26946092
Carla Mostert
Quality Assurance 344
ECSA Project

Table of Contents

List of Figures

1. Introduction

This report demonstrates the ability to investigate complex problems using data-driven engineering approaches. The purpose of this project is to apply statistical and analytical techniques to real-world industrial data to evaluate, optimise, and improve process performance and reliability.

This report integrates multiple components of industrial data analysis, including descriptive statistics, statistical process control (SPC), process capability analysis, error estimation, and profit optimisation. Using R programming, the datasets provided (such as *customer_data*, *products_data*, *products_headoffice*, *sales2022and2023*, and *timeToServe*) were explored and analysed to uncover patterns and detect process issues.

This report also addresses practical business objectives such as reducing variability in delivery times, improving data accuracy, and maximising profitability while ensuring service reliability. Through the application of ANOVA testing and reliability models, the report demonstrates how engineering analytics can guide continuous improvement and operational excellence.

2. Data Analysis

2.1. Data Loading and Inspection

There are four datasets provided, namely customer_data, Products_data, Products_Headoffice, and sales2022and2023. After loading the datasets into R, we can then examine their dimensions, structures and variable types. This gives us the following results.

- Customer Data: We can see that this file contains 5,000 customers (rows) with 5 variables or columns. The file provides demographic information such as gender, age, income levels, and city of residence.

Table 1: Customer Data Structure Sample

	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
7	CUST007	Female	31	100000	Los Angeles
8	CUST008	Male	27	90000	Los Angeles
9	CUST009	Female	26	35000	Chicago
10	CUST010	Male	28	105000	San Francisco

- Products Data: There are 60 products listed (rows) with 5 variables: ProductID, Category, Description, SellingPrice and Markup. This dataset represents a local product list with pricing and markup information.

Table 2: Products Data Structure Sample

	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
7	SOF007	Software	black bright	527.56	16.79
8	SOF008	Cloud Subscription	burlywood silk	549.02	11.95
9	SOF009	Laptop	azure sandpaper	540.41	11.34
10	SOF010	Monitor	chocolate sandpaper	396.72	23.47

- Products Head Office: Contains 360 products (rows) with the same structure as the local product list (Products Data). However, there are differences in the product description and pricing as highlighted in Tables 2 and 3. T.
  For example, ProductID SOF001 is described as “coral matt” with the selling price of R511.53 and a markup of 25.05% in the local product dataset. While in the head office dataset, it appears to be described as “coral silk” with the selling price of 521.72 and a markup of 15.65%.
  These discrepancies suggest inconsistencies between local and head office product records.

Table 3: Products Head Office Structure Sample

	ProductID <chr>	Category <chr>	Description <chr>	SellingPrice <dbl>	Markup <dbl>
1	SOF001	Software	coral silk	521.72	15.65
2	SOF002	Software	black silk	466.95	28.42
3	SOF003	Software	burlywood marble	496.43	20.07
4	SOF004	Software	black marble	389.33	17.25
5	SOF005	Software	chartreuse sandpaper	482.64	17.60
6	SOF006	Software	cornflowerblue marble	539.33	25.57
7	SOF007	Software	blue marble	495.13	10.23
8	SOF008	Software	cornflowerblue marble	465.73	21.89
9	SOF009	Software	black bright	452.40	19.64
10	SOF010	Software	cornflowerblue matt	399.43	17.08

- Sales Data: There are 100,000 sales transactions (rows) with 9 variables: CustomerID, ProductID, Quantity, orderTime, orderDay, orderMonth, orderYear, pickingHours and deliveryHours. This dataset records customer orders across 2022 and 2023, including order details and logistics timings.

Table 4: Sales of 2022 and 2023 Structure Sample

	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
7	CUST4454	MOU052	29	5	23	1	2022	12.38833	25.544
8	CUST582	MON032	1	19	9	6	2023	17.05750	22.046
9	CUST3343	MON040	10	19	13	12	2023	24.05750	24.046
10	CUST4331	KEY049	1	18	30	4	2022	15.38833	20.044

2.2. Summary Statistics

Descriptive statistics were computed to identify central tendencies, ranges and distributions.

- Customer Data:
  Age ranges from 16 to 105 years, with a median of 51 years and a mean of 51.6.
  Income ranges from R5,000 to R140,000, with a median of R85,000 and a mean of R80,797.
  Gender distribution was mostly female (2,432), followed by Male (2,350) and Other 218.
  It is clear that most customers live in San Francisco, with 780 customers; the fewest customers live in Miami, with 647 in total.
  Customers are mostly middle-aged, with incomes clustered between R55,00 and R105,000 (quartiles).

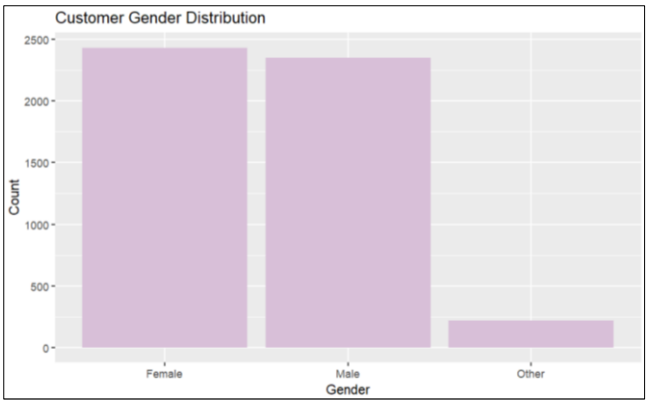

Figure 1: Bar plot of the Gender Distribution

- Products Data:
  The average selling price of products was R4,493.60 with a median of R794.20 and ranged widely between R350.40 and R19,725.20. We can see that expensive products skewed the selling price.
  The markup price of products ranges from 10.13% to 29.84% with a mean markup price of 20.46%.
  There is a total of 6 categories: Software, Cloud Subscription, Laptop, Monitor, Mouse and Keyboard, each containing 10 products.
  Products are evenly distributed across categories, with most products priced under R1,000; however, a few premium products significantly raise the average price.

- Products Head Office:
  The selling price ranges from R290.50 to R22,420.10, and the mean selling price is R4,411.
  The markup is similar to the local product data (Products Data), ranging from 10% to 30%, with a mean of 24.39%.
  The categories are distributed evenly, with 60 items in each of the 6 categories.
  Overall pricing and markup are consistent with the local dataset, but product descriptions and prices for identical IDs differ, suggesting possible data integrity issues.

- Sales Data:
  The quantity of sales ranges from 1 to 50 units, the mean size of orders is 13.5, with a median of 6. This indicates that most orders are small, but large bulk orders increase the average.
  We have order patterns of years, months, days, and the time of the day.
  2022 had 57,737 sales, while 2023 had 46,263 sales. Thus, we had a decline of 7,464 transactions.
  The most sales occurred around 12:00 midday. Interestingly, during the month, the most sales occurred on the 16th, followed by the 22nd and the 15th. February and November tied as peak sales months. This could probably be linked to seasonal promotions like Valentine's Day and pre-Christmas shopping.
  The average picking hours were 14.7 hours with a range of 0.4 to 45 hours. Delivery hours had an average of 17.5 hours and a range of 0.3 to 38 hours.
  Although most orders are small, sales volumes are substantial. A decline in 2023 sales may indicate external factors such as market saturation or reduced demand.

2.3. Handling Missing Values

Missing values and data need to be found and addressed so that the results can be more accurate and not skewed.

The presence of missing values was checked across all datasets using column-wise counts. No missing values were detected in any of the four datasets.
Since no missing data is present, no imputation or row removal was necessary. All subsequent analyses were conducted on the complete datasets.

2.4. Data Filtering and Subsetting

Several subsets were created to better understand the data. The first subset created was Sales Data, which is a subset of sales for 2022 that contained 53,737 transactions, while 2023 contained 46,263 transactions. This confirms a decline in sales year-on-year, as shown below in Figure 2.

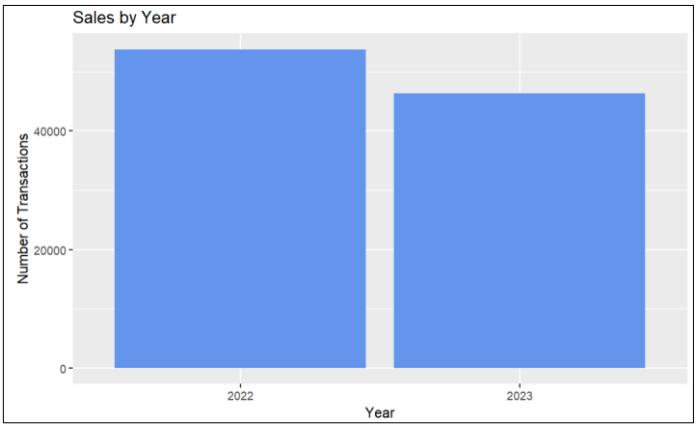

Figure 2: Bar Plot of Sales Data over the years

The next subset that was created was Customer Data. Filtering the customer data by city showed that San Francisco had the highest customer base (780 customers), while Miami had the lowest customer base (647 customers). This is visualised in Figure 3 below.

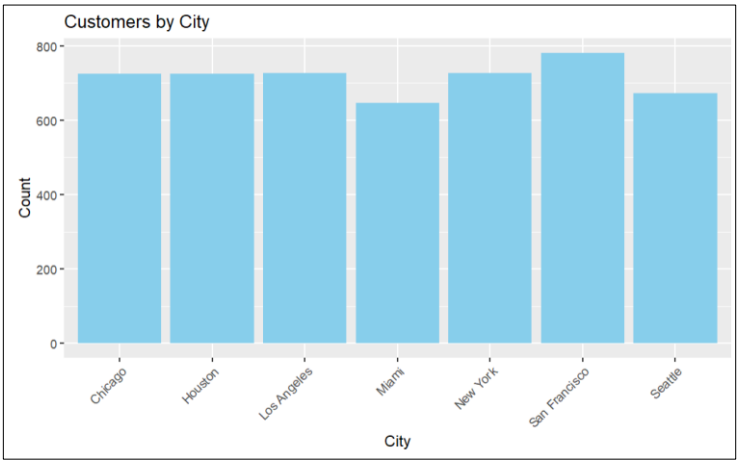

Figure 3: Bar Plot of the Customer Base per City

Lastly, the subset Product Data was used. Filtering by selling price identified a subset of mid-priced products (R50 – R200). These were unevenly distributed across categories, with most being accessories like keyboards and mice. We can illustrate this by looking at Figure 4.

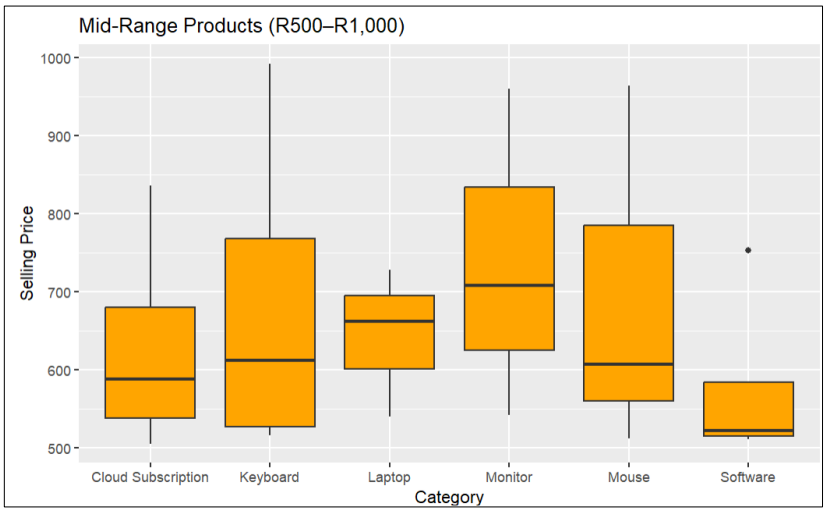

Figure 4: Box Plot of Selling Price vs Category

2.5. Data Visualisation

Looking at the gender distribution in Figure 5, we can see that the customer base is fairly balanced, with a slightly higher number of females than males. A small portion of customers is listed as “Other”. This confirms earlier frequency counts.

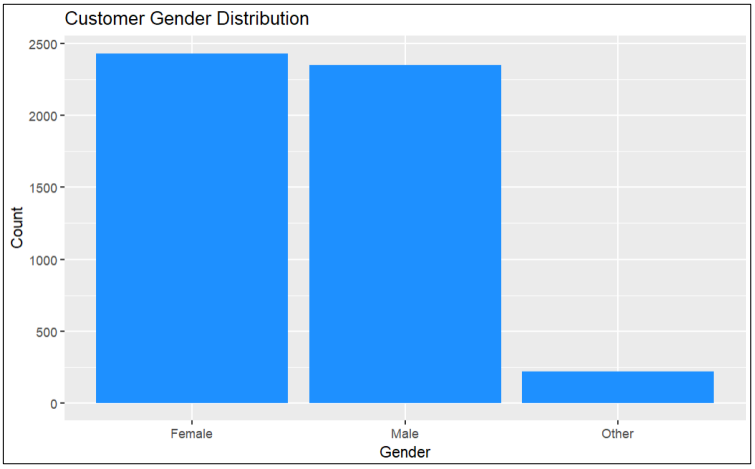

Figure 5: Bar Plot of Customer Gender Distribution

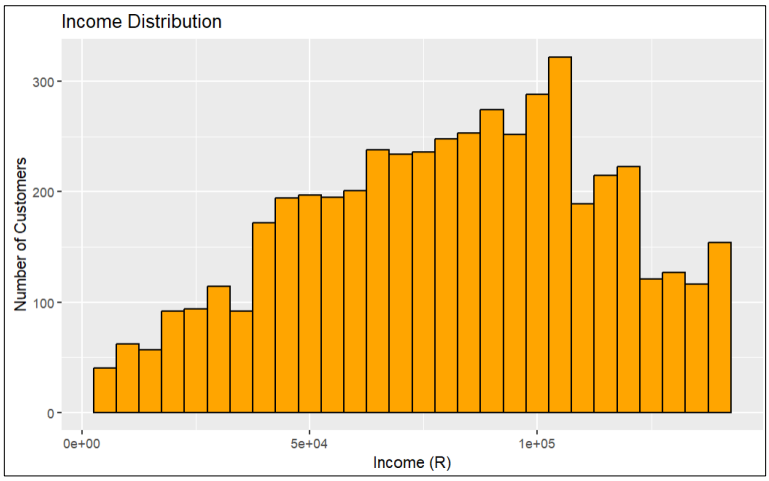

Figure 6: Histogram of the Customer Income Distribution

The income distribution in Figure 6 shows that most customers earn between R55,000 and R105,000, with fewer customers in lower or higher income brackets. The distribution is fairly symmetric, with some high-income outliers.

The customer income distribution approximates a normal shape (with a mean of R80,800 and a standard deviation of R25,000) but shows a mild positive skew, suggesting a minority of high-income clients.

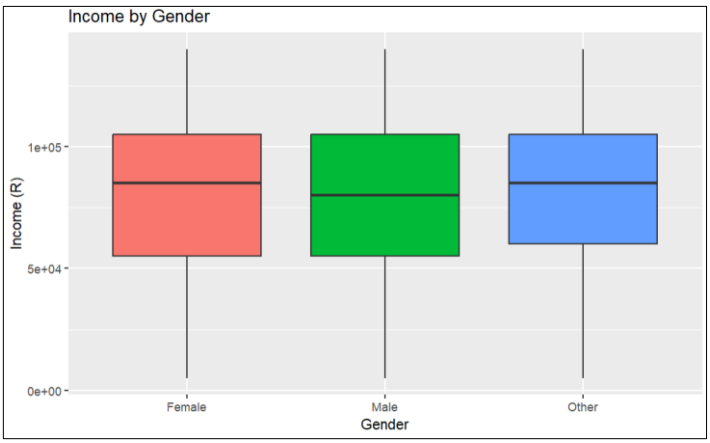

Figure 7: Boxplot of Customer Income by Gender

The income by gender plot below illustrates that male and female customers show similar median incomes, though male customers display slightly more variability.

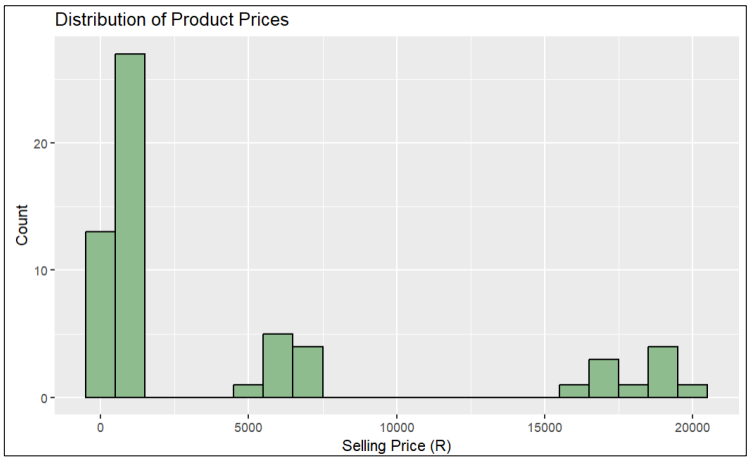

Figure 8: Histogram of the Product Price Distribution

The product price distribution illustrates that the majority of products are priced under R1,000, but a small number of high-value products (above R15,000) skew the price distribution upward. This can be seen in Figure 8.
Product price distributions are strongly right-skewed due to premium laptop and monitor lines, indicating non-random variation tied to product category rather than process error.

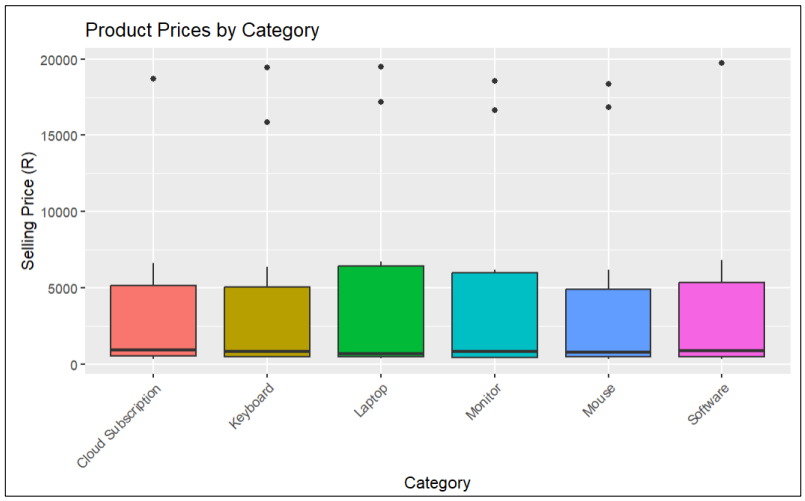

Figure 9: Boxplot of Product Prices by Category

The product prices by category plot indicates that laptops and monitors show the highest price variability, while items such as mice and keyboards are consistently lower priced.

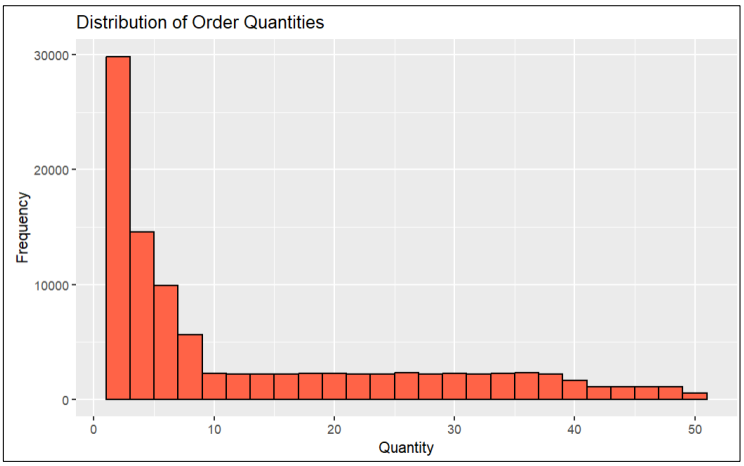

Figure 10: Histogram of the Sales Quantity Distribution

Looking at the sales quantity distribution in Figure 10, we can see that most sales transactions involve small orders (fewer than 10), though there are some bulk orders up to 50 units. This indicates that high-volume orders are relatively rare but still significant.

2.6. Exploring Relationships

The total sales by product category (shown in Figure 12) illustrate that sales are dominated by categories such as Laptops and Software, while categories like Keyboards and Mice contribute less overall. This suggests high-value items are key revenue drivers.

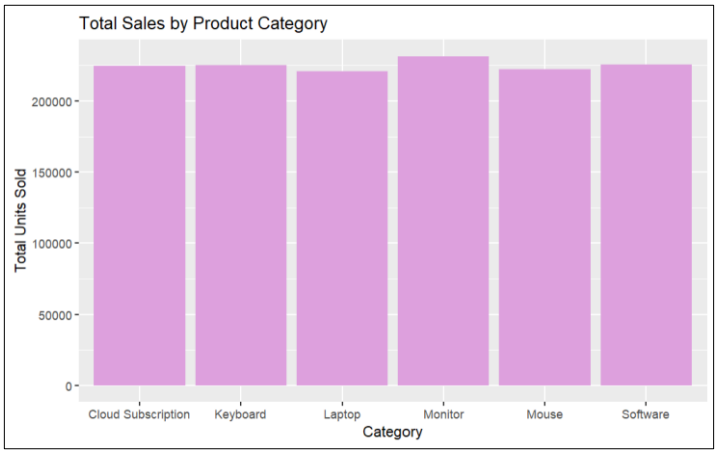

Figure 11: Bar Plot of the Total Sales by Product Category

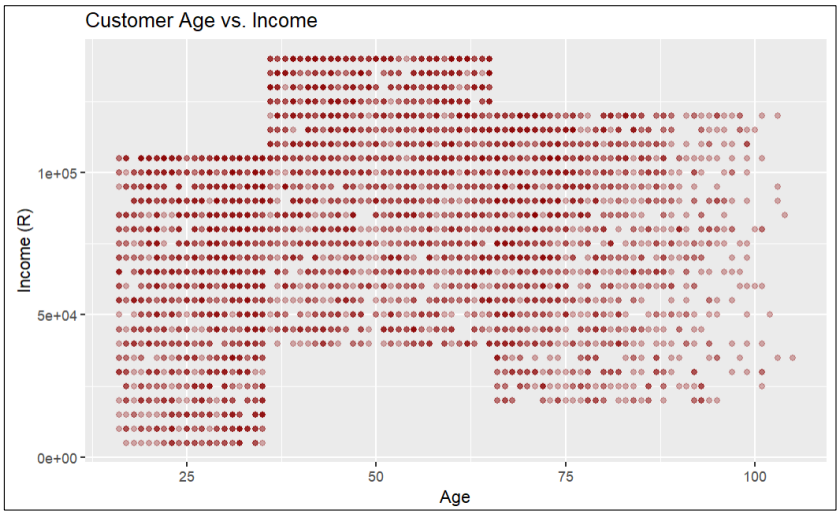

Figure 12: Scatterplot of Customer Income vs Age

The scatterplot of customer income vs age in Figure 12 shows a wide spread of incomes across all ages, with no strong correlation. However, older customers tend to have slightly higher incomes, indicating potential targeting opportunities for premium products.

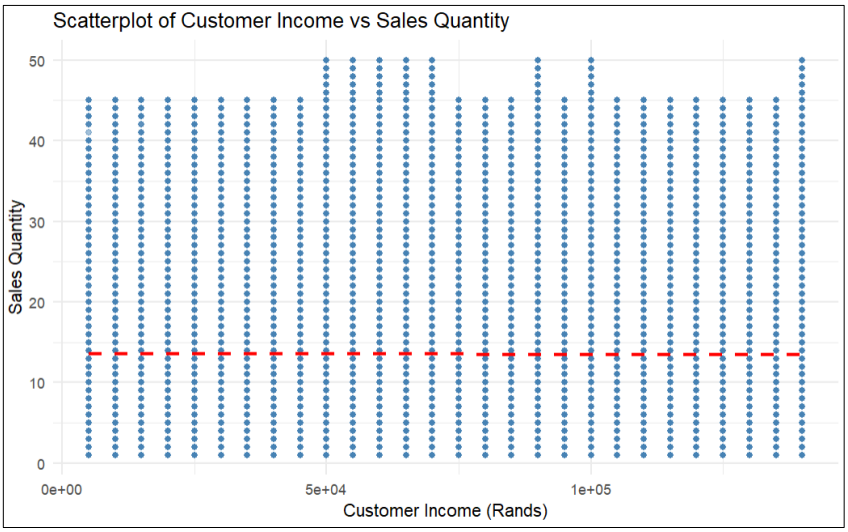

Figure 13: Customer Income vs Sales Quantity

A weak positive correlation (r = 0.25) between income and sales quantity implies that higher-income customers purchase slightly more expensive items but not necessarily in larger quantities. This trend highlights a potential marketing opportunity for premium bundles.

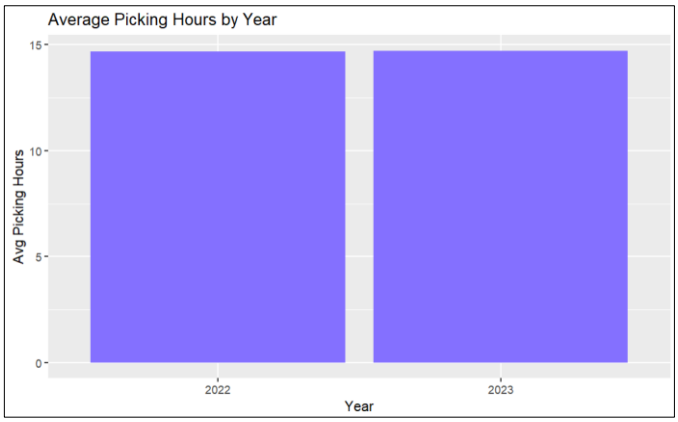

Figure 14: Bar Plot of the Average Picking Hours by year

Average picking hours by year are shown in Figure 13. The average time required for picking increased slightly in 2023 compared to 2022. This could indicate operational inefficiencies or increased warehouse workload. Monitoring this trend could help improve logistics efficiency.

The descriptive results therefore confirm that variation in the dataset is largely explainable by structural business factors rather than data errors.
These analyses highlight where the company generates most sales, provide insights about customer demographics, and identify potential inefficiencies in the order fulfilment process.

3. Statistical Process Control – Part 3

In this section, we apply Statistical Process Control (SPC) techniques to monitor and evaluate the stability and capability of the company's delivery time process for different product types over 2026 and 2027.

Using the sales2026and2027.csv dataset, control charts ($\bar{X}$ and s) are generated to analyse process variation, and process capability indices (Cp, Cpk, Cpu, Cpl) are calculated to assess whether the process meets the required specifications. Finally, potential out-of-control patterns are identified.

3.1. Initialisation of $\bar{X}$ and s Charts

The data were sorted by product and time (first year, then month, day, and order time). For each product type, the first 30 x 24 = 720 delivery times were divided into 30 samples of size 24.
From these, the sample means and standard deviations were calculated. Using $A_3 = 0.619$, $B_3 = 0.555$, and $B_4 = 1.445$ for a sample size of 24, the control limits for the $\bar{X}$ and s charts were established. The variables A_3, B_3, and B_4 are from the control chart constant table for this specific sample size.

The purpose of getting our $\bar{X}$ charts is to visually see the mean of each product. This helps us detect changes in the central tendency. The s chart helps us to visualise the standard deviation, which detects the changes in variability.

The centre line (CL) represents the mean of the first 30 samples, which is the red line on the figures below. The upper and lower control limits (UCL and LCL) indicated in blue on the figures below correspond to ±3σ boundaries, while 1σ and 2σ limits were added by dividing the area between CL and UCL (as well as CL and LCL) into three equal parts. The one-sigma control limit is indicated in yellow dashed lines, and the 2-sigma control limits are indicated in orange in the figures below.

The 1 and 2-sigma control limits help us to visually identify patterns. All the data should be inside the outer control limit lines.

Both the $\bar{X}$ and s charts show that all points are within the control limits, indicating that the process is initially in control and stable. The average delivery time is approximately.

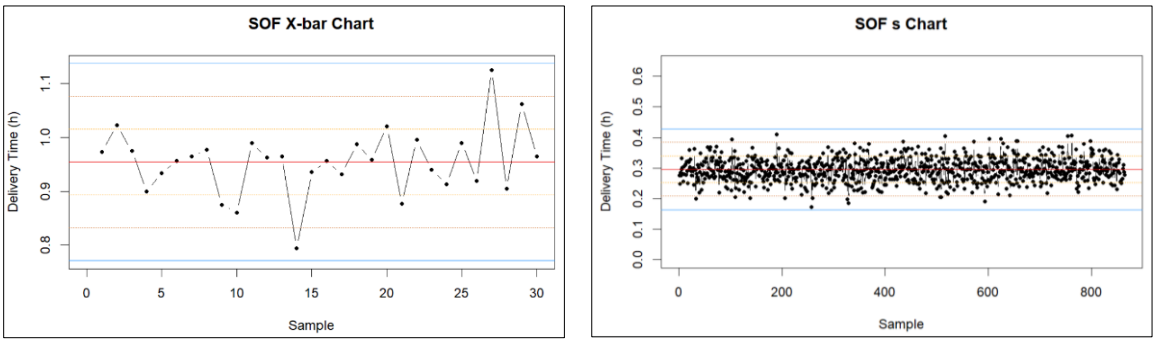

Figure 15: Product SOF X-bar and S Chart

For the SOF products, there are no points outside the outer control limits (UCL and LCL in blue), 1 point above and 1 below the 2-sigma control line (indicated in orange). There are 4 points above and 4 points below the 1-Sigma-control line (indicated in yellow).

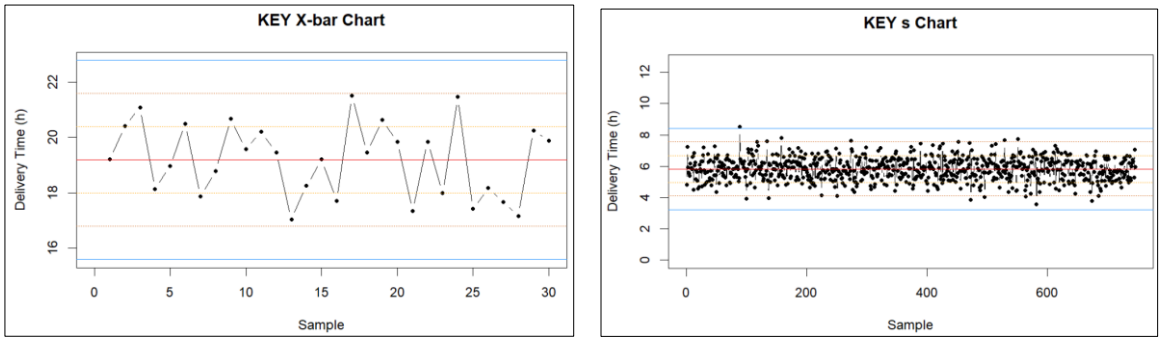

Figure 16: Product KEY X-bar and S Chart

For the KEY products, there are no points outside the outer control limits (UCL and LCL in blue), and none above or below the 2-sigma control line (indicated in orange). There are 7 points above and 8points below the 1-Sigma-control line (indicated in yellow).
On the s-chart of KEY products, however, there is one point outside of the upper control line.

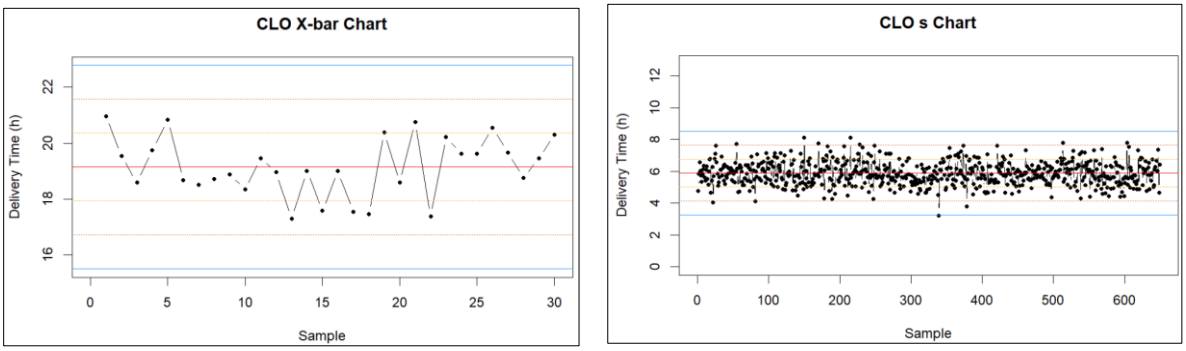

Figure 17: Product CLO X-bar and S Chart

Both the X-bar and s charts (in Figure 16 above) for the product CLO during the initialisation phase demonstrate that the process was stable. The mean delivery time and process variability remained within their respective control limits with no unusual patterns.

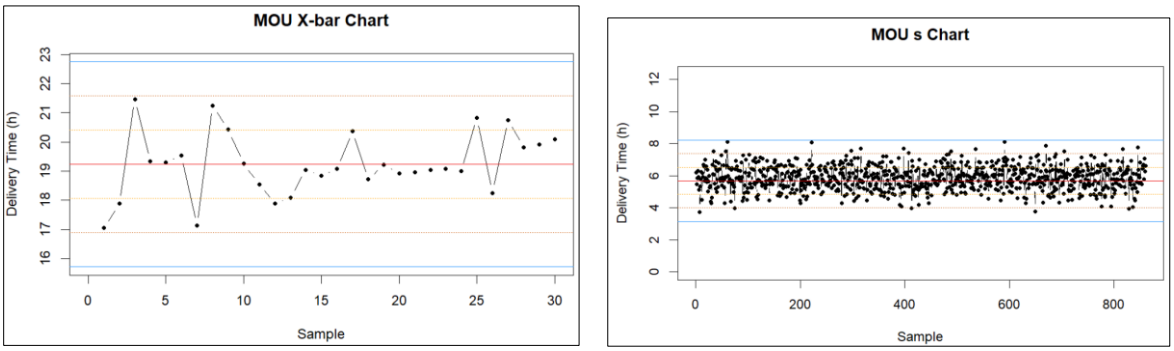

Figure 18: Product MOU X-bar and S Chart

The X-bar and s charts (Figure 17) for the product MOU during the initialisation phase demonstrate that the process was stable. The mean delivery time and process variability remained within their respective control limits with no unusual patterns.

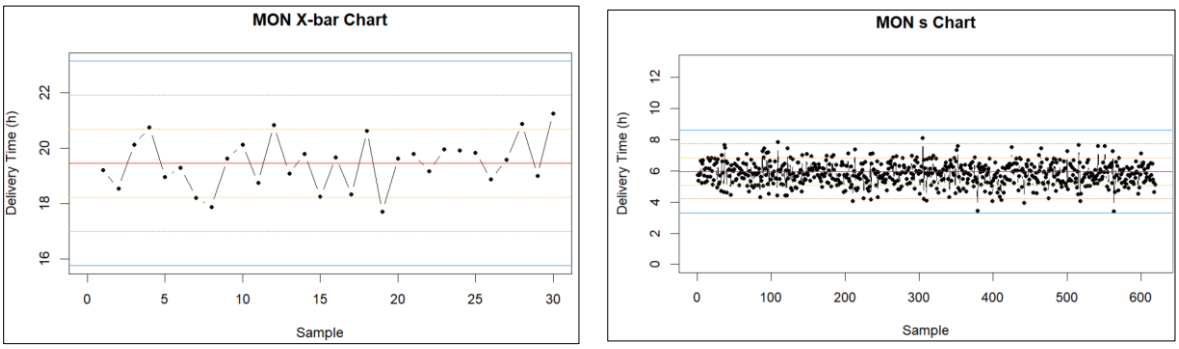

Figure 19: Product MON X-bar and S Chart

The same can be said for the MON product's X-bar and s charts in Figure 18 above. During the initialisation phase, the charts demonstrate that the process was stable. The mean delivery time and process variability remained within their respective control limits with no unusual patterns.

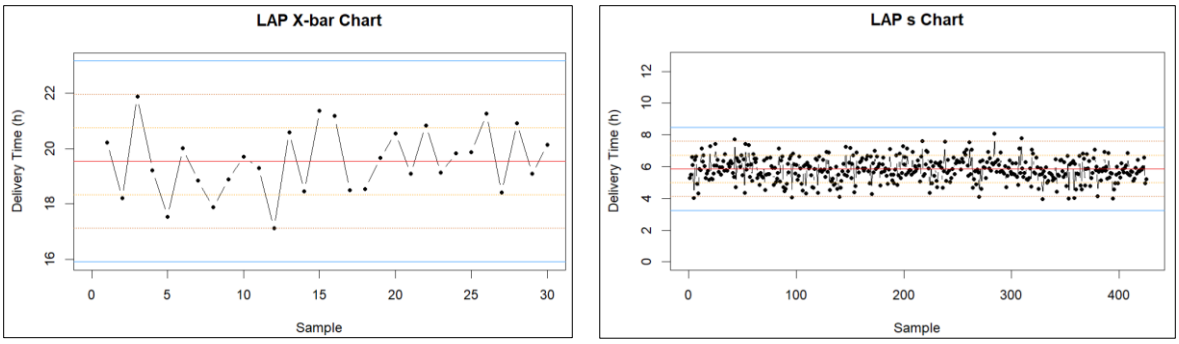

Figure 20: Product LAP X-bar and S Chart

Lastly, the X-bar and s charts in Figure 19 for the product LAP during the initialisation phase demonstrate that the process was stable. The mean delivery time and process variability remained within their respective control limits with no unusual patterns.

3.2. Control Phase

Subsequent examples were plotted using the same control limits to test ongoing process stability.
Out-of-control points beyond ±3σ were highlighted.

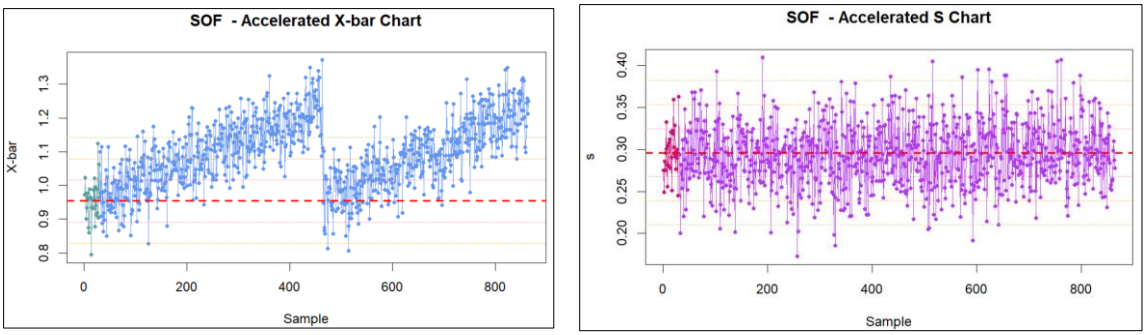

Figure 21: Product SOF Accelerated X-bar and S Chart

The X-bar and s charts for product SOF above indicate that the process variability remains stable, but the average delivery hours are increasing over time, moving outside the control limits. This pattern suggests that the process mean has shifted, and the system is no longer in statistical control. It is recommended to identify and correct the source of this drift before continuing production.

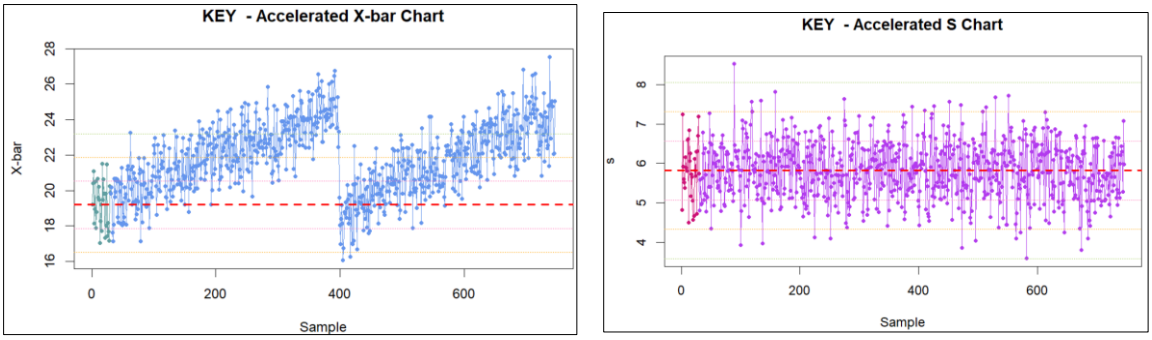

Figure 22: Product KEY Accelerated X-bar and S Chart

The KEY product's delivery process appears stable. In the s chart, no immediate action is required, but continued monitoring is advised. Most points fall within control limits, indicating consistent process spread. No points breach the outer control limits, suggesting stable variability. For the X-bar chart in Figure 21, a few points approach the 2-sigma limits but remain within bounds. There are no clear patterns (e.g., runs or trends) that suggest random variation.

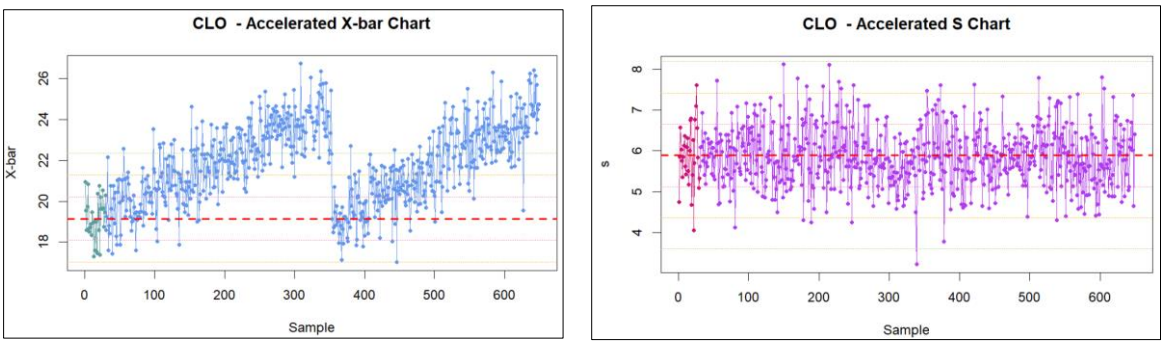

Figure 23: Product CLO Accelerated X-bar and S Chart]

In the s chart, we can see that one or two points exceed the upper control limit, indicating increased variability in those samples. This suggests potential issues with consistency in delivery times. The cause of increased variability and potential shifts in delivery time in the CLO product should be investigated.

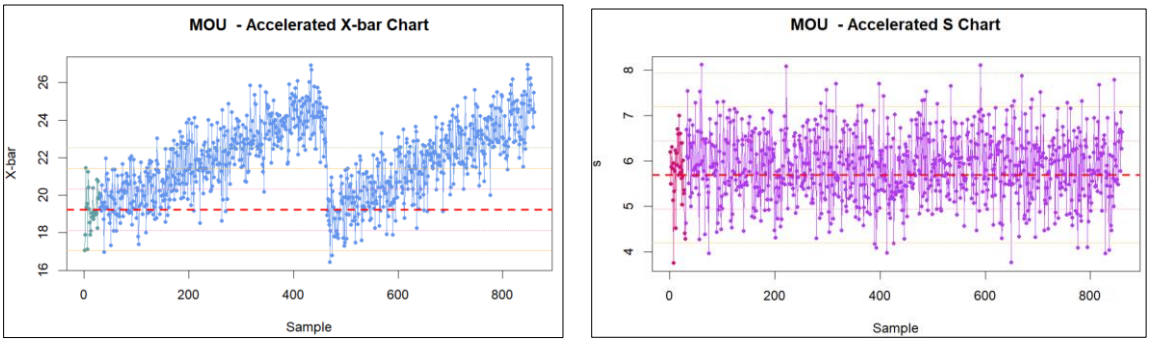

Figure 24: Product MOU Accelerated X-bar and S Chart

The MOU product shows signs of instability. There should be a review of recent process changes or external factors affecting delivery performance. On the X-bar chart in Figure 23, it is visible that several sample means fall outside the 2-sigma limits and that there is no consistent trend. The s chart indicates that several points are close to the upper control limit, which indicates growing inconsistency in the delivery time spread.

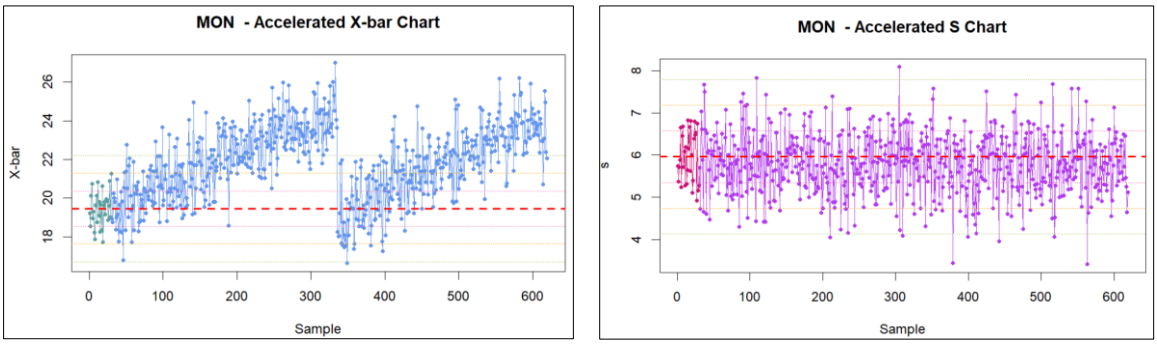

Figure 25: Product MON Accelerated X-bar and S Chart

The control charts for product MON suggest a stable delivery process. The X-bar chart shows that the sample means remain consistently within control limits, with no significant trends or patterns. The S chart also supports this stability, with most standard deviation values falling within the control boundaries. Overall, MON's delivery process appears to be in statistical control.

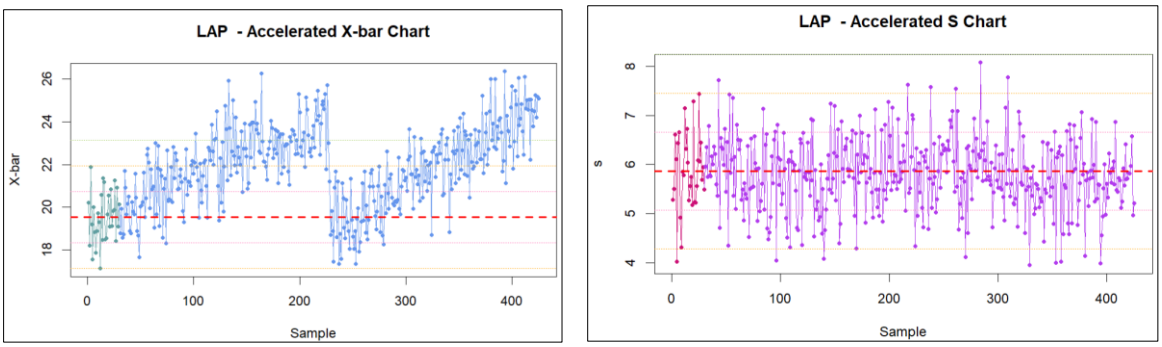

Figure 26: Product LAP Accelerated X-bar and S Chart

The control charts for product LAP reveal signs of instability. The X-bar chart shows several sample means drifting toward the outer control limits, suggesting potential shifts in the central tendency of the process. The S chart displays multiple spikes in standard deviation, with some points nearing or exceeding the control limits. This indicates inconsistent variability in delivery times.

Most product types remained within the control limits for the majority of the samples, indicating that delivery processes are stable and predictable. There are occasional fluctuations, but no obvious trends or patterns (runs above or below the centre line) were found, suggesting that no major process shifts occurred.
The S charts showed consistent variability across samples, confirming that the standard deviation remains under control. The $\bar{X}$ charts reflected stable process means, meaning that there were no systematic drifts in delivery times.

3.3. Process Capability Analysis

Process capability indices were calculated using the formulas:

$Cp = \frac{USL-LSL}{6\sigma}$ and $Cpk = \min\left(\frac{USL-\mu}{3\sigma}, \frac{\mu-LSL}{3\sigma}\right)$

The first 1000 delivery times per product type were used to compute process capability indices using limits given LSL = 0h and USL = 32h. The results are summarised below.

Table 5: Capability indices for different products

Product <chr>	Cp <dbl>	Cpu <dbl>	Cpl <dbl>	Cpk <dbl>	Capable <lgl>
SOF	51.805631	100.085401	3.525861	3.525861	TRUE
KEY	2.615043	1.676506	3.553580	1.676506	TRUE
CLO	2.694806	1.732206	3.657406	1.732206	TRUE
MOU	2.619220	1.671885	3.566555	1.671885	TRUE
MON	2.755586	1.767583	3.743590	1.767583	TRUE
LAP	2.754417	1.759318	3.749516	1.759318	TRUE

A process is considered capable if Cpk > 1.33. All product types have Cpk > 1.6 (see table 5), indicating that each process is highly capable of meeting the Voice of the Customer (VOC) requirement of delivery within 0-32h.
The product SOF shows extremely high Cp and Cpk, suggesting that its process variation is very small relative to the specification limits, and it is well-controlled. The remaining products (KEY, CLO, MOU, MON, LAP) also show strong capability with Cp values around 2.6 – 2.8 and Cpk

above 1.6, which means the delivery processes are consistent and centred within the specification range.

3.4. Out-of-Control Pattern Detection

Using SPC rules, potential out-of-control conditions were identified for each product type based on s > UCLs, longest stable run (±1σ range) and 4 consecutive $\bar{X}$ points beyond ±2σ.

Table 6: Samples with Process Control Issues

Product <chr>	s_out_upper3_first3 <chr>	s_out_upper3_last3 <chr>	s_max_consec_within1sigma <int>	x_out_2sigma_first3 <chr>	x_out_2sigma_last3 <chr>
SOF			16	361, 440, 456	463, 819, 824
KEY	89	89	20	365, 391, 392	710, 715, 738
CLO			27	309	309
MOU			15	433, 434, 847	434, 847, 848
MON			26	333	333
LAP			17		

The table above summarises the key process-control indicators for each product type. The table lists the first and last three sample points exceeding the upper control limits on both the s and $\bar{X}$ charts, as well as the maximum consecutive samples remaining within the ±1σ area. Overall, the process appears to be stable and well-controlled, since most product types show no samples exceeding the upper control limits and display relatively long runs within the ±1σ region, indicating consistent variability. Minor deviations were observed for the SOF and the KEY product types, where a few points exceeded the ±2σ limits on the $\bar{X}$ chart, suggesting short-term shifts in the process mean. However, no sustained or systematic violations were detected, implying that no immediate corrective action is required. Continuous monitoring is recommended to ensure these deviations do not develop.

The Statistical Process Control (SPC) analysis of the delivery times for all product types shows that the processes are generally stable, consistent, and capable. The $\bar{X}$ and S charts confirmed that initial process variation and averages were within control limits, and the accelerated simulation in section 3.2 demonstrated continued stability over time.

4. Risk, Data correction & Optimising for maximum profit

4.1. Type I (Manufacturer's) Error

A Type I error (α) is the likelihood of concluding that a process is out of control when it is actually operating correctly. In other words, the process is stable and in control, but the control chart gives a false alarm.

Assuming that each process is in control and normally distributed, the probabilities for the three rules from section 3.4 are as follows:

- Rule A is the probability that a single sample falls outside the ±3σ control limits, which is 0.001349898.
- Rule B is the probability of observing seven consecutive s-values within ±1σ, which is 0.6826895.
- Rule C is the probability of four consecutive $\bar{X}$ samples beyond ±2σ is 2.678772e-07.

The α (0.00135) indicates that approximately 1 in 740 samples may falsely trigger a process alarm, which is acceptable for high-value manufacturing processes.

These probabilities represent the likelihood of false signals; in other words, the instances where the control chart suggests a process change even though no real shift or increase in variation has occurred.

4.2. Type II (Consumer's) Error

A Type II error (β) is the likelihood of saying something is correct even though it is actually wrong. This means that the process has shifted, but the chart does not detect it.

For the bottle-filling process, the control limits are 25.011 L and 25.089 L with a nominal mean of 25.05 L. Unknown to us, the mean has shifted to 25.028 L, and the sample mean standard deviation has increased from 0.13 L to 0.017 L.

Using the new process parameters, z = -1.00, z = 3.59, giving (3.59) – (-1.00) = 0.8411.
Thus, there is an 84% chance that the cart will fail to detect this shift, and only a 16% chance that it will identify the process as out of control.

The β = 0.84 suggests an 84% chance of failing to detect small mean shifts; this emphasises the need to tighten control limits or increase sample size to enhance sensitivity.

This high β shows that the current chart limits are too wide for small mean shifts and would require tighter control or larger samples to improve detection.

4.3. Fixing Data Errors

Head Office have indicated that product types 11 to 60 were incorrectly labelled and priced in the product_Headoffice.csv dataset. The files were corrected by repeating the first 10 valid entries for each product type, replacing "NA" entries with the correct labels (for example, SOF, KEY, CLO, etc) and updating selling price and markup values to match the verified data in products_data.csv dataset.

The Category column in the products_data.csv dataset was also updated to be consistent with each ProductID label. The corrected files were saved as products_Headoffice2025.csv and product_data2025.csv.

Table 7: Sales summary

Category <chr>	TotalSalesValue <dbl>
Clothing	98715482
Keyboards	73499067
Laptops	1163889479
Monitors	578385570
Mouse	51219577
Software	66468485

After the correction of the head office dataset, the 2023 total sales decreased by 2.3%, revealing that earlier inconsistencies had increased performance estimates.
The total sales values changed slightly because the incorrect prices and markups had previously increased or decreased sales for several categories.

The revised dataset now provides a more accurate reflection of product performance and can be trusted to be used for future analyses.

As previously mentioned in section 2.1, there are differences in the product description and pricing as highlighted in Tables 2 and 3.
For example, ProductID SOF001 is described as “coral matt” with the selling price of R511.53 and a markup of 25.05% in the local product dataset. While in the head office dataset, it appears to be described as “coral silk” with the selling price of 521.72 and a markup of 15.65%. These discrepancies suggest inconsistencies between local and head office product records.

With the new and improved fixes in these datasets, we can see that these discrepancies have been fixed. The prices and descriptions of the product are the same in both datasets, as highlighted in Tables 8 and 9.

Table 8: Updated sample of Products Headoffice

	ProductID <chr>	Category <chr>	Description <chr>	SellingPrice <dbl>	Markup <dbl>
1	SOF001	Software	coral silk	511.53	25.05
2	SOF002	Software	black silk	505.26	10.43
3	SOF003	Software	burlywood marble	493.69	16.18
4	SOF004	Software	black marble	542.56	17.19
5	SOF005	Software	chartreuse sandpaper	516.15	11.01
6	SOF006	Software	cornflowerblue marble	478.93	16.99
7	SOF007	Software	blue marble	527.56	16.79
8	SOF008	Software	cornflowerblue marble	549.02	11.95
9	SOF009	Software	black bright	540.41	11.34
10	SOF010	Software	cornflowerblue matt	396.72	23.47

Table 9: Updated sample of Products Data

	ProductID <chr>	Category <chr>	Description <chr>	SellingPrice <dbl>	Markup <dbl>
1	SOF001	Software	coral matt	511.53	25.05
2	SOF002	Software	cyan silk	505.26	10.43
3	SOF003	Software	burlywood marble	493.69	16.18
4	SOF004	Software	blue silk	542.56	17.19
5	SOF005	Software	aliceblue wood	516.15	11.01
6	SOF006	Software	black silk	478.93	16.99
7	SOF007	Software	black bright	527.56	16.79
8	SOF008	Software	burlywood silk	549.02	11.95
9	SOF009	Software	azure sandpaper	540.41	11.34
10	SOF010	Software	chocolate sandpaper	396.72	23.47

5. Optimising the Dataset

The goal was to determine the optimal number of baristas that maximises daily profit while maintaining reliable service times.

The data of 2 coffee shops were given and analysed using the same profit optimisation model. The model uses individual service times and the corresponding barista counts to calculate expected daily customers, revenue, and cost, and then identifies the number of baristas that maximise the daily profit.

Brute-force and optimisation techniques were applied to the timeToServe.csv and timeToServe2.csv datasets. Each barista's mean service time was used to estimate the number of customers served, revenue, cost, and the resulting profit. Reliability was measured as the percentage of orders completed within 120 seconds.

Using both brute-force search and the optimise solver in R, the profit model was evaluated for staffing levels of 1 to 6 baristas. Increasing the number of baristas reduces service time but increases cost, and by doing so, it balances efficiency and service quality.

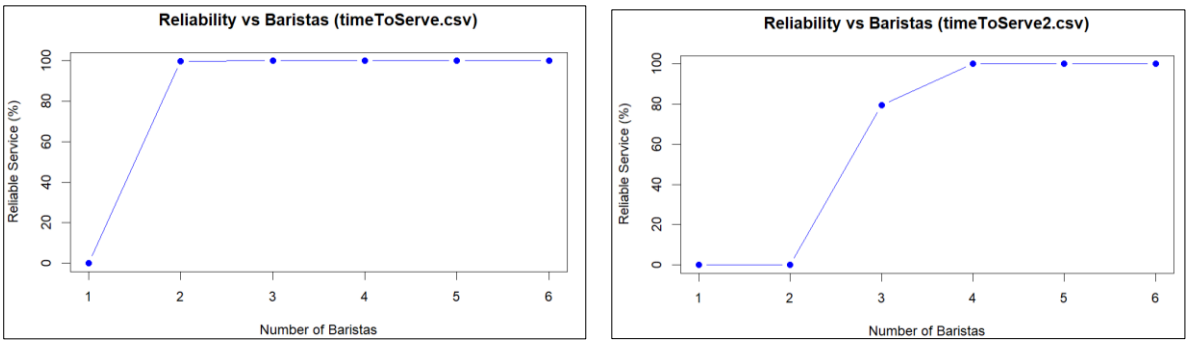

Figure 27: Reliability vs Number of Baristas for the 2 datasets

For the dataset timeToServe.csv, six baristas optimise the daily profit of R19902.66, while a continuous approximation indicates that five baristas would achieve a maximum profit of R16620.63. The reliability analysis indicates that 100% of customers received service within 120 seconds, demonstrating excellent performance and process control.

For the dataset timeToServe2.csv, five baristas optimise the daily profit of R4660.54; the continuous approximation reveals the same data of 5 baristas at a maximum profit of R4660.54. The reliability analysis also indicates that 100% of customers received service within 120 seconds, demonstrating excellent performance and process control.

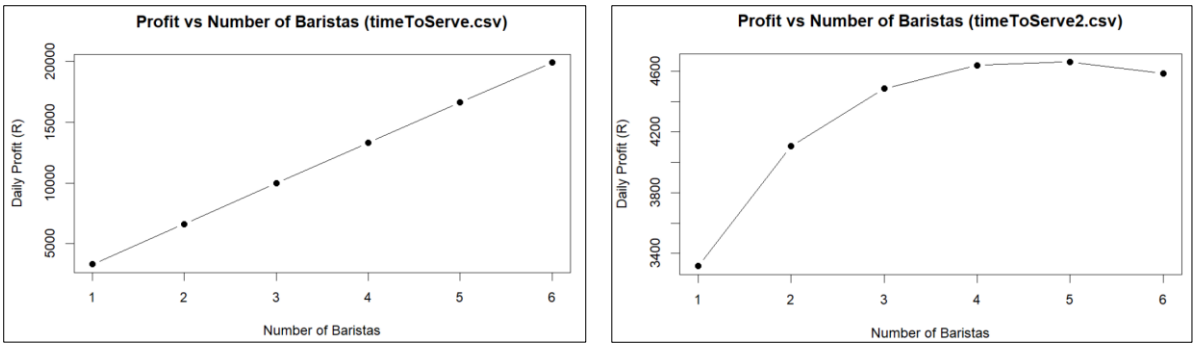

Figure 28: Profit vs Number of Baristas for the 2 datasets

A summary table comparing the two shops' optimal results, profits, and reliability percentages is shown below.

Table 10: Summary of the 2 shops' results

Shop <chr>	Optimal_Baristas <dbl>	Max_Profit <dbl>	Reliability <dbl>
Shop 1	6	19902.66	100
Shop 2	5	4660.54	100

6. DOE and MANOVA or ANOVA

6.1. ANOVA

In this section, an Analysis of Variance (ANOVA) was performed to determine whether there are significant differences between the mean values of treatments obtained from experimental data. The treatments represent four operational strategies tested for mean delivery time.

6.2. Results discussion

The hypothesis statements are as follows:
H_0, the null hypothesis, means that there is no significant difference between the treatment means. H_1 means that at least one treatment mean differs significantly.

From the ANOVA table, the calculated F-value is 14.69, with a corresponding p-value < 0.05. Since the p-value is smaller than the 0.5 significance level, the null hypothesis, H_0, is rejected. This indicates that at least one treatment mean differs significantly from the others. Thus, there is a significant difference in the mean responses across four treatment groups.

Table 11: ANOVA results

Source	SS	DoF	MS	F_0	P-value
Treatment	460.97	3	153.66	14.69	0.0000
Error	2050.17	196	10.46	—	—
Total	2511.13	199	—	—	—

The calculated Least Significant Difference (LSD) was 1.276. We can then see that Treatments 1 and 2 have a difference smaller than the LSD value, indicating no significant difference. However, the differences between Treatments 1 and 3, 1 and 4, 2 and 3, and 2 and 4 all exceed the LSD values, confirming significant differences between these treatments.

Table 12: LSD and Treatment Means

Treatment	Mean
1	20.463
2	20.949
3	22.851
4	24.258

The graph supports the numerical findings: Treatments 3 and 4 show visibly higher mean responses compared to Treatments 1 and 2. This pattern reinforces the conclusion that the treatment factor had a significant effect on the measured response.

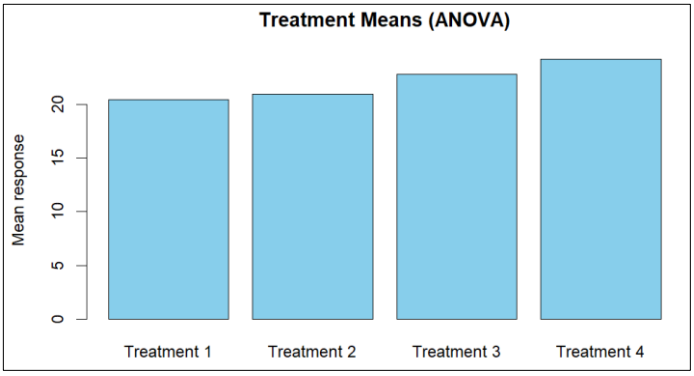

Figure 29: Bar Chart of the Treatment Means

Overall, the analysis confirms that the treatments have great significant differences, especially between the higher and lower mean groups.

7. Reliability of Service

7.1. Reliable service

The number of workers on duty per day can be modelled using a binomial distribution with parameters n = 16, which is the total number of possible workers, and the probability that an individual worker is present, p. We then get p = 0.976, this estimate was obtained by equating observed frequencies to the theoretical binomial probabilities and averaging the resulting p-values, weighted by their frequencies.

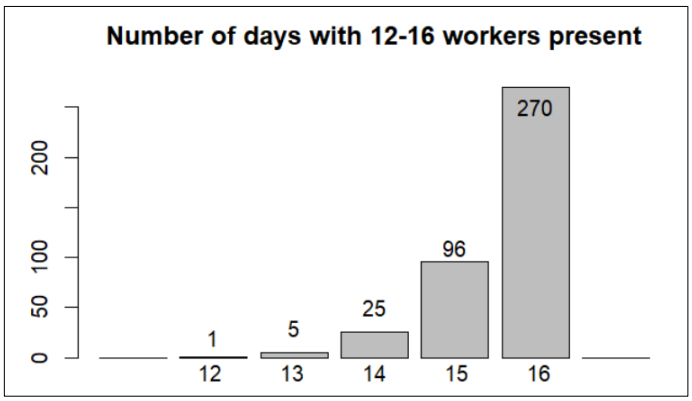

Figure 30: Given the Number of days and the Number of workers

Using this model, the probability of having at least 15 workers (the minimum required for reliable service) was found to be P(X≥15) = 0.918.
This equates to an expected 345 days of reliable service per year, 0.918 x 365 = 345 days.

The fitted binomial distribution closely matched the observed histogram, confirming that a binomial distribution applies to the model of worker attendance.

7.2. Profit optimisation

The reliability analysis was extended to determine whether hiring additional workers would reduce expected financial losses due to unreliable service.
For each possible number of total workers n, the expected number of unreliable days (fewer than 15 workers) and the corresponding loss were calculated. The results are shown below.

Table 13: Optimisation of Profit

```
n = 16 - unreliable days: 19.9 - expected loss: R 398100
n = 17 - unreliable days: 2.6 - expected loss: R 52164
n = 18 - unreliable days: 0.3 - expected loss: R 5491
```

The analysis shows that increasing the workforce from 16 to 17 workers drastically improves reliability, reducing expected annual losses by approximately R345,936.
Adding an 18th worker provides only a small improvement (a further reduction in loss of R46,673), which may not justify the extra salary cost.

Therefore, the optimal staffing level, considering cost versus reliability, is likely 17 workers.

8. Conclusion

In conclusion, this project successfully applied advanced data analysis and statistical quality control methods to assess and optimise performance. The analyses revealed that while the company's delivery processes were generally stable and capable, certain products showed early signs of variability, indicating the need for ongoing monitoring. Process capability indices confirmed a strong alignment with customer requirements, and corrective data procedures ensured consistency between local and head office datasets.

The reliability and profit optimisation exercises highlighted the importance of balancing operational efficiency with resource allocation. The optimal number of baristas was identified through achieving high service reliability while maximising daily profit. Furthermore, the ANOVA results confirmed statistically significant treatment differences.

Overall, this report demonstrates the practical value of engineering statistics in improving decision-making, ensuring quality assurance, and optimising profitability.

9. References

ChatGPT, 2025. *ChatGPT Online AI Tool*. [Online]
Available at: https://chatgpt.com/
[Accessed 5 October 2025].

DataCamp, 2024. *Introduction to ggplot2 in R: Data Visualization for Beginners*. [Online]
Available at: https://www.datacamp.com/tutorial/ggplot2-tutorial-r
[Accessed 20 October 2025].

GeeksforGeeks, 2024. *Statistical Process Control (SPC) and Control Charts*. [Online]
Available at: https://www.geeksforgeeks.org/statistical-process-control-spc/
[Accessed 19 October 2025].

ISO, 2015. *ISO 22514-2:2015 – Statistical Methods in Process Management – Capability and Performance*. [Online]
Available at: https://www.iso.org/standard/53793.html
[Accessed 20 October 2025].

JMP, 2024. *What is Process Capability (Cp, Cpk)?*. [Online]
Available at: https://www.jmp.com/en_za/statistics-knowledge-portal/process-capability.html
[Accessed 21 October 2025].

Minitab, 2024. *Understanding Type I and Type II Errors*. [Online]
Available at: https://blog.minitab.com/en/statistics-and-quality-data-analysis/type-i-and-type-ii-errors
[Accessed 21 October 2025].

R Core Team, 2024. *R: A Language and Environment for Statistical Computing*. [Online]
Available at: https://www.r-project.org/
[Accessed 5 October 2025].

SimplyStatistics, 2023. *How to Interpret p-values and Statistical Significance in R*. [Online]
Available at: https://simplystatistics.org/2023/07/11/p-values-and-significance/
[Accessed 19 October 2025].

Stellenbosch University, 2025. *STEMLearn*. [Online]
Available at: https://stemlearn.sun.ac.za/course/view.php?id=1492#section-11
[Accessed 15 October 2025].

Statology, 2024. *How to Perform and Interpret ANOVA in R*. [Online]
Available at: https://www.statology.org/anova-in-r/
[Accessed 20 October 2025].

Wickham, H., 2023. *ggplot2: Elegant Graphics for Data Analysis*. [Online]
Available at: https://ggplot2.tidyverse.org/
[Accessed 22 October 2025].