 2022 2022 2022 2023 2023 2022 ...
 $ pickingHours : num  17.7 38.4 14.7 41.4 15.7 ...
 $ deliveryHours: num  24.5 31.5 21.5 24.5 24 ...
```

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

6 rows

```
 CustomerID         ProductID          Quantity       orderTime       orderDay       orderMonth      orderYear
Length:100000      Length:100000      Min.   : 1.0   Min.   : 1.00   Min.   : 1.0   Min.   : 1.000   Min.   :2022
Class :character   Class :character   1st Qu.: 3.0   1st Qu.: 9.00   1st Qu.: 8.0   1st Qu.: 4.000   1st Qu.:2022
Mode  :character   Mode  :character   Median : 6.0   Median :13.00   Median :15.0   Median : 6.000   Median :2022
                                      Mean   :13.5   Mean   :12.93   Mean   :15.5   Mean   : 6.448   Mean   :2022
                                      3rd Qu.:23.0   3rd Qu.:17.00   3rd Qu.:23.0   3rd Qu.: 9.000   3rd Qu.:2023
                                      Max.   :50.0   Max.   :23.00   Max.   :30.0   Max.   :12.000   Max.   :2023

 pickingHours     deliveryHours
Min.   : 0.4259   Min.   : 0.2772
1st Qu.: 9.3908   1st Qu.:11.5460
Median :14.0550   Median :19.5460
Mean   :14.6955   Mean   :17.4765
3rd Qu.:18.7217   3rd Qu.:25.0440
Max.   :45.0575   Max.   :38.0460
```

The sales dataset contains 100,000 transactions with customer and product identifiers, order timing (hour, day, month, year), quantities, and process metrics for picking and delivery. Orders span 2022 and 2023, quantities are right-skewed (median 6, mean ≈13.5, max 50), and process times vary widely (picking ~14–15 hours on average; delivery ~17–20 hours), indicating notable operational variability. The order date fields were first corrected and standardised, then the records were re-ordered chronologically. The corrected view is shown in the image below. For clearer analysis and visualisation, the cleaned products dataset was then merged with the newly ordered sales data, allowing prices, markups, and product categories to be examined alongside transaction volumes and timing.

CustomerID <chr>	ProductID <chr>	Quantity <dbl>	orderTime <dbl>	orderDay <dbl>	orderMonth <dbl>	orderYear <dbl>	pickingHours <dbl>	deliveryHours <dbl>	Gender <chr>
CUST3795	MOU059	4	1	1	1	2022	16.3883333	9.5440	Male
CUST2337	KEY049	7	1	1	1	2022	10.3883333	18.5440	Male
CUST3281	SOF009	5	1	1	1	2022	0.4258889	0.6772	Female
CUST3721	CLO019	47	1	1	1	2022	11.3883333	19.5440	Female
CUST4015	KEY045	1	1	1	1	2022	12.3883333	15.5440	Female
CUST4364	SOF005	2	10	1	1	2022	0.5592222	0.5272	Female

6 rows | 1-10 of 19 columns

A summary of the new merged data is then printed to inspect:

```
  CustomerID         ProductID           Quantity      orderTime       orderDay      orderMonth       orderYear
 Length:100000      Length:100000      Min.   : 1.0   Min.   : 1.00   Min.   : 1.0   Min.   : 1.000   Min.   :2022
 Class :character   Class :character   1st Qu.: 3.0   1st Qu.: 9.00   1st Qu.: 8.0   1st Qu.: 4.000   1st Qu.:2022
 Mode  :character   Mode  :character   Median : 6.0   Median :13.00   Median :15.0   Median : 6.000   Median :2022
                                       Mean   :13.5   Mean   :12.93   Mean   :15.5   Mean   : 6.448   Mean   :2022
                                       3rd Qu.:23.0   3rd Qu.:17.00   3rd Qu.:23.0   3rd Qu.: 9.000   3rd Qu.:2023
                                       Max.   :50.0   Max.   :23.00   Max.   :30.0   Max.   :12.000   Max.   :2023

  pickingHours     deliveryHours        Gender               Age            Income          City
 Min.   : 0.4259   Min.   : 0.2772   Length:100000      Min.   : 16.00   Min.   :  5000   Length:100000
 1st Qu.: 9.3908   1st Qu.:11.5460   Class :character   1st Qu.: 33.00   1st Qu.: 55000   Class :character
 Median :14.0550   Median :19.5460   Mode  :character   Median : 51.00   Median : 85000   Mode  :character
 Mean   :14.6955   Mean   :17.4765                      Mean   : 51.57   Mean   : 80699
 3rd Qu.:18.7217   3rd Qu.:25.0440                      3rd Qu.: 69.00   3rd Qu.:105000
 Max.   :45.0575   Max.   :38.0460                      Max.   :105.00   Max.   :140000

  Age_Bracket      Category          Description         SellingPrice        Markup         CostPrice
 65<     :29962   Length:100000      Length:100000      Min.   :  350.4   Min.   :10.13   Min.   :  323.3
 26-35   :17207   Class :character   Class :character   1st Qu.:  493.7   1st Qu.:16.18   1st Qu.:  477.5
 56-65   :13859   Mode  :character   Mode  :character   Median :  627.9   Median :20.44   Median :  601.8
 36-45   :13343                                         Mean   : 3243.8   Mean   :20.42   Mean   : 3223.3
 46-55   :12676                                         3rd Qu.: 5346.1   3rd Qu.:25.56   3rd Qu.: 5316.4
 (Other):12672                                         Max.   :19725.2   Max.   :29.84   Max.   :19713.5
 NA's    :  281
```

The merged table combines customer, product, price, and operations fields in one structure, so segmenting by city, category, or age bracket is now straightforward. Data quality is high, with only 281 ages missing, and the variables cover wide operational and financial ranges: income spans 5,000 to 140,000; selling price runs from about 350 to nearly 19,725; cost price from about 323 to 19,713; and markup from roughly 10 to 29.8. Operational times also show broad dispersion, with picking between ~0.43 and ~45.06 hours and delivery between ~0.28 and ~38.05 hours. This integrated view enables direct analysis of profitability, service performance, and customer mix across products and cities without further joins.

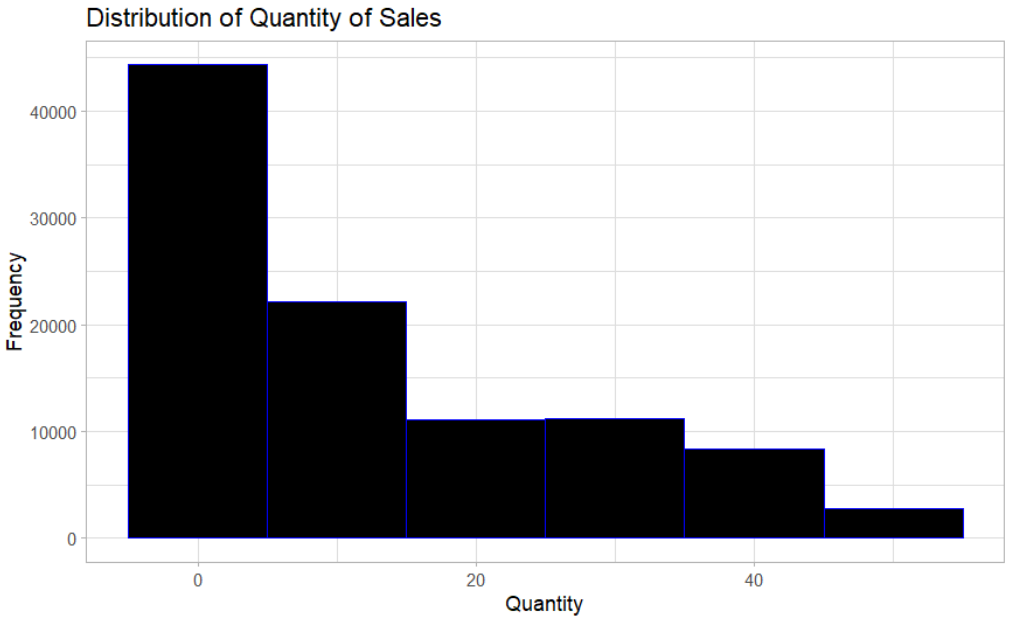


Figure 10

This histogram shows the distribution of order quantities. It is strongly right skewed: most transactions involve only a few units, while large orders are uncommon. The long upper tail, extending to roughly 50+ units, indicates occasional bulk purchases that pull the average above the median. Operationally, this pattern suggests inventory and picking processes should be

optimized for small, frequent orders, with separate handling or planning for intermittent high-volume buys.

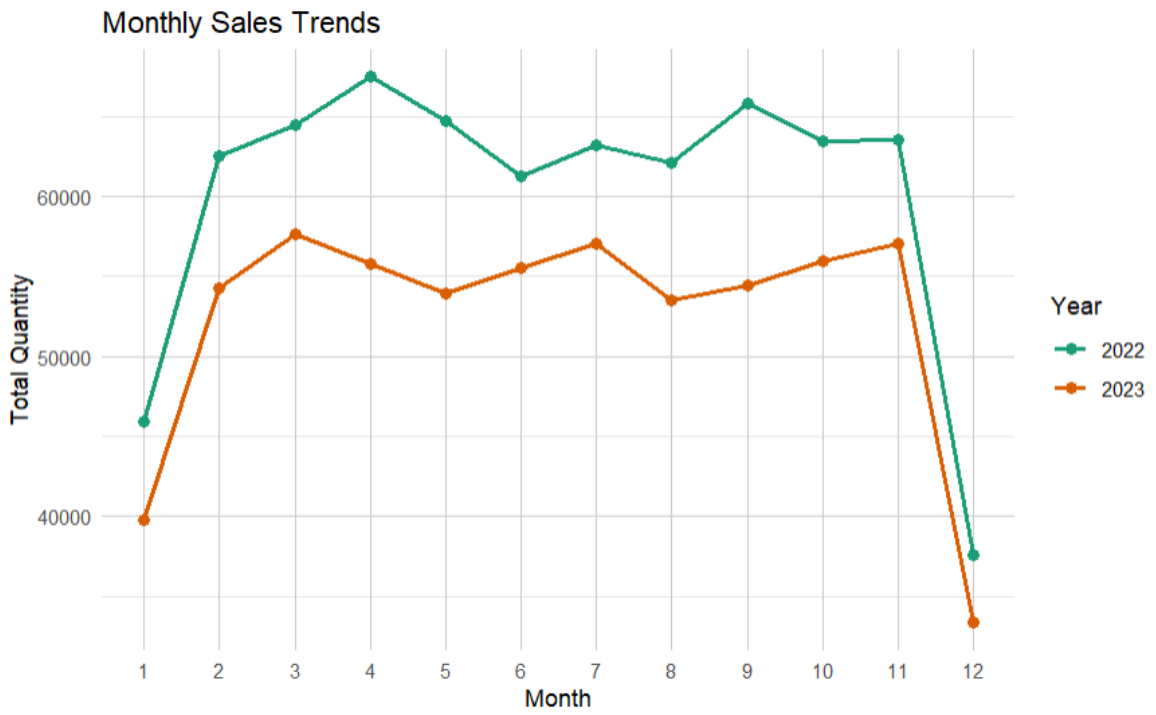


Figure 11

This line chart tracks total monthly quantity for 2022 and 2023. In both years sales climb sharply after January, but 2022 remains consistently higher than 2023 by roughly five to ten thousand units across most months. The 2022 series peaks around April, softens through May–June, recovers in late summer and early autumn, and shows another local high around September before tapering. The 2023 series follow a flatter mid-year path with smaller swings, rising into March, easing through May, and then hovering near the mid-50,000s until November. Both years drop steeply in December, which likely reflects a seasonal slowdown or a partial month in the data extract. The larger amplitude in 2022 suggests more volatility or stronger promotional spikes, whereas 2023 appears steadier but at a lower level. Operationally, the pattern points to higher demand planning needs from February to April and again around late summer, with reduced volumes in December.

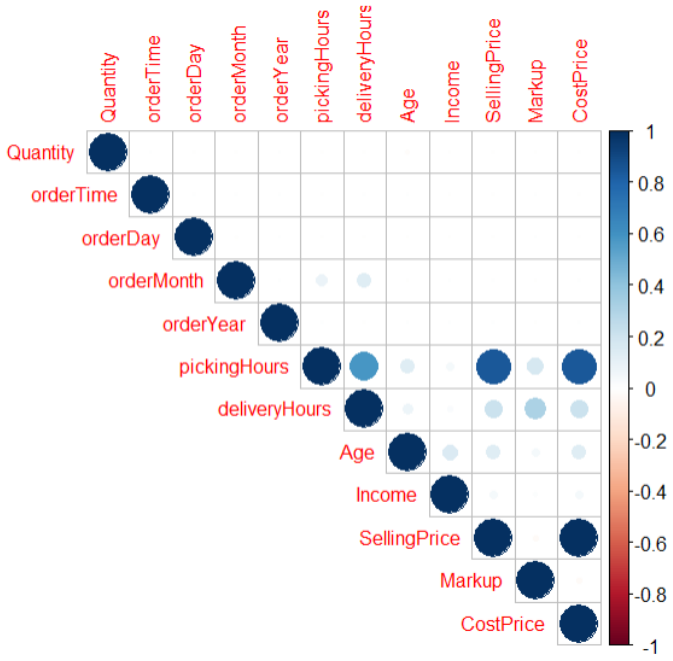


Figure 12

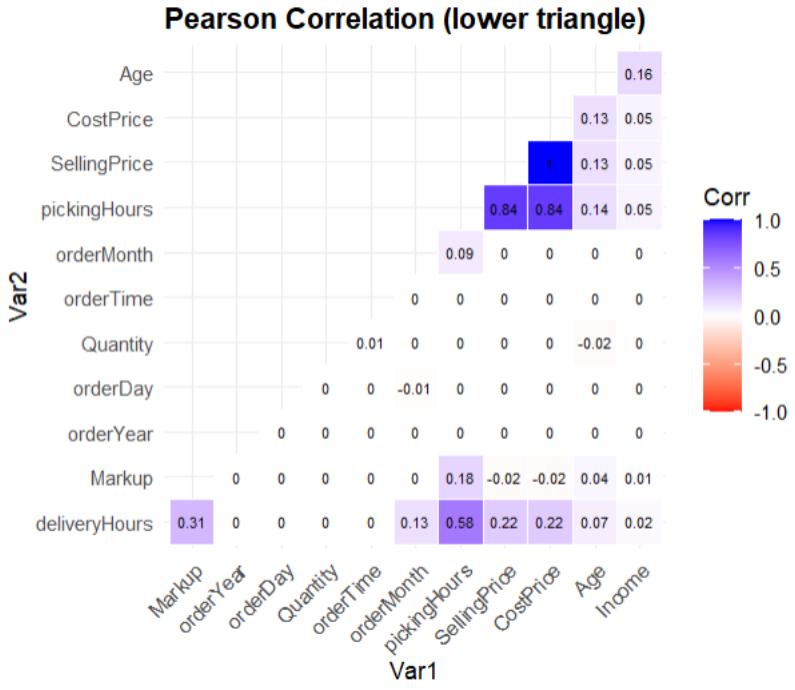


Figure 13

These two correlation heatmaps tell a consistent story about the linear relationships in the merged data. The strongest association is between selling price and cost price ($r \approx 1.00$), which is expected and confirms that the pricing fields are internally consistent. Operational times move together as well: picking hours and delivery hours show a moderate positive link ($r \approx 0.55–0.60$), indicating that slower picking tends to coincide with slower delivery. Markup is only weakly related to prices and costs ($r \approx 0.13–0.18$), so higher-priced items carry somewhat larger absolute markups, but the effect is small. Demographics show limited connections to financials, with age and income having a weak positive relationship ($r \approx 0.16$). Calendar variables (order time, day, month, year) and quantity have correlations near zero with most other fields, suggesting they do not drive pricing or process durations in a linear way. Taken together, both plots

reinforce that cost and price co-move tightly, the two operational time measures are moderately aligned, and most other pairs exhibit little linear dependence.

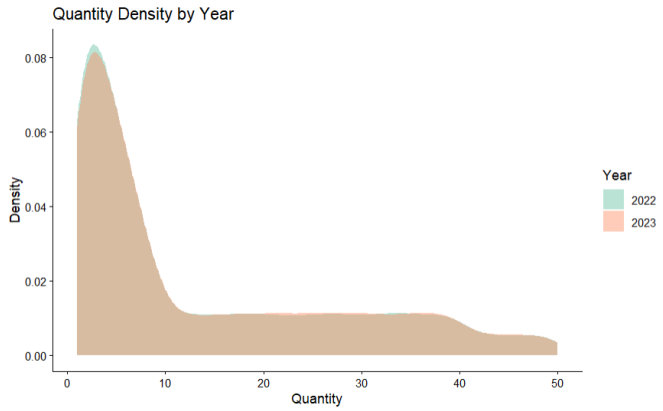


Figure 14

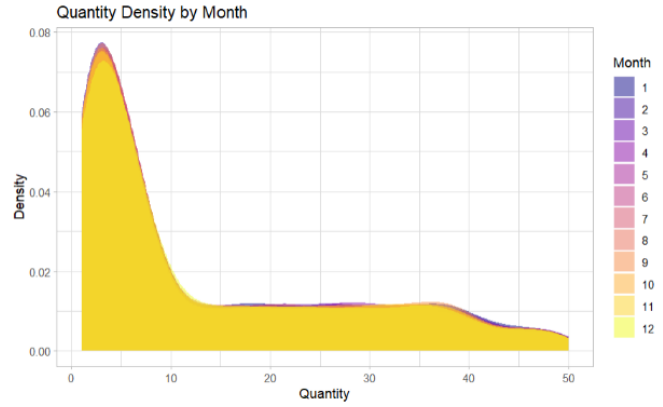


Figure 15

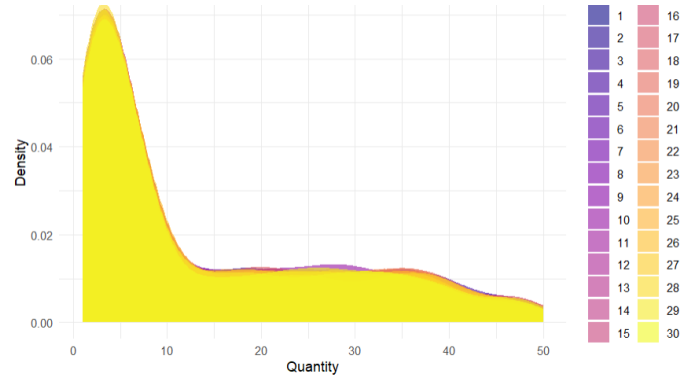


Figure 16

These three density plots examine the distribution of order quantity across years, months, and days, and the shapes are essentially identical in all three views. The year plot shows 2022 and 2023 overlapping almost perfectly. Most orders are for a few units, densities fall sharply after about five units, and a long thin tail extends to approximately 50. The month plot repeats the same curve for every month, and the day plot does the same across all calendar days, with no segment showing a meaningfully different order-size profile. When densities are "identical," it means the whole distribution, not just the average, matches across time periods, so the probability of seeing any order size is effectively the same in January as in July, and on the 1st as on the 20th. Practically, this implies that order size is right-skewed and time-invariant: seasonality may change how many orders occur, but not how big each order tends to be. That stability lets operations and inventory teams use a single, pooled order-size distribution for planning, simplifying picking strategies and capacity models. It also suggests promotions or calendar effects did not materially shift basket sizes in this data. Improvements should therefore target volume (number of orders) or conversion rather than expecting large changes in units per order.

Figure 17

This figure is a scatter plot of picking hours (x-axis) versus delivery hours (y-axis), coloured by product category and faceted by year. Across both 2022 and 2023 the cloud shows a consistent, weak-to-moderate positive association: orders that take longer to pick also tend to take longer to deliver. Categories separate in intuitive ways. Software clusters at the lower-left with near-zero picking and delivery times, reflecting digital fulfilment. Physical goods occupy higher ranges, with laptops generally showing the longest picking and delivery times, monitors in a mid-range, and items such as keyboards, mice, and cloud subscriptions tending to the lower end among physical or service items. The similarity of the two panels suggests the operational profile is stable year-over-year, so performance gains are more likely to come from process improvements (especially in picking for laptop workflows) rather than from seasonal changes.

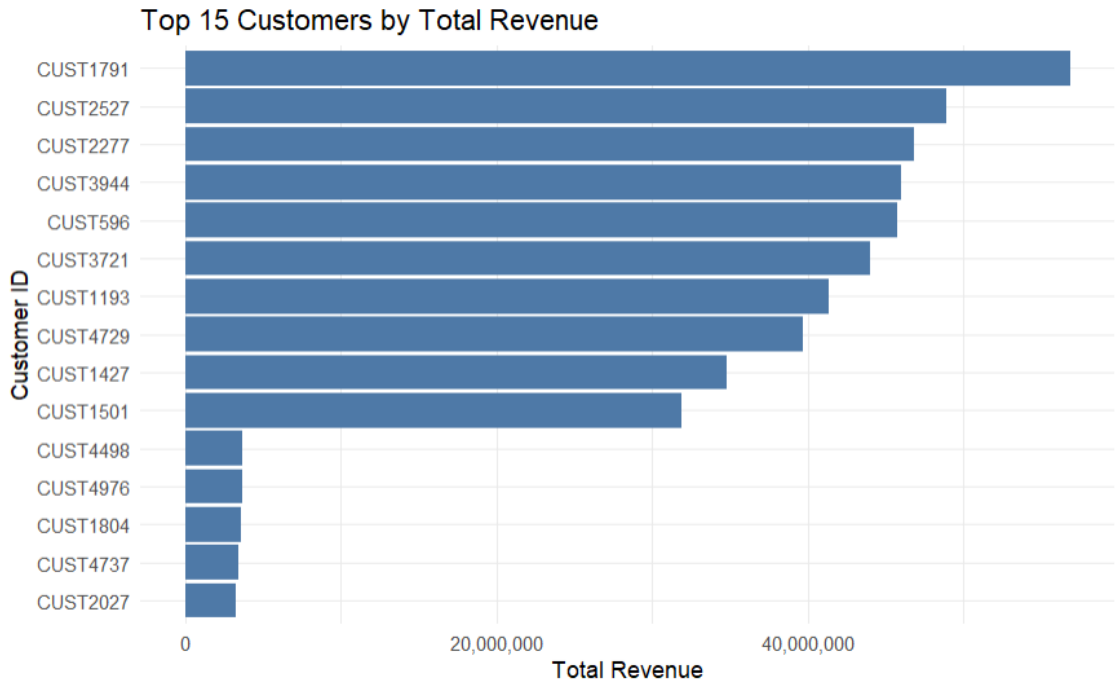


Figure 18

This horizontal bar chart ranks the top 15 customers by total revenue. Revenue is highly concentrated: CUST1791 generates a markedly larger amount than any other account. A compact second tier (roughly CUST2527 through CUST596) sits below the leader with similar totals, followed by a noticeable step-down through CUST1501. The final five customers contribute relatively little compared with the leaders, forming a long tail. The pattern implies dependence on a few key accounts, so retention and risk monitoring for those customers are critical. At the same time, the clustered mid-tier represents an opportunity for targeted upsell and expansion to reduce concentration risk and broaden the revenue base.

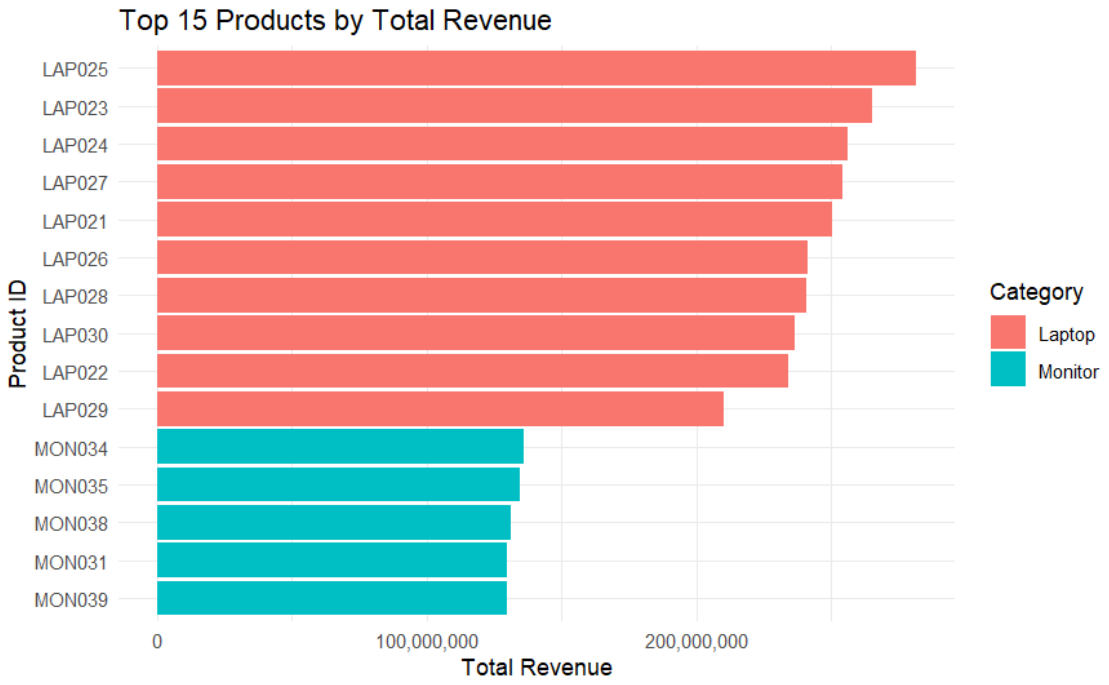


Figure 19

This bar chart ranks the top 15 products by total revenue, with colour indicating category. Laptops dominate the list, occupying the entire top ten and generating substantially higher revenue per SKU (stock keeping unit) than the monitors that make up the remaining five positions. The lead laptop (LAP025) sits well ahead of the rest, followed by a tight cluster of other laptop models, while monitor revenues form a lower, more uniform tier. This concentration implies that category performance is driven primarily by a handful of laptop SKUs, which is great for focus but increases dependency risk if any of those models face supply or demand shocks. Operationally, stocking, forecasting, and promotional planning should prioritise these laptop leaders, with cross sell opportunities into monitors. Because earlier results showed laptops and monitors have relatively low margin percentages, the revenue dominance here does not automatically translate to profit leadership.

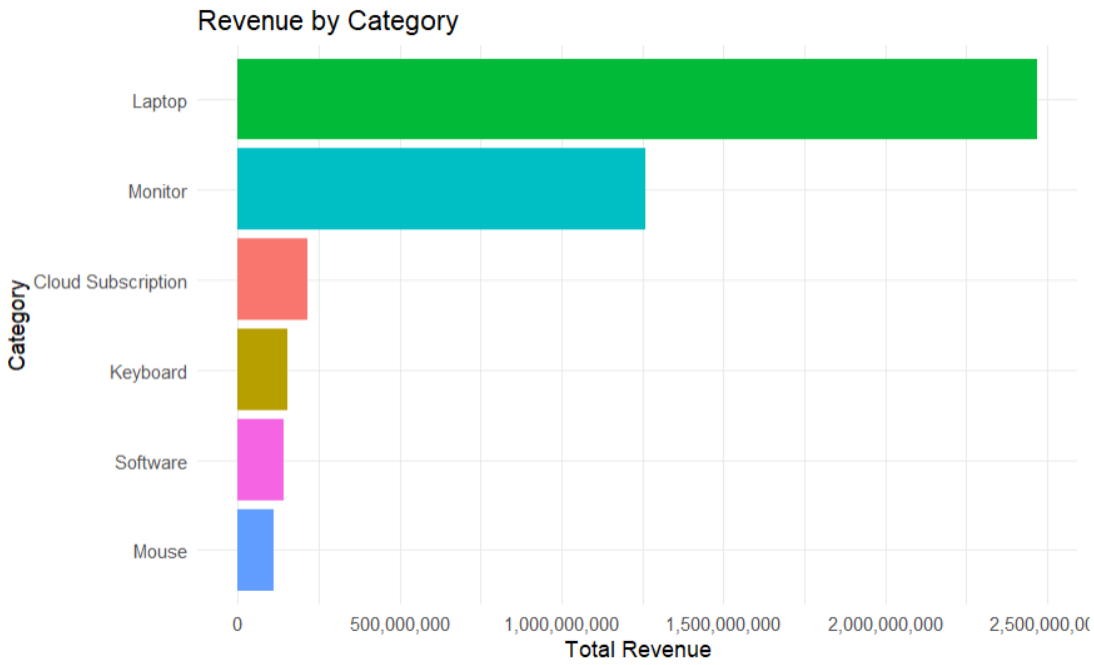


Figure 20

This bar chart summarizes total revenue by category. Laptops dominate overall sales by a wide margin, with monitors a clear but distant second, while cloud subscriptions, keyboards, software, and mice each contribute only a small share. The pattern is highly concentrated and Pareto-like, implying that most revenue comes from two hardware categories. Operationally, forecasting, inventory, and promotional effort should prioritize laptops and monitors because they drive volume and supply risk. At the same time, the smaller categories offer room for growth through bundling and cross-sell, especially alongside laptop purchases. Because earlier analysis showed relatively low margin percentages in some hardware lines, this view should be paired with margin metrics to ensure that revenue leadership translates into profit.

3. Statistical Process Control

3.1. Initialisation of X-charts and s-charts

To initialise the statistical process control study for delivery times, the data was first converted into a time-ordered stream for each product type by sorting from oldest to newest using year, month, day, and order time. That stream was partitioned into consecutive non-overlapping subgroups of twenty-four deliveries so the analysis would mimic data arriving in real time. The first thirty subgroups, which comprise seven hundred and twenty observations, were reserved to build the baseline. For each subgroup the mean and the standard deviation were calculated. From these values the grand mean of the subgroup means, and the average within-subgroup standard deviation were obtained. Using the standard SPC constants for a subgroup size of twenty-four, the centre line and three-sigma control limits were fixed for both charts. The X-bar chart uses the grand mean as the centre line with limits equal to the grand mean plus or minus A3 multiplied by the average standard deviation. The S chart uses the average standard deviation as the centre line with limits equal to B3 multiplied by the average standard deviation and B4 multiplied by the average standard deviation. One-sigma and two-sigma guide bands were added by dividing the distance between the centre line and the three-sigma limits into thirds. This initial thirty by twenty-four window constitutes Phase 1 and is used only to estimate stable limits that remain fixed for subsequent monitoring.

Two representative charts for the MOU product group are shown. The X-bar chart plots the mean delivery time for each consecutive sample of twenty-four orders against the sample number. The dashed blue line is the centre line, the purple dashed lines are the three-sigma control limits, the green dashed lines indicate the two-sigma guides, and the light-blue dashed lines indicate the one-sigma guides. The S chart displays the within-sample standard deviation for the same subgroups, with the red dashed line as the centre line and the purple dashed lines as the three-sigma control limits. The complete set of charts for all product groups is presented in Appendix A as graphs one through twelve. Across these examples and across the full appendix, no X-bar points exceed the three-sigma limits, and the S charts also remain within their control limits. This behaviour indicates that within the Phase 1 window the process means and dispersion for delivery time were stable and in statistical control for the product groups reviewed.

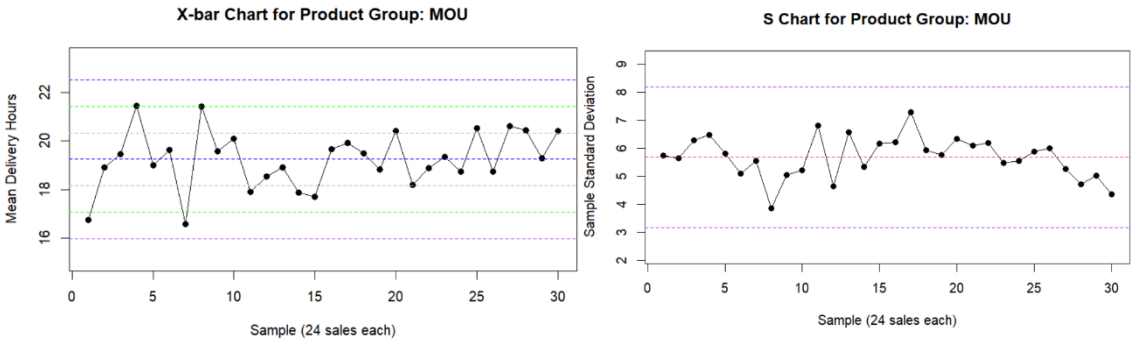


Figure 22

Figure 21

3.2. Ongoing SPC of Delivery Times Using X-bar and s Charts

For Part 3.2 the analysis moved from limit-setting to ongoing monitoring. After fixing the Phase-1 limits from the first thirty subgroups, the remaining observations were processed chronologically in the same subgroup size of twenty-four, creating samples numbered 31, 32, and so on. For each new subgroup the mean and the standard deviation were computed and plotted on the X-bar and S charts against the fixed Phase-1 centre lines and control limits. Interpretation followed the standard SPC workflow: the S chart was reviewed first to confirm that within-subgroup variability remained in control; only then was the corresponding X-bar point considered reliable. Any subgroup that crossed a three-sigma limit was flagged, and additional sensitising rules were used to detect subtler shifts, such as runs or clusters near the two-sigma bands and persistent trends above or below the centre line. This monitoring step provides product managers with timely signals about when the delivery-time process is behaving normally and when investigation or adjustment is warranted. We now evaluate the results from these monitoring charts; all figures are presented in Appendix B as graphs 13 through 18.

Graph 13: X-bar and S charts: MOU

This figure shows Phase-2 monitoring for the MOU product group using subgroups of twenty-four deliveries. The top panel is the X-bar chart, which plots the subgroup means against the fixed Phase-1 centre line and the 1σ, 2σ, and 3σ reference bands. Early in the series the means fluctuate tightly around the centre line, but beginning roughly after sample 200 there is a sustained upward shift: many points sit above the +2σ band and several are flagged in red, indicating out-of-control signals. Around sample 500 the mean drops back toward the centre line before drifting upward again through the final third of the series, where additional red points reappear. These step changes and runs above the upper bands are consistent with real shifts in the process mean rather than random noise.

The bottom panel is the S chart for the same subgroups, with the centre line at the average within-subgroup standard deviation and three-sigma limits shown by the dashed purple lines. Most points remain comfortably within the limits and cluster near a standard deviation of roughly five to six hours, with only a few brief spikes toward the upper limit. Because the dispersion is largely in control, the alerts seen on the X-bar chart are best interpreted as changes in average delivery time rather than explosions in variability.

Taken together, this graph indicates that the MOU delivery process experienced periods where the average time increased materially while variability stayed broadly stable. That pattern points to assignable causes affecting level, such as workload surges, routing or carrier changes, or staffing constraints, rather than general loss of control in spread. These episodes merit investigation and, if confirmed, targeted corrective action.

Graph 14: X-bar and S charts: KEY

This figure tracks delivery performance for the KEY product group in Phase-2 monitoring, with each point summarising twenty-four consecutive orders. In the X-bar chart (top), the subgroup means are plotted against the fixed Phase-1 limits and the 1σ, 2σ, and 3σ guide bands. The series begins close to the centre line but drifts upward through the first 200 to 350 samples, where many points sit above the +2σ band and several exceed the 3σ limit (flagged in red). Around sample 400 there is a clear downward step, after which the mean starts another gradual climb; by the final third, the pattern repeats with persistent points near or above the upper bands and additional out-of-control signals. These runs and step changes indicate real shifts in the average delivery time rather than random variation.

The S chart (bottom) shows the within subgroup standard deviation fluctuating around a roughly constant level, with most points comfortably inside the control limits and only occasional spikes. Because dispersion remains largely in control, the red signals on the X-bar chart are best interpreted as changes in process level rather than widening spread. Operationally, the KEY stream appears to experience episodic increases in average delivery time, potentially due to workload peaks, routing or carrier changes, or temporary capacity constraints, while the underlying variability is stable. Investigation should focus on the timing of the upward shifts (e.g., the periods surrounding the rises before ~350 and after ~600) to isolate assignable causes and restore the mean to its baseline.

Graph 15: X-bar and S charts: SOF

The SOF X-bar chart (top) tracks subgroup means for consecutive batches of twenty-four deliveries against the fixed Phase-1 limits and the 1σ/2σ/3σ guide bands. Average delivery time is short, roughly around one hour, but there is a clear upward drift over the first half of the series, followed by a step down near sample ~450 and then another gradual rise. Many points in the mid and late segments sit above the +2σ band and several are flagged beyond the 3σ limit (red), indicating sustained shifts in the process mean rather than random noise. These episodes suggest changes in operating conditions, such as workload peaks, batching or release timing, or temporary capacity constraints, that push the mean higher even though the overall level remains near one hour.

The S chart (bottom) shows within-subgroup dispersion clustering near about 0.30 hours and staying well inside the 3σ limits with only occasional spikes. Because variability is largely stable, the out-of-control signals on the X-bar chart are best interpreted as level shifts in the mean delivery time, not loss of control in spread. Operationally, SOF is a fast stream with tight variability, but it exhibits recurring upward movements in average delivery time that merit targeted investigation and corrective action.

Graph 16: X-bar and S charts: CLO

The X-bar chart shows the subgroup means for CLO plotted against fixed Phase-1 limits with 1σ, 2σ, and 3σ guide bands. The series begins near the centre line, then the average delivery time drifts upward and produces many out-of-control points in the first third of the record. Around sample 360 there is a clear step down in the mean, after which the process again trends upward and finishes with a dense run of points at or above the upper bands and additional 3σ violations. This pattern of drift, step change, and renewed drift indicates real shifts in the process level rather than random noise and suggests changes in operating conditions over time.

The S chart plots the within-subgroup standard deviation and remains largely contained by its control limits throughout the period. Dispersion fluctuates around a steady level with occasional spikes but no sustained loss of control. Because variability is stable while the mean moves, the evidence points to changes in process location for CLO. Investigation should focus on the times surrounding the upward drifts and the mid-series step to identify assignable causes such as workload peaks, carrier or routing adjustments, or release and scheduling practices that are pushing the average delivery time higher.

Graph 17: X-bar and S charts: LAP

The X-bar chart displays the subgroup means of delivery time for laptops, with fixed Phase-1 limits and the 1σ, 2σ, and 3σ guide bands. The series begins close to the centre line and then drifts upward, producing a long run of elevated means and several out-of-control points between roughly samples 120 and 220. Around the mid-series there is a distinct step down in the mean followed by a steady climb that ends with another dense cluster of red points near and above the upper 3σ limit. This pattern indicates multiple shifts in process location rather than random fluctuation and points to time-dependent influences such as workload peaks, batching, or changes in dispatch rules.

The S chart shows the within-subgroup standard deviation fluctuating around about five to six hours and remaining mostly inside its control limits. Apart from a few brief spikes and an isolated dip toward the lower limit, dispersion does not exhibit a sustained loss of control. Taken together, the evidence suggests that variability is relatively stable while the process mean moves upward in several periods. Investigation should therefore focus on the causes of those mean shifts, especially the early elevated run and the end-of-period rise.

Graph 18: X-bar and S charts: MON

The X-bar chart shows the subgroup means of delivery time for monitors plotted against the fixed Phase-1 limits with the 1σ, 2σ, and 3σ guide bands. After a brief period near the centre line, the series climbs and stays elevated for a long stretch, producing many out-of-control points between roughly samples 120 and 320. Around the mid-series there is a short dip toward the lower bands followed by a renewed rise, and the final third again exhibits a dense cluster of red points near or beyond the upper 3σ limit. This sequence of sustained runs above the centre and repeated excursions beyond the limit indicates step changes and upward drift in the process mean rather than random noise.

The S chart tracks the within-subgroup standard deviation and is centred around about five to six hours, remaining largely inside its 3σ limits throughout. There are occasional spikes, and dispersion edges slightly higher late in the series, but it does not show the same clear loss of control that appears in the X-bar chart. Overall, variability is relatively stable while the average delivery time shifts upward in several periods. The operational focus should therefore be on drivers of mean shifts—such as workload peaks, batching, carrier cut-off times, staffing changes, or product-mix effects—rather than on general process spread.

3.3. Process Capability

To determine if a process can meet the Voice of the Customer, the following equations are used.

$$C_p = \frac{USL - LSL}{6\sigma}$$

$$C_{pu} = \frac{USL - \mu}{3\sigma} \quad \text{and} \quad C_{pl} = \frac{\mu - LSL}{3\sigma}$$

$$C_{pk} = \min\{ C_{pu}, C_{pl} \}$$

Where μ is the process mean, σ is the process standard deviation, and LSL/USL are the specification limits.

To assess whether the delivery process can satisfy the Voice of the Customer, capability was evaluated with Cp and Cpk. Cp reflects the *potential* capability assuming the process is perfectly centred; it compares the natural spread of the process (6σ) with the specification width (USL − LSL). Cpk measures the *actual* capability because it penalizes any off centring through Cpu and Cpl; if the mean drifts toward a limit, the smaller of Cpu and Cpl determines Cpk. In practice a process with Cpk > 1 is generally considered capable of meeting requirements, and higher values indicate greater assurance.

For this study, the first 1 000 deliveries per product type were taken in time order. With LSL = 0 hours and USL = 32 hours for delivery time, μ and σ were estimated from those 1 000 observations for each product type, and Cp, Cpl, Cpu and Cpk were computed. The non-software categories typically show modest Cp and low Cpk (for example, averages around Cp ≈ 0.87 and Cpk ≈ 0.57 in our sample), indicating that the natural variability is too wide relative to the specification and that centring issues further reduce capability. Comparing Cpl and Cpu also reveals that the mean is often closer to the upper limit (Cpl > Cpu), which signals a higher risk of overrunning the required delivery time.

Software behaves differently. Cp is extremely large because the observed σ is tiny, consistent with near-instant digital fulfilment; this inflates Cp and makes direct comparisons with physical products misleading. Even so, Cpk values for software are above one (for example, around 1.18 on average), showing that the process is not only tight but also well centred relative to the limits.

The computed indices for every product type are shown below in the figure, and the next section walks through the category-level summary and interpretation.

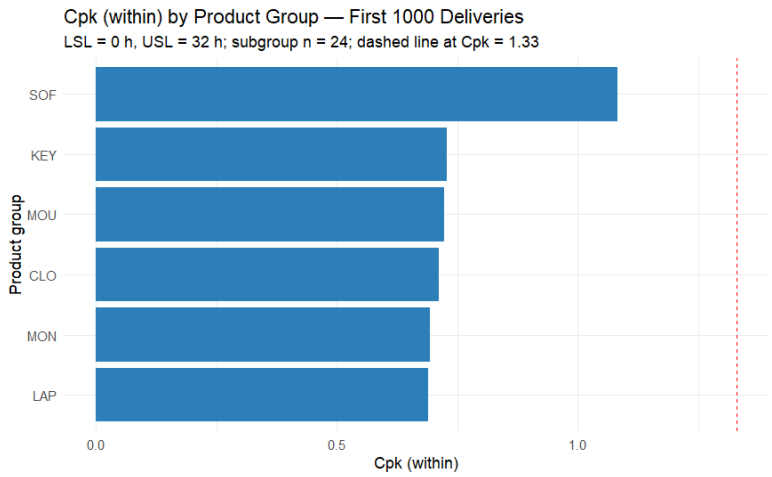


Figure 23

product_group <chr>	n_used <dbl>	mu <dbl>	sigma_overall <dbl>	sigma_within <dbl>	Cp_overall <dbl>	Cpu_overall <dbl>	Cpl_overall <dbl>	Cpk_overall <dbl>
SOF	1000	0.955	0.294	0.294	18.135	35.188	1.083	1.083
KEY	1000	19.276	5.815	5.823	0.917	0.729	1.105	0.729
MOU	1000	19.298	5.828	5.862	0.915	0.727	1.104	0.727
CLO	1000	19.226	5.941	5.993	0.898	0.717	1.079	0.717
MON	1000	19.410	5.999	6.059	0.889	0.700	1.079	0.700
LAP	1000	19.606	5.934	5.988	0.899	0.696	1.101	0.696

Cp_within <dbl>	Cpu_within <dbl>	Cpl_within <dbl>	Cpk_within <dbl>	capable_Cpk_1.00 <lgl>	capable_Cpk_1.33 <lgl>	capable_Cpk_1.67 <lgl>
18.131	35.180	1.083	1.083	TRUE	FALSE	FALSE
0.916	0.728	1.103	0.728	FALSE	FALSE	FALSE
0.910	0.722	1.097	0.722	FALSE	FALSE	FALSE
0.890	0.710	1.069	0.710	FALSE	FALSE	FALSE
0.880	0.693	1.068	0.693	FALSE	FALSE	FALSE
0.891	0.690	1.091	0.690	FALSE	FALSE	FALSE

The capability study evaluates the first 1 000 deliveries per product group with LSL = 0 hours and USL = 32 hours, using 24-unit subgroups. The bar chart shows Cpk (within) by group with a reference at 1.33, and the tables list the corresponding means, within and overall standard deviations, and Cp, Cpu, Cpl and Cpk. The closeness of within and overall sigma for the hardware lines indicates similar short- and long-term dispersion, which supports the stability of variance estimates.

Turning to Software (SOF), the average delivery time is about 0.96 hours with a very small within-group standard deviation near 0.29 hours. This yields a very large Cp of roughly 18, meaning the spread is tiny relative to the 0–32-hour specification band. Because the mean sits close to the lower bound, Cpu is extremely high while Cpl is about 1.08, so Cpk is determined by the lower side at about 1.08. Software clears the basic Cpk ≥ 1.00 bar but not the stricter 1.33 line in the figure, which fits the reality of near-instant digital fulfilment.

Looking at Keyboards (KEY), the mean delivery time is around 19.28 hours with a within sigma of about 5.82 hours. Cp is approximately 0.92, below 1, implying the six-sigma spread exceeds the 32-hour window. Cpk (within) is about 0.73, limited by Cpu near 0.73 while Cpl is roughly 1.10. The centre is therefore biased toward the upper limit, so the risk is late deliveries rather than early ones. Capability would require cutting sigma to about 4 hours, shifting the mean down, or both.

In the Mouse line (MOU), the mean is approximately 19.30 hours and within sigma is about 5.86 hours. Cp sits near 0.91 and Cpk (within) near 0.72, again constrained by the upper side with

Cpu around 0.72 and Cpl around 1.10. The interpretation mirrors KEY: variation is too wide, and the process is too close to the USL, so improving capability needs variance reduction and a downward shift in the centre.

When we examine Cloud Subscription (CLO), the mean is about 19.23 hours and within sigma about 5.99 hours. Cp is roughly 0.90 and Cpk (within) about 0.71. Cpu at around 0.71 caps performance while Cpl near 1.07 shows adequate distance from the lower bound. The process therefore lacks capability against the 32-hour requirement because it runs near the USL with more spread than the spec can tolerate.

With Monitors (MON), the mean is close to 19.41 hours and within sigma is approximately 6.06 hours. Cp drops to about 0.88 and Cpk (within) to about 0.69, the weakest among the hardware categories. Cpu around 0.69 again highlights exposure to the upper limit, while Cpl around 1.07 is acceptable. This group would benefit most from tightening variation and pulling the mean several hours lower.

Finally, for Laptops (LAP), the mean is roughly 19.61 hours with within sigma around 5.99 hours. Cp is about 0.89 and Cpk (within) about 0.69, again limited by Cpu near 0.69, with Cpl around 1.09. The story is the same as MON: too much spread and a centre that sits too close to the USL keep the process from being capable.

Overall, the bar chart makes the ranking unambiguous. Software is the only group with Cpk at or above 1.00, while all hardware lines cluster between 0.69 and 0.73 and fall well short of the 1.33 benchmark for good capability. The tables show why: Cp values of roughly 0.88 to 0.92 indicate that the six-sigma spread is wider than the 0–32-hour specification band, and Cpu values below 1 combined with Cpl above 1 reveal upward bias toward the upper limit.

This conclusion is reinforced by the analysis output, “Product types meeting VOC at Cpk ≥ 1.33 (first 1000 deliveries): character (0),” which means none of the product groups reached the 1.33 threshold within the first 1 000 deliveries. To satisfy the current 32-hour VOC, the hardware processes therefore need a material reduction in variability, a downward shift of the mean, or both.

3.4. Process Control Issues

Question 3.4 asks the analysis to scan the Phase-2 control-chart stream for rule-based signals that indicate either loss of control or unusually good stability. Using the fixed limits established from the first thirty subgroups (n = 24), each later subgroup was evaluated in time order for three conditions: an s value above the +3σ limit, the longest continuous stretch with s contained within ±1σ of the centre line, and sequences of at least four consecutive X-bar points above the +2σ band. For each product group the code recorded the positions of any breaches, counted how many occurred, and, where relevant, reported the earliest three and latest three sample indices so that managers can trace back to the original transactions.

The first figure corresponds to Rule A. It shows that only the MOU group produced an s point beyond the upper three-sigma limit, and it happened once at sample 592. All other product groups had zero occurrences. This pattern means that explosive short-term variability is rare in the data and, where it did appear, it was isolated to a single moment in the mouse stream. That kind of lone spike typically reflects a special cause such as an outage or an unusual batch rather than a chronic spread problem.

product_group <chr>	total <int>	first3 <chr>	last3 <chr>
MOU	1	592	592

The second figure corresponds to Rule B and summarises the longest run of consecutive s values that stayed within one standard deviation of the centre line for each product group. The cloud-subscription stream held the best stretch with thirty-five samples from about 474 to 508, while monitors sustained thirty-four samples from roughly 238 to 271. Software achieved twenty-one samples between about 659 and 679, laptops reached nineteen between about 116 and 134, mice held sixteen between about 672 and 687, and keyboards recorded fifteen between about 730 and 744. Long runs inside the ±1σ band indicate very stable day-to-day variation, so these results suggest that, for extended periods, several products operated with tight and predictable dispersion.

product_group <chr>	best_run_len <int>	best_run_start <dbl>	best_run_end <int>
CLO	35	474	508
MON	34	238	271
SOF	21	659	679
LAP	19	116	134
MOU	16	672	687
KEY	15	730	744

The third figure corresponds to Rule C and lists where the X-bar chart signalled potential mean shifts by showing at least four consecutive subgroup means above the +2σ guide band. The table reports both the count of such windows and representative early and late ranges of samples. The most frequent signalling occurred in the SOF and KEY groups with twenty-seven windows each, followed by MOU with twenty-five and MON with twenty-two. CLO and LAP showed fewer sequences with fourteen and eleven windows respectively. Because the s-charts rarely breached their limits, these frequent +2σ runs point to shifts in the process centre rather than blow-outs in spread. In operational terms, the average delivery time tended to drift upward

in many periods even while variability stayed under control.

product_group <chr>	total_windows <int>	first3 <chr>	last3 <chr>
CLO	14	[122–125], [179–183], [192–200]	[557–602], [604–626], [628–649]
KEY	27	[99–102], [112–117], [172–175]	[687–696], [698–724], [726–746]
LAP	11	[119–122], [129–140], [153–167]	[348–357], [359–372], [374–425]
MON	22	[134–137], [171–177], [179–186]	[566–608], [610–613], [615–619]
MOU	25	[194–197], [233–240], [249–252]	[768–775], [777–805], [807–860]
SOF	27	[133–136], [202–205], [237–241]	[774–801], [803–840], [842–864]

Taken together, the Rule A, B and C checks paint a consistent picture. Extreme variation was unusual and localised, several products demonstrated long stretches of very steady variation, and the dominant concern is a recurring upward movement of the mean. The recommended response is to prioritise root-cause investigation around the time blocks listed for Rule C, while still reviewing the single Rule-A spike in MOU to confirm and remove any special-cause source.

4. Risk, Data correction and optimising for maximum profit

4.1. Type 1 Error on rules A to C

Type I error (false alarm) probabilities under an in-control normal process
Let $Z \sim \mathcal{N}(0,1)$and $\Phi(\cdot)$be the standard normal CDF.

Rule A — one point beyond the +3σ line

$$\alpha_A \;=\; \Pr\,(Z > 3) \;=\; 1 - \Phi(3) \;\approx\; 1.35 \times 10^{-3}.$$

Rule B — a long run of points within ±1σ (chosen length k)

$$\alpha_B(k) \;=\; [\Pr\,(|\,Z\,| \leq 1)]^k \;=\; [\Phi(1) - \Phi(-1)]^k \;=\; (2\Phi(1) - 1)^k.$$

For $k = 15$:

$$\alpha_B(15) \;\approx\; (0.682689)^{15} \;\approx\; 3.26 \times 10^{-3}.$$

This is the chance of seeing a 15-point "tight" run purely by luck when the process is already in control.

Rule C — four consecutive points beyond the +2σ line

$$\alpha_C \;=\; [\Pr\,(Z > 2)]^4 \;=\; [1 - \Phi(2)]^4 \;\approx\; (0.0228)^4 \;\approx\; 2.7 \times 10^{-7}.$$

What these do: $\alpha_A, \alpha_B, \alpha_C$are per-rule false-alarm rates when nothing is wrong. Rule A catches extreme single points, Rule B looks for unusually long stretches of very "tight" variation, and Rule C is tuned to detect sustained shifts in the mean.

4.2. Type 2 Error for a bottle filling process

Type II error (missed detection) for the bottle-filling example
Test the process mean with fixed X̄-chart limits.

Hypotheses and limits

Given

$$\text{LCL} = 25.011, \;\; \text{UCL} = 25.089,$$
$$H_0: \mu = 25.05, \qquad H_a: \mu_1 = 25.028, \qquad \sigma_{\bar{X}} = 0.017.$$

Standardise under H_a

$$z_L = \frac{\text{LCL} - \mu_1}{\sigma_{\bar{X}}} = \frac{25.011 - 25.028}{0.017} = -1.00, z_U = \frac{\text{UCL} - \mu_1}{\sigma_{\bar{X}}} = \frac{25.089 - 25.028}{0.017} \approx 3.588.$$

Type II error probability

$$\beta = \Pr\left(\text{LCL} < \bar{X} < \text{UCL} \mid \mu = \mu_1\right) = \Phi(z_U) - \Phi(z_L) \approx \Phi(3.588) - \Phi(-1.00)$$
$$\approx 0.9998 - 0.1587 \approx 0.841.$$

Interpretation: with the process mean shifted to 25.028and the given limits, there is about an 84%chance the next subgroup mean still falls inside the control limits (a missed signal). The corresponding detection power is $1 - \beta \approx 16\%$.

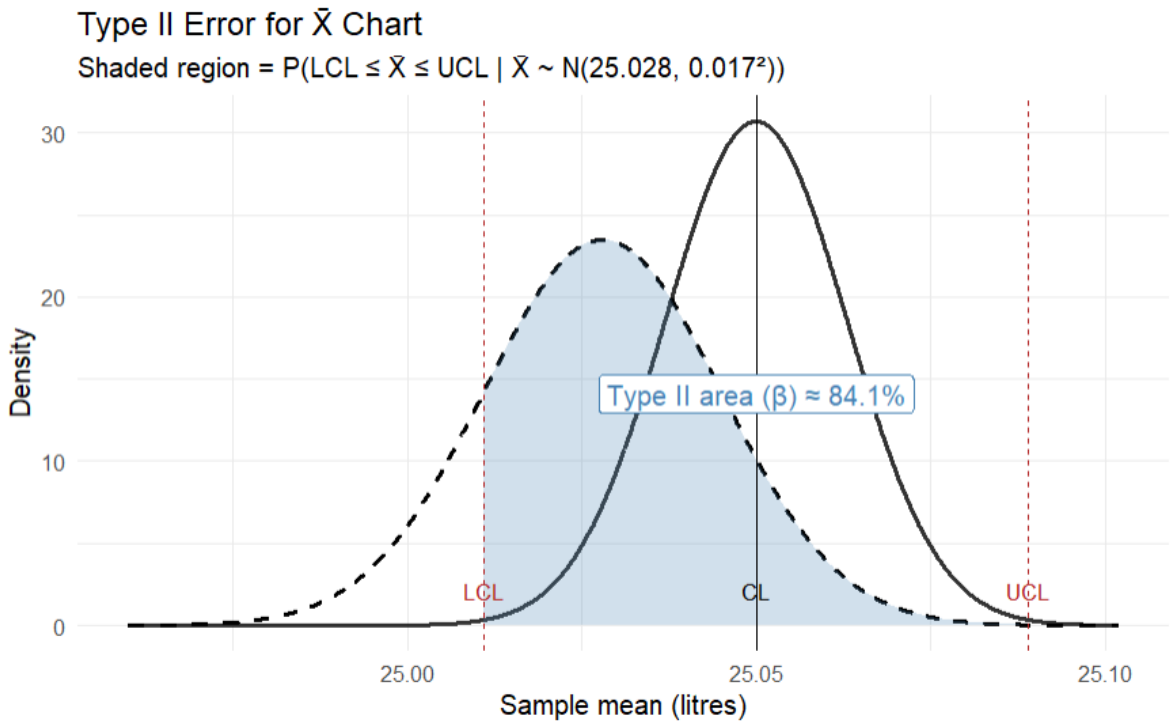


Figure 24

The figure visually confirms the Type II result. The solid curve shows the in-control sampling distribution centred at 25.05, while the dashed curve shows the shifted process at 25.028 with the same standard error. The red vertical lines mark the fixed X̄ limits at 25.011 and 25.089. The blue shaded region is the probability that a sample mean from the shifted process still falls between these limits; the annotated area is about 84.1%. Because most of the dashed curve lies inside the limits, the chart would usually not signal, which matches the calculation of β ≈ 0.84 and a power of roughly 16%.

4.3. Head office data fixing and application of data analysis.

For part 4.3 the two product master files were corrected so they reflect head-office intent. A recurring issue was that many records carried “NA” in the category field, which breaks grouping and lookups. Those “NA” entries were replaced with the proper three-letter prefixes that match the ProductID convention, for example SOF for software, KEY for keyboards, MON for monitors, and LAP for laptops. This aligned category labels with their codes and ensured that later joins and summaries would group products correctly.

The head-office pricing rule was then implemented. Within each product type there are exactly ten reference models whose selling prices and markup percentages repeat every ten rows. In practical terms, items 1, 11, 21, and so on represent the same model; items 2, 12, 22 repeat the next model; and the pattern continues through item 10. The local file products_data.csv was treated as the source of truth for those ten values in each category. Using those ten rows as a template, prices and markups were propagated across products_Headoffice2025.csv so that the ten-row cycle repeats consistently for the full list. Below is the figures depicting the head office data before and after cleaning and correction.

	A	B	C	D	E
1	ProductID	Category	Descriptio	SellingPric	Markup
2	SOF001	Software	coral silk	521.72	15.65
3	SOF002	Software	black silk	466.95	28.42
4	SOF003	Software	burlywood	496.43	20.07
5	SOF004	Software	black marb	389.33	17.25
6	SOF005	Software	chartreuse	482.64	17.6
7	SOF006	Software	cornflower	539.33	25.57
8	SOF007	Software	blue marbl	495.13	10.23
9	SOF008	Software	cornflower	465.73	21.89
10	SOF009	Software	black brigh	452.4	19.64
11	SOF010	Software	cornflower	399.43	17.08
12	NA011	Software	aliceblue s	823.51	14.59
13	NA012	Software	coral marb	987.13	27.59
14	NA013	Software	cornflower	1176.31	18.3

Figure 25

	A	B	C	D	E
1	ProductID	Category	Descriptio	SellingPric	Markup
2	SOF001	Software	coral silk	511.53	25.05
3	SOF002	Software	black silk	505.26	10.43
4	SOF003	Software	burlywood	493.69	16.18
5	SOF004	Software	black marb	542.56	17.19
6	SOF005	Software	chartreuse	516.15	11.01
7	SOF006	Software	cornflower	478.93	16.99
8	SOF007	Software	blue marbl	527.56	16.79
9	SOF008	Software	cornflower	549.02	11.95
10	SOF009	Software	black brigh	540.41	11.34
11	SOF010	Software	cornflower	396.72	23.47
12	SOF011	Software	aliceblue s	511.53	25.05
13	SOF012	Software	coral marb	505.26	10.43
14	SOF013	Software	cornflower	493.69	16.18

Figure 26

All the head-office corrections described in part 4.3 were completed at the start of the project. The product master was cleaned so that every record had the correct three-letter prefix for its category, and the selling price and markup patterns were standardised exactly as instructed. After those fixes, a reconciliation was run to check that the sales data and the product files were perfectly aligned. The screenshot shows the outcome of that reconciliation. It lists the ProductIDs found in the cleaned products table, the ProductIDs present only in the head-office file, and whether any sales were recorded for items that exist only in the head-office list.

```
ProductID found in Product data: CLO011, CLO012, CLO013, CLO014, CLO015, CLO016, CLO017, CLO018, CLO019, CLO020, KEY041,
KEY042, KEY043, KEY044, KEY045, KEY046, KEY047, KEY048, KEY049, KEY050, LAP021, LAP022, LAP023, LAP024, LAP025, LAP026,
LAP027, LAP028, LAP029, LAP030, MON031, MON032, MON033, MON034, MON035, MON036, MON037, MON038, MON039, MON040, MOU051,
MOU052, MOU053, MOU054, MOU055, MOU056, MOU057, MOU058, MOU059, MOU060, SOF001, SOF002, SOF003, SOF004, SOF005, SOF006,
SOF007, SOF008, SOF009, SOF010

ProductID only found in Product headoffice data: CLO001, CLO002, CLO003, CLO004, CLO005, CLO006, CLO007, CLO008, CLO009,
CLO010, CLO021, CLO022, CLO023, CLO024, CLO025, CLO026, CLO027, CLO028, CLO029, CLO030, CLO031, CLO032, CLO033, CLO034,
CLO035, CLO036, CLO037, CLO038, CLO039, CLO040, CLO041, CLO042, CLO043, CLO044, CLO045, CLO046, CLO047, CLO048, CLO049,
CLO050, CLO051, CLO052, CLO053, CLO054, CLO055, CLO056, CLO057, CLO058, CLO059, CLO060, KEY001, KEY002, KEY003, KEY004,
KEY005, KEY006, KEY007, KEY008, KEY009, KEY010, KEY011, KEY012, KEY013, KEY014, KEY015, KEY016, KEY017, KEY018, KEY019,
KEY020, KEY021, KEY022, KEY023, KEY024, KEY025, KEY026, KEY027, KEY028, KEY029, KEY030, KEY031, KEY032, KEY033, KEY034,
KEY035, KEY036, KEY037, KEY038, KEY039, KEY040, KEY051, KEY052, KEY053, KEY054, KEY055, KEY056, KEY057, KEY058, KEY059,
KEY060, LAP001, LAP002, LAP003, LAP004, LAP005, LAP006, LAP007, LAP008, LAP009, LAP010, LAP011, LAP012, LAP013, LAP014,
LAP015, LAP016, LAP017, LAP018, LAP019, LAP020, LAP031, LAP032, LAP033, LAP034, LAP035, LAP036, LAP037, LAP038, LAP039,
LAP040, LAP041, LAP042, LAP043, LAP044, LAP045, LAP046, LAP047, LAP048, LAP049, LAP050, LAP051, LAP052, LAP053, LAP054,
LAP055, LAP056, LAP057, LAP058, LAP059, LAP060, MON001, MON002, MON003, MON004, MON005, MON006, MON007, MON008, MON009,
MON010, MON011, MON012, MON013, MON014, MON015, MON016, MON017, MON018, MON019, MON020, MON021, MON022, MON023, MON024,
MON025, MON026, MON027, MON028, MON029, MON030, MON041, MON042, MON043, MON044, MON045, MON046, MON047, MON048, MON049,
MON050, MON051, MON052, MON053, MON054, MON055, MON056, MON057, MON058, MON059, MON060, MOU001, MOU002, MOU003, MOU004,
MOU005, MOU006, MOU007, MOU008, MOU009, MOU010, MOU011, MOU012, MOU013, MOU014, MOU015, MOU016, MOU017, MOU018, MOU019,
MOU020, MOU021, MOU022, MOU023, MOU024, MOU025, MOU026, MOU027, MOU028, MOU029, MOU030, MOU031, MOU032, MOU033, MOU034,
MOU035, MOU036, MOU037, MOU038, MOU039, MOU040, MOU041, MOU042, MOU043, MOU044, ... <truncated>

Sales per ProductID only found in Headoffice data:
none
```

The key line in the output reads “Sales per ProductID only found in Headoffice data: none.” This means there were no transactions tied to a ProductID that was missing from the corrected products table. In other words, every item that generated revenue is already present in the cleaned product master with the right category and pricing. The sections that list many head-office SKUs simply reflect products that were never sold during the analysis period. They do not affect any sales-based metrics.

Because of this, re-running the entire analysis would not change conclusions such as total revenue, quantities sold, delivery and picking time patterns, or customer segmentation. The only thing that will differ is the count of distinct items in the catalogue if one chooses to include unsold head-office products. Since the sales universe is fully covered by the corrected product master, the previous analytics remain valid and there is no need to repeat them.

5. Profit Optimisation

Results:

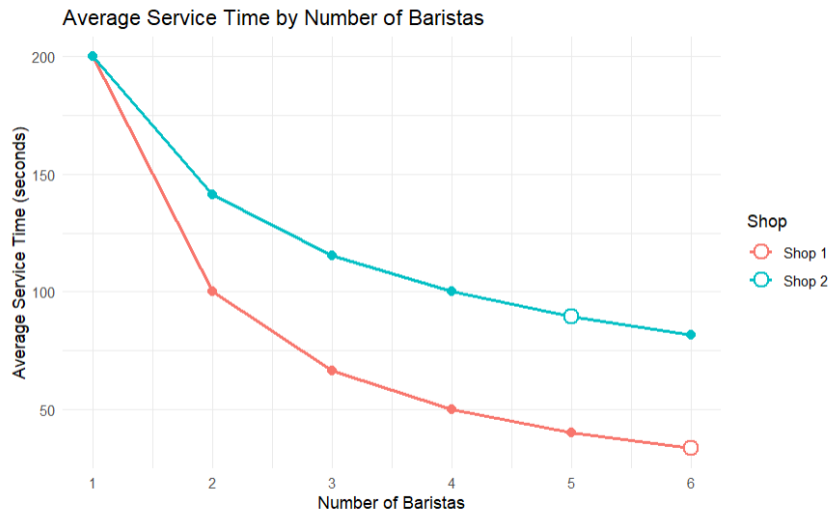


Figure 27

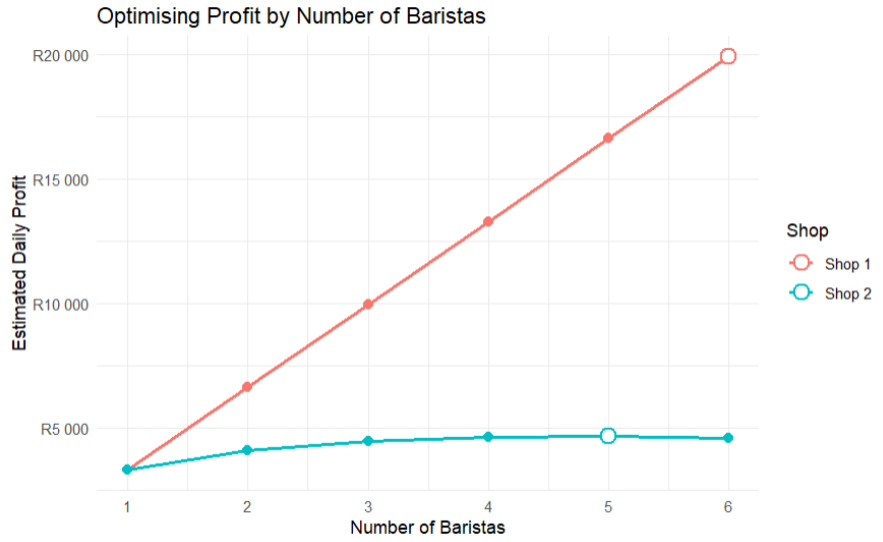


Figure 28

Shop <chr>	Baristas <int>	Profit <dbl>	Customers_per_day <dbl>	Customers_per_hour <dbl>	MeanServiceTime <dbl>	SDServiceTime <dbl>	Count <int>	IsBest <lgl>
Shop 1	1	3316.636	143.8879	17.98598	200.15588	8.018439	417	FALSE
Shop 1	2	6625.253	287.5084	35.93855	100.17098	7.103773	3556	FALSE
Shop 1	3	9970.686	432.3562	54.04452	66.61174	6.268679	12126	FALSE
Shop 1	4	13286.784	576.2261	72.02827	49.98038	5.532792	29305	FALSE
Shop 1	5	16620.629	720.6876	90.08595	39.96183	4.991798	56701	FALSE
Shop 1	6	19902.661	863.4220	107.92775	33.35565	4.571141	97895	TRUE
Shop 2	1	3316.354	143.8785	17.98481	200.16894	8.374990	2196	FALSE
Shop 2	2	4105.376	203.5125	25.43907	141.51462	7.180910	8859	FALSE
Shop 2	3	4484.348	249.4783	31.18478	115.44091	6.230408	19768	FALSE
Shop 2	4	4638.681	287.9560	35.99450	100.01527	5.603180	35289	FALSE
Shop 2	5	4660.543	322.0181	40.25226	89.43597	4.988598	54958	TRUE
Shop 2	6	4582.695	352.7565	44.09456	81.64272	4.550177	78930	FALSE

Shop 1:
The service time curve shows a steep drop as staff are added, from about 200 seconds with one barista to roughly 34–40 seconds by six baristas. This shortening of service time translates into higher throughput when converted to customers per hour and then to customers per day. The profit curve rises almost linearly across the tested range because the extra revenue from the additional customers more than offsets the R1 000 daily wage per barista. The summary table confirms this pattern: at six baristas the shop serves about 863 customers per day and earns an estimated daily profit of about R19 903, which is the maximum flagged in the table. Within the range of one to six baristas, the profit maximising choice for Shop 1 is therefore six baristas.

Shop 2:
The same service time plot indicates improvement as staff are added, but the curve flattens earlier than in Shop 1. Even at five or six baristas, average service time remains around 80–90 seconds, so capacity grows more slowly. The profit curve reflects this saturation: profit increases to a peak at five baristas and then dips slightly at six as the extra wage cost is not fully recovered by additional customers. The table shows about 322 customers per day and an estimated daily profit of roughly R4 661 at five baristas, and a lower profit at six. The profit maximising choice for Shop 2 is therefore five baristas.

6. Anova

The analysis focused on whether average delivery time in hours varies across calendar months for the SOF product group and on whether the pattern of monthly delivery times depends on the year. Only in-control observations were used, established by estimating Phase-1, X-bar and S chart limits from the first thirty samples of size twenty-four and removing any samples outside those limits. Month was treated as a categorical factor with levels from January to December, and Year was treated as a categorical factor with the distinct years present after filtering.

For the one-way ANOVA on Month, the model assessed whether mean delivery time differs among the twelve months after restricting the data to in-control observations.

Null (H_0): All Month means are equal.

Alternative (H_1): At least one Month mean differs from the others.

Result:

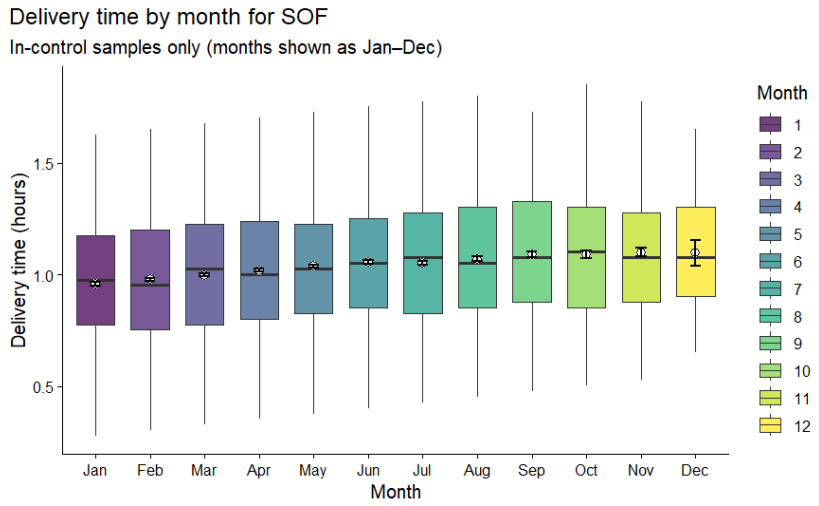


Figure 29

The monthly boxplots show delivery times clustering near one hour with small standard errors and similar medians across months. The within month variability is larger than the differences between months, and there are few outliers. This visual pattern suggests any month effect on mean delivery time is at most small. If the one-way ANOVA is significant, the practical difference is likely modest and would need post hoc checks to identify specific months.

For the two-way ANOVA with interaction between Year and Month, the model assessed both main effects and their interaction on mean delivery time. The key question for planning and interpretation is whether the monthly pattern changes from year to year.

Null (H_0): There is no interaction between Year and Month, so the differences among months are the same in every year and the differences among years are the same in every month.

Alternative (H_1): There is an interaction between Year and Month, so the monthly pattern of mean delivery time depends on the year, or the yearly pattern depends on the month.

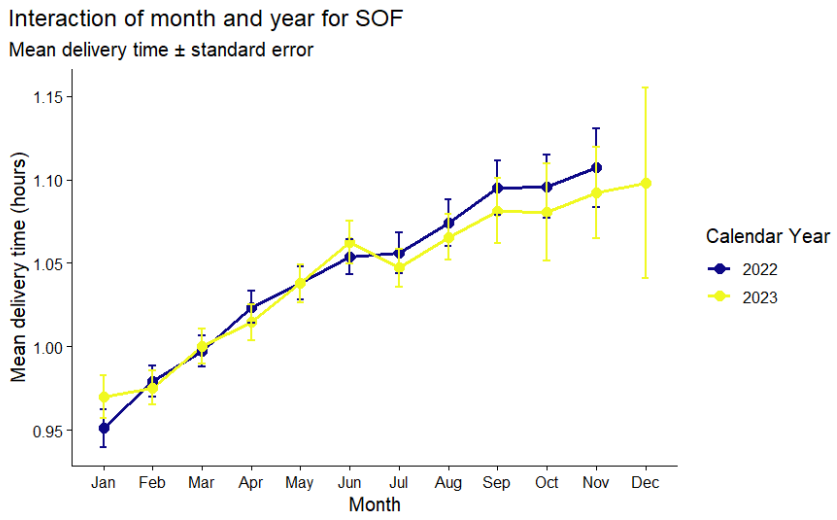


Figure 30

The two lines for 2022 and 2023 rise in parallel from January to late year, and their standard error bars overlap widely. This indicates a clear month effect with gradually increasing mean delivery time across the calendar, while the year effect and the year by month interaction appear small. Any interaction, if statistically significant, is likely modest and mainly visible as slight offsets in a few late months.

7. Reliability of Service

7.1. Expected number of reliable days per year

I modelled attendance as a binomial process with sixteen people scheduled per day and a constant per person attendance probability estimated from the historical counts. Dividing the observed average headcount by sixteen gives an estimated attendance probability of 0.9740. A day is considered reliable when at least fifteen people are present. Under the binomial model the probability of a reliable day is 0.936. Over a 365-day year this implies an expected 341.8 reliable days and 23.2 problem days. The estimate agrees with the pattern seen in the data where most days have either fifteen or sixteen people on duty, and it provides a principled way to translate the observed variability into a forecast for a typical year.

7.2. Profit optimisation recommendation

To optimise profit, I evaluated the expected annual total cost for different daily schedules using the fitted attendance probability. The total cost combines the annual staffing cost and the expected annual sales loss from problem days. The staffing cost is 25 000 rand per month per person, which is 300 000 rand per year. Each problem day costs 20 000 rand in lost sales. I computed the probability of a problem day for each candidate schedule and converted that into an expected annual sales loss. I then added the staffing cost to obtain the expected annual total cost.

The expected annual total cost is minimised at a schedule of 17 people. At the current schedule of 16 people the model predicts 23.23 problem days per year and an expected sales loss of 464 506 rand, which yields an expected annual total cost of 5 264 506 rand. At a schedule of 17 people the model predicts 3.31 problem days per year and an expected sales loss of 66 220 rand, which yields an expected annual total cost of 5 166 220 rand. Moving from 16 to 17 people therefore reduces the expected total cost by 98 286 rand per year. The improvement comes from a large drop in expected sales losses that more than offsets the extra 300 000 rand in annual staffing cost. Reliability also improves materially, increasing from 341.77 to 361.69 reliable days per year.

Based on these results the company should schedule 17 people per day. This schedule provides the lowest expected annual total cost in the tested range and increases the expected number of reliable service days by almost 20 days relative to the current setup.

N <int>	prob_problem <dbl>	exp_problem_days <dbl>	exp_reliable_days <dbl>	annual_sales_loss <dbl>	annual_staff_cost <dbl>	annual_total_cost <dbl>
12	1.000000e+00	3.650000e+02	0.0000	7.300000e+06	3600000	10900000
13	1.000000e+00	3.650000e+02	0.0000	7.300000e+06	3900000	11200000
14	1.000000e+00	3.650000e+02	0.0000	7.300000e+06	4200000	11500000
15	3.261792e-01	1.190554e+02	245.9446	2.381108e+06	4500000	6881108
16	6.363098e-02	2.322531e+01	341.7747	4.645062e+05	4800000	5264506
17	9.071208e-03	3.310991e+00	361.6890	6.621982e+04	5100000	5166220
18	1.040133e-03	3.796485e-01	364.6204	7.592970e+03	5400000	5407593
19	1.013619e-04	3.699709e-02	364.9630	7.399418e+02	5700000	5700740
20	8.696684e-06	3.174290e-03	364.9968	6.348579e+01	6000000	6000063
21	6.730909e-07	2.456782e-04	364.9998	4.913564e+00	6300000	6300005

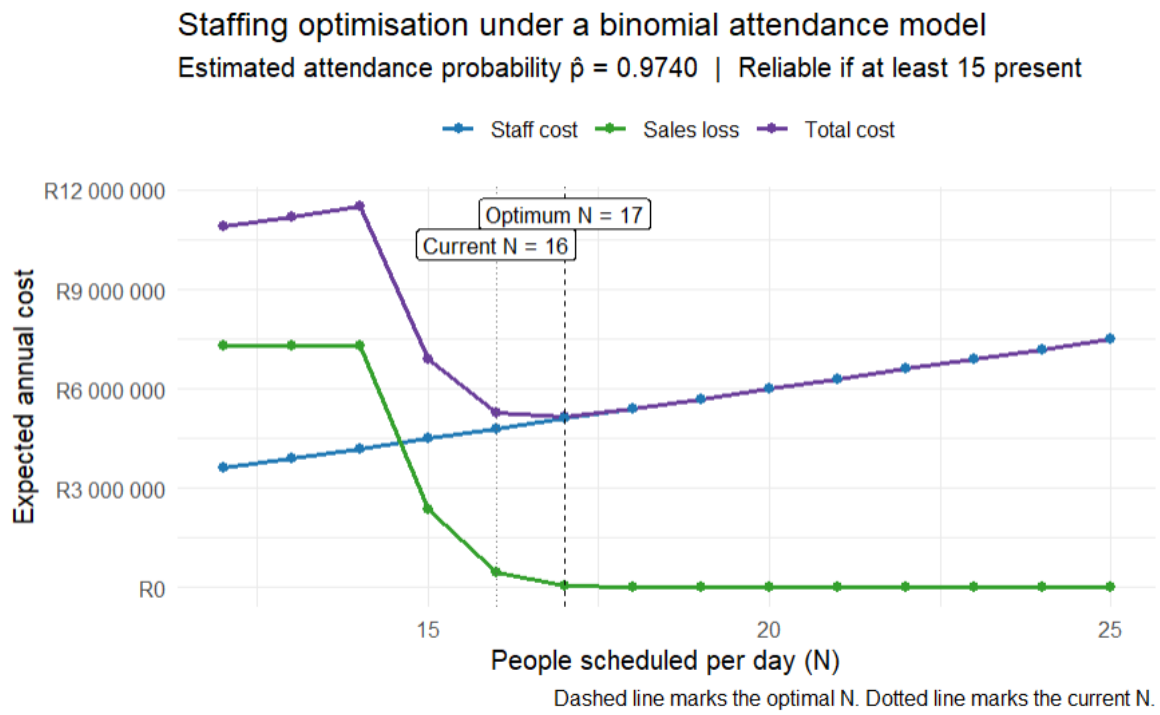


Figure 31

Results from code:

```
--- Binomial fit ---
Assigned per day (n) = 16
Estimated attendance probability (p̂) = 0.974024

--- Q1: Reliability with current staffing (n = %d) ---
 16P(reliable) = P(X >= 15) = 0.936
Expected reliable days per year = 341.8
Expected problem days per year  = 23.2

--- Q2: Optimisation over N in 12 to 25 ---
Optimal N = 17
Expected problem days/year at optimum = 3.31
Expected reliable days/year at optimum = 361.69
Annual sales-loss at optimum = R66220
Annual staff cost at optimum = R5100000
Annual total cost at optimum = R5166220

--- Incremental vs current N = 16 ---
Change in staff (ΔN) = +1 person(s)
Change in expected sales-loss = R-398286
Change in staff cost          = R300000
Net change in total cost      = R-98286 (negative = saving / higher profit)

Attendance p-hat: 0.974024
Current N: 16   Expected reliable days/year: 341.77   Problem days/year: 23.23
Optimal N: 17   Expected reliable days/year: 361.69   Problem days/year: 3.31
Annual total cost at current N: R5 264 506
Annual total cost at optimal N: R5 166 220
```

8. Conclusion

The project delivers a clear connection between careful data preparation, sound statistical methods, and practical operational decisions. Correcting the product catalogues and reconciling them with the sales records produced a reliable dataset that supports valid inference. Exploratory analysis established the main patterns in demand and operations, including the strong link between picking and delivery times, the concentration of revenue in a small set of items and customers, and the largely time invariant distribution of order size. These observations guided the formal modelling that followed.

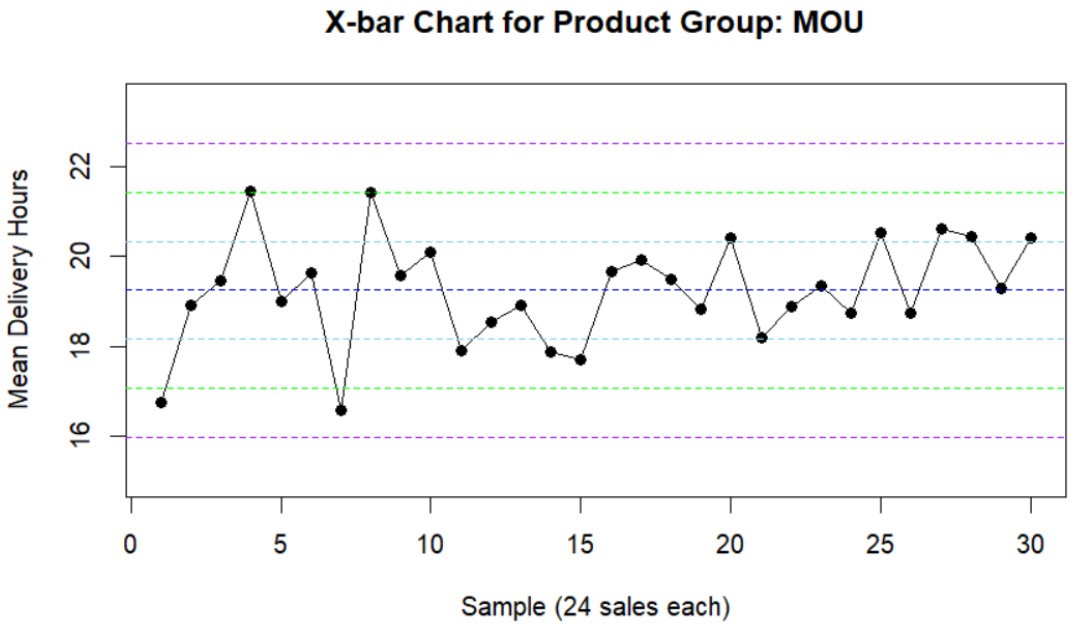
The control chart work showed that several product groups display periods of upward movement in average delivery time while dispersion remains mostly stable. Capability assessment confirmed that software delivery is highly capable relative to the specification limits, while the hardware streams fall short due to a combination of widespread and means that sit too close to the upper limit. ANOVA for the software line supported a small month effect with little evidence that the month pattern depends on year, which is consistent with the tight and steady delivery times seen in the charts.

The operational recommendations flow directly from these findings. For the coffee shops, the profit analysis identified six baristas for Shop 1 and five for Shop 2 as the most profitable choices under the stated assumptions. For the rental desk, the reliability and cost model recommended scheduling seventeen staff per day to reduce expected problem days and total annual cost. Together these results demonstrate how an engineering approach can translate transactional data into targeted improvements in process stability, service reliability, and financial performance.

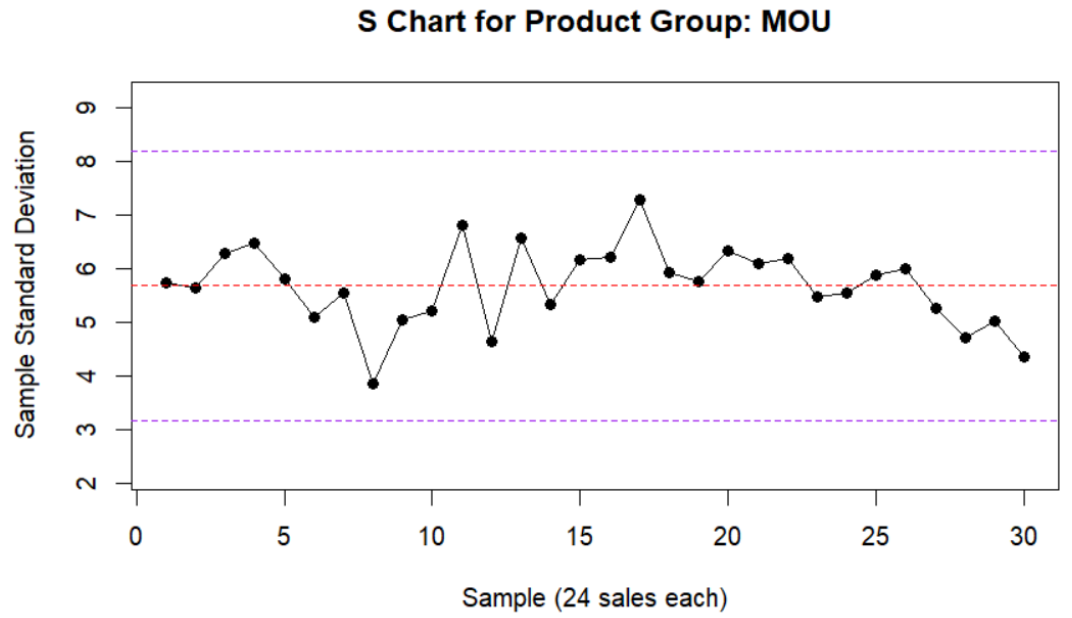
References

- Illowsky, B. & Dean, S. (2018) *Introductory Statistics*. Houston, TX: OpenStax, Rice University. Available at: https://openstax.org/details/books/introductory-statistics [Accessed 02 October 2025].
- Staff, C. (2025). *Navigating Your Data Analyst Career Path: From Entry-Level to Expert*. [online] Coursera. Available at: https://www.coursera.org/articles/rstudio?msockid=1345c5b2d6e0605a1524d72cd7926190 [Accessed 02 Oct. 2025].
- Crabtree, M. (2023). *What is Data Analysis? An Expert Guide With Examples*. [online] Datacamp.com. Available at: https://www.datacamp.com/blog/what-is-data-analysis-expert-guide.
- Berardinelli, C. (2013). *The Complete Guide to Understanding Control Charts*. [online] isixsigma.com. Available at: https://www.isixsigma.com/control-charts/a-guide-to-control-charts/.
- Frost, J. (2023). *Control Chart: Uses, Example, and Types*. [online] Statistics By Jim. Available at: https://statisticsbyjim.com/graphs/control-chart/.
- croft, D. (2023). *Guide: Process Capability Analysis (Cp, Cpk) - Learn Lean Sigma*. [online] www.learnleansigma.com. Available at: https://www.learnleansigma.com/guides/process-capability-analysis-cp-cpk-pp-ppk/.
- GeeksforGeeks (2024). *Type I and Type II Errors*. [online] GeeksforGeeks. Available at: https://www.geeksforgeeks.org/maths/type-i-and-type-ii-errors/.
- Amplitude (2024). *Type 1 and Type 2 Errors Explained - Differences and Examples | Amplitude*. [online] Amplitude.com. Available at: https://amplitude.com/explore/experiment/type-1-and-type-2-errors-explained.
- Sthda.com. (2025). *MANOVA Test in R: Multivariate Analysis of Variance - Easy Guides - Wiki - STHDA*. [online] Available at: https://www.sthda.com/english/wiki/manova-test-in-r-multivariate-analysis-of-variance.
- Hassan, M. (2024). ANOVA (Analysis of variance) - Formulas, Types, and Examples. [online] Research Method. Available at: https://researchmethod.net/anova/.
- OpenAI. (2025). *ChatGPT. [online] Available at: https://chatgpt.com/ [Accessed 1 October 2025].

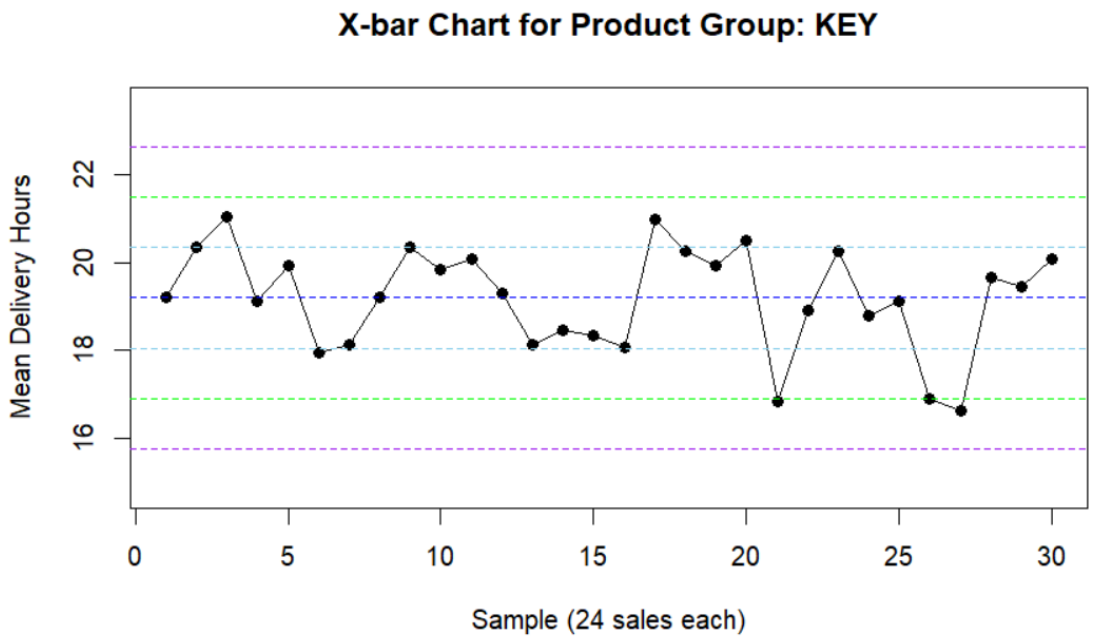
Appendix A



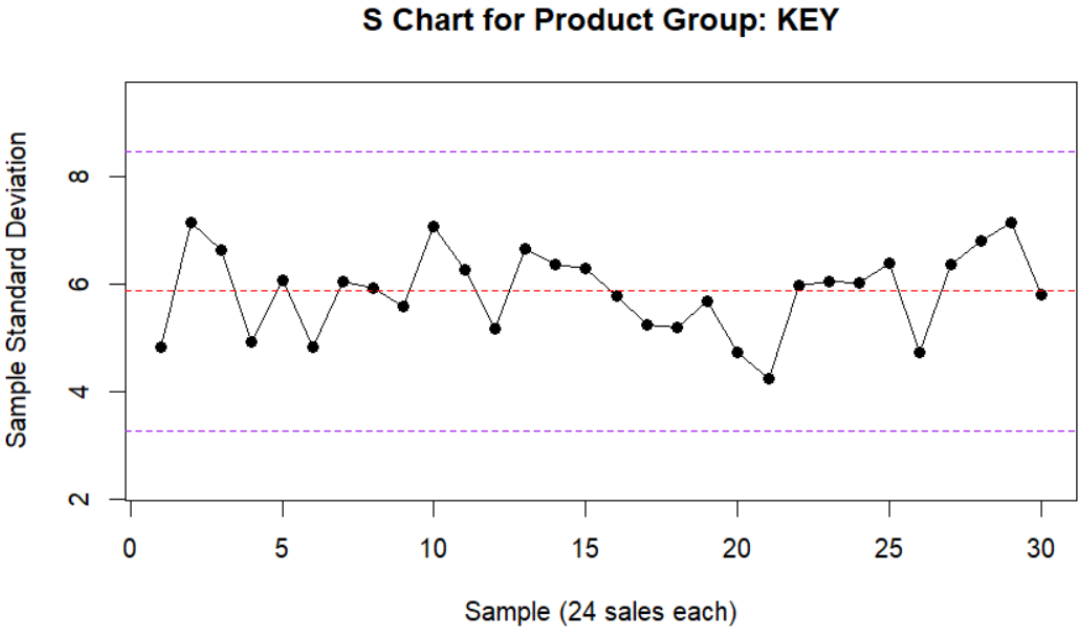
Graph 1



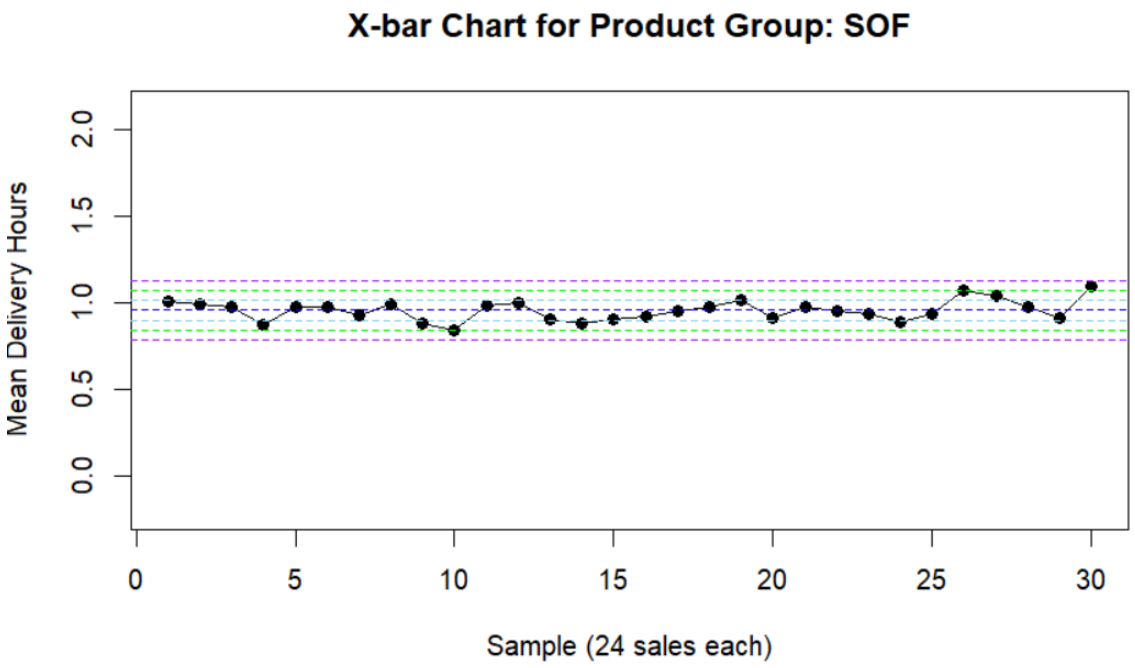
Graph 2



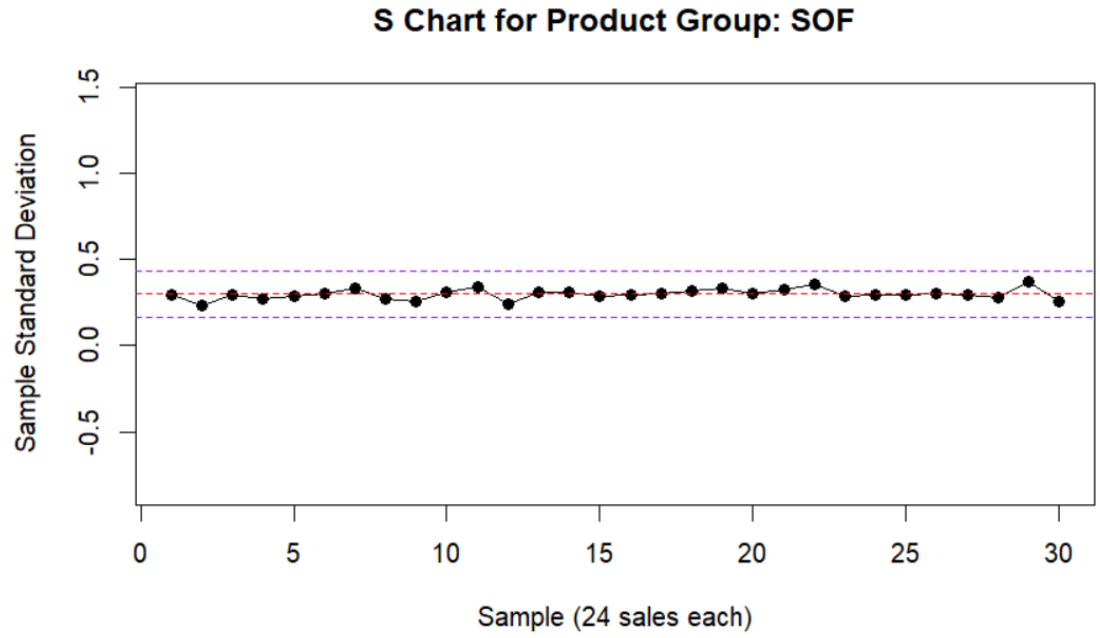
Graph 3



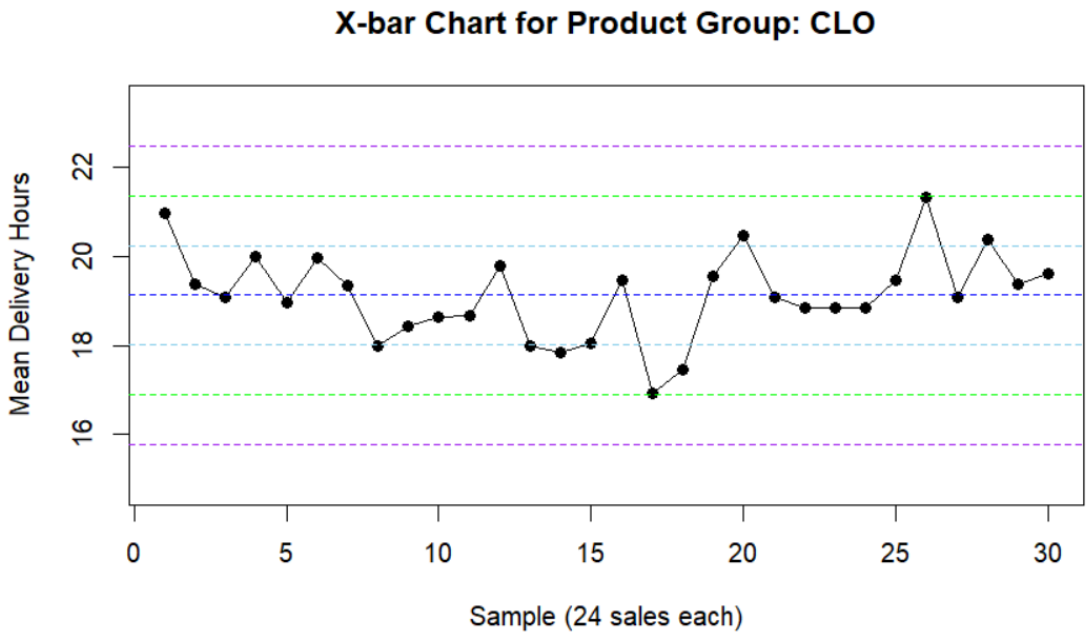
Graph 4



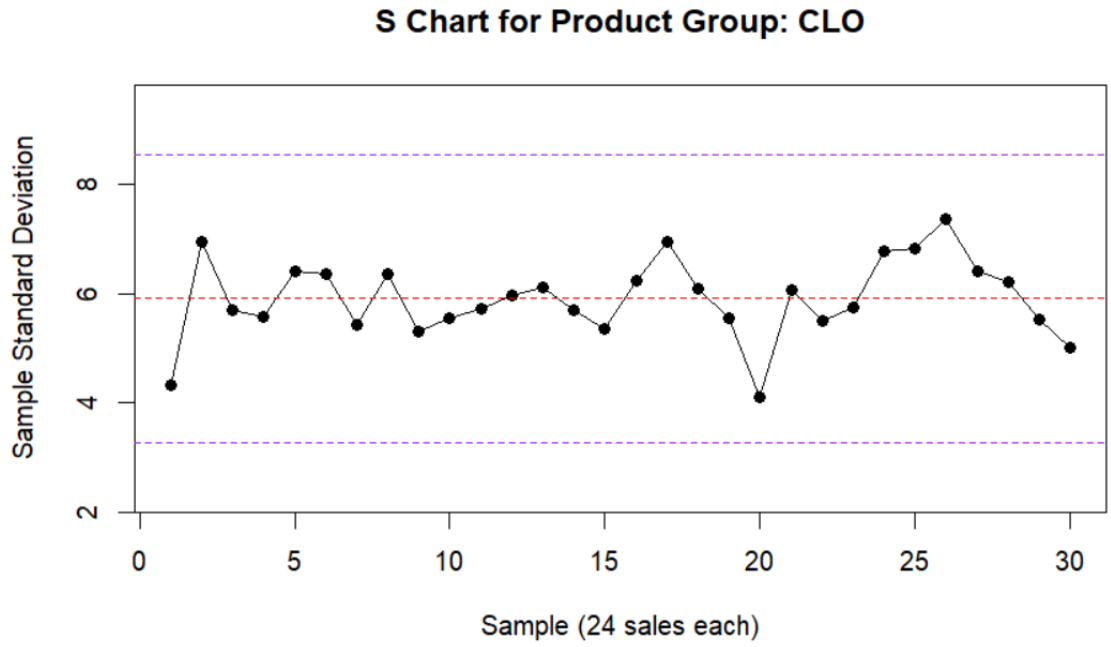
Graph 5



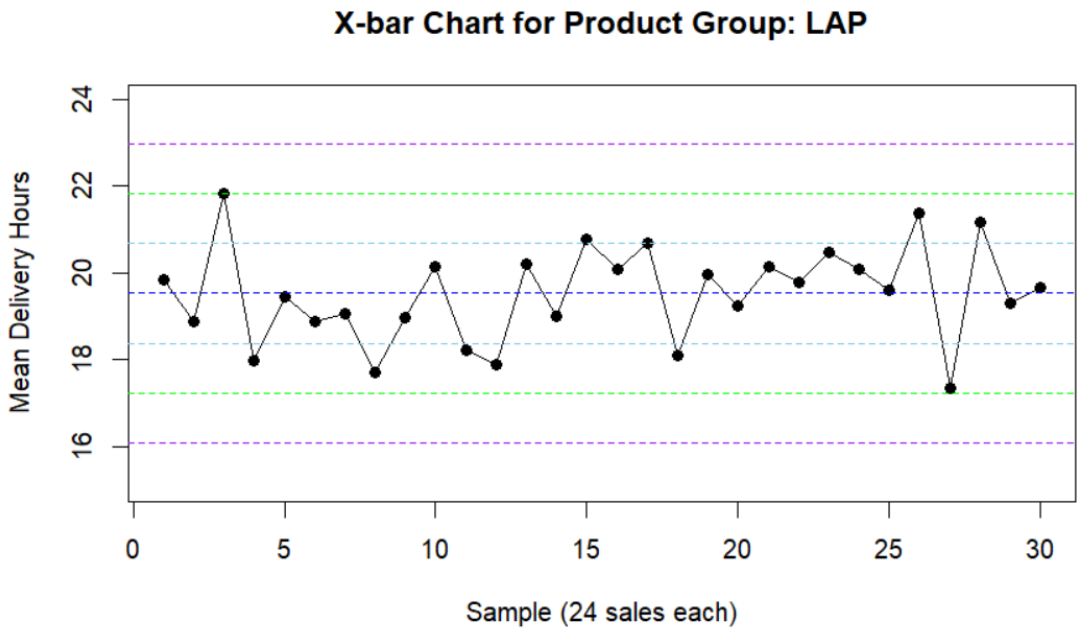
Graph 6



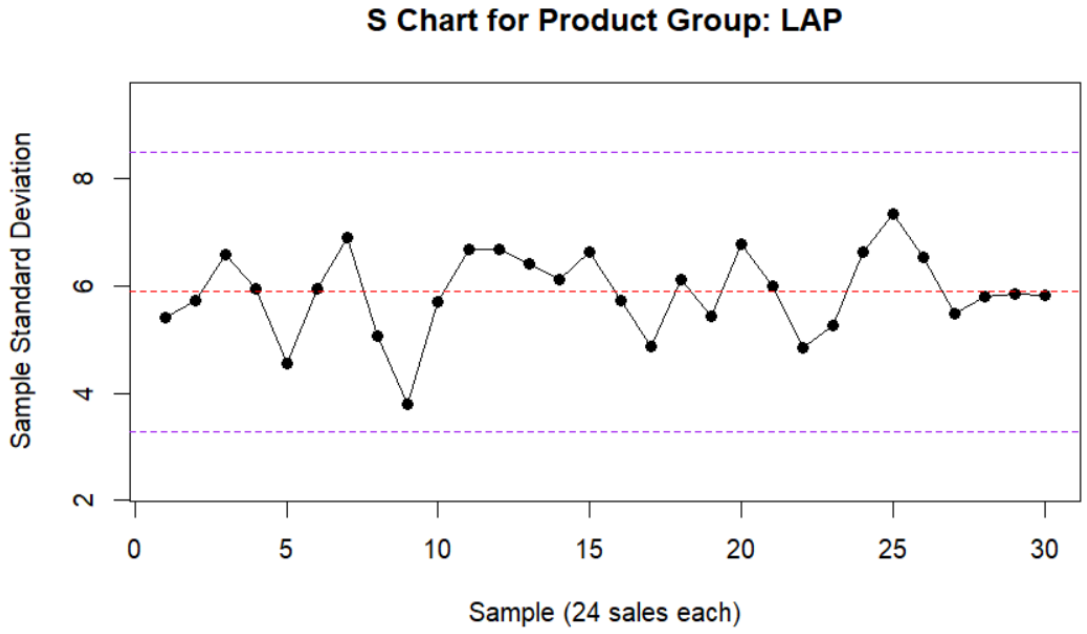
Graph 7



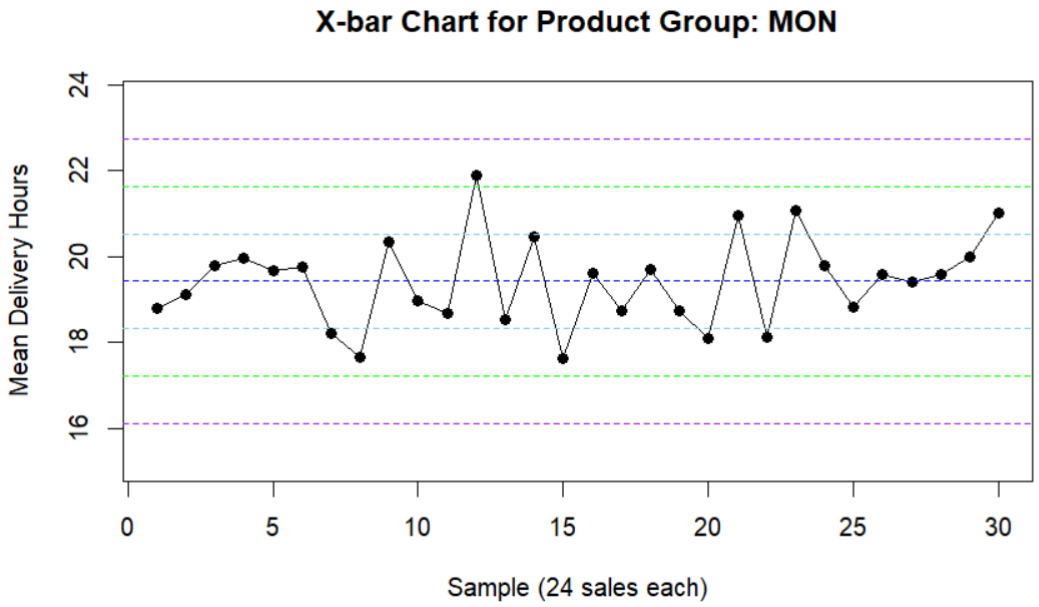
Graph 8



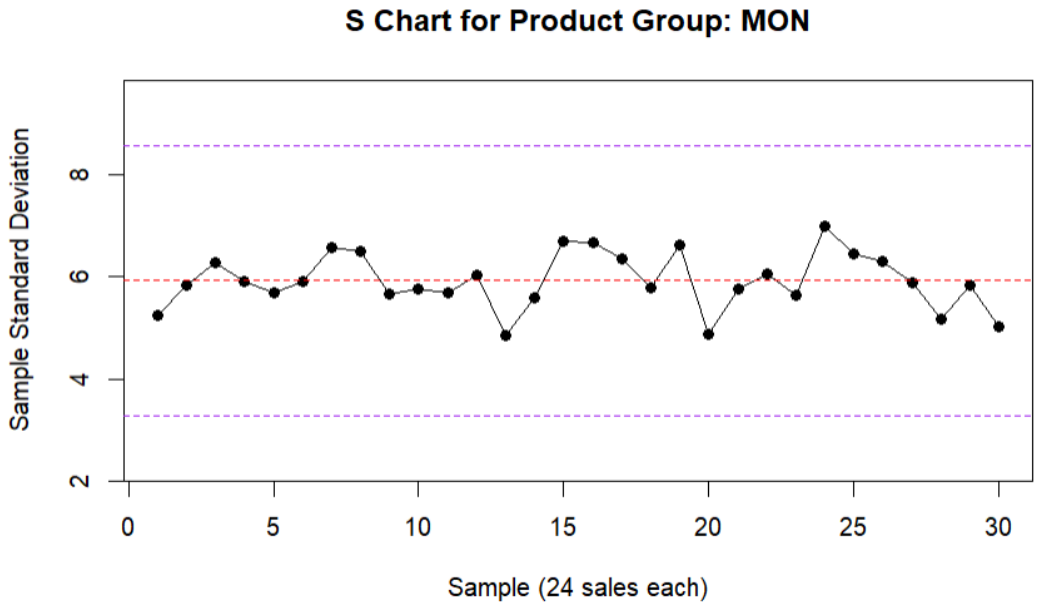
Graph 9



Graph 10

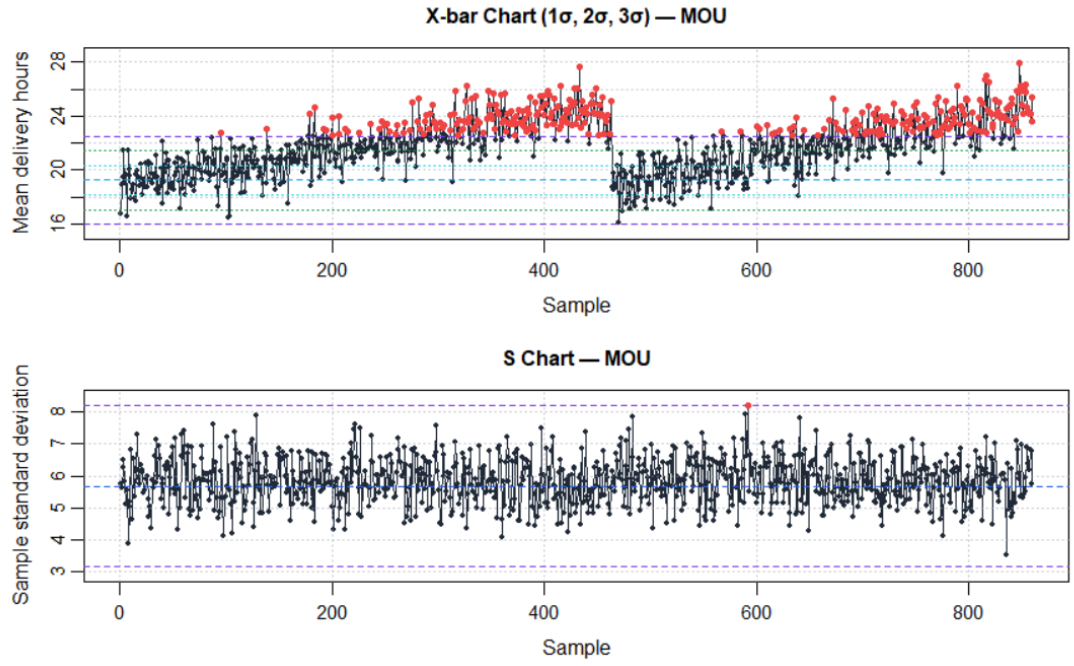


Graph 11

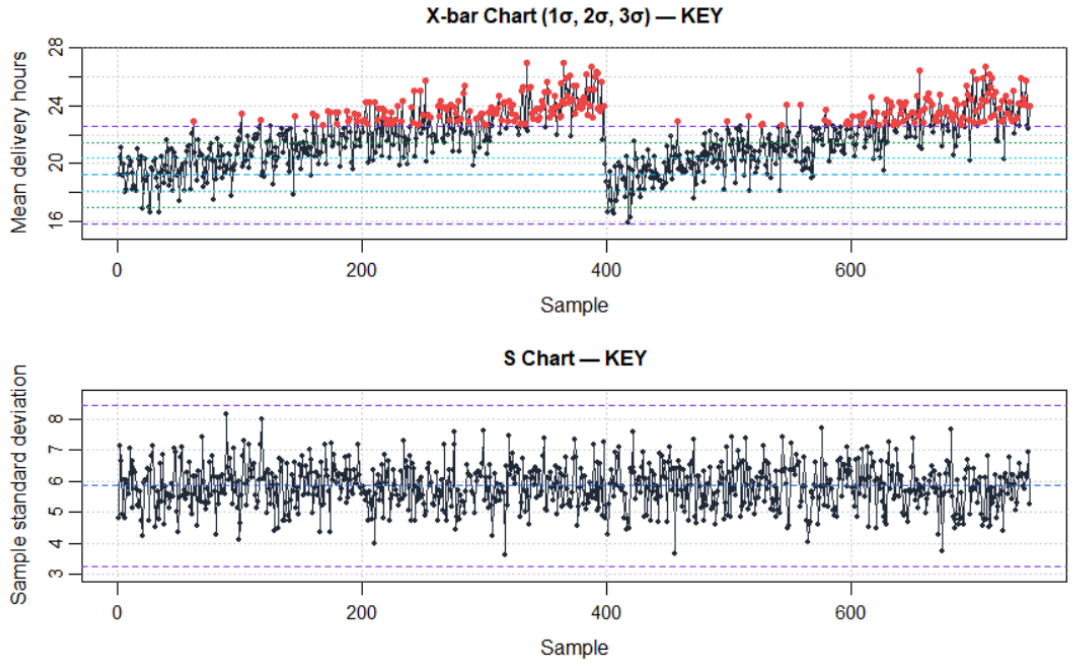


Graph 12

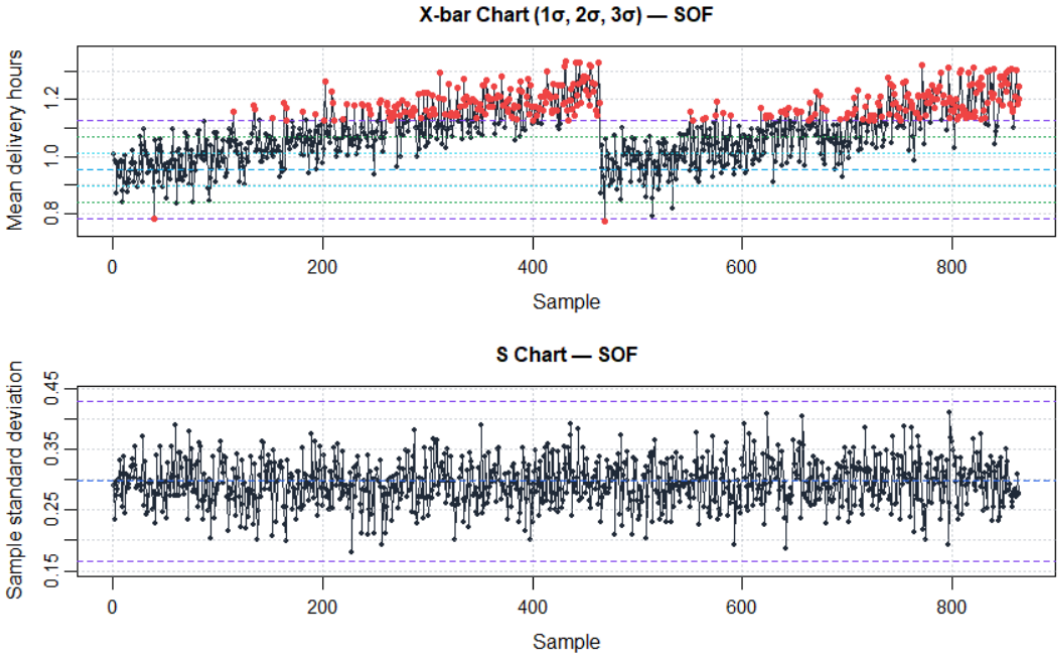
Appendix B



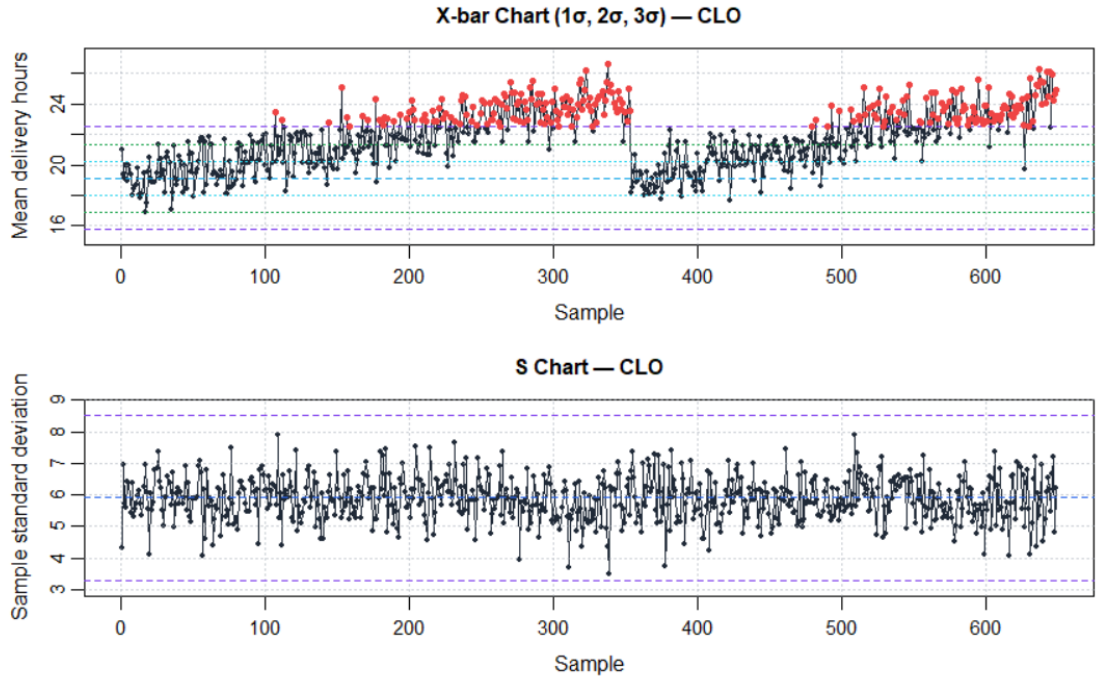
Graph 13



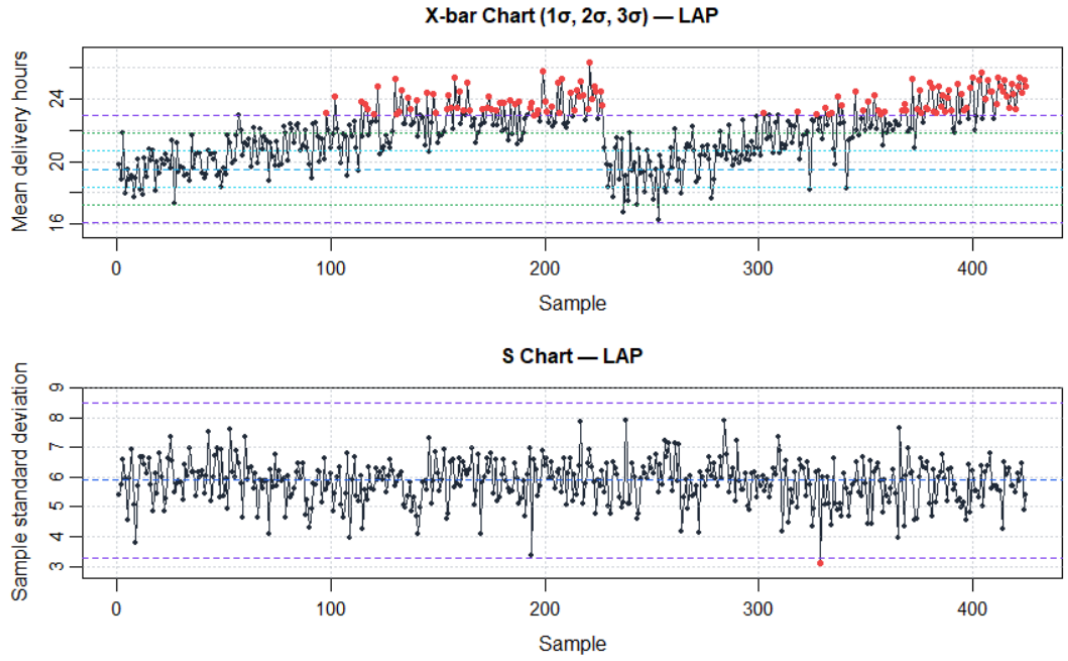
Graph 14



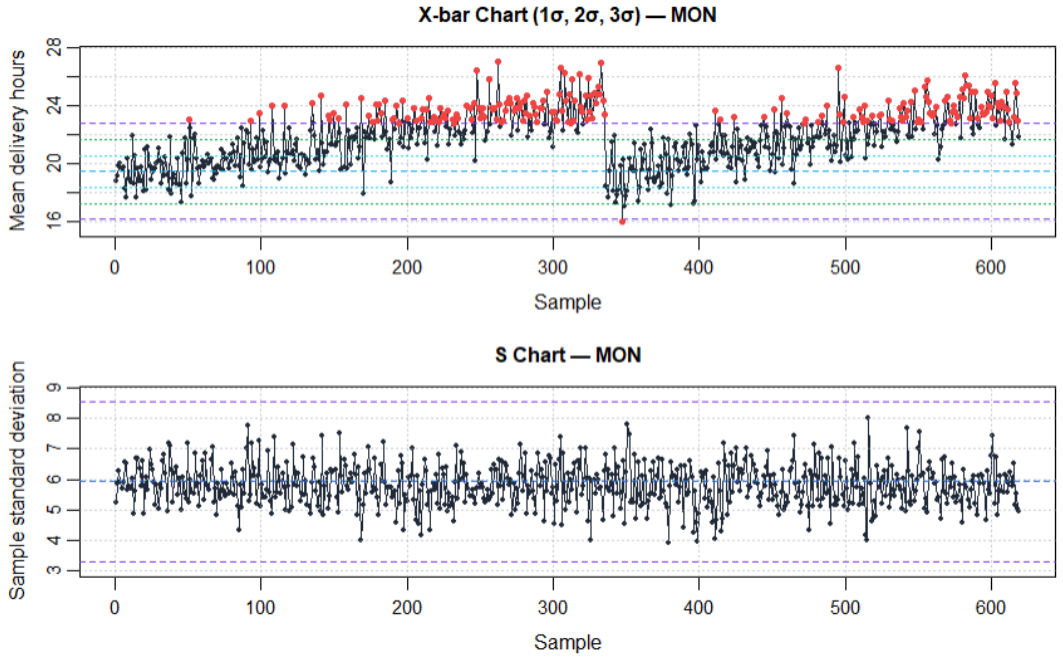
Graph 15



Graph 16



Graph 17



Graph 18