Quality Assurance

ECSA Report

Lien Hoogenboezem [27128466]
DEPARTMENT OF INDUSTRIAL ENGINEERING,
STELLENBOSCH UNIVERSITY

Table of Contents

List of Figures:

List of Tables:

Introduction

This report serves as an Engineering Council of South Africa (ACSA) Graduate Attribute 4 (GA4) submission for the Stellenbosch University Industrial Engineering program. It demonstrates competence in applying quantitative engineering science principles to solve complex problems. The primary objective of this report is to evaluate operational datasets that cover customer sales, product logistics and service times. The datasets are evaluated using statistical optimization techniques to improve data quality, control processes and maximize profit.

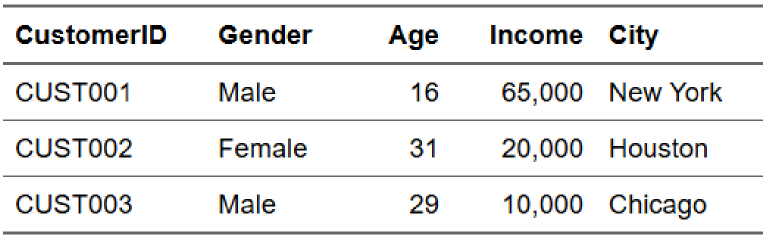
The analysis is structured into seven main parts. This report begins with a Basic Data Analysis using descriptive statistics to understand the data structure and quality issues. This leads to Statistical Process Control (SPC) to establish control limits and assess process capabilities for delivery times. Risk and Data Correction addresses data discrepancies, calculates Type I and Type II errors, and re-evaluates the descriptive statistics with corrected data. The report then applies Optimization techniques to determine optimal barista staffing (part 5) and car rental personnel levels (part 7). It focuses on balancing operational costs, reliability and revenue. Finally, Design of Experiments (DOE) and Analysis of Variance (ANOVA) is used to test for significant differences in the performance of processes. The findings provide actionable insights for process stability, data governance and strategic resource allocation.

1. Basic Data Analysis

Data loading and inspection

Customer Data

Table 1: Customer Data Preview

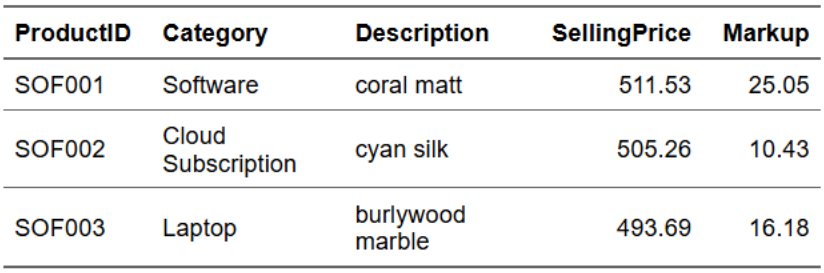


CustomerID	Gender	Age	Income	City
CUST001	Male	16	65,000	New York
CUST002	Female	31	20,000	Houston
CUST003	Male	29	10,000	Chicago

The customer data contains 5 columns. The first column, *CustomerID*, contains the unique ID of each customer. The second column, *Gender*, contains a "Male", "Female" or "Other" characteristic of the customer in question. The *Age* column is a numeric variable that contains the customer's age. The *Income* column contains the numeric values of the customer's income, and the *City* column contains the city that the customer resides in.

Products Data

Table 2: Products Data Preview

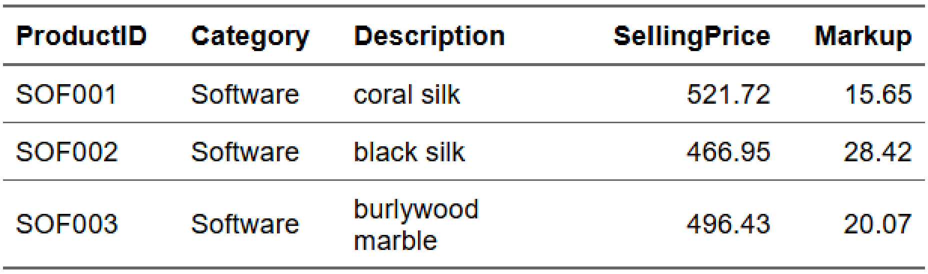


ProductID	Category	Description	SellingPrice	Markup
SOF001	Software	coral matt	511.53	25.05
SOF002	Cloud Subscription	cyan silk	505.26	10.43
SOF003	Laptop	burlywood marble	493.69	16.18

The Products Data dataset contains 5 columns. *ProductID* contains the unique ID of each product in a character format. The *Category* column contains the category that the product is defined under by the company. The *Description* column is a textual field that contains a description of the product in question. The *SellingPrice* column contains the amount that the customer pays for the product, and the *Markup* column contains the amount that the company is profiting off the specific product per purchase.

Products Head Office

Table 3: Products Head Office Preview

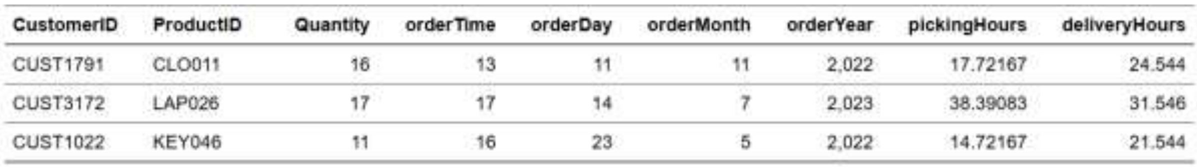


ProductID	Category	Description	SellingPrice	Markup
SOF001	Software	coral silk	521.72	15.65
SOF002	Software	black silk	466.95	28.42
SOF003	Software	burlywood marble	496.43	20.07

The Products HeadOffice dataset contains the exact same details as the *ProductsData* dataset.

Sales 2022 and 2023 Data

Table 4: Sales 2022 and 2023 Preview



CustomerID	ProductID	Quantity	orderTime	orderDay	orderMonth	orderYear	pickingHours	deliveryHours
CUST1791	CLO011	16	13	11	11	2,022	17.72167	24.544
CUST3172	LAP026	17	17	14	7	2,023	38.39083	31.546
CUST1022	KEY046	11	16	23	5	2,022	14.72167	21.544

The Sales2022and2023 dataset contains 9 columns, which outline the following: the *CustomerID* column and the *ProductID* column will indicate which customer bought which product from the catalogue. The *Quantity* outlines how many of the product the customer ordered. The *orderTime*, *orderDay*, *OrderMonth* and *OrderYear* columns outline the specific time that the order was placed with the company. The *PickingHours* column has details about how long it took the company to pick and pack the order, and the *deliveryHours* column outlines how long it took for the order to be delivered to the customer.

Missing values

Using the *is.na()* function, it is confirmed that there are no missing data in any of the datasets.

Mismatching Values

On inspection, although the *ProductID* values align across both datasets, there are major inconsistencies in the associated attributes. The *Category* and *Description* fields are not standardized between *ProductsData* and *ProductsHeadOffice*. Most importantly, the *SellingPrice* and *Markup* values differ for the same *ProductID* across the two datasets. This raises concerns about data governance and consistency between branch-level and head-office systems. Such discrepancies need to be reconciled before reliable analysis or decision-making can be performed.

The following plot depicts the difference in the *SellingPrice* column in the products in the *ProductsHeadOffice* and *ProductsData* datasets that have the same *ProductID*.

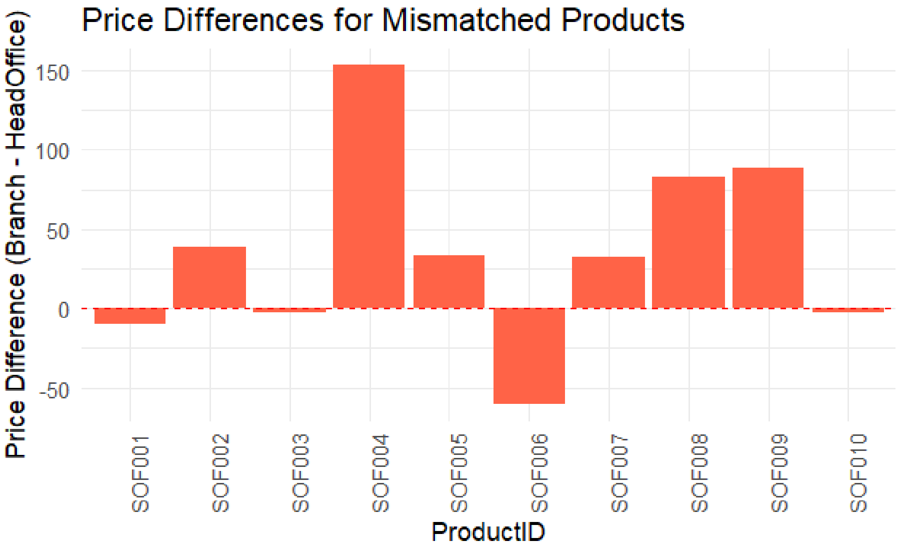


Figure 1: Visual figure of Data quality issues

Data Visualizations

Customer Data

The age distribution of the customer can be viewed with the following plot:

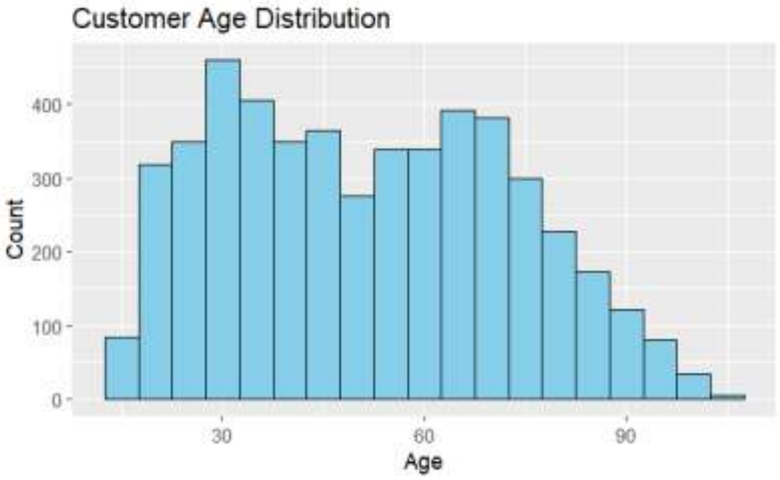


Figure 2:

The distribution of age is bimodal, with peaks around 30 years old and 65 years old. The majority of this company's customers are middle aged. This company should use this age-based data to target marketing at these age groups on the applicable platforms.

The following plot depicts the distribution across genders for the company's customers:

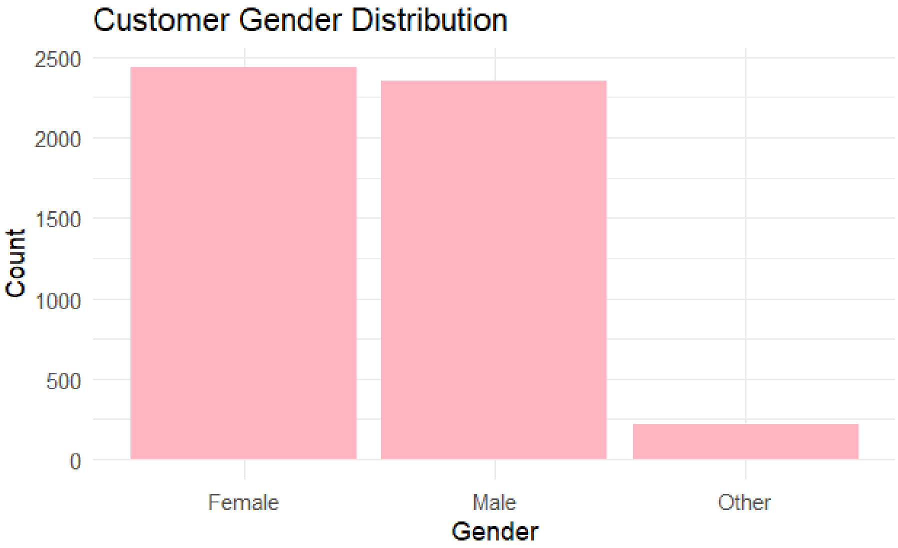


Figure 3

The distribution between male and female customers is approximately even. The “other” class has very little data. Therefore, there is no opportunity for targeted marketing toward a single gender.

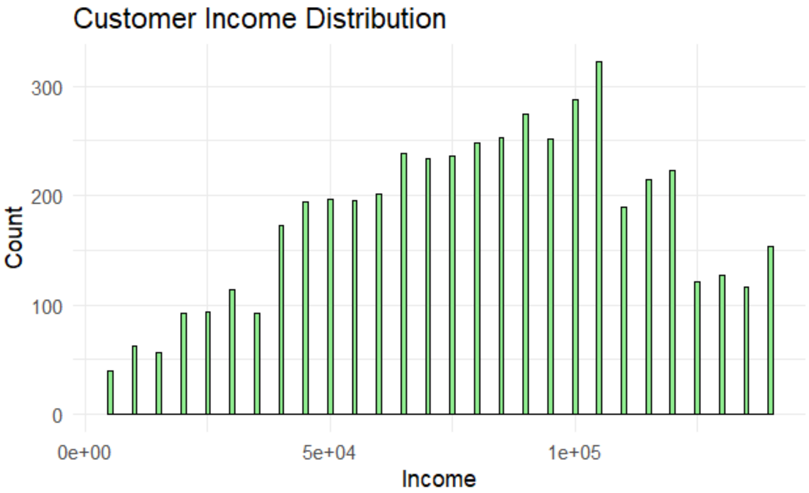


Figure 4

The distribution of the income for different customers infers that the customers buying products from this company are middle to high income customers.

The following plot depicts the correlation of the customerdata.

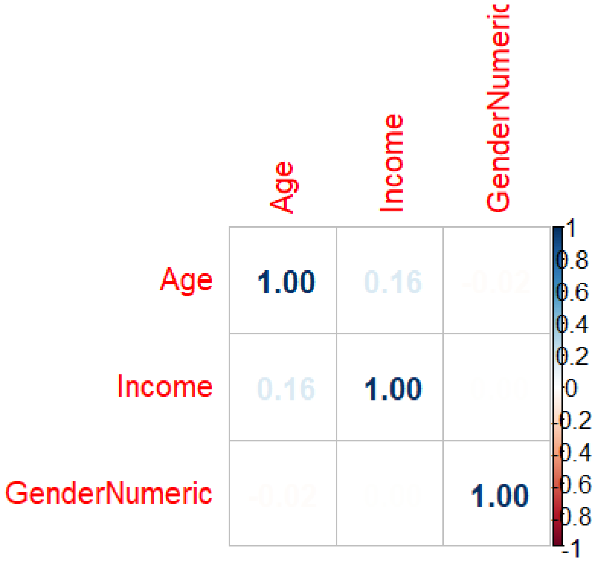


Figure 5: Customer Data Correlation plot

There is a very slight correlation between the customer's age and their income, but it is not significant enough to extrapolate information from. Other than the Income and age correlation, there are no other correlations.

Product Data

By viewing the following plot, it can be seen that the company has an equal number of different products, with no variation at all.

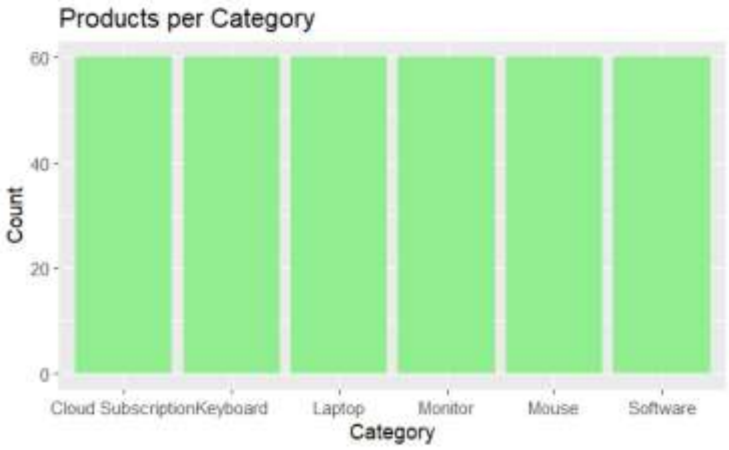


Figure 6: Distribution of products by category

The following plot depicts how long it takes to pack and then deliver the different products by their categories.

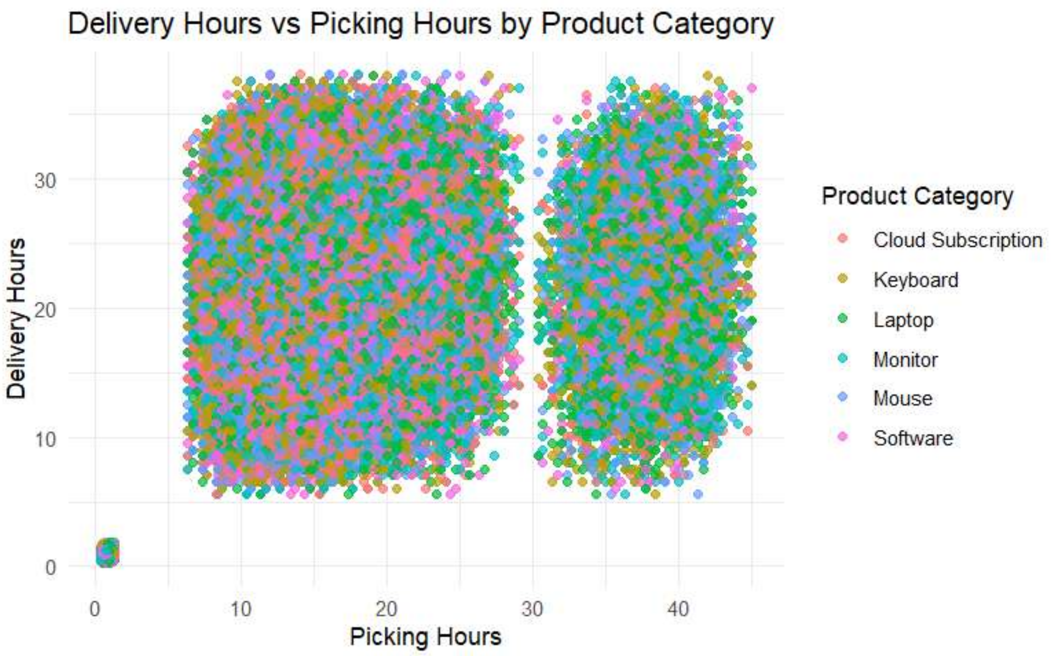
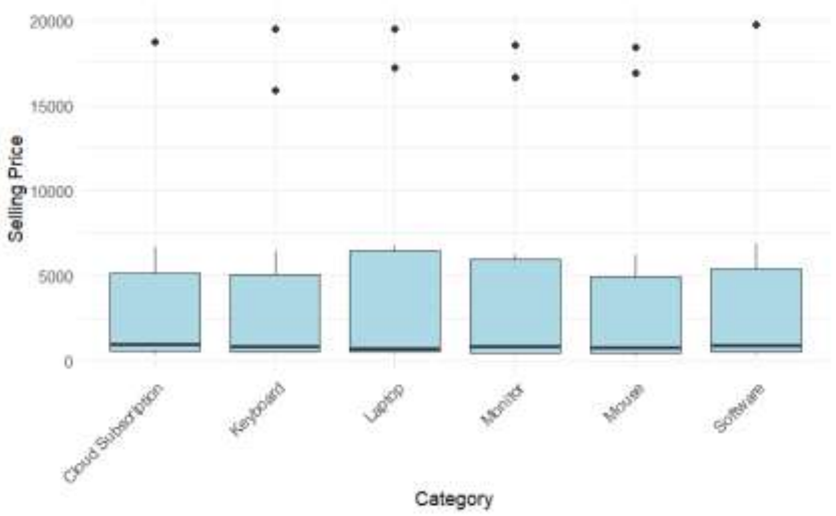


Figure 7:

This plot shows a slight positive correlation, meaning that it takes longer to deliver a product if the picking time is long. The type of product being packed does not seem to have any effect on the picking and delivery times. The monitors and laptops take longer to pack and deliver.

The following boxplot depicts the distribution of the prices per product category.

Figure 8: Selling Price by Category



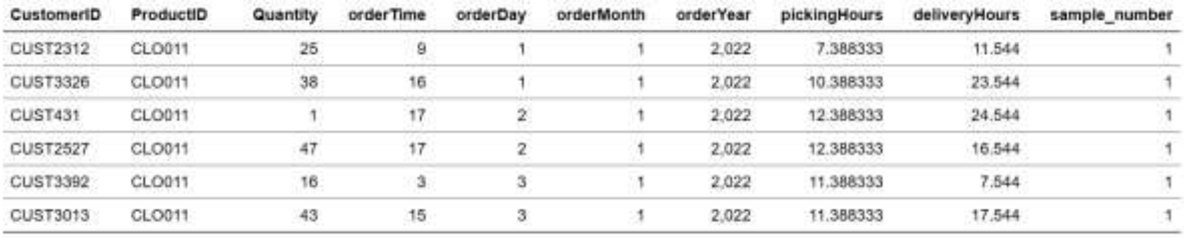
The majority of products within each class are all lower on the price range. There are significant outliers for each category that could indicate incorrect data.

3. Statistical Process Control

To prepare the *sales2026and2026* data for Statistical Process Control (SPC), the forecasted sales data was ordered and sorted chronologically by ProductID, followed by the order year, month, day and time. The data being ordered in this way simulates real-time data collection.

Next, the data for each product was divided into samples of 24 consecutive delivery times to adhere to the standard SPC subgroup sizing. Only the first 30 samples of the 720 delivery records per product were used to establish the initial X-bar and s control limits. The table below shows a sample of the subgroups:

Table 5: Table showcasing subgroups of sales2026and2027



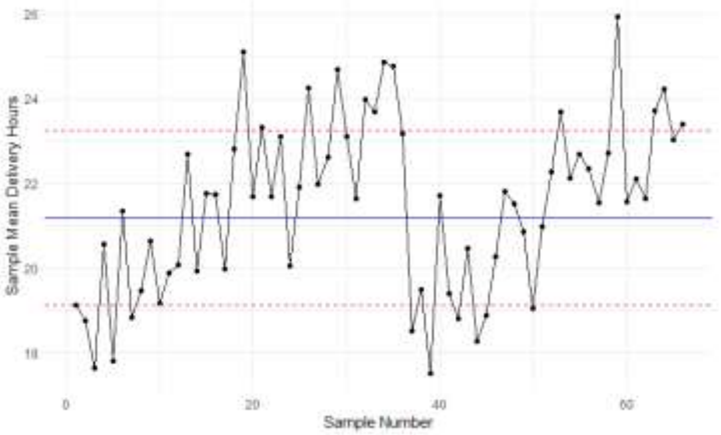
CustomerID	ProductID	Quantity	orderTime	orderDay	orderMonth	orderYear	pickingHours	deliveryHours	sample_number
CUST2312	CLO011	25	9	1	1	2,022	7.388333	11.544	1
CUST3326	CLO011	38	16	1	1	2,022	10.388333	23.544	1
CUST431	CLO011	1	17	2	1	2,022	12.388333	24.544	1
CUST2527	CLO011	47	17	2	1	2,022	12.388333	16.544	1
CUST3392	CLO011	16	3	3	1	2,022	11.388333	7.544	1
CUST3013	CLO011	43	15	3	1	2,022	11.388333	17.544	1

This structured sampling is critical for accurately modelling process behaviour and setting the control thresholds in further analyses.

Then, the centre lines for the X-bar chart and the s control chart are calculated using the first 30 samples. Using the standard SPC formulas and constants for n=24, the 3-sigma, 2-sigma and 1 sigma control limits are calculated for both charts. The product CLO011 is isolated for the following graphs:

The X-bar control chart is presented as follows:

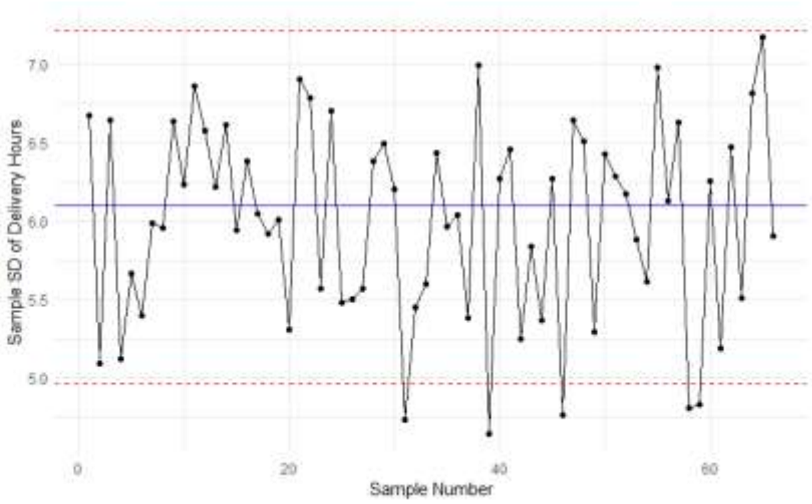
Figure 9: X-Bar control chart for CLO011



This chart shows that most sample means are within the control limits, but there are many points that are close to or outside the upper control limit. This implies that there is a shift or special cause variation affecting the process mean.

The s chart for product CLO011:

Figure 10: S control chart for CLO011



This chart shows that most of the standard deviations are within the control limits, indicating thay the process variability for this product is relatively stable.

After the control limits for the first 30 samples are established, each product is monitored by drawing samples of 24. The X-bar and S charts for all products (figures 11 and 12) show the ongoing process control, with each new sample checked against these estalished control limits.

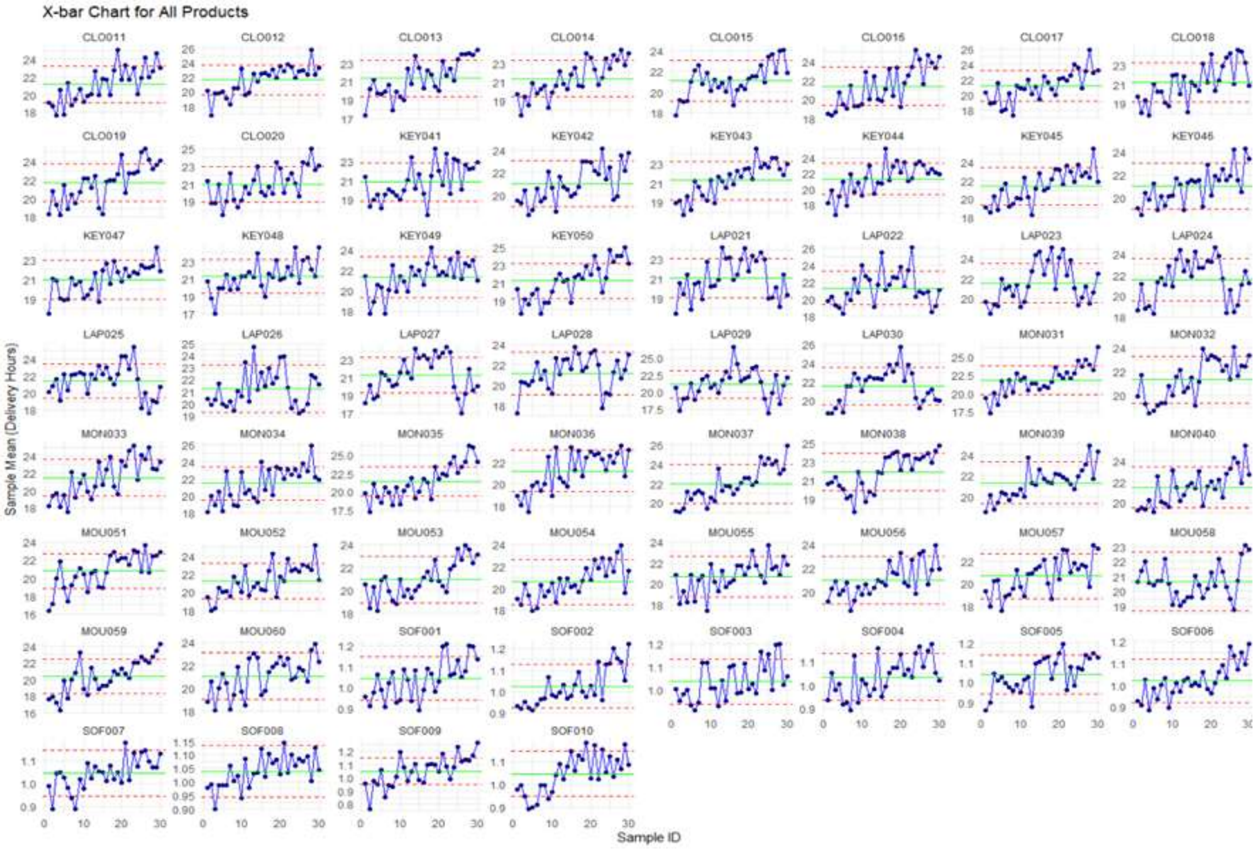


Figure 11: X-Bar plots for all products

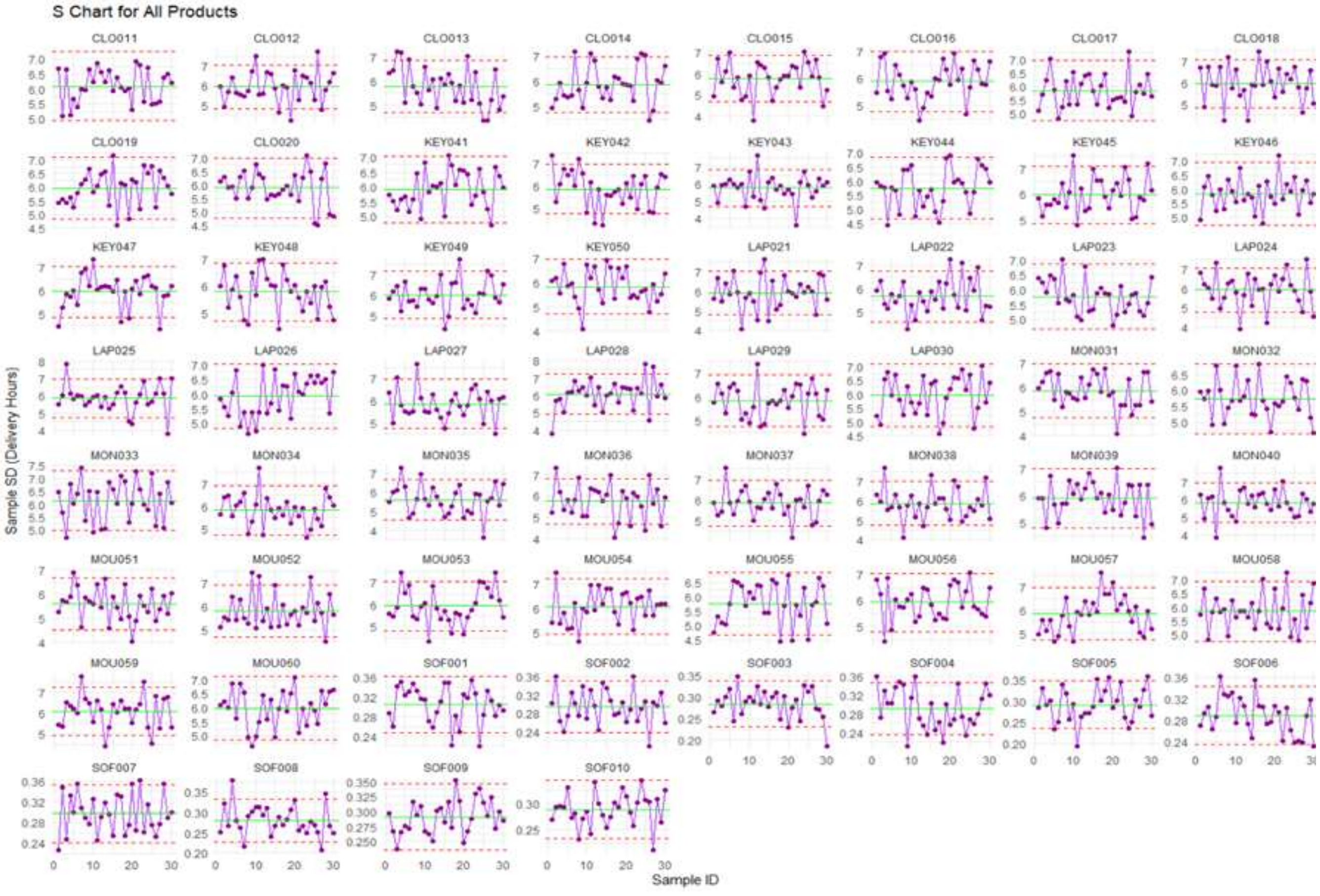


Figure 12: S chart for all products

Real-time Process monitoring with Control Charts

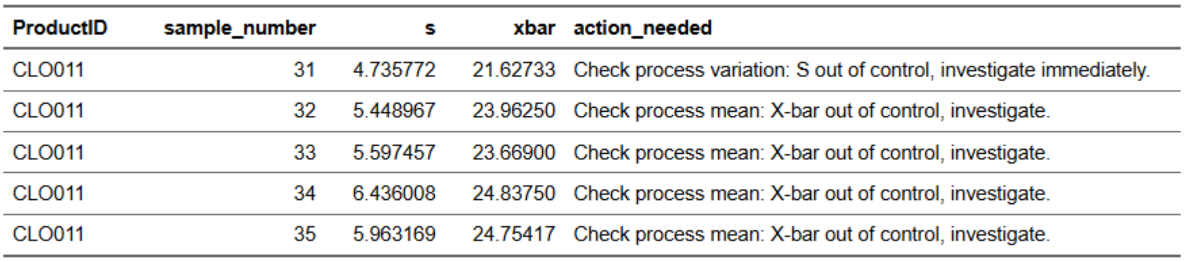
In real world applications, the process data (such as delivery times) will be observed and recorded as the events occur. Therefore, the control charts must be updated and manually checked in real time using the new information as it becomes available.

When manually checking, the relevant process manager should first check the S chart to see if the sample standard deviation is outside of the control limits. If it is, the process is unstable and checking the X-bar chart for that sample would also show the instability. However, if the S chart is in control, the X-bar chart should be inspected to see if the sample mean is within the limits. If it is not within the limits, it signals a potential process shift or special cause variation.

Whenever either of the real-time charts are flagged as unstable and out of control, the relevant manager should be notified to identify the cause and rectify it.

The following table shows a simulated real-time summary of samples where the process was flagged as out of control. This will give managers a warning and guidance when samples are flagged.

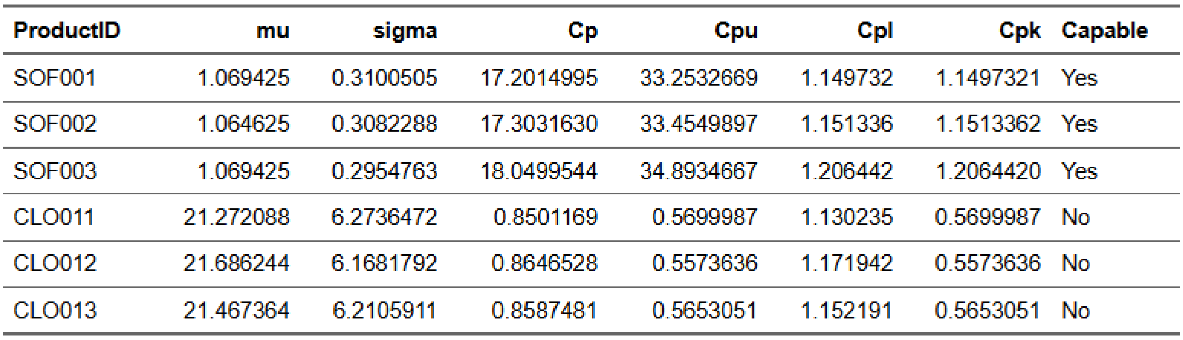
Table 6: Real-time flagging



ProductID	sample_number	s	xbar	action_needed
CLO011	31	4.735772	21.62733	Check process variation: S out of control, investigate immediately.
CLO011	32	5.448967	23.96250	Check process mean: X-bar out of control, investigate.
CLO011	33	5.597457	23.66900	Check process mean: X-bar out of control, investigate.
CLO011	34	6.436008	24.83750	Check process mean: X-bar out of control, investigate.
CLO011	35	5.963169	24.75417	Check process mean: X-bar out of control, investigate.

Process capability indices were calculated for each product type (using the first 1000 deliveries). The table below summarizes the Cp, Cpu, Cpl and Cpk values. Product types where the Cpk is greater than one is considered capable of meeting the VOC requirements for delivery times.

Table 7: Process Capability indices for delivery times by product type



ProductID	mu	sigma	Cp	Cpu	Cpl	Cpk	Capable
SOF001	1.069425	0.3100505	17.2014995	33.2532669	1.149732	1.1497321	Yes
SOF002	1.064625	0.3082288	17.3031630	33.4549897	1.151336	1.1513362	Yes
SOF003	1.069425	0.2954763	18.0499544	34.8934667	1.206442	1.2064420	Yes
CLO011	21.272088	6.2736472	0.8501169	0.5699987	1.130235	0.5699987	No
CLO012	21.686244	6.1681792	0.8646528	0.5573636	1.171942	0.5573636	No
CLO013	21.467364	6.2105911	0.8587481	0.5653051	1.152191	0.5653051	No

Samples that show process control

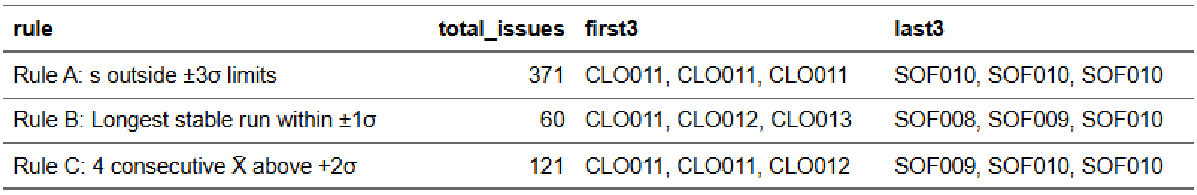
To evaluate the process's stability, three standard SPC rules were applied across all the product types and is shown in table 6.

Rule A: Identified samples with standard deviations that exceed the bounds of 3 sigma control limits.

Rule B: Shows the longest consecutive runs of samples with S values within the 1 sigma limits, indicating that they have consistent process performance. The top 3 products with the longest consecutive runs of samples within the limits are: SOF008 with a consecutive run of 8 samples, KEY049 with a consecutive run of 6 samples and MOU058 with a consecutive run of 6 samples as well.

Rule C: Shows instances where four consecutive sample's means exceeded the 2-sigma limit.

Table 8: SPC rules



rule	total_issues	first3	last3
Rule A: s outside ±3σ limits	371	CLO011, CLO011, CLO011	SOF010, SOF010, SOF010
Rule B: Longest stable run within ±1σ	60	CLO011, CLO012, CLO013	SOF008, SOF009, SOF010
Rule C: 4 consecutive $\bar{X}$ above +2σ	121	CLO011, CLO011, CLO012	SOF009, SOF010, SOF010

Together, these analyses help differentiate between instances where there is random variation and instances where there are systematic issues.

4. Risk, Data correction and Optimising for maximum profit

4.1

A Type I error occurs when a control chart signals that the process is out of control, when it is not out of control in reality. Under the assumption that the process is normally distributed and in control (H_0: the process mean and variation are stable), the theoretical probabilities of the abovementioned "false alarms" can be estimated using statistical rules. Since a normal distribution is symmetric about the mean, the probability that a single point lies above or below the centreline is equal. Therefore, the probability of finding 7 consecutive samples above the centreline is $0.5^7 = 0.0078$. This is the probability of a false alarm for rule B. For rule A (A point beyond the upper 3-sigma limit), the probability of false alarms is 0.0027 or 0.27 %. For rule C (four consecutive points beyond the upper 2-sigma limit), the probability of a false alarm is incredibly small at 0.000027%. These probabilities are theoretical and represent the probability of a false signal under perfectly in-control situations.

4.2

A Type II error occurs when a process has shifter out of control (H_a is true) but the control chart doesn't signal that it did. For the bottle-filling process, the X-bar chart has a target mean of 25.05 litres and control limits of 25.001 (LCLO) and 25.089 (UCL). Unknown to us, the process mean has shifted to 25.028 L and the standard deviation increased from 0.013 to 0.017 litres.

In this context

A Type II error occurs when the process has actually shifted out of control (H_a is true), but the control chart fails to signal it. For the bottle-filling process, the x̄ chart has a target mean of 25.05 L and control limits of 25.011 L (LCL) and 25.089 L (UCL). Unknown to us, the process mean has shifted to 25.028 L and the standard deviation has increased from 0.013 L to 0.017 L.

Standardizing the control limits with respect to the new process mean and standard deviation:

$$Z_{\text{UCL}} = \frac{25.089 - 25.028}{0.017} \approx 3.588, \quad Z_{\text{LCL}} = \frac{25.011 - 25.028}{0.017} \approx -1$$
$$\beta = \Phi(3.588) - \Phi(-1) \approx 0.9998 - 0.1587 = 0.8411$$

Therefore, the likelihood of failing to detect the shift (Type II error) is approximately **84.1%**, indicating that the x̄ chart is not very sensitive to this small shift in the process mean. This calculation was made to verify the answer gained from R (0.8412).

4.3.

After the initial analysis was completed in week 1, several data inconsistencies were identified. Particularly, the mismatches between ProductsData and ProductsHeadoffice datasets in the SellingPrice and Markup fields were noted. These discrepancies were corrected by reconstructing the head office dataset so that each product's pricing and markup match the verified data. The corrected datasets are now referred to as productsdata2025 and productsheadoffice2025. The initial analysis will now be performed on the new datasets:

Data Quality Issues

After the same mismatch analysis is performed on the new dataset, the following visual for price mismatches was accrued:

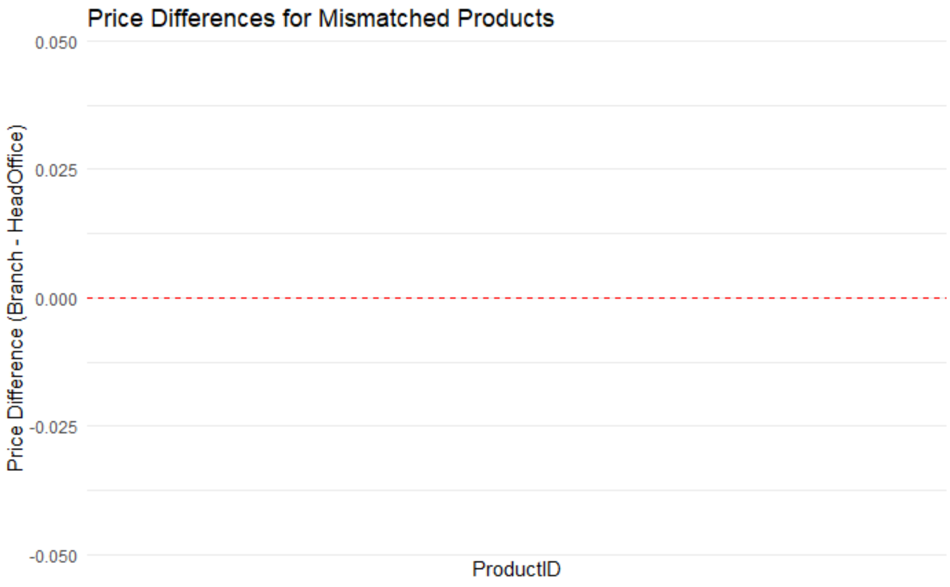


Figure 13: A visual of the updated dataset's mismatches

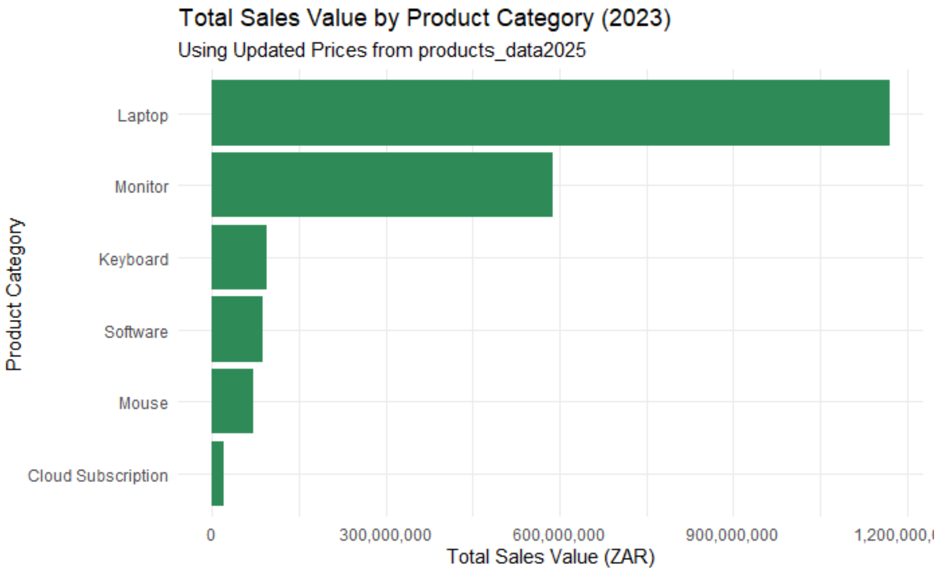
This graph displays how the inconsistencies in the datasets were corrected entirely.

Data Visualizations

The customer data will remain the same dataset. Therefore, there is no need for re-analysis.

Products Data

The following figure of the sales by product category shows significant change after the datasets were updated:



Delivery and Picking Time

The updated scatter plot of the time it takes to deliver and the time it takes to pick the product is shown in the following plot:

Figure 14: Updated figure of total sales per product category

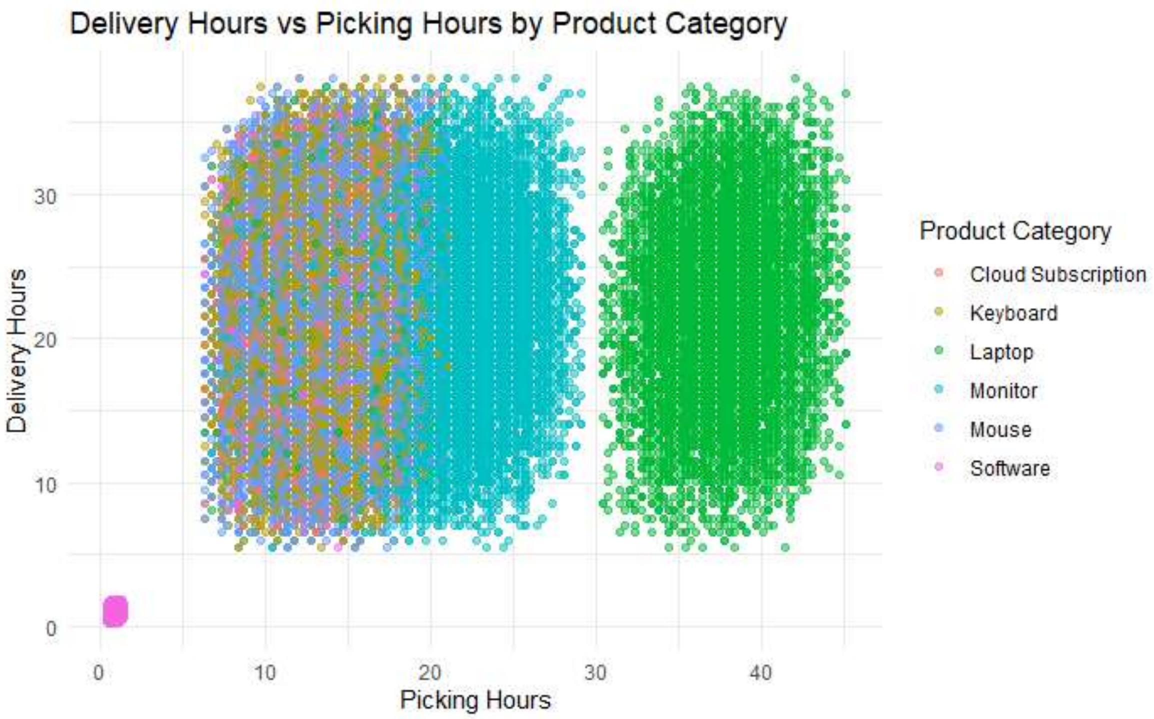


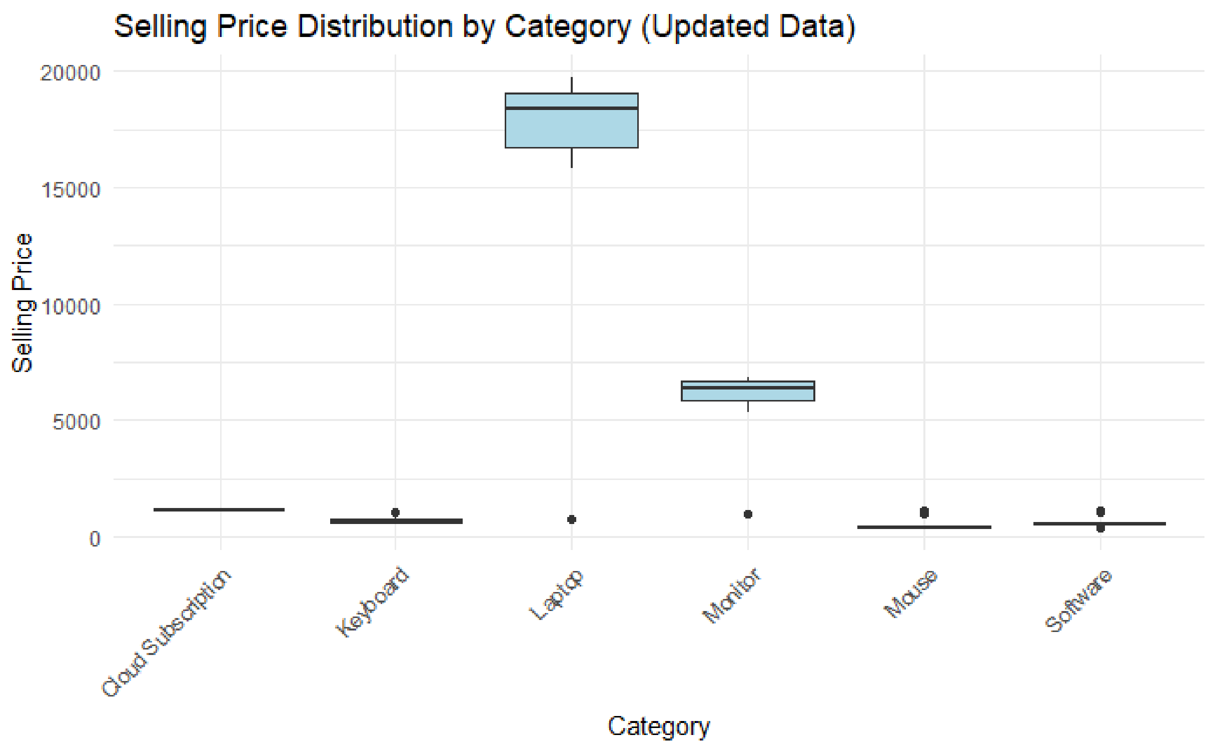
Figure 15: Updated delivery hours compared to the picking time by product type

The updated information clearly shows how there are clearer clusters formed. The clusters indicate that, in general, it takes longer to pick laptops out for customers, and that there is a large variation in the time it takes to deliver any product to the customer. The only exception to this delivery time trend is the time it takes to pick and deliver a cloud subscription. The picking and delivery time for this product is notably faster than other products, due to the intangible nature of the product. It does not need to be packaged and shipped. This figure also shows that monitors take moderately longer to be picked than the other products.

Selling Price

The following boxplot depicts the distribution of the prices per product category using the corrected head office data.

Figure 16: Selling Price by Category with the corrected data.



The corrected data shows a segmentation in pricing. Laptops are the highest -priced category (by median) with low variability. Monitors forms second tier in pricing, indicating that they are more moderately priced than laptops. The remaining four categories are all low-priced with very little price variation. These low-cost products include cloud subscriptions, keyboards, mouses and software. Notably, both laptops and monitors have outliers that are low in price value, indicating that the company sells them below their typical price range.

Using this figure in combination with a box plot of how each product is marked up will further reveal valuable insights. This box plot can be seen in figure 17

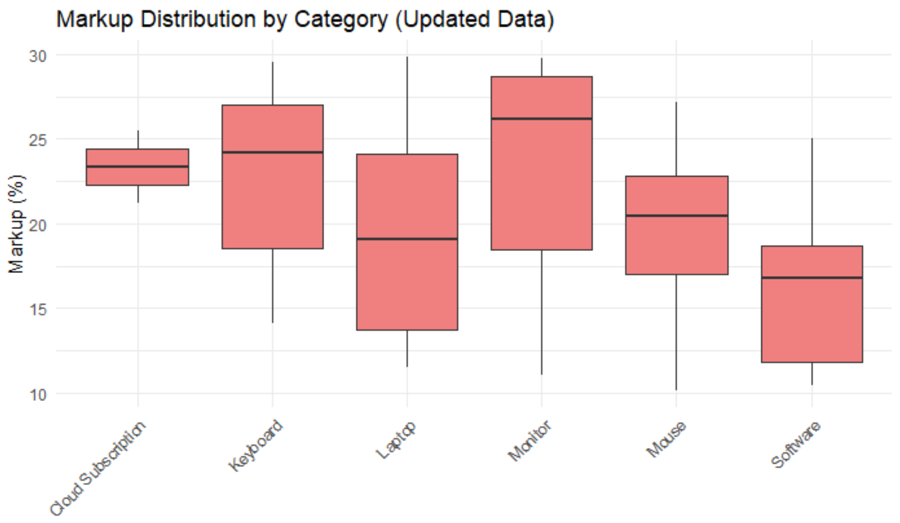


Figure 17: Box plot of price markup per category

The products with the highest overall markup percentage are monitors, keyboards and cloud subscriptions. The cloud subscription markup shows that it has the smallest variation, indicating that the company is very consistent pricing and markup value for these products. This could be due to the low variable cost connected to subscription services. The products with the lowest markups are software and laptops.

5.

Optimizing Barista Staffing for Maximum Profit

To determine the optimal amount of baristas for shop 1 and shop 2, the individual service times must be analysed from the timeToServe and timeToServe2 datasets. The goal is to maximise profit, while maintaining reliable service. The maximum amount of baristas allowed is 6 per shift.

The methodology followed to obtain these results are as follows:

1. Data Preparation:
   Both datasets contain two columns namely V1 and V2. V1 demotes the baristas (and was renamed to Baristas) on shift, and V2 denotes the time to serve each customer in seconds (renamed to ServiceTime). Additionally, all columns were converted to numeric.
2. Defining Reliable Service:
   Both the shop 1 and shop 2 datasets were inspected, and it was decided calculate the reliability threshold value based on the mean and 3 standard deviations. Therefore, shop 1 has a reliability threshold of approximately 90 seconds and shop 2 has a reliability threshold of approximately 150 seconds. Based on these thresholds, a metric called ReliableRate was calculated as the percentage of customers served within the threshold.
3. Profit Calculation:
   The values used for the profit calculation are as follows:

   Revenue per customer served = R30

   Personnel cost per barista per day = R1000

   Using these values and the following calculations, the daily metrics were derived for each shop based on their staffing level ranging from 1 barista to 6 baristas.

   Gross Revenue Per day = customers served per day × R30

   Net Profit Per day = Gross Revenue per day – Personnel Cost Per day

   Annual Profit = Net Profit Per Day × 365
4. Optimization:
   The optimal number of baristas were identified as the staffing level that maximized annual profit, while still maintaining a good service level.

The results that the model obtained were as follows:

Shop 1

The optimal number of baristas were 6 per day, corresponding to an annual profit of R746 850. The gross revenue per day would be R8 046, while still maintaining a service level of 100%. The following line plot displays gross revenue, personnel cost and net profit vary with the number of baristas.

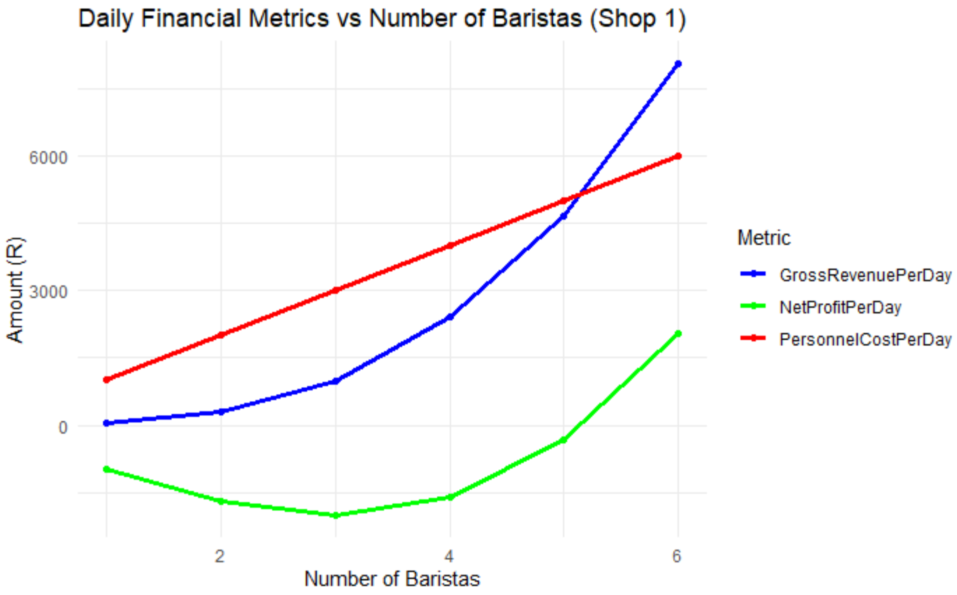


Figure 18: Financial metrics of shop 1

The following plot shows the time it takes to serve a customer based on the amount of baristas there are:

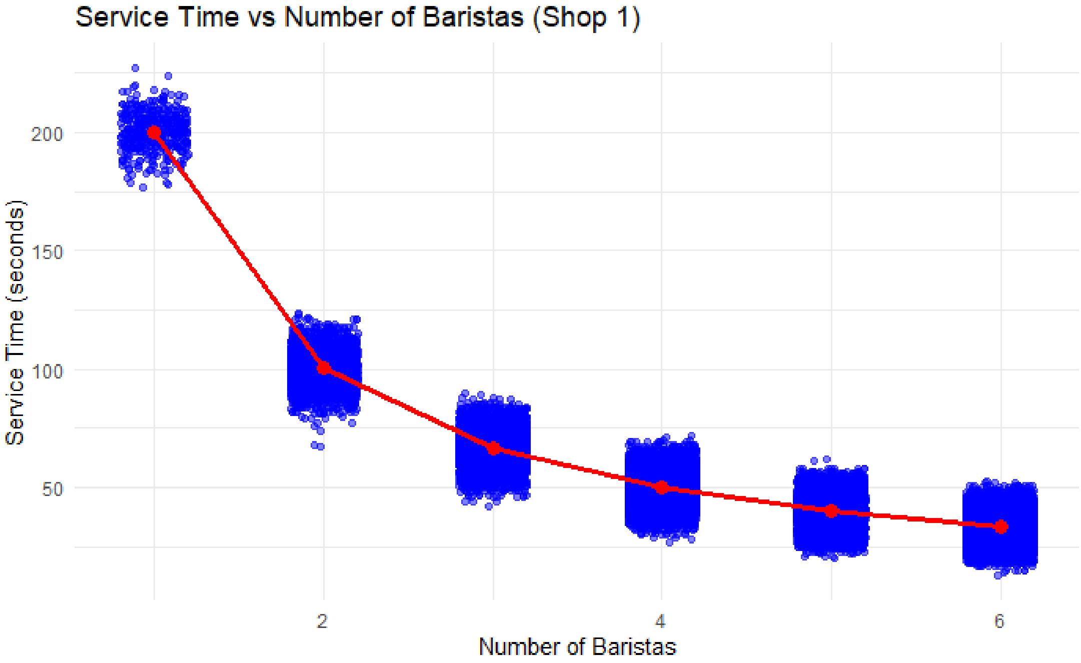


Figure 19: Service time for shop 1 based on the number of baristas employed

Shop 2

The optimal number of baristas found were 6, corresponding to an annual profit of R 177 900. The daily gross revenue for this amount of baristas would be R6 487, with a service level of 100% as well. The following line plot displays gross revenue, personnel cost and net profit vary with the number of baristas:

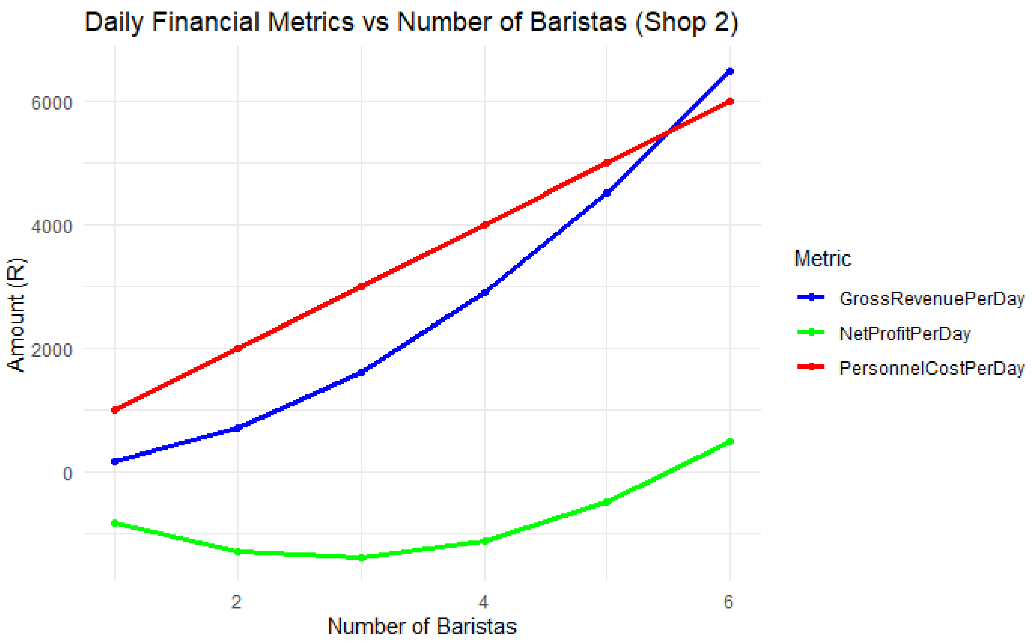


Figure 20: Financial metrics of shop 2

The following plot shows the time it takes to serve a customer based on the amount of baristas there are:

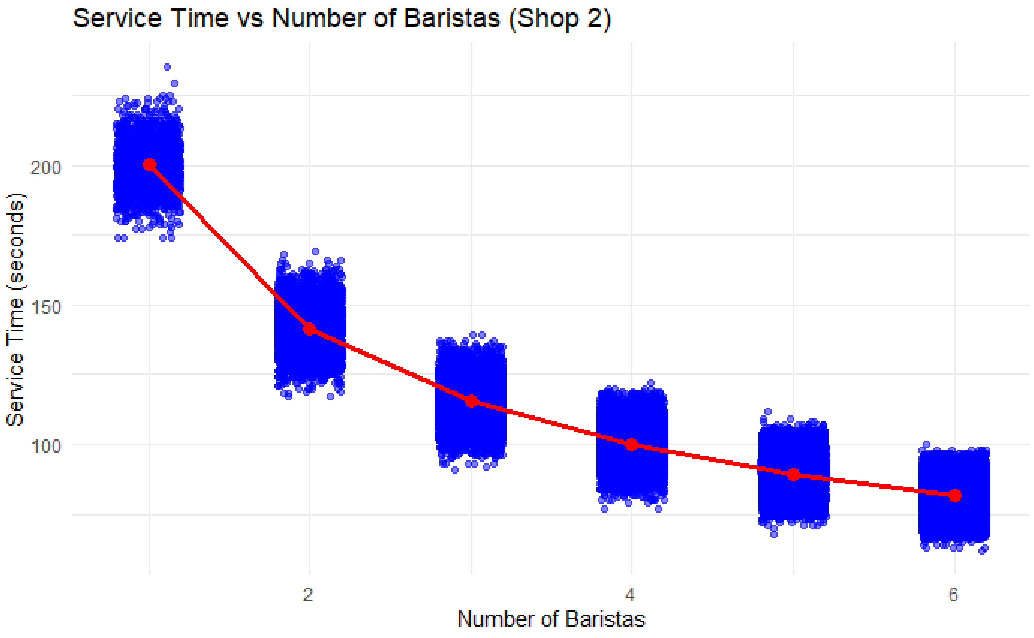


Figure 21:Service time for shop 1 based on the number of baristas employed

The approach for this optimization model has a standard financial approach: Maximize *Revenue – Cost*. The optimal decision (6 baristas) occurs when the marginal gain in revenue from more customers being served outweighs the marginal increase in personnel cost.

This model is conceptually different from the Taguchi Loss Function. The Taguchi approach is centered around quality loss and states that deviation from the target value (not just those outside the limits of the specifications) results in a loss to society. If the Taguchi model were imp

In this case, optimizing to maximize the financial profit provides the best operational decision. However, a Taguchi-based approach would enable us to continuously improve in speed beyond the 100% reliability threshold. It will quantify customer inconvenience and long-term brand damage as financial loss.

6.

6.1 Design of Experiments (DOE) and ANOVA

Design of Experiments (DOE) is a systematic method that us used to plan experiments so that we can draw valid and fully objective conclusions efficiently. DOE studies how changes in one or more input factors (independent variables) will affect the output response (dependent variable). It helps identify which factors have the biggest influence on the response, how all the different factors interact and which combination off parameters will produce the optimal results.

Analysis of Variance (ANOVA) is a statistical technique used in DOE to test if the means of several samples are different from one another. The null hypothesis (H_0) is that all sample means are equal, while the alternative hypothesis (H_1) is that at least one sample mean differs. If the p-value is smaller than the chosen significance level α (e.g., 0.05), H_0 will be rejected. Then, we can conclude that the slight difference in methods for each sample has an effect on the mean response.

When there are multiple dependent variables, the Multivariate Analysis of Variance (MANOVA) is used. MANOVA tests if the means of a group of samples differ across all samples simultaneously.

The output of the sample ANOVA table is:

Table 9: ANOVA sample results

Source	Degrees of Freedom	Sum of Squares	Mean Square	F Value	Pr(>F)
group	2	975.4148	487.70742	22.93855	0.00000004913035
Residuals	57	1,211.9041	21.26147		

The p-value (4.91×10^{-8}) is smaller than 0.05, indicating that there is a statistically significant difference between the sample means. Therefore, we reject the null hypothesis and conclude that at least one treatment produces a different mean response. The following box plot supports this visually, showing that treatment 3 has the highest mean, followed by treatment 2 and treatment 1.

This shows how ANOVA can be used to compare multiple treatments in a designed experiment to find out if changing the dependent variables lead to meaningful differences in the result.

6.2

Based on the SPC results in part 3, product CLO011 was selected for further analysis because its process variability remained relatively stable. This means that its performance is suitable to be compared to other years. The purpose of this analysis is to determine if there is a statistically significant difference in the mean delivery times between years 2022 and 2023.

H_0: There is no significant difference in the mean delivery times of product CLO011 between 2022 and 2023

H_1: There is a significant difference in the mean delivery times of product CLO011 between 2022 and 2023.

To test these hypotheses, we use a one-way ANOVA analysis in R that uses the delivery time as the response variable and the year (2022 and 2023) as the independent factor.

The resulting ANOVA table summary is presented as follows:

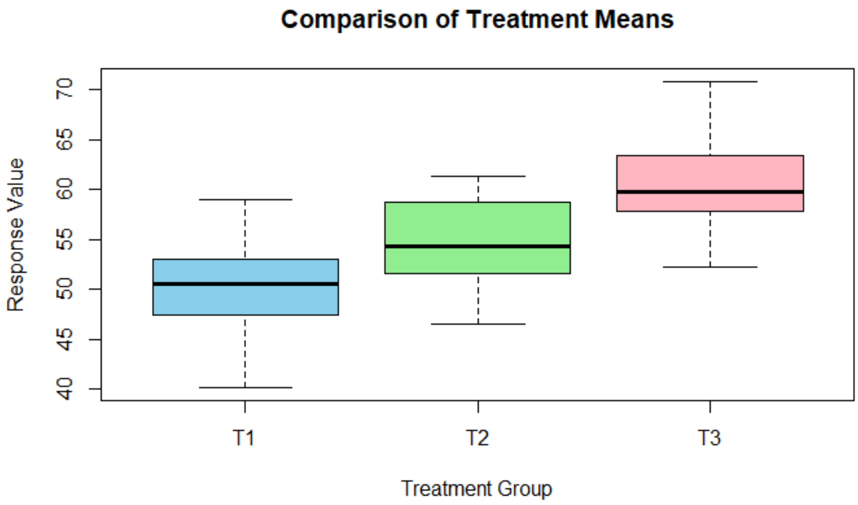
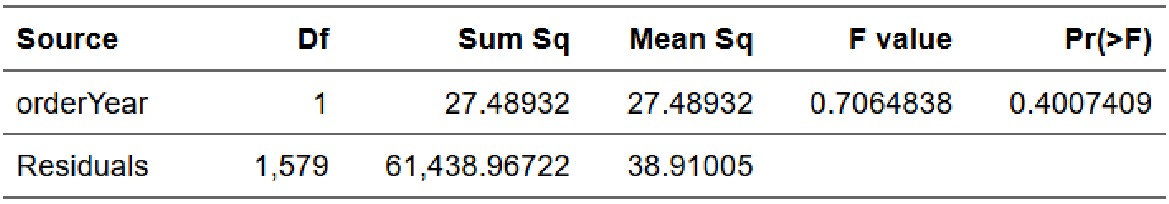


Figure 22: Box plot of different treatment means

Table 10: One-Way ANOVA results for product CLO011 (2026 and 2027)



Source	Df	Sum Sq	Mean Sq	F value	Pr(>F)
orderYear	1	27.48932	27.48932	0.7064838	0.4007409
Residuals	1,579	61,438.96722	38.91005		

The p-value for the year factor is 0.401, which is greater than our chosen significance level of 0.05. Therefore, we fail to reject H_0. This indicates that there is no statistical difference in the mean delivery times for CLO011 between 2022 and 2023. This result is confirmed by the following box plots of delivery times by year for CLO011:

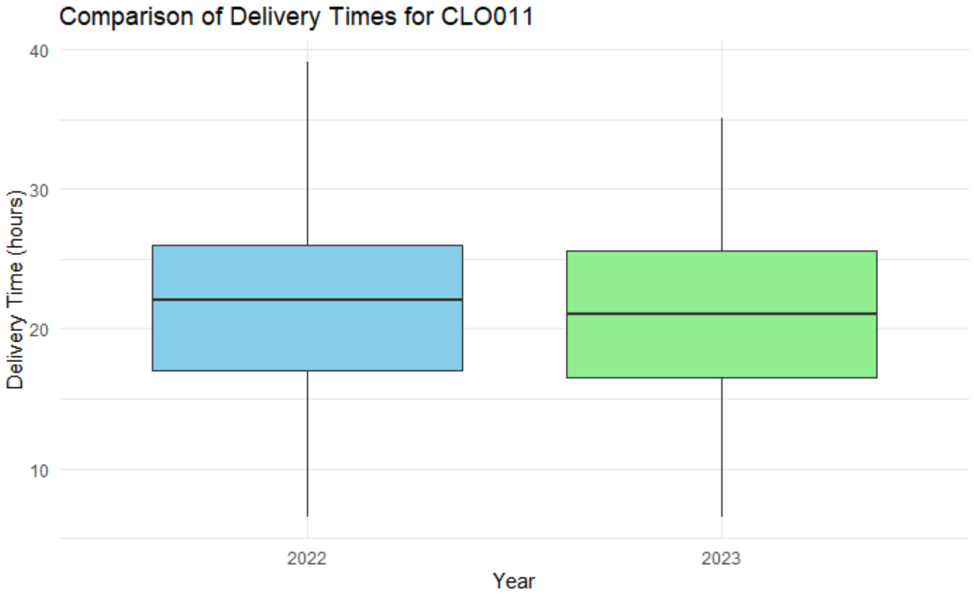


Figure 23: Box plot of delivery times for CLO011

The ox plots confirms that the delivery times by year show overlapping distributions and similar medians. This supports the ANOVA conclusion that the mean of the processes have not changed significantly. The delivery process for CLO011 remained statistically stable across the years 2022 and 2023. This aligns with the SPC analysis in part 3, which showed that most sample means and standard deviations were stable and within the control limits.

7.

7.1

The following graph (Figure 24: Distribution of staff for a car rental agency) shows the number of days with 12 to 16 workers on duty at a car rental agency over a total of 397 recorded days.

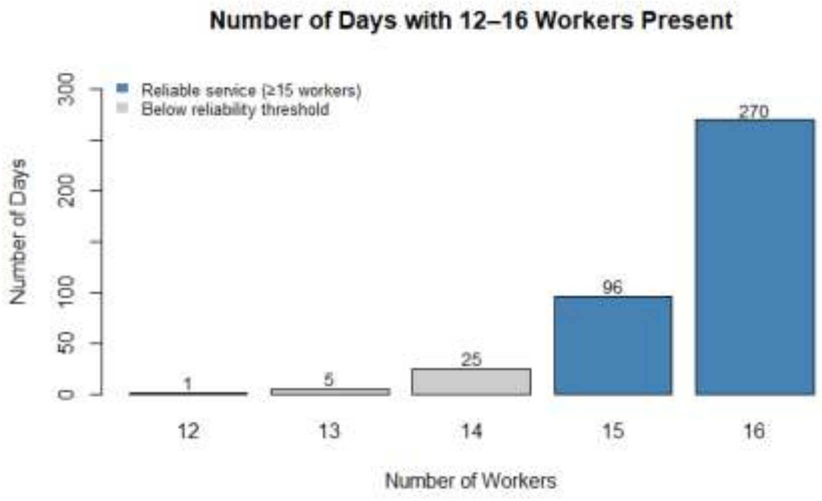


Figure 24: Distribution of staff for a car rental agency

To evaluate service reliability, we assume that having 15 workers on duty ensures that all service counters and support roles are sufficiently staffed. Having enough staff present will minimize delays and ensure customer satisfaction. Days with fewer than 15 workers are considered to deliver a lower level of service reliability.

Using the following criterion:

- Days with 15 workers: 96 days.
- Days with 16 workers: 270 days.
- Total reliable days: 366 days.
- Total days: 397 days.

The expected percentage of reliable service days is therefore

$$Reliability\ Rate = \frac{366}{397} \times 100 = 92.17\%$$

Based on the staffing data, the car rental agency can expect customer satisfaction and reliable service for 92% of days in a year. This indicates that the current staffing levels that were given are generally sufficient, but could be improved by ensuring that staff on duty does not drop below 15 workers.

7.2

To optimize staffing and maximise profit, we model the number of workers who arrive for duty each day as a binomial random variable. Now, if the company schedule S_0 workers on a day, and each worker independently arrives with a probability *p,* the actual number present that day *(K)* follows:

$$K \sim \text{Binomial}(S_0, p).$$

We observed an actual staff count for n = 397 days with the distribution shown in Figure 24. The average number of workers per day is calculated as:

$$\bar{K} = \frac{\sum(k_i \times n_i)}{\text{total days}} = \frac{6187}{397} \approx 15.5844$$

Assuming that 16 workers were scheduled each day ($S_0 = 16$), the expected value of $K$is $E[K] = S_0 p$. The attendance probability can then be estimated as:

$$\hat{p} = \frac{\bar{K}}{S_0} = \frac{15.5844}{16} \approx 0.9740$$

This means that each scheduled employee is present approximately 97.4% of the time. The same calculations were implemented in R and verified these results. To assess precision, the standard error and 95% confidence interval was computed and the following results were obtained: The lower bound of the confidence interval is 0.970, while the upper bound is 0.978. These values indicate very high attendance reliability.

The company experiences service problems when there are fewer than 15 employees on duty, resulting in an average loss of R20 000 in daily sales.

Each additional employee that is scheduled costs R25 000 per month. Therefore, the total monthly cost can be modelled as:

$$\text{Monthly Cost}(S) = (S \times 25{,}000) + [30 \times 20{,}000 \times P(K < 15)],$$

where $P(K < 15)$is the probability that fewer than 15 employees arrive, calculated from the Binomial distribution. This model was implemented in R, and the following results were obtained and are visualized in Figure 25

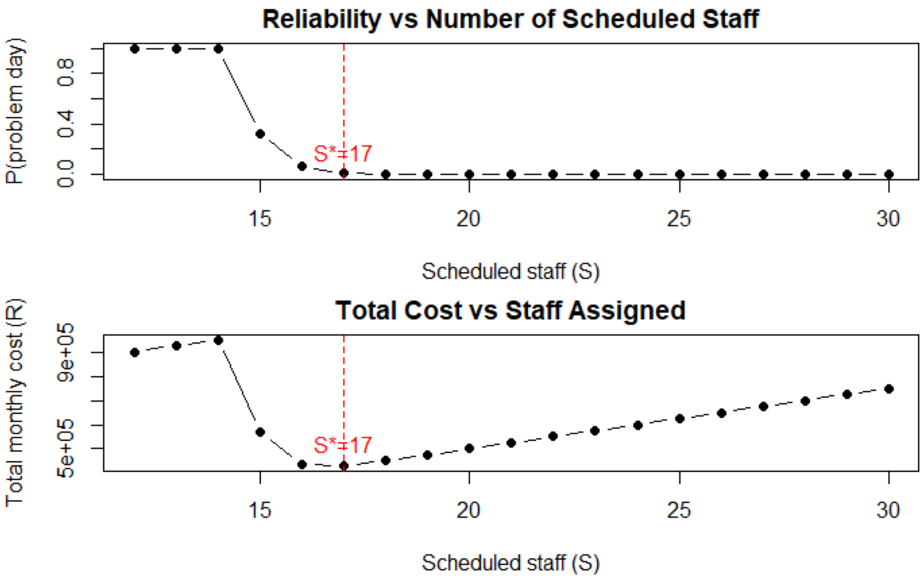


Figure 25: Modelling results for car rental agency.

The optimal number of scheduled workers was found to be S* = 17, with a corresponding reliability rate of 99.1% and an expected monthly total cost of R430 443. The optimal model results reduce the monthly cost by approximately R7 736 per month compared to when only 16 workers were scheduled.

This result implies that there is room for improvement for the car rental company's service reliability at a minimal additional cost. By increasing the scheduled staff members from 16 to 17 employees, the likelihood of there being a day with low service quality decreases from 7% to under 1%. This leads to smoother operations and improved customer satisfaction. In operational terms, scheduling one additional employee acts as a form of "capacity insurance" that ensures that there is consistent service delivery even when attendance fluctuates.

Conclusion

This report was successful at applying fundamental industrial engineering and quality assurance principles to a comprehensive business dataset. The initial analysis revealed data quality issues, specifically the mismatch between the pricing and markup values of the product database and the headoffice database. These data quality issues were systematically corrected and therefore ensure the integrity of all analyses after the correction.

The application of SOC showed that while process variability remained stable for most product, there are products that are frequently out of control. This highlights the need for immediate management investigation into the special causes of variation. Process capability indices confirmed that while non-physical products were highly capable, the delivery processes for physical goods were not capable of meeting the VOC standards of 32 hours.

Furthermore, the optimization models provided clear strategic guidance. For the coffee shop, 6 baristas must be scheduled at both shops to maximize profits, and for the car rental agency, 17 employees must be on duty in a day. The latter recommendation was driven by a binomial model, and it reduced the probability of costly service failures to under 1%, demonstrating the value of predictive capacity planning. Finally, ANOVA confirmed the mean delivery times for product CLO011 did not significantly change between 2022 and 2023. The integrated approach of data correction, process monitoring, risk assessment and profit optimization provides a robust framework for improving quality, reducing operational loss, and supporting data-driven management decision.

References

- ECSA Project Brief. (2025). *Preamble to the Engineering Counsel of South Africa (ECSA) report that proves graduate attribute 4 (ECSA GA4) in 2025* (ProjectECSA2025Final.pdf). Stellenbosch University, Department of Industrial Engineering.
- Stellenbosch University, Department of Industrial Engineering**.** (2025). *Statistical Methods in Quality Assurance Part 1 summary* (Study Material: QA344 Statistics.pdf).