Thomas Bothwell

27210510

ECSA project-Quality assurance

Stellenbosch University

Content

Introduction:

This project focuses on applying data analytics and statistical quality control techniques to evaluate and improve business performance across different operational areas. Using R, multiple datasets were analysed to extract insights into customer behaviour, product sales, and process efficiency. The project includes descriptive statistics, revenue analysis, and statistical process control (SPC) to monitor process stability, as well as an optimisation study aimed at maximising profit for a service-based operation. Through these analyses, data-driven decisions are demonstrated as key tools for ensuring process reliability, profitability, and continuous improvement within an engineering and business context.

Part 1.2: Data analysis

1. **Introduction**

The following is a data analytical report of a company's sales insight. Four datasets need to be analyzed and they are the company's: customers, products at the branch, Sales for 2022 and 2023 and A master product list from the company's HQ.

2. **Inspection of data:**

2.1 sales2022and2023

- Dimensions 100,000 × 9
- Structure: Transaction-level dataset covering 2022–2023 sales.
- Column Names (colnames): CustomerID, ProductID, Quantity, orderTime, orderDay, orderMonth, orderYear, pickingHours, deliveryHours.
- Insight: This is the central fact table, capturing all sales transactions along with order timing and fulfilment details.

2.2 products_data

- Dimensions (dim): 60 × 5
- Structure: product catalogue.
- Column Names ProductID, Category, Description, SellingPrice, Markup.
- Insight: Contains detailed information about products at branch level, including descriptions, categories, and pricing margins.

2.3 products_Headoffice

- Dimensions (dim): 360 × 5
- Structure: Official head office product catalogue.
- Column Names (colnames): ProductID, Category, Description, SellingPrice, Markup.
- Insight: Provides the standardized reference dataset for product details, ensuring alignment between branch-level and head office product information.

2.4 customer_data

- Dimensions: 5,000 × 5
- Structure: Customer dataset with demographic and geographic details.
- Column Names: CustomerID, Gender, Age, Income, City.
- Insight: Captures customer attributes, enabling segmentation and profiling for sales and marketing analysis.

3. Summary of data

3.1 Sales:

	vars <dbl>	n <dbl>	mean <dbl>	sd <dbl>	median <dbl>	trimmed <dbl>	mad <dbl>	min <dbl>	max <dbl>
CustomerID*	1	1e+05	2492.34	1444.58	2503.00	2491.19	1862.15	1.00	5000.00
ProductID*	2	1e+05	32.44	18.03	35.00	32.82	23.72	1.00	60.00
Quantity	3	1e+05	13.50	13.76	6.00	11.46	5.93	1.00	50.00
orderTime	4	1e+05	12.93	5.50	13.00	13.12	5.93	1.00	23.00
orderDay	5	1e+05	15.50	8.65	15.00	15.50	10.38	1.00	30.00
orderMonth	6	1e+05	6.45	3.28	6.00	6.45	4.45	1.00	12.00
orderYear	7	1e+05	2022.46	0.50	2022.00	2022.45	0.00	2022.00	2023.00
pickingHours	8	1e+05	14.70	10.39	14.05	13.54	6.92	0.43	45.06
deliveryHours	9	1e+05	17.48	10.00	19.55	17.78	8.90	0.28	38.05

	skim_variable <chr>	n_missing <int>	complete_rate <dbl>	min <int>	max <int>	empty <int>	n_unique <int>	whitespace <int>
1	CustomerID	0	1	7	8	0	5000	0
2	ProductID	0	1	6	6	0	60	0

	skim_variable <chr>	n_missing <int>	complete_rate <dbl>	mean <dbl>	sd <dbl>	p0 <dbl>	p25 <dbl>	p50 <dbl>
1	Quantity	0	1	13.50347	13.7601316	1.0000000	3.000000	6.000
2	orderTime	0	1	12.93230	5.4951268	1.0000000	9.000000	13.000
3	orderDay	0	1	15.49683	8.6465055	1.0000000	8.000000	15.000
4	orderMonth	0	1	6.44813	3.2834460	1.0000000	4.000000	6.000
5	orderYear	0	1	2022.46273	0.4986115	2022.0000000	2022.000000	2022.000
6	pickingHours	0	1	14.69547	10.3873345	0.4258889	9.390833	14.055
7	deliveryHours	0	1	17.47646	9.9999440	0.2772000	11.546000	19.546

The sales dataset is well-structured and complete, with no missing values. It captures the full span of customer and product interactions across 2022 and 2023. Transaction quantities show a strong right skew, where most sales involve only a few units, but some large bulk orders raise the mean far above the median. Operational measures such as picking and delivery hours also display wide variation, suggesting potential inefficiencies or differences in order types and logistics. Customer and product IDs map consistently to expected ranges (5,000 customers and 60 products), making this dataset reliable for integration with customer and product reference tables. Overall, it provides a solid foundation for analysing both sales performance and operational efficiency, but variability in order sizes and process times should be carefully considered in downstream analysis

3.2 Products_Data

	vars	n	mean	sd	median	trimmed	mad	min	max
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
ProductID*	1	60	30.50	17.46	30.50	30.50	22.24	1.00	60.00
Category*	2	60	3.50	1.72	3.50	3.50	2.22	1.00	6.00
Description*	3	60	16.40	10.08	16.00	16.21	13.34	1.00	35.00
SellingPrice	4	60	4493.59	6503.77	794.18	3189.25	525.72	350.45	19725.18
Markup	5	60	20.46	6.07	20.34	20.51	7.31	10.13	29.84

5 rows | 1-10 of 13 columns

	skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
	<chr>	<int>	<dbl>	<int>	<int>	<int>	<int>	<int>
1	ProductID	0	1	6	6	0	60	0
2	Category	0	1	5	18	0	6	0
3	Description	0	1	9	21	0	35	0

	skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75
	<chr>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	SellingPrice	0	1	4493.59283	6503.770150	350.45	512.1825	794.185	6416.6600
2	Markup	0	1	20.46167	6.072598	10.13	16.1400	20.335	25.7075

The products dataset is compact, with 60 products described across five attributes. The identifiers are consistent, ranging cleanly from 1 to 60. Categories and descriptions are complete with no missing values, though some overlap exists with only 18 unique categories and 35 unique product descriptions, indicating multiple products share categories or similar naming. Selling prices vary widely, from as low as about 350 to nearly 20,000, with a mean near 4,500. This spread suggests a mix of lower-value, fast-moving products and high-value items that significantly influence average revenue. Markup values cluster more tightly (mean ≈ 20.5, range 10–30), reflecting stable pricing margins across the product set. Overall, the dataset is clean and consistent, offering a reliable foundation for analysing product-level performance and profitability, with particular attention to price segmentation between low- and high-value goods.

3.3 Product_Headoffice

	vars	n	mean	sd	median	trimmed	mad	min	max
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
ProductID*	1	360	69.39	23.22	72.00	71.89	22.24	1.00	110.00
Category*	2	360	3.50	1.71	3.50	3.50	2.22	1.00	6.00
Description*	3	360	30.69	17.32	29.50	30.77	22.98	1.00	60.00
SellingPrice	4	360	4410.96	6463.82	797.22	3054.23	515.75	290.52	22420.14
Markup	5	360	20.39	5.67	20.58	20.43	6.66	10.06	30.00

	skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
	<chr>	<int>	<dbl>	<int>	<int>	<int>	<int>	<int>
1	ProductID	0	1	5	6	0	110	0
2	Category	0	1	5	18	0	6	0
3	Description	0	1	9	24	0	60	0

3 rows

	skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75
	<chr>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	SellingPrice	0	1	4410.9619	6463.822788	290.52	495.9375	797.215	5843.332
2	Markup	0	1	20.3855	5.665949	10.06	15.8400	20.580	24.845

The head office product catalogue is larger and more comprehensive, containing 360 records across five attributes. Product IDs range from 1 to 110, confirming a broader scope than the branch-level dataset. Categories (18 unique) and descriptions (60 unique) indicate a well-structured but somewhat overlapping classification system. Selling prices cover a wide spectrum, from roughly 290 to over 22,000, with a mean near 4,400 and a median below 800. This again suggests that while many products are low- to mid-priced, a small number of high-value items heavily influence the average. Markup percentages are stable (mean ≈ 20.4, range 10–30), reflecting consistent company-wide pricing policies. The data is complete with no missing values, making it a strong reference for validating branch-level product information. Overall, this dataset provides a standardized foundation for aligning local product lists with corporate-level catalogues and ensuring consistent reporting.

3.4 Customers

Description: df [5 × 13]

	vars <dbl>	n <dbl>	mean <dbl>	sd <dbl>	median <dbl>	trimmed <dbl>	mad <dbl>	min <dbl>	max <dbl>
CustomerID*	1	5000	2500.50	1443.52	2500.5	2500.50	1853.25	1	5000
Gender*	2	5000	1.56	0.58	2.0	1.52	1.48	1	3
Age	3	5000	51.55	21.22	51.0	50.88	26.69	16	105
Income	4	5000	80797.00	33150.11	85000.0	81665.00	37065.00	5000	140000
City*	5	5000	3.99	2.00	4.0	3.99	2.97	1	7

5 rows | 1-10 of 13 columns

	skim_variable <chr>	n_missing <int>	complete_rate <dbl>	min <int>	max <int>	empty <int>	n_unique <int>	whitespace <int>
1	CustomerID	0	1	7	8	0	5000	0
2	Gender	0	1	4	6	0	3	0
3	City	0	1	5	13	0	7	0

3 rows

	skim_variable <chr>	n_missing <int>	complete_rate <dbl>	mean <dbl>	sd <dbl>	p0 <dbl>	p25 <dbl>	p50 <dbl>	p75 <dbl>
1	Age	0	1	51.5538	21.2161	16	33	51	68
2	Income	0	1	80797.0000	33150.1067	5000	55000	85000	105000

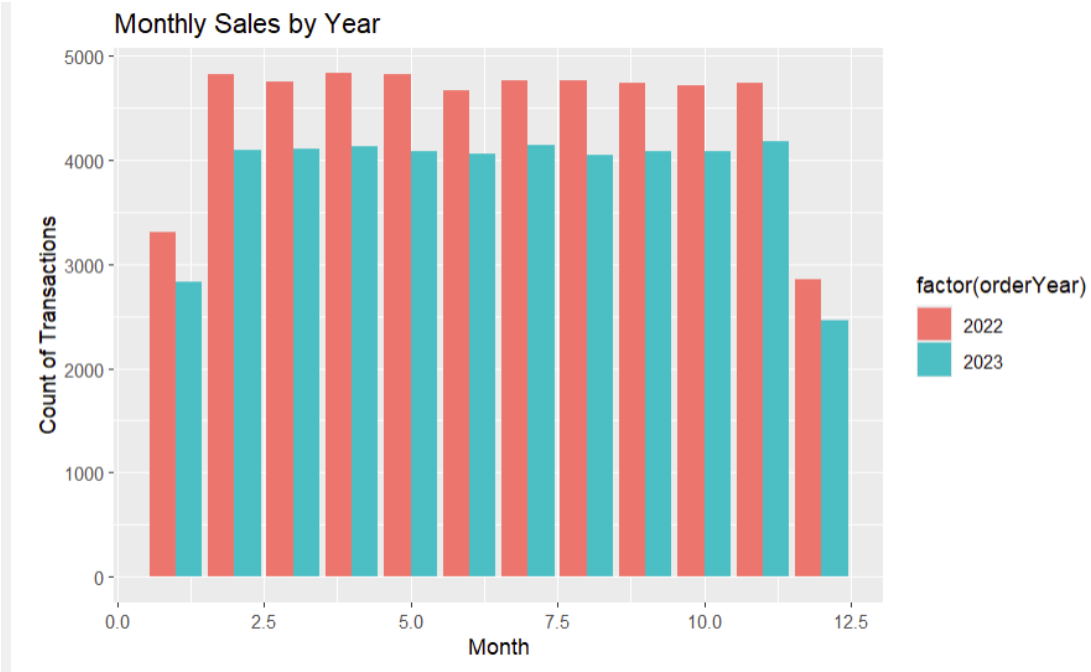
The customer dataset includes 5,000 individuals with complete demographic and geographic information. Gender is coded into three categories, with a balanced distribution but one group being slightly larger. Ages range from 16 to 105, with an average of 52 years, showing a broad customer base that spans young to elderly buyers. Income varies widely, from 5,000 to 140,000, with a mean around 80,800 and a median of 85,000. This indicates a mix of lower-income and higher-income customers, but with a central tendency toward middle-to-upper income brackets. City is coded into seven categories, evenly distributed, representing different geographic markets. No missing values are present, and the data appears consistent and reliable. Overall, the dataset enables rich segmentation of customers by age, income, gender, and location, making it highly valuable for targeted marketing and sales strategy.

4. Handling missing values

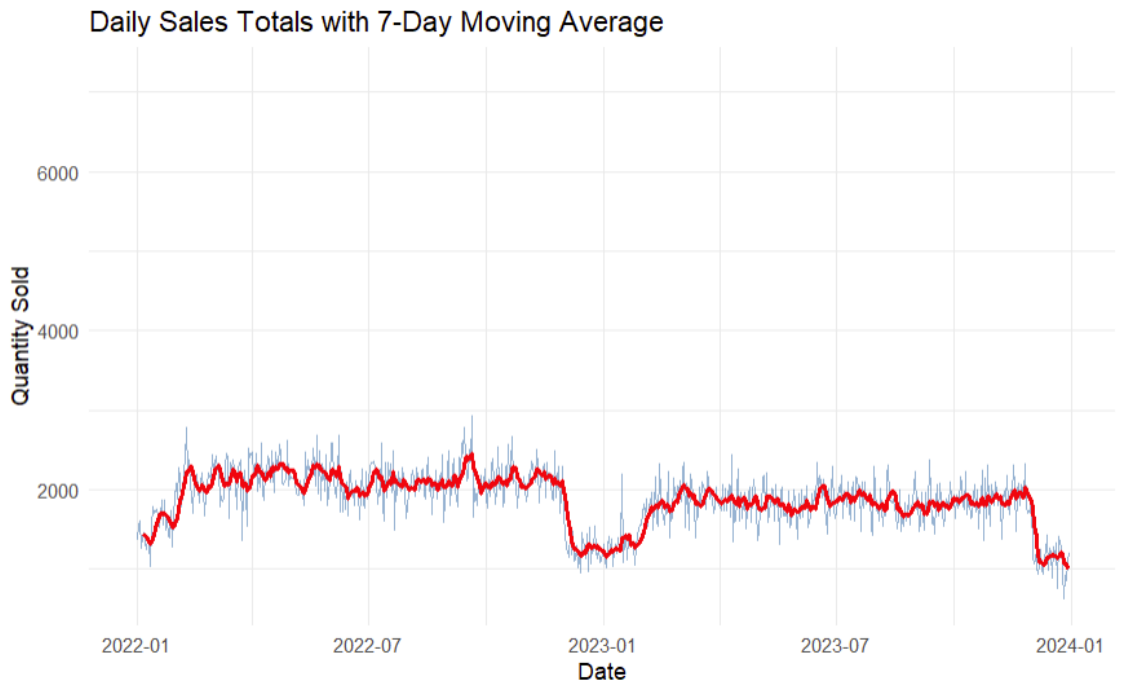
The inspection confirmed that no missing values are present in any of the datasets. As a result, no imputation, removal, or filtering of records was required.

5. Data Visualization and Exploratory Data Analysis (EDA)

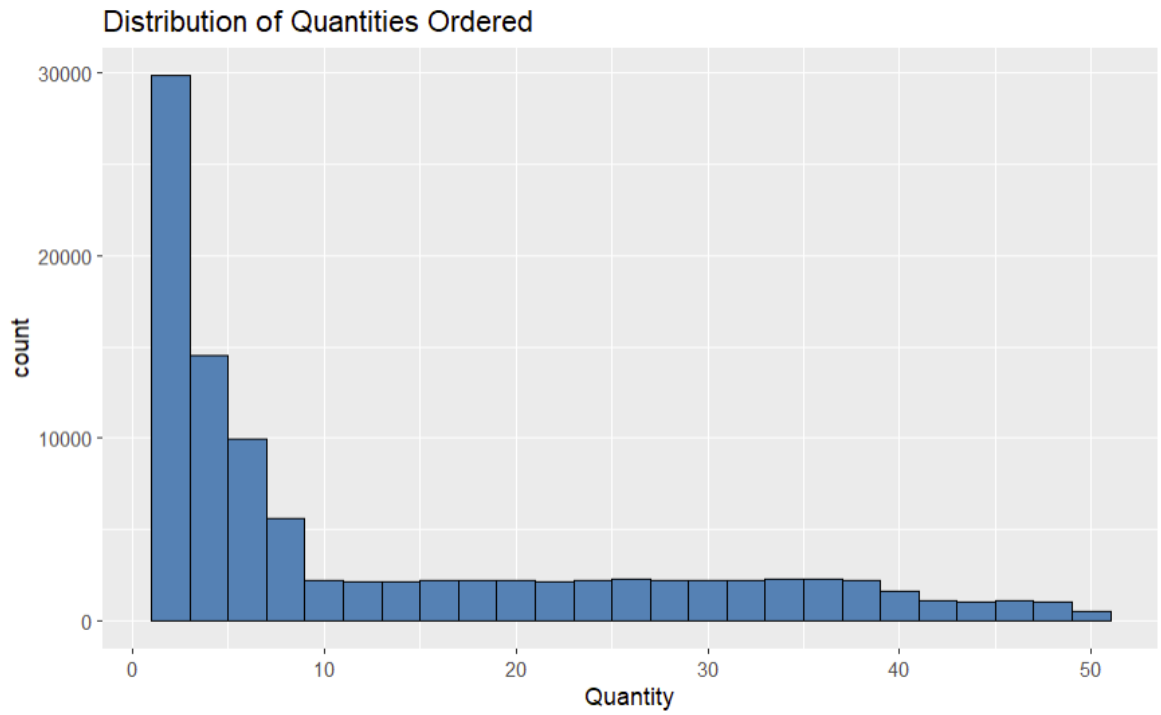
5.1 Monthly sales/year



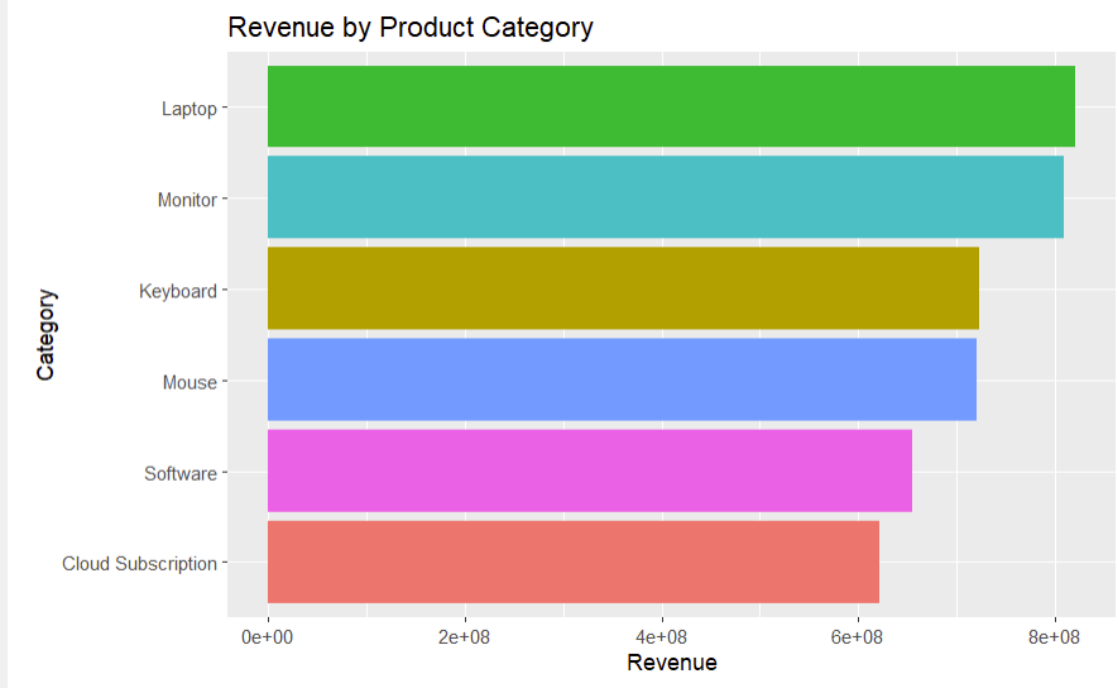
5.2


Daily Sales Totals with 7-Day Moving Average
Quantity Sold
6000
4000
2000
2022-01
2022-07
2023-01
2023-07
2024-01
Date

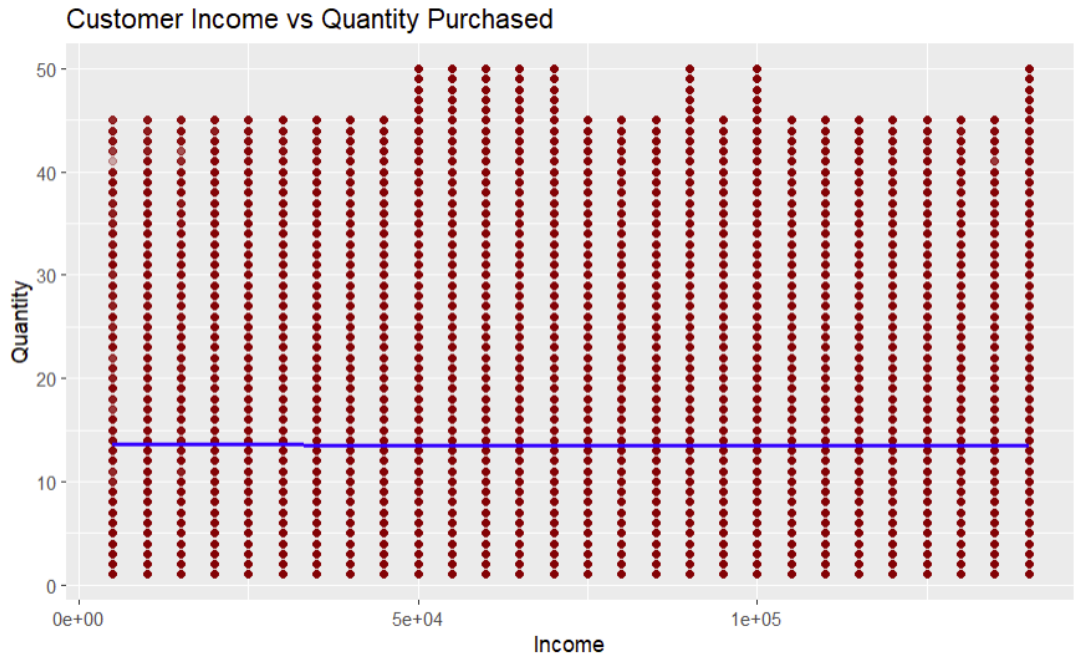
5.3


Distribution of Quantities Ordered
count
30000
20000
10000
0
0
10
20
30
40
50
Quantity

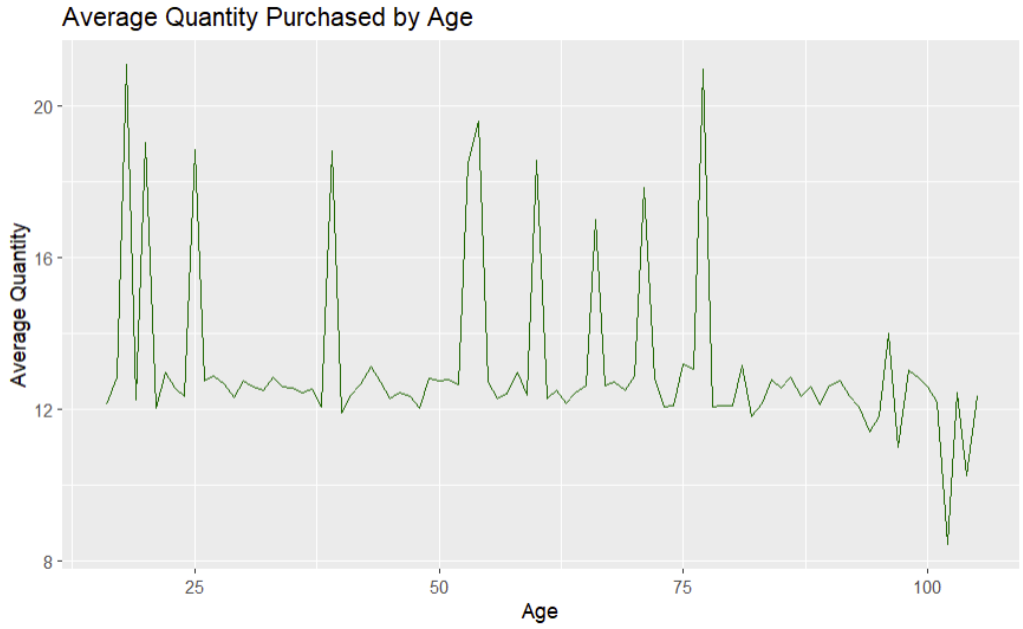
5.4


Revenue by Product Category
Laptop
Monitor
Keyboard
Mouse
Software
Cloud Subscription
Category
0e+00
2e+08
4e+08
6e+08
8e+08
Revenue

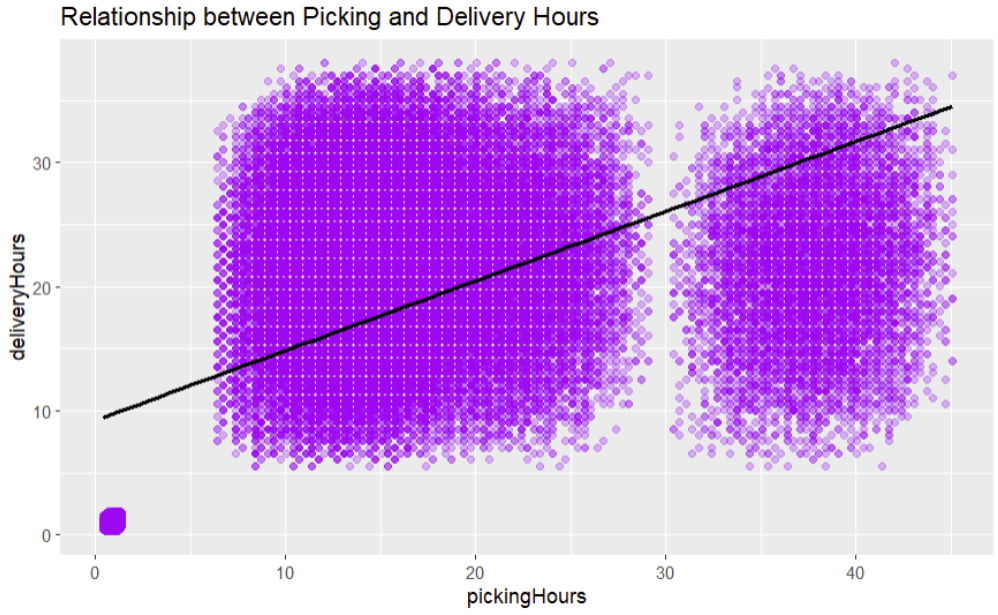
5.5


Customer Income vs Quantity Purchased
50
40
30
20
10
0
Quantity
0e+00
5e+04
1e+05
Income

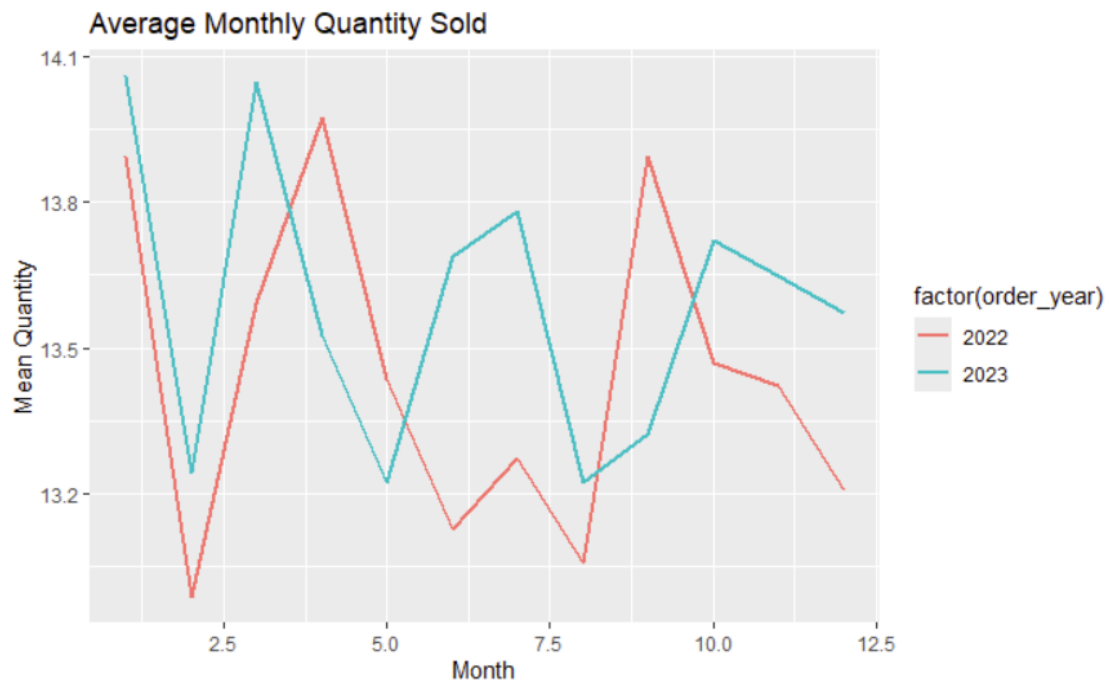
5.6


Average Quantity Purchased by Age
Average Quantity
20
16
12
8
25
50
75
100
Age

5.7


Relationship between Picking and Delivery Hours
deliveryHours
30
20
10
0
0
10
20
30
40
pickingHours

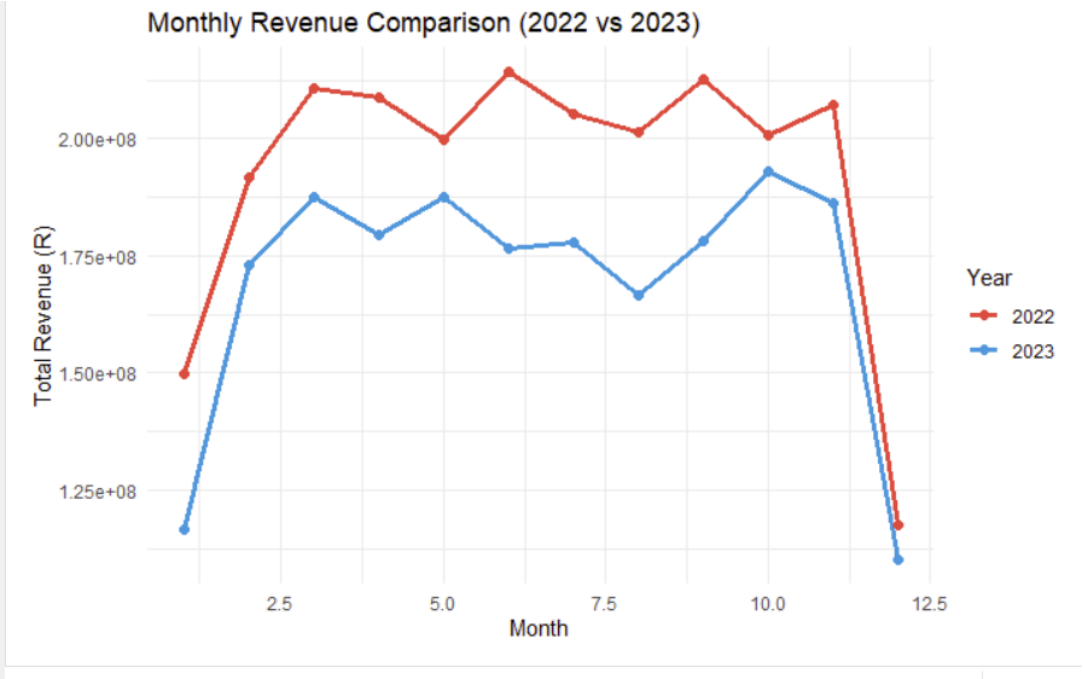
5.8


Average Monthly Quantity Sold
Mean Quantity
14.1
13.8
13.5
13.2
2.5
5.0
7.5
10.0
12.5
Month
factor(order_year)
2022
2023

5.9


quantity
picking_hours
delivery_hours
quantity
picking_hours
delivery_hours
1.00
0.00
0.00
0.00
1.00
0.58
0.00
0.58
1.00
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

5.10



6. Observations and comments

Observations

1. **Sales Performance:**
    - Transaction volumes are consistent throughout most of 2022, with a noticeable dip at the start and end of each year.
    - The 7-day moving average confirms steady sales with minor seasonality but highlights short periods of demand spikes.
2. **Order Quantities:**
    - Most transactions involve small quantities (median = 6 units), but a few large orders skew the distribution upward.
    - This indicates a mix of small retail customers and occasional bulk buyers.
3. **Product Categories:**
    - High revenues are concentrated in categories such as **Laptops and Monitors**, while Software and Cloud Subscriptions lag.
    - Pricing is highly varied, but markups remain stable (10–30%), reflecting consistent pricing policies.

4. **Customer Segments:**
    - The customer base spans a wide age range (16–105 years) with an average around 52 years.
    - Income levels are skewed toward middle- to high-income brackets, and geographic representation is spread across 7 cities.
    - However, the “Income vs Quantity” relationship shows that high income does not necessarily translate into higher purchase volumes.
5. **Operational Metrics:**
    - Picking and delivery hours are positively correlated but display wide variability.
    - Some clusters of outliers suggest inefficiencies or inconsistencies in the order fulfilment process.
6. **Sales decrease**
    - There is an obvious decrease in the number of sales from 2022 to 2023
    - This indicates performance decrease

7. **Revenue by product**
    - Revenue generated from different products are evenly spread, but laptops and monitors generate the most revenue.
    - This could be caused by slightly lower quantities sold or reduced average selling prices.
8. **Correlation**
    - The operational efficiency appears decoupled from order size but internally connected between the picking and delivery phases.
    - Improving efficiency in the picking stage could positively influence delivery performance, since these two are moderately correlated.
9. **Seasonal trends**
    - Both years exhibit similar seasonal patterns, with noticeable fluctuations in monthly sales volumes rather than a steady trend.
    - The data indicates that revenue peaks occur around April and October, while December consistently experiences a drop, likely due to holiday shutdowns or reduced operations.

Part 3: Statistical Process Control (SPC)

This section presents the Statistical Process Control (SPC) analysis of the delivery times for each product type from a 2026 and 2027 sales data set. The purpose was to evaluate whether each delivery process is stable and capable of meeting the customer's Voice of the Customer requirement of 0 ≤ delivery time ≤ 32 hours. The data were ordered chronologically by year, month, day, and order time, and samples of 24 deliveries were formed for each product type.

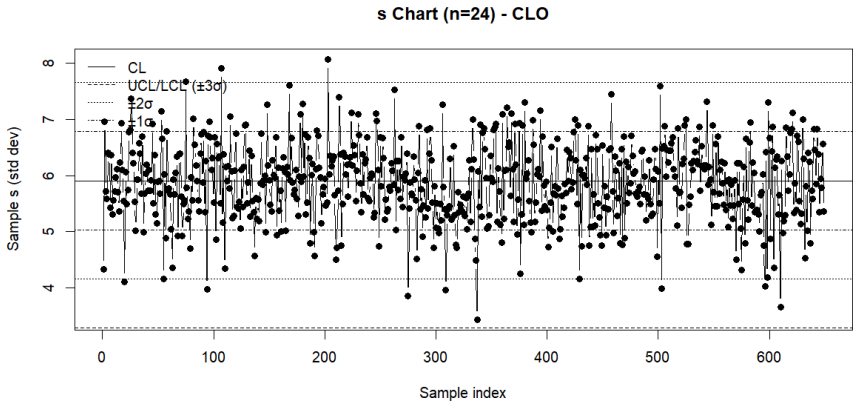
3.1) Initialisation of the control charts

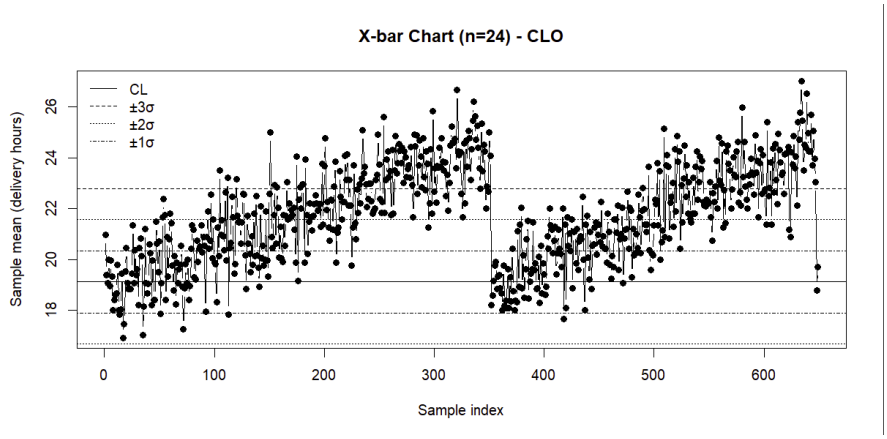
The first 30 samples of 24 deliveries for each product type were used to initialise the $\bar{X}$– and s– charts. These Phase I samples were used to calculate the centre lines and the ±1σ, ±2σ, and ±3σ control limits. There are 6 product types meaning that there are 12 graphs.

Example of product CLO s-chart and x-chart.

3.2 Interpretation of given data:

Product CLO:

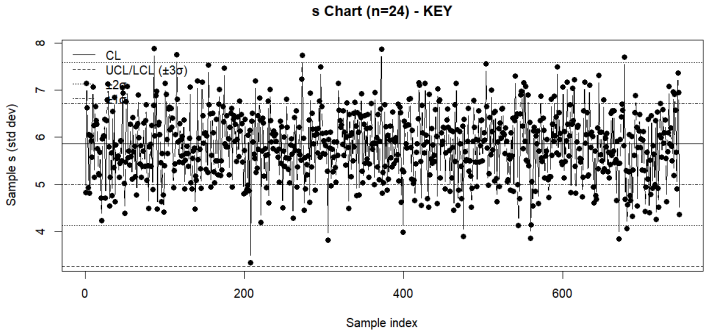




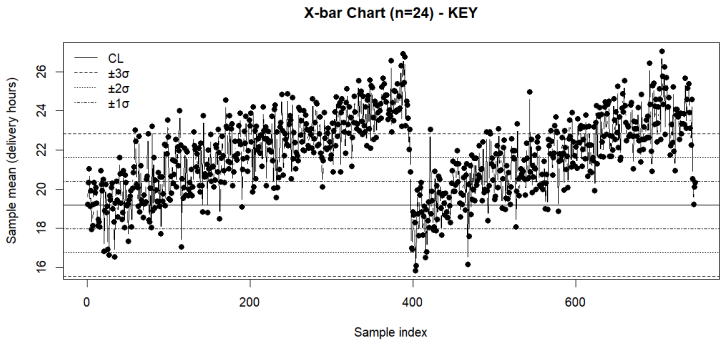
The S-chart (top) indicates high variation with many points exceeding the LCL/UCL boundaries.

The X–chart (bottom) shows the subgroup means of delivery times across consecutive samples. The centre line represents the average mean delivery time, with 1σ, 2σ, and 3σ control limits indicated. A gradual upward trend in the subgroup means can be observed between samples 0–250 and again after 450, suggesting a slow increase in mean delivery times over time. Despite these trends, most points remain within the ±3σ limits, meaning the process is still statistically stable but showing a shift in the mean.

Product Key



.

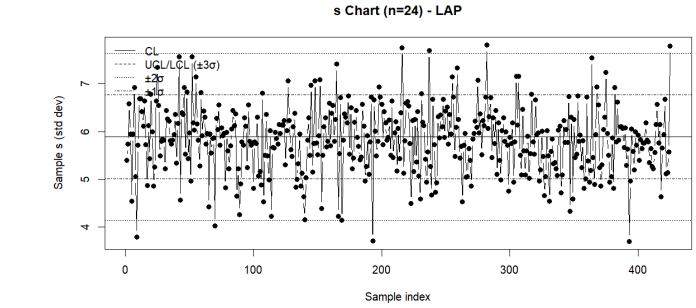


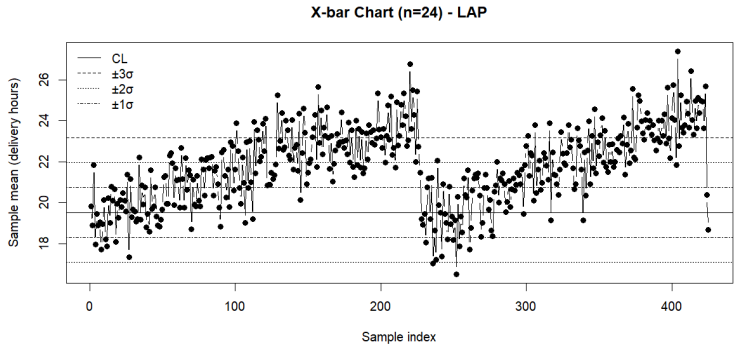
The S chart shows that process variation remains stable over time, with most subgroup standard deviations staying within the control limits. There are no clear trends or runs outside the ±3 standard deviations limits, which suggests that the variability of the KEY product's delivery times is under control. This indicates consistent performance in terms of process spread.

However, the X-bar chart reveals a visible trend and potential shifts in the process mean. The sample means gradually increase and then drop sharply around the midpoint before climbing again. This pattern indicates a change in central tendency, possibly due to an operational adjustment, seasonal influence, or process drift over time. Although most points remain within control limits, the presence of systematic patterns rather than random fluctuation suggests that the process mean is not fully stable, even if variability is.

Overall, the process for the KEY product is in control in terms of variation (S chart) but shows instability in the mean (X-bar chart), which should be investigated to identify underlying causes such as workload changes, scheduling shifts, or equipment performance differences

Product LAP:



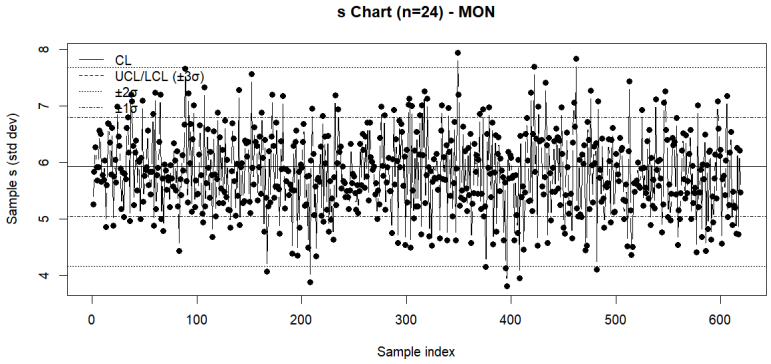


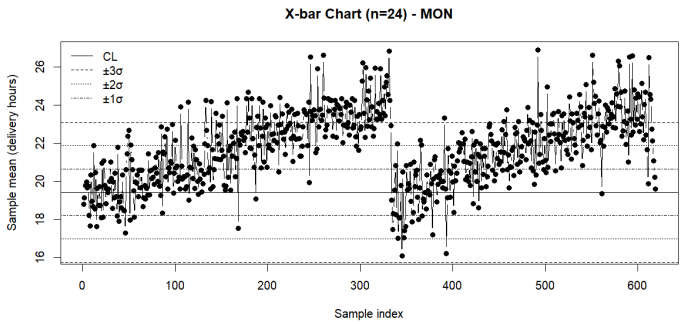
The S chart shows that process variation is mostly consistent, with only minor fluctuations around the control limits. This suggests that variation in delivery times is generally stable.

However, the X-bar chart indicates several samples with means exceeding the upper 3σ control limit, signalling that the process mean is not stable. There is also a clear centre jump where the average delivery time shifts noticeably upward, followed by recurring out-of-control points.

Overall, the LAP process exhibits stable variation but poor mean control, with frequent violations beyond ±3σ. This suggests that external factors or operational inconsistencies are affecting delivery performance and that process adjustment is required.

Product MON:



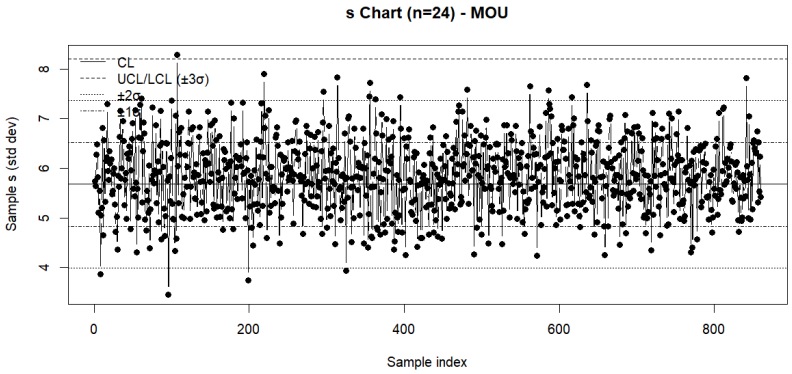


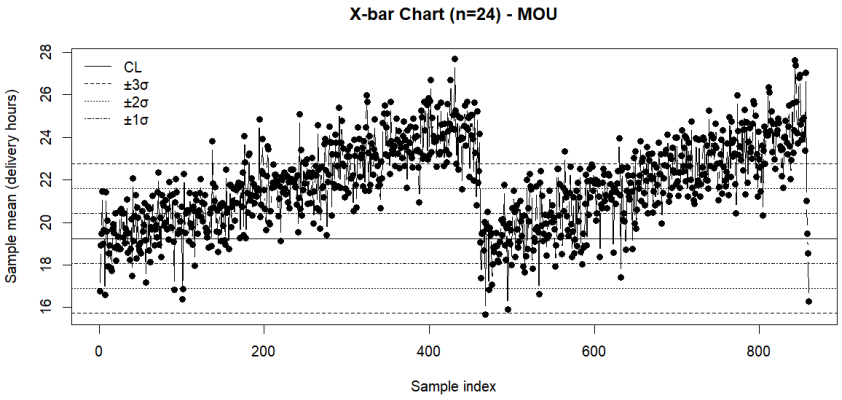
The s chart shows that delivery-time variation remained consistent within the ±3 standard deviation limits, indicating a stable and predictable process. No major out-of-control points were detected, suggesting that variation is driven mainly by common causes.

The X-bar chart, however, reveals a gradual upward trend in the mean delivery time. Although the process remains statistically in control, the mean has shifted over time, indicating a loss of efficiency. This trend suggests systematic issues—such as increased workload or scheduling delays—that should be investigated.

Overall, the process variation is stable, but the mean performance has drifted upward. Process adjustments are needed to recentre the system and maintain delivery-time targets.

Product MOU:

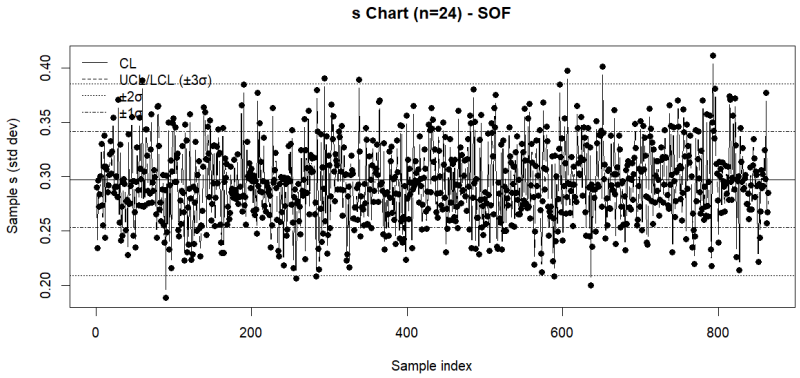


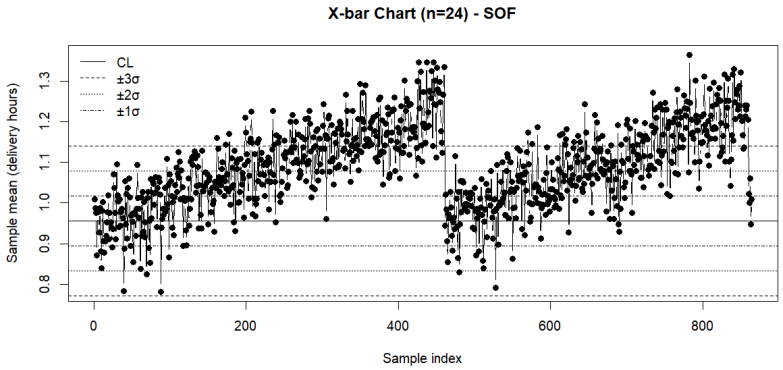


The S chart shows that process variation is generally consistent, with most points remaining within the control limits. This indicates that the spread of delivery times is stable, and the process variability is under control.

The X-bar chart shows a noticeable centre jump and several samples falling outside the ±3σ limits, particularly in the middle section of the timeline. This suggests that while variation is stable, the process mean shifts significantly over time, likely due to operational or scheduling changes.

Product SOF:





The S chart shows a highly stable process with very little variation in standard deviation. All points fall well within the control limits, indicating that process spread is consistent and tightly controlled.

In the X-bar chart, there is a clear centre jump midway through the data and some minor clustering near the upper control limits, but no significant points exceed ±3σ. This suggests that while the process mean shifts slightly over time, overall performance remains within acceptable control limits.

In summary, the SOF process demonstrates excellent stability and capability, with both the mean and variation well maintained throughout the period.

Below is a summary of the process capability of each product:

Product Type	Mean (μ)	Std Dev (s)	Cp	Cpu	Cpl	Cpk	Capable?
CLO	19.226	5.9408	0.8977	0.7167	1.0788	0.7167	No
KEY	19.276	5.8152	0.9171	0.7294	1.1049	0.7294	No
LAP	19.6135	5.9585	0.8950	0.6929	1.0972	0.6929	No
MON	19.41	5.9989	0.8891	0.6959	1.0785	0.6959	No
MOU	19.2975	5.8276	0.9152	0.7266	1.1038	0.7266	No
SOF	0.9554	0.2941	18.1352	35.1876	1.0829	1.0829	Yes

A process is considered capable if Cp and Cpk is larger than 1. Unfortunately, only product SOF is consistently meeting the delivery time specification of 0<x<32 hours. This is due to high variation in all product delivery times except product SOF. Thus, only SOF meets the VOC.

3.3) Process control issues:

3.3.1) The rules

A. 1 s sample outside of the upper +3 sigma-control limits for all product types (if many, list only the first 3 and last 3 and total number identified).

B. Find the most consecutive samples of s between the -1 and +1 sigma-control limits for all product types. This signifies good control.

C. 4 consecutive X-bar samples outside of the upper, second control limits for all product types (if many, list only the first 3 and last 3 and total number identified) just this

3.3.2) Issues

Product_Type <chr>	RuleA_s_above_3sigma <int>	RuleA_First_Last <chr>	RuleB_Longest_Run <int>	RuleC_Run_Count <int>	RuleC_Starts <chr>
CLO	0	None	28	228	165 ... 642
KEY	0	None	17	234	97 ... 737
LAP	0	None	23	110	115 ... 418
MON	0	None	36	165	132 ... 606
MOU	1	107 ... 107	19	265	209 ... 851
SOF	0	None	22	261	129 ... 854

The Rule A analysis found only one instance (Product MOU, sample 107) where the sample standard deviation exceeded the +3σ control limit, indicating a single abnormal increase in process spread. All other products remained within the control limits, showing stable variation.

Rule B results show that Product MON achieved the longest continuous run (36 samples) within the ±1σ limits, demonstrating strong process consistency. Most other products exhibited stable runs between 17 and 28 samples, reflecting good but slightly less consistent control.

Rule C identified multiple occurrences of four consecutive subgroup means above the +2σ limit, suggesting minor mean shifts. Products MOU and SOF had the highest counts of such runs (265 and 261 respectively), while CLO and MON showed fewer, more evenly spaced occurrences.

Overall, the control-chart diagnostics indicate that process variation is well controlled across all product types, with only isolated special-cause events. Product SOF remains the most stable and capable process, while MOU may require investigation into the cause of its single +3σ spike.

Part 4:

This section handles the probabilities of type 1 and type 2 errors in the previous section (part 3).

Then

4.1 Likelihood of type 1 error (false alarm):

For rule A:

The control limits are set at 3 standard deviations from the centreline.
So, the chance that any one sample's statistic randomly falls beyond +3 standard deviations—even when nothing is wrong—is given by the area under the normal curve to the right of 3:

$a=P(Z>3)$

=0,00135

For rule B:

Rule B asks you to find the longest run of samples whose s-values lie between +1 and -1 standard deviation. That run length shows how stable and consistent the process variation is. Thus, a Type 1 error is not possible.

For rule C:

Four consecutive $\bar{X}$ (sample means) above the +2 standard deviation line on the X-chart.

The probability that one group falls above 2 standard deviations

$P(Z>2) = 0.0228$

The probability that four consecutive groups are above the +2 standard deviation line

$P=0.0228^4$

P=**0.00000027**

4.2 Likelihood of type 2 errors

With the mean shifted to 25.028 Land the sampling Standard deviation increased to 0.017, about 84% of subgroup means would still land between the old control limits, so the chart would usually miss the shift.

The probability of a type two error is:

B=P(25,011<X<25.089)

=0.159

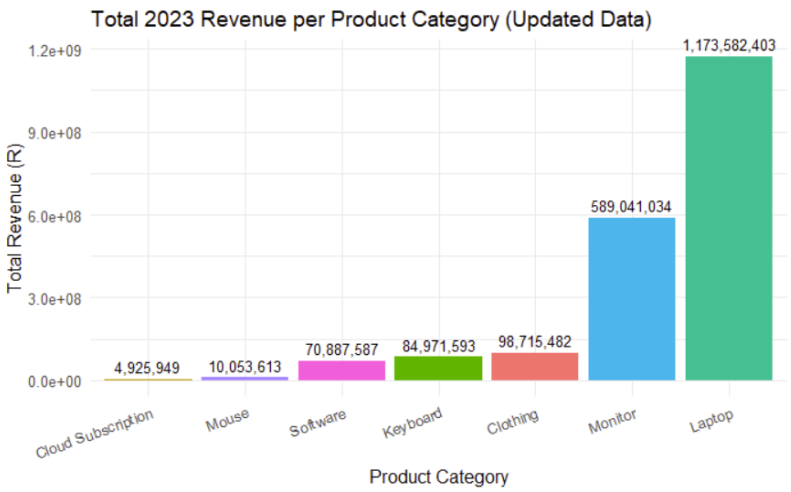
The likelihood of a type 1 error is very low in comparison with a type 2 error

Type I error is low because ±3 standard deviations limits make false alarms very rare. Type II error is high because a small mean shift and larger variation make it hard for the chart to detect the real change.

Part 4.3: Re-analysis of Part 1.2

category <chr>	total_revenue <dbl>
Laptop	1173582403
Monitor	589041034
Clothing	98715482
Keyboard	84971593
Software	70887587
Mouse	10053613
Cloud Subscription	4925949

7 rows



Key observations:

- Laptop and Monitor categories dominate sales, together accounting for nearly 75% of total 2023 revenue.
    - Laptop revenue increased substantially to R1.17 billion, suggesting that earlier data had understated the selling price for this category.
    - Monitors follow at R589 million, confirming consistent demand across both years.
- Mid-tier categories such as Clothing, Keyboard, and Software each contribute between R70–R100 million, indicating moderate but steady performance.
- Cloud Subscription and Mouse products remain the smallest contributors, though these figures are now accurate after the price corrections.
- The earlier (before-correction) chart showed compressed revenue ranges across categories — meaning pricing inconsistencies had flattened the differences. The corrected version restores realistic proportionality, clearly showing that Laptops and Monitors generate far higher sales.

Part 5: Optimisation of coffee shops:

How it was done:

It was estimated that each customer on average would spend 30 units while a barista would be paid 1000 units. Then an estimated customers per day was calculated if operating hours was equal to 10 hours/day:

average customers/day= (operating hours x 3600)/ (average service time)

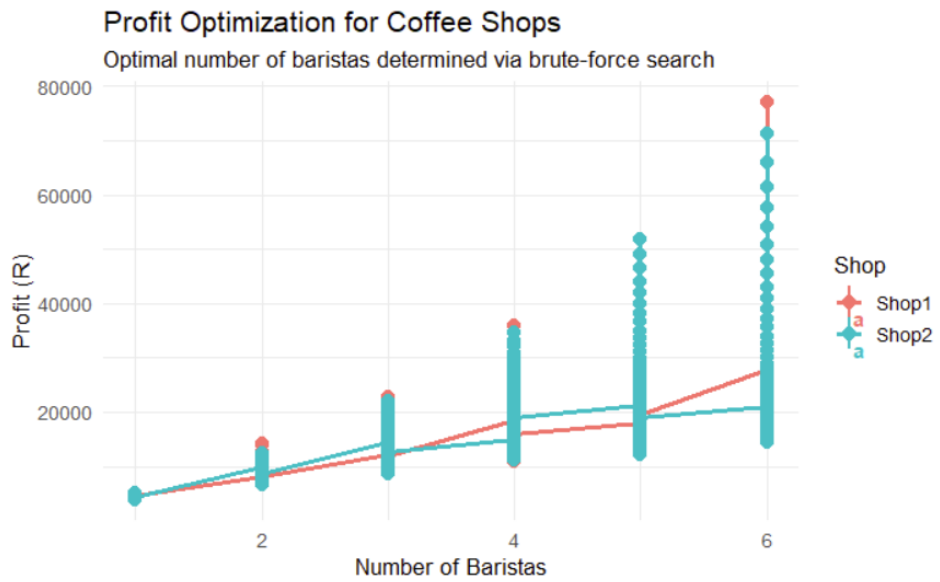
For every number of baristas, the average revenue was calculated for both shops.

Average Revenue/day= (average customers/day x 30) - (1000 x # baristas)

Shop <chr>	Baristas <int>	MeanServiceTime_s <dbl>	CustomersPerDay <dbl>	Revenue_R <dbl>	StaffCost_R <dbl>	Profit_R <dbl>
Shop1	2	100.17	359.4	10781.6	2000	8781.6
Shop1	3	66.61	540.4	16213.4	3000	13213.4
Shop1	4	49.98	720.3	21608.5	4000	17608.5
Shop1	5	39.96	900.9	27025.8	5000	22025.8
Shop1	6	33.36	1079.3	32378.3	6000	26378.3
Shop2	2	99.94	360.2	10806.6	2000	8806.6
Shop2	3	66.70	539.7	16191.9	3000	13191.9
Shop2	4	50.05	719.2	21577.4	4000	17577.4
Shop2	5	40.01	899.8	26993.7	5000	21993.7
Shop2	6	33.32	1080.3	32409.7	6000	26409.7

It was found that for both shops the optimal number of baristas was 6.

Shop <chr>	Baristas <int>	Profit_R <dbl>
Shop1	6	26378.3
Shop2	6	26409.7



Part 6: Manova tests

Because it is a multivariate test Manova is chosen.

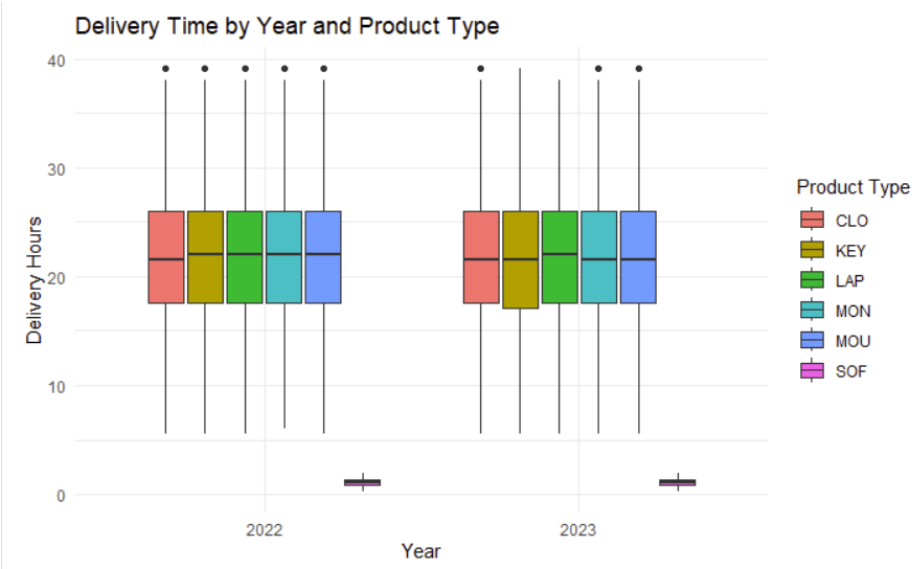
6.1 Test:

H_0: The mean of delivery-related times does not differ by year nor product type

H_1: At least one of the mean vectors differs across years or product types.

MANOVA 1 Summary (Wilks' Lambda Test)

	Df	Wilks_Lambda	Approx_F	Num_Df	Den_Df	P_value	Significa
as.factor(orderYear)	1	0.99991	4.49546	2	99987	0.01116	*
as.factor(substr(ProductID, 1, 3))	5	0.02669	102414.85393	10	199974	0.00000	***
as.factor(orderYear):as.factor(substr(ProductID, 1, 3))	5	0.99989	1.13091	10	199974	0.33396	ns
Residuals	99988	NA	NA	NA	NA	NA	ns



The p value for delivery time between year 2026 and 2027:

Λ=0.99991

P=0.011

➔ P<0.05

P value for delivery time between product types:

Λ=0.02

P=0.000000006

➔ P<0.05

Conclusion:

The analysis showed that both the year of the orders and the product type had a clear influence on delivery performance. Delivery and turnaround times changed between the two years, suggesting that process efficiency or external conditions may have affected performance. Product type also had a strong effect, meaning that different products consistently required different handling or delivery times. However, the way performance changed from one year to the next was similar for all products, showing that improvements or delays affected every category in roughly the same way. Overall, the results indicate that delivery performance evolved over time and varied across products, but the general trend of change was consistent throughout the system.

Part 7: Binomial

7.1 Estimated reliable days in a year

Service is reliable when there are 15 or more workers:

P= (270+96)/397

P=0.9219

Estimated reliable days in a year= 0.9219 x 365

=336.5

=337 days

7.2 Optimise profit

Modelled as a Binomial problem:

Take:

N= 16

P=0.9219

$$P(X \geq 15) = \sum_{n=16}^{16} \binom{n}{k} x^k a^{n-k}$$

$$0.922 = \sum_{n=16}^{16} \binom{16}{15} (p)^{16}(1-p)^{1}$$

p=0.966 each individual worker has a 96.6% chance of pitching

Loss/problem day=20000

Hiring new staff member=25000/month

In a 30-day month:

No extra staff -> No money spent on hiring new staff

Number of non-reliable days= (1-0.922) x30

=2.34

Average money lost=2.34x20000

=46800

Thus, more staff is needed.

Table comparing Number of workers and annual expected loss

Number of Workers	**Reliability (% of reliable days)**	**Problem Probability (1 – Reliability)**	**Expected Problem Days (out of 365)**	**Expected Annual Loss (R20 000/day)**	**Extra Staffing Cost (R25 000 per month per worker)**	**Estimated Annual Profit (Rands)**
16	92.2 %	7.8 %	28	R 560 000	R 0	**Base profit**
17	99.5 %	0.5 %	2	R 40 000	R 300 000	**↑ Net gain ≈ R 220 000**
18	99.9 %	0.1 %	< 1	R 20 000	R 600 000	**↓ Net loss ≈ R –380 000**

Adding one extra worker increases reliability to nearly 100 percent while only slightly increasing staff costs, making 17 workers the most cost-effective option. Hiring 18 workers provides almost no additional improvement in reliability and reduces overall profit due to higher wage expenses. Therefore, maintaining 17 workers per day offers the best balance between reliability and profitability.

Conclusion:

The project successfully demonstrated how data analytics and statistical process control can be used to monitor, evaluate, and optimise business performance. Through descriptive statistics, key sales and operational trends were identified, while SPC techniques such as $\bar{X}$ and S charts provided insight into process stability and variation. The optimisation analysis further illustrated how data-driven decision-making can improve efficiency and profitability by determining ideal operating conditions. Overall, the project highlights the value of integrating statistical tools and R programming into quality assurance and management practices to support continuous improvement and evidence-based decision-making.

References:

ChatGPT (2025) *Data analysis explanations, R code generation, and project writing assistance for ECSA Quality Assurance Project.* OpenAI, 22 October 2025. Available at: https://chat.openai.com

Montgomery, D.C. (2020) *Introduction to Statistical Quality Control.* 8th ed. Hoboken, NJ: John Wiley & Sons.

Evans, J.R. and Lindsay, W.M. (2020) *Managing for Quality and Performance Excellence.* 12th ed. Boston, MA: Cengage Learning.

Field, A., Miles, J. and Field, Z. (2012) *Discovering Statistics Using R.* London: SAGE Publications.

Ross, S.M. (2017) *Introduction to Probability and Statistics for Engineers and Scientists.* 6th ed. London: Academic Press.

Slack, N., Brandon-Jones, A. and Burgess, N. (2022) *Operations Management.* 10th ed. Harlow: Pearson Education.

Kume, H. (1993) *Statistical Methods for Quality Improvement.* 2nd ed. New York: CRC Press.